Unsupervised extraction of functional gene expression signatures in the bacterial pathogen *Pseudomonas aeruginosa* 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

While the large sets of publicly available gene expression data contain substantial information about relationships between mRNA expression profiles and genetic background, environment, and cellular state, cross experiment comparisons of public data are challenged by technical noise that masks biological signals. We previously showed that one could reveal biological signatures within compendia of expression data using an unsupervised neural network algorithm, called ADAGE, which excels at detecting patterns in noisy datasets. Here, we show that the generation and integration of multiple ADAGE models, resulting in an ensemble ADAGE (eADAGE), better identified biological pathways. For the bacterium *Pseudomonas aeruginosa*, our analysis found that on the order of 1000 samples were needed to build pathway-level gene expression signatures. The *P. aeruginosa* gene expression compendium contains experiments performed in 78 different media, and we used eADAGE to identify expression signatures associated with medium-type across multiple experiments. We identified a subset of media, including several complex media that were not designed to limit phosphate, in which *P. aeruginosa* exhibited a phosphate starvation response controlled by PhoB. Furthermore, while it was expected that PhoB activates the phosphate starvation response in low phosphate, our analyses found that PhoB was also active in moderate phosphate concentrations and predicted that activity required a second stimulus provided by a sensor kinase, KinB, which was validated in subsequent experiments including a screen of a histidine kinase knock out collection confirmed the specificity of its role in the activation of the Pho regulon. Algorithms that extract biological signal from large collections of public gene expression data, such as eADAGE, can highlight opportunities to discover mechanisms that are currently unrecognized from public data.

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 (Schnoes *et al*, 2013; Gillis & Pavlidis, 2013; Greene & Troyanskaya, 2012). 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 some activation function that determines whether or not it turns on given some input. For the final layer of the network, the output is designed to accomplish some task, and the entire network is trained by grading the quality of the output. The process of training adjusts the edge weight that each node of the network provides to the other. The denoising autoencoder (DA) is a type of unsupervised neural network method in which noise is added to the input data and the neural network is trained to produce the original input at the final layer (Vincent *et al*, 2008). This approach has properties that make it 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. The first step in training a DA is to add random noise to input data. The DA is then trained to process the data through neural network nodes to reconstruct the original data without noise. This step helps a DA to learn features that are 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 and that if we trained multiple models, each would represent a local minimum. In this work, we aim to identify patterns that are stable across many of these neural networks. Yu's work on stability emphasizes the importance of patterns that are stable across statistical models in the process of discovery (Yu, 2013). In this case with gene expression data, each neural network is a statistical model. While run-to-run variability obscures some biological features within individual models, stable cross-model patterns 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 aerguinosa* 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 & 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

Analysis of the effects of model size on pathway characterization of ADAGE models from *P. aeruginosa* gene expression.

Determining the optimal structure of a neural network is challenging. In the case of an ADAGE model with a single hidden layer we term the number of nodes in the hidden layer to be the model size. We evaluated the model size through both a data-driven heuristic and a knowledge-driven heuristic. Importantly, the data-driven heuristic requires no curated pathway information and can be applied even when such resources are unavailable for an organism. During ADAGE training, neural networks are trained to reconstruct the input from data with noise added. The reconstruction error can be used to estimate model sizes that can be supported by the available *P. aeruginosa* gene expression data. The reconstruction error quickly decreases as model size increases and reaches a floor at model size of approximately 300 (Figure 1A). Further increasing model size does not improve reconstruction, suggesting that the available data are insufficient to support larger models.

While ADAGE models are constructed without the use of any curated information, we can use experimentally-derived knowledge of gene functions to provide heuristic information about the number and types of pathways captured by a model and determine how this varies with model size. We define a functional signature learned by an ADAGE model as a set of genes that contribute the highest positive or highest negative weights to a specific node (see methods for detail). Therefore, one node results in two gene signatures, one on each high weight side. These high-weight (HW) genes are often involved in a common biological process as demonstrated by the fact that there is often a statistically significant enrichment in specific KEGG pathways within each signature. For models of different sizes (10-1000 nodes), we determined the number of KEGG pathways significantly associated with at least one gene signature in a model, referred to as KEGG pathway coverage for that model, and found that KEGG pathway coverage increased as model size increased until a model size of approximately 300 (Figure 1B). The number of pathways per node (including pathways associated with both the positive and negative signatures in a node) for all nodes with at least one associated KEGG pathway decreased as model size increased (Figure EV1), suggesting that multiple pathways were grouped in small models and were separated into more discrete features in large models with more nodes. Though this method was unsupervised, we inferred that methods that extracted signatures corresponding to known pathways better captured biological signals in the compendium. Therefore, considering the data-driven and knowledge-driven heuristics together, we identified a 300-node neural network model as most appropriate for the existing *P. aeruginosa* gene expression compendium.

Analysis of the effects of sample number in the training set on ADAGE models

The expression compendium contains 1051 samples from 125 experiments. We aimed to identify the amount of data required to saturate the method's ability to discover biologically supported signatures and to identify how far the compendium could be reduced before performance dropped precipitously. To assess this, we performed a subsampling analysis in which we trained ADAGE models on randomly selected sets of 100, 200, 500, and 800 expression profiles. We examined the number of KEGG pathways associated with at least one gene signature (pathway coverage) as a function of the size of the training set (Figure 1C). In the 50-node models, the size used in (Tan *et al*, 2016b), the average KEGG pathway coverage at each training size increased significantly up to 500 samples (Tukey's HSD adjusted p-values < 0.05 between models trained with 100, 200, and 500 samples), but differences beyond 500 training samples were not significant (Tukey's HSD adjusted p values > 0.05 between models trained with 500, 800, and 1051 samples). For 300-node models, pathway coverage showed significant increases (Figure 1C) between the models constructed with 100, 200, 500, and 800 samples (Tukey's HSD adjusted p-values < 0.05) but not between 800 and 1051 (Tukey's HSD adjusted p-value > 0.05). The slower increase in pathway coverage when sample size is relatively large suggests redundancy in the compendium, potentially due to biological replicates or experiments probing similar processes. This highlights the importance of data that capture diverse processes.

Using the subsampling strategy, we also evaluated the reconstruction error of each model on its training set and a randomly chosen held out test set of 200 samples. As sample size

increased, training reconstruction errors increased slightly while testing reconstruction errors dropped dramatically (Figure 1D). We fitted exponential models between sample size and the differences of training and testing errors (R^2= 0.78 for 50-node models and R^2= 0.83 for 300-node models). We extrapolated from these models to predict that testing errors would approximately match training errors when sample size was 782 for 50-node models and 1076 for 300-node models. These results suggested that smaller models were less sensitive to sample size, likely because they have fewer parameters to fit and also that our 1051 sample compendium was sufficient to train a 300-node model.

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

Individual ADAGE models capture a local minimum, with models of the same size capturing different pathways. This occurs because each ADAGE model is initialized with random weights, and the training processes are sensitive to these 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 from different models, we applied consensus clustering on nodes' weight vectors. Our previous ADAGE analyses showed that HW genes 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 primary 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.

eADAGE models exhibited greater KEGG pathway coverage than those generated by other methods. We evaluated ADAGE, corADAGE, and eADAGE for the number of covered KEGG pathways (Figure 2B). Both corADAGE and eADAGE covered significantly more KEGG pathways than ADAGE (t-test p-value of 1.04e-6 between corADAGE (n=10) and ADAGE (n=1000) and t-test p-value of 1.41e-6 between eADAGE (n=10) and ADAGE (n=1000)). Moreover, eADAGE models covered, on average, 10 more pathways than corADAGE (t-test p-value of 1.99e-3, n=10 for both groups), confirming the critical roles of an ADAGE node's HW gene signatures in defining biological pathways. Genes that participate in multiple pathways can influence pathway enrichment analysis, a factor termed pathway crosstalk (Donato *et al*, 2013). If eADAGE signatures tended to include genes that participated in many pathways, this could also drive the increase in number of observed pathways. To control for this, we performed crosstalk correction (Donato *et al*, 2013). After correction, the total number of covered pathways dropped approximately by half (Figure EV2), but eADAGE still covered significantly more pathways than corADAGE (t-test p-value of 1.60e-3) and ADAGE (t-test p-value of 6.16e-07). These results suggested that eADAGE effectively integrates multiple models to more broadly capture pathway signals embedded in diverse gene expression compendia.

We next evaluated how specifically and completely signatures learned by the models capture known KEGG pathways. We use each gene signature's FDR corrected p-value for enrichment of a KEGG pathway as a combined measure, as this captures both the sensitivity and specificity. If

a pathway was significantly associated with multiple gene signatures in a model, we only considered its most significant association. We found that 71% of pathways were more significantly enriched (had lower median p-values) in corADAGE models (n=10) when compared to individual ADAGE models (n=100) (Figure EV3). This increased to 87% for eADAGE (n=10). We also directly compared eADAGE and corADAGE by this measure and observed that 74% of pathways were more significantly enriched in eADAGE. Our earlier evaluation of pathway-based heuristics showed that different pathways were best captured at different model sizes (Figure EV3). We next compared the 300-node eADAGE model to individual models of each size. Although the 300-node eADAGE models were constructed only from 300-node ADAGE models, we found that 69% pathways were more significantly enriched (i.e. lower median p-values) in eADAGE models than ADAGE models of any size, including those with more nodes than the eADAGE models (Figure EV3). Three example pathways that are best captured either when model size is small, large, or in the middle are all well captured in the 300-node eADAGE model (Figure 2C). These results demonstrate that eADAGE's ensemble modeling procedure captures signals across model sizes more effectively than individual ADAGE and corADAGE models. Thus eADAGE more completely and precisely captures the gene expression signatures of biological pathways.

We designed eADAGE to provide a more robust analysis framework than individual ADAGE models. To assess this, we examined the percentage of models that covered each pathway (coverage rate) between ADAGE and eADAGE (Figure EV4). The pathways covered by each individual ADAGE model were highly variable. Most KEGG pathways were covered by less than half of individual models but more than half of eADAGE models (Figure EV5), suggesting that eADAGE models were more robust than individual ADAGE models. We excluded all pathways always covered by both individual ADAGE and eADAGE models and observed that 72% of the remaining pathways were covered more frequently by eADAGE than ADAGE. This suggests that their associations are stabilized through the ensemble construction procedures. In summary, these comparisons of eADAGE and ADAGE reveal that not only are more pathways captured more specifically, but also those that are captured are captured more consistently.

Principle component analysis (PCA) and independent component analysis (ICA) have been previously used to extract biological features and build functional gene sets (Engreitz *et al*, 2010; Raychaudhuri *et al*, 2000; Gong *et al*, 2007; Alter *et al*, 2000; Raychaudhuri *et al*, 2000; Lutter *et al*, 2009; Frigyesi *et al*, 2006; Chen *et al*, 2008; Roden *et al*, 2006; Ma & Kosorok, 2009). We performed PCA and generated multiple ICA models from the same *P. aeruginosa* expression compendium and evaluated their KEGG pathway coverage following the same procedures used for eADAGE. eADAGE substantially and significantly outperforms PCA in terms of pathway coverage (Figure 2D). We observed that low-order PCs tend to be associated with more pathways than high-order PCs, which is consistent with the higher variance explained by low-order PCs. ICA and eADAGE covered a similar number of pathways at the significance cutoff of FDR 0.05. However, we observed that eADAGE represented KEGG pathways more precisely than ICA. Specifically, among pathways significantly enriched in either approach, 68% pathways exhibited more significant enrichment in eADAGE. Increasing the significance threshold for pathway coverage demonstrates the advantage of eADAGE (Figure 2D).

Pathway databases provide a means to compare unsupervised methods for signature discovery. Not all pathways will be regulated at the transcriptional level, but those that are may be extracted from gene expression data. The unsupervised eADAGE method revealed signatures that corresponded to *P. aeruginosa* KEGG pathways better than PCA, ICA, ADAGE, and corADAGE. It had higher pathway coverage (breadth), covered pathways more specifically (depth), and more consistently than existing methods (robustness).

Elucidating functional signatures that are indicative of growth medium

Analysis of differentially expressed genes is widely used to analyze single experiments, but crosscutting signatures are required to reveal general response patterns from large-scale compendia. Signature-based analyses can suggest mechanisms such as crosstalk and novel regulatory networks. However, in order for this to be effective, these signatures must be robust and comprehensive. By capturing biological pathways more completely and robustly, eADAGE enables the analysis of signatures, including those that don't correspond to any KEGG pathway, across the entire compendium of *P. aeruginosa.*

Gene expression experiments have been used to investigate a diverse set of questions about *P. aeruginosa* biology, and these experiments have used many different media to emphasize different phenotypes. Our manual annotation showed that 78 different base media were used across the gene expression compendium (Table EV1). While the compendium contains 125 different experiments, it is exceedingly rare for investigators to use multiple base lab media within the same experiment. There were only two examples in the entire compendium (Table EV1). Further, other than LB, which is used in 43.6% (458/1051) of the samples in the compendium, most media are only represented by a handful of samples. Comparing samples across the compendium can help shed light on the influence of medium on *P. aeruginosa* even if the medium wasn't widely used.

For biological evaluation, we built a single new eADAGE model with 300 nodes. The model's weight matrix (Table EV2), positive and negative gene signatures for each node (Table EV3), and signature activities for each sample in the compendium (Table EV4) are provided. To compare the *P. aeruginosa* response to the different media in the compendium, we looked for signatures that were differentially active in samples grown in different media and media groups (see methods). Table EV5 lists signatures that are active in a specific medium above a stringent threshold and Table EV6 lists signatures that are most active in a group of media (a complete list of signature-media group associations is in Table EV7).

Distinct aspects of the response to low phosphate are captured among the most active signatures

The two signatures with the highest pan-media activation scores are Node164pos and Node108neg (Table EV6), therefore we examined them to better understand their roles in the transcriptional response to different media. To evaluate the basis for the high activation scores, we examined their underlying activities across all media (Node164pos is shown as an example in Figure 3A). This revealed that they were highly active in King's A medium, Peptone medium,

and NGM+<0.1mM phosphate (NGMlowP), but not in NGM+25mM phosphate (NGMhighP). The difference in their activities between NGMlowP and NGMhighP suggested that these signatures respond to phosphate concentrations. Interestingly, the other two media (Peptone and King's A) in which *P. aeruginosa* gene expression leads to consistently high activities also had low phosphate concentrations (0.4 mM) relative to other media in the compendium. For example, commonly used LB has a phosphate concentration of approximately 4.5 mM (Bertani, 2004) and many others have concentrations above 20 mM.

Consistent with the observation that Node164pos is only active in low phosphate media, a KEGG pathway enrichment analysis of Node164pos genes suggested a strong enrichment in phosphate acquisition related pathways (Table EV6). Many of the phosphate responsive genes in Node164pos are known to be under the control of PhoB, a transcription factor in the PhoR-PhoB two-component system that responds to conditions of low phosphate in *P. aeruginosa* (Santos-Beneit, 2015; Blus-Kadosh *et al*, 2013; Bielecki *et al*, 2015), and *phoB* itself is in this signature. Enrichment analysis using a previously defined PhoB regulon (see methods) showed that this signature has the largest overlap with the PhoB regulon in comparison to all other gene signatures learned by this eADAGE model (FDR q-value of 8.1e-29 in hypergeometric test). The transcript levels of genes in Node164pos are higher in Peptone, King's A, and NGMlowP medium relative to the other samples in the compendium including NGMhighP (Figure 3B).

Among the highest weight genes in Node164pos is a gene that encodes alkaline phosphatase (PhoA), an enzyme with an activity that can be easily measured using a colorimetric assay. As expected, PhoA activity (blue color) was high when *P. aeruginosa* was grown on NGMlowP and not when grown on NGMhighP (Figure 4A). The same trend was observed in another medium, MOPS, with the same low and high phosphate concentrations. Also consistent, PhoA was not active on the phosphate-replete medium LB. Although King's A and Peptone are not considered phosphate-limited media, their phosphate concentrations were low enough to provoke PhoA activity, as predicted by Node164pos's signature-medium relationship (Figure 4B). Furthermore, PhoA activity was dependent on PhoB and the PhoB-activating histidine kinase PhoR, which is consistent with previous publications (Bielecki *et al*, 2015). These results provide striking evidence that low phosphate media, including Peptone and King's A, induced PhoB activity as predicted by the eADAGE analysis and previous characterizations of the *P. aeruginosa* phosphate response.

Unlike Node164pos, Node108neg is not characterized by KEGG pathways associated with phosphate starvation (Table EV6). Of the thirty-two genes in Node108neg, more than half of the genes are shared with Node164pos, suggesting a relationship between the two signatures. Node108neg was also enriched of PhoB regulated genes (FDR q-value of 5.2e-9 in hypergeometric test), but contains a much narrower set of PhoB regulon. Of the PhoB-regulated genes present in Node108neg, we found almost all of the genes (six of seven) that were also regulated by TctD, a transcriptional repressor described by Haussler and colleagues (Bielecki et al, 2015). TctD binding to these promoters prevents their activation by PhoB. Therefore, we predicted that Node108neg represented a group of PhoB-activated genes that were also negatively regulated by TctD. This is consistent with the observation that

Node108neg was the most differentially active signature in the RNA-Seq experiment that compares Δ*tctD* and the wild type (E-GEOD-64056). eADAGE learned the relationship between PhoB and TctD directly from an expression compendium that did not contain any samples of *tctD* mutants, suggesting that the signature was derived from samples with differences in TctD activity. This demonstrates the value of KEGG-independent eADAGE signatures for the analysis of large-scale compendia.

We evaluated whether the PhoB and TctD signals were also extracted by PCA, ICA, or ADAGE using ten models constructed by each method except PCA, which is deterministic and produces a single model models. The PhoB regulon was captured with less fidelity by ICA and ADAGE than by eADAGE, as reflected by the smaller overlap with the PhoB regulon (Table EV8). PCA captured a strong PhoB signal in its 19th principle component. However, it did not learn the subtler TctD signal. Component19pos was the only signature that was highly active in the low phosphate media group (Table EV9). In summary, the other methods were able to capture some of this signature but in a manner that was less complete or failed to separate TctD.

Cross-compendium analysis of Node164pos activity reveals a role for the histidine kinase KinB in the regulation of PhoB

Interestingly, Node164pos activity exhibited a wide spread in PIA medium (0.8 mM phosphate), with six samples having high activities and the other six samples having low activities (Figure 3A). Each set of six samples came from a different study. All of the strains in the first sample set, published in 2012, in which Node164pos was low used a PAO1 *kinB*::Gm^{R} mutant background (Damron *et al*, 2012). The second, published in 2013, used a PAO1 strain with *kinB* intact and showed high Node164pos activity on PIA (Damron *et al*, 2013). The fact that *kinB* mutant samples in PIA show significantly lower Node164pos activity suggests that KinB may be a regulator of PhoB on PIA.

As predicted by Node164pos activity, PhoA activity was evident and was KinB dependent on PIA medium (Figure 4B). Notably, PhoA activity was still dependent on PhoB and PhoR as it was on Peptone and King's A. Over time, the Δ*phoR* mutant developed PhoA activity on all three media, but the Δ*kinB* mutant on PIA did not (Figure 4C). Recovery of PhoA activity in the Δ*phoR* mutant suggests that there are PhoR-independent paths for PhoB activation. The co-dependence on KinB and PhoR suggest that these kinases do not perform redundant functions but rather regulate PhoB in conjunction with each other. To determine if the deletion of kinases non-specifically altered PhoB activation, we screened 63 in-frame deletion mutants each lacking a histidine kinase (Table EV9) to identify mutants altered in PhoA activation on PIA. Other than Δ*phoR*, Δ*kinB* was the only strain lacking PhoA activity on NGMlowP. This suggests that PhoA activity is not altered by changes in generic kinase activity but rather the interaction with KinB is specific.

Notably, although PIA, Peptone and King's A contain low phosphate levels compared to rich media such as LB, PIA has a higher phosphate concentration (0.8mM) than do peptone and King's A (0.4mM). To test whether the moderately low level of phosphate in PIA provokes KinB regulation of PhoA, we conducted a titration experiment in MOPS minimal medium. PhoA

activity was consistent with that on NGM and other media. Further, while the PA14 wild type (WT) showed PhoA activity at 0.5 mM, Δ*kinB* did not (Figure 4D). This shows KinB regulation of PhoB in moderately low phosphate and not in lower phosphate. Noting that PIA has a phosphate concentration of 0.8 mM and our titration experiment in MOPS showed *kinB*-sensitivity at 0.5 mM, it is likely that the precise concentration at which KinB regulates PhoA depends on the background medium. Node164pos activity across the compendium suggested, and experimental evidence confirmed, that KinB regulates PhoA at moderately low phosphate levels. To our knowledge, KinB has not been previously implicated in the activation of PhoB, although cross-compendium eADAGE analysis of two PIA experiments was capable of revealing this relationship.

In summary, eADAGE effectively extracted biologically meaningful features, accurately indicated their activity in multiple media spanning numerous independent experiments, and revealed a novel regulatory mechanism. By summarizing gene-based expression information into biologically relevant features, eADAGE greatly simplifies analyses that cut across large gene expression compendia.

Discussion

For unsupervised approaches, it has been very challenging to determine the appropriate model complexity. In supervised learning problems, predictors can be assessed through cross-validation accuracies. To our knowledge, there is not yet a similar well-established approach to estimate appropriate model size for unsupervised feature construction of gene expression data. Here we develop heuristics that target two aspects of the problem: the model needs to be well supported by the amount of available data and the extracted features should well resemble known biological processes. Our data-driven heuristics can be applied to organisms for which gene-process annotations are lacking. We expect that additional data will support larger models, especially data that measure experimental conditions that are not tested in the existing compendium.

We also contribute a novel eADAGE algorithm. This algorithm combines multiple ADAGE models into one ensemble model to address model variability due to stochasticity and local minima. The algorithm is inspired by consensus clustering, which reconciles the differences in cluster assignments in multiple runs. Comparable approaches have also been applied for ICA, where researchers have used the centrotypes of multiple ICA models as the final model (Frigyesi *et al*, 2006). The ICA centrotype approach for ADAGE corresponds to corADAGE, and our comparison of eADAGE and corADAGE shows that eADAGE not only covers more biological pathways, but also results in cleaner representations of biological pathways. The results of our direct comparison suggest that placing particular emphasis on the genes most associated with a particular signature may be a useful property for other unsupervised feature construction algorithms in biology. While we believe the eADAGE algorithm can help improve the biological interpretability of neural networks, it should be clear that it is not designed for increasing prediction accuracies in supervised learning problems.

Our eADAGE method revealed patterns that were detectable from a large data compendium containing experiments performed in 78 different media, but that were not necessarily evident in individual experiments. For example, our cross-compendium analysis of expression patterns specific to certain media determined those media in which *P. aeruginosa* had high PhoB activity. PhoB is a global regulator, and thus an awareness of the different states of the PhoB regulon in different media will likely provide important insight into medium-specific phenotypes. For example, King's A and PIA are known to stimulate robust production of colorful secondary metabolites (King *et al*, 1954) called phenazines and separate studies have shown that PhoB can also influence phenazine levels (Jensen *et al*, 2006). Future studies will reveal whether or not the low phosphate levels in these media contribute to this characteristic phenotype.

Using eADAGE, we uncovered a subtle aspect of the phosphate starvation response that depends on a histidine kinase not previously associated with PhoB. Bacteria have evolved many mechanisms to insulate response pathways from each other (Podgornaia & Laub, 2013), but cross-talk, wherein a sensor kinase from one pathway phosphorylates a response regulator from another, has been suggested to explain the complexity of signaling networks, including that of PhoB (Ninfa *et al*, 1988; Fisher *et al*, 1995). Thus, it is possible that both PhoR and KinB directly activate PhoB. While there are many examples of one kinase partially compensating when the cognate kinase is absent, it is more challenging to find conditions where two kinases are needed for full response regulator activation (Verhamme *et al*, 2002). Alternatively, KinB may influence PhoB activity indirectly by regulating activities that affect PhoB levels, phosphorylation state, or protein-protein interactions. Future work will further address the mechanistic aspects of this model. 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. Without the eADAGE analysis across multiple experiments in different media, we would not have found this nuanced mechanism.

The eADAGE model also includes a signature for PhoB interactions with another transcriptional regulator, TctD. While this node was differentially active in samples grown on PIA, Peptone or King's A agar due to the presence of multiple PhoB-regulated genes in its signature, we do not have any evidence that TctD is involved in the PhoB responses studied here. Future studies may leverage subtle and previously unrecognized differences in Node108neg activity in published datasets to learn more about the range of conditions in which TctD-PhoB interactions are most relevant.

In the 300-node eADAGE model used for medium analysis, 81% genes in *P. aeruginosa* PAO1 genome are included in at least one gene signature. Among all signatures, 59% have a KEGG pathways associated with them. Among the remaining 41% of signatures, 17 of them had ten or fewer genes and may not have associated with KEGG pathways for this reason. Some others were associated with known pathways that have not yet been annotated in KEGG, such as Node174pos which encodes genes encoded by the P2 phage or Node150neg which contains

genes involved in the formation of surface associated communities referred to as biofilms. Thus eADAGE can group functionally linked genes in ways that may facilitate the identification or annotation of pathways.

There are now abundant public gene expression data. 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 that robustly integrate these diverse datasets in a manner that is not tied to only aspects of biology that are well understood. We expect that robust unsupervised data integration methods, like eADAGE, will play a key role in this process.

Materials and Methods

Data processing

We followed the same procedures for data collection, processing, and normalization from (Tan *et al*, 2016b) and updated the *P. aeruginosa* gene expression compendium to include newly uploaded datasets on GPL84 platform from the ArrayExpress database (Rustici *et al*, 2013) on 31 July 2015. The updated *P. aeruginosa* compendium contains 125 datasets with 1051 individual genome-wide assays. Processed expression values of the Δ*tctD* RNAseq dataset were downloaded from ArrayExpress (E-GEOD-64056) and normalized to the range of the compendium using TDM (Thompson *et al*, 2016). We provide the *P. aeruginosa* expression compendium (Dataset EV1) along with all the code used in this paper (Tan *et al*, 2016a). The eADAGE repository is also tracked under version control at https://bitbucket.org/greenelab/eadage.

Construction of ADAGE models

We constructed ADAGE models as described in (Tan *et al*, 2016b). To summarize the process and outputs, we constructed a denoising autoencoder for the gene expression compendium. Denoising autoencoders model the data in a lower dimension than the input space, and the models are trained with random gene expression measurements set to zero. Thus an ADAGE model must learn gene-gene dependencies to fill in this missing information. Once the ADAGE model is trained, each node in the hidden layer contains a weight vector. These positive and negative weights represent the strength of each gene's connection to that node.

Gene signatures as sign-specific high-weight gene sets

In previous work (Tan *et al*, 2016b) we defined high-weight (HW) genes as those in the extremes of the weight distribution on the positive or negative side of a node. Here, we use a more granular definition that accounts for sign specificity. Each node's gene weights are approximately normal and centered at zero in ADAGE models (Tan *et al*, 2015, 2016b). We defined positive HW genes as those that were more than 2.5 standard deviations from the mean on the positive side, and negative HW genes as those that were more than 2.5 standard deviations from the mean on the negative side. After this split, a model with n nodes provides 2n gene signatures. Because a node is simply named by the order that it occurs in a model, we named two gene signatures derived from one node as "NodeXXpos" and "NodeXXneg".

KEGG pathway enrichment analysis

To evaluate the biological relevance of gene signatures extracted by an ADAGE model, we tested how they relate to known KEGG pathways (Kanehisa, 2000). We tested a signature's association with each KEGG pathway using hypergeometric test and corrected the p-value by the number of KEGG pathways we tested following the Benjamini–Hochberg procedure. We used a false discovery rate of 0.05 as the significance cutoff.

Genes can be annotated to multiple pathways. To control for this effect in our analysis, we also performed a parallel analysis after applying crosstalk correction as described in (Donato *et al*, 2013). This approach uses expectation maximization to map each gene to the pathway in which it has the greatest predicted impact. A gene-to-pathway membership matrix, defined using KEGG pathway annotations, initially makes the assumption that each gene's role in all of its assigned pathways remains constant independent of context. We then applied pathway crosstalk correction using genes' weights for each node in the ADAGE model. We used the expectation maximization algorithm to maximize the log-likelihood of observing the membership matrix given each node's weight vector. This process inferred an underlying gene-to-pathway impact matrix and iteratively estimated the probability that a particular gene g contributed the greatest fraction of its impact to some pathway P. Upon convergence, we assigned each gene to the pathway in which it had the maximum impact. The resulting pathway definitions do not share genes. We then used these corrected definitions for an analysis parallel to the KEGG process described above.

Reconstruction error calculation

The training objective of ADAGE is to take a sample with added noise and return the originally measured expression values. The error between the reconstructed data and the initial data is the 'reconstruction error.' To summarize the difference over all genes we used cross-entropy between the original sample and the reconstruction, which has been widely used with these methods and in this domain (Vincent *et al*, 2008; Tan *et al*, 2016b). This matches the statistic used during training of the model. To calculate reconstruction error for a model, we use the mean reconstruction error across samples.

Model size and sample size heuristics

One important parameter of a denoising autoencoder model is the number of nodes in the hidden layer, which we refer to as the model size. To evaluate the impact of model size and choose the most appropriate size, we built 100 ADAGE models at each model size of 10, 50, 100, 200, 300, 500, 750, and 1000, using different random seeds. The random seed determines the initialization statuses of the weight matrix and bias vectors in ADAGE construction and thus different random seeds will result in training stopped at different local minimums. Other training parameters were kept the same and set to the values identified as suitable for a gene expression compendium (Tan *et al*, 2015). In total, 800 ADAGE models with 100 at each model size were generated in the model size evaluation experiment.

To evaluate the impact of sample size on the performance of ADAGE models, we randomly generated subsets of the *P. aeruginosa* expression compendium with sample size of 100, 200,

500, and 800. We then trained 100 ADAGE models at each sample size, each with a different combination of 10 different random subsets and 10 different random training initializations. To evaluate each model, we randomly selected 200 samples not used during training as its testing set. We performed this subsampling analysis at model size 50 and 300. In total, 800 ADAGE models were built in the sample size evaluation experiment.

Construction of eADAGE models

We constructed ensemble ADAGE (eADAGE) models by combining many individual ADAGE models in to a single model. For each eADAGE model we combined 100 individual ADAGE models. The 100 models were trained with identical parameters but distinct random seeds. For an eADAGE model of size 300, we trained 100 individual models with 300 nodes each, which provided 30000 total nodes. Each node has a weight vector. We have previously observed that high-weight genes provided the most information to each node (Tan *et al*, 2016b), so we calculated a weighted Pearson correlation between each node's weight vectors. Our weighted Pearson correlation used (|node1 weight|+|node2 weight|)/2 as the weight function for each gene. We compared this to an unweighted Pearson correlation (corADAGE) as well a baseline ADAGE model.

After calculating correlation (weighted for eADAGE and unweighted for corADAGE), we converted the correlation to distance by calculating (1- correlation)/2. This provided a 30000*30000 distance matrix storing distances between every two nodes. We clustered this distance matrix using the Partition Around Medoids (PAM) clustering algorithm (Park & Jun, 2009).We implemented clustering in R using the ConsensusClusterPlus package (Wilkerson & Hayes, 2010) from Bioconductor with the ppam function from Sprint package to perform parallel PAM (Piotrowski *et al*, 2011). We set the number of clusters to match the individual ADAGE model (e.g. 300) allowing for direct comparison between the eADAGE and ADAGE methods.

Clustering assigned each node to a cluster ranging from 1 to 300. We combined nodes assigned to the same cluster by calculating the average of their weight vectors. These 300 averaged vectors formed the weight matrix of the eADAGE model. Because the ensemble model is built from the weight matrices of individual models, it does not have the parameters that form the bias vectors. We built 10 eADAGE and 10 corADAGE models from 1000 ADAGE models with each ensemble model built upon 100 different individual models. The individual eADAGE model used for biological analysis in this work was constructed with random seed 123, which was arbitrarily chosen before model construction and evaluation.

PCA and ICA model construction

We constructed PCA and ICA models and defined each model's weight matrix following the same procedures in (Tan *et al*, 2016b). To compare with the 300-node eADAGE, we generated models of matching size (300 components). For ICA, we evaluated 10 replicates. PCA provides a single model. PCA and ICA models were evaluated through the KEGG pathway enrichment analysis described above.

Media annotation of the *P. aeruginosa* compendium

A team of *P. aeruginosa* biologists annotated the media for all samples in the compendium by referring to information associated with each sample in the ArrayExpress (Rustici *et al*, 2013) and/or GEO (Edgar, 2002) databases and along with the original publication, if reported. Each sample was annotated by two curators separately. Conflicting annotations, if they occurred, were resolved by a third curator. The media annotation for all samples in the compendium were provided in Table EV1.

Activity calculation for a gene signature

We calculated a signature's activity for a specific sample as $A = W \cdot E/N$, in which W is the weight vector of genes in that signature, E is a vector of genes' expression values after zero-one normalization in that sample, and N is the number of genes. It can be viewed as an averaged weighted sum of genes' expression levels. We normalized a signature's activity by the number of genes (N) in that signature, because different signatures have different number of genes.

Identification of signatures activated across media

We calculated an activation score to identify gene signatures with dramatically elevated or reduced activity in a specific medium. We grouped samples by their medium annotation. For each gene signature and medium combination, we calculated the absolute difference between the mean activity of the signature for samples in that medium as well as the mean activity across the remainder of samples in the compendium. We divided this difference in the means by the range of activity for all samples across the compendium. This score captures the proportion by which the mean activity in a medium differs relative to the total difference across the compendium. We termed this ratio the activation score.

To identify the most specifically active signatures for each medium, we constructed a table for all pairs with an activation score greater than or equal to 0.4 (Table EV5). This was highly stringent: it captured only the top 2.4% of the potential signature-medium pairs. To identify pan-media signatures, we limited signatures to those that were active in multiple media (greater or equal to 0.4) and averaged their activation scores (Table EV7). These signatures exhibit parallel patterns for multiple media across multiple distinct experiments.

Definition of the PhoB regulon

A PhoB regulon for the PAO1 genome was adapted from the PhoB regulon of PA14 in (Bielecki *et al*, 2015) in order to be comparable to models built with PAO1 genome. Of the 187 genes in the PA14 regulon, 160 were in the PAO1 reference genome (www.pseudomonas.com).

Strains and Media

Strains used were WT, Δ*phoB* (DH2633, O'Toole lab collection), Δ*phoR* (DH2516) and Δ*kinB* (DH2517), all in the PA14 background. All strains were maintained on LB with 1.5% agar and grown at 37 °C. For cross-media and phosphate concentration comparisons, BCIP assays (see methods below) were conducted on different base media with 1.5% agar (Fisher): King's A (Pancreatic Digest of Gelatin (Difco) 20g/L; $MgCl_2$ 1.4g/L; K_2SO_4 10g/L; Glycerol 10ml/L) (King *et al*, 1954), LB (Tryptone (Fisher) 10g/L; Yeast Extract (Fisher) 5g/L; NaCl 5g/L) (Bertani, 2004),

MOPS (morpholinepropanesulfonic acid 40mM; Glucose 20 ml/L; K_2SO_4 2.67mM; K_2HPO_2 0mM, 25mM or 0.1 – 1 mM) (Neidhardt *et al*, 1974), NGM (Pancreatic Digest of Gelatin 2.5g/L; Cholesterol 5mg/L; NaCl 3g/L; $MgSO_4$ 1mM; $CaCl_2$ 1mM; KCl 25mM; Potassium Phosphate buffer pH6 0 or 25 mM) (Zaborin *et al*, 2009), Peptone (Pancreatic Digest of Gelatin 10g/L; $MgSO_4$ 1.5g/L; K_2SO_4 10g/L) (Lundgren *et al*, 2013), Pseudomonas Isolation Agar (PIA, prepared as per instructions, BioWorld).

BCIP assay

Various media were supplemented with 5-bromo-4-chloro-3-indolyl phosphate (BCIP) DMF solution to a final concentration of 60 µg/mL. BCIP assay plates were inoculated with 5 µl of overnight *P. aeruginosa* culture in LB broth. Colonies were grown for 16 hours at 37 °C then matured at room temperature until imaging. Images were collected 16 and 32 hours post inoculation.

Screen of a histidine kinase mutant collection

Molecular techniques to construct the histidine kinase (HK) knock out collection were carried out as previously described (Ha *et al*, 2014). For each strain in the HK collection, a BCIP assay was performed on PIA. Plates were struck with an overnight *P. aeruginosa* culture concentrated two-fold by centrifugation. Plates were incubated at 37 °C 12-16 hours and matured at room temperature for an additional 12-16 hours alkaline phosphatase activity was determined qualitatively, based on blue color.

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 acknowledge Gregory Way and René Zelaya for helpful code review.

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 constructed and contributed the histidine kinase knock out collection. JT, GD, KMC, DAH and CSG wrote the manuscript, and KAL, CEP, KMC, KD, BP and MTL provided critical feedback.

Conflict of interest

The authors have no conflicts of interest to report.

Reference

Alter O, Brown PO & Botstein D (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. U. S. A.* **97:** 10101–6

Bengio Y, Courville A & Vincent P (2013) Representation Learning: A Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **35:** 1798–1828

723 Bertani G (2004) Lysogeny at mid-twentieth century: P1, P2, and other experimental systems. *J.*
724 *Bacteriol.* **186:** 595–600
725 Bielecki P, Jensen V, Schulze W, Gödeke J, Strehmel J, Eckweiler D, Nicolai T, Bielecka A, Wille T,
726 Gerlach RG & Häussler S (2015) Cross talk between the response regulators PhoB and TctD
727 allows for the integration of diverse environmental signals in *Pseudomonas aeruginosa*.
728 *Nucleic Acids Res.* **43:** 6413–25
729 Blus-Kadosh I, Zilka A, Yerushalmi G & Banin E (2013) The effect of *pstS* and *phoB* on quorum
730 sensing and swarming motility in *Pseudomonas aeruginosa*. *PLoS One* **8:** e74444
731 Chen L, Xuan J, Wang C, Shih I-M, Wang Y, Zhang Z, Hoffman E, Clarke R, Devore J, Peck R,
732 Tusher V, Tibshirani R, Chu G, Storey J, Xiao W, Leek J, Tompkins R, Davis R, Conesa A,
733 Nueda M, et al (2008) Knowledge-guided multi-scale independent component analysis for
734 biomarker identification. *BMC Bioinformatics* **9:** 416
735 Damron FH, Barbier M, McKenney ES, Schurr MJ & Goldberg JB (2013) Genes required for and
736 effects of alginate overproduction induced by growth of *Pseudomonas aeruginosa* on
737 Pseudomonas isolation agar supplemented with ammonium metavanadate. *J. Bacteriol.*
738 **195:** 4020–36
739 Damron FH, Owings JP, Okkotsu Y, Varga JJ, Schurr JR, Goldberg JB, Schurr MJ & Yu HD (2012)
740 Analysis of the *Pseudomonas aeruginosa* regulon controlled by the sensor kinase KinB and
741 sigma factor RpoN. *J. Bacteriol.* **194:** 1317–30
742 Donato M, Xu Z, Tomoiaga A, Granneman JG, Mackenzie RG, Bao R, Than NG, Westfall PH,
743 Romero R & Draghici S (2013) Analysis and correction of crosstalk effects in pathway
744 analysis. *Genome Res.* **23:** 1885–93
745 Edgar R (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data
746 repository. *Nucleic Acids Res.* **30:** 207–210
747 Engreitz JM, Daigle BJ, Marshall JJ & Altman RB (2010) Independent component analysis: mining
748 microarray data for fundamental human gene expression modules. *J. Biomed. Inform.* **43:**
749 932–44
750 Fisher SL, Jiang W, Wanner BL & Walsh CT (1995) Cross-talk between the histidine protein
751 kinase VanS and the response regulator PhoB. Characterization and identification of a VanS
752 domain that inhibits activation of PhoB. *J. Biol. Chem.* **270:** 23143–9
753 Frigyesi A, Veerla S, Lindgren D, Höglund M, Quackenbush J, Jutten C, Herault J, Chiappetta P,
754 Roubaud M, Torrésani B, Hyvärinen A, Oja E, Liebermeister W, Lee S, Batzoglou S,
755 Martoglio A, Miskin J, Smith S, MacKay D, Saidi S, et al (2006) Independent component
756 analysis reveals new and biologically significant structures in microarray data. *BMC*
757 *Bioinformatics* **7:** 290
758 Gillis J & Pavlidis P (2013) Assessing identity, redundancy and confounds in Gene Ontology
759 annotations over time. *Bioinformatics* **29:** 476–82
760 Gong T, Xuan J, Wang C, Li H, Hoffman E, Clarke R & Wang Y (2007) Gene module identification
761 from microarray data using nonnegative independent component analysis. *Gene Regul.*
762 *Syst. Bio.* **1:** 349–63
763 Greene CS & Troyanskaya OG (2012) Accurate evaluation and analysis of functional genomics
764 data and methods. *Ann. N. Y. Acad. Sci.* **1260:** 95–100
765 Ha D-G, Richman ME & O'Toole GA (2014) Deletion mutant library for investigation of
766 functional outputs of cyclic diguanylate metabolism in *Pseudomonas aeruginosa* PA14.

767 *Appl. Environ. Microbiol.* **80:** 3384–93
768 Jensen V, Lons D, Zaoui C, Bredenbruch F, Meissner A, Dieterich G, Munch R & Haussler S (2006)
769 RhlR Expression in Pseudomonas aeruginosa Is Modulated by the Pseudomonas Quinolone
770 Signal via PhoB-Dependent and -Independent Pathways. *J. Bacteriol.* **188:** 8601–8606
771 Jiang Y, Oron TR, Clark WT, Bankapur AR, D'Andrea D, Lepore R, Funk CS, Kahanda I, Verspoor
772 KM, Ben-Hur A, Koo DCE, Penfold-Brown D, Shasha D, Youngs N, Bonneau R, Lin A,
773 Sahraeian SME, Martelli PL, Profiti G, Casadio R, et al (2016) An expanded evaluation of
774 protein function prediction methods shows an improvement in accuracy. *Genome Biol.* **17:**
775 184
776 Kanehisa M (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28:**
777 27–30
778 KING EO, WARD MK & RANEY DE (1954) Two simple media for the demonstration of pyocyanin
779 and fluorescin. *J. Lab. Clin. Med.* **44:** 301–7
780 Lundgren BR, Thornton W, Dornan MH, Villegas-Peñaranda LR, Boddy CN & Nomura CT (2013)
781 Gene PA2449 is essential for glycine metabolism and pyocyanin biosynthesis in
782 *Pseudomonas aeruginosa* PAO1. *J. Bacteriol.* **195:** 2087–100
783 Lutter D, Langmann T, Ugocsai P, Moehle C, Seibold E, Splettstoesser WD, Gruber P, Lang EW &
784 Schmitz G (2009) Analyzing time-dependent microarray data using independent
785 component analysis derived expression modes from human macrophages infected with *F.*
786 *tularensis holartica*. *J. Biomed. Inform.* **42:** 605–611
787 Ma S & Kosorok MR (2009) Identification of differential gene pathways with principal
788 component analysis. *Bioinformatics* **25:** 882–9
789 Monti S, Tamayo P, Mesirov J & Golub T (2003) Consensus clustering: a resampling-based
790 method for class discovery and visualization of gene expression microarray data. *Mach.*
791 *Learn.* **52:** 91–118
792 Neidhardt FC, Bloch PL & Smith DF (1974) Culture medium for enterobacteria. *J. Bacteriol.* **119:**
793 736–47
794 Ninfa AJ, Ninfa EG, Lupas AN, Stock A, Magasanik B & Stock J (1988) Crosstalk between bacterial
795 chemotaxis signal transduction proteins and regulators of transcription of the Ntr regulon:
796 evidence that nitrogen assimilation and chemotaxis are controlled by a common
797 phosphotransfer mechanism. *Proc. Natl. Acad. Sci. U. S. A.* **85:** 5492–6
798 Park H-S & Jun C-H (2009) A simple and fast algorithm for K-medoids clustering. *Expert Syst.*
799 *Appl.* **36:** 3336–3341
800 Piotrowski M, Forster T, Dobrezelecki B, Sloan TM, Mitchell L, Ghazal P, Mewsissen M, Petrou S,
801 Trew A & Hill J (2011) Optimisation and parallelisation of the partitioning around medoids
802 function in R. In *2011 International Conference on High Performance Computing &*
803 *Simulation* pp 707–713. IEEE
804 Podgornaia AI & Laub MT (2013) Determinants of specificity in two-component signal
805 transduction. *Curr. Opin. Microbiol.* **16:** 156–62
806 Raychaudhuri S, Stuart JM & Altman RB (2000) Principal components analysis to summarize
807 microarray experiments: application to sporulation time series. *Pac. Symp. Biocomput.*:
808 455–66
809 Roden JC, King BW, Trout D, Mortazavi A, Wold BJ, Hart CE, Tavazoie S, Hughes J, Campbell M,
810 Cho R, Church G, Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E,

Lander E, Golub T, Eisen M, et al (2006) Mining gene expression data by interpreting principal components. *BMC Bioinformatics* **7:** 194

Rustici G, Kolesnikov N, Brandizi M, Burdett T, Dylag M, Emam I, Farne A, Hastings E, Ison J, Keays M, Kurbatova N, Malone J, Mani R, Mupo A, Pedro Pereira R, Pilicheva E, Rung J, Sharma A, Tang YA, Ternent T, et al (2013) ArrayExpress update--trends in database growth and links to data analysis tools. *Nucleic Acids Res.* **41:** D987-90

Santos-Beneit F (2015) The Pho regulon: a huge regulatory network in bacteria. *Front. Microbiol.* **6:** 402

Schnoes AM, Ream DC, Thorman AW, Babbitt PC & Friedberg I (2013) Biases in the experimental annotations of protein function and their effect on our understanding of protein function space. *PLoS Comput. Biol.* **9:** e1003063

Tan J, Doing G, Lewis KA, Price CE, Chen KM, Cady KC, Perchuk B, Laub MT, Hogan DA & Greene CS (2016a) eADAGE-1.0.0rc2. *Zenodo*

Tan J, Hammond JH, Hogan DA & Greene CS (2016b) ADAGE-Based Integration of Publicly Available *Pseudomonas aeruginosa* Gene Expression Data with Denoising Autoencoders Illuminates Microbe-Host Interactions. *mSystems* **1:** e00025-15

Tan J, Ung M, Cheng C & Greene CS (2015) Unsupervised feature construction and knowledge extraction from genome-wide assays of breast cancer with denoising autoencoders. *Pac. Symp. Biocomput.* **20:** 132–43

Thompson JA, Tan J & Greene CS (2016) Cross-platform normalization of microarray and RNA-seq data for machine learning applications. *PeerJ* **4:** e1621

Verhamme DT, Arents JC, Postma PW, Crielaard W & Hellingwerf KJ (2002) Investigation of in vivo cross-talk between key two-component systems of *Escherichia coli*. *Microbiology* **148:** 69–78

Vincent P, Larochelle H, Bengio Y & Manzagol P-A (2008) Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine learning - ICML '08* pp 1096–1103. New York, New York, USA: ACM Press

Wanner BL & Chang BD (1987) The phoBR operon in Escherichia coli K-12. *J. Bacteriol.* **169:** 5569–74

Wilkerson MD & Hayes DN (2010) ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* **26:** 1572–3

Yu B (2013) Stability. *Bernoulli* **19:** 1484–1500

Zaborin A, Romanowski K, Gerdes S, Holbrook C, Lepine F, Long J, Poroyko V, Diggle SP, Wilke A, Righetti K, Morozova I, Babrowski T, Liu DC, Zaborina O & Alverdy JC (2009) Red death in *Caenorhabditis elegans* caused by *Pseudomonas aeruginosa* PAO1. *Proc. Natl. Acad. Sci. U. S. A.* **106:** 6327–32

Figure Legends

Figure 1: Knowledge- and data-driven heuristics for ADAGE

A Data-driven model size heuristics on reconstruction error. As model size increases, the reconstruction error drops quickly at the beginning and levels out at size 300. The red line goes through the median reconstruction errors at each model size.

B Knowledge-driven model size heuristics on pathway coverage. As model size increases, pathway coverage also increases at first and then levels out at size 300. The red line goes through the median value at each model size.
C Knowledge-driven sample size heuristics on pathway coverage. For 50-node models, pathway coverage increases with sample size and peaks at 500 samples. 300-node models cover more pathways than 50-node modes in general and maintain a slow growing trend of pathway coverage at the maximum sample size.
D Data-driven sample size heuristics on reconstruction error. In both 50- and 300-node models, the reconstruction errors on the test set get closer to the reconstruction errors on the train set as sample size increases.

Figure 2: The construction and performance of eADAGE.
A eADAGE construction workflow. 100 individual ADAGE models were built using the same input dataset (step 1). Nodes from all models were extracted (step 2) and clustered based on the similarities in their associated weight vectors (step 3). Nodes derived from different models were rearranged by their clustering assignments (step 4). Weight vectors from nodes in the same cluster were averaged and thus becoming the final weight vector of a newly constructed node in an eADAGE model (step5).
B Pathway coverage comparison between individual ADAGE and ensemble ADAGE. eADAGE models (n=10) covers significantly more pathways than both corADAGE (n=10) and ADAGE (n=1000).
C The enrichment significance of three example pathways in different models. The three pathways show different trends as model size increases in individual ADAGE, however, their median significance levels in eADAGE are comparable or better than all individual models with different sizes. The grey dotted line indicates FDR q-value of 0.05 in pathway enrichment.
D Comparison among PCA, ICA, and eADAGE in pathway coverage at different significance levels. eADAGE outperforms PCA at all significance levels. eADAGE and ICA show similar pathway coverage at the cutoff q-value = 0.05. However, ICA covers less pathways than eADAGE as the significance cutoff becomes more stringent.

Figure 3: Node164pos is active in a NGM+<0.1phosphate, peptone, King's A, and PIA media
A Activity of Node164pos in each medium type. NGM+<0.1phosphate, peptone, and King's A media have evident elevation in Node164pos's activity. PIA medium show a wide range in Node164pos's activity. All other media have very low activities.
B Gene expression heatmaps of genes in Node164pos across samples in NGM+<0.1phosphate, peptone, King's A, and PIA media. Heatmap color range is determined by the Z-scored gene expression of all samples in the compendium (Figure EV6). These genes are highly expressed in all samples grown on NGM + <0.1mM phosphate, peptone, King's A, and half of samples on PIA, but not expressed in samples grown on NGM + 25mM phosphate.

Figure 4: PhoA activity, as seen by the colorimetric BCIP assay in various media
A PhoA activity, as seen by the blue-colored product of BCIP cleavage, is dependent on low phosphate concentrations, *phoB, phoR* and, in NGM, *kinB*.

B PhoA is active in King's A, Peptone and PIA and is dependent on *phoB* and *phoR* on King's A and peptone but dependent on *kinB* as well on PIA at 16 hours.
C PhoA is active in King's A, Peptone and PIA and is dependent on *phoB,* but no longer *phoR,* while still dependent on *kinB* on PIA after 32 hours.
D PhoA activity is dependent on phosphate concentrations < 0.6 mM, *phoB, phoR* and *kinB* as well at 0.5 mM phosphate in MOPS. Concentration 0.2 mM (not shown) mimics 0.1mM and concentrations 0.7mM – 0.9mM (not shown) mimic 1.0 mM.

Expanded View Figure Legends

Figure EV1: The relationship between model size and the number of KEGG pathways a node significantly associated with. Pathways associated with positive and negative signatures of a node were added together. When model is small, one node needs to account for multiple KEGG pathways. As model size grows, more nodes become available and pathways also tend to spread into different nodes.

Figure EV2: Pathway coverage comparison between individual ADAGE and ensemble ADAGE after correcting pathway crosstalk effects. eADAGE models (n=10) covers significantly more pathways than both corADAGE (n=10) and ADAGE (n=1000).

Figure EV3: The association significance of each KEGG pathway in the 300-node eADAGE models (n = 10), 300-node corADAGE models (n=10) and ADAGE models with different number of nodes (n = 100 for each model size).

Figure EV4: The coverage rate of each KEGG pathway in 300-node ADAGE models (n=1000) and 300-node eADAGE (n=10) models.

Figure EV5: The distribution of KEGG pathway coverage rates in 300-node ADAGE models (n=1000) and 300-node eADAGE models (n=10). eADAGE shows a higher density in distribution on the high coverage end.

Figure EV6: Z-scored gene expression heatmap of genes in signature Node164pos across all samples in the compendium.

Expanded View Datasets and Tables

Dataset EV1: *Pseudomonas aeruginosa* gene expression compendium. Gene expression values in each sample have been background corrected and quantile normalized.

Table EV1: Medium annotation for each sample in the compendium. Some medium names are abbreviated and their actual ingredients are provided in a separate sheet in the excel file.

Table EV2: Weight matrix of the eADAGE model highlighted in this paper. The weight matrix defines how much each gene contributes to each node.

Table EV3: Genes in each signature in the eADAGE model. Filtering the weight matrix provides these gene sets. Specifically, genes in a signature are those that were more than 2.5 standard deviations from the mean on the positive side, or those that were more than 2.5 standard deviations from the mean on the negative side.

Table EV4: The activity of each signature for each sample in the compendium. A signature has high activity when its gene weights are required to reconstruct the gene expression.

Table EV5: A complete list of signatures activated in one medium with activation scores higher than or equal to 0.4.

Table EV6: Top 5 eADAGE signatures that were active in a group of media. Each signature was annotated by its percentage of uncharacterized genes, its associated KEGG pathways, and a manual inspection of genes in the signature.

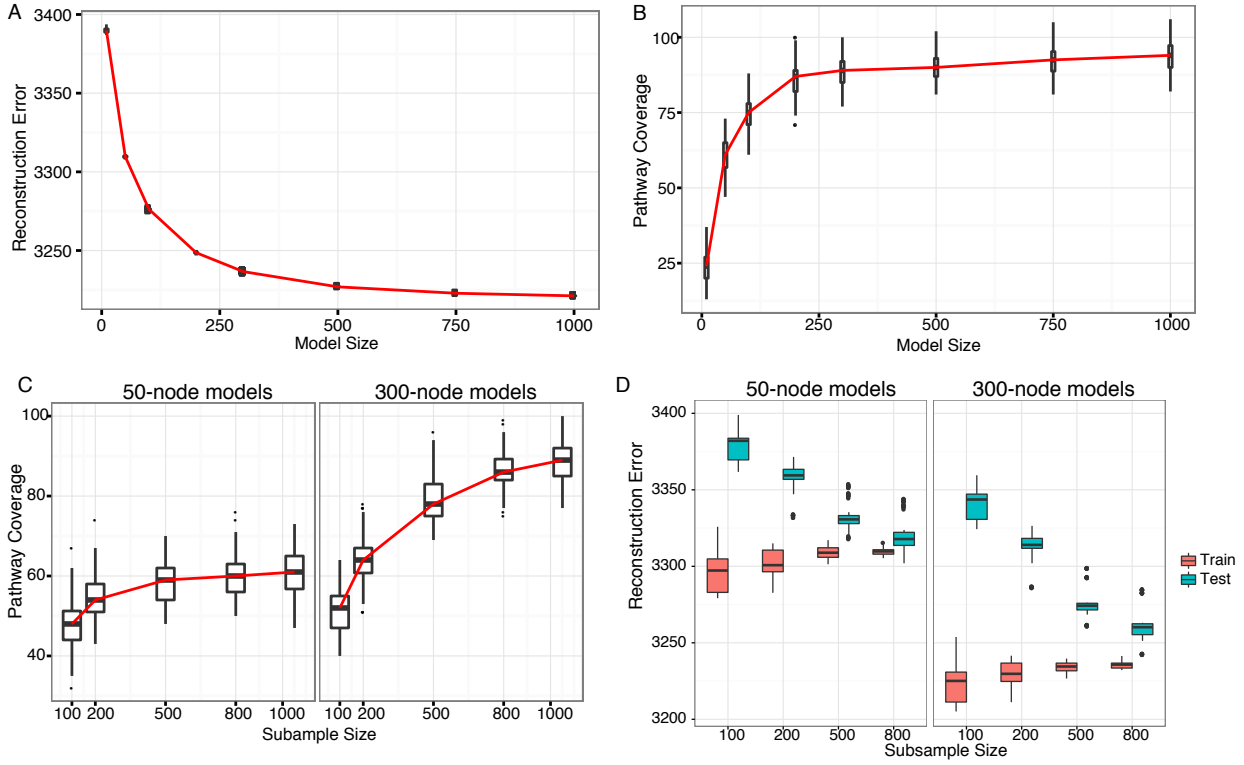
Table EV7: A complete list of eADAGE signatures activated in a group of media.

Table EV8: Comparison of the extraction of the PhoB regulon signal by PCA, ICA, ADAGE, and eADAGE models.

Table EV9: A complete list of PCA signatures activated in a group of media.

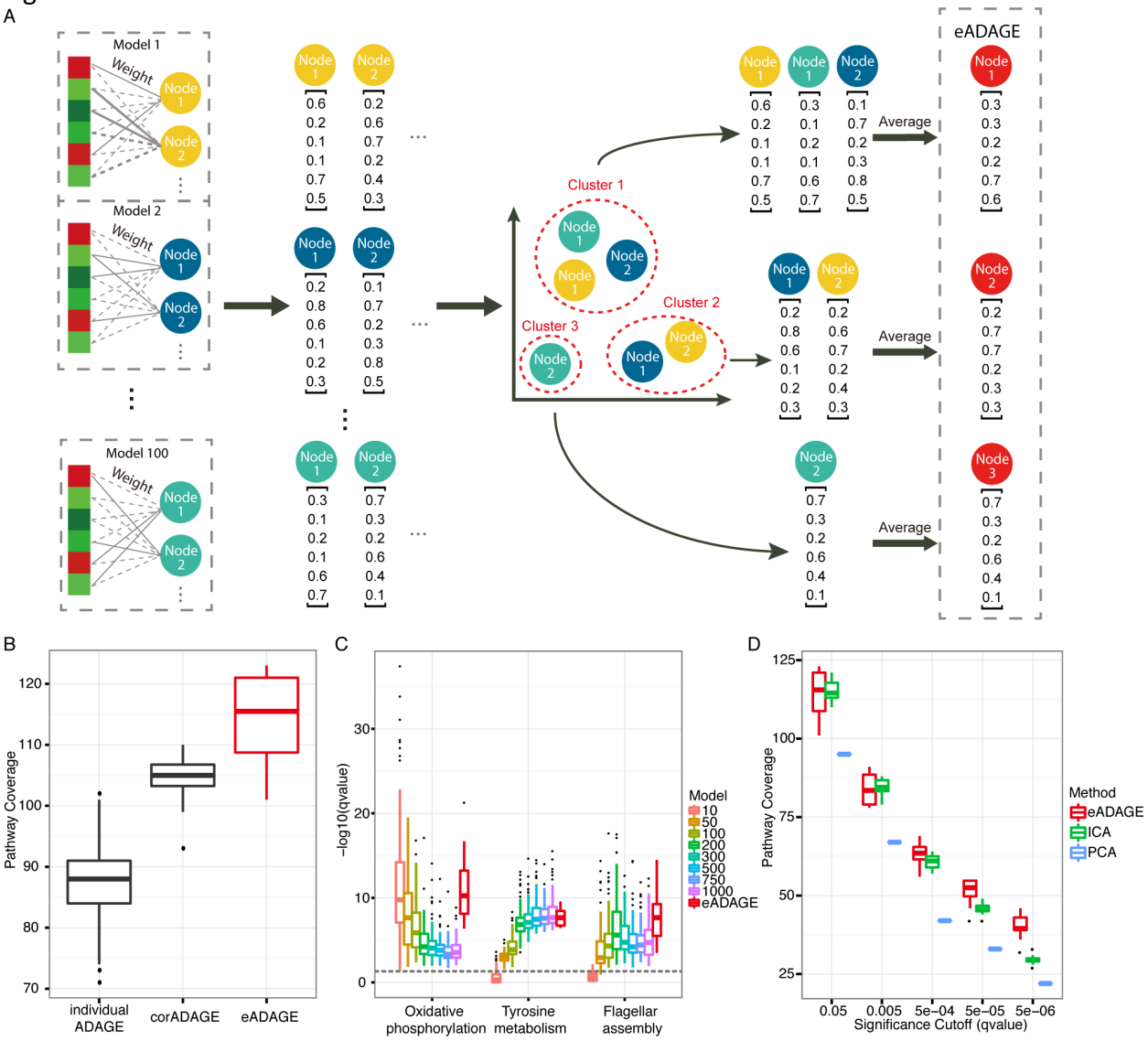
Table EV10: Library of histidine kinase deletion mutants in PA14 used for a BCIP screen in PIA. All strains in the collection of PA14 histidine kinase mutants described by the number from the Hogan lab collection (DH number), the locus of the deleted gene, gene name (if available) and position in the storage plate.

967 Figure 1



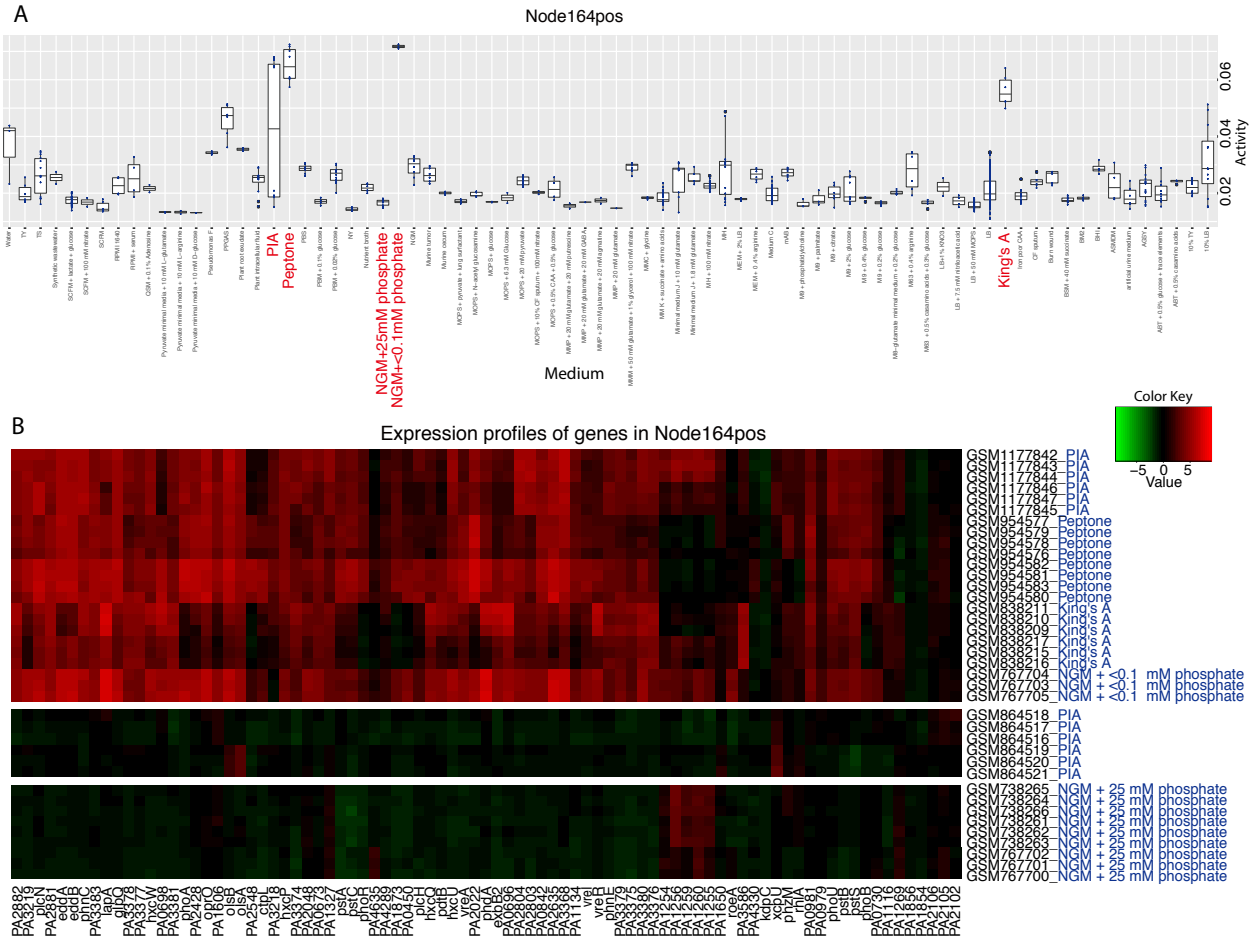
968

969 Figure 2



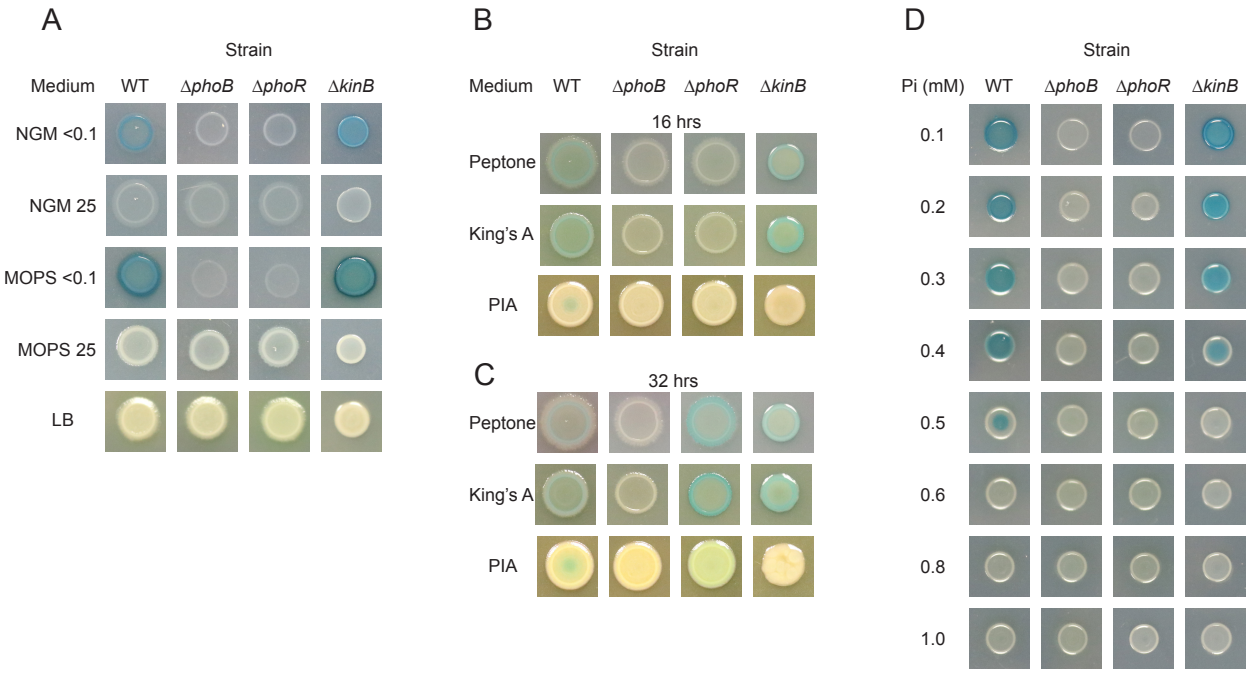
970

971 Figure 3



972

973 Figure 4



974