System-wide automatic extraction of functional signatures in *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

Running title: Extracting signatures from public data

Abstract

Abundant public expression data capture gene expression across diverse conditions. These steady state mRNA measurements could reveal the transcriptional consequences of cells' genetic backgrounds or their responses to the environment. However, public data remain relatively untapped, in part because extracting biological signal as opposed to technical noise remains challenging. Here we introduce a procedure, termed eADAGE, that performs unsupervised integration of public expression data using an ensemble of neural networks as well as heuristics that, given a dataset, help users identify an appropriate level of model complexity. This ensemble modeling approach captures biological pathways more clearly than existing methods, enabling analyses that span entire public gene expression compendia such as that for the bacterium *Pseudomonas aeruginosa*. These analyses reveal a previously undiscovered feature of the phosphate starvation response apparent in public data: a sensor kinase, KinB, that is required for full activation of the response to phosphate at intermediate concentrations. Our molecular validation experiments confirm this role of KinB and our screen of a histidine kinase knock out collection confirmed the prediction's specificity. Public data are captured from a broad range of conditions in diverse organism backgrounds and may provide a unique opportunity to identify these subtle and context-specific regulatory interactions. Algorithms that extract biological signal from these data, such as eADAGE, can highlight opportunities to discover mechanisms that are apparent from but unrealized in 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 (Vincent *et al*, 2008). Denoising autoencoders (DAs), an unsupervised neural network method, are well suited to gene expression data (Tan *et al*, 2015). DAs are 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). Investigators can use models trained on public

data to reexamine their own data to identify the processes or pathways that are particularly perturbed in their experiment.

As a result of a training process that uses stochastic gradient descent, each DA model represents a local minimum. Because the number of parameters is large, there are many distinct models that can explain the data. One way to address this challenge is by finding patterns that are stable across multiple models, which Yu highlights for its importance in statistical inference (Yu, 2013). Even if run-to-run variability in models obscures some biological features, these stable cross-model patterns may clearly resolve biological pathways. To directly target stability, we introduce an unsupervised ensemble modeling procedure that improves model robustness by constructing an ensemble neural network model. Our approach is inspired by consensus clustering (Monti *et al*, 2003), which now has become a standard part of clustering applications for biological datasets.

While this new generation of unsupervised data integration methods provides a powerful lens through which to study complex biological systems, it also raises questions: how many biological signatures are in the data; and how many samples are needed to find those features? Heuristics should include data-driven analyses that can be applied in any setting, as well as knowledge-based heuristics that are suitable in cases where pathway annotations are available. We introduce both data- and knowledge-based heuristics to suggest an appropriate model size for a given data compendium.

We previously constructed a 50-node ADAGE (Analysis using Denoising Autoencoders for Gene Expression) model of *P. aeruginosa* because this balanced the need for breadth of the model with the limits of manual annotation required for the initial validation of the method (Tan *et al*, 2016b). Our heuristics indicate that the existing *P. aeruginosa* compendium can support 300 node models, which we use in this work. We compare ADAGE models with models generated through an ensemble form of ADAGE (eADAGE) and find that eADAGE models have greater stability. The eADAGE models are not only more robust, but also more successfully capture pathway features in the data. While ADAGE allows comparison of pathway activities within an experiment, eADAGE models also enable analyses that cut across an organism's gene expression compendium. We analyze *P. aeruginosa*'s responses to multiple types of media with eADAGE. This cross-compendium analysis reveals a gene expression signature for the response to low-phosphate in certain media, which our molecular validation confirms. Our analyses of this signature across the compendium revealed a previously undiscovered link between KinB and PhoB activity. PhoB is known to mediate the low phosphate response and is canonically regulated by a histidine kinase, PhoR. However, our molecular validation shows PhoA activity, which indicates PhoB activity, independent of PhoR in some conditions and dependent on KinB in some conditions. Our experiments confirm that KinB regulates PhoB activity at physiologically relevant phosphate concentrations and that this activity is specific to KinB among kinases.

In summary, eADAGE more precisely and robustly represents biological processes and pathways. It allows 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.

While ADAGE models are constructed without the use of any curated information, we can use experimentally-derived knowledge on 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 1A). 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.

Methods that use pathway coverage to identify the appropriate model size for a collection of samples require curated pathways, but a parallel analysis can be performed without this requirement. During ADAGE training, neural networks are trained to reconstruct the input from data with noise added, and calculating the reconstruction error does not require any curated information. The reconstruction error can also be used to estimate model sizes that can be supported by the available data. The reconstruction error quickly decreases as model size increases and reaches a plateau at model size of approximately 300 (Figure 1B). Further increasing model size does not improve reconstruction, suggesting that the available data are insufficient to support larger models. Therefore, considering the knowledge-driven and data-driven heuristics together, we identified a 300-node 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 next 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.a* 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

174 eADAGE signatures tended to include genes that participated in many pathways, this could also
175 drive the increase in number of observed pathways. To control for this, we performed crosstalk
176 correction (Donato *et al*, 2013). After correction, the total number of covered pathways
177 dropped approximately by half (Figure EV2), but eADAGE still covered significantly more
178 pathways than corADAGE (t-test p-value of 1.60e-3) and ADAGE (t-test p-value of 6.16e-07).
179 These results suggested that eADAGE effectively integrates multiple models to more broadly
180 capture pathway signals embedded in diverse gene expression compendia.
181
182 We next evaluated how specifically and completely signatures learned by the models capture
183 known KEGG pathways. We use each gene signature's FDR corrected p-value for enrichment of
184 a KEGG pathway as a combined measure, as this captures both the sensitivity and specificity. If
185 a pathway was significantly associated with multiple gene signatures in a model, we only
186 considered its most significant association. We found that 71% of pathways were more
187 significantly enriched (had lower median p-values) in corADAGE models (n=10) when compared
188 to individual ADAGE models (n=100) (Figure EV3). This increased to 87% for eADAGE (n=10). We
189 also directly compared eADAGE and corADAGE by this measure and observed that 74% of
190 pathways were more significantly enriched in eADAGE. Our earlier evaluation of pathway-based
191 heuristics showed that different pathways were best captured at different model sizes (Figure
192 EV3). We next compared the 300-node eADAGE model to individual models of each size.
193 Although the 300-node eADAGE models were constructed only from 300-node ADAGE models,
194 we found that 69% pathways were more significantly enriched (i.e. lower median p-values) in
195 eADAGE models than ADAGE models of any size, including those with more nodes than the
196 eADAGE models (Figure EV3). Three example pathways that are best captured either when
197 model size is small, large, or in the middle are all well captured in the 300-node eADAGE model
198 (Figure 2C). These results demonstrate that eADAGE's ensemble modeling procedure captures
199 signals across model sizes more effectively than individual ADAGE and corADAGE models. Thus
200 eADAGE more completely and precisely captures the gene expression signatures of biological
201 pathways.
202
203 We designed eADAGE to provide a more robust analysis framework than individual ADAGE
204 models. To assess this, we examined the percentage of models that covered each pathway
205 (coverage rate) between ADAGE and eADAGE (Figure EV4). The pathways covered by each
206 individual ADAGE model were highly variable. Most KEGG pathways were covered by less than
207 half of individual models but more than half of eADAGE models (Figure EV5), suggesting that
208 eADAGE models were more robust than individual ADAGE models. We excluded all pathways
209 always covered by both individual ADAGE and eADAGE models and observed that 72% of the
210 remaining pathways were covered more frequently by eADAGE than ADAGE. This suggests that
211 their associations are stabilized through the ensemble construction procedures. In summary,
212 these comparisons of eADAGE and ADAGE reveal that not only are more pathways captured
213 more specifically, but also those that are captured are captured more consistently.
214
215 Principle component analysis (PCA) and independent component analysis (ICA) have been
216 previously used to extract biological features and build functional gene sets (Engreitz *et al*, 2010;
217 Raychaudhuri *et al*, 2000; Gong *et al*, 2007; Alter *et al*, 2000; Raychaudhuri *et al*, 2000; Lutter *et*

218 *al*, 2009; Frigyesi *et al*, 2006; Chen *et al*, 2008; Roden *et al*, 2006; Ma & Kosorok, 2009). We
219 performed PCA and generated multiple ICA models from the same *P. aeruginosa* expression
220 compendium and evaluated their KEGG pathway coverage following the same procedures used
221 for eADAGE. eADAGE substantially and significantly outperforms PCA in terms of pathway
222 coverage (Figure 2D). We observed that low-order PCs tend to be associated with more
223 pathways than high-order PCs, which is consistent with the higher variance explained by low-
224 order PCs. ICA and eADAGE covered a similar number of pathways at the significance cutoff of
225 FDR 0.05. However, we observed that eADAGE represented KEGG pathways more precisely
226 than ICA. Specifically, among pathways significantly enriched in either approach, 68% pathways
227 exhibited more significant enrichment in eADAGE. Increasing the significance threshold for
228 pathway coverage demonstrates the advantage of eADAGE (Figure 2D).
229
230 Taken together, the eADAGE method outperformed PCA, ICA, ADAGE, and corADAGE in
231 capturing KEGG pathways of *P. aeruginosa*. eADAGE had higher pathway coverage, covered
232 pathways more specifically, and more robustly than existing methods. Though the primary goal
233 of eADAGE is not function prediction, it captures pathway signals in transcriptional data
234 significantly more effectively than existing unsupervised data integration methods. These
235 results clearly indicate that eADAGE extracts biological patterns from gene expression
236 compendia.
237
238 **Elucidating functional signatures that are indicative of growth medium**
239 Gene expression analysis experiments are generally designed to include only a small number of
240 variables. Thus, it is not surprising that in the majority of publically available *P. aeruginosa* gene
241 expression experiments, the growth medium was held constant while there was variation in
242 other parameters such as strain, environmental parameters (i.e. pH, temperature, growth
243 phase, or gaseous atmosphere), or culture amendments (antibiotics, etc.). Though it was rare
244 for the base medium to vary within a single experiment, our manual annotation showed that 78
245 different base media were used across the gene expression compendium (Table EV1). These
246 media exhibit great diversity in the number of different medium components, the abundance of
247 different nutrients, and the form in which nutrients were provided. Some medium groups
248 contained numerous samples from multiple experiments other media types were unique to a
249 single experiment.
250
251 For biological evaluation, we built a single new eADAGE model with 300 nodes. In this model
252 we identified cross-media signals of biological importance. The model's weight matrix (Table
253 EV2), positive and negative gene signatures for each node (Table EV3) and signature activities
254 for each sample in the compendium (Table EV4) are provided. For each gene signature, we
255 used its mean activity for samples prepared from cells grown in a given medium compared to
256 the mean activity of the signature across all the remaining samples in the compendium
257 normalized to the total range to provide an activation score. We then further reduced this set
258 by filtering to all signature-medium relationships with an activation score greater than or equal
259 to 0.4 (Table EV5), which included only the top 2.4% of potential pairs. To capture pan-media
260 patterns we identified signatures that had multiple signature-media activation scores above this
261 threshold and averaged their activation scores. Table EV6 lists the media groups on which gene

262 signatures were most differentially active (a complete list of signature-media associations is in
263 Table EV7).
264
265 The signature Node164pos had the highest pan-media activation score. To evaluate the basis
266 for the high activation score, we examined the signature's underlying activity across media. This
267 revealed that Node164pos was highly active in King's A medium, peptone medium, and
268 NGM+<0.1mM phosphate (NGMlowP), but not in NGM+25mM phosphate (NGMhighP) (Figure
269 3A). The difference in Node164pos activity between NGMlowP and NGMhighP suggested that
270 the genes in this signature respond to phosphate concentrations. Consistent with this proposal,
271 a KEGG pathway enrichment analysis of Node164pos genes suggested a strong enrichment in
272 genes involved in phosphate acquisition (Table EV6). The other two media in which *P.*
273 *aeruginosa* gene expression leads to consistently high Node164pos activity (Peptone and King's
274 A) also had low phosphate concentrations (0.4 mM) relative to other media in the compendium.
275 For example, commonly used LB has a phosphate concentration of approximately 4.5 mM
276 (Bertani, 2004) and many others have concentrations above 20 mM.
277
278 Many of the phosphate responsive genes in Node164pos are known to be under the control of
279 the transcription factor PhoB in *P. aeruginosa* (Santos-Beneit, 2015; Blus-Kadosh *et al*, 2013;
280 Bielecki *et al*, 2015) and *phoB* itself is in this signature. Analysis of genes in Node164pos finds
281 that this signature has the largest overlap with the PhoB regulon (see methods) in comparison
282 to all other gene signatures learned by this eADAGE model (FDR q-value of 8.1e-29 in
283 hypergeometric test). The transcript levels of genes in Node164pos are higher in peptone,
284 King's A, and NGM+<0.1 phosphate medium relative to the other samples in the compendium
285 including NGM+ 25 mM phosphate (Figure 3B).
286
287 Among the highest weight genes in Node164pos is a gene that encodes alkaline phosphatase
288 (PhoA), an enzyme with an activity that can be easily measured using a colorimetric assay. As
289 expected, PhoA activity (blue color) was high when *P. aeruginosa* was grown on NGMlowP and
290 not when grown on NGMhighP (Figure 4A). The same trend was observed in another medium,
291 MOPS, with the same high and low phosphate concentrations. Also consistent, PhoA was not
292 active on the phosphate replete medium LB. Furthermore, PhoA activity was dependent on
293 PhoB and the PhoB-activating histidine kinase PhoR, which is consistent with previous
294 publications (Bielecki *et al*, 2015). These results provide striking evidence that low phosphate
295 media induced PhoB activity as predicted by the eADAGE analysis and previous
296 characterizations of the *P. aeruginosa* phosphate response.
297
298 Interestingly, Node164pos activity exhibited a wide spread in PIA medium (0.8 mM phosphate),
299 with six samples having high activities and the other six samples having low activities (Figure
300 3A). Each set of six samples came from a different study. All of the strains in the first sample set,
301 published in 2012, in which Node164pos was low used a PAO1 *kinB*::Gm^R mutant background
302 (Damron *et al*, 2012). The second, published in 2013, used a PAO1 strain with *kinB* intact and
303 showed high Node164pos activity on PIA (Damron *et al*, 2013). The fact that *kinB* mutant
304 samples in PIA show significantly lower Node164pos activity suggests that KinB may be a
305 regulator of PhoB on PIA.

306
307 Expectedly, PhoA activity was high in peptone and King's A media, as in NGMlowP. Further, as
308 predicted by eADAGE Node164pos activity, PhoA activity was evident and was KinB dependent
309 on PIA medium (Figure 4B). Notably, PhoA activity was still dependent on PhoB and PhoR as it
310 was on peptone and King's A. Over time the Δ*phoR* mutants developed PhoA activity on all
311 three media, but the Δ*kinB* on PIA did not (Figure 4C). Recovery of PhoA activity in Δ*phoR*
312 mutants suggests that there are PhoR-independent paths for PhoB activation. The co-
313 dependence on KinB and PhoR suggest that these kinases do not perform redundant functions
314 but rather regulate PhoB in conjunction with each other. To determine if the deletion of kinases
315 non-specifically altered PhoB activation, we screened 63 in-frame deletion mutants each lacking
316 a histidine kinase (Table EV8) to identify mutants altered in PhoA activation on PIA. Other than
317 Δ*phoR*, Δ*kinB* was the only strain lacking PhoA activity on NGMlowP. This suggests that PhoA
318 activity is not altered by changes in generic kinase activity but rather the interaction with KinB is
319 specific.
320
321 Notably, although PIA, peptone and King's A contain low phosphate levels compared to rich
322 media such as LB, PIA has a higher phosphate concentration (0.8mM) than do peptone and
323 King's A (0.4mM). To test whether the moderately low level of phosphate in PIA provokes KinB
324 regulation of PhoA, we conducted a titration experiment in MOPS minimal medium. PhoA
325 activity was consistent with that on NGM and other media. Further, while the PAO1 wild type
326 (WT) showed PhoA activity at 0.5 mM, PAO1 Δ*kinB* did not (Figure 4D). This shows KinB
327 regulation of PhoB in moderately low phosphate and not in lower phosphate. Noting that PIA
328 has a phosphate concentration of 0.8 mM and our titration experiment in MOPS showed *kinB*-
329 sensitivity at 0.5 mM, it is likely that the precise concentration at which KinB regulates PhoA
330 depends on the background medium. Node164pos activity across the compendium suggested,
331 and experimental evidence confirmed, that KinB regulates PhoA at moderately low phosphate
332 levels. To our knowledge, KinB has not been previously implicated in the activation of PhoB,
333 although cross-compendium eADAGE analysis of two PIA experiments was capable of revealing
334 this relationship.
335
336 In summary, eADAGE effectively extracted biologically meaningful features, accurately
337 indicated their activity in multiple media spanning numerous independent experiments, and
338 revealed a novel regulatory mechanism. By summarizing gene-based expression information
339 into biologically relevant features, eADAGE greatly simplifies analyses that cut across large gene
340 expression compendia.
341

342 ## Discussion

343 For unsupervised approaches, it has been very challenging to determine the appropriate model
344 complexity. In supervised learning problems, predictors can be assessed through cross-
345 validation accuracies. To our knowledge, there is not yet a similar well-established approach to
346 estimate appropriate model size for unsupervised feature construction of gene expression data.
347 Here we develop heuristics that target two aspects of the problem: the model needs to be well
348 supported by the amount of available data and the extracted features should well resemble
349 known biological processes. Our data-driven heuristics can be applied to organisms for which

350 gene-process annotations are lacking. We expect that additional data will support larger models,
351 especially data that measure experimental conditions that are not tested in the existing
352 compendium.
353
354 We also contribute a novel eADAGE algorithm. This algorithm combines multiple ADAGE
355 models into one ensemble model to address model variability due to stochasticity and local
356 minima. The algorithm is inspired by consensus clustering, which reconciles the differences in
357 cluster assignments in multiple runs. Comparable approaches have also been applied for ICA,
358 where researchers have used the centrotypes of multiple ICA models as the final model
359 (Frigyesi *et al*, 2006). The ICA centrotype approach for ADAGE corresponds to corADAGE, and
360 our comparison of eADAGE and corADAGE shows that eADAGE not only covers more biological
361 pathways, but also results in cleaner representations of biological pathways. The results of our
362 direct comparison suggest that placing particular emphasis on the genes most associated with a
363 particular signature may be a useful property for other unsupervised feature construction
364 algorithms in biology.
365
366 Our eADAGE model revealed patterns that were obvious in the analysis of the large data
367 compendium containing experiments performed in 78 different media, but that were not
368 necessarily evident in individual experiments. For example, our cross-compendium analysis of
369 expression patterns specific to certain media determined those media in which *P. aeruginosa*
370 was in a phosphate limited, PhoB-ON state. PhoB is a global regulator, and thus an awareness
371 of the different states of the PhoB regulon in different media will likely provide important
372 insight into medium-specific phenotypes of relevance. Using eADAGE, we uncovered a subtle
373 aspect of the phosphate starvation response that depends on a histidine kinase not previously
374 associated with the PhoB-dependent phosphate response pathway. Bacteria have evolved
375 many mechanisms to insulate response pathways from each other (Podgornaia & Laub, 2013),
376 but cross-talk, wherein a sensor kinase from one pathway phosphorylates a response regulator
377 from another, has been suggested to explain the complexity of signaling networks, including
378 that of PhoB (Ninfa *et al*, 1988; Fisher *et al*, 1995). Thus, it is possible that both PhoR and KinB
379 directly activate PhoB. While there are many examples of one kinase partially compensating
380 when the cognate kinase is absent, it is more challenging to find conditions where two kinases
381 are needed for full response regulator activation (Verhamme *et al*, 2002). Alternatively, KinB
382 may influence PhoB activity indirectly by regulating activities that affect PhoB levels, PhoB
383 dephosphorylation, or PhoB localization, or protein-protein interactions. Future work will
384 further address the mechanistic aspects of this model. We propose that moderate levels of
385 phosphate, like those in PIA, provide a niche for crosstalk: the activity of PhoR is low enough
386 that the interaction with KinB is needed for full PhoB activity on this medium. Together, PhoR
387 and KinB may enable a more sensitive and effective response to phosphate limitation. Without
388 the eADAGE analysis across multiple media, we would not have found this nuanced mechanism.
389
390 In the 300-node eADAGE model used for medium analysis, 81% genes in *P.a.* PAO1 genome are
391 included in at least one gene signature. Among all signatures, 59% have a KEGG pathways
392 associated with them. Among the remaining 41% of signatures, 17 of them had ten or fewer
393 genes and may not have associated with KEGG pathways for this reason. Some others were

394 associated with known pathways that have not yet been annotated in KEGG, such as
395 Node174pos which encodes genes encoded by the P2 phage or Node150neg which contains
396 genes involved in the formation of surface associated communities referred to as biofilms. Thus
397 eADAGE can group functionally linked genes in ways that may facilitate the identification or
398 annotation of pathways.
399
400 There are now abundant public gene expression data. Cross-compendium analyses provide the
401 opportunity to efficiently use existing data to identify regulatory patterns that are evident
402 across multiple experiments, datasets, and labs. To tap this potential, we will require algorithms
403 that robustly integrate these diverse datasets in a manner that is not tied to only aspects of
404 biology that are well understood. We expect that robust unsupervised data integration
405 methods, like eADAGE, will play a key role in this process.
406

407 ## Materials and Methods

408 **Data processing**

409 We followed the same procedures for data collection, processing, and normalization from (Tan
410 *et al*, 2016b) and updated the *P. aeruginosa* gene expression compendium to include newly
411 uploaded datasets on GPL84 platform from the ArrayExpress database (Rustici *et al*, 2013) on
412 31 July 2015. The updated *P. aeruginosa* compendium contains 125 datasets with 1051
413 individual genome-wide assays. We provide the *P. aeruginosa* expression compendium (Dataset
414 EV1) along with all the code used in this paper (Tan *et al*, 2016a). The eADAGE repository is also
415 tracked under version control at https://bitbucket.org/greenelab/eadage.
416

417 **Construction of ADAGE models**

418 We constructed ADAGE models as described in (Tan *et al*, 2016b). To summarize the process
419 and outputs, we constructed a denoising autoencoder for the gene expression compendium.
420 Denoising autoencoders model the data in a lower dimension than the input space, and the
421 models are trained with random gene expression measurements set to zero. Thus an ADAGE
422 model must learn gene-gene dependencies to fill in this missing information. Once the ADAGE
423 model is trained, each node in the hidden layer contains a weight vector. These positive and
424 negative weights represent the strength of each gene's connection to that node.
425

426 **Gene signatures as sign-specific high-weight gene sets**

427 In previous work (Tan *et al*, 2016b) we defined high-weight (HW) genes as those in the
428 extremes of the weight distribution on the positive or negative side of a node. Here, we use a
429 more granular definition that accounts for sign specificity. Each node's gene weights are
430 approximately normal and centered at zero in ADAGE models (Tan *et al*, 2015, 2016b). We
431 defined positive HW genes as those that were more than 2.5 standard deviations from the
432 mean on the positive side, and negative HW genes as those that were more than 2.5 standard
433 deviations from the mean on the negative side. After this split, a model with n nodes provides
434 2n gene signatures. Because a node is simply named by the order that it occurs in a model, we
435 named two gene signatures derived from one node as "NodeXXpos" and "NodeXXneg".
436

437 **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

482 combination of 10 different random subsets and 10 different random training initializations. To
483 evaluate each model, we randomly selected 200 samples not used during training as its testing
484 set. We performed this subsampling analysis at model size 50 and 300. In total, 800 ADAGE
485 models were built in the sample size evaluation experiment.
486
487 **Construction of eADAGE models**
488 We constructed ensemble ADAGE (eADAGE) models by combining many individual ADAGE
489 models in to a single model. For each eADAGE model we combined 100 individual ADAGE
490 models. The 100 models were trained with identical parameters but distinct random seeds. For
491 an eADAGE model of size 300, we trained 100 individual models with 300 nodes each, which
492 provided 30000 total nodes. Each node has a weight vector. We have previously observed that
493 high-weight genes provided the most information to each node (Tan *et al*, 2016b), so we
494 calculated a weighted Pearson correlation between each node's weight vectors. Our weighted
495 Pearson correlation used (|node1 weight|+|node2 weight|)/2 as the weight function for each
496 gene. We compared this to an unweighted Pearson correlation (corADAGE) as well a baseline
497 ADAGE model.
498
499 After calculating correlation (weighted for eADAGE and unweighted for corADAGE), we
500 converted the correlation to distance by calculating (1- correlation)/2. This provided a
501 30000*30000 distance matrix storing distances between every two nodes. We clustered this
502 distance matrix using the Partition Around Medoids (PAM) clustering algorithm (Park & Jun,
503 2009).We implemented clustering in R using the ConsensusClusterPlus package (Wilkerson &
504 Hayes, 2010) from Bioconductor with the ppam function from Sprint package to perform
505 parallel PAM (Piotrowski *et al*, 2011). We set the number of clusters to match the individual
506 ADAGE model (e.g. 300) allowing for direct comparison between the eADAGE and ADAGE
507 methods.
508
509 Clustering assigned each node to a cluster ranging from 1 to 300. We combined nodes assigned
510 to the same cluster by calculating the average of their weight vectors. These 300 averaged
511 vectors formed the weight matrix of the eADAGE model. Because the ensemble model is built
512 from the weight matrices of individual models, it does not have the parameters that form the
513 bias vectors. We built 10 eADAGE and 10 corADAGE models from 1000 ADAGE models with
514 each ensemble model built upon 100 different individual models. The individual eADAGE model
515 used for biological analysis in this work was constructed with random seed 123, which was
516 arbitrarily chosen before model construction and evaluation.
517
518 **PCA and ICA model construction**
519 We constructed PCA and ICA models and defined each model's weight matrix following the
520 same procedures in (Tan *et al*, 2016b). To compare with the 300-node eADAGE, we generated
521 models of matching size (300 components). For ICA, we evaluated 10 replicates. PCA provides a
522 single model. PCA and ICA models were evaluated through the KEGG pathway enrichment
523 analysis described above.
524
525 **Media annotation of the *P. aeruginosa* compendium**

526 A team of *P. aeruginosa* biologists annotated the media for all samples in the compendium by
527 referring to information associated with each sample in the ArrayExpress (Rustici *et al*, 2013)
528 and/or GEO (Edgar, 2002) databases and along with the original publication, if reported. Each
529 sample was annotated by two curators separately. Conflicting annotations, if they occurred,
530 were resolved by a third curator. The media annotation for all samples in the compendium
531 were provided in Table EV1.
532
533 **Activity calculation for a gene signature**
534 We calculated a signature's activity for a specific sample as $\mathrm{A} = \mathrm{W} \cdot \mathrm{E} \ / \ \mathrm{N}$, in which W is the
535 weight vector of genes in that signature, E is a vector of genes' expression values after zero-one
536 normalization in that sample, and N is the number of genes. It can be viewed as an averaged
537 weighted sum of genes' expression levels. We normalized a signature's activity by the number
538 of genes (N) in that signature, because different signatures have different number of genes.
539
540 **Identification of signatures activated across media**
541 We calculated an activation score to identify gene signatures with dramatically elevated or
542 reduced activity in a specific medium. We grouped samples by their medium annotation. For
543 each gene signature and medium combination, we calculated the difference between the mean
544 activity of the signature for samples in that medium as well as the mean activity across the
545 remainder of samples in the compendium. We divided this absolute difference in the means by
546 the range of activity for all samples across the compendium. This score captures the proportion
547 by which the mean activity in a medium differs relative to the total difference across the
548 compendium. We termed this ratio the activation score.
549
550 To identify the most specifically active signatures for each medium, we constructed a table for
551 all pairs with an activation score greater than or equal to 0.4 (Table EV5). This was highly
552 stringent: it captured only the top 2.4% of the potential signature-medium pairs. To identify
553 pan-media signatures, we limited signatures to those that were active in multiple media
554 (greater or equal to 0.4) and averaged their activation scores. These signatures exhibit parallel
555 patterns for multiple media across multiple distinct experiments.
556
557 **Definition of the PhoB regulon**
558 A PhoB regulon for the PAO1 genome was adapted from the PhoB regulon of PA14 in (Bielecki
559 *et al*, 2015) in order to be comparable to ADAGE models. Of the 187 genes in the PA14 regulon,
560 160 were in the PAO1 reference genome (www.pseudomonas.com).
561
562 **Strains and Media**
563 Strains used were WT, Δ*phoB* (DH2633, O'Toole lab collection), Δ*phoR* (DH2516) and Δ*kinB*
564 (DH2517), all in the PA14 background. All strains were maintained on LB with 1.5% agar and
565 grown at 37 °C. For cross-media and phosphate concentration comparisons, BCIP assays (see
566 methods below) were conducted on different base media with 1.5% agar (Fisher): King's A
567 (pancreatic digest of gelatin (Difco) 20g/L; $MgCl_2$ 1.4g/L; K_2SO_4 10g/L; glycerol 10ml/L) (King *et al*,
568 1954), LB (Tryptone (Fisher) 10g/L; Yeast Extract (Fisher) 5g/L; NaCl 5g/L) (Bertani, 2004), MOPS
569 (morpholinepropanesulfonic acid 40mM; Glucose 20 ml/L; K_2SO_4 2.67mM; K_2HPO_2 0mM, 25mM

570 or 0.1 – 1 mM) (Neidhardt *et al*, 1974), NGM (Pancreatic digest of gelatin 2.5g/L; cholesterol
571 5mg/L; NaCl 3g/L; $MgSO_4$ 1mM; $CaCl_2$ 1mM; KCl 25mM; Potassium Phosphate buffer pH6 0 or
572 25 mM) (Zaborin *et al*, 2009), Peptone (Pancreatic digest of gelatin 10g/L; $MgSO_4$ 1.5g/L; K_2SO_4
573 10g/L) (Lundgren *et al*, 2013), Pseudomonas Isolation Agar (PIA, prepared as per instructions,
574 BioWorld).
575
576 **BCIP assay**
577 Various media were supplemented with 5-bromo-4-chloro-3-indolyl phosphate (BCIP) DMF
578 solution to a final concentration of 60 µg/mL. BCIP assay plates were inoculated with 5 µl of
579 overnight *P.a* culture in LB broth. Colonies were grown for 16 hours at 37 °C then matured at
580 room temperature until imaging. Images were collected 16 and 32 hours post inoculation.
581
582 **Screen of a histidine kinase knock out collection**
583 Molecular techniques to construct the histidine kinase (HK) knock out collection were carried
584 out as previously described (Ha *et al*, 2014). For each strain in the HK collection, a BCIP assay
585 was performed on PIA. Plates were struck with an overnight *P.a.* culture concentrated two-fold
586 by centrifugation. Plates were incubated at 37 °C 12-16 hours and matured at room
587 temperature for an additional 12-16 hours alkaline phosphatase activity was determined
588 qualitatively, based on blue color.
589

590 ## Acknowledgements

591 This work was supported in part by a grant from the Gordon and Betty Moore Foundation
592 (GBMF 4552) to CSG. This work was supported by National Institutes of Health (NIH) grant RO1-
593 AI091702 to DAH. MTL is an investigator of the Howard Hughes Medical Institute. This work
594 was supported by a pilot grant from the Cystic Fibrosis Foundation (STANTO15R0) to CSG and
595 DAH. The authors acknowledge Gregory Way and René Zelaya for helpful code review.
596

597 ## Author contributions

598 JT, DAH and CSG conceived and designed the research. JT, GD and KMC performed
599 computational analyses. GD, KAL and CEP performed molecular experiments. KC, BP and MTL
600 constructed and contributed the histidine kinase knock out collection. JT, GD, KMC, DAH and
601 CSG wrote the manuscript, and KAL, CEP, KMC, KD, BP and MTL provided critical feedback.
602

603 ## Conflict of interest

604 The authors have no conflicts of interest to report.
605

606 ## Reference

607 Alter O, Brown PO & Botstein D (2000) Singular value decomposition for genome-wide
608 expression data processing and modeling. *Proc. Natl. Acad. Sci. U. S. A.* **97:** 10101–6
609 Bertani G (2004) Lysogeny at mid-twentieth century: P1, P2, and other experimental systems. *J.*
610 *Bacteriol.* **186:** 595–600
611 Bielecki P, Jensen V, Schulze W, Gödeke J, Strehmel J, Eckweiler D, Nicolai T, Bielecka A, Wille T,
612 Gerlach RG & Häussler S (2015) Cross talk between the response regulators PhoB and TctD

613 allows for the integration of diverse environmental signals in *Pseudomonas aeruginosa*.
614 *Nucleic Acids Res.* **43:** 6413–25
615 Blus-Kadosh I, Zilka A, Yerushalmi G & Banin E (2013) The effect of *pstS* and *phoB* on quorum
616 sensing and swarming motility in *Pseudomonas aeruginosa*. *PLoS One* **8:** e74444
617 Chen L, Xuan J, Wang C, Shih I-M, Wang Y, Zhang Z, Hoffman E, Clarke R, Devore J, Peck R,
618 Tusher V, Tibshirani R, Chu G, Storey J, Xiao W, Leek J, Tompkins R, Davis R, Conesa A,
619 Nueda M, et al (2008) Knowledge-guided multi-scale independent component analysis for
620 biomarker identification. *BMC Bioinformatics* **9:** 416
621 Damron FH, Barbier M, McKenney ES, Schurr MJ & Goldberg JB (2013) Genes required for and
622 effects of alginate overproduction induced by growth of *Pseudomonas aeruginosa* on
623 Pseudomonas isolation agar supplemented with ammonium metavanadate. *J. Bacteriol.*
624 **195:** 4020–36
625 Damron FH, Owings JP, Okkotsu Y, Varga JJ, Schurr JR, Goldberg JB, Schurr MJ & Yu HD (2012)
626 Analysis of the *Pseudomonas aeruginosa* regulon controlled by the sensor kinase KinB and
627 sigma factor RpoN. *J. Bacteriol.* **194:** 1317–30
628 Donato M, Xu Z, Tomoiaga A, Granneman JG, Mackenzie RG, Bao R, Than NG, Westfall PH,
629 Romero R & Draghici S (2013) Analysis and correction of crosstalk effects in pathway
630 analysis. *Genome Res.* **23:** 1885–93
631 Edgar R (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data
632 repository. *Nucleic Acids Res.* **30:** 207–210
633 Engreitz JM, Daigle BJ, Marshall JJ & Altman RB (2010) Independent component analysis: mining
634 microarray data for fundamental human gene expression modules. *J. Biomed. Inform.* **43:**
635 932–44
636 Fisher SL, Jiang W, Wanner BL & Walsh CT (1995) Cross-talk between the histidine protein
637 kinase VanS and the response regulator PhoB. Characterization and identification of a VanS
638 domain that inhibits activation of PhoB. *J. Biol. Chem.* **270:** 23143–9
639 Frigyesi A, Veerla S, Lindgren D, Höglund M, Quackenbush J, Jutten C, Herault J, Chiappetta P,
640 Roubaud M, Torrésani B, Hyvärinen A, Oja E, Liebermeister W, Lee S, Batzoglou S,
641 Martoglio A, Miskin J, Smith S, MacKay D, Saidi S, et al (2006) Independent component
642 analysis reveals new and biologically significant structures in microarray data. *BMC*
643 *Bioinformatics* **7:** 290
644 Gillis J & Pavlidis P (2013) Assessing identity, redundancy and confounds in Gene Ontology
645 annotations over time. *Bioinformatics* **29:** 476–82
646 Gong T, Xuan J, Wang C, Li H, Hoffman E, Clarke R & Wang Y (2007) Gene module identification
647 from microarray data using nonnegative independent component analysis. *Gene Regul.*
648 *Syst. Bio.* **1:** 349–63
649 Greene CS & Troyanskaya OG (2012) Accurate evaluation and analysis of functional genomics
650 data and methods. *Ann. N. Y. Acad. Sci.* **1260:** 95–100
651 Ha D-G, Richman ME & O'Toole GA (2014) Deletion mutant library for investigation of
652 functional outputs of cyclic diguanylate metabolism in *Pseudomonas aeruginosa* PA14.
653 *Appl. Environ. Microbiol.* **80:** 3384–93
654 Jiang Y, Oron TR, Clark WT, Bankapur AR, D'Andrea D, Lepore R, Funk CS, Kahanda I, Verspoor
655 KM, Ben-Hur A, Koo DCE, Penfold-Brown D, Shasha D, Youngs N, Bonneau R, Lin A,
656 Sahraeian SME, Martelli PL, Profiti G, Casadio R, et al (2016) An expanded evaluation of

657 protein function prediction methods shows an improvement in accuracy. *Genome Biol.* **17:**
658 184
659 Kanehisa M (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28:**
660 27–30
661 KING EO, WARD MK & RANEY DE (1954) Two simple media for the demonstration of pyocyanin
662 and fluorescin. *J. Lab. Clin. Med.* **44:** 301–7
663 Lundgren BR, Thornton W, Dornan MH, Villegas-Peñaranda LR, Boddy CN & Nomura CT (2013)
664 Gene PA2449 is essential for glycine metabolism and pyocyanin biosynthesis in
665 *Pseudomonas aeruginosa* PAO1. *J. Bacteriol.* **195:** 2087–100
666 Lutter D, Langmann T, Ugocsai P, Moehle C, Seibold E, Splettstoesser WD, Gruber P, Lang EW &
667 Schmitz G (2009) Analyzing time-dependent microarray data using independent
668 component analysis derived expression modes from human macrophages infected with *F.*
669 *tularensis holartica*. *J. Biomed. Inform.* **42:** 605–611
670 Ma S & Kosorok MR (2009) Identification of differential gene pathways with principal
671 component analysis. *Bioinformatics* **25:** 882–9
672 Monti S, Tamayo P, Mesirov J & Golub T (2003) Consensus clustering: a resampling-based
673 method for class discovery and visualization of gene expression microarray data. *Mach.*
674 *Learn.* **52:** 91–118
675 Neidhardt FC, Bloch PL & Smith DF (1974) Culture medium for enterobacteria. *J. Bacteriol.* **119:**
676 736–47
677 Ninfa AJ, Ninfa EG, Lupas AN, Stock A, Magasanik B & Stock J (1988) Crosstalk between bacterial
678 chemotaxis signal transduction proteins and regulators of transcription of the Ntr regulon:
679 evidence that nitrogen assimilation and chemotaxis are controlled by a common
680 phosphotransfer mechanism. *Proc. Natl. Acad. Sci. U. S. A.* **85:** 5492–6
681 Park H-S & Jun C-H (2009) A simple and fast algorithm for K-medoids clustering. *Expert Syst.*
682 *Appl.* **36:** 3336–3341
683 Piotrowski M, Forster T, Dobrezelecki B, Sloan TM, Mitchell L, Ghazal P, Mewsissen M, Petrou S,
684 Trew A & Hill J (2011) Optimisation and parallelisation of the partitioning around medoids
685 function in R. In *2011 International Conference on High Performance Computing &*
686 *Simulation* pp 707–713. IEEE
687 Podgornaia AI & Laub MT (2013) Determinants of specificity in two-component signal
688 transduction. *Curr. Opin. Microbiol.* **16:** 156–62
689 Raychaudhuri S, Stuart JM & Altman RB (2000) Principal components analysis to summarize
690 microarray experiments: application to sporulation time series. *Pac. Symp. Biocomput.***:**
691 455–66
692 Roden JC, King BW, Trout D, Mortazavi A, Wold BJ, Hart CE, Tavazoie S, Hughes J, Campbell M,
693 Cho R, Church G, Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E,
694 Lander E, Golub T, Eisen M, et al (2006) Mining gene expression data by interpreting
695 principal components. *BMC Bioinformatics* **7:** 194
696 Rustici G, Kolesnikov N, Brandizi M, Burdett T, Dylag M, Emam I, Farne A, Hastings E, Ison J,
697 Keays M, Kurbatova N, Malone J, Mani R, Mupo A, Pedro Pereira R, Pilicheva E, Rung J,
698 Sharma A, Tang YA, Ternent T, et al (2013) ArrayExpress update--trends in database
699 growth and links to data analysis tools. *Nucleic Acids Res.* **41:** D987-90
700 Santos-Beneit F (2015) The Pho regulon: a huge regulatory network in bacteria. *Front. Microbiol.*

701 **6:** 402
702 Schnoes AM, Ream DC, Thorman AW, Babbitt PC & Friedberg I (2013) Biases in the
703 experimental annotations of protein function and their effect on our understanding of
704 protein function space. *PLoS Comput. Biol.* **9:** e1003063
705 Tan J, Doing G, Lewis KA, Price CE, Chen KM, Cady KC, Perchuk B, Laub MT, Hogan DA & Greene
706 CS (2016a) eADAGE-1.0.0rc1. *Zenodo*
707 Tan J, Hammond JH, Hogan DA & Greene CS (2016b) ADAGE-Based Integration of Publicly
708 Available *Pseudomonas aeruginosa* Gene Expression Data with Denoising Autoencoders
709 Illuminates Microbe-Host Interactions. *mSystems* **1:** e00025-15
710 Tan J, Ung M, Cheng C & Greene CS (2015) Unsupervised feature construction and knowledge
711 extraction from genome-wide assays of breast cancer with denoising autoencoders. *Pac.*
712 *Symp. Biocomput.* **20:** 132–43
713 Verhamme DT, Arents JC, Postma PW, Crielaard W & Hellingwerf KJ (2002) Investigation of in
714 vivo cross-talk between key two-component systems of *Escherichia coli*. *Microbiology* **148:**
715 69–78
716 Vincent P, Larochelle H, Bengio Y & Manzagol P-A (2008) Extracting and composing robust
717 features with denoising autoencoders. In *Proceedings of the 25th International Conference*
718 *on Machine learning - ICML '08* pp 1096–1103. New York, New York, USA: ACM Press
719 Wilkerson MD & Hayes DN (2010) ConsensusClusterPlus: a class discovery tool with confidence
720 assessments and item tracking. *Bioinformatics* **26:** 1572–3
721 Yu B (2013) Stability. *Bernoulli* **19:** 1484–1500
722 Zaborin A, Romanowski K, Gerdes S, Holbrook C, Lepine F, Long J, Poroyko V, Diggle SP, Wilke A,
723 Righetti K, Morozova I, Babrowski T, Liu DC, Zaborina O & Alverdy JC (2009) Red death in
724 *Caenorhabditis elegans* caused by *Pseudomonas aeruginosa* PAO1. *Proc. Natl. Acad. Sci. U.*
725 *S. A.* **106:** 6327–32
726

727 ## Figure Legends

728 **Figure1: Knowledge- and data-driven heuristics for ADAGE**

729 A Knowledge-driven model size heuristics on pathway coverage. As model size increases,
730 pathway coverage also increases at first and then levels out at size 300. The red line goes
731 through the median value at each model size.
732 B Data-driven model size heuristics on reconstruction error. As model size increases, the
733 reconstruction error drops quickly at the beginning and levels out at size 300. The red line goes
734 through the median reconstruction errors at each model size.
735 C Knowledge-driven sample size heuristics on pathway coverage. For 50-node models,
736 pathway coverage increases with sample size and peaks at 500 samples. 300-node models
737 cover more pathways than 50-node modes in general and maintain a slow growing trend of
738 pathway coverage at the maximum sample size.
739 D Data-driven sample size heuristics on reconstruction error. In both 50- and 300-node
740 models, the reconstruction errors on the test set get closer to the reconstruction errors on the
741 train set as sample size increases.
742

743 **Figure 2: The construction and performance of eADAGE.**

744 A eADAGE construction workflow. 100 individual ADAGE models were built using the same
745 input dataset (step 1). Nodes from all models were extracted (step 2) and clustered based on
746 the similarities in their associated weight vectors (step 3). Nodes derived from different models
747 were rearranged by their clustering assignments (step 4). Weight vectors from nodes in the
748 same cluster were averaged and thus becoming the final weight vector of a newly constructed
749 node in an eADAGE model (step5).
750 B Pathway coverage comparison between individual ADAGE and ensemble ADAGE.
751 eADAGE models (n=10) covers significantly more pathways than both corADAGE (n=10) and
752 ADAGE (n=1000).
753 C The enrichment significance of three example pathways in different models. The three
754 pathways show different trends as model size increases in individual ADAGE, however, their
755 median significance levels in eADAGE are comparable or better than all individual models with
756 different sizes. The grey dotted line indicates FDR q-value of 0.05 in pathway enrichment.
757 D Comparison among PCA, ICA, and eADAGE in pathway coverage at different significance
758 levels. eADAGE outperforms PCA at all significance levels. eADAGE and ICA show similar
759 pathway coverage at the cutoff q-value = 0.05. However, ICA covers less pathways than
760 eADAGE as the significance cutoff becomes more stringent.
761
762 **Figure 3: Node164pos is active in a NGM+<0.1phosphate, peptone, King's A, and PIA media**
763 A Activity of Node164pos in each medium type. NGM+<0.1phosphate, peptone, and King's
764 A media have evident elevation in Node164pos's activity. PIA medium show a wide range in
765 Node164pos's activity. All other media have very low activities.
766 B Gene expression heatmaps of genes in Node164pos across samples in
767 **NGM+<0.1phosphate, peptone, King's A, and PIA media**. Heatmap color range is determined
768 by the Z-scored gene expression of all samples in the compendium (Figure EV6). These genes
769 are highly expressed in all samples grown on NGM + <0.1mM phosphate, peptone, King's A, and
770 half of samples on PIA, but not expressed in samples grown on NGM + 25mM phosphate.
771
772 **Figure 4: PhoA activity, as seen by the colorimetric BCIP assay in various media**
773 A PhoA activity, as seen by the blue-colored product of BCIP cleavage, is dependent on
774 low phosphate concentrations, *phoB, phoR* and, in NGM, *kinB*.
775 B PhoA is active in King's A, Peptone and PIA and is dependent on *phoB* and *phoR* on
776 King's A and peptone but dependent on *kinB* as well on PIA at 16 hours.
777 C PhoA is active in King's A, Peptone and PIA and is dependent on *phoB,* but no longer
778 *phoR,* while still dependent on *kinB* on PIA after 32 hours.
779 D PhoA activity is dependent on phosphate concentrations < 0.6 mM, *phoB, phoR* and *kinB*
780 as well at 0.5 mM phosphate in MOPS. Concentration 0.2 mM (not shown) mimics 0.1mM and
781 concentrations 0.7mM – 0.9mM (not shown) mimic 1.0 mM.
782
783 ## Expanded View Figure Legends
784 Figure EV1: The relationship between model size and the number of KEGG pathways a node
785 significantly associated with. Pathways associated with positive and negative signatures of a
786 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.

829 Table EV6: Top 5 eADAGE signatures that were active in a group of media. Each signature was
830 annotated by its percentage of uncharacterized genes, its associated KEGG pathways, and a
831 manual inspection of genes in the signature.
832
833 Table EV7: A complete list of signatures activated in a group of media.
834
835 Table EV8: Library of histidine kinase deletion mutants in PA14 used for a BCIP screen in PIA. All
836 strains in the collection of PA14 histidine kinase mutants described by the number from the
837 Hogan lab collection (DH number), the locus of the deleted gene, gene name (if available) and
838 position in the storage plate.

Figure 1

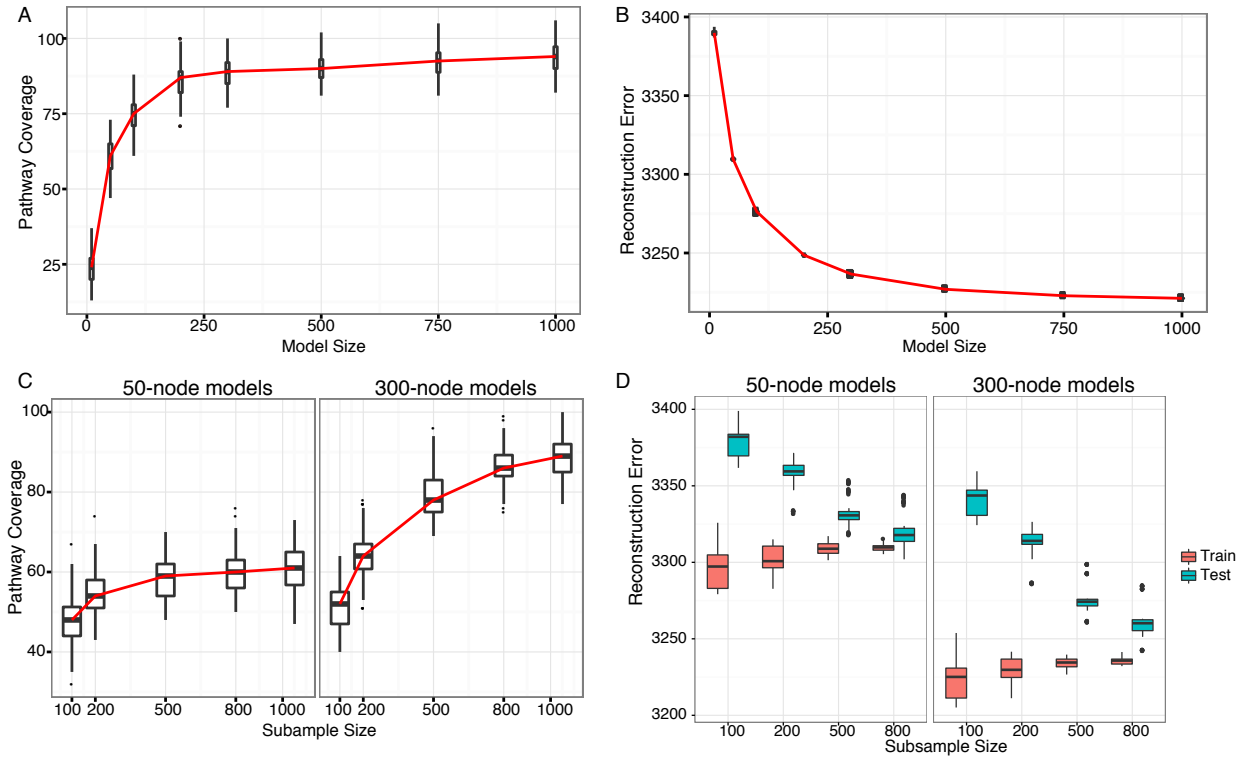

A
Pathway Coverage
Model Size
B
Reconstruction Error
Model Size
C
50-node models
300-node models
Pathway Coverage
Subample Size
D
50-node models
300-node models
Reconstruction Error
Subsample Size
Train
Test

Figure 2

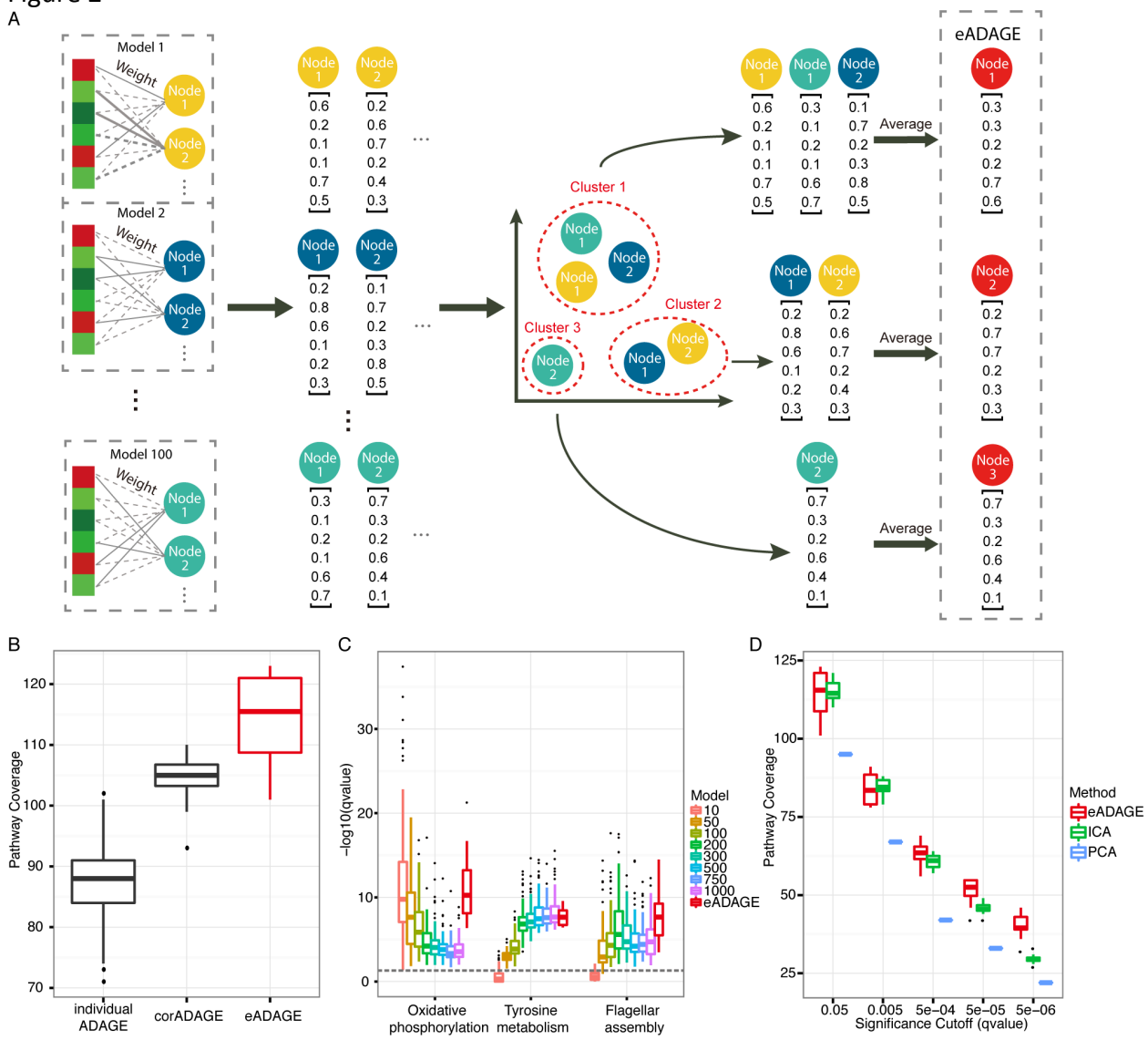
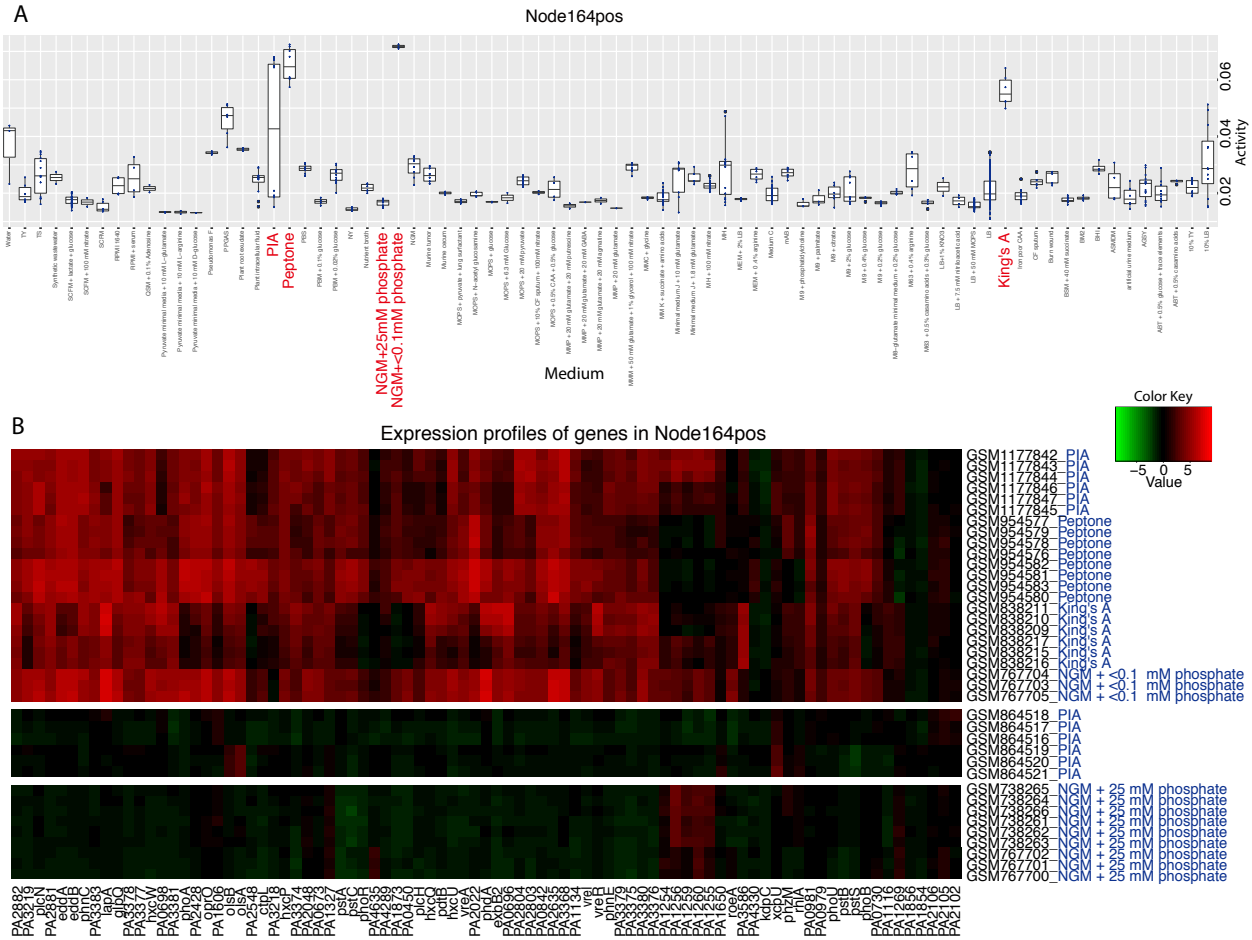
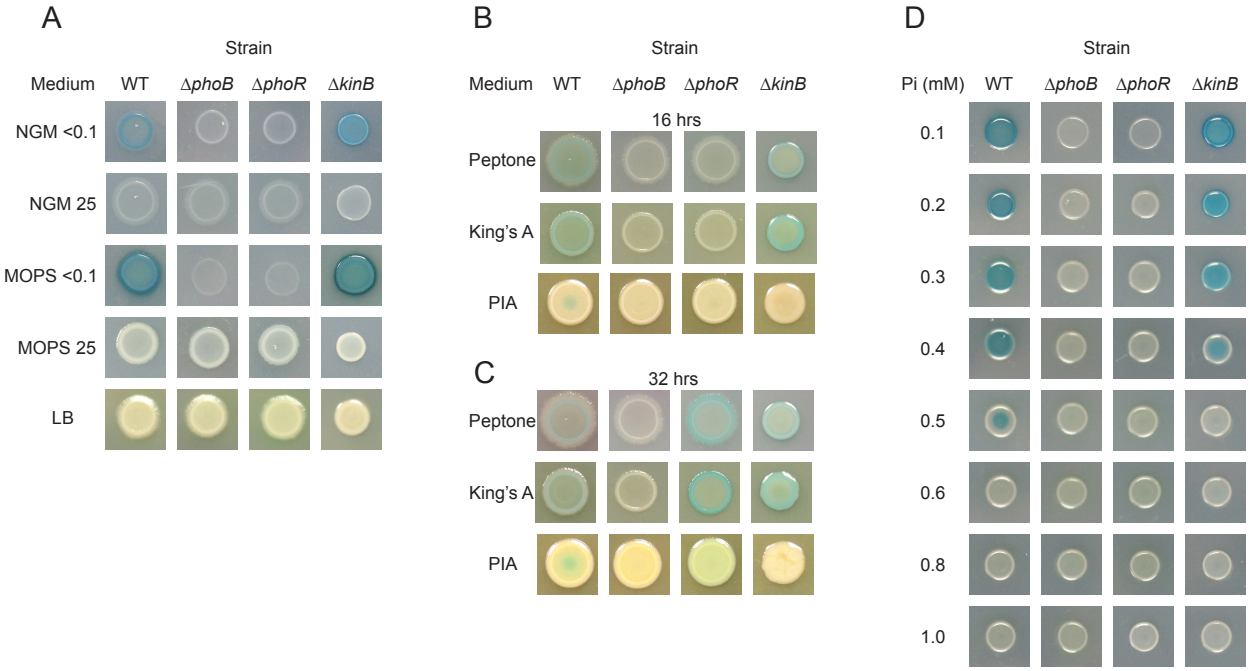

A
Model 1
Weight
Node 1
Node 2
Model 2
Model 100
Cluster 1
Cluster 2
Cluster 3
Average
eADAGE
Node 3
B
Pathway Coverage
individual ADAGE
corADAGE
eADAGE
C
−log10(qvalue)
Oxidative phosphorylation
Tyrosine metabolism
Flagellar assembly
Model
10
50
100
200
300
500
750
1000
eADAGE
D
Pathway Coverage
Significance Cutoff (qvalue)
Method
eADAGE
ICA
PCA

Figure 3


A
Node164pos
Activity
0.02
0.04
0.06
PIA
Peptone
King's A
NGM+25mM phosphate
NGM+<0.1mM phosphate
Medium
B
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

845 Figure 4



846