Generative neural networks separate common and specific transcriptional responses

Alexandra J. Lee[1,2], Dallas L. Mould[3], Jake Crawford[1,2], Dongbo Hu[2], Rani K. Powers[4], Georgia Doing[3], James C. Costello[5], Deborah A. Hogan[3], Casey S. Greene[2,6,7]

[1] Genomics and Computational Biology Graduate Program, University of Pennsylvania, Philadelphia, PA, USA

[2] Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania, Philadelphia, PA, USA

[3] Department of Microbiology and Immunology, Geisel School of Medicine, Dartmouth, Hanover, NH, USA

[4] Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, MA, USA

[5] Department of Pharmacology, University of Colorado School of Medicine, Denver, CO, USA

[6] Center for Health AI, University of Colorado School of Medicine, Denver, CO, USA

[7] Department of Biochemistry and Molecular Genetics, University of Colorado School of Medicine, Denver, CO, USA

Abstract:

Genome-wide transcriptome profiling identifies genes that are prone to differential expression across contexts (“common DEGs”), as well as specific changes relevant to the experimental manipulation. Distinguishing common DEGs from those that are specifically changed in a context of interest will allow more efficient inference of relevant mechanisms and a more systematic understanding of the biological process under scrutiny. Currently, common changes can only be identified through the laborious manual curation of highly controlled experiments, an inordinately time-consuming and impractical endeavor. Here we pioneer a method for identifying common patterns using generative neural networks. This method produces a background set of transcriptomic experiments from which a gene and pathway-specific null distribution can be generated. By comparing the set of differentially expressed genes found in a target experiment against the background set, common results can easily be separated from specific ones. This “Specific cOntext Pattern Highlighting In Expression data” (SOPHIE) method is broadly applicable to new platforms or any species with a large collection of unannotated gene expression data. We apply SOPHIE to diverse datasets including human, including human cancer, and bacterial datasets. SOPHIE recapitulates previously described common DEGs, and our molecular validation indicates it detects highly specific, but low magnitude, biologically relevant, transcriptional changes. SOPHIE’s measure of specificity can complement log fold change activity generated from traditional differential expression analyses by, for example, filtering the set of changed genes to identify those that are specifically relevant to the experimental condition of interest. Consequently, these results can inform future research directions.

Introduction:

Genome-wide gene expression analysis allows investigators to examine how gene expression changes under the tested experimental stimulus or varies across different states. This is useful for comparison, for example, of patient samples from different individuals with the same disease. When interpreting the results of these analyses, attention tends to focus on controlling false discoveries [3–6] – i.e. differential gene expression patterns that arise due to noise or variation during measurement. In addition to false discoveries, however, certain genes were found to be commonly differentially expressed across a diverse panel of environmental stresses.[7] The response of this collection of genes was termed the environmental stress response (ESR). Despite the ESR being described more than two decades ago[7], compared to false discoveries, less attention has been paid to controlling for these commonly changed genes. These findings include differential expression changes that are observed across experiments regardless of the experimental manipulation. While these common findings are actual changes, not false discoveries, they provide little contextual information or insight into the biological process being queried as they are observed in many unrelated experiments. Not knowing which discoveries are common versus specific can lead to misinterpretations or lack of specificity in interpreting the results, so it is important to account for these different types of findings in addition to correcting for false discoveries. Both gene-based[1,7] and pathway-based[2] analyses can return common results.

However, controlling for common findings is inordinately time-consuming, limiting the use of protocols that would identify them. Current methods rely on manual curation to generate a background set of experiments. These experiments are analyzed to identify genes and pathways that are common based on the frequency at which they are differentially expressed in the background experiments.[1,2] Switching to a new measurement platform, experimental design, analytical approach, incorporating new data, or examining a different organism requires re-curation in the new context to derive an appropriate background distribution. Even when data are readily available, curating and analyzing hundreds of experiments requires a significant time investment.

We introduce a method based on latent space transformation in multi-layer neural networks that makes it possible to automate the analysis of commonly differentially expressed genes ("common DEGs"), termed Specific cOntext Pattern Highlighting In Expression data (SOPHIE). This approach requires enough gene expression data to generate synthetic measurements; however, the data do not need to be curated by experimental design, removing a usually time-consuming step. Such data are readily available through NCBI Gene Expression Omnibus (GEO), Short Read Archive (SRA), and other repositories. Many are already processed for reuse through projects such as recount2[8] or ARCHS4[9]. Because SOPHIE relies on generating synthetic data that match a user-selected template experiment, it can be applied to arbitrary downstream analytical workflows, which could be differential expression (DE) analysis, pathway analysis, or other methods, to provide a background distribution of common findings. Without the need for manual curation, SOPHIE can expand the list of genes for follow-up by identifying genes that are context-specific, but have subtle signals and are thus understudied. SOPHIE can also filter the list of genes for functional validation if we limit our list of genes to those that are both differentially expressed and highly specific. Overall, SOPHIE's specificity score can be a complementary indicator of activity compared to the traditional log fold change measure and can help drive future analyses.

We applied SOPHIE to identify common DEGs in human cancer cell line microarray data and the results are consistent with prior microarray-based methods. Furthermore, we find consistent common DEGs using human RNA-seq data, demonstrating that SOPHIE is robust across platforms. SOPHIE is also generalizable as shown by application to a different organism,

Pseudomonas aeruginosa (*P. aeruginosa*). SOPHIE analysis of alternative carbon utilization in *P. aeruginosa*[10] revealed gene expression changes that were specific to different levels in the hierarchy of the carbon catabolite repression cascade. This analysis revealed a context-specific response to arginine metabolism, which would be undetected in a traditional differential expression analysis due to its low magnitude. Based on our SOPHIE results, we hypothesize that these selected genes are specific to arginine catabolism. Experimental data support the prediction that arginine catabolism is specifically perturbed by some, but not all mutants, in the pathway. This demonstrates that SOPHIE can successfully identify candidate genes that are specifically relevant to the context in question, and difficult to uncover through previously developed analyses.

Results:

Latent space transformation supports template-based differential expression analysis

In order to generate a background set of transcriptome experiments, we trained a generative neural network, in this case a variational autoencoder (VAE), on all RNA-seq samples in SRA from recount2[8]. This dataset, which included measurements for 17,555 genes across 49,651 human samples, provided a wide and diverse representation of gene expression patterns to learn from (Figure 1A). Using a previously reported[11] model architecture, we found that training and validation set losses for this model stabilized after 40 epochs, which is the number of times the VAE was trained over the input dataset (Figure 1B). Using this VAE, we sought to create synthetic gene expression experiments that follow the general patterns observed in real experiments with specific differences introduced.

Intuitively, we can think of the latent space as capturing key gene expression patterns across the entire compendium. Imagine an experiment of interest that tests a binary response to an experimental manipulation: this represents a pre-defined 'template set'. We used the above VAE to generate new samples from a pre-selected template set of samples from recount2 (Figure 1C). Using the trained VAE, we encoded the template samples into the latent space and shifted these samples to create new simulated samples. Further details can be found in the Methods. When we encoded this template experiment into the latent space, the relative positioning of samples captures the change between the two conditions, and the absolute position represents the biological context of this specific experiment. When we shifted the template experiment in the latent space, we created a new experiment with the same magnitude of change within the latent space but within different contexts. Then, we applied the decoder network to the shifted representation to generate new genome-wide measurements. Using a small molecule treated human adult erythroid cell experiment (SRP061689)[12] as a template, we simulated three experiments (Figure 1D). Compared to the original template experiment, the simulated experiments had a reduced number of differentially expressed genes, which was due to the VAE shrinking the variance of the simulated data. This variance shrinkage property was previously observed in Akrami et al.[13] In general, the latent space shifting preserves differences between groups, but this signal can be distorted depending on where the experiment was shifted to (i.e. the new location can cause the experiment to have a more compressed difference between groups). Overall, the three simulated experiments in figure 1D demonstrate that our simulation can generate experiments with similar structures but with different sets of differentially expressed genes. This observation is consistent with a previous report by Lee et al., which used this same VAE approach to simulate an experiment and found that the new experiment contained related groups of differentially expressed genes that were distinct from

the original template experiment.[11] SOPHIE uses this VAE approach to simulate realistic-looking transcriptome experiments that serve as a background set for analyzing common versus specific transcriptional signals.

Simulation-based common DEGs recapitulate curation-derived ones

Studying common differential expression has been challenging because it requires extensive manual curation. We sought to compare the common DEGs by SOPHIE with those identified in a prior report. This prior study curated 2,456 human microarray datasets from the GPL570 (Affymetrix Human Genome U133 Plus 2.0 Array) platform[1] to identify common DEGs.[1] This study provided a list of genes ranked based on how frequently they were identified as differentially expressed across approximately 600 experiments, which we refer to as the Crow et al. results. We trained a VAE on a different collection of microarray data that accompanied another prior report of commonly differentially expressed pathways.[2] This second dataset we refer to as the Powers et al. results, which included 442 differential comparisons (2,812 human microarray datasets) testing the response of small-molecule treatments in cancer cell lines. For this analysis, we selected an arbitrary template experiment (GSE11352 examined estradiol exposure in breast cancer cells[14]), generated 25 simulated experiments through latent space transformation, and calculated differential expression association statistics (DE association statistics) for each experiment comparing treated versus untreated cells (Figure 2A). We calculated the percentile of genes by their median $\log_2$ fold change across the 25 simulated experiments. Finally, we intersected the set of genes in this dataset with those in Crow et al. and examined their concordance using Spearman correlation (Figure 2B). Results between Crow et al. and our VAE trained on Powers et al. were concordant, particularly for the genes in the highest and lowest percentiles, the most and least commonly differentially expressed genes respectively. In particular, there was a significant (p-value=1e-49) over-representation of SOPHIE identified common DEGs within the common changes that Crow et al. identified.

In general, transcriptome analysis approaches do not effectively translate between different platforms (RNA-seq, microarray) and datasets. To demonstrate that SOPHIE easily extends to new platforms, we trained a VAE on human RNA-seq data from recount2 to identify common DEGs. We selected an arbitrary template experiment from recount2 (SRP012656 examined non-small cell lung adenocarcinoma tumors[15]), simulated 25 new experiments, and calculated differentially expressed genes using DESeq2. For this template experiment, primary non-small cell lung adenocarcinoma tumors were compared to adjacent normal tissues for 6 never-smoker Korean female patients. We again calculated the percentile for genes by their median $\log_2$ fold change in the simulated data, intersected with the Crow et al. set, and examined concordance (Figure 2C). We found, again, a significant (p-value= 2e-15) over-representation of SOPHIE-identified common DEGs shared with the Crow et al. analysis. Since the common findings from Crow et al. were based on a manually curated set of experiments, extending their analysis to use a new platform would require re-curation, which would be time-consuming. Thus, it is advantageous that SOPHIE need only retrain on a new dataset to extend its capabilities to new datasets and platforms.

We also noticed a set of genes in the bottom right corner of Figure 2C with a high percentile score were common DEGs in RNA-seq but not in Crow et al. We did not observe a corresponding set in the upper left corner, suggesting that RNA-seq captured the microarray-based common DEGs, but prior microarray-based reports lacked certain RNA-seq specific ones. This subset of genes was specifically differentially expressed in RNA-seq and not in array data, suggesting that platform differences underlie this effect. Some preliminary experiments showed that these RNA-seq common DEGs tended to have a lower expression compared to

those common genes identified using both the array and RNA-seq platform (Figure S1). The VAE appeared to artificially boost the expression of these RNA-seq common DEGs so that they were found to be differentially expressed. Unlike the array data, the RNA-seq data has a larger variance and so the effects of the VAE are more pronounced, affecting genes in the outliers of the compendium distribution, which are these RNA-seq commonly changed genes.

Finally, when we extended this analysis to a different organism, *P. aeruginosa*, we observed the same concordance (R^2 = 0.449) between SOPHIE-generated percentiles compared to those generated using a manually validated dataset, GAPE (Figure 2D).[16] GAPE contained a collection of 73 array experiments from the GPL84 platform. We found a significant over-representation (p=1e-139) of SOPHIE identified common DEGs within the GAPE set of common DEGs.

Having shown that SOPHIE's commonly changed gene percentiles can recapitulate percentiles of genes using a manually curated dataset (Crow et al. or GAPE), we next examined the robustness of these common patterns. We compared SOPHIE percentiles from different simulations using the same template experiment and showed that we get a very strong correlation (R^2 = 0.907), especially for high and low percentile genes (Figure 2E). The genes in the middle percentiles are more sensitive to changes so the signal is less clear. This noise is more pronounced when we compare the percentiles generated using two different template experiments (Figure 2F). Overall, we observe consistent common DEG percentiles across different template experiments (R^2=0.572). From this set of analyses, we have validated that SOPHIE was able to identify commonly changed genes previously reported by Crow et al and GAPE. While Crow et al. and GAPE rely on having a manually curated dataset, SOPHIE identified these genes in a more scalable and automated way, leveraging existing gene expression data to simulate a background set of experiments to use as a reference.

Simulation-based commonly differentially expressed pathways recapitulate common-derived ones

In addition to common DEGs, we also examined common differentially expressed pathways. While there is some variation between the ranking of common DEGs, grouping genes into pathways may find more robust common signals. For this analysis we used a set of common differentially expressed pathways reported by Powers et al. We used the VAE trained on the data from that same report, to simulate 25 new experiments from the same template experiment used previously (GSE11352), and used GSEA[17] to identify pathways enriched in differentially expressed genes (Figure 3A). We compared the percentile of pathways determined using data simulated from the VAE with those reported by Powers et al. and found strong concordance (R^2= 0.65, Figure 3B).

SOPHIE can also be applied using other pathway analysis methods. We easily extended SOPHIE to use multiple different enrichment methods (Figure 3C) and examined the common findings. We selected 4 enrichment methods (GSEA, GSVA, CAMERA, ORA) from Geistlinger et al.[18] We selected methods if 1) they could be applied to both RNA-seq and array data and 2) they covered a wide range of statistical performance measures including runtime, the number of gene sets found to be statistically significant and the type of method – self-contained versus competitive. Overall, the percentile of common pathways enriched varied between enrichment methods, likely due to the different assumptions and modeling procedures (Figure 3D, S2). Therefore, scientists will need to use a method-specific common correction approach. Similar to our analysis of common DEGs, compared to Powers et al., SOPHIE can automatically identify

commonly changed pathways. Additionally, SOPHIE can be easily customized to use different enrichment methods depending on the analysis.

Common DEGs may correspond to hyperresponsive pathways

We next examined how the genes that are commonly differentially expressed are related to previously reported transcriptional patterns in order to gain insight into the role of these commonly changed genes. We identified common DEGs using recount2, which is a heterogeneous compendium of human gene expression data containing a range of different types of experiments and tissue types. The recount2 data was decomposed into latent variables (LV), representing gene expression modules, some of which were aligned with known curated pathways, in prior work.[19] In these latent variables, genes had some weighted contribution, and we found that the median number of genes with non-zero weight was 2,824. We divided genes into a set of common DEGs, which were genes that were in the 60th percentile and above in our recount2 analysis (Figure 2C), and all other genes. We found that the commonly changed genes had non-zero weight to roughly the same number of latent variables as other genes (Figure 4A, p-value = 0.239 comparing the median between gene groups). However, common DEGs were found among the highest weights (the 98th percentile and above for each latent variable) for fewer latent variables than other genes (Figure 4B). Taken together, these results suggest that common DEGs contribute to as many latent variables as other genes (i.e. have a non-zero weight), but common genes occur less frequently among the highest weight genes. Overall, the wide coverage across latent variables but lack of high weight contributions indicate that common DEGs mainly contribute to a few pathways. Given the small number of pathways, one possibility for why these genes are commonly changed might be related to a few hyper-responsive pathways.

Since these latent variables tend to be associated with particular biological processes, we wanted to test if there were any latent variables, and thereby processes, that contained a large fraction of common DEGs. If there exist latent variables that were primarily composed of common DEGs, this might lend insight into the role of commonly changed genes. For this analysis, we ranked latent variables by the proportion of commonly shifted genes at the 98th percentile and above. Overall, many of these latent variables were associated with immune responses, signaling, and metabolism. This finding is consistent with the hypothesis that these commonly changed genes are related to hyper-responsive pathways. One example latent variable, that contained a high proportion of commonly changed genes compared to other genes (proportion of commonly changed genes > 0.5), was LV61 (Figure 4C, Table S1). This latent variable included pathways related to immune response (Neutrophils), signaling (DMAP ERY2), and wound healing (megakaryocyte platelet production).

We performed a similar analysis to examine common patterns in *P. aeruginosa* data. Again, we leveraged an existing model. Tan et al. previously created a low dimensional representation of the *P. aeruginosa* compendium using a denoising autoencoder, called eADAGE, where some of the latent variables were found to be associated with KEGG pathways and other biological sources of variation.[20,21] Using this existing eADAGE model, we created a gene-gene similarity network where the correlation between the eADAGE representation was used to connect genes. After performing a community detection analysis, we discovered that commonly changed genes tended to cluster in fewer communities compared to other genes (Figure 4D). Furthermore, commonly changed genes had a slightly higher median degree in the eADAGE similarity network compared to other genes (Figure 4E). These observations were consistent with an analysis that found a set of virulence-related transcriptional regulators that target multiple pathways.[22] Together, these data suggest that, like the patterns we observed in the human

dataset, there are relatively few communities that commonly changed genes contribute. These few communities containing commonly changed genes were highly connected to other communities. Therefore, we conclude that these common-driven communities might correspond to hyper-responsive pathways.

SOPHIE-identified genes specific to arginine catabolism

In general, differential expression analyses often aim to understand the genetic causes and downstream consequences of gene expression. However using traditional p-values and log fold change criteria, such datasets often contain hundreds of genes, many of which are not specific to the phenotype of interest. Using SOPHIE, we can filter these differentially expressed gene sets to identify those that are specific to the context of the experiment. Likewise, for experimental conditions that uncover fewer novel genes of interest, SOPHIE can highlight those that show modest, but specific changes that would be missed by traditional DE analysis. To illustrate these points, we chose the experiment E-GEOD-33245 as a template because it investigated metabolic decision making, a process known as carbon catabolite repression, that is important for *P. aeruginosa* pathogenicity[23]. This decision depends on a complex mechanism involving both transcriptional and translational regulation that results in both direct and indirect effects on the transcriptome. A previous analysis by Sonnleitner et al.[10], suggested that the production of catabolic enzymes and transporters is controlled by the translational co-repressor Crc (Figure 5A). Crc activity can be inhibited by *crcZ*, a small RNA, which sequesters the Crc protein[24]. CbrB meanwhile controls levels of the *crcZ* small RNA.

We focused on the comparisons between WT and isogenic Δ*cbrB* and *Δcrc* mutants. In the absence of the transcription factor CbrB or the translational co-repressor Crc, 156 and 149 genes were differentially expressed ($|\log_2 FC| > 1$, FDR-adj p-value < 0.05), respectively, relative to wild type. Of these DE genes, we focused first on those that had a low z-score, indicating a high likelihood of it being part of a common response. *ArcB* (z-score: 1.09) was identified as the eighth-most frequently differentially expressed gene in the GAUGE-annotated Δ*crc* DE dataset, with DE in 40 out of 73 studies. More broadly, genes considered commonly differentially expressed by SOPHIE and GAPE accounted for 79 and 30 of the differentially expressed genes in Δ*cbrB* and *Δcrc* comparisons respectively. Both comparisons included the genes *pqsA, nosZ*, *pqsE*, and *ccoP2*. Though CbrB and Crc are part of the same metabolic regulatory pathway, SOPHIE found genes involved in arginine catabolism (*argA*) and arginine transport (*aotJQMP)* changed by less than 2-fold in both samples. However, the specificity (high ranked z-score, Table 1) was high in Δ*cbrB* but not Δ*crc* (Figure 5B). We constructed *P. aeruginosa* strain PA14 mutants Δ*cbrB* and Δ*crc* and found that only Δ*cbrB* was defective for arginine catabolism (Figure 5D). This result supports the model that arginine metabolism is specifically regulated by CbrB, consistent with published data by other studies[25,26], and highlights the utility of SOPHIE to drive the prioritization of genes for follow-up analysis of candidate differentially expressed genes. This method is particularly powerful for those genes that do not change very much but do so more than in the background simulated experiments (i.e. specific genes). It is appreciated that small expression changes can have biological significance, but we often choose not to pursue these genes because it is easier to start with those that show the largest difference in expression. However, SOPHIE provides strong confidence scores that highlight biologically important, but less studied genes for further analysis. By leveraging publicly available data, SOPHIE identified

candidate specific genes. Independently, we experimentally validated that these genes played a specific role in the context of the target experiment. SOPHIE can therefore successfully predict biologically relevant gene targets that further our mechanistic understanding and drive future analyses.

Discussion

We introduce a method, SOPHIE, named after one of the main characters from Hayao Miyazaki's animated film *Howl's moving castle*. Sophie's outward appearance as an old woman, despite being a young woman that has been cursed, demonstrates that initial observation can be misleading. This is the idea behind our approach, which allows users to identify specific gene expression signatures that can be masked by common background patterns.

SOPHIE automatically identified commonly differentially expressed genes and pathways using public gene expression compendia. SOPHIE returned consistent genes and pathways, by percentile, compared to previous results using both human[2,8] and bacterial[21] datasets. Furthermore, experimental validation confirmed a group of genes that SOPHIE predicted to show context-specific differential expression. In contrast to using a manually curated dataset, SOPHIE can be easily extended to new contexts and can generate a background distribution of experiments for any organism with public data available. These background experiments define a set of genes and pathways that are commonly changed across many different experimental conditions. These background sets of changes, provide context to individual experiments, highlighting specific gene expression changes and thus giving insight into mechanisms relevant to specific contexts including disease conditions.

Compared to prior work using manually curated datasets[1,2], SOPHIE demonstrates consistent results but using an automated process. In short, SOPHIE identifies the same common patterns but in a fast and scalable way. However, there was a subset of genes that were specifically differentially expressed using SOPHIE but not found using the manually curated background. In one case, SOPHIE is using RNA-seq while the manually curated data is based on hybridization technology (microarray). Some initial experiments showed that this inconsistency is likely due to platform differences and how the VAE handled these two different data types. In addition to platform differences, the context of the background dataset was shown to influence the commonly changed pathways detected. We found that commonly altered pathways were more sensitive to different contexts compared to commonly changed genes. One speculation for this observation comes from recent work from Sazali et al.[27] Given that information flows from a stimulation that activates proteins within pathways, and these proteins regulate gene expression, a context-specific signal will eventually lead to changes in gene expression. Thinking about the flow of information, measuring pathway activity (pathway enrichment) will be more sensitive to context compared to measuring differential expression in individual genes. Since the genes are regulated as a group, we would expect to see coordinated changes in expression that are correlated with the specific context. Examining the expression of individual genes wouldn't necessarily reveal this correlation with context. Using SOPHIE, we identified commonly changed cancer pathways, that are not necessarily commonly altered pathways in general datasets. Overall, SOPHIE results are consistent with previous findings, but we also identified differences that might indicate there exists a hierarchy of common changes depending on the platform and context.

Building on the discovery of these common signals, we also examined the potential role of these commonly changed genes. These common DEGs appear to contribute to a small number of hyperresponsive pathways (Figure 4). This supports the observation that genes found to be

differentially expressed across different contexts may not be informative about the experimental manipulation of interest. Therefore, considering specificity can be complementary to using log fold change activity to study biological processes.

One limitation is that our template experiments test two conditions, but there are many different types of experiments (e.g. time course). Do commonly changed genes vary based on experimental design? To answer this question, we would need to curate more experiments testing different experimental designs. We would also need to determine how to group samples to perform a differential expression analysis or develop a new metric to define how many genes change. Another limitation to our study is that we are using a random linear shift to simulate experiments. While this linear shift is using a location drawn from the known distribution of gene expression data, this shift currently doesn't allow us to vary or shift along certain axes, such as tissue type or drug. If SOPHIE could be extended to simulate background experiments along a specific axis, like tissue type, then we could ask if there are different sets of commonly changed genes that come up as we vary along the tissue axis versus the drug axis. To answer this question we would need to have a deeper understanding of the structure of the latent space and what is being captured. These questions can help lead to an improved understanding of common signals and the type of correction that might be needed. Additionally, while SOPHIE is mostly portable, more work needs to be done to define the optimal neural network architecture for different data types – i.e. different platforms. Depending on the data type, there likely exists some optimal neural network architecture that preserves the underlying structure in the data. Therefore, some additional training of the VAE is required before applying SOPHIE to datasets of interest.

SOPHIE is a powerful tool that can be used to drive how we study mechanisms underlying different cellular states and diseases. With SOPHIE, we can identify commonly changed genes that might be useful for diagnostic[28] and detection[29] purposes. We can also identify specific signals that point to possible treatment options[30]. In general, studies trying to uncover these genetic mechanisms tend to focus on prominent biological signals – those genes that are strongly differentially expressed. However, with SOPHIE we can start to glean information about those genes that are subtle but specifically relevant to the biology in question. Overall, SOPHIE is a powerful tool that can complement existing traditional analyses to separate specific versus common differentially expressed genes and pathways. These context-specific genes and pathways include both subtle changes that are largely unexplored and prominent changes that might point to areas of treatment and biomarker development. In general, SOPHIE can easily be applied across a range of different datasets to help drive discovery and further understanding of mechanisms.

The best way to deploy SOPHIE in practice will depend on the scientific question and the ease with which leads can be validated. The software associated with this paper is available on github (https://github.com/greenelab/generic-expression-patterns) and users can modify the notebooks for their own analysis following the instructions in the README file.

Methods

Gene expression datasets

We used three complementary gene expression compendia in this work. Two were sets of assays of human samples, one via microarray and the other via RNA-seq profiling. The third was a collection from the microbe *Pseudomonas aeruginosa*.

The first human compendium that we used contains gene expression data from Powers et al.[2] We downloaded the dataset from synapse on (October 7, 2020). This dataset contains samples from the Gene Expression Omnibus (GEO) measured on Affymetrix Human Genome U133 Plus 2.0 Array. Samples were selected based on the following criteria: having at least 2 replicates per condition and containing a vehicle control. The dataset included 442 experiments testing the response of small-molecule treatments in cancer cell lines. Samples were processed using the *rma* library to convert probe intensity values from the *.cel* files to $\log_2$ base gene expression measurements, and these gene expression values were then normalized to 0-1 range per gene. This resulted in an expression matrix that contains 6,763 genes and 2,410 samples.

The second human compendium that we used includes human RNA-seq data from recount2[8]. We downloaded all SRA data in recount2 as RangedSummarizedExperiment (RSE) objects for each project id using the recount library in Bioconductor (version 1.12.0). Raw reads were mapped to genes using Rail-RNA[31], which includes exon-exon splice junctions. Each RSE contained counts summarized at the gene level using the Gencode v25 (GRCh38.p7, CHR) annotation provided by Gencode.[32] These RSE objects include coverage counts as opposed to read counts, so we applied the scale_counts function to scale by sample coverage (average number of reads mapped per nucleotide). The compendium contained 49,651 samples with measurements for 58,129 genes. Our goal was to compare percentiles with ones provided by Crow et al.[1], which required us to map the ensembl gene ids in recount2 to HGNC symbols. We used the intersection of genes between the recount2 and Crow et al. sets. This resulted in a gene expression matrix of 49,651 samples and 17,755 genes. We then normalized gene expression values to a 0-1 range per gene. This recount2 compendium contained a heterogeneous set of gene expression experiments – 31 tissue types (i.e. blood, lung), 57 cell types (i.e. stem, HeLa), multiple experimental designs (i.e. case-control, time-series).

The last compendium contained *P. aeruginosa* gene expression data that was collected and processed as described in Lee et al.[11] The dataset was originally downloaded from the ADAGE[21] GitHub repository (https://github.com/greenelab/adage/tree/master/Data_collection_processing). Raw microarray data (measured on the release of the GeneChip *P. aeruginosa* genome array and the time of data freeze in 2014) were downloaded as .cel files. Then *rma* was used to convert probe intensity values from the *.cel* files to $\log_2$ base gene expression measurements. These gene expression values were then normalized to 0-1 range per gene. The resulting matrix contained 989 samples and 5,549 genes that represent a wide range of gene expression patterns including characterization of clinical isolates from cystic fibrosis infections, differences between mutant versus WT, response to antibiotic treatment, microbial interactions, and the adaptation from water to GI tract infection.

SOPHIE: Specific cOntext Pattern Highlighting In Expression

Train VAE: We built a multi-layer variational autoencoder (VAE) that extended from the previously published Tybalt model.[33] The model was built using Keras (version 2.3.1) with a TensorFlow backend (version 1.15.4). The structure of the VAE is composed of an encoder neural network, which compresses the input gene expression data into 30 latent space features, and a decoder neural network, which decompresses the data back into raw gene expression space. The architecture of the encoder and decoder neural networks includes an intermediate

layer with 2,500 features and a hidden layer of 30 latent space features with a rectified linear unit (ReLU) activation function to combine weighted nodes from the previous layer. The VAE was optimized using binary cross-entropy loss, which included the reconstruction loss as well as a Kullbach-Leibler (KL) divergence term to constrain the latent space to follow a normal distribution. We used the same neural network architecture used in Lee et al. due to the success. We used the same strategy outlined in Lee et al.[11] to train the VAE. We performed a 90:10 split of the data for training and validation. The hyperparameters were manually adjusted based on a visual inspection of the validation loss outputs. Our optimal hyperparameter settings were: learning rate of 0.001, a batch size of 10, warmups set to 0.01. We trained 3 VAE models using recount2 (40 epochs), Powers et al. (40 epochs), and the *P. aeruginosa* (100 epochs) compendia.

Simulate gene expression experiments: Our simulation approach was an extension of the experiment-level simulation approach from Lee et al.[11] We selected a template experiment from our compendium (SRP012656 from recount2, GSE11352 from Powers et al., and E-GEOD-33245 from *P. aeruginosa*). We simulated a new experiment by linearly shifting the selected template experiment to a new location in the latent space. This new location was randomly sampled from the distribution of the low dimensional representation of the trained gene expression compendium. The vector that connects the template experiment and the new location was added to the template experiment to create a new simulated experiment. This process was repeated 25 times to create 25 simulated experiments based on the single template experiment.

Differential expression analysis: For the recount2 compendium we used the DESeq module in the DESeq2 library[34] to calculate differential expression values for each gene comparing the two different conditions in the selected template experiment (SRP012656). The template experiment contained primary non-small cell lung adenocarcinoma tumors and adjacent normal tissues of 6 never-smoker Korean female patients. The differential expression analysis compared tumor vs normal. Following a similar procedure for the array-based datasets (*P. aeruginosa* compendium and the Powers et al. compendium) we used the eBayes module in the limma library[35] to calculate differential gene expression values for each gene. The output statistics include $\log_2$ fold change between the two conditions tested and p-values adjusted by Benjamini-Hochberg's method to control for false discovery rate (FDR). The template experiment we used for the Powers et al. compendium is GSE11352, which examined the transcriptional response of MCF7 breast cancer cells to estradiol treatment. So the differential expression analysis compared samples untreated versus treated. The template experiment we used to the *P. aeruginosa* compendium is E-GEOD-33245, contained multiple comparisons examining the CbrAB system. The two we focused on for our analysis compared WT vs *cbrB* and *crc* mutants in LB media.

For the *P. aeruginosa* experiment, differentially expressed genes were those with FDR adjusted cutoff (using Benjamini-Hochberg correction) < 0.05 and $\log_2$ absolute value fold-change >1, which are thresholds frequently used in practice.

Calculate specificity of each gene (z-score): Using the association statistics from the differential expression analysis, we calculated a score to indicate if a gene was specifically differentially expressed in the template experiment. We calculated a z-score for each gene using the following formula:

$$z - score\ of\ gene\ A = \frac{log_2FC\ gene\ A\ in\ template\ experiment - \ mean(log_2FC\ gene\ A\ in\ simulated\ experiments)}{var(log_2FC\ gene\ A\ in\ simulated\ experiments)}$$

Higher z-scores indicate a gene is specifically differentially expressed in the template experiment in reference to the null set of experiments (i.e. 25 simulated experiments).

Enrichment analysis (EA)

The goal of EA is to detect coordinated changes in prespecified sets of related genes (i.e. those genes in the same pathway or share the same GO term).

Our primary method was GSEA, for which we used the fgsea module from the fgsea library.[17,36] The method first ranks all genes based on the DE association statistics. In this case, we used the log_2 fold change. An enrichment score (ES) is defined as the maximum distance from the middle of the ranked list. Thus, the enrichment score indicates whether the genes contained in a gene set are clustered towards the beginning or the end of the ranked list (indicating a correlation with the change in expression). The statistical significance of the ES is estimated by a phenotypic-based permutation test to produce a null distribution for the ES (i.e. scores based on permuted phenotype). Each pathway was output with statistics including a Benjamini-Hochberg adjusted p-value. The pathways used in this analysis were the Hallmark pathways for the recount2 compendium and the Powers et al. compendium. For *P. aeruginosa* compendium, we used the KEGG pathways used in Tan et al.[21] These pathways can be found in the associated repository: https://github.com/greenelab/adage/blob/master/Node_interpretation/pseudomonas_KEGG_terms.txt

Other methods we used included: Gene Set Variation Analysis (GSVA)[37], Correlation Adjusted Mean Rank gene set test (CAMERA)[38], and Over-Representation Analysis (ORA). GSVA is a self-contained gene set test that estimates the variation of gene set enrichment over the samples independent of any class label. We used the gsva function from the gsva library. CAMERA is a competitive gene set test that performs the same rank-based test procedure as GSEA but also estimates the correlation between genes instead of treating genes independently. For CAMERA, we used the camera function that is part of the limma library.[39] Last, ORA is a method that uses the hypergeometric test to determine if there a significant over-representation of a pathway in the selected set of DEGs. Here we used the clusterProfiler[40] library but there are multiple options for this analysis.

Comparison of gene percentiles

We wanted to compare the percentile of human genes identified using SOPHIE (trained on recount2 and Powers et al. datasets) with the percentile found from Crow et al., which identified a set of genes as common DEGs based on how frequently they were found to be DE across 635 manually curated experiments. In their paper, they ranked genes as 0 if they were not commonly DE and 1 if there were commonly DE. Our genes were ranked from 1 to 17,754 based on their median absolute log_2 fold change value across the 25 simulated experiments. We linearly scaled the gene ranks to be a percentile from 0 to 100. Finally, we applied Spearman correlation to compare the ranks for each gene (Figure 2B, C).

We performed this same correlation analysis comparing SOPHIE trained on the *P. aeruginosa* compendium with percentiles generated from the GAPE project from the Stanton lab (https://github.com/DartmouthStantonLab/GAPE).[16] The GAPE dataset contained ANOVA statistics generated for 73 *P. aeruginosa* microarray experiments using the Affymetrix platform GPL84. We downloaded the differential expression statistics for 73 array experiments from the associated repository (https://github.com/DartmouthStantonLab/GAPE/blob/main/Pa_GPL84_refine_ANOVA_List_unzip.rds). For each experiment, we identified differentially expressed genes using $\log_2$ fold change > 1 and FDR < 0.05. We then calculated the percentile per gene based on the proportion that they were found to be differentially expressed. We compared these GAPE percentiles against those found by SOPHIE (Figure 2D).

We also compared percentiles of genes amongst two SOPHIE-generated results. This included comparing percentiles generated from two SOPHIE runs using the same template experiment (Figure 2E) and SOPHIE generated for two different template experiments (Figure 2F).

Comparison of pathway percentiles

We wanted to compare the percentile of pathways identified using SOPHIE (trained on recount2 and Powers et al. datasets) with the percentile based on the Powers et al. data. There was no pathway ranking provided in the publication, so we defined a reference ranking by calculating the fraction of the 442 experiments that a given pathway was found to be significant (FDR corrected p-value using Benjamini-Hochberg method <0.05) and used these rank pathways and then converted the ranking to a percentile as described above. We used the Hallmarks_qvalues_GSEAPreranked.csv file from https://www.synapse.org/#!Synapse:syn11806255. The file contains the q-values for the test: given the enrichment score (ES) of the experiment is significant compared to the null distribution of enrichment scores, where the null set is generated from permuted gene sets. Our percentile is based on the median Benjamini-Hochberg adjusted p-value across the simulated experiments. We compared our percentile versus the reference percentile using the Spearman correlation.

Latent variable analysis

The goal of this analysis was to examine why genes were found to be commonly differentially expressed – we sought to answer the question: are commonly changed genes found in more PLIER latent variables (LV)[41] compared to specific genes? The PLIER model performed a matrix factorization of the same recount2 gene expression data to get two matrices: loadings (Z) and latent matrix (B). The loadings (Z) were constrained to aligned with curated pathways and gene sets specified by prior knowledge to ensure that some but not all latent variables capture known biology. For this analysis, we focused on the Z matrix, which is a weight matrix that has dimensions gene by LV. For this analysis, common DEGs were above the 60th percentile (approximately the top 40% of genes were selected based on the distribution seen in Figure 4B) using the SOPHIE Trained on recount2. We calculated the coverage of common DEGs versus other genes across these PLIER latent variables. For each gene we calculated two values: 1) how many LVs the gene was present in (i.e. has a nonzero weight value according to the Z matrix), 2) how many LVs the gene was high weight in, using the 98th quantile for the LV distribution as the threshold.

Network analysis

In order to examine associations between commonly changed genes and pathways or functional modules in *P. aeruginosa*, we constructed a network of gene-gene interactions. Nodes in this network represent *P. aeruginosa* genes, and edges represent correlations between the eADAGE weight vectors of the two genes they connect. We constructed the network using the ADAGEpath R package, described in more detail in the associated manuscript.[21] To form the final network, we removed all edges (correlations) with a value between -0.5 and 0.5, and took the absolute value of the remaining edges (so negative edge weights became positive).

There are many existing methods to partition a network into well-connected, non-overlapping subnetworks, often referred to as communities. Using our gene similarity network, we sought to answer the question: Do commonly changed genes tend to occupy fewer network communities than a similar set of random genes, or do they tend to spread out across comparatively many communities? We chose two representative methods to divide the network into communities: (1) the Louvain method[42], as implemented in the *python-igraph* package[43], and (2) the "planted partition" model[44] (data not shown), as implemented in the *graph-tool* Python package[45]. In order to make a meaningful comparison between common and non-common DEGs, we sampled an equal number of both gene categories. This meant that the non-common DEGs were approximately degree-matched with the common DEGs (i.e., for each commonly changed gene we sampled a specific differentially expressed gene with approximately the same network degree). We performed this sampling procedure 1000 times. We then counted the number of communities containing at least one commonly changed gene, and compared this count to the distribution across the 1000 samples of the number of communities containing at least one sampled non-commonly changed gene.

In addition, we used the same eADAGE gene similarity network to compute several metrics describing individual network nodes, which we then compared between common and non-common DEGs. For both sets of genes, we calculated: (1) node degree, (2) edge weight, (3) betweenness centrality[46] (4) PageRank centrality[47]. For each of these metrics, we used the implementations in the *graph-tool* Python package. In contrast to the other metrics, betweenness centrality treats edge weights as "costs" (lower = better, as opposed to correlation or similarity measures where higher = better), so for the betweenness centrality calculation we transformed all edge weights by setting edge cost = 1 - correlation.

Strain Construction

Plasmids for making in-frame deletions of *cbrB* and *crc* were made using a *Saccharomyces cerevisiae* recombination technique previously described.[48] The arabinose-inducible *cbrB* expression vector was made using Gibson cloning. All plasmids were sequenced at the Molecular Biology Core at the Geisel School of Medicine at Dartmouth and maintained in *E. coli*. In frame-deletions constructs were introduced into *P. aeruginosa* by conjugation via S17/lambda pir *E. coli.* Merodiploids were selected by drug resistance and double recombinants were obtained using sucrose counter-selection and genotype screening by PCR. The *cbrB* and empty expression vectors were introduced into *P. aeruginosa* by electroporation and selected by drug resistance.

P. aeruginosa experiment

Bacteria were maintained on LB (lysogeny broth) with 1.5% agar. For strains harboring expression plasmids, 300 ug/mL Carbenicillin or 60 ug/mL Gentamycin was added. Yeast strains for cloning were maintained on YPD (yeast peptone dextrose) with 2% agar. Planktonic

cultures (5 mL) were grown on roller drums at 37° from single colonies for 16 h in LB (under antibiotic selection for the appropriate strains). The 16 h LB cultures were normalized to $OD_{600\ nm}$ = 1 in 2 mL, and a 250 µL aliquot of the normalized culture was used to inoculate three 5 mL cultures of M63 medium containing 10 mM arginine as a sole carbon source under inducing conditions (0.2% arabinose) for a starting $OD_{600\ nm}$ = 0.05. Inoculated cultures were grown at 37° C on the roller drum and cellular density ($OD_{600\ nm}$) was monitored using a Spec20 every hour for 8 hours. Each data point is representative of the average of the 3 replicates per day for 3 independent days.

Software

All scripts used in these analyses are available in the GitHub repository (https://github.com/greenelab/generic-expression-patterns) under an open-source license to facilitate reproducibility of these findings (BSD 3-Clause). The repository's structure is described in the Readme file. The notebooks that perform the validation experiment for common DEGs and pathways can be found in "human_general_analysis" (SOPHIE trained on recount2), "human_cancer_analysis" (SOPHIE trained on Powers et al.), and "pseudomonas_analysis" (SOPHIE trained on the *P. aeruginosa* compendium) directories. The notebooks that explore why genes are commonly differentially expressed can be found in "LV_analysis" directory. The notebooks for the network analysis can be found in the "network_analysis" directory. All supporting functions to run these notebooks can be found in "generic_expression_patterns_modules" directory. The virtual environment was managed using conda (version 4.6.12), and the required libraries and packages are defined in the environment.yml file. Additionally, scripts to simulate gene expression experiments using the latent space shifting approach are available as a separate module, called *ponyo*, and can be installed from PyPi (https://github.com/greenelab/ponyo). The Readme file describes how users can re-run the analyses associated with this manuscript or analyze their own data using this method. An example of how to apply SOPHIE to a new dataset can be found in "new_experiment" directory. All simulations were run on a CPU.

Acknowledgements:

The authors would like to thank David Nicholson, Ben Heil, and Milton Pividori for reviewing the software associated with this work and providing valuable feedback. We would like to thank Ben Heil for coming up with the name of this method. This work was supported by Grant GBMF4552 from the Gordon Betty Moore Foundation, NIH NHGRI R01, CA241747, CA231978, Cystic Fibrosis Foundation (CFF) HOGAN19G0 awarded to Deborah A. Hogan. Support for the project was also provided by DartCF through NIH NIDDK grant P30 DK117469 and the CFF Research Development Program (CF RDP) through CFF grant STANTO19R0 and bioMT through NIH NIGMS grant P20 GM113132. Finally we would like to thank Sabbi Lall from Live Sciences Editors for providing editing advice.

Figure Legends:

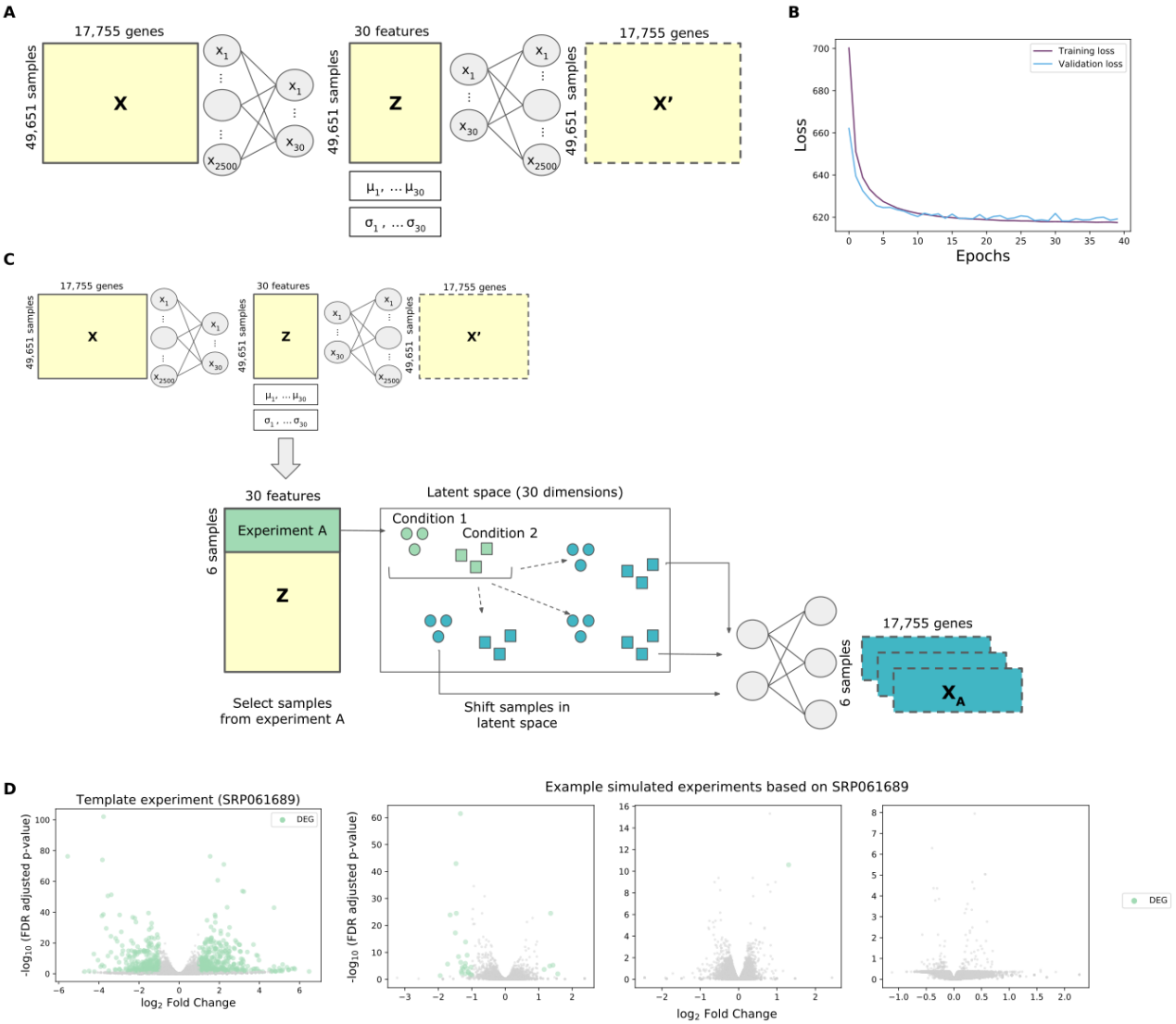


Figure 1: Using a VAE, we can simulate gene expression data. A) Architecture of the VAE. The input data is compressed into an intermediate layer of 2,500 features and then into a hidden layer of 30 latent features. Each latent feature follows a normal distribution with mean μ and variance σ. The input dimensions of the recount2 compendium have 49,651 samples and 17,755 genes. B) Training loss (purple) and validation loss (blue) plotted per epoch during training. C) Workflow to simulate gene expression experiments starting with a template experiment (green) and shifting the experiment in the latent space to generate a new simulated experiment (blue). D) Volcano plot of the original experiment SRP061689 (left) and 3 example simulated experiments using SRP061689 as a template (right) with differentially expressed genes highlighted in green. Differentially expressed genes (DEG) were selected as those that satisfied adjusted p-value < 0.05 and absolute $\log_2$ fold change > 1.

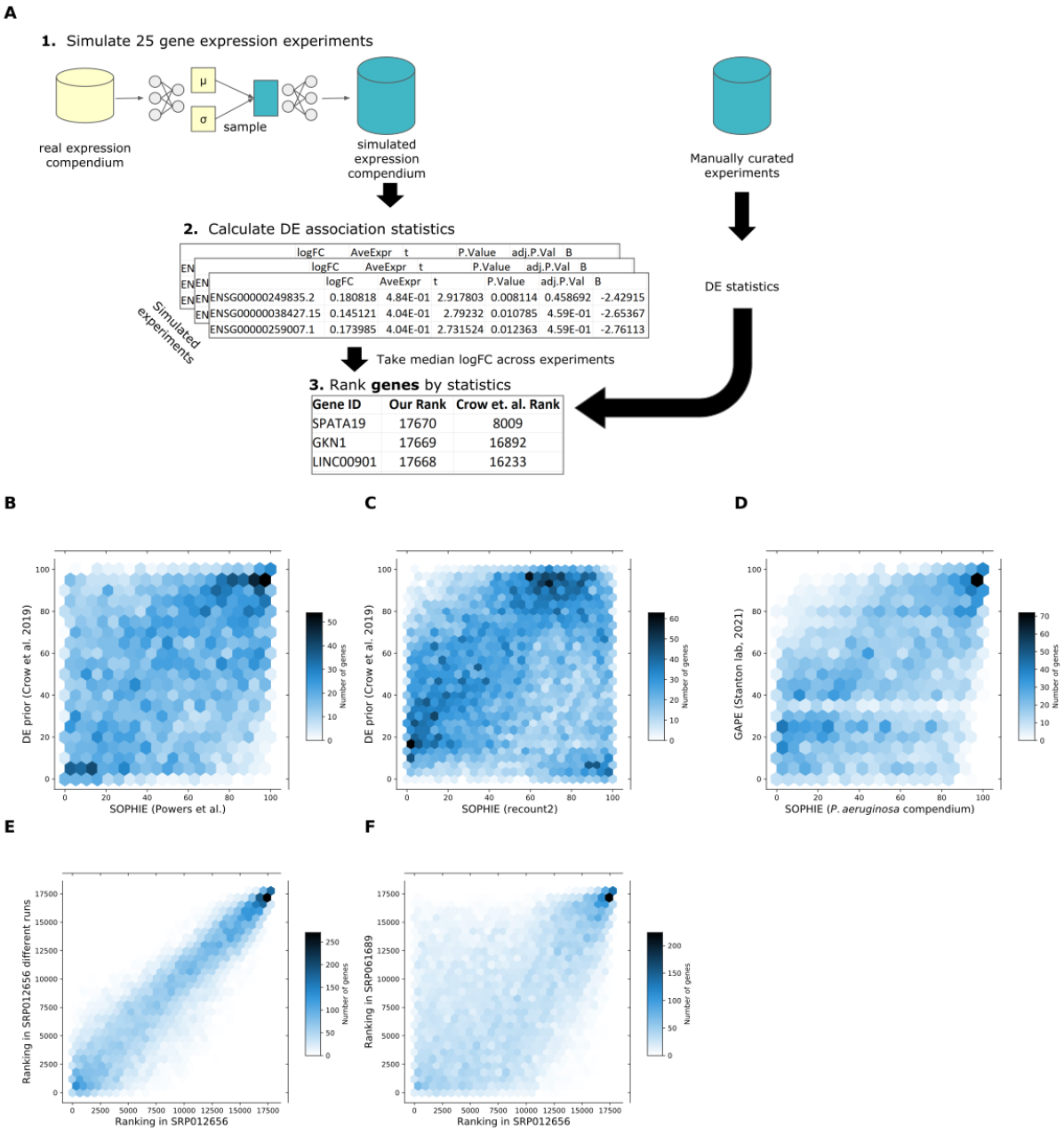


Figure 2: SOPHIE finds the same commonly shifted genes as previously identified using a manually derived dataset. A) Workflow describing how genes were ranked by how common they are differentially changed using SOPHIE versus using a manually curated set of experiments. B) Spearman correlation between gene percentiles using our simulated method trained on Powers et al. (array) using GSE11352 as a template (x-axis) versus percentiles using manually curated experiments from Crow et al. (y-axis) with significant over-representation of SOPHIE common DEGs in Crow et al. commonly changed genes (p-value=1e-49). C) Spearman correlation between gene percentiles using our simulated method trained on recount2 (RNA-seq) using SRP012656 as a template (x-axis) versus percentile using manually curated experiments from Crow et al. (y-axis) with significant over-representation of SOPHIE common DEGs in Crow et al.

commonly changed genes (p-value=2e-15). D) Spearman correlation between gene percentile using our simulated method trained on the *P. aeruginosa* compendium (array) using E-GEOD-33245 as a template (x-axis) versus percentile using manually curated experiments from GAPE. (y-axis) with significant over-representation of SOPHIE common DEGs in GAPE commonly changed genes (p-value=1e-139). E) Spearman correlation (R^2 = 0.907) between gene percentiles generated by SOPHIE using two runs of the same experiment (SRP012656) and F) Spearman correlation (R^2=0.572) between gene percentiles generated by SOPHIE using two different template experiments (SRP012656 and SRP061689).

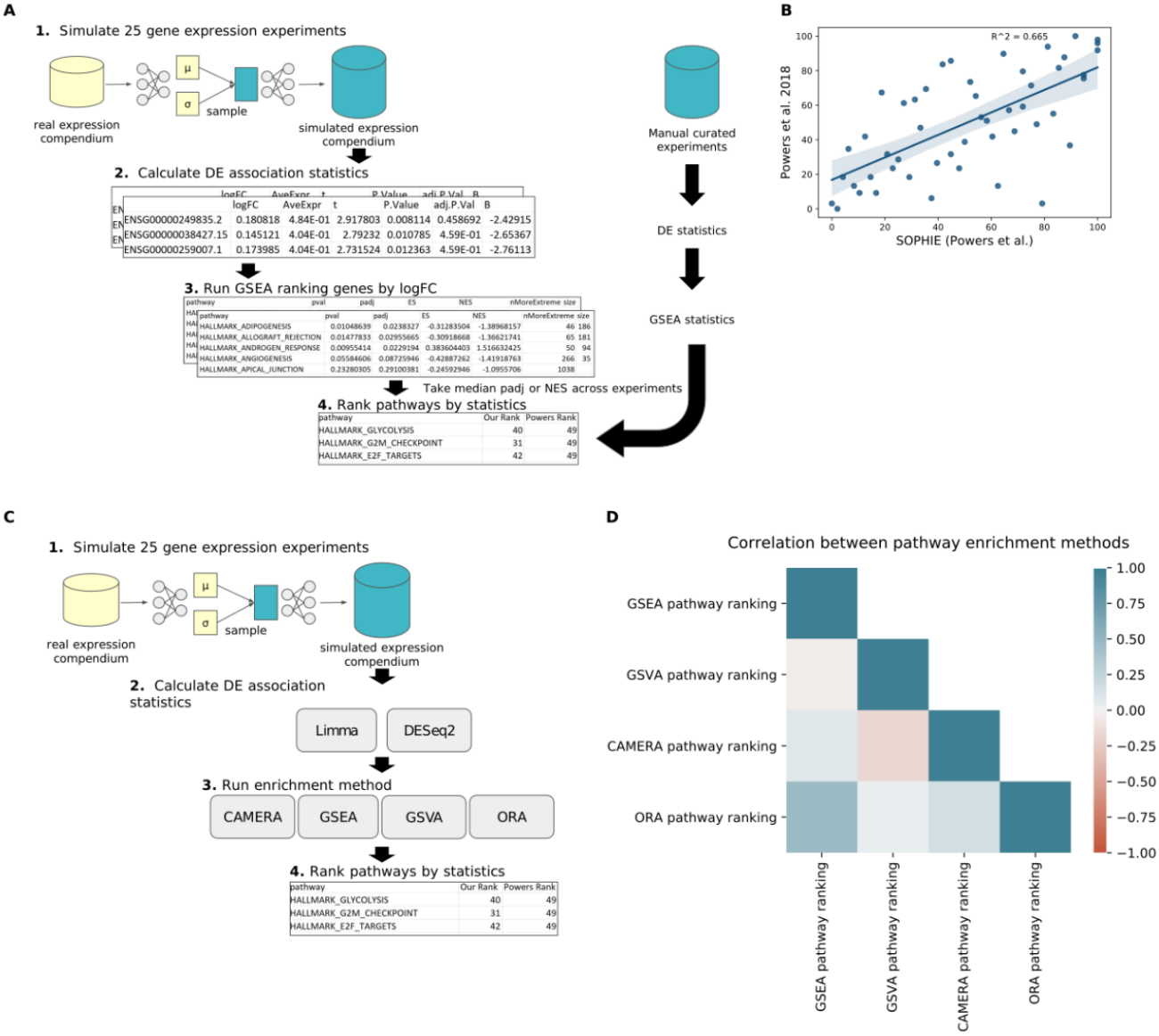


Figure 3: SOPHIE identifies the same commonly changed pathways previously found using manual curation. A) Workflow describing how pathways were ranked by how commonly shifted they are using SOPHIE versus using a manually curated set of experiments. B) Correlation between pathway percentiles using our simulated method trained on Powers et al. compendium (x-axis) versus percentiles obtained from Powers et al. (y-axis). C) Workflow describing how the pipeline can be easily extended to plug in different enrichment methods. D) Correlation of

pathway percentiles between different enrichment methods (GSEA, GSVA, CAMERA, ORA) using RNA-seq data.

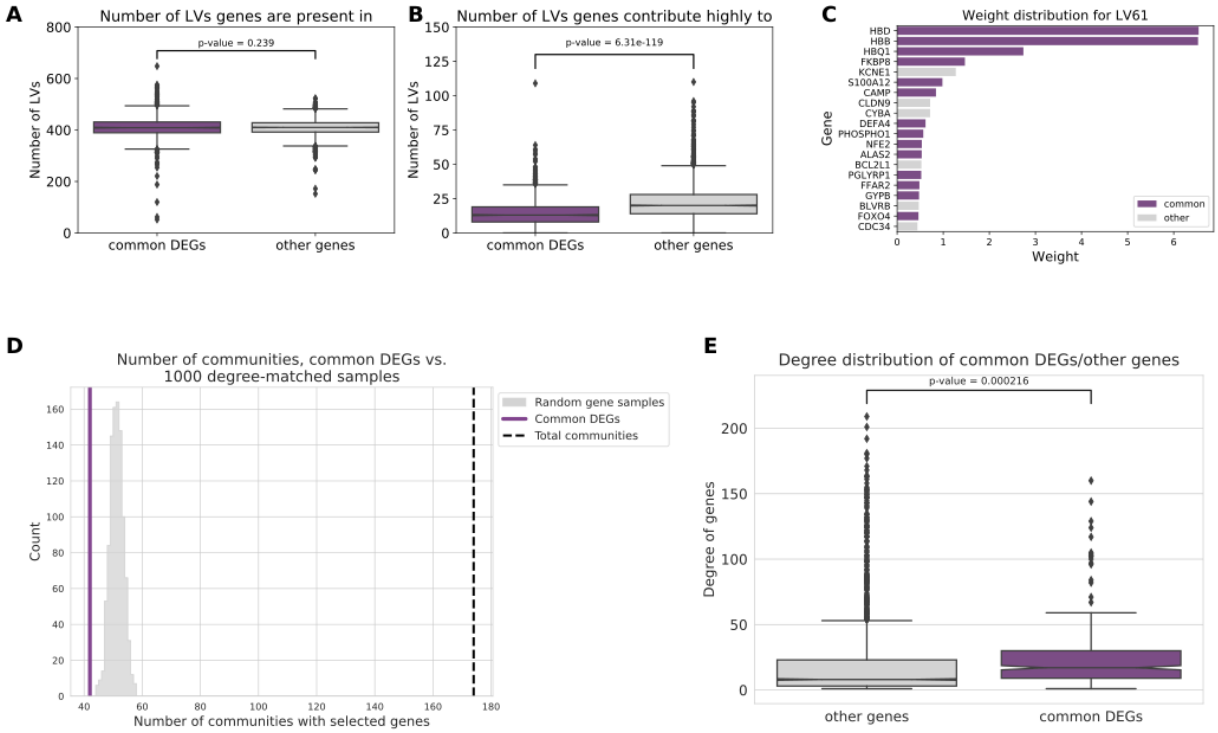


Figure 4: Common DEGs may contribute to a few hyperresponsive pathways. A) Number of human PLIER latent variables (LVs) commonly changed genes and other genes are present in (t-test p-value=0.239). B) Number of human PLIER latent variables commonly changed genes and other genes have a high weight score in (t-test p-value=6.31e-119). C) Distribution of top-weighted human genes in example LV61, which was found to contain a high proportion of high weight commonly changed genes. D) The number of communities with at least one commonly changed *P. aeruginosa* gene (purple) compared to the distribution of the number of communities with at least one non-commonly changed gene across 1000 samplings (grey) with the total number of communities marked by the black dashed line. E) Distribution of the degree of commonly changed *P. aeruginosa* genes (purple) compared to other genes (grey).

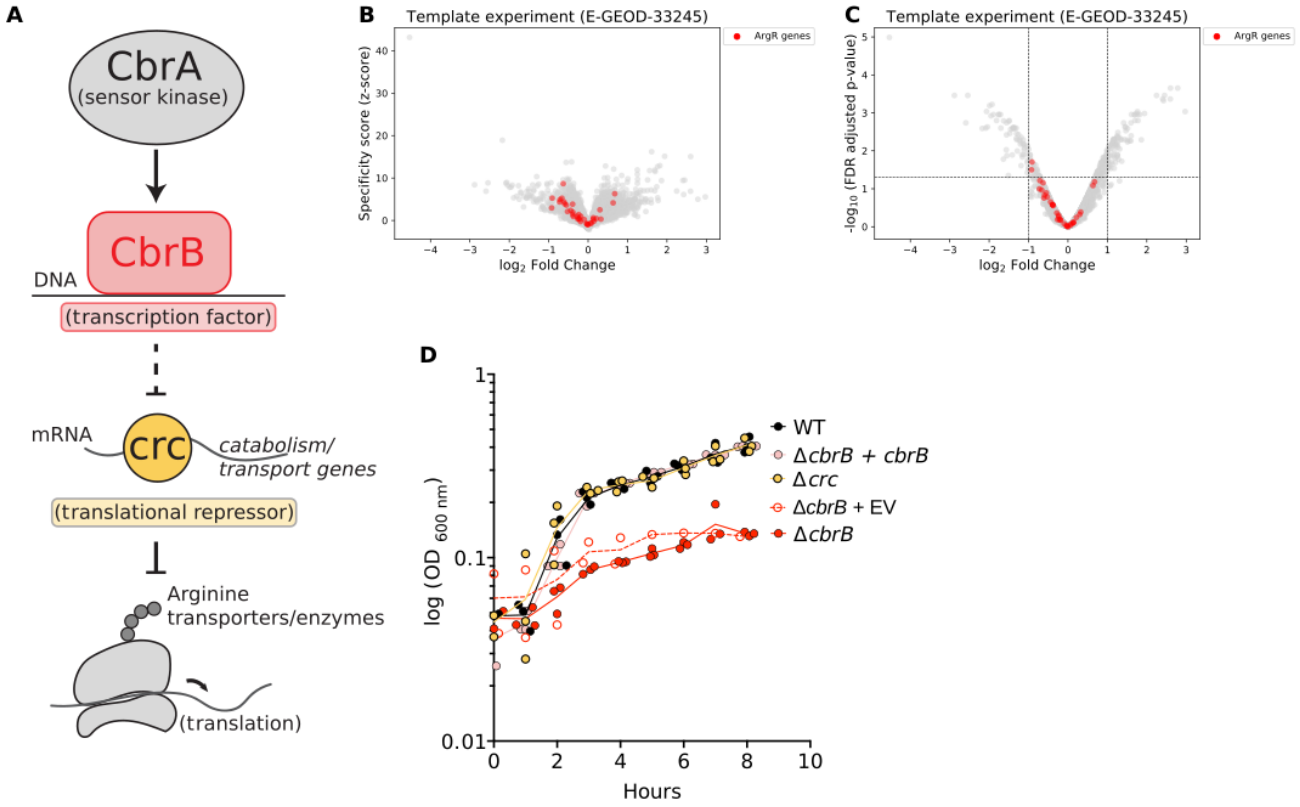
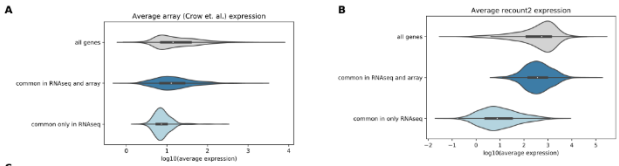


Figure 5: SOPHIE can identify genes with specific expression shifts in experiments. A) Model of CbrAB system. Volcano plot with $\log_2$ fold change versus B) z-score or C) adjusted p-values where genes regulated by ArgR are highlighted in red. D) Growth curves for *P. aeruginosa* in 10 mM arginine using WT (black), *cbrB* mutant (filled red), *cbrB* mutant with an empty expression vector (empty red), *cbrB* mutant with extrachromosomal complementation (pink), and *crc* mutant (yellow).



C

Average expression of RNA-seq only common DEGs

log10(average expression)

D

Average expression of RNA-seq/array common DEGs

Figure S1: Commonly changed genes found in RNA-seq but not array data indicate platform-specific shifts. A) Average gene expression for all genes in Crow et al. array dataset (grey), genes commonly found to be changed in both RNA-seq using SOPHIE and array dataset using Crow et al. (dark blue), genes commonly found to be differentially expressed only in RNA-seq dataset (light blue). B) Average gene expression for all genes in recount2 RNA-seq dataset (grey), genes commonly found to be differentially expressed in both RNA-seq using SOPHIE and array dataset using Crow et al. (dark blue), genes commonly found to be differentially expressed only in RNA-seq dataset (light blue). C) Average gene expression of genes commonly found to be differentially expressed only in RNA-seq dataset in template experiment (grey) compared to simulated experiment (light blue). D) Average gene expression of genes commonly found to be shifted in both RNA-seq and array datasets in template experiment (grey) compared to simulated experiment (dark blue).

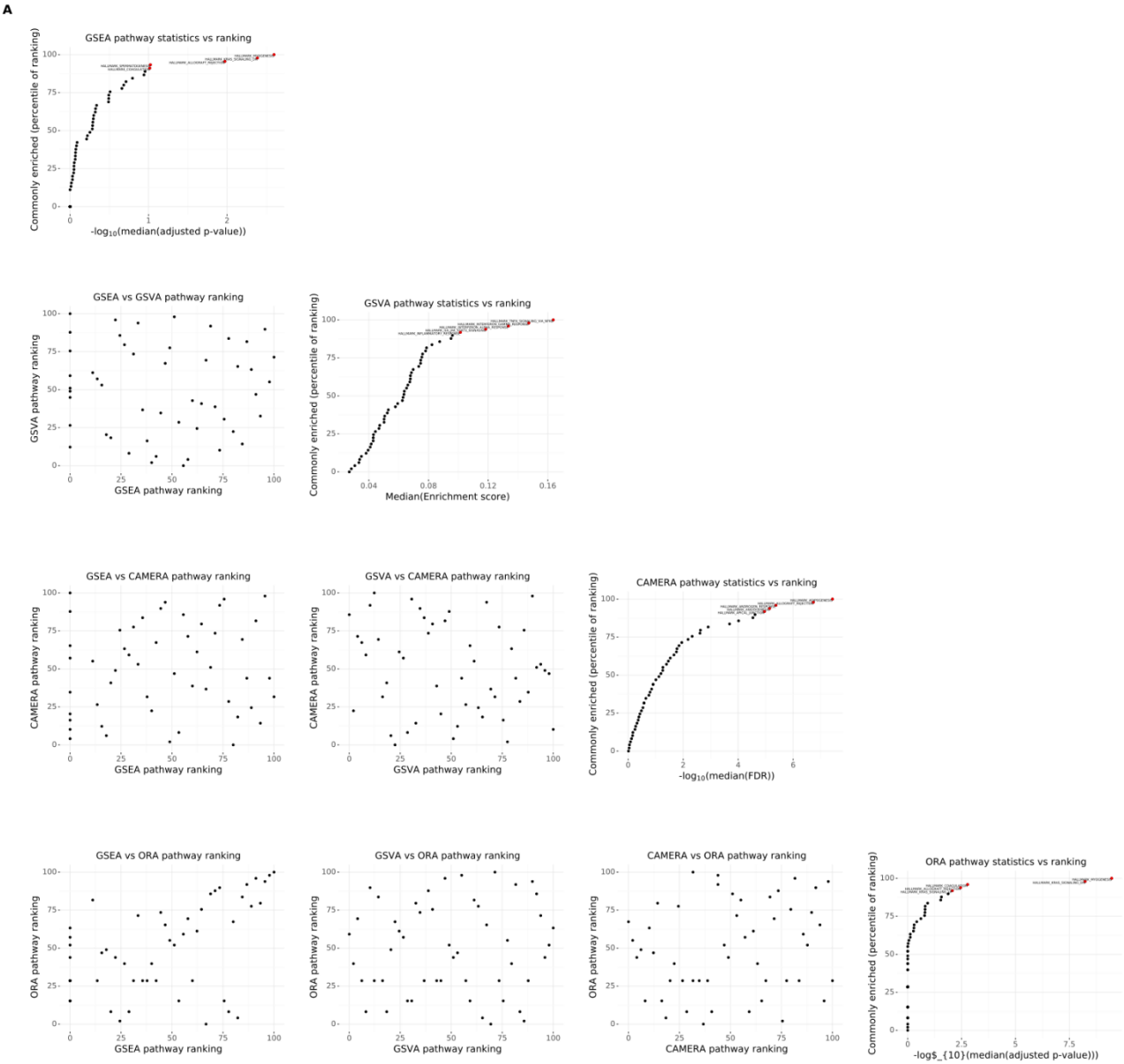


Figure S2: Different pathway enrichment methods will find different commonly enriched pathways. Scatterplot showing the correlation of pathway percentiles between different enrichment methods (GSEA, GSVA, CAMERA, ORA) using RNA-seq data.

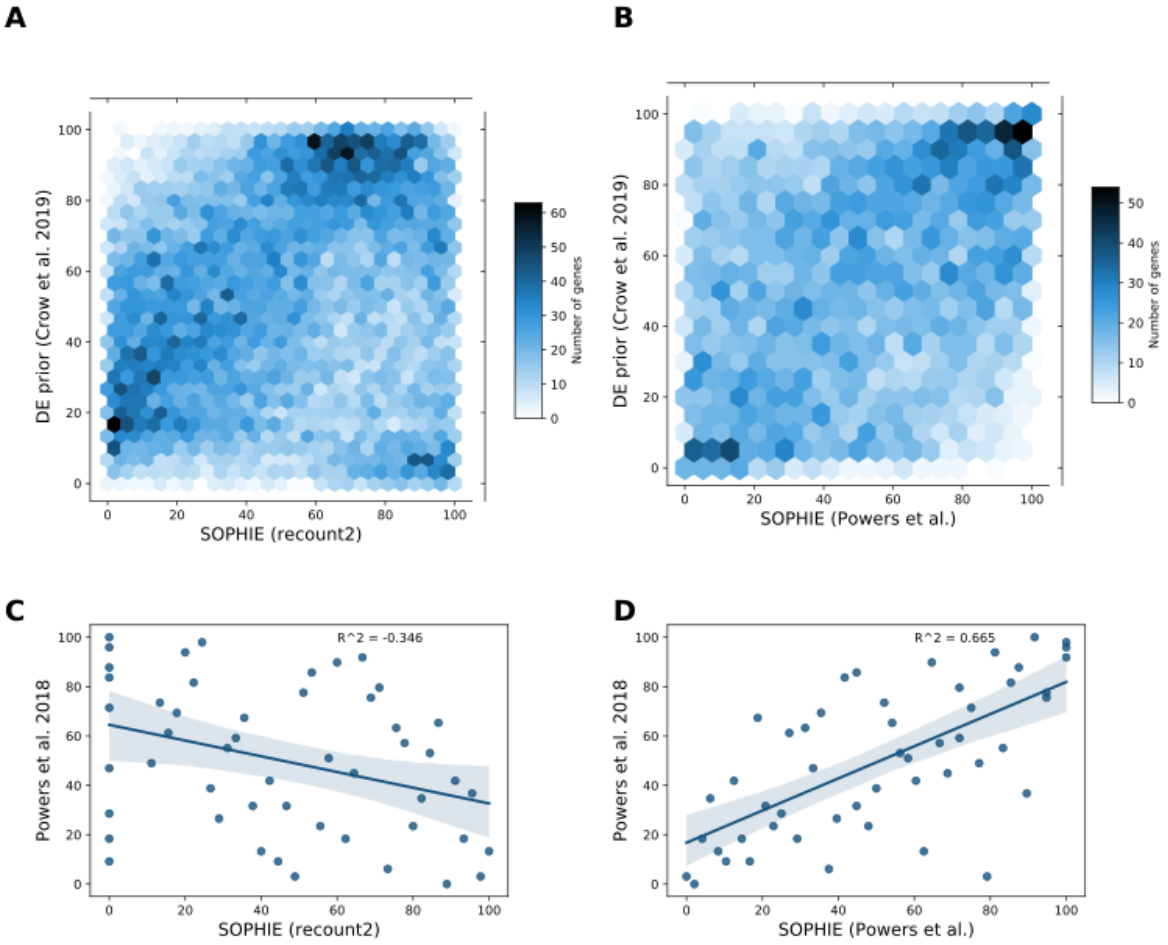


Figure S3: Commonly changed pathways are sensitive to experimental context. A) Correlation between gene percentiles using manually curated experiments from Crow et al. (y-axis) versus using our simulated method (x-axis) trained on recount2 (left) or Powers et al. compendium (right). B) Correlation between pathway percentiles using manually curated experiments obtained from Powers et al. (y-axis) versus our simulated method (x-axis) trained on recount2 (left) or Powers et al. compendium (right).

Gene Name	Gene Number	log (Δcrc/WT)	adjusted p-value	Z score	log(ΔcbrB/WT)	adjusted p-value	Z score
argA	PA5204	-0.181731828	0.908351211	1.237275138	0.756581996	0.077254116	8.449776846
aotJ	PA0888	-0.238474744	0.84970096	0.728557016	-0.71352001	0.060655286	4.919430215
aotP	PA0892	0.264663612	0.811504229	0.86782479	-0.595324546	0.147157567	3.917027549
aotQ	PA0889	0.10590825	0.948478128	-1.443774538	-0.442326784	0.222073814	2.37561046
aotM	PA0890	0.247726508	0.80833403	-1.078908312	-0.523785083	0.164420943	2.095615369

Table 1: Differential association statistics for genes regulated by the transcription factor ArgR that were found to be specific by SOPHIE.

Table S1: Human pathways associated with latent variables that contain a high (> 50%) proportion of high-weight commonly changed genes.

References:

1. Crow M, Lim N, Ballouz S, Pavlidis P, Gillis J. Predictability of human differential gene expression. *Proc Natl Acad Sci*. 2019;116(13):6491 LP - 6500. doi:10.1073/pnas.1802973116

2. Powers RK, Goodspeed A, Pielke-Lombardo H, Tan A-C, Costello JC. GSEA-InContext: identifying novel and common patterns in expression experiments. *Bioinformatics*. 2018;34(13):i555-i564. doi:10.1093/bioinformatics/bty271

3. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B*. 1995;57(1):289-300.

4. Storey JD, Taylor JE, Siegmund D. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J R Stat Soc Ser B (Statistical Methodol*. 2004;66(1):187-205.

5. Reiner A, Yekutieli D, Benjamini Y. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*. 2003;19(3):368-375.

6. Schurch NJ, Schofield P, Gierliński M, et al. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *Rna*. 2016;22(6):839-851.

7. Gasch AP, Spellman PT, Kao CM, et al. Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell*. 2000;11(12):4241-4257. doi:10.1091/mbc.11.12.4241

8. Collado-Torres L, Nellore A, Kammers K, et al. Reproducible RNA-seq analysis using recount2. *Nat Biotechnol*. 2017;35(4):319-321. doi:10.1038/nbt.3838

9. Lachmann A, Torre D, Keenan AB, et al. Massive mining of publicly available RNA-seq data from human and mouse. *Nat Commun*. 2018;9(1):1366. doi:10.1038/s41467-018-03751-6

10. Sonnleitner E, Valentini M, Wenner N, Haichar F el Z, Haas D, Lapouge K. Novel Targets of the CbrAB/Crc Carbon Catabolite Control System Revealed by Transcript Abundance in *Pseudomonas aeruginosa*. *PLoS One*. 2012;7(10):e44637. https://doi.org/10.1371/journal.pone.0044637.

11. Lee AJ, Park Y, Doing G, Hogan DA, Greene CS. Correcting for experiment-specific

variability in expression compendia can remove underlying signals. *Gigascience*. 2020;9(11). doi:10.1093/gigascience/giaa117

12. Renneville A, Van Galen P, Canver MC, et al. EHMT1 and EHMT2 inhibition induces fetal hemoglobin expression. *Blood*. 2015;126(16):1930-1939. doi:10.1182/blood-2015-06-649087

13. Akrami H, Joshi AA, Aydore S, Leahy RM. Addressing Variance Shrinkage in Variational Autoencoders using Quantile Regression. *arXiv Prepr arXiv201009042*. 2020.

14. Lin C-Y, Vega VB, Thomsen JS, et al. Whole-genome cartography of estrogen receptor alpha binding sites. *PLoS Genet*. 2007;3(6):e87. doi:10.1371/journal.pgen.0030087

15. Kim SC, Jung Y, Park J, et al. A high-dimensional, deep-sequencing study of lung adenocarcinoma in female never-smokers. *PLoS One*. 2013;8(2):e55596. doi:10.1371/journal.pone.0055596

16. Li Z, Koeppen K, Holden VI, et al. GAUGE-Annotated microbial transcriptomic data facilitate parallel mining and high-throughput reanalysis to form data-driven hypotheses. David LA, ed. *mSystems*. 2021;6(2):e01305-20. doi:10.1128/mSystems.01305-20

17. Korotkevich G, Sukhov V, Sergushichev A. Fast gene set enrichment analysis. *bioRxiv*. January 2019:60012. doi:10.1101/060012

18. Geistlinger L, Csaba G, Santarelli M, et al. Toward a gold standard for benchmarking gene set enrichment analysis. *Brief Bioinform*. 2021;22(1):545-556. doi:10.1093/bib/bbz158

19. Taroni JN, Grayson PC, Hu Q, et al. MultiPLIER: a transfer learning framework for transcriptomics reveals systemic features of rare disease. *Cell Syst*. 2019;8(5):380-394.

20. Chen KM, Tan J, Way GP, Doing G, Hogan DA, Greene CS. PathCORE-T: identifying and visualizing globally co-occurring pathways in large transcriptomic compendia. *BioData Min*. 2018;11:14. doi:10.1186/s13040-018-0175-7

21. Tan J, Hammond JH, Hogan DA, Greene CS. ADAGE-Based integration of publicly available *Pseudomonas aeruginosa* gene expression data with denoising autoencoders illuminates microbe-host interactions. *mSystems*. 2016;1(1). doi:10.1128/mSystems.00025-15

22. Huang H, Shao X, Xie Y, et al. An integrated genomic regulatory network of virulence-related transcriptional factors in Pseudomonas aeruginosa. *Nat Commun*. 2019;10(1):2931. doi:10.1038/s41467-019-10778-w

23. Yeung ATY, Janot L, Pena OM, et al. Requirement of the *Pseudomonas aeruginosa* CbrA sensor kinase for full virulence in a murine acute lung infection model. *Infect Immun*. 2014;82(3):1256-1267. doi:10.1128/IAI.01527-13

24. Sonnleitner E, Abdou L, Haas D. Small RNA as global regulator of carbon catabolite repression in *Pseudomonas aeruginosa*; *Proc Natl Acad Sci*. December 2009:pnas.0910308106. doi:10.1073/pnas.0910308106

25. Nishijyo T, Haas D, Itoh Y. The CbrA-CbrB two-component regulatory system controls the utilization of multiple carbon and nitrogen sources in Pseudomonas aeruginosa. *Mol Microbiol*. 2001;40(4):917-931. doi:10.1046/j.1365-2958.2001.02435.x

26. Li W, Lu C-D. Regulation of Carbon and Nitrogen Utilization by CbrAB and NtrBC Two-Component Systems in Pseudomonas aeruginosa *J Bacteriol*. 2007;189(15):5413 LP - 5420. doi:10.1128/JB.00432-07

27. Szalai B, Saez-Rodriguez J. Why do pathway methods work better than they should? *bioRxiv*. January 2020:2020.07.30.228296. doi:10.1101/2020.07.30.228296

28. Grützmann R, Boriss H, Ammerpohl O, et al. Meta-analysis of microarray data on pancreatic cancer defines a set of commonly dysregulated genes. *Oncogene*. 2005;24(32):5079-5088. doi:10.1038/sj.onc.1208696

29. Zhang JD, Berntenis N, Roth A, Ebeling M. Data mining reveals a network of early-response genes as a consensus signature of drug-induced *in vitro* and *in vivo* toxicity. *Pharmacogenomics J*. 2014;14(3):208-216. doi:10.1038/tpj.2013.39

30. Swindell WR, Sarkar MK, Liang Y, Xing X, Gudjonsson JE. Cross-Disease Transcriptomics: Unique IL-17A Signaling in Psoriasis Lesions and an Autoimmune PBMC Signature. *J Invest Dermatol*. 2016;136(9):1820-1830. doi:https://doi.org/10.1016/j.jid.2016.04.035

31. Nellore A, Collado-Torres L, Jaffe AE, et al. Rail-RNA: scalable analysis of RNA-seq splicing and coverage. *Bioinformatics*. 2017;33(24):4033-4040. doi:10.1093/bioinformatics/btw575

32. Frankish A, Diekhans M, Ferreira A-M, et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res*. 2019;47(D1):D766-D773. doi:10.1093/nar/gky955

33. Way GP, Greene CS. Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. *Pac Symp Biocomput*. 2018;23:80-91.

34. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12):550. doi:10.1186/s13059-014-0550-8

35. Smyth Gordon K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*. 2004;3(1):1-25.

36. Sergushichev AA. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. *bioRxiv*. January 2016:60012. doi:10.1101/060012

37. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics*. 2013;14:7. doi:10.1186/1471-2105-14-7

38. Wu D, Smyth GK. Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Res*. 2012;40(17):e133-e133. doi:10.1093/nar/gks461

39. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;43(7):e47-e47.

40. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omi a J Integr Biol*. 2012;16(5):284-287.

41. Taroni JN, Grayson PC, Hu Q, et al. MultiPLIER: a transfer learning framework reveals systemic features of rare autoimmune disease. *bioRxiv*. 2018:395947.

42. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. *J Stat Mech theory Exp*. 2008;2008(10):P10008.

43. Csardi G, Nepusz T. The igraph software package for complex network research. *InterJournal, complex Syst*. 2006;1695(5):1-9.

44. Zhang L, Peixoto TP. Statistical inference of assortative community structures. *Phys Rev Res*. 2020;2(4):43271.

45. P. Peixoto T. The graph-tool python library. figshare. Dataset. 2014. https://doi.org/10.6084/m9.figshare.1164194.v14.

46. Freeman LC. A set of measures of centrality based on betweenness. *Sociometry*. 1977:35-41.

47. Page L, Brin S, Motwani R, Winograd T. *The PageRank Citation Ranking: Bringing Order to the Web.* Stanford InfoLab; 1999.

48. Shanks RMQ, Caiazza NC, Hinsa SM, Toutain CM, O'Toole GA. *Saccharomyces cerevisiae*-based molecular tool kit for manipulation of genes from gram-negative bacteria. *Appl Environ Microbiol*. 2006;72(7):5027-5036. doi:10.1128/AEM.00682-06