Computational and Structural Biotechnology Journal 20 (2022) 4315–4324

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Using genome-wide expression compendia to study microorganisms

Alexandra J. Lee [a], Taylor Reiter [b], Georgia Doing [c], Julia Oh [c], Deborah A. Hogan [d], Casey S. Greene [b,*]

[a] Genomics and Computational Biology Graduate Program, University of Pennsylvania, Philadelphia, PA, USA
[b] Department of Biochemistry and Molecular Genetics, University of Colorado School of Medicine, Denver, CO, USA
[c] The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA
[d] Department of Microbiology and Immunology, Geisel School of Medicine, Dartmouth, Hanover, NH, USA

ARTICLE INFO

Article history:
Received 10 March 2022
Received in revised form 7 August 2022
Accepted 7 August 2022
Available online 10 August 2022

Keywords:
Transcriptomics
Compendia
Machine learning
Microbiology

ABSTRACT

A gene expression compendium is a heterogeneous collection of gene expression experiments assembled from data collected for diverse purposes. The widely varied experimental conditions and genetic backgrounds across samples creates a tremendous opportunity for gaining a systems level understanding of the transcriptional responses that influence phenotypes. Variety in experimental design is particularly important for studying microbes, where the transcriptional responses integrate many signals and demonstrate plasticity across strains including response to what nutrients are available and what microbes are present. Advances in high-throughput measurement technology have made it feasible to construct compendia for many microbes. In this review we discuss how these compendia are constructed and analyzed to reveal transcriptional patterns.

Contents

1. Introduction

Genome-wide transcriptional profiling measures the expression of all genes within a given sample [1,2]. This profile captures a snapshot of an organism's cellular state – what genes are active and how much they change in response to an environmental condition [3] or stimulus [4]. Consequently, transcriptional patterns can reveal the biological processes and possible mechanisms that contribute to traits including virulence [5–8], antibiotic resistance [9–11], metabolic versatility [12–14] and adaption [15]. These traits are of interest because they pertain to anthropocentric processes like microbial infection and bioreactor design [16,17],

* Corresponding author.
E-mail address: Casey.S.Greene@cuanschutz.edu (C.S. Greene).

https://doi.org/10.1016/j.csbj.2022.08.012

inform our understanding of biological mechanisms in multicellular eukaryotes [18], and underlie ecological cycles of biotransformations [19]. Therefore, transcriptomic studies are commonly used to examine trait-associated genes and their regulation.

Early experiments revealed the importance of transcriptional regulation in microbes. For example, experiments in the model organisms *Escherichia coli* (*E. coli*) and *Saccharomyces cerevisiae* (*S. cerevisiae*) revealed that common gene expression responses were elicited by different environmental stressors [20,21]. Studies in *Pseudomonas aeruginosa* (*P. aeruginosa*), an opportunistic gram-negative pathogen, found transcription factors, including the global regulator LasR, that control the expression of a number of extracellular factors that contribute to virulence [6,22,23] including proteases [5,24]. Overall, by studying the global transcriptome response, we can start to understand the mechanisms of traits of interest.

Microbial transcription in response to microbial interactions and environmental cues is complex. Genome organization affects transcription through diverse mechanisms, including factors like 3-dimensional organization [25], gene proximity [26], and promoter location [27]. Transcriptional regulators interact with internal and external cues to achieve transcription programs that reflect their environment. For example, in microbial quorum sensing (QS), a cell–cell communication process that allows microbes to respond to population density through signal molecules, microbes produce and respond to signals that facilitate adaptation to varying conditions [28,29]. For example, farnesol, a QS signaling molecule produced by *Candida albicans* (*C. albicans*), inhibits biofilm formation [30]. Similarly, the accessory gene regulator locus in *Staphylococcus aureus (S. aureus)*, which encodes the QS system, mediates the development of biofilm in response to nutrient availability [31]. QS regulators can also impact environmental responses by regulating other transcription factors such as the oxygen-sensitive Anr in *P. aeruginosa* [32]. Anr activity is higher in QS-defective strains that lack function of the LasR QS regulator. Thus, *lasR* mutants (LasR-), which are frequently isolated from CF patients, are more fit in microoxic conditions than their LasR + counterparts [33].

In addition to environmental cues, microbes tend to grow in polymicrobial communities where they sense and transcriptionally respond to other microbes. Both competitive and cooperative behaviors [34,35] influence phenotypes [36,37], eliciting interactions like the production of public goods, (cross-feeding) [38], resource consumption, interference competition [39], or coordinate production of phenotypes (increased virulence and antibiotic resistance [40]). For example, in co-infection of *S. aureus* and *P. aeruginosa*, *P. aeruginosa* exoproducts can select for *S. aureus* small colony variants that are aminoglycosides resistant [40]. Finally, these microbe-microbe interactions are also dependent on environmental factors. Doing *et al.* found that *P. aeruginosa* produced antifungal phenazines against *C. albicans*, but that this antagonistic interaction depends on phosphate availability and *C. albicans* fermentation [41,42]. Even two different genotypes can influence each other as in citrate cross-feeding found by Mould *et al.* [43].

Given the context-specific nature of transcription, leveraging data across many experiments allows researchers to study how microbes regulate transcription of different genes and pathways across different conditions – to gain a more systems level understanding of the transcriptome. Gene expression compendia, which are integrated collections of experiments, are one solution for examining transcriptional patterns across different contexts. In this review we describe how these compendia are constructed and the challenges faced as well as highlight analyses using compendia to reveal patterns of interest.

2. Construction of microbial expression compendia

For the purposes of this review, we defined an expression compendium to be a heterogeneous collection containing hundreds to thousands of samples that span multiple experiments assembled from data collected for diverse purposes. The range of experiments in work that we identified as meeting these criteria starts from 8 experiments and goes upwards of 100 experiments, which ensures that the compendium contained enough samples to apply system-level computational tools. Notable existing microbial compendia can be found in Table 1**.** The construction of each of these compendia began with the collection of relevant gene expression experiments from public repositories like ArrayExpress [44], Gene Expression Omnibus (GEO) [45], Sequence Read Archive (SRA) [46] and others [47,48]. Experiments of interest were then downloaded from these public repositories. In the case of the compendia represented in Table 1, all experiments (i.e. samples deposited together) within a given compendia were measured on the same platform to avoid bias and maintain a uniform reference. Additional filtering of samples were optionally performed to ensure that removal of spurious random correlation between genes [49]. Next, the samples were normalized to allow for cross sample comparison. Overall, filtering, consistency in platform and normalization ensure that the compendium data is uniformly processed.

There are different normalization techniques available depending on the technology. As an example, the *P. aeruginosa* RNA-seq compendium started with expression profiles downloaded from SRA and then median-ratio (MR) normalized [49,50]. The authors evaluated well-known RNA-seq normalizations, transcripts per million (TPM) and trimmed mean of means (TMM) [51], which corrected for spurious correlations; however correlations between random pairs of genes were still elevated compared to using MR normalization, which was their preferred strategy. These RNA-seq normalization methods address systematic variation, including differences in library size (i.e. sequencing depth) [52] and gene length [53], allowing for between sample and gene comparisons. Similarly, there also exist systematic variation in measurements using array technology though the sources are different and include differences in preparation protocol (i.e., total quantity of starting RNA, dye labeling) or differences in processing (i.e., different scanners or runs). One of the well-established normalization methods for the Affymetrix GeneChip system, which most of the compendia in Table 1 used, is RMA [54] which is a quantile method. In comparison to other single label normalization methods, Bolstad *et al.* [55] reported that RMA successfully reduced bias at reasonable compute speed compared to other global normalization methods. A similar review of two-color array technology, performed by Yang *et al.* [56], showed that different global or location-based normalization methods should be performed depending on the set of control spots. In a couple cases, where the compendium integrated across different platforms, such as two different array technologies or combining array and RNA-seq, studies used quantile normalization [57–59]. Regardless of the technology used, expression levels between samples can vary due to technical reasons, mentioned above, and so it's important to use normalization methods to adjust for these differences in order to compare between two gene expression profiles for applications such as gene function prediction, transcription regulatory network (TRN) inference and feature extraction.

Most of the existing compendia in Table 1 did not apply batch correction. In one case, where the compendium combined array and RNA-seq data, ComBat [58] was applied. While normalization is necessary in the context of compendia and facilitates cross-sample comparisons, batch correction is an optional step, and its application depends on the experiments included in the compendia. See section 'Challenges integrating across experiments' for a discussion of batch correction.

As more transcriptome data are generated, repositories like refine.bio [59], COLOMBOS [60,61] PILGRM [62], and M^{3D} [63] are being developed to provide easily downloadable compendia where

Table 1
Examples of existing microbial compendia.

Compendium	Organism	Description	No. experiments	No. Samples	No. genes	Platform
P. aeruginosa compendium [80]	*P. aeruginosa*	Compendium containing *P. aeruginosa* array data downloaded from ArrayExpress archived in 2014. It includes a mixture of different strain types, media, experimental stimuli.	109	950	5,549	Affymetrix platform GPL84
P. aeruginosa RNA-seq compendium [49]	*P. aeruginosa*	Compendium containing *P. aeruginosa* RNA-seq data downloaded from GEO and SRA in 2021. It includes a mixture of different strain types, media, experimental stimuli	> 100	2,333	5,563 (PAO1) 5,887 (PA14)	RNA-seq
EcoMAC [57]	*E. coli*	Compendium containing *E. coli* array data downloaded from GEO, ASAP database, ArrayExpress. It includes different strains, media and tests different environmental and genetic perturbations.	127	2,198	4,189	Affymetrix E. Coli Genome 2.0 Array GPL 3154; Affymetrix Ecoli Antisense Array GPL 199
EcoGEC [58]	*E. coli*	Compendium containing *E. coli* gene array data from EcoMAC plus RNA-seq data downloaded from GEO. It includes different strains, media and tests different environmental and genetic perturbations.	144	2,262	4,166	Affymetrix E. Coli Genome 2.0 Array; Affymetrix Ecoli Antisense Array; RNA-seq
Unnamed [139]	*E. coli*	Compendium containing E. coli gene array data downloaded from GEO, ArrayExpress and Stanford Microarray Database. It includes a mixture of different experimental conditions	74	870	NA	Affymetrix; P33; spotted cDNA/DNA; spotted oligonucleotides
Unnamed [68]	*E. coli*	Compendium containing E. coli RNA-seq data downloaded from GEO. It includes a mixture of different experimental conditions	21	278	3,923	RNA-seq
Unnamed [69]	*S. aureus*	Compendium containing S. auerus RNA-seq data downloaded form GEO combined with RNA-seq data generated by this publication. It includes expression profiles exposed to various media conditions, antibiotics, nutrient sources, and other stressors	8	109	2,581	RNA-seq
Unnamed [140]	*S. cerevisiae*	Compendium containing *S. cerevisiae* array data was a combination of perturbation experiments downloaded from PUMAdb and experiments generated by a genetic screen comparing mutant or compound-treated culture vs wild-type or mock-treated culture. Growth conditions for the screen were consistent across experiment.	>151	1,909	>2000	two-color cDNA microarray hybridization assay
Refine.bio [59]	Many prokaryotes	Database containing processed compendia for multiple prokaryotes including P. aerguinosa, E. Coli and sacch. The data for these compendia were downloaded from SRA, GEO and ArrayExpress.	~40 to > 500	~300 to ~ 13,000	~5000	microarray; RNA-seq

*Note in SRA, samples are referred to as "Experiment" and a group of samples forming an experiment are referred to as a "Study".

the data has been uniformly processed for different bacterial species. In general, the abundance of data has facilitated the generation of compendia to study transcriptional patterns across experiments.

3. Why use compendia: Benefits and applications of using compendia

3.1. Systems-level models

The construction of compendia, which contain hundreds to thousands of samples, has opened the door to the development of computational approaches, especially machine learning methods that have been successful at prediction tasks [64] and pattern extraction [65] in computer science, to discover transcriptional patterns in microbes.

Compendia can contribute to helping us gain a systems-level understanding of microbial biology. One major goal for systems biology is to model how information is encoded, specifically to reverse engineer the hierarchy of the transcriptomic regulatory network (TRN) [66–71]. Knowing the organization of a regulatory network allows us to control or optimize parts of the system, a necessary step for many biotechnological advances [72–75]. This task requires a large amount of heterogeneous data, which compendia provide, to identify shared patterns looking across a variety of interventions [76].

Dimensionality reduction methods can also be deployed to extract key patterns in data and reveal the transcriptional relationships between sets of genes [77]. Applying dimensionality reduction models to compendia allows users to study changes in gene sets and reveal more subtle and possibly undiscovered signals that could be masked by strong signals (i.e. a large fraction of genes representing the same pathway) [78,79]. For example, a denoising autoencoder trained on a *P. aeruginosa* compendium, ADAGE, captured regulation patterns and biological processes [80]. Tan *et al.* showed that co-operonic genes were weighted highly in the same latent variables and, similarly, KEGG gene sets were enriched in some latent variables. They also showed that function prediction using the ADAGE weight matrix was more accurate compared to using a randomly permuted gene weight matrix. Furthermore, the latent representation of the gene expression data detected existing subtle expression differences [80] and also revealed a new aspect of low phosphate response that depends on the media [81]. These latent variables were also shown to detect pathway-pathway relationships – i.e. pathways that co-occur in the same latent variable [82]. A similar dimensionality reduction analysis was performed applying a sparse autoencoder to a yeast (*S. cerevisiae*) compendium, where Chen *et al.* found latent variables represented pathways and other layers of biological abstractions such as a transcription factor complexes and signaling pathways [83]. In other studies, applying independent component analysis (ICA) to a compendium of transcriptome data revealed transcription modules [68–70]. Specifically, Poudel *et al.* identified differentially active modules in *S. aureus* that varied based on the growth different media, which revealed metabolic regulators that respond to shifts in nutrients available [69]. These unsupervised approaches summarize patterns in the expression compendia that can abstract different layers of a biological system that are useful for understanding the interaction between different molecular processes as well as generating new hypothesis. Webtools were developed to facilitate the exploration of the summarized data, like ADAGE [84], as well as to search through the experiments available in compendia such as the ones found in COLOMBOS [61,85,86], PILGRM [62] and others [87] in order to direct future research.

3.2. Methodologies to leverage compendia

With the breadth of transcriptional patterns captured by compendia, recent approaches have been developed that demonstrate how compendia can be used to put new experiments in the context of existing ones as well as to leverage the aggregation of patterns available to study genomic patterns. Lee *et al.* developed a general framework for distinguishing between common and experiment-specific differentially expressed genes, called SOPHIE (Specific cOntext Pattern Highlighting In Expression data) [88]. This approach compares gene expression changes in their target experiment with changes in a background set of experiments thereby allowing researchers to interpret and prioritize patterns in differentially expressed genes. The authors demonstrated that SOPHIE successfully prioritized genes with small differences in expression that were directly due to the perturbation being studied and not due to condition-specific secondary effects. In general, reanalysis and mining of the experiments within these compendia can be facilitated by tools like SOPHIE [88] or algorithms like GAUGE [89], which automate sample group detection for downstream statistical analyses. Overall, approaches like SOPHIE can find patterns that generalize across compendia.

3.3. Condition-specific responses

Transcriptional profiling is a snapshot of an organism's state, which includes numerous diverse processes. Understanding the information that is captured in these profiles is important to understand how microbes that sense and respond to their environment. For example, Kim *et al.* inferred *E. coli* cellular and environmental state, like growth phase or aerobic conditions, from a gene expression compendium and identified pathways that are associated with the genes that are most predictive of these cellular states [58]. In other examples, studies also used gene expression to annotate the functional roles of genes [90–94]. For example, Troyanskaya *et al.*, introduced a method called MAGIC, which predicts if two proteins are functionally related using multiple data types including gene expression data. They demonstrated that MAGIC function predictions were consistent with GO terms using *S. cerevisiae* expression data [92]. Overall, by using these compendia to make predictions we can learn what genes are involved in different environmental conditions or processes, which can improve our understanding of microbial condition-specific responses.

Importantly, the identification of conditional regulons requires the study of a response of interest across multiple conditions. The diversity of condition-specific responses has been elucidated in targeted studies that have examined expression profiles in response to multiple stimuli such as various stressors [95]. However, the comprehensive mapping of condition-specific responses is often beyond the scope of an individual experiment. The reanalysis and *meta*-analysis of publicly available data revealed subsets due to the natural differences in how separate groups studied related phenomena in a way that informed each other [81,93,96]. For example, through compendium-wide analysis of the low phosphate response, Tan *et al*. identified a condition-specific element of the low phosphate signaling cascade [81]. In another example, Huttenhower *et al.*, developed an approach that provided condition-specific context for gene function predictions. They suggest a novel connection between *S. cerevisiae* sporulation response and the introduction of xylose metabolism genes [93]. These results would not have stood out from any individual experiment but was clear when the larger compendium was analyzed.

3.4. Inspiration from non-microbial expression compendia

Non-microbial gene expression compendia have also been generated and used for a variety of purposes, many of which may inspire future endeavors for microbial compendia [97–103]. A human-based gene, ortholog, or k-mer based tool could facilitate rapid searches of the microbial compendia to identify samples

from different experiments with similar expression profiles. Transfer learning has also successfully transferred knowledge contained in publicly available data sets and databases to rare disease samples [99,102]. Such methods could be applied to better unravel pathway-level patterns for rare microbial species. Lastly, human compendia have been leveraged to identify alternative splicing [100], lessons which may be applied to the discovery of polycistronic transcripts directly from RNA-seq reads. Further research is needed to explore how lessons learned from human transcriptome compendia can best apply to microbial transcriptomics.

These studies demonstrate that the versatile data that is available in compendia provides a valuable resource to gain a systems level understanding of transcriptional signaling as well as to make predictions. Additionally a low dimensional representation of compendia capture transcriptional patterns that can reveal coordinated activity of gene sets and pathways as well as allows researchers to generate new hypotheses [41,81]. Finally new methods are being developed to further leverage the benefits of compendia to improve different types of analyses.

4. Challenges integrating across experiments

While compendia are rich community resources that can be leveraged to gain new insights into transcription, two challenges make integration across experiments a difficult endeavor: batch effects and strain variation. Batch effects introduced by technical sources (lab that produced the data, sequencing depth) or biological sources (experimental conditions) can either obscure or highlight biological signals, while strain-level genome differences can lead to reduced detection of transcription due to incomplete read mapping.

4.1. Batch effects

In general, batch effects can disrupt detection of biological signal [104–107]. Consequently, it might be expected that compendia, which integrate many different types of experiments together, require batch correction. However, a recent study by Lee *et al.* [108] examined the effect of technical sources of variability in a compendium setting. They simulated gene expression compendia with varying amounts of technical variability and assessed the ability to detect the original underlying structure in the data after noise was added and then after batch correction was applied. In general, they found that for compendium with a few sources of technical variation batch correction can be effective, however with many more sources of technical variation batch correction isn't necessary and can even start to remove some of the desired biological signal. If correction is applied to a compendium where the experiment-specific noise is largely independent, more of the biological information is removed since biological signals are consistent while noise is experiment specific.

In the case where a compendium contains a few sources of technical variability, like different platforms [58], the dominant signal is the variability between platforms and applying batch correction methods should recover the underlying biological signal. In contrast, in the case where a compendium contains many sources of variability, like many different types of experiments each contributing independent sources of noise, then the aggregation of each experiment-specific source of variability washes out from the underlying biological signal that is consistent across experiments. In this scenario, applying batch correction methods will remove more of the biological signal.

For the cases where batch correction is effective, commonly established methods like Limma [109] and ComBat [110] allow scientists to set sources of variability as covariates [58]. Limma removes technical noise by first fitting a linear model, using *lmFit*, which describes the relationship between the input gene expression and the experimental design labels such as batch assignments and covariates. The resulting model is a coefficient matrix that contains weights for the contribution of the noise component contained in the total observed gene expression matrix. This estimated contribution can be subtracted out from the input expression data. Similarly, Combat also assumes that the input gene expression signal contains an additive batch effect component that can be removed by estimating the batch effect using empirical bayes and subtracting this out.

4.2. Strain variation

Microbial strain variation further hinders integration across experiments. Strain variation refers to genomic variation that occurs at the sub-species level and can take the form of single nucleotide variants (SNVs) or indels of different sizes, distinct complements of accessory genes, and genomic rearrangements [111]. While strain variation is a critical component of understanding a species' ultimate phenotypic variation, each form of variation causes distinct challenges for integrating expression data across strain types. For example, SNVs decrease the average nucleotide identity between the reference sequence used for read quantification and the sample, which can decrease mapping rates non-uniformly across samples [112]. Similarly, the reference sequence may not contain the same set of genes as is present in the sample. This is because many microbial species have a large number of accessory genes, genes which are not universal within that species. For example, accessory genes comprise $\sim$ 20 % of the genome for some staphylococci [113–115]. When the reference sequence does not contain the same genes as are present in a sample, this can lead to decreased mapping rates and unobserved gene expression [116]. Lastly, genomic rearrangements or insertions may cause difficulties for counting spanning reads that are present in a sample but not represented in a reference [117]. However, integrating strain variation is important not only to understand within-species phenotypic diversity, but also because accessory genes can modify function of the core genome [118].

Even given these challenges, different approaches have been developed to take advantage of publicly available microbial expression data sets in the face of strain variation. For example, *P. aeruginosa* has five major lineages detected upon genome analyses of over a thousand strains [119] with two major clades that many strains belong to including the widely studied strains PAO1 and PA14 [120]. Strains PAO1 and PA14 contain different sets of accessory genes. One common solution is to only consider core genes since they are shared across strain type [121–125]. In order to include accessory genes, separate compendia can be generated so that major strain types (PAO1 and PA14) are separated but there are PAO1-specific genes within the PAO1 compendium [49,126].

Most compendia are comprised of a single strain of microorganism (Table 1). This can be achieved by relying on the metadata associated with experiments available in the data repository or using information provided in publications to collect experiments from a single strain. However metadata are notoriously incompletely recorded [127] and difficult to harmonize across studies [128], which may lead to inappropriate inclusion or exclusion of samples in a compendia. Notably, less than half of the publicly available microbial RNA-seq data has been submitted to the GEO or Array Express or Expression Atlas. These three platforms provide detailed and standardized *meta*-data that can be accessed programmatically and easily used in high throughput computational analyses [129]. An alternative approach is to verify the strain annotation using taxonomy assignments provided in the SRA Run Browser analysis tab, or to perform assignment using with a tool like

sourmash gather, which selects the minimum set of reference genomes in a database necessary to cover the reads in a sample [130].

Alternatively, a pangenome could be used as a reference so that core genes are collapsed across strain types while accessory genes are included in the analysis [116]. Using these pangenomes as a reference balances computational cost and fidelity to sample genomes, and can take advantage of databases designed to address similar problems for metagenomic sample processing [131]. This approach was pioneered for the analysis of *S. aureus* strains directly from metatranscriptomes, as no reference genome was available with which to perform read quantification. This approach may be successful for building species-wide compendia but needs further research. Indeed, one substantial draw back would be the negation of spanning reads, as pangenomes are typically built from genes and not operons. The increasing use of metatranscriptomics to contextualize a species' function presents computational challenges but also opportunities to identify unique transcriptional signatures in their native and highly complex environment, such as *Haemophilus influenzae* during viral infection, or *S. aureus'* host defense response in the nares [132,133].

Overall, despite some challenges to constructing compendia, there are existing solutions that make compendia analysis possible and the benefits of the biological discoveries we can glean make it worth it. Additionally, it is worth noting that a recent analysis of a normalized and quality filtered compendium of over two thousand samples found strong gene expression correlations between co-regulated genes, even without batch correction, for genes present in the reference genome [49].

5. Future directions

As more transcriptomic data is generated, it is feasible to construct compendia for a wide array of organisms that include measurements of diverse conditions. Consequently, developing and systematizing strategies for how compendia can be analyzed to best learn from the broad array of conditions that they represent is critical. Our review integrates and synthesizes current strategies in this area. Looking ahead, hurdles still exist.

One methodological concern with using compendia, which integrate multiple experiments, is: when do we batch correct and what methods would be the most effective? Batch effects are the systemic sources of variance present in and between individual experiments. These sources of variance, which include noise generated by different experimental designs or different labs running the experiment, can confound the biological patterns that we are interested in detecting. Despite the presence of hundreds of technical sources of variability in compendia, previous compendia analyses have successfully extracted relevant biological patterns [70,80,81,83]. One hint for why these approaches work comes from a study by Lee *et al.* that found that a consistent biological signal was preserved in compendia when technical noise was independent between experiments. Applying batch correction in this setting was harmful – not helpful. Efforts to correct for technical artifacts in this setting would best focus on those that span multiple experiments: an area that remains relatively under-explored.

Current studies have done a lot of work to integrate bulk transcriptomics data in microbes using dimensionality reduction methods [68–70,81–84]. In the future, we can extend this work to new data types and computational models. One possible avenue to explore would be integrating data from multiple resolutions. Single-cell data provide an opportunity to examine variability between cells; however, technical limitations mean that the earliest single-cell data are likely to focus on the easiest to assay settings. It may require many years before the available single-cell data measure as many conditions as existing bulk assays. Bringing these resolutions together in compendia will require certain foundational work: for example, of the type performed by Doing *et al.* [49] to assess the mapping, quantification and normalization which remains to be done for microbes in the single-cell context.

Often these strategies are used in an interpretive manner, but emerging approaches may support genome-wide predictions of physiological state. One area of interest is latent space arithmetic, which is an approach simulating new samples using vector arithmetic to manipulate samples in an encoded space, to predict the response to perturbations [134,135]. This strategy might reveal the effect of a perturbation in a specific never-before tested setting. Perhaps it could inform the design cell-state-specific targets, providing increased specificity for anti-microbial agents by targeting microbes in more virulent states.

All compendia that we examined focused on one genome or transcriptome at a time. The future might include pan-genome approaches [116] to identify strain specific genes that induce novel transcriptional effects in different contexts. The methods and work that we discuss present the potential for a major conceptual shift in microbiology: instead of examining transcriptional profiles from an individual experiment in isolation, investigators studying compendia analyze patterns across experiments to reveal novel relationships that further our understanding of systems-level biology. Work to date has focused on a relatively narrow slice of what is possible, both from the point of view of data types and analytical approaches, and it has already been a fruitful strategy to identify and understand novel biological mechanisms.

6. Discussion

With advancements in high throughput sequencing technology more transcriptome data has become available, presenting opportunities for integration of diverse experiments into compendia. Recent successes of computational methods, especially unsupervised machine learning approaches, have demonstrated that biologically meaningful patterns can be extracted from microbial compendia. Given these recent advances, as well as tools developed in the analysis of human expression compendia, we anticipate development in the computational tool space will continue to drive biological discovery from microbial compendia.

While computational approaches for using heterogeneous compendia have been around for approximately 15 years [79], there remains work to be done to evaluate the computational methods that are most suitable for capturing the transcriptional patterns in compendia. Given the success to date of unsupervised learning methods [68–70,81,83,84,88], and the work that has been done in this space in human expression compendia [136,137], we anticipate that future development and evaluation of these methods will prove useful in the analysis of microbial expression compendia. A comprehensive analysis using human compendia showed that different models and model architectures captured different pathways, revealing that the use of multiple analysis methods led to more complete biological representations [136]. Similarly, there has been some assessment of microbial compendia examining pathway representation using different forms of dimensionality reduction methods [81,83] and expression changes captured using variational autoencoders [108]. However, an equivalent comprehensive evaluation as undertaken in human compendia is needed to assess the information captured in microbial compendia – what types of signals are captured when the model architecture, regularization, penalty functions, connectivity between layers is varied? This information will determine what model, or range of models, are appropriate for downstream analyses. As new feature extraction models continue to be developed to improve the information captured by and the interpretability of these models, such

as through the incorporation of prior information [138], such assessment becomes important to help guide researchers on the computational strategy they use.

Microbial gene expression compendia have proven to be a fruitful resource for studying systems-level changes and have been leveraged to infer TRNs [66], make predictions about phenotypes [58], and reveal coordinated gene sets [70,80,81,83]. Furthermore, compendia have been shown to improve the analysis of individual experiments [88] and to reveal specific genomic patterns [126]. The advancements in computational tools and webtools, which have made the information in some existing compendia easily accessible, is opening the door to new avenues of research, situating the study of transcription in a global context.

CRediT authorship contribution statement

Alexandra J. Lee: Conceptualization, Investigation, Project administration, Writing – original draft, Writing – review & editing. **Taylor Reiter:** Conceptualization, Investigation, Writing – original draft, Writing – review & editing. **Georgia Doing:** Writing – review & editing. **Julia Oh:** Writing – review & editing. **Deborah A. Hogan:** Funding acquisition, Writing – review & editing. **Casey S. Greene:** Conceptualization, Funding acquisition, Supervision, Writing – review & editing.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Casey S. Greene is a consultant for Arcadia Science, which aims to use non-traditional model organisms to make discoveries and develop new technologies.

Acknowledgements

This work was supported by grants from the Gordon and Betty Moore Foundation (GBMF4552 to CSG) and Cystic Fibrosis Foundation (HOGAN19GO to DAH and GREENE21GO to CSG).

References:

[1] Schulze A, Downward J. Navigating gene expression using microarrays — a technology review. Nat Cell Biol 2001;3(8):E190–5. https://doi.org/10.1038/35087138.
[2] Conway TKG. Microarray expression profiling: capturing a genome-wide portrait of the transcriptome. Mol Microbiol 2003;47(4):879–89. https://doi.org/10.1046/j.1365-2958.2003.03338.x.
[3] Lee LJ, Barrett JA, Poole RK. Genome-Wide Transcriptional Response of Chemostat-Cultured Escherichia coli to Zinc. J Bacteriol 2005;187(3):1124–34. https://doi.org/10.1128/JB.187.3.1124-1134.2005.
[4] Poole K, Krebes K, McNally C, Neshat S. Multiple antibiotic resistance in Pseudomonas aeruginosa: evidence for involvement of an efflux operon. J Bacteriol 1993;175(22):7363–72.
[5] Yanagihara K, Tomono K, Kaneko Y, et al. Role of elastase in a mouse model of chronic respiratory *Pseudomonas aeruginosa* infection that mimics diffuse panbronchiolitis. J Med Microbiol 2003;52(Pt 6):531–5. https://doi.org/10.1099/jmm.0.05154-0.
[6] Gambello MJ, Kaye S, Iglewski BH. LasR of *Pseudomonas aeruginosa* is a transcriptional activator of the alkaline protease gene (apr) and an enhancer of exotoxin A expression. Infect Immun 1993;61(4):1180–4.
[7] Wang R, Braughton KR, Kretschmer D, et al. Identification of novel cytolytic peptides as key virulence determinants for community-associated MRSA. Nat Med 2007;13(12):1510–4. https://doi.org/10.1038/nm1656.
[8] Jin L, Chen D, Liao S, et al. Transcriptome analysis reveals downregulation of virulence-associated genes expression in a low virulence Verticillium dahliae strain. Arch Microbiol 2019;201(7):927–41. https://doi.org/10.1007/s00203-019-01663-7.
[9] Whiteley M, Bangera MG, Bumgarner RE, et al. Gene expression in Pseudomonas aeruginosa biofilms. Nature 2001;413(6858):860–4. https://doi.org/10.1038/35101627.
[10] Prosser BL, Taylor D, Dix BA, Cleeland R. Method of evaluating effects of antibiotics on bacterial biofilm. Antimicrob Agents Chemother 1987;31(10):1502–6. https://doi.org/10.1128/AAC.31.10.1502.
[11] Alterations in kinetic properties of penicillin-binding proteins of penicillin-resistant Streptococcus pneumoniae. doi:10.1128/AAC.30.1.57.
[12] Sonnleitner E, Valentini M, Wenner N, el Z Haichar F, 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.
[13] Larimer FW, Chain P, Hauser L, et al. Complete genome sequence of the metabolically versatile photosynthetic bacterium Rhodopseudomonas palustris. Nat Biotechnol 2004;22(1):55–61. https://doi.org/10.1038/nbt923.
[14] Keller NP, Hohn TM. Metabolic Pathway Gene Clusters in Filamentous Fungi. Fungal Genet Biol 1997;21(1):17–29. https://doi.org/10.1006/fgbi.1997.0970.
[15] Sentausa E, Basso P, Berry A, et al. Insertion sequences drive the emergence of a highly adapted human pathogen. Microb. Genomics 2019;6(9). https://doi.org/10.1099/mgen.0.000265. mgen000265.
[16] Hong J, Li W, Lin B, Zhan M, Liu C, Chen BY. Deciphering the effect of salinity on the performance of submerged membrane bioreactor for aquaculture of bacterial community. Desalination 2013;316:23–30. https://doi.org/10.1016/j.desal.2013.01.015.
[17] Lefebvre O, Moletta R. Treatment of organic pollution in industrial saline wastewater: A literature review. Water Res 2006;40(20):3671–82. https://doi.org/10.1016/j.watres.2006.08.027.
[18] Ho YH, Gasch AP. Exploiting the yeast stress-activated signaling network to inform on stress biology and disease signaling. Curr Genet 2015;61(4):503–11. https://doi.org/10.1007/s00294-015-0491-0.
[19] Galachyants YP, Zakharova YR, Volokitina NA, Morozov AA, Likhoshway YV, Grachev MA. De novo transcriptome assembly and analysis of the freshwater araphid diatom Fragilaria radians, Lake Baikal. Sci Data 2019;6(1):183. https://doi.org/10.1038/s41597-019-0191-6.
[20] Metabolomic and transcriptomic stress response of Escherichia coli. *Mol Syst Biol*. 2010;6(1):364. doi:10.1038/msb.2010.18.
[21] 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):4241–57. https://doi.org/10.1091/mbc.11.12.4241.
[22] Lee J, Zhang L. The hierarchy quorum sensing network in *Pseudomonas aeruginosa*. Protein Cell 2015;6(1):26–41. https://doi.org/10.1007/s13238-014-0100-x.
[23] Miyazaki S, Matsumoto T, Tateda K, Ohno A, Yamaguchi K. Role of exotoxin A in inducing severe *Pseudomonas aeruginosa* infections in mice. J Med Microbiol 1995;43(3):169–75. https://doi.org/10.1099/00222615-43-3-169.
[24] Gambello MJ, Iglewski BH. Cloning and characterization of the *Pseudomonas aeruginosa* lasR gene, a transcriptional activator of elastase expression. J Bacteriol 1991;173(9):3000–9. https://doi.org/10.1128/jb.173.9.3000-3009.1991.
[25] El Houdaigui B, Forquet R, Hindré T, et al. Bacterial genome architecture shapes global transcriptional regulation by DNA supercoiling. Nucleic Acids Res 2019;47(11):5648–57. https://doi.org/10.1093/nar/gkz300.
[26] Junier I, Hérisson J, Képès F. Genomic Organization of Evolutionarily Correlated Genes in Bacteria: Limits and Strategies. J Mol Biol 2012;419(5):369–86. https://doi.org/10.1016/j.jmb.2012.03.009.
[27] Chowdhury WP, Satyshur KA, Keck JL, Kiley PJ. Minor Alterations in Core Promoter Element Positioning Reveal Functional Plasticity of a Bacterial Transcription Factor. *mBio*. Published online November 2, 2021. doi:10.1128/mBio.02753-21.
[28] Albuquerque P, Casadevall A. Quorum sensing in fungi – a review. Med Mycol 2012;50(4):337–45. https://doi.org/10.3109/13693786.2011.652201.
[29] Venturi V. Regulation of quorum sensing in Pseudomonas. FEMS Microbiol Rev 2006;30(2):274–91. https://doi.org/10.1111/j.1574-6976.2005.00012.x.
[30] Ramage G, Saville SP, Wickes BL, López-Ribot JL. Inhibition of Candida albicans Biofilm Formation by Farnesol, a Quorum-Sensing Molecule. Appl Environ Microbiol 2002;68(11):5459–63. https://doi.org/10.1128/AEM.68.11.5459-5463.2002.
[31] Boles BR, Horswill AR. agr-Mediated Dispersal of Staphylococcus aureus Biofilms. PLOS Pathog 2008;4(4):e1000052.
[32] Ye RW, Haas D, Ka JO, et al. Anaerobic activation of the entire denitrification pathway in Pseudomonas aeruginosa requires Anr, an analog of Fnr. J Bacteriol 1995;177(12):3606–9. https://doi.org/10.1128/jb.177.12.3606-3609.1995.
[33] Clay ME, Hammond JH, Zhong F, et al. Pseudomonas aeruginosa lasR mutant fitness in microoxia is supported by an Anr-regulated oxygen-binding hemerythrin. Proc Natl Acad Sci U S A 2020;117(6):3167–73. https://doi.org/10.1073/pnas.1917576117.
[34] Abisado RG, Benomar S, Klaus JR, Dandekar AA, Chandler JR. Bacterial Quorum Sensing and Microbial Community Interactions. Garsin DA, ed *mBio* 2018;9(3). https://doi.org/10.1128/mBio.02331-17.
[35] D'hoe K, Vet S, Faust K, et al. Integrated culturing, modeling and transcriptomics uncovers complex interactions and emergent behavior in a three-species synthetic gut community. Morgan X, Garrett WS, eds. *eLife*. 2018;7:e37090. doi:10.7554/eLife.37090.
[36] Mottola C, Mendes JJ, Cristino JM, Cavaco-Silva P, Tavares L, Oliveira M. Polymicrobial biofilms by diabetic foot clinical isolates. Folia Microbiol (Praha) 2016;61(1):35–43. https://doi.org/10.1007/s12223-015-0401-3.
[37] Moura A, Tacão M, Henriques I, Dias J, Ferreira P, Correia A. Characterization of bacterial diversity in two aerated lagoons of a wastewater treatment plant using PCR–DGGE analysis. Microbiol Res 2009;164(5):560–9. https://doi.org/10.1016/j.micres.2007.06.005.

[38] Cavaliere M, Feng S, Soyer OS, Jiménez JI. Cooperation in microbial communities and their biotechnological applications. Environ Microbiol 2017;19(8):2949–63. https://doi.org/10.1111/1462-2920.13767.
[39] Ghoul M, Mitri S. The Ecology and Evolution of Microbial Competition. Trends Microbiol 2016;24(10):833–45. https://doi.org/10.1016/j.tim.2016.06.011.
[40] Hoffman LR, Déziel E, D'Argenio DA, et al. Selection for Staphylococcus aureus small-colony variants due to growth in the presence of Pseudomonas aeruginosa. Proc Natl Acad Sci U S A 2006;103(52):19890–5. https://doi.org/10.1073/pnas.0606756104.
[41] Doing G, Koeppen K, Occipinti P, Harty CE, Hogan DA. Conditional antagonism in co-cultures of Pseudomonas aeruginosa and Candida albicans: An intersection of ethanol and phosphate signaling distilled from dual-seq transcriptomics. PLoS Genet 2020;16(8):e1008783.
[42] Gibson J, Sood A, Hogan DA. Pseudomonas aeruginosa-Candida albicans Interactions: Localization and Fungal Toxicity of a Phenazine Derivative. Appl Environ Microbiol 2009;75(2):504–13. https://doi.org/10.1128/AEM.01037-08.
[43] Mould DL, Botelho NJ, Hogan DA. Intraspecies Signaling between Common Variants of Pseudomonas aeruginosa Increases Production of Quorum-Sensing-Controlled Virulence Factors. *mBio*. Published online August 25, 2020. doi:10.1128/mBio.01865-20.
[44] Rustici G, Kolesnikov N, Brandizi M, et al. ArrayExpress update–trends in database growth and links to data analysis tools. *Nucleic Acids Res*. 2013;41 (Database issue):D987-990. doi:10.1093/nar/gks1174.
[45] Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 2002;30(1):207–10. https://doi.org/10.1093/nar/30.1.207.
[46] Kodama Y, Shumway M, Leinonen R. on behalf of the International Nucleotide Sequence Database Collaboration. The sequence read archive: explosive growth of sequencing data. Nucleic Acids Res 2012;40(D1):D54–6. https://doi.org/10.1093/nar/gkr854.
[47] Glasner JD, Liss P, Plunkett G, et al. ASAP, a systematic annotation package for community analysis of genomes. Nucleic Acids Res 2003;31(1):147–51.
[48] Leinonen R, Akhtar R, Birney E, et al. The European Nucleotide Archive. *Nucleic Acids Res*. 2011;39(Database issue):D28-31. doi:10.1093/nar/gkq967.
[49] Doing G, Lee AJ, Neff SL, et al. Computationally efficient assembly of a Pseudomonas aeruginosa gene expression compendium. Published online January 25, 2022:2022.01.24.477642. doi:10.1101/2022.01.24.477642.
[50] Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol 2010;11(10):R106. https://doi.org/10.1186/gb-2010-11-10-r106.
[51] Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol 2010;11(3):R25. https://doi.org/10.1186/gb-2010-11-3-r25.
[52] Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 2008;5 (7):621–8. https://doi.org/10.1038/nmeth.1226.
[53] Oshlack A, Wakefield MJ. Transcript length bias in RNA-seq data confounds systems biology. Biol Direct 2009;4(1):14. https://doi.org/10.1186/1745-6150-4-14.
[54] Irizarry RA, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostat Oxf Engl 2003;4(2):249–64. https://doi.org/10.1093/biostatistics/4.2.249.
[55] Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinforma Oxf Engl 2003;19(2):185–93. https://doi.org/10.1093/bioinformatics/19.2.185.
[56] Yang YH, Dudoit S, Luu P, et al. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res 2002;30(4):e15.
[57] Carrera J, Estrela R, Luo J, Rai N, Tsoukalas A, Tagkopoulos I. An integrative, multi-scale, genome-wide model reveals the phenotypic landscape of Escherichia coli. Mol Syst Biol 2014;10(7):735. https://doi.org/10.15252/msb.20145108.
[58] Kim M, Zorraquino V, Tagkopoulos I. Microbial Forensics: Predicting Phenotypic Characteristics and Environmental Conditions from Large-Scale Gene Expression Profiles. PLOS Comput Biol 2015;11(3):e1004127.
[59] Casey S. Greene, Dongbo Hu, Richard W. W. Jones, et al. refine.bio: a resource of uniformly processed publicly available gene expression datasets. https://www.refine.bio.
[60] Moretto M, Sonego P, Dierckxsens N, et al. COLOMBOS v3.0: leveraging gene expression compendia for cross-species analyses. Nucleic Acids Res 2016;44 (D1):D620–3. https://doi.org/10.1093/nar/gkv1251.
[61] Engelen K, Fu Q, Meysman P, et al. COLOMBOS: access port for cross-platform bacterial expression compendia. PLoS ONE 2011;6(7):e20938.
[62] Greene CS, Troyanskaya OG. PILGRM: an interactive data-driven discovery platform for expert biologists. *Nucleic Acids Res*. 2011;39(Web Server issue): W368-W374. doi:10.1093/nar/gkr440.
[63] Faith JJ, Driscoll ME, Fusaro VA, et al. Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata. Nucleic Acids Res 2008;36(suppl_1):D866–70. https://doi.org/10.1093/nar/gkm815.
[64] Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. In: *Advances in Neural Information Processing Systems*. Vol 25. Curran Associates, Inc.; 2012. Accessed December 21, 2021. https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.
[65] Vincent P, Larochelle H, Bengio Y, Manzagol PA. Extracting and composing robust features with denoising autoencoders. In: *Proceedings of the 25th International Conference on Machine Learning - ICML '08*. ACM Press; 2008:1096-1103. doi:10.1145/1390156.1390294.
[66] Marbach D, Costello JC, Küffner R, et al. Wisdom of crowds for robust gene network inference. Nat Methods 2012;9(8):796–804. https://doi.org/10.1038/nmeth.2016.
[67] Ishchukov I, Wu Y, Van Puyvelde S, Vanderleyden J, Marchal K. Inferring the relation between transcriptional and posttranscriptional regulation from expression compendia. BMC Microbiol 2014;14(1):14. https://doi.org/10.1186/1471-2180-14-14.
[68] Sastry AV, Gao Y, Szubin R, et al. The Escherichia coli transcriptome mostly consists of independently regulated modules. Nat Commun 2019;10(1):5536. https://doi.org/10.1038/s41467-019-13483-w.
[69] Poudel S, Tsunemoto H, Seif Y, et al. Revealing 29 sets of independently modulated genes in Staphylococcus aureus, their regulators, and role in key physiological response. Proc Natl Acad Sci 2020;117(29):17228–39. https://doi.org/10.1073/pnas.2008413117.
[70] Rajput A, Tsunemoto H, Sastry AV, et al. *Machine Learning of* Pseudomonas Aeruginosa *Transcriptomes Identifies Independently Modulated Sets of Genes Associated with Known Transcriptional Regulators*. Bioinformatics 2021. https://doi.org/10.1101/2021.07.28.454220.
[71] Simoes R de M, Emmert-Streib F. Bagging Statistical Network Inference from Large-Scale Gene Expression Data. *PLOS ONE*. 2012;7(3):e33624. doi:10.1371/journal.pone.0033624.
[72] Ling H, Chen B, Kang A, Lee JM, Chang MW. Transcriptome response to alkane biofuels in Saccharomyces cerevisiae: identification of efflux pumps involved in alkane tolerance. Biotechnol Biofuels 2013;6(1):95. https://doi.org/10.1186/1754-6834-6-95.
[73] Ibraim IC, Parise MTD, Parise D, et al. Transcriptome profile of Corynebacterium pseudotuberculosis in response to iron limitation. BMC Genomics 2019;20(1):663. https://doi.org/10.1186/s12864-019-6018-1.
[74] An experimentally supported model of the Bacillus subtilis global transcriptional regulatory network. *Mol Syst Biol*. 2015;11(11):839. doi:10.15252/msb.20156236.
[75] Kang A, Chang MW. Identification and reconstitution of genetic regulatory networks for improved microbial tolerance to isooctane. Mol Biosyst 2012;8 (4):1350–8. https://doi.org/10.1039/C2MB05441H.
[76] Hecker M, Lambeck S, Toepfer S, van Someren E, Guthke R. Gene regulatory network inference: Data integration in dynamic models—A review. Biosystems 2009;96(1):86–103. https://doi.org/10.1016/j.biosystems.2008.12.004.
[77] Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell 2013;35(8):1798–828. https://doi.org/10.1109/TPAMI.2013.50.
[78] Myers CL, Barrett DR, Hibbs MA, Huttenhower C, Troyanskaya OG. Finding function: evaluation methods for functional genomic data. BMC Genomics 2006;7(1):187. https://doi.org/10.1186/1471-2164-7-187.
[79] Hibbs MA, Hess DC, Myers CL, Huttenhower C, Li K, Troyanskaya OG. Exploring the functional landscape of gene expression: directed search of large microarray compendia. Bioinformatics 2007;23(20):2692–9. https://doi.org/10.1093/bioinformatics/btm403.
[80] 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):e00025–e115. https://doi.org/10.1128/mSystems.00025-15.
[81] Tan J, Doing G, Lewis KA, et al. Unsupervised extraction of stable expression signatures from public compendia with an ensemble of neural networks. Cell Syst 2017;5(1):63–71.e6. https://doi.org/10.1016/j.cels.2017.06.003.
[82] 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(1):14. https://doi.org/10.1186/s13040-018-0175-7.
[83] Chen L, Cai C, Chen V, Lu X. Learning a hierarchical representation of the yeast transcriptomic machinery using an autoencoder model. BMC Bioinf 2016;17 (1):S9. https://doi.org/10.1186/s12859-015-0852-1.
[84] Tan J, Huyck M, Hu D, Zelaya RA, Hogan DA, Greene CS. ADAGE signature analysis: differential expression analysis with data-defined gene sets. BMC Bioinf 2017;18:512. https://doi.org/10.1186/s12859-017-1905-4.
[85] Frontiers | VESPUCCI: Exploring Patterns of Gene Expression in Grapevine | Plant Science. Accessed February 10, 2022. https://www.frontiersin.org/articles/10.3389/fpls.2016.00633/full.
[86] Fu Q, Lemmens K, Sanchez-Rodriguez A, et al. Directed Module Detection in a Large-Scale Expression Compendium. In: van Helden J, Toussaint A, Thieffry D, eds. *Bacterial Molecular Networks: Methods and Protocols*. Methods in Molecular Biology. Springer; 2012:131-165. doi:10.1007/978-1-61779-361-5_8.
[87] Neff SL, Hampton TH, Puerner C, et al. CF-Seq, An Accessible Web Application for Rapid Re-Analysis of Cystic Fibrosis Pathogen RNA Sequencing Studies. Published online March 7, 2022:2022.03.07.483313. doi:10.1101/2022.03.07.483313.
[88] Lee AJ, Mould DL, Crawford J, et al. Generative neural networks separate common and specific transcriptional responses. Published online May 24, 2021:2021.05.24.445440. doi:10.1101/2021.05.24.445440.

[89] 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. *mSystems*. Published online March 23, 2021. doi:10.1128/mSystems.01305-20.
[90] Zhou N, Jiang Y, Bergquist TR, et al. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. Genome Biol 2019;20(1):244. https://doi.org/10.1186/s13059-019-1835-8.
[91] Modi SR, Camacho DM, Kohanski MA, Walker GC, Collins JJ. Functional characterization of bacterial sRNAs using a network biology approach. Proc Natl Acad Sci U S A 2011;108(37):15522–7. https://doi.org/10.1073/pnas.1104318108.
[92] Troyanskaya OG, Dolinski K, Owen AB, Altman RB, Botstein D. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). Proc Natl Acad Sci U S A 2003;100(14):8348–53. https://doi.org/10.1073/pnas.0832373100.
[93] Huttenhower C, Hibbs M, Myers C, Troyanskaya OG. A scalable method for integration and functional analysis of multiple microarray datasets. Bioinformatics 2006;22(23):2890–7. https://doi.org/10.1093/bioinformatics/btl492.
[94] Huttenhower C, Flamholz AI, Landis JN, et al. Nearest Neighbor Networks: clustering expression data based on gene neighborhoods. BMC Bioinf 2007;8:250. https://doi.org/10.1186/1471-2105-8-250.
[95] Spoto M, Fleming E, Nzutchi YO, Guan C, Oh J. Large-scale CRISPRi and transcriptomics of Staphylococcus epidermidis identify genetic factors implicated in commensal-pathogen lifestyle versatility. Published online April 29, 2021:2021.04.29.442003. doi:10.1101/2021.04.29.442003.
[96] Guan Y, Dunham MJ, Troyanskaya OG, Caudy AA. Comparative gene expression between two yeast species. BMC Genomics 2013;14(1):33. https://doi.org/10.1186/1471-2164-14-33.
[97] Li H, Rukina D, David FPA, et al. Identifying gene function and module connections by the integration of multispecies expression compendia. Genome Res 2019;29(12):2034–45. https://doi.org/10.1101/gr.251983.119.
[98] Zhu Q, Wong AK, Krishnan A, et al. Targeted exploration and analysis of large cross-platform human transcriptomic compendia. Nat Methods 2015;12(3):211–4. https://doi.org/10.1038/nmeth.3249.
[99] 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.e4. https://doi.org/10.1016/j.cels.2019.04.003.
[100] Zhang D, Guelfi S, Garcia-Ruiz S, et al. Incomplete annotation has a disproportionate impact on our understanding of Mendelian and complex neurogenetic disorders. *Sci Adv*. Published online June 2020. doi:10.1126/sciadv.aay8299.
[101] Gu J, Dai J, Lu H, Zhao H. Comprehensive Analysis of Ubiquitously Expressed Genes in Human, From a Data-Driven Perspective. Published online February 10, 2021:2021.02.09.430465. doi:10.1101/2021.02.09.430465.
[102] Oh S, Geistlinger L, Ramos M, et al. GenomicSuperSignature: interpretation of RNA-seq experiments through robust, efficient comparison to public databases. Published online May 27, 2021:2021.05.26.445900. doi:10.1101/2021.05.26.445900.
[103] Lin CX, Li HD, Deng C, Guan Y, Wang J. TissueNexus: a database of human tissue functional gene networks built with a large compendium of curated RNA-seq data. Nucleic Acids Res 2022;50(D1):D710–8. https://doi.org/10.1093/nar/gkab1133.
[104] Leek JT, Storey JD. Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. PLoS Genet 2007;3(9):e161.
[105] Espín-Pérez A, Portier C, Chadeau-Hyam M, van Veldhoven K, Kleinjans JCS, de Kok TMCM. Comparison of statistical methods and the use of quality control samples for batch effect correction in human transcriptome data. PLoS ONE 2018;13(8):e0202947.
[106] Tseng GC, Oh MK, Rohlin L, Liao JC, Wong WH. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. Nucleic Acids Res 2001;29(12):2549–57. https://doi.org/10.1093/nar/29.12.2549.
[107] Kerr MK, Martin M, Churchill GA. Analysis of variance for gene expression microarray data. J Comput Biol J Comput Mol Cell Biol 2000;7(6):819–37. https://doi.org/10.1089/10665270050514954.
[108] 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). https://doi.org/10.1093/gigascience/giaa117.
[109] Smyth GK, Speed T. Normalization of cDNA microarray data. Methods San Diego Calif 2003;31(4):265–73. https://doi.org/10.1016/s1046-2023(03)00155-5.
[110] Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostat Oxf Engl 2007;8(1):118–27. https://doi.org/10.1093/biostatistics/kxj037.
[111] Van Rossum T, Ferretti P, Maistrenko OM, Bork P. Diversity within species: interpreting strains in microbiomes. Nat Rev Microbiol 2020;18(9):491–506. https://doi.org/10.1038/s41579-020-0368-1.
[112] Price A, Gibas C. The quantitative impact of read mapping to non-native reference genomes in comparative RNA-Seq studies. PLoS ONE 2017;12(7):e0180904. https://doi.org/10.1371/journal.pone.0180904.
[113] Tettelin H, Masignani V, Cieslewicz MJ, et al. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome". Proc Natl Acad Sci 2005;102(39):13950–5. https://doi.org/10.1073/pnas.0506758102.
[114] Zhou W, Spoto M, Hardy R, et al. Host-Specific Evolutionary and Transmission Dynamics Shape the Functional Diversification of Staphylococcus epidermidis in Human Skin. Cell 2020;180(3):454–470.e18. https://doi.org/10.1016/j.cell.2020.01.006.
[115] Conlan S, Mijares LA, Becker J, et al. Staphylococcus epidermidis pan-genome sequence analysis reveals diversity of skin commensal and hospital infection-associated isolates. Genome Biol 2012;13(7):R64. https://doi.org/10.1186/gb-2012-13-7-r64.
[116] Chaves-Moreno D, Wos-Oxley ML, Jáuregui R, Medina E, Oxley APA, Pieper DH. Application of a Novel "Pan-Genome"-Based Strategy for Assigning RNAseq Transcript Reads to Staphylococcus aureus Strains. PLoS ONE 2015;10(12):e0145861.
[117] Chung M, Adkins RS, Mattick JSA, et al. FADU: a Quantification Tool for Prokaryotic Transcriptomic Analyses. *mSystems*. Published online January 12, 2021. doi:10.1128/mSystems.00917-20.
[118] van Opijnen T, Dedrick S, Bento J. Strain Dependent Genetic Networks for Antibiotic-Sensitivity in a Bacterial Pathogen with a Large Pan-Genome. PLoS Pathog 2016;12(9):e1005869. https://doi.org/10.1371/journal.ppat.1005869.
[119] Freschi L, Vincent AT, Jeukens J, et al. The Pseudomonas aeruginosa Pan-Genome Provides New Insights on Its Population Structure, Horizontal Gene Transfer, and Pathogenicity. Genome Biol Evol 2019;11(1):109–20. https://doi.org/10.1093/gbe/evy259.
[120] Freschi L, Jeukens J, Kukavica-Ibrulj I, et al. Clinical utilization of genomics data produced by the international Pseudomonas aeruginosa consortium. *Front Microbiol*. 2015;6. Accessed January 24, 2022. https://www.frontiersin.org/article/10.3389/fmicb.2015.01036.
[121] Palma M, DeLuca D, Worgall S, Quadri LEN. Transcriptome analysis of the response of *Pseudomonas aeruginosa* to hydrogen peroxide. J Bacteriol 2004;186(1):248–52. https://doi.org/10.1128/JB.186.1.248-252.2004.
[122] Ochsner UA, Wilderman PJ, Vasil AI, Vasil ML. GeneChip® expression analysis of the iron starvation response in Pseudomonas aeruginosa: identification of novel pyoverdine biosynthesis genes. Mol Microbiol 2002;45(5):1277–87. https://doi.org/10.1046/j.1365-2958.2002.03084.x.
[123] Aspedon A, Palmer K, Whiteley M. Microarray analysis of the osmotic stress response in *Pseudomonas aeruginosa*. J Bacteriol 2006;188(7):2721–5. https://doi.org/10.1128/JB.188.7.2721-2725.2006.
[124] Finck-Barbançon V, Goranson J, Zhu L, et al. ExoU expression by *Pseudomonas aeruginosa* correlates with acute cytotoxicity and epithelial injury. Mol Microbiol 1997;25(3):547–57. https://doi.org/10.1046/j.1365-2958.1997.4891851.x.
[125] Nunn D, Bergman S, Lory S. Products of three accessory genes, pilB, pilC, and pilD, are required for biogenesis of *Pseudomonas aeruginosa* pili. J Bacteriol 1990;172(6):2911–9. https://doi.org/10.1128/jb.172.6.2911-2919.1990.
[126] Lee AJ, Doing G, Neff SL, Reiter T, Hogan DA, Greene CS. Compendium-wide analysis of P. aeruginosa core and accessory genes reveal more nuanced transcriptional patterns. Published online April 15, 2022:2022.04.14.488429. doi:10.1101/2022.04.14.488429.
[127] Bhandary P, Seetharam AS, Arendsee ZW, Hur M, Wurtele ES. Raising orphans from a metadata morass: A researcher's guide to re-use of public 'omics data. Plant Sci 2018;267:32–47. https://doi.org/10.1016/j.plantsci.2017.10.014.
[128] Gonçalves RS, Musen MA. The variable quality of metadata about biological samples used in biomedical experiments. Sci Data 2019;6(1):. https://doi.org/10.1038/sdata.2019.21190021.
[129] Davis S, Meltzer PS. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics 2007;23(14):1846–7. https://doi.org/10.1093/bioinformatics/btm254.
[130] Irber L, Brooks PT, Reiter T, et al. Lightweight compositional analysis of metagenomes with FracMinHash and minimum metagenome covers. Published online January 17, 2022:2022.01.11.475838. doi:10.1101/2022.01.11.475838.
[131] Zhou W, Gay N, Oh J. ReprDB and panDB: minimalist databases with maximal microbial representation. Microbiome 2018;6(1):15. https://doi.org/10.1186/s40168-018-0399-2.
[132] Rajagopala SV, Bakhoum NG, Pakala SB, et al. Metatranscriptomics to characterize respiratory virome, microbiome, and host response directly from clinical samples. Cell Rep Methods 2021;1(6):. https://doi.org/10.1016/j.crmeth.2021.100091100091.
[133] Chaves-Moreno D, Wos-Oxley ML, Jáuregui R, Medina E, Oxley AP, Pieper DH. Exploring the transcriptome of Staphylococcus aureus in its natural niche. Sci Rep 2016;6(1):33174. https://doi.org/10.1038/srep33174.
[134] Chow YL, Singh S, Carpenter AE, Way GP. Predicting drug polypharmacology from cell morphology readouts using variational autoencoder latent space arithmetic. PLOS Comput Biol 2022;18(2):e1009888.
[135] Lotfollahi M, Wolf FA, Theis FJ. scGen predicts single-cell perturbation responses. Nat Methods 2019;16(8):715–21. https://doi.org/10.1038/s41592-019-0494-8.
[136] Way GP, Zietz M, Rubinetti V, Himmelstein DS, Greene CS. Compressing gene expression data using multiple latent space dimensionalities learns

complementary biological representations. Genome Biol 2020;21(1):109. https://doi.org/10.1186/s13059-020-02021-3.
[137] Way GP, Greene CS. Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. Pac Symp Biocomput Pac Symp Biocomput 2018;23:80–91.
[138] Gut G, Stark SG, Rätsch G, Davidson NR. PmVAE: Learning Interpretable Single-Cell Representations with Pathway Modules. Bioinformatics 2021. https://doi.org/10.1101/2021.01.28.428664.
[139] Lemmens K, De Bie T, Dhollander T, et al. DISTILLER: a data integration framework to reveal condition dependency of complex regulons in Escherichia coli. Genome Biol 2009;10(3):R27. https://doi.org/10.1186/gb-2009-10-3-r27.
[140] Sherlock G, Hernandez-Boussard T, Kasarskis A, et al. The Stanford Microarray Database. Nucleic Acids Res 2001;29(1):152–5.