




A Comprehensive Coexpression Network Analysis in *Vibrio cholerae*

Cory D. DuPai,^a  Claus O. Wilke,^{a,b} Bryan W. Davies^{a,c}

^aInstitute for Cellular and Molecular Biology, University of Texas at Austin, Austin, Texas, USA

^bDepartment of Integrative Biology, University of Texas at Austin, Austin, Texas, USA

^cDepartment of Molecular Biosciences, University of Texas at Austin, Austin, Texas, USA

ABSTRACT Research into the evolution and pathogenesis of *Vibrio cholerae* has benefited greatly from the generation of high-throughput sequencing data to drive molecular analyses. The steady accumulation of these data sets now provides a unique opportunity for *in silico* hypothesis generation via coexpression analysis. Here, we leverage all published *V. cholerae* RNA sequencing data, in combination with select data from other platforms, to generate a gene coexpression network that validates known gene interactions and identifies novel genetic partners across the entire *V. cholerae* genome. This network provides direct insights into genes influencing pathogenicity, metabolism, and transcriptional regulation, further clarifies results from previous sequencing experiments in *V. cholerae* (e.g., transposon insertion sequencing [Tn-seq] and chromatin immunoprecipitation sequencing [ChIP-seq]), and expands upon microarray-based findings in related Gram-negative bacteria.

IMPORTANCE Cholera is a devastating illness that kills tens of thousands of people annually. *Vibrio cholerae*, the causative agent of cholera, is an important model organism to investigate both bacterial pathogenesis and the impact of horizontal gene transfer on the emergence and dissemination of new virulent strains. Despite the importance of this pathogen, roughly one-third of *V. cholerae* genes are functionally unannotated, leaving large gaps in our understanding of this microbe. Through coexpression network analysis of existing RNA sequencing data, this work develops an approach to uncover novel gene-gene relationships and contextualize genes with no known function, which will advance our understanding of *V. cholerae* virulence and evolution.

KEYWORDS *Vibrio cholerae*, computational biology

Since the completion of the first *Vibrio cholerae* genome sequence in 2000, over 1,000 *V. cholerae* isolates have been sequenced (1, 2). These sequences have allowed for the development of sophisticated phylogeographic models, which emphasize the importance of controlling the spread of virulent and antibiotic-resistant *V. cholerae* strains to lower disease burden, in addition to fighting endemic local strains (2–6). The integration of hundreds of genomes paired with temporal and geographic information into ever-growing phylogenies enables analyses using selection models to predict future population trends and derive biologically meaningful insights into *V. cholerae* evolution (7, 8). By developing treatment and vaccination strategies based on phylogenetic models (9), organizations and governments can more efficiently leverage limited resources and more effectively prevent disease spread in line with the World Health Organization's goal of eradicating cholera by 2030 (10).

Alongside advances in genomics research, the *V. cholerae* and broader bacterial biology communities have benefited greatly from other next-generation sequencing (NGS) technologies. Targeted sequencing experiments have been essential in mapping

Citation DuPai CD, Wilke CO, Davies BW. 2020.

A comprehensive coexpression network analysis in *Vibrio cholerae*. mSystems 5:e00550-20. <https://doi.org/10.1128/mSystems.00550-20>.

Editor Sergio Baranzini, University of California, San Francisco

Copyright © 2020 DuPai et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

Address correspondence to Bryan W. Davies, bwdavies@austin.utexas.edu.

Received 17 June 2020

Accepted 24 June 2020

Published 7 July 2020

complex virulence pathways, illuminating a novel interbacterial defense system, and expanding our knowledge of the role of noncoding RNAs (ncRNAs) in the *Vibrio* life cycle (11–17). Further discoveries, such as transcription factor-mediated transposon insertion bias (18) and the role of cAMP receptor protein in host colonization (19), have benefited from composite research strategies utilizing multiple technologies. Similarly, meta-analyses utilizing pooled data from multiple experiments are empowered by the increasing availability of high-quality bacterial NGS data sets. Expression data are particularly amenable to such pooling and can be used to accurately group genes into functional modules based on their coexpression (20). In bacteria, weighted gene coexpression network analysis (WGCNA) (21) has been successfully used to underscore biologically important genes and gene-gene relationships via “guilt-by-association” approaches (22, 23). These studies have taken advantage of larger and larger heterogeneous microarray data sets to provide novel biological insights via existing data.

Despite major advances in sequencing technologies and research strategies, most of the over two dozen existing transcriptome sequencing (RNA-seq) experiments in *V. cholerae* have been limited to targeted approaches that involve quantifying the differential abundance of genetic material across a few conditions. Via these approaches, it is nearly impossible to generalize about any change in expression observed in one experiment to other treatment conditions, and analyses are limited to a few pathways or genes of interest. In contrast, meta-analyses such as WGCNA can uncover much broader relationships throughout the genome by combining information from multiple data sets. As there is no existing coexpression analysis in *V. cholerae* to date, the accumulation of over 300 publicly available RNA-seq samples from targeted RNA-seq experiments represents a heretofore untapped resource for the cholera community.

Motivated by the success of pooled genetic sequencing analyses, our current work utilizes all publicly available *V. cholerae* RNA-seq-based expression-level data to generate a coexpression network. We expand upon existing bacterial WGCNA approaches by integrating broader sequencing data (including chromatin immunoprecipitation sequencing [ChIP-seq] and transposon insertion sequencing [Tn-seq]) and multiple annotation platforms into our analysis. Our network ultimately contributes information on connections across all *V. cholerae* genes, including the roughly 1,500 predicted but functionally unannotated genetic elements that account for some 37% of the genome. More specifically, we implicate new loci in virulence regulation and clearly demonstrate a powerful and accurate approach to hypothesis generation via our described network.

RESULTS

Gene network generation. To generate our coexpression analysis in *V. cholerae*, we applied our WGCNA pipeline to analyze 27 *V. cholerae* RNA sequencing experiments deposited in NCBI’s Sequence Read Archive (SRA) in addition to two novel experiments. The RNA sequencing samples are derived from experiments exploring a range of important *V. cholerae* processes including intestinal colonization, quorum sensing, and stress response. In total, our network includes 300 individual RNA-seq samples (see Table S1 in the supplemental material). All samples were mapped to a recently inferred *V. cholerae* transcriptome derived from the N16961 reference genome (1, 13). This reference was chosen because the majority (293) of samples were collected from strain N16961 or the closely related strains C6706 and A1552.

Figure 1 outlines the process used to generate our coexpression network with a small subset of genes. The five included loci are known to be involved in cysteine metabolism with loci VC0384 to VC0386 and loci VC0539 and VC0540 falling within two separate operons. Following normalization of mapped transcripts (Fig. 1A), a weighted gene coexpression network analysis was performed using WGCNA, as follows (21). First, a Pearson correlation matrix was calculated for expression levels of all genes (Fig. 1B). This correlation matrix clearly captures strong relationships between coexpressing genes but can produce background noise from unrelated gene pairs and underlying gene structures (i.e., operons). We limit this noise by calculating a topological overlap matrix (TOM) (24) that weights pairwise coexpression data based on each gene’s

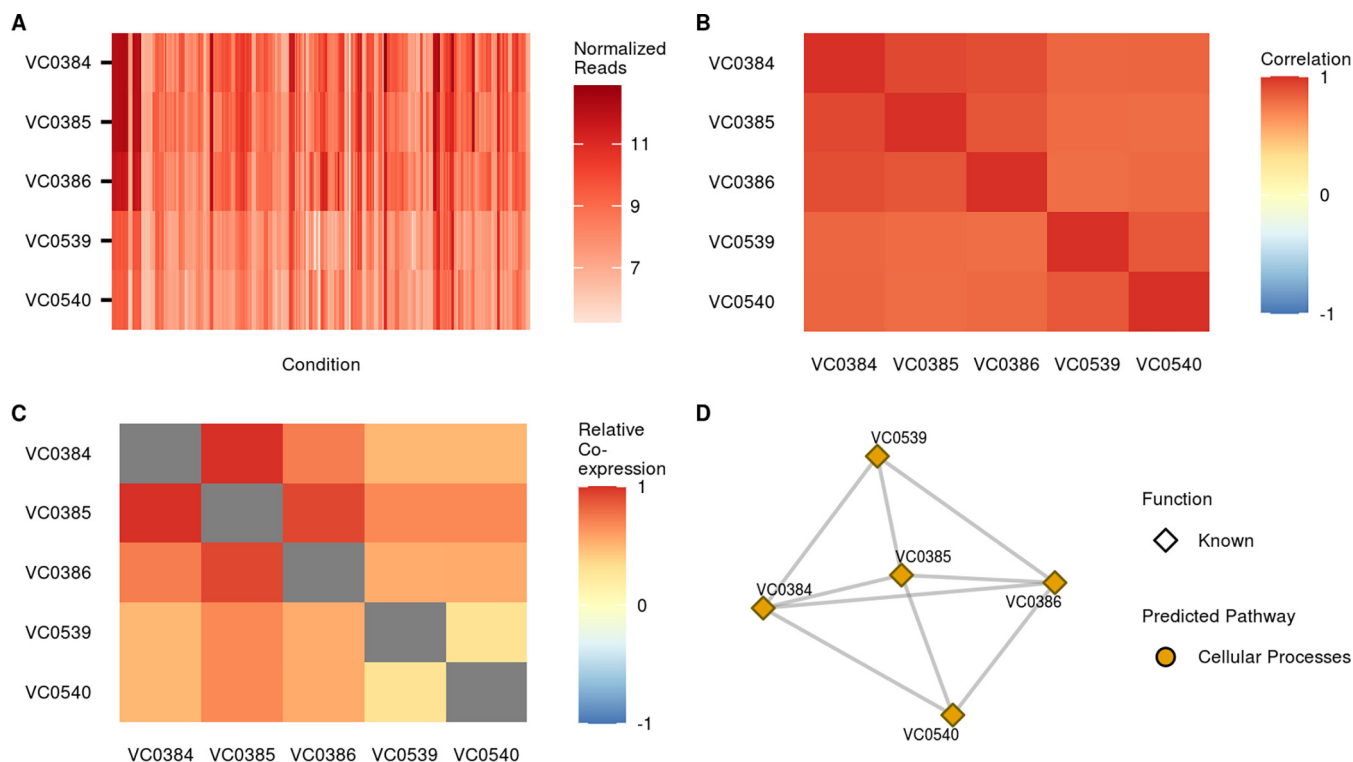


FIG 1 General outline of network construction. To explain the overall WGCNA process, we have chosen a subset of genes that are involved in the same core process, cysteine metabolism. Loci VC0394 to VC0386 are predicted to fall within one operon while loci VC0539 and VC0540 are predicted to be in another. (A) Normalized (log₂) expression reads for the same genes across multiple conditions supply the basis for our coexpression analysis. (B) Correlations are calculated from the normalized counts shown in panel A for every pair of genes. (C) An adjacency matrix (not shown) was calculated from the correlations shown in panel B and ultimately used to produce a topological overlap matrix (TOM) that supplies network edge weights with less noise than the raw correlation matrix. While the signal of coexpressing pairs is dampened slightly, this step greatly decreases spurious relationships as it favors transcripts which coexpress with similar sets of genes rather than potentially noisy direct correlations. (D) The final network groups transcripts that tightly coexpress while indicating what pathways they are involved in. In this example, all genes significantly coexpress, with the exception of VC0539 and VC0540 despite their colocalization within the same operon. After network construction, information was added to label genes based on their function and essentiality under virulence and growth conditions.

interactions with all other genes (Fig. 1C). In this way, the relationships between genes that fall within the same subnetwork are favored while signals from less tightly coregulated genes are abated. This TOM, after filtering is performed for normalized values greater than 0.1, is used to construct an accurate coexpression network that captures biologically meaningful relationships while minimizing background noise (Fig. 1D).

In addition to coexpression data, our network and analyses incorporated information from multiple other sources. Our network includes predicted pathway annotations and gene functional knowledge from the NCBI Biosystems database as well as the DAVID, Panther, and KEGG databases (25–28). Operon structure was inferred using Operon-mapper (29). Additionally, importance labels were applied to genes with no known function which have been implicated as playing a role in intestinal colonization or *in vitro* growth via *Tn*-seq-based essentiality experiments (14, 30). Information from ChIP-seq binding assays and microarray results were incorporated in downstream analyses to substantiate network-derived relationships. By combining all of these data sources, we were able to develop and analyze an informative network of coexpressing genes that provides both qualitative and quantitative information about relationships between transcripts across 49 gene clusters covering the entire *V. cholerae* genome (Data Set S1 and S2).

A network of novel, unexpected, and informative interactions. As many functionally related bacterial genes are coexpressed in operons such as the operon of VC0384 to VC0386 above, we sought to discover if operon structure was a contributing

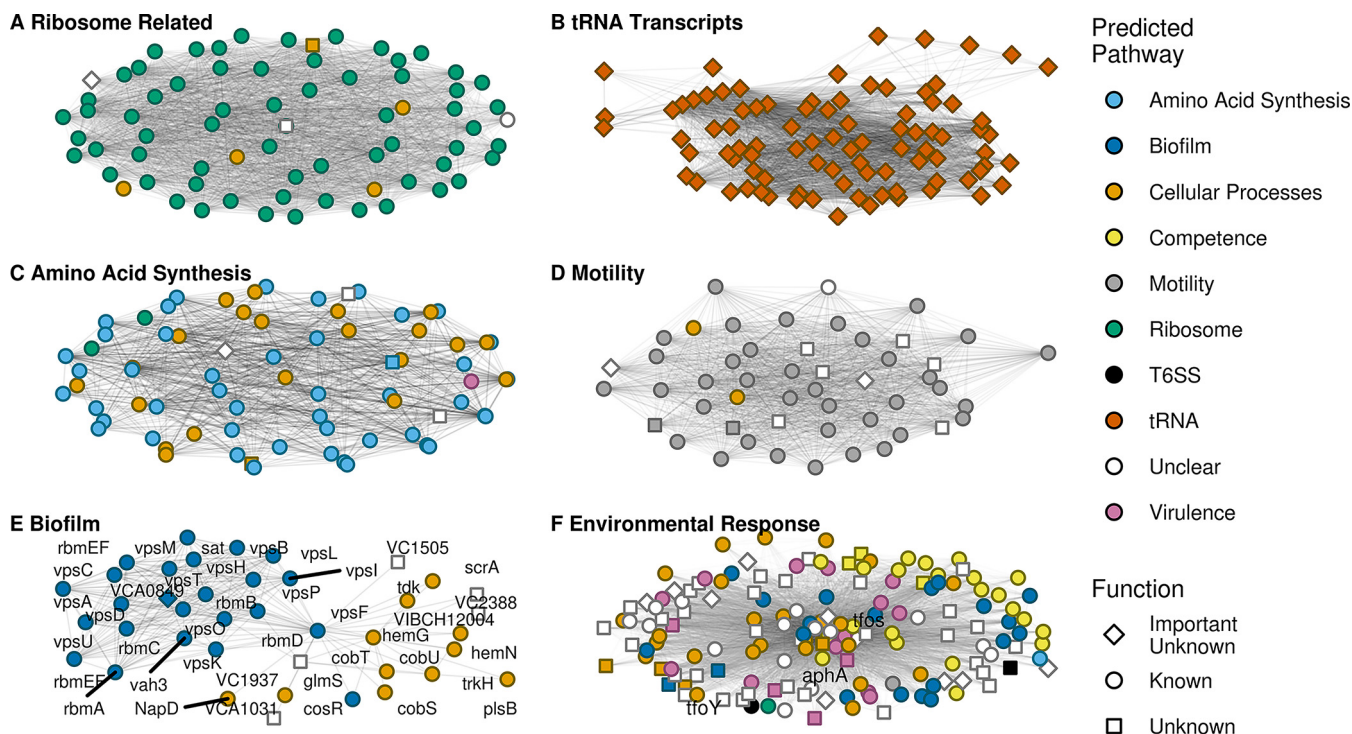


FIG 2 Subnetworks recapitulating known results. The depicted subnetworks each contain transcripts that are known to be largely involved in one or more related biological process(es). For each subnetwork, the nodes represent transcripts while the edges represent a coexpression relationship of at least 0.1 between transcripts. (A to F) Subnetworks involved in the following core processes: ribosome related, tRNA transcripts, amino acid synthesis, motility, biofilm, and environmental response.

factor to our network or specific subnetworks. Indeed, gene pairs predicted to fall within the same operon did show significantly higher average normalized coexpression than their nonoperon counterparts (0.186 versus 0.147; $P < 0.001$), and some subnetworks, such as the ribosome-related subnetwork (Fig. 2A), contain a high proportion of intraoperon gene pairs (Fig. S1). However, across our full network only 0.2% of all coexpressing gene pairs fell within the same operon, and no subnetwork had a majority of such pairs (Fig. S1). Moreover, our overall network captured information on relationships with roughly one-third of unannotated *V. cholerae* genes (Fig. S2), providing insight into functional roles that are not obvious based on gene homology or known operon structure.

Genes in known pathways cluster together and contextualize genes of unknown function. As a demonstration of the accuracy of our approach, we have highlighted several clusters that recapitulate known interactions between transcripts involved in highly conserved, well-studied cellular processes (Fig. 2). The correct grouping of transcripts encoding ribosomal proteins, tRNAs, and amino acid synthesis proteins into significantly coexpressing subnetworks provided a positive control for our overall network (Fig. 2A to C). Importantly, our analysis clustered together genes known to be involved in more specialized processes such as motility and biofilm formation (Fig. 2D and E), with corresponding Gene Ontology (GO) (31) and KEGG (27) pathway terms enriched for genes within these subnetworks (Fig. 3 and Table S2).

In addition to capturing relationships between genes involved in specific pathways, our approach can also accurately group genes involved in interconnected processes that share overlapping regulation, as seen in the environmental sensing subnetwork (Fig. 2F). This subnetwork includes high-level transcriptional regulators, such as Apha, TfoS, and TfoY, with known roles mediating the complex interplay between quorum sensing, natural competence, type VI secretion, and other related pathways (32–37). As each of these transcription factors is involved in a multitude of cellular processes and

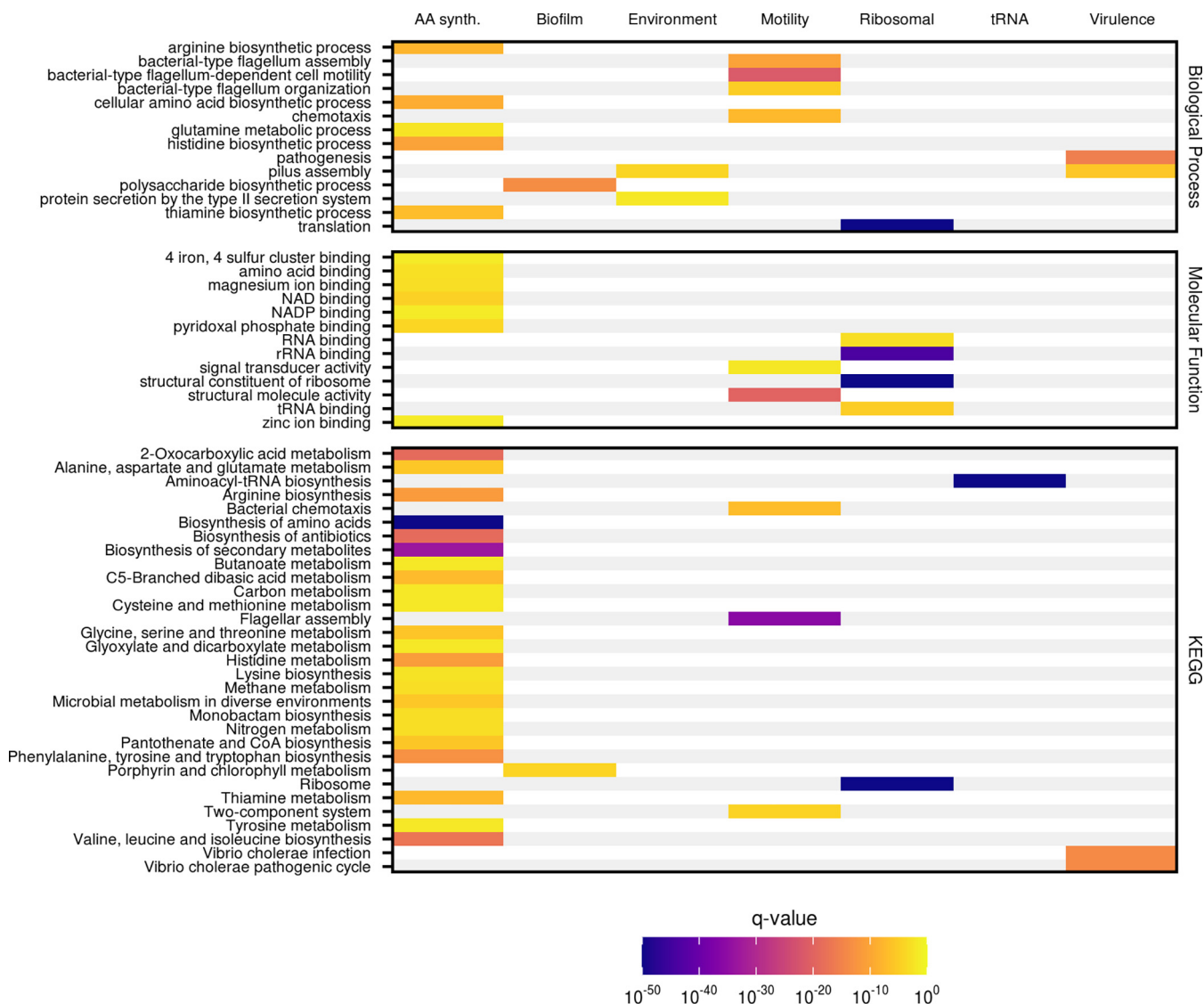


FIG 3 Significantly enriched GO and KEGG terms for specific subnetworks. The indicated terms are significantly enriched within highlighted pathways, with the color indicating the significance of the enrichment as determined via the false-discovery rate-adjusted *P* value (*q* value). The terms are divided by database and, for Gene Ontology (GO) terms, GO domain, as indicated to the right.

significantly coexpresses with hundreds of other genes, our analysis describes their closest connections under parameters designed to find meaningful and practically interpretable relationships. By altering these parameters (significance cutoffs, minimum number of genes per cluster, clustering algorithm, etc.), analysis of the overall network can be fine-tuned to focus in on specific biological processes or explore the nodes that drive connections between processes that are necessary for *V. cholerae* to adapt and survive in diverse environments.

The subnetworks outlined in Fig. 2 support the utility of our analysis in powering the inference of gene function based on guilt by association (38). Because each of these gene clusters contains coexpressing genes that are involved in the same biological process, it can be assumed that unannotated genes in the same cluster are likely involved in the same process. Such links, while not definitive on their own, can be used with other data to hint at gene functions. For example, the genes with known function shown in Fig. 2E are primarily involved in biofilm formation (39, 40). This clustering of biofilm genes suggests that the few genes with no known function in this subnetwork may be involved in the same process. Two of these unannotated transcripts, VC1937

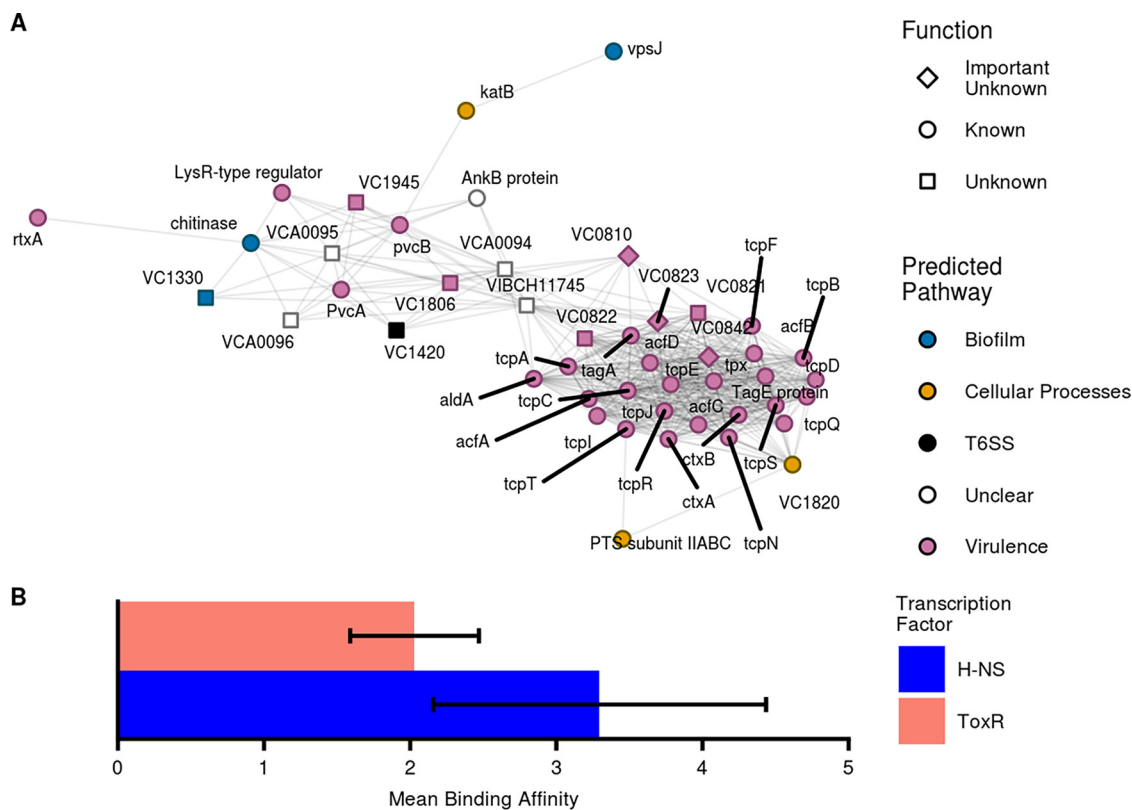


FIG 4 Virulence-related subnetwork. (A) This subnetwork contains a majority of genes that are predicted to be involved in virulence-related pathways, providing clues to the genes with no known functions, such as those at loci VCA0094 to VCA0096. PTS subunit IIABC, phosphotransferase system fructose-specific transporter subunit IIABC; T6SS, type 6 secretion system. (B) Mean binding affinity (\log_2 fold change in occupancy compared to level of the loading control) for different virulence-associated transcription factors near loci VCA0094 to VCA0096. Both H-NS and ToxR show significant binding preferences for this region. Error bars indicate standard deviations from the means.

and VC2388, are, per GO cellular component location labels, integral membrane components. Further, the VC2388 locus is directly upstream of a Vcr084, a short RNA involved in quorum sensing which is essential for biofilm formation (41). Taken together, this evidence suggests that VC1937 and VC2388 may play a role in some of the complex membrane restructuring necessary for biofilm formation. In facilitating such guilt-by-association approaches to novel hypothesis generation, our coexpression network serves as a highly efficient substitute for more traditional screening assays.

A virulence subnetwork suggests novel gene functions. While the biofilm-associated subnetwork (Fig. 2E) presents a relatively simple example of the functional insights our coexpression data can yield, the virulence-related subnetwork (Fig. 4A) represents a more complex case in which genes of known function provide clues to the role of unannotated genes. The majority of transcripts in this module originate from within the virulence-related ToxR regulon that consists principally of genes on *V. cholerae* pathogenicity island 1 (VPI-1) (VC0809 to VC0848) and cholera toxin subunits A and B (*ctxAB*, VC1456, and VC1457) (42). Other genes in this subnetwork, such as *vpsJ*, VC1806, VC1810, and chitinase, are predominately localized to virulence islands and other areas of the genome under tight control of the known virulence regulator ToxR, ToxT, or H-NS as determined via ChIP-seq and/or RNA-seq (43–45). Genes in this subnetwork are also enriched for virulence-related GO and KEGG terms, such as “pathogenesis” and “*Vibrio cholerae* infection” (Fig. 3). The clustering of such genes with well-characterized interactions into a cohesive subnetwork is further validation of our ability to generate accurate coexpression maps of related genes. The association of uncharacterized genes in these clusters suggests that they may also play a role in *V.*

cholerae virulence and generates hypotheses about the function of unknown genes within this module.

Many of the important transcripts with unknown function are expected to coexpress with known virulence genes because they fall within VPI-1 (VC0810, VC0821 to VC0823, and VC0842) or VPI-2 (VC1806 and VC1810) or are proximal to other virulence genes (VC1945) (46, 47). However, our analysis also identified genes such as VCA0094 to VCA0096 which are on a completely different chromosome than the rest of the subnetwork and do not neighbor any known virulence elements.

A major benefit of our approach is that we incorporate additional regulatory data such as ChIP-seq and Tn-seq into our coexpression analysis, allowing us to verify the association between VCA0094 to VCA0096 and virulence pathways using existing experimental data. Tn-seq analysis has previously identified VCA0094 and VCA0095 as essential for infection of a rabbit intestine (14), suggesting that these loci play a role in virulence. Because transcripts for these genes coexpress with genes regulated by ToxT, ToxR, and H-NS, we also probed existing ChIP-seq binding data sets (12, 19, 43) to see if any of these well-studied transcription factors bind near loci VCA0094 to VCA0096. While ToxT binding was not observed near this site (data not shown), our analysis identified significant peaks in the promoter region of VCA0094 for both ToxR and H-NS, as calculated via reanalysis of existing binding data from Kazi et al. (43). Both peaks showed large and significant increases in binding affinity (\log_2 fold change in average occupancy) compared to levels of the input controls (Fig. 4B). H-NS showed a clear binding peak in the region of the VCA0094 promoter that extended in a diffuse manner to the VCA0095 transcription start site while ToxR binding covered a similar region but was more diffuse throughout (data not shown). Collectively, these results indicate virulence-related functions for the products of the transcripts of VCA0094 to VCA0096. Although the exact mechanistic role of these genes remains elusive, we have nevertheless demonstrated the ability of our pipeline to generate meaningful hypotheses by incorporating existing data from a multitude of sources.

Coexpression data provides an accurate *in silico* complement to RNA-seq. In addition to the guilt-by-association inference described above, coexpression analysis can provide a partial substitute or complement to RNA-seq experiments. Novel, meaningful genetic relationships can be found in a coexpression network by focusing on the transcripts that are coregulated with a gene of interest.

We can apply a network-based approach in lieu of new RNA-seq-based experiments to identify genes which coexpress with *rpoS* (VC0534) and are similarly involved in the bacterial stress response. As our network utilizes only RNA-seq-based transcriptomics studies and as none of these studies involves direct manipulation of *rpoS*, we can compare existing microarray data involving an *rpoS* (VC0534) deletion mutant (48) to determine how accurate our approach is. When an absolute coexpression cutoff of 0.1 is applied, 272 genes are identified as having a relationship with *rpoS* expression in both our network analysis and the *rpoS* mutant microarray data (Fig. 5A). This represents nearly two-thirds of genes identified as differentially expressed in the original microarray study. While our network links far more genes with *rpoS* than the microarray approach, this is in line with recent RNA-seq-based work that found that 23% of the *Escherichia coli* genome is regulated by RpoS (49). Additionally, all of the flagellum- and chemotaxis-related proteins highlighted as particularly informative in the original study were identified by our analysis (Fig. 5B), and relevant values (i.e., network coexpression and microarray-derived log fold change in expression) for the 273 shared transcripts have a Spearman correlation of -0.516 . This accuracy was achieved without any direct genetic manipulation of the *rpoS* locus in the RNA-seq data sets used to generate our coexpression network and serves as a testament to the potential utility and versatility of our approach.

Our approach to isolating genetic interactions also has advantages over transcriptomics-focused sequencing. As seen in Fig. 5A, our network-based analysis identified far more genes associated with *rpoS*. This is likely because RNA-seq-based

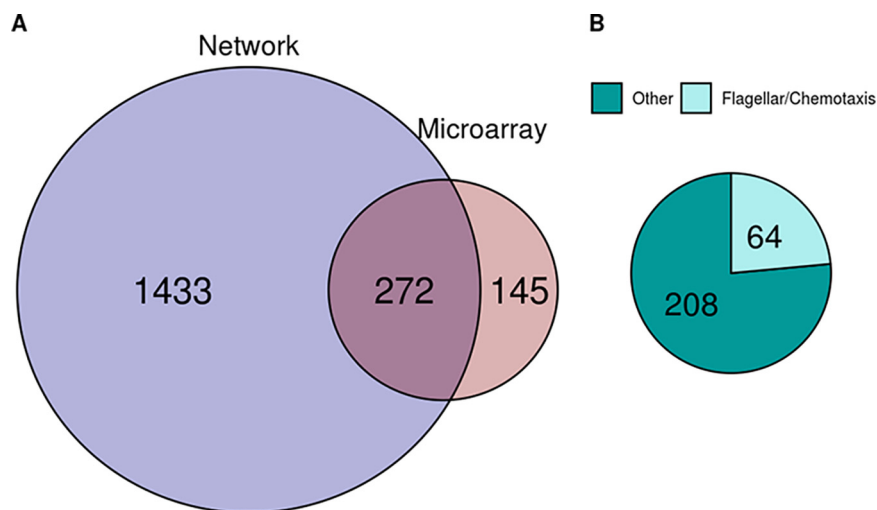


FIG 5 Comparing RpoS microarray data to data of coexpressing genes in our WGCNA. (A) Overlap of genes with expression patterns related to the pattern of *rpoS* expression as identified via our network analysis (blue) and existing microarray data (red). The overlapping region identifies 272 genes that are common between the two analyses. (B) Breakdown of shared genes (overlapping region in panel A). All of the flagellar and chemotaxis genes highlighted as particularly important in the microarray data set are identified by both methods.

approaches can identify a broader range of gene transcripts as they are not limited by restrictive microarray probes (50). Separate from differences in underlying technology, coexpression networks are also more likely to detect genes regulating a target's expression than traditional transcriptomics experiments, which largely capture downstream responses to changes in a target's expression (51, 52). Thus, a coexpression network can provide an alternative perspective to complement or clarify transcriptomics data.

DISCUSSION

We have successfully constructed the first *V. cholerae* coexpression network through a computationally inexpensive process that is simple, easily expanded upon, and straightforward to implement in other organisms. Our network effectively identifies canonical gene clusters related to specific molecular pathways or functions, such as those corresponding to tRNAs or biofilm proteins. We have also outlined two use cases for the data provided and have shown the accuracy of both approaches using existing data. Additionally, we have included relevant network files as well as raw read counts across RNA-seq conditions (see Data Sets S1 and S2 and Table S3 in the supplemental material) alongside all code used in our analysis (see Materials and Methods) to encourage broad usage of these data.

Our results have proven both the utility and accuracy of our approach despite in-depth analysis limited to a few genes across 5 of the 49 observed gene clusters. Furthermore, our work with the virulence subnetwork supports previously published research loosely implicating genes VCA0094 to VCA0096 in virulence and virulence-related functions. All three transcripts have shown up in screens focusing on biofilm development (53) and the SOS response (13). From a mechanistic perspective, protein homology analysis via NCBI's Conserved Domain Database (54) indicates that VCA0094 possesses a DNA-binding transcriptional regulator domain while VCA0096 contains domains that implicate it in protein activation via proteolysis. These data combined with our novel findings hint at the potential biological importance of this genomic locus.

When viewed through the lens of a specific gene of interest, coexpression data are in large part analogous to the differential expression data produced by RNA-seq experiments. While RNA-seq offers finer assay control and can be tailored more exactly

to suit a specific research question, there are both technical and practical limitations that may make such an approach impractical. Whether an experimenter is interested in examining the role of an essential locus or is limited by available resources, our coexpression analysis presents a fast, free, and faithful alternative for probing genetic interactions, as outlined in our analysis of *rpoS* above.

Major motivations for this work include the successful implementation of bacterium-focused, microarray-based coexpression networks and the lack of clear functional knowledge for a large portion of *V. cholerae* genes. In addition to simpler guilt-by-association studies (22, 23), coexpression networks have helped to elucidate relationships in diverse microbial communities (55–58) and to enable comparisons across strains and species (59–61). These works as well as the relative dearth of knowledge about the *V. cholerae* genome (roughly two-thirds of genes are annotated whereas around 86% percent of all *E. coli* genes are annotated [62]) and the growing abundance of *V. cholerae*-focused NGS data served as the impetus for this research.

The calculated coexpression network, though accurate, could be improved via the inclusion of more experiments and more extensive SRA annotations. Our somewhat limited pooled data set consisting of 300 samples is an order of magnitude below the few thousand samples necessary to derive the most faithful coexpression estimates (63). Though sample size will improve as more *V. cholerae* RNA-seq experiments are published, more samples may also increase the risk posed by batch effects which cause spurious correlations among genes through technical variation (64, 65). The diverse structure of our current data helps to minimize the impact of batch effects, but this would be offset by the future inclusion of larger data sets from single experiments. While automated sample clustering methods (66–68) can effectively group overly correlated samples, there is no way to know if the correlation is biological (i.e., meaningful) or technical (i.e., noise) in origin. Similarly, manual curation of batch annotations is also difficult since few SRA records are extensively annotated with detailed experimental conditions (e.g., bacterial growth stage or exact medium used). Thus, careful consideration may be necessary when expanding and generalizing this analysis to include future data.

The mapping of raw reads to a transcriptome derived from a single reference genome presents a limitation to our current work. While this approach is reasonable given the similarity of the vast majority of included strains to our reference, a more elaborate comparative transcriptomic strategy (69, 70) would be ideal if more diverse samples are included in future analyses. This is especially true when we consider the inclusion of expression data from clinical samples which are likely to have much more genomic variability than the closely related lab cultured strains used to construct our network. On the other hand, because comparative transcriptomics requires defining homologous alleles across all strains analyzed (71), such an approach would greatly increase the difficulty of incorporating strains without an assembled genome.

In summary, our coexpression network can drive functional hypotheses for unannotated genes in *V. cholerae*. As the *Vibrio* community steadily adds high-quality data from increasingly sophisticated sequencing experiments to public databases, our imputed network can only improve, providing ever-deeper insights into the *V. cholerae* genome. At the same time, highly annotated transcript-based coexpression networks can empower research with related technologies (e.g., single-cell transcriptomics and dual RNA-seq) and research into a host of other clinically relevant bacteria, such as *Pseudomonas aeruginosa* or *Staphylococcus aureus*, for which there are over 2,000 and 1,400 RNA-seq experiments, respectively, in the SRA.

MATERIALS AND METHODS

Data collection and processing. All RNA and ChIP sequencing data were downloaded from the Sequence Read Archive (SRA) (72) and converted to compressed fastq files using the SRA Toolkit (<https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software>) (see Table S1 in the supplemental material for details on included experiments). RNA-seq samples were selected by searching the SRA on 10 September 2019 for the organism and strategy terms “*vibrio cholerae*” and “rna seq,” respectively, resulting in 326 initial samples including the 34 novel samples from this publication. Samples were

mapped to a recently inferred *V. cholerae* transcriptome derived from the N16961 reference genome (1, 13) using Kallisto, version 0.45.1 (73). This reference was chosen because the majority (293) of samples were collected from strain N16961 or the closely related C6706 and A1552 strains. Twenty-six low-quality samples with <50% of reads mapping to the reference transcriptome were discarded before further analysis, leaving 300 samples used for further analysis.

For ChIP-seq analysis, accession numbers were identified via the relevant publications (12, 19, 43), and sequences were downloaded from the SRA and converted to fastq files as described above. Raw reads were mapped to the same N16961 reference genome using Bowtie 2, version 2.3.5.1 (74). From this mapping, peaks were identified using MACS2, version 2.1.2, with the parameter *extsize* set at 225 (various sizes from 150 to 500 were tested with little observable difference in the peaks identified) (75), and differential binding and significance were calculated using DiffBind, version 2.12.0 (76).

Processed Tn-seq data were collected directly from published data sets. *In vitro* essentiality and semiessentiality labels were derived from Table S1 in Chao et al. (30), with the original labels of domain essential and sick genes replaced with essential and semiessential, respectively. We used Table S2 from Fu et al. (14) to label genes involved in host infection, with any gene exhibiting a \log_2 fold change of less than -3 deemed essential and any gene with a \log_2 fold change between -1 and -3 deemed semiessential.

Network construction. Figure 1 highlights the process used to generate our coexpression network. Kallisto-derived reads were first imported into R via *tximport* (77) and then normalized using *DESeq2*, version 1.24.0 (78), resulting in values that are comparable across conditions and experiments. Following normalization, a weighted gene coexpression network analysis was performed using *WGCNA* (21). This process is highlighted with a subset of data in Fig. 1 and consists of the sequential calculation of a Pearson correlation matrix, adjacency matrix with power $\beta = 6$, and, ultimately, a topological overlap matrix (TOM) (24) from normalized gene expression counts across conditions. We further filtered this TOM to exclude samples with weighted coexpression of <0.1 for all analyses included in Results.

Predicted pathway annotations and gene functional knowledge were derived from the NCBI Biosystems database as well as the DAVID, Panther, and KEGG databases (25–28). Genes for which functional knowledge is lacking and which are identified as essential or semiessential in either Tn-seq data set are labeled in network visualizations as “important unknown.” Operon predictions were inferred using *Operon-mapper* (29).

Data availability. Information on the 34 novel samples are available in the SRA database under accession numbers [SRR10905341](https://www.ncbi.nlm.nih.gov/sra/SRR10905341) to [SRR10905344](https://www.ncbi.nlm.nih.gov/sra/SRR10905344), [SRR10905351](https://www.ncbi.nlm.nih.gov/sra/SRR10905351), [SRR10905362](https://www.ncbi.nlm.nih.gov/sra/SRR10905362), and [SRR10905369](https://www.ncbi.nlm.nih.gov/sra/SRR10905369) to [SRR10905396](https://www.ncbi.nlm.nih.gov/sra/SRR10905396) and under BioProject accession number PRJNA601792. SRA accession numbers and information on all included samples can be found in Table S1 in the supplemental material. A full, unfiltered network graph is provided in Data Set S1, with the corresponding node labels given in Data Set S2. Raw, unnormalized read counts are also provided in Table S3. All data analysis and figure generation were performed using the R programming language, with code available at <https://doi.org/10.5281/zenodo.3572870>.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

FIG S1, TIF file, 0.3 MB.

FIG S2, TIF file, 0.3 MB.

TABLE S1, XLSX file, 0.02 MB.

TABLE S2, XLSX file, 0.03 MB.

TABLE S3, XLSX file, 8.8 MB.

DATA SET S1, TXT file, 84.4 MB.

DATA SET S2, TXT file, 0.1 MB.

REFERENCES

- Heidelberg JF, Eisen JA, Nelson WC, Clayton RA, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Umamay L, Gill SR, Nelson KE, Read TD, Tettelin H, Richardson D, Ermolaeva MD, Vamathevan J, Bass S, Qin H, Dragoi I, Sellers P, McDonald L, Utterback T, Fleischmann RD, Nierman WC, White O, Salzberg SL, Smith HO, Colwell RR, Mekalanos JJ, Venter JC, Fraser CM. 2000. DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. *Nature* 406:477–483. <https://doi.org/10.1038/35020000>.
- Weill F-X, Domman D, Njamkepo E, Almesbahi AA, Naji M, Nasher SS, Rakesh A, Assiri AM, Sharma NC, Kariuki S, Pourshafie MR, Rauzier J, Abubakar A, Carter JY, Wamala JF, Seguin C, Bouchier C, Malliavin T, Bakshi B, Abulmaali HHN, Kumar D, Njoroge SM, Malik MR, Kiiru J, Luquero FJ, Azman AS, Ramamurthy T, Thomson NR, Quilici M-L. 2019. Genomic insights into the 2016–2017 cholera epidemic in Yemen. *Nature* 565:230–233. <https://doi.org/10.1038/s41586-018-0818-3>.
- Greig DR, Schaefer U, Octavia S, Hunter E, Chattaway MA, Dallman TJ, Jenkins C. 2018. Evaluation of whole-genome sequencing for identification and typing of *Vibrio cholerae*. *J Clin Microbiol* 56:e00831-18. <https://doi.org/10.1128/JCM.00831-18>.
- Domman D, Chowdhury F, Khan AI, Dorman MJ, Mutreja A, Uddin MI, Paul A, Begum YA, Charles RC, Calderwood SB, Bhuiyan TR, Harris JB, LaRocque RC, Ryan ET, Qadri F, Thomson NR. 2018. Defining endemic cholera at three levels of spatiotemporal resolution within Bangladesh. *Nat Genet* 50:951–955. <https://doi.org/10.1038/s41588-018-0150-8>.
- Weill F-X, Domman D, Njamkepo E, Tarr C, Rauzier J, Fawal N, Keddy KH, Salje H, Moore S, Mukhopadhyay AK, Bercion R, Luquero FJ, Ngandjio A, Dosso M, Monakhova E, Garin B, Bouchier C, Pazzani C, Mutreja A, Grunow R, Sidikou F, Bonte L, Breurec S, Damian M, Njanpop-Lafourcade B-M, Sapriel G, Page A-L, Hamze M, Henkens M, Chowdhury G, Mengel M, Koeck J-L, Fournier J-M, Dougan G, Grimont PAD, Parkhill J, Holt KE, Piarroux R, Ramamurthy T, Quilici M-L, Thomson NR. 2017. Genomic history of the seventh pandemic of cholera in Africa. *Science* 358:785–789. <https://doi.org/10.1126/science.aad5901>.
- Domman D, Quilici M-L, Dorman MJ, Njamkepo E, Mutreja A, Mather AE,

- Delgado G, Morales-Espinosa R, Grimont PAD, Lizárraga-Partida ML, Bouchier C, Aanensen DM, Kuri-Morales P, Tarr CL, Dougan G, Parkhill J, Campos J, Cravioto A, Weill F-X, Thomson NR. 2017. Integrated view of *Vibrio cholerae* in the Americas. *Science* 358:789–793. <https://doi.org/10.1126/science.aao2136>.
7. Li Z, Pang B, Wang D, Li J, Xu J, Fang Y, Lu X, Kan B. 2019. Expanding dynamics of the virulence-related gene variations in the toxigenic *Vibrio cholerae* serogroup O1. *BMC Genomics* 20:360. <https://doi.org/10.1186/s12864-019-5725-y>.
 8. Rahman MH, Biswas K, Hossain MA, Sack RB, Mekalanos JJ, Faruque SM. 2008. Distribution of genes for virulence and ecological fitness among diverse *Vibrio cholerae* population in a cholera endemic area: tracking the evolution of pathogenic strains. *DNA Cell Biol* 27:347–355. <https://doi.org/10.1089/dna.2008.0737>.
 9. Lessler J, Moore SM, Luquero FJ, McKay HS, Grais R, Henkens M, Mengel M, Dunoyer J, M'bangombe M, Lee EC, Djingarey MH, Sudre B, Bompangue D, Fraser RSM, Abubakar A, Perea W, Legros D, Azman AS. 2018. Mapping the burden of cholera in sub-Saharan Africa and implications for control: an analysis of data across geographical scales. *Lancet* 391:1908–1915. [https://doi.org/10.1016/S0140-6736\(17\)33050-7](https://doi.org/10.1016/S0140-6736(17)33050-7).
 10. Global Task Force on Cholera Control. 2017. Ending cholera: a global roadmap to 2030. World Health Organization, Geneva, Switzerland.
 11. Herzog R, Peschek N, Fröhlich KS, Schumacher K, Papenfort K. 2019. Three autoinducer molecules act in concert to control virulence gene expression in *Vibrio cholerae*. *Nucleic Acids Res* 47:3171–3183. <https://doi.org/10.1093/nar/gky1320>.
 12. Davies BW, Bogard RW, Young TS, Mekalanos JJ. 2012. Coordinated regulation of accessory genetic elements produces cyclic di-nucleotides for *V. cholerae* virulence. *Cell* 149:358–370. <https://doi.org/10.1016/j.cell.2012.01.053>.
 13. Krin E, Pierlé SA, Sismeiro O, Jagla B, Dillies M-A, Varet H, Irazoki O, Campoy S, Rouy Z, Cruveiller S, Médigue C, Coppée J-Y, Mazel D. 2018. Expansion of the SOS regulon of *Vibrio cholerae* through extensive transcriptome analysis and experimental validation. *BMC Genomics* 19:373. <https://doi.org/10.1186/s12864-018-4716-8>.
 14. Fu Y, Waldor MK, Mekalanos JJ. 2013. Tn-Seq analysis of *Vibrio cholerae* intestinal colonization reveals a role for T6SS-mediated antibacterial activity in the host. *Cell Host Microbe* 14:652–663. <https://doi.org/10.1016/j.chom.2013.11.001>.
 15. Mandlik A, Livny J, Robins WP, Ritchie JM, Mekalanos JJ, Waldor MK. 2011. RNA-Seq-based monitoring of infection-linked changes in *Vibrio cholerae* gene expression. *Cell Host Microbe* 10:165–174. <https://doi.org/10.1016/j.chom.2011.07.007>.
 16. Kamp HD, Patimalla-Dipali B, Lazinski DW, Wallace-Gadsden F, Camilli A. 2013. Gene fitness landscapes of *Vibrio cholerae* at important stages of its life cycle. *PLoS Pathog* 9:e1003800. <https://doi.org/10.1371/journal.ppat.1003800>.
 17. Pukatzki S, Ma AT, Sturtevant D, Krastins B, Sarracino D, Nelson WC, Heidelberg JF, Mekalanos JJ. 2006. Identification of a conserved bacterial protein secretion system in *Vibrio cholerae* using the Dictyostelium host model system. *Proc Natl Acad Sci U S A* 103:1528–1533. <https://doi.org/10.1073/pnas.0510322103>.
 18. Kimura S, Hubbard TP, Davis BM, Waldor MK. 2016. The nucleoid binding protein H-NS biases genome-wide transposon insertion landscapes. *mBio* 7:e01351-16. <https://doi.org/10.1128/mBio.01351-16>.
 19. Manneh-Roussel J, Haycocks JRJ, Magán A, Perez-Soto N, Voelz K, Camilli A, Krachler A-M, Grainger DC. 2018. cAMP receptor protein controls *Vibrio cholerae* gene expression in response to host colonization. *mBio* 9:e00966-18. <https://doi.org/10.1128/mBio.00966-18>.
 20. Saelens W, Cannoodt R, Saeys Y. 2018. A comprehensive evaluation of module detection methods for gene expression data. *Nat Commun* 9:1090. <https://doi.org/10.1038/s41467-018-03424-4>.
 21. Langfelder P, Horvath S. 2008. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9:559. <https://doi.org/10.1186/1471-2105-9-559>.
 22. Jiang J, Sun X, Wu W, Li L, Wu H, Zhang L, Yu G, Li Y. 2016. Construction and application of a co-expression network in *Mycobacterium tuberculosis*. *Sci Rep* 6:28422. <https://doi.org/10.1038/srep28422>.
 23. Liu W, Li L, Long X, You W, Zhong Y, Wang M, Tao H, Lin S, He H. 2018. Construction and analysis of gene co-expression networks in *Escherichia coli*. *Cells* 7:19. <https://doi.org/10.3390/cells7030019>.
 24. Li A, Horvath S. 2007. Network neighborhood analysis with the multi-node topological overlap measure. *Bioinformatics* 23:222–231. <https://doi.org/10.1093/bioinformatics/btl581>.
 25. Geer LY, Marchler-Bauer A, Geer RC, Han L, He J, He S, Liu C, Shi W, Bryant SH. 2010. The NCBI BioSystems database. *Nucleic Acids Res* 38:D492–D496. <https://doi.org/10.1093/nar/gkp858>.
 26. Huang DW, Sherman BT, Lempicki RA. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 4:44–57. <https://doi.org/10.1038/nprot.2008.211>.
 27. Kanehisa M, Goto S. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 28:27–30. <https://doi.org/10.1093/nar/28.1.27>.
 28. Mi H, Dong Q, Muruganujan A, Gaudet P, Lewis S, Thomas PD. 2010. PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. *Nucleic Acids Res* 38:D204–D210. <https://doi.org/10.1093/nar/gkp1019>.
 29. Taboada B, Estrada K, Ciria R, Merino E. 2018. Operon-mapper: a web server for precise operon identification in bacterial and archaeal genomes. *Bioinformatics* 34:4118–4120. <https://doi.org/10.1093/bioinformatics/bty496>.
 30. Chao MC, Pritchard JR, Zhang YJ, Rubin EJ, Livny J, Davis BM, Waldor MK. 2013. High-resolution definition of the *Vibrio cholerae* essential gene set with hidden Markov model-based analyses of transposon-insertion sequencing data. *Nucleic Acids Res* 41:9033–9048. <https://doi.org/10.1093/nar/gkt654>.
 31. The Gene Ontology Consortium. 2019. The Gene Ontology resource: 20 years and still GOing strong. *Nucleic Acids Res* 47:D330–D338. <https://doi.org/10.1093/nar/gky1055>.
 32. Rutherford ST, van Kessel JC, Shao Y, Bassler BL. 2011. AphA and LuxR/HapR reciprocally control quorum sensing in vibrios. *Genes Dev* 25:397–408. <https://doi.org/10.1101/gad.2015011>.
 33. Haycocks JRJ, Warren GZL, Walker LM, Chlebek JL, Dalia TN, Dalia AB, Grainger DC. 2019. The quorum sensing transcription factor AphA directly regulates natural competence in *Vibrio cholerae*. *PLoS Genet* 15:e1008362. <https://doi.org/10.1371/journal.pgen.1008362>.
 34. Zhang Y, Gao H, Osei-Adjei G, Zhang Y, Yang W, Yang H, Yin Z, Huang X, Zhou D. 2017. Transcriptional regulation of the type VI secretion system 1 genes by quorum sensing and ToxR in *Vibrio parahaemolyticus*. *Front Microbiol* 8:2005. <https://doi.org/10.3389/fmicb.2017.02005>.
 35. Yamamoto S, Mitobe J, Ishikawa T, Wai SN, Ohnishi M, Watanabe H, Izumiya H. 2014. Regulation of natural competence by the orphan two-component system sensor kinase ChiS involves a non-canonical transmembrane regulator in *Vibrio cholerae*. *Mol Microbiol* 91:326–347. <https://doi.org/10.1111/mmi.12462>.
 36. Dalia AB, Lazinski DW, Camilli A. 2014. Identification of a membrane-bound transcriptional regulator that links chitin and natural competence in *Vibrio cholerae*. *mBio* 5:e01028-13. <https://doi.org/10.1128/mBio.01028-13>.
 37. Metzger LC, Stutzmann S, Scrignari T, Van der Henst C, Matthey N, Blokesch M. 2016. Independent regulation of type VI secretion in *Vibrio cholerae* by TfoX and TfoY. *Cell Rep* 15:951–958. <https://doi.org/10.1016/j.celrep.2016.03.092>.
 38. van Dam S, Vósa U, van der Graaf A, Franke L, de Magalhães JP. 2017. Gene co-expression analysis for functional classification and gene–disease predictions. *Brief Bioinform* 19:575–592.
 39. Silva AJ, Benitez JA. 2016. *Vibrio cholerae* biofilms and cholera pathogenesis. *PLoS Negl Trop Dis* 10:e0004330. <https://doi.org/10.1371/journal.pntd.0004330>.
 40. Teschler JK, Zamorano-Sánchez D, Utada AS, Warner CJA, Wong GCL, Linington RG, Yildiz FH. 2015. Living in the matrix: assembly and control of *Vibrio cholerae* biofilms. *Nat Rev Microbiol* 13:255–268. <https://doi.org/10.1038/nrmicro3433>.
 41. Papenfort K, Förstner KU, Cong J-P, Sharma CM, Bassler BL. 2015. Differential RNA-seq of *Vibrio cholerae* identifies the VqmR small RNA as a regulator of biofilm formation. *Proc Natl Acad Sci U S A* 112:E766–E775. <https://doi.org/10.1073/pnas.1500203112>.
 42. Weber GG, Klose KE. 2011. The complexity of ToxT-dependent transcription in *Vibrio cholerae*. *Indian J Med Res* 133:201–206.
 43. Kazi MI, Conrado AR, Mey AR, Payne SM, Davies BW. 2016. ToxR antagonizes H-NS regulation of horizontally acquired genes to drive host colonization. *PLoS Pathog* 12:e1005570. <https://doi.org/10.1371/journal.ppat.1005570>.
 44. Dorman MJ, Dorman CJ. 2018. Regulatory hierarchies controlling virulence gene expression in *Shigella flexneri* and *Vibrio cholerae*. *Front Microbiol* 9:2686. <https://doi.org/10.3389/fmicb.2018.02686>.
 45. Ayala JC, Wang H, Silva AJ, Benitez JA. 2015. Repression by H-NS of genes required for the biosynthesis of the *Vibrio cholerae* biofilm matrix is modulated by the second messenger cyclic diguanylic acid. *Mol Microbiol* 97:630–645. <https://doi.org/10.1111/mmi.13058>.

46. Boyd EF, Jermyn WS, Boyd EF. 2002. Characterization of a novel *Vibrio* pathogenicity island (VPI-2) encoding neuraminidase (nanH) among toxigenic *Vibrio cholerae* isolates. *Microbiology* 148:3681–3693. <https://doi.org/10.1099/00221287-148-11-3681>.
47. Karaolis DKR, Johnson JA, Bailey CC, Boedeker EC, Kaper JB, Reeves PR. 1998. A *Vibrio cholerae* pathogenicity island associated with epidemic and pandemic strains. *Proc Natl Acad Sci U S A* 95:3134–3139. <https://doi.org/10.1073/pnas.95.6.3134>.
48. Nielsen AT, Dolganov NA, Otto G, Miller MC, Wu CY, Schoolnik GK. 2006. RpoS controls the *Vibrio cholerae* mucosal escape response. *PLoS Pathog* 2:e109. <https://doi.org/10.1371/journal.ppat.0020109>.
49. Wong GT, Bonocora RP, Schep AN, Beeler SM, Lee Fong AJ, Shull LM, Batachari LE, Dillon M, Evans C, Becker CJ, Bush EC, Hardin J, Wade JT, Stoebel DM. 2017. Genome-wide transcriptional response to varying RpoS levels in *Escherichia coli* K-12. *J Bacteriol* 199:e00755-16. <https://doi.org/10.1128/JB.00755-16>.
50. Russo G, Zegar C, Giordano A. 2003. Advantages and limitations of microarray technology in human cancer. *Oncogene* 22:6497–6507. <https://doi.org/10.1038/sj.onc.1206865>.
51. Serin EAR, Nijveen H, Hilhorst HWM, Ligterink W. 2016. Learning from co-expression networks: possibilities and challenges. *Front Plant Sci* 7:444. <https://doi.org/10.3389/fpls.2016.00444>.
52. Koschmann J, Bhar A, Stegmaier P, Kel AE, Wingender E. 2015. Upstream Analysis: an integrated promoter-pathway analysis approach to causal interpretation of microarray data. *Microarrays (Basel)* 4:270–286. <https://doi.org/10.3390/microarrays4020270>.
53. Mueller RS, McDougald D, Cusumano D, Sodhi N, Kjelleberg S, Azam F, Bartlett DH. 2007. *Vibrio cholerae* strains possess multiple strategies for abiotic and biotic surface colonization. *J Bacteriol* 189:5348–5360. <https://doi.org/10.1128/JB.01867-06>.
54. Marchler-Bauer A, Bo Y, Han L, He J, Lanczycki CJ, Lu S, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, Gwadz M, Hurwitz DL, Lu F, Marchler GH, Song JS, Thanki N, Wang Z, Yamashita RA, Zhang D, Zheng C, Geer LY, Bryant SH. 2017. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res* 45:D200–D203. <https://doi.org/10.1093/nar/gkw1129>.
55. Duran-Pinedo AE, Paster B, Teles R, Frias-Lopez J. 2011. Correlation network analysis applied to complex biofilm communities. *PLoS One* 6:e28438. <https://doi.org/10.1371/journal.pone.0028438>.
56. Geng H, Tran-Gyamfi MB, Lane TW, Sale KL, Yu ET. 2016. Changes in the structure of the microbial community associated with *Nannochloropsis salina* following treatments with antibiotics and bioactive compounds. *Front Microbiol* 7:1155. <https://doi.org/10.3389/fmicb.2016.01155>.
57. Meisel JS, Sfyroera G, Bartow-McKenney C, Gimblet C, Bugayev J, Horwinski J, Kim B, Brestoff JR, Tyldsley AS, Zheng Q, Hodgkinson BP, Artis D, Grice EA. 2018. Commensal microbiota modulate gene expression in the skin. *Microbiome* 6:20. <https://doi.org/10.1186/s40168-018-0404-9>.
58. Jackson MA, Bonder MJ, Kuncheva Z, Zierer J, Fu J, Kurilshikov A, Wijmenga C, Zhernakova A, Bell JT, Spector TD, Steves CJ. 2018. Detection of stable community structures within gut microbiota co-occurrence networks from different human populations. *PeerJ* 6:e4303. <https://doi.org/10.7717/peerj.4303>.
59. Hosseinkhan N, Mousavian Z, Masoudi-Nejad A. 2018. Comparison of gene co-expression networks in *Pseudomonas aeruginosa* and *Staphylococcus aureus* reveals conservation in some aspects of virulence. *Gene* 639:1–10. <https://doi.org/10.1016/j.gene.2017.10.005>.
60. Peña-Castillo L, Mercer RG, Gurinovich A, Callister SJ, Wright AT, Westbye AB, Beatty JT, Lang AS. 2014. Gene co-expression network analysis in *Rhodobacter capsulatus* and application to comparative expression analysis of *Rhodobacter sphaeroides*. *BMC Genomics* 15:730. <https://doi.org/10.1186/1471-2164-15-730>.
61. Wang J, Wu G, Chen L, Zhang W. 2013. Cross-species transcriptional network analysis reveals conservation and variation in response to metal stress in cyanobacteria. *BMC Genomics* 14:112. <https://doi.org/10.1186/1471-2164-14-112>.
62. Keseler IM, Mackie A, Santos-Zavaleta A, Billington R, Bonavides-Martínez C, Caspi R, Fulcher C, Gama-Castro S, Kothari A, Krummenacker M, Latendresse M, Muñoz-Rascado L, Ong Q, Paley S, Peralta-Gil M, Subhraveti P, Velázquez-Ramírez DA, Weaver D, Collado-Vides J, Paulsen I, Karp PD. 2017. The EcoCyc database: reflecting new knowledge about *Escherichia coli* K-12. *Nucleic Acids Res* 45:D543–D550. <https://doi.org/10.1093/nar/gkw1003>.
63. Ballouz S, Verleyen W, Gillis J. 2015. Guidance for RNA-seq co-expression network construction and analysis: safety in numbers. *Bioinformatics* 31:2123–2130. <https://doi.org/10.1093/bioinformatics/btv118>.
64. Li S, Tighe SW, Nicolet CM, Grove D, Levy S, Farmerie W, Viale A, Wright C, Schweitzer PA, Gao Y, Kim D, Boland J, Hicks B, Kim R, Chhangawala S, Jafari N, Raghavachari N, Gandara J, Garcia-Reyero N, Hendrickson C, Roberson D, Rosenfeld JA, Smith T, Underwood JG, Wang M, Zumbo P, Baldwin DA, Grills GS, Mason CE. 2014. Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. *Nat Biotechnol* 32:915–925. <https://doi.org/10.1038/nbt.2972>.
65. Bin Goh WW, Wang W, Wong L. 2017. Why batch effects matter in omics data, and how to avoid them. *Trends Biotechnol* 35:498–507. <https://doi.org/10.1016/j.tibtech.2017.02.012>.
66. Leek JT, Storey JD. 2007. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet* 3:1724–1735. <https://doi.org/10.1371/journal.pgen.0030161>.
67. Alter O, Brown PO, Botstein D. 2000. Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci U S A* 97:10101–10106. <https://doi.org/10.1073/pnas.97.18.10101>.
68. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA. 2010. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet* 11:733–739. <https://doi.org/10.1038/nrg2825>.
69. Cohen O, Doron S, Wurtzel O, Dar D, Edelheit S, Karunker I, Mick E, Sorek R. 2016. Comparative transcriptomics across the prokaryotic tree of life. *Nucleic Acids Res* 44:W46–W53. <https://doi.org/10.1093/nar/gkw394>.
70. Chang YM, Lin HH, Liu WY, Yu CP, Chen HJ, Wartini PP, Kao YY, Wu YH, Lin JJ, Lu MYJ, Tu SL, Wu SH, Shiu SH, Ku MSB, Li WH. 2019. Comparative transcriptomics method to infer gene coexpression networks and its applications to maize and rice leaf transcriptomes. *Proc Natl Acad Sci U S A* 116:3091–3099. <https://doi.org/10.1073/pnas.1817621116>.
71. Rodríguez-García A, Sola-Landa A, Barreiro C. 2017. RNA-Seq-based comparative transcriptomics: RNA preparation and bioinformatics. *Methods Mol Biol* 1645:59–72. https://doi.org/10.1007/978-1-4939-7183-1_5.
72. Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database Collaboration. 2011. The Sequence Read Archive. *Nucleic Acids Res* 39:D19–D21. <https://doi.org/10.1093/nar/gkq1019>.
73. Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* 34:525–527. <https://doi.org/10.1038/nbt.3519>.
74. Langmead B, Salzberg SL. 2012. Bowtie 2. *Nat Methods* 9:357–359. <https://doi.org/10.1038/nmeth.1923>.
75. Gaspar JM. 2018. Improved peak-calling with MACS2. *bioRxiv* <https://doi.org/10.1101/496521>.
76. Ross-Innes CS, Stark R, Teschendorff AE, Holmes KA, Ali HR, Dunning MJ, Brown GD, Gojis O, Ellis IO, Green AR, Ali S, Chin S-F, Palmieri C, Caldas C, Carroll JS. 2012. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature* 481:389–393. <https://doi.org/10.1038/nature10730>.
77. Sonesson C, Love MI, Robinson MD. 2015. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res* 4:1521. <https://doi.org/10.12688/f1000research.7563.2>.
78. Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15:550. <https://doi.org/10.1186/s13059-014-0550-8>.