A comprehensive co-expression network analysis in *Vibrio cholerae*

Cory D. DuPai[a], Claus O. Wilke[a,b], and Bryan W. Davies[a,c]

[a] Institute for Cellular and Molecular Biology, The University of Texas at Austin, Austin, TX, USA

[b] Department of Integrative Biology, The University of Texas at Austin, Austin, TX, USA

[c] Department of Molecular Biosciences, The University of Texas at Austin, Austin, TX, 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 datasets now provides a unique opportunity for *in silico* hypothesis generation via co-expression analysis. Here we leverage all published *V. cholerae* RNA-sequencing data, in combination with select data from other platforms, to generate a gene co-expression 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. Tn-seq and ChIP-seq), and expands upon micro-array 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 this importance, roughly one third of *V. cholerae* genes are functionally un-annotated, leaving large gaps in our understanding of this microbe. Through co-expression 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.

Introduction

Since the completion of the first *Vibrio cholerae* genome sequence in 2000, over a thousand *V. cholerae* isolates have been sequenced (1, 2). These sequences has 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 phylogenic 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 complex virulence pathways, illuminating a novel interbacterial defense system, and expanding our knowledge of the role of non-coding RNA (ncRNA) 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 datasets. Expression data is particularly amenable to such pooling and can be used to accurately group genes into functional modules based on their co-expression (20). In bacteria, weighted gene co-expression 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 datasets to provide novel biological insights via existing data.

Despite major advances in sequencing technologies and research strategies, most of the over two dozen existing RNA-seq experiments in *V. cholerae* have been limited to targeted approaches that involve quantifying the differential abundance of genetic material across a handful of conditions. Via these approaches, any change in expression observed in one experiment is nearly impossible to generalize 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 entire genome by combining information from multiple datasets. As there is no existing co-expression analysis in *V. cholerae* to date, the accumulation of over 300 publicly available RNA-seq

70 samples from targeted RNA-seq experiments represents a heretofore untapped resource
71 for the cholera community.

72 Motivated by the success of pooled genetic sequencing analyses, our current work
73 utilizes all publicly available *V. cholerae* RNA-seq based expression-level data to generate a
74 co-expression network. We expand upon existing bacterial WGCNA approaches by
75 integrating broader sequencing data (including ChIP-seq and Tn-seq) and multiple
76 annotation platforms into our analysis. Our network ultimately contributes information on
77 connections across all *V. cholerae* genes, including the roughly 1500 predicted but
78 functionally un-annotated genetic elements that account for some 37% of the genome.
79 More specifically, we implicate new loci in virulence regulation and clearly demonstrate a
80 powerful and accurate approach to hypothesis generation via our described network.
81

82 **Results**

83 **Gene network generation**

84 To generate our co-expression analysis in *V. cholerae,* we applied our WGCNA pipeline to
85 analyze twenty-seven *V. cholerae* RNA sequencing experiments deposited in NCBI's
86 Sequence Read Archive (SRA) in addition to two novel experiments. The RNA sequencing
87 samples are derived from experiments exploring a range of important *V. cholerae* processes
88 including intestinal colonization, quorum sensing, and stress response. In total, our
89 network includes 300 individual RNA-seq samples (supplementary table S1). All samples
90 were mapped to a recently inferred *V. cholerae* transcriptome derived from the N16961
91 reference genome (1, 13). This reference was chosen because the majority (293) of
92 samples were collected from strains N16961 or the closely related C6706 and A1552.

Figure 1 outlines the process used to generate our co-expression network with a small subset of genes. Loci VC0384–VC0386 are known to be involved in cysteine metabolism while the two genomically adjacent loci VC0383 and VC0388 do not share this function. Following normalization of mapped transcripts (Fig. 1A), a weighted gene co-expression network analysis was performed using WGCNA (21). First, a Pearson correlation matrix is calculated for expression levels of all genes (Fig. 1B). This correlation matrix clearly captures strong relationships between co-expressing genes such as VC0384–VC0386 but can produce background noise from un-related gene pairs. We limit this noise by calculating a topological overlap matrix (TOM) (24) that weights pairwise co-expression data based on each gene's interactions with all other genes (Fig. 1C). In this way, the relationships between genes that fall within the same subnetwork, i.e. VC0384-86, are favored while the signal from unrelated genes, i.e. VC0383 and VC0388, is abated. This TOM, after filtering for normalized values greater than 0.1, is used to construct an accurate co-expression network that captures biologically meaningful relationships (Fig. 1D).

In addition to co-expression data, our network and analyses incorporate 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). Additionally, importance labels are 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, 29). 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 co-

116 expressing genes that provides both qualitative and quantitative information about
117 relationships between transcripts across forty-nine gene-clusters covering the entire *V.*
118 *cholerae* genome (Supp. Data S1-2).
119
120 **Genes in known pathways cluster together and contextualize genes of unknown**
121 **function**
122 As proof of the accuracy of our approach, we have highlighted four clusters that
123 recapitulate known interactions between transcripts involved in specific pathways or
124 cellular processes (Fig. 2). The correct grouping of transcripts encoding products such as
125 ribosomal proteins, amino acid synthesis proteins, and tRNA transcripts that have largely
126 known functions and are involved in well-studied, highly conserved cellular processes
127 provides a positive control for the validity of our network clusters (Fig. 2A–C). Likewise,
128 the clustering of genes known to be involved in more specialized processes such as biofilm
129 formation (Fig. 2D) further underscores the success of our approach.
130 The subnetworks mentioned above also support the utility of our analysis in powering
131 guilt-by-association based inference of gene function (30). Because each of these gene
132 clusters contain co-expressing genes that are involved in the same biological process, it can
133 be assumed that unannotated genes in the same cluster are likely involved in the same
134 process. Such links, while not definitive on their own, can be used with other data to hint at
135 gene functions. For example, genes with known function in Fig. 2D are primarily involved
136 in biofilm formation (31, 32). This clustering of biofilm genes suggests that the few genes
137 with no known function in this subnetwork may be involved in the same process. Two of
138 these unannotated transcripts, VC1937 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 (33). 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 co-expression network serves as a highly efficient substitute for more traditional screening assays.

A virulence subnetwork suggests novel gene functions

While the biofilm associated subnetwork presents a relatively simple example of the functional insights our co-expression data can yield, the virulence-related subnetwork (Fig. 3A) 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 the *V. cholerae* pathogenicity island 1 (VC0809–VC0848) and cholera toxin sub-units A and B (*ctxAB*, VC1456 and VC1457) (34). 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 regulators ToxR, ToxT, or H-NS as determined via ChIP and/or RNA-seq (35–37). The clustering of such genes with well-characterized interactions into a cohesive subnetwork is further validation of our ability to generate accurate co-expression maps of related genes. The association of uncharacterized genes in these clusters suggests 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 co-express with known virulence genes because they fall within vibrio pathogenicity island (VPI)-1 (VC0810, VC0821–VC0823, VC0842) or VPI-2 (VC1806, VC1810), or are proximal to other virulence genes (VC1945) (38, 39). However, our analysis also identified genes such as VCA0094–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 and Tn-seq into our co-expression analysis, allowing us to verify the association between VCA0094–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 co-express with genes regulated by ToxT, ToxR, and H-NS, we also probed existing ChIPseq binding datasets (12, 19, 35) to see if any of these well-studied transcription factors bind near the VCA0094-96 loci. 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 re-analysis of existing binding data from (35). Both peaks showed a large and significant increase in binding affinity ($\log_2$ fold change in average occupancy) when compared against input controls (Fig. 3B). H-NS showed a clear binding peak in the region of the VCA0094 promoter that extended in a diffuse manner to the VCA0095 TSS 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 VCA0094–VCA0096 transcripts. Although the exact mechanistic role of these genes remains elusive, we have nevertheless demonstrated

185 the ability of our pipeline to generate meaningful hypotheses by incorporating existing data
186 from a multitude of sources.
187
188 **Co-expression data provides an accurate *in silico* complement to RNA-seq**
189 In addition to the guilt-by-association inference described above, co-expression analysis
190 can provide a partial substitute or complement to RNA-seq experiments. Novel, meaningful
191 genetic relationships can be found in a co-expression network by focusing on the
192 transcripts that are co-regulated with a gene of interest.
193 We can apply a network-based approach in lieu of new RNA-seq based experiments to
194 identify genes which co-express with *rpoS* (VC0534) and are similarly involved in bacterial
195 stress response. As our network utilizes only RNA-seq based transcriptomics studies and
196 none of these studies involves direct manipulation of *rpoS*, we can compare existing
197 microarray data involving an *rpoS* (VC0534) deletion mutant (40) to determine how
198 accurate our approach is. When applying an absolute co-expression cutoff of 0.1, 273 genes
199 are identified as having a relationship with *rpoS* expression in both our network analysis
200 and the *rpoS* mutant microarray data (Fig. 4A). This represents nearly two-thirds of genes
201 identified as differentially expressed in the original microarray study. While our network
202 links far more genes with *rpoS* than the microarray approach, this is in line with recent
203 RNA-seq based work that found that 23% of the E. coli genome is regulated by RpoS (41).
204 Additionally, all of the flagella and chemotaxis related proteins highlighted as particularly
205 informative in the original study are identified by our analysis (Fig. 4B) and relevant values
206 (i.e. network co-expression and microarray-derived log fold change in expression) for the
207 273 shared transcripts have a Spearman correlation of -0.516. This accuracy is achieved

208 without any direct genetic manipulation of the *rpoS* locus in the RNA-seq datasets used to
209 generate our co-expression network and serves as a testament to the potential utility and
210 versatility of our approach.

211 Our approach to isolating genetic interactions also has advantages over
212 transcriptomics-focused sequencing. As seen in Fig. 4A, our network-based analysis
213 identifies far more genes associated with *rpoS*. This is likely because RNAseq-based
214 approaches are can identify a broader range of gene transcripts as they are not limited by
215 restrictive microarray probes (42). Separate from differences in underlying technology, co-
216 expression networks are also more likely to detect genes regulating a target's expression
217 than traditional transcriptomics experiments which largely capture downstream responses
218 to changes in a target's expression (43, 44). Thus, a co-expression network can provide an
219 alternative perspective to complement or clarify transcriptomics data.

220

221 **Discussion**

222 We have successfully constructed the first *V. cholerae* co-expression network through a
223 computationally inexpensive process that is simple, easily expanded upon, and
224 straightforward to implement in other organisms. Our network effectively identifies
225 canonical gene clusters related to specific molecular pathways or functions, such as those
226 corresponding to rRNAs or biofilm proteins. We have also outlined two use-cases for the
227 data provided and have shown the accuracy of both approaches either experimentally or
228 using existing data. Additionally, we have included relevant network files as well as raw
229 read counts across RNA-seq conditions (Supp. Data S1-2 & Supp. Table S2) alongside all

code used in our analysis (see Materials and Methods) to encourage broad usage of this data.

Our results have proven both the utility and accuracy of our approach despite in-depth analysis limited to a handful of genes across five of the forty-nine observed gene clusters. Furthermore, our work with the virulence subnetwork supports previously published research loosely implicating genes VCA0094–96 in virulence and virulence related functions. All three transcripts have shown up in screens focusing on biofilm development (45), and SOS response (13). From a mechanistic perspective, protein homology analysis via NCBI's Conserved Domain Database (46) 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, co-expression data is 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 co-expression 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 bacterial-focus, microarray-based co-expression networks and the lack of clear functional knowledge for a large portion of *V. cholerae* genes. Besides more simple guilt-by-association studies

(22, 23), co-expression networks have helped to elucidate relationships in diverse microbial communities (47–50) and enable comparisons across strains and species (51–53). These works as well as the relative dearth of knowledge about the *V. cholerae* genome (roughly two third of genes are annotated compared to around 86% percent of all *E. coli* genes (54)) and the growing abundance of *V. cholerae* focused NGS data served as the impetus for this research.

The calculated co-expression network, though accurate, could be improved via the inclusion of more experiments and more extensive SRA annotations. Our somewhat limited pooled dataset consisting of three hundred samples is an order of magnitude off from the few thousand samples necessary to derive the most faithful co-expression estimates (55). 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 (56, 57). 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 datasets from single experiments. While automated sample clustering methods (58–60) 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. Likewise, manual curation of batch annotations is also difficult since few SRA records are extensively annotated with detailed experimental conditions (e.g. bacterial growth stage, 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

276 similarity of the vast majority of included strains to our reference, a more elaborate
277 comparative transcriptomic strategy (61, 62) would be ideal if more diverse samples are
278 included in future analyses. This is especially true when considering the inclusion of
279 expression data from clinical samples which are likely to have much more genomic
280 variability than the closely related lab cultured strains used to construct our network. On
281 the other hand, because comparative transcriptomics requires defining homologous alleles
282 across all strains analyzed (63), such an approach would greatly increase the difficulty of
283 incorporating strains without an assembled genome.
284 In summary, our co-expression network can drive functional hypotheses for
285 unannotated genes in *V. cholerae*. As the Vibrio community steadily adds high quality data
286 from increasingly sophisticated sequencing experiments to public databases our imputed
287 network can only improve, providing ever deeper insights into the *V. cholerae* genome. At
288 the same time, highly annotated transcript-based co-expression networks can empower
289 research with related technologies (e.g. single cell transcriptomics and dual RNA-seq) and
290 research into a host of other clinically relevant bacteria, such as *Pseudomonas aeruginosa*
291 or *Staphylococcus aureus* which have over 2000 and 1400 RNA-seq experiments in SRA
292 respectively.
293

294 **Materials and Methods**

295 **Data Collection and Processing**

296 All RNA and ChIP sequencing data were downloaded from the Sequence Read Archive
297 (SRA)(64) and converted to compressed fastq files using the SRA toolkit
298 (https://www.ncbi.nlm.nih.gov/sra/docs/toolkitsoft/) (see Table S1 for details on

included experiments). RNA-seq samples were selected by searching the SRA on Sept 10th, 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 (PRJNA601792). Samples were mapped to a recently inferred *V. cholerae* transcriptome derived from the N16961 reference genome (1, 13) using Kallisto version 0.45.1 (65). This reference was chosen because the majority (293) of samples were collected from strains N16961 or the closely related C6706 and A1552. 26 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, 35) and sequences were downloaded from SRA and converted to fastq files as above. Raw reads were mapped to the same N16961 reference genome using Bowtie 2 version 2.3.5.1 (66). From this mapping, peaks were identified using MACS2 version 2.1.2 with an extsize of 225 (various sizes from 150 to 500 were tested with little observable difference in peaks identified) (67) and differential binding and significance were calculated using DiffBind version 2.12.0 (68).

Processed Tn-seq data were collected directly from published datasets. *In vitro* essentiality and semi-essentiality labels were derived from Chao et al. 2013 Table S1 (29), with the original labels of domain essential and sick genes replaced with essential and semi-essential respectively. We used Table S2 from Fu, Waldor, and Mekalanos 2013 (14) to label genes involved in host infection, with any gene exhibiting a $\log_2$ fold change less than negative three deemed essential and any gene with a $\log_2$ fold change between negative one and negative three deemed semi-essential.

Network Construction

Figure 1 highlights the process used to generate our co-expression network. Kallisto derived reads were first imported into R via tximport (69), then normalized using DESeq2 version 1.24.0 (70), resulting in values that are comparable across conditions and experiments. Following normalization, a weighted gene co-expression network analysis was performed using WGCNA (21). This process is highlighted with a subset of data in Figure 1 and consists of the sequential calculation of a Pearson correlation matrix, adjacency matrix with power ß=6, and, ultimately, topological overlap matrix (TOM) (24) from normalized gene expression counts across conditions. We further filtered this TOM to exclude samples with weighted co-expression <0.1 for all analysis included in the Results section.

Predicted pathway annotations and gene functional knowledge are derived from the NCBI Biosystems database as well as DAVID, Panther, and KEGG databases (25–28). Genes lacking functional knowledge which are identified as essential or semi-essential in either Tn-seq dataset are labeled in network visualizations as “important.”

Data Availability

SRA accession numbers and information on included samples can be found in Supplementary Table S1. A full, unfiltered network graph is provided in Supplementary File S1 with the corresponding node labels in Supplementary File S2. Raw, un-normalized read counts are also provided in Supplementary Table S2 All data analysis and figure generation were done using the R programming language, with code available at DOI: 10.5281/zenodo.3572870.

References

1. Heidelberg JF, Eisen JA, Nelson WC, Clayton RA, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Umayam 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, Fleishmann 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.

2. 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, Bakhshi 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.

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.

4. 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.

5. 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 (80-) 358:785–789.

6. 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.

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.

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.

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 (London, England) 391:1908–1915.

10. 2017. WHO | Ending Cholera. WHO.

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.

392 12. Davies BW, Bogard RW, Young TS, Mekalanos JJ. 2012. Coordinated Regulation of
393 Accessory Genetic Elements Produces Cyclic Di-Nucleotides for V. cholerae Virulence.
394 Cell 149:358–370.
395 13. Krin E, Pierlé SA, Sismeiro O, Jagla B, Dillies M-A, Varet H, Irazoki O, Campoy S, Rouy Z,
396 Cruveiller S, Médigue C, Coppée J-Y, Mazel D. 2018. Expansion of the SOS regulon of
397 Vibrio cholerae through extensive transcriptome analysis and experimental validation.
398 BMC Genomics 19:373.
399 14. Fu Y, Waldor MK, Mekalanos JJ. 2013. Tn-Seq Analysis of Vibrio cholerae Intestinal
400 Colonization Reveals a Role for T6SS-Mediated Antibacterial Activity in the Host. Cell
401 Host Microbe 14:652–663.
402 15. Mandlik A, Livny J, Robins WP, Ritchie JM, Mekalanos JJ, Waldor MK. 2011. RNA-Seq-
403 based monitoring of infection-linked changes in Vibrio cholerae gene expression. Cell
404 Host Microbe 10:165–174.
405 16. Kamp HD, Patimalla-Dipali B, Lazinski DW, Wallace-Gadsden F, Camilli A. 2013. Gene
406 Fitness Landscapes of Vibrio cholerae at Important Stages of Its Life Cycle. PLoS Pathog
407 9:e1003800.
408 17. Pukatzki S, Ma AT, Sturtevant D, Krastins B, Sarracino D, Nelson WC, Heidelberg JF,
409 Mekalanos JJ. 2006. Identification of a conserved bacterial protein secretion system in
410 Vibrio cholerae using the Dictyostelium host model system. Proc Natl Acad Sci 103:1528
411 LP – 1533.
412 18. Kimura S, Hubbard TP, Davis BM, Waldor MK. 2016. The Nucleoid Binding Protein H-NS
413 Biases Genome-Wide Transposon Insertion Landscapes. MBio 7:e01351-16.
414 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.

20. Saelens W, Cannoodt R, Saeys Y. 2018. A comprehensive evaluation of module detection methods for gene expression data. Nat Commun 9:1090.

21. Langfelder P, Horvath S. 2008. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 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.

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.

24. Li A, Horvath S. 2006. Network neighborhood analysis with the multi-node topological overlap measure. Bioinformatics 23:222–231.

25. Geer LY, Marchler-Bauer A, Geer RC, Han L, He J, He S, Liu C, Shi W, Bryant SH. 2009. The NCBI BioSystems database. Nucleic Acids Res 38:D492–D496.

26. Huang DW, Sherman BT, Lempicki RA. 2008. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4:44.

27. Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28:27–30.

28. Mi H, Dong Q, Muruganujan A, Gaudet P, Lewis S, Thomas PD. 2009. PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. Nucleic Acids Res 38:D204–D210.

29. 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

438 model-based analyses of transposon-insertion sequencing data. Nucleic Acids Res
439 41:9033–9048.
440 30. van Dam S, Võsa U, van der Graaf A, Franke L, de Magalhães JP. 2017. Gene co-expression
441 analysis for functional classification and gene–disease predictions. Brief Bioinform
442 19:575–592.
443 31. Silva AJ, Benitez JA. 2016. Vibrio cholerae Biofilms and Cholera Pathogenesis. PLoS Negl
444 Trop Dis 10:e0004330.
445 32. Teschler JK, Zamorano-Sánchez D, Utada AS, Warner CJA, Wong GCL, Linington RG, Yildiz
446 FH. 2015. Living in the matrix: assembly and control of Vibrio cholerae biofilms. Nat Rev
447 Microbiol 13:255–68.
448 33. Papenfort K, Förstner KU, Cong J-P, Sharma CM, Bassler BL. 2015. Differential RNA-seq
449 of Vibrio cholerae identifies the VqmR small RNA as a regulator of biofilm formation.
450 Proc Natl Acad Sci 112:E766–E775.
451 34. Weber GG, Klose KE, Klose. 2011. The complexity of ToxT-dependent transcription in
452 Vibrio cholerae. Indian J Med Res 133:201–6.
453 35. Kazi MI, Conrado AR, Mey AR, Payne SM, Davies BW. 2016. ToxR Antagonizes H-NS
454 Regulation of Horizontally Acquired Genes to Drive Host Colonization. PLOS Pathog
455 12:e1005570.
456 36. Dorman MJ, Dorman CJ. 2018. Regulatory Hierarchies Controlling Virulence Gene
457 Expression in Shigella flexneri and Vibrio cholerae. Front Microbiol.
458 37. Ayala JC, Wang H, Silva AJ, Benitez JA. 2015. Repression by H-NS of genes required for
459 the biosynthesis of the Vibrio cholerae biofilm matrix is modulated by the second
460 messenger cyclic diguanylic acid. Mol Microbiol 97:630–645.

461 38. Boyd EF, Jermyn WS. 2002. Characterization of a novel Vibrio pathogenicity island (VPI-
462 2) encoding neuraminidase (nanH) among toxigenic Vibrio cholerae isolates.
463 Microbiology 148:3681–3693.
464 39. Karaolis DKR, Johnson JA, Bailey CC, Boedeker EC, Kaper JB, Reeves PR. 1998. A Vibrio
465 cholerae pathogenicity island associated with epidemic and pandemic strains. Proc Natl
466 Acad Sci 95:3134 LP – 3139.
467 40. Nielsen AT, Dolganov NA, Otto G, Miller MC, Wu CY, Schoolnik GK. 2006. RpoS controls
468 the Vibrio cholerae mucosal escape response. PLoS Pathog 2:e109–e109.
469 41. Wong GT, Bonocora RP, Schep AN, Beeler SM, Lee Fong AJ, Shull LM, Batachari LE, Dillon
470 M, Evans C, Becker CJ, Bush EC, Hardin J, Wade JT, Stoebel DM. 2017. Genome-Wide
471 Transcriptional Response to Varying RpoS Levels in Escherichia coli K-12. J Bacteriol
472 199:e00755-16.
473 42. Russo G, Zegar C, Giordano A. 2003. Advantages and limitations of microarray
474 technology in human cancer. Oncogene 22:6497–6507.
475 43. Serin EAR, Nijveen H, Hilhorst HWM, Ligterink W. 2016. Learning from Co-expression
476 Networks: Possibilities and Challenges. Front Plant Sci 7:444.
477 44. Koschmann J, Bhar A, Stegmaier P, Kel AE, Wingender E. 2015. "Upstream Analysis": An
478 Integrated Promoter-Pathway Analysis Approach to Causal Interpretation of Microarray
479 Data. Microarrays (Basel, Switzerland) 4:270–286.
480 45. Mueller RS, Mcdougald D, Cusumano D, Sodhi N, Kjelleberg S, Azam F, Bartlett DH. 2007.
481 Vibrio cholerae Strains Possess Multiple Strategies for Abiotic and Biotic Surface
482 Colonization. J Bacteriol 189:5348–5360.
483 46. Marchler-Bauer A, Bo Y, Han L, He J, Lanczycki CJ, Lu S, Chitsaz F, Derbyshire MK, Geer

RC, Gonzales NR, Gwadz M, Hurwitz DI, 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.

47. Duran-Pinedo AE, Paster B, Teles R, Frias-Lopez J. 2011. Correlation Network Analysis Applied to Complex Biofilm Communities. PLoS One 6:e28438.

48. 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.

49. Meisel JS, Sfyroera G, Bartow-McKenney C, Gimblet C, Bugayev J, Horwinski J, Kim B, Brestoff JR, Tyldsley AS, Zheng Q, Hodkinson BP, Artis D, Grice EA. 2018. Commensal microbiota modulate gene expression in the skin. Microbiome 6:20.

50. 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.

51. 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.

52. 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.

507 53. Wang J, Wu G, Chen L, Zhang W. 2013. Cross-species transcriptional network analysis
508 reveals conservation and variation in response to metal stress in cyanobacteria. BMC
509 Genomics 14:112.
510 54. Keseler IM, Mackie A, Santos-Zavaleta A, Billington R, Bonavides-Martínez C, Caspi R,
511 Fulcher C, Gama-Castro S, Kothari A, Krummenacker M, Latendresse M, Muñiz-Rascado
512 L, Ong Q, Paley S, Peralta-Gil M, Subhraveti P, Velázquez-Ramírez DA, Weaver D, Collado-
513 Vides J, Paulsen I, Karp PD. 2017. The EcoCyc database: reflecting new knowledge about
514 Escherichia coli K-12. Nucleic Acids Res2016/11/29. 45:D543–D550.
515 55. Ballouz S, Verleyen W, Gillis J. 2015. Guidance for RNA-seq co-expression network
516 construction and analysis: Safety in numbers. Bioinformatics 31:2123–2130.
517 56. Li S, Tighe SW, Nicolet CM, Grove D, Levy S, Farmerie W, Viale A, Wright C, Schweitzer
518 PA, Gao Y, Kim D, Boland J, Hicks B, Kim R, Chhangawala S, Jafari N, Raghavachari N,
519 Gandara J, Garcia-Reyero N, Hendrickson C, Roberson D, Rosenfeld JA, Smith T,
520 Underwood JG, Wang M, Zumbo P, Baldwin DA, Grills GS, Mason CE. 2014. Multi-platform
521 assessment of transcriptome profiling using RNA-seq in the ABRF next-generation
522 sequencing study. Nat Biotechnol 32:915–925.
523 57. Goh WW Bin, Wang W, Wong L. 2017. Why Batch Effects Matter in Omics Data, and How
524 to Avoid Them. Trends Biotechnol. Elsevier Ltd.
525 58. Leek JT, Storey JD. 2007. Capturing heterogeneity in gene expression studies by
526 surrogate variable analysis. PLoS Genet 3:1724–1735.
527 59. Alter O, Brown PO, Botstein D. 2000. Singular value decomposition for genome-Wide
528 expression data processing and modeling. Proc Natl Acad Sci U S A 97:10101–10106.
529 60. 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.

61. 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.

62. 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.

63. Rodríguez-García A, Sola-Landa A, Barreiro C. 2017. RNA-Seq-Based Comparative Transcriptomics: RNA Preparation and Bioinformatics BT - Microbial Steroids: Methods and Protocols, p. 59–72. *In* Barredo, J-L, Herráiz, I (eds.), . Springer New York, New York, NY.

64. Leinonen R, Sugawara H, Shumway M, Collaboration INSD. 2011. The sequence read archive. Nucleic Acids Res2010/11/09. 39:D19–D21.

65. Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol 34:525–527.

66. Langmean B, Salzberg SL. 2012. Bowtie 2. Nat Methods 9:357–359.

67. Gaspar JM. 2018. Improved peak-calling with MACS2. bioRxiv 496521.

68. 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.

553 69. Soneson C, Love MI, Robinson MD. 2015. Differential analyses for RNA-seq: transcript-
554 level estimates improve gene-level inferences. F1000Research 4:1521.
555 70. Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion
556 for RNA-seq data with DESeq2. Genome Biol 15:550.

557

558 Legends:

559 Figure 1: General outline of network construction.

560 1A) Normalized (log2) expression reads for the same genes across multiple conditions supply
561 the basis for our co-expression analysis. In this small example, it is clear that genes VC0384-
562 VC0386 have a very similar expression pattern across conditions. 1B) Correlations are
563 calculated from the normalized counts in A for every pair of genes. The pattern seen in A
564 becomes much clearer when looking at the correlation. 1C) An adjacency matrix (not shown) is
565 calculated from the correlations in B and ultimately used to produce a topological overlap
566 matrix (TOM) that supplies network edge weights with less noise than the raw correlation
567 matrix. While the single of co-expressing pairs is dampened slightly, this step greatly decreases
568 spurious relationships as it favors transcripts which coexpress with similar sets of genes
569 rather than potentially noisy direct correlations. 1D) The final network groups transcripts that
570 tightly co-express while indicating what pathway they are involved in. This network also
571 includes functional and essentiality based knowledge. In this case, the three genes involved in
572 cysteine metabolism (VC0383-VC0385, *cysHIJ*) form a subnetwork while the other genes do
573 not meet our 0.10 co-expression cutoff.

574

575 Figure 2: Sub-networks recapitulating known results

The four depicted subnetworks each contain subsets of transcripts which are known to be largely involved in the same biological process. For each subnetwork, the nodes represent transcripts while the edges represent a co-expression relationship of at least 0.1 between transcripts. A) This sub-network consists completely of tRNA transcripts. B) These transcripts are almost completely related to ribosomal structure and/or function. C) These transcripts play a role in amino acid synthesis. D) This sub-network contains a majority of transcripts that play a role in biofilm formation in addition to unrelated genes.

Figure 3: Virulence related subnetwork.

3A) 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 locus VCA0094-VCA0096. 3B) Mean binding affinity (log2 fold change in occupancy compared to loading control) for different virulence-associated transcription factors near the VCA0094-96 locus. Both HNS and TOXR show a significant binding preference for this region. Error bars indicate standard deviation from the mean.

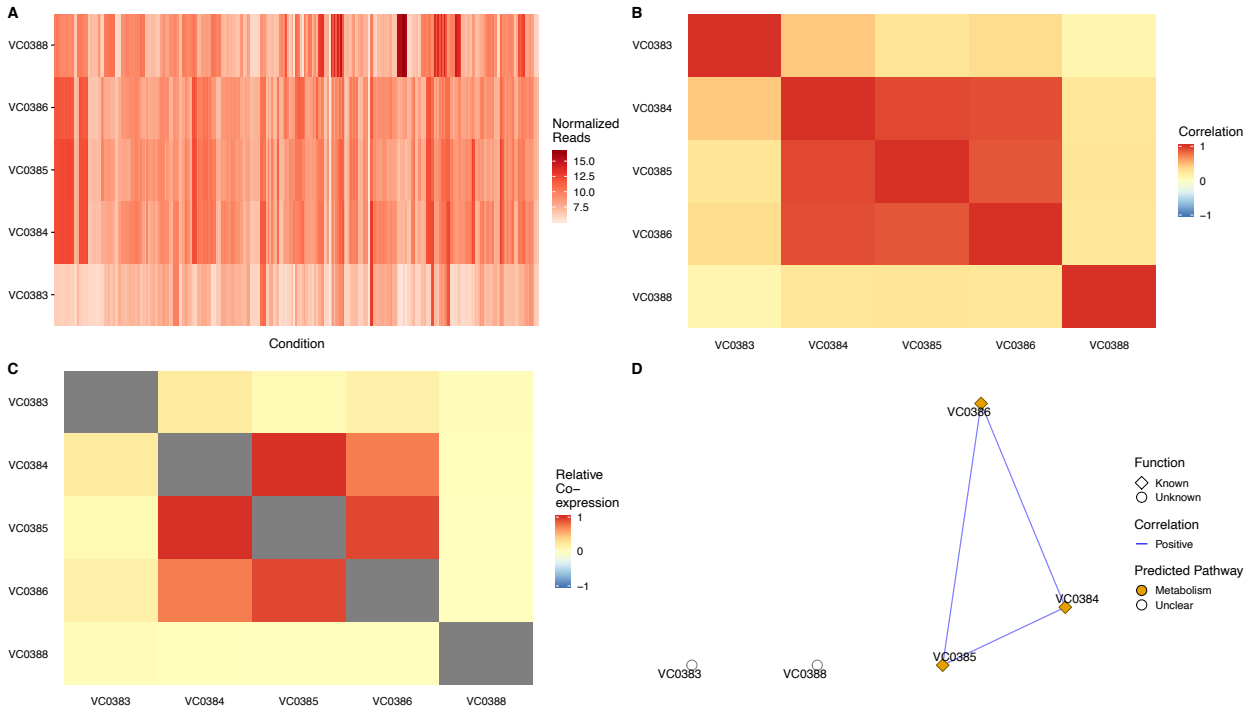
Figure 4: Comparing RpoS microarray data to co-expressing genes in our WGCNA

A) Overlap of genes with expression pattern related to *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 A). All of the flagellar and chemotaxis genes highlighted as particularly important in the microarray dataset are identified by both methods.

599

600

601 Figure 1



602

603

604

605

606

607

608

609

610

611

612

Figure 2

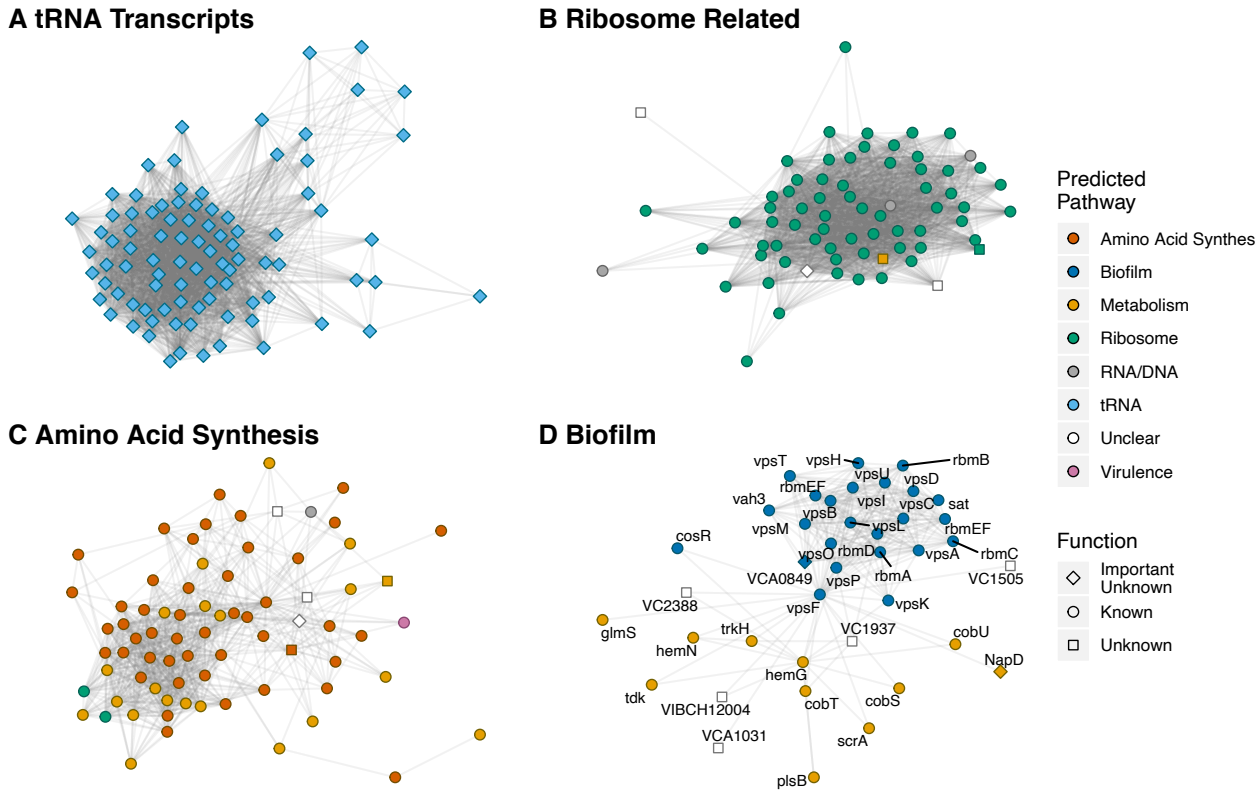
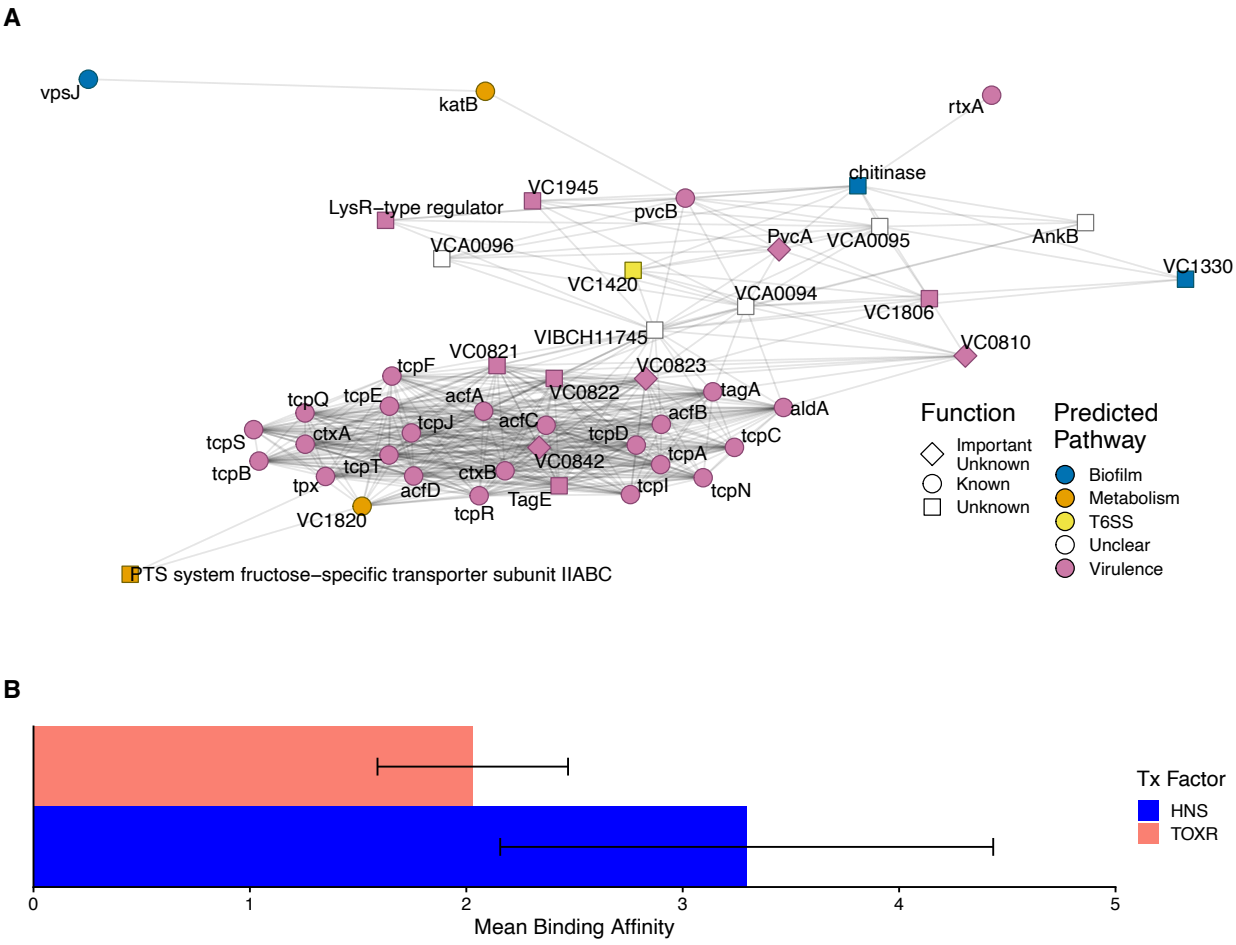


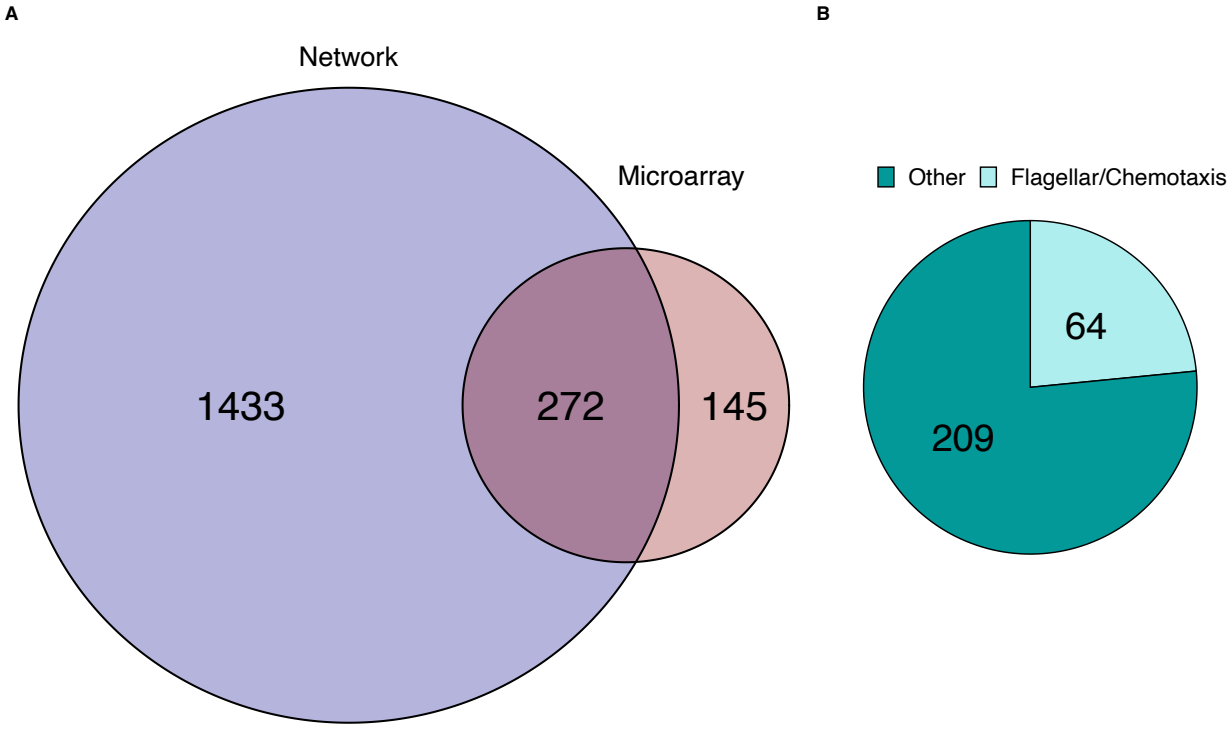
Figure 3



636

637

638 Figure 4



639