Check for updates

# Mr.Vc v2: An updated version of database with increased data of transcriptome and experimental validated interactions

Zhiyuan Zhang[1,2], Guozhong Chen[3], Wajid Hussain[4], Zixin Qin[3], Juntong Liu[1,2], Yang Su[1,2], Hao Zhang[1]*and Mingquan Ye[1,2]*

[1]School of Medical Information, Wannan Medical College, Wuhu, China, [2]Research Center of Health Big Data Mining and Applications, Wannan Medical College, Wuhu, China, [3]College of Life Sciences and Technology, Huazhong University of Science and Technology, Wuhan, China, [4]Advanced Biomaterials and Tissue Engineering Center, College of Life Sciences and Technology, Huazhong University of Science and Technology, Wuhan, China

Mr.Vc is a database of curated *Vibrio cholerae* transcriptome data and annotated information. The main objective is to facilitate the accessibility and reusability of the rapidly growing *Vibrio cholerae* omics data and relevant annotation. To achieve these goals, we performed manual curation on the transcriptome data and organized the datasets in an experiment-centric manner. We collected unknown operons annotated through text-mining analysis that would provide more clues about how *Vibrio cholerae* modulates gene regulation. Meanwhile, to understand the relationship between genes or experiments, we performed gene co-expression analysis and experiment-experiment correlation analysis. In additional, functional module named "Interactions" which dedicates to collecting experimentally validated interactions about *Vibrio cholerae* from public databases, MEDLINE documents and literature in life science journals. To date, Mr.Vc v2, which is significantly increased from the previous version, contains 107 microarray experiments, 106 RNA-seq experiments, and 3 Tn-seq projects, covering 56,839 entries of DEGs (Differentially Expressed Genes) from transcriptomes and 7,463 related genes from Tn-seq, respectively. and a total of 270,129 gene co-expression entries and 11,990 entries of experiment-experiment correlation was obtained, in total 1,316 entries of interactions were collected, including 496 protein-chemical signaling molecule interactions, 472 protein−protein interactions, 306 TF (Transcription Factor)-gene interactions and 42 *Vibrio cholerae*-virus interactions, most of which obtained from 402 literature through text-mining analysis. To make the information easier to access, Mr.Vc v2 is equipped with a search widget, enabling users to query what they are interested in. Mr.Vc v2 is freely available at http://mrvcv2.biownmc.info.

## Introduction

Cholera is a notorious and devastating diarrheal disease, until now had caused seven epidemics in history globally and is still endemic in many parts of the world, especially developing countries like Asia, South America, and Africa (Faruque et al., 1998; Liu et al., 2008; Xia et al., 2017).

*Vibrio cholerae* is the causative agent of cholera resulting in 23,000 to 143,000 people dying worldwide annually (Qin et al., 2020). However, the pathogenesis of *Vibrio cholerae* is still unclear. Growing evidence suggests that *Vibrio cholerae* can rapidly modulate its gene transcriptional expression in response to the switches of different environments for better survival and infection. Therefore, studying the gene expression profile of *Vibrio cholerae* under various conditions is important to fully expose and understand the genetic mechanism of *Vibrio cholerae*.

In recent years, the transcriptional sequencing is powerful enough to study general gene expression profiles, sequencing of *Vibrio cholerae* transcriptomes rapidly increased the number and total volume of *Vibrio cholerae* transcriptome data. At present, most of the raw sequencing data has been deposited into several general-purpose databases, such as European Nucleotide Archive (ENA) (Amid et al., 2020)[1] and NCBI Sequence Read Archive (SRA) (Kodama et al., 2012).[2] In addition, several other public resources, including MicrobesOnline (Dehal et al., 2010), PubMLST (Jolley et al., 2018), BioCyc (Paley and Karp, 2017), and PATRIC (Wattam et al., 2014), collected the processed transcriptome data, microbial genome and metabolic pathway information and then organized them according to experimental conditions and organisms for the purposes to greatly promote data reuse. However, obstacles to the reusability and accessibility of the rapidly growing *Vibrio cholerae* transcriptome data remain, especially the inaccurate data sets and/or incomplete data. For example, MicrobesOnline, which integrated vast amounts of microbial genetic information, did not update transcriptome data since 2012 and only collected 42 high-throughput *Vibrio cholerae* microarray data, not collecting RNA-seq data, under different experimental conditions, deriving from seven published papers. Additionally, MicrobesOnline did not compute *p*-values in differential gene expression analysis due to lacking technical/ biological replicates. PubMLST, which integrated population sequence data with provenance and phenotype information for over 100 different microbial species and genera, focus on the analysis of molecular typing and microbial genome diversity, and did not collect microbial transcriptome data, including *Vibrio cholerae*. BioCyc, a database for collection of the genome and metabolic pathways of organisms, provided annotations, essentiality and reactions of gene or protein, however, transcriptome data were not considered for collection in BioCyc.

Although identified DEGs from transcriptome data are useful to discover novel genes for phenotypes, obtaining a comprehensive

regulatory network is meaningful (Saint-André, 2021). To date, several databases, including STRING (Szklarczyk et al., 2021), CollecTF (Kiliç et al., 2014), STITCH (Szklarczyk et al., 2016), and MVP (Gao et al., 2018), had provided data of protein–protein, TF (Tanscription Factor)-gene, protein-chemical signaling molecule, and *Vibrio cholerae*–virus interactions. These data will better help to understand *Vibrio cholerae* and useful for researchers to work on *Vibrio cholerae*. However, some interaction data presented in databases were obtained through prediction rather than experimental verification, which cannot ensure the data quality.

In 2019, we introduced Mr.Vc v1 as an online database of curated microarray and RNA-seq of *Vibrio cholerae* to facilitate the reusability and accessibility of the rapidly increasing *Vibrio cholerae* omics data and relevant annotation (Zhang et al., 2019). We collected data from 145 high-throughput gene expression experiments of *Vibrio cholerae* from 49 journal articles and the detailed annotation for 3,834 genes of *Vibrio cholerae* (*Vibrio cholerae O1 biovar eltor str. N16961*), we also collected relevant information including which operons they may belong to and possible interaction partners of their protein products. Mr.Vc is the first comprehensive data repository dedicated to *Vibrio cholerae* and could provide convenience for all researchers in related fields.

In current study, we are going to introduce an updated version of Mr.Vc. In this new version, we extended more annotation information, such as operon annotation through text-mining analysis, collected more transcriptome projects, samples, experiments, and performed extensive analysis, including differential gene expression analysis, operon member expression visualization, gene co-expression analysis, and experiment-experiment correlation analysis. Most importantly, we added an "Interactions" functional module in Mr.Vc v2, which listed collected experimentally validated interaction entries about *Vibrio cholerae*. We obtained them from public databases including STRING, STITCH, CollecTF, MVP, literature-describing interactions referring to *Vibrio cholerae*, and text-mining results from MEDLINE documents and literature in life science journals. Additionally, we manually categorized these interaction data into four interaction types, including protein-chemical signaling molecule interaction, protein–protein interaction, TF (Transcription Factor)-gene interaction, and *Vibrio cholerae*-virus interaction. These interaction data will provide us with more clues into the regulatory network and mechanism in *Vibrio cholerae* (Kathuria and Chattopadhyay, 2018). Mr.Vc v2 is equipped with a search widget, allowing experimental biologists and medical scientists to quick and easy finding what they are interested.

## Materials and methods

### Data collection and pre-processing

To give users a clear overview of the data collection, pre-processing, and integration in Mr.Vc v2, we provided the detailed workflow (Figure 1).

For data collection, as shown in Figure 1A, we searched recently updated RNA-seq projects in the NCBI BioProject

---
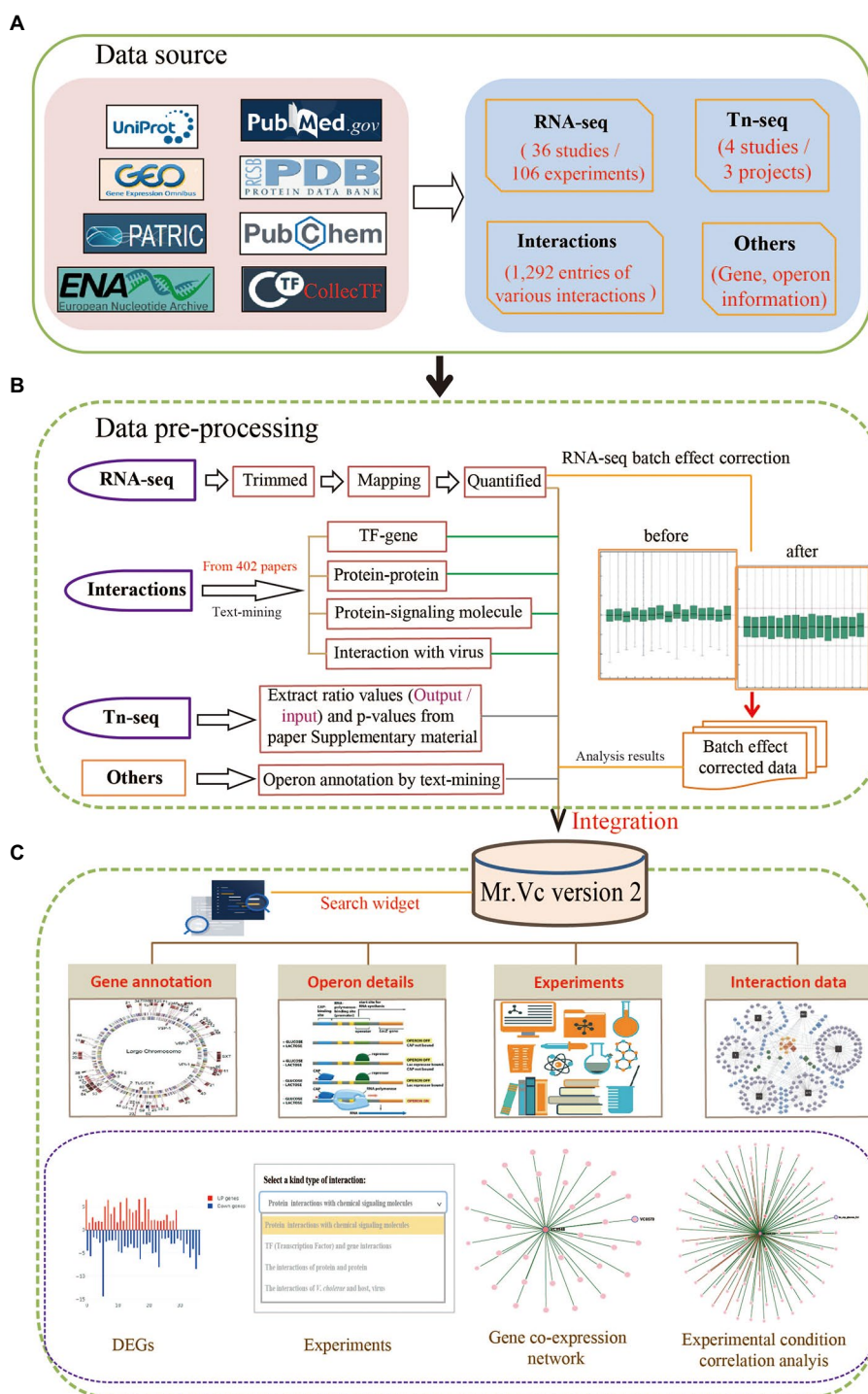
FIGURE 1
Overview of Mr.Vc v2. **(A)** The data source. All data including the raw sequencing data, information on genomic annotation, literature, Tn-seq data, and experimentally validated interactions were collected from public databases, such as NCBI GEO, ENA, PATRIC, Uniprot, collecTF, PDB, PubMed, and so on. **(B)** The data pre-processing. The RNA-seq raw data perform quality control, mapping, transcript quantification, and normalization to obtain high-quality data for further analysis. The resulting operon and interaction data were obtained from MEDLINE documents and literature in life science journals through text-mining analysis. The Tn-seq data were directly extracted from supplemental materials of literature. **(C)** The data integration. Mr.Vc v2 integrated polytype data, which was placed in the "Genes," "Operons," "Experiments," and "Interactions" pages and provided a search widget.

database[3] and publications in NCBI PubMed[4] using "*Vibrio cholerae*" as the keyword. Projects with public raw sequencing data, 106 RNA-seq datasets were collected for further analysis. The raw sequencing reads were downloaded from EBI ENA (European Nucleotide Archive,[5]) and NCBI SRA (Sequence Read Archive,[6]) using command line tools from enaBrowserTools[7] and SRA-Tools[8] facilitated by Aspera (a high-speed data transfer tool). As the related meta-data of corresponding experiments, projects and literature were obtained from NCBI PubMed and GEO databases.

For the processing of raw sequencing reads, we used FastQC[9] to evaluate the overall quality of the downloaded data, followed by the Trim_galore to remove sequencing vectors and low-quality bases (Utturkar et al., 2020). Salmon is the latest computational algorithm for transcript quantification from RNA-seq data that could make the expression data compared across experiments and projects (Patro et al., 2017). In this study, we performed transcript quantification using Salmon, which adopted TPM (Transcript Per Million) for normalization, a better unit for RNA abundance than RPKM and FPKM since it respects the invariance property and is proportional to the average relative RNA molar concentration (Zhao et al., 2020). We searched Tn-seq projects in the NCBI PubMed database and downloaded corresponding supplemental materials of literature. These data included experiment details and the ratio value of output/input.

We performed text-mining analysis for *Vibrio cholerae* operon annotation using a customized Python pipeline by searching for functional descriptions of operon member genes (e.g., gene symbol, locus ID, common name, and aliases) in the titles and abstracts of MEDILNE documents and literature in life science journals available from the NCBI PubMed database. Additionally, for each operon, we visualized the expression profile of its member genes using TPM (Transcript Per Million) values in 110 different experiments. For example, operon "OP437" is involved in the function of bacterial motility, whose member genes expression trends were intuitively observed through visualization[10] that could help us deduce a positive or negative regulatory relationship between genes.

In addition, we also had a special focus on experimentally validated interactions about *Vibrio cholerae* in this study. We retrieved related interaction entries by following steps: (1) searching for literature with the keyword: "*Vibrio cholerae*"; (2) Filtering sentence from literature, and one sentence containing two of these: *Vibrio cholerae* gene name, chemical molecule name, and virus name. (3) This sentence should also contain one of the following words/phrases: *interaction, altered, associated, caused, confer, contribute, association, downregulate, elevate, implicated, increase, induce, influence, interact, involved, lead to, link, mediate, modulate, overexpressed, reduce, regulate, related, relationship, treat, binding, environmental signals, target, pathogenic, pathogenesis, progression, and transcriptional regulation.* (4) Extracting these sentences from literature as supporting evidence and finally identified 1,316 entries of different interaction types from 402 literature shown in (Table 1), including (1) a total 472 protein–protein interactions, all of which were experimentally validated; (2) 496 protein-chemical signaling molecule interactions that included protein information extracted from Uniprot and details of chemical molecule extracted from PubChem (Kim et al., 2021) and PDB database (Burley et al., 2017); (3) 306 TF-gene interactions, most of which obtained from collecTF database; (4) 42 *Vibrio cholerae* -virus interactions, including 30 entries extracted from literature and 13 from MVP database.

## Data analyses

We analyzed transcriptome data and out of a total 106 RNA-seq experiments, 318 samples, in which the expression abundances were normalized as TPM values. For RNA-seq experiments, we used a cutoff of $|\log2\ FC| > 1.5$ (FC, fold change) and $p$-value $<0.05$ to define differentially expressed genes between experiments. The 33,180 differentially expressed gene entries were extracted, representing *Vibrio cholerae* gene expression under 106 different experimental conditions. We performed gene co-expression analysis and experiment-experiment correlation analysis using an in-house Python script to calculate pearson's correlation coefficients, spearman's correlation coefficients, and $p$-value between genes and between experiments, separately.

---

TABLE 1  Data summary in Mr.Vc v2 database.

| Mr.Vc | Version 1 | Version 2 |
| --- | --- | --- |
| Genes | 3,834 | 3,998 |
| Well-annotated operons | 415 | 600 |
| Literature | 49 | 402 |
| Transcriptome experiment | 145 | 213 |
| DEGs | 25,937 | 56,839 |
| Tn-seq projects | 0 | 3 |
| Top 5% differentially expressed genes | 0 | 161 |
| Experimentally validated interactions | 0 | 1,316 |
| Gene co-expression entries | 0 | 270,129 |
| Entries of experiment-experiment correlation | 0 | 11,990 |
| Text-mining analysis | No | Yes |
| Data visualization | No | Yes |

Pearson's correlation coefficient have a greater statistical power than spearman's coefficients, however which requires that the statistical data should conform to normal distribution. Transcriptome data tend to conform to negative binomial distribution, causing poor reliability of pearson's correlation coefficient analysis. And if using spearman's coefficients, the statistical power is not good. So we provided both spearman's coefficients and pearson's correlation coefficients for complementation on Mr.Vc v2 database. We used a cutoff of pearson or spearman >0.8 and $p$-value <0.05 to filter data of correlation, 270,129 gene–gene correlation entries and 11,990 experiment-experiment correlation entries were obtained. In order to make more comprehensive use of these two coefficients of correlation, we calculated their average value for users' reference. In addition, we calculated the $p$-values using pearson's and spearman's coefficients, respectively, and then compared the calculated $p$-value, finally outputted the larger $p$-value (< 0.05) to page.

## Propose of "Top 5% differentially expressed gene" hypothesis

During the analysis of *Vibrio cholerae* transcriptome data, we found that some genes tended to be differentially expressed in most experiments. These observations indicated multiple functions of genes. To seek out these potential genes and their biological significance, we proposed a hypothesis that 5% of genes in bacteria are active in most conditions, which defined as "Top 5% differentially expressed gene." We calculated the numbers of the experimental condition of each gene when differentially expressed in experiments, and ranked the corresponding gene by the calculated number of experimental conditions. Then, we calculated the number of "Top 5% differentially expressed gene" in *Vibrio cholerae* that used the product of 0.05 and numbers of gene except for essential genes, which is stable in gene expression and less biological significance, in total obtained 161 genes. Finally, according to the ranking result, we regarded the top 161 genes as "Top 5% differentially expressed gene." For more details of "Top 5% differentially expressed gene," please go to "Help" page.[11]

## Database design and implementation

Mr.Vc v2 was designed as a relational database. All data were loaded into a MySQL database. The frontend of the website was coded using JavaScript and HTML, while the backend was coded using PHP with a Slim framework to support queries to the MySQL database and provide representational state transfer (REST) application programming interfaces (APIs) for programmable access to our data. The AngularJS framework was used to bride the front- and back-ends. Echarts.js and plotly.js used for visualizations at the front end. The website hosted on an Apache server.

# Results and discussion

## Overview of Mr.Vc v2

In Mr.Vc v2, we updated gene and operon annotation information, including 3,842 protein-coding genes, 98 tRNAs, 16 rRNA, 42 pseudo genes, 161 Top 5% differentially expressed genes, and 600 well-annotated operons. We also provided external links to public databases such as KEGG (Kanehisa et al., 2017), NCBI Entrez Gene,[12] OGEE (Cao et al., 2019), Uniprot (UniProt, Consortium, 2021) and MicrobesOnline to allow users to explore in more details of these genes and operons;

We analyzed 106 RNA-seq datasets and performed differentially expressed gene analysis, experiment-experiment correlation analysis, and gene co-expression analysis, in total obtaining 33,180 DEGs, 11,990 entries of experiment-experiment correlation, and 270,129 entries of gene co-expression. In addition, Mr.Vc v2 added 3 Tn-seq projects and 1,316 entries of experimentally validated interaction through public databases and text-mining analysis, including protein–protein interaction, protein-chemical signaling molecule interaction, TF-gene interaction, and *Vibrio cholerae*-virus interaction. All data shown in Table 1.

## Top 5% differentially expressed genes shared by experiments

We found that experiments collected in our database shared some differentially expressed genes with other experiments, which suggested these genes might play an important role in multiple phenotypes. Therefore, the concept of "Top 5% differentially expressed gene" was proposed and according to the hypothesis, we ranked differentially expressed genes in 213 experiments based on the frequency of the gene's appearance, in total obtaining 161 Top 5% differentially expressed genes (Figure 2). For example, Top 5% differentially expressed gene VCA1028, an outer membrane protein, was differentially expressed in 67 experiments (~31% of all), including gene deletion background, stress, nutritional condition, and human infection experiment. These data indicated this gene might be involved in several signaling pathways and functions. A previous study reported that VCA1028 was associated with virulence and indirectly regulated by ToxT (Thomson et al., 2015). However, there was no research to discuss its relationship with nutrition and stress. Therefore, the appearance of a "Top 5% differentially expressed gene" could remind researchers of its multi-function.

**FIGURE 2**
Summary of Top 5% differentially expressed genes. The treemap plot summarizes Top 5% differentially expressed genes. The number below the gene name is the number of experimental conditions that the gene was found to be differentially expressed, including up- and down-regulated. Mr.Vc v2 calculates the numbers of the experimental condition of each gene when differentially expressed in experiments, and ranks the corresponding gene by the calculated number of experimental conditions. A total of 161 Top 5% differentially expressed genes were obtained, according to ranking analysis. The Top 5% differentially expressed gene of the last rank was differentially expressed in 40 experimental conditions.

## Case study

In *Vibrio cholerae, toxT* and *tcpN* used to refer to the same gene (VC0838). In order to make users pain-free for these synonyms, we added "Aliases" data for genes.[13] The *Vibrio cholerae toxT* gene (VC0838) was took as an example to represent how to use Mr.Vc v2 database for extraction of related information. The users have the access to five pages including "Genes," "Operons," "Experiments," "Interactions" and "Download." On the Mr.Vr v2 database functions, an additional page for instructions and a home page. On the "Genes" page, all gene records were listed, according to their gene locus. Users can find the individual gene information, including gene ID, description, gene location, gene orientation, gene length, and gene essentiality (Figure 3A). In addition, users can click the "VC0838" link redirect to detailed gene information, which includes a list of DEGs in experimental conditions collected in the database (Figure 3C), and the gene

co-expression network (Figure 3B). We found that *toxT* is differentially expressed in 28 experiments, which are mainly related to hosting infection and oxidative stress, up-regulated in 9 experiments, and down-regulated in 19 experiments. Additionally, the gene co-expression network suggested that *toxT* is involved in cholera pathogenesis. These observations were consistent with previous findings that ToxT is a transcriptional activator to regulate virulence factors (Baranova et al., 2020; Stone and Withey, 2021).

On the "Operons" page, users can directly search VC0838 to find the corresponding operon, where a table used to provide a summary report, including the Operon ID, and putative operon function. Users can expand the table by clicking the '+' sign before the "Operon ID" to view information on its member genes (Figure 3D).

The "Experiments" page summarized information on 3 Tn-seq projects and 213 transcriptome experiments, including a brief summary of experimental conditions, DEGs (Differentially Expressed Genes), and corresponding reference(s). If users want to know in which experiments VC0838 is differentially expressed, Searching on "Genes" page is a better choice, because "Experiments" page does not provide the summary report in a gene-centric manner.
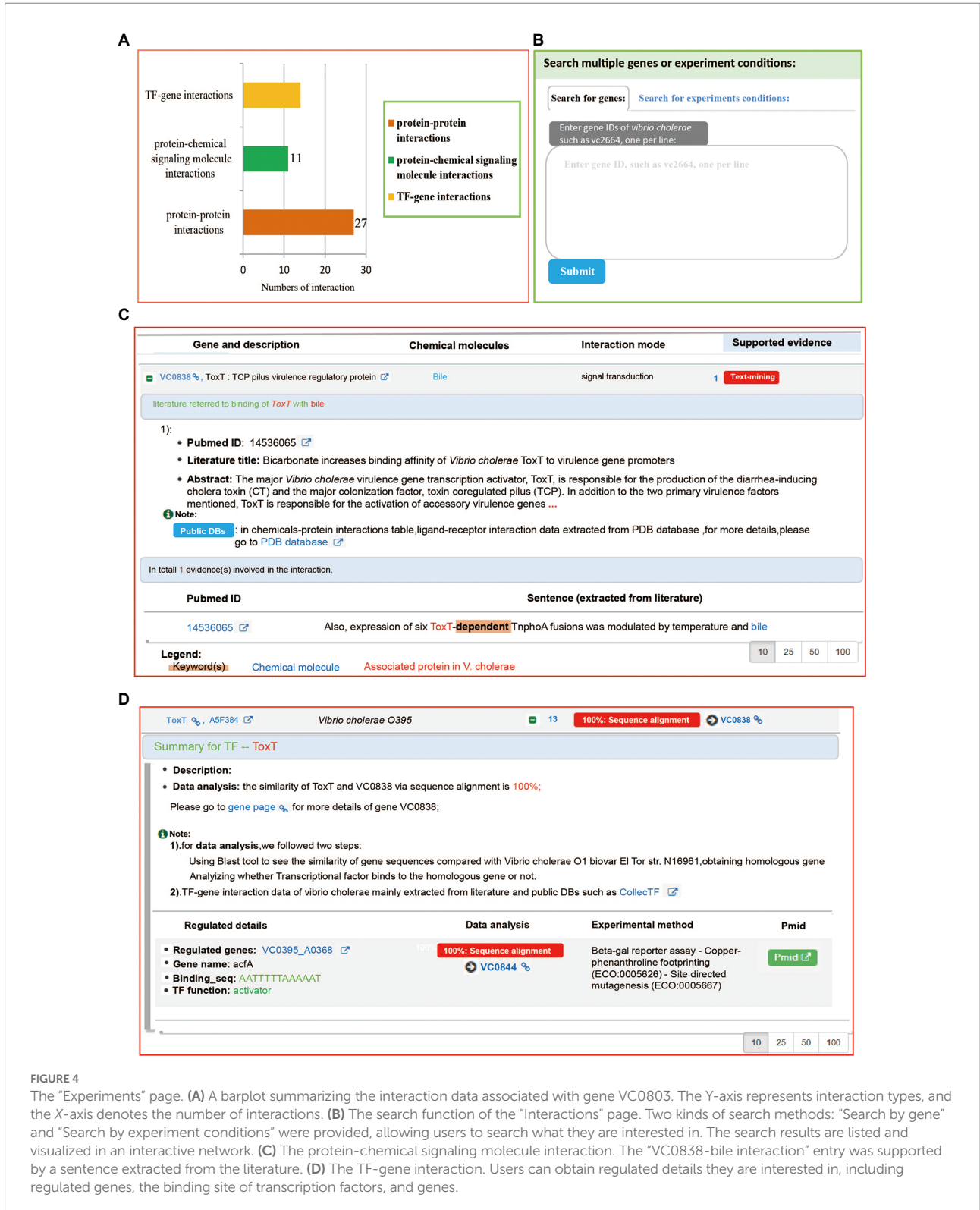
---

13   http://mrvcv2.biownmc.info/genes/

**A**

**Genomic information:**

- **Essentiality**: *Essential* gene, please go to Ogee database for more details
- **Genomic coordinates(start - end):** locating at 899896 - 900726 on choromosome
- and gene length: 831 , operon ID: OP166 ➡
- Annotation source: reviewed in Uniprot

*External link:* [OGEE:VC0838] [KEGG] [MicrobesOnline] [STRING database]

**B**



**C**

In total **28** experiment(s) in which VC0838 has differentially expressed. [Microarrays, RNA-seq data]

| Experiment conditions | Log2 FoldChange | Adjust p_value | Omics type(s) |
|---|---|---|---|
| ID: 113<br>WT OD-2 vs. KO hns OD-2 | 7.90633 | 8.27e-55 | RNA-seq |
| ID: 78<br>*V.cholerae* infection in mouse model vs. strain cultured in LB | 5.5 | 1.51e-38 | RNA-seq |
| ID: 164<br>growth: OD600 0.2 vs. OD600 2.5 | -2.48034 | 9.21e-27 | RNA-seq |
| ID: 167<br>KO varA; growth: OD600 0.2 vs. KO varA; growth: OD600 2.5 | 2.32142 | 5.68e-16 | RNA-seq |
| ID: 76<br>*V.cholerae* infection in Rabbit model vs. strain cultured in LB | 6.48 | 1.83e-15 | RNA-seq |
| ID: 200<br>oppB mutated;BCM treatment vs.oppB mutated;H2O treatment | 4.04237 | 4.29e-15 | RNA-seq |
| ID: 182<br>adding compound MMV675968 vs. DMSO | -4.44352 | 8.43e-8 | RNA-seq |

[10] 25 50 100

**D**

Search operon IDs and their functions:                                                    [Clear search]

[Except for... ☐]  vc0838                                                                        🔍

| Operon ID | Putative functions | Nr. Member genes |
|---|---|---|
| ➖ OP166 🔗☑ | Involved in TCP pilus biogenesis,virulence,colonization | [Public DBs] [6] |

**Member genes:**

| Gene ID and description | Genomic coordinates(start-end) | Strand | COG |
|---|---|---|---|
| VC0843 🔗 - tagE protein ☑ | 904648 - 905556 | + | COG0739 **M** |
| VC0842 🔗 - hypothetical protein ☑ | 904166 - 904639 | + | COG1714 **S** |
| VC0841 🔗 - accessory colonization factor AcfC ☑ | 903380 - 904150 | + | COG4588 **R** |
| VC0840 🔗 - accessory colonization factor AcfB ☑ | 901494 - 903374 | + | COG0840 **NT** |
| VC0839 🔗 - Type 4 prepilin-like proteins leader peptide-processing enzyme ☑ | 900726 - 901487 | + | COG1989 **NOU** |
| VC0838 🔗 - TCP pilus virulence regulatory protein ☑ | 899896 - 900726 | + | COG2207 **K** |

[10] 25 50 100

**FIGURE 3**

The "Genes and "Operons" page. **(A)** A summary of *Vibrio cholerae* gene VC0838. The information of gene annotation including gene ID, description, gene location, gene orientation, gene length, and gene essentiality was provided. **(B)** The gene co-expression network. Mr.Vc v2 filtered and obtained gene co-expression entries, according to the calculated pearson's correlation coefficients (>0.8), Spearman's correlation coefficients (>0.8), and *p*-value (<0.05) between genes. There are 14 genes associated with gene VC0838 in our database. **(C)** A list of experiments in which gene VC0838 is differentially expressed in. A total of 28 experiments were collected, including Microarray and RNA-seq experiments. **(D)** The operon information. Searching genes to match to corresponding operon page, which listed putative function and member genes information of operon.

**FIGURE 4**
The "Experiments" page. **(A)** A barplot summarizing the interaction data associated with gene VC0803. The Y-axis represents interaction types, and the *X*-axis denotes the number of interactions. **(B)** The search function of the "Interactions" page. Two kinds of search methods: "Search by gene" and "Search by experiment conditions" were provided, allowing users to search what they are interested in. The search results are listed and visualized in an interactive network. **(C)** The protein-chemical signaling molecule interaction. The "VC0838-bile interaction" entry was supported by a sentence extracted from the literature. **(D)** The TF–gene interaction. Users can obtain regulated details they are interested in, including regulated genes, the binding site of transcription factors, and genes.

On the "Interactions" page, all interaction records categorized into four groups according to their interaction type, including protein-chemical signaling molecule interaction, protein–protein interaction, TF (Transcription Factor)-gene interaction, and *Vibrio cholerae*-virus interaction.

Users can obtain data of experiment-validated interaction with VC0838. We found that a total 51 interaction records are associated with VC0838, including 27 protein–protein interactions, 11 protein-chemical signaling molecule interactions, and 13 TF-gene interactions (Figure 4A).

Regarding details of protein–protein and protein-chemical signaling molecule interactions, Mr.Vc v2 provided reference(s), sentence(s) describing interaction, and keyword(s) extracted from the sentence (Figure 4C). For TF-gene interaction details, a summary was provided that included information on transcription factor function, regulated genes, binding sequence, and experiment method (Figure 4D). Additionally, a search widget was integrated, users can search genes or experimental conditions that they are interested in (Figure 4B), to obtain gene co-expression or experiment-experiment correlation network.

In addition to the above pages, on the "Download" page, all data in Mr.Vc v2 are downloadable as excel files at the "Downloads" page. Users can click the corresponding file name to download the data tables.

## Conclusion

In this study, we introduced Mr.Vc v2, an updated version of the online database of curated omics data and annotation information. Updates since the last version include increased numbers of RNA-seq experiments/samples, experimental validated interactions, and results of data analysis. So far, Mr.Vc v2 includes 3,998 genes, 2,366 operons, 213 transcriptome experiments, 3 Tn-seq projects, and 1,316 entries of experimentally validated interaction. Additionally, through data analysis, Mr.Vc v2 obtained 56,739 DEGs, 270,129 entries of gene co-expression, and 11,990 entries of experiment-experiment correlation. We believe that Mr.Vc v2 is expected to be a highly useful and important database for biologists and bioinformaticians studying *Vibrio cholerae*. In the future, we aim to update Mr.Vc v2 regularly to provide up-to-date content and include more functionalities.

## Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: http://mrvcv2.biownmc.info/download/.

## Author contributions

ZZ, MY, and HZ designed the study. ZZ and GC collected the data. ZZ analyzed the data prepared and the first draft. WH reviewed and edited the final draft. All authors approved the final submission.

## Funding

## Acknowledgments

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

Amid, C., Alako, B. T. F., Kadhirvelu, V. B., Burdett, T., Burgin, J., Fan, J., et al. (2020). The European nucleotide archive in 2019. *Nucleic Acids Res.* 48, D70–D76. doi: 10.1093/nar/gkz1063

Baranova, D. E., Willsey, G. G., Levinson, K. J., Smith, C., Wade, J., and Mantis, N. J. (2020). Transcriptional profiling of vibrio cholerae O1 following exposure to human anti- lipopolysaccharide monoclonal antibodies. *Pathog. Dis.* 78:ftaa 029. doi: 10.1093/femspd/ftaa029

Burley, S. K., Berman, H. M., Kleywegt, G. J., Markley, J. L., Nakamura, H., and Velankar, S. (2017). Protein data Bank (PDB): the single global macromolecular structure archive. *Methods Mol. Biol.* 1607, 627–641. doi: 10.1007/978-1-4939-7000-1_26

Cao, H., Ma, Q., Chen, X., and Xu, Y. (2019). DOOR: a prokaryotic operon database for genome analyses and functional inference. *Brief. Bioinform.* 20, 1568–1577. doi: 10.1093/bib/bbx088

Dehal, P. S., Joachimiak, M. P., Price, M. N., Bates, J. T., Baumohl, J. K., Chivian, D., et al. (2010). MicrobesOnline: an integrated portal for comparative and functional genomics. *Nucleic Acids Res.* 38, D396–D400. doi: 10.1093/nar/gkp919

Faruque, S. M., Albert, M. J., and Mekalanos, J. J. (1998). Epidemiology, genetics, and ecology of toxigenic vibrio cholerae. *Microbiol. Mol. Biol. Rev.* 62, 1301–1314. doi: 10.1128/MMBR.62.4.1301-1314.1998

Gao, N. L., Zhang, C., Zhang, Z., Hu, S., Lercher, M. J., Zhao, X.-M., et al. (2018). MVP: a microbe-phage interaction database. *Nucleic Acids Res.* 46, D700–D707. doi: 10.1093/nar/gkx1124

Jolley, K. A., Bray, J. E., and Maiden, M. C. J. (2018). Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res.* 3:124. doi: 10.12688/wellcomeopenres.14826.1

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 45, D353–D361. doi: 10.1093/nar/gkw1092

Kathuria, R., and Chattopadhyay, K. (2018). Vibrio cholerae cytolysin: multiple facets of the membrane interaction mechanism of a β-barrel pore-forming toxin. *IUBMB Life* 70, 260–266. doi: 10.1002/iub.1725

Kiliç, S., White, E. R., Sagitova, D. M., Cornish, J. P., and Erill, I. (2014). Collec TF: a database of experimentally validated transcription factor-binding sites in bacteria. *Nucleic Acids Res.* 42, D156–D160. doi: 10.1093/nar/gkt1123

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., et al. (2021). PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Res.* 49, D1388–D1395. doi: 10.1093/nar/gkaa971

Kodama, Y., Shumway, M., and Leinonen, R. (2012). International nucleotide sequence database collaboration. The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Res.* 40, D54–D56. doi: 10.1093/nar/gkr854

Liu, Z., Miyashiro, T., Tsou, A., Hsiao, A., Goulian, M., and Zhu, J. (2008). Mucosal penetration primes vibrio cholerae for host colonization by repressing quorum sensing. *Proc. Natl. Acad. Sci. U. S. A.* 105, 9769–9774. doi: 10.1073/pnas.0802241105

Paley, S., and Karp, P. D. (2017). Update notifications for the BioCyc collection of databases. *Database (Oxford)* 2017:bax086. doi: 10.1093/database/bax086

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* 14, 417–419. doi: 10.1038/nmeth.4197

Qin, Z., Yang, X., Chen, G., Park, C., and Liu, Z. (2020). Crosstalks between gut microbiota and vibrio Cholerae. *Front. Cell. Infect. Microbiol.* 10:582554. doi: 10.3389/fcimb.2020.582554

Saint-André, V. (2021). Computational biology approaches for mapping transcriptional regulatory networks. *Comput. Struct. Biotechnol. J.* 19, 4884–4895. doi: 10.1016/j.csbj.2021.08.028

Stone, J. B., and Withey, J. H. (2021). H-NS and ToxT inversely control cholera toxin production by binding to overlapping DNA sequences. *J. Bacteriol.* 203:e0018721. doi: 10.1128/JB.00187-21

Szklarczyk, D., Gable, A. L., Nastou, K. C., Gable, A. L., Nastou, K. C., Lyon, D., et al. (2021). The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* 49, D605–D612. doi: 10.1093/nar/gkaa1074

Szklarczyk, D., Santos, A., Von Mering, C., Jensen, L. J., Bork, P., and Kuhn, M. (2016). STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data. *Nucleic Acids Res.* 44, D380–D384. doi: 10.1093/nar/gkv1277

Thomson, J. J., Plecha, S. C., and Withey, J. H. (2015). A small unstructured region in vibrio cholerae ToxT mediates the response to positive and negative effectors and ToxT proteolysis. *J. Bacteriol.* 197, 654–668. doi: 10.1128/JB.02068-14

UniProt, Consortium (2021). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* 49, D480–D489. doi: 10.1093/nar/gkaa1100

Utturkar, S., Dassanayake, A., Nagaraju, S., and Brown, S. D. (2020). Bacterial differential expression analysis methods. *Methods Mol. Biol.* 2096, 89–112. doi: 10.1007/978-1-0716-0195-2_8

Wattam, A. R., Abraham, D., Dalay, O., Disz, T. L., Driscoll, T., Gabbard, J. L., et al. (2014). PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res.* 42, D581–D591. doi: 10.1093/nar/gkt1099

Xia, X., Larios-Valencia, J., Liu, Z., Xiang, F., Kan, B., Wang, H., et al. (2017). OxyR-activated expression of Dps is important for vibrio cholerae oxidative stress resistance and pathogenesis. *PLoS One* 12:e0171201. doi: 10.1371/journal.pone.0171201

Zhang, Z., Chen, G., Hu, J., Hussain, W., Fan, F., Yang, Y., et al. (2019). Mr.vc: a database of microarray and RNA-seq of vibrio cholerae. *Database (Oxford)* 2019:baz069. doi: 10.1093/database/baz069

Zhao, S., Ye, Z., and Stanton, R. (2020). Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols. *RNA* 26, 903–909. doi: 10.1261/rna.074922.120