TITLE: Bacteria.guru: comparative transcriptomics and co-expression database for bacterial pathogens

AUTHORS: Peng Ken Lim¹, Emilia E Davey¹, Sean Wee¹, Wei Song Seetoh¹, Jong Ching Goh¹, Xinghai Zheng¹, Sean Kia Ann Phang¹, Eugene Sheng Kai Seah¹, Janice Wan Zhen Ng¹, Xavier Jia Hui Wee¹, Aloysius Jun Hui Quek¹, Jordan JingHeng Lim¹, Edbert Edric Rodrigues¹, Heesoo Lee¹, Chin Yong Lim¹, Wei Zhi Tan¹, Yuet Ruh Dan¹, Bronson Lee¹, Samuel En Le Chee¹, Zachary Ze En Lim¹, Jia Sheng Guan¹, Ivan Jia Le Tan¹, Trinidad Jeremiah Arong¹, Marek Mutwil^{1*}

¹School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore, 637551, Singapore

*Corresponding author:

Marek Mutwil

School of Biological Sciences,

Nanyang Technological University, 60 Nanyang Drive,

637551, Singapore,

Singapore

Email: mutwil@ntu.edu.sg

Abstract

Summary: The bacterial kingdom comprises unicellular prokaryotes able to establish symbioses from mutualism to parasitism. To combat bacterial pathogenicity, we need an enhanced understanding of gene function and regulation, which will mediate the development of novel antimicrobials. Gene expression can predict gene function, but there lacks a database enabling expansive inter- and intraspecific exploration of gene expression profiles and co-expression networks for bacteria. To address this, we integrated the genomic and transcriptomic data of the 17 most notorious and studied bacterial pathogens, creating bacteria.guru, an interactive database that can identify, visualize, and compare gene expression profiles, co-expression networks, functionally enriched clusters, and gene families across species. Through illustrating antibiotic resistance mechanisms in *P. aeruginosa*, we demonstrate that bacteria.guru could potentially aid the discovery of multi-faceted antibiotic targets. Hence, we believe bacteria.guru will facilitate future bacterial research.

Availability: The database and co-expression networks are freely available from <https://bacteria.guru/>. The sample annotations are found in the supplemental data.

Introduction

Bacteria are ubiquitous unicellular organisms that constitute a diverse kingdom. Despite their fundamental role in the biosphere, the synergistic relationship between the rapid evolution and acute virulence of the pathogenic variety poses the urgency for novel antimicrobials (Davies and Davies, 2010), especially considering the negligent use of antibiotics. The combination of these factors precipitated an epidemic of antibiotic resistance; for example, following the introduction of β -lactams, which inhibit cell wall peptidoglycan synthesis, many bacterial pathogens increased the production of β -lactamases, rendering these medications ineffective. As bacteria continue to evolve these novel resistance mechanisms, our inability to combat them becomes increasingly exposed.

Several methodologies can be used to study genes essential for virulence and antibiotic resistance, such as sequence similarity, gene expression, and co-expression network analyses (Van dam et al., 2018; Jiang et al., 2016). Since predicting gene function from gene sequence alone can result in only a partial or incorrect prediction (Gerlt and Babbitt, 2000; Rhee and Mutwil, 2014), other methods to deduce gene function have emerged. Co-expression analysis is based on the observation that genes which exhibit similar expression profiles across different growth conditions and genotypes are often functionally related (Mutwil et al., 2011), and the utility of co-expression networks has exponentiated our understanding of gene function in multiple kingdoms of life. These networks comprise genes transcriptionally coordinated across various environmental conditions and growth stages, wherein nodes correspond to genes, and edges represent their significant co-expression relationships (Ruan et al., 2010). Co-expression networks are now used to impart insight into the evolution and interspecific conservation of gene modules (Ruprecht et al., 2017), gene function, gene regulation, and subcellular localization of gene products (Mutwil et al., 2011).

Yet, there lacks a platform enabling expansive exploration of inter-and intraspecific gene expression for bacteria. To address this, we present bacteria.guru, a database comprising RNA-sequencing data for 17 notorious bacterial pathogens. In addition to identifying and analyzing co-expression networks and conserved gene modules, bacteria.guru enables the within-and across species analysis of gene expression profiles, gene families, ontology terms, and possible gene-cluster redundancy.

Implementation

Bacteria.guru is based on the CoNekT framework (Proost and Mutwil, 2018). The database supplies gene pages, each containing functional annotations, cDNA and protein sequences, the assigned gene family and corresponding phylogenetic tree, expression profiles, the co-expression neighborhood and cluster, significantly similar neighborhoods, and Gene Ontology

information (<https://bacteria.guru/features>). For a description of the exhaustive construction process of bacteria.guru, refer to the Supplementary Methods.

Figure 1. Antibiotic metabolic process analysis in *Pseudomonas aeruginosa*. A) Average expression profile snippet of cluster 12 for *P. aeruginosa*. The x-axis and y-axis represent different samples (capturing genotype and growth condition) and gene expression levels (as transcript per million), respectively. The bars indicate the average expression, while the points represent the maximum/minimum expression in a given experiment. B) Co-expression cluster 12 from *P. aeruginosa* enriched for GO:0016999 (Antibiotic Metabolic Process). Nodes, colored shapes, and gray edges represent family genes, orthogroups, and co-expressed genes, respectively. For concision, only co-expressed genes involved in the discussed processes are shown. C) Conserved clusters involved in virulence in *Pseudomonas aeruginosa* (cluster 12, left) and *Vibrio cholerae* (cluster 4, right). Dashed edges connect genes belonging to the same orthogroup. D) The donut chart indicates the proportion of genes belonging to *Pseudomonas aeruginosa* (4 genes), *Escherichia coli* (1 gene), *Klebsiella pneumoniae* (1 gene), and *Salmonella enterica* (1 gene) in the OG_02_0001636 orthogroup. E) Gene tree of the

OG_02_0001636 orthogroup (*P. aeruginosa* paralogs comprise the upper clade). The color coding between the donut charts and gene trees is conserved.

To demonstrate how our database can be used to generate novel insights, we use *Pseudomonas aeruginosa* (*P. aeruginosa*) as a case study. Among the most common ICU-acquired infections in critically ill patients (Pachori et al., 2019), the pathogen is not only multidrug-resistant but was also recently discovered to have adaptive resistance capabilities (Taylor et al., 2014). To better comprehend antibiotic resistance in *P. aeruginosa*, 'antibiotic metabolic process' (GO:0016999) was queried using the 'Find enriched clusters' tool (<https://bacteria.sbs.ntu.edu.sg/search/enriched/clusters>), where a cluster containing 238 genes, cluster 12 (<https://bacteria.sbs.ntu.edu.sg/cluster/view/467>), was identified to be significantly enriched (p-value < 0.05). Analysis of Cluster 12 revealed gene members involved in several key processes, which include phenazine synthesis, and quorum sensing (QS), alongside antibiotic metabolism (Fig. 1B), which implicates these processes in *P. aeruginosa* resistance mechanisms. Production of phenazine has been proposed to alter antibiotic susceptibility (Schiessl et al., 2019) and affect swarming motility (Ramos et al., 2010) at early stages of biofilm formation, whereas QS serves as an important regulatory mechanism for coordinating biofilm formation (Subramani et al., 2019). By exploring these interrelations, we could determine whether adaptive antibiotic resistance is achieved through the mechanism of biofilm aggregation. The presence of several proteins related to flagellum-hooking, fimbrial assembly, and cell-adhesion in the cluster, which are essential in biofilm formation (Haiko and Westerlund-Wikström, 2013; Ruer et al., 2007), supports this premise; in fact, Miller and colleagues observed biofilm-induction by aminoglycoside antibiotics in *P. aeruginosa* (Hoffman et al., 2005), demonstrating how our database can be used to unravel potential antibiotic-induced epigenetic changes.

Using the 'similar clusters' tool, one can identify similar clusters in other bacterial species for further analysis. Here we identified and compared a similar cluster in *Vibrio cholerae* to cluster 12, with a Jaccard index of 0.062 (Fig. 1C). To illustrate the database's ability to provide ortholog information on genes, a gene encoding for SdiA, an autoinducer receptor, from cluster 12 (ID: CRO08731) was selected. Clicking on the gene's family (OG_02_0001636) located on the gene page (Fig. 1D), followed by the corresponding phylogenetic tree link, disclosed orthologs found in three other bacterial species (Fig. 1E). Furthermore, the presence of four *sdiA* paralogs denotes a critical role of QS in *P. aeruginosa* antibiotic resistance. With the interplay of QS in formation and regulation of resistance mechanisms necessitating further study (Zhao et al., 2020), additional scrutiny of this biological process overlap could illuminate the nexus of microbial resistance.

Conclusion

The need to better comprehend antibiotic resistance motivated our construction of bacteria.guru. The database offers inter-and intraspecific analyses of co-expressed gene networks and neighbourhoods, expression profiles, gene families, phylograms, and ontology terms in 17 pathogens. We demonstrated how bacteria.guru can be used to study *P. aeruginosa*

antibiotic metabolism, but we envision that its function can be extended to explore any biological pathway in these bacteria.

SUPPLEMENTARY INFORMATION

Supplementary Methods 1. Tools and analyses used to construct bacteria.guru

Supplementary tables

Table S1. Genome versions used to build the database

Table S2. Annotations of the RNA-seq samples used in the study

Citations

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature biotechnology*, 34(5), 525–527.

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature methods*, 12(1), 59–60.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, 10, 421.

Davies, J., & Davies, D. (2010). Origins and evolution of antibiotic resistance. *Microbiology and molecular biology reviews* : MMBR, 74(3), 417–433.

Emms, D. M., & Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome biology*, 16(1), 157.

Gerlt, J. A., & Babbitt, P. C. (2000). Can sequence determine function?. *Genome biology*, 1(5), REVIEWS0005.

Haiko, J., & Westerlund-Wikström, B. (2013). The role of the bacterial flagellum in adhesion and virulence. *Biology*, 2(4), 1242–1267.

Hew, B., Tan, Q. W., Goh, W., Ng, J., & Mutwil, M. (2020). LSTrAP-Crowd: prediction of novel components of bacterial ribosomes with crowd-sourced analysis of RNA sequencing data. *BMC biology*, 18(1), 114.

Hoffman, L. R., D'Argenio, D. A., MacCoss, M. J., Zhang, Z., Jones, R. A., & Miller, S. I. (2005). Aminoglycoside antibiotics induce bacterial biofilm formation. *Nature*, 436(7054), 1171–1175.

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*. *Scientific reports*, 6, 28422.

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S. Y., Lopez, R., & Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics* (Oxford, England), 30(9), 1236–1240.

Mutwil, M., Klie, S., Tohge, T., Giorgi, F. M., Wilkins, O., Campbell, M. M., Fernie, A. R., Usadel, B., Nikoloski, Z., & Persson, S. (2011). PlaNet: combined sequence and expression comparisons across plant networks derived from seven species. *The Plant cell*, 23(3), 895–910.

Mutwil, M., Usadel, B., Schütte, M., Loraine, A., Ebenhöf, O., & Persson, S. (2010). Assembly of an interactive correlation network for the Arabidopsis genome using a novel heuristic clustering algorithm. *Plant physiology*, 152(1), 29–43.

Pachori, P., Gothalwal, R., & Gandhi, P. (2019). Emergence of antibiotic resistance *Pseudomonas aeruginosa* in intensive care unit; a critical review. *Genes & diseases*, 6(2), 109–119.

Price, M. N., Dehal, P. S., & Arkin, A. P. (2009). FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular biology and evolution*, 26(7), 1641–1650.

Proost, S., & Mutwil, M. (2018). CoNekT: an open-source framework for comparative genomic and transcriptomic network analyses. *Nucleic acids research*, 46(W1), W133–W140.

Ramos, I., Dietrich, L. E., Price-Whelan, A., & Newman, D. K. (2010). Phenazines affect biofilm formation by *Pseudomonas aeruginosa* in similar ways at various scales. *Research in microbiology*, 161(3), 187–191.

Rhee, S. Y., & Mutwil, M. (2014). Towards revealing the functions of all genes in plants. *Trends in plant science*, 19(4), 212–221.

Ruan, J., Dean, A. K., & Zhang, W. (2010). A general co-expression network-based approach to gene expression analysis: comparison and applications. *BMC systems biology*, 4, 8.

Ruer, S., Stender, S., Filloux, A., & de Bentzmann, S. (2007). Assembly of fimbrial structures in *Pseudomonas aeruginosa*: functionality and specificity of chaperone-usher machineries. *Journal of bacteriology*, 189(9), 3547–3555.

Ruprecht, C., Vaid, N., Proost, S., Persson, S., & Mutwil, M. (2017). Beyond Genomics: Studying Evolution with Gene Coexpression Networks. *Trends in plant science*, 22(4), 298–307.

Schiessl, K. T., Hu, F., Jo, J., Nazia, S. Z., Wang, B., Price-Whelan, A., Min, W., & Dietrich, L. (2019). Phenazine production promotes antibiotic tolerance and metabolic heterogeneity in *Pseudomonas aeruginosa* biofilms. *Nature communications*, 10(1), 762.

Subramani R., Jayaprakashvel M. (2019) Bacterial Quorum Sensing: Biofilm Formation, Survival Behaviour and Antibiotic Resistance. In: Bramhachari P. (eds) Implication of Quorum Sensing and Biofilm Formation in Medicine, Agriculture and Food Industry. Springer, Singapore.

Taylor, P. K., Yeung, A. T., & Hancock, R. E. (2014). Antibiotic resistance in *Pseudomonas aeruginosa* biofilms: towards the development of novel anti-biofilm therapies. *Journal of biotechnology*, 191, 121–130.

Van Dam, S., Vösa, U., van der Graaf, A., Franke, L., & de Magalhães, J. P. (2018). Gene co-expression analysis for functional classification and gene-disease predictions. *Briefings in bioinformatics*, 19(4), 575–592.

Zhao, X., Yu, Z., & Ding, T. (2020). Quorum-Sensing Regulation of Antimicrobial Resistance in Bacteria. *Microorganisms*, 8(3), 425.

Escherichia coli

Salmonella enterica

Klebsiella pneumoniae

Pseudomonas aeruginosa

- CR008731

CRP58504

CRP78559

CRP50785

AGT23578

AAL20862

- AAC74983