

BIC: a database for the transcriptional landscape of bacteria in cancer

Kai-Pu Chen¹, Chia-Lang Hsu², Yen-Jen Oyang¹, Hsuan-Cheng Huang^{3,*} and Hsueh-Fen Juan^{1,4,5,*}

¹Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei 106, Taiwan, ²Department of Medical Research, National Taiwan University Hospital, Taipei 100, Taiwan, ³Institute of Biomedical Informatics, National Yang Ming Chiao Tung University, Taipei 112, Taiwan, ⁴Department of Life Science, National Taiwan University, Taipei 106, Taiwan and ⁵Center for Computational and Systems Biology, National Taiwan University, Taipei 106, Taiwan

Received August 13, 2022; Revised September 27, 2022; Editorial Decision September 28, 2022; Accepted October 03, 2022

ABSTRACT

Microbial communities are massively resident in the human body, yet dysbiosis has been reported to correlate with many diseases, including various cancers. Most studies focus on the gut microbiome, while the bacteria that participate in tumor microenvironments on site remain unclear. Previous studies have acquired the bacteria expression profiles from RNA-seq, whole genome sequencing, and whole exon sequencing in The Cancer Genome Atlas (TCGA). However, small-RNA sequencing data were rarely used. Using TCGA miRNA sequencing data, we evaluated bacterial abundance in 32 types of cancer. To uncover the bacteria involved in cancer, we applied an analytical process to align unmapped human reads to bacterial references and developed the BIC database for the transcriptional landscape of bacteria in cancer. BIC provides cancer-associated bacterial information, including the relative abundance of bacteria, bacterial diversity, associations with clinical relevance, the co-expression network of bacteria and human genes, and their associated biological functions. These results can complement previously published databases. Users can easily download the result plots and tables, or download the bacterial abundance matrix for further analyses. In summary, BIC can provide information on cancer microenvironments related to microbial communities. BIC is available at: <http://bic.jhlab.tw/>.

INTRODUCTION

The human microbiota massively lives, varies in our bodies, and is diverse in different body sides (1,2). It was esti-

mated that a human body harbors more than three trillion bacterial members, similar to the number of human cells (3). Host–microbiome interactions impact multiple physiological processes and disease susceptibilities. The human microbiota plays an important role in human health, such as maintaining homeostasis, immunity and inflammation (4,5). Most microbial studies focus on the gut microbiome and related diseases, such as inflammatory bowel disease (IBD) and depression and anxiety (6). Furthermore, studies have shown that the microbial compositions are different and associated with cancer (7,8).

While many studies focus on the gut microbiome derived from patients' stool (9–11), the bacteria that participate in the on-site tumor microenvironments remain unclear. Dohlman *et al.* and Poore *et al.* have acquired the bacteria expression profiles from RNA-seq, whole genome sequencing (WGS), and whole exon sequencing (WXS) in The Cancer Genome Atlas (TCGA) (12,13). However, the small-RNA sequencing data are not used. We developed an analytical approach using the small-RNA sequencing data of colorectal cancer (CRC) tissue samples to study cancer-associated microbiome in CRC and observed similar results compared to other studies using 16S rDNA sequencing (14).

There are certain benefits in using miRNA-seq compared to WGS, WXS, and RNA-seq. First, small RNAs (sRNAs) have been found to play regulatory roles in both bacteria and bacterial infectious diseases (15,16). Compared to WGS and WXS, sRNAs were transcribed and functional in either bacteria or hosts. Only a small fraction of total RNA was polyadenylated and appeared transiently in bacteria (17,18). In many RNA-seq studies, RNAs were extracted and reverse-transcribed to cDNAs through poly-A tails. Most bacterial RNAs without poly-A tails will be filtered in RNA-seq data. Compared to RNA-seq, miRNA-seq which is processed without poly-A filter-

*To whom correspondence should be addressed. Tel: +886 2 3366 4536; Email: yukijuan@ntu.edu.tw
Correspondence may also be addressed to Hsuan-Cheng Huang. Tel: +886 2 2826 7357; Email: hsuancheng@nycu.edu.tw

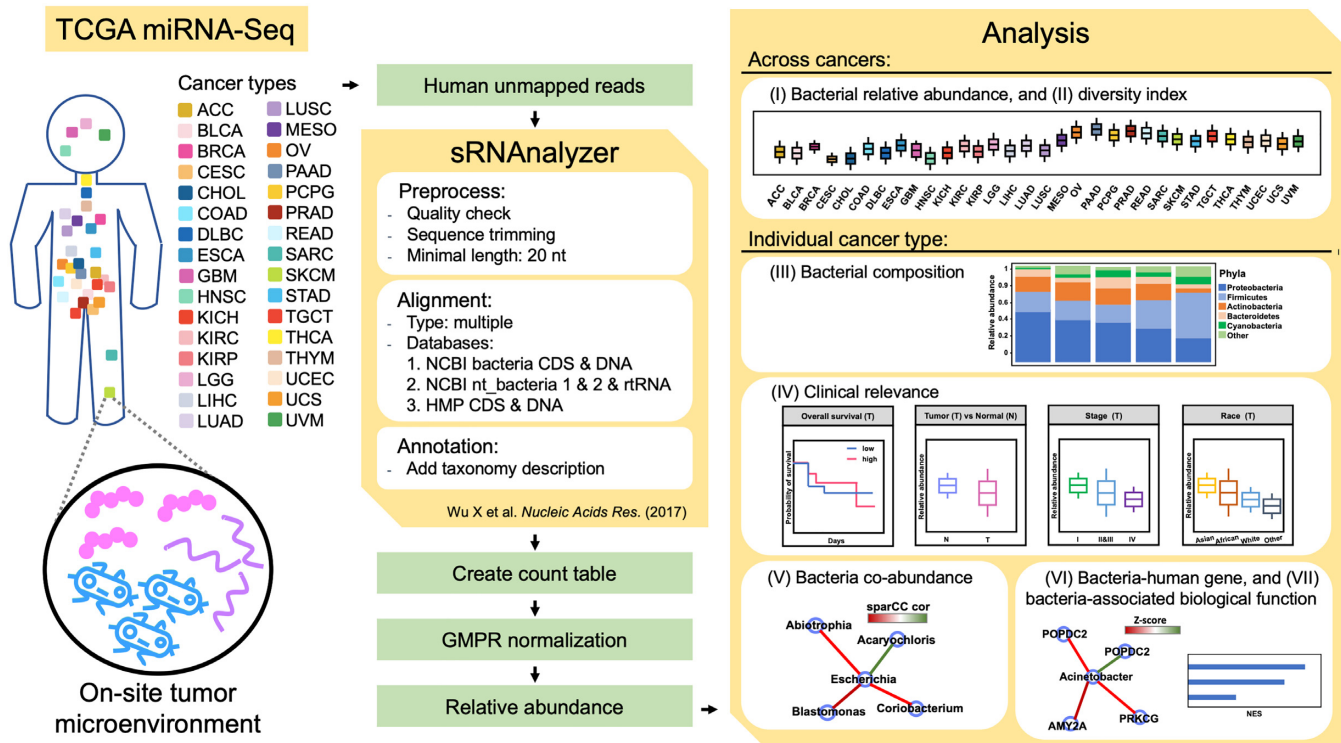


Figure 1. An overview of BIC analysis workflow. We downloaded miRNA-seq data from TCGA and used sRNAAnalyzer for read processing. We parsed and merged count tables, conducted GMPR normalization, and produced the bacterial relative abundance matrices of each taxonomic level in our scripts. We provide seven modules in the BIC Analyses panel for users to query and download results.

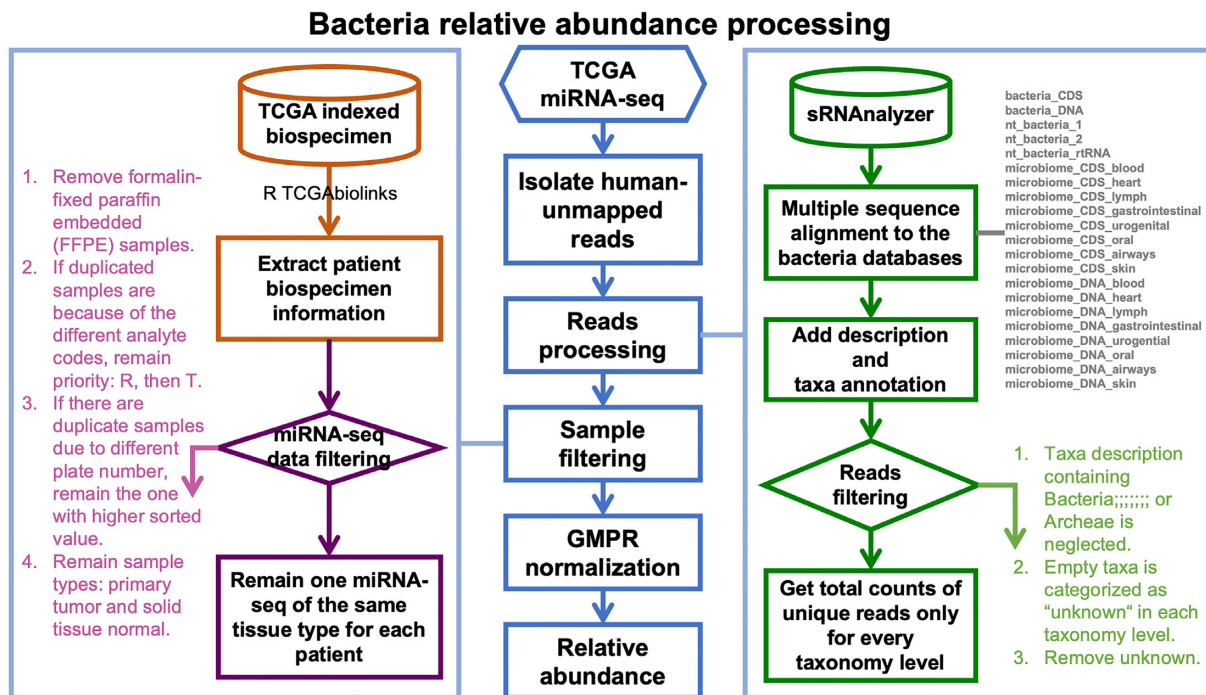


Figure 2. Workflow of data processes to produce bacterial relative abundance matrices.

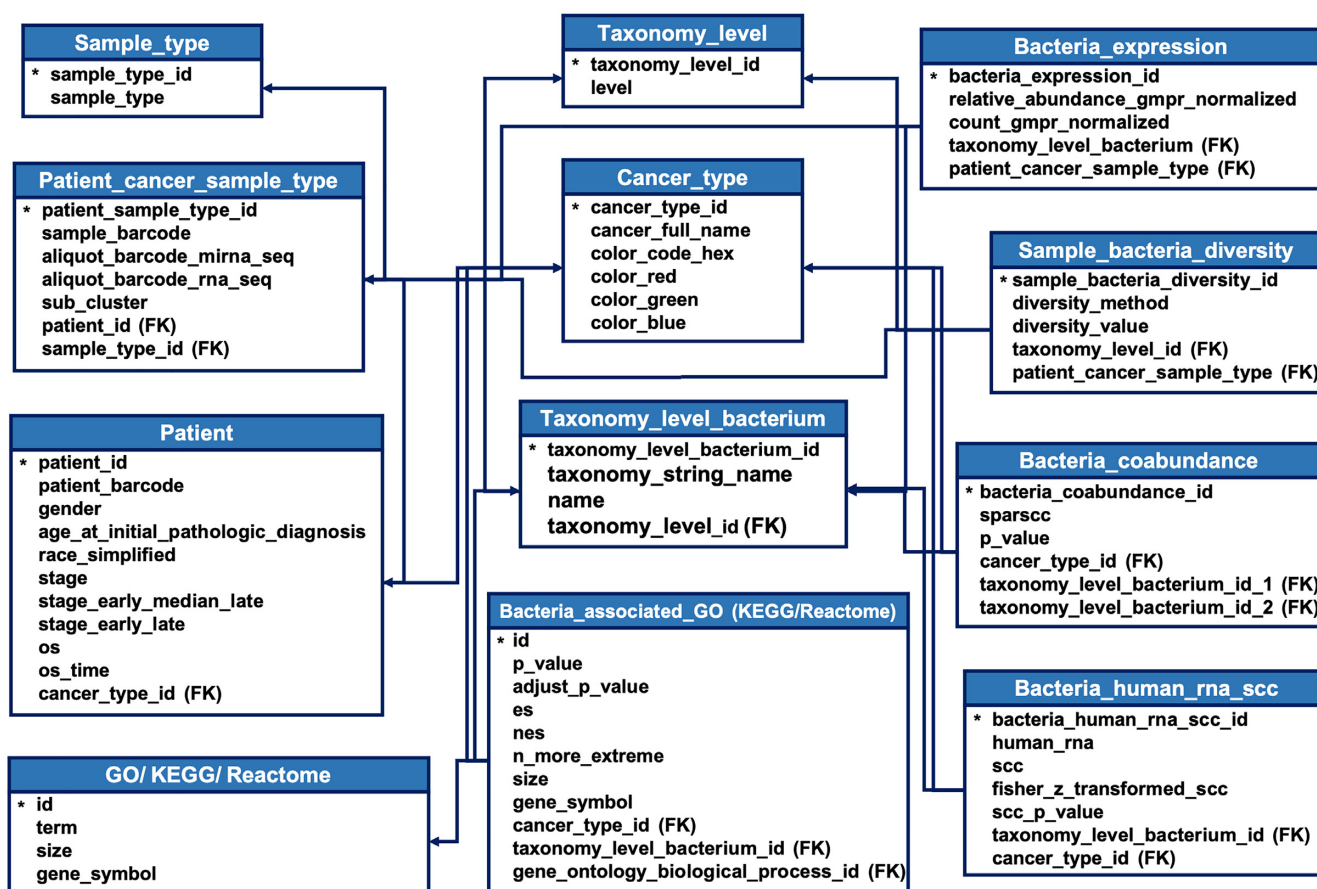


Figure 3. Data tables saved in PostgreSQL. All the precomputed analysis results were saved in PostgreSQL and the primary key of each table was labeled with a star symbol (*).

ing could have a chance to identify bacteria not found in RNA-seq.

Using TCGA miRNA sequencing data, we evaluated tissue-resident bacterial abundance in 32 types of cancer. We aligned unmapped human reads to bacterial references by sRNAAnalyzer and merged them for each taxonomic rank of 32 cancer types (14,19). The bacterial relative abundance and sample diversity were compared across different cancer types. We parsed all the data and developed the BIC database for the transcriptional landscape of bacteria in cancer. BIC provides the following information: (i) relative abundance of bacteria, (ii) bacterial diversity, (iii) bacterial composition, (iv) clinical relevance, (v) bacterial coabundance network, (vi) bacteria-correlated human gene expression network and (vii) bacteria-associated biological function (Figure 1). Users can easily query and browse the analysis plots and result tables, or download the bacterial expression matrices for further analyses.

DATA COLLECTION

The TCGA miRNA-seq BAM files were retrieved from the NCI Genomic Data Commons (GDC) using the GDC Data Transfer Tool (20). Human RNA expression profiles (EBPlusPlusAdjustPAN-CAN_IlluminaHiSeq_RNASeqV2.geneExp.tsv), tumor

stages, races, survival events, and time (TCGA-CDR-Supplemental Table S1.xlsx) were downloaded from the Supplemental Data in PanCanAtlas Publications (<https://gdc.cancer.gov/about-data/publications/pancanatlas>).

The gene symbols in the RNA expression profiles were renamed according to org.Hs.eg.db (version 3.6.0) (21). Only samples from primary tumors and their adjacent normal tissues were used. We acquired the biospecimen information using TCGAbiolinks (version 2.17.3) (22).

DATA PROCESSING AND INTEGRATION

Bacteria relative abundance matrixes

We used SAMtools (version 1.3.1) to extract the unmapped reads from human miRNA-seq BAM files and stored them in FASTQ format (23). sRNAAnalyzer scripts ('pre-process.pl', 'align.pl', 'desProfile.pl' and 'taxProfile.pl') were used for read preprocessing, alignment, taxonomy annotation (19). We set the minimal read length to 20 nucleotides and mapped the reads to multiple references, but did not allow any mismatch to obtain the highest alignment accuracy. The references used in alignment were provided by sRNAAnalyzer, including CDS and DNA of bacteria, nt_bacteria, and microbiomes. After taxonomy annotation by the sRNAAnalyzer scripts, we reassigned the reads mapped to multi-

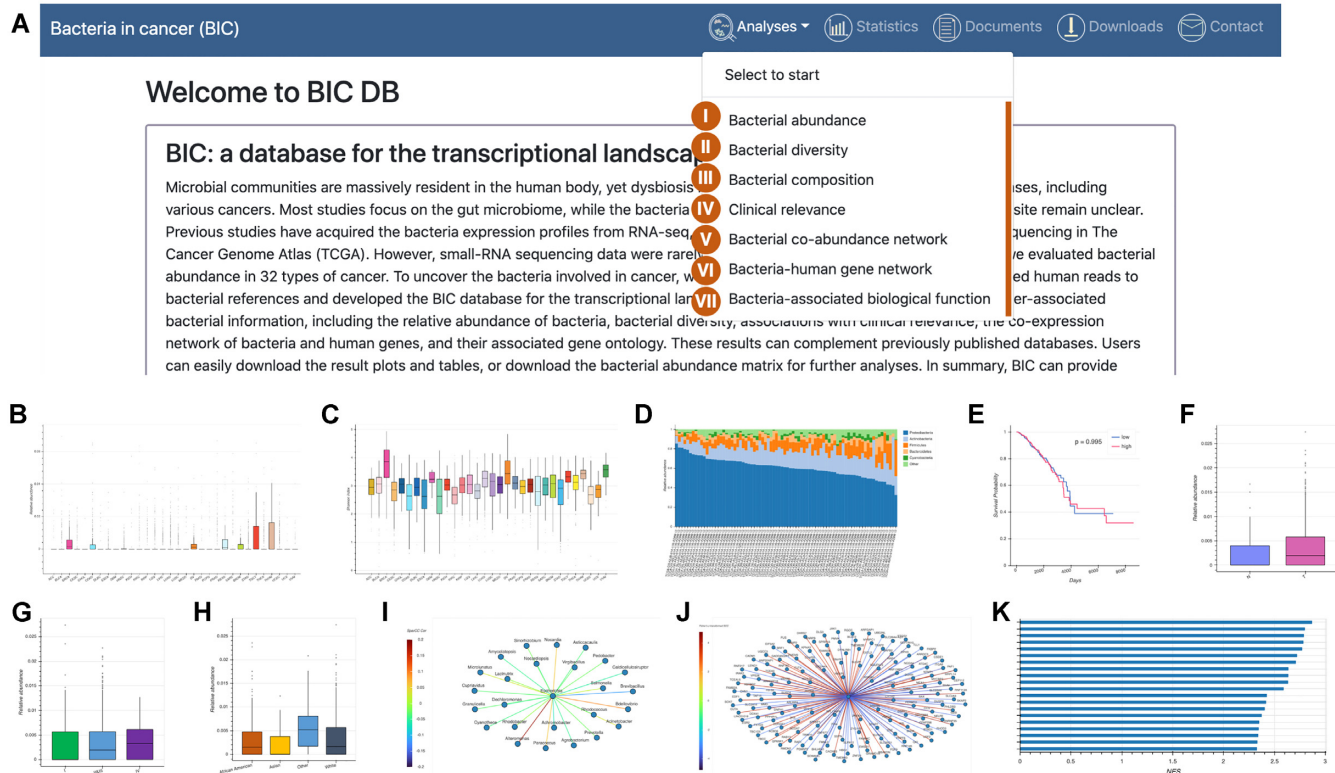


Figure 4. BIC user interface and analysis modules. (A) The user interface and analysis modules. (B, C) Modules I and II show the relative abundance and diversity of bacteria across cancers. (D) Module III shows the bacteria composition. (E–H) Module IV shows the clinical relevance, such as overall survival, and relative abundance compared in different groups (tumor versus adjacency normal; tumor stages; races). (I) Module V shows the bacteria co-abundance network. (J) Module VI shows the bacteria-correlated human gene expression network. (K) Module VII shows the bacteria-associated biological functions.

ple species to their common higher-level taxa and generated the read matrixes at different taxonomic levels (14). The processed read counts in each data processing step are summarized in Supplementary Figure S1. We identified 1617 genera, 303 families, 126 orders, 56 classes, and 47 phyla from 10362 samples (9709 patients) across 32 cancer types. Since the count matrixes were sparse, we applied the geometric mean of pairwise ratio (GMPR), a robust normalization method for zero-inflated data, to produce normalized count tables (24). To keep all 10 362 samples, the intersection numbers of the phylum, class, order, family, and genus count tables were set to 3, 5, 5, 6 and 5, respectively. Normalized count matrices were transformed into relative abundance matrices. The relative bacterial abundance of each taxonomy level was used for all subsequent analyses in BIC. An overview of these processes is shown in Figure 2. Detailed bacterial references and processing scripts are available on GitHub.

Precomputed analysis data and database construction

Based on bacterial relative abundances, we calculated the bacterial diversity in each taxonomy level for every kind of cancer. Vegan (version 2.5–7, <https://CRAN.R-project.org/package=vegan>) was used to calculate the Shannon, Gini-Simpson, and inverse Simpson indices (25). Bacteria with a prevalence (nonzero count) of $\geq 20\%$ in the

individual type of cancer were used to analyze the co-abundance of bacteria, the correlation with host gene expression, and the associated biological function. We applied SparCC, a method designed for compositional data, to calculate bacterial co-abundance relationships and establish the co-abundance networks for individual cancer types (26,27). The function `sparccboot` in `SpiecEasi` (version 1.1.0) was used to acquire SparCC correlation coefficients and empirical p-values of the bacterial co-abundance with 10 000 times of bootstraps (28). Spearman correlation coefficients (SCC) were calculated for the bacterial correlation with human gene expression using common samples between bacteria and tissue transcriptome data. Only human genes that were measured with nonzero counts in $\geq 20\%$ of the samples were considered. To correct for the sample size effect, we applied Fisher's z-transformation for SCC. To reveal the possible biological processes in which the queried bacteria are involved, we performed gene set enrichment analysis (FGSEA, version 1.12.0) for bacteria-correlated gene expression ranked in the descending order of the corrected z-score. The gene sets of biological processes annotated by Gene Ontology (GO, [c5.go.bp.v7.2](https://www.ebi.ac.uk/ontology/ontologies/go)) (29), KEGG ([c2.cp.kegg.v7.5.1](https://www.genome.jp/kegg/)) (30) and Reactome ([c2.cp.reactome.v7.5.1](https://www.ebi.ac.uk/ontology/ontologies/reactome)) (31) were downloaded from the Molecular Signatures Database (32,33). These analyses were performed with R scripts (version 3.6.0) (34) and all the precomputed analysis data are stored in Post-

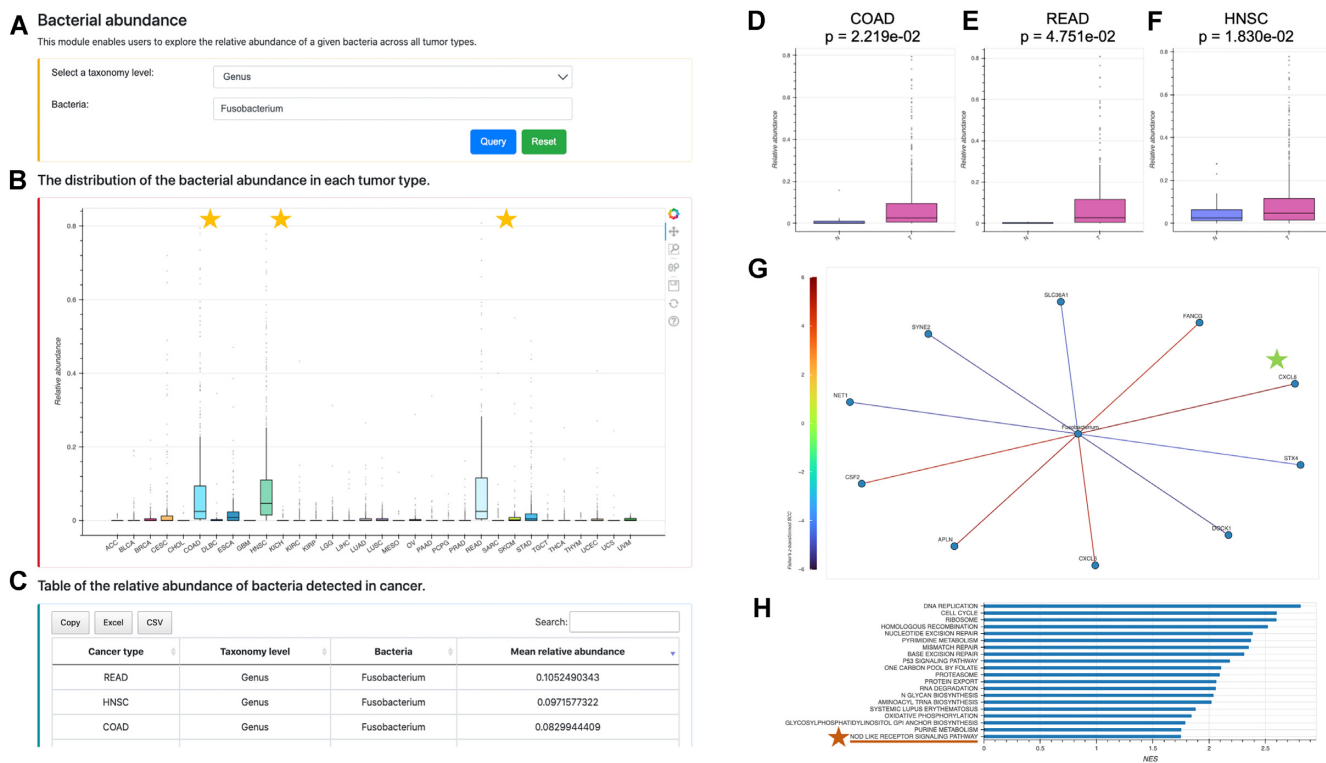


Figure 5. Query examples of the analysis modules in BIC. (A–C) Query of *Fusobacterium* at the genus level and the output distribution plot and tables for the abundance of *Fusobacterium* across cancer types in the bacteria abundance module. (D–F) The relative abundance of *Fusobacterium* in tumor is significantly higher than in the adjacent normal tissues in COAD, READ and HNSC with the clinical relevance module. (G) The bacteria-human gene network module displays the z -scores between *Fusobacterium* and the top 10 correlated genes in COAD. (H) The bacteria-associated biological function module displays the top 20 *Fusobacterium*-associated KEGG pathways in COAD.

greSQL (version 13.3). The tables deposited in PostgreSQL are shown in Figure 3.

Web application framework

The BIC web application framework (Supplementary Figure S2) was constructed using Python (version 3.6.8) (35) and Django (version 3.2.3, <https://djangoproject.com>). The analyses of clinical relevance were performed under Django, including overall survival and bacterial abundance comparison of different groups, such as tumor (T) versus adjacent normal (N), tumor stages and races. The survival analysis was implemented using lifelines (version 0.26.3) (36). Calculations of statistical P -values (Kruskal–Wallis and Wilcoxon ranksum tests) in different groups were implemented by `kruskal` and `ranksums` in `scipy` (version 1.5.4) (37). Plots were produced by `bokeh` (version 2.3.3, <http://www.bokeh.pydata.org>).

USER INTERFACE AND USE CASES

Figure 4 shows the user interface and all the analyses provided by BIC. Modules I and II enable users to query the bacterial relative abundance and diversity indexes or evenness of the selected taxonomy level across all cancer types. Modules III to VII allow users to find the bacterial composition, clinical relevance, co-abundance, correlated human

gene expression, and inferred biological processes of the queried bacteria under specified taxonomy level of the selected cancer type. Users can easily save the output plots and tables for their queried analyses.

Figure 5 illustrates an example of how users can investigate the genus *Fusobacterium* in cancer. For CRC and head and neck cancer, *Fusobacterium* is known to be associated with cancer progression (38,39). With the Bacterial abundance module, users can query *Fusobacterium* at the genus level (Figure 5A) and observe that the relative abundances of *Fusobacterium* are remarkably high in COAD (colon adenocarcinoma), READ (rectum adenocarcinoma), and HNSC (head and neck squamous cell carcinoma) (Figure 5B, C). Furthermore, the Clinical relevance module shows that *Fusobacterium* is more abundant in tumor than in adjacent normal tissues in these three types of cancer (Figure 5D–F). In the Bacteria-human gene network module, *CXCL8* is the top gene positively correlated with *Fusobacterium* in COAD (Figure 5G). *CXCL8* has been found to play an important role in CRC (40,41). With the Bacteria-associated biological function module, users can view the most significant KEGG pathways correlated with the abundance of *Fusobacterium* in COAD (Figure 5H). Among many cancer-related pathways, the NOD-like receptor signaling pathway has previously been reported to be related to the onset of CRC (40,41).

CONCLUSION

We have developed a user-friendly database, BIC, for bacterial profiles derived from TCGA miRNA-seq data in 32 types of cancer. BIC allows comparisons of the relative abundance and diversities of bacteria in different types of cancer. BIC also provides the bacterial composition, clinical relevance, co-abundance network, correlated human gene expression network, and associated gene ontologies, for each type of cancer. With the comprehensive characterization of bacteria in tissues of different cancers, BIC can greatly facilitate the exploration of bacterial functions and mechanisms in tumor microenvironments. We believe that our database will be a valuable resource for understanding the interactions between humans and microbes in cancer formation.

DATA AVAILABILITY

BIC is freely accessible at: <http://bic.jhlab.tw/>. The entire BIC data collection can be downloaded from the website. The source codes of BIC data processing, database construction, and web application are available at GitHub https://github.com/Kai-Pu/BIC_production.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Yue-Hua Tu for advice and discussion of the BIC web framework. We thank the National Center for High-performance Computing (NCHC) for providing computational and storage resources.

FUNDING

Ministry of Science and Technology, Taiwan [MOST 109-2221-E-002-161-MY3, 109-2320-B-002-017-MY3, 109-2221-E-010-011-MY3]; Ministry of Education (the Higher Education Sprout Project) [NTU-110L8808, NTU-CC-109L104702-2]. Funding for open access charge: Ministry of Science and Technology, Taiwan.
Conflict of interest statement. None declared.

REFERENCES

- Costello, E.K., Lauber, C.L., Hamady, M., Fierer, N., Gordon, J.I. and Knight, R. (2009) Bacterial community variation in human body habitats across space and time. *Science*, **326**, 1694–1697.
- Cho, I. and Blaser, M.J. (2012) The human microbiome: at the interface of health and disease. *Nat. Rev. Genet.*, **13**, 260–270.
- Sender, R., Fuchs, S. and Milo, R. (2016) Revised estimates for the number of human and bacteria cells in the body. *PLoS Biol.*, **14**, e1002533.
- Belkaid, Y. and Hand, T.W. (2014) Role of the microbiota in immunity and inflammation. *Cell*, **157**, 121–141.
- Lee, J.Y., Tsolis, R.M. and Baumber, A.J. (2022) The microbiome and gut homeostasis. *Science*, **377**, eabp9960.
- Foster, J.A. and McVey Neufeld, K.A. (2013) Gut-brain axis: how the microbiome influences anxiety and depression. *Trends Neurosci.*, **36**, 305–312.
- Arthur, J.C., Perez-Chanona, E., Muhlbauer, M., Tomkovich, S., Uronis, J.M., Fan, T.J., Campbell, B.J., Abujamel, T., Dogan, B., Rogers, A.B. *et al.* (2012) Intestinal inflammation targets cancer-inducing activity of the microbiota. *Science*, **338**, 120–123.
- Urbaniak, C., Gloor, G.B., Brackstone, M., Scott, L., Tangney, M. and Reid, G. (2016) The microbiota of breast tissue and its association with breast cancer. *Appl. Environ. Microbiol.*, **82**, 5039–5048.
- Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D. *et al.* (2012) A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, **490**, 55–60.
- Amato, K.R., Arrieta, M.C., Azad, M.B., Bailey, M.T., Brossard, J.L., Bruggeling, C.E., Claud, E.C., Costello, E.K., Davenport, E.R., Dutilh, B.E. *et al.* (2021) The human gut microbiome and health inequities. *Proc. Natl. Acad. Sci. U.S.A.*, **118**, e2017947118.
- Henrick, B.M., Rodriguez, L., Lakshminanth, T., Pou, C., Henckel, E., Arzoomand, A., Olin, A., Wang, J., Mikes, J., Tan, Z. *et al.* (2021) Bifidobacteria-mediated immune system imprinting early in life. *Cell*, **184**, 3884–3898.
- Dohlman, A.B., Arguijo Mendoza, D., Ding, S., Gao, M., Dressman, H., Iliev, I.D., Lipkin, S.M. and Shen, X. (2021) The cancer microbiome atlas: a pan-cancer comparative analysis to distinguish tissue-resident microbiota from contaminants. *Cell Host Microbe*, **29**, 281–298.
- Poore, G.D., Kopylova, E., Zhu, Q., Carpenter, C., Fraccio, S., Wandro, S., Kosciolk, T., Janssen, S., Metcalf, J., Song, S.J. *et al.* (2020) Microbiome analyses of blood and tissues suggest cancer diagnostic approach. *Nature*, **579**, 567–574.
- Lee, W.H., Chen, K.P., Wang, K., Huang, H.C. and Juan, H.F. (2020) Characterizing the cancer-associated microbiome with small RNA sequencing data. *Biochem. Biophys. Res. Commun.*, **522**, 776–782.
- Storz, G., Vogel, J. and Wassarman, K.M. (2011) Regulation by small RNAs in bacteria: expanding frontiers. *Mol. Cell*, **43**, 880–891.
- Gonzalez Plaza, J.J. (2020) Small RNAs as fundamental players in the transference of information during bacterial infectious diseases. *Front. Mol. Biosci.*, **7**, 101.
- Sarkar, N. (1996) Polyadenylation of mRNA in bacteria. *Microbiology (Reading)*, **142**, 3125–3133.
- Hajnsdorf, E. and Kaberdin, V.R. (2018) RNA polyadenylation and its consequences in prokaryotes. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **373**, 20180166.
- Wu, X., Kim, T.K., Baxter, D., Scherler, K., Gordon, A., Fong, O., Etheridge, A., Galas, D.J. and Wang, K. (2017) sRNAAnalyzer—a flexible and customizable small RNA sequencing data analysis pipeline. *Nucleic Acids Res.*, **45**, 12140–12151.
- Grossman, R.L., Heath, A.P., Ferretti, V., Varmus, H.E., Lowy, D.R., Kibbe, W.A. and Staudt, L.M. (2016) Toward a shared vision for cancer genomic data. *N. Engl. J. Med.*, **375**, 1109–1112.
- Carls, M. (2019) org.Hs.eg.db: Genome wide annotation for Human. R package version 3.10.10.
- Colaprico, A., Silva, T.C., Olsen, C., Garofano, L., Cava, C., Garolini, D., Sabedot, T.S., Malta, T.M., Pagnotta, S.M., Castiglioni, I. *et al.* (2016) TCGAAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.*, **44**, e71.
- Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M. *et al.* (2021) Twelve years of SAMtools and BCftools. *Gigascience*, **10**, giab008.
- Chen, L., Reeve, J., Zhang, L., Huang, S., Wang, X. and Chen, J. (2018) GMPR: a robust normalization method for zero-inflated count data with application to microbiome sequencing data. *PeerJ*, **6**, e4600.
- Oksanen, J., Guillaume Blanchet, J., Friendly, M., Kindt, R., Legendre, P., McGlenn, D., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P. *et al.* (2020) vegan: Community Ecology Package. R package version 2.5-7.
- Chen, L., Collij, V., Jaeger, M., van den Munckhof, I.C.L., Vich Vila, A., Kurilshikov, A., Gacesa, R., Sinha, T., Oosting, M., Joosten, L.A.B. *et al.* (2020) Gut microbial co-abundance networks show specificity in inflammatory bowel disease and obesity. *Nat. Commun.*, **11**, 4018.
- Friedman, J. and Alm, E.J. (2012) Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.*, **8**, e1002687.
- Kurtz, Z.D., Müller, C.L., Miraldi, E.R., Littman, D.R., Blaser, M.J. and Bonneau, R.A. (2021) SpiecEasi: Sparse Inverse Covariance for Ecological Statistical Inference. R package version 1.1.0.

29. Gene Ontology Consortium (2021) The gene ontology resource: enriching a GOLD mine. *Nucleic Acids Res.*, **49**, D325–D334.
30. Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M. and Tanabe, M. (2021) KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res.*, **49**, D545–D551.
31. Gillespie, M., Jassal, B., Stephan, R., Milacic, M., Rothfels, K., Senff-Ribeiro, A., Griss, J., Sevilla, C., Matthews, L., Gong, C. *et al.* (2022) The reactome pathway knowledgebase 2022. *Nucleic Acids Res.*, **50**, D687–D692.
32. Sergushichev, A. A. (2016) An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. bioRxiv doi: <https://doi.org/10.1101/060012>, 20 June 2016, preprint: not peer reviewed.
33. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 15545–15550.
34. R. C. Team (2019) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
35. Van Rossum, G. and Fred, L. (1995) *Python Reference Manual*. Centrum voor Wiskunde en Informatica Amsterdam.
36. Davidson-Pilon, C. (2019) lifelines: survival analysis in python. *J. Open Source Softw.*, **4**, 1317.
37. Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J. *et al.* (2020) SciPy 1.0: fundamental algorithms for scientific computing in python. *Nat. Methods*, **17**, 261–272.
38. Wu, J., Li, Q. and Fu, X. (2019) *Fusobacterium nucleatum* contributes to the carcinogenesis of colorectal cancer by inducing inflammation and suppressing host immunity. *Transl. Oncol.*, **12**, 846–851.
39. Bronzato, J. D., Bomfim, R. A., Edwards, D. H., Crouch, D., Hector, M. P. and Gomes, B. (2020) Detection of fusobacterium in oral and head and neck cancer samples: a systematic review and meta-analysis. *Arch. Oral. Biol.*, **112**, 104669.
40. Bie, Y., Ge, W., Yang, Z., Cheng, X., Zhao, Z., Li, S., Wang, W., Wang, Y., Zhao, X., Yin, Z. *et al.* (2019) The crucial role of CXCL8 and its receptors in colorectal liver metastasis. *Dis. Markers*, **2019**, 8023460.
41. Velloso, F. J., Trombetta-Lima, M., Anschau, V., Sogayar, M. C. and Correa, R. G. (2019) NOD-like receptors: major players (and targets) in the interface between innate immunity and cancer. *Biosci. Rep.*, **39**, BSR20181709.