Article

Not peer-reviewed version

Computational Identification of Specific Genes and Those Functions on the Gut Microbiota

Kazuki Arita , Yoshio Nakano , Satoka Kasai , Kazumi Yoshizawa , Satoru Miyazaki *

Posted Date: 14 January 2025

doi: 10.20944/preprints202501.1007.v1

Keywords: 16S rRNA amplicon analysis; microbiota; M. schaedleri; Sequence homology; PSI-BLAST; gene neighborhood; gene functional prediction; Peptidase M23; USV; hyperalgesia

Article

Computational Identification of Specific Genes and Those Functions on the Gut Microbiota

Kazuki Arita [1], Yoshio Nakano [1], Satoka Kasai [2], Kazumi Yoshizawa [2] and Satoru Miyazaki [1,*]

[1] Laboratory of Bioinformatics, Department of Pharmacy, Faculty of Pharmaceutical Sciences, Tokyo University of Science, Chiba 278–8501, Japan

[2] Laboratory of Pharmacology and Therapeutics, Department of Pharmacy, Faculty of Pharmaceutical Sciences, Tokyo University of Science, Chiba 278–8501, Japan

* Correspondence: Author: smiyazak@rs.tus.ac.jp Tel.: 04-7121-3630

Abstract: The gut microbiota interacts with the brain via the intestines and is related to diseases such as depression. Metagenome analysis, which measures bacterial genes, is commonly used in the analysis of bacterial flora. However, only a small portion of bacterial genes have known functions, and most have unknown functions, so the information is insufficient with existing analysis methods. Therefore, we developed an analysis method that combines "16S rRNA amplicon analysis data" and "prediction information obtained by database search" to enable the analysis of genes with unknown functions. Using this method, we were able to add information to the gene of bacteria with unknown functions and show part of the mechanism by which intestinal bacteria affect mouse diseases. We applied this method to the intestinal flora of mice that show hyperalgesia due to ultrasound irradiation. *M. scheadleri* was found to be increased in the ultrasound-irradiation group (USV). Adapting the analysis method to the *M. scheadleri* genome, we were able to predict the function of proteins specifically produced by *M. scheadleri*. The specifically produced protein may have the function of Peptidase M23 in addition to the function related to the membrane obtained by the usual search.

Keywords: 16S rRNA amplicon analysis; microbiota; *M. schaedleri*; Sequence homology; PSI-BLAST; gene neighborhood; gene functional prediction; Peptidase M23; USV; hyperalgesia;

1. Introduction

Humans coexist with a diverse array of bacteria, forming a complex microbial community that has co-evolved with its host over millennia [1]. These symbiotic bacteria are instrumental in maintaining human health, particularly within the digestive system, where they contribute to metabolic processes and the maturation of the immune system. The relationship between humans and their microbiota is characterized by intricate interactions that sustain homeostasis and confer physiological advantages. However, disturbances to this delicate equilibrium may result in a range of immune, inflammatory, and metabolic disorders[2]. The extensive genetic diversity observed among bacteria, which significantly surpasses the genetic variation found within the human population, carries important implications for human health and disease. An understanding of bacterial polymorphisms and their effects on the human host provides valuable insights into pathogens, cellular processes, and potential therapeutic interventions [3].

So far, the two principal methodologies, amplification sequencing, and shotgun sequencing, that have been employed to analyze the intestinal microbiota and all DNA present in the sample. On the other hand, there are some limitations to shotgun sequencing. High costs and data scalability are major issues[4]. When assigning genes to specific bacterial taxa, a significant barrier arises when the query sequence lacks a corresponding reference in existing databases. This absence of taxonomic context complicates the classification of metagenomic data [5]. Limitations of short read lengths,

 doi:10.20944/preprints202501.1007.v1

sequencing errors, and lack of taxonomically related reference sequences make metagenomic shotgun sequence classification unsuccessful. [6]

As indicated by prior research, PICRUSt2 serves as a predictive metagenomic analysis tool that estimates the constituent genes based on amplicon analysis data. This software predicts the functional composition of metagenomes utilizing 16S rRNA gene sequences and reference genome databases, allowing for accurate estimations of gene family abundances across various microbial communities[7]. Tax4Fun2 is an R package that predicts functional profiles and functional gene redundancy in prokaryotic communities from 16S rRNA gene sequences[8]. However, this method is unable to yield results for genes with unknown functions and lacks sensitivity regarding the reliable detection of functional changes associated with variations in bacterial populations. It is estimated that a significant proportion of proteins within prokaryotic genomes—ranging from 20% to 50%—remain uncharacterized.

Various computational methods have been developed to predict the functions of hypothetical proteins in microbial genomes. These approaches integrate multiple bioinformatics tools and databases to improve prediction accuracy. Common techniques include sequence similarity searches using BLAST, domain identification with InterProScan and Pfam, and ortholog detection using COGs[9–11]. Functional interaction networks have also been employed to predict protein functions and analyze their properties [12]. These methods have been applied to predict functions for hypothetical proteins in important pathogens like *Mycobacterium tuberculosis* and *Vibrio parahaemolyticus*, revealing potential roles in virulence, pathogenicity, and host adaptation[11,12].

In response to these limitations, we have developed a novel methodology for predicting the functions of bacterial proteins by integrating database searches for functional annotation with amplicon analysis. This approach enables the prediction of functions for proteins of unknown function by considering sequence homology and protein interactions. Furthermore, it can anticipate the implications of variations in bacterial populations by assessing the predicted functions of identified bacteria. Previous studies have suggested that exposure to ultrasound can increase and decrease bacteria[13]. It is also known that mice use ultrasonic vocalizations to communicate their emotional states and to engage in social interactions.[14] In this study, we also applied our novel methodology to analyze the changes in the intestinal flora of mice that exhibited heightened pain sensitivity due to exposure to ultrasonic frequencies.

2. Materials and Methods

2.1. Sample Preparation

2.1.1. Animals

Adult C57BL/6N mice (male, 7 weeks old) were obtained from Japan SLC (Shizuoka, Japan) and used for behavioral and biochemical studies. The mice were housed in the Laboratory Animal Research Center at Tokyo University of Science under standard conditions (23°C ± 1°C, 12-h light/dark cycle, lights on at 8:00 a.m.). The mice were housed in standard mouse cages with sawdust bedding and had free access to food and water. This study was conducted by the "Basic Guidelines for the Implementation of Animal Experiments at Academic Research Institutions, etc." under the jurisdiction of the Ministry of Education, Culture, Sports, Science and Technology. All protocols were approved by the Tokyo University of Science Animal Experiment Committee (Approval Nos. Y21036 and Y22028).

2.1.2. Recordings of Cries

The mouse was fixed with a fixation device, and the mouse's tail was pinched with a 5 cm long artery clip to record its cries (or vocalizations) during pain responses. Its cries were recorded with an UltraSoundGate (Avisoft Bioacoustics, Nordbahn, Germany). The ultrasonic regions (over 20 kHz)

 doi:10.20944/preprints202501.1007.v1

of the obtained sound sources were used in this experiment as psychological stress sound sources (sound stress).

2.1.3. Von Frey Test

Naive mice, different from those used for the recordings of cries, were exposed to sound stress at 80 dB for 4h (day 0) in a soundproof box, and tactile thresholds were measured after 1, 3, and 7 days. The von Frey filaments evaluated the sound stress-induced hyperalgesia.

2.1.4. 16S rRNA Amplicon Analysis

The contents of the large intestine of mice exposed to acoustic stress for 4 hours a day were collected and subjected to 16S rRNA amplicon analysis (Illumina 16S Metagenomic Sequencing Library Preparation Protocol[15]). 16S rRNA amplicon was conducted by Hokkaido System Science Co. (Hokkaido, Japan).

2.2. New Application Methods to Identify the Specific Genes from Amplicon Data

2.2.1. Getting Data from NCBI RefSeq and Creating a Database

Using the NCBI Datasets command line tool (CLI) v12[16], we obtained the FAA file of type strain data for all bacteria from NCBI RefSeq[17]. After removing duplicate data, we created a database via BLAST+[18]. We have created a database (BSDB0 to BSDB9) that divides these data into 10 parts. We used data obtained on December 6th, 2024. We accessed the NCBI FTP site to retrieve the FASTA file of proteins produced by bacteria in the RefSeq database and created a database (DB2) using BLAST+. We used data obtained on April 20, 2023.

2.2.2. Classification of Bacterial-Specific Amino Acid Sequences Using BLASTP Search

We focused on bacteria species whose presence rate changed more than 1% between healthy mice as control and mice in psychological stress conditions. Then we analyzed the proteins produced specifically by bacteria that showed significant changes in relative abundance. The genome data for the standard strains of the relevant bacteria were obtained from NCBI Genome in FASTA and TSV formats.

Standard Protein BLAST[19] (BLASTP) searches were performed on the acquired amino acid sequences via BLAST+, and sequences that matched only a single bacterial species were classified as bacterial-specific sequences. The BLASTP search parameters were as follows: an expected value threshold of 1e-10, BSDB0~9, and a maximum target sequence length of 20×10DB. All other parameters were left in their default settings.

We integrated the search results and extracted amino acid sequences only present in related bacteria. We also classified amino acid sequences that had not been characterized for function.

2.2.3. Functional Prediction by Sequence Analysis Using PSI-BLAST and InterPro Search

In Section 2.2.2, we set out to predict the functions of proteins with unknown roles by conducting sequence analysis using PSI-BLAST[20]. We performed PSI-BLAST searches focused on these uncharacterized proteins to forecast the functions that lacked annotations. The parameters for the search were as follows: Database: DB2, expected value threshold: 0.05, PSI-BLAST threshold: 0.0001, Number of iterations: 3, Maximum target sequences: 100, with all other settings maintained at their default values.

Furthermore, we employed InterProScan[21] to investigate domains by utilizing InterPro[22] to detect the presence of specific sequence domains. Subsequently, the results from both InterPro and PSI-BLAST were combined and visualized graphically. We filtered the outcomes of our sequence function predictions based on established criteria for acceptance.

((Positive > 40% AND Identity hits / Subject Total Length > 20%) OR (InterPro hits = “yes”)) (1)

 doi:10.20944/preprints202501.1007.v1

2.2.4. Functional Prediction of Protein-Protein Interactions Using Clusters Based on Gene Neighborhood

We sought to predict the functions of proteins whose roles remained undetermined in Section 2.2.3, utilizing genetic information as a foundation. Prokaryotic genomes, particularly those of bacteria, typically exhibit a circular structure wherein adjacent genes can influence each other's expression simultaneously. This configuration is referred to as the operon structure, and it is well established that proteins encoded by genes within an operon are likely to share similar functions.

Recent studies have introduced innovative strategies for predicting protein functions and interactions through protein-protein interaction (PPI) networks[23]. Prokaryotic genomes, particularly those of bacteria, typically exhibit a circular structure wherein adjacent genes can influence each other's expression simultaneously[24]. The STRING database is a comprehensive resource for protein-protein interactions, integrating experimental data, computational predictions, and literature-based information[25]. The STRING database uses gene neighborhoods of up to 300 bases as a factor in predicting protein interactions[25].

We identified the genes associated with each protein to utilize this information for inferring protein function. We then analyzed the locations of these genes, creating gene clusters with known functions that extended 100 bases before and after each gene. Because only the gene neighborhood is used as a predictor, the value is set more strictly than that used in STRING. We assessed whether the protein-coding genes that could not be inferred in Section 3.4 were assigned to the established gene clusters. The protein with the closest inter-gene distance was chosen as the representative sequence for those transferred.

2.2.5. Gene Ontology Enrichment Analysis by DAVID

We created a list of the extracted proteins identified in 2.2.2, plus the adopted sequences identified in 2.2.3~2.2.4. We used DAVID (The Database for Annotation, Visualization, and Integrated Discovery) [26] to perform gene ontology enrichment analysis on these lists. The list of extracted proteins identified in 2.2.2 and the adopted sequences identified in 2.2.3-2.2.4 have been added. The list was submitted using the RefSeq protein ID.

3. Result

3.1. 16S rRNA Amplicon Analysis Result

We obtained data on the percentage of bacterial composition within each sample. Bacterial data from healthy mice were treated as Ctrl, and bacterial data from ultrasonically irradiated mice and evaluated for hyperalgesia with the von Frey test were treated as USV (Figure 1, S1).

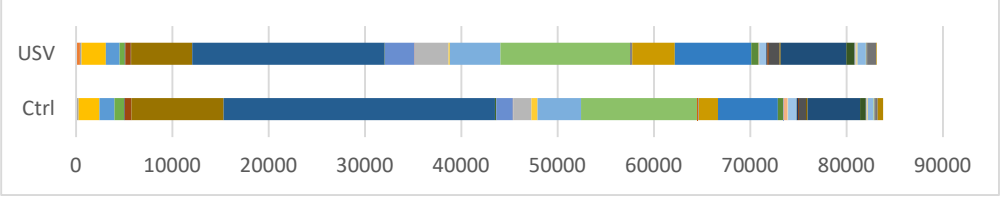


Figure 1. Bacterial composition from read counts at the Species Level.

We identified seven species of bacteria at the species level. We then tested for differences in the sample proportion for these species and found that *Mucispirillum schaedleri (M. schaedleri)* had increased significantly. (Table 1, Table S1)

Table 1. The composition ratio of bacteria identified at the species level in the samples.

 doi:10.20944/preprints202501.1007.v1

Organisms	*C. colinum*	*D. C21_c20*	*P. distasonis*	*B. pullicaecorum*	*R. gnavus*	*B. acidifaciens*	*M. schaedleri*	**total**
Ctrl read counts	11	37	65	133	912	1563	1970	83774
USV read counts	0	13	64	106	685	1448	3505	83116
Ctrl ratio (%)	0.013	0.042	0.078	0.159	1.089	1.866	2.232	
USV ratio (%)	0	0.016	0.077	0.128	0.824	1.742	4.217	
p-value	9.10.E-04	7.43.E-04	9.65.E-01	9.15.E-02	2.81.E-08	5.78.E-02	0.00.E+00	

[1] Normal mice were used as Ctrl and mice whose hyperalgesia was measured by the von Frey test after ultrasound irradiation were used as USV. We determined the proportion of bacteria identified at the species level from the total number of reads in each. The p-value was determined by the test of the difference between the mother ratios to determine if the bacterial composition ratios changed significantly.

We focused on *M. schaedleri*, which was present in more than 1% of the samples in the control and ultrasound-exposed groups and significantly increased by ultrasound exposure.

3.2. Analysis of Bacterial Genome Data Result

For the *M. schaedleri* genome data, we applied the developed analysis method and obtained the results as follows.

3.2.1. Classification of Bacterial-Specific Amino Acid Sequences Using BLASTP Search

We obtained the type strain genome annotation of *M. schaedleri* in TSV format. We received the amino acid sequence in fasta format. 2188 genes were annotated, of which 2118 were protein-coding genes. After eliminating duplicates, 2065 types of proteins were reported. As a result of performing a BLASTP query on the 2065 proteins produced by *M. schaedleri*, it was found that 303 of the proteins had amino acid sequences characteristic of *M.schaedleri.* As a result of performing a gene ontology enrichment analysis on these using DAVID, it was found that there are many proteins with transmembrane functions present in the outer membrane. Of the 303 proteins, 253 were “hypothetical proteins” with unknown functions. RefSeq Protein IDs were obtained for the 50 proteins with known functions. Functional prediction was performed for the remaining 253 proteins using method 2.2.2 (Table S2).

3.2.2. Functional Prediction by Sequence Analysis Using PSI-BLAST and InterPro Search

After analysis of Method 2.2.2, 46 of the 253 sequences met the filtering criteria. These were added to the analysis, and as in Result 3.2.2, many terms related to membrane proteins were found. (Table S3)

3.2.3. Functional Prediction of Protein-Protein Interactions Using Clusters Based on Gene Neighborhood

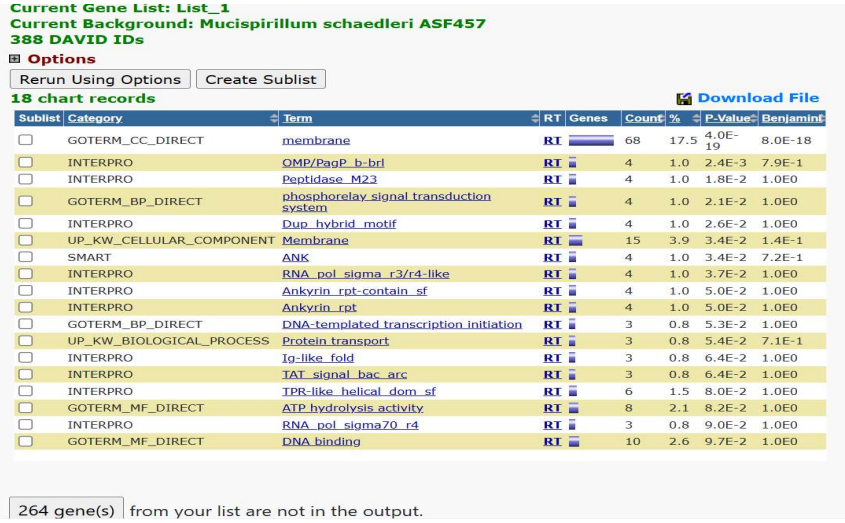
After the analysis was run, 439 clusters were generated. As a result of mapping the target protein-coding genes to the clusters in 2.2.3, it was found that 114/144 genes belonged to known functional clusters. Of these, 109 clusters contained the protein-coding gene under analysis. In total, 136 protein-coding genes could belong to clusters of known functions. The most proximal genes in the clusters were extracted from these 136 genes, and 116 genes were found to be in the clusters, excluding duplicates (Table S4).

3.2.4. Gene Ontology Enrichment Analysis by DAVID

In addition to the 303 proteins classified in 3.2.1, we created a list that included the predicted results of proteins that were considered to have similar functions in 3.2.2 and 3.2.3 (Table S5). We

 doi:10.20944/preprints202501.1007.v1

performed a Gene Ontology enrichment analysis using DAVID for these. As a result, we found that in addition to membrane proteins, the functions of Peptidase M23, Ankyrin repeat, and Tetratricopeptide domain were enriched (Figure 2).



Current Gene List: List_1
Current Background: Mucispirillum schaedleri ASF457
388 DAVID IDs

Options

Rerun Using Options | Create Sublist

18 chart records — Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	GOTERM_CC_DIRECT	membrane	RT		68	17.5	4.0E-19	8.0E-18
	INTERPRO	OMP/PagP_b-brl	RT		4	1.0	2.4E-3	7.9E-1
	INTERPRO	Peptidase_M23	RT		4	1.0	1.8E-2	1.0E0
	GOTERM_BP_DIRECT	phosphorelay signal transduction system	RT		4	1.0	2.1E-2	1.0E0
	INTERPRO	Dup_hybrid_motif	RT		4	1.0	2.6E-2	1.0E0
	UP_KW_CELLULAR_COMPONENT	Membrane	RT		15	3.9	3.4E-2	1.4E-1
	SMART	ANK	RT		4	1.0	3.4E-2	7.2E-1
	INTERPRO	RNA_pol_sigma_r3/r4-like	RT		4	1.0	3.7E-2	1.0E0
	INTERPRO	Ankyrin_rpt-contain_sf	RT		4	1.0	5.0E-2	1.0E0
	INTERPRO	Ankyrin_rpt	RT		4	1.0	5.0E-2	1.0E0
	GOTERM_BP_DIRECT	DNA-templated transcription initiation	RT		3	0.8	5.3E-2	1.0E0
	UP_KW_BIOLOGICAL_PROCESS	Protein transport	RT		3	0.8	5.4E-2	7.1E-1
	INTERPRO	Ig-like_fold	RT		3	0.8	6.4E-2	1.0E0
	INTERPRO	TAT_signal_bac_arc	RT		3	0.8	6.4E-2	1.0E0
	INTERPRO	TPR-like_helical_dom_sf	RT		6	1.5	8.0E-2	1.0E0
	GOTERM_MF_DIRECT	ATP hydrolysis activity	RT		8	2.1	8.2E-2	1.0E0
	INTERPRO	RNA_pol_sigma70_r4	RT		3	0.8	9.0E-2	1.0E0
	GOTERM_MF_DIRECT	DNA binding	RT		10	2.6	9.7E-2	1.0E0

264 gene(s) from your list are not in the output.

Figure 2. Results of Gene ontology enrichment analysis by DAVID.

4. Discussion

M. schaedleri, which increased in the ultrasound irradiation group, is a Gram-negative anaerobic bacterium with a spiral flagellum that is mainly observed in the intestines of mice. It has been reported that it is more abundant in the intestinal mucosa than in feces. [27,28] Although it is present in small quantities in human feces, the presence of 16S rRNA genes has been confirmed in colon mucus, suggesting its presence. *M. schaedleri* suppresses the growth of other bacteria by competing with pathogenic microorganisms such as *Salmonella enterica* serovar Typhimurium. [29] On the other hand, *M. schaedleri* can hurt the host by proliferating excessively. *M. schaedleri* has been reported to increase inflammatory conditions in the intestine. [27] The mechanism by which *M. schaedleri* increases in these inflammatory conditions is thought to be that *M. schaedleri* can inhabit environments rich in nitrate in inflammatory conditions, and that it has a rich supply of genes that adapt to oxidative stress. [27,30–33]It has also been reported that *M. schaedleri* increases due to social stress.[34] In this study, no outward signs of inflammatory conditions were observed. The results of the brain DNA microarray suggest that inflammation is occurring in the brain. It has been reported that neuroinflammation in the brain causes chronic pain and hyperalgesia. [35]In this study, outer membrane proteins were found to be abundant in proteins specific to *M. schaedleri*.[36] Gram-negative bacteria can release membrane proteins in the form of outer membrane vesicles as carriers of toxins, DNA, etc. OMVs can cause inflammatory diseases due to the presence of lipopolysaccharide (LPS) on their surface.[37] LPS causes inflammatory diseases via the innate immune system, such as TLR4[38]. Peptidase M23 can destroy the bacterial cell wall by hydrolyzing peptidoglycan and eliminating the bacteria.[39] This process has been demonstrated to cause the diffusion of LPS in the bacterial outer membrane into the surrounding environment, resulting in inflammation of the host's immune cells and nerve cells. Furthermore, the other report showed that inflammation caused by LPS may increase pain sensitivity.[40] In addition, it has been confirmed that ultrasound can increase or decrease the number of bacteria. The results of this study revealed that ultrasound irradiation can directly or indirectly increase the level of *M. schaedleri*.[41] The results suggest that the action of *M. schaedleri* peptidase M23 increases the concentration of LPS in the host, causing inflammation in both the host's immune cells and nerve cells. This LPS-induced inflammation likely contributes in part to the increased pain sensitivity observed in the host. Recent

 doi:10.20944/preprints202501.1007.v1

studies have shown that some bacteria contain Peptidase M23 in outer membrane vesicles. [42] This also suggests that peptidase M23 may contribute to LPS secretion via outer membrane vesicles. ANK repeat proteins are translocated into host cells via the type IV secretion system in prokaryotes. The translocated proteins are associated with evading the immune system and host cell-mediated gene expression. Class II chaperones of the type III secretion system, which are TPR-containing proteins, inject bacterial effectors into host cells by creating pores in the cell membrane of eukaryotic proteins.[43] These results also suggest that *M. schaedleri* affects host cells.

Our method succeeded in predicting and assigning annotations to 162 protein genes of the 253 genes whose functions were previously unknown. The most distinctive feature of our method focused on unique genes which only exist on that bacteria genome sequence. In our method, it was very helpful to predict the features for unknown or hypothetical genes by use of genes in neighborhood locus with annotations.

DNA microarray data of brain tissues for those mice also suggested that deferentially expression genes were related to inflammatory diseases, such as Ptgs2, Lcn2, and Cxcl1. We then performed a Gene Ontology enrichment analysis using DAVID for deferentially expression genes and extracted the term GO:0032496 response to lipopolysaccharide. This result suggests that there is a causal relationship between the data obtained from gut microbes and the data in the host brain via LPS.

We found that the proteins produced by *M. schaedleri* are registered in the AlphaFold Protein Structure Database.[44] Although our proposed methods only use primary structures of proteins, our method can be expanded to utilize 3D structure information.

We believe that further insights can be gained by predicting functions based on these predicted structures using structural comparisons.

We applied our method to analyze *M. schaedleri*. However, we also identified more than 20 types of bacteria classified only at the genus or family level. Therefore, as future trials, we will apply this method to unique genes on genus and family-level taxa all at once.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org, Table S1: Taxonomy table from 16S rRNA amplicon analysis data; www.mdpi.com/xxx/s2, Table S2: Results of classification of bacterial-specific amino acid sequences using the *M. schaedleri* genome and BLASTP search; www.mdpi.com/xxx/s3, Table S3: Result of Functional prediction by sequence analysis using PSI-BLAST and InterPro search www.mdpi.com/xxx/s4, Table S4: Result of Functional prediction of protein-protein interactions using clusters based on gene neighborhood; www.mdpi.com/xxx/s5, Table S5: List of Refseq protein IDs for analysis in DAVID ;.

Author Contributions: Conceptualization, K.Arita., Y.Nakano., and S.Miyazaki.; methodology, K.Arita.; software, K.Arita.; validation, S.Miyazaki., Y.Nakano. and K.Yoshizawa.; formal analysis, K.Arita.; investigation, K.Arita.; resources, S.Kasai., K.Yoshizawa.; data curation, K.Arita.; writing—original draft preparation, K.Arita.; writing—review and editing, S.Miyazaki.; visualization, K.Arita.; supervision, S.Miyazaki.; project administration, S.Miyazaki.; funding acquisition, S.Miyazaki.

Funding: This research was funded by FUJIMIC, Inc., Tokyo, Japan.

Data Availability Statement: 16S rRNA amplicon analysis data: The raw data supporting the conclusions of this article will be made available by the authors on request. Bacterial genome data: The original data presented in the study are openly available in Reference Sequence Database (RefSeq) at http://www.ncbi.nlm.nih.gov/RefSeq/.

Conflicts of Interest: The study was supported by a Grant from FUJIMIC, Inc., Tokyo, Japan.

References

1. Dethlefsen, L.; McFall-Ngai, M.; Relman, D. An ecological and evolutionary perspective on human–microbe mutualism and disease. *Nature* **2007**, *449*, 811–818. https://doi.org/10.1038/nature06245.
2. Neish, A.S. Microbes in gastrointestinal health and disease. *Gastroenterology* **2009**, *136*(1), 65–80. https://doi.org/10.1053/j.gastro.2008.10.080.
3. Blaser, M.J.; Musser, J.M. Bacterial polymorphisms and disease in humans. *J. Clin. Invest.* **2001**, *107*(4), 391–392. https://doi.org/10.1172/JCI11993.

4. New, F.N.; Brito, I.L. What Is Metagenomics Teaching Us, and What Is Missed? *Annu. Rev. Microbiol.* **2020**, *74*, 117–135. https://doi.org/10.1146/annurev-micro-012520-072314.
5. Mande, S.S.; Mohammed, M.H.; Ghosh, T.S. Classification of metagenomic sequences: methods and challenges. *Brief Bioinform.* **2012**, *13*(6), 669–681. https://doi.org/10.1093/bib/bbs054.
6. Douglas, G.M.; Maffei, V.J.; Zaneveld, J.R.; Yurgel, S.N.; Brown, J.R.; Taylor, C.M.; Huttenhower, C.; Langille, M.G.I. PICRUSt2 for prediction of metagenome functions. *Nat. Biotechnol.* **2020**, *38*(6), 685–688. https://doi.org/10.1038/s41587-020-0548-6.
7. Wemheuer, F.; Taylor, J.A.; Daniel, R. Tax4Fun2: prediction of habitat-specific functional profiles and functional redundancy based on 16S rRNA gene sequences. *Environ. Microbiome* **2020**, *15*, 11. https://doi.org/10.1186/s40793-020-00358-7.
8. Wood, V.; Lock, A.; Harris, M.A.; Rutherford, K.; Bähler, J.; Oliver, S.G. Hidden in plain sight: what remains to be discovered in the eukaryotic proteome? *Open Biol.* **2019**, *9*(2), 180241. https://doi.org/10.1098/rsob.180241.
9. An, A.A.K.; Shanmughavel, P. Computational Annotation for Hypothetical Proteins of *Mycobacterium Tuberculosis*. *J. Comput. Sci. Syst. Biol.* **2008**, *1*, 50–62.
10. Sanmukh, S.G.; Paunikar, W.N.; Meshram, D.B.; Ghosh, T.K. In silico Function Prediction for Hypothetical Proteins in *Vibrio Parahaemolyticus* Chromosome II. *Data Min. Knowl. Eng.* **2011**, *3*, 404–432.
11. Rajadurai, C.P.; Subazini, T.K.; Kumar, G. An Integrated Re-Annotation Approach for Functional Predictions of Hypothetical Proteins in Microbial Genomes. *Curr. Bioinform.* **2011**, *6*, 450–461.
12. Mazandu, G.K.; Mulder, N.J. Function Prediction and Analysis of *Mycobacterium Tuberculosis* Hypothetical Proteins. *Int. J. Mol. Sci.* **2012**, *13*, 7283–7302.
13. Pitt, W.G.; Ross, S.A. Ultrasound Increases the Rate of Bacterial Cell Growth. *Biotechnol. Prog.* **2003**, *19*, n.p.
14. Premoli, M.; Memo, M.; Bonini, S.A. Ultrasonic vocalizations in mice: relevance for ethologic and neurodevelopmental disorders studies. *Neural Regen. Res.* **2021**, *16*(6), 1158–1167. https://doi.org/10.4103/1673-5374.300340.
15. Amplicon, P.C.R.; P.C.R. Clean-Up; P.C.R. Index. 16s Metagenomic Sequencing Library Preparation. Illumina: San Diego, CA, USA, 2013, pp. 21.
16. O'Leary, N.A.; Cox, E.; Holmes, J.B.; Anderson, W.R.; Falk, R.; et al. Exploring and retrieving sequence and metadata for species across the tree of life with NCBI Datasets. *Sci. Data* **2024**, *11*(1), 732. https://doi.org/10.1038/s41597-024-03571-y.
17. O'Leary, N.A.; Wright, M.W.; Brister, J.R.; Ciufo, S.; Haddad, D.; et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **2016**, *44*(D1), D733–D745. https://doi.org/10.1093/nar/gkv1189.
18. Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; et al. BLAST+: architecture and applications. *BMC Bioinform.* **2009**, *10*, 421. https://doi.org/10.1186/1471-2105-10-421.
19. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410.
20. Altschul, S.F.; Madden, T.L.; Schaffer, A.A.; Zhang, J.; Zhang, Z.; et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402.
21. Jones, P.; Binns, D.; Chang, H.Y.; Fraser, M.; Li, W.; et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **2014**, *30*(9), 1236–1240. https://doi.org/10.1093/bioinformatics/btu031.
22. Paysan-Lafosse, T.; Blum, M.; Chuguransky, S.; Grego, T.; et al. InterPro in 2022. *Nucleic Acids Res.* **2023**, *51*(D1), D418–D427. https://doi.org/10.1093/nar/gkac993.
23. Zhou, J.; Xiong, W.; Wang, Y.; Guan, J. Protein Function Prediction Based on PPI Networks: Network Reconstruction vs Edge Enrichment. *Front. Genet.* **2021**, *12*, n.p.
24. Rogozin, I.B.; Makarova, K.S.; Murvai, J.; Czabarka, É.; et al. Connected gene neighborhoods in prokaryotic genomes. *Nucleic Acids Res.* **2002**, *30*(10), 2212–2223.
25. Szklarczyk, D.; Kirsch, R.; Koutrouli, M.; Nastou, K.; et al. The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* **2023**, *51*(D1), D638–D646.

26. Sherman, B.T.; Panzade, G.; Imamichi, T.; Chang, W. DAVID ortholog: An integrative tool to enhance functional analysis through orthologs. *Bioinformatics* **2024**, btae615. https://doi.org/10.1093/bioinformatics/btae615.
27. Herp, S.; Durai Raj, A.C.; Salvado Silva, M.; Woelfel, S.; Stecher, B. The human symbiont *Mucispirillum schaedleri*: causality in health and disease. *Med. Microbiol. Immunol.* **2021**, *210*(4), 173–179. https://doi.org/10.1007/s00430-021-00702-9.
28. Robertson, B.R.; O'Rourke, J.L.; Neilan, B.A.; et al. *Mucispirillum schaedleri* gen. nov., sp. nov., a spiral-shaped bacterium colonizing the mucus layer of the gastrointestinal tract of laboratory rodents. *Int. J. Syst. Evol. Microbiol.* **2005**, *55*, 1199–1204. https://doi.org/10.1099/ijs.0.63472-0.
29. Herp, S.; Brugiroux, S.; Garzetti, D.; et al. *Mucispirillum schaedleri* Antagonizes *Salmonella* Virulence to Protect Mice against Colitis. *Cell Host Microbe* **2019**, *25*, 681–694.e8. https://doi.org/10.1016/j.chom.2019.03.004.
30. Lundberg, J.O.; Hellstrom, P.M.; Lundberg, J.M.; Alving, K. Greatly increased luminal nitric oxide in ulcerative colitis. *Lancet* **1994**, *344*, 1673. https://doi.org/10.1016/s0140-6736(94)90460-x.
31. Singer, I.I.; et al. Expression of inducible nitric oxide synthase and nitrotyrosine in colonic epithelium in inflammatory bowel disease. *Gastroenterology* **1996**, *111*, 871. https://doi.org/10.1016/s0016-5085(96)70055-0.
32. Enocksson, A.; Lundberg, J.; Weitzberg, E.; Norrby-Teglund, A.; Svenungsson, B. Rectal nitric oxide gas and stool cytokine levels during the course of infectious gastroenteritis. *Clin. Diagn. Lab. Immunol.* **2004**, *11*(2), 250. https://doi.org/10.1128/CDLI.11.2.250-254.2004.
33. Winter, S.E.; Winter, M.G.; Xavier, M.N.; Thiennimitr, P.; Poon, V.; et al. Host-derived nitrate boosts growth of *E. coli* in the inflamed gut. *Science* **2013**, *339*(6120), 708–711. https://doi.org/10.1126/science.1232467.
34. Werbner, M.; Barsheshet, Y.; Werbner, N.; Zigdon, M.; et al. Social-Stress-Responsive Microbiota Induces Stimulation of Self-Reactive Effector T Helper Cells. *mSystems* **2019**, *4*(4), e00292-18. https://doi.org/10.1128/mSystems.00292-18.
35. Echeverria-Villalobos, M.; Tortorici, V.; Brito, B.E.; Ryskamp, D.; Uribe, A.; Weaver, T. The role of neuroinflammation in the transition of acute to chronic pain and the opioid-induced hyperalgesia and tolerance. *Front. Pharmacol.* **2023**, *14*, 1297931. https://doi.org/10.3389/fphar.2023.1297931.
36. Chen, S.; Lei, Q.; Zou, X.; Ma, D. The role and mechanisms of gram-negative bacterial outer membrane vesicles in inflammatory diseases. *Front. Immunol.* **2023**, *14*, 1157813. https://doi.org/10.3389/fimmu.2023.1157813.
37. Vanaja, S.K.; Russo, A.J.; Behl, B.; Banerjee, I.; Yankova, M.; et al. Bacterial Outer Membrane Vesicles Mediate Cytosolic Localization of LPS and Caspase-11 Activation. *Cell* **2016**, *165*(5), 1106–1119. https://doi.org/10.1016/j.cell.2016.04.015.
38. Kim, H.J.; Kim, H.; Lee, J.H.; et al. Toll-like receptor 4 (TLR4): new insight immune and aging. *Immun. Ageing* **2023**, *20*, 67. https://doi.org/10.1186/s12979-023-00383-3.
39. Razew, A.; Schwarz, J.N.; Mitkowski, P.; Sabala, I.; Kaus-Drobek, M. One fold, many functions—M23 family of peptidoglycan hydrolases. *Front. Microbiol.* **2022**, *13*, 1036964. https://doi.org/10.3389/fmicb.2022.1036964.
40. Kovács, K.J.; Papic, J.C.; Larson, A.A. Movement-evoked hyperalgesia induced by lipopolysaccharides is not suppressed by glucocorticoids. *Pain* **2008**, *136*(1–2), 75–84. https://doi.org/10.1016/j.pain.2007.06.017.
41. Pitt, W.G.; Ross, S.A. Ultrasound increases the rate of bacterial cell growth. *Biotechnol. Prog.* **2003**, *19*(3), 1038–1044. https://doi.org/10.1021/bp0340685.
42. Wang, Y.; Deng, J.; Wang, X.; Zhang, L.; Xu, Y.; et al. Isolation, identification, and proteomic analysis of outer membrane vesicles of *Riemerella anatipestifer* SX-1. *Poult. Sci.* **2024**, *103*(6), 103639. https://doi.org/10.1016/j.psj.2024.103639.
43. Loy, A.; Pfann, C.; Steinberger, M.; Hanson, B.; Herp, S.; et al. Lifestyle and Horizontal Gene Transfer-Mediated Evolution of *Mucispirillum schaedleri*, a Core Member of the Murine Gut Microbiota. *mSystems* **2017**, *2*(3), e00171-16. https://doi.org/10.1128/msystems.00171-16.
44. Mihaly, V.; Anyango, S.; Deshpande, M.; Nair, S.; et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **2022**, *50*(D1), D439–D444. https://doi.org/10.1093/nar/gkab1061.

 doi:10.20944/preprints202501.1007.v1