Machine Learning Aided Construction of the Quorum Sensing Communication Network for Human Gut Microbiota

Shengbo Wu
Tianjin University
Jie Feng
Tianjin University
Chunjiang Liu
Tianjin University
Hao Wu
Tianjin University
Zekai Qiu
Tianjin University
Jianjun Ge
Tianjin University
Shuyang Sun
Tianjin University
Xia Hong
Tianjin University
Yukun Li
Tianjin University
Xiaona Wang
Tianjin University
Aidong Yang
University of Oxford
Fei Guo
Tianjin University
Jianjun Qiao (jianjunq@tju.edu.cn)
Tianjin University

Article

Keywords: Quorum sensing, microbial languages, gut microbiota, microbial interactions, machine learning, microbial social network

Posted Date: August 12th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-755166/v1

Version of Record: A version of this preprint was published at Nature Communications on June 2nd, 2022. See the published version at https://doi.org/10.1038/s41467-022-30741-6.

Abstract

Quorum sensing (QS) is a cell-cell communication mechanism that connects members in various microbial systems. Conventionally, limited QS entries were collected for specific microbes, which is far from being able to fully depict QS-based complex microbial interactions in human gut microbiota. In this study, we proposed a systematic workflow including three modules and the use of machine learning-based ensemble classifiers to collect reported QS entries, expand the QS repository, and mine new potential QS proteins. Furthermore, we developed the Quorum Sensing of Human Gut Microbes (QSHGM) database (http://www.qshgm.lbci.net/) including 28,567 redundancy removal QS entries, to bridge the gap between QS repositories and human gut microbiota. With the help of QSHGM, various QS-based microbial interactions could be predicted and a comprehensive QS communication network was further constructed and analysed for 818 human gut microbes. This work contributes to the establishment of the QS communication network which may form one of the key knowledge maps of the human gut microbiota, supporting future applications such as potential therapies to gut diseases.

Highlights

A systematic workflow is built for QS entries collecting, expanding, and mining.

Four machine learning-based classifiers are trained for QS entries prediction.

QSHGM database is developed for various uses such as QS interactions prediction.

A QS communication network of gut microbiota is constructed for diverse applications.

Main Text

Human gut microbiota is a dynamic and complex microbial system[1] that links to the pathogen colonization resistance[2], immune system regulation[3], and human health maintenance[4]. Recent breakthroughs in high-throughput screening and multi-omics technologies have enabled the detection and quantification of the microbiota composition[5] in the human gut system. More and more research suggests that engineering the gut microbiota and regulating the microbial interactions[6, 7] can be viewed as potential novel therapeutics for treating diverse gut diseases[8].

Quorum sensing (QS), a population-level communication mechanism, has huge potential to be engineered for regulating microbial interactions and developing future therapies[9, 10]. Generally, there are diverse QS signals termed as microbial languages for intraspecies (N-Acyl homoserine lactones, AHLs; diffusible signal factors, DSFs; 4-hydroxy-2-alkylquinolines, HAQs; cholera autoinducer 1, CAI-1; auto-inducing peptides, AIPs; dialkylresorcinols; photopyrones)[11, 12] and interspecies (autoinducer 2, AI-2; indole) communications[13, 14]. The above QS languages in natural microbial systems such as gut microbiota play a significant role in the QS-based microbial interactions, which are closely relevant to various diseases[15]. For example, N-(3-oxodecanoyl)-L-homoserine lactone, a common AHL-type signal, can be incorporated into the host plasma membrane, leading to spontaneous TNFR1 trimerization and neutrophil apoptosis, thus providing a better colonization for *Pseudomonas aeruginosa* in the host[16]. DSF analogues were verified to strengthen the mucosal barrier and reduce antibiotic tolerance of *P. aeruginosa*[17]. Different hosts can utilize the aryl hydrocarbon receptor (AhR) to “listen in” the concentration of the HAQs from *P. aeruginosa* to regulate immune responses dynamically[18]. CAI-1 from *V. cholerae* can be designed to be recognized by an engineered *L. lactis* specifically in the gut, and the lactic acid from the engineered strain can repress the infection of *V. cholerae* in turn[19]. AI-2 produced by *Ruminococcus obeum* could repress several colonization factors of *Vibrio cholerae*, thus restricting the colonization of *V. cholerae*, which leads to diarrhoeal diseases[20]. Furthermore, indole has been confirmed to increase the expression of anti-inflammatory genes, elicit proinflammatory effects, affect the immune system of hosts, and decrease pathogen colonization[14, 21]. The evidence stated above suggests that manipulations of the level of diverse QS languages such as AI-2[22] in the QS communication networks play an important role in diverse host-centric applications for gut microbial systems[23]. Therefore, it is essential to construct a comprehensive QS database, which includes the collections of human gut microbes and QS entries repository, to bridge the gap between existing QS repositories and human gut microbiota.

Some existing databases relevant to gut microbiota or diverse QS systems have been constructed separately to provide data integration and interpretation for relevant researches. With respect to gut microbiota, the gutMEGA database[24] contains thousands of gut microbiota composition (metagenomic sequences), phenotypes, and experimental information. GMrepo[25] focuses on the annotated human gut metagenomes to facilitate the development of human metagenomic data. BIO-ML[26] includes 7,758 gut bacterial isolates, 3,632 genome sequences, and diverse longitudinal multi-omics data. Particularly, VMH[27] is a database that has integrated thousands of metabolites, reactions, human genes, microbes (818 strains), microbial genes, and food items that link to hundreds of gut diseases and nutritional data. With regard to QS, repositories of limited QS systems in Gram-negative and Gram-positive bacteria have previously been curated to form SigMol[28] and Quorumpeps[29], respectively. P2CS[30, 31] was constructed and updated for two-component system (TCS), which is a typical QS system that is composed of a histidine kinase receptor and a response regulator partner. Furthermore, we have previously developed the QSIdb database[32] to expand the potential QS interference molecules for different QS systems. We applied a pipeline including SMILES-based algorithms and docking-based validations to obtain a potential QS interference molecules dataset (73,073 compounds) from the existing compounds in the PubChem database. Note that some recent databases such as gutMDisorder[33] have linked the human microbiota and many macro-environmental factors together to describe the intervention and regulation for various diseases. In addition, exogenous active substances and endogenous host factors were also collected for human microbiota into MASI[34] and GIMICA[35], respectively, to provide the information of the interactions of various substances and gut microbiota.

While gut microbiota and QS systems have been curated in various databases, they have largely been collected separately so far, which may limit the understanding of QS-based complex microbial interactions in human gut microbiota. Furthermore, existing studies have often focused on using limited reported QS entries; novel QS entries mining and comprehensive integration to form a relatively complete network is yet to be further explored. Although some biological networks such as protein–protein interaction networks have been relatively mature, they cannot decipher complex microbial cell-cell communications. In this study, we developed a systematic workflow including QS collecting, expanding, and mining modules to construct a comprehensive QS repository for human gut microbiota. In the QS collecting module, we curated the annotated QS entries carefully for each component in human gut microbiota to form a repository of reported QS entries. Information gathering in this module was also combined with Machine learning (ML) algorithms including random forest (RF), k-nearest neighbour (KNN), support vector machine (SVM), and deep neural network (DNN) to develop four ensemble classifiers, which were then used in the QS expanding module to nominate further candidates of human gut QS signal molecules and receptors from existing (general-purpose) QS databases. These candidates were finally analysed in the QS mining module, where protein annotation, functional analysis and homologous modelling were combined to re-annotate and mine new QS entries. These have led to a comprehensive QS database of human gut microbiota (QSHGM, http://www.qshgm.lbci.net/) including the reported (21,410) and extended (7,157) QS entries, which offers user-friendly browsing and searching functions to support various applications. With the help of QSHGM, we can predict complex QS-based interactions for different microbial consortia and further proposed a QS communication network for the first time to visualize and decipher intricate QS-based microbial interactions for human gut microbiota. Finally, we identified key challenges and suggest directions for the QS communication network and how we can engineer them to provide more future applications.

Results

The systematic workflow for collecting, expanding, and mining QS entries. To construct a comprehensive QS repository for human gut microbiota, we developed a systematic workflow which includes three modules (QS collecting, QS expanding, and QS mining modules) and four ensemble classifiers based on ML algorithms (Fig. 1). In the QS collecting module, we firstly obtained 213 recognized QS entries (Dataset I) from SigMol and Quorumpeps databases, and curated their corresponding amino acid sequences from the UniProt database. In parallel, we manually searched the 818 gut microbes from the VMH database[27] (Dataset II) to collect reported QS entries which are termed “positive samples” (Dataset III). The search was based on four commonly used QS annotations, i.e., “quorum sensing”, “LuxR”, “two-component”, and “tryptophanase”. The negative samples (Dataset IV) were then obtained by removing proteins from typical proteomes in dataset II, such as *Escherichia coli* and *Pseudomonas aeruginosa* (more details in Method section), that conform to QS cluster rules. These rules were developed based on Dataset I through sequence analysis, including evolution analysis, QS-relevant protein annotations, and amino acid sequence descriptors comparison (more details in Method section). In the QS expanding module, we obtained an extended protein dataset (Dataset V) from the results of the local BLASTP[36] on the datasets I and II with the criteria of the *E* value[37] being smaller than 10^{-5}, which is commonly used in sequence

alignment to obtain homologs. Four different ML algorithms (SVM, RF, KNN, and DNN) were used to construct ensemble classifiers, which were trained and validated based on the above positive (III) and negative samples (IV). After excluding from dataset V those which are already collected as the reported QS entries in dataset, the remaining QS entries (Dataset VII) were then classified by the four ML-based ensemble classifiers stated above. The output of these classifiers was further processed in the QS mining module, where the potential QS entries predicted by the classifiers, which had not previously been discovered and annotated, were mined and sorted out manually with the help of the functional analysis and homologous modelling that were supported by UniProt[38], NCBI (https://www.ncbi.nlm.nih.gov/) and Phyre[2] databases[39].

Reported and annotated QS entries. There are 84 autoinducer synthases and 129 QS receptors in dataset I. With respect to autoinducer synthases, we divided them into seven types, i.e., AHLs, DSFs, AI-2, indole, HAQs, CAI-1, and others. As a result, AHLs synthases account for the vast majority, which among other possibilities can be divided into two protein families, LuxI (from *Vibrio fischeri*) and YenI (from *Yersinia enterocolitica*) (Fig. 2A). With regard to QS receptors, we also divided them into seven types, i.e., LuxR type, TCS type, CAI-1 receptor, AI-2 receptor, DSFs receptor, HAQs receptor, and other receptors (Fig. 2B). LuxR and TCS type receptors account for the vast majority of QS receptors. Similarly, LuxR type receptors can be roughly divided into two protein families, LuxR (from *V. fischeri*) and YenR (from *Y. enterocolitica*). Note that the evolutionary trees of AHLs synthases and their receptors counterpart are in a high similarity (Fig. 2A and 2B), part of which was also identified by Gray et al[40]. This indicates that there is a coevolution for AHLs synthases and their corresponding receptors.

There are 1,640, 5,921, 15,703, and 66 QS entries for “quorum sensing”, “LuxR”, “two-component”, and “tryptophanase”, respectively (Fig. 2C). LuxR-type and TCS QS entries account for the vast majority, which are 25.38% and 67.31%, respectively. We have also shown the distribution of QS entries for each strain based on the seven-strain simplified human microbiomes (SIHUMIs) used by Colosimo et al[41] (Fig. 2D). This indicates that LuxR and TCS type QS entries account for the vast majority of QS entries in these strains. Furthermore, we noted that there are certain overlaps in the distribution of the four QS entries. For example, there are seven QS entries (P69409, P0ACZ6, P0AGA8, P66798, P0AF30, P0AEL9, and Q8XE66) shared by both LuxR and TCS receptors in the *E. coli* O157:H7 strain (Fig. 1E). This suggests potential crosstalk of LuxR type and TCS QS systems. In addition, we have counted and distributed the total QS entries of the 818 gut microbes from the VMH database[27] to form a better picture of the QS repository in human gut microbiota (Fig. 2F). As a result, we found that about 90% strains contain less than 60 QS entries, and only seven strains have more than 150 QS entries, which have been listed in Fig. 2F. This distribution will be revisited after extended QS entries are included (see below).

Expanded and new QS entries. We conducted 5-fold cross validation to test classifiers, where the accuracy, prediction, recall, and F1 score (more details were listed in method section) were applied to evaluate their performances. The RF classifier achieves the highest accuracy and F1 score among the four classifiers, which indicates that the RF classifier achieves the best performance, followed by KNN, SVM, and DNN. To obtain more details of the positive entries predicted by different classifiers, we have manually checked their annotations and categorized the proteins into four types, i.e., QS irrelevant, autoinducer synthases, QS receptors, and uncharacterized proteins. The results show that QS receptors account for the vast majority, followed by the autoinducer synthases (Fig 3B).

In addition to the collection of the confirmed autoinducer synthases and QS receptors, we have further analysed the details for the uncharacterized proteins (534 entries) from the positive ones predicted by the three better performed classifiers (RF, SVM, KNN), in order to mine more QS relevant proteins. As a result, we have re-annotated the 534 entries and grouped them into nine protein clusters manually (Fig. 3C), in which the histidine kinase (a major component in a TCS) occupied the majority. Note that there were another 28 entries that are vaguely described without specific protein annotations (Fig. 3C). As listed in Table 1, these entries were further explored and re-annotated based on the web BLASTP of NCBI database or Phyre[2]. There were 20 proteins (Table 1, upper) that can be re-annotated based on the BLASTP results from NCBI. Except U2J6M1 and C0C5Y6, there is much potential for the other 18 proteins to be QS proteins. ArsR, a component of ArsRS TCS, regulates the acid adaptation and biofilm formation of the pathogen *Helicobacter pylori* in human gut[42]. Beta-ketoacyl-ACP synthase III catalyzes the condensation reaction of fatty acid synthesis, which indicates that there is potential for *Prevotella bivia* to produce Dialkylresorcinols just like the function of DarB from *Photorhabdus asymbiotica*[43]. The histidine kinase, LuxR family regulator, and Rgg/GadR/MutR family regulator are important parts of TCS, LuxR-type, Rgg-based QS systems[44], respectively.

There are eight entries (Table 1, lower) that have no specific annotations or classifications in NCBI or UniProt database. We submitted these protein sequences to Phyre[2] to investigate the 2D and 3D structures of their models, their domain compositions and model quality. A0A4Y4IIW5 and A0A5C4P2T9 are signalling protein and AgrC (belonging to Agr QS system[45]) family protein, respectively. This indicates that *Lysinibacillus fusiformis* and *Streptococcus salivarius* may have some protein components of the *agr* QS system, thus producing and/or responding to the same QS signalling peptide as common pathogen *Staphylococcus aureus*. The other six of them are templated on the AimR transcriptional regulator, which is the intracellular signal peptide receptor for the QS-based communication between viruses that guides lysis–lysogeny decisions[46]. This suggests that different Bacillus phages may "listen in" diverse bacterial hosts, such as *Bacillus amyloliquefaciens*, *Bacillus mycoides*, *Bacillus thuringiensis*, and *Bacillus atrophaeus*, to coordinate lysis–lysogeny decisions.

Table 1. Results of 28 expanded entries without existing annotations.

Strains	TaxID	Entry	Template	Query Cover	Percent identity	New annotations	Sources
Halococcus morrhuae	931277	M0MA34	WP_004054989.1	100%	100%	ArsR subfamily of regulator	Web BLASTP
Clostridium hylemonae	553973	C0C300	WP_006443816.1	100%	100%	Autoinducer 2 ABC transporter	Web BLASTP
Prevotella bivia	868129	I4Z9V6	WP_036847997.1	80%	80.39%	Beta-ketoacyl-ACP synthase III	Web BLASTP
Enterococcus caccae	1158612	R3TYZ5	WP_069646785.1	100%	80.80%	Histidine kinase	Web BLASTP
Lactobacillus ruminis	525362	E7FSN7	WP_003695050.1	98%	98.96%	Histidine kinase	Web BLASTP
Streptococcus peroris	888746	E8KCS5	WP_070888551.1	100%	99.58%	Histidine kinase	Web BLASTP
Streptococcus parauberis	1348	A0A3E1JFV3	WP_116486843.1	100%	100%	Histidine kinase	Web BLASTP
Hungatella hathewayi	566550	D3ADP6	PXX46370.1	98%	92.45%	LuxR family regulator	Web BLASTP
Enterococcus cecorum	1121864	S1R0J3	WP_047242627.1	100%	97.31%	Rgg/GadR/MutR family regulator	Web BLASTP
Enterococcus cecorum	1121864	S1R7E8	WP_171336239.1	98%	93.70%	Rgg/GadR/MutR family regulator	Web BLASTP
Streptococcus constellatus	1035184	U2ZME3	WP_022525523.1	100%	100%	Rgg/GadR/MutR family regulator	Web BLASTP
Streptococcus equinus	525379	E8JR85	WP_029875994.1	97%	97.20%	Rgg/GadR/MutR family regulator	Web BLASTP
Streptococcus intermedius	1095731	U2XPZ3	WP_003032153.1	100%	100%	Rgg/GadR/MutR family regulator	Web BLASTP
Candidatus Melainabacteria	2052166	A0A3S0FWU1	MBI4533416.1	80%	47.68%	Sensor histidine kinase	Web BLASTP
Candidatus Melainabacteria	2052166	A0A431KQ57	MBI5174129.1	79%	47.28%	Sensor histidine kinase	Web BLASTP
Coriobacteriales bacterium	2491116	A0A437UTJ5	WP_130811315.1	99%	43.81%	Sensor histidine kinase	Web BLASTP
Lactobacillus amylolyticus	585524	D4YTV9	EST03116.1	97%	36.63%	Sensor histidine kinase	Web BLASTP
Alistipes putredinis	445970	B0MUZ2	OKY96599.1	100%	96%	Tryptophanase	Web BLASTP
Sphingobacterium paucimobilis	1346330	U2J6M1	WP_021069213.1	100%	100%	DoxX family, membrane protein YphA	Web BLASTP
Clostridium hylemonae	553973	C0C5Y6	WP_006444869.1	100%	100%	Sugar ABC transporter protein	Web BLASTP
Strains	**TaxID**	**Entry**	**Template**	**Confidence**	**Coverage**	**New annotations**	**Sources**
Bacillus amyloliquefaciens	1390	A0A5C8IUS9	c5xybB	100%	97%	AimR transcriptional regulator	Phyre[2]
Bacillus mycoides	1405	A0A1W6AJT8	c5zvvA	100%	90%	AimR transcriptional regulator	Phyre[2]

Bacillus thuringiensis	56955	A0A243M9P9	c5zw5A	100%	95%	AimR transcriptional regulator	Phyre[2]
Bacillus amyloliquefaciens	1390	A0A5C8IY56	c5zvvA	100%	99%	AimR transcriptional regulator	Phyre[2]
Bacillus atrophaeus	720555	A0A0H3E1W6	c5zvvA	99.90%	98%	AimR transcriptional regulator	Phyre[2]
Bacillus atrophaeus	720555	A0A0H3E2G4	c5zw5A	100%	100%	AimR transcriptional regulator	Phyre[2]
Lysinibacillus fusiformis	28031	A0A4Y4IIW5	c6mfvC	100%	90%	Signaling protein (tetratricopeptide repeat)	Phyre[2]
Streptococcus salivarius	1304	A0A5C4P2T9	c4bxiA	99.90%	33%	ATP binding domain of AgrC	Phyre[2]

To sum up, with the help of the proposed systematic workflow (Fig. 1), we obtained a comprehensive QS repository including the manually collected 21,410 positive samples and the extended 7,157 ones for 818 gut microbes, and the total 28,567 QS entries are composed of 1,882 QS synthases and 26,685 receptors. There was a 33.43% increase of QS entries for the comprehensive QS repository (Fig. 3D) from the the previous annotation-based QS collections (Fig. 2F). Furthermore, included in the extended entries, we have re-annotated 534 proteins and mined eight new potential QS proteins with the help of functional analysis and homologous modelling. This is of great significance to the further exploration of the related QS mechanism and their applications.

QSHGM browsing and searching. To enable user-friendly browsing and searching for QS entries identified in this work, we constructed a comprehensive QS database of human gut microbiota (QSHGM), which is freely available at: http://www.qshgm.lbci.net/. A user-friendly “‘browse” option allows to explore the QS data including the annotated QS and extended QS entries. In the “browse” option, a query box is provided in which the user can enter the query on the basis of “All”, “Synthases” or “Receptors” for the browsing of QS entries. By “Synthases”, one can query QS entries according to nine QS languages: AHLs, CAI-1, Dialkylresorcinols, Photopyrones, DSFs, HAQs, AIPs, Indole, and AI-2. As an example, we have illustrated part of browsing results for AHLs language in Fig. 4, and the output displays information of the QS entries, fielded by Entry, Genus, Species, Strain, Taxonomic identifier (TaxID), Protein annotations, conventional abbreviations of QS signals (Languages), and Link Address.

QSHGM also includes “Search” searching facilities for different QS entries. In the search option, a query box is provided in which the user can enter the query on the basis of “Microbes”, “Synthases” or “Receptors” for the searching of QS entries. By “Microbes”, one can query QS entries according to different options: Entry (e.g., J7JCP9), Name (e.g., *Pseudomonas aeruginosa*), or TaxID (e.g., 208964). The output displays information of the QS entries, fielded by Entry, Organism from WMH, TaxID from Uniprot (Uniprot), Proteome ID, substitute organism (Organism), substitute organism TaxID (Organism ID), all protein counts (All Proteins), QS entries counts for the strain (Counts), and Protein annotations. The “Synthases” is provided according to different options: Entry (e.g., J7JCP9) or Languages (e.g., AHLs). By “Receptors”, one can search with different options: Entry (e.g., P25084) or Annotations (e.g., Histidine kinase). The output displays information of QS synthase and QS receptors, fielded by Entry, Genus, Species, Strain, TaxID, Protein annotations, and Languages. Note that search type allows users to retrieve either an exact match or the match containing the query.

QS-based microbial interactions prediction. QS-based microbial interactions play an essential role in deciphering complex interactions of natural microbial systems and dynamically manipulating diverse synthetic microbial consortia. According to the collected data in the QSHGM database, we can predict various potential pairwise QS-based microbial interactions. For example, we predicted AI-2-based communication between *E. coli* O157:H7 and *Bacteroides pectinophilus* ATCC 43243 (Fig. 5A), which is in line with the previously reported observation that AI-2 produced by *E. coli* can influence the Bacteriodetes[47]. Furthermore, TnaA (encoding indole) was previously reported in *E. coli*[48] and *Enterobacteriaceae*[49], which is also indicated by the QSHGM database, suggesting that there will be indole-based interaction between these two microbes. Therefore, we predicted that a microbial consortium including *E. coli*

O157:H7, *B. pectinophilus* ATCC 43243 and *E. bacterium* 9_2_54FAA can be regulated by manipulating the concentration level of AI-2 and indole (Fig. 5B). Furthermore, we can predict more sophisticated interaction networks. When introducing the *P. aeruginosa* PAO1 into the above three-strain consortium, there will be complex microbial cell-cell communications based on AI-2, AHLs and indole (Fig. 5C), in which the interactions between *P. aeruginosa* PAO1 and *E. coli* were reported and validated previously[50, 51]. When adding *Burkholderia cepacia* GG4 to the above four-strain consortium, we can also predict the complex QS-based interaction network for a five-strain consortium that communicates with AI-2, indole, AHLs, HAQs, and DSFs (Fig. 5D), which included a previously validated HAQs-based interaction between *P. aeruginosa* and *B. cepacia* GG4[52]. To sum up, QS-based interaction predictions stated above have been partially verified in the corresponding experiments from other reported researches. Therefore, it has huge potential to predict more complex QS-based interaction networks including multi-component strains based on diverse QS languages.

QS communication network construction for the human gut microbiota. Microbes communicate via various QS signals, and it is possible to construct a cell-cell communication network among different gut microbes based on diverse QS languages, which we termed as "QS communication network". With the help of the comprehensive QS repository in the QSHGM database, we constructed a QS communication network for the 818 gut microbes based on the "speaking" of the above nine QS languages (Fig. 6A). This intricate network visualizes the complex QS-based communications and interactions among human gut microbiota. Different microbes are linked together through various languages to form a microbial communication network, and the connections could be used to regulate the microbial interactions between themselves and the surrounding ones. As shown in Fig. 6A, most of the strains produce the signal AI-2 (567, 69.3% of 818 gut microbes) as the communication language, followed by HAQs (332, 40.6%), DSFs (325, 39.7%), CAI-1 (259, 31.7%), Dialkylresorcinols (129, 15.8%), Photopyrones (107, 13.1%), indole (77, 9.4%), AHLs (64, 7.7%), and AIPs (22, 2.7%).

Note that multiple microbes can speak one common language which is in line with the interspecies crosstalk[53]. Taking six typical languages (AHLs, CAI-1, HAQs, DSFs, Indole, and AI-2) as example, we found that there are 64, 40, 22 and 5 species sharing two, three, four, and five QS languages, respectively (Fig. 6B). AI-2 also ranks first with the highest genus-level counts (138 genus) than the other languages, which is in line with what has been broadly observed[13]. Many overlaps of the languages being spoken (between different microbes) include AI-2 or indole for various genus, which also indicates that both of them are widely recognizable languages playing a major role for inter-specie communications[54, 55]. We found that those traditionally often considered as intraspecies languages (AHLs, CAI-1, HAQs, and DSFs) may also be involved in some interspecies communications. In addition, the crosstalk of different QS languages for various microbes implies the redundancy of microbial languages that is potentially helpful for the stability of natural microbial systems.

The QS communication network was constructed based on the 818 human gut microbes, which include mainly Firmicutes (79), Actinobacteria (36), Proteobacteria (69), Bacteroidetes (16) and others (10). We have collected and sorted the nine QS languages for 210 microbes at the genus level, shown by the heatmap representation in Fig. 6C to gain a better understanding of the QS communication network (Fig. 6A). It has previously been reported that AHLs are only found in Proteobacteria[56], AIPs exist mostly in Firmicutes[12], and other QS languages are distributed in-homogeneously in the whole genus-level microbes[57], with which the QS communication network predicted by the QSHGM database agrees. Surprisingly, there are no highly similar distributions of QS languages within the same genus-level microbes. On the contrary, taking the distribution of QS languages in Actinobacteria as an example, the language distributions are quite different between its members (Fig. 6C, cyan). This suggests that the existence and evolution of autoinducer synthases in microbes might have not been strictly familial at the genus level, but are more likely to be related to a variety of factors, such as environmental factors and spatial distributions[58-60]. To sum up, these predicted patterns of distribution of QS languages between these microbes suggests the diversity of the microbial communication languages, the complexity of cell-cell communication, and the redundancy of QS-based interactions among human gut microbiota.

Discussion

The construction of several types of biological networks including protein–protein interaction networks, genome-scale metabolic networks, gene regulatory networks, and gut microbial co-abundance networks has been relatively mature. However, these networks do not directly reveal the structure of microbial cell-cell communications. The QS communication network we established for the first time is thus of great potential for the investigation of the complex QS-based interactions in human gut microbiota. On the one hand, the QS communication network can give us a better understanding of QS-based microbial communication principles which determine

the dynamics of the complex ecological systems. On the other hand, this network will do much help for developing potential therapies for the pathogen-relevant diseases by manipulating different QS languages or strains (probiotics). Thanks to the large scale of the data established in this work, potential useful details for the QS-based communications among different gut microbes can be obtained in our QSHGM database.

The QS communication network we presented above (Fig 6A) is a bipartite network involving two types of nodes, namely QS languages and microbes. This network can be projected to a one-mode network that visualizes microbe-microbe interactions directly. The giant network would consist of 801 nodes connected via 190,580 edges (Fig. S1). The largest degree in the giant network is 771, while its average degree is 237.93. This network is characterized by a large number of highly connected nodes, while only a small number of nodes have few connections, which suggests that microbes are not limited to communicate within their own species but are capable of 'listening in' and 'broadcasting to' unrelated species for the good of the population[61]. The dense QS network is similar to other microbial interaction networks that carry high degrees for individual strains[62, 63]. Key nodes in this network were selected from 5% of the total nodes (40 nodes, Table S1) of the network with large degree and high betweenness centrality[64]. Note that all of the 40 key nodes are Firmicutes, Bacteroidetes, or Proteobacteria, (Table S1) which are known to be dominating species of the human gut microbiota[65, 66]. For illustrative purposes, a complete graph of the 40 key nodes is shown in Fig. 7A, where there is a link between every pair of nodes. While such a dense network more likely approximates a theoretical maximum set of QS-based interactions it nevertheless indicates excellent microbial communications among the core microbes.

It is worth noting that the network shown in Fig. 6A and Fig. 7A illustrate diverse language connections, which lacks the further interactions between QS languages sender and receiver. By differentiating QS signals producing and receiving with the help of both QS synthases and receptors, there is potential to construct a directed QS network. Taking the seven-strain simplified human microbiomes from Colosimo et al[41] as an example, we constructed a typical small QS communication network that includes QS languages producing and receiving (Fig. 7B). QS languages receiving (Fig. 7B, right) is more complicated than language producing (Fig. 7B, left), which indicates that some microbes can receive a particular QS signal without producing it. This phenomenon is consistent with the previously observed QS cheating behaviour in certain microbes, such as *P. aeruginosa*[67] and *E. coli*[68]. The reliable construction of directed QS networks still faces many challenges, such as the huge network scale, multi-layer control structures, complex QS crosstalk, intricate social cheating, diverse environmental factors, different spatial distributions, and insufficient QS entries for many uncultured microbes. Nevertheless, we expect that the further comprehensive QS communication networks including QS languages producing and receiving will receive increasing attention from future research which will be engaged in developing more knowledge and technologies for various gut microbes, aiming to construct the valuable comprehensive QS communication networks which can be regarded as one of the key knowledge maps of the human gut system.

Conclusion

Various QS-based interactions play an essential role in the regulation of homeostatic states, metabolism, and immunity responses for the human gut system. Therefore, constructing a comprehensive QS database for the human gut microbiota is highly desirable for making gut microbiology more predictable and for developing potential therapies for diverse gut diseases. In this work, we developed a systematic workflow including QS collecting, QS expanding, and QS mining modules to construct a comprehensive QS repository for the human gut microbiota. machine learning algorithms including SVM, RF, KNN, and DNN were combined with protein annotations, functional analysis and homologous modelling to facilitate the efficiency of data collection and mining. As a result, we established the QSHGM database (http://www.qshgm.lbci.net/, with browsing and searching functions) which contains 28,567 redundancy removal QS synthases (1,882) and receptors (26,685) entries for 818 gut microbes.

With the help of QSHGM database, users can predict many QS-based interactions for various microbial consortia based on diverse QS languages. We constructed a QS communication network to visualize and decipher intricate QS-based microbial interactions for human gut microbiota. We found that the distribution of QS languages in microbes is not strictly familial at the genus level, but is more likely to be related to other factors. There are significant genus-level overlaps between microbes on what are commonly regarded as intraspecies languages, which suggests that these languages may also be involved in some interspecies communications. The predicted sharing of various subsets of the QS languages between microbes supports the notions of the diversity of the microbial language and the redundancy of cell-cell communications, which are helpful for maintaining the stability of natural microbial systems. This work contributes to the construction of the QS communication network for human gut microbiota

that may form one of the key knowledge maps of the human gut system in the future. Such a network holds huge potential for improving our understanding of the dynamics and resilience of the gut microbiology and for developing applications such as potential therapies.

Methods

Data acquisition. QS is a common mechanism which includes autoinducer synthase and relevant QS receptors[69]. For most Gram-negative bacteria, the autoinducer produced by the autoinducer synthase accumulates in the culture with the cell density increasing; When the concentration of the autoinducer reaches a certain threshold, it will diffuse back into strain and be recognized and bonded by the QS receptor to be a complex to activate or inhibit the transcription of downstream genes[54]. The autoinducer synthases and receptors for Gram-negative bacteria from SigMol (http://bioinfo.imtech.res.in/manojk/sigmol), and QS receptors for Gram-positive bacteria from Quorumpeps (http://quorumpeps.ugent.be) are utilized as the validated QS proteins in our research. Their corresponding amino acid sequences are obtained from UniProt (https://www.uniprot.org/)[38]. In addition, 818 gut microbes from virtual metabolic human database (www.vmh.life)[27] are regarded as the human gut microbiota in this study, and their corresponding proteomes are also obtained from UniProt.

Feature extraction and classifiers development. The secondary and tertiary structure of a protein depends on its amino acid sequence[70]. Therefore, a large amount of physiochemical and structural descriptors extracted from amino acid sequences have been widely used to predict drug-target interactions[71], variable-length antiviral peptides[72], and protein-ligand binding sites[73] etc. In this study, the information of amino acids in protein sequences was calculated, and ML algorithms (SVM[74], RF[75] and KNN[76]) and deep learning algorithm (DNN[77]) were trained on the carefully curated positive and negative samples to develop different classifiers. We calculated the frequency of each amino acid type in each QS related protein sequence. The frequencies of all 20 natural amino acids are the percent of the number of amino acid type divided by the length of a protein sequence[78].

Positive and negative samples construction. With the help of the evolution analysis of amino acid sequences of autoinducer synthases and receptors, we collected the reported and annotated QS proteins for 818 gut microbes as the positive samples. In the QS expanding module, we did an evolutionary analysis for the validated QS entries to propose a possible cluster rules for negative samples collection with the help of MEGA X[79], and iTOL[80]. The evolutionary history was inferred using the Neighbor-Joining method[81]. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree[82]. The evolutionary distances were computed using the Poisson correction method and are in the units of the number of amino acid substitutions per site[79]. We constructed negative samples by removing QS-related components from typical Gram-negative bacteria (*Aliivibrio fischeri*, *Escherichia coli*, *Pseudomonas aeruginosa*, *Salmonella typhimurium*, and *Vibrio parahaemolyticus*) and Gram-positive bacteria (*Bacillus subtilis*, *Staphylococcus aureus*, and *Lactococcus lactis*), and removing proteins that directly and indirectly associated with QS, i.e., cluster rules, such as quorum sensing, luxR, two-component, homoserine-lactone synthase, histidine kinase, biofilm, autoinducer, bacteriocin, competence, virulence, signal, sensor, response, regulator, membrane, binding, transcriptional activator etc.

ML-based classifiers. All the four classifiers were applied to predict that whether the input amino acid sequences are QS entries or not with the output being 1 (yes) or 0 (no), respectively. Classifiers were trained and validated based on the positive and negative samples, and then tested on the dataset V (Fig. 2). Performances of the four ML-based classifiers were measured based on the accuracy, precision, recall, and F1 score, which are defined as follows [83].

$$\text{Accuracy} = (TN+TP) / (TN+FP+FN+TP) \quad (1)$$

$$\text{Precision} = TP / (TP+FP) \quad (2)$$

$$\text{Recall} = TP / (TP+FN) \quad (3)$$

$$F1 = 2 * \text{Precision} * \text{Recall} / (\text{Precision} + \text{Recall}) \quad (4)$$

where TP represents true positives, TN denotes true negatives, and FP and FN are false positives and false negatives, respectively. F1 score is the harmonic mean of prediction and recall. The higher the F1 score is, the better performance the classifier will be of.

SVM is a commonly used supervised ML algorithm in proteins prediction[83, 84]. The basic idea of SVM is to find the separated hyperplane in a very high-dimension feature space that can correctly partition the training data set[74]. SVM can also integrate kernel functions, which make it to be a nonlinear classifier. In this study, for our results we applied the radial basis function (RBF) with standard deviation $\sigma = 0.125$ and set regularization parameter $C = 4$ to train the positive and negative samples.

K-nearest-neighbour (KNN) is also a traditional classification method when there is little or no prior knowledge about the distribution of the data[85]. The principle behind KNN is to find k training positive and negative samples nearest in the distance to the new point, and predict the label from these samples. Firstly, the distance between the test sample point and each other sample point is calculated, then each distance will be sorted and k points with the smallest distance will be selected, and the categories of K points will be compared and classified. We used a MultiScheme package in WEKA to choose between 12 KNN models (1, 3, 5, 10, 20, 30, 50, 100, 150, 200, 250, 300) and the KNN with $k = 5$ yielded the best result.

Random forest (RF) is a classification algorithm that uses a set of decision trees[75]. Each decision tree is constructed by using a sample of training data, and each segmentation candidate set is a subset of random characteristics. RF has been proven to have excellent performance in classification tasks[77, 86, 87]. In this study, positive and negative samples are randomly selected from the original data to construct the sub-training set to generate the decision tree. At each node, we randomly selected the n child variables ($n \ll N$) from the N input variables. The optimal segmentation coefficients on these N sub-variables are used to segment the nodes. The n value remains constant during the growth of the forest. For new samples, the classification results can be obtained by voting on these decision trees. N is generally taken as the square root of the dimension of the eigenvector of the input samples. Here, we set *n_estimators* = 122 (the number of trees in the forest), and *max_depth* = 55 (the maximum depth of the tree).

Neural Networks (NN) plays an essential role in biomedicine[77], antiviral peptides prediction[72], protein-RNA interaction[88], and protein data mining[89]. For regular neural networks, the most common layer type is the fully connected layer in which neurons between two adjacent layers are fully pairwise connected. In the input layer, there are a certain number of neurons corresponding to input features. In the first layer (one-to-one layer), the same number of neurons are used, and each is connected to one neuron from the input layer. Then we added two hidden layers after the one-to-one layer. The first hidden layer is fully connected with one-to-one layer and second hidden layer is fully connected with the first hidden layer. The last layer is an output layer which only has two neurons. Batch normalization was applied to one-to-one layer and each hidden layer to accelerate the training process. SGD optimizer was used to train the DNN model and learning rate was fixed as 0.01.

Declarations

Availability

QSHGM, a database of 28,567 redundancy removal QS synthases (1,882) and receptors (26,685) entries for 818 gut microbes, which is freely available at: (http://www.qshgm.lbci.net/).

We will continuously update the database QSHGM.

Acknowledgement

This study was supported by the National Key Research and Development Project of China (No.2019YFA0905600, 2017YFD0201400), the National Natural Science Foundation of China (61772362), the Funds for Creative Research Groups of China (21621004), and National Key Research and Development Program of China (No. 2020YFA0907900).

Author contributions

J.Q. conceived the project. F. G. and A. Y. designed the project. S.W. conducted the systematic workflow and relevant analytical calculations. J. F. trained the reported data with different ML algorithms and constructed the database. H. W, Z. Q, J. G, S. S, X. H., Y. L, X. W. collected the reported and annotated QS entries. All authors analysed the results. S.W. wrote the manuscript. C.L., A. Y., F. G., and J.Q. edited the manuscript.

Conflict of interest

The authors declare no competing financial interests.

References

1. Donaldson GP, Lee SM, Mazmanian SK. Gut biogeography of the bacterial microbiota. *Nat. Rev. Microbiol.* **14**, 20–32 (2016).
2. Bäumler AJ, Sperandio V. Interactions between the microbiota and pathogenic bacteria in the gut. *Nature* **535**, 85–93 (2016).
3. Schluter J, *et al.* The gut microbiota is associated with immune cell dynamics in humans. *Nature* **588**, 303–307 (2020).
4. Neurath MF. Host-microbiota interactions in inflammatory bowel disease. *Nat Rev Gastroenterol Hepatol* **17**, 76–77 (2020).
5. Almeida A, *et al.* A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* **39**, 105–114 (2021).
6. Sung J, *et al.* Global metabolic interaction network of the human gut microbiota for context-specific community-scale analysis. *Nat. Commun.* **8**, 15393 (2017).
7. Goyal A, Wang T, Dubinkina V, Maslov S. Ecology-guided prediction of cross-feeding interactions in the human gut microbiome. *Nat. Commun.* **12**, 1335 (2021).
8. Fan Y, Pedersen O. Gut microbiota in human metabolic health and disease. *Nat. Rev. Microbiol.* **19**, 55–71 (2021).
9. Defoirdt T. Quorum-sensing systems as targets for antivirulence therapy. *Trends Microbiol.* **26**, 313–328 (2017).
10. Wu S, Liu J, Liu C, Yang A, Qiao J. Quorum sensing for population-level control of bacteria and potential therapeutic applications. *Cell. Mol. Life Sci.* **77**, 1319–1343 (2020).
11. Papenfort K, Bassler BL. Quorum sensing signal-response systems in gram-negative bacteria. *Nat. Rev. Microbiol.* **14**, 576–588 (2016).
12. Monnet V, Gardan R. Quorum-sensing regulators in gram-positive bacteria: 'Cherchez le peptide'. *Mol. Microbiol.* **97**, 181–184 (2015).
13. Pereira CS, Thompson JA, Xavier KB. Ai-2-mediated signalling in bacteria. *FEMS Microbiol. Rev.* **37**, 156–181 (2013).
14. Zarkan A, Liu J, Matuszewska M, Gaimster H, Summers DK. Local and universal action: The paradoxes of indole signalling in bacteria. *Trends Microbiol.* **28**, 566–577 (2020).
15. Stephens K, Bentley WE. Synthetic biology for manipulating quorum sensing in microbial consortia. *Trends Microbiol.* **28**, 633–643 (2020).
16. Song D, *et al.* Pseudomonas aeruginosa quorum-sensing metabolite induces host immune cell death through cell surface lipid domain dissolution. *Nat. Microbiol.* **4**, 97–111 (2018).
17. An SQ, *et al.* Modulation of antibiotic sensitivity and biofilm formation in pseudomonas aeruginosa by interspecies signal analogues. *Nat. Commun.* **10**, 2334 (2019).
18. Moura-Alves P, *et al.* Host monitoring of quorum sensing during pseudomonas aeruginosa infection. *Science* **366**, eaaw1629 (2019).
19. Mao N, Cubillos-Ruiz A, Cameron DE, Collins JJ. Probiotic strains detect and suppress cholera in mice. *Sci. Transl. Med.* **10**, eaao2586 (2018).
20. Hsiao A, *et al.* Members of the human gut microbiota involved in recovery from vibrio cholerae infection. *Nature* **515**, 423–426 (2014).
21. Lee JH, Wood TK, Lee J. Roles of indole as an interspecies and interkingdom signaling molecule. *Trends Microbiol.* **23**, 707–718 (2015).
22. Sedlmayer F, Hell D, Muller M, Auslander D, Fussenegger M. Designer cells programming quorum-sensing interference with microbes. *Nat. Commun.* **9**, 1822–1835 (2018).
23. Wu S, Xu C, Liu J, Liu C, Qiao J. Vertical and horizontal quorum sensing-based multicellular communications. *Trends Microbiol.* (2021). doi: 10.1016/j.tim.2021.04.006.
24. Zhang Q, *et al.* Gutmega: A database of the human gut metagenome atlas. *Brief Bioinform* **22**, bbaa082 (2021).
25. Wu S, *et al.* Gmrepo: A database of curated and consistently annotated human gut metagenomes. *Nucleic Acids Res.* **48**, D545-D553 (2020).

26. Poyet M, *et al.* A library of human gut bacterial isolates paired with longitudinal multiomics data enables mechanistic microbiome research. *Nat. Med.* **25**, 1442–1452 (2019).
27. Noronha A, *et al.* The virtual metabolic human database: Integrating human and gut microbiome metabolism with nutrition and disease. *Nucleic Acids Res.* **47**, D614-D624 (2019).
28. Rajput A, Kaur K, Kumar M. Sigmol: Repertoire of quorum sensing signaling molecules in prokaryotes. *Nucleic Acids Res.* **44**, 634–639 (2016).
29. Wynendaele E, *et al.* Quorumpeps database: Chemical space, microbial origin and functionality of quorum sensing peptides. *Nucleic Acids Res.* **41**, D655-659 (2013).
30. Barakat M, Ortet P, Whitworth DE. P2cs: A database of prokaryotic two-component systems. *Nucleic Acids Res.* **39**, D771-D776 (2011).
31. Ortet P, Whitworth DE, Santaella C, Achouak W, Barakat M. P2cs: Updates of the prokaryotic two-component systems database. *Nucleic Acids Res.* **43**, D536-D541 (2015).
32. Wu S, Liu C, Feng J, Yang A, Guo F, Qiao J. Qsidb: Quorum sensing interference molecules. *Brief Bioinform*, **22**, bbaa218 (2020).
33. Cheng L, Qi C, Zhuang H, Fu T, Zhang X. Gutmdisorder: A comprehensive database for dysbiosis of the gut microbiota in disorders and interventions. *Nucleic Acids Res.* **48**, D554-d560 (2020).
34. Zeng X, *et al.* Masi: Microbiota-active substance interactions database. *Nucleic Acids Res.* **49**, D776-D782 (2021).
35. Tang J, *et al.* Gimica: Host genetic and immune factors shaping human microbiota. *Nucleic Acids Res.* **49**, D715-D722 (2021).
36. Ye J, McGinnis S, Madden TL. Blast: Improvements for better sequence analysis. *Nucleic Acids Res.* **34**, W6-W9 (2006).
37. Kerfeld CA, Scott KM. Using blast to teach "e-value-tionary" concepts. *PLoS Biol* **9**, e1001014 (2011).
38. Bairoch A, *et al.* The universal protein resource (uniprot). *Nucleic Acids Res.* **33**, D154-159 (2005).
39. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10**, 845–858 (2015).
40. Gray KM, Garey JR. The evolution of bacterial luxi and luxr quorum sensing regulators. *Microbiology (Reading)* **147**, 2379–2387 (2001).
41. Colosimo DA, *et al.* Mapping interactions of microbial metabolites with human g-protein-coupled receptors. *Cell. Host. Microbe.* **26**, 273–282 e277 (2019).
42. Servetas SL, Carpenter BM, Haley KP, Gilbreath JJ, Gaddy JA, Merrell DS. Characterization of key helicobacter pylori regulators identifies a role for arsrs in biofilm formation. *J. Bacteriol.* **198**, 2536–2548 (2016).
43. Brameyer S, Kresovic D, Bode HB, Heermann R. Dialkylresorcinols as bacterial signaling molecules. *Proceedings of the National Academy of Sciences* **112**, 572–577 (2015).
44. Parashar V, Aggarwal C, Federle MJ, Neiditch MB. Rgg protein structure-function and inhibition by cyclic peptide compounds. *Proc Natl Acad Sci U S A* **112**, 5177–5182 (2015).
45. Yang T, Talgan Y, Paharik AE, Horswill AR, Blackwell HE. Structure-function analyses of a staphylococcus epidermidis autoinducing peptide reveals motifs critical for agrc-type receptor modulation. *ACS Chem. Biol.* **11**, 1982 (2016).
46. Erez Z, *et al.* Communication between viruses guides lysis–lysogeny decisions. *Nature* **541**, 488–493 (2017).
47. Bivar Xavier K. Bacterial interspecies quorum sensing in the mammalian gut microbiota. *C R Biol* **341**, 297–299 (2018).
48. Wang D, Ding X, Rather PN. Indole can act as an extracellular signal in escherichia coli. *J. Bacteriol.* **183**, 4210–4216 (2001).
49. Jaglin M, *et al.* Indole, a signaling molecule produced by the gut microbiota, negatively impacts emotional behaviors in rats. *Front. Neurosci.* **12**, 216 (2018).
50. Nguyen Y, *et al.* Structural and mechanistic roles of novel chemical ligands on the sdia quorum-sensing transcription regulator. *Mbio* **6**, e02429-02414 (2015).
51. Chu W, *et al.* Indole production promotes escherichia coli mixed-culture growth with pseudomonas aeruginosa by inhibiting quorum signaling. *Appl. Environ. Microbiol.* **78**, 411 (2012).
52. Chapalain A, *et al.* Interplay between 4-hydroxy-3-methyl-2-alkylquinoline and n-acyl-homoserine lactone signaling in a burkholderia cepacia complex clinical strain. *Front Microbiol* **8**, 1021 (2017).

53. Wellington S, Greenberg EP. Quorum sensing signal selectivity and the potential for interspecies cross talk. *mBio* **10**, e00146-00119 (2019).
54. Wang S, Payne GF, Bentley WE. Quorum sensing communication: Molecularly connecting cells, their neighbors, and even devices. *Annu. Rev. Chem. Biomol. Eng.* **11**, 447–468 (2020).
55. Kumar P, Lee J-H, Lee J. Diverse roles of microbial indole compounds in eukaryotic systems. *Biol Rev Camb Philos Soc*, (2021). doi: 10.1111/brv.12765.
56. Case RJ, Labbate M, Kjelleberg S. Ahl-driven quorum-sensing circuits: Their frequency and function among the proteobacteria. *ISME J* **2**, 345–349 (2008).
57. Whiteley M, Diggle SP, Greenberg EP. Progress in and promise of bacterial quorum sensing research. *Nature* **551**, 313–320 (2017).
58. Vrancken G, Gregory AC, Huys GRB, Faust K, Raes J. Synthetic ecology of the human gut microbiota. *Nat. Rev. Microbiol.* **17**, 754–763 (2019).
59. Consortium THMP. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
60. Faust K, *et al.* Microbial co-occurrence relationships in the human microbiome. *PLoS Comput. Biol.* **8**, e1002606 (2012).
61. Atkinson S, Williams P. Quorum sensing and social networking in the microbial world. *J R Soc Interface* **6**, 959–978 (2009).
62. Venturelli OS, *et al.* Deciphering microbial interactions in synthetic human gut microbiome communities. *Mol. Syst. Biol.* **14**, e8157 (2018).
63. Faust K, Raes J. Microbial interactions: From networks to models. *Nat. Rev. Microbiol.* **10**, 538–550 (2012).
64. Ran J, *et al.* Construction and analysis of the protein-protein interaction network related to essential hypertension. *BMC Syst. Biol.* **7**, 32 (2013).
65. Eckburg PB, *et al.* Diversity of the human intestinal microbial flora. *Science* **308**, 1635 (2005).
66. Zou Y, *et al.* 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. *Nat. Biotechnol.* **37**, 179–185 (2019).
67. Sandoz KM, Mitzimberg SM, Schuster M. Social cheating in pseudomonas aeruginosa quorum sensing. *Proc. Natl Acad. Sci. USA* **104**, 15876–15881 (2007).
68. Y Y, MA M-Y, TJ D, AP B, PE W, HJ D. Structure of the escherichia coli quorum sensing protein sdia: Activation of the folding switch by acyl homoserine lactones. *J Mol Biol* **355**, 262–273 (2006).
69. Hawver LA, Jung SA, Ng WL. Specificity and complexity in bacterial quorum-sensing systems. *FEMS Microbiol. Rev.* **40**, 738–752 (2016).
70. Zhao B, *et al.* Describeprot: Database of amino acid-level protein structure and function predictions. *Nucleic Acids Res.* **49**, D298-D308 (2021).
71. Ding Y, Tang J, Guo F. Identification of drug–target interactions via dual laplacian regularized least squares with multiple kernel fusion. *Knowledge-Based Systems* **204**, 106254 (2020).
72. Li J, Pu Y, Tang J, Zou Q, Guo F. Deepavp: A dual-channel deep neural network for identifying variable-length antiviral peptides. *IEEE J Biomed Health Inform* **24**, 3012–3019 (2020).
73. Ding Y, Tang J, Guo F. Identification of protein-ligand binding sites by sequence information and ensemble classifier. *J. Chem. Inf. Model.* **57**, 3149–3161 (2017).
74. Cortes C, Vapnik V. Support-vector networks. *Machine learning* **20**, 273–297 (1995).
75. Breiman L. Random forests. *Machine learning* **45**, 5–32 (2001).
76. Peterson LE. K-nearest neighbor. *J Scholarpedia* **4**, 1883 (2009).
77. Wainberg M, Merico D, Delong A, Frey BJ. Deep learning in biomedicine. *Nat. Biotechnol.* **36**, 829–838 (2018).
78. Chen Z, *et al.* Ifeature: A python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics* **34**, 2499–2502 (2018).
79. Kumar S, Stecher G, Li M, Knyaz C, Tamura K, Battistuzzi FU. Mega x: Molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018).

80. Letunic I, Bork P. Interactive tree of life (itol) v4: Recent updates and new developments. *Nucleic Acids Res.* **47**, W256-W259 (2019).
81. Saitou N, Nei M. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**, 406–425 (1987).
82. Felsenstein J. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* **39**, 783–791 (1985).
83. Meng C, Guo F, Zou Q. Cwly-svm: A support vector machine-based tool for identifying cell wall lytic enzymes. *Comput. Biol. Chem.* **87**, 107304 (2020).
84. Liu B, Li CC, Yan K. Deepsvm-fold: Protein fold recognition by combining support vector machines and pairwise sequence similarity scores generated by deep learning networks. *Brief Bioinform* **21**, 1733–1741 (2020).
85. Royce TE, Rozowsky JS, Gerstein MB. Toward a universal microarray: Prediction of gene expression through nearest-neighbor probe sequence identification. *Nucleic Acids Res.* **35**, e99 (2007).
86. Radivojevic T, Costello Z, Workman K, Garcia Martin H. A machine learning automated recommendation tool for synthetic biology. *Nat. Commun.* **11**, 4879 (2020).
87. Degenhardt F, Seifert S, Szymczak S. Evaluation of variable selection methods for random forests and omics data sets. *Brief Bioinform* **20**, 492–503 (2019).
88. Lam JH, *et al.* A deep learning framework to predict binding preference of rna constituents on protein surface. *Nat. Commun.* **10**, 4941 (2019).
89. Shi Q, Chen W, Huang S, Wang Y, Xue Z. Deep learning for mining protein data. *Brief Bioinform* **22**, 194–218 (2021).

Figures

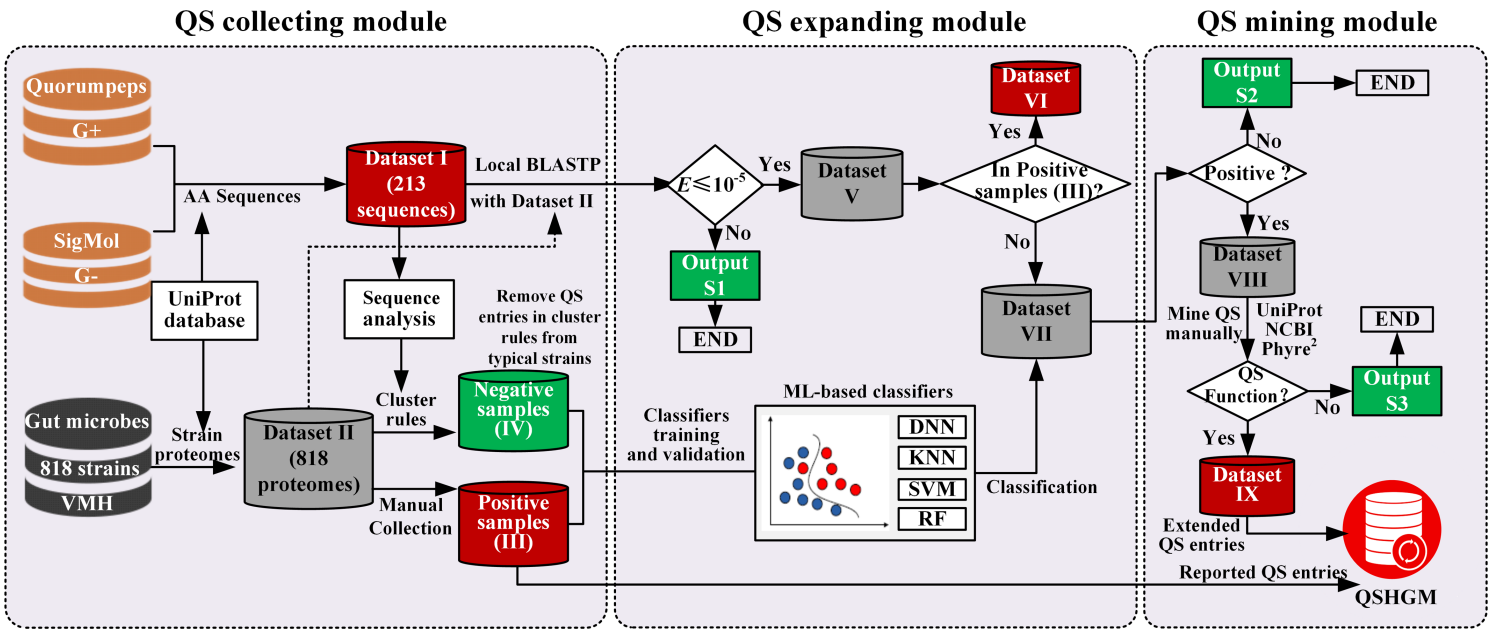


Figure 1

Schematic diagram of the systematic workflow including three modules. There are nine engaged datasets in our systematic workflow, i.e., 213 validated QS entries from Gram-positive (G+) and Gram-negative (G-) microbes (Dataset I), 818 proteomes for the gut microbiota from VMH and UniProt (Dataset II), positive samples collected manually from dataset I (Dataset III), negative samples obtained from dataset I (Dataset IV), results of local BLASTP with E ≤ 10-5 (Dataset V), overlaps of the reported QS entries in dataset III and V (Dataset VI), proteins dataset excluded dataset VI for dataset V (Dataset VII), positive ones classified by different ML-based classifiers (Dataset VIII), and potential QS entries (Dataset IX). There are another three abandoned datasets in the workflow of the systematic workflow, i.e., protein dataset with E ⩾ 10-5 (Output S1), negative ones classified by ML-based classifiers (Output S2), and proteins without QS functions (Output S3).

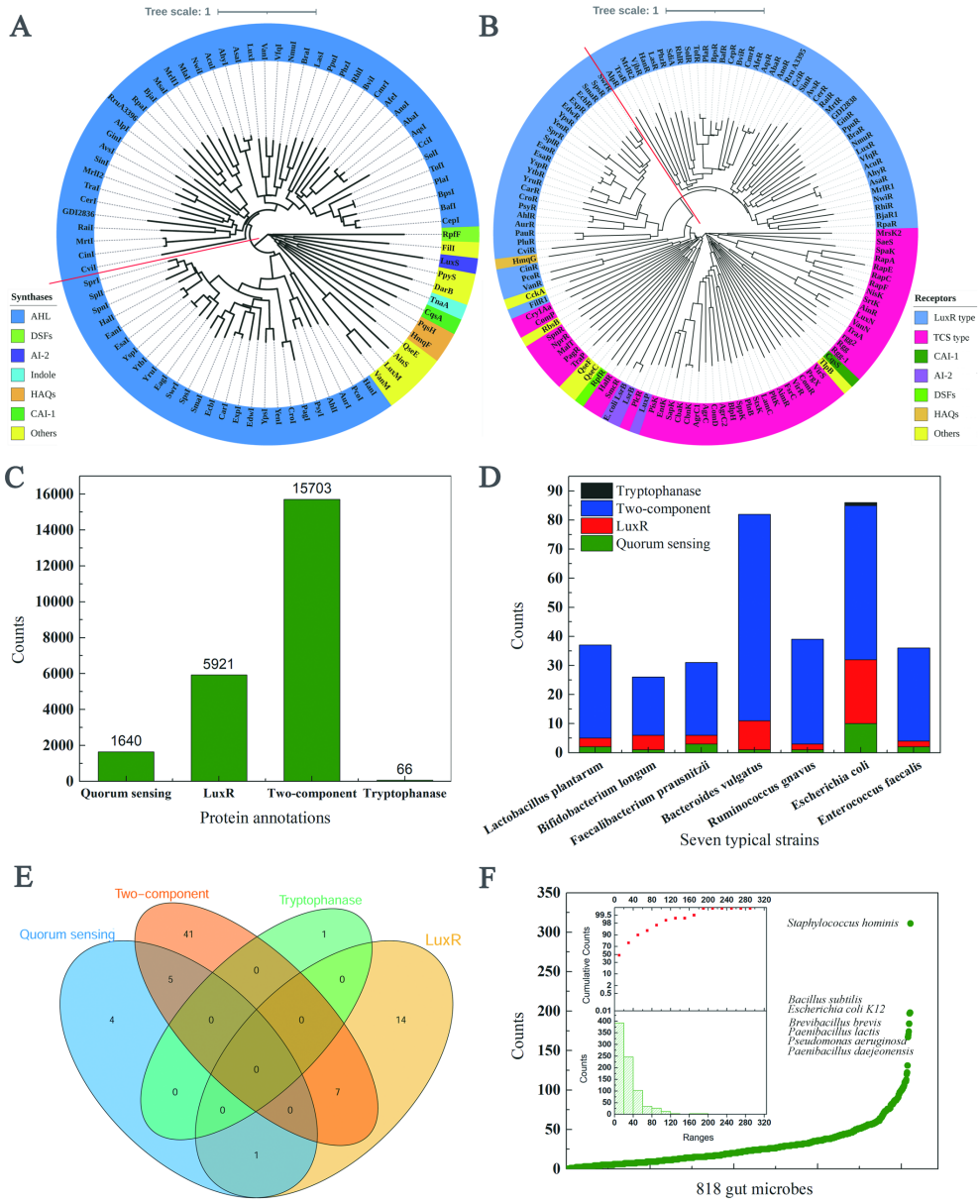


Figure 2

Results of collections of the reported and annotated QS entries. Evolutionary trees of QS synthases (A) and receptors (B). A. The optimal tree with the sum of branch length = 40.33 is shown. This analysis involves 84 amino acid sequences, and there are a total of 1,374 positions in the final dataset. B. The optimal tree with the sum of branch length = 91.14 is shown. This analysis involves 129 amino acid sequences, and there are a total of 1,010 positions in the final dataset. C. Total QS entries with four protein annotations, i.e., “quorum sensing”, “LuxR”, “two-component”, and “tryptophanase”. D. QS entries distribution of the seven-strain simplified human gut microbes used by Colosimo et al41. E. The overlap of the four types of QS entries in Escherichia coli O157:H7 strain. F. QS entries counts distribution of 818 human gut microbes from the VMH database27.

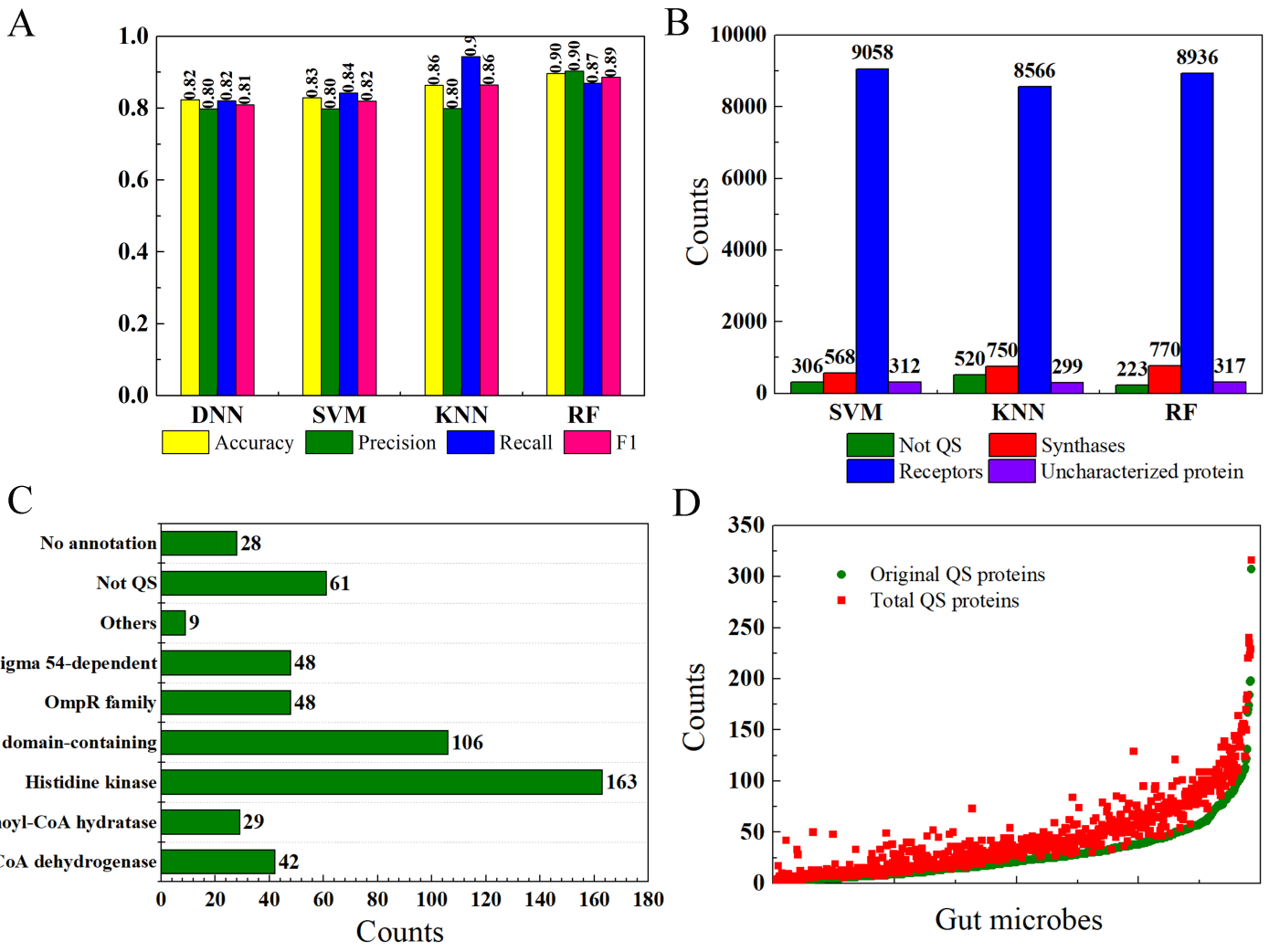
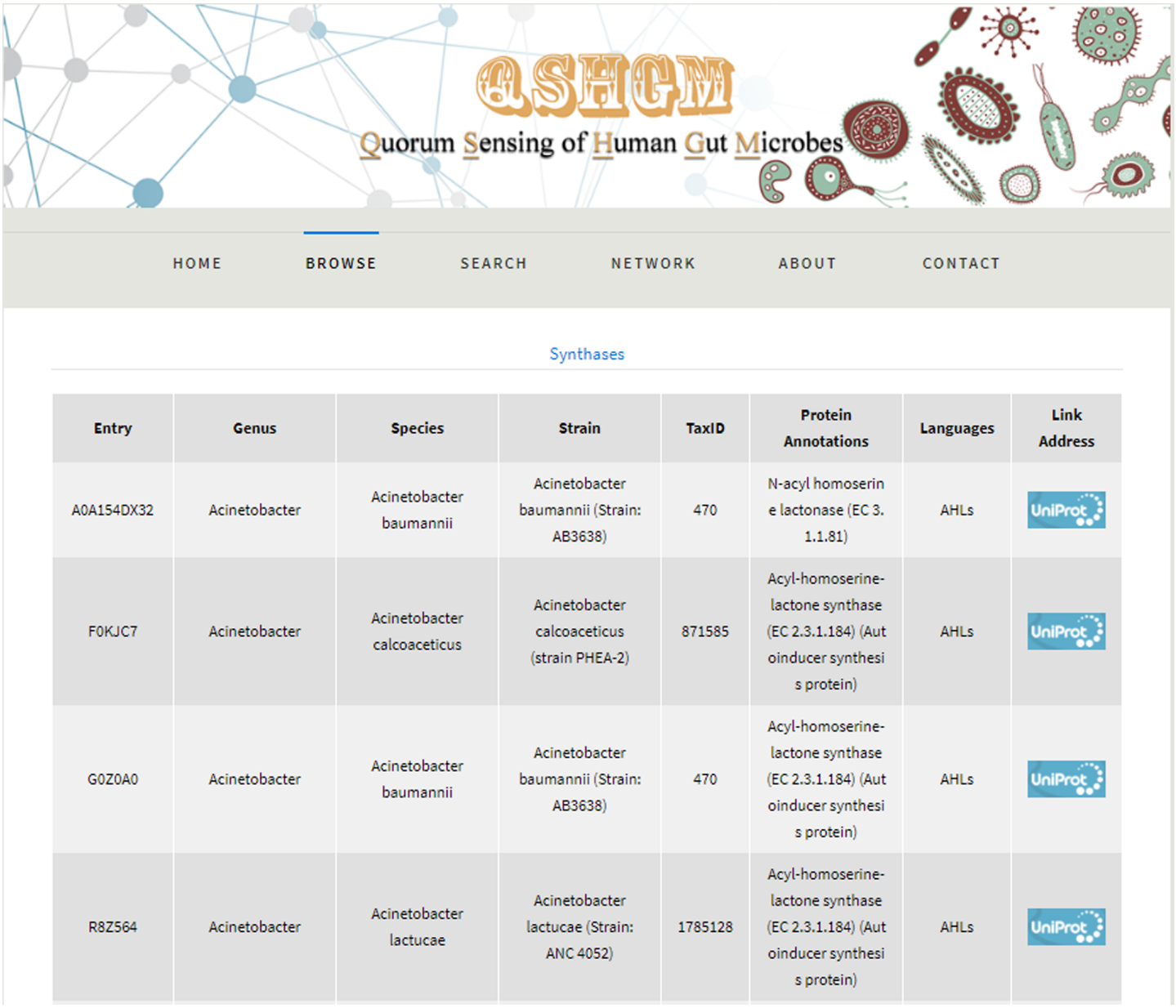


Figure 3

Results of the expansion and mining for QS entries based on the proposed systematic workflow. A. Accuracy, precision, recall, and F1 score of the four classifiers based on SVM, KNN, RF, and DNN algorithms. B. Groups classification for the positive ones of SVM, KNN, and RF classifiers based on protein annotations; C. Results of the protein clusters of 534 re-annotated protein entries. D. Distribution of the total 28,567 redundancy removal QS entries in 818 gut microbes.



Entry	Genus	Species	Strain	TaxID	Protein Annotations	Languages	Link Address
A0A154DX32	Acinetobacter	Acinetobacter baumannii	Acinetobacter baumannii (Strain: AB3638)	470	N-acyl homoserine lactonase (EC 3.1.1.81)	AHLs	UniProt
F0KJC7	Acinetobacter	Acinetobacter calcoaceticus	Acinetobacter calcoaceticus (strain PHEA-2)	871585	Acyl-homoserine-lactone synthase (EC 2.3.1.184) (Autoinducer synthesis protein)	AHLs	UniProt
G0Z0A0	Acinetobacter	Acinetobacter baumannii	Acinetobacter baumannii (Strain: AB3638)	470	Acyl-homoserine-lactone synthase (EC 2.3.1.184) (Autoinducer synthesis protein)	AHLs	UniProt
R8Z564	Acinetobacter	Acinetobacter lactucae	Acinetobacter lactucae (Strain: ANC 4052)	1785128	Acyl-homoserine-lactone synthase (EC 2.3.1.184) (Autoinducer synthesis protein)	AHLs	UniProt

Figure 4

Part of browsing results for AHLs language in QSHGM database.

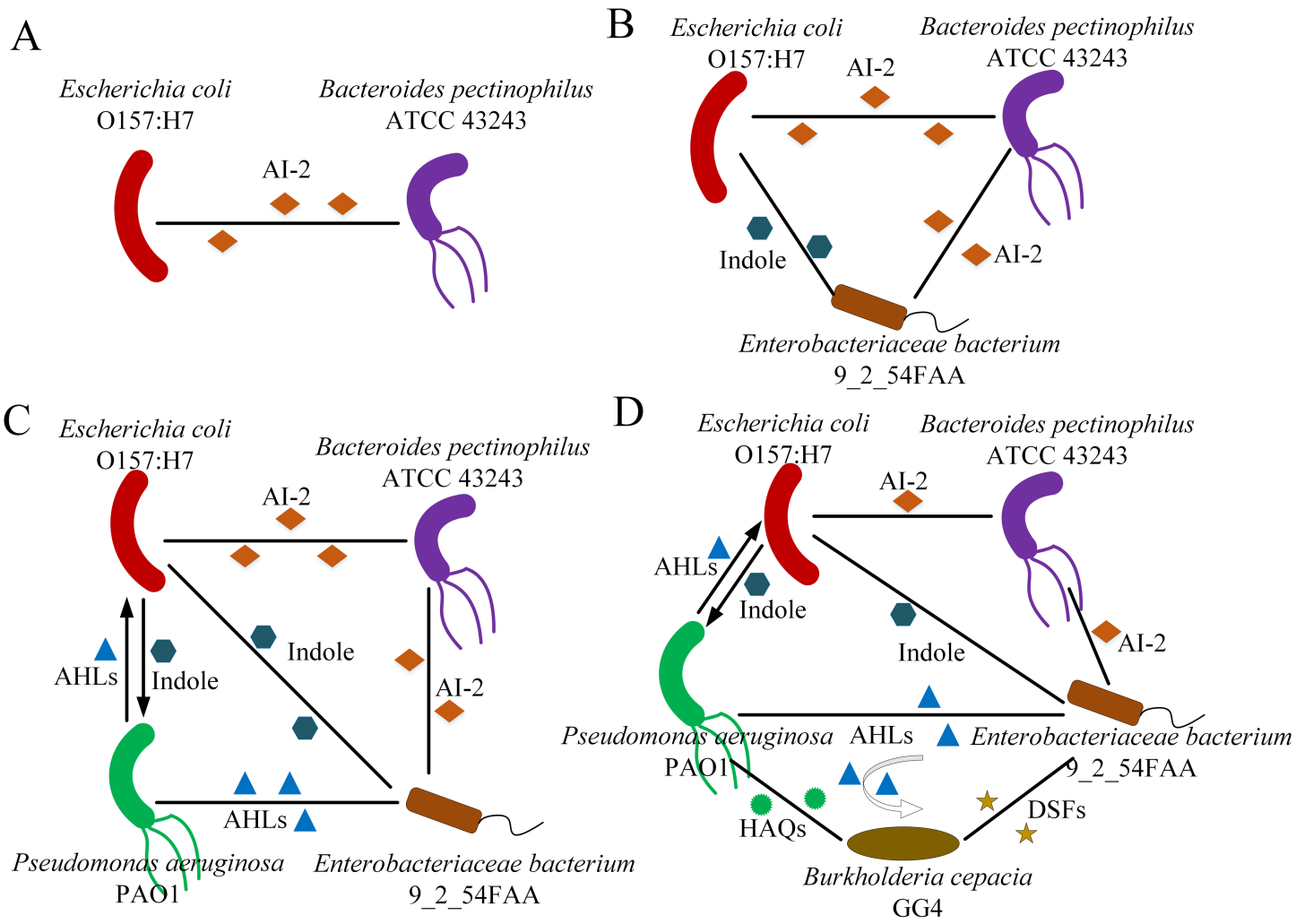


Figure 5

QS-based interactions predictions for two-strain (A), three-strain (B), four-strain (C), and five-strain (D) consortium based on diverse QS languages.

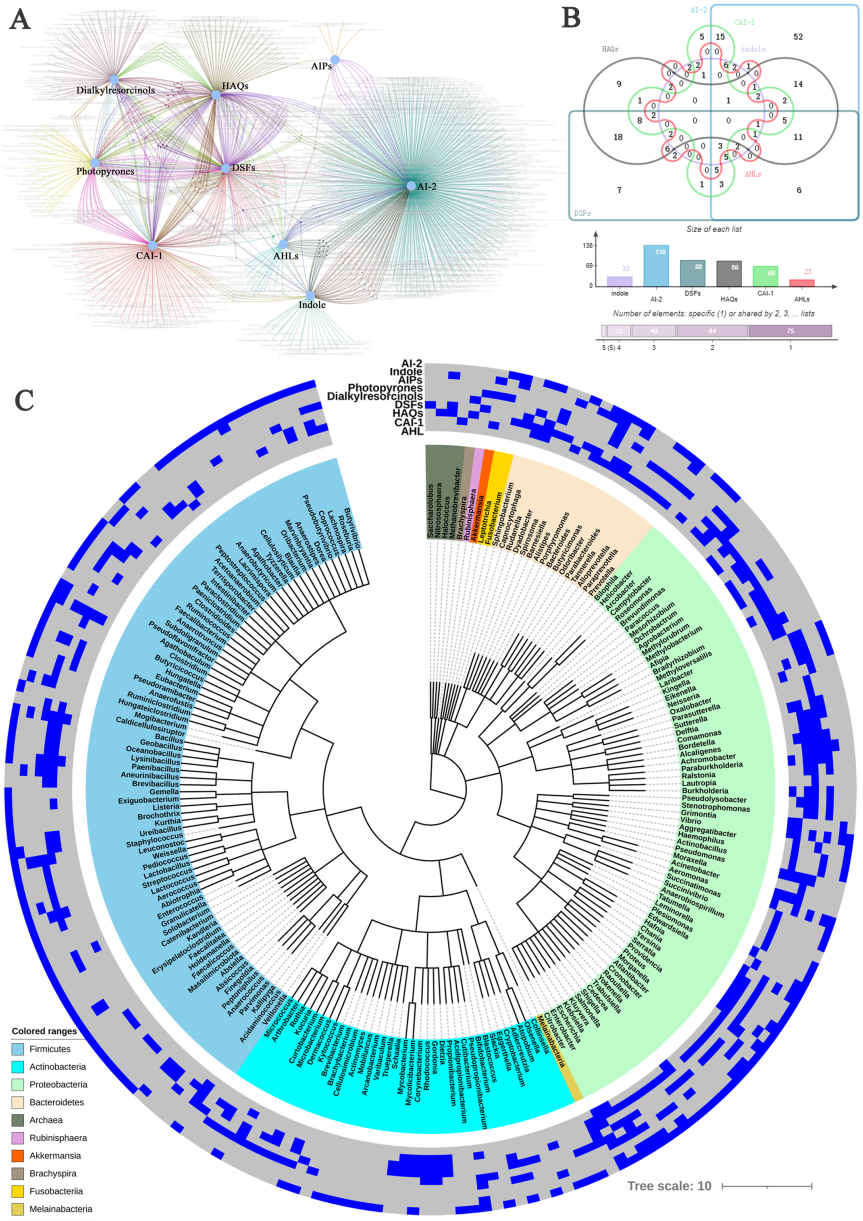


Figure 6

QS communication network for human gut microbiota based on QS languages. A. QS communication network for 818 human gut microbes based on nine languages. Note that the network diagram was generated using EVenn (http://www.ehbio.com/test/venn). B. Microbial genus distribution for six typical QS languages, i.e., AHLs, CAI-1, HAQs, DSFs, Indole, and AI-2. C. Hierarchical clustering of nine QS languages found in 210 human gut microbial genus. The constructions are classified into 10 genus-level clusters based on their phyla and taxonomy. Microbial genus from Firmicutes is coloured in blue; Actinobacteria, cyan; Proteobacteria, green; Bacteroidetes, yellowish. Heatmap on the outermost layer indicates QS languages distribution in each cluster, existence is coloured in blue; no existence, grey.

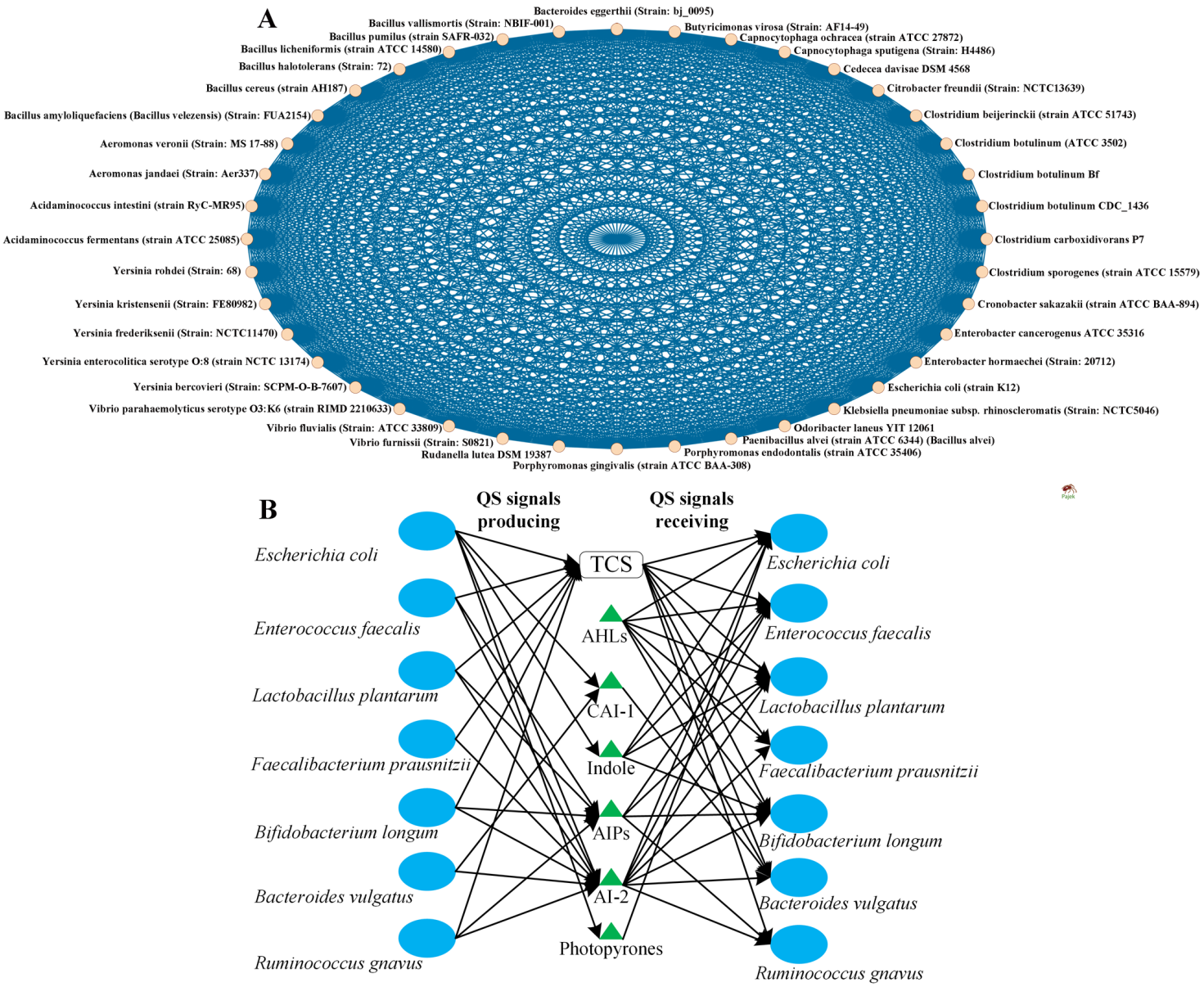


Figure 7

Extended QS-based microbial interaction networks. A. a complete graph that links every node to every other node for the 40 core microbes; B. A typical small QS communication network that includes QS signals producing and receiving for the seven human gut species form Colosimo et al41.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarymaterial.docx