














Machine learning aided construction of the quorum sensing communication network for human gut microbiota

Shengbo Wu ^{1,2}, Jie Feng ³, Chunjiang Liu ^{1,2}, Hao Wu ⁴, Zekai Qiu ¹, Jianjun Ge ¹, Shuyang Sun ¹, Xia Hong ¹, Yukun Li ¹, Xiaona Wang ¹, Aidong Yang ⁵✉, Fei Guo ⁶✉ & Jianjun Qiao ^{1,4,7}✉

Quorum sensing (QS) is a cell-cell communication mechanism that connects members in various microbial systems. Conventionally, a small number of QS entries are collected for specific microbes, which is far from being able to fully depict communication-based complex microbial interactions in human gut microbiota. In this study, we propose a systematic workflow including three modules and the use of machine learning-based classifiers to collect, expand, and mine the QS-related entries. Furthermore, we develop the Quorum Sensing of Human Gut Microbes (QSHGM) database (<http://www.qshgm.lbcj.net/>) including 28,567 redundancy removal entries, to bridge the gap between QS repositories and human gut microbiota. With the help of QSHGM, various communication-based microbial interactions can be searched and a QS communication network (QSCN) is further constructed and analysed for 818 human gut microbes. This work contributes to the establishment of the QSCN which may form one of the key knowledge maps of the human gut microbiota, supporting future applications such as new manipulations to synthetic microbiota and potential therapies to gut diseases.

¹School of Chemical Engineering and Technology, Tianjin University, Tianjin 300072, China. ²State Key Laboratory of Chemical Engineering, Tianjin University, Tianjin 300072, China. ³School of Computer Science and Technology, College of Intelligence and Computing, Tianjin University, Tianjin 300350, China. ⁴Zhejiang Shaoxing Research Institute of Tianjin University, Shaoxing 312300, China. ⁵Department of Engineering Science, University of Oxford, Oxford OX1 3PJ, UK. ⁶School of Computer Science and Engineering, Central South University, Changsha 410083, China. ⁷Key Laboratory of Systems Bioengineering, Ministry of Education (Tianjin University), Tianjin 300072, China. ✉email: aidong.yang@eng.ox.ac.uk; guofei@csu.edu.cn; jianjunq@tju.edu.cn

Human gut microbiota is a dynamic and complex microbial system¹ that links to the pathogen colonization resistance², immune system regulation³, and human health maintenance⁴. Recent breakthroughs in high-throughput screening and multi-omics technologies have enabled the detection and quantification of the microbiota composition⁵ 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 diseases⁸.

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 interactions, which are closely relevant to various diseases¹⁵. For example, *N*-(3-oxodecanoyl)-L-homoserine lactone, a common AHL-type signal, plays an important role in the modulation of the gut immune system by inducing neutrophils apoptosis¹⁶ and attenuating innate immune responses via disruption of NF- κ B signaling¹⁷, thus providing better colonization for *Pseudomonas aeruginosa* in the host. DSF analogs were verified to strengthen the mucosal barrier and reduce antibiotic tolerance of *P. aeruginosa*¹⁸. Different hosts can utilize the aryl hydrocarbon receptor (AhR) to “listen in” the concentration of the HAQs from *P. aeruginosa* to regulate immune responses dynamically¹⁹. 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 turn²⁰. AI-2 produced by *Ruminococcus obeum* could repress several colonization factors of *Vibrio cholerae*, thus restricting the colonization of *V. cholerae*, which leads to diarrheal diseases²¹. 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^{14,22}. The evidence stated above suggests that manipulations of the level of diverse QS languages such as AI-2²³ in microbial communication play an important role in diverse host-centric applications for gut microbial systems. Therefore, in our previous study²⁴, we have proposed the “QS communication network” (QSCN), a unifying concept for vertical and horizontal QS-based interactions implemented through producing, transducing, and responding to QS signaling molecules, to indicate its important role in host-centric probiotic manipulations and various practical applications of synthetic microbial consortia. QSCN calls for a comprehensive QS database, which includes the collections of human gut microbes and QS repository, to bridge the gap between existing QS-related 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 research. With respect to the gut microbiota, the gutMEGA database²⁵ contains thousands of gut microbiota compositions (metagenomic sequences), phenotypes, and experimental information. GMrepo²⁶ focuses on the annotated human gut metagenomes to facilitate the development of human metagenomic data. BIO-ML²⁷ includes 7,758 gut bacterial isolates, 3632 genome sequences, and diverse longitudinal multi-omics data. Particularly, VMH²⁸ 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 SigMol²⁹ and Quorumpeps³⁰, respectively. P2CS^{31,32} was constructed and updated for a two-component system (TCS), which is a typical communicating system that is composed of a histidine kinase receptor and a response regulator partner. Furthermore, we have previously developed the QSIdb database³³ 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 gutMDisorder³⁴ have linked the human microbiota and many macro-environmental factors together to describe the intervention and regulation of various diseases. In addition, exogenous active substances and endogenous host factors were also collected for human microbiota into MASI³⁵ and GIMICA³⁶, respectively, to provide information on 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 communication-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 integration to form a relatively complete network is yet to be further explored. Although some biological networks such as metabolite-based interaction networks have been relatively mature, they cannot decipher complex regulatory relationships among microorganisms, thus leading to the incompleteness of microbial interaction networks.

In this study, we aim to address the above deficiencies through a combination of various methods (a framework diagram can be found in Supplementary Fig. 1). We firstly developed a systematic workflow including entry collecting, expanding, and mining modules to construct a QS repository for human gut microbiota. In the collecting module, due to the intricate overlaps on QS and two-component system (TCS) entries, we curated the annotated QS and TCS (QS&TCS) entries carefully for each component in the human gut microbiota to form a repository of reported QS entries as inclusively as possible. Information gathering in this module was also combined with machine learning (ML) algorithms including random forest (RF), k-nearest neighbor (KNN), support vector machine (SVM), and deep neural network (DNN) to develop four classifiers, which were then used in the expanding module to nominate further candidates of human gut QS entries from existing (general-purpose) QS databases. These candidates were finally analysed in the mining module, where protein annotation, functional analysis, and homologous modeling were combined to re-annotate and mine QS entries. These have led to a QS database of human gut microbiota (QSHGM, <http://www.qshgm.lbc.net/>) including the reported (21,383) and extended (7184) QS entries, which offers user-friendly browsing and searching functions to support various applications. With the help of QSHGM, we can search complex regulation-based interactions for different microbial consortia and further constructed a QSCN to visualize and decipher intricate QS-based interactions for human gut microbiota. Finally, we identified key challenges and suggest directions for the QSCN and how we can engineer them to provide more future applications.

Results

The systematic workflow for QSHGM. We developed a systematic workflow which includes three modules (collecting,

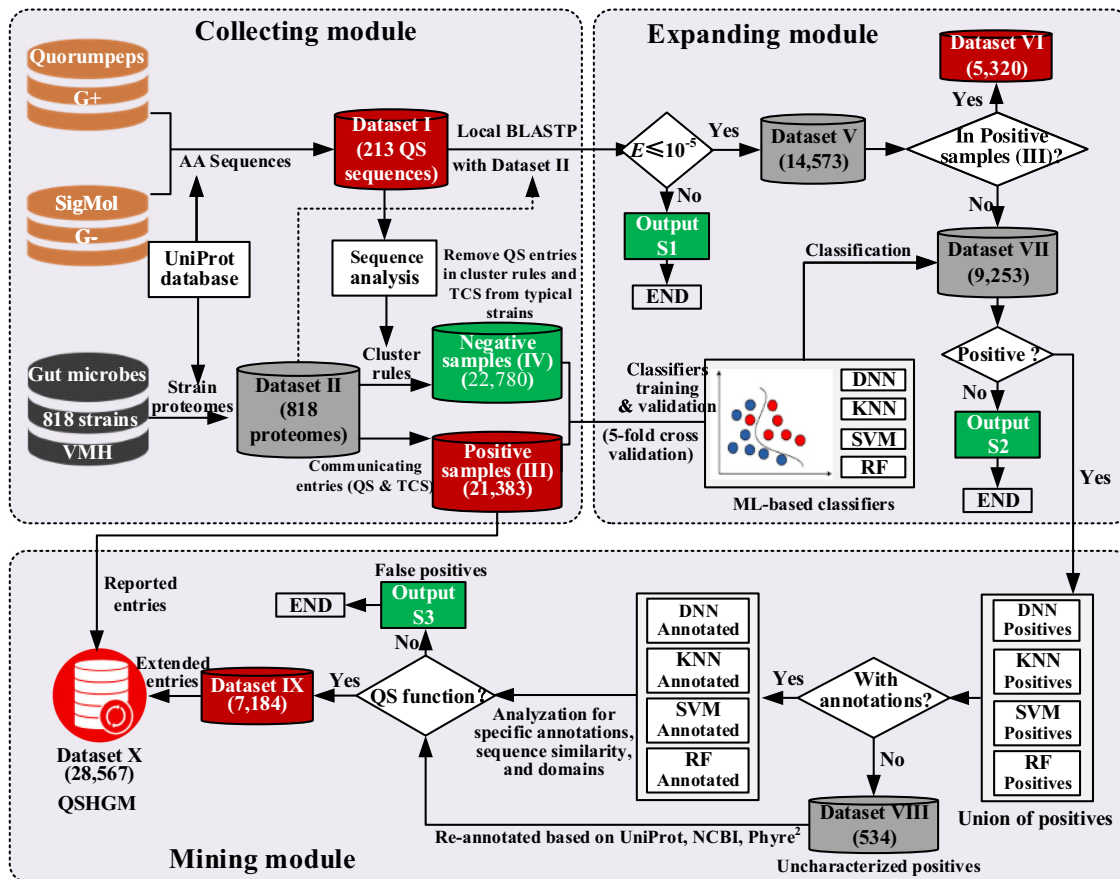


Fig. 1 Schematic diagram of the systematic workflow including three modules. There are ten engaged datasets in our systematic workflow, i.e., 213 validated QS entries from Gram-positive (G+) and Gram-negative (G-) microbes (Dataset I) (Supplementary Data 1), 818 proteomes for the gut microbiota from VMH and UniProt (Dataset II, <https://pan.baidu.com/s/1o46nn1b7L5nvCqgpwW7Zlw>. Password: tfnx), positive samples collected manually from dataset I (Dataset III) (Supplementary Data 2), negative samples obtained from dataset I (Dataset IV) (Supplementary Data 3), results of local BLASTP with $E \leq 10^{-5}$ (Dataset V) (Supplementary Data 4), overlaps of the reported QS entries in dataset III and V (Dataset VI) (Supplementary Data 5), proteins dataset excluded dataset VI for dataset V (Dataset VII) (Supplementary Data 6), uncharacterized positives classified by different ML-based classifiers (Dataset VIII) (Supplementary Data 7), extended QS entries (Dataset IX) (Supplementary Data 8), and total data for QSHGM (Dataset X) (Supplementary Data 10). There are another three abandoned datasets in the workflow of the systematic workflow, i.e., protein datasets with $E > 10^{-5}$ (Output S1), negative ones classified by ML-based classifiers (Output S2), and proteins without QS functions (false positives) (Output S3, Supplementary Data 9). Details of the above datasets are provided in Supplementary Table 1. Note that positive/negative/mixed datasets are colored in red/green/gray, respectively.

expanding, and mining modules) and four classifiers based on ML algorithms to construct a QS repository for human gut microbiota (Fig. 1). In the 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. TCS entries play an important role in microbial communications, which overlap with QS, but it is difficult to separate them clearly. In this work, we started by manually searching the 818 gut microbes from the VMH database²⁸ (Dataset II) to collect reported both QS and TCS (QS&TCS) entries which are termed “positive samples” (Dataset III, 21,383 entries) to cover the reported QS entries as inclusively as possible for constructing a comprehensive microbial communication database. The manual search was based on commonly used QS (“quorum sensing”, “LuxR”, “tryptophanase”) or TCS (“two-component”) annotations. The negative samples (Dataset IV, 22,780 entries) were then obtained by removing QS&TCS entries 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 expanding module, we obtained an extended dataset (Dataset V, 14,573 entries) from the results of the local BLASTP³⁷ on Dataset I and II with the criteria of the E value³⁸ being smaller than 10^{-5} , which is commonly used in the sequence alignment to obtain homologs. Four different ML algorithms (DNN, SVM, RF, and KNN) were used to construct classifiers, which were trained and validated based on the above positive (III) and negative samples (IV) to obtain more potential QS entries. After excluding from Dataset V those which were already collected as the reported QS&TCS entries in dataset III (Dataset VI, 5320 entries), the remaining entries (Dataset VII, 9,253 entries) were then classified by the four ML-based classifiers stated above. The output of these classifiers was further processed in the mining module, where the union of the four positives predicted by the four classifiers were divided into uncharacterized positives (Dataset VIII, 534 entries) and annotated positives. The uncharacterized positives were re-annotated, mined, and sorted out manually with the help of UniProt³⁹, NCBI (<https://www.ncbi.nlm.nih.gov/>) and Phyre² databases⁴⁰. Furthermore, we

conducted the function analysis by checking their specific annotations, sequence similarity, and domains (see more details in Supplementary Data 11) for the annotated/re-annotated union of positives to decide whether the entry has a QS function (true positives, Dataset IX, 7,184 entries) or not (false positives, Output S3, 438 entries), if so, whether it is a QS synthase or a QS receptor. A combination of manual curation, BLASTP-based expanding, and multiple ML-based classifications helped us obtain as many potential QS entries as possible. Finally, the extended QS entries and the reported QS&TCS entries were combined together to form the QSHGM (Dataset X, 28,567 entries) database (<http://www.qshgm.lbc.net/>).

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, b), part of which was also identified by Gray et al.⁴¹. This indicates that there is coevolution for AHLs synthases and their corresponding receptors.

There are 1640, 5921, 66, and 15,703 entries for “quorum sensing”, “LuxR”, “tryptophanase”, and “two-component”, respectively (Fig. 2c). LuxR-type and TCS entries account for the vast majority, which are 25.38 and 67.31%, respectively. We have also shown the distribution of QS&TCS entries for each strain based on the seven-strain simplified human microbiomes (SIHUMIs) used by Colosimo et al.⁴² (Fig. 2d). This verified that LuxR-type QS and TCS entries account for the vast majority of QS&TCS entries in these strains. Furthermore, we noted that there are certain overlaps in the distribution of the four entries. For example, there are seven entries (P69409, P0ACZ6, P0AGA8, P66798, P0AF30, P0AEL9, and Q8XE66) in the *E. coli* O157:H7 strain (Fig. 2e), which are both LuxR-type and TCS receptors. In addition, we have counted and distributed the total QS&TCS entries of the 818 gut microbes from the VMH database²⁸ to form a better picture of the QS repository in human gut microbiota (Fig. 2f). According to the cumulative distribution curve for the statistics (Fig. 2f subgraph), we found that about 90% of strains contain less than 60 QS&TCS entries, and only seven strains have more than 150 entries. This distribution will be revisited after extended QS entries are included (see below).

Expanded QS entries. The amino acid composition (AAC) calculates the frequency of each amino acid type in a protein sequence. The frequencies of all 20 natural amino acids are the percent of the number of amino acid types divided by the length of a protein sequence⁴³ (more details in the Method section). We calculated the frequency of each amino acid type in each entry sequence as the protein features, and we conducted a fivefold cross-validation to train classifiers using the positive (Dataset III) and negative samples (Dataset IV), where the average accuracy, prediction, recall, and F1 score (more details were listed in method section) were applied to evaluate their performances. The results show that the performances of the DNN, SVM, KNN, and

RF classifiers were not very different, with the RF-based classifier being slightly more prominent (Fig. 3a). We then manually checked the annotations of the predicted results from the classification of the four ML-based classifiers on Dataset VII and divided the four positives into annotated positives and uncharacterized positives (Dataset VIII, 534 entries) (Fig. 3b), which were analysed further for their specific overlaps (Supplementary Fig 2) (see more details in Supplementary Data 12). In order to obtain as many potential QS entries as possible, it was helpful to combine the four positives from the four classifiers together to form a union.

With the help of the functional analysis (Supplementary Data 11), we then re-annotated the 534 uncharacterized 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 were vaguely described without specific protein annotations (Fig. 3c). As listed in Table 1 and Table 2, these entries were further explored and re-annotated based on the web BLASTP of the NCBI database and Phyre², respectively. There were 20 proteins (Table 1) that can be re-annotated based on the BLASTP results from NCBI. Except for 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 the human gut⁴⁴. 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*⁴⁵. The histidine kinase, LuxR family regulator, and Rgg/GadR/MutR family regulator are important parts of TCS, LuxR-type, and Rgg-based QS systems⁴⁶, respectively.

There were eight entries (Table 2) that have no specific annotations or classifications in NCBI or UniProt database. We submitted these protein sequences to Phyre² to investigate the 2D and 3D structures of their models, their domain compositions, and model quality. A0A4Y4IIW5 and A0A5C4P2T9 are signaling proteins and AgrC (belonging to Agr QS system⁴⁷) family proteins, 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 signaling 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 decisions⁴⁸. 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.

Furthermore, we have further conducted the function analysis and checked their specific annotations, sequence similarity, and domains for the annotated/re-annotated union of positives (Supplementary Data 11) to decide whether the entry has a QS function (true positives, Dataset IX, 7184 entries, more details in Supplementary Data 8) or not (false positives, Output S3, 438 entries, more details in Supplementary Data 9). Finally, the reported QS&TCS and extended QS entries were combined together to be Dataset X (28,567 entries). To sum up, with the help of the proposed systematic workflow (Fig. 1), we obtained a comprehensive QS repository including the manually collected 21,383 positive samples (Database III) and the extended 7184 ones (Database IX) for 818 gut microbes, and the total 28,567 entries (Database X) are composed of 1882 QS synthases and 26,685 receptors. There was a 33.60% increase in extended entries (Database IX) for Dataset X (Fig. 3d) from the previous

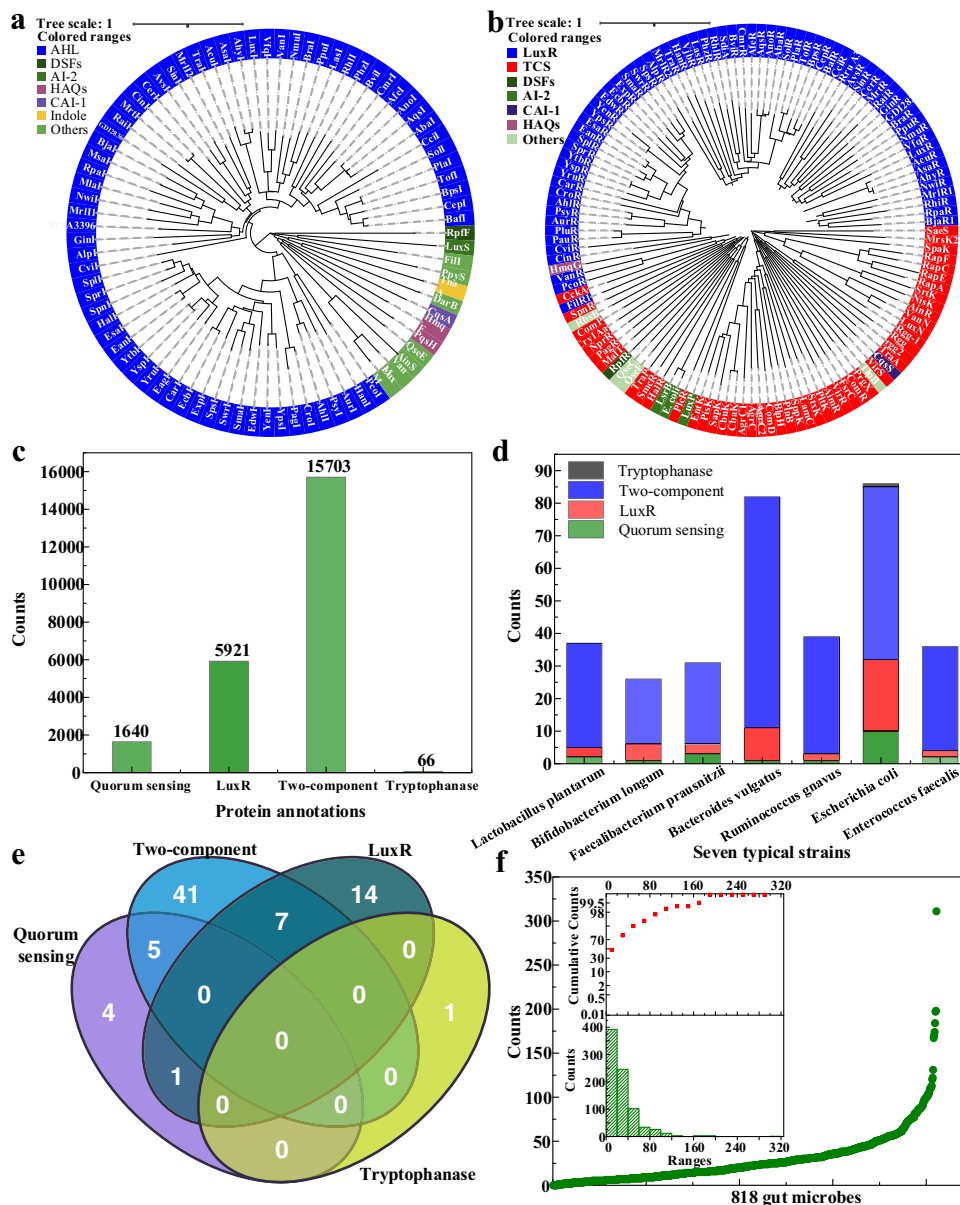


Fig. 2 Results of collections of the reported and annotated QS entries. Evolutionary trees of QS synthases (a) and receptors (b). **a** The optimal tree with the sum of branch length = 40.33 is shown. This analysis involves 84 amino acid sequences, and there are a total of 1374 positions in the final 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 1010 positions in the final dataset. **c** Total QS&TCS entries with four protein annotations, i.e., “quorum sensing”, “LuxR”, “two-component”, and “tryptophanase”. **d** QS&TCS entries distribution of the seven-strain simplified human gut microbes used by Colosimo et al.⁴² **e** The overlap of the four types of QS&TCS entries in *Escherichia coli* O157:H7 strain. **f** QS&TCS entries count distribution of 818 human gut microbes from the VMH database²⁸. Note that the subgraph indicates the cumulative distribution curve for the statistics of the collected QS&TCS entries. The upper and lower insets show the probabilities and histogram, respectively. “Ranges” in the subgraph shows the range of the number of QS&TCS entries contained in each strain. The “Counts” and “Cumulative counts” in the subgraph represents the specific number of strains and proportion, respectively. Source data are provided as a Source Data file.

annotation-based collections (Database III) (Fig. 2f). Note that we have mined eight potential QS proteins (Table 2) with the help of functional analysis and homologous modeling, which is of great significance for the further exploration of the related QS mechanism and their applications. To enable user-friendly browsing and searching for entries identified in this work, we constructed a comprehensive QS-related database of human gut microbiota (QSHGM), which is freely available at: <http://www.qshgm.lbcj.net/>. There is a simple user guide for QSHGM browsing and searching (Supplementary Fig 3) in the supplementary information.

QS-based interactions prediction. QS-based interactions play an essential role in deciphering complex interactions of natural microbial systems and dynamically manipulating diverse synthetic microbial consortia. The collected data in the QSHGM can enable the prediction of the existence of QS-based microbial interaction by querying whether any pairwise microbes can speak the same QS language. For example, due to speaking the AI-2 language, we predicted AI-2-based communication between *E. coli* O157:H7 and *Bacteroides pectinophilus* ATCC 43243 (Fig. 4a), which is in line with the previously reported observation that AI-2 produced by *E. coli* can influence the Bacteroidetes⁴⁹.

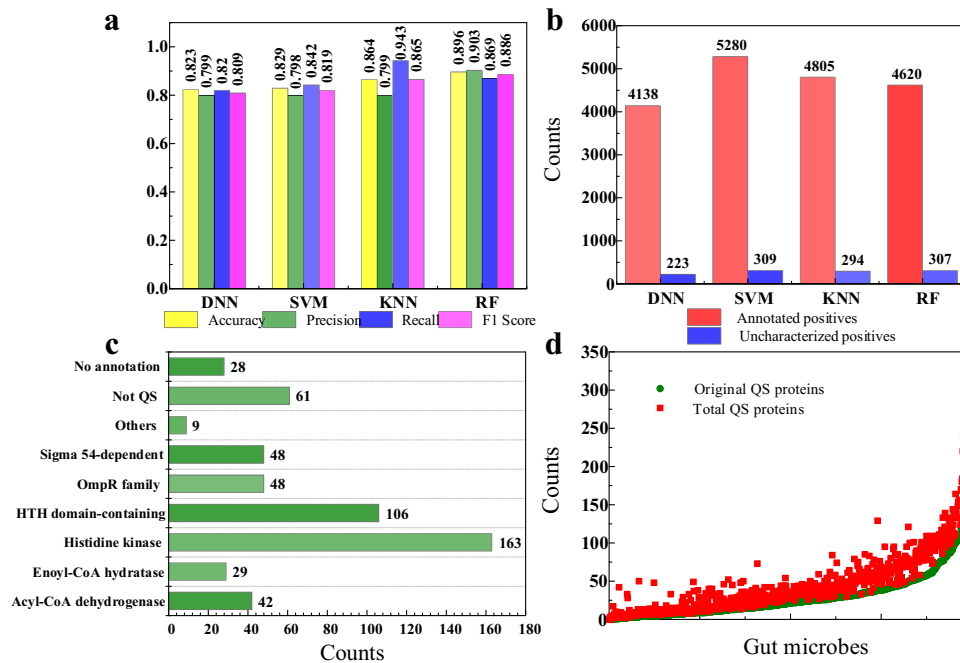


Table 2 Results of another eight expanded entries without existing annotations based on Phyre².

Strains	TaxID	Entry	Template	Confidence	Coverage	Annotations
<i>Bacillus amyloliquefaciens</i>	1390	A0A5C8IU59	c5xybB	100%	97%	AimR transcriptional regulator
<i>Bacillus mycoides</i>	1405	A0A1W6AJT8	c5zvva	100%	90%	AimR transcriptional regulator
<i>Bacillus thuringiensis</i>	56955	A0A243M9P9	c5zw5A	100%	95%	AimR transcriptional regulator
<i>Bacillus amyloliquefaciens</i>	1390	A0A5C8IY56	c5zvva	100%	99%	AimR transcriptional regulator
<i>Bacillus atrophaeus</i>	720555	A0A0H3E1W6	c5zvva	99.90%	98%	AimR transcriptional regulator
<i>Bacillus atrophaeus</i>	720555	A0A0H3E2G4	c5zw5A	100%	100%	AimR transcriptional regulator
<i>Lysinibacillus fusiformis</i>	28031	A0A4Y4IIV5	c6mfvc	100%	90%	Signaling protein (tetratricopeptide repeat)
<i>Streptococcus salivarius</i>	1304	A0A5C4P2T9	c4bxiA	99.90%	33%	ATP binding domain of AgrC

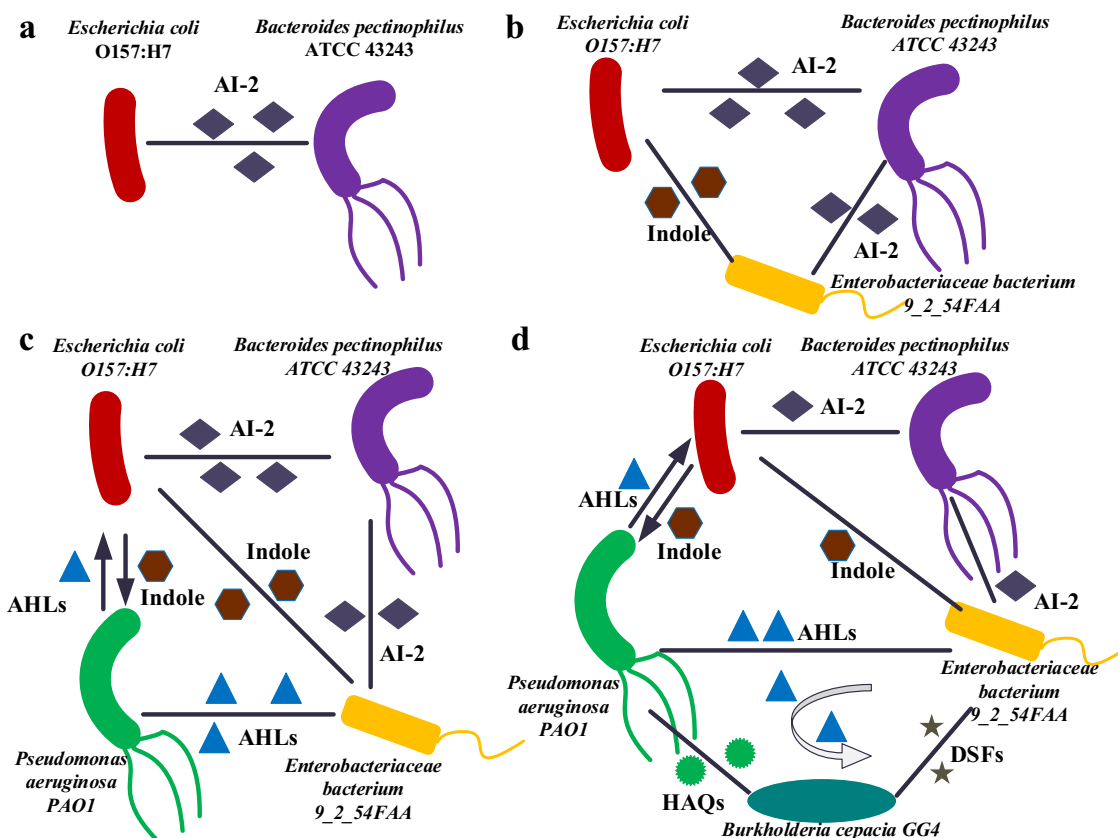


Fig. 4 QS-based communication predictions for various microbial consortium. a Two-strain communication based on AI-2; **b** three-strain communication based on AI-2 and indole; **c** four-strain communication based on AI-2, indole, and AHLs; **d** five-strain communication based on diverse QS languages.

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. Microbes communicate via various QS signals (also termed as microbial languages), 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” (QSCN). Based on a review of previous studies (Supplementary Table 2), we decided to focus on the common nine QS languages, i.e., AHLs, DSFs, HAQs, CAI-1, AIPs, Dialkylresorcinols, Photopyrones, indole, and AI-2 to construct the proposed QSCN. With the help of the QSHGM and several hypotheses (details given in Supplementary Table 3), we firstly constructed an undirected QSCN for the 818 gut microbes based on the “speaking” of the above nine QS languages (Fig. 5a) (Supplementary Data 13). This intricate network visualizes complex QS-based communications among human

gut microbiota. Different microbes are linked together through various languages to form a microbial communication network, and connections could be used to regulate microbial interactions between themselves and the surrounding ones. Most of strains produce 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 crosstalk⁵⁵. Taking six typical languages (AHLs, CAI-1, HAQs, DSFs, Indole, and AI-2) as examples, we found that there are 64, 40, 22, and 5 species sharing two, three, four, and five QS languages, respectively (Fig. 5b). 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 observed¹³. Many overlaps of AI-2 or indole being spoken among different microbes, which also indicates that both of them are widely recognizable languages playing a major

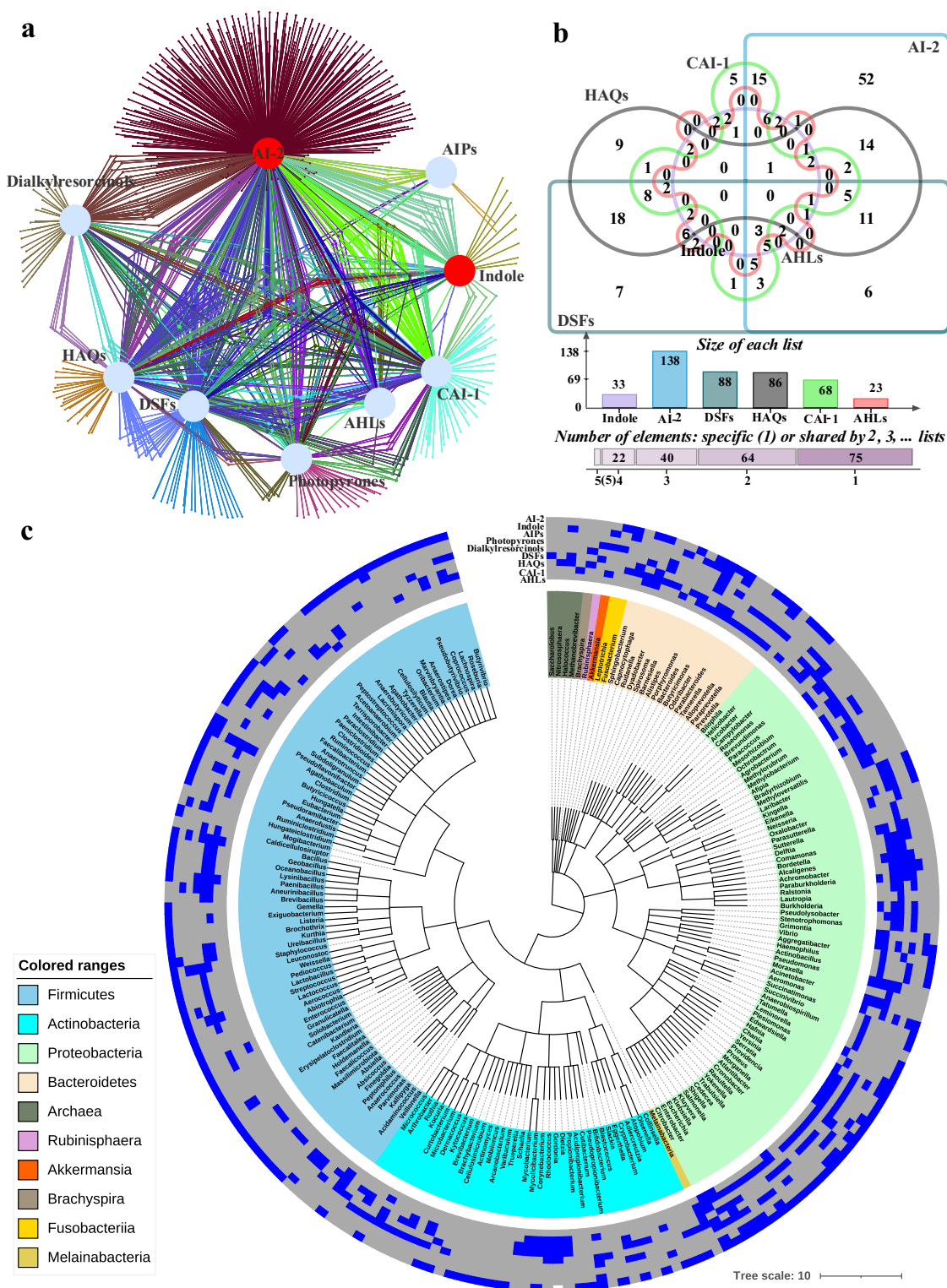


Fig. 5 QSCN for human gut microbiota based on diverse QS languages. **a** QSCN for 818 human gut microbes based on nine languages (AHLs, DSFs, HAQs, CAI-1, AIPs, Dialkylresorcinols, Photopyrones, indole, and AI-2). Generally recognized intraspecies and interspecies languages are marked in blue and red, respectively. Note that the network diagram was generated using EYenn⁶⁴. **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 ten genus-level clusters based on their phyla and taxonomy. Microbial genus from Firmicutes is colored in blue; Actinobacteria, cyan; Proteobacteria, green; Bacteroidetes, yellowish. Heatmap on the outermost layer indicates QS languages distribution in each cluster, existence is colored in blue; no existence, gray. Source data are provided as a Source Data file.

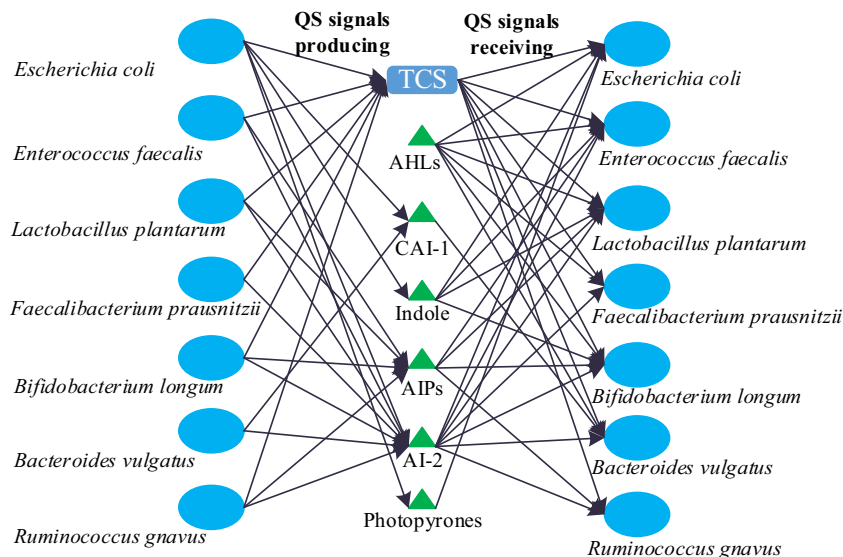


Fig. 6 Typical small QSCN that includes QS signals producing and receiving for seven human microbes. Note that the seven-strain simplified human microbiomes are taken from Colosimo et al.⁴².

role in interspecies communications^{56,57}. We found that those traditionally often considered intraspecies languages (AHLs, CAI-1, HAQs, and DSFs) may also be involved in some interspecies communications. Like Scott et al.⁵⁸, we also realized that the crosstalk of different QS languages implies the redundancy of microbial languages that is potentially helpful for the stability of natural microbial systems.

The QSCN 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. 5c to gain a better understanding of the QSCN (Fig. 5a). As in previous studies, we also found that AHLs exist only in Proteobacteria⁵⁹, AIPs exist mostly in Firmicutes¹², and other QS languages are distributed in-homogeneously in the whole genus-level microbes⁶⁰. Surprisingly, there is no highly similar distribution of QS languages within the same genus-level microbes. For example, the distribution of QS languages in Actinobacteria is quite different (Fig. 6c, cyan). This suggests that the existence and evolution of QS synthases in microbes might have not been strictly conserved at the genus level, but are more likely to be related to some other factors, such as environmental factors and spatial distributions^{61–63}. To sum up, the distribution of QS languages suggests the diversity of the microbial languages, the complexity of cell-cell communication, and the redundancy of QS-based interactions among human gut microbiota.

The QSCN we presented above (Fig. 5a) is an undirected and bipartite network involving two types of nodes, namely QS languages and microbes. This network can be projected to a one-mode network that visualizes microbial communication-based interactions directly. The giant network would consist of 801 nodes connected via 190,580 edges (Supplementary Fig. 4). The largest degree in the giant network is 771, while its average degree is 237.93. The dense QS network is similar to other microbial interaction networks that carry high degrees for individual strains^{65,66}. Key nodes in this network were selected from 5% of the total nodes (40 nodes, Supplementary Table 4) of the network with a large degree and high betweenness centrality⁶⁷. Note that all the 40 key nodes are Firmicutes, Bacteroidetes, or Proteobacteria, (Supplementary Table 4) which are known to be dominating species of the human gut microbiota^{68,69}. Therefore,

QSCN can be projected to a one-mode network and shrunk further to be the complete graph with 40 core gut microbes (Supplementary Fig. 5). 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. As such, what is visualized here is essentially a sub-network with a particularly high “density”, not representative of the entire network of the whole gut environment. We would also like to point out that, although each microbe can produce so many QS languages in the 40 core gut microbes, the specific intensity for each language cannot be provided in this work; the intention is that we determine the existence of the QS-based communications (as done in this work), and then to investigate its corresponding intensity (future work), eventually bringing a comprehensive understanding of the communication-based microbial interactions.

Discussion

This work has been based on several hypotheses on microbial composition, language types, TCS function, non-cheating ecology, and QS crosstalk, which may be addressed in the future to improve the accuracy and completeness of the QSCN (Supplementary Table 3). Besides, the large number of links in our QSHGM Database and the QSCN means that it is inevitable that there will be some false positive relations. Even if there is no problem at the level of individual nodes, the relations we have predicted were not necessarily always true in reality. Note that many TCS entries possess QS functionality (Supplementary Data 1, Supplementary Fig. 8), but not all of them would do so, which would apply to a portion of the TCS entries collected into our Dataset III that was built with the intention of collecting as many potentially QS-relevant entries as we can, let alone these entries would still be relevant to inter-cellular communication. To mine more potential QS entries, we combined manual curations, BLASTP-based expansion, and ML-based classifications together in this work along with minimizing false positives as possible. On the other hand, QS links we predicted based on the database would be “possibilities”, not reality, and still require experimental verification. We offered a tool to allow users with various applications in mind to see the “possibilities” in the first place, which allows them to subsequently focus their experimental verification.

Note that short peptides (such as AIPs) and proteins are not generally placed together for sequence BLASTP and functional analysis, because proteins generally have a fixed structure while short peptides do not. AIPs sequences can also easily lead to the increasing of false positives from the BLASTP process. The four ML-based classifiers were trained on AA frequencies, which were not accurate for the prediction of the short sequences such as the AIPs (about 5–30 amino acids), of which the physicochemical properties⁷⁰, the information on amino acids combinations with fixed length⁴³, and even the composition of common amino acids were not complete. Therefore, to increase the reliability of the expansion, we have removed the signal peptides in the BLASTP-related datasets (I and VII), thus leading to sparse edges for the “AIPs” node in our QSCN (Fig. 5a). This calls for a more accurate method to cover more aforementioned amino acid features for short sequences to mine the potential signal peptides in the future to make the QSCN more complete. Furthermore, the nine QS languages studied in this work (Supplementary Table 2) did not include all existing QS signals, such as Autoinducer-3 (AI-3, with unclear synthase sequence)⁷¹, let alone new QS languages that would be discovered in the future. Considering the QS crosstalk widely exists in nature⁵⁵, we also hypothesized that microbes that speak the same type of languages (such as AHLs) can communicate with each other. Future works should be conducted to quantify the specific intensity of diverse QS crosstalk for the same type of languages, such as the AHLs with different side chains. Therefore, more AIPs, some novel QS entries, and their corresponding weighted networks of different QS languages for more gut microbes will gradually be updated in our QSHGM and QSCN.

Bipartite (Fig. 5a) or one-mode QSCN (Supplementary Fig. 4) illustrates diverse language connections, which however lacks the further interactions between QS languages senders and receivers. By differentiating QS signals producing and receiving with the help of both QS synthases and receptors, there is potential to construct a directed and more precise QSCN. Taking the seven-strain simplified human microbiomes from Colosimo et al⁴² as an example, we constructed a typical small precise QSCN that includes QS languages producing and receiving (Fig. 6). QS language receiving (Fig. 6, right) is more complicated than language producing (Fig. 6, 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 behavior in certain microbes, such as *P. aeruginosa*⁷² and *E. coli*⁷³. However, the reliable construction of the directed and precise 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, and different spatial distributions, and insufficient QS entries for many uncultured microbes. Nevertheless, we expect that the further directed and precise QSCN 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 precise QSCN which can be regarded as one of the key knowledge maps of the human gut system.

Microbial communities and their functions are shaped by both metabolic interactions and communication-based regulations. Microbe–microbe interactions based on the exchange of metabolites have received much attention in microbial ecology^{65,74}. At the same time, various two-strain or three-strain synthetic consortia have been constructed by implementing QS for stabilizing the microbial ecosystem (more details in Supplementary Table 5). As we proposed earlier, QS-based communication networks (QSCNs) can be vertically and horizontally applied to the regulations in natural microbial systems and synthetic microbial

consortia, respectively²⁴, and they play different roles than metabolic integration networks (MINs) (more details in Supplementary Table 6). On the other hand, a QSCN and a MIN can be co-present and function collectively in a microbial ecosystem (Supplementary Fig. 6). One such example is from our earlier work⁷⁵, where we developed combinational QS devices for automatic dynamic control in a cross-feeding cocultivation of a synthetic community, to achieve the optimization of the system which simultaneously involved QS communication, cell growth competition, and cooperative production. More recently, a methodology was proposed for designing robust synthetic communities that include competition for nutrients, and use QS to control amensal bacteriocin interactions⁷⁶, which can be considered as a more generalized example of how the combination of QSCNs and MINs could lead to desirable designs of engineered microbial consortia.

To illustrate the potential of complementary use of the QSCN constructed in this work for the gut community and a MIN, here we consider the work of Venturelli et al⁶⁵ on a simplified human microbiota consortium (SIHUMI). By comparing the inferred total interaction network (Supplementary Fig. 7a) with the MIN (Supplementary Fig. 7b), it was recognized that the former was significantly denser than the latter with several prominent inter-microbial links not associated with the exchange of metabolites, which were considered to be possibly mediated by signaling molecules instead⁶⁵. Applying our QSHGM database to the SIHUMI community, we have obtained a bipartite QSCN (Supplementary Fig. 7c), which shows specific QS-based communications that offer plausible mechanisms for links that the MIN could not explain (Supplementary information, Section 5). Thus, we consider our QSHGM database as a tool that can facilitate the identification of possible QS-based inter-microbial interactions which may complement metabolic exchanges in a complex community in explaining an observed community structure; such possible interactions can be tested based on the microbe-QS signal pairs suggested by the database through e.g. detecting and manipulating the excretion/reception of the specific QS signaling molecules involved.

Various QS-based interactions play an essential role in the regulation of homeostatic states, metabolism, and immune responses in 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 collecting, expanding, and 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 modeling to facilitate the efficiency of data collection and mining. As a result, we established the QSHGM (<http://www.qshgm.lbc.net/>, with browsing and searching functions) which contains 28,567 redundancy removal entries for 818 human gut microbes.

With the help of the QSHGM, users can search many QS-based interactions for various microbial consortia based on diverse QS languages. We constructed a QSCN to visualize and decipher intricate QS-based interactions for human gut microbiota. We found that the distribution of QS languages in microbes is not strictly conserved 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 suggest 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 microbial language and

the redundancy of cell-cell communications, which are helpful for maintaining the stability of natural microbial systems. The QSCN can be projected to a one-mode network; a fraction of which is a sub-network representing potentially very “dense” communications of 40 core gut microbes. This work contributes to the construction of the QSCN 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 gut microbiology and for developing applications such as potential therapies. For the QSCN to be more effective and more widely applicable, further work is needed to identify the strengths of diverse QS-based interactions and combine it with other types of connections, particularly those captured by microbial interaction networks, to achieve reliable, quantitative predictions for microbial ecosystems.

QSHGM and QSCN can not only give us a better understanding of QS-based microbial communication principles but also will do much help in providing new manipulations to synthetic microbiota and developing potential therapies (Supplementary Fig 9). 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. At the strain level, QSHGM and QSCN will provide user-friendly data searching, and assist the scientific community in various interferences and manipulations of QS systems to alleviate antimicrobial resistance, inhibit pathogenic bacteria, and develop new QS-based synthetic gene circuits for various applications. At the community level, the communication-based regulations can be visualized for human gut microbiota, users can search many QS-based communications for various microbial consortia including multi-component strains based on diverse QS languages. The predicted communications will provide guidance for consortia-based therapies or constructing new synthetic microbial consortia. Furthermore, QSHGM furnishes high-throughput data for large-scale QS-relevant statistical analysis.

Methods

Data acquisition. QS is a common mechanism which includes autoinducer synthase and relevant QS receptors⁷⁷. 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 genes⁵⁶. 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/>)³⁹. Note that QS entries in Sigmol and QuorumPeps database without corresponding amino acid sequences were discarded in this study. We have also made some choices about the entries from Sigmol and QuorumPeps database to improve the efficiency of Local BLAST. For example, we kept only one related entry for the same QS entries, such as the *S*-ribosylhomocysteine lyase (LuxS), Acyl-homoserine-lactone synthase LuxM, and accessory gene regulator C (AgrC). The autoinducer peptides (AIPs), such as Nisin precursor peptide (NisA) and competence stimulating peptide AgrD, were not considered in the BLASTP-related work. Because the signal peptide sequence of Gram-positive bacteria is relatively short (about 10–30 amino acids), which will be easy to increase the false positives for the BLASTP. Therefore, to increase the reliability of the ML-based classifiers, we have removed the signal peptides in the BLASTP-related datasets (I and VII) by more exact matches (with lower *E*-value) and manual checking of their protein annotations or the open reading frames (ORF) in the NCBI database. In addition, 818 gut microbes from a virtual metabolic human database (www.vmh.life)²⁸ 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 sequence⁷⁸. In this study, the information of amino acids in protein sequences was calculated with the help of the iFeature package, ML algorithms (SVM⁷⁹, RF⁸⁰, and KNN⁸¹), and deep learning

algorithms (DNN⁸²) 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 types divided by the length of a protein sequence⁴³, which is listed as follows:

$$f(t) = N(t)/N, t \in \{A, C, D, \dots, Y\} \quad (1)$$

where $N(t)$ is the number of amino acid type t , while N is the length of a protein or peptide sequence.

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 positive samples. The finally obtained positive samples (Dataset III) were the arrays of 21,383 amino acid frequencies of the collected QS entries. In the collecting module, we did an evolutionary analysis for the validated QS entries to propose a possible cluster rule for negative samples collection with the help of MEGA X⁸³ and iTOL⁸⁴. The evolutionary history was inferred using the Neighbor-Joining method⁸⁵. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree⁸⁶. The evolutionary distances were computed using the Poisson correction method and are in the units of the number of amino acid substitutions per site⁸³. 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. Subsequently, we got an output array of 22,780 (negative Dataset IV) amino acid frequencies, which were calculated from amino acids sequences of proteins of the above eight organisms after removing QS-related entries.

ML-based classifiers. The amino acid composition (AAC) calculates the frequency of each amino acid type in a protein or peptide sequence. We calculated AAC in each entry sequence as the protein features. “model.py” was created for training samples with SVM, KNN, and RF (random forest). “nn.py” was the script used for training samples with Neural Network. Classifiers were trained and validated based on the positive and negative samples, and then tested on dataset VII (Fig. 2). Performances of the four ML-based classifiers were measured based on the accuracy, precision, recall, and *F1* score, which are defined as follows⁸⁷.

$$\text{Accuracy} = (TN + TP)/(TN + FP + FN + TP) \quad (2)$$

$$\text{Precision} = TP/(TP + FP) \quad (3)$$

$$\text{Recall} = TP/(TP + FN) \quad (4)$$

$$F1 = 2 * \text{Precision} * \text{Recall}/(\text{Precision} + \text{Recall}) \quad (5)$$

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.

All the four classifiers were applied to predict whether the input amino acid sequences are QS entries or not with the output being 1 (yes) or 0 (no), respectively. SVM is a commonly used supervised ML algorithm in protein prediction⁸⁷. The basic idea of SVM is to find the separated hyperplane in a very high-dimension feature space that can correctly partition the training dataset⁷⁹. SVM can also integrate kernel functions, which makes 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. The GridSearchCv code⁸⁸ was used to select and determine the optimal combination of hyper-parameters automatically to achieve the best performance.

K-nearest-neighbor (KNN) is also a traditional classification method when there is little or no prior knowledge about the distribution of the data⁸⁹. 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, and 300) and the KNN with $k = 5$ yielded the best result.

Random forest (RF) is a classification algorithm that uses a set of decision trees⁸⁰. 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 tasks⁸². 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 $max_depth = 55$ (maximum depth of the tree). Other hyper-parameters were also generated and selected with the help of GridSearchCV⁸⁸.

Neural networks (NN) play an essential role in biomedicine⁸², antiviral peptides prediction, protein–RNA interaction⁹⁰, and protein data mining. 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 the one-to-one layer and the 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 a one-to-one layer and each hidden layer to accelerate the training process. SGD optimizer was used to train the DNN model and the learning rate was fixed as 0.01. Default values of the other hyper-parameters of the DNN model were set to default ones without tuning.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

About 28,567 redundancy removal entries for 818 gut microbes generated in this study have been deposited in our QSHGM database, which is freely available at: (<http://www.qshgm.lbcj.net/>). We will continuously update the database QSHGM. Computer-readable tables generated in this study are provided in Supplementary Data 1–13. More details for the relevant data of QS entries from Gram-negative microbes, QS entries from Gram-positive microbes, 818 human gut microbes, and their corresponding proteomes can be searched in SigMol²⁹, Quorumpeps³⁰, VMH²⁸, and UniProt³⁹ databases, respectively.

Code availability

We also used Python 3.7 to write the method and analyse the collected data to construct ensemble classifiers. The codes have been provided in a GitHub repository at: <https://github.com/guofei-tju/qshgm-code>⁹¹.

Received: 27 July 2021; Accepted: 17 May 2022;

Published online: 02 June 2022

References

- Donaldson, G. P., Lee, S. M. & Mazmanian, S. K. Gut biogeography of the bacterial microbiota. *Nat. Rev. Microbiol.* **14**, 20–32 (2016).
- Baumler, A. J. & Sperandio, V. Interactions between the microbiota and pathogenic bacteria in the gut. *Nature* **535**, 85–93 (2016).
- Schluter, J. et al. The gut microbiota is associated with immune cell dynamics in humans. *Nature* **588**, 303–307 (2020).
- Neurath, M. F. Host-microbiota interactions in inflammatory bowel disease. *Nat. Rev. Gastroenterol. Hepatol.* **17**, 76–77 (2020).
- Almeida, A. et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* **39**, 105–114 (2021).
- Sung, J. et al. Global metabolic interaction network of the human gut microbiota for context-specific community-scale analysis. *Nat. Commun.* **8**, 15393 (2017).
- 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).
- Fan, Y. & Pedersen, O. Gut microbiota in human metabolic health and disease. *Nat. Rev. Microbiol.* **19**, 55–71 (2021).
- Defoirdt, T. Quorum-sensing systems as targets for antivirulence therapy. *Trends Microbiol.* **26**, 313–328 (2017).
- 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).
- Papenfert, K. & Bassler, B. L. Quorum sensing signal-response systems in gram-negative bacteria. *Nat. Rev. Microbiol.* **14**, 576–588 (2016).
- Monnet, V. & Gardan, R. Quorum-sensing regulators in gram-positive bacteria: ‘Cherchez le peptide’. *Mol. Microbiol.* **97**, 181–184 (2015).
- Pereira, C. S., Thompson, J. A. & Xavier, K. B. AI-2-mediated signalling in bacteria. *FEMS Microbiol. Rev.* **37**, 156–181 (2013).
- Zarkan, A., Liu, J., Matuszewska, M., Gaimster, H. & Summers, D. K. Local and universal action: The paradoxes of indole signalling in bacteria. *Trends Microbiol.* **28**, 566–577 (2020).
- Stephens, K. & Bentley, W. E. Synthetic biology for manipulating quorum sensing in microbial consortia. *Trends Microbiol.* **28**, 633–643 (2020).
- Tateda, K. et al. The *Pseudomonas aeruginosa* autoinducer N-3-oxododecanoyl homoserine lactone accelerates apoptosis in macrophages and neutrophils. *Infect. Immun.* **71**, 5785–5793 (2003).
- Kravchenko, V. V. et al. Modulation of gene expression via disruption of NF- κ B signaling by a bacterial small molecule. *Science* **321**, 259–263 (2008).
- An, S. Q. et al. Modulation of antibiotic sensitivity and biofilm formation in *Pseudomonas aeruginosa* by interspecies signal analogues. *Nat. Commun.* **10**, 2334 (2019).
- Moura-Alves, P. et al. Host monitoring of quorum sensing during *Pseudomonas aeruginosa* infection. *Science* **366**, eaaw1629 (2019).
- Mao, N., Cubillos-Ruiz, A., Cameron, D. E. & Collins, J. J. Probiotic strains detect and suppress Cholera in mice. *Sci. Transl. Med.* **10**, eaao2586 (2018).
- Hsiao, A. et al. Members of the human gut microbiota involved in recovery from *Vibrio cholerae* infection. *Nature* **515**, 423–426 (2014).
- Lee, J. H., Wood, T. K. & Lee, J. Roles of indole as an interspecies and interkingdom signaling molecule. *Trends Microbiol.* **23**, 707–718 (2015).
- Sedlmayer, F., Hell, D., Muller, M., Auslander, D. & Fussenegger, M. Designer cells programming quorum-sensing interference with microbes. *Nat. Commun.* **9**, 1822–1835 (2018).
- Wu, S., Xu, C., Liu, J., Liu, C. & Qiao, J. Vertical and horizontal quorum-sensing-based multicellular communications. *Trends Microbiol.* **29**, 1130–1142 (2021).
- Zhang Q., et al. gutMEGA: A database of the human gut metagenome atlas. *Brief Bioinform.* **22**, (2021).
- Wu, S. et al. GMrepo: A database of curated and consistently annotated human gut metagenomes. *Nucleic Acids Res.* **48**, D545–D553 (2020).
- 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).
- 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).
- Rajput, A., Kaur, K. & Kumar, M. SigMol: Repertoire of quorum sensing signaling molecules in prokaryotes. *Nucleic Acids Res.* **44**, 634–639 (2016).
- Wynendaale, E. et al. Quorumpeps database: chemical space, microbial origin and functionality of quorum sensing peptides. *Nucleic Acids Res.* **41**, D655–D659 (2013).
- Barakat, M., Ortet, P. & Whitworth, D. E. P2CS: A database of prokaryotic two-component systems. *Nucleic Acids Res.* **39**, D771–D776 (2011).
- Ortet, P., Whitworth, D. E., Santaella, C., Achouak, W. & Barakat, M. P2CS: updates of the prokaryotic two-component systems database. *Nucleic Acids Res.* **43**, D536–D541 (2015).
- Wu, S. et al. QSIdb: quorum sensing interference molecules. *Brief. Bioinform.* **22**, bbaa218 (2021).
- 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).
- Zeng, X. et al. MASI: microbiota-active substance interactions database. *Nucleic Acids Res.* **49**, D776–D782 (2021).
- Tang, J. et al. GIMICA: host genetic and immune factors shaping human microbiota. *Nucleic Acids Res.* **49**, D715–d722 (2021).
- Ye, J., McGinnis, S. & Madden, T. L. BLAST: improvements for better sequence analysis. *Nucleic Acids Res.* **34**, W6–W9 (2006).
- Kerfeld, C. A. & Scott, K. M. Using BLAST to teach “e-value-tionary” concepts. *PLoS Biol.* **9**, e1001014 (2011).
- Bairoch, A. et al. The universal protein resource (uniprot). *Nucleic Acids Res.* **33**, D154–D159 (2005).
- Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. E. The Phyre² web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10**, 845–858 (2015).
- Gray, K. M. & Garey, J. R. The evolution of bacterial LuxI and LuxR quorum sensing regulators. *Microbiology (Read.)* **147**, 2379–2387 (2001).
- Colosimo, D. A. et al. Mapping interactions of microbial metabolites with human G-protein-coupled receptors. *Cell. Host. Microbe* **26**, 273–282 e277 (2019).
- 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).
- Servetas, S. L. et al. Characterization of key *Helicobacter pylori* regulators identifies a role for arsrs in biofilm formation. *J. Bacteriol.* **198**, 2536–2548 (2016).
- Brameyer, S., Kresovic, D., Bode, H. B. & Heermann, R. Dialkylresorcinols as bacterial signaling molecules. *Proc. Natl Acad. Sci. USA* **112**, 572–577 (2015).

46. Parashar, V., Aggarwal, C., Federle, M. J. & Neiditch, M. B. Rgg protein structure-function and inhibition by cyclic peptide compounds. *Proc. Natl Acad. Sci. USA* **112**, 5177–5182 (2015).
47. Yang, T., Talgan, Y., Paharik, A. E., Horswill, A. R. & Blackwell, H. E. Structure-function analyses of a *Staphylococcus epidermidis* autoinducing peptide reveals motifs critical for AgrC-type receptor modulation. *ACS Chem. Biol.* **11**, 1982 (2016).
48. Erez, Z. et al. Communication between viruses guides lysis-lysogeny decisions. *Nature* **541**, 488–493 (2017).
49. Bivar Xavier, K. Bacterial interspecies quorum sensing in the mammalian gut microbiota. *C. R. Biol.* **341**, 297–299 (2018).
50. Wang, D., Ding, X. & Rather, P. N. Indole can act as an extracellular signal in *Escherichia coli*. *J. Bacteriol.* **183**, 4210–4216 (2001).
51. Jaglin, M. et al. Indole, a signaling molecule produced by the gut microbiota, negatively impacts emotional behaviors in rats. *Front. Neurosci.* **12**, 216 (2018).
52. Nguyen, Y. et al. Structural and mechanistic roles of novel chemical ligands on the SdiA quorum-sensing transcription regulator. *mBio* **6**, e02429–02414 (2015).
53. 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).
54. 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).
55. Wellington, S. & Greenberg, E. P. Quorum sensing signal selectivity and the potential for interspecies cross talk. *mBio* **10**, e00146–00119 (2019).
56. Wang, S., Payne, G. F. & Bentley, W. E. Quorum sensing communication: Molecularly connecting cells, their neighbors, and even devices. *Annu. Rev. Chem. Biomol. Eng.* **11**, 447–468 (2020).
57. Kumar, P., Lee, J.-H. & Lee, J. Diverse roles of microbial indole compounds in eukaryotic systems. *Biol. Rev. Camb. Philos. Soc.* **96**, 2522–2545 (2021).
58. Scott, S. R. et al. A stabilized microbial ecosystem of self-limiting bacteria using synthetic quorum-regulated lysis. *Nat. Microbiol.* **2**, 17083 (2017).
59. Case, R. J., Labbate, M. & Kjelleberg, S. AHL-driven quorum-sensing circuits: their frequency and function among the proteobacteria. *ISME J.* **2**, 345–349 (2008).
60. Whiteley, M., Diggle, S. P. & Greenberg, E. P. Progress in and promise of bacterial quorum sensing research. *Nature* **551**, 313–320 (2017).
61. Vrancken, G., Gregory, A. C., Huys, G. R. B., Faust, K. & Raes, J. Synthetic ecology of the human gut microbiota. *Nat. Rev. Microbiol.* **17**, 754–763 (2019).
62. Consortium THMP. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
63. Faust, K. et al. Microbial co-occurrence relationships in the human microbiome. *PLoS Comput. Biol.* **8**, e1002606 (2012).
64. Chen, T., Zhang, H., Liu, Y., Liu, Y.-X. & Huang, L. EVenn: easy to create repeatable and editable Venn diagrams and venn networks online. *J. Genet. Genomics* **48**, 863–866 (2021).
65. Venturelli, O. S. et al. Deciphering microbial interactions in synthetic human gut microbiome communities. *Mol. Syst. Biol.* **14**, e8157 (2018).
66. Faust, K. & Raes, J. Microbial interactions: from networks to models. *Nat. Rev. Microbiol.* **10**, 538–550 (2012).
67. Ran, J. et al. Construction and analysis of the protein-protein interaction network related to essential hypertension. *BMC Syst. Biol.* **7**, 32 (2013).
68. Eckburg, P. B. et al. Diversity of the human intestinal microbial flora. *Science* **308**, 1635 (2005).
69. Zou, Y. et al. 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. *Nat. Biotechnol.* **37**, 179–185 (2019).
70. Lee, T. Y., Lin, Z. Q., Hsieh, S. J., Bretana, N. A. & Lu, C. T. Exploiting maximal dependence decomposition to identify conserved motifs from a group of aligned signal sequences. *Bioinformatics* **27**, 1780–1787 (2011).
71. Sperandio, V., Torres, A. G., Jarvis, B., Nataro, J. P. & Kaper, J. B. Bacteria-host communication: The language of hormones. *Proc. Natl Acad. Sci. USA* **100**, 8951–8956 (2003).
72. Sandoz, K. M., Mitzimberg, S. M. & Schuster, M. Social cheating in *Pseudomonas aeruginosa* quorum sensing. *Proc. Natl Acad. Sci. USA* **104**, 15876–15881 (2007).
73. Yao, Y. et al. 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).
74. Heirendt, L. et al. Creation and analysis of biochemical constraint-based models using the COBRA toolbox v.3.0. *Nat. Protoc.* **14**, 639–702 (2019).
75. Wu, S. et al. Combinational quorum sensing devices for dynamic control in cross-feeding cocultivation. *Metab. Eng.* **67**, 186–197 (2021).
76. Karkaria, B. D., Fedorec, A. J. H. & Barnes, C. P. Automated design of synthetic microbial communities. *Nat. Commun.* **12**, 672 (2021).
77. Hawver, L. A., Jung, S. A. & Ng, W. L. Specificity and complexity in bacterial quorum-sensing systems. *FEMS Microbiol. Rev.* **40**, 738–752 (2016).
78. Zhao, B. et al. Describeprot: database of amino acid-level protein structure and function predictions. *Nucleic Acids Res.* **49**, D298–D308 (2021).
79. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).
80. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
81. Peterson, L. E. K-nearest neighbor. *J. Scholarpedia* **4**, 1883 (2009).
82. Wainberg, M., Merico, D., Delong, A. & Frey, B. J. Deep learning in biomedicine. *Nat. Biotechnol.* **36**, 829–838 (2018).
83. Kumar, S. et al. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018).
84. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* **47**, W256–W259 (2019).
85. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425 (1987).
86. Felsenstein, J. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**, 783–791 (1985).
87. 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).
88. Pedregosa, F. et al. Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
89. Royce, T. E., Rozowsky, J. S. & Gerstein, M. B. Toward a universal microarray: prediction of gene expression through nearest-neighbor probe sequence identification. *Nucleic Acids Res.* **35**, e99 (2007).
90. Lam, J. H. et al. A deep learning framework to predict binding preference of RNA constituents on protein surface. *Nat. Commun.* **10**, 4941 (2019).
91. Wu, S. et al. Machine learning aided construction of the quorum sensing communication network for human gut microbiota, guofei-tju/qshgm-code: qshgm. zenodo. 6534482 <https://doi.org/10.5281/zenodo.6534482> (2022).

Acknowledgements

This study was supported by the National Key Research and Development Project of China (No.2019YFA0905600, J.Q.), the National Natural Science Foundation of China (62172296, F.G.), the Funds for Creative Research Groups of China (21621004, J.Q.), and National Key Research and Development Program of China (No. 2020YFA0907900, J.Q.).

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., and 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.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-022-30741-6>.

Correspondence and requests for materials should be addressed to Aidong Yang, Fei Guo or Jianjun Qiao.

Peer review information *Nature Communications* thanks Liang Tian and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022