Phylogenetic analysis of quorum sensing systems in bacteria

Marina Giannakara, Vassiliki Lila Koumandou

Genetics Laboratory, Department of Biotechnology, Agricultural University of Athens, Athens, Greece

Corresponding author:
Vassiliki Lila Koumandou
Iera Odos 75, Athens 11855, Greece
Email address: koumandou@aua.gr

Abstract

Quorum sensing (QS) is a cell-to-cell communication system that enables bacteria to coordinate their gene expression depending on their population density, via the detection of small molecules called autoinducers. In this way bacteria can act collectively to initiate processes like bioluminescence, virulence and biofilm formation. Autoinducers are detected by receptors, some of which are part of Two Component Signal Transduction Systems (TCS), which comprise of a sensor histidine kinase (usually membrane-bound) and a cognate response regulator. Different QS systems are used by different bacterial taxa, and their relative evolutionary relationships have not been extensively studied. To address this, we used the KEGG database to identify all the QS receptors and response regulators that are part of TCS and collected their amino acid sequences from different species. In order to discover their evolutionary relationships: (i) we compared the combinations of the highly conserved domains in the different receptors and response regulators using the SMART and KEGG databases, and (ii) we constructed and compared phylogenetic trees, based on neighbor-joining and maximum likelihood methods. For both the QS receptors and the response regulators, our analysis indicates certain close evolutionary relationships, highlight a common evolutionary history, and which can inform future applications, such as the design of novel inhibitors for pathogenic QS systems.

Introduction

The mechanism and importance of quorum sensing

Quorum sensing is a mechanism responsible for a variety of group behaviors and interactions in bacteria. It regulates numerous mechanisms, including pathogenicity, biofilm formation, mobility, sporulation, conjugal plasmid transfer, bioluminescence, resistance to antibiotics, and antibiotics production (Bandara et al., 2012; Ng and Bassler, 2009). Despite the differences between various regulatory components and their molecular mechanisms, all quorum sensing systems regulate their target genes in a population-density dependent manner. An extracellular molecule (autoinducer, AI) is secreted by a group of bacteria and when the number of the bacteria rises, so does the number of these molecules produced in the surrounding area. When the density of the AIs overcomes a specific threshold, they bind to a receptor protein, which can be found either on the membrane or in the cytoplasm of the cell. The detection of the AIs activates a signal transduction cascade, altering the gene expression of the cells all over the population (Ng and Bassler, 2009).

The quorum sensing mechanism is different between Gram-positive and Gram-negative bacteria. Gram-positive bacteria use modified oligopeptides as AIs (autoinducing peptides, AIPs), which bind to membrane-bound receptors- Histidine Kinases (HKs) when they exceed a density threshold. This leads to the phosphorylation of the HK protein and its cognate receptor, which then regulates the target genes. The AIPs can also be transported into the cell and interact with cytoplasmic AIP receptors, which act as transcription regulators to regulate gene expression (Ng and Bassler, 2009; Rutherford and Bassler, 2012). Gram-negative bacteria use AHLs (acyl-homoserine-lactones) or SAM-products (S-adenosylmethionine)(Papenfort and Bassler, 2016). A second group of AIs are DPD derived molecules (4,5-dihydroxy-2,3-pentanedione) and are collectively known as AI-2 (Bandara et al., 2012). Those small molecules can diffuse through the cell membrane and bind to LuxR-type cytoplasmic receptors, which function as transcription factors. However, in some Gram negative bacteria the AIs are detected by Histidine Kinases and the process is similar to that in the Gram positive bacteria (Rutherford and Bassler, 2012). Further signal molecules taking part in quorum sensing include protein molecules that are not described as AIPs, pheromones, sulfate and phosphate, fucose, oxygen, quinolones, γ-aminobutiric acid ("KEGG PATHWAY: map02024," 2019) or even adrenaline and noradrenaline produced by the gut bacteria in humans, known as AI-3 (Reading and Sperandio, 2006).

Since quorum sensing affects many bacterial processes, it is an interesting subject of study which can have numerous applications, most important of which is the control of pathogenicity via QS inhibition, given the increasing spread of antibiotic-resistant bacteria. It has been suggested that targeting QS imposes less selective pressure, since it does not kill bacteria but only hinders the production of virulence factors (Whiteley et al., 2017). QS inhibition (Quorum Quenching, QQ) can be applied to fields such as hospital infections, phytopathogen control in agriculture and producing new preservatives in the food industry (Sreenivasulu, 2015; Tiwari et al., 2016). The gene products involved in QS mechanisms are

consequently possible targets in new antimicrobial strategies and the investigation of their evolutionary relationships can contribute to that direction (Lerat and Moran, 2004).

Two-Component System types

The Histidine Kinases that contribute to QS in both Gram positive and negative bacteria are part of the Two Component Signal Transduction (or Component) Systems (TCSs). TCS is the main mechanism in bacteria for responding to environmental stimuli and prevails across the entire bacterial kingdom (Capra and Laub, 2012). Although there is a wide variety in TCS mechanisms, they all fall into two main categories: The canonical and the multi-step TCS.

Canonical TCS: His-Asp

A typical TCS comprises two proteins: a transmembrane Histidine Kinase (HK) and a Response Regulator (RR). The HK is a dimeric protein and it includes 3 domains: The Sensor, the HisKA (Histidine Kinase domain or Dimerization domain) and the HATPase (ATP kinase binding domain). The sensor domain binds to the AI and this interaction results in an ATP-dependent autophosphorylation of the conserved histidine of HisKA. The phosphoryl group is then transferred to a protein called Response Regulator (RR). The RR contains the Response Regulator domain (or REC or Receiver domain) in the N-terminal region and the Effector domain in the C-terminus (Jacob et al., 2014; Kaserer and West, 2010). The first domain includes the conserved aspartate residue, which is phosphorylated (Kaserer and West, 2010). This leads to conformational changes in the structure of the RR (West and Stock, 2001), which is responsible for controlling quorum sensing-associated behaviors through the Effector domain either via protein-protein interactions or by binding to DNA for gene regulation (Mascher et al., 2006).

Multi-step TCS: His-Asp-Asp

A more complicated TCS mechanism usually includes a hybrid HK (HHK), a histidine phosphotransferase (HPt) and its cognate RR. The structure of the HHK is similar to HK but additionally includes a conserved Asp residue in the C-terminus, namely a REC domain. The binding of the signaling molecule induces the autophosorylation of the His of the HisKA domain and then the phosphoryl-group is transferred intramolecularly to the Asp residue of the HHK. Then the phosphotransfer continues to the HPt protein and from there to the RR (Capra and Laub, 2012; Kaserer and West, 2010).

Categories of HKs and RRs

The classification of the HKs and the RRs into groups varies and depends on the comparison criteria. In the KEGG database the HKs and their cognate RRs are divided into separate families (Supplementary Fig. S1), however the criterion for this classification is not obvious. Given the structure of the proteins and more specifically their domain organization,

the HKs can be divided into 4 groups: 3 groups with proteins containing the intracellular domains PAS, HAMP or GAF, respectively, and a group with none of these domains. The RRs are divided into 5 groups according to their Effector domain (Capra and Laub, 2012).

Another way of classification indicates that the family of HKs consists of 11 subfamilies, a conclusion that resulted after a multiple alignment of 348 HKs and comparison of the "Homology boxes". These are groups of highly conserved amino acids, which are assumed to be important in substrate binding, catalysis or the structure of the protein: the H- and X-boxes are parts of the HisKA domain, and the N-,D- F-and G-boxes are part of the HATPase_c domain. The D- box is also part of the REC domain. In the same way there were distinct groups for the RRs, by comparing their Receiver domains (Grebe and Stock, 1999).

In a study of Frost and Kim, the proteins were separated into 5 types (Types I-IV and the chemosensor HKs group, CheA) based on the inferred phylogenetic tree using the whole genome of 22 bacterial species and 4 archaea. More specifically, the criterion for their division was the H-box (HisKA) and its secondary structure, and the kinase domain (ATPase) (Forst and Kim, 2015). Grebe and Stock classified the RRs into 8 subfamilies (R_A-R_H) by comparing the Receiver domains (REC) (Grebe and Stock, 1999). A previous classification was conducted by Hakenbeck and Stock based on the DNA-binding domains of RRs (Effector domain) (Hakenbeck and Stock, 1996). Others have proposed that the comparison of the RRs for their classification should be based only on the Effector domain, due to the fact that the Receiver domains of different RRs are highly conserved and therefore not very informative for classification (Galperin, 2006).

Evolution of TCS

It is suggested that the more variable the environment, the higher the number of the TCS of the organisms living in it. The genes of the HKs and their cognate RRs are usually found on the same operon. The various TCS pathways are the result of lateral gene transfer and/or duplication. The duplication can happen to all of the genes of the operon or to one of them; in the second case the resulting TCS pathway consists of more than one HK and responds to more than one extracellular signal, or it includes more than one RR and gives multiple responses for a specific signal. Also, due to the modular nature of HKs the domain shuffling can lead to new HKs and the fusion of the HK and RR genes of an operon can create hybrid HKs. The function of TCS pathways relies on molecular recognition and consequently the amino acids have extensively co-evolved as a means to prevent the disruption of signaling by mutations. Co-evolution also prevents crosstalk with other pathways. Concerning the origin of HKs, it has been suggested that they come from ATPases of the GHKL superfamily: Hsp90, the mismatch repair protein MutL or type II topoisomerases (Capra and Laub, 2012). On the other hand, the origin of the RRs remains unclear; the evolution of RRs can be the result of changes in the DNA-binding or RNA-polymerase interaction sites, it includes lateral gene transfer and gene duplication, while domain shuffling and rearrangement have also led to new forms of RRs (Capra and Laub, 2012).

Evolution of QS

Most of the phylogenetic studies of QS mainly focus on the produced AIs or on specific pathways, including LuxI/LuxR and LuxS/LuxQ. Lerat and Moran conducted a phylogenetic analysis for the proteins LuxI, LuxR and LuxS, which shows that these proteins are ancient in many bacterial species and that they appeared very early in the evolutionary path of the bacteria. It was also found that in most cases the gene pairs of the inducer and its cognate receptor are located next to each other on the chromosomes and they preserve their pairwise function, a sign of common evolutionary history. The existence of homologous proteins in some genomes is representative of horizontal transfer and duplication of the genes, both of which can alter the regulation of different gene targets (Lerat and Moran, 2004). In a research of the ComQXPA pathway of Dogsa et. al, the phylogenetic tree of HKs proteins in 60 genomes containing comQXPA locus clusters all the ComP proteins from various organisms together, instead of gathering them with HKs of the same organism. Therefore, it is concluded that ComP evolved from the other HKs before the appearance of the modern bacterial species and thus the ComQXPA pathways have an ancient origin (Dogsa et al., 2014).

In a study of the Vibrionaceae family, it was found that the various inducers (AHLs) do not show any correlation with the species' geographical distribution. This result indicates that AHLs had a worldwide distribution throughout their evolution, since there was no identification of specific AHL to the environment of the organism which produces it (Rasmussen et al., 2014). Another survey, focusing on thermophilic bacteria, showed that they use AIs-2 for QS communication. The phylogenetic tree comparison of LuxS and the 16S rRNA of thermophiles and mesophiles show that LuxS (which produces the autoinducer) of mesophilic bacteria may originate from thermophiles. Also, the LuxS proteins of thermophilic bacteria within a phylum were evolutionarily closer compared to LuxS of different phyla (Kaur et al., 2018).

In this study we examine the evolutionary relationships between different QS systems. To achieve this, we conducted phylogenetic and comparative analysis of the components of these pathways. We used the KEGG database to identify all of the QS proteins that are also part of the TCS, and we collected their amino acid sequences from various bacterial species. The KEGG and SMART databases were used to compare the combinations of highly conserved domains in HKs and RRs. Phylogenetic analysis was based on Neighbor-Joining and Maximum Likelihood methods.

Methods

Choice of species and proteins

We used the **KEGG Database** (Kyoto Encyclopedia of Genes and Genomes, https://www.kegg.jp/ or https://www.genome.jp/kegg/), to select the amino acid sequences

used in this study. First, the Two Component System protein types that are also part of the Quorum Sensing system were found using the pathway maps of the KEGG PATHWAY database (maps ko02020 and ko02024 respectively). Each protein type is attributed to a specific orthology group accession number. By entering each identifier in the KEGG ORTHOLOGY database section, we were able to gather information about the amino acid sequences of each protein type in different bacterial organisms ("Genes" section) and the lineages in which each protein type is present ("Taxonomy").

Distribution of QS proteins

Using the KEGG database and the QS map (map ko2020), we checked for the presence of the protein subunits of different QS pathways in selected groups of bacteria. The chosen groups of bacteria include various Phyla, Classes and Genera, as shown in Supplementary Table S1.

Conserved domains

One amino acid sequence for each protein type was chosen randomly to identify the conserved domains of it, by searching in the "motifs" section in the KEGG GENES database for each amino acid sequence. We also used the SMART database (Simple Modular Architecture Research Tool, http://smart.embl-heidelberg.de/), in order to identify additional domains in our selected amino acid sequences not shown in KEGG. The domains of the SMART and KEGG GENES databases that were finally taken into consideration for our research, were those with an E-value of at least 10^{-5} which were also mentioned as "Confidently predicted domains" in the SMART database. In general, most of the domains were the same in the two databases. Next, using the **Pfam** database (https://pfam.xfam.org/), we found the groups (clans) in which these domains are classified (Supplementary Table S2) as well as information about their functions. For further information we used the Conserved Domain Database (**CDD**) in NCBI (https://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml).

Taxa and families representation

For the phylogenetic analysis, the number of the collected amino acid sequences for each protein type was kept as low as possible so as to have a manageable amount of data but at the same time include a representative portion of the different taxa and families which include each protein type. As a result, the average number of the amino acid sequences for each protein type was about 20. Supplementary Tables S3A and S3B show the representation percentages of the taxa, families and species out of the total available ones in the KEGG database.

Phylogenetic analysis

Multiple alignments of the sequences were created using the MUSCLE alignment

program (MUltiple Sequence Comparison by Log- Expectation, available at the EBI webpage: https://www.ebi.ac.uk/Tools/msa/muscle/. We kept the default settings for the following amino acid sequence alignments:

- each HK family
- all the HKs
- each RR family
- all the RRs

Additionally to the FASTA results of the alignment, we also saved the Nexus format for the phylogenetic trees, which was inferred with the Neighbor Joining Method and the results of the alignments were visualized with the Jalview program (https://www.jalview.org/).

The multiple alignment for all the HK proteins consisted of 347 amino acid sequences and had a total length of 1919 amino acids. There were several clusters of conserved residues mixed with gaps or non-conserved residues towards the N-terminal region of the alignment (41-121, 202-662) and in some parts as we approach the C-terminal part (949-1411). The rest of the alignment contained mostly gaps or non-conserved parts. The trimmed alignment had 455 amino acids (Supplementary Fig. S2A – S2C).

The total length of the alignment for all the RR proteins was 670 amino acids and 293 amino acid sequences were used. The alignment contained conserved clusters alternating with gaps or non-conserved residues (65-224, 287-528). Trimming resulted in an alignment of 274 amino acids (Supplementary Fig. S3A – S3C).

The substitution model for each multiple alignment was inferred via the ProtTest3.4 program (https://github.com/ddarriba/prottest3/releases) (Abascal et al., 2005). In order to compute the likelihood scores, the following settings were used:

− Substitution matrices to test: VT, Blosum62, JTT, Dayhoff, WAG
− Rate variation: I, G, I+G
− Amino acid frequencies: unchecked empirical

The results were evaluated by the Akaike information criterion (AIC).

The data of the multiple alignments were used in the Phylogeny.fr webpage (http://www.phylogeny.fr/ -also available at https://ngphylogeny.fr/). We used the "A la Carte" function, which enables various adjustments for the construction of the phylogenetic trees: we chose the Maximum Likelihood method (PhyML), the SH-like Approximate Likelihood-Ratio Test (aLRT) as the statistical test for branch support and we also set a number of 100 bootstraps for the bootstrapping procedure. For each multiple alignment we chose the appropriate substitution model. Due to limitations in the amount of data, we were not able to use this program for the multiple alignment of all the HKs (maximum number of sequences is 200 and maximum length of sequences is 2000 for proteins) (Dereeper et al., 2008).

The phylogenetic trees for all the protein families together (2 trees: HKs and RRs) were also constructed using the CIPRES Science Gateway (https://www.phylo.org/portal2/login.action), using the RAxML code (Randomized Axelerated Maximum Likelihood): *RAxML-HPC BlackBox (8.2.12) - Phylogenetic tree*

inference using maximum likelihood/rapid bootstrapping on XSEDE (Stamatakis, 2014). Before entering the multiple alignments to the program, we removed any positions in the alignments which included gaps or non-conserved residues. For this purpose, we used Jalview to detect those positions and then the Mesquite program to do the trimming (http://www.mesquiteproject.org/). All the phylogenetic trees were visualized with the Figtree program (download available at https://github.com/rambaut/figtree/releases). The results were saved as PNG images and were further edited with Microsoft Draw. Most of the nodes were labelled as white, grey or black bullets, if their bootstrap value was 50%-80%, 80%-95% and 95%-10% respectively. The raw and masked muscle alignments, as well as the phylogenetic trees in nexus format are available in the Supplementary dataset.

Results

Distributions of QS proteins

The distribution of each protein type is clearly demonstrated by Fig. 1. All pathways are fully present in at least one bacterial group, but many groups of bacteria have some of the subunits of each pathway. In detail, all Lux pathways are fully present in γ-proteobacteria. Also, α-proteobacteria, deinococci, chloroflexi, cyanobacteria, actinobacteria, clostridia and bacilli (and other Spirochaetes) possess most or all proteins of the Lsr protein pathway**.** The Lux pathway contains two different protein types: a pathway of HKs and an RR, and the cytoplasmic Lsr proteins; both pathways are activated by the AI produced by LuxS. The UvrY pathways (BarA/UvrY and SdiA) are fully present in γ-proteobacteria and almost all in β-proteobacteria. Several proteins of all stages of one or both of the pathways are also found in δ-proteobacteria, Nitrospira and Chrysiogenetes. Like in the Lux pathway, there are two different types of receptors, an HK (BarA) and a cytoplasmic protein (SdiA). Concerning the NarL pathway, γ-proteobacteria have all proteins of the pathway and one of the receptors in α-, β- proteobacteria, chloroflexi, cyanobacteria, actinobacteria, clostridia and bacilli. All three proteins of the DesR pathway are present in α-,β-,γ- actinobacteria, clostridia and bacilli. Only γ-proteobacteria possess all proteins of the Rcs pathway and β-proteobacteria and Bacteroidetes contain the HK and RR of the pathway. The Agr pathway (both AgrC/AgrA and SaeS/SaeR) is fully present in bacilli. A-proteobacteria and clostridia contain the Agr proteins. All proteins of the Com(XQPASK) pathway are found in bacilli and all proteins up to ComA are in clostridia. Also, the HK/RR pair ComP/ComA is present in chloroflexi. Similarly, the proteins of Com(ABDEX) are all found in bacilli, whereas the HK/RR pair ComD/ComE is present in clostridia. All the proteins of the Nis pathway are present in both clostridia and bacilli. Also, bacteroidetes and actinobacteria contain all of the proteins except one and two respectively. Other bacterial groups contain parts of the pathway but no other contains the sensing system (both the HK and RR). Bacteroidetes, actinobacteria, clostridia and bacilli contain both the CiaH and CiaR proteins.

All three proteins of the Fus pathway are in γ-proteobacteria, however, the HK/RR pair

is found also in α- and β-proteobacteria, bacteroidetes, deinococci, actinobacteria and Other Spirochaetes. Concerning the Qse pathway, γ-proteobacteria contain all of the proteins (QseC/QseB and QseE(GlrK)/QseF(GlrR)). QseC/QseB are present in bacteroidetes, actinobacteria and aquificae. QseE(GlrK)/QseF(GlrR) are found in β- and δ-proteobacteria and nitrospira. No bacterial group contains all the proteins of the Rpf pathway, since RpfB is absent in α- and β- proteobacteria. However, it should be pointed out that although RpfB is present in the KEGG map for Quorum Sensing, it is not included in Pathway or Brite in KEGG ORTHOLOGY. Also, the HK/RR pair RpfC/RpfG are found in nitrospira.
Out of all the aforementioned bacterial groups, there are several other groups of bacteria in which these proteins are located, but usually alone without other proteins of the same pathway, hence these incidents are not mentioned.

Fig. 1 shows that the majority of the complete pathways are found in α-, β- and γ-proteobacteria, clostridia and bacilli. A significant number of proteins are also found in δ- and ε-proteobacteria, bacteroidetes and actinobacteria. Also, based on the average bacterial groups for all the proteins in each pathway, the most widespread proteins are those of the UvrY, DesR, Fus, Nis pathways and the Lsr proteins of the Lux pathway. In the least bacterial groups are found the proteins of the two Com pathways and the Rcs and Agr pathways. Looking at each protein separately the most widespread proteins are CsrA, QseB, DesA, FusR, malt and LuxS (Fig. 2 and Supplementary Table S4). Interestingly, those belong to different pathways (CsrA:UvrY, FusR: Fus, QseB:Qse, DesA:DesR, malt and LuxS: Lux pathway). The Lux and UvrY pathways contain the most widespread as well as some of the most limited protein types (LuxQ, Qrr and CsrB). There is also a difference even in the distribution between the HKs and their cognate RRs: In most cases, the RRs are found in more bacterial groups than their cognate HKs (NarL, ComA, NisR, FusR, QseB, GlrR, and RpfG).

Conserved domains

In the Pfam database, the domains are organized into larger groups called “clans”. The NCBI database characterizes most domains as “superfamilies”. Table 1 contains the clans and their domains found in the studied amino acid sequences and Supplementary Table S2 shows the accessions for each domain in PFAM and InterPro, as well as the entry of the superfamily (NCBI-CDD) and pfam clan they belong to. The following domains do not belong to any clan: MASE1, PilJ, Hpt, Sigma54_AID, Trans_reg_C

<u>Conserved domains in HKs:</u> The length of the HKs varies from 406 (AgrC) to 949 (RcsC) amino acids. The longest HKs are the hybrid ones (NarL family: BarA, RcsC, Lux family: LuxN, LuxQ and CqsS, “Other”: RpfC) which also contain more conserved domains. Differences in the length of the sequences can also be found within the proteins of the same family. The HKs LuxQ and RcsC function as a complex with another protein (LuxP and RcsF respectively). Most HKs contain 1-3 conserved domains and up to 6 (Fig. 3). The common trait is the existence of either HATPase_c or HATPase_c_5, which are also the only conserved domains

in the protein NisK and the proteins of the LytTR family. HAMP is the second most frequent domain and it is found in HKs of the NarL family (NarX, NarQ, BarA). The Response_reg and Hpt domains are located in the C-terminal region of hybrid HKs, but only Hpt can alternatively be found as an independent protein (within the Rcs pathway). Concerning the N-terminal region, only few proteins contain conserved domains, most of which belong to the NarL family (QseC, NarX, NarQ, FusK).

RRs Conserved domains: The average length of RRs is significantly shorter that the HKs. It varies from 207 to 455 amino acids. Most RRs contain 2 conserved domains (Fig. 3) except for LuxO (3 domains). The N-terminal regions always contain the Response_reg domain, which is a common structure trait for all the RRs. The conserved domain on the C-terminal region is different in each RR family. The only exception is the protein GlrR (NtrC family), which has the same conserved domain as LuxO (Lux family).

An overview of the conserved domains of both HKs and RRs is present in Table 2. HATPase_c is the most common conserved domain, as it is found in 15 protein types. HisKA is the second most common with 9 occurrences.

Phylogenetic trees by family

Given the families formed by the HKs and their cognate receptors, we constructed four phylogenetic trees for the HKs and three for the RRs. The HKs GlrK (NtrC family), RpfC ("Other"), their cognate RRs GlrR and RpfG and also LuxO are the only members of their family, hence no phylogenetic tree was inferred.

The NarL family

The HKs of the NarL family are organized in distinct clades in the PhyML tree (Fig. 4A). Each clade consists of amino acid sequences of the same protein type. The same structure is observed in the phylogenetic tree of their cognate RRs (Fig. 4B). Although both trees have some amino acid sequences away from the clade of their protein type (e.g. marrComP, mdnBarA, paejNarQ and anxNarX), the phenomenon is more prevalent in the RRs. Notably, the structure of the phylogenetic tree shows three pairs of HK types (RcsC-BarA, ComP-FusK and NarQ-NarX), which group closer to each other than to the other protein types. This pattern is also seen for their cognate RRs.

RcsC-BarA and RcsB-UvrY: Most of the clades of each HK type are supported by high bootstrap values and they are monophyletic. The clades of the RR types are also supported by good bootstrap values, but most of them are below 95%. The clade of the RcsC proteins is closer to BarA than the rest of the protein types and their common node has 96% bootstrap value. Both clades are supported by high bootstrap values (97% and 93% respectively). Only mdnBarA is outside of the BarA clade. Their cognate RRs (RcsB and UvrY) also create a common clade with 87% bootstrap value. However, these are not too distant from the rest of the clades; in particular, they are close to FusR RRs but without significant bootstrap

support (below 50%). The RcsB RRs are monophyletic in both trees with 100% bootstrap support, except for sequences abamRcsB and hee1RcsB. The UvrY proteins also form a monophyletic clade, but with lower (58%) bootstrap support (excluding the nodes of the first two proteins-sbxUvrY and mmarUvrY which have very low bootstrap support), while the sequences dunUvrY, cbaeUvrY and maesUvrY fall outside the UvrY clade.

ComP-FusK and ComA-FusR: The clade of all the ComP HKs has 88% bootstrap support, but excludes marrComP. The ComP clade groups, with high bootstrap support (91%), with the FusK clade. The latter includes most FusK proteins, with 93% bootstrap support. Concerning their cognate RRs, the ComA and FusR clades are also monophyletic. The node for all the ComA sequences has a 94% bootstrap value and the FusR clade has 95% bootstrap value, but excludes the sequences spirFusR and sphnFusR. Concerning their position, in contrast to the ComP and FusK, the ComA and FusR clades do not form a monophyletic group.

NarQ-NarX and NarP-NarL: The NarQ and NarX proteins are gathered in monophyletic clades. The NarX proteins are separated in two clades, one of which shares a common node with the NarQ clade. All three clades have high bootstrap values (NarQ: 98%, NarX: 98%, and 100%), as well as the common clade of the NarX and NarQ proteins (98%), with the exception of sequences paejNarQ and anxNarX, which lie outside these clades. However, the distribution of the cognate receptors NarL and NarP is more complicated. They are divided into several clades (sometimes including both protein types), yet these clades are placed close to each other and their common node has a 90% bootstrap value. In detail, the NarP sequences are grouped into two distant main clades; a larger one, neighboring with the NarLs and a smaller one. Their bootstrap value is 50% and 59% respectively. There are more sequences outside of these clades than for their cognate HK sequences, and from different species than in the HKs tree: miiNarP, polaNarP, tamNarP and schNarP. Many NarL sequences are dispersed in the phylogenetic tree. The NarLs are gathered in several small clades close to each other. Most of their nodes demonstrate bootstrap values of at least 80%. The main NarP and NarL clades group close to each other, similar to their cognate HKs. However, the bootstrap value of their common node is higher in the HK tree.

The OmpR family

All different HK types of the OmpR family are grouped into separate clades. None of the HK sequences is misplaced in a clade of a different protein type. Their cognate RRs follow a similar pattern, except for sequences phmQseB and aprsQseB, which lie outside the QseB clade. There are two related groups in both phylogenetic trees: CiaH - QseC and NisK – SaeS and their cognate RRs, CiaR - QseB and NisR – SaeR (Fig. 5A and 5B).

CiaH-QseC and CiaR-QseB: The majority of the CiaH sequences stems out of a common node and the taxa which appear earlier on the main clade are soxCiaH and cooCiaH. The bootstrap value for all the CiaHs is 97%. The QseC sequences are distinctly separated from the rest of

the RR types and they share a common node of 92% bootstrap value with the CiaH clade. The clade for all the QseC HKs is statistically supported by a 100 % bootstrap value. The cognate RR of CiaH -CiaR- is also monophyletic, although psorCiaR branches early on the clade. The common node of all the CiaR sequences is 99%. The QseB sequences are more dispersed than the QseC. The tscQseB, ftsQseB and rtrQseB sequences are found in an earlier branch with a low bootstrap value. Except for these sequences, the main QseB clade is supported by an 89% bootstrap value. However, phmQseB and aprsQseB are found between the two clades of the NisR proteins with a 93% bootstrap value; both of these sequences show significant differences in the conserved regions.

NisK -SaeS and NisR-SaeR: The NisK sequences are organized into one main clade, with a high bootstrap value (99%) and they have a common node of 92% bootstrap value with the SaeS proteins. The SaeS sequences are placed into one main clade, with 99% bootstrap support. The NisR clade is also monophyletic (87% bootstrap value), except for spyaNisR, which forms a clade with two distant QseB sequences among the NisR sequences. The SaeR sequences are grouped into one main clade with a 100% bootstrap value, and their common node with the NisR clades has 96% bootstrap support.

The Lux family HKs

The sequences of each HK type of the Lux family (LuxN, LuxQ and CqsS) are grouped in distinct clades (Fig. 6). The LuxN sequences have one common node, which is divided into four groups of branches. The nodes are supported by a high bootstrap value (80% minimum). Most LuxQ branches appear near the core of the tree. The most distant LuxQ sequences are dhyLuxQ, daLuxQ, oceLuxQ and vgaLuxQ. The bootstrap values for all the nodes of the LuxQ sequences are 50% minimum and the majority of them have at least a 95% bootstrap value. PhyML categorizes all CqsS sequences in one clade; excluding fsoCqsS, the bootstrap value for the CqsS clade is 92%. As there is only one RR for all the Lux family HKs (LuxO), there is no phylogenetic tree for the RRs of the Lux family.

The LytTR family

LytTR family – HKs: The different types of HKs of the LytTR family, as well as their cognate RRs, are clearly separated from each other forming monophyletic clades and both phylogenetic trees have a similar overall appearance (Fig. 7A and 7B).

AgrC-AgrC2 and AgrA-AgrA2: The AgrC sequences form several clades, all of which are close to each other on the phylogenetic tree (with the exception of sndAgrC). The bootstrap values of the nodes vary; however, the majority is above 80%. The two AgrC2 sequences are found in a distinct clade, which derives from the AgrC clade, with 92% bootstrap support. The closest AgrC protein to them is rimAgrC. Concerning the RRs, the AgrA sequences do not form one distinct clade, but smaller ones, with fewer sequences each than for AgrC,

gathered on one side of the phylogenetic tree. The AgrA2 sequences belong to the AgrA clade, but with insignificant bootstrap value. The closest AgrA protein differs from the HKs (cssAgrA).

ComD and ComE: Most of the ComD sequences are gathered in one clade, with 97% bootstrap support, apart from the clade of lsjComD and lhiComD and the sequence laliComD, which forms a clade with the HKs asanAgrA and lps1AgrA. The std2ComD and smb1ComD sequences appear a lot earlier than the rest of the ComD proteins and they are closer to AgrC, with 97% bootstrap value. The node of the ComE clade is statistically supported by a high bootstrap value (99%). The std2ComE and smb1ComE sequences branch early, as do their cognate HKs, but with a low bootstrap value.

Phylogenetic trees for all the families together

RaxML phylogenetic tree with all the HK families

In general, the members each of HK family form one main clade. The clades of each HK protein type are also mostly monophyletic (Fig. 8). The sequences of the LytTR family, as well as the RpfC ("Other") and GlrK HK types (NtrC family) are each gathered in one clade, whereas the protein sequences of the Lux as well as the NarL family are found in more than one clades. The following groups of families are closer to each other: 1) LytTR and NarL (NarQ, NarX, FusK and ComP), 2) OmpR and NtrC, 3) Lux, NarL (BarA, RcsC) and "Other" (RpfC). However, the families of the groups 2) and 3) are also very close to each other.

NarL family: The sequences of this family form three main clades: the RcsC clade with an 80% bootstrap value, the BarA clade with a 63% bootstrap value and a third -and larger- one which includes the NarQ, NarX, FusK and ComP sequences. The FusK sequences form a monophyletic clade and derive from a common node with 100% bootstrap value (excluding mpurFusK, djaFusK, amgFusK as they have very low bootstrap values). All the ComP sequences also derive from a common node, but with very low bootstrap value. The NarQ main clade has a 58% bootstrap value. There are also two main clades for the NarX proteins, which are supported by a 55% and 86% bootstrap value respectively. Some NarQ and NarX sequences fall close to, but outside their protein type clades. In general, the grouping of the clades of the NarL family mirrors the PhyML tree for the HKs of the NarL family alone, resulting in three pairs: FusK-ComP, BarA-RcsC, and NarX-NarQ.

OmpR family: The HK types of this family are separated into two main clades: one main clade for the SaeS and NisK sequences and another for QseC and CiaH, which recapitulates the structure of the phylogenetic tree of the OmpR family alone (Fig. 5). The clades of each protein type of the OmpR family are monophyletic and supported by high bootstrap values (SaeS: 92%, CiaH: 89%, NisK: 74% and QseC: 96%), although their common node has low bootstrap support.

Lux family: Two distinct monophyletic clades and a smaller mixed one are formed by the HKs of the Lux family: one clade exclusively consists of CqsS and LuxN sequences and it is statistically supported by a 100% bootstrap value (excluding fsoCqsS which is distant from the rest of the CqsS sequences and it has 57% bootstrap value) and the other is made of only LuxQ sequences with a 92% bootstrap value. The mixed clade consists of LuxQ and one RcsC sequences (oceLuxQ, dsaLuxQ, dhyLuxQ and psegRcsC). In contrast to the phylogenetic tree of the Lux family alone, in this tree CqsS and LuxN are grouped together, while the LuxQ sequences are placed away from the rest of the members of the Lux family. They are closer to the BarA sequences and are part of a bigger clade including the RpfC and RcsC sequences, but with low bootstrap value.

LytTR family: All the sequences of this family derive from a common node with 89% bootstrap value. The ComD sequences share a common ancestor. The exceptions are the same as in the PhyML tree of the OmpR family: laliComD is in one clade with lps1AgrC, while asanAgrC and the sequences lsjComD and lhiComD form a distinct clade. The difference is that both of these clades have a common node (with low bootstrap value), whereas the phylogenetic tree of the OmpR family places them distant from each other. The AgrC sequences form many small clades, all of which are close to the ComD clades. They also include the two AgrC2 sequences. Their bootstrap values vary from low to high. The closest clade to the LytTR family is the clade of the ComA sequences of the NarL family, but without significant support for their common node.

NtrC family (GlrK proteins-QseE): All the sequences of the GlrK (QseE) HK are in one clade, with 57% bootstrap value. The closest clade to them is that of QseC and CiaH, although the bootstrap value for their common node is very low.

"Other" (RpfC sequence): All RpfC sequences are placed in one clade, with 80% bootstrap value. The neighboring clade consists of RcsC sequences of the NarL family, but the bootstrap value for their common clade is very low.

PhyML and RaxML phylogenetic trees with all the RR families

In general, both phylogenetic trees of the RRs for all the families together show similar results. It is clear that the sequences of each RR type remain together, like in the phylogenetic tree of each family. The closely related RR types of the trees of each family are mostly the same in the PhyML and RaxML trees of all the families together.

PhyML phylogenetic tree

The RR families are distinctly separated from each other in the phylogenetic tree, as well as the protein types within each family. Only few sequences fall outside of the clade of their

family or RR type (Fig. 9). An overview of the tree reveals a closer relationship between the following family groups: 1) NtrC, Lux and "Other" (RpfG) 2) OmpR and group 1, 3) NarL and LytTR.

NarL family: The clades contain only RRs of the NarL family. Most of the ComA, FusR, UvrY and RcsB proteins are monophyletic. In contrast, the NarP and NarL sequences are divided into several clades, which are either of the same protein type or mixed. The RcsB sequences form one monophyletic clade with 95% bootstrap value and only abamRcsB is among different protein types. Closest to the RcsB clade are the sphnFusR and dunUvrY sequences and also the common clade of the FusR and UvrY RRs.The FusR RRs are all in one monophyletic clade of 85% bootstrap value except sphnFusR, which is near the RcsB clade. The FusR clade is close to the UvrY proteins, having a common node of 76% bootstrap value. The UvrY clade is also monophyletic with 97% bootstrap value. However, multiple UvrY proteins lie outside this clade (tjeUvrY, dunUvrY and the clades of mmarUvrY, mcbUvrY, noaUvrY and cbaeUvrY, sbxUvrY). All the ComA sequences create a monophyletic clade of 41% bootstrap value, their closest neighbor being a clade of NarP RRs with a bootstrap value of 95% for their common node. There are two clades containing exclusively NarP RRs: one with 52% bootstrap value, which shares a common node of 95% with the ComA clade and another one next to NarL sequences (73% bootstrap value). The sequences of schNarP, ssilnarP, miiNarP and rspNarP lie outside these two main clades. Concerning the NarL RRs, they create several small clades, either with only NarL sequences or mixed ones with NarP, UvrY, FusR or RcsB sequences.

OmpR family: While there is one main clade for the OmpR family, it consists of smaller clades, each of them containing one different protein type. The bootstrap value is 99% for the SaeR sequences, 97% for the CiaRs, 45% for the NisRs and 88% for the QseBs. Additionally, there are pairs of protein types which group close to each other: SaeR-NisR and CiaR-QseB, with 99% bootstrap value. The taxon of aprsQseB is much distant from the rest QseB sequences and the ftsQseB is the earliest of all the QseB sequences. In addition, phmQseB lies outside of the QseB clade.

Lux (LuxO protein): The LuxO sequences are very close to the NtrC family and their common node is statistically supported by a high bootstrap value (98%). The common node of the Lux clade has 55% bootstrap value and the hamLuxO sequence is the most distant taxon.

LytTR family: All the members of the LytTR family are gathered together in a clade of 100% bootstrap value. The ComE sequences form one clade with 99% bootstrap support, except srqComE, std2ComE and smb1ComE, which form a clade of 59% bootstrap among AgrA sequences. The AgrA sequences are spread in several clades within the LytTR clade, which are statistically supported by bootstrap values over 50%. The AgrA2 branches belong to the

clade of the cbbAgrA and cadAgrA sequences with 83% bootstrap value.

NtrC (GlrR protein): All sequences have a common node of a 98% bootstrap value, except for obgGlrR which falls outside this clade.

"Other" (RpfG protein): All the RpfG sequences form one clade with 100% bootstrap value. The closest RRs to them are the LuxO- GlrR sequences, but their common node has low bootstrap value (22%).

RaxML phylogenetic tree

The relative position among the different families of the RRs is similar in the RaxML (Fig. 10) and PhyML (Fig. 9) phylogenetic trees, although the distance between their clades is bigger in the PhyML tree. Similar to the PhyML tree: 1) the RpfG, the Lux and NtrC families are closer to each other, 2) the OmpR family is near the LytTR, Lux, NtrC and RpfC families and 3) the LytTR clade is closer to the NarL family.

NarL family: The NarL family members create a sum of clades, with a common node of very low bootstrap value (below 50%). Most clades consist mostly of sequences of one particular NarL protein type. The bootstrap values for each RR type clade are lower than in the phyML tree: FusR: 55% (spirFusR excluded, which has a very low bootstrap value), RcsB: 66% and ComA: 84%, while the UvrY RRs clade has a very low bootstrap value. The positions of all the family members are very close to each other, to the extent that there are no defined pairs of protein types. RcsB is the most distant RR type from the rest of the proteins. Similarly to the PhyML tree, abamRcsB is in a distant branch from the RcsB sequences. The clade of the UvrY sequences has, in general, nodes of low bootstrap value and five sequences are placed outside of it: tjeUvrY, dunUvrY and the clade of mmarUvrY, mcbUvrY and noaUvrY, which is among the NarL sequences. In contrast to the PhyML tree, sbxUvrY and cbaeUvrY are part of the UvrY clade. All the FusR and ComA RRs each form one clade, except sphnFusR (like in the PhyML tree), which is right before the clade of the RcsB proteins.

Most NarPs are divided to two main clades, a larger one with 65% bootstrap value and a smaller one with a low bootstrap value. The rest of the sequences are among NarL sequences. The NarL sequences are widely distributed in the clades of the NarL family, like in the PhyML tree. One of the largest clades (64% bootstrap value) has a common node with NarP sequences, with low bootstrap value. The rest of the NarL sequences are among NarL and UvrY sequences. The vbhNarL sequence is closer to the clade of the LytTR family.

OmpR family: The members of the OmpR family are gathered in one clade with 72% bootstrap value and there is one individual clade for each protein type, as in the PhyML tree.

However, the bootstrap values of these clades are lower than in the PhyML tree (QseB: 59%, CiaR: 99%, SaeR: 87% and NisR: 54%). The SaeR-NisR grouping appears like in the PhyML tree, but with 57% bootstrap value. Closely related are also the CiaR and QseB clades but their common node has a low bootstrap value. Also, phmQseB is placed outside the OmpR family clade, as in the PhyML tree.

Lux family: Although the Lux family is close to NtrC, the bootstrap value of their common node is low in the RaxML tree (37%). An even lower bootstrap value is found for the common node of all the LuxO sequences (17%), but the majority of the bootstrap values within the LuxO clade is 50-80%. Whereas the PhyML tree places hamLuxO away from the rest of the LuxO sequences, this doesn't happen in the RaxML tree.

LytTR family: The clade for all the RRs of the LytTR family is closer to the NarL family members. Several nodes within the LytTR family are statistically supported by a high bootstrap value, especially the early branches (AgrA sequences). Similar to the PhyML tree, the different RR types belong to separate clades, except std2ComE and smb1ComE. The ComE clade has 100% bootstrap value. There are many smaller AgrA clades, which start -in contrast to the PhyML tree- at nodes of low bootstrap value. The AgrA2 sequences have a common node with the AgrA clade, but with very low bootstrap value (17%).

NtrC family (GlrR protein): The GlrR sequences are monophyletic; however, the early nodes have a very low bootstrap value. The obgGlrR sequence branches out a lot earlier than the rest of the GlrR sequences with a very low bootstrap value (12%). In contrast to PhyML, obgGlrR is in the GlrR clade, but it has a low bootstrap value and the RaxML tree places dyeGlrR out of the NtrC clade, on the LuxO clade, with 60% bootstrap value.

"Other" (RpfG protein): The RpfG RRs form one clade of 100% bootstrap value, like in the PhyML tree. They are monophyletic and include no sequences of other RR type. The closest RR to them is not LuxO, as indicated by the PhyML tree, but the common clade of GlrR (NtrC family) and LuxO RRs, but again without significant statistical support.

Discussion

The QS mechanism includes a multitude of signal molecules and pathways, which aim to coordinate the activities of a bacterial population by changing their gene expression according to the population density. The fact that this mechanism affects various processes makes this mechanism a potential target in antibiotics research. Few phylogenetic analyses on the bacterial QS exist, and given the multiple pathways contributing to QS, they focus only on the signaling molecules, or specific pathways. Among them, the LuxI/R pathway

(Gram positive) has been mainly studied. Based on the lack of phylogenetic analyses of a wider spectrum for QS proteins, we chose using the KEGG database the proteins which belong to QS pathways and take part in a widespread signal transduction system, the Two Component System (TCS). This mechanism includes pairs of a Histidine Kinase (HK) and its cognate Response Regulator (RR). The studied QS pathways are found in a wide variety of both Gram-positive and Gram-negative bacteria.

QS protein conservation

The overall conclusion based on Fig. 1 is that α-, β-, γ-proteobacteria, clostridia and bacilli possess the highest number of the QS proteins. The most widespread proteins belong to various pathways and families and they have different roles, indicating no clear relationship between these characteristics and the number of the bacterial groups containing them. The most widespread proteins are CsrA, FusR, QseB, DesA, malt and LuxS. The least widespread proteins are found in the two Com pathways, the Rcs and Agr pathways as well as the Lux pathway. These results suggest that the wide presence of one protein doesn't mean a similar distribution for the rest of the proteins of the same pathway. This contradiction is pointed out by the fact that the Lux and UvrY pathways contain the most widespread as well as some of the most limited protein types (LuxQ, Qrr and CsrB). Also, there is a difference even in the distribution of the HK/RR pairs, between the HK and its cognate RR. Examining each protein family of our study separately, it seems that the most widespread HK/RR pairs of QS are those of the NarL, OmpR and NtrC families, whereas the Lux and LytTR are the least common.

Conserved Domains

The HK-RR pairs belong to 6 families, as defined by KEGG: OmpR, LytTR, NarL, Lux, NtrC and "Other". Both in the OmpR and the LytTR family, the different protein types consist of a small number of common conserved domains. On the other hand, HKs of the NarL and Lux families show common domains and a significant variety of other domains. In contrast to the OmpR and LytR families, the NarL and Lax also include hybrid HKs. The RRs of all families together contain common domains (all of them have the Helix-Turn-Helix motif).

The Lux family consists exclusively of Hybrid HKs, and all 3 of them activate a common phosphotransfer protein and RR. The NtrC contains only the GlrK HK, which has the same conserved domains with the HKs of the OmpR family, a characteristic being confirmed by the PhyML phylogenetic tree, as both of the families are found close to each other. However, its cognate RR (GlrK) has various common domains with the RR of the Lux family. This is also evident in the PhyML and RaxML phylogenetic trees. Therefore, we conclude that similarities can be found not only among the protein types within a family, but also among different family members, which indicate a possible common evolution pathway.

A comparison of the Grebe and Stock study about the classification of the HKs (Grebe

and Stock, 1999) with the present study leads to the following conclusions (only some of the HKs mentioned here were also included in that previous study): The HKs ComP, NarQ and NarX were classified in the group HPK7, CiaH in HPK1a, NisK in HPK3c and BarA, RcsC, RpfC in HPK1b. Therefore, this way of classification differs from the KEGG database and our phylogenetic trees results, as CiaH and NisK (OmpR family members) belong to separate groups and RpfC (hybrid HK) is placed with the hybrid HKs of the NarL family (although hybrid HKs were placed in other HPK groups as well).

In the study of Forst and Kim, BarA and RcsC were characterized as Type IB and NarX and NarQ as Type III HKs (Forst and Kim, 2015). However, no members of other families included in our study were present in that analysis, so we cannot conclude if the comparison criteria used by Forst and Kim led to a different classification or if their classification is more detailed and separated the members of one family to more than one groups.

Capra and Laub classified the HKs into 5 groups depending on the presence of PAS, HAMP and GAF domains (Capra and Laub, 2012). It is difficult to directly compare our results with their analysis as no PAS and GAF domains were found in our study (results by KEGG and SMART). Nevertheless, all three HKs with a HAMP domain (NarX, NarQ, BarA) were grouped to the same family (NarL), in agreement with the Capra and laub study. Similarly, the effector domains for the classification of the RRs were also different in the research of Capra and Laub and our study, as no GGDEF and Methyltransfer domains were present. However, the LuxO and GlrR proteins possessing the AAA_5 domain, were grouped close to each other in our phylogenetic trees, complying with the classification criterion of Capra and Laub.

Concerning the classification of the RRs according Hakenbeck and Stock (Hakenbeck and Stock, 1996) it is the same as the data of KEGG and our study. CiaR and NisR were classified to the OmpR family and ComA, NarP, NarL and RcsB were classified to the FixJ family. The results of Grebe and Stock are similar(Grebe and Stock, 1999): CiaR and NisR were classified to the R_{A1} group and ComA and RcsB to R_E, but NarP and NarL were classified to a different group (R_{C1}). In this paper, AgrA and ComE are also mentioned and grouped to R_D, both of which belong to the LytTR family in our study.

Phylogenetic trees

In all the inferred phylogenetic trees of each family (PhyML, RaxML), the amino acid sequences of both HKs and RRs are grouped in separate distinct clades according to their protein type and in many cases, with high bootstrap value. The same happens in the phylogenetic trees of all families together, where amino acid sequences of different family and protein types belong to distinct clades.

The different protein types of the OmpR family demonstrate the same relation in all

the phylogenetic trees (of the OmpR family alone and the tree with all the HK families together). There are two pairs of HKs: CiaH-QseC and SaeS-NisK. Their cognate receptors follow the same relation in all the three RRs phylogenetic trees: the SaeR clade is closer to NisR and the CiaR clade is closer to QseB.

The protein members of the LytTR family have separate clades for the AgrC -ComD and ArgA-ComE sequences in all trees, although the RRs show more exceptions. The AgrC2 and AgrA2 sequences belong to the AgrC and AgrA clade, which is confirmed by all the inferred phylogenetic trees.

Regarding the Lux HKs, while the PhyML phylogenetic tree of the family places each HK type far apart from each other, the RaxML tree for all the HKs shows a similarity between the CqsS and LuxN. It also places the LuxQ sequences closer to other families (BarA, RpfC and RcsC sequences). Their common RR (LuxO) is in the same clade with GlrR in both RaxML and PhyML trees of all RRs together.

The closer protein types for the amino acid sequences RpfC and GlrK (NtrC family) are also different for the HKs and their RRs. In the PhyML tree of all the HKs, the RpfC sequences are closer to the NarL family (RcsC) and GlrK is closer to the OmpR family (the common clade of QseC and CiaH HKs). On the other hand, GlrR is closer to LuxO in both PhyML and RaxML trees and RpfG sequences are next to clade of LuxO and GlrR.

The members of the NarL family form pairs of HKs that are closely related: 1) NarX and NarQ, 2) BarA and RcsC, 3) ComP and FusK. Out of these pairs, the first is closer to the third in both PhyML and RaxML phylogenetic trees. The cognate receptors of the NarL family demonstrate in the PhyML tree of the NarL family a similar grouping to the HKs from which they are activated. However, this grouping was different in the PhyML tree of all families, as the following pairs were formed: FusR-UvrY, RcsB-common clade of FusR-UvrY, ComA-NarP. The NarL and NarP sequences had various clades and didn't have a specific relation with any RR type. Also, the RaxML tree for all the families together did not make any of these pairs clear, as they were all very close to each other.

The relations among the HKs are also reflected by the relations of their cognate receptors, indicating a common evolutionary history, as has already been suggested by former studies on other proteins of the TCS. Notably, a comparison of all the three phylogenetic trees of all HKs and of all RRs suggests three clusters of the QS protein families: 1) NtrC, Lux and "Other", 2) OmpR with group 1 and 3) NarL and LytTR. Therefore, a stronger connection between different types of HKs in each family is obvious, as they are always placed in close clades in all phylogenetic trees. Their cognate RRs demonstrate the same pattern. These close evolutionary relationships highlight a common evolutionary history, and can inform future applications, such as the design of novel inhibitors for pathogenic QS systems.

References

Abascal, F., Zardoya, R., Posada, D., 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21: 2104–2105.

Bandara, H.M.H.N., Lam, O.L.T., Jin, L.J., Samaranayake, L., 2012. Microbial chemical signaling: A current perspective. Crit. Rev. Microbiol. 38, 217–249. https://doi.org/10.3109/1040841X.2011.652065

Capra, E.J., Laub, M.T., 2012. Evolution of Two-Component Signal Transduction Systems. Annu. Rev. Microbiol. 66, 325–347. https://doi.org/10.1146/annurev-micro-092611-150039

Dogsa, I., Choudhary, K.S., Marsetic, Z., Hudaiberdiev, S., Vera, R., Pongor, S., Mandic-Mulec, I., 2014. ComQXPA quorum sensing systems may not be unique to Bacillus subtilis: A census in prokaryotic genomes. PLoS One 9, e96122. https://doi.org/10.1371/journal.pone.0096122

Dereeper, A., Guignon, V., Blanc, G., Audic, S., Buffet, S., Chevenet, F., Dufayard, J.F., Guindon, S., Lefort, V., Lescot, M., Claverie, J.M., Gascuel, O., 2008. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res. 36. https://doi.org/10.1093/nar/gkn180

Forst, S., Kim, D., 2015. Genomic analysis of the histidine kinase family in bacteria and archaea. Microbiology 147, 1197–1212. https://doi.org/10.1099/00221287-147-5-1197

Galperin, M.Y., 2006. Structural classification of bacterial response regulators: Diversity of output domains and domain combinations. J. Bacteriol. 188, 4169–4182. https://doi.org/10.1128/JB.01887-05

Grebe, T.W., Stock, J.B., 1999. The histidine protein kinase superfamily. Adv. Microb. Physiol. 41, 139–227. https://doi.org/10.1016/s0065-2911(08)60167-8

Hakenbeck, R., Stock, J.B., 1996. Analysis of two-component signal transduction systems involved in transcriptional regulation. Methods Enzymol. 273, 281–300. https://doi.org/10.1016/s0076-6879(96)73026-4

Jacob, S., Foster, A.J., Yemelin, A., Thines, E., 2014. Histidine kinases mediate differentiation, stress response, and pathogenicity in Magnaporthe oryzae. Microbiol. Open 3, 668–687. https://doi.org/10.1002/mbo3.197

Kaserer, A.O., West, A.H., 2010. Histidine kinases in two-component signaling pathways, in: Handbook of Cell Signaling, 2/E. Elsevier Inc., pp. 581–586. https://doi.org/10.1016/B978-0-12-374145-5.00078-4

Kaur, A., Capalash, N., Sharma, P., 2018. Quorum sensing in thermophiles: prevalence of autoinducer-2 system. BMC Microbiol. 18, 62.

KEGG PATHWAY: map02024 [WWW Document], 2019. URL https://www.genome.jp/kegg-bin/show_pathway?map02024

Lerat, E., Moran, N.A., 2004. The Evolutionary History of Quorum-Sensing Systems in Bacteria. Mol. Biol. Evol. 21, 903–913. https://doi.org/10.1093/molbev/msh097

Mascher, T., Helmann, J.D., Unden, G., 2006. Stimulus Perception in Bacterial Signal-Transducing Histidine Kinases. Microbiol. Mol. Biol. Rev. 70, 910–938. https://doi.org/10.1128/mmbr.00020-06

Ng, W.-L., Bassler, B.L., 2009. Bacterial Quorum-Sensing Network Architectures. Annu. Rev. Genet. 43, 197–222. https://doi.org/10.1146/annurev-genet-102108-134304

Papenfort, K., Bassler, B.L., 2016. Quorum sensing signal-response systems in Gram-negative bacteria. Nat. Rev. Microbiol. 14, 576–588. https://doi.org/10.1038/nrmicro.2016.89

Rasmussen, B.B., Nielsen, K.F., Machado, H., Melchiorsen, J., Gram, L., Sonnenschein, E.C., 2014. Global and Phylogenetic Distribution of Quorum Sensing Signals, Acyl Homoserine Lactones, in the Family of Vibrionaceae. Mar. Drugs 12, 5527–5546. https://doi.org/10.3390/md12115527

Reading, N.C., Sperandio, V., 2006. Quorum sensing: The many languages of bacteria. FEMS Microbiol. Lett. 254, 1–11. https://doi.org/10.1111/j.1574-6968.2005.00001.x

Rutherford, S.T., Bassler, B.L., 2012. Bacterial quorum sensing: Its role in virulence and possibilities for its control. Cold Spring Harb. Perspect. Med. 2:a012427. https://doi.org/10.1101/cshperspect.a012427

Sreenivasulu, Y., 2015. THE POTENTIAL OF MICROBIAL WEALTH FOR PHYTOPATHOGEN MANAGEMENT AND PLANT GROWTH ENHANCEMENT. J. Biol. Nat. 2(1), 6–15.

Stamatakis, A., 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Tiwari, R., Karthik, K., Rana, R., Malik, Y.S., Dhama, K., Joshi, S.K., 2016. Quorum sensing inhibitors/antagonists countering food spoilage bacteria-need molecular and pharmaceutical intervention for protecting current issues of food safety. Int. J. Pharmacol. 12, 262–271. https://doi.org/10.3923/ijp.2016.262.271

West, A.H., Stock, A.M., 2001. Histidine kinases and response regulator proteins in two-component signaling systems. Trends Biochem. Sci. 26, 369–376. https://doi.org/10.1016/S0968-0004(01)01852-7

Whiteley, M., Diggle, S.P., Greenberg, E.P., 2017. Progress in and promise of bacterial quorum sensing research. Nature 551, 313–320. https://doi.org/10.1038/nature24624

Table 1: Categorization of the conserved domains of all the HK and RR types in this study.

Clans	Domains					
HTH (Helix-Turn-Helix)	HTH_8	Sigma70_r4	GerE	Sigma54_CBD	Sigma54_CBD	Sigma54_DBD
His kinase A	HATPase_c	HATPase_c_5	HisKA	HisKA_3		
P-loop-NTPase	AAA	AAA_5	Sigma54_activitat	Sigma54_activ_2		
Cache-like domains	2CSK_N	DUF2222	LuxQ-periplasm			
Chey-like superfamily	Response_reg	RcsD_ABL	RcsC			
HD-PDEase	HD	HD_5				
Clp crotonase	ECH_1	ECH_2				
Periplas BP	Peripla_BP					
HAMP	HAMP					
Tudor	LytTR					
YbjQ-like	RcsF					

Table 2: Occurence of the conserved domains of all the HKs and RRs in the different protein types and families.

Domain	Protein Types	Family
HATPase_c	15	All of the families
HATPase_c_5	3	LytTR
HisKA	9	OmpR, NarL, Lux, Other, NtrC
HisKA_3	4	NarL
HAMP	3	NarL
Response_reg	3	NarL, Lux, Other
Hpt	2	NarL, Other
Trans_reg_C	4	OmpR
LytTR	3	LytTR
GerE	5	NarL
Sigma70_r4_2	4	NarL
Sigma54_activat	2	LuxO, NtrC
Sigma54_activ_2	2	LuxO, NtrC

Figure legends

Figure 1: Presence of the protein subunits of different QS pathways in selected groups of bacteria. Colored sectors indicate presence, empty sectors indicate absence, based on the KEGG taxonomy database for each orthology group.

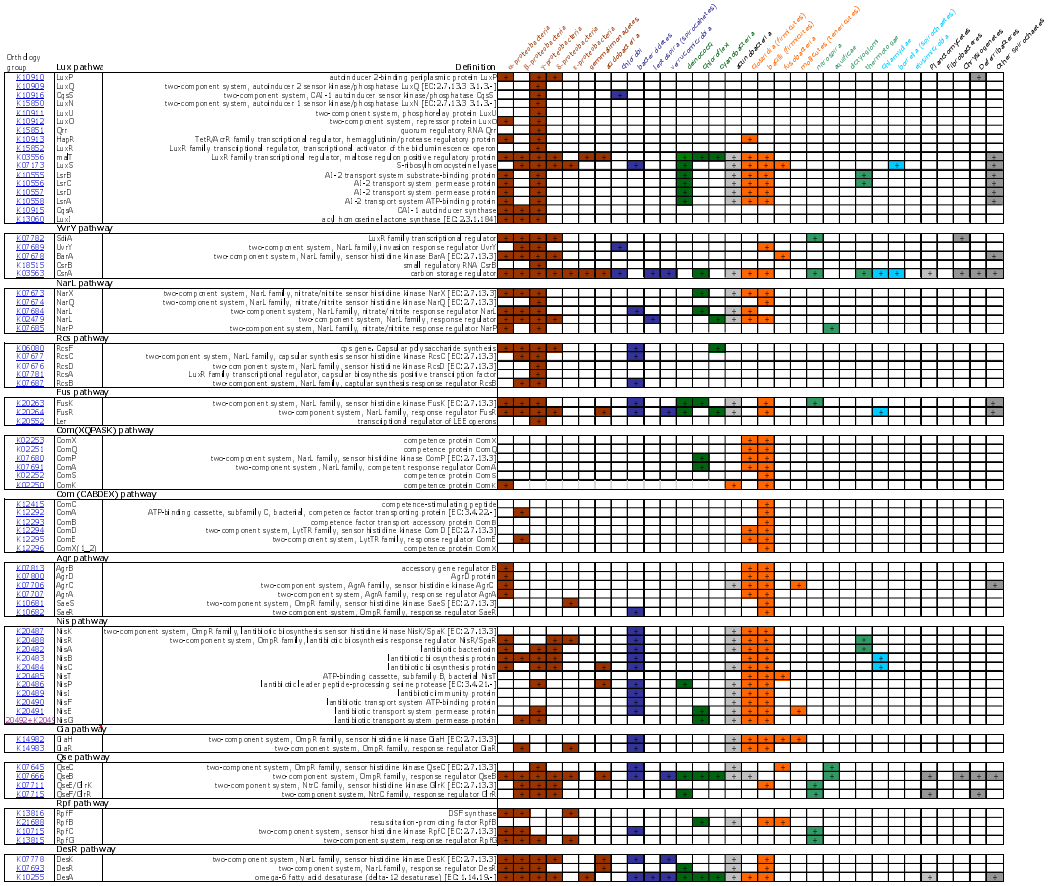


Figure 2: The number of bacterial groups in which each protein is found. The different colors indicate the family of each HK and RR in our study; proteins involved in the same pathway are indicated with a lighter shade of the same color. Blue: Lux family, Green: NarL family, Yellow: LytTR family, Orange: OmpR family, Purple: NtrC family, Grey: "Other", Black: Des pathway.

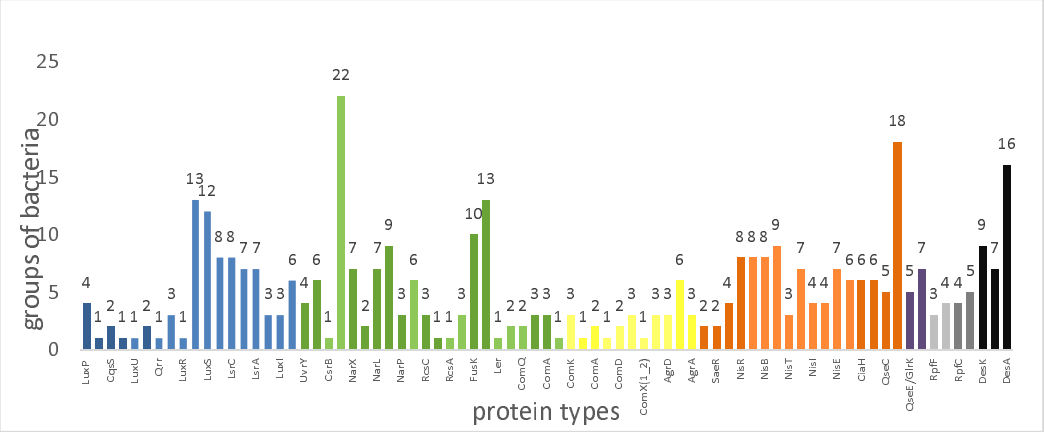


Figure 3: An overview of the position of the conserved domains in the HKs and RRs of this study. The similar colors of domains indicate their presence in a common clan of the Pfam database. The pathway drawings are from KEGG PATHWAY map ko02020, edited with Microsoft Draw.

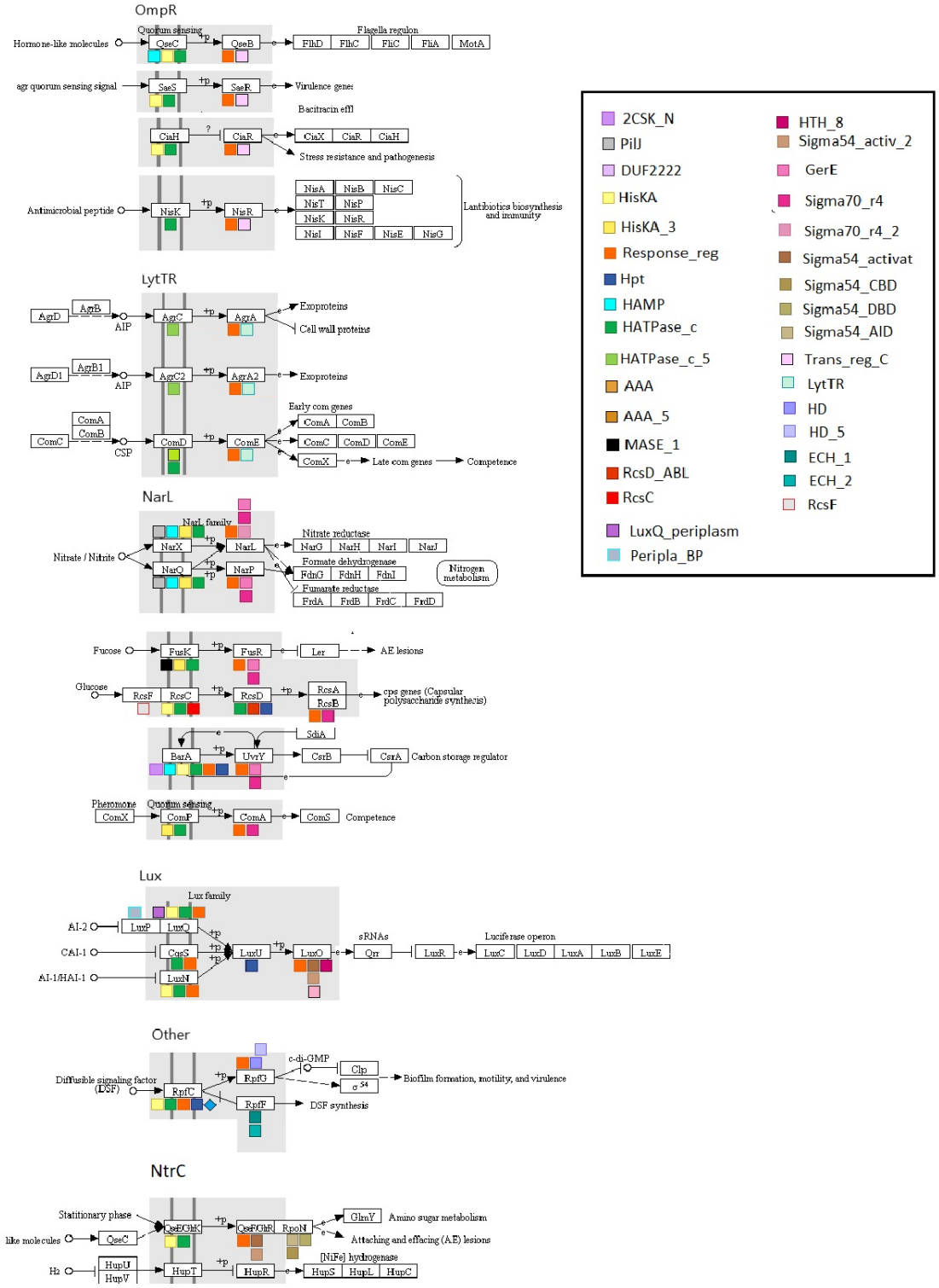


Figure 4: Phylogenetic analysis of the NarL family. (A) Histidine Kinases. (B) Response Regulators. Both trees reconstructed using PhyML. The color of the bullets indicates the bootstrap value of the nodes. White: 50% -80%, Grey: 80%-95%, Black: over 95%

Figure 5: Phylogenetic analysis of the OmpR family. (A) Histidine Kinases. (B) Response Regulators. Both trees reconstructed using PhyML. The color of the bullets indicates the bootstrap value of the nodes, as in Figure 4.

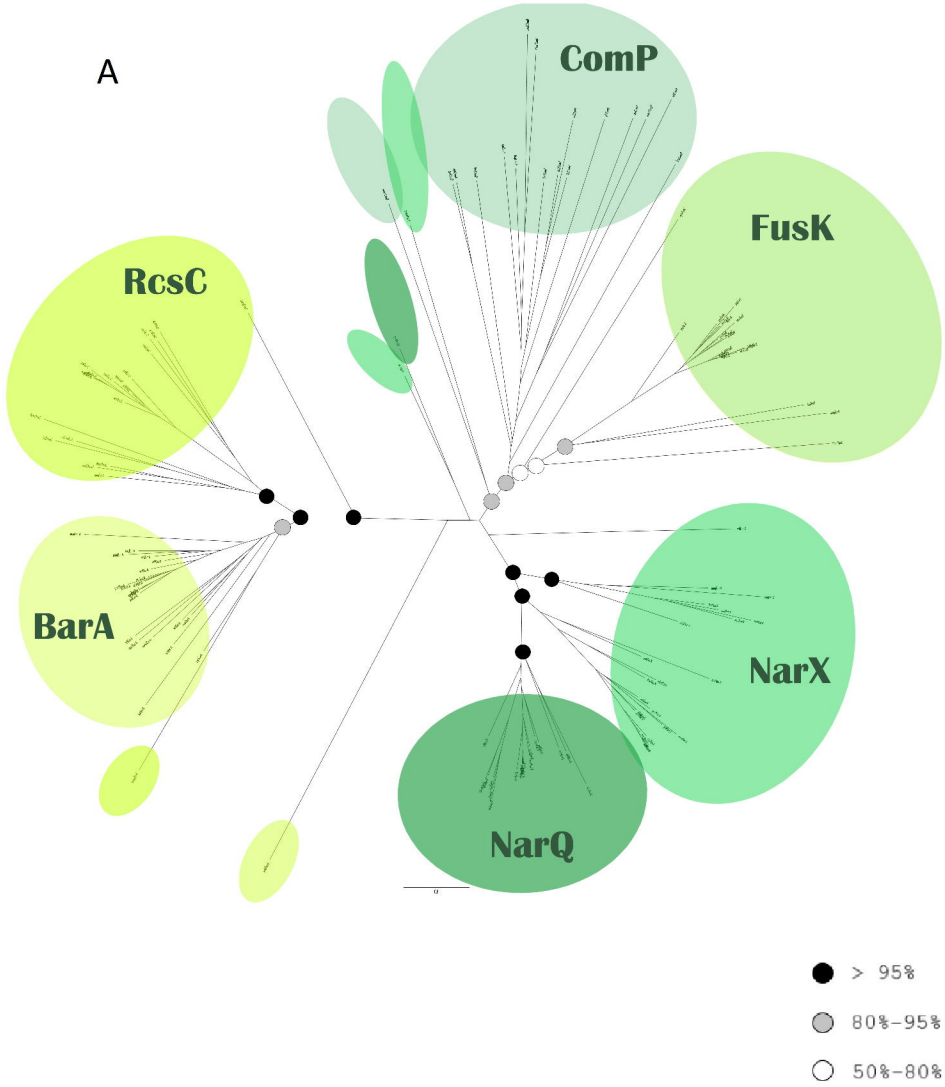
Figure 6: Phylogenetic analysis of the Lux family histidine kinases. Tree reconstructed using PhyML. The color of the bullets indicates the bootstrap value of the nodes, as in Figure 4.

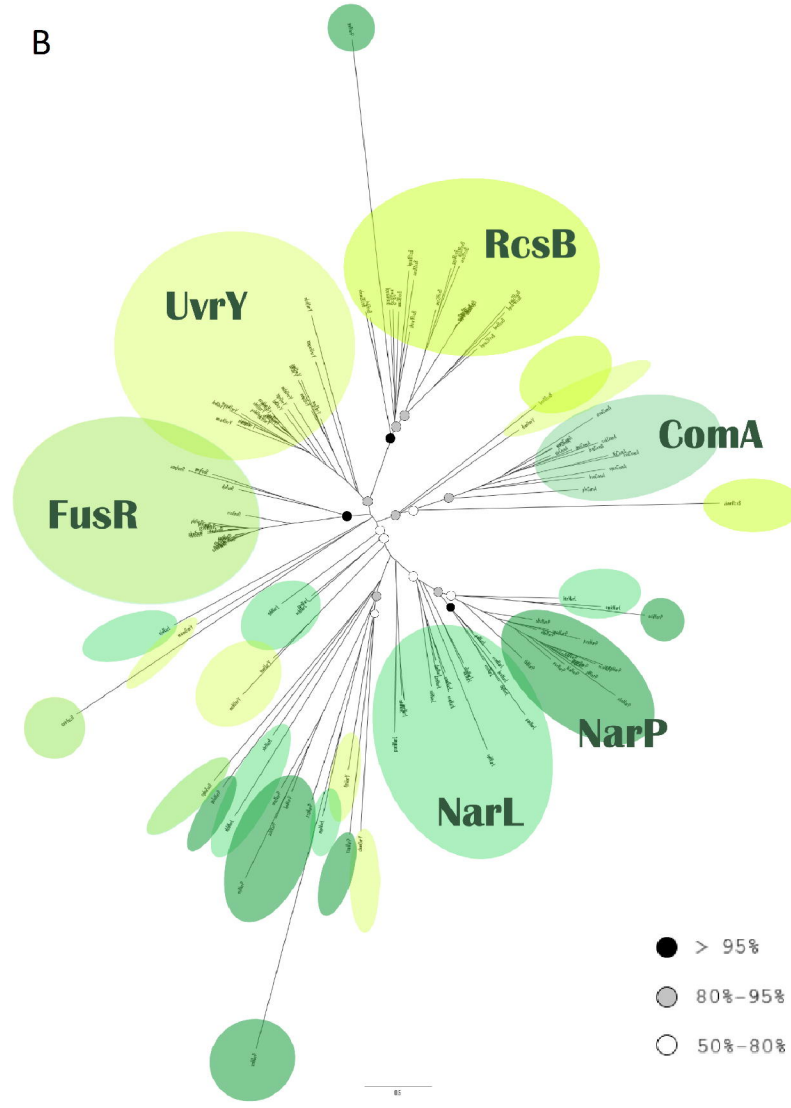
Figure 7: Phylogenetic analysis of the LytR family. (A) Histidine Kinases. (B) Response Regulators. Both trees reconstructed using PhyML. The color of the bullets indicates the bootstrap value of the nodes, as in Figure 4.

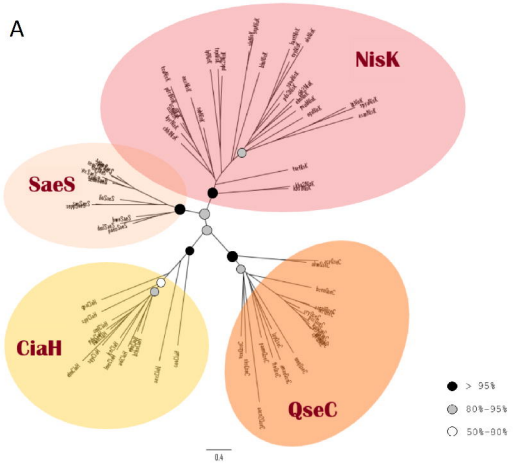
Figure 8: Phylogenetic analysis of all histidine kinase families. Tree reconstructed using RaxML. The bullets indicate the bootstrap value of the nodes, as in Figure 4.

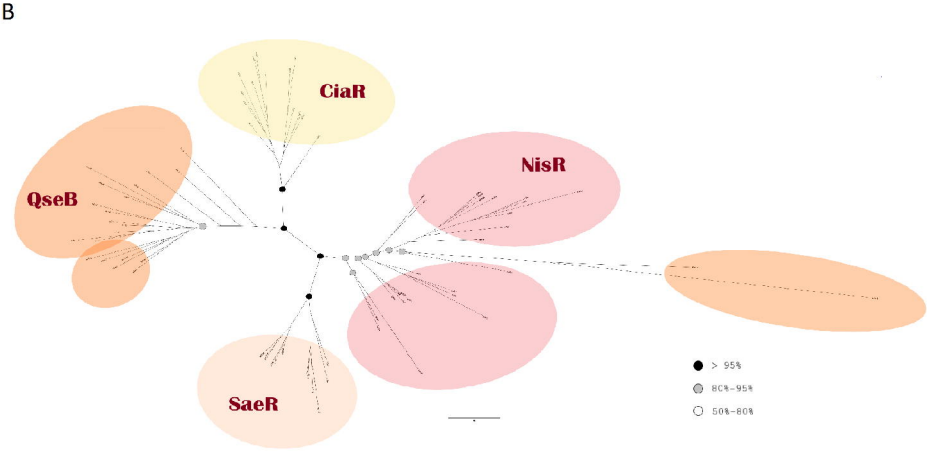
Figure 9: Phylogenetic analysis of all response regulator families. Tree reconstructed using PhyML. The bullets indicate the bootstrap value of the nodes, as in Figure 4.

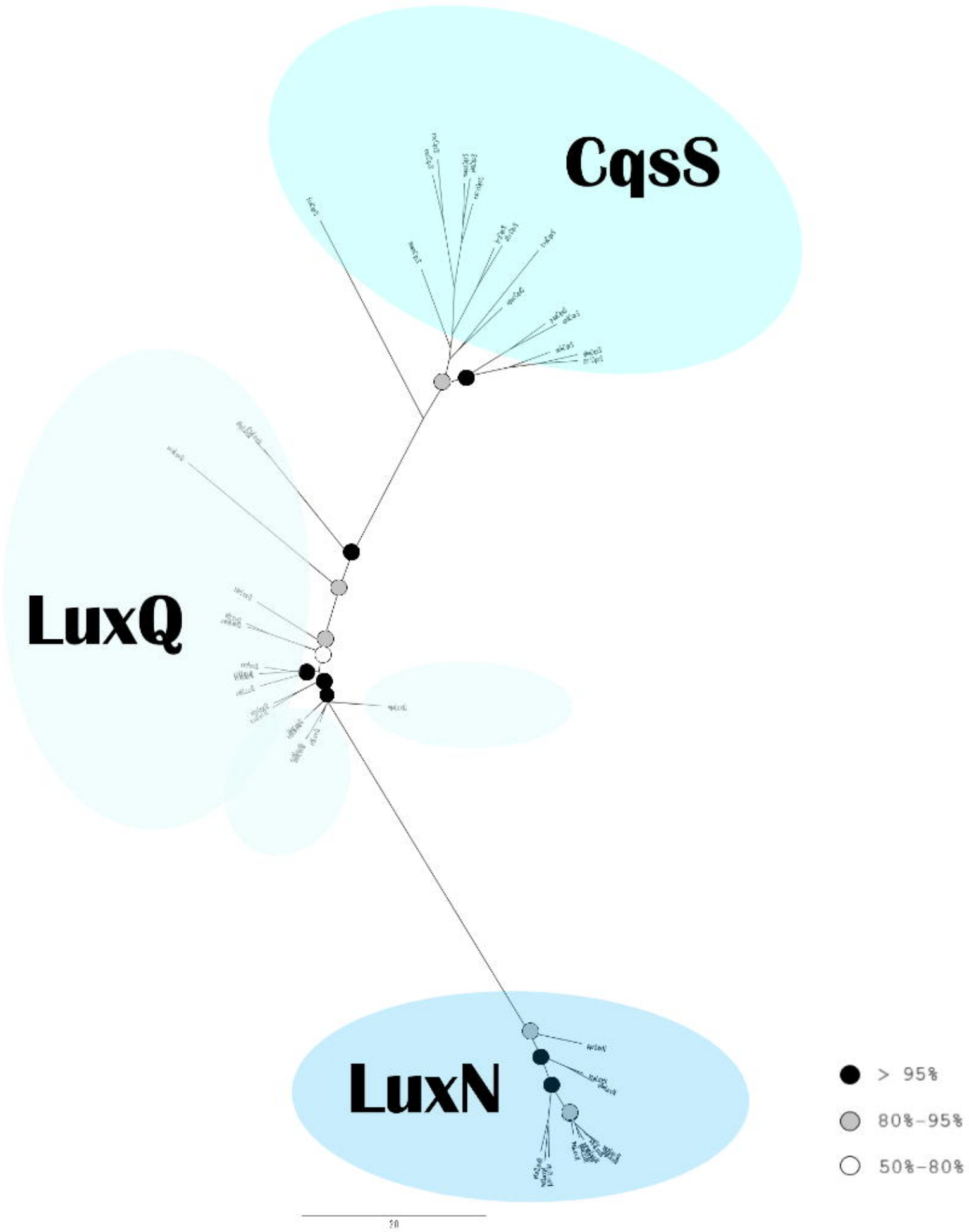
Figure 10: Phylogenetic analysis of all response regulator families. Tree reconstructed using RaxML. The bullets indicate the bootstrap value of the nodes, as in Figure 4.

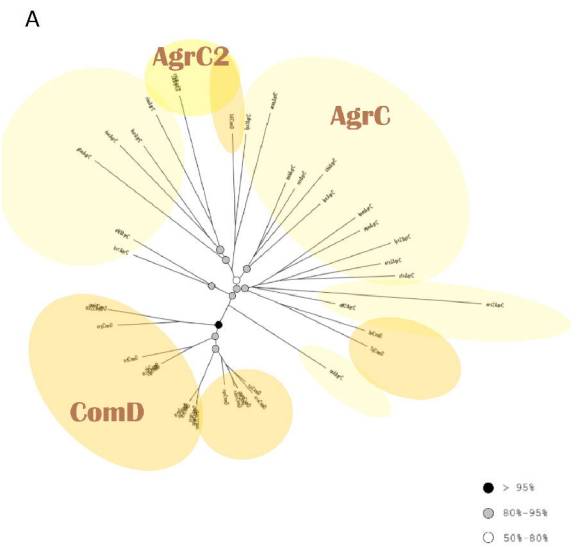

A
ComP
FusK
RcsC
BarA
NarX
NarQ
> 95%
80%–95%
50%–80%

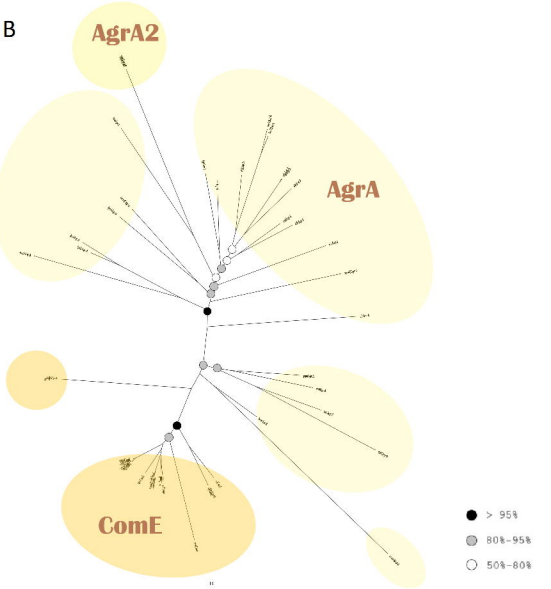

B
RcsB
UvrY
ComA
FusR
NarP
NarL
> 95%
80%–95%
50%–80%

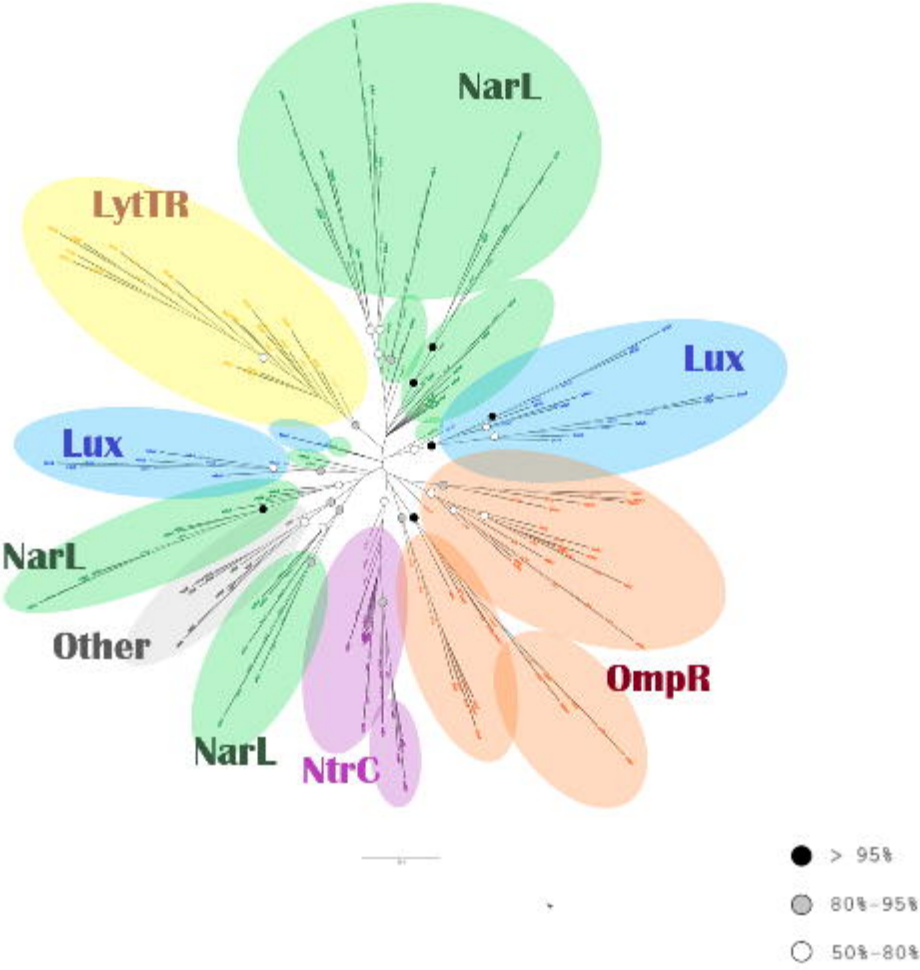

A
NisK
SaeS
CiaH
QseC
> 95%
80%-95%
50%-80%
0.4

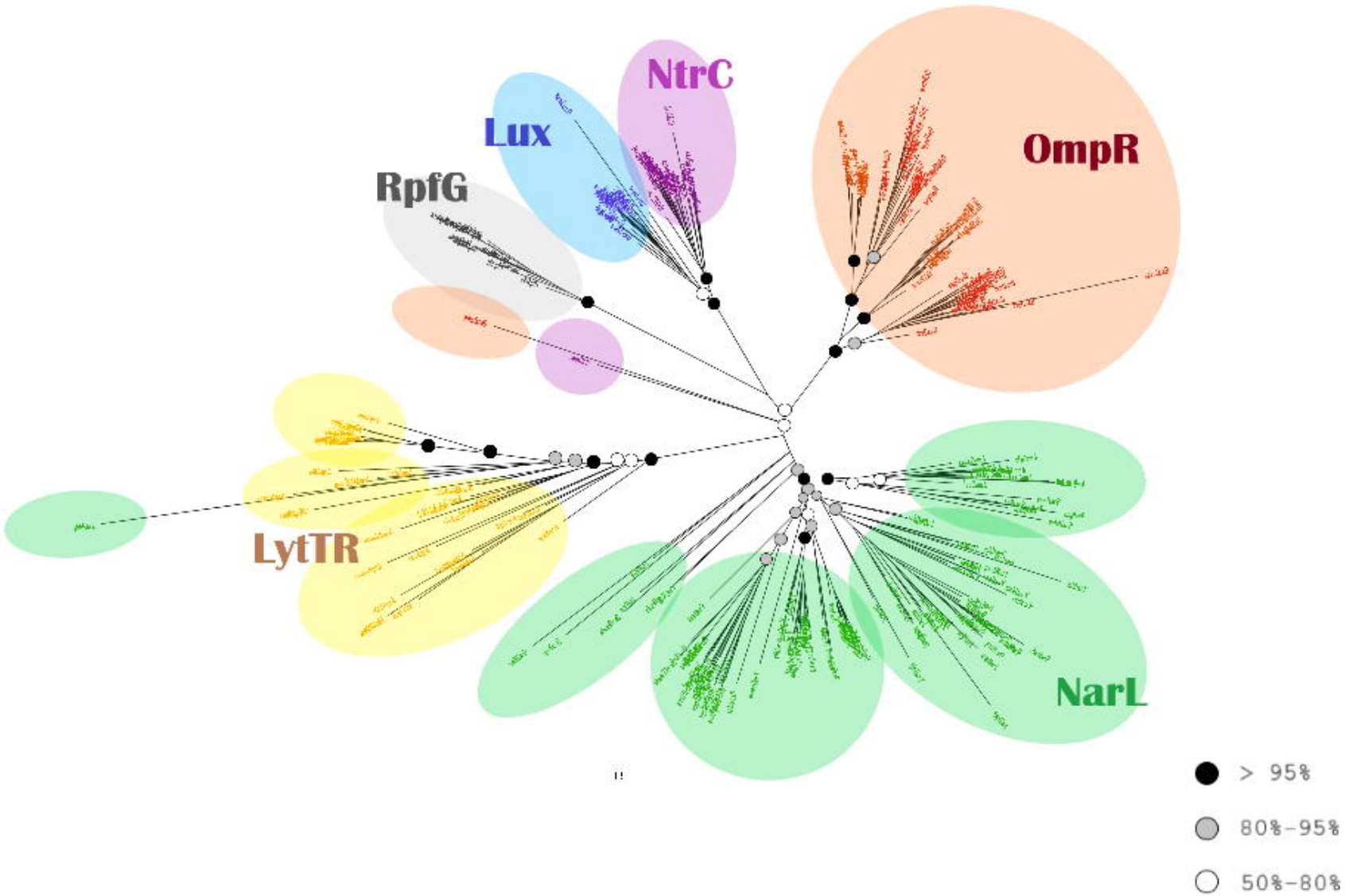

B
CiaR
QseB
NisR
SaeR
> 95%
80%-95%
50%-80%

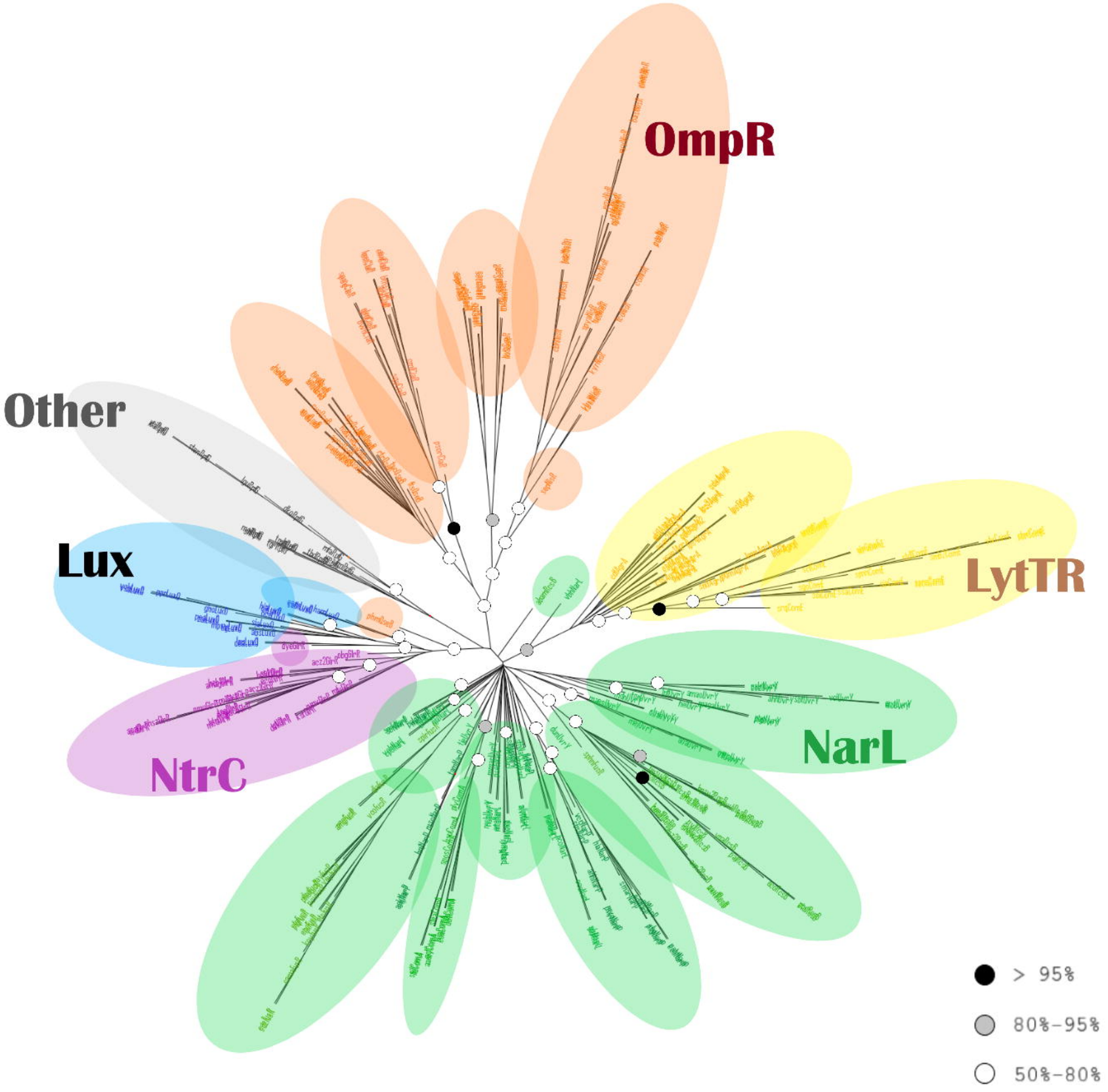

CqsS
LuxQ
LuxN
> 95%
80%–95%
50%–80%
10


A
AgrC2
AgrC
ComD
> 95%
80%-95%
50%-80%


B
AgrA2
AgrA
ComE
> 95%
80%-95%
50%-80%


NarL
LytTR
Lux
Lux
NarL
Other
OmpR
NarL
NtrC
> 95%
80%-95%
50%-80%


NtrC
Lux
RpfG
OmpR
LytTR
NarL
> 95%
80%–95%
50%–80%


OmpR
Other
Lux
LytTR
NarL
NtrC
> 95%
80%–95%
50%–80%
200.0