Evolution of two-component quorum sensing systems

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 (usually membrane-bound) sensor histidine kinase (HK) and a cognate response regulator (RR). 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 HKs and RRs that are part of TCS and examined their conservation across microbial taxa. We compared the combinations of the highly conserved domains in the different families of receptors and response regulators using the SMART and KEGG databases, and we also carried out phylogenetic analyses for each family, and all families together. The distribution of the different QS systems across taxa, indicates flexibility in HK/RR pairing, and highlights the need for further study of the most abundant systems. For both the QS receptors and the response regulators, our analysis indicates close evolutionary relationships between certain families, highlighting a common evolutionary history which can inform future applications, such as the design of novel inhibitors for pathogenic QS systems.

Keywords

Quorum sensing
Two Component Systems
Histidine Kinase
Response regulator
Conserved domains
Molecular Evolution
Phylogenetics

42 **Introduction**

43 The mechanism and importance of quorum sensing

44 Quorum sensing is a mechanism responsible for regulating a variety of group behaviors
45 and interactions in bacteria, including pathogenicity, biofilm formation, mobility,
46 sporulation, conjugal plasmid transfer, bioluminescence, resistance to antibiotics, and
47 antibiotics production (Bandara et al., 2012; Ng and Bassler, 2009). All quorum sensing
48 systems, despite in their molecular mechanisms and regulatory components, regulate their
49 target genes in a population-density dependent manner. An extracellular molecule
50 (autoinducer, AI) is secreted by a group of bacteria and when the number of the bacteria
51 rises, so does the number of these molecules produced in the surrounding area. When the
52 density of the AIs overcomes a specific threshold, they bind to a receptor protein, either on
53 the membrane or in the cytoplasm of the cell. The detection of the AIs activates a signal
54 transduction cascade, altering the gene expression of the cells all over the population (Ng
55 and Bassler, 2009).

56 The quorum sensing mechanism is different between Gram-positive and Gram-
57 negative bacteria. Gram-positive bacteria use modified oligopeptides as AIs (autoinducing
58 peptides, AIPs), which bind to membrane-bound receptors - Histidine Kinases (HKs) - when
59 they exceed a density threshold. This leads to the phosphorylation of the HK protein and its
60 cognate response regulator (RR), which then regulates the target genes. The AIPs can also be
61 transported into the cell and interact with cytoplasmic AIP receptors, which act as
62 transcription regulators to regulate gene expression (Ng and Bassler, 2009; Rutherford and
63 Bassler, 2012). Gram-negative bacteria use AHLs (acyl-homoserine-lactones) or SAM-
64 products (S-adenosylmethionine)(Papenfort and Bassler, 2016). A second group of AIs are
65 DPD (4,5-dihydroxy-2,3-pentanedione) derived molecules, collectively known as AI-2
66 (Bandara et al., 2012). Those small molecules can diffuse through the cell membrane and
67 bind to LuxR- type cytoplasmic receptors, which function as transcription factors. However,
68 in some Gram negative bacteria the AIs are detected by Histidine Kinases and the process is
69 similar to that in the Gram positive bacteria (Rutherford and Bassler, 2012). Other inducer
70 molecules taking part in quorum sensing include protein molecules that are not described as
71 AIPs, pheromones, sulfate and phosphate, fucose, oxygen, quinolones, γ-aminobutiric acid
72 (https://www.genome.jp/kegg-bin/show_pathway?map02024) or even adrenaline and
73 noradrenaline produced by the gut bacteria in humans, known as AI-3 (Reading and
74 Sperandio, 2006).

75 Since quorum sensing affects many bacterial processes, it is an interesting subject of
76 study which can have numerous applications, most important of which is the control of
77 pathogenicity via QS inhibition, given the increasing spread of antibiotic-resistant bacteria. It
78 has been suggested that targeting QS imposes less selective pressure, since it does not kill
79 bacteria but only hinders the production of virulence factors (Whiteley et al., 2017). QS
80 inhibition (Quorum Quenching, QQ) can be applied to fields such as hospital infections,
81 phytopathogen control in agriculture and for 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 which 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

118 separate families (Supplementary Fig. S1); however, the criterion for this classification is not
119 obvious. In previous studies, the HKs and RRs were classified based on their conserved
120 domains. For example, the HKs were divided into 11 subfamilies, based on multiple
121 alignment of 348 HKs and comparison of the "Homology boxes", i.e. groups of highly
122 conserved amino acids, which are assumed to be important in substrate binding, catalysis or
123 the structure of the protein: the H- and X-boxes are parts of the HisKA domain, and the N-,D-
124 F-and G-boxes are part of the HATPase_c domain; the D- box is also part of the REC domain
125 (Grebe and Stock, 1999). Alterntively, given the structure of the proteins and more
126 specifically their domain organization, the HKs can be divided into 4 groups: 3 groups with
127 proteins containing the intracellular domains PAS, HAMP or GAF, respectively, and a group
128 with none of these domains (Capra and Laub, 2012). In a more recent study, the proteins
129 were separated into 5 types (Types I-IV and the chemosensor HKs group, CheA) based on the
130 inferred phylogenetic tree using the whole genome of 22 bacterial species and 4 archaea.
131 More specifically, the criterion for their division was the H-box (HisKA) and its secondary
132 structure, and the kinase domain (ATPase) (Forst and Kim, 2015).
133 Based on the DNA-binding domains (Effector domain) of RRs, they were classified into
134 three subfamilies: OmpR, Fix J and NtrC (Hakenbeck and Stock, 1996), or more recently into
135 5 groups (Capra and Laub, 2012). By comparing the Receiver domains (REC), the RRs were
136 classified into 8 subfamilies (R_A-R_H) (Grebe and Stock, 1999), however, it has been proposed
137 that the comparison of the RRs for their classification should be based only on the Effector
138 domain, due to the fact that the Receiver domains of different RRs are highly conserved and
139 therefore not very informative for classification (Galperin, 2006).

140 Evolution of TCS

141 It is suggested that the more variable the environment, the higher the number of the
142 TCS of the organisms living in it. The genes of the HKs and their cognate RRs are usually
143 found on the same operon. The various TCS pathways are the result of lateral gene transfer
144 and/or duplication. The duplication can happen to all of the genes of the operon or to one of
145 them; in the second case, the resulting TCS pathway consists of more than one HK and
146 responds to more than one extracellular signal, or it includes more than one RR and gives
147 multiple responses for a specific signal. Also, due to the modular nature of HKs, domain
148 shuffling can lead to new HKs and the fusion of the HK and RR genes of an operon can create
149 hybrid HKs. The function of TCS pathways relies on molecular recognition and consequently
150 the amino acids have extensively co-evolved as a means to prevent the disruption of
151 signaling by mutations. Co-evolution also prevents crosstalk with other pathways.
152 Concerning the origin of HKs, it has been suggested that they come from ATPases of the
153 GHKL superfamily: Hsp90, the mismatch repair protein MutL or type II topoisomerases
154 (Capra and Laub, 2012; Koretke et al., 2000). On the other hand, the origin of the RRs
155 remains unclear beyond a general structural similarity to P-loop NTPases (Koretke et al.,
156 2000). The evolution of RRs can be the result of changes in the DNA-binding or RNA-
157 polymerase interaction sites; it includes lateral gene transfer and gene duplication, while

158 domain shuffling and rearrangement have also led to new forms of RRs (Capra and Laub,
159 2012).

160 Evolution of QS

161 Most of the phylogenetic studies of QS have mainly focused on the produced AIs or on
162 specific pathways, mainly LuxI/LuxR and LuxS/LuxQ. A phylogenetic analysis for the proteins
163 LuxI, LuxR and LuxS, showed that these proteins are ancient in many bacterial species and
164 that they appeared very early in the evolutionary path of the bacteria. It was also found that
165 in most cases the gene pairs of the inducer and its cognate response regulator are located
166 next to each other on the chromosomes and they preserve their pairwise function, a sign of
167 common evolutionary history (Lerat and Moran, 2004). The existence of homologous
168 proteins in some genomes is representative of horizontal transfer and duplication of the
169 genes, both of which can alter the regulation of different gene targets (Lerat and Moran,
170 2004). In a study of the ComQXPA pathway, the phylogenetic tree of HK proteins in 60
171 firmicute genomes containing the comQXPA locus clustered all the ComP proteins from
172 various organisms together, instead of gathering them with HKs of the same organism.
173 Therefore, it concluded that ComP evolved from the other HKs before the appearance of the
174 modern bacterial species and thus the ComQXPA pathways have an ancient origin (Dogsa et
175 al., 2014).
176 In a study of the Vibrionaceae family, it was found that the various inducers (AHLs) do
177 not show any correlation with the species' geographical distribution. This result indicates
178 that AHLs had a worldwide distribution throughout their evolution, since there was no
179 identification of specific AHL to the environment of the organism which produces it
180 (Rasmussen et al., 2014). Another survey, focusing on thermophilic bacteria, showed that
181 they use AIs-2 for QS communication. The phylogenetic tree comparison of LuxS and the 16S
182 rRNA of thermophiles and mesophiles show that LuxS (which produces the autoinducer) of
183 mesophilic bacteria may originate from thermophiles. Also, the LuxS proteins of
184 thermophilic bacteria within a phylum were evolutionarily closer compared to LuxS of
185 different phyla (Kaur et al., 2018).

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

193

Methods

Distribution of QS 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"). We checked for the presence of the protein subunits of different QS pathways in selected groups of bacteria, including 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 on it, by searching in the "motifs" section in the KEGG GENES database. We also used SMART (Simple Modular Architecture Research Tool, http://smart.embl-heidelberg.de/) to identify additional domains in our selected amino acid sequences not shown in KEGG. Both SMART and KEGG also give the E-value of each conserved domain; the domains 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 SMART. 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 function. 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. Accession numbers and abbreviations for all sequences used are in the Supplementary dataset.

229 Phylogenetic analysis

230 Multiple alignments of the amino acid sequences were created using the MAFFT
231 (Multiple Alignment using Fast Fourier Transform, https://www.ebi.ac.uk/Tools/msa/mafft/)
232 (Madeira et al., 2019) and MUSCLE (MUltiple Sequence Comparison by Log- Expectation,
233 https://www.ebi.ac.uk/Tools/msa/muscle/) (Edgar, 2004), alignment programs, with default
234 settings, for:
235 • each HK family
236 • all the HKs
237 • each RR family
238 • all the RRs
239 The substitution model for each multiple alignment was inferred via the ProtTest3.4
240 program (https://github.com/ddarriba/prottest3/releases) (Pearson, 2013). In order to
241 compute the likelihood scores, the following settings were used:
242 – Substitution matrices to test: VT, Blosum62, JTT, Dayhoff, WAG
243 – Rate variation: I, G, I+G
244 – Amino acid frequencies: unchecked empirical
245 The results were evaluated by the Akaike information criterion (AIC). More specifically, the
246 substitution model for the MUSCLE multiple alignments of the HKs and RRs was VT and for
247 the MAFFT alignments was VT for HKs and WAG for RRs.
248 Phylogenetic reconstruction for each protein family was done with PhyML at
249 http://www.phylogeny.fr/ (also available at https://ngphylogeny.fr/) (Dereeper et al., 2008).
250 We used the "A la Carte" function, with the following settings: "remove gaps in alignment"
251 (alignment curation), Maximum Likelihood method (PhyML), SH-like Approximate
252 Likelihood-Ratio Test (aLRT) as the statistical test for branch support and 100 bootstraps. For
253 each multiple alignment we chose the WAG substitution model as the best one according to
254 the ProtTest results (WAG, JTT and Dayhoff were the given options in the PhyML settings).
255 The phylogenetic trees for all the protein families together (2 trees: HKs and RRs) were
256 constructed in the CIPRES Science Gateway (https://www.phylo.org/portal2/login.action),
257 using RAxML (Randomized Axelerated Maximum Likelihood): *RAxML-HPC BlackBox (8.2.12) -*
258 *Phylogenetic tree inference using maximum likelihood/rapid bootstrapping on XSEDE*
259 (Stamatakis, 2014). Before entering the multiple alignments into the program, we removed
260 any positions not conserved in >50% of the sequences. For this purpose, we used Jalview
261 (https://www.jalview.org/) to detect those positions and then the Mesquite program to do
262 the trimming (http://www.mesquiteproject.org/). The proper substitution model was
263 chosen for each multiple alignment and the bootstrap number was set to 100.
264 The multiple alignment for all the HK proteins consisted of 347 sequences and had a
265 total length of 2410 amino acids using MAFFT or 1919 using MUSCLE. In both alignments,
266 there were several clusters of conserved residues mixed with gaps or non-conserved
267 residues towards the N-terminal region of the alignment (41-867) (34-1073, 1295-1781) and
268 near the C-terminal part (949-1411). The trimmed MAFFT alignment had 520 amino acids

269 (Supplementary Fig. S2), while the trimmed MUSCLE alignment had 455 amino acids
270 (Supplementary Fig. S3). The multiple alignment for all the RR proteins consisted of 293
271 sequences and had a total length of 751 amino acids using MAFFT or 670 using MUSCLE.
272 Both the MUSCLE and MAFFT alignments contained conserved clusters alternating with gaps
273 or non-conserved residues (65-224, 287-528). Trimming resulted in an alignment of 229
274 amino acids using MAFFT (Supplementary Fig. S4), and 274 using MUSCLE (Supplementary
275 Fig. S5).
276 All the phylogenetic trees were visualized with the Figtree program
277 (https://github.com/rambaut/figtree/releases). The results were saved as PNG images and
278 were further edited with Microsoft Draw. Nodes were labelled as white, grey or black
279 bullets, if their bootstrap value was 50%-80%, 80%-95% and 95%-100% respectively. The raw
280 and masked alignments, as well as the phylogenetic trees in nexus format are available in
281 the Supplementary dataset. The alignments for each family and all families together are
282 available in the files "HK alignments" and "RR alignments". The alignments of all HKs and all
283 RRs are presented as they were trimmed in Mesquite: in the added taxon, "C" and "I"
284 represent the removed and included positions. The trees in nexus format are available in the
285 "HKs trees" and "RRs trees" files.

286

287 **Results**

288 Distributions of QS proteins

289 The distribution of each protein type across bacterial taxa is demonstrated in Fig. 1;
290 bacterial groups are colored based on (Jun et al., 2010). All pathways are fully present in at
291 least one bacterial group, but many groups of bacteria have some of the subunits of each
292 pathway. All Lux pathways are fully present in γ-proteobacteria. The Lux system contains a
293 pathway of HKs and an RR, and the cytoplasmic Lsr proteins; both pathways are activated by
294 the AI produced by LuxS. Deinococci, actinobacteria, clostridia, bacilli, α- and γ-
295 proteobacteria, and Spirochaetes possess most or all proteins of the Lsr protein pathway.
296 The UvrY pathways (BarA/UvrY and SdiA) are fully present in γ-proteobacteria. Like in the
297 Lux pathway, there are two different types of receptors, an HK (BarA) and a cytoplasmic
298 protein (SdiA). Several proteins of one or both of the pathways are also found in α- β-and δ-
299 proteobacteria, chlorobi, bacilli, *Nitrospira*, Chrysiogenetes and Spirochaetes. Concerning
300 the NarL pathway, γ-proteobacteria have all proteins of the pathway; α- and β-
301 proteobacteria, chloroflexi, cyanobacteria, actinobacteria, clostridia and bacilli have one of
302 the receptors. Only γ-proteobacteria possess all proteins of the Rcs pathway; bacteroidetes
303 and β-proteobacteria contain the HK and RR of the pathway. All three proteins of the Fus
304 pathway are in γ-proteobacteria; however, the HK/RR pair is found also in α- and β-
305 proteobacteria, bacteroidetes, deinococci, actinobacteria and Spirochaetes. All proteins of
306 the Com(XQPASK) pathway are found in bacilli and most proteins are found in clostridia.
307 Also, the HK/RR pair ComP/ComA is present in deinococci. The proteins of Com(ABDEX) are

308 all found in bacilli, whereas the HK/RR pair ComD/ComE is present in clostridia. The Agr
309 pathway (both AgrC/AgrA and SaeS/SaeR) is fully present in bacilli, while α-proteobacteria
310 and clostridia contain the Agr proteins only. All the proteins of the Nis pathway are present
311 in both clostridia and bacilli, while bacteroidetes, chloroflexi and actinobacteria contain all
312 but two of the proteins. Other bacterial groups contain parts of the pathway but not the full
313 sensing system (they lack either one or both the HK and RR). Bacteroidetes, actinobacteria,
314 clostridia and bacilli contain both the CiaH and CiaR proteins. Concerning the Qse pathway,
315 γ-proteobacteria contain all of the proteins (QseC/QseB and QseE(GlrK)/QseF(GlrR)).
316 QseC/QseB are present in bacteroidetes, actinobacteria and aquificae.
317 QseE(GlrK)/QseF(GlrR) are found in β- and δ-proteobacteria and *Nitrospira*. No bacterial
318 group contains all the proteins of the Rpf pathway: Nitrospira, α- and β-proteobacteria
319 contain the HK/RR pair RpfC/RpfG. All three proteins of the DesR pathway are present in α-
320 ,β- and γ- proteobacteria, actinobacteria and bacilli. For details of all other bacterial groups
321 which contain only certain proteins of each pathway, please refer to Fig. 1.
322 Fig. 1 shows that the majority of the complete pathways are found in α-, β- and γ-
323 proteobacteria, clostridia and bacilli. A significant number of proteins are also found in δ-
324 and ε-proteobacteria, bacteroidetes and actinobacteria. Also, based on the number of
325 bacterial groups which contain proteins of each pathway, the most widespread proteins are
326 those of the UvrY, DesR, Fus, Nis pathways and the Lsr proteins of the Lux pathway. A more
327 imited distribution across taxa is seen for the proteins of the two Com pathways and the Rcs
328 and Agr pathways. Looking at each protein separately the most widespread proteins are
329 CsrA, QseB, DesA, FusR, malt and LuxS (Fig. 2). Interestingly, those belong to different
330 pathways (CsrA: UvrY, FusR: Fus, QseB: Qse, DesA: DesR, malt and LuxS: Lux pathway). The
331 Lux and UvrY pathways contain the most widespread as well as some of the most limited
332 protein types (LuxQ, Qrr and CsrB). There is also a difference in the distribution of the HKs
333 and their cognate RRs: In most cases, the RRs are found in more bacterial groups than their
334 cognate HKs (NarL, ComA, NisR, FusR, QseB, GlrR, and RpfG).

335 Conserved domains

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

342 <u>Conserved domains in HKs:</u> The length of the HKs varies from 406 (AgrC) to 949 (RcsC) amino
343 acids. The longest HKs are the hybrid ones (NarL family: BarA, RcsC, Lux family: LuxN, LuxQ
344 and CqsS, "Other": RpfC) which also contain more conserved domains. Differences in the
345 length of the sequences can also be found within the proteins of the same family. The HKs
346 LuxQ and RcsC function as a complex with another protein (LuxP and RcsF, respectively).

347 Most HKs contain 1-3 conserved domains and up to 6 (Fig. 3). Comparing the number of the
348 conserved domains, the members of the families OmprR, LytTR and NtrC (GlrK) have the
349 lowest number of conserved domains (1 or 2). In contrast, the HKs of the NarL, Lux and
350 "Other" (RfC) families have more than 2 conserved domains (Fig.3).

351 The common trait is the existence of either HATPase_c or HATPase_c_5, which are also
352 the only conserved domains in the protein NisK and the proteins of the LytTR family.
353 According to NCBI-CDD, HATPase domains (Histidine kinase-like ATPase domains), also
354 referred to as GHKL ATPase domains, are found in ATP-binding proteins, including histidine
355 kinases. HAMP (Histidine kinase, Adenylyl cyclase, Methyl-accepting protein, and
356 Phosphatase domain) is the second most frequent domain, found in HKs of the NarL family
357 (NarX, NarQ, BarA). It relays various input signals but no common mechanism of HAMP has
358 yet been established (NCBI-CDD). According to InterPro, one or more copies of the domain
359 can be found in association with other domains, including HisKA. It has been suggested that
360 its role is the regulation of the phosphorylation or methylation of homodimeric receptors by
361 transmitting the conformational changes in periplasmic ligand-binding domains to
362 cytoplasmic signalling kinase and methyl-acceptor domains. According to a study of Koretke
363 et al., HAMP is the "linker region" and it is widely conserved in HKs as well as in
364 chemoreceptors, bacterial nucleotidyl cyclases, and phosphatases. Its wide distribution is
365 notable, as about a quarter of the HKs in the SMART database contain this domain and about
366 half of all HAMP domains are found in HKs (Koretke et al., 2003). The Response_reg and Hpt
367 domains are located in the C-terminal region of hybrid HKs, but only Hpt can alternatively be
368 found as an independent protein (within the Rcs pathway). The Response_reg (Response
369 regulator receiver domain or phosphoacceptor receiver-REC) receives the signal from the HK
370 and Hpt (histidine-containing phosphotransfer domain), which mediates phosphotransfer
371 reactions in Multi-step TCS signaling systems. Concerning the N-terminal region, only few
372 proteins contain conserved domains, most of which belong to the NarL family (QseC, NarX,
373 NarQ, FusK). These are Cache-like sensory domains (2CSK_N, DUF2222 and LuxQ-periplasm),
374 PilJ (Type IV pili methyl-accepting chemotaxis transducer), Peripla_BP_4
375 (Periplasmic_Binding domain) and MASE1 (PFAM: predicted integral membrane sensory
376 domain in HKs, diguanylate cyclases and other bacterial signaling proteins).

377 <u>Conserved domains in RRs:</u> The average length of RRs is significantly shorter that the HKs. It
378 varies from 207 to 455 amino acids. Most RRs contain 2 conserved domains except for LuxO,
379 which has 3 domains (Fig. 3). The N-terminal regions always contain the Response_reg
380 domain, which is a common structure trait for all the RRs. The conserved domain on the C-
381 terminal region is different in each RR family. The only exception is the protein GlrR (NtrC
382 family), which has the same conserved domain as LuxO (Lux family). An overview of the
383 conserved domains of both HKs and RRs is present in Table 2. HATPase_c is the most
384 common conserved domain, as it is found in 15 protein types. HisKA is the second most
385 common, with 9 occurrences.

386 Concerning the variety of the Effector domains, the majority has a DNA-binding

387 activity: HTH_8, Sigma70_r4, Sigma70_r4_2, GerE, and Sigma54_DBD (DNA-Binding Domain)
388 are HTH (Helix-Turn-helix) domains. Although Sigma54_CBD (Core Binding Domain) is also
389 considered as an HTH domain, it interacts directly with the core RNA polymerase, forming an
390 enhancer-dependent holoenzyme. The centre of this domain is slightly similar to a HTH
391 motif, which may represent a DNA-binding domain. Also, some domains are P-loop-NTPases:
392 Sigma54_activat, Sigma54_activ_2 interact with the sigma-54 factor of RNA polymerase;
393 AAA and AAA_5 (ATPases Associated to a variety of cellular Activities) are described as a
394 dynein-related subfamily and although their role is yet unclear, the AAA+ superfamily to
395 which AAA ATPases belong, is associated with chaperone-like functions. Sigma_54_AID
396 (Sigma_54_Activator interacting domain) is not classified into any clan. It is necessary for the
397 interaction of Sigma54 RNAP holoenzyme with the activator and it can also inhibit
398 transcription initiation prior to interaction with the activator.

399 Phylogenetic trees by family

400 Given the families formed by the HKs and their cognate RRs, we constructed four
401 phylogenetic trees for the HKs and three for the RRs. The HKs GlrK (NtrC family), RpfC
402 ("Other"), their cognate RRs GlrR and RpfG, as well as LuxO are the only members of their
403 family, hence no phylogenetic tree was inferred. Unless otherwise stated, only well-
404 supported clades, with >90% bootstrap support, are discussed below. Phylogenetic
405 reconstructions were done with PhyML using MAFFT alignments, but the major differences
406 between the trees created from the MAFFT and MUSCLE alignments (Supplementary Fig. S6-
407 S9) are also mentioned, wherever they apply.

408 *The NarL family*

409 The HKs of the NarL family are organized in distinct clades in the PhyML tree (Fig. 4A).
410 Each clade consists of amino acid sequences of the same protein type, supported by high
411 bootstrap values. The same structure is observed in the phylogenetic tree of their cognate
412 RRs (Fig. 4B); the clades of the RR types are also supported by good bootstrap values, but
413 most of them are below 95%. Although both trees have some amino acid sequences away
414 from the clade of their protein type (e.g. marrComP, mdnBarA, paejNarQ and anxNarX), the
415 phenomenon is more prevalent in the RRs. Notably, the structure of the phylogenetic tree
416 shows three pairs of HK types (RcsC-BarA, ComP-FusK and NarQ-NarX), which group closer to
417 each other than to the other protein types. This pattern is also seen for their cognate RRs.

418 The clade of the RcsC proteins is closer to BarA than the rest of the protein types. On
419 the other hand, their cognate RRs (RcsB and UvrY) are apart in the PhyML tree: although the
420 UvrY proteins do not create a distinct pair with any other RR type, their clade is closer to
421 FusR and the common clade of NarL and NarP. The RcsB RRs group together, except for the
422 sequence abamRcsB. The UvrY proteins also form a clade (86% bootstrap support, excluding
423 the first node of maesUvrY, nppNarL and ssilNarP sequences); outside of it sbxUvrY, dunUvrY
424 and cbaeUvrY fall in the RcsB clade, mmarUvrY falls in the NarL clade, and tjeUvrY in the

425 NarP clade. A comparison with the same PhyML tree from the MUSCLE alignment
426 (Supplementary Fig. S6B) shows that although the distribution of the NarL and NarP RRs is in
427 multiple dispersed clades in the tree, the UvrY and RcsB proteins clearly form a common
428 clade (87% bootstrap value). Also, the ComA clade is closer to FusR than to RcsB.

429 The ComP clade has a 50% bootstrap value, excluding kbsComP and marrComP, and it
430 groups with the FusK clade (72% bootstrap value). In contrast, the clades of their cognate
431 RRs, ComA and FusR, do not group together. The ComA sequences are not directly related to
432 any other clade but they are close to RcsB in the tree.

433 The NarX proteins are separated in two main clades, both of which share a common
434 node with the NarQ clade; spswNarX branches out earlier. However, the distribution of their
435 cognate RRs NarL and NarP is more complicated. They are divided into several clades
436 (sometimes including both protein types), yet most of these clades are placed close to each
437 other and their common node has a 79% bootstrap value. In detail, the NarP sequences are
438 grouped into two distant main clades; a larger one, neighboring with the NarLs and a smaller
439 distant one. There are more sequences outside of these clades than is seen for their cognate
440 HK sequences, and from different species than in the HKs tree. The NarLs are gathered in
441 several small clades close to each other, most of which demonstrate bootstrap values of at
442 least 80%, but many NarL sequences are dispersed throughout the phylogenetic tree.

443 *The OmpR family*

444 All different HK types of the OmpR family are grouped into separate clades. None of
445 the HK sequences is misplaced in a clade of a different protein type, except spyaNisK, which
446 falls in the CiaH clade. Their cognate RRs follow a similar pattern, except for sequences
447 phmQseB and aprsQseB, which lie outside the QseB clade. There are two related groups in
448 both phylogenetic trees: CiaH - QseC and NisK – SaeS and their cognate RRs, CiaR - QseB and
449 NisR – SaeR (Fig. 5A and 5B). The majority of the CiaH sequences stem from a common node,
450 with soxCiaH and cooCiaH branching out earlier. The QseC sequences are distinctly
451 separated from the rest of the RR types and they share a common node with the CiaH clade.
452 The cognate RRs follow the same pattern: The clade of the QseB sequences (85% bootstrap
453 value) forms a common clade with the CiaR sequences. The NisK sequences are organized
454 into one main clade and they have a common node with all the SaeS proteins. The NisR
455 proteins also group together with the clade of the SaeR sequences.

456 *The Lux family HKs*

457 The sequences of each HK type of the Lux family (LuxN, LuxQ and CqsS) are grouped
458 into distinct clades (Fig. 6). The LuxN sequences have one common node. On the other side
459 of the tree, LuxQ create a common clade, although dhyLuxQ, dsaLuxQ, oceLuxQ are more
460 diverged. The CqsS sequences branch out at the center of the tree in 3 clades, with fsoCqsS
461 more diverged. On the contrary, most LuxQ branches in the PhyML tree of MUSCLE

462 (Supplementary Fig. S8) appear near the core of the tree; it seems that the clades CqsS and
463 LuxQ switch positions in the MAFFT and MUSCLE versions of the tree. As there is only one RR
464 type for all the Lux family HKs (LuxO), there is no phylogenetic tree for the RRs of the Lux
465 family.

466 *The LytTR family*

467 The different types of HKs of the LytTR family, as well as their cognate RRs, are clearly
468 separated from each other and both phylogenetic trees have a similar overall appearance
469 (Fig. 7A and 7B). The AgrC sequences form several clades, all of which stem from a common
470 node on the phylogenetic tree. The two AgrC2 sequences are found in a distinct clade within
471 the AgrC clade, closest to rimAgrC. Similar to AgrC HKs, the AgrA sequences do not form one
472 distinct clade, but smaller ones, gathered on one side of the phylogenetic tree. The AgrA2
473 sequences are found in a distinct clade near the AgrA clades closest to cdfAgrA (77%
474 bootstrap value). Most of the ComD sequences are gathered in one clade, apart from the
475 lhiComD, lsjComD and laliComD sequences, which are placed among AgrC sequences. The
476 ComE RRs create a main common clade, with the exception of an earlier smaller clade:
477 sequences std2ComE and smb1ComE branch early (58% bootstrap support) as do their
478 cognate HKs.

479

480 Phylogenetic trees of all the families together

481 *Phylogenetic analysis of all the HK families*

482 The two phylogenetic trees of all the HKs and RRs families together are discussed
483 below, based on the MAFFT alignment. Major differences with the phylogenetic trees based
484 on the MUSCLE alignments are also mentioned. In general, the members of each HK family
485 tend to form one main clade, although the bootstrap values are not always significant. The
486 clades of each HK protein type are also mostly monophyletic (Fig. 8). The sequences of the
487 LytTR family, as well as the RpfC ("Other") and GlrK HK types (NtrC family) are each gathered
488 in one clade, whereas the protein sequences of the Lux as well as the NarL family are found
489 in more than one clades. The following 3 groups of families are discernible: 1) LytTR and
490 NarL (NarQ, NarX, FusK and ComP), 2) OmpR and NtrC, 3) Lux, NarL (BarA, RcsC) and "Other"
491 (RpfC).

492 <u>NarL family:</u> The sequences of this family form three main clades: the RcsC clade, the BarA
493 clade (97% bootstrap value excluding the early branches) and a third -and larger- one which
494 includes the NarQ, NarX, FusK and ComP sequences (Fig. 8). Within this larger group, the
495 FusK sequences cluster together (56%) and create a common clade with the NarQ and NarP
496 clades. All the ComP sequences also derive from a common node, but with very low
497 bootstrap value. NarQ sequences form a well-supported main clade, while NarX forms two
498 main clades with 87% and 97% bootstrap support, respectively. Only one NarQ and three

499 NarX sequences fall outside their protein type clades. In general, the grouping of the clades
500 of the NarL family is very similar to the tree for the HKs of the NarL family alone (Fig. 4): the
501 BarA-RcsC, and NarX-NarQ pairs are recapitulated, and the only difference is the positions of
502 FusK, which is closer to ComP in Fig. 4, while it shares a common node with the NarQ/NarX
503 clade in Fig. 8.

504 OmpR family: The HK types of this family are separated into three main clades: one for QseC
505 and CiaH one clade for the SaeS and one for the NisK sequences (Fig. 8). The tree from the
506 MUSCLE alignment (Supplementary Fig. S10) recapitulates the exact structure of the
507 phylogenetic tree of the OmpR family alone (Fig. 5), as SaeS and NisK sequences create a
508 common clade. The clades of each protein type of the OmpR family are supported by high
509 bootstrap values, although their common node with the rest of the phylogenetic tree has
510 low bootstrap support.

511 Lux family: Two distinct clades and a smaller mixed one are formed by the HKs of the Lux
512 family: one clade consists of CqsS and LuxN sequences (with dhyLuxQ and dsaLuxQ at the
513 root of the clade) and neighbors with the OmpR, LytR and NarL clades and the other is made
514 of only LuxQ sequences near the BarA clade of the NarL family (Fig. 8). This pattern is also
515 seen in the tree based on the MUSCLE (Supplementary Fig. S10), and follows the groupings
516 of the phylogenetic tree of the Lux family alone (Fig. 6), as CqsS and LuxN are grouped
517 together, while the LuxQ sequences are placed away from the rest of the members of the
518 Lux family.

519 LytTR family: All the sequences of this family derive from a common node (56% bootstrap
520 support) (Fig. 8). The ComD sequences form one clade (88%) within the AgrC clade. The AgrC
521 sequences form many small clades, including the two AgrC2 sequences. The closest clade to
522 the LytTR family is the overall clade of the ComP-NarX-NarQ-FusK sequences of the NarL
523 family (62% bootstrap support).

524 NtrC family (GlrK proteins-QseE): All the sequences of the GlrK (QseE) HK are in one clade,
525 with 73% bootstrap support (Fig. 8). The closest clade to them is that of SaeS (of the OmpR
526 family), but with insignificant bootstrap support; in the tree based on the MUSCLE
527 alignment, the NtrC family groups more clearly with the OmpR family, and in particular the
528 common clade of QseC and CiaH, although again with very low boostrap support.

529 "Other" (RpfC sequence): All RpfC sequences are placed in one clade, with 74% bootstrap
530 value. The neighboring clade consists of RcsC sequences of the NarL family (Fig. 8), but the
531 bootstrap value for their common clade is very low.
532

533 *Phylogenetic analysis of all the RR families*

534 The RR families are distinctly separated from each other in the phylogenetic tree. Also, it is
535 clear that the sequences of each RR type within each family remain together, like in the
536 phylogenetic tree of each family. Very few sequences fall outside of the clade of their family
537 or RR type (Fig. 9). An overview of the tree reveals a closer relationship between the
538 following family groups: 1) the RpfG and NarL families, 2) the OmpR, NtrC and Lux families
539 and 3) the LytTR clade, which is closer to group 1). In the tree based on the MUSCLE
540 alignment (Supplementary Fig. S11), the position of the RpfG (Other) group is closer to NtrC
541 and Lux families; however, the position of this group relative to other families is not
542 supported by high bootstraps in either of the trees.

543 NarL family: The NarL family members create a sum of clades, each consisting mostly of
544 sequences of one particular NarL protein type. The positions of all the family members are
545 very close to each other, to the extent that there are no defined pairs of protein types. The
546 RcsB and ComA clades branch later in the NarL family. All RcsB and ComA RRs are grouped in
547 their own RR type clade, with the exception of abamRcsB. The clade of the UvrY sequences
548 has, in general, nodes of low bootstrap value and five sequences are placed outside of it:
549 tjeUvrY, dunUvrY and mmarUvrY. All the FusR RRs form one clade, except spirFusR and
550 sphnFusR, which are right before the clade of the RcsB proteins. Most NarPs are divided into
551 two main clades, one with NarL sequences and a second one near the ComA sequences. The
552 NarL sequences are widely distributed in the clades of the NarL family, forming several small
553 clades. One group of these clades is close to the NarP sequences and another group is placed
554 among the ComA and FusR clades.

555 OmpR family: The members of the OmpR family are grouped together (86% bootstrap value)
556 and there is one individual clade for each protein type, supported by significant bootstrap
557 values (except NisR: 63%). SaeR and NisR group together with 72% bootstrap value and CiaR
558 with QseB (58% bootstrap). This recapitulates the pattern seen in Fig 5B. Only phmQseB is
559 distant in the tree, near th RpfG clade (but with inisgnficant bootstrap support).

560 Lux family: The Lux family (LuxO) (78%) is close to NtrC, forming a common clade (52%). This
561 clade shares a common node with the OmpR family (69%).

562 LytTR family: The clade for all the RRs of the LytTR family is closer to the NarL and RpfG
563 family members and it consists of many smaller clades or individual branches. The different
564 RR types belong to separate clades where the AgrA2 sequences branch out earlier of all and
565 they have a common node with the AgrA clade (100%), wheras the ComE is the last clade.

566 NtrC family (GlrR protein): The GlrR sequences group together; however, the early nodes
567 have a very low bootstrap value. The obgGlrR sequence branches out early with a very low

568 bootstrap value (32%), while dyeGlrR lies outside the NtrC clade, on the LuxO clade, with
569 78% bootstrap value.

570 "Other" (RpfG protein): The RpfG RRs form one clade with 96% bootstrap value, which
571 includes no sequences of other RR type. The closest RR family is NarL, complying with the
572 structure of the tree of all the HKs (Fig. 8).

573

574 **Discussion**

575 The QS mechanism includes a multitude of signal molecules and pathways, which aim
576 to coordinate the activities of a bacterial population by changing gene expression according
577 to the population density. The fact that this mechanism affects various processes makes this
578 mechanism a potential target in antibiotics research. Few phylogenetic analyses on the
579 bacterial QS exist, and given the multiple pathways contributing to QS, studies have so far
580 focused only on τhe signaling molecules, or on specific pathways. Among them, the LuxI/R
581 pathway (Gram positive) has been mainly studied. Based on the lack of a wider phylogenetic
582 study of QS proteins, we chose to analyze proteins of the QS pathways which belong to a
583 widespread signal transduction system, the Two Component System (TCS). This mechanism
584 includes pairs of a Histidine Kinase (HK) and its cognate Response Regulator (RR). The
585 studied QS pathways are found in a wide variety of both Gram-positive and Gram-negative
586 bacteria.

587 QS protein conservation

588 Looking at the distribution of the different QS pathways across bacterial taxa, the
589 overall conclusion is that α-, β-, γ-proteobacteria, clostridia and bacilli possess the highest
590 number of the QS proteins (Fig. 1). The most widespread proteins are CsrA, FusR, QseB,
591 DesA, malt and LuxS. These belong to various pathways and families and they have different
592 roles, indicating no clear relationship between these characteristics and the number of the
593 bacterial groups containing them. The least widespread proteins are found in the two Com
594 pathways, the Rcs and Agr pathways, as well as the Lux pathway. These results suggest that
595 the wide presence of one protein doesn't mean a similar distribution for the rest of the
596 proteins of the same pathway. This contradiction is pointed out by the fact that the Lux and
597 UvrY pathways contain both the most widespread and some of the most limited protein
598 types (LuxQ, Qrr and CsrB). Also, there is a difference even in the distribution of the HK/RR
599 pairs, i.e. between the HK and its cognate RR. This distribution may simply be the result of
600 inadequacies in genome annotation, or it may be an indication of flexibility in the system,
601 meaning that HKs can function with other than their "cognate" RRs, in essence leading to
602 mixed pathways. Although TCS pathways demonstrate a high specificity overall, such "cross-
603 talk" between different TCS pathways is possible (Agrawal et al., 2016). In the case of
604 "orphan" HKs or RRs with no cognate protein (Freedman et al., 2019; Wang et al., 2009), an

605 interaction between an orphan protein with a non-cognate one can also occur (Wang et al.,
606 2009). Also, given the QS map of KEGG (Fig. 3), QseC can interact with GlrR (QseF), which is
607 not only a non-cognate RR, but it also belongs to a different family (NtrC); interestingly, QseC
608 and GlrK both bind to Adrenaline and Noradrenaline. All of the interactions depicted in the
609 QS maps of KEGG (maps ko02020 and ko02024) involve proteins of the OmpR family. QseB
610 inhibits the cognate pair of FusK/FusR of the NarL family, CiaR inhibits ComC, ComD and
611 ComE proteins of the LytTR family and SaeS is activated by the agr QS signal.
612 Examining each protein family of our study separately, it seems that the most
613 widespread HK/RR pairs of QS are those of the NarL, OmpR and NtrC families, whereas the
614 Lux and LytTR are the least common, although the Lux pathway is the most well-known and
615 well-studied QS pathway.
616

617 Conserved Domains

618 The HK-RR pairs belong to 6 families, as defined by KEGG: OmpR, LytTR, NarL, Lux, NtrC
619 and "Other". In both the OmpR and the LytTR family, the different protein types consist of a
620 small number of common conserved domains. On the other hand, HKs of the NarL and Lux
621 families show common domains and a number of variable domains; notably, the NarL and
622 Lux families also include hybrid HKs. The RRs of all families contain common domains (all of
623 them have the Helix-Turn-Helix motif).

624 The Lux family consists exclusively of Hybrid HKs, and all 3 of them activate a common
625 phosphotransfer protein and RR. The NtrC contains only the GlrK HK, which has the same
626 conserved domains with the HKs of the OmpR family, a characteristic which is also evident in
627 their close relationship on the phylogenetic tree. However, its cognate RR (GlrK) has various
628 common domains with the RR of the Lux family. This is also evident in the phylogenetic
629 trees. Therefore, we conclude that similarities can be found not only among the protein
630 types within a family, but also among different family members, which indicate a possible
631 common evolutionary history.

632 There is partial agreement of our results with previous sudies on the classification of
633 HKs and RRs. The differences are mostly due to differences in the selection of proteins
634 included in each study, and due to differences in classification based on domain structure
635 alone versus phylogenetic analysis. For example, looking at the classification of the HKs in
636 the Grebe and Stock study, ComP, NarQ and NarX were classified in group HPK7, CiaH in
637 HPK1a, NisK in HPK3c and BarA, RcsC, RpfC in HPK1b (Grebe and Stock, 1999). However, only
638 some of the HKs mentioned here were also included in that study, and this way of
639 classification differs from the KEGG database and our phylogenetic trees results, as CiaH and
640 NisK (OmpR family members) belong to separate groups and RpfC (hybrid HK) is placed with
641 the hybrid HKs of the NarL family (although hybrid HKs were placed in other HPK groups as
642 well).

643 In the study of Forst and Kim, BarA and RcsC were characterized as Type IB and NarX
644 and NarQ as Type III HKs (Forst and Kim, 2015). However, no members of other families
645 included in our study were present in that analysis, so we cannot conclude if the comparison
646 criteria used by Forst and Kim led to a different classification or if their classification is more
647 detailed and separated the members of one family into multiple groups. Capra and Laub
648 classified the HKs into 4 groups depending on the presence of PAS, HAMP and GAF domains
649 or the absence of these domains (Capra and Laub, 2012). According to the authors, most HKs
650 have at least one domain between the transmembrane part of the protein and the Histidine
651 phosphotransfer domain, with PAS, HAMP and GAF being "by far the most common" ones.
652 However, our studied proteins had no PAS and GAF domains, with the exception of RcsC, in
653 which the PAS domain had a very high E-value in KEGG and SMART and thus it didn't meet
654 our criterion to be included in the conserved domains of our study. Nevertheless, all three
655 HKs with a HAMP domain (NarX, NarQ, BarA) were grouped to the same family (NarL), in
656 agreement with the Capra and laub study. What contradicts their classification is the fact
657 that the NarL family contains both HKs with HAMP and others without a HAMP domain,
658 which are supposed to belong to different groups according to Capra and Laub. Similarly, the
659 effector domains for the classification of the RRs were also different in the research of Capra
660 and Laub, as no GGDEF and Methyltransfer domains were present in the proteins of our
661 study. However, the LuxO and GlrR proteins possessing the AAA_5 domain, were grouped
662 close to each other in our phylogenetic trees, complying with the classification criterion of
663 Capra and Laub.

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

671 Phylogenetic trees

672 In all the inferred phylogenetic trees of each family, the amino acid sequences of both
673 HKs and RRs are grouped in separate distinct clades according to their protein type and in
674 many cases, with high bootstrap value. The same happens in the phylogenetic trees of all
675 families together, with distinct clades for sequences of different family and protein types.
676 Also, the relation among protein types (either HKs or RRs) is generally the same in the tree of
677 each family and of all families together. In most cases, the relations among the HKs are also
678 reflected by the relations of their cognate RRs, indicating a common evolutionary history, as
679 has already been suggested by former studies on other proteins of the TCS (Koretke et al.,
680 2000).

681 The members of the NarL family form pairs of HKs that are closely related, NarX-NarQ,
682 BarA-RcsC and ComP-FusK in both the tree of the NarL family alone and the tree of all the HK
683 families together. In both trees the BarA-RcsC group is more distant. The tree for all the
684 families together placed all RR types of the NarL family very close to each other: NarL and
685 NarP clades are clustered together and UvrY sequences are placed close to RcsB, being in
686 agreement with the trees of their cognate HKs. However, in the RR tree of the NarL family
687 alone, only NarL and NarP demonstrate a similar grouping to the HKs from which they are
688 activated: ComA is closer to RcsB (instead of FusR), and UvrY is closer to the clade of
689 NarL/NarP (instead of RcsB).

690 The different protein types of the OmpR family demonstrate the same relation in all
691 the phylogenetic trees (of the OmpR family alone and the tree with all the HK families
692 together). There are two pairs of HKs: CiaH-QseC and SaeS-NisK. Their cognate RRs follow
693 the same relation in all the RRs phylogenetic trees: the SaeR clade is closer to NisR and the
694 CiaR clade is closer to QseB. Regarding the Lux HKs, both the tree of the Lux family and the
695 tree for all the HKs show a similarity between the CqsS and LuxN. The latter also places the
696 LuxQ sequences in the clade of other families (BarA, RpfC and RcsC sequences). Their
697 common RR (LuxO) is in the same clade with GlrR in the tree of all RRs. The protein members
698 of the LytTR family have separate clades for the AgrC -ComD and ArgA-ComE sequences in all
699 trees. The AgrC2 and AgrA2 sequences belong to the AgrC and AgrA clade, which is
700 confirmed by all the inferred phylogenetic trees. The position of the RpfC/RpfG and
701 GlrK/GlrR sequences (NtrC family) are slighlty different for the HKs and their RRs. The RpfC
702 sequences are closer to RcsC, whereas RpfG is closer to NarL and NarP. GlrK is closer to the
703 OmpR family but GlrR is in the same clade with LuxO.

704 A comparison of the phylogenetic trees of all HKs and all RRs suggests two clusters of
705 the QS protein families: 1) OmpR-NtrC-Lux,and 2) NarL and LytTR, as well as clusters of
706 protein types per family as outlined above. These evolutionary relationships highlight a
707 common evolutionary history, and can inform future applications, such as the design of
708 novel inhibitors for pathogenic QS systems. In fact, there have already been numerous
709 scientific papers about Quorum Quenching, including for some of the proteins of our study.
710 Diol-containing compounds, boronic acids and sulfones have been suggested as antagonists
711 of the signaling molecules which bind with LuxP and as potential solutions to biofilm
712 formation (Brackman and Coenye, 2014). Other molecules inhibiting the QS pathways of
713 Gram-negative pathogens have been reported, including for the HKs and RRs of the Lux
714 family and RpfC/RpfG, by blocking the receptors, inhibiting the production of the signaling
715 molecules, or by degrading them (Defoirdt, 2018). The inhibition of the Agr pathway of
716 *Staphylococcus aureus* has also been studied, with natural and synthetic substances
717 targeting either AgrC or AgrA: the cyclodepsipeptide Solonamide B of *Photobacterium*
718 *halotolerans* interferes with the binding of AIPs to AgrC, whereas ω-hydroxyanodin from
719 *Penicillium restrictum* binds to AgrA and therefore inhibits the expression of the regulated
720 genes (Scoffone et al., 2019). Savirin (*S. aureus* virulence inhibitor) is another small molecule

721 that selectively targets AgrA (Sully et al., 2014). The synthetic small molecule inhibitor
722 LED209 is selective to QseC (Curtis et al., 2014). The genes ComD (HK) and ComE (RR) was
723 also found to be downregulated by a small molecule containing a 2-aminoimidazole subunit
724 (2A4) along with five other biofilm-related genes of *S. mutans* (Liu et al., 2011). The
725 antiobiotic mupirocin when used in sub-inhibitory concentrations on high-level mupirocin-
726 resistant MRSA strains (Methicillin-resistant Staphylococcus) interferes with the agrA and
727 saeRS genes (Jin et al., 2018). Our prediction is that such inhibitors are more likely to also
728 interfere with protein types similar to the original target protein, as highlighted by the
729 phylogeneti analysis, than with more distant ones, e.g. an inhibitor for a Lux family HK may
730 also affect NtrC family HKs. On the other hand, pathway cross-talk in species with multiple
731 QS systems may hinder these inhibitory effects. Broader inhibition approaches have also
732 been suggested, with the ATP-binding domain of HKs (HATPase domain) being the main
733 target for TCS pathways inhibition (Wilke et al., 2015; Worthington et al., 2013).

734 Two-component signal transduction arose early in bacterial evolution after their
735 separation from the last common ancestor with significant diversification throughout the
736 subsequent bacterial speciation; HKs arose from ATPases of the GHKL superfamily and
737 pyruvate dehydrogenase kinases (PDKs), while the origin of the RRs remains unclear (Capra
738 and Laub, 2012; Koretke et al., 2000). Regarding the origin of two-component quorum
739 sensing systems, a phylogenetic analysis for the proteins LuxI, LuxR and LuxS in firmicutes,
740 proteobacteria and actinobacteria, indicated that these proteins appeared very early in the
741 evolutionary path of the bacteria, but with multiple occassions of horizontal gene transfer
742 resulting in their present-day distribution (Lerat and Moran, 2004). In a study of the
743 ComQXPA pathway in firmicutes, the phylogenetic tree of HK proteins clustered all the ComP
744 proteins from various organisms together, instead of gathering them with other HKs of the
745 same organism, indicating an early origin of ComP (Dogsa et al., 2014). The phylogenetic
746 analyses presented here are unrooted and assume an early origin of these pathways in
747 bacterial evolution. We present the distribution and evolutionary relationships of different
748 HK and RR protein types and families, but a larger-scale analysis of all available sequences
749 for two-component QS systems and other HK/RR pairs would be needed to confidently
750 address the issue of the root of quorum sensing.

751

752 **References**

753 Agrawal, R., Sahoo, B.K., Saini, D.K., 2016. Cross-talk and specificity in two-component signal
754 transduction pathways. Future Microbiol. 11, 685–697. https://doi.org/10.2217/fmb-
755 2016-0001

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

759 Brackman, G., Coenye, T., 2014. Quorum Sensing Inhibitors as Anti-Biofilm Agents. Curr.
760 Pharm. Des. 21, 5–11. https://doi.org/10.2174/1381612820666140905114627

761 Capra, E.J., Laub, M.T., 2012. Evolution of Two-Component Signal Transduction Systems.
762 Annu. Rev. Microbiol. 66, 325–347. https://doi.org/10.1146/annurev-micro-092611-
763 150039

764 Curtis, M.M., Russell, R., Moreira, C.G., Adebesin, A.M., Wang, C., Williams, N.S., Taussig, R.,
765 Stewart, D., Zimmern, P., Lu, B., Prasad, R.N., Zhu, C., Rasko, D.A., Huntley, J.F., Falck,
766 J.R., Sperandio, V., 2014. QseC inhibitors as an antivirulence approach for gram-
767 negative pathogens. MBio 5. https://doi.org/10.1128/mBio.02165-14

768 Defoirdt, T., 2018. Quorum-Sensing Systems as Targets for Antivirulence Therapy. Trends
769 Microbiol. 26, 313–328. https://doi.org/10.1016/j.tim.2017.10.005

770 Dereeper, A., Guignon, V., Blanc, G., Audic, S., Buffet, S., Chevenet, F., Dufayard, J.F.,
771 Guindon, S., Lefort, V., Lescot, M., Claverie, J.M., Gascuel, O., 2008. Phylogeny.fr: robust
772 phylogenetic analysis for the non-specialist. Nucleic Acids Res. 36, W465–W469.
773 https://doi.org/10.1093/nar/gkn180

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

778 Edgar, R.C., 2004. MUSCLE: Multiple sequence alignment with high accuracy and high
779 throughput. Nucleic Acids Res. 32, 1792–1797. https://doi.org/10.1093/nar/gkh340

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

782 Freedman, J.C., Li, J., Mi, E., McClane, B.A., 2019. Identification of an important orphan
783 histidine kinase for the initiation of sporulation and enterotoxin production by
784 Clostridium perfringens type F strain SM101. MBio 10, 1–17.
785 https://doi.org/10.1128/mBio.02674-18

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

789 Grebe, T.W., Stock, J.B., 1999. The histidine protein kinase superfamily. Adv. Microb. Physiol.
790 41, 139–227. https://doi.org/10.1016/s0065-2911(08)60167-8

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

794 Jacob, S., Foster, A.J., Yemelin, A., Thines, E., 2014. Histidine kinases mediate differentiation,
795 stress response, and pathogenicity in Magnaporthe oryzae. Microbiol. Open 3, 668–

796 687. https://doi.org/10.1002/mbo3.197

797 Jin, Y., Li, M., Shang, Y., Liu, L., Shen, X., Lv, Z., Hao, Z., Duan, J., Wu, Y., Chen, C., Pan, J., Yu,
798 F., 2018. Sub-inhibitory concentrations of mupirocin strongly inhibit alpha-toxin
799 production in high-level mupirocin-resistant MRSA by down-regulating agr, saeRS, and
800 sarA. Front. Microbiol. 9, 1–8. https://doi.org/10.3389/fmicb.2018.00993

801 Jun, S.R., Sims, G.E., Wu, G.A., Kim, S.H., 2010. Whole-proteome phylogeny of prokaryotes
802 by feature frequency profiles: An alignment-free method with optimal feature
803 resolution. Proc. Natl. Acad. Sci. U. S. A. 107, 133–138.
804 https://doi.org/10.1073/pnas.0913033107

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

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

810 Koretke, K.K., Lupas, A.N., Warren, P. V, Rosenberg, M., Brown, J.R., 2000. Evolution of Two-
811 Component Signal Transduction. Mol Biol Evol. 17, 1956–1970.

812 Koretke, K.K., Volker, C., Bower, M.L., Lupas, A.N., 2003. Molecular Evolution of Histidine
813 Kinases. Histidine Kinases Signal Transduct. 483–506. https://doi.org/10.1016/B978-
814 012372484-7/50024-2

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

817 Liu, C., Worthington, R.J., Melander, C., Wu, H., 2011. A new small molecule specifically
818 inhibits the cariogenic bacterium Streptococcus mutans in multispecies biofilms.
819 Antimicrob. Agents Chemother. 55, 2679–2687. https://doi.org/10.1128/AAC.01496-10

820 Madeira, F., Park, Y.M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey,
821 A.R.N., Potter, S.C., Finn, R.D., Lopez, R., 2019. The EMBL-EBI search and sequence
822 analysis tools APIs in 2019. Nucleic Acids Res. 47, W636–W641.
823 https://doi.org/10.1093/nar/gkz268

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

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

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

831 Pearson, W.R., 2013. Selecting the Right Similarity-Scoring Matrix. Curr Protoc Bioinforma.
832 43, 3.5.1–3.5.9. https://doi.org/10.1002/0471250953.bi0305s43

833 Rasmussen, B.B., Nielsen, K.F., Machado, H., Melchiorsen, J., Gram, L., Sonnenschein, E.C.,
834 2014. Global and Phylogenetic Distribution of Quorum Sensing Signals, Acyl Homoserine
835 Lactones, in the Family of Vibrionaceae. Mar. Drugs 12, 5527–5546.
836 https://doi.org/10.3390/md12115527

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

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

842 Scoffone, V.C., Trespidi, G., Chiarelli, L.R., Barbieri, G., Buroni, S., 2019. Quorum sensing as
843 antivirulence target in cystic fibrosis pathogens. Int. J. Mol. Sci. 20.
844 https://doi.org/10.3390/ijms20081838

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

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

850 Sully, E.K., Malachowa, N., Elmore, B.O., Alexander, S.M., Femling, J.K., Gray, B.M., DeLeo,
851 F.R., Otto, M., Cheung, A.L., Edwards, B.S., Sklar, L.A., Horswill, A.R., Hall, P.R., Gresham,
852 H.D., 2014. Selective Chemical Inhibition of agr Quorum Sensing in Staphylococcus
853 aureus Promotes Host Defense with Minimal Impact on Resistance. PLoS Pathog. 10.
854 https://doi.org/10.1371/journal.ppat.1004174

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

859 Wang, W., Shu, D., Chen, L., Jiang, W., Lu, Y., 2009. Cross-talk between an orphan response
860 regulator and a noncognate histidine kinase in streptomyces coelicolor. FEMS
861 Microbiol. Lett. 294, 150–156. https://doi.org/10.1111/j.1574-6968.2009.01563.x

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

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

867 Wilke, K.E., Francis, S., Carlson, E.E., 2015. Inactivation of multiple bacterial histidine kinases
868 by targeting the ATP-binding domain. ACS Chem. Biol. 10, 328–335.
869 https://doi.org/10.1021/cb5008019

870 Worthington, R.J., Blackledge, M.S., Melander, C., 2013. Small-molecule inhibition of
871 bacterial two-component systems to combat antibiotic resistance and virulence. Future
872 Med. Chem. 5, 1265–1284. https://doi.org/10.4155/fmc.13.58

873
874

Figure and table 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. “Other” Spirochaetes refer toTreponema, Spirochaeta, Leptospira and Salinispira.

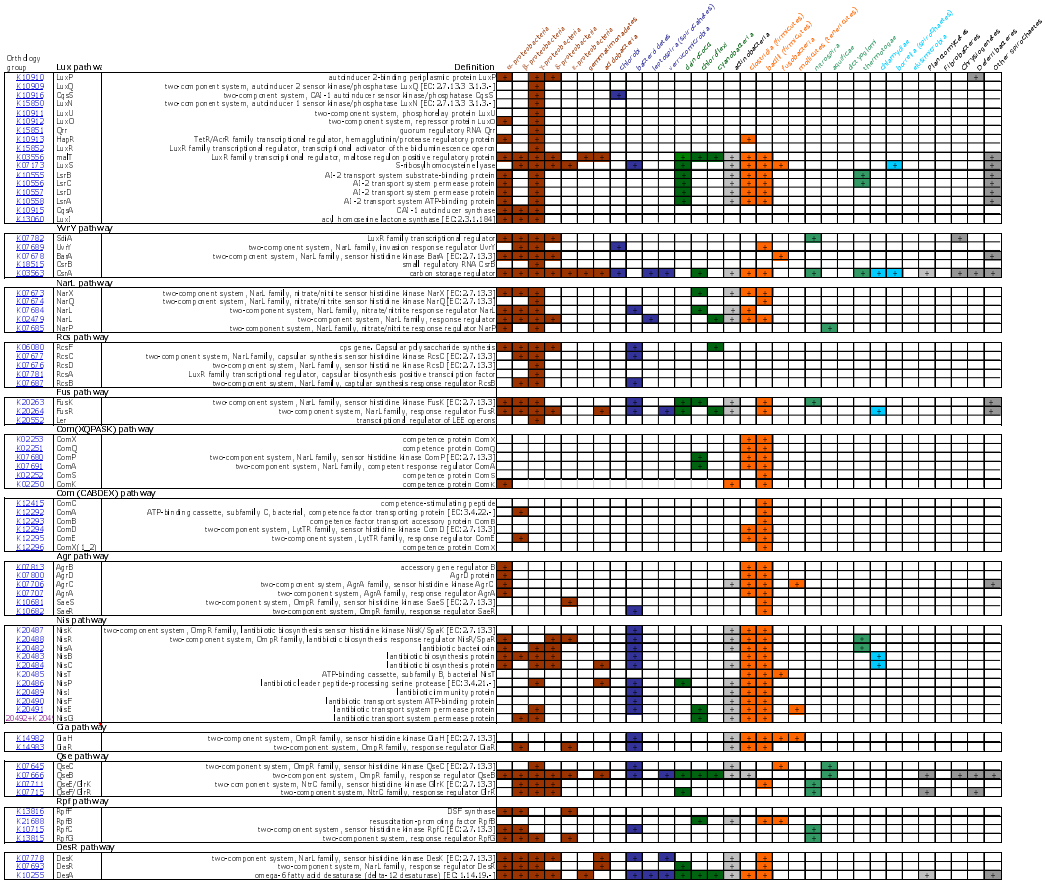


Figure 2: The number of bacterial groups in which each protein is found. The different colors indicate the family of each HK and RR in our study; proteins involved in the same pathway are indicated with a lighter shade of the same color. Blue: Lux family, Green: NarL family, Yellow: LytTR family, Orange: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

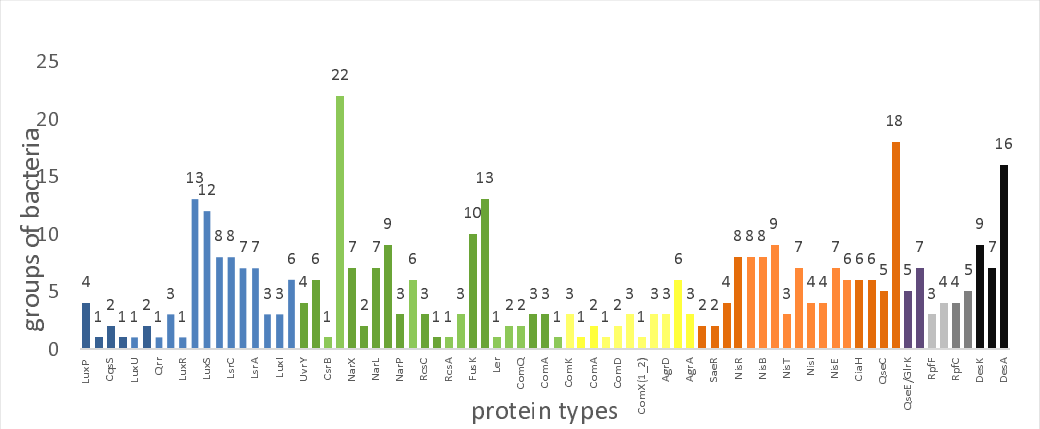


Figure 3: An overview of the position of the conserved domains in the HKs and RRs of this study. The similar colors of domains indicate their presence in a common clan of the Pfam database. The pathway drawings are from KEGG PATHWAY map ko02020, edited with Microsoft Draw.

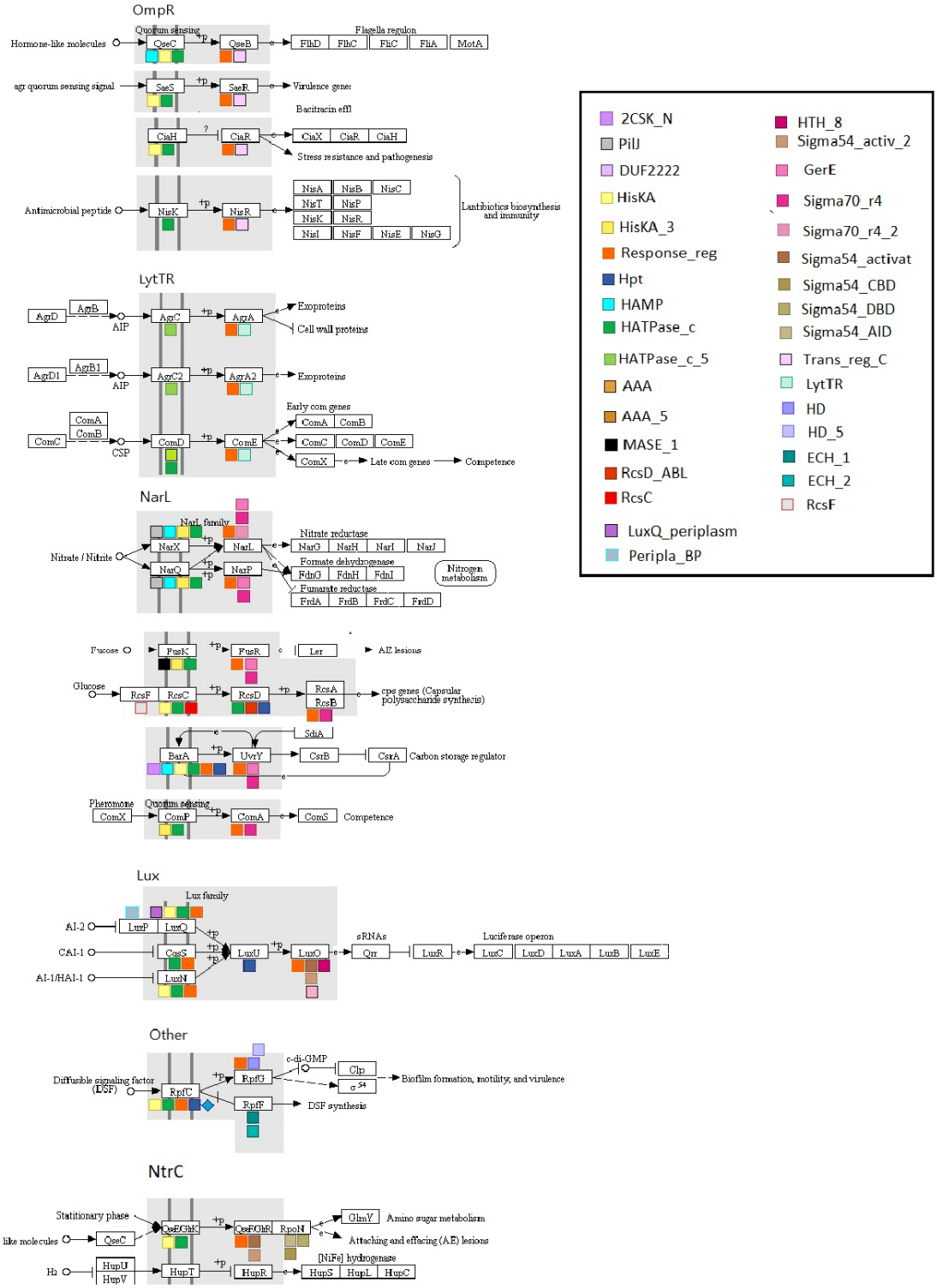


Figure 4: Phylogenetic analysis of the NarL family. (A) Histidine Kinases. (B) Response Regulators. Both trees reconstructed using PhyML based on the MAFFT alignment. The color of the bullets indicates the bootstrap value of the nodes. White: 50% -80%, Grey: 80%-95%, Black: over 95%

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

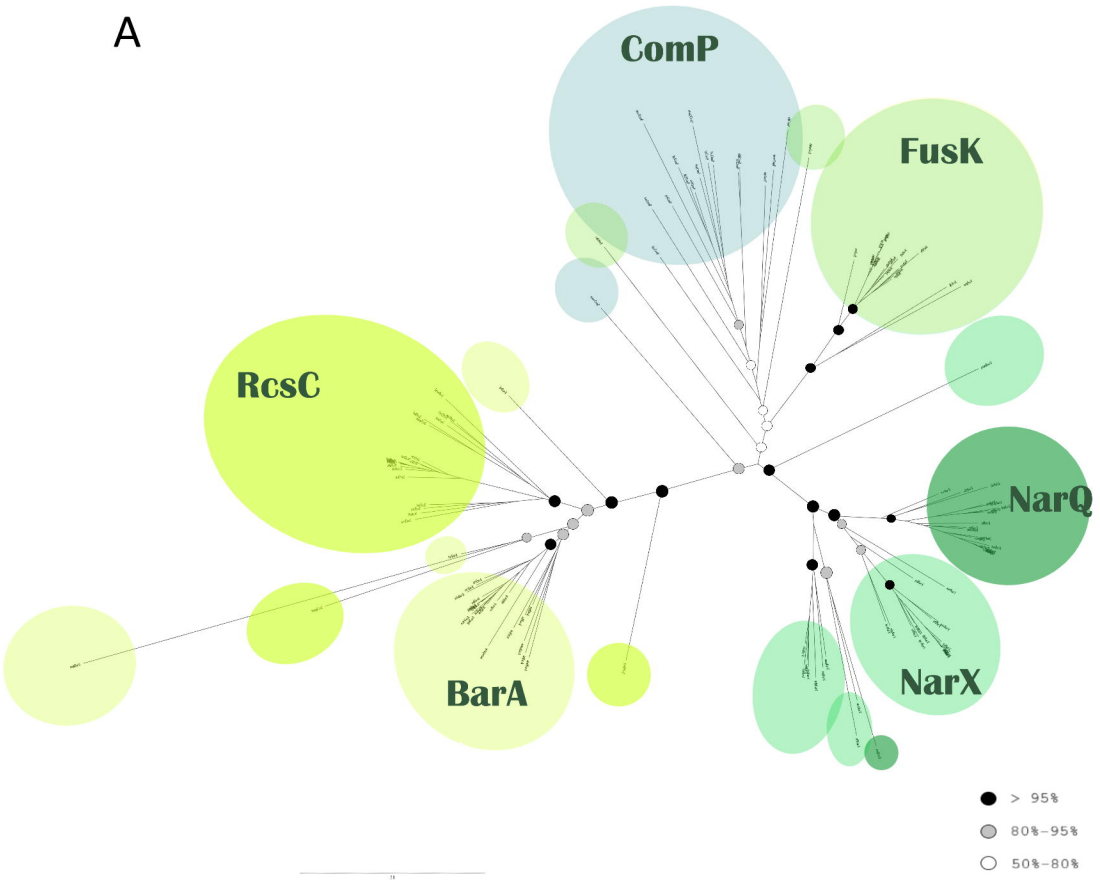
Figure 9: Phylogenetic analysis of all response regulator families. Tree reconstructed using RaxML based on the MAFFT alignment. The bullets indicate the bootstrap value of the nodes, as in Figure 4.

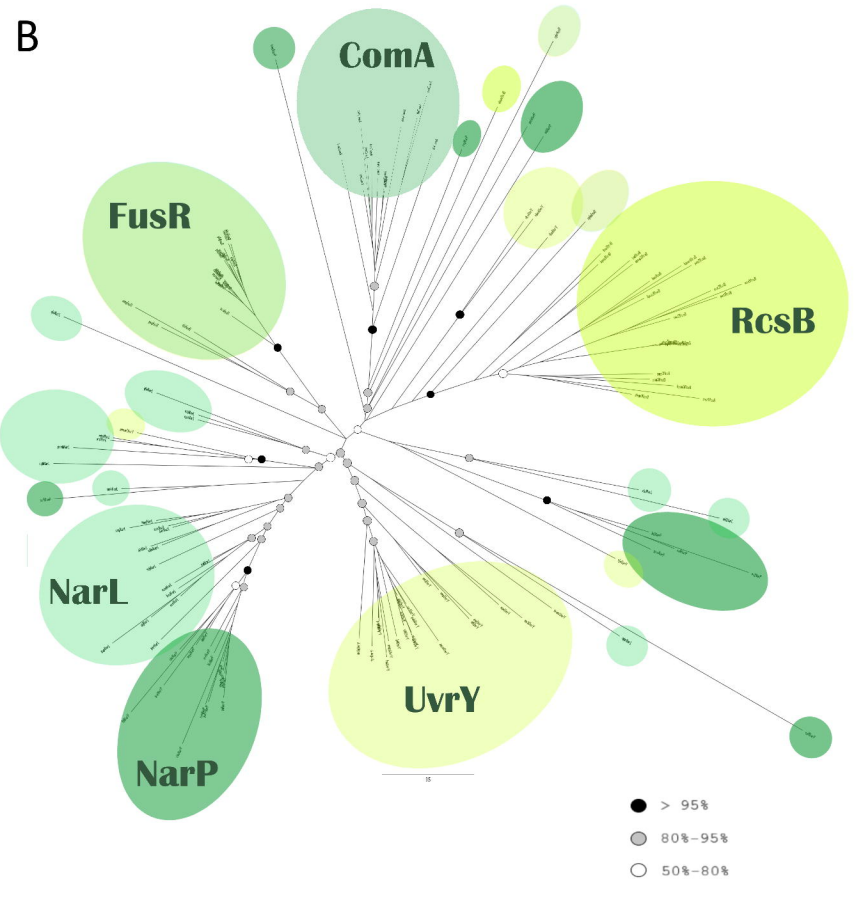
Table 1: Categorization of the conserved domains of all the HK and RR types in this study.

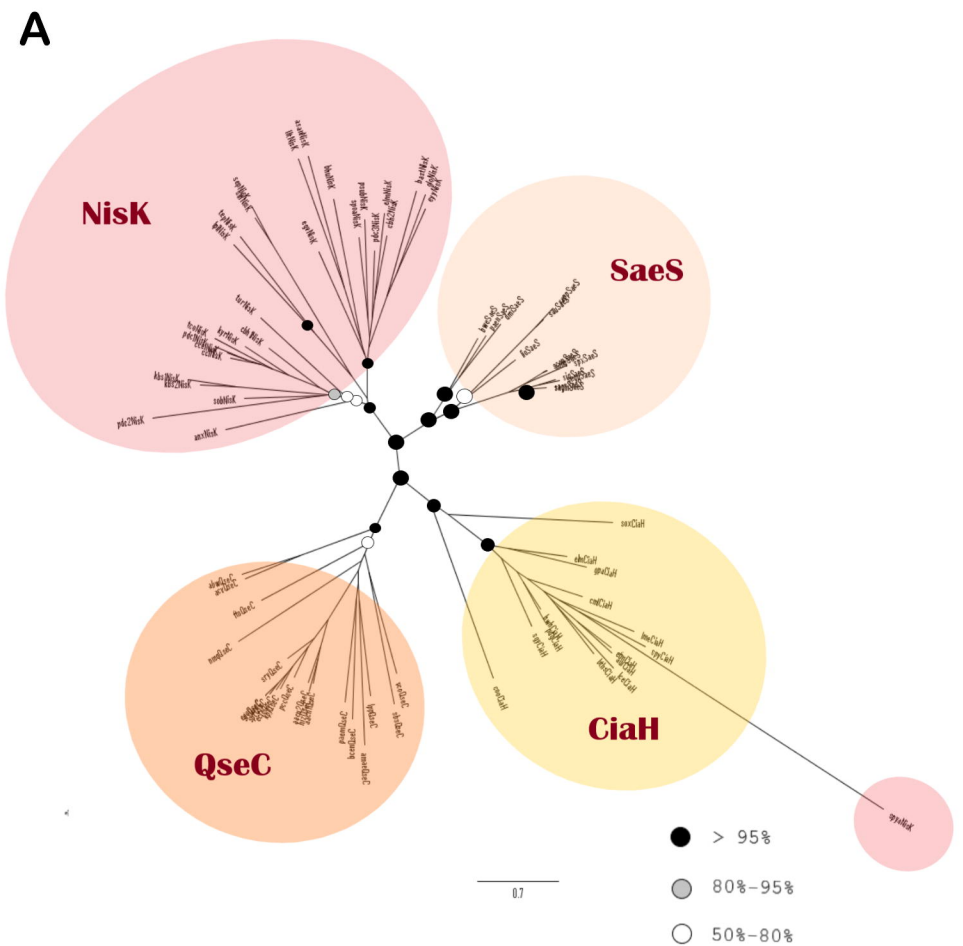
Clans	Domains					
HTH (Helix-Turn-Helix)	HTH_8	Sigma70_r4	Sigma70_r4_2	GerE	Sigma54_CBD	Sigma54_DBD
His kinase A	HATPase_c	HATPase_c_5	HisKA	HisKA_3		
P-loop-NTPase	AAA	AAA_5	Sigma54_activitat	Sigma54_activ_2		
Cache-like domains	2CSK_N	DUF2222	LuxQ-periplasm periplasmic sensor domain			
Chey-like superfamily	Response_reg Receiver domain of RRs	RcsD_ABL Found in the RcsD HK, recognizes the RcsB effector domain.	RcsC			
HD-PDEase	HD	HD_5				
Clp crotonase	ECH_1 Enoyl-CoA hydratase	ECH_2 (Enoyl-CoA hydratase				
Periplas BP	Peripla_BP					
HAMP	HAMP					
Tudor	LytTR DNA-binding domain					
YbjQ-like	RcsF lipoprotein					

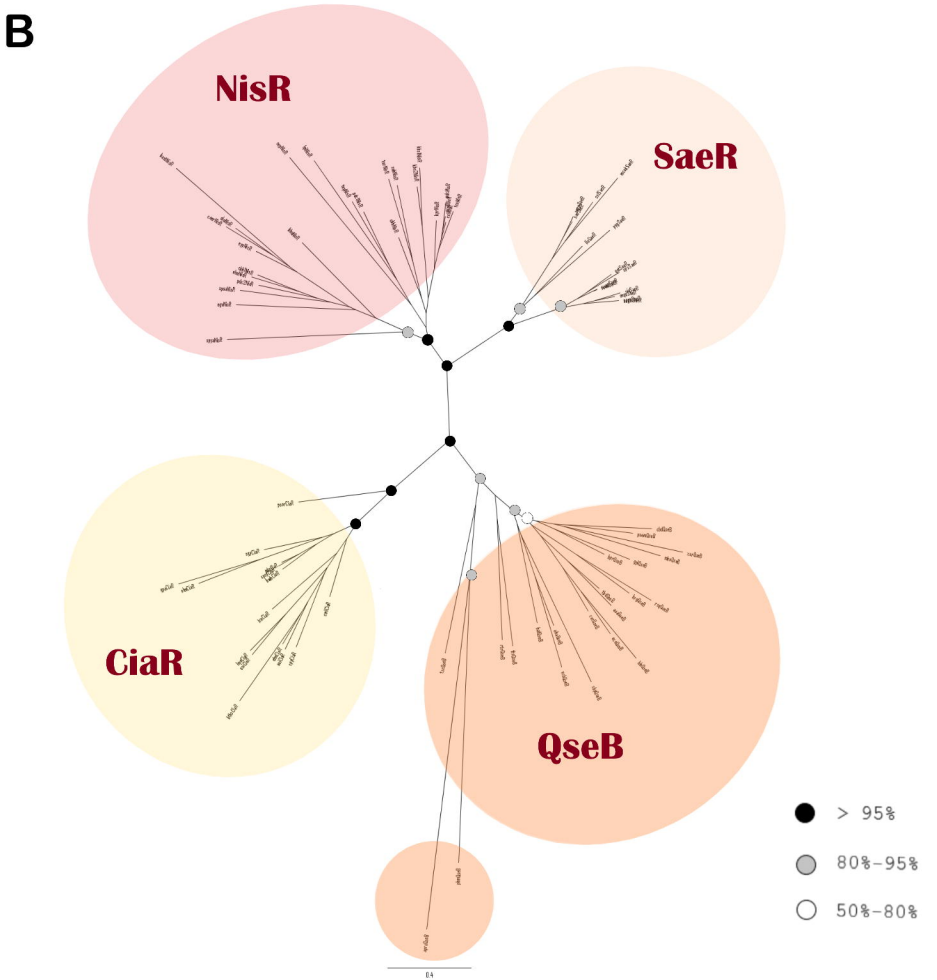
Table 2: Occurence of the conserved domains of all the HKs and RRs in the different protein types and families.

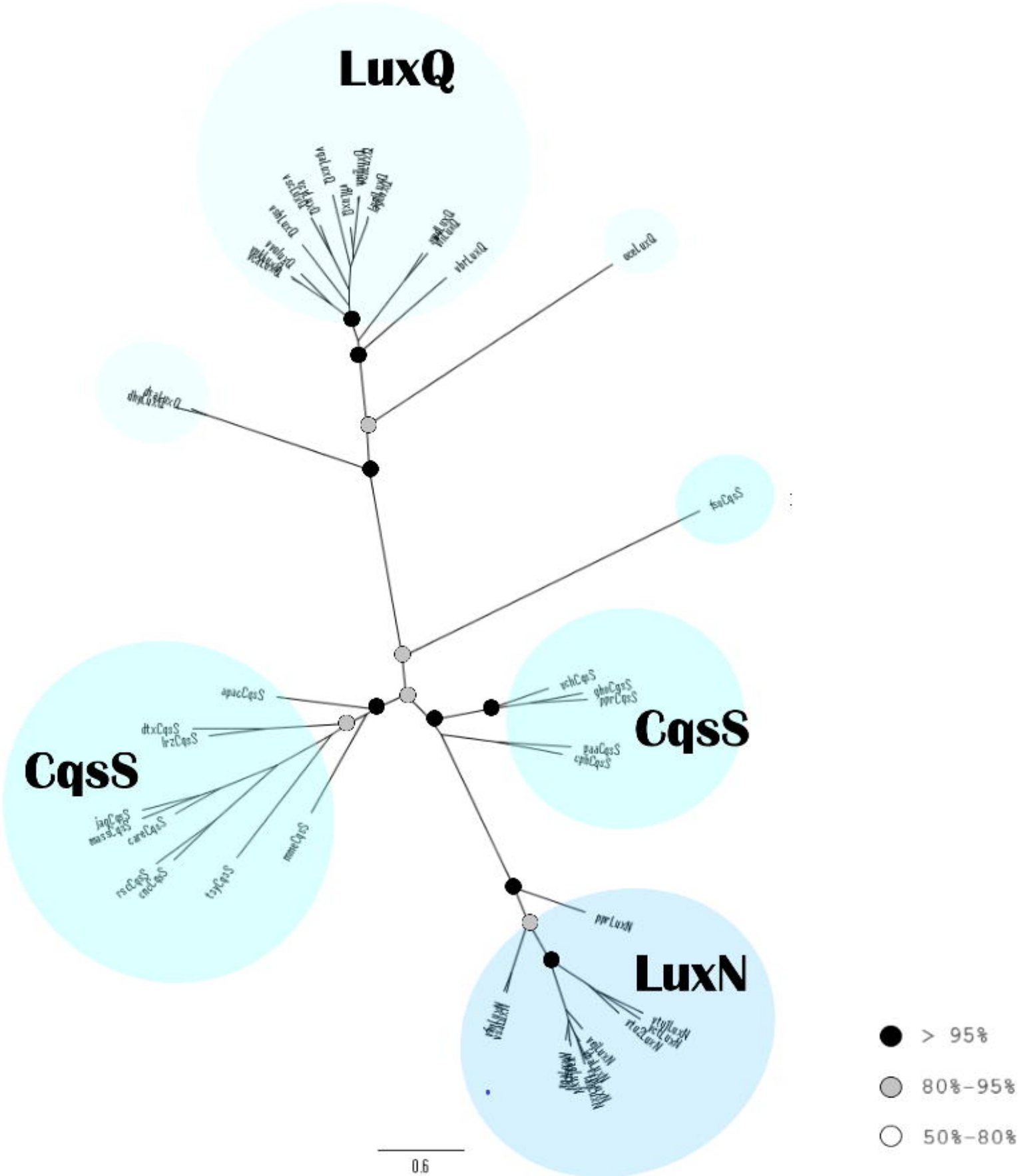
Domain	Protein Types	Family
HATPase_c	15	All of the families
HATPase_c_5	3	LytTR
HisKA	9	OmpR, NarL, Lux, Other, NtrC
HisKA_3	4	NarL
HAMP	3	NarL
Response_reg	3	NarL, Lux, Other
Hpt	2	NarL, Other
Trans_reg_C	4	OmpR
LytTR	3	LytTR
GerE	5	NarL
Sigma70_r4_2	4	NarL
Sigma54_activat	2	LuxO, NtrC
Sigma54_activ_2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

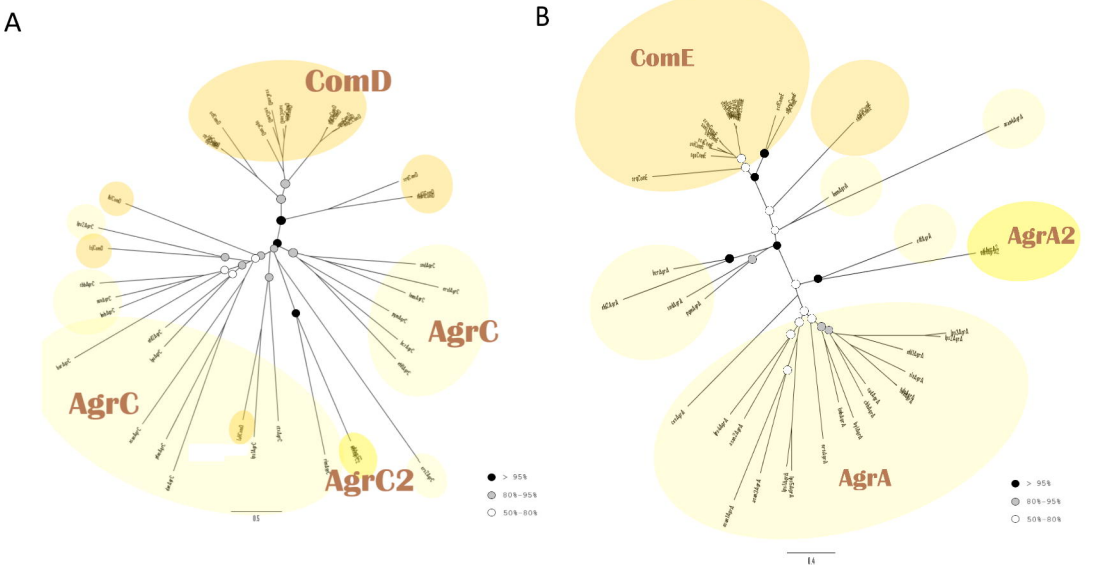

A
ComP
FusK
RcsC
NarQ
BarA
NarX
> 95%
80%–95%
50%–80%

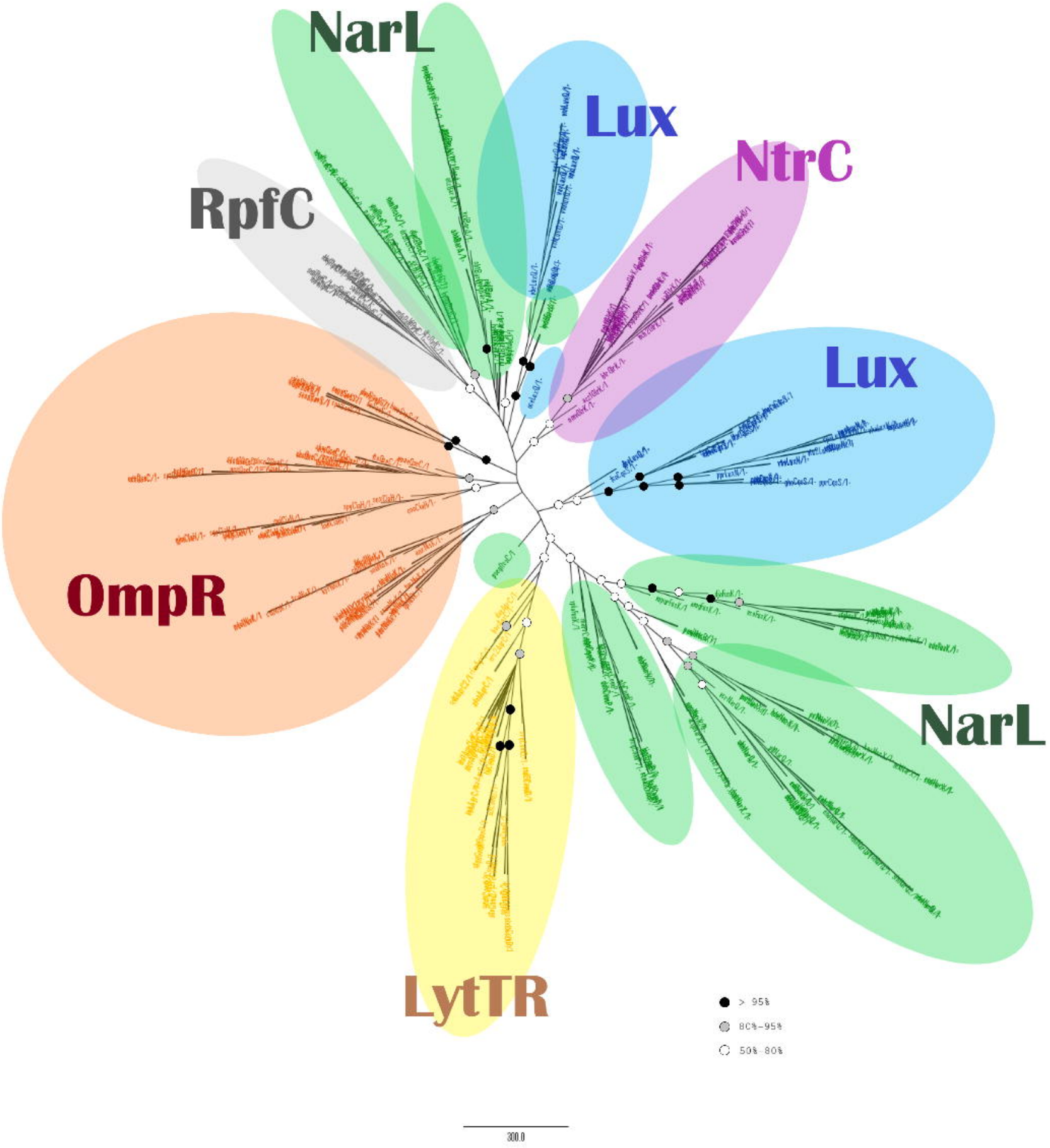

B
ComA
FusR
RcsB
NarL
UvrY
NarP
> 95%
80%–95%
50%–80%

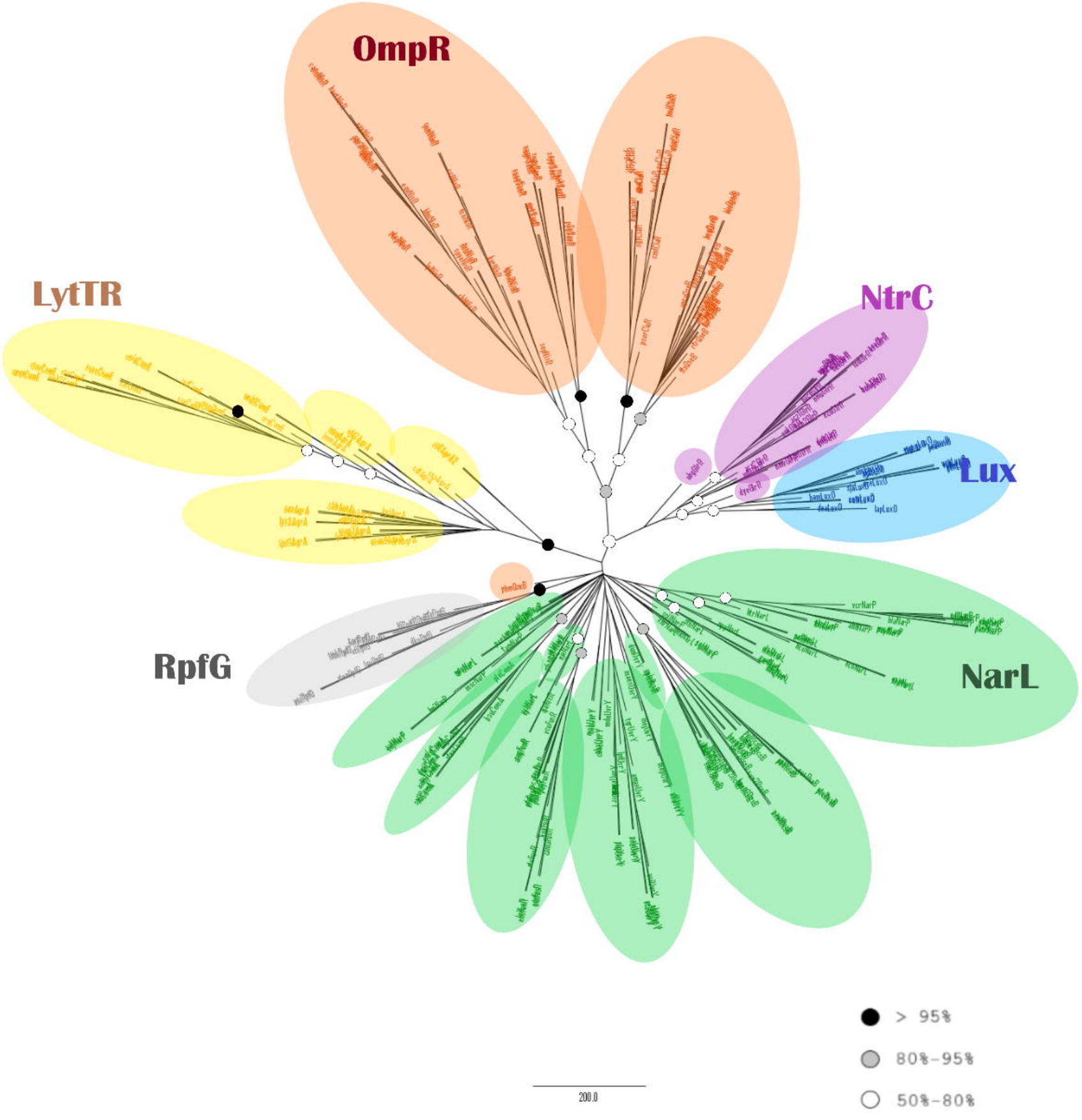

A
NisK
SaeS
QseC
CiaH
> 95%
80%-95%
50%-80%
0.7


B
NisR
SaeR
CiaR
QseB
> 95%
80%-95%
50%-80%
0.4


LuxQ
CqsS
CqsS
LuxN
> 95%
80%–95%
50%–80%
0.6


A
ComD
AgrC
AgrC
AgrC2
> 95%
80%-95%
50%-80%
B
ComE
AgrA2
AgrA
> 95%
80%-95%
50%-80%


NarL
Lux
NtrC
RpfC
Lux
OmpR
NarL
LytTR
> 95%
80%–95%
50%–80%
300.0


OmpR
LytTR
NtrC
Lux
RpfG
NarL
> 95%
80%-95%
50%-80%
200.0