Novel Two-Component System-Like Elements Reveal Functional Domains Associated with Restriction–Modification Systems and paraMORC ATPases in Bacteria

Daniel Bellieny-Rabelo[1,2,†], Willem J.S. Pretorius[1,2], and Lucy N. Moleleki[1,2,*]

[1]Department of Biochemistry, Genetics and Microbiology, University of Pretoria, Gauteng, South Africa

[2]Forestry and Agricultural Biotechnology Institute, University of Pretoria, Gauteng, South Africa

†Present address: Department of Ecology and Environmental Sciences, Umeå University, Umeå, Sweden

*Corresponding author: E-mail: lucy.moleleki@up.ac.za.

Accepted: 4 February 2021

Abstract

Two-component systems (TCS) are important types of machinery allowing for efficient signal recognition and transmission in bacterial cells. The majority of TCSs utilized by bacteria is composed of a sensor histidine kinase (HK) and a cognate response regulator (RR). In the present study, we report two newly predicted protein domains—both to be included in the next release of the Pfam database: Response_reg_2 (PF19192) and HEF_HK (PF19191)—in bacteria which exhibit high structural similarity, respectively, with typical domains of RRs and HKs. Additionally, the genes encoding for the novel predicted domains exhibit a 91.6% linkage observed across 644 genomic regions recovered from 628 different bacterial strains. The remarkable adjacent colocalization between genes carrying Response_reg_2 and HEF_HK in addition to their conserved structural features, which are highly similar to those from well-known HKs and RRs, raises the possibility of Response_reg_2 and HEF_HK constituting a new TCS in bacteria. The genomic regions in which these predicted two-component systems-like are located additionally exhibit an overrepresented presence of restriction–modification (R–M) systems especially the type II R–M. Among these, there is a conspicuous presence of C-5 cytosine-specific DNA methylases which may indicate a functional association with the newly discovered domains. The solid presence of R–M systems and the presence of the GHKL family domain HATPase_c_3 across most of the HEF_HK-containing genes are also indicative that these genes are evolutionarily related to the paraMORC family of ATPases.

Key words: two-component system, MORC, GHKL ATPases, DNA modification methylases, DNA restriction–modification enzymes, histidine kinase.

Significance

Bacteria are pervasive in virtually all human-inhabited environments, therefore impacting human health and the economy. Hence, understanding molecular mechanisms in bacteria that orchestrate fine-tuned responses to environmental challenges is a powerful resource to manipulate these organisms in different capacities to human's best interest. In this study, we predict two novel protein domains that are likely involved in environmental responses by bacteria. The availability of these sequence patterns in online databases will enable their detection in future studies which may shed light on its exact function.

Introduction

Monitoring environmental conditions is determinant for the success of microbial organisms. To address this demand, biochemical cascades that enable programmatic responses to diverse stimuli have evolved across the three domains of life (Lodish et al. 2008; Alberts 2017). A dominant signaling archetype in nature is comprised of a receptor/transmitter protein paired with a ligand-binding regulator to be affected downstream, collectively known as two-component systems (TCS) (Hoch and Silhavy 1995; Stock et al. 2000). Specifically, in bacteria, preferable conservation of histidine-kinase (HK) as the transmitter module in receptor proteins has been recurrently documented since the discovery of TCSs roughly three decades ago (Ninfa and Magasanik 1986; Nixon et al. 1986; Koretke et al. 2000; Wuichet et al. 2010). HKs commonly associate with intracellular response regulators (RR), which transduce signals and generate responses through diverse mechanisms. Such mechanisms range from binding DNA or proteins to enzymatic catalysis (Romling et al. 2005; Batchelor et al. 2008; Gao and Stock 2009). As the genomic TCS collection in a single cell may comprise dozens of duplets, interaction specificity among the cognate partners of a TCS is critical to avoid cross-talk with other pathways (Galperin 2005; Skerker et al. 2005).

In this context, the canonical transduction chain in this system depends on five key domains. The first domain is a highly conserved catalytic ATP-binding domain (CA) typified by the HATPase domains (Pfam: PF02518; PF13581; PF13589; PF14501) which integrate the C-terminal regions of all known HKs and harbor a characteristic nucleotide-binding α/β sandwich known as the Bergerat fold (Bergerat et al. 1997; Dutta and Inouye 2000). The proteins harboring this structural fold are classified into the GHKL superfamily (Dutta and Inouye 2000). The second domain is a dimerization domain containing the active His site for phosphorylation (DHp). Together, DHp and CA domains comprise the two well-conserved regions in HK proteins. The prototypic DHp domain harbors the active site His in a double α-helical structure that is part of a four-helix bundle termed H-box (Karniol and Vierstra 2004). Structural variants of DHp have been described over the years, which are commonly called HiskA domains (Pfam: PF00512; PF07568; PF07730; PF07536) (Bilwes et al. 1999; Grebe and Stock 1999; Karniol and Vierstra 2004). The third well-conserved domain in TCS is the receiver domain (REC), which is commonly found in the RRs, and features a highly conserved Asp residue within its active site (Pfam: PF00072) (West and Stock 2001). This active site Asp is typically found within a characteristic $(\beta\alpha)_5$ fold conserved in RR proteins (Stock et al. 1989). The remaining two domains among the aforementioned five typical domains in TCS are highly variable and comprise: 1) a sensor/input domain (in HKs) and 2) effector/output (in RRs), respectively responsible for recognizing specific signals, and eliciting responses. The high-sequence variability observed in these domains enables a myriad of signal/target recognition patterns (Gao and Stock 2009).

The functional classification of HKs usually provides necessary insight that allows the initial characterization of the TCS role. This classification is generally inferred employing domain architecture, which encompasses transmembrane regions (TMRs) and sensor domain arrangement within the sequences (Mascher et al. 2006). Three major groups of HKs can be determined following this criterion. The largest among these groups is represented by the periplasmic-sensing HKs carrying an extracellular sensor typically flanked by two TMRs (Mascher et al. 2006). Members of this group commonly harbor a so-called "linker" domain (e.g., HAMP, or PAS) flanked by the TMR2 and the DHp domain in the protein sequence (Aravind and Ponting 1999). The second major group includes both soluble and membrane-anchored sensors typified by inward-facing input domains (Mascher et al. 2006). Among these, NtrB along with its cognate NtrC comprise a well-characterized TCS model responsible for gene regulation under nitrogen deprivation scenarios in *Escherichia coli* (Ninfa and Magasanik 1986; Jiang et al. 2003). The role of another soluble HK termed HoxJ has also been established in *Ralstonia eutropha* as being part of a regulation mechanism for hydrogenase genes in response to H_2 (Lenz and Friedrich 1998; Kleihues et al. 2000). Interestingly, *hoxJ* is invariably found adjacently to *hoxBC* genes, which encode two proteins that help assemble the sensory/transduction complex HoxJ-BC necessary for proper H_2 perception (Mascher et al. 2006). This is unsurprising given the typical colocalization of functionally related genes in the same gene-neighborhoods found in prokaryotic genomes (Qi et al. 2010). HKs of the third group are characterized by the conservation of 2–20 TMRs that actively enable stimulus perception often associated with the membrane itself (e.g., mechanical stress) (Mascher et al. 2003, 2006).

In this report, we predict two novel domains structurally related respectively to DHp and REC in bacteria. In addition, the extensive colocalization observed in genes that encode these domains supports the possibility of these domains comprising a novel TCS. These genes also tend to be neighbored by restriction–modification (R–M) systems, which provided further insights on their putative biological role.

Results

Establishing Homologous Sets and Inspecting Genomic Contexts

On the course of a large-scale gene expression analysis conducted by our group on potato tubers using the highly virulent plant pathogen *Pectobacterium brasiliense* strain PBR 1692 (*Pb*1692), three adjacent uncharacterized genes were detected exhibiting significant activation (Bellieny-Rabelo et al. 2019, 2020).

The context in which these genes were up-regulated involves different stresses that may indicate their recruitment to either support or participate in bacterial competition, or interaction with the eukaryotic host. One of these three genes (*PCBA_RS21900, locus tag in the previous version of the genome from NCBI; AED-0003799 updated locus tag from* https://asap.genetics.wisc.edu/, last accessed February 25, 2021) was identified as a member of the GHKL superfamily. The encoded product of *PCBA_RS21900* exhibits two HATPase domains (Pfam: PF02518; PF13589) (Dutta and Inouye 2000). These domains are typically found in Hsp90 chaperones, DNA-gyrases, MutL, and HK family members (Dutta and Inouye 2000). The other two neighboring genes (*PCBA_RS05725-05720, previous* version's locus tags; *AED-0003798 updated locus tag that merges those from the previous version*), contrarily, did not harbor any known conserved domains. With further inspection of the *PCBA_RS05725-05720* homologs in other strains from the same organism (i.e., LMG21371, PcbHPI01) through nucleotide alignment, we found that those two hypothetical entries probably comprise in reality one single gene, which may result from a minor misassembly and/or gene-prediction error in the *Pb*1692 genome. Curiously, these homologs from LMG21371 and PcbHPI01 strains were also neighbored by GHKL superfamily members. Hence, aiming to improve the accuracy in the subsequent computational analyses involving sequence comparisons, the sequence from *Pb* LMG21371 (*KS44_RS10365*) was used instead of those from *Pb*1692. The goal here was to provide an assessment of the frequency of genomic colocalization between *KS44_RS10365*-related genes and GHKL superfamily members in other bacterial genomes.

To perform an initial gene-neighborhood screening on KS44_RS10365-related sequences, we collected a wide range of similar protein sequences from prokaryotes in an extensive search on the NCBI database (see Materials and Methods for details). From the initial sequence search, 677 bacterial strains returned positive hits for KS44_RS10365, and their respective genomic data were obtained to conduct subsequent analyses (supplementary table S1, Supplementary Material online). Next, to predict the orthology relationship among the 441,897 sequences collected from the 677 strains, a sequence clustering step was conducted using OrthoMCL (Li et al. 2003). The sequence clustering analysis produced 23,397 distinct orthologous groups (OG) in total, which were populated by ranges of 2–990 sequences. This approach was paralleled by conserved domains inspection (Eddy 2009) in the same protein data set. Conserved domains assessment was used to provide functional annotations, and to carefully check the cohesiveness of domain architectures within the clustered sequences in the OGs.

The integration of domains and OGs annotation with genomic coordinates enabled extensive gene-neighborhood analysis on the genes encoding the OG_4 members, which harbors KS44_RS10365 and its predicted orthologs. The OG_4 sequences spread through 628 different strains in which 30 conserve two to three paralogs, totaling 664 proteins (supplementary table S2, Supplementary Material online). Out of these 664 paralogs, 20 were in regions unsuitable for contextual analysis due to incompletion of the respective genome assemblies, which resulted in the lack of necessary data in the region of interest. In the 644 analyzed regions (after the exclusion of the 20 unsuitable regions), we observed a remarkable linkage of 91.6% between OG_4 and GHKL superfamily members (OG_5), in a distance of up to three genes. Next, by scanning a broader region including ten genes up/downstream to OG_4 homologs, the HATPase-containing genes (GHKL) were among the most prevalently detected domains (fig. 1). In the OG_4 members' genomic neighborhoods, we also found the dominant presence of DNA_methylase (Pfam: PF00145) containing genes, frequently located upstream of the OG_4 homologs. This strong colocalization raised the possibility of the association of OG_4 members with R–M systems, which will be addressed in detail in a subsequent topic. Other domains associated with two-component signal transduction systems, such as HiskA (Pfam: PF00512) and Response_reg (Pfam: PF00072) were also frequent in the OG_4 members' regions (fig. 1). The presence of immunity- and polymorphic toxin-related domains such as CdiI_3 (Pfam: PF18616) and Ntox28 (Pfam: PF15605) in OG_4 members' vicinity indicates that these genes can frequently be found in association with toxin–antitoxin systems. Additionally, several genes located in these regions for which no known domains could be detected were grouped in orthologous groups, such as the members of OG_20, OG_16, OG_24, OG_8, and OG_294 (fig. 1). The traceable evolutionary relationship among these uncharacterized genes can potentially drive the discovery of novel functionalities in subsequent studies. Together, these observations revealed a strong colocalization pattern among OG_4 and OG_5 (GHKL-containing) members in a wide range of bacteria as well as the most important functional groups to which they frequently associate. Below, we lay emphasis on sequence analysis to examine other conserved features in OG_4/5 groups aiming to determine which biological role they might undertake.

Domain Architectures Characterization

As a preliminary step, we performed a sequence analysis step using the OG_5 sequences in comparison with the major GHKL families (i.e., Hsp90, DNA-gyrases, MutL, and HK). This revealed that OG_5 sequences only produced results when aligned with HKs. This preliminary observation, in combination with the remarkable trend for colocalization of OG_4 and 5 indicated that OG_5 and OG_4 could comprise a TCS. Hence, taking advantage of diverse REC and DHp domain profiles publicly available (Finn et al. 2010), members of

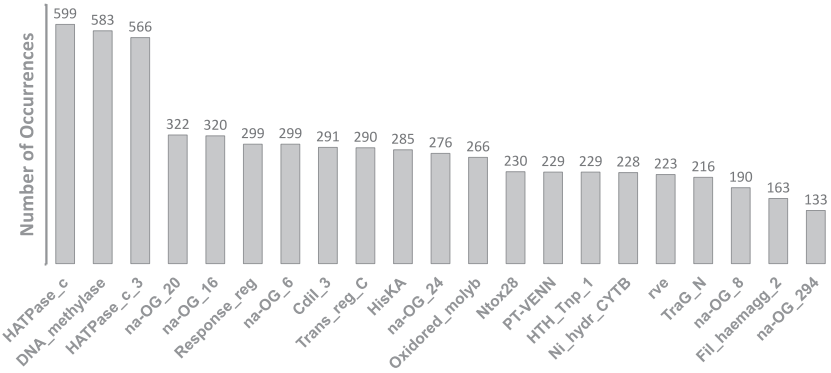

Fig. 1.—Top 1% best-represented protein domains encoded in OG_4 members' genomic regions. In total 628 genomes containing OG_4 genes were analyzed ten genes up- and downstream. The functional domains encoded by each locus in that range were computed and the total amounts are displayed in the bar plot. Domain names were detected by the hmmscan tool and are annotated according to the Pfam database. For each gene product, in the absence of known domains (na—not applicable), the respective orthologous group label (prefix: OG_; suffix: numeric tag) is displayed.

OG_4 and OG_5 were analyzed along with those profiles to verify the possible presence of similar features in the multiple-sequences alignments (MSA). These analyses revealed two previously unknown conserved regions respectively in OG_4 and OG_5 bearing strong similarities with REC and DHp domains.

The most pronounced feature of REC domains is the active site Asp residue (Stock et al. 1989), which was remarkably conserved with 100% consensus found across representative OG_4 sequences (fig. 2*A*). The active site Asp was also consistently conserved within the C-termini ends of β_3. It has also been reported that two consecutive negatively charged residues located N-terminal to the active site are involved in cofactor coordination (Stock et al. 1993). These residues were strikingly conserved in the C-terminal end of β_1 both in OG_4 and in the canonical REC structure (fig. 2*A* and *B*). The Thr column located at the C-termini of β_4 comprises another residue known to be conserved in REC domains, which was also shown to be highly conserved in the OG_4 representatives exhibiting a high consensus level within this group. Additionally, the C-terminal end of β_5 has a conserved Lys residue in the canonical REC, which was also observed in the OG_4 MSA. Although there was a lowly supported split of β_5 structure into 2 β sheets, here, the C-terminal end of the second sheet seems highly preserved in OG_4 sequences exhibiting strong conservation of the Lys residue similarly to the REC domain (fig. 2*A*). This same β_5 split probably reflects the difficulty in secondary structure prediction for some sequences in the MSA, which is not frequently observed when some of the sequences are inspected individually (fig. 2*B*). Notably, the tolerance in OG_4 sequences toward loop extensions interspersing the β/α structures, especially between α_1/β_1 and α_5/β_5, have presumably deepened the evolutionary distance to the canonical Response_reg (REC domain) resulting in the solid evolutionary separation between these motifs (fig. 2*A* and *C*). Although the OG_4 conserved motif and Response_reg have diverged, the topology found in these domains exhibits remarkable similarity in the characteristic $(\beta\alpha)_5$ regulatory fold. This observed $(\beta\alpha)_5$ also exhibits key amino acid residues in identical positions to those reported in the canonical domain structure of REC domains.

On the other hand, domain alignments of OG_5 members suggested a close relatedness between a conserved region within these sequences and the HiskA domain (Pfam: PF00512). Thus, by framing OG_5 and HiskA-containing representative sequences, the remarkable conservation of core His residue followed by a negatively charged residue (Glu or Asp) was observed (fig. 3*A*). The two predicted α-helices in the analysis comprised the characteristic DHp found in H-boxes, in which α_1 harbored the core His residue (fig. 3*A* and *B*). Four randomly picked models of OG_5 sequences depicted the presence of the core His residue in α_1 (fig. 3*B*). Such strong conservation of H-box helices constituted important functional evidence because this structure is responsible for both dimerization among HKs and phosphorylation of the active site Asp residue in the cognate RR. The phylogenetic reconstruction of H boxes from representative sequences of four canonical DHp domain families (see Materials and Methods for details) suggested that the domain found in OG_5 sequences might have evolved from HiskA (fig. 3*C*). These two closely related domains observably conserved a negatively charged residue (Glu or Asp) adjacent to the core His (fig. 3*A*). Since HKs attachment to the membrane, or lack thereof, is intrinsically associated with their function, we quantitatively compiled TMR predictions on representative sequences carrying those same four DHp domains (i.e., HiskA, HiskA_2, HiskA_3, and HWE_HK). The results revealed a strong preference by HiskA_3 representatives to conserve six TMRs, whereas most of HWE_HK, HiskA_2, and OG_5 representative members tend to be found in soluble HKs (fig. 3*D*). In OG_5 proteins specifically, none of the sequences exhibited positive TMR predictions, which suggests that these group members function as soluble proteins. Taken together, these results strongly suggest that members of OG_4/5

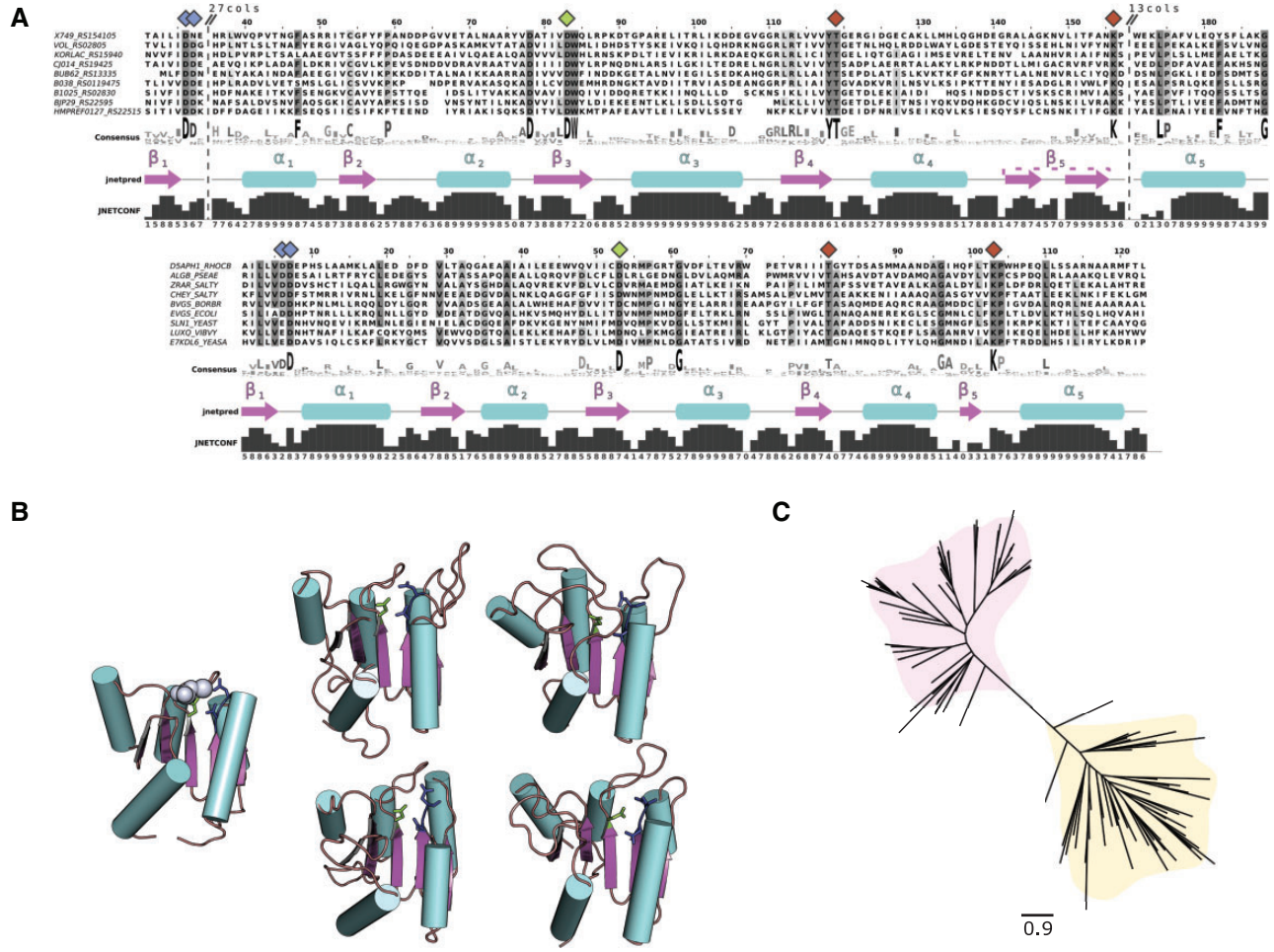

FIG. 2.—Multiple sequence alignment, structure scaffold, and phylogenetic reconstruction of a conserved motif found in OG_4 members. (*A*) Multiple sequences alignments depicting the conserved regions in the REC domains in representative OG_4 (top) and Response_reg-containing (bottom) (Pfam: PF00072) proteins. Dashed lines in the alignments represent nonconserved loops hidden from the alignment. The numbers on the top indicate the number of amino acid columns not shown. Diamond shapes on top of specific columns represent key amino acids in the typical REC domain structure. Corresponding colors of diamond shapes in the top and bottom alignments represent homologous amino acids in both OG_4 and Response_reg-containing sequences. Numbered secondary structure (ss) predicted by JPred are respectively represented below each alignment. Confidence levels of ss predictions provided by JPred are displayed below the graphical ss representations. (*B*) Structures predicted by RaptorX represented as cartoons depict the canonical REC domain from CheY (PDB ID: 2FKA) with the Mg^{2+} highlighted in light purple on the left, and four examples of the newly described DHp domains from OG_4 representatives on the right: BJP29_RS22595 from *Bacteroides fragilis*; H147_RS0115650 from *Loktanella vestfoldensis*; B038_RS0119475 from *Martelella mediterranea*; VOL_RS02805 from *Vibrio owensii*. Three key Asp amino acids known to play a role in Mg^{2+} coordination are highlighted in the cartoons using the same color patterns as in the correspondent diamond shapes in (*A*). (*C*) Phylogenetic reconstruction of the Response_reg domains (yellow) from Pfam seed sequences (PF00072) and the predicted domains in OG_4 sequences (pink) is represented as an unrooted radial tree.

groups could comprise a cytoplasmic TCSs. The two respective novel domains found in OG_4 and OG_5 will be hereafter referred to as Response_reg_2 and HEF_HK, and the sequences carrying these domains RR- and HK-like, respectively.

The Association of RR- and HK-Like with R–M Systems

Next, to gain further insight into the members of both families through "guilty by association" inference, we conducted in-depth genomic context analysis in these RR-like-harboring regions. By using the previously obtained 644 genomic regions from 628 bacterial strains carrying RR-like orthologs as a reference, we first examined the functional domains encoded by their neighboring genes. This step aimed to inspect a genomic range of ten neighboring genes up- and downstream of the RR-like members. Besides, based on the previous assessment showing the solid presence of C-5 cytosine-specific DNA methylase (C5 Mtase) in the RR-like vicinity, we laid special emphasis on identifying R–M elements. To keep this analysis stringent from the functional prediction perspective (see Materials and Methods for details), we used the "gold standard" library of R–M-associated proteins from

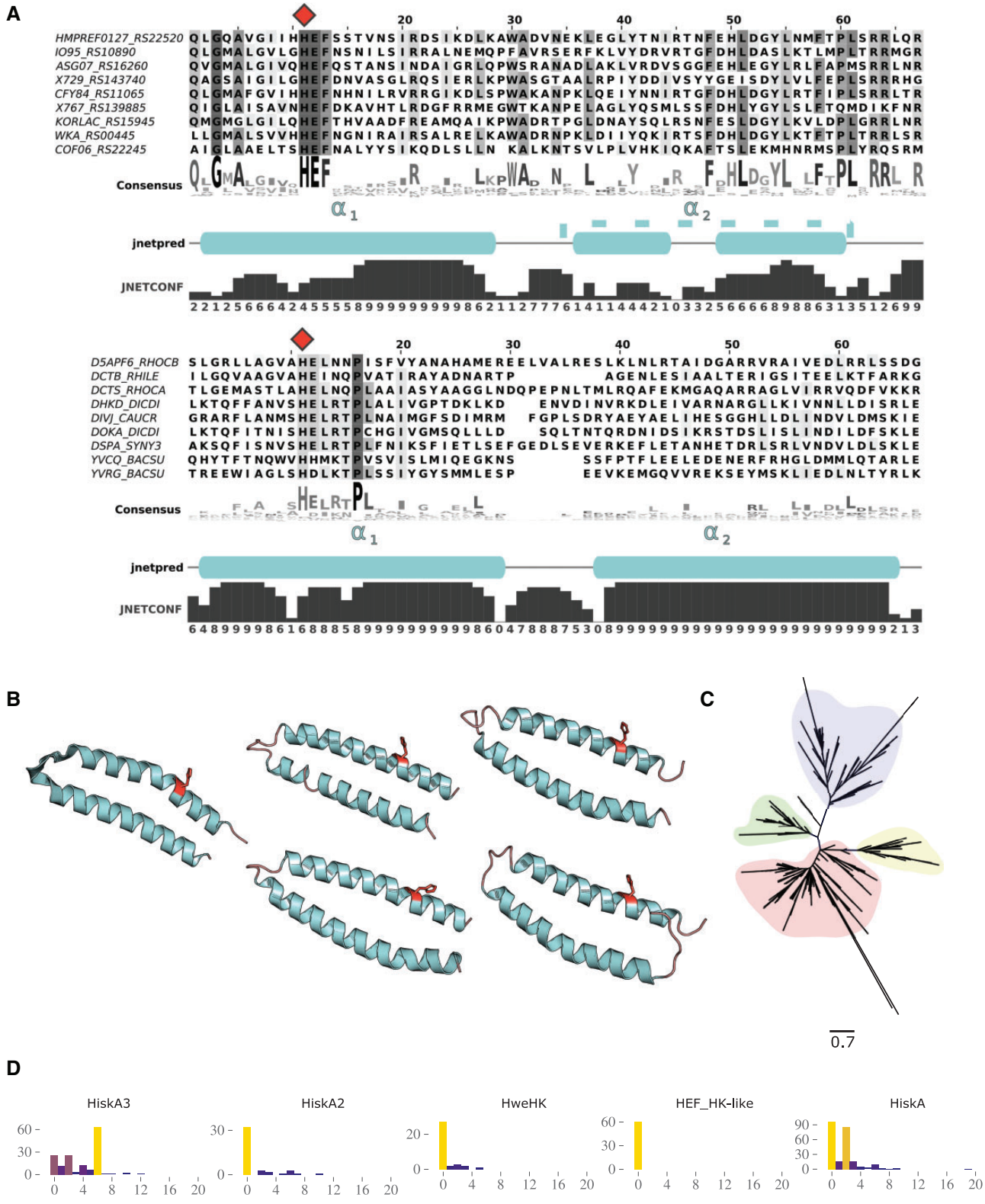

Fig. 3.—Multiple sequence alignment, structure scaffold, and phylogenetic reconstruction of a conserved motif found in OG_5 members. The MSA representation follows the same pattern as in figure 1. (*A*) Red diamond shapes on top of MSA indicate the core His residue in both groups. (*B*) Structural scaffolds represent the α-helices of the canonical HiskA domain (left) (PDB_ID: 5B1N) and four representatives of the newly predicted DHp domain (right): CFY84_RS11065 from *Acinetobacter apis*; HMPREF0127_RS22520 from *Bacteroides* sp. 1_1_30; WKA_RS00445 from *Escherichia coli*; KORLAC_RS15945 from *Kordiimonas lacus*. The core His residue in all five scaffolds are highlighted in red stick representations. (*C*) Phylogenetic reconstruction of HiskA domain variants, including HiskA (red), HiskA2 (green), HWE_HK (green), HiskA3 (blue) from Pfam seed sequences (respectively: PF00512, PF07568, PF07536, PF07730), and the predicted DHp domain (yellow). (*D*) Each histogram depicts the frequency of transmembrane occurrences distribution found in representative sequences from each domain family predicted by TopCons.

the REBASE database (Roberts et al. 2015). The REBASE gold standard set comprises only experimentally characterized components that have been associated with either restriction or modification functions. Aiming to determine which known functional domains are associated with R–M systems, we scanned the entire REBASE gold standard protein set using the Pfam database as a reference. Subsequently, by inspecting the domains encoded by the genes neighboring the RR-like encoding genes and comparing these with the previous results, we found a consistent presence of R–M-related domains.

Next, to evaluate the correlation of R–M systems elements in close proximity to RR-like genes, a computational simulation step was implemented. This technique shuffles the entire gene pools from known genomes, generating "pseudogenomes." By analyzing the results from these randomized pseudogenomes, we could infer up to which frequency particular gene associations might occur by chance, as well as overrepresented associations. Here, we utilized 25 randomly picked complete bacterial genomes from our initial set of 628 RR-like-containing organisms to conduct the simulations. From each of the 25 genomes, 2,000 pseudogenomes were generated. Next, the RR-like neighboring genes (ten genes up- and downstream) from each of the 50,000 simulations were inquired for the occurrence of R–M domains. The number of R–M domains in these regions was computed for each strain. For the vast majority of strains analyzed individually, the pseudogenomes analysis showed that up to three R–M domain occurrences should be expected by chance in the RR-like gene neighborhoods (fig. 4*A*). Furthermore, we found that between five and seven R–M domains were found within a maximum distance of ten genes from the RR-like encoding gene in 41.9% of the 628 bacterial genomes analyzed, whereas 9.9–15% of the 50,000 pseudogenomes exhibited between five and seven R–M domains in the same genomic range. On the other hand, only 9.7% of the 628 genomes exhibit up to two R–M-related domains in the RR-like neighboring regions, whereas 49.3–56.8% exhibits the same frequency of R–M domains in the pseudogenomes. These observations further underscored the conspicuous presence of R–M systems elements in the genomic contexts of the newly described RR-like orthologs. On top of these comparisons, statistical analyses showed that for all strains, the frequency of R–M domains in RR-like neighborhoods exhibited significant contrast (P value <0.0001) compared with the frequency found in the real data set (fig. 4*B*). The nonrandom genomic association of RR-Like and R–M-related genes strongly suggests a mechanistic relationship between these themes in a vast range of bacteria.

By further inspecting the RR-like regions (ten genes up- and downstream) with the REBASE annotation support, we found the prominent presence of type I and especially type II R–M system elements (fig. 5*A*). Although functional domains can often be coopted to perform their roles in different proteins from different systems (and can thus be classified into more than one R–M system type), types I and II are still the most prevalent in these regions. Several nuclease domains from type II R–M systems were found in these vicinities, such as MvaI_BcnI (Pfam: PF15515), ParBc (Pfam: PF02195), EcoRII-C (Pfam: PF09019), ResIII (PF04851), HSDR (PF04313), and variants of the HNH domain (fig. 5*B* and supplementary table S2, Supplementary Material online). Specifically, the HNH_2 (Pfam: PF13391) is the best-represented endonuclease domain in RR-like neighborhoods. The HNH_2 can be found in seven different genera of Gram-negative bacteria: Three from the Bacteroidetes (*Chryseobacterium*, *Flavobacterium*, and *Hymenobacter*) and four from the Proteobacteria (*Klebsiella*, *Pectobacterium*, *Salmonella*, and Yersinia) phyla (fig. 5*B*). Besides the type II-associated domains, the widespread presence of the Vsr endonuclease domain (Pfam: PF03852) was also observed. The Vsr-containing genes encode the very short patch repair (VSR) proteins. VSRs are often classified as V genes, and play a role in correcting mispairs in the DNA that arise following spontaneous deamination of 5-methylcytosine (Bhagwat and Lieb 2002). In this analysis, we found Vsr-containing genes in 20 different genera, from four different phyla (Actinobacteria, Proteobacteria, Bacteroidetes, and Verrucomicrobia) encompassing: four γ-proteobacteria (*Acinetobacter*, *Azotobacter*, *Pseudomonas*, *Shewanella*), five α-proteobacteria (*Agrobacterium*, *Mesorhizobium*, *Pleomorphonas*, *Rhizobium*, *Sphingomonas*), two δ-proteobacteria (*Desulfovibrio*, *Geobacter*), five bacteroidias (*Bacteroides*, *Hallella*, *Hymenobacter*, *Porphyromonas*, *Prevotella*), one chlorobium (*Chlorobium*), one flavobacterium (*Chryseobacterium*), one actinobacterium (*Diaminobutyricimonas*), and one verrucomicrobiae (*Akkermansia*). The most prevalent genomic organization of VSR-encoding genes across the different taxa was that in which VSR genes were adjacently located downstream to the RR-like genes (fig. 5*B*). These analyses underscored the solid association of RR-like with type II R–M systems, which may be preferentially recruited to work along with the newly discovered TCS-like system.

Phyletic Distribution of Protein Architecture Containing the Newly Characterized HK-Like

The next step was to consolidate the newly found domain profiles based on the representative sequences respectively from OG_4 and OG_5. Consolidation of the new profiles enabled us to perform broad domain-oriented searches of Response_reg_2 and HEF_HK conserved motifs against large databases using hmmsearch (Potter et al. 2018). By conducting this sensitive search, we aimed to test the results from the previous analysis and assess the prevalence of the new domains across the bacterial lineages.

Hence, Ensembl, UniProtKB, and Reference proteomes (uniprotrefprot) were inspected for the presence of the two

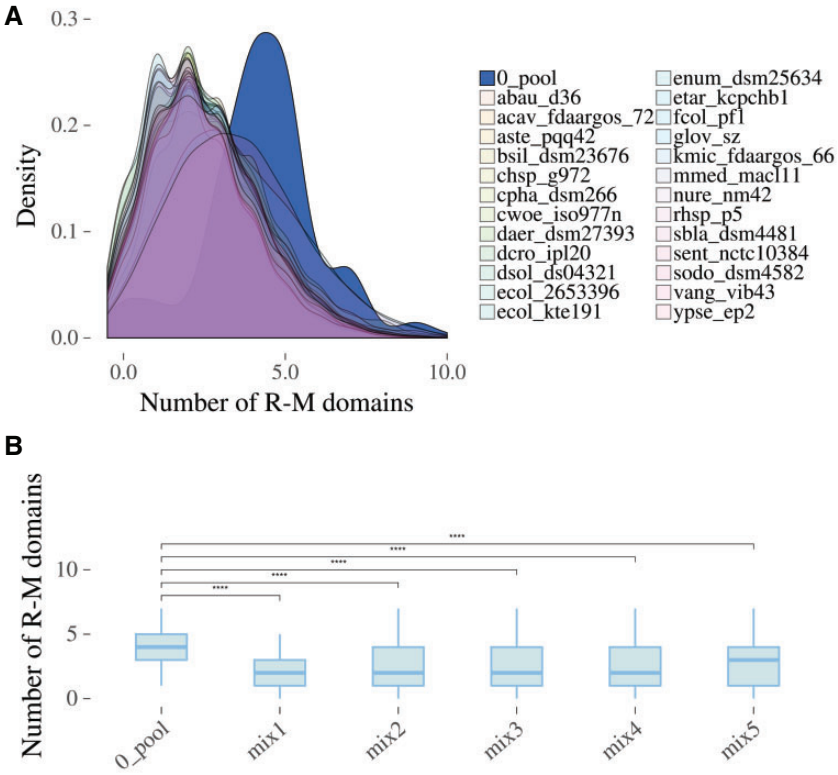

Fig. 4.—Distribution of R–M-related domains encoded in RR-like gene neighborhoods. (*A*) The density plot compares the distribution of genes encoding R–M domains found in close proximity to RR-like genes between the data set representing 628 actual bacterial genomes (0_pool) and the respective 2000 shuffled versions (pseudogenomes) of 25 different genomes gleaned from the 0_pool. Species and strain names are abbreviated as follows: *Acinetobacter baumannii* (abau_d36), *Aeromonas caviae* (acav_fdaargos_72), *Alteromonas stellipolaris* (aste_pqq42), *Brevibacterium siliguriense* (bsil_dsm23676), *Chryseobacterium* sp. G972 (chsp_g972), *Chlorobium phaeobacteroides* (cpha_dsm266), *Conexibacter woesei* (cwoe_iso977n), *Diaminobutyricimonas aerilata* (daer_dsm27393), *Devosia crocina* (dcro_ipl20), *Dickeya solani* (dsol_ds04321), *Escherichia coli* (ecol_2653396), *Escherichia coli* KTE191 (ecol_kte191), *Endozoicomonas numazuensis* (enum_dsm25634), *Edwardsiella tarda* (etar_kcpchb1), *Flavobacterium columnare* (fcol_pf1), *Geobacter lovleyi* (glov_sz), *Klebsiella michiganensis* (kmic_fdaargos_66), *Martelella mediterranea* (mmed_macl11), *Nitrosomonas ureae* (nure_nm42), *Rhodovulum* sp. P5 (rhsp_p5), *Shimwellia blattae* (sbla_dsm4481), *Salmonella enterica* subsp. *enterica* serovar Senftenberg (sent_nctc10384), *Serratia odorifera* (sodo_dsm4582), *Vibrio anguillarum* (vang_vib43), *Yersinia pseudotuberculosis* (ypse_ep2). (*B*) The box plots depict the statistical analysis using one-tailed Student's *t*-test between the 0_pool and five compilations ("mix") of the 25 pseudogenomes previously simulated. Each mix contains five strains. *t*-Test results are represented as follows: ns ($P > 0.05$), * ($P \leq 0.05$), ** ($P \leq 0.01$), *** ($P \leq 0.001$), **** ($P \leq 0.0001$). The mix data sets include the following above mentioned genomes: mix1—abau_d36, acav_fdaargos, aste_pqq42, bsil_dsm23676, chsp_g972; mix2—cwoe_iso977n, dear_dsm27393, dcro_ipl20, dsol_ds04321, ecol_2653396; mix3—ecol_kte191, enum_dsm25634, etar_kcpchb1, fcol_pf1, glov_sz; mix4—kmic_fdaargos, mmed_macl11, nure_nm42, rhsp_p5, sbla_dsm4481; mix5—cpha_dsm266, sent_nctc10384, sodo_dsm4582, vang_vib43, ypse_ep2.

domains, which returned respectively from these databases: 529, 163, and 50 hits for Response_reg_2 and 199, 139, and 52 hits for HEF_HK. As expected, based on our first analysis, the Response_reg_2 domains were detected in proteins lacking any other currently known domains. On the other hand, the HEF_HK proteins exhibited diversity in terms of associated known domains. These sequences contained GHKL superfamily motifs (i.e., HATPase_c, and HATPase_c_3), which strikingly confirmed the pattern observed in the first section of this study. Furthermore, the variable presence of functional domains could be exploited to glean additional information on the phyletic distribution of the HEF_HK-containing proteins. To achieve that, all HEF_HK significant hits found by hmmsearch in the three databases mentioned above were acquired and then merged with the OG_5 sequences from the previous analyses. Next, this merged data set was filtered to remove redundant sequences, which typically belong to different strains of the same species. This step prevented the inflated representation of species that were prevalent in public databases. This resulted in a data set including unique sequences to be further analyzed. Next, aiming to ensure that only bona fide HEF_HK-containing proteins were to be included, we performed an additional in-house domain profile recognition on the unique sequences set with a stringent setting (e-value: 1e-05) using hmmscan (supplementary table S3, Supplementary Material online). The results from the 222 unique bona fide HEF_HK-containing sequences showed that most of these were found in organisms from the

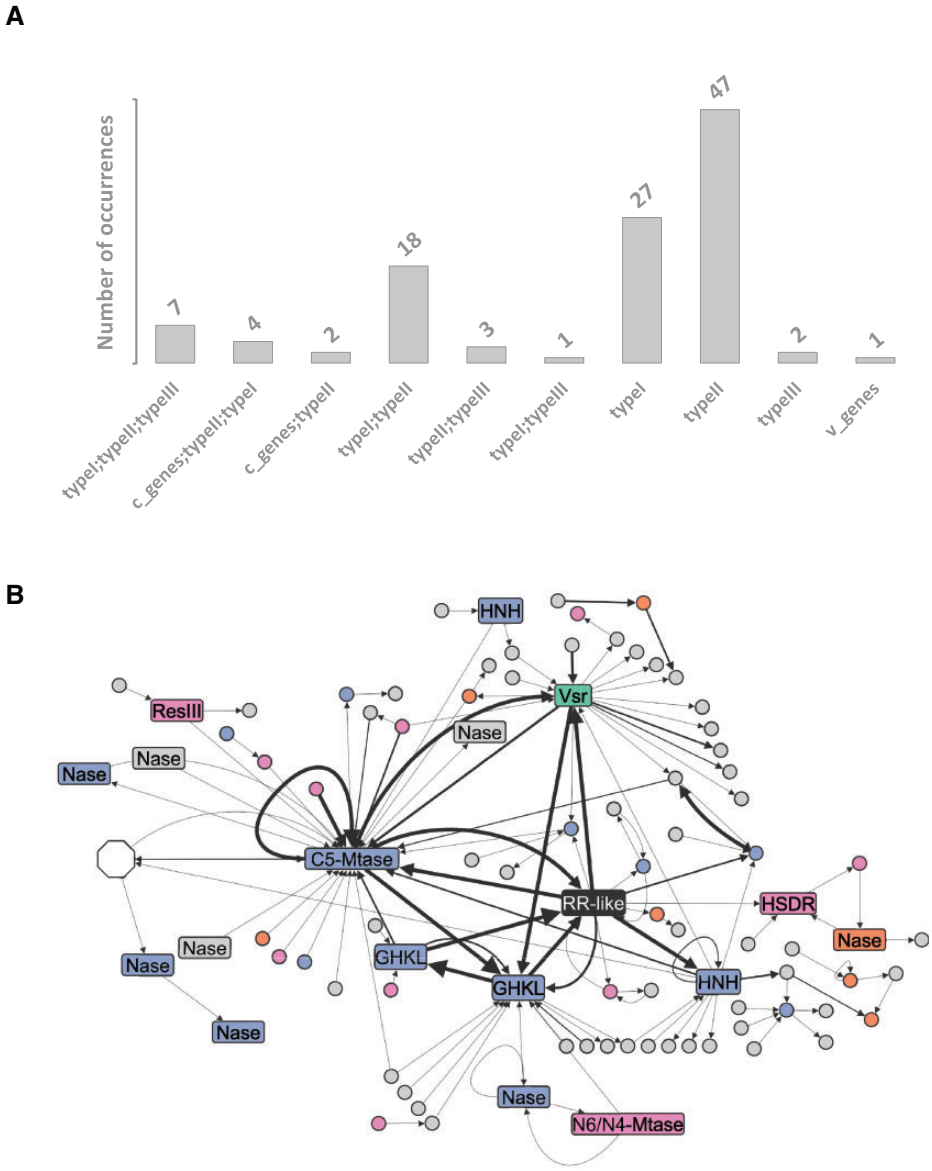

Fig. 5.—Classification of R–M systems identified in RR-like genomic regions. (*A*) The occurrence of distinct R–M types found in ten genes up- and downstream of RR-like genes according to REBASE based on protein domains analysis. (*B*) Network representation using a subset (3 genes up- and downstream) of the analysis presented in (*A*) for improved visualization. The network shows the encoded domains with at least one R–M-related neighboring domain in the RR-like vicinities. Neighboring domains are linked with an arrow representing the order of appearance (5′ -> 3′) in the respective genomic regions. Line thickness is proportional to the number of genera in which a particular domain neighborhood was found. Different types of R–M systems according to the REBASE database are highlighted in blue (type II), orange (type I), green (v genes), or pink (domains assigned to more than one R–M type). A white hexagonal shape represents 57 different non-R–M domains neighboring the C5-Mtase domain. Different gene families are labeled/abbreviated as follows: The Response_reg_2 encoding genes (RR-like), GHKL family domains (GHKL), nuclease domains (Nase), HNH family domains (HNH), very short patch repair domains (Vsr), DNA_methylase domain encoded in C-5 cytosine-specific DNA methylase (C5-Mtase), N6_N4_Mtase domain encoded by N-6 adenine-specific DNA methylase or N-4 adenine-specific DNA methylase (N6/N4-Mtase), and ResIII or HSDR domains encoded by restriction enzymes (respectively ResIII and HSDR).

Proteobacteria phylum (69.8%). Furthermore, it underscored the remarkable association of this domain with two other domains: 1) the HATPase_c_3, found within N-terminal regions (83.3%), and/or the 2) HATPase_c within C-terminal regions (93.7%) (supplementary table S3, Supplementary Material online). In this context, the predominant domain organization observed is the one including HATPase_c_3+HEF_HK+HATPase_c (group 1), represented in 174 out of 222 sequences (fig. 6*A*). Followed by group 2 (34/222), which has lost the N-terminal HATPase_c_3 (i.e.,

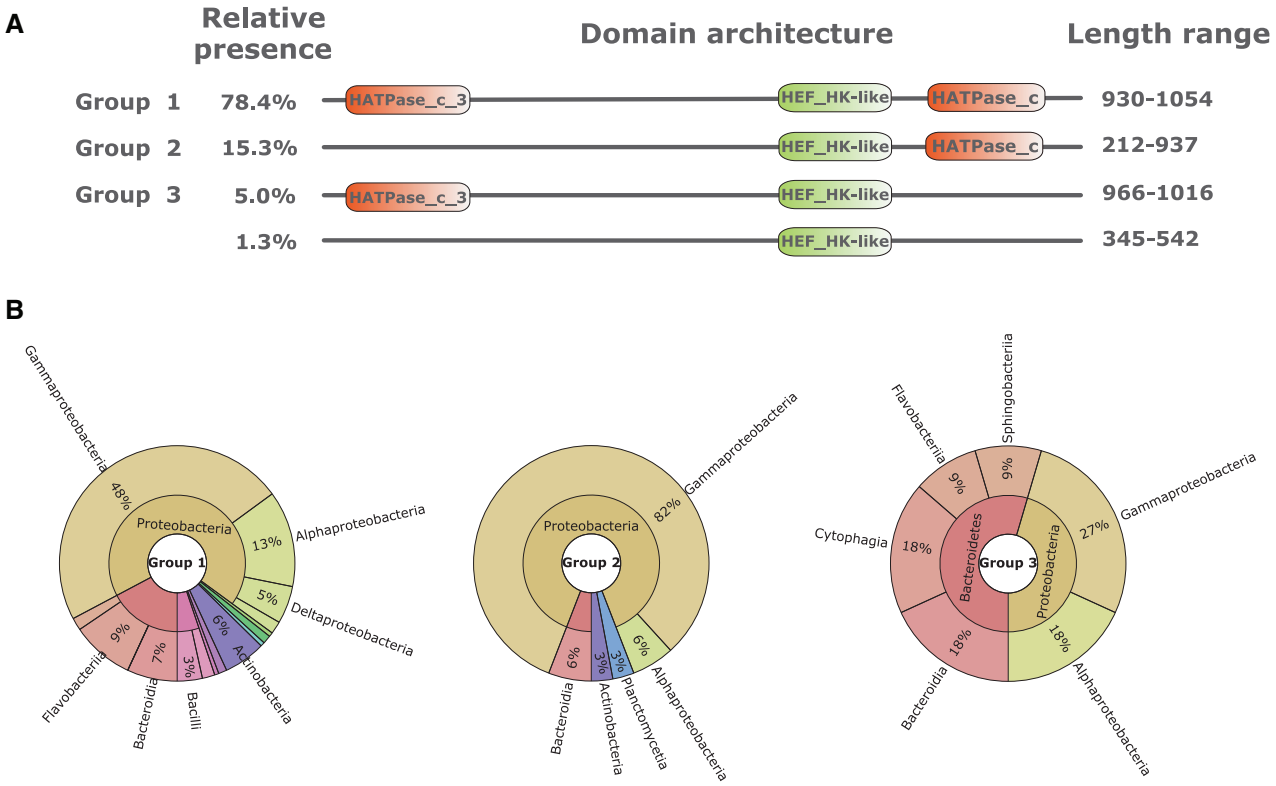

Fig. 6.—HK-like protein domain architectures presence in bacterial lineages. (*A*) Different architectures found in the 222 unique bona fide HEF_HK-containing sequences retrieved from Ensembl, UniProtKB, and Reference proteomes (uniprotrefprot). The three prevalent architectures are labeled as Group 1, 2, and 3 and their relative presence in the entire set of sequences is adjacent to the group labels. Domains recovered from the Pfam database are represented in orange, whereas the newly consolidated HEF_HK appears in green. The longest and shortest sequence lengths of each group are depicted adjacent to each architecture. (*B*) The phyletic distribution of the three prevalent architecture groups is depicted by composite pie charts.

HEF_HK+HATPase_c). Group 3 (11/222), instead, comprised those architectures in which the C-terminal HATPase_c was lost (i.e., HATPase_c_3+HEF_HK). As the most unusual architecture, group 4 (3/222) conserved only the HEF_HK, with no other known domain.

Notably, group 1 was the only architecture found in HEF_HK-containing sequences conserved in Betaproteobacteria, Deltaproteobacteria, or in the Firmicutes phylum (fig. 6*B* and supplementary table S3, Supplementary Material online). Numerous orders of the Alphaproteobacteria class were shown to conserve any of the three major groups (i.e., 1, 2, and 3). Among these orders, Rhodobacterales seemed to obligatorily conserve the C-terminal signature (HATPase_c), compared with the N-terminal HATPase_c_3 may not be essential. Contrarily, in the Rhizobiales order, the N-terminal HATPase_c_3 seems to be mandatory. The Gammaproteobacteria class encompasses most of the organisms carrying HEF_HK-containing sequences (51.8%). Within this class, bacteria from Yersiniaceae (including *Yersinia* and *Serratia* spp.) and Pectobacteriaceae (including *Pectobacterium* and *Dickeya* spp.) families are constricted to group 1. Conversely, Moraxellaceae and Vibrionaceae families may conserve either group 1 or 2, meaning that the C-terminal HATPase_c is essential in these organisms' sequences. These results underscore the wide diversity of organisms in which the HEF_HK domain is conserved and the contrasting conservation for the companion HATPase domains in different phyletic groups.

Discussion

In this study, two novel conserved domains were characterized: HEF_HK and Response_reg_2. The newly described domains conserve characteristic secondary structure organization canonically described DHp and REC domains found respectively in HKs and RRs. In HK proteins, although several DHp domain variants have been described, some key features must remain unaltered such as the structure of the kinase core and the adjacent nucleotide-binding fold (Robinson et al. 2000). The archetypal organization among the highly conserved domains of HK proteins includes five motifs known as H, N, G1, F, and G2 boxes (Tanaka et al. 1998). The H box harbors the core His residue and remains typically conserved adjacently to the C-terminal nucleotide-binding pocket which encompasses the N, G1, F, and G2 boxes (Tomomori et al. 1999). As a highly conserved structure, the C-terminal

CA domain containing the nucleotide-binding pocket could be promptly detected in the sequences from our acquired data set by exhibiting the HATPase signature. Within these sequences, the structure reminiscent of an H box could be detected adjacently to the C-terminal HATPase_c domain (fig. 6*A*). Of the sequences carrying a HEF_HK domain, 83.4% also conserved an N-terminal HATPase_c_3 (Pfam: PF13589) domain. The HATPase_c_3 domain is classified into the GHKL superfamily similarly to HATPase_c (Pfam: CL0025). By looking at the complete set of sequences carrying the HATPase_c_3 domains, the strikingly strong preference for N-terminal localization was clear, contrary to the C-terminal preference observed for the HATPase_c domain (Finn et al. 2010). The N-terminal localization of HK-associated HATPase_c_3 occupies the classic site of sensor domains in HK proteins (Szurmant et al. 2007). Intriguingly, none of the canonical DHp domains can be found in HATPase_c_3 containing proteins (Finn et al. 2010). These observations could suggest specific recruitment of HATPase_c_3 domains to play a role as a sensor in the newly predicted HEF_HK containing HKs. Since the HEF_HK containing proteins are predominantly soluble (fig. 2*D*), it is possible to speculate that HATPase_c_3+ HEF_HK containing proteins could be involved in interactions with intracellular concentrations of nucleoside triphosphate molecules (NTP).

Characterization of TCSs utilizing sequence analysis and extensive comparative genomics has been largely exploited in the last decades, comprising a powerful strategy for the identification of new systems (Mascher 2006; Wecke et al. 2006; Revilla-Guarinos et al. 2013). Besides revealing the solid linkage between genes encoding HEF_HK- and Response_reg_2-containing products, this approach also sheds light on the striking conservation of upstream genes to this duplet encoding methyltransferases (supplementary table S2, Supplementary Material online). The DNA_methylase domains integrate the DNA methyltransferases (DNMT) protein family which became specialized in cytosine-specific methylation (Li et al. 2013). In this context, DNA methylation in prokaryotes has been largely associated with R–M systems, which typically contain two opposing utilities: endonucleases (performs double-strand DNA cleavage) and DNA methyltransferase (prevents double-strand cleavage by the cognate endonuclease) (Marinus and Casadesus 2009). The linkage between R–M systems and predicted ATPase-encoding genes has been previously observed by (Furuta et al. 2010), however, the presence of a predicted response regulator has not been revealed thus far.

The modification systems are effective barriers against bacterial lineages of varied descent which contain different epigenetic identities as defined by their R–M systems (Kobayashi 2001). Since some R–M systems can occasionally target the host genome if eliminated, they are typically referred to as selfish elements as they promote their survival in the cell lineages (Kobayashi 2001; Casadesús and Low 2006; Mruk and Kobayashi 2014). Thus, if the concentration of M proteins is not sufficient to protect the host DNA against cognate R endonucleases in the cell, this may cause cell death (Fukuda et al. 2008). The overrepresented presence of R–M-related domains in the vicinity of HEF_HK-/Response_reg_2-containing genes observed in a wide range of bacterial genomes corroborates that the association between these themes does not occur by chance (fig. 5 and supplementary table S2, Supplementary Material online). Hence, one might speculate that these newly discovered domains could be involved in signaling cascades controlling, directly or indirectly, R–M systems.

The N-terminal presence of the GHKL domain HATPase_c_3 in HEF_HK-containing proteins is a well-established characteristic feature of the Microrchidia (MORC) ATPase family, originally described as a necessary gene for the completion of mammalian spermatogenesis (Watson et al. 1998). Subsequent investigations revealed the presence of MORC family members in prokaryotes (Iyer et al. 2008), as well as in the major crown group lineages of eukaryotes (Iyer et al. 2008; Koch et al. 2017). Although mechanistic details concerning the majority of MORC's functions remain elusive, they constitute a widespread family primarily involved in diverse processes associated with epigenetic regulation (Li et al. 2013). In *Arabidopsis thaliana*, for example, AtMORC1 and AtMORC6 have been implicated in gene silencing through chromatin condensation (Moissiard et al. 2012). Another MORC1 (SlMORC1 from *Solanum lycopersicum*) was further reported to exhibit similarities with type II topoisomerases (Manohar et al. 2017). Furthermore, the ability to form dimers through the N-terminal GHKL domain was elucidated in the murine MORC3 protein (Li et al. 2016). In contrast to the recent advances in the understanding of eukaryotic MORCs, the molecular and biochemical aspects of the bacterial MORCs remain poorly understood. Hence, it is currently unclear whether the focus in chromatin superstructure manipulation is conserved across the eukaryotic and prokaryotic members. In prokaryotes, the MORCs were classified as part of larger radiation that includes several GHKL proteins termed paraMORCs (Iyer et al. 2008). The paraMORCs were shown to remain preferably in genomic regions coinhabited by R–M systems, endonucleases, and helicases (Iyer et al. 2008).

The striking similarity of sequence and contextual signatures between the paraMORCs and HEF_HK-encoding genes strongly suggests that this newly found HK-like family may have branched from the paraMORCs through the acquisition of the C-terminal HATPase_c and HEF_HK domains. These findings open an opportunity for further evaluation of the probable regulatory outspread of these genes in a vast number of bacterial taxa, such as via transcriptomics and/or proteomics of mutant strains. Besides, the ability to ascertain which environmental cues or chemicals can trigger a response transduced by this system would also comprise and important

step toward its characterization. Altogether, this study revealed two novel protein conserved domains and shed light on the possibility of a TCS-like system undertaking a regulatory role mechanistically linked to R–M systems in a large variety of bacterial lineages.

Materials and Methods

Sequence Search, Orthology, and Gene-Neighborhood Screenings

The initial search for RR-similar sequences was conducted by using TBlastN 2.8.0+ (Gertz et al. 2006) using default parameters in five iterations on the RefSeq Genomes database available on NCBI webpage (https://www.ncbi.nlm.nih.gov, last accessed February 25, 2021), allowing up to 1,000 positive hits restricted to Bacteria (taxid: 2). The query used for this search is 600 aa long and was obtained from *Pectobacterium carotovorum* subsp. *brasiliense* (KS44_RS10365). A total of 684 positive hits returned, for which 628 came from bacterial strains with support of genome-wide information. These data sets were then obtained, enabling subsequent gene-neighborhood screenings in the strains in which the hits were found. Full records in GenBank format of the 628 entries representing their respective genomic constructs (complete genome, scaffold, or contig) containing the KS44_RS10365-similar sequences were downloaded from the RefSeq database. Additional ecological and taxonomical information on these 628 strains was then gleaned from different sources, including NCBI taxonomy-, Uniprot-, and PATRIC bacterial databases (Sayers et al. 2009; Wattam et al. 2014; The UniProt Consortium 2017). Taxonomic records of interest include the bacterial class, order, and family.

Following the preliminary sequence search, two distinct methods were carried out aiming to provide relevant information on the obtained sequences for subsequent gene-neighborhood screening. First, the OrthoMCL (Li et al. 2003) pipeline was applied to predict homology relationships among the sequences by orthologous clustering based on MCL (inflation = 1.5). Each orthologous group (OG) is numerically labeled for easy identification (e.g., OG_1, OG_2, etc.). Second, all sequences were scanned for known conserved domains using the HMMER3 package (Eddy 2009) using default parameters supported by the PFAM domains database (Finn et al. 2010). To combine both sources of information in the gene-neighborhood screenings, the genomic coordinates were then integrated into these results by custom PERL scripts (https://www.perl.org, last accessed February 25, 2021). The identification of R–M-related domains in this analysis was achieved by retrieving the gold standard sequences from the REBASE database (Roberts et al. 2015) and performing in-house detection of functional domains using HMMER just as described above. This information was integrated into the gene-neighborhood screenings, and the resulting network was then generated by using Cytoscape software (Shannon et al. 2003).

Sequence Alignment, Phylogeny, and Domain Analysis

The orthologous sequences of both OG_4 (RRs) and OG_5 (HKs) were aligned by Clustal Omega (Sievers and Higgins 2018) and PROMALS3D (Pei et al. 2008), and visualized by Jalview (Waterhouse et al. 2009) in comparison with canonical REC and DHp domains obtained from Pfam database (https://pfam.xfam.org/, last accessed February 25, 2021). All alignments performed involving canonical REC and DHp domains use the Pfam "seed" entries available in the online repository for each described domain, which includes only representative sequences carrying the respective domains. This approach aimed to comparatively assess the conservation of key RR and HK residues in the sequences from OG_2 and OG_4. Secondary structures were also inferred in the alignments by using JPred4 (Drozdetskiy et al. 2015), and membrane topology predicted by the TopCons server (Tsirigos et al. 2015). To establish a comparison with OG_2 sequences, we obtained the domain families from Pfam clan CL0304 seed sequences carrying a highly conserved Asp residue (PFAM: PF00072—Response_reg; PF16359—RcsD_ABL; PF09456—RcsC) to perform alignments with OG_2 sequences. Aside from the Response_reg domain, the other combined alignments with OG_2 sequences showed poor quality and were hence discarded. Sequences from OG_4 were aligned with domains families from Pfam Clan CL0025 seed sequences carrying highly conserved His residue (PFAM: PF00512—HiskA; PF07568—HiskA_2; PF07730—HiskA_3; PF07536—HWE_HK), resulting in a good quality combined alignment. Individual alignments including only sequences carrying Response_reg_2 and HEF_HK domains respectively were then performed to determine the limits of each domain. These limits were determined based on secondary structure analyses performed using JPred. The Response_reg_2 and HEF_HK domains were then cut from the alignments, and filtered to avoid sequence redundancy, and then supplied to SMS (Lefort et al. 2017) for evolutionary model selection. Next, the domain sequences were supplied to FastTree (Price et al. 2010) to reconstruct the domains phylogeny. The cut domains from Response_reg_2 and HEF_HK were converted into HMM-profiles, which were consolidated by *hmmbuild* and *hmmpress* from the HMMER3 package. After the domain profiles were consolidated (Response_reg_2 and HEF_HK), these were used in a high sensitivity search in public databases through *hmmersearch* under default parameters (Potter et al. 2018). The positive matches were obtained, merged with the sequences from the respective orthologous groups, and then filtered to avoid sequence redundancies. The nonredundant sets of Response_reg_2 and HEF_HK were then subjected to in-house domain scanning under stringent criteria (e-value

<1e-05) using *hmmscan*, to avoid low-confidence domain predictions. The 3D model predictions of the conserved motifs in RR-like and HK-like were carried out by using RaptorX (Kallberg et al. 2014) which selects the most suitable PDB structure for each analyzed sequence and uses it as a reference. Four randomly picked representative sequences from both the RR-like (BJP29_RS22595—PDB ID: 2ZWM; H147_RS0115650—PDB ID: 1A2O; B038_RS0119475 PDB ID: 3Q9S; VOL_RS02805 PDB ID: 3Q9S) and the HK-like (CFY84_RS11065—PDB ID: 5OO7; HMPREF0127_RS22520—PDB ID: 5OO7; WKA_RS00445—PDB ID: 5OO7; KORLAC_RS15945—PDB ID: 5JRW) groups were analyzed using this method. Subsequent visualization, image rendering of domain models used Pymol (https://pymol.org/, last accessed February 25, 2021), in combination with EnvZ (PDB ID: 5B1N) and CheY (PDB ID: 2FKA) obtained from PDB online repository (Berman et al. 2000) for comparison.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

We would like to acknowledge the expert assistance of Dr Divine Y. Shyntum with insights on microbial genetics and genomics in the early stages of this work. This research was funded by the National Research Foundation (NRF), South Africa through Competitive Funding for Rated Researchers (CFRR) 98993. W.J.S.P. studentship was supported by The Research Technology Fund (RTF) 98654. D.B.-R. was supported by the University of Pretoria Postdoctoral Fellowship. Any findings and/or recommendations expressed here are those of the author(s) and the NRF does not accept any liability in this regard.

Author Contributions

Conceptualization: D.B.-R. and L.N.M. Methodology: D.B.-R., W.J.S.P., and L.N.M. Formal analysis: D.B.-R. and W.J.S.P. Writing—original draft preparation: D.B.-R. and W.J.S.P. Writing—review and editing: D.B.-R and L.N.M. Funding acquisition: L.N.M. All authors have read and agreed to the published version of the manuscript.

Data Availability

The protein domain profiles described in this study are available online on the PFAM database and can be found respectively on these addresses: https://xfamsvn.ebi.ac.uk/svn/pfam/trunk/Data/Families/PF19191/ (last accessed February 25, 2021) (HEF_HK) and https://xfamsvn.ebi.ac.uk/svn/pfam/trunk/Data/Families/PF19192/ (last accessed February 25, 2021) (Response_reg_2).

Literature Cited

Alberts B. 2017. Molecular biology of the cell. New York: Garland Science.
Aravind L, Ponting CP. 1999. The cytoplasmic helical linker domain of receptor histidine kinase and methyl-accepting proteins is common to many prokaryotic signalling proteins. FEMS Microbiol Lett. 176(1):111–116.
Batchelor JD, et al. 2008. Structure and regulatory mechanism of *Aquifex aeolicus* NtrC4: variability and evolution in bacterial transcriptional regulation. J Mol Biol. 384(5):1058–1075.
Bellieny-Rabelo D, Nkomo NP, Shyntum DY, Moleleki LN. 2020. Horizontally acquired quorum-sensing regulators recruited by the PhoP regulatory network expand the host adaptation repertoire in the phytopathogen *Pectobacterium brasiliense*. mSystems 5(1):e00650–19.
Bellieny-Rabelo D, et al. 2019. Transcriptome and comparative genomics analyses reveal new functional insights on key determinants of pathogenesis and interbacterial competition in *Pectobacterium* and *Dickeya* spp. Appl Environ Microbiol. 85(2):e02050–02018.
Bergerat A, et al. 1997. An atypical topoisomerase II from Archaea with implications for meiotic recombination. Nature 386(6623):414–417.
Berman HM, et al. 2000. The Protein Data Bank. Nucleic Acids Res. 28(1):235–242.
Bhagwat AS, Lieb M. 2002. Cooperation and competition in mismatch repair: very short-patch repair and methyl-directed mismatch repair in *Escherichia coli*. Mol Microbiol. 44(6):1421–1428.
Bilwes AM, Alex LA, Crane BR, Simon MI. 1999. Structure of CheA, a signal-transducing histidine kinase. Cell 96(1):131–141.
Casadesús J, Low D. 2006. Epigenetic gene regulation in the bacterial world. Microbiol Mol Biol Rev. 70(3):830–856.
Drozdetskiy A, Cole C, Procter J, Barton GJ. 2015. JPred4: a protein secondary structure prediction server. Nucleic Acids Res. 43(W1):W389–W394.
Dutta R, Inouye M. 2000. GHKL, an emergent ATPase/kinase superfamily. Trends Biochem Sci. 25(1):24–28.
Eddy SR. 2009. A new generation of homology search tools based on probabilistic inference. Genome Inform. 23(1):205–211.
Finn RD, et al. 2010. The Pfam protein families database. Nucleic Acids Res. 38(suppl_1):D211–D222.
Fukuda E, Kaminska KH, Bujnicki JM, Kobayashi I. 2008. Cell death upon epigenetic genome methylation: a novel function of methyl-specific deoxyribonucleases. Genome Biol. 9(11):R163.
Furuta Y, Abe K, Kobayashi I. 2010. Genome comparison and context analysis reveals putative mobile forms of restriction-modification systems and related rearrangements. Nucleic Acids Res. 38(7):2428–2443.
Galperin MY. 2005. A census of membrane-bound and intracellular signal transduction proteins in bacteria: bacterial IQ, extroverts and introverts. BMC Microbiol. 5(1):35.
Gao R, Stock AM. 2009. Biological insights from structures of two-component proteins. Annu Rev Microbiol. 63(1):133–154.
Gertz EM, et al. 2006. Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. BMC Biol. 4(1):41.
Grebe TW, Stock JB. 1999. The histidine protein kinase superfamily. Adv Microb Physiol. 41:139–227.
Hoch JA, Silhavy TJ. 1995. Two-component signal transduction. Washington: ASM Press.
Iyer LM, Abhiman S, Aravind L. 2008. MutL homologs in restriction-modification systems and the origin of eukaryotic MORC ATPases. Biol Direct. 3(1):8.

Iyer LM, Anantharaman V, Wolf MY, Aravind L. 2008. Comparative genomics of transcription factors and chromatin proteins in parasitic protists and other eukaryotes. Int J Parasitol. 38(1):1–31.
Jiang P, Pioszak A, Atkinson MR, Peliska JA, Ninfa AJ. 2003. New insights into the mechanism of the kinase and phosphatase activities of *Escherichia coli* NRII (NtrB) and their regulation by the PII protein. In: Inouye M, Dutta R, editors. Histidine kinases in signal transduction. San Diego: Elsevier. p. 143–164.
Kallberg M, Margaryan G, Wang S, Ma J, Xu J. 2014. RaptorX server: a resource for template-based protein structure modeling. Methods Mol Biol. 1137:17–27.
Karniol B, Vierstra RD. 2004. The HWE histidine kinases, a new family of bacterial two-component sensor kinases with potentially diverse roles in environmental signaling. J Bacteriol. 186(2):445–453.
Kleihues L, Lenz O, Bernhard M, Buhrke T, Friedrich B. 2000. The H2 sensor of *Ralstonia eutropha* is a member of the subclass of regulatory [NiFe] hydrogenases. J Bacteriol. 182(10):2716–2724.
Kobayashi I. 2001. Behavior of restriction-modification systems as selfish mobile elements and their impact on genome evolution. Nucleic Acids Res. 29(18):3742–3756.
Koch A, et al. 2017. MORC proteins: novel players in plant and animal health. Front Plant Sci. 8:1720.
Koretke KK, Lupas AN, Warren PV, Rosenberg M, Brown JR. 2000. Evolution of two-component signal transduction. Mol Biol Evol. 17(12):1956–1970.
Lefort V, Longueville JE, Gascuel O. 2017. SMS: smart model selection in PhyML. Mol Biol Evol. 34(9):2422–2424.
Lenz O, Friedrich B. 1998. A novel multicomponent regulatory system mediates H2 sensing in *Alcaligenes eutrophus*. Proc Natl Acad Sci U S A. 95(21):12474–12479.
Li DQ, Nair SS, Kumar R. 2013. The MORC family: new epigenetic regulators of transcription and DNA damage response. Epigenetics 8(7):685–693.
Li L, Stoeckert CJ Jr, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13(9):2178–2189.
Li S, Du J, et al. 2013. Functional and structural characterization of DNMT2 from *Spodoptera frugiperda*. J Mol Cell Biol. 5(1):64–66.
Li S, et al. 2016. Mouse MORC3 is a GHKL ATPase that localizes to H3K4me3 marked chromatin. Proc Natl Acad Sci U S A. 113(35):E5108–E5116.
Lodish H, et al. 2008. Molecular cell biology. New York: Macmillan.
Manohar M, et al. 2017. Plant and human MORC proteins have DNA-modifying activities similar to type II topoisomerases, but require one or more additional factors for full activity. Mol Plant Microbe Interact. 30(2):87–100.
Marinus MG, Casadesus J. 2009. Roles of DNA adenine methylation in host-pathogen interactions: mismatch repair, transcriptional regulation, and more. FEMS Microbiol Rev. 33(3):488–503.
Mascher T. 2006. Intramembrane-sensing histidine kinases: a new family of cell envelope stress sensors in Firmicutes bacteria. FEMS Microbiol Lett. 264(2):133–144.
Mascher T, Helmann JD, Unden G. 2006. Stimulus perception in bacterial signal-transducing histidine kinases. Microbiol Mol Biol Rev. 70(4):910–938.
Mascher T, Margulis NG, Wang T, Ye RW, Helmann JD. 2003. Cell wall stress responses in *Bacillus subtilis*: the regulatory network of the bacitracin stimulon. Mol Microbiol. 50(5):1591–1604.
Moissiard G, et al. 2012. MORC family ATPases required for heterochromatin condensation and gene silencing. Science 336(6087):1448–1451.
Mruk I, Kobayashi I. 2014. To be or not to be: regulation of restriction-modification systems and other toxin-antitoxin systems. Nucleic Acids Res. 42(1):70–86.
Ninfa AJ, Magasanik B. 1986. Covalent modification of the glnG product, NRI, by the glnL product, NRII, regulates the transcription of the glnALG operon in *Escherichia coli*. Proc Natl Acad Sci U S A. 83(16):5909–5913.
Nixon BT, Ronson CW, Ausubel FM. 1986. Two-component regulatory systems responsive to environmental stimuli share strongly conserved domains with the nitrogen assimilation regulatory genes ntrB and ntrC. Proc Natl Acad Sci U S A. 83(20):7850–7854.
Pei J, Kim B-H, Grishin NV. 2008. PROMALS3D: a tool for multiple protein sequence and structure alignments. Nucleic Acids Res. 36(7):2295–2300.
Potter SC, et al. 2018. HMMER web server: 2018 update. Nucleic Acids Res. 46(W1):W200–W204.
Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 – approximately maximum-likelihood trees for large alignments. PLoS One 5(3):e9490.
Qi M, Sun F-J, Caetano-Anollés G, Zhao Y. 2010. Comparative genomic and phylogenetic analyses reveal the evolution of the core two-component signal transduction systems in Enterobacteria. J Mol Evol. 70(2):167–180.
Revilla-Guarinos A, et al. 2013. Characterization of a regulatory network of peptide antibiotic detoxification modules in *Lactobacillus casei* BL23. Appl Environ Microbiol. 79(10):3160–3170.
Roberts RJ, Vincze T, Posfai J, Macelis D. 2015. REBASE – a database for DNA restriction and modification: enzymes, genes and genomes. Nucleic Acids Res. 43(D1):D298–D299.
Robinson VL, Buckler DR, Stock AM. 2000. A tale of two components: a novel kinase and a regulatory switch. Nat Struct Biol. 7(8):626–633.
Romling U, Gomelsky M, Galperin MY. 2005. C-di-GMP: the dawning of a novel bacterial signalling system. Mol Microbiol. 57(3):629–639.
Sayers EW, et al. 2009. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 37(Database):D5–D15.
Shannon P, et al. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13(11):2498–2504.
Sievers F, Higgins DG. 2018. Clustal Omega for making accurate alignments of many protein sequences. Protein Sci. 27(1):135–145.
Skerker JM, Prasol MS, Perchuk BS, Biondi EG, Laub MT. 2005. Two-component signal transduction pathways regulating growth and cell cycle progression in a bacterium: a system-level analysis. PLoS Biol. 3(10):e334.
Stock AM, et al. 1993. Structure of the magnesium-bound form of CheY and mechanism of phosphoryl transfer in bacterial chemotaxis. Biochemistry 32(49):13375–13380.
Stock AM, Mottonen JM, Stock JB, Schutt CE. 1989. Three-dimensional structure of CheY, the response regulator of bacterial chemotaxis. Nature 337(6209):745–749.
Stock AM, Robinson VL, Goudreau PN. 2000. Two-component signal transduction. Annu Rev Biochem. 69(1):183–215.
Szurmant H, White RA, Hoch JA. 2007. Sensor complexes regulating two-component signal transduction. Curr Opin Struct Biol. 17(6):706–715.
Tanaka T, et al. 1998. NMR structure of the histidine kinase domain of the *E. coli* osmosensor EnvZ. Nature 396(6706):88–92.
The UniProt Consortium. 2017. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 45(D1):D158–D169.
Tomomori C, et al. 1999. Solution structure of the homodimeric core domain of *Escherichia coli* histidine kinase EnvZ. Nat Struct Biol. 6(8):729–734.
Tsirigos KD, et al. 2015. The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. Nucleic Acids Res. 43(W1):W401–W407.
Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ. 2009. Jalview Version 2 – a multiple sequence alignment editor and analysis workbench. Bioinformatics 25(9):1189–1191.

Watson ML, et al. 1998. Identification of morc (microrchidia), a mutation that results in arrest of spermatogenesis at an early meiotic stage in the mouse. Proc Natl Acad Sci U S A. 95(24):14361–14366.
Wattam AR, et al. 2014. PATRIC, the bacterial bioinformatics database and analysis resource. Nucleic Acids Res. 42(D1):D581–D591.
Wecke T, Veith B, Ehrenreich A, Mascher T. 2006. Cell envelope stress response in *Bacillus licheniformis*: integrating comparative genomics, transcriptional profiling, and regulon mining to decipher a complex regulatory network. J Bacteriol. 188(21):7500–7511.
West AH, Stock AM. 2001. Histidine kinases and response regulator proteins in two-component signaling systems. Trends Biochem Sci. 26(6):369–376.
Wuichet K, Cantwell BJ, Zhulin IB. 2010. Evolution and phyletic distribution of two-component signal transduction systems. Curr Opin Microbiol. 13(2):219–225.

Associate editor: Nancy Moran