Supporting Information

Ortega et al. 10.1073/pnas.1708842114

SI Methods

Data Sources and Bioinformatics Software. Sequences of chemotaxis proteins and associated information were obtained from the MiST 2.2 database (35). Multiple sequence alignments were built using the L-INS-I algorithm from the MAFFT v4.182 package (57). A phylogenetic tree for the phylogenetic profile was generated using RAxML 8.1.21 (58) and the cladogram of chemoreceptor consensus sequences of the kinase-interacting subdomain was produced using the UPGMA (unweighted pair group method with arithmetic mean) algorithm in PAUP* v4.0 (59), with the following settings: bootstrap nreps = 500 search = UPGMA and criterion set to distance. Both were visualized using Figtree v1.4.3. RPS BLAST was used as part of the pipeline to build the reference tree and BLASTp used to build the phylogenetic profiles were from the BLAST command-line applications v2.2.29+. Information content logos were built using Weblogo 3.0 (60). A phylogenetic profile was made using custom scripts with NetworkX v1.9.1 and Numpy v1.10.4 modules for Python v2.7.12 and figures were produced using d3.js. Chemoreceptors were assigned to heptad classes using previously described hidden Markov models (39) and HMMER v2.0 (61). Kinases and adaptors were assigned to chemotaxis classes using previously described hidden Markov models (8) and HMMER v3.0 (62). The reference tree in Fig. S1 was built using the alignment of 30 universal gene markers, as previously described (63). Protein sequences of each marker were aligned using L-INS-I from MAFFT and then concatenated. The final alignment was used to infer a phylogeny using 1,000 rapid bootstrap analysis and search for best-scoring maximum-likelihood tree (option –f a) from RAxML under the PROTCATILG substitution model.

Genome Context and Gene-Expression Analyses. Potential gene fusion events and gene neighborhoods of chemotaxis genes were visualized and analyzed using the MiST database (35). Expression data for chemotaxis genes was obtained from the ADAGE-based integrated network, which was built for a compendium of *Pseudomonas aeruginosa* gene-expression data that included 950 Affymetrix GeneChip arrays from 109 experiments (38). Genes were considered to be coexpressed if they were found as high-weight (outside 2 SDs) in the same network node.

Pipeline to Construct COGs. Initial COGs were built as previously described (37) using 10E-150, 10E-140, and 10E-130 e-value cutoffs for chemoreceptors, histidine kinases, and adaptors, respectively. In addition, only hits with more than 95% query coverage were considered for chemoreceptors and adaptors and more than 25% for histidine kinases. In the case of adaptors, only the part of the sequence matching the CheW domain, as defined by the hidden Markov model PF01584 in the Pfam database (64), was considered. COGs were merged into a subfamily based on: (*i*) similarity between COGs, (*ii*) gene neighborhood, and (*iii*) domain architecture. Similarity between COGs for each of the three protein families considered—chemoreceptors, adaptors, and histidine kinases—were calculated using a distance matrix composed by the average BLAST e-value between all members of one COG against another to all pairs of COGs. A cladogram was generated based on this matrix using the neighbor software from the PHYLIP package (65) with the UPGMA algorithm and randomized input order of entries. During analysis, the phylogenetic profile columns were organized based in COG similarity and similar COGs with complementary profiles, with similar gene neighborhood and similar domain architecture, were merged into a single group. The final grouping and domain architecture information for each protein family is available in Dataset S2. Each sequence header follows the following pattern: Xx.yyy.NNN-locus-accession_number—chemoreceptor_heptad_classification [A, B]. Xx represents the first two letters of the genus, yyy represents the first three letters of the species, and NNN represents the identification number of the genome in the MiST database. The A and B are related to the original grouping of the sequence and current group of the sequence.

Similarity of Kinase-Interacting Subdomain Between COGs of Chemoreceptors. To build the similarity cladogram in Fig. 2, we aligned the full sequences of chemoreceptors for each COG independently, using L-INS-I, and trimmed the sequences to include only the kinase-interacting region, as previously described (40). A custom script was used to calculate the consensus sequence. The consensus sequences were aligned with the sequence of *Escherichia coli* Tsr and corresponding residues were mapped onto the 3D crystal structure of the Tsr trimer of dimers (PDB ID code 1QU7). The RASA in Fig. 2 was calculated using the sasa function in the VMD 1.9.3 (66).

Identification of Conserved Methylation Sites. Consensus methylation site sequence, [ASTG]-[ASTG]-x(2)-[EQ]-[EQ]-x(2)-[ASTG]-[ASTG], that was previously derived from various chemoreceptor classes (39) was used to screen multiple sequence alignments of chemoreceptor COGs to identify putative methylation sites.

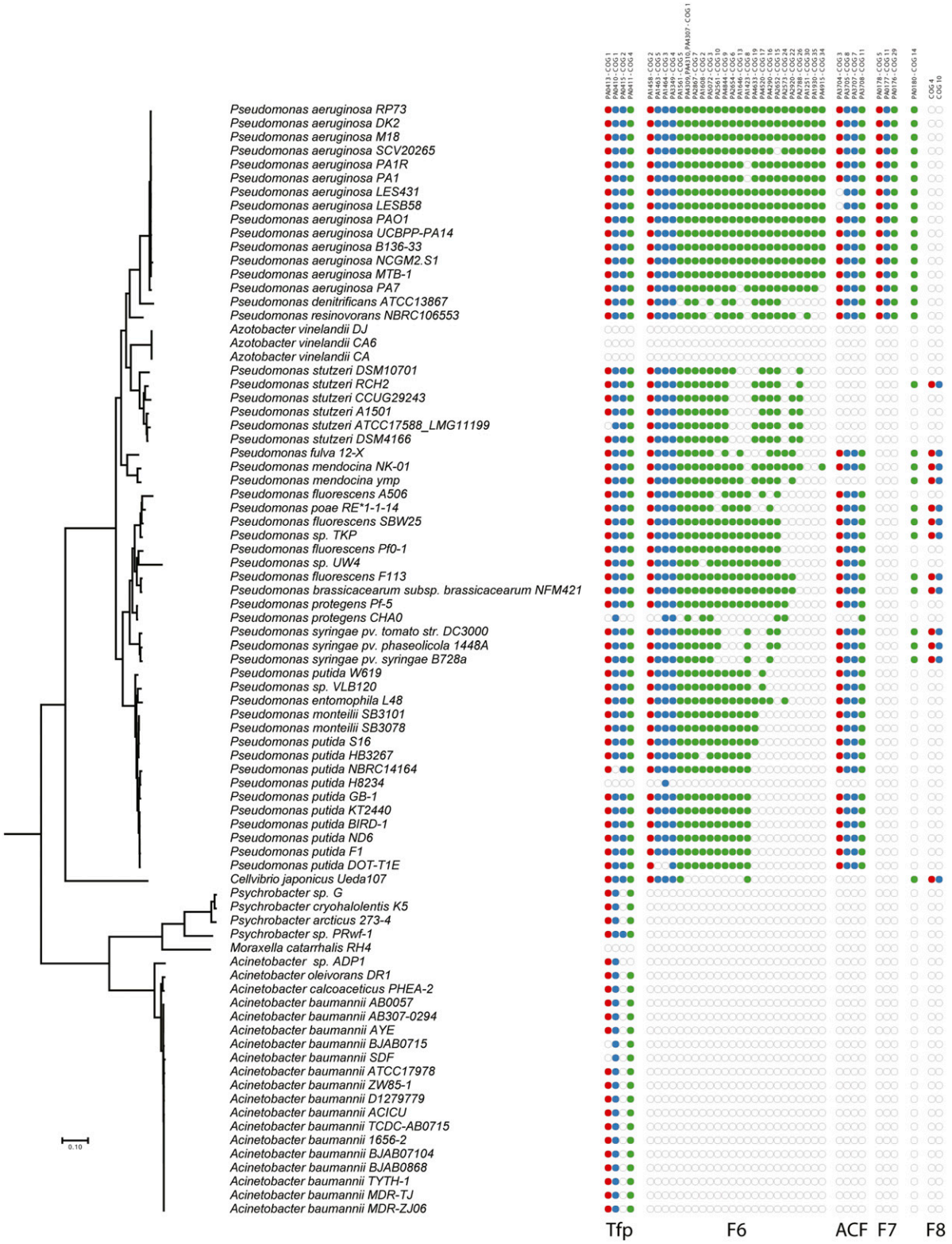

Fig. S1. A phylogenetic profile of chemotaxis proteins in *Pseudomonadales*. A ribosomal protein tree (*Left*) shows relationships between species. The panel of circles represent the presence of genes from clusters of orthologous groups of kinases (red), adaptors (blue), and chemoreceptors (green).

Fig. S2. Coexpression patterns of chemotaxis pathways and chemoreceptors in *P. aeruginosa* PAO1. Underlying data extracted from Tan et al. (38) mSystems 1: e00025-15. Color code: highly up-regulated genes are shown in green and highly down-regulated genes are in purple.

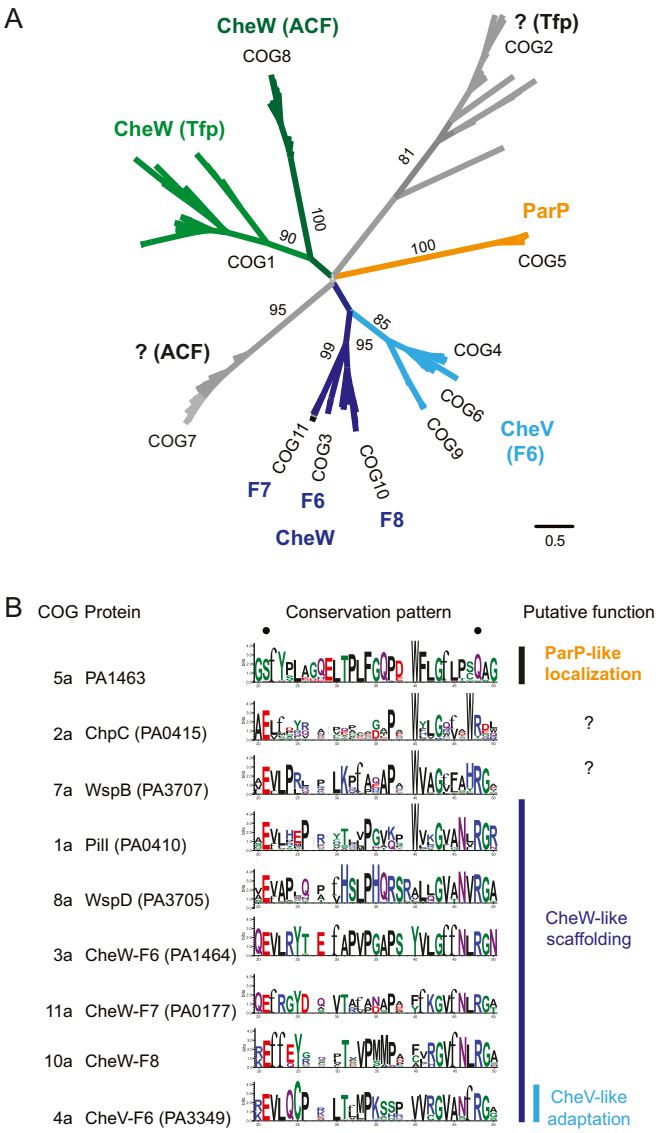

Fig. S3. Relationships between adaptor proteins from *Pseudomonas* chemosensory pathways. (*A*) A maximum-likelihood phylogenetic tree based on adaptor COGs (Dataset S2). (*B*) Conservation of the E38-R62 stabilizing salt bridge in adaptor proteins. Positions corresponding to E38 and R62 in the *E. coli* CheW protein are marked with black dots.

Table S1. Chemosensory apparatus of *P. aeruginosa* PAO1

Gene cluster	Pathway*	Chemotaxis class†	Locus tag	Gene name	Chemotaxis definition	Chemoreceptor heptad class‡
Cluster I	Che I	F6	PA1456	*cheY*	CheY	
			PA1457	*cheZ*	CheZ	
			PA1458	*cheA*	CheA	
			PA1459	*cheB1*	CheB	
			PA1460	*motA*	MotA	
			PA1461	*motB*	MotB	
			PA1462		ParA-like	
			PA1463		CheW-like	
			PA1464	*cheW*	CheW	
Cluster V	Che I	F6	PA3348	*cheR1*	CheR	
			PA3349		CheV	
Cluster II	Che II	F7	PA0173	*cheB2*	CheB	
			PA0174	*cheD*	CheD	
			PA0175	*cheR2*	CheR	
			PA0176	*mcpB, aer2, tlpG*	MCP	36H
			PA0177		CheW	
			PA0178		CheA	
			PA0179		CheY	
			PA0180	*cttP, mcpA*	MCP	Uncategorized
Cluster III	Wsp	ACF	PA3702	*wspR*	Response regulator	
			PA3703	*wspF, cheB3*	CheB	
			PA3704	*wspE*	CheA	
			PA3705	*wspD*	CheW	
			PA3706	*wspC*	CheR	
			PA3707	*wspB*	CheW	
			PA3708	*wspA*	MCP	40H
Cluster IV	Chp	TFP	PA0408	*pilG*	Response regulator	
			PA0409	*pilH*	Response regulator	
			PA0410	*pilI*	CheW	
			PA0411	*pilJ*	MCP	40H
			PA0412	*pilK*	CheR	
			PA0413	*chpA*	CheA	
			PA0414	*chpB*	CheB	
			PA0415	*chpC*	CheW	
Orphans			PA1251		MCP	40H
			PA1423	*bdlA*	MCP	40H
			PA1561	*aer*	MCP	40H
			PA1608		MCP	24H
			PA1646		MCP	40H
			PA1930	*mcpS*	MCP	24H
			PA2561	*ctpH*	MCP	40H
			PA2573		MCP	40H
			PA2652		MCP	40H
			PA2654	*tlpQ*	MCP	40H
			PA2788		MCP	40H
			PA2867		MCP	40H
			PA2920		MCP	40H
			PA4290		MCP	Uncategorized
			PA4307	*pctC*	MCP	40H
			PA4309	*pctA*	MCP	40H
			PA4310	*pctB*	MCP	40H
			PA4520		MCP	40H
			PA4633		MCP	40H
			PA4844	*ctpL*	MCP	40H
			PA4915		MCP	40H
			PA5072	*mcpK*	MCP	40H

*Pathway definitions according to Hickman et al. (6) and Whitchurch et al. (19).
†Chemotaxis class definition according to Wuichet and Zhulin (8), where a letter F followed by the class number (1–17) denotes pathways controlling flagella-mediated motility, TFP is type IV pili-mediated motility, and ACF is alternative cellular functions, such as gene expression.
‡Chemoreceptor heptad class definition according to Alexander and Zhulin (39), where the number of helical heptad in signaling domain defines the class.

Table S2. Correlation between chemotaxis systems and chemoreceptor heptad classes in genomes with a single chemotaxis pathway

	Chemotaxis systems (no. of genomes)											
Heptad class (no. of genomes)	F1 (167)	F2 (7)	F3 (20)	F4 (2)	F5 (32)	F6 (25)	F7 (96)	F8 (2)	F14 (8)	F15 (2)	ACF (12)	Tfp (30)
24H (74)	21		8		6	18	21					
28H (95)	1		94									
36H (1,061)	3		1				1,037					20
38H (247)	0				247							
40H (661)	11		36	20	4	529	2		3	16	8	32
44H (1,097)	1,089	1						7				
48H (72)		72										
Unclassified (1,086)	668	13	100	2	72	48	128	4	38	4	3	6

Data from the MiST database (35).

Other Supporting Information Files

Dataset S1 (PDF)
Dataset S2 (PDF)
Dataset S3 (PDF)
Dataset S4 (PDF)