**UNIVERSIDAD POLITÉCNICA DE MADRID**

Escuela Técnica Superior de Ingeniería Agronómica, Alimentaria y de Biosistemas

Centro de Investigación en Biotecnología y Genómica de Plantas




# Ecology and Evolution of Chemoreceptors: a global genomic and metagenomic survey

## DOCTORAL THESIS

Submitted for the degree of Doctor by:

**Claudia Sanchis López**

BSc. in Biotechnology

MSc. in Computational Biology

Madrid, 2024





UNIVERSIDAD POLITÉCNICA DE MADRID
Escuela Técnica Superior de Ingeniería Agronómica, Alimentaria y de Biosistemas
Centro de Investigación en Biotecnología y Genómica de Plantas

**Doctoral Degree in Biotechnology and Genetic Resources of Plants and Associated Microorganisms**

# Ecology and Evolution of Chemoreceptors: a global genomic and metagenomic survey

## DOCTORAL THESIS

Submitted for the degree of Doctor by:

**Claudia Sanchis López**
BSc. in Biotechnology
MSc. in Computational Biology

Under the supervision of:
Dra. Emilia López Solanilla
Dr. Jaime Huerta Cepas

Madrid, 2024

Title: Ecology and Evolution of Chemoreceptors: a global genomic and metagenomic survey

Author: Claudia Sanchis López

Doctoral Programme: Doctoral Degree in Biotechnology and Genetic Resources of Plants and Associated Microorganisms

Thesis Supervision:

Dra. Emilia López Solanilla, Catedrática de Universidad, Escuela Técnica Superior de Ingeniería Agronómica, Alimentaria y de Biosistemas, Universidad Politécnica de Madrid (UPM). (Supervisor)

Dr. Jaime Huerta Cepas, Investigador Científico, Centro de Biotecnología y Genómica de Plantas, Universidad Politécnica de Madrid (UPM) – Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA-CSIC)(Supervisor)

External Reviewers:

Thesis Defense Committee:

Thesis Defense Date:

This thesis has been partially supported by research grant FPU19/06635 from the Ministerio de Universidades, and research grants PID2021-125673OB-I00 and PID2021-127210NB-I00 from the Ministerio de Ciencia e Innovación.


Nutriagobies tía

# Abstract

Chemoreception, the ability of microorganisms to detect and respond to environmental signals, is essential for their adaptability, survival, and ecological interactions. Chemosensory pathways enable microbes to perceive environmental changes and regulate behaviors such as nutrient acquisition, predator evasion, and host interactions. Central to these processes are chemoreceptor proteins (CRs), which initiate signal transduction by recognizing stimuli through their ligand-binding domains (LBDs), regulating cellular responses. Despite their importance, much of the diversity and ecological roles of CRs remain uncharacterized, particularly in uncultured and low-abundance microorganisms. This thesis aims to address these gaps by studying CRs at genomic and metagenomic scales, exploring their biodiversity, ecological relevance, and evolutionary significance.

The main objectives of this thesis are: (1) to develop a methodology for automated CR characterization and functional classification, (2) to analyze CR prevalence and specificity across prokaryotic phylogeny, including uncultured microorganisms, (3) to investigate the evolutionary history of CRs on a global scale, and (4) to create a novel method to study CR diversity in low-abundance taxa from environmental samples.

In Chapter 1, we analyzed over 82,000 CR sequences from 11,806 representative microbial genomes. A novel homology-based clustering approach was developed to classify CRs by their LBDs, offering a more exhaustive functional classification system than traditional methods. This revealed thousands of novel LBD types, indicating that CR diversity is far from fully explored. Furthermore, we found that LBD specificity for plant-associated lifestyles is more influenced by ecological factors than phylogenetic signals, emphasizing environmental adaptation in shaping CR profiles.

In Chapter 2, we extended the analysis to environmental data, compiling what we believe to be the largest and most comprehensive CR catalog to date, GlobalCR. This catalog includes >1.4 million CR sequences from 800,000 metagenome-assembled genomes (MAGs) and 229,000 CRs from 41,000 reference genomes. Extracellular sensing domains exhibited immense diversity and strong habitat-specific associations, suggesting significant adaptive value. Taxonomic analysis revealed that many CRs belong to previously undescribed lineages with high CR content, indicating novel adaptations. A phylogeny of 400,000 unique CR sequences, from both cultured and uncultured organisms, provided insights into the evolutionary history of CRs, horizontal gene transfer events, and functional diversification. Notably, evidence of co-evolution between heptad classes and CheA histidine kinases, key components of chemosensory pathways, was identified.

In Chapter 3, we developed a novel capture-based targeted sequencing method to investigate low-abundance chemosensory microorganisms in environmental samples. This approach identified up to two orders of magnitude more CRs per sample compared to standard metagenomic sequencing. Applied to plant rhizosphere samples, it uncovered a hidden biodiversity of chemosensitive bacteria, including latent pathogens and symbionts, and numerous previously unknown microorganisms representing new bacterial lineages.

Overall, this thesis advances the ecological, functional, and evolutionary understanding of microbial chemosensory systems and provides a valuable resource of data and methods to support future exploration and analysis in this field.

# Resumen

La quimiorrecepción, la capacidad de los microorganismos para detectar y responder a señales ambientales, es fundamental para su adaptabilidad, supervivencia e interacciones ecológicas. Las vías quimiosensoriales permiten a los microbios percibir cambios en su entorno y regular comportamientos como la adquisición de nutrientes, la evasión de depredadores y las interacciones con hospedadores. En el centro de estos procesos están los quimiorreceptores (CRs), que inician la transducción de señales a través de sus dominios de unión a ligandos (LBDs). A pesar de su importancia, gran parte de la diversidad y los roles ecológicos de los CRs siguen sin caracterizarse, especialmente en microorganismos no cultivados o de baja abundancia. Esta tesis aborda estas lagunas estudiando los CRs a escalas genómicas y metagenómicas, explorando su biodiversidad, relevancia ecológica y significancia evolutiva.

Los objetivos son: (1) desarrollar una metodología para la caracterización automatizada de los CRs, (2) examinar la prevalencia y distribución de CRs en especies procariotas conocidas y desconocidas, (3) investigar patrones ecológicos y evolutivos de los CRs utilizando datos metagenómicos y (4) diseñar un método innovador para explorar la diversidad de CRs en taxones raros.

En el Capítulo 1, se realizó un análisis filogenómico de 82,000 secuencias de CR extraídas de 11,806 genomas. Se desarrolló un nuevo enfoque de agrupamiento basado en homología de novo para clasificar los CRs según sus LBDs, lo que reveló miles de tipos de LBDs novedosos y destacó una diversidad no saturada de CRs. El análisis mostró que la especificidad de los LBDs para estilos de vida asociados a plantas está más influida por factores ecológicos que filogenéticos, subrayando la adaptación ambiental en los perfiles de CR.

En el Capítulo 2, el análisis se amplió a datos ambientales, generando el mayor catálogo de CRs hasta la fecha (GlobalCR). Este incluye >1.4 millones de CRs de 800,000 genomas ensamblados a partir de metagenomas (MAGs) y 229,000 CRs de 41,000 genomas de referencia. La diversidad de dominios sensoriales extracelulares resultó ser inmensa y fuertemente vinculada a hábitats naturales, sugiriendo un valor adaptativo significativo. El análisis taxonómico reveló muchos CRs en linajes no descritos con alto contenido de CR, indicando adaptaciones novedosas. Una filogenia completa de 400,000 CRs únicos de organismos cultivados y no cultivados permitió investigar la historia evolutiva de los CRs, eventos potenciales de transferencia horizontal de genes y la diversificación funcional. En particular, se identificó la coevolución de las clases de heptadas con las quinasas de histidina CheA, reflejando adaptaciones en las vías quimiosensoriales.

En el Capítulo 3, se desarrolló un nuevo método de secuenciación dirigida basado en captura para investigar microorganismos quimiosensibles en taxones de baja abundancia. Este enfoque identificó hasta dos órdenes de magnitud más CRs por muestra que la secuenciación metagenómica estándar. Aplicado a muestras de la rizosfera de plantas, descubrió una biodiversidad oculta de bacterias quimiosensibles, incluidos patógenos latentes y simbiontes, así como numerosos microorganismos desconocidos con CRs en linajes bacterianos nuevos.

En resumen, esta tesis avanza en la comprensión ecológica, funcional y evolutiva de los sistemas quimiosensoriales microbianos, ofreciendo un recurso público de datos y métodos para impulsar futuras investigaciones en este campo.

# Table of Contents


Abstract . . . . . v
Resumen . . . . . vii
List of Figures . . . . . xii
List of Tables . . . . . xviii
Abbreviations and acronyms . . . . . xxi

**Introduction** **1**

1 Microbial chemoperception: sensing and responding to the environment . . . . . 3
1.1 Microbial communities are everywhere . . . . . 3
1.2 Mechanisms of enviromental sensing: signal transduction systems . . . . . 3
1.3 A sophisticated solution: Chemoreceptor-based signaling pathways . . . . . 5
1.4 Chemoreceptors, first line of the system . . . . . 6
Cytoplasmic signaling domain: MCPsignal . . . . . 8
Sensory domain: The Ligand Binding Domain (LBD) . . . . . 9
LBD classification . . . . . 10
1.5 Characterization of LBDs: a major challenge in CR research . . . . . 14
2 Ecological role of chemoreceptors . . . . . 16
2.1 Chemoperception in human-associated communities . . . . . 17
2.2 Chemoperception in aquatic environments . . . . . 18
2.3 Chemoperception in plant-associated environments: phytopathogens and symbionts . . . . . 19
Chemosensitive microorganisms in the Rhizosphere . . . . . 19
Chemosensitive microorganisms in the Phyllosphere . . . . . 20
2.4 Challenges in chemoreceptors research across ecological niches . . . . . 21
3 Chemoperception from a genomic and metagenomic perspective . . . . . 22
3.1 First steps in understanding the microbial chemoperception: from phenotype to genotype . . . . . 22

3.2 DNA sequencing becomes the new microscope: Genomics provides a comprehensive view of chemosensitive systems . . . 23
3.3 The emerging role of metagenomics in chemoperception research: challenges and opportunities . . . 24
Environmental DNA: the uncultivated majority . . . 25
Major advances and data integration in metagenomics . . . 27
3.4 Low abundance microorganisms: beyond standard metagenomics . . . 30

**Objectives** **31**

**Materials and methods** **35**

1 Prevalence and Specificity of Chemoreceptors Profiles in Plant-Associated Bacteria 37
1.1 Chemoreceptors (CRs) sequence retrieval . . . 37
1.2 Ligand binding domain (LBD) extraction . . . 37
1.3 Clustering of LBD sequences . . . 38
1.4 Construction of the plant-associated bacteria (PAB) databases . . . 38
1.5 Degree of Plant Specificity (DPS) . . . 38
1.6 Phylogenetic tree reconstruction and visualization . . . 38
1.7 Phylogenetic signal tests . . . 39
2 Adaptative value and evolutionary history of bacterial chemoreceptors revealed by global metagenomic survey . . . 40
2.1 GlobalCR sequence retrieval . . . 40
2.2 Taxonomy annotation and taxonomy distribution of CR genes . . . 40
2.3 Heptad class and CheA genes annotation . . . 41
2.4 MCPsignal tree . . . 41
2.5 Putative Horizontal Gene Transfer detection . . . 41
2.6 Habitat distribution of CRs genes . . . 42
2.7 Datasets of LBDs and MCPsignal subfamilies . . . 42
2.8 Cumulative curves of domain subfamilies . . . 43
2.9 Relative abundance matrix of subfamilies . . . 43
2.10 Species relative abundance matrix . . . 44
2.11 Random Forest Classification of Biome Prediction . . . 44
2.12 UMAP and HDBSCAN . . . 44
2.13 Hierarchical clustering of HDBSCAN clusters . . . 45
2.14 Prevalence and specificity of subfamilies per biome . . . 46
2.15 Subfamilies trees . . . 46
3 Unveiling the Repertoire of Microbial Chemoreceptor Genes in the Rare Biosphere . 47

3.1 MetaCR sequence retrieving . . . 47
3.2 metaCR ecological annotation . . . 47
3.3 Probe set design . . . 47
3.4 Sample collection and DNA extraction . . . 48
3.5 DNA Fragmentation . . . 49
3.6 Target capture experiments . . . 49
3.7 Bulk shotgun metagenomics sequencing . . . 51
3.8 Read processing: trimming, demultiplexing, assembly, and deduplication . . . 51
3.9 Detection of on-target reads . . . 52
3.10 Sequencing methods comparison (Nanopore *vs.* Illumina) . . . 52
3.11 Mapping to GTDB r207 . . . 52
3.12 Phylogenetic distribution of the known CRs detected . . . 53
3.13 Taxonomic profiling of metagenomic samples . . . 53
3.14 Phylogenetic delineation of novel CRs . . . 53

**Results** **55**

**Chapter 1: Prevalence and Specificity of Chemoreceptors Profiles in Plant-Associated Bacteria** **57**

1.1 Towards a global catalog of chemoreceptors in plant-associated bacteria . . . 59
1.2 Classifying chemoreceptors according to their Ligand Binding Domains . . . 60
1.3 Identifying PAB-specific Ligand Binding Domains . . . 63
1.4 Phylogenetic vs. Ecological signal in PAB specific Ligand Binding Domains . . . 68
1.5 Ligand Binding Domains profiles per genome . . . 71
1.6 Discussion . . . 71

**Chapter 2: Adaptive value and evolutionary history of bacterial chemoreceptors revealed by global metagenomic survey** **75**

2.1 A global catalog of chemoreceptors from cultured and uncultured microorganism . . . 77
2.2 Novel lineages of chemosensing microorganisms . . . 79
2.3 Origin and evolution of CRs . . . 84
2.4 Detecting HGT in CRs across metagenomic data . . . 86
2.5 An open space of bacterial chemosensors . . . 98
2.6 Free living microbial communities are enriched chemosensitive species . . . 98
2.7 Extracellular chemosensing domains are an adaptative value in bacterial evolution . . . 100

2.8 Biome specific sensing capabilities . . . 102
2.9 Discussion . . . 106
Evolutionary history of CRs revealed by the MCPsignal domain . . . 107
Ecological signal revealed by Ligand Bindings Domains (LBD) . . . 108

**Chapter 3: Unveiling the Repertoire of Microbial Chemoreceptor Genes in the Rare Biosphere** **113**

3.1 A comprehensive catalog of CRs in the global microbiome . . . 115
3.2 Target capture sequencing boosts CR gene detection across habitats . . . 116
3.3 Identification of CR sensing domains from rare taxa . . . 118
3.4 Unveiling the hidden biodiversity of low-abundance chemosensitive organisms 121
3.5 A vast space of novel chemosensitive species . . . 124
3.6 Discussion . . . 127

**General Discussion** **129**

**References** **139**

**Annexe I: Supplementary Material** **169**

1 Prevalence and Specificity of Chemoreceptors Profiles in Plant-Associated Bacteria 171
2 Adaptive value and evolutionary history of bacterial chemoreceptors revealed by global metagenomic survey . . . 174
3 Unveiling the repertoire of microbial chemoreceptor genes in the rare biosphere . . . 182

**Annexe II** **189**

# List of Figures

1 **Signal transduction systems: One component system (OS), Two components systems (TCS) and Chemoreceptor-based signaling pathway.** . . . . . 4

2 **Classification of chemoreceptors according to membrane topology and sensing domain position.** Adapted from Lacal *et al.* (2010)[42], Ud-dDin and Roujeinikova (2017)[39] and Cha *et al.* (2022)[43]. . . . . 7

3 **Representation of the canonical chemoreceptor organization.** The extracytosolic sensing domain (LBD) is flanked by two transmembrane regions (TM). The signaling domain (MCPsignal) is in the cytoplasm, with the linker HAMP domain. The MCPsignal can be subdivided in the Methylation Helix (MH), connected by the Flexible Bundle (FB) to the Signal Subdomain (SSD). The major heptad classes of the signaling domain are represented. . . . . 9

4 **Sensory repertoire distribution.** The distribution of each family of LBDs is shown, while some of the structural models within each LBD family are represented. Adapted from Ortega *et al.* (2017). . . . . 10

5 **Evidence of chemoperception importance across different environments.** a) Human-associated habitats, such as the gastrointestinal tract and the lungs. b) ocean environments, nutrient hotspots, phytoplankton association or symbiosis with the Hawaiian squid. c) soil and plant, with enriched areas around plant roots (rhizosphere), or on the aerial part of the plant (phyllosphere). Bacterial names enclosed in dark-lined squares indicate pathogens. . 16

6 **Prokaryotic reference genomes increase over the years.** extracted from prokaryotes.txt of GenBank downloaded in June 2024. a) Number of prokaryotes reference genomes in NCBI, currently 19,328. b) Number of genomes available in RefSeq database. . . . . 25

8 **Shotgun metagenomics pipeline.** . . . . . 28

1.1 **Schematic view of the bioinformatics pipeline used to identify CRs that are potentially relevant for plant association.** From a set of 11,806 representative prokaryotic genomes, 82,277 protein sequences were mined using HMM-based searches against the MCPsignal Pfam domain (PF00015). CR topology was analysed by predicting transmembrane regions (TMs) and Pfam domains. Based on the topological analysis, LBD regions were predicted and a set of 72,480 LBD sequences was obtained. Clustering of LBDs based on sequence homology (20% minimum sequence identity with at least 50% sequence coverage) resulted in 5,149 clusters or subfamilies of LBDs, of which 1,842 contained a single sequence. To study a possible link between the LBD profiles and plant-associated lifestyle, a manually-curated subset of 960 representative species of Plant-Associated Bacteria (PAB) was generated, including Phytopathogens (119) and Symbionts (191) subsets. The determination of the proportion of PAB LBDs present in each cluster allowed to assign the Degree of Plant Specificity (DPS) value for each LBD subfamily. Subsequent analysis of high-DPS clusters identifies LBD clusters that are potentially important for bacterial plant association. Furthermore, the validation of the high-DPS clusters as good ecological indicators was corroborated by measuring their phylogenetic signal. A detailed step-by-step description of the process can be found in the Materials and Methods section. . . . . 61

1.2 **Length distribution of the LBDs.** The analysis was conducted on 72,480 LBDs and the predominant LBD types within each of the main peaks are indicated. Only LBDs shorter than 500 amino acids are represented. . . . . 64

1.3 **Visual representation of the abundance of the LBD families.** The outer donut of the chart represents the distribution of each LBD type and its relative abundance (in percentage of sequences), and the number of clusters with at least 90% of their sequences sharing the same LBD type, as defined by the Pfam signature. The LBDs are sorted according to the number of clusters within each LBD type. The inner donut of the chart represents all the clusters included within each LBD category, indicating the number of sequences contained in each subfamily. All singletons are merged in the last section of each LBD type (e.g., LBDs classified as 'UNKNOWN' have 1,242 singletons, that is, clusters containing only one sequence). 'Mixed clusters' are those that do not reach the 90% threshold of sequences with the same Pfam model per cluster. 'Low abundance LBDs' include those LBD types that group into less than 12 'compact clusters'. . . . . 65

1.4 **Phylogenetic signal detection in plant-associated LBD clusters.** (A) Proportion of the significant phylogenetic signal among LBD clusters enriched in PAB-broad, PAB-phyto and PAB-symb species within two thresholds ($\geq$50%, $\geq$80%). The significance through a p-value test with 100 iterations, p-value $\geq$0.05 rejects the null hypothesis of a phylogenetic signal (see Materials and Methods). (B) LBD cluster #549 (in red) distributed according to the chemotactic species phylogeny (5,763 representative species). The green leaves of the tree represent the PAB species. DPS, $\delta$ and p-values for this LBD cluster are represented in the lower-left table. (C) Phylogeny representation of LBD cluster #549, containing 27 LBD sequences distributed across 4 orders (numbered 1-4 in the tree). The domain architecture prediction is shown for each of the CRs. . . . . 70

1.5 **Distribution of the high plant specific LBDs in the PAB species profiles per genome.** (A) Calculation of the proportion of high DPS-broad ($\geq$80%) LBDs in the total number of LBDs present in each species. The graph illustrates the distribution of the number of species according to the proportional ranges, plotting the species count as PAB-symb (yellow), PAB-phyto (brown) and the rest of PAB (green). (B) Absolute number of LBDs with a DPS-broad $\geq$80% in each PAB genome. The graph illustrates the distribution of the number of species as an absolute count of high DPS LBDs ranges. The species count is plotted as PAB-symb (yellow), PAB-phyto (brown) and the rest of PAB (green). . . . . 72

2.1 **Distribution of CRs across the prokaryotic tree, in reference genomes and Metagenome Assembly Genomes.** The mean number of CRs per chemosensitive species is calculated for each phyla within the Genome Taxonomy Database tree r207, and plotted in a barplot for the metagenomic (orange) and reference genomic (purple) data. Scale is showing mean CRs lines in both datasets (8.3 CRs/MAGs and 12.3 CRs/ref. species). The 28 chemosensitive uncultivated-only phyla are highlighted in blue. . . . . 81

2.2 **Phylogenetic tree reconstruction of CRs across globalCR.** The tree includes a total number of 428,179 unique MCPsignal with the heptad subtype metadata annotated, as well as the co-ocurrence pattern with the different CheA histidine kinases within the same genome. Collapsed nodes in red are duplication branches, and the last common ancestor name is shown. The background of the branches is of an heptad color only if more than 90% of the sequences are from the same heptad. . . . . 87

2.3 **Phylogenetic distribution of heptad subtypes at phylum level.** A) The stacked bars represent the phyla distribution of the total sequences in that heptad subtype. . . . . 88

2.3 **Phylogenetic distribution of heptad subtypes at phylum level.** B) GTDB tree collapsed at the phylum level, with bar plots indicating the proportion of chemosensitive genus within each phylum associated with each heptad subtype. . . . . 89

2.4 **Correlation matrix of MCPsignal heptads and CheA subtypes across species clusters in globalCR.** The heatmap displays Pearson correlation coefficients between MCPsignal heptads and CheA subtypes based on their mean abundances across species clusters. Color intensity indicates the strength and direction of correlations, with values ranging from -1 (negative correlation) to +1 (positive correlation). . . . . 90

2.5 **Cumulative curve of sensing domain homology clusters across globalCR.** The curves show the cumulative number of subfamilies or clusters for each domain as randomly shuffled sequences are added. A) Subfamilies are grouped by their dominant Pfam domain (comprising more than half of the sequences) or classified as "UNKNOWN" if no Pfam domain is detected. B) Sensing domain subfamilies are categorized as extracellular or intracellular (without predicted transmembrane domains). MCPsignal subfamilies are also included. . . 99

2.6 **Distribution of chemoperception across biomes.** A) The percentage of chemosensitive species is calculated across clusters, counting the proportion of SPIRE species clusters with at least one CRs in the MAGs out of the total number of species clusters annotated in each habitat. B) Box plots of the mean number of CRs per species cluster. Each species cluster number of CRs is the average of all the MAGs within the species cluster. The mean number is plotted in red dots for each habitat, and the global mean is plotted in the red dash line (8.8 CRs/species). . . . . 101

2.7 **Projection of the ecological sub-structure of datasets applying UMAP dimensional reduction.** UMAP projections of the relative abundance matrix for each domain subfamily, colored by ecological label. The full UMAP projection is shown on the left, while only HDBSCAN clusters are shown on the right (with noise uncolored). A) MCPsignal domains, B) Intracellular sensing domains, C) Extracellular sensing domains, D) Taxonomic structure based on mOTUs[211] relative abundance. . . . . 103

2.8 **Hierarchical clustering of samples grouped by their sensing capabilities.** Clusters identified by HDBSCAN algorithm in the UMAP visualization of samples grouped by extracellular sensing domain are shown in each row. The hierarchical clustering group them by similarity. The barplot represents the relative abundance of the ecological origin (biome) of the samples within each cluster. The heatmap panel is showing the biome specificity of the prevalent sensing subdomains subfamilies for each HDBSCAN cluster (row) subfamilies. . . . . 105

2.9 **Proposed Origins and Evolutionary Dynamics of Chemoreceptors (CRs).** Heptads not found as compact groups (harboring at least 10% of sequences of that subtype) are not included in the scheme. Detected duplication events are marked in red, while potential undetected events, possibly due to gene loss, specialization, or horizontal gene transfer (HGT), are indicated in pink. Dashed branches suggest proposed gene loss. . . . . 109

3.1 **MetaCR distribution and diversity in the sensing domain profile across habitats.** a) Total count of CR genes retrieved from each habitat; b) Total number of samples of the source catalog Global Microbial Gene Catalog v1 (GMGCv1[186]) from each habitat; c) Normalized distribution of CRs per Gpb sequenced in the samples of origin per habitat with mean values in red; outliers are not shown (>200 CRs/Gbp); d) Sensing profiles for each habitat in terms of ligand binding domains (LBDs) showing the relative abundance of the most abundant (>3%) LBDs in each habitat. Low-abundance LBDs are grouped in the "Others" category. . . . . 117

3.2 **Schematic representation of the molecular protocols implemented for long-read targeted metagenomics and shotgun metagenomics for the detection of chemoreceptors in metagenomic samples.** . . . . . 119

3.3 **Long-read targeted metagenomics sequencing approach boosts detection of CRs across diverse samples.** Total count of a) MCPsignal domains detected in Pool1; b) MCPsignal domains normalized by sequencing depth at 40Mbp (average of 10 randomized sets for Pool-1); c) MCPsignal domains detected for each sample in Pool-2; d) MCPsignal domains after normalization at 374Mbp (average of 10 randomized sets for Pool-2); and e) the 70 most frequent domain architectures detected in Pool-2. . . . . 120

3.4 **Detection of CRs with long-read targeted metagenomics across the entire microbial phylogeny.** GTDB r207 archaeal and bacterial trees concatenated and collapsed at order level. Target capture (purple) and shotgun (orange) sequencing bars represent the total count (log scale) of different CRs detected at 100% nts identity unambiguously per order. Branches with more than 30 CRs are highlighted. . . . . 123

3.5 **Revealing the repertoire of low abundance known CRs detected by long-read targeted metagenomics.** Number of different CRs detected (unique hit at 100% nts identity) by target capture (purple) or shotgun (orange) sequencing from among the total repertoire contained in GTDB r207 (gray), showing a) the top species with more CRs detected by either approach; b) plant pathogenic species only; and c) plant symbiont species only. . . . . 125

3.6 **Last common ancestor of novel CRs.** The number and the last common ancestor of the nodes with more than 100 sequences in each protein identity range are shown. The calculation of the last common ancestor was done based on the best hit taxonomic annotations of the target capture reads within the nodes in the tree. . . . . 126

S1.1 **Phylogenetic signal detection in plant-associated LBD clusters using the phylogeny of chemotactic species tree.** Proportion of significant phylogenetic signal among LBD clusters enriched in PAB-broad, PAB-phyto and PAB-symb species within two thresholds ($\geq$50%, $\geq$80%). The reference phylogeny used in this analysis excluded any species without CRs in their genome. . . . . 173

S2.1 **Detail on phylogenetic tree reconstruction of CRs across globalCR (Figure 2.2).** . . . 175

S2.1 **Detail on phylogenetic tree reconstruction of CRs across globalCR (Figure 2.2).** . . . 176
S2.2 **F1 score of Random Forest Classifier of different dataset across environments.** The average number of the F1 score and the standard error bars is calculated for each 10 iteration of the subsampling for each biome and dataset. . . . . 177
S2.3 **Enrichment of sequenced chemoreceptors across biomes.** Box plots of the ratio number of CRs sequenced per giga base pairs of each sample distributed across biomes. The mean within each habitat is plotted in red dots, and the mean of all the samples is plotted in the red dash line (6.4 CRs/Gb). . . . . 178
S2.4 **Specificity of sensing domains in HDBSCAN predicted clusters of sensing capabilities in grained ecological labels.** . . . . . 179
S3.1 **Number of CR genes assembled per Gbp in 13,174 metagenomes from GMGCv1.** The histogram represents the distribution of mean CR genes detected per Gbp across the metagenomics samples found in GMGCv1. . . . . 183
S3.2 **CR count normalization in metaCR.** a) CRs per GBp calculation after subsampling 100 times the same sequencing depth (samples until fulfilling 50 GBp) in each habitat. b) CRs per GBp after subsampling 100 times the same number of samples (20) in each habitat. Outliers (>200 CRs/GBp) are not shown. . . . . 184
S3.3 **Identity distribution of the best-hit against GTDB.** a) Pool-2 CRs are mapped using BLASTp against MCPsignal domains from Genome Taxonomy Database r207. b) Pool-2 putative LBDs are mapped using BLASTp against LBDs extracted from Genome Taxonomy Database r207. The identity distribution of the best hit is shown. . . . . 185
S3.4 **Host-plant contamination rate.** The bars show the percentage of reads aligning with the reference genomes of the host plants (tomato, *Solanum lycopersicum* L; potato, *Solanum tuberosum* L) for the samples from Pool-1. . . . . 186

# List of Tables


1 **Summary of main LBDs.** Selection of LBDs of interest for this work from Ortega *et al.* analysis[50]. The description of each LBD model included in this table has been extracted from InterPro[59]. . . . . 12

2 **GTDB release number of lineages per rank.** . . . . . 29

1.1 **List of PAB with the highest number of predicted CRs.** P, Phytopathogen. S, Plant Symbiont. 62

1.2 **CRs predicted to be involved in plant-bacteria interaction.** . . . . . 67

1.3 **Distribution of LBD types among clusters with high DPS ($\geq$50%)** . . . . . 68

2.1 **Top lineages with highest mean CR content per chemosensitive species.** The top 25 lineages for phylum and order are shown, with the average content of CRs per each species cluster containing at least one CR within each lineage. . . . . 82

2.2 **Unknown chemosensitive species clusters across globalCR at different rank levels.** The number of different species clusters without a classification with gtdb-tk using r207 released are counted by the last rank that could be classified. In the second column the number of different GTDB assignments to all the species is illustrated, together with the gtdb-tk classifications for the top 5 spire clusters with the highest CRs content. . . . . 83

2.3 **Putative Horizontal Gene Transfer Events across globalCR.** . . . . . 92

2.3 **Putative Horizontal Gene Transfer Events across globalCR.** . . . . . 93

2.3 **Putative Horizontal Gene Transfer Events across globalCR.** . . . . . 94

2.3 **Putative Horizontal Gene Transfer Events across globalCR.** . . . . . 95

2.3 **Putative Horizontal Gene Transfer Events across globalCR.** . . . . . 96

2.3 **Putative Horizontal Gene Transfer Events across globalCR.** . . . . . 97

3.1 **Taxonomy assignment of known CRs from Genome Taxonomy Database (GTDB) of sequences identified in target capture and shotgun metagenomics.** Known CRs are unambiguously assigned to species based on comparisons using different identity thresholds to the CRs contained in the GTDB r207 database. The number of sequences assigned at species rank is shown along with the number of mapped species in parenthesis. Sequences that meet the identity thresholds but do not map to a single species are indicated as not species-specific. Sequences that are not mapped are shown, along with their percentage of the total sequences. . . . . 122
3.2 **Delineation of novel CRs groups from target capture experiments.** A phylogenetic tree of MCPsignal domains was constructed using protein sequences from GTDB and captured sequences. Nodes were classified based on different protein identity ranges and minimum number of sequences per node, focusing on identifying novel or divergent CR groups derived exclusively from the capture experiments. . . . . 124
S2.1 **Sample distribution of SPIRE.** . . . . . 180
S2.2 **Distribution of heptad subtypes across globalCR.** . . . . . 180
S2.3 **Mean number and presence of CheA subtypes across globalCR.** . . . . . 181
S3.1 **Summary statistics for sequencing Pool-1 and Pool-2.** . . . . . 187
S3.2 **Number of captured MCPsignal domains clustered at different identity thresholds.** . 188

# Abbreviations and acronyms

**ACF** Alternative Cellular Function

**BLAST** Basic Local Alignment Search Tool

**bp** Base Pair

**CR** Chemoreceptor

**DNA** Deoxyribonucleic Acid

**DPS** Degree of Plant Specificity

**ETE** Environment for Tree Exploration

**Gbp** Giga Base Pair

**GMGC** Global Microbial Gene Catalog

**GTDB** Genome Taxonomy Database

**H** Heptad

**HAMP** Histidine kinase, adenylyl cyclase, methyl-accepting chemotaxis protein, phosphatase

**HDBSCAN** Hierarchical Density-Based Spatial Clustering of Applications with Noise

**HGT** Horizontal Gene Transfer

**HMM** Hidden Markov Model

**LBD** Ligand Binding Domain

**MAG** Metagenome-Assembled Genomes

**MCP** Methyl-accepting Chemotaxis Protein

**NCBI** National Center for Biotechnology Information

**NGS** Next Generation Sequencing

**nt** Nucleotide

**OCS** One Component System

**ONT** Oxford Nanopore Technology

**ORF** Open Reading Frame

**PAB** Plant Associated Bacteria

**PCR** Polymerase Chain Reaction

**Pfam** Protein Families Database

**RR** Response Regulator

**SD** Signaling Domain

**SPIRE** Searchable Planetary-scale Microbiome REsource

**SSD** Signaling Subdomain

**TCS** Two Component System

**Tfp** Type-IV Pili

**TM** Transmembrane

**UMAP** Uniform Manifold Approximation and Projection

**UPM** Universidad Politecnica de Madrid

# Introduction

# 1 Microbial chemoperception: sensing and responding to the environment

## 1.1 Microbial communities are everywhere

Microbial life is the invisible backbone of Earth's ecosystems, existing in every part of the planet and influencing all aspects of life as we know it[1]. The contributions of microorganisms* range from very visible processes, such as the production of the food we consume, to the complexity of orchestrating biogeochemical cycles[2]–[4]. Microbes shape the survival of macroscopic life, including humans[5], [6], animals[7] and plants[8], and also contribute to detoxifying the environment from pollutants[9]. Additionally, their involvement in biotechnological and therapeutic innovations[10] demonstrate their importance for global health and sustainability.

All of these microbial functions are performed in a wide range of ecosystems, from nutrient-rich areas to extreme conditions of heat, cold, acidity, and alkalinity found in different habitats such as seawater, soil, host-associated, geothermal or polar environments[11]–[13]. To survive in this vast array of natural settings, microorganisms have evolved sophisticated mechanisms for sensing and responding to a multitude of external stimuli. These sensing systems represent an ancient adaptation strategy for detecting and transducing signals evolved from simple one-component systems, already present in the last common ancestor of all modern life forms of life[14], to more sophisticated chemosensory pathways extended across prokaryotes.

## 1.2 Mechanisms of enviromental sensing: signal transduction systems

Prokaryotic organisms, including bacteria and archaea, need to monitor changes in their environment in order to survive. Successful adaptation depends on their ability to sense, interpret and react to external stimuli[15]. The environmental cues that microorganisms can detect are diverse and include temperature, pH, oxygen tension, osmolarity, nutrients, toxins, and other chemicals[16]. After signal recognition, signal transduction systems lead to a cellular response such as control of metabolism, quorum sensing, biofilm formation, or motility[17]. Signal transduction pathways

*Microorganisms include prokaryotes (bacteria and archaea), viruses, and some eukaryotes (fungi, protists and algae). However, for the scope of this thesis, the term is used to refer to bacteria and archaea exclusively.

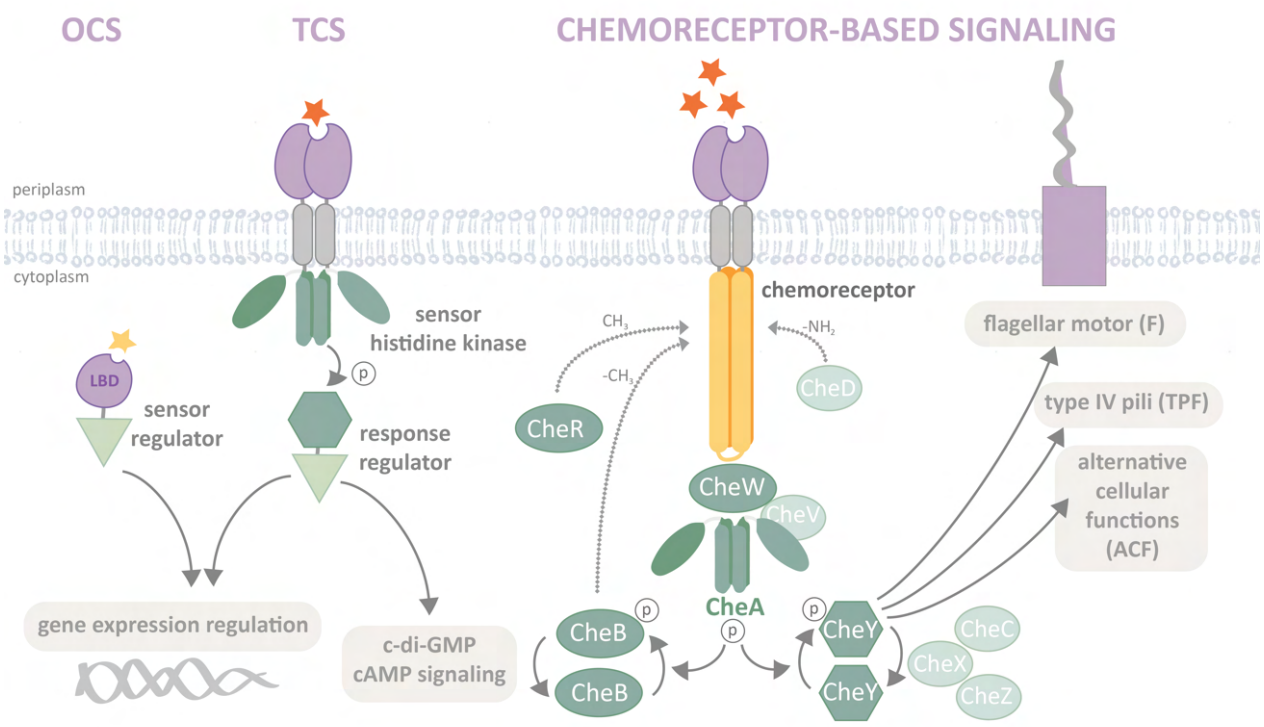



**Figure 1: Signal transduction systems: One component system (OS), Two components systems (TCS) and Chemoreceptor-based signaling pathway.**

can be classified into three categories of increasing complexity: one-component systems, two-component systems and the specialized chemosensory system[18]. The simplest form of signaling is the one-component system. These systems are highly distributed across prokaryotes, possibly encoded by the last common ancestor of archaea and bacteria[14]. It consists of a single protein that contains an input and an output domain. The input sensory domain is responsible for detecting internal or external signals, while the output regulatory domain initiates a cellular response, primarily regulating gene expression (Figure 1)[14]. This system has a fundamental role in adaptation to internal conditions as it is mostly sensing cytosolic molecules[14].

Two-component systems (TCS) represent a more complex system, probably evolved from one-component systems[19]. They are present in the three domains of life, but with the highest prevalence in bacterial genomes[20]. In TCS, signal detection and response are divided between two proteins: a sensor histidine kinase and a response regulator. The sensor kinase contains an input domain, detecting environmental signals, and a transmitter domain that phosphorylates the response regulator. The response regulator activates cellular responses through its output domain (Figure 1)[20], [21]. The evolution into TCS had the major advantage of the detection of extracellular

signals[22], a crucial adaptation for microbes interacting with their environments. It has been shown they are involved in the perception of physicochemical conditions, and recognizing quorum sensing signals[23]–[25].

Lastly, chemoreceptor-based signaling cascades or chemosensory systems are a specialized subset of two-component systems, with a sophisticated adaptation to environmental sensing[18]. In chemosensory pathways, chemical gradients are processed by dedicated chemoreceptors (CRs), also called methyl-accepting chemotactic proteins (MCPs). The signal recognition by the input domain of chemoreceptors triggers the signaling cascade that finally enables a cellular response. This pathway involves a complex molecular machinery (Figure 1), which will be explained in detail in the next section, and it is capable of sensing a wide range of signals thanks to the diversity of chemoreceptors' input domain. The canonical chemosensory system is known as chemotaxis, where the motility (*-taxis*) is the cellular response[16]. In the advance of the field other responses than motility (biofilm formation, regulation of concentrations of second messengers, etc.) using the same core pathway were reported, and the systems are broadly classified as chemosensory or chemoreceptor-based signaling systems[18], [26].

## 1.3 A sophisticated solution: Chemoreceptor-based signaling pathways

Chemotaxis is the process that allows prokaryotes to modulate their motility in response to chemical gradients, and it is the canonical function of chemoreceptor-based systems. This process is mediated by a complex network of interacting proteins that facilitate the movement of microbial communities toward attractants or away from repellents[27]. The core chemotaxis machinery is divided into two main pathways[28]: the excitation pathway, which directly influences bacterial motility[29], and the adaptation pathway, which allows for a "molecular memory" mechanism so the response of the system can be adjusted to persistent stimuli[30].

The excitation pathway is a rapid signal transduction, and initiates with the detection of environmental signals by methyl-accepting chemotaxis proteins (MCPs), typically known as the chemoreceptors (CRs). When an environmental signal is perceived, the responsible CRs change their conformation and initiate a cascade of reactions through the scaffold protein (CheW), which activates the histidine kinase CheA[31]. Activation of CheA leads to phosphorylation of the response regulator CheY (Figure 1), which increases its affinity for the flagellar motor and changes the swimming performance. The system is reset by dephosphorylation of CheY-P, involving several accessory phosphatases (CheC, CheX, CheZ)[32].

The adaptation pathway adjusts the signal processing of chemoreceptors by changing their methylation state. The regulation is performed by CheR, a methyltransferase that methylates specific glutamate residues of the chemoreceptors, and CheB, a methylesterase that is activated by CheA through phosphorylation[33]. This methylation feedback loop enables the adaptation to different chemical stimuli or variations in the concentrations. Additionally, in some pathways, CheD is involved in the deamidation prior to methylation, contributing to this adaptation mechanism and sometimes enhancing dephosphorylation activity of CheC. Some of the chemotaxis pathways include an additional component of the sensory complexes, CheV, which contain the CheW-like scaffolding domain and the CheY-like regulatory domain[18], [33].

The canonical chemotaxis system has been well described mainly in model organisms such as *Escerichia coli*, but research in recent years has revealed that the more complex additional machinery of the chemotaxis systems may regulate processes other than motility[18], [26]. For this reason, chemotaxis is broadly classified as a chemosensory system and these systems have been subdivided into different functional categories. The first category is the canonical chemotaxis system that regulates flagellar motility (F), the second category regulates type IV pili motility (TFP)[34] and, finally, alternative cellular functions are grouped into a third category (ACF). These ACF pathways regulate processes such as biofilm formation[35], cell development[36] or exopolysaccharide production[37] among others. All categories share the core of signaling genes, but are diversified with different phosphorylatable components downstream of CheA. The abundance of these systems and the diversity of their phosphorylatable components suggest a significant complexity in Che-like ACF signal transduction that is not fully understood yet[38].

## 1.4 Chemoreceptors, first line of the system

Chemoreceptors are the first line in chemosensory pathways, being the signal-sensing proteins of the systems[39]. They are also known as MCPs (Methyl-accepting Chemotaxis Proteins), due to their potential for being methylated. The term chemoreceptor (CR) will be used in this work, because most of the called MCPs have not been experimentally tested for methylation[40]. CRs keep the input-output domain nature from the one and two component system origin. One canonical chemoreceptor consists of a ligand binding domain or LBD (input domain), transmembrane (TM) helices, and a cytoplasmic signaling domain (output domain), which interacts with the histidine kinases of the signaling pathway[38]. Sometimes, the cytoplasmic signaling domain contains the HAMP domain (Histidine kinase, Adenylyl cyclase, Methyl-accepting and Phosphatase region), as the transmitter of the signal[41].

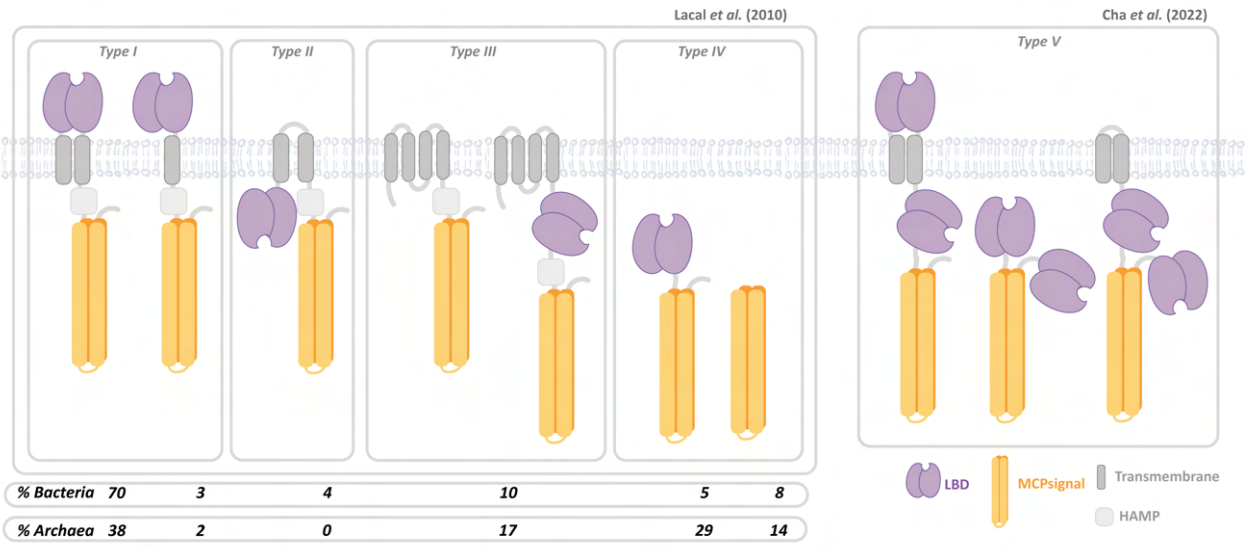



**Figure 2: Classification of chemoreceptors according to membrane topology and sensing domain position.** Adapted from Lacal *et al.* (2010)[42], Ud-dDin and Roujeinikova (2017)[39] and Cha *et al.* (2022)[43].

Although the majority of CRs belong to the canonical arrangement, their topology is very diverse[42]. While the signaling domain is always cytoplasmic, the sensing domain can be located on the extracellular or intracellular space (or be absent in some chemoreceptors)[40]. The presence and number of transmembrane regions can differ from zero to eight[39], which allows classifying CRs into several known topological classes. This type of CRs classification was first published in 2010 by Lacal *et al.*[42] and covered a relatively small number of sequences (3,524) compared to today's numbers, so there may be more topological classes yet to be discovered.

In Lacal's study[42] (Figure 2), the major class identified in Bacteria and Archaea was the two-transmembrane type of chemoreceptors, namely Type-Ia, where the LBD is located between the transmembrane domains and exposed to the periplasm. containing 70% of bacterial CRs). Type-I groups cover the majority of bacterial CRs ( 70% of the total), and given the position of its sensory domain, are mostly specialized in sensing extracellular signals. This class has two subgroups, Type-Ia and Type-Ib, with two and one TMs respectively. Type-Ia is described as the predominant topology, and it is classified in two clusters according to its LBD size.

A second major class of CRs is the Type-II, which has been described as the lowest in abundance, and also presents two TM regions with its LBD located in the cytosol, at the N-terminal end, so mostly specialized in sensing internal cellular signals. The following class, Type-III, represents 10% of bacterial sequences, and they are characterized by the large number of TMs present in its topology (from two to eight)[39]. Lastly, the cytoplasmic CRs Type-IV, lacking transmembrane

regions, are responsible for sensing the state of the cell. According to Lacal *et al.*, 13% of bacterial chemoreceptors are within this group, and most of them without LBD (Type-IVb)[42]. In a recent study of CRs diversity in Archaea[43], a new class was added to the classification system, termed class V (Figure 2), based on the observation of numerous cytoplasmic CRs containing a putative N-terminal LBD, and some also containing a C-terminal LBD.

This high diversity in structural arrangement strongly suggests the wide arrays of signals that can be measured by the CRs, and the constant evolution of these proteins for covering them. Based on their study of 244 bacterial genomes, Lacal *et al.* (2010) found that the average number of CRs per genome in chemosensitive bacteria is 14[42], [44]. More recent findings show that the number of CRs within a genome can vary enormously, ranging from one chemoreceptor, to ninety[44].

### Cytoplasmic signaling domain: MCPsignal

The cytoplasmic signal domain is the output domain of CRs. This domain is known as MCPsignal and is ubiquitous in CRs, serving as the exclusive criterion for annotating them[38]. It can be subdivided into a methylation subdomain (MH) connected to the signaling subdomain (SSD) across a flexible bundle (FB) (Figure 3).

This domain is essential for propagating the signal from the external environment to the intracellular response mechanisms of the system. The structure is highly conserved and consists of a long $\alpha$-helical hairpin that forms a coiled-coil four-helix bundle upon dimerization. The evolutionary pattern of this domain was defined by Alexander and Zhulin in 2007[38] for the insertion/deletions of seven-residue stretches, known as heptads (H), in both coils of the hairpin. These heptads, which correspond exactly to two-helical turns, help to stabilize the coiled-coil conformation, allowing the proper alignment and interaction of receptor dimers and trimers of dimers. CRs are classified based on the number of heptads, but the structure of the domain does not change between classes. The heptad classes are divided into seven major classes (44, 40, 38, 36, 34, 28, and 24 H), and five minor classes (48, 42, 52, 58, and 64 H). From the 2,125 sequences analyzed for establishing the heptad classes, 90% of them were included in the seven major ones. The heptads are located across the methylation subdomain and the flexible bundle, which could show a moderate degree of divergence compared to the highly conserved signaling subdomain (SSD) (Figure 3).

The methylation subdomain (MH) contains a consensus sequence for methylation, normally has two consecutive Glx (E/Q) residues[45]. The CheR enzyme methylates the residues while CheB demethylates it in the adaptation pathway to persistent stimuli. Methylation sites in CRs can vary in location and number. Connecting the methylation subdomain with the signaling subdomain is

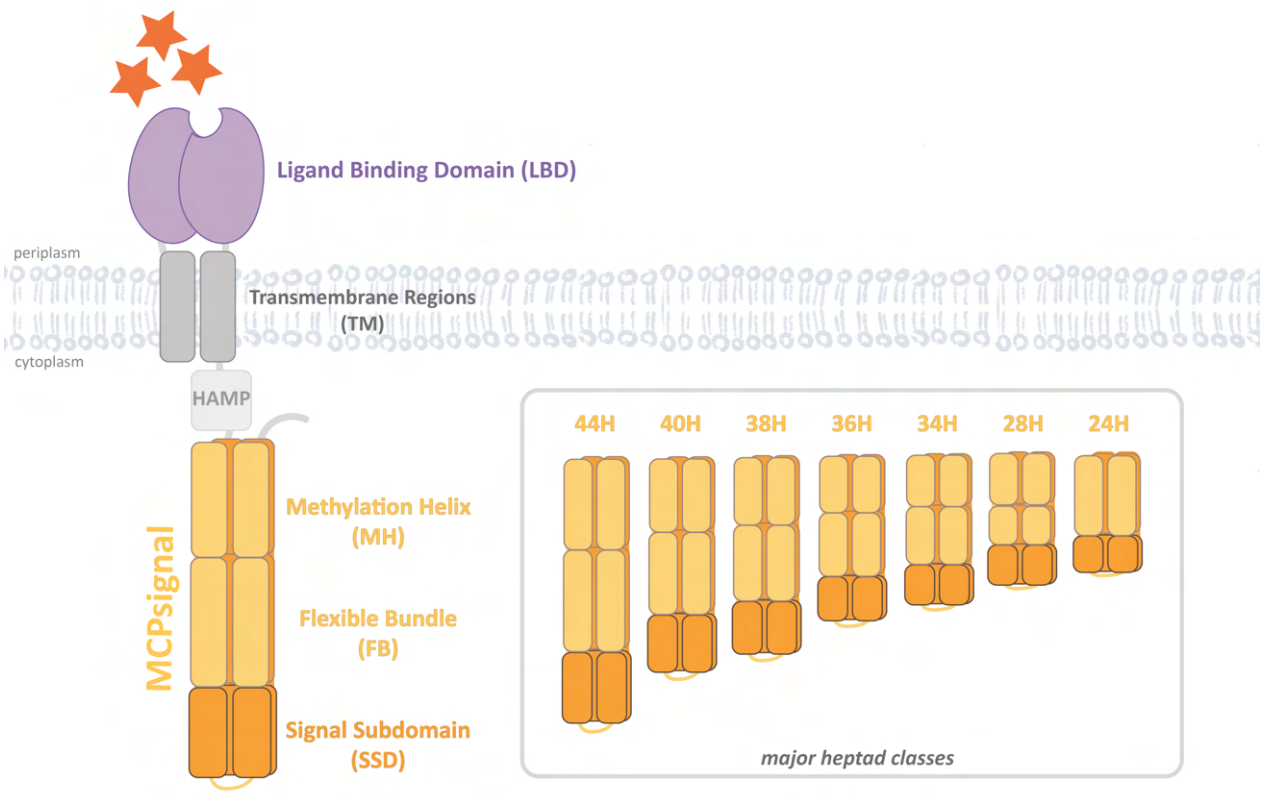



**Figure 3: Representation of the canonical chemoreceptor organization.** The extracytosolic sensing domain (LBD) is flanked by two transmembrane regions (TM). The signaling domain (MCPsignal) is in the cytoplasm, with the linker HAMP domain. The MCPsignal can be subdivided in the Methylation Helix (MH), connected by the Flexible Bundle (FB) to the Signal Subdomain (SSD). The major heptad classes of the signaling domain are represented.

the flexible bundle (FB). This subdomain is more variable in amino acid sequence, but contains conserved glycine residues that form a hinge and are essential for the trimers of dimers conformation of the chemoreceptor complexes[33]. Finally, the signaling subdomain (SSD) is the most conserved subdomain and interacts directly with the histidine kinases CheA and the scaffold protein CheW. It is present in all CRs at the tip of the cytoplasmic signal domain. This subdomain changes between two conformations, depending on the activated or inactivated state of the chemoreceptor.

## Sensory domain: The Ligand Binding Domain (LBD)

The input domain, also known as Ligand Binding Domain (LBD), is a conserved component across all the signal transduction systems (Figure 1). In one component systems, the LBD is sensing mainly cytoplasmic signals and modulating the affinity of the output domain to the gene promoter.

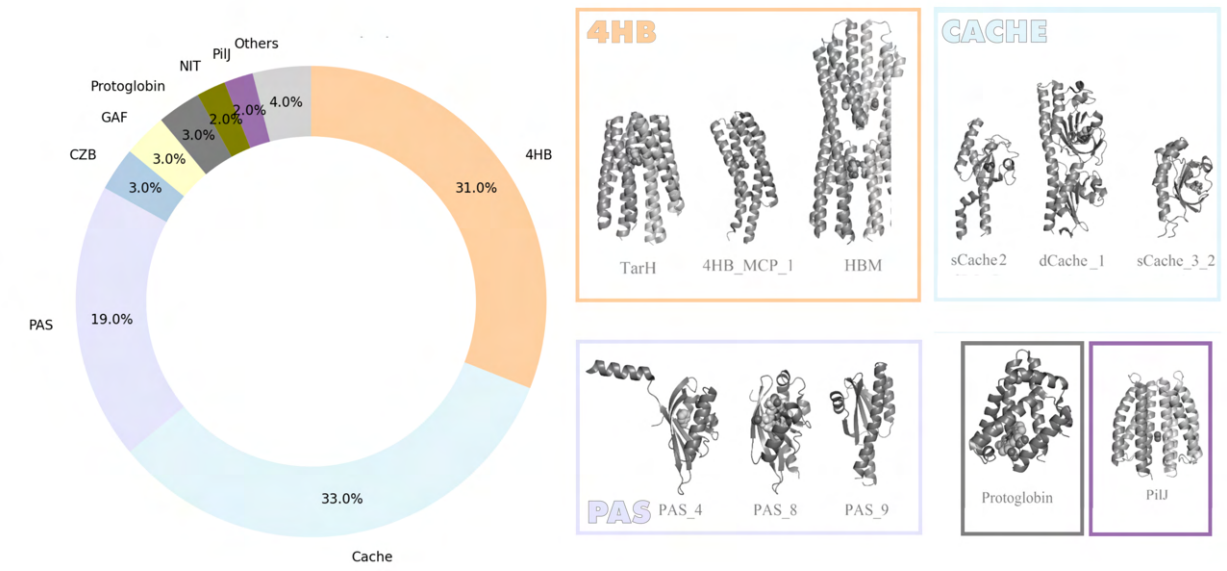



**Figure 4: Sensory repertoire distribution.** The distribution of each family of LBDs is shown, while some of the structural models within each LBD family are represented. Adapted from Ortega *et al.* (2017).

In TCS and Chemoreceptor-based signaling pathways, the LBD recognizes mainly extracellular stimuli, and it causes piston-like movements of transmembrane regions that modifies the autokinase activity[46]. This domain harbors the information for the recognition of a wide range of stimuli, from nutrients to toxins[47], [48]. While the MCPsignal domain is used to identify the CRs, their LBDs allow them to be classified on the basis of their function[47], [48].

The sensory domains are highly divergent, with a wide structural and sequence diversity. In contrast with the cytoplasmic signaling domain or other sensor histidine kinase with a high degree of conservation, LBDs are rapidly evolving domains[49].

### *LBD classification*

The most representative analysis of the diversity and abundance of sensory domains was done by Ortega *et al.* (2017) on a dataset of 15,872 LBD sequences[50]. In this work, they identified more than 80 different domains working as LBD, and only a few of them were found frequently, whereas the majority of them were uncommon (Figure 4).

The classification of LBDs types is done based on sequence signatures, such as the domains in the Pfam database. Most of the LBDs types can be subclassified by the structural perspective in $\alpha/\beta$ domains or four-helix bundle folds. Both structural types can be found with mono- and bimodular arrangements. The most abundant group is $\alpha/\beta$ domains Cache-type, where dCache_1 is the most

prevalent. Within Cache domains, several types are identified and grouped in sCACHE (single Cache) and dCACHE (double Cache, with two Cache domains present)[51]. This superfamily is widespread across bacteria, and have also been identified in archaea and many eukaryotes like humans[52]. Although the evolution of dCache domains were hypothesized to happen by duplication or fusion of monomodular sCache[53], a recent work suggests it could be the opposite way, so the monomodular sensing domain has evolved from the bimodular[54]. The known ligands for Cache domains include amino acids, sugars, organic acids, and nucleotides[48].

The other predominant superfamily of LBDs is 4HB (4-Helix-Bundle), present in all signal transduction systems. This family only sense extracytosolic signals, and according to Ortega *et al.* compromises the 31% of known LBDs (Figure 4). The study of this family is characterized for the Tar and Tsr CRs of *E. coli*, two well studied sensors which directly detect L-Asp or maltose, and L-Ser or quorum sensing signal respectively among other signals[33]. The Pfam models defining this superfamily are 4HB_MCP_1 (most abundant), TarH and Chase3, as well as HBM as a bimodular 4-helix bundle (Figure 4).

In the third place of abundance are PAS-domains containing CRs (Figure 4), also $\alpha/\beta$ domains with the exceptional feature of being exclusively cytosolic (Type-IV topology)[55]. This superfamily has a similar fold as Cache-type domains, but are defined by different sequence signatures. They are also in all the signal transduction systems in prokaryotes, and with the particularity of being spread across eukaryotes[56]. PAS-domains sense the cell oxygen or redox state in the cytosol. They can be bound to heme, FAD or flavin mononucleotide (FMN), and are involved in aerotaxis or energytaxis responses[50], [57], [58].

This three superfamilies are the 80% of the LBDs described by Ortega *et al.* (2017)[50], while other Pfam domains as CZB, GAF (found in cGMP-specific phosphodiesterases, adenylyl cyclases, and FhlA), protoglobin, NIT (nitrate and nitrite sensing), and PilJ represent 2% or 3% each one (Figure 4)[50] (they are summarized in Table 1.1). The majority of LBDs types are found in low abundance (less than 2% of total presence), and some of them can be found in enzymes and transporters with a known or predicted specificity. Finding them as putative LBDs could indicate a recent evolution in the adaptation to some conditions[50].

**Table 1: Summary of main LBDs.** Selection of LBDs of interest for this work from Ortega *et al.* analysis[50]. The description of each LBD model included in this table has been extracted from InterPro[59].

| LBD Superfamily | Description | LBD Model |
|---|---|---|
| 4HB | Four helix bundles that operate as a ubiquitous sensory module in prokaryotic signal-transduction. The 4HB_MCP_1 recognises citrate and TCA cycle intermediates. The aspartate receptor, TarH, is well studied in *E. coli*. The CHASE domain binds cytokinine. HBM is a double 4HB sensing Di- and tricarboxylic acids. | 4HB_MCP_1<br>TarH<br>CHASE3<br>HBM |
| CACHE-like | Cache domain family is an extracellular protein domain recognising a wide range of proteins, including the animal dihydropyridine-sensitive voltage-gated $Ca^{2+}$ channel alpha-2delta subunit, and various bacterial chemotaxis receptors. The majority of known ligands for dCache_1 domains are amino acids, whereas single Cache domains are described to sense organic acids. | dCache_1<br>Cache_2-Cache_3<br>dCache_2<br>dCache_3<br>sCache_2<br>sCache_3_2<br>sCache_3_3<br>CHASE4<br>CHASE8 |
| PAS | Per-Arnt-Sim (PAS) domain is a protein domain found in all kingdoms of life. They are described as intracellular domains, with a conserved three-dimensional structure (PAS fold). | PAS_3<br>PAS_9 |
| Other superfamily/ not assigned superfamily | The GAF domain is named after some of the proteins it is found in: cGMP-specific phosphodiesterases, adenylyl cyclases and FhlA. They present PAS-like Fold, and are described as sensing domain of a wide range of proteins from all species. | GAF |

*Continued on next page*

| LBD Superfamily | Description | LBD Model |
|---|---|---|
| | PilZ domain binds the ubiquitous bacterial second messenger cyclic-di-GMP (c-di-GMP), which is associated with the regulation of biofilm formation, the control of exopolysaccharide synthesis, flagellar- and pili-based motility, gene expression, interactions of bacteria with eukaryotic hosts and multicellular behaviour in diverse bacteria. | PilZ |
| | This zinc binding domain is found in bacterial proteins binding zinc with high affinity, with a Kd in the femtomolar range. This domain also binds the chemoattractant HOCl at a site very close to that of zinc. It has been shown that zinc participates in HOCl sensing by forming a redox 'Cys-Zn switch' that reacts towards HOCl. | CZB |
| | PilJ domain is linked to motility, being found on many type IV pili methyl-accepting chemotaxis transducer proteins. | PilJ |
| | Nitrate and nitrite binding domain involved in gene expression regulation, cellular motility and enzyme activity. | NIT |
| | Carbohydrate-binding module (CBM) is a domain found in carbohydrate-active enzymes. The CBM_4_9 sensing domain is linked to cellulose binding, while CBM27 binds to beta-1,4-mannooligosaccharides, carob galactomannan, and konjac glucomannan. | CBM_4_9<br>CBM27 |
| | This domain is named after three conserved tyrosines, and is part of the periplasmic sensor domain binding to unsaturated disaccharides. Its function is not known. | Y_Y_Y |
| | Novel variant of PAS-like Fold domains. It seems to detect small hydrocarbon derivatives as propanediol. | PocR |

*Continued on next page*

| LBD Superfamily | Description | LBD Model |
|---|---|---|
| | This protein family has not been functionally characterized. | DUF3365 |
| | FIST is a novel sensory domain, present in Bacteria, Archaea and Eukarya. It is suggested that FIST binds small ligands, such as amino acids. | FIST |
| | Protoglobin is part of the haemoglobin superfamily, and it acts as an oxygen sensor domain, or also can recognise cyanide. | Protoglobin |
| | Periplasmic binding protein (PBPs) are formed by two large lobes that close around the bound ligand. | Phosphonate-bd<br>Peripla_BP_4<br>SBP_bac_5<br>SBP_bac_3 |

## 1.5 Characterization of LBDs: a major challenge in CR research

Despite the significant progress in understanding LBDs, their structural and functional characterization is the major bottleneck in the characterization of CRs, as the majority of them remain unknown[51]. The central challenge in understanding chemoreceptor function lies in the identification and classification of their sensing domains, which are essential for the detection of environmental stimuli[50]. CRs sense pH state[60], redox state[61], [62] , or temperature[63] and a broad spectrum of ligands—including amino acids, organic acids, sugars, polyamines, nucleotides, and inorganic ions[48], [50]— but the majority of LBDs cannot be functionally annotated. Some experiments strongly support the functional implication of the sensing domains in the chemosensory pathways. For instance, the same LBD can be shared by different signaling receptors, supporting the idea that LBDs are independent modules that have been recombined during evolution to create novel proteins. Another example is how chimeric receptors, with LBDs from different signaling domains, maintain the function of the LBD[64], [65]. This evidence points to the need for functional annotation of LBDs to understand the evolutionary forces that have shaped chemosensory pathways.

One of the main reasons for the lack of functionally annotation of LBDs is the extensive sequence divergence within and across domain families[50], [55]. Current domain databases, such as Pfam, do not adequately capture the diversity of LBD subfamilies[50], [55]. In addition, a substantial

proportion of sensing domains fall outside the scope of existing domain classifications, so, despite their potential functional relevance, they are excluded from most analyses.

Finally, this challenge is amplified by the wide array of chemoreceptor topologies and their multidomain nature, adding complexity to the identification of sensing domains. Addressing this gap requires more robust methods to classify and analyze these uncharacterized sensing domains, which would enhance our understanding of chemoreceptor diversity and function. In Chapter 1 we address this challenge with the description of a *de novo* homology-based clustering approach, classifying LBDs type and recovering previously unknown subtypes.

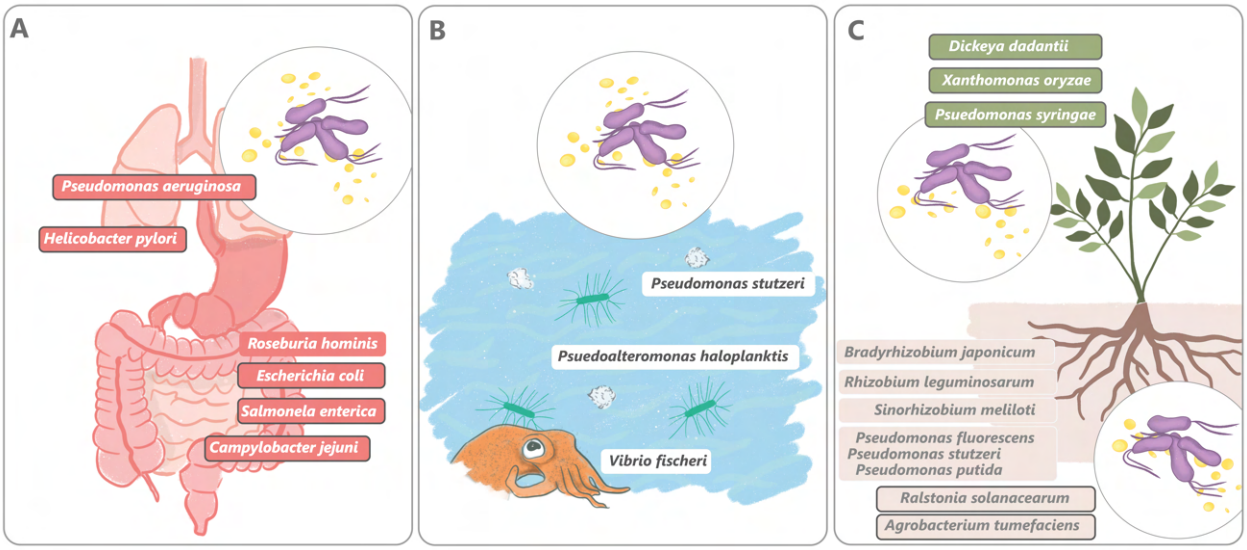



**Figure 5: Evidence of chemoperception importance across different environments.** a) Human-associated habitats, such as the gastrointestinal tract and the lungs. b) ocean environments, nutrient hotspots, phytoplankton association or symbiosis with the Hawaiian squid. c) soil and plant, with enriched areas around plant roots (rhizosphere), or on the aerial part of the plant (phyllosphere). Bacterial names enclosed in dark-lined squares indicate pathogens.

# 2 Ecological role of chemoreceptors

Chemoreceptor-mediated responses to environmental stimuli are fundamental to the survival and ecological fitness of microorganisms. The ecological role of chemoperception has been described in several natural settings, where the establishment of a host-interaction, or the ability to reach carbon sources in limiting environments depends on the chemoperception of the microbial communities[66]. These implications go from modulating the health or disease of human, animal or plant hosts, to mediating in the major biogeochemical cycles. The extensive variability in sensing domains observed across microbes is, in fact, a strong indirect evidence of the high diversity and complexity of the environments to which prokaryotes have adapted to.

In this regard, Lacal *et al.* (2010) first established a correlation between environmental complexity and the number of CRs in the genomes of prokaryotes inhabiting those environments[42]. For example, bacteria living in confined specific niches have fewer CRs than bacteria living in open environments. In this section, we summarized the importance of chemoperception in several complex and high impact ecosystems such as human associations, aquatic environments, and plant-associated niches (Figure 5); where the ecological role of chemoreceptor-based signaling pathways of several model organisms has been well described.

## 2.1 Chemoperception in human-associated communities

The human body is a well-studied, diverse and complex habitat where microbial colonization and interactions are essential and the role of chemosensory activity of some pathogens and symbionts has been characterized.[67]. In the human priority list of pathogens elaborated by the World Health Organization, which considers resistance to antibiotics and other factors, 60% of the pathogens possess a large array of chemosensory pathways[68].

Indeed, the gastrointestinal tract is a complex and dynamic environment with an enormous diversity of microorganisms[69]. Thousands of microbial species survive in the conditions of peristaltic mixing, varying pH levels, and with a steep of chemical and physical gradients. Gut microbial communities are known for the impact on host health, with implications in the immunity[70], drug metabolism[71], heart disease[72] or liver cirrhosis[73] among others. In this environment, some gastrointestinal pathogens rely on chemotaxis to achieve fitness and pathogenicity.

For example, *Vibrio cholerae*, the diarrheal pathogen that causes cholera, harbors 44 CRs and uses chemotaxis to locate nutrient-rich zones and avoid hostile conditions, facilitating its survival and pathogenicity[74], [75]. Similarly, *Escerichia coli*, *Salmonella enterica* and *Campylobacter jejuni* use chemotactic responses to navigate the gut environment, aiding their ability to cause gastrointestinal diseases. *S enterica* expresses CRs that detect serine, nitrate, and other host-derived compounds, allowing it to locate favorable niches within the intestines[76]. And *C. jejuni* has CRs that sense amino acids[77] and blood components[78].

The stomach presents a harsh environment with highly acidic conditions, where *Helicobacter pylori*, a chemotactic pathogen causing gastritis or peptic ulcers, uses chemosensory pathways to navigate the acidic environment and locate the protective mucus layer lining the stomach epithelium. Urea detection allows *H. pylori* to survive by using it as a substrate for urease to buffer the acidic environment[79]. Additionally, this bacteria uses CRs to detect blood components like lactate, which are indicative of tissue damage[80]. Omeprazole, a commonly used medication, has been shown to interfere in the chemotactic response of this bacterium[81], providing a proof of concept for chemotaxis-based therapeutic strategies.

The lungs are an environment where non-motile pathogens are more prevalent. Despite this, pathogens such as *Pseudomonas aeruginosa* show chemotactic behaviors conferring fitness advantages in early infection phases, especially in cystic fibrosis patients[82]. *P. aeruginosa* utilizes CRs to detect host signals, including inflammatory mediators like acetylcholine[83] or histamine[84], which help the bacteria localize infection sites within the lungs.

Apart from human pathogens, the gut microbiota harbors numerous symbiotic bacteria establishing beneficial relationships. As an example, *Roseburia hominis* uses chemotaxis for colonizing the host cells[85].

## 2.2 Chemoperception in aquatic environments

Oceans are the largest habitat on the planet, covering 70% of the surface and accounting for 97% of the Earth water[86]. Prokaryotic cells constitute the bulk of ocean biomass with the average presence of 1 billion microorganisms in every liter of seawater[87]. Marine microorganisms control the major biogeochemical cycles occurring in the ocean, and constitute the basis of the food web, as they access the dissolved organic matter (DOM) making it available for higher trophic levels. In addition, marine bacteria establish symbiotic and pathogenic relationships with marine animal or plant species. These major functions occur in the vast and dilute ocean, where chemosensory pathways are critical to locate nutrient-rich microzones or interact with other organisms to adapt to these environmental conditions[87].

Nutrient patches are common in the normally poor ocean waters, providing ephemeral low concentration spots of dissolved organic matter (DOM)[88], [89]. These patches originate from different scenarios such as phytoplankton degradation[90], cell lysate[88] or organic matter leaking from particles[91]; and represent a great energy source for those microorganisms that can quickly sense and move towards them. The rapid chemotactic response of *Pseudoalteromonas haloplanktis* to exposure to nutrient patches suggests that this strategy is a strong driver of the carbon cycling in the ocean, allowing the aggregation of these bacteria in nutrient-rich areas[92].

Mutualistic relationships between phytoplankton and bacteria are essential in aquatic environments as they control nutrient cycling and biomass production at the base of the food web[93]. The chemotaxis towards phytoplankton-derived DOM has been demonstrated using in situ chemotaxis assays with natural marine microbial communities including *Vibrionaceae*, *Flavobacteriaceae*, *Alteromonadaceae* and *Pseudomonadaceae* families[94]. Similarly, it has been shown that the heterotrophic diazotroph *Pseudomonas stutzeri* uses its chemotaxis system not only to move towards phytoplankton-derived DOM[95], but also to low-oxygen regions where their N-fixation pathways are optimal, contributing to the nitrogen cycle in the ocean[96].

Chemolithoautotrophic bacteria (*Hydrogenovibrio*, *Thiomicrorhabdus* and *Thiomicrospira* genus) present in hydrothermal vents or marine sediments harbor a large number of CRs despite the small size of their genomes[97].

A classic example of marine chemotaxis in symbiosis is the relationship between *Vibrio fischeri* and the Hawaiian bobtail squid (*Euprymna scolopes*)[98]. Chemotaxis enables the bacteria to colonize the light organ of the squid[99], following the secretion of chemical attractants by the squid and allowing the subsequent production of bioluminescence by the bacteria when the population is big enough, which is essential for the squid's nocturnal activities.

## 2.3 Chemoperception in plant-associated environments: phytopathogens and symbionts

Plants represent complex habitats colonized by diverse microbial communities, which are intricately linked to their health and productivity[100]. These microorganisms play critical roles in agriculture, where plant-associated bacteria can significantly impact crop yields[101]. Some bacteria associated with plants are pathogenic, and produce considerable losses in agricultural production. Other microorganisms are symbionts which promote plant growth and enhance the acquisition of carbon or nitrogen sources to the plant, among other functions. Plants and the microbial communities associated with them are suggested to work as a holobiont, where the microbiome is called the "second genome" of the plant[102]. Plant-microbe associations are found in different environments such as rhizosphere (soil adjacent to plant roots)[103], phyllosphere (aerial portion of the plant)[104] or endosphere (interior or the plant)[105].

The role of chemosensory pathways in the plant-microbiome interaction has been described in several model organisms living in the rhizosphere and phyllosphere. In the case of phytopathogenic bacteria, chemosensory genes seem to play an important role in the colonization and success of the infection. Among the 10 most relevant phytopathogenic bacteria[106], 90% presented chemosensory genes. Furthermore, in the analysis of 25 bacterial species of plant pathogens, Matilla and Krell (2018)[107] determined the mean number of chemoreceptor per genome as 33 genes, more than double of the average 14 bacterial CRs described by Lacal *et. al.* (2010)[42].

### Chemosensitive microorganisms in the Rhizosphere

The rhizosphere is one of the most complex environments on Earth, where sophisticated plant-microbiome-soil interactions occur[108] . The composition of the microbial communities is locally influenced by the plant, with the possibility of modifying the soil pH, soil structure or oxygen availability, and providing carbon-rich exudates as energy source[108], [109]. The bulk soil communities

had been evolutionarily selected to interact with the roots, and are the first line of defense and the gateway for uptaking nutrients for the plant[110]. In the context of these interactions, the chemosensory pathway has been studied as a key feature for establishing the association in pathogenic[107] and nonpathogenic relationships[111].

In the case of symbiotic relationships, some well known bacteria rely on chemoreceptor-base signaling pathways to establish nitrogen-fixing symbiosis with their hosts. For example, *Sinorhizobium meliloti* shows chemotaxis activity towards carboxylates detected in the alfalfa root, or towards proline, which is in high concentration in the host seeds[112]. *Rhizobium leguminosarum* can detect flavonoid inducers of the symbiotic nodulation genes[113], and a mutation of some CRs reduces the ability to form nodules on peas[114]. *Bradyrhizobium japonicum* chemotactically finds soybean roots to form nitrogen-fixing nodules[115].

Similar observations have been made for *Pseudomonas putida*[116], *Pseudomonas fluorescens*[117] or *Pseudomonas stutzeri*[118], which efficiently colonize roots using chemosensory pathways. The metabolic versatility of *Pseudomonas* species is reflected in the large number of CRs harbored in their genomes (more than 25), as well as their ability to detect an extensive repertoire of chemoeffectors (organic and amino acids, sugars, fatty acids, peptides, metal ions, herbicides, etc.)[119].

The rhizosphere is also the habitat for chemosensitive phytopathogenic bacteria. An emblematic example is the species *Ralstonia solanacearum*, which is attracted by the tomato host, and the mutation in their CRs reduce bacterial infection in the plant[120]. Similarly, *Agrobacterium tumefaciens* can detect potential entries to the plant tissues by sensing phenolic compounds excreted from the wounded areas[121], [122].

### Chemosensitive microorganisms in the Phyllosphere

In the phyllosphere, the above ground part of the plants, many specific challenges exist, including overcoming nutrient limitation, ultraviolet light, temperature changes, competition with other microbes or the lack of water[123]. In contrast to the rhizosphere, the life cycle of a perennial plant comprises the change of leaves every year, so it is considered a short-lived environment[104]. Under these conditions, epiphytes (organisms that colonize the surface of plants) get a fitness advantage from having flagellar motility or forming biofilms[124]. The epiphytic lifestyle also represents the initial stage of foliar colonization by many bacterial phytopathogens, preceding the entry into the leaf apoplast via wounds or natural plant openings (e.g., stomata)[107]. In the transition

from epiphytic to endophytic stage of some bacteria, the chemosensory pathways are defined as a key process.

In *Xanthomonas oryzae* the epiphytic entry and virulence is dependent on the chemotaxis system that guides the bacteria towards components of the xylem sap[125]. In the case of *Pseudomonas syringae*, classified as the most relevant phytopathogenic bacteria[106], the mutation of CRs reduces the virulence in tomato plants[126], [127]. Similarly, *Dickeya dadantii* shows chemotaxis to jasmonate and xilose, which is essential for the entry through wounded tissues in *Arabidopsis*[128], and mutation on nitrate/nitrite chemoreceptor reduces the entry in potato plants[129].

## 2.4 Challenges in chemoreceptors research across ecological niches

The ecological role of CRs is critical for bacterial survival and adaptation in a wide range of environments, from human-associated habitats to oceans and plant associations. However, our understanding of the global ecological and evolutionary patterns of chemoperception remains largely incomplete. Current knowledge comes from niche-specific studies and many useful but also anecdotal examples like the ones described previously, which refer to a few model organisms. Today, we know that this represents a very small fraction of the microbial biosphere. Bacterial communities are highly complex, diverse and dynamic; and the ecological roles of chemoreception could only be addressed by studying these communities as a whole. Thousands of yet uncultured microorganisms populate natural habitats, and we still ignore the prevalence and potential novelty of chemosensory systems in such environments, leaving even the most basic questions unanswered:

- What are the global patterns of chemoreceptor profiles across different prokaryotic species and environments?
- What mechanisms drive the evolution of CRs in microorganisms occupying distinct ecological niches?
- How does the chemoreceptor repertoire relate to specific microbial lifestyles?

These questions are crucial for advancing our understanding of chemoperception and its role in microbial ecology and are explored in Chapters 1 and 2 of this thesis.

# 3 Chemoperception from a genomic and metagenomic perspective

The molecular and cellular mechanisms of chemosensory pathways have been studied in detail on model species such as Escherichia coli, whose chemotaxis is considered one of the best-understood behaviors in biology[33], [130], [131]. Advances in capillary assays[132], microfluidic experiments[133], [134], or computational modeling[135] have become strategic tools to deepen in the study of chemosensory pathways. However, these methods are primarily intended to investigate single, or small groups of organisms. Despite recent advances in genomic and metagenomic approaches, the scale and diversity of the chemosensitive organisms remains barely explored.

On one side, thousands of bacterial and archaeal genomes are fully sequenced and populate public databases, providing an excellent ground for investigating the prevalence, diversity and evolution of chemosensory pathways on all known major microbial lineages. Furthermore, shotgun metagenomic sequencing offers a robust approach to studying microbial communities directly from environmental samples without culturing, expanding the research of chemotaxis towards a yet unknown corpus of novel uncultured species adapted to diverse natural habitats. To date, only a few metagenomic studies have examined chemosensory pathways at the metagenomic scale, both of which emerged during the course of this thesis[136], [137]. This section provides an overview of -omic research performed on chemotactic microorganisms, and explores how recent technological advances in the genomic and metagenomic fields can be used to expand our understanding of chemosensory pathways, addressing crucial gaps in current research.

## 3.1 First steps in understanding the microbial chemoperception: from phenotype to genotype

The first insights into microbial behavior came from the ability to see the invisible microorganisms thanks to the microscope. Leeuwenhoek reported seeing his oral microbiota for the first time in 1676[138], where he already noticed bacterial motility. The importance of microorganisms emerged 200 years later when Robert Koch defined their role in major diseases[139] and developed a method for culturing them in nutrient-rich media. This boosted the development of microbiology in the nineteenth century[140], and led to the first systematic studies of how bacteria moved in a directed way, publishing the first works on chemotaxis[141].

Technological advances in microbiology, such as the development of growth media and the improvement in the microscope, allowed scientists to study the behavior of motile bacteria in more detail[142]–[144]. These early methods enabled researchers to explore how bacteria move in response to specific stimuli, such as oxygen, light, or chemical attractants. The improvement of capillary assays and the advances in molecular biology led to the first major studies of *Escherichia coli* around 1960, setting the bases for modeling the chemosensory pathways[145]. This period marked the beginning of systematic studies into bacterial chemoperception[146].

However, the paradigm of how to study microorganisms changed in 1944 with the discovery of the genetic information within the DNA sequence[147]. By analyzing the DNA sequences of organisms, scientists could infer evolutionary relationships and functional capabilities, advancing our understanding of microbial communities beyond what was possible with traditional culture techniques. DNA sequencing not only enabled the identification and characterization of genes involved in chemotaxis but also illuminated the genetic underpinnings of previously unrecognized pathways. For example, decades of genetic and biochemical studies had identified four chemotaxis transducers in *Escherichia coli*, but the discovery of a fifth transducer (Aer) only became possible during the genomic era when its gene was revealed through the *E. coli* genome sequencing project[148]–[151].

## 3.2 DNA sequencing becomes the new microscope: Genomics provides a comprehensive view of chemosensitive systems

How genetic information is stored and replicated laid the foundation for modern molecular genetics. After the demonstration that DNA carried genetic information, the further investigations led to the possibility of sequencing the strands[152][153]. The improvement and development of the sequencing technology by Sanger[154] allowed the first prokaryotic genome sequence of *Haemophilus influenzae* in 1995 [155], followed by many others. In this context, a new field of research, termed genomics, emerged, enabling the comprehensive study of the entire DNA content of an organism. Under this new genomic era, the knowledge of chemosensory pathways increased, not only in the canonical *Escherichia coli* pathway [156], but also on other model species such as *Bacillus subtilis* [157] or *Campylobacter jejuni* [158]. Notably, the whole-genome sequencing of twelve microbial species revealed over 100 chemoreceptor sequences in public databases by 2000[148], enabling broader comparative studies and a better understanding of the chemotaxis system beyond *E. coli*.

While the Sanger sequencing method represented a significant advance in the field, it was also highly complex and costly to scale up, providing a relatively slow progress. In fact, the advances in

sequencing technologies over the next 40 years were critical in establishing more solid insights on the molecular basis of microbial chemoperception and interactions with the environment. In 2005, Next Generation Sequencing (NGS) transformed genomics by enabling faster, cost-effective sequencing with millions of parallel short reads[147]. This was achieved through advances in microfabrication, imaging, and computational power[159]. Key platforms included Roche/454, Illumina/Solexa, and ABI/SOLiD.

NGS allowed complex DNA to be sequenced in parallel, surpassing Sanger methods. These high-throughput technologies have exponentially increased the number of prokaryotic genomes sequenced that are publicly available, enabling deeper analysis on microbial diversity and genomic content. In the field of chemoperception, genome-wide studies have identified a wide range of CRs and signaling pathways in bacteria and archaea. The most comprehensive genomic analysis in CRs was performed in 2007 with the study of 2,125 MCPsignal domains of 152 prokaryotic genomes[38], followed by the work by Lacal *et al.* in 2010, which expanded the analysis to 3,524 CRs of 262 strains[42]. In the case of the whole chemosensory pathway, the study of Wuichet and Zhulin in 2010 [18] established the basis of the classification of chemosensory systems, which was based on 450 prokaryotic genomes. The latest work on sensing domains included 15,000 LBD sequences and was done by Ortega *et al.* in 2017 [50], and dedicated databases exist that catalog genes involved in signal transduction systems on increasing number of bacterial genomes every version (MIST database[137].). However, the reference genome database such as the NCBI have grown exponentially during the last decade (Figure 6), currently holding more than 20,000 full-length prokaryotic genomes. This includes hundreds of microbial lineages and novel species for which chemotaxis systems have never been explored. Moreover, we still lack a global perspective on the prevalence, specificity and evolution of chemosensory pathways across the entire prokaryotic phylogeny [26].

Today, the most widely used platform remains to be Illumina[147]. Further developments, like Helicos Oxford Nanopore and PacBio technologies shape the third generation of NGS, and are capable of sequencing single molecules and longer reads. The potential applications and benefits of this new generation of technologies remain fully unexplored to CR research.

## 3.3 The emerging role of metagenomics in chemoperception research: challenges and opportunities

While genomics has expanded our understanding of chemosensory pathways in culturable organisms, the vast majority of microorganisms remain inaccessible through culture-based tech-

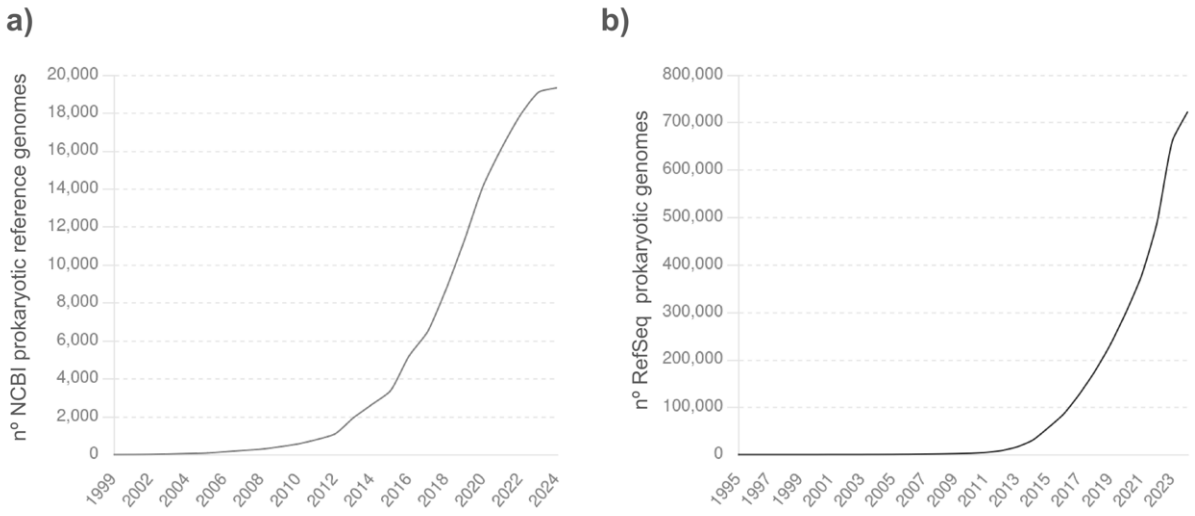



**Figure 6: Prokaryotic reference genomes increase over the years.** extracted from prokaryotes.txt of GenBank downloaded in June 2024. a) Number of prokaryotes reference genomes in NCBI, currently 19,328. b) Number of genomes available in RefSeq database.

niques[160]. One of the earliest evidence of this phenomenon was known as the "great plate anomaly" [161], which revealed that less than 1% of microorganisms in a sample can be cultured. This represented a complete shift in our understanding of microbiology in general, with the majority of microorganisms having been previously unisolated and unstudied.

Despite the vast increase in metagenomic data, the study of chemosensory systems in uncultured microorganisms is still in its very early stages. Most CRs identified to date come from culturable bacteria, limiting our understanding of how chemoperception operates in the majority of microbial life. During the progress of this thesis, the first few insights of chemosensory pathways in metagenomic data arose, such as the recent incorporation of metagenomic sources in the MIST 4.0 database[137] and the identification of CRs from human gut microbiome [136]. A recent study applied in situ chemotaxis assays to natural marine microbial communities, offering a glimpse of chemoperception in uncultured microorganisms, but still distinct from large-scale metagenomic efforts [94]. In the following sections, we briefly introduce the latest advances and current state of metagenomics research, illustrating its importance in the global understanding of the microbial world and its potential impact on the field of chemoperception.

## Environmental DNA: the uncultivated majority

The great advances in the last decades such as polymerase chain reaction (PCR) and next-generation sequencing (NGS) with massive parallel sequencing, unlocked the exploration of complex

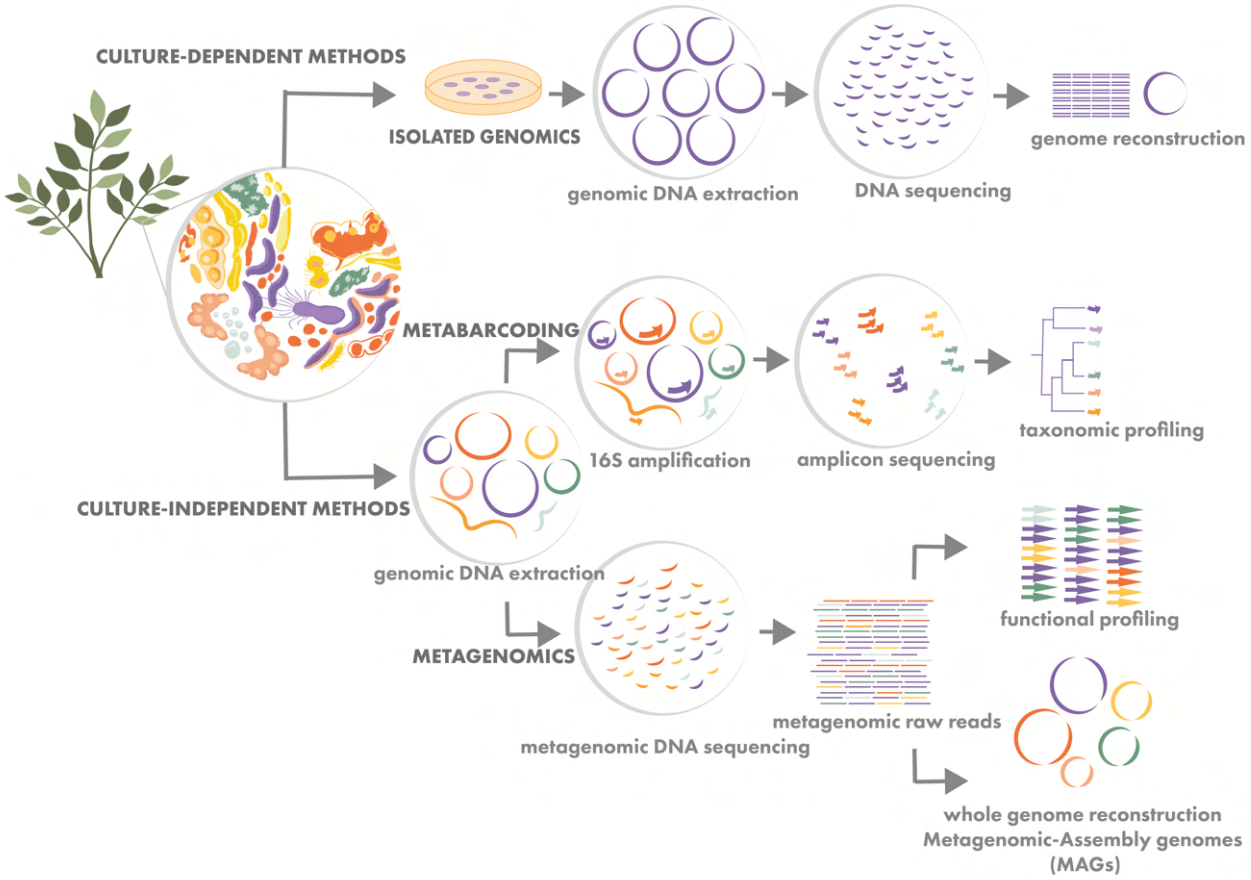



**Figure 7: Culture dependent and independent methodology.**

samples of DNA extracted directly from the environment. These new approaches bypassed the need of culturing isolated microorganisms, and the whole microbial communities could be studied directly from their natural habitats (Figure 7). These techniques are known as environmental DNA approaches (eDNA).

The first approach for studying the microbial communities directly from the environment was the study of the 16s ribosomal RNA (rRNA) as a phylogenetic marker [162]. This technological advance, known as metabarcode, allowed the first classification of prokaryotes based on this technology in 1977, changing completely our knowledge of the microbial world with the discovery of the Archaea domain as a separated group from Bacteria and Eukarya[163]. In the subsequent years, thanks to the invention of the PCR (1986), the 16S rRNA gene could be amplified directly from environmental DNA, which unveiled a great diversity of uncultured microorganisms. The resolution of the 16S taxonomic profiling was illustrated with the increasing number of prokaryotic phyla described, from the 12 phyla described in 1987[164] to the 194 phyla listed in the Genome

Taxonomy Database nowadays (125 of these being uncultivated taxa[165]). This was a great advance in the understanding of microbial communities, but it was limited to providing information about phylogenetic relationships, without revealing the functional capabilities of the organisms in question.

Metagenomics, introduced in 1998 by Handelsman *et al.* [166], is the methodology used to characterize the entire collection of genes and genomes of a microbial community. The methodology used is shotgun metagenomics, which consists of non-targeted sequencing of all DNA extracted from environmental samples (see Figure 8)[167]. Metagenomics provided a global view of the genetic content of all DNA in a sample, shedding light on the functional and metabolic diversity of the microbial community. The initial studies early demonstrated the power of metagenomics to explore the microbial diversity. For instance, in 2000 the bacteriorhodopsin gene was first discovered in DNA samples of seawater[168]. In 2004, 1.2 million novel genes and 148 novel phylotypes were identified in a single study of the microbial diversity of Sargasso Sea surface water[169].

### Major advances and data integration in metagenomics

In early metagenomic approaches, the environmental DNA was cloned into vectors and sequenced with Sanger methodology[170], which was really expensive ($500 per Mb)[171]. However, with the advent of next-generation sequencing (NGS) and improved bioinformatics techniques, the field became more accessible and an explosion of studies across a wide range of environments enabled the discovery of the functional and phylogenetic diversity of all domains of life [172]–[175].

Since then, metagenomic research has led to significant discoveries about the ecological roles of microbial communities [176], [177]. Major findings include the identification of uncultured bacterial phyla with endosymbiotic behavior [178] or species that perform complete nitrification of ammonia [178], [179]. Findings from hydrothermal vents on the ocean floor have reinterpreted our knowledge of the relationship between prokaryotes and eukaryotes, with the discovery of an archaeal superphylum called Asgard that contains genes previously thought to be exclusive to eukaryotes [180], [181]. In addition, some important discoveries are related to human disease, such as the strong association of the microbiome with inflammatory bowel disease [182] or with neurological disorders [183]. Metagenomics has also enabled advances in diagnostics [184] and personalized medicine [185] with major implications for human health.

Efforts to integrate the vast amounts of data generated by individual studies have led to the development of comprehensive databases. The Global Microbial Gene Catalog (GMGC)[186], published in 2022, compiles data from over 13,000 public metagenomic studies, creating a non-

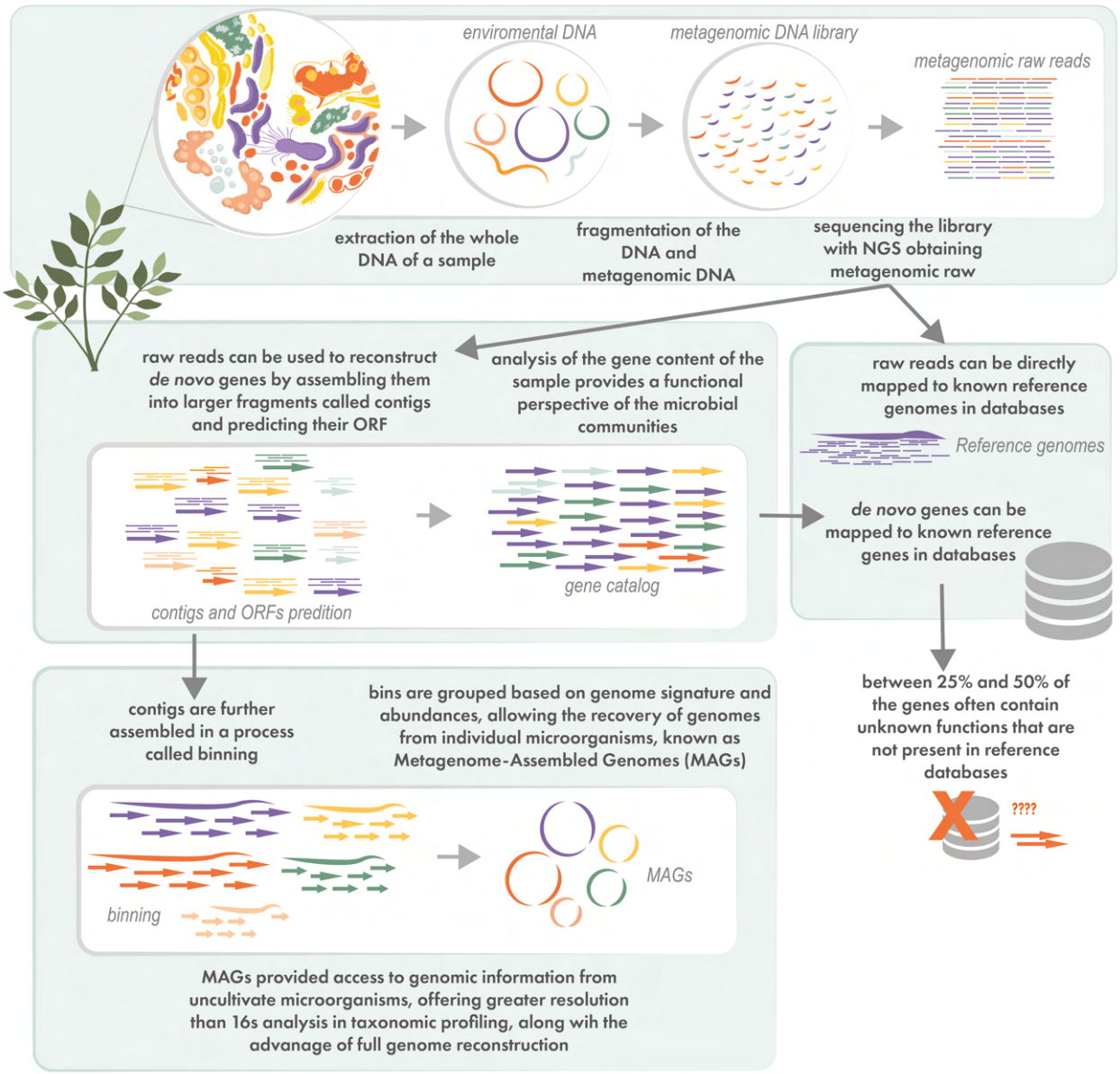



**Figure 8: Shotgun metagenomics pipeline.**

redundant gene catalogue of more than 303 million genes. Another recent initiative, SPIRE (a Searchable, Planetary-scale mIcrobiome REsource)[187], released in 2024, is a comprehensive multi-habitat genomic dataset covering the entire prokaryotic phylogeny. It integrates 99,146 metagenomic samples from 739 studies across environments, using a 'microntology' framework to describe microbial habitats. SPIRE includes 35 billion predicted protein sequences and 1.16 million MAGs, clustered into 133,402 species-level genomes.

In terms of taxonomic classification, the Genome Taxonomy Database (GTDB) has made important contributions in the field, merging cultured and uncultured taxonomic species[165]. Unlike traditional taxonomies such as NCBI, the GTDB framework provides a comprehensive classification performed in all sequenced prokaryotic genomes and high quality MAGs, based on evolutionary distances calculated from the conserved marker genes and on the reconstruction of a phylogenetic tree. The number of species has increased by 358% in the last 5 years (Table 2), representing almost a 6-fold increase in the current number of reference genomes in NCBI (Figure 6).

**Table 2: GTDB release number of lineages per rank.**

| Rank | 2019 | 2020 | 2021 | 2022 | 2023 | 2024 |
|---|---|---|---|---|---|---|
| Phylum | 128 | 129 | 146 | 166 | 181 | 194 |
| Class | 332 | 369 | 407 | 477 | 548 | 602 |
| Order | 912 | 1,020 | 1,279 | 1,571 | 1,772 | 2,006 |
| Family | 2,207 | 2,558 | 3,222 | 4,070 | 4,772 | 5,434 |
| Genus | 7,906 | 9,428 | 12,888 | 16,686 | 20,739 | 24,959 |
| Species | 24,706 | 31,910 | 47,894 | 65,703 | 85,205 | 113,104 |

Despite the rapid growth in the number of environmental sequences, generating petabytes of data in recent years[188], between 85-99% of the microbial diversity is still unknown [160]. Almost every newly sequenced metagenome reveals some novel MAGs and thousands of unknown novel genes[186]. This ongoing discovery highlights the vast uncharted territory in microbial ecology and the potential to uncover novel genes related to chemoperception. For instance, a novel gene identified from a metagenomic catalog did not map to any known gene but was located within a chemotaxis operon and was experimentally validated to play a role in bacterial motility[189]. The amount of new species and ecological data provided by metagenomics holds a vast potential in uncovering the full diversity and ecological significance of microbial chemosensory systems, moving beyond the cultured-based methods limitations. In Chapter 2 of this thesis we explore the potential of metagenomic studies in the knowledge of chemoperception field.

## 3.4 Low abundance microorganisms: beyond standard metagenomics

Microorganisms harboring chemosensory systems are commonly found in low abundance in natural settings[190], [191]. Although metagenomics has greatly advanced our understanding of microbial communities and their ecological significance, one of the main limitations is that sequencing data usually represent the most abundant species in the microbial community. Even attempts of deep sequencing of the sample seem to provide only a glimpse of the genetic variability present in the less abundant microorganisms[192]. The fraction of these microorganisms that are invisible to current methods is known as 'rare biosphere' [193], and it has a great impact on the microbiome responses to environmental changes[194]. It is well known that rare biosphere can modulate major biogeochemical cycles of the environment and strongly influence the tolerance of natural habitats to climate change[195], [196]. The extent and ecological significance of the functional and phylogenetic biodiversity contributed by low abundance microorganisms on a global scale has not been systematically investigated, and it is a completely unexplored field in the context of chemoperception. Chapter 3 of this thesis offers the first approach to deepen into the analysis of low-abundance chemosensitive species.

# Objectives

Despite the importance of chemoperception in microbial communities, most CRs described to date belong to cultivable bacteria and lack functional and ecological information. However, the number of sequenced prokaryotic genomes and metagenomes, which are likely to include many chemosensing species, has increased exponentially over the last decade, and up to 99% of the functional and taxonomic biodiversity revealed remains unknown[160]. Furthermore, it is hypothesised that a large number of chemosensing microorganisms exist at very low abundances in natural habitats, which hinders their in-depth analysis using current metagenomic approaches.

We hypothesise that with the recent availability of genomic and metagenomic samples and modern third-generation sequencing technologies, a global analysis of CRs across different ecosystems and along the entire prokaryotic phylogeny, including low abundance organisms, is possible.

Furthermore, given the multi-domain nature of CRs, we argue that ecological pressures and evolution should be studied separately for each domain, as they may represent different functional entities.

With this antecedents, in this thesis we address four specific objectives:

1. To develop a robust and automated methodology for the characterisation and functional classification of CRs based on the different domains at the genomic and metagenomic scale.

2. To study the prevalence and specificity of CRs across the prokaryotic phylogeny (cultured and uncultured microorganisms) in different habitats, with a particular focus on plant-associated microorganisms.

3. To study the evolutionary history of CRs on a global scale, including uncultivated chemosensitive microorganisms.

4. To develop a novel method to study low abundance chemosensitive species in environmental samples.

# Materials and methods

# 1 Prevalence and Specificity of Chemoreceptors Profiles in Plant-Associated Bacteria

## 1.1 Chemoreceptors (CRs) sequence retrieval

82,277 CR sequences were obtained from the genomes of 11,806 representative species in the proGenomes2 database [197]. The representative species in proGenomes2 are the result of a phylogeny-based classification of all RefSeq [198] genomes, where species delineation is based on a systematic phylogenetic threshold (i.e., <95% divergence in 40 universal marker genes) rather than relying on the NCBI taxonomic names. Although this might lead to inconsistencies with the current NCBI Taxonomy names for strains and species, it better represents the genomic definition of species, as well as providing a standardized classification system [199], [200]. To identify CRs in our set of representative genomes, all the sequences matching the MCPsignal Pfam domain signature (PF00015) were retrieved using HMMER 3.1b2 [201], Pfam-A 31.0 [202] and the specific gathering threshold provided for the MCPsignal HMM Pfam model. Multiple hits were resolved by retaining the match with the highest bit score. In analogy to previous studies [38], [44], [50] the presence of an MCPsignal domain in the sequence was the only criterion used for CR identification.

## 1.2 Ligand binding domain (LBD) extraction

For each CR sequence, transmembrane regions (TMs) were predicted using TMHMM2 [203]. The position of the TM region(s) was used to infer the putative extracellular LBD regions, which were subsequently annotated using the Pfam domain database. When no significant Pfam matches were found, LBD sequences were labelled as “unknown”. Two different topologies of extracellular LBDs were considered: (a) sequence regions flanked by two TM regions, and (b) sequence regions located between one TM and the N-terminal sequence. In both cases, sequences shorter than 30 amino acids were discarded. Intracellular LBD regions, as well as potentially overlooked extracellular LBDs (e.g., due to undetected TMs), were inferred based on the detection of Pfam domains other than the MCPsignal and HAMP domains. Pfam mappings were performed using HMMER [201] searches as implemented in eggnog-mapper version 2.0.5 [201], [204]. When more than two domains mapped to the same region, the best hit was selected. The final dataset contained 72,480 LBD sequences.

## 1.3 Clustering of LBD sequences

*De novo* homology-based clustering of the 72,480 LBD sequences was inferred using MMseqs2 [205] with an e-value threshold of 0.01, 20% minimum identity, and 50% minimum query coverage. These parameters were chosen to maximize remote homology detection and to infer LBD clusters with broad phylogenetic divergence (i.e., distant homologues) while still grouping sequences with a common evolutionary origin. The MMseqs2 command used was “mmseqs cluster -c 0.2 –minseqid 0.5 –cov-mode 2”.

## 1.4 Construction of the plant-associated bacteria (PAB) databases

A curated list of PAB was manually curated from the 11,806 representative species. As a first filter, we used the habitat information (i.e., “host plant-associated” label) provided by proGenomes2, which is based on the PATRIC database [197]. The resulting list was reviewed manually to exclude uncertain or incorrectly annotated entries by checking their metadata and associated literature. Additionally, we included other known plant-associated species on the list that were missed by the PATRIC database but that were considered PAB based on published data. In total, we identified 960 reference species (PAB-broad) that could be considered related to the plant-environment. From this list, we extracted two sub-databases (Table S1): phytopathogens (PAB-phyo, 119 members) and plant symbionts (PAB-symb, 192 members).

## 1.5 Degree of Plant Specificity (DPS)

A specificity score for plant-association was calculated for each LBD type based on the proportion of PAB species present in each LBD cluster. We calculated three score values, which we refer to as the Degree of Plant Specificity (DPS), depending on the PAB reference database used: DPSbroad, the proportion of PAB-broad species in each LBD cluster; DPS-phyto, the proportion of PAB-phyto species; and DPSsymb, the proportion of PAB-symb species.

## 1.6 Phylogenetic tree reconstruction and visualization

Multiple sequence alignments were built for each cluster using Clustal Omega v1.2.4[206], and phylogeny was inferred by IQ-Tree v1.6.12 using the default parameters[207]. The trees were

further analysed and visualized using ETE3 v3.0[208], with custom Python scripts integrating the annotations of each sequence for its taxonomy, domain architecture, sequence alignment and plant-specificity prediction (DPS).

## 1.7 Phylogenetic signal tests

The phylogenetic signal tests were performed using the $\delta$-approach[209], a phylogenetic analogue of the Shannon entropy that measures the degree of phylogenetic signal between a categorical trait (trait vector) and a phylogeny (metric-tree). We used the $\delta$-approach to specifically test the null hypothesis that a given taxonomic distribution of a LBD follows the phylogenetic signal of the underlying species, which provided us with a p-value for each LBD cluster. We applied 100 iterations per test and set the p-value threshold at 0.05.

The species phylogeny used as a reference in all the tests was reconstructed using the ETE3[208] supermatrix-based workflow and a concatenated alignment of 40 universal marker genes [210] extracted from the 11,806 species-representative genomes using the FetchMG tool [211]. Multiple sequence alignments were inferred using Clustal Omega v1.2.4 [206] and phylogenetic reconstruction was performed with Fasttree v2.1[212]. Moreover, an alternative species phylogeny including only genomes with at least one CR was reconstructed using the same methodology. As the $\delta$-statistics has poor sensitivity in detecting the phylogenetic signal for small taxon sample sizes (<20 taxa), LBD clusters mapping to reference phylogenetic tree nodes smaller than 20 leaves were discarded from the analysis (Table S2).

# 2 Adaptative value and evolutionary history of bacterial chemoreceptors revealed by global metagenomic survey

## 2.1 GlobalCR sequence retrieval

Sequences containing the MCPsignal domain (Pfam:PF00015) were retrieved using HMMER version 3.2.1 [213], [214] with an e-value of 0.0001. Then, LBD characterization was carried out using the pipeline described in Sanchis-López *et al.* [215]. The resulting catalog is available at https://github.com/compgenomicslab/GlobalCR.

A total of 1,414,005 CRs sequences were obtained from the SPIRE resource[187], a large multi-habitat environmental genomic dataset covering 99,146 metagenomic samples from 739 studies. This catalogue comprises 35 billion predicted protein sequences and 1.16 million newly constructed metagenome-assembled genomes (MAGs) clustered into 133,402 species-level genomes. We used CheckM2[216] to filter medium and high-quality MAGs, selecting those with $\geq$50% completeness, $\leq$10% contamination and passing the GUNC[217] test.

A total of 229,160 CRs sequences were obtained from the non-redundant representative genomes set of proGenomes3[218], which contains 41,171 representative genomes per species cluster.

## 2.2 Taxonomy annotation and taxonomy distribution of CR genes

Taxonomic classification was obtained from SPIRE resource, where all SPIRE MAGs were mapped using gtdb-tk v2.11 [219] against release r207[165]. Each species cluster in SPIRE had assigned a consensus taxonomy based on majority vote. ProGenomes3 species representatives were also mapped with the gtdb-tk classifier against r207.

The number of CRs was annotated by calculating the average number of CRs within the MAGs for each species cluster, assigning this average to the cluster. For each taxonomic rank, the mean number of CRs was determined by assigning the average CR count from the chemosensitive species clusters to the respective rank.

## 2.3 Heptad class and CheA genes annotation

Sequences containing Chemosensory Pathways hidden Markov models (HMMs) from the AGFam database (https://github.com/biowonks/agfam/), as described by Wuichet and Zhulin[18] and the study by Alexander and Zhulin[38] , were retrieved using HMMER[213], [214] version 3.2.1 with an e-value threshold of 0.0001. For each sequence, the best hit was selected, which allowed the assignment of CRs to their respective heptad classes, and the identification of CheA genes and their subtypes.

From SPIRE, 1,427,478 CRs were assigned to a heptad class, and 1,195,221 CheA genes were identified in the same MAGs as the CRs. Although a total of 3,571,268 CheA genes were found in MAGs, only those co-occurring with CRs were considered. For proGenomes3 sequences, only heptad subtypes were assigned, with 228,969 CRs mapped to specific heptad subtypes.

The prevalence of each gene across all species in the catalogue was calculated, as well as the mean number of genes per species for each subtype. The species-level gene count was determined by averaging the counts across all MAGs within each species cluster. Co-occurrence between genes was assessed by calculating the percentage of species clusters where both models were present in a two-against-two comparison.

## 2.4 MCPsignal tree

MCPsignal tree was reconstructed with 428,179 unique MCPsignal representative sequences after MMseqs2 [205] homology clustering at 99% of protein identity and 99% of coverage ( -c 0.99 –min-seq-id 0.99 –cov-mode 2). FAMSA version 2.2.2[220] was employed for the protein alignment, and FastTree version 2.1.11[212] for reconstructing the phylogenetic tree. Each representative sequence was annotated with the heptad class and taxonomic information for the cluster of MCPsignal domains represented. Events of duplication were detected when two nodes were overlapping in species in more than 10% of the total species on each node. Visualization of the gene tree and the annotations was performed using ETE version 4.0.0-beta2[208] and TreeProfiler[221].

## 2.5 Putative Horizontal Gene Transfer detection

The 1,643,155 CR sequences in globalCR were clustered using MMseqs2 [205] at 99% protein identity and 99% coverage (-c 0.99 –min-seq-id 0.99 –cov-mode 2). This clustering produced

516,960 groups of highly similar CR genes. To identify clusters containing nearly identical sequences from distantly related lineages, taxonomic assignments were inspected using gtdb-tk annotations. Clusters were annotated for potential horizontal gene transfer (HGT) if they contained sequences from at least two distinct taxonomic orders (or higher ranks, such as class, phylum, or domain) and had at least two supporting sequences per lineage, to reduce the possible false events for a bad taxonomic classification. Based on these criteria, 39 clusters were identified as potential HGT events.

## 2.6 Habitat distribution of CRs genes

Only CRs from SPIRE were analyzed on the basis of their habitats. The simplified microontology provided by SPIRE classifies each sample in one unique habitat with the samples classified with more than one label grouped in "multiple" class. The habitat labels are Air, Anthropogenic: Food Production, Anthropogenic: Industrial and Built Environments, Anthropogenic: Waste Handling, Aquatic: Extreme Environments, Aquatic: Freshwater, Aquatic: Saltwater, Host-Associated: Animal intestine/rumen, Host-Associated: Animal others, Host-Associated: Animal skin, Host-Associated: Human intestine, Host-Associated: Human mouth, Host-Associated: Human others, Host-Associated: Human skin,Host-Associated: Plant, Terrestrial: Agricultural and Wet Environments, Terrestrial: Extreme Environments and Terrestrial: Soil and Dusty Environments.

The ratio of CRs genes per giga base pairs sequenced was calculated only taking into account samples with more than 2 giga base pairs sequenced. The mean number of CRs per species was calculated for each habitat taking into account the species clusters within each sample.

## 2.7 Datasets of LBDs and MCPsignal subfamilies

A total of 1,283,529 sensing domains were annotated and extracted using the methodology of Chapter 1 (Sanchis-López *et al.* [215]). Intracellular sensing domains were annotated as those with no transmembrane domain predicted within the gene, as described in Collins *et al.* [40]. From all the LBDs, 307,228 were intracellular.

*De novo* homology-based clustering of the 1,283,529 LBD sequences was inferred using MMseqs2 [205] with an e-value threshold of 0.001, 30% minimum identity, and 50% minimum query coverage. These parameters were chosen to maximize remote homology detection and to infer LBD clusters with broad phylogenetic divergence (i.e. distant homologues) while still grouping sequences with a

common evolutionary origin. The MMseqs2 command used was “mmseqs cluster -c 0.3 –min-seq-id 0.5 –cov-mode 2”. A total of 42,375 clusters or subfamilies of LBDs were delineated, with 17,086 of them represented by singletons.

For building the datasets of extra and intracellular subfamilies, the subfamilies were classified as intracellular if more than half of the sequences within the cluster were annotated as intracellular in our catalog. A total of 1,797 subfamilies were intracellular, and 40,580 extracellular. The same *de novo* homology-based clustering was applied to the MCPsignal domain of 1,643,155 CRs, and resulted in 3,393 subfamilies of this domain.

## 2.8 Cumulative curves of domain subfamilies

To plot cumulative curves showing the growth in the number of subfamilies as more sequences are analyzed, sequences were randomly shuffled. The curve was generated by plotting the number of detected clusters as the number of analyzed sequences increased in random order. For each sensing domain subtype among the top 10 most abundant, clusters were assigned to a domain if at least half of the sequences within the cluster belonged to that domain. A total of 4,240 out of 42,375 clusters had mixed ligand-binding domains (LBDs), and 627 clusters had no clear majority domain. Among the detected subfamilies, 22,840 belonged to UNKNOWN LBDs, 6,134 to 4HB_MCP_1, 5,860 to dCache_1, 589 to TarH, 411 to sCache_2, 281 to PilZ, 126 to PAS_3, 82 to CZB, and 31 to PocR.

## 2.9 Relative abundance matrix of subfamilies

To determine the relative abundance of sequence clusters across multiple datasets, we first mapped each sample to itself to calculate the depth of coverage for each contig. Depth files were generated for each sample, and the total average depth for each contig was calculated. For each contig, the relative abundance was computed by dividing its average depth by the sum of the average depths for all contigs within the sample. This relative abundance data was then used to calculate cluster-level abundances.

For each sample, the relative abundances of all sequences belonging to the same cluster were summed to generate a final relative abundance matrix. In this matrix, each row represented a subfamily and each column represented a sample. The values corresponded to the total relative abundance of sequences in each cluster for the respective sample.

This pipeline was applied to the sequence clustering of each dataset: extracellular sensing domain, intracellular sensing domain and MCPsignal domain.

## 2.10 Species relative abundance matrix

The matrix was built using mOTUs v3 [222] relative abundances of each sample. The samples of SPIRE are in rows, and each mOTUs detected is in a column.

## 2.11 Random Forest Classification of Biome Prediction

Random Forest classifiers were applied to predict the presence or absence of specific biomes using relative abundance matrices from different datasets (extracellular sensing domains, intracellular sensing domains, MCPsignal domains and species-level mOTUs classification) as input. For each biome, a binary classification model was constructed, where the positive class indicated the presence of the biome, and the negative class indicated its absence. To mitigate class imbalances, up to 300 samples per biome when possible were randomly subsampled. Biomes with less than 100 samples were discarded. The dataset was split into training (70%) and testing (30%) sets, and 10 iterations were performed to ensure statistical robustness.

A Random Forest model with 100 decision trees was trained on each subsampled dataset. The performance of the models was evaluated using accuracy, precision, recall, and F1-score for each biome. For each metric, the mean and standard deviation were calculated across the 10 iterations, providing a summary of the model's performance.

## 2.12 UMAP and HDBSCAN

The relative abundance matrix was first log-transformed by applying a natural logarithm to non-zero values using the formula $\log x + 1$. The log-transformed matrix was then normalized using MinMaxScaler from scikit-learn[223] python package, scaling the values between 0 and 1. After normalization, the matrix was transposed for further analysis. UMAP dimensionality reduction was then applied using Bray-Curtis dissimilarity as the distance metric. The UMAP was performed across a range of parameters, with 'n_neighbors' set between 20 and 100 in steps of 10, and 'min_dist' values of 0, 0.01, 0.05, and 0.1. All-zero rows and columns were removed. The resulting 2D embeddings were saved as CSV files for each combination of parameters.

HDBSCAN clustering was then applied to the UMAP embeddings to identify clusters of subfamilies. Three values of ‘min_cluster_size’ (100, 200, 500) and ‘min_samples’ (10, 50, 100) were tested. For each combination, the clustering labels were saved to a CSV file, and the clustering quality was evaluated using the Davies-Bouldin score. The best-performing model, determined by the lowest Davies-Bouldin score, was selected as the optimal configuration.

## 2.13 Hierarchical clustering of HDBSCAN clusters

The HDBSCAN clusters were further analyzed using hierarchical clustering based on Bray-Curtis dissimilarity. First, the relative abundance matrix was transposed to align the samples with their corresponding HDBSCAN cluster labels. The cluster labels were then merged with the abundance matrix, and the average relative abundance for each cluster was calculated by averaging the values within each cluster. To prepare the data for hierarchical clustering, a log transformation (log(x+1)) was applied to non-zero values in the abundance matrix, followed by normalization using MinMaxScaler to rescale values between 0 and 1. The Bray-Curtis dissimilarity metric was then used to calculate the compositional differences between clusters. The resulting dissimilarity matrix was used for hierarchical clustering to assess the relationships and structure among the HDBSCAN clusters.

Hierarchical clustering was subsequently performed on the Bray-Curtis dissimilarity matrix using the average linkage method. Biome metadata were mapped from a microontology file, assigning each sample to a corresponding biome. HDBSCAN clusters were then associated with biomes, and the relative abundance of each biome within each cluster was calculated as the proportion of samples belonging to each biome. The relative abundance matrix was grouped by HDBSCAN clusters, and the biome distributions within each cluster were calculated and normalized for visualization.

A dendrogram was generated based on the hierarchical clustering results, and stacked bar plots were created to represent the biome distribution within each cluster. The number of samples per cluster was visualized alongside the biome abundance using a consistent color scheme for biomes. The final figure combined the hierarchical clustering with the biome abundance and sample counts, showing the relationships between clusters and their associated biomes.

## 2.14 Prevalence and specificity of subfamilies per biome

For each subfamily of sensing domains, the percentage of specificity was calculated counting the number of samples from each biome over the total number of samples in that subfamily. The prevalence across the biome of the sensing subfamily was assigned with the percentage of samples in that sensing subfamily over the total number of samples for that biome in the CRs catalogue.

## 2.15 Subfamilies trees

For each subfamily of sensing domains with at least three sequences, the alignment of the protein sequences of the domain was performed using FAMSA version 2.2.2[220], and the phylogenetic tree was reconstructed using FastTree version 2.1.11[212]. A total of 15,894 phylogenetic trees of sensing subfamilies are included in globalCR catalog.

# 3 Unveiling the Repertoire of Microbial Chemoreceptor Genes in the Rare Biosphere

## 3.1 MetaCR sequence retrieving

A total of 149,326 CRs sequences were obtained from the GMGCv1[186] resource (http://gmgc.embl.de), a novel metagenomic catalog combining 84,029 genomes from ProGenomes2[224] and 302,655,267 unigenes from 13,174 public metagenomic samples from 14 different environments. Sequences containing the MCPsignal domain (Pfam:PF00015) were retrieved using HMMER version 3.2.1 [213], [214] with an e-value of 0.0001. Then, LBD characterization was carried out using the pipeline described in Sanchis-López *et al.* [215]. MetaCR catalogue and metadata information are available at https://github.com/compgenomicslab/metaCR.

## 3.2 metaCR ecological annotation

Ecological annotation of the CRs open reading frames (ORFs) was based on the GMGCv1 annotation of the samples where the ORFs were assembled. As for the origin of the samples, we maintained the 14-environment encompassing human-associated microbial ecosystems (gut, nose, vagina, oral, skin), animal-associated (cat gut, mouse gut, dog gut, pig gut), and environmental (soil, built-environment, marine, wastewater, and freshwater). The number of CRs per habitat was calculated aggregating the information of the samples within each environment.

## 3.3 Probe set design

MCPsignal domains of the 149,326 collected sequences were fragmented into 80-mers (i.e., k-mers of length 80), hereafter probes, at 20 nt intervals (i.e., 60 nts of overlap) and clustered using CD-HIT version 4.8.1[225] (with 80% identity and 75% coverage). From 3,249,805 initial k-mers, we obtained 1,354,437 clusters from which we took one representative k-mer as a probe (i.e., 21.4 probes per MCPsignal domain). This set of probes was mapped (using BLASTn from BLAST+ version 2.9.0[226]; min % identity 80%, min % query coverage 90%) to the initial MCPsignal sequences. Further probes were retrieved (same procedure and parameters as above) from regions without any probes mapping to them. To remove promiscuous and nonspecific 80-mers, we mapped our

probes to Progenomes2[224] and to a custom database of organellar genomes using once again BLASTn. Probes that either hit off-target sequences (promiscuous) or a broad range of targets (nonspecific) were discarded. The final set of probes (1,327,258 80-mers) was synthesized as a Roche NimbleGen EZ Developer DNA probe set (Roche Sequencing Solutions, Madison, WI, USA).

## 3.4 Sample collection and DNA extraction

Bulk soil, rhizosphere, and phyllosphere samples were collected from two different crops: tomato (Solanum lycopersicum L.; greenhouses of HM. Clause Iberica, Almería, Spain) and potato (Solanum tuberosum L.; crops from Udapa S.C., Vitoria-Gasteiz, Spain). Bulk soil was sampled in the fraction adjacent to the plant, but not adhered to the roots, and stored at -80ºC. Rhizospheric soil was collected from the root, after hand-shaking elimination of the non-adhering soil fraction, by immersion and washing in a solution of PBS and Trition X-100 (0.1%) for 10 min. Then, rhizospheric soil was sieved and collected by centrifugation at 4000 x g for 10 min at 4ºC. Supernatant was discarded and the rhizospheric soil pellet was stored at -80ºC. For the phyllosphere fraction, leaves were collected from the plants, placed in ziplock bags and stored at 4ºC for 24 h prior to DNA extraction. DNA from Bulk and Rhizospheric soil was extracted using Power Soil DNA Isolation kit (MO BIO Laboratories, Inc., Carlsbad, CA, United States) following manufacturer's instructions. DNA from phyllosphere samples was extracted following the protocol described by Suda *et al.*[227].

For the lake sample, a volume of 125 mL of water was collected (Casa de Campo Lake, Madrid, Spain) and filtered through a 0.22 µm filter. Then, the filter was stored at -20ºC. For DNA extraction, the filter was cutted into pieces using sterilized scissors and processed using commercial Power Soil DNA Isolation kit (MO BIO Laboratories, Inc., Carlsbad, CA, United States) following manufacturer's instructions.

The rumen sample was collected from a Holstein cow in her first lactation using a custom-built mechanical device to elevate the cow's snout. Then, approximately 100 ml of ruminal content was extracted by inserting a tube connected to a mechanical pump (Vacubrand ME 2SI, Wertheim, Germany) via the esophagus. The sample was filtered using four layers of sterile cheesecloth to separate the solid elements, and the liquid portion was immediately frozen using liquid nitrogen vapors. These frozen samples were then conveyed to the lab in containers filled with liquid nitrogen and kept at -80 ℃ until analysis. The practice was performed in accordance with EU Directive 2010/63/EU for animal experiments and approved by the ethics committee (NEIKER-OEBA-2017-004).

For DNA extraction, the sample was thawed and homogenized using a blender and microbial DNA was extracted using the commercial kit “DNeasy PowerSoil” (Qiagen, Valencia, CA, USA).

A human fecal sample was collected from an adult asymptomatic individual and ca. 5 g were stored within 5 min at -80℃. The sample was processed by grinding it to powder with a sterile mortar and pestle under liquid nitrogen. DNA from an aliquot of ca. 1 g was extracted using the AllPrep DNA/RNA Kit (Qiagen) with the following modification: the powder was added to a 2 mL tube pre-filled with 100 µm Zirconium beads (OPS Diagnostics LLC) and 0.6 mL of RLT buffer complemented with 10 µL beta-mercaptoethanol (Sigma-Aldrich) and disrupted twice using the mixer mill Retsch MM400 at 30 Hz for 3 min with 5 min incubations in-between. Cell debris and beads were pelleted by centrifugation at full speed before transferring the supernatant to a DNA-binding column. The remaining steps followed the manufacturer’s instructions. The DNA was sent to Novogene for library preparation and shotgun metagenomic sequencing.

3.5 Samples pooling Pool-1: This pool included DNA samples from the human gut, rumen, lake, phyllosphere of potato (phyllpot), phyllosphere of tomato (phylltom), three biological replicates of the rhizosphere of potato (rhizpot1, rhizpot2, rhizpot3), three biological replicates of the soil of potato (soilpot1, soilpot2, soilpot3), and three biological replicates of the soil of tomato (soiltom1, soiltom2, soiltom3).

Pool-2: This pool included DNA from three biological replicates of the potato rhizosphere (rhizpot1, rhizpot2, rhizpot3) and three biological replicates of the tomato rhizosphere (rhiztom1, rhiztom2, rhiztom3). All samples in Pool-2 were different from those in Pool-1.

## 3.5 DNA Fragmentation

DNA fragmentation in Pool-1 was performed using enzymatic digestion NEBNext dsDNA Fragmentase (M0384), resulting in DNA sizes 1.5 Kb. Fragmentation in Pool-2 was done using Covaris M220 ultrasonicator shearing protocol for 1.5 Kb.

## 3.6 Target capture experiments

The target capture methodology described below was inspired by the work of Bethune *et al.* 2019[228], modified as necessary, and applied to both the Pool-1 and Pool-2 experiments.

*Quality Control and DNA Purification*: Fluorometric DNA quantitation was done using a Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific, Waltham, Massachusetts, USA). DNA fragment size distribution and molarity were profiled using the Agilent 4150 TapeStation System (Agilent Technologies, Santa Clara, CA, USA). DNA clean-up and size selection was performed using Agencourt AMPure XP beads (Beckman Coulter).

*Library Preparation*: DNA libraries were prepared according to the NEBNext Ultra II DNA Library Prep Kit for Illumina (E7645, E7103) manual with in-house modifications to obtain longer DNA sizes. The starting amount of fragmented DNA used per sample was 300 ng. All the steps of the manufacturer's protocol were performed at half the recommended volume to reduce costs[229]. The size selection step was modified to retain the longer sequences, with a single step adding 0.5x AMPure XP beads (Beckman Coulter, Brea, CA, USA) to discard shorter fragments. Samples were indexed using NEBNext Multiplex Oligos for Illumina (Dual Index Primers Set 1, E7600). Thermal cycling parameters for enrichment of Adaptor-ligated DNA were 6 cycles of 130 s extension each. Clean-up after PCR was performed using AMPure XP beads (0.65x). Finally, all samples of each pool were mixed at equimolar concentration in Pool-1 and Pool-2.

*Hybridization and enrichment*: From the equimolar pools, 1µg of DNA was taken for capture experiments. Target enrichment experiments were performed according to the Roche NimbleGen Inc. protocol "SeqCap EZ HyperCap Workflow User's Guide v2.3" with in-house modifications. DNA library pools from the previous step were hybridized with the SeqCap EZ probe set for 24h at 49.5ºC. After washing and recovery of the captured multiplexed DNA sample, amplification of the DNA was done using the Kapa HiFi HotStart ReadyMix 22 cycles and an extension time of 3 min.

*Long-read sequencing of Pool-1*: The captured and enriched DNA was sequenced in a GridION using a R.9.4.1 flow cell (Oxford Nanopore Technologies, ONT, Oxford, UK). The ligation sequencing kit used was SQK-LSK110, with end prep and adapter ligation performed according to the manufacturer's instructions. Base-calling was performed using the ONT software Guppy version 4.2.2 in high accuracy mode with quality score (QS) filter >9 and read length >50 bp. A total of 5.56 Gb passed (7,019,121 reads), out of 6.91 Gb sequenced. N50 of 875 pb.

*Long-read sequencing of Pool-2*: The captured and enriched DNA was sequenced in a MinION using an ONT R.10.4.1 flow cell. The ligation sequencing kit used was SQK-LSK114, with end prep and adapter ligation performed according to the manufacturer's instructions on the multiplexed NEBNext Ultra II genomic libraries (with Illumina adapters). High accuracy base-calling was performed using the ONT software Guppy version 4.2.2 with filter QS >10 and read length >200 bp. A total of 14.93 Gb passed (10,738,522 reads) out of 17.28 Gb sequenced. N50 of 1.65 kb.

## 3.7 Bulk shotgun metagenomics sequencing

DNA samples of Pool-1 and Pool-2 (except human gut) were sequenced with Illumina using Nextera DNA XT libraries and sequencing with NovaSeq 6000 PE150 (2 x 150bp) 8Gb per sample. For Pool-1, although multiple biological replicates were available, only one sample was selected for sequencing, no biological replicates were sequenced.

## 3.8 Read processing: trimming, demultiplexing, assembly, and deduplication

*Assembly of long-read sequences for Pool-1 and Pool-2:* Nanopore basecalled fastq reads were trimmed with porechop version 0.2.4. Demultiplexing was done with custom scripts, searching for the NEBNext Multiplex Oligos unique sequences (which include Illumina adaptors). A round of BlASTn from BLAST+ version 2.9.0[226] with the whole index sequence was performed, then the 8 unique nts for each index were identified by text similarity, giving a score to each index per sequence and finally classified based on the best scores and index combination of the samples. In Pool-1, 3,352,198 reads could be demultiplexed, out of the 6,938,358 trimmed. For Pool-2, 9,053,747 reads were correctly demultiplexed, out of 10,738,705 trimmed reads. To remove duplicate reads, representative clusters at 99% of identity were identified using CD-HIT version 4.8.1 ( -c 0.99 -aL 0.95 -aS 1).

*Bulk shotgun metagenomics sequences:* Obtained Illumina reads were inspected with FastQC version 0.11.9[230] and MultiQC version 1.10.1[231]. Reads were trimmed using Trimmomatic version 0.38[232] (-phred33 LLUMINACLIP:$EBROOTTRIMMOMATIC/adapters/NexteraPE-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:50). The trimmed Illumina reads were subjected to de novo assembly using MEGAHIT version 1.2.8[232], [233] (–presets meta-large). To remove duplicate reads and to reduce redundancy in the dataset, the assembly output was processed using BBMap version 38.87[234] dedup.sh (minidentity=99).

*Host-plant contamination* For measuring the plant host contamination in the samples sequenced with shotgun metagenomics, the reads were aligned to the reference genomes of tomato (*Solanum lycopersicum* L, GCF_000188115.4) and potato (*Solanum tuberosum* L, GCF_000226075.1) using minimap2 version 2.24-r1122[235] and SAMtools version 1.12[236].

## 3.9 Detection of on-target reads

In order to mine the potential on-target reads/contigs (capture/shotgun metagenomics experiments), first we performed a search against a few MCPsignal representative domains extracted from PfamA-32-A (Supp. materials), using eggNOG-mapper version 5[237] (-m diamond –itype metagenome). The whole read/contig containing ORFs predicted as potentially on-target were translated into the six possible frames with transeq version 6.5.7[238] (-frame 6 -clean). Domains were extracted from protein sequences using HMMER version 3.2.1 (hmmsearch against Pfam-32-A database). Then, 'MCPsignal' domains were extracted and counted from protein reads, and only one reading frame was kept for each read according to the best MCPsignal hit. In addition, for each on-target read or contig, we extracted Pfam domains and constructed a domain architecture profile based on the two domains immediately before and after the 'MCPsignal' if present.

## 3.10 Sequencing methods comparison (Nanopore *vs.* Illumina)

The on-target ratio was normalized by base pair by performing ten random subsampling runs, for the same amount of pb given the depth of sequencing for each experiment (40Mb Pool-1, 374Mb Pool-2). In the shotgun samples, the subsampling was done with seqtk version 1.3[239], while for the captured demultiplexed pools, subsampling was done with a custom python script. After the subsampling, the starting raw reads were processed for detecting the on-target reads following the same pipeline described above.

## 3.11 Mapping to GTDB r207

MCPsignal domains (PF00015) were extracted from the proteome of GTDB r207[165] following the same methodology described in Sanchis-López *et al.* [165], [215] Later on, the domain sequence (in nts) was mapped against the target reads (shotgun or capture analyses) using Blastn from BLAST+ version 2.9.0[226] (-task blastn -evalue 0.0001 -max_target_seqs 10000). The output hits were filtered by sequence identity (100%, $\geq$98%, and $\geq$80%), and by query coverage ($\geq$0%). The assignment of a known CR from GTDB was only made when all the hits were 100% identical to the same and unique CR, in order to avoid ambiguity. In case of having more than one unique CR as significant hits, the assignment of the last common ancestor of the CR was calculated over all the taxonomies of significant hits.

## 3.12 Phylogenetic distribution of the known CRs detected

The GTDBr207 tree of bacteria and archaea was concatenated and annotated using ETE Toolkit version 4.0.0-beta2[208] and TreeProfiler[221], and subsequently collapsed to the order level. The sum of the different known CRs detected, by order for each analysis, was calculated and placed in the tree. iTOL[240] was used for tree visualization.

## 3.13 Taxonomic profiling of metagenomic samples

mOTUS version 3.0.1[222] was run using -g3 (at least three marker genes) as standard parameter for profiling the taxonomy present in the samples sequenced with bulk metagenomics methodology.

## 3.14 Phylogenetic delineation of novel CRs

The MCPsignal tree was built using protein sequences of the domain extracted from GTDB r207 (305,906 domains), and the MCPsignal domains from shotgun and target capture experiments. A total of 616,738 protein sequences were aligned using FAMSA version 2.2.2[220], and the phylogenetic tree was reconstructed using FastTree version 2.1.11[212]. Each CR sequence from our experiments was annotated with the percentage of identity to the best hit in GTDB r207, as well as the taxonomic information of that best-hit. Visualization of the gene tree and the annotations was performed using ETE version 4.0.0-beta2[208] and TreeProfiler[221] .

For identifying the nodes containing novel sequences of CRs, the average protein identity of the captured sequences per node was annotated. When the node contains a minimum number of sequences from target capture experiments, the support is greater than 0.8, and the mean best hit protein identity is within the ranges settled, the node is counted.

# Results

# Chapter 1

## Prevalence and Specificity of Chemoreceptors Profiles in Plant-Associated Bacteria

The main results of this chapter are found in the paper “Prevalence and Specificity of Cemoreceptors Profiles in Plant-Associated Bacteria” (Annexe II)[215]. Traditionally, much of our understanding of bacterial chemoreception has focused on a few model organisms, with limited insight into the diversity and ecological roles of chemoreceptors (CRs) in different environments. Despite their importance in processes such as colonization and infection, particularly in plant-associated bacteria, significant gaps remain in understanding the functional diversity and evolutionary dynamics of CRs. How are these chemoreceptors distributed across various bacterial lineages? What roles do they play in specific ecological contexts, such as plant-microbiome interactions? How do ligand-binding domains (LBDs) diversify to meet the demands of different lifestyles? In this chapter, we aim to address these questions by developing a robust methodology for characterizing and classifying CR sensing domains, and by exploring the prevalence and distribution of CRs in cultured prokaryotic species. These efforts will help us better understand the adaptive evolution of CRs and their contribution to niche-specific bacterial behaviors, with a particular focus on plant-associated environments.

## 1.1 Towards a global catalog of chemoreceptors in plant-associated bacteria

In order to maximize the coverage of our analysis, we first built a comprehensive catalogue of CRs detected across the entire prokaryotic phylogeny (Figure 1.1). Species genomes were retrieved from the proGenomes v2 Databases[224]. Unlike the NCBI Taxonomy database, which is not an authoritative source for nomenclature or classification[241], proGenomes2 data do not rely on taxonomic names to identify species. Instead, each species-representative genome in proGenomes is delineated based on the evolutionary distances calculated between universally conserved genes present in nearly all organisms[199], [241]. To establish links between CRs and the plant-associated lifestyle, we compiled three manually curated databases of PAB (see Materials and Methods): i) PAB-broad, a reference database of 960 organisms found in multiple plant environments including leaves, roots and rhizospheric soil; ii) PAB-phyto, a subset database of 119 species including only known phytopathogens, and iii) PAB-symb, which groups 192 plant symbionts. Using HMM-based searches, we then mined all the sequences containing the MCPsignal domain in the 11,806 species-representative genomes from the proGenomes database, compiling a global catalogue of 82,277 CR sequences from 5,546 genomes (Table S1.3). This confirms the broad distribution of CRs, with 47% of the representative genomes containing at least one chemotactic receptor.

PAB species possessed almost twice as many CRs per genome (22.86) than those species not classified as plant-associated (12.94), with the subset of phytopathogens showing the highest number (27.29). No CRs were predicted in 178 out of the 960 PAB genomes, indicating that more than 81% of PABs possess at least one CR gene, a percentage largely superior to the bacterial average. From all species considered in this study, 36 PAB genomes stood out by their high content of CRs (Table 1.1), most notably: i) 14 genomes from the *Pseudomonas* genus (49-60 CRs), including the well-known plant pathogens *P. syringae* and *P. savastanoi*[242], and ii) 9 genomes from the *Herbaspirillum* genus (52-67 CRs), a group of of beta-proteobacteria that endophytically colonize gramineous species thereby promoting their growth[243].

## 1.2 Classifying chemoreceptors according to their Ligand Binding Domains

As the ecological relevance of CRs is mostly defined by their LBD region, we explored whether sequence segments corresponding to the LBD, rather than the full-length CR sequences, were related to a plant-associated lifestyle. To maximise the number of LBD sequences included in our analysis and not limit this to known LBD types from the Pfam database[50], we inferred LBDs based on the domain architecture of each CR. First, we extracted LBD sequences from the whole set of 82,277 CRs. Next, and given the high variability in the domains that could be considered as LBDs, we identified putative LBDs using three different strategies: i) detecting sequence regions matching any known domain other than the MCPsignal or HAMP; ii) locating sequence regions flanked by two TM regions; and iii) taking domains between the N-terminal and a single TM region. In total, we retrieved 72,480 putative LBD sequences, which could be fit into three main groups based on their length (Figure 1.2). The first group includes LBDs with a size between 60 and 110 amino acids, containing 21% of all the LBDs detected. The most abundant LBD family within this size range was PAS_3. The second group, comprising LBDs from 130 to 200 amino acids, contained over 45% of all LBDs and included 4HB_MCP_1 as the predominant family. The third group, comprising LBD lengths between 220 and 299 amino acids, covers 26% of all LBDs and has dCache_1 as the most abundant LBD family. Only 8% of all the LBDs detected fell outside these three size ranges, and the three most abundant LBDs were 4HB_MCP_1 (17.6%), dCache_1 (15.5%) and PAS_3 (9.2%).

We next investigated whether LBDs could be classified into broader sequence homology clusters, each representing a group of LBD sequences sharing a common evolutionary origin. Using relaxed homology thresholds (e-value $\leq$10-3, 50% coverage, 20% amino acid identity), we grouped all 72,480 LBD sequences into 5,149 family clusters (Table S1.4), of which 3,307 contain more than

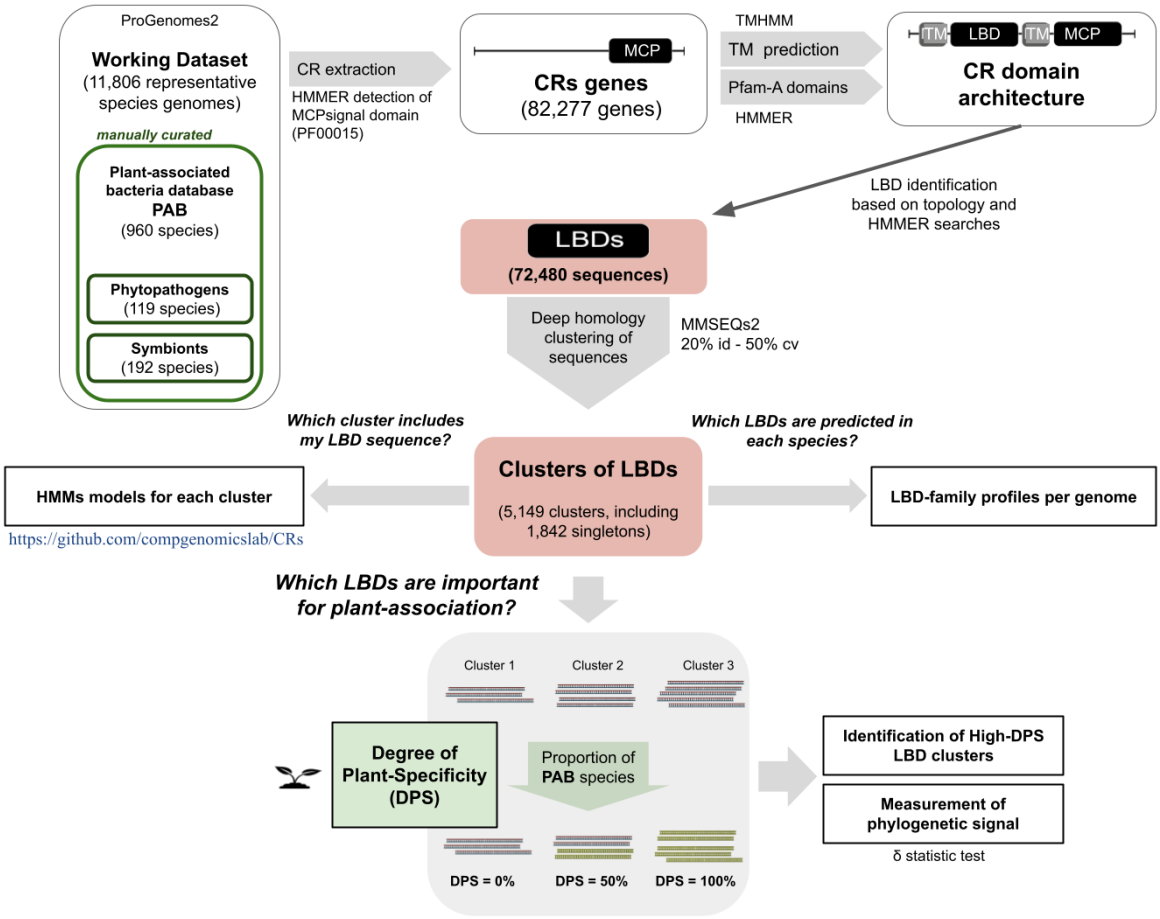



**Figure 1.1: Schematic view of the bioinformatics pipeline used to identify CRs that are potentially relevant for plant association.** From a set of 11,806 representative prokaryotic genomes, 82,277 protein sequences were mined using HMM-based searches against the MCPsignal Pfam domain (PF00015). CR topology was analysed by predicting transmembrane regions (TMs) and Pfam domains. Based on the topological analysis, LBD regions were predicted and a set of 72,480 LBD sequences was obtained. Clustering of LBDs based on sequence homology (20% minimum sequence identity with at least 50% sequence coverage) resulted in 5,149 clusters or subfamilies of LBDs, of which 1,842 contained a single sequence. To study a possible link between the LBD profiles and plant-associated lifestyle, a manually-curated subset of 960 representative species of Plant-Associated Bacteria (PAB) was generated, including Phytopathogens (119) and Symbionts (191) subsets. The determination of the proportion of PAB LBDs present in each cluster allowed to assign the Degree of Plant Specificity (DPS) value for each LBD subfamily. Subsequent analysis of high-DPS clusters identifies LBD clusters that are potentially important for bacterial plant association. Furthermore, the validation of the high-DPS clusters as good ecological indicators was corroborated by measuring their phylogenetic signal. A detailed step-by-step description of the process can be found in the Materials and Methods section.

**Table 1.1: List of PAB with the highest number of predicted CRs.** P, Phytopathogen. S, Plant Symbiont.

| TaxId | Biosample | Representative species | Number of CRs | PAB-phyto | PAB-symb |
|---|---|---|---|---|---|
| 1078773 | SAMN04334956 | *Herbaspirillum rubrisubalbicans* M1 | 67 | | |
| 1144319 | SAMN00839627 | *Herbaspirillum* sp. CF444 | 67 | | |
| 964 | SAMN03779333 | *Herbaspirillum seropedicae* | 65 | | |
| 193 | SAMN02982994 | *Azospirillum lipoferum* | 65 | | |
| 286727 | SAMN02982917 | *Azospirillum oryzae* | 64 | | |
| 346179 | SAMN03785417 | *Herbaspirillum rhizosphaerae* | 62 | | |
| 864073 | SAMN02471292 | *Herbaspirillum frisingense* GSF30 | 62 | | |
| 237610 | SAMN05860868 | *Pseudomonas psychrotolerans* | 60 | | |
| 288000 | SAMN02598359 | *Bradyrhizobium* sp. BTAi1 | 60 | | S |
| 92645 | SAMN06130964 | *Herbaspirillum frisingense* | 59 | | |
| 1175306 | SAMN02469572 | *Herbaspirillum* sp. GW103 | 59 | | |
| 1121033 | SAMN02440867 | *Azospirillum halopraeferens* DSM 3675 | 58 | | |
| 169679 | SAMN05170519 | *Clostridium saccharobutylicum* | 58 | | |
| 29438 | SAMN03837775 | *Pseudomonas savastanoi* | 57 | P | |
| 1262470 | SAMN03010392 | *Herbaspirillum hiltneri* N3 | 55 | | |
| 582667 | SAMN05192568 | *Methylobacterium pseudosasicola* | 54 | | |
| 50340 | SAMN05216581 | *Pseudomonas fuscovaginae* | 54 | | |
| 1001585 | SAMN02603190 | *Pseudomonas mendocina* NK-01 | 54 | | |
| 1749078 | SAMN04216969 | *Pseudomonas* sp. EpS/L25 | 53 | | |
| 1190415 | SAMN05216593 | *Pseudomonas asturiensis* | 53 | | |
| 50340 | SAMN03100370 | *Pseudomonas fuscovaginae* | 53 | | |
| 129140 | SAMN03976254 | *Pseudomonas syringae* pv. tagetis | 52 | P | |
| 294 | SAMN04992557 | *Pseudomonas fluorescens* | 52 | | |
| 1855289 | SAMN05216319 | *Duganella* sp. CF402 | 52 | | |
| 1144342 | SAMN00839653 | *Herbaspirillum* sp. YR522 | 52 | | |
| 47885 | SAMN03365871 | *Pseudomonas oryzihabitans* | 51 | | |
| 205918 | SAMN02604347 | *Pseudomonas syringae* pv. syringae B728a | 51 | P | |
| 1907416 | SAMN05880558 | *Aeromonas* sp. RU39B | 51 | | |
| 693986 | SAMN03075686 | *Methylobacterium oryzae* CBMB20 | 50 | | |
| 1736267 | SAMN04151647 | *Pseudomonas* sp. Leaf127 | 50 | | |
| 114615 | SAMEA3138227 | *Bradyrhizobium* sp. ORS 278 | 50 | | S |
| 1028989 | SAMD00019511 | *Pseudomonas* sp. StFLB209 | 50 | | |
| 80867 | SAMN04009978 | *Acidovorax avenae* | 50 | P | |
| 1122963 | SAMN02440654 | *Pleomorphomonas oryzae* DSM 16300 | 50 | | |
| 223283 | SAMN02604017 | *Pseudomonas syringae* pv. tomato str. DC3000 | 49 | P | |
| 1245469 | SAMD00061052 | *Bradyrhizobium oligotrophicum* S58 | 49 | | S |

1 sequence. This de novo clustering approach might not be adequate for a detailed functional characterization of LBDs, as single residue changes have been shown to modify LBD ligand affinities[79], [112], [244]. Nevertheless, each of our LBD clusters could be interpreted as an independent LBD type, with implicit levels of functional and ecological conservation. In fact, our approach consistently recovered all known LBD types and distributed them into 2,068 compact clusters where 90% of their members belonged to the same Pfam domain family (Table S1.5). Moreover, our clustering strategy allowed us to split large LBD families into finely grained sub-categories (Figure 1.3). For example, despite 4HB_MCP and Cache-like being present at similar levels in the initial CR sequence database, the number of derived clusters differ significantly, namely 20.9% for 4HB_MCP as compared to 8.3% for Cache-like. In the case of 4HB_MCP_1, the 10,034 sequences group into 856 different clusters compared to the 283 clusters for the 9,162 dCache_1 sequences, indicating higher sequence conservation in the latter. The situation is even more drastic in the case of PAS_3 LBDs, where 2,675 sequences group into just 21 clusters (Table S1.5), indicating a very low degree of diversity.

Notably, an important fraction (45%) of the LBD clusters inferred could not be confidently associated with any previous family of Pfam domains, since more than 90% of their LBD sequences did not match to any known domain signature, suggesting the existence of a large number of unknown LBD types.

## 1.3 Identifying PAB-specific Ligand Binding Domains

To identify LBD families specific to a plant-associated lifestyle, we analysed each LBD cluster and calculated the corresponding percentage of PAB species there-in, which we referred to as the Degree of Plant Specificity (DPS, see Materials and Methods). For each LBD cluster, we calculated three DPS values, based on three databases of PAB species: i) DPS-broad, calculated based on the PAB-broad reference database; ii) DPS-phyto, based on the PAB-phyto subset; and iii) DPS-symb, using the PAB-symb subgroup as a reference. In all cases, the DPS values ranged from 0% (no LBD family observed in the corresponding PAB database) to 100% (the LBD cluster only includes species from a given reference PAB database). From the 3,307 LBD non-singleton clusters, we identified 419 and 139 clusters with a DPS-broad score $\geq$50%, and $\geq$80%, respectively. Similarly, many LBD clusters showed high specificity in the stricter PAB reference databases (Table S1.6).

To further validate our findings we crosslinked our predictions with experimental data from previous studies[245]–[249]. In particular, we found that CRs with increased expression in planta, and

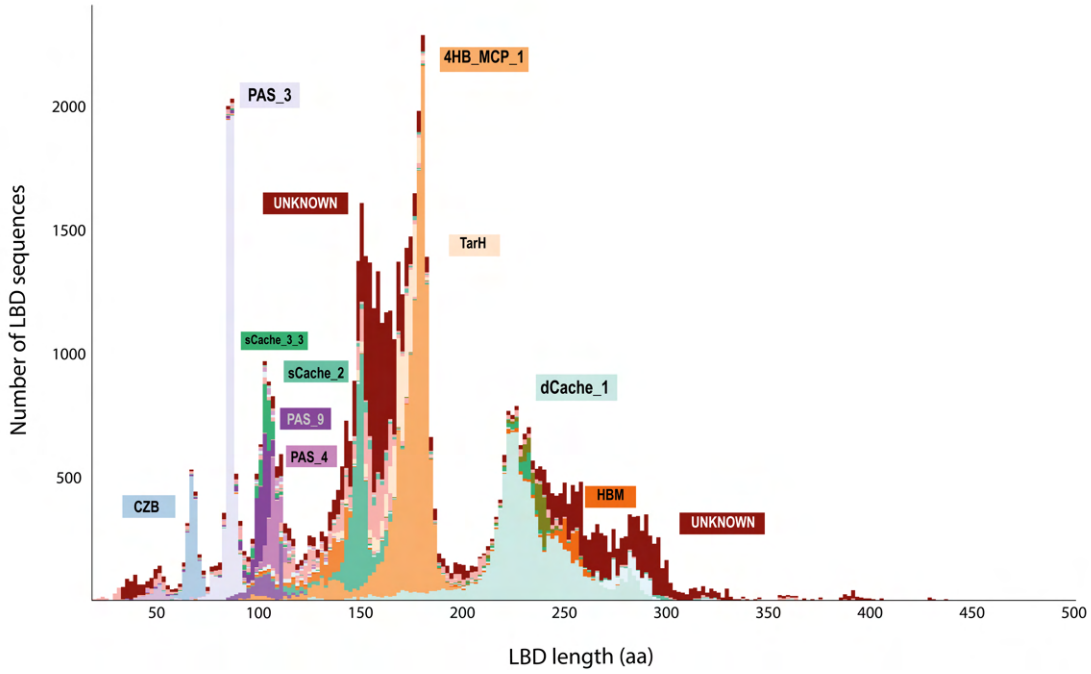



**Figure 1.2: Length distribution of the LBDs.** The analysis was conducted on 72,480 LBDs and the predominant LBD types within each of the main peaks are indicated. Only LBDs shorter than 500 amino acids are represented.

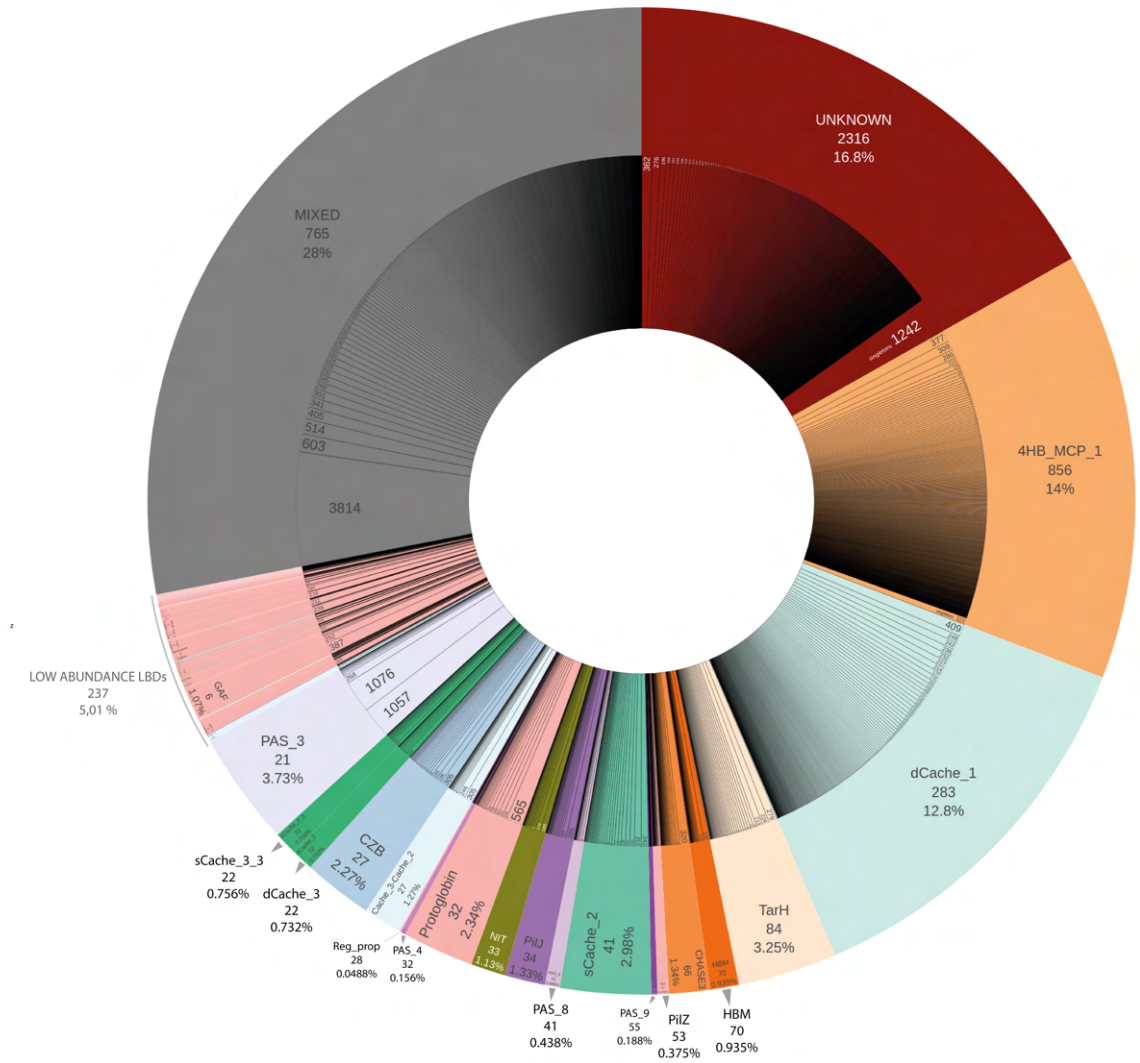



**Figure 1.3: Visual representation of the abundance of the LBD families.** The outer donut of the chart represents the distribution of each LBD type and its relative abundance (in percentage of sequences), and the number of clusters with at least 90% of their sequences sharing the same LBD type, as defined by the Pfam signature. The LBDs are sorted according to the number of clusters within each LBD type. The inner donut of the chart represents all the clusters included within each LBD category, indicating the number of sequences contained in each subfamily. All singletons are merged in the last section of each LBD type (e.g., LBDs classified as 'UNKNOWN' have 1,242 singletons, that is, clusters containing only one sequence). 'Mixed clusters' are those that do not reach the 90% threshold of sequences with the same Pfam model per cluster. 'Low abundance LBDs' include those LBD types that group into less than 12 'compact clusters'.

particularly those required for full bacterial virulence, belonged to high DPS clusters (Table 1.2). This list includes CRs that are up-regulated in: *Dickeya dadantii* 3937 and *Pectobacterium carotovorum* WPP14, two soft rot bacterial strains[245]; *Dickeya dianthicola* RNS04.9, that grows on macerated potato tubers[246]; and *Xanthomonas fragariae*, that grows on strawberry leaves[247]. Similarly, we found several CRs with very high DPS values (80%) that were shown to be relevant in *Xanthomonas citri* virulence[248], or required for fitness of *Pseudomonas savastanoi* pv. savastanoi in olive knots[249]. Taken together, these data support the validity of our approach to identify CRs that are relevant for a plant-associated lifestyle.

Interestingly, we also found that many PAB-specific clusters (41.75%) are formed by proteins of unknown LBD type, suggesting the presence of a significant number of uncharacterized LBD types. Excluding unknown LBD families, the most common domains among high-DPS clusters are 4HB_MCP_1 (26%), TarH (4.5%) and HBM (4%) (Table 1.3). It is remarkable that the three domain families form four-helix bundle structures (53, 54). The case of the HBM and TarH domains is particularly interesting, as the majority of sequences that belonged to these categories concentrated in very few high-DPS clusters: 57.0% (516/906) of all HBM sequences are grouped into 23 high-DPS clusters, and 36.7% (1,042/2,840) of all TarH sequences are grouped into 26 high-DPS clusters. This indicates a strong association of the TarH and HBM domains with the plant-associated lifestyle. By contrast, despite being the second most abundant LBD domain in bacteria (Table S7), the dCache_1 domain was not very abundant in PAB .

**Table 1.2: CRs predicted to be involved in plant-bacteria interaction.**

| CR gene ID | LBD type | DPS (%) | DPS-phyto (%) | DPS-symb (%) | Cluster no. | AA identity (%) | Bacterial species and strain | Reference |
|---|---|---|---|---|---|---|---|---|
| ABF-0014824 | TarH | 100 | 100 | 0 | 932 | 100 | *Dickeya dadantii* 3937 | [245] |
| ABF-0015168 | TarH | 68.09 | 40.43 | 0 | 179 | 100 | *Dickeya dadantii* 3937 | [245] |
| ABF-0016115 | HBM | 63.64 | 48.48 | 0 | 409 | 100 | *Dickeya dadantii* 3937 | [245] |
| ABF-0016585 | TarH | 51.09 | 14.60 | 0.73 | 42 | 100 | *Dickeya dadantii* 3937 | [245] |
| ABF-0017097 | UNKNOWN | 80.00 | 80.00 | 0 | 838 | 100 | *Dickeya dadantii* 3937 | [245] |
| ABF-0017674 | 4HB_MCP_1 | 100 | 87.50 | 0 | 630 | 100 | *Dickeya dadantii* 3937 | [245] |
| ABF-0019851 | TarH | 61.22 | 42.86 | 0 | 123 | 100 | *Dickeya dadantii* 3937 | [245] |
| ABF-0019855 | TarH | 61.22 | 42.86 | 0 | 123 | 100 | *Dickeya dadantii* 3937 | [245] |
| ABF-0020431 | sCache_2 | 34.55 | 12.73 | 0 | 233 | 100 | *Dickeya dadantii* 3937 | [245] |
| DDI_0843 | dCache_1 | 38.61 | 10.13 | 1.27 | 51 | 100 | *Dickeya dianthicola* RNS04.9 | [246] |
| DDI_0932 | sCache_2 | 34.55 | 12.73 | 0 | 233 | 100 | *Dickeya dianthicola* RNS04.9 | [246] |
| DDI_1647 | TarH | 61.22 | 42.86 | 0 | 123 | 88.97 | *Dickeya dianthicola* RNS04.9 | [246] |
| DDI_1649 | TarH | 61.22 | 42.86 | 0 | 123 | 100 | *Dickeya dianthicola* RNS04.9 | [246] |
| DDI_2258 | HBM | 100 | 100 | 0 | 792 | 100 | *Dickeya dianthicola* RNS04.9 | [246] |
| DDI_4092 | 4HB_MCP_1 | 100 | 87.50 | 0 | 630 | 100 | *Dickeya dianthicola* RNS04.9 | [246] |
| ADT-0000027 | HBM | 63.64 | 48.48 | 0 | 409 | 96.03 | *Pectobacterium carotovorum* WPP14 | [245] |
| ADT-0000661 | sCache_2 | 38.46 | 12.82 | 0 | 160 | 99.30 | *Pectobacterium carotovorum* WPP14 | [245] |
| ADT-0001320 | TarH | 51.09 | 14.60 | 0.73 | 42 | 98.84 | *Pectobacterium carotovorum* WPP14 | [245] |
| ADT-0001602 | TarH | 56.99 | 29.03 | 0 | 116 | 94.15 | *Pectobacterium carotovorum* WPP14 | [245] |
| ADT-0001887 | TarH | 61.22 | 42.86 | 0 | 123 | 97.59 | *Pectobacterium carotovorum* WPP14 | [245] |
| ADT-0002104 | TarH | 100 | 100 | 0 | 932 | 100 | *Pectobacterium carotovorum* WPP14 | [245] |
| ADT-0003152 | TarH | 68.09 | 40.43 | 0 | 179 | 91.15 | *Pectobacterium carotovorum* WPP14 | [245] |
| ADT-0003245 | 4HB_MCP_1 | 100 | 87.50 | 0 | 630 | 97.40 | *Pectobacterium carotovorum* WPP14 | [245] |
| ADT-0003418 | UNKNOWN | 80.00 | 80.00 | 0 | 838 | 95.60 | *Pectobacterium carotovorum* WPP14 | [245] |
| PSA3335_17610 | UNKNOWN | 87.50 | 50.00 | 0 | 835 | 100 | *Pseudomonas savastanoi* NCPPB3335 | [249] |
| XAC1892 | UNKNOWN | 100 | 100 | 0 | 846 | 86.77 | *Xanthomonas citri* subsp. citri XHG3 | [248] |
| XAC2448 | 4HB_MCP_1 | 39.23 | 14.62 | 0 | 77 | 98.88 | *Xanthomonas citri* subsp. citri XHG3 | [248] |
| NBC2815_01024 | 4HB_MCP_1 | 92.00 | 84.00 | 0 | 549 | 100 | *Xanthomonas fragariae* IPO 3485 | [247] |
| NBC2815_02005 | 4HB_MCP_1 | 60.53 | 42.11 | 0 | 353 | 100 | *Xanthomonas fragariae* IPO 3485 | [247] |
| NBC2815_02008 | 4HB_MCP_1 | 88.46 | 88.46 | 0 | 273 | 100 | *Xanthomonas fragariae* IPO 3485 | [247] |
| NBC2815_02009 | 4HB_MCP_1 | 82.14 | 75.00 | 0 | 340 | 100 | *Xanthomonas fragariae* IPO 3485 | [247] |

**Table 1.3: Distribution of LBD types among clusters with high DPS (≥50%)**

| LBD type | Number of clusters | % of clusters over total | Number of sequences | Average number of sequences per cluster |
|---|---|---|---|---|
| UNKNOWN | 243 | 41.75 | 2053 | 8.45 |
| 4HB_MCP_1 | 151 | 25.95 | 2642 | 17.50 |
| TarH | 26 | 4.47 | 1042 | 40.08 |
| HBM | 23 | 3.95 | 516 | 22.43 |
| CHASE3 | 9 | 1.55 | 142 | 15.78 |
| PilZ | 7 | 1.20 | 39 | 5.57 |
| PAS_9 | 6 | 1.03 | 7 | 1.17 |
| sCache_2 | 4 | 0.69 | 245 | 61.25 |
| NIT | 4 | 0.69 | 79 | 19.75 |
| Protoglobin | 3 | 0.52 | 153 | 51 |
| PAS_8 | 3 | 0.52 | 9 | 3 |
| PAS_3 | 3 | 0.52 | 82 | 27.33 |
| dCache_1 | 3 | 0.52 | 151 | 50.33 |
| Cache_3-Cache_2 | 3 | 0.52 | 79 | 26.33 |
| PAS_4 | 2 | 0.34 | 4 | 2 |
| CHASE4 | 2 | 0.34 | 3 | 1.50 |
| Usher | 1 | 0.17 | 1 | 1 |
| Tox-URI2 | 1 | 0.17 | 1 | 1 |
| SURF1 | 1 | 0.17 | 1 | 1 |
| SOR_SNZ | 1 | 0.17 | 1 | 1 |
| sCache_3_3 | 1 | 0.17 | 2 | 2 |
| Porin_4 | 1 | 0.17 | 1 | 1 |
| Peripla_BP_5 | 1 | 0.17 | 1 | 1 |
| PAS_7 | 1 | 0.17 | 19 | 19 |
| PapC_N | 1 | 0.17 | 1 | 1 |
| Glyco_hydro_2_N | 1 | 0.17 | 1 | 1 |
| Glyco_hydro_106 | 1 | 0.17 | 1 | 1 |
| FHIPEP | 1 | 0.17 | 1 | 1 |
| DUF4077 | 1 | 0.17 | 5 | 5 |
| dCache_3 | 1 | 0.17 | 71 | 71 |
| CBS | 1 | 0.17 | 2 | 2 |
| Asparaginase | 1 | 0.17 | 1 | 1 |
| ABC_tran | 1 | 0.17 | 1 | 1 |
| 5TM-5TMR_LYT | 1 | 0.17 | 20 | 20 |
| TOTAL | 510 | | 7377 | |

## 1.4 Phylogenetic vs. Ecological signal in PAB specific Ligand Binding Domains

Intrigued by the potential ecological significance of PAB-specific LBD clusters, we further tested whether their taxonomic distribution is due to the phylogenetic signal of the underlying species, or if it might be driven by additional ecological factors. To address this issue, we reconstructed the complete phylogeny of the 11,806 species considered here (see Materials and Methods), and used it to assess the taxonomic distribution of each individual LBD cluster. Using the $\delta$-approach [209], we found that the majority (75.7%) of plant-associated LBD types (DPS≥50%) did not follow the expected phylogenetic signal. By contrast, the taxonomic distribution of most PAB specific LBDs was scattered over the global bacterial phylogeny (Figure 1.4). This observation was consistent for

the three PAB reference databases considered in this study, using stricter DPS cutoffs, and even when the species lacking CR genes were excluded from the analysis (Supplementary Figure S1.1).

Overall, the lack of phylogenetic signal for most of the LBD clusters, together with the fact that the LBDs tested are enriched in PAB species, suggests that the evolution of the sensory machinery of bacterial species might be at least partially driven by ecological pressures. This should allow the use of particular LBD clusters, even if functionally undefined, as lifestyle biomarkers. This issue is best illustrated by the LBD cluster #549 (Figure 1.4B and C), which contains 27 CRs from broadly distributed bacterial families and orders, while retaining a high plant-association signal (DPS-broad >80%).

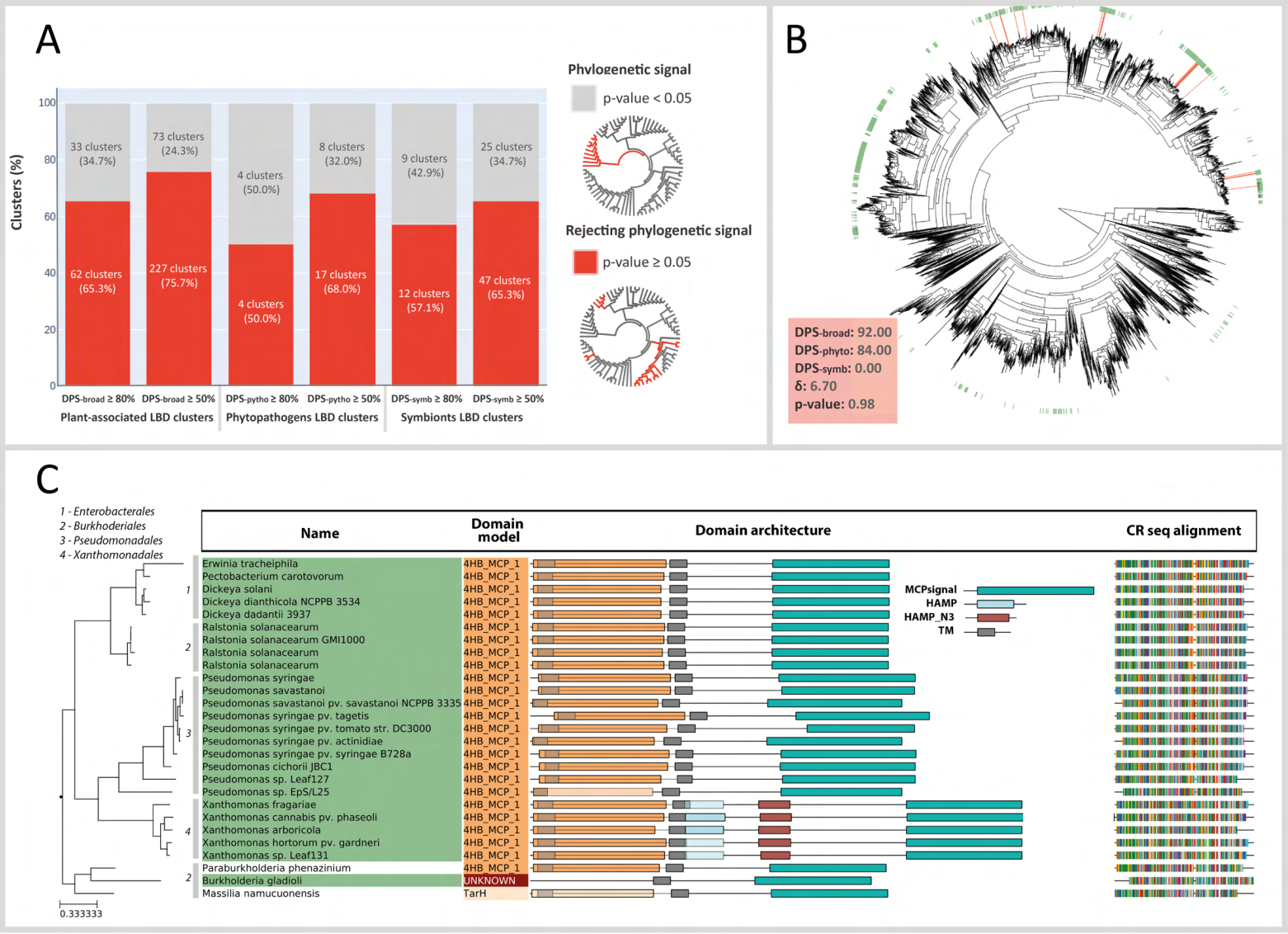



**Figure 1.4: Phylogenetic signal detection in plant-associated LBD clusters.** (A) Proportion of the significant phylogenetic signal among LBD clusters enriched in PAB-broad, PAB-phyto and PAB-symb species within two thresholds ($\geq$50%, $\geq$80%). The significance through a p-value test with 100 iterations, p-value $\geq$0.05 rejects the null hypothesis of a phylogenetic signal (see Materials and Methods). (B) LBD cluster #549 (in red) distributed according to the chemotactic species phylogeny (5,763 representative species). The green leaves of the tree represent the PAB species. DPS, $\delta$ and p-values for this LBD cluster are represented in the lower-left table. (C) Phylogeny representation of LBD cluster #549, containing 27 LBD sequences distributed across 4 orders (numbered 1-4 in the tree). The domain architecture prediction is shown for each of the CRs.

## 1.5 Ligand Binding Domains profiles per genome

To investigate whether the profile of LBD clusters per genome could be informative about the plant-associated bacterial lifestyle, we studied the full repertoire of CRs among different PAB species. The genomes from the PAB species not only contained more CRs than those of non-PAB species, but also, many of their CRs could be considered highly specific to plant-related environments. In fact, assessing the LBD profiles per genome showed that microorganisms with a pronounced plant-associated lifestyle (i.e., PAB-symb and PAB-phyto) harbour more specific CRs than other PAB species (Figure 1.5). On average, 28% and 20% of plant-symbiont and plant-phytopathogen CRs, respectively, are highly specific (DPS-broad > 80%). By contrast, other PAB with a less pronounced plant-associated lifestyle, like non-symbionts and non-phytopathogens plant-associated species, contained significantly fewer specific CRs (6%) (Figure 1.5). Taken together, this information reinforces the idea that the repertory of CRs has been partially shaped by niche adaptation, with more specialized adaptations leading to more specific CRs.

## 1.6 Discussion

In the present study, we carried out a comprehensive phylogenomic analysis of the full repertoire of CRs from a wide collection of microbial genomes, classifying them according to their LBDs. To maximize the representativeness of our study, we used more than 82,000 species-level CR sequences from 11,000 species-representative genomes, significantly expanding the scope of previous works [39], [50], [250], both in terms of the number of sequences examined and the phylogenetic coverage. To achieve this, we developed a novel method to extract LBDs and classified them based on a de novo homology-based clustering approach, departing from the traditional classification of CRs centred around their general protein topology [39], [148], [250], [251] or on known LBD domain searches [50]. This approach allowed us to identify many new potential LBD types, suggesting that the chemosensing landscape remains largely unexplored. Additionally, we believe that our strategy delineating large LBD families into finely grained sub-categories could provide further information (Figure 1.3). Moreover, by classifying CRs based on their putative LBD

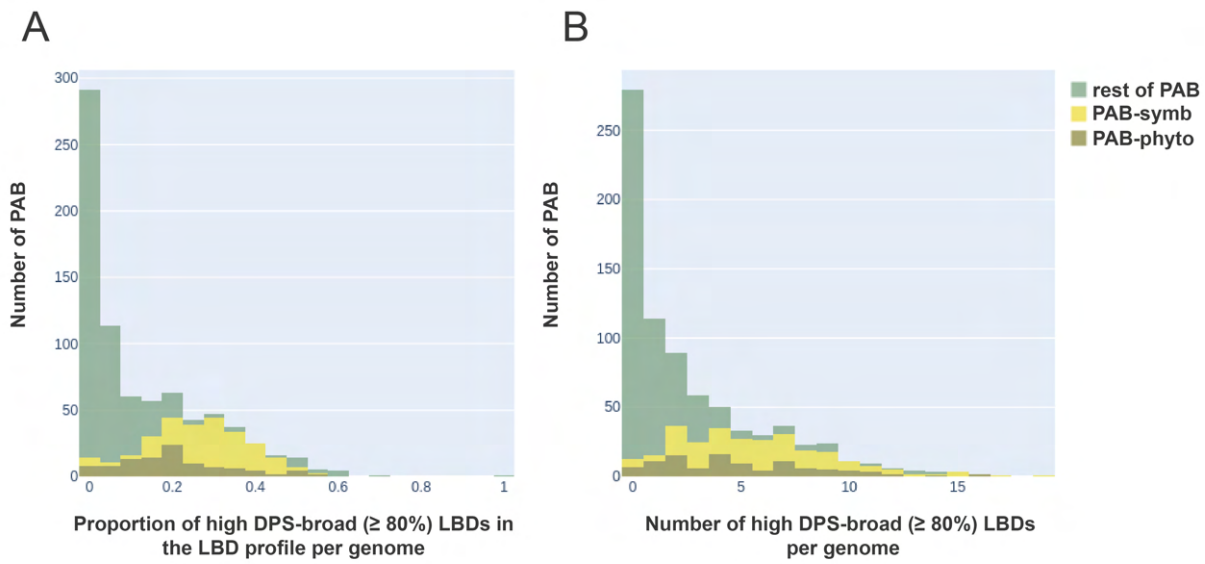



**Figure 1.5: Distribution of the high plant specific LBDs in the PAB species profiles per genome.** (A) Calculation of the proportion of high DPS-broad ($\geq$80%) LBDs in the total number of LBDs present in each species. The graph illustrates the distribution of the number of species according to the proportional ranges, plotting the species count as PAB-symb (yellow), PAB-phyto (brown) and the rest of PAB (green). (B) Absolute number of LBDs with a DPS-broad $\geq$80% in each PAB genome. The graph illustrates the distribution of the number of species as an absolute count of high DPS LBDs ranges. The species count is plotted as PAB-symb (yellow), PAB-phyto (brown) and the rest of PAB (green).

type, for the first time we were able to quantify to what extent the chemosensory activity of PAB is linked to lifestyle.

Considering the enormous variety of LBDs at sensor proteins, establishing the forces that have driven their evolution is an important question that was never specifically addressed. To our knowledge we present here the first clear demonstration showing that environmental factors play an important role in the selection and evolution of LBDs. We found that the specificity of LBDs to a plant-associated lifestyle could not be explained by just a phylogenetic signal, since the taxonomic distribution of most PAB-specific LBD types was scattered over the microbial phylogeny, that at times covered different orders and phyla. This indicates that the selection of the certain CRs might indeed be guided by ecological factors, opening the possibility of identifying lifestyle biomarkers. We also found that bacterial species more tightly associated to plant environments (such as plant symbionts and phytopathogens) tend to have stronger lifestyle specificity signals in their CR repertory. For instance, plant-symbionts had the largest number of PAB-specific LBDs per genome, followed by phytopathogens, with both showing significantly higher ratios than generic soil microbiota. It appears likely that even stronger links between the chemosensory capabilities of bacteria and their lifestyle will be detected in the future as more data becomes available on new organisms (e.g., via metagenomics sequencing) and data on their niche-adaptation (i.e., plant-tissue specificity).

These findings thus offer a number of research opportunities in the field of single transduction. Firstly, it can be explored whether similar relationships can be observed in CRs of bacteria with a different lifestyle such as for example those that inhabit or infect the human intestine. Another interesting issue that requires to be addressed is the question as to whether similar LBD types are shared by members of different sensor protein families. Major families of these receptors are sensor histidine kinases, chemoreceptors, adenylate, diadenylate and diguanylate cyclases, certain cAMP, c-di-AMP and c-di-GMP phosphodiesterases, as well as Ser/Thr/Tyr protein kinases and phosphoprotein phosphatases [252]. As the different sensor proteins of a given strain are exposed to the same signals, it appears plausible that the same LBD types might be present in members of different sensor protein families. Several examples have been reported in this direction, such as the specific sensing of nitrate by PilJ type LBDs of the NarQ type sensor kinases [253], the

McpN chemoreceptor [254], and the PAS domain, universally found in different signal transduction systems [252]. It would be of interest to estimate the global occurrence of such cases.

Overall, we believe that our study provides a comprehensive resource for future studies on bacterial chemoreception and that it sets the basis for the identification of novel CRs relevant for bacteria-plant interaction.

# Chapter 2

## Adaptive value and evolutionary history of bacterial chemoreceptors revealed by global metagenomic survey

To date, almost everything we know about bacterial chemoreception is based on the analysis of cultured or isolated species. However, environmental sampling in recent decades has revealed a vast number of uncultured microorganisms whose biology is poorly understood. Many unknown species are thought to possess chemosensitive properties, but even basic questions about them remain unanswered. How many new chemosensitive species exist in the natural environment? How common are they in different habitats? Do they encode novel sensing domains or transduction systems? How did this large superfamily of genes evolve throughout the prokaryotic phylogeny? These and many other questions remain unanswered despite recent metagenomic data that would, in principle, allow a comprehensive global view of chemoreception. In this chapter, we address this challenge by analyzing the entire landscape of chemoreceptors observed in both cultured and uncultured microorganisms, combining the most recent and comprehensive data available from genomics and metagenomics studies.

## 2.1 A global catalog of chemoreceptors from cultured and uncultured microorganism

We created a comprehensive catalog of CRs, hereafter referred to as globalCR, by mining two global genomic and metagenomic resources: proGenomes3[218] and SPIRE[187]. ProGenomes provides a comprehensive dataset of 41,171 fully sequenced reference genomes from known (cultured or isolated) bacterial and archaeal species. On the other hand, SPIRE represents a searchable, planetary-scale mIcrobiome REsource and contains over 35 billion microbial genes (bacteria and archaea) assembled from 99,146 million metagenomes and 739 studies covering a wide array of habitats. In addition, SPIRE provides 814,153 high-quality metagenome-assembled genomes, enabling better taxonomic assignments and genomic analysis of the observed uncultivated microorganisms.

To build a unified CR resource from SPIRE and proGenomes, we mined both resources using the same methodology, as described in Chapter 1 and detailed in Sanchis-López *et al.*[215]. Briefly, we identified significant HMM hits for the MCPsignal domain (Pfam[255]) across all protein sequences in each resource. This approach yielded 1,414,005 ORFs from SPIRE and 229,160 ORFs from

the non-redundant set in proGenomes3, resulting in a unified catalog (globalCR) with 1,643,165 entries. Compared to previous research on genome-scale CR catalogs, globalCR offers a 6-fold increase over CR counts in public databases like UniProt[256] and a 20-fold increase over our prior chemoreceptor analysis in Chapter 1.

Next, we taxonomically annotated each CR based on its source genome or MAG. To this end, we conducted *de novo* taxonomic annotation using GTDB-tk[219] and the GTDB r207 database[165], which comprises 317,542 reference genomes/MAGs from 62,291 bacterial and 3,412 archaeal species. This approach ensured that all CRs in the catalog were labeled with a normalized and standardized taxonomy. MAGs lacking species-level classification in GTDB were annotated to the most specific common ancestor possible.

To facilitate in-depth evolutionary analysis of CRs in the catalog, we further grouped CR genes into distinct subfamilies. Given the multidomain nature of CRs and the unique evolutionary and functional roles of each domain type, we classified all CRs based on (i) their sensing domain, or ligand-binding domain (LBD), and (ii) their signal transduction domain, or MCPsignal domain.

For the first classification (clustering CRs by their sensing domain), we applied the methodology developed in Sanchis-López *et al.*, (2021)[215], to identify and extract the putative ligand-binding domain (LBD) sequence region within each CR gene. An LBD region was defined from the translated CR gene sequence if it met either of the following criteria: (i) it contained a sequence region encoding at least 33 amino acids (subsequence) that matched a Pfam domain family other than HAMP, without overlapping the MCPsignal domain; or (ii) it was flanked by two transmembrane (TM) regions or located between a TM region and the N-terminus. Using these criteria, we extracted and performed deep-homology clustering on all LBD sequences, resulting in 42,375 clusters and 17,086 singletons. Since the LBD region of CRs determines their sensing capabilities and specificity, we consider each generated cluster to represent a distinct functional subfamily of CRs. Additionally, we labeled LBDs as intracellular if no transmembrane regions were detected along the full CR gene sequence, and as extracellular (i.e., exposed to the external environment) if transmembrane regions were present.

Similarly, we extracted the MCPsignal domain from each CR gene and clustered them using the same deep-homology clustering method as for the LBDs. Given that the MCPsignal domain is universally present in CRs, highly conserved, and functionally integral to signal transduction within the cell, we consider MCPsignal-based clusters to best represent the evolutionary history of CRs across species. Finally, to facilitate ecological studies of all CRs identified in environmental samples, we calculated the relative abundance of each LBD and MCPsignal cluster across all 99,146 samples and biomes represented in the SPIRE database.

Overall, our final global CR catalog offers a comprehensive resource of CR genes in both cultured and uncultured microorganisms. It includes taxonomic classification for each CR, predicted habitat distribution, relative abundance across habitats, and fine-grained subfamily classifications that account for the intrinsic multidomain structure of CRs and potential domain-specific evolutionary variations. The globalCR catalog, along with all associated metadata, is publicly available at https://github.com/compgenomicslab/globalCR.

## 2.2 Novel lineages of chemosensing microorganisms

Taxonomic and phylogenetic analysis of the globaCR catalog revealed that chemotactic microorganisms are widespread in the uncultivated data space. To investigate the distribution of chemoreception throughout the prokaryotic phylogeny, we estimated the number of chemotactic species—defined as those containing at least one chemotactic receptor (CR) from our globaCR catalog— at different taxonomic levels. For this analysis, we used the entire GTDB phylogeny as a reference, which represents all cultivated and uncultivated species identified to date. Additionally, we calculated the prevalence of chemotactic species within each clade, as well as the maximum and average number of CRs per species.

Figure 2.1 provides a visual overview of these mappings across the entire GTDB phylogeny, collapsed to the phylum level (other collapsed trees are available as supplementary material). The figure shows that CR-containing species are widely distributed across the phylogeny. Notably, only 28 of the 189 phyla present in GTDB r207 lacked species with chemotactic capabilities. Conversely, several CR-rich nodes were identified, including some previously reported lineages enriched for

CR capabilities. In addition, we uncovered numerous uncultivated lineages not previously reported as potentially chemotactic. These newly discovered chemotactic lineages span 28 phyla (blue background in Figure 2.1), 72 classes, 240 orders, 531 families and 1,493 genera.

In addition, some uncultivated-only lineages emerged as chemosensing hubs with remarkably high CR counts per genome. For example, while the average number of CRs per species was 12.33 for reference isolated species and 8.78 for uncultivated species, the uncultivated phylum *Moduliflexota* had an average of 61.76 CRs per species (Table 2.1). These results underscore the importance of studying uncultivated lineages and highlight significant knowledge gaps in our current dataset.

Finally, given that many CRs in the globalCR catalog belonged to metagenome-assembled genomes (MAGs) that could not be confidently assigned to any reference prokaryotic species in the GTDB—and therefore not mapped to GTDB reference tree—, we used lower-level taxonomic classifications to estimate the number of novel uncultivated chemotactic species present in our dataset. To do this, we used unsupervised species-level clustering of all MAGs in the SPIRE database to estimate how many species clusters were unclassifiable at species-level using GTDB-tk and included at least one MAG with a CR gene. Table 2.2 summarizes the number of novel chemotactic species found, grouped by their last classifiable common ancestor. Among them, we identified a novel chemotactic species from an unknown bacterial phylum that harbored up to 30 CRs in its genome—more than three times the global average observed in metagenomic species.

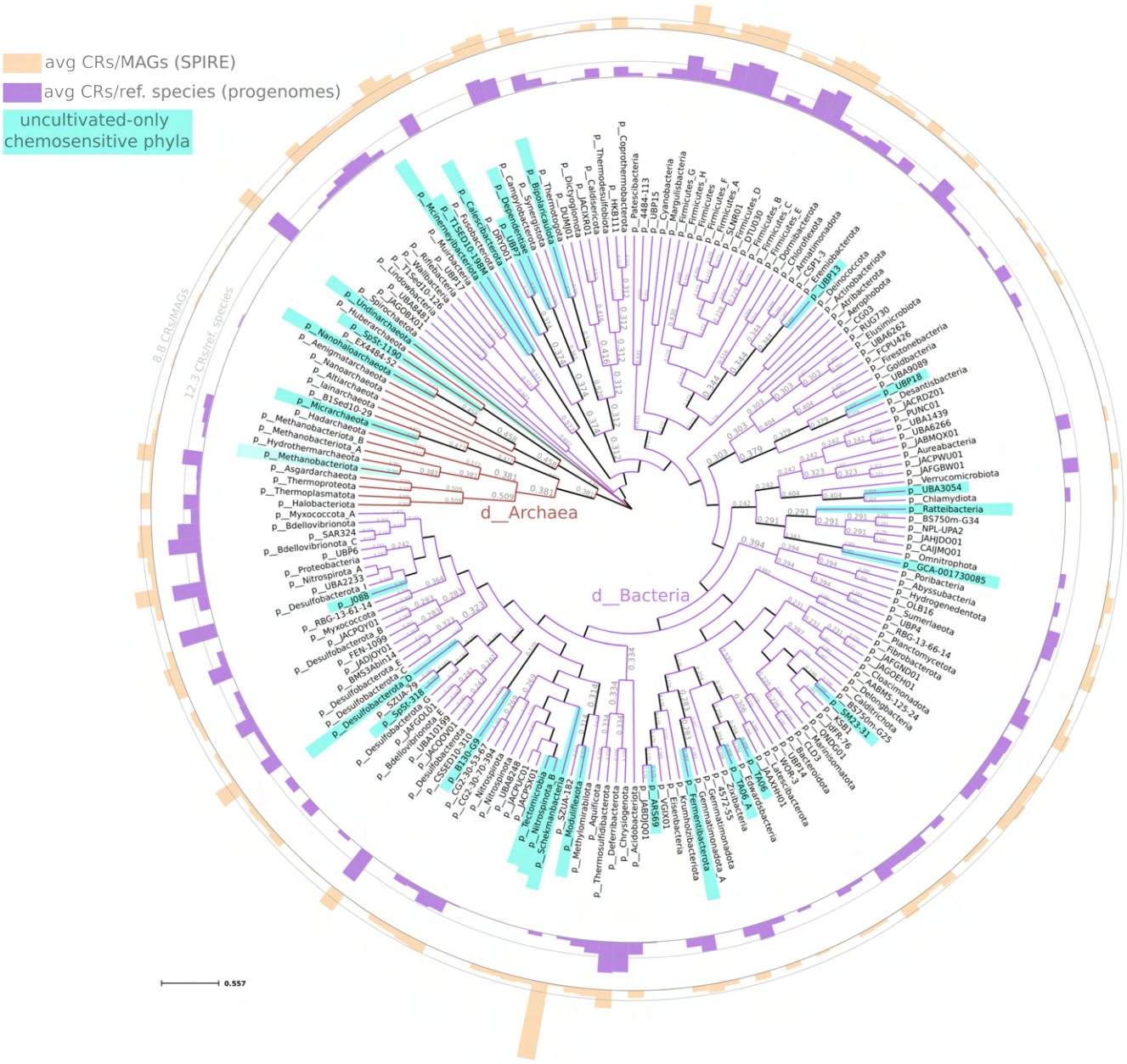



**Figure 2.1: Distribution of CRs across the prokaryotic tree, in reference genomes and Metagenome Assembly Genomes.** The mean number of CRs per chemosensitive species is calculated for each phyla within the Genome Taxonomy Database tree r207, and plotted in a barplot for the metagenomic (orange) and reference genomic (purple) data. Scale is showing mean CRs lines in both datasets (8.3 CRs/MAGs and 12.3 CRs/ref. species). The 28 chemosensitive uncultivated-only phyla are highlighted in blue.

**Table 2.1: Top lineages with highest mean CR content per chemosensitive species.** The top 25 lineages for phylum and order are shown, with the average content of CRs per each species cluster containing at least one CR within each lineage.

| phylum | mean CRs per SPIRE species | mean CRs per proGenomes species | order | mean CRs per SPIRE species | mean CRs per proGenomes species |
|---|---|---|---|---|---|
| ***Moduliflexota*** | 61.76 | - | ***Moduliflexales*** | 61.76 | - |
| *Desulfobacterota_I* | - | 28.72 | *Azospirillales* | 54.79 | 58.04 |
| *Chrysiogenota* | - | 23.61 | *UMGS1260* | 47.50 | 61.00 |
| *Margulisbacteria* | 21.61 | 19.22 | *Elsterales* | 41.87 | 48.53 |
| ***T1SED10-198M*** | 18.99 | - | *Propionisporales* | 36.38 | 39.86 |
| *Proteobacteria* | 18.37 | 26.19 | *Rhodospirillales* | 34.95 | 43.07 |
| *Desulfobacterota_C* | 18.23 | 13.19 | *Sporomusales* | 34.88 | 42.45 |
| ***Desulfobacterota_F*** | 17.53 | - | *Halobacteroidales* | 34.35 | 32.05 |
| *Desulfobacterota* | 16.14 | 24.82 | *Spirochaetales_E* | 29.68 | 29.11 |
| *Firmicutes_A* | 15.87 | 21.74 | *GWE2-31-10* | 28.27 | 30.27 |
| *Firmicutes_F* | 15.66 | 23.15 | *Caldalkalibacillales* | 28.00 | 16.69 |
| *Spirochaetota* | 15.10 | 23.22 | *UBA4802* | 26.65 | 31.87 |
| *JADJOY01* | - | 15.00 | *Pseudomonadales* | 26.15 | 30.41 |
| *Firmicutes_C* | 14.96 | 26.91 | *Leptospirales* | 25.99 | 22.84 |
| *Deferribacterota* | 14.86 | 24.48 | *Kapabacteriales* | 25.71 | 36.50 |
| *Synergistota* | 14.09 | 15.61 | *GWF2-39-127* | 25.67 | 28.00 |
| *Firmicutes_E* | 13.94 | 7.00 | *Bacillales_A* | 25.64 | 28.93 |
| *Firmicutes* | 13.68 | 19.40 | ***SHVW01*** | 25.07 | - |
| *Fusobacteriota* | 13.34 | 15.71 | *Symbiobacteriales* | 24.70 | 11.11 |
| *4572-55* | - | 13.31 | *LC-3* | 24.29 | |
| *Asgardarchaeota* | 12.89 | 14.87 | *SZUA-567* | 23.98 | 15.12 |
| *T1Sed10-126* | 12.68 | 4.00 | ***Selenomonadales_A*** | 23.64 | - |
| *Thermotogota* | 12.41 | 12.31 | *Ferrovibrionales* | 23.62 | 26.03 |
| *Firmicutes_B* | 12.28 | 22.61 | *Magnetococcales* | 23.18 | 31.13 |
| *Campylobacterota* | 12.19 | 18.36 | *JC444* | 22.92 | 30.65 |

**Table 2.2: Unknown chemosensitive species clusters across globalCR at different rank levels.** The number of different species clusters without a classification with gtdb-tk using r207 released are counted by the last rank that could be classified. In the second column the number of different GTDB assignments to all the species is illustrated, together with the gtdb-tk classifications for the top 5 spire clusters with the highest CRs content.

| Rank | nº unknown species clusters | Number different gtdb mapping and top 5 lineages in CR content |
|---|---|---|
| **Phylum** | 2 | 1<br>d__Bacteria;p__;c__;o__;f__;g__;s__,spire_v1_095_000562990 (30 CRs)<br>d__Bacteria;p__;c__;o__;f__;g__;s__,spire_v1_095_000529458 (6 CRs) |
| **Class** | 11 | 10<br>d__Bacteria;p__Fibrobacterota;c__;o__;f__;g__;s__,spire_v1_095_000529778 (9 CRs)<br>d__Bacteria;p__Calditrichota;c__;o__;f__;g__;s__,spire_v1_095_000342247 (9 CRs)<br>d__Bacteria;p__Bacteroidota;c__;o__;f__;g__;s__,spire_v1_095_000615922 (8 CRs)<br>d__Bacteria;p__GCA-001730085;c__;o__;f__;g__;s__,spire_v1_095_000627810 (7 CRs)<br>d__Bacteria;p__Calditrichota;c__;o__;f__;g__;s__,spire_v1_095_000342247 (6 CRs) |
| **Order** | 79 | 45<br>d__Bacteria;p__Chloroflexota;c__UBA5177;o__;f__;g__;s__,spire_v1_095_000322964 (27 CRs)<br>d__Bacteria;p__RBG-13-61-14;c__RBG-13-61-14;o__;f__;g__;s__,spire_v1_095_000300596 (19 CRs)<br>d__Bacteria;p__RBG-13-61-14;c__RBG-13-61-14;o__;f__;g__;s__,spire_v1_095_000338872 (18 CRs)<br>d__Bacteria;p__RBG-13-61-14;c__RBG-13-61-14;o__;f__;g__;s__,spire_v1_095_000627422 (17 CRs)<br>d__Bacteria;p__RBG-13-61-14;c__RBG-13-61-14;o__;f__;g__;s__,spire_v1_095_000645466 (16 CRs) |
| **Family** | 252 | 153<br>d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Peptostreptococcales;f__;g__;s__,spire_v1_095_000370048 (37 CRs)<br>d__Bacteria;p__Campylobacterota;c__Campylobacteria;o__Campylobacterales;f__;g__;s__,spire_v1_095_000411721 (32 CRs)<br>d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Tissierellales;f__;g__;s__,spire_v1_095_000278696 (32 CRs)<br>d__Bacteria;p__Campylobacterota;c__Campylobacteria;o__Campylobacterales;f__;g__;s__,spire_v1_095_000411720 (26 CRs) |
| **Genus** | 2,730 | 840<br>d__Bacteria;p__Campylobacterota;c__Campylobacteria;o__Campylobacterales;f__JACAEZ01;g__;s__,spire_v1_095_000275611 (68 CRs)<br>d__Bacteria;p__Spirochaetota;c__UBA4802;o__UBA4802;f__UBA5550;g__;s__,spire_v1_095_000514294 (64 CRs)<br>d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pseudomonadales;f__Nitrincolaceae;g__;s__,spire_v1_095_000371673 (60 CRs)<br>d__Bacteria;p__Spirochaetota;c__UBA4802;o__UBA4802;f__UBA5550;g__;s__,spire_v1_095_000514262 (60 CRs)<br>d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__GCF-002020875;f__GRL18;g__;s__,spire_v1_095_000514374 (59 CRs) |
| **Species** | 12,925 | 3,224<br>d__Bacteria;p__Moduliflexota;c__Moduliflexia;o__Moduliflexales;f__Moduliflexaceae;g__Moduliflexus;s__,spire_v1_095_000456502 (75 CRs)<br>d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Azospirillales;f__Azospirillaceae;g__Azospirillum;s__,spire_v1_095_000283837 (74 CRs)<br>d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Clostridiales;f__Clostridiaceae;g__Clostridium;s__,spire_v1_095_000600210 (74 CRs)<br>d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhodospirillales;f__Magnetospirillaceae;g__Phaeospirillum;s__,spire_v1_095_000542334 (73 CRs)<br>d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Azospirillales;f__Azospirillaceae;g__Azospirillum;s__,spire_v1_095_000283836 (73 CRs) |

## 2.3 Origin and evolution of CRs

To explore the evolutionary history of CR genes across all known and unknown species, we performed a phylogenetic analysis of all observed 1,643,165 CR sequences. Given the extensive diversity and variability within the CR gene family, we focused on the MCPsignal domain, which, as noted above, is universally present across CR types and is more conserved than the ligand-binding domain. This approach allowed us to reconstruct the evolutionary history of CR genes across species, minimizing the variability introduced by the ligand-binding regions, which are likely to be under different selective pressures.

To this end, we extracted the MCPsignal domain region from each CR in the catalog and clustered them at 99% protein identity, resulting in a set of 428,179 unique MCPsignal domain sequences. We then used FAMSA[220] to compute a multiple sequence alignment (MSA) of all MCPsignal domain sequences and inferred a phylogenetic tree using the approximate maximum likelihood method implemented in FastTree2[212]. While this approach may not use the most precise methods available for phylogenetic reconstruction, it allowed us to construct what we believe to be the first comprehensive evolutionary tree of CRs across all prokaryotic species.

Next, using the taxonomic annotations available for each CR in our catalog, we identified gene duplication events in the resulting phylogeny and explored their functional implications. Our findings indicate that these duplication events often lead to functional subspecialization within the signal transduction system. For example, MCPsignal domains exhibit distinct structural configurations, typically classified by the number of heptads—regions of seven consecutive amino acid residues that correspond to two $\alpha$-helical turns[38]. As shown in Figure 2.2, certain heptad classes display clear monophyletic distributions in the phylogenetic tree, such as 34H, 36H, and 38H, while others, like 44H and 40H, are more dispersed. Detailed examination of the phylogenetic distribution of each heptad type also revealed specialization patterns across classes. In Archaea, for instance, the 44H class dominates, representing 95.4% of all archaeal chemoreceptors, while only 4.6% belong to other subclasses. Similarly, as shown in Figure 2.3, the 48H heptad class appears almost exclusive to the *Spirochaetota* phylum (95.9%), particularly within the *Treponemataceae* family (82.1%). Likewise, the 44H class, one of the most abundant heptad classes (Supplementary Table

2), is primarily associated with the *Firmicutes_A* phylum (82.96%), especially the *Lachnospiraceae* family (78.95%) (Figure 2.3).

Some heptad classes are closely linked to specific signaling pathways through interactions with different histidine kinase CheA subtypes, each associated with unique cellular responses. These CheaA subtypes have been categorized by their roles, including Alternative Cellular Function (CheA-ACF), Type IV pili motility (CheA-Tfp), and various flagellar motility systems (CheA-F1 to CheA-F17)[18]. To investigate the relationship between CheA subtypes and evolutionary events in the MCPsignal phylogeny, we conducted a phylogenetic profiling analysis across all genomes and MAGs in our globalCR catalog. Our results show that CheA-ACF is the most widespread pathway, present in 87.3% of all chemotactic species (Supplementary Table 3), with an average of 3.8 copies per genome. The Type IV motility system follows, present in 73% of species. Among flagellar motility subtypes, CheaA-F7 is the most common (61.6%), followed by CheA-F1 (42.4%) and CheA-F9 (34.9%).

To further analyze the relationship between heptad classes and CheA subtypes, we performed a phylogenetic profiling and correlation analysis of these features across the MCPsignal phylogenetic tree. Using TreeProfiler[221], we mapped the distribution of heptad classes, CheA subtypes, and other associated traits across the entire MCPsignal domain tree. While direct assignment of a heptad class to a specific CheA subtype is often feasible when only one CheA type is present in a species, the correlation patterns of heptads and CheA subtypes also provide insights into the relationships between chemoreceptor structures and signaling pathways (Figure 2.4). For instance, we confirmed an association between the 38H heptad and CheA-F5 motility systems, with an additional link to 42H and 58H. Additionally, CheA-F3 was associated with the 28H heptad, CheA-F1 with 44H, and CheA-F10 with 64H.

We integrated the CheA profiles into the chemoreceptor phylogenetic tree (Figure 2.2), which revealed several co-evolutionary patterns. For example, the 38H heptad consistently co-evolves with the CheA-F5 subtype (Figure 2.2a and zoom in Supplementary Figure 2.1a). Additionally, within the *Campylobacter* order, a cluster of 40H heptads, including some sequences from the 28H type, shows co-evolution with the F3 system (Figure 2.2b and zoom in Supplementary Figure 2.1b). The tree also indicates that the 48H heptad emerged as a duplication of the 44H class

and subsequently specialized within the *Treponema* order, co-evolving with CheA-F2 systems (Figure 2.2c and zoom in Supplementary Figure 2.1c). Another notable pattern was observed in the *Desulfobacterales* order, where a unique 40H chemoreceptor co-evolved with the F4 system, the only F4 system in the entire tree (Figure 2.2c and zoom in Supplementary Figure 2.1c). This order also contains the majority of 42H heptads, which appear to have diverged from 38H in the tree (Figure 2.2a and zoom in Supplementary Figure 2.2a). Finally, the rare 64H heptad class shows a strong association with the CheA-F10 and CheA-F12 systems (Figure 2.4), primarily within the *Geobacterales* order (Supplementary Figure 2.1d).

## 2.4 Detecting HGT in CRs across metagenomic data

In this section we systematically identified potential horizontal gene transfer (HGT) events affecting chemoreceptors across metagenomic data. To do this, we searched for chemoreceptor proteins that are nearly identical (99% protein identity and coverage) but are distributed across distant microbial lineages. As a quality filter, we only considered putative HGT events where at least two sequences per lineage were present within a cluster, and these lineages had to span different orders to rule out vertical inheritance as a source of the similarity. From the 516,960 clusters analyzed, we identified 39 candidates classified as putative HGT events (Table 2.3).

Notably, 28 of the 39 putative HGT events occurred in host-associated environments, particularly the human intestine. The *Lachnospiraceae* family was involved in 31 of the 39 putative HGT events, representing 80% of the total, while *Bacteroidaceae* was involved in 30 events (77%). Of these, 29 HGT (74%) occurred between *Lachnospiraceae* and *Bacteroidaceae*, with 18 events (46%) occurring exclusively between the two families, and the remaining 11 events involving other families as well. The majority of these HGT occurred between genera such as *Agathobacter*, *Roseburia*, *Blautia_A* and *Lachnospira* within *Lachnospiraceae*, and *Bacteroides* and *Phocaeicola* within *Bacteroidaceae*. This highlights the prominent role of these two families, particularly in the human gut, in the acquisition and sharing of chemoreceptor genes. In addition, other families such as *Erysipelotrichaceae* (5 putative HGT events, 13%) and *Ruminococcaceae* (2 putative HGT

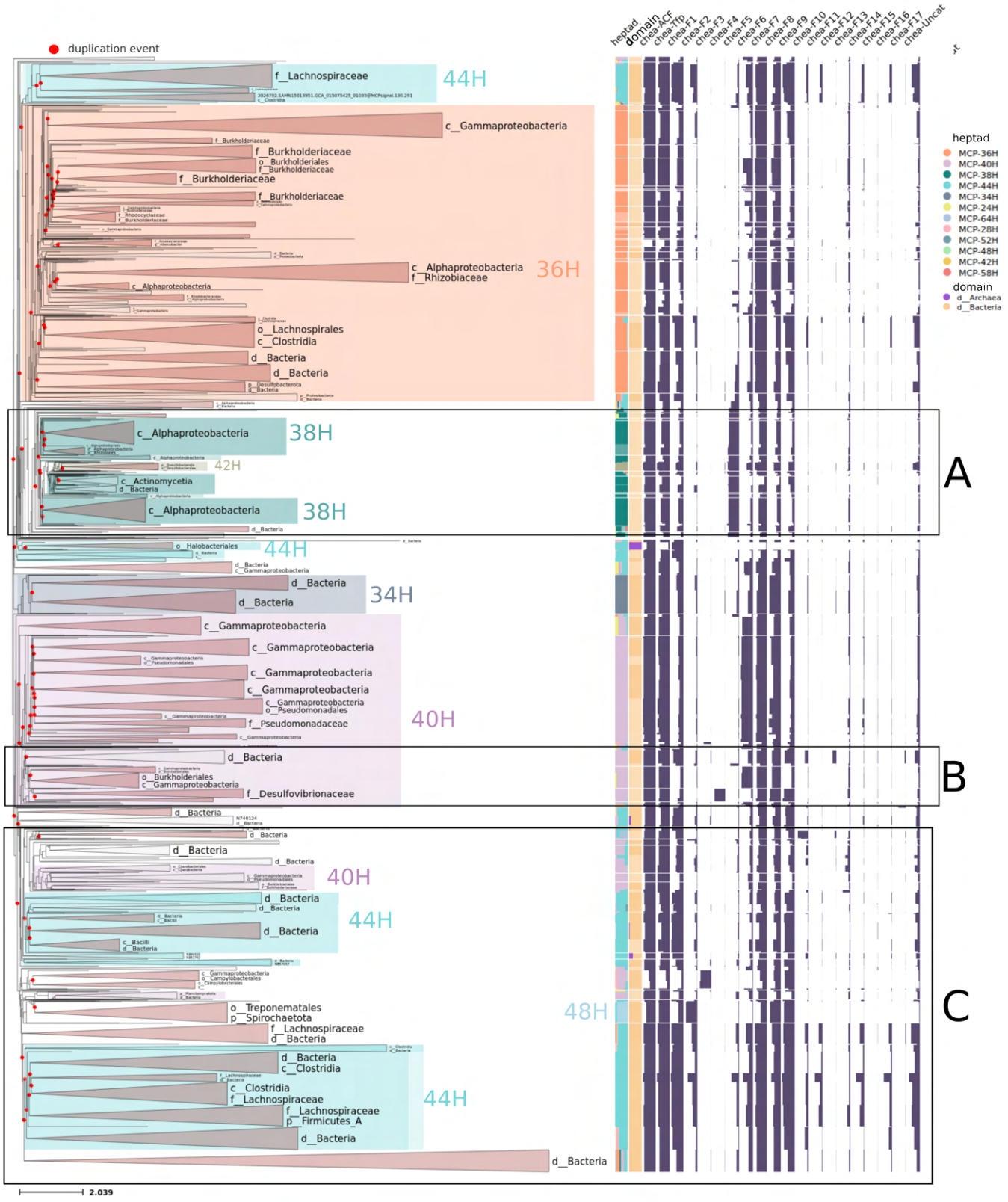



**Figure 2.2: Phylogenetic tree reconstruction of CRs across globalCR.** The tree includes a total number of 428,179 unique MCPsignal with the heptad subtype metadata annotated, as well as the co-ocurrence pattern with the different CheA histidine kinases within the same genome. Collapsed nodes in red are duplication branches, and the last common ancestor name is shown. The background of the branches is of an heptad color only if more than 90% of the sequences are from the same heptad.

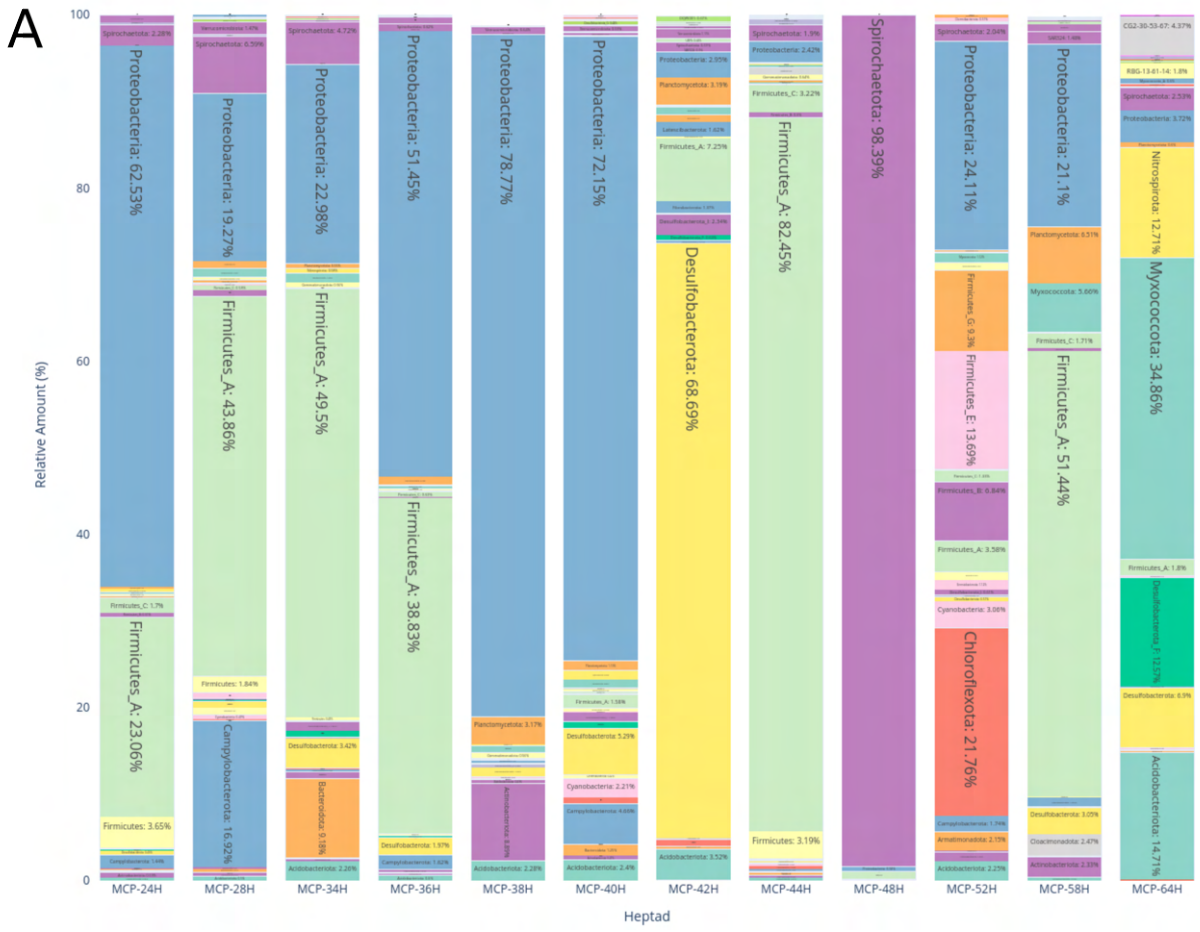



**Figure 2.3: Phylogenetic distribution of heptad subtypes at phylum level.** A) The stacked bars represent the phyla distribution of the total sequences in that heptad subtype.

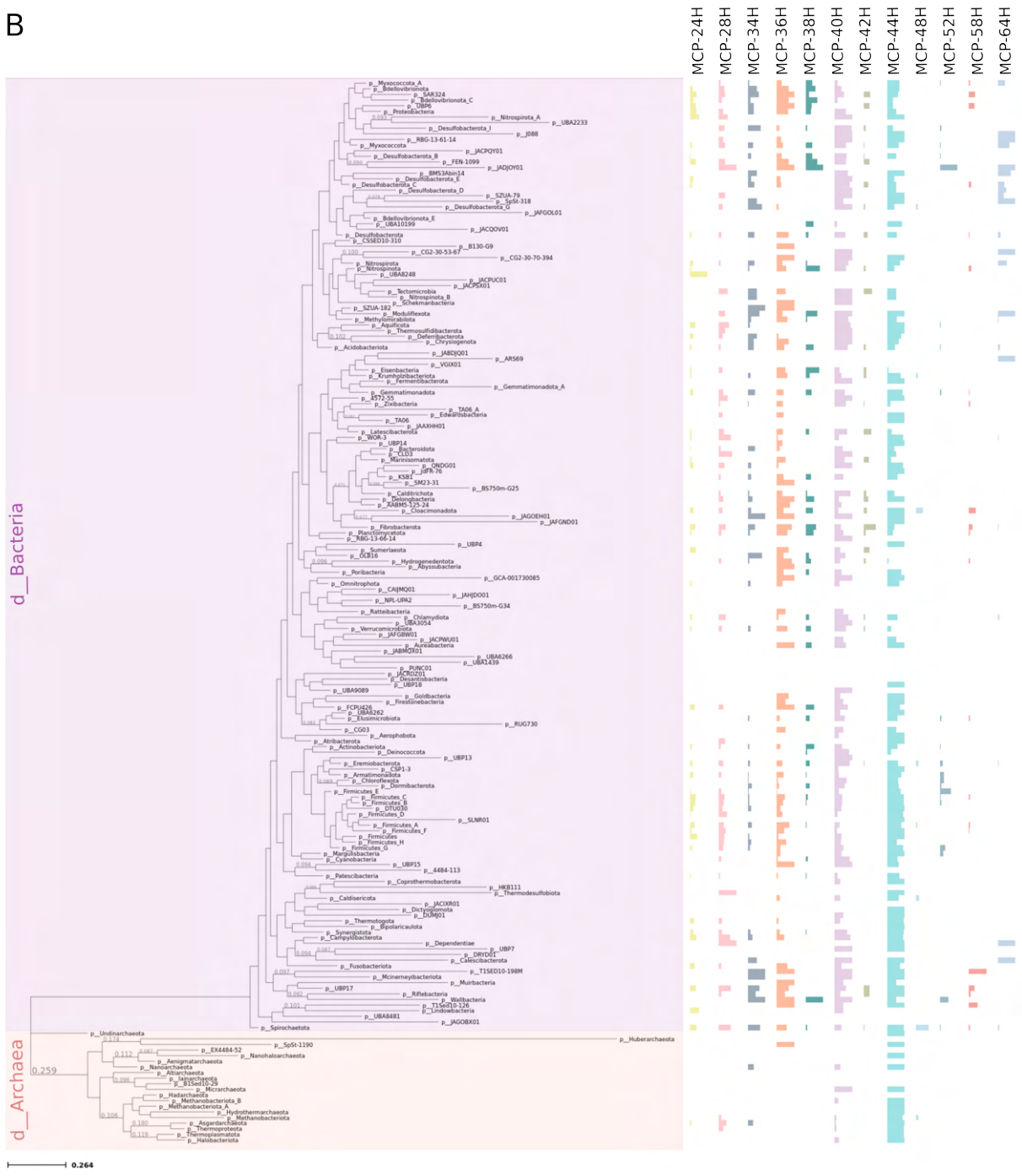



**Figure 2.3: Phylogenetic distribution of heptad subtypes at phylum level.** B) GTDB tree collapsed at the phylum level, with bar plots indicating the proportion of chemosensitive genus within each phylum associated with each heptad subtype.

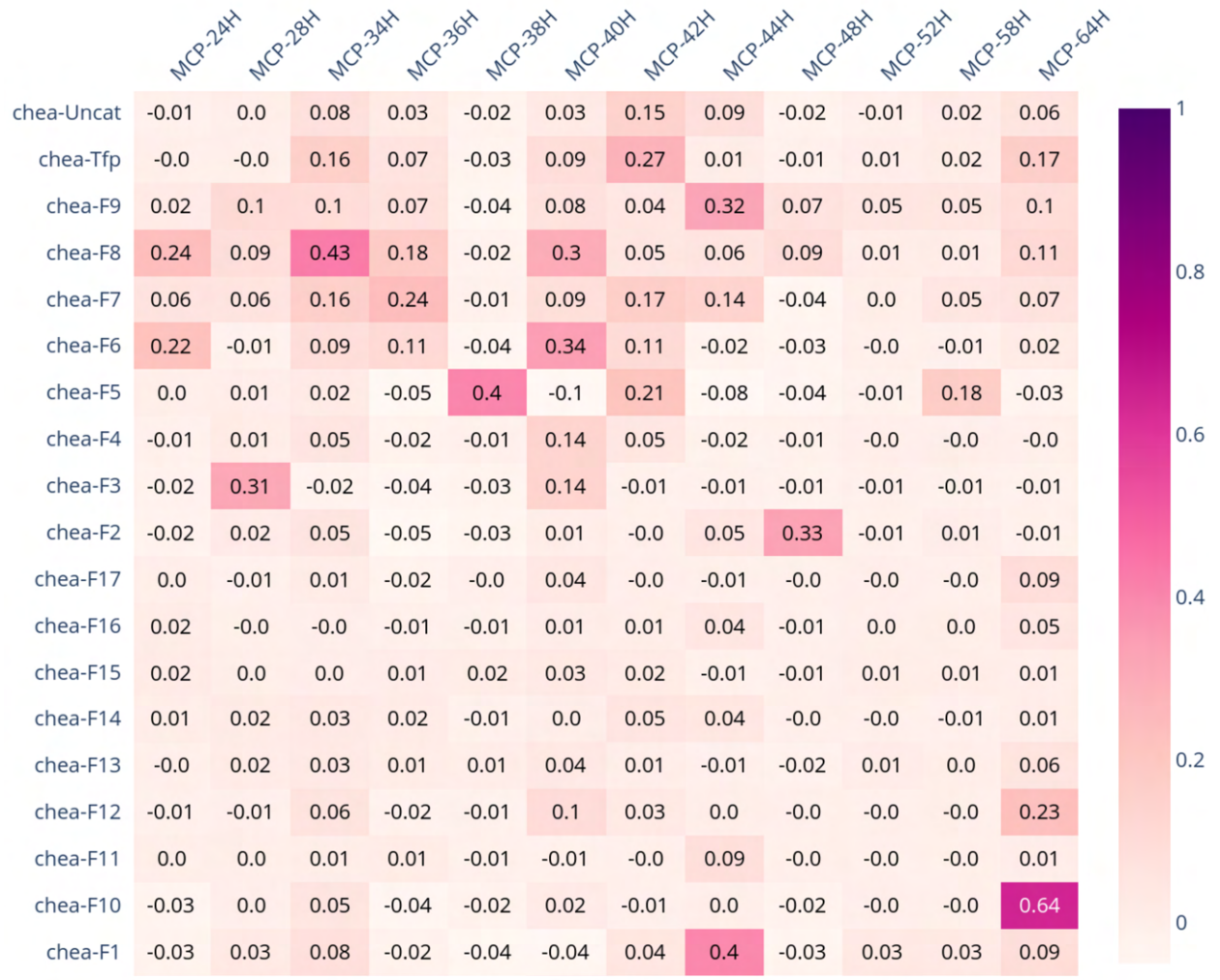



**Figure 2.4: Correlation matrix of MCPsignal heptads and CheA subtypes across species clusters in globalCR.** The heatmap displays Pearson correlation coefficients between MCPsignal heptads and CheA subtypes based on their mean abundances across species clusters. Color intensity indicates the strength and direction of correlations, with values ranging from -1 (negative correlation) to +1 (positive correlation).

events, 5%) were also involved, although less frequently, in environments such as the human gut and rumen.

A specific example, putative_HGT_1, illustrates an interesting event, involving 3 phyla, 5 orders, and 7 families. The chemoreceptors of this potential event harbor CHASE3 ligand-binding domain, and have been found in terrestrial, plant-associated and aquatic environments. The CHASE3 domain has previously been shown to bind compounds like pyrene[257] (a polycyclic aromatic hydrocarbon and environmental pollutant) and acetoin[258] (a volatile compound involved in establishing symbiotic relationships).

**Table 2.3: Putative Horizontal Gene Transfer Events across globalCR.**

| cluster name | Phyla | Order | Family | Genus | LBD | habitat |
|---|---|---|---|---|---|---|
| putative_HGT_1 | p__Acidobacteriota, p__Actinobacteriota, p__Proteobacteria | o__Acidimicrobiales, o__Acidobacteriales, o__Bryobacterales, o__Solirubrobacterales, o__Steroidobacterales | f__Acidobacteriaceae, f__Bog-793, f__Bryobacteraceae, f__Koribacteraceae, f__RAAP-2, f__Solirubrobacteraceae, f__Steroidobacteraceae | g__Bog-209, g__Bog-375, g__Bog-473, g__FEN-1191, g__Fen-455, g__Palsa-288, g__Palsa-343, g__Palsa-744, g__RAAP-2 | CHASE3 | Aquatic: Freshwater Terrestrial: Agricultural and Wet Environments Terrestrial: Extreme Environments Terrestrial: Soil and Dry Environments |
| putative_HGT_2 | p__Actinobacteriota, p__Firmicutes, p__Firmicutes_C | o__, o__Acidaminococcales, o__Lactobacillales, o__Selenomonadales | f__, f__Acidaminococcaceae, f__Enterococcaceae, f__Selenomonadaceae, f__Streptococcaceae | g__, g__Enterococcus_I, g__RZYP01, g__Streptococcus | UNKNOWN | Anthropogenic: Industrial and Built Environments Anthropogenic: Waste Handling Aquatic: Freshwater |
| putative_HGT_3 | p__Bacteroidota, p__Desulfobacterota_G | o__Bacteroidales, o__Syntrophorhabdales | f__Syntrophorhabdaceae, f__UBA932, f__VadinHA17 | g__DMER64, g__LD21, g__MWEV01, g__SR-FBR-E99, g__Syntrophorhabdus, g__UBA5609 | | Anthropogenic: Waste Handling |
| putative_HGT_4 | p__Bacteroidota, p__Firmicutes, p__Firmicutes_A | o__Bacteroidales, o__Erysipelotrichales, o__Lachnospirales | f__Bacteroidaceae, f__Erysipelotrichaceae, f__Lachnospiraceae | g__Acetatifactor, g__Agathobacter, g__Anaerostipes, g__Bacteroides, g__Bacteroides_F, g__Bariatricus, g__Blautia_A, g__Bulleidia, g__Dorea_A, g__KLE1615, g__Lachnospira, g__Mediterraneibacter, g__Phocaeicola, g__Prevotella, g__Roseburia | UNKNOWN | Anthropogenic: Industrial and Built Environments Anthropogenic: Waste Handling Host-Associated: Human intestine |
| putative_HGT_5 | | | | g__Agathobacter, g__Anaerostipes, g__Bacteroides, g__Blautia_A, g__COE1, g__Coprococcus, g__Holdemanella, g__Lachnospira, g__Mediterraneibacter, g__Phocaeicola, g__Prevotella, g__Roseburia | UNKNOWN | Host-Associated: Human intestine |
| putative_HGT_6 | | o__Bacteroidales, o__Erysipelotrichales, o__Lachnospirales, o__Oscillospirales | f__Bacteroidaceae, f__Erysipelatoclostridiaceae, f__Erysipelotrichaceae, f__Lachnospiraceae, f__Ruminococcaceae | g__Agathobacter, g__Blautia_A, g__Butyrivibrio_A, g__Catenibacterium, g__Lachnospira, g__Prevotella, g__Ruminococcus, g__UBA2868, g__UBA636 | UNKNOWN | Host-Associated: Animal intestine/rumen |

## Table 2.3: Putative Horizontal Gene Transfer Events across globalCR.

| cluster name | Phyla | Order | Family | Genus | LBD | habitat |
|---|---|---|---|---|---|---|
| putative_HGT_7 | | o__Bacteroidales, o__Erysipelotrichales, o__Peptostreptococcales | f__Anaerovoracaceae, f__Bacteroidaceae, f__Erysipelotrichaceae | g__Bacteroides, g__Emergencia, g__Longicatena, g__NSJ-61 | 4HB_MCP_1 | Host-Associated: Animal intestine/rumen Host-Associated: Human intestine |
| putative_HGT_8 | p__Bacteroidota, p__Firmicutes_A | o__Bacteroidales, o__Lachnospirales | f__Bacteroidaceae, f__Lachnospiraceae | g__1XD8-76, g__An181, g__CAG-95, g__Phocaeicola | 4HB_MCP_1 | Host-Associated: Animal intestine/rumen |
| putative_HGT_9 | | | | g__AM51-8, g__Acetatifactor, g__Agathobacter, g__Bacteroides, g__Blautia_A, g__COE1, g__KLE1615, g__Phocaeicola, g__Roseburia | | Anthropogenic: Industrial and Built Environments Host-Associated: Human intestine |
| putative_HGT_10 | | | | g__Acetatifactor, g__Agathobacter, g__Anaerostipes, g__Bacteroides, g__Blautia_A, g__Butyrivibrio_A, g__CAG-127, g__CAG-882, g__CAG-95, g__COE1, g__Faecalimonas, g__KLE1615, g__Lachnospira, g__Mediterraneibacter, g__RUG115, g__Roseburia | sCache_2 | Anthropogenic: Industrial and Built Environments Anthropogenic: Waste Handling Host-Associated: Human intestine |
| putative_HGT_11 | | | | g__Acetatifactor, g__Agathobacter, g__Anaerostipes, g__Bacteroides, g__Blautia_A, g__Coprococcus, g__Coprococcus_A, g__Lachnospira, g__Roseburia | 4HB_MCP_1 | Anthropogenic: Industrial and Built Environments Anthropogenic: Waste Handling Host-Associated: Animal intestine/rumen Host-Associated: Human intestine |
| putative_HGT_12 | | | | g__Acetatifactor, g__Agathobacter, g__Bacteroides, g__Blautia_A, g__CAG-127, g__Phocaeicola, g__RUG115, g__Roseburia, g__TF01-11 | UNKNOWN | Anthropogenic: Waste Handling Host-Associated: Human intestine |
| putative_HGT_13 | | | | g__Acetatifactor, g__Agathobacter, g__Bacteroides, g__Blautia_A, g__CAG-603, g__Faecalimonas, g__KLE1615, g__Lachnospira, g__Phocaeicola, g__Roseburia, g__Schaedlerella | UNKNOWN | Host-Associated: Human intestine |

**Table 2.3: Putative Horizontal Gene Transfer Events across globalCR.**

| cluster name | Phyla | Order | Family | Genus | LBD | habitat |
|---|---|---|---|---|---|---|
| putative_HGT_14 | | | | g__Acetatifactor, g__Agathobacter, g__Blautia_A, g__CAG-194, g__CAG-632, g__CAG-95, g__Dorea_A, g__Mediterraneibacter, g__Phocaeicola, g__Prevotella, g__RUG115, g__Roseburia, g__TF01-11, g__UBA11774 | 4HB_MCP_1 | Host-Associated: Animal intestine/rumen Host-Associated: Human intestine |
| putative_HGT_15 | | | | g__Acetatifactor, g__Agathobacter, g__Blautia_A, g__Dorea_A, g__Lachnospira, g__Phocaeicola, g__Roseburia | | Anthropogenic: Industrial and Built Environments Host-Associated: Human intestine |
| putative_HGT_16 | | | | g__Agathobacter, g__Anaerostipes, g__Bacteroides, g__Blautia_A, g__CAG-95, g__Eubacterium_F, g__Faecalimonas, g__Fusicatenibacter, g__KLE1615, g__Mediterraneibacter, g__Phocaeicola, g__RUG115, g__Roseburia, g__UMGS1375 | | Host-Associated: Human intestine |
| putative_HGT_17 | | | | g__Agathobacter, g__Anaerostipes, g__Blautia_A, g__Phocaeicola | UNKNOWN | Host-Associated: Human intestine |
| putative_HGT_18 | | | | g__Agathobacter, g__Bacteroides, g__Bariatricus, g__CAG-127, g__Coprococcus, g__Dorea_A, g__Lachnospira, g__Phocaeicola, g__Roseburia, g__TF01-11 | UNKNOWN | Host-Associated: Human intestine |
| putative_HGT_19 | | | | g__Agathobacter, g__Bacteroides, g__Blautia_A, g__Phocaeicola, g__RUG115, g__Roseburia, g__TF01-11 | UNKNOWN | Anthropogenic: Waste Handling Host-Associated: Human intestine |
| putative_HGT_20 | | | | g__Agathobacter, g__Bacteroides, g__Lachnospira, g__Phocaeicola, g__Roseburia | d_Cache_1 | Host-Associated: Human intestine |
| putative_HGT_21 | | | | g__Agathobacter, g__Bacteroides, g__Phocaeicola, g__Roseburia | | Host-Associated: Human intestine |
| putative_HGT_22 | | | | g__Agathobacter, g__Bacteroides, g__Prevotella, g__RUG115, g__Roseburia, g__UBA11774 | UNKNOWN | Host-Associated: Human intestine Host-Associated: Human skin |

## Table 2.3: Putative Horizontal Gene Transfer Events across globalCR.

| cluster name | Phyla | Order | Family | Genus | LBD | habitat |
|---|---|---|---|---|---|---|
| putative_HGT_23 | | | | g__Agathobacter, g__Blautia_A, g__CAG-127, g__Coprococcus, g__Mediterraneibacter, g__Phocaeicola, g__Roseburia | UNKNOWN | Anthropogenic: Waste Handling Host-Associated: Human intestine |
| putative_HGT_24 | | | | g__Agathobacter, g__Coprococcus, g__Eubacterium_F, g__KLE1615, g__Lachnospira, g__Mediterraneibacter, g__Phocaeicola | 4HB_MCP_1 | Host-Associated: Human intestine |
| putative_HGT_25 | | | | g__Bacteroides, g__Lachnospira | UNKNOWN | Host-Associated: Human intestine |
| putative_HGT_26 | | | f__Bacteroidaceae, f__Lachnospiraceae, f__Rikenellaceae | g__43-108, g__Acetatifactor, g__Agathobacter, g__Alistipes, g__Mediterraneibacter, g__RUG115, g__Roseburia | 4HB_MCP_1 | Host-Associated: Human intestine |
| putative_HGT_27 | | | f__Bacteroidaceae, f__Lachnospiraceae, f__Tannerellaceae | g__Agathobacter, g__Anaerostipes, g__Bacteroides, g__Blautia_A, g__CAG-127, g__Coprococcus, g__Eubacterium_F, g__KLE1615, g__Lachnospira, g__Mediterraneibacter, g__Parabacteroides, g__Phocaeicola, g__RUG115, g__Roseburia, g__Schaedlerella, g__TF01-11 | | Anthropogenic: Industrial and Built Environments Anthropogenic: Waste Handling Host-Associated: Human intestine |
| putative_HGT_28 | | | f__Bacteroidaceae, f__Lachnospiraceae, f__UBA932 | g__, g__Agathobacter, g__Bariatricus, g__Blautia_A, g__Coprococcus_A, g__Eubacterium_J, g__Prevotella, g__RC9, g__UBA629, g__Weimerbacter | sCache_2 | Host-Associated: Animal intestine/rumen |
| putative_HGT_29 | | o__Bacteroidales, o__Lachnospirales, o__Oscillospirales | f__Acutalibacteraceae, f__Bacteroidaceae, f__Lachnospiraceae | g__Acetatifactor, g__Agathobacter, g__Blautia_A, g__Lachnospira, g__Phocaeicola, g__Pseudoruminococcus, g__RUG115, g__UBA11774 | UNKNOWN | Anthropogenic: Waste Handling Host-Associated: Human intestine |

**Table 2.3: Putative Horizontal Gene Transfer Events across globalCR.**

| cluster name | Phyla | Order | Family | Genus | LBD | habitat |
|---|---|---|---|---|---|---|
| putative_HGT_30 | | o__Bacteroidales, o__Lachnospirales, o__Oscillospirales | f__Bacteroidaceae, f__Lachnospiraceae, f__Rikenellaceae, f__Ruminococcaceae | g__1XD42-69, g__AM51-8, g__Acetatifactor, g__Agathobacter, g__Alistipes, g__Anaerostipes, g__Bacteroides, g__Bacteroides_F, g__Bariatricus, g__Blautia, g__Blautia_A, g__Butyribacter, g__Butyrivibrio_A, g__CAG-127, g__CAG-882, g__CAG-95, g__COE1, g__Dorea_A, g__Eubacterium_F, g__Faecalimonas, g__KLE1615, g__Lachnoclostridium_B, g__Lachnospira, g__Mediterraneibacter, g__Phocaeicola, g__Prevotella, g__RUG115, g__Roseburia, g__Ruminococcus_D, g__TF01-11 | 4HB_MCP_1 | Host-Associated: Human intestine |
| putative_HGT_31 | | o__Bacteroidales, o__Lachnospirales, o__Peptostreptococcales | f__Anaerovoracaceae, f__Bacteroidaceae, f__Lachnospiraceae | g__, g__Bacteroides, g__Blautia, g__CAG-95, g__Clostridium_Q, g__Enterocloster, g__Muricomes, g__SFDP01, g__Schaedlerella, g__UBA11774 | UNKNOWN | Host-Associated: Animal intestine/rumen Host-Associated: Human intestine |
| putative_HGT_32 | | o__Bacteroidales, o__Lachnospirales, o__UBA1212 | f__Lachnospiraceae, f__Muribaculaceae, f__Paludibacteraceae, f__UBA1255 | g__, g__Faecalimonas, g__HGM12587, g__RGIG4034, g__UBA1181, g__UBA3282, g__UMGS1537 | dCache_1 | Host-Associated: Animal intestine/rumen |
| putative_HGT_33 | p__Bacteroidota, p__Firmicutes_A, p__Firmicutes_C | o__Bacteroidales, o__Lachnospirales, o__Selenomonadales | f__Bacteroidaceae, f__Lachnospiraceae, f__Selenomonadaceae | g__Agathobacter, g__Anaerostipes, g__Bacteroides, g__Blautia_A, g__Lachnospira, g__Megamonas, g__Phocaeicola, g__RUG115, g__Roseburia | UNKNOWN | Anthropogenic: Waste Handling Host-Associated: Human intestine |
| putative_HGT_34 | | o__Bacteroidales, o__Lachnospirales, o__Veillonellales | f__Bacteroidaceae, f__Dialisteraceae, f__Lachnospiraceae | g__Agathobacter, g__Bacteroides, g__Butyrivibrio_A, g__Dialister, g__Phocaeicola | | Host-Associated: Animal intestine/rumen Host-Associated: Human intestine |
| putative_HGT_35 | | | | | sCache_3_3 | Host-Associated: Human intestine |
| putative_HGT_36 | p__Bacteroidota, p__Proteobacteria | o__Flavobacteriales, o__Rhodobacterales | f__Crocinitomicaceae, f__Flavobacteriaceae, f__Rhodobacteraceae, f__Salibacteraceae | g__MED-G52, g__MS024-2A, g__Planktomarina, g__UBA6770, g__UBA952 | UNKNOWN | Aquatic: Freshwater Aquatic: Saltwater Terrestrial: Agricultural and Wet Environments |

## Table 2.3: Putative Horizontal Gene Transfer Events across globalCR.

| cluster name | Phyla | Order | Family | Genus | LBD | habitat |
|---|---|---|---|---|---|---|
| putative_HGT_37 | p__Desulfobacterota_B, p__Proteobacteria | o__Binatales, o__Rhizobiales | f__Beijerinckiaceae, f__Binataceae, f__Xanthobacteraceae | g__, g__BOG-931, g__Binatus, g__Roseiarcus | PilZ | Aquatic: Freshwater Terrestrial: Agricultural and Wet Environments Terrestrial: Soil and Dry Environments |
| putative_HGT_38 | p__Firmicutes, p__Firmicutes_A | o__Erysipelotrichales, o__Lachnospirales | f__Erysipelatoclostridiaceae, f__Erysipelotrichaceae, f__Lachnospiraceae | g__, g__ASTD01, g__Catenibacterium, g__UBA2868 | dCache_1 | Host-Associated: Animal intestine/rumen |
| putative_HGT_39 | p__Firmicutes_A, p__Firmicutes_B | o__Lutisporales, o__Peptostreptococcales, o__Syntrophomonadales | f__JAAYPU01, f__Lutisporaceae, f__Syntrophomonadaceae | g__BRH-c25, g__JAAYBS01, g__JAAYPU01, g__UBA1368, g__UBA4844 | dCache_2 | Anthropogenic: Waste Handling |

## 2.5 An open space of bacterial chemosensors

Using the LBD (ligand-binding domain) subfamily classification, we analyzed the variability and biodiversity of CRs within the globaCR dataset. Of the 42,375 clusters inferred, a significant proportion were clearly associated with known Pfam domains, with the majority classified as well-characterized protein-sensing domains, such as 4HB_MCP_1 (223,872 sequences) and dCache_1 (162,428 sequences). Interestingly, many other CRs were grouped into LBD clusters labeled as unknown, suggesting the presence of potentially novel sensing domains.

To further investigate the biodiversity of both known and unknown sensing domains in nature, we analyzed the number of LBD clusters inferred per CR sequence. The cumulative curve of clusters per observed CR genes (Figure 2.5) shows that well-known LBD-like Pfam domains, such as 4HB_MCP_1 and dCache_1, exhibited a nearly flat saturated curve, suggesting that a discovery plateau may have been reached, with few new subfamilies left to be identified in nature (Figure 2.5). In contrast, the unknown LBD category showed a strikingly different trend, indicating a largely unexplored variability within unknown CR sensor domains. This immense variability was predominantly attributed to LBDs tagged as extracellular (Figure 2.5b), suggesting that this open space in bacterial and archaeal chemosensors is likely shaped by adaptation to different habitats and environmental conditions.

## 2.6 Free living microbial communities are enriched chemosensitive species

Although chemosensitive microorganisms are found throughout nature, their distribution, prevalence, and biodiversity across various habitats remain largely understudied. To explore broad environmental trends across 19 biomes, we used ecological annotations in our global CR catalog to calculate both the proportion of chemosensitive species in each habitat (Figure 2.6a) and the average number of CRs present per genome (Figure 2.6b).

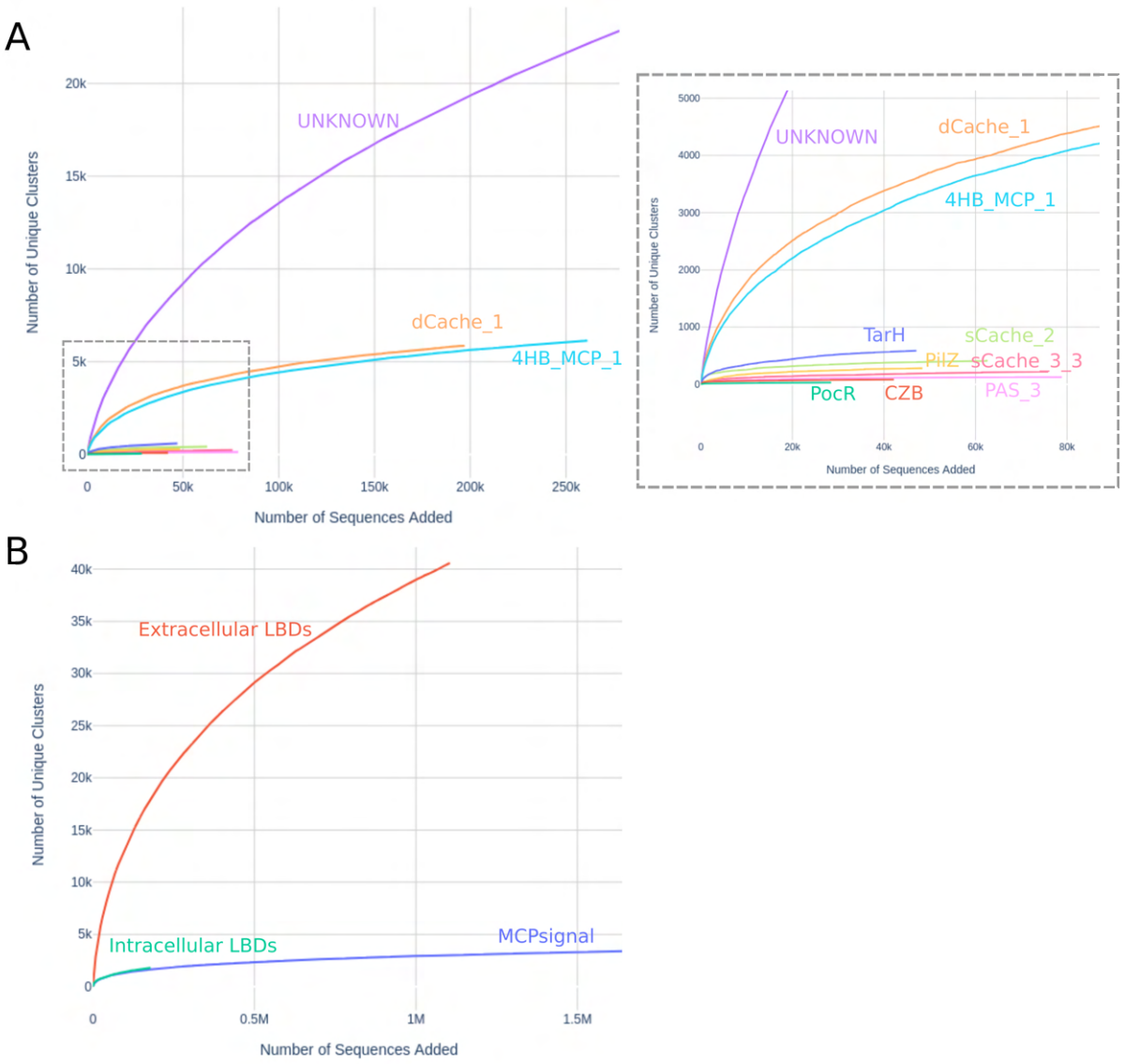



**Figure 2.5: Cumulative curve of sensing domain homology clusters across globalCR.** The curves show the cumulative number of subfamilies or clusters for each domain as randomly shuffled sequences are added. A) Subfamilies are grouped by their dominant Pfam domain (comprising more than half of the sequences) or classified as "UNKNOWN" if no Pfam domain is detected. B) Sensing domain subfamilies are categorized as extracellular or intracellular (without predicted transmembrane domains). MCPsignal subfamilies are also included.

Our findings reveal that host-associated microbial communities—particularly those in physically confined niches such as the gut, mouth, or rumen—contain a lower proportion of chemosensitive microorganisms. In contrast, free-living communities, such as those in air, water, plants, and soil, harbor a significantly higher fraction of chemosensitive species. Interestingly, host-associated but non-confined microbiota, like those on human and animal skin, exhibited a higher proportion of chemosensitive bacteria.

We observed a similar trend in the average number of CRs per species across different biomes. In general, free-living environments displayed a greater CR count per species than confined ecological niches. Notably, bacterial species found in surface-associated environments—such as the plant associated, human skin, and built environments (e.g., hospital and urban surfaces)—were significantly more chemosensitive (indicated by a higher number of CRs per genome) than those from other habitats.

## 2.7 Extracellular chemosensing domains are an adaptative value in bacterial evolution

The diversity and rapid evolution of sensing domains, particularly those located extracellularly, suggest a high adaptive significance for prokaryotes, likely driven by ecological pressures that require detecting and responding to a dynamic array of environmental signals. To explore potential links between microbial sensing capabilities and habitat distribution, we quantified the relative abundance of CR genes across 99,146 environmental samples in the SPIRE dataset (see Methods). Specifically, we assessed habitat prevalence of CRs based on their sensing domains—clustering by ligand-binding domain (LBD) for sensing and by MCPsignal domain for transduction—differentiating between extracellular and intracellular sensing domains.

The multidimensional analysis of CR relative abundances across these groupings (Figure 2.7) reveals that CRs clustered by their signal transduction domain (MCPsignal) display the lowest ecological specificity (Figure 2.7), followed by those with intracellular sensing domains (Figure 2.7b). In striking contrast, clustering based on extracellular sensor domains shows a strong ecological signal, effectively distinguishing between different habitats (Figure 2.7c). Interestingly, the clusters

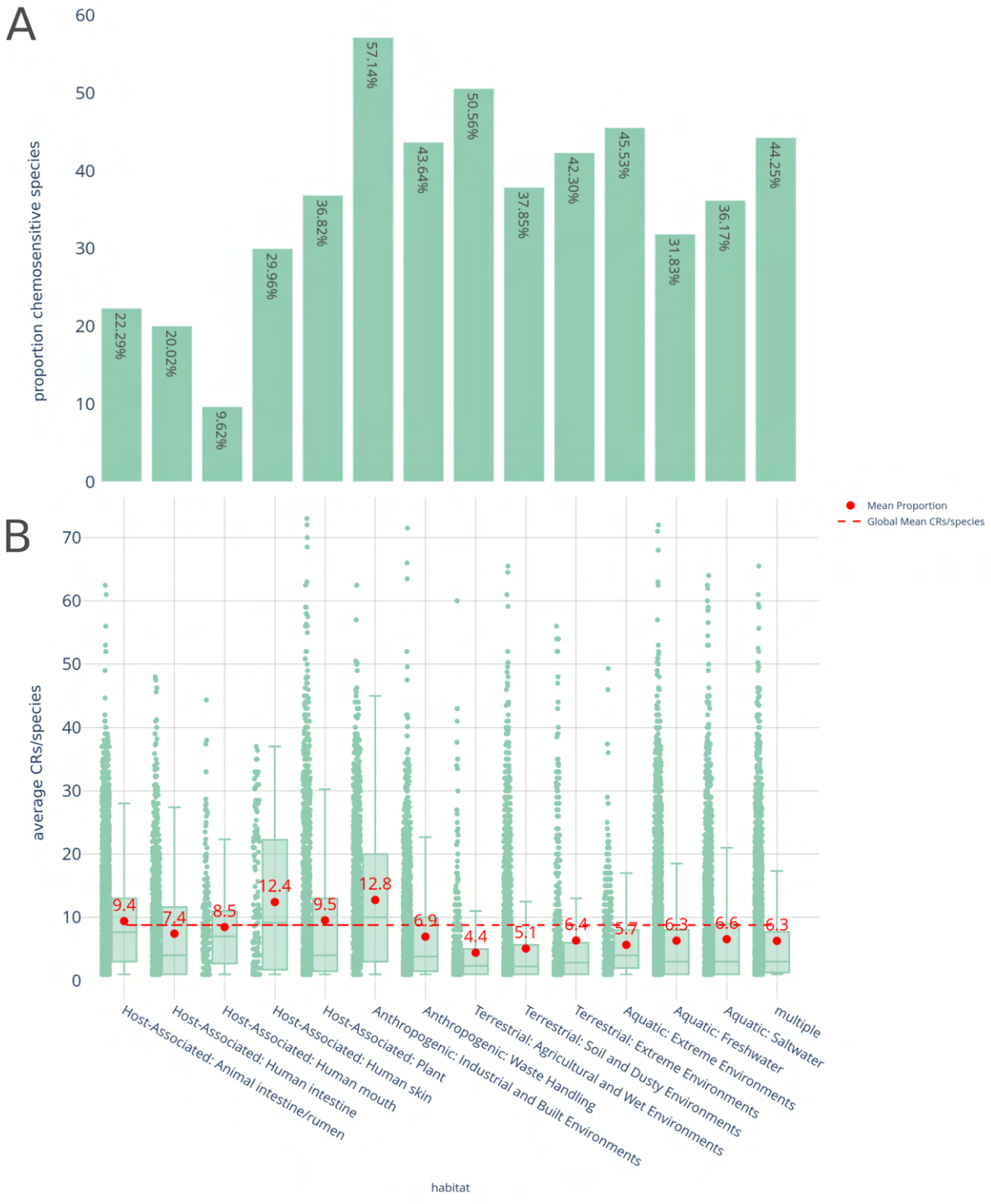



**Figure 2.6: Distribution of chemoperception across biomes.** A) The percentage of chemosensitive species is calculated across clusters, counting the proportion of SPIRE species clusters with at least one CRs in the MAGs out of the total number of species clusters annotated in each habitat. B) Box plots of the mean number of CRs per species cluster. Each species cluster number of CRs is the average of all the MAGs within the species cluster. The mean number is plotted in red dots for each habitat, and the global mean is plotted in the red dash line (8.8 CRs/species).

derived from extracellular LBD analysis resemble the habitat separations observed in taxonomic analyses, such as UMAPs based on species composition inferred with the mOTUs tool[211]. This finding is notable, as MCPsignal domains and intracellular sensors are generally more conserved and would typically be expected to retain a stronger phylogenetic signal than the highly variable and diverse extracellular LBDs. This result is further validated by Random Forest classifiers trained on the same datasets, which demonstrate that extracellular LBDs have a significantly higher predictive value for ecological conditions (Supplementary Figure 2.2).

The observation that only extracellular sensor domains show clear ecological associations—while the signal transduction and intracellular sensor domains of the same CR genes do not—suggests that these associations stem from the adaptive value of CR chemosensors, rather than from neutral processes like genetic drift or historical colonization patterns by closely related species.

## 2.8 Biome specific sensing capabilities

To further examine the relationships between clusters of sensing capabilities and their correspondence to different ecological environments, we applied hierarchical clustering to the HDBSCAN clusters. This clustering was based on Bray-Curtis dissimilarities between samples (see Methods), enabling us to visualize relationships on a broader scale and address UMAP's limitations in capturing global patterns.

The resulting dendrogram, displayed in Figure 2.8, includes a bar plot illustrating the relative abundance of samples from each ecological environment. The clusters group samples from similar ecologies, with some clusters corresponding to specific biomes, such as the human gut (e.g., clusters 4 and 20), the rumen (cluster 5), the human mouth (cluster 26), and saltwater environments (cluster 29). Other clusters represent environments with shared sensing capabilities, including aquatic and wastewater environments (cluster 35) or a mix of terrestrial, plant-associated, and aquatic samples (clusters 36 and 40).

Building on this framework, we investigated whether certain sensor subfamilies are prevalent and specific to different environments. We sought environmental CR markers within each cluster

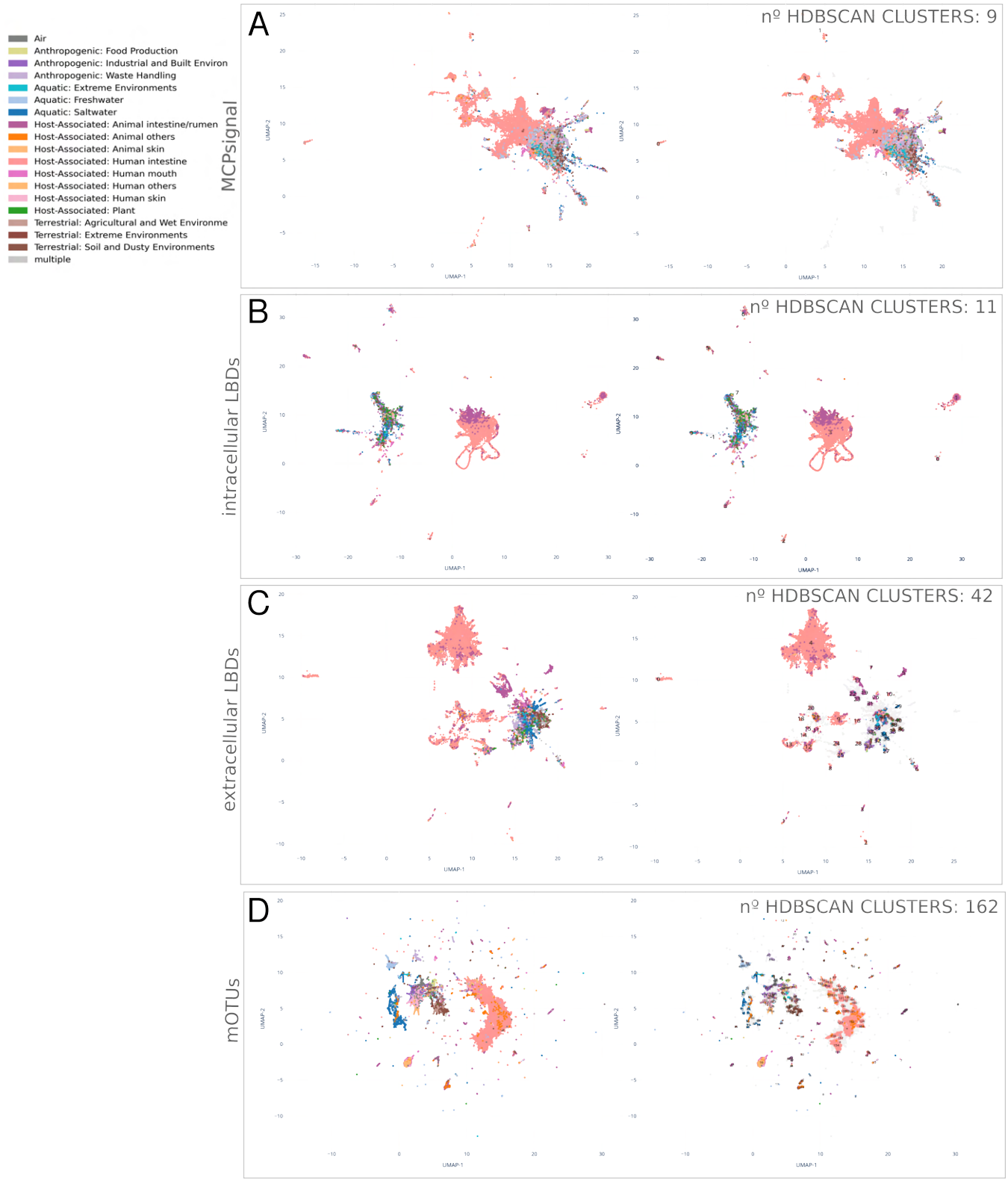



**Figure 2.7: Projection of the ecological sub-structure of datasets applying UMAP dimensional reduction.** UMAP projections of the relative abundance matrix for each domain subfamily, colored by ecological label. The full UMAP projection is shown on the left, while only HDBSCAN clusters are shown on the right (with noise uncolored). A) MCPsignal domains, B) Intracellular sensing domains, C) Extracellular sensing domains, D) Taxonomic structure based on mOTUs[211] relative abundance.

identified in our multidimensional analysis, identifying sensing subfamilies present in each cluster and calculating their prevalence. We defined prevalent subfamilies as those found in at least 50% of the samples within a cluster, resulting in 192 identified prevalent sensing subfamilies. For each of those subfamilies, we calculated its specificity to each biome, representing the percentage of samples within each biome containing the subfamily.

In Figure 2.7b, we show the mean specificity of the sensing subfamilies across biomes within the HDBSCAN clusters. These results reveal a strong alignment between the specificity of sensing subfamilies and the ecological origins of the samples. This finding indicates that prevalent sensing subfamilies are not only abundant but also highly specific to particular environments, suggesting that these sensing capabilities are closely linked to the ecological roles of microbial communities.

We identified several sensing domains that are highly prevalent and specific within large HDBSCAN clusters, many of which are linked to particular environments. Notable examples include the CZB, NIT, and several UNKNOWN subfamilies, all specific to saltwater environments. Additionally, certain sensing subfamilies prevalent in the order *Gemmatiomonadales*, associated with soil, exhibited GAF sensing domains, while the rumen environment showed a high abundance of *Treponematales* with prevalent dCache_1 sensing domains.

In the human gut, we observed a PilZ-like sensing domain with remarkable prevalence and specificity, present in 62% of the 20,000 human gut samples screened and scarcely detected in other microbial environments.

To further investigate the unique features of samples grouped by sensing capabilities in the HDBSCAN clusters, we applied a more detailed micro-ontology approach to assess sensing domain specificity (Supplementary Figure 2.4). Preliminary results revealed that clusters 8 and 16 were specific to infant or premature birth gut samples, while cluster 26 was enriched in samples from the child gut. Cluster 29, associated with aquatic saltwater samples, showed specificity for pelagic environments. Further analysis of these examples and their sensing domains are ongoing.

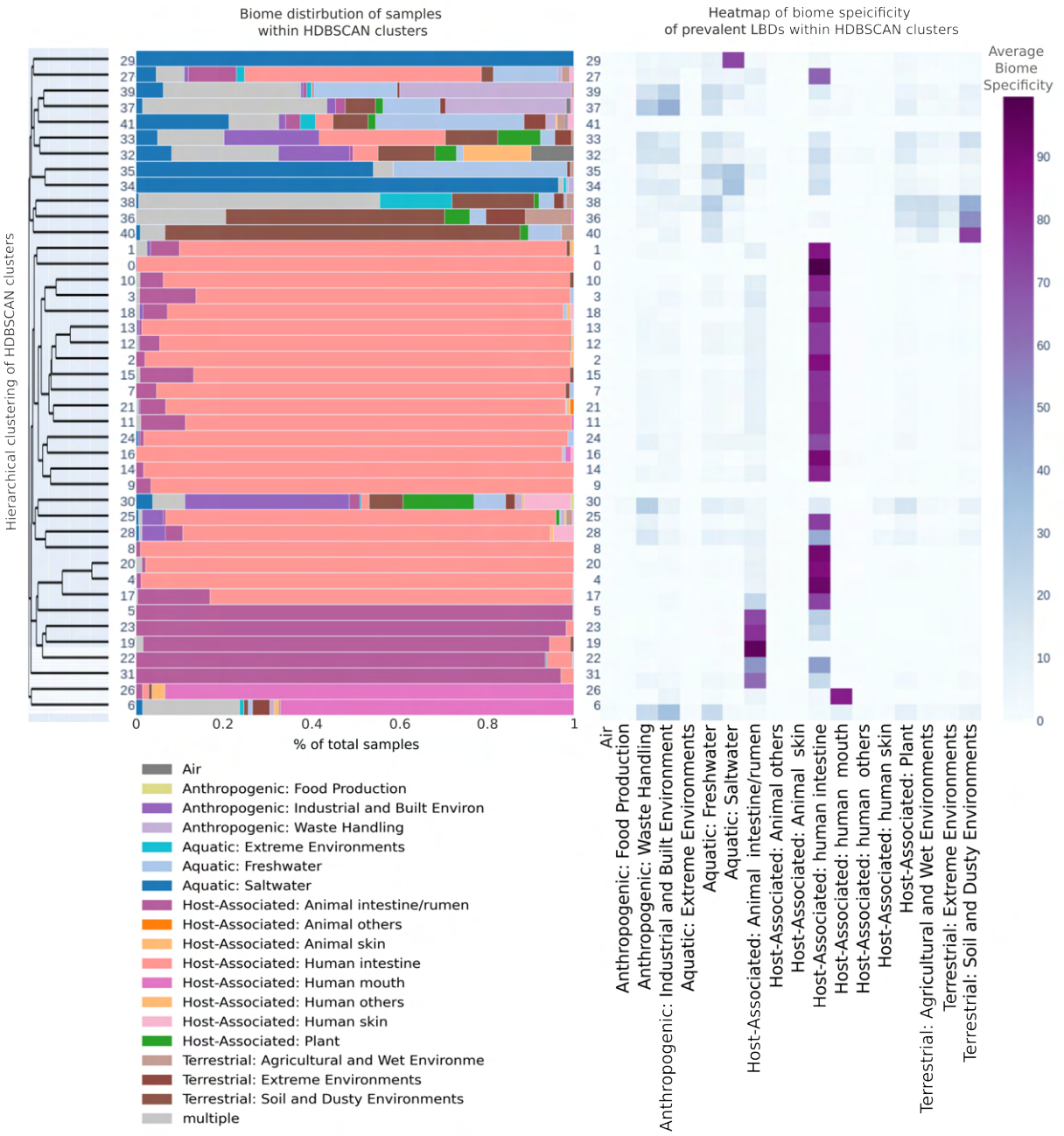



**Figure 2.8: Hierarchical clustering of samples grouped by their sensing capabilities.** Clusters identified by HDBSCAN algorithm in the UMAP visualization of samples grouped by extracellular sensing domain are shown in each row. The hierarchical clustering group them by similarity. The barplot represents the relative abundance of the ecological origin (biome) of the samples within each cluster. The heatmap panel is showing the biome specificity of the prevalent sensing subdomains subfamilies for each HDBSCAN cluster (row) subfamilies.

## 2.9 Discussion

In this study, we performed a comprehensive analysis of chemoreceptors (CRs) by integrating all genomic and metagenomic data available to date. While previous efforts[137] have paved the way for the integration of environmental data into the study of chemosensory pathways, a vast amount of data generated over the last decades has remained unexplored. In our study, we included 1.6 million CR sequences from more than 800,000 genomes of uncultivated microorganisms and 40,000 representative reference species. To our knowledge, this is the most comprehensive survey of CRs to date and the first to allow the study of CR evolution, ecology and function on a global planetary scale.

We found that the landscape of chemosensory microorganisms is extremely large and variable across different environments, including hundreds of new species yet to be classified. We also identified many uncultured bacterial lineages that, despite possessing a large number of CRs, have been systematically excluded from chemosensing-oriented genomic surveys. This systematic exclusion may be due to the potential risk of performing genome-scale analyses on metagenome-based assemblies, as the quality of metagenomics data might distort results. However, our analyses demonstrate that current metagenomic data, and in particular high- and medium-quality metagenome-based genome assemblies (MAGs), are sufficiently robust to reliably study chemoreceptor mechanisms in unculturable microorganisms.

For instance, this study allowed us to characterize the chemosensing capabilities of the fully uncultured *Moduliflexota* bacterial phyla; which, strikingly, represent a group of microorganisms with the highest number of CRs per genome known to date. By exploring the functional repertoire of all sensing domains present in *Moduliflexota* CRs, we also found a large variability and a high rate of unclassifiable ligand binding domains. We hypothesize that such a large and variable set of CRs might be the result of adaptation of *Modulliflexota* species to diverse extreme environments like hypersaline soda lake sediments[259], which are described as responsible for bulking in wastewater treatment[260].

Moreover, our systematic analysis of CR sequences, which took into account their multidomain nature, revealed radically different evolutionary and ecological patterns associated with their various domains.

### Evolutionary history of CRs revealed by the MCPsignal domain

The signal transduction domain (MCPsignal), which is present in all CRs, displayed a highly conserved pattern across all organisms and a strong vertical phylogenetic signal. Despite their great conservation, structural variations within the MCPsignal domain, known as heptads, have been used to classify CRs into distinct classes and hypothesize about their potential different functionality[38]. However, very little was known about the evolutionary history of such heptad classes and their potential subtypes and diversification.

By building a massive phylogenetic tree out of 400,000 unique MCPsignal sequences, we were able to reconstruct the evolutionary steps that gave rise to each heptad class and trace their emergence. Our data showed that heptad classes originated, primarily, by duplication events, with some classes emerging specifically in particular clades.

Interestingly, the distribution of heptad-based clades in the resulting phylogenetic tree correlated with the presence/absence matrix of different CheA gene subtypes, suggesting potential functional associations between heptad classes and different signal transduction systems. Such links have been documented in the past for certain heptad types[18], [261], but no comprehensive analysis was undertaken to date. Our phylogenetic and functional profiling analyses (Figure 2.2) confirmed the association between CheA subtypes and heptad classes, also revealing new potential functional links.

For example, the 48H heptad class, exclusively identified in *Treponema* species, showed a strong association with the F2 CheA systems. While this direct connection had not been previously reported, it aligns with earlier studies that highlighted the specialized arrays formed within F2 systems[262]. Similar findings were observed in *Campylobacter* species, which have been widely studied as a model lineage for exploring the plasticity of chemosensory systems[263]. Our analysis revealed *Campylobacter*-specific 28H-like chemoreceptors are associated with F3 systems. Lastly,

we found that the *Desulfobacterota*-specific CheA F4 system is strongly linked to a subset of 40H-like CRs in such lineage.

Moreover, our phylogenetic analysis provided insights into the origins of each heptad class. Based on its widespread prevalence and functional diversity, the 44H heptad class appears to be the closest to the ancestral MCPsignal domain. Not only is the 44H class present in the majority of both cultured and uncultured prokaryotic lineages, but it also correlates strongly with the presence of the CheA F1 system (Figure 2.3b), which has been proposed as the most generic and primary chemosensory system[264]. Assuming this origin, Figure 2.9 presents a schematic representation of the proposed evolutionary history of chemoreceptors (CRs) based on their MCPsignal domains. The diagram illustrates a scenario in which an ancestral 44H-like CR gene underwent a duplication event, creating two main lineages, where both 44H-like copies were retained. This suggests that two orthologous groups might exist for the 44H-like CR genes, with potential functional sub-specialization between them. Each of the two 44H-like lineages that emerged following the proposed ancestral duplication is thought to have given rise to all other heptad classes through successive, and sometimes lineage-specific, duplication events. One branch diverged into the 34H, 36H, and 38H classes, and later to the 42H class (specific to *Desulfobacterales*) and the 58H class (found in *Magnetococcales*). The other branch led to the 40H group, followed by the emergence of 28H (mostly in class *Bacilli*), 48H (specific to *Tremonema*) and 64H (found in *Geobacterales*) groups. Importantly, our analysis revealed that only various orthologs are likely to exist within each heptad class, resulting in the formation of distinct subclasses. These subclasses should be considered when investigating the functional and phylogenetic relationships of chemoreceptors.

### Ecological signal revealed by Ligand Bindings Domains (LBD)

In contrast to the weak ecological patterns observed for the MCPsignal domain, the analysis of ligand-binding domain (LBD) clusters revealed strong environmental associations. This chapter builds on and extends the findings of Sanchis-López *et al.* [215] (Chapter 1), where we demonstrated the ecological significance of chemosensory systems by linking CR distribution to environmental lifestyles.

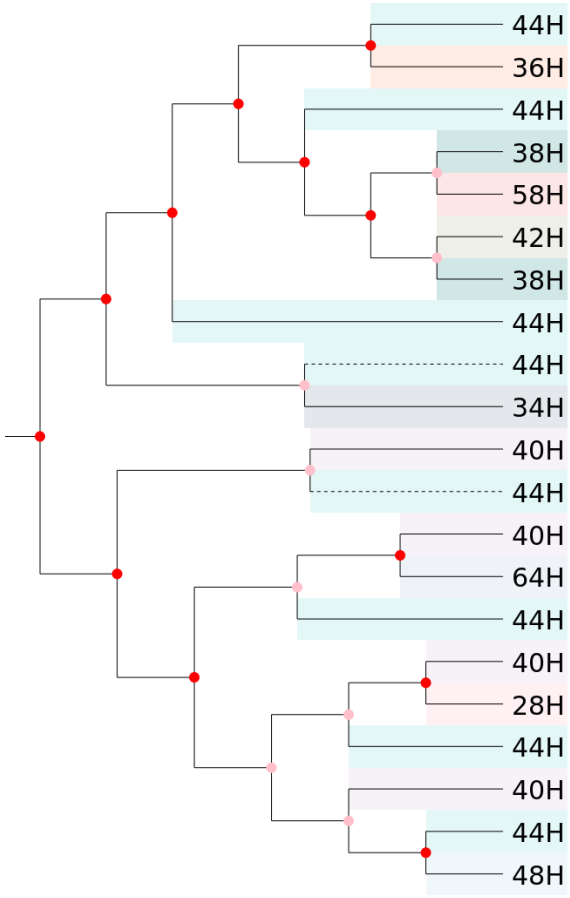



**Figure 2.9: Proposed Origins and Evolutionary Dynamics of Chemoreceptors (CRs).** Heptads not found as compact groups (harboring at least 10% of sequences of that subtype) are not included in the scheme. Detected duplication events are marked in red, while potential undetected events, possibly due to gene loss, specialization, or horizontal gene transfer (HGT), are indicated in pink. Dashed branches suggest proposed gene loss.

We found that, while the MCPsignal domain sequence is highly conserved across chemoreceptors (CRs) and microorganisms, LBD sequences exhibit huge variability, often presenting little to no homology with one another. In fact, a deep-homology clustering analysis of all observed LBD sequences, including both genomic and metagenomic data, identified more than 40,000 distinct clusters, underscoring the immense biodiversity of microbial sensing capabilities.

Interestingly, the diversity of LBD sequences differed significantly between well-characterized and previously unknown sensing domains. Known LBD types, such as 4HB_MCP_1 and dCache_1, showed a plateau in observed sequence variability. In contrast, clusters of unknown LBDs—likely representing novel sensing domain types—revealed a clearly open diversity landscape. This trend was particularly pronounced for unknown LBDs identified as extracellular, further supporting the idea of an untapped reservoir of environmental sensors in nature.

We observed that dynamic environments, such as industrial and urban areas, harbor species with a higher abundance of chemoreceptors (CRs) compared to stable, nutrient-rich environments like the human gut. Additionally, terrestrial, aquatic, and plant-associated biomes displayed a greater proportion of chemosensitive species, indicating that environmental pressures play a significant role in driving CR evolution.

In fact, using dimensionality reduction and clustering methods, we linked sample clusters and their sensing capabilities, finding that extracellular LBDs exhibited strong ecological associations that distinguished between biomes. In contrast, MCPsignal domains and intracellular sensors showed no such patterns. This finding strongly suggests an adaptive advantage for extracellular sensing domains, indicating that selective pressures—rather than phylogeny—shape the repertoire of sensing capabilities in organisms. However, further analyses are needed to fully validate this hypothesis.

Regarding potential limitations of our study, it is important to note that clustering accuracy might be influenced by sequencing depth and the limitations of ecological terms and annotations, which could result in misclassifications or oversimplifications, particularly in poorly characterized environments. For instance, the mean number of CRs per Gbp is approximately six (Supplementary Figure 2.3), translating to around 48 CRs in an 8 Gbp sample. This number may be insufficient to capture the

complete CR repertoire, as some species possess over 70 CRs. While this limitation likely does not affect the broader patterns observed, it underscores the need to study low-abundance genes to ensure that CR profiles are comprehensively represented and accurately linked to their ecological niches.

Despite potential confounders, our methodology and results demonstrated that the LBD repertoire within microbial communities is closely linked to their respective habitats, particularly through certain prevalent and highly specific LBD types. For example, we identified a cluster of PilZ sensing domains as the most prevalent LBD subfamily in the catalog, appearing in 62% of all 22,000 human intestinal samples but being almost entirely absent in other biomes. This remarkable specificity suggests that the PilZ domain is highly adapted to the human gut environment. The PilZ domain is indeed predicted to sense c-di-GMP, a second messenger in prokaryotes that regulates various cellular processes, including motility, extracellular polysaccharide production, and biofilm formation[265].

In conclusion, this chapter serves as a foundational step toward understanding the global diversity and evolution of chemoreceptors. It opens up a wealth of research opportunities, from in-depth investigations into newly identified domains to comprehensive studies on the evolution and adaptation of chemosensory systems across diverse environments.

# Chapter 3

## Unveiling the Repertoire of Microbial Chemoreceptor Genes in the Rare Biosphere

In this Chapter, we address the challenge of deciphering the low abundance microorganisms functionality by developing a novel target capture sequencing approach that enables the exploration of CRs genes in the rare microbial biosphere. Capture and target enrichment methods are central to genomic research, allowing the selective isolation of specific genomic regions from complex samples. This technique uses custom-designed oligonucleotide probes that hybridize to target sequences, enabling targeted sequencing and in-depth analysis of regions of interest. Compared to whole genome sequencing, this approach is cost-effective, increases the depth of coverage, and reduces the complexity of data analysis. It has been widely applied in model organisms for genetic studies and functional genomics[266], [267], and rarely in metagenomics [268], where the tolerance of the method to capture divergent sequences is particularly useful for covering a wide range of microbial variants. However, challenges remain, including the complexity of probe design, potential capture biases, and technology limitations to cover wide phylogenetic ranges and capture and enrich highly variable genomic regions such as CRs sensing domains.

## 3.1 A comprehensive catalog of CRs in the global microbiome

To investigate the biodiversity of CRs across different environments, we analyzed a recent multi-habitat metagenomic dataset (GMGC v1[186]) based on 84,029 prokaryotic reference genomes and over 2.3 billion environmental genes assembled from 13,174 metagenomes across 14 biomes*. We followed the approach described in Sanchis-López *et al.*[215] to identify chemosensitive proteins via highly conserved MCPsignal domains (Pfam:PF00015)[213], mining the GMGC resource and obtaining a total of 149,326 sequences containing the MCPsignal domain, 91,776 of which were full length genes. On average, 21 CR genes were detected per sample, depending on sequencing depth (4.9 assembled chemoreceptor genes per sequenced Gbp; Supplementary Figure 3.1).

Soil and built-environment ecosystems were the largest contributors of unique CRs to the catalog (Figure 3.1a), despite representing a low number of samples in the GMGC dataset (Figure 3.1b). In fact, the rate of unique CR discovery per sequencing depth unit (Figure 3.1c) was clearly lower in

*Although the SPIRE[187] resource and the globalCR catalog previously described would have been more appropriate for this analysis, it should be noted that SPIRE has only recently been released and was not available at the time of this project.

less biodiverse ecosystems such as the human gut microbiome (1.47 CRs/Gbp) than in soil (18.05 CRs/Gbp) and built-environment (28.41 CRs/Gbp).

Then, each CR gene was analyzed to identify potential sensing domains. Since CR sensing regions typically consist of LBDs located near the N-terminus of the sequence, we identified putative LBDs using three different strategies: (i) By matching any known Pfam domain except MCPsignal or HAMP, (ii) by checking for the presence of two flanking transmembrane regions, and (iii) extracting the sequence domain between the N-terminus and a single transmembrane (TM) region. When no significant Pfam matches were found, LBD sequences were labeled as “unknown”. This approach yielded putative LBDs for 80,298 CR genes (54,9% of the total). From an ecological perspective, the distribution of LBD types varied across habitats, with the unknown LBD category being prevalent in all biomes (33% on average). Soil was the habitat with the highest diversity of LBD types (289 different domains), followed by the human gut (238), and the marine environment (211).

To the best of our knowledge, this metagenomic CR catalog (metaCR) represents the largest repository of CR genes described to date and the first to estimate the biodiversity of chemoreceptors from most abundant uncultured organisms across habitats.

## 3.2 Target capture sequencing boosts CR gene detection across habitats

To facilitate the exploration of the chemoreceptor repertoire in low-abundance uncultured taxa, we developed a novel, PCR-free, long-read targeted metagenomics approach (Figure 3.2). The method focuses on capturing and enriching long DNA fragments that contain the highly conserved MCPsignal domain.

Using our metaCR catalog as a reference, we designed a custom panel of 1.3 million probes that covers all the sequence variability observed in the MCPsignal domain of bacteria and archaea. To this end, we first extracted the MCPsignal domain sequence from all 149,326 CRs in the metaCR catalog and generated 80 nucleotides (nts) long probes covering each MCPsignal sequence region observed, using a sliding window of 20 nts. This yielded a total of 3,429,805 probes,

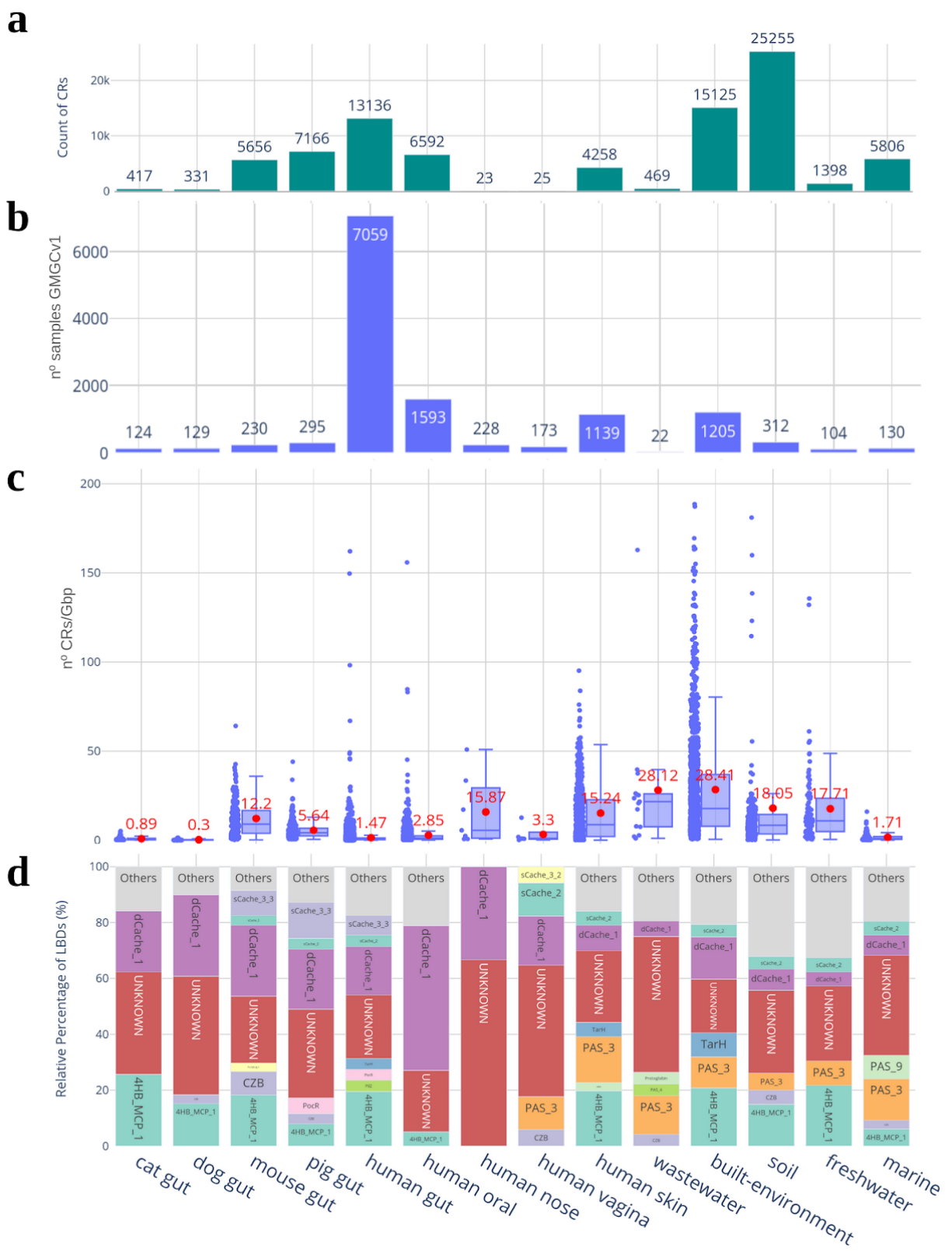



**Figure 3.1: MetaCR distribution and diversity in the sensing domain profile across habitats.** a) Total count of CR genes retrieved from each habitat; b) Total number of samples of the source catalog Global Microbial Gene Catalog v1 (GMGCv1[186]) from each habitat; c) Normalized distribution of CRs per Gpb sequenced in the samples of origin per habitat with mean values in red; outliers are not shown (>200 CRs/Gbp); d) Sensing profiles for each habitat in terms of ligand binding domains (LBDs) showing the relative abundance of the most abundant (>3%) LBDs in each habitat. Low-abundance LBDs are grouped in the “Others” category.

which were clustered at 80% identity to reduce unnecessary redundancy, resulting in 1,323,163 cluster-representative probes.

To evaluate the efficiency of this optimized capture-enrichment kit, we compared it to results obtained using conventional shotgun metagenomic sequencing of microbiome communities from different ecosystems, including rhizosphere, phyllosphere, bulk soil, lake water, human gut, and animal rumen. Following standard procedures in metagenomic sequencing, DNA was extracted from each sample and sequenced using short-read Illumina PE150 technology at a depth of 8 Gbp per sample. In parallel, DNA extracted from the same samples was fragmented at an average length of 1kb and multiplexed into a pooled library (Pool-1), which was subjected to a capture enrichment protocol using our customized hybridization panel, followed by long-read sequencing with Oxford Nanopore Technologies (ONT). As standard capture-enrichment protocols are incompatible with ONT multiplexing kits[269], we developed a custom ad hoc protocol (inspired on Bethune *et al.* 2019[228]) that combines illumina-based indexing as well as nanopore library-preparation and sequencing (see Methods section for details). This facilitated multiplexed long-read target capture sequencing of CRs.

Despite differences among environments (Supplementary Table 3.1), the number of unique CRs identified in capture-based experiments increased manifold compared to bulk metagenomics data (Figures 3.3a and 3.3b), ranging from 3-fold (lake) to 1119-fold (phyllosphere) increases.

## 3.3 Identification of CR sensing domains from rare taxa

To assess whether target capture sequencing of CRs could be further optimized to boost CR detection in rare uncultured organisms and elucidate the domain architecture of the CR repertoire, we created a new pooled library (Pool-2) consisting of six ecologically homogenous rhizosphere samples to minimize potential sequencing imbalances due to the heterogeneity of the multiplexed samples. DNA from these samples was fragmented at 1.5 Kb, to be able to sequence longer reads than for samples in Pool-1 and thereby facilitate the capture of multi-domain CRs. Target capture sequencing was successful for all six samples and produced a detection rate of CR sequences per sample that was several orders of magnitude higher than with bulk metagenomic sequencing

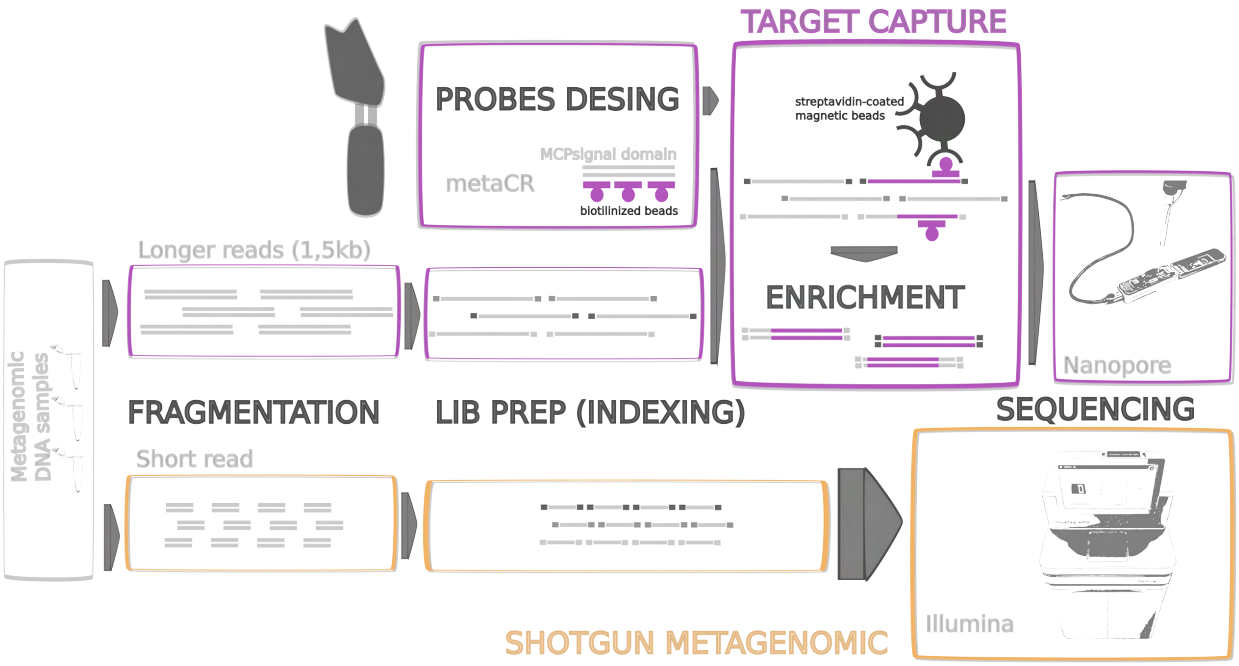



**Figure 3.2: Schematic representation of the molecular protocols implemented for long-read targeted metagenomics and shotgun metagenomics for the detection of chemoreceptors in metagenomic samples.**

(Figures 3.3c and 3.3d). This improvement was evident both in the sheer volume of CR genes identified per sample, which yielded an average 212-fold increase (Figure 3.3c), and in the efficiency of CR gene assembly per Gbp, with an average 924-fold increase (Figure 3.3d). In other words, target capture enrichment boosted CR gene discovery in the rare biosphere by two orders of magnitude compared to shotgun metagenomics sequencing, both with regard to total outcome per sample and when normalizing by equal sequencing depth. Importantly, this significant increase in CR gene detection is unrelated to the high rate of duplicates (Supplementary Table 3.2).

Apart from these quantitative improvements, target capture sequencing also yielded CR genes that were longer than those obtained by assembling regular shotgun metagenomic data (1,300 nts vs 521 nts). The ability to sequence long CR gene fragments greatly enhances the study of chemoreception, as sensing domains are typically located near the gene's N-terminus and therefore difficult to reach by short gene assemblies (Figure 3.3e). In fact, captured CR sequences were longer and contained more detectable domains close to the MCPsignal region than the bulk metagenomic CR gene assemblies (28,645 CR sequences matching 2,794 unique known Pfam

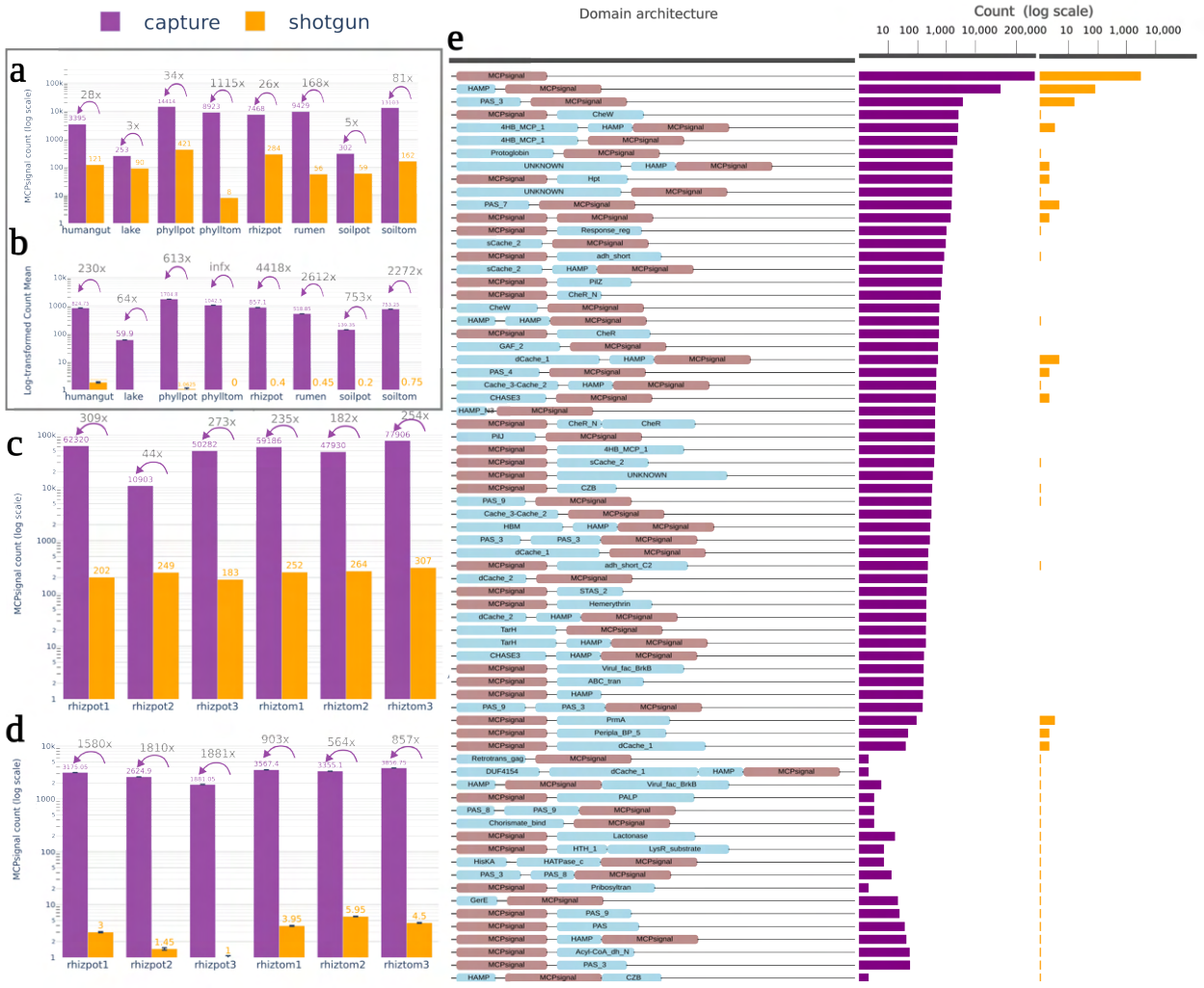



**Figure 3.3: Long-read targeted metagenomics sequencing approach boosts detection of CRs across diverse samples.** Total count of a) MCPsignal domains detected in Pool1; b) MCPsignal domains normalized by sequencing depth at 40Mbp (average of 10 randomized sets for Pool-1); c) MCPsignal domains detected for each sample in Pool-2; d) MCPsignal domains after normalization at 374Mbp (average of 10 randomized sets for Pool-2); and e) the 70 most frequent domain architectures detected in Pool-2.

domains vs 77 sequences and 94 Pfam). Notably, target capture sequencing of Pool-2 enabled us to identify 14,690 CR genes with at least one known LBD (i.e., a potential sensing domain). In contrast, bulk metagenomics sequencing only identified 63 CR gene assemblies with known LBDs.

Overall, our results highlight the significant functional variability and extensive range of sensing domains found in rare uncultured taxa. Additionally, they demonstrate that long-read target capture sequencing is a highly effective and powerful tool to analyze these taxa.

## 3.4 Unveiling the hidden biodiversity of low-abundance chemosensitive organisms

To investigate the phylogenetic biodiversity of the detected low-abundance chemosensitive organisms, we generated taxonomic profiles based on the taxonomic assignments inferred from BLAST searches of the CR genes' MCPsignal domain against the genomes of 65,703 representative prokaryotic species from the GTDB r207 database[165]. To ensure accurate taxonomic annotations, we called the presence of a species only if an identical (100% nucleotide identity) and not ambiguous (second best hit <10% nucleotide identity) CR sequence match was found in the species genome.

Under such criteria, target capture sequencing allowed us to taxonomically annotate 3,764 CR genes from 608 distinct species (Table 3.1), covering 19 bacterial and 2 archaeal phyla and highlighting the effectiveness of our target capture sequencing approach for uncovering hidden biodiversity in highly complex environments (Figure 3.4). In contrast, regular shotgun sequencing led to the taxonomic annotation of 12 CR genes from six species. From all capture-based species these detections, 203 species were supported by two or more CRs matches per genome in the case of capture-based data, while only two species were supported under the same threshold in shotgun-based data (Figure 3.5a).

Importantly, target capture sequencing led to the detection of 26 known plant pathogens and symbionts whose low abundance difficulties their identification using standard methods (Figure 3.5b and 3.5c). Relevant examples include the pathogens *Pantoea agglomerans*, *Dickeya dadantii*, and

*Pseudomonas syringae*, whose presence was supported by 100% nt identity matches for various CR genes per genome in capture-based data, but were undetected using on the standard shotgun approach. Similarly the symbiotic species *Esnifer adhaerens*, *Rhizobium rhizoryzae*, *Sinorhizobium saheli*, and *Sinorizobium meliloti*, were identified using the capture-based approach, but went undetected using bulk metagenomics. Overall, only four out of these 26 latent organisms were detectable using bulk metagenomics data, either by CR profiling or through taxonomic profiling of phylogenetic marker genes[222] (orange circles and names respectively in Figure 3.5).

**Table 3.1: Taxonomy assignment of known CRs from Genome Taxonomy Database (GTDB) of sequences identified in target capture and shotgun metagenomics.** Known CRs are unambiguously assigned to species based on comparisons using different identity thresholds to the CRs contained in the GTDB r207 database. The number of sequences assigned at species rank is shown along with the number of mapped species in parenthesis. Sequences that meet the identity thresholds but do not map to a single species are indicated as not species-specific. Sequences that are not mapped are shown, along with their percentage of the total sequences.

| | **100%** | | **98-100%** | | **80%** | |
|---|---|---|---|---|---|---|
| | **Capture** | **Shotgun** | **Capture** | **Shotgun** | **Capture** | **Shotgun** |
| **species specific** | 3,764 (608) | 12 (6) | 28,203 (1,507) | 74 (28) | 6,387 (1,785) | 59 (49) |
| **not species specific** | 350 | 0 | 2.959 | 3 | 289.377 | 1.133 |
| **not mapped** | 305,259 (98,67%) | 1,448 (99,18%) | 278,214 (89,93%) | 1,383 (94,73%) | 13,608 (4,40%) | 268 (18,36%) |

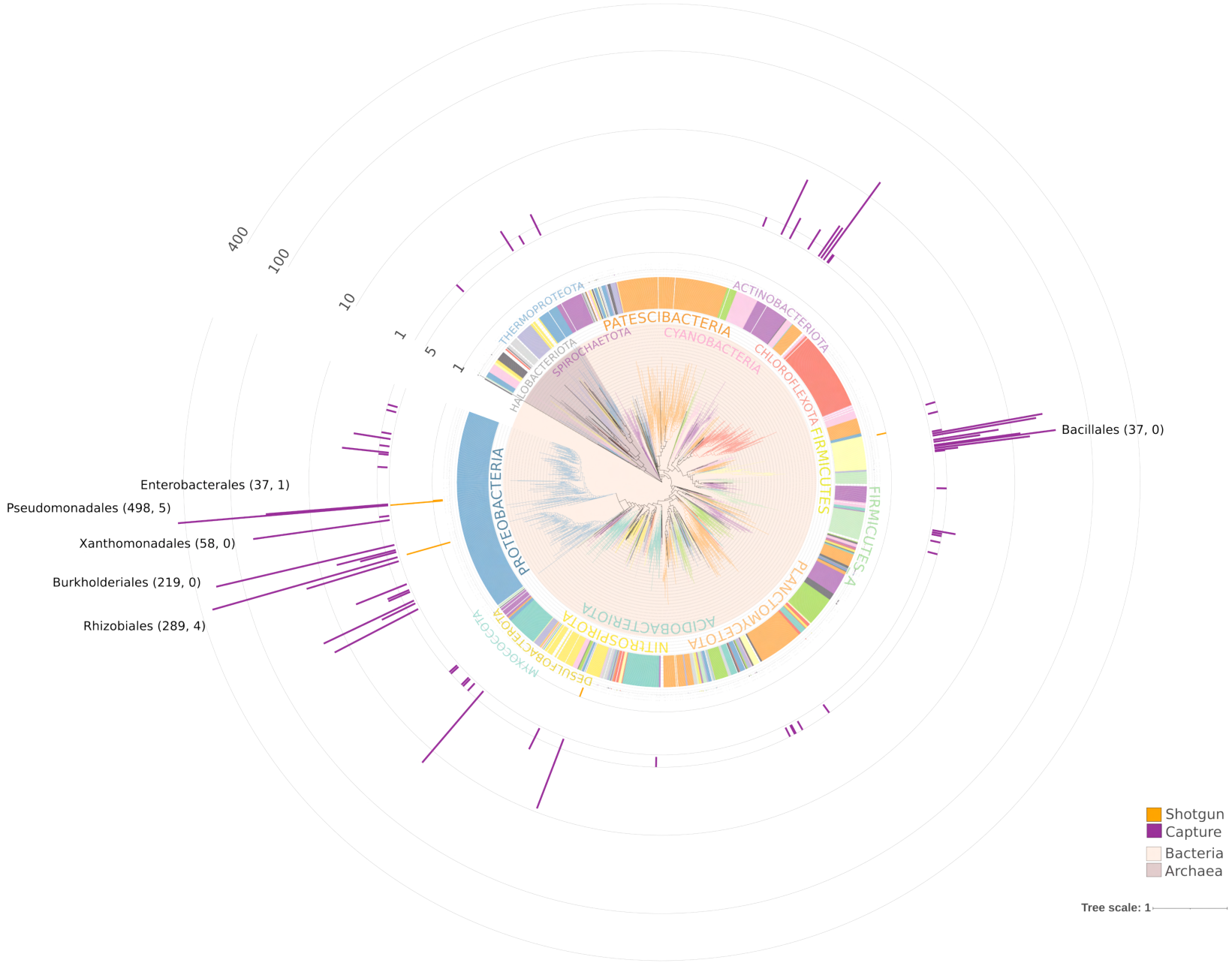



**Figure 3.4: Detection of CRs with long-read targeted metagenomics across the entire microbial phylogeny.** GTDB r207 archaeal and bacterial trees concatenated and collapsed at order level. Target capture (purple) and shotgun (orange) sequencing bars represent the total count (log scale) of different CRs detected at 100% nts identity unambiguously per order. Branches with more than 30 CRs are highlighted.

## 3.5 A vast space of novel chemosensitive species

Despite improving the detection of known chemosensitive species, the full majority of CRs observed in capture-based sequencing results remained in the range of 90-99% nucleotide against known references (Table 3.1, Supplementary Figure 3.3a), suggesting potential intra-species or intra-genus variability. However, 1,106 CRs showed much greater divergence as they did not map against any known gene. Therefore, we performed a phylogenetic analysis including both all known CR sequences available to date and the newly uncovered CRs from the rare microbial biosphere. The delineation of potential novel CRs from the phylogeny is shown in Table 3.2, where some groups were detected in different protein identity ranges. The Figure 3.6 is showing the last common ancestor of these new groups detected with at least 100 sequences across all the identity ranges.

Finally, we explored potential new sensing domains in the rare microbial biosphere. Sensing domains were extracted from the GTDB r207 database. After filtering the putative LBDs of Pool-2 with those matching the domains acting as sensing domains according to GTDB data, a total of 18,985 sensing domains were retrieved. From those, 2,638 (13,9%) did not map to any known LBD (from GTDB), with the majority of these domains within the “unknown” LBD classification (927 sequences, 35,14% of the no mapped ones). In addition, 7,364 (38,8%) from the mapped ones were significantly divergent ($\leq$ 70% protein sequence similarity) (Supplementary Figure 3.3b).

**Table 3.2: Delineation of novel CRs groups from target capture experiments.** A phylogenetic tree of MCPsignal domains was constructed using protein sequences from GTDB and captured sequences. Nodes were classified based on different protein identity ranges and minimum number of sequences per node, focusing on identifying novel or divergent CR groups derived exclusively from the capture experiments.

| % protein identity range of target capture CRs to GTDB CRs | min nº sequences | | | |
|---|---|---|---|---|
| | **50** | **100** | **150** | **200** |
| **0-65** | 23 | 11 | 7 | 3 |
| **65-70** | 57 | 21 | 10 | 5 |
| **70-75** | 123 | 48 | 31 | 21 |
| **75-80** | 200 | 109 | 70 | 56 |
| **80-85** | 187 | 95 | 66 | 49 |
| **85-90** | 3 | 3 | 3 | 3 |
| **90-95** | 261 | 149 | 102 | 80 |
| **95-100** | 240 | 124 | 86 | 71 |




**Figure 3.5: Revealing the repertoire of low abundance known CRs detected by long-read targeted metagenomics.** Number of different CRs detected (unique hit at 100% nts identity) by target capture (purple) or shotgun (orange) sequencing from among the total repertoire contained in GTDB r207 (gray), showing a) the top species with more CRs detected by either approach; b) plant pathogenic species only; and c) plant symbiont species only.

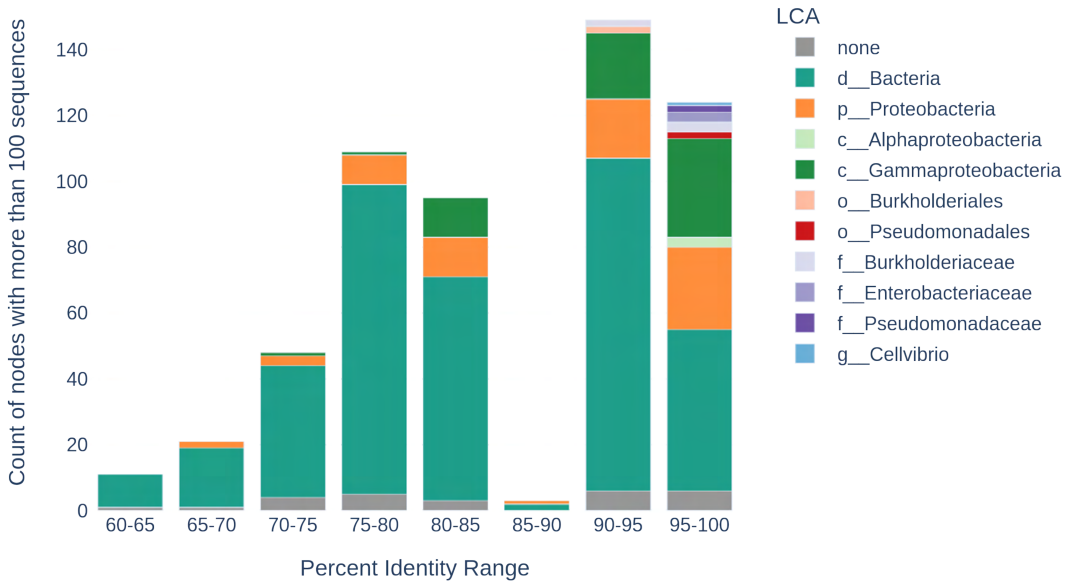



**Figure 3.6: Last common ancestor of novel CRs.** The number and the last common ancestor of the nodes with more than 100 sequences in each protein identity range are shown. The calculation of the last common ancestor was done based on the best hit taxonomic annotations of the target capture reads within the nodes in the tree.

## 3.6 Discussion

Although low-abundance chemosensitive bacteria and archaea are known to play key roles in shaping community structure[190], [191] and environmental interactions in microbial communities, their study by metagenomic sequencing has been hampered by a lack of phylogenetic depth and resolution. Here, we developed a novel target capture sequencing method to reveal the chemoreceptor repertoire of low-abundance taxa in natural microbial communities, overcoming PCR-dependent methods.

We based our capture enrichment panel on a broad set of DNA probes designed from a comprehensive catalog of CR genes observed in the global microbiome, covering sequence variability across the prokaryotic phylogeny on 13,174 public metagenomic samples from 14 habitats. This custom setup allowed us to explore the pool of CR genes present in an environmental sample, regardless of their abundance and phylogenetic affiliation, in a single experiment. Our approach facilitated the study of CRs in natural microbial communities at unprecedented depth and scale, and in a variety of microbial habitats. The number of unique CRs detected in the 6 rhizosphere samples included in Pool-2 exceeds the total number of CRs detected from shotgun sequencing data across the 13,000 metagenomic samples scanned to create the capture panel. This highlights the enormous biodiversity of chemosensory microorganisms hidden in the rare microbial biosphere. In particular, because hybridization capture is typically tolerant to sequence divergence levels of up to 20-25% nucleotide identity, our approach allowed the discovery of thousands of new CRs from unknown species. Target capture sequencing was particularly efficient in complex microbial habitats, where host DNA is typically overabundant and tends to bias sequencing results. This is the case of the tomato plant phyllosphere included in Pool1 (Supplementary Figure 3.4), where plant DNA is difficult to deplete, making it difficult to explore the microbial communities present in plant leaves.

However, a disadvantage of capture enrichment is that probes cannot realistically target the full range of possible sequences derived from highly variable regions of DNA. This was an important challenge in our study because the capture domains of CR are extremely variable and nearly impossible to target with a single enrichment panel. Therefore, we used an approach that combined

a capture panel targeting only the conserved MCPsignal domain with a custom protocol for long-fragment multiplexed target sequencing using ONT. This allowed us to sequence the untargeted sensing domains of the captured CR genes. The discovery rate of sensing domains was actually much higher when using long-fragment targeted sequencing than when using gene assemblies from regular metagenomics data.

While the performance of the capture method was very successful, we also noticed several limitations. First, capture experiments on heterogeneous pools of samples (i.e. from very different environments as in pool 1) affect the efficiency of the method and can bias the results towards certain samples. In addition, the high error rate of the old ONT flow cell chemistry (R9.4.1) makes phylogenetic classification of CRs difficult. The new ONT chemistry combined with homogeneous multiplexed sampling (e.g. Pool-2 in our study) clearly resulted in more efficient capture and lower sequencing error rates. This allowed us to taxonomically classify many CRs at the species level, which is useful for the detection of latent pathogens and many potentially novel chemosensing microorganisms.

Our study not only provides a novel method for unveiling a vast biodiversity of CRs, but also a comprehensive catalog of CRs genes observed across habitats, which serves as a reference for future studies on the functional, ecological, and evolutionary aspects of chemoreception in microorganisms.

# General Discussion

Genomic exploration of chemoreceptor (CR) diversity and function has been hampered by limited taxonomic representation, insufficient ecological context, and methodological challenges in recovering CRs from complex microbial communities. This work addresses many of these gaps through an integrative genomic and metagenomic perspective that advances our understanding of CR diversity, evolution and ecological importance. On a technical level, one of the main goals of this thesis was to establish a domain-centric methodology that allows for a better functional and evolutionary classification of CRs by treating each of their domain types as a distinct evolutionary and functional unit. Thus, all the work presented in this thesis is based on the hypothesis that the highly conserved transduction domain (MCPsignal) of CRs is subject to different evolutionary and ecological constraints than the highly variable sensing domain (LBD), an idea that generally applies to all multi-domain proteins[270]. We established this domain-centric methodology in Chapter 1 by creating an automated bioinformatics pipeline that was used in all our subsequent analyses and addressed different research questions.

For instance, this approach allowed us to consider the evolutionary history of each domain independently, leading to a more precise analysis of CRs and enabling the reclassification of sensing domains into fine-grained subfamilies. As expected, this unsupervised domain-centric methodology overcame the limitations of using existing Pfam HMM models[55] as the only reference, which we indeed found to lack representation for many LBDs[271]. It also enabled us to uncover previously uncharacterized LBD subtypes, demonstrating that the diversity of extracellular sensing domains, especially those classified as "unknown," remains far from saturation when new genomes are analyzed. We expect our pipeline and the resulting catalog of sensory subfamilies will serve as a resource for advancing the functional characterization and experimental validation of CRs in future work.

Another key objective of this work was to investigate the prevalence and global distribution of CRs in both cultured and uncultured prokaryotic species. In Chapter 2, we found that the biodiversity and sequence variability of CRs, and in particular their extracellular LBDs, is far from reaching a technical plateau. Such an open space of variability was not due to small sequence variations in the sequenced CRs, but in the number of deep homology clusters obtained from a domain-aware analysis of all CR genes, suggesting that differences are due to functional divergence. In total, we

identified 1.6 million CR sequences from known and unknown species. This expansion represents a major step forward in the study of CRs and their sensing domains, as it has revealed a large number of unknown LBDs and chemosensitive microorganisms. For example, by profiling CRs in both well-characterized and previously unknown chemosensitive microbial lineages, we were able to uncover numerous uncultivated taxa, such as the phylum *Moduliflexota*, which was found to harbor a rich repertoire of novel LBD types. This stresses the importance of including uncultivated bacteria in the analysis of CRs.

With regard to the ecological importance of CRs, this work provides compelling evidence for a strong link between the repertoire of CRs available in a microbiota community and the environment it inhabits. First, we have shown that plant-associated bacteria possess a highly specific set of CRs, the species abundance of which cannot be explained by their phylogenetic distribution alone. For example, in Chapter 1 we described a strong association between different sensor domains and a plant-associated lifestyle, particularly in plant pathogens and symbionts. This pattern was further analysed and extended in Chapter 2, where we integrated massive metagenomic data into our ecological analyses. Our results not only showed that ecological associations are broader, but also led us to determine that it is the extracellular LBD regions of CRs the ones that confer the stronger adaptive value to the environment. These associations were such that it was possible to predict the source biome of a metagenomic sample based on its LBD profiles, and even to identify LBDs with high prevalence and specificity for particular habitats, opening up new avenues for the identification of CR-based biomarkers. However, validating particular LBDs as biomarkers and confirming their adaptive value might require additional experiments.

While LBDs showed a strong ecological signal, the MCPsignal domain remained largely uninformative at the ecological level. However, having identified over 1.6M MCPsignal sequence variants in cultured and uncultured organisms, we were able to study the evolutionary history of CRs at an unprecedented phylogenetic scale. In Chapter 2, we showed that different MCPsignal structural subtypes (i.e. heptads) were likely generated by nested duplication events, some of which were lineage-specific. This not only allows us to trace the evolutionary origin of specific heptad types, but also provides valuable evolutionary information to guide, and probably improve, such classification. For example, we found a strong co-occurrence pattern between histidine kinase cheA ACF subtypes

and the emergence of certain heptad types, suggesting a possible functional link. While in silico predictions provide valuable insights, experimental validation is crucial, as predicted functional relationships may not always work as expected 275.

Finally, in Chapter 3, we explored the potential extent and variability of chemotactic microorganisms in rare, low abundance microorganisms. Although it has been postulated that chemotactic species are often found at low relative abundances within microbial communities, current sequencing technologies have not been able to systematically test and explore such a hypothesis. For example, we estimated that the average number of CRs detected per sequenced gigabase pair (Gbp) in a metagenomic sample was approximately six. Considering that some species may harbor more than 70 CRs, it is clear that traditional shotgun strategies may fail to represent the full diversity of chemosensing microorganisms.

To address the limitations of conventional metagenomics in studying low-abundance microorganisms, we developed a novel target capture sequencing method that significantly improves the detection of chemoreceptors (CRs) in natural microbial communities. Our approach, which bypasses PCR-dependent biases, allowed us to study CR genes regardless of their abundance or phylogenetic affiliation, achieving detection levels up to two orders of magnitude higher than traditional methods. As shown in Chapter 3, this method allowed us to capture CRs from latent pathogens in the samples that were not detected by standard metagenomic methods. We validated our novel method in different habitats, with particular emphasis on complex environments such as the rhizosphere, demonstrating that the hidden space of chemosensitive microorganisms is vast and includes numerous unknown species with novel sensing capabilities. Taken together with the observations in Chapter 2, our results show that the diversity of chemosensitive species and the functional variability of their CR genes is greater than originally thought, with no sign of a plateau in our environmental observations.

In its whole, the three chapters presented in this thesis updated and extended current knowledge of the origin, evolution and ecological implications of microbial chemoreceptors on a global scale, providing also new tools, protocols and databases that we expect will be valuable to address future questions.

We anticipate that the resources and biological insights obtained, combined with advances in protein structure prediction and computational docking, would enhance our ability to investigate ligand binding mechanisms and more precisely study the molecular functional and adaptive roles of individual chemoreceptors. Advances in protein structure prediction tools, such as AlphaFold [272], open up the possibility of creating a structural catalog of receptor domains that could be cross-linked with all evolutionary and ecological ligands derived from our globalCR dataset. The construction of such a structural catalog is currently underway in our laboratories and we expect it to provide clues to the molecules sensed by each CR. For example, recent research on Cache domains in gut-dwelling bacteria has shown that these domains have a well-defined ligand-binding pocket[136] that can be confidently modelled by AlphaFold. Advances in computational docking, such as DiffDock[273], offer promising opportunities for large-scale ligand screening, potentially leading to the identification of target ligands for unknown sensing domains.

# Conclusions

1. We have developed an automated method to automatically identify and analyse chemoreceptor (CR) genes from massive genomic and metagenomic data, allowing us to classify CRs differently according to their transduction (MCPsignal) and sensing (ligand binding) domains.

2. We show that CRs are widespread in prokaryotes, but plant-associated bacteria, and in particular plant pathogens and symbionts, possess a highly specific, ecologically driven repertoire of CRs and LBDs.

3. By exploring CRs across 99,146 metagenomes, we show that free-living and biodiverse environments, such as soil and aquatic habitats, harbor higher proportions of chemosensing species compared to confined ecosystems, such as the human gut.

4. Extracellular LBDs, but not the transducing MCP signalling domain or intracellular LBDs, carry strong ecological signals that are likely to represent adaptive value and make them potential global biomarkers of microbial lifestyle.

5. Phylogenetic analysis of over 428,179 unique MCPsignal domain sequences shows that different structural CR subtypes (heptads) have emerged and diversified through duplication events and correlate with subfunctionalization events.

6. Taxonomic and phylogenetic analysis of over 1.6 million CRs sequences observed in different habitats revealed hundreds of new species and lineages with chemosensory capabilities, including both cultured and uncultured microorganisms, significantly expanding the landscape of CR genes.

7. By developing a novel long-read target capture sequencing method to explore low-abundance microbes in environmental samples, we discovered a vast hidden diversity of CRs from rare, often unknown, taxa.

8. We show that our target capture sequencing method increases CR detection rates by orders of magnitude compared to bulk metagenomics, and also enables the sequencing of nearly full-length multi-domain CR genes.

9. Our target capture sequencing method proved to be highly effective for the identification of latent chemosensitive microorganisms, such as pathogens, that are undetectable by traditional methods.

10. The bioinformatic CR detection method, including domain-based clustering and classification of CR sequences, together with a global database of CRs across habitats, is available as a public resource for future analysis.

# References


[1] C. J. Hurst, *Microbes: The Foundation Stone of the Biosphere*. Springer International Publishing, 2021.

[2] "Unveiling the crucial role of soil microorganisms in carbon cycling: A review", *Sci. Total Environ.*, vol. 909, p. 168 627, Jan. 2024.

[3] M. M. M. Kuypers, H. K. Marchant, and B. Kartal, "The microbial nitrogen-cycling network", en, *Nat. Rev. Microbiol.*, vol. 16, no. 5, pp. 263–276, Feb. 2018.

[4] "Microorganisms and their roles in fundamental biogeochemical cycles", *Curr. Opin. Biotechnol.*, vol. 22, no. 3, pp. 456–464, Jun. 2011.

[5] L. Gao, T. Xu, G. Huang, S. Jiang, Y. Gu, and F. Chen, "Oral microbiomes: More and more importance in oral cavity and whole body", en, *Protein Cell*, vol. 9, no. 5, p. 488, May 2018.

[6] A. Heintz-Buschart and P. Wilmes, "Human gut microbiome: Function matters", en, *Trends Microbiol.*, vol. 26, no. 7, pp. 563–574, Jul. 2018.

[7] A. E. Douglas, "Simple animal models for microbiome research", en, *Nat. Rev. Microbiol.*, vol. 17, no. 12, pp. 764–775, Aug. 2019.

[8] "A review on the plant microbiome: Ecology, functions, and emerging trends in microbial application", *Journal of Advanced Research*, vol. 19, pp. 29–37, Sep. 2019.

[9] "Microbial bioremediation: A potential tool for restoration of contaminated areas", in *Microbial Biodegradation and Bioremediation*, Elsevier, Jan. 2014, pp. 1–21.

[10] C. Y. Chiu and S. A. Miller, "Clinical metagenomics", en, *Nat. Rev. Genet.*, vol. 20, no. 6, pp. 341–355, Mar. 2019.

[11] P. Amato, M. Joly, L. Besaury, *et al.*, “Active microorganisms thrive among extremely diverse communities in cloud water”, *PLoS One*, vol. 12, no. 8, e0182869, Aug. 2017.

[12] A. C. Schuerger, D. J. Smith, D. W. Griffin, *et al.*, “Science questions and knowledge gaps to study microbial transport and survival in asian and african dust plumes reaching north america”, en, *Aerobiologia*, vol. 34, no. 4, pp. 425–435, Oct. 2018.

[13] D. J. Smith, D. W. Griffin, R. D. McPeters, P. D. Ward, and A. C. Schuerger, “Microbial survival in the stratosphere and implications for global dispersal”, en, *Aerobiologia*, vol. 27, no. 4, pp. 319–332, Mar. 2011.

[14] “One-component systems dominate signal transduction in prokaryotes”, *Trends Microbiol.*, vol. 13, no. 2, pp. 52–56, Feb. 2005.

[15] “Chapter 3 diversity in bacterial chemotactic responses and niche adaptation”, in *Advances in Applied Microbiology*, vol. 66, Academic Press, Jan. 2009, pp. 53–75.

[16] G. H. Wadhams and J. P. Armitage, “Making sense of it all: Bacterial chemotaxis”, en, *Nat. Rev. Mol. Cell Biol.*, vol. 5, no. 12, pp. 1024–1037, Dec. 2004.

[17] M. A. Matilla, E. Monteagudo-Cascales, J. P. Cerna-Vargas, V. M. Gumerov, I. B. Zhulin, and T. Krell, “Is it possible to predict signal molecules that are recognized by bacterial receptors?”, en, *Environ. Microbiol.*, vol. 25, no. 1, pp. 11–16, Jan. 2023.

[18] K. Wuichet and I. B. Zhulin, “Origins and diversification of a complex signal transduction system in prokaryotes”, en, *Sci. Signal.*, Jun. 2010.

[19] E. J. Capra and M. T. Laub, “Evolution of two-component signal transduction systems”, *Annu. Rev. Microbiol.*, vol. 66, 2012.

[20] P. M. Wolanin, P. A. Thomason, and J. B. Stock, “Histidine protein kinases: Key signal transducers outside the animal kingdom”, en, *Genome Biol.*, vol. 3, no. 10, pp. 1–8, Sep. 2002.

[21] A. M. Stock, V. L. Robinson, and P. N. Goudreau, “Two-Component signal transduction”, *Annu. Rev. Biochem.*, vol. 69, no. Volume 69, 2000, pp. 183–215, Jul. 2000.

[22] I. B. Zhulin, A. N. Nikolskaya, and M. Y. Galperin, “Common extracellular sensory domains in transmembrane receptors for diverse signal transduction pathways in bacteria and archaea”, en, *J. Bacteriol.*, Jan. 2003.

[23] M. B. Clarke, D. T. Hughes, C. Zhu, E. C. Boedeker, and V. Sperandio, “The QseC sensor kinase: A bacterial adrenergic receptor”, *Proc. Natl. Acad. Sci. U. S. A.*, vol. 103, no. 27, Jul. 2006.

[24] F. F. Wang and W. Qian, “The roles of histidine kinases in sensing host plant and cell-cell communication signal in a phytopathogenic bacterium”, *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, vol. 374, no. 1767, Mar. 2019.

[25] P. D. Straight and R. Kolter, “Interspecies chemical communication in bacterial development”, *Annu. Rev. Microbiol.*, vol. 63, 2009.

[26] “Diversity of bacterial chemosensory systems”, *Curr. Opin. Microbiol.*, vol. 61, pp. 42–50, Jun. 2021.

[27] “Responding to chemical gradients: Bacterial chemotaxis”, *Curr. Opin. Cell Biol.*, vol. 24, no. 2, pp. 262–268, Apr. 2012.

[28] P. A. Spiro, J. S. Parkinson, and H. G. Othmer, “A model of excitation and adaptation in bacterial chemotaxis”, en, *Proceedings of the National Academy of Sciences*, vol. 94, no. 14, pp. 7263–7268, Jul. 1997.

[29] V. Sourjik and H. C. Berg, “Receptor sensitivity in bacterial chemotaxis”, en, *Proceedings of the National Academy of Sciences*, vol. 99, no. 1, pp. 123–127, Jan. 2002.

[30] C. H. Hansen, R. G. Endres, and N. S. Wingreen, “Chemotaxis in escherichia coli: A molecular model for robust precise adaptation”, *PLoS Comput. Biol.*, vol. 4, no. 1, e1, Jan. 2008.

[31] M. Li and G. L. Hazelbauer, “Selective allosteric coupling in core chemotaxis signaling complexes”, en, *Proceedings of the National Academy of Sciences*, vol. 111, no. 45, pp. 15 940–15 945, Nov. 2014.

[32] D. Martín-Mora, M. Fernández, F. Velando, *et al.*, “Functional annotation of bacterial signal transduction systems: Progress and challenges”, en, *Int. J. Mol. Sci.*, vol. 19, no. 12, p. 3755, Nov. 2018.

[33] J. S. Parkinson, G. L. Hazelbauer, and J. J. Falke, “Signaling and sensory adaptation in escherichia coli chemoreceptors: 2015 update”, en, *Trends Microbiol.*, vol. 23, no. 5, p. 257, May 2015.

[34] C. B. Whitchurch, A. J. Leech, M. D. Young, *et al.*, “Characterization of a complex chemosensory signal transduction system which controls twitching motility in pseudomonas aeruginosa”, en, *Mol. Microbiol.*, vol. 52, no. 3, pp. 873–893, May 2004.

[35] J. W. Hickman, D. F. Tifrea, and C. S. Harwood, “A chemosensory system that regulates biofilm formation through modulation of cyclic diguanylate levels”, en, *Proceedings of the National Academy of Sciences*, vol. 102, no. 40, pp. 14 422–14 427, Oct. 2005.

[36] J. E. Berleman and C. E. Bauer, “Involvement of a che-like signal transduction cascade in regulating cyst cell development in rhodospirillum centenum”, en, *Mol. Microbiol.*, vol. 56, no. 6, pp. 1457–1466, Jun. 2005.

[37] W. Liu, Y. Sun, R. Shen, *et al.*, “A chemotaxis-like pathway of azorhizobium caulinodans controls flagella-driven motility, which regulates biofilm formation, exopolysaccharide biosynthesis, and competitive nodulation”, en, *Mol. Plant. Microbe. Interact.*, May 2018.

[38] R. P. Alexander and I. B. Zhulin, “Evolutionary genomics reveals conserved structural determinants of signaling and adaptation in microbial chemoreceptors”, en, *Proceedings of the National Academy of Sciences*, vol. 104, no. 8, pp. 2885–2890, Feb. 2007.

[39] A. I. Salah Ud-Din and A. Roujeinikova, “Methyl-accepting chemotaxis proteins: A core sensing element in prokaryotes and archaea”, en, *Cell. Mol. Life Sci.*, vol. 74, no. 18, pp. 3293–3303, Apr. 2017.

[40] K. D. Collins, J. Lacal, and K. M. Ottemann, “Internal sense of direction: Sensing and signaling from cytoplasmic chemoreceptors”, en, *Microbiol. Mol. Biol. Rev.*, vol. 78, no. 4, p. 672, Dec. 2014.

[41] P. Ames, Q. Zhou, and J. S. Parkinson, “HAMP domain structural determinants for signalling and sensory adaptation in tsr, the escherichia coli serine chemoreceptor”, *Mol. Microbiol.*, vol. 91, no. 5, Mar. 2014.

[42] J. Lacal, C. García-Fontana, F. Muñoz-Martínez, J. Ramos, and T. Krell, “Sensing of environmental signals: Classification of chemoreceptors according to the size of their ligand binding regions”, en, *Sensing of environmental signals: Classification of chemoreceptors according to the size of their ligand binding regions*, vol. 12, no. 11, pp. 2873–2884, 2010.

[43] G. Cha, Y. Liu, Q. Yang, L. Bai, L. Cheng, and W. Fan, “Comparative genomic insights into chemoreceptor diversity and habitat adaptation of archaea”, en, *Appl. Environ. Microbiol.*, Oct. 2022.

[44] V. M. Gumerov, D. R. Ortega, O. Adebali, L. E. Ulrich, and I. B. Zhulin, “MiST 3.0: An updated microbial signal transduction database with an emphasis on chemosensory systems”, en, *Nucleic Acids Res.*, vol. 48, no. D1, pp. D459–D464, Nov. 2019.

[45] “Kinetics of receptor modification. the multiply methylated aspartate receptors involved in bacterial chemotaxis”, *J. Biol. Chem.*, vol. 261, no. 23, pp. 10 814–10 820, Aug. 1986.

[46] D. Yu, X. Ma, Y. Tu, and L. Lai, “Both piston-like and rotational motions are present in bacterial chemoreceptor signaling”, en, *Sci. Rep.*, vol. 5, no. 1, pp. 1–8, Mar. 2015.

[47] “Accessing nutrients as the primary benefit arising from chemotaxis”, *Curr. Opin. Microbiol.*, vol. 75, p. 102 358, Oct. 2023.

[48] M. A. Matilla, F. Velando, D. Martín-Mora, E. Monteagudo-Cascales, and T. Krell, “A catalogue of signal molecules that interact with sensor kinases, chemoreceptors and transcriptional regulators”, en, *FEMS Microbiol. Rev.*, vol. 46, no. 1, fuab043, Aug. 2021.

[49] J. A. Gavira, V. M. Gumerov, M. Rico-Jiménez, *et al.*, “How bacterial chemoreceptors evolve novel ligand specificities”, en, *MBio*, Feb. 2020.

[50] Á. Ortega, I. B. Zhulin, and T. Krell, “Sensory repertoire of bacterial chemoreceptors”, en, *Microbiol. Mol. Biol. Rev.*, vol. 81, no. 4, Dec. 2017.

[51] “Chemoreceptor-based signal sensing”, *Curr. Opin. Biotechnol.*, vol. 45, pp. 8–14, Jun. 2017.

[52] V. M. Gumerov, E. P. Andrianova, M. A. Matilla, *et al.*, “Amino acid sensor conserved from bacteria to humans”, en, *Proceedings of the National Academy of Sciences*, vol. 119, no. 10, e2110415119, Mar. 2022.

[53] A. A. Upadhyay, A. D. Fleetwood, O. Adebali, R. D. Finn, and I. B. Zhulin, “Cache domains that are homologous to, but different from PAS domains comprise the largest superfamily of extracellular sensors in prokaryotes”, en, *PLoS Comput. Biol.*, vol. 12, no. 4, Apr. 2016.

[54] E. Monteagudo-Cascales, J. A. Gavira, J. Xing, *et al.*, “Bacterial sensor evolved by decreasing complexity”, en, May 2024.

[55] J. Xing, V. M. Gumerov, and I. B. Zhulin, “Origin and functional diversification of PAS domain, a ubiquitous intracellular sensor”, en, *Science Advances*, Sep. 2023.

[56] E. Jonasch, F. Donskov, O. Iliopoulos, *et al.*, “Belzutifan for renal cell carcinoma in von Hippel-Lindau disease”, *N. Engl. J. Med.*, vol. 385, no. 22, Nov. 2021.

[57] M. V. Airola, D. Huh, N. Sukomon, *et al.*, “Architecture of the soluble receptor aer2 indicates an in-line mechanism for PAS and HAMP domain signaling”, *J. Mol. Biol.*, vol. 425, no. 5, Mar. 2013.

[58] Z. Xie, L. E. Ulrich, I. B. Zhulin, and G. Alexandre, “PAS domain containing chemoreceptor couples dynamic changes in metabolism with chemotaxis”, *Proc. Natl. Acad. Sci. U. S. A.*, vol. 107, no. 5, Feb. 2010.

[59] T. Paysan-Lafosse, M. Blum, S. Chuguransky, *et al.*, “Interpro in 2022”, *Nucleic Acids Research*, vol. 51, no. D1, pp. D418–D427, Nov. 2022. DOI: 10.1093/nar/gkac993. eprint: https://academic.oup.com/nar/article-pdf/51/D1/D418/48440814/gkac993.pdf.

[60] P. Tohidifar, M. J. Plutz, G. W. Ordal, and C. V. Rao, “The mechanism of bidirectional ph taxis in bacillus subtilis”, *J. Bacteriol.*, vol. 202, no. 4, Jan. 2020.

[61] G. Alexandre, “Coupling metabolism and chemotaxis-dependent behaviours by energy taxis receptors”, *Microbiology*, vol. 156, no. 8, pp. 2283–2293, Aug. 2010.

[62] D. Samanta, J. Widom, P. P. Borbat, J. H. Freed, and B. R. Crane, “Bacterial energy sensor aer modulates the activity of the chemotaxis kinase CheA based on the redox state of the flavin cofactor”, *J. Biol. Chem.*, vol. 291, no. 50, Dec. 2016.

[63] S. Bi and V. Sourjik, “Stimulus sensing and signal processing in bacterial chemotaxis”, *Curr. Opin. Microbiol.*, vol. 45, Oct. 2018.

[64] J. A. Reyes-Darias, Y. Yang, V. Sourjik, and T. Krell, “Correlation between signal input and output in PctA and PctB amino acid chemoreceptor of pseudomonas aeruginosa”, en, *Mol. Microbiol.*, vol. 96, no. 3, pp. 513–525, May 2015.

[65] S. Bi, A. M. Pollard, Y. Yang, F. Jin, and V. Sourjik, “Engineering hybrid chemotaxis receptors in bacteria”, en, Jun. 2016.

[66] J. M. Keegstra, F. Carrara, and R. Stocker, “The ecological roles of bacterial chemotaxis”, en, *Nat. Rev. Microbiol.*, vol. 20, no. 8, pp. 491–504, Mar. 2022.

[67] “Bacterial chemotaxis in human diseases”, *Trends Microbiol.*, vol. 31, no. 5, pp. 453–467, May 2023.

[68] “Discovery, research, and development of new antibiotics: The WHO priority list of antibiotic-resistant bacteria and tuberculosis”, *Lancet Infect. Dis.*, vol. 18, no. 3, pp. 318–327, Mar. 2018.

[69] K. Hou, Z.-X. Wu, X.-Y. Chen, *et al.*, “Microbiota in health and diseases”, en, *Signal Transduction and Targeted Therapy*, vol. 7, no. 1, pp. 1–28, Apr. 2022.

[70] N. Arpaia, C. Campbell, X. Fan, *et al.*, “Metabolites produced by commensal bacteria promote peripheral regulatory t-cell generation”, en, *Nature*, vol. 504, no. 7480, pp. 451–455, Nov. 2013.

[71] P. Spanogiannopoulos, T. S. Kyaw, B. G. H. Guthrie, *et al.*, “Host and gut bacteria share metabolic pathways for anti-cancer drug metabolism”, en, *Nature Microbiology*, vol. 7, no. 10, pp. 1605–1620, Sep. 2022.

[72] Z. Jie, H. Xia, S.-L. Zhong, *et al.*, “The gut microbiome in atherosclerotic cardiovascular disease”, en, *Nat. Commun.*, vol. 8, no. 1, pp. 1–12, Oct. 2017.

[73] N. Qin, F. Yang, A. Li, *et al.*, “Alterations of the human gut microbiome in liver cirrhosis”, en, *Nature*, vol. 513, no. 7516, pp. 59–64, Jul. 2014.

[74] O. Irazoki, J. ter Beek, L. Alvarez, *et al.*, “D-amino acids signal a stress-dependent run-away response in vibrio cholerae”, en, *Nature Microbiology*, vol. 8, no. 8, pp. 1549–1560, Jun. 2023.

[75] D. R. Ortega, A. Kjær, and A. Briegel, “The chemosensory systems of vibrio cholerae”, en, *Mol. Microbiol.*, vol. 114, no. 3, p. 367, Sep. 2020.

[76] F. Rivera-Chávez, C. A. Lopez, L. F. Zhang, *et al.*, “Energy taxis toward Host-Derived nitrate supports a salmonella pathogenicity island 1-independent mechanism of invasion”, en, *MBio*, Jul. 2016.

[77] M. F. Khan, M. A. Machuca, M. M. Rahman, *et al.*, “Structure–Activity relationship study reveals the molecular basis for specific sensing of hydrophobic amino acids by the campylobacter jejuni chemoreceptor tlp3”, en, *Biomolecules*, vol. 10, no. 5, p. 744, May 2020.

[78] C. J. Day, R. M. King, L. K. Shewell, *et al.*, “A direct-sensing galactose chemoreceptor recently evolved in invasive strains of campylobacter jejuni”, en, *Nat. Commun.*, vol. 7, no. 1, pp. 1–8, Oct. 2016.

[79] E. Goers Sweeney, J. N. Henderson, J. Goers, *et al.*, “Structure and proposed mechanism for the ph-sensing helicobacter pylori chemoreceptor TlpB”, en, *Structure*, vol. 20, no. 7, pp. 1177–1188, Jul. 2012.

[80] H. Hanyu, K. A. Engevik, A. L. Matthis, K. M. Ottemann, M. H. Montrose, and E. Aihara, “Helicobacter pylori uses the TlpB receptor to sense sites of gastric injury”, en, *Infect. Immun.*, Jul. 2019.

[81] M. Azevedo-Vethacke, D. Garten, C. Groll, and S. Schreiber, “Specific therapeutic schemes of omeprazole affect the orientation of helicobacter pylori”, en, *Antimicrob. Agents Chemother.*, vol. 53, no. 8, p. 3511, Aug. 2009.

[82] E. Faure, K. Kwong, and D. Nguyen, “Pseudomonas aeruginosa in chronic lung infections: How to adapt within the host?”, en, *Front. Immunol.*, vol. 9, 2018.

[83] M. A. Matilla, F. Velando, A. Tajuelo, *et al.*, “Chemotaxis of the human pathogen pseudomonas aeruginosa to the neurotransmitter acetylcholine”, en, *MBio*, Mar. 2022.

[84] A. Corral-Lugo, M. A. Matilla, D. Martín-Mora, *et al.*, "High-Affinity chemotaxis to histamine mediated by the TlpQ chemoreceptor of the human pathogen pseudomonas aeruginosa", en, *MBio*, Nov. 2018.

[85] A. M. Patterson, I. E. Mulder, A. J. Travis, *et al.*, "Human gut symbiont roseburia hominis promotes and regulates innate immunity", en, *Front. Immunol.*, vol. 8, p. 281 394, Sep. 2017.

[86] M. A. Charette and W. H. F. Smith, "The volume of earth's ocean", *Oceanography*, vol. 23, no. 2, pp. 112–114, Oct. 2015.

[87] R. Stocker and J. R. Seymour, "Ecology and physics of bacterial chemotaxis in the ocean", en, *Microbiol. Mol. Biol. Rev.*, vol. 76, no. 4, p. 792, Dec. 2012.

[88] N. Blackburn, T. Fenchel, and J. Mitchell, "Microscale nutrient patches in planktonic habitats shown by chemotactic bacteria", en, *Science*, vol. 282, no. 5397, pp. 2254–2256, Dec. 1998.

[89] F. Azam, "Microbial control of oceanic carbon flux: The plot thickens", en, *Science*, vol. 280, no. 5364, pp. 694–696, May 1998.

[90] J. G. Mitchell, A. Okubo, and J. A. Fuhrman, "Microzones surrounding phytoplankton form the basis for a stratified marine microbial ecosystem", en, *Nature*, vol. 316, no. 6023, pp. 58–59, Jul. 1985.

[91] T. Kiørboe and G. A. Jackson, "Marine snow, organic solute plumes, and optimal chemosensory behavior of bacteria", en, *Limnol. Oceanogr.*, vol. 46, no. 6, pp. 1309–1318, Sep. 2001.

[92] R. Stocker, J. R. Seymour, A. Samadani, D. E. Hunt, and M. F. Polz, "Rapid chemotactic response enables marine bacteria to exploit ephemeral microscale nutrient patches", en, *Proceedings of the National Academy of Sciences*, vol. 105, no. 11, pp. 4209–4214, Mar. 2008.

[93] J. R. Seymour, S. A. Amin, J. B. Raina, and R. Stocker, "Zooming in on the phycosphere: The ecological interface for phytoplankton-bacteria relationships", *Nature microbiology*, vol. 2, May 2017.

[94] J.-B. Raina, B. S. Lambert, D. H. Parks, *et al.*, “Chemotaxis shapes the microscale organization of the ocean’s microbiome”, en, *Nature*, vol. 605, no. 7908, pp. 132–138, Apr. 2022.

[95] S. Hallstrøm, J.-B. Raina, M. Ostrowski, *et al.*, “Chemotaxis may assist marine heterotrophic bacterial diazotrophs to find microzones suitable for N2 fixation in the pelagic ocean”, en, *ISME J.*, vol. 16, no. 11, p. 2525, Nov. 2022.

[96] M. Bentzon-Tilia, H. Farnelid, K. Jürgens, and L. Riemann, “Cultivation and isolation of n2-fixing bacteria from suboxic waters in the baltic sea”, en, *FEMS Microbiol. Ecol.*, vol. 88, no. 2, pp. 358–371, Apr. 2014.

[97] K. M. Scott, J. Williams, C. M. B. Porter, *et al.*, “Genomes of ubiquitous marine and hypersaline hydrogenovibrio, thiomicrorhabdus and thiomicrospira spp. encode a diversity of mechanisms to sustain chemolithoautotrophy in heterogeneous environments”, *Environ. Microbiol.*, vol. 20, no. 8, Aug. 2018.

[98] E. G. Ruby and L. M. Asato, “Growth and flagellation of vibrio fischeri during initiation of the sepiolid squid light organ symbiosis”, en, *Arch. Microbiol.*, vol. 159, no. 2, pp. 160–167, Feb. 1993.

[99] M. J. Mandel, A. L. Schaefer, C. A. Brennan, *et al.*, “Squid-Derived chitin oligosaccharides are a chemotactic signal during colonization by vibrio fischeri”, en, *Appl. Environ. Microbiol.*, Jul. 2012.

[100] P. Trivedi, J. E. Leach, S. G. Tringe, T. Sa, and B. K. Singh, “Plant–microbiome interactions: From community assembly to plant health”, en, *Nat. Rev. Microbiol.*, vol. 18, no. 11, pp. 607–621, Aug. 2020.

[101] J. W. Kloepper, R. Lifshitz, and R. M. Zablotowicz, “Free-living bacterial inocula for enhancing crop productivity”, en, *Trends Biotechnol.*, vol. 7, no. 2, pp. 39–44, Feb. 1989.

[102] P. Vandenkoornhuyse, A. Quaiser, M. Duhamel, A. Le Van, and A. Dufresne, “The importance of the microbiome of the plant holobiont”, en, *New Phytol.*, vol. 206, no. 4, pp. 1196–1206, Jun. 2015.

[103] P. R. Hirsch and T. H. Mauchline, “Who’s who in the plant root microbiome?”, *Nat. Biotechnol.*, vol. 30, no. 10, Oct. 2012.

[104] J. A. Vorholt, “Microbial life in the phyllosphere”, en, *Nat. Rev. Microbiol.*, vol. 10, no. 12, pp. 828–840, Nov. 2012.

[105] S. Compant, M. C. Cambon, C. Vacher, B. Mitter, A. Samad, and A. Sessitsch, “The plant endosphere world - bacterial life within plants”, *Environ. Microbiol.*, vol. 23, no. 4, Apr. 2021.

[106] J. Mansfield, S. Genin, S. Magori, *et al.*, “Top 10 plant pathogenic bacteria in molecular plant pathology”, en, *Mol. Plant Pathol.*, vol. 13, no. 6, pp. 614–629, Aug. 2012.

[107] M. A. Matilla and T. Krell, “The effect of bacterial chemotaxis on host infection and pathogenicity”, en, *FEMS Microbiol. Rev.*, vol. 42, no. 1, fux052, Oct. 2017.

[108] R. Mendes, P. Garbeva, and J. M. Raaijmakers, “The rhizosphere microbiome: Significance of plant beneficial, plant pathogenic, and human pathogenic microorganisms”, *FEMS Microbiol. Rev.*, vol. 37, no. 5, Sep. 2013.

[109] Y. Yan, E. E. Kuramae, M. de Hollander, P. G. Klinkhamer, and J. A. van Veen, “Functional traits dominate the diversity-related selection of bacterial communities in the rhizosphere”, *ISME J.*, vol. 11, no. 1, Jan. 2017.

[110] D. Bulgarelli, K. Schlaeppi, S. Spaepen, V. L. van Themaat E, and P. Schulze-Lefert, “Structure and functions of the bacterial microbiota of plants”, *Annu. Rev. Plant Biol.*, vol. 64, 2013.

[111] J.-B. Raina, V. Fernandez, B. Lambert, R. Stocker, and J. R. Seymour, “The role of microbial motility and chemotaxis in symbiosis”, en, *Nat. Rev. Microbiol.*, vol. 17, no. 5, pp. 284–294, Mar. 2019.

[112] B. A. Webb, S. Hildreth, R. F. Helm, and B. E. Scharf, “Sinorhizobium meliloti chemoreceptor McpU mediates chemotaxis toward host plant exudates through direct proline sensing”, en, *Appl. Environ. Microbiol.*, vol. 80, no. 11, pp. 3404–3415, Jun. 2014.

[113] J. Manuel Munoz Aguilar, A. M. Ashby, A. J. M. Richards, G. J. Loake, M. D. Watson, and C. H. Shaw, “Chemotaxis of rhizobium leguminosarum biovar phaseoli towards flavonoid inducers of the symbiotic nodulation genes”, *Microbiology*, vol. 134, no. 10, pp. 2741–2746, Oct. 1988.

[114] C. K. Yost, P. Rochepeau, and M. F. Hynes, “Rhizobium leguminosarum contains a group of genes that appear to code for methyl-accepting chemotaxis proteins”, *Microbiology*, vol. 144 ( Pt 7), Jul. 1998.

[115] W. M. Barbour, D. R. Hattermann, and G. Stacey, “Chemotaxis of bradyrhizobium japonicum to soybean exudates”, en, *Appl. Environ. Microbiol.*, vol. 57, no. 9, p. 2635, Sep. 1991.

[116] C. García-Fontana, J. A. Reyes-Darias, F. Muñoz-Martínez, *et al.*, “High specificity in CheR methyltransferase function: CheR2 OF PSEUDOMONAS PUTIDA IS ESSENTIAL FOR CHEMOTAXIS, WHEREAS CheR1 IS INVOLVED IN BIOFILM FORMATION *”, en, *J. Biol. Chem.*, vol. 288, no. 26, pp. 18 987–18 999, Jun. 2013.

[117] S. Oku, A. Komatsu, Y. Nakashimada, T. Tajima, and J. Kato, “Identification of pseudomonas fluorescens chemotaxis sensory proteins for malate, succinate, and fumarate, and their involvement in root colonization”, *Microbes Environ.*, vol. 29, no. 4, 2014.

[118] G. Vardar, P. Barbieri, and T. K. Wood, “Chemotaxis of pseudomonas stutzeri OX1 and burkholderia cepacia G4 toward chlorinated ethenes”, en, *Appl. Microbiol. Biotechnol.*, vol. 66, no. 6, pp. 696–701, Jul. 2004.

[119] I. Sampedro, R. E. Parales, T. Krell, and J. E. Hill, “Pseudomonas chemotaxis”, en, *FEMS Microbiol. Rev.*, vol. 39, no. 1, pp. 17–46, Nov. 2014.

[120] A. Hida, S. Oku, T. Kawasaki, Y. Nakashimada, T. Tajima, and J. Kato, “Identification of the mcpa and mcpm genes, encoding methyl-accepting proteins involved in amino acid and l-malate chemotaxis, and involvement of McpM-mediated chemotaxis in plant infection by ralstonia pseudosolanacearum (formerly ralstonia solanacearum phylotypes I and III)”, *Appl. Environ. Microbiol.*, vol. 81, no. 21, Nov. 2015.

[121] D. Parke, L. N. Ornston, and E. W. Nester, “Chemotaxis to plant phenolic inducers of virulence genes is constitutively expressed in the absence of the ti plasmid in agrobacterium tumefaciens”, *J. Bacteriol.*, vol. 169, no. 11, Nov. 1987.

[122] A. M. Ashby, Watson, G. J. Loake, and C. H. Shaw, “Ti plasmid-specified chemotaxis of agrobacterium tumefaciens C58C1 toward vir-inducing phenolic compounds and

soluble factors from monocotyledonous and dicotyledonous plants", *J. Bacteriol.*, vol. 170, no. 9, Sep. 1988.

[123] S. E. Lindow and M. T. Brandl, "Microbiology of the phyllosphere", *Appl. Environ. Microbiol.*, vol. 69, no. 4, Apr. 2003.

[124] S. Santamaría-Hernando, J. P. Cerna-Vargas, P. M. Martínez-García, *et al.*, "Blue-light perception by epiphytic pseudomonas syringae drives chemoreceptor expression, enabling efficient plant infection", en, *Mol. Plant Pathol.*, vol. 21, no. 12, pp. 1606–1619, Dec. 2020.

[125] R. K. Verma, B. Samal, and S. Chatterjee, "Xanthomonas oryzae pv. oryzae chemotaxis components and chemoreceptor mcp2 are involved in the sensing of constituents of xylem sap and contribute to the regulation of virulence-associated functions and entry into rice", en, *Mol. Plant Pathol.*, vol. 19, no. 11, p. 2397, Nov. 2018.

[126] S. Santamaría-Hernando, Á. López-Maroto, C. Galvez-Roldán, *et al.*, "Pseudomonas syringae pv. tomato infection of tomato plants is mediated by GABA and l-pro chemoperception", *Mol. Plant Pathol.*, vol. 23, no. 10, Oct. 2022.

[127] J. P. Cerna-Vargas, S. Santamaría-Hernando, M. A. Matilla, *et al.*, "Chemoperception of specific amino acids controls phytopathogenicity in pseudomonas syringae pv. tomato", en, *MBio*, Oct. 2019.

[128] M. Antunez-Lamas, E. Cabrera, E. Lopez-Solanilla, *et al.*, "Bacterial chemoattraction towards jasmonate plays a role in the entry of dickeya dadantii through wounded tissues", en, *Mol. Microbiol.*, vol. 74, no. 3, pp. 662–671, Nov. 2009.

[129] C. Gálvez-Roldán, J. P. Cerna-Vargas, J. J. Rodríguez-Herva, T. Krell, S. Santamaría-Hernando, and E. López-Solanilla, "A Nitrate-Sensing Domain-Containing chemoreceptor is required for successful entry and virulence of dickeya dadantii 3937 in potato plants", en, *Phytopathology®*, Mar. 2023.

[130] D. R. Brumley, F. Carrara, A. M. Hein, G. I. Hagstrom, S. A. Levin, and R. Stocker, "Cutting through the noise: Bacterial chemotaxis in marine microenvironments", en, *Front. Mar. Sci.*, vol. 7, p. 529 369, Jul. 2020.

[131] R. G. Endres and N. S. Wingreen, “Accuracy of direct gradient sensing by single cells”, en, *Proc. Natl. Acad. Sci. U. S. A.*, vol. 105, no. 41, pp. 15 749–15 754, Oct. 2008.

[132] D. A. Brown and H. C. Berg, “Temporal stimulation of chemotaxis in escherichia coli”, en, *Proceedings of the National Academy of Sciences*, vol. 71, no. 4, pp. 1388–1392, Apr. 1974.

[133] M. M. Salek, F. Carrara, V. Fernandez, J. S. Guasto, and R. Stocker, “Bacterial chemotaxis in a microfluidic t-maze reveals strong phenotypic heterogeneity in chemotactic sensitivity”, en, *Nat. Commun.*, vol. 10, no. 1, pp. 1–11, Apr. 2019.

[134] J. R. Seymour, R. Simó, T. Ahmed, and R. Stocker, “Chemoattraction to dimethylsulfoniopropionate throughout the marine microbial food web”, en, *Science*, Jul. 2010.

[135] L. Jiang, Q. Ouyang, and Y. Tu, “Quantitative modeling of escherichia coli chemotactic motion in environments varying in space and time”, *PLoS Comput. Biol.*, vol. 6, no. 4, e1000735, Apr. 2010.

[136] P. A. Ross, W. Xu, E. Jalomo-Khayrova, *et al.*, “Framework for exploring the sensory repertoire of the human gut microbiota”, en, *MBio*, Jun. 2024.

[137] V. M. Gumerov, L. E. Ulrich, and I. B. Zhulin, “MiST 4.0: A new release of the microbial signal transduction database, now with a metagenomic component”, en, *Nucleic Acids Res.*, vol. 52, no. D1, pp. D647–D653, Oct. 2023.

[138] A. Van Leeuwenhoek, “Observations, communicated to the publisher by mr. antony van leewenhoeck, in a dutch letter of the 9th octob. 1676. here english’d: Concerning little animals by him observed in rain-well-sea- and snow water; as also in water wherein pepper had lain infused”, en, *Philosophical Transactions of the Royal Society of London*, Mar. 1677.

[139] “Robert koch and the ‘golden age’ of bacteriology”, *Int. J. Infect. Dis.*, vol. 14, no. 9, e744–e751, Sep. 2010.

[140] M. Wainwright and J. Lederbergl, *History of microbiology*, en, 1992.

[141] T. W. Engelmann, *Neue methode zur untersuchung der sauerstoffausscheidung pflanzlicher und thierischer organismen*, Accessed: 2024-11-21, 1881.

[142] D. B. Roszak, D. J. Grimes, and R. R. Colwell, “Viable but nonrecoverable stage of salmonella enteritidis in aquatic systems”, *Can. J. Microbiol.*, vol. 30, no. 3, Mar. 1984.

[143] J. D. Oliver, L. Nilsson, and S. Kjelleberg, “Formation of nonculturable vibrio vulnificus cells and its relationship to the starvation state”, en, *Appl. Environ. Microbiol.*, vol. 57, no. 9, p. 2640, Sep. 1991.

[144] R. R. Colwell, P. Brayton, D. Herrington, B. Tall, A. Huq, and M. M. Levine, “Viable but non-culturable vibrio cholerae O1 revert to a cultivable state in the human intestine”, *World J. Microbiol. Biotechnol.*, vol. 12, no. 1, Jan. 1996.

[145] G. L. Hazelbauer, “Bacterial chemotaxis: The early years of molecular studies”, en, *Annu. Rev. Microbiol.*, vol. 66, p. 285, 2012.

[146] R. Fu and H. Feng, “Deciphering bacterial chemorepulsion: The complex response of microbes to environmental stimuli”, en, *Microorganisms*, vol. 12, no. 8, Aug. 2024.

[147] J. Shendure, S. Balasubramanian, G. M. Church, *et al.*, “DNA sequencing at 40: Past, present and future”, *Nature*, vol. 550, no. 7676, Oct. 2017.

[148] I. B. Zhulin, *The superfamily of chemotaxis transducers: From physiology to genomics and back*, 2001.

[149] S. I. Bibikov, R. Biran, K. E. Rudd, and J. S. Parkinson, “A signal transducer for aerotaxis in escherichia coli”, *Journal of bacteriology*, vol. 179, no. 12, Jun. 1997.

[150] F. R. Blattner, G. Plunkett, C. A. Bloch, *et al.*, “The complete genome sequence of escherichia coli K-12”, *Science (New York, N.Y.)*, vol. 277, no. 5331, Sep. 1997.

[151] A. Rebbapragada, M. S. Johnson, G. P. Harding, *et al.*, “The aer protein and the serine chemoreceptor tsr independently sense intracellular energy levels and transduce oxygen, redox, and energy signals for escherichia coli behavior”, en, *Proceedings of the National Academy of Sciences*, vol. 94, no. 20, pp. 10 541–10 546, Sep. 1997.

[152] R. Wu and A. D. Kaiser, “Structure and base sequence in the cohesive ends of bacteriophage lambda DNA”, *J. Mol. Biol.*, vol. 35, no. 3, Aug. 1968.

[153] M. Cobb and N. Comfort, *What rosalind franklin truly contributed to the discovery of DNA's structure*, en, http://dx.doi.org/10.1038/d41586-023-01313-5, Accessed: 2024-6-25, Apr. 2023.

[154] F. Sanger, S. Nicklen, and A. R. Coulson, "DNA sequencing with chain-terminating inhibitors", *Proc. Natl. Acad. Sci. U. S. A.*, vol. 74, no. 12, Dec. 1977.

[155] R. D. Fleischmann, Adams, O. White, *et al.*, "Whole-genome random sequencing and assembly of haemophilus influenzae rd", *Science*, vol. 269, no. 5223, Jul. 1995.

[156] "The superfamily of chemotaxis transducers: From physiology to genomics and back", vol. 45, pp. 157–198, Jan. 2001.

[157] H. Szurmant and G. W. Ordal, "Diversity in chemotaxis mechanisms among the bacteria and archaea", en, *Microbiol. Mol. Biol. Rev.*, 2004.

[158] J. Marchant, B. Wren, and J. Ketley, "Exploiting genome sequence: Predictions for mechanisms of campylobacter chemotaxis", en, *Trends Microbiol.*, vol. 10, no. 4, pp. 155–159, Apr. 2002.

[159] J. M. Heather and B. Chain, "The sequence of sequencers: The history of sequencing DNA", *Genomics*, vol. 107, no. 1, Jan. 2016.

[160] K. G. Lloyd, A. D. Steen, J. Ladau, J. Yin, and L. Crosby, "Phylogenetically novel uncultured microbial cells dominate earth microbiomes", en, *mSystems*, Sep. 2018.

[161] J. T. Staley and A. Konopka, "Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats", *Annu. Rev. Microbiol.*, vol. 39, 1985.

[162] S. J. Sogin, M. L. Sogin, and C. R. Woese, "Phylogenetic measurement in procaryotes by primary structural characterization", *J. Mol. Evol.*, vol. 1, no. 2, Jun. 1972.

[163] C. R. Woese and G. E. Fox, "Phylogenetic structure of the prokaryotic domain: The primary kingdoms", *Proc. Natl. Acad. Sci. U. S. A.*, vol. 74, no. 11, Nov. 1977.

[164] N. R. Pace, "The small things can matter", *PLoS Biol.*, vol. 16, no. 8, e3000009, Aug. 2018.

[165] D. H. Parks, M. Chuvochina, C. Rinke, A. J. Mussig, P.-A. Chaumeil, and P. Hugenholtz, "GTDB: An ongoing census of bacterial and archaeal diversity through a

phylogenetically consistent, rank normalized and complete genome-based taxonomy", en, *Nucleic Acids Res.*, vol. 50, no. D1, pp. D785–D794, Jan. 2022.

[166] J. Handelsman, M. R. Rondon, S. F. Brady, J. Clardy, and R. M. Goodman, "Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products", *Chem. Biol.*, vol. 5, no. 10, Oct. 1998.

[167] C. Quince, A. W. Walker, J. T. Simpson, N. J. Loman, and N. Segata, "Shotgun metagenomics, from sampling to analysis", en, *Nat. Biotechnol.*, vol. 35, no. 9, pp. 833–844, Sep. 2017.

[168] O. Béjà, L. Aravind, E. V. Koonin, *et al.*, "Bacterial rhodopsin: Evidence for a new type of phototrophy in the sea", en, *Science*, vol. 289, no. 5486, pp. 1902–1906, Sep. 2000.

[169] J. C. Venter, K. Remington, J. F. Heidelberg, *et al.*, "Environmental genome shotgun sequencing of the sargasso sea", en, *Science*, vol. 304, no. 5667, pp. 66–74, Apr. 2004.

[170] "Metagenomics: Microbial diversity through a scratched lens", *Curr. Opin. Microbiol.*, vol. 15, no. 5, pp. 605–612, Oct. 2012.

[171] M. Kircher and J. Kelso, "High-throughput DNA sequencing – concepts and limitations", en, *Bioessays*, vol. 32, no. 6, pp. 524–536, Jun. 2010.

[172] A. Meyerdierks, M. Kube, I. Kostadinov, *et al.*, "Metagenome and mRNA expression analyses of anaerobic methanotrophic archaea of the ANME-1 group", en, *Environ. Microbiol.*, vol. 12, no. 2, pp. 422–439, Feb. 2010.

[173] J. Qin, R. Li, J. Raes, *et al.*, "A human gut microbial gene catalogue established by metagenomic sequencing", en, *Nature*, vol. 464, no. 7285, pp. 59–65, Mar. 2010.

[174] E. A. Dinsdale, R. A. Edwards, D. Hall, *et al.*, "Functional metagenomic profiling of nine biomes", en, *Nature*, vol. 452, no. 7187, pp. 629–632, Mar. 2008.

[175] M. L. Cuvelier, A. E. Allen, A. Monier, *et al.*, "Targeted metagenomics and ecology of globally important uncultured eukaryotic phytoplankton", en, *Proceedings of the National Academy of Sciences*, vol. 107, no. 33, pp. 14 679–14 684, Aug. 2010.

[176] A. Escalas, L. Hale, J. W. Voordeckers, *et al.*, “Microbial functional diversity: From concepts to applications”, en, *Ecol. Evol.*, vol. 9, no. 20, pp. 12 000–12 016, Oct. 2019.

[177] S. Louca, M. F. Polz, F. Mazel, *et al.*, “Function and functional redundancy in microbial systems”, *Nature ecology & evolution*, vol. 2, no. 6, Jun. 2018.

[178] C. T. Brown, L. A. Hug, B. C. Thomas, *et al.*, “Unusual biology across a group comprising more than 15% of domain bacteria”, *Nature*, vol. 523, no. 7559, Jul. 2015.

[179] M. A. H. J. van Kessel, D. R. Speth, M. Albertsen, *et al.*, “Complete nitrification by a single microorganism”, en, *Nature*, vol. 528, no. 7583, pp. 555–559, Nov. 2015.

[180] A. Spang, J. H. Saw, S. L. Jørgensen, *et al.*, “Complex archaea that bridge the gap between prokaryotes and eukaryotes”, *Nature*, vol. 521, no. 7551, May 2015.

[181] K. Zaremba-Niedzwiedzka, E. F. Caceres, J. H. Saw, *et al.*, “Asgard archaea illuminate the origin of eukaryotic cellular complexity”, *Nature*, vol. 541, no. 7637, Jan. 2017.

[182] J. Lloyd-Price, C. Arze, A. N. Ananthakrishnan, *et al.*, “Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases”, en, *Nature*, vol. 569, no. 7758, pp. 655–662, May 2019.

[183] J. F. Cryan, K. J. O’Riordan, K. Sandhu, V. Peterson, and T. G. Dinan, “The gut microbiome in neurological disorders”, *Lancet Neurol.*, vol. 19, no. 2, Feb. 2020.

[184] A. Jamiu and C. Chaguza, “Metagenomics to improve infectious disease diagnostics in low- and middle-income countries”, en, *Nature Microbiology*, vol. 9, no. 5, pp. 1150–1151, May 2024.

[185] K. Ratiner, D. Ciocan, S. K. Abdeen, and E. Elinav, “Utilization of the microbiome in personalized medicine”, en, *Nat. Rev. Microbiol.*, vol. 22, no. 5, pp. 291–308, Dec. 2023.

[186] L. P. Coelho, R. Alves, Á. R. Del Río, *et al.*, “Towards the biogeography of prokaryotic genes”, en, *Nature*, vol. 601, no. 7892, pp. 252–256, Jan. 2022.

[187] T. S. B. Schmidt, A. Fullam, P. Ferretti, *et al.*, “SPIRE: A searchable, planetary-scale microbiome REsource”, en, *Nucleic Acids Res.*, vol. 52, no. D1, pp. D777–D783, Oct. 2023.

[188] L. Richardson, B. Allen, G. Baldi, *et al.*, “MGnify: The microbiome sequence data analysis resource in 2023”, en, *Nucleic Acids Res.*, vol. 51, no. D1, pp. D753–D759, Dec. 2022.

[189] Á. Rodríguez del Río, J. Giner-Lamia, C. P. Cantalapiedra, *et al.*, “Functional and evolutionary significance of unknown genes from uncultivated taxa”, en, *Nature*, vol. 626, no. 7998, pp. 377–384, Dec. 2023.

[190] R. Sachdeva, B. J. Campbell, and J. F. Heidelberg, “Rare microbes from diverse earth biomes dominate community activity”, en, May 2019.

[191] S. Yooseph, K. H. Nealson, D. B. Rusch, *et al.*, “Genomic and functional adaptation in surface ocean planktonic prokaryotes”, en, *Nature*, vol. 468, no. 7320, pp. 60–66, Nov. 2010.

[192] A. P. Alivisatos, M. J. Blaser, E. L. Brodie, *et al.*, “A unified initiative to harness earth’s microbiomes”, en, *Science*, Oct. 2015.

[193] M. L. Sogin, H. G. Morrison, J. A. Huber, *et al.*, “Microbial diversity in the deep sea and the underexplored “rare biosphere””, en, *Proceedings of the National Academy of Sciences*, vol. 103, no. 32, pp. 12 115–12 120, Aug. 2006.

[194] M. D. J. Lynch and J. D. Neufeld, “Ecology and exploration of the rare biosphere”, en, *Nat. Rev. Microbiol.*, vol. 13, no. 4, pp. 217–229, Mar. 2015.

[195] M. Delgado-Baquerizo, D. J. Eldridge, V. Ochoa, B. Gozalo, B. K. Singh, and F. T. Maestre, “Soil microbial communities drive the resistance of ecosystem multifunctionality to global change in drylands across the globe”, en, *Ecol. Lett.*, vol. 20, no. 10, pp. 1295–1305, Oct. 2017.

[196] S. E. Evans and M. D. Wallenstein, “Climate change alters ecological strategies of soil bacteria”, en, *Ecol. Lett.*, vol. 17, no. 2, pp. 155–164, Feb. 2014.

[197] D. R. Mende, I. Letunic, O. M. Maistrenko, *et al.*, *Progenomes2: An improved database for accurate and consistent habitat, taxonomic and functional annotations of prokaryotic genomes*, 2019.

[198] W. Li, K. R. O'Neill, D. H. Haft, *et al.*, "RefSeq: Expanding the prokaryotic genome annotation pipeline reach with protein family model curation", en, *Nucleic Acids Res.*, vol. 49, no. D1, pp. D1020–D1028, Jan. 2021.

[199] D. R. Mende, S. Sunagawa, G. Zeller, and P. Bork, *Accurate and universal delineation of prokaryotic species*, 2013.

[200] D. H. Parks, M. Chuvochina, D. W. Waite, *et al.*, *A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life*, 2018.

[201] S. R. Eddy, "Accelerated profile HMM searches", en, *PLoS Comput. Biol.*, vol. 7, no. 10, e1002195, Oct. 2011.

[202] S. El-Gebali, J. Mistry, A. Bateman, *et al.*, "The pfam protein families database in 2019", en, *Nucleic Acids Res.*, vol. 47, no. D1, pp. D427–D432, Jan. 2019.

[203] A. Krogh, B. Larsson, G. von Heijne, and E. L. L. Sonnhammer, *Predicting transmembrane protein topology with a hidden markov model: Application to complete genomes11edited by f. cohen*, 2001.

[204] J. Huerta-Cepas, K. Forslund, L. P. Coelho, *et al.*, *Fast Genome-Wide functional annotation through orthology assignment by eggNOG-Mapper*, 2017.

[205] M. Steinegger and J. Söding, *MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets*, 2017.

[206] F. Sievers, A. Wilm, D. Dineen, *et al.*, *Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega*, 2011.

[207] L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, and B. Q. Minh, "IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies", en, *Mol. Biol. Evol.*, vol. 32, no. 1, pp. 268–274, Jan. 2015.

[208] J. Huerta-Cepas, F. Serra, and P. Bork, "ETE 3: Reconstruction, analysis, and visualization of phylogenomic data", en, *Mol. Biol. Evol.*, vol. 33, no. 6, pp. 1635–1638, Jun. 2016.

[209] R. Borges, J. P. Machado, C. Gomes, A. P. Rocha, and A. Antunes, *Measuring phylogenetic signal between categorical traits and phylogenies*, 2019.

[210] F. D. Ciccarelli, T. Doerks, C. von Mering, C. J. Creevey, B. Snel, and P. Bork, "Toward automatic reconstruction of a highly resolved tree of life", en, *Science*, vol. 311, no. 5765, pp. 1283–1287, Mar. 2006.

[211] A. Milanese, D. R. Mende, L. Paoli, *et al.*, "Microbial abundance, activity and population genomic profiling with mOTUs2", en, *Nat. Commun.*, vol. 10, no. 1, p. 1014, Mar. 2019.

[212] M. N. Price, P. S. Dehal, and A. P. Arkin, "FastTree 2–approximately maximum-likelihood trees for large alignments", en, *PLoS One*, vol. 5, no. 3, e9490, Mar. 2010.

[213] R. D. Finn, P. Coggill, R. Y. Eberhardt, *et al.*, "The pfam protein families database: Towards a more sustainable future", en, *Nucleic Acids Res.*, vol. 44, no. D1, pp. D279–D285, Dec. 2015.

[214] S. C. Potter, A. Luciani, S. R. Eddy, Y. Park, R. Lopez, and R. D. Finn, "HMMER web server: 2018 update", en, *Nucleic Acids Res.*, vol. 46, no. W1, W200–W204, Jun. 2018.

[215] C. Sanchis-López, J. P. Cerna-Vargas, S. Santamaría-Hernando, *et al.*, "Prevalence and specificity of chemoreceptor profiles in Plant-Associated bacteria", en, *mSystems*, vol. 6, no. 5, e0095121, Oct. 2021.

[216] A. Chklovski, D. H. Parks, B. J. Woodcroft, and G. W. Tyson, "CheckM2: A rapid, scalable and accurate tool for assessing microbial genome quality using machine learning", en, *Nat. Methods*, vol. 20, no. 8, pp. 1203–1212, Jul. 2023.

[217] A. Orakov, A. Fullam, L. P. Coelho, *et al.*, "GUNC: Detection of chimerism and contamination in prokaryotic genomes", *Genome Biol.*, vol. 22, no. 1, Jun. 2021.

[218] A. Fullam, I. Letunic, T. S. B. Schmidt, *et al.*, "Progenomes3: Approaching one million accurately and consistently annotated high-quality prokaryotic genomes", en, *Nucleic Acids Res.*, vol. 51, no. D1, pp. D760–D766, Nov. 2022.

[219] P.-A. Chaumeil, A. J. Mussig, P. Hugenholtz, and D. H. Parks, "GTDB-Tk v2: Memory friendly classification with the genome taxonomy database", en, *Bioinformatics*, vol. 38, no. 23, pp. 5315–5316, Oct. 2022.

[220] S. Deorowicz, A. Debudaj-Grabysz, and A. Gudyś, “FAMSA: Fast and accurate multiple sequence alignment of huge protein families”, en, *Sci. Rep.*, vol. 6, no. 1, pp. 1–13, Sep. 2016.

[221] Z. Deng, A. Hernández-Plaza, and J. Huerta-Cepas, “TreeProfiler: A command-line tool for computing and visualizing phylogenetic profiles against large trees”, en, Sep. 2023.

[222] H.-J. Ruscheweyh, A. Milanese, L. Paoli, *et al.*, “Cultivation-independent genomes greatly expand taxonomic-profiling capabilities of mOTUs across various environments”, en, *Microbiome*, vol. 10, no. 1, pp. 1–12, Dec. 2022.

[223] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, “Scikit-learn: Machine learning in python”, *J. Mach. Learn. Res.*, vol. 12, no. 85, pp. 2825–2830, 2011.

[224] D. R. Mende, I. Letunic, O. M. Maistrenko, *et al.*, “Progenomes2: An improved database for accurate and consistent habitat, taxonomic and functional annotations of prokaryotic genomes”, en, *Nucleic Acids Res.*, vol. 48, no. D1, pp. D621–D625, Jan. 2020.

[225] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, “CD-HIT: Accelerated for clustering the next-generation sequencing data”, en, *Bioinformatics*, vol. 28, no. 23, p. 3150, Dec. 2012.

[226] C. Camacho, G. Coulouris, V. Avagyan, *et al.*, “BLAST+: Architecture and applications”, *BMC Bioinformatics*, vol. 10, Dec. 2009.

[227] W. Suda, M. Oto, S. Amachi, H. Shinoyama, and M. Shishido, “A direct method to isolate DNA from phyllosphere microbial communities without disrupting leaf tissues”, en, *Microbes Environ.*, vol. 23, no. 3, pp. 248–252, 2008.

[228] K. Bethune, C. Mariac, M. Couderc, *et al.*, “Long-fragment targeted capture for long-read sequencing of plastomes”, en, *Appl. Plant Sci.*, vol. 7, no. 5, e1243, May 2019.

[229] H. Hale, E. M. Gardner, J. Viruel, L. Pokorny, and M. G. Johnson, “Strategies for reducing per-sample costs in target capture sequencing for phylogenomics and population genomics in plants”, en, *Appl. Plant Sci.*, vol. 8, no. 4, e11337, Apr. 2020.

[230] "Babraham bioinformatics - FastQC a quality control tool for high throughput sequence data", Accessed: 2024-4-15.

[231] P. Ewels, M. Magnusson, S. Lundin, and M. Käller, "MultiQC: Summarize analysis results for multiple tools and samples in a single report", en, *Bioinformatics*, vol. 32, no. 19, pp. 3047–3048, Jun. 2016.

[232] A. M. Bolger, M. Lohse, and B. Usadel, "Trimmomatic: A flexible trimmer for illumina sequence data", en, *Bioinformatics*, vol. 30, no. 15, p. 2114, Aug. 2014.

[233] D. Li, C.-M. Liu, R. Luo, K. Sadakane, and T.-W. Lam, "MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de bruijn graph", en, *Bioinformatics*, vol. 31, no. 10, pp. 1674–1676, Jan. 2015.

[234] B. Bushnell, "BBMap: A fast, accurate, Splice-Aware aligner", en, Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States), Tech. Rep. LBNL-7065E, Mar. 2014.

[235] H. Li, "Minimap2: Pairwise alignment for nucleotide sequences", en, *Bioinformatics*, vol. 34, no. 18, pp. 3094–3100, May 2018.

[236] P. Danecek, J. K. Bonfield, J. Liddle, *et al.*, "Twelve years of SAMtools and BCFtools", en, *Gigascience*, vol. 10, no. 2, giab008, Feb. 2021.

[237] J. Huerta-Cepas, D. Szklarczyk, D. Heller, *et al.*, "eggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses", en, *Nucleic Acids Res.*, vol. 47, no. D1, pp. D309–D314, Nov. 2018.

[238] F. Madeira, M. Pearce, A. R. N. Tivey, *et al.*, "Search and sequence analysis tools services from EMBL-EBI in 2022", en, *Nucleic Acids Res.*, vol. 50, no. W1, W276–W279, Jul. 2022.

[239] *GitHub - lh3/seqtk: Toolkit for processing sequences in FASTA/Q formats*, en, https://github.com/lh3/seqtk, Accessed: 2024-4-15.

[240] I. Letunic and P. Bork, "Interactive tree of life (iTOL) v6: Recent updates to the phylogenetic tree display and annotation tool", en, *Nucleic Acids Res.*, gkae268, Apr. 2024.

[241] C. L. Schoch, S. Ciufo, M. Domrachev, *et al.*, “NCBI taxonomy: A comprehensive update on curation, resources and tools”, en, *Database*, vol. 2020, Jan. 2020.

[242] M. W. Silby, C. Winstanley, S. A. C. Godfrey, S. B. Levy, and R. W. Jackson, *Pseudomonasgenomes: Diverse and adaptable*, 2011.

[243] J. I. Baldani, B. Pot, G. Kirchhof, *et al.*, *Emended description of herbaspirillum; inccusion of [pseudomonas] rubrisubalbicans, a mild plant pathogen, as herbaspirillum rubrisubalbicans comb. nov.; and classification of a group of clinical isolates (EF group 1) as herbaspirillum species 3*, 1996.

[244] S. Bi, D. Yu, G. Si, *et al.*, *Discovery of novel chemoeffectors and rational design of escherichia coli chemoreceptor specificity*, 2013.

[245] B. Ma, A. O. Charkowski, J. D. Glasner, and N. T. Perna, “Identification of host-microbe interaction factors in the genomes of soft rot-associated pathogens dickeya dadantii 3937 and pectobacterium carotovorum WPP14 with supervised machine learning”, en, *BMC Genomics*, vol. 15, p. 508, Jun. 2014.

[246] Y. Raoul des Essarts, J. Pédron, P. Blin, E. Van Dijk, D. Faure, and F. Van Gijsegem, “Common and distinctive adaptive traits expressed in dickeya dianthicola and dickeya solani pathogens when exploiting potato plant host”, en, *Environ. Microbiol.*, vol. 21, no. 3, pp. 1004–1018, Mar. 2019.

[247] J. Puławska, M. Kałużna, W. Warabieda, J. F. Pothier, M. Gétaz, and J. M. van der Wolf, “Transcriptome analysis of xanthomonas fragariae in strawberry leaves”, en, *Sci. Rep.*, vol. 10, no. 1, p. 20 582, Nov. 2020.

[248] C. Wei, T. Ding, C. Chang, C. Yu, X. Li, and Q. Liu, *Global regulator PhoP is necessary for motility, biofilm formation, exoenzyme production, and virulence of xanthomonas citri subsp. citri on citrus plants*, 2019.

[249] I. M. Matas, L. Lambertsen, L. Rodríguez-Moreno, and C. Ramos, “Identification of novel virulence genes and metabolic pathways required for full fitness of pseudomonas savastanoi pv. savastanoi in olive (olea europaea) knots”, en, *New Phytol.*, vol. 196, no. 4, pp. 1182–1196, Dec. 2012.

[250] J. Lacal, C. García-Fontana, F. Muñoz-Martínez, J.-L. Ramos, and T. Krell, *Sensing of environmental signals: Classification of chemoreceptors according to the size of their ligand binding regions*, 2010.

[251] K. Wuichet, R. P. Alexander, and I. B. Zhulin, *Comparative genomic and protein sequence analyses of a complex system controlling bacterial chemotaxis*, 2007.

[252] M. Y. Galperin, "What bacteria want", en, *Environ. Microbiol.*, vol. 20, no. 12, pp. 4221–4229, Dec. 2018.

[253] I. Gushchin, I. Melnikov, V. Polovinkin, *et al.*, "Mechanism of transmembrane signaling by sensor histidine kinases", en, *Science*, vol. 356, no. 6342, Jun. 2017.

[254] D. Martín-Mora, Á. Ortega, M. A. Matilla, S. Martínez-Rodríguez, J. A. Gavira, and T. Krell, "The molecular mechanism of nitrate chemotaxis via direct ligand binding to the PilJ domain of McpN", en, *MBio*, vol. 10, no. 1, Feb. 2019.

[255] J. Mistry, S. Chuguransky, L. Williams, *et al.*, "Pfam: The protein families database in 2021", en, *Nucleic Acids Res.*, vol. 49, no. D1, pp. D412–D419, Oct. 2020.

[256] The UniProt Consortium, A. Bateman, M.-J. Martin, *et al.*, "UniProt: The universal protein knowledgebase in 2023", en, *Nucleic Acids Res.*, vol. 51, no. D1, pp. D523–D531, Nov. 2022.

[257] "Finding novel chemoreceptors that specifically sense and trigger chemotaxis toward polycyclic aromatic hydrocarbons in novosphingobium pentaromativorans US6-1", *J. Hazard. Mater.*, vol. 416, p. 126 246, Aug. 2021.

[258] "The volatile organic compound acetoin enhances the colonization of azorhizobium caulinodans ORS571 on sesbania rostrata", *Sci. Total Environ.*, vol. 912, p. 169 006, Feb. 2024.

[259] C. D. Vavourakis, A.-S. Andrei, M. Mehrshad, R. Ghai, D. Y. Sorokin, and G. Muyzer, "A metagenomics roadmap to the uncultured genome diversity in hypersaline soda lake sediments", en, *Microbiome*, vol. 6, no. 1, pp. 1–18, Sep. 2018.

[260] Y. Sekiguchi, A. Ohashi, D. H. Parks, T. Yamauchi, G. W. Tyson, and P. Hugenholtz, "First genomic insights into members of a candidate bacterial phylum responsible for wastewater bulking", en, *PeerJ*, vol. 3, e740, Jan. 2015.

[261] D. R. Ortega, A. D. Fleetwood, T. Krell, C. S. Harwood, G. J. Jensen, and I. B. Zhulin, “Assigning chemoreceptors to chemosensory pathways in pseudomonas aeruginosa”, *Proceedings of the National Academy of Sciences of the United States of America*, vol. 114, no. 48, p. 12 809, Nov. 2017.

[262] A. R. Muok, D. R. Ortega, K. Kurniyati, *et al.*, “Atypical chemoreceptor arrays accommodate high membrane curvature”, *Nat. Commun.*, vol. 11, no. 1, Nov. 2020.

[263] R. Mo, S. Zhu, Y. Chen, Y. Li, Y. Liu, and B. Gao, “The evolutionary path of chemosensory and flagellar macromolecular machines in campylobacterota”, *PLoS Genet.*, vol. 18, no. 7, e1010316, Jul. 2022.

[264] M. A. Berry, E. P. Andrianova, and I. B. Zhulin, “Diverse domain architectures of CheA histidine kinase, a central component of bacterial and archaeal chemosensory systems”, en.

[265] Y. Sun, Z. Xie, F. Sui, X. Liu, and W. Cheng, “Identification of cbp1, a c-di-GMP binding chemoreceptor in azorhizobium caulinodans ORS571 involved in chemotaxis and nodulation of the host plant”, en, *Front. Microbiol.*, vol. 10, p. 436 858, Apr. 2019.

[266] K. Witek, F. Jupe, A. I. Witek, D. Baker, M. D. Clark, and J. D. G. Jones, “Accelerated cloning of a potato late blight–resistance gene using RenSeq and SMRT sequencing”, en, *Nat. Biotechnol.*, vol. 34, no. 6, pp. 656–660, Apr. 2016.

[267] L. Mamanova, A. J. Coffey, C. E. Scott, *et al.*, “Target-enrichment strategies for next-generation sequencing”, en, *Nat. Methods*, vol. 7, no. 2, pp. 111–118, Jan. 2010.

[268] L. Manoharan, S. K. Kushwaha, K. Hedlund, and D. Ahrén, “Captured metagenomics: Large-scale targeting of genes based on ‘sequence capture’ reveals functional diversity in soils”, en, *DNA Res.*, vol. 22, no. 6, pp. 451–460, Oct. 2015.

[269] T. Karamitros and G. Magiorkinis, “A novel method for the multiplexed target enrichment of MinION next generation sequencing libraries using PCR-generated baits”, en, *Nucleic Acids Res.*, vol. 43, no. 22, e152, Dec. 2015.

[270] J.-H. Han, S. Batey, A. A. Nickson, S. A. Teichmann, and J. Clarke, “The folding and evolution of multidomain proteins”, en, *Nature Reviews Molecular Cell Biology*, vol. 8, no. 4, pp. 319–330, Mar. 2007.

[271] Á. Ortega, I. B. Zhulin, and T. Krell, “Sensory repertoire of bacterial chemoreceptors”, en, *Microbiology and Molecular Biology Reviews*, Oct. 2017.

[272] J. Jumper, R. Evans, A. Pritzel, *et al.*, “Highly accurate protein structure prediction with AlphaFold”, en, *Nature*, vol. 596, no. 7873, pp. 583–589, Jul. 2021.

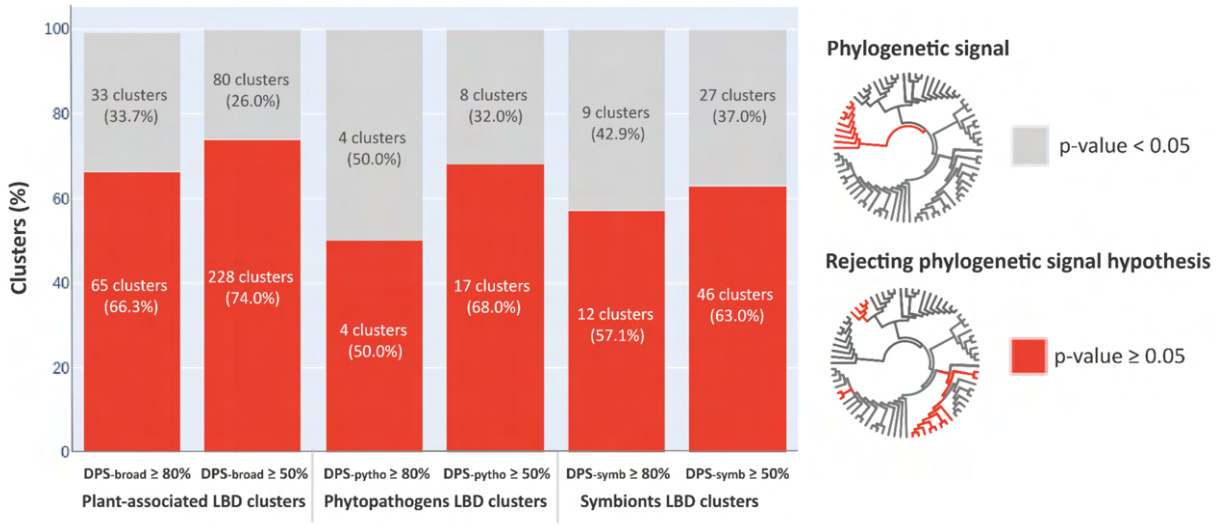
[273] G. Corso, H. Stärk, B. Jing, R. Barzilay, and T. Jaakkola, “DiffDock: Diffusion steps, twists, and turns for molecular docking”, Oct. 2022. eprint: 2210.01776.

# Annexes

# Annexe I

## Supplementary Material

# 1 Prevalence and Specificity of Chemoreceptors Profiles in Plant-Associated Bacteria




**Supplementary Figure S1.1: Phylogenetic signal detection in plant-associated LBD clusters using the phylogeny of chemotactic species tree.** Proportion of significant phylogenetic signal among LBD clusters enriched in PAB-broad, PAB-phyto and PAB-symb species within two thresholds (≥50%, ≥80%). The reference phylogeny used in this analysis excluded any species without CRs in their genome.

Given the dimensions of these Tables, they are available in an external link, or in the supplementary material found in the paper of this chapter.

Supplementary Table 1.1. **List of the bacterial TaxIds included in the Plant-Associated Bacteria (PAB) database.**

Supplementary Table 1.2. Delta-statistic values to calculate the phylogenetic signal of the LBD clusters.

Supplementary Table 1.3. List of the 5,546 species-representative genomes in the proGenomes2 database having at least one CR.

Supplementary Table 1.4. List of all CRs identified in 5,546 species-representative genomes obtained from the proGenomes2 database indicating their cluster classification, and the amino acid coordinates for their predicted domains.

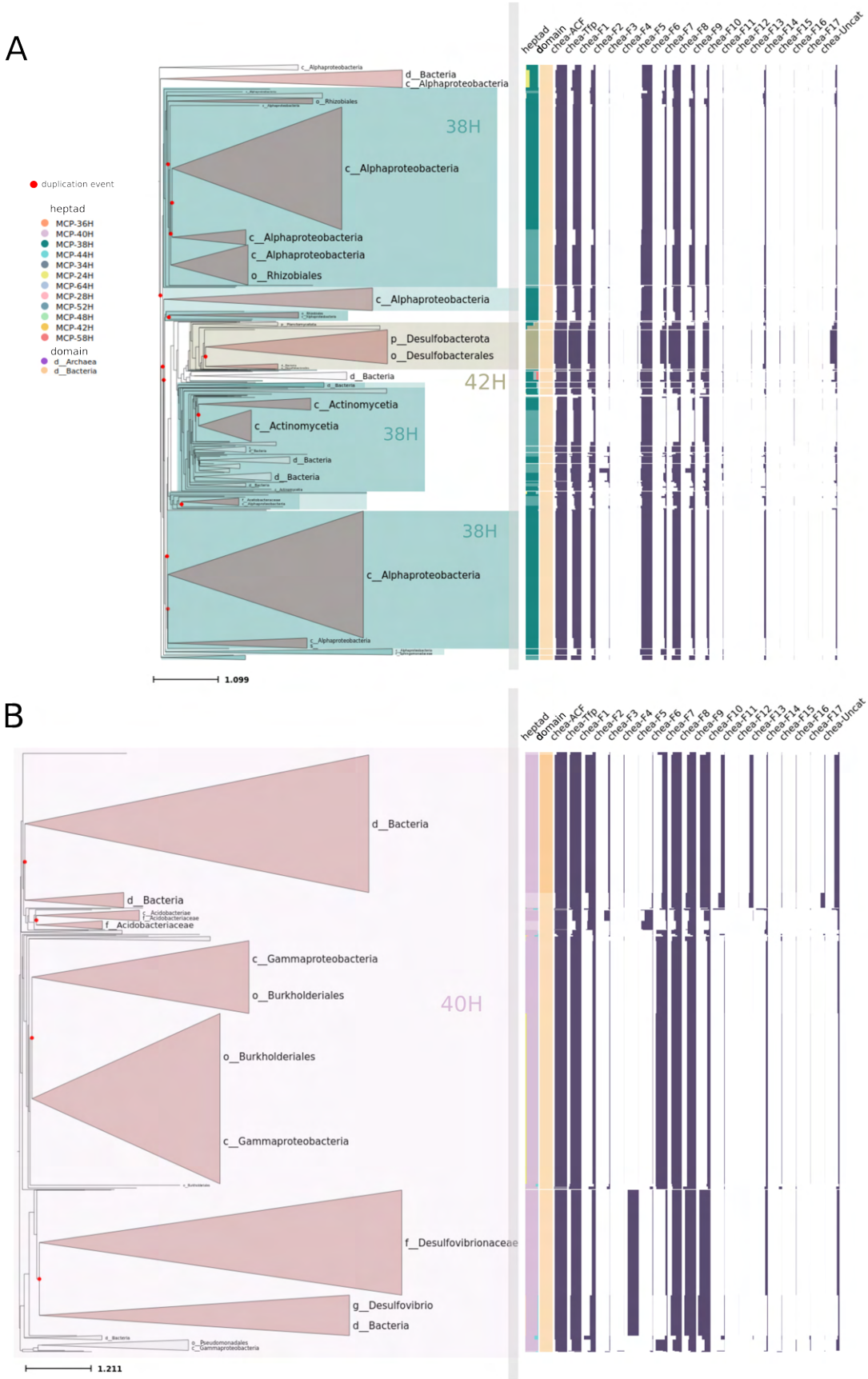
Supplementary Table 1.5. Distribution of LBD types among 'compact' LBD clusters (at least 90% of their sequences sharing the same LBD type).

Supplementary Table 1.6. List of the 5,149 clusters of LBDs indicating their DPS-broad, DPS-phyto

and DPS-symb.

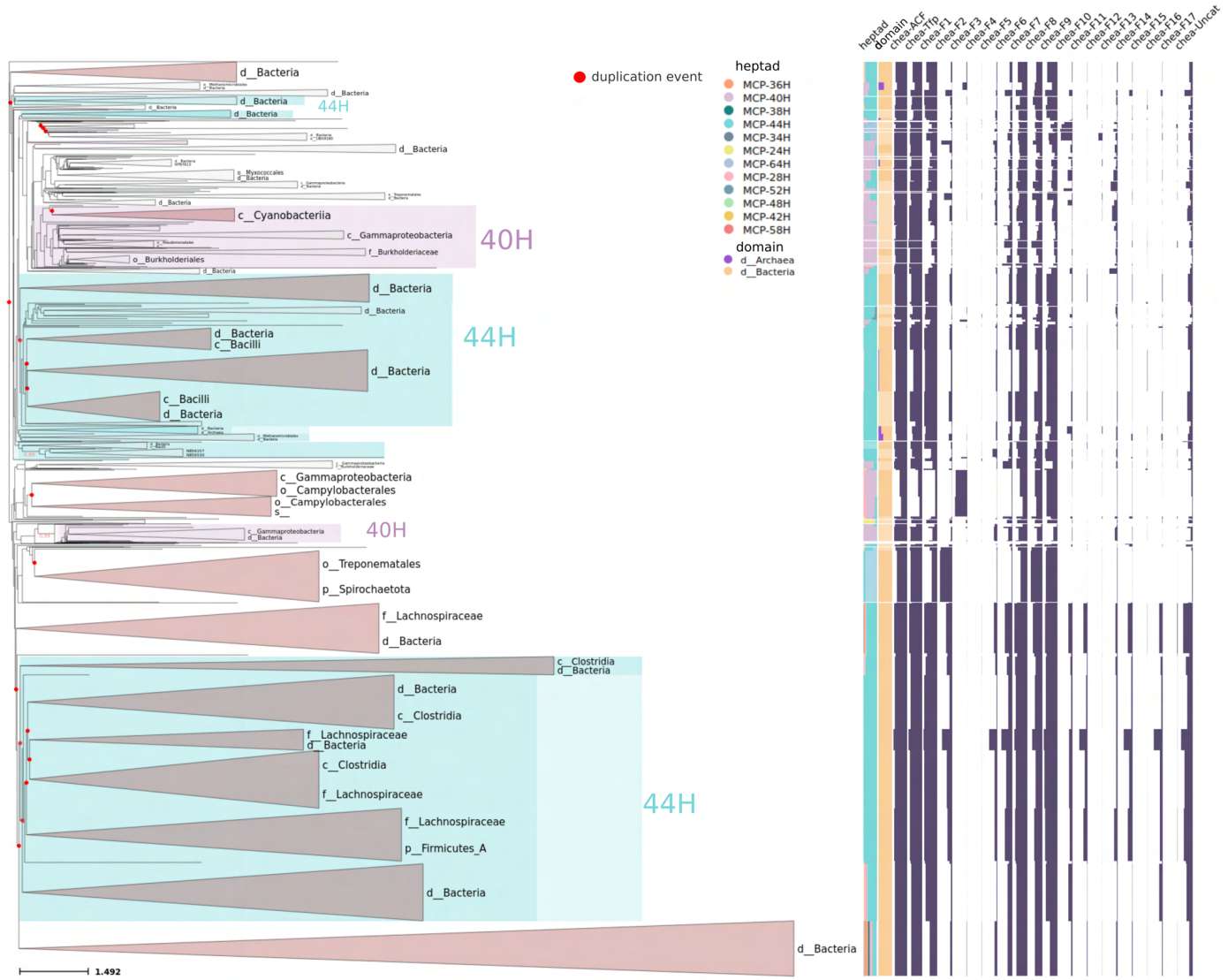
Supplementary Table 1.7. Distribution of LBD types and domain models in the total LBD dataset.

# 2 Adaptive value and evolutionary history of bacterial chemoreceptors revealed by global metagenomic survey




**Supplementary Figure S2.1: Detail on phylogenetic tree reconstruction of CRs across globalCR (Figure 2.2).**

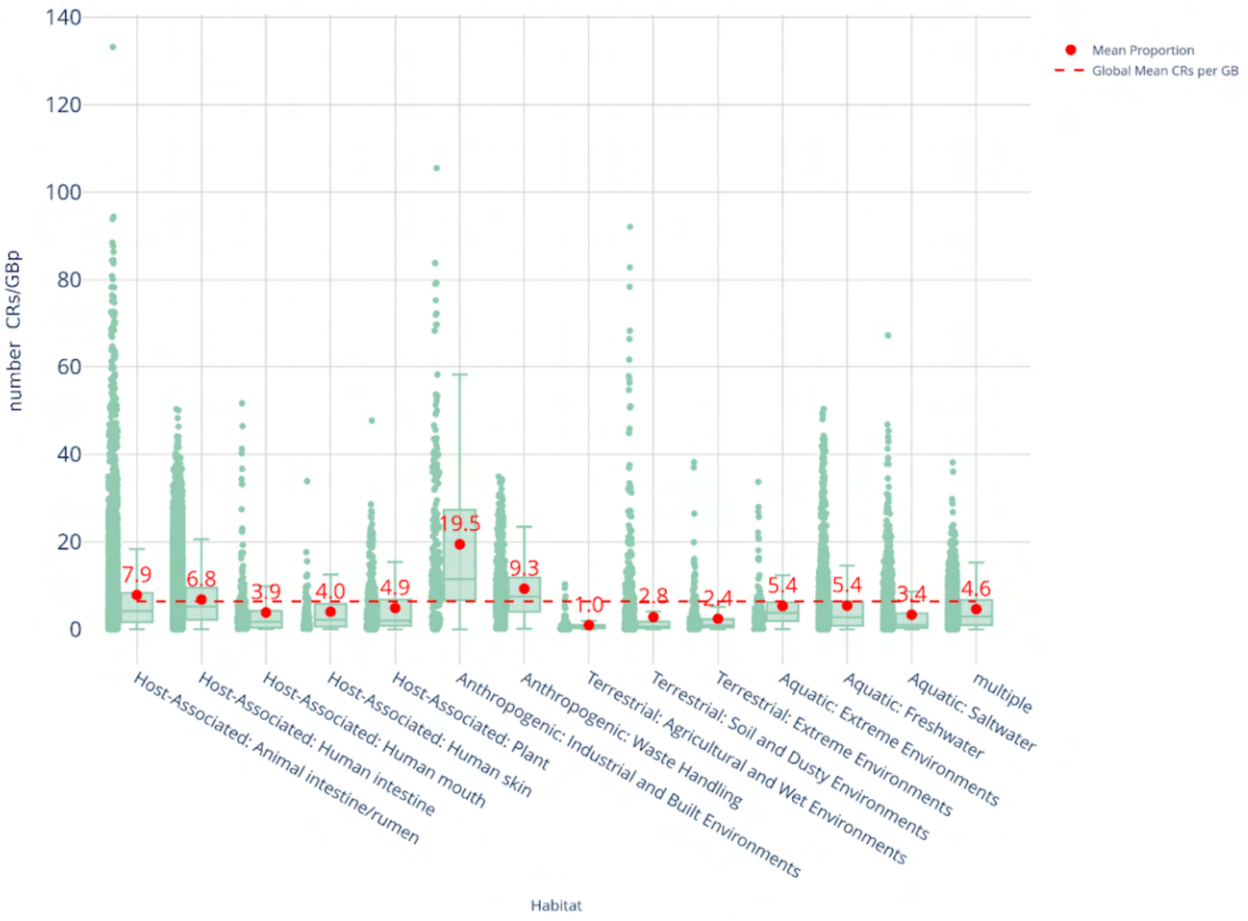
C




**Supplementary Figure S2.1: Detail on phylogenetic tree reconstruction of CRs across globalCR (Figure 2.2).**




**Supplementary Figure S2.2: F1 score of Random Forest Classifier of different dataset across environments.** The average number of the F1 score and the standard error bars is calculated for each 10 iteration of the subsampling for each biome and dataset.




**Supplementary Figure S2.3: Enrichment of sequenced chemoreceptors across biomes.** Box plots of the ratio number of CRs sequenced per giga base pairs of each sample distributed across biomes. The mean within each habitat is plotted in red dots, and the mean of all the samples is plotted in the red dash line (6.4 CRs/Gb).




**Supplementary Figure S2.4: Specificity of sensing domains in HDBSCAN predicted clusters of sensing capabilities in grained ecological labels.**

**Supplementary Table S2.1: Sample distribution of SPIRE.**

| | SPIRE total count samples | Samples with at least one CR | % of samples with at least one CR |
|---|---|---|---|
| **Host-Associated: Human intestine** | 47206 | 22268 | 47.17 |
| **Host-Associated: Animal intestine/rumen** | 8172 | 3901 | 47.74 |
| **Aquatic: Saltwater** | 7175 | 1730 | 24.11 |
| **Aquatic: Freshwater** | 6126 | 1606 | 26.22 |
| **Anthropogenic: Industrial and Built Environments** | 5217 | 600 | 11.50 |
| **Host-Associated: Human mouth** | 4506 | 459 | 10.19 |
| **Host-Associated: Human skin** | 3858 | 167 | 4.33 |
| **Terrestrial: Soil and Dusty Environments** | 3805 | 1254 | 32.96 |
| **multiple** | 2507 | 1480 | 59.03 |
| **Host-Associated: Animal others** | 1554 | 20 | 1.29 |
| **Host-Associated: Human others** | 1170 | 118 | 10.09 |
| **Host-Associated: Plant** | 1154 | 429 | 37.18 |
| **Anthropogenic: Waste Handling** | 1074 | 654 | 60.89 |
| **Terrestrial: Extreme Environments** | 922 | 338 | 36.66 |
| **Air** | 906 | 39 | 4.30 |
| **Terrestrial: Agricultural and Wet Environments** | 529 | 160 | 30.25 |
| **Aquatic: Extreme Environments** | 440 | 137 | 31.14 |
| **Anthropogenic: Food Production** | 248 | 48 | 19.35 |
| **Host-Associated: Animal skin** | 48 | 5 | 10.42 |

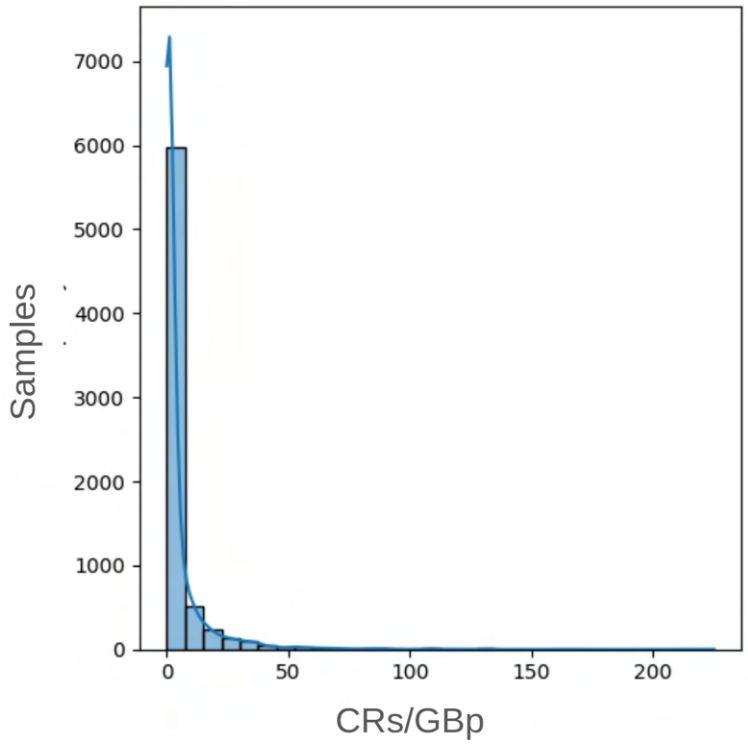
**Supplementary Table S2.2: Distribution of heptad subtypes across globalCR.**

| | % Presence across all species | nº seqs | % total seqs |
|---|---|---|---|
| **MCP-36H** | 39,39 | 355521 | 21,77 |
| **MCP-40H** | 53,10 | 237921 | 14,57 |
| **MCP-44H** | 46,00 | 805012 | 49,31 |
| **MCP-38H** | 17,23 | 73882 | 4,53 |
| **MCP-34H** | 18,02 | 69044 | 4,23 |
| **MCP-48H** | 2,13 | 26118 | 1,60 |
| **MCP-24H** | 11,07 | 24966 | 1,53 |
| **MCP-28H** | 10,97 | 26033 | 1,59 |
| **MCP-42H** | 2,13 | 5667 | 0,35 |
| **MCP-64H** | 3,91 | 3715 | 0,23 |
| **MCP-52H** | 0,92 | 2609 | 0,16 |
| **MCP-58H** | 0,79 | 2221 | 0,14 |

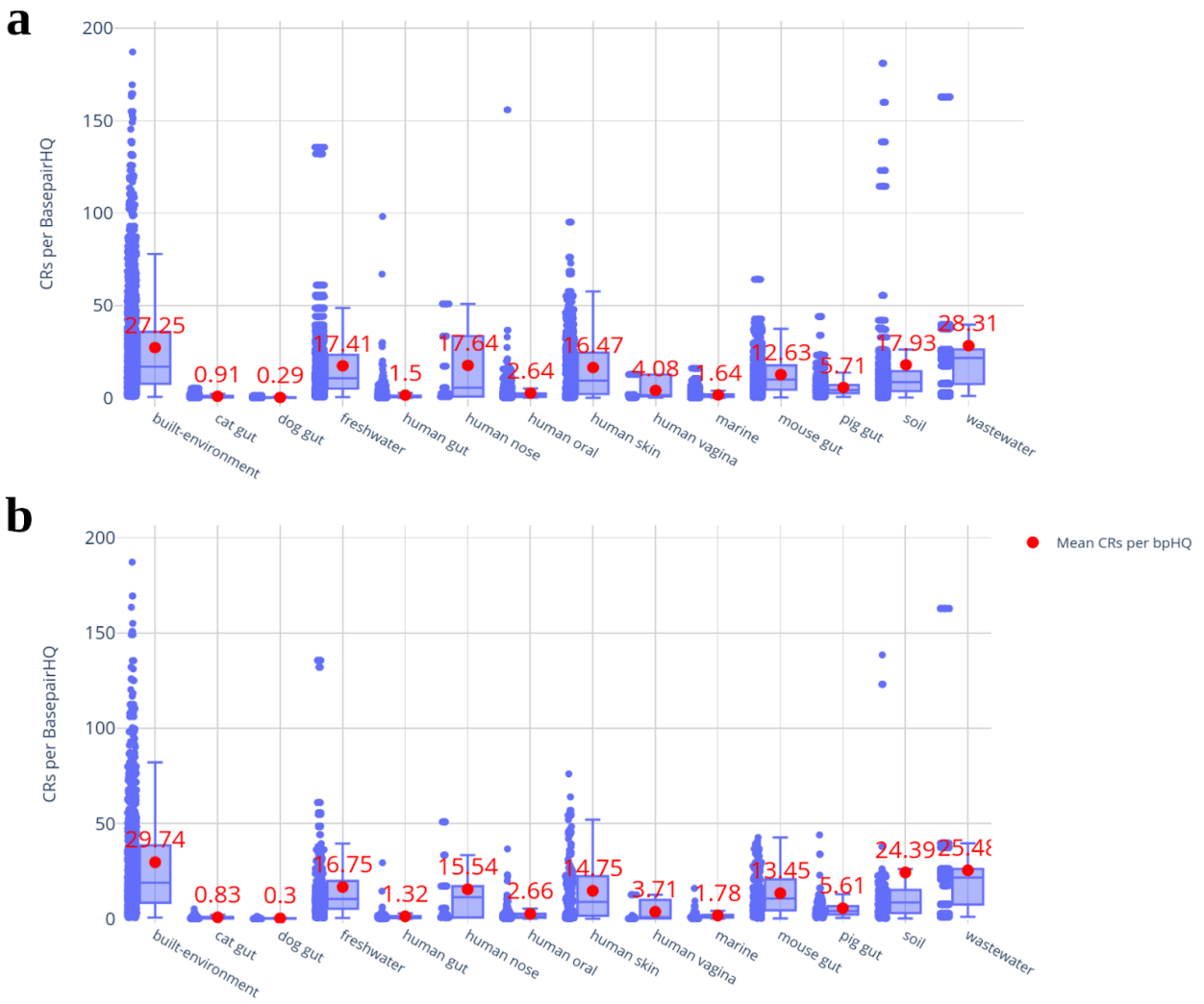
**Supplementary Table S2.3: Mean number and presence of CheA subtypes across globalCR.**

| | Mean Count | % Presence across all species |
|---|---|---|
| **chea-ACF** | 3.81 | 87.32 |
| **chea-Tfp** | 1.90 | 72.94 |
| **chea-F7** | 1.09 | 61.64 |
| **chea-F1** | 0.55 | 42.40 |
| **chea-F9** | 0.43 | 34.90 |
| **chea-F8** | 0.41 | 33.54 |
| **chea-F6** | 0.20 | 19.30 |
| **chea-F5** | 0.19 | 17.46 |
| **chea-Uncat** | 0.13 | 13.68 |
| **chea-F13** | 0.08 | 7.72 |
| **chea-F2** | 0.07 | 7.72 |
| **chea-F10** | 0.05 | 5.22 |
| **chea-F3** | 0.01 | 1.54 |
| **chea-F14** | 0.01 | 1.04 |
| **chea-F4** | 0.01 | 0.59 |
| **chea-F12** | 0.00 | 0.50 |
| **chea-F17** | 0.00 | 0.40 |
| **chea-F15** | 0.00 | 0.42 |
| **chea-F16** | 0.00 | 0.49 |
| **chea-F11** | 0.00 | 0.22 |

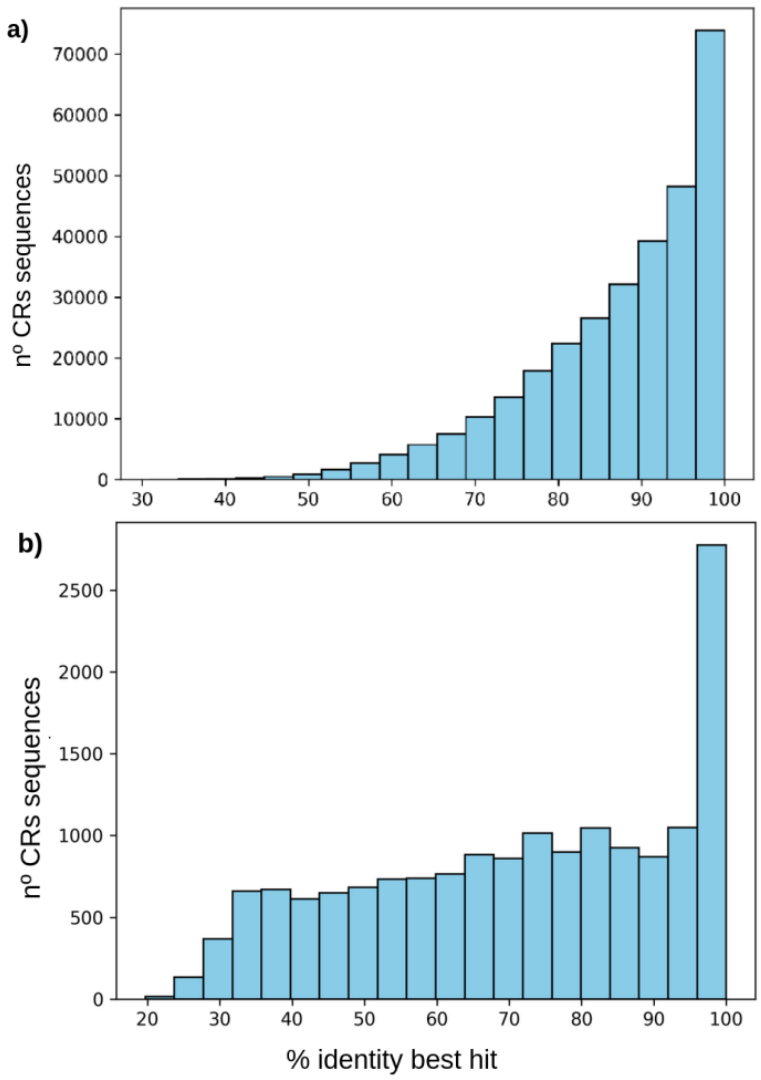
# 3 Unveiling the repertoire of microbial chemoreceptor genes in the rare biosphere



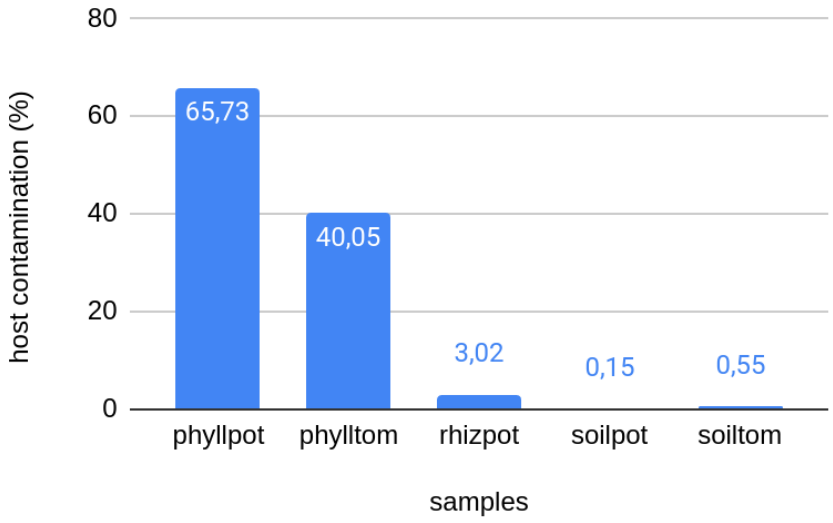

**Supplementary Figure S3.1: Number of CR genes assembled per Gbp in 13,174 metagenomes from GMGCv1.** The histogram represents the distribution of mean CR genes detected per Gbp across the metagenomics samples found in GMGCv1.




**Supplementary Figure S3.2: CR count normalization in metaCR.** a) CRs per GBp calculation after subsampling 100 times the same sequencing depth (samples until fulfilling 50 GBp) in each habitat. b) CRs per GBp after subsampling 100 times the same number of samples (20) in each habitat. Outliers (>200 CRs/GBp) are not shown.




**Supplementary Figure S3.3: Identity distribution of the best-hit against GTDB.** a) Pool-2 CRs are mapped using BLASTp against MCPsignal domains from Genome Taxonomy Database r207. b) Pool-2 putative LBDs are mapped using BLASTp against LBDs extracted from Genome Taxonomy Database r207. The identity distribution of the best hit is shown.




**Supplementary Figure S3.4: Host-plant contamination rate.** The bars show the percentage of reads aligning with the reference genomes of the host plants (tomato, *Solanum lycopersicum* L; potato, *Solanum tuberosum* L) for the samples from Pool-1.

**Supplementary Table S3.1: Summary statistics for sequencing Pool-1 and Pool-2.**

| Pool1 | target capture + Oxford Nanopore sequencing | | | | shotgun sequencing | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **sample** | **reads** | **pb** | **MCPsignal domains** | **MCPsignal pb** | **reads** | **pb** | **MCPsignal domains** | **MCPsignal pb** | **fold increase** |
| **humangut** | **140.399** | **40,793,491** | **3.395** | **2,004,391** | **151,168,388** | **22,675,258,200** | **121** | **49.209** | **28** |
| **lake** | **141.691** | **41,650,504** | **253** | **145.928** | **78,967,238** | **11,845,085,700** | **91** | **32.712** | **3** |
| **phyllpot** | **276.886** | **101,326,889** | **14.414** | **9,836,668** | **71,637,176** | **10,745,576,400** | **423** | **150.033** | **34** |
| **phylltom** | **296.52** | **101,830,341** | **8.923** | **5,981,389** | **80,550,122** | **12,082,518,300** | **8** | **2.373** | **1.115** |
| **rhizpot1** | **274.051** | **103,969,085** | **7.468** | **5,197,515** | **71,816,876** | **10,772,531,400** | **286** | **91.371** | **26** |
| rhizpot2 | 170.252 | 36,273,989 | 1.981 | 1,308,885 | - | - | - | - | - |
| rhizpot3 | 121.341 | 45,084,923 | 3.266 | 1,946,865 | - | - | - | - | - |
| **rumen** | **700.151** | **232,653,990** | **9.429** | **5,270,820** | **59,018,526** | **8,852,778,900** | **56** | **17.502** | **168** |
| soilpot1 | 58.635 | 23,801,088 | 904 | 631.645 | - | - | - | - | - |
| soilpot2 | 8.413 | 2,386,966 | - | - | - | - | - | - | - |
| **soilpot3** | **43.302** | **10,570,087** | **302** | **155.773** | **54,453,648** | **8,168,047,200** | **59** | **17.088** | **5** |
| soiltom1 | 169.528 | 50,592,717 | 2.393 | 1,507,102 | - | - | - | - | - |
| soiltom2 | 359.38 | 139,338,169 | 6.437 | 4,770,078 | - | - | - | - | - |
| **soiltom3** | **591.649** | **221,563,132** | **13.103** | **8,344,121** | **71,749,210** | **10,762,381,500** | **162** | **52.044** | **81** |
| **Pool2** | | | | | | | | | |
| **rhizpot1** | **1,775,465** | **2,327,937,974** | **62.509** | **6,983,813,922** | **68,290,510** | **10,311,867,010** | **202** | **20,951,441,766** | **309** |
| **rhizpot2** | **279.457** | **374,497,672** | **10.947** | **1,123,493,016** | **72,484,148** | **10,945,106,348** | **249** | **3,370,479,048** | **44** |
| **rhizpot3** | **2,165,422** | **3,191,533,103** | **50.491** | **9,574,599,309** | **75,411,306** | **11,387,107,206** | **185** | **28,723,797,927** | **273** |
| **rhiztom1** | **1,588,117** | **1,931,610,176** | **59.336** | **5,794,830,528** | **68,478,212** | **10,340,210,012** | **253** | **17,384,491,584** | **235** |
| **rhiztom2** | **1,331,120** | **1,648,128,976** | **48.033** | **4,944,386,928** | **71,252,910** | **10,759,189,410** | **264** | **14,833,160,784** | **182** |
| **rhiztom3** | **1,914,166** | **2,376,522,004** | **78.056** | **7,129,566,012** | **71,627,784** | **10,815,795,384** | **307** | **21,388,698,036** | **254** |

**Supplementary Table S3.2: Number of captured MCPsignal domains clustered at different identity thresholds.**

| % id aa of clustering | nº MCPsignal Pool-1 | nº MCPsignal Pool-2 |
|---|---|---|
| **100** | 57.065 | 287.271 |
| **98** | 56.78 | 269.694 |
| **95** | 55.972 | 256.652 |
| **90** | 54.628 | 243.211 |
| **0** | 57.39 | 309.372 |

# Annexe II

AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems®

RESEARCH ARTICLE

# Prevalence and Specificity of Chemoreceptor Profiles in Plant-Associated Bacteria

Claudia Sanchis-López,[a] Jean Paul Cerna-Vargas,[a] Saray Santamaría-Hernando,[a] Cayo Ramos,[b] Tino Krell,[c] Pablo Rodríguez-Palenzuela,[a,d] Emilia López-Solanilla,[a,d] Jaime Huerta-Cepas,[a] José J. Rodríguez-Herva[a,d]

[a]Centro de Biotecnología y Genómica de Plantas (CBGP), Universidad Politécnica de Madrid (UPM) - Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA), Madrid, Spain

[b]Área de Genética, Facultad de Ciencias, Instituto de Hortofruticultura Subtropical y Mediterránea La Mayora, Universidad de Málaga-Consejo Superior de Investigaciones Científicas (IHSM-UMA-CSIC), Málaga, Spain

[c]Department of Environmental Protection, Estación Experimental del Zaidín, Consejo Superior de Investigaciones Científicas, Granada, Spain

[d]Departamento de Biotecnología-Biología Vegetal, Escuela Técnica Superior de Ingeniería Agronómica, Alimentaria y de Biosistemas, Universidad Politécnica de Madrid (UPM), Madrid, Spain

**ABSTRACT** Chemosensory pathways are among the most abundant prokaryotic signal transduction systems, allowing bacteria to sense and respond to environmental stimuli. Signaling is typically initiated by the binding of specific molecules to the ligand binding domain (LBD) of chemoreceptor proteins (CRs). Although CRs play a central role in plant-microbiome interactions such as colonization and infection, little is known about their phylogenetic and ecological specificity. Here, we analyzed 82,277 CR sequences from 11,806 representative microbial species covering the whole prokaryotic phylogeny, and we classified them according to their LBD type using a *de novo* homology clustering method. Through phylogenomic analysis, we identified hundreds of LBDs that are found predominantly in plant-associated bacteria, including several LBDs specific to phytopathogens and plant symbionts. Functional annotation of our catalogue showed that many of the LBD clusters identified might constitute unknown types of LBDs. Moreover, we found that the taxonomic distribution of most LBD types that are specific to plant-associated bacteria is only partially explained by phylogeny, suggesting that lifestyle and niche adaptation are important factors in their selection. Finally, our results show that the profile of LBD types in a given genome is related to the lifestyle specialization, with plant symbionts and phytopathogens showing the highest number of niche-specific LBDs. The LBD catalogue and information on how to profile novel genomes are available at https://github.com/compgenomicslab/CRs.

**IMPORTANCE** Considering the enormous variety of LBDs at sensor proteins, an important question resides in establishing the forces that have driven their evolution and selection. We present here the first clear demonstration that environmental factors play an important role in the selection and evolution of LBDs. We were able to demonstrate the existence of LBD families that are highly enriched in plant-associated bacteria but show a wide phylogenetic spread. These findings offer a number of research opportunities in the field of single transduction, such as the exploration of similar relationships in chemoreceptors of bacteria with a different lifestyle, like those inhabiting or infecting the human intestine. Similarly, our results raise the question whether similar LBD types might be shared by members of different sensor protein families. Lastly, we provide a comprehensive catalogue of CRs classified by their LBD region that includes a large number of putative new LBD types.

**KEYWORDS** MCP, chemoreceptor, chemotaxis, methyl-accepting chemotaxis protein, plant-associated bacteria

**Citation** Sanchis-López C, Cerna-Vargas JP, Santamaría-Hernando S, Ramos C, Krell T, Rodríguez-Palenzuela P, López-Solanilla E, Huerta-Cepas J, Rodríguez-Herva JJ. 2021. Prevalence and specificity of chemoreceptor profiles in plant-associated bacteria. mSystems 6:e00951-21. https://doi.org/10.1128/mSystems.00951-21.

**Editor** Marnix Medema, Wageningen University

**Copyright** © 2021 Sanchis-López et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license.

Address correspondence to Jaime Huerta-Cepas, huerta.jaime@inia.es, or José J. Rodríguez-Herva, jj.rodriguez@upm.es.

**Received** 12 August 2021
**Accepted** 2 September 2021
**Published** 21 September 2021



Downloaded from https://journals.asm.org/journal/msystems on 03 December 2024 by 83.36.16.221.

To ensure cell survival, bacteria have to adapt to changing environmental conditions (1). For this, bacterial cells are equipped with an array of different signal transduction systems that sense different environmental stimuli, such as osmolarity, oxygen tension, temperature, pH, light, nutrients, toxins, and other chemicals (2). Chemosensory pathways represent one of the primary bacterial signal transduction mechanisms, and more than half of all the bacterial genomes contain signaling genes (3). Most chemosensory pathways appear to mediate chemotaxis (3), whereas others have been associated with type IV pilus-based motility (4) or alternative cellular functions such as the control of second messenger levels (4, 5).

In a canonical chemosensory pathway, signals are perceived by binding specific molecules to the ligand binding domain (LBD) of chemoreceptors (CRs), which modulates the activity of the CheA autokinase and the subsequent transphosphorylation to the CheY response regulator. In canonical CRs, the extracytosolic LBD is flanked by two transmembrane (TM) regions, a cytosolic HAMP domain, and a signaling domain (MCPsignal). While the CR signaling domain (MCPsignal) is highly conserved, LBDs are rapidly evolving domains (6), which reflects the wide variety of chemoeffectors to be sensed. To date, more than 80 different LBD families have been identified (7, 8), and new types of LBDs continue to be discovered (9). The thermodynamic parameters for ligand binding to the individual CRs are very similar to those for binding to specific LBDs (10, 11), supporting the idea that the molecular determinants for signal recognition by CRs are located in the LBD. Further evidence of this came from the construction of chimeric receptors recombining LBDs with other signaling domains (e.g., autokinase domains), where the LBD was proved to define the function of the chimera (12, 13). Thus, while the conserved MCPsignal domain can be used to identify CRs, their LBDs allow them to be classified on the basis of their function (7, 8).

On the other hand, there is evidence suggesting that the genomic repertory of CRs is related to bacterial lifestyle (14, 15). For instance, it has been shown that plant-associated bacteria (PAB) possess a particularly large number of CRs (8, 16), indicating that chemosensory signaling is indeed an important requisite for plant-bacterium interactions. This is of particular relevance for plant pathogens and symbionts, for which it has been shown that flagellum-mediated chemotaxis is required for optimal virulence or symbiosis establishment (17–25). Plants represent complex habitats for colonization by different kinds of microorganisms, and PAB species can colonize the plant rhizosphere, phyllosphere, or endosphere (26). Motile sensory behavior has been shown to play a key role in the establishment of plant-microbe interactions, since bacteria that can sense and rapidly navigate toward niches optimal for growth and survival will have a clear competitive advantage (27–29). These considerations are valid for both pathogenic and nonpathogenic relationships between microorganisms and plants (8, 16). Similarly, microbial inhabitants of the phyllosphere, comprising the aerial part of plants, have to deal with the challenges of life on leaf surfaces, where flagellar motility confers advantages in terms of epiphytic fitness (30). The epiphytic lifestyle also represents the initial stage of foliar colonization by many bacterial phytopathogens, preceding entry into the leaf apoplast via wounds or natural plant openings (e.g., stomata) (30). However, despite their biological significance, the function and cognate signal have been determined for only a limited number of CRs from PAB, and very little information exists on their phylogenetic and ecological specificity.

In order to study those LBD types most tightly coupled to the plant-associated lifestyle, here we comprehensively identified the CR genes in all known bacterial lineages and classified them according to their LBDs, with a particular focus on the LBD types linked to a plant-associated lifestyle. As such, we employed a novel *de novo* methodology to extract putative LBD regions from all CR sequences and group them into homology-based clusters (i.e., putative LBD types). This analysis allowed us to identify hundreds of LBD types highly specific for PAB species, many of them unknown. We further found that the taxonomic distribution of the majority of PAB-specific LBD clusters is only partially explained by phylogeny, suggesting that niche and host adaptation



Downloaded from https://journals.asm.org/journal/msystems on 03 December 2024 by 83.36.16.221.

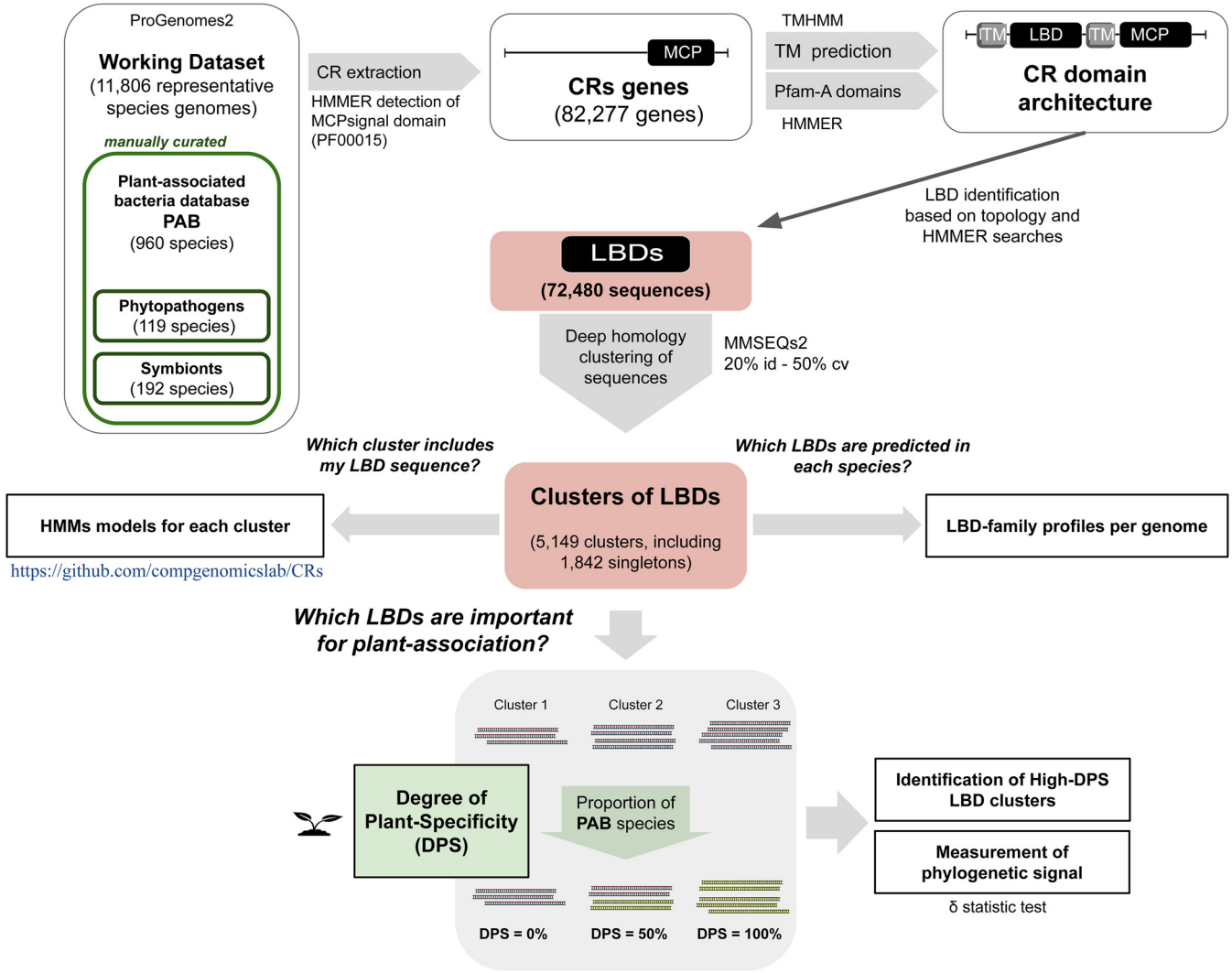



**FIG 1** Schematic view of the bioinformatics pipeline used to identify CRs that are potentially relevant for plant association. From a set of 11,806 representative prokaryotic genomes, 82,277 protein sequences were mined using HMM-based searches against the MCPsignal Pfam domain (PF00015). CR topology was analyzed by predicting transmembrane regions (TMs) and Pfam domains. Based on the topological analysis, LBD regions were predicted and a set of 72,480 LBD sequences was obtained. Clustering of LBDs based on sequence homology (20% minimum sequence identity with at least 50% sequence coverage) resulted in 5,149 clusters or subfamilies of LBDs, of which 1,842 contained a single sequence. To study a possible link between the LBD profiles and plant-associated lifestyle, a manually curated subset of 960 representative species of plant-associated bacteria (PAB) was generated, including phytopathogen (119) and symbiont (192) subsets. The determination of the proportion of PAB LBDs present in each cluster allowed us to assign the degree of plant specificity (DPS) value for each LBD subfamily. Subsequent analysis of high-DPS clusters identified LBD clusters that are potentially important for bacterium-plant associations. Furthermore, the validation of the high-DPS clusters as good ecological indicators was corroborated by measuring their phylogenetic signal. A detailed step-by-step description of the process can be found in Materials and Methods.

might have played relevant roles for their selection. Together, these results form a solid basis for the design of experiments aimed at identifying CRs that are essential for plant-microbe interactions and virulence.

## RESULTS

**Towards a global catalogue of chemoreceptors in plant-associated bacteria.** In order to maximize the coverage of our analysis, we first built a comprehensive catalogue of CRs detected across the entire prokaryotic phylogeny (Fig. 1). Species genomes were retrieved from the proGenomes v2 databases (31). Unlike the NCBI Taxonomy database, which is not an authoritative source for nomenclature or classification (32), proGenomes2 data do not rely on taxonomic names to identify species. Instead, each species-representative genome in proGenomes is delineated based on the evolutionary distances calculated between universally conserved genes present in



Downloaded from https://journals.asm.org/journal/msystems on 03 December 2024 by 83.36.16.221.

**TABLE 1** List of PAB with the highest number of predicted CRs

| TaxId | Biosample | Representative species | No. of CRs | PAB-phyto[a] | PAB-symb[b] |
|---|---|---|---|---|---|
| 1078773 | SAMN04334956 | *Herbaspirillum rubrisubalbicans* M1 | 67 | | |
| 1144319 | SAMN00839627 | *Herbaspirillum* sp. CF444 | 67 | | |
| 964 | SAMN03779333 | *Herbaspirillum seropedicae* | 65 | | |
| 193 | SAMN02982994 | *Azospirillum lipoferum* | 65 | | |
| 286727 | SAMN02982917 | *Azospirillum oryzae* | 64 | | |
| 346179 | SAMN03785417 | *Herbaspirillum rhizosphaerae* | 62 | | |
| 864073 | SAMN02471292 | *Herbaspirillum frisingense* GSF30 | 62 | | |
| 237610 | SAMN05860868 | *Pseudomonas psychrotolerans* | 60 | | |
| 288000 | SAMN02598359 | *Bradyrhizobium* sp. BTAi1 | 60 | | S |
| 92645 | SAMN06130964 | *Herbaspirillum frisingense* | 59 | | |
| 1175306 | SAMN02469572 | *Herbaspirillum* sp. GW103 | 59 | | |
| 1121033 | SAMN02440867 | *Azospirillum halopraeferens* DSM 3675 | 58 | | |
| 169679 | SAMN05170519 | *Clostridium saccharobutylicum* | 58 | | |
| 29438 | SAMN03837775 | *Pseudomonas savastanoi* | 57 | P | |
| 1262470 | SAMN03010392 | *Herbaspirillum hiltneri* N3 | 55 | | |
| 582667 | SAMN05192568 | *Methylobacterium pseudosasicola* | 54 | | |
| 50340 | SAMN05216581 | *Pseudomonas fuscovaginae* | 54 | | |
| 1001585 | SAMN02603190 | *Pseudomonas mendocina* NK-01 | 54 | | |
| 1749078 | SAMN04216969 | *Pseudomonas* sp. EpS/L25 | 53 | | |
| 1190415 | SAMN05216593 | *Pseudomonas asturiensis* | 53 | | |
| 50340 | SAMN03100370 | *Pseudomonas fuscovaginae* | 53 | | |
| 129140 | SAMN03976254 | *Pseudomonas syringae* pv. tagetis | 52 | P | |
| 294 | SAMN04992557 | *Pseudomonas fluorescens* | 52 | | |
| 1855289 | SAMN05216319 | *Duganella* sp. CF402 | 52 | | |
| 1144342 | SAMN00839653 | *Herbaspirillum* sp. YR522 | 52 | | |
| 47885 | SAMN03365871 | *Pseudomonas oryzihabitans* | 51 | | |
| 205918 | SAMN02604347 | *Pseudomonas syringae* pv. syringae B728a | 51 | P | |
| 1907416 | SAMN05880558 | *Aeromonas* sp. RU39B | 51 | | |
| 693986 | SAMN03075686 | *Methylobacterium oryzae* CBMB20 | 50 | | |
| 1736267 | SAMN04151647 | *Pseudomonas* sp. Leaf127 | 50 | | |
| 114615 | SAMEA3138227 | *Bradyrhizobium* sp. ORS 278 | 50 | | S |
| 1028989 | SAMD00019511 | *Pseudomonas* sp. StFLB209 | 50 | | |
| 80867 | SAMN04009978 | *Acidovorax avenae* | 50 | P | |
| 1122963 | SAMN02440654 | *Pleomorphomonas oryzae* DSM 16300 | 50 | | |
| 223283 | SAMN02604017 | *Pseudomonas syringae* pv. tomato DC3000 | 49 | P | |
| 1245469 | SAMD00061052 | *Bradyrhizobium oligotrophicum* S58 | 49 | | S |

[a]P, phytopathogen.
[b]S, plant symbiont.

nearly all organisms (32, 33). To establish links between CRs and the plant-associated lifestyle, we compiled three manually curated databases of PAB (see Materials and Methods): (i) PAB-broad, a reference database of 960 organisms found in multiple plant environments including leaves, roots, and rhizospheric soil; (ii) PAB-phyto, a subset database of 119 species including only known phytopathogens; and (iii) PAB-symb, which groups 192 plant symbionts. Using HMM-based searches, we then mined all the sequences containing the MCPsignal domain in the 11,806 species-representative genomes from the proGenomes database, compiling a global catalogue of 82,277 CR sequences from 5,546 genomes (see Data Set S3 in the supplemental material). This confirms the broad distribution of CRs, with 47% of the representative genomes containing at least one chemotactic receptor.

PAB species possessed almost twice as many CRs per genome (22.86) as those species not classified as plant associated (12.94), with the subset of phytopathogens showing the highest number (27.29). No CRs were predicted in 178 out of the 960 PAB genomes, indicating that more than 81% of PABs possess at least one CR gene, a percentage largely superior to the bacterial average (47%). From all species considered in this study, 36 PAB genomes stood out by their high content of CRs (Table 1), most notably the following: (i) 14 genomes from the *Pseudomonas* genus (49 to 60 CRs), including the well-known plant pathogens *P. syringae* and *P. savastanoi* (34), and (ii) 9 genomes from the *Herbaspirillum* genus (52 to 67 CRs), a group of betaproteobacteria that endophytically colonize gramineous species, thereby promoting their growth (35).



Downloaded from https://journals.asm.org/journal/msystems on 03 December 2024 by 83.36.16.221.

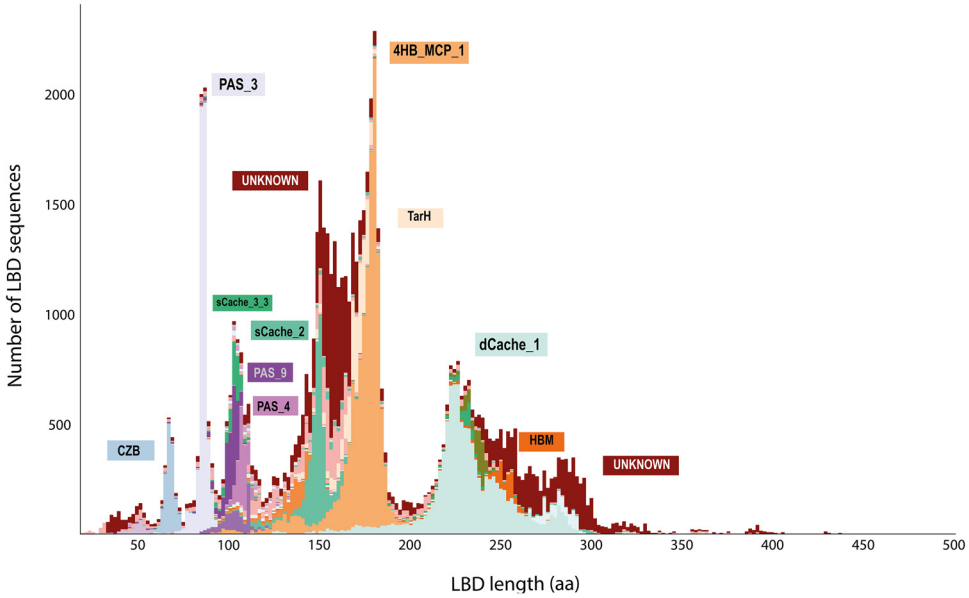



**FIG 2** Length distribution of the LBDs. The analysis was conducted on 72,480 LBDs, and the predominant LBD types within each of the main peaks are indicated. Only LBDs shorter than 500 amino acids (aa) are represented.

**Classifying chemoreceptors according to their ligand binding domain.** As the ecological relevance of CRs is mostly defined by their LBD region, we explored whether sequence segments corresponding to the LBD, rather than the full-length CR sequences, were related to a plant-associated lifestyle. To maximize the number of LBD sequences included in our analysis and not limit this to known LBD types from the Pfam database (7), we inferred LBDs based on the domain architecture of each CR. First, we extracted LBD sequences from the whole set of 82,277 CRs. Next, and given the high variability in the domains that could be considered LBDs, we identified putative LBDs using three different strategies: (i) detecting sequence regions matching any known domain other than the MCPsignal or HAMP, (ii) locating sequence regions flanked by two TM regions, and (iii) taking domains between the N-terminus and a single TM region. In total, we retrieved 72,480 putative LBD sequences, which could be fitted into three main groups based on their length (Fig. 2). The first group includes LBDs with a size between 60 and 110 amino acids, containing 21% of all the LBDs detected. The most abundant LBD family within this size range was PAS_3. The second group, comprising LBDs from 130 to 200 amino acids, contained over 45% of all LBDs and included 4HB_MCP_1 as the predominant family. The third group, comprising LBD lengths between 220 and 299 amino acids, covers 26% of all LBDs and has dCache_1 as the most abundant LBD family. Only 8% of all the LBDs detected fell outside these three size ranges, and the three most abundant LBDs were 4HB_MCP_1 (17.6%), dCache_1 (15.5%), and PAS_3 (9.2%).

We next investigated whether LBDs could be classified into broader sequence homology clusters, each representing a group of LBD sequences sharing a common evolutionary origin. Using relaxed homology thresholds (E value $\leq 10^{-3}$, 50% coverage, 20% amino acid identity), we grouped all 72,480 LBD sequences into 5,149 family clusters (Data Set S4), of which 3,307 contain more than 1 sequence. This *de novo* clustering approach might not be adequate for a detailed functional characterization of LBDs, as single residue changes have been shown to modify LBD ligand affinities (36–38). Nevertheless, each of our LBD clusters could be interpreted as an independent LBD type, with implicit levels of functional and ecological conservation. In fact, our approach consistently recovered all known LBD types and distributed them into 2,068 compact clusters where 90% of their members belonged to the same Pfam domain family (Table S1). Moreover, our clustering strategy allowed us to split large LBD



Downloaded from https://journals.asm.org/journal/msystems on 03 December 2024 by 83.36.16.221.

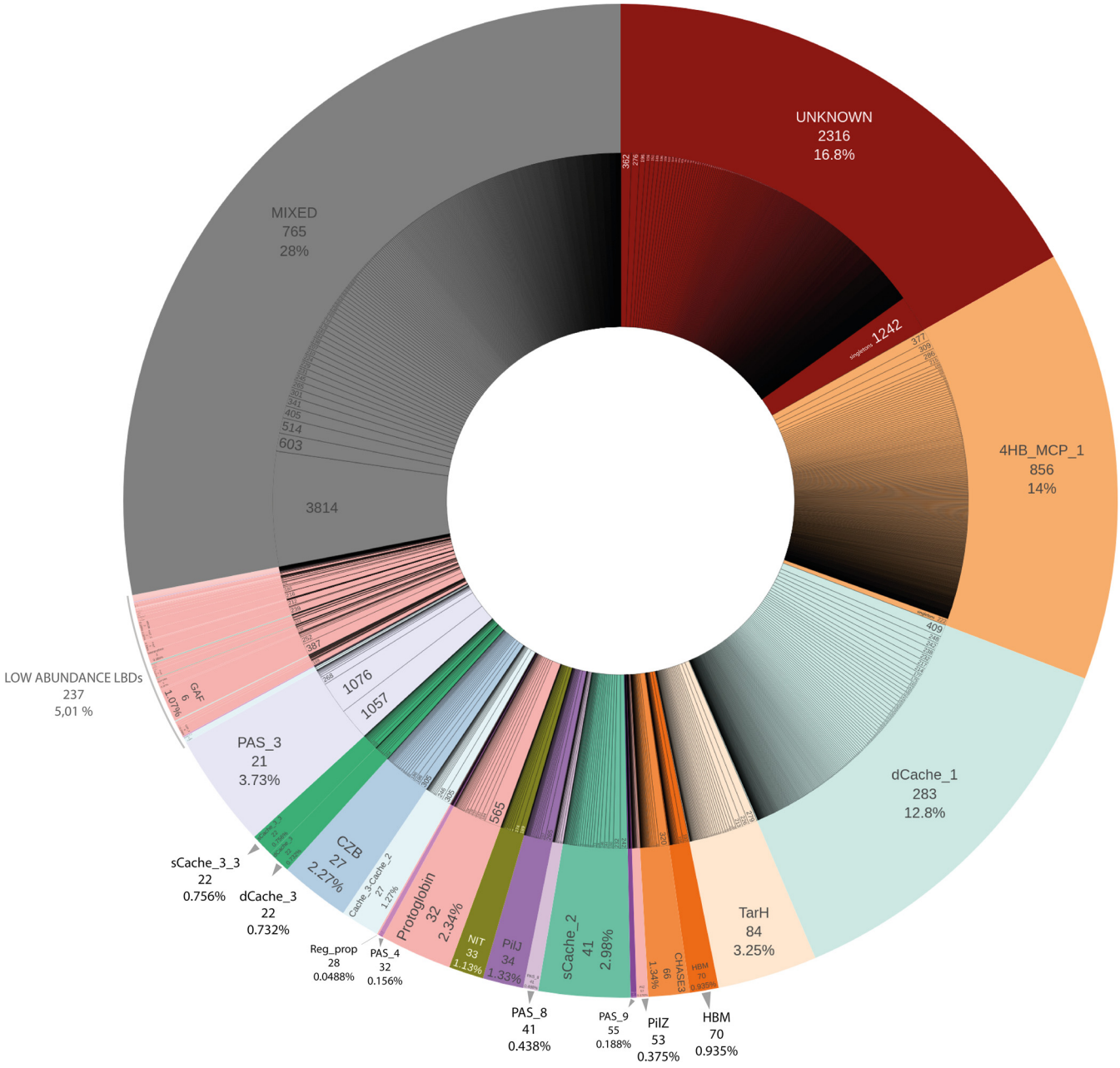



**FIG 3** Visual representation of the abundance of the LBD families. The outer donut of the chart represents the distribution of each LBD type and its relative abundance (in percentage of sequences), and the number of clusters with at least 90% of their sequences sharing the same LBD type, as defined by the Pfam signature. The LBDs are sorted according to the number of clusters within each LBD type. The inner donut of the chart represents all the clusters included within each LBD category, indicating the number of sequences contained in each subfamily. All singletons are merged in the last section of each LBD type (e.g., LBDs classified as "Unknown" have 1,242 singletons, that is, clusters containing only one sequence). "Mixed clusters" are those that do not reach the 90% threshold of sequences with the same Pfam model per cluster. "Low-abundance LBDs" include those LBD types that group into fewer than 12 "compact clusters."

families into finely grained subcategories (Fig. 3). For example, despite 4HB_MCP and Cache-like being present at similar levels in the initial CR sequence database, the number of derived clusters differs significantly, namely, 20.9% for 4HB_MCP compared to 8.3% for Cache-like. In the case of 4HB_MCP_1, the 10,034 sequences group into 856 different clusters compared to the 283 clusters for the 9,162 dCache_1 sequences, indicating higher sequence conservation in the latter. The situation is even more drastic in the case of PAS_3 LBDs, where 2,675 sequences group into just 21 clusters (Table S1), indicating a very low degree of diversity.



Downloaded from https://journals.asm.org/journal/msystems on 03 December 2024 by 83.36.16.221.

**TABLE 2** CRs predicted to be involved in plant-bacterium interactions

| CR gene ID (in the original source) | LBD type | DPS (%) | DPS-phyto (%) | DPS-symb (%) | Cluster no. | Amino acid identity (%) to representative LBDs from the database | Bacterial species and strain | Reference |
|---|---|---|---|---|---|---|---|---|
| ABF-0014824 | TarH | 100 | 100 | 0 | 932 | 100 | *Dickeya dadantii* 3937 | 39 |
| ABF-0015168 | TarH | 68.09 | 40.43 | 0 | 179 | 100 | *Dickeya dadantii* 3937 | 39 |
| ABF-0016115 | HBM | 63.64 | 48.48 | 0 | 409 | 100 | *Dickeya dadantii* 3937 | 39 |
| ABF-0016585 | TarH | 51.09 | 14.60 | 0.73 | 42 | 100 | *Dickeya dadantii* 3937 | 39 |
| ABF-0017097 | Unknown | 80.00 | 80.00 | 0 | 838 | 100 | *Dickeya dadantii* 3937 | 39 |
| ABF-0017674 | 4HB_MCP_1 | 100 | 87.50 | 0 | 630 | 100 | *Dickeya dadantii* 3937 | 39 |
| ABF-0019851 | TarH | 61.22 | 42.86 | 0 | 123 | 100 | *Dickeya dadantii* 3937 | 39 |
| ABF-0019855 | TarH | 61.22 | 42.86 | 0 | 123 | 100 | *Dickeya dadantii* 3937 | 39 |
| ABF-0020431 | sCache_2 | 34.55 | 12.73 | 0 | 233 | 100 | *Dickeya dadantii* 3937 | 39 |
| DDI_0843 | dCache_1 | 38.61 | 10.13 | 1.27 | 51 | 100 | *Dickeya dianthicola* RNS04.9 | 40 |
| DDI_0932 | sCache_2 | 34.55 | 12.73 | 0 | 233 | 100 | *Dickeya dianthicola* RNS04.9 | 40 |
| DDI_1647 | TarH | 61.22 | 42.86 | 0 | 123 | 88.97 | *Dickeya dianthicola* RNS04.9 | 40 |
| DDI_1649 | TarH | 61.22 | 42.86 | 0 | 123 | 100 | *Dickeya dianthicola* RNS04.9 | 40 |
| DDI_2258 | HBM | 100 | 100 | 0 | 792 | 100 | *Dickeya dianthicola* RNS04.9 | 40 |
| DDI_4092 | 4HB_MCP_1 | 100 | 87.50 | 0 | 630 | 100 | *Dickeya dianthicola* RNS04.9 | 40 |
| ADT-0000027 | HBM | 63.64 | 48.48 | 0 | 409 | 96.03 | *Pectobacterium carotovorum* WPP14 | 39 |
| ADT-0000661 | sCache_2 | 38.46 | 12.82 | 0 | 160 | 99.30 | *Pectobacterium carotovorum* WPP14 | 39 |
| ADT-0001320 | TarH | 51.09 | 14.60 | 0.73 | 42 | 98.84 | *Pectobacterium carotovorum* WPP14 | 39 |
| ADT-0001602 | TarH | 56.99 | 29.03 | 0 | 116 | 94.15 | *Pectobacterium carotovorum* WPP14 | 39 |
| ADT-0001887 | TarH | 61.22 | 42.86 | 0 | 123 | 97.59 | *Pectobacterium carotovorum* WPP14 | 39 |
| ADT-0002104 | TarH | 100 | 100 | 0 | 932 | 100 | *Pectobacterium carotovorum* WPP14 | 39 |
| ADT-0003152 | TarH | 68.09 | 40.43 | 0 | 179 | 91.15 | *Pectobacterium carotovorum* WPP14 | 39 |
| ADT-0003245 | 4HB_MCP_1 | 100 | 87.50 | 0 | 630 | 97.40 | *Pectobacterium carotovorum* WPP14 | 39 |
| ADT-0003418 | Unknown | 80.00 | 80.00 | 0 | 838 | 95.60 | *Pectobacterium carotovorum* WPP14 | 39 |
| PSA3335_17610 | Unknown | 87.50 | 50.00 | 0 | 835 | 100 | *Pseudomonas savastanoi* NCPPB3335 | 43 |
| XAC1892 | Unknown | 100 | 100 | 0 | 846 | 86.77 | *Xanthomonas citri* subsp. *citri* XHG3 | 42 |
| XAC2448 | 4HB_MCP_1 | 39.23 | 14.62 | 0 | 77 | 98.88 | *Xanthomonas citri* subsp. *citri* XHG3 | 42 |
| NBC2815_01024 | 4HB_MCP_1 | 92.00 | 84.00 | 0 | 549 | 100 | *Xanthomonas fragariae* IPO 3485 | 41 |
| NBC2815_02005 | 4HB_MCP_1 | 60.53 | 42.11 | 0 | 353 | 100 | *Xanthomonas fragariae* IPO 3485 | 41 |
| NBC2815_02008 | 4HB_MCP_1 | 88.46 | 88.46 | 0 | 273 | 100 | *Xanthomonas fragariae* IPO 3485 | 41 |
| NBC2815_02009 | 4HB_MCP_1 | 82.14 | 75.00 | 0 | 340 | 100 | *Xanthomonas fragariae* IPO 3485 | 41 |

Notably, an important fraction (45%) of the LBD clusters inferred could not be confidently associated with any previous family of Pfam domains, since more than 90% of their LBD sequences did not match to any known domain signature, suggesting the existence of a large number of unknown LBD types.

**Identifying PAB-specific ligand binding domains.** To identify LBD families specific to a plant-associated lifestyle, we analyzed each LBD cluster and calculated the corresponding percentage of PAB species therein, which we referred to as the degree of plant specificity (DPS; see Materials and Methods). For each LBD cluster, we calculated three DPS values, based on three databases of PAB species: (i) DPS-broad, calculated based on the PAB-broad reference database; (ii) DPS-phyto, based on the PAB-phyto subset; and (iii) DPS-symb, using the PAB-symb subgroup as a reference. In all cases, the DPS values ranged from 0% (no LBD family observed in the corresponding PAB database) to 100% (the LBD cluster includes only species from a given reference PAB database). From the 3,307 LBD nonsingleton clusters, we identified 419 and 139 clusters with a DPS-broad score of $\geq$50%, and $\geq$80%, respectively. Similarly, many LBD clusters showed high specificity in the stricter PAB reference databases (Data Set S5).

To further validate our findings, we cross-linked our predictions with experimental data from previous studies (39–43). In particular, we found that CRs with increased expression *in planta*, and particularly those required for full bacterial virulence, belonged to high-DPS clusters (Table 2). This list includes CRs that are upregulated in *Dickeya dadantii* 3937 and *Pectobacterium carotovorum* WPP14, two soft-rot bacterial strains (39); *Dickeya dianthicola* RNS04.9, which grows on macerated potato tubers (40); and *Xanthomonas fragariae*, which grows on strawberry leaves (41). Similarly, we



Downloaded from https://journals.asm.org/journal/msystems on 03 December 2024 by 83.36.16.221.

**TABLE 3** Distribution of LBD types among clusters with high DPS (≥50%)

| LBD type | No. of clusters | % of clusters over total[a] | No. of LBD sequences with the indicated domain[b] | Avg no. of LBD sequences per cluster |
|---|---|---|---|---|
| Unknown | 243 | 41.75 | 2,053 | 8.45 |
| 4HB_MCP_1 | 151 | 25.95 | 2,642 | 17.50 |
| TarH | 26 | 4.47 | 1,042 | 40.08 |
| HBM | 23 | 3.95 | 516 | 22.43 |
| CHASE3 | 9 | 1.55 | 142 | 15.78 |
| PilZ | 7 | 1.20 | 39 | 5.57 |
| PAS_9 | 6 | 1.03 | 7 | 1.17 |
| sCache_2 | 4 | 0.69 | 245 | 61.25 |
| NIT | 4 | 0.69 | 79 | 19.75 |
| Protoglobin | 3 | 0.52 | 153 | 51 |
| PAS_8 | 3 | 0.52 | 9 | 3 |
| PAS_3 | 3 | 0.52 | 82 | 27.33 |
| dCache_1 | 3 | 0.52 | 151 | 50.33 |
| Cache_3-Cache_2 | 3 | 0.52 | 79 | 26.33 |
| PAS_4 | 2 | 0.34 | 4 | 2 |
| CHASE4 | 2 | 0.34 | 3 | 1.50 |
| Usher | 1 | 0.17 | 1 | 1 |
| Tox-URI2 | 1 | 0.17 | 1 | 1 |
| SURF1 | 1 | 0.17 | 1 | 1 |
| SOR_SNZ | 1 | 0.17 | 1 | 1 |
| sCache_3_3 | 1 | 0.17 | 2 | 2 |
| Porin_4 | 1 | 0.17 | 1 | 1 |
| Peripla_BP_5 | 1 | 0.17 | 1 | 1 |
| PAS_7 | 1 | 0.17 | 19 | 19 |
| PapC_N | 1 | 0.17 | 1 | 1 |
| Glyco_hydro_2_N | 1 | 0.17 | 1 | 1 |
| Glyco_hydro_106 | 1 | 0.17 | 1 | 1 |
| FHIPEP | 1 | 0.17 | 1 | 1 |
| DUF4077 | 1 | 0.17 | 5 | 5 |
| dCache_3 | 1 | 0.17 | 71 | 71 |
| CBS | 1 | 0.17 | 2 | 2 |
| Asparaginase | 1 | 0.17 | 1 | 1 |
| ABC_tran | 1 | 0.17 | 1 | 1 |
| 5TM-5TMR_LYT | 1 | 0.17 | 20 | 20 |
| Total | 510 | | 7,377 | |

[a]Percentages are calculated over the total number of LBD clusters with at least 90% of their sequences sharing the same LBD type. These total clusters comprise more than 88% of the total number of clusters in this work.
[b]Sum of the total number of sequences sharing the same domain type found in the indicated clusters.

found several CRs with very high DPS values (80%) that were shown to be relevant in *Xanthomonas citri* virulence (42) or required for fitness of *Pseudomonas savastanoi* pv. savastanoi in olive knots (43). Taken together, these data support the validity of our approach to identify CRs that are relevant for a plant-associated lifestyle.

Interestingly, we also found that many PAB-specific clusters (41.75%) are formed by proteins of unknown LBD type, suggesting the presence of a significant number of uncharacterized LBD types. Excluding unknown LBD families, the most common domains among high-DPS clusters are 4HB_MCP_1 (26%), TarH (4.5%), and HBM (4%) (Table 3). It is remarkable that the three domain families form four-helix bundle structures (37, 38). The case of the HBM and TarH domains is particularly interesting, as the majority of sequences that belonged to these categories concentrated in very few high-DPS clusters: 57.0% (516/906) of all HBM sequences are grouped into 23 high-DPS clusters, and 36.7% (1,042/2,840) of all TarH sequences are grouped into 26 high-DPS clusters. This indicates a strong association of the TarH and HBM domains with the plant-associated lifestyle. In contrast, despite being the second most abundant LBD in bacteria (Table S2), the dCache_1 domain was not very abundant in PAB.

**Phylogenetic versus ecological signal in PAB-specific ligand binding domains.** Intrigued by the potential ecological significance of PAB-specific LBD clusters, we



Downloaded from https://journals.asm.org/journal/msystems on 03 December 2024 by 83.36.16.221.

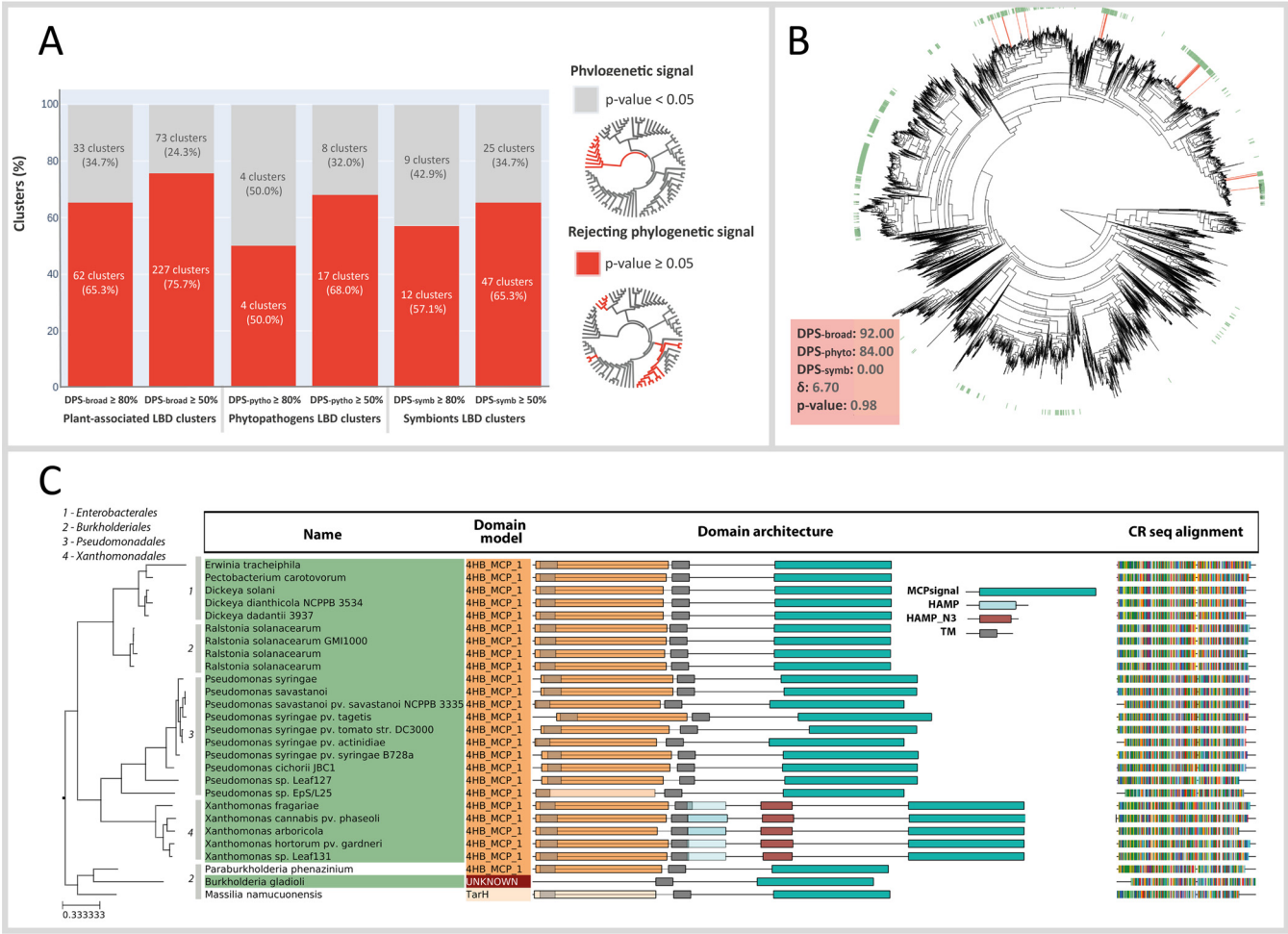



**FIG 4** Phylogenetic signal detection in plant-associated LBD clusters. (A) Proportion of the significant phylogenetic signal among LBD clusters enriched in PAB-broad, PAB-phyto, and PAB-symb species within two thresholds (≥50%, ≥80%). The significance through a *P* value test with 100 iterations, *P* value ≥0.05 rejects the null hypothesis of a phylogenetic signal (see Materials and Methods). (B) LBD cluster 549 (in red) distributed according to the chemotactic species phylogeny (5,763 representative species). The green leaves of the tree represent the PAB species. DPS, $\delta$, and *P* values for this LBD cluster are represented in the lower left box. (C) Phylogeny representation of LBD cluster 549, containing 27 LBD sequences distributed across 4 orders (numbered 1 to 4 in the tree). The domain architecture prediction is shown for each of the CRs.

further tested whether their taxonomic distribution is due to the phylogenetic signal of the underlying species, or if it might be driven by additional ecological factors. To address this issue, we reconstructed the complete phylogeny of the 11,806 species considered here (see Materials and Methods) and used it to assess the taxonomic distribution of each individual LBD cluster. Using the $\delta$-approach (44), we found that the majority (75.7%) of plant-associated LBD types (DPS ≥50%) did not follow the expected phylogenetic signal. In contrast, the taxonomic distribution of most PAB-specific LBDs was scattered over the global bacterial phylogeny (Fig. 4). This observation was consistent for the three PAB reference databases considered in this study, using stricter DPS cutoffs, and even when the species lacking CR genes were excluded from the analysis (Fig. S1).

Overall, the lack of phylogenetic signal for most of the LBD clusters, together with the fact that the LBDs tested are enriched in PAB species, suggests that the evolution of the sensory machinery of bacterial species might be at least partially driven by ecological pressures. This should allow the use of particular LBD clusters, even if functionally undefined, as lifestyle biomarkers. This issue is best illustrated by the LBD cluster 549 (Fig. 4B and C), which contains 27 CRs from broadly distributed bacterial families and orders, while retaining a high plant-association signal (DPS-broad >80%).



Downloaded from https://journals.asm.org/journal/msystems on 03 December 2024 by 83.36.16.221.

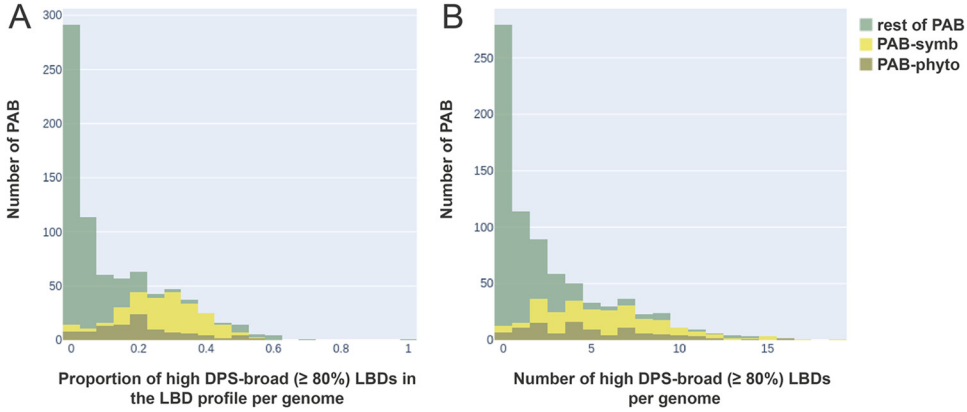



**FIG 5** Distribution of the high-plant-specificity LBDs in the PAB species profiles per genome. (A) Calculation of the proportion of high-DPS-broad (≥80%) LBDs in the total number of LBDs present in each species. The graph illustrates the distribution of the number of species according to the proportional ranges, plotting the species count as PAB-symb (yellow), PAB-phyto (brown), and the rest of PAB (green). (B) Absolute number of LBDs with a DPS-broad value of ≥80% in each PAB genome. The graph illustrates the distribution of the number of species as an absolute count of high-DPS LBD ranges. The species count is plotted as PAB-symb (yellow), PAB-phyto (brown), and the rest of PAB (green).

**LBD profiles per genome.** To investigate whether the profile of LBD clusters per genome could be informative about the plant-associated bacterial lifestyle, we studied the full repertoire of CRs among different PAB species. The genomes from the PAB species not only contained more CRs than those of non-PAB species, but also, many of their CRs could be considered highly specific to plant-related environments. In fact, assessing the LBD profiles per genome showed that microorganisms with a pronounced plant-associated lifestyle (i.e., PAB-symb and PAB-phyto) harbor more specific CRs than other PAB species (Fig. 5). On average, 28% and 20% of plant-symbiont and plant-phytopathogen CRs, respectively, are highly specific (DPS-broad >80%). In contrast, other PAB with a less pronounced plant-associated lifestyle, like nonsymbiont and nonphytopathogen plant-associated species, contained significantly fewer specific CRs (6%) (Fig. 5). Taken together, this information reinforces the idea that the repertoire of CRs has been partially shaped by niche adaptation, with more specialized adaptations leading to more specific CRs.

## DISCUSSION

In the present study, we carried out a comprehensive phylogenomic analysis of the full repertoire of CRs from a wide collection of microbial genomes, classifying them according to their LBDs. To maximize the representativeness of our study, we used more than 82,000 species-level CR sequences from 11,000 species-representative genomes, significantly expanding the scope of previous works (7, 15, 45), in terms of both the number of sequences examined and the phylogenetic coverage. To achieve this, we developed a novel method to extract LBDs and classified them based on a *de novo* homology-based clustering approach, departing from the traditional classification of CRs centered around their general protein topology (15, 45–47) or on known LBD domain searches (7). This approach allowed us to identify many new potential LBD types, suggesting that the chemosensing landscape remains largely unexplored. Additionally, we believe that our strategy delineating large LBD families into finely grained subcategories could provide further information (Fig. 3). Moreover, by classifying CRs based on their putative LBD type, for the first time we were able to quantify to what extent the chemosensory activity of PAB is linked to lifestyle.

Considering the enormous variety of LBDs at sensor proteins, establishing the forces that have driven their evolution is an important question that was never specifically addressed. To our knowledge, we present here the first clear demonstration showing that environmental factors play an important role in the selection and



Downloaded from https://journals.asm.org/journal/msystems on 03 December 2024 by 83.36.16.221.

evolution of LBDs. We found that the specificity of LBDs to a plant-associated lifestyle could not be explained by just a phylogenetic signal, since the taxonomic distribution of most PAB-specific LBD types was scattered over the microbial phylogeny, which at times covered different orders and phyla. This indicates that the selection of the certain CRs might indeed be guided by ecological factors, opening the possibility of identifying lifestyle biomarkers.

We also found that bacterial species more tightly associated with plant environments (such as plant symbionts and phytopathogens) tend to have stronger lifestyle specificity signals in their CR repertory. For instance, plant symbionts had the largest number of PAB-specific LBDs per genome, followed by phytopathogens, with both showing significantly higher ratios than generic soil microbiota. It appears likely that even stronger links between the chemosensory capabilities of bacteria and their lifestyle will be detected in the future as more data become available on new organisms (e.g., via metagenomics sequencing) and on their niche adaptation (i.e., plant-tissue specificity).

These findings thus offer a number of research opportunities in the field of signal transduction. First, it can be explored whether similar relationships can be observed in CRs of bacteria with a different lifestyle, such as for example those that inhabit or infect the human intestine. Another interesting issue that needs to be addressed is the question whether similar LBD types are shared by members of different sensor protein families. Major families of these receptors are sensor histidine kinases; chemoreceptors; adenylate, diadenylate, and diguanylate cyclases; and certain cAMP, c-di-AMP, and c-di-GMP phosphodiesterases, as well as Ser/Thr/Tyr protein kinases and phosphoprotein phosphatases (48). As the different sensor proteins of a given strain are exposed to the same signals, it appears plausible that the same LBD types might be present in members of different sensor protein families. Several examples have been reported in this direction, such as the specific sensing of nitrate by PilJ-type LBDs of the NarQ-type sensor kinases (49), the McpN chemoreceptor (50), and the PAS domain, universally found in different signal transduction systems (48). It would be of interest to estimate the global occurrence of such cases.

Overall, we believe that our study provides a comprehensive resource for future studies on bacterial chemoreception and that it sets the basis for the identification of novel CRs relevant for bacterium-plant interactions.

## MATERIALS AND METHODS

**Chemoreceptor (CR) sequence retrieval.** From the genomes of 11,806 representative species in the proGenomes2 database (31), 82,277 CR sequences were obtained. The representative species in proGenomes2 are the result of a phylogeny-based classification of all RefSeq (51) genomes, where species delineation is based on a systematic phylogenetic threshold (i.e., $<$95% divergence in 40 universal marker genes) rather than relying on the NCBI taxonomic names. Although this might lead to inconsistencies with the current NCBI Taxonomy names for strains and species, it better represents the genomic definition of species, as well as providing a standardized classification system (33, 52). To identify CRs in our set of representative genomes, all the sequences matching the MCPsignal Pfam domain signature (PF00015) were retrieved using HMMER 3.1b2 (53), Pfam-A 31.0 (54), and the specific gathering threshold provided for the MCPsignal HMM Pfam model. Multiple hits were resolved by retaining the match with the highest bit score. In analogy to previous studies (7, 55, 56), the presence of an MCPsignal domain in the sequence was the only criterion used for CR identification.

**Ligand binding domain (LBD) extraction.** For each CR sequence, transmembrane regions (TMs) were predicted using TMHMM2 (57). The position of the TM region(s) was used to infer the putative extracellular LBD regions, which were subsequently annotated using the Pfam domain database. When no significant Pfam matches were found, LBD sequences were labeled as "unknown." Two different topologies of extracellular LBDs were considered: (i) sequence regions flanked by two TM regions and (ii) sequence regions located between one TM and the N-terminal sequence. In both cases, sequences shorter than 30 amino acids were discarded. Intracellular LBD regions, as well as potentially overlooked extracellular LBDs (e.g., due to undetected TMs), were inferred based on the detection of Pfam domains other than the MCPsignal and HAMP domains. Pfam mappings were performed using HMMER (53) searches as implemented in eggNOG-mapper v.2.0.5 (53, 58). When more than two domains mapped to the same region, the best hit was selected. The final data set contained 72,480 LBD sequences.

**Clustering of LBD sequences.** *De novo* homology-based clustering of the 72,480 LBD sequences was inferred using MMseqs2 (59) with an E value threshold of 0.01, 20% minimum identity, and 50% minimum query coverage. These parameters were chosen to maximize remote homology detection and



Downloaded from https://journals.asm.org/journal/msystems on 03 December 2024 by 83.36.16.221.

to infer LBD clusters with broad phylogenetic divergence (i.e., distant homologues) while still grouping sequences with a common evolutionary origin. The MMseqs2 command used was "mmseqs cluster -c 0.2 –min-seq-id 0.5 –cov-mode 2".

**Construction of the databases for plant-associated bacteria (PAB).** A curated list of PAB was manually curated from the 11,806 representative species. As a first filter, we used the habitat information (i.e., "host plant-associated" label) provided by proGenomes2, which is based on the PATRIC database (31). The resulting list was reviewed manually to exclude uncertain or incorrectly annotated entries by checking their metadata and associated literature. Additionally, we included other known plant-associated species on the list that were missed by the PATRIC database but that were considered PAB based on published data. In total, we identified 960 reference species (PAB-broad) that could be considered related to the plant environment. From this list, we extracted two subdatabases (see Data Set S1 in the supplemental material): phytopathogens (PAB-phyto, 119 members) and plant symbionts (PAB-symb, 192 members).

**Degree of plant specificity (DPS).** A specificity score for plant association was calculated for each LBD type based on the proportion of PAB species present in each LBD cluster. We calculated three score values, which we refer to as the degree of plant specificity (DPS), depending on the PAB reference database used: DPS-broad, the proportion of PAB-broad species in each LBD cluster; DPS-phyto, the proportion of PAB-phyto species; and DPS-symb, the proportion of PAB-symb species.

**Phylogenetic tree reconstruction and visualization.** Multiple sequence alignments were built for each cluster using Clustal Omega v1.2.4 (60), and phylogeny was inferred by IQ-Tree v1.6.12 using the default parameters (61). The trees were further analyzed and visualized using ETE3 v3.0 (62), with custom Python scripts integrating the annotations of each sequence for its taxonomy, domain architecture, sequence alignment, and plant-specificity prediction (DPS).

**Phylogenetic signal tests.** The phylogenetic signal tests were performed using the $\delta$-approach (44), a phylogenetic analogue of the Shannon entropy that measures the degree of phylogenetic signal between a categorical trait (trait vector) and a phylogeny (metric-tree). We used the $\delta$-approach to specifically test the null hypothesis that a given taxonomic distribution of an LBD follows the phylogenetic signal of the underlying species, which provided us with a *P* value for each LBD cluster. We applied 100 iterations per test and set the *P* value threshold at 0.05.

The species phylogeny used as a reference in all the tests was reconstructed using the ETE3 (62) supermatrix-based workflow and a concatenated alignment of 40 universal marker genes (63) extracted from the 11,806 species-representative genomes using the FetchMG tool (64). Multiple sequence alignments were inferred using Clustal Omega v1.2.4 (60), and phylogenetic reconstruction was performed with FastTree v2.1 (65). Moreover, an alternative species phylogeny including only genomes with at least one CR was reconstructed using the same methodology. As the $\delta$-statistic has poor sensitivity in detecting the phylogenetic signal for small taxon sample sizes ($<$20 taxa), LBD clusters mapping to reference phylogenetic tree nodes smaller than 20 leaves were discarded from the analysis (Data Set S2).

## SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

**FIG S1**, TIF file, 0.5 MB.

**TABLE S1**, PDF file, 0.1 MB.

**TABLE S2**, PDF file, 0.04 MB.

**DATA SET S1**, XLSX file, 0.1 MB.

**DATA SET S2**, XLSX file, 0.1 MB.

**DATA SET S3**, XLSX file, 0.4 MB.

**DATA SET S4**, XLSX file, 6.9 MB.

**DATA SET S5**, XLSX file, 0.3 MB.

## ACKNOWLEDGMENTS

This research has been supported by grants PGC2018-098073-A-I00 MCIU/AEI/FEDER, UE (to J.H.-C.), BIO2016-76779-P (to T.K.), AGL2017-82492-C2-1-R (to C.R.), and RTI2018-095222-B-I00 (to E.L.-S.) from the Ministerio de Ciencia, Innovación y Universidades, Spain, as well as grant P18-FR-1621 (to T.K.) from the Junta de Andalucía. C.S.-L. was supported by the FPU program (FPU19/06635, MICINN-Spain), and J.P.C.-V. by the FPI program (BES-2016-076452, MINECO-Spain).

## REFERENCES

1. Miller LD, Russell MH, Alexandre G. 2009. Diversity in bacterial chemotactic responses and niche adaptation. Adv Appl Microbiol 66:53–75. https://doi.org/10.1016/S0065-2164(08)00803-4.
2. Wadhams GH, Armitage JP. 2004. Making sense of it all: bacterial chemotaxis. Nat Rev Mol Cell Biol 5:1024–1037. https://doi.org/10.1038/nrm1524.
3. Wuichet K, Zhulin IB. 2010. Origins and diversification of a complex signal transduction system in prokaryotes. Sci Signal 3:ra50. https://doi.org/10.1126/scisignal.2000724.
4. Whitchurch CB, Leech AJ, Young MD, Kennedy D, Sargent JL, Bertrand JJ, Semmler ABT, Mellick AS, Martin PR, Alm RA, Hobbs M, Beatson SA, Huang B, Nguyen L, Commolli JC, Engel JN, Darzins A, Mattick JS. 2004.



Downloaded from https://journals.asm.org/journal/msystems on 03 December 2024 by 83.36.16.221.

Characterization of a complex chemosensory signal transduction system which controls twitching motility in *Pseudomonas aeruginosa*. Mol Microbiol 52:873–893. https://doi.org/10.1111/j.1365-2958.2004.04026.x.
5. Hickman JW, Tifrea DF, Harwood CS. 2005. A chemosensory system that regulates biofilm formation through modulation of cyclic diguanylate levels. Proc Natl Acad Sci U S A 102:14422–14427. https://doi.org/10.1073/pnas.0507170102.
6. Gavira JA, Gumerov VM, Rico-Jiménez M, Petukh M, Upadhyay AA, Ortega A, Matilla MA, Zhulin IB, Krell T. 2020. How bacterial chemoreceptors evolve novel ligand specificities. mBio 11:e03066-19. https://doi.org/10.1128/mBio.03066-19.
7. Ortega Á, Zhulin IB, Krell T. 2017. Sensory repertoire of bacterial chemoreceptors. Microbiol Mol Biol Rev 81:e00033-17. https://doi.org/10.1128/MMBR.00033-17.
8. Matilla MA, Krell T. 2018. The effect of bacterial chemotaxis on host infection and pathogenicity. FEMS Microbiol Rev 42:fux052. https://doi.org/10.1093/femsre/fux052.
9. Elgamoudi BA, Andrianova EP, Shewell LK, Day CJ, King RM, Taha Rahman H, Hartley-Tassell LE, Zhulin IB, Korolik V. 2021. The *Campylobacter jejuni* chemoreceptor Tlp10 has a bimodal ligand-binding domain and specificity for multiple classes of chemoeffectors. Sci Signal 16:eabc8521. https://doi.org/10.1126/scisignal.abc8521.
10. Milligan DL, Koshland DE, Jr. 1993. Purification and characterization of the periplasmic domain of the aspartate chemoreceptor. J Biol Chem 268:19991–19997. https://doi.org/10.1016/S0021-9258(20)80684-X.
11. Clarke S, Koshland DE, Jr. 1979. Membrane receptors for aspartate and serine in bacterial chemotaxis. J Biol Chem 254:9695–9702. https://doi.org/10.1016/S0021-9258(19)83572-X.
12. Bi S, Pollard AM, Yang Y, Jin F, Sourjik V. 2016. Engineering hybrid chemotaxis receptors in bacteria. ACS Synth Biol 5:989–1001. https://doi.org/10.1021/acssynbio.6b00053.
13. Reyes-Darias JA, Yang Y, Sourjik V, Krell T. 2015. Correlation between signal input and output in PctA and PctB amino acid chemoreceptor of *Pseudomonas aeruginosa*. Mol Microbiol 96:513–525. https://doi.org/10.1111/mmi.12953.
14. Alexandre G, Greer-Phillips S, Zhulin IB. 2004. Ecological role of energy taxis in microorganisms. FEMS Microbiol Rev 28:113–126. https://doi.org/10.1016/j.femsre.2003.10.003.
15. Lacal J, García-Fontana C, Muñoz-Martínez F, Ramos J-L, Krell T. 2010. Sensing of environmental signals: classification of chemoreceptors according to the size of their ligand binding regions. Environ Microbiol 12:2873–2884. https://doi.org/10.1111/j.1462-2920.2010.02325.x.
16. Scharf BE, Hynes MF, Alexandre GM. 2016. Chemotaxis signaling systems in model beneficial plant–bacteria associations. Plant Mol Biol 90:549–559. https://doi.org/10.1007/s11103-016-0432-4.
17. Antunez-Lamas M, Cabrera E, Lopez-Solanilla E, Solano R, González-Melendi P, Chico JM, Toth I, Birch P, Pritchard L, Liu H, Rodriguez-Palenzuela P. 2009. Bacterial chemoattraction towards jasmonate plays a role in the entry of *Dickeya dadantii* through wounded tissues. Mol Microbiol 74:662–671. https://doi.org/10.1111/j.1365-2958.2009.06888.x.
18. Cerna-Vargas JP, Santamaría-Hernando S, Matilla MA, Rodríguez-Herva JJ, Daddaoua A, Rodríguez-Palenzuela P, Krell T, López-Solanilla E. 2019. Chemoperception of specific amino acids controls phytopathogenicity in *Pseudomonas syringae* pv. tomato. mBio 10:e01968-19. https://doi.org/10.1128/mBio.01868-19.
19. Hida A, Oku S, Kawasaki T, Nakashimada Y, Tajima T, Kato J. 2015. Identification of the *mcpA* and *mcpM* genes, encoding methyl-accepting proteins involved in amino acid and L-malate chemotaxis, and involvement of McpM-mediated chemotaxis in plant infection by *Ralstonia pseudosolanacearum* (formerly *Ralstonia solanacearum* phylotypes I and III). Appl Environ Microbiol 81:7420–7430. https://doi.org/10.1128/AEM.01870-15.
20. Kumar Verma R, Samal B, Chatterjee S. 2018. *Xanthomonas oryzae* pv. oryzae chemotaxis components and chemoreceptor Mcp2 are involved in the sensing of constituents of xylem sap and contribute to the regulation of virulence-associated functions and entry into rice. Mol Plant Pathol 19:2397–2415. https://doi.org/10.1111/mpp.12718.
21. Yao J, Allen C. 2006. Chemotaxis is required for virulence and competitive fitness of the bacterial wilt pathogen *Ralstonia solanacearum*. J Bacteriol 188:3697–3708. https://doi.org/10.1128/JB.188.10.3697-3708.2006.
22. Antúnez-Lamas M, Cabrera-Ordóñez E, López-Solanilla E, Raposo R, Trelles-Salazar O, Rodríguez-Moreno A, Rodríguez-Palenzuela P. 2009. Role of motility and chemotaxis in the pathogenesis of *Dickeya dadantii* 3937 (ex *Erwinia chrysanthemi* 3937). Microbiology 155:434–442. https://doi.org/10.1099/mic.0.022244-0.
23. Río-Álvarez I, Muñoz-Gómez C, Navas-Vásquez M, Martínez-García PM, Antúnez-Lamas M, Rodríguez-Palenzuela P, López-Solanilla E. 2015. Role of *Dickeya dadantii* 3937 chemoreceptors in the entry to Arabidopsis leaves through wounds. Mol Plant Pathol 16:685–698. https://doi.org/10.1111/mpp.12227.
24. Raina J-B, Fernandez V, Lambert B, Stocker R, Seymour JR. 2019. The role of microbial motility and chemotaxis in symbiosis. Nat Rev Microbiol 17:284–294. https://doi.org/10.1038/s41579-019-0182-9.
25. Kamoun S, Kado CI. 1990. Phenotypic switching affecting chemotaxis, xanthan production, and virulence in *Xanthomonas campestris*. Appl Environ Microbiol 56:3855–3860. https://doi.org/10.1128/aem.56.12.3855-3860.1990.
26. Jones P, Garcia BJ, Furches A, Tuskan GA, Jacobson D. 2019. Plant host-associated mechanisms for microbial selection. Front Plant Sci 10:862. https://doi.org/10.3389/fpls.2019.00862.
27. Yuan J, Zhang N, Huang Q, Raza W, Li R, Vivanco JM, Shen Q. 2015. Organic acids from root exudates of banana help root colonization of PGPR strain *Bacillus amyloliquefaciens*. Sci Rep 5:13438. https://doi.org/10.1038/srep13438.
28. Gupta Sood S. 2003. Chemotactic response of plant-growth-promoting bacteria towards roots of vesicular-arbuscular mycorrhizal tomato plants. FEMS Microbiol Ecol 45:219–227. https://doi.org/10.1016/S0168-6496(03)00155-7.
29. Bulgarelli D, Schlaeppi K, Spaepen S, van Themaat EVL, Schulze-Lefert P. 2013. Structure and functions of the bacterial microbiota of plants. Annu Rev Plant Biol 64:807–838. https://doi.org/10.1146/annurev-arplant-050312-120106.
30. Vorholt JA. 2012. Microbial life in the phyllosphere. Nat Rev Microbiol 10:828–840. https://doi.org/10.1038/nrmicro2910.
31. Mende DR, Letunic I, Maistrenko OM, Schmidt TSB, Milanese A, Paoli L, Hernández-Plaza A, Orakov AN, Forslund SK, Sunagawa S, Zeller G, Huerta-Cepas J, Coelho LP, Bork P. 2020. proGenomes2: an improved database for accurate and consistent habitat, taxonomic and functional annotations of prokaryotic genomes. Nucleic Acids Res 48(D1):D621–D625. https://doi.org/10.1093/nar/gkz1002.
32. Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, Leipe D, Mcveigh R, O'Neill K, Robbertse B, Sharma S, Soussov V, Sullivan JP, Sun L, Turner S, Karsch-Mizrachi I. 2020. NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database (Oxford) 2020:baaa062. https://doi.org/10.1093/database/baaa062.
33. Mende DR, Sunagawa S, Zeller G, Bork P. 2013. Accurate and universal delineation of prokaryotic species. Nat Methods 10:881–884. https://doi.org/10.1038/nmeth.2575.
34. Silby MW, Winstanley C, Godfrey SAC, Levy SB, Jackson RW. 2011. *Pseudomonas* genomes: diverse and adaptable. FEMS Microbiol Rev 35:652–680. https://doi.org/10.1111/j.1574-6976.2011.00269.x.
35. Baldani JI, Pot B, Kirchhof G, Falsen E, Baldani VLD, Olivares FL, Hoste B, Kersters K, Hartmann A, Gillis M, Dobereiner J. 1996. Emended description of *Herbaspirillum*; inclusion of [*Pseudomonas*] *rubrisubalbicans*, a mild plant pathogen, as *Herbaspirillum rubrisubalbicans* comb. nov.; and classification of a group of clinical isolates (EF Group 1) as *Herbaspirillum* species 3. Int Syst Bacteriol 46:802–810. https://doi.org/10.1099/00207713-46-3-802.
36. Bi S, Yu D, Si G, Luo C, Li T, Ouyang Q, Jakovljevic V, Sourjik V, Tu Y, Lai L. 2013. Discovery of novel chemoeffectors and rational design of *Escherichia coli* chemoreceptor specificity. Proc Natl Acad Sci U S A 110:16814–16819. https://doi.org/10.1073/pnas.1306811110.
37. Goers Sweeney E, Henderson JN, Goers J, Wreden C, Hicks KG, Foster JK, Parthasarathy R, Remington SJ, Guillemin K. 2012. Structure and proposed mechanism for the pH-sensing *Helicobacter pylori* chemoreceptor TlpB. Structure 20:1177–1188. https://doi.org/10.1016/j.str.2012.04.021.
38. Webb BA, Hildreth S, Helm RF, Scharf BE. 2014. *Sinorhizobium meliloti* chemoreceptor McpU mediates chemotaxis toward host plant exudates through direct proline sensing. Appl Environ Microbiol 80:3404–3415. https://doi.org/10.1128/AEM.00115-14.
39. Ma B, Charkowski AO, Glasner JD, Perna NT. 2014. Identification of host-microbe interaction factors in the genomes of soft rot-associated pathogens *Dickeya dadantii* 3937 and *Pectobacterium carotovorum* WPP14 with supervised machine learning. BMC Genomics 15:508. https://doi.org/10.1186/1471-2164-15-508.
40. Raoul des Essarts Y, Pédron J, Blin P, Van Dijk E, Faure D, Van Gijsegem F. 2019. Common and distinctive adaptive traits expressed in *Dickeya dianthicola* and *Dickeya solani* pathogens when exploiting potato plant host. Environ Microbiol 21:1004–1018. https://doi.org/10.1111/1462-2920.14519.



Downloaded from https://journals.asm.org/journal/msystems on 03 December 2024 by 83.36.16.221.

41. Puławska J, Kałużna M, Warabieda W, Pothier JF, Gétaz M, van der Wolf JM. 2020. Transcriptome analysis of *Xanthomonas fragariae* in strawberry leaves. Sci Rep 10:20582. https://doi.org/10.1038/s41598-020-77612-y.
42. Wei C, Ding T, Chang C, Yu C, Li X, Liu Q. 2019. Global regulator PhoP is necessary for motility, biofilm formation, exoenzyme production, and virulence of *Xanthomonas citri* subsp. citri on citrus plants. Genes (Basel) 10:340. https://doi.org/10.3390/genes10050340.
43. Matas IM, Lambertsen L, Rodríguez-Moreno L, Ramos C. 2012. Identification of novel virulence genes and metabolic pathways required for full fitness of *Pseudomonas savastanoi* pv. savastanoi in olive (*Olea europaea*) knots. New Phytol 196:1182–1196. https://doi.org/10.1111/j.1469-8137.2012.04357.x.
44. Borges R, Machado JP, Gomes C, Rocha AP, Antunes A. 2019. Measuring phylogenetic signal between categorical traits and phylogenies. Bioinformatics 35:1862–1869. https://doi.org/10.1093/bioinformatics/bty800.
45. Salah Ud-Din AIM, Roujeinikova A. 2017. Methyl-accepting chemotaxis proteins: a core sensing element in prokaryotes and archaea. Cell Mol Life Sci 74:3293–3303. https://doi.org/10.1007/s00018-017-2514-0.
46. Wuichet K, Alexander RP, Zhulin IB. 2007. Comparative genomic and protein sequence analyses of a complex system controlling bacterial chemotaxis. Methods Enzymol 422:1–31. https://doi.org/10.1016/S0076-6879(06)22001-9.
47. Zhulin IB. 2001. The superfamily of chemotaxis transducers: from physiology to genomics and back. Adv Microb Physiol 45:157–198. https://doi.org/10.1016/s0065-2911(01)45004-1.
48. Galperin MY. 2018. What bacteria want. Environ Microbiol 20:4221–4229. https://doi.org/10.1111/1462-2920.14398.
49. Gushchin I, Melnikov I, Polovinkin V, Ishchenko A, Yuzhakova A, Buslaev P, Bourenkov G, Grudinin S, Round E, Balandin T, Borshchevskiy V, Willbold D, Leonard G, Büldt G, Popov A, Gordeliy V. 2017. Mechanism of transmembrane signaling by sensor histidine kinases. Science 356:eaah6345. https://doi.org/10.1126/science.aah6345.
50. Martín-Mora D, Ortega Á, Matilla MA, Martínez-Rodríguez S, Gavira JA, Krell T. 2019. The molecular mechanism of nitrate chemotaxis via direct ligand binding to the PilJ domain of McpN. mBio 10:e02334-18. https://doi.org/10.1128/mBio.02334-18.
51. Li W, O'Neill KR, Haft DH, DiCuccio M, Chetvernin V, Badretdin A, Coulouris G, Chitsaz F, Derbyshire MK, Durkin AS, Gonzales NR, Gwadz M, Lanczycki CJ, Song JS, Thanki N, Wang J, Yamashita RA, Yang M, Zheng C, Marchler-Bauer A, Thibaud-Nissen F. 2021. RefSeq: expanding the prokaryotic genome annotation pipeline reach with protein family model curation. Nucleic Acids Res 49:D1020–D1028. https://doi.org/10.1093/nar/gkaa1105.
52. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, Hugenholtz P. 2018. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol 36:996–1004. https://doi.org/10.1038/nbt.4229.
53. Eddy SR. 2011. Accelerated profile HMM searches. PLoS Comput Biol 7:e1002195. https://doi.org/10.1371/journal.pcbi.1002195.
54. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, Sonnhammer ELL, Hirsh L, Paladin L, Piovesan D, Tosatto SCE, Finn RD. 2019. The Pfam protein families database in 2019. Nucleic Acids Res 47:D427–D432. https://doi.org/10.1093/nar/gky995.
55. Alexander RP, Zhulin IB. 2007. Evolutionary genomics reveals conserved structural determinants of signaling and adaptation in microbial chemoreceptors. Proc Natl Acad Sci U S A 104:2885–2890. https://doi.org/10.1073/pnas.0609359104.
56. Gumerov VM, Ortega DR, Adebali O, Ulrich LE, Zhulin IB. 2020. MiST 3.0: an updated microbial signal transduction database with an emphasis on chemosensory systems. Nucleic Acids Res 48:D459–D464. https://doi.org/10.1093/nar/gkz988.
57. Krogh A, Larsson B, von Heijne G, Sonnhammer ELL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305:567–580. https://doi.org/10.1006/jmbi.2000.4315.
58. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C, Bork P. 2017. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. Mol Biol Evol 34:2115–2122. https://doi.org/10.1093/molbev/msx148.
59. Steinegger M, Söding J. 2017. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol 35:1026–1028. https://doi.org/10.1038/nbt.3988.
60. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7:539. https://doi.org/10.1038/msb.2011.75.
61. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32:268–274. https://doi.org/10.1093/molbev/msu300.
62. Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. Mol Biol Evol 33:1635–1638. https://doi.org/10.1093/molbev/msw046.
63. Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P. 2006. Toward automatic reconstruction of a highly resolved tree of life. Science 311:1283–1287. https://doi.org/10.1126/science.1123061.
64. Milanese A, Mende DR, Paoli L, Salazar G, Ruscheweyh H-J, Cuenca M, Hingamp P, Alves R, Costea PI, Coelho LP, Schmidt TSB, Almeida A, Mitchell AL, Finn RD, Huerta-Cepas J, Bork P, Zeller G, Sunagawa S. 2019. Microbial abundance, activity and population genomic profiling with mOTUs2. Nat Commun 10:1014. https://doi.org/10.1038/s41467-019-08844-4.
65. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS One 5:e9490. https://doi.org/10.1371/journal.pone.0009490.

Downloaded from https://journals.asm.org/journal/msystems on 03 December 2024 by 83.36.16.221.