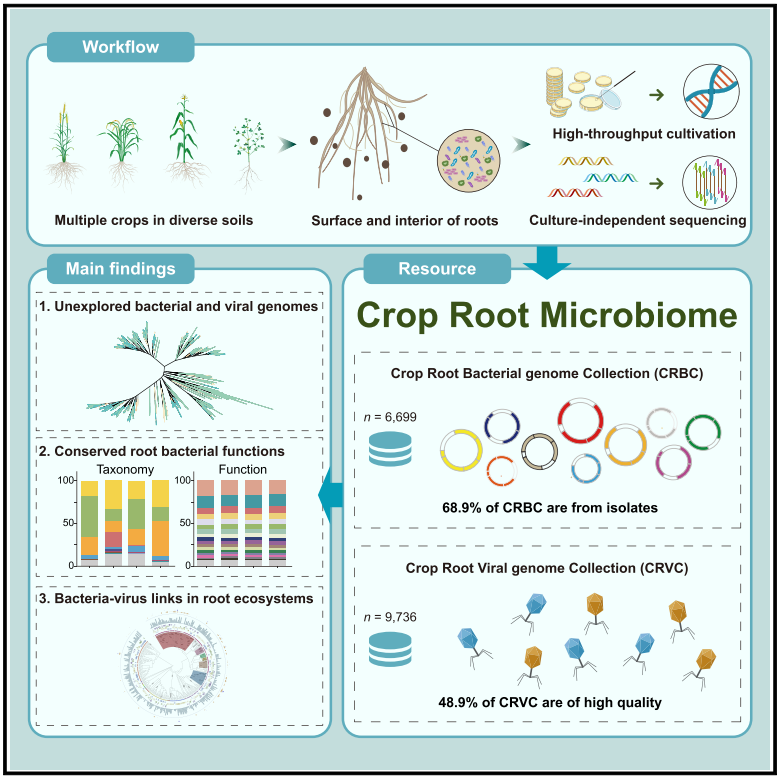
Cell

Crop root bacterial and viral genomes reveal unexplored species and microbiome patterns

Graphical abstract



Highlights

- Crop root bacterial (CRBC) and viral (CRVC) genome categories were established
- We found 1,817 undefined bacterial species and 1,572 unreported viral genera
- Despite varied taxa, root microbiome functions remain conserved across plants and soils
- A total of 27% of abundant bacteria in crop root microbiomes are linked to phages

Authors

Rui Dai, Jingying Zhang, Fang Liu, ..., Asaf Levy, Klaus Schlaeppi, Yang Bai

Correspondence

ybai@pku.edu.cn

In brief

Comprehensive bacterial and viral genome collections from crop roots uncover a large number of unexplored species and highlight conserved patterns within root microbiomes. Genomes and isolates are accessible at www.cropmicrobiome.com.

Dai et al., 2025, Cell *188*, 1–19
May 1, 2025
https://doi.org/10.1016/j.cell.2025.02.013

CellPress

Please cite this article in press as: Dai et al., Crop root bacterial and viral genomes reveal unexplored species and microbiome patterns, Cell (2025), https://doi.org/10.1016/j.cell.2025.02.013

Cell

CellPress
OPEN ACCESS

Resource

Crop root bacterial and viral genomes reveal unexplored species and microbiome patterns

Rui Dai,[1,2,3,4,11] Jingying Zhang,[2,3,11] Fang Liu,[1,11] Haoran Xu,[1,3,4] Jing-Mei Qian,[1,3,4] Shani Cheskis,[5] Weidong Liu,[1,3,4] Binglei Wang,[2] Honghui Zhu,[6] Lotte J.U. Pronk,[7] Marnix H. Medema,[7] Ronnie de Jonge,[8,9] Corné M.J. Pieterse,[8] Asaf Levy,[5] Klaus Schlaeppi,[10] and Yang Bai[2,3,12,*]

[1]Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China
[2]Peking-Tsinghua Center for Life Sciences, State Key Laboratory of Gene Function and Modulation Research, Peking-Tsinghua-NIBS Graduate Program, School of Life Sciences, Peking University, Beijing 100871, China
[3]CAS-JIC Centre of Excellence for Plant and Microbial Science, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China
[4]College of Advanced Agricultural Sciences, University of Chinese Academy of Sciences, Beijing 100101, China
[5]Department of Plant Pathology and Microbiology, Institute of Environmental Science, The Faculty of Agriculture, Food, and Environment, The Hebrew University of Jerusalem, Rehovot, Israel
[6]State Key Laboratory of Applied Microbiology Southern China, Institute of Microbiology, Guangdong Academy of Sciences, Guangzhou 510070, China
[7]Bioinformatics Group, Wageningen University & Research, 6708 PB Wageningen, the Netherlands
[8]Plant-Microbe Interactions, Department of Biology, Science for Life, Utrecht University, 3584 CH Utrecht, the Netherlands
[9]AI Technology for Life, Department of Information and Computing Sciences, Science for Life, Utrecht University, 3584 CC Utrecht, the Netherlands
[10]Department of Environmental Sciences, University of Basel, Basel 4056, Switzerland
[11]These authors contributed equally
[12]Lead contact
*Correspondence: ybai@pku.edu.cn
https://doi.org/10.1016/j.cell.2025.02.013

SUMMARY

Reference genomes of root microbes are essential for metagenomic analyses and mechanistic studies of crop root microbiomes. By combining high-throughput bacterial cultivation with metagenomic sequencing, we constructed comprehensive bacterial and viral genome collections from the roots of wheat, rice, maize, and *Medicago*. The crop root bacterial genome collection (CRBC) significantly expands the quantity and phylogenetic diversity of publicly available crop root bacterial genomes, with 6,699 bacterial genomes (68.9% from isolates) and 1,817 undefined species, expanding crop root bacterial diversity by 290.6%. The crop root viral genome collection (CRVC) contains 9,736 non-redundant viral genomes, with 1,572 previously unreported genus-level clusters in crop root microbiomes. From these, we identified conserved bacterial functions enriched in root microbiomes across soils and host species and uncovered previously unexplored bacteria-virus connections in crop root ecosystems. Together, the CRBC and CRVC serve as valuable resources for investigating microbial mechanisms and applications, supporting sustainable agriculture.

INTRODUCTION

Plant roots act as a hub for recruiting a diverse array of microorganisms from the surrounding soil.[1] These microorganisms colonize both the surface and interior of roots and, along with their microbial genomes, are collectively known as the root microbiome, which plays a critical role in influencing the growth and health of host plants, especially agronomic crops.[2–7] Recently, functional investigations and community profiling of root bacteria have emerged as a cutting-edge area in crop root microbiome research,[7–10] highlighting the urgent need for a comprehensive collection of bacterial genomes specific to crop root ecosystems. Such genome collections, particularly those derived from cultivated isolates, are invaluable resources for exploring the functional and biological potential of crop root microbiomes.

The absence of a comprehensive bacterial genome collection presents a significant challenge for in-depth research on the crop root microbiome. Marker gene-based amplicon sequencing technology captures the taxonomic diversity of microbiomes[11–14] but does not address their functional diversity. Bacterial genomes encode massive functional diversity, which can vary even between strains of the same species.[15–17] While shotgun metagenomic techniques can identify functional gene diversity within root microbes and allow for more precise taxonomic classification, they rely on comprehensive, high-quality reference genomes to be effective.[18,19]

Please cite this article in press as: Dai et al., Crop root bacterial and viral genomes reveal unexplored species and microbiome patterns, Cell (2025), https://doi.org/10.1016/j.cell.2025.02.013

Strain and genome collections of root bacteria are limited for agronomically relevant crops. Pioneering studies in model plants, such as *Arabidopsis thaliana* and *Lotus japonicus*, have generated nearly 800 bacterial genomes.[20–22] However, these resources from model plants are poorly suitable for crop research, as bacterial colonization exhibits host specificity.[22,23] Bacteria derived from one plant species can trigger strong immune responses in other plants, reducing their colonization.[22] To fully understand these microbial functions, genomic data derived from crop root microbiomes are crucial.[6,15,24,25] However, existing genomic resources related to the crop root microbiome primarily rely on small-scale datasets derived from divergent methodologies, and dominated with traditional agricultural species like *Rhizobium* and *Bacillus*.[4,26,15,24,27–29] In addition, limited access to isolates and inadequately organized metadata have hindered the full potential of these resources for the broader scientific community. Given the growing commercial interest in developing probiotics for agricultural applications—such as biocontrol agents, biofertilizers, and biostimulant inocula—the absence of a systematic collection of crop root bacteria represents a missed opportunity for advancing crop microbiome research.

Cultivation methods capture only a limited range of root bacterial taxa.[5,20] Metagenome-assembled genomes (MAGs) provide an alternative and complementary approach to accurately acquire genome information from uncultured root microbes.[30,31] However, obtaining MAGs for the root microbiomes of crops and model plants presents significant challenges. These include the high proportion of host plant genome sequences in root metagenomic data and the vast differences in microbial abundances within the root microbiome, both of which significantly increase sequencing costs, often 10–20 times higher than for environmental samples. These factors make it less feasible to acquire MAGs for microbes colonizing root surfaces and interiors.[6,25] As a result, comprehensive bacterial genome collections for crop root microbiome research remain scarce, and functional information for both cultivable and uncultivable bacteria is limited. This gap hinders mechanistic studies and genome-level analyses of the crop root microbiome.

In addition to research on root bacterial microbiomes, the study of root viromes has attracted considerable interest, given that viruses are widespread and play significant roles in microbial communities.[32,33] Numerous viral genomes have been discovered in the human gut, soils, and marine environments, often exhibiting distinct ecological specificities.[34–37] Across ecosystems, viruses, especially phages that infect bacteria, show intrinsic interaction with bacteria, profoundly influencing the population and life activities of their hosts. However, the distribution and prevalence of these viruses within plant root microbiomes, particularly in crops, remain unclear. Phages shape the genomes of bacteria and have led to the evolution of many antiviral systems.[38,39] Some phages possess the ability to regulate host behaviors, which not only benefits phage survival but also assists the host in, e.g., resisting infection by other phages.[40] Due to the limited genomic and metagenomic data available for crop root microbiomes, a dedicated viral genome collection for crop roots has yet to be developed. Establishing a systematic and comprehensive genome collection of crop root viruses is indispensable for advancing research in crop root microbiome studies.

In this work, we combined high-throughput bacterial cultivation with shotgun metagenomic sequencing to establish the crop root bacterial genome collection (CRBC) and the crop root viral genome collection (CRVC), comprehensive databases of bacterial and viral genomes from roots of multiple crop species. Using these resources, we evaluated the phylogenetic diversity and genome novelty of CRBC and CRVC, explored conserved characteristics of crop root microbiomes, and examined the interactions between bacteria and viruses within the crop root microbiome. The resources will advance research on crop root microbiomes, enable strain-level mechanistic research, and provide a foundation for developing microbial applications to support sustainable agricultural practices, and are accessible on our website (www.cropmicrobiome.com).

RESULTS

The CRBC comprises 6,699 genomes of root bacteria

To infer the genome sequences and functional repertoire of the bacterial root microbiome, we established a systematic and comprehensive collection of bacterial genomes from crop roots. This collection includes genomes of both cultivated bacterial strains and uncultured bacteria obtained through metagenomic assembly (Table S1A). The CRBC is derived from the roots of wheat, rice, maize, and *Medicago* grown in agricultural soils. Our sampling strategy enriched for root-associated microbiota, capturing microbes residing on the rhizoplane and within the endosphere (STAR Methods). To retrieve the culturable fraction of the crop root bacterial microbiome, we conducted extensive isolation of bacterial strains for whole-genome sequencing. To complement this and cover the unculturable fraction, we included MAGs by performing deep, culture-independent metagenomic sequencing on crop root samples (Figure 1).

The CRBC dataset comprises 6,699 high-quality bacterial genomes. We obtained 4,618 bacterial genomes (5.5 Tbp in raw reads) from pure root bacterial cultures. Following high-throughput bacterial isolation and selection of representative bacterial strains (STAR Methods), we identified 1,496 root bacterial genomes from wheat, 1,287 from rice, 1,056 from maize, and 779 from *Medicago* (Table S1A). Additionally, 2,081 MAGs of root bacteria were assembled from 332 root metagenomic samples across 14 datasets, spanning various soils and crop species (Tables S1A–S1D). These MAGs, representing a range of genomes of uncultured CRBC bacteria, complement the cultured bacterial genomes, adding 21 phyla and 147 genera not found in cultured isolates (Table S1E). Notably, the taxonomic diversity within the 6,699 genomes (68.9% from isolates) is extensive, spanning 27 phyla, 49 classes, and 113 orders. This diversity includes numerous genomes within prominent root bacterial phyla, such as Proteobacteria (3,659 genomes), Actinobacteriota (1,720 genomes), Bacteroidota (395 genomes), and Firmicutes (433 genomes). An impressive 79.1% of these genomes are of high quality, with an average completeness of 98.9% and an average contamination of 0.9%, aligning with standards in the Genome Taxonomy Database (GTDB),[41] the most

Please cite this article in press as: Dai et al., Crop root bacterial and viral genomes reveal unexplored species and microbiome patterns, Cell (2025), https://doi.org/10.1016/j.cell.2025.02.013

CellPress
OPEN ACCESS

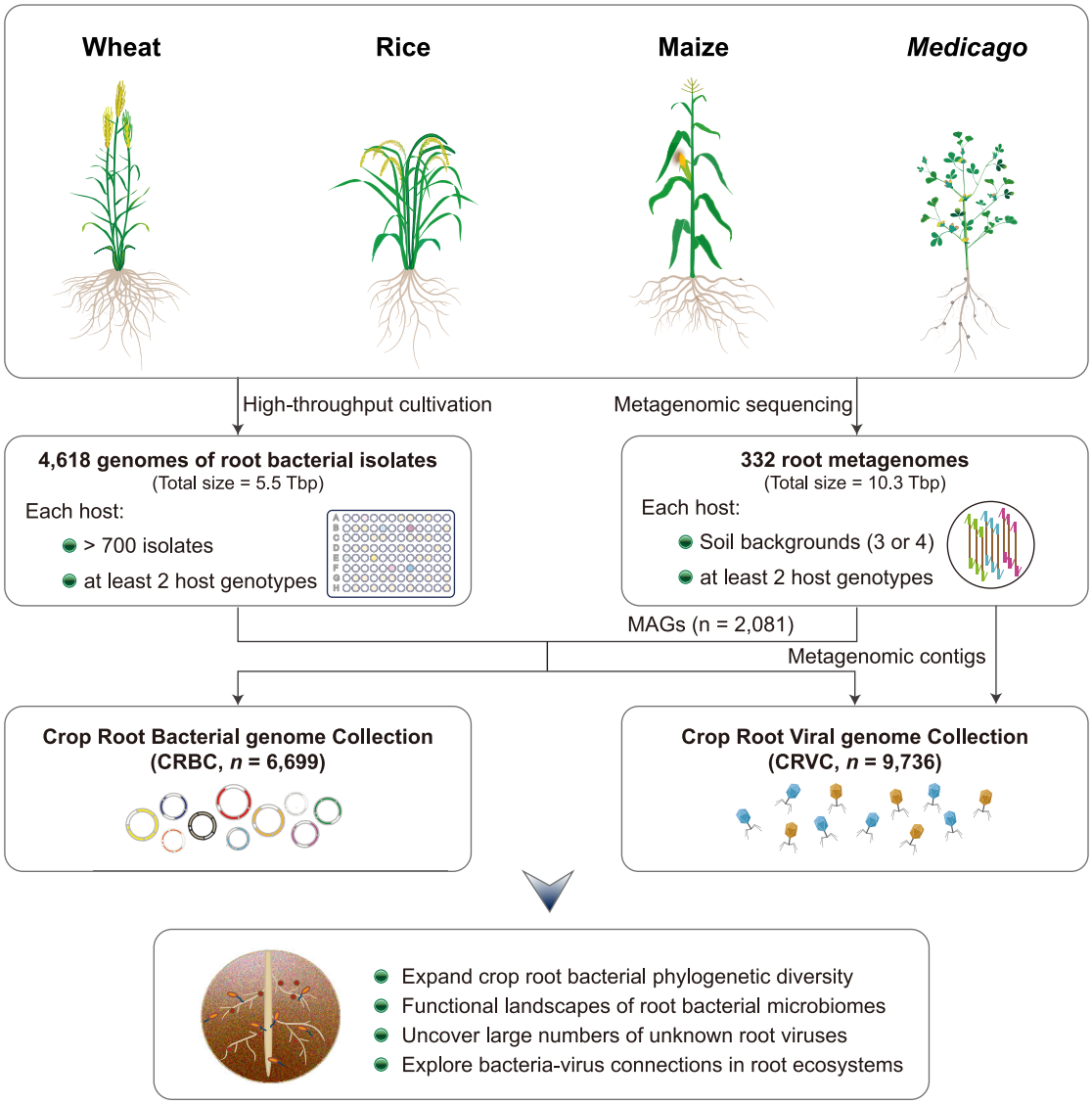


Figure 1. Design and overview of the CRBC and CRVC

Based on comprehensive root-derived bacterial isolates and root metagenomic data, we established the crop root bacterial genome collection (CRBC) and the crop root viral genome collection (CRVC) from wheat, rice, maize, and *Medicago*. The CRBC encompasses 6,699 root bacterial genomes. The database includes 4,618 genomes (5.5 Tbp of raw reads) from bacterial isolates from crop roots. Each crop comprises over 700 isolates. The database also includes 2,081 MAGs (10.3 Tbp of raw reads) obtained from 332 root metagenomic samples. Each crop species is represented by at least two genotypes and spans three or four soil backgrounds. We also reconstructed 9,736 viral genomes by using bacterial genomes and root metagenomic contigs and defined them as the CRVC.
See also Figure S1.

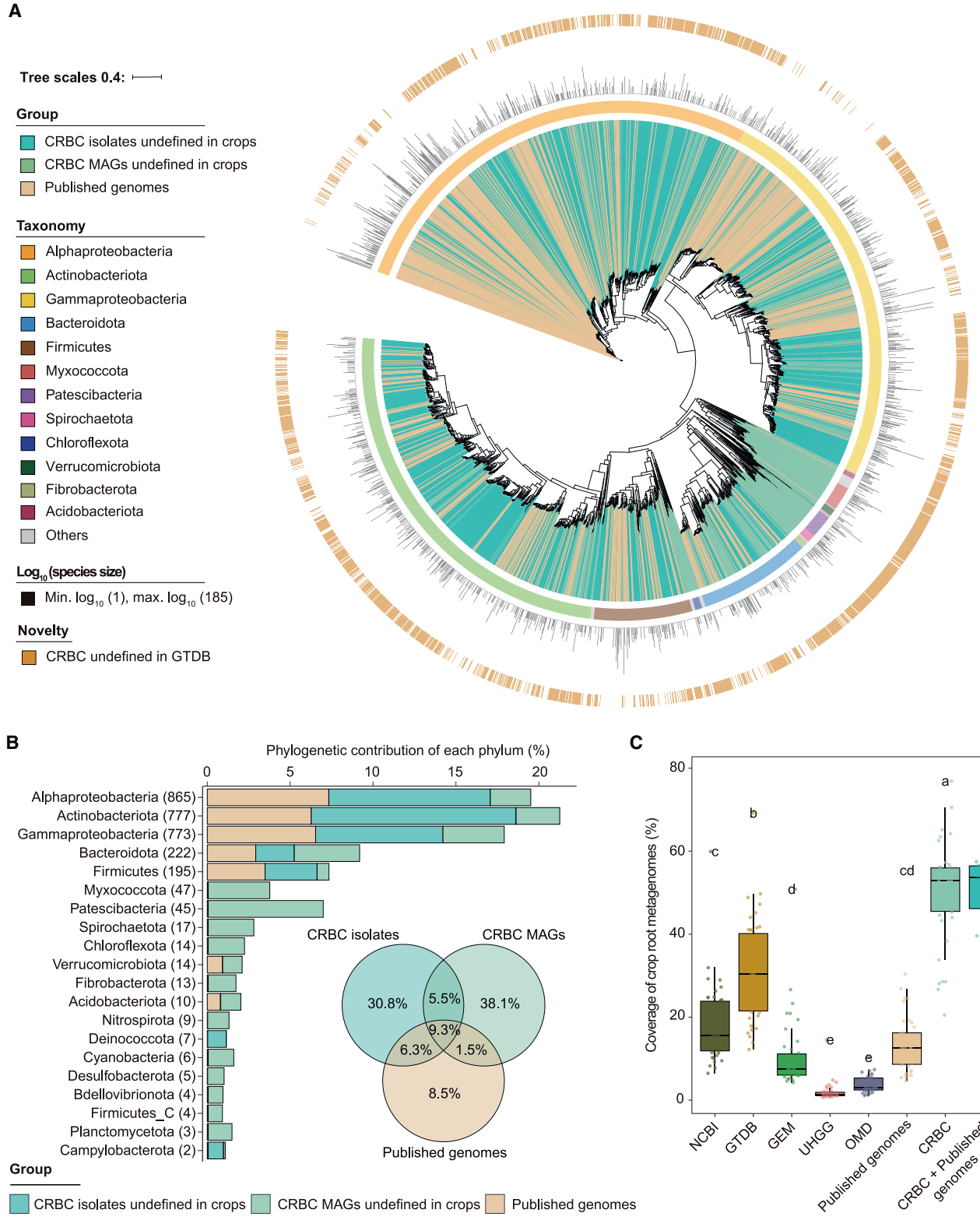
comprehensive database of over 300,000 prokaryotic genomes from diverse ecosystems (Figures S1A–S1C; Table S1A; STAR Methods). With its high-quality and expansive genomic representation, the resources of the CRBC will serve as a valuable foundation for both microbiological and genomic studies on the crop root microbiome.

The CRBC enhances crop root bacterial genome diversity and metagenomic read coverage

To uncover novel bacterial species within the CRBC, we systematically surveyed crop root bacterial genomes across different public databases, including NCBI,[42] the Integrated Microbial Genomes and Microbiomes (IMG/M),[43] European Nucleotide Archive (ENA),[44] and GTDB.[41] We found that fewer than 1% of the bacterial genomes were derived from crop roots (Table S1F), with 3,073 genomes having a quality score (QS) > 50, representing only 6 phyla and 33 orders (Figures S1D and S1E; Table S1G; STAR Methods).

We compared the CRBC genomes with publicly available crop root genomes and found that the CRBC substantially increased the number and diversity of crop root bacterial species. Using an average nucleotide identity (ANI) threshold of 95%, the combined set of genomes (6,699 + 3,073) clustered into a total of

Please cite this article in press as: Dai et al., Crop root bacterial and viral genomes reveal unexplored species and microbiome patterns, Cell (2025), https://doi.org/10.1016/j.cell.2025.02.013



(legend on next page)

Please cite this article in press as: Dai et al., Crop root bacterial and viral genomes reveal unexplored species and microbiome patterns, Cell (2025), https://doi.org/10.1016/j.cell.2025.02.013

3,044 distinct species (Figure 2A). Remarkably, the CRBC contributed significantly to this diversity, encompassing 2,318 species, 2.8 times the species richness of publicly available crop root bacterial genomes. The CRBC included 2,212 bacterial species not previously reported in crop roots (Figure S1F). Compared with genomes in GTDB, CRBC genomes contained 1,817 undefined bacterial species, including 1,329 isolates and 488 MAGs, which exhibited an average ANI of less than 95% with GTDB genomes (Figures 2A and S1G). Of these species, the following ranks could not be assigned, phylum (1), class (1), order (5), family (22), and 177 species with an unknown genus compared with the GTDB (Table S2A). Similar trends were obtained by using the relative evolutionary divergence (RED) value to infer the taxonomic ranks (Table S2A). On average, each CRBC genome contains 14.9% uncharacterized genes, and these undefined bacterial species substantially expand the tree of life, spanning 26 phyla, 47 classes, and 182 families (Tables S1A and S2A).

The CRBC expands the phylogenetic diversity (PD) of crop root bacteria, increasing that of publicly available crop root bacteria by 290.6% (Figure 2B; Table S2B; STAR Methods). Genomes derived from CRBC isolates notably expanded the phylogeny of bacterial phyla commonly found in the root microbiome, such as Alphaproteobacteria, Actinobacteriota, Gammaproteobacteria, Bacteroidota, and Firmicutes. Specifically, four essential root bacterial families—Burkholderiaceae, Sphingomonadaceae, Microbacteriaceae, and Nocardioidaceae—each containing over 100 species, offer valuable resources for targeted genomic investigations (Table S2C). In contrast, the CRBC MAGs predominantly expanded the diversity of phyla associated with uncultured bacteria in crop roots, such as Myxococcota, Patescibacteria, and Fibrobacterota. Intriguingly, a combined analysis of bacterial genomes and crop root metagenomic data revealed that bacteria corresponding to the MAGs were estimated to have lower growth rates than cultured bacterial isolates (Figure S1H; Table S2D; STAR Methods), underscoring the importance of MAGs in capturing less abundant bacterial species.

Then, we assessed the read coverage in root metagenomic data by comparing the CRBC with other public databases, including NCBI RefSeq,[42] GTDB,[41] Genomes from Earth's Microbiomes catalog (GEM),[45] Unified Human Gastrointestinal Genome collection (UHGG),[46] Ocean Microbiomics Database (OMD),[31] and public genomes from crop root bacteria. Using 42 metagenome samples from crop roots and 37 samples from the crop rhizosphere, which were not used in the CRBC construction (Table S2E; STAR Methods), we aligned the reads from these metagenomes to each reference database to determine read coverage. The CRBC achieved significantly higher coverage of metagenomic data (mean coverage: 50.2%) compared with the other reference databases (Figure 2C). Notably, root metagenomes had higher coverage with CRBC than rhizosphere data (Figure S1I). Reference genomes from human and ocean ecosystems covered only a small fraction (below 4%) of root metagenomic sequences (Table S2E). Although GEM included rhizosphere microbial genomes and aimed to cover diverse soil ecosystems, its mean coverage was only 10.6%, highlighting the distinct microbial compositions between rhizosphere soils and root-associated microbiomes. Finally, the most comprehensive collections of reference genomes, NCBI and GTDB, achieved mean coverages of 18.2% and 31.1%, respectively, both considerably lower than the CRBC. Taken together, these findings indicate the importance of niche-specific genome reference databases and highlight the CRBC's substantial improvement in supporting crop root metagenome analyses.

CRBC genomes harbor diverse functions and metabolite genes for crop growth benefits

Taking advantage of the genomic resource, we systematically examined all high-quality, non-redundant genomes in the CRBC and published crop root bacteria for the coexistence pattern of their capacity to encode genes associated with plant growth-promoting (PGP) functions. We observed a widespread distribution of genomic traits related to nutrient utilization,[47–51] the biosynthesis of plant growth hormones,[52–56] and resistance to biotic and abiotic stresses[57–59] among crop root bacteria

Figure 2. The CRBC dramatically expands the PD of publicly available crop root bacterial genomes

(A) The CRBC contains 1,817 undefined bacterial species. The maximum-likelihood phylogenetic tree contains 3,044 representative species of the CRBC and crop root bacteria in public databases (NCBI, IMG/M, and ENA; STAR Methods). The clades associated with phylogenetic branches of 2,212 CRBC unique bacterial species, which show ANI below 95% when compared with published genomes derived from crop roots, are designated as undefined in crops and highlighted in blue (n = 1,702, contributed by CRBC isolates) or green (n = 510, contributed by CRBC MAGs). Published bacterial species from crop roots (n = 832) are indicated in beige. The initial inner circle illustrates the taxonomy of representative species at the phylum level. Proteobacteria are displayed at the class level due to excessive species numbers. Black bars represent $\log_{10}$-transformed numbers of non-redundant genomes (ANI of <99.9%) within each representative bacterial species. The outermost circle indicates 1,817 undefined bacterial species (green, n = 1,817) in the CRBC when compared with bacterial genomes from GTDB with the threshold of 95% ANI.

(B) The CRBC dramatically extends the PD of crop root bacterial genomes in public databases. The bar plot illustrates the proportional contribution of the CRBC and published crop root bacteria to the PD of crop root bacterial genomes, with the top 20 most abundant phyla displayed. Bars are colored by the contributions of genomes, with CRBC isolates undefined in crops (blue), CRBC MAGs undefined in crops (green), and published species from crop roots in beige. The numbers of representative species in each phylum are specified in brackets following the phylum names. The inset Venn diagram illustrates the percentage of total PD of all genomes in (A). See also Table S2B. Note that the CRBC expands the PD of crop root bacterial genomes in public databases by 290.6% (30.8% + 5.5% + 38.1%)/(6.3% + 9.3% + 1.5% + 8.5%).

(C) The CRBC shows a higher coverage of crop root metagenomic reads than public genome resources. The boxplot shows proportions of metagenomic reads of 42 root samples unrelated with the CRBC construction (wheat, n = 12; rice, n = 12; maize, n = 6; *Medicago*, n = 12) aligned to genomes in public databases and the CRBC. The abbreviations are: NCBI, NCBI RefSeq database; GTDB, Genome Taxonomy Database; GEM, the Genome from Earth's Microbiomes catalog; UHGG, the Unified Human gastrointestinal Genome collection; OMD, the Ocean Microbiomics Database (adjusted $p < 0.05$, Kruskal-Wallis rank-sum test and Dunn's test). See also Table S2E.

See also Figure S1.

Please cite this article in press as: Dai et al., Crop root bacterial and viral genomes reveal unexplored species and microbiome patterns, Cell (2025), https://doi.org/10.1016/j.cell.2025.02.013

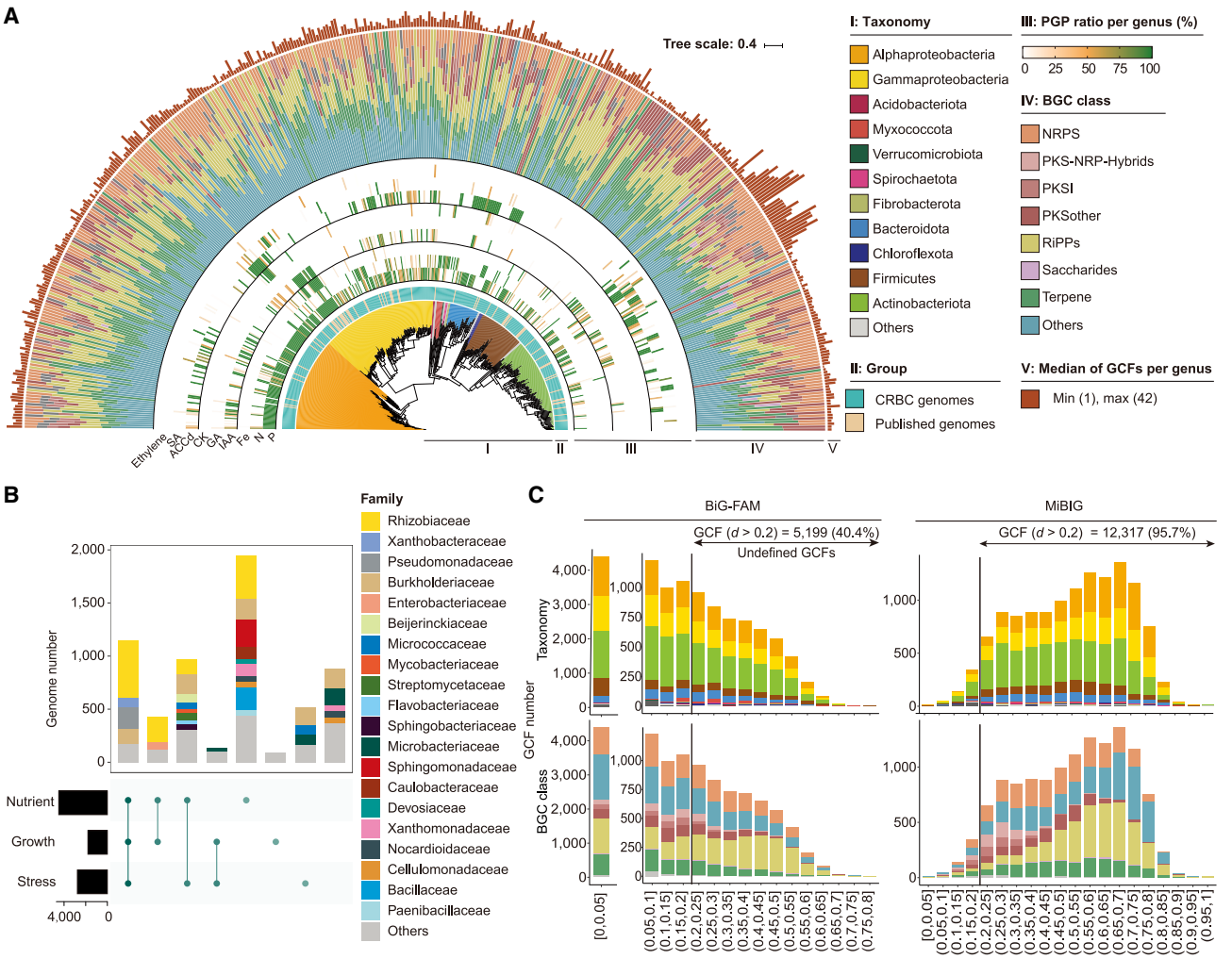


Figure 3. The CRBC harbors diverse beneficial functions and numerous metabolite gene clusters

(A) Distribution of PGP functions and BGCs. The phylogenetic tree is constructed using the high-quality representative genomes of each bacterial genus in the CRBC and public databases (STAR Methods). Phylum level taxonomy is colored (I). Genera containing CRBC genomes are labeled in blue (II). For each genus, plant growth-promoting (PGP) functions are illustrated as the percentage of genomes containing all the necessary genes for each function (III; STAR Methods). Bacterial PGP functions are categorized into three groups: nutrient utilization, growth promotion, and stress tolerance, according to the KEGG database and prior literatures. Nutrient utilization includes phosphorus nutrition (P), nitrogen fixation (N), and siderophore biosynthesis (Fe). Growth promotion includes biosynthesis of indole-3-acetic acid (IAA), gibberellin (GA), and cytokinin (CK). Stress tolerance includes the biosynthesis of 1-aminocyclopropane-1-carboxylic acid deaminase (ACCd), salicylic acid (SA), and ethylene. The BGC compositions are presented as the average of genomes at the genus level (IV). BGC classes are shown with the following abbreviations: non-ribosomal peptide synthetases (NRPSs); polyketide synthase (PKS); ribosomally synthesized and post-translationally modified peptide (RiPP). The outermost bar chart (section V) shows the median number of BGCs in all genomes of each genus. See also Table S3C.

(B) Coexistence patterns of PGP functions within individual bacterial genomes. The intersection scheme with vertical lines illustrates co-existent functions within individual genomes. The number and taxonomy of genomes for each coexistence pattern are shown in the stacked bar plot in the above panel. The number of genomes for each PGP group is shown in the lower left.

(C) The CRBC encodes a large number of undefined BGCs. All 48,643 BGCs identified in genomes of the CRBC, and public databases are clustered into 12,865 GCFs. Comparison with computationally predicted (BiG-FAM, left) and experimentally validated (MIBiG, right) BGCs revealed 5,199 undefined GCFs with an average cosine distance greater than 0.2, 81.1% of which are uniquely contributed by CRBC genomes. The upper panels are colored according to the taxonomy of genomes encoding BGCs, whereas the bottom panels are colored according to BGC classes.

See also Figure S2.

(Figure 3A; Tables S3A–S3C). Among the 6,109 high-quality bacteria genomes, 5,231 encoded at least one of the tested PGP functions, implying a significant role for root bacteria in supporting crop growth (Figure S2A; Table S3C). More than 43.8% of these genomes encode at least two categories of PGP functions, and 18.7% encode all three categories, with dominant representation from the Rhizobiaceae, Xanthobacteraceae, Pseudomonadaceae, and Burkholderiaceae (Figure 3B). Further analysis identified over 1,000 bacterial genomes with both nitrogen-fixation and phosphorus solubilization capabilities, especially in

Please cite this article in press as: Dai et al., Crop root bacterial and viral genomes reveal unexplored species and microbiome patterns, Cell (2025), https://doi.org/10.1016/j.cell.2025.02.013

CellPress
OPEN ACCESS

Figure 4. Conserved genetic features of the root bacterial microbiomes across multiple crop species grown in diverse soils

(A) Taxonomic composition of root bacterial metagenomes shows dramatic variation among 14 metagenomic datasets. The taxonomy of bacterial genes in 332 root metagenomic samples was classified by Kraken2 with the GTDB database implemented with crop root bacterial genomes (STAR Methods). Bacteria are displayed at the phylum level. The top 11 most abundant phyla are displayed, while phyla of lower abundance are shown as others. Proteobacteria are displayed at the class level due to excessive abundance.

(B) Convergent functions of root bacterial metagenomes among 14 metagenomic datasets. Bacterial genes are annotated according to the KEGG database and are summarized at the KEGG pathway level 2 (superpathway). The top 17 most abundant superpathways are displayed, while the functions of lower abundance annotated by the KEGG are shown as others.

(C) Consistent enrichment patterns of genetic functions in the root bacterial metagenomes compared with soils across multiple crops grown in diverse soils. The Venn diagrams show the overlap of KEGG bacterial functions that are enriched in root microbiomes. Functions enriched in the roots of all four crop species are highlighted in green (root, $n = 332$; soil, $n = 75$, adjusted $p < 0.05$, Wilcoxon rank-sum test). See also Table S4I.

(D and E) Abundance and KO distribution of functions consistently enriched in the root microbiomes. Boxplots (D) show the cumulative abundance of top 11 abundant bacterial functions that are consistently enriched in roots of wheat, rice, maize, and *Medicago* compared with corresponding soils. The distribution (E) of KOs in each function are colored according to their frequency of enrichment (green) or depletion (brown) in the roots of crop species. The colors from light to dark

(legend continued on next page)

Please cite this article in press as: Dai et al., Crop root bacterial and viral genomes reveal unexplored species and microbiome patterns, Cell (2025), https://doi.org/10.1016/j.cell.2025.02.013

Rhizobiaceae and Burkholderiaceae (Figure S2B). Bacillaceae were notable for their dual capabilities in phosphorus solubilization and siderophores biosynthesis. Notably, 18 genomes encode functions related to phosphorus, nitrogen, and siderophore utilization, with 8 genomes from the *Klebsiella* genus (Table S3C). In contrast, functions related to plant growth hormone biosynthesis or resistance to biotic and abiotic stresses rarely co-occurred in a single bacterium (Figures S2C and S2D). This bacterial resource offers a valuable opportunity to investigate the genomic coexistence of these functions and to identify promising candidates for future agricultural applications.

We next identified biosynthetic gene clusters (BGCs) within crop root bacteria, which are related with crop growth and health.[60–62] In total, we identified 48,643 BGCs across the CRBC and publicly available crop root bacterial genomes (Figure 3A; Table S3D). The predominant categories of BGCs were ribosomally synthesized and post-translationally modified peptides (RiPPs), non-ribosomal peptide synthetase (NRPS), and terpenes, representing 26.1%, 15.4%, and 14.3% of the clusters, respectively (Figure S2E). These bacteria exhibited diverse biosynthetic capabilities (Figures S2F–S2H), averaging 8 BGCs per genome. Next, we clustered the identified BGCs into 12,865 gene cluster families (GCFs) (Table S3E; STAR Methods). Strikingly, 5,199 (40.4%) of the identified GCFs did not show close similarity to any computationally predicted BGCs from the biosynthetic gene cluster families (BiG-FAM) database,[63] while 12,317 (95.7%) of GCFs showed no close similarity to experimentally validated BGCs from the Minimum Information about a Biosynthetic Gene cluster (MIBiG) database[64] (Figure 3C; STAR Methods). These undefined GCFs were mainly contributed by Proteobacteria and Actinobacteriota, encoded functions related to RiPPs, NRPS, polyketide synthase (PKS), and terpenes. The most significant differences in our GCFs compared with known databases were in biosynthetic families related to RiPPs, predominantly originating from the genera *Dyadobacter*, *Paenibacillus*, and *Arthrobacter* (Tables S3D and S3E). These BGCs present an untapped reservoir of secondary metabolites with presumably specialized functions in crop root bacteria, potentially offering new avenues for agricultural applications.

Conserved genetic characteristics of root microbiomes across multiple crop species grown in diverse soils

The scarcity of bacterial genomes and the interference of host DNA hinder our understanding of root ecosystems. To investigate the genomic patterns of root microbes, we systematically explored conserved features within 332 deeply sequenced root metagenomes (median depth of 30.4 Gbp) from 14 datasets across diverse crop species and soil backgrounds, including wheat, rice, maize, and *Medicago*. Each crop included at least two genotypes and three to four soils to ensure diverse sample representation. On average, 82.0% of the sequencing reads were host derived (Table S1D). After removing host reads, we annotated microbial genes using our genome resources and publicly available microbial reference genomes. Rarefaction analysis showed that the non-redundant microbial gene reservoirs for each crop had reached saturation (Figure S3A; Table S3F). The number of non-redundant genes reproducibly detected in the root microbiomes was 13.2–55.4 times of those observed in their respective hosts (Figure S3B; Table S3G; STAR Methods). On average, 23.8% of root microbial genes could not be annotated by Kyoto Encyclopedia of Genes and Genomes (KEGG) (Table S3H), underscoring the extensive genetic potential of the root microbiome. Notably, we found a high abundance of bacteria in root microbiomes, comprising an average of 95.7% of the overall microbial relative abundance (Figure S3C; Table S3I). In contrast, fungi, archaea, viruses, and protists represented much smaller fractions of the root microbiome, averaging 0.74%, 0.15%, 3.21%, and 0.16%, respectively. These findings highlight the dominance of bacteria in root microbiomes and underpin the importance of developing systematic genome reference collections for crop root bacteria.

We next identified the conserved features of bacteria within the crop root microbiome using our root metagenomic datasets across multiple host plant species grown in diverse soil backgrounds (Tables S4A–S4F). For each root dataset, we collected 4–6 soil samples from unplanted soils alongside root sampling. Notably, although the taxonomic compositions of root microbiomes were strongly influenced by soil backgrounds and host species, the compositions of superpathway (2^{nd} level of KEGG pathway) functions in root microbiomes were similar across all 14 datasets (Figures 4A, 4B, and S3D), suggesting the presence of conserved genetic features in multiple crop root ecosystems. Further, we compared the microbiomes of roots and surrounding soils in each dataset and found that, within each crop species, a higher proportion of microbial taxa was influenced by soil backgrounds compared with the proportion of microbial functions affected by soils (Figures S3E–S3H; Tables S4G–S4J). A total of 11 microbial functions were conservatively enriched in root microbiomes of all four crops grown in various soils (Figure 4C; Table S4I). These functions are related to ATP-binding cassette (ABC) transporters, two-component systems, biofilm formation, bacterial chemotaxis, and flagellar assembly (Figure 4D). Their enrichment patterns were confirmed by analysis at the KEGG functional orthologs (KOs) level (Figure 4E; Table S4J). Notably, compared with KOs depleted in roots, most root-enriched KOs were associated with functions related to cell motility, membrane transport, and signal transduction (Figures 4F and 4G), genes belonging to which have been previously identified in random barcoded transposon mutant sequencing in the laboratory in single

correspond to the number of crop species enriched or depleted, ranging from 0 to 4. The horizontal bars within the boxes represent median values. The tops and bottoms of the boxes represent the 75th and 25th percentiles, respectively. The upper and lower whiskers extend to data no more than 1.5× the interquartile range from the upper edge and lower edge of the box, respectively (root, $n = 332$; soil, $n = 75$, adjusted $p < 0.05$, Wilcoxon rank-sum test). See also Table S4J.
(F and G) Functional proportion of KOs enriched or depleted in roots of four host species. Doughnut charts indicate KOs consistently enriched (F) or depleted (G) in the roots of wheat, rice, maize, and *Medicago* compared with corresponding soils. KOs are presented at the superpathway level (root, $n = 332$; soil, $n = 75$, adjusted $p < 0.05$, Wilcoxon rank-sum test).
See also Figures S3 and S4.

Please cite this article in press as: Dai et al., Crop root bacterial and viral genomes reveal unexplored species and microbiome patterns, Cell (2025), https://doi.org/10.1016/j.cell.2025.02.013

plant species,[65,66] or genomic exploration of root microbes.[21,67] Our metagenomic findings underscore the significance and widespread presence of these functions in the root ecosystem across diverse soil backgrounds and crop species.

Genome references are essential for linking functional genes to specific taxa. Therefore, we searched for the crop root bacterial taxa that predominantly harbor the identified root-enriched functional KOs (Table S4K). Overall, phylogenetically related genera exhibited similar functional patterns (Figure S4A). Proteobacteria were particularly enriched in root-specific functions, aligning with their dominance in the crop root microbiome. Genomes from genera such as *Rhizobacter*, *Variovorax*, *Acidovorax*, and *Rubrivivax* encoded over 130 root-enriched KOs (Table S4L), and Gammaproteobacteria were enriched in functions related to cellular community (Figure S4B). In addition, KOs related to cell motility, including bacterial chemotaxis and flagellar assembly, were abundant in Proteobacteria, Firmicutes, and several Actinobacteriota, such as Microbacteriaceae and Nocardioidaceae. However, these functions were less prevalent in Bacteroidota and most other Actinobacteriota (Table S4L). This analysis identified key determinants of root-associated bacteria and highlights essential taxon-to-function links, offering potential targets to enhance bacterial competence within crop root ecosystems.

The CRVC reveals unreported viral genus-level clusters and enhanced viral genetic diversity in crop root ecosystems

Viruses intrinsically interact with other microbes and play pivotal roles in natural ecosystems.[68] However, a systematic collection of viral genomes is currently lacking in crop root microbiome research. Based on CRBC and crop bacterial genomes in published databases, along with 29.5 million contigs derived from root metagenomes in this study, we systematically retrieved viruses within crop root ecosystems. Using a hybrid analysis involving geNomad,[69] VirSorter,[70] and DeepVirFinder,[71] we identified 9,736 non-redundant viral genomes with completeness of over 50%, collectively referred to as the CRVC (Figure S5A; Table S5A; STAR Methods). Most (95.3%) of these viral genomes are identified as bacteriophages, with 71.0% exhibiting lysogenic behavior (Figure S5B; Table S5A), reflecting our pipeline's focus on discovering viral genomes from bacterial genomes and metagenomes. Nearly half of the genomes within the CRVC are of high quality, with 723 complete viral genomes identified based on direct terminal repeats (DTRs) or inverted terminal repeats (ITRs), and most of the CRVC genomes belonged to *Caudoviricetes* (Figures S5C and S5D; Table S5A). The virus sizes vary extensively, ranging from 1,501 to 477,160 bp, with a median length of 40,590 bp (Figure S5C). Notably, 65.5% of the viral genomes in the CRVC were derived from the CRBC and our root metagenomes (Figure S5D).

To identify novel viral taxa within the CRVC, we clustered all CRVC viral genomes with those from public databases (NCBI RefSeq,[42] Metagenomic Gut Virus catalog [MGV],[35] Global Ocean Viromes 2.0 [GOV 2.0],[37] and viruses from Integrated Microbial Genomes/Virus [IMG/VR][36] root and soil ecosystems) using the Minimum Information about an Uncultivated Virus Genome (MIUViG) proposed standard thresholds of 95% ANI over 85% alignment fraction.[72] We identified 7,653 species-level clusters (viral operational taxonomic units [vOTUs]) in the CRVC, of which 92.8% are not reported in existing databases (Figure S5E; Table S5B). Between 12.3% and 36.1% of CRVC proteins were successfully annotated using KEGG,[51] Pfam,[73] and VOGDB[74] databases, while 57.6% of proteins exhibiting unknown functions (Figures S5F and S5G; Table S5C), underscoring CRVC's vast genetic potential. To assign higher taxonomic ranks, we grouped genomes into genus-level categories based on whole-genome gene-sharing profiles (STAR Methods). Notably, 1,572 (50.8%) of the 3,097 viral genus-level clusters in the CRVC did not cluster with viral genomes from other published databases (Figure 5B; Table S5D; STAR Methods). To minimize technical biases from bacterial isolation when comparing databases, we compared metagenome-derived viruses in the CRVC with those in public databases. We found that the CRVC shared more viral clusters with IMG/VR root and soil environments than with human and ocean environments (Figure S5H), suggesting an ecosystem-specific nature of viral distribution.

Using genomes from the CRVC, we investigated viral genetic diversity within crop root metagenomes. High intra-population genetic diversity (microdiversity) enhances a viral population's ability to adapt to environmental changes, ensuring its evolutionary potential.[75–77] Microdiversity was assessed by aligning metagenomic sequencing reads to a representative genome and calculating the coverage diversity of each nucleotide position. Quantitative analysis revealed 2,690 viral species-level clusters in crop root metagenomes (Table S5E). These viruses were categorized into four groups based on distribution patterns: rare, single-crop regional, single-crop multizonal, and multi-crops multizonal (STAR Methods). Most viruses (82.4%) were consistently detected, with 31.8%–58.5% appearing across multiple locations but within a single crop species (Figure S5I). Notably, 123 viruses were found across different locations and crop species, accounting for an average of 13.8% of the total viral abundance (Table S5E). Microdiversity analysis showed higher genetic diversity in viruses with broader distribution, especially those infecting diverse crop species and spanning geographic locations (Figure 5C), suggesting that ecological niche differentiation drives viral selective variation. Lytic viruses were more abundant in root environments than temperate viruses, but temperate viruses exhibited greater microdiversity (Figure S5J; Tables S5E and S5F), suggesting the interactions with bacteria may shape their evolutionary pathways.

Phage-bacteria interactions in crop root ecosystems

Utilizing genome and metagenome resources, we first investigated the abundance of viruses in crop root ecosystems. Metagenomic data from roots and their corresponding soils were annotated using the CRVC and IMG/VR databases (STAR Methods). We observed a significantly higher relative abundance of phages in root microbiomes than in the corresponding soils (Figure 6A; Table S5E; relative abundance, $mean_{root}$ = 0.00645%, $mean_{soil}$ = 0.00002%). This pattern was consistent across 14 datasets representing diverse soil sources and crop species. Similar results were obtained when using only the IMG/VR database as reference (Figure S6A). This analysis

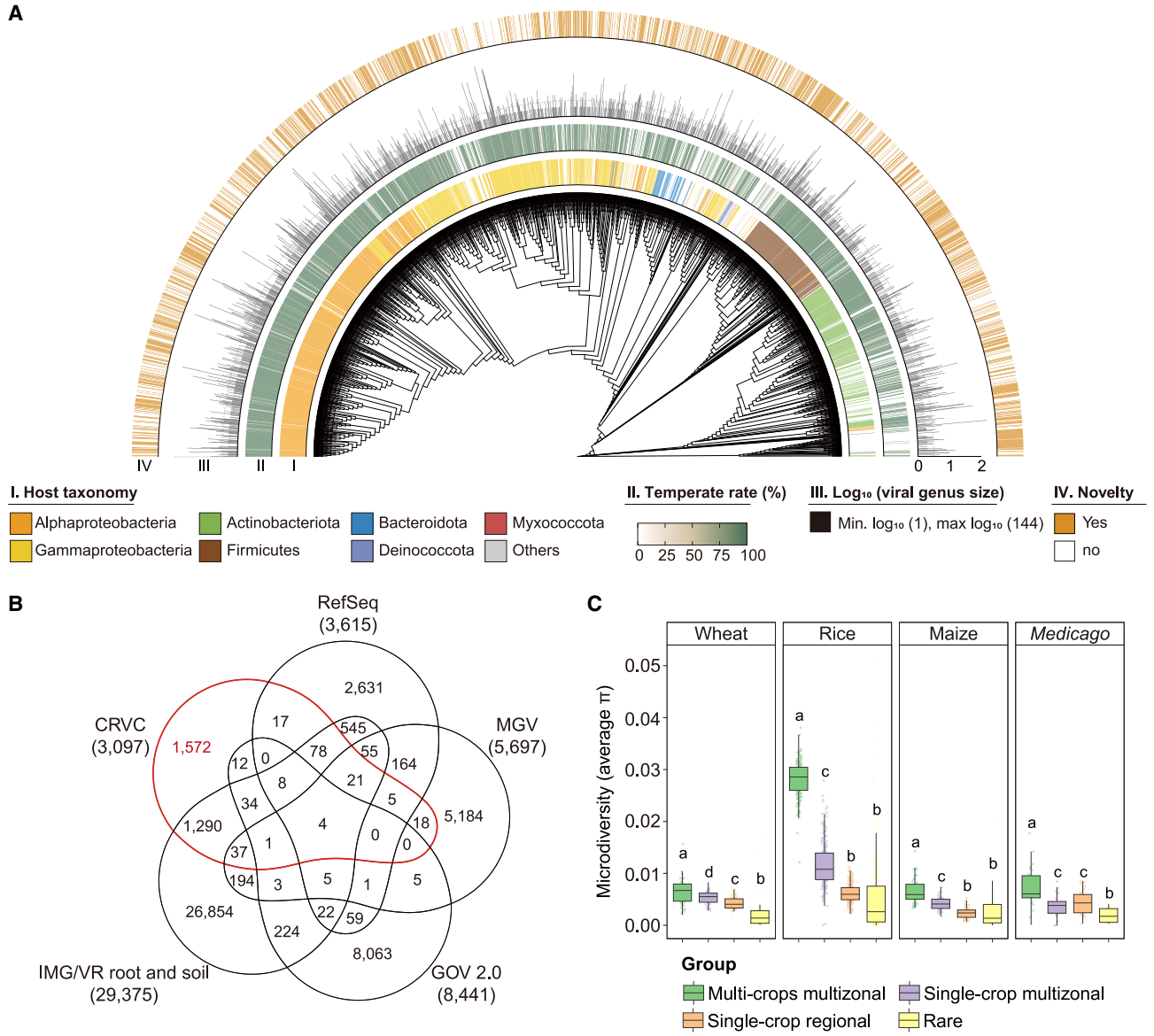


Figure 5. The CRVC unveils 1,572 unreported viral genus-level clusters and unique patterns in crop root microbiomes

(A) The CRVC reveals a striking diversity of unreported viral genus-level clusters and connections with crop root bacterial genomes. The hierarchical clustering dendrogram was built based on the clustered viral protein families in the CRVC using the Jaccard distance with the average clustering model (STAR Methods). The first circle (I) indicates the taxonomy of host bacteria for each viral cluster at the genus level, showing that 2,199 (71.0%) of CRVC clusters exhibit connections with crop root bacteria. Connections were determined according to viral contigs and CRISPR spacers detected in bacterial genomes. The second inner circle represents the proportion of temperate phages within each viral genus (II). Black bars show the $\log_{10}$-transformed genome numbers in each viral genus, ranging from 1 to 144 (III). In the outermost circle (IV), the gene-sharing network analysis reveals 1,572 unreported viral genus-level clusters in the CRVC compared with publicly available databases (STAR Methods).

(B) Half of the CRVC genus-level clusters are not reported in public databases. The Venn diagram illustrates overlaps of the CRVC genus-level clusters with the public viral databases including RefSeq, MGV, GOV2, and IMG/VR root and soil (STAR Methods).

(C) Viruses identified in root metagenomes across various crop species and geographical locations exhibit a higher microdiversity. Boxplots illustrate the microdiversity of viruses detected in roots of multiple crops and locations are higher compared with those detected in a single crop or location (adjusted $p < 0.05$, Kruskal-Wallis rank-sum test and Dunn's test; STAR Methods). Using prevalence characteristics, phages are categorized into four prevalence groups: stable presence across roots of multiple crops in multiple locations (multi-crops multizonal, green), within roots of a single crop species in multiple locations (single-crop multizonal, purple), within roots of a single crop species in a single location (single-crop regional, orange), and viruses with prevalence lower than 10% of samples (rare, yellow). Wheat, $n = 47$; rice, $n = 212$; maize, $n = 45$; *Medicago*, $n = 28$.

See also Figure S5.

Please cite this article in press as: Dai et al., Crop root bacterial and viral genomes reveal unexplored species and microbiome patterns, Cell (2025), https://doi.org/10.1016/j.cell.2025.02.013

CellPress
OPEN ACCESS

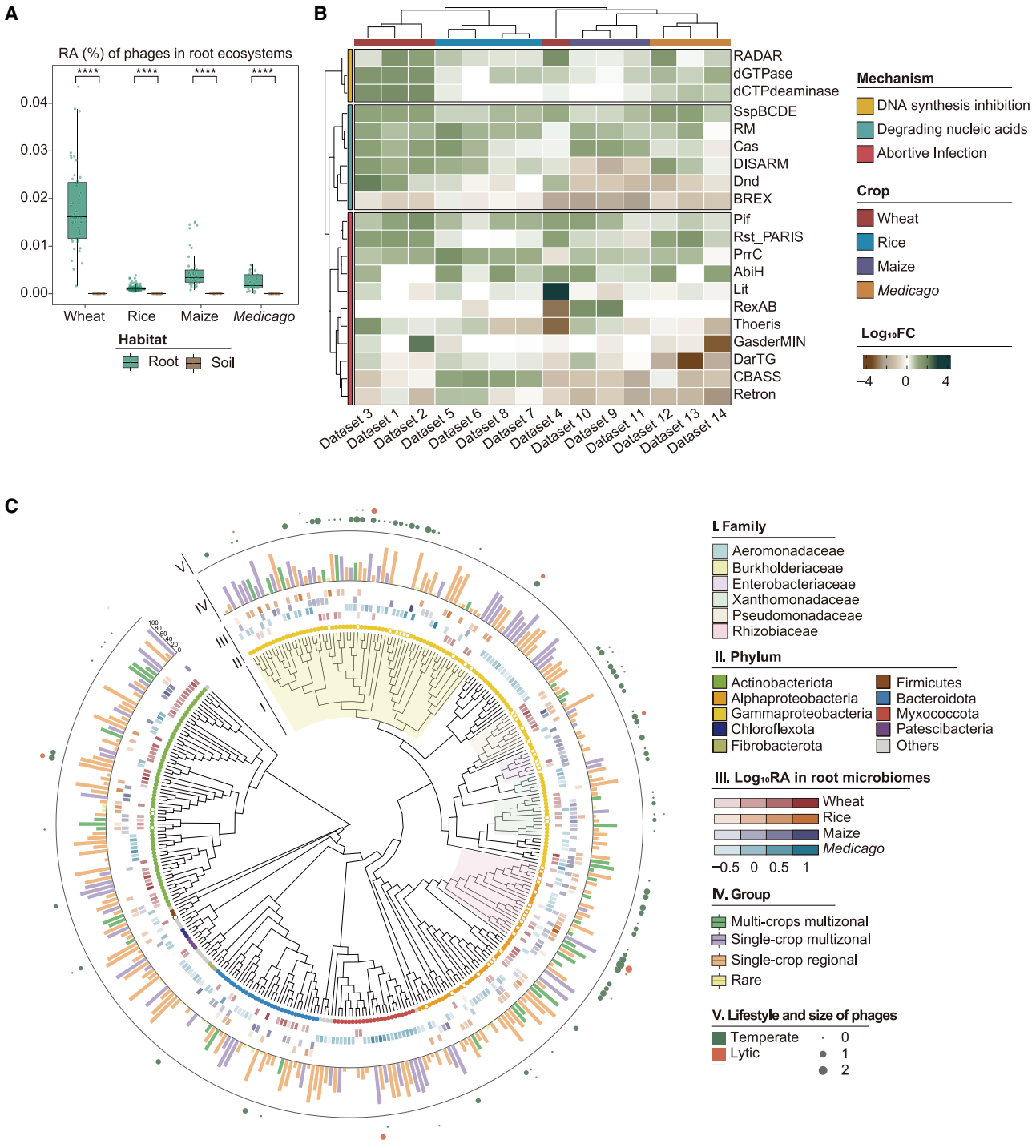


Figure 6. Phage-bacteria connections within crop root ecosystems

(A) Phage abundances in the root microbiomes are higher than those in corresponding soils. The box plot shows the relative abundances of phages in root and soil metagenomic samples from wheat, rice, maize, and *Medicago* (root, $n = 332$; soil, $n = 75$. ****adjusted $p < 0.0001$, Wilcoxon rank-sum test).

(B) The bacterial antiviral systems are enriched in the root microbiomes of multiple crops grown in diverse soil backgrounds. The heatmap illustrates the normalized $\log_{10}$-transformed fold changes in enrichment (green) and depletion (brown) of abundance concerning bacterial antiviral systems within the root microbiomes of wheat, rice, maize, and *Medicago* compared with their corresponding soils. Bacterial antiviral systems are categorized into three groups based on mechanisms: DNA synthesis inhibition, degrading nucleic acids, and abortive infection. Metagenomic datasets were clustered using the ward.D2 model.

(legend continued on next page)

Please cite this article in press as: Dai et al., Crop root bacterial and viral genomes reveal unexplored species and microbiome patterns, Cell (2025), https://doi.org/10.1016/j.cell.2025.02.013

highlights the adaptability of viruses colonizing the surfaces and interiors of roots.

Consistently, we observed that 94.0% of root bacterial genomes possessed at least one bacterial antiviral defense system, with a median of five defense families per genome (Figure S6B; Table S6A). In crop root microbiomes, most of the bacterial antiviral systems were more enriched compared with those in soils. Specifically, systems that inhibit DNA synthesis and degrade nucleic acids were consistently enriched in the roots across 14 datasets. For example, the R-M systems and CRISPR-Cas systems, two widespread and typical immune systems, exhibited convergent enrichment patterns in root microbiomes across crop species and soil sources (Figures 6B and S6C; Table S6B), suggesting the adaptability of root bacteria to phage-rich environments. In summary, the enrichment of both phages and bacterial antiviral defense systems within crop root microbiomes highlights the presence of complex and largely unexplored interactions between phages and root bacteria.

Next, we systematically investigated phage-bacteria interactions across root microbiomes from multiple crops and soil sources. We found that a total of 5,885 (60.2%) of the 9,772 crop root bacterial genomes had connections with phages, with phylum-specific proportions ranging from 50.9% to 83.5% (Figure S6D). These connections were identified through CRISPR spacer matches or the presence of phages (Table S6C), with most phages showing specific connections with distinct bacterial species (Figure S6E). Additionally, 27.0% of abundant and prevalent root bacteria exhibited connections with phages, as confirmed by both genomic matches and metagenomic co-occurrence (Figure 6C; Table S6D). These phage connections exhibited a preference for specific bacterial families, including Burkholderiaceae, Rhizobiaceae, Xanthomonadaceae, and Pseudomonadaceae (Figure 6C; Tables S6E and S6F; STAR Methods). Interestingly, phage-bacteria pairs confirmed by genomic matches exhibited significantly stronger associations within root ecosystems than those identified solely by co-occurrence analysis (Figure S6F; Table S6G), indicating that pairs confirmed by genomic matches are more reliable. Together, these findings underscore the pivotal role of the CRBC and CRVC resources in unraveling the intricate network of connections within the crop root microbiome.

DISCUSSION

Bacteria colonizing the surface and, more intimately, the interior of roots have significant impacts on crop growth and health.[1,2] Obtaining the genomic content of these bacteria and the corresponding isolates is essential for deciphering the mechanisms behind their interactions with host plants and for developing applications that benefit crop ecosystems. In this study, we established a systematic CRBC, comprising 6,699 genomes (68.9% from cultivated isolates) from three major grain crops wheat, rice, and maize, and the important forage legume *Medicago*, using two complementary methods: high-throughput bacterial cultivation and metagenome-derived MAGs (Figure 1). The CRBC contains 1,817 undefined bacterial species compared with all current public databases (Figure 2A), not only addressing the shortage of publicly available crop root bacterial genomes but also massively extending the diversity of bacterial genomes from all ecosystems.[31,46,45] Interestingly, 43.8% of crop root bacterial genomes encode at least two distinct PGP functional categories including nutrient utilization, biosynthesis of plant growth hormones, and stress resistance (Figure 3B)—functions that have garnered significant attention in recent years.[78,79] These genomic resources and extensive isolates provide unique opportunities to explore the roles and molecular mechanisms by which the root microbiome influences crop growth and health.[80–82]

Root metagenomics, specifically in the rhizoplane and endosphere, presents unique challenges compared with sequencing rhizosphere soil and gut microbiomes due to the high proportion of host DNA in root samples, which significantly raises the costs of capturing microbial taxa. These constraints have resulted in a shortage of crop root microbiome data, with limited samples in small-scale studies,[83] thus limiting our understanding of root metagenome patterns. In this study, we performed deep sequencing of 332 root metagenomic samples from 14 datasets and found that on average, 82.0% are host reads. Notably, bacteria dominate root microbiomes, constituting an average of 95.7% of the total microbial abundance, while fungi, archaea, viruses, and protists represented much smaller fractions (Figure S3C; Table S3I). This pattern is consistent with the bacterial dominance in the gut microbiome,[84,85] highlighting the importance of bacteria in the interactions between hosts and their associated microbiomes.

The CRBC represents a significant step forward in accurately quantifying and characterizing bacterial functional patterns within and across crop root ecosystems. Marker-gene-based profiling has already demonstrated that the taxonomic composition of the root microbiome is shaped by environmental factors and host plants,[2,6,11,12,14] such as the increase of Actinobacteriota in low-water soil environments.[9,80] Our comparison between metagenomes of roots and soils confirmed this taxonomic variation (Figure 4A). However, we were surprised to observe that root metagenome functions were conserved across 14 datasets, regardless of soil sources, host species, and genotypes

(C) The genome-level analysis reveals the broad connections of bacteria and phages in crop root ecosystems. The maximum-likelihood phylogenetic tree contains 330 bacterial species that are in the top ten relative abundances in the root microbiome of each metagenomic sample (n = 332; STAR Methods). The bacterial families showing intensive connections with phages are colored in the phylogenetic branches (I). The phylum information of each bacterial species is color-coded (II). Undefined crop root bacterial species are in solid points while published crop bacterial species are in hollow points. Log_{10}-transformed mean relative abundance of each bacterial species in the root microbiomes of each crop is shown in the heatmap, with each crop species color-coded (III). The prevalence of each bacterial species in crop root microbiomes are shown in the bar plot and categorized and color-coded into four groups (IV). The out layers (V) illustrate the genome-level connections of phages with each bacterial species in crop root microbiomes. The size of points reflects the log_{10}-transformed number of phage species interacting with each bacterial species. Temperate phages and lytic phages are color-coded in green and red, respectively. See also Table S6E.

See also Figure S6.

Please cite this article in press as: Dai et al., Crop root bacterial and viral genomes reveal unexplored species and microbiome patterns, Cell (2025), https://doi.org/10.1016/j.cell.2025.02.013

(Figure 4B). Genes associated with specific conserved functions, such as cell motility, secretion systems, and signaling transduction, have previously been identified in random barcoded transposon mutant sequencing in the laboratory in single plant species,[65,66] or genomic exploration of root microbes.[21,67] Our metagenomic results from diverse soil backgrounds, host plant species, and genotypes provide an ecological support for the significance and prevalence of these functions in crop root ecosystems. Given that similar conserved patterns are observed in the human gut microbiome across individuals with different backgrounds,[86,87] we hypothesize that microbiome-host interactions share conserved molecular principles. Our work provides valuable insights into potential targets for microbiome manipulation in agricultural applications.

Viruses have intrinsic connections with bacteria and are increasingly recognized for their roles in agriculture. However, their genomes and ecological patterns remain understudied in the crop root microbiome compared with other ecosystems.[88–91] Bacteria-phage connections represent an underexplored aspect in the crop root ecosystem, primarily due to a lack of genomic resources that limit systematic analysis. Our metagenomic samples from diverse crops and soil backgrounds revealed that nearly 30% of the prevalent root bacterial species across multiple crop species grown in various soils contained phage genomes or CRISPR spacer matches with phage genomes (Figure 6C). These connections showed a preference for specific bacterial families, including the abundant Rhizobiaceae, Xanthomonadaceae, and Pseudomonadaceae, which are known to play roles in crop growth and health.[92–95] Interestingly, viruses found across multiple hosts or locations exhibited significantly greater genetic diversity (Figure 5C). A similar pattern was observed in ocean viral data,[37] supporting the interpretation that the genetic diversity may contribute to viral speciation and adaptability in diverse environments.

Our genomes are valuable resources for exploring novel mechanisms and functional potentials within the crop root microbiome. For example, we have identified a type II-C CRISPR-Cas system promoting factor based on a *Chryseobacterium* genome in the CRBC.[96] The PcrIIC1 protein within the *CbCas9* (*Chryseobacterium* sp. Cas9, CbCas9) gene cluster mediates the dimerization of CbCas9, allowing it to tolerate mismatches in the spacer and protospacer adjacent motif (PAM) sequences, thereby improving bacterial immunity against phages.[96] This finding is in line with our observation that bacterial antiviral systems and viruses are both enriched in root microbiomes compared with soils (Figures 6A and 6B). In addition, by leveraging complete genome sequences, we examined the presence and coexistence of microbial functions at the genome level, insights that fragmented metagenomic sequencing cannot provide, thereby expanding our knowledge of potential beneficial bacteria. We also identified a significant number of undefined BGCs from diverse phylogenetic origins (Figure 3C), highlighting the genetic potential of the CRBC. Based on our bacterial and viral resources, we are establishing an international crop microbiome repository, which other scientists can continuously expand and benefit. This repository will serve as a valuable resource of genomes and isolates for the global crop root microbiome research community.

Together, the CRBC and CRVC genomic resources will serve an important role similar to that of well-known and widely used human and environmental microbiome genome resources,[31,46,88] advancing crop root microbiome research into mechanistic, genome-level understanding. The availability of a large number of isolates with accompanying genomic information will particularly stimulate functional and mechanistic studies of the crop root microbiome, ultimately benefiting applied agricultural sciences.

Limitations of the study

While the CRBC and CRVC provide comprehensive coverage of the dominant bacteria and viruses in the crop root microbiome, additional microbial genomes are needed particularly for the low abundant members of crop root ecosystems. The undefined species in this study are based solely on genome comparison; therefore, further investigation is still needed on their phenotypic and chemotaxonomic characteristics as well as ecological roles. These species also represent some rare cases of previously published isolates that lack genomes or associated metadata. Many bacterial and viral species-level clusters currently contain only one or two genomes, limiting the ability to identify their pangenomes and intraspecies diversity. Expanding microbial genomes to include crops grown in more diverse soil types and from various continents will also be crucial. Increased global cooperation to store and share root microbial genomes and isolates, with the standardized procedure, will be important and required in the future to advance root microbiome research to the genome and strain level.

RESOURCE AVAILABILITY

Lead contact

For further information concerning the genomes and bacterial isolates, please contact Dr. Yang Bai from Peking University (ybai@pku.edu.cn).

Materials availability

Details on CRBC bacterial isolate strains can be accessed through the Guangdong Microbial Culture Collection Center (GDMCC; https://gdmcc.net/#/index). Accession IDs for each isolate are listed in Table S1A. Furthermore, all cultivated bacterial strains are available upon request by contacting Dr. Yang Bai (www.cropmicrobiome.com; ybai@pku.edu.cn).

Data and code availability

The computational scripts for analysis in this work are available on GitHub (https://github.com/Bailab-pku/CRBC-and-CRVC). The bacterial and viral genomes, along with annotation files, and other related data are accessible on Zenodo (accessions: 14091751, 14095420, 13918137, and 13939322). Raw sequencing data for bacterial genomes and root metagenomes are available on NCBI (accession PRJNA1183633 and PRJNA1184367), and maize rhizosphere metagenome data are available on ENA (accession PRJEB77048). The genomic data in this work are also accessible through our website (http://www.cropmicrobiome.com/). The genomic datasets used in this manuscript are all publicly accessible. For details, please refer to the "deposited data" section in the key resources table.

ACKNOWLEDGMENTS

We thank Prof. Paul Schulze-Lefert (Max Planck Institute for Plant Breeding Research, Cologne, Germany) and Dr. Simon Roux (DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, USA) for their suggestions for improving the work. We thank Luisa Arias Giraldo and Nico Louwen

for their help with the design of the web portal. This work was supported by the National Key Research and Development Program of China (grant nos. 2022YFF1001800 to Y.B.), the Strategic Priority Research Program of the Chinese Academy of Sciences (grant nos. XDA24020000 to Y.B. and XDA28030202 and XDA0450000 to J.Z.), The National Natural Science Foundation of China (grant nos. 32430003 to Y.B. and 32322002 to J.Z.), the Joint CAS-MPG Research Project (grant no. HZXM20225001MI to Y.B.), the CAS Project for Young Scientists in Basic Research (grant no. YSBR-078 to Y.B.), the Scientific and Technological Innovation Project of China Academy of Chinese Medical Sciences (grant no. CI2024C007YN to Y.B.), WISH-ROOTS (2022-EJPSOILS-WISHROOTS-CAS to Y.B.), the Hainan Excellent Talent Team, Disciplinary Breakthrough Project of MOE, and the New Cornerstone Science Foundation through the XPLORER PRIZE to Y.B.

AUTHOR CONTRIBUTIONS

Y.B. designed and supervised the project. R.D. and F.L. performed the informatic analysis. S.C., A.L., and R.d.J. checked and supported the analysis. J.Z., H.X., J.-M.Q., and W.L. cultivated bacterial strains and collected crop root metagenomic samples. B.W., H.Z., L.J.U.P., and M.H.M. participated in the website construction. K.S. provided crop rhizosphere metagenomes in Switzerland. R.D., J.Z., F.L., A.L., R.d.J., C.M.J.P., K.S., and Y.B. wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Plant materials and growth conditions
- METHOD DETAILS
  - Bacterial isolation, cultivation and genomic DNA extraction
  - Whole-genome sequencing and *de novo* assembly of bacterial isolates
  - Harvesting and metagenomic sequencing of root microbiome samples
  - Quality control, assemble and gene prediction of metagenomic samples
  - Binning and refinement of MAGs
  - Collecting genomes of crop root bacteria from public databases
  - Quality control, dereplication, and species clustering of bacterial genomes
  - Comparison of bacterial species within the CRBC and representative species in the GTDB database
  - Taxonomic classification and phylogenetic analysis of root bacterial genomes
  - Reads coverage of root and rhizosphere metagenomic samples using different genome databases
  - Functional annotation and BGC and antiviral system identification of bacterial genomes
  - PGP functions in bacterial genomes
  - Comparison of BGCs in crop root bacterial genomes and public databases
  - Taxonomic classification and functional annotation of crop root metagenomic samples
  - Prediction and quality control of viral genomes
  - Gene prediction, functional annotation, and protein clustering of viral genomes
  - Clustering of viral genomes at the species and genus level
  - Taxonomic classification of the CRVC
  - Host prediction of the CRVC
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Quantification of bacterial species and evaluation of bacterial growth rate
  - Quantification of metagenomic reads and feature table generation
  - Rarefaction analysis and evaluation of microbial genetic potential
  - Quantification of viruses and antiviral defense systems in metagenomic samples
  - Microdiversity calculation of viral genomes
  - Definition of prevalence groups for bacterial and viral genomes
  - Statistic analysis and data visualization

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2025.02.013.

Received: June 29, 2024
Revised: November 14, 2024
Accepted: February 16, 2025
Published: March 12, 2025

REFERENCES

1. Bakker, P.A.H.M., Pieterse, C.M.J., de Jonge, R., and Berendsen, R.L. (2018). The soil-borne legacy. Cell *172*, 1178–1180. https://doi.org/10.1016/j.cell.2018.02.024.
2. Russ, D., Fitzpatrick, C.R., Teixeira, P.J.P.L., and Dangl, J.L. (2023). Deep discovery informs difficult deployment in plant microbiome science. Cell *186*, 4496–4513. https://doi.org/10.1016/j.cell.2023.08.035.
3. Trivedi, P., Leach, J.E., Tringe, S.G., Sa, T., and Singh, B.K. (2020). Plant-microbiome interactions: from community assembly to plant health. Nat. Rev. Microbiol. *18*, 607–621. https://doi.org/10.1038/s41579-020-0412-1.
4. Wang, F., Zhang, H., Liu, H., Wu, C., Wan, Y., Zhu, L., Yang, J., Cai, P., Chen, J., and Ge, T. (2024). Combating wheat yellow mosaic virus through microbial interactions and hormone pathway modulations. Microbiome *12*, 200. https://doi.org/10.1186/s40168-024-01911-z.
5. Zhang, J., Liu, Y.X., Zhang, N., Hu, B., Jin, T., Xu, H., Qin, Y., Yan, P., Zhang, X., Guo, X., et al. (2019). NRT1.1B is associated with root microbiota composition and nitrogen use in field-grown rice. Nat. Biotechnol. *37*, 676–684. https://doi.org/10.1038/s41587-019-0104-4.
6. Bulgarelli, D., Garrido-Oter, R., Münch, P.C., Weiman, A., Dröge, J., Pan, Y., McHardy, A.C., and Schulze-Lefert, P. (2015). Structure and function of the bacterial root microbiota in wild and domesticated barley. Cell Host Microbe *17*, 392–403. https://doi.org/10.1016/j.chom.2015.01.011.
7. He, X., Wang, D., Jiang, Y., Li, M., Delgado-Baquerizo, M., McLaughlin, C., Marcon, C., Guo, L., Baer, M., Moya, Y.A.T., et al. (2024). Heritable microbiome variation is correlated with source environment in locally adapted maize varieties. Nat. Plants *10*, 598–617. https://doi.org/10.1038/s41477-024-01654-7.
8. Carrión, V.J., Perez-Jaramillo, J., Cordovez, V., Tracanna, V., De Hollander, M., Ruiz-Buck, D., Mendes, L.W., van Ijcken, W.F.J., Gomez-Exposito, R., Elsayed, S.S., et al. (2019). Pathogen-induced activation of disease-suppressive functions in the endophytic root microbiome. Science *366*, 606–612. https://doi.org/10.1126/science.aaw9285.
9. Xu, L., Naylor, D., Dong, Z., Simmons, T., Pierroz, G., Hixson, K.K., Kim, Y.M., Zink, E.M., Engbrecht, K.M., Wang, Y., et al. (2018). Drought delays development of the sorghum root microbiome and enriches for monoderm bacteria. Proc. Natl. Acad. Sci. USA *115*, E4284–E4293. https://doi.org/10.1073/pnas.1717308115.
10. Yu, P., He, X., Baer, M., Beirinckx, S., Tian, T., Moya, Y.A.T., Zhang, X., Deichmann, M., Frey, F.P., Bresgen, V., et al. (2021). Plant flavones enrich rhizosphere Oxalobacteraceae to improve maize performance under nitrogen deprivation. Nat. Plants *7*, 481–499. https://doi.org/10.1038/s41477-021-00897-y.

11. Edwards, J., Johnson, C., Santos-Medellín, C., Lurie, E., Podishetty, N.K., Bhatnagar, S., Eisen, J.A., and Sundaresan, V. (2015). Structure, variation, and assembly of the root-associated microbiomes of rice. Proc. Natl. Acad. Sci. USA *112*, E911–E920. https://doi.org/10.1073/pnas.1414592112.

12. Peiffer, J.A., Spor, A., Koren, O., Jin, Z., Tringe, S.G., Dangl, J.L., Buckler, E.S., and Ley, R.E. (2013). Diversity and heritability of the maize rhizosphere microbiome under field conditions. Proc. Natl. Acad. Sci. USA *110*, 6548–6553. https://doi.org/10.1073/pnas.1302837110.

13. Bulgarelli, D., Rott, M., Schlaeppi, K., Ver Loren van Themaat, E., Ahmadinejad, N., Assenza, F., Rauf, P., Huettel, B., Reinhardt, R., Schmelzer, E., et al. (2012). Revealing structure and assembly cues for Arabidopsis root-inhabiting bacterial microbiota. Nature *488*, 91–95. https://doi.org/10.1038/nature11336.

14. Lundberg, D.S., Lebeis, S.L., Paredes, S.H., Yourstone, S., Gehring, J., Malfatti, S., Tremblay, J., Engelbrektson, A., Kunin, V., del Rio, T.G., et al. (2012). Defining the core Arabidopsis thaliana root microbiome. Nature *488*, 86–90. https://doi.org/10.1038/nature11237.

15. Garrido-Oter, R., Nakano, R.T., Dombrowski, N., Ma, K.W., AgBiome, T., McHardy, A.C., and Schulze-Lefert, P. (2018). Modular traits of the Rhizobiales root microbiota and their evolutionary relationship with symbiotic rhizobia. Cell Host Microbe *24*, 155–167.e155. https://doi.org/10.1016/j.chom.2018.06.006.

16. Rosconi, F., Rudmann, E., Li, J., Surujon, D., Anthony, J., Frank, M., Jones, D.S., Rock, C., Rosch, J.W., Johnston, C.D., et al. (2022). A bacterial pan-genome makes gene essentiality strain-dependent and evolvable. Nat. Microbiol. *7*, 1580–1592. https://doi.org/10.1038/s41564-022-01208-7.

17. Touchon, M., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S., Bidet, P., Bingen, E., Bonacorsi, S., Bouchier, C., Bouvet, O., et al. (2009). Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. PLoS Genet. *5*, e1000344. https://doi.org/10.1371/journal.pgen.1000344.

18. Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., Beghini, F., Manghi, P., Tett, A., Ghensi, P., et al. (2019). Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. Cell *176*, 649–662.e20. https://doi.org/10.1016/j.cell.2019.01.001.

19. Forster, S.C., Kumar, N., Anonye, B.O., Almeida, A., Viciani, E., Stares, M.D., Dunn, M., Mkandawire, T.T., Zhu, A., Shao, Y., et al. (2019). A human gut bacterial genome and culture collection for improved metagenomic analyses. Nat. Biotechnol. *37*, 186–192. https://doi.org/10.1038/s41587-018-0009-7.

20. Bai, Y., Müller, D.B., Srinivas, G., Garrido-Oter, R., Potthoff, E., Rott, M., Dombrowski, N., Münch, P.C., Spaepen, S., Remus-Emsermann, M., et al. (2015). Functional overlap of the Arabidopsis leaf and root microbiota. Nature *528*, 364–369. https://doi.org/10.1038/nature16192.

21. Levy, A., Gonzalez, I.S., Mittelviefhaus, M., Clingenpeel, S., Paredes, S.H., Miao, J., Wang, K., Devescovi, G., Stillman, K., Monteiro, F., et al. (2018). Genomic features of bacterial adaptation to plants. Nat. Genet. *50*, 138–150. https://doi.org/10.1038/s41588-017-0012-9.

22. Wippel, K., Tao, K., Niu, Y., Zgadzaj, R., Kiel, N., Guan, R., Dahms, E., Zhang, P., Jensen, D.B., Logemann, E., et al. (2021). Host preference and invasiveness of commensal bacteria in the Lotus and Arabidopsis root microbiota. Nat. Microbiol. *6*, 1150–1162. https://doi.org/10.1038/s41564-021-00941-9.

23. Thoenen, L., Kreuzer, M., Pestalozzi, C., Florean, M., Mateo, P., Züst, T., Wei, A., Giroud, C., Rouyer, L., Gfeller, V., et al. (2024). The lactonase BxdA mediates metabolic specialisation of maize root bacteria to benzoxazinoids. Nat. Commun. *15*, 6535. https://doi.org/10.1038/s41467-024-49643-w.

24. Grubbs, K.J., Bleich, R.M., Santa Maria, K.C., Allen, S.E., Farag, S., AgBiome, T., Shank, E.A., and Bowers, A.A. (2017). Large-scale bioinformatics analysis of bacillus genomes uncovers conserved roles of natural products in bacterial physiology. mSystems *2*, e00040–e00017. https://doi.org/10.1128/mSystems.00040-17.

25. Nayfach, S., and Pollard, K.S. (2016). Toward accurate and quantitative comparative metagenomics. Cell *166*, 1103–1116. https://doi.org/10.1016/j.cell.2016.08.007.

26. Cavassim, M.I.A., Moeskjær, S., Moslemi, C., Fields, B., Bachmann, A., Vilhjálmsson, B.J., Schierup, M.H., W Young, J.P., and Andersen, S.U. (2020). Symbiosis genes show a unique pattern of introgression and selection within a Rhizobium leguminosarum species complex. Microb. Genom. *6*, e000351. https://doi.org/10.1099/mgen.0.000351.

27. Coombs, J.T., and Franco, C.M.M. (2003). Isolation and identification of Actinobacteria from surface-sterilized wheat roots. Appl. Environ. Microbiol. *69*, 5603–5608. https://doi.org/10.1128/AEM.69.9.5603-5608.2003.

28. Greenlon, A., Chang, P.L., Damtew, Z.M., Muleta, A., Carrasquilla-Garcia, N., Kim, D., Nguyen, H.P., Suryawanshi, V., Krieg, C.P., Yadav, S.K., et al. (2019). Global-level population genomics reveals differential effects of geography and phylogeny on horizontal gene transfer in soil bacteria. Proc. Natl. Acad. Sci. USA *116*, 15200–15209. https://doi.org/10.1073/pnas.1900056116.

29. Epstein, B., Abou-Shanab, R.A.I., Shamseldin, A., Taylor, M.R., Guhlin, J., Burghardt, L.T., Nelson, M., Sadowsky, M.J., and Tiffin, P. (2018). Genome-wide association analyses in the model rhizobium Ensifer meliloti. mSphere *3*, e00386–e00318. https://doi.org/10.1128/mSphere.00386-18.

30. Almeida, A., Mitchell, A.L., Boland, M., Forster, S.C., Gloor, G.B., Tarkowska, A., Lawley, T.D., and Finn, R.D. (2019). A new genomic blueprint of the human gut microbiota. Nature *568*, 499–504. https://doi.org/10.1038/s41586-019-0965-1.

31. Paoli, L., Ruscheweyh, H.J., Forneris, C.C., Hubrich, F., Kautsar, S., Bhushan, A., Lotti, A., Clayssen, Q., Salazar, G., Milanese, A., et al. (2022). Biosynthetic potential of the global ocean microbiome. Nature *607*, 111–118. https://doi.org/10.1038/s41586-022-04862-3.

32. Wang, X., Wei, Z., Yang, K., Wang, J., Jousset, A., Xu, Y., Shen, Q., and Friman, V.P. (2019). Phage combination therapies for bacterial wilt disease in tomato. Nat. Biotechnol. *37*, 1513–1520. https://doi.org/10.1038/s41587-019-0328-3.

33. Kauffman, K.M., Hussain, F.A., Yang, J., Arevalo, P., Brown, J.M., Chang, W.K., VanInsberghe, D., Elsherbini, J., Sharma, R.S., Cutler, M.B., et al. (2018). A major lineage of non-tailed dsDNA viruses as unrecognized killers of marine bacteria. Nature *554*, 118–122. https://doi.org/10.1038/nature25474.

34. Paez-Espino, D., Eloe-Fadrosh, E.A., Pavlopoulos, G.A., Thomas, A.D., Huntemann, M., Mikhailova, N., Rubin, E., Ivanova, N.N., and Kyrpides, N.C. (2016). Uncovering earth's virome. Nature *536*, 425–430. https://doi.org/10.1038/nature19094.

35. Nayfach, S., Páez-Espino, D., Call, L., Low, S.J., Sberro, H., Ivanova, N.N., Proal, A.D., Fischbach, M.A., Bhatt, A.S., Hugenholtz, P., et al. (2021). Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. Nat. Microbiol. *6*, 960–970. https://doi.org/10.1038/s41564-021-00928-6.

36. Camargo, A.P., Nayfach, S., Chen, I.A., Palaniappan, K., Ratner, A., Chu, K., Ritter, S.J., Reddy, T.B.K., Mukherjee, S., Schulz, F., et al. (2023). IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata. Nucleic Acids Res. *51*, D733–D743. https://doi.org/10.1093/nar/gkac1037.

37. Gregory, A.C., Zayed, A.A., Conceição-Neto, N., Temperton, B., Bolduc, B., Alberti, A., Ardyna, M., Arkhipova, K., Carmichael, M., Cruaud, C., et al. (2019). Marine DNA viral macro- and microdiversity from pole to pole. Cell *177*, 1109–1123.e14. https://doi.org/10.1016/j.cell.2019.03.040.

38. Doron, S., Melamed, S., Ofir, G., Leavitt, A., Lopatina, A., Keren, M., Amitai, G., and Sorek, R. (2018). Systematic discovery of antiphage defense

systems in the microbial pangenome. Science *359*, eaar4120. https://doi.org/10.1126/science.aar4120.

39. Georjon, H., and Bernheim, A. (2023). Publisher Correction: The highly diverse antiphage defence systems of bacteria. Nat. Rev. Microbiol. *21*, 833. https://doi.org/10.1038/s41579-023-00987-y.
40. Al-Shayeb, B., Sachdeva, R., Chen, L.X., Ward, F., Munk, P., Devoto, A., Castelle, C.J., Olm, M.R., Bouma-Gregson, K., Amano, Y., et al. (2020). Clades of huge phages from across earth's ecosystems. Nature *578*, 425–431. https://doi.org/10.1038/s41586-020-2007-4.
41. Parks, D.H., Chuvochina, M., Rinke, C., Mussig, A.J., Chaumeil, P.-A., and Hugenholtz, P. (2022). GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. Nucleic Acids Res. *50*, D785–D794. https://doi.org/10.1093/nar/gkab776.
42. O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. *44*, D733–D745. https://doi.org/10.1093/nar/gkv1189.
43. Chen, I.A., Chu, K., Palaniappan, K., Pillay, M., Ratner, A., Huang, J., Huntemann, M., Varghese, N., White, J.R., Seshadri, R., et al. (2019). IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. Nucleic Acids Res. *47*, D666–D677. https://doi.org/10.1093/nar/gky901.
44. Yuan, D., Ahamed, A., Burgin, J., Cummins, C., Devraj, R., Gueye, K., Gupta, D., Gupta, V., Haseeb, M., Ihsan, M., et al. (2024). The European nucleotide archive in 2023. Nucleic Acids Res. *52*, D92–D97. https://doi.org/10.1093/nar/gkad1067.
45. Nayfach, S., Roux, S., Seshadri, R., Udwary, D., Varghese, N., Schulz, F., Wu, D., Paez-Espino, D., Chen, I.M., Huntemann, M., et al. (2021). A genomic catalog of earth's microbiomes. Nat. Biotechnol. *39*, 499–509. https://doi.org/10.1038/s41587-020-0718-6.
46. Almeida, A., Nayfach, S., Boland, M., Strozzi, F., Beracochea, M., Shi, Z.J., Pollard, K.S., Sakharova, E., Parks, D.H., Hugenholtz, P., et al. (2021). A unified catalog of 204,938 reference genomes from the human gut microbiome. Nat. Biotechnol. *39*, 105–114. https://doi.org/10.1038/s41587-020-0603-3.
47. Kuypers, M.M.M., Marchant, H.K., and Kartal, B. (2018). The microbial nitrogen-cycling network. Nat. Rev. Microbiol. *16*, 263–276. https://doi.org/10.1038/nrmicro.2018.9.
48. McGrath, J.W., Chin, J.P., and Quinn, J.P. (2013). Organophosphonates revealed: new insights into the microbial metabolism of ancient molecules. Nat. Rev. Microbiol. *11*, 412–419. https://doi.org/10.1038/nrmicro3011.
49. Zeng, J., Tu, Q., Yu, X., Qian, L., Wang, C., Shu, L., Liu, F., Liu, S., Huang, Z., He, J., et al. (2022). PCycDB: a comprehensive and accurate database for fast analysis of phosphorus cycling genes. Microbiome *10*, 101. https://doi.org/10.1186/s40168-022-01292-1.
50. Dai, Z., Liu, G., Chen, H., Chen, C., Wang, J., Ai, S., Wei, D., Li, D., Ma, B., Tang, C., et al. (2020). Long-term nutrient inputs shift soil microbial functional profiles of phosphorus cycling in diverse agroecosystems. ISME J. *14*, 757–770. https://doi.org/10.1038/s41396-019-0567-9.
51. Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. *28*, 27–30. https://doi.org/10.1093/nar/28.1.27.
52. Nett, R.S., Montanares, M., Marcassa, A., Lu, X., Nagel, R., Charles, T.C., Hedden, P., Rojas, M.C., and Peters, R.J. (2017). Elucidation of gibberellin biosynthesis in bacteria reveals convergent evolution. Nat. Chem. Biol. *13*, 69–74. https://doi.org/10.1038/nchembio.2232.
53. Salazar-Cerezo, S., Martinez-Montiel, N., Garcia-Sanchez, J., Perez-y-Terron, R., and Martinez-Contreras, R.D. (2018). Gibberellin biosynthesis and metabolism: a convergent route for plants, fungi and bacteria. Microbiol. Res. *208*, 85–98. https://doi.org/10.1016/j.micres.2018.01.010.
54. Kieber, J.J., and Schaller, G.E. (2018). Cytokinin signaling in plant development. Development *145*, dev149344. https://doi.org/10.1242/dev.149344.
55. Frébortová, J., and Frébort, I. (2021). Biochemical and structural aspects of cytokinin biosynthesis and degradation in bacteria. Microorganisms *9*, 1314. https://doi.org/10.3390/microorganisms9061314.
56. Zhang, P., Jin, T., Kumar Sahu, S., Xu, J., Shi, Q., Liu, H., and Wang, Y. (2019). The distribution of tryptophan-dependent indole-3-acetic acid synthesis pathways in bacteria unraveled by large-scale genomic analysis. Molecules *24*, 1411. https://doi.org/10.3390/molecules24071411.
57. Glick, B.R., Cheng, Z., Czarny, J., and Duan, J. (2007). Promotion of plant growth by acc deaminase-producing soil bacteria. Eur. J. Plant Pathol. *119*, 329–339. https://doi.org/10.1007/s10658-007-9162-4.
58. Ding, P., and Ding, Y. (2020). Stories of salicylic acid: a plant defense hormone. Trends Plant Sci. *25*, 549–565. https://doi.org/10.1016/j.tplants.2020.01.004.
59. Kerbarh, O., Ciulli, A., Howard, N.I., and Abell, C. (2005). Salicylate biosynthesis: overexpression, purification, and characterization of Irp9, a bifunctional salicylate synthase from Yersinia enterocolitica. J. Bacteriol. *187*, 5061–5066. https://doi.org/10.1128/JB.187.15.5061-5066.2005.
60. Bhattacharyya, P.N., and Jha, D.K. (2012). Plant growth-promoting rhizobacteria (PGPR): emergence in agriculture. World J. Microbiol. Biotechnol. *28*, 1327–1350. https://doi.org/10.1007/s11274-011-0979-9.
61. Hayat, R., Ali, S., Amara, U., Khalid, R., and Ahmed, I. (2010). Soil beneficial bacteria and their role in plant growth promotion: a review. Ann. Microbiol. *60*, 579–598. https://doi.org/10.1007/s13213-010-0117-1.
62. Wang, X., Zhang, J., Lu, X., Bai, Y., and Wang, G. (2024). Two diversities meet in the rhizosphere: root specialized metabolites and microbiome. J. Genet. Genomics *51*, 467–478. https://doi.org/10.1016/j.jgg.2023.10.004.
63. Kautsar, S.A., Blin, K., Shaw, S., Weber, T., and Medema, M.H. (2021). BiG-FAM: the biosynthetic gene cluster families database. Nucleic Acids Res. *49*, D490–D497. https://doi.org/10.1093/nar/gkaa812.
64. Terlouw, B.R., Blin, K., Navarro-Muñoz, J.C., Avalon, N.E., Chevrette, M.G., Egbert, S., Lee, S., Meijer, D., Recchia, M.J.J., Reitz, Z.L., et al. (2023). MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters. Nucleic Acids Res. *51*, D603–D610. https://doi.org/10.1093/nar/gkac1049.
65. Cole, B.J., Feltcher, M.E., Waters, R.J., Wetmore, K.M., Mucyn, T.S., Ryan, E.M., Wang, G., Ul-Hasan, S., McDonald, M., Yoshikuni, Y., et al. (2017). Genome-wide identification of bacterial plant colonization genes. PLoS Biol. *15*, e2002860. https://doi.org/10.1371/journal.pbio.2002860.
66. Georgoulis, S.J., Shalvarjian, K.E., Helmann, T.C., Hamilton, C.D., Carlson, H.K., Deutschbauer, A.M., and Lowe-Power, T.M. (2021). Genome-wide identification of tomato xylem sap fitness factors for three plant-pathogenic Ralstonia species. mSystems *6*, e0122921. https://doi.org/10.1128/mSystems.01229-21.
67. Vannier, N., Mesny, F., Getzke, F., Chesneau, G., Dethier, L., Ordon, J., Thiergart, T., and Hacquard, S. (2023). Genome-resolved metatranscriptomics reveals conserved root colonization determinants in a synthetic microbiota. Nat. Commun. *14*, 8274. https://doi.org/10.1038/s41467-023-43688-z.
68. Piel, D., Bruto, M., Labreuche, Y., Blanquart, F., Goudenège, D., Barcia-Cruz, R., Chenivesse, S., Le Panse, S., James, A., Dubert, J., et al. (2022). Phage-host coevolution in natural populations. Nat. Microbiol. *7*, 1075–1086. https://doi.org/10.1038/s41564-022-01157-1.
69. Camargo, A.P., Roux, S., Schulz, F., Babinski, M., Xu, Y., Hu, B., Chain, P.S.G., Nayfach, S., and Kyrpides, N.C. (2024). Identification of mobile genetic elements with geNomad. Nat. Biotechnol. *42*, 1303–1312. https://doi.org/10.1038/s41587-023-01953-y.

70. Roux, S., Enault, F., Hurwitz, B.L., and Sullivan, M.B. (2015). VirSorter: mining viral signal from microbial genomic data. PeerJ *3*, e985. https://doi.org/10.7717/peerj.985.

71. Ren, J., Song, K., Deng, C., Ahlgren, N.A., Fuhrman, J.A., Li, Y., Xie, X., Poplin, R., and Sun, F. (2020). Identifying viruses from metagenomic data using deep learning. Quant. Biol. *8*, 64–77. https://doi.org/10.1007/s40484-019-0187-4.

72. Roux, S., Adriaenssens, E.M., Dutilh, B.E., Koonin, E.V., Kropinski, A.M., Krupovic, M., Kuhn, J.H., Lavigne, R., Brister, J.R., Varsani, A., et al. (2019). Minimum information about an uncultivated virus genome (MIUViG). Nat. Biotechnol. *37*, 29–37. https://doi.org/10.1038/nbt.4306.

73. Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., et al. (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.

74. Trgovec-Greif, L., Hellinger, H.J., Mainguy, J., Pfundner, A., Frishman, D., Kiening, M., Webster, N.S., Laffy, P.W., Feichtinger, M., and Rattei, T. (2024). VOGDB-database of virus orthologous groups. Viruses *16*, 1191. https://doi.org/10.3390/v16081191.

75. Siranosian, B.A., Tamburini, F.B., Sherlock, G., and Bhatt, A.S. (2020). Acquisition, transmission and strain diversity of human gut-colonizing crAss-like phages. Nat. Commun. *11*, 280. https://doi.org/10.1038/s41467-019-14103-3.

76. Shah, S.A., Deng, L., Thorsen, J., Pedersen, A.G., Dion, M.B., Castro-Mejía, J.L., Silins, R., Romme, F.O., Sausset, R., Jessen, L.E., et al. (2023). Expanding known viral diversity in the healthy infant gut. Nat. Microbiol. *8*, 986–998. https://doi.org/10.1038/s41564-023-01345-7.

77. Mavrich, T.N., and Hatfull, G.F. (2017). Bacteriophage evolution differs by host, lifestyle and genome. Nat. Microbiol. *2*, 17112. https://doi.org/10.1038/nmicrobiol.2017.112.

78. Compant, S., Clément, C., and Sessitsch, A. (2010). Plant growth-promoting bacteria in the rhizo- and endosphere of plants: their role, colonization, mechanisms involved and prospects for utilization. Soil Biol. Biochem. *42*, 669–678. https://doi.org/10.1016/j.soilbio.2009.11.024.

79. Finkel, O.M., Salas-González, I., Castrillo, G., Conway, J.M., Law, T.F., Teixeira, P.J.P.L., Wilson, E.D., Fitzpatrick, C.R., Jones, C.D., and Dangl, J.L. (2020). A single bacterial genus maintains root growth in a complex microbiome. Nature *587*, 103–108. https://doi.org/10.1038/s41586-020-2778-7.

80. Santos-Medellín, C., Liechty, Z., Edwards, J., Nguyen, B., Huang, B., Weimer, B.C., and Sundaresan, V. (2021). Prolonged drought imparts lasting compositional changes to the rice root microbiome. Nat. Plants *7*, 1065–1077. https://doi.org/10.1038/s41477-021-00967-1.

81. Wagner, M.R., Tang, C., Salvato, F., Clouse, K.M., Bartlett, A., Vintila, S., Phillips, L., Sermons, S., Hoffmann, M., Balint-Kurti, P.J., et al. (2021). Microbe-dependent heterosis in maize. Proc. Natl. Acad. Sci. USA *118*, e2021965118. https://doi.org/10.1073/pnas.2021965118.

82. Zhang, L., Zhang, M., Huang, S., Li, L., Gao, Q., Wang, Y., Zhang, S., Huang, S., Yuan, L., Wen, Y., et al. (2022). A highly conserved core bacterial microbiota with nitrogen-fixation capacity inhabits the xylem sap in maize plants. Nat. Commun. *13*, 3361. https://doi.org/10.1038/s41467-022-31113-w.

83. Ofek-Lalzar, M., Sela, N., Goldman-Voronov, M., Green, S.J., Hadar, Y., and Minz, D. (2014). Niche and host-associated functional signatures of the root surface microbiome. Nat. Commun. *5*, 4950. https://doi.org/10.1038/ncomms5950.

84. Yan, Q., Li, S., Yan, Q., Huo, X., Wang, C., Wang, X., Sun, Y., Zhao, W., Yu, Z., Zhang, Y., et al. (2024). A genomic compendium of cultivated human gut fungi characterizes the gut mycobiome and its relevance to common diseases. Cell *187*, 2969–2989.e24. https://doi.org/10.1016/j.cell.2024.04.043.

85. Zou, Y., Xue, W., Luo, G., Deng, Z., Qin, P., Guo, R., Sun, H., Xia, Y., Liang, S., Dai, Y., et al. (2019). 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. Nat. Biotechnol. *37*, 179–185. https://doi.org/10.1038/s41587-018-0008-8.

86. Human; Microbiome; Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. Nature *486*, 207–214. https://doi.org/10.1038/nature11234.

87. Lozupone, C.A., Stombaugh, J.I., Gordon, J.I., Jansson, J.K., and Knight, R. (2012). Diversity, stability and resilience of the human gut microbiota. Nature *489*, 220–230. https://doi.org/10.1038/nature11550.

88. Camarillo-Guerrero, L.F., Almeida, A., Rangel-Pineros, G., Finn, R.D., and Lawley, T.D. (2021). Massive expansion of human gut bacteriophage diversity. Cell *184*, 1098–1109.e9. https://doi.org/10.1016/j.cell.2021.01.029.

89. Brum, J.R., Ignacio-Espinoza, J.C., Roux, S., Doulcier, G., Acinas, S.G., Alberti, A., Chaffron, S., Cruaud, C., de Vargas, C., Gasol, J.M., et al. (2015). Ocean plankton. Patterns and ecological drivers of ocean viral communities. Science *348*, 1261498. https://doi.org/10.1126/science.1261498.

90. Ma, B., Wang, Y., Zhao, K., Stirling, E., Lv, X., Yu, Y., Hu, L., Tang, C., Wu, C., Dong, B., et al. (2024). Biogeographic patterns and drivers of soil viromes. Nat. Ecol. Evol. *8*, 717–728. https://doi.org/10.1038/s41559-024-02347-2.

91. Kwak, M.-J., Kong, H.G., Choi, K., Kwon, S.-K., Song, J.Y., Lee, J., Lee, P.A., Choi, S.Y., Seo, M., Lee, H.J., et al. (2018). Rhizosphere microbiome structure alters to enable wilt resistance in tomato. Nat. Biotechnol. *36*, 1100–1109. https://doi.org/10.1038/nbt.4232.

92. Compant, S., Nowak, J., Coenye, T., Clément, C., and Ait Barka, E. (2008). Diversity and occurrence of Burkholderia spp. in the natural environment. FEMS Microbiol. Rev. *32*, 607–626. https://doi.org/10.1111/j.1574-6976.2008.00113.x.

93. Mercado-Blanco, J., and Bakker, P.A.H.M. (2007). Interactions between plants and beneficial Pseudomonas spp.: exploiting bacterial traits for crop protection. Antonie Leeuwenhoek *92*, 367–389. https://doi.org/10.1007/s10482-007-9167-1.

94. Masson-Boivin, C., Giraud, E., Perret, X., and Batut, J. (2009). Establishing nitrogen-fixing symbiosis with legumes: how many rhizobium recipes? Trends Microbiol. *17*, 458–466. https://doi.org/10.1016/j.tim.2009.07.004.

95. Ryan, R.P., Vorhölter, F.-J., Potnis, N., Jones, J.B., Van Sluys, M.-A., Bogdanove, A.J., and Dow, J.M. (2011). Pathogenomics of Xanthomonas: understanding bacterium–plant interactions. Nat. Rev. Microbiol. *9*, 344–355. https://doi.org/10.1038/nrmicro2558.

96. Zhang, S., Sun, A., Qian, J.M., Lin, S., Xing, W., Yang, Y., Zhu, H.Z., Zhou, X.Y., Guo, Y.S., Liu, Y., et al. (2024). Pro-CRISPR PcrIIC1-associated Cas9 system for enhanced bacterial immunity. Nature *630*, 484–492. https://doi.org/10.1038/s41586-024-07486-x.

97. Basenko, E.Y., Pulman, J.A., Shanmugasundram, A., Harb, O.S., Crouch, K., Starns, D., Warrenfeltz, S., Aurrecoechea, C., Stoeckert, C.J., Jr., Kissinger, J.C., et al. (2018). FungiDB: an integrated bioinformatic resource for fungi and oomycetes. J. Fungi (Basel) *4*, 39. https://doi.org/10.3390/jof4010039.

98. Quiza, L., Tremblay, J., Greer, C.W., Hemmingsen, S.M., St-Arnaud, M., Pozniak, C.J., and Yergeau, E. (2021). Rhizosphere shotgun metagenomic analyses fail to show differences between ancestral and modern wheat genotypes grown under low fertilizer inputs. FEMS Microbiol. Ecol. *97*, fiab071. https://doi.org/10.1093/femsec/fiab071.

99. Cole, J.R., Wang, Q., Fish, J.A., Chai, B., McGarrell, D.M., Sun, Y., Brown, C.T., Porras-Alfaro, A., Kuske, C.R., and Tiedje, J.M. (2014). Ribosomal database project: data and tools for high throughput rRNA analysis. Nucleic Acids Res. *42*, D633–D642. https://doi.org/10.1093/nar/gkt1244.

100. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. *45*, D353–D361. https://doi.org/10.1093/nar/gkw1092.

101. Siddell, S.G., Smith, D.B., Adriaenssens, E., Alfenas-Zerbini, P., Dutilh, B.E., Garcia, M.L., Junglen, S., Krupovic, M., Kuhn, J.H., Lambert, A.J., et al. (2023). Virus taxonomy and the role of the international committee on taxonomy of viruses (ICTV). J. Gen. Virol. *104*, 001840. https://doi.org/10.1099/jgv.0.001840.

102. Mihara, T., Nishimura, Y., Shimizu, Y., Nishiyama, H., Yoshikawa, G., Uehara, H., Hingamp, P., Goto, S., and Ogata, H. (2016). Linking virus genomes with host taxonomy. Viruses *8*, 66. https://doi.org/10.3390/v8030066.

103. Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L., Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J., et al. (2021). Pfam: the protein families database in 2021. Nucleic Acids Res. *49*, D412–D419. https://doi.org/10.1093/nar/gkaa913.

104. Tesson, F., Hervé, A., Mordret, E., Touchon, M., d'Humières, C., Cury, J., and Bernheim, A. (2022). Systematic and quantitative view of the antiviral arsenal of prokaryotes. Nat. Commun. *13*, 2561. https://doi.org/10.1038/s41467-022-30269-9.

105. Zhang, J., Liu, Y.X., Guo, X., Qin, Y., Garrido-Oter, R., Schulze-Lefert, P., and Bai, Y. (2021). High-throughput cultivation and identification of bacteria from the plant root microbiota. Nat. Protoc. *16*, 988–1012. https://doi.org/10.1038/s41596-020-00444-7.

106. Edgar, R.C. (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics *26*, 2460–2461. https://doi.org/10.1093/bioinformatics/btq461.

107. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.

108. Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A., and Korobeynikov, A. (2020). Using SPAdes de novo assembler. Curr. Protoc. Bioinformatics *70*, e102. https://doi.org/10.1002/cpbi.102.

109. Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., and Tyson, G.W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. *25*, 1043–1055. https://doi.org/10.1101/gr.186072.114.

110. Uritskiy, G.V., DiRuggiero, J., and Taylor, J. (2018). MetaWRAP-a flexible pipeline for genome-resolved metagenomic data analysis. Microbiome *6*, 158. https://doi.org/10.1186/s40168-018-0541-1.

111. Wood, D.E., Lu, J., and Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. Genome Biol. *20*, 257. https://doi.org/10.1186/s13059-019-1891-0.

112. Lu, J., Rincon, N., Wood, D.E., Breitwieser, F.P., Pockrandt, C., Langmead, B., Salzberg, S.L., and Steinegger, M. (2022). Metagenome analysis using the Kraken software suite. Nat. Protoc. *17*, 2815–2839. https://doi.org/10.1038/s41596-022-00738-y.

113. Shen, W., Le, S., Li, Y., and Hu, F. (2016). SeqKit: a cross-platform and ultrafast toolkit for fasta/q file manipulation. PLoS One *11*, e0163962. https://doi.org/10.1371/journal.pone.0163962.

114. Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de bruijn graph. Bioinformatics *31*, 1674–1676. https://doi.org/10.1093/bioinformatics/btv033.

115. Hyatt, D., Chen, G.-L., LoCascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics *11*, 119. https://doi.org/10.1186/1471-2105-11-119.

116. Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics *28*, 3150–3152. https://doi.org/10.1093/bioinformatics/bts565.

117. Kang, D.D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., and Wang, Z. (2019). MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ *7*, e7359. https://doi.org/10.7717/peerj.7359.

118. Wu, Y.-W., Simmons, B.A., and Singer, S.W. (2016). MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics *32*, 605–607. https://doi.org/10.1093/bioinformatics/btv638.

119. Chan, P.P., Lin, B.Y., Mak, A.J., and Lowe, T.M. (2021). TRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. Nucleic Acids Res. *49*, 9077–9096. https://doi.org/10.1093/nar/gkab688.

120. Olm, M.R., Brown, C.T., Brooks, B., and Banfield, J.F. (2017). dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. ISME J. *11*, 2864–2868. https://doi.org/10.1038/ismej.2017.126.

121. Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S., and Phillippy, A.M. (2016). Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. *17*, 132. https://doi.org/10.1186/s13059-016-0997-x.

122. Marçais, G., Delcher, A.L., Phillippy, A.M., Coston, R., Salzberg, S.L., and Zimin, A. (2018). MUMmer4: a fast and versatile genome alignment system. PLoS Comput. Biol. *14*, e1005944. https://doi.org/10.1371/journal.pcbi.1005944.

123. Chaumeil, P.-A., Mussig, A.J., Hugenholtz, P., and Parks, D.H. (2019). GTDB-Tk: a toolkit to classify genomes with the genome taxonomy database. Bioinformatics *36*, 1925–1927. https://doi.org/10.1093/bioinformatics/btz848.

124. Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods *14*, 417–419. https://doi.org/10.1038/nmeth.4197.

125. Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. Nat. Methods *12*, 59–60. https://doi.org/10.1038/nmeth.3176.

126. Blin, K., Shaw, S., Kloosterman, A.M., Charlop-Powers, Z., van Wezel, G.P., Medema, M.H., and Weber, T. (2021). antiSMASH 6.0: improving cluster detection and comparison capabilities. Nucleic Acids Res. *49*, W29–W35. https://doi.org/10.1093/nar/gkab335.

127. Navarro-Muñoz, J.C., Selem-Mojica, N., Mullowney, M.W., Kautsar, S.A., Tryon, J.H., Parkinson, E.I., De Los Santos, E.L.C., Yeong, M., Cruz-Morales, P., Abubucker, S., et al. (2020). A computational framework to explore large-scale biosynthetic diversity. Nat. Chem. Biol. *16*, 60–68. https://doi.org/10.1038/s41589-019-0400-9.

128. Kautsar, S.A., van der Hooft, J.J.J., de Ridder, D., and Medema, M.H. (2021). BiG-SLiCE: a highly scalable tool maps the diversity of 1.2 million biosynthetic gene clusters. GigaScience *10*, giaa154. https://doi.org/10.1093/gigascience/giaa154.

129. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359. https://doi.org/10.1038/nmeth.1923.

130. Olm, M.R., Crits-Christoph, A., Bouma-Gregson, K., Firek, B.A., Morowitz, M.J., and Banfield, J.F. (2021). inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. Nat. Biotechnol. *39*, 727–736. https://doi.org/10.1038/s41587-020-00797-0.

131. Gao, Y., and Li, H. (2018). Quantifying and comparing bacterial growth dynamics in multiple metagenomic samples. Nat. Methods *15*, 1041–1044. https://doi.org/10.1038/s41592-018-0182-0.

132. Madeira, F., Madhusoodanan, N., Lee, J., Eusebi, A., Niewielska, A., Tivey, A.R.N., Lopez, R., and Butcher, S. (2024). The EMBL-EBI job bispatcher sequence analysis tools framework in 2024. Nucleic Acids Res. *52*, W521–W525. https://doi.org/10.1093/nar/gkae241.

133. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

134. Putri, G.H., Anders, S., Pyl, P.T., Pimanda, J.E., and Zanini, F. (2022). Analysing high-throughput sequencing data in python with HTSeq 2.0. Bioinformatics *38*, 2943–2945. https://doi.org/10.1093/bioinformatics/btac166.

135. Nayfach, S., Camargo, A.P., Schulz, F., Eloe-Fadrosh, E., Roux, S., and Kyrpides, N.C. (2021). CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nat. Biotechnol. *39*, 578–585. https://doi.org/10.1038/s41587-020-00774-7.

136. Kieft, K., Zhou, Z., and Anantharaman, K. (2020). VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. Microbiome *8*, 90. https://doi.org/10.1186/s40168-020-00867-0.

137. Steinegger, M., and 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.

138. Enright, A.J., Van Dongen, S., and Ouzounis, C.A. (2002). An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. *30*, 1575–1584. https://doi.org/10.1093/nar/30.7.1575.

139. Couvin, D., Bernheim, A., Toffano-Nioche, C., Touchon, M., Michalik, J., Néron, B., Rocha, E.P.C., Vergnaud, G., Gautheret, D., and Pourcel, C. (2018). CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. Nucleic Acids Res. *46*, W246–W251. https://doi.org/10.1093/nar/gky425.

140. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). Blast+: architecture and applications. BMC Bioinformatics *10*, 421. https://doi.org/10.1186/1471-2105-10-421.

141. Finn, R.D., Clements, J., and Eddy, S.R. (2011). HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. *39*, W29–W37. https://doi.org/10.1093/nar/gkr367.

142. Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., et al. (2019). Welcome to the tidyverse. J. Open Source Software *4*, 1686. https://doi.org/10.21105/joss.01686.

143. Pedersen, T.L. (2025). patchwork: the composer of plots. https://github.com/thomasp85/patchwork.

144. Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics *32*, 2847–2849. https://doi.org/10.1093/bioinformatics/btw313.

145. Garnier, S., Ross, N., Rudis, R., Camargo, A.P., Sciaini, M., and Scherer, C. (2024). viridis(Lite) – colorblind-friendly color maps for R. https://sjmgarnier.github.io/viridis/.

146. Oksanen, J., Simpson, G., Blanchet, F., Kindt, R., Legendre, P., Minchin, P., O'Hara, R., Solymos, P., Stevens, M., Szoecs, E., et al. (2025). vegan: community ecology package. https://github.com/vegandevs/vegan.

147. Kassambara, A. (2023). rstatix: pipe-friendly framework for basic statistical tests. https://rpkgs.datanovia.com/rstatix/.

148. Xu, S., Li, L., Luo, X., Chen, M., Tang, W., Zhan, L., Dai, Z., Lam, T.T., Guan, Y., and Yu, G. (2022). Ggtree: a serialized data object for visualization of a phylogenetic tree and annotation data. Imeta *1*, e56. https://doi.org/10.1002/imt2.56.

149. Xu, S., Dai, Z., Guo, P., Fu, X., Liu, S., Zhou, L., Tang, W., Feng, T., Chen, M., Zhan, L., et al. (2021). ggtreeExtra: compact visualization of richly annotated phylogenetic data. Mol. Biol. Evol. *38*, 4039–4042. https://doi.org/10.1093/molbev/msab166.

150. Revell, L.J. (2024). phytools 2.0: an updated R ecosystem for phylogenetic comparative methods (and other things). PeerJ *12*, e16505. https://doi.org/10.7717/peerj.16505.

151. Graves, S., Piepho, H.-P., Selzer, L., and Dorai-Raj, S. (2024). multcompView: visualizations of paired comparisons. https://github.com/lselzer/multcompview.

152. Gao, C.H., Chen, C., Akyol, T., Dusa, A., Yu, G., Cao, B., and Cai, P. (2024). ggVennDiagram: intuitive venn diagram software extended. Imeta *3*, e177. https://doi.org/10.1002/imt2.177.

153. Pohlert, T. (2024). PMCMRplus: calculate pairwise multiple comparisons of mean rank sums extended. https://CRAN.R-project.org/package=PMCMRplus.

154. Neuwirth, E. (2022). RColorBrewer: colorbrewer palettes. https://CRAN.R-project.org/package=RColorBrewer.

155. Wang, H., Xu, X., Vieira, F.G., Xiao, Y., Li, Z., Wang, J., Nielsen, R., and Chu, C. (2016). The power of inbreeding: NGS-based GWAS of rice reveals convergent evolution during rice domestication. Mol. Plant *9*, 975–985. https://doi.org/10.1016/j.molp.2016.04.018.

156. Bowers, R.M., Kyrpides, N.C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T.B.K., Schulz, F., Jarett, J., Rivers, A.R., Eloe-Fadrosh, E.A., et al. (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat. Biotechnol. *35*, 725–731. https://doi.org/10.1038/nbt.3893.

157. Sayers, E.W., Bolton, E.E., Brister, J.R., Canese, K., Chan, J., Comeau, D.C., Connor, R., Funk, K., Kelly, C., Kim, S., et al. (2022). Database resources of the national center for biotechnology information. Nucleic Acids Res. *50*, D20–D26. https://doi.org/10.1093/nar/gkab1112.

158. Parks, D.H., Rinke, C., Chuvochina, M., Chaumeil, P.A., Woodcroft, B.J., Evans, P.N., Hugenholtz, P., and Tyson, G.W. (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat. Microbiol. *2*, 1533–1542. https://doi.org/10.1038/s41564-017-0012-7.

159. Faith, D.P. (1992). Conservation evaluation and phylogenetic diversity. Biol. Conserv. *61*, 1–10. https://doi.org/10.1016/0006-3207(92)91201-3.

160. Hu, L., Robert, C.A.M., Cadot, S., Zhang, X., Ye, M., Li, B., Manzo, D., Chervet, N., Steinger, T., van der Heijden, M.G.A., et al. (2018). Root exudate metabolites drive plant-soil feedbacks on growth and defense by shaping the rhizosphere microbiota. Nat. Commun. *9*, 2738. https://doi.org/10.1038/s41467-018-05122-7.

161. Suzek, B.E., Huang, H., McGarvey, P., Mazumder, R., and Wu, C.H. (2007). UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics *23*, 1282–1288. https://doi.org/10.1093/bioinformatics/btm098.

162. UniProt Consortium (2007). The universal protein resource (UniProt). Nucleic Acids Res. *35*, D193–D197. https://doi.org/10.1093/nar/gkl929.

163. Bin Jang, H., Bolduc, B., Zablocki, O., Kuhn, J.H., Roux, S., Adriaenssens, E.M., Brister, J.R., Kropinski, A.M., Krupovic, M., Lavigne, R., et al. (2019). Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. Nat. Biotechnol. *37*, 632–639. https://doi.org/10.1038/s41587-019-0100-8.

164. Dion, M.B., Oechslin, F., and Moineau, S. (2020). Phage diversity, genomics and phylogeny. Nat. Rev. Microbiol. *18*, 125–138. https://doi.org/10.1038/s41579-019-0311-5.

165. Korem, T., Zeevi, D., Suez, J., Weinberger, A., Avnit-Sagi, T., Pompan-Lotan, M., Matot, E., Jona, G., Harmelin, A., Cohen, N., et al. (2015). Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples. Science *349*, 1101–1106. https://doi.org/10.1126/science.aac4812.

166. Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature *490*, 55–60. https://doi.org/10.1038/nature11450.

Please cite this article in press as: Dai et al., Crop root bacterial and viral genomes reveal unexplored species and microbiome patterns, Cell (2025), https://doi.org/10.1016/j.cell.2025.02.013

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial strains		
CRBC isolates	This paper	http://www.cropmicrobiome.com; https://gdmcc.net/#/index
Biological samples		
Root and soil samples of *Triticum aestivum* L.	This paper	N/A
Root and soil samples of *Oryza sativa* L.	This paper	N/A
Root and soil samples of *Zea mays* L.	This paper	N/A
Root and soil samples of *Medicago truncatula*	This paper	N/A
Deposited data		
Bacterial genomes of CRBC	This paper	http://www.cropmicrobiome.com; Zenodo: 14091751
Viral genomes of CRVC	This paper	http://www.cropmicrobiome.com; Zenodo: 14091751
Gene catalogs and KEGG annotations	This paper	http://www.cropmicrobiome.com; Zenodo: 14095420
BGCs of CRBC	This paper	http://www.cropmicrobiome.com; Zenodo: 13918137 and 13939322
Raw sequencing of root metagenomes and genomes from cultured bacteria	This paper	NCBI: PRJNA1183633 and PRJNA1184367
Maize rhizosphere metagenomic samples	This paper	ENA: PRJEB77048
NCBI bacterial genomes	NCBI	https://ftp.ncbi.nlm.nih.gov/genomes/genbank
NCBI RefSeq bacterial and viral genomes	NCBI	https://ftp.ncbi.nlm.nih.gov/genomes/refseq/
IMG/M bacterial genome dataset	JGI[43]	https://img.jgi.doe.gov/cgi-bin/m/main.cgi
IMG/VR Viral Resources v4	JGI[36]	https://img.jgi.doe.gov/cgi-bin/vrer/main.cgi
FungiDB	Basenko et al.[97]	https://fungidb.org/fungidb/app
Chickpea root bacterial symbionts	Greenlon et al.[28]	NCBI: PRJNA453501
Rhizobiales genome collection	Garrido-Oter et al.[15]	ENA: PRJEB26998
Lotus japonicus root culture collection	Wippel et al.[22]	ENA: PRJEB37696
Rhizobium leguminosarum collection	Cavassim et al.[26]	NCBI: PRJNA510726
Ensifer spp collection	Epstein et al.[29]	NCBI: PRJNA401434
Bacillus genomes collection	Grubbs et al.[24]	NCBI: PRJNA400804
Wheat rhizosphere metagenomic samples	Quiza et al.[98]	NCBI: PRJNA643787
Genome Taxonomy Database (GTDB)	GTDB team[41]	https://gtdb.ecogenomic.org/
GEM database	Nayfach et al.[45]	https://portal.nersc.gov/GEM
UHGG database	Almeida et al.[46]	http://ftp.ebi.ac.uk/pub/databases/metagenomics/mgnify_genomes/
OMG database	Paoli et al.[31]	https://microbiomics.io/ocean/
RDP database	Cole et al.[99]	https://mothur.org/wiki/rdp_reference_files/
BiG-FAM database	Kautsar et al.[63]	https://bigfam.bioinformatics.nl
MIBiG database	Terlouw et al.[64]	https://mibig.secondarymetabolites.org
KEGG database	Kanehisa et al.[100]	https://www.genome.jp/kegg/

(*Continued on next page*)

Please cite this article in press as: Dai et al., Crop root bacterial and viral genomes reveal unexplored species and microbiome patterns, Cell (2025), https://doi.org/10.1016/j.cell.2025.02.013

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
MGV database	Nayfach et al.[35]	https://portal.nersc.gov/MGV/
GOV 2.0 database	Gregory et al.[37]	https://datacommons.cyverse.org/browse/iplant/home/shared/iVirus/GOV2.0
ICTV database	Siddell et al.[101]	https://ictv.global/
Virus-Host Dtabase	Mihara et al.[102]	https://www.genome.jp/virushostdb/
Pfam database	Mistry et al.[103]	http://pfam.xfam.org/
VOGDB database	Trgovec-Greif et al.[74]	http://vogdb.org
DefenseFinder Structures DB	Tesson et al.[104]	https://defensefinder.mdmlab.fr/wiki/structure
Oligonucleotides		
27F (AGAGTTTGATCCTGGCTCAG)	Zhang et al.[105]	16S primers for Sanger sequencing
1492R (TACGGCTACCTTGTTACGACTT)	Zhang et al.[105]	16S primers for Sanger sequencing
799F (AACMGGATTAGATACCCKG)	Zhang et al.[105]	First step primers for high throughput identification
1193R (ACGTCATCCCCACCTTCC)	Zhang et al.[105]	First step primers for high throughput identification
Software and algorithms		
USEARCH	Edgar et al.[106]	http://www.drive5.com/usearch/
Trimmomatic v0.39	Bolger et al.[107]	https://github.com/usadellab/Trimmomatic
SPAdes v3.14.0	Prjibelski et al.[108]	https://github.com/ablab/spades
CheckM v1.1.8	Parks et al.[109]	https://github.com/Ecogenomics/CheckM
MetaWRAP v1.3.2	Uritskiy et al.[110]	https://github.com/bxlab/metaWRAP
KneadData v0.7.6	The huttenhower lab	https://github.com/biobakery/kneaddata
Kraken2 v2.1.1	Wood et al.[111]	https://github.com/DerrickWood/kraken2
Kraken tool extract_kraken_reads.py	Lu et al.[112]	https://github.com/jenniferlu717/KrakenTools/blob/master/extract_kraken_reads.py
SeqKit	Shen et al.[113]	https://bioinf.shenwei.me/seqkit/
MEGAHIT v1.2.9	Li et al.[114]	https://github.com/voutcn/megahit
Prodigal v2.6.3	Hyatt et al.[115]	https://github.com/hyattpd/Prodigal
CD-HIT v.4.8.1	Fu et al.[116]	https://sites.google.com/view/cd-hit
MetaBAT2	Kang et al.[117]	https://bitbucket.org/berkeleylab/metabat
MaxBin2	Wu et al.[118]	http://sourceforge.net/projects/maxbin/
tRNAscan-SE v 2.0.12	Chan et al.[119]	http://trna.ucsc.edu/tRNAscan-SE/
barrnap v0.9	Torsten Seemann	https://github.com/tseemann/barrnap
dRep v3.2.2	Olm et al.[120]	https://github.com/MrOlm/drep
Mash v2.3	Ondov et al.[121]	https://github.com/marbl/mash
MUMmer v4.0.0	Marcais et al.[122]	https://github.com/mummer4/mummer
GTDB-Tk v2.0.0	Chaumeil et al.[123]	https://github.com/Ecogenomics/GTDBTk
Salmon v1.9.0	Patro et al.[124]	https://combine-lab.github.io/salmon/
DIAMOND v2.0.15.153	Buchfink et al.[125]	https://github.com/bbuchfink/diamond
antiSMASH v6.1.0	Blin et al.[126]	https://antismash.secondarymetabolites.org/#!/start
BiG-SCAPE v1.1.5	Navarro-Muñoz et al.[127]	https://github.com/medema-group/BiG-SCAPE
DefenseFinder v1.2.2	Tesson et al.[104]	https://defensefinder.mdmlab.fr/
BiG-SLICE v1	Kautsar et al.[128]	https://github.com/medema-group/bigslice
Bowtie 2 v2.4.5	Langmead and Salzberg[129]	https://bowtie-bio.sourceforge.net/bowtie2/index.shtml

(*Continued on next page*)

Please cite this article in press as: Dai et al., Crop root bacterial and viral genomes reveal unexplored species and microbiome patterns, Cell (2025), https://doi.org/10.1016/j.cell.2025.02.013

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
InStrain v1.8.0	Olm et al.[130]	https://github.com/MrOlm/inStrain/
DEMIC	Gao and Li[131]	https://sourceforge.net/projects/demic/
EMBOSS Transeq v6.6.0.0	Madeira et al.[132]	https://www.ebi.ac.uk/jdispatcher/st/emboss_transeq
SAMtools v1.9	Li et al.[133]	https://www.htslib.org/
HTseq v2.0.2	Putri et al.[134]	https://htseq.readthedocs.io/en/latest/
geNomad v1.5.2	Camargo et al.[69]	https://portal.nersc.gov/genomad
VirSorter v1.0.6	Roux et al.[70]	https://github.com/simroux/VirSorter
DeepVirFinder v1.0	Ren et al.[71]	https://github.com/jessieren/DeepVirFinder
CheckV v1.0.1	Nayfach et al.[135]	https://bitbucket.org/berkeleylab/checkv/src/master/
Prodigal v2.11.0-gv	Camargo et al.[69]	https://portal.nersc.gov/genomad
VIBRANT v1.2.1	Kieft et al.[136]	https://github.com/AnantharamanLab/VIBRANT
MMSeqs2 v13.45111	Steinegger and Söding[137]	https://github.com/soedinglab/MMseqs2
mcl v14-137	Enright et al.[138]	https://micans.org/mcl/
CRISPRCasFinder v2.0.3	Couvin et al.[139]	https://github.com/dcouvin/CRISPRCasFinder
BLAST	Camacho et al.[140]	https://blast.ncbi.nlm.nih.gov/Blast.cgi
CoverM v0.6.1	Aroney et al.	https://github.com/wwood/CoverM
hmmsearch v3.3.2	Finn et al.[141]	https://www.ebi.ac.uk/Tools/hmmer/search/hmmsearch
tidyverse (R package)	Wickham et al.[142]	https://www.tidyverse.org/
patchwork (R package)	Thomas Lin Pedersen[143]	https://github.com/thomasp85/patchwork
ComplexHeatmap (R package)	Gu et al.[144]	https://github.com/jokergoo/ComplexHeatmap
viridis (R package)	Garnier et al.[145]	https://cran.r-project.org/web/packages/viridis/
vegan (R package)	Oksanen et al.[146]	https://github.com/vegandevs/vegan
rstatix (R package)	Alboukadel Kassambara[147]	https://rpkgs.datanovia.com/rstatix/
ggtree (R package)	Xu et al.[148]	https://www.bioconductor.org/packages/ggtree
ggtreeExtra (R package)	Xu et al.[149]	https://bioconductor.org/packages/release/bioc/html/ggtreeExtra.html
phytools (R package)	Liam J. Revell[150]	https://cran.r-project.org/web/packages/phytools/index.html
multcompView (R package)	Graves et al.[151]	https://github.com/lselzer/multcompview
ggVennDiagram (R package)	Gao et al.[152]	https://github.com/gaospecial/ggVennDiagram/
PMCMRplus (R package)	Thorsten Pohlert[153]	https://cran.r-project.org/web/packages/PMCMRplus/index.html
RColorBrewer (R package)	Erich Neuwirth[154]	https://cran.r-project.org/web/packages/RColorBrewer/index.html

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Plant materials and growth conditions

Healthy wild-type cultivars of wheat (*Triticum aestivum* L.), rice (*Oryza sativa* L.), maize (*Zea mays* L.), and *Medicago* (*Medicago truncatula*) were used for bacterial isolation and cultivation. Wheat cultivars, including BenDiHuangHuaMai, China Spring, Jing 411, RHT, and Xiaoyan 54, were cultivated at Changping Farm in Beijing, China. Rice cultivars, including Nipponbare and IR24, were grown on

Please cite this article in press as: Dai et al., Crop root bacterial and viral genomes reveal unexplored species and microbiome patterns, Cell (2025), https://doi.org/10.1016/j.cell.2025.02.013

the same farm as wheat but in a separate field. Maize cultivars, including Ji 53 and Zong 3, were cultivated in Dongying, Shandong, China. *Medicago* cultivars, including Jemalong A17 and R108, were grown in the greenhouse using agricultural soil collected from Changping Farm in Beijing, China (Tables S1A–S1C).

For field experiments, wheat cultivars, including Fangnong 16, Jimai 78, Jimai 22, Xiaoyan 54, and Zhoumai 18, were cultivated in fields at Changping, Beijing, and Yangling, Shanxi. Rice cultivars, including Nipponbare, IR24, and materials from the USDA rice mini-core collection,[155] were grown in fields at Hainan, Hefei and Beijing in China. Maize cultivars, including inbred lines Chang 7-2, Jing 724, Zheng 58, Zi 330, 5237, as well as hybrid varieties Jingke 968, HG968, Jingnongke 728, Ximeng 1919, Ximeng 3358, and Ximeng 6, were cultivated in fields at Shangzhuang, Beijing and Quzhou, Hebei. *Medicago* cultivars, including Jemalong A17 and R108, were grown in the greenhouse with agricultural soil collected from Shangzhuang, Changping Farm, Beijing, and Sipin, Jilin (Table S1D).

For both bacterial isolation and field sampling, fresh roots were collected from healthy crops harvested at the vegetative stage of growth, see following details.

METHOD DETAILS

Bacterial isolation, cultivation and genomic DNA extraction

Bacteria were isolated from the roots of wheat, rice, maize, and *Medicago* (Table S1A). Fresh roots were collected from healthy crops grown in agricultural soils. Three independent plants for each crop genotype were harvested at the vegetative stage. Roots were immersed in 30 mL of sterile 1× PBS to remove loosely adhering soil particles and vortexed at 180 rpm for 15 min; this process was repeated three times. The roots were cut into 2 mm sections and mixed to maximize bacterial species diversity. Visible nodules and nodule initials in *Medicago* roots were removed. A total of 0.02 g of mixed tissues was immersed in 200 μL of 10 mM $MgCl_2$ and ground into a homogeneous mixture. The homogenate was diluted to different concentrations using 10% TSB and R2A liquid media. The diluted homogenate was distributed into 96-well, sterile cell culture plates and cultivated for 16–20 days at room temperature.

Cultivated bacterial DNA was extracted to amplify the V5-V7 regions of the 16S rRNA genes via two-sided barcoded PCR.[5,105] The DNA extraction from bacterial cultures was carried out through a lysis procedure: 6 μL of bacterial culture was mixed with 10 μL of Buffer I, which is composed of 25 mM NaOH and 0.2 mM EDTA at a pH of 12, and the mixture was then incubated at 95°C for 30 minutes, then the pH was neutralized by adding 10 μL of Buffer II (containing 40 mM Tris-HCl at pH 7.5). To identify and track bacterial isolates in 96-well plates, a two-step PCR approach was employed, utilizing degenerate primers 799F and 1193R that were customized with unique barcodes for each well and plate to amplify the V5-V7 variable regions of the bacterial 16S rRNA genes. In the first PCR phase, 3 μL of the lysed bacterial DNA was amplified in a 30 μL reaction mixture that included 0.75 U of HS-Taq DNA polymerase, 10× buffer, 0.2 mM dNTPs (supplied by Takara), and 0.1 μM each of the forward (799F) and reverse (1193R) primers (obtained from Life Technologies). The PCR conditions for this stage consisted of an initial denaturation at 94°C for 2 minutes, followed by 25 cycles of 94°C for 30 seconds, 55°C for 30 seconds, and 72°C for 1 minute, with a final extension step at 72°C for 5 minutes. The PCR products from each well were then diluted 40-fold to prepare for the second PCR stage. In the second PCR stage, 3 μL of the diluted first-stage PCR product was amplified in a 30 μL reaction mixture that contained 0.75 U of HS-Taq DNA Polymerase, 10× buffer, 0.2 mM dNTPs, 0.1 μM of one of the 96 barcoded forward primers,[105] and 0.1 μM of the reverse barcoded primer. The PCR cycling conditions for this stage were: an initial denaturation at 94°C for 2 minutes, followed by 25 cycles of 94°C for 30 seconds, 55°C for 30 seconds, and 72°C for 1 minute, and a final extension at 72°C for 5 minutes. After amplification, the PCR products were purified using the Agencourt AMPure XP Kit from Beckman Coulter GmbH and the Wizard® SV Gel and PCR Clean-Up System from Promega. The DNA concentrations were measured using the PicoGreen dsDNA Assay Kit from Life Technologies, and the samples were pooled in equal volumes.

Purified PCR products were subjected to Illumina sequencing. After bacterial identification based on 16S rRNA marker genes using USEARCH,[99,106] bacteria were selected from target wells in the 96-well cell culture plates and further purified on 1/2 TSB or R2A agar plates. After three continuous purifications, single colonies were transferred to a liquid medium and cultured for 3–5 days. Liquid cultures were then validated by Sanger sequencing with 27F and 1492R primers. To select taxonomically diverse isolates, we chose bacterial strains with a threshold of $\leq$ 97% 16S rRNA gene identity within each crop species for whole-genome sequencing. At the same time and to capture intraspecies genetic diversity, we also included replicated bacteria that shared $\geq$ 97% 16S rRNA gene identity but were isolated from independent crop roots or different soils. Key criteria for replicated strains were that they must represent independent host colonization events. For bacterial genomic DNA extraction, 500 μL of bacterial solution was added to a Lysing Matrix E tube and homogenized twice at 7,200 rpm for 30 s with a Precellys Evolution homogenizer (Bertin Technologies). DNA was extracted using a FastDNA Spin kit (MP Biomedicals), according to the manufacturer's instructions. DNA concentrations were measured using a PicoGreen dsDNA Assay kit (Life Technologies). In total, 1 μg of DNA from each bacterial strain was used for whole-genome sequencing.

Whole-genome sequencing and *de novo* assembly of bacterial isolates

For each bacterial strain, genomic DNA was sequenced on an Illumina NovaSeq 6000 platform using 150-base pair (bp) paired-end reads, with 1 Gbp of raw data sequenced per genome to ensure ~20× sequencing depth. In total, 5.5 Tbp of raw reads were collected. Clean reads with adapters removed were filtered by quality analysis using Trimmomatic v0.39[107] with options 'SLIDINGWINDOW:4:20' and 'MINLEN:100' and assembled using SPAdes v3.14.0[108] with options '–isolate, –careful, –cov-cutoff auto' to form

contigs, and only contigs longer than 1,000 bp were retained for subsequent analysis. After assembly, genome purities were evaluated by using the lineage_wf workflow in CheckM v1.1.8.[109] Multi-isolates with contamination of > 5% were split into separate genomes using the binning module of MetaWRAP v1.3.2.[110] In total, we obtained 4,620 bacterial genomes with less than 5% contamination from pure isolates (wheat, 1,497; rice, 1,287; maize, 1,057; *Medicago*, 779). The majority (99.1%) of these genomes are of high quality, surpassing the 90.0% completeness threshold, and having contamination levels lower than 5.0% (Table S1A; STAR Methods).

Harvesting and metagenomic sequencing of root microbiome samples

In total, 332 root samples were *in situ* collected from wheat, rice, maize, and *Medicago* grown in 9 field soils. To ensure the representativeness and diversity of the crop root microbiome, at least two plant genotypes of each crop species were grown in three to four different soil backgrounds and were categorized into 14 datasets (Table S1D). All plants were harvested at the vegetative stage. For wheat, rice, and maize plants, individual root tissue was dug out from the field and gently shaken to remove loosely adhering soil. Root microbiome samples were then washed and collected as described previously.[5] *Medicago* plants were grown in a greenhouse in agricultural soils from geographically different locations. Visible nodules and nodule initials in *Medicago* roots were removed. Roots were then washed and collected the same way as described before. As a control, for each dataset, four to six bulk soil samples were collected from corresponding fields or pots without plants. In total, 75 bulk soil samples were collected. Both root and bulk soil samples were frozen in liquid nitrogen and stored at –80 °C. Root and soil samples were smashed for DNA extraction using a FastDNA SPIN kit (MP Biomedicals). DNA was fragmented by sonication following Illumina's instructions to generate PCR-free libraries for shotgun sequencing.

Quality control, assemble and gene prediction of metagenomic samples

In total, 10.3 Tbp of raw reads were generated across 407 root and soil metagenomic samples. After removing adapter, barcode sequences, and low-quality reads, host crop reads were depleted using the KneadData v0.7.6 profile (https://github.com/biobakery/kneaddata) with the options 'SLIDINGWINDOW:4:20, MINLEN:50, and –very-sensitive'. To further remove host crop reads, reads were taxonomically classified by Kraken2 v2.1.1[112] with the default database, and reads classified as host crop species were filtered using the Kraken tool extract_kraken_reads.py.[111] For root samples, the percentage of clean microbial reads varied across datasets, with an average of 17.9% (Table S1D). To achieve relatively balanced coverage of microbial diversity, cleaned reads were normalized to no more than 6 Gbp or 8 Gbp for root and soil samples using SeqKit tool,[113] respectively. For each dataset, clean reads from roots or soil were *de novo* co-assembled into contigs using MEGAHIT v1.2.9[114] under meta large mode with a minimum contig length threshold of 500 bp. Prokaryotic coding sequences were predicted from the assembled contigs using Prodigal v2.6.3[115] with the option '-p meta'. To remove redundant gene sequences, predicted sequences were clustered into non-redundant (NR) genes at 95% identity and 90% coverage of the shorter gene using CD-HIT v.4.8.1[116] with options '-c 0.95, -as 0.9, -G 0, -g 1, -d 0, and -l 150'. Considering their limited genetic information, NR genes shorter than 150 bp were removed. The resulting NR genes for both root and soil subsets were then further merged by clustering to generate an integrated NR gene reference for later use as the reference for metagenomic read quantification.

Binning and refinement of MAGs

MAGs were reconstructed from microbial contigs of 332 root metagenomic samples using MetaWRAP v1.3.2.[110] *De novo* metagenomic assembly was also performed on individual samples according to the previously mentioned methods to identify unique microbes. Assembled contigs with lengths of over 1,000 bp were binned into individual draft genomes based on sequence composition and coverage information using MetaBAT2[117] and MaxBin2[118] as implemented in the binning module of MetaWRAP with default parameters. To improve the quality of the assembled genomes, the bin_refinement module was used on the assembled and binned genomes with a contamination cut-off below 10% and a completeness cut-off over 50%.[156]

Collecting genomes of crop root bacteria from public databases

To uncover novel bacterial species within the CRBC, we systematically collected crop root bacterial genomes across different public databases, including NCBI,[157] IMG/M,[43] and ENA[44] databases. A total of 5,441 plant root-related bacterial genomes were obtained from public databases using metadata from April 2023. For details, in NCBI, we scanned the NCBI Datasets genome table (https://www.ncbi.nlm.nih.gov/genome/browse#!/overview/) and BioProject and BioSample information from NCBI plant-associated environmental packages. Keywords such as root, rhizosphere, rhizoplane, endophyte, endosphere, and nodule were used to filter the root-associated genomes. As a result, we obtained 3,168 root-related bacterial genomes from NCBI. For the IMG/M system (https://img.jgi.doe.gov/), 2,273 root-related bacterial genomes were downloaded based on the criterion of ecosystem type belonging to "Root", which also includes genomes from rhizosphere and nodules. Additionally, 294 bacterial genomes from lotus roots under accession number PRJEB37696 were downloaded from the ENA database. We then screened the crop hosts of these root-related bacterial genomes by both their English names and scientific names, e.g. rice (*Oryza sativa*), sorghum (*Sorghum bicolor*), maize (*Zea mays*), and barley (*Hordeum vulgare*), according to the World Crops Database (https://world-crops.com). In total 3,073 quality-checked crop root-related bacterial genomes were obtained from public databases (Table S1G).

Please cite this article in press as: Dai et al., Crop root bacterial and viral genomes reveal unexplored species and microbiome patterns, Cell (2025), https://doi.org/10.1016/j.cell.2025.02.013

CellPress
OPEN ACCESS

Quality control, dereplication, and species clustering of bacterial genomes

The quality of all 10,973 crop root bacterial genomes was assessed by CheckM v1.1.8[109] using the lineage_wf workflow with default parameters. In total, 9,772 genomes passed the quality criterion of a score exceeding 50 (QS = completeness – 5 × contamination)[158] and were retained for further analysis. These genomes consisted of 4,618 isolates and 2,081 MAGs from the CRBC and 3,073 bacterial genomes from public databases (Tables S1A and S1G). Prediction of tRNA and rRNA in bacterial genomes was conducted using tRNAscan-SE v 2.0.12[119] and barrnap v0.9 (https://github.com/tseemann/barrnap) with default parameters, respectively. High-quality bacterial genomes met the following specific criteria: completeness > 90%, contamination < 5%, and the presence of *23S*, *16S*, and *5S rRNA* genes along with at least *18 tRNAs*.[156] Yet, the challenge of assembling *rRNA* genes from short-read sequencing data has led to difficulty in obtaining the *5S rRNA* gene. For subsequent functional analysis, we defined genomes with completeness > 90% and contamination < 5% as high quality.

We dereplicated the 9,772 genomes using an ANI ≥ 99.9% and coverage ≥ 30% using dRep v3.2.2.[120] The options used for dereplication were '-strW 0, -centW 0, -nc 0.3, and -sa 0.999'. Subsequently, clustering at the species level was performed with an ANI of ≥ 95% and coverage of ≥ 30%. For species-level clustering, ANI was set to -sa 0.95. To designate representatives for NR genomes and species, the genome with the highest quality was selected, prioritizing cultured genomes over MAGs. The scoring criteria were based on the following formula: completeness – 5 × contamination + 0.5 × log (N_{50}). In total, 2,212 species lacking published crop root genomes were identified as undefined bacterial species from crop roots within the CRBC (Table S2A).

Comparison of bacterial species within the CRBC and representative species in the GTDB database

To explore the novelty of CRBC genomes compared with bacterial genomes across different habitats, we compared all the genomes within the CRBC to 62,291 reference prokaryotic species in GTDB Release 207 (8 April 2022).[41] We first converted the GTDB reference species and CRBC into sketch files by Mash v2.3[121] with default parameters. We then used mash dist to estimate the pairwise Mash distances between CRBC genomes and GTDB representative species. For each CRBC genome, we aligned the genome to the best-matched reference GTDB genome with the lowest Mash distance using DNAdiff v1.3 from MUMmer v4.0.0[122] to calculate the genome coverage (AlignedBases) and ANI (AvgIdentity). We considered a CRBC genome to match the reference GTDB if the coverage was ≥ 30% and ANI was ≥ 95%. CRBC species lacking a match to the GTDB were defined as undefined species across various habitats (Table S2A).

To infer the higher taxonomic ranks of the undefined CRBC genomes, we employed a strategy based on relative evolutionary divergence (RED) to delineate taxa above the species level, as recommended by GTDB-tk.[123] First, we used the de_novo_wf workflow to construct a *de novo* phylogenetic tree for the CRBC and GTDB reference genomes. The tree was then annotated, and the taxonomic ranks were inferred using GTDB-tk. Potential unknown taxa were labeled according to the rank inference results. The analysis revealed that these species encompass the following potential unique taxa: 1 phylum, 1 class, 3 orders, 15 families, and 101 genera (Table S2A).

Taxonomic classification and phylogenetic analysis of root bacterial genomes

Taxonomic classification of all bacteria was conducted using GTDB-Tk v2.0.0,[123] the classify_wf workflow, and the unsplit bacterial tree with the option '–full_tree'. The GTDB-Tk infer method was used to construct a maximum-likelihood phylogenetic tree *de novo* using representatives of root bacterial species or genera. For tree construction, we used a concatenated multiple sequence alignment of 120 bacterial phylogenetically informative bacterial markers, analyzed under the WAG model for amino-acid substitution. Phylogenetic diversity of the species tree was calculated by the sum of total branch lengths,[159] and PD of each phylum was calculated by the sum of branch lengths of the representative genomes belonging to the phylum (Table S2B). The phylogenetic gain (PG) of the CRBC compared with that of publicly available root genomes was calculated by the formula[158] $PG_{CRBC} = PD_{total} - PD_{Pub}$. Phylogenetic analysis revealed the limited species diversity present in public genomes of crop root bacteria, contributing merely 25.6% to the overall phylogenetic diversity (PD) of all crop root bacteria.

Reads coverage of root and rhizosphere metagenomic samples using different genome databases

Read coverage analysis of metagenomic data was conducted using Salmon v1.9.0[124] with options '–libType A, –meta, and –validateMappings', and mapping rates were used to evaluate read utilization efficiency of diverse genome databases. We focused on root-related metagenomic samples including 42 root metagenome samples not used for CRBC construction (wheat, $n = 12$; rice, $n = 12$; maize, $n = 6$; *Medicago*, $n = 12$) and 37 rhizosphere metagenome samples sourced from previous studies[98,160] (wheat, $n = 29$; maize grown in Europe [PRJEB77048], $n = 8$, two independent sequencing from four rhizosphere samples; Table S2E). To maximize comparability, host reads were removed and the metagenomic data were normalized to 1 Gbp per sample. Bacterial genome reference databases included NCBI RefSeq (September, 2022), GTDB Release 207,[41] GEM,[45] UHGG[46] and OMD.[31] Coverages of crop bacterial genomes from the CRBC and published databases was evaluated individually and then integrated for analysis (Table S2E).

Functional annotation and BGC and antiviral system identification of bacterial genomes

Coding sequences for all crop root bacterial genomes were predicted using Prodigal v2.6.3[115] with options '-p single, -f gff, and -a, -d'. These predicted protein sequences were compared with prokaryotic orthologs of the KEGG database[100] (v.102 release April 2021) using the DIAMOND BLASTP v2.0.15.153[125] with options '–outfmt 6, –max-target-seqs 1, –evalue 1e-5, –sensitive, –block-size 6,

and –index-chunks 1', and the identity threshold was set to 50% (Table S4L). Secondary metabolite BGCs were identified using antiSMASH[126] v6.1.0 with the option '–cb-knownclusters'. Because fragmented BGCs are likely to introduce bias for genome comparison, BGCs were subsequently screened to retain those predicted from contigs longer than 5 kb,[31] resulting in a total of 48,643 BGCs, with 67% predicted to be complete by antiSMASH (Table S3D). These BGCs were divided into BGC classes according to BiG-SCAPE v1.1.5.[127] For individual genomes, BGC composition was calculated as the percentage of BGCs that belong to each class. The prediction of bacterial antiviral defense systems within bacterial genomes was performed using DefenseFinder v1.2.2[104] with default parameters (Table S6B). Defense system families in each genome were calculated by summing the number of unique subsystems.

PGP functions in bacterial genomes

To elucidate the PGP functions encoded within the CRBC and published crop root bacterial genomes, bacterial PGP biological processes and genes/orthologs involved were compiled from previous studies. These included genes involved in nutrient utilization, such as phosphorus nutrition,[48–50] nitrogen fixation,[47] siderophore biosynthesis,[51] biosynthesis of plant growth hormones including IAA[52,53,56], GA,[52,53,56] and CK,[54,55] and resistance to biotic and abiotic stresses by biosynthesis of ACC deaminase,[57] salicylic acid,[58,59] and ethylene[51] (Tables S3A and S3B). For PGP functional annotation, we collected and curated PGP-related proteins within the KEGG database, encompassing those encoded across eukaryotes, prokaryotes, and viruses. Additionally, proteins involved in bacterial GA biosynthesis, such as experimentally-validated copalyl diphosphate synthase and kaurene synthase[52,53] that were not incorporated by KEGG, were manually downloaded and added into this PGP-related protein reference. To ensure robustness and accuracy in profiling PGP features, we used high-quality, non-redundant genomes within crop root bacterial genomes (n = 6,109). Based on BLAST results, as mentioned earlier, PGP genes encoded in each genome were summarized at the KO level as 0/1 binary data. A PGP feature was considered present when all the orthologs involved in a specific PGP process were encoded within a genome (Table S3C). The conservation of specific PGP functions at the genus level was assessed by determining the percentage of NR genomes encoding that particular function.

Comparison of BGCs in crop root bacterial genomes and public databases

To quantify BGC novelty and richness across phylogenies, BGCs were further clustered into 12,865 GCFs by computing all-against-all cosine distance using BiG-SLICE v1[128] and a 0.2 distance threshold. We compared these BGCs with the public BGC databases, including the computationally predicted BGCs in the BiG-FAM database[63] and experimentally validated BGCs from the MIBiG v3.1 database.[64] For each of these BGCs, the minimum cosine distance (d_{min}) to public BGCs was defined as its distance against the public database. The average distance of a BGC within a GCF was considered its final cosine distance ($\bar{d}_{min}$). A GCF was defined as undefined when its $\bar{d}_{min}$ against the public database was greater than 0.2. To illustrate the extent of GCF in terms of BGC classes and taxonomy, the $\bar{d}_{min}$ calculated between crop root GCFs and each of these public databases was displayed along the gradient distance from 0 to 1.00 with an interval of 0.05 (Table S3E). GCFs containing BGCs that belong to several different BGC classes were annotated as ''Multiple classes''. At the genus level, BGC composition and GCF richness were summarized as median values across all high-quality NR genomes under a specific genus. When a genus included BGCs from genomes of the CRBC, the genus was annotated as CRBC present.

Taxonomic classification and functional annotation of crop root metagenomic samples

To optimize microbial gene identification for metagenomic analysis, NR genes were taxonomically classified with Kraken2 v2.1.1[112] using the default database incorporated with the CRBC, published crop root bacteria, and publicly available reference databases including GTDB,[41] IMG/VR v4,[36] FungiDB,[97] and Ensemble protists in addition to the default Kraken built-in database. Briefly, the reference genome fasta files were downloaded and formatted to fit with Kraken2 format requirements and were integrated into Kraken2 default libraries (Tables S3F–S3I). For functional annotation, the NR gene set was first translated into protein sequences using EMBOSS Transeq v6.6.0.0[132] and annotated by DIAMOND BLASTP against the KEGG protein databases.

Prediction and quality control of viral genomes

We predicted viral sequences from 9,772 bacterial genomes and assembled contigs of 332 crop root metagenomes. To achieve this, we used geNomad v1.5.2,[69] which uses a hybrid framework that integrates both alignment-free and gene-based models to identify viral contigs. Additionally, to ensure comprehensive predictions, we used two different algorithms for viral sequence identification, i.e., VirSorter v1.0.6[70] and DeepVirFinder v1.0,[71] which rely on reference genome datasets and deep learning for viral sequence prediction, respectively. For geNomad (v1.3), we used end-to-end execution with the conservative preset and the default database. For VirSorter, Viromedb (–db 2) was used as a reference to predict prophage and viral fragments from genomes, and viral contigs classified as category 1, 2, 4, or 5 were retained. Contigs were scored by DeepVirFinder with default parameters, and those with a P value < 0.05 were considered potential viral sequences. Predicted viral contigs were further filtered using a length threshold of 1 kb.[35] Genome quality, gene classification and identification, and trimming of host regions were conducted with CheckV v1.0.1[135] using the end-to-end workflow. Finally, 16,546 viral genomes with host gene content below 30% and genome quality classified as medium or above (medium-quality, high-quality, and complete) were retained. To prevent redundant sequences predicted by the

different algorithms, we deduplicated the viral sequences based on 100% sequence similarity and 100% coverage of short sequences. This process resulted in a final set of predicted 9,736 unique crop root viral genomes (Table S5A).

Gene prediction, functional annotation, and protein clustering of viral genomes

For a comprehensive comparison, we obtained viral genomes from public databases of representative ecosystems, including NCBI RefSeq, the Metagenomic Gut Virus catalogue (MGV),[35] Global Ocean Viromes 2.0 (GOV2.0),[37] and IMG/VR v4.[36] Within IMG/VR, viruses originating from roots and soil were considered representative of plant root and soil environments. To ensure comparability, only genomes with CheckV quality classified as medium or above were retained, with the exception of those from NCBI RefSeq with confirmed completeness. Viral proteins from the NCBI virus were directly downloaded, whereas for other databases we used Prodigal v2.11.0-gv[69] to predict coding sequences with default parameters, followed by translation into proteins. In total, 587,992 genes were predicted from the CRVC (Table S5C). Viral life cycle prediction and functional annotation were conducted using VIBRANT v1.2.1[136] with option '-f prot -virome' using reference databases including KEGG, Pfam,[103] and VOGDB[74] (http://vogdb.org/). Protein clustering was conducted using MMSeqs2 v13.45111[137] with thresholds of 50% average amino acid identity (AAI) and 80% coverage against UniRef50[161,162] and options '–cov-mode 1, -c 0.8, –cluster-mode 2, –min-seq-id 0.5, and –kmer-per-seq 80'. Finally, the CRVC dataset contained 200,069 protein clusters, of which only 33,323 were clustered with previously published viral sequences.

Clustering of viral genomes at the species and genus level

To assess the diversity and novelty of our viral genomes, all viruses were clustered at the species and genus levels (Tables S5B and S5D). According to MIUViG standards[72] a threshold of 95% ANI and coverage of 85% of the shorter sequence were used for species-level (vOTU) clustering. Similarity and coverage between genomes were estimated using scripts from the CheckV database.[135] All-vs-all local comparisons between viral sequences were performed using the blastn package from BLAST v2.5.0[140] with options '-max_target_seqs 30000 and -perc_identity 90'. The script cluster.py from the MGV database was used to perform species-level clustering using a greedy, centroid-based algorithm. A total of 7,653 species-level clusters were identified in the CRVC.

At the genus level, clustering was performed based on protein similarity and the protein share network between genomes.[163,164] Viruses from the CRVC and public databases were represented by representative genomes at the species level. Initially, all-vs-all protein alignments were conducted using DIAMOND BLASTP v2.0.15.153[125] with options '–evalue 1e-5, –max-target-seqs 10,000, –query-cover 50, and –subject-cover 50'. Subsequently, we calculated the number of shared proteins between genomes, the proportion of proteins shared, and the average AAI of shared proteins between each genome pair. The score between genome pairs was calculated based on $\min_{cov} \times \text{mean}_{aai}$, and the score was used as the network edge between the genome pairs. Finally, genome pairs were filtered based on the set threshold, and mcl v14-137[138] was used to perform clustering with the option '–abc -I'. We benchmarked combinations of different filtering thresholds, including protein sharing ratios of 10%, 15%, 20%, 25%, and 30%, and note that when the viral genome was too large, at least 20 proteins needed to be shared between them. We also tested average AAI values of 20%, 30%, 40%, 50%, 60%, and 70% as well as MCL inflation factors of 1.1, 1.2, 1.4, 2.0, 4.0, and 6.0. We evaluated their combinations by comparing them against the taxonomy and clustering results derived from the RefSeq viruses, which served as our mock dataset, using Adjusted Mutual Information scores. Finally, we selected a protein network sharing threshold at average AAI of 30% and an inflation factor of 4.0 for genus-level clustering, with precision and recall thresholds set to 0.85 and 0.88, respectively. Viral species- or genus-level clusters that could not be clustered with other viruses were defined as unreported.

Taxonomic classification of the CRVC

We performed taxonomic classification on the CRVC by integrating protein annotation and genus-level clustering results. First, following the taxonomy of the ICTV Master Species List MSL38 v1 (number 22),[101] we curated corresponding RefSeq representative virus sequences to serve as classification standards. For taxonomic classification, if the genome in the CRVC could be clustered at the genus level with known reference viruses and the taxonomic consistency of reference viruses within the cluster exceeded 50%, we assigned the taxonomic label to that cluster. For viruses that did not cluster with any reference viral genomes, we used MMSeqs2[137] based on the lowest common ancestor algorithm to classify the taxonomy of the genomes. The custom reference viral databases used by MMSeq2 were built on the proteins of RefSeq viral genomes. This construction yielded a classification rate of 94.9% at the family level with an accuracy of 99.9% by comparing the results against the mock taxonomy derived from RefSeq data.

Host prediction of the CRVC

For predicting the domain-level hosts of the CRVC, we defined the viral host domain based on taxonomic statistics from the Virus-Host DB.[102] We established connections between crop root bacterial genomes and phages based on hits of predicted phage sequences and CRISPR spacers identified within bacterial genomes. The predicted phage sequences included explicit prophages and phage fragments in 9,772 bacterial genomes. Prophages were defined by VirSorter categories 4 and 5 and geNomad's provirus labels, and we excluded contigs in MAGs if more than 50% genes were predicted viral according to CheckV. For phage fragments without prophage structure, we only included those predicted from isolates and excluded MAGs to eliminate false positives caused by binning errors. Spacer sequences within bacteria were predicted using CRISPRCasFinder v2.0.3[139] with default parameters, retaining spacer sequences in CRISPR arrays with evidence level 4 (high confidence). In total, we obtained 18,417 phage sequences

Please cite this article in press as: Dai et al., Crop root bacterial and viral genomes reveal unexplored species and microbiome patterns, Cell (2025), https://doi.org/10.1016/j.cell.2025.02.013

from 5,624 bacterial genomes and 27,952 spacer sequences from 644 bacterial genomes. To establish phage–bacterial connections at the within the bacterial and viral genomes, we used BLAST to match these sequences to viral genomes of the CRVC and IMG/VR sequences originating from root or soil environments. The alignment options for phage sequences were '-max_target_seqs 30000 and -perc_identity 95', and for spacer sequences, the alignment options were 'blastn-short, -dust no, -word_size 7, -perc_identity 100, -max_hsps 1, and -max_target_seqs 100000'. For phage sequences, a BLAST length of $\geq$ 1 kb and an ANI of $\geq$ 96% were considered positive BLAST hits. For CRISPR spacers, only spacers matched across the whole length of spacers were considered as positive BLAST hits (Table S6A). The definition of temperate phages was based on the classification from VIBRANT and the identification of prophages.

QUANTIFICATION AND STATISTICAL ANALYSIS

Quantification of bacterial species and evaluation of bacterial growth rate

To ensure the accuracy of read mapping, 3,044 species-level representative genomes were used as references. Reads of 332 root metagenomic samples were aligned to this genome reference using Bowtie 2 v2.4.5[129] under 'end-to-end and –very sensitive' mode. Subsequently, InStrain v1.8.0[130] was used to calculate the coverage and breadth of each sample against the bacterial genomes. A breadth value of > 0.5 was considered to indicate the presence of a genome in the sample (Table S6D). The top ten most abundant bacterial species in each crop root metagenomic sample were considered to be highly abundant. The peak-to-trough ratio[165] of the copy numbers of DNA near the replication origins and copy numbers of bacteria terminus points in a sample was calculated to measure genome growth dynamics. Evaluation of bacterial growth rate was conducted using Dynamic Estimator of Microbial Communities (DEMIC)[131] with default parameters (Table S2D).

Quantification of metagenomic reads and feature table generation

Host-depleted, high-quality reads of metagenomic root samples were mapped against corresponding NR gene references using Bowtie 2. A positive match was defined by an alignment identity of $\geq$ 95% and an alignment length $\geq$ 60% of read length. All positive matches from the Bowtie 2 SAM output were next extracted and the resulting filtered SAM files were sorted and converted into BAM format using SAMtools v1.9.[133] A non-redundant gene count table was generated using HTseq v2.0.2[134] with options '–mode=union, –stranded=no, –type=CDS, –idattr=gene_id, –nonunique=none, –secondary-alignment=ignore, and –supplementary-alignment =ignore'. This gene count table was then correlated to the microbiome according to genes assignment to microbes by Kraken 2, and further normalized by the gene length and count.[166] The relative abundance table was scaled to 10^9 per sample (Tables S4A–S4F). Taxonomic and functional feature count tables were generated based on non-redundant gene annotation. The function feature tables were further filtered to exclude unannotated features and were normalized within each sample. Considering the presence of some features (e.g., KOs) with extremely low abundance, accurate repeats between technical replicates were not feasible. To reduce the interference of these features on differential abundance analysis, we defined the measurable KOs by evaluating the repetition of deeply sequenced root metagenomic samples. We sequenced five root samples three times each and compared the consistency of the relative abundance tables by calculating the Pearson correlation coefficients between each pair of technical repeats at different abundance thresholds. The correlation coefficient curve indicated that the repeatability of KO abundance was satisfactory when the threshold of relative abundance reached 3.0×10^{-5} (normalized to 1). In terms of the functional pathway table, the technical repeatability was high (r = 0.98), so no abundance cutoff was set. To ensure accuracy, functions that were exclusively from animals or plants were excluded (Table S4E).

Rarefaction analysis and evaluation of microbial genetic potential

To evaluate the representativeness of non-redundant genes for crop root metagenomic samples, we calculated the gene richness as the number of included root samples increased. For each specific sample number gradient, a random shuffle of the root samples was repeated 20 times (Figure S3A; Tables S3F). To evaluate the microbial genetic diversity for each crop species, we filtered non-redundant microbial genes in at least 70% of the root samples within each dataset, integrated the gene counts across each crop species, and defined them as representative non-redundant microbial genes (Figure S3B; Table S3G). The ratio of representative NR microbial gene entries to the gene number of the crop genome was used to measure microbial genetic diversity for each crop.

Quantification of viruses and antiviral defense systems in metagenomic samples

The representative genomes at the species level in the CRVC and IMG/VR were selected as references. Host-removed metagenomic sequences were aligned to the reference genomes using Bowtie 2 in end-to-end mode with options '–non-deterministic and –very-sensitive'. SAMtools was used to sort and convert the files to BAM format. The filter mode of CoverM v0.6.1 (https://github.com/wwood/CoverM) was used to filter reads with an alignment identity threshold of 95%. Metagenomic aligned bases and coverage of viral genomes were calculated using CoverM with options '-m covered_bases and -m covered_fraction', respectively. Viruses with coverage of $\geq$ 70% or aligned bases of $\geq$ 5 kb were considered present in the metagenomic sample. Virus abundance was computed using CoverM with option '-m trimmed_mean', which calculated the average read depth after removing the 5% of bases with the highest and lowest genome coverage (tpmeans). The obtained tpmean value was divided by the number of reads in the sample, yielding the relative virus abundance. The relative abundance of phages in each sample was calculated as the sum of the relative

Please cite this article in press as: Dai et al., Crop root bacterial and viral genomes reveal unexplored species and microbiome patterns, Cell (2025), https://doi.org/10.1016/j.cell.2025.02.013

abundances of all bacteriophages in the sample. For quantification of defense systems in crop root and soil samples, the protein profiles from DefenseFinder[104] were used to annotate translated NR genes using hmmsearch v3.3.2[141] with options '–cut_ga and noali'. Based on the annotations and previously generated relative abundance table of NR genes, a feature table of defense-related proteins was generated. The relative abundance of each antiviral defense system was calculated as the sum of the abundances of mandatory proteins[104] within that system.

Microdiversity calculation of viral genomes

Metagenomic data were used to analyze the intrapopulation genetic diversity (microdiversity) of viral genomes within sampled environments. We used InStrain v1.8.0[130] to calculate the average nucleotide diversity (π) of each viral genome in each sample.[37] For accuracy, only genomes with coverage of at least 5× across all base pairs were used for calculation. Microdiversity within each prevalence group described in the next section was defined as the mean π value of all viruses belonging to that prevalence group in the sample.

Definition of prevalence groups for bacterial and viral genomes

Genomes were grouped based on the prevalence characteristics of bacterial and viral genomes in crop root metagenomic samples. Prevalence refers to the frequency at which a genome was detected in samples from a single host crop at a single location. Genomes with a prevalence of less than 10% or those found in only a single sample if the sample number was below 10 were defined as rare. The remaining genomes were considered stably detected in root ecosystems. Further categorization included the following three groups: stable presence across roots of multiple crop species at multiple locations (multi-crop multizonal), stable presence in roots of a single crop species across multiple locations (single-crop multizonal), and stable presence in roots of a single crop species at a single location (single-crop regional). For bacteria, there were 60 species in the multi-crop multizonal group, 99 in the single-crop multizonal group, 455 in the single-crop regional group, and 331 in the rare group. For viruses, there were 123, 281, 1,311, and 975 species in each group, respectively.

Statistic analysis and data visualization

Comparisons between two groups were conducted using a two-tailed Wilcoxon rank-sum tests. For comparisons among multiple groups, a Kruskal–Wallis rank sum test was used, followed by a Dunn's test for pairwise comparisons. Multiple comparisons were adjusted using the false discovery rate (FDR) method, with significance considered when the adjusted *P* value was less than 0.05. Asterisks were used to denote the level of statistical significance (*, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$; ****, $P < 0.0001$). Multiple groups were labeled with letters based on adjusted *P* values. A correlation analysis was conducted using the cor.test() function in R with Spearman's rho, which calculated correlation coefficients and *P* values. Similar to the previous procedure, multiple comparisons were adjusted using the FDR method, with significance defined as an adjusted *P* value of less than 0.05. Downstream data processing and visualization were performed using R 4.3.1 and packages including tidyverse, reshape2, rstatix, vegan, PMCMRplus, and multicompView for data processing and statistical analysis and ggforce, ggpubr, ggVennDiagram, UpSetR, ComplexHeatmap, viridis, RcolorBrewer, and patchwork for visualization. Visualization of phylogenetic trees was conducted using ape, picant, pytools, ggtree, ggtreeExtra, and ggnewscale. Note that tree-branch lengths in Figures 5A and 6C do not reflect taxonomic distance.

CellPress
OPEN ACCESS

Supplemental figures

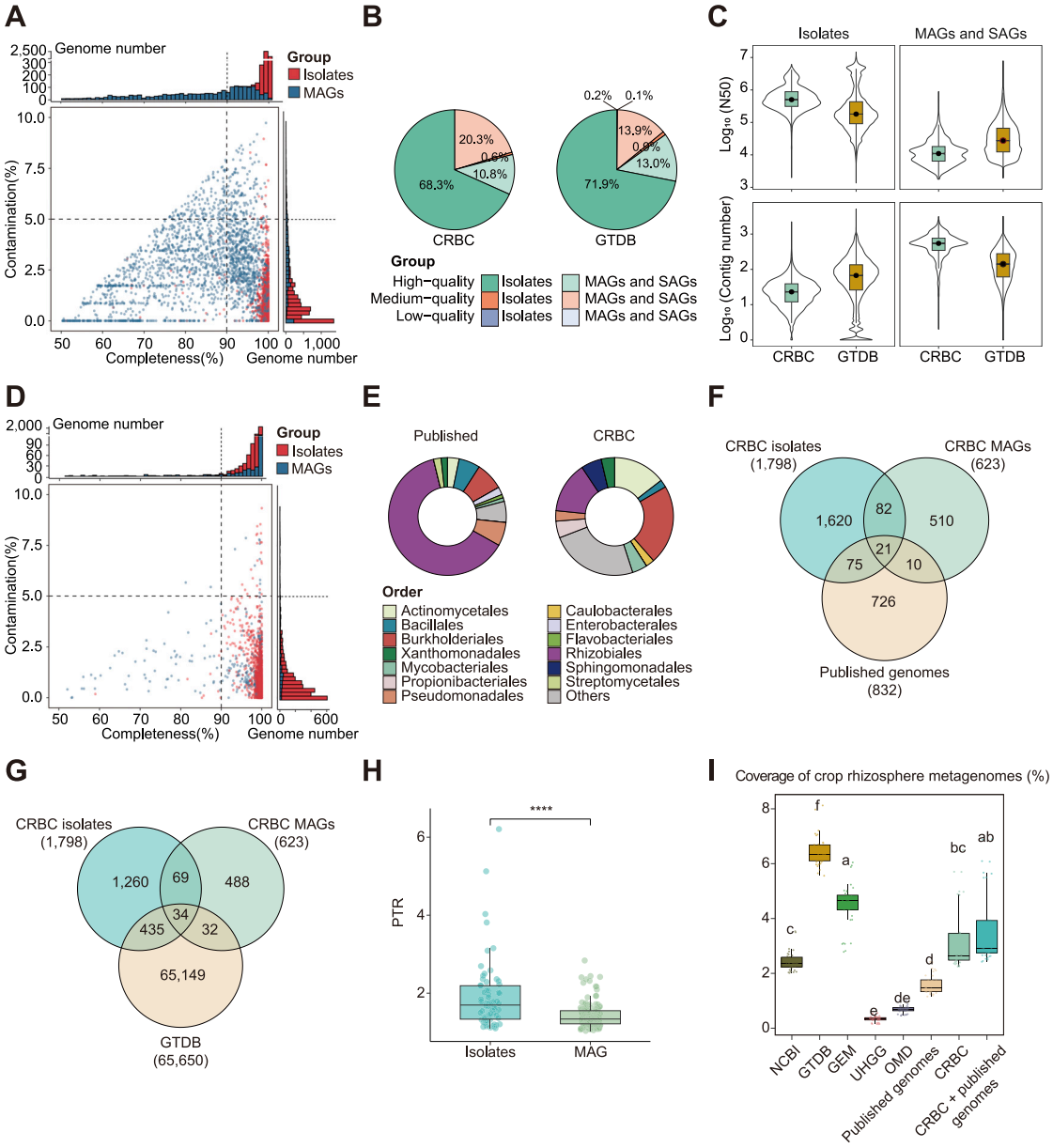


Figure S1. Characteristics of the CRBC and publicly available genomes, related to Figures 1 and 2

(A) Genome quality of the CRBC. The 6,699 genomes in the CRBC are displayed according to the genome completeness and contamination rates. Genome numbers are color-coded based on their sources, with isolates represented in red and MAGs in dark blue.

(B) Distribution of genome quality in the CRBC and GTDB. The pie chart shows the composition of genomes at different quality levels.

(C) Genomes in the CRBC show comparable quality to those in the GTDB. The boxplots show $\log_{10}$-transformed N50 length (top) and contig number (bottom) in each genome of the CRBC and GTDB. Isolates are shown in the left and MAGs are shown in the right, respectively.

(D) Quality of published crop root bacterial genomes. The 3,073 published genomes of crop root bacteria are displayed according to the genome completeness and contamination rates.

(E) The taxonomic composition of published crop root bacterial genomes (left) and the CRBC (right) at the order level. The bacterial taxonomy is displayed at the order level.

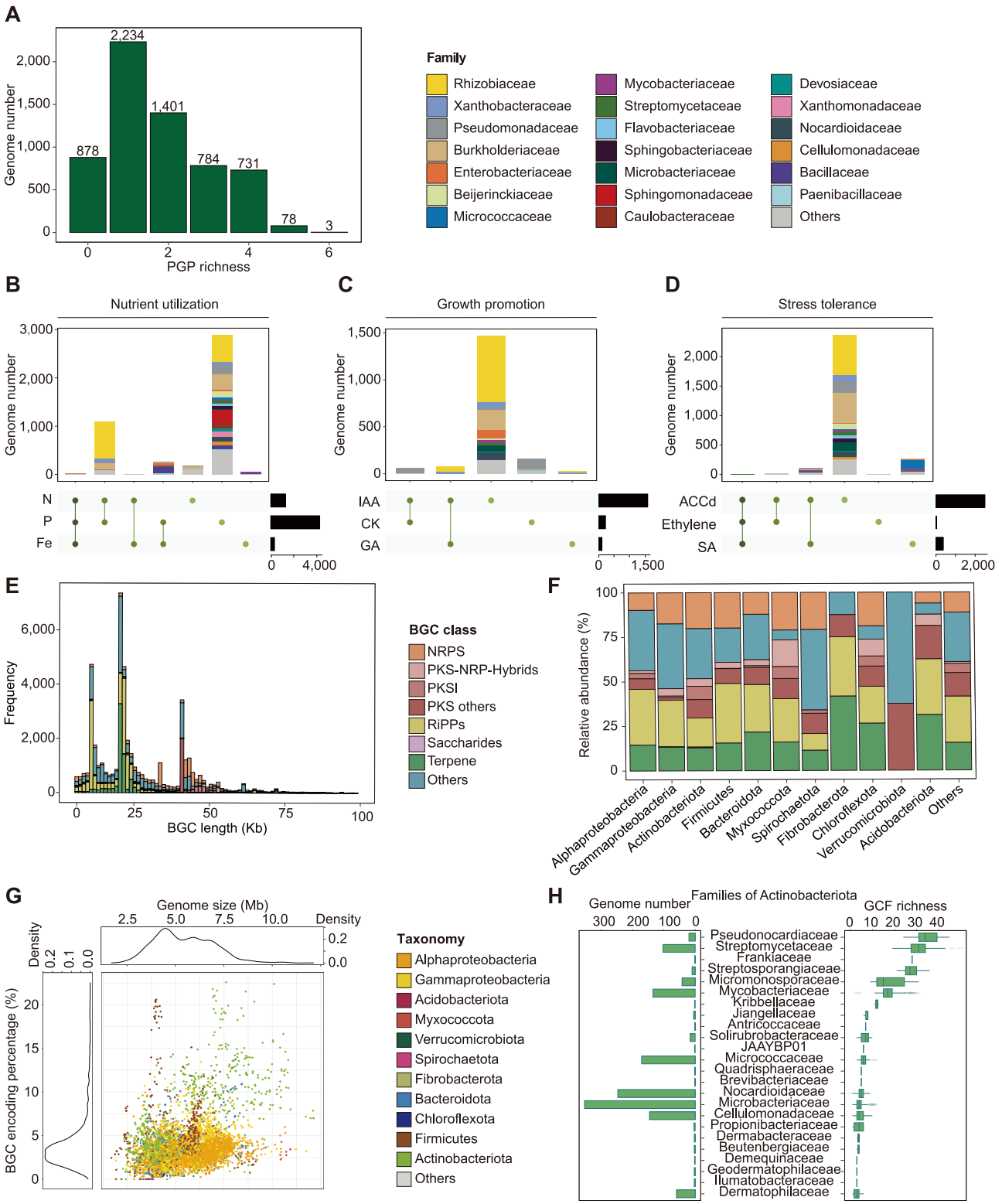
(legend continued on next page)

(F) Overlap of bacterial species in the CRBC and published crop root bacteria. The species are color-coded based on their genomic sources, with CRBC isolates represented in blue, CRBC MAGs in green and published crop root genomes in beige. Note that 76.9% ([1,620+82]/2,212) of genomes not reported before in crop roots are from the CRBC isolates.

(G) Overlap of CRBC genomes with GTDB. The comparison between 6,699 CRBC genomes and over 60,000 GTDB representative genomes shows a total of 1,817 undefined CRBC species, with 1,329 identified from CRBC isolates and 448 from CRBC MAGs.

(H) The CRBC MAGs showed lower growth rates than the CRBC isolates within and between taxa. The bacterial isolates were analyzed by DEMIC based on the ratio of the coverage at the origin and terminus (peak-to-trough ratio [PTR]) of bacterial genomes within crop root microbial populations (STAR Methods, adjusted $p < 0.05$, Wilcoxon rank-sum test).

(I) The reference genomes showed source specificity in rhizosphere metagenomic data. The boxplot shows proportions of metagenomic reads of crop rhizosphere samples in maize ($n = 8$) and wheat ($n = 29$) aligned to genomes in public databases and the CRBC. The abbreviations are: NCBI, NCBI RefSeq database; GTDB, Genome Taxonomy Database; GEM, the Genome from Earth's Microbiomes catalog; UHGG, the Unified Human Gastrointestinal Genome collection; OMD, the Ocean Microbiomics Database (adjusted $p < 0.05$, Kruskal-Wallis rank-sum test and Dunn's test).



(legend on next page)

Figure S2. Characteristics of PGP functions and BGCs in the CRBC and published crop root bacterial genomes, related to Figure 3

(A) Distribution of genomes across their PGP functions. Bar plots show the number of PGPs in each genome. Note that among the 6,109 high-quality genomes of crop root bacteria, 5,231 genomes were found to encode at least one tested PGP function.

(B) Coexistence patterns of nutrient utilization functions within individual bacterial genomes. The number and taxonomy of genomes for each coexistence pattern are shown in the stacked bar plot in the upper panel. The lower panel with vertical lines illustrates co-existent functions (nitrogen fixation, phosphorus and siderophore biosynthesis) within individual genomes. The number of genomes for each PGP group is shown in the lower right panel.

(C) Coexistence patterns of crop growth functions within individual bacterial genomes. The number and taxonomy of genomes for each coexistence pattern are shown in the stacked bar plot in the upper panel. The lower panel with vertical lines illustrates co-existent functions (biosynthesis of IAA, CK, and GA) within individual genomes. The number of genomes for each PGP group is shown in the lower right panel.

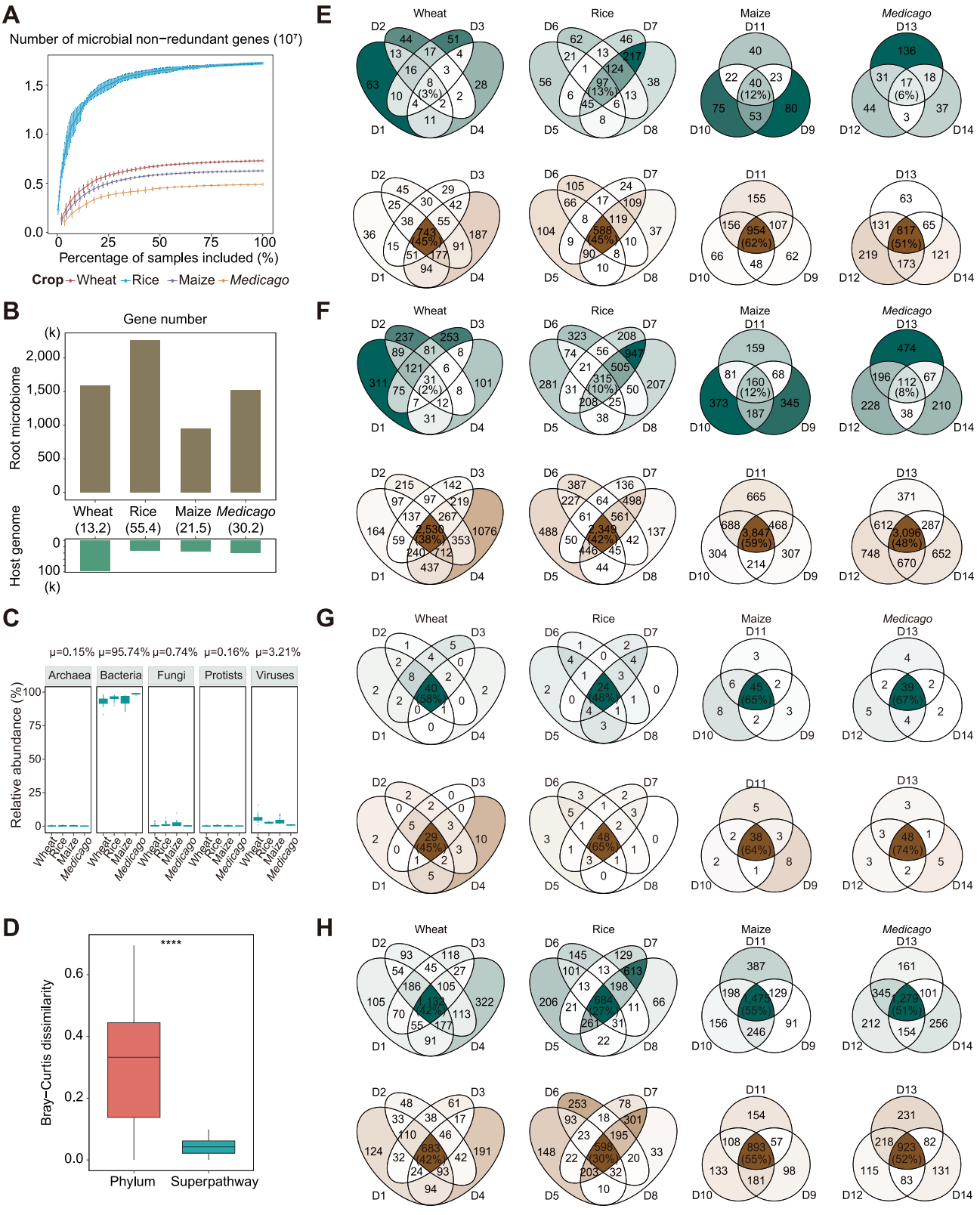
(D) Coexistence patterns of stress tolerance functions within individual bacterial genomes. The number and taxonomy of genomes for each coexistence pattern are shown in the stacked bar plot in the upper panel. The lower panel with vertical lines illustrates co-existent functions (biosynthesis of ACCd, ethylene and SA) within individual genomes. The number of genomes for each PGP group is shown in the lower right panel.

(E) Distribution of BGC classes according to the length of gene clusters. The bars are color-coded based on BGC classes. The lengths of NRPS are relatively higher than other BGC classes.

(F) Composition of BGC classes in each bacterial phylum. Proteobacteria are displayed at the class level due to their excessive species number.

(G) Relationship between the BGCs percentage in the genome and the corresponding genome size. Each dot represents a bacterial genome. The scatter plot illustrates the BGC percentage and genome size in each genome. The density of data in each axis is shown on the left or top of the scatter plot. Note that Actinobacteriota and Firmicutes have a higher proportion of genes encoding BGCs.

(H) Distribution of genome number and GCF richness of families in Actinobacteriota. The bar plot and boxplot show the genome number (left) and GCF number (right) in each family of Actinobacteriota. Families are arranged in descending order according to the median of GCF richness.



(legend on next page)

Figure S3. Taxonomic and functional patterns of root metagenomes from multiple crops in diverse soils, related to Figure 4

(A) Accumulation curves indicate that the number of non-redundant microbial genes in the root metagenomes of each crop species is approaching saturation. Data from wheat, rice, maize, and *Medicago* are presented in different colors. Each observation was sampled 20 times without replacement.

(B) The gene count comparison between the root metagenomes and host crop genomes. The bar plot shows gene count of non-redundant genes within the root metagenomes (upper) compared with corresponding crops (lower). The number with bracket under the host name represents the ratio of gene numbers between the root microbiome and the corresponding host.

(C) Relative abundance of archaea, bacteria, fungi, protists, and viruses in the root metagenomes of wheat, rice, maize, and *Medicago*. Note that bacteria contribute an average of 95.7% of overall microbial relative abundance within the root metagenomes.

(D) Taxonomic and functional composition of root metagenomic samples. Boxplot showing pairwise Bray-Curtis dissimilarity with root metagenomes of multiple crops grown in diverse soils based on taxonomic and functional compositions (n = 332, adjusted $p < 0.05$, Wilcoxon rank-sum test).

(E) Families enriched or depleted in root metagenomes of each crop species derived from diverse soils. The Venn diagram illustrates overlaps of bacterial families enriched (upper, green) and depleted (bottom, brown) in roots of wheat, rice, maize, and *Medicago* from multiple datasets. Each area is color-coded according to the number of functions in the area.

(F) Genera enriched or depleted in root metagenomes of each crop species derived from diverse soils. The Venn diagram illustrates overlaps of bacterial genera enriched (upper, green) and depleted (bottom, brown) in roots of wheat, rice, maize, and *Medicago* from multiple datasets. Each area is color-coded according to the number of functions in the area.

(G) Microbial functions enriched or depleted in root metagenomes of each crop species derived from diverse soils. The Venn diagram illustrates overlaps of independent datasets reflecting the functions consistently enriched (upper, green) and depleted (bottom, brown) in roots of wheat, rice, maize, and *Medicago*. Each area is color-coded according to the number of functions in the area.

(H) Functional KOs enriched or depleted in root metagenomes of each crop species derived from diverse soils. The Venn diagram illustrates overlaps of independent datasets reflecting the functional KOs consistently enriched (upper, green) and depleted (bottom, brown) in roots of wheat, rice, maize, and *Medicago*. Each area is color-coded according to the number of functions in the area.

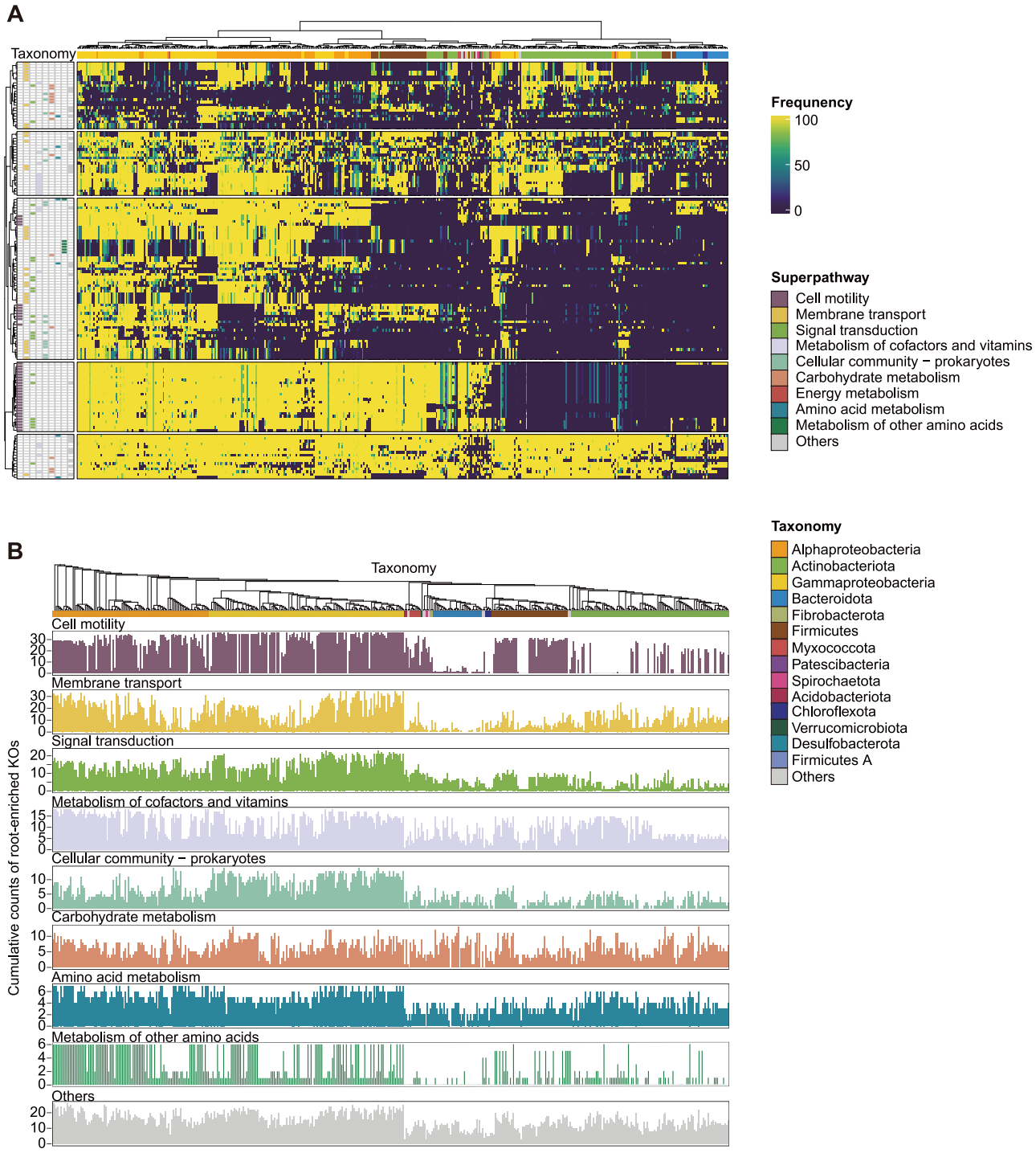


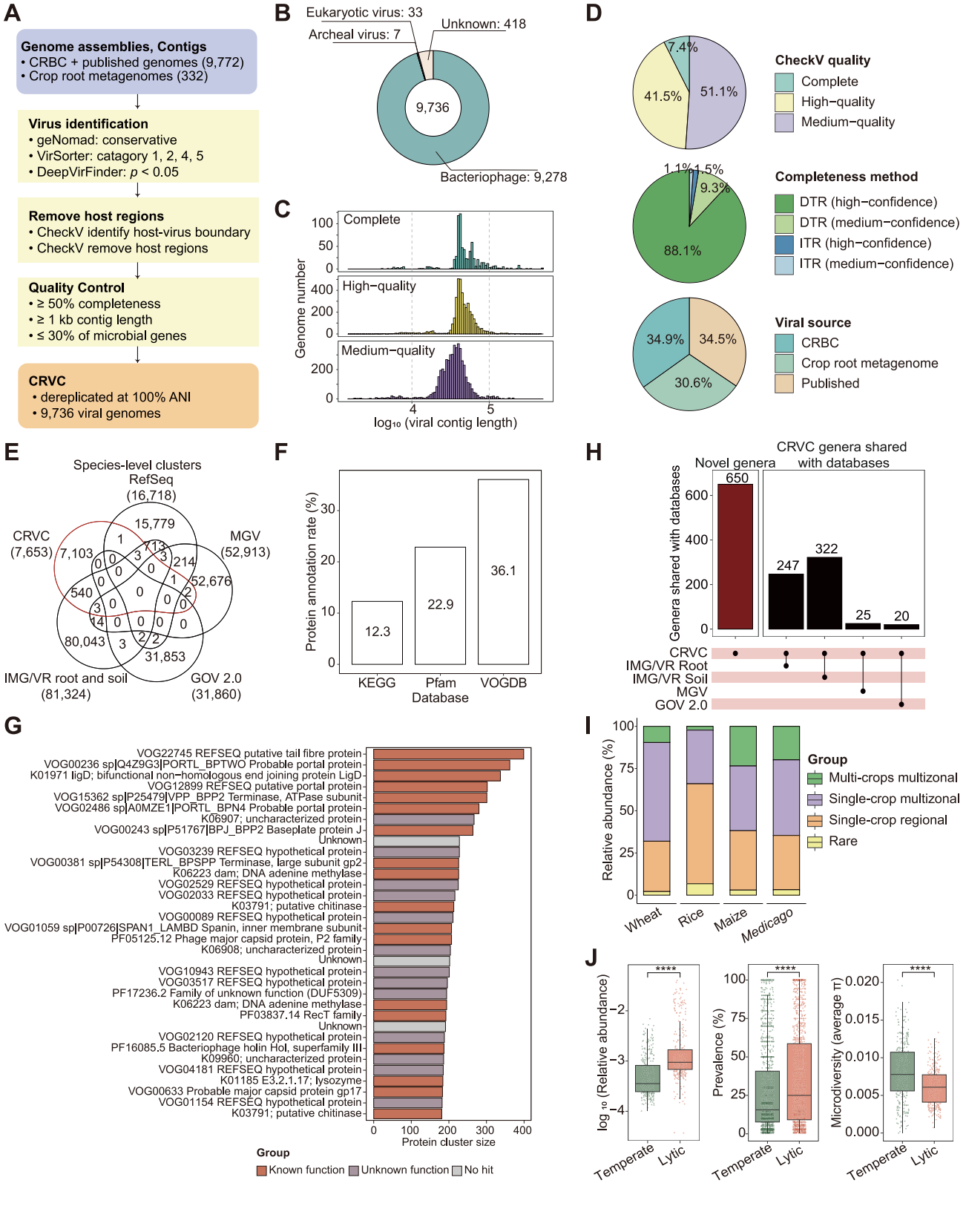
Figure S4. Distribution of 146 root-enriched KOs in the crop root bacterial genomes, related to Figure 4

(A) Clustering of 146 root-enriched KOs in the crop root bacterial genomes. The central heatmap shows the frequency of root-enriched KOs across crop root bacterial genomes. The frequency is calculated based on the percentage of genomes containing the KO in each genus. The distributions of KOs and genomes are clustered separately using the ward.D2 method. KOs are clustered into 5 groups according to their distribution in crop root bacteria and colored according to the

(legend continued on next page)

superpathways to which they belong. The taxonomy information of genomes is presented at the phylum level, with Proteobacteria at the class level due to their excessive abundance. Note that phylogenetically related genera exhibited similar functional patterns of root enrichment.

(B) Functional landscape of 146 root-enriched KOs in the genomes of crop root bacteria. Bar plots show the cumulative counts of root-enriched KOs that are prevalent in over 50% of genomes within each bacterial genus. KOs are grouped and color-coded based on their KEGG superpathways. Bacteria genera are arranged according to their phylogeny.



(legend on next page)

Figure S5. Quality, protein content, and distribution of the CRVC, related to Figure 5

(A) Workflow of CRVC construction. Viral contigs were predicted from 9,772 bacterial genomes in the CRBC and published databases and 332 crop root metagenomes, using a combination of software tools including geNomad, VirSorter, and DeepVirFinder. The removal of host regions and estimation of viral completeness were conducted using CheckV. Viral genomes underwent quality control and dereplication at 100% ANI, yielding 9,736 non-redundant microbial viral genomes designated as the CRVC. Note that 65.5% of CRVC genomes are derived from CRBC genomes and our crop root metagenomes.

(B) The composition of the CRVC. The doughnut chart illustrates the proportion and number of viruses in the CRVC according to the domain of hosts. The CRVC comprises 9,736 non-redundant viruses, including 9,278 bacteriophages, 7 archaeal viruses, 33 eukaryotic viruses, and 418 entries with unclear hosts (STAR Methods).

(C) Genome quality and length distribution of the CRVC. The bar plots show the length distribution of CRVC genomes. CRVC genome completeness is categorized into three quality groups: 723 complete genomes, 4,041 high-quality genomes (>90% completeness), and 4,972 medium-quality genomes (50%–90% completeness). Note that nearly half (48.9%) of the genomes in the CRVC are over 90% completeness (high-quality and complete). The median length of viral contigs is 40.6 kbp.

(D) Overview of CRVC quality. The upper pie chart represents the distribution of the CRVC genomes across different quality categories as evaluated by CheckV. Note all of the CRVC genomes exceed the medium-quality level with completeness over 50%. The middle pie chart illustrates the distribution of 723 complete CRVC genomes acquired through the direct terminal repeats (DTRs) or inverted terminal repeats (ITRs) methods, with methods and confidence levels color-coded for clarity. The bottom pie chart illustrates the proportion of the CRVC genomes derived from the CRBC genomes, our crop root metagenomes, and published crop root bacteria, each color-coded for clarity.

(E) Majority of the CRVC species-level clusters are not reported in public databases. The Venn diagram illustrates overlaps of the CRVC species-level clusters (vOTUs) with the public viral databases including, RefSeq, MGV, GOV2, and IMG/VR. In the IMG/VR database, viral genomes sourced from both root and soil environments were incorporated into the analysis.

(F) Annotation rate of the CRVC proteins by different reference databases. The bar plot shows the percentage of viral proteins in the CRVC annotated using the KEGG, Pfam, VOGDB databases, respectively. In total, we predicted 587,992 proteins within the CRVC, of which the annotation rates for these databases ranged from 12.3% to 36.1%. Note that the low annotation rates across various reference databases suggest that a significant portion of the CRVC comprises proteins with unknown functions.

(G) Functional annotations of the largest 34 protein clusters within CRVC. The bar plot shows the number of proteins in each abundant protein cluster, and only protein clusters with numbers greater than 180 are shown. The annotation of each protein cluster is color-coded according to annotation categories: known, unknown functions, and no hit. Note that a substantial proportion of protein clusters are either with unknown functions or lacking matches in databases.

(H) Comparison between the CRVC and public viral databases using genomes derived from metagenomes. The bar plot shows the number of metagenome-derived CRVC viral clusters that are shared with publicly available databases at the genus level. Bar plots showing that 650 metagenome-derived unreported CRVC clusters whose genomes are distinct from those in published viral databases, including viruses from IMG/VR, MGV, and GOV2. The CRVC exhibits a higher shared presence of viruses with IMG/VR soil and roots than other ecosystems, suggesting the ecosystem specificity of viral distribution.

(I) Viral distribution in crop root habitats. The stacked bar plot illustrates the composition of viral categories in the root microbiomes of each crop species. Using prevalence characteristics, viruses are categorized into four prevalence groups: stable presence across roots of multiple crops grown in multiple locations (multi-crops multizonal, green), within roots of a single crop species in multiple locations (single-crop multizonal, purple), within roots of a single crop species in a single location (single-crop regional, orange), and viruses with prevalence lower than 10% of samples (rare, yellow).

(J) The characteristics of temperate and lytic viruses in crop root microbiomes. The bar plots illustrate the $\log_{10}$-transformed relative abundance (left), prevalence (middle), and microdiversity (right) of temperate and lytic viruses in 332 crop root metagenomic samples (adjusted $p < 0.05$, Wilcoxon rank-sum test).

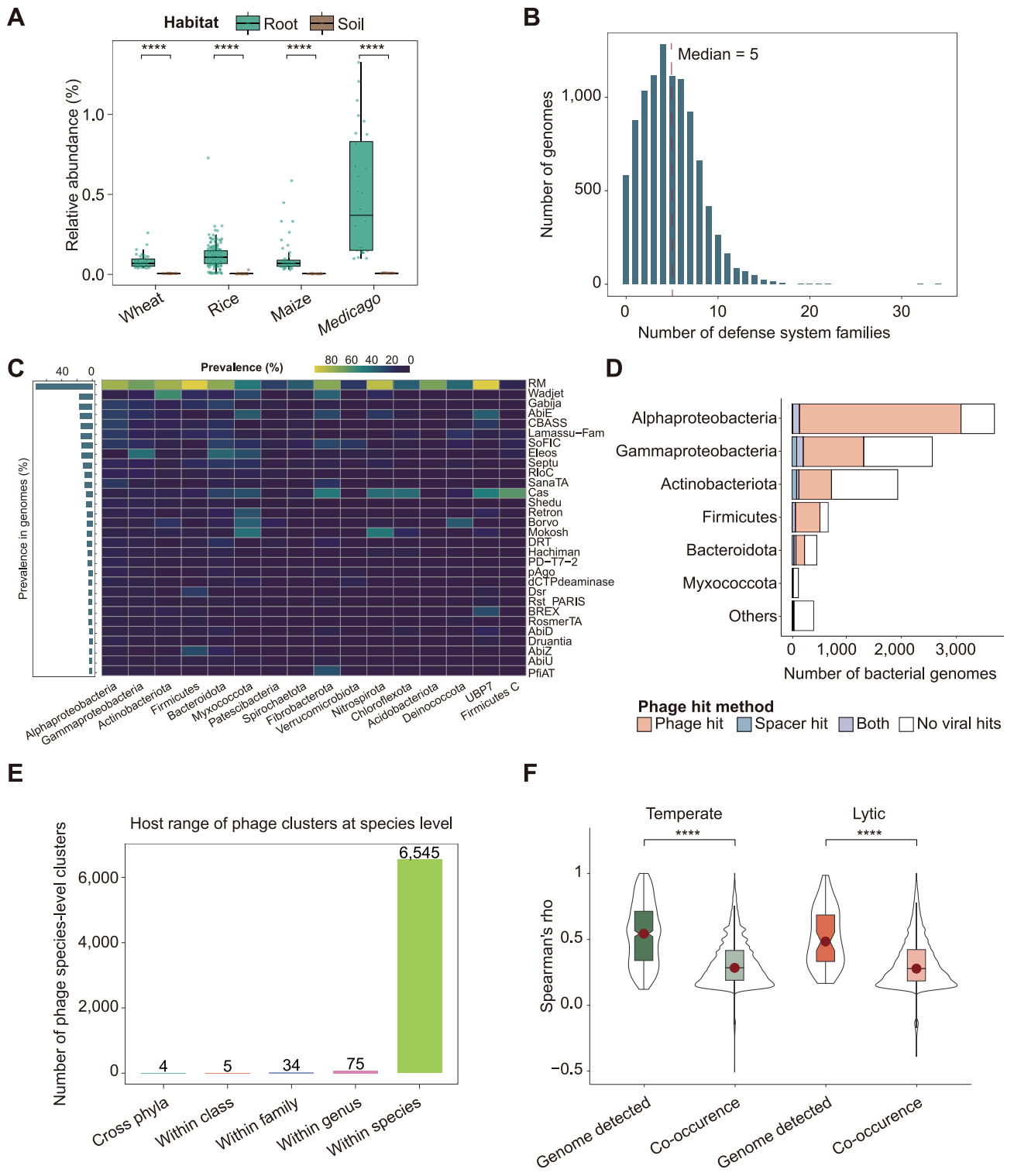


Figure S6. Phage-bacteria connections within bacterial genomes and crop root ecosystems, related to Figure 6

(A) Phage abundances in the root microbiomes of multiple crops are higher than those in corresponding soils based on the IMG/VR version 4 database. The boxplot shows the median and quartiles of phage relative abundance in root metagenomic samples from wheat, rice, maize, *Medicago*, and corresponding soils. Statistic differences are assessed using the Wilcoxon rank-sum test, root, $n = 332$; soil, $n = 75$. **** represents adjusted $p < 0.0001$.

(legend continued on next page)

(B) The distribution of antiviral defense system families in the bacterial genomes. The histogram shows the distribution of the number of defense system families in each bacterial genome.
(C) The prevalence of bacterial defense systems in crop root bacterial genomes. The heatmap illustrates the prevalence of each bacterial defense system at the phylum level. Proteobacteria are shown at the class level due to their excessive abundance. The bar plot (left) showing the prevalence of each bacterial defense system in all crop root bacteria.
(D) Substantial proportion of genomes within major crop root bacteria phyla exhibit connections with phage genomes. Bar plot shows the number and proportion of crop root bacterial genomes with phage hits according to phage contigs and CRISPR spacers detected in bacterial genomes. Each bar represents the genome number within each bacterial phylum (class level for Proteobacteria). Bacterial genomes with phage hits are colored orange, those with CRISPR spacer hits are colored blue, and genomes with phage hits detected by both methods are shown in purple.
(E) Large majority of phage species-level clusters show specific connections with bacteria. The bar plot illustrates the host range of phage species-level clusters at different taxonomic levels. Note that most phage clusters have a narrow host range and can only interact with bacteria within a single genus or species.
(F) The phage-bacteria connections identified by genomes show higher correlation efficiency than links only determined by co-occurrence in crop root ecosystems. The boxplot illustrates the Spearman correlation coefficients of temperate phage-bacteria connections detected via genomes (dark green) or via co-occurrence (light green), as well as the Spearman correlation coefficients of lytic phage-bacteria connections detected via genomes (dark red) or via co-occurrence (light red). Statistic differences are assessed using the Wilcoxon rank-sum test, root, $n = 332$. **** represents adjusted $p < 0.0001$.