mSystems | Open Peer Review | Ecology | Research Article

Global analyses of biosynthetic gene clusters in phytobiomes reveal strong phylogenetic conservation of terpenes and aryl polyenes

Arijit Mukherjee,[1,2] Hitesh Tikariha,[1,2] Aditya Bandla,[2,3] Shruti Pavagadhi,[2,3] Sanjay Swarup[1,2,3]

AUTHOR AFFILIATIONS See affiliation list on p. 13.

ABSTRACT There are gaps in our understandings on how did the evolutionary relationships among members of the phytobiomes shape their ability to produce tremendously complex specialized metabolites under the influence of plant host. To determine these relationships, we investigated the phylogenetic conservation of biosynthetic gene clusters (BGCs) on a global collection of 4,519 high-quality and nonredundant (out of 12,181) bacterial isolates and metagenome-assembled genomes from 47 different plant hosts and soil, by adopting three independent phylogenomic approaches (*D*-test, Pagel's λ, and consenTRAIT). We report that the BGCs are phylogenetically conserved to varying strengths and depths in their different classes. We show that the ability to produce specialized metabolites qualifies as a complex trait, and the depth of conservation is equivalent to ecologically relevant complex microbial traits. Interestingly, terpene and aryl polyene BGCs had the strongest phylogenetic conservation in the phytobiomes, but not in the soil microbiomes. Furthermore, we showed that terpenes are largely uncharacterized in phytobiomes and pinpointed specific clades that harbor potentially novel terpenes. Taken together, this study sheds light on the evolution of specialized metabolites' biosynthesis potential in phytobiomes under the influence of plant hosts and presents strategies to rationally guide the discovery of potentially novel classes of metabolites.

IMPORTANCE This study expands our understandings of the biosynthetic potential of phytobiomes by using such worldwide and extensive collection of microbiomes from plants and soil. Apart from providing such vital resource for the plant microbiome researchers, this study provides fundamental insights into the evolution of biosynthetic gene clusters (BGCs) in phytobiomes under the influence of plant host. Specifically, we report that the strength of phylogenetic conservation in microbiomes varies for different classes of BGCs and is influenced as a result of plant host association. Furthermore, our results indicate that biosynthetic potential of specialized metabolites is deeply conserved equivalent to other complex and ecologically relevant microbial traits. Finally, for the most conserved class of specialized metabolites (terpenes), we identified clades harboring potentially novel class of molecules. Future studies could focus on plant–microbe coevolution and interactions through specialized metabolites building upon these findings.

KEYWORDS biosynthetic gene clusters, phylogenetic conservation, plant microbiome, squalene hopene cyclase, squalene phytoene synthase, terpenes

Plant-associated microbiomes, or phytobiomes, provide critical services for plant growth, development, and health (1–3). Host plants and their phytobiomes collectively produce diverse specialized metabolites of immense importance, influencing

Editor Aurélie Deveau, UMR1136 INRA Université de Lorraine, Champenoux, France

Address correspondence to Sanjay Swarup, sanjay@nus.edu.sg.

Arijit Mukherjee and Hitesh Tikariha contributed equally to this article. Author order was decided alphabetically.

The authors declare no conflict of interest.

See the funding table on p. 13.

Received 21 April 2023
Accepted 30 May 2023
Published 6 July 2023

the chemical ecology of different niches (4, 5). Importantly, these metabolites shape the outcome of "microbe–microbe" and "plant–microbe" interactions. Specific classes of molecules, such as benzoxazinoids, phenylpropanoids, and terpenes secreted from plant roots facilitate root microbiome assembly (5–8). Similarly, terpenes, benzenoids, and methanol from leaf exudates shape leaf microbiomes (9). Moreover, specialized microbial metabolites also modulate plant growth and defense (10, 11).

The specialized metabolites from microbes are encoded by numerous families of biosynthetic gene clusters (BGCs) (12). BGCs are highly complex in terms of their genetic organization, as they comprise enzymes, regulators, and transporters (13, 14). They are grouped into seven major classes based on their chemical structures, which are further grouped into subcategories based on both specific chemical moieties and gene cluster similarity (15–17). For example, the terpene class consists of the subcategories hemiterpenes, triterpenes, and tetraterpenes, which also have correspondingly different BGCs (18). On the other hand, structurally conserved metabolites such as aryl polyenes do not have any evident subdivision (19). Such complexity of BGCs limits our understanding of their influence on the "chemical ecology" with respect to phytobiome function.

Phylogenomic approaches, such as phylogenetic trait analysis, can be used to determine the consensus between microbial clades and their associated traits (20). Phylogenetic conservation of specific traits has been used in classical ecology (21), and now, with refined tools, it can be applied in microbial ecology (22). Recently, by the use of this approach, phylogenetic conservation of carbon assimilation (23), responses to nitrogen addition (24), and several other functional traits (22) in microbes have been well characterized. Since specialized metabolites directly govern plant–microbe and microbe–microbe interactions, the ability to produce these metabolites can also be considered an "*effect trait*" (20). Therefore, this approach can be adopted to investigate the association of specific categories of specialized metabolites with microbial clades within and between different microbial ecosystems.

Here, we investigated how phylogenetic relationships of microbial members within and between plant and soil ecosystems have shaped their potential of secondary metabolite biosynthesis. First, we collated phytobiome data sets from cultured bacterial isolates (referred as "plant isolates") and metagenome-assembled genomes (referred as "plant MAGs") and included soil-associated cultured bacterial isolates (referred as "soil isolates") and MAGs (referred as "soil MAGs") as reference. Next, we asked the following questions: (i) Do phylogenetically related microbial members have similar secondary metabolite biosynthesis potential in phytobiomes? (ii) If so, what are the strength and depth of the phylogenetic conservation of different BGC classes among phytobiomes compared to the soil microbiomes? and (iii) How could this approach be used as a tool to identify novel microbial clades that are associated with different subgroups of specialized metabolites? To address these questions, we first predicted the BGCs of 4,519 high-quality and nonredundant genomes (both isolates and MAGs from plant and soil) and investigated their phylogenetic distribution. Next, we calculated the strength and depth of the phylogenetic conservation of BGCs of the four data sets by the use of three independent statistical approaches. Finally, for the most conserved BGC class in the plant isolates, we applied sequence-based analyses to identify microbial clades harboring potentially novel secondary metabolites. Overall, these approaches led to the identification of the patterns of phylogenetic conservation of BGCs in plant and soil microbiomes and have potential application in the guided discovery of novel specialized metabolites.

MATERIALS AND METHODS

Study inclusion

The genome sequences of plant-derived isolates were searched and retrieved from two main sources: the JGI (Joint Genome Institute—https://jgi.doe.gov/) and PATRIC databases (25). We followed the following criteria to select for plant-derived isolates:

JGI: filtering criteria—Analysis Project → Genome Analysis (Isolate), Study Ecosystem → Host-associated, Project, Sequencing Strategy → Whole Genome Sequencing, Study Ecosystem Category → Plants; PATRIC: filtering criteria—Filtered by plant name, excluded plasmid and poor-quality genomes (data included until 31 December 2020). Isolates for which either the plant host name was missing or niche information was unavailable were removed. Finally, we merged the plant-derived isolates after removing duplicated (using assembly number) and poor-quality genomes (>5% contamination or <95% complete). This resulted in 4,931 genomes of plant-derived isolates (Table S2). Similarly, genome sequence of soil isolates was retrieved from JGI: filtering criteria—Sequencing strategy → Whole Genome Sequencing, Study Ecosystem Category → Terrestrial; Habitat → soil. Source of isolates, such as glaciers, greenhouse, or having possible plant influence, was excluded. Finally, we merged the soil isolates after removing duplicated (using assembly number) and poor-quality genomes (>5% contamination or <95% complete). This resulted in 2,572 soil isolates.

The genome FASTA files were then downloaded via their genome assembly number using Bioinformatics Tools (bit) (https://github.com/AstrobioMike/bit#bioinformatics-tools-bit) (26). Missing metadata were added manually [located from the NCBI database and the literature, genome size using Quast (v5.0.2) (27), and completeness using CheckM (v1.2.0) (28)]. Taxonomic assignment of the genomes was performed using GTDB-Tk v1.5.1 (29). Furthermore, all the genomes were dereplicated at 98% ANI (30) with fraction alignment of 0.6 using dREP v2.6.2 (31). This cutoff was chosen based on a recent benchmark that demonstrated 98% ANI as an appropriate threshold for generating sub-species level representation (30) of genomes, and in our case, it best represented nonredundant genomes while preserving the microheterogeneity of BGCs. Finally, the resulting 1,395 genomes from plant isolates and 1,768 soil isolates were used for further analyses.

Plant and soil metagenome-assembled genomes (MAGs) were considered from the Earth Microbiome project (32) and our recently published study (33). First, we filtered MAGs belonging to "terrestrial ecosystem" and "soil ecosystem" types to consider them as soil MAGs. Furthermore, this list of soil MAGs was manually curated to remove MAGs derived from the rhizosphere to exclude plant influence. For plant MAGs, we collated the MAGs from both the Earth Microbiome data set and the data set provided by Bandla et al. (33). Bulk and compost soil MAGs from Bandla et al. were excluded from the list of plant MAGs. Both data sets were curated to include only high-quality (≥70% completeness and ≤5% contamination) MAGs. Furthermore, the MAGs in each data set were dereplicated using the same criteria (as used for isolates), 98% ANI with a fraction alignment of 0.6, resulting in 573 plant MAGs and 783 soil MAGs.

BGC prediction

We used antiSMASH (v5.1.2) (34) (Command: antismash --genefinding-tool prodigal --asf --cb-general --cb-subclusters --cb-knownclusters –pfam2go) for BGC predictions. The antiSMASH results were processed using BiG-SCAPE (v1.0.1; using GCF or Gene Cluster Family clustering threshold of 0.3) (35) and MIBiG (15) reference database version 1.4 for clustering the sequences into systematic BGC categories. The BGCs were classified into 11 classes as per the output from BiG-SCAPE (and as per their abundance in the whole data set, some were moved into and out from class Others), i.e., nonribosomal peptides (NRPS), ribosomally synthesized and posttranslationally modified peptides (RiPPs), terpenes, aryl polyenes, beta-lactones, hserlactone, siderophores, PKS_NRP_hybrids, polyketide synthase other (PKS_other), PKSI, and others. The output table was parsed to obtain the final table of the number of BGCs in each class for each genome across all four data sets.

Phylogenetic tree and BGC distribution

An unrooted phylogenetic tree of both isolates and MAGs was constructed separately using the GTDB-Tk v1.5.1 *de novo* option, which uses FastTree (36) method for tree

construction. To obtain a more robust maximum likelihood (ML) tree, the supermatrix of all the sequences for each data set from GTDB-Tk output was trimmed separately using BMGE (Block Mapping and Gathering with Entropy) v1.12 (37). The trimmed alignment was used to reconstruct the best consensus tree by IQ-TREE 2.2.0.3 (38) with Q.yeast+F+R10 as substitution model (determined by model finder), branch support calculated by SH-like approximate likelihood ratio test with 1,000 bootstrap replicates along with 1,000 ultrafast bootstrap approximation. The phylogenetic tree for plant isolates was rooted at the Desulfobacterota phylum. For plant MAGs, soil isolates, and soil MAGs, the trees were rooted at the Phylum Cyanobacteriota. The BGC count and metadata were overlaid over the phylogenetic tree using iTOL (39) to visualize the BGC distribution.

The phylogenetic distance among plant isolate members was calculated based on the "cophenetic distance," whereas BGC dissimilarity was calculated based on the "Bray-Curtis" (40) distance from their BGC count table. The strength (Mantel's *r*) and significance (*P* value based on 1,000 permutations) of the correlation between these two-distance matrices were calculated based on Mantel's test using the "ade4" package in R (41).

Principal coordinate analysis (PCoA) was performed based on "Bray-Curtis" distance from their BGC count table. Statistical significance was calculated based on PERMANOVA using the "vegan" package in R (42).

We estimated the differential enrichment of BGCs among taxa by modeling the count of BGCs in each genome. The "negative binomial" distribution was fitted to model the count of BGCs against genus. Model selection was performed based on residual diagnostics and comparing the Akaike information criteria of the models using the "DHARMA" package in R (43). Differential enrichment for BGCs has been shown for the top 7, and only significantly abundant genera in isolates (plant and soil) data sets.

Phylogenetic conservation of BGCs

Phylogenetic conservation of BGCs was inferred based on three independent approaches: the *D*-test of Fritz and Purvis (21), lambda (λ) statistics of Pagel (44), and consenTRAIT of Martiny et al. (22). Notably, these approaches are different in principle. The *D*-test calculates the strength of phylogenetic conservation from binary traits, whereas Pagel's λ calculates phylogenetic signal from continuous traits. On the other hand, consenTRAIT calculates the genetic depth of conservation from binary traits. Each of these approaches has its own advantages and limitations; for example, consenTRAIT considers the clades with mixed responses, in contrast to *D*-test (24). However, both consenTRAIT and *D*-test/Pagel's λ calculate different attributes of phylogenetic conservation—genetic depth and strength of phylogenetic conservation, respectively. The raw count table of BGCs was converted to a binary table by setting counts greater than or equal to 1 to 1.

D-statistics for phylogenetic conservation were calculated based on binary count table of BGCs, using the default parameters of phylo.D function ("caper" package in R) (45). We permuted the tips of the tree 1,000 times based on a random evolution model and Brownian mode (BM) of evolution to estimate the statistical significance. D-statistics with P_{random} <0.05 and $P_{Brownian}$ <0.05 were considered statistically significant. Pagel's λ for continuous trait values were calculated using the "phylosig" function of "geiger" package (46) in R with default parameters. Statistical significance was calculated based on likelihood ratio test with random evolution model (λ = 0). We performed consenTRAIT to determine the genetic depth at which the traits are conserved. consenTRAIT finds clades at which 90% of the descendants have the trait of interest (here, each of the BGCs) and calculates the mean genetic depth (τ_D) of those clades. We used the getTraitDepth function from the "castor" package in R, and the parameters used were as follows: min_fraction = 0.9, count_singletons = TRUE, singleton_resolution =0, weighted = FALSE, as described by Martiny et al. (22). Statistical significance was determined by permuting the tree tips 1,000 times and when the proportion of times τ_D was less than the observed

TABLE 1 Mean genetic depth of conservation of individual BGC classes among the four data sets[a]

BGC category	Plant isolates	Soil isolates	Plant MAGs	Soil MAGs
Aryl polyene	**0.02**	0.021	0.04	0.049
Beta-lactone	**0.018**	**0.031**	**0.052**	0.051
Hserlactone	**0.023**	0.027	0.038	0.044
NRPS	**0.026**	0.036	0.051	0.057
RiPPs	**0.032**	**0.041**	**0.061**	0.061
Siderophore	**0.018**	0.028	0.043	0.037
Terpenes	**0.036**	**0.05**	**0.069**	**0.07**
PKS_NRP_Hybrids	**0.022**	0.027	0.054	0.054
PKS_other	**0.022**	**0.034**	0.039	0.056
PKSI	**0.016**	0.029	**0.067**	0.059
Others	**0.019**	0.033	**0.06**	0.05
Average	0.0229	0.0325	0.0522	0.0535

[a]Bold numbers indicate statistically significant mean genetic depth (τ_D) of conservation ($P < 0.05$).

τ_D. The mean, maximum, and minimum genetic depths for each of the BGCs were subsequently calculated and shown in Table 1.

Terpene network and domain analysis

We investigated the amino acid sequence similarity network of only terpenes in phytobiomes (plant isolates and plant MAGs) since terpenes were found to be more abundant and diverse than aryl polyenes. The network file of terpenes containing GCF information of both plant isolates and plant MAGs obtained from BiG-SCAPE was overlaid with taxonomic information (at the class level) of the genomes. The network was then visualized using Cytoscape (v3.8.2) (47). The nodes were colored according to the bacterial class and MiBIGs referenced BGCs.

The Pfam domain list belonging to terpenes was extracted from the BiG-SCAPE output and was sorted based on the number of hits. Among all the domains, only those belonging to core biosynthetic genes and with high abundance were considered for further analysis. The presence/absence data of these domains were used to visualize their distribution in the phylogenetic tree of the plant isolates. To identify differences in the sequences with SQS/PSY (squalene/phytoene synthase) domain, the sequences were extracted from the BiG-SCAPE output and aligned using MAFFT (v7.475) (48). Since more than 80% of gaps were found after alignment, BMGE (v1.12) (37) was used to extract regions from the multiple sequence alignment for phylogenetic inference. The BMGE output was then used for phylogenetic tree construction by RAxML (v8.2.12) (49) with the following parameters: GAMMA model of rate heterogeneity and maximum likelihood (ML) estimate of alpha-parameter with 100 bootstraps. The consensus bootstrap tree was generated using the majority consensus option in RAxML. The resulting tree was then visualized using iTOL. The same method used for Pfam domains was applied to extract, align, and visualize SQS-PSY domains.

RESULTS

Phylogenetic distribution of BGCs in phytobiomes

We mined publicly available databases to generate a comprehensive catalog of isolates and MAGs from plants and soil microorganisms (Tables S2, S5, S8, and S11). Taken together, this collection represented a total of 4,931 plant isolates, 1,523 plant MAGs, 2,572 soil isolates, and 1,316 soil MAGs after manual curation and confirmation of the source of isolation from the plant or soil environment (see Materials and Methods). Our collection of phytobiomes is broad, covering 47 different plant hosts across 90 countries (Fig. S1A and C; Tables S2 and S5). The quality profiles of all the data sets, including genome size, completeness, and contamination, are shown in Fig. S1B. The genomes of plant and soil isolates were larger and near complete (median completeness

= 99.59% and 99.54%, respectively; median genome size = 5.74 Mb and 6.16 Mb, respectively) in comparison with those of the plant and soil MAGs (median completeness = 82.93% and 88.81%, respectively; median genome size = 3.69 Mb and 3.49 Mb, respectively). Our collection of phytobiomes covers the most abundant phyla (Proteobacteria, Actinobacteria, Firmicutes, and Bacteroidetes at a ratio of 66:18:9:5) in plant environments, as suggested in previous studies (50, 51). Notably, Firmicutes were underrepresented in MAGs compared to the isolates in both plant and soil (Fig. S1D). However, the collection of plant MAGs covered the most abundant taxa found in the plant environment.

Considering only high-quality and nonredundant genomes, we report a total of 12,916, 23,152, 2,318, and 2,307 predicted BGCs from the plant isolates, soil isolates, plant MAG, and soil MAG data sets, respectively (Tables S3, S12, S6, and S9). NRPS was found to be the most abundant BGC, with an average BGC counts per genome of 2.047, 2.959, 0.916, and 0.706 for plant isolates, soil isolates, plant MAGs, and soil MAGs, respectively (Tables S3, S12, S6, and S9). The number of total BGCs per genome varied considerably among members (Fig. 1A) and among the four data sets (Fig. S1E). We observed characteristic patterns in the distribution of different BGC classes among taxonomic members of plant isolates (Fig. 1). Visualization of the distribution of BGCs in plant isolates revealed that members of the class Actinomycetes (*Streptomyces* sp.) encode a higher number of BGCs per genome compared to those of other taxa and represent a metabolically dynamic clade (Fig. 1A and inset A-1). This was also reflected in the BGC profiles of soil isolates (Fig. 2, inset A-1). Moreover, members of the class Alphaproteobacteria (such as *Mesorhizobium*, *Bradyrhizobium*, and *Rhizobium*) encoded a higher number of homoserine-lactones compared to other members of their taxonomic group (Fig. 1A and inset A-2). However, this trend was particularly absent in soil isolates, suggesting differential patterns of BGC distribution in soil and plant isolates (Fig. 2 and inset A-2). Additionally, the principal coordinate analysis (PCoA) of the BGC profile showed that members of the most abundant taxonomic groups (top 7 abundant genera in plant isolates) occupy distinct biosynthetic space in the coordinate plot, suggesting that BGCs are phylogenetically clustered (Fig. 1B). Similarly, the biosynthetic space of soil isolates also displayed a distinct clustering in the coordinate plot (Fig. 2B).

Differential enrichment analysis revealed a parallel trend of BGC distribution, as observed in Fig. 1A. For example, *Streptomyces* (phylum Actinomycetes) encode higher number of BGCs involving NRPS, PKS-other, PKS-NRP hybrids, siderophores, and terpenes in their genomes (negative binomial; $P < 0.05$ and Fig. 1C) compared to those of other taxa. We noted that members of the *Bacillus_A* genus possess equal proportions of NRPS and RiPP BGCs but have a lower number of BGCs involving aryl polyenes and PKS-other in their genomes compared to those of other taxa (Fig. 1C). Whereas this trend was not particularly reflected in soil isolates, suggesting that certain BGCs are differentially distributed in phytobiomes (Fig. 2C). The detailed results for differential analyses of BGCs for plant and soil isolates are presented in Tables S15 and 16, respectively.

The strength and depth of phylogenetic conservation vary for different BGCs among phytobiomes

Since we observed that the biosynthetic space of plant isolates is taxonomically distinct, we next tested whether there is a relationship between their phylogenetic distance and their BGC profiles using dissimilarity measures. Overall, there was a moderate yet significant correlation between BGC profile dissimilarity and phylogenetic distance (Mantel's $r = 0.31$; $P = 0.0009$; Table S14), suggesting that BGCs may be phylogenetically conserved. To investigate this further, we considered the individual BGC classes and sought to determine what the strength and depth of their phylogenetic conservation were among the phytobiome clades. We first investigated phylogenetic conservation in plant isolates and then used plant MAGs as an independent data set for validation. Next, we compared this pattern of phylogenetic conservation of BGCs with that of the soil isolates and soil MAGs separately, to understand how their conservation differs between

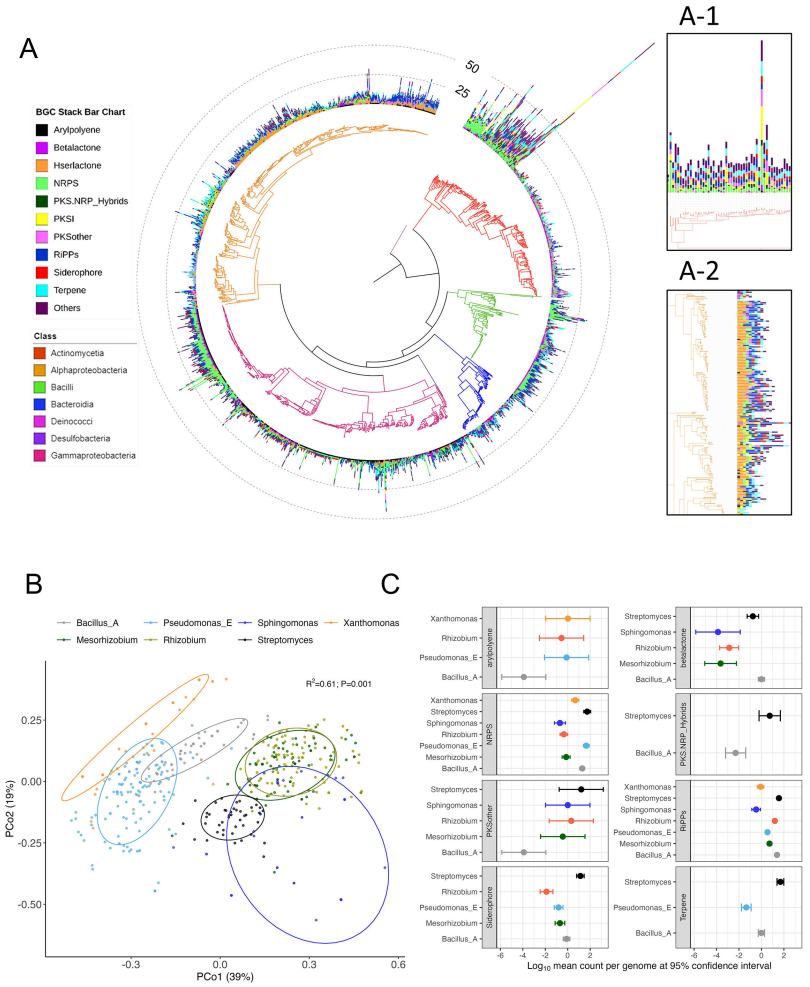


FIG 1 Phylogenetic distribution of BGCs in plant isolates. (A) Phylogenetic tree displaying distribution of BGCs among members of plant isolates. Stacked bars in the outer rings indicate the count of different BGCs (shown in different colors) encoded by individual members. Taxonomic information (at the class level) is shown in colors of the phylogenetic tree branches. Inset A-1 highlights that Actinomycetes phylum possess overall higher number of BGCs and inset A-2 highlights that Alphaproteobacteria encode higher number of hserlactones in their genomes. (B) Principal coordinate plot based on BGC profile dissimilarity of genomes of plant isolates. Ellipses show the parametric smallest area around the mean that contains 80% of the probability mass of each genus. Statistical significance was calculated based on PERMANOVA. Taxonomic groups occupy distinct regions of the biosynthetic space as indicated by distinct clusters of members of each genus. Data have been shown for the top 7 abundant genera in plant isolates. (C) Forest plot displaying differential enrichment of BGC categories among the top 7 abundant genera in plant isolates. The $\log_{10}$ mean count per genome (dots) and 95% confidence interval (upper and lower bars) are shown only for differentially abundant (negative binomial; $P < 0.05$) BGC categories relative to Bacillus_A.

plant and soil environments. The *D*-test of Fritz and Purvis (21) showed that all BGC classes display phylogenetic signals among plant isolates (Fig. 3A-1). The *D*-test statistic, a measure of the strength of phylogenetic signals, ranged from 0.06 to 0.46 (average *D*-value for all BGCs = 0.22; P_{random} <0.05), indicating that the strength of phylogenetic signals differed for the different BGC classes (Fig. 3A-1). We also calculated Pagel's λ, a measure of phylogenetic signal on continuous traits data, and found that λ values range from 0.80 to 0.99 (average λ value for all BGCs = 0.89; likelihood test $P < 0.00001$) (Table S19). These suggest that all BGCs display phylogenetic signal in plant isolates data set.

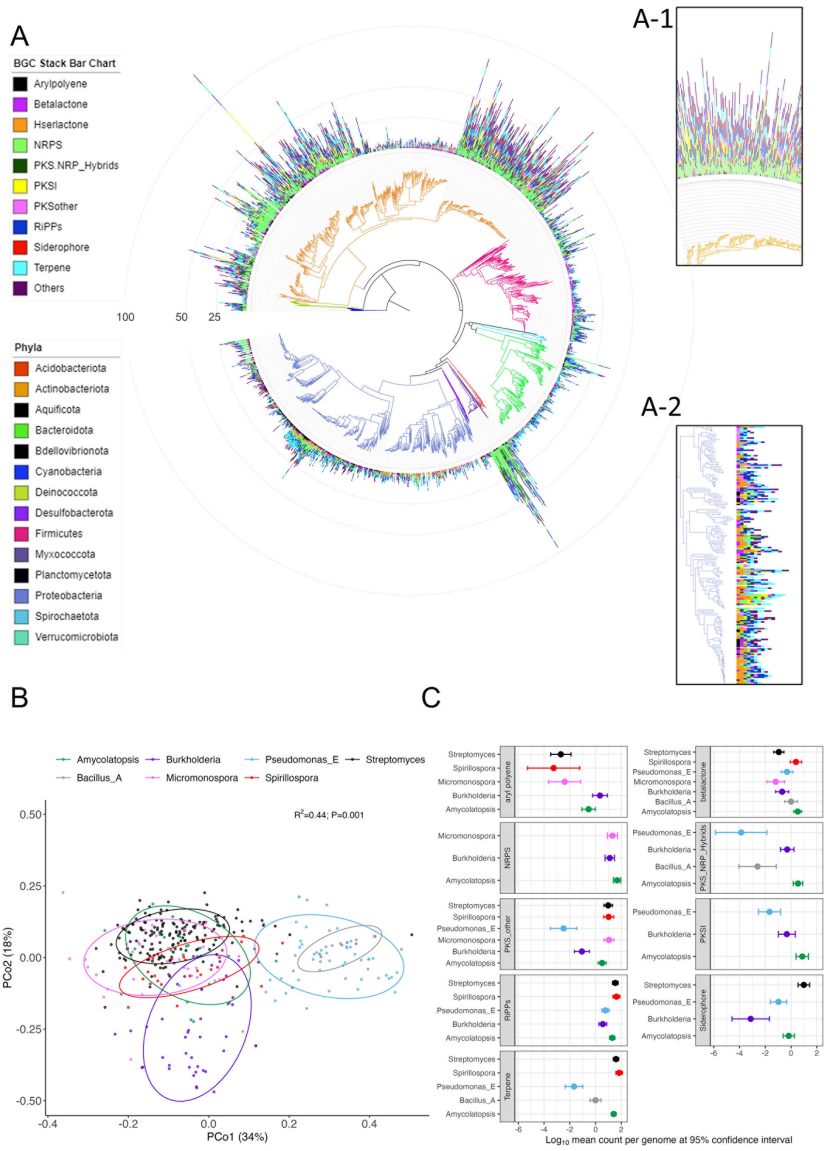


FIG 2 Phylogenetic distribution of BGCs in soil isolates. (A) Phylogenetic tree displaying the distribution of BGCs among members of soil isolates. Stacked bars in the outer rings indicate the count of different BGCs (shown in different colors) encoded by genomes of individual members. Taxonomic information (at the class level) is shown in colors of the phylogenetic tree branches. Inset A-1 highlights that Actinomycetes phylum possess overall higher number of BGCs and inset A-2 highlights that Alphaproteobacteria lacking characteristic presence of hserlactones compared to that of plant isolates. (B) Principal coordinate plot based on BGC profile dissimilarity of genomes of soil isolates. Ellipses show the parametric smallest area around the mean that contains 80% of the probability mass of each genus. Statistical significance was calculated based on PERMANOVA. Taxonomic groups occupy distinct regions of the biosynthetic space as indicated by distinct clusters of members of each genus. Data have been shown for the top 7 abundant genera in soil isolates. (C) Forest plot displaying differential enrichment of BGC categories among the top 7 abundant genera in soil isolates. The $\log_{10}$ mean count per genome (dots) and 95% confidence interval (upper and lower bars) are shown only for differentially abundant (negative binomial; $P < 0.05$) BGC categories relative to Amycolatopsis.

Furthermore, we applied the consenTRAIT approach (52) to estimate their mean genetic depth (τ_D) of conservation. The consenTRAIT results demonstrated that BGCs are conserved in plant isolates, with a mean genetic depth ranging from 0.016 to 0.036 (average $\tau_D = 0.0229$) ($P < 0.05$) (Table 1). Taken together, the *D*-test and consenTRAIT

results established that BGCs are phylogenetically conserved among the phytobiomes. We applied all three approaches in plant MAGs and found that the BGCs are indeed phylogenetically conserved (average *D*-value for all BGCs = 0.77; P_{random} <0.05), albeit the strength of the conservation was slightly weaker compared to plant isolates (Fig. 3A). The mean genetic depth (τ_D) of the conservation of BGCs ranged from 0.038 to 0.069 (average τ_D = 0.0522) in the plant MAGs, which was similar to that in the plant isolates (Table 1).

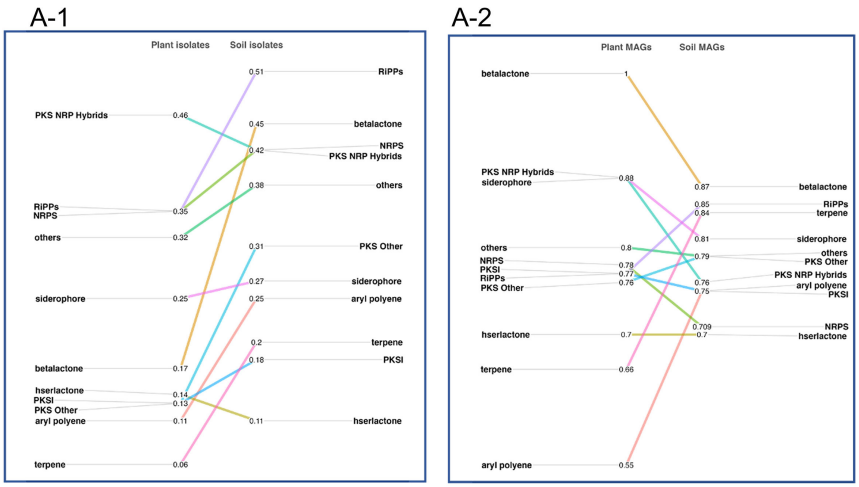
Terpene and aryl polyenes display stronger phylogenetic signals in phytobiomes

Noticeably, among the eleven BGC classes tested, the terpenes and aryl polyene classes had the strongest phylogenetic signals (*D* = 0.06 and 0.11, respectively; P_{random} <0.05) in the plant isolates. This pattern of terpene and aryl polyene classes with the strongest phylogenetic signal was also present in plant MAGs (*D* = 0.66 and 0.55, respectively; P_{random} <0.05) but not in the soil isolates and MAGs (Fig. 3A). Visualization of the distribution of strongly (terpene and aryl polyene BGC classes) and weakly conserved BGC classes (PKS-NRP hybrids and beta-lactone) in the phylogenetic tree of both plant isolates and plant MAGs further confirmed our claim (Fig. 3B and Fig. S2A for plant isolates and plant MAGs, respectively). These results support that terpene and aryl polyene BGCs are strongly phylogenetically conserved among phytobiomes. Between these two, terpene BGCs were found to be more predominant among phytobiomes (74.6% and 34.5% of the members in plant isolates and 29.3% and 24.9% of the members in plant MAGs possess terpene and aryl polyene biosynthetic capacity, respectively).

Terpene-related sequences are phylogenetically clustered and mostly uncharacterized in phytobiomes

To improve our understanding of the phylogenetic conservation of terpenes, we further studied (i) the relationships among terpene-related sequences (GCFs) from different bacterial classes and (ii) the functional potential of these sequences. The sequence similarity network of terpene-related sequences in the plant isolates and plant MAGs revealed that only a few of terpene BGCs (12 out of 180 and 1 out of 44 in plant isolates and plant MAGs, respectively) had reference annotations (from the MIBiG database). The largest networks, especially those of Alphaproteobacteria, Bacilli, and Bacteroidia, did not harbor any of the previously well-characterized terpene BGCs (Fig. 4A). None of the clusters from Bacilli and Bacteroidia contain any previously characterized terpene BGCs, as further described below. Hence, members of these classes have the functional potential to synthesize novel terpenoids. Interestingly, in the well-studied Actinomycetia, previously characterized terpene-related sequences were indeed present in the largest clusters but not in smaller clusters (Fig. 4A). On the other hand, the terpene BGCs from plant MAGs were mostly singletons reflecting the possibility of diverse and novel molecules being encoded by these genomes (Fig. S2B).

This widespread occurrence of uncharacterized terpene clusters prompted us to investigate the domains of their core genes to understand their relationship with each other and their distribution in the different microbial clades. As expected, squalene/phytoene synthase (SQS-PSY), which is involved in the first step of tri- and tetraterpene biosynthesis, was the most widely distributed (~83% of members) (Fig. 4B). Interestingly, SQS-PSY is absent in specific clades of Bacilli (e.g., *B. altitudinis*, *B. safensis*, and *B. cereus*, among others). Bacilli, in general, did not possess carotenoid BGCs, while the majority possessed Squalene Hopene Cyclase (SHC) clusters. The well-characterized Actinomycetia (*Streptomyces* spp.) harbored all core terpene domains, indicating their potential to produce diverse types of tri- and tetraterpenes. Both sequence similarity network and domain analyses suggested that Alphaproteobacteria potentially can produce distinct tri- and/or tetra-terpenes via their unique SQS-PSY domains (Fig. 4A and B; Fig. S3). In summary, based on sequence similarity networks and the domain distribution of terpene



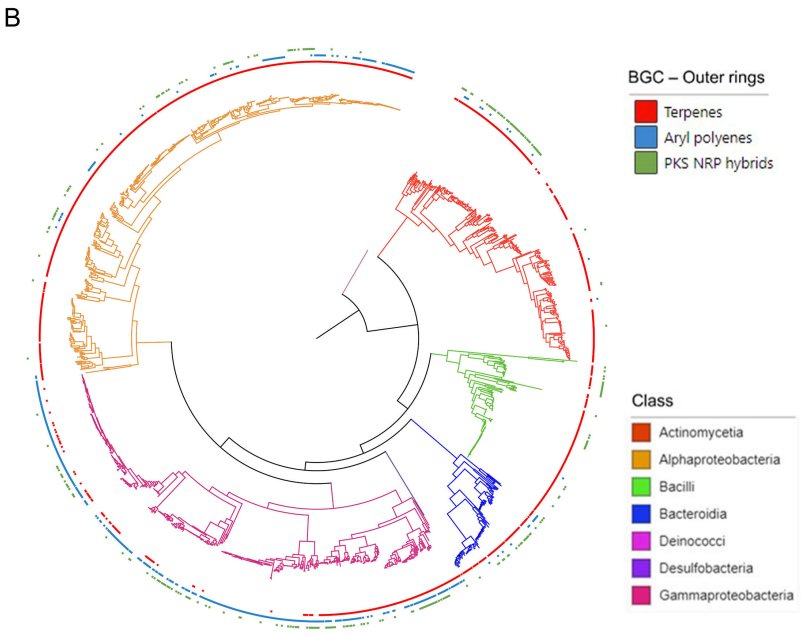


FIG 3 Phylogenetic conservation of Biosynthetic gene clusters. (A) Slope chart displaying the strength of phylogenetic conservation in isolates (A-1) and MAGs (A-2) data sets. A lower *D*-value indicates stronger phylogenetic conservation. The D-statistics were statistically significant (P_{random} <0.05 and $P_{Brownian}$ <0.05) for all BGC classes, except for beta-lactone in soil isolates and plant MAGs data sets. (B) Phylogenetic tree showing presence/absence pattern of terpenes, aryl polyenes, and PKS-NRP hybrids in plant isolates genomes. Stronger phylogenetic conservation of terpenes and aryl polyenes compared to PKS-NRP hybrids are shown by their three respective outer rings (red, blue, and green, respectively). The taxonomic classes of the plant isolates genomes are shown in different colors of the phylogenetic tree branches.

clusters, it is evident that phytobiomes are rich in potentially novel biosynthetic genes in previously understudied clades.

DISCUSSION

We investigated the relationship between microbial phylogeny and the biosynthetic potential of specialized metabolites in a worldwide collection of isolates and MAGs from diverse plant hosts. This is the first attempt to investigate the biosynthetic potential of phytobiomes using such global-scale resources, expanding on previous reports based on

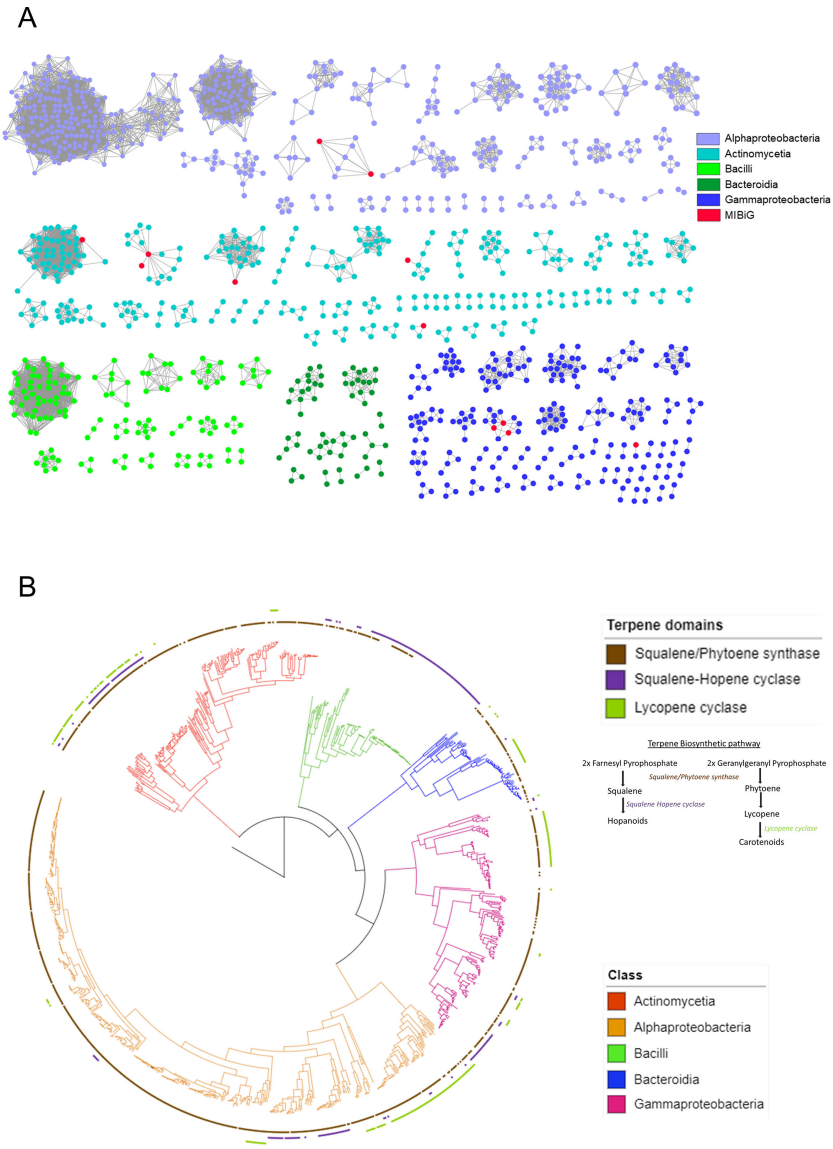


FIG 4 Sequence similarity network and domain distribution of terpenes in plant isolates. (A) Sequence similarity network of terpenes GCFs found in the plant isolates genomes (singletons are not shown). Each node represents individual terpene BGC, colored according to taxonomy. Nodes with existing reference (MIBiG) are highlighted in red. (B) Phylogenetic tree showing presence/absence pattern of terpene biosynthetic domains in plant isolates genomes. Distribution patterns of Squalene/Phytoene synthase, Squalene-hopene cyclase, and Lycopene cyclase are shown in separate colors of the outer rings (brown, violet, and green, respectively). The taxonomic level (class) of plant isolates genomes is shown in different branch colors of the phylogenetic tree. The terpene biosynthetic pathway for hopanoid and carotenoid biosynthesis is drawn with major enzyme involved in it.

either MAGs or isolates from specific plant hosts or microbial taxa (32, 53–55). This resource would help in generating experimentally verifiable hypotheses on how specific microbial groups can contribute to chemical ecology in different niches and environments. The framework for generating such hypotheses is reported here, and this resource could be used to reliably reveal conserved patterns of BGCs in both isolates and MAGs based on the choice of retaining only high-quality and nonredundant genomes. However, owing to the inherent differences in genomic characteristics and in BGC contents between isolates and MAGs, we recommend separate use of the two data sets.

The overall approach used here could be extended to other ecologically relevant microbial traits, such as antibiotic resistance (ARDB) (56), biogeochemical cycling of nutrients (57), and those involving carbohydrate active enzymes (CAZymes) (58).

Plants and microbes have co-evolved for millennia, which has led to their traits being highly influenced by both plant–microbe and microbe–microbe interactions (59, 60). The differential phylogenetic signal strength of several BGC classes in the phytobiomes shown in this study is suggestive of a strong plant host selection pressure in shaping the evolution of these traits in phytobiomes. Since selection pressures act distinctly on different BGC classes (13, 61), the evolutionary processes leading to differences in their strength of conservation under host influence may be somewhat different and independent. The well-recognized role of specialized metabolites in microbe–microbe communications, signaling, and antagonism also suggests that microbe–microbe and plant–microbe interaction components together contribute to the complex eco-evolution of BGCs in phytobiomes (62, 63).

Considering BGCs as microbial traits, this approach presents an opportunity to measure and compare their complexity with that of other microbial traits. The genetic depth of conservation of BGCs (τ_D = 0.022–0.053) is equivalent to that of other complex microbial traits, such as methanogenesis (τ_D = 0.042) and sulfate reduction (τ_D = 0.039) (64). Since trait complexity is correlated with the number of genes underlying a particular trait (65), the highly diverse components of metabolic machinery, such as enzymes, transporters, regulators, and accessory proteins, together contribute to establish BGCs as a complex microbial trait (66).

The phylogenetic signal strength of microbial traits provides clues to the mode of their inheritance (vertical/horizontal inheritance) among community members in an ecosystem. Terpene and aryl polyene BGC classes displayed stronger conservation than did other BGC classes in the phytobiomes, suggesting that vertical inheritance has a strong influence on their distribution. However, we did not specifically investigate this in this work but provide testable hypothesis for such patterns of inheritance among microbes. One explanation of their similar strength of conservation could be the result of their functionally convergent roles in antioxidative processes, which is imperative for bacterial colonization of plant roots (19, 67, 68). The strong conservation of terpenes in phytobiomes coincides with the phylogenetic conservation of terpenes in members of the plant kingdom (69). Furthermore, a recent report suggested that members of different bacterial clades respond differently toward terpenes secreted from plant roots (5). Taken together, these findings will provide important clues to the eco-evolutionary trajectories of terpenes in this "holobionts" entity (70).

Phylogenomic approaches can facilitate discoveries of molecules with potentially novel biochemical properties through rationally guided hypothesis generation (13). Here, we pinpoint such microbial clades to search for potentially novel terpene classes. First, Bacteroidia and Bacilli, the major abundant taxa in phytobiomes, harbor uncharacterized terpenes, directing future studies to focus on these clades to search for functionally novel terpenes. Second, specific members of Alphaproteobacteria form a distinct lineage of SQS/PSY domains, indicating that these domain-containing enzymes could possess novel biochemical properties. We also noted an instance of clade-specific biochemical adaptations that could shed light on the evolutionary mechanisms of terpenes in phytobiomes. For instance, the absence of SQS/PSY but the presence of SHC in specific clades of Bacilli could suggest that throughout the course of coevolution with plants, they either have evolved to take up squalene from other phytobiome members or have lost the capacity to produce hopanoids.

In conclusion, this study presents a novel approach through which chemical and microbial ecology is combined via a phylogenomics framework. One potential application of this approach, though investigated little here, is in understanding the coevolution of hosts and microbiomes through specialized metabolites. This can be further investigated in subsequent studies. The second application, as demonstrated here with terpenes, is in rationally guiding the discovery of novel metabolites classes that can

potentially be produced by distinct clades of host-associated microbiomes. This could be particularly helpful in determining the roles of widespread cryptic metabolic pathways for natural product discovery.

ACKNOWLEDGMENTS

The project is supported by the National Research Foundation, Prime Minister's Office, Singapore, with Grant ID NRF-CRP16-2015-04, titled as "Novel integrated agrotechnologies, plant nutrients and microbials for improved production of green leafy vegetables in Singapore." A.M. is a recipient of SCELSE-NUS Graduate fellowship. The authors thank the authors whose data contributed to this analysis. The authors also acknowledge the use of ChatGPT for grammatical corrections in this manuscript.

A.M. and H.T. contributed equally to the overall manuscript. The author orders were determined alphabetically. A.M. contributed to conceiving, phylogenomics framework development, statistical analyses, manuscript drafting, and review. H.T. contributed to conceiving, network analyses, biosynthetic gene clusters analyses, manuscript drafting, and review. A.B. provided inputs for statistical methods. S.P. contributed to developing framework and reviewing the manuscript. S.S. provided inputs for conceptual development, manuscript drafting, and review. All authors read and approved the final manuscript.

The authors declare no competing interests.

AUTHOR AFFILIATIONS

[1]Department of Biological Sciences, National University of Singapore, Singapore, Singapore
[2]Singapore Centre for Environmental Life Sciences Engineering, National University of Singapore, Singapore, Singapore
[3]NUS Environmental Research Institute, National University of Singapore, Singapore, Singapore

AUTHOR ORCIDs

Arijit Mukherjee http://orcid.org/0000-0003-3329-3535

FUNDING

Funder	Grant(s)	Author(s)
National Research Foundation Singapore (NRF)	NRF-CRP16-2015-04	Sanjay Swarup

AUTHOR CONTRIBUTIONS

Arijit Mukherjee, Conceptualization, Data curation, Formal analysis, Methodology, Visualization, Writing – original draft, Writing – review and editing | Hitesh Tikariha, Conceptualization, Data curation, Formal analysis, Methodology, Software, Visualization, Writing – original draft, Writing – review and editing | Aditya Bandla, Methodology | Shruti Pavagadhi, Project administration, Writing – review and editing | Sanjay Swarup, Conceptualization, Funding acquisition, Supervision, Writing – original draft, Writing – review and editing

DATA AVAILABILITY STATEMENT

Data used in the study were collected from public repositories and can be retrieved by using the information provided in the Materials and Methods. All codes to reproduce the analysis and generate figures of this manuscript have been deposited in GitHub. Software versions and nondefault parameters used have been appropriately specified where required.

ADDITIONAL FILES

The following material is available online.

Supplemental Material

Supplemental figures (mSystems00387-23-S0001.pdf). Figures S1 to S3.
Supplemental tables (mSystems00387-23-S0002.xlsx). Tables S1 to S19.

Open Peer Review

PEER REVIEW HISTORY (review-history.pdf). An accounting of the reviewer comments and feedback.

REFERENCES

1. Turner TR, James EK, Poole PS. 2013. The plant microbiome. Genome Biol 14:209. https://doi.org/10.1186/gb-2013-14-6-209
2. Vandenkoornhuyse P, Quaiser A, Duhamel M, Le Van A, Dufresne A. 2015. The importance of the microbiome of the plant holobiont. New Phytol 206:1196–1206. https://doi.org/10.1111/nph.13312
3. Trivedi P, Leach JE, Tringe SG, Sa T, Singh BK. 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. Blair PM, Land ML, Piatek MJ, Jacobson DA, Lu T-YS, Doktycz MJ, Pelletier DA. 2018. Exploration of the biosynthetic potential of the populus microbiome. mSystems 3:1–17. https://doi.org/10.1128/mSystems.00045-18
5. Huang AC, Jiang T, Liu Y-X, Bai Y-C, Reed J, Qu B, Goossens A, Nützmann H-W, Bai Y, Osbourn A. 2019. A specialized metabolic network selectively modulates *Arabidopsis* root microbiota. Science 364:eaau6389. https://doi.org/10.1126/science.aau6389
6. Chen Q, Jiang T, Liu YX, Liu H, Zhao T, Liu Z, Gan X, Hallab A, Wang X, He J, Ma Y, Zhang F, Jin T, Schranz ME, Wang Y, Bai Y, Wang G. 2019. Recently duplicated sesterterpene (C25) gene clusters in *Arabidopsis thaliana* modulate root microbiota. Sci. China Life Sci 62:947–958. https://doi.org/10.1007/s11427-019-9521-2
7. Cotton TEA, Pétriacq P, Cameron DD, Meselmani MA, Schwarzenbacher R, Rolfe SA, Ton J. 2019. Metabolic regulation of the maize rhizobiome by benzoxazinoids. ISME J 13:1647–1658. https://doi.org/10.1038/s41396-019-0375-2
8. Voges MJEEE, Bai Y, Schulze-Lefert P, Sattely ES. 2019. Plant-derived coumarins shape the composition of an *Arabidopsis* synthetic root microbiome. Proc Natl Acad Sci U S A 116:12558–12565. https://doi.org/10.1073/pnas.1820691116
9. Shakir S, Zaidi SS-E-A, de Vries FT, Mansoor S. 2021. Plant genetic networks shaping phyllosphere microbial community. Trends Genet 37:306–316. https://doi.org/10.1016/j.tig.2020.09.010
10. Niu DD, Liu HX, Jiang CH, Wang YP, Wang QY, Jin HL, Guo JH. 2011. The plant growth-promoting rhizobacterium *Bacillus* cereus Ar156 induces systemic resistance in *Arabidopsis thaliana* by simultaneously activating salicylate- and jasmonate/ethylene-dependent signaling pathways. Mol Plant Microbe Interact 24:533–542. https://doi.org/10.1094/MPMI-09-10-0213
11. Aghdam SA, Brown AMV. 2021. Deep learning approaches for natural product discovery from plant endophytic microbiomes. Environ Microbiome 16:6. https://doi.org/10.1186/s40793-021-00375-0
12. Cimermancic P, Medema MH, Claesen J, Kurita K, Wieland Brown LC, Mavrommatis K, Pati A, Godfrey PA, Koehrsen M, Clardy J, Birren BW, Takano E, Sali A, Linington RG, Fischbach MA. 2014. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. Cell 158:412–421. https://doi.org/10.1016/j.cell.2014.06.034
13. Chevrette MG, Gutiérrez-García K, Selem-Mojica N, Aguilar-Martínez C, Yañez-Olvera A, Ramos-Aboites HE, Hoskisson PA, Barona-Gómez F. 2020. Evolutionary dynamics of natural product biosynthesis in bacteria. Nat Prod Rep 37:566–599. https://doi.org/10.1039/c9np00048h
14. Challis GL. 2008. Genome mining for novel natural product discovery. J Med Chem 51:2618–2628. https://doi.org/10.1021/jm700948z
15. Kautsar SA, Blin K, Shaw S, Navarro-Muñoz JC, Terlouw BR, van der Hooft JJJ, van Santen JA, Tracanna V, Suarez Duran HG, Pascal Andreu V, Selem-Mojica N, Alanjary M, Robinson SL, Lund G, Epstein SC, Sisto AC, Charkoudian LK, Collemare J, Linington RG, Weber T, Medema MH. 2020. Mibig 2.0: a repository for biosynthetic gene clusters of known function. Nucleic Acids Research. https://doi.org/10.1093/nar/gkz882
16. Blin K, Shaw S, Steinke K, Villebro R, Ziemert N, Lee SY, Medema MH, Weber T. 2019. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. Nucleic Acids Res 47:W81–W87. https://doi.org/10.1093/nar/gkz310
17. Tyc O, Song C, Dickschat JS, Vos M, Garbeva P. 2017. The ecological role of volatile and soluble secondary metabolites produced by soil bacteria. Trends Microbiol 25:280–292. https://doi.org/10.1016/j.tim.2016.12.002
18. Avalos M, Garbeva P, Vader L, van Wezel GP, Dickschat JS, Ulanova D. 2022. Biosynthesis, evolution and ecology of microbial terpenoids. Nat Prod Rep 39:249–272. https://doi.org/10.1039/d1np00047k
19. Schöner TA, Gassel S, Osawa A, Tobias NJ, Okuno Y, Sakakibara Y, Shindo K, Sandmann G, Bode HB. 2016. Aryl Polyenes, a highly abundant class of bacterial natural products, are functionally related to antioxidative carotenoids. Chembiochem 17:247–253. https://doi.org/10.1002/cbic.201500474
20. Martiny JBH, Jones SE, Lennon JT, Martiny AC. 2015. Microbiomes in light of traits: a phylogenetic perspective. Science 350:aac9323. https://doi.org/10.1126/science.aac9323
21. Fritz SA, Purvis A. 2010. Selectivity in mammalian extinction risk and threat types: a new measure of phylogenetic signal strength in binary traits. Conserv Biol 24:1042–1051. https://doi.org/10.1111/j.1523-1739.2010.01455.x
22. Martiny AC, Treseder K, Pusch G. 2013. Phylogenetic conservatism of functional traits in microorganisms. ISME J 7:830–838. https://doi.org/10.1038/ismej.2012.160
23. Van Assche A, Álvarez-Pérez S, de Breij A, De Brabanter J, Willems KA, Dijkshoorn L, Lievens B. 2017. Phylogenetic signal in phenotypic traits related to carbon source assimilation and chemical sensitivity in *Acinetobacter* species. Appl Microbiol Biotechnol 101:367–379. https://doi.org/10.1007/s00253-016-7866-0
24. Isobe K, Allison SD, Khalili B, Martiny AC, Martiny JBH. 2019. Phylogenetic conservation of bacterial responses to soil nitrogen addition across continents. Nat Commun 10:2499. https://doi.org/10.1038/s41467-019-10390-y
25. Wattam AR, Davis JJ, Assaf R, Boisvert S, Brettin T, Bun C, Conrad N, Dietrich EM, Disz T, Gabbard JL, Gerdes S, Henry CS, Kenyon RW, Machi D, Mao C, Nordberg EK, Olsen GJ, Murphy-Olson DE, Olson R, Overbeek R, Parrello B, Pusch GD, Shukla M, Vonstein V, Warren A, Xia F, Yoo H, Stevens RL. 2017. Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center. Nucleic Acids Res 45:D535–D542. https://doi.org/10.1093/nar/gkw1017
26. Lee M 2023. bit : a multipurpose collection of bioinformatics tools [version 1; peer review : 2 not approved] 1–7.
27. Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: Quality assessment tool for genome assemblies. Bioinformatics 29:1072–1075. https://doi.org/10.1093/bioinformatics/btt086

28. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 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
29. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. 2019. GTDB-TK: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics 36:1925–1927. https://doi.org/10.1093/bioinformatics/btz848
30. Olm MR, Crits-Christoph A, Bouma-Gregson K, Firek BA, Morowitz MJ, Banfield JF. 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
31. Olm MR, Brown CT, Brooks B, Banfield JF. 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
32. Nayfach S, Roux S, Seshadri R, Udwary D, Varghese N, Schulz F, Wu D, Paez-Espino D, Chen I-M, Huntemann M, Palaniappan K, Ladau J, Mukherjee S, Reddy TBK, Nielsen T, Kirton E, Faria JP, Edirisinghe JN, Henry CS, Jungbluth SP, Chivian D, Dehal P, Wood-Charlson EM, Arkin AP, Tringe SG, Visel A, IMG/M Data Consortium, Abreu H, Acinas SG, Allen E, Allen MA, Alteio LV, Andersen G, Anesio AM, Attwood G, Avila-Magaña V, Badis Y, Bailey J, Baker B, Baldrian P, Barton HA, Beck DAC, Becraft ED, Beller HR, Beman JM, Bernier-Latmani R, Berry TD, Bertagnolli A, Bertilsson S, Bhatnagar JM, Bird JT, Blanchard JL, Blumer-Schuette SE, Bohannan B, Borton MA, Brady A, Brawley SH, Brodie J, Brown S, Brum JR, Brune A, Bryant DA, Buchan A, Buckley DH, Buongiorno J, Cadillo-Quiroz H, Caffrey SM, Campbell AN, Campbell B, Carr S, Carroll J, Cary SC, Cates AM, Cattolico RA, Cavicchioli R, Chistoserdova L, Coleman ML, Constant P, Conway JM, Mac Cormack WP, Crowe S, Crump B, Currie C, Daly R, DeAngelis KM, Denef V, Denman SE, Desta A, Dionisi H, Dodsworth J, Dombrowski N, Donohue T, Dopson M, Driscoll T, Dunfield P, Dupont CL, Dynarski KA, Edgcomb V, Edwards EA, Elshahed MS, Figueroa I, Flood B, Fortney N, Fortunato CS, Francis C, Gachon CMM, Garcia SL, Gazitua MC, Gentry T, Gerwick L, Gharechahi J, Girguis P, Gladden J, Gradoville M, Grasby SE, Gravuer K, Grettenberger CL, Gruninger RJ, Guo J, Habteselassie MY, Hallam SJ, Hatzenpichler R, Hausmann B, Hazen TC, Hedlund B, Henny C, Herfort L, Hernandez M, Hershey OS, Hess M, Hollister EB, Hug LA, Hunt D, Jansson J, Jarett J, Kadnikov VV, Kelly C, Kelly R, Kelly W, Kerfeld CA, Kimbrel J, Klassen JL, Konstantinidis KT, Lee LL, Li W-J, Loder AJ, Loy A, Lozada M, MacGregor B, Magnabosco C, Maria da Silva A, McKay RM, McMahon K, McSweeney CS, Medina M, Meredith L, Mizzi J, Mock T, Momper L, Moran MA, Morgan-Lang C, Moser D, Muyzer G, Myrold D, Nash M, Nesbø CL, Neumann AP, Neumann RB, Noguera D, Northen T, Norton J, Nowinski B, Nüsslein K, O'Malley MA, Oliveira RS, Maia de Oliveira V, Onstott T, Osvatic J, Ouyang Y, Pachiadaki M, Parnell J, Partida-Martinez LP, Peay KG, Pelletier D, Peng X, Pester M, Pett-Ridge J, Peura S, Pjevac P, Plominsky AM, Poehlein A, Pope PB, Ravin N, Redmond MC, Reiss R, Rich V, Rinke C, Rodrigues JLM, Rodriguez-Reillo W, Rossmassler K, Sackett J, Salekdeh GH, Saleska S, Scarborough M, Schachtman D, Schadt CW, Schrenk M, Sczyrba A, Sengupta A, Setubal JC, Shade A, Sharp C, Sherman DH, Shubenkova OV, Sierra-Garcia IN, Simister R, Simon H, Sjöling S, Slonczewski J, Correa de Souza RS, Spear JR, Stegen JC, Stepanauskas R, Stewart F, Suen G, Sullivan M, Sumner D, Swan BK, Swingley W, Tarn J, Taylor GT, Teeling H, Tekere M, Teske A, Thomas T, Thrash C, Tiedje J, Ting CS, Tully B, Tyson G, Ulloa O, Valentine DL, Van Goethem MW, VanderGheynst J, Verbeke TJ, Vollmers J, Vuillemin A, Waldo NB, Walsh DA, Weimer BC, Whitman T, van der Wielen P, Wilkins M, Williams TJ, Woodcroft B, Woolet J, Wrighton K, Ye J, Young EB, Youssef NH, Yu FB, Zemskaya TI, Ziels R, Woyke T, Mouncey NJ, Ivanova NN, Kyrpides NC, Eloe-Fadrosh EA. 2021. A genomic catalog of earth's microbiomes. Nat Biotechnol 39:499–509. https://doi.org/10.1038/s41587-020-0718-6
33. Bandla A, Pavagadhi S, Sridhar Sudarshan A, Poh MCH, Swarup S. 2020. 910 Metagenome-assembled genomes from the phytobiomes of three urban-farmed leafy Asian greens. Sci Data 7:278. https://doi.org/10.1038/s41597-020-00617-9
34. Blin K, Shaw S, Steinke K, Villebro R, Ziemert N, Lee SY, Medema MH, Weber T. 2019. Antismash 5.0: updates to the secondary metabolite genome mining pipeline. Nucleic Acids Res 47:W81–W87. https://doi.org/10.1093/nar/gkz310
35. Navarro-Muñoz JC, Selem-Mojica N, Mullowney MW, Kautsar SA, Tryon JH, Parkinson EI, De Los Santos ELC, Yeong M, Cruz-Morales P, Abubucker S, Roeters A, Lokhorst W, Fernandez-Guerra A, Cappelini LTD, Goering AW, Thomson RJ, Metcalf WW, Kelleher NL, Barona-Gomez F, Medema MH. 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
36. Price MN, Dehal PS, Arkin AP. 2010. Fasttree 2 - approximately maximum-likelihood trees for large alignments. PLoS One 5:e9490. https://doi.org/10.1371/journal.pone.0009490
37. Criscuolo A, Gribaldo S. 2010. BMGE (block mapping and gathering with entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol Biol 10:210. https://doi.org/10.1186/1471-2148-10-210
38. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol 37:1530–1534. https://doi.org/10.1093/molbev/msaa131
39. Letunic I, Bork P. 2019. Interactive tree of life (iTOL) V4: recent updates and new developments. Nucleic Acids Res 47:W256–W259. https://doi.org/10.1093/nar/gkz239
40. Ricotta C, Podani J. 2017. On some properties of the Bray-Curtis dissimilarity and their ecological meaning. Ecol Complex31:201–205. https://doi.org/10.1016/j.ecocom.2017.07.003
41. Dray S, Dufour AB. 2007. The Ade4 package: implementing the duality diagram for ecologists . J. Stat. Soft 22:1–20. https://doi.org/10.18637/jss.v022.i04
42. Dixon P. 2003. Computer program review VEGAN, a package of R functions for community ecology. J Veg Sci14:927–930. https://doi.org/10.1111/j.1654-1103.2003.tb02228.x
43. Hartig F. 2016. Dharma: residual diagnostics for hierarchical (multi-level/mixed) regression models
44. Pagel M. 1999. Inferring the historical patterns of biological evolution. Nature 401:877–884. https://doi.org/10.1038/44766
45. Orme CDL. 2012. The caper package: comparative analyses in phylogenetics and evolution in R. Available from: See http://caper.r-forge.r-project.org/. Http://CaperR-ForgeR-ProjectOrg/ 1–36
46. Harmon LJ, Weir JT, Brock CD, Glor RE, Challenger W. 2008. GEIGER: investigating evolutionary radiations. Bioinformatics 24:129–131. https://doi.org/10.1093/bioinformatics/btm538
47. Saito R, Smoot ME, Ono K, Ruscheinski J, Wang PL, Lotia S, Pico AR, Bader GD, Ideker T. 2012. A travel guide to cytoscape plugins. Nat Methods 9:1069–1076. https://doi.org/10.1038/nmeth.2212
48. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780. https://doi.org/10.1093/molbev/mst010
49. Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. 2019. Raxml-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics 35:4453–4455. https://doi.org/10.1093/bioinformatics/btz305
50. Levy A, Salas Gonzalez I, Mittelviefhaus M, Clingenpeel S, Herrera Paredes S, Miao J, Wang K, Devescovi G, Stillman K, Monteiro F, Rangel Alvarez B, Lundberg DS, Lu T-Y, Lebeis S, Jin Z, McDonald M, Klein AP, Feltcher ME, Rio TG, Grant SR, Doty SL, Ley RE, Zhao B, Venturi V, Pelletier DA, Vorholt JA, Tringe SG, Woyke T, Dangl JL. 2017. Genomic features of bacterial adaptation to plants. Nat Genet 50:138–150. https://doi.org/10.1038/s41588-017-0012-9
51. Dastogeer KMG, Tumpa FH, Sultana A, Akter MA, Chakraborty A. 2020. Plant Microbiome–an account of the factors that shape community composition and diversity. Current Plant Biology 23:100161. https://doi.org/10.1016/j.cpb.2020.100161
52. Malik AA, Martiny JBH, Brodie EL, Martiny AC, Treseder KK, Allison SD. 2020. Defining trait-based microbial strategies with consequences for soil carbon cycling under climate change. ISME J 14:1–9. https://doi.org/10.1038/s41396-019-0510-0
53. Adamek M, Alanjary M, Sales-Ortells H, Goodfellow M, Bull AT, Winkler A, Wibberg D, Kalinowski J, Ziemert N. 2018. Comparative genomics reveals phylogenetic distribution patterns of secondary metabolites in *Amycolatopsis* species. BMC Genomics 19:426. https://doi.org/10.1186/s12864-018-4809-4

54. Chevrette MG, Carlos-Shanley C, Louie KB, Bowen BP, Northen TR, Currie CR. 2019. Taxonomic and metabolic Incongruence in the ancient genus *Streptomyces*. Front Microbiol 10:2170. https://doi.org/10.3389/fmicb.2019.02170
55. Helfrich EJN, Vogel CM, Ueoka R, Schäfer M, Ryffel F, Müller DB, Probst S, Kreuzer M, Piel J, Vorholt JA. 2018. Bipartite interactions, antibiotic production and biosynthetic potential of the *Arabidopsis* leaf microbiome. Nat Microbiol 3:909–919. https://doi.org/10.1038/s41564-018-0200-0
56. Liu B, Pop M. 2009. ARDB - antibiotic resistance genes database. Nucleic Acids Res 37:D443–7. https://doi.org/10.1093/nar/gkn656
57. Anantharaman K, Brown CT, Hug LA, Sharon I, Castelle CJ, Probst AJ, Thomas BC, Singh A, Wilkins MJ, Karaoz U, Brodie EL, Williams KH, Hubbard SS, Banfield JF. 2016. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. Nat Commun 7:13219. https://doi.org/10.1038/ncomms13219
58. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. 2014. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res 42:D490–D495. https://doi.org/10.1093/nar/gkt1178
59. Cordovez V, Dini-Andreote F, Carrión VJ, Raaijmakers JM. 2019. Ecology and evolution of plant microbiomes. Annu Rev Microbiol 73:69–88. https://doi.org/10.1146/annurev-micro-090817-062524
60. Tan J, Kerstetter JE, Turcotte MM. 2021. Eco-evolutionary interaction between Microbiome presence and rapid Biofilm evolution determines plant host fitness. Nat Ecol Evol 5:670–676. https://doi.org/10.1038/s41559-021-01406-2
61. Firn R. 2009. Nature's chemicals. Oxford University Press. Available from: https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780199566839.001.0001/acprof-9780199566839
62. Medema MH. 2018. Computational genomics of specialized metabolism: from natural product discovery to microbiome ecology. mSystems 3:1–5. https://doi.org/10.1128/mSystems.00182-17
63. Van Goethem MW, Osborn AR, Bowen BP, Andeer PF, Swenson TL, Clum A, Riley R, He G, Koriabine M, Sandor L, Yan M, Daum CG, Yoshinaga Y, Makhalanyane TP, Garcia-Pichel F, Visel A, Pennacchio LA, O'Malley RC, Northen TR. 2021. Long-read metagenomics of soil communities reveals phylum-specific secondary metabolite dynamics. Commun Biol 4:1302. https://doi.org/10.1038/s42003-021-02809-4
64. Martiny AC, Treseder K, Pusch G. 2013. Phylogenetic conservatism of functional traits in microorganisms. ISME J 7:830–838. https://doi.org/10.1038/ismej.2012.160
65. Amend AS, Martiny AC, Allison SD, Berlemont R, Goulden ML, Lu Y, Treseder KK, Weihe C, Martiny JBH. 2016. Microbial response to simulated global change is phylogenetically conserved and linked with functional potential. ISME J 10:109–118. https://doi.org/10.1038/ismej.2015.96
66. Borsetto C, Amos GCA, da Rocha UN, Mitchell AL, Finn RD, Laidi RF, Vallin C, Pearce DA, Newsham KK, Wellington EMH. 2019. Microbial community drivers of PK/NRP gene diversity in selected global soils. Microbiome 7:78. https://doi.org/10.1186/s40168-019-0692-8
67. Fones H, Preston GM. 2012. Reactive oxygen and oxidative stress tolerance in plant pathogenic pseudomonas. FEMS Microbiol Lett 327:1–8. https://doi.org/10.1111/j.1574-6968.2011.02449.x
68. Tzipilevich E, Russ D, Dangl JL, Benfey PN. 2021. Plant immune system activation is necessary for efficient root colonization by auxin-secreting beneficial bacteria. Cell Host Microbe 29:1507–1520. https://doi.org/10.1016/j.chom.2021.09.005
69. Zhang Y, Deng T, Sun L, Landis JB, Moore MJ, Wang H, Wang Y, Hao X, Chen J, Li S, Xu M, Puno P-T, Raven PH, Sun H. 2021. Phylogenetic patterns suggest frequent multiple origins of secondary metabolites across the seed-plant 'tree of life'. Natl Sci Rev 8. https://doi.org/10.1093/nsr/nwaa105
70. Berg G, Grube M, Schloter M, Smalla K. 2014. Unraveling the plant microbiome: Looking back and future perspectives. Front. Microbiol 5:1–7. https://doi.org/10.3389/fmicb.2014.00148