Received: 12 October 2025 | Accepted: 22 March 2026

DOI: 10.1002/aps3.70064

APPLICATION ARTICLE

Applications in Plant Sciences

Off-target metagenomics: Leveraging whole genome sequencing to study the bacteriome of the liverwort *Calasterella californica*

Ixchel S. González-Ramírez[1,2] | **Michael J. Song**[3] | **Elijah C. Mehlferber**[4] | **Brent D. Mishler**[2]

[1]Missouri Botanical Garden, St. Louis, Missouri, USA

[2]Department of Integrative Biology, University of California, Berkeley, California, USA

[3]Skyline College, San Bruno, California, USA

[4]Georgia Institute of Technology, Atlanta, Georgia, USA

Correspondence

Ixchel S. González-Ramírez, Department of Integrative Biology, University of California, Berkeley, California, USA.
Email: ixchel_gonzalezrmz@berkeley.edu

This article is part of the special issue "Branching out: Resolving plant evolution through phylogenetic networks."

Funding information

University of California, Grant/Award Number: UC Award ID RSI-19-690224; CONACHYT, Grant/Award Number: Doctoral fellowship 709967

Abstract

Premise: The recovery of non-target organism reads, especially when whole organisms are sampled, constitutes a great opportunity for studying microbial communities. The increase in whole genome sequencing feasibility and the development of new marker-based pipelines enable the use of short reads to study bacterial communities associated with organisms.

Methods: We utilized population genomic data of the liverwort *Calasterella californica* obtained through the California Conservation Genomics Project to characterize the composition of its associated bacterial communities and explore its variation across the geographic space.

Results: The bacterial communities associated with *C. californica* were dominated by the methanotroph *Methylobacterium* and other Hyphomicrobiales, a group that includes well-known plant symbionts. While diversity metrics of bacteria composition were similar across localities, we found significant differences in the relative abundance of a few taxa across California regions, likely driven by differences in precipitation and temperature seasonality.

Discussion: Our results support previous observations that liverwort bacterial communities are not randomly assembled, suggesting a potential role of the plant in determining community composition, an emerging pattern that deserves more attention. The novel off-target metagenomics approach can be applied to any population-level resequencing where whole organisms are sequenced, opening the door to exciting avenues of microbiome research using repurposed data from landscape genomics.

KEYWORDS

bryophyte, California Conservation Genomics Project (CCGP), liverwort, MetaPhlAn, *Methylobacterium*, microbiome

An important opportunity for leveraging existing sequencing data lies in the recovery of non-host reads that are inadvertently captured alongside the target organism during whole genome sequencing (WGS). The standard protocols used for WGS extraction do not distinguish between the DNA of the target organism and other organisms that might be present in the sample—such as parasitic or symbiont microorganisms—which are often considered contaminants (Steinegger and Salzberg, 2020). In larger hosts, whose extractions are usually performed on samples of internal organ tissue (Wood et al., 2022; Benham et al., 2023; Supple et al., 2025), we might expect a relatively low proportion of contaminant reads. Performing WGS on small organisms, like insects, small invertebrates, and small plants such as bryophytes, often involves using complete individual(s) as a sample, which increases the proportion of non-host reads

Ixchel S. Gonzalez-Ramirez and Michael J. Song contributed equally to this work.

that we might expect because the initial proportion of host tissue is smaller, and also because sequencing whole organisms includes environment-interfacing tissues. Given the growing appreciation for studying holobionts—i.e., an organism and all the associated organisms living within and around it (Vandenkoornhuyse et al., 2015)—we can leverage WGS data of whole organisms to provide insights into bacterial communities, symbiosis, and co-evolution between hosts and microorganisms (Song et al., 2024, 2025a), in an approach that we refer to as "off-target metagenomics."

To demonstrate the practicality of the off-target metagenomics approach, we repurposed data generated by the California Conservation Genomics Project (CCGP), originally intended to create genomic maps across the state to influence land management and conservation policy (Beninde et al., 2022; Fiedler et al., 2022; Shaffer et al., 2022; Toffelmier et al., 2022), to explore the composition and variability of the microbiome associated with a liverwort species. The CCGP collected and sequenced samples for hundreds of individuals of more than 100 species across California. Many of these focal taxa represent an ideal opportunity to analyze the non-target reads to study the composition of microbial communities and their potential variation across the geographic space, especially for those taxa where organisms were sequenced whole (Blair et al., 2022; Huang et al., 2022; Grether et al., 2023; Mead et al., 2024; González-Ramírez et al., 2025a; Song et al., 2025b). Off-target metagenomics is possible thanks to the recent development of a metagenomic profiling software, MetaPhlAn4 (Segata et al., 2012; Truong et al., 2015; Blanco-Míguez et al., 2023), that uses short reads from metagenomic samples to perform a taxonomic profiling based on "species-level genome bins" (SGBs). In short, rather than classifying all reads individually, this marker-based approach maps reads to a curated database of clade-specific markers and uses the read coverage of these markers to estimate taxon abundances (Segata et al., 2012). Using SGBs and the marker-based approach facilitates the profiling of undescribed microbial communities in comparison to other approaches like reference-based computational ones, which are limited by the number of reference genomes available, or metagenome-assembled genomes (MAGs), which require high coverage for all the taxa (Blanco-Míguez et al., 2023). Together, available WGS datasets for which we expect to observe high proportions of off-target DNA and analytical tools like MetaPhlAn allow for streamlined research of bacterial communities (Härer and Rennison, 2023).

In this study, we focused on the bacterial community of the liverwort *Calasterella californica* (Hampe ex Austin) D.G. Long & T.X. Zheng (Figure 1; González-Ramírez et al., 2025a). Liverworts are small plants that live in close proximity to the ground and have a high proportion of environment-interfacing tissue, making it an ideal system to apply the off-target metagenomics approach. Liverworts are known to have important associations with fungi (Bidartondo and Duckett, 2010), and in recent years, pioneering work on the bacterial component of liverwort microbiomes (e.g., Kutschera and Koopmann, 2005; Alcaraz et al., 2018; Wicaksono et al., 2023; Young et al., 2025) has pointed to potentially stronger bacteria–liverwort associations than previously thought, highlighting the need for further investigation. Furthermore, *C. californica*'s broad geographic and environmental distribution range—from wet redwood forests to hot and dry deserts (González-Ramírez et al., 2025b)—raises the question of whether the microbiome varies across this set of conditions. By applying the off-target metagenomics approach to the extensive WGS dataset of *C. californica* generated by the CCGP, we (1) characterize the composition of the bacterial community associated with a non-model species of liverwort, and (2) describe the variation in the composition of the bacteriome across a broad geographic and environmental range. We discuss the implications of our results in understanding liverwort–bacterial community dynamics, as well as a proof of concept of the ability of off-target metagenomics to fill the gap on the study of microbial communities associated with non-model organisms.

METHODS

Study system

The liverwort species *Calasterella californica* (Figure 1) is the only species in the recently described genus *Calasterella* D.G. Long & T.X. Zheng, in the family Aytoniaceae (Long and Zheng, 2023; González-Ramírez et al., 2025b). As a complex thalloid liverwort, *C. californica* possesses a dorsiventral thallus and scales and rhizoids on the ventral surface; it does not produce true leaves, stems, or roots. The thalli of *C. californica* are around 7 mm wide and less than a couple centimeters long, and often form tight circular clusters. The clusters are usually around 10 cm wide, but in ideal conditions they can extend for several meters. *Calasterella californica* plants live for multiple years, but during dry seasons the thalli are folded inside their ventral scales and therefore are inconspicuous. *Calasterella californica* is dioicous, a trait that easily differentiates it from other California liverworts when plants are reproductive; however, when the plants are not reproductive, differentiating them (from, e.g., *Reboulia hemisphaerica* (L.) Raddi or *Asterella bolanderi* (Austin) Underw.) often requires careful examination of the epidermis and ventral scales under a microscope.

Original sampling of *Calasterella californica*

The original sampling of *C. californica* was conducted as part of the CCGP (Shaffer et al., 2022), a large collaborative project that aimed to understand the genetic diversity of multiple California species. *Calasterella californica*'s sampling included 110 collections that represent the breadth of its geographic and environmental distribution range. The sampling activities primarily targeted localities with previous herbarium or

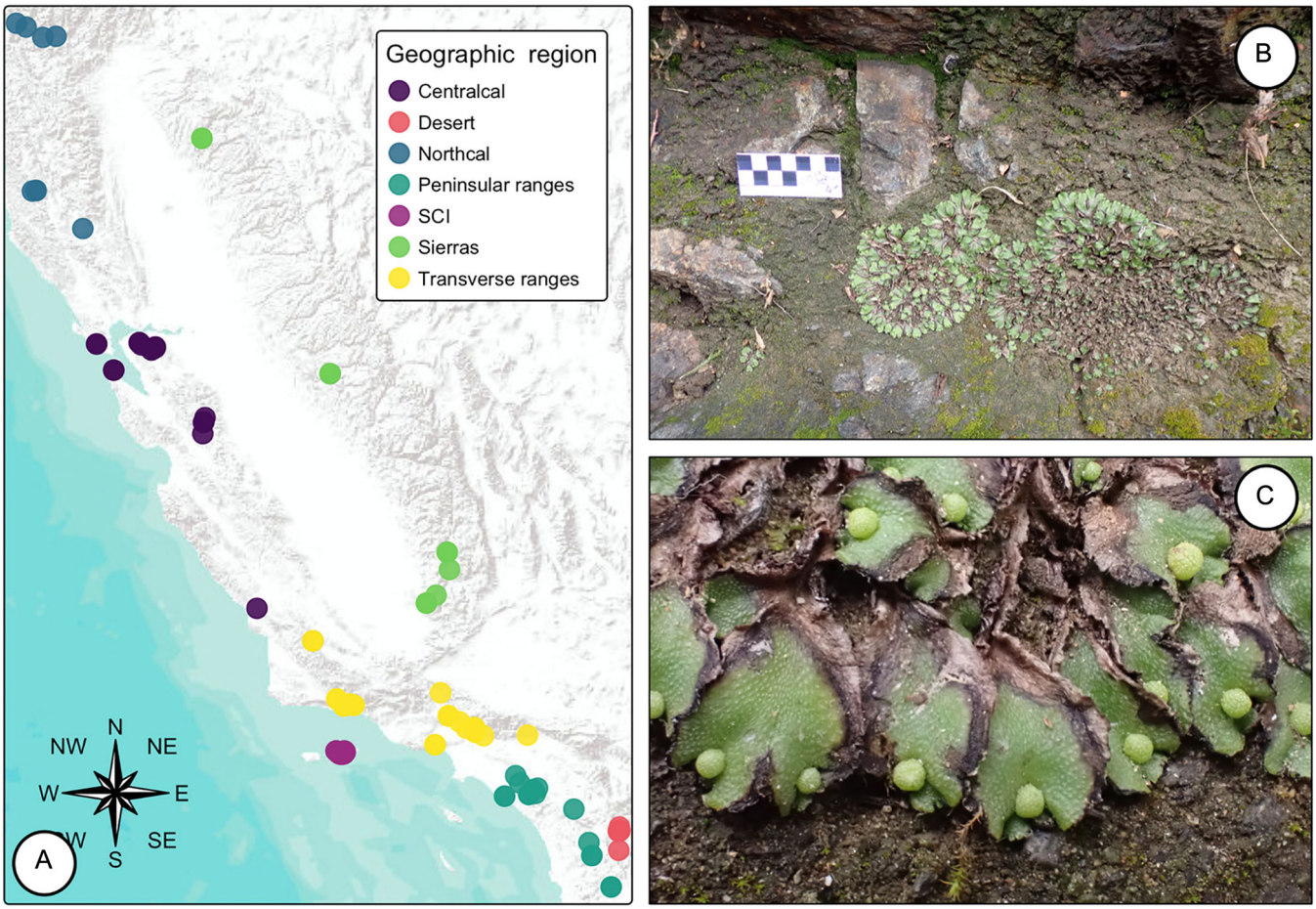

FIGURE 1 Distribution and habit of *Calasterella californica*. (A) Collection locations of *Calasterella californica* samples for this study; dots are color coded according to geographic region. (B) Typical morphology of *C. californica*, growing on bare and exposed soil or rocks and forming circular mats. (C) Close-up image of female thalli of *C. californica* displaying the proximity of this plant with the soil.

verifiable iNaturalist (https://www.inaturalist.org/) records, and National Forest land (González Ramírez, 2024). At each locality, one to two clumps of liverworts were collected by removing complete clusters of thalli together with the adjacent layer of soil with a knife, and they were placed in a hard plastic container (15.5 cm × 10 cm × 6 cm). The liverworts were kept alive in their container under direct sun until they were transported to the Mishler Bryolab at the University of California, Berkeley. Immediately upon arrival at the lab, the samples were processed. The processing consisted of (1) corroborating the field taxonomic determination of the samples that were not reproductive, (2) isolating and cleaning (by manually removing soil particles and most of the rhizoids under a Zeiss Stemi DV4 microscope [Zeiss, Oberkochen, Germany]) the largest 1–3 thalli of each collection for DNA extraction, (3) taking morphological measurements of thalli (relevant for the population genomics study), and (4) preparing a herbarium specimen. All the herbarium specimens were later deposited in the UC Berkeley herbarium collection. The number of thalli used for DNA extraction depended on the size of the thalli and aimed to secure a similar amount of tissue across samples. When more than one thallus was used, we selected thalli that, according to the growth pattern, appeared to be part of the same individual. The material isolated for DNA extraction was placed in paper coin envelopes with silica gel beads and stored at room temperature for up to 30 days until enough samples were accumulated for a 12-sample batch for extraction. Previous microbiome studies have shown that short-term storage conditions have a surprisingly small impact on bacterial diversity and community structure as long as the storage strategies are consistent across samples and that care is taken to limit the amount of freeze–thaw cycles (reviewed in Goodrich et al., 2014).

Leveraging samples for studying bacterial communities associated with *Calasterella californica*

We utilized 97 samples to characterize *C. californica*'s bacteriome across the geographic space. Of the original 110 samples, 13 samples were excluded from subsequent study: three non-reproductive samples that were found to be a different species of liverwort upon inspection under the microscope and

10 samples that were flagged as yielding very low sequencing depth during the population genomics work (González Ramírez, 2024, manuscript in prep.). In this study, we used geographic region as a proxy to characterize the environmental conditions that liverworts live in, as these regions largely reflect shared environmental conditions and geological history (geographic regions of individual samples are recorded in Table S4 in Appendix S1 [see Supporting Information]). Seven geographic regions were defined as follows. (1) The Northern California region is characterized by higher precipitation and cooler temperatures, with vegetation being predominantly conifer forests. (2) The Sierras region spans the interior mountain range of California (east of the Central Valley) and is characterized by a shared geological history, marked seasonal variation, and higher altitude. (3) The Central California region spans the coastal region (west of the Central Valley) from the Bay Area to the Transverse Ranges. This region is characterized by coastal scrublands (as opposed to more perennial conifer forests in Northern California) and comparatively less seasonal variation than samples from more inland regions. In Southern California, we differentiate between the (4) Transverse Ranges region, characterized by higher altitude, precipitation, and seasonality compared to the more coastal (5) Peninsular Ranges region. (6) The Santa Cruz Island region is very similar to the Central California region in terms of vegetation, but the strong oceanic influence decreases the seasonal variation in temperature and rainfall. Finally, (7) the desert region corresponds to the flat area east of the Transverse Ranges that belongs to the Sonoran Desert and is characterized by a dry and warm climate (Figure 1, Figure S3 in Appendix S1).

DNA extraction, sequencing, and quality checks

For each sample, total genomic and symbiont DNA was extracted using either the Qiagen DNeasy Plant Pro kit (Qiagen, Germantown, Maryland, USA) or a cetyltrimethylammonium bromide (CTAB) protocol. Library preparation and sequencing were carried out at the UC Davis DNA Technologies and Expression Analysis Core Laboratory. Libraries were sequenced using an Illumina NovaSeq X 300 (PE150; Illumina, San Diego, California, USA) flow cell. All the samples displayed a bimodal read GC-content distribution that characterizes the presence of both bacterial (high GC content) and liverwort (low GC content) DNA.

Read processing and bacteriome characterization using MetaPhlAn

All the paired-end files were quality checked and cleaned from adapters using Trimmomatic (Bolger et al., 2014). We applied a lenient filtering scheme focusing on removing adapters in order to keep as many reads as possible (ILLUMINACLIP:TruSeq3-PE.fa:2:30:10:2:True LEADING:3 TRAILING:3 MINLEN:36). This lenient filtering approach is congruent with that suggested by the MetaPhlAn developers (Segata et al., 2012). Next, for each sample, the paired-end reads were mapped to the *C. californica* IGR150-G1 reference assembly (González-Ramírez et al., 2025a) using BWA-MEM (Li, 2013). Using SAMtools (Li et al., 2009) with the flag -f 12, we kept only the reads whose mate was also unmapped to the liverwort reference genome. These unmapped reads constituted the input to study the microbiome. To characterize the *C. californica* metagenome composition, we used MetaPhlAn 4 (Blanco-Míguez et al., 2023), a recently developed pipeline that directly uses reads from shotgun sequence data to characterize microbial communities by comparing these reads to a database of SGBs. This pipeline outputs a list of the bacterial taxa present, as well as an estimate of the relative abundances of each bacterial taxon (Blanco-Míguez et al., 2023). We ran MetaPhlAn on the Savio computing cluster at UC Berkeley Research Computing using the most recent reference database of SGBs provided by the MetaPhlAn team of developers (mpa_vJan25_CHOCOPhlAnSGB_202503) and employing 20 cores simultaneously for each sample. The number of reads at each step of this workflow is detailed in Table S4 (Appendix S1).

Bacteriome diversity and composition

All the statistical analyses and visualizations were performed using the microeco package (Liu et al., 2026) in R (R Core Team, 2026). The output obtained from MetaPhlAn was formatted as input for microeco using the R package file2meco (Liu et al., 2022). First, we conducted a series of statistical tests to understand if the proportion of classified (or better referred to as "used") reads was affected by sample collection date, geographic region, sequencing batch, or collection date (detailed methods and results are presented in sections 1–3 of Appendix S1).

We quantified the α-diversity of each sample by computing multiple diversity indices (Shannon Diversity index [Shannon, 1948], Simpson Diversity index [Simpson, 1949], and Pielou's index [Pielou, 1966]) and coverage, as implemented in microeco. We tested the potential marginal effect of geographic region, collection month, and collection date on these diversity metrics using analysis of variance (ANOVA) tests (Fisher, 1928) and employing a multiple comparisons significance adjustment (detailed methods and results are provided in section 3 of Appendix S1). We visualized the composition of samples at the order level using barplots per sample and pooled barplots per geographic region. To quantify the contribution of different taxa across samples, we also computed prevalence as the proportion of samples in which a taxon is present and the mean abundance of each taxon across samples (see section 7 of Appendix S1). Prevalence and mean abundance were quantified at the order and family taxonomic levels. Bacterial families and orders that were prevalent in more than 75% of the samples are considered to be part of the

core microbiome of *C. californica*; the rest of the taxonomic groups are referred to as non-core taxa.

Next, to characterize the similarity of the bacteriome across samples, we computed Bray–Curtis dissimilarity matrices at different taxonomic levels (genus, family, and order) and tested for the effect of geographic region, month, and year using PERMANOVAs (Anderson, 2017) with 999 permutations. Because geographic region had a significant effect, we evaluated which specific taxa were significantly over- or under-represented across geographic regions using a differential abundance test at the family and genus taxonomic levels. We assessed significance using a linear discriminant analysis effect size (LefSE) as implemented in microeco (Segata et al., 2012). Before using this test, we applied an arc-sine square-root transformation of the abundances to account for the use of relative abundances.

Effect of climatic variables on microbial community structure

Because macro-environmental variables are highly correlated with geographic regions and could be drivers of composition differences in the *C. californica* bacteriome, we evaluated the potential correlation of climatic variables on the composition of *C. californica*'s bacteriome. For this, we extracted the values of the bioclimatic layers from the BioClim database (Fick and Hijmans, 2017) for each collection point using the R packages sf (Pebesma, 2018) and raster (Hijmans, 2025). Of the 19 variables in BioClim, we selected a subset of variables with relatively low correlation and high variation within our geographic scope (Figure S2 in Appendix S1). We used a canonical correspondence analysis (CCA) to visualize the relationship between the climatic variables and the microbial composition of *C. californica* samples at the family level (Ter Braak, 1986), and we tested for the significance of this relationship using a Mantel test (Mantel, 1967) implemented in microeco.

Functional composition

To characterize the functional profile of the bacteriome associated with *C. californica*, we used the FAPROTAX v1.2.10 database (Louca et al., 2016a, 2016b) to link the taxonomic composition of the bacteriome of each sample to functional diversity using the microeco built-in function. We tested for differential abundance of these functions in different geographic regions using LefSE, with the relative abundance of the different functions in the bacteriome as input.

Reproducibility

The Bash, Python, and R code used to perform the cleaning, microbiome pipeline and statistical analyses, and visualizations is available in the GitHub repository: https://github.com/ixchelgzlzr/CCal_microbiome. Output from MetaPhlAn is also provided so that statistical analyses can be reproduced and independently tested.

RESULTS

Bioinformatics

On average, there were 30,600,000 reads before filtering out liverwort reads, ranging from 13,028,666 to 45,988,969 reads. After filtering out the reads originating from *C. californica*, there was an average of ~19,000,000 reads (61% of the total), with the smallest sample containing 3,835,559 reads and the largest containing 36,016,561 reads. After running the MetaPhlAn microbiome profiling pipeline, an average of 4% of reads were used for taxonomic profiling of the microbiome; this is an average of ~640,000 reads per sample, with the smallest sample containing 70,519 reads and the largest containing 3,610,608 reads (Table S4 in Appendix S1).

For our dataset, the running time on a standard laptop using eight cores was about seven hours per sample. This timeline would have been restrictive, so the pipeline was performed in UC Berkeley's Savio computing cluster. In our particular cluster environment, we observed that even with 384 GB of total RAM available for 20 cores, we faced memory constraints when trying to parallelize per sample. We found that we obtained the best performance when assigning all the cores and RAM memory to a single sample and then running the analysis sequentially for all samples. With this setup, each sample took about 30 min (20 cores simultaneously with 384 GB of RAM).

Composition of the bacterial microbiome of *Calasterella californica*

Across the 97 samples of *C. californica*, the MetaPhlAn pipeline identified 751 species-level bacterial operational taxonomic units (OTUs) belonging to 338 different genera, 141 families, 87 orders, 51 classes, and 15 phyla (Table S5 in Appendix S1). We identified eight core orders (Hyphomicrobiales, Pseudonocardiales, Pseudomonadales, Burkholderiales, Micrococcales, Cytophagales, Acetobacterales, and Propionibacteriales; Figure 2) and eight core families (Methylobacteriaceae, Pseudonocardiaceae, Pseudomonadaceae, Microbacteriaceae, Nitrobacteriaceae, Cytophagaceae, Rhizobiaceae, and Roseomonadaceae) of bacteria that were present in more than 75% of samples of *C. californica* (see section 7 of Appendix S1). Overall, the most prevalent taxa were also the taxa with the highest abundance. A few exceptions are the orders Mycobacteriales, Oscillatoriales, and Micromonosporales and the families Noracardiaceae and Comamonadaceae; these were relatively abundant when present but were not prevalent enough across samples to be considered part of the core

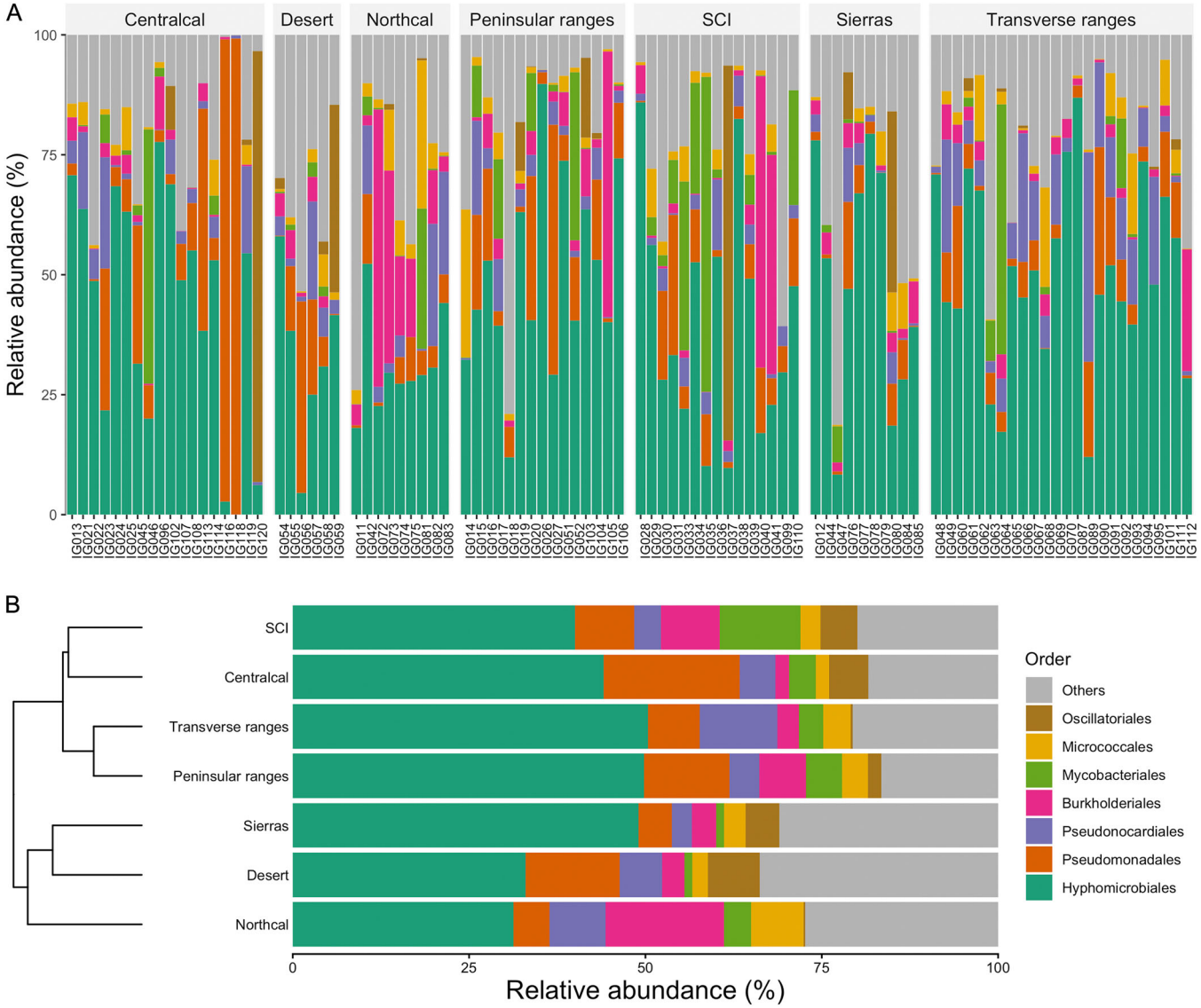

FIGURE 2 Composition of the bacteriome of *Calasterella californica*. (A) Relative abundance of the seven most abundant orders of bacteria across all the samples of *C. californica*. The samples are arranged by the geographic region where they were collected. (B) Bacteriome composition grouped by geographic region. The relative abundances are obtained from group averages. The dendrogram on the left reflects similarity among regions based on a hierarchical clustering calculated from Euclidian distances. SCI, Santa Cruz Island region.

microbiome (see section 7 of Appendix S1). At both the order and family level, the microbiome was largely dominated by one taxon. The order Hyphomicrobiales had a mean relative abundance of more than 40%, while the next most abundant order had a relative abundance of ~10%. Similarly, the family Methylobacteriaceae had a mean relative abundance of almost 30%, while the second most abundant family had a mean relative abundance of ~10% (see section 7 of Appendix S1).

According to an ANOVA, there were no significant differences in α-diversity metrics across geographic regions (Table S1 in Appendix S1). When accounting for the marginal effect of the month and year of collection on diversity metrics (Table S2 in Appendix S1), we detected a significant effect of year only on the Simpson Diversity index ($P = 0.04$). In terms of the composition of bacterial communities, we found a significant effect of the geographic region across all taxonomic levels evaluated (genus, family, and order) according to a PERMANOVA test (Figure 3A). While statistically significant ($P < 0.01$), the effect of geographic region was weak, explaining only around 10% of the variation observed (Figure 3A). Further inspection of the specific taxa that vary across geographic regions with a LefSE analysis revealed 18 families and 50 genera of bacteria that have significantly differential abundance in different geographic regions (Figure 4). Of the taxa with significant differential abundance across regions, only one family (Pseudonocardiaceae) and two orders (Pseudonocardiales and Propionibacteriales) were part of the core microbiome of *C. californica* (see section 7 of Appendix S1). The

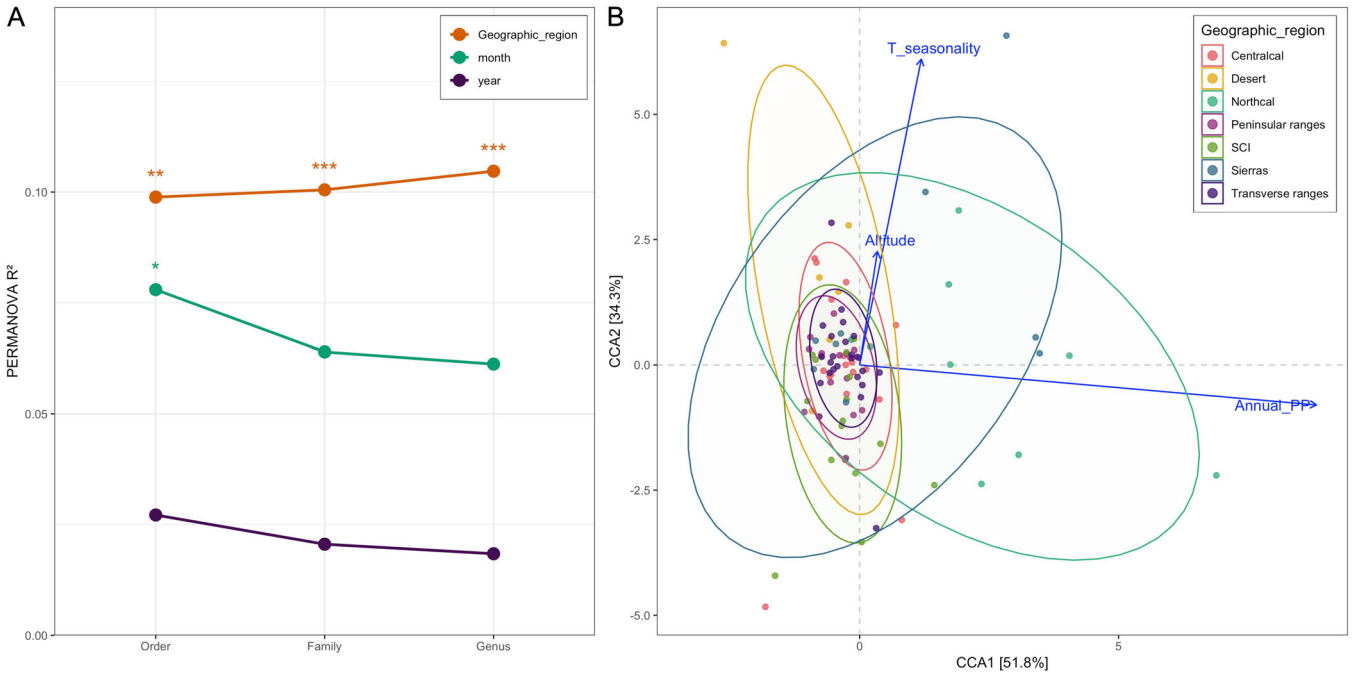

FIGURE 3 Effect of geographic region, collection date, and environmental variables on microbiome composition. (A) Marginal effects of geographic region, collection year, and collection month on bacterial community composition across taxonomic levels (order, family, genus), assessed using PERMANOVA on Bray–Curtis dissimilarities (999 permutations). Points indicate the proportion of variance (R^2). Asterisks denote levels of significance (*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$). (B) Canonical correspondence analysis (CCA) of bacterial community composition at the family level constrained by macroclimatic variables (temperature seasonality, altitude, and annual precipitation). Points represent individual samples colored by geographic region, ellipses indicate the dispersion of samples within regions, and blue arrows show the direction and relative strength of climatic gradients in ordination space. SCI, Santa Cruz Island region.

PERMANOVA test also detected a significant effect of the month of collection only at the order level ($P < 0.5$; Figure 3A, Table S2 in Appendix S1).

Effect of macro-climatic variables in bacteriome composition

The three macro-climatic variables that we selected to test for the effect of climate on bacteriome composition were temperature seasonality, altitude, and annual precipitation. These three variables explained most of the variation across geographic regions and captured the variation in other highly correlated variables (Figure S3 in Appendix S1). A biplot visualization of the CCA (Figure 3B) shows a significant association of annual temperature with the main ordination axis according to a permutation analysis ($\chi^2 = 0.17$, $F = 2.39$, $P = 0.002$), while the second ordination axis has a marginal association with temperature seasonality ($\chi^2 = 0.097$, $F = 1.35$, $P = 0.055$; Figure 3). Additionally, a Mantel test also supports a weak but significant effect of annual precipitation and temperature seasonality (but not altitude) in the composition of the bacterial communities of *C. californica* at the family level (Mantel test: temperature seasonality, altitude, annual precipitation; Pearson correlation coefficient: 0.11, 0.07, 0.18; adjusted *P*-values: 0.006, 0.064, 0.006, respectively).

Functional composition

The functional composition analyses show that there is a relatively high proportion of bacteria associated with energy sourcing, specifically performing the functions of aerobic chemoheterotrophy and anaerobic chemoheterotrophy. Other functions that are relatively well represented in the bacteriome of *C. californica* are methylotrophy, methanotrophy, hydrocarbon degradation, ureolysis, and methanol oxidation (Figure 5). All of these highly represented functions seem to be common across all the samples. A differential abundance LefSE analysis performed for these different functions showed significant differences in 20 categories (Figure S5 in Appendix S1); however, none of the taxa in any of these categories comprised more than 4% of the total relative abundance (Figure S4 in Appendix S1).

DISCUSSION

Practical insights of the off-target metagenomics approach

By definition, the off-target metagenomics approach uses WGS datasets collected for different purposes. Consequently, researchers using this approach lack the ability to control the sampling design. Thus, when using this

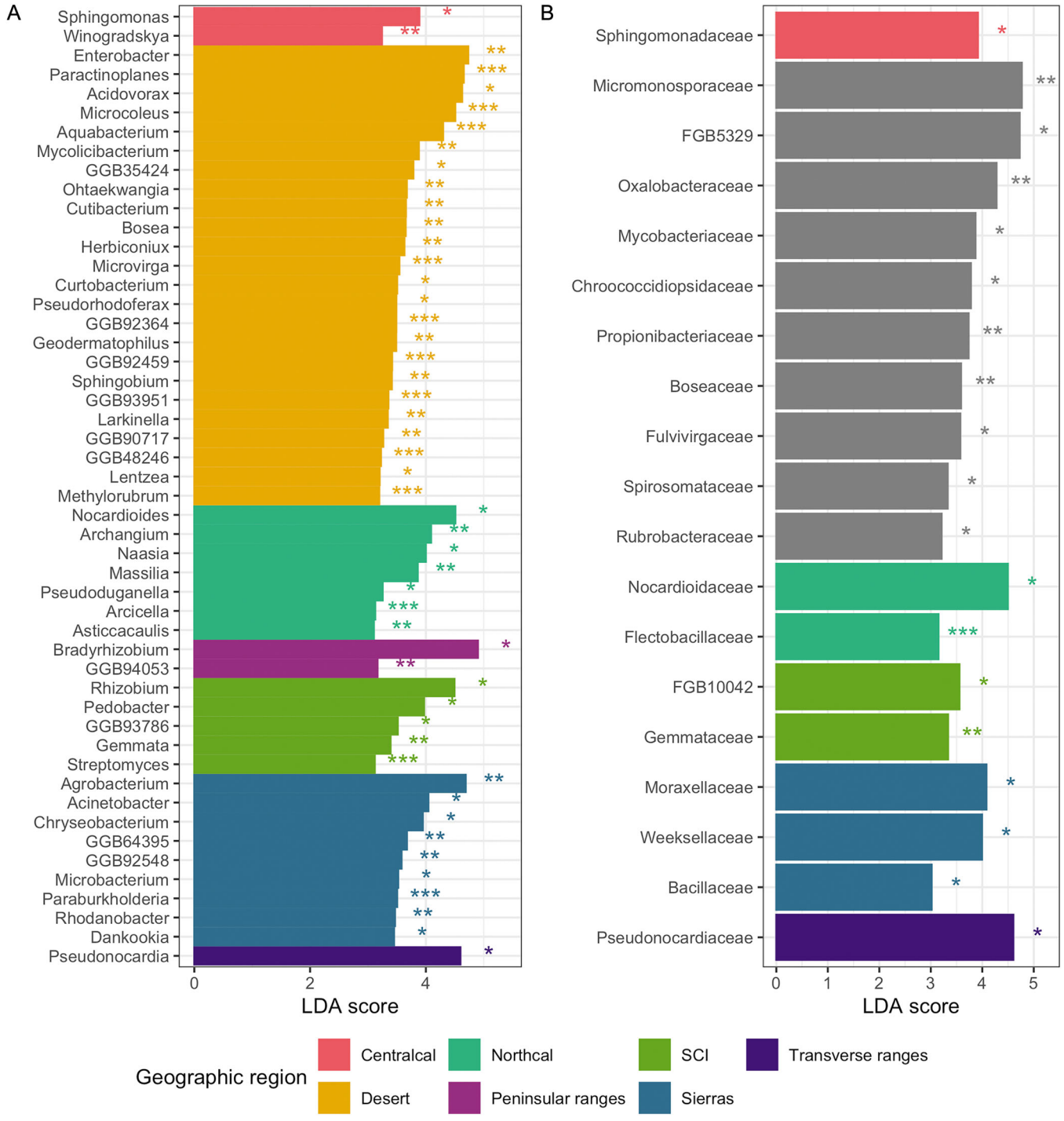

FIGURE 4 Significant differential abundance of bacterial (A) genera and (B) families across geographic regions. There are 18 families and 50 genera of bacteria with significantly different abundances across geographic regions. Differential abundance was determined based on LDA scores >3. Significance values are indicated with asterisks (*$P < 0.5$, **$P < 0.1$, ***$P < 0.01$). When differential abundance is characteristic of a single region, the bar is colored by that region; otherwise, the bar is gray, signifying that more than one region has differential abundance of this taxon. SCI, Santa Cruz Island region.

approach, it is important to statistically check for the potential effect of recorded variables. In our dataset, we tested for the potential effect of variables like collection date (month and year), sequencing batch, and plant sex on the number of reads used by the pipeline (see section 3 of Appendix S1) and found none. We also included collection date as a potential explanatory variable for microbiome composition in our statistical models (PERMANOVA). We found that, while collection date (month) has a small effect on community composition (measured as Bray–Curtis distances between samples), the effect is weak and inconsistent across taxonomic

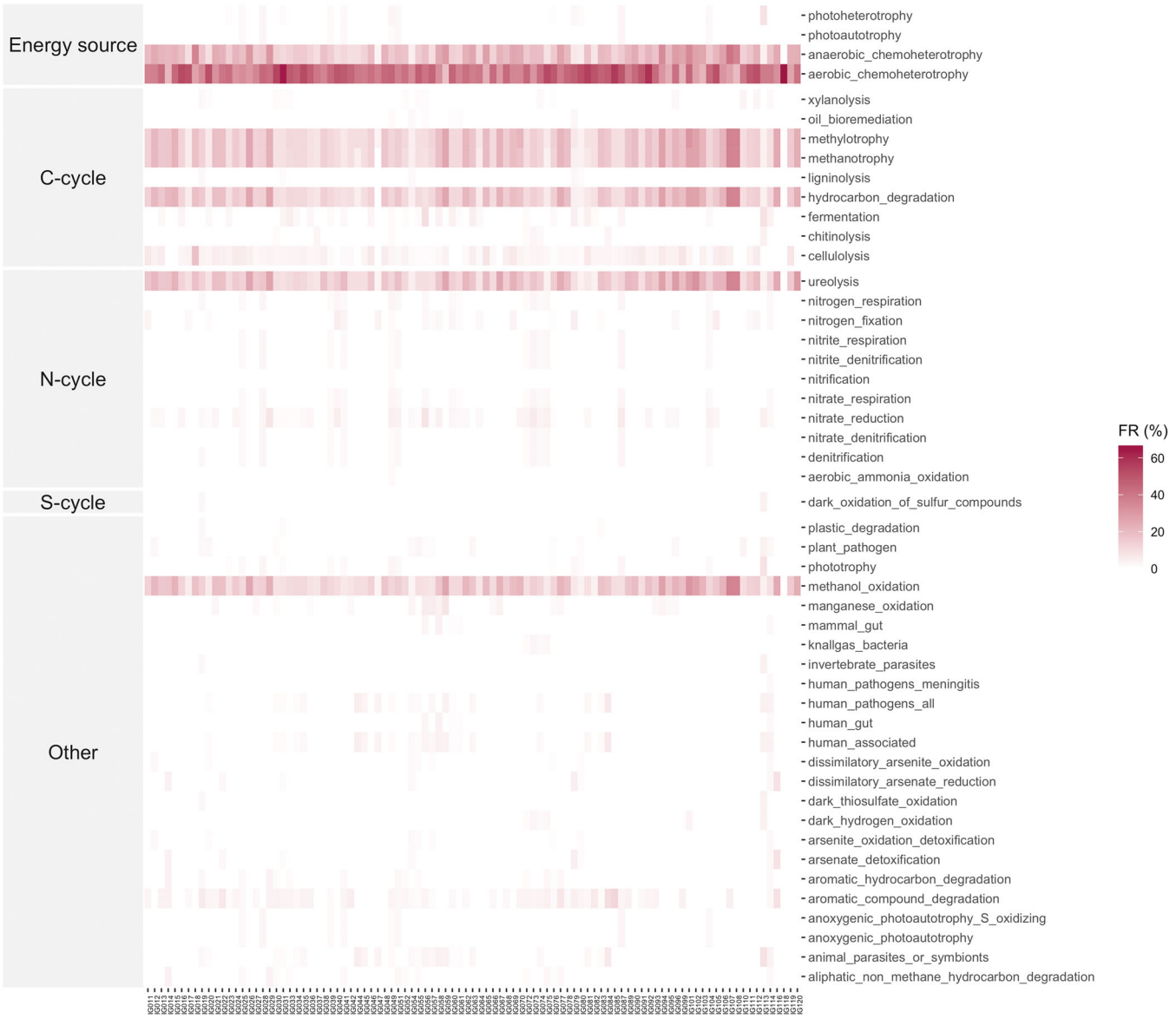

FIGURE 5 Heat map showing the functional composition of the bacteriome of *Calasterella californica*, showing the relative abundance of bacterial functions in the community of bacteria associated with 97 samples of *C. californica*. C-cycle, carbon cycle; N-cycle, nitrogen cycle; S-cycle, sulfur cycle.

scales (it was only observed at the order level). We focused on the comparatively stronger effect of geographic region on bacterial communities but acknowledge that temporal variation of microbiome composition might deserve a focused study in the future.

Another feature of this approach is the small proportion of off-target (i.e., non-host related) reads that are used for metagenomic characterization, e.g., in our analyses only 5% of reads were used. While this proportion of used reads might be surprising in a classic metagenomic approach, it is normal and expected in MetaPhlAn 4's marker-based approach, as only clade-specific markers are present in the comparison database. Reads that do not map to markers are reported as "unclassified" but are not conceptually considered failed classifications; rather, they are expected in this framework and are excluded from abundance estimation (Segata et al., 2012; Truong et al., 2015; Blanco-Míguez et al., 2023). In fact, our reported number of classified reads is comparable to the benchmarking classifications of MetaPhlAn (7–8%; Segata et al., 2012). The slightly smaller proportion of used reads in our study (5% vs. 8%) might be due to the understudied nature of our system, with potentially many bacterial taxa not being represented in MetaPhlAn's SGB database. The difficulty in characterizing the bacterial communities of understudied systems is shared among all metagenomic studies due to the incompleteness of the reference databases, which are biased against bacteria exclusive of understudied environments. It is important to mention that the SGB database has been continuously updated (Blanco-Míguez et al., 2023) since MetaPhlAn was first developed in 2012; therefore, we can expect our ability to classify bacteria to increase over time.

The microbial communities of a widespread California liverwort

In this paper, we characterize the composition of the bacterial community associated with *C. californica*. Because whole thalli were sequenced, we lack the ability to distinguish between endophytic bacteria and bacteria living on the plant surface; therefore, the best way to think about our results is to conceptualize the described bacterial community as part of the *C. californica* holobiont. The pattern of composition that we observed in the microbial community of this liverwort is two-fold. On one hand, we found a core group of bacterial taxa that are remarkably prevalent and abundant across samples and geographic regions, making the bacterial communities of this liverwort very similar across the distributional range of this plant. At the same time, non-core bacterial taxa vary across samples in a pattern that correlates with geographic region and climatic variables (e.g., annual precipitation and seasonality).

A stable set of core microbial taxa

Despite living in contrasting regions across California (Figure 1, Figure S3 in Appendix S1), the bacterial communities associated with *C. californica* do not show significant differences in any of the α-diversity metrics (Table S1 in Appendix S1) across geographic regions. *Calasterella californica* bacteriomes are largely dominated by bacteria in the order Hyphomicrobiales (Figure 2, section 7 of Appendix S1), with important contributions from a few other orders, such as Pseudomonadales, Pseudonocardiales, Burkholderiales, Mycobacteriales, Micrococcales, and Oscillatoriales (Figure 2). These results are congruent with previous work in liverworts of the genus *Riccia* L., whose bacterial communities are seemingly unaffected by environmental conditions. For example, Wicaksono et al. (2023), using a traditional targeted 16S sequencing approach, found a stable bacterial microbiome in *Riccia* individuals growing in contrasting soil types. Similarly, Wiśniewski et al. (2025), using nanopore long-read sequencing, identified a core microbiome that was significantly enriched on some bacterial taxa in comparison to the surrounding soil. While comparative studies of the bacterial communities associated with liverworts are still scarce, these three studies together provide growing evidence that liverwort-associated bacterial communities are largely non-random and somewhat stable assemblies, regardless of the environmental conditions in which individual liverworts occur. This observation particularly contrasts with patterns observed in angiosperms, where bacterial community composition is strongly shaped by both host genotype and environment (Wagner et al., 2016; Wei and Tan, 2023; He et al., 2024).

Geographic variation of non-core bacterial taxa

While the primary components of the microbiome of *C. californica* are the same across samples, the core family Pseudonocardiaceae had significant differential abundance across regions. However, rather than characterizing a single region, this differential abundance specifically differentiated the Transverse Ranges and Northern California regions from the Santa Cruz Island and the Sierras regions (see section 6 of Appendix S1). Bacteria in this family are diverse in their biology and are associated with soil and plants, and further studies are needed to understand the correlation between this taxon abundance and the geographic region. Furthermore, when scrutinizing the identity of non-core taxa, we found some differences across geographic regions (Figure 4) that are potentially associated with variation in annual precipitation and temperature seasonality (Figure 3). For example, a higher abundance of photosynthetic bacteria in the families Oscillatoriaceae, Calotrichaceae, and Coleofasciculaceae was correlated with higher annual precipitation. Desert samples had many significantly differentially abundant genera (Figure 4), many of which—as reflected in the functional profile (Figure S5 in Appendix S1)—are plant pathogens and human, animal, and gut associated (Figure 4). One potential explanation is simply that the samples from the desert were collected near seasonal streams influenced by runoff and touristic human activities. Another potential explanation could be a mechanistic response, similar to what has been observed in some angiosperms, to water availability (e.g., Chao et al., 2025) or diseases (e.g., Gao et al., 2021). Further work is needed to understand the cause of differences in the non-core bacteria of *C. californica*'s microbiome in the desert and Santa Cruz Island regions, the two most environmentally unique sampled regions.

The functional composition of *Calasterella californica*'s bacteriome

Among the taxa that were consistently abundant in *C. californica* bacterial communities were taxa in the order Hyphomicrobiales. Two representatives of this order, *Methylobacterium* and *Rhizobium*, were also found to be main components of the microbiome of the model liverwort *Marchantia polymorpha* L. (Alcaraz et al., 2018), where they play an important role promoting the growth of this plant (Kutschera and Koopmann, 2005). The order Hyphomicrobiales contains many genera that are known to be beneficial to plants not only for stimulating plant growth, but also as nitrogen-fixers and root nodulation promoters (Lindström and Mousavi, 2020). While our dataset precludes assessing the location of this bacteria, it is likely that, as in *Marchantia* L., Hyphomicrobiales live on the surface of *C. californica* thalli, potentially having a close mutualistic interaction in which bacteria use plant-secreted methanol as source of carbon while serving regulatory purposes for the plant.

The functional profiles of *C. californica* bacteriomes were largely conserved across samples (Figure 5), with high relative abundance of bacteria that perform methanol oxidation, methanotrophy and methylotrophy, and aerobic

chemoheterotrophy. These results highlight the role that methylotrophic bacteria play as the dominant taxon in the *C. californica* microbiome. Whether and how liverworts actively maintain high proportions of these bacteria in their microbiome, in a dynamic and heterogeneous environment, warrants more investigation.

The significance of the off-target metagenomics approach

While the off-target metagenomics approach, by definition, is not the optimal approach to study microbiomes, the results obtained with this approach are complementary to other targeted approaches and are particularly valuable in identifying questions that require more targeted research than can be addressed with standard microbiome-focused approaches. Furthermore, the congruent results of our study with previous research on liverwort microbiomes using 16S rRNA metabarcoding (Alcaraz et al., 2018; Wicaksono et al., 2023) are interesting not only because they imply general patterns of microbiome composition across liverworts, but also because they provide indirect support that 16S rRNA metabarcoding and metagenomic approaches like MetaPhlAn discover patterns that are consistent with each other. Currently, there has been an increasing emphasis on incorporating long-read sequencing technology, which has facilitated the assembly of high-quality metagenome-assembled genomes and allowed for strain-level resolution and metabolic profiling (Han et al., 2024). Nonetheless, these approaches remain both cost-prohibitive and computationally expensive. The ability to directly use shotgun sequencing data to study microbiomes facilitates the study of bacteria associated with non-model organisms, at a time when increasing numbers of WGS datasets are being generated.

CONCLUSIONS

The use of MetaPhlAn in our study exemplifies how sequencing produced for plant genomics work can be leveraged to obtain microbiome metagenomics information. Our results demonstrate that meaningful metagenomic insights can be gained from utilizing these types of landscape genomic datasets, and these insights are especially impactful for relatively understudied groups and groups where we might expect a high proportion of non-target reads in WGS, such as liverworts. Our results are consistent with previous knowledge on liverwort bacterial microbiomes generated through standard methods for the study of microbiomes, pointing to the efficacy of these marker-based response strategies in exploratory studies that lead to targeted research questions on the assembly of microbial communities. In the case of liverwort bacteriomes, our results highlight the stability of major components of *C. californica*'s microbiome, while conversely revealing some variation associated with the geographic region in which the liverwort occurs. It is increasingly exciting to conduct sampling and in vitro studies aimed at understanding the potentially active role that liverworts play in the assembly of its bacterial communities.

AUTHOR CONTRIBUTIONS

I.S.G.R. and M.J.S. designed the study, performed the analyses, and wrote the first draft of the manuscript. E.C.M. contributed to analyses and editing of subsequent versions of the manuscript. B.D.M. supervised the work and edited the manuscript. All authors approved the final version of the manuscript.

ACKNOWLEDGMENTS

The authors thank Carl Rothfels for his constructive comments and edits to the manuscript, and *Applications in Plant Sciences*'s Senior Associate Editor Gregory J. Pec and Reviewing Editor Shan Wong and two anonymous reviewers for their thoughtful and constructive feedback on this manuscript. We also thank the California Conservation Genomics Project, with funding provided to the University of California by the State of California, State Budget Act of 2019 (UC Award ID RSI-19-690224). I.S.G.R. was supported by a UC MEXUS-CONACyT fellowship (number 709967), a Plant Science Fellowship from Oak Spring Garden Foundation, and the Philomathia Graduate Fellowship in Environmental Sciences at UC Berkeley. This research used the Savio computing cluster resource provided by the UC Berkeley Research Computing program at the University of California, Berkeley (supported by the UC Berkeley Chancellor, Vice Chancellor for Research, and Chief Information Officer).

DATA AVAILABILITY STATEMENT

All the data and code used for the analyses are hosted in the GitHub repository: https://github.com/ixchelgzlzr/CCal_microbiome.

ORCID

Ixchel S. González-Ramírez https://orcid.org/0000-0003-3667-5360
Michael J. Song https://orcid.org/0000-0001-5961-753X
Elijah C. Mehlferber https://orcid.org/0000-0002-9958-3247
Brent D. Mishler https://orcid.org/0000-0001-5727-4916

REFERENCES

Alcaraz, L. D., M. Peimbert, H. R. Barajas, A. E. Dorantes-Acosta, J. L. Bowman, and M. A. Arteaga-Vázquez. 2018. *Marchantia* liverworts as a proxy to plants' basal microbiomes. *Scientific Reports* 8: 12712.

Anderson, M. J. 2017. Permutational multivariate analysis of variance (PERMANOVA). In N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri, and J. L. Teugels [eds.], Wiley StatsRef: Statistics Reference Online. Available at https://doi.org/10.1002/9781118445112.stat07841.

Benham, P. M., C. Cicero, M. Escalona, E. Beraut, M. P. A. Marimuthu, O. Nguyen, M. W. Nachman, and R. C. K. Bowie. 2023. A highly contiguous genome assembly for the California quail (*Callipepla californica*). *Journal of Heredity* 114: 418–427.

Beninde, J., E. Toffelmier, and H. B. Shaffer. 2022. A brief history of population genetic research in California and an evaluation of its utility for conservation decision-making. *Journal of Heredity* 113: 604–614.

Bidartondo, M. I., and J. G. Duckett. 2010. Conservative ecological and evolutionary patterns in liverwort–fungal symbioses. *Proceedings of the Royal Society B: Biological Sciences* 277: 485–492.

Blair, K., S. Rose, J. Hull, M. Escalona, A. Finger, S. E. K. Joslin, R. Sahasrabudhe, et al. 2022. The reference genome of the vernal pool tadpole shrimp, *Lepidurus packardi*. *Journal of Heredity* 113: 706–711.

Blanco-Míguez, A., F. Beghini, F. Cumbo, L. J. McIver, K. N. Thompson, M. Zolfo, P. Manghi, et al. 2023. Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nature Biotechnology* 41: 1633–1644.

Bolger, A. M., M. Lohse, and B. Usadel. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30: 2114–2120.

Chao, C., D. Xu, B. Jiang, X. Lu, C. Yu, Y. Wang, H. Wang, J. Li, and J. Zhu. 2025. Precipitation-driven restructuring of rhizosphere microbiota enhances alpine plant adaptation. *Frontiers in Plant Science* 16: e1641511.

Fick, S. E., and R. J. Hijmans. 2017. WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology* 37: 4302–4315.

Fiedler, P. L., B. Erickson, M. Esgro, M. Gold, J. M. Hull, J. M. Norris, B. Shapiro, et al. 2022. Seizing the moment: The opportunity and relevance of the California Conservation Genomics Project to state and federal conservation policy. *Journal of Heredity* 113: 589–596.

Fisher, R. A. 1928. Statistical methods for research workers. Oliver and Boyd, Edinburgh, United Kingdom.

Gao, M., C. Xiong, C. Gao, C. K. M. Tsui, M.-M. Wang, X. Zhou, A.-M. Zhang, and L. Cai. 2021. Disease-induced changes in plant microbiome assembly and functional adaptation. *Microbiome* 9: e187.

González Ramírez, I. S. 2024. Studies on the biology and evolution of liverworts: From genomes to biogeography. Ph.D. thesis, University of California, Berkeley, California, USA.

González-Ramírez, I. S., M. Escalona, C. Miller, N. Chumchim, M. Marimuthu, O. Nguyen, M. J. Song, and B. D. Mishler. 2025a. A genome assembly for the California endemic liverwort *Calasterella californica*. *Journal of Heredity* 116: 389–396.

González-Ramírez, I. S., D. G. Long, L. Briscoe, and B. D. Mishler. 2025b. Typification of *Fimbraria californica* (*Calasterella californica*, Aytoniaceae). *Madroño* 72: 37–39.

Goodrich, J. K., S. C. Di Rienzi, A. C. Poole, O. Koren, W. A. Walters, J. G. Caporaso, R. Knight, and R. E. Ley. 2014. Conducting a microbiome study. *Cell* 158: 250–262.

Grether, G. F., J. Beninde, E. Beraut, N. Chumchim, M. Escalona, Z. G. MacDonald, C. Miller, et al. 2023. Reference genome for the American rubyspot damselfly, *Hetaerina americana*. *Journal of Heredity* 114: 385–394.

Han, Y., J. He, M. Li, Y. Peng, H. Jiang, J. Zhao, Y. Li, and F. Deng. 2024. Unlocking the potential of metagenomics with the PacBio high-fidelity sequencing technology. *Microorganisms* 12: e2482.

Härer, A., and D. J. Rennison. 2023. The biogeography of host-associated bacterial microbiomes: Revisiting classic biodiversity patterns. *Global Ecology and Biogeography* 32: 931–944.

He, X., D. Wang, Y. Jiang, M. Li, M. Delgado-Baquerizo, C. McLaughlin, C. Marcon, et al. 2024. Heritable microbiome variation is correlated with source environment in locally adapted maize varieties. *Nature Plants* 10: 598–617.

Hijmans, R. 2025. raster: Geographic data analysis and modeling. R package version 3.6-32. https://github.com/rspatial/raster.

Huang, Y., M. Escalona, G. Morrison, M. P. A. Marimuthu, O. Nguyen, E. Toffelmier, H. B. Shaffer, and A. Litt. 2022. Reference genome assembly of the big berry manzanita (*Arctostaphylos glauca*). *Journal of Heredity* 113: 188–196.

Kutschera, U., and V. Koopmann. 2005. Growth in liverworts of the Marchantiales is promoted by epiphytic *Methylobacteria*. *Naturwissenschaften* 92: 347–349.

Li, H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv 1303.3997 [preprint]. Available at https://doi.org/10.48550/arXiv.1303.3997 [posted 16 March 2013; accessed 19 May 2026].

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, et al. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* 25: 2078–2079.

Lindström, K., and S. A. Mousavi. 2020. Effectiveness of nitrogen fixation in rhizobia. *Microbial Biotechnology* 13: 1314–1335.

Liu, C., X. Li, F. R. P. Mansoldo, J. An, Y. Kou, X. Zhang, J. Wang, et al. 2022. Microbial habitat specificity largely affects microbial co-occurrence patterns and functional profiles in wetland soils. *Geoderma* 418: e115866.

Liu, C., F. R. P. Mansoldo, H. Li, A. B. Vermelho, R. J. Zeng, X. Li, and M. Yao. 2026. A workflow for statistical analysis and visualization of microbiome omics data using the R microeco package. *Nature Protocols* 21: 1300–1324.

Long, D. G., and T.-X. Zheng. 2023. A new subfamily Calasterelloideae and new genus *Calasterella* for a phylogenetically and morphologically distinct member of the Aytoniaceae. *Phytotaxa* 606: 225–230.

Louca, S., S. M. S. Jacques, A. P. F. Pires, J. S. Leal, D. S. Srivastava, L. W. Parfrey, V. F. Farjalla, and M. Doebeli. 2016a. High taxonomic variability despite stable functional structure across microbial communities. *Nature Ecology & Evolution* 1: 0015.

Louca, S., L. W. Parfrey, and M. Doebeli. 2016b. Decoupling function and taxonomy in the global ocean microbiome. *Science* 353: 1272–1277.

Mantel, N. 1967. The detection of disease clustering and a generalized regression approach. *Cancer Research* 27: 209–220.

Mead, A., S. T. Fitz-Gibbon, M. Escalona, E. Beraut, S. Sacco, M. P. A. Marimuthu, O. Nguyen, and V. L. Sork. 2024. The genome assembly of island oak (*Quercus tomentella*), a relictual island tree species. *Journal of Heredity* 115: 221–229.

Pebesma, E. 2018. Simple features for R: Standardized support for spatial vector data. *The R Journal* 2018: e626. https://digitalcommons.unl.edu/r-journal/626

Pielou, E. C. 1966. The measurement of diversity in different types of biological collections. *Journal of Theoretical Biology* 13: 131–144.

R Core Team. 2026. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Website http://www.R-project.org/ [accessed 19 May 2026].

Segata, N., L. Waldron, A. Ballarini, V. Narasimhan, O. Jousson, and C. Huttenhower. 2012. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods* 9: 811–814.

Shaffer, H. B., E. Toffelmier, R. B. Corbett-Detig, M. Escalona, B. Erickson, P. Fiedler, M. Gold, et al. 2022. Landscape genomics to enable conservation actions: The California Conservation Genomics Project. *Journal of Heredity* 113: 577–588.

Shannon, C. E. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27: 379–423.

Simpson, E. H. 1949. Measurement of diversity. *Nature* 163: 688.

Song, M. J., S. Pirro, S. Lahmeyer, K. Kyle, and M. White. 2024. The complete chloroplast sequence of the invasive feathered mosquito fern and a draft genome of its endosymbiont reveals that rates of genome erosion vary across the genus. *American Fern Journal* 114: 334–336.

Song, M. J., F. Freund, C. M. Tribble, E. Toffelmier, C. Miller, H. B. Shaffer, F.-W. Li, and C. J. Rothfels. 2025a. The nitrogen-fixing fern *Azolla* has a complex microbiome characterized by varying degrees of co-phylogenetic signal. *American Journal of Botany* 112: e70010.

Song, M. J., Y. C. Rizzieri, F.-W. Li, F. Freund, M. Escalona, E. Toffelmier, C. Miller, et al. 2025b. The genome assembly of the duckweed fern, *Azolla caroliniana*. *Journal of Heredity* 116(5): 691–701.

Steinegger, M., and S. L. Salzberg. 2020. Terminating contamination: Large-scale search identifies more than 2,000,000 contaminated entries in GenBank. *Genome Biology* 21: e115.

Supple, M. A., M. Escalona, N. Alexandre, M. R. Buchalski, S. P. D. Riley, J. A. Dellinger, T. W. Vickers, et al. 2025. A chromosome-level genome assembly of the mountain lion, *Puma concolor*. *Journal of Heredity* 116: 479–487.

Ter Braak, C. J. F. 1986. Canonical correspondence analysis: A new eigenvector technique for multivariate direct gradient analysis. *Ecology* 67: 1167–1179.

Toffelmier, E., J. Beninde, and H. B. Shaffer. 2022. The phylogeny of California, and how it informs setting multispecies conservation priorities. *Journal of Heredity* 113: 597–603.

Truong, D. T., E. A. Franzosa, T. L. Tickle, M. Scholz, G. Weingart, E. Pasolli, A. Tett, et al. 2015. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature Methods* 12: 902–903.

Vandenkoornhuyse, P., A. Quaiser, M. Duhamel, A. Le Van, and A. Dufresne. 2015. The importance of the microbiome of the plant holobiont. *New Phytologist* 206: 1196–1206.

Wagner, M. R., D. S. Lundberg, T. G. Del Rio, S. G. Tringe, J. L. Dangl, and T. Mitchell-Olds. 2016. Host genotype and age shape the leaf and root microbiomes of a wild perennial plant. *Nature Communications* 7: e12151.

Wei, N., and J. Tan. 2023. Environment and host genetics influence the biogeography of plant microbiome structure. *Microbial Ecology* 86: 2858–2868.

Wicaksono, W. A., B. Semler, M. Pöltl, C. Berg, G. Berg, and T. Cernava. 2023. The microbiome of *Riccia* liverworts is an important reservoir for microbial diversity in temporary agricultural crusts. *Environmental Microbiome* 18: e46.

Wiśniewski, P., M. Maździarz, J. Sawicki, and K. Krawczyk. 2025. Taxonomic and functional differentiation of soil and thallus microbiomes in *Riccia sorocarpa*. *Biologia* 80: 2695–2706.

Wood, D. A., J. Q. Richmond, M. Escalona, M. P. A. Marimuthu, O. Nguyen, S. Sacco, E. Beraut, et al. 2022. Reference genome of the California glossy snake, *Arizona elegans occidentalis*: A declining California species of special concern. *Journal of Heredity* 113: 632–640.

Young, B. C., B. Thiers, J. F. White, and L. Struwe. 2025. Endophytic bacteria discovered in oil body organelles of the liverworts *Marchantia polymorpha* and *Radula complanata*. *American Journal of Botany* 112: e70017.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Appendix S1. Supporting Information for: "Off-target metagenomics: Leveraging whole genome sequencing to study the bacteriome of the liverwort *Calasterella californica*."

How to cite this article: González-Ramírez, I. S., M. J. Song, E. C. Mehlferber, and B. D. Mishler. 2026. Off-target metagenomics: Leveraging whole genome sequencing to study the bacteriome of the liverwort *Calasterella californica*. *Applications in Plant Sciences* 14(3): e70064. https://doi.org/10.1002/aps3.70064