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

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

[1]*Missouri Botanical Garden, St. Louis, MO*

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

[3]*Skyline College, San Bruno, CA*

[4]Georgia Institute of Technology, Atlanta, GA 30332

[a]*These authors contributed equally to the work*

[*]*E-mail: igonzalez@mobot.org*

January 27, 2026

Manuscript received _____; revision accepted _____.

Total word count: 3749

Short title: González-Ramírez et al.—*Calasterella californica* bacteriome

ABSTRACT

Premise of study: 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 pipelines and databases enable the use 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.

Key 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 was 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.

Conclusions: 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. Our novel off-target metagenomics approach can be applied to any population level re-sequencing where whole organisms are sequenced, opening the door to exciting avenues of microbiome research using re-purposed data from landscape genomics.

Keywords: Bryophyte; California Conservation Genomics Project; CCGP; Liverwort; MetaPhlan; *Methylobacterium*; Microbiome

INTRODUCTION

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. 2024), we might expect a relatively low proportion of contaminant reads. But performing WGS on small organisms, like insects, small invertebrates, and plants such as bryophytes, often involves using complete individual(s) as a sample, which increases the proportion of non-host reads that we might expect because of 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 organism 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 host and micro-organism (Song et al. 2025a, 2024), in an approach that we refer to as "off-target metagenomics".

To demonstrate the practicality of the off-target metagenomics approach, we re-purpose 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 (Shaffer et al. 2022; Toffelmier et al. 2022; Fiedler et al. 2022; Beninde et al. 2022), to explore 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 (Mead et al. 2024; Grether et al. 2023; Blair et al. 2022; Huang et al. 2022; 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, (Blanco-Míguez et al. 2023) that matches short-reads from metagenomic samples to "species level genome bins" (SGBs). Using SGBs as matching bins improves 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, population level datasets, and analytical tools like MetaPhlan, allow for research into microbiome geographic variation which has hitherto been greatly overlooked (Härer and Rennison 2023), especially for understudied taxa like liverworts.

In this study, we focused on the bacterial community of the liverwort *Calasterella californica* (Fig. 1, González-Ramírez et al. (2025a)). Liverworts are small plants with comparatively little tissue differentiation and many cells in direct contact with the environment. Like many plants, liverworts are known to have important associations with microorganisms, particularly fungi. *C. californica* has a geographical distribution, occurring in contrasting environments that range from wet redwood forests to the dry Anza-Borrego desert in Southern California (González-Ramírez et al. 2025b), making it an ideal system to investigate the variation of its associated bacterial community across the geographic space. In recent years, there has been an effort to understand the microbial communities associated with liverworts (*e.g.*, Wicaksono et al. 2023; Alcaraz et al. 2018; Kutschera and Koopmann 2005; Young et al. 2025) that point to potentially stronger associations than previously thought, and highlight the need for further investigation. By applying the off-target metagenomics approach to *C. californica* we provide new evidence on the potential tight relationships of liverworts and their bacteriomes, and address the question of how much these bacterial communities vary across the geographic space.

METHODS

Sampling.—The original sampling of *C. californica* consisted of 110 samples collected as part of the CCGP project (González-Ramírez et al. 2025a). This sampling focused on representing the broad geographic space and environmental conditions in which this liverwort occurs. The localities where the samples were collected were assigned to different California geographic regions

that largely reflect shared environmental conditions and shared geological history (Individual samples geographic regions are recorded in Sup. Tab. S.2). The Northern California region is characterized by higher precipitation and cooler temperatures. The Sierras region spans the interior mountain range of California, east of the Central Valley and it is characterized by a shared geological history, marked seasonal variation, and higher altitude. Central California samples span 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 Transverse Ranges, characterized by higher altitude, precipitation, and seasonality than the more coastal Peninsular Ranges. 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, 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 (Fig. 1 and Sup. Fig. S2).

We leveraged 97 of the CCGP samples to characterize *C. californica*'s bacteriome across the geographic space. Only 97 were used out of 110 liverwort collections because the remaining 13 samples did not pass the sequencing quality checks or were found to be a different liverwort species. Each liverwort sample was collected in a different geographic location. At each location, one to two clumps of liverworts were collected in hard plastic containers and transported to the lab. From each collection, the largest 1-3 thalli were isolated and cleaned by manually removing soil particles and most of the rhizoids and used for DNA extraction. The number of thalli used was based on securing 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. 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)).

DNA-extraction, sequencing, and quality checks.—For each sample, total genomic and symbiont DNA of one to four thalli was extracted using either the Qiagen DNeasy Plant Pro kit (California, USA) or a CTAB protocol. Library preparation and sequencing was carried out at the UC Davis DNA Technologies & Expression Analysis Core Laboratory. Libraries were sequenced using an Illumina NovaSeqX 300 (PE150) flow-cell. During ealy sequencing quality checks, all the samples displayed a bimodal GC-content distribution that characterizes the presence of both bacterial (high GC-content), and liverwort (low GC-content) DNA.

Reads processing and bacteriome characterization using MetaPhlAn.—All the paired-end files were quality checked and cleaned from adaptors using `Trimmommatic` (Bolger et al. 2014). 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). Next, 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, and these reads were used as the input to study the microbiome. To characterize the *C. californica* metagenome composition of each liverwort, we used `MetaPhlAn`, a recently developed pipeline that uses a compendium of metagenome-assembled genomes and reference genomes of microbes to characterize microbial communities and estimate relative abundances from shotgun sequence data (Blanco-Míguez et al. 2023). This approach differs from other common microbiome-targeting approaches such as 16S rRNA/18S rRNA/ITS sequencing or the assembly and annotation of full metagenome assembled genomes (MAGs) in that it does not attempt to assemble the microbial markers or genomes, but instead directly uses the short-reads to characterize the metagenome. We ran MetaPhlan on the Berkeley Computer Cluster SAVIO 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 a each step of this workflow is detailed in Sup. Tab. S.2.

Bacteriome composition and diversity.—All the statistical analyses and visualizations were performed using the `microeco` package (Liu et al. 2025) in R (R Core Team 2000). The output obtained from MetaPhlan was formatted as input for microeco using the R package `file2meco`

(Liu et al. 2022). To quantify the α-diversity for each sample we used multiple standard diversity indexes: Shannon (1948), Simpson (1949), Pielou (1966), and coverage, as implemented in `microeco`. We tested the potential effect of geographic region on these diversity metrics using an Analysis of Variance (ANOVA) (Fisher 1928).

To characterize the similarity among the bacteriomes of the different samples of *C. californica*, we calculated a Bray–Curtis dissimilarity matrix and computed a Non-Metric multidimensional scaling (NMDS) ordination. To assess whether samples belonging to the same geographic region were more similar in composition we performed a PerMANOVA, as implemented in `microeco` (Anderson 2001). Furthermore, we evaluated whether or not 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. 2011). 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.—Since macro-environmental variables might affect the bacteriome of *C. californica* by determining the pool of bacteria available in the environment, and because this liverwort occurs in a wide range of climatic conditions, we evaluated the effect 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 et al. 2013). Of the 19 variables in BioClim we selected a subset of variables with relatively low correlation and high variation within our geographic scope (Sup. Fig. S1. We used a Correspondence Analysis (CCA) to visualize the relationship between the climatic variables and the microbial composition of *C. californica* samples at the taxonomic family level (Ter Braak 1986), and we tested for the significance of this relationship, using a Mantel test implemented in `microeco` (Mantel 1967).

Functional composition.—In order to characterize the functional profile of the bacteriome associated to *C. californica*, we used the `FAPROTAX v1.2.10.` database (Louca et al. 2016a,b) to link the taxonomic composition of the bacteriome of each sample to functional diversity using the `microeco` built in function. Finally, we tested for differential abundance of these functions in different geographic regions using a linear discriminant analysis effect size (LefSE) with the relative abundance of the different functions in the bacteriome as input.

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.

RESULTS

Bioinformatics

Ninety-seven of the original 110 *C. californica* passed quality control and were used in our study (Sup. Tab. S.2). The number of sequences before filtering steps were on average around 30,600,000 reads. The smallest sample contained 13,028,666 reads and the largest contained 45,988,969 reads. After filtering out host *C. californica* reads, we were left with around 19,000,000 reads on average (61% of the total), with the smallest sample containing 3,835,559 reads and the largest 36,016,561 reads. When we profiled the microbiome using MetaPhlan, on average 96.63 percent of reads were unclassified and not used for downstream analysis. Our final analysis on classified microbiome reads is composed of on average around 640,000 reads per sample with the smallest sample containing 70,519 reads and the largest 3,610,608 reads (Sup. Tab. S.2).

For our dataset, the running time on a standard laptop using eight cores was about seven hours per sample. In our particular cluster environment (SAVIO, at UC Berkeley), 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 running the analysis sequentially for all samples. With this set up, each sample took about 30 min (20 cores simultaneously with 384 GB of RAM).

Composition of the bacterial microbiome of C. californica

Across the 97 samples of *C. californica*, the MetaPhlan pipeline identified 751 species level bacterial OTUs, belonging to 338 different genera, 141 families, 87 orders, 51 classes, and 15 phyla (Sup. Tab. S.3). Overall, the bacterial microbiome of *C. californica* is dominated by bacteria in the order Hypomicrobioales, often accompanied by other bacteria orders like Pseudomononadales, Pseudonocardiales, Burkholderiales, Mycobacteriales, Micrococcales, and Oscillatoriales (Fig. 2).

An ANOVA test did not find any of these α-diversity metrics to be significantly different between geographic regions (Sup. Tab. S.1). Nevertheless, a PERMANOVA test ($P < 001$, $R^2 = 0.127$,$F = 2.19$ $df = 6$) found that there is a weak but significant effect of the geographic region on explaining the patterns of β-diversity of the bacteriome of *C. californica* (Fig. 3). The linear discriminant effect size (LefSE) analysis revealed that there are 18 families and 50 genera of bacteria that have significantly differential abundance in different geographic regions (Fig. 4).

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 (via affecting the pool of bacteria in the environment) 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 (Fig. S2). A biplot visualization of the correspondence analysis shows an significant association of annual temperature with the main ordination axis according to a permutation analysis ($\chi^2 = 0.17$, $F = 2.39$, $P = 0.002$) , which also correlates with the abundance of the taxa shown in black arrows: Oscillatoriaceae, Sanguibacteriaceae, Calotrichaceae, Aestuariivirgaceae, Coleofasciculaceae, Dinobryaceae, Perlucidibacaceae, Leadbetterellaceae, FGB26132, and FGB52975 (Fig. 5). And the second ordination axis has a marginal association with temperature seasonality ($\chi^2 = 0.097$, $F = 1.35$, $P = 0.055$; Fig. 5). 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* (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 shows 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 (Fig. 6). 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 (Fig. S4). However, none of these taxa in any category comprised more than four percent of the total relative abundance (Fig. S3).

DISCUSSION

Despite living in contrasting regions across California (Fig. 1 and Fig. S2), the bacterial communities associated with *C. californica* are similar in their α-diversity metrics across all geographic regions (Tab. S.1). This overall similarity extends to their composition—*i.e.*, all bacterial communities are dominated by bacteria in the order Hyphomicrobiales, with important contributions from a few other orders, such as Pseudomononadales, Pseudonocardiales, Burkholderiales, Mycobacteriales, Micrococcales, and Oscillatoriales (Fig. 2). These results are congruent with previous work in liverworts of the genus *Riccia*, whose bacterial communities are not affected 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 surrounding soil. While comparative studies on the bacterial communities associated with liverworts are still scarce, these three studies together provide growing evidence that liverwort-associated bacterial communities are non-random and stable regardless of the environmental

conditions in which individual liverworts occur. This is in contrast 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).

Among the taxa that were consistently abundant in *C. californica* bacterial communities were bacteria 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* (Alcaraz et al. 2018), where they play an important role promoting the growth of this plant (Kutschera and Koopmann 2005). Hyphomicrobiales contain 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). The functional profiles of the bacteriomes were also very conserved across samples and geographic regions (Fig. 6) with high relative abundance of bacteria that perform methanol oxidation, methano- and methylotrophy, and aerobic chemoheterotrophy. These results highlight the role of that methylotrophic bacteria play as the dominant taxon in the *C. californica* microbiome. Whether and how liverworts actively maintain high proportion of these bacteria in their microbiome, in a dynamic and heterogeneous environment, warrants more investigation.

While the primary components of the microbiome of *C. californica* are the same across samples, when scrutinizing the identity of less abundant taxa, we found some differences across geographic regions (Fig. 4) potentially associated with variation in annual precipitation and temperature seasonality (Fig. 5). For example, higher abundance of photosynthetic bacteria in the families Oscillatoriaceae, Calotrichaceae, and Coleofasciculaceae was with higher annual precipitation. Desert samples had many significantly differentially abundant genera (Fig. 4), many of which—as reflected in the functional profile (Fig. S4)—are plant pathogens and human, animal, and gut associated (Fig. 4). One potential explanation is simply that the samples from the desert were collected near seasonal streams influenced by runoff and touristic human activities. On the other hand, previous work in angiosperms has documented shifts of bacteria composition associated with water availability (*e.g.*, Chao et al. 2025) and disease response (*e.g.*, Gao et al. 2021), suggesting some mechanisms that might be affecting these secondary bacterial components of

C. californica microbiome. Further work would need to be done in order to understand the cause of relative abundance differences in minor components of *C. californica*'s microbiome in the most extreme sampled localities.

Using a novel off-target meta-genomics approach, we were able to characterize the bacterial composition of the *C. californica* microbiome. Although a large fraction of off-target (*i.e.*, non-host related) reads were unclassified, this reflects MetaPhlan's focus on bacterial metagenomes, while we expect to have reads associated with other groups such as fungi, small metazoans, and viruses. With adequate analytical tools, these non-bacterial reads represent exciting opportunities to investigate other organisms associated with this plant. A real limitation of this approach—which is shared among all metagenomic studies—is the incompletness of the reference databases, which are potentially underrepresenting the diversity of bacteria in uncommonly studied environments. This characteristic of our databases highlights the need for continuous work on building larger genomic datasets (whether they are MAGs, 16S, SGBs).

While our approach, by definition, leverages data that was not aimed to study bacterial communities, results obtained with this approach are complementary to other microbiome targeted approaches, and particularly valuable to identify questions that require more targeted research. Furthermore, the congruent results of our study with previous research on liverwort microbiome studies that used 16S rRNA metabarcoding (Wicaksono et al. 2023; Alcaraz et al. 2018) is not only interesting insofar that 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. Additionally, shotgun sequencing, such as our study, has been found to have more power in identifying rare taxa than 16S (Durazzi et al. 2021), as it is not limited by primer amplification bias (Campanaro et al. 2018), and has greater taxonomic resolution than 16S (Laudadio et al. 2018).

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. At the same time, population genomic studies that use short-read shotgun resequencing approaches are becoming more common. 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 made from utilizing these types of landscapes genomic datasets, and these insights are especially impactful for groups relatively understudied, as are liverworts. As the field moves toward long-read sequencing, this paper shines a spotlight to a potentially large body of microbiome research that can investigated that would otherwise be unused.

Our results are consistent with previous knowledge on liverworts bacterial microbiomes generated through standard methods to study microbiomes, pointing to the efficacy of these data-mining strategies to do exploratory studies that lead to targeted research questions on the assembly of microbial communities. In the case of liverwort bacteriomes, it is increasingly exciting to conduct sampling and *in vitro* studies aimed to understanding the likely active role that liverworts play in the assembly of its bacterial communities.

Acknowledgements

The authors thank Carl Rothfels for his constructive comments and edits to the manuscript. Thank you to two anonymous reviewers and two editors for their thoughtful and constructive feedback of this manuscript. Thank you to 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]. IGR was supported by a UC Mexus-CONACyT fellowship (number 709967), a Plant Science Fellowship by Oak Spring Garden Foundation, and the Philomathia Graduate Fellowship in Environmental Sciences, at UC Berkeley. This research used the Savio computational cluster resource provided by the Berkeley Research Computing program at the University of California, Berkeley (supported by the UC Berkeley Chancellor, Vice Chancellor for Research, and Chief Information Officer).

Author contributions

IGR and MJS designed the study, performed the analyses and wrote the first draft of the manuscript. ECM contributed to analyses and edition of following versions of the manuscript. BDM supervised the work and edited the manuscript.

Data availability statement

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

REFERENCES

Alcaraz, L. D., Peimbert, M., Barajas, H. R., Dorantes-Acosta, A. E., Bowman, J. L., and Arteaga-Vázquez, M. A. (2018). Marchantia liverworts as a proxy to plants' basal microbiomes. *Scientific Reports*, 8(1):12712.

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, 26(1):32–46.

Benham, P. M., Cicero, C., Escalona, M., Beraut, E., Marimuthu, M. P., Nguyen, O., Nachman, M. W., and Bowie, R. C. (2023). A highly contiguous genome assembly for the california quail (*Callipepla californica*). *Journal of Heredity*, 114(4):418–427.

Beninde, J., Toffelmier, E., and Shaffer, H. B. (2022). A brief history of population genetic research in california and an evaluation of its utility for conservation decision-making. *Journal of Heredity*, 113(6):604–614.

Blair, K., Rose, S., Hull, J., Escalona, M., Finger, A., Joslin, S. E., Sahasrabudhe, R., Marimuthu, M. P., Nguyen, O., Chumchim, N., Morris, E. R., et al. (2022). The reference genome of the vernal pool tadpole shrimp, *Lepidurus packardi*. *Journal of Heredity*, 113(6):706–711.

Blanco-Míguez, A., Beghini, F., Cumbo, F., McIver, L. J., Thompson, K. N., Zolfo, M., Manghi, P., Dubois, L., Huang, K. D., Thomas, A. M., et al. (2023). Extending and improving metagenomic taxonomic profiling with uncharacterized species using `MetaPhlAn 4`. *Nature biotechnology*, 41(11):1633–1644.

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120.

Campanaro, S., Treu, L., Kougias, P. G., Zhu, X., and Angelidaki, I. (2018). Taxonomy of anaerobic digestion microbiome reveals biases associated with the applied high throughput sequencing strategies. *Scientific Reports*, 8(1):1926.

Chao, C., Xu, D., Jiang, B., Lu, X., Yu, C., Wang, Y., Wang, H., Li, J., and Zhu, J. (2025). Precipitation-driven restructuring of rhizosphere microbiota enhances alpine plant adaptation. *Frontiers in Plant Science*, 16:1641511.

Durazzi, F., Sala, C., Castellani, G., Manfreda, G., Remondini, D., and De Cesare, A. (2021). Comparison between 16s rrna and shotgun sequencing data for the taxonomic characterization of the gut microbiota. *Scientific Reports*, 11(1):3030.

Fick, S. E. and Hijmans, R. J. (2017). Worldclim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, 37(12):4302–4315.

Fiedler, P. L., Erickson, B., Esgro, M., Gold, M., Hull, J. M., Norris, J. M., Shapiro, B., Westphal, M., Toffelmier, E., and Shaffer, H. B. (2022). Seizing the moment: the opportunity and relevance of the california conservation genomics project to state and federal conservation policy. *Journal of Heredity*, 113(6):589–596.

Fisher, R. A. (1928). *Statistical methods for research workers*. Number 5. Oliver and Boyd.

Gao, M., Xiong, C., Gao, C., Tsui, C. K., Wang, M.-M., Zhou, X., Zhang, A.-M., and Cai, L. (2021). Disease-induced changes in plant microbiome assembly and functional adaptation. *Microbiome*, 9(1):187.

González-Ramírez, I. S., Escalona, M., Miller, C., Chumchim, N., Marimuthu, M., Nguyen, O., Song, M. J., and Mishler, B. D. (2025a). A genome assembly for the california endemic liverwort *Calasterella californica*. *Journal of Heredity*, 116(3):389–396.

González-Ramírez, I. S., Long, D. G., Briscoe, L., and Mishler, B. D. (2025b). Typification of *Fimbraria californica* (*Calasterella californica*, Aytoniceae). *Madroño*, 72(1):37–39.

Goodrich, J. K., Di Rienzi, S. C., Poole, A. C., Koren, O., Walters, W. A., Caporaso, J. G., Knight, R., and Ley, R. E. (2014). Conducting a microbiome study. *Cell*, 158(2):250–262.

Grether, G. F., Beninde, J., Beraut, E., Chumchim, N., Escalona, M., MacDonald, Z. G., Miller, C., Sahasrabudhe, R., Shedlock, A. M., Toffelmier, E., et al. (2023). Reference genome for the american rubyspot damselfly, *Hetaerina americana*. *Journal of Heredity*, 114(4):385–394.

Han, Y., He, J., Li, M., Peng, Y., Jiang, H., Zhao, J., Li, Y., and Deng, F. (2024). Unlocking the potential of metagenomics with the pacbio high-fidelity sequencing technology. *Microorganisms*, 12(12):2482.

Härer, A. and Rennison, D. J. (2023). The biogeography of host-associated bacterial microbiomes: Revisiting classic biodiversity patterns. *Global Ecology and Biogeography*, 32(6):931–944.

He, X., Wang, D., Jiang, Y., Li, M., Delgado-Baquerizo, M., McLaughlin, C., Marcon, C., Guo, L., Baer, M., Moya, Y. A., et al. (2024). Heritable microbiome variation is correlated with source environment in locally adapted maize varieties. *Nature Plants*, 10(4):598–617.

Hijmans, R. J., Van Etten, J., Mattiuzzi, M., Sumner, M., Greenberg, J., Lamigueiro, O., Bevan, A., Racine, E., and Shortridge, A. (2013). Raster package in r. *Version https://mirro rs sjtug sjtu edu cn/cran/web/packa ges/rast er/raster pdf*.

Huang, Y., Escalona, M., Morrison, G., Marimuthu, M. P., Nguyen, O., Toffelmier, E., Shaffer, H. B., and Litt, A. (2022). Reference genome assembly of the big berry manzanita (*Arctostaphylos glauca*). *Journal of Heredity*, 113(2):188–196.

Kutschera, U. and Koopmann, V. (2005). Growth in liverworts of the marchantiales is promoted by epiphytic methylobacteria. *Naturwissenschaften*, 92(7):347–349.

Laudadio, I., Fulci, V., Palone, F., Stronati, L., Cucchiara, S., and Carissimi, C. (2018). Quantitative assessment of shotgun metagenomics and 16s rdna amplicon sequencing in the study of human gut microbiome. *OMICS: A Journal of Integrative Biology*, 22(4):248–254.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *arXiv preprint arXiv:1303.3997*.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, . G. P. D. P. (2009). The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079.

Lindström, K. and Mousavi, S. A. (2020). Effectiveness of nitrogen fixation in rhizobia. *Microbial biotechnology*, 13(5):1314–1335.

Liu, C., Li, X., Mansoldo, F. R., An, J., Kou, Y., Zhang, X., Wang, J., Zeng, J., Vermelho, A. B., and Yao, M. (2022). Microbial habitat specificity largely affects microbial co-occurrence patterns and functional profiles in wetland soils. *Geoderma*, 418:115866.

Liu, C., Mansoldo, F. R., Li, H., Vermelho, A. B., Zeng, R. J., Li, X., and Yao, M. (2025). A workflow for statistical analysis and visualization of microbiome omics data using the r microeco package. *Nature Protocols*, pages 1–25.

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

Louca, S., Parfrey, L. W., and Doebeli, M. (2016b). Decoupling function and taxonomy in the global ocean microbiome. *Science*, 353(6305):1272–1277.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer research*, 27(2_Part_1):209–220.

Mead, A., Fitz-Gibbon, S. T., Escalona, M., Beraut, E., Sacco, S., Marimuthu, M. P., Nguyen, O., and Sork, V. L. (2024). The genome assembly of island oak (quercus tomentella), a relictual island tree species. *Journal of Heredity*, 115(2):221–229.

Pebesma, E. (2018). Simple features for r: standardized support for spatial vector data.

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

Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S., and Huttenhower, C. (2011). Metagenomic biomarker discovery and explanation. *Genome Biology*, 12(6):R60.

Shaffer, H. B., Toffelmier, E., Corbett-Detig, R. B., Escalona, M., Erickson, B., Fiedler, P., Gold, M., Harrigan, R. J., Hodges, S., Luckau, T. K., et al. (2022). Landscape genomics to enable conservation actions: the california conservation genomics project. *Journal of Heredity*, 113(6):577–588.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Simpson, E. H. (1949). Measurement of diversity. *Nature*, 163(4148):688–688.

Song, M. J., Freund, F., Tribble, C. M., Toffelmier, E., Miller, C., Bradley Shaffer, H., Li, F.-W., and Rothfels, C. J. (2025a). The nitrogen-fixing fern *Azolla* has a complex microbiome characterized by varying degrees of cophylogenetic signal. *American Journal of Botany*, 112(3):e70010.

Song, M. J., Pirro, S., Lahmeyer, S., Kyle, K., and White, M. (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(4):334–336.

Song, M. J., Rizzieri, Y. C., Li, F.-W., Freund, F., Escalona, M., Toffelmier, E., Miller, C., Shaffer, H. B., Nguyen, O., Marimuthu, M. P., et al. (2025b). The genome assembly of the duckweed fern, *Azolla caroliniana*. *Journal of Heredity*, page esaf022.

Steinegger, M. and Salzberg, S. L. (2020). Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in genbank. *Genome Biology*, 21(1):115.

Supple, M. A., Escalona, M., Alexandre, N., Buchalski, M. R., Riley, S. P., Dellinger, J. A., Vickers, T. W., Sahasrabudhe,

R., Nguyen, O., Fairbairn, C. W., et al. (2024). A chromosome-level genome assembly of the mountain lion, *Puma concolor*. *Journal of Heredity*, page esae063.

Team, R. C. (2000). R language definition. *Vienna, Austria: R foundation for statistical computing*, 3(1):116.

Ter Braak, C. J. (1986). Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology*, 67(5):1167–1179.

Toffelmier, E., Beninde, J., and Shaffer, H. B. (2022). The phylogeny of california, and how it informs setting multi-species conservation priorities. *Journal of Heredity*, 113(6):597–603.

Vandenkoornhuyse, P., Quaiser, A., Duhamel, M., Le Van, A., and Dufresne, A. (2015). The importance of the microbiome of the plant holobiont. *New Phytologist*, 206(4):1196–1206.

Wagner, M. R., Lundberg, D. S., Del Rio, T. G., Tringe, S. G., Dangl, J. L., and Mitchell-Olds, T. (2016). Host genotype and age shape the leaf and root microbiomes of a wild perennial plant. *Nature Communications*, 7(1):12151.

Wei, N. and Tan, J. (2023). Environment and host genetics influence the biogeography of plant microbiome structure. *Microbial Ecology*, 86(4):2858–2868.

Wicaksono, W. A., Semler, B., Pöltl, M., Berg, C., Berg, G., and Cernava, T. (2023). The microbiome of *Riccia* liverworts is an important reservoir for microbial diversity in temporary agricultural crusts. *Environmental Microbiome*, 18(1):46.

Wiśniewski, P., Maździarz, M., Sawicki, J., and Krawczyk, K. (2025). Taxonomic and functional differentiation of soil and thallus microbiomes in *Riccia sorocarpa*. *Biologia*, pages 1–12.

Wood, D. A., Richmond, J. Q., Escalona, M., Marimuthu, M. P., Nguyen, O., Sacco, S., Beraut, E., Westphal, M., Fisher, R. N., Vandergast, A. G., et al. (2022). Reference genome of the california glossy snake, *Arizona elegans occidentalis*: a declining california species of special concern. *Journal of Heredity*, 113(6):632–640.

Young, B. C., Thiers, B., White, J. F., and Struwe, L. (2025). Endophytic bacteria discovered in oil body organelles of the liverworts *Marchantia polymorpha* and *Radula complanata*. *American Journal of Botany*, 112(3):e70017.

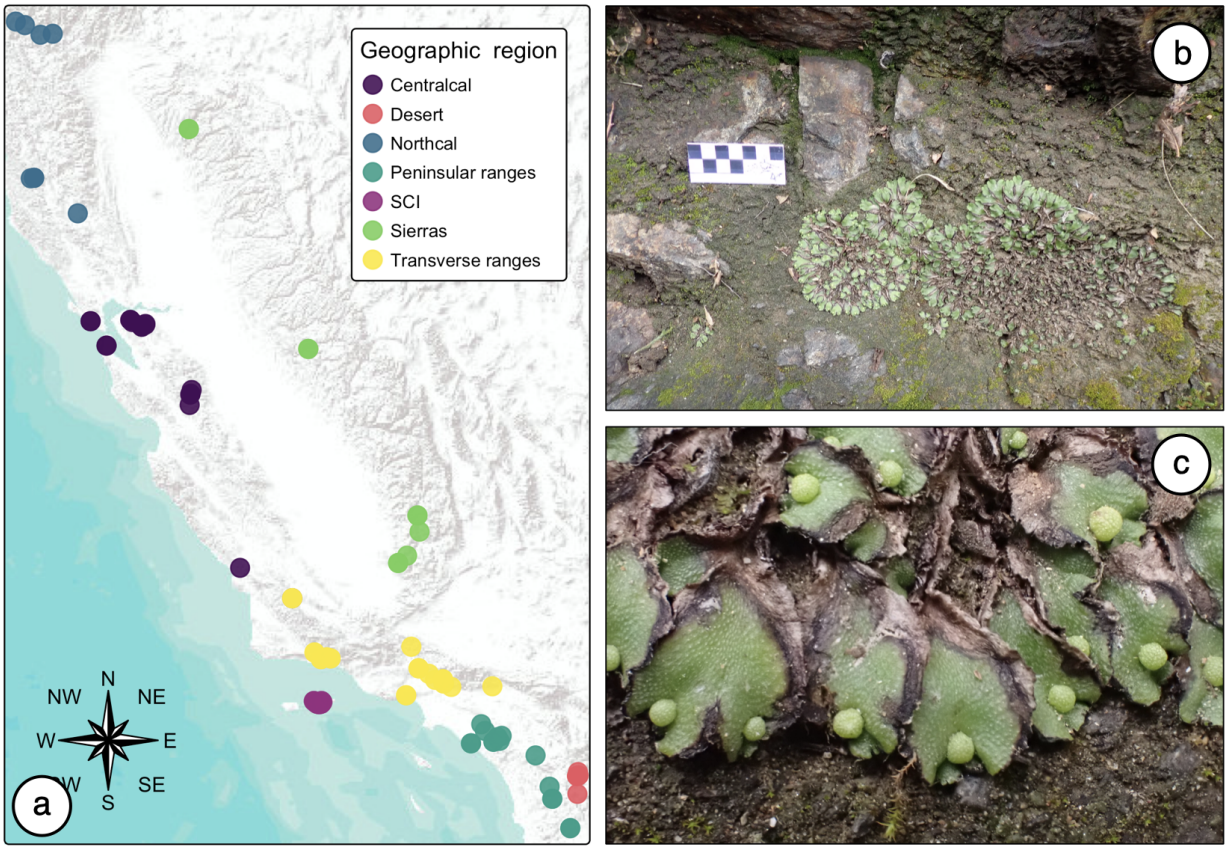

Figure 1: a. Collection locations of *Calasterella californica* samples for this study, dots are color coded by their California geographic region. **b.** Typical morphology, *C. californica* grows on bare and exposed soil or rocks, forming circular mats. **c.** Close up to female thalli of *C. californica* displaying the proximity of this plant with the soil.

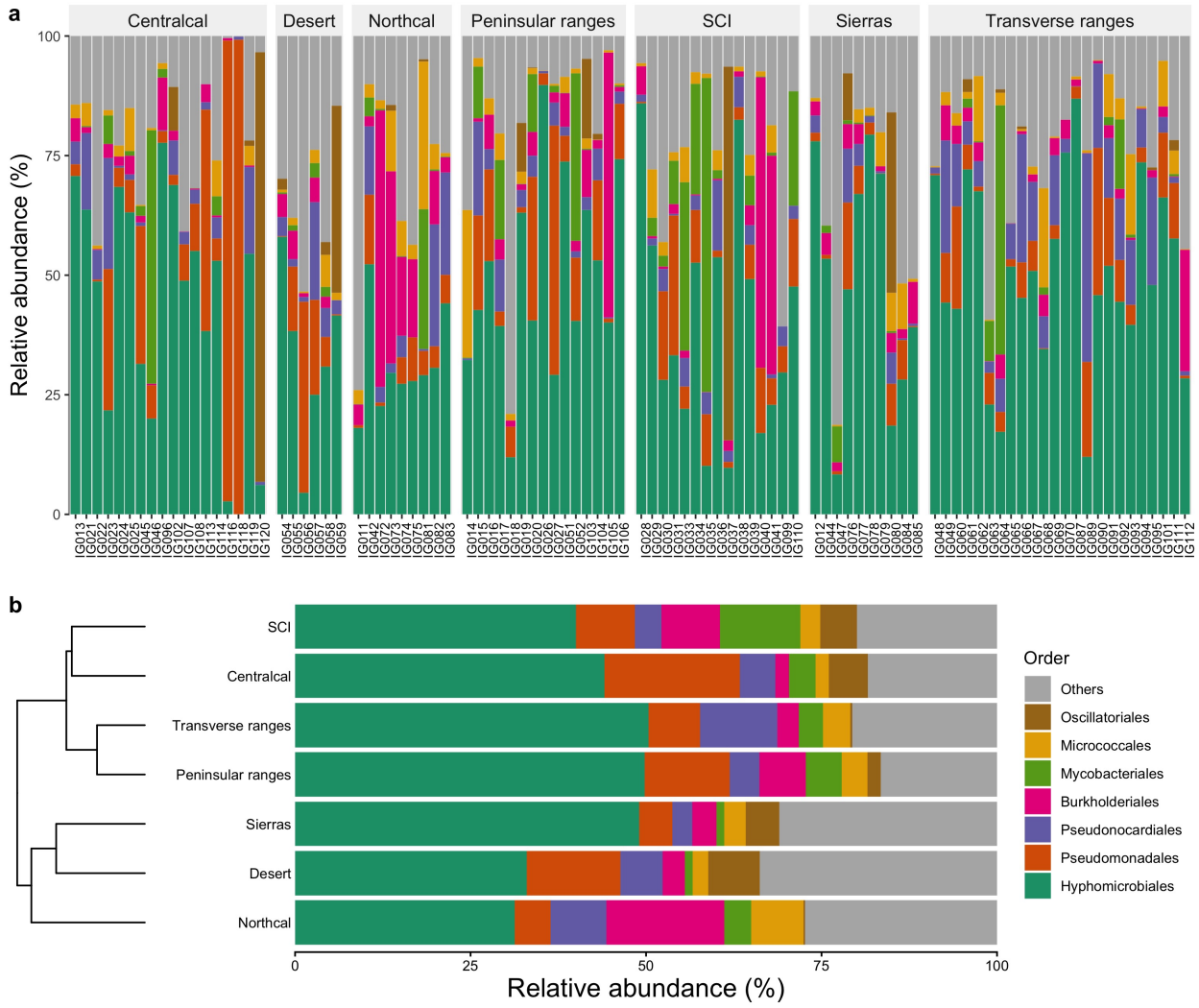

Figure 2: Composition of the bacteriome of *C. californica*. a. Relative abundance of the seven most abundant orders of bacteria across all the samples of *C. californica*. The samples are arranged by the geographic region where they were collected. **b.** Composition grouped by geographic region. The 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.

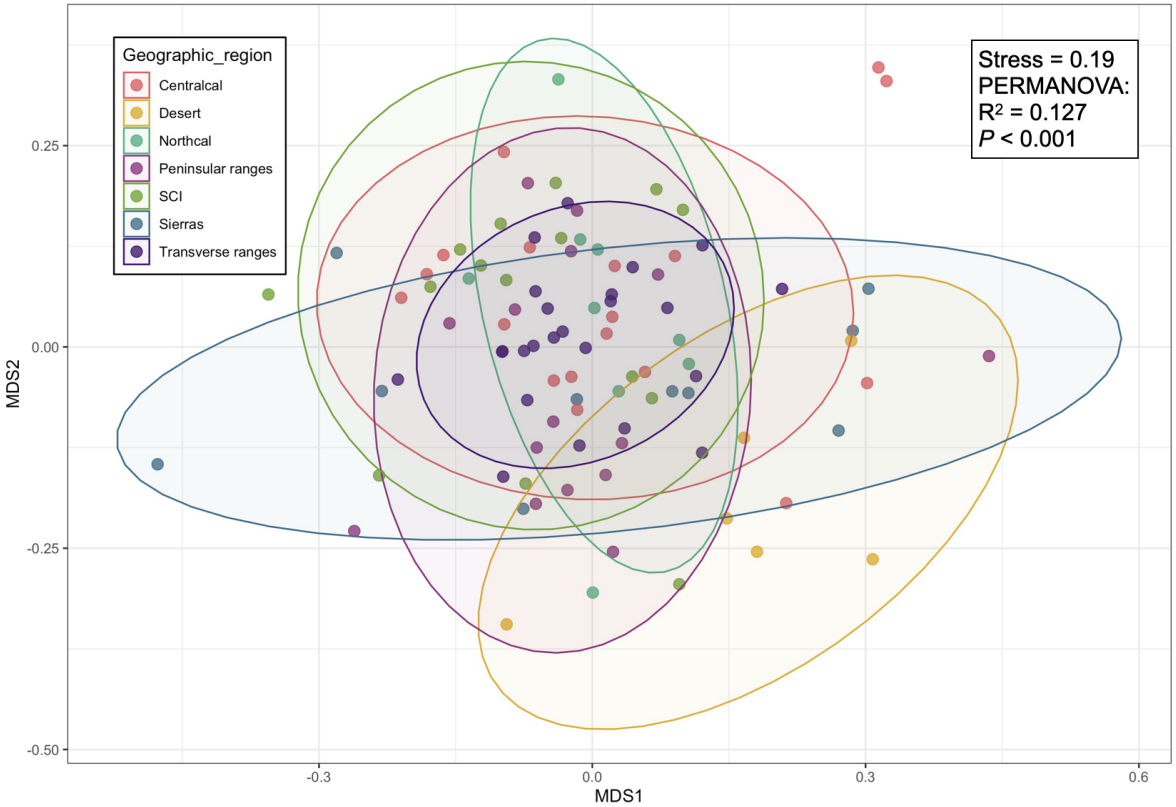

Figure 3: Non-metric multidimensional scaling (NMDS) ordination of *C. californica* associated bacterial communities based on Bray–Curtis dissimilarities. Each point represents a different sample color-coded by the geographic region it was collected. While the biplot ordination does not show a strong arrangement by geographic region, a PERMANOVA test finds a weak but significant effect of the region on the composition of the bacterial community.

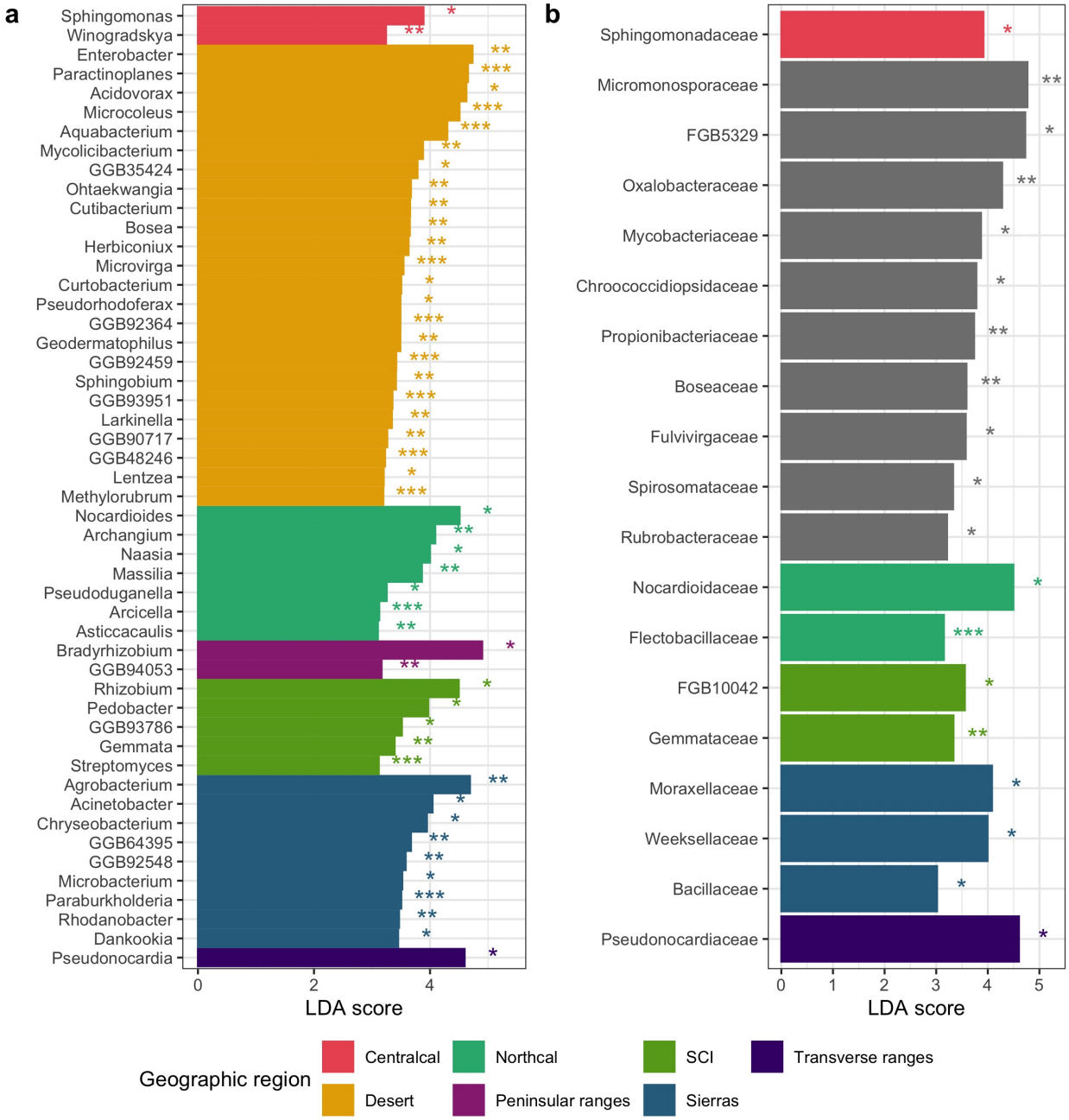

Figure 4: Significant differential abundance of bacterial (a) genera and (b) families across geographic regions. There are 18 families and 50 genera of bacteria with significantly different abundances across geographic regions. Differential abundance was determined based on LDA scores > 3. Significance values are signaled with asterisks such as $P < 0.5*$; $P < 0.1**$, $P < 0.01***$. When differential abundance is characteristic of a single region, the bar is colored by such region, otherwise the bar is grey. Grey bars imply that more than one region has differential abundance of this taxon.

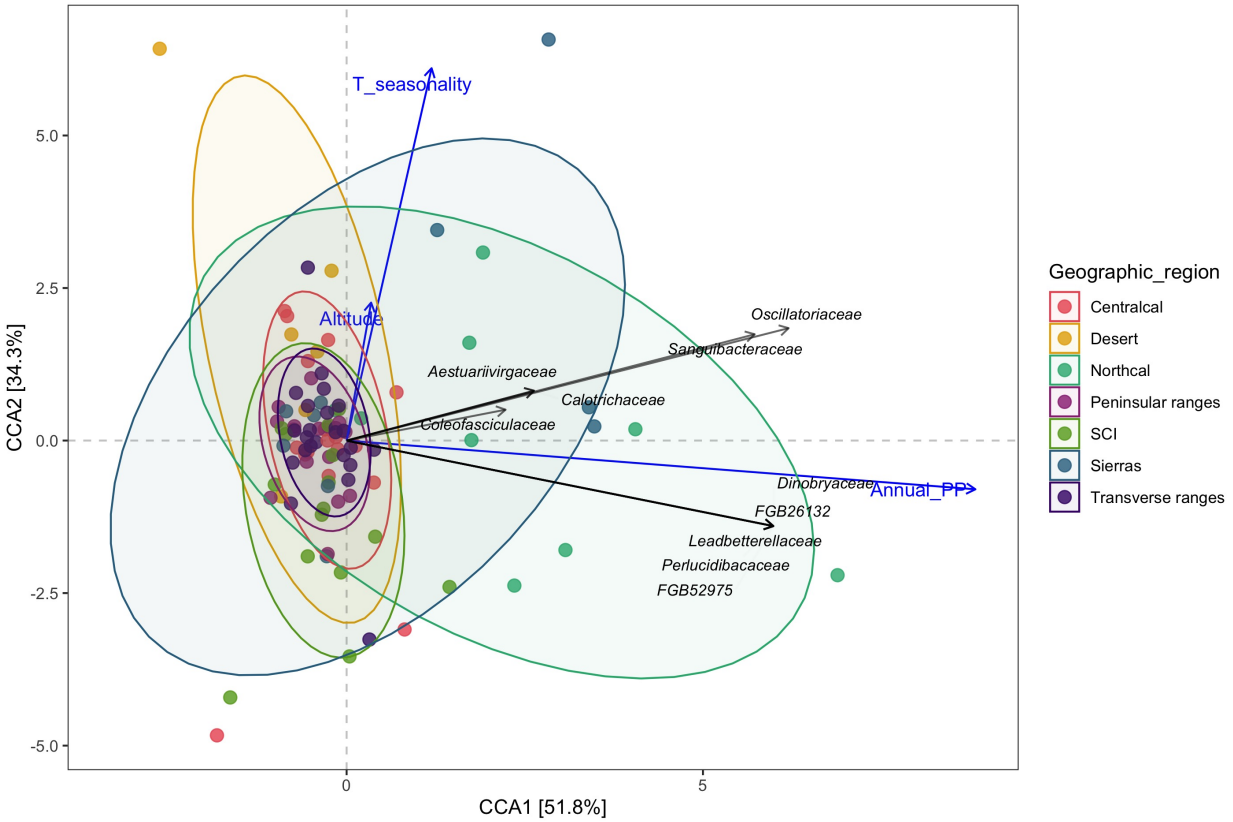

Figure 5: Correspondence Analysis (CCA) of the bacterial communities of *C. californica*. The blue arrows show the directionality of the variation of climatic variables in the multivariate space. The black arrows show the main axis of variation of bacterial families.

Figure 6: Functional composition of the bacteriome of *C. californica*. Relative abundance of bacterial functions in the community of bacteria associated to 97 samples of *C. californica*.