

Functional exploration of natural product  
metabolism and metatranscriptomic analyses in  
uncultivated soil bacteria from the Talabre-Lejía  
transect (Atacama Desert)

by

Constanza M. Andreani-Gerard

submitted in partial satisfaction of the requirements for the degree of

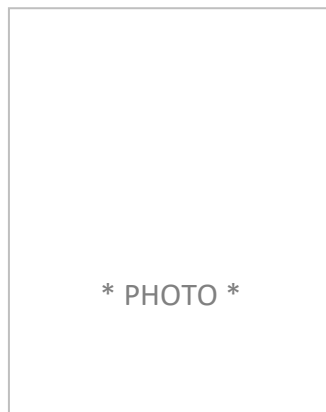
Master of Science

in

Biology



Facultad de Ciencias  
Escuela de Postgrado  
Universidad de Chile  
September 30<sup>th</sup>, 2022



Licenciada en Agronomía y Ciencias de los Recursos Naturales en la P. Universidad Católica de Chile (2016) con mención en Ciencias Vegetales y candidata, mediante la presente tesis, al grado de magíster en Ciencias Biológicas en la Universidad de Chile (2022). Con experiencia previa ligada a la educación ambiental (huertos urbanos, vermicompostaje y manejo de otros residuos domiciliarios) y al desarrollo de productos (abonos fermentados, biofertilizantes foliares y muebles de plástico reciclado), hoy se advoca profesionalmente a las ciencias bioinformáticas, en particular, a la metagenómica microbiana y al metabolismo de productos naturales. Desempeña los cargos de analista de bases de datos biológicas para el consorcio CEODOS Chile, vinculada vía el Centro de Regulación del Genoma (CGR), y de asistente de investigación en proyecto FONDECYT del Laboratorio de Bioinformática y Expresión Génica (LBEG) del Instituto de Nutrición y Tecnología de los Alimentos (INTA).

FACULTAD DE CIENCIAS  
UNIVERSIDAD DE CHILE

INFORME DE APROBACIÓN

TESIS DE MAGÍSTER

Se informa a la Escuela de Postgrado de la Facultad de Ciencias que la Tesis de Magíster presentada por la candidata:

Constanza María Andreani Gerard,

ha sido aprobada por la Comisión de Evaluación de tesis como requisito para optar al grado de **Magíster en Ciencias Biológicas** en el examen de Defensa Privada rendido el día 20 de octubre de 2022.

Director de Tesis:

Dr. Mauricio González

Comisión de Evaluación de la Tesis:

Dra. Inmaculada Vaca

Dr. Andrés Marcoleta

To my parents, Marcelo and Claudia,  
and to my grandmother, María Luisa.

## Acknowledgements

This project was possible thanks to the contributions of several people and institutions in my formative process. Thanks to:

Dr. Mauricio González for taking me in as his student, for encouraging me to accept new challenges, for his trust by giving me autonomy in the definition and execution of the experimental design presented, and for allowing me to discover that my curiosity is rare and that scientific research is a way of making art.

Dr. Christian Hodar for his technical advice regarding biostatistical methods and for teaching me to robustly discern between approaches for large-scale data management.

Dr. Alexis Gaete for his mentoring and support in my familiarization with the wet lab.

The entire group of the Bioinformatics and Genetic Expression Laboratory (LBEG) of the Nutrition and Food Technology Institute (INTA) for their generosity by sharing with me the biological material hereby studied, product of the hard work of many researchers over the years.

Dr. Alejandro Maass for giving me access to the computational resources (Leftraru) of the National Laboratory for High Performance Computing (NLHPC) to execute several bioinformatic steps.

Ricardo Palma for his support in software installations, database downloads and linux navigation tips.

Dr. Dante Travisany for giving me access to a server harbouring FastQC, a licenced tool for quality control of reads, and for being the one introducing me to RNA-seq data handling back in 2020.

My colleagues Silver Ceballos and Glen Yupanki for selflessly giving me their critical opinion, involved assessment, and unmeasured time.

My brothers Federico Andreani, for making the black screen a delightful and less lonely immersion, and Domingo Andreani, for reading this manuscript upon very short notice and for reminding me to trust myself.

Finally, my family and friends for their caring companionship throughout the years.

This study was funded by ANID FONDECYT grants 1201278 and by the Centre for Genome Regulation (CGR), a Millennium Institute Project supported by the ANID Millennium Scientific Initiative (ICN2021-044, Chile).

## Abstract

Soils are heterogeneous and complex environments, and bacterial competition, cooperation, and communication dynamics there are mediated through the secretion of natural products (NPs) synthesized in specialized pathways allowing diverse survival strategies to arise in oligotrophic conditions. Due to the vast array of thriving mechanisms, classification of compound classes and comparative approaches turn out necessary means to assess specialized metabolism (SM) in evolutionary and ecological contexts. Complementing current advances in the field of genome mining with transcriptional data enables better understanding of how often these metabolites are produced in natural settings and what stimuli set their production off. Here, a dataset of 190 biosynthetic gene clusters (BGCs) of mainly unknown functions was obtained from six metagenomic samples of the Atacama Desert revealing site- and/or phylum-specific behaviours. *Acidobacteria* was found to be the most abundant and the most SM-enriched taxa in these soils. A complete hybrid region of ~130 kb was fully predicted making it one of the largest to be directly recovered from an environmental sample. This is the first report of NP-encoding bacteria from *Lentisphaerae* and *Spirochaetes* taxa. Examination of functional annotation of essential and accessory specialized genes showed association between biosynthetic classes of compounds and categories of orthologs such as NRPS/T1PKS with transposases and binding protein-dependant transport systems, and terpenes with the arsenic regulator ArsR and the iron transporter TonB. Manual curation of protein family domains of reference BGCs from gene cluster families (GCFs) suggest that shared biological functions of the Talabre-Lejía transect are mainly advocated to antibiotic biosynthesis, nitrogen metabolism, oxidative stress, and metal resistance.

This study provides insights on functional co-occurrence patterns of NP-encoding repertoires obtained from genomic analyses of 53 metagenome-assembled drafts (MAGs) recovered from a rare natural environment throughout a scalable pipeline constructed upon considerations for researchers with no bioinformatic background.

# Table of contents

<b>Lists of figures and tables</b> .....	<b>8</b>
<b>Introduction</b> .....	<b>9</b>
Hypothesis and objectives of the study .....	<b>12</b>
<b>Background of biological-derived data</b>	
Obtention of MAGs .....	<b>13</b>
Obtention of RNA reads .....	<b>15</b>
<b>Methods</b>	
Genome-resolved functional comparative analyses of specialized genes .....	<b>17</b>
Natural product-oriented transcriptional analyses .....	<b>18</b>
<b>Results</b>	
Taxonomical and geographical (un)specificities of BGCs repertoires ...	<b>22</b>
Adjustments to quality-filtering parameters for RNA-seq input data ..	<b>30</b>
<b>Discussion</b>	
Natural product biosynthetic capabilities serve as marks for enlightening ecological and evolutionary patterns .....	<b>37</b>
A specialized and bioinformatically viable metatranscriptomic pipeline .....	<b>43</b>
<b>Conclusions</b> .....	<b>47</b>
<b>References</b> .....	<b>49</b>
<b>Extended data</b> .....	<b>57</b>
<b>Supplementary information</b> .....	<b>61</b>

## List of figures

Figure M1 .....	18
Figure M2 .....	21
Figure 1 .....	23
Figure 2 .....	25
Figure 3 .....	26
Figure 4 .....	27
Figure 5 .....	29
Figure 6 .....	31
Figure 7 .....	32
Figure 8 .....	35
Figure 9 .....	36
Supplementary Figure E1 .....	57
Supplementary Figure E2 .....	59
Supplementary Figure S1 .....	61
Supplementary Figure S2 .....	62

## List of tables

Table M1 .....	19
Table 1 .....	24
Table 2 .....	28
Table 3 .....	30
Tables 4 and 5 .....	34
Supplementary Table E1 .....	58
Supplementary Tables E2 – E9 .....	60
Supplementary Table S1 .....	61
Supplementary Tables S2 – S4 .....	62
Supplementary Tables S5 and S6 .....	63



## Introduction

Soils are highly heterogeneous and complex environments, and bacteria that inhabit them play multiple intra- and inter-specific ecological roles accounting for competition, cooperation, and communication dynamics through the secretion of specialized metabolites, of which most remain uncharacterized<sup>[1,2]</sup>. Known biological functions performed by these highly optimized molecules<sup>[3]</sup> include induction of motility and/or biofilm production<sup>[4,5]</sup>, tolerance to various forms of abiotic stress through, for example, pigments and ectoines<sup>[6-8]</sup>, cycling of nutrients by capturing them or making them bioavailable<sup>[9,10]</sup>, and inhibition of growth via antibiotics and antifungals<sup>[11,12]</sup>, among others. The latter are main actors in predator and prey dynamics in the *microbial jungle*<sup>[1]</sup> providing an advantage to producer strains when competing against susceptible ones for the same pool of resources<sup>[13]</sup>. Notwithstanding, it is also known that antimicrobial compounds might act as signalling molecules<sup>[14]</sup> and stimulators of sporulation<sup>[15]</sup> at subinhibitory concentrations. Put this way, characterizing these natural products (NPs) as “secondary” seems inaccurate as pointed out many times<sup>[16,17]</sup>, even though it is mainstream slang.

Hence, specialized metabolite-mediated interactions can influence evolutionary fitness landscapes by closely referring to adaptative strategies. Congruently, pathways involved in natural product biosynthesis show more restricted taxonomic distributions and admit a greater metabolic diversity compared to those involved in autonomous growth<sup>[18]</sup>. Such diversity is, at least in part, product of enzymatic promiscuity which offers proteins more flexibility in responding to different and evolving environments through the acceptance of more than one substrate<sup>[19]</sup>. This idea is clearer when considering that the ability to accept more ligands requires fewer mutations for substrate specificities or stability of protein configurations to be altered and, thus, for new functions to be coined<sup>[20]</sup>. From a genomic lens, structural modularity of genes

encoding the enzymes involved in specialized metabolism somehow favours the maintenance of such mutations by consisting of several adjacent genes accounting for essential steps in biosynthesis and many times for regulation, transport, tailoring and other accessory functions as well<sup>[21]</sup>. Thus, even if subtle, modifications that these biosynthetic gene clusters (BGCs) might undergo and their cascading effects on possibly co-regulated<sup>[22]</sup> enzymatic mechanisms can change the final compound. Moreover, the combinatorial nature of tailoring steps, in which intermediate products can influence the molecular characteristics of the biosynthetic pathway itself<sup>[23,24]</sup>, extends promiscuity above described for enzymes to the reactions they catalyse. Good example of this is the case of ribosomally synthesized and post-translationally modified peptides (RiPPs), in which the final number of synthesized molecules varies depending on the original substrate(s) indistinctly from the enzymes of the pathway itself<sup>[25,26]</sup>. Other classes of enzymes such as nonribosomal peptide synthases (NRPSs) and polyketide synthases (PKSs) exploit modular architecture with multiple repeats of domains and exhibit genetic mechanisms of duplications, natural hybridizations, insertions, and deletions of these gene modules<sup>[21,27]</sup>. Striking structural similarities between NRPSs and PKSs allow the formation of clusters containing genes encoding enzymes of both classes<sup>[28]</sup>. These hybrids regions can sometimes extend over 100,000 bp in length (“superclusters”) and translate into thousands of amino acids (“megaenzymes”)<sup>[29]</sup> as is the case of rapamycin from *Streptomyces hygroscopicus* isolated from Rapa Nui Island in 1995 by a British group<sup>[30]</sup> before any legal regulation regarding origins of biological resources was considered in Chile.

Precisely due to the diverse array of thriving mechanisms, classification of compound classes and comparative approaches turn out necessary means to assess specialized metabolism in evolutionary and ecological contexts<sup>[16,31]</sup>. Main classes of specialized metabolites are assigned according to few clearly defined types of core enzymes<sup>[2]</sup>, encoded by genomic building blocks with distinguishable features, making it data of highly predictive quality. Thereby, genome mining of BGCs has been organically established as key step<sup>[32]</sup> in pretty much every current pipeline advocated to natural product research. Fast accumulation of high-throughput genomic and metagenomic

data led to the emergence of several BGC-specific databases<sup>[33-35]</sup>. The former allowed a structured and standardized platform for sharing data with the scientific community and for the development of custom bioinformatic pipelines upon functional annotation of specialized genes for study-specific downstream analyses, theme of major relevance when dealing with environmental samples —as the ones processed in this study— given the lack of dedicated tools<sup>[29]</sup>. Namely, obstacles arise from immensely larger datasets than those obtained from organisms isolated in laboratory conditions, where each sequencing delivers information of one genome, as environmental datasets such as those that would be obtained from a handful of dirt or a cup of sea water harbour information of hundreds, if not thousands. This way, even though genomic analyses from cultivated bacteria enable functional validation of *in silico* predictions, the study of whole metagenomes and draft genomes assembled from metagenomes (MAGs) enable more accurate description of a community's functionality<sup>[36]</sup> while having higher computing requirements. Besides, little is known about how often specialized metabolites are produced or how the environment regulates their production<sup>[31]</sup> turning metatranscriptomic data into critical biological input to understand transcriptional behaviours of BGCs in natural settings. Lastly, a new era of antibiotic discovery begun when teixobactin was isolated in 2015 by being the first antimicrobial obtained from uncultivated bacteria<sup>[37]</sup> stepping ahead of culture-based techniques that still have to overcome tight regulation of BGC expression and difficulties related to mimicking natural stimuli.

To this extend, the Talabre-Lejía transect where MAGs examined in this project were recovered from constitutes an ideal natural laboratory in the middle of the Atacama Desert (see Background 1.1) to explore functional diversity of specialized metabolites synthesized by non-model bacteria. The aim of this study is to track whether co-occurrence patterns in natural product genomic repertoires of draft organisms assembled from rare and taxonomically unique environments can reveal ecological and/or evolutionary dynamics through a scalable pipeline constructed upon considerations for researchers with no bioinformatic background such as broad-use and up-to-date databases, open source tools with visualization interfaces when possible, clear manuals and available discussion and tutorials on the internet.

## Hypothesis

“Co-occurrence patterns in natural product genomic repertoires of draft organisms assembled from rare and taxonomically unique environments can reveal ecological and/or evolutionary dynamics of soil bacterial communities through the recovery of functional information and comparative approaches.”

## Objectives

### General

Assessment of functional co-occurrence and co-expression patterns of NP-encoding BGCs upon available biological information obtained from previous DNA and RNA sequencing efforts of the LBEG research group through the construction of a bioinformatic pipeline customized for data originated from environmental samples in the pursuit of giving insight into ecological and/or evolutionary dynamics of soil bacterial communities of the Talabre Lejía transect.

### Specifics

- a. Prediction and annotation of BGCs.
- b. Genome-resolved comparative analyses on functional information of core biosynthetic specialized capabilities with regards of taxonomy and geographic origins.
- c. Description of biosynthetic classes of specialized pathways according to the occurrence of accessory modules with regards of gene functionality.
- d. Identification of functional redundancy of BGCs or fragments of them and construction of networks of genomes harbouring similar predicted compounds.
- e. Ranking of genomes, BGCs and specialized genes according to their transcriptional activity related to NP biosynthesis.
- f. Assessment of co-expression patterns between types of core enzymes.
- g. Visualization of networks accounting for *in situ* active regulation of BGCs.

## Background of the input data

This section was included aiming to clearly distinguish the efforts of previous works of the Bioinformatics and Gene Expression Laboratory research group from the methodology proposed and executed throughout this thesis project. Figures and tables obtained from biological-derived data and/or constructed upon this input after quality filters are duly declared (see [Supplementary information](#)).

### 1. Obtention of metagenome-assembled genomes

#### 1.1. *Sites and sample collection*

Bulk soil (BS, plant-free soil) was sampled during April of 2014 after the rainy season at six sites along the Talabre-Lejía transect (23°50'S 67°69'W) exhibiting contrasting features of precipitations, temperature, and vegetation belts ([Supplementary Figure S1 and Table S1](#)). BS samples (100 g) were collected in triplicate at 10 cm depth from the ground and stored in dry ice for metagenomic sequencings until their arrival to the laboratory<sup>[79]</sup>. Soil physicochemical and nutritional analytical protocols have been reported previously<sup>[38]</sup>. Mean annual precipitation (MAP) and mean annual temperature (MAT) data were obtained from Díaz *et al.*<sup>[39]</sup>. Briefly, Site 1 represents the lower elevation (S1; 2,870 masl) and, consistently with the rainfall gradient that increases with altitude, exhibited the lowest MAP and the highest MAT values. Soil samples from here were rich in K and Na while achieving low measurements for NH<sub>4</sub>, NO<sub>3</sub>, P and Fe. The medium elevation site (S2; 3,870 masl) showed intermediate values of MAP and MAT, and soil samples were rich in nitrogen and reported the highest content in organic matter. The rest of sites (S3-S6) are located above 4,300 masl, in an area nested at the summit of Lascar volcano, the most active

volcano of the northern Chilean Andes<sup>[40]</sup>, and were exposed to lower temperatures and higher precipitations than sites from lower and medium elevations. S6 soil sample was collected from the Lejía lagoon's very shore and, consistently, exhibited 15 to 40-folds salt concentration than the other three sites located at the same altitude<sup>[41]</sup>.

### 1.2. *Metagenomic shotgun sequencing*

DNA extraction was carried out with NucleoSpin Food kit (Macherey-Nagel) as previously described by Mandakovic *et al.*<sup>[41]</sup>. DNA obtained from triplicate soil samples was pooled to obtain one representative DNA sample per site. Sequencing was carried out by MR DNA (www.mrdnlab.com, Shallowater, TX, USA) on a Miseq platform (Illumina, San Diego, CA) in an overlapping 2 x 150 bp configuration. A total of 832 x 10<sup>6</sup> good quality reads were obtained after filtering (91,9%; 117 Gb). A summary of sequencing information is available in [Supplementary Table S3](#). Annotation of the six metagenomes with Prodigal 2.6.2<sup>[42]</sup> yielded, after discarding redundancy (14%), 6,232,633 genes mainly assigned to bacteria (99,05%) according to the NCBI database. Representation of archaeal taxa was low (0,92%) while of viruses and eukaryotes was almost inexistant ([Supplementary Table S4](#)).

### 1.3. *Binning*

Metagenomes were assembled with IDBA-UD 1.1.1<sup>[43]</sup> after quality checking and adaptor trimming on raw reads. All samples were assembled separately or as pair of samples when they were spatially close (S3+S4 and S4+S5) with the purpose of increasing the chance of assembling larger scaffolds. Hybrid assemblies were performed following Albertsen *et al.*'s protocol<sup>[44]</sup> with modifications according to Alneberg *et al.*<sup>[45]</sup>. Scaffolds were binned by CONCOCT<sup>[44]</sup> according to GC-content, k-mers frequency (tetramers), and reads abundances across samples into 226 bins that recruited approximately 17% of total assembled scaffolds<sup>[79]</sup>. Bin completeness and contamination were evaluated by checking for single-copy genes sets by CheckM

1.0.13<sup>[46]</sup>. Accepted bins were re-assembled into 114 MAGs using Velvet 1.2.10<sup>[47]</sup>. This collection contained 328,282 non-redundant genes, out of which 88,9% had a match with metagenomic reads (identity 95%; BLASTn e-value  $\leq 1e-100$ ). A summary of assembly information is available in [Supplementary Table S5](#). Taxonomic assignment of MAGs was performed with PhyloPhlAn 1.0<sup>[48]</sup> and its standard database of marker genes. Raw abundances of MAGs were determined as the mapping coverage of the reads obtained from the six metagenomes sequenced. Mapping was carried out by Bowtie2 2.3.4.3<sup>[49]</sup> and coverage were calculated with BEDTools 2.17.0<sup>[50]</sup>. A summary of MAGs genomic information is available in [Supplementary Table S2](#).

## 2. Obtention of RNA reads from soil samples

### 2.1. *Sample collection*

Detailed methodology regarding sample collection is described in González *et al.* (in preparation)<sup>[79]</sup>. Briefly, bulk soil (BS, plant-free soil) was sampled during April of 2017 after the rainy season from two of the six sites just described (S1 and S5; [Supplementary Figure S1](#)) to recover regulatory information. Three biological replicates of BS (50 g) were collected at 10 cm depth from the ground and stored in dry ice until their arrival to the laboratory for metatranscriptomic sequencings.

### 2.2. *Metatranscriptomic shotgun sequencing*

RNA was extracted from samples using the RNeasy PowerSoil kit (Qiagen) following the manufacturer's protocol with the following modifications: soil samples were of 15 g instead of 2 g and 5 mL of PowerBead solution were added instead of 15 mL. The integrity of the RNA was evaluated by electrophoresis in denaturing agarose gel and quantification was checked by fluorometry with the Qubit RNA Assay kit (Life Technologies). DNA contamination was removed using Baseline-ZERO DNase (Epicentre)

following manufacturer's instructions and then purification was conducted with RNA Clean & Concentrator 5-columns (Zymo Research). DNA-free RNA samples were amplified with the QuantiTect Whole Transcriptome (Qiagen) and the Nextera DNA Sample Preparation kits. Final concentration of all libraries ([Supplementary Table S6](#)) was measured using the Qubit dsDNA HS Assay Kit (Life Technologies), and the average library size was determined using the Agilent 2100 Bioanalyzer (Agilent Technologies). Libraries were then pooled in equimolar ratios of 2nM, and 10pM and clustered using the cBot for paired-end sequencing using the HiSeq 2500 system (Illumina) for 300 cycles.

### 2.3. Data pre-processing

Trimming of raw reads was executed using Trimmomatic 0.38<sup>[51]</sup> with specifications for paired-end sequences, TruSeq3 optative method for adapter removal and the Phred-33 mode. Leading and trailing parameters were set to 10 bases to be cut off at the start and at end of every read, respectively.



## Methods

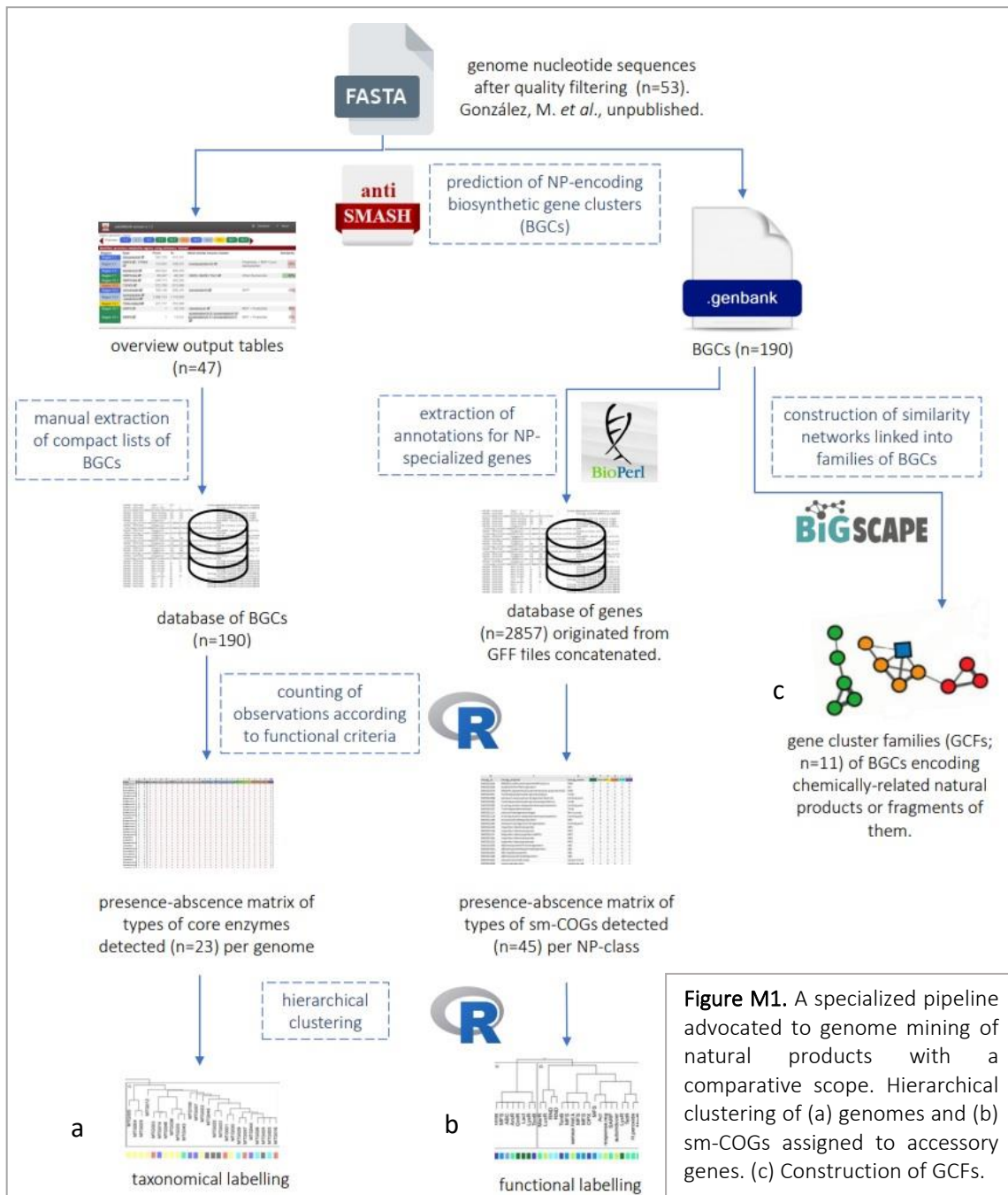
### Genome-resolved comparative analyses on functional information of biosynthetic gene clusters

MAGs with completeness above 70% and contamination under 10% were considered for functional analyses. Specialized metabolite-encoding biosynthetic gene clusters (BGCs) prediction, annotation, and comparison against the MIBiG repository<sup>[34]</sup> were performed with antiSMASH 5.1.2<sup>[52]</sup> webserver in relaxed mode and KnownClusterBlast option on. The tool's basic options retrieve gene classifications according to secondary metabolite categories of orthologs groups (sm-COGs) constructed upon protein families (PFAM<sup>[53]</sup>) hidden Markov models. Biosynthetic core detection rules and -if applicable- their conversion into those of the latest version, MiBiG classification and biosynthetic classes as defined by BiG-SCAPE (Figure 2B) are available in Supplementary Table E2.

'Overview' output tables were fused and BGCs of lengths shorter than 5,000 bp were filtered out (Supplementary Table E3), while genbank output files were converted into GFF format with the script bp\_genbank2gff3.pl of Bio::DB::GFF module<sup>[54]</sup> calling --split and --noinfer flags, and then parsed (Supplementary Table E5). For downstream statistical analyses, these two databases were transformed into matrixes of frequencies of: (i) core biosynthetic capabilities of BGCs per genome (Supplementary Table E4), and (ii) regulatory and transport sm-COGs in BGCs per biosynthetic classes (Supplementary Table E6). Distances of the scaled presence/absence matrixes were calculated with the Manhattan metric and then hierarchical clustering was conducted in R<sup>[55]</sup> with base function 'hclust' and linked by Ward's minimum variance method (Figure M1).

Natural product diversity was explored throughout BiG-SCAPE 1.1.1<sup>[56]</sup> which groups BGCs into gene cluster families (GCFs) with a highly similar predicted metabolite

chemotype. The flags --mix, --hybrids-off and --mibig were called. Default 'auto' mode accounts for comparisons between complete and fragmented BGCs by setting 'glocal' option instead of 'global' when at least one of the regions has one or both of its neighbourhoods located at a contig edge. Resulting sequence similarity networks were visualized with Cytoscape 3.9.1<sup>[57]</sup>. Manual inspection of GCFs of at least two members was conducted throughout the analysis of PFAM domains of reference BGCs delivered by the tool and their conversion into GO terms<sup>[58]</sup> using 'pfam2go' function of ragp package<sup>[59]</sup> in R.



**Figure M1.** A specialized pipeline advocated to genome mining of natural products with a comparative scope. Hierarchical clustering of (a) genomes and (b) sm-COGs assigned to accessory genes. (c) Construction of GCFs.

Qualities of draft genomes were assigned according to [Table M1](#). Relative abundances correspond to individual raw abundances over the sum of genomes kept after quality filter per site ([Supplementary Table S2](#)).

completeness	contamination < 5%	contamination < 10%
70% - 80%	medium	low
80% - 90%	medium-high	medium
90% - 100%	high	medium-high

**Table M1.** Draft genomes quality classification criteria.

## Natural product-oriented pipeline for handling transcriptional data from environmental samples

Indexation of the genomic references and mapping of RNA reads were performed via STAR 2.7.10<sup>[60]</sup> with `--alignIntronMax` and `--genomeSAIndexNbases` parameters fixed for bacterial specifications. ‘CDS’ was defined to be considered as the exon feature. Two strategies were approached to deal with defining transcriptional activity from BGCs retrieved from the previous chapter ([Figure M2](#)).

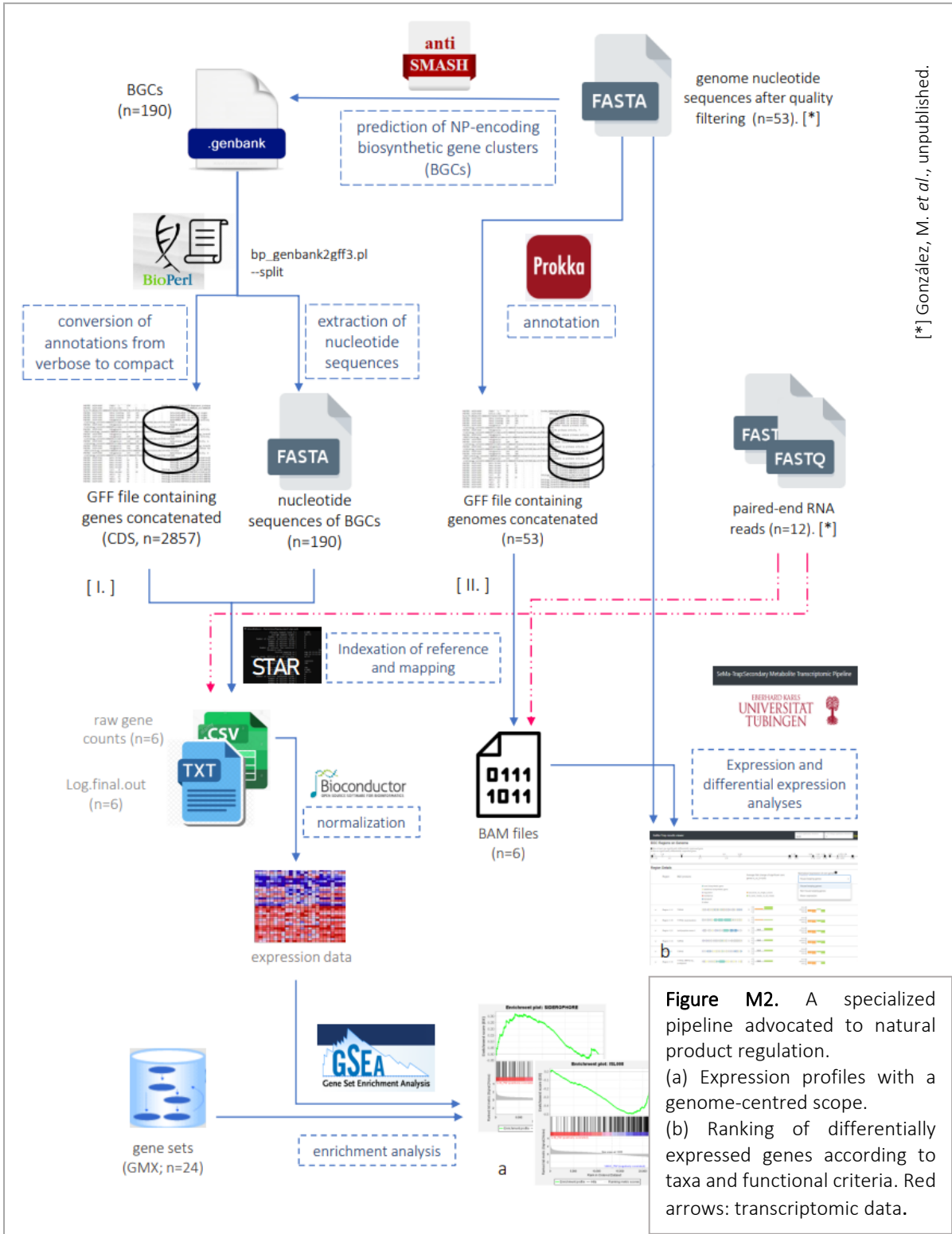
The first consisted in indexing as the reference for alignment all specialized metabolite-encoding regions concatenated. To do so, merged fasta files and GFFs ([Supplementary Table E5](#)) retrieved with the `--split` option flagged for `bp_genbank2gff3.pl` script were delivered as input. RNA reads were uncompressed during execution with the `--readFilesCommand zcat` command. Gene counts obtained from the mapping step were normalized with DESeq2<sup>[61]</sup> considering only genes that had in average at least 5 counts in each sample, and were then ranked with GSEA 4.3.0<sup>[62]</sup> selecting as contrasting conditions samples from S1 and S5. One thousand permutations were requested and, considering that the number of samples hereby studied did not satisfy the minimum for using ‘phenotype label’ mode (at least seven per condition), gene sets were chosen instead for permuting while FDR thresholds were

tightened from 15% to 5%, as recommended for these cases at the tool's website. Gene Set Enrichment Analysis was executed twice for taxonomical and functional classifications by dividing the data into lists of genes according to the phylum of genomes and the biosynthetic core function of BGCs harbouring each gene, respectively. Sets of genes were constructed following the considerations for minimum and maximum default sizes (15 and 500 genes, respectively) and for avoiding repeating genes between lists.

The second strategy considered comparing the expression of genes annotated to encode functions involved in natural product biosynthesis with the expression of housekeeping genes by including in this methodology the recently launched transcriptomic pipeline specific for secondary metabolites, Sema-Trap<sup>[63]</sup>. This software accepts both SRA accessions to download transcriptomic information from the NCBI database and BAM files that can be directly uploaded from local machines to the webserver. Main advantage of Sema-Trap relies on giving consensus about how to manage scoring of clustered genes so their unitary nature is preserved in the ranked results. Such consensus consists in weighting each gene's score with the BGC's average by simply multiplying them. Since RNA data from BioProject PRJNA291433 is not publicly available yet, with the aim of trying this new tool, full genomes were annotated with Prokka 1.13<sup>[64]</sup> and resulting GFFs together with the concatenated fasta files were used as input in the indexation step (again with STAR 2.7.10). Same concatenated nucleotide sequences of genomes and the BAM files (sorted by coordinate) that resulted from the mappings were then delivered as input for Sema-Trap.

Corroboration of input biological-derived data qualities was reassessed with FastQC 0.11.9<sup>[65]</sup> after trimming was conducted with Cutadapt 2.8<sup>[66]</sup> (both tools wrapped in Trim Galore!) detailing options for paired-end reads from Illumina technologies over the already-trimmed sequences (see [section 2 of Background](#)). A minimum read length of 70 bp was set and a phred quality value of 30 was requested, meaning that the probability of getting a wrong base is 1 in 1000 (precision of 99.9%).

Flexibilizations of STAR parameters --outFilterScoreMinOverRead (henceforth: Score) and --outFilterMatchNminOverRead (henceforth: Match) were evaluated from default values 0.66 down to 0.20 in two samples for 22 combinations of cut-offs and in all samples for the reduced critical four cut-offs.



**Figure M2.** A specialized pipeline advocated to natural product regulation. (a) Expression profiles with a genome-centred scope. (b) Ranking of differentially expressed genes according to taxa and functional criteria. Red arrows: transcriptomic data.

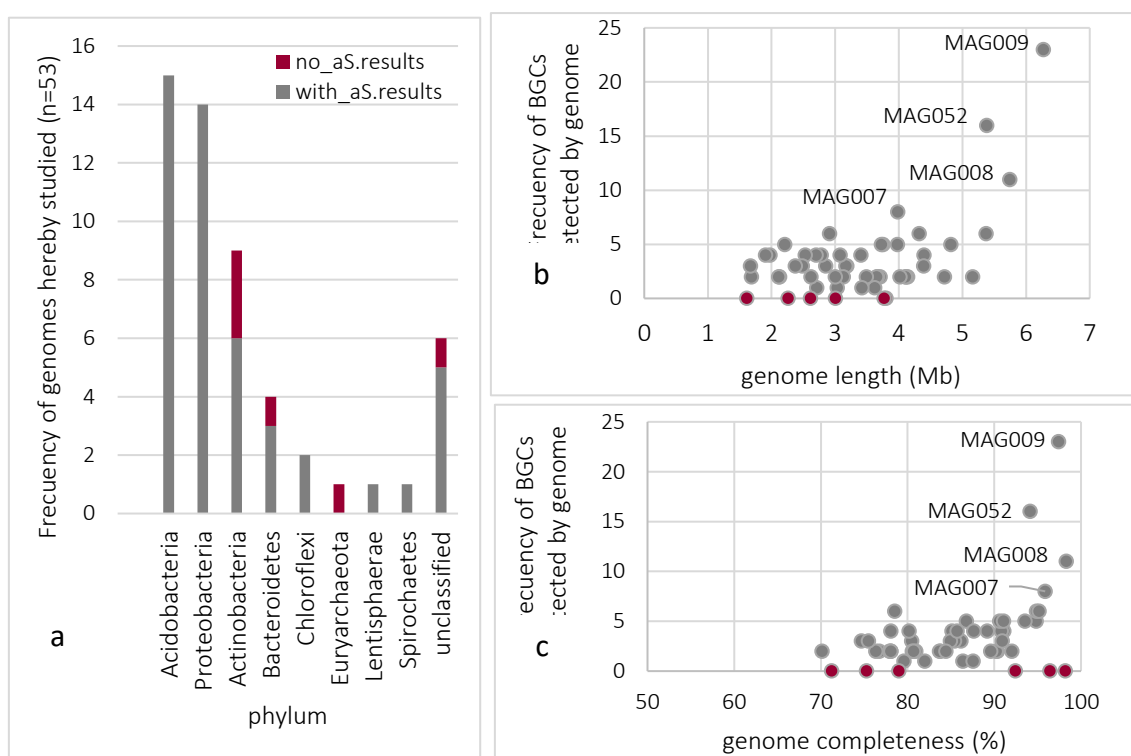
## Results

### Natural product genomic repertoires demonstrate taxonomical and geographic specificities in soil bacteria

Out of all 114 MAGs binned out from BioProject PRJNA291433<sup>[79]</sup>, 53 (46.5%) passed the quality thresholds for completeness and contamination, and were considered for functional exploration. In average, kept genomes yielded 3,409,889 bp in length, completeness and contamination indexes of 85.6% and 3.7%, respectively, and a N50 of 31,270 (Supplementary Table S2). Of these, 32 (56.1%) were classified as high or medium-high quality draft genomes according to criteria defined in Table M1, and 47 (88.7%) threw at least one predicted region of 5,000 bp or longer related to specialized metabolite biosynthesis. Those six that had no antiSMASH results belonged to *Actinobacteria*, *Bacteroidetes*, *Euryarchaeota* or unclassified taxas (Figure 1A), compressed high, medium, and low draft-qualityes (Figure 1C) and constituted relatively small genomes (Figure 1B). The three most dominant phyla across sampling sites were *Acidobacteria*, *Proteobacteria* and *Actinobacteria* which recruited ~85% of overall relative abundance (Supplementary Figure S2 A). A closer observation of the data disaggregated by sites shows that rather than unabdundant, *Euryarchaeota* appears to be restricted to S2, while *Chloroflexi* is to S3 and S4, and *Bacteroidetes*, *Spirochaetes* and *Lenthisphaerae* are to S6. Soils of S3 were deprived of *Proteobacteria* while soils of S6 were so of *Acidobacteria* and *Actinobacteria* (Supplementary Figure S2 B).

The dataset resulted in 190 NP-encoding regions (henceforth: the TLT dataset) out of which ~15% were complete —meaning that they were fully allocated in the contigs and not truncated on their edges achieving an uninterrupted prediction— with an average length of 26,852 bp. More than half of BGCs had no homologue genes when searched against the MIBiG repository (Figure 2B and Supplementary Table E3) and about 90% were identified as single candidate clusters as only 22 conformed hybrids, 15 of them nominated as ‘intervealed’ or ‘chemical hybrid’<sup>[52]</sup>. Four genomes —all drafts

of high qualities belonging to *Acidobacteria*— accounted for ~30% of the BGC collection: MAG007, MAG008, MAG009 and MAG052 with genome sizes ranging from 4.0 to 6.3 Mb (Figure 1). Highest BGCs counts were found in MAG009 with 23 regions of 26,530 bp average in length, 5-folding the average frequency of regions per genome with results (4.0). The longest BGC of the dataset corresponded to a complete hybrid region of 131,091 bp found in MAG008 with ~10% of its genes matching half of those of the nostopeptolide A2-producing gene cluster (BGC0001028, Supplementary Table E3).



**Figure 1.** (a) Taxonomic distribution of MAGs with completeness  $\geq 70\%$  and contamination  $\leq 10\%$  ( $n=53$ ), and incidence of genome (b) completeness and (c) length over frequency of detected BGCs ( $n=190$ ). aS: antiSMASH

Among the 59 core biosynthetic types available at the fifth version of antiSMASH (Supplementary Table E2), 23 were detected among the TLT dataset. Predominant biosynthetic classes were NRPS, RiPPs and terpenes (Figure 2B) and were mainly contributed by *Acidobacteria* and *Proteobacteria*. *Bacteroidetes* and *Chloroflexi* become relevant for terpene and type I polyketide synthase (T1PKS) biosynthetic types, respectively. *Lentisphaerae* contributed only one BGC to the collection (NRPS-like) while *Spirochaetes* and unclassified taxa threw no regions related to NRPS, T1PKS or

hybrids of these mechanisms (Figure 2C). A total of 2,857 coding sequences were predicted averaging 15 genes per BGC, out of which 331 (11.6%) were annotated as core for NP biosynthesis, 484 (16.9%) were termed as ‘biosynthetic additional’ and 1788 (62.6%) had no functional information whatsoever. Remaining fraction of genes (8.9%) were assigned to ‘transport’ (80), ‘regulatory’ (100) and ‘other’ (75) sm-COGs. All seven resistance genes found corresponded to an ABC transport-related protein (SMCOG1288). Beside core genes, 237 genes had a predicted molecule (Supplementary Table E5). Of these, relevant specialized metabolite domains most frequently found (~30%) were Radical SAM (PF04055), Peptidases (S8, S9, S41, M16, M42, M50 and C39), a *SnoaL*-like polyketide cyclase (PF07366), the DegT/DnrJ/EryC/StrS aminotransferase family (PF01041) and a nitroreductase (PF00881).

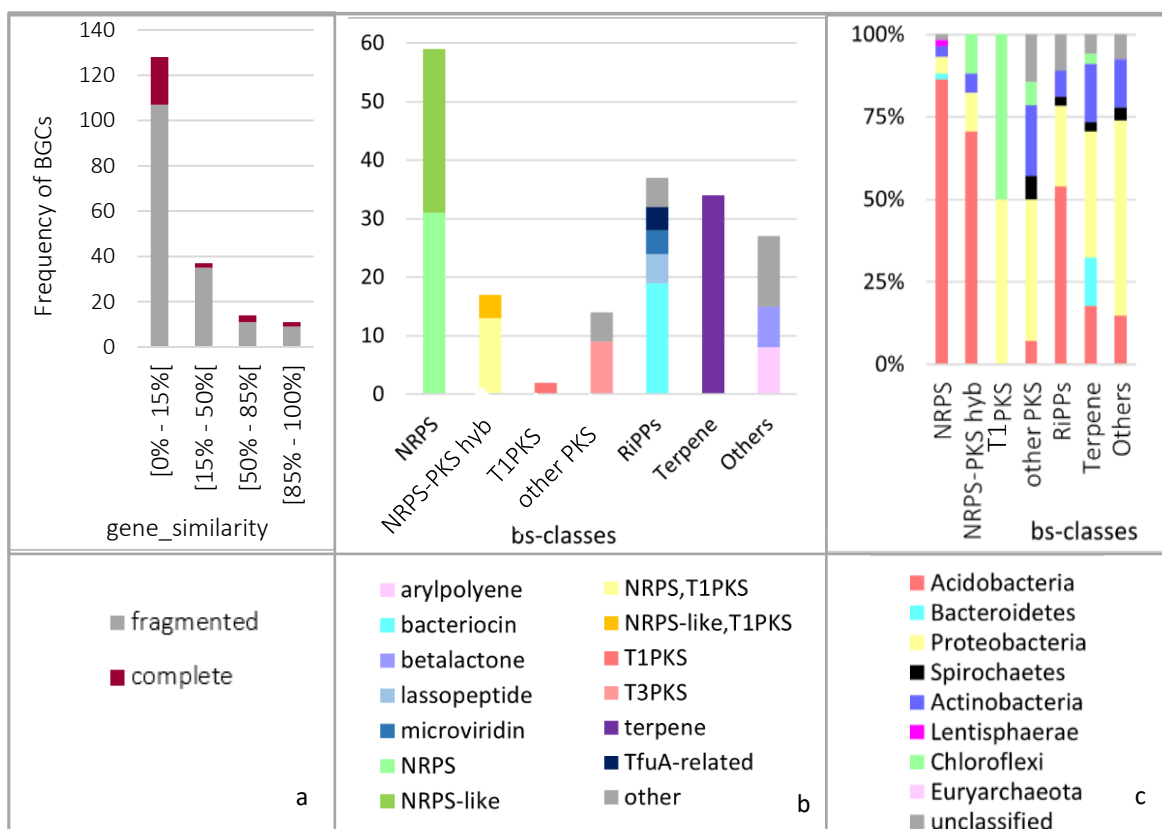
phylum	average n° of bgcs	total n° of bgcs	contribution (%)	relative abundance (%)	ranking
Acidobacteria	6.27	94	49.47	36.38	+ 13.09
Spirochaetes	4.00	4	2.11	0.54	+ 1.57
Proteobacteria	3.57	50	26.32	25.03	+ 1.29
Lentisphaerae	1.00	1	0.53	0.45	+ 0.08
Bacteroidetes	1.50	6	3.16	3.32	- 0.16
Chloroflexi	2.50	5	2.63	2.87	- 0.24
unclassified	1.83	11	5.79	6.40	- 0.61
Euryarchaeota	0.00	0	0.00	1.55	- 1.55
Actinobacteria	2.11	19	10.00	23.45	- 13.45

**Table 1.** Summary of gene cluster predictions reviewed by taxonomic group: average number of BGCs was calculated upon the whole genomes collection, including those with no results (n=53); ranking is the difference between phyla percentual contributions of BGCs calculated upon the overall sum (n=190) and their relative abundance.

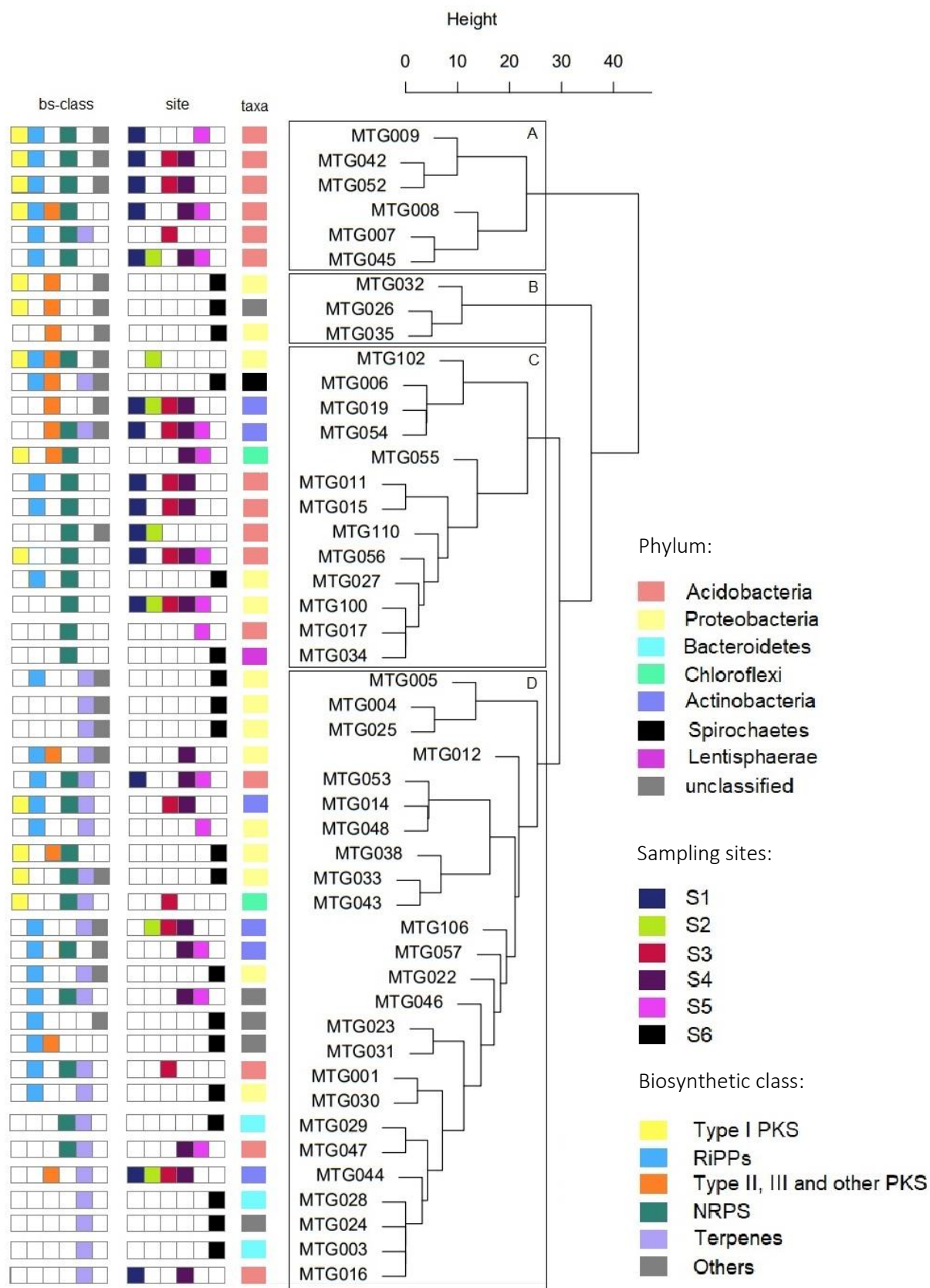
Hierarchical clustering with regards of present NP-encoding capacities in genomes allowed to divide the data in four groups (Figure 3). The first and most distinct group is composed of only *Acidobacteria* with mainly wide distributions (S1-S5) and is characterized for being the only group containing both NRPS and NRPS-like-encoding BGCs. Bacteriocins, microviridins and/or lanthipeptides are present in all five genomes of this group (Figure 3A). Group B consists of one unclassified and two *Proteobacteria* genomes from S6 and have arylpolyene and heterocyst glycolipid synthase-like PKS (hglE-KS) biosynthetic types as markers (Figure 3B). Groups C corresponds to diverse



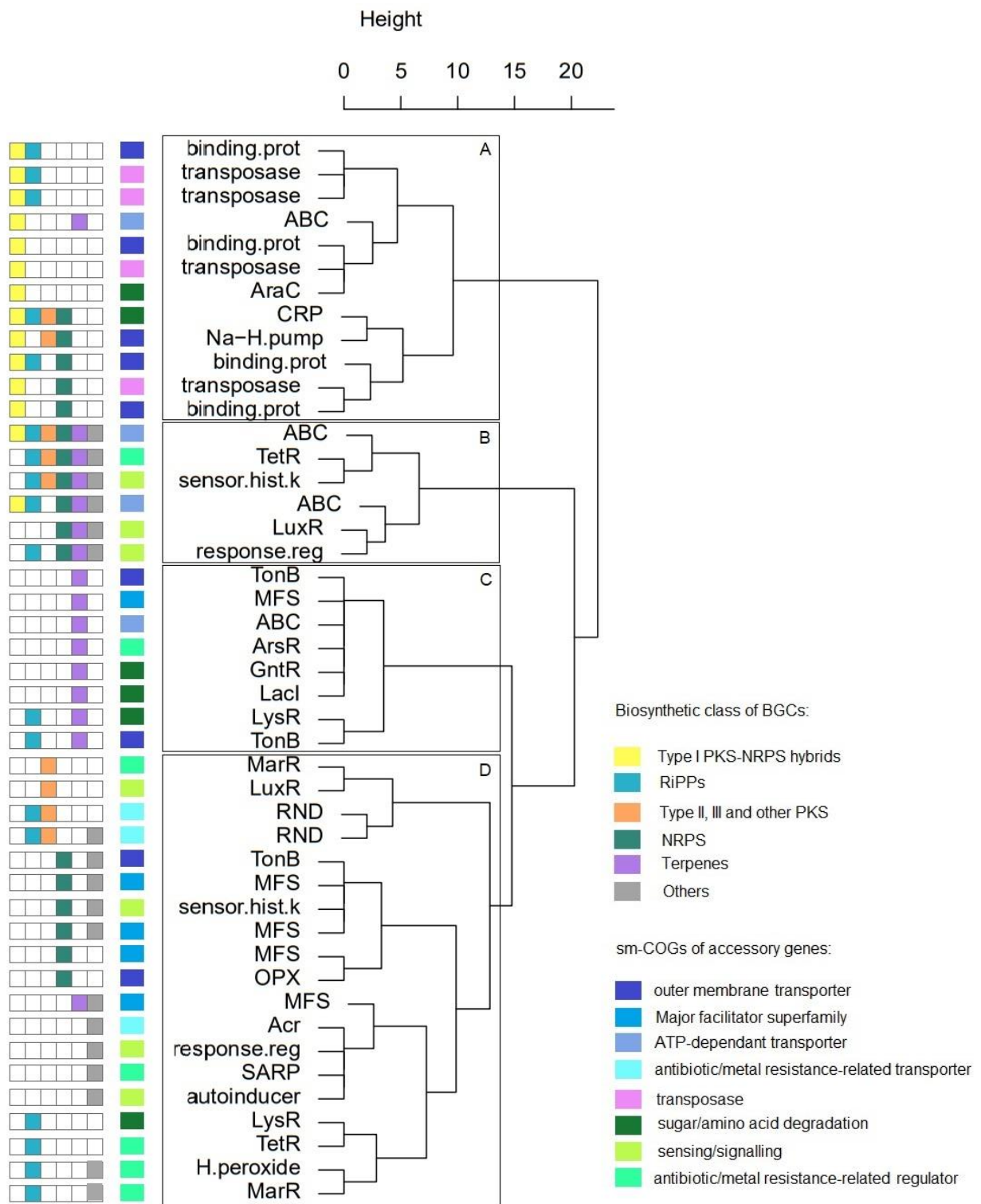
taxa from both restricted and generalist geographic distributions. Members from the upper branch share metabolic capacities for encoding betalactones and type III polyketide synthases (T3PKS); while the lower branch is dominated by NRPS-like and distinguishes MAG055 (*Chloroflexi*) for being the only genome in the collection holding a type II polyketide synthase (T2PKS) gene cluster (Figure 3C). Group D is mainly defined by the presence of terpenoid pathways. The upper branch of this group refers to *Proteobacteria* from S6 that possess regions related to ectoine biosynthesis, while the middle branch (including MAG012) refers to genomes with NRPS, T1PKS and/or lassopeptide-encoding BGCs. Lastly the lower branch hosts genomes with unique capacities among the dataset such as siderophore (MAG106), linear azol(in)e-containing peptide (LAP, MAG046), N-acetyl-glutaminy-glutamine amide (NAGGN, MAG057) and other non-classified metabolite (MAG057) production (Figure 3D).



**Figure 2.** Frequencies of BGCs (n=190) by (a) similarity ranges of closest match as delivered by KnownClusterBlast, and (b) biosynthetic classes disaggregated by biosynthetic types. (c) Taxonomic contribution of predicted BGCs per genome by biosynthetic class.



**Figure 3.** Hierarchical clustering of genomes with antiSMASH results (n=47) based on presence or absence of functional capacities (n=23) over 144 observations. Sites are coloured if relative abundance was above 0.1%.



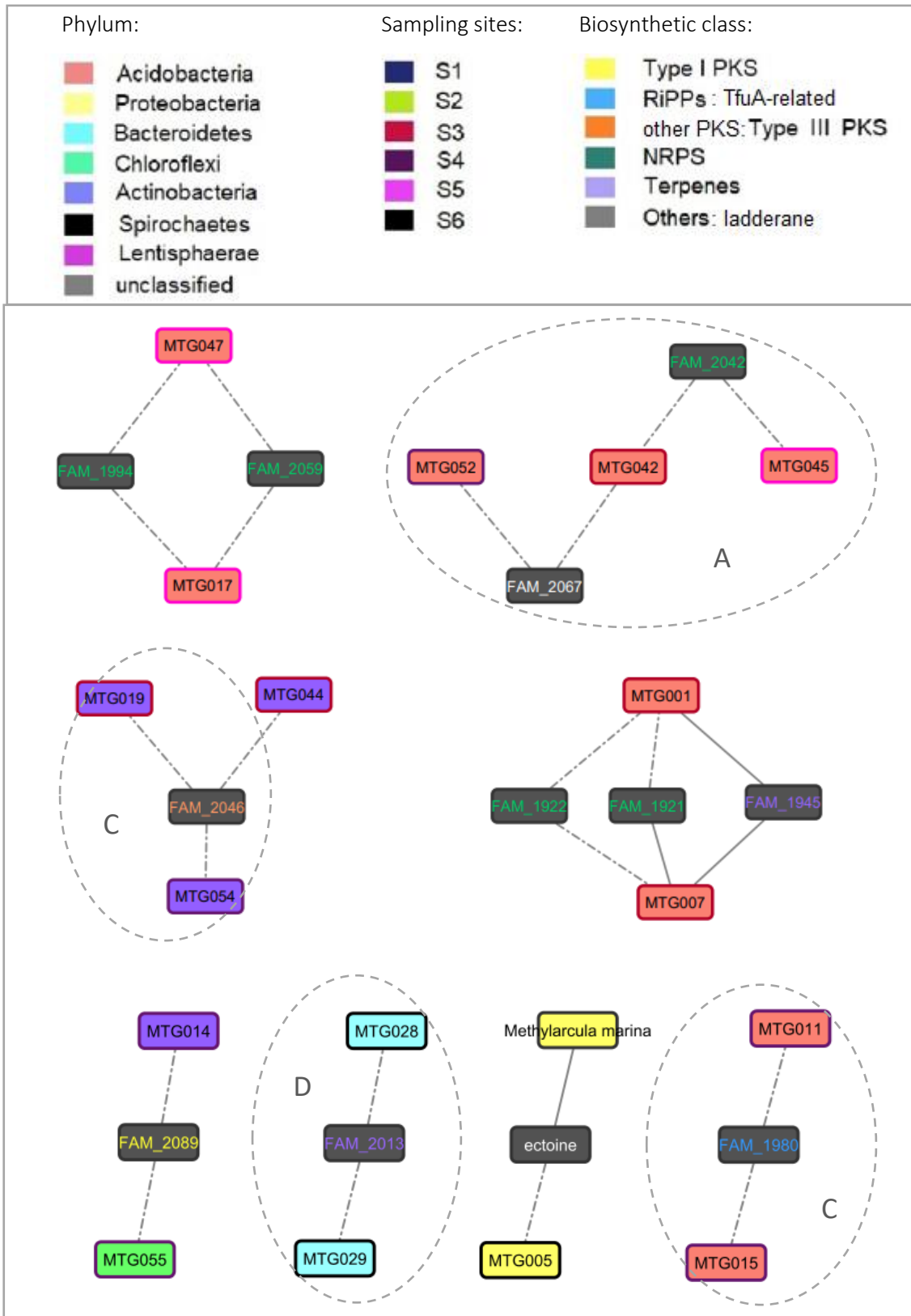
**Figure 4.** Hierarchical clustering of transport and regulatory sm-COGs (n=45) assigned to accessory genes based on their presence or absence in BGCs according to their biosynthetic classes (n=6) over 92 observations. Distances were calculated with manhattan method and linked by ward.D2 distance. Functional labels were assigned manually. CRP: catabolite repressor protein; binding.prot: binding protein-dependant transport systems; NA-H.pump: sodium/hydroxygen exchanger; sensor.hist.k: sensor histidine kinase; response.reg: unspecified response regulator; RND: resistance, nodulation, and cell division superfamily of transporters; OPX: outer membrane polysaccharide export; H.peroxide: hydroxygen-peroxide sensitive repressor; Acr: arsenic resistance transporter; SARP: Streptomyces antibiotic regulatory protein.

Parallely, hierarchical clustering of sm-COGs assigned to accessory genes with regards of the biosynthetic core capacity of the BGCs they belonged to revealed functional organization (Figure 4). Namely, transposases and binding protein-dependant transporter domains were found associated to T1PKS/NRPS hybrids (Figure 4A). Group B harbours sensing-related functions related to several NP-pathways (Figure 4B). Regulators of lactose, lysine and gluconate metabolisms and ABC transporters have as common factor terpene biosynthetic capacities (Figure 4C). All antibiotic and/or metal resistance-related transporters and regulators, excepting ArsR and one of the two TetR sm-COGs included, are clustered together in group D, which is mainly constituted and subdivided by NRPS, RiPPs, other PKSs (non-T1PKS), and Others classes (Figure 4D).

Exploration of biosynthetic diversity assessed with BiG-SCAPE revealed that 88% of BGCs from the transect were singletons, this is, had no structurally similar NP associated within the genomes hereby studied (Table 2). Gene cluster families that were built show shared NRPS-derived metabolites within *Acidobacteria*, while three *Actinobacteria* house the same (or nearly the same) T3PKS (Figure 5). Family 2089 is the only one that disrupts taxonomic specificity at phylum level —within the TLT dataset— by being detected in the genomic contexts of MAG055 (*Chloroflexi*) and of MAG014 (*Actinobacteria*). Families 1921, 1922 and 1945 link genomes that were exclusive to S3 and Family 2013 is harboured by genomes that were so to S6 (Figures 3 and 5). Only one detected region matched an experimentally verified ectoine-encoding BGC from the MiBIG repository (BGC0000860) which was found in MAG005 (*Proteobacteria*).

biosynthetic class	GCFs	BGCs in GCFs	singletons
NRPS	5	10	49
other PKS	1	3	11
hybrid NRPS-PKS	1	2	15
RiPPs	1	2	35
terpenes	2	4	30
T1PKS	0	0	2
Others	1	2	25
saccharides	0	0	0
total	11	23	167

**Table 2.** BiG-SCAPE output summary of biosynthetic gene cluster families.



**Figure 5.** Sequence similarity networks of GCFs with at least two members as delivered by BiG-SCAPE. Edges are the BGCs, and grey nodes are the GCFs they were grouped in. Border of coloured nodes: MAGs' preferred site; filling of coloured nodes: phylum; font of grey nodes: biosynthetic class of shared predicted natural product; Edge of continuous line: complete BGC. Dotted circles: same functional group in [Figure 3](#).

GCF	biosynthetic type	possible function
FAM_2067	Ladderane	possible antibiotic
FAM_1921	NRPS	DNA-repair sub-module
FAM_1922	NRPS	possible siderophore
FAM_1994	NRPS-like	possible antitoxin
FAM_2042	NRPS-like	unknown metal-related
FAM_2059	NRPS-like	possible cocaine-hydrolase
FAM_2089	hybrid NRPS-PKS	possible antibiotic
FAM_2046	T3PKS	oxidative stress resistance
FAM_1945	Terpene	oxidative stress resistance
FAM_2013	Terpene	unknown metal-related
FAM_1980	TfuA-related	possible antibiotic

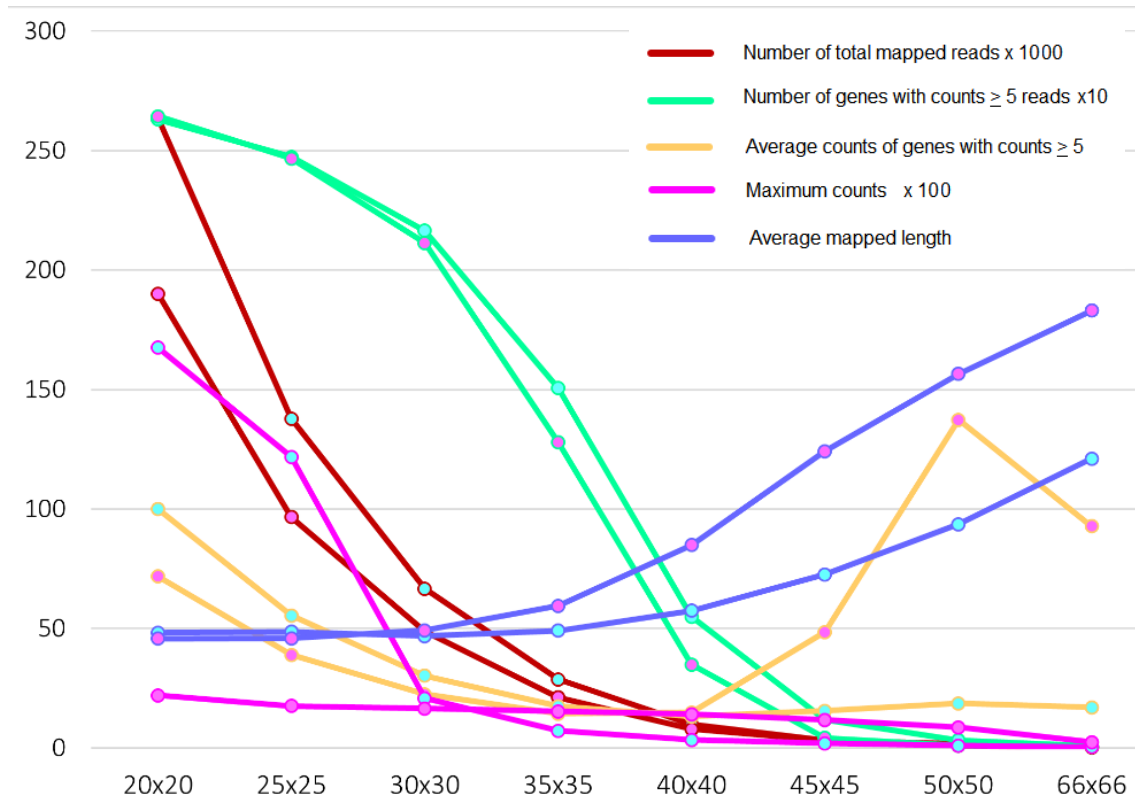
**Table 3.** GCFs summary and possible functions.

### Input biological data showed high sensitivity to adjustments of quality-filtering parameters for matched RNA reads

Concatenation of BGCs resulted in a GFF file containing 2,857 coding sequences (CDS) after the conversion of antiSMASH output genes originally retrieved in genbank format (Figure M2 and Supplementary Table E5). When the mapping step was executed with STAR's default values for filtering parameters with regards of the minimal number of matched bases and the minimal mapping scores —both normalized over read length (66%)— proper indexation of genes was observed but no counts were retrieved appearing '0' for every gene in every sample. After some research in forums, deep study of customizable commands listed in the tool's manual and consulting with experienced RNA-seq data analysts, a flexibilization of these parameters appeared to be the only chance of changing the no-result result.

To do so, a broad first immersion was conducted by trying equal thresholds for Match and Score parameters in two random samples, one representing each site. Additional to the default requirement for considering a mapped read such if it has at least two thirds of its length matching the reference, seven other cut-offs ranging from 20% to 50% were overviewed to define where the mapping statistics critically decreased. The immediate cut-off before the one in which counts became zero would

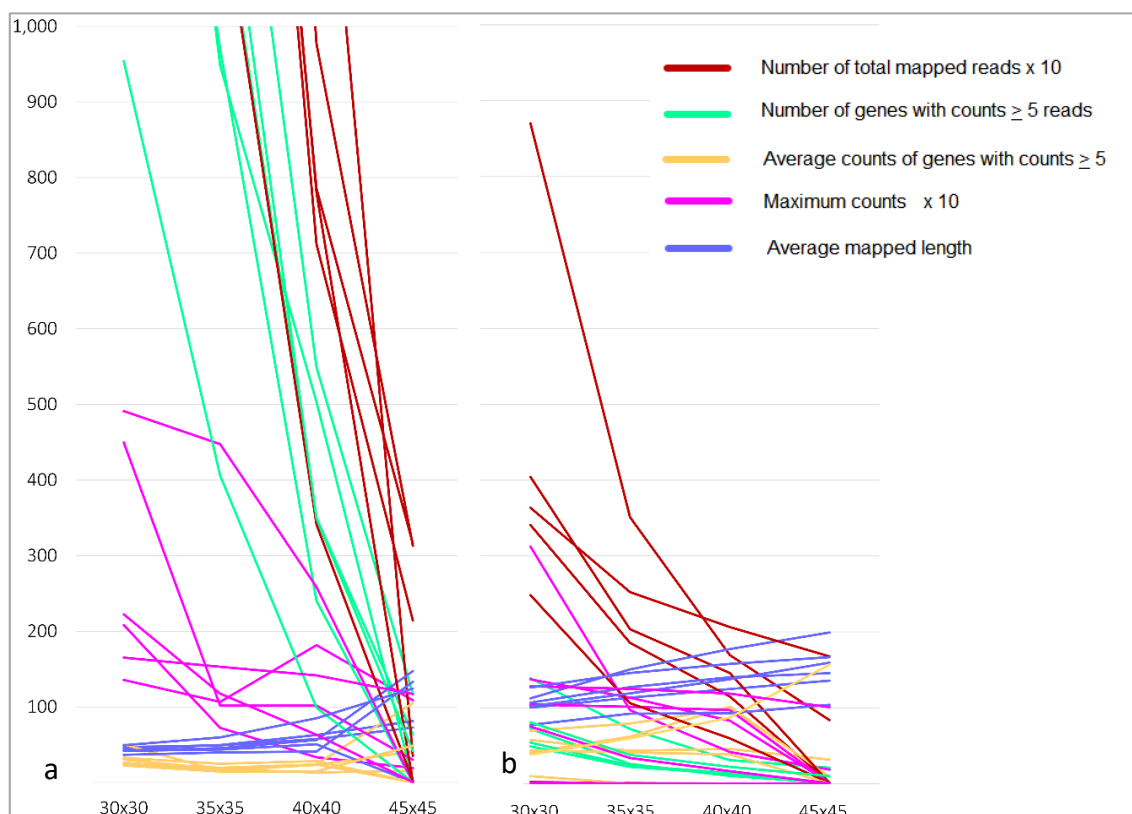
be chosen for downstream analyses in the aim of optimizing the highest possible values for the parameters above described. Mapping statistics were created upon the Log.final.out files delivered in STAR's output for each sample; of them, the following were informative: number of total mapped reads, number of genes with five or more mapped reads, average number of mapped reads for genes with five or more reads, maximum counts of reads and average mapped length (Figure 6).



**Figure 6.** RNA reads mapping behaviour for eight cut-off trials evaluating Match x Score (%) over read minimum length for two samples before Trim Galore processing. Cyan dots: ID#4; pink dots: ID#9).

Main observation retrieved from these indicators was the steep drop of total mapped genes and of the numbers of genes with five or more counts from over 250,000 and 2,500, respectively, to practically none at 50%. An abnormal peak in the average counts of reads mapping to genes was observed for sample #9 at the 50% requirement; however, when this issue was inspected, it was an artefact due to the very few genes with considered counts. Next, the 45% requirement was evaluated, and it was noted that the average mapped lengths appeared to achieve values closer to 100 bp implying

a theoretical resemblance to the default quality threshold by covering about 2/3 of the 150 bp-long reads. The latter made the 45% cut-offs the highest further considered and, therefore, the optimal. Lastly, when Match and Score parameters were set to 30%, a fall of disruptively high maximum counts in sample #4 was observed which made this threshold the lowest further considered. A deeper exploration trying twenty-two combinations of different values between the two parameters hereby analysed was conducted revealing a tendency with an almost identical dropping-behaviour to the one just described ([Supplementary Figure E1](#)). This way, in a narrowing effort, requirements between 30% and 45% were applied to all samples of the study corroborating the defined critical interval and bringing higher resolution to the dramatic sensibility of the data ([Figure 7A](#)). A particular mapping statistic ('Unmapped: too short', [Supplementary Table E1 subtable \(i\)](#)) incepted the first clue regarding poor sequencing qualities and allowed justified review of previous steps. The only one manageable from a bioinformatic scope was the checking of input RNA reads, ends for which Trim Galore! was included in the pipeline aiming to discard technical impossibility for downstream analyses.



**Figure 7.** RNA reads mapping behaviour for four cut-off trials evaluating Match x Score (%) over read minimum length for six samples (a) before and (b) after Trim Galore! Processing.









FastQC reported all samples to pass quality thresholds for ‘per base sequence quality’, ‘per tile sequence quality’ and ‘per sequence quality score’ statistics, while all were rejected by means of ‘per base sequence content’ and ‘sequence duplication levels’ and received a warning for ‘sequence length distribution’ (Supplementary Figure E2). Only two statistics varied between samples: ‘per sequence GC content’ (Figures 8 and 9) and ‘overrepresented sequences’. It’s worth noting that if metatranscriptomic samples were deduplicated, in average, only 28.9% of reads would remain (Table 5). Also in average, 27.4% of reads still had adapters while 9.5% were low quality before the second trimming was applied (Table 4). When mapping statistics of Trim Galore-reprocessed reads were compared to those of Trimmomatic-processed sequences (Figure 7B) a demurer slope in the number of total reads was observed while the number of genes with mapped reads dropped considerably. At the same time, disruptive maximum high counts previously obtained were now softened. At the lowest threshold, average mapped lengths of Trim Galore-derived reads resemble the values retrieved at the highest threshold of the untrimmed experiment. Average number of mapped reads in genes with considered counts increased, metric probably achieved by the withdraw of short sequences. A shrinkage of uniquely mapped reads from 0.47% to 0.04% was observed at the 30% cut-off while at 45% the decrease was thinner (from 0.02% to 0.01%; all averages). Mapped lengths statistics also reflect the quality-filtering of the data by increasing from 15.2% to 37.3% and from 39.1% to 54.3% at the minimum and maximum critical thresholds, respectively. Furthermore, sample #7 —before Trim Galore! processing— threw the following message when ran with the STAR aligner set to 45%: “FATAL ERROR in reads input: quality string length is not equal to sequence length (...). SOLUTION: fix your fastq file.” Finally, values ranging up to 0.21% for the multimapping index in the reads trimmed once were crushed down to zero when double-trimmed (Supplementary Table E1, subtables (i) and (ii)). All taken together provided the context for defining Trim Galore-resulting trimmed RNA sequences as best available input for downstream analyses. However, comparison of cut-offs at differential expression stages wasn’t possible as the only threshold that retrieved a result in DESeq analyses (Match and Score of 30%) threw 23 up-regulated genes in S1 that were not sufficient for GSEA to process, also retrieving errors in every case.

Even though results gathered so far settled an ominous portent for the second strategy, considering that Sema-Trap uses other tool for mapping, it was executed looking for new outcomes. Unfortunately, all jobs failed due to very few reads mapping to the reference, as informed by the tool’s support team via email after asking for assistance.

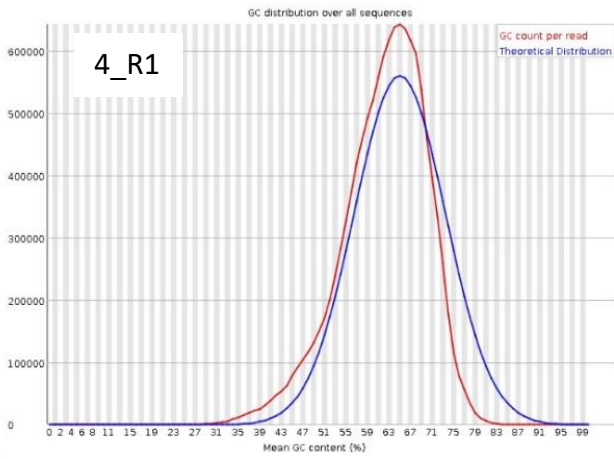
stats	4_S1_i	5_S1_ii	6_S1_iii	7_S5_i	8_S5_ii	9_S5_iii
total reads processed	14,011,630	13,293,545	17,777,500	12,916,790	13,331,868	14,142,688
% reads with adapters (R1-R2)	26.6 – 24.5	28.5 – 26.6	28.1 – 25.7	25.6 – 23.8	33.6 – 31.9	27.6 – 25.8
% quality-trimmed (R1-R2)	7.2 – 11.4	6.4 – 9.3	6.9 – 11.2	8.9 – 12.8	8.9 – 14.4	7.0 – 10.0

**Table 4.** Trim Galore! reports summary for six paired RNA samples. R1: pair 1; R2: pair 2.

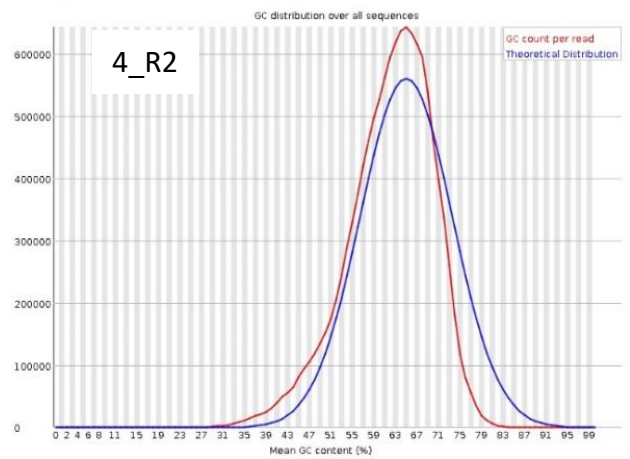
Stats	4_S1_i	5_S1_ii	6_S1_iii	7_S5_i	8_S5_ii	9_S5_iii
total sequences	11,760,578	11,440,094	14,883,642	10,545,073	10,726,661	12,085,268
poor quality seq	0	0	0	0	0	0
% GC content	61	59	59	59	48	58
% remaining seq if deduplicated (R1-R2)	41.8 - 43.1	27.1 – 28.3	14.2 – 15.0	29.9 – 31.0	9.6 – 10.5	47.4 – 48.7
overrepresented seq						

**Table 5.** FastQC reports summary for six paired RNA samples after Trim Galore! processing.

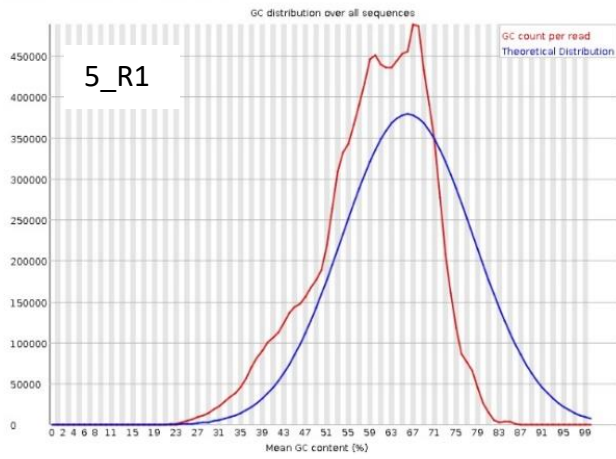
1 Per sequence GC content



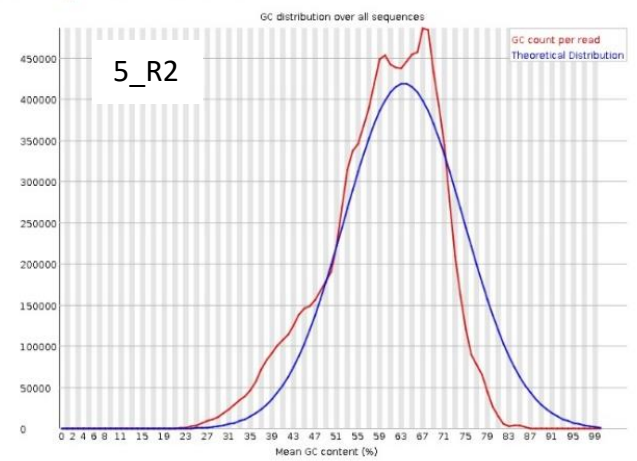
1 Per sequence GC content



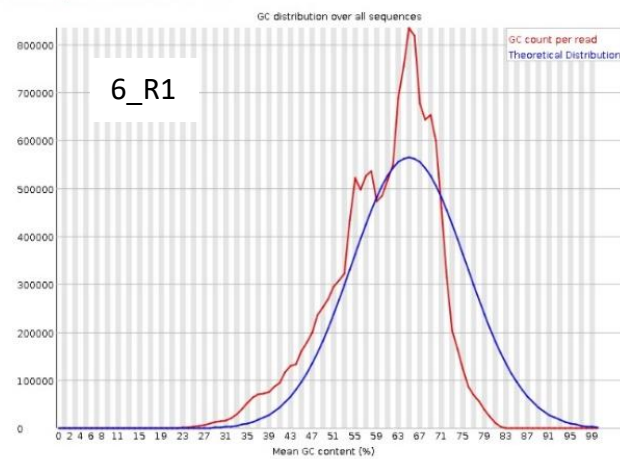
2 Per sequence GC content



1 Per sequence GC content



2 Per sequence GC content



2 Per sequence GC content

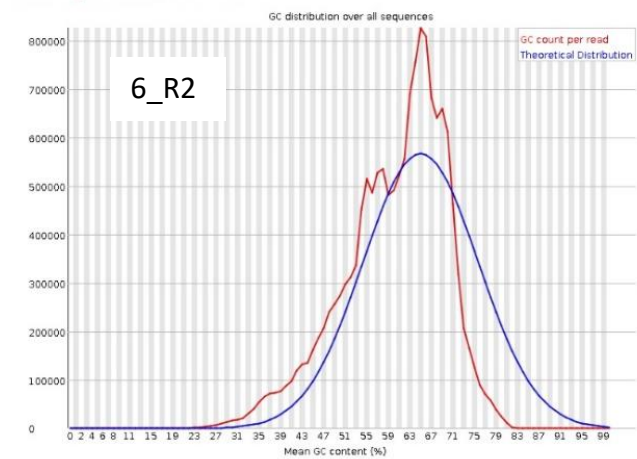
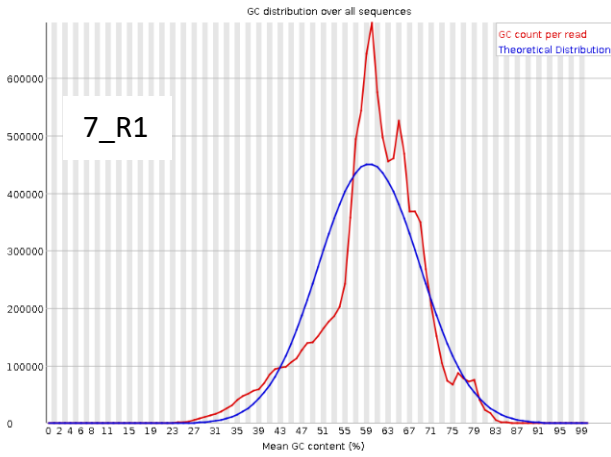
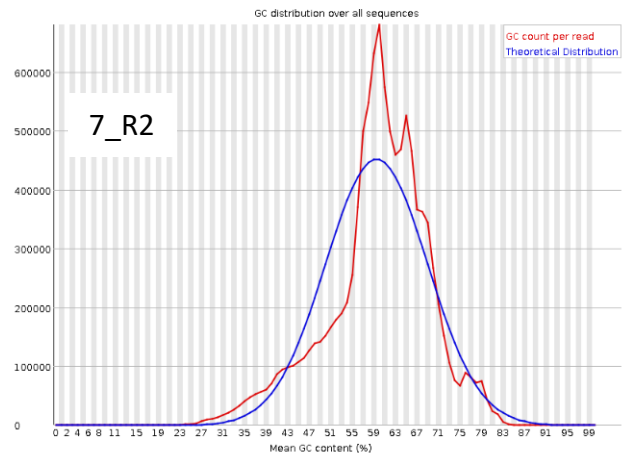


Figure 8. Trim Galore! report from the GC content module for samples from S1. R1: read pair 1; R2: read pair 2.

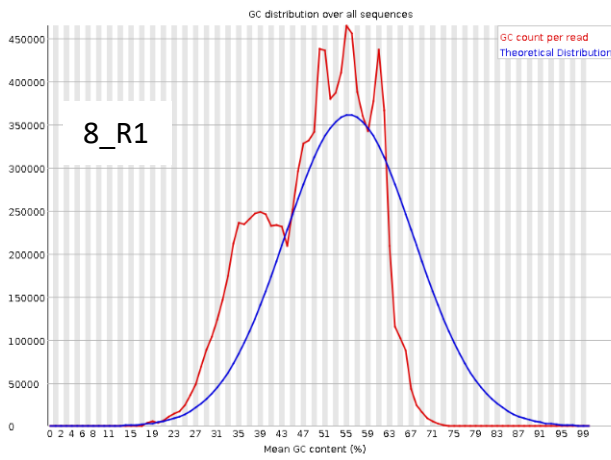
✖ Per sequence GC content



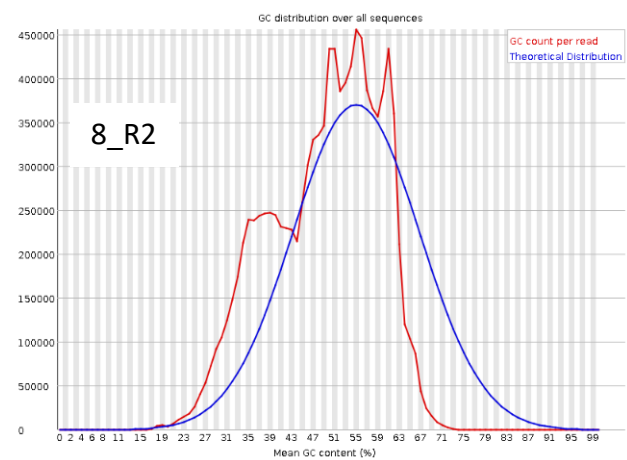
✖ Per sequence GC content



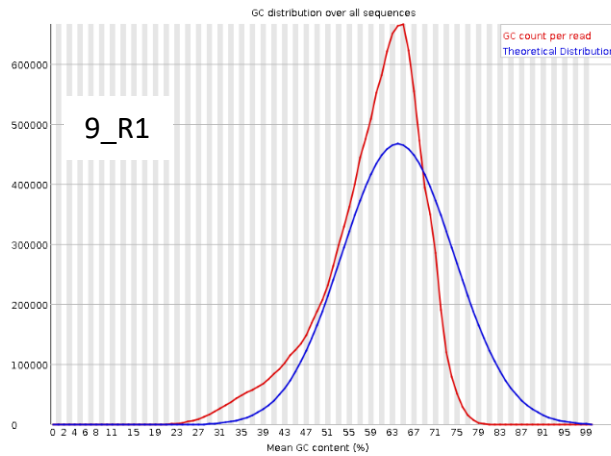
✖ Per sequence GC content



✖ Per sequence GC content



✖ Per sequence GC content



✖ Per sequence GC content

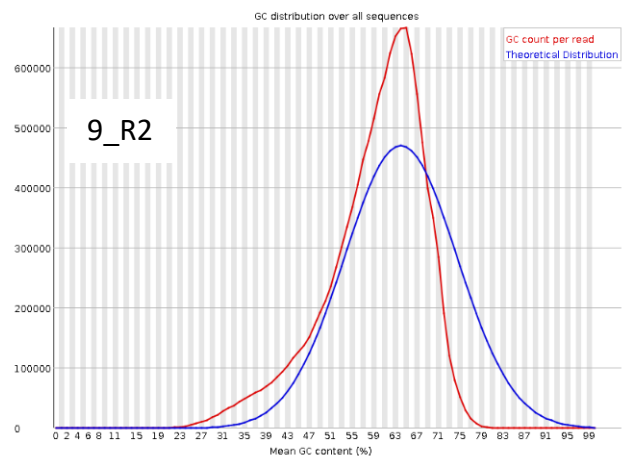


Figure 9. Trim Galore! report from the GC content module for samples from S5. R1: read pair 1; R2: read pair 2.

## Discussion

### Natural product biosynthetic capabilities serve as marks for enlightening ecological and evolutionary patterns in uncultivated soil bacteria

The taxonomic composition of MAGs changed among the six sampling sites highlighting S6 as notably distinct (Supplementary Figure S2 B). Observed overall most abundant phyla (Supplementary Figure S2 A) agree with previously described for bacterial communities from the Atacama Desert<sup>[67,68]</sup> but —within the collection— *Actinobacteria* rather than dominant was underrepresented and, instead, *Acidobacteria* prevailed delivering good representation of this underexplored phylum<sup>[69]</sup>, ubiquitous and diverse in soils<sup>[70]</sup>. Examination of presence of genomes across the transect (defined as a relative abundance of at least 0.1%) evidenced that ~30% presented a generalist geographic distribution covering lower, middle, and high elevations (S1-S5) while ~55% of MAGs were found to be exclusive to a unique site (Figure 3) which supports the conceiving of spatial gradients as dominant drivers of microbial diversity in salars<sup>[71]</sup> and boulders<sup>[72]</sup> of the Atacama Desert. S6 appears to be a mutually exclusive character with regards of all other sites which could be explained by the extreme salt conditions its soils are exposed to<sup>[41]</sup>. This is consistent with extremely high sample-specificity among 115 MAGs from salt crusts of the Atacama Desert described by Finstad *et al.*<sup>[73]</sup> and halobacteria in general<sup>[74]</sup>. Notably, except for MAG100, all *Proteobacteria* behave in a site-specific manner.

With the aim of exploring whether functional repertoires of MAGs respond at some level to environmental heterogeneity, genomic analyses were carried out. Contributions of taxonomic groups regarding specialized metabolite-encoding capabilities compared to their relative abundance revealed *Acidobacteria* as superiorly enriched (Table 1), comment that was made for the first time a few years ago by Crits-

Christoph *et al.* in MAGs reconstructed from grassland soils<sup>[75]</sup>, and has been made again by the same group in soils of a vernal pool<sup>[76]</sup> and of a forested hillslope<sup>[77]</sup>, all in northern California, USA. Both MAGs with unusually high counts, this is, with 15 or more BGCs as defined in the last quoted studies (MAG009 and MAG052), and the one containing the longest BGC in the dataset (MAG008) belonged to *Acidobacteria* and were clustered together in the most distinct group of genomes (Figure 3A) when natural product co-occurrence patterns were assessed. It should be noted that region 2 from MAG008 is full-length and longer than the longest BGC retrieved from a long-read sequencing-based effort published last year<sup>[78]</sup> making it one of the longest complete BGCs identified directly from a soil metagenome. As of the submission to evaluation of this thesis, no reports of natural product encoding *Lentisphaerae* or *Spirochaetes* were found in literature.

Most common functions hereby predicted (Figure 2B) agree with predominant biosynthetic classes retrieved in other large-scale researches focusing on specialized metabolism<sup>[76,77,80,81]</sup>. On one hand, high counts for NRPS are probably result of fragmented assemblies of these large iterative enzymes from short read sequencing<sup>[82]</sup>. On the other, many terpenes are volatile organic compounds and, therefore, functional in a range of soil moistures as they can travel not only through liquid-filled but also through air-filled soil pores<sup>[77]</sup>. This trait may explain why they are so prevalent in these arid soils (Figure 3D). Furthermore, it has been shown that bacteria can use some terpenes as antibiotics<sup>[2]</sup> and to communicate with each other and with fungi<sup>[83]</sup>; however, most of their ecological functions remain poorly understood. Lastly, bacteriocins are known to be a common feature among bacteria<sup>[84]</sup> and have recently been found to represent the vast majority of phage-encoded BGCs<sup>[85]</sup>.

It's worth mentioning the discrepancy between: (i) KnownClusterBlast output that threw at least a partial match for almost half of BGCs when searched for homologies against the MiBIG repository (Figure 2A and Supplementary Table E3), and (ii) the output of BiG-SCAPE that only linked the ectoine encoding region 255 of MAG005 (Figure 5) to one from the moderately halophilic methylotrophic *Methylophilum marina* isolated from

the Azov Sea, Russia<sup>[86]</sup>. This is explained by the definition of the similarity index delivered by antiSMASH which does not report sequence similarity but the percentual number of genes of the matched sequence that had a homologue in the query. Even though there is a warning regarding the latter at the PDF version of the tool's manual, it has been mostly missed in result interpretation. On the other hand, BiG-SCAPE requests a higher threshold in order to link a BGC with other by including scoring indexes that take into account adjacency and percentage of shared types of PFAM domains and by separately optimizing these for each biosynthetic class<sup>[56]</sup>. Keeping up with state-of-art<sup>[87]</sup> then, it's not uncommon to find few matches<sup>[88]</sup> or none<sup>[76]</sup> with experimentally verified BGCs, at least when dealing with environmental samples. Manual revision of PFAM domains in region 255 evidenced that this ectoine lacked a protein family domain from the reference (acetyltransf\_7) and contained several others. For instance, instead of a IclR helix-turn-motif, it had a MarR-type one (HTH\_27) which has been described to be involved in mechanisms that respond to aromatic compounds<sup>[89]</sup>. This, summed to recent results showing high carotenoid diversity in halo-prokaryotes<sup>[90]</sup>, support the close relationship observed between MAG004, MAG005 and MAG0025, all *Proteobacteria* from S6 harbouring ectoines and terpenes (Figure 3D). Also, region 255 holds a sugar-specific transcriptional regulator of the trehalose/maltose ABC transporter (TrmB) characterized from a hyperthermophilic archaeon<sup>[91]</sup>. Members of the TrmB family have been described as repressors in distinct methanogenic pathways from methylotrophs<sup>[92]</sup> further linking this genome's adaptative strategies to *M. marina's*. Additionally, a peptidase (S8), a penicillinase repressor, and a metalloregulator of the DtxR/MntR family were found in the neighbourhoods of region 255 allowing to hypothesize that strategies to survive during osmotic up and down-shifts of this bacterium may interact with antibiotic and/or metal-related pathways<sup>[93,94,135]</sup>.

Relevant molecules were defined as such that were not predicted for core genes and, after manual revision, did not constitute a common building feature within or across functions (i.e. adh\_short, PP-binding, AMP-binding; Supplementary Table E5). Among these distinct accessory products, many of them were related to proteolytic and nitroreductase activities which are main actors in soil organic nitrogen recycling<sup>[95]</sup>. It

has been reported that extracellular peptidase profiles vary among different soils and that metallopeptidases contribute between 30% and 50% of the overall proteolytic activity in these ecosystems<sup>[96]</sup> which might be part of the explanation to the multiple different types of peptidases found in the TLT dataset. Other frequent relevant molecules were related to oxidative mechanisms and antibiotic biosynthesis<sup>[97-99]</sup>.

Overall, co-occurrences of core functional patterns show at some level taxonomic (Figure 3A) and geographic specificities (Figure 3B) and unspecificities (Figure 3C). Conspicuous evidence of the latter is clade B that groups only genomes from S6 that share arylpolyene and hglE-K capabilities, reported to participate in biofilm formation, protection from oxidative stress<sup>[100]</sup> and nitrogen fixation<sup>[101]</sup>. Another example yet to review of this is the grouping of *Bacteroidetes* at the bottom (Figure 3D) or the distinction of the only genomes of the collection restricted to S2 (MAG102) and S4 (MAG012) even though no rare biosynthetic functions were seen in their repertoires (Supplementary Table E4) suggesting that combinations of biosynthetic features can also constitute or be interpreted as ecological marks.

Examination of functional annotations of accessory genes revealed (i) well-known associations between transporters and regulators involved in natural product biosynthesis and (ii) unreported ones between these and the nature of the metabolic products synthesized at biosynthetic class level. For instance, transposases are enzymes that catalyse transposon integration into and excision from bacterial chromosomes and plasmids, and mediate bacterial conjugation<sup>[102]</sup> which is known to use trans-membrane transporters as gatekeepers<sup>[103]</sup>. More, some transposons have self-conjugative potential and can transfer directly to another cell without having to hitchhike on a plasmid<sup>[104]</sup>. All this taken together allows to suggest that the strict relation observed in Figure 4A of transposases with periplasmic (SMCOG1085 and SMCOG1118) and extracellular-solute(SMCOG1068)/-ligand(SMCOG1282) binding proteins might be indicative of hybrid mechanisms and perhaps RiPPs as possible shapers of bacterial evolution. Another remarkable case is the association of outer membrane iron transporters (TonB-dependant) and the arsenical resistance operon repressor (ArsR) since it has already been described in literature at genomic<sup>[105,106]</sup> and metagenomic<sup>[107]</sup>



lenses. Thus, the results obtained in this study (Figure 4C) point terpenes of the Talabre-Lejía transect as candidate participants of the recently introduced ArsR-regulated arsenic stimulon<sup>[108]</sup>. Regarding sugar metabolism, hierarchical clustering of sm-COGs separated the arabinose regulator (AraC) from those involved in lactose (LacI) and gluconate (GntR) utilization which is outstanding considering evidence supporting arabinose as an antagonist mechanism of lactose in *E. coli*<sup>[109]</sup> and of gluconate in *Vibrio cholerae*<sup>[110]</sup>. Different regulation triggers among hybrids and terpenes might account for such antagonism. A recent study on 20 Illumina-sequenced metagenomes from octocoral microbiomes revealed that PFAM domains, categorical groups of orthologs and biosynthetic core metabolic capacities show significant differences between origins of samples<sup>[111]</sup> further reinforcing the notion that functional information has still great unveiled potential to be recovered from mainstream bioinformatic tools' outputs for the elucidation of taxonomic and geographic behaviours of bacterial communities and their metabolic pathways.

Comparison of architectural relationships between BGCs revealed most of the regions predicted to be singletons (Table 2) meaning that are rare or unique gene clusters that encode unknown enzymes and pathways<sup>[56]</sup>. Sequence similarity networks evidenced shared NP-encoding capabilities to be mostly restricted to phylogenetic boundaries (Figure 5) which is hard to interpret considering the relatively small size of the dataset and that most large-scale studies from environmental samples including gene cluster families approaches deliver results with no resolution for taxonomical assignment. The only two works found that showed this information were based in the RefSeq database<sup>[112]</sup> and the MiBIG repository<sup>[113]</sup> carrying a probable bias, but still a similar extension of this tendency could be observed in both. More importantly, reports of population-specific patterns in diversity of BGCs and phylum-specific transcriptional profiles were reported in were reported in 94 isolates of *Aspergillus flavus*<sup>[130]</sup> and almost 3000 BGCs retrieved from biocrusts<sup>[78]</sup>, respectively, making this light-weight comment one with an encouraging context. Notably, five of nine pairs of genomes linked by a common gene cluster family (Figure 5) belonged to the same functional group (Figure 3). Three out of seven networks showed a generalist distribution (S1-S5), two

were found only in high elevations (S3-S5) and, of the two left, one network was exclusively restricted to S3 (MAG001 and MAG007) and the other was so to S6 (MAG028 and MAG029). The T3PKS-encoding Family 2046 links genomes conjugating geographic and functional criteria by being found in MAG019 and MAG054 that were clustered together in group C and by recruiting another *Actinobacteria* (MAG044, [Figure 3D](#)) that inhabit same sites as the others suggesting that the inclusion of evolutionary inquiries in genomic-fuelled pipelines strengthens the contexts for paving the path towards better understanding bacterial ecology and natural product discovery.

Manual inspection of PFAM domains of the reference BGC from each gene cluster family retrieved, their converted term into GO nomenclature and subsequent key word extraction from original databases ([Supplementary Table E8](#)) were conducted with the aim of naively describing possible functions of these shared metabolites. Congruently with all previous results, most protein family domains found were related to antibiotic biosynthesis, oxidative stress and/or (heavy-)metal resistance ([Table 3 and Supplementary Table E9](#)). The NRPS-like encoding Family 2059 is particularly interesting as it contains a X-Pro dipeptidyl-peptidase C-terminal non-catalytic domain and a X-Pro dipeptidyl-peptidase domain from the S15 family. Both domains can be found at the configuration of cocaine esterase (CocE) which grants bacteria the ability to use cocaine as sole source of carbon and nitrogen<sup>[114]</sup>. This is an unexpected curiosity since cultivation areas of *Erythroxylum coca* during late Inca times reported in Chile (Tacna, Pisagua and Azapa Valley<sup>[115, 116]</sup>) are over 600 km north from the Talabre-Lejía transect and in coastal areas. New chemical evidence of cocaine use in a 1000-year-old ritual bundle was reported in the Sora River valley in south-western Bolivia and, even though it's also a couple of hundreds of kilometres north, the sampling location —also called “La Cueva del Chileno” — refers to the Lípez highlands (3,890 masl) near the Uyuni salar<sup>[117]</sup> which might be a better reference for this study ([Supplementary Figure E3](#)).

Summing up, even though analysing the unassembled metagenomic data would overcome size-derived interpretative restrictions resulting from dealing with few biological samples, genome-resolved functional explorations allow to intuit ecological

and evolutionary patterns of draft organisms —that would be missed at metagenomic scale— with customizable level of detail.

## A bioinformatically viable metatranscriptomic pipeline for assessing natural product metabolism

Considering that bins from which MAGs of BioProject PRJNA291433 were assembled from recruited less than a fifth of all metagenomic scaffolds (~17%)<sup>[79]</sup>, that the collection of accepted drafts consisted of about half of those bins and that genomes hereby studied constitute half of that collection (see [Background section 1](#)), unmapped reads were expected at higher rates than those reported for complete indexes<sup>[118]</sup>. Nevertheless, it was surprising to find at relatively loose mapping parameters that this index was above 99% (from 99.1% to 99.9%, averages per cut-off) and mostly due to short read lengths ([Supplementary Table E1 subtable \(i\)](#)). The first speculation of why such tearing output was retrieved had to do with the artificial concatenation of nucleotide sequences of BGCs, perhaps constituting an invisible object to study at the massive scale of metatranscriptomic data. This was addressed by evaluating same statistics but for the concatenation of whole genomes and the double-trimmed RNA reads. Despite it was shown that the unmapped fraction was lowered to 96.6% in average when filtering thresholds of 30% were applied, the rates caught up with prior trials with the tightening of the parameters up to 45% (99.2%, [Supplementary Table E1 subtable \(iii\)](#)). Biologically, most answers found in literature refer to poor qualities of sequences<sup>[119]</sup>. Bioinformatically, this could be explained if reads had been soft-clipped beyond the threshold, this is: when the aligner disregards the ends of reads —if too many gaps—, the mapped length of the middle part is shorter than the requested<sup>[60]</sup>. The issue is concerning as this is precisely the context where results were obtained from: a flexibilization of matching and scoring minimum values with regards of read length. However, distributions of sequence lengths show that most of reads concentrate at full or nearly full lengths ([Supplementary Figure E2](#)) suggesting that extremely low mappings are more probable to be result of a very complex community composition that required

deeper sequencing coverages summed to the three years-apart samplings for metagenomic and metatranscriptomic inquiries rather than of the degradation of RNA sequences between samplings and sequencings. Nevertheless, at designing stages of every RNA-based experiment, considerations related to the preservation of samples, such as solutions for field stabilization and storage<sup>[120,121]</sup>, should be given special attention to.

Notably, the 22-combinations trial showed that the sensibility of the data to Match and Score adjustments was mostly invariant for differential settings demonstrating close relationship among these parameters ([Supplementary Figure E1](#)). Possible useful applications not overviewed here can have place in more advanced pipelines or in experiments with better mapping statistics translating into a more visually sensible input data.

Regarding quality control, reports mainly classify the data as ‘normal’ or ‘abnormal’, where a sample is considered normal if it is random and diverse, as described at FastQC’s project website. Thus, rather than fixing expected results, modules should serve as spotters for biases<sup>[65]</sup>. For instance, among results that were common between samples, ‘per base sequence content’ is reported as abnormal but wild calling of bases was only observed for the first twenty positions of reads and was then stabilized parallelly ([Supplementary Figure E2](#)), consistent with GC contents close to 60% ([Table 5](#)). In particular, ‘per sequence GC content’ was the most evident module to analyse highlighting samples #6 ([Figure 8](#)), #7 and #8 ([Figure 9](#)) as notably below sufficient, same samples pointed out by ‘overrepresented sequences’ with a warning ([Table 5](#)). Overrepresentation of sequences is frequent for highly duplicated libraries<sup>[121]</sup> as is the case of the studied ones in which redundancy accounts in average for more than a 70% of total reads ([Table 4](#)). Overall, assessments over the qualities of the input data suggest that results obtained from these biological samples might lack accuracy and can mean very difficult interpretation<sup>[119]</sup>. If transferred background methodology as described in [section 2 of Background](#) is complete, further considerations accounting for the high diversity and complexity of the objects of study could improve mapping results, for example, taxonomical and functional classification of reads<sup>[134]</sup>. It’s worth

noting that when the pipeline's branch of concatenated BGCs was fully tried using data originated from overly loose mapping parameters (Match and Score of 15%, [Supplementary Figure E1](#)), processable gene counts and statistically significant results were retrieved from mappings and enriched gene sets, respectively, making it bioinformatically viable.

Despite not retrieving results for Sema-Trap, this tool deserves to be discussed as it comes to fulfil a very much needed space in already solid networks of databases<sup>[33,34,35,122]</sup> and pipelines<sup>[52,56,123,124]</sup> advocated to natural product discovery from a genomic scope. Natural product research, besides of huge interest for industrial purposes<sup>[125-128]</sup>, is a sound way to unveiling niche ecologies by setting specialized metabolites as footprints<sup>[129]</sup> of microbial interactions and adaptative strategies. For these matters, complementing genome mining with transcriptional data is more and more often in the field<sup>[31,75,78,131-133]</sup>. However, until now, approaches taken for defining expression of clusters of genes —even though properly justified— lacked a reference. The latter allowed several different methodologies to arise ranging from simple ones that, for example, used average values of expression of “key” biosynthetic genes<sup>[31]</sup> to others with refined statistics but that falter when setting “100 random genes” as the base of comparisons for permutation-based analyses<sup>[75]</sup>. Against this backdrop, Sema-Trap's scoring metric considers housekeeping genes for ranking the transcription levels of BGCs, allowing other options as well such as the mean of all genes and customization for taxonomical groups if known<sup>[63]</sup>. Hence, since the first strategy here proposed only served the purpose of ranking individual genes, getting by without a fixed notion of what basal expression was, Sema-Trap was added to the pipeline originally proposed.

Other advantage of this new tool is its output window with interactive customizable significance threshold options that don't have to be set *a priori* which allows deeper understanding of the (in)sensibility of expression to adjustments of these in a single run. However, the fact that this manner of visualization is also fixed for local executions is a major impasse for large-scale data management requirements regarding systematization as is the case of metagenomic-based analyses. Related to this, loading

of the input data and running Sema-Trap was pretty slow for the TLT dataset (several hours for loose mapping parameters) which further evidences that computing resources must be first optimized and considerations regarding trackable IDs among runs should be taken if handling transcriptional information originated from environmental samples is considered in future developments of the tool. If we are already talking about potential bugs, one was specially upsetting as, on one hand, Sema-Trap allows two modes for analysis accounting for expression (single condition) and differential expression (two conditions) and, on the other, single condition experimental option is only available for SRA-based submissions, being restricted —for now— to public data. Plus, error messages are undetailed. Also, it can be foreseen that the unmodifiable structure of the pipeline may dissuade usage for large datasets if antiSMASH has been previously run considering that it's easier to replicate the scoring metric than dealing with new nomenclature.

Technical utilities for the output of the transcriptomic scoped pipeline are expected to allow (i) ranking of genomes according to their transcriptional activity of NP-encoding BGCs, (ii) ranking of expressed BGCs, (iii) ranking of differentially expressed genes from enriched gene sets constructed upon taxonomical and functional criteria, (iv) unveiling of co-expression patterns between types of core enzymes, and (v) the construction of networks accounting for *in situ* active regulation of specialized metabolites biosynthesis.

Finally, transcriptomic-based analyses downstream of the genomic approaches evidenced that first steps in every pipeline account for great part of reproducibility capacities. For instance, if annotation files were delivered as input for antiSMASH, names of genes would be trackable. However, arbitrary renaming of scaffolds when summiting jobs to the tool's webserver restricts insertion of results in whole genome-based pipelines with a functional lens, unless manual labelling (column 'original\_scaffold' in [Supplementary Table E5](#)). Other critical consideration is the checking of description lines of the input fasta files as these constitute the only entrance for identifying BGCs later, so if datasets containing several genomes are processed, unique IDs for contigs are required, within and across files.

## Conclusions

Main observations derived from the results retrieved in this project relate to the overlooked potential of functional annotations retrieved through basic options in mainstream genomic pipelines. Particularly, in natural product research this information can be translated into co-occurrence patterns that, if sufficient objects to study, may serve as enlightening spotters of both spread and unique metabolic capabilities to be fixed as targets when exploring ecological and evolutionary dynamics of microbial communities. Genome-resolved logics demonstrated to be useful for the description of functional repertoires as it allows to break the data down in a categorical manner and, thus, to ask the same question many times for more specific features than when characterizing whole communities, which sets a clearer context for interpretation.

Regarding the Talabre-Lejía transect, composition of MAGs changed at high taxonomic ranks among the six sampling sites leaning towards a sample-specific behaviour. *Acidobacteria* prevailed by means of relative abundance and compared frequency of BGCs and gathers most distinct genomes in the dataset. Results here obtained supports previous evidence of NRPS, terpenes and bacteriocins as predominant biosynthetic classes predicted in soil microbiomes and environmental samples in general. Most common biological functions of specialized metabolites examined here are mainly advocated to antibiotic biosynthesis, nitrogen metabolism, oxidative stress, and metal resistance. Functional exploration allowed to visualize previously unreported associations between transporters and regulators involved in natural product biosynthesis and the nature of the metabolic products synthesized at biosynthetic class level.

Overall, the studied samples evidenced highly complex bacterial communities that required higher depth of sequencing coverage to be properly assessed. Genomic repertoires of specialized metabolites —or fragments of them— can constitute distinctive features and be interpreted as consistent biological footprints to track niche adaptative strategies. Inspections of regulatory dynamics are expected to sharpen these

ecological marks observed at genomic level and, thus, to contribute paving the path towards better understanding of *in situ* competitive and cooperative interactions of microorganisms.

Finally, regarding analyses of transcriptomic data from environmental samples, even though subject of major interest nowadays in the NP field and recent advances in pipelines specifically dedicated to dealing with BGCs, it's still early times for mainstream and automatized processing, while quality preservation of biological samples and experimental designing are key for successful informative data management.



## References

- [1] Hibbing, M. *et al.* (2009). "Bacterial competition: surviving and thriving in the microbial jungle." *Nat Rev Microbiol.* 8(1):15-25.
- [2] Tyc, O. *et al.* (2017). "The ecological role of volatile and soluble secondary metabolites produced by soil bacteria." *Trends Microbiol.* 25(4):280-92.
- [3] Yan, Y. *et al.* (2020). "Recent developments in self-resistance gene directed natural product discovery." *Nat Prod Rep.* 37(7):879-92.
- [4] Brescia, F. *et al.* (2020). "The rhizosphere signature on the cell motility, biofilm formation and secondary metabolite production of a plant-associated *Lysobacter* strain." *Microbiol Res.* 234:126424.
- [5] Arnaouteli, S. *et al.* (2021). "Bacillus subtilis biofilm formation and social interactions." *Nat Rev Microbiol.* 19(9):600-14.
- [6] Pavan, M.E. *et al.* (2020). "Melanin biosynthesis in bacteria, regulation and production perspectives." *Appl Microbiol Biotechnol.* 104:1357-70.
- [7] Singh, S. *et al.* (2021). "Microbial melanin: Recent advances in biosynthesis, extraction, characterization, and applications." *Biotechnol Adv.* 53:107773.
- [8] Czech, L. *et al.* (2018). "Role of the extremolytes ectoine and hydroxyectoine as stress protectants and nutrients: genetics, phylogenomics, biochemistry, and structural analysis." *Genes (Basel)* 9(4):177.
- [9] Jones, S.E. *et al.* (2019). "Streptomyces volatile compounds influence exploration and microbial community dynamics by altering iron availability." *mBio* 10(2):e00171-19.
- [10] Emami, S. *et al.* (2019). "Effect of rhizospheric and endophytic bacteria with multiple plant growth promoting traits on wheat growth." *Environ Sci Pollut. Res* 26:19804-13.
- [11] Someya, N. *et al.* (2020). "Diversity of antibiotic biosynthesis gene-possessing rhizospheric Fluorescent Pseudomonads in Japan and their biocontrol efficacy." *Microbes Environ.* 35(2):ME19155.
- [12] Gutiérrez-Santa Ana, A. *et al.* (2020). "Volatile emission compounds from plant growth-promoting bacteria are responsible for the antifungal activity against *F. solani*". *3 Biotech.* 10:292.
- [13] Korp, J. *et al.* (2016). "Antibiotics from predatory bacteria." *Beilstein J Org Chem.* 12:594-607.
- [14] Cornforth, D.M. & Foster, K.R. (2013). "Competition sensing: the social side of bacterial stress responses." *Nat Rev Microbiol.* 11:285-93.
- [15] Ueda, K. *et al.* (2000). "Wide distribution of interspecific stimulatory events on antibiotic production and sporulation among *Streptomyces* species." *J Antibiot.* 53:979-98.
- [16] Jensen, P.R. (2016). "Natural Products and the Gene Cluster Revolution." *Trends Microbiol.* 24(12):968-77.

- [17] Garagounis, C. *et al.* (2021). "Unraveling the roles of plant specialized metabolites: using synthetic biology to design molecular biosensors." *New Phytologist* 231(4):1338-52.
- [18] Chevrette, M.G. *et al.* (2020). "Evolutionary dynamics of natural product biosynthesis in bacteria." *Nat Prod Rep.* 37:566-99.
- [19] Peracchi, A. (2018). "The limits of enzyme specificity and the evolution of metabolism." *Trends Biochem Sc.* 43(12):984-96.
- [20] Copley, S.D. (2017). "Shining a light on enzyme promiscuity." *Curr Op Struct Biol.* 47:167-75.
- [21] Fischbach, M.A. *et al.* (2007). "The evolution of gene collectives: How natural selection drives chemical innovation." *PNAS* 105(12):4601-08.
- [22] Medema, M.H. *et al.* (2021). "Mining genomes to illuminate the specialized chemistry of life". *Nat Rev Genetics.* 22(9):553-71.
- [23] Firn, R.D. & Jones, C.G. (2003). "Natural products – a simple model to explain chemical diversity". *Nat Prod Rep.* 20:382-91.
- [24] Firn, R.D. & Jones, C.G. (2009). "A Darwinian view of metabolism: molecular properties determine fitness." *J Exp Bot.* 60(3):719-26.
- [25] Russel, A.H. & Truman, A.W. (2020). "Genome mining strategies for ribosomally synthesised and post-translationally modified peptides." *Comput Struct Biotechnol J.* 18:1838-51.
- [26] Sardar, D. & Schmidt, E.W. (2016). "Combinatorial biosynthesis of RiPPs: docking with marine life." *Curr Op Chem Biol.* 31:15-21.
- [27] Jenke-Kodama, H. *et al.* (2006). "Natural Biocombinatorics in the polyketide synthase genes of the actinobacterium *Streptomyces avermitilis*". *PLoS Comput Biol.* 2(9):e132.
- [28] Du, L. *et al.* (2001). "Hybrid peptide–polyketide natural products: biosynthesis and prospects toward engineering novel molecules." *Metabol Eng.* 3:78-95.
- [29] Dror, B. *et al.* (2020) "State-of-the-art methodologies to identify antimicrobial secondary metabolites in soil bacterial communities: a review." *Soil Biology and Biochemistry* 147:107838.
- [30] Schwecke, T. *et al.* (1995) "The biosynthetic gene cluster for the polyketide immunosuppressant rapamycin." *PNAS* 92(17):7839-43.
- [31] Amos, G.C. *et al.* (2017). "Comparative transcriptomics as a guide to natural product discovery and biosynthetic gene cluster functionality." *PNAS* 114(52):E11121-30.
- [32] Blin, K. *et al.* (2017). "Recent development of antiSMASH and other computational approaches to mine secondary metabolite biosynthetic gene clusters." *Briefings in Bioinformatics* 20(4):1103-13.
- [33] Palaniappan, K. *et al.* (2020). "IMG-ABC v.5.0: an update to the IMG/Atlas of Biosynthetic Gene Clusters Knowledgebase". *Nucl Acids Res.* 48(D1):D422-30.
- [34] Kautsar, S.A. *et al.* (2019). "MIBiG 2.0: a repository for biosynthetic gene clusters of known function". *Nucl Acids Res.* 48(D1):D454-48.
- [35] Blin, K. *et al.* (2021). "The antiSMASH database version 3: increased taxonomic coverage and new query features for modular enzymes." *Nucl Acids Res.* 49(D1):D639-43.

- [36] Chen, L. *et al.* (2020). "Accurate and complete genomes from metagenomes." *Genome Res.* 30(3):315-33.
- [37] Shykla, B. *et al.* (2022). "Teixobactin kills bacteria by a two-pronged attack on the cell envelope." *Nature* 608:390-96.
- [38] Mandakovic, D. *et al.* (2018) "Structure and co-occurrence patterns in microbial communities under acute environmental stress reveal ecological factors fostering resilience." *Sci Rep.* 8:5875.
- [39] Díaz, F.P. *et al.* (2016). "Nitrogen cycling in an extreme hyperarid environment inferred from  $\delta^{15}\text{N}$  analyses of plants, soils and herbivore diet." *Sci Rep.* 6:22226.
- [40] Tassi, F. *et al.* (2009). "The magmatic- and hydrothermal-dominated fumarolic system at the Active Crater of Lascar volcano, northern Chile." *Bull Volcanol* 71:171-83.
- [41] Mandakovic, D. *et al.* (2018) "Microbiome analysis and bacterial isolation from Lejía Lake soil in Atacama Desert." *Extremophiles* 22:665-73.
- [42] Hyatt, D. *et al.* (2012). "Gene and translation initiation site prediction in metagenomic sequences." *Bioinformatics* 28:2223-30.
- [43] Peng, Y. *et al.* (2012). "IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth." *Bioinformatics* 28:1420-28.
- [44] Albertsen, M. *et al.* (2013). "Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes." *Nat Biotechnol.* 31:533-38.
- [45] Alneberg, J. *et al.* (2014). "Binning metagenomic contigs by coverage and composition." *Nat Methods.* 11:1144-46.
- [46] Parks, D.H. *et al.* (2015). "CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes." *Genome Res.* 25:1043-55.
- [47] Zerbino, D.R. & Birney, E. (2008). "Velvet: Algorithms for de novo short read assembly using de Bruijn graphs." *Genome Res.* 18:821-29.
- [48] Segata, N. *et al.* (2013). "PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes." *Nat Commun.* 4:2304.
- [49] Langmead, B. & Salzberg, S.L. (2012). "Fast gapped-read alignment with Bowtie 2." *Nat Methods.* 9:357-59.
- [50] Quinlan, A.R. & Hall, I.M. (2010). "BEDTools: A flexible suite of utilities for comparing genomic features." *Bioinformatics* 26:841-42.
- [51] Bolger, A.M. *et al.* (2014). "Trimmomatic: A flexible trimmer for Illumina sequence data." *Bioinformatics* 30(15):2114-20.
- [52] Blin, K. *et al.* (2019). "antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline." *Nucl Acids Res.* 47(W1):W81-87.
- [53] Mistry, J. *et al.* (2021). "Pfam: The protein families database in 2021." *Nucl Acids Res.* 47(D1):D412-19.
- [54] Stajich, J.E. *et al.* (2002). "The Bioperl toolkit: Perl modules for the life sciences." *Genome Res.* 12(10):1611-18.

- [55] R Core Team (2022). "R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria." [www.R-project.org/](http://www.R-project.org/)
- [56] Navarro-Muñoz, J.C. *et al.* (2020). "A computational framework to explore large-scale biosynthetic diversity." *Nat Chem Biol.* 16:60-68.
- [57] Cline, M.S. *et al.* (2007). "Integration of biological networks and gene expression data using Cytoscape." *Nat Protocols.* 2:2366-82.
- [58] The Gene Ontology Consortium (2021). "The Gene Ontology resource: enriching a GOLD mine." *Nucleic Acids Res.* 49(D1):D325-34.
- [59] Dragičević, M.B. *et al.* (2020). "ragp: Pipeline for mining of plant hydroxyproline-rich glycoproteins with implementation in R." *Glycobiology* 30(1):19-35
- [60] Dobin, A. *et al.* (2013). "STAR: ultrafast universal RNA-seq aligner." *Bioinformatics* 29(1):15-21.
- [61] Love, M.I. *et al.* (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." *Genome Biol* 15:550.
- [62] Subramanian, A. *et al.* (2005). "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles." *PNAS* 102(43):15545-50.
- [63] Mungan, M.D., *et al.* (2022). "Secondary Metabolite Transcriptomic Pipeline (SeMa-Trap), an expression-based exploration tool for increased secondary metabolite production in bacteria." *Nucl Acids Res.* 50(W1):W682-89. 5 July 2022.
- [64] Seemann, T. (2014). "Prokka: rapid prokaryotic genome annotation." *Bioinformatics* 30(14):2068-69.
- [65] Andrews, S. (2010). "FastQC: A Quality Control Tool for High Throughput Sequence Data".
- [66] Martin, M. (2011). "Cutadapt removes adapter sequences from high-throughput sequencing reads." *EMBnet.journal* 17(1):10-12.
- [67] Crits-Christoph, A. *et al.* (2013). "Colonization patterns of soil microbial communities in the Atacama Desert." *Microbiome* 1:28.
- [68] Marasco, R. *et al.* (2022). "The plant rhizosphere–root niche is an edaphic "mini-oasis" in hyperarid deserts with enhanced microbial competition." *ISME Commun.* 2:47.
- [69] Kalam, S. *et al.* (2020). "Recent understanding of soil Acidobacteria and their ecological significance: a critical review." *Front Microbiol.* 11:580024.
- [70] Kielak, A.M. *et al.* (2016) "The Ecology of Acidobacteria: moving beyond genes and genomes." *Front Microbiol.* 7:744.
- [71] Reid, R.P. *et al.* (2021). "Electrical conductivity as a driver of biological and geological spatial heterogeneity in the Puquios, Salar de Llamara, Atacama Desert, Chile." *Sci Rep.* 11:12769.
- [72] Hwang, Y. *et al.* (2021) "Leave no stone unturned: individually adapted xerotolerant Thaumarchaeota sheltered below the boulders of the Atacama Desert hyperarid core." *Microbiome* 9:234.
- [73] Finstad, K.M. *et al.* (2017). "Microbial community structure and the persistence of cyanobacterial populations in salt crusts of the hyperarid Atacama Desert from genome-resolved metagenomics." *Front Microbiol.* 8:1435

- [74] Martijn, J. *et al.* (2020). "Hikarchaea demonstrate an intermediate stage in the methanogen-to-halophile transition." *Nat Commun.* 11:5490.
- [75] Crits-Christoph, A. *et al.* (2018). "Novel soil bacteria possess diverse genes for secondary metabolite biosynthesis." *Nature* 558:440-44.
- [76] Crits-Christoph, A. *et al.* (2022). "A widely distributed genus of soil Acidobacteria genomically enriched in biosynthetic gene clusters." *ISME Commun.* 2:70.
- [77] Sharrar, A.M. *et al.* (2020). "Bacterial secondary metabolite biosynthetic potential in soil varies with phylum, depth, and vegetation type." *mBio* 11:e00416-20.
- [78] Van Goethem, M.W. *et al.* (2021). "Long-read metagenomics of soil communities reveals phylum-specific secondary metabolite dynamics." *Nat Comm Biol.* 4:1302.
- [79] González, M. *et al.* (unpublished). "Genomic analysis provides insights into the functional capacity of soil bacteria communities inhabiting an altitudinal gradient in the Atacama Desert."
- [80] Chen, R. *et al.* (2020). "Discovery of an Abundance of Biosynthetic Gene Clusters in Shark Bay Microbial Mats." *Front Microbiol.* 11:1950.
- [81] Cuadrat, R.R.C. *et al.* (2018). "Recovering genomics clusters of secondary metabolites from lakes using genome-resolved metagenomics." *Front Microbiol.* 9:251.
- [82] Olm, M.R. *et al.* (2020). "Consistent metagenome-derived metrics verify and delineate bacterial species boundaries." *mSystems* 5:e00731-19.
- [83] Schmidt, R. *et al.* (2017). "Fungal volatile compounds induce production of the secondary metabolite sodorifen in *Serratia plymuthica* PRI-2C". *Sci Rep.* 7:862.
- [84] de Oliveira, I.M.F. (2022). "Whole-genome sequencing and comparative genomic analysis of antimicrobial producing *Streptococcus lutetiensis* from the rumen. *Microorganisms* 10:551.
- [85] Dragos, *et al.* (2021). "Phages carry interbacterial weapons encoded by biosynthetic gene clusters." *Curr Biol.* 31:3479–89.
- [86] Reshetnikov, A. *et al.* (2011). "Diversity and phylogeny of the ectoine biosynthesis genes in aerobic, moderately halophilic methylotrophic bacteria." *Extremophiles* 15:653-63.
- [87] Kenshole, E. *et al.* (2021). "Natural product discovery through microbial genome mining." *Curr Op Chem Biol.* 60:47-54.
- [88] Rego, A. *et al.* (2021). "Secondary metabolite biosynthetic diversity in Arctic Ocean metagenomes." *Microb Genom.* 7(12):000731.
- [89] Alekshun, M.N. & Levy, .SB. (1999). "The mar regulon: multiple resistance to antibiotics and other toxic chemicals." *Trends Microbiol.* 7(10):410-3.
- [90] Serrano, S. *et al.* (2022). "Haloarchaea have a high genomic diversity for the biosynthesis of carotenoids of biotechnological interest." *Res Microbiol.* 173(3):103919.
- [91] Lee, S. *et al.* (2003). "TrmB, a sugar-specific transcriptional regulator of the trehalose/maltose ABC transporter from the hyperthermophilic archaeon *Thermococcus litoralis*." *Jour Biol Chem.* 278(2):983-90.
- [92] Reichlen, M.J. *et al.* (2012). "MreA functions in the global regulation of methanogenic pathways in *Methanosarcina acetivorans*." *MBio* 3(4):e00189

- [93] Wittman, V. *et al.* (1993). "Functional domains of the penicillinase repressor of *Bacillus licheniformis*." *J Bacteriol.* 175(22):7383-90.
- [94] Guedon, E. & Helmann, J.D. (2003). "Origins of metal ion selectivity in the DtxR/MntR family of metalloregulators." *Mol Microbiol.* 48(2):495-506.
- [95] Zhou, Z. *et al.* (2020). "Genome diversification in globally distributed novel marine Proteobacteria is linked to environmental adaptation." *ISME J* 14: 2060–77.
- [96] Nguyen, T. *et al.* (2019). "Contribution of different catalytic types of peptidases to soil proteolytic activity." *Soil Biol Biochem.* 138:107578.
- [97] Wang, S. *et al.* (2020). "Studies of lincosamide formation complete the biosynthetic pathway for lincomycin A." *PNAS* 117(49):24794-801.
- [98] Sofia, H.J. *et al.* (2001). "Radical SAM, a novel protein superfamily linking unresolved steps in familiar biosynthetic pathways with radical mechanisms: functional characterization using new analysis and information visualization methods." *Nucleic Acids Res.* 29(5):1097-106.
- [99] Sultana, A. *et al.* (2004). "Structure of the polyketide cyclase SnoaL reveals a novel mechanism for enzymatic aldol condensation." *EMBO J.* 23(9):1911-21.
- [100] Johnston, I. *et al.* (2020) "Identification of essential genes for *Escherichia coli* aryl polyene biosynthesis and function in biofilm formation." *npj Biofilms Microbiomes* 7:56.
- [101] Campbell, E.L. *et al.* (1997). "A polyketide-synthase-like gene is involved in the synthesis of heterocyst glycolipids in *Nostoc punctiforme* strain ATCC 29133." *Arch Microbiol.* 167(4):251-58.
- [102] Roberts, A.P. *et al.* (2008). "Revised nomenclature for transposable genetic elements." *Plasmid* 60(3):167-73.
- [103] Gomis-Rüth, F.X. & Coll, M. (2001). "Structure of TrwB, a gatekeeper in bacterial conjugation." *Int J Biochem Cell Biol.* 33(9):839-43.
- [104] Guglielmini, J. *et al.* (2011). "The repertoire of ICE in prokaryotes underscores the unity, diversity, and ubiquity of conjugation." *PLoS Genet* 7(8): e1002222.
- [105] Kaur, A. *et al.* (2021). "Discerning the role of a functional arsenic-resistance cassette in the evolution and adaptation of a rice pathogen." *Microbial Genomics* 7(7):000608.
- [106] Mohapatra, B. *et al.* (2019). "Comparative genome analysis of arsenic reducing, hydrocarbon metabolizing groundwater bacterium *Achromobacter* sp. KAs 3-5T explains its competitive edge for survival in aquifer environment." *Genomics* 111(6):1604-19.
- [107] Das, S. *et al.* (2017). "A metagenomic approach to decipher the indigenous microbial communities of arsenic contaminated groundwater of Assam." *Genomics Data* 12:89-96.
- [108] Rawle, R. *et al.* (2021). "Introducing the ArsR-Regulated Arsenic Stimulon." *Front Microbiol.* 12:630562.
- [109] Ammar, E.M. *et al.* (2018). "Regulation of metabolism in *Escherichia coli* during growth on mixtures of the non-glucose sugars: arabinose, lactose, and xylose." *Sci Rep* 8:609.
- [110] Golder, T. *et al.* (2020). "Nonmetabolizable arabinose inhibits *Vibrio cholerae* growth in m9 medium with gluconate as the sole carbon source." *Jpn J Infect Dis.* 73(5):343-48.
- [111] Keller-Costa, T. *et al.* (2021). "Metagenomic insights into the taxonomy, function, and dysbiosis of prokaryotic communities in octocorals." *Microbiome* 9:72.

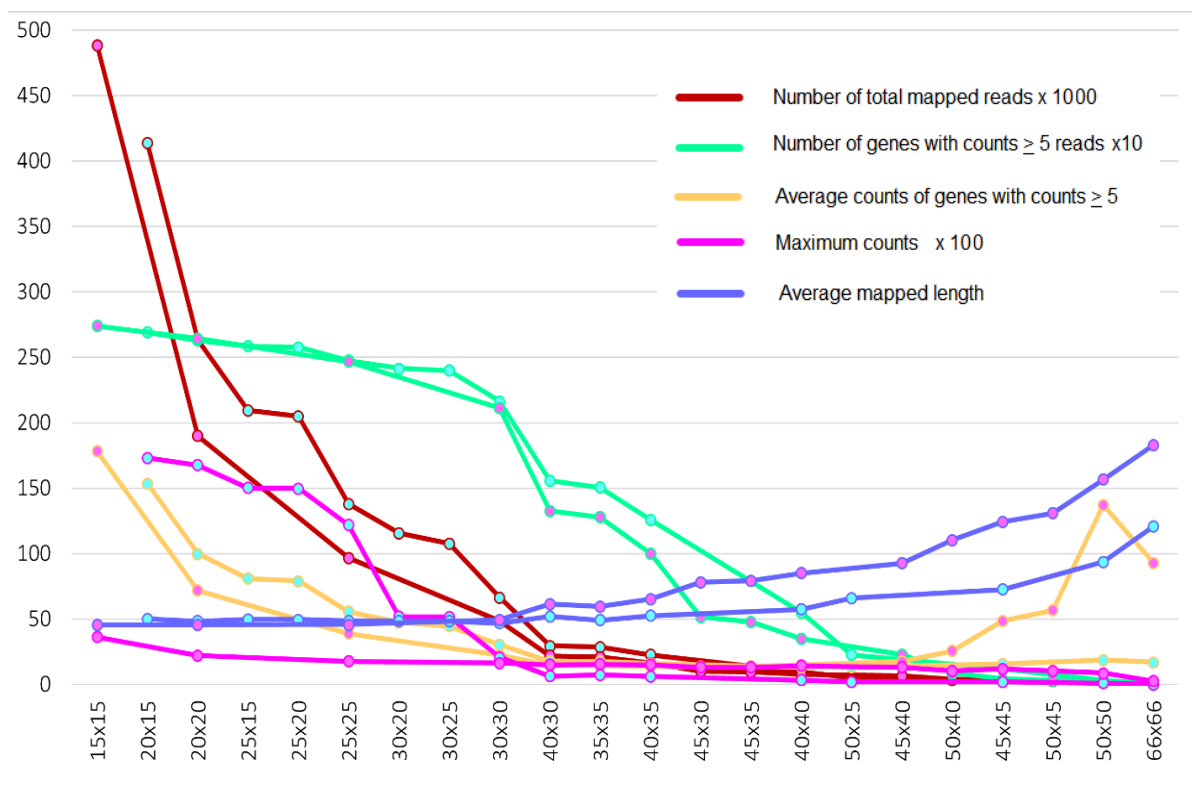
- [112] Li, S. & Horsman, P. (2022). "An inventory of early branch points in microbial phosphonate biosynthesis." *Microbial Genomics* 8:000781.
- [113] Crits-Christoph, A. *et al.* (2021). "Transporter genes in biosynthetic gene clusters predict metabolite characteristics and siderophore activity." *Genome Res.* 31:239-50.
- [114] Bresler, M.M. *et al.* (2000). "Gene cloning and nucleotide sequencing and properties of a cocaine esterase from *Rhodococcus* sp. strain MB1." *Appl Environ Microbiol.* 66(3):904-8.
- [115] Rostworowski, M.D.C. (1973). "Plantaciones prehispánicas de coca en la vertiente del Pacífico." *Revista del Museo Nacional* 39:193-224.
- [116] Rivera, M.A. *et al.* (2005). "Antiquity of coca-leaf chewing in the South Central Andes: a 3,000 year archaeological record of coca-leaf chewing from Northern Chile." *Jour Psychoact Drugs.* 37(4):455-58.
- [117] Miller, M.J. *et al.* (2019). "Chemical evidence for the use of multiple psychotropic plants in a 1,000-year-old ritual bundle from South America." *PNAS* 116(23):11207-12.
- [118] Jacksch, S. *et al.* (2021). "Metatranscriptomic Analysis of Bacterial Communities on Laundered Textiles: A Pilot Case Study." *Microorganisms* 9:1591.
- [119] Dobin, A. & Gingeras, T.R. (2015). "Mapping RNA-seq reads with STAR." *Curr Protoc Bioinform.* 51:11.14.1-19.
- [120] Westermann, A.J. (2017). "Resolving host-pathogen interactions by dual RNA-seq." *Plos Pathog.* 13(2):e1006033.
- [121] Chung, M. *et al.* (2021). "Best practices on the differential expression analysis of multi-species RNA-seq." *Genome Biol.* 22:121.
- [122] Flissi, A. *et al.* (2020). "Norine: update of the nonribosomal peptide resource." *Nucl Acids Res.* 48(D1):D465-69.
- [123] Sélem-Mojica, N. *et al.* (2019). "EvoMining reveals the origin and fate of natural product biosynthetic enzymes." *Microb Genom.* 5(12):e000260.
- [124] Ziemert, N. *et al.* (2012). "The Natural Product domain seeker NaPDoS: A phylogeny based bioinformatic tool to classify secondary metabolite gene diversity." *PLoS ONE* 7(3): e34064.
- [125] Merarchi, M. *et al.* (2021). "Natural products and phytochemicals as potential anti-SARS-CoV-2 drugs." *Phytoterapy Res.* 35(19):5384-96.
- [126] Alves, C. & Diederich, M. (2021). "Marine natural products as anticancer agents." *Marine Drugs.* 19(8):447.
- [127] Huang, H. *et al.* (2018). "Abyssomicin monomers and dimers from the marine-derived *Streptomyces koyangensis* SCSIO 5802." *J Nat Prod.* 81(8):1892-98.
- [128] Nowruzzi, B. *et al.* (2020). "The cosmetic application of cyanobacterial secondary metabolites." *Algal Res.* 49:101959.
- [129] Hoffmann, T. *et al.* (2018). "Correlating chemical diversity with taxonomic distance for discovery of natural products in myxobacteria." *Nat Commun.* 9:803.
- [130] Drott, M.T. (2021). "Microevolution in the pansecondary metabolome of *Aspergillus flavus* and its potential macroevolutionary implications for filamentous fungi." *PNAS* 118(21):e2021683118.

- [131] Fierer, N. (2017). "Embracing the unknown: disentangling the complexities of the soil microbiome." *Nat Rev Microbiol.* 15:579-90.
- [132] Galambos, D. *et al.* (2019). Genome-resolved metagenomics and metatranscriptomics reveal niche differentiation in functionally redundant microbial communities at deep-sea hydrothermal vents." *Environ Microbiol* 21(11), 4395-410.
- [133] Graham-Taylor, C.*et al.* (2020). "A detailed in silico analysis of secondary metabolite biosynthesis clusters in the genome of the broad host range plant pathogenic fungus *Sclerotinia sclerotiorum*." *BMC Genomics* 21:7.
- [134] Martínez, X. *et al.* (2016). "MetaTrans: an open-source pipeline for metatranscriptomics." *Sci Rep* 6:26447.
- [135] Li, H. *et al.* (2016). "Characterization of a new S8 serine protease from marine sedimentary photobacterium sp. A5-7 and the function of its protease-associated domain." *Front Microbiol* 7: 2016.



## Extended data

**Supplementary Figure E1.** STAR cut-off trials (n=22) for `-outFilterMatchNminOverLread` and `(x) -outFilterScoreMinOverLread` parameters in two RNA-seq samples. Cyan dots: ID #4; pink dots: ID #9.



**Supplementary Table E1.** STAR cut-off statistics for concatenated (i) BGCs before trimming and for (ii) BGCs and (iii) genomes after trimming of samples from (a) S1 and (b) S5. TG: Trim Galore.

(a)

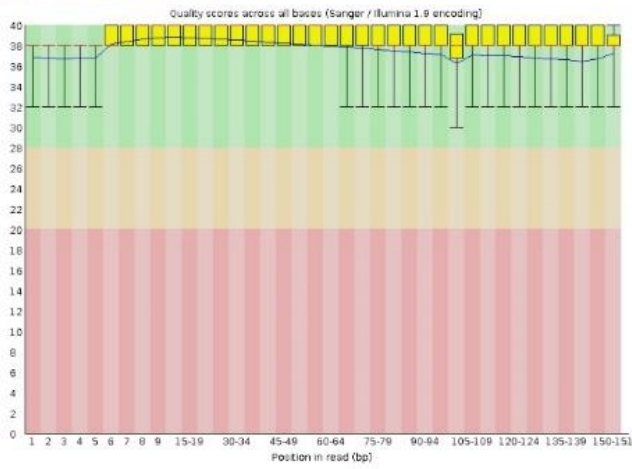
sample.ID_site_triplicate	4_S1_i				5_S1_ii				6_S1_iii			
(i) ref index: bgcs	30x30	35x35	40x40	45x45	30x30	35x35	40x40	45x45	30x30	35x35	40x40	45x45
%_uniquely.mapped	0.54%	0.23%	0.08%	0.03%	0.46%	0.20%	0.07%	0.02%	0.69%	0.29%	0.10%	0.00%
%_mapped.length	16.4%	17.2%	20.1%	25.4%	15.5%	17.2%	21.9%	28.5%	14.9%	16.0%	19.8%	52.1%
mismatch.rate	8.80%	8.49%	8.42%	8.04%	8.28%	7.89%	7.53%	7.02%	8.55%	7.74%	6.91%	1.51%
%_multimapping	0.18%	0.06%	0.01%	0.00%	0.16%	0.05%	0.01%	0.00%	0.21%	0.07%	0.01%	0.00%
%_unmapped:too.short	99.01%	99.44%	99.64%	99.70%	99.23%	99.60%	99.77%	99.82%	98.80%	99.34%	99.59%	99.98%
(ii) ref index: bgcs (TG)	30x30	35x35	40x40	45x45	30x30	35x35	40x40	45x45	30x30	35x35	40x40	45x45
%_uniquely.mapped	0.09%	0.03%	0.02%	0.01%	0.05%	0.03%	0.02%	0.01%	0.03%	0.02%	0.01%	0.01%
%_mapped.length	35.8%	40.6%	44.6%	48.7%	37.9%	45.1%	49.6%	51.6%	40.6%	54.3%	64.0%	71.9%
mismatch.rate	8.67%	7.75%	7.01%	6.41%	7.99%	6.76%	5.99%	5.44%	6.11%	3.94%	2.71%	2.00%
%_multimapping	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
%_unmapped:too.short	99.90%	99.95%	99.97%	99.98%	99.94%	99.96%	99.97%	99.98%	99.95%	99.97%	99.98%	99.98%
(iii) ref index: genomes (TG)	30x30	35x35	40x40	45x45	30x30	35x35	40x40	45x45	30x30	35x35	40x40	45x45
%_uniquely.mapped	2.05%	1.31%	0.88%	0.60%	2.13%	1.31%	0.85%	0.54%	1.30%	0.72%	0.41%	0.26%
%_mapped.length	46.3%	51.6%	56.4%	60.9%	44.7%	49.8%	54.6%	59.4%	43.1%	49.0%	54.8%	60.2%
mismatch.rate	6.23%	5.35%	4.63%	4.00%	5.89%	5.14%	4.47%	3.79%	6.67%	5.88%	4.97%	4.16%
%_multimapping	0.42%	0.23%	0.14%	0.08%	0.33%	0.18%	0.11%	0.06%	0.28%	0.16%	0.08%	0.04%
%_unmapped:too.short	97.53%	98.46%	98.99%	99.32%	97.55%	98.50%	99.04%	99.40%	98.42%	99.13%	99.50%	99.70%

(b)

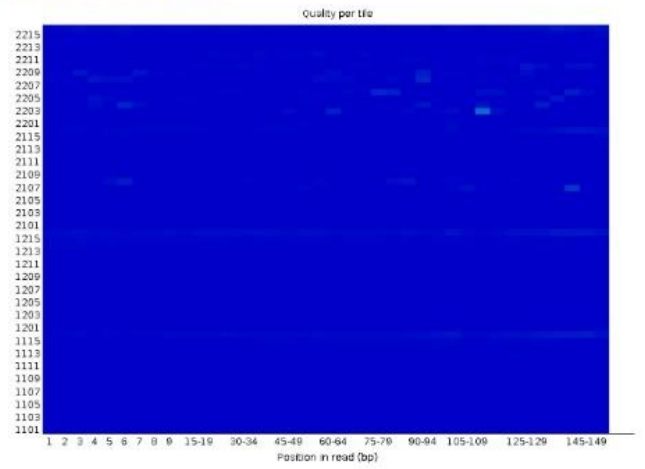
sample.ID_site_triplicate	7_S5_i				8_S5_ii				9_S5_iii			
(i) ref index: bgcs	30x30	35x35	40x40	45x45	30x30	35x35	40x40	45x45	30x30	35x35	40x40	45x45
%_uniquely.mapped	0.42%	0.19%	0.07%	ERROR	0.29%	0.08%	0.03%	0.00%	0.41%	0.18%	0.07%	0.03%
%_mapped.length	14.5%	15.1%	17.5%	ERROR	12.7%	13.9%	14.2%	46.3%	17.1%	20.6%	29.5%	43.1%
mismatch.rate	8.40%	7.97%	7.59%	ERROR	8.62%	10.29%	9.60%	4.66%	7.30%	6.29%	5.07%	3.81%
%_multimapping	0.14%	0.05%	0.01%	ERROR	0.10%	0.02%	0.00%	0.00%	0.14%	0.05%	0.01%	0.00%
%_unmapped:too.short	99.13%	99.45%	99.61%	ERROR	99.10%	99.40%	99.46%	99.95%	99.33%	99.65%	99.80%	99.85%
(ii) ref index: bgcs (TG)	30x30	35x35	40x40	45x45	30x30	35x35	40x40	45x45	30x30	35x35	40x40	45x45
%_uniquely.mapped	0.03%	0.01%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.05%	0.04%	0.03%	0.02%
%_mapped.length	36.5%	42.6%	49.3%	57.1%	27.9%	33.1%	33.4%	37.4%	45.0%	51.9%	56.2%	59.3%
mismatch.rate	7.97%	6.54%	5.45%	4.50%	10.45%	10.08%	9.48%	5.77%	4.58%	3.63%	3.13%	2.83%
%_multimapping	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
%_unmapped:too.short	99.94%	99.96%	99.97%	99.97%	99.96%	99.96%	99.96%	99.96%	99.94%	99.96%	99.97%	99.97%
(iii) ref index: genomes (TG)	30x30	35x35	40x40	45x45	30x30	35x35	40x40	45x45	30x30	35x35	40x40	45x45
%_uniquely.mapped	6.44%	4.83%	3.41%	1.02%	3.21%	1.96%	1.15%	0.64%	3.17%	2.28%	1.68%	1.21%
%_mapped.length	45.1%	47.6%	49.4%	54.8%	44.5%	48.4%	52.3%	56.5%	50.1%	54.9%	59.2%	63.6%
mismatch.rate	5.20%	4.84%	4.64%	4.61%	6.78%	6.21%	5.60%	4.87%	5.76%	5.23%	4.79%	4.43%
%_multimapping	0.40%	0.21%	0.12%	0.07%	0.28%	0.14%	0.07%	0.04%	0.39%	0.28%	0.19%	0.14%
%_unmapped:too.short	93.16%	94.96%	96.47%	98.91%	96.51%	97.89%	98.78%	99.32%	96.45%	97.44%	98.13%	98.65%

Supplementary Figure E2. Trim Galore! module reports common for all samples.

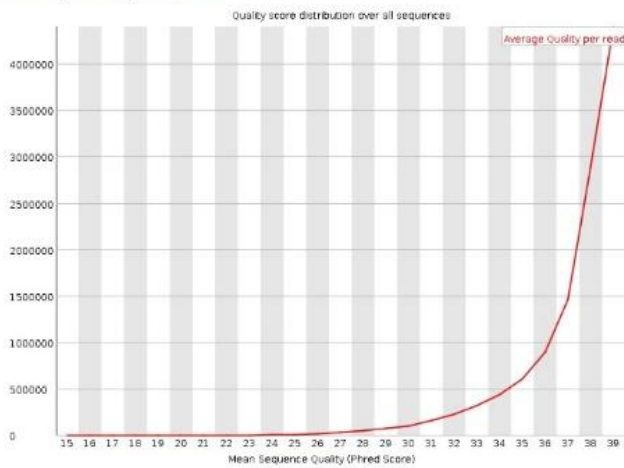
✔ Per base sequence quality



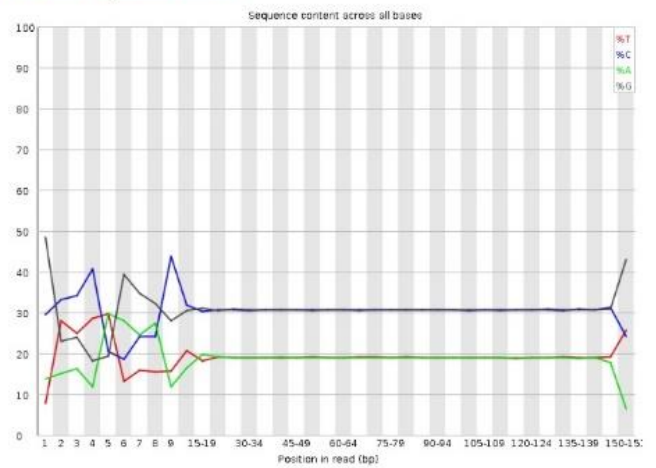
✔ Per tile sequence quality



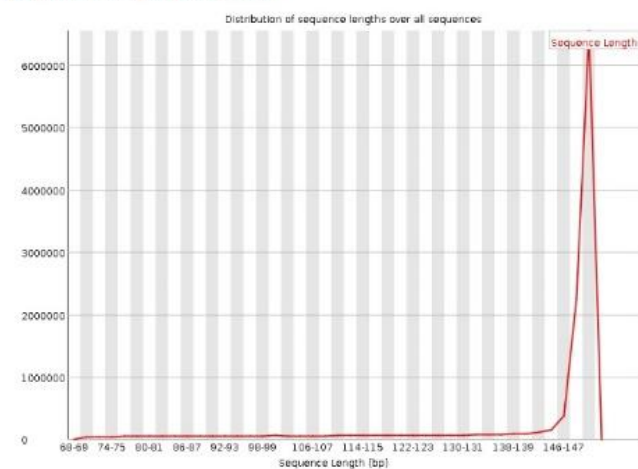
✔ Per sequence quality scores



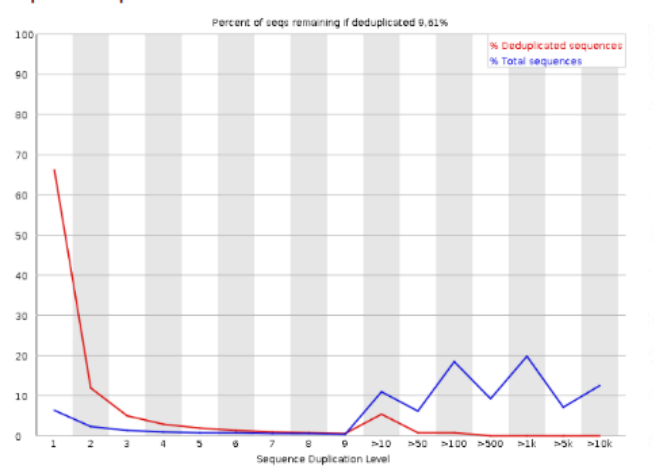
✘ Per base sequence content



📊 Sequence Length Distribution



✘ Sequence Duplication Levels



**Supplementary Figure E3.** Screenshot of Google Maps for political regions XV, I and II of Chile. Red circles show reported locations for cultivation areas of *E. coca*. Triangles show sampling locations above 4000 masl. Green: Talabre-Lejía transect. Purple: “Cueva del Chileno” at the Lípez highlands, Bolivia.



Following extended material is available upon request via email to [cm.andreani.g@gmail.com](mailto:cm.andreani.g@gmail.com):

**Supplementary Table E2.** antiSMASH versions 5 and 6, MiBIG repository and BiG-SCAPE nomenclature conversion: definitions of biosynthetic core types and classes.

**Supplementary Table E3.** Biosynthetic gene clusters database (n=190).

**Supplementary Table E4.** Frequencies of core biosynthetic types (n=23) detected per genome.

**Supplementary Table E5.** Specialized metabolite genes database (n=2857).

**Supplementary Table E6.** Frequencies of transport and regulatory sm-COGs (n=45) assigned to genes according to their belonging to BGCs per biosynthetic class.

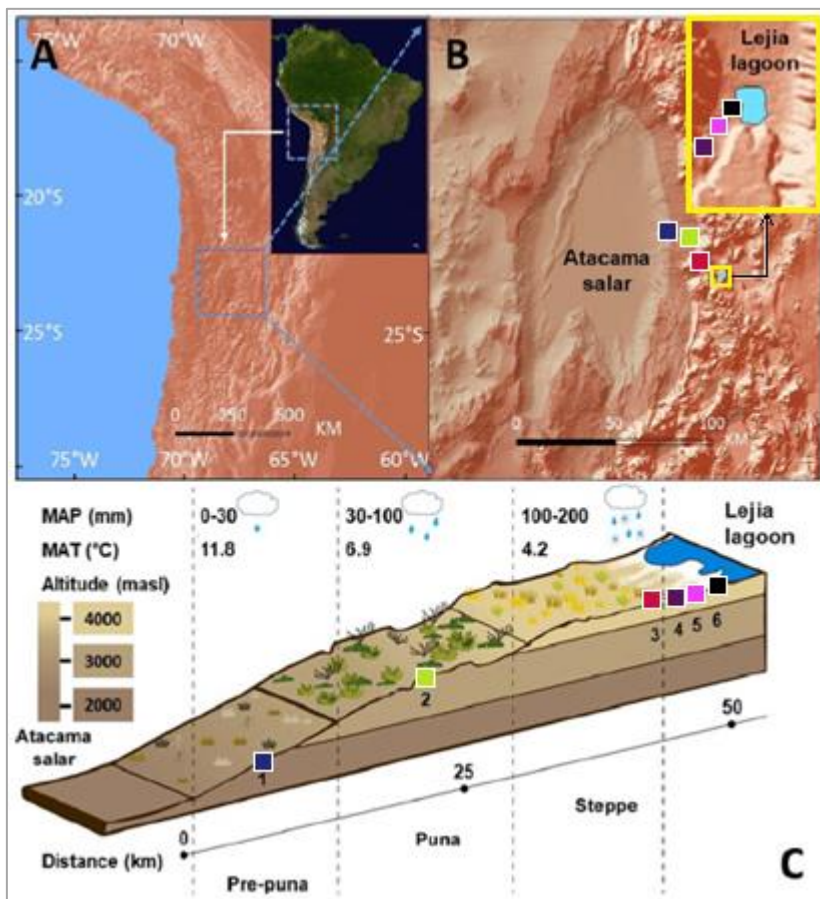
**Supplementary Table E7.** Gene cluster families database (n=178).

**Supplementary Table E8.** Conversion of PFAM domains (n=149) detected in BGCs belonging to a GCF of at least two members into GO terms.

**Supplementary Table E9.** PFAM domains found in reference BGC of each GCF database.

## Supplementary information

**Supplementary Figure S1.** Talabre-Lejía transect geographical information (adapted from González, M. *et al.*; unpublished).

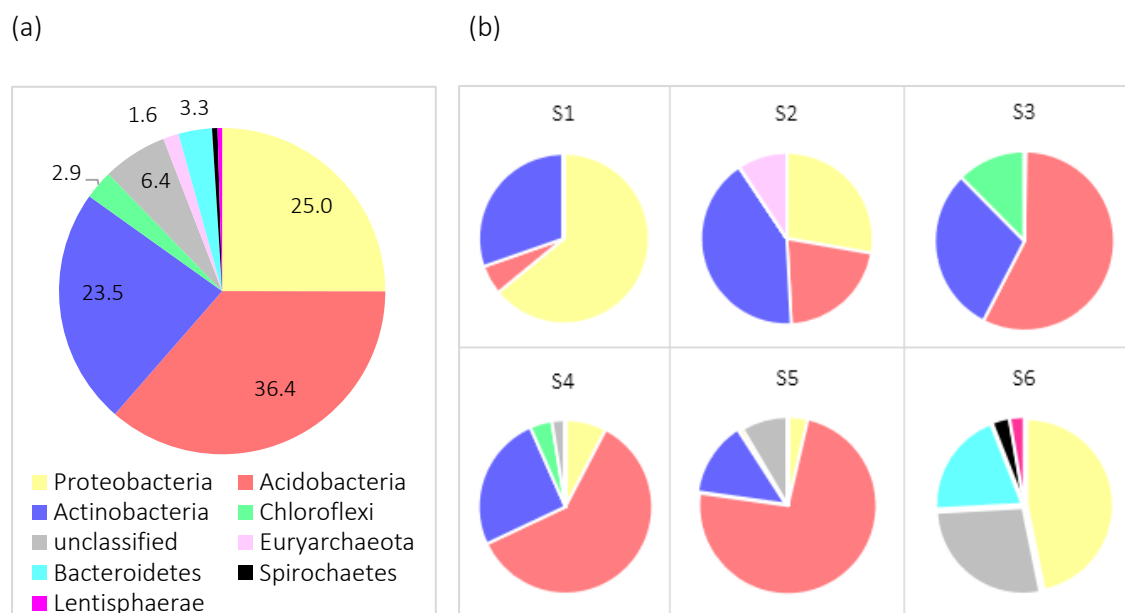


**Supplementary Table S1.** Talabre-Lejía transect soil physicochemical information (González, M. *et al.*; unpublished).

	S1	S2	S3	S4	S5	S6
altitude (masl)	2,870	3,870	4,480	4,480	4,480	4,314
TLT sites (Díaz et al., 2016)	TLT18	TLT08	TLT01	TLT01	TLT01	Lejía Lagoon
vegetation belt	pre-puna	puna	estepa	estepa	estepa	estepa
physicochemical data						
MAT (°C)	11.8	6.9	4.2	4.2	4.2	8.5
MAP (mm/año)	15.0	75.1	161.9	161.9	161.9	161.9
pH	8.1 ± 0.3	6.1 ± 0.15	5.82 ± 0.09	7.54 ± 0.12	8.485 ± 0.005	8.52 ± 0.02
electric conductivity (mS/cm)	0.06 ± 0.01	0.1 ± 0.02	0.05 ± 0.005	0.06 ± 0	0.13 ± 0	1.94 ± 0.15

**Supplementary Table S2** (adapted from González, M. *et al.*; unpublished) is available upon request via email to [cm.andreani.g@gmail.com](mailto:cm.andreani.g@gmail.com).

**Supplementary Figure S2.** Overall sum of (a) average and (b) site-specific relative abundances (%) of studied genomes (n=53) by phylum. Constructed upon total raw reads per genome as informed at Supplementary Table S2.



**Supplementary Table S3.** Metagenomic sequencing information (González, M. *et al.*; unpublished).

sample	filtered reads	data (Gb)	average length (bp)
S1	162,035,390	23.2	143.4
S2	166,329,234	24.0	144.2
S3	153,290,713	21.0	137.0
S4	141,107,489	19.5	138.2
S5	127,793,777	17.7	138.3
S6	81,854,686	11.9	145.7
summary	832,411,289	117.3	141.1

**Supplementary Table S4.** Percentages of annotated genes from each metagenome assigned to domains of life (González, M. *et al.*; unpublished).

sample	bacteria	archaea	eukaryota	viruses
S1	98.6	1.4	0	0
S2	97.8	2.2	0	0
S3	99.3	0.7	0	0
S4	99.7	0.3	0	0
S5	99.2	0.8	0	0
S6	99.7	0.1	0.1	0.1
summary	99.05	0.92	0.02	0.02

**Supplementary Table S5.** Metagenomes assembly information (González, M. *et al.*; unpublished).

metagenomes	S1	S2	S3	S4	S5	S3 + S4	S4 + S5	S6
total length (Mb)	460	655	823	830	678	1,542	1,685	592
number of scaffolds	578,420	727,989	893,733	892,653	714,882	1,656,500	1,832,894	640,971
average length (pb)	795	900	920	929	948	931	919	924
max length (pb)	231,815	174,227	306,936	459,495	1,145,089	1,145,083	567,355	368,304
N50 (pb)	851	988	1,283	1,334	1,384	1,343	1,290	1,426
N50 (%)*	75%	80%	87%	88%	90%	89%	87%	1
L50	436,427	579,801	776,204	788,942	640,269	ND	ND	573,308
% of used reads	17.8%	21.0%	38.3%	33.2%	32.2%	34.5%	36.6%	48.3%
% GC	67.0%	65.4%	67.3%	66.6%	64.7%	65.7%	66.9%	53.8%

**Supplementary Table S6.** Total RNA sequencing information (González, M. *et al.*; unpublished).

site	triplicate	sample ID	soil samples (ng/uL)	RNA Qubit (ng/uL)	dscDNA Qubit (ng/uL)	final library (ng/μL)	library size (bp)
S1	i	4	9.02	4.2	608.0	3.40	1011
S1	ii	5	6.94	too low	784.0	6.00	1035
S1	iii	6	6.86	too low	1448.0	2.46	650
S5	i	7	24.8	19.6	416.0	4.98	984
S5	ii	8	7.62	5.9	388.0	4.14	1096
S5	iii	9	27.2	22.4	700.0	4.14	1010