Research Networking Programmes

Short Visit Grant ☒ or Exchange Visit Grant ☐

(please tick the relevant box)

Scientific Report

The scientific report (WORD or PDF file – maximum of eight A4 pages) should be submitted online within one month of the event. It will be published on the ESF website.

Proposal Title: Metagenomic analysis of the seagrass microbiome

Application Reference N°: 6349

1) Purpose of the visit

Seagrasses are important ecosystem engineers in coastal areas, distributed worldwide. Unlike the extensive research that has been performed on terrestrial plants, the microbial consortia in association with these marine angiosperms is so far poorly studied. A diverse cocktail of bacteria live and interact in the rhizosphere of these plants, thriving on root exudates and using the large array of electron acceptors present in marine sediments for microbial respiration. These bacteria can be beneficial for the plant by fixing nitrogen or suppressing pathogens, but can also be harmful, such as the group of sulfate-reducing bacteria, which produce the phytotoxic gas sulfide. The toxicity generated by the production of sulfide is biologically counteracted mainly by sulfur-oxidizing bacteria, another group of bacteria strongly involved in the sulfur cycle in marine sediments. The aim of this visit was to perform a comparative analysis of four metagenomes of the rhizosphere of the seagrass Zostera marina and surrounding bulk sediment sampled in Portugal and France, looking in particular at genes involved in the sulfur cycle.

2) Description of the work carried out during the visit

The whole period of this visit was dedicated to the analysis of the metagenomes using several bioinformatic tools, described below.

Both paired-ends of all 4 datasets were initially trimmed using trimfastq.py, with a probability value of 0.05, above which base-calling error was considered too high. Fastq files were converted to fasta using convertseq and paired-ends were immediately combined and intercolated with the program velvet-shuffleSequences_fasta.py.
(Whenever settings are not detailed, default values were used.)

****GC%****

The percentage of GC (GC%) of each metagenome was computed on 1 million randomly selected sequences using gcseq.

****16S rRNA****

The phylogenetic assignment of 16S rRNA reads was also performed. A subsample of 10 million reads was computed on rnabait in order to identify potential 16S rRNA sequences, which were further aligned in parallel-ssu-align. All 16S rRNA reads assigned to prokaryotes were then blasted against the greengenes database, and reads smaller than 90 nucleotides were manually filtered and removed from the tables, as well as hits classified with less than 80% identity to the query sequence.

****Assembly of metagenomes****

The reads from all 4 metagenomes were assembled in contigs using IDBA_UD, an iterative de Bruijn Graph Assembler. A minimum k value of 70 and maximum 100 was used, with increments of 10 k-mers on each iteration. A pre-correction flag was used to correct reads before assimbling, based on their alignment with the contigs.

****Annotation****

Annotation was performed using Prodigal, a gene prediction software considered to be tolerant to minor errors originated during the assembly. Inicially, annotation was computed on contigs bigger than 1kb, however due to the small size and small number of contigs, this process ended up being performed on all the contigs assembled.

At this point of the analysis, all tables resulting from the annotation were visually analysed, and strong differences between the 4 datasets were detected. These differences stressed the need to follow a different approach for each metagenome, described below.

****Zm Portugal****

Phylogenetic analysis of genes involved in the sulfur cycle was performed. Protein sequences annotated as dsrAB, sox, aprAB, sqr, ATP sulfurylase, hdr and fcc (key proteins in the sulfur cycle) were manually selected, and proteins bigger than 250 bp were added to a file containing published protein sequences from the same families. The file containing sequences retrieved from the literature and the ones collected in this study was used to build a phylogenetic tree in makefasttree with a replication of 100 bootstraps.

****Bulk sediment France****

Contigs annotated as Vibrio splendidus and Eudoraea adriatica which were larger than 5kb, had a query coverage and similarity above 80 and 90%, respectivelly, were selected. The reference genome of V. splendidus LGP32 was added to this list of contigs, and tetranucleotide frequency was performed using the program compseq from the packag e EMBOSS. A Principal Component Analysis (PCA) was then built using cca-compseq-combiner.pl.

In order to investigate if the contigs associated with V. splendidus indeed belong to that species, Average Nucleotide Identity (ANI) was calculated using the abovementioned reference genome of V. splendidus LGP32 in the program ani.

3) Description of the main results obtained

****GC plots****

Microbial composition of the 4 metagenomes in terms of GC content was found to be different between sites. Both Z. marina rhizosphere and bulk sediment from Portugal accounted only with one peak of GC content between 60 and 65%, while both samples from France had two peaks clearly separated, one at approximately 45% and another one at 55% (Figure 1).

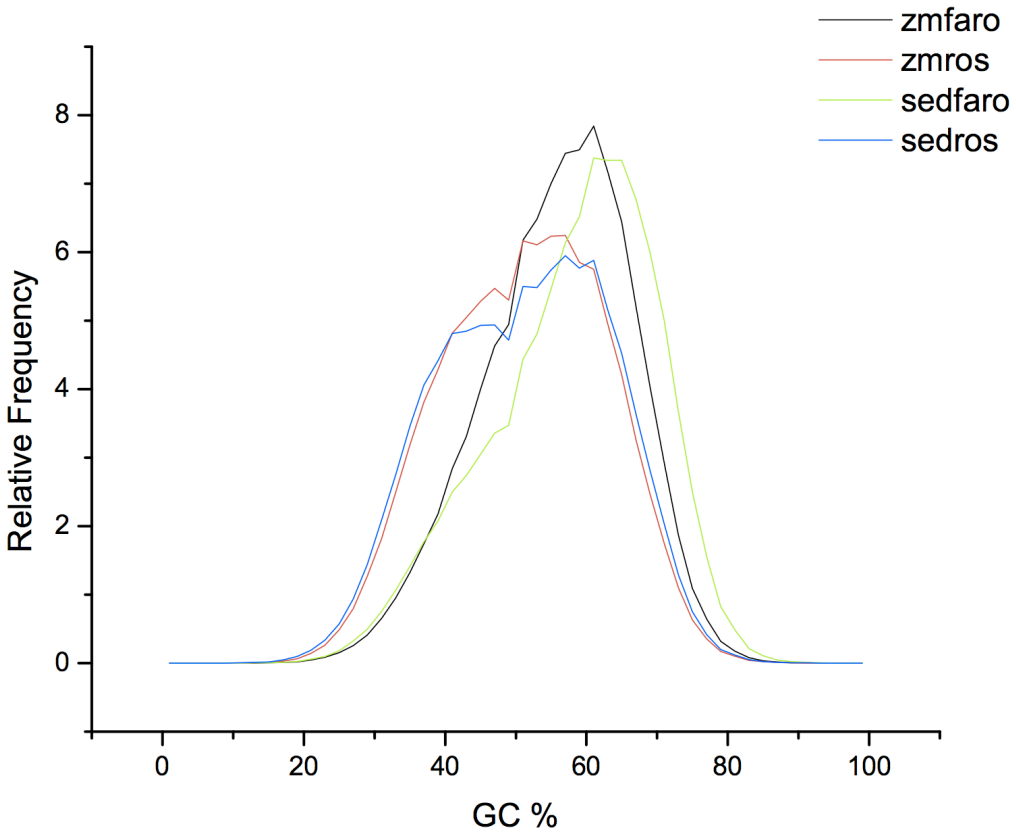
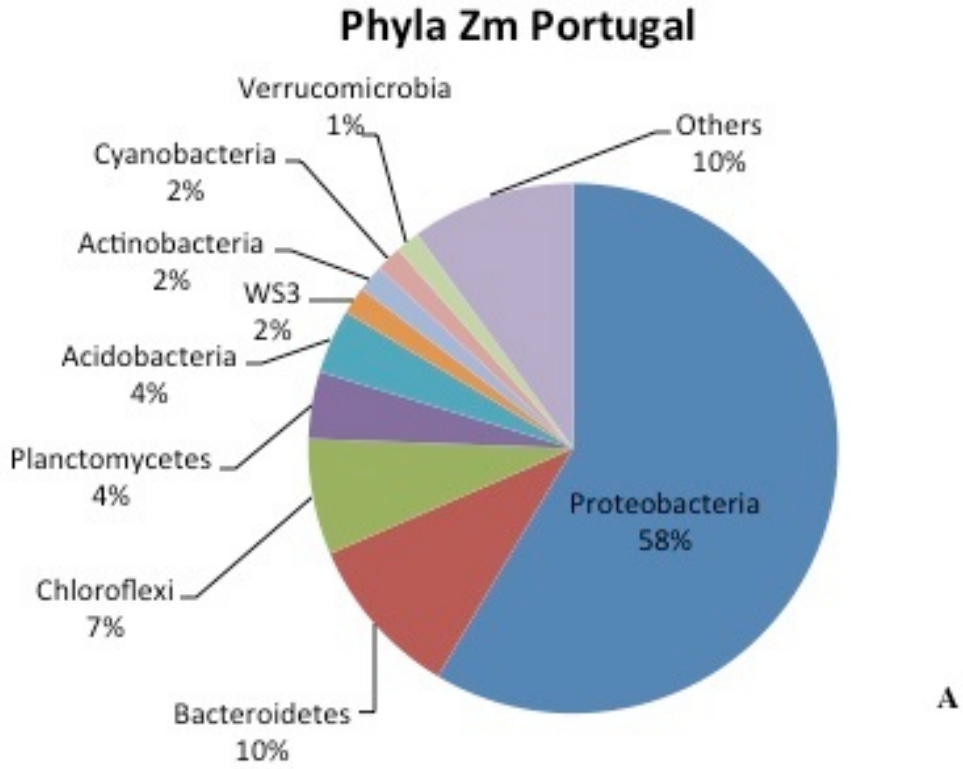


Figure 1. GC content profiles of the the 4 metagenomes. Rhizosphere from Z. marina and bulk sediment sampled in Portugal (zmfaro, black; sedfaro, light green), and in France (zmros, red; sedros, blue).

****16S rRNA****

So far, RNA analysis was only completed for the metagenome of the rhizosphere of Portugal. The analysis with rnabait and the greengenes database revealed that the rhizosphere of Z. marina is dominated by Proteobacteria (58%), followed by Bacteroidetes (10%) and Chloroflexi (7%, Figure 2a).

The phylum Proteobacteria was mainly represented by Gammaproteobacteria (31%) and Deltaproteobacteria (24%), which were also the overall most abundant classes in this metagenome (Figure 2b). The former was highly dominated by members of the order Chromatiales and to a lesser extent by Thiotrichales (57 and 15%, respectively). The order Desulfobacterales accounted for the highest abundance of Deltaproteobacteria, followed by Myxococcales (78 and 6%, respectivelly).



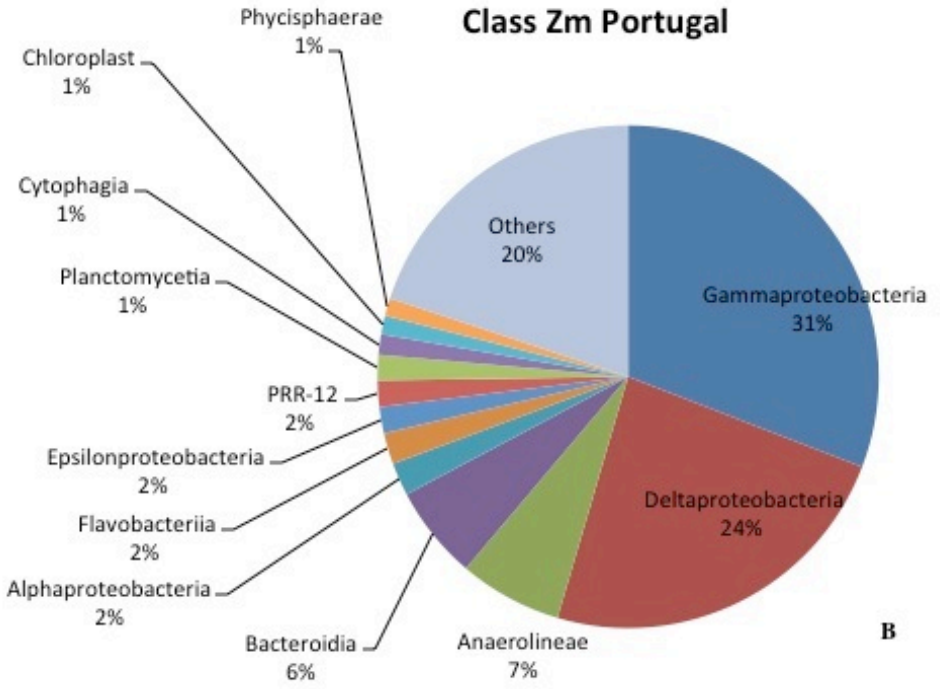


Figure 2. Classification of 16S rRNA sequences selected from 10 million random raw reads of the rhizosphere of Zostera marina sampled in Portugal. A total of 1835 reads with a query length above

90 nucleotides and identity above 80% were analysed and blasted against the greengenes database.

Two families were identified within the order Desulfobacterales, Desulfobacteraceae (75%) and Desulfobulbaceae (25%). Regarding the most abundant order of Gammaproteobacteria, Chromatiales, 64% of the reads assigned to this order were unclassified, 33% belonged to the family Chromatiaceae, 2% to Ectothiorhodospiraceae and 0.3% to Halothiobacillaceae.

****Assembly of metagenomes****

A summary of the assembly of the four metagenomes can be found in table 1.

Table 1. Summary of assembled contigs using IDBA_UD. All metagenomes are described, with emphasis on contigs larger than 1kb, and on all the contigs built.

	Rhizosphere Portugal		Sediment Portugal	
	All	**> 1kb**	**All**	**> 1kb**
# sequences	24 912	2 685	3 464	490
# bases	16 669 913	4 281 816	2 550 647	961 718
Smallest	221	1 000	207	1001
Largest	10 119	10 119	13 879	13 879
AVG length	669	1 594	763	1 962

	Rhizosphere France		Sediment France	
	All	**> 1kb**	**All**	**> 1kb**
# sequences	9 552	1 102	21 504	5 440
# bases	6 533 981	2 003 543	21 022 162	11 568 122
Smallest	205	1 000	200	1000
Largest	54 757	54 757	30 987	30 987
AVG length	684	1 818	997	2 126

****Annotation****

So far it was only possible to evaluate the phylogeny of dsr proteins. A total of 72 contigs with an average length of 795.5 basepairs were selected and a crude phylogeny was built (Figure 3).

Figure 3. Phylogeny of dsr proteins from the rhizosphere of Z. marina from Faro. Only 6 proteins were selected due to the cutoff of 250 bp. The green cluster correspond to the reverse dsrB proteins and the blue to the reverse dsrA, both associated to sulfur oxidation. On the top of the tree, black sequences are relative to dsrA proteins, which are coupled with sulfate reduction.

****Bulk sediment France****

The PCA resulting from the tetranucleotide frequency analysis resulted in two clear clusters, in which V. splendidus and E. adriatica were clearly separated (Figure 4).

Average Nucleotide Identity revealed that the contigs in close proximity with the reference genome of V. splendidus belong to a different species (ANI<0.95).

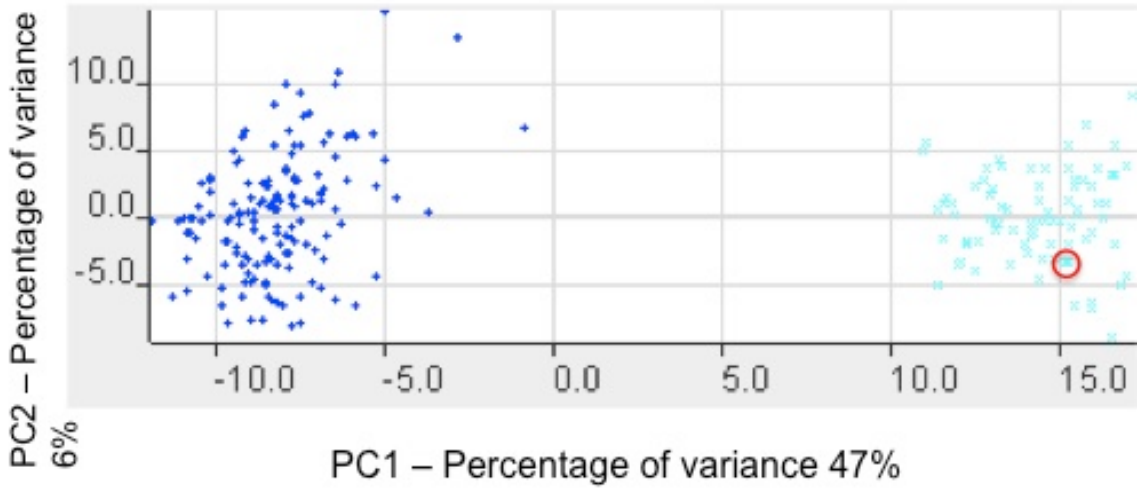


Figure 4. Principal component analysis of the tetranucleotide frequency between contigs larger than 5kb annotated as E. adriatica (dark blue) and V. splendidus (light blue). Enclosed in red is the reference genome of V. splendidus LGP32.

The analysis of metagenomic datasets is considered to be the bottleneck of this type of approach, especially in environments for which little is known. Besides, highly diverse environments are also difficult to study due to their complexity.

4) Future collaboration with host institution (if applicable)

I will finish the metagenome analysis in Amsterdam, and once everything is finished, I will publish the results with Prof. Dr. Rodríguez-Valera.

5) **Projected publications / articles resulting or to result from the grant *(ESF must be acknowledged in publications resulting from the grantee's work in relation with the grant)***

Publication in a peer-reviewed journal about the diversity of sulfur genes present in the rhizosphere of Z. marina sampled in Portugal.

6) Other comments (if any)