Bachelor’s Degree in Bioinformatics (UPF-UPC-UB)

Final Degree Project

Strain detection and genetic diversity in Mediterranean marine microbial populations

Sergio Rodríguez Llana[1]

Scientific director: Ramiro Logares Haurié[2]

[1]ESCI-UPF, Passeig de Pujades, 1, 08003 Barcelona. [2]Ecology of Marine Microbes, Institut de Ciències del Mar (ICM-CSIC), Passeig Marítim de la Barceloneta, 37, 08003 Barcelona.

Abstract

Motivation: Marine microbial populations play a key role in the ecology of the seas. The adaptive potential of these species is key to understanding the evolution of their populations across time and space. This means to comprehend and predict how they may react to certain environmental conditions, biotic and abiotic factors, and how their population size may change as a consequence. Here, we investigate the strain formation within some abundant population genomes exploring genetic diversity, looking into genetic variations present mainly in hot water seasons, and observing which are the primary protein functions that are being affected.
Results: Our strain detection analysis shows the presence of different clusters of the same species that in some cases seem to group by water temperature. We observed how one MAG of *Pelagibacter (SAR11)* contained more genetic diversity than MAGs of *Acidimicrobiales, Flavobacteriaceae* and *Marinisomatia* (*SAR406 clade*). We detected genetic variations (mainly SNPs) affecting proteins related to different functional categories, standing out genetic information processing and carbohydrate metabolism, that were shared and abundant in the 4 analysed MAGs. We also performed a variant annotation useful for protein adaptation analysis.
Supplementary information: Supplementary code scripts are available at GitHub: https://github.com/SergioRLL/FDP.git

1. Introduction

Marine microbiomes play an important role in the biochemical processes that take place in such environments. The understanding of these events is fundamental to preserve the correct functioning of the planet and may allow us to predict ecological responses driven by changes in environmental conditions. For this purpose, there is a growing interest in omics studies that aim to reveal the functional and taxonomic diversity of ocean microorganisms. Nevertheless, the diversity of microorganisms at the population level (variation that is found in the genome of the same species) has not yet been deeply studied. Obtaining more knowledge in this field will give us important insight into the evolutionary capacities of marine organisms and their adaptive potential to different ecological niches, which is indeed crucial to understand the ecosystem function and community reactions.

Ecology is the branch of biology that seeks to explain life processes, interactions and adaptations, as well as the abundance and distribution of organisms and biodiversity in the context of the environment. Adaptations appear in organisms due to changes in the environment. Environmental factors can be classified into two main groups: biotic factors and abiotic factors, both of them make up the ecosystem. In biology, abiotic components include non-living chemical and physical factors, while biotic factors include living components of the ecosystem: the organisms that live in it.

It is without a doubt that oceans are huge biological environments and they constitute very important ecosystems and they play a very important role and contribution to the marine and the global biosphere. For this reason, the information gained in the microbial diversity field is key to understanding and anticipating changes in these environments, not only for species diversity but also for the ecological niche as the role these species represent in the ecosystem. Ecological niches describe how populations of organisms are adapted to environmental conditions (for instance, resource availability and competitors, temperature, and salinity levels). How marine microbial communities are affected by these ecological variables may be determined by populations' genetic diversity. For example, there might be strains of a certain population that have developed resistance to low or high temperatures while another does not, and therefore the first population may grow at a higher rate than the second one affecting the ecosystem diversity and niches.

To study genetic diversity in microbial populations early approaches involved isolating multiple cells from the same population and then performing genome sequencing and/or phenotypic analysis. Recent approaches have been using metagenome analysis, which involves extracting and sequencing DNA directly from the environment. Then, the resulting DNA is assembled and binned into genomes *in silico*. This technique allows for the analysis of all taxa in microbial populations, including the identification of genetic variants and their frequencies inside the population. Therefore, using metagenomic tools we have the potential to investigate the evolutionary dynamics that lead to strain selection.

Among the most frequent questions that arise when investigating microbial population adaptation, we wonder which are the most important environmental factors selecting for strains. It is known that strong candidates for this are temperature, salinity concentration, and nutrient abundance due to geographical location and the depth in which the organisms live in the water. In addition, microbial interaction also plays an important role as a biotic factor. It is also common knowledge that the Mediterranean Sea is a highly diverse marine ecosystem that hosts 7-10% of the world's marine biodiversity and it has been defined as "under siege" due to the pressure of multiple human activities [1]. Therefore, it will be necessary to understand how these human activities, and therefore climate change, impact the biodiversity of the Mediterranean. A first approach may be to see how strains of microbial populations behave concerning water temperature and how these populations react to seasonal changes. Furthermore, this analysis would be interesting to compare with data from the global ocean as the two environments present different features.

In order to analyze population genomic parameters, researchers have been developing pipelines that deal with intra-population genetic diversity to cluster populations of MAGs (Metagenome-Assembled Genomes) according to ANI values (Average Nucleotide Identity) using hierarchical clustering.

It is a hard task to evaluate in bacteria which are the potential traits of an organism that could undergo a better survival in unfavorable conditions. A first approach may be to observe which are the most relevant genetic variations and see the functions that they might have affected.

Objective

This scientific project is focused on investigating the genomic diversity of microbial populations using two time series of 7 years from the coast of the Mediterranean where samples are obtained each month. These genomes were obtained from two marine observatories: BBMO (Blanes Bay Microbial Observatory) and SOLA (from Bay of Banyuls-sur-mer). To carry out this task, we performed an analysis of SNPs (Single Nucleotide Polymorphisms) and MNPs (Multiple Nucleotide Polymorphisms) to explore the genetic variability of the strains inside the populations.

We analysed the genomic diversity of 4 population genomes, analyzing the possibility of the presence of different strains for each of them. In that case, we aimed to focus on the differences that define those strains and try to explore the phenotypic changes caused by the corresponding genetic variations.

2. Material and Methods

2.1 Sample generation

We used 84 samples collected from BBMO (Blanes Bay Microbial Observatory) and 90 from SOLA (from Bay of Banyuls-sur-mer). Samples were collected once a month for a period of 7 years in both marine observatories, however, in the case of SOLA some samples were extracted during different days of the same month the same year, thus, there are 6 more samples collected from SOLA. From these samples, MAGs were assembled and used as reference genomes for this project. First, 3 year monthly samples (Jan 2011 - Dec 2013) from BBMO were assembled all together into one co-assembly using the Ray assembler [2]. Then, 34 high-quality MAGs were obtained from the co-assembly by binning with the MetaWrap pipeline [3]. From them, we selected 4 abundant genomes to perform the analyses and test this methodology.

2.2 Read mapping, gene prediction and detection of variants

We firstly generated and prepared the required data, containing the information regarding sequence mapping to the genome assemblies, gene prediction, and detection of variants.

We used *Prodigal* [4] for gene prediction for our 4 assembled genomes, getting the sequences of the genes as well as gene features contained in GFF files.

Using *Bowtie 2* [5] we mapped the forward and reverse reads (R1 & R2) obtained from the metagenomic samples with the purpose of obtaining the alignments in SAM format files. Afterwards, We used *SAMTOOLS* [6] to convert the SAM files to BAM files (the compressed binary version of the alignments) and index them. We concatenated the 4 MAGs putting them together into a single FASTA file in order to be interpreted by *inStrain* in the next step along with an STB file, a text file that indicates to which MAG each scaffold belongs.

Subsequently, we used *Freebayes* [7] to generate files (.VCF) containing the variants observed in the alignment files for future detection of SNPs, MNPs, indels (insertions and deletions), and other types of genetic variations.

2.3 inStrain pipeline

In this project, we have been testing different pipelines that can detect the presence of several strains within MAGs. Here we present the results obtained from *InStrain* [8]. In particular, *InStrain* uses metagenomic paired reads to profile intra-population genetic diversity across whole genomes and compares microbial populations in a microdiversity-aware manner.

First, we used the *inStrain profile* module to generate objects called “IS_profile” (see Figure 1). This module is described to be as the heart of the pipeline as its output provides information regarding locations of divergent sites, that is to say, positions in the genome where either an SNV (Single Nucleotide Variant) or SNS (Single Nucleotide Substitution) is present, the number of read-pairs that passed the program filters for each scaffold, the linkage between SNV pairs, etc. through a series of tables and figures. The module uses as input the previously created FASTA file containing the genomes and the sample alignments (BAM files). The program takes all the alignments for every sample and keeps the ones mapping to the genomes with sufficient quality while the others are filtered out. Then, using only the filtered reads, microdiversity metrics are computed on a scaffold-to-scaffold basis, including the coverage and nucleotide diversity at each position along the scaffold, and identification of SNVs and SNSs. It also calculates the linkage between divergent sites on the same read pair, that is to say, a measure of how likely two divergent sites are inherited together on the same read pair. Additionally, it computes scaffold-level properties like the overall coverage, breadth of coverage, ANI between the reads and the reference genome, and the expected breadth of coverage based on that true coverage.

Afterwards, we used *InStrain compare*, another *InStrain* module to compare multiple inStrain profiles making pairwise comparisons between them. The module compares genomes that have been profiled by multiple different metagenomic mappings. Here we executed *inStrain compare* using the previously generated profiles and default parameters which forced the program to only compare genomes that are sufficiently detected and skip clusters that have near 0% overlap. As the output for this process, different tables are produced summarizing the differences between the inStrain profiles at genome and scaffold levels, including coverage overlap, compared base count, and information about SNPs used to compute ANI (Average Nucleotide Identity) among compared bases between two scaffolds. In addition, and most importantly for our case, the result of clustering the pairwise comparison data is used to generate strain-level clusters using hierarchical clustering. Moreover, the pipeline also produces figures of the dendrograms clustering the samples according to ANI and shared genome coverage.

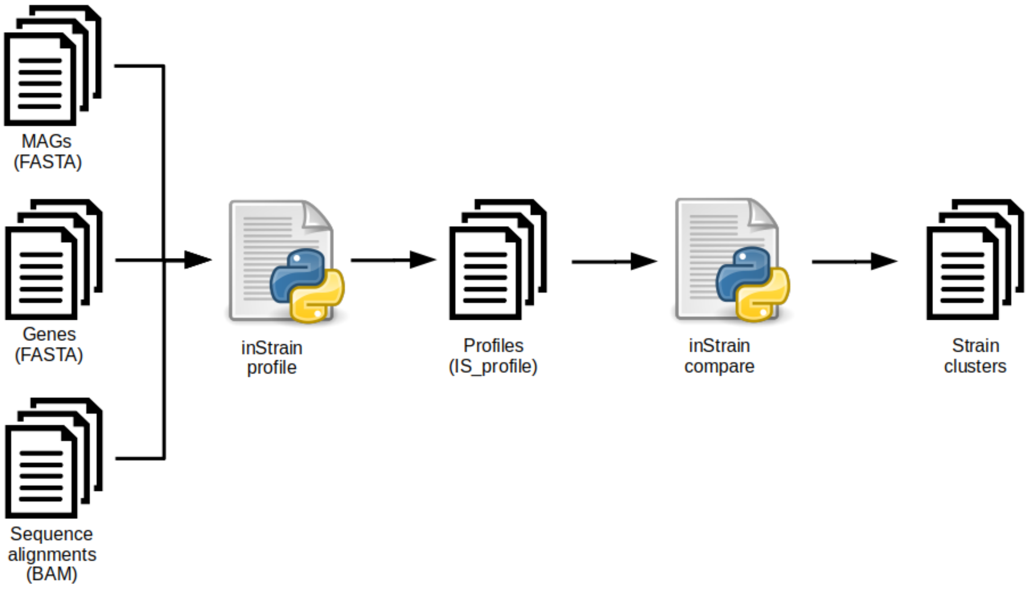


Figure 1. Overview of *inStrain* pipeline. A graphical representation of the main procedure that we followed, using the main *inStrain* modules.

2.4 Visualization of genetic variants

After observing the results obtained by *inStrain*, especially the strain clustering, we hypothesized that there could be a strain grouping by temperature, as it is usually the most decisive environmental factor in microbial life [9]. Therefore, we decided to classify the samples by temperature and season and to try to determine potential genetic variations that may be causing this strain differentiation. Thus, these variations might cause significant phenotypic variations that undergo strain selection.

For this task, we used the *Integrative Genomics Viewer* (*IGV*) [10], a high-performance interactive tool that allows the visualization of different types of genomic data. We observed the 4 MAGs separately with their respective predicted genes and variants, showing some variants that were present only during hot water seasons: summer and autumn. Due to the low abundance of 3 out of these 4 bins (consensus genome corresponding to some operational taxonomical unit) during these seasons, we thought that genetic variations occurring mainly or exclusively in these seasons may be significant for explaining the strain differentiation, allowing the species to live in hot environments. For this reason, we decided to make a list of potential significant variations, including SNPs, MNPs and indels, with their respective predicted genes.

2.5 Gene functional analysis and variation effect

Finally, we performed a protein functional analysis in which we used BLASTX from NCBI [11] and KEGG's BlastKOALA [12,13] to search for homologous proteins and to seek after functions and metabolic pathways in the genes we selected due to the presence of potentially relevant DNA variations occurring exclusively or mostly in genomes from hot water samples. We used EMBOSS Transeq [14] to translate genes' nucleic acid sequences to the corresponding amino acid sequences.

To further estimate the effect of the genetic variations, mainly SNPs and MNPs, on the transcripts we ran the SnpEff pipeline [15]. This program provides tools for genetic variant annotations and predicts the effects of genetic variants on genes and proteins such as amino acid changes. It is also interesting because the tool performs an impact prediction based on the variation effect type which allows to categorize and prioritize variants. In our case, the MAGs assembly was at the contig level, we did not have chromosome information and therefore we had to create a new SnpEff species database for every MAG.

In this way, we were able to determine the functions of the genes and also the effect that the detected variants were causing on them.

Supplementary code

Supplementary code scripts are available at GitHub: https://github.com/SergioRLL/FDP.git

3. Results and Discussion

3.1 Strain detection and clustering

With the *inStrain* pipeline, we were able to cluster the different metagenomic sequence reads according to the variants detected for each of the 4 MAGs. The strain definition and differentiation was the result of clustering the pairwise comparison data from the metagenomics samples performed by *inStrain compare*. From the output of *inStrain compare*, we were able to see the different number of predicted strains for every MAG. In Figure 2 we can observe the *Pelagibacter* MAG dendrogram comparing all samples based on population ANI and shared bases. The dendrograms for the other 3 MAGs can be found in Supplementary Figure 1. In depth, we found that the hierarchical clustering algorithm determined 68 strains (clusters) for *Pelagibacter* (bin208), 18 for *Marinisomatia* (*SAR406 clade*) (bin228), and only 3 for *Flavobacteriaceae* (bin216) and *Acidimicrobiales* (bin120). A table showing which metagenomic sample belongs to each strain for every MAG can be found in Supplementary Table 1. These results can tell us about the genetic diversity that is found within populations of these microbial organisms. Among these 4 species, we can state that *Pelagibacter* (*SAR11*) has more divergent sites than the rest of the other species populations. Thus, *Pelagibacter* species contain more genetic diversity and may have more adaptive potential than the others, however, we can relate that with the fact that *Pelagibacter* has a higher abundance than the rest of the species.

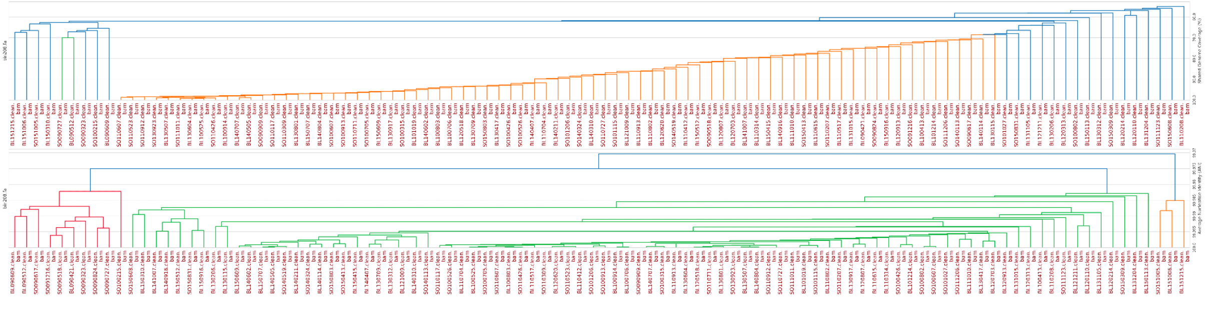


Figure 2. Dendrogram of clustered samples after mapping them to the selected *Pelagibacter* MAG. The figure shows the metagenomic samples mapped to the *Pelagibacter* MAG that passed the filters. Then, they were compared by Shared Genome Coverage (above) and Average Nucleotide Identity (below). We can see the main clusters in different colors.

In order to confirm this hypothesis, we can take a look at the abundance of the *Acidimicrobiales* and *Pelagibacter* MAGs among the samples from BBMO and see the number of reads that we find per sample. Observing Figure 3, we can state, by looking at the number of reads (RPKG) and the number of SNPs we found at each season, that the *Pelagibacter* population is abundant during the whole year while *Acidimicrobiales* is not. Moreover, we find that the latter species is found mostly in cold water seasons (winter) with low abundance the rest of the year, while the *Pelagibacter* MAG seems to be a bit more abundant in hot seasons (summer) although it is significantly present in all of them. This result is reinforced by the fact that *Pelagibacter* (*SAR11*) is the most abundant plankton species found in the ocean, therefore, we should expect that it has high genomic variability. Regarding the other 2 species, we find that *Marinisomatia* and *Flavobacteriaceae* are also species that seem to behave similarly to *Acidimicrobiales*, appearing almost exclusively during cold seasons. Graphs for *Marinisomatia* and *Flavobacteriaceae* are shown in Supplementary Figure 2.

Furthermore, by comparing the number of strains obtained using *inStrain* and *Anvi'o* [16] metagenomic pipelines, we think that perhaps they are similar enough to state that both approaches are reliable in terms of estimating the number of clustered metagenomic samples, and therefore, the number of strains. We have to take into account that this is a number that depends on the way the hierarchical clustering is done, and when determining the distance threshold that defines the number of strains.

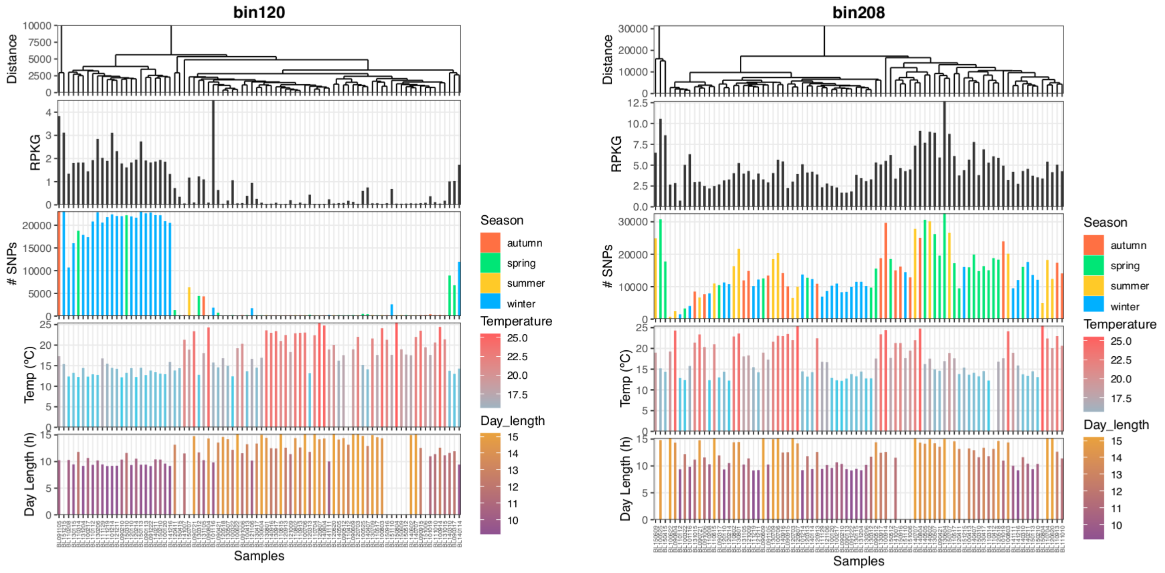


Figure 3. Dendrograms of BBMO samples based on SNP composition for *Acidimicrobiales* (bin120) and *Pelagibacter* (bin208) MAGs. Dendrogram showing the clustering of the samples based on the SNP composition of each MAG using Euclidean distance. Abundance data in RPKG (Reads Per Kilobase Million) and number of SNPs in each sample is also provided. Environmental factors including temperature and day length (number of sun hours) are also plotted for each sample. Figures made using R *ggplot2* package, data obtained from the Anvi'o software.

3.2 Detection of relevant genetic variations

Using *IGV's* ability to combine genome sequences, genes (using in our case GFF format) and DNA variants information, it was possible to perform a visualization of gene regions and observe where SNPs, MNPs, and indels were occurring. By inspecting those regions and the VCF files, we realized that the main types of detected genetic variations corresponded to mainly SNPs and MNPs, with few cases of insertions and deletions. Unfortunately, for this analysis, we only used BBMO samples because SOLA samples were not mapped at this point to the MAGs yet. We grouped the metagenomic samples according to their season and water temperature. While observing in *IGV Marinisomatia*, *Flavobacteriaceae*, and *Acidimicrobiales* MAGs we detected that the majority of the variations appeared mainly in samples collected during cold water seasons (winter and spring), therefore, strains living in cold water environments. We inferred that this fact made a lot of sense owing to the high abundance of these organisms during those seasons and the low abundance that they presented during the rest of the year.

With this idea in mind, we decided to make a selection of genetic variations present mainly or only during summer and autumn such as the ones shown in Figure 4 with the hypothesis that the effects of these variations in their corresponding genes, are important to understand the key gene or protein functions that undergo adaptation for these species in hot water environments. Some of these variants were also occurring in non-coding regions, we noticed them but we decided to discard them as it may be difficult to determine their possible effects without a precise annotation.

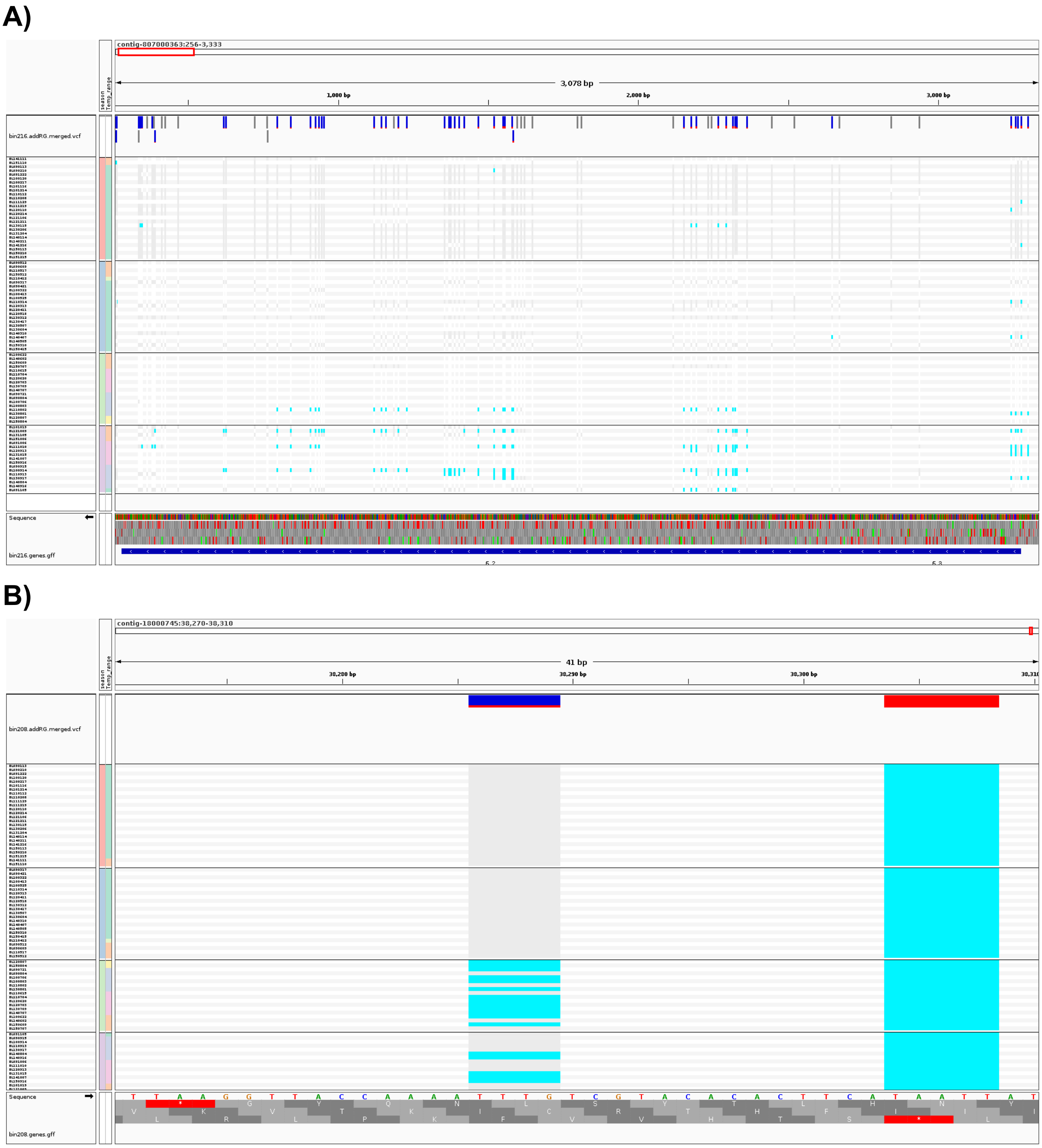


Figure 4. Genetic variants distribution visualization in IGV. Section of *Flavobacteriaceae* (bin216) (A) and *SAR11 Pelagibacter* (bin208) (B) MAGs. The upper region displays genomic coordinates (bp). Below it, we can see the variants track, where we can see their distribution across the genome, indicating the corresponding source sample. In light-blue we can observe the alternative nucleotides differing from the reference ones (e.g. SNPs or MNPs). The MAG reference sequence is shown below with colors representing nucleotides. Under it, we can find the corresponding codon translation with the 3 possible frames. At the bottom, we also see the predicted genes by Prodigal colored in dark-blue. The 2 colored columns next to the sample names on the left side show the added sample attributes. For seasons, red corresponds to winter, blue to spring, green to summer, and purple to autumn.

In total, we selected 43 genes for *Pelagibacter*, 82 for *Acidimicrobiales*, 42 for *Flavobacteriaceae*, and 45 for *Marinisomatia* having relevant genetic variations, those appearing only or mainly in hot water samples.

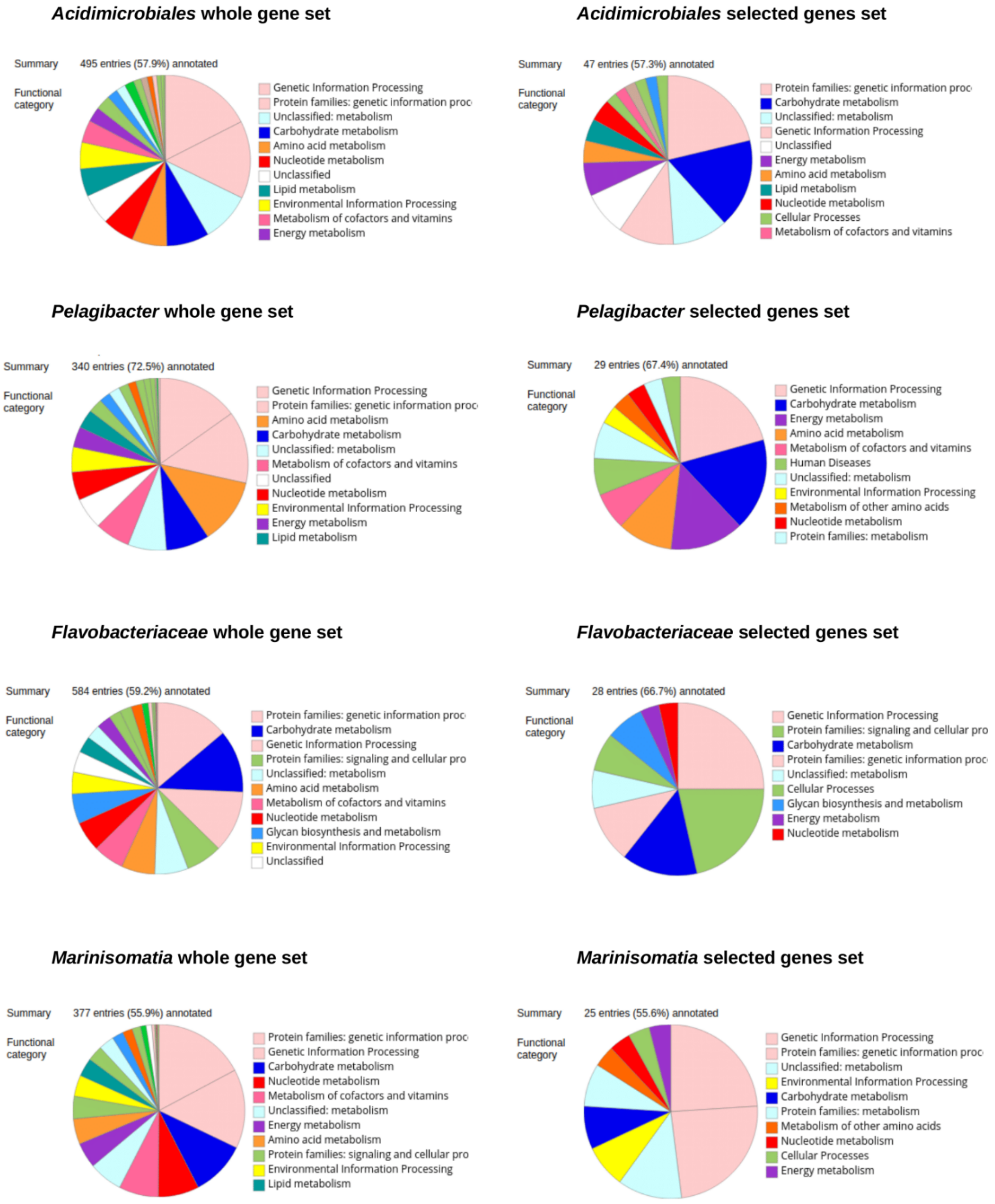
3.3 Functional analysis

After detecting these genetic variations, we performed a BLASTX search at the NCBI's non-redundant protein sequences (nr) database. We also tried to find hits in other databases such as UniProtKB/Swiss-Prot, however, as our data belong to a metagenomic project and bacterial genomes were not deeply investigated and annotated, there were few results and the percentage of identity was low. For some of the genes, we could only find hits for hypothetical proteins predicted in other metagenomic studies but they have not been deeply annotated.

On the other hand, we also used KEGG's BlastKOALA tool to make a search and find hits in KO (KEGG Orthology) database [17]. In this step, we did a search for both the selected gene sets (with those genes containing the relevant genetic variations) and the whole sets of predicted genes for the 4 MAGs.

As we expected in Figure 5, there is no highly predominant functional category in neither the whole gene sets nor the selected gene sets. Thus, we can confirm that there is not only one functional category that is key for adaptability and that there are many of them that are intervening with this process. Moreover, looking at the selected gene sets' pie charts, we see that genetic information processing categories (skin-colored sections including both genetic information processing and protein families: genetic information processing) and carbohydrate metabolism represent nearly half of the functions found in all of them. For this reason, we looked in-depth at these categories to search for similarities and differences between them.

Interestingly, the four MAGs had SNPs in some of the functional pathways within the selected gene sets defined, including some cases where the affected protein, the one with the relevant SNPs, is the same. For example, we found that preprotein translocase subunit SecA [EC:7.4.2.8], a protein involved in the protein export process in the bacterial secretion system interacting with the SecYEG preprotein conducting channel, had relevant SNPs in both *Flavobacteriaceae* and *Marinisomatia* genomes. An image showing protein export through the bacterial Sec pathway can be found in Supplementary Figure 3. SecA has a central role in coupling the hydrolysis of ATP to the transfer of proteins into and across the cell membrane, serving both as a receptor for the preprotein-SecB complex and as an ATP-driven molecular motor driving the stepwise translocation of polypeptide chains across the membrane [18].


***Acidimicrobiales* whole gene set**

***Acidimicrobiales* selected genes set**

***Pelagibacter* whole gene set**

***Pelagibacter* selected genes set**

***Flavobacteriaceae* whole gene set**

***Flavobacteriaceae* selected genes set**

***Marinisomatia* whole gene set**

***Marinisomatia* selected genes set**

Figure 5. Pie chart of KEGG's KO functional categories. On the left, we can observe the results obtained for the whole set of genes of each MAG and on the right for the selected genes containing

interesting genetic variants. On top of each pie chart, we can see the number of proteins annotated, *i.e.*, the ones having a significant KO match in the database.

Another case we found interesting was the modification of the RNA degradosome in both *Pelagibacter* and *Marinisomatia*. In this case, different proteins were affected: polyribonucleotide nucleotidyltransferase [EC:2.7.7.8] in *Pelagibacter*, and transcription termination factor Rho in *Marinisomatia*, both involved in different types of RNA degradosome. It is known that its presence has been maintained throughout the evolution of many bacterial species due most likely to its diverse contributions in global cellular regulation. It also has been experimentally demonstrated that the presence of RNA degradosome is a selective benefit for *E. coli* [19]. Thus, some variations in those species' degradosomes may be positively selected for their evolution. An image showing the bacterial RNA degradosome structure can be found in Supplementary Figure 4.

Other interesting functions that we noticed include genetic variants present in bacterial DNA-directed RNA polymerase β and β' subunits [EC:2.7.7.6] of *Acidimicrobiales*, modifications in glycolysis/gluconeogenesis in *Acidimicrobiales*, *Pelagibacter* and *Marinisomatia*, as well as other carbohydrate metabolism processes in some of them.

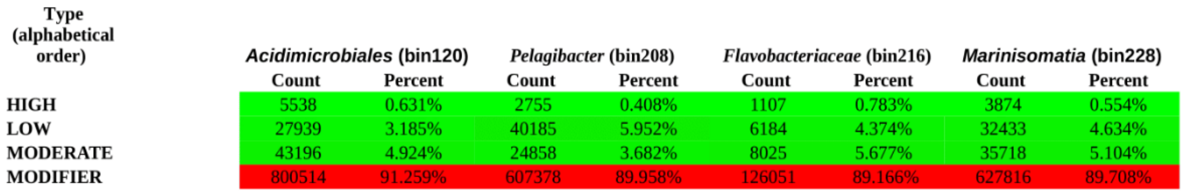
Unfortunately, arround 40% of the genes selected having interesting genetic variations did not get a significant hit in the KO database and therefore we were not able to see the functions related to them. However, it would be interesting for future studies to get more insight into them.

3.4 Analysis of genetic variations effect

Finally, after using SnpEff to see what type of effect the DNA variations were causing in the genes, we obtained files with every variant annotated with their corresponding effect prediction, such as missense variants or stop codon gained. In addition, we got information on every MAG and gene and the number of variants according to their impact. For instance, a "high" impact is assigned when the variant produces a STOP codon or a frameshift caused by insertions or deletions, and a "low" impact is assigned in the case of a synonymous variant.

After observing the results, included in Table 1, we noticed that *Flavobacteriaceae* has the largest number of high and moderate impact variants and that *Pelagibacter* has the largest number of low impact variants. Nevertheless, for all of the MAGs, close to 90% of the variants were assigned a "modifier" impact, these are usually non-coding variants or variants affecting non-coding genes, where predictions are difficult or there is no evidence of impact. Therefore, around 90% of the variants were defined as just upstream or downstream gene variants because it was difficult to predict their effect beyond that. However, we got useful information on the other 10% of the variants for which an impact was predicted and thus, some of these annotated variants may be interesting to analyse in depth, relating them to structural and functional protein modifications that may play a role in those species adaptability.

Files showing resulting summary statistics can be found in the Github repository. They contain information regarding the number of variants by type, the number of effects by functional class, type and region, base and codon changes, etc.



Type (alphabetical order)	*Acidimicrobiales* (bin120)		*Pelagibacter* (bin208)		*Flavobacteriaceae* (bin216)		*Marinisomatia* (bin228)	
	Count	Percent	Count	Percent	Count	Percent	Count	Percent
HIGH	5538	0.631%	2755	0.408%	1107	0.783%	3874	0.554%
LOW	27939	3.185%	40185	5.952%	6184	4.374%	32433	4.634%
MODERATE	43196	4.924%	24858	3.682%	8025	5.677%	35718	5.104%
MODIFIER	800514	91.259%	607378	89.958%	126051	89.166%	627816	89.708%

Table 1. Number of effects by impact type.

4. Conclusions

In this project, we obtained more insights on genetic variability regarding Mediterranean marine bacterial species. So far, we have explored the genetic diversity within 4 population genomes that are abundant in samples extracted from Blanes Bay Microbial Observatory and SOLA during a period of 7 years. We detected the presence of several strains for every population genome tested in *inStrain*, a recently developed pipeline, and we discovered that *Pelagibacter* (SAR11) has higher genetic variability than *Acidimicrobiales*, *Flavobacteriaceae,* and *Marinisomatia* species. We saw how these detected strains may cluster according to water temperature and that, with the exception of *Pelagibacter*, the species were a lot less abundant during hot water seasons: summer and autumn.

Using a visualization tool, *IGV* (*Integrative Genomics Viewer*), we detected relevant genetic variations (SNPs and MNPs) occurring only or mainly in hot water seasons. We selected a set of predicted genes that presented those kinds of variations and we searched for their identity and functionality in different protein databases. From the results of KEGG's KO significant hits, we further analyzed the functional categories with more hits in them within the selected gene sets previously defined: genetic information processing and carbohydrate metabolism. Inside those categories, we found several proteins with relevant SNPs present in hot water environments that may be underlying selective pressures acting on those species' phenotypes and undergoing adaptation changes, that would be interesting to investigate in further detail.

However, this is just the beginning of the project, since it would be interesting to perform the same analysis using the same dataset but including future samples, as 7 years is not a very long time to observe a clear genotypic adaptation. Moreover, further analysis should also be done using the rest of the MAGs to try to see a general pattern of adaptation/evolution shared among all the species or groups of them.

Acknowledgements

I would like to express my deepest gratitude to Ramiro Logares and Francisco Latorre for all the support and encouragement that they have given me during my internship and the development of my Final Degree Project at the Institut de Ciències del Mar (ICM-CSIC). Thank you very much for your advice and everything you have taught me during this process. It has been a very nice experience and a very enjoyable stay in your laboratory group.

I would also like to extend my thanks to the ESCI-UPF Bioinformatics Degree coordinators and professors for the great experience and learning I have had coursing the degree.

To my family for encouraging and supporting me in my life decisions.

Last but not least, I would like to deeply thank my classmates and friends for all the good times and memories we have shared together and I am sure we will continue to share in the future.

Again, thank you all!

Supplementary Material

Supplementary material is available in an attached document and at Supplementary material.zip.

References

1. Piroddi, C., Coll, M., Liquete, C., Macias, D., Greer, K., Buszowski, J., Steenbeek, J., Danovaro, R., & Christensen, V. (2017). Historical changes of the Mediterranean Sea ecosystem: Modelling the role and impact of primary productivity and fisheries changes over time. *Scientific Reports*, *7*(1), 1–18. https://doi.org/10.1038/srep44491
2. Boisvert, S., Raymond, F., Godzaridis, É., Laviolette, F., & Corbeil, J. (2012). Ray Meta: Scalable de novo metagenome assembly and profiling. *Genome Biology*, *13*(12). https://doi.org/10.1186/gb-2012-13-12-r122
3. Uritskiy, G. V., Diruggiero, J., & Taylor, J. (2018). MetaWRAP - A flexible pipeline for genome-resolved metagenomic data analysis 08 Information and Computing Sciences 0803 Computer Software 08 Information and Computing Sciences 0806 Information Systems. *Microbiome*, *6*(1). https://doi.org/10.1186/s40168-018-0541-1
4. Hyatt, D., Chen, G. L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, *11*. https://doi.org/10.1186/1471-2105-11-119

5. Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4). https://doi.org/10.1038/nmeth.1923
6. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16). https://doi.org/10.1093/bioinformatics/btp352
7. Garrison, E., & Marth, G. (2012). *Haplotype-based variant detection from short-read sequencing*. http://arxiv.org/abs/1207.3907
8. Olm, M. R., Crits-Christoph, A., Bouma-Gregson, K., Firek, B. A., Morowitz, M. J., & Banfield, J. F. (2021). inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. *Nature Biotechnology*, *39*(6), 727–736. https://doi.org/10.1038/s41587-020-00797-0
9. Giner, C. R., Balagué, V., Krabberød, A. K., Ferrera, I., Reñé, A., Garcés, E., Gasol, J. M., Logares, R., & Massana, R. (2019). Quantifying long-term recurrence in planktonic microbial eukaryotes. *Molecular Ecology*, *28*(5), 923–935. https://doi.org/10.1111/mec.14929
10. Thorvaldsdóttir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, *14*(2). https://doi.org/10.1093/bib/bbs017
11. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. In *Nucleic Acids Research* (Vol. 25, Issue 17). https://doi.org/10.1093/nar/25.17.3389
12. Kanehisa, M., Sato, Y., & Morishima, K. (2016). BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *Journal of Molecular Biology*, *428*(4). https://doi.org/10.1016/j.jmb.2015.11.006
13. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, *44*(D1). https://doi.org/10.1093/nar/gkv1070
14. Madeira, F., Park, Y. M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A. R. N., Potter, S. C., Finn, R. D., & Lopez, R. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Research*, *47*(W1), W636–W641. https://doi.org/10.1093/nar/gkz268
15. Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, *6*(2). https://doi.org/10.4161/fly.19695
16. Eren, A. M., Esen, O. C., Quince, C., Vineis, J. H., Morrison, H. G., Sogin, M. L., & Delmont, T. O. (2015). Anvi'o: An advanced analysis and visualization platform for 'omics data. *PeerJ*, *2015*(10). https://doi.org/10.7717/peerj.1319
17. Mao, X., Cai, T., Olyarchuk, J. G., & Wei, L. (2005). Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. *Bioinformatics*, *21*(19). https://doi.org/10.1093/bioinformatics/bti430
18. Fröderberg, L., Houben, E. N. G., Baars, L., Luirink, J., & De Gier, J. W. (2004). Targeting and translocation of two lipoproteins in Escherichia coli via the SRP/Sec/YidC pathway. *Journal of Biological Chemistry*, *279*(30), 31026–31032. https://doi.org/10.1074/jbc.M403229200
19. Górna, M. W., Carpousis, A. J., & Luisi, B. F. (2012). From conformational chaos to robust regulation: The structure and function of the multi-enzyme RNA degradosome. *Quarterly Reviews of Biophysics*, *45*(2), 105–145. https://doi.org/10.1017/S003358351100014X