Note

This report contains a brief background about the tools and techniques used to do the different analysis, the commands I used and the answers to the questions.

1. METAGENOMICS

Shotgun metagenomic sequencing of the prokaryotic microbiome was run in two samples:

- One taking during the high temperature episodes
- Right after the episodes, when the temperature is back to normal and there is a bloom of algae

To deal with metagenomic data, we are going to use the metagenomic operational taxonomic units (mOTUs). The tool performs taxonomic profiling of known (sequenced) and unknown microorganisms at species-level resolution from shotgun sequencing data. The method clusters single-copy phylogenetic marker gene sequences from metagenomes and reference genomes into mOTUs to quantify their abundances in shotgun metagenomic samples [1]. The procedure is the following: with our two samples, using both forward and reverse reads, we perform a motus profiling and we merge results (optional) [2]:

```
$ motus profile -f /home/2019_2020/data/metagenomics/hotspring-hightemp.1.fq.gz -r /home/2019_2020/data/metagenomics/hotspring-hightemp.2.fq.gz -o hightemp_test.motus -n hightemp
$ motus profile -f /home/2019_2020/data/metagenomics/hotspring-normaltemp.1.fq.gz -r /home/2019_2020/data/metagenomics/hotspring-normaltemp.2.fq.gz -o normaltemp_test.motus -n normaltemp
$ motus merge -i hightemp_test.motus,normaltemp_test.motus -o results.motus
```

The resulting profile reports the relative abundance for each mOTU (ref + meta). The taxonomic profiles were generated, and the resulting profile contained the relative abundance for each mOTU (ref + meta). mOTUs uses 40 marker genes to generate the taxonomic profile [1].

Questions

What is the most abundant organism in high temperature?

```
$ cat results.motus | sort -t$'\t' -k2 -r | head -2
```

The most abundant organism in high temperature is *Aquifex aeolicus.*

- What's its relative abundance?

  Its relative abundance is 0.6189655514 (61.9%).

- How do you interpret the abundance number obtained?

  The abundance number obtained is the proportion of total reads aligned to an organism (in this case, *Aquifex aeolicus*) out of total reads aligned to all organisms.

- Is it a novel or known species? If possible, describe the most important features of such species.

  It is a known specie. *Aquifex aeolicus* is a hyperthermophile with an optimal growth temperature between 85 – 95ºC. It is a chemolithoautotrophy and thus, it uses inorganic carbon sources and inorganic chemical sources. [3].

What's the level of the most abundant organism in normal temperature?

```
$ cat results.motus | sort -t$'\t' -k3 -r | head -2
```

The most abundant organism in normal temperature is *Pelagibacteraceae species incertae sedis*, with a relative abundance of 0.2385013523 (23.85%).

- Is the high-abundant species in the high-temp sample detected here?

  ```
  $ cat results.motus | sort -t$'\t' -k2 -r | head -2
  ```

  No, it isn't. *Aquifex aeolicus* does not appear at normal temperature.

Why no algaes are not observed in the normal-temperature condition?

Algae is a diverse group of photosynthetic eukaryotic organisms. Algaes do not appear in the metagenomic analysis, neither at normal temperature nor high temperature, because a shotgun metagenomic prokaryotic sequencing was carried out and to analyze these sequences mOTUs database was used, which includes only prokaryotic organism. Therefore, eukaryotic organisms could not be detected.

Briefly describe your hypothesis explaining the differences observed between high and normal temp samples.

The most remarkable difference between high and normal temperature samples is the presence of *Aquifex aeolicus* at high temperatures but not at normal temperatures. *Aquifex aeolicus* can only grow at high temperatures, as its optimal growth temperature is between 85 – 95 ºC. Therefore, in this specific case, it can only grow after the high temperature episodes which we know they reach a temperature of around 90ºC.

The bloom of algae occurs after the high temperature episodes, after the presence of this organism when the temperature is back to normal. Therefore, it is possible that *Aquifex aeolicus* changes the environment somehow to favor the growth of algae.

2. GENOMICS

After the metagenomics analysis, sequenced cDNA samples from both normal and high temperature conditions (two biological replicates) are going to be analyzed. The quality checking is already done, so the resulting ones are high-quality reads in fast format.

In this step of the analysis, we are exploring our data to check how many samples do we have and what can we found in them, using basic linux commands.

Questions

1. How many samples do you have?

```
$ cd data/reads/
$ ls -l
```

We have four samples which are two biological replicates for both normal and high temperature conditions: hightemp01, hightemp02, normtemp01 and normtemp02.

2. How many reads do you have in each of your samples?

```
$ cd data/reads/
$ grep -c "^>" hightemp_01.fasta
$ grep -c "^>" hightemp_02.fasta
$ grep -c "^>" normal_01.fasta
$ grep -c "^>" normal_02.fasta
```

Sample	Number of reads
hightemp01	291814
hightemp02	289637
normtemp01	290331
normtemp02	291324

3. What kind of reads are they? (e.g. paired-end reads, mate-pair, single-end…)

Although in each sequence we have information about mate-pair coordinates (mate1 and mate2), there is just one file per sample and replicate and within the same file there are not sequences which belong to either one pair or the other. Therefore, the reads we are dealing with are single end.

4. Are all the reads of the same length?

```
$ grep -v "^[#>]" hightemp_01.fasta | awk '{print length}' | uniq
$ grep -v "^[#>]" hightemp_02.fasta | awk '{print length}' | uniq
$ grep -v "^[#>]" normal_01.fasta | awk '{print length}' | uniq
$ grep -v "^[#>]" normal_02.fasta | awk '{print length}' | uniq
```

Yes, all of the reads have the same length (100 nucleotides).

5. Just from the files you have been provided, could you say something about reads orientation (5' to 3', 3' to 5')? And what about DNA strand (forward or reverse strand)?

According to the standard notation for DNA sequences, all of the reads provided should be 5' to 3. Regarding the DNA strand orientation, as there is not information provided in the file, we assume that there will be both forward and reverse strands.

6. Are there any additional comments you would like to do about your reads?

The information available for all the sequences is the number of read followed by the name of the gene followed by the mate1 and mate2 coordinates. However, it is strange to have information about the name of each gene as reads have not been mapped yet and also it is strange to have information about mate coordinates when it seems we are dealing with single end reads.

3. READ MAPPING

Before performing other downstream analyses (variant calling, expression analysis, etc.) we need to map our reads to the reference. To do it, we are going to use bowtie2 [4].

First, an index of the reference genome is created:

```
$ mkdir 3.readmapping
$ mkdir 3.readmapping/index
$ bowtie2-build bowtie2-build data/refs/genome.fna 3.readmapping/index/genome_index
```

Next, samples are mapped to the reference genome, storing mapping statistics:

```
$ (bowtie2 -x 3.readmapping/index/genome_index -f data/reads/hightemp_01.fasta -S 3.readmapping/ht1_alg.sam) 2> 3.readmapping/ht1_alg_metrics.log

$ (bowtie2 -x 3.readmapping/index/genome_index -f data/reads/hightemp_02.fasta -S 3.readmapping/ht2_alg.sam) 2> 3.readmapping/ht2_alg_metrics.log

$ (bowtie2 -x 3.readmapping/index/genome_index -f data/reads/normal_01.fasta -S 3.readmapping/nt1_alg.sam) 2> 3.readmapping/nt1_alg_metrics.log

$ (bowtie2 -x 3.readmapping/index/genome_index -f data/reads/normal_02.fasta -S 3.readmapping/nt2_alg.sam) 2> 3.readmapping/nt2_alg_metrics.log
```

SAM files are converted to BAM files (same alignment information but in binary). We are using samtools [5] to do this step:

```
$ cd 3.readmapping/
$ mkdir alg_bam
$ cd alg_bam/
$ samtools view -bS alg_sam/ht1_alg.sam > alg_bam/ht1_alg.bam
$ samtools view -bS alg_sam/ht2_alg.sam > alg_bam/ht2_alg.bam
$ samtools view -bS alg_sam/nt1_alg.sam > alg_bam/nt1_alg.bam
$ samtools view -bS alg_sam/nt2_alg.sam > alg_bam /nt2_alg.bam
```

Questions

1. How many records are in your mapping (.sam/.bam) files? How many different reads are in your mapping (.sam/.bam) files? How these numbers compare with the number of reads in your original samples and with the alignment statistics (stats from bowtie2)?

The number of reads was searched within the sam files as they can be read directly without needing tools to view binary files such as bam.

```
# Number of records
$       grep -c "^read" ht1_alg.sam
$       grep -c "^read" ht2_alg.sam
$       grep -c "^read" nt1_alg.sam
$       grep -c "^read" nt2_alg.sam

# Number of different reads

$       grep "^read" ht1_alg.sam | uniq | wc -l
$       grep "^read" ht2_alg.sam | uniq | wc -l
$       grep "^read" nt1_alg.sam | uniq | wc -l
$       grep "^read" nt2_alg.sam | uniq | wc -l
# To check statistics from bowtie2

$       cat ht1_alg_metrics.log
$       cat ht2_alg_metrics.log
$       cat nt1_alg_metrics.log
$       cat nt2_alg_metrics.log
```

Sample	Number of records	Number of different reads	Number of reads (bowtie2 statistics)
High temperature 1	291814	291814	291814
High temperature 2	289637	289637	289637
Normal temperature 1	290331	290331	290331
Normal temperature 2	291324	291324	291324

The number of records, the number of different reads and the number of reads from bowtie2 statistics are the same for each sample. All reads from our samples are stored in the sam files generated and there are not equal reads in our samples.

2. How many reads map to a single location and how many to more than one (multiple mapping reads)? How do you think that multiple mapping reads could affect downstream analyses (variant calling and RNAseq)?

```
$ cd 3.readmapping/
$ cat ht1_alg_metrics.log
$ cat ht2_alg_metrics.log
$ cat nt1_alg_metrics.log
$ cat nt2_alg_metrics.log
```

Sample	Number of reads mapping to a **single location** (percentage)	Number of reads mapping to **multiple locations** (percentage)
High temperature 1	289275 (99.13%)	2522 (0.86%)
High temperature 2	287101 (99.12%)	2529 (0.87%)
Normal temperature 1	287957 (99.18%)	2366 (0.81%)
Normal temperature 2	289007 (99.20%)	2307 (0.79%)

Multiple mapping reads could affect to downstream analysis in different ways. It could affect to RNA seq by overestimating the presence of a sequence, as the same sequence mapping to multiple locations would be counted multiple times. It could also affect variant calling by detecting more variants that with just one sequence would not have been detected.

3. Could you use these mappings to perform an analysis of Copy Number Variation [6]?

Copy number variation (CNV) is a type of structural variation: specifically, it is a type of duplication or deletion event that affects a considerable number of base pairs. Therefore, as structural variations involve a larger number of base pairs and our reads were 100 nucleotides long, our mappings seem to be useless to perform a CNV analysis.
We could use the multiple mapping reads to detect duplications (or more repetitions, depending if the reads are mapping to two or more).

4. VARIANT CALLING

To perform variant calling we are going to use our bam files containing the information of the alignments. We are going to use samtools to do this. First, files obtained in the previous step of the analysis are organized in several directories.

Then, bam files are sorted:

```
$ cd 3.readmapping/
$ mkdir sorted_bam
$ samtools sort alg_bam/ht1_alg.bam -o sorted_bam/ht1_sorted.bam
$ samtools sort alg_bam/ht2_alg.bam -o sorted_bam/ht2_sorted.bam
$ samtools sort alg_bam/nt1_alg.bam -o sorted_bam/nt1_alg.bam
$ samtools sort alg_bam/nt2_alg.bam -o sorted_bam/nt2_alg.bam
```

After, bam files are merged into a single one, to be able to detect the variants in our samples:

```
$ cd sorted_bam/
$ samtools merge -f merged.bam ht1_sorted.bam ht2_sorted.bam nt1_sorted.bam nt2_sorted.bam
```

Finally, variant calling is performed using `bcftools mpileup` and `bcftools call`:

```
$ mkdir 4.variantcalling
$ bcftools mpileup -f data/refs/genome.fna 3.readmapping/sorted_bam/merged.bam > 4.variantcalling/raw.vcf
$ cd 4.variantcalling/
$ bcftools call -vc raw.vcf > calls.vcf
```

The bcf file was transformed into a “tab” separated file, containing the variant ID, its position, the reference allele, the alternative allele, the variant quality and the depth of coverage of the variant.

```
$ bcftools view calls.vcf | grep -v "^#" | cut -f 1,2,4,5,6,8 | sed 's#DP=\([0-9][0-9]*\).*#\1#' > calls.tsv
```

Questions

1. How many variants did you obtain? How many are SNPs, how many insertions and how many deletions?

```
$ cd 4.variantcalling/
$ grep "INDEL"  calls.vcf
```

There are not indels in our variants, i.e., there are not neither insertions nor deletions. Therefore, all the variants obtained are going to be SNPs and each one of them is going to be stored in a different line in our "tab" separated file.

```
$ wc -l calls.tsv
```

There are 34 SNPs.

2. How many variants have quality greater or equal than 10?

```
$ awk '{if($5>10)print$5}' calls.tsv | wc -l
```

There are 7 variants which have a quality greater than 10.

3. How many variants have depth of coverage greater or equal than 10?

```
$ awk '{if($6>10)print$6}' calls.tsv | wc -l
```

There are 11 variants have a coverage greater or equal than 19.

4. Identify the variant with best quality. Could this variant be affecting a gene? Which gene did you find, if any? Without actually checking it, could you give an example of how your variant could be affecting a gene product (e.g. a protein)?

```
$ sort -k6,6n calls.tsv | tail -1
```

ID	Position	Reference allele	Alternative allele	Quality	Depth
NC_000918.1	135330	A	C	221.999	247

```
$ cd data/refs/
$ grep -E '\<13[0-9]{4}\>' genes.gff
```

When executing previous commands, it was found that there was a gene including position 135330. This gene starts in the position 134585 and finishes in the position 136105, and it is nifA. Therefore, the variant is affecting a gene, specifically, nifA.

There are ways by which a variant which may cause different alterations in the protein. The variant may cause:

- Synonymous mutation: the amino acid does not change, and the protein has its function unaltered.
- Non-Synonymous mutation, which furthermore may be:
  - Sense mutation: in this case, the impact on the function of the variant will depend on the amino acid changed and its relevance for the protein function.
  - Non-sense mutations: causing a premature codon-stop and truncation of the protein. The impact of the variant on the function will depend on the truncated products formed.

nifA is a transcriptional factor which binds nucleotides. The variant may be causing a higher affinity of a lower affinity of the binding.

5. Repeat the variant calling using only a single .bam file, instead of merging them. What is the main difference you find in the results when you use only one file?

```
$ bcftools mpileup -f data/refs/genome.fna 3.readmapping/sorted_bam/ht1_sorted.bam > 4.variantcalling/rawHT1.vcf
$ bcftools call -vc 4.variantcalling/rawHT1.vcf > 4.variantcalling/callsHT1.vcf
$ bcftools view 4.variantcalling/callsHT1.vcf | grep -v "^#" | cut -f 1,2,4,5,6,8 | sed 's#DP=\([0-9][0-9]*\).*#\1#' > 4.variantcalling/callsHT1.tsv
$ wc -l 4.variantcalling/callsHT1.tsv
```

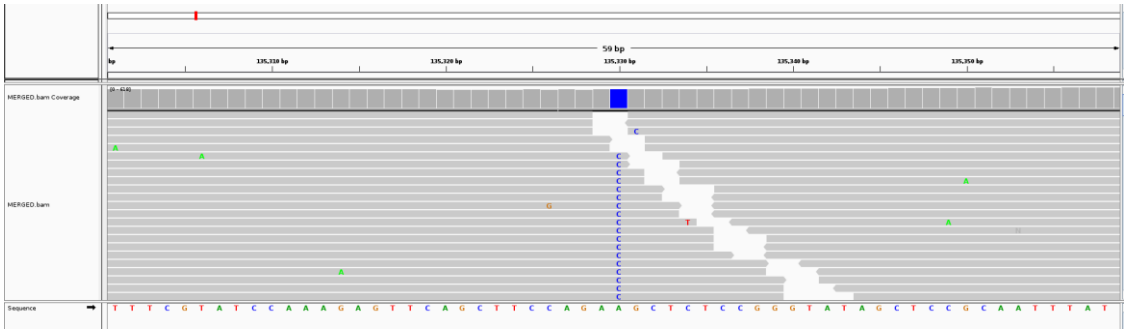
When repeating the variant calling with a single bam file, more variants are obtained. In this case, we are obtaining 397 variants in one single file vs 34 variants obtained in the file merging all samples. This is because when multiple files are merged, the variants obtained will be those common to all files.

6. Download, to your local machine, the files with the mappings and the fasta file of the genome. Use them to locate in IGV the best variant you have and capture the image of the variant. Note that you will likely need the indexes of the mappings (.bai) and genome (.fai) files.

To visualize the variant, the Integrative Genomics Viewer (IGV) was used [7].

```
$ samtools faidx genome.fna
$ samtools index merged.bam
```

The variant is shown in blue:



As it was expected, in the visualization we can see that the reference allele was A (Sequence, bottom of the image) and the alternative allele is C (variant observant in most of the reads).

5. DIFFERENTIAL EXPRESSION ANALYSIS

After the variant calling analysis, differential expression analysis is performed with the sorted bam files. First, we do a read count using the tool htseq-count and we join all count results into a single file:

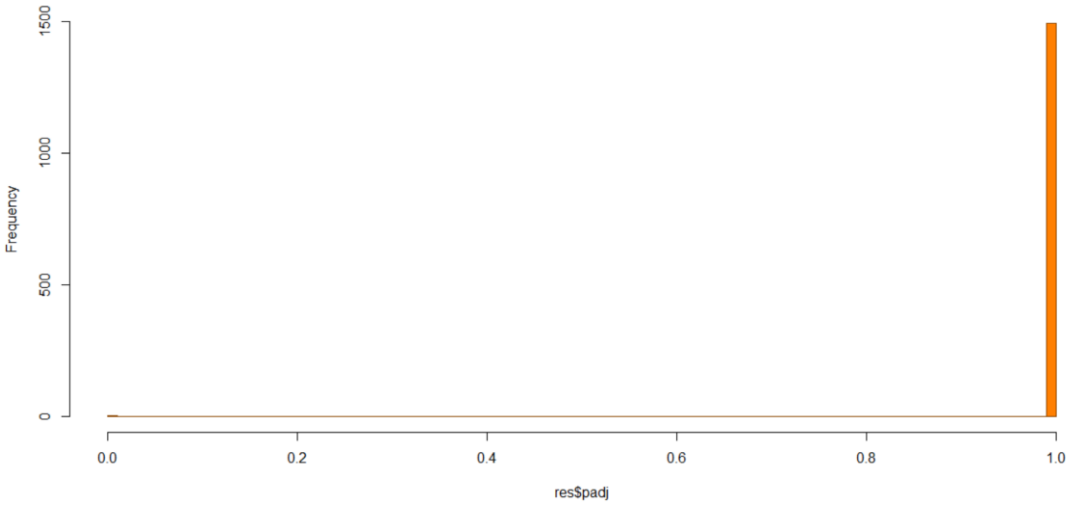
```
$ mkdir 5.diffexp/count

$ htseq-count -m union -i Name -t gene -f bam 5.diffexp/sorted_bam/ht1_sorted.bam data/refs/genes.gff > 5.diffexp/counts/ht1.count

$ htseq-count -m union -i Name -t gene -f bam 5.diffexp/sorted_bam/ht2_sorted.bam data/refs/genes.gff > 5.diffexp/counts/ht2.count

$ htseq-count -m union -i Name -t gene -f bam 5.diffexp/sorted_bam/nt1_sorted.bam data/refs/genes.gff > 5.diffexp/counts/nt1.count

$ htseq-count -m union -i Name -t gene -f bam 5.diffexp/sorted_bam/nt2_sorted.bam data/refs/genes.gff > 5.diffexp/counts/nt2.count


$ join ht1.count ht2.count | join - nt1.count | join - nt2.count > counts.txt
```

Finally, we are going to use Bioconductor package DESeq2 [8] in R to analyze the results. The R script to load our data and analyze it is the following:

```
## LOADING DATA ##
counts = read.table("counts.txt", header=F, row.names=1) # Load the raw counts table
colnames = c("Normal","Normal","High","High") # names for column names
my.design <- data.frame(row.names = colnames( counts ),
                group = c("Normal","Normal","High","High")
) # our experiment design for DESeq2 analysis
## installing DESeq2 ## # (if necessary)
# if (!requireNamespace("BiocManager", quietly = TRUE))
#  install.packages("BiocManager")
# BiocManager::install(version = "3.10")
# BiocManager::install("DESeq2")
library("DESeq2") # Load the DESeq2 package
#import data matrix from NtcA results to generate DeSeqDataSet
dds <- DESeqDataSetFromMatrix(countData = counts, colData = my.design, design = ~ group + group:group)
#DEF analysis
dds <- DESeq(dds)
resultsNames(dds)
res <- results(dds, contrast=c("group","Normal","High"))
res
```

```
res = na.omit (res)
##Histogram of p-values from the call to nbinomTest.
hist(res$pvalue, breaks=100, col="skyblue", border="slateblue", main="")
##Histogram of p-adj from the call to nbinomTest.
hist(res$padj, breaks=100, col="darkorange1", border="darkorange4", main="")
#plotMA
plotMA(res, main="DESeq2", ylim=c(-8,8))
#identifie genes in the plot
idx <- identify(res$baseMean, res$log2FoldChange)
rownames(res)[idx]
#rlogtranformation for PCA analysis
rld <- rlog(dds)
head(assay(rld), 100)
plotPCA(rld, intgroup=c("Location", "group"))
#heatmap
resOrdered <- rld[order(res$log2FoldChange),]
select_genes<-rownames(subset(resOrdered, res$padj< 0.01))
mat<-assay(rld)[select_genes,]
mat<-mat-rowMeans(mat)
heatmap(mat)
library("genefilter")
rld <- rlog(dds, blind=FALSE)
topVarGenes <- head(order(rowVars(assay(rld)),decreasing=TRUE),50)
heatmap( assay(rld)[ topVarGenes, ], scale="row")
##filter for significant genes,
##according to some chosen threshold for the false dicovery rate (FDR)
resSig = res[res$padj < 0.01, ]
## QUESTION 2 ##
filtered_padj <- subset(res, res$padj < 0.01)[c("pvalue", "padj", "log2FoldChange")]
```

1. Have a look to the p-adj histogram obtained (res$padj), what does this result mean?



There are very few genes with a very low p-value (close to 0), while most of the genes have a very high p-value (close to 1). Therefore, only very few genes have a differential expression.

2. How many genes showed a statistical (p-adj < 0.01) differential expression? The results has to be justified with a table showing all the altered genes (including, p-val, p-adj, fold change).

There are 4 genes which have a statistical differential expression (p-adj < 0.01) in normal and high temperature conditions: NP_213724.1, NP_213881.1, NP_Unk01 and NP_Unk02.

	p-value	p-adj	log2FoldChange
NP_213724.1	7.06392169180146e-72	2.6436726931567e-6	-3.01090249464287
NP_213887.1	2.27060888562002e-154	3.39910150177317e-151	-3.40402069638887
NP_Unk01	3.26044073489232e-72	1.62695992671127e-69	2.23940090115434
NP_Unk02	1.58018733549178e-74	1.1827702206156e-71	2.29494749439629

The two unknown genes are being overexpressed in normal conditions are: NP_Unk01, whose expression is increased by a factor of $2^{2.394}$ and NP_Unk02, whose expression is increased by a factor of $2^{2.295}$.

3. Taking all this data together, what can you say about the statistical significance of your DEA? Do you feel confident about your differentially expressed genes?

Taking all this data together, it is very likely that there are only 4 genes which have a differential expression significantly different. I feel confident about these genes as they have very low p-values and p-adjusted values.

6. FUNCTIONAL PREDICTION

The differential expression results gave you an idea about potential important upregulated/overexpressed genes. In this step of the analysis, we are going to use online databases and bioinformatic resources (NCBI Blast, STRING-DB).

First, we need to extract the sequence of each over expressed genes (NP_Unk01 and NP_Unk02) and they are stored in individual fasta files:

```
$ grep -A1 "NP_Unk01" /home/2019_2020/data/phylo/novel_proteome.faa > NP_Unk01.faa
$ grep -A1 "NP_Unk02" /home/2019_2020/data/phylo/novel_proteome.faa > NP_Unk02.faa
```

Questions

Has any other strain of the same abundant species been sequenced? (i.e. whole genome). Report it if so.

No, it has not. The only strain of *Aquifex aeolicus* which has been sequenced is VF5. [9]

Do all the overexpressed genes have any close homolog in similar strains or in other species/lineages?

To address this question, a blastp [10] was performed with the 2 amino acid sequences from the overexpressed genes, in fasta format against non-redundant protein sequences database. The results were the following.:

For NP_Unk01, there was not a close homologue. In fact, the best hit belongs to a methanogenic archeon mixed culture (53.80% identity). The best hit belonging to a single species is with Methanomassiliicoccus luminyensis (45.41% identity, E-value = 5e-154)

For NP_Unk02, the best hit was *Aquifex aeolicus*, and the best hit belonging to other species was *Hydrogenivirga caldilitoris* (85.77% identity)

*Note: not reported e-values are 0.

Do the overexpressed genes have any known molecular function (inferred from homologs)? Which function?

As mentioned in the previous exercise, NP_Unk01 does not have a close homologue. Nevertheless, the best hit is a **nitrogenase iron protein NifH**, which forms part of an enzymatic complex involved in nitrogen fixation.

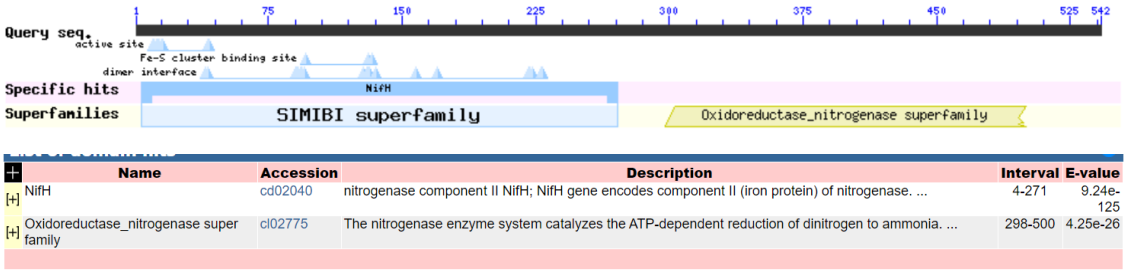
For NP_Unk02, the best hit is a transcriptional factor, more specifically is **sigma-54-dependent fis family transcriptional regulator**.

Do the overexpressed genes have any known domain?

To address this question, a tool from NCBI was used to search conserved domains [11].

NP_Unk01

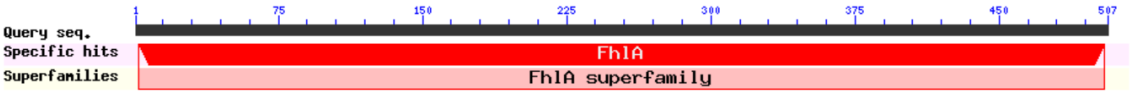
- NifH: nitrogenase component II NifH
- Oxidoreductase nitrogenase super family



Name	Accession	Description	Interval	E-value
[+] NifH	cd02040	nitrogenase component II NifH; NifH gene encodes component II (iron protein) of nitrogenase. ...	4-271	9.24e-125
[+] Oxidoreductase_nitrogenase super family	cl02775	The nitrogenase enzyme system catalyzes the ATP-dependent reduction of dinitrogen to ammonia. ...	298-500	4.25e-26

NP_Unk02

- FhIA: Transcriptional regulation containing GAF, AAA-type ATPase and DNA-binding Fis domains



Name	Accession	Description	Interval	E-value
[+] FhlA	COG3604	Transcriptional regulator containing GAF, AAA-type ATPase, and DNA-binding Fis domains ...	2-505	0e+00

Are the overexpressed genes functionally related? (i.e. protein-protein interactions)

To address this question, STRING database was used. This database contains functional association protein networks [12]. Both NP_Unk01 and NP_Unk02 sequences were searched agains *Aquifex aeolicus* and neither of them gave an annotated protein interaction network. This may have different explanations: (1) The interactions have neither been described nor annotated yet; (2) The proteins are not interacting directly and they are not functionally related; (3) The proteins are not interacting directly but they are functionally related indirectly.

However, searching in PubMed database [13], using as keywords “NifH” and “sigma-54” there are some papers describing a relationship between NifH and sigma-54. Transcriptional factor sigma-54 is necessary for the expression of some genes involved in nitrogen fixation including NifH [14]. This is not a protein-protein interaction, but a protein-DNA interaction. Although this interaction is not described in *Aquifex aeolicus*, it shows that the functions are related.

Could those function you inferred be related with the bloom of algae observed in the hot spring after high temperature? What would be such relationship? Briefly elaborate your hypothesis.

The overexpression of sigma-54 transcription factor lead to an overexpression of NifH, involved in nitrogen fixation. The rapid growth of *Aquifex aeolicus* during the high temperature episodes and the stimulation of nitrogen fixation increase the levels available of nitrogen in the hot spring. This condition favors the bloom of algae, as stated in this reference "The potential for blooms comes from nutrient pollution, an overabundance of the essential plant nutrients nitrogen and phosphorus" [15]. If we also consider that bloom of algae is also favored by warm temperatures [16], the conditions after the high temperature episodes (warm temperature and increased availability of nitrogen) are optimal for the bloom of algae.

7. PHYLOGENETIC ANALYSIS

The last step of the analysis is to perform a phylogenetic analysis of the upregulated genes comparing them against other prokaryotic proteomes.

```
$ mkdir 7.phylogenia
$ cp /home/2019_2020/data/phylo/* 7.phylogenia/
$ cd 7.phylogenia/
```

First, we run a blast search for each over expressed protein against all reference proteomes. Extract hits with e-value <= 0.00001 (tip: you can use blast parameters for this)

To do it, we have to create a blast database:

```
$ makeblastdb -dbtype prot -in all_ref_proteomes.faa
```

The database built contains 88473 sequences. Then, we run the blastp for the overexpressed genes (NP_Unk01 and NP_Unk02).

```
$ blastp -task blastp -query NP_Unk01.faa -db all_ref_proteomes.faa -outfmt 6 -evalue 0.00001 > blastp_results_NP_Unk01.txt

$ blastp -task blastp -query NP_Unk02.faa -db all_ref_proteomes.faa -outfmt 6 -evalue 0.00001 > blastp_results_NP_Unk02.txt
```

We extract the sequences in fasta format with the python script provided:

```
$ python extract_seqs_from_blast_result.py blastp_results_NP_Unk01.txt all_ref_proteomes.faa > blastp_results_NP_Unk01.faa

$ python extract_seqs_from_blast_result.py blastp_results_NP_Unk02.txt all_ref_proteomes.faa > blastp_results_NP_Unk02.faa
```

Build a phylogenetic tree out of the fasta file, using clustalo (a tool to perform alignments) [17] and iqtree (a tool to construct trees) [18].

To build the alignments, we execute the following command:

```
$ clustalo -i blastp_results_NP_Unk01.faa > homologs_NP_Unk01.alg

$ clustalo -i blastp_results_NP_Unk02.faa > homologs_NP_Unk02.alg
```

Then, once the alignment files are built, we build the trees:

```
$ iqtree -s homologs_NP_Unk01.alg -fast

$ iqtree -s homologs_NP_Unk02.alg -fast
```

Four files are generated: iqtree, treefile, mldist and log.

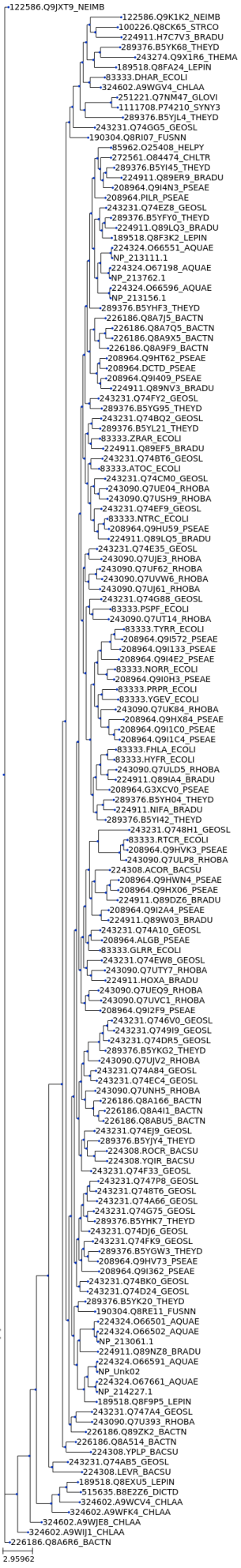
The last step of the analysis is using python and ete3 package [19] to visualize and manage the previous trees.

The following steps have been done using python. The python jupyter notebook can be found in my directory: pablo.gomez.sacristan/7.phylogenia, as "Tree visualization.ipynb"

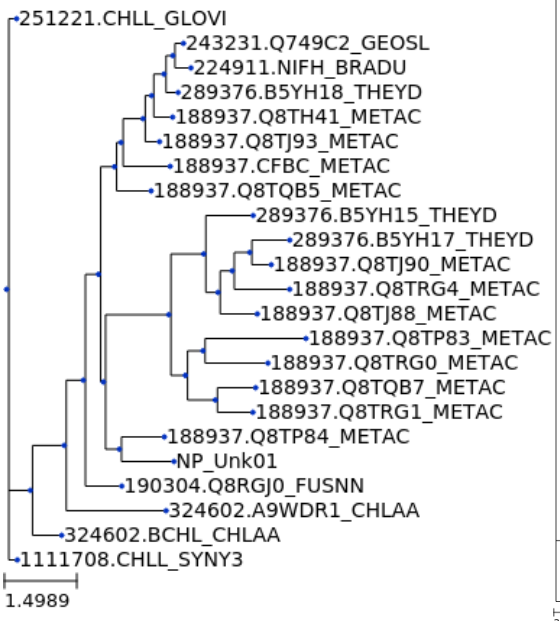
Once the tree is constructed, we visualize it before relabeling the branches (page 22 of this report).

Then, the tree is relabeled and visualized it again and interpret it. Relabeled tree 1 and 2 are shown in pages 23 and 24 of this report respectively. As they contain a lot of information, an image amplifying the region where either NP_Unk01 or NP_Unk02 are located.

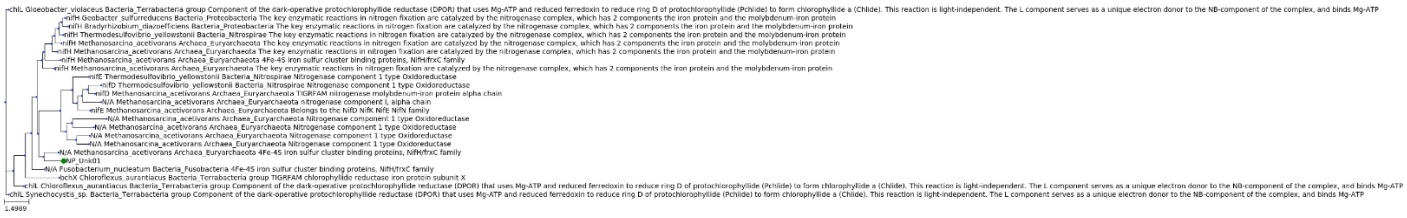
Tree 2 (NP_Unk02)

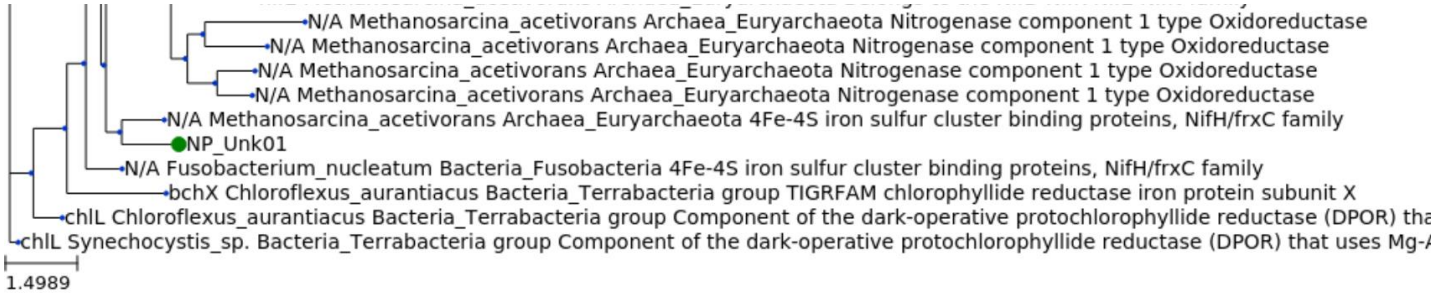


Tree 1 (NP_Unk01)

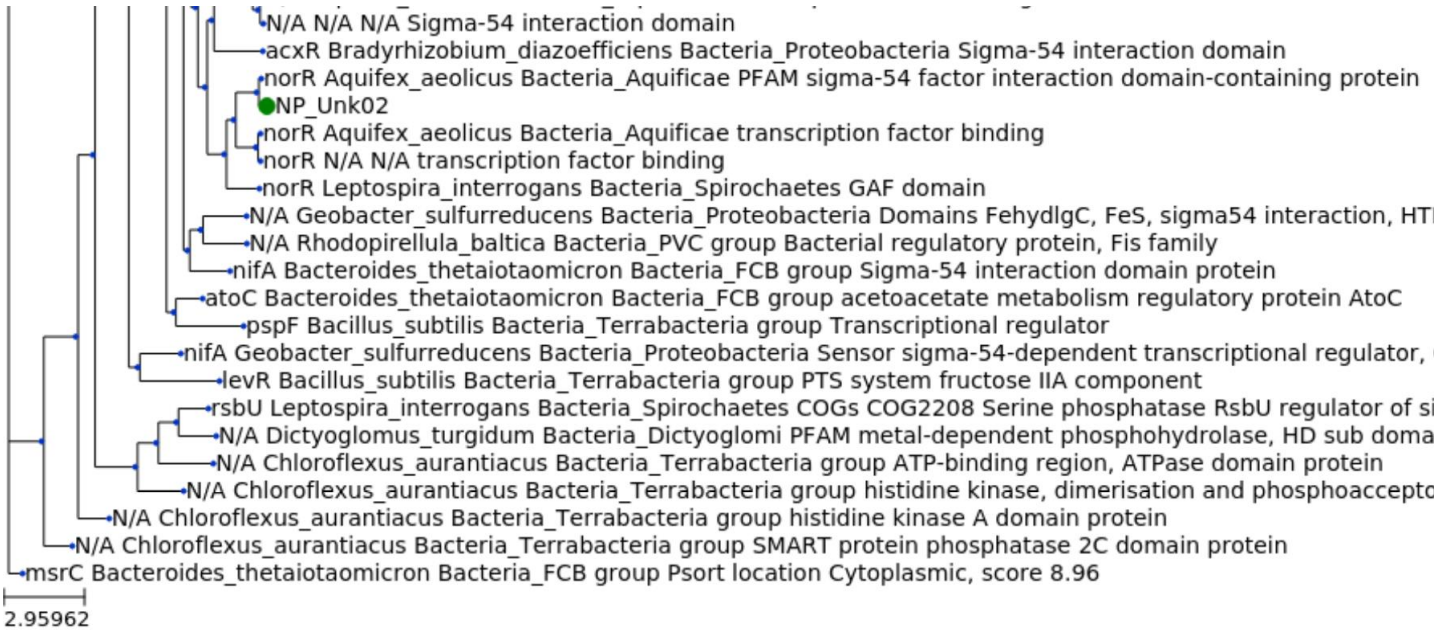


Tree 1 (NP_Unk01)





Tree 2 (NP_Unk02)



Questions

For each overexpressed gene…

1. What is the closest ortholog from a phylogenetic point of view? From what species?

For NP_Unk01, the closest ortholog is 4Fe-4S iron sulfur cluster binding protein, from *Methanosarcina acetivorans*.

For NP_Unk02, the closest ortholog is norR from Leptospira interrogans.

2. Do orthology assignment support your previous functional annotations? (you might need to look up the functional annotation (i.e. gene names) of close orthologs)

Yes, it does. The closest orthologs of NP_Unk01belong to NifH/frxC family and the closest orthologs of NP_Unk02 are transcription factors.

3. Are all genes present in the reference proteome of the same species? Why not? Do they all over expressed genes share the same evolutionary history?

No, they are not. The reference proteome must have different species because if not it is not possible to detect orthologs.

No, they do not. The over expressed genes have different evolutionary histories, as NP_Unk01 comes from a speciation event and NP_Unk02 from a duplication event.

Taking all the project together, what's your best hypothesis for the effect observed in the hot spring?

The metagenomics analysis allowed us to discover that there was an organism, *Aquifex aeolicus*, growing in a very significant way during the high temperature episodes compare to the normal temperature conditions. The differential expression analysis, which was performed after processing our sequences (genome exploring and read mapping), revealed that there were two genes, NP_Unk01 and NP_Unk02 overexpressed as consequence of the high temperature conditions. The functional prediction and the phylogenetic analysis revealed that the most probable functions for our unknown genes were related to nitrogen fixation and activation of the transcription of a protein involved in this process. In addition, variant calling analysis revealed that there was a very good quality variant related to a gene also involved in nitrogen fixation. Therefore, the final hypothesis taking all data together is that during high temperature episodes, the increase in the levels of *Aquifex aeolicus* and the activation of the expression of genes involved in nitrogen fixation result in the increase of the nitrogen available in the environment. Then, when temperature decrease, a bloom of algae appears as the conditions for this to happen are optimal.

REFERENCES

[1] https://bio.tools/motus
[2] https://motu-tool.org/tutorial.html
[3] Deckert, G., Warren, P. V., Gaasterland, T., Young, W. G., Lenox, A. L., Graham, D. E., ... & Huber, R. (1998). The complete genome of the hyperthermophilic bacterium Aquifex aeolicus. Nature, 392(6674), 353.
[4] http://bowtie-bio.sourceforge.net/bowtie2/
[5] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... & Durbin, R. (2009). The sequence alignment/map format and SAMtools. Bioinformatics, 25(16), 2078-2079.
Genomics
[6] https://en.wikipedia.org/wiki/Copy-number_variation
[7] James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, Jill P. Mesirov. Integrative Genomics Viewer. Nature Biotechnology 29, 24–26 (2011).
[8] Love MI, Huber W, Anders S (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." Genome Biology, 15, 550. doi: 10.1186/s13059-014-0550-8.
[9] https://www.ncbi.nlm.nih.gov/genome/1049?genome_assembly_id=300501
[10] Kent, W. J. (2002). BLAT—the BLAST-like alignment tool. Genome research, 12(4), 656-664.
[11] https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi
[12] https://string-db.org/
[13] https://www.ncbi.nlm.nih.gov/pubmed/
[14] Buck, M., & Cannon, W. (1992). Activator-independent formation of a closed complex between σ54-holoenzyme and nifH and nifU promoters of Klebsiella pneumoniae. Molecular microbiology, 6(12), 1625-1630.
[15] https://www.sjrwmd.com/education/algae/#why-do-algal-blooms-occur
[16] https://www.neefusa.org/nature/water/algal-blooms-are-blooming
[17] Madeira, F., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., ... & Lopez, R. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic acids research.
[18] L.-T. Nguyen, H.A. Schmidt, A. von Haeseler, B.Q. Minh (2015) IQ-TREE: A fast and effective stochastic algorithm for estimating maximum likelihood phylogenies. Mol. Biol. Evol., 32:268-274. https://doi.org/10.1093/molbev/msu300
[19] Jaime Huerta-Cepas, François Serra and Peer Bork. "ETE 3: Reconstruction, analysis and visualization of phylogenomic data." Mol Biol Evol (2016) doi: 10.1093/molbev/msw046