

# **Final project**

## Genomic Data Analysis and Visualization (GDAV) 2021/2022

Sandra Alonso Paz  
sandra.apaz@alumnos.upm.es

February 1, 2022



### **1 Project description**

Your lab has identified an interesting effect in a hot spring located in Iceland:

Coinciding with the activity of a nearby volcano, the hot spring undergoes events of very high temperature. You noticed that, after such episodes of high temperature (close to 90 degrees!), a bloom of algae living in the same environment happens.

You wrote a grant proposing to investigate this effect and, lucky you!, you got a very generous funding from Tyrell Corp. Your grant proposed to perform an in depth genomic and metagenomic exploration of this singular hot spring ecosystem, including 8 main work packages:

1. Metagenomic analysis
2. Genome analysis (basic checks)
3. Genome analysis (read mapping)
4. Genome analysis (variant calling)
5. Differential expression analysis
6. Functional analysis of the responsible genes
7. Phylogenetic analysis of the responsible genes
8. Conclusions

## 2 Metagenomics

As a first step, you decide to run shotgun metagenomic sequencing of the prokaryotic microbiome in two kind of samples, obtained at different times:

1. One sample taken during the high temperature episodes.
2. another sample taken right after the episodes, when the temperature is back to normal and there is a bloom of algae.

You extracted the DNA present in each sample, and sent for Illumina Sequencing. After a few weeks, the sequencing results from your two metagenomic samples just arrived!

The raw read files (reverse and forward) were produced by Illumina pair-end sequencing and are now located in your computing server:

```
ls /final_project/
./metagenomics-hotspring-hightemp.1.fq.gz
./metagenomics-hotspring-hightemp.2.fq.gz
(forward and reverse reads from the high temperature sample)

./metagenomics-hotspring-normaltemp.1.fq.gz
./metagenomics-hotspring-normaltemp.2.fq.gz
(forward and reverse reads from the normal temperature sample)
```

### 2.1 Tasks and questions

#### 2.1.1 Use the tool mOTUs to perform a taxonomic profiling of your samples.

(Warning: Check mOTUs parameters to process both forward and reverse read files!)

In the metagenomics samples we were given we have sequences from many different genes and species. In order to know which species is the predominant in the sample, we must perform a sample's profiling. Specifically, I will use taxonomic profiling. This study will contain a list of detected taxa and their estimated relative abundances as well as the various diversity indices [1].

For performing this analysis I have made use of mOTUs tool [2] in both samples (normal and high temperature). Focusing on the parameters used, I utilized -f for the forward file, -r for the reverse file and -o for the output file or motus file. This final file will have .motus extension.

#### Commands:

```
» motus profile -f metagenomics-hotspring-normaltemp.1.fq.gz
-r metagenomics-hotspring-normaltemp.2.fq.gz -o ./metagenomics/normaltemp.motus

» motus profile -f metagenomics-hotspring-hightemp.1.fq.gz
-r metagenomics-hotspring-hightemp.2.fq.gz -o ./metagenomics/hightemp.motus
```

### 2.1.2 What is the most abundant organism in high-temperature?

As the result motus file was not sorted, I arranged it by the relative abundance which refers to the evenness of distribution of individuals among species in a community [3].

The command used is shown below. I used grep command for selecting and -v option for deleting headers ("^#"). Then I joined the sort command for arranging the file by its second column (-k2) and in reverse order (nr) for showing the highest value on the top. Finally, I showed just the first result (highest relative abundance) using head command.

**Command:** » `grep -v "^#" hightemp.motus | sort -t$' t' -k2nr | head -n1`

**Output :** Aquifex aeolicus [ref\_mOTU\_v25\_10705] 0.8468378081

#### 1. What's its relative abundance?

The most abundant organism is Aquifex aeolicus whose relative abundance is 84.68%. This means that out of all the genomes identified, 84.68% has been identified specifically as Aquifex aeolicus.

#### 2. Is it a novel or known species?

Aquifex aeolicus was one of the earliest diverging species, and is one of the most thermophilic bacteria known [4].

#### 3. Has the same organism been sequenced before? (i.e. there is a public whole genome sequence in current databases). Provide the sources supporting your answer.

Aquifex aeolicus is the first hyperthermophilic bacterium to have its genome sequence completely determined. It is also the first obligate chemolithoautotrophic bacterium to be completely sequenced [5]. However, according to NCBI database, only the strain VF5 of Aquifex aeolicus is publicly accessible [6].

#### 4. If possible, describe the most important features of such species.

Aquifex aeolicus is a bacteria that belongs to the Aquificae Phylum which constitute an important component of the microbial communities at elevated temperatures [7]. It grows at 85°C under a H<sub>2</sub>, CO<sub>2</sub>, O<sub>2</sub> and mineral salts atmospheres, for example near underwater volcanoes or hot springs. Although this organism grows until 95°C, the extreme thermal limit of the Bacteria, only a few specific indications of thermophily are apparent from the genome [8].

Focusing on its metabolism, it is a chemoautotrophic species (an organism which uses an inorganic carbon source for biosynthesis and an inorganic chemical energy source). Metabolic flexibility seems to be reduced as a result of the limited genome size [4].

### 2.1.3 What are the most abundant organisms in normal-temperature (i.e.>1%)?

This time I applied the same command options for the second file sample (normal temperature sample):

**Command:** » `grep -v "^#" normaltemp.motus | sort -t$' t' -k2nr | head -n1`

**Output :** Methanococcus maripaludis [ref\_mOTU\_v25\_01426] 0.0480790524

The most abundant organism is Methanococcus maripaludis whose relative abundance is 4.807%.

1. **Are they also present in the high-temperature samples? At which relative abundance?**

For this task I used grep command to find out which line contained this species name:

**Command:** » `grep 'Methanococcus maripaludis' hightemp.motus`

**Output :** `Methanococcus maripaludis [ref_mOTU_v25_01426] 0.1317584768`

Yes, *Methanococcus maripaludis* is also present in high temperatures with a relative abundance of 13.17%.

2.1.4 **In your opinion, which is the condition (high or normal temperature) with a greater level of alpha biodiversity?**

As both files have the same number of lines and therefore contains the same number of species we can compare both. For this task I have used two commands in order to count the lines in which the relative abundance is over 0.0 which means that the species is represented in the sample.

Firstly I selected the file by cat command and then I counted the lines (`wc -l`) whose relative abundance is not 0.0 (`grep -v`).

**Command:** » `cat hightemp.motus | grep -v '0.0000000000' | wc -l`

**Output :** 19

**Command:** » `cat normaltemp.motus | grep -v '0.0000000000' | wc -l`

**Output :** 233

In the light of these results, there is a greater level of alpha biodiversity in the normal temperature environment because I obtained a higher number of different species found in this sample.

This is reasonable because most of known bacteria are inactivated in temperatures ranging above 49°C [9]. Once that temperature is reached, only thermophile bacteria belongs in the ecosystem and therefore the biodiversity decreases.

2.1.5 **Can you detect any algae organisms in the normal-temperature? Report if so. If not, why not?**

No algae is observed either in normal or high temperature, we only have detected bacteria. There are two main explanations:

- I have developed shotgun metagenomic sequencing of the prokaryotic microbiome. This means that we have only obtained prokaryotic sequences. As eukaryotic organisms, algae is not included on the results.
- Concretely, I have used [mOTU](#) which identifies species by their 16S RNA sequence (a component only present in prokaryotic organisms) [10]. As eukaryotic organisms, algae do not have the 16S ribosomal RNA and therefore cannot be identified using mOTU.

**2.1.6 Briefly describe the main differences observed between high and normal temperature samples. Is there any biological insight you may want to comment on?**

As any other environmental factor, temperature changes affect environmental conditions and consequently, its living beings. In this scenario, temperature rises and consequently, increases survival opportunities of *Aquifex aeolicus* which as a thermophilic organism, perfectly grows at high temperature and small oxygen concentrations. This is why we found an 84% of relative abundance. Thanks to the likely environment conditions and to its chemoautotrophic metabolism, during high temperatures conditions it can easily be fed.

However, when temperature becomes lower, *Aquifex aeolicus* is no longer under its optimal conditions for growth, and its abundance is reduced. Consequently, nutrients left behind by *Aquifex aeolicus* could be used by other kinds of bacteria which prefer lower temperatures to grow. A good example is *Methanococcus maripaludis* whose relative abundance raises as temperature get normal again.

### 3 Genome Analysis (basic checks)

You are very excited with your preliminary findings that one specific organism is very abundant in high-temperature episodes. To further characterize it, your lab isolated the organism and

1. sequenced its whole genome.
2. performed a RNAseq analysis of samples from both cultures at normal-temperature and high-temperature conditions, two biological replicates each

The sequencing company already assembled the genome for you and performed quality checking for the reads in all samples, providing us only high quality reads in Fasta format.

You sit in front of your computer. Your coffee cup is still smoking and everybody is quietly hitting the keyboard in the lab. You open a bash terminal and 'ls' the directory where you left ready both reference data and the raw sequencing reads. There they are. First things first... you want to be sure what you are dealing with. You will need to unzip your data and begin to analyze.

#### 3.1 Tasks

##### 3.1.1 Check your RNAseq samples (folder RNAseq/) and answer the following questions

**Command:** » `cd RNAseq/`

**Command:** » `ls -l`

**Output :** total 43900

```
-rw-r--r--. 1 sandra.apaz sandra.apaz 5867401 Jan 12 16:09 hightemp01.r1.fq.gz
-rw-r--r--. 1 sandra.apaz sandra.apaz 5868305 Jan 12 16:09 hightemp01.r2.fq.gz
-rw-r--r--. 1 sandra.apaz sandra.apaz 5870385 Jan 12 16:09 hightemp02.r1.fq.gz
-rw-r--r--. 1 sandra.apaz sandra.apaz 5871104 Jan 12 16:09 hightemp02.r2.fq.gz
-rw-r--r--. 1 sandra.apaz sandra.apaz 5365704 Jan 12 16:10 normal01.r1.fq.gz
-rw-r--r--. 1 sandra.apaz sandra.apaz 5364988 Jan 12 16:09 normal01.r2.fq.gz
-rw-r--r--. 1 sandra.apaz sandra.apaz 5366424 Jan 12 16:09 normal02.r1.fq.gz
-rw-r--r--. 1 sandra.apaz sandra.apaz 5363958 Jan 12 16:09 normal02.r2.fq.gz
```

Firstly it is mandatory to unzip every file. For this task I have used gzip command with -d option for decompressing files:

**Command:** » `gzip -d hightemp01.r1.fq.gz1`

**Command:** » `gzip -d hightemp01.r2.fq.gz1`

**Command:** » `gzip -d hightemp02.r1.fq.gz1`

**Command:** » `gzip -d hightemp02.r2.fq.gz1`

**Command:** » `gzip -d normaltemp01.r1.fq.gz1`

**Command:** » `gzip -d normaltemp01.r2.fq.gz1`

**Command:** » `gzip -d normaltemp02.r1.fq.gz1`

**Command:** » `gzip -d normaltemp02.r2.fq.gz1`

## 3.2 Questions

### 3.2.1 How many samples do you have?

**Command:** » `ls`

**Output:** `hightemp01.r1.fq hightemp02.r1.fq normal01.r1.fq normal02.r1.fq  
hightemp01.r2.fq hightemp02.r2.fq normal01.r2.fq normal02.r2.fq`

There are 4 samples. Although there are 8 files in the RNAseq folder, every sample is formed by two files. For instance, `hightemp01.r1.fq` and `hightemp01.r2.fq` are the forward and reverse file for `hightemp01` sample.

### 3.2.2 How many reads do you have in each of your samples?

As every read starts with '@', I executed a `grep` command joined to `wc` command to count the lines which included an @ at the beginning of the line. In fact, forward and reverse files from the same sample will have the same number of reads. Therefore, I included below just one command per sample:

**Command:** » `grep '^ @' 'hightemp01.r1.fq' | wc -l`

**Output:** `318693`

**Command:** » `grep '^ @' 'hightemp02.r1.fq' | wc -l`

**Output:** `318693`

**Command:** » `grep '^ @' 'normal01.r1.fq' | wc -l`

**Output:** `288742`

**Command:** » `grep '^ @' 'normal02.r1.fq' | wc -l`

**Output:** `288742`

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

As I mentioned in section 3.2.1, I had been given two files (forward and reverse) for each sample, therefore, this reads should be mate-paired or paired-end. Nevertheless, to be sure about it, let's see the first line of one of the fastaq files:

**Command:** » `head -n1 hightemp01.r1.fq`

**Output:** `@AQUIFEX_00001_33_357_0:0:0_0:0:0_0/1`

This first line reveals a lot of information. For instance, we know the specie id (`AQUIFEX_00001_33_357_0`), the flowcell lane, tile number within the flowcell lane, 'x'-coordinate and 'y'-coordinates of the cluster within the tile and finally, the member of the pair which could be /1 or /2. In our case, this read is member 1 as it ends in /1. This attribute is only included in paired-end or mate-pair reads only.

In order to classify these reads in paired-end or mate-pair, I needed to do a deep research. Paired-end reads are typically short (50-300) reads, most of them often use Illumina HiSeq, MiSeq or NovaSeq protocols. Both pairs originate from a single fragment which is sequenced from either end. In contrast, mate pairs arise from a fragment that is circularized before sequencing [11].

Having this in mind, and as we know they are short reads and the lab has used Illumina for sequencing, current reads are paired-end reads.





[illegible]

### 3.2.6 Are there any additional comments you would like to make about your reads?

While reviewing every read of each sample, I realized that every single read shared the same forth line. This line encodes the quality values for the sequence in the second line of each read. Each character of the fourth line represents a Q score value for each sequence letter present in the second line of the read. Higher Q scores indicate smaller probability of error and lower Q scores can result in a significant portion of the reads being unusable. They may also lead to increased false-positive variant calls, resulting in inaccurate conclusions [12].

Sharing the same quality value for all the reads seems quite suspicious. Continuing, I searched ASCII Codes and Q-Scores of the different possible symbols [13]. In our case the unique symbol present is "I" which has the best Q-score (Q40) in the Illumina quality Score Encoding. Over Q30, the inference base call accuracy is 99.9%.

Concluding, this seems a very idyllic situation where all the reads reached the maximum Q score and therefore indicates that it has a tiny probability of error.

## 4 Genome Analysis (read mapping)

After checking your reads, you'd like to perform several downstream analyses, including variant calling, expression analysis, etc. You decide to begin by mapping your reads to the assembled genome.

### 4.1 Task

#### 4.1.1 First, you will need to create an index of the reference genome using BWA INDEX.

Firstly I created a new directory for working and copy the genome.fasta file inside of it. This will prevent the given file from being corrupted.

```
Command: » mkdir mapping
Command: » cp genome.fasta ./mapping/genome.fasta
Command: » cd mapping
Command: » bwa index genome.fasta
Command: » ls
Output: genome.fasta.amb genome.fasta.ann genome.fasta.bwt
genome.fasta.pac genome.fasta.sa
```

#### 4.1.2 Next, map your samples to the reference genome using BWA MEM. (Note: check the bwa mem options for mapping the type of reads you have: single-end, paired-end, ...).

In order to map my reads which are paired-end, I used the bwa mem command using the reference genome generated from bwa index and both files of reads for each sample. I saved it in a SAM file called bwa.sam and I added an error file which will be written in error case [14].

```
Command: » bwa mem genome.fasta ../RNAseq/hightemp01.r1.fq
../RNAseq/hightemp01.r2.fq > bwa_h_1.sam 2>bwa.error
Command: » bwa mem genome.fasta ../RNAseq/hightemp02.r1.fq
../RNAseq/hightemp02.r2.fq > bwa_h_2.sam 2>bwa.error
Command: » bwa mem genome.fasta ../RNAseq/normal01.r1.fq
../RNAseq/normal01.r2.fq > bwa_n_1.sam 2>bwa.error
Command: » bwa mem genome.fasta ../RNAseq/normal02.r1.fq
../RNAseq/normal02.r2.fq > bwa_n_2.sam 2>bwa.error
```

#### 4.1.3 Finally, create BAM files for your mappings using SAMTOOLS and remove the SAM files afterwards. The final BAM file should contain the original header information which is present in the SAM file (tip: check the parameters available for 'samtools view' [14]).

For this task I firstly activated samtools environment. Continuing, I made use of samtools view command with -h option to keep headers of the SAM file in the output BAM file. In addition I included -b option for setting the output file type as BAM file.

```

Command: » conda activate samtools
Command: » samtools view -h -b bwa_h_1.sam > bwa_h_1.bam
Command: » samtools view -h -b bwa_h_2.sam > bwa_h_2.bam
Command: » samtools view -h -b bwa_n_1.sam > bwa_n_1.bam
Command: » samtools view -h -b bwa_n_2.sam > bwa_n_2.bam

```

In order to check if the final BAM file contains the original header information which was present in the SAM file I used again the samtools view command. This time I included -H option for just showing the headers.

```

Command: » samtools view -H bwa_h_1.sam
Output: @SQ SN:Aquifex_genome LN:1556396
@PG ID:bwa PN:bwa VN:0.7.17-r1188 CL:bwa mem genome.fasta
../RNAseq/hightemp01.r1.fq ../RNAseq/hightemp01.r2.fq

```

```

Command: » samtools view -H bwa_h_1.bam
Output: @SQ SN:Aquifex_genome LN:1556396
@PG ID:bwa PN:bwa VN:0.7.17-r1188 CL:bwa mem genome.fasta
../RNAseq/hightemp01.r1.fq ../RNAseq/hightemp01.r2.fq

```

## 4.2 Questions

**4.2.1** How many records (“mappings”) are in your mapping (BAM) files? How many different reads are in your mapping (BAM) files? How do these numbers compare to each other?

For counting the number of records, I used samtools view option -c [15]:

```

Command: » samtools view -c bwa_h_1.bam
Output: 637890
Command: » samtools view -c bwa_h_2.bam
Output: 637950
Command: » samtools view -c bwa_n_1.bam
Output: 577503
Command: » samtools view -c bwa_n_2.bam
Output: 577498

```

For counting the number of different reads I selected the first column of each entry with cut -f1 and counted them (wc -l) without repetition (uniq).

```

Command: » samtools view bwa_h_1.bam | cut -f1 | uniq | wc -l
Output: 318693
Command: » samtools view bwa_h_2.bam | cut -f1 | uniq | wc -l
Output: 318693
Command: » samtools view bwa_n_1.bam | cut -f1 | uniq | wc -l
Output: 288742
Command: » samtools view bwa_n_2.bam | cut -f1 | uniq | wc -l
Output: 288742

```

As I have created the BAM file using two different Fastq files (forward and reverse reads) there are around half of unique reads. For instance, there are 637890 reads in bwa\_h\_1.bam and 318693 are unique (99.92%). Therefore I can conclude that there are 318693 complete reads for the high temperature sample and 288742 for the normal temperature sample.

**4.2.2 Comment about how the previous numbers compare with the number of reads in your original samples: are those numbers different? What could be the reason for the differences, if any? Could the differences, if any, affect downstream analyses?).**

As seen in section 3.2.2, there are almost the same number of reads in the initial Fastq files and in the BAM files obtained in section 4.2.1. However, there is a tiny difference between them (less than 1%). This might be due to multiple mapping reads which are sequences that map more than one time on the genome.

Although there are methods to solve this problem, in our case, multiple mapping reads represent less than 1% of all the reads so we can continue with our study without affecting downstream analyses.

However, if this difference were higher, and for instance, if these sequences contain some special features and they are mapped more than one time, when we continue with the downstream analysis, that feature could seem to be more abundant than it is in reality. Consequently, this could end in wrong conclusions.

**4.2.3 Are these reads and mappings appropriate to perform an analysis of Copy Number Variation? Explain why.**

Copy number variation (CNV) is a phenomenon in which sections of the genome are repeated and the number of repeats in the genome varies between individuals. As the information we are working with, comes from a transcription process we cannot determine if it comes from just one gene that is over-expressed or from a multiple reading of the same gene. So to conclude, it is not appropriate to perform an analysis of Copy Number Variation.

## 5 Genome Analysis (variant calling)

Using your mappings, you will carry out a variant calling analysis. Maybe some mutation is related to the sudden proliferation of these organisms...

### 5.1 Tasks

**5.1.1** To obtain the most statistical power of your data, you will use all your samples to perform variant calling. First, sort the mappings of each of your samples (tip: you should use ‘samtools sort’ for that).

As mentioned, I made use of samtools sort command. Consequently I obtained 4 sorted BAM files:

**Commands:**

```
» samtools sort bwa_h_1.bam > bwa_h_1_sorted.bam
» samtools sort bwa_h_2.bam > bwa_h_2_sorted.bam
» samtools sort bwa_n_2.bam > bwa_n_2_sorted.bam
» samtools sort bwa_n_1.bam > bwa_n_1_sorted.bam
```

**5.1.2** Then, you decide to merge the mappings from all your samples into a single BAM file (tip: use ‘samtools merge’ for that). Use the option to transfer the header information from one of your BAM files to the final BAM merged file (tip: check ‘samtools merge’ help, and look for the ‘-h FILE’ option).

In order to merge all the samples mappings into a single BAM file, I used samtools merge command with -h option for merging @SQ headers of input files into the specified header. Final result are two files called all\_hightemp.bam and all\_normaltemp.bam.

**Commands:**

```
» samtools merge -h bwa_h_1.bam -h bwa_h_2.bam all_hightemp.bam
bwa_h_1_sorted.bam bwa_h_2_sorted.bam
» samtools merge -h bwa_n_1.bam -h bwa_n_2.bam all_normaltemp.bam
bwa_n_1_sorted.bam bwa_n_2_sorted.bam
» ls -l
```

**Output:**

```
-rw-rw-r-. 1 sandra.apaz sandra.apaz 27300226 Jan 14 10:16 all_hightemp.bam
-rw-rw-r-. 1 sandra.apaz sandra.apaz 25463420 Jan 14 10:16 all_normaltemp.bam
-rw-rw-r-. 1 sandra.apaz sandra.apaz 18436587 Jan 14 10:04 bwa_h_1.bam
-rw-rw-r-. 1 sandra.apaz sandra.apaz 14359625 Jan 14 10:05 bwa_h_1_sorted.bam
-rw-rw-r-. 1 sandra.apaz sandra.apaz 18439892 Jan 14 10:04 bwa_h_2.bam
-rw-rw-r-. 1 sandra.apaz sandra.apaz 14365118 Jan 14 10:05 bwa_h_2_sorted.bam
-rw-rw-r-. 1 sandra.apaz sandra.apaz 16653530 Jan 14 10:05 bwa_n_1.bam
-rw-rw-r-. 1 sandra.apaz sandra.apaz 13381756 Jan 14 10:06 bwa_n_1_sorted.bam
-rw-rw-r-. 1 sandra.apaz sandra.apaz 16647332 Jan 14 10:05 bwa_n_2.bam
-rw-rw-r-. 1 sandra.apaz sandra.apaz 13368583 Jan 14 10:06 bwa_n_2_sorted.bam
```

**5.1.3** Next, perform the variant calling using ‘bcftools mpileup’ and ‘bcftools call’ with the merged BAM file. Check the parameters used in the practice lessons and apply those, except ‘-ploidy 1’ (which is not working as expected in the current version in the server).

First of all it is mandatory to generate an id for for the reference genome using samtools faidx. Then, I turned to bcftools environment and execute bcftools mpileup to transform the BAM file in VCF format. I used -f option for selecting the faidx-indexed reference file in the FASTA format. This process finish with the creation of two VCF files called all\_hightemp.vcf and all\_normaltemp.vcf.

Finally I executed bcftools call command for both VCF files. Including option -mv (multiallelic-caller, variants-only), -Ob (Output compressed BCF) and -o (output file)

**Commands:**

```
» samtools faidx ../genome.fasta
» conda activate bcftools
» bcftools mpileup -f ../genome.fasta all_hightemp.bam > all_hightemp.vcf
» bcftools mpileup -f ../genome.fasta all_normaltemp.bam > all_normaltemp.vcf
» bcftools call -mv -Ob -o call_h.bcf all_hightemp.vcf
» bcftools call -mv -Ob -o call_n.bcf all_normaltemp.vcf
```

**5.1.4** Finally, you decide to repeat the previous step, but for the 4 sorted BAM files separately, instead of merging them. For that, use ‘bcftools mpileup’ with the 4 BAM files as input in a single run, and then ‘bcftools call’ as usual.

For this performing this task I have used the same commands and options as in last section.

**Commands:**

```
» bcftools mpileup -f ../genome.fasta bwa_h_1_sorted.bam > bwa_h_1.vcf
» bcftools mpileup -f ../genome.fasta bwa_h_2_sorted.bam > bwa_h_2.vcf
» bcftools mpileup -f ../genome.fasta bwa_n_1_sorted.bam > bwa_n_1.vcf
» bcftools mpileup -f ../genome.fasta bwa_n_2_sorted.bam > bwa_n_2.vcf
```

**Commands:**

```
» bcftools call -mv -Ob -o call_h_1.bcf bwa_h_1.vcf
» bcftools call -mv -Ob -o call_h_2.bcf bwa_h_2.vcf
» bcftools call -mv -Ob -o call_n_1.bcf bwa_n_1.vcf
» bcftools call -mv -Ob -o call_n_2.bcf bwa_n_2.vcf
```

**Command:**

```
» ls -l
```

**Output:**

```
all_hightemp.bam bwa_h_1.bam bwa_h_2_sorted.bam bwa_n_1.vcf
call_h_1.bcf call_n_2.bcf all_hightemp.vcf bwa_h_1_sorted.bam
bwa_h_2.vcf bwa_n_2.bam call_h_2.bcf call_n.bcf all_normaltemp.bam
bwa_h_1.vcf bwa_n_1.bam bwa_n_2_sorted.bam call_h.bcf
all_normaltemp.vcf bwa_h_2.bam bwa_n_1_sorted.bam bwa_n_2.vcf
call_n_1.bcf
```

## 5.2 Questions

### 5.2.1 How many variants did you expect to obtain, if any? How many variants did you actually obtain? How many are SNPs, how many insertions and how many deletions?

I expected to obtain more variants in normal temperatures than in high temperatures because as this specie is not in optimal conditions, it will try to adapt in several ways to survive.

However as I show bellow, there are the same variants in both samples and all of them are SNPs.

Variants obtained in high temperatures:

**Command:** `» bcftools view -H call_h.bcf | wc -l`

**Output:** 3

Variants obtained in normal temperatures:

**Command:** `» bcftools view -H call_n.bcf | wc -l`

**Output:** 3

Number of INDELS in high temperatures:

**Command:** `» bcftools view -H call_h.bcf |cut -f8 |tr ";" " t" | cut -f1 |grep "INDEL"`

**Output:** 0

Number of INDELS in normal temperatures:

**Command:** `» bcftools view -H call_n.bcf |cut -f8 |tr ";" " t" | cut -f1 |grep "INDEL"`

**Output:** 0

Number of SNPs in high temperatures:

**Command:** `» bcftools stats call_h.bcf`

**Output:**

#	SN	id	key	value
	SN	0	number of samples:	1
	SN	0	number of records:	3
	SN	0	number of no-ALTs:	0
	SN	0	number of SNPs:	3
	SN	0	number of MNPs:	0
	SN	0	number of indels:	0
	SN	0	number of others:	0

Number of SNPs in normal temperatures:

**Command:** `» bcftools stats call_n.bcf`

**Output:**

#	SN	id	key	value
	SN	0	number of samples:	1
	SN	0	number of records:	3
	SN	0	number of no-ALTs:	0
	SN	0	number of SNPs:	3
	SN	0	number of MNPs:	0
	SN	0	number of indels:	0
	SN	0	number of others:	0

Therefore, there is not any insertion or deletions and there are 3 SNPs in each sample.

### 5.2.2 How many variants have quality greater than 100?

As we know, BCF files header has this format [16]:

```
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT
```

Therefore I sorted by the sixth column and obtain this:

```
Command: » bcftools view -H call_h.bcf | sort -k6,6gr | head
```

Output:

Aquifex_genome	1265109	.	C	A	187
Aquifex_genome	1265734	.	T	A	23.434
Aquifex_genome	1265735	.	T	A	23.434

```
Command: » bcftools view -H call_n.bcf | sort -k6,6gr | head
```

Output:

Aquifex_genome	1265109	.	C	A	222
Aquifex_genome	1265734	.	T	A	5.04598
Aquifex_genome	1265735	.	T	A	5.04598

So, there is only one variant over 100 in the high temperature sample (187) and other one in the normal temperature sample (222).

### 5.2.3 How many variants have depth of coverage greater than 100?

This information is included in the same file but in the eight column. As it contains extra data, firstly I have separated it by tabs and then look for the field we are looking for (DP). Then I used awk for introducing the condition needed (DP>100) and count all the lines which accomplished this condition (wc -l).

```
Command: » bcftools view -H call_h.bcf | cut -f8 | tr ';' ' ' | cut -f1 |  
sed -e 's/DP=//g' | awk 'if($1>=100)print $1' | wc -l
```

Output:1

```
Command: » bcftools view -H call_n.bcf | cut -f8 | tr ';' ' ' | cut -f1 |  
sed -e 's/DP=//g' | awk 'if($1>=100)print $1' | wc -l
```

Output:1

Concluding, there is just one variant with depth coverage over 100 in each sample.

### 5.2.4 Compare the output file you obtain with the merged BAM file (step 3) and with the 4 samples (step 4). What differences do you find? What advantages and disadvantages do you think each of those methods have?

```
Command: bcftools view -H call_h_1.bcf | cut -f1-8
```

Output:

Aquifex_genome	1265109	.	C	A	210	.	DP=249;
Aquifex_genome	1265734	.	T	A	36.4154	.	DP=11;
Aquifex_genome	1265735	.	T	A	36.4154	.	DP=8;

```
Command: bcftools view -H call_h_2.bcf | cut -f1-8
```

Output:

Aquifex_genome	1265109	.	C	A	183	.	DP=249;
Aquifex_genome	1265734	.	T	A	23.4340	.	DP=15;
Aquifex_genome	1265735	.	T	A	23.4340	.	DP=9;



Although variants are shared between the two samples from high temperature, they do not share quality nor depth value. This is why the study done with these two merge files has made an average of this data in order to approximate both values.

**Command:** `bcftools view -H call_n_1.bcf | cut -f1-8`

**Output:**

```
Aquifex_genome      1265109      .      C      A      222      .      DP=115;
```

**Command:** `bcftools view -H call_n_2.bcf | cut -f1-8`

**Output:**

```
Aquifex_genome      1265109      .      C      A      222      .      DP=81;
Aquifex_genome      1265734      .      T      A      5.04598      .      DP=3;
Aquifex_genome      1265735      .      T      A      5.04598      .      DP=2;
```

In this case I have obtained 1 variant of the first sample of normal temperatures and 3 from the second sample. Focusing on the quality of each variant, we can observe that for the first one (Aquifex\_genome 1265109) is the same. Therefore, its depth is the sum of the depth values of this variant from each file. On the other hand, the other two variants present on the step 4 file are only on the second sample of normal temperatures, therefore it is directly incorporated to the step 4 file.

Focusing on the advantages of using variant information from several samples would give us a more realistic view about the presence of the mutations shared across the population. ON the other hand, although using variant information from just one sample could end in an incomplete general study, this could give us more information about a specific variant in a specific population avoiding including extra information unneeded from other populations.

**5.2.5 Identify the variant with the best quality. Could this variant be affecting a gene? (tip: compare the position of the variant with the positions of the genes in the genome GFF file). Which gene did you find, if any? Give an example for how your variant could be affecting a gene (there is no need to actually check its effect).**

As we saw in section 5.2.2, quality value is placed in the sixth column of the bcf file. Therefore, I have ordered this column and show only the first value (the highest one).

**Command:** `> bcftools view -H call_h.bcf | sort -k6,6gr | cut -f2,4,5 | head -1`

**Output:** `1265109 C A`

**Command:** `> bcftools view -H call_n.bcf | sort -k6,6gr | cut -f2,4,5 | head -1`

**Output:** `1265109 C A`

The best quality variant is shared by both conditions, it is located in position 1265109 from the reference genome. After looking for the position in the genome.gff file, I found this match:

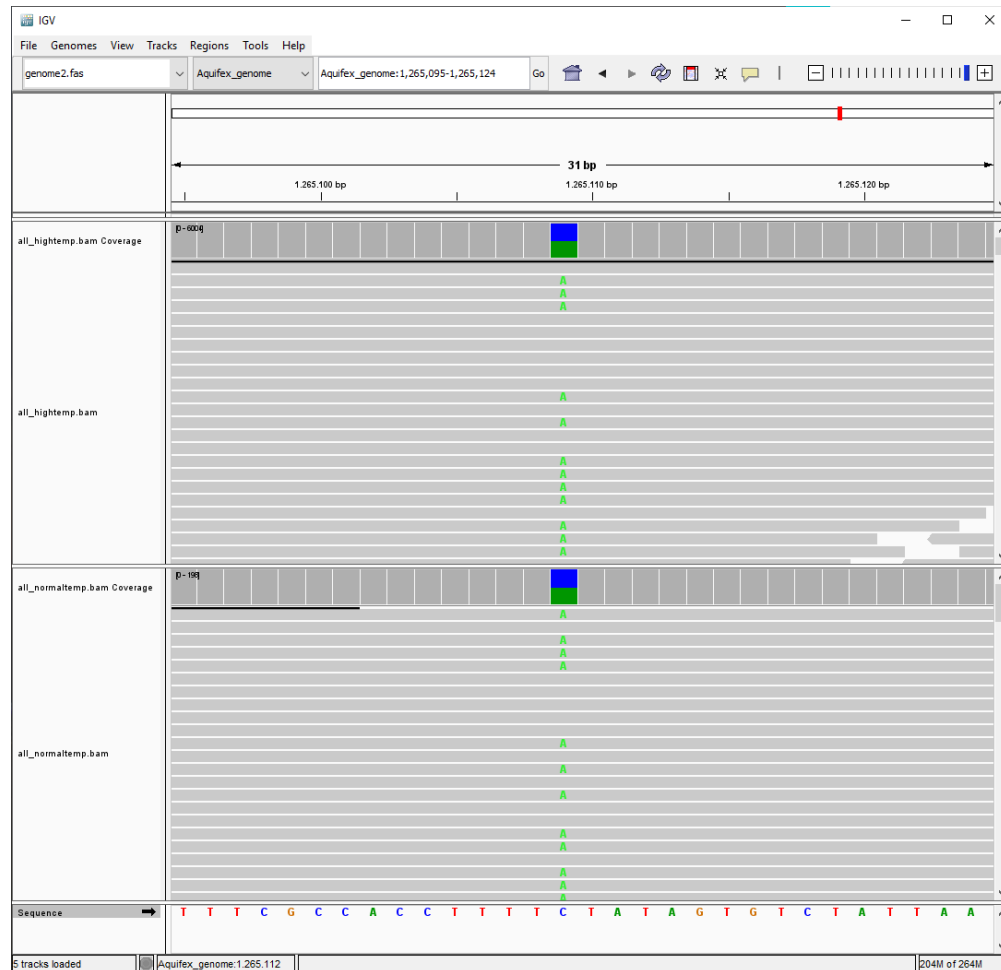
**Command:** `> awk -F "t" '$4<=1265109' genome.gff | awk -F "t" '$5>=1265109'`

**Output:**

```
Aquifex_genome Prodigal:002006 CDS 1264237 1265730 . - 0
ID=AQUIFEX_01423;Name=nifA;gene=nifA;inference=ab initio prediction:Prodigal:002006,
similar to AA sequence:UniProtKB:P03027;locus_tag=AQUIFEX_01423;product=Nif-specific
regulatory protein
```

About the gene affected, we can see that these position corresponds to a NifA family transcriptional regulator. As we are studying bacteria which prefer growing in high temperatures, this mutation could be related to the increase of the fixation rate or turning this process more efficient.

**5.2.6 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. Paste below the image you captured as an answer for this question.**



## 6 Differential expression analysis

Given that you have expression data for each sample, you can compare the expression differences between the samples grown under normal and high-temperature conditions, which could provide additional information about important genes involved.

### 6.1 Tasks

**6.1.1 For Differential Expression Analysis (DEA)** you need to start using the sorted bam files generated previously. First of all, you need to do a “read-count” using ‘htseq-count’. Some important parameters: you have to use ‘-i locus\_tag’ and ‘-t CDS’: (tip: you need to use a .gff file to count the reads). Once you have the four count files, it is necessary to merge all together and save it as ‘count.txt’ file (tip: Remember ‘join’ command). You should finally obtain something like this in your count.txt file (Remember to remove last lines with summary results from htseq count, \_\_no\_features, \_\_unambiguous, etc.):

Example:

Command:

```
» cat count.txt
```

Output: GeneID normal normal high high (The header it's not necessary)

```
AQUIFEX_00001 456 402 448 476
```

```
AQUIFEX_00002 154 174 152 124
```

```
AQUIFEX_00003 0 0 0 0
```

```
AQUIFEX_00004 134 136 136 158
```

```
AQUIFEX_00005 119 116 135 116
```

```
AQUIFEX_00006 0 0 0 0
```

Firstly, it is mandatory to create index files for all the BAM files:

Command:

```
» samtools index bwa_h_1_sorted.bam
```

```
» samtools index bwa_h_2_sorted.bam
```

```
» samtools index bwa_n_1_sorted.bam
```

```
» samtools index bwa_n_2_sorted.bam
```

Then I did the read count using htseq-count and the given parameters.

Command:

```
» htseq-count -i locus_tag -t CDS -f bam bwa_h_1_sorted.bam genome.gff > hightemp_count_1.txt
```

```
» htseq-count -i locus_tag -t CDS -f bam bwa_h_2_sorted.bam genome.gff > hightemp_count_2.txt
```

```
» htseq-count -i locus_tag -t CDS -f bam bwa_n_1_sorted.bam genome.gff > normaltemp_count_1.txt
```

```
» htseq-count -i locus_tag -t CDS -f bam bwa_n_2_sorted.bam genome.gff > normaltemp_count_2.txt
```

After creating the text file counts.txt, I wrote manually the headers. Then, I join all the counts files with join command:

```
Command: » join normaltemp_count_1.txt normaltemp_count_2.txt  
| join - hightemp_count_1.txt | join - hightemp_count_2.txt » counts.txt
```

**6.1.2** Now you can use your counts to perform the DEA analysis. For that use the Bioconductor package DESeq2, using these loading data parameters:

Command: `> cat DESeq2.R`  
`# Loading Data in R #`

```
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
```

Be sure you write properly the contrast analysis in DEF section :

**DEF section**

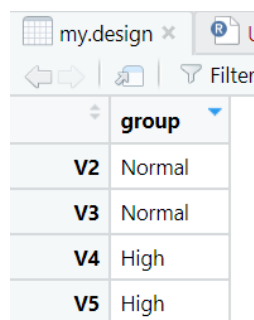
```
res <- results(dds, contrast=c("group","High","Normal"))
```

If you inspect in R the data frame of your experiment design (my.design variable) it has looks like this:

	group
V2	Normal
V3	Normal
V4	High
V5	High

**Important note:** `#rlogtranformation` for the PCA analysis section of the DESeq2.R script used in the practical session has to be silenced or you will have an error message!!.

The DESeq2 script can be consulted in [this](#) GitHub repository. Here I show the data frame of my experimental design.



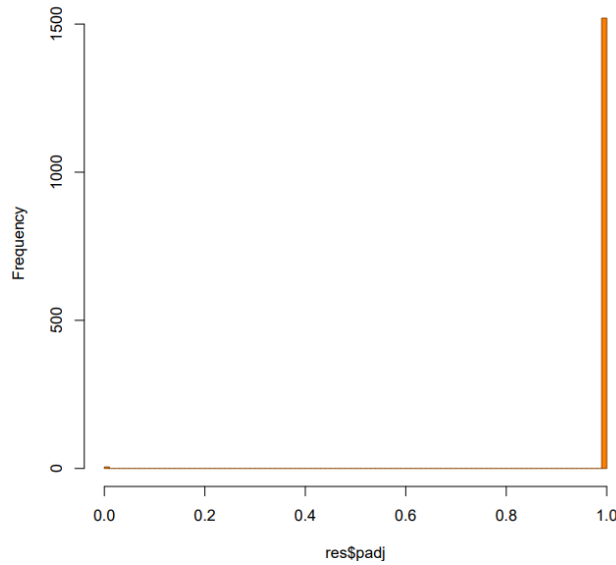
	group
V2	Normal
V3	Normal
V4	High
V5	High

## 6.2 Questions

**6.2.1** Have a look at the p-adj histogram obtained (`res$padj`), what does this result mean?

DESeq.R script determines which genes are significantly different. For each of these genes, it returns 3 parameters: the p-value, the adjusted p-value and the fold change. All of them can be consulted in the pdf stored in [this](#) GitHub repository.

In this section I will focuss in `padj`. The histogram for the adjusted p-value shows that most of the adjusted p-values are close to one. This means that most of the differences in expression are not significant. Only few of them are near to 0 value and therefore have significant expression levels between the samples.



### 6.2.2 How many genes showed a statistical ( $p\text{-adj} < 0.01$ ) differential expression?

After filtering the "res" DataFrame which can be consulted [here](#), I selected the ones with lowest p-adj. There are 5 genes with lower values of p-adj than 0.01:

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
AQUIFEX_01423	10026,82486	-5,065390163	0,045215702	-112,0272378	0	0
AQUIFEX_01759	3035,031837	-4,899035173	0,079631325	-61,52145766	0	0
AQUIFEX_01761	2790,984673	-5,093165425	0,087456047	-58,23685832	0	0
AQUIFEX_01760	159,3731739	10,76155543	1,451206195	7,415593641	1,21081E-13	4,61622E-11
AQUIFEX_01754	124,900585	10,40992754	1,453954366	7,159734708	8,08333E-13	2,46542E-10

Table 1: Genes which showed a statistical ( $p\text{-adj} < 0.01$ ) differential expression

### 6.2.3 Annotate these genes using the Name and product fields of the genome.gff file. Generate a table showing all the altered genes, their description, name, p-val, p-adj and fold change.

Using the command shown in section 5.2.5, I could annotate each name and product field on the table below. However for the last to genes I could not find its name and the product shown was "hypothetical protein". They might be unknown.

Gene	Name	product	log2FoldChange	pvalue	padj
AQUIFEX_01423	nifA	Nif-specific regulatory protein	-5,065390163	0	0
AQUIFEX_01759	nifB	FeMo cofactor biosynthesis protein NifB	-4,899035173	0	0
AQUIFEX_01761	nifH1	Nitrogenase iron protein 1	-5,093165425	0	0
AQUIFEX_01760		hypothetical protein	10,76155543	1,21081E-13	4,61622E-11
AQUIFEX_01754		hypothetical protein	10,40992754	8,08333E-13	2,46542E-10

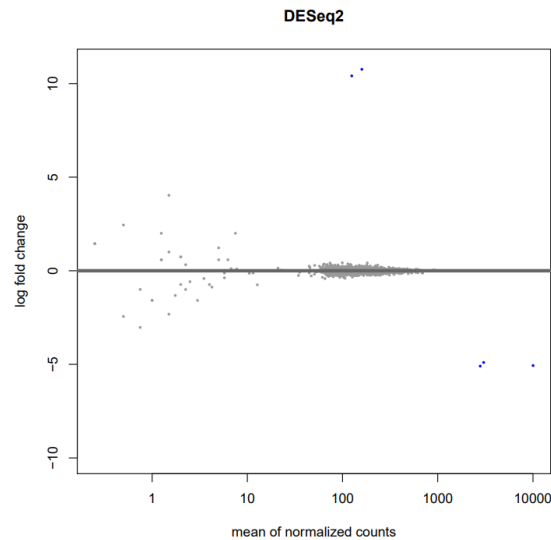
Table 2: Genes which showed a statistical ( $p\text{-adj} < 0.01$ ) differential expression and its annotation

#### 6.2.4 Have a look at the MA plot. What do the dots mean?, and the blue ones?. Which conclusions can you obtain from this figure?

In the MA plots we are able to see graphically how dots (genes) are located depending on the variation on their expression and the number of counts.

In this sense, grey dots are considered non significance due its low variation or small number of counts. They are all located in a triangle shape which has its base in the y-axis and its height is the horizontal line which crosses  $y = 0$ .

On the other hand, we can also observe 5 blue genes (dots) which are out of this triangle shape. These are the genes shown in section 6.2.3 and due its p-adj is lower than 0.01 are considered significatives.



#### 6.2.5 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?

Considering the threshold p-value I used (0.01) and the strong restrictions applied to the significance level, personally I trust that these 5 genes are expressed differently in samples from high temperature and in normal samples.

If we take a look at the expression of these genes in each sample we will be able to see in which conditions they are overexpressed.

Gene	NormalTemp1	NormalTemp2	HighTemp1	HighTemp2
AQUIFEX_01423	608	556	19478	19537
AQUIFEX_01759	184	210	5978	5790
AQUIFEX_01761	154	164	5472	5394
AQUIFEX_01760	340	298	0	0
AQUIFEX_01754	268	232	0	0

Table 3: Genes which showed a statistical ( $p\text{-adj} < 0.01$ ) differential expression and its expression values

In the light of the resulted table above, the relevant genes which may explain the bloom of bacteria in high temperatures are AQUIFEX\_01423, AQUIFEX\_01759 and AQUIFEX\_0176 because they have the highest values in the two last columns.

## 7 Functional analysis

The differential expression results gave you an idea about potentially important overexpressed genes. But, what are those genes doing? Why are they important?

### 7.1 Tasks

#### 7.1.1 Extract the sequence of each over expressed gene out from the assembled proteome (proteome.faa)

For this task I made use of grep command to find each gene in the proteome.faa file. I used the -A option to show the next lines (1 in my case) after the match was found.

```
Command: > cat proteome.faa | grep -A 1 'AQUIFEX_01423'
Output:  MERLKLKEINVVNEITKILTGDYTFEESLKEVLKVLVSYLGVHSFIAIREGNTLRIASSYGYFLNKDVAFKKG
EGITGKVFQRGIPLVIPNVKHNSAFANKTGIGRLLTEKHALIAAPIKVGGEVKGVITIFKEFSDESLENFYQTINVIGNLLG
MFFKLREKFESEKKAWEQKRELRELSEKYSLHGLIGSKVMRELIDTIEKVAKTDSVLILGESGTGKSILIARIHYESR
REKPFLTVNCASIPETLLESELFGYEKGAFGTAYTTKKGKFELANGGTIFLDEIGDMPLSLQAKLLRALQEREIERLGSEKPI
KVDVRIITATNKDLEKLKVEGKFREDLYYRINVVPLYVPPLRERKEDIPIILIEYYLKELNKKYGKEVYLSSEDALEVLLEYDYP
GNIRELINILERVVILNNGRVTSSSELPPELLRKREKRTGDLPKFIEETEEKRIIEALEKTGYVKSRAAKLLGYTLRQLDYRIKK
YGIELKKF
```

```
Command: > cat proteome.faa | grep -A 1 'AQUIFEX_01759'
Output:  MVIIMKEDKMSKFAHITSVHPCYNEKLHHKIGRVHLPVAPKCNIACRFCKRSIGDESVCEDRPGVS
RHIMKPAEVENYLNLDLEKSPNIKVAGIAGPGDSL FNKETFTLEVLKEKFPDLVRCLSTNGLLLPKYAERLAELGVKTVTVT
VNAVDPKIQAQIVDWIYYEGKVYRGEEGAELIKNQIEGVKKLAEFDVAVKINTVLIPEINMEHIVEIAKTFEKDAFVHNVIP
LIPMYKMENLRKPTCDEISNIRDTAEEYLHQFRACQQRADAAGLITEHKHLEGEKGGKLDIYDLKHFSH
```

```
Command: > cat proteome.faa | grep -A 1 'AQUIFEX_01761'
Output:  MVRKIAIYGKGGIGKSTTTQNTVAAMAHFHDKKVFIHGCDPKADSTRILHKGQQTMMDTLREKG
EDECTPDKVIEVGFGGVKCVESGGPEPGVGCAGRGVITAITLMEQHGVYEDDLDFVFFDVLGDVVCGGFAMPVRDGKADEIYVV
ASGEMMALYAANNICKGMVKYAEQSGVRLGGIICNSRNVGDGELDLLQEFCDKIGTQLIHVPRDNIVQKAEFQKKAVVDYDDTC
NQALEYKELARKIIENENLVIPTPMTMDELEELTSKYGFLD
```

```
Command: > cat proteome.faa | grep -A 1 'AQUIFEX_01760'
Output:  MDSNIDVEILSILSEASAPVGAKIIADSLKDRGYDIGERAVRYHLKVLNENSLTKKLGYSGREITEK
GIEELEKANISFRIGSVFSQVIEKLYLSDFPSKVLINTAKFEGDYKTIKEMVLRSEAGYSVGDYLNKKKGNTVSVETLCSIT
FDNFLKNGIIPTEPYGGIVKFEDYEPVNFEGVIDFKSSSIDPLVAFIMQGKTDVIGVINGEGLVPANFRVIPKSSEKQFENI
LKKDMLNSVLAYGTENVLGMNLNPEQIGVVLVGGTLPVPHESGYTADISAATQLKDISSMEKKTGKFLEAKKKKGKFKVTPV
LSKMLSKMQTINYDIEDKKGNVVNTAKIPIEYKEEAINALKDSYENKLAISDRLKVECDKFLNAYTICSLTVDGVFLKNKIP
VIPYGGILEVKADKKRFIEAIDYEGTSLDPHEVFFNKADGKNYILAGIRKVPMSASEKLIELNEKLGWNSIIIEIGRPNNDICG
VRVEKCMFGITTIGGTNPFANIRKNNIPVEMKTLHKSIDYSELTHYDDI
```

```
Command: > cat proteome.faa | grep -A 1 'AQUIFEX_01754'
Output:  MRNFLFLLLVISLPLFGGQRIMDKTLENGVKVIIKETKGRGLVSGVIFFGGVHGEERGETQLLFT
MLLKGSKNYPNASAVSYPFEKYGGYIYSSSEDDFSEIGFSTKVEGLKEGLKVIRDIIQNPLFKEEVLELEKRNQIVAIRSKRER
GMSYAYEELRTLTYKGTPYEYSSLGKDEDVERVSREDLIRRFNQIKKGENVVVLVGDFKAEDVPLLEEAFSDIPKGFELSS
VNKKIEKNEVKRVKREGTQATILCAFNAPPKSKDYFVFKVYNAVLGEGMTSKLFKVLREEKGAYATYSFYPTRYSSPRLFAY
VGTSPKKNALQDLIKVVSEGRVSEEDVELAKRKIIIGDFLLDHQTRLRQAWYLGFFEVMGLGWKMDEEYTKRISEVKREEVEE
AVRKYIDFHHCVVVEP
```

### 7.1.2 Save each sequence in FASTA format in an individual file and search for functional annotations in as many databases as you consider relevant (e.g. PFAM, PHMMER, eggNOG, KEGG, NCBI Blast, NCBI Taxonomy, STRING-DB, Uniprot, Ensembl, SMART, etc)

Using the command showed before, I introduced the output in an individual FASTA file. Then, I used NCBI Blast, eggNOG and PFAM for finding out their functionality.

#### Commands:

```
» cat ../proteome.faa | grep -A 1 'AQUIFEX_01754' > AQUIFEX_01754.faa
» cat ../proteome.faa | grep -A 1 'AQUIFEX_01423' > AQUIFEX_01423.faa
» cat ../proteome.faa | grep -A 1 'AQUIFEX_01759' > AQUIFEX_01759.faa
» cat ../proteome.faa | grep -A 1 'AQUIFEX_01761' > AQUIFEX_01761.faa
» cat ../proteome.faa | grep -A 1 'AQUIFEX_01760' > AQUIFEX_01760.faa
» ls -l
```

#### Output:

```
total 20
-rw-rw-r-. 1 sandra.apaz sandra.apaz 513 Jan 15 17:25 AQUIFEX_01423.faa
-rw-rw-r-. 1 sandra.apaz sandra.apaz 435 Jan 15 17:24 AQUIFEX_01754.faa
-rw-rw-r-. 1 sandra.apaz sandra.apaz 317 Jan 15 17:25 AQUIFEX_01759.faa
-rw-rw-r-. 1 sandra.apaz sandra.apaz 552 Jan 15 17:26 AQUIFEX_01760.faa
-rw-rw-r-. 1 sandra.apaz sandra.apaz 291 Jan 15 17:26 AQUIFEX_01761.faa
```

### 7.1.3 Do overexpressed genes have any known molecular/cellular function? Report if so, and briefly describe how you inferred those functions (i.e. functional information sources).

As far as I consider, not all the differentially expressed proteins are of our interest. As we saw in last section there were two differentiate groups: the first one contains AQUIFEX\_01423, AQUIFEX\_01759 and AQUIFEX\_01761 where the ones overexpressed in high temperatures and the second one which contains AQUIFEX\_01760 and AQUIFEX\_01754 which were not expressed at all in high temperatures. In this sense, the first group is highly interesting for the current study because the relative abundance of Aquifex in high temperature samples is considerable and therefore, the proteins which are overexpressed in these condition might be responsible for the rise of survival in this conditions.

However, in this section I will come up with a deep search of the first group of proteins and additionally, a superficial search of the second group just in case they could provide some additional information for the study.

As we have the FASTA file of all of them. I performed a BLAST search of each file and I obtained the best hits (all over 97%) from which I could infer the protein from which the sequence belongs, and therefore, know the molecular function by searching it in UniProt and NCBI. Carrying out this process I came up with these information:

- **AQUIFEX\_01423:** After performing the BLAST search, I obtained a hit whose identity was 100% which means a perfect match. The organism to which this gene belongs is Aquifex aeolicus (strain VF5) which is the organism I am studying. As a result of this process I obtained the gene name (ntrC2) and afterwards I looked for it in NCBI where I found its id (1193468) and description [17].



The gene name is ntrC2 [18] which is the transcriptional activator for nitrogen-regulated promoters and, as a response regulator, belongs to the protein family of two-component systems. The activity of all response regulators is modulated by phosphorylation of the conserved N-terminal receiver domain. Phosphorylation of the dimeric NtrC has two consequences [19]:

1. A strong increase in the cooperative binding of NtrC to two adjacent binding sites and
2. Activation of NtrC as an ATPase

- **AQUIFEX\_01759:** Again, I performed a BLAST search where I obtained A 97.3% identity hit. The organism to which this gene belongs is *Methanococcus maripaludis* (strain S2 / LL). Although it is not the organism we are studying, the identity value is quite high so we can assume that AQUIFEX\_01759 can have the same functionality. Once I obtained the gene name (MMP0658), I looked for it in NCBI where I found its id (2761676) and description [20].

The gene name is MMP0658 which encodes a FeMo cofactor biosynthesis protein NifB. It is Involved in the biosynthesis of the iron-molybdenum cofactor (FeMo-co or M-cluster) found in the dinitrogenase enzyme of the nitrogenase complex in nitrogen-fixing microorganisms. NifB catalyzes the crucial step of radical SAM-dependent carbide insertion that occurs concomitant with the insertion of a 9th sulfur and the rearrangement/coupling of two [4Fe-4S] clusters into a [8Fe-9S-C] cluster, the precursor to the M-cluster.[21]

- **AQUIFEX\_01761:** Firstly, I performed a BLAST search where I obtained a match hit with a 100% identity. The organism to which this gene belongs is *Methanococcus maripaludis* (*Methanococcus deltae*). Although it is not the organism we are studying, the identity value is the highest one so we can assume that AQUIFEX\_01759 can have the same functionality. Once I obtained the gene name, I looked for it in NCBI where I found its id (5327360) and description [22].

The gene name is nifH which encodes a Nitrogenase iron protein [23]. The nitrogenase enzyme system catalyzes the ATP (adenosine triphosphate)-dependent reduction of dinitrogen to ammonia during the process of nitrogen fixation. Nitrogenase consists of two proteins: the iron (Fe)-protein, which couples hydrolysis of ATP to electron transfer, and the molybdenum-iron (MoFe)-protein, which contains the dinitrogen binding site [24].

- **AQUIFEX\_01760:** After performing a BLAST search, I obtained a 99.8% hit. The organism to which this gene belongs is *Methanococcus maripaludis* (strain S2 / LL) and the gene name is nrpR which encodes a Global nitrogen regulator NrpR [25]. It is a transcriptional repressor of nitrogen fixation and assimilation genes. Binds to two tandem operators in the glnA and nif promoters, thereby blocking transcription of the genes. Under nitrogen limitation, binding of the intracellular nitrogen metabolite 2-oxoglutarate to NrpR decreases the binding affinity of NrpR to DNA, leading to initiation of transcription.
- **AQUIFEX\_01754:** Results of BALST searc showed a match hit hith a 100% identity. The organism to which this gene belongs is *Aquifex aeolicus* (strain VF5). It is ymxG which encodes processing protease. Its main molecular functions are metal ion binding and peptidase activity. [26]

#### **7.1.4 Do the overexpressed genes have any known domain? Which ones?**

We can find information about the protein domains in UniProt links included in section 7.1.3. All the proteins mentioned have known domains, and below we can see some of them:

AQUIFEX\_01423: Sigma-54 factor interaction. [18]

AQUIFEX\_01759: Radical SAM core.[21]

AQUIFEX\_01761: Belongs to the NifH/BchL/ChlL family.[23]

AQUIFEX\_01760: NRD 1 and NRD 2.[25]

AQUIFEX\_01754: Peptidase\_M16 and Peptidase\_M16\_C. [26]

#### **7.1.5 Are the overexpressed genes functionally related with each other? (i.e. protein interactions). Describe your sources.**

In order to look for interaction between the first group proteins (AQUIFEX\_01423, AQUIFEX\_01759 and AQUIFEX\_01761) I looked for them in STRING and BioGRID databases. However, I could not find relevant annotated protein-protein interactions in Aquifex aeolicus organism. However, this does not mean that this indication does not exist.

Consequently, I searched through scientific papers and found that their domains are quite related. In this sense, several authors claimed that the Sigma-54 domain controls the expression of NifB [27][28] and NifH [29][30] genes.

#### **7.1.6 Could those genes and functions be related to the bloom of algae observed in the hot spring after the high-temperature episodes? Why? Briefly elaborate your hypothesis**

Most algae require for their growth water, light, CO<sub>2</sub>, and mineral salts, among which are mainly some source of nitrogen such as nitrate or ammonium and a source of phosphorus that is usually some inorganic phosphate [31]. In this sense, thanks to overexpression of nitrogen fixation genes of Aquifex, nitrogen will increase its fixation in the environment. Although this high temperature does not allow algae to grow, when temperatures low there will be a higher concentration of nitrogen in the environment that will be used by algae to grow.

## 8 Phylogenetic analysis

The functional analysis of the overexpressed genes gave you an idea about the biological processes happening in the hot spring during the high temperature episodes. Intrigued by this unusual biological phenomenon, you decide to refine the functional inferences and investigate the evolutionary origin of the overexpressed genes in the isolated genome.

To do so, you decide to perform an in depth phylogenetic analysis comparing each overexpressed gene against their homologs in other prokaryotic genomes.

Your reference set of organisms includes the **public** genome of the same species you isolated, and 6 other bacteria and archaea that are known to be related to the biological processes you identified previously:

CLOPA - *Clostridium pasteurianum*  
9AQUI - *Hydrogenivirga caldilitoris*  
METV3 - *Methanococcus voltae*  
NOSS1 - *Nostoc* sp.  
AQUAE - *Aquifex aeolicus* (strain VF5)  
METMP - *Methanococcus maripaludis*  
RHOCB - *Rhodobacter capsulatus*

You have already downloaded the complete proteome of all 7 species from UNIPROT and put them together into a FASTA file (`all_reference_proteomes.faa`), which is in your server.

### 8.1 Tasks

For each over expressed gene, perform a standard phylogenetic workflow:

#### 8.1.1 Run a blast search for each over expressed protein against all reference proteomes. Extract hits with e-value $\leq 0.001$ (tip: you can use blast parameters for this)

Firstly, I created a new folder for saving future output files inside. Then, I made a blast database containing the proteomes of the 7 species (`all_reference_proteomes.faa`) and, as I did not include -out option, the resulting database will have the same name.

**Commands:**

```
» mkdir phylo
» cd phylo

» makeblastdb -dbtype prot -in ../all_reference_proteomes.faa
» cat ../proteome.faa | grep -A 1 'AQUIFEX_01423' > AQUIFEX_01423.faa
» cat ../proteome.faa | grep -A 1 'AQUIFEX_01759' > AQUIFEX_01759.faa
» cat ../proteome.faa | grep -A 1 'AQUIFEX_01761' > AQUIFEX_01761.faa
» cat ../proteome.faa | grep -A 1 'AQUIFEX_01760' > AQUIFEX_01760.faa
» cat ../proteome.faa | grep -A 1 'AQUIFEX_01754' > AQUIFEX_01754.faa
```

Secondly, I used `blastp` command to search for all the homologs of each sequence file. I use an evalue threshold of 0.001. This is a parameter that describes the number of hits one can "expect" to see by chance when searching a database of a particular size [32]. Finally I saved the result in a txt file.

**Commands:**

```
» blastp -task blastp -query AQUIFEX_01423.faa -db ../all_reference_proteomes.faa
-outfmt 6 -evaluate 0.001 > AQUIFEX_01423.txt
» blastp -task blastp -query AQUIFEX_01759.faa -db ../all_reference_proteomes.faa
-outfmt 6 -evaluate 0.001 > AQUIFEX_01759.txt
» blastp -task blastp -query AQUIFEX_01761.faa -db ../all_reference_proteomes.faa
-outfmt 6 -evaluate 0.001 > AQUIFEX_01761.txt
» blastp -task blastp -query AQUIFEX_01760.faa -db ../all_reference_proteomes.faa
-outfmt 6 -evaluate 0.001 > AQUIFEX_01760.txt
» blastp -task blastp -query AQUIFEX_01754.faa -db ../all_reference_proteomes.faa
-outfmt 6 -evaluate 0.001 > AQUIFEX_01754.txt
```

### 8.1.2 Create a FASTA file with all the sequences of selected hits (Tip1: you can use the `extract_sequences_from_blast_result.py` used in our practical exercises, Tip2: Include the query protein in the same FASTA file!)

For this task I executed the python script `extract_sequences_from_blast_result.py` which can be consulted in [this](#) GitHub repository.

Results are different FASTA files which contains the sequences of the closest homolog of each initial sequence.

**Commands:**

```
» python ../extract_sequences_from_blast_result.py AQUIFEX_01423.txt
../all_reference_proteomes.faa » AQUIFEX_01423.faa
» python ../extract_sequences_from_blast_result.py AQUIFEX_01759.txt
../all_reference_proteomes.faa » AQUIFEX_01759.faa
» python ../extract_sequences_from_blast_result.py AQUIFEX_01761.txt
../all_reference_proteomes.faa » AQUIFEX_01761.faa
» python ../extract_sequences_from_blast_result.py AQUIFEX_01760.txt
../all_reference_proteomes.faa » AQUIFEX_01760.faa
» python ../extract_sequences_from_blast_result.py AQUIFEX_01754.txt
../all_reference_proteomes.faa » AQUIFEX_01754.faa
```

### 8.1.3 Build a phylogenetic tree out of the FASTA file (suggested tools: MAFFT, iqtree) (Tip: for iqtree, you can run a fast workflow with `-m LG` and `-fast`)

Before inferring a phylogenetic tree, homologous sequences need to be aligned. There are multiple programs to do it: ClustalOmega, MAFFT, MUSLE, etc. Here, we will use MAFFT, which has a very simple command line.

**Command:**

```
» mafft AQUIFEX_01423.faa > AQUIFEX_01423.alg
» mafft AQUIFEX_01759.faa > AQUIFEX_01759.alg
» mafft AQUIFEX_01761.faa > AQUIFEX_01761.alg
» mafft AQUIFEX_01760.faa > AQUIFEX_01760.alg
» mafft AQUIFEX_01754.faa > AQUIFEX_01754.alg
```

Command:

#### 8.1.4 Visualize the result (suggested tools: [etetoolkit.org/treeview](http://etetoolkit.org/treeview), `ete3`)

```
--AQUIFEX_01759
+-----tr|A0A1D9N142|A0A1D9N142_CLOPA
+-----sp|Q8YQ66|MOAA_NOSS1
+-----tr|A0A497XQ08|A0A497XQ08_9AQUI
+-----sp|Q67929|MOAA_AQUAE
+-----tr|D7DS11|D7DS11_METV3
+-----sp|Q6LZ03|MOAA_METHP
+-----tr|A0A1D9NZX3|A0A1D9NZX3_CLOPA
+-----sp|P20627|NIFB_NOSS1
+-----tr|D5ANH7|D5ANH7_RHOCB
+-----tr|Q632L8|Q632L8_CLOPA
+-----tr|A0A1D9N710|A0A1D9N710_CLOPA
+-----tr|D7DS17|D7DS17_METV3
+-----tr|Q6LZH0|Q6LZH0_METHP
```

```

**AQUIFEX_01423
+-----tr|A0A1D9N7N0|A0A1D9N7N0_CLOPA
+-----tr|O66581|O66501_AQUAE
+-----tr|A0A1D9N8L1|A0A1D9N8L1_CLOPA
+-----tr|A0A1D9M7Z9|A0A1D9M7Z9_CLOPA
+-----tr|A0A1D9N929|A0A1D9N929_CLOPA
+-----tr|A0A1D9N5D3|A0A1D9N5D3_CLOPA
+-----tr|A0A1D9N6U0|A0A1D9N6U0_CLOPA
+-----tr|A0A1D9N7Z8|A0A1D9N7Z8_CLOPA
+-----tr|Q8YR90|Q8YR90_NOS51
+-----tr|D5AVB4|D5AVB4_RHOCB
+-----sp|P09432|NTRC_RHOCB
+-----tr|D5ANR6|D5ANR6_RHOCB
+-----tr|D5AS32|D5AS32_RHOCB
+-----tr|D5AMA7|D5AMA7_RHOCB
+-----tr|D5APH1|D5APH1_RHOCB
+-----tr|D5AU6|D5AU6_RHOCB
+-----tr|A0A497XPD5|A0A497XPD5_9AQUI
+-----tr|O67198|O67198_AQUAE
+-----tr|A0A497XW11|A0A497XW11_9AQUI
+-----tr|O66551|O66551_AQUAE
+-----tr|A0A497XSL5|A0A497XSL5_9AQUI
+-----tr|O66596|O66596_AQUAE
+-----tr|A0A497XQV2|A0A497XQV2_9AQUI
+-----tr|A0A1D9N3I3|A0A1D9N3I3_CLOPA
+-----tr|A0A1D9N167|A0A1D9N167_CLOPA
+-----tr|A0A497XPP3|A0A497XPP3_9AQUI
+-----tr|O66502|O66502_AQUAE
+-----tr|D5AN35|D5AN35_RHOCB
+-----tr|A0A497XQD2|A0A497XQD2_9AQUI
+-----tr|O66591|O66591_AQUAE
+-----sp|D5ANR9|N1FA_RHOCB
+-----tr|D5ANH8|D5ANH8_RHOCB
+-----tr|A0A497XW95|A0A497XW95_9AQUI
***tr|O67661|O67661_AQUAE

```

29

## 8.2 Questions

(Answer the questions referring to each over expressed gene/protein)

### 8.2.1 What is the closest ortholog of each overexpressed gene from a phylogenetic point of view? From what species?

In order to visualize every tree, I used the online tool [itol](#) I could visualize all the trees by introducing them in *newick* format. Figures obtained can be seen in section 8.2.6.

Firstly, it is mandatory to know the definition of ortholog genes. They are genes of different species that have evolved through speciation events only. In this sense, by studying above trees I found the closest orthologs for each overexpressed gene (highlighted in green color).

For AQUIFEX\_01423 the closest ortholog is A0A497XW95 9AQUI .

For AQUIFEX\_01759 the closest ortholog is Q6LZH0 METMP.

For AQUIFEX\_01761 the closest ortholog is NIFH METMP.

For AQUIFEX\_01760 the closest ortholog is NRPR METMP.

For AQUIFEX\_01754 the closest ortholog is A0A497XMY7 9AQUI.

Most of the closest ortholog of the overexpressed genes are from the specie *Methanococcus maripaludis* (Mnemonic METMP). It is a specie of methanogen. It is anaerobic, weakly motile, non-spore-forming, Gram-negative, and a pleomorphic coccoid-rod averaging 1.2 by 1.6  $\mu$ m is size.

On the other hand, for AQUIFEX\_01423 and AQUIFEX\_01754 closest orthologs are from the same specie, *Hydrogenivirga caldilitoris* (Mnemonic 9AQUI) which is a species with a mean length of 2  $\mu$ m and width of approximately 0.3  $\mu$ m. It is important to highlight that the temperature range for growth is 55–77.5°C (optimum 75°C) [33].

### 8.2.2 Do orthology assignments support your previous functional annotations? Briefly describe why. (Tip: the sequence IDs of orthologs are in Uniprot format, you can search for their functional information online)

For this task I have searched the orthologs in the all\_reference\_Proteomes.faa using cat and grep commands.

```
Command: > cat all_reference_proteomes.faa | grep "|A0A497XW95_9AQUI"
```

```
Output: >tr|A0A497XW95|A0A497XW95_9AQUI Nif-specific regulatory protein
OS=Hydrogenivirga caldilitoris OX=246264 GN=BCF55_1332 PE=4 SV=1
```

```
Command: > cat all_reference_proteomes.faa | grep "Q6LZH0_METMP"
```

```
Output: >tr|Q6LZH0|Q6LZH0_METMP FeMo cofactor biosynthesis protein NifB
OS=Methanococcus maripaludis (strain S2 / LL) OX=267377 GN=MMP0658 PE=3 SV=1
```

```
Command: > cat all_reference_proteomes.faa | grep "NIFH_METMP"
```

```
Output: >sp|POCW57|NIFH_METMP Nitrogenase iron protein
OS=Methanococcus maripaludis (strain S2 / LL) OX=267377 GN=nifH PE=3 SV=1
```

```
Command: > cat all_reference_proteomes.faa | grep "NRPR_METMP"
```

```
Output: >sp|Q6LZL7|NRPR_METMP Global nitrogen regulator NrpR
OS=Methanococcus maripaludis (strain S2 / LL) OX=267377 GN=nrpR PE=1 SV=1
```

```
Command: > cat all_reference_proteomes.faa | grep "A0A497XMY7_9AQUI"
```

```
Output: >tr|A0A497XMY7|A0A497XMY7_9AQUI Putative Zn-dependent peptidase
OS=Hydrogenivirga caldilitoris OX=246264 GN=BCF55_0553 PE=4 SV=1
```

All the orthologs from overexpressed genes in hightemperatures (AQUIFEX\_01423, AQUIFEX\_01759 and AQUIFEX\_01761) supports their functional annotations shown in section 6.2.3.

For example, gene A0A497XW95\_9AQUI which contains A0A497XW95 protein has a Nif-specific regulatory function as it was proposed in section 6.2.3 which contains nifA, a protein with an Nif-specific regulatory function.

On the other hand, for overexpressed genes in normal temperatures I could not find out its function and I inferred them by searching it orthologs. Therefore with this process I have obtained the same results. However, taking into account the good result obtained with the previous overexpressed genes (hight temperatures) I highly trust that their function should be the similar as for its orthologs.

### **8.2.3 Are all the overexpressed genes present in the public genome of your organism? (Tip: remember that you isolated and sequenced the novo your organisms. The genome you got is not necessarily identical to the public genome in databases)**

According to NCBI genomes, only the strain VF5 of Aquifex aeolicus has been completely sequenced.

By performing a BLAST I obtained a 100% hit with Aquifex aeolicus (strain VF5) for overexpressed genes AQUIFEX\_01423 and AQUIFEX\_01754. Therefore, they are contained in the public genome.

However, the remaining overexpressed genes did not hit with Aquifex aeolicus species and therefore they are not contained in the public genome.

### **8.2.4 What's the most probable origin for each overexpressed gene? Explain why (for each gene).**

After consulting the principal characteristics of the main species contained in the phylogenetic tree of each gene I came up with this conclusions:

- AQUIFEX\_01423 phylogenetic tree contains Clostridium pasteurianum, Rhodobacter capsulatus and Hydrogenivirga caldilitoris. Although all of them are bacteria whose main function is metabolizing environment substances, their growth environments are quite different.

For the first one which could fix free nitrogen from the air, would need a non-water environment. For the second one which can obtain energy through photosynthesis would prefer aqueous environments such as those around natural water sources or in sewage. Finally, Hydrogenivirga caldilitoris need water environments for growing.

In this sense, the common ancestor could be a metabolic bacteria which lives near an aqueous environment and after several years started to diverge into different species which can live in these different environments.

- AQUIFEX\_01761 phylogenetic tree divides plenty of species from one concrete gene (Q6LY48), from the species Methanococcus maripaludis. After consulting in NBCI Database, I found that it is a global nitrogen fixer [34]. Therefore, as they share this function, a possible common ancestor could be a microorganism with this specific function.
- AQUIFEX\_01760 phylogenetic tree only contains 2 different species, Methanococcus voltae and Methanococcus maripaludis. Focusing on the last one, our gene only diverged from it in a short period of time before. In this sense we need to search for similarities between

Methanococcus voltae and our gene. Consequently, the common ancestor of these genes could be a hyperthermophilic microorganism, as all shared this feature.

- AQUIFEX\_01754 phylogenetic tree mainly contains Nostoc sp which is Bacteria that fix atmospheric nitrogen and sequester carbon in soils [35]. In this sense they share one of their principal functions. Therefore I would suppose that the common ancestor is a Bacteria which can fix atmospheric nitrogen.

### 8.2.5 Is there any relationship or interesting finding between those genes and the microbial communities you profiled in the metagenomics work package?

As explained in section 2.1.2, the most abundant organism in normal temperature environments was Methanococcus maripaludis whose relative abundance was 4.8% and in high temperature environment, was the second most abundant organism. This means that this species shares with Aquifex aeolicus in some way the climate likely characteristics to grow. This is why most of the orthologs from our genes are from Methanococcus maripaludis species. Consequently, is the most representative species in the phylogenetic trees of the overexpressed genes of Aquifex aeolicus

### 8.2.6 Include a screenshot of all the trees you obtained (using iTOL, etetoolkit.com/treeview, or any other graphical software).

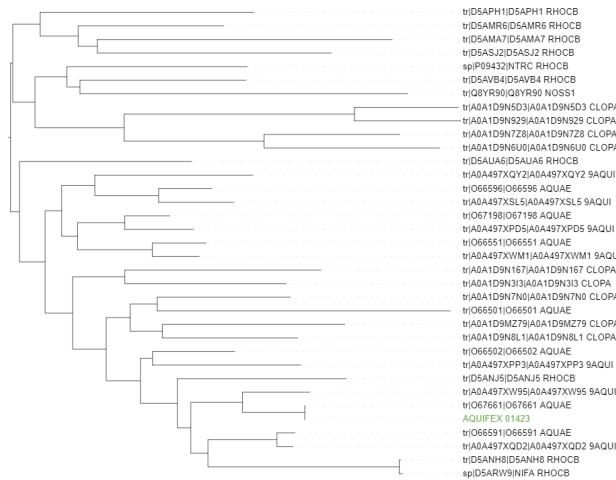


Figure 3: Tree obtained from AQUIFEX\_01423 using itol tool

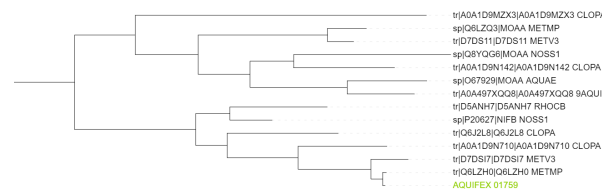


Figure 4: Tree obtained from AQUIFEX\_01759 using itol tool



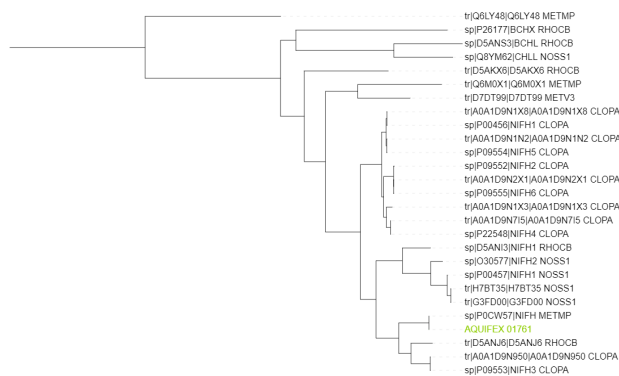


Figure 5: Tree obtained from AQUIFEX\_01761 using itol tool

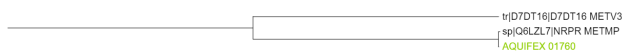


Figure 6: Tree obtained from AQUIFEX\_01760 using itol tool

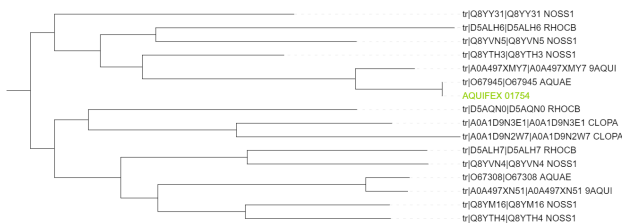


Figure 7: Tree obtained from AQUIFEX\_01754 using itol tool

## 9 Conclusion

My hypothesis for the effect observed in the hot spring after the high-temperature event is the following one:

At the light of the metagenomics analysis results, we could observe how the diversity of the environment changes as temperature does. When temperature rises it increases survival opportunities of *Aquifex aeolicus* which perfectly grows at high temperature and small oxygen concentrations. This is why we found 84% of relative abundance. Thanks to the likely environment conditions and to its chemoautotrophic metabolism, during high temperatures conditions it can easily fix nutrients. However, when temperature becomes lower, *Aquifex aeolicus* is no longer under its optimal conditions for growth, and its abundance is highly reduced. Consequently, nutrients left behind by *Aquifex aeolicus* could be used by other kinds of bacteria which prefer lower temperatures to grow. In this sense the biodiversity of the ecosystem increases. A good example is *Methanococcus maripaludis* which is the most abundant and has around 4% of relative abundance.

Continuing with the workflow and after getting familiar with the read characteristics, I made a variant calling analysis from which I obtained a particular mutation with a relevant higher quality than all others. This mutation was placed in the 1265109 position and corresponds to a NifA family transcriptional regulator. This finding made me think that probably, as we are studying bacteria which prefer growing in high temperatures, this mutation could be increasing the fixation rate or turning this process more efficient in high temperatures environment.

The differential expression analysis showed that 5 genes were overexpressed ( $p\text{-adj} < 0.01$ ) and 3 of them were highly overexpressed in high temperatures environment. Finally, this analysis confirmed that the most overexpressed in high temperatures was AQUIFEX\_01423 gene which is in charge of nitrogen fixation.

This research was later on confirmed by the phylogenetic analysis that showed that the closest orthologs are ones with the same inferred function as our protein sequences. Most part of these orthologs were from the *Methanococcus maripaludis* species (which was found in the initial samples of normal and high temperature).

To sum up, the effect observed in the hot spring is caused by the nitrogen fixation efficiency of this *Aquifex aeolicus*. When it finds optimal conditions (high temperature) it reaches the best efficiency. Otherwise, when temperature gets a normal value and therefore is no longer an optimal environment for *Aquifex aeolicus*, it stops growing and other bacteria take its place. However, normal temperatures and nitrogen left by *Aquifex aeolicus*, which is one of the main mineral salts needed by algae [31], creates a suitable environment for algae to grow.

## References

- [1] Seven Bridges. *Taxonomic Profiling of Metagenomics Samples*. URL: <https://www.sevenbridges.com/taxonomic-profiling-of-metagenomics-samples/>.
- [2] AlessioMilanese unode LucasPaoli hjruscheweyh and SuShiAtGit. *mOTUS tool*. URL: <https://github.com/motu-tool/mOTUs>.
- [3] Britannica. *mrelative abundance*. URL: <https://www.britannica.com/science/relative-abundance>.
- [4] Gerard Deckert et al. "The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*". In: *Nature* 392.6674 (1998), pp. 353–358.
- [5] Ronald V Swanson. "Genome of *Aquifex aeolicus*". In: *Hyperthermophilic Enzymes Part A*. Vol. 330. Methods in Enzymology. Academic Press, 2001, pp. 158–169.
- [6] NCBI. *Aquifex aeolicus* VF5. URL: <https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi>.
- [7] Marianne Guiral et al. "Chapter Four - The Hyperthermophilic Bacterium *Aquifex aeolicus*: From Respiratory Pathways to Extremely Resistant Enzymes and Biotechnological Applications". In: *Advances in Bacterial Respiratory Physiology*. Ed. by Robert K. Poole. Vol. 61. Advances in Microbial Physiology. Academic Press, 2012, pp. 125–194.
- [8] Marianne Guiral et al. "Hyperthermostable and oxygen resistant hydrogenases from a hyperthermophilic bacterium *Aquifex aeolicus*: Physicochemical properties". In: *International Journal of Hydrogen Energy* 31.11 (2006). IHEC 2005 and COST Action 841 Final Meeting, pp. 1424–1431. ISSN: 0360-3199.
- [9] Sam Maddy Ph.D: Psycholoanalysis Human Behavior. *At what temperature are most bacteria and micro organisms dead?* URL: <https://www.quora.com/At-what-temperature-are-most-bacteria-and-micro-organisms-dead>.
- [10] Wikipedia. *ARN ribosomal 16S definition*. URL: [https://es.wikipedia.org/wiki/ARN\\_ribosomal\\_16S](https://es.wikipedia.org/wiki/ARN_ribosomal_16S).
- [11] GenoMax. *Difference between mate-pair, paired-end and long read*. URL: <https://www.biostars.org/p/467845/#467852>.
- [12] Inc Illumina. *Q Score Definition*. URL: <https://emea.illumina.com/science/technology/next-generation-sequencing/plan-experiments/quality-scores.html>.
- [13] Base Space Online Help. *Quality Score Encoding*. URL: [https://support.illumina.com/help/BaseSpace\\_OLH\\_009008/Content/Source/Informatics/BS/QualityScoreEncoding\\_swBS.htm](https://support.illumina.com/help/BaseSpace_OLH_009008/Content/Source/Informatics/BS/QualityScoreEncoding_swBS.htm).
- [14] Samtools Manual Page. *samtools view – views and converts SAM/BAM/CRAM files*. URL: <http://www.htslib.org/doc/samtools-view.html>.
- [15] Metagenomics wiki. *Number of reads in bam file*. URL: <https://www.metagenomics.wiki/tools/samtools/number-of-reads-in-bam-file>.
- [16] EMBL-EBI. *Understanding VCF format*. URL: <https://www.ebi.ac.uk/training/online/courses/human-genetic-variation-introduction/variant-identification-and-analysis/understanding-vcf-format/#:~:text=VCF%5C%20is%5C%20the%5C%20standard%5C%20file,for%5C%20variation%5C%20archives%5C%20like%5C%20EVA..>
- [17] NCBI. *ntnC2 NtrC family transcriptional regulator [ Aquifex aeolicus VF5 ]*. URL: <https://www.ncbi.nlm.nih.gov/gene/1193468>.
- [18] Uniprot Database. *Transcriptional regulator (NtrC family)*. URL: <https://www.uniprot.org/uniprot/O67661>.

- [19] Inga Mettke, Ulrike Fiedler, and Verena Weiss. “Mechanism of activation of a response regulator: interaction of NtrC-P dimers induces ATPase activity”. In: *Journal of bacteriology* 177.17 (1995), pp. 5056–5061.
- [20] NCBI. *MMP<sub>R</sub>S03455radicalSAMprotein[MethanococcusmaripaludisS2]*. URL: <https://www.ncbi.nlm.nih.gov/gene/?term=MMP0658>.
- [21] Uniprot Database. *FeMo cofactor biosynthesis protein NifB*. URL: <https://www.uniprot.org/uniprot/Q6LZH0>.
- [22] NCBI. *nifH nitrogenase iron protein [Methanococcus aeolicus Nankai-3]*. URL: <https://www.ncbi.nlm.nih.gov/gene/5327360>.
- [23] Uniprot Database. *Nitrogenase iron protein*. URL: <https://www.uniprot.org/uniprot/POCW56>.
- [24] MM Georgiadis et al. “Crystallographic structure of the nitrogenase iron protein from Azotobacter vinelandii”. In: *Science* 257.5077 (1992), pp. 1653–1659.
- [25] Uniprot Database. *Global nitrogen regulator NrpR*. URL: <https://www.uniprot.org/uniprot/Q6LZL7>.
- [26] Uniprot Database. *Procesing protease*. URL: <https://www.uniprot.org/uniprot/067945>.
- [27] E Morett and L Segovia. “The sigma 54 bacterial enhancer-binding protein family: mechanism of action and phylogenetic relationship of their functional domains”. In: *Journal of bacteriology* 175.19 (1993), pp. 6067–6074.
- [28] Fabiane GM Rego et al. “The expression of nifB gene from Herbaspirillum seropedicae is dependent upon the NifA and RpoN proteins”. In: *Canadian journal of microbiology* 52.12 (2006), pp. 1199–1207.
- [29] Eduardo Santero et al. “Role of integration host factor in stimulating transcription from the  $\sigma$ 54-dependent nifH promoter”. In: *Journal of molecular biology* 227.3 (1992), pp. 602–620.
- [30] Enrique Morett and Martin Buck. “In vivo studies on the interaction of RNA polymerase- $\sigma$ 54 with the Klebsiella pneumoniae and Rhizobium meliloti nifH promoters: the role of NIFA in the formation of an open promoter complex”. In: *Journal of molecular biology* 210.1 (1989), pp. 65–77.
- [31] Microalga technology. *Ingeniería de Procesos aplicada a la Biotecnología de Microalgas*. URL: <https://w3.ual.es/~jfernand/ProcMicro70801207/tema-1---generalidades/1-3-nutrientes.html#:~:text=Las%5C%20microalgas%5C%20requieren%5C%20para%5C%20su,suele%5C%20ser%5C%20alg%5C%C3%BA%5C%20fosfato%5C%20inorg%5C%C3%A1l%5C%20Alnico..>
- [32] BLAST. *The Expect value (E)*. URL: [https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE\\_TYPE=BlastDocs&DOC\\_TYPE=FAQ#:~:text=The%5C%20Expect%5C%20value%5C%20\(E\)%5C%20is,describes%5C%20the%5C%20random%5C%20background%5C%20noise..](https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=FAQ#:~:text=The%5C%20Expect%5C%20value%5C%20(E)%5C%20is,describes%5C%20the%5C%20random%5C%20background%5C%20noise..)
- [33] Satoshi Nakagawa et al. “Hydrogenivirga caldilitoris gen. nov., sp. nov., a novel extremely thermophilic, hydrogen-and sulfur-oxidizing bacterium from a coastal hydrothermal field”. In: *International journal of systematic and evolutionary microbiology* 54.6 (2004), pp. 2079–2084.
- [34] Nishu Goyal, Zhi Zhou, and Iftekhar A Karimi. “Metabolic processes of Methanococcus maripaludis and potential applications”. In: *Microbial cell factories* 15.1 (2016), pp. 1–19.
- [35] IFAS Extension. *BIOLOGÍA Y MANEJO DE NOSTOC (CYANOBACTERIA) EN VIVEROS E INVERNADEROS*. URL: <https://edis.ifas.ufl.edu/publication/AG432/>.