Final project

Genomic Data Analysis and Visualization (GDAV) **2023-2024**

Jaime Huerta Cepas y Carlos Pérez Cantalapiedra

Name:	Jacob Gonzalez Isa
Email:	jacob.gonzalez.isa@alumnos.upm.es

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 research grant proposing to investigate this effect and, lucky you!, you received a very generous funding from Tyrell Corporation. Your grant proposed to perform an in depth genomic and metagenomic exploration of this singular hot spring ecosystem, which involved eight work packages:

1. Metagenomic analysis
2. RNA-seq samples and read mapping
3. Variant calling
4. Differential expression analysis
5. Functional analysis
6. Phylogenetic analysis
7. Conclusions

1. Metagenomics

As a first step, you decide to run shotgun metagenomic sequencing of the microbiome in two conditions obtained at different times: 1) one sample taken during the high temperature episodes, and 2) another sample taken right *after* the episodes, when the temperature is back to normal and the bloom of algae has started.

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

The raw read files (reverse and forward) were produced by Illumina pair-end sequencing and are now located in your home folder in at the 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)
```

Tasks

1. Use the tool mOTUs to perform a taxonomic profiling of your samples
   (Warning: process both forward and reverse read files!)

Questions

1.1 What is the most abundant organism in high-temperature? What is its relative abundance in the sample?

In high temperature, 10 species were detected, being the most abundant ***Aquifex aeolicus*** [ref_mOTU_v31_10705], with a relative abundance of **0.9050938825** (90%).

1.2 Has the most-abundant organism in high-temp been sequenced before? (i.e. there is a public whole genome sequence in current databases). Provide the sources supporting your answer.

Yes, it has been sequenced before. First published in 1998.

Deckert, G., Warren, P., Gaasterland, T. *et al.* The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus* . *Nature* **392**, 353–358 (1998). https://doi.org/10.1038/32831

The Bioproject accesion is PRJNA215.
https://www.ncbi.nlm.nih.gov/bioproject/PRJNA215/

The registration date was 25-Feb-2003

1.3 If possible, describe the most important features of the most-abundant organism in high-temp.

Hyperthermophilic organism that is one of the deepest branching organism.

Aquifex aeolicus strain VF5*. This hyperthermophilic organism grows optimally at 95 C on hydrogen, oxygen, carbon dioxide, and mineral salts. It is one of the earliest diverging bacteria known (deepest branching) based on 16S ribosomal RNA sequencing and phylogenetic analyses. Many of the enzymes this organism produces are studied due to their heat stable nature. The genome sequence of this organism will provide information on the mechanisms involved in survival at high temperatures.*
https://www.ncbi.nlm.nih.gov/bioproject/57765

1.4 Do you still detect the same organism if you run mOTUs with a more stringent threshold for the number of marker-gene detections (at least 4 genes)? What is the estimated relative abundance in such a case?

In this case, only three organisms are detected. The most abundant is still the same, although with a higher relative abundance in this case: 0.9154098245

_____
cat 01metagenomics/results/hightemp_report_threshold4.txt
Aquifex aeolicus [ref_mOTU_v31_10705] 0.9154098245
Methanococcus maripaludis [ref_mOTU_v31_01426] 0.0842476155
unassigned 0.0003425600

1.5 What is the most abundant organism in normal temperature? What is its relative abundance in the sample?

Again, ***Aquifex aeolicus***, but now with a smaller relative abundance: 0.0333818811

1.6 is this species also present in the high-temperature samples? At which relative abundance?

Yes, it was present at a higher relative abundance (90%)

1.7 Which is the condition (high or normal temperature) with a greater level of alpha biodiversity?

The normal-temperature condition. A simple measure would be the number of total

species. For high temperature, no more than 10 species. For normal temperature, around 190.
This makes sense, as normal-temperature conditions do not require specialized metabolism or traits and, consequently, diverse species can thrive.

1.8 Do you detect any eukaryotic algae in the high or normal temperature sample? If so, report. If not, explain why not?

No, no eukaryotic algae were detected.

According to mOTUs paper, during the preprocessing for the obtention of the mOTUs database, *"Genes were predicted on length-filtered (≥ 500 bp) scaffolded contigs (hereafter scaffolds) using Prodigal (v2.6.3)".*

Prodigal is a prokaryotic-oriented annotator, then the mOTUs database does not contain eukaryotic genes apparently.

https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-022-01410-z

2. RNA-seq samples and read mapping

You are very excited with your preliminary findings that one specific organism is very abundant in high-temperature episodes. To further characterise it, your lab isolated the organism and i) sequenced its whole genome and ii) performed a RNAseq assay with samples at normal- and high-temperature conditions, two biological replicates each. The sequencing company provided you with an already assembled genome and clean, high quality RNAseq reads in FastQ 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 terminal and `cd` to the directory where you left both the reference data and the sequencing reads. First things first. You want to be sure what you are dealing with, how is your data…

Tasks

1. check your RNAseq samples (folder `RNAseq/`) and answer the following questions

Questions

2.1 How many samples do you have?

4 samples (2 conditions and 2 replicates per condition).
Each sample has two read files (r1 and r2).

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

Below, the output of 'seqkit stats'. In bold, the number of reads for each sample.

file	format	type	num_seqs	sum_len	min_len	avg_len	max_len
RNAseq/hightemp01.r1.fq.gz	FASTQ	DNA	**318,719**	31,871,900	100	100	100
RNAseq/hightemp01.r2.fq.gz	FASTQ	DNA	**318,719**	31,871,900	100	100	100
RNAseq/hightemp02.r1.fq.gz	FASTQ	DNA	**318,719**	31,871,900	100	100	100
RNAseq/hightemp02.r2.fq.gz	FASTQ	DNA	**318,719**	31,871,900	100	100	100
RNAseq/normal01.r1.fq.gz	FASTQ	DNA	**288,742**	28,874,200	100	100	100
RNAseq/normal01.r2.fq.gz	FASTQ	DNA	**288,742**	28,874,200	100	100	100
RNAseq/normal02.r1.fq.gz	FASTQ	DNA	**288,742**	28,874,200	100	100	100
RNAseq/normal02.r2.fq.gz	FASTQ	DNA	**288,742**	28,874,200	100	100	100

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

These are short-reads with Old Illumina naming.
Paired-end, two files with Read 1 (R1) and Read 2 (R2). This is the conventional naming by Illumina.

Anyways, the infer_experiment.py script from RSeQC (https://rseqc.sourceforge.net/) can help us to find out what type of reads and experiment we are dealing with.

The output was:

This is PairEnd Data
Fraction of reads failed to determine: 0.0051
Fraction of reads explained by "1++,1--,2+-,2-+": 0.4986
Fraction of reads explained by "1+-,1-+,2++,2--": 0.4963

From the documentation of RSeQC:
***Interpretation**: 1.72% of total reads were mapped to genome regions that we cannot determine the "standness of transcripts" (such as regions that having both strands transcribed). For the remaining 98.28% (1 - 0.0172 = 0.9828) of reads, half can be explained by "1++,1–,2+-,2-+", while the other half can be explained by "1+-,1-+,2++,2–". We conclude that this is **NOT a strand specific dataset** because "strandness of reads" was independent of "standness of transcripts"*

Our experiment is not stand-specific.

NOTE: in further stages of the exercise (when using *bamtools stats*, for instance) further reports support this.

2.4 Are all the reads of the same length?

Each condition (hightemp and normaltemp), have the same read length, but there are differences between condition in terms of number of reads.

file	format	type	num_seqs	sum_len	min_len	avg_len	max_len
RNAseq/hightemp01.r1.fq.gz	FASTQ	DNA	318,719	31,871,900	100	100	100
RNAseq/hightemp01.r2.fq.gz	FASTQ	DNA	318,719	31,871,900	100	100	100
RNAseq/hightemp02.r1.fq.gz	FASTQ	DNA	318,719	31,871,900	100	100	100
RNAseq/hightemp02.r2.fq.gz	FASTQ	DNA	318,719	31,871,900	100	100	100
RNAseq/normal01.r1.fq.gz	FASTQ	DNA	288,742	28,874,200	100	100	100
RNAseq/normal01.r2.fq.gz	FASTQ	DNA	288,742	28,874,200	100	100	100
RNAseq/normal02.r1.fq.gz	FASTQ	DNA	288,742	28,874,200	100	100	100
RNAseq/normal02.r2.fq.gz	FASTQ	DNA	288,742	28,874,200	100	100	100

2.5 Just from the files you have been provided, could you say something about the orientation of the reads: are they forward or reverse; coding or template strand; 5' to 3' or 3' to 5'?

All the reads in R1 and R2 are **5' → 3'** because that is how DNA sequences are represented in bioinformatics by convention (in FASTA, FASTQ or whatever format).

However, depending on the "strandedness" of an RNA-seq experiment, R1 and R2 might be **forward or reverse**. *Stranded RNA library preparation preserves the directionality of the transcript by distinguishing the first and second strands of cDNA*.(https://www.azenta.com/blog/stranded-versus-non-stranded-rna-seq). In fr-firststrand

case, R1 would map to forward and R2 to reverse, whereas the opposite would occur in fr-secondstrand (see image below). https://www.biostars.org/p/344264/
In our case, which is unstranded, the reads are simply a mix.

Regarding **coding and template**, this also depends on the strandedness of the paired-end data. Notice that, if we had "fr-firsttrand" stranded data, the R1 would be aligned to the template and the R2 to the coding.

To summarize, we cannot say anything about forward/reverse and coding/template from our unstranded data. The image below is a very good reference:

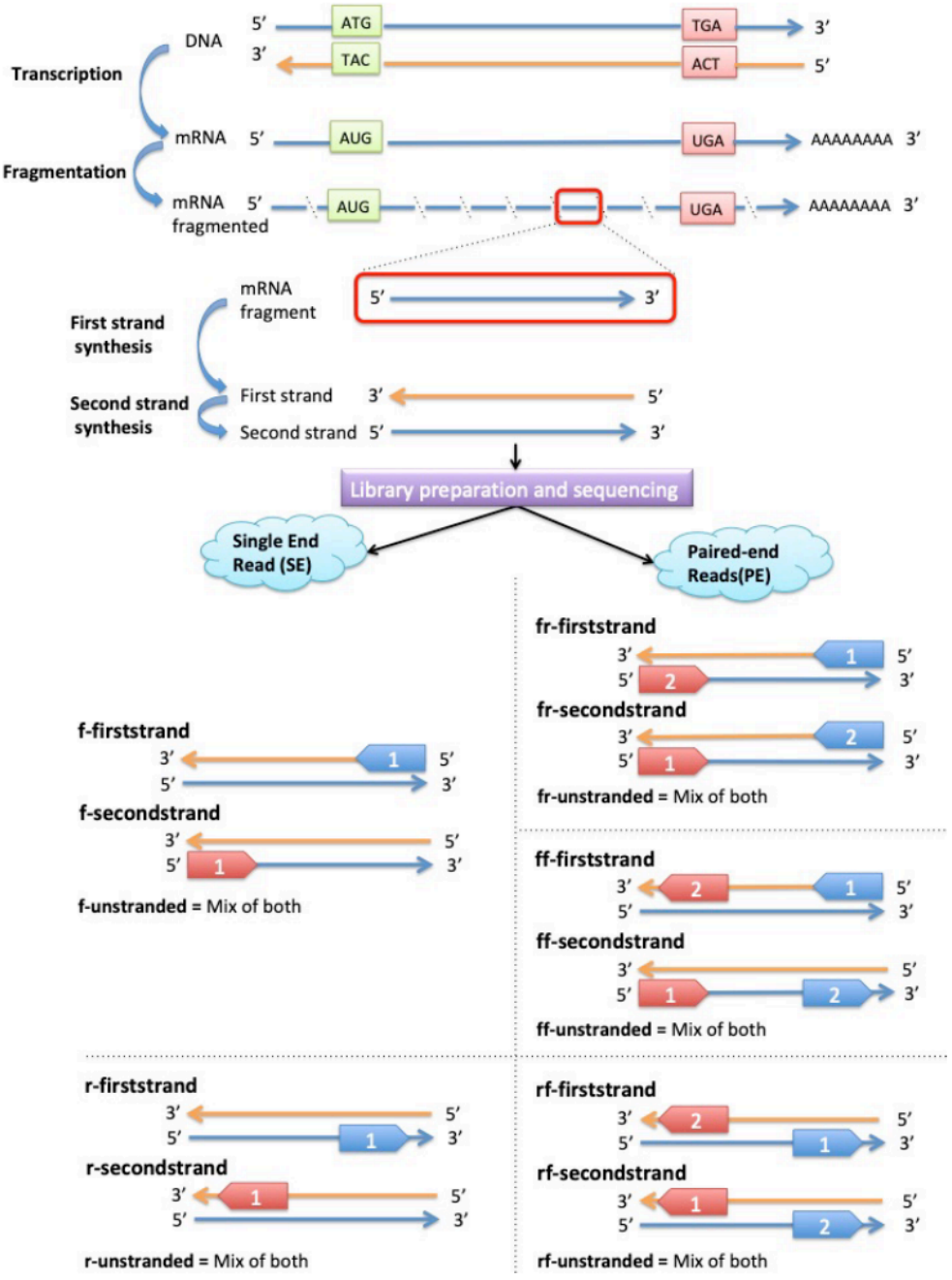
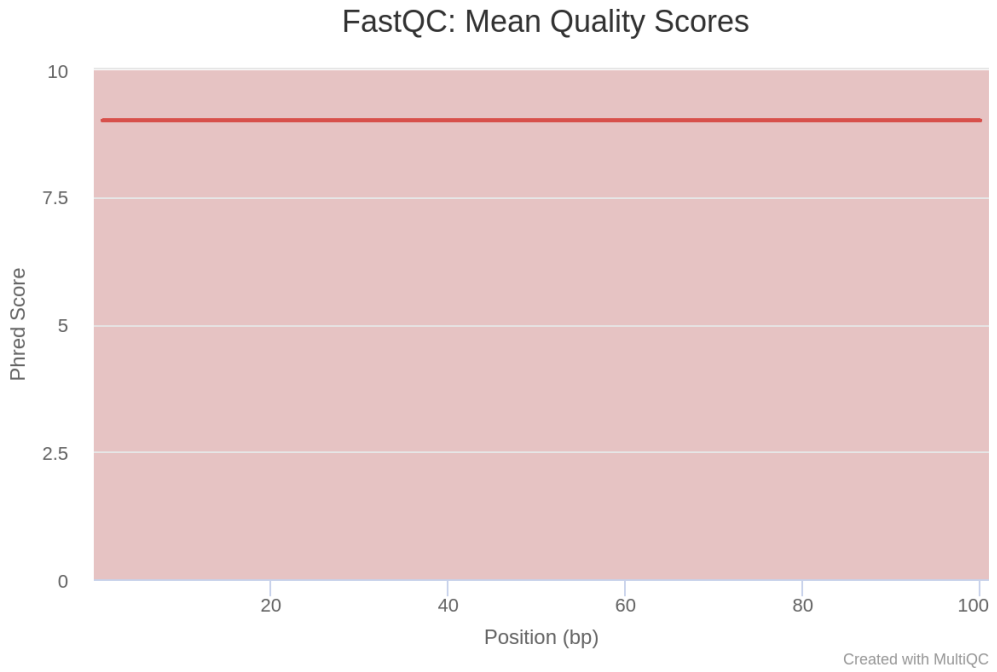


Figure 1: Overview of the different library types.

2.6 Regarding the quality of the reads, is there something that calls your attention?

Yes, that all PHRED scores are the same (I, which is quality of 9, quite low). Below, the report for all the samples shown in MultiQC.



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 ...

Tasks

2. First, you will need to **create an index of the reference genome** using BWA INDEX. (Note that for this work we are using BWA Version: 0.7.17-r1188)
   Done

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

4. Finally, **create BAM files for your mappings** using SAMTOOLS (we are using samtools version 1.9) and remove the SAM files afterwards.
   Done

Questions

2.7 How many mappings are there in your BAM files? How many different reads are present in your BAM files? Are these previous numbers equal or different? Can you explain why? Discuss also these numbers compared to the original number of reads you had in each sample. (You may use `samtools flagstat` and/or `samtools stats` to help you find the answers).

The number of mapped reads:

normal01	577484
normal02	577484
hightemp02	637438
hightemp01	637438

The number of total reads: (this include secondary alignment, that's is why there are more)

normal01	577508
normal02	577503
hightemp01	638004
hightemp02	637997

_____

Compared with the original number of reads,

318,719 *(for hightemp condition)*
288,742 *(for normaltemp condition)*

the number of mapped reads is twice as much. This is because now, in each .bam, reads are coming from R1 and R2, so it is doubled.

2.8 Are these mappings appropriate to perform an analysis of Copy Number Variation (https://en.wikipedia.org/wiki/Copy-number_variation)? Explain why.

#rname	startpos	endpos	numreads	covbases	coverage	meandepth	meanbaseq	meanmapq
Aquifex_genome	1	1556396	638004	1429486	**91.8459**	**40.9445**	40	59.5

Above, the mean depth and coverage. However, this is more related to SNP variant calling, and CNV is a quite special type of structural variation.

Indeed, RNA-seq is not a good strategy for analyzing Copy Number Variation. Reads mapping to a specific gene is associated to gene expression and not necessarily to more copies of a gene. Moreover, some CNVs (such as satellite arrays) are not transcribed and RNA-seq cannot detected

Detecting CNAs from RNA-Seq is a challenging task because coverage is highly variable owing to differences in gene expression.
(https://academic.oup.com/bioinformatics/article/37/22/4023/6300509)

However, some recent methods have made efforts to detect CNVs with RNA-seq.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9177690/

3. 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...

Tasks

1. Before performing variant calling, we need to **sort the mappings** of each of our samples (tip: you should use `` `samtools sort` `` for that). Remove the previous BAM files once you have the sorted ones.

   I did it before

2. Then, you decide to **merge the mappings** from all your samples into a single BAM file (tip: use `` `samtools merge` `` for that). **Don't remove** any BAM file in this step.

   Done

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. We are using bcftools version 1.9 here.

   Done, calls.bcf created

4. Now, repeat the previous step (task 3), but without '--ploidy 1' (i.e. using the default value, which is for diploids). Save the variant calls to a different file than in the previous task, so that we can compare results.

   Done, calls_diploid.bcf created

5. Finally, you decide to **repeat** the previous step (task 4), 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 in the previous task, but save it to a different file so that we can compare results.

   Done, calls_4files.bcf created

   Done, calls_4files_diploid.bcf created

Questions

3.1 How many variants did you expect to identify, if any? How many variants did you actually detect? How many are SNPs, how many insertions and how many deletions?

I would not expect any (or many) variants. This is because the genome we are working with was sequenced *de novo*. This is not the reference genome found in the databases. Therefore, if we are doing RNA-seq on the newly sequenced organism, we would expect no variants.

All the variants are SNPs. Curiously, the “diploid” files have one more SNP.

file_name	**variants**	**SNPs**	**indels**
calls	2	2	0
calls_diploid	3	3	0
calls_4files	2	2	0
calls_4files_diploid	3	3	0

It is okay to find variants anyways. We might check the raw data from the whole genome sequencing to check if the “called bases” had less quality in the assembly. Yet if we find everything okay, I would propose two hypotheses:

1. The organism suffered a mutation when growing in culture (supposing that thre was a time interval between sample collection for DNA sequencing and RNA sequencing, which would not have been the most appropriate experimental approach)
2. It is an RNA editing event (the mRNA is modified)

3.2 How many variants have quality greater than 100?

There are **two**.

_____
calls_diploid:

Aquifex_genome 1265060 . C T **188** PASS DP=249;VDB=3.5208e-10;SGB=-0.693147;RPB=0.944775;MQB=1;BQB=0.984769;MQ0F=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=136,0,112,0;MQ=60 GT:PL 0/1:221,0,249

calls_4files:

Aquifex_genome 1265060 . C T **950** PASS DP=674;VDB=1.01192e-06;SGB=-101.196;RPB=0.969374;MQB=1;MQSB=1;BQB=0.985692;MQ0F=0;ICB=0.166667;HOB=0.5;AC=4;AN=8;DP4=294,55,272,45;MQ=60 GT:PL 0/1:223,0,249 0/1:255,0,255 0/1:255,0,255 0/1:255,0,255

3.3 How many variants have depth of coverage greater than 100?

There are 2 (actually the same)

_____
calls_diploid:

Aquifex_genome 1265060 . C T 188 PASS **DP=249**;VDB=3.5208e-10;SGB=-0.693147;RPB=0.944775;MQB=1;BQB=0.984769;MQ0F=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=136,0,112,0;MQ=60 GT:PL 0/1:221,0,249

Aquifex_genome 1265060 . C T 950 PASS

calls_4files:

DP=674;VDB=1.01192e-06;SGB=-101.196;RPB=0.969374;MQB=1;MQSB=1;BQB=0.985692;MQ0F=0;ICB=0.166667;HOB=0.5;AC=4;AN=8;DP4=294,55,272,45;MQ=60 GT:PL 0/1:223,0,249 0/1:255,0,255 0/1:255,0,255 0/1:255,0,255

3.4 Compare the output you obtained using the merged data with '--ploidy 1' (task 3) and without it (task 4), and with the one using 4 independent samples (task 5). What differences do you observe? What is the cause of these differences? What advantages and disadvantages do you think have using merged vs independent samples?

In exercise 3.1 it was already clear that with ploidy=2, one more variant was detected.
When looking the filtered reads by quality (and by depth)
calls
0

calls_diploid
1

calls_4files
0

calls_4files_diploid
1

So it appears that t**he highest quality variant is only detected when ploidy = 2!**

Some discussion on the –ploidy parameter:
https://github.com/samtools/bcftools/issues/1523
When –ploidy 1, *"you are telling the program that there is only one expected allele at a position and if there are more, they should be considered sequencing and alignment errors"*

On the right, I explicitly point the difference with the highest quality variant. We observe that this SNP is "heterozygous", which means that not all the reads have the SNP. When –ploidy 1 (call_4files.vcf and calls.vcf), if not every read has the SNP ("T"), this will be considered a sequencing error.

Notice that, despite having a variant with high depth, if this is "heterozygous" and –ploidy is set to one, it will not be called.

Notice, however, the case in which –ploidy 1 is calling SNPs.
In this case, every read has the SNP (“A”), so the variant is called.

———

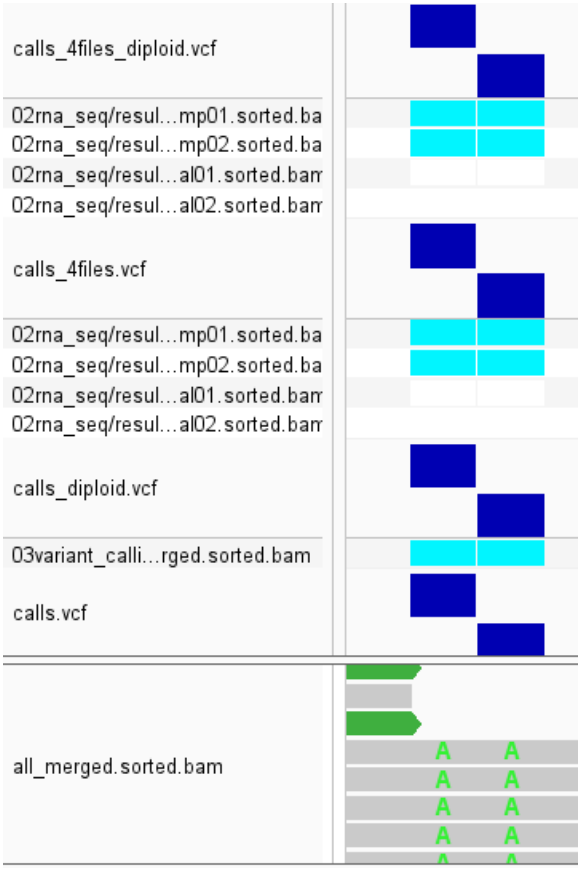
Regarding **merged vs independent** samples, the image on the right can help us.
-Independent samples:

- PRO: They tell you about the quality of the variant calling in each sample. In our case, the normal-temperature samples had lower quality in the SNP (on the right).
- CON: if samples have low depth and/or coverage.

-Merged sample:

- PRO: low depth can be mitigated because, in the end, you are increasing the number of reads in your alignment.
- CON: if a variant is only present in one condition, you would miss it.

Assuming a good depth, I would go for independent samples, as it is more informative.



3.5 Focusing on the variant with the best quality, does this variant look homozygous or heterozygous? Also, 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 about how your variant could be affecting a gene phenotype (just an example, even if it is not the case of this exercise).

calls_diploid
Aquifex_genome 1265060 . C T 188 PASS DP=249;VDB=3.5208e-10;SGB=-0.693147;RPB=0.944775;MQB=1;BQB=0.984769;MQ0F=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=136,0,112,0;MQ=60 **GT:PL** **0/1:**221,0,249

calls_4files_diploid
Aquifex_genome 1265060 . C T 950 PASS DP=674;VDB=1.01192e-06;SGB=-101.196;RPB=0.969374;MQB=1;MQSB=1;BQB=0.985692;MQ0F=0;ICB=0.166667;HOB=0.5;AC=4;AN=8;DP4=294,55,272,45;MQ=60 **GT:**PL **0/1:**223,0,249 0/1:255,0,255 0/1:255,0,255 0/1:255,0,255

The best quality variant is found in calls_4files_diploid.bcf (950 quality)

It looks heterozygous (0/1), carrying one of each the REF and ALT alleles (seems a transversion C > T). But this is not a true heterozygousity, since bacteria are haploid. Indeed, C>T is a typical signature of RNA editing. https://pubmed.ncbi.nlm.nih.gov/32729074/

A point mutation can affect in many ways, from being silent (due to the degeneration of the genetic code in the third nucleotide of the codon) to non-sense (a stop codon arises, then truncating the protein, which is likely to have detrimental effects as the gene is considered to be knocked-out if the stop codon appears early in the gene sequence).

In the case of the exercise, I am doubtful. The SNP called is C>T, however, since the gene is in the reverse strand, would it be a GGG > GAG mutation?

Below I attach an image showing in more detail the mutation.
Since this is an unstranded RNA-seq, I am not very sure if a CCC > CTC would be GGG > CAC. If that were the case, we would be dealing with a missense mutation: GGG > GAG, which in aminoacid would be Gly > Glu.

3.6 Download, to your local machine, the files with the mappings and the fasta file of the genome. Use them to visualise **the best variant** you have using IGV, and capture the image of the variant. Note that you will likely need the indexes of the mappings (.bai) and genome (.fai) files. Capture the IGV image with the variant, and paste it below as an answer for this question.

I am visualizing the best variant (they are the same, but I am checking the effect of --ploidy in the results).
Also, I just uploaded to IGV the all_merged.bam/bai because it allows a clear visualization.

4. Differential expression analysis

Next, you will compare the differences in gene expression between the samples grown under normal and high-temperature conditions. Hopefully, this could give some clue about genes involved in the environmental changes observed.

Tasks

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). Finally, you should obtain a ***<sample_id>.count*** file from each of your samples, as shown below.

Example:

```
> head -4 normal01.count
AQUIFEX_00001 456
AQUIFEX_00002 154
AQUIFEX_00003 0
AQUIFEX_00004 134
```

2. Now, you can use these counts to perform the DEA. For that use the Bioconductor and R packages DESeq2 and tibble in the next R script:

```
#!/usr/bin/env Rscript
#

library(DESeq2);
library(tibble);

# Create an object with the directory containing HTseq counts:
directory <- "WRITE YOUR PATH TO YOUR .count FILES"
#list.files(directory)
sampleFiles <- list.files(directory, pattern = ".*count")
sampleFiles

# create a vector of sample names. Ensure these are in the same order as the
"sampleFiles" object!
sampleNames <- c( "hightemp01", "hightemp02","normal01","normal02")

# create a vector of conditions. again, mind that they are ordered correctly!
replicate <- c("Rep1","Rep2","Rep1","Rep2")

# create a vector of conditions. again, mind that they are ordered correctly!
sampleCondition <- c("hightemp","hightemp","normal", "normal")

# now create a data frame from these three vectors.
sampleTable <- data.frame(
              sampleName = sampleNames,
              fileName = sampleFiles,
              condition = sampleCondition,
              replicate = replicate)

sampleTable
```

If you run the R script up to this point, the data frame of your experiment design (sampleTable) should look like this:

sampleName	fileName	condition	replicate
hightemp01	hightemp01.count	hightemp	Rep1
hightemp02	hightemp02.count	hightemp	Rep2
normal01	normal01.count	normal	Rep1
normal02	normal02.count	normal	Rep2

Next, make the DESeq2 object using the counts and metadata. Be sure you write properly the reference level:

```
## Make DESeq2 object from counts and metadata

ddsHTSeq <- DESeqDataSetFromHTSeqCount(
              sampleTable = sampleTable,
              directory = directory,
              design = ~condition)

# specify the reference level:

ddsHTSeq$condition <- relevel(ddsHTSeq$condition, ref = "normal")
```

Add a minimal pre-filtering to keep only rows that have at least 10 reads total.

```
# sum counts for each gene across samples

sumcounts <- rowSums(counts(ddsHTSeq))

# get genes with summed counts greater than 10; remove lowly expressed genes

keep <- sumcounts > 10

# keep only the genes for which the vector "keep" is TRUE

ddsHTSeq_filter <- ddsHTSeq[keep,]
```

Run `DESeq` *and* `results` *functions. If you want, check the contrast design ("condition_hightemp_vs_normal") using the 'resultsNames(dds)' function.*

```
dds <- DESeq(ddsHTSeq_filter)

# get results table

res <- results(dds, pAdjustMethod="BH")

summary(res)

# check out the first few lines

head(res)

mcols(res, use.names = T)
```

Finally, add the code to create the output files:

```
##Create normalized read counts

normalized_counts <- counts(dds, normalized=TRUE)

normalized_counts_mad <- apply(normalized_counts, 1, mad)
normalized_counts <- normalized_counts[order(normalized_counts_mad, decreasing=T), ]

#DESeq get results table

Res_A_X_total <- results(dds, name="condition_hightemp_vs_normal", pAdjustMethod="BH")
Res_A_X_total <- Res_A_X_total[order(Res_A_X_total$padj),]
Res_A_X_total <- data.frame(Res_A_X_total)
Res_A_X_total  <- rownames_to_column(Res_A_X_total, var = "ensembl_gene_id")

Res_A_X_total$sig <- ifelse(Res_A_X_total$padj <= 0.05, "yes", "no")

Res_A_X_total_0.05 <- subset(Res_A_X_total, padj <= 0.05)

# Export output files

write.csv(normalized_counts,"deseq2_normcounts.csv")
write.csv(Res_A_X_total, "deseq2_results_total.csv")
write.csv(Res_A_X_total_0.05, "deseq2_results_padj_0.05.csv")
```

Questions

4.1 Describe the fields in the “deseq2_results_padj_0.05.csv” file.

All definitions are in the link(https://support.illumina.com/help/BS_App_RNASeq_DE_OLH_1000000071939/Content/Source/Informatics/Apps/DESeq2ResultFile_swBS.htm), but I will try my best to define them with my own words.

"ensembl_gene_id": unique identifiers for each gene in the analysis (derived from the Ensembl database apparently, although if you look for the identifiers in the database you will not obtain any results)

"baseMean": “ the average of normalized counts across all samples for a particular gene. It represents the average expression level of the gene, providing a measure of its overall abundance across conditions

"log2FoldChange": the effect size estimate. It represents the magnitude and direction of change in expression between different conditions. A positive value indicates upregulation, and a negative value indicates downregulation. log2(fold change) = log2(expression value in condition A) / log2(expression value in condition B)

"lfcSE": The standard error of the log2 fold change. It reflects the uncertainty or variability associated with the log2 fold change estimate. lfcSE = standard deviation (logfc of each sample) / sqrt(number of samples)

"stat": The test statistic is calculated to assess the significance of the log2 fold change. It is often used in hypothesis testing to determine whether the observed fold change is statistically significant. By default DESeq2 uses the Wald test: “*For the Wald test, stat is the Wald statistic: the log2FoldChange divided by lfcSE, which is compared to a standard Normal distribution to generate a two-tailed pvalue.” (https://www.biostars.org/p/9559224/)*

"pvalue": The p-value is a measure of the evidence against a null hypothesis. In the context of

differential expression analysis, a low p-value indicates that the observed differences are unlikely to be due to random chance.

"padj": The adjusted p-value, in this case calculated with the Benjamini-Hochberg method. It accounts for multiple testing and provides a corrected p-value to control the false discovery rate.

"sig": This column likely indicates whether the gene is considered statistically significant based on a chosen significance threshold (e.g., adjusted p-value < 0.05). It might contain "yes" or "no" to indicate significance.

4.2 How many genes were differentially expressed genes (DEGs) with p-adj < 0.05? Among these, which ones would be up- and down-regulated?

Five DEGs out of 1508
3 upregulated
AQUIFEX_01423
AQUIFEX_01759
AQUIFEX_01761

2 downregulated
AQUIFEX_01723
AQUIFEX_01749

4.3 Use the genome.gff file to retrieve the potential function of the DEGs, if any. Which are those functions?

Nif-specific regulatory protein
hypothetical protein
hypothetical protein
FeMo cofactor biosynthesis protein NifB
Nitrogenase iron protein 1

These are all related with nitrogen fixation

4.4. Based on the previous results, is there a common function or process which could be differentially regulated under the environmental conditions under study. Which is such a process, if any? Which DEGs are involved? Are those DEGs up- or down-regulated or both?

The common function which is differentially regulated is nitrogen metabolism.

The nitrogen fixation-related proteins are upregulated (AQUIFEX_01423, AQUIFEX_01759, AQUIFEX_01761)

The ‘hypothetical protein’ are downregulated (AQUIFEX_01723, AQUIFEX_01749)

5. Comparative and Functional analysis

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

Tasks

1. Extract the sequence of each over expressed gene out from the assembled proteome of your species isolate (`proteome.faa`)
   Done, with my custom extrac_fasta.py
2. Save each sequence in FASTA format in an individual file and search for functional annotations in as many databases as you consider relevant.
   Done, with my custom divide_fasta.py

Questions:

5.1 Provide functional information of the overexpressed genes obtained in the previous exercise? Include the information about what database/resource did you use to infer each functional term.

GeneName	**Functional description**	**Protein Domains (by *de novo* detection)**	**KEGG Pathway / Module**	**Gene Name (based on closest homolog with curated functional information (i.e. SwissProt))**
AQUIFEX_01423	EggNOG-mapper transcription factor binding BLAST: sigma-54-dependent Fis family transcriptional regulator (BLAST)	InterPro: gaf_1 SIGMA54_INTERACT_4 HTH_8 EggNOG-mapper: GAF_2 HTH_8 Sigma54_activat	EggNOG-mapper: ko02020 ko05132 map02020 map05132 Very generic terms “gene activation”, “trancription factor”, “two-component system”	BLAST-Uniprot: (I am using this because the Uniprot ID is typically an input STRING) **O67661_AQUAE** (Transcriptional regulator (NtrC family)
AQUIFEX_01759	EggNOG-mapper FeMo cofactor biosynthesis protein NifB BLAST: Nitrogen fixation protein NifB	InterPro: RADICAL_SAM EggNOG-mapper: SAM_adeno_trans	EggNOG-mapper: K02585 nifB; nitrogen fixation protein NifB (no pathway nor module, unclassified)	**Q6LZH0_METMP** (FeMo cofactor biosynthesis protein NifB)
AQUIFEX_01761	EggNOG-mapper Nitrogenase iron protein 1	InterPro: NIFH_FRXC_3 EggNOG-mapper:	EggNOG-mapper: ko00625 ko00910	**P0CW56** (Nitrogenase iron protein NifH)

	BLAST: nitrogenase iron protein	Fer4_NifH	ko01100 ko01120 map00625 map00910 map01100 map01120 M00175 (nitrogen fixation)	

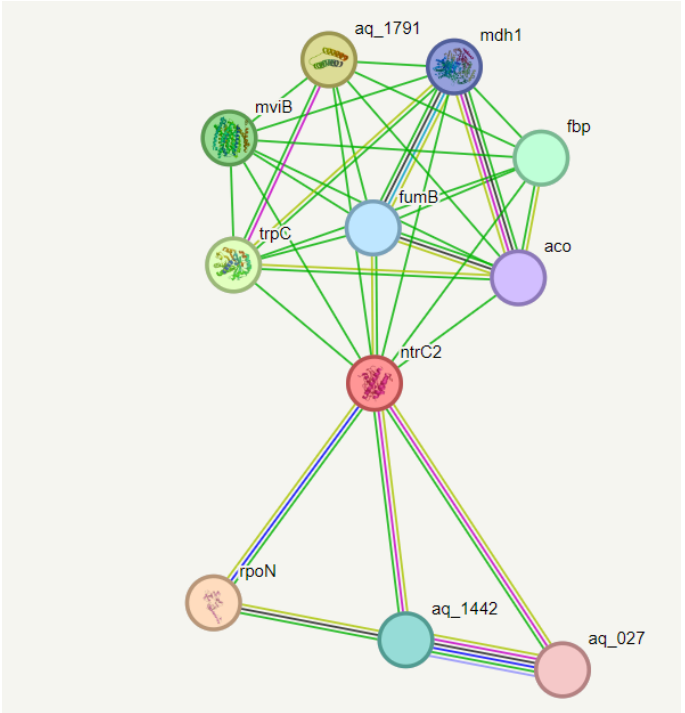
5.2 Are these genes functionally related? Are they involved in protein-protein interactions?

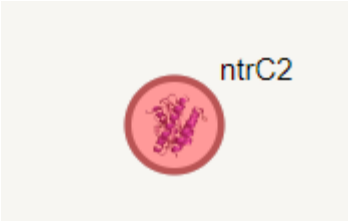
Yes, they are involved in the nitrogen cycle.

I looked up in STRING to look for protein-protein interactions.

O67661_AQUAE

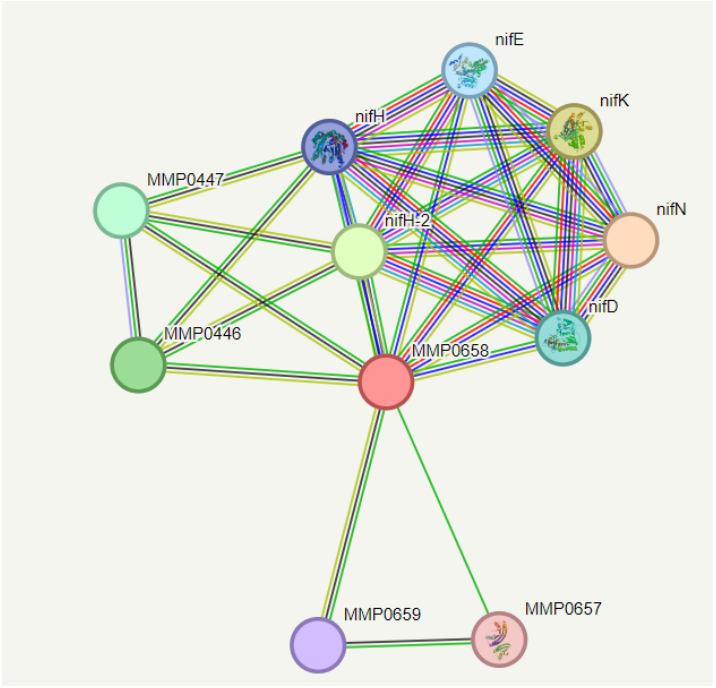
Although it might look that it is involved in protein-protein interactions, these are predicted based on co-expression, text mining, co-occurence, etc. When the network type is set to “physical”, the protein is alone (on the right).

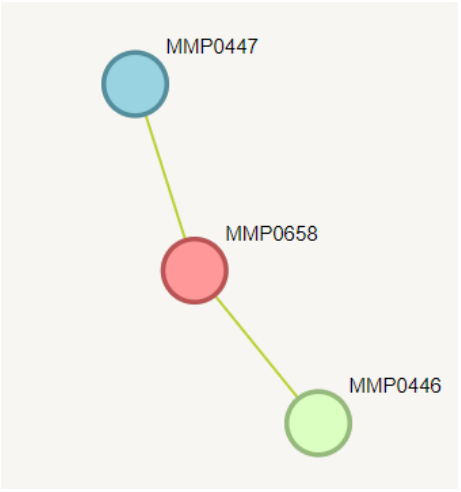




Q6LZH0_METMP

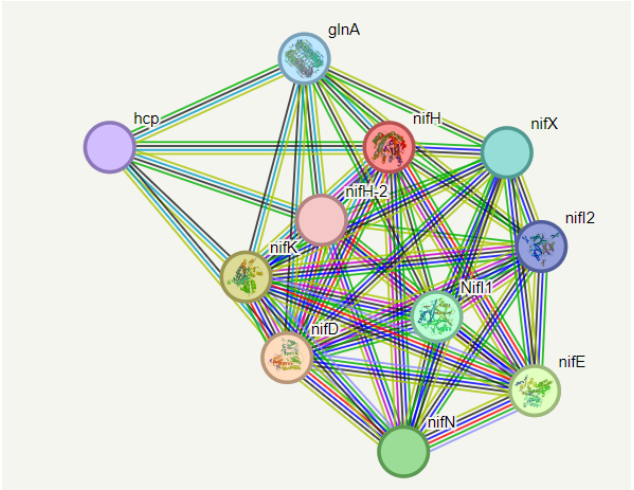
Same as before. Physical interactions are inferred (on the right) with text mining

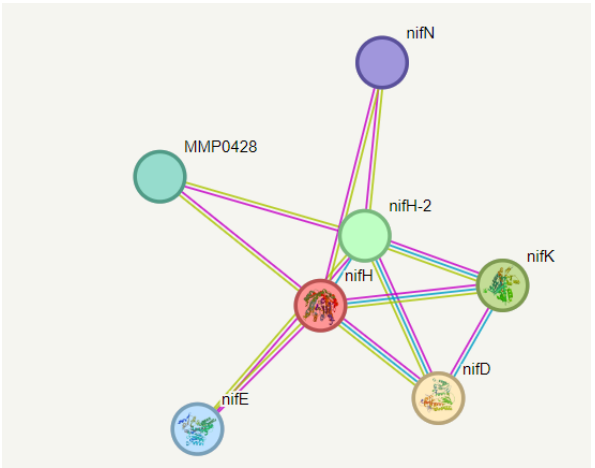




P0CW57

This is the most curated protein, since it is in SwissProt (the others were in TrEMBL). Consequently, in this case this protein has been experimentally studied and protein-protein interaction confirmed (on the right). All of this interactors have to do with nitrogen fixation.





5.3 How could these genes be related with the bloom of algae events?

Nitrogen is the primary nutrient for plants and algae. Then, an increase in nitrogen fixation means the movement of nitrogen from air (N2) to water (NH3). Available nitrogen in water is a perfect environment for algae overgrowth.

This is a well-studied phenomenon. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3566065/

6. Phylogenetic and comparative analysis

The functional analyis of the overexpressed genes gave you an idea about the biological processes happening in the hotspring 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 contains **7 public** genomes, including of the public genome of the same species you isolated and various 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
```

The 7 proteomes are already consolidated in a FASTA file (`all_reference_proteomes.faa`), which is already in your server.

Tasks

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

1. Run a blast search for *each* over expressed protein against all reference proteomes
2. Extract hits with e-value <= 0.001 (tip: you can use blast parameters for this)

Done

3. 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 FASTA file you used for phylogenetic reconstruction!)

Done

4. 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)

Done

5. Visualize the resulting (suggested tools: etetoolkit.org/treeview, ete3)

Done

Questions

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

6.1 What is the closest ortholog of each overexpressed gene based on your phylogenetic analysis? From what species?

AQUIFEX_01423 A0A497XW95_9AQUI (Hydrogenivirga caldilitoris)
AQUIFEX_01759 Q6LZH0_METMP (Methanococcus maripaludis)
AQUIFEX_01761 P0CW57|NIFH_METMP (Methanococcus maripaludis)

6.2 Do orthology assignments support your previous functional annotations? Briefly describe why. (Tip: the sequence IDs of orthologs in the 7 public proteomes are in Uniprot format, you can search for their functional information online)

For AQUIFEX_01423, its closest ortholog:
A0A497XW95_9AQUI -> Nif-specific regulatory protein

Molecular Function ATP bindingSource:UniProtKB-KW
Molecular Function ATP hydrolysis activity (Source:InterPro)
Molecular Function sequence-specific DNA binding (Source:InterPro)
Biological Process regulation of DNA-templated transcription (Source:InterPro)

Yes, it partially supports the previous functional annotation, although the ATP-related activities of the ortholog do not match.

_____

For AQUIFEX_01759, its closest ortholog:
Q6LZH0_METMP FeMo cofactor biosynthesis protein NifB

Indeed, in the previous exercise I already determined that the AQUIFEX_01759 was identical to Q6LZH0_METMP with a 100% of similarity and coverage.

Thus, the orthology assignment fully supports the previous annotation

_____

For AQUIFEX_01761, its closest ortholog:
P0CW57|NIFH_METMP Nitrogenase iron protein

The case is similar.

Indeed, in the previous exercise, I already determined that the AQUIFEX_01761 was similar to Q6LZH0_METMP with a 100% of similarity and coverage.

Thus, the orthology assignment fully supports the previous annotation

6.3 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, which is the one you included in the phylogenetic analysis)

AQUIFEX_01423
Yes, it is present. It is identiical to O67661_AQUAE

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident
☑	sigma-54-dependent Fis family transcriptional regulator [Aquifex aeolicus]	Aquifex aeolicus	978	978	100%	0.0	100.00%

_____
AQUIFEX_01759
No, it is not present. It is almost identical to Q6LZH0_METMP (from Methanococcus maripaludis.)

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☐	nitrogen fixation protein NifB [Methanococcus sp.]	Methanococcus sp.	616	616	100%	0.0	99.34%	306	MDK2928943.1
☐	FeMo cofactor biosynthesis protein NifB [Methanococcus maripaludis]	Methanococcus maripaludis	614	614	100%	0.0	98.67%	306	AVB75737.1

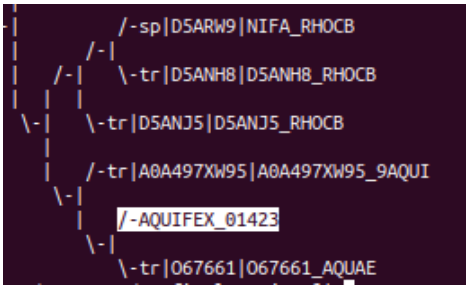
_____
AQUIFEX_01761
No, it is not present. It is identical to P0CW57|NIFH_METMP (Methanococcus maripaludis).

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	nitrogenase iron protein [Methanococcus maripaludis]	Methanococcus maripalu…	564	564	100%	0.0	100.00%	275	WP_011170797.1

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

AQUIFEX_01423 -> Knowing that AQUIFEX_01423=O67661_AQUAE, when looking at the phylogenetic tree, the evolutionary trajectory is clearer.

Looking at this clade (on the right), we can elucidate the most recent evolutionary events, which was a **speciation event** that separated AQUAE and 9AQUI. In addition, notice that there was also a speciation that separated RHOCB from 9AQUI and AQUAE. Interestingly, within the RHOCB lineage, the gene suffered several duplications, unlike in the other species



____

AQUIFEX_01759
AQUIFEX_01761

For both genes, the origin seems to be the same, a Horizontal Gene Transfer (HGT).
This is supported by the absence of this gene in the reference genome of Aquifex and the high similarity (and phylogenetic clustering) with *M.maripaludis* genes .

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

Yes, *M.maripaludis* is the second most abundant species. The co-existence of these species further supports HGT as the best hypothesis.

Indeed this is a common phenomenon in hotsprings:

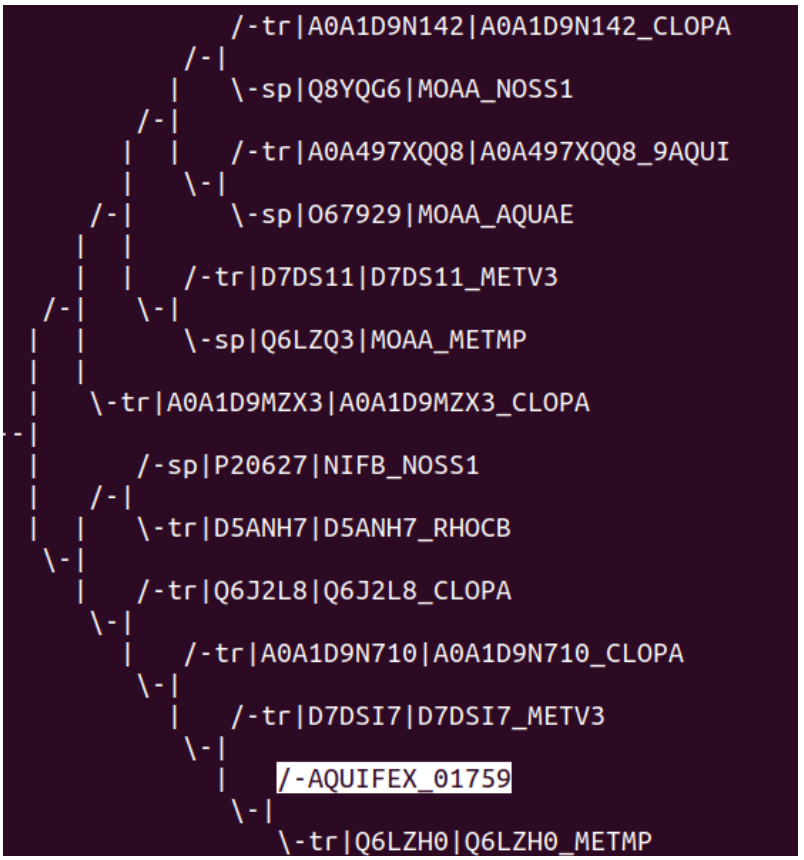
Effect of the environment on horizontal gene transfer between bacteria and archaea
https://pubmed.ncbi.nlm.nih.gov/28975058/

There is also evidence for HGT of nitrogen fixation gene clusters:
Horizontal transfer of the nitrogen fixation gene cluster in the cyanobacterium Microcoleus chthonoplastes
https://www.nature.com/articles/ismej200999

6.6 Include a screenshot of all the trees you obtained after rooting (using ete3, iTOL, etetoolkit.com/treeview, or any other graphical software).

AQUIFEX_01423

AQUIFEX_01759



AQUIFEX_01761

7. Conclusion

Tasks

1. Taking all your results and analyses, try to formulate an integrative interpretation of what's going on in this hotspring.

7.1 Briefly explain your hypothesis about the global effect observed in the hot spring, trying to interpret and connect the results obtained in all the previous steps.

Summary:
Aquifex aeolicus is a hyperthermophilic bacteria that has succeeded in adapting to a hotspring in Iceland, being the most abundant organism in hot events.

Through RNA-seq, we found that three nitrogen metabolism-related genes were upregulated at high temperature.

With phylogenomics, it was shown that two out of these three were acquired via HGT from a methanogenic archaeaA (*Methanococcus maripaludis*).

Interpretation and hypothesis:

Aquifex aeolicus and *Methanococcus maripaludis* coexist in a hot spring environment, along with other microorganisms.

Aquifex acquired nitrogen-related genes from *M.maripaludis* which conferred an advantageous trait which was selected: the ability to use the nitrogen metabolism as energy source in an environment deprived of nutrients.

Combined with its hyperthermophilic features, Aquifex has become exceptionally well-adapted to hot springs.

The nitrogen metabolic capabilities of Aquifex have another effect. Apart from being a source of energy for this bacteria, it also makes the environment nitrogen (ammonia)-rich. Then, after the high-temperature event, the spring is colonized by algae, which are colonizers of nitrogen-rich habitats.

Further views/experiments:

If we have some funding left, I would:

1. Isolate and resequence *Methanococcus maripaludis.* Would we find HGTs occurring on the opposite way (from Aquifex to Methanococcus)?
2. Collaborate with algae/eukaryotic experts to do eukaryotic metagenomics. We might find interesting algae species that show signs of relationships with Aquifex (HGTs, symbiosis, etc.).