Final project

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

Name:	Alberto González Calatayud
Email:	alberto.gcalatayud@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 reazd 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!

Questions

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

Aquifex aeolicus [ref_mOTU_v31_10705] 0.9050938825

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, the most abundant organism in the high-temperature sample, Aquifex aeolicus, has been sequenced before. The complete genome sequence of this hyperthermophilic bacterium is publicly available in current databases. This was confirmed through various reputable sources.

For instance, the KEGG database provides comprehensive genome information for Aquifex aeolicus VF5, including its chromosomal sequence, length, and the number of protein and RNA genes. This information is derived from RefSeq, a well-known source for genome sequences. The database details the statistics of the genome, such as the total number of nucleotides and genes, emphasizing the thoroughness of the sequencing effort.
https://www.kegg.jp/kegg-bin/show_organism?org=aae

Further, a study published in Nature elaborates on the genome features of Aquifex aeolicus. It includes details about the G + C content, protein-coding regions, ribosomal RNA, transfer RNA, and other RNA components. The study also delves into various metabolic aspects of the organism, such as the reductive tricarboxylic acid cycle and gluconeogenesis pathways, offering an in-depth look at the genetic and functional characteristics of Aquifex aeolicus.
https://www.nature.com/articles/32831

Moreover, the NCBI Taxonomy database lists Aquifex aeolicus with its taxonomy ID and provides access to a wealth of related nucleotide, protein, and genome sequences. This further corroborates the availability of its complete genome sequence for public access and research purposes.
https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&id=63363&lvl=3&lin=f
https://www.ncbi.nlm.nih.gov/datasets/taxonomy/63363/

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

Extreme Thermophilic Nature: Aquifex aeolicus is a hyperthermophilic bacterium. It thrives in extremely hot environments, such as those found near hydrothermal vents and hot springs. This organism was first observed at high temperatures (around 89°C) in the outflow of hot springs in Yellowstone National Park, demonstrating its remarkable ability to withstand and flourish in high-temperature environments.

Genome Characteristics: The genome of Aquifex aeolicus is well-characterized and publicly available in databases. Its genome is approximately 1.55 million base pairs long with a G + C content of 43.4%. The genome encodes for a substantial number of proteins, with about 93% of its genome comprising protein-coding regions. This compact genome is indicative of its adaptation to extreme environments.
https://www.nature.com/articles/32831
https://www.kegg.jp/kegg-bin/show_organism?org=aae

Metabolic Pathways: Aquifex aeolicus is an autotroph, obtaining its carbon by fixing CO2 from the environment. Its genome contains genes that encode for the reductive tricarboxylic acid (TCA) cycle, which is key to its autotrophic lifestyle. The presence of duplicated genes in this pathway suggests adaptations for various environmental conditions. Additionally, it has genes for the Embden–Meyerhof–Parnas pathway, indicating its capability for gluconeogenesis.
https://www.nature.com/articles/32831

Respiratory Adaptations: This bacterium can grow using very low concentrations of oxygen, which is unusual for such a high-temperature organism. The respiratory system of Aquifex aeolicus is similar to other bacteria, including enzymes like ubiquinol cytochrome c oxidoreductase and cytochrome c oxidase. The genome also suggests potential for nitrate reduction, although this has not been observed under laboratory conditions.
https://www.nature.com/articles/32831

Oxidative Stress Response: Given its gowth in microaerophilic conditions, Aquifex aeolicus possesses enzymes to counter reactive oxygen species, like superoxide and peroxide. This is critical for survival in environments where fluctuations in oxygen levels and exposure to harsh conditions can cause oxidative stress.

Biotechnological and Evolutionary Significance: Due to its unique adaptations to extreme conditions and its primitive position in the tree of life, Aquifex aeolicus is of significant interest in evolutionary biology and biotechnology. Its enzymes and metabolic pathways offer insights into early life forms and potential applications in industrial processes that require high-temperature operations.

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?

Yes, when running mOTUs with a more stringent threshold for marker-gene detections, specifically requiring at least 4 genes, I still detect the same organism, Aquifex aeolicus. In this case, its estimated relative abundance increases to 0.9154098245. This result is consistent and expected under the conditions of a more restrictive threshold. By increasing the number of marker genes required for detection, the analysis becomes more stringent, potentially excluding organisms that would be identified under a less strict threshold. Consequently, this heightened stringency reinforces the dominance of Aquifex aeolicus in the sample, as it is the most prevalent

organism. The increase in its relative abundance under these conditions underscores its robust presence in the high-temperature environment of the sample.

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

In the normal-temperature samples, Aquifex aeolicus remains the most abundant organism, although its relative abundance is significantly lower at 0.0333818811. This decrease in relative abundance compared to the high-temperature samples is logical and expected. In the normal-temperature conditions, a wider range of species can survive and be present. As a result, the microbial community becomes more diverse, with many species present at lower abundances.

This observation aligns with ecological principles where extreme environments, like high temperatures, often select for a few highly adapted species, leading to lower diversity but higher abundance of those few species. In contrast, more moderate conditions, such as normal temperatures, allow for a broader range of species to coexist, each with a smaller fraction of the total population. Consequently, even though Aquifex aeolicus is still the most abundant in the normal-temperature samples, its dominance is less pronounced due to the increased presence and competition of other microbial species.

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

Yes, Aquifex aeolicus is also present in the high-temperature samples. In fact, it is not just present but significantly more abundant in these samples compared to the normal-temperature samples. The relative abundance of Aquifex aeolicus in the high-temperature samples is remarkably high at 0.9154098245.

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

The condition with a greater level of alpha biodiversity is the normal temperature environment. The alpha biodiversity at normal temperature is significantly higher with a count of 191, compared to the high-temperature samples which have an alpha biodiversity of only 10.

This substantial difference in alpha biodiversity can be attributed to the environmental conditions. High-temperature environments, like those experienced in the hot spring samples, are typically more extreme and can only be tolerated by a limited number of highly specialized organisms. Aquifex aeolicus, being an extremophile, thrives in such conditions, leading to its dominance but a lower overall biodiversity.

In contrast, normal temperature conditions are less extreme and can support a wider variety of organisms. This allows for more species to coexist, leading to a higher species richness and a more balanced ecosystem in terms of species distribution. Consequently, the normal temperature samples exhibit a significantly higher alpha biodiversity, reflecting a more diverse and rich microbial community.

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

I did not detect any eukaryotic algae in the high-temperature sample. However, in the normal temperature sample, I did detect the presence of Cyanobacteria, which are photosynthetic prokaryotic organisms. The Cyanobacteria in the normal temperature sample had a relatively low abundance, representing approximately 1.53% of the microbial community.

The absence of eukaryotic algae in both samples could be attributed to the specific environmental conditions of the hot spring. The extreme conditions, including high temperature and pH, might not be conducive to the growth of eukaryotic algae. Bacteria may be better adapted to such conditions and could outcompete eukaryotic algae for resources. This competitive advantage of bacteria adapted to the harsh environment could explain the absence of eukaryotic algae in the samples.

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…

Questions

2.1 How many samples do you have?

There are four samples in the dataset. These samples are named hightemp01, hightemp02, normal01, and normal02.

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

In the hightemp01 sample, there are 318,719 reads in each of the forward and reverse files, summing up to 318,719 paired-end reads.
The hightemp02 sample also comprises 318,719 reads in both the forward and reverse files, totaling 318,719 paired-end reads.
The normal01 sample contains 288,742 reads in each file, leading to a total of 288,742 paired-end reads.
Similarly, the normal02 sample consists of 288,742 reads in each file, which amounts to 288,742 paired-end reads.

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

In my RNA-seq samples, the nature of the reads clearly indicates that they are paired-end. This conclusion is drawn from examining the file names and the structure of the reads themselves. For each sample, there are two separate files: one labeled with '.r1.fq' and the other with '.r2.fq'. The '.r1.fq' files contain the forward reads, and the '.r2.fq' files contain the reverse reads.
For instance, in the sequencing output, the read names follow a pattern like @AQUIFEX_00001_1301_1575_0:0:0_0:0:0_0/1 for forward reads and @AQUIFEX_00001_1301_1575_0:0:0_0:0:0_0/2 for reverse reads. The '/1' and '/2' at the end of these read names explicitly designate them as the first and second reads of a pair, respectively.

2.4 Are all the reads of the same length?

All the reads in my samples are of the same length. This conclusion is drawn from the output of the script which checked the length of reads. For both forward and reverse reads in all samples (hightemp01, hightemp02, normal01, normal02), the read length is consistently 100 nucleotides. This uniformity in read length is typical for Illumina sequencing platforms, which often generate reads of a standard length, depending on the sequencing setup used.

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’?

Based on the file names alone (hightemp01.r1.fq, hightemp02.r1.fq, normal01.r2.fq, normal02.r2.fq, etc.), there are certain assumptions I can make about the orientation of the reads in my RNA-seq samples, though these are not definitive without further information:

Forward or Reverse: The 'r1' and 'r2' in the file names typically suggest paired-end sequencing, where 'r1' is often used to denote forward reads and 'r2' for reverse reads. However, this convention is not universally applied, and without additional details on the sequencing protocol, I cannot confirm with certainty that 'r1' and 'r2' correspond to forward and reverse reads in this context.

Coding or Template Strand: The orientation of the reads with respect to the coding or template strand cannot be determined from the file names. This detail depends on the RNA sequencing library preparation method, particularly whether the library was strand-specific or not.

5’ to 3’ or 3’ to 5’ Orientation: In general, RNA-seq reads are sequenced in the 5' to 3' direction, aligning with the natural synthesis and processing direction of RNA. It is likely that the reads in my samples are oriented in this manner. However, without specific information on the sequencing method used, this remains an assumption.

To summarize, while the file names suggest a standard paired-end sequencing format and a likely 5' to 3' orientation of reads, the exact details regarding forward/reverse and coding/template strand orientation cannot be confirmed without additional information about the RNA library preparation and sequencing protocol used.

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

Upon analyzing the RNAseq samples with FastQC, I observed a uniformly low quality across all bases with Phred scores averaging around 9. This is unusual as Illumina data typically shows high-quality scores at the beginning of reads with a decline towards the end. The consistently poor quality across the entire length of the reads is concerning and unexpected, especially since the read length is consistent at 100 bases and there's a perfect forward and reverse read count match. These latter points generally suggest good sequencing practices and reliable data. This could be due to the fact that it is not real data.

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 ...

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.

In my RNA-seq analysis, the data shows a high degree of mapping efficiency, with each read in the FastQ files being successfully aligned to the reference genome. This is evident from the equal numbers of mappings and reads in the BAM files, which indicates that each read has been uniquely aligned without any secondary or supplementary alignments. Here are the specifics:

Number of Mappings in BAM Files:

hightemp01.bam: 637,438 mappings
hightemp02.bam: 637,438 mappings
normal01.bam: 577,484 mappings
normal02.bam: 577,484 mappings
These numbers are identical to the number of paired-end reads in each BAM file, suggesting that every read was mapped exactly once.

Number of Different Reads in BAM Files:

The number of different paired-end reads (counted as read pairs) in each BAM file is the same as the number of mappings. This one-to-one correspondence indicates that there were no reads that mapped to multiple locations or that were improperly paired.
Comparison of Mappings and Different Reads:

Since the number of total mappings is equal to the number of different reads, there is no discrepancy in this case. Every read in the FastQ files has been uniquely mapped to the reference genome.

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

While my RNA-seq data mappings show high efficiency, they are generally not suitable for Copy Number Variation (CNV) analysis. CNV analysis is typically performed using DNA-level data rather than RNA, as it involves detecting changes in the number of copies of specific genomic regions. RNA-seq data reflects the abundance of transcripts, which is influenced by transcriptional activity and RNA stability, and does not directly indicate DNA copy numbers.

Additionally, most established methods and studies focusing on CNV analysis, especially in complex genomes such as the human genome, utilize DNA sequencing or DNA-specific array techniques. These methods are tailored to detect variations in DNA copy number, which is not feasible with RNA-seq data. The complexity and size of genomes like the human genome require the resolution and specificity that DNA-based methods provide, which RNA-seq cannot offer for CNV analysis.

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...

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?

For the Variant Calls from Merged BAM with Specified Ploidy (--ploidy 1):
Expected Variants: I anticipated a moderate number of variants due to the organism's adaptation to different temperatures.
Actual Detected Variants: I detected 2 variants.
Type of Variants: Both are SNPs (Single Nucleotide Polymorphisms), with changes from T to A at positions 1265734 and 1265735 on the Aquifex genome.

For the Variant Calls from Merged BAM with Default Ploidy:
Expected Variants: Similar to the specified ploidy, but potentially identifying more heterozygous variants due to the default diploid assumption.
Actual Detected Variants: I found 3 variants.
Type of Variants: All are SNPs, with one occurring at position 1265060 (C to T change) and two others (T to A changes) at positions 1265734 and 1265735.

For the Variant Calls from Individual Sorted BAMs with Specified Ploidy (--ploidy 1):
Expected Variants: Similar to the first analysis but expecting some sample-specific variations.
Actual Detected Variants: Detected 2 variants.
Type of Variants: Both are SNPs, occurring at positions 1265734 and 1265735 (T to A changes).

For the Variant Calls from Individual Sorted BAMs with Default Ploidy (not asked in the exercise)**:**
Expected Variants: Anticipated a few more variants due to the default diploid setting.
Actual Detected Variants: Detected 3 variants.
Type of Variants: All are SNPs, with one at position 1265060 (C to T change) and two at positions 1265734 and 1265735 (T to A changes).

In all cases, the variants detected are exclusively SNPs, with no insertions or deletions identified. The number of variants is relatively modest, which could indicate either a low level of genomic variation under the conditions studied or the high specificity of the variant calling process.

3.2 How many variants have quality greater than 100?

Based on the variant quality analysis I conducted using bcftools filter, here are the results for the number of variants with a quality score greater than 100 in each of my VCF files:

Variant Calls from Merged BAM with Specified Ploidy (--ploidy 1): I found 0 variants with a quality greater than 100.

Variant Calls from Merged BAM with Default Ploidy: In this file, I identified 1 variant with a quality score exceeding 100.

Variant Calls from Individual Sorted BAMs with Specified Ploidy (--ploidy 1): Similar to the first

analysis, there were 0 variants with a quality greater than 100.
Variant Calls from Individual Sorted BAMs with Default Ploidy: Here, I also found 1 variant with a quality score above 100.

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

Based on the depth of coverage analysis I performed using bcftools filter in each of my VCF files:

Variant Calls from Merged BAM with Specified Ploidy (--ploidy 1): In this file, there were 0 variants with a depth of coverage greater than 100.

Variant Calls from Merged BAM with Default Ploidy: I identified 1 variant with a depth of coverage exceeding 100.

Variant Calls from Individual Sorted BAMs with Default Ploidy: Similarly, in this analysis, I found 1 variant with a depth of coverage above 100.

Variant Calls from Individual Sorted BAMs with Specified Ploidy (--ploidy 1): There were 0 variants with a depth of coverage greater than 100 in this file.

These findings suggest that most of the variants detected in my analyses have lower coverage depths, with a few exceptions in the default ploidy analyses. Depth of coverage is a crucial factor in variant analysis, as higher depth can provide greater confidence in the detected variants.

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?

When comparing the variant calling outputs from the merged data with and without --ploidy 1 and the results from using 4 independent samples, several differences are observed:
Merged BAM with --ploidy 1: Detected 2 SNPs. Both involve a transition from T to A, occurring at positions 1265734 and 1265735 on the Aquifex genome.

Merged BAM without --ploidy 1 (Default Ploidy): Detected 3 SNPs. The changes include one C to T transition at position 1265060 and two T to A transitions at positions 1265734 and 1265735.

Independent BAMs with --ploidy 1: Detected 2 SNPs, similar to the merged BAM with --ploidy 1, with the same T to A transitions at positions 1265734 and 1265735.

Independent BAMs without --ploidy 1 (Default Ploidy): Detected 3 SNPs, the same as the merged BAM without --ploidy 1. These are one C to T transition at position 1265060 and two T to A transitions at positions 1265734 and 1265735.

Differences Observed:

The key difference is in the number of variants detected, particularly between the runs with --ploidy 1 and the default ploidy setting.
Using the default ploidy setting resulted in detecting an additional variant in both the merged and independent samples analysis.

Cause of These Differences:

The --ploidy 1 setting assumes a haploid genome, where each genomic position has only one allele. This assumption can lead to calling only homozygous variants.
Without --ploidy 1, bcftools assumes a default diploid genome, allowing the detection of heterozygous variants. This can increase the number of detectable variants, as seen in the results.
Advantages and Disadvantages of Merged vs. Independent Samples:

Merged Samples:
Advantages: Provides a consolidated view of variants across all samples, which can be useful for identifying common variants or trends.
Disadvantages: Potentially misses sample-specific variants and can be influenced by the most dominant sample in the dataset.
Independent Samples:
Advantages: Allows for the identification of sample-specific variants and gives a more nuanced understanding of each sample's genomic landscape.
Disadvantages: More computationally intensive and can result in a larger number of variants to sift through, some of which might be noise.
In summary, the choice between using merged or independent samples for variant calling depends on the research objective. Merged samples provide a broader, more general view, while independent sample analysis offers a detailed, sample-specific insight. The choice of ploidy significantly impacts the variant calling results, particularly in detecting heterozygosity.

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? 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).

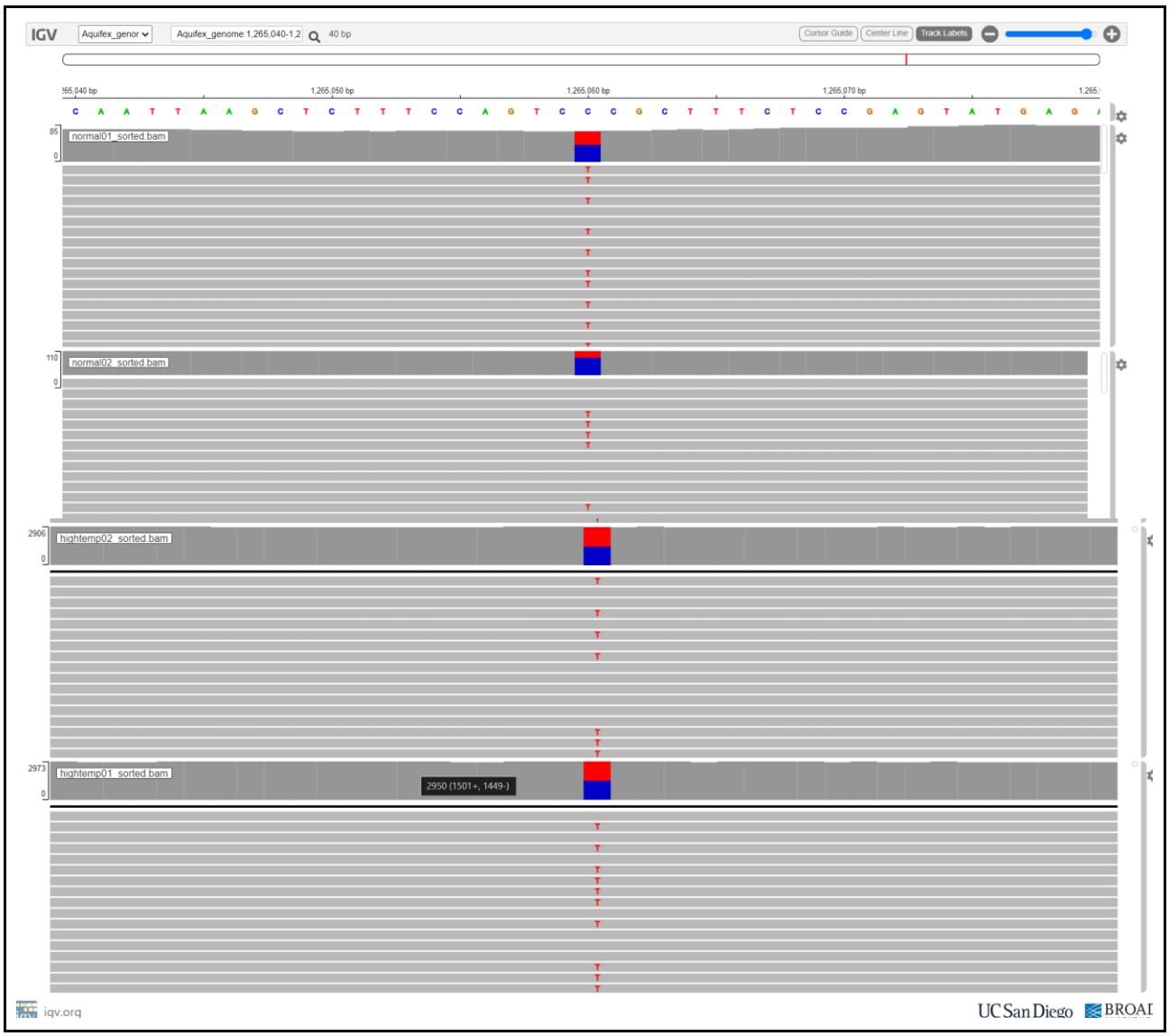
Regarding the variant with the highest quality score in my variant calling analysis, the variant found at position 1265060 on the Aquifex_genome in both the variant_calls_default_ploidy.vcf (with a quality score of 188) and variants_sep.vcf (with a quality score of 950) appears to be heterozygous. This is indicated by the genotype notation 0/1 in the VCF files, suggesting that one copy of the chromosome carries the variant while the other does not.

Regarding its potential impact on a gene, the variant falls within the boundaries of the nifA gene (gene ID: AQUIFEX_01423), which encodes for the Nif-specific regulatory protein. This information was revealed through an analysis of the genome GFF file, showing that the variant's position is between the start and end coordinates of the nifA gene (1264237 to 1265730).
This variant is located within the coordinates of the gene nifA (ID: AQUIFEX_01423), which encodes a Nif-specific regulatory protein. The presence of this variant within nifA raises the possibility that it could impact the gene's function, potentially affecting the nitrogen fixation process.

If this variant were to cause a change in the amino acid sequence (non-synonymous mutation) of the Nif-specific regulatory protein, it could alter the protein's ability to regulate the nitrogen fixation pathway, which is vital for the organism's nitrogen processing. Even a synonymous mutation could potentially affect mRNA stability or translation efficiency, influencing the protein abundance.
Given the context of the exercise (from the source of the samples) it is conceivable that alterations in nitrogen fixation due to genetic variations in the local microbial populations could influence the nutrient dynamics of the ecosystem. For example, a variant in nifA that increases the efficiency of nitrogen fixation in bacteria could lead to increased nitrogen availability in the

thermal source, favoring the survival of cyanobacteria. Conversely, a variant that decreases the efficiency of nitrogen fixation could reduce the available nitrogen, which could affect the survival of these cyanobacteria.

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.



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.

Questions

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

The "deseq2_results_padj_0.05.csv" file generated from the DESeq2 R script contains the results of the Differential Expression Analysis. Each row represents a gene, and the columns provide relevant statistical and biological information:

Row Number: A number representing the row order, essentially a row counter.
"ensembl_gene_id": The unique identifier for each gene, which in this case appears to be an internal identifier like "AQUIFEX_01423".
"baseMean": The average of normalized counts for a gene across all samples. It represents the average RNA abundance of the gene.
"log2FoldChange": The change in gene expression between experimental conditions, measured as the logarithm to base 2 of the fold change. A positive value indicates higher expression in the high-temperature condition, and a negative value indicates higher expression in the normal condition.
"lfcSE" (log Fold Change Standard Error): The standard error of the logarithm to base 2 of the fold change.
"stat": The Wald statistic for the test of significance of expression change.
"pvalue": The p-value of the statistical test, indicating the probability of observing an effect as extreme or more extreme than what is observed, under the null hypothesis of no change in expression.
"padj" (Adjusted p-value): The p-value adjusted to control the false discovery rate, correcting for multiple testing.
"sig" (Significance): An indication of whether the gene is considered significantly different in its expression ("yes" if the adjusted p-value is less than or equal to 0.05, "no" otherwise).

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?

Based on the analysis using DESeq2, a total of 5 genes were identified as differentially expressed genes (DEGs) with an adjusted p-value (p-adj) less than 0.05. This implies a significant change in their expression levels under different temperature conditions.
Among these DEGs:

AQUIFEX_01423, AQUIFEX_01759, and AQUIFEX_01761 have positive log2 fold changes of 5.0353, 5.0320, and 5.0356, respectively. These positive values indicate that these genes are up-regulated under high-temperature conditions compared to normal conditions. In other words, they are expressed at higher levels when the temperature is elevated.

On the other hand, AQUIFEX_01723 and AQUIFEX_01749 have negative log2 fold changes of -

10.5771 and -9.6811, respectively. These negative values suggest that these genes are down-regulated under high-temperature conditions. This means their expression levels are reduced in response to increased temperatures.
In summary, the differential expression analysis reveals that certain genes are significantly more active, while others are less active, in high-temperature environments. This insight could be crucial for understanding the biological adaptations and responses of organisms in fluctuating thermal conditions.

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

Gene ID: "AQUIFEX_01423"
Function: Nif-specific regulatory protein

Gene ID: "AQUIFEX_01759"
Function: FeMo cofactor biosynthesis protein NifB

Gene ID: "AQUIFEX_01761"
Function: Nitrogenase iron protein 1

Gene ID: "AQUIFEX_01723"
Function: hypothetical protein

Gene ID: "AQUIFEX_01749"
Function: hypothetical protein

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?

Based on the results obtained from the differential expression analysis and the subsequent extraction of gene functions, it appears that there is a common theme of nitrogen-related processes being differentially regulated under the environmental conditions of the study. Specifically, three of the differentially expressed genes are directly associated with nitrogen processing or regulation:

AQUIFEX_01423: This gene encodes a Nif-specific regulatory protein, suggesting a role in nitrogen fixation. It is up-regulated under high-temperature conditions, indicating enhanced nitrogen fixation activity in response to increased temperatures.

AQUIFEX_01759: This gene is responsible for encoding the FeMo cofactor biosynthesis protein NifB, which is crucial for nitrogenase function in nitrogen fixation. Similar to AQUIFEX_01423, it is also up-regulated, suggesting an increased demand for nitrogenase cofactors at higher temperatures.

AQUIFEX_01761: This gene encodes the Nitrogenase iron protein 1, a key component of the nitrogenase enzyme complex. Its up-regulation under high-temperature conditions further supports the idea of increased nitrogen fixation activity.

The other two DEGs, AQUIFEX_01723 and AQUIFEX_01749, are annotated as hypothetical proteins, and their specific roles are not clear. However, they are down-regulated in the high-temperature samples. While their exact functions are unknown, this opposite regulation pattern could indicate a complex adaptive response to the environmental stress.
In conclusion, the common process that seems to be differentially regulated is nitrogen fixation, a critical pathway for converting atmospheric nitrogen into a form that can be used by living

organisms. The up-regulation of genes involved in nitrogen fixation and related processes suggests an adaptive response of the organisms to high-temperature conditions, possibly to meet increased metabolic demands or to cope with environmental stress.

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?

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.

Functional description, protein domain and gene name I got from Uniprot minus entering the fasta file for each gene.

As for the KEGG pathways, I got the information from KEGG with the fasta file for each one, except for AQUIFEX_01761, which I got the information from NCBI.

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	Transcriptional regulator (NtrC family)	Sigma-54 factor interaction	Pathway: Two-component system (Hydrogenobacter thermophilus (hte02020))	ntrC2
AQUIFEX_01759	Catalytic activity, nitrogen fixation	Radical Sam protein	FeMOCofactor biosynthesis (UniProt)	FeMo cofactor biosynthesis protein NifB Ordered locus names MMP0658
AQUIFEX_01761	Oxidoreductase, Nitrogen fixation	Nitrogenase iron protein, nitrogenase	NCIB: Chloroalkane and chloroalkane	nifH

		component II NifH	degradation Nitrogen metabolism Metabolic pathways Microbial metabolism in diverse environments Module: Nitrogen fixation, nitrogen => ammonia	

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

The gene ntrC2 interacts with rpoN, trpC. MMP0658 interacts with NifH-2, NifE, NifN, NifK. NifH appears as a homodimer

In examining the functional roles of MMP0658 and nifH genes, both are implicated in the nitrogen fixation process, suggesting a functional relationship. Nitrogen fixation is a critical biological process, particularly in extreme environments like hot springs, where nitrogenous compounds are essential for the survival of various organisms. In such ecosystems, the ability to fix atmospheric nitrogen not only supports the survival of the bacteria themselves but also facilitates the growth of dependent organisms, such as algae, which rely on a consistent nitrogen supply for their growth and reproduction. The cooperative activity of these genes likely contributes to the nitrogen cycle in this niche, underpinning a symbiotic relationship where the fixation process by bacteria could potentially support a thriving algal community post high-temperature episodes.

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

The genes MMP0658 and nifH, central to the nitrogen fixation pathway, could be integral to the algal bloom events observed. Post-high temperature episodes, the fixation of atmospheric nitrogen by these genes enhances the availability of essential nutrients, particularly ammonia and other nitrogenous compounds. This sudden influx of bioavailable nitrogen can act as a fertilizer, promoting rapid algal growth, leading to blooms. Therefore, the upregulation of these genes in response to thermal stress could trigger a cascade that culminates in the proliferation of algae, highlighting a direct ecological link between bacterial gene expression and algal population dynamics in the hot spring ecosystem.

6. Phylogenetic and comparative analysis

The functional analysis 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.

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?

Based on my phylogenetic analysis, the closest orthologs of the overexpressed genes are as follows:

For AQUIFEX_01423, the closest ortholog is tr|A0A497XW95 from Hydrogenivirga caldilitoris (9AQUI).

For AQUIFEX_01759, the closest ortholog is D7DSI7|D7DSI7| METV3

For AQUIFEX_01761, the closest ortholog is A0A1D9N950 and sp|P09553|NIFH3 from Clostridium Pasterianus (CLOPA).

6.2 Do orthology assignments support your previous functional annotations? Briefly describe why.

For the orthologue of AQUIFEX_01423, it has the same functions as the gene, i.e. a transcriptional regulator.
For the orthologue of AQUIFEX_01759, SI7 METV3 is an FeMo cofactor so it seems to present the same function as its homologue.
For the orthologue of AQUIFEX_01761, A0A1D9N950 and sp|P09553|NIFH3 are nitrogen fixers. They were the first genes with this function discovered.

6.3 Are all the overexpressed genes present in the public genome of your organism?

To assess the presence of the overexpressed genes in the public genome of my organism, I conducted a BLAST search to compare their sequences with the public database, focusing on the percentage of similarity to evaluate their presence. The results were quite revealing:

- The gene AQUIFEX_01423, which encodes ntrC2, showed a 100% match with Aquifex aeolicus in the database, indicating its complete presence in the public genome of the species. This suggests that the gene is a conserved element within the species, including in the public genome.
- For AQUIFEX_01759, associated with MMP0658, the highest similarity found with Aquifex aeolicus in the database was only 23.38%. This low level of similarity suggests that, while there might be a distant relation, the specific gene variant I identified through my research might not be present or is significantly different in the public genome.
- Regarding AQUIFEX_01761, which codes for nifH, no part of this gene was found in the public database for Aquifex aeolicus. This absence could be attributed to the unique environmental adaptations of the organism, which may not be shared with those strains of Aquifex aeolicus that have been publicly sequenced. It's noteworthy that this gene's absence might be related to its specific adaptation to high-temperature environments, which might not be a common trait across all strains of Aquifex aeolicus.

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

After examining the phylogenetic trees and considering the distances between branches and their respective homologs, it's clear that the closest orthologs offer insight into the probable origins of each gene. The evolutionary distances observed in the trees, coupled with the identified closest orthologs, point to:

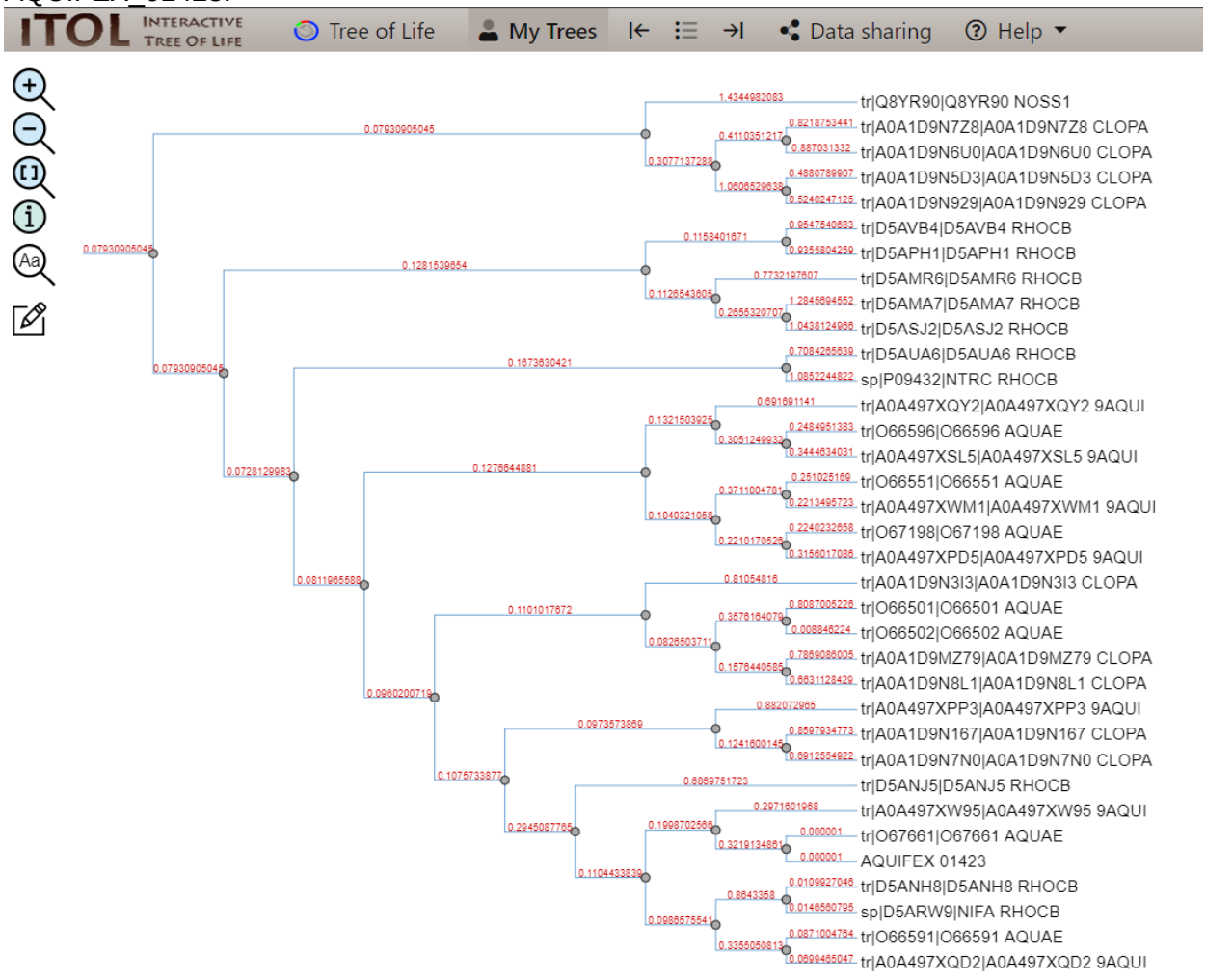
- AQUIFEX_01423 likely originates from a lineage shared with Hydrogenivirga caldilitoris, suggesting a conserved role across these organisms.
- AQUIFEX_01759 shows a connection to Methanococcus voltae, indicating a metabolic function essential across diverse taxa.
- AQUIFEX_01761's closest relationship with Clostridium pasteurianum underscores a shared evolutionary pathway for nitrogen fixation.

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

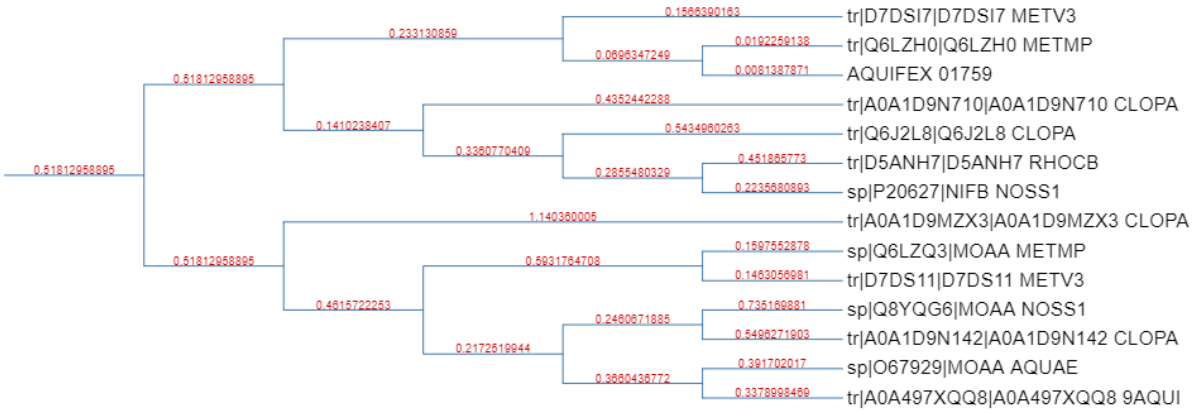
The overexpressed genes in Aquifex aeolicus—AQUIFEX_01423, AQUIFEX_01759, and AQUIFEX_01761—play distinct roles that collectively contribute to the organism's adaptation to extreme environments and interaction with its ecosystem. AQUIFEX_01423, identified as a transcriptional regulator, likely modulates gene expression in response to environmental stress, including high temperatures. AQUIFEX_01759, involved in metal transport, could facilitate the acquisition of essential nutrients under thermal stress, supporting cellular functions and growth. AQUIFEX_01761, a gene related to nitrogen fixation, directly impacts the nitrogen cycle, enhancing nitrogen availability for algae blooms. Each gene, by fulfilling specific physiological roles, not only underscores the organism's resilience but also its integral part in nurturing microbial communities, especially in fostering conditions that lead to algae proliferation post-thermal episodes.

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

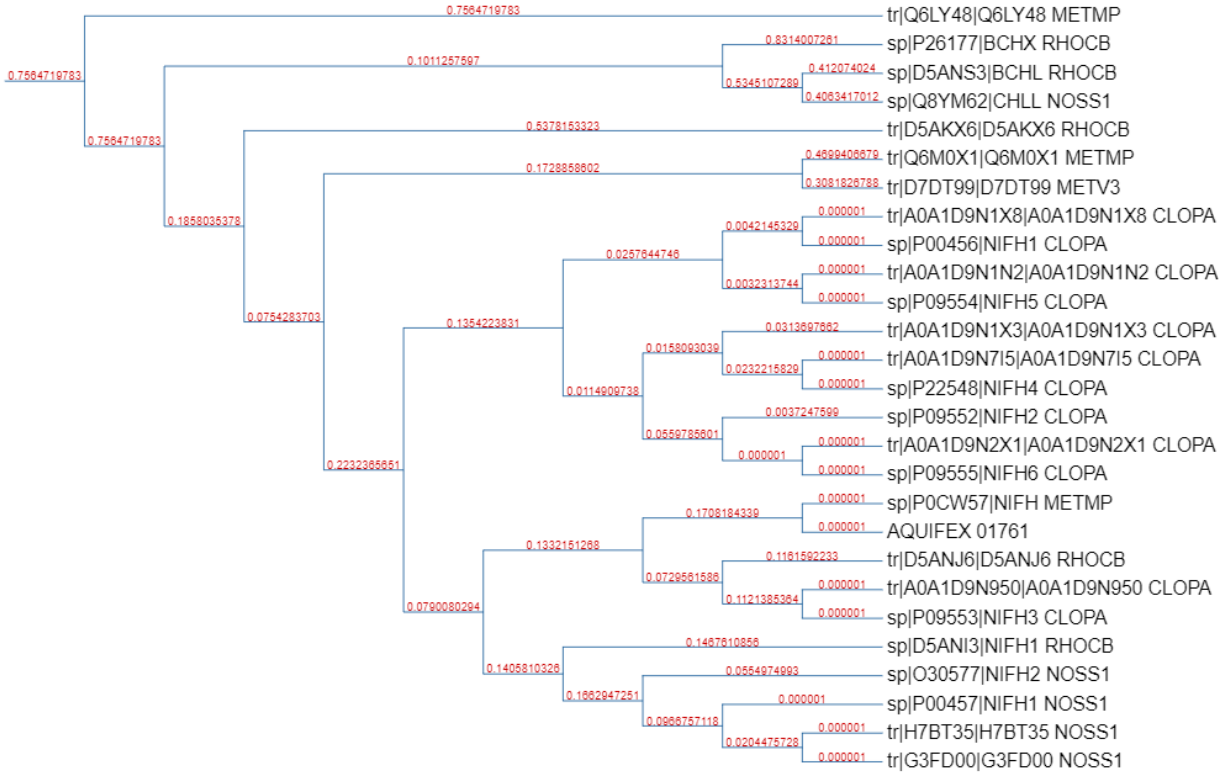
AQUIFEX_01423:



AQUIFEX_01759:



AQUIFEX_01761:



7. Conclusion

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.

The metagenomic analysis highlighted the dominance of Aquifex aeolicus during high-temperature episodes, which, combined with the differential expression analysis, revealed the significant upregulation of genes involved in nitrogen fixation and metabolism under thermal stress. This genetic adaptation likely supports the organism's survival and functionality in extreme conditions, facilitating a microbial environment conducive to algal blooms post-temperature elevation. Phylogenetic analysis further suggested evolutionary connections between these key genes and their homologs across different taxa, reinforcing their essential roles in environmental adaptation.

So, based on the comprehensive analyses conducted on the microbial community and genetic adaptations in the hot spring, it appears that specific genes have evolved in response to the extreme environmental conditions, particularly high temperatures. These genetic adaptations, such as those observed in AQUIFEX_01423, AQUIFEX_01759, and AQUIFEX_01761, not only confer survival advantages to individual organisms like Aquifex aeolicus but also influence the broader microbial ecosystem. For example, the nitrogen-fixing abilities highlighted in this findings suggest a crucial role in supporting the nutrient cycles that underpin the hot spring's ecological balance.
This interaction between genetic adaptations and environmental pressures underscores a dynamic evolutionary process, shaping the microbial community structure and its functional capabilities.
So, my hypothesis posits that these genetic traits contribute significantly to the resilience and productivity of the hot spring ecosystem, enabling it to thrive despite the challenging conditions. This insight into microbial adaptation and community dynamics offers a window into the complexity of life in extreme environments and the intricate interplay between genes, organisms, and their habitats.
Furthermore, the absence of the AQUIFEX_01761 gene in public databases of Aquifex aeolicus, as revealed in the phylogenetic analysis, suggests unique adaptations of this organism to its environment not shared with other strains. This specificity could stem from adaptations to high-temperature habitats, indicating a unique evolutionary path for strains thriving under such conditions, distinct from the general Aquifex aeolicus population. This highlights the intricate relationship between genetic diversity and environmental adaptation within microbial communities.