Q1) How many total sample files do we have?

A1: There are two ways to do this:

- List and count the number of *.fasta files we have in our folder

    -
      ```
      ls *.fasta | wc
      ```
- Count the lines in the metadata file. Each row in the metadata file corresponds to one sample. Remember to subtract 1 as the first row contains the headers

    -
      ```
      wc metadata-file-for-osd-subset-210615.txt
      ```

Q2) How many sequences does the sample from OSD station 10 contain?

A2: Since we are dealing with fasta files we can do this by counting the number of “>” characters in the file using the grep command

-
  ```
  grep -c “>” OSD10.comb.qc.masked.dedup.subsample.fasta
  ```

Q3) How many samples of each type are there in each of the different Province code categories?

A3: Again we can use the metadata file for this. The prov_code is in column 12 in the file and we can use the sort and uniq commands to to get a count of each type.

-
  ```
  cut -f 12 metadata-file-for-osd-subset-210615.txt | sort |uniq -c
  ```

Q4) In the STAMP analysis of the Metaphlan results, do you see any separation in the samples when the PCA is coloured by Depth?

A4: No.

Q5) Do you see any separation in the samples when the PCA is coloured by the province codes? If so, describe which PC axis differentiates these samples.

A5: Yes, there is a slight separation between the Arctic samples and the Northwest Atlantic samples. The separation is seen in the PC1 v/s PC2 and occurs partly along PC1.

Q6) In a “multiple group test” using ANOVA with no multiple test correction how many genera are statistically significant?

A6: 2

Q7) How many are still significant in the “two group test” using White's non-parametric t-test without and with Benjamini-hochberg FDR for multiple test correction?

A7: Without Benjamini-Hochberg FDR – 22

With – 20

Q8) What are the top 3 Modules present in the 1m sample from the Bedford basin (station 152)?

A8: You can find this out from the modules.spf file that we generated.

You can view the modules.spf file using 'less' and manually count the column number where the OSD152 1m sample occurs. In this case it is 13.

```
less modules.spf
```

Now we will cut out that column (which contains the abundances pf the modules for OSD152 1m) and sort it based on the abundance values and list out the top 10 most abundant values

```
cut -f 13 modules.spf | sort -nr |head
```

Next we will take the top most abundant value (8.9726e-05 in this case) and grep the 1st column (which contains the pathway name) using this value

```
cut -f 1,13 modules.spf |grep 8.9726e-05
```

You can do the same for the top 3 abundance values. The corresponding top 3 module names are as follows:

M00185: Sulfate transport system

M00039: Lignin biosynthesis, cinnamate => lignin

M00235: Arginine/ornithine transport system

Q9) In the STAMP analysis of the Humann results using a two group test with no multiple test correction applied how many significant differences are seen between the Arctic and Northwest Atlantic samples?

A9: 919

Q10) What happens when the p-value cut-off is lowered to 0.01 for Q9?

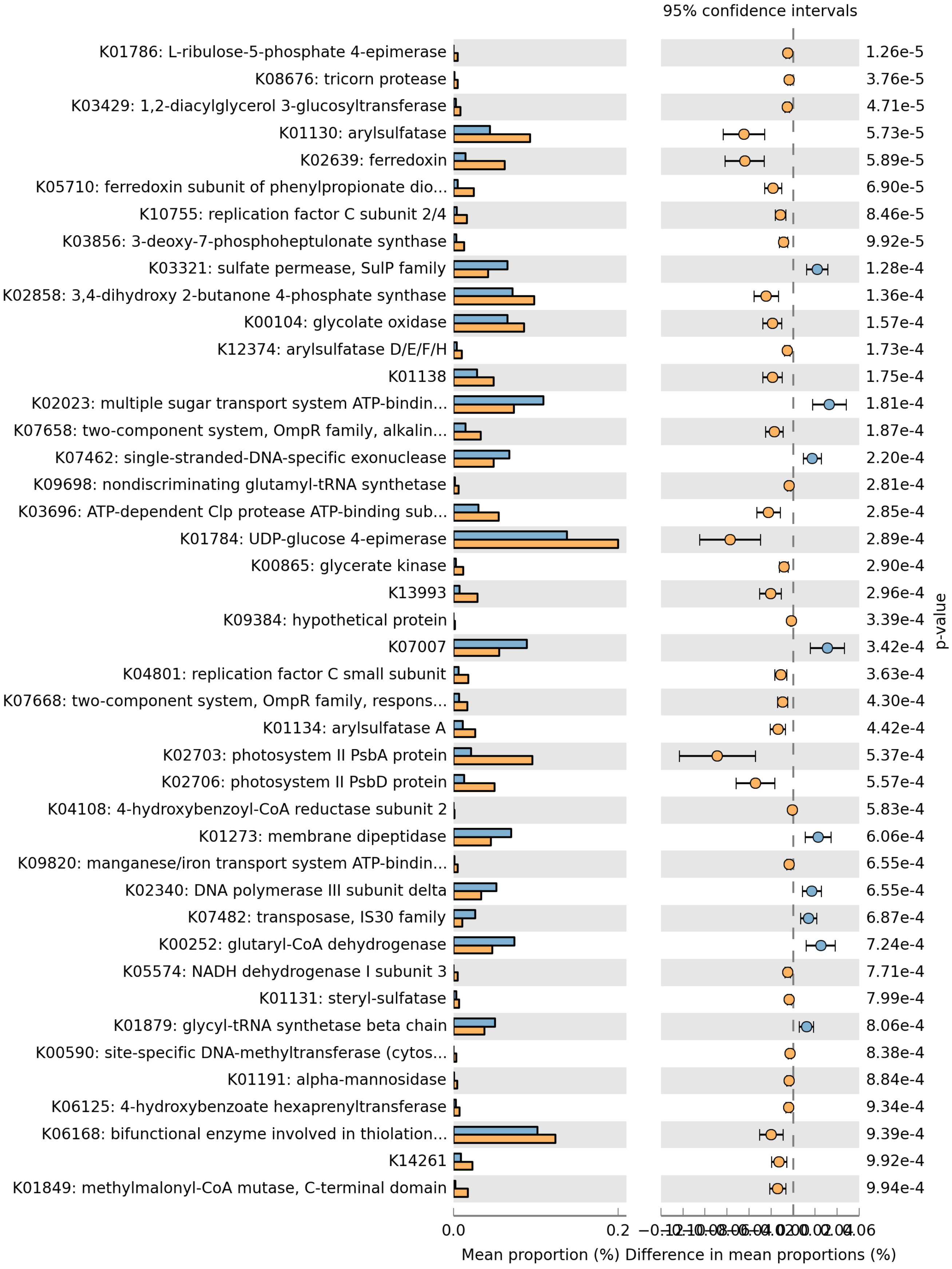
A10: The significantly different modules decrease to 272

Q11) What is the most significantly different KEGG orthology group? What is the p-value for this KO?

A11: K01786: L-ribulose-5-phosphate 4-epimerase . P-value: 1.26e-5

Q12) Change the p-value to 0.001 and create an “Extended error bar” plot and save the image as a .png using the File->Save Plot option.

A12:


95% confidence intervals
K01786: L-ribulose-5-phosphate 4-epimerase
K08676: tricorn protease
K03429: 1,2-diacylglycerol 3-glucosyltransferase
K01130: arylsulfatase
K02639: ferredoxin
K05710: ferredoxin subunit of phenylpropionate dio...
K10755: replication factor C subunit 2/4
K03856: 3-deoxy-7-phosphoheptulonate synthase
K03321: sulfate permease, SulP family
K02858: 3,4-dihydroxy 2-butanone 4-phosphate synthase
K00104: glycolate oxidase
K12374: arylsulfatase D/E/F/H
K01138
K02023: multiple sugar transport system ATP-bindin...
K07658: two-component system, OmpR family, alkalin...
K07462: single-stranded-DNA-specific exonuclease
K09698: nondiscriminating glutamyl-tRNA synthetase
K03696: ATP-dependent Clp protease ATP-binding sub...
K01784: UDP-glucose 4-epimerase
K00865: glycerate kinase
K13993
K09384: hypothetical protein
K07007
K04801: replication factor C small subunit
K07668: two-component system, OmpR family, respons...
K01134: arylsulfatase A
K02703: photosystem II PsbA protein
K02706: photosystem II PsbD protein
K04108: 4-hydroxybenzoyl-CoA reductase subunit 2
K01273: membrane dipeptidase
K09820: manganese/iron transport system ATP-bindin...
K02340: DNA polymerase III subunit delta
K07482: transposase, IS30 family
K00252: glutaryl-CoA dehydrogenase
K05574: NADH dehydrogenase I subunit 3
K01131: steryl-sulfatase
K01879: glycyl-tRNA synthetase beta chain
K00590: site-specific DNA-methyltransferase (cytos...
K01191: alpha-mannosidase
K06125: 4-hydroxybenzoate hexaprenyltransferase
K06168: bifunctional enzyme involved in thiolation...
K14261
K01849: methylmalonyl-CoA mutase, C-terminal domain
1.26e-5
3.76e-5
4.71e-5
5.73e-5
5.89e-5
6.90e-5
8.46e-5
9.92e-5
1.28e-4
1.36e-4
1.57e-4
1.73e-4
1.75e-4
1.81e-4
1.87e-4
2.20e-4
2.81e-4
2.85e-4
2.89e-4
2.90e-4
2.96e-4
3.39e-4
3.42e-4
3.63e-4
4.30e-4
4.42e-4
5.37e-4
5.57e-4
5.83e-4
6.06e-4
6.55e-4
6.55e-4
6.87e-4
7.24e-4
7.71e-4
7.99e-4
8.06e-4
8.38e-4
8.84e-4
9.34e-4
9.39e-4
9.92e-4
9.94e-4
p-value
0.0
0.2
−0.12−0.10−0.08−0.06−0.04−0.02 0.00 0.02 0.04 0.06
Mean proportion (%)
Difference in mean proportions (%)