Metagenomics

So Crescent cluster was decided to carry out the practicals, and thus the first task was to make the subfolders; data, scripts, tools, results and resources and then install the SRA-toolset and to configure it (as modules FastQC, MultiQC, and QIIME2 is already present).

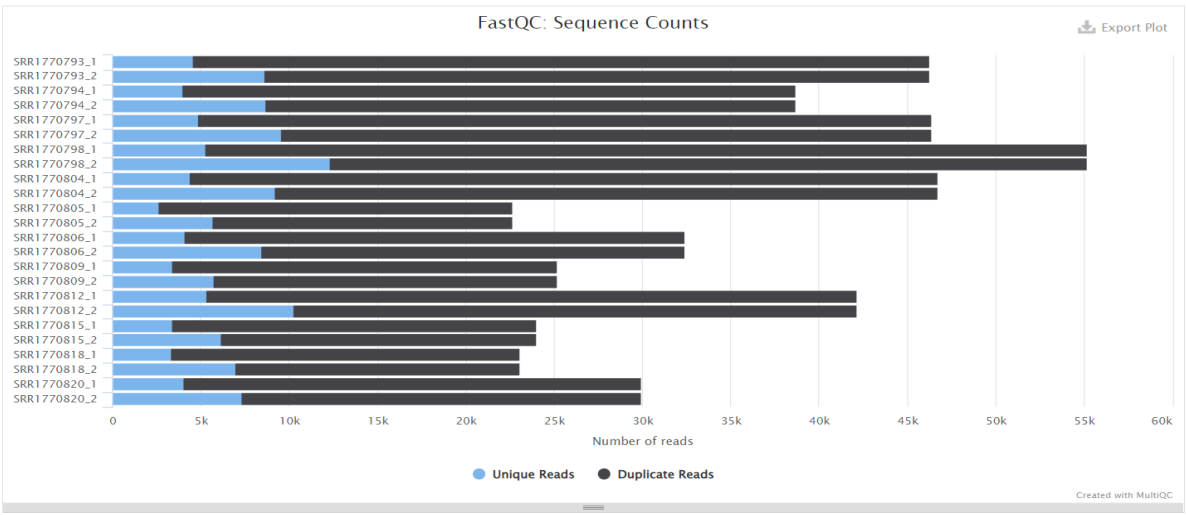
One key point to note here is that the samples (12) provided are all control samples! Our aim is to do metagenomic analysis on these samples.

Task 1 (Downloading data)

The SRA-toolset was installed in the 'tools' subfolder and the file "samples.txt" was updated with the correct SRA id's along with changes in their respective name, group, type and country. Now the script was run to download the 12 samples in their FASTQ formats in the 'data' subfolder, resulting a total of 24 files (forward + reverse reads).

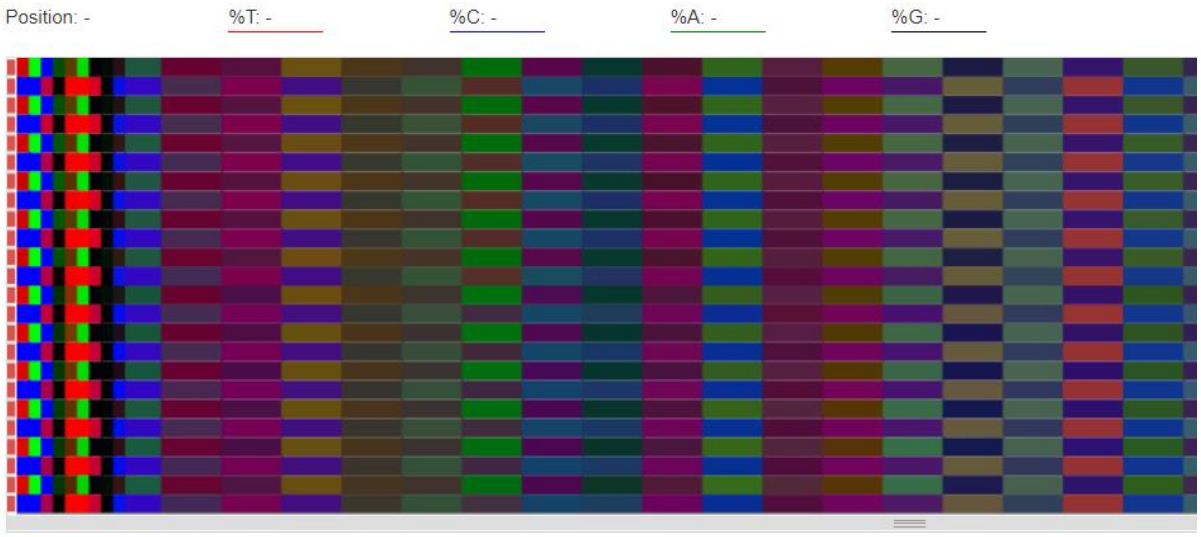
Task 2 (FastQC + MultiQC)

The quality control check for the individual files was done and a MultiQC was performed. From the sequence count plot, it was seen that we have a high duplicated reads as compared with unique reads which indicates good sequencing depth (the number of molecules sequenced). Also, higher duplicated reads are from Montane grassland South Africa as compared with the Semiarid grassland (Australia).



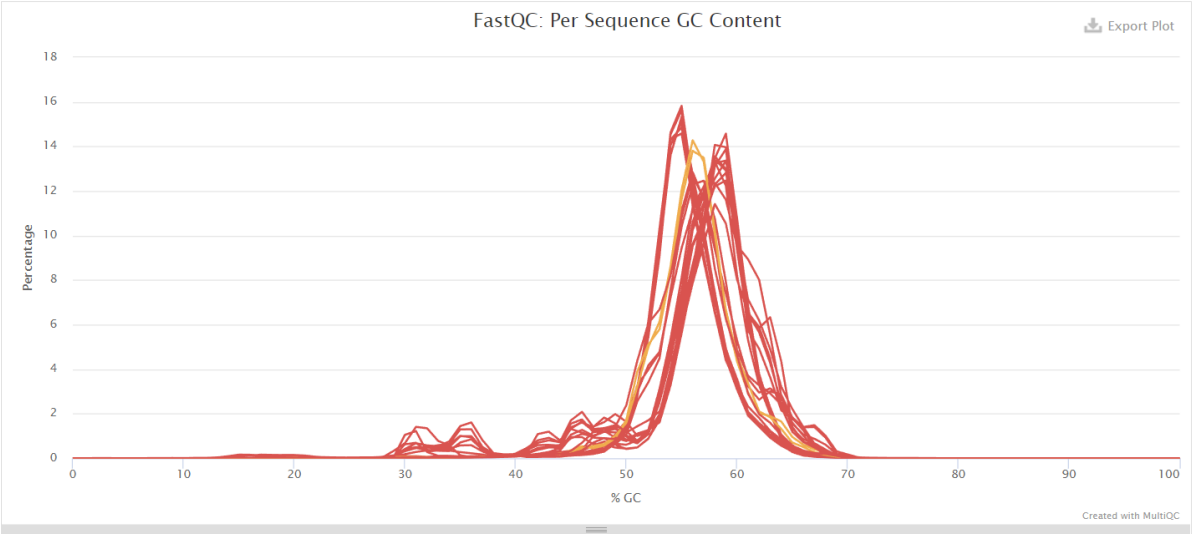
Sample Name	% Dups
SRR1770804_1	90.6%
SRR1770798_1	90.4%
SRR1770793_1	90.2%
SRR1770794_1	89.7%
SRR1770797_1	89.5%
SRR1770805_1	88.4%
SRR1770812_1	87.4%
SRR1770806_1	87.3%
SRR1770809_1	86.5%
SRR1770820_1	86.5%
SRR1770815_1	85.9%
SRR1770818_1	85.5%
SRR1770793_2	81.4%
SRR1770804_2	80.3%
SRR1770797_2	79.4%
SRR1770798_2	77.7%
SRR1770794_2	77.6%
SRR1770809_2	77.1%
SRR1770812_2	75.6%
SRR1770820_2	75.6%
SRR1770805_2	74.8%
SRR1770815_2	74.4%
SRR1770806_2	74.0%
SRR1770818_2	69.8%

While the mean quality scores for all the 24 samples were good (>30). But the per base sequence content model failed for all the samples as the bases A,G,C,T in the reads should reflect the base composition of the entire genomes and this model fails when the difference between the bases is >20% which it clearly is when we open a specific column from the heatmap



(should be mud brown for homogeneous composition)

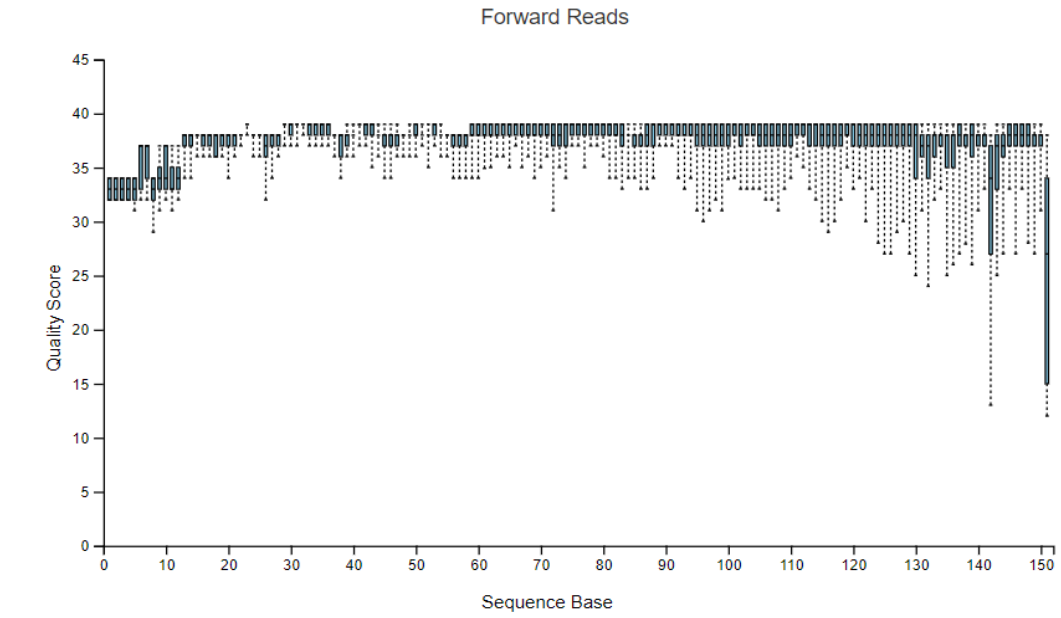
Also the per sequence GC content has <u>two peaks</u> (bimodal) but the secondary peak is not too far from the main distribution which suggests that it could be due to <u>adapters</u>, trimming will do the job.

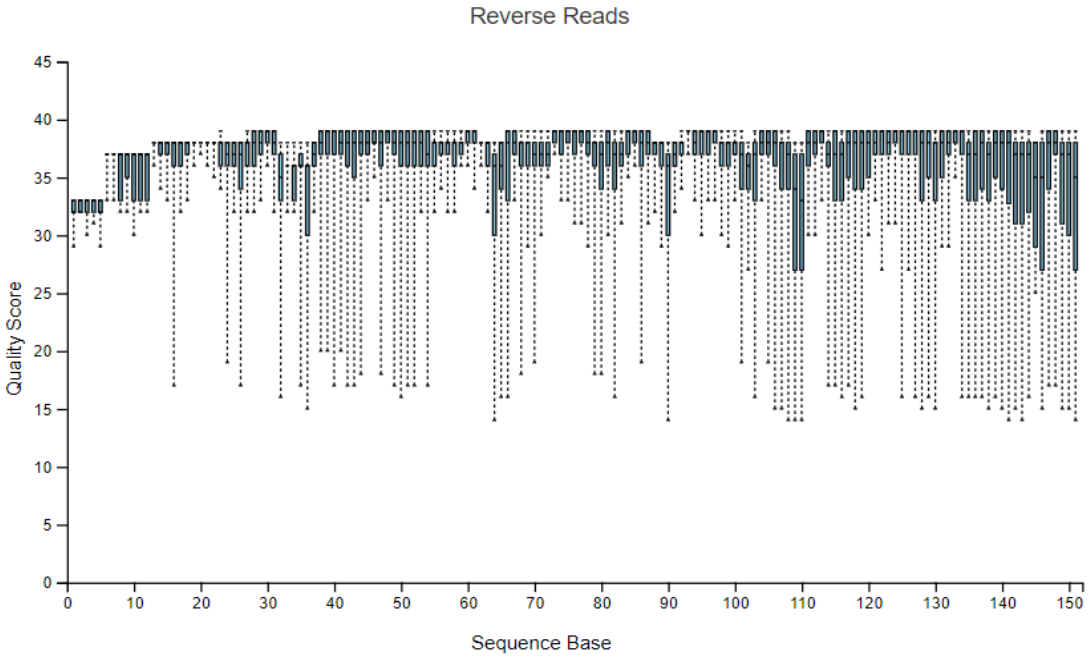


<u>Task 3 (QIIME2 QC and Trim)</u>

Before running the script, the file " source_samples.txt" was updated with proper SRA id's and the filepaths. Then the script was run with QIIME2 module and the data was imported after which the amplicon primers were removed. The quality plots and sequence length was analysed. The main aim for demultiplexing is to monitor the quality of the reads. The minimum number of reads from the sample are 22671 while the maximal number of reads is 55210. From the plots below,we know that the quality is quite good for the forward reads while for reverse reads it is not, as prior to read 2

being sequenced, the size of the clusters decreases during bridge amplification at the paired-end turnaround stage (amplification problem).





Task 4 (Denoising)

In this step, denoising with DADA2 plugin was done for quality filtering, paired-end read joining and chimera checking, basically to reduce the OTUs. And we get three output files;

summary of the denoised results (stats), sequence variants which are joined pair end reads (sequences), and all samples feature count (table).

From the ASV table we can see that every row gives us the samples information, like the "input" represents the number of reads present in the demultiplexes sequence visualisation, the "filter" column gives the number of reads left after the filtering through DADA2 algorithm (SRR1770798 had the highest reads). Then we have the number of reads left after the denoise in the "denoised" column, while merged is only performed with the paired end reads. Then we have the "non-chimeric" column which basically excludes the union of two different clusters (or reads) that were combined during PCR.

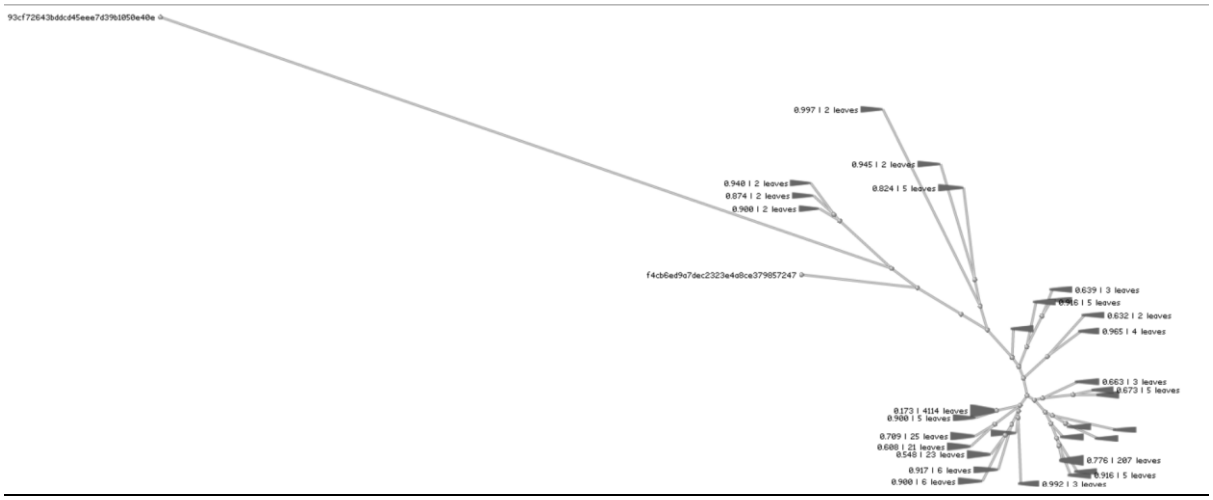
sample-id #q2:types	input numeric	filtered numeric	percentage of input passed filter numeric	denoised numeric	merged numeric	percentage of input merged numeric	non-chimeric numeric	percentage of input non-chimeric numeric
SRR1770798	55210	51551	93.37	49201	33001	59.77	29304	53.08
SRR1770797	46379	44536	96.03	42179	28237	60.88	26248	56.59
SRR1770793	46286	44711	96.6	42294	27381	59.16	24889	53.77
SRR1770804	46725	44826	95.94	42477	27414	58.67	24261	51.92
SRR1770812	42177	40221	95.36	37554	22263	52.78	21563	51.13
SRR1770794	38683	36706	94.89	34769	21227	54.87	19207	49.65
SRR1770820	29925	28780	96.17	26736	17610	58.85	17128	57.24
SRR1770806	32428	30404	93.76	28315	17474	53.89	16845	51.95
SRR1770815	24037	22944	95.45	21108	13390	55.71	13171	54.79
SRR1770805	22671	21538	95	20058	13292	58.63	12788	56.41
SRR1770809	25195	24189	96.01	22306	12900	51.2	12358	49.05
SRR1770818	23070	21614	93.69	19858	11754	50.95	11413	49.47

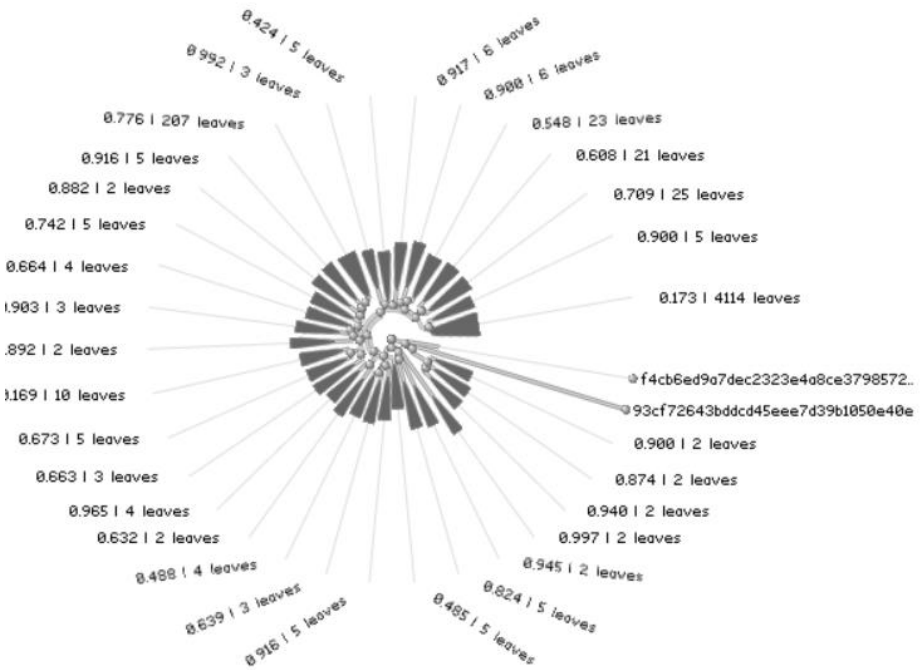
(Visualisation of table.qzc)

Also the total number of features found were 4494 while the number of samples being 12. To see the feature ID and the respective sequence, the "s04_rep_seqs_dada2.qzv" can be visualised in the qiime viewer.

Feature ID	Sequence Length	Sequence
6c231f101e665040912942268b6ef003	253	TACAGAGGTCTCAAGCGTTGTTCGGATTCATTGGGCGTAAAGGGCGCGTAGGCGGCGGGGTAAGTCTGGTGTGAAATCTAGGAGCTCAACTCCTAAACTGCACTGGATACTGCTCTGCTAGAGGACTGGAGAGGAGACTGGA
c30a2474092ab045472dc4844742b3c2	253	TACAGAGGGTGCAAGCGTTGTTCGGAATTATTGGGCGTAAAGGGTGCGTAGGCGGTGCGGTAAGTCTTCTGTGAAACCTCCGGGCTTAACTCGGAGCCTGCAGGGGAAACTGCCGTGCTGGAGTATGGGAGAGGTGAGTGGA
bae4ee50aa231e85edbbb7dae26231f2	253	TACAGAGGTCTCAAGCGTTGTTCGGATTCATTGGGCGTAAAGGGCGCGTAGGCGGCGGGGTAAGTCGGGTGTGAAATCTCAGAGCTCAACTCTGAAACTGCACTCGATACTGCTCTGCTCGAGGACTGGAGAGGAGACTGGA
ab0184f9b865f34e92ea87ed83be9079	253	TACAGAGGTCTCAAGCGTTGTTCGGATTCATTGGGCGTAAAGGGTGTGTAGGTGGCGCGCTAAGTCGGGTGTGAAATTTCGGAGCTTAACTCCGAAACTGCATTCGATACTGGCGTGCTTGAGGACTGGAGAGGAGACTGGA

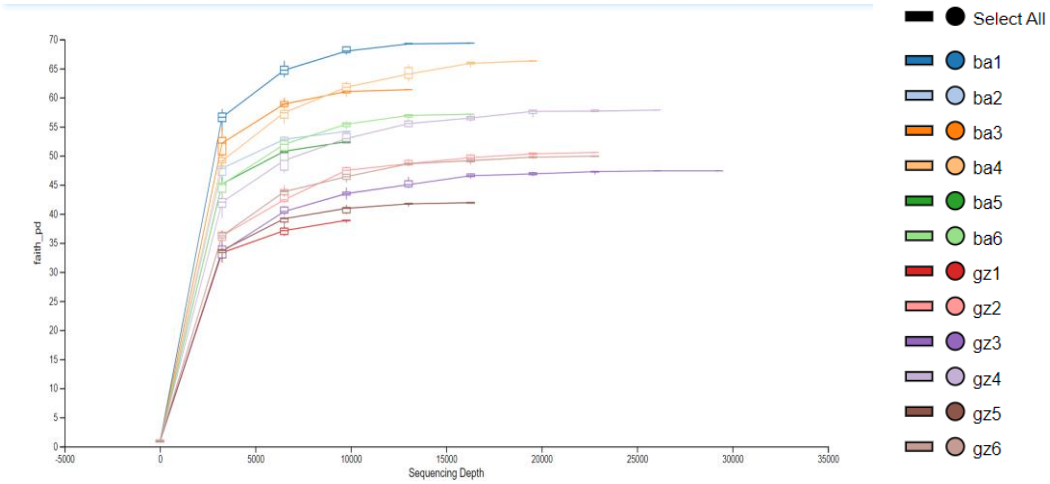
Task 5 (Phylogeny)



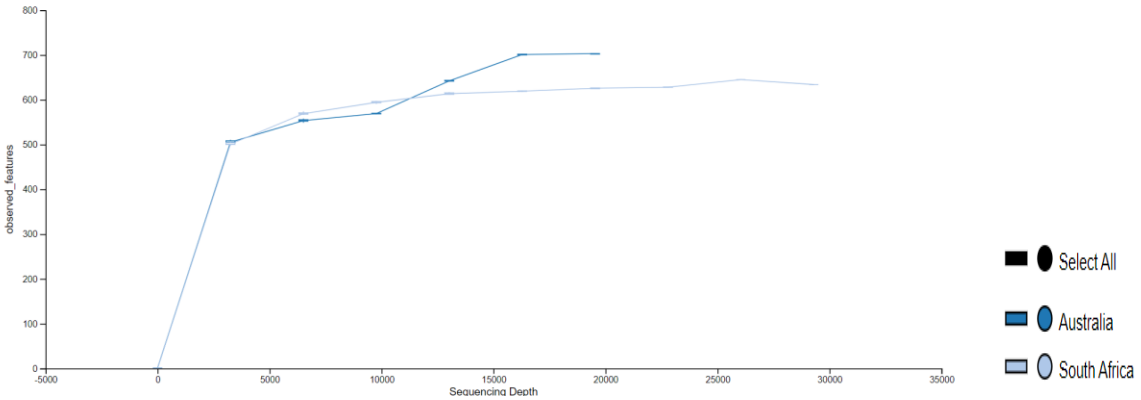


(Rarefaction)

The rarefaction step was done to normalize the richness of the detected features from the given samples. The value for “max depth” was 29304 (from maximum frequency per sample). Although in “shannon index” the diversity is calculated be measuring number of OTUs in the sample, but the scaling is done on the basis of evenness of the community.



The micro biodata from ba1 is more diverse than the gz6 as in ba1 the number of OTUs increases much more than in other samples.



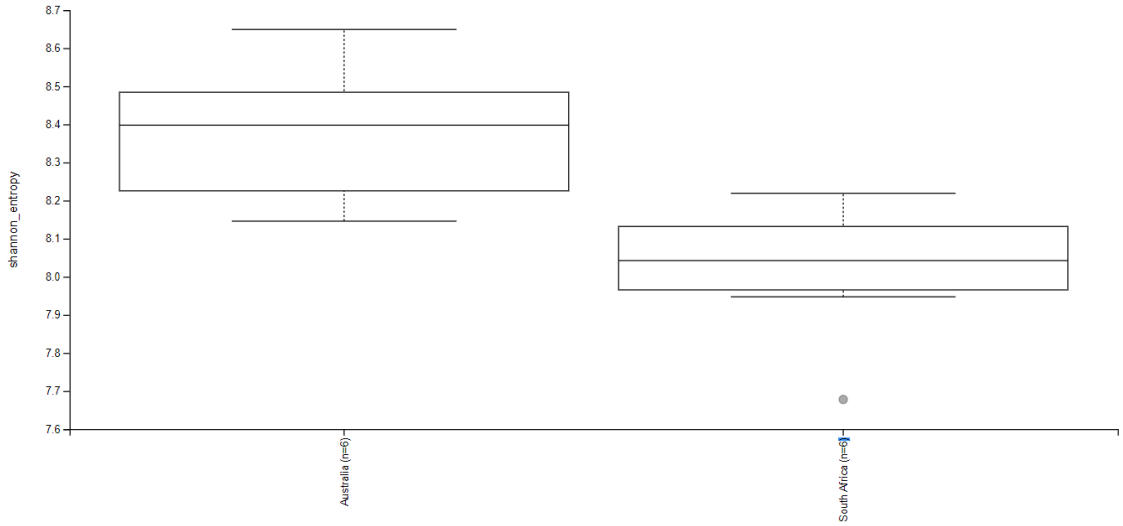
The observed features in samples from South Africa are almost constant while the observed features in samples from Australia tend to have higher diversity and there is also a possibility of common OTUs between the two as we can see from the intersection.

(Diversity metrics)

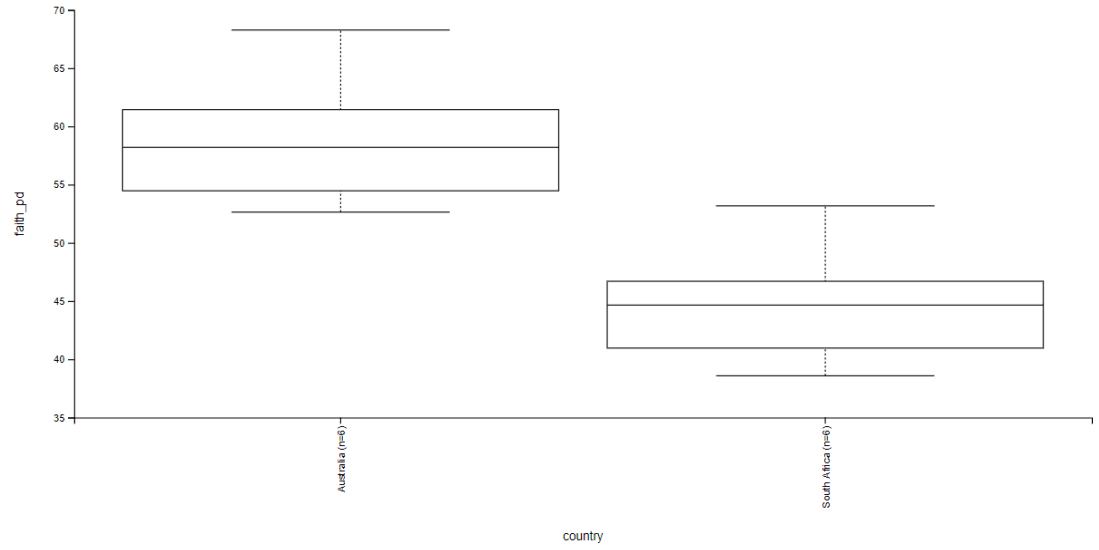
In this sub task, the phylogenetic tree and the non-rare field table from the previous tasks are taken as input to produce the alpha and beta diversity metrics. The sampling depth that was taken is 9768 which we got from the last column where the values per sample stopped appearing. Alpha diversity contains only one value (sample specific);" evenness" , "faith_PD", "shannon". While beta diversity is a kind of matrix (all samples against each other);" bray_curtis", "jaccard_distance", "unifrac (unweighted and weighted)" . Along with these subfolders, the artefacts files were also stored as the output.

Task 6 (Alpha Diversity)

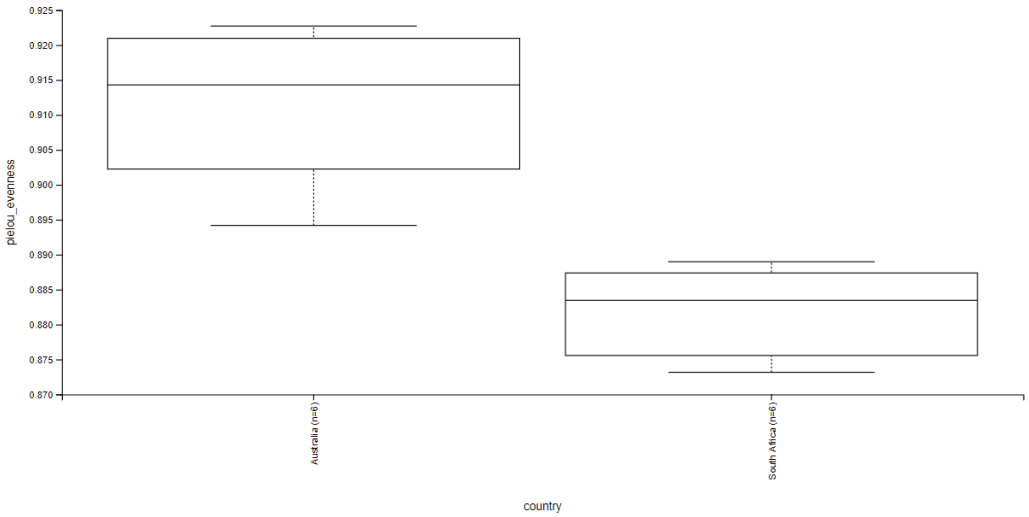
In this task we run plugin for every alpha diversity group (as mentioned in **Diversity metrics**) and produce the visualisation file for the same (.qza). [The boxplot representation includes 50% of the distribution values] . For shannon index, it refers to the uncertainty that is involved in identification of a species that is picked from a given sample randomly. And the boxplots were made according to the column "country", as different grassland types could be in same country but variable as a country is unique (unique climate, temperature etc). From the boxplot it was concluded that identifying species from Australia's sample is more uncertain than that of South Africa's sample if we compare the medians. While the H value was 5.76 and the p value was 0.016 (<0.05) which is significant.



In faiths phylogenetic diversity, which refers to the phylogenetic branches covered by the community meaning that higher number of branches means more diversity; thus we can conclude that in Australia (Semiarid grassland) the diversity is more rich as compared with that of South Africa (Montane grasslands) when we compare the medians of both the boxplots. While the H value is 7.41 and the p value is 0.006 (<0.05) which is significant.

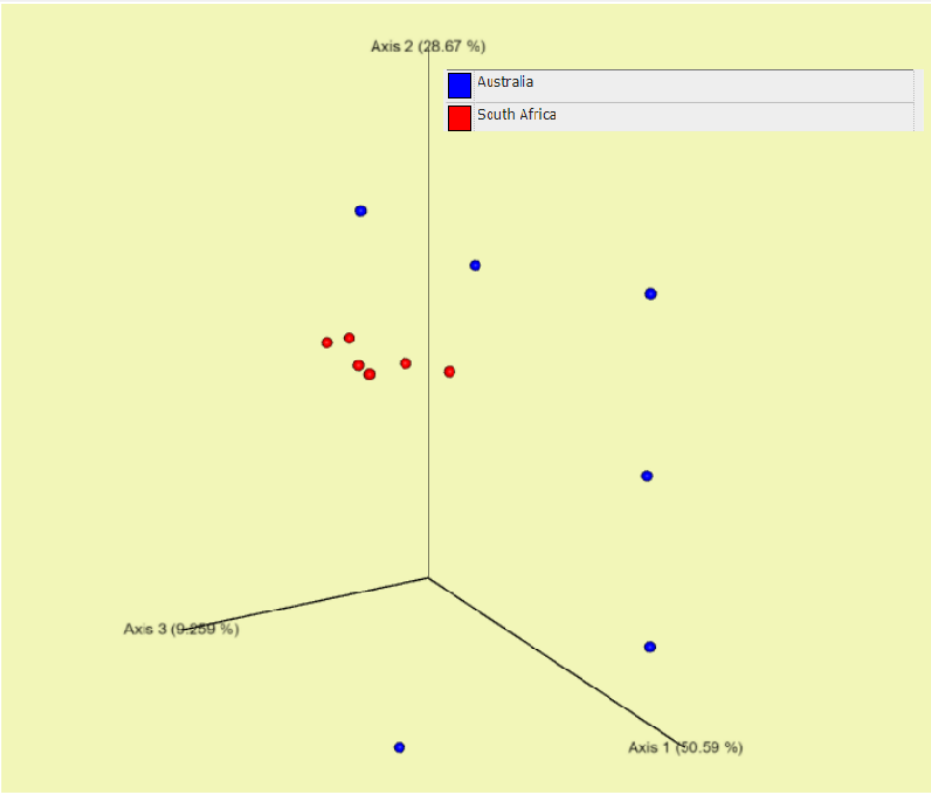


Pielou's evenness measure species diversity along with its richness, mathematically it is shannon index/shannon index h max. So it has to vary between 0 to 1. From the box plot we know that pielou's evenness for Semiarid grasslands (Australia)> Montane grassland (South Africa) which means that microbiota in the former is more abundant than in latter (types of species + number of species). Also the h value is 8.307 and p value is 0.0039 (<0.05) which is statistically significant.



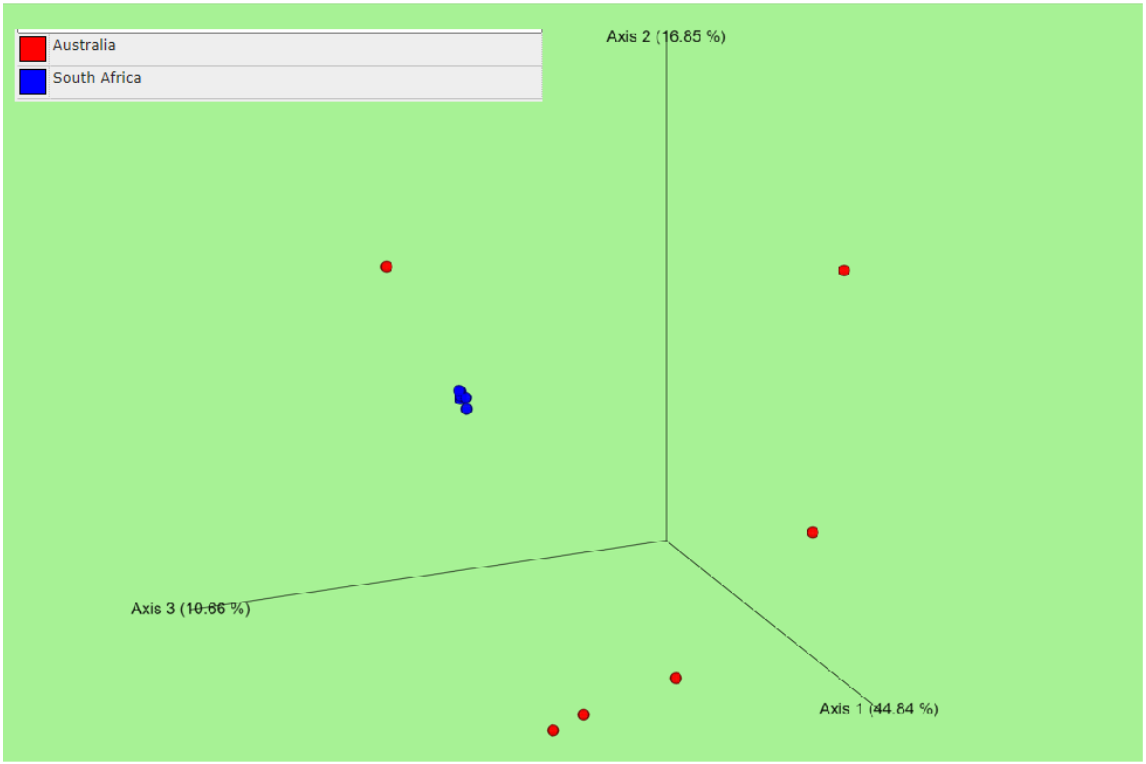
Task 7 (Beta Diversity)

In order to make PCOA plots for beta diversity knowing the distances is mandatory, for which two approaches were taken; weighted unifrac and the bray Curtis distances. In the former, the species abundance information and weights which the branch length will use was taken into account. Thus after tweaking with the graphs outlook we can see that the axis 1 (largest data change) and the axis 2 (largest proportion of remaining changes) explained over 50.6 % and 28.6 % variance of abundance in our micro biodata at the species level.



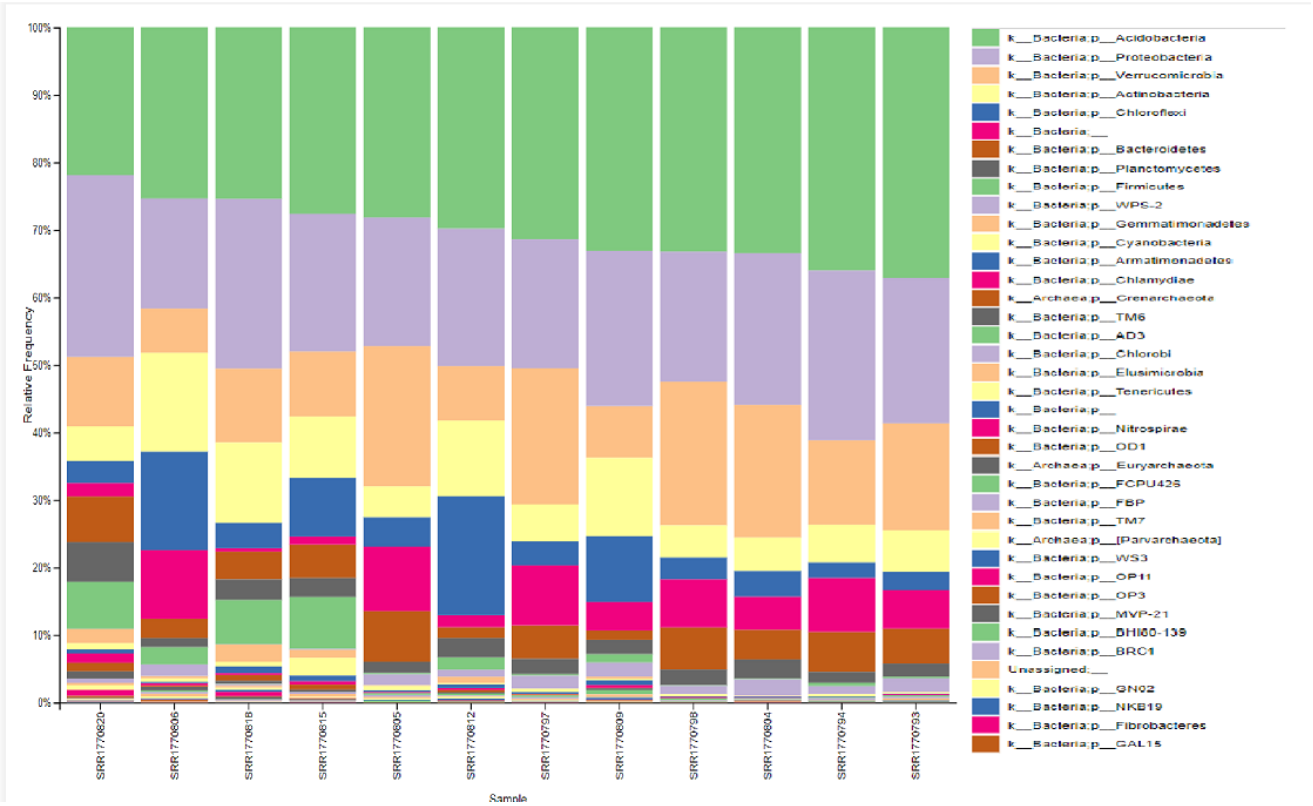
For the bray Curtis plot the main idea is to use the dissimilarity matrix and not the variance matrix, and from the plot it can be seen that the samples from South Africa are clustered

quite close as compared with the samples from Australia. So the former region shares similar composition (in terms of species) while the latter has more varied composition.

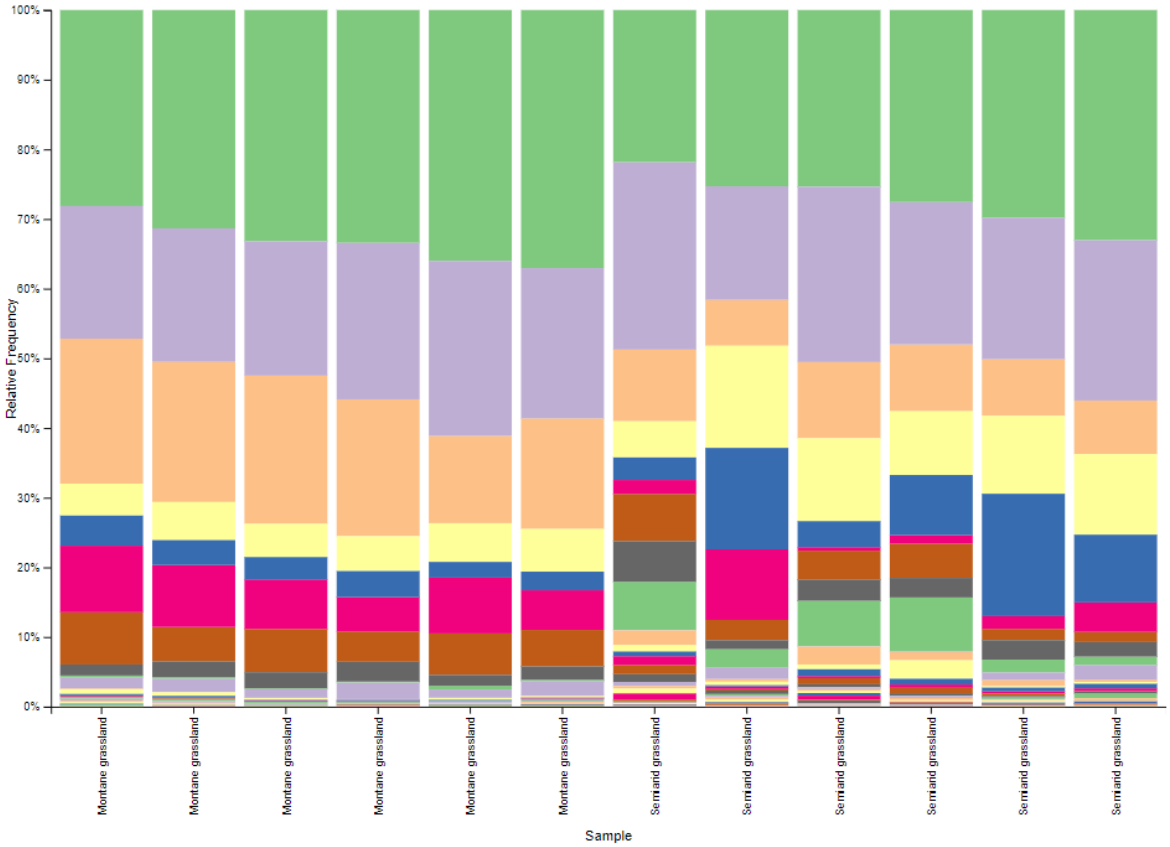


Task 8 (Taxonomy)

First the classifier (515F/806R trained) was downloaded into the resource folder and the script was run to do three tasks; assigning taxonomy to sequences, taxonomies to each ASV, and making taxonomy bar plots. For convenience, the taxonomic classification was kept till level 2 (phylum).



On the basis of level 1 taxonomical classification , one observation was the number of bacteria is >> archaea in all the samples. From the 2 level of taxonomy, it was observed that from the bacteria's, acidobacteria (green) was abundantly found in samples all across the land types (from the samples),followed by proteobacteria (grey). The names of all ASV and the taxonomical classification is provided in the file.



Proteomics

It was given that a veterinary researcher carried out proteomics experiment and the lead (peptide) that was found was LPPNTQINESPRAELSVTERTLEPPTQSPSPPPRLS

Task 1 (protein identification)

So the peptide sequence was given and our goal was to identify the protein that it came from, so the first logical step was to do a "BLAST" search. So quick "BLASTp" was performed as it is faster when target sequence identity>50% along with some pre-processing steps. It was an experimental assumed attempt as the peptide was found in animal so higher chances of better alignment (larger datasets). Few tweaking was done on the datasets ; non-redundant and swiss prot, which gave us these results:-

Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Mus musculus	39.3	39.3	94%	1e-04	58.82%	288	Q02242.1

(swiss prot)

(nr protein sequence)

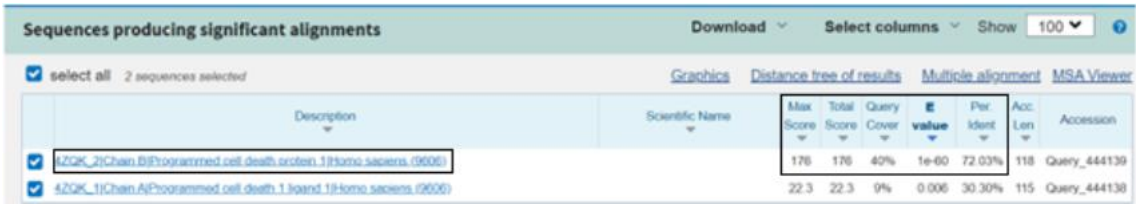
Canis lupus dingo was selected as the query coverage and the percentage of identity is 100% from the nr dataset along with the E value also low (better match) as compared with *mus musculus* (from swiss prot). Another strong reason was the different specie (vulpes) also popping as the search results, and it was found that *canis lupus and vulpes vulpes* have the same family (DNA barcoding of three species (Canis aureus, Canis lupus and Vulpes vulpes) of Canidae - PubMed (nih.gov)) but *vulpes* was not chosen due to its less score.

So the protein was “programmed cell death isoform X1” from *canis lupus dingo.*

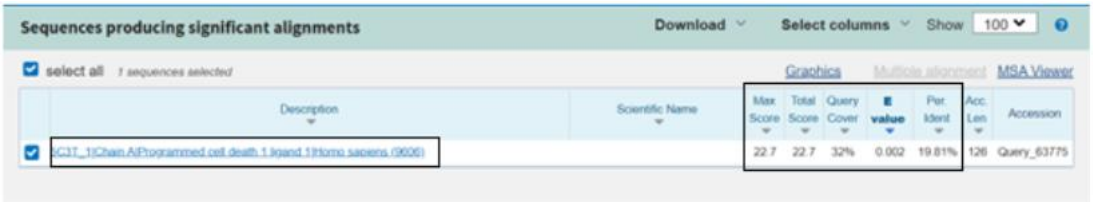
Just for the sake of curiosity, the fasta sequence of *dingo* and *familiaris* was compared and the only difference was in one ammino acid base (Proline and leucine).

Task 2 (Homology modelling)

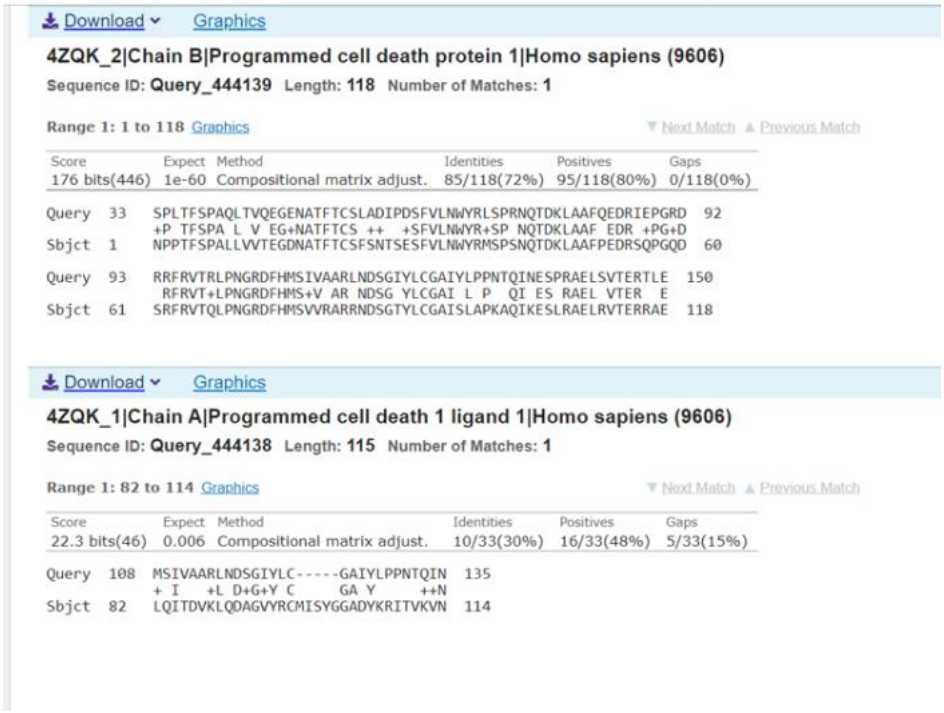
So the given research paper (Zak et al. 2015) was studied and their were two proteins that were identified; 4ZQK and 5C3T both found in humans. From these two, selection of our template will be done by analysing with the “pairwise” blast. After the pairwise blast, the max score, E value, Query cover, and the percentage of identity was compared with 4ZQK clearly being the better alternative, with 72.03 % sequence identity. This is for the Chain B as Chain A has only 30% sequence identity.



(4ZQK description)



(5C3T description)



Download Graphics

4ZQK_2|Chain B|Programmed cell death protein 1|Homo sapiens (9606)

Sequence ID: **Query_444139** Length: **118** Number of Matches: **1**

Range 1: 1 to 118 Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
176 bits(446)	1e-60	Compositional matrix adjust.	85/118(72%)	95/118(80%)	0/118(0%)

```
Query  33   SPLTFSPAQLTVQEGENATFTCSLADIPDSFVLNWYRLSPRNQTDKLAAFQEDRIEPGRD  92
            +P TFSPA L V EG+NATFTCS ++  +SFVLNWYR+SP NQTDKLAAF EDR +PG+D
Sbjct  1    NPPTFSPALLVVTEGDNATFTCSFSNTSESFVLNWYRMSPSNQTDKLAAFPEDRSQPGQD  60

Query  93   RRFRVTRLPNGRDFHMSIVAARLNDSGIYLCGAIYLPPNTQINESPRAELSVTERTLE  150
             RFRVT+LPNGRDFHMS+V AR NDSG YLCGAI L P  QI ES RAEL VTER  E
Sbjct  61   SRFRVTQLPNGRDFHMSVVRARRNDSGTYLCGAISLAPKAQIKESLRAELRVTERRAE  118
```

Download Graphics

4ZQK_1|Chain A|Programmed cell death 1 ligand 1|Homo sapiens (9606)

Sequence ID: **Query_444138** Length: **115** Number of Matches: **1**

Range 1: 82 to 114 Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
22.3 bits(46)	0.006	Compositional matrix adjust.	10/33(30%)	16/33(48%)	5/33(15%)

```
Query  108  MSIVAARLNDSGIYLC-----GAIYLPPNTQIN  135
            + I   +L D+G+Y C     GA Y     ++N
Sbjct  82   LQITDVKLQDAGVYRCMISYGGADYKRITVKVN  114
```

(4ZQK alignment)

For the modelling , SWISS-MODEL workspace was used and the fasta sequence of the target protein (programmed cell death isoform X1) along with the template protein (4ZQK) in its pdb format was submitted with the title and email id. After sometime, the model that was generated is shown below:-

(colour scheme based on confidence of class)