Całość chce przeanalizować za pomocą Qiime2 i DADA2

Użyłem takich dwóch tabel do analizy

sample_data.tsv

sample-id	forward-absolute-filepath	reverse-absolute-filepath
A10	/media/Data1/Bartosz_Fotchki_16S_Nano/test/A10_S9_L001_R1_001.fastq.gz	/media/Data1/Bartosz_Fotchki_16S_
A12	/media/Data1/Bartosz_Fotchki_16S_Nano/test/A12_S10_L001_R1_001.fastq.gz	/media/Data1/Bartosz_Fotchki_16S_
A1	/media/Data1/Bartosz_Fotchki_16S_Nano/test/A1_S1_L001_R1_001.fastq.gz	/media/Data1/Bartosz_Fotchki_16S_
A2	/media/Data1/Bartosz_Fotchki_16S_Nano/test/A2_S2_L001_R1_001.fastq.gz	/media/Data1/Bartosz_Fotchki_16S_
a3	/media/Data1/Bartosz_Fotchki_16S_Nano/test/a3_S3_L001_R1_001.fastq.gz	/media/Data1/Bartosz_Fotchki_16S_
a4	/media/Data1/Bartosz_Fotchki_16S_Nano/test/a4_S4_L001_R1_001.fastq.gz	/media/Data1/Bartosz_Fotchki_16S_
a5	/media/Data1/Bartosz_Fotchki_16S_Nano/test/a5_S5_L001_R1_001.fastq.gz	/media/Data1/Bartosz_Fotchki_16S_
a6	/media/Data1/Bartosz_Fotchki_16S_Nano/test/a6_S6_L001_R1_001.fastq.gz	/media/Data1/Bartosz_Fotchki_16S_
a7	/media/Data1/Bartosz_Fotchki_16S_Nano/test/a7_S7_L001_R1_001.fastq.gz	/media/Data1/Bartosz_Fotchki_16S_
a8	/media/Data1/Bartosz_Fotchki_16S_Nano/test/a8_S8_L001_R1_001.fastq.gz	/media/Data1/Bartosz_Fotchki_16S_
B10	/media/Data1/Bartosz_Fotchki_16S_Nano/test/B10_S20_L001_R1_001.fastq.gz	/media/Data1/Bartosz_Fotchki_16S_
B12	/media/Data1/Bartosz_Fotchki_16S_Nano/test/B12_S21_L001_R1_001.fastq.gz	/media/Data1/Bartosz_Fotchki_16S_
b1	/media/Data1/Bartosz_Fotchki_16S_Nano/test/b1_S11_L001_R1_001.fastq.gz	/media/Data1/Bartosz_Fotchki_16S_
b2	/media/Data1/Bartosz_Fotchki_16S_Nano/test/b2_S12_L001_R1_001.fastq.gz	/media/Data1/Bartosz_Fotchki_16S_
b3	/media/Data1/Bartosz_Fotchki_16S_Nano/test/b3_S13_L001_R1_001.fastq.gz	/media/Data1/Bartosz_Fotchki_16S_
b4	/media/Data1/Bartosz_Fotchki_16S_Nano/test/b4_S14_L001_R1_001.fastq.gz	/media/Data1/Bartosz_Fotchki_16S_
b5	/media/Data1/Bartosz_Fotchki_16S_Nano/test/b5_S15_L001_R1_001.fastq.gz	/media/Data1/Bartosz_Fotchki_16S_
b6	/media/Data1/Bartosz_Fotchki_16S_Nano/test/b6_S16_L001_R1_001.fastq.gz	/media/Data1/Bartosz_Fotchki_16S_
b7	/media/Data1/Bartosz_Fotchki_16S_Nano/test/b7_S17_L001_R1_001.fastq.gz	/media/Data1/Bartosz_Fotchki_16S_
b8	/media/Data1/Bartosz_Fotchki_16S_Nano/test/b8_S18_L001_R1_001.fastq.gz	/media/Data1/Bartosz_Fotchki_16S_
B9	/media/Data1/Bartosz_Fotchki_16S_Nano/test/B9_S19_L001_R1_001.fastq.gz	/media/Data1/Bartosz_Fotchki_16S_

sample_metadata.tsv

sample-id	group	number
A10	A	10
A12	A	12
A1	A	1
A2	A	2
a3	A	3
a4	A	4
a5	A	5
a6	A	6
a7	A	7
a8	A	8
B10	B	10
B12	B	12
b1	B	1
b2	B	2
b3	B	3
b4	B	4
b5	B	5
b6	B	6
b7	B	7
b8	B	8

Wyłączyłem Undetermined_S0_L001_R1_001.fastq.gz i Undetermined_S0_L001_R2_001.fastq.gz
W pierwszej analizie wziąłem wszystko i te pliki wyraźnie odstawały, np. pod kątem liczby odczytów

Jak rozumiem mamy tutaj do czynenia z paired ends i próbami po demultiplexingu, więc można od razu przejść do wycinania odczytów o słabej jakości (adapterów w oparciu o fastqc nie zauważyłem).

Generalnie wydaje mi się, że w tym miejscu powinienem zwrócić uwagę na jeszcze jakieś rzeczy, np. dowiedzieć się czy są primery i jakie mają sekwencje, ale nie wiem jak to zrobić.

```
qiime tools import \
   --type 'SampleData[PairedEndSequencesWithQuality]' \
   --input-path sample_data.tsv \
   --output-path paired-end-demux.qza \
   --input-format PairedEndFastqManifestPhred33V2

qiime demux summarize \
   --i-data paired-end-demux.qza \
   --o-visualization demux.qzv
```

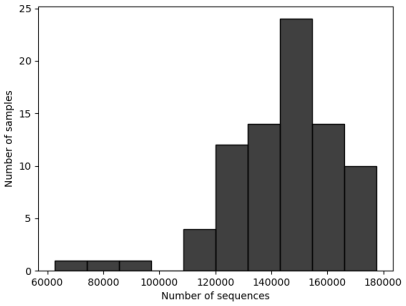
Jak rozumiem taka liczba odczytów jak poniżej jest wystarczająca. Gdyby minimum był w okolicach 10k to dopiero wtedy należy się martwić, szczególnie jeśli interesuje nas wykrycie bakterii, które występują w nieznacznych ilościach.

Zastanawiam się jednak, czy na podstawie poniższych barplotów nie należy rozważyć wyłączenia trzech prób, które odstają pod kątem liczby odczytów.

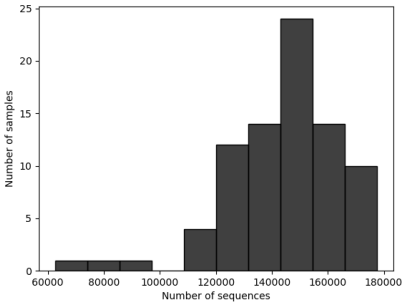
Demultiplexed sequence counts summary

	forward reads	reverse reads
Minimum	62533	62533
Median	146838.0	146838.0
Mean	144384.888889	144384.888889
Maximum	177675	177675
Total	11695176	11695176

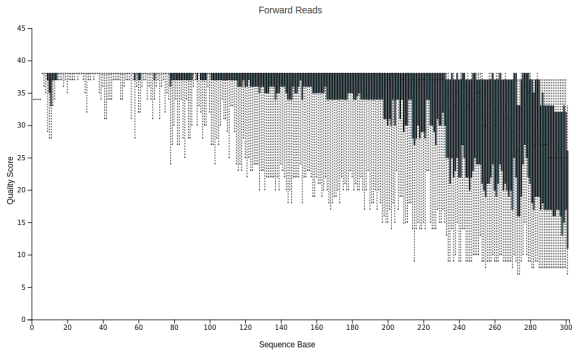
Forward Reads Frequency Histogram

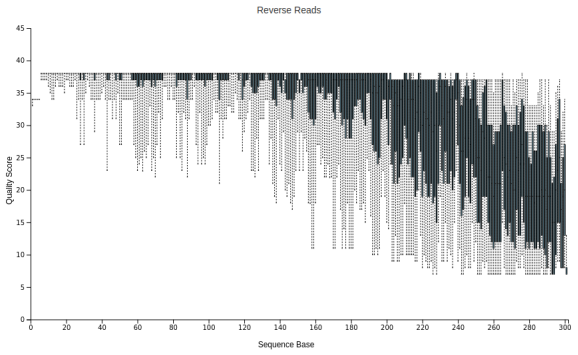


Reverse Reads Frequency Histogram



Forward i reverse przyciąłem na pozycjach 20 i 250.
Przypuszczam jednak, że z lewej strony można niczego nie przycinać. Czy takie przycinanie na oko jest wystarczajaco dobre? Wiem, że są programy, które automatycznie dobierają parametry (np. FIGARO) ale jeszcze niczego takiego nie testowałem






```
qiime dada2 denoise-paired \
     --i-demultiplexed-seqs paired-end-demux.qza \
     --p-trim-left-f 20 \
     --p-trim-left-r 20 \
     --p-trunc-len-f 250 \
     --p-trunc-len-r 250 \
     --p-n-threads 12 \
     --o-representative-sequences rep-seqs.qza \
```

```
      --o-table table.qza \
      --o-denoising-stats dadastats.qza

qiime metadata tabulate \
      --m-input-file dadastats.qza \
      --o-visualization denoising-stats.qzv

qiime feature-table summarize \
      --i-table table.qza \
      --o-visualization table.qzv \
      --m-sample-metadata-file sample_metadata.tsv

qiime feature-table tabulate-seqs \
      --i-data rep-seqs.qza \
      --o-visualization rep-seqs.qzv
```

Mam wrażenie, że trochę za mało zostało tych odczytów i te procenty powinny się wahać w okolicach 80%

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
A1	147724	99580	67.41	98555	94281	63.82	91412	61.88
A10	146503	97930	66.85	96880	93020	63.49	89436	61.05
A12	147396	101960	69.17	101250	97873	66.4	92630	62.84
A2	142939	94507	66.12	93429	88654	62.02	85922	60.11
a3	156229	103715	66.39	102554	96474	61.75	93657	59.95
a4	131957	89786	68.04	88858	84535	64.06	82233	62.32
a5	145324	100749	69.33	99911	96099	66.13	90692	62.41
a6	116879	81152	69.43	80515	77380	66.21	72430	61.97
a7	148766	97614	65.62	96545	92087	61.9	88410	59.43
a8	130887	86846	66.35	85998	82018	62.66	79654	60.86

Table summary

Metric	Sample
Number of samples	81
Number of features	1,952
Total frequency	7,069,637

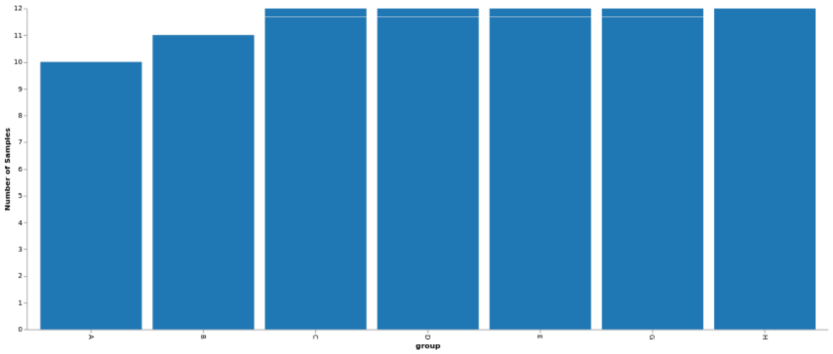
Frequency per sample

	Frequency
Minimum frequency	22,120.0
1st quartile	79,617.0
Median frequency	88,774.0
3rd quartile	94,906.0
Maximum frequency	111,247.0
Mean frequency	87,279.46913580247

Frequency per sample detail (csv | html)

Frequency per feature

	Frequency
Minimum frequency	1.0
1st quartile	8.0
Median frequency	77.0
3rd quartile	660.0
Maximum frequency	327,927.0
Mean frequency	3,621.7402663934427



Plot Controls

Save as SVG | Save as PNG | View Source

Metadata Category

group

Sampling Depth

0

(zero implies no even sampling)
Retained 7,069,637 (100.00%) features in 81 (100.00%) samples at the specifed sa
depth.

Sample ID	Feature Count
g1	111247
C11	108132
C12	108048
g2	106594
e6	105038
E8	104213
e1	104116
B12	103643
c2	102895
g6	101549

Sequence Length Statistics

Download sequence-length statistics as a TSV

Sequence Count	Min Length	Max Length	Mean Length	Range	Standard Deviation
1952	272	437	406.3	165	10.69

Seven-Number Summary of Sequence Lengths

Download seven-number summary as a TSV

Percentile:	2%	9%	25%	50%	75%	91%	98%
Length* (nts):	399	400	400	402	419	421	425

*Values rounded down to nearest whole number.

Sequence Table

To BLAST a sequence against the NCBI nt database, click the sequence and then click the *View report* button on the resulting page.

Download your sequences as a raw FASTA file

Click on a Column header to sort the table.

Feature ID	Sequence Length	Sequence
b00a2611a89f9b584f640dd873f32010	420	GGAATATTGGTCAATGGGCGATGGCCTGAACCAGCCAAGTAGCGTGAAGGATGAAGGTTCTATGGATTGTAAACTTCTTTTATAAAGGAATAAAGTGAGGCACGTGTGCCTTTTTGTATGTACTTTATGAATAAGGATCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGATCCGAGCGTTATCCGGATTTAT
7b865e5faff9350082fece8ac01787a2	420	GGAATATTGGTCAATGGGCGAGAGCCTGAACCAGCCAAGTAGCGTGAAGGATGACTGCCCTATGGGTTGTAAACTTCTTTTATAAAGGAATAAAGTCGGGTATGGATACCCGTTTGCATGTACTTTATGAATAAGGATCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGATCCGAGCGTTATCCGGATTTAT
627243fcc4bc104918ab912d2653e020	403	GGAATATTGCACAATGGGGGAAACCCTGATGCAGCAACGCCGCGTGAAGGAAGACGGTTTTCGGATTGTAAACTTCTTTTCTTAGTGAAGAAACAAATGACGGTAGCTAAGGAATAAGCATCGGCTAACTACGTGCCAGCAGCCGCGGTAATACGTAGGATGCGAGCGTTATCCGGATTTACTGGGTGTAAAGGGAGCG
a7f318369a46065ae032dddf5502a890	420	GGAATATTGGTCAATGGGCGAGAGCCTGAACCAGCCAAGTAGCGTGAAGGATGACTGCCCTATGGGTTGTAAACTTCTTTTATAAAGGAATAAAGTCGGGTATGTATACCCGTTTGCATGTACTTTATGAATAAGGATCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGATCCGAGCGTTATCCGGATTTAT
6174afb7991fc60d1336439ae545b673	419	GGAATATTGGTCAATGGGCGAGAGCCTGAACCAGCCAAGTCGCGTGAGGGAAGACGGTCCTATGGATTGTAAACCTCTTTTGCCGGGGAGCAAAGAGCGGTACGTGTACCGCGCCGAGAGTACCCGGAGAAAAAGCATCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGATCGCGAGCGTTATCCGGATTTAT
f1fb4c263704d6a492ca9667dd383a64	400	GGGATATTGCACAATGGGGGAACCCTGATGCAGCGACGCCGCGTGGGTGAAGAAGCGCTCCGGCGCGTAAAGCCCTGTCAGCAGGGAAGAAGGTGACGGTACCTGACCAAGAAGCCCCGGCTAACTACGTGCCAGCAGCCGCGGTAATACGTAGGGGGCAAGCGTTATCCGGATTTACTGGGTGTAAAGGGGGCGCAG
d38d4706c512115fa8f1a576c268c65c	420	GGAATATTGGTCAATGGGCGAGAGCCTGAACCAGCCAAGTCGCGTGAAGGATGAAGGATCTATGGTTTGTAAACTTCTTTTATATGGGAATAAAGTGAGGAACGTGTTCCTTTTTGTATGTACCATATGAATAAGCATCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGATGCGAGCGTTATCCGGATTTAT
6e77c78161a2866f2e5784627b21f3f7	420	GGAATATTGGTCAATGGACGCAAGTCTGAACCAGCCATGCCGCGTGCAGGGAAGACGGCTCTATGAGTTGTAAACTGCTTTTGTATTAGGGTAAACTCAGGTACGTGTACCTGACTGAAAGTATAATACGAATAAGGATCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGATCCAAGCGTTATCCGGATTTAT
d28457e5212cdfb10be4a94691efdd61	419	GGAATATTGGTCAATGGGCGAGAGCCTGAACCAGCCAAGTCGCGTGAGGGAAGACGGTCCTATGGATTGTAAACCTCTTTTGCCGGGGAGCAAAGAGCGGCACGTGTGCCGCGCCGAGAGTACCCGGAGAAAAAGCATCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGATCGCGAGCGTTATCCGGATTTAT
863bb4ecd7466237ce3ad9c9f7584bf2	420	GGAATATTGGTCAATGGGCGATGGCCTGAACCAGCCAAGTAGCGTGAAGGATGAAGGTTCTATGGATTGTAAACTTCTTTTATAAAGGAATAAAGTGGGGCACGTGTGCCTTTTTGTATGTACTTTATGAATAAGGATCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGATCCGAGCGTTATCCGGATTTAT
5c1fbdf5ff54c1109924b7e136c089b0	420	GGAATATTGGTCAATGGACGGGAGTCTGAACCAGCCAAGTAGCGTGAAGGATGAAGGTTCTATGGATTGTAAACTTCTTTTATAAAGGAATAAAGTGGGGCACGTGTGCCTTTTTGTATGTACTTTATGAATAAGGATCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGATTCGAGCGTTATCCGGATTTAT
63094ae95c6a0687443c88e79e4da234	420	GGAATATTGGTCAATGGGCGAAGGCCTGAACCAGCCAAGTAGCGTGAAGGATGAAGGTTCTATGGATTGTAAACTTCTTTTATAAAGGAATAAAGTGAGGCACGTGTGCCTTTTTGTATGTACTTTATGAATAAGGATCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGATCCGAGCGTTATCCGGATTTAT
8a11d6cab7c6932c404e54734d982ff2	420	GGAATATTGGTCAATGGACGAGAGTCTGAACCAGCCAAGTAGCGTGAAGGATGACTGCCCTATGGGTTGTAAACTTCTTTTATATGGGAATAAAGTGCAGTATGTATACTGTTTTGTATGTACCATATGAATAAGGATCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGATCCGAGCGTTATCCGGATTTAT
8e4a7f0b1083e15fbe31a1ee62b4d5e0	420	GGAATATTGGTCAATGGGCGAAGGCCTGAACCAGCCAAGTAGCGTGAAGGATGAAGGTTCTATGGATTGTAAACTTCTTTTATAAAGGAATAAAGTGGGGCACGTGTGCCTTTTTGTATGTACTTTATGAATAAGGATCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGATCCGAGCGTTATCCGGATTTAT
3752fb0ed9f09ad9bcf6e860a0543d01	400	GGAATATTGGGCAATGGAGGCAACTCTGACCCAGCCATGCCGCGTGAGTGAAGAAGGTTTTCGGATTGTAAAGCTCTTTCGGATGTGACGATGATGACGGTAGCATCTAAAGAAGCCCCGGCTAACTTCGTGCCAGCAGCCGCGGTAATACGAAGGGGGCGAGCGTTGTTCGGAATTACTGGGCGTAAAGGGTGTGTAG
053339c8d82c891a55928adeb303ee1a	403	GGAATATTGGGCAATGGGCGCAAGCCTGACCCAGCAACGCCGCGTGAAGGAAGAAGGCTTTCGGGTTGTAAACTTCTTTTCTCAGGGACGAAGCAAGTGACGGTACCTGAGGAATAAGCCACGGCTAACTACGTGCCAGCAGCCGCGGTAATACGTAGGTGGCAAGCGTTATCCGGATTTACTGGGTGTAAAGGGCGTG
66d2497029774ac97fad1f4328b997ca	420	GGAATATTGGTCAATGGGCGTAAGCCTGAACCAGCCAAGTCGCGTGAGGGATGAAGGTTCTATGGATCGTAAACCTCTTTTATAAGGGAATAAAGTGCGGGACGTGTCCCGTTTTGTATGTACCTTATGAATAAGGATCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGATCCGAGCGTTATCCGGATTTAT
4c5b965bc83a4ad8ace0b4a2b7291b17	420	GGAATATTGGTCAATGGACGCGAGTCTGAACCAGCCAAGTAGCGTGAAGGATGACTGCCCTATGGGTTGTAAACTTCTTTTATATGGGAATAAAGTGGTCCACGTGTGGACTTTTGTATGTACCATATGAATAAGGATCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGATCCGAGCGTTATCCGGATTTAT

Przypisanie wykrytych sekwencji do określonych taksonów na podstawie sekwencji referencyjnych (baza SILVA). Nie wiem jak ten etap poprawnie zdefiniować. Mamy AVS i przypisu

```
qiime tools import \
      --type FeatureData[Sequence] \
```

```
     --input-path
$HOME/SILVA_132_QIIME_release/rep_set/rep_set_16S_only/99/silva_13
2_99_16S.fna \
     --output-path 99_otus_16S

qiime tools import \
     --type FeatureData[Taxonomy] \
     --input-path
$HOME/SILVA_132_QIIME_release/taxonomy/16S_only/99/majority_taxono
my_7_levels.txt \
     --input-format HeaderlessTSVTaxonomyFormat \
     --output-path majority_taxonomy_7_levels
```

Używam gotowego klasyfikatora, które są wbudowane do qiime

```
qiime feature-classifier classify-consensus-blast \
     --i-query rep-seqs.qza \
     --i-reference-taxonomy majority_taxonomy_7_levels.qza \
     --i-reference-reads 99_otus_16S.qza \
     --o-classification taxonomy \
     --p-perc-identity 0.90 \
     --p-maxaccepts 1
```

Nie wiem, czy powinienem coś odfiltrowywać. Używam tego na wszelki wypadek.

```
qiime taxa filter-table \
     --i-table table.qza \
     --i-taxonomy taxonomy.qza \
     --o-filtered-table filtered-table.qza


qiime taxa barplot \
     --i-table table.qza \
     --i-taxonomy taxonomy.qza \
     --m-metadata-file sample_metadata.tsv \
     --o-visualization taxa-bar-plots.qzv
```

Budowanie drzew

```
qiime phylogeny align-to-tree-mafft-fasttree \
      --i-sequences rep-seqs.qza \
      --o-alignment aligned-rep-seqs.qza \
      --o-masked-alignment masked-aligned-rep-seqs.qza \
      --o-tree unrooted-tree.qza \
      --o-rooted-tree rooted-tree.qza

qiime diversity core-metrics-phylogenetic \
      --i-phylogeny rooted-tree.qza \
      --i-table table.qza \
      --p-sampling-depth 1000 \
      --m-metadata-file sample-metadata.tsv \
      --output-dir core-metrics-results
```

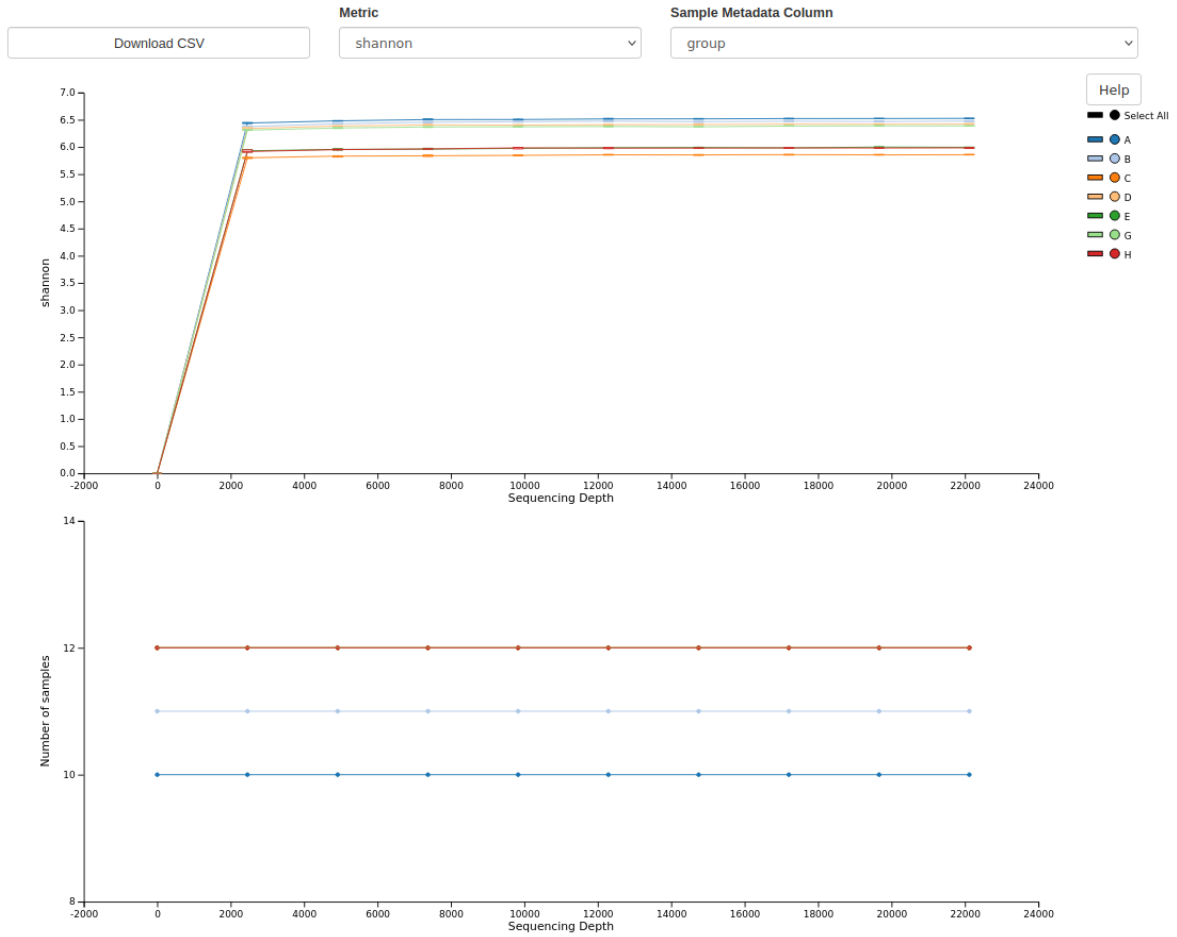
Różnorodność alfa i beta

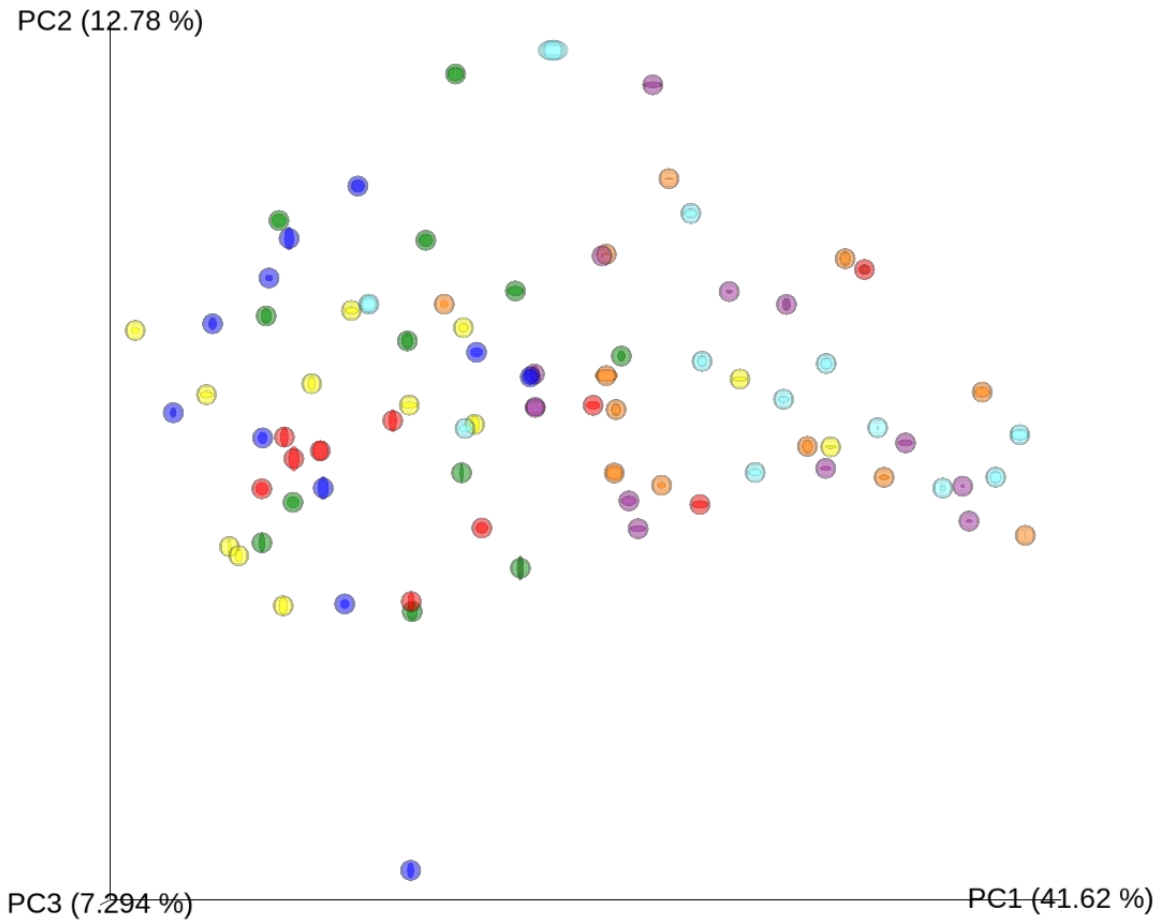
```
qiime diversity alpha-rarefaction \
      --i-table table.qza \
      --i-phylogeny rooted-tree.qza \
      --p-max-depth 22120 \
      --m-metadata-file sample_metadata.tsv \
      --o-visualization alpha-rarefaction.qzv

qiime diversity beta-rarefaction \
      --i-table feature_table_samples.qza \
      --p-metric weighted_unifrac \
      --p-clustering-method nj \
      --m-metadata-file metadata_for_qiime2.txt \
      --p-sampling-depth 21383 \
      --i-phylogeny phylogenetic_tree/rooted_tree.qza \
      --o-visualization beta_rarefaction/weighted_unifrac.qzv
```

Alpha rarefaction

The following metadata columns have been omitted because they didn't contain categorical data, or the column consisted only of missing values: **number**




PC2 (12.78 %)
PC3 (7.294 %)
PC1 (41.62 %)