

Полузёрова Анастасия

Магистратура ТюмГУ

Практическое задание 3

QIIME 2

```
In [1]: !conda --version
```

```
conda 23.5.0
```

```
In [2]: !ls data_met/sequences
```

Abacumov-B-13_S13_L001_R1_001.fastq.gz	Abacumov-B-39_S39_L001_R1_001.fastq.g
z	
Abacumov-B-13_S13_L001_R2_001.fastq.gz	Abacumov-B-39_S39_L001_R2_001.fastq.g
z	
Abacumov-B-14_S14_L001_R1_001.fastq.gz	Abacumov-B-3_S3_L001_R1_001.fastq.gz
Abacumov-B-14_S14_L001_R2_001.fastq.gz	Abacumov-B-3_S3_L001_R2_001.fastq.gz
Abacumov-B-15_S15_L001_R1_001.fastq.gz	Abacumov-B-40_S40_L001_R1_001.fastq.g
z	
Abacumov-B-15_S15_L001_R2_001.fastq.gz	Abacumov-B-40_S40_L001_R2_001.fastq.g
z	
Abacumov-B-16_S16_L001_R1_001.fastq.gz	Abacumov-B-49_S49_L001_R1_001.fastq.g
z	
Abacumov-B-16_S16_L001_R2_001.fastq.gz	Abacumov-B-49_S49_L001_R2_001.fastq.g
z	
Abacumov-B-1_S1_L001_R1_001.fastq.gz	Abacumov-B-4_S4_L001_R1_001.fastq.gz
Abacumov-B-1_S1_L001_R2_001.fastq.gz	Abacumov-B-4_S4_L001_R2_001.fastq.gz
Abacumov-B-25_S25_L001_R1_001.fastq.gz	Abacumov-B-50_S50_L001_R1_001.fastq.g
z	
Abacumov-B-25_S25_L001_R2_001.fastq.gz	Abacumov-B-50_S50_L001_R2_001.fastq.g
z	
Abacumov-B-26_S26_L001_R1_001.fastq.gz	Abacumov-B-51_S51_L001_R1_001.fastq.g
z	
Abacumov-B-26_S26_L001_R2_001.fastq.gz	Abacumov-B-51_S51_L001_R2_001.fastq.g
z	
Abacumov-B-27_S27_L001_R1_001.fastq.gz	Abacumov-B-52_S52_L001_R1_001.fastq.g
z	
Abacumov-B-27_S27_L001_R2_001.fastq.gz	Abacumov-B-52_S52_L001_R2_001.fastq.g
z	
Abacumov-B-28_S28_L001_R1_001.fastq.gz	Abacumov-B-61_S61_L001_R1_001.fastq.g
z	
Abacumov-B-28_S28_L001_R2_001.fastq.gz	Abacumov-B-61_S61_L001_R2_001.fastq.g
z	
Abacumov-B-2_S2_L001_R1_001.fastq.gz	Abacumov-B-62_S62_L001_R1_001.fastq.g
z	
Abacumov-B-2_S2_L001_R2_001.fastq.gz	Abacumov-B-62_S62_L001_R2_001.fastq.g
z	
Abacumov-B-37_S37_L001_R1_001.fastq.gz	Abacumov-B-63_S63_L001_R1_001.fastq.g
z	
Abacumov-B-37_S37_L001_R2_001.fastq.gz	Abacumov-B-63_S63_L001_R2_001.fastq.g
z	
Abacumov-B-38_S38_L001_R1_001.fastq.gz	Abacumov-B-64_S64_L001_R1_001.fastq.g
z	
Abacumov-B-38_S38_L001_R2_001.fastq.gz	Abacumov-B-64_S64_L001_R2_001.fastq.g
z	

```
In [3]: !qiime tools import \
  --type 'SampleData[PairedEndSequencesWithQuality]' \
  --input-path data_met/sequences \
  --input-format CasavaOneEightSingleLanePerSampleDirFmt \
  --output-path demux-paired-end.qza
```

Imported data_met/sequences as CasavaOneEightSingleLanePerSampleDirFmt to demux-paired-end.qza

Deblur

```
In [4]: #фильтрация последовательностей на основе оценки качества
!qiime quality-filter q-score \
  --i-demux demux-paired-end.qza \
  --o-filtered-sequences demux-filtered.qza \
  --o-filter-stats demux-filter-stats.qza
```

Saved SampleData[SequencesWithQuality] to: demux-filtered.qza
Saved QualityFilterStats to: demux-filter-stats.qza

```
In [5]: #с помощью deblur объединяются похожие последовательности и удаляются ошибки
!qiime deblur denoise-16S \
  --i-demultiplexed-seqs demux-filtered.qza \
  --p-trim-length 120 \
  --o-representative-sequences rep-seqs-deblur.qza \
  --o-table table-deblur.qza \
  --p-sample-stats \
  --o-stats deblur-stats.qza
```

Saved FeatureTable[Frequency] to: table-deblur.qza
Saved FeatureData[Sequence] to: rep-seqs-deblur.qza
Saved DeblurStats to: deblur-stats.qza

```
In [6]: # табулирование статистики фильтрации для визуализации.
# Результаты сохраняются в файле demux-filter-stats.qzv.
!qiime metadata tabulate \
  --m-input-file demux-filter-stats.qza \
  --o-visualization demux-filter-stats.qzv
!qiime deblur visualize-stats \
  --i-deblur-stats deblur-stats.qza \
  --o-visualization deblur-stats.qzv
```

Saved Visualization to: demux-filter-stats.qzv
Saved Visualization to: deblur-stats.qzv

Оценить эффективность фильтрации и качество данных можно с помощью визуализации полученной таблицы:

qiime2view	File: deblur-stats.qzv	Visua
------------	------------------------	-------

Per-sample Deblur stats

Click on a Column header to sort the table.

Mouse over a Column header to get a description.

	sample-id	reads-raw	fraction-artifact-with-minsize	fraction-artifact	fraction-missed-reference	unique-reads-derep	reads-derep	unique-reads-deblur	reads-deblur	unique-reads-hit-artifact	reads-hit-artifact	unique-reads-chimeric	reads-chimeric	unique-reads-hit-reference	reads-hit-reference	unique-reads-missed-reference	reads-missed-reference
0	Abacumov-B-2	45925	0.506739	0.0	0.002075	3457	22653	2389	15922	0	0	300	502	1132	13380	5	32
1	Abacumov-B-40	28529	0.498125	0.0	0.000819	2203	14318	1582	10044	0	0	162	271	896	8742	3	8
2	Abacumov-B-4	40053	0.497416	0.0	0.001618	3163	20130	2313	14610	0	0	258	392	1140	12324	5	23
3	Abacumov-B-1	43208	0.496922	0.0	0.001564	3325	21737	2323	15748	0	0	231	398	1102	13228	6	24
4	Abacumov-B-39	26750	0.486953	0.0	0.000735	2123	13724	1496	9737	0	0	120	207	866	8552	2	7
5	Abacumov-B-38	32977	0.464657	0.0	0.001741	2575	17654	1772	12358	0	0	171	296	937	10711	4	21
6	Abacumov-B-64	37802	0.460769	0.0	0.022824	2858	20384	1732	13625	0	0	167	262	874	11726	16	305
7	Abacumov-B-3	54892	0.449446	0.0	0.001985	4052	30221	2734	21691	0	0	307	532	1208	18477	7	42
8	Abacumov-B-37	31422	0.448730	0.0	0.001599	2525	17322	1747	12142	0	0	155	263	940	10588	4	19
9	Abacumov-B-62	39664	0.446879	0.0	0.013072	3007	21939	1869	14850	0	0	154	239	945	12869	15	191

Add metadata

In [7]: `!sed 's/,/\t/g' map.csv > map.tsv #преобразование файла метаданных map.csv в`

In [8]: `!sed -i 's/Filename/#SampleID/' map.tsv #замена"Filename" на "SampleID"`

In [11]: `# Создание сводной информации о таблице признаков
!qiime feature-table summarize \
--i-table table-deblur.qza \
--o-visualization table.qzv \
--m-sample-metadata-file map.tsv

Создание таблицы с информацией о последовательностях (включая идентификаторы)
!qiime feature-table tabulate-seqs \
--i-data rep-seqs-deblur.qza \
--o-visualization rep-seqs.qzv`

Saved Visualization to: table.qzv

Saved Visualization to: rep-seqs.qzv

qiime2view

File: table.qzv

VisualizationDetailsProvenance

OverviewInteractive Sample DetailFeature Detail

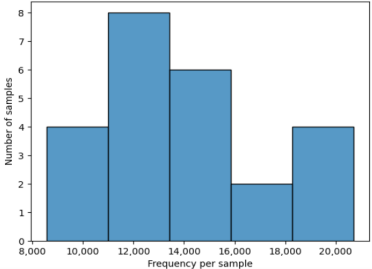
Table summary

Metric	Sample
Number of samples	24
Number of features	3,606
Total frequency	339,008

Frequency per sample

	Frequency
Minimum frequency	8,552.0
1st quartile	12,174.5
Median frequency	13,460.0
3rd quartile	16,066.75
Maximum frequency	20,720.0
Mean frequency	14,125.333333333334

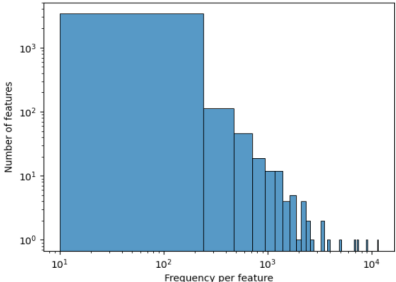
Frequency per sample detail (csv | html)



Frequency per feature

	Frequency
Minimum frequency	10.0
1st quartile	15.0
Median frequency	27.0
3rd quartile	65.0
Maximum frequency	11,821.0
Mean frequency	94.01220188574598

Frequency per feature detail (csv | html)



qiime2view

File: rep-seqs.qzv

VisualizationDetailsProvenance

Sequence Length Statistics

Download sequence-length statistics as a TSV

Sequence Count	Min Length	Max Length	Mean Length	Range	Standard Deviation
3696	120	120	120.0	0	0.0

Seven-Number Summary of Sequence Lengths

Download seven-number summary as a TSV

Percentile:	2%	9%	25%	50%	75%	91%	98%
Length* (nts):	120	120	120	120	120	120	120

*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
01ebc570168778cc37b84b2f5289c5	120	GTGCCAGCAGCCGCGTAAAGCAGGAGGATGCGAGCGTTATCCGGAATTAGTTGGGCGTAAAGGCGTACAGGCGGCTGATCCCGCTGTCAAGCGCTGGGGCTCAACCTCTGAAAGGC
d9d2d0ac3b34949ba064c0fbcd68ef	120	GTGCCAGCAGCCGCGGTAATACGAGAGGAGGCTAGCGTTGTTCCGAATTACTGGGCGTAAAGCGTACAGGCGGCTTGTAGATTAGAGGTGAAGGCCCGGGGCTCAACTCCGGAATTGC
f3781fdb2dfb24889739e0611361ab7	120	GTGCCAGCAGCCGCGGTAAAGCAGAACCGTGCAGCGTTGTTCCGAATTACTGGGCTTAAAGGCGCGTAGAGCGGACACGAGCTAGGGTGAATCTTTCAAGCTTAACCTGGAAGATGC
9eeff2a8666729f5e2d2a4e27b18d	120	GTGCCAGCAGCCGCGGTAATACGAGAGGTCAGCGTTATCCGGAATTACTGGGTTTAAAGGTCGTAGGCGGCGAGTAAAGTCAAGTGGTGAATCTTCAAGCTTAACCTGGAAGATGC
91f1f733cc05e1e11f1c8aa3af5c68db	120	GTGCCAGCAGCCGCGGTAATACGAGAGGAGGCGAGCGTTGTTCCGAATTATTGGGCGTAAAGGCGCGTAGAGTGGTTCGCGAGCTTGTGTGAATCTTCAAGCTCAACTGGAAGTCTGC
c1cb5b1e5b5cd768a876e75f883d0	120	GTGCCAGCAGCCGCGGTAATACGAGAGGAGGCTAGCGTTGTTCCGAATTACTGGGCGTAAAGCGCGCGTAGGCGGCTTGTAGATTAGAGGTGAAGGCCGAGGCTCAACTCCGGAATTGC
32b1e1d359eabdbb88e17551b210526	120	GTGCCAGCAGCCGCGGTAATACGAGAGGTCGAGCGTTGTTCCGAATTATTGGGCGTAAAGGCGCGTAGAGTGGCGGTCAGTCAAGTGGTGAAGCGCGGGGCTCAACCCCGCGTCGCG
fb84e58bcd0a620a672bcd9d19d6293a	120	GTGCCAGCAGCCGCGGTAATACGAGAGGTCAGAGCGTTATCCGGAATTACTGGGTTTAAAGGTCGTAGGAGGCGAGTAAAGTCAAGTGGTGAATCTTCAAGCTTAACCTGGAAGATGC
1ed59ad9e7fd11ac852428096e25e021	120	GTGCCAGCAGCCGCGGTAATACGAGAGGTCAGCGTTGTTCCGGAATTATTGGGTTTAAAGGTCGTAGAGTGGCGTCTTAAGTCTGGTTTGAAGCAGCGCGGCTCAACCTGCTGATGTG
224f92393dfb745260985559268a71	120	GTGCCAGCAGCCGCGGTAATACGAGAGGTCAGCGTTGTTCCGGAATTACTGGGCGTAAAGGCGCGTAGAGCGGCGTGTACGACCGCGGTGAAGCCCCCGGCTCAACTGGGAGGGTC

Add tree

```
In [12]: # QIIME выравнивает последовательности и строит филогенетическое дерево:
!qiime phylogeny align-to-tree-mafft-fasttree \
  --i-sequences rep-seqs-deblur.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
```

Saved FeatureData[AlignedSequence] to: aligned-rep-seqs.qza
Saved FeatureData[AlignedSequence] to: masked-aligned-rep-seqs.qza
Saved Phylogeny[Unrooted] to: unrooted-tree.qza
Saved Phylogeny[Rooted] to: rooted-tree.qza

Diversity Analysis

Анализ разнообразия микробиома на уровне альфа- и бета-разнообразия

In [14]: *#QIIME рассчитывает различные метрики альфа- и бета-разнообразия:*

```
!qiime diversity core-metrics-phylogenetic \  
  --i-phylogeny rooted-tree.qza \  
  --i-table table-deblur.qza \  
  --p-sampling-depth 8000 \  
  --m-metadata-file map.tsv \  
  --output-dir core-metrics-results
```

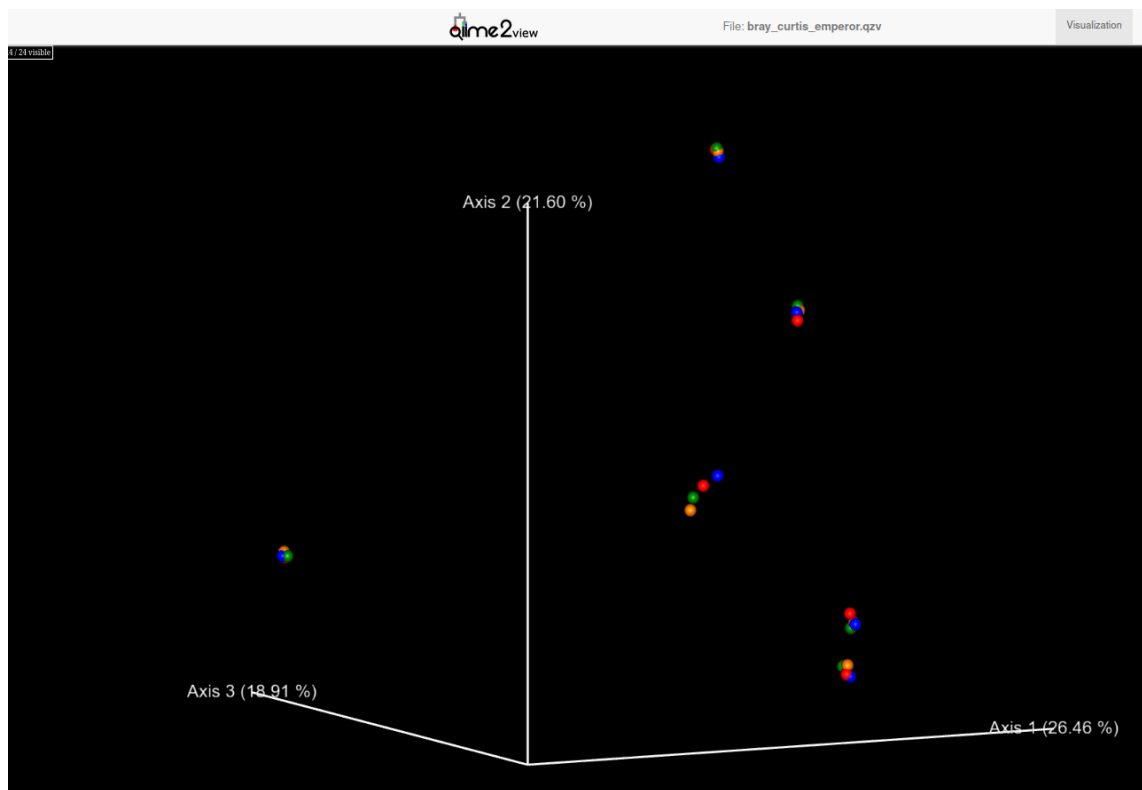
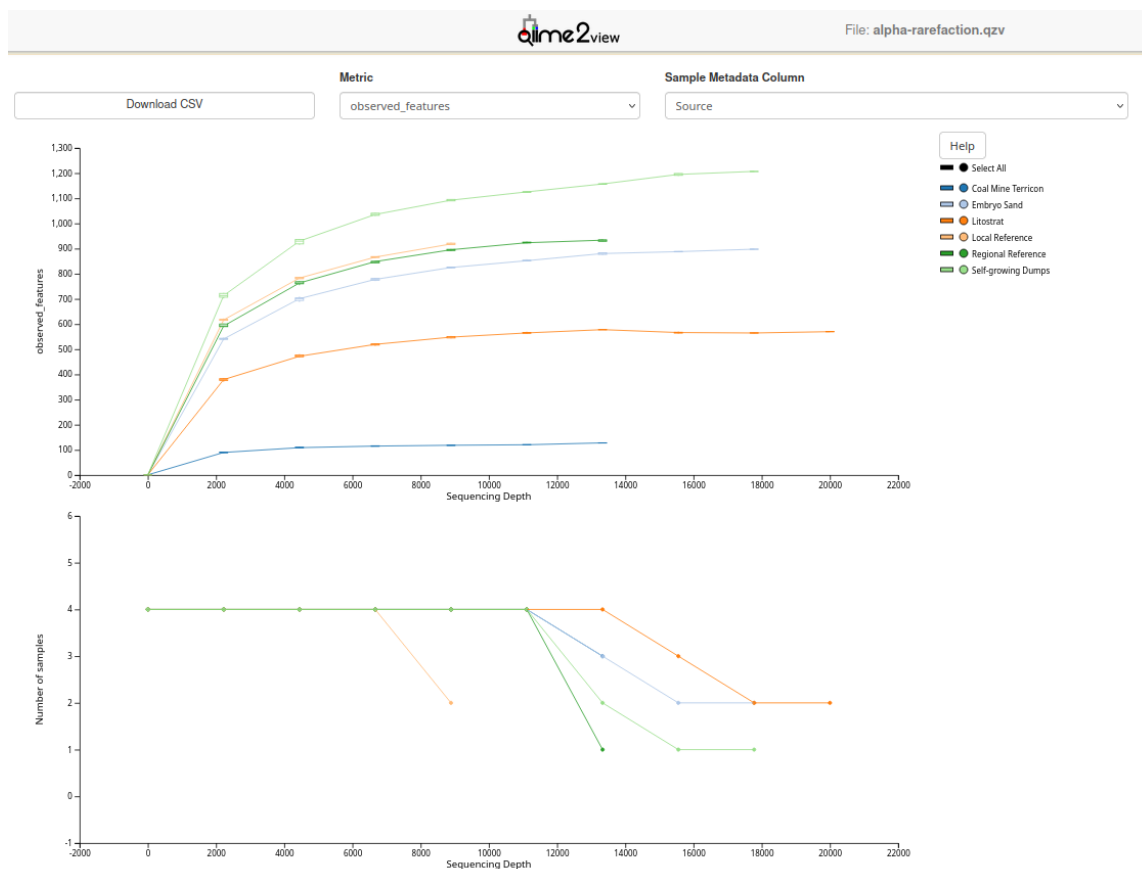
Saved FeatureTable[Frequency] to: core-metrics-results/rarefied_table.qza
Saved SampleData[AlphaDiversity] to: core-metrics-results/faith_pd_vector.qza
Saved SampleData[AlphaDiversity] to: core-metrics-results/observed_features_vector.qza
Saved SampleData[AlphaDiversity] to: core-metrics-results/shannon_vector.qza
Saved SampleData[AlphaDiversity] to: core-metrics-results/evenness_vector.qza
Saved DistanceMatrix to: core-metrics-results/unweighted_unifrac_distance_matrix.qza
Saved DistanceMatrix to: core-metrics-results/weighted_unifrac_distance_matrix.qza
Saved DistanceMatrix to: core-metrics-results/jaccard_distance_matrix.qza
Saved DistanceMatrix to: core-metrics-results/bray_curtis_distance_matrix.qza
Saved PCoAResults to: core-metrics-results/unweighted_unifrac_pcoa_results.qza
Saved PCoAResults to: core-metrics-results/weighted_unifrac_pcoa_results.qza
Saved PCoAResults to: core-metrics-results/jaccard_pcoa_results.qza
Saved PCoAResults to: core-metrics-results/bray_curtis_pcoa_results.qza
Saved Visualization to: core-metrics-results/unweighted_unifrac_emperor.qzv
Saved Visualization to: core-metrics-results/weighted_unifrac_emperor.qzv
Saved Visualization to: core-metrics-results/jaccard_emperor.qzv
Saved Visualization to: core-metrics-results/bray_curtis_emperor.qzv

Visualise alpha-diversity

In [15]:

```
!qiime diversity alpha-rarefaction \  
  --i-table table-deblur.qza \  
  --i-phylogeny rooted-tree.qza \  
  --p-max-depth 20000 \  
  --m-metadata-file map.tsv \  
  --o-visualization alpha-rarefaction.qzv
```

Saved Visualization to: alpha-rarefaction.qzv

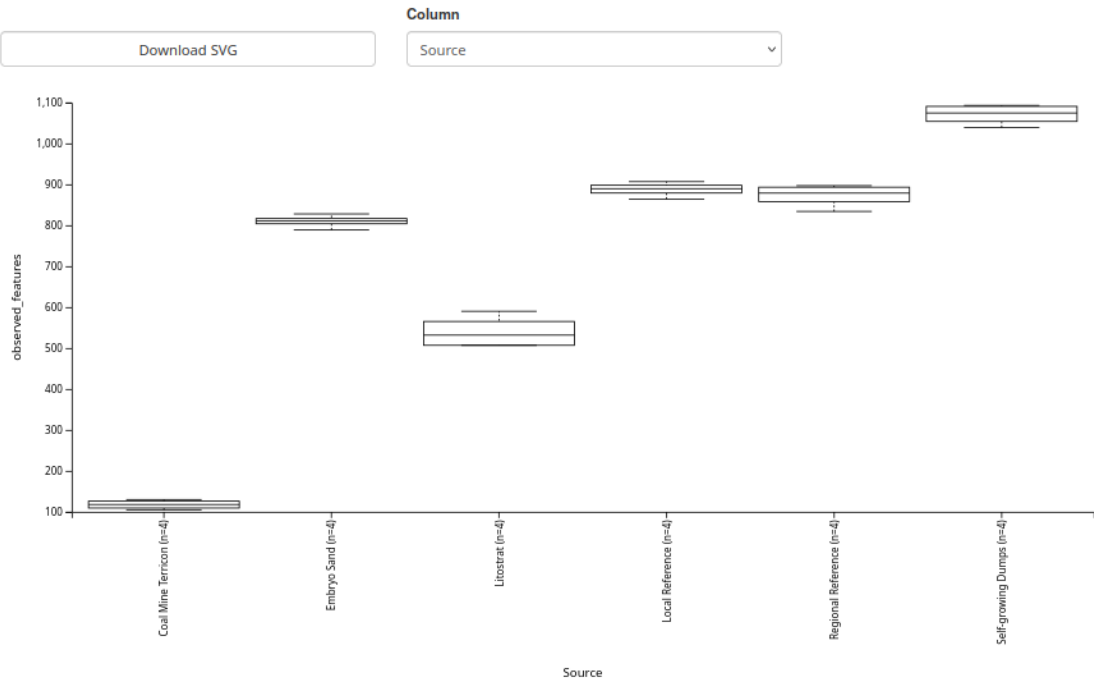


```
In [16]: #статистическая оценка значимости альфа-разнообразия между группами (на основе
!qiime diversity alpha-group-significance \
--i-alpha-diversity core-metrics-results/observed_features_vector.qza \
--m-metadata-file map.tsv \
--o-visualization core-metrics-results/observed_features-significance.qzv
```

```
!qiime diversity alpha-group-significance \
  --i-alpha-diversity core-metrics-results/evenness_vector.qza \
  --m-metadata-file map.tsv \
  --o-visualization core-metrics-results/evenness-group-significance.qzv
```

Saved Visualization to: core-metrics-results/observed_features-significance.qzv
 Saved Visualization to: core-metrics-results/evenness-group-significance.qzv

Alpha Diversity Boxplots



[Download raw data as TSV](#)

Kruskal-Wallis (all groups)

Result	
H	21.80948238364506
p-value	0.0005691073139090535

Kruskal-Wallis (pairwise)

[Download CSV](#)

		H	p-value	q-value
Group 1	Group 2			
Coal Mine Terricon (n=4)	Embryo Sand (n=4)	5.333333	0.020921	0.024140
	Litostrat (n=4)	5.333333	0.020921	0.024140
	Local Reference (n=4)	5.333333	0.020921	0.024140
	Regional Reference (n=4)	5.333333	0.020921	0.024140
	Self-growing Dumps (n=4)	5.333333	0.020921	0.024140
Embryo Sand (n=4)	Litostrat (n=4)	3.000000	0.083265	0.089212
	Local Reference (n=4)	5.333333	0.020921	0.024140
	Regional Reference (n=4)	0.333333	0.563703	0.563703
	Self-growing Dumps (n=4)	5.333333	0.020921	0.024140
Litostrat (n=4)	Local Reference (n=4)	5.333333	0.020921	0.024140
	Regional Reference (n=4)	5.333333	0.020921	0.024140
	Self-growing Dumps (n=4)	5.333333	0.020921	0.024140
Local Reference (n=4)	Regional Reference (n=4)	5.333333	0.020921	0.024140
	Self-growing Dumps (n=4)	5.333333	0.020921	0.024140
Regional Reference (n=4)	Self-growing Dumps (n=4)	5.333333	0.020921	0.024140

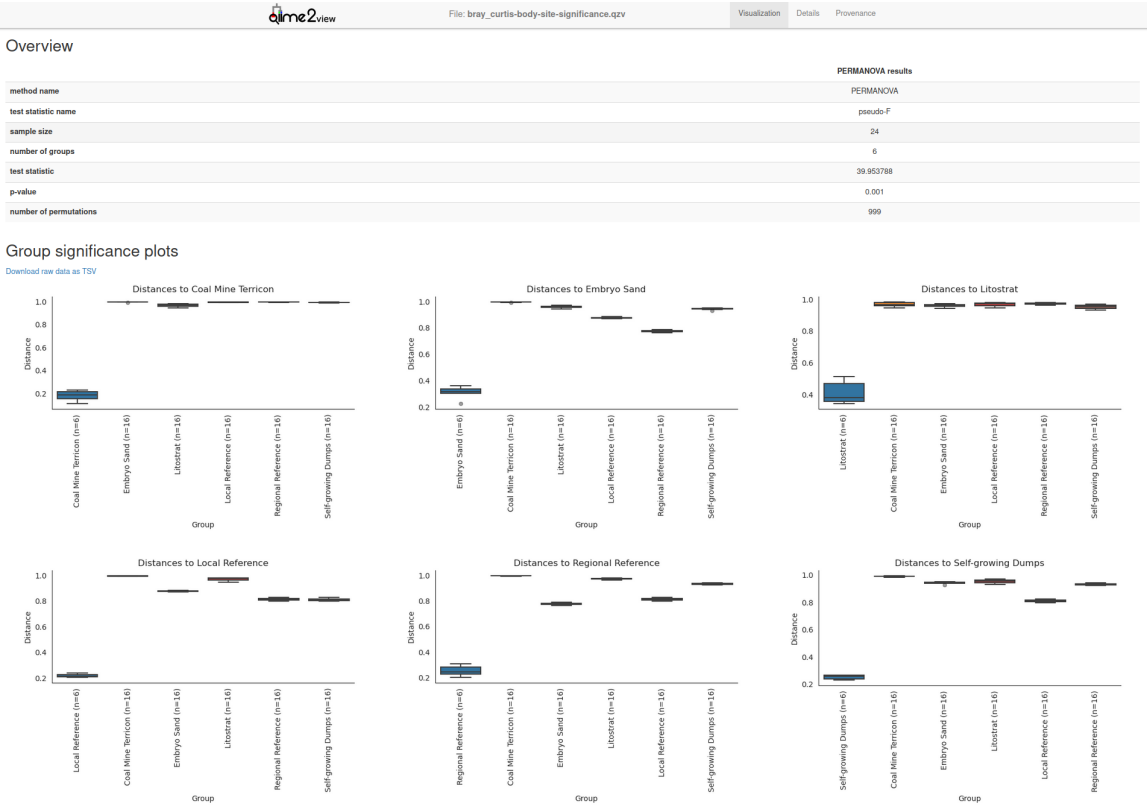
Тест Kruskal-Wallis: Нулевая гипотеза в этом тесте заключается в том, что распределения значений во всех группах равны.

На основании полученных результатов можно сделать вывод, что существует статистически значимая разница между группами. Значение p-value меньше уровня значимости 0.05, что позволяет отвергнуть нулевую гипотезу. Это означает, что распределения значений в разных группах статистически отличаются, и существует вероятность, что эти различия не случайны.

Между группами "Coal Mine Terricon" и "Embryo Sand" наблюдается статистически значимая разница (p-value = 0.020921), с учетом поправки на множественные сравнения (q-value = 0.024140). Между группами "Litostrat" и "Embryo Sand" также наблюдается статистически значимая разница (p-value = 0.083265), но после коррекции на множественные сравнения эта разница становится менее значимой (q-value = 0.089212). Между другими парами групп также наблюдаются значения статистики H, p-value и q-value, которые указывают на различия между группами, но точность этих различий может быть несколько ниже из-за коррекции на множественные сравнения.

```
In [17]: # Beta-diversity significance
!qiime diversity beta-group-significance \
  -i distance-matrix core-metrics-results/bray_curtis_distance_matrix.qza \
  -m metadata-file map.tsv \
  -m metadata-column Source \
  -o visualization core-metrics-results/bray_curtis-body-site-significance.
  -p pairwise
```

Saved Visualization to: core-metrics-results/bray_curtis-body-site-significance.qzv



		Sample size	Permutations	pseudo-F	p-value	q-value
Group 1	Group 2					
Coal Mine Terricon	Embryo Sand	8	999	56.592534	0.022	0.038077
	Litostrat	8	999	32.713719	0.036	0.038571
	Local Reference	8	999	93.124470	0.031	0.038077
	Regional Reference	8	999	75.902149	0.027	0.038077
	Self-growing Dumps	8	999	76.843551	0.030	0.038077
Embryo Sand	Litostrat	8	999	23.862476	0.031	0.038077
	Local Reference	8	999	39.064722	0.030	0.038077
	Regional Reference	8	999	26.174828	0.021	0.038077
	Self-growing Dumps	8	999	40.902108	0.033	0.038077
Litostrat	Local Reference	8	999	30.616369	0.029	0.038077
	Regional Reference	8	999	28.365940	0.031	0.038077
	Self-growing Dumps	8	999	27.351967	0.021	0.038077
Local Reference	Regional Reference	8	999	43.515447	0.025	0.038077
	Self-growing Dumps	8	999	43.960127	0.039	0.039000
Regional Reference	Self-growing Dumps	8	999	50.620822	0.023	0.038077

Анализ PERMANOVA подтверждает, что существует статистически значимая разница между группами на основе анализа многомерных данных.