

## Задача 1.

Результаты fastqc:

- [Basic Statistics](#)
- [Per base sequence quality](#)
- [Per sequence quality scores](#)
- [Per base sequence content](#)
- [Per sequence GC content](#)
- [Per base N content](#)
- [Sequence Length Distribution](#)
- [Sequence Duplication Levels](#)
- [Overrepresented sequences](#)
- [Adapter Content](#)

Пройдусь по несоответствиям:

Per base sequence content

видны высоченные пики гуанина, а его много в праймере, так что это может говорить о секвенировании в основном праймеров.

Per sequence GC content

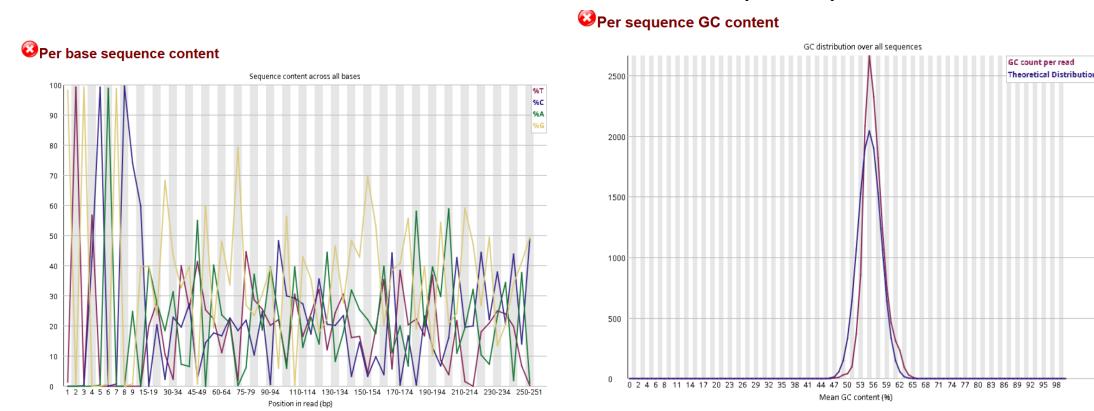
график сильно не похож на теоретический, видно высокий узкий пик, это может говорить о том, что он не показывает реальный GC-состав бактерий, а GC-состав праймера (а он не разнообразный)

Sequence Duplication Levels

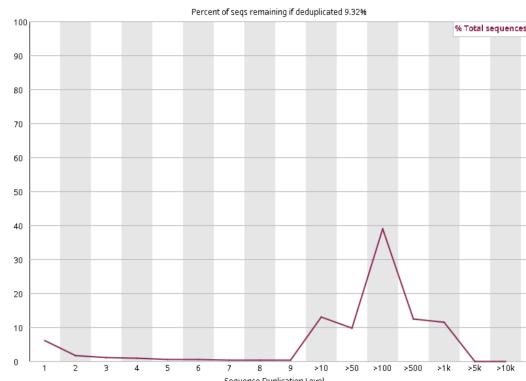
percent of segs remaining if deduplicated 9.32% - очень мало останется ридов, если уберем дубликаты. очень высокий уровень дублирования -> праймеры.

Overrepresented sequences

видно много последовательностей, похожих на наш праймер.



### Sequence Duplication Levels



### Overrepresented sequences

| Sequence  | Count | Percentage         | Possible Source |
|---|-------|--------------------|-----------------|
| GTGTCAGCCGCCGGTAACTGTAGGGTGCAGCGTTAACGGAAATTAC    | 1818  | 6.70569574985181   | No Hit          |
| GTGCCAGCCGCCGGTAACTGTAGGGTCAAGCGTTAACGGAAATTAC    | 1328  | 4.319976289270895  | No Hit          |
| GTGTCAGCCGCCGGTAACTACAGAGGGTCAAGCGTTAACGGAAATTAC  | 896   | 3.319582074688797  | No Hit          |
| GTGCCAGCCGCCGGTAACTACAGAGGGTCAAGCGTTAACGGAAATTAC  | 699   | 2.589656194427979  | No Hit          |
| GTGTCAGCCGCCGGTAACTAGCTAGGGGGCAAGCGTTAACGGAAATTAC | 688   | 2.53778974510966   | No Hit          |
| GTGTCAGCAGCCGGTAACTACGTAGGGTCAAGCGTTAACGGAAATTAC  | 613   | 2.2718432720806164 | No Hit          |
| GTGTCAGCCGCCGGTAACTACGAAGGGGGTAGCGTTCTGGGAATTAC   | 502   | 1.8598103141671607 | No Hit          |
| GTGCCAGCCGCCGGTAACTAGCTAGGGGGCAAGCGTTAACGGAAATTAC | 49    | 1.8198574985180796 | No Hit          |
| GTGTCAGCCGCCGGTAACTACGTAGGGGTGGAGCGTTAACGGAAATTAC | 478   | 1.7708950800237107 | No Hit          |
| GTGCCAGCCGCCGGTAACTACGTAGGGGGCAAGCGTTAACGGAAATTAC | 437   | 1.6189982216953174 | No Hit          |
| GTGTCAGCCGCCGGTAACTACGTAGGGGGCAAGCGTTAACGGAAATTAC | 421   | 1.574540646235921  | No Hit          |
| GTGCCAGCCGCCGGTAACTACGTAGGGGGCAAGCGTTAACGGAAATTAC | 381   | 1.4115293420272674 | No Hit          |
| GTGCCAGCCGCCGGTAACTACGTAGGGGGCAAGCGTTAACGGAAATTAC | 365   | 1.3522525192649675 | No Hit          |
| GTGCCAGCCGCCGGTAACTACGTAGGGGGCAAGCGTTAACGGAAATTAC | 349   | 1.2929756965026673 | No Hit          |

результат задачи 1:

```
(qiime2-amplicon-2024.10) aandreeva@frontend-1-2-13:~/hw/hw_15/metagenome/qza$ ls
soil_ASV_table.qza soil_reads.dada2.stats.qza soil_reads.qza soil_rep_seq.qza
```

Задача 2.

--p-trim-left 25 нужен, чтобы обрезать праймеры (они примерно такой длины и не информативны)

qiime2

qiime2-view

File: soil\_reads.dada2.stats.qzv ×

Visualization Citations Provenance Metadata

Download metadata TSV file

This file won't necessarily reflect dynamic sorting or filtering options based on the interactive table below.

Search:

| sample-id<br>[qzotypes] | input<br>[numeric] | filtered<br>[numeric] | percentage of input passed filter<br>[numeric] | denoised<br>[numeric] | non-chimeric<br>[numeric] | percentage of input non-chimeric<br>[numeric] |
|-------------------------|--------------------|-----------------------|--|-----------------------|---------------------------|---|
| SRR17307258             | 26992              | 26932                 | 99.78  | 26472                 | 25395                     | 94.08   |
| SRR17307262             | 5996               | 5926                  | 98.83  | 5754                  | 5731                      | 95.58   |
| SRR17307269             | 10387              | 10322                 | 99.37  | 9859                  | 9859                      | 94.92   |
| SRR17307273             | 23451              | 23405                 | 99.8   | 23023                 | 20966                     | 89.4  |
| SRR17307278             | 24834              | 24783                 | 99.79  | 24230                 | 23607                     | 95.06   |
| SRR17307316             | 10833              | 10723                 | 98.98  | 10193                 | 10189                     | 94.06   |
| SRR17307364             | 8661               | 8605                  | 98.35  | 8270                  | 8270                      | 95.49   |
| SRR17307380             | 5128               | 5087                  | 99.2   | 4830                  | 4809                      | 93.78   |
| SRR17307392             | 25424              | 25374                 | 99.8   | 24617                 | 24547                     | 96.55   |
| SRR17307400             | 9513               | 9351                  | 98.3   | 9127                  | 9077                      | 95.42   |
| SRR17307404             | 8614               | 8551                  | 99.27  | 8158                  | 8141                      | 94.51   |
| SRR17307406             | 22460              | 22420                 | 99.82  | 21987                 | 21190                     | 94.35   |

Задача 3.

около 93-95% остается после всех фильтраций (Percentage of input non-chimeric) по количеству non-chimeric:

самый большой: 45518

самый маленький: 4866

на самом деле процент сохранения довольно большой, это подозрительно с учетом того, что мы анализируем метагеномные данные(а они шумные), так что это ещё раз что основная часть данных - артефакты праймеров.

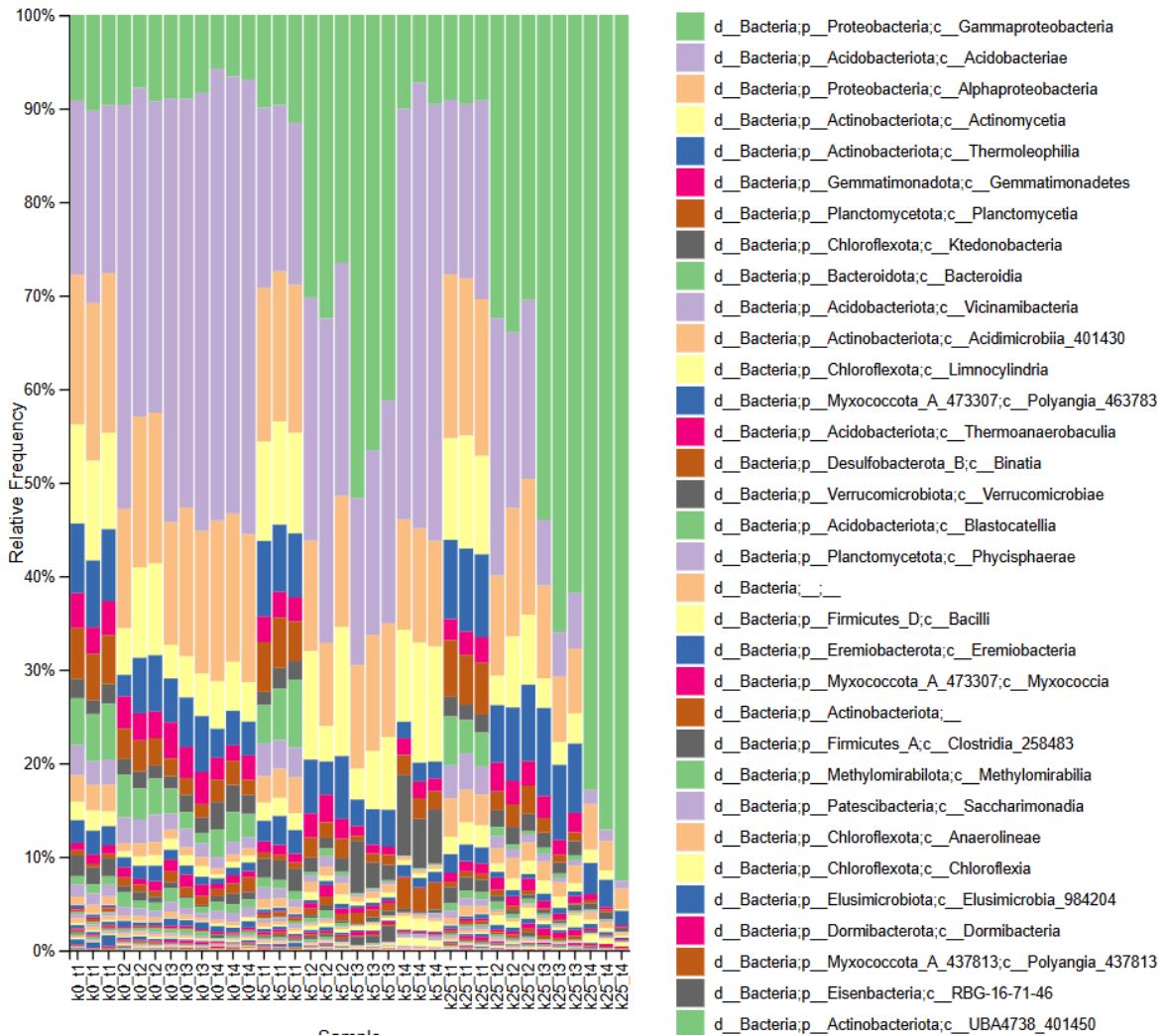
результат 3 задачи:

```
(qiime2-amplicon-2024.10) aandreeva@frontend-1-2-13:~/hw/hw_15/metagenome/qza$ ls
(qiime2-amplicon-2024.10) aandreeva@frontend-1-2-13:~/hw/hw_15/metagenome/qza$ ls
soil_ASV_table.qza soil_reads.dada2.stats.qza soil_reads.qza soil_rep_seq.qza soil_taxonomy.qza
```

Задача 4.

классификатор, обученный на V4, содержит достаточно таксономической информации для надёжной аннотации; классификатор, обученный на том же гипервариабельном регионе, даёт точнее результаты, чем полный классификатор.

Задача 5.



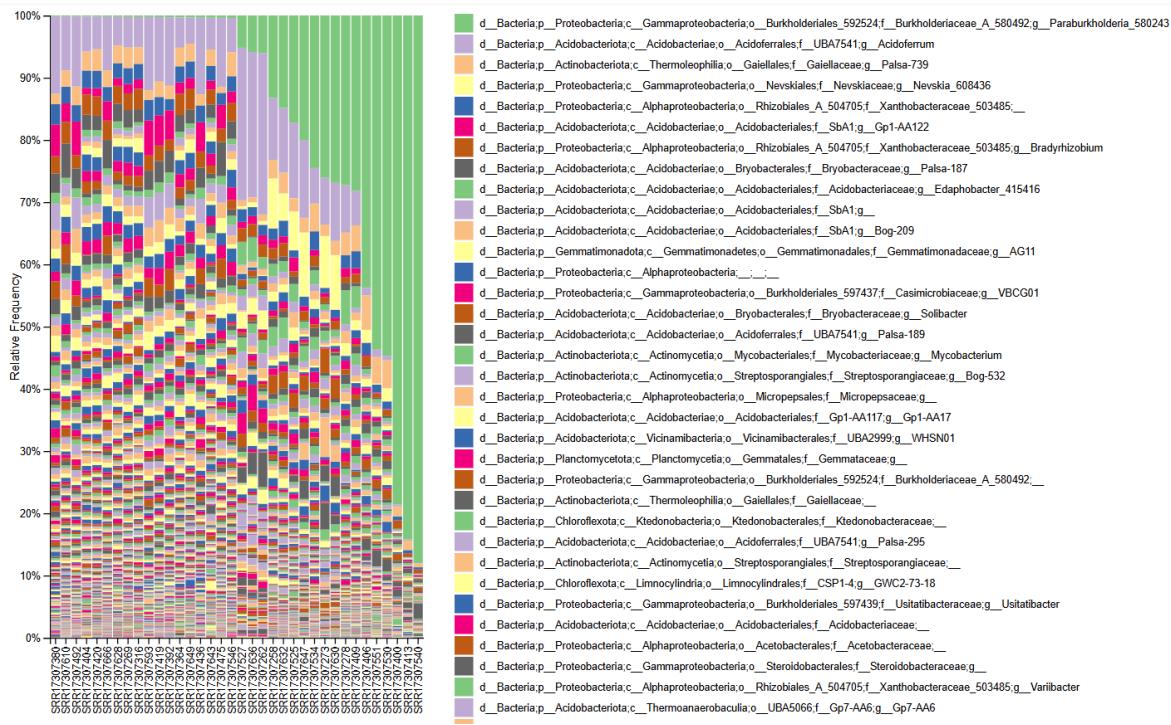
В первые образцы преобладает фиолетовый и оранжевый цвета. Сходство между ними довольно высокое — композиция бактериальных классов почти одинаковая, что логично для исходной, незагрязнённой среды.

Со временем видно увеличение зелёных фрагментов (Gammaproteobacteria), снижение фиолетового и оранжевого. Это указывает на сдвиг в структуре микробного сообщества, возможно из-за изменения условий среды.

Самые частые классы:

Первая временная точка: Acidobacteriia (фиолетовый) и Alphaproteobacteria (оранжевый).

Последняя временная точка: Gammaproteobacteria (зелёный), иногда Actinobacteria (жёлтый), видно доминирование этих классов после воздействия.



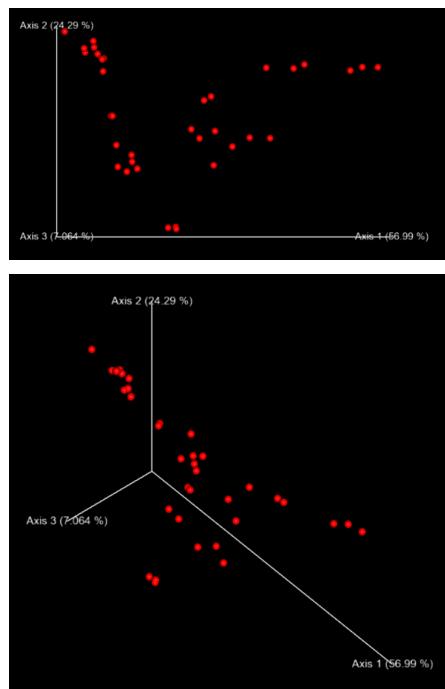
самым частым родом в загрязнённых образцах является Paraburkholderia.

Paraburkholderia - группа бактерий, которые хорошо разлагают углеводороды.

поскольку керосин состоит в основном из углеводородов, Paraburkholderia будут активно расти и доминировать в загрязнённой среде, используя керосин в качестве питательной среды.

Задача 6.

взяла X = 4830



видно разделение микробных сообществ: образцы первой временной точки (исходные) сгруппированы в левой верхней части (вдоль axis 2). после загрязнения все образцы смешаются вправо вдоль Axis 1 ( 56.99% дисперсии, получается тут основное

смещение). признаков восстановления к исходному состоянию не выявлено — ни один образец не возвращается к начальной группе.