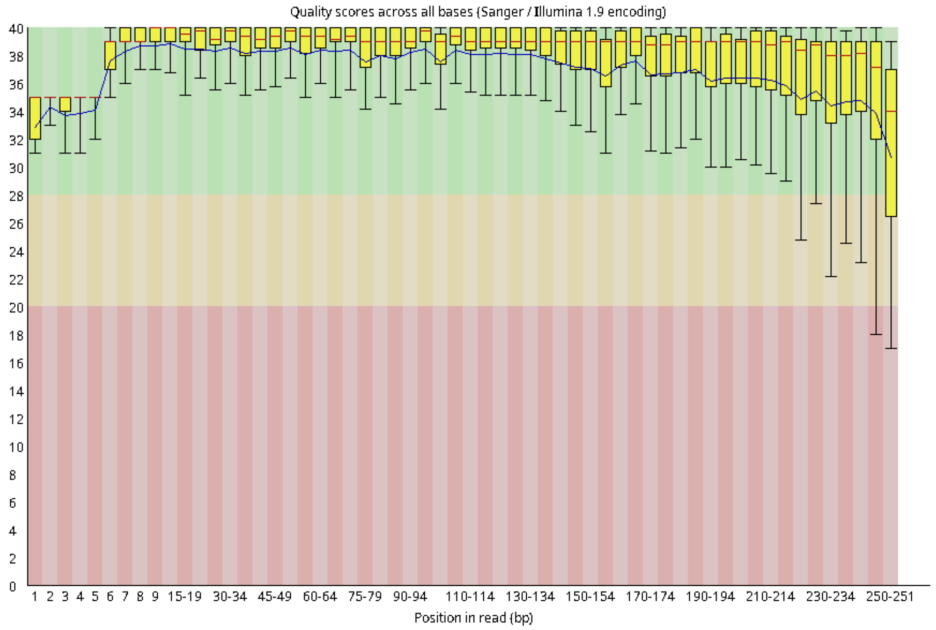
(**Задача 1**) Запустите fastqc на одном из образцов по выбору, опишите полученные результаты.
Выбрал /projects/mipt_dbmp_biotechnology/metagenomes/soil/raw_reads/SRR17307258_1.fastq

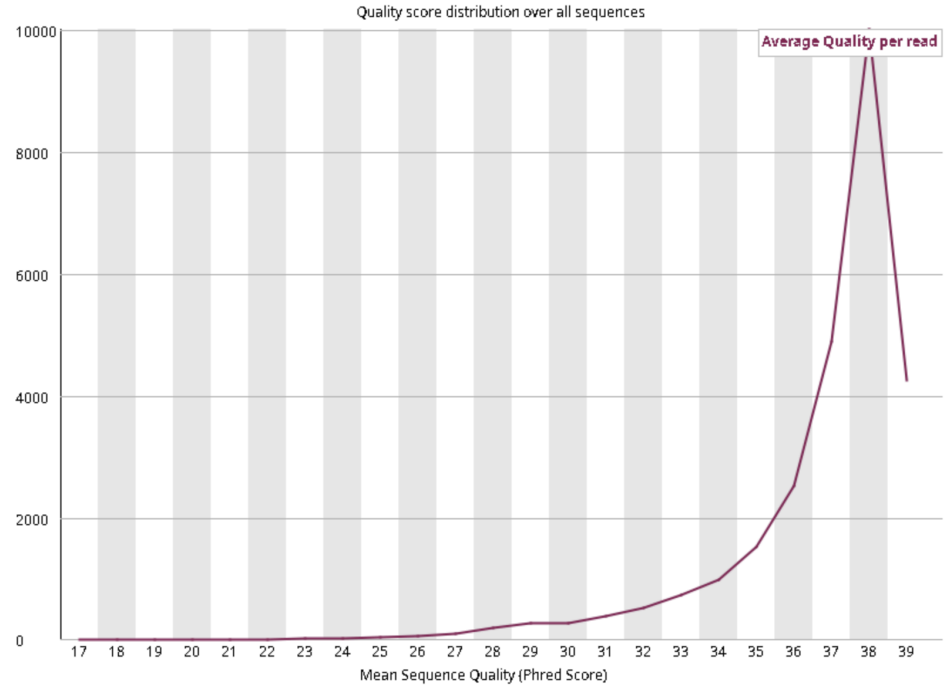
Качество прочтений хорошее, в среднем выше Q37. Ближе к концу качество слегка снижается, что является нормальным явлением. Наличие адаптерных последовательностей заметно, однако их количество невелико и находится в пределах допустимого.

Per base sequence quality



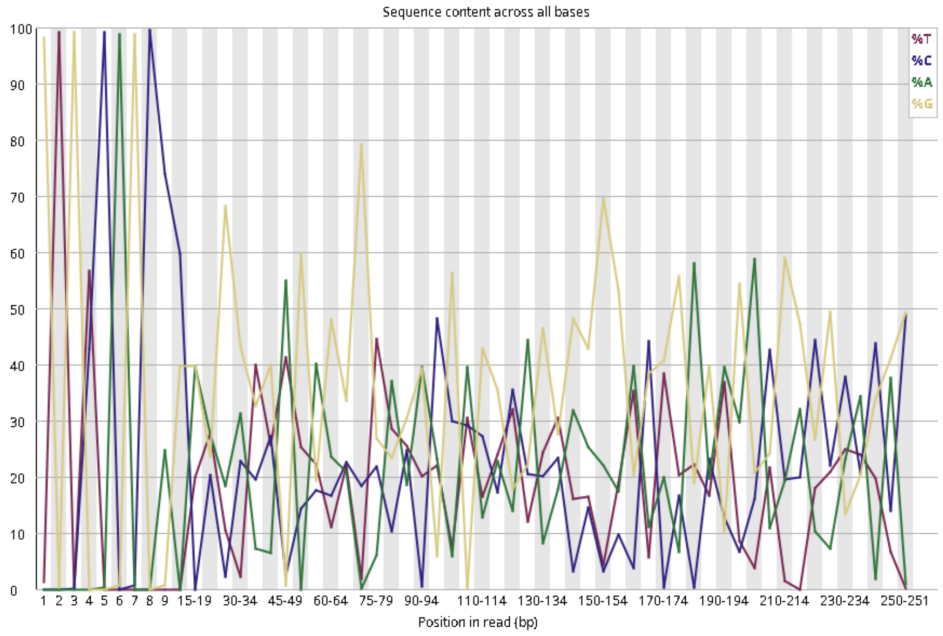
Пик среднего качества прочтений приходится на Q38, что является высоким показателем.

Per sequence quality scores



Такой профиль характерен для 16S rRNA: первые ~16 нуклеотидов соответствуют праймеру, поэтому на этих позициях один из нуклеотидов встречается почти на 100%, а остальные почти отсутствуют. Далее следует вариабельная область 16S, где каждая позиция представляет смесь последовательностей разных микроорганизмов.

Per base sequence content

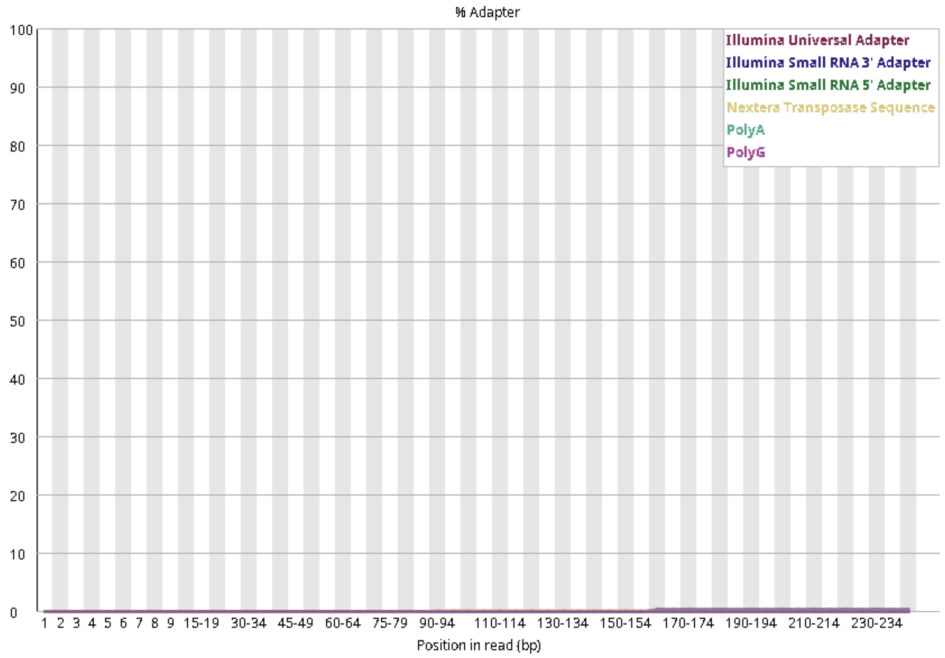


Overrepresented sequences встречаются, например, последовательность GTGTCAGCCGCCGCGGTAATACGTAGGGTGCAAGCGTTAATCGGAATTAC составляет 6.7%, что является нормой, так как это консервативный регион 16S rRNA.

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GTGTCAGCCGCCGCGGTAATACGTAGGGTGCAAGCGTTAATCGGAATTAC	1810	6.705690574985181	No Hit
GTGCCAGCCGCCGCGGTAATACGTAGGGTGCAAGCGTTAATCGGAATTAC	1328	4.919976289270895	No Hit
GTGTCAGCCGCCGCGGTAATACAGAGGGTGCAAGCGTTAATCGGATTTAC	896	3.319502074688797	No Hit
GTGCCAGCCGCCGCGGTAATACAGAGGGTGCAAGCGTTAATCGGATTTAC	699	2.589656194427979	No Hit
GTGTCAGCCGCCGCGGTAAGACGTAGGGGGCCAGCGTTGTTCGGAATTAC	685	2.537788974510966	No Hit
GTGTCAGCAGCCGCGGTAATACGTAGGGTGCAAGCGTTAATCGGAATTAC	613	2.2710432720806164	No Hit

Adapter Content



(**Задача 2**)
Зачем мы используем параметр **--p-trim-left 25**?

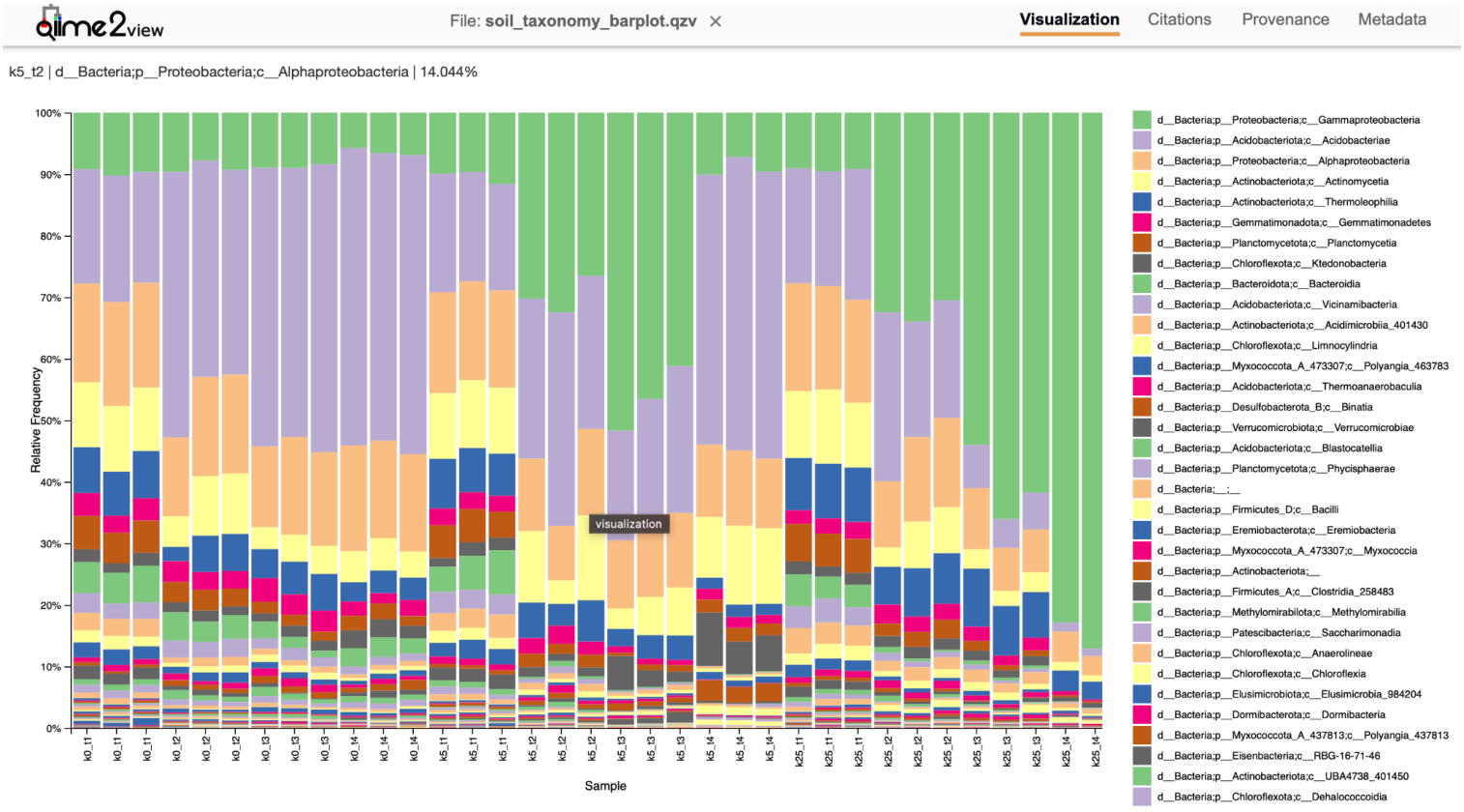
--p-trim-left 25 \ отрезаем 25 нуклеотидов слева у прочтения

Используем, чтобы удалить: сам праймер/шапку/артефакты праймерной амплификации и участки с низким качеством в начале прочтения, что улучшает качество входных данных и снижает ложные вариации, вызванные остатками праймеров.

(**Задача 3**)
Какая примерно доля исходных прочтений остаётся после всех стадий фильтрации?
+- 94%
Сколько прочтений осталось в самом большом и в самом маленьком образцах?
SRR17307475: 43518 и SRR17307380: 4809
Что полученные результаты на ваш взгляд говорят о качестве данных?
Даже в самом «слабом» образце сохраняется свыше 97% прочтений после всех этапов обработки, что свидетельствует о стабильности данных и отсутствии заметного ухудшения их качества.

(**Задача 4**)
Хотя праймеры амплифицировали регион V4–V5, мы использовали классификатор, обученный только на V4. Это работает, потому что V4 — самая информативная часть для идентификации бактерий и полностью входит в наш ампликон. То есть классификатор видит ту же ключевую последовательность, на которой он обучался, и может надежно определить таксономию до рода или семейства. Для наших целей этого вполне достаточно, а готовый V4-классификатор экономит время и ресурсы.

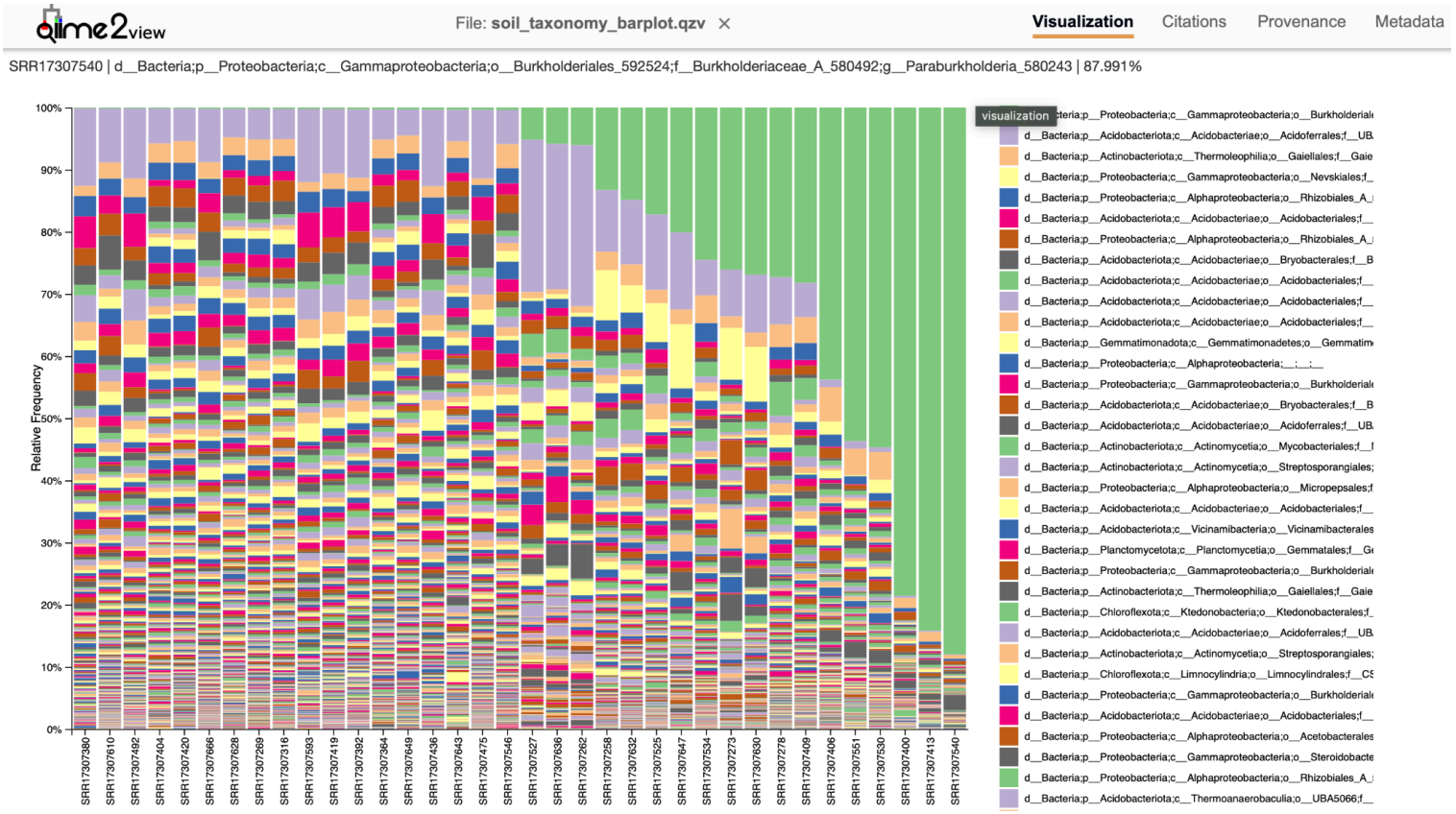
(**Задача 5**)



Авторы работы поместили образцы почвы в закрытые ёмкости и обработали их разным количеством керосина: контроль (k0 – без керосина), слабое загрязнение (k5 – 5 г/кг) и сильное загрязнение (k25 – 25 г/кг). На день 3 (d3, time point t1), 90 (d90, t2), 180 (d180, t3) и 360 (d360, t4) из ёмкостей отбирали немного почвы и выделяли тотальное ДНК с помощью кита DNeasy PowerSoil.

График распределения бактериальных классов (Level 3) показывает, как изменяется соотношение бактериальных классов в зависимости от загрязнения керосином и времени. На протяжении всего эксперимента доминируют классы Gammaproteobacteria (Proteobacteria), Acidobacteriae (Acidobacteriota), Alphaproteobacteria (Proteobacteria), Actinomycetia (Actinobacteriota) и Thermoleophilia (Actinobacteriota). В первой временной точке образцы имеют пренебрежимо малые различия.

Как меняется состав сообществ со временем?
В k0 заметно увеличивается соотношение Acidobacteriae к дургим бактериальным классам.
В k5 увеличивается доля Gammaproteobacteria (Proteobacteria), в которой есть виды, умеющие разлагать углеводороды, в том числе компоненты керосина. Например, представители родов Pseudomonas и Acinetobacter часто встречаются в местах загрязнения нефтью и участвуют в её разложении. Поэтому увеличение доли этих бактерий в k5 вполне ожидаемо. Я так понимаю, между k5_t3 и k5_t4 количество керосина в почве начинает существенно уменьшаться, что уменьшает долю(за ненадобностью такого количества) Gammaproteobacteria (Proteobacteria) и увеличвается доля Acidobacteriae (Acidobacteriota), которые предпочитают стабильные, кислые, богатые органикой почвы.
Аналогично просиходит в k25, только намного агрессивнее. Доля Gammaproteobacteria (Proteobacteria) стремительно увеличивается. Я бы предположил, что когда-то позже, после расщепления продуктов керосина, снова начала бы расти доля Acidobacteriae (Acidobacteriota), тк первые подготовили бы для них более пригодную среду обитания.
Очевидно, наблюдается закономерность: бактерии, устойчивые к керосину, увеличивают свою долю, тогда как чувствительные классы снижаются, пока состояние почвы не возвращается к более естественному.



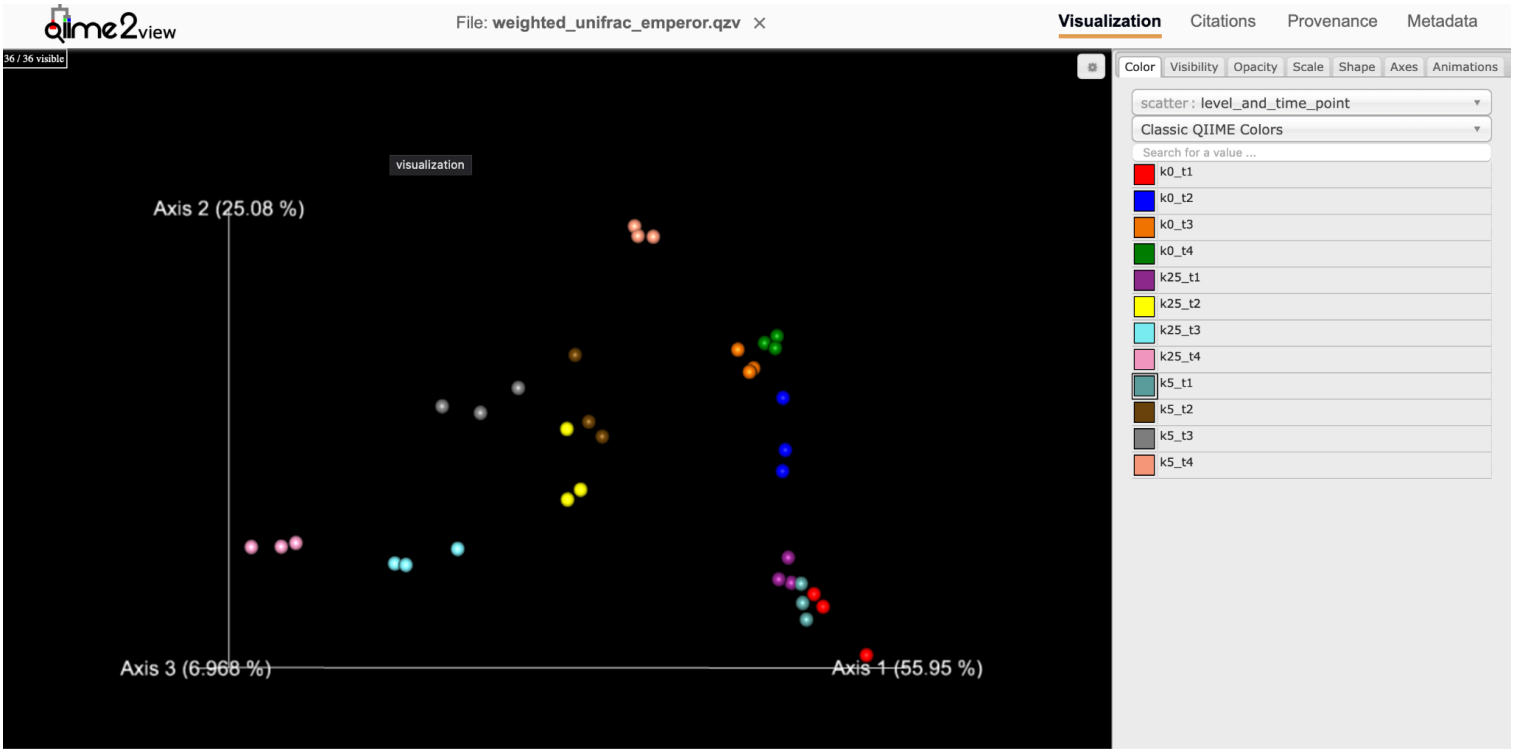
Частый род в загрязнённых образцах: Gammaproteobacteria, что ожидаемо при присутствии керосина.

qiime feature-table summarize \
  --i-table qza/soil_ASV_table.qza \
  --o-visualization soil_ASV_table_summary.qzv

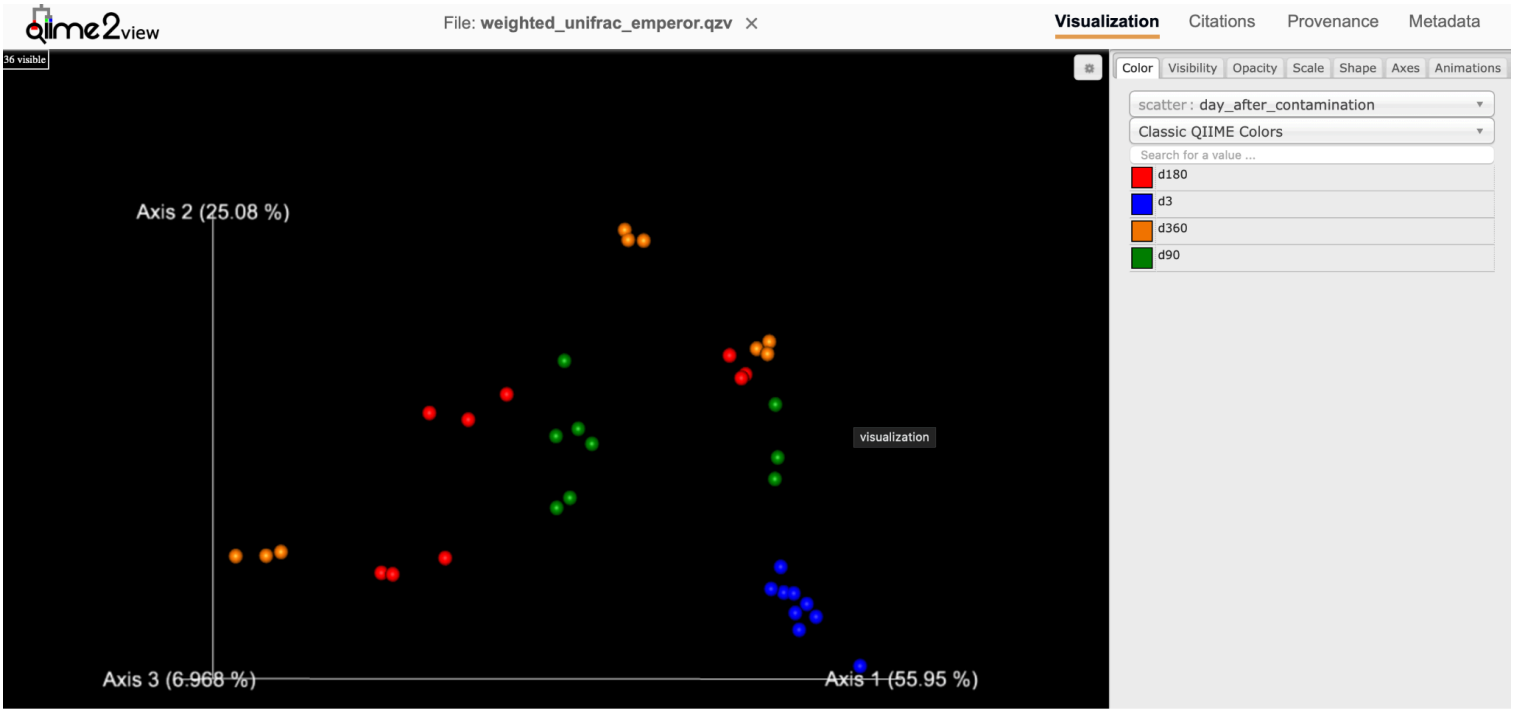
**qiime diversity core-metrics-phylogenetic **
** --i-phylogeny qza/soil_rooted_tree.qza **

```
--i-table qza/soil_ASV_table.qza \
--p-sampling-depth 4809 \
--m-metadata-file /projects/mbf_rsmu_bioinformatics_masters/metagenomes/soil/soil_metadata_full.tsv \
--output-dir soil_core_metrics_results
```

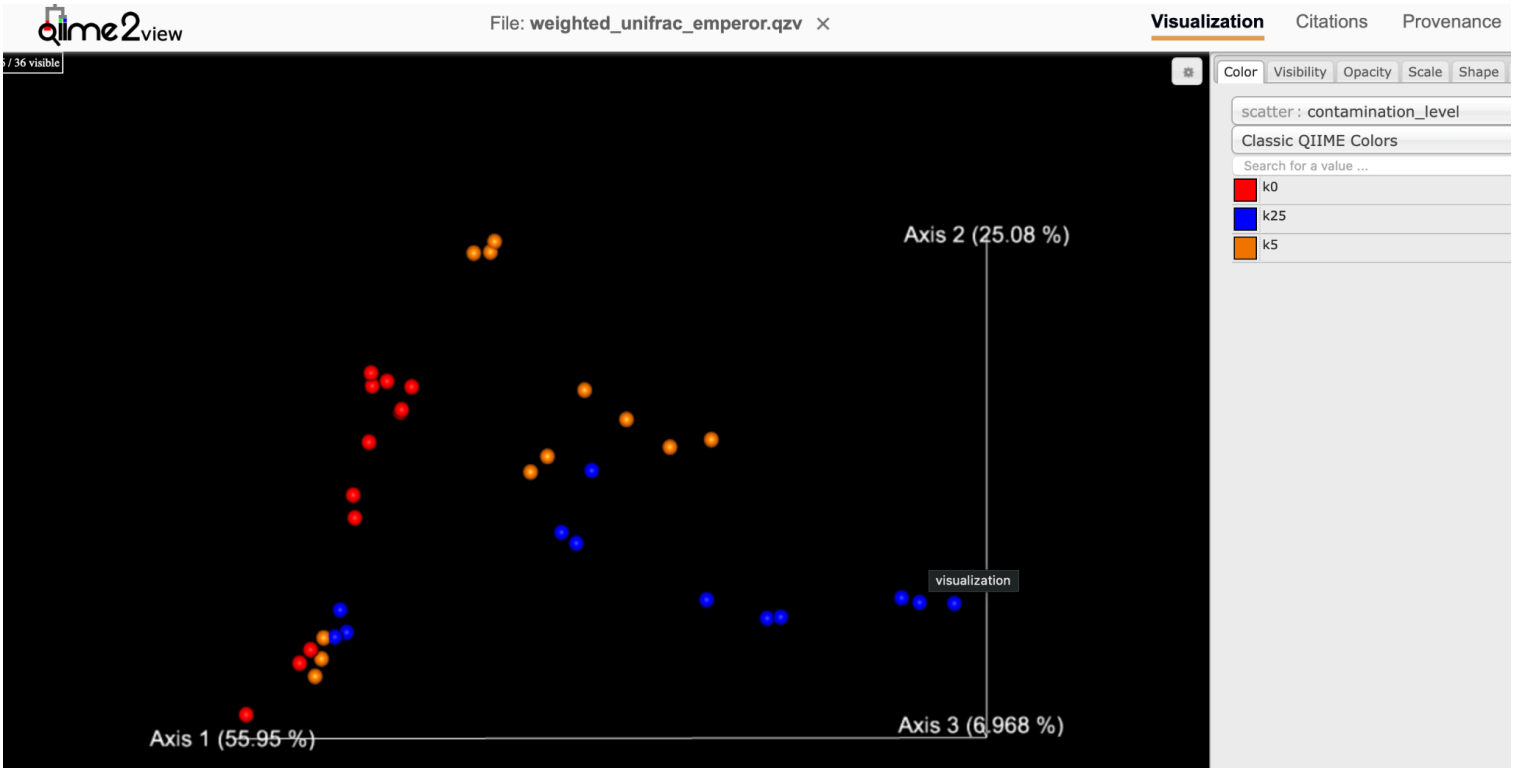
(**Задача 6**)



Опишите наблюдаемые результаты. Похожи ли образцы в первой временной точке?
В первой временной точке образцы образуют группу точек справа внизу(2-ая картинка - синим цветом), что указывает на сходство сообществ.



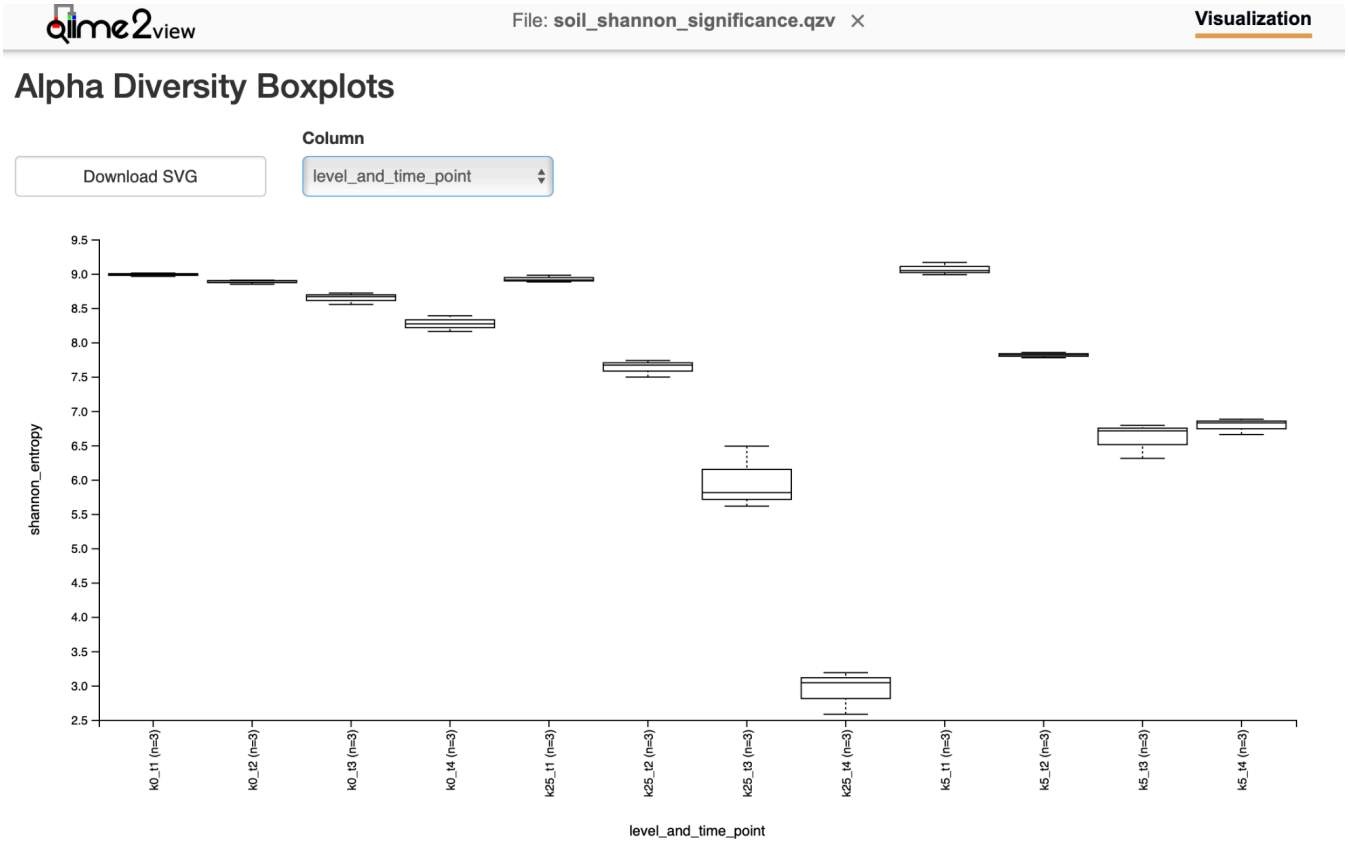
Как меняется их положение со временем (учтите, что самые сильные изменения показаны вдоль первой оси графика Axis1, более слабые вдоль второй Axis2, и ещё более слабые вдоль Axis3)?



k25 заметно меняет состав со временем, тк сильное смещение по Axis1
Моя гипотеза, насчет восстанавления бактериального сообщества в образце k5 с t3 на t4 подтвердилась, тк k5_t4 ближе к начальной позиции, чем k5_t3(нагляднее на 1-ой картинке).

Есть ли признаки восстановления сообщества после слабого или сильного загрязнения со временем?
Да, после слабого загрязнения есть признаки восстановления. Но в k25 бактериальное сообщество ушло еще дальше по Axis1 от начальной позиции, что говорит о том, что образец далек от восстановления.

(**Задача 7**)
Опишите как меняется разнообразие бактериального сообщества со временем при разном уровне загрязнения образцов.



Чем выше индекс Шеннона, тем более разнообразное и сбалансированное сообщество бактерий.

В наших образцах наблюдаются те же тенденции, что и на предыдущих графиках: в контрольных образцах k0 разнообразие слабо падает с течением времени, сообщество остаётся относительно стабильным. При сильном загрязнении k25 разнообразие резко снижается с течением времени, на точках t3 и t4 барплоты становятся очень широкими, что указывает на сильную вариабельность между повторными образцами и деградацию сообщества. При слабом загрязнении k5 разнообразие снижается до t3, где барплот широкий, а затем слегка растет к t4, при этом барплот становится уже, что говорит о частичном восстановлении бактериального сообщества после воздействия керосина.