Практикум 14

Для выполнения данного практикума был задан код доступа SRR4240358 (проект по секвенированию бактерии *Buchnera aphidicola* str. Tuc7). Вся дальнейшая работа осуществлялась в директории:
/mnt/scratch/NGS/bratzveron/pr14

1. Подготовка чтений

Для начала был скачан архив с чтениями по данному коду доступа:

```
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR424/008/SRR4240358/SRR4240358.fastq.gz
```

Далее он был разархивирован:

```
gunzip SRR4240358.fastq.gz
```

Сводка про исходные данные:

Accession: **SRR4240358**
Тип чтений: **одиночные (single-end)**
Платформа: **Illumina**
Длина чтений: **39 bp**
Исходный файл: SRR4240358.fastq

Затем была выполнена подготовка чтений, на первом этапе нужно было удалить возможные остатки адаптеров. Для этого сперва был создан файл adapters.fa, где объединены все адаптеры из директории /mnt/scratch/NGS/adapters, следующей командой:

```
cat /mnt/scratch/NGS/adapters/*.fa > adapters.fa
```

На втором этапе проводим удаление адаптеров с помощью Trimmomatic:

```
trimmomatic SE \
SRR4240358.fastq \
SRR4240358.trimmed.fastq \
ILLUMINACLIP:adapters.fa:2:7:7
```

Параметры:
SE — одиночные чтения
ILLUMINACLIP:adapters.fa:2:7:7 — поиск и удаление остатков адаптеров:
2 — максимум несовпадений
7 — порог палиндромного совпадения
7 — порог простого совпадения

Не сработало. Посмотрела, что в окружении trimmomatic как команда не прописан в PATH, поэтому нужно запускать через java -jar:

```
java -jar /usr/share/java/trimmomatic.jar SE \
SRR4240358.fastq \
SRR4240358.trimmed.fastq \
ILLUMINACLIP:adapters.fa:2:7:7
```

Исходных чтений: **10 543 839**
Сохранённых чтений: **10 368 884 (98.34%)**
Удалённых чтений: **174 955 (1.66%)**

Процент чтений, содержащих остатки адаптерных последовательностей, составил: **1.66%**
*(174 955 / 10 543 839) × 100% = 1.66%

После была проведена дополнительная очистка чтений. С правых (3') концов удалялись нуклеотиды с качеством ниже 20 и сохранялись только чтения длиной не менее 32 нуклеотидов:

```
java -jar /usr/share/java/trimmomatic.jar SE \
 SRR4240358.trimmed.fastq \
 SRR4240358.filtered.fastq \
 TRAILING:20 \
 MINLEN:32
```

Параметры:
TRAILING:20 — удаление нуклеотидов с конца чтения с качеством < 20
MINLEN:32 — отбрасывание чтений короче 32 нт после обрезки

Входных чтений: **10 368 884**
Сохранённых чтений: **8 016 437 (77.31%)**
Удалённых чтений: **2 352 447 (22.69%)**

Размеры файлов определялись командой:

```
ls -lh SRR4240358.trimmed.fastq SRR4240358.filtered.fastq
```

Размер файла **до очистки** (SRR4240358.trimmed.fastq): **1.1 GB**
Размер файла **после очистки** (SRR4240358.filtered.fastq): **826 MB**

Вывод: фильтрация по качеству и длине привела к удалению около 22.7% чтений, что обусловлено низким качеством нуклеотидов на 3'-концах и сокращением длины некоторых чтений ниже порогового значения. В результате было получено более 8 млн высококачественных одиночных чтений длиной не менее 32 нуклеотидов.

2. Подготовка k-меров

Исходные данные:
Файл чтений: **SRR4240358.filtered.fastq**
Тип чтений: **короткие, одиночные (short, single-end)**
Длина чтений после очистки: ≥ **32 нт**
Выбранная длина k-мера (**hash_length**): **31** (максимально возможная)

Для изучения доступных параметров и формата запуска была выполнена команда:

```
velveth -help
```

Из справки программы было установлено:
первый аргумент — имя выходной директории
второй аргумент — длина k-мера (hash_length)
далее указывается тип данных (-short) и формат входного файла (-fastq)
программа предназначена для подготовки k-меров перед сборкой (velvetg)

На основе файла очищенных чтений были подготовлены k-меры длины **31:**

```
velveth velvet_31 31 -short -fastq SRR4240358.filtered.fastq
```

velvet_31 — директория, в которую будут записаны выходные файлы
31 — длина k-мера (hash_length)

-short — одиночные короткие чтения
-fastq — формат входного файла
SRR4240358.filtered.fastq — файл с очищенными чтениями

В результате работы программы была создана директория velvet_31
В ней были сформированы служебные файлы Velvet (Sequences, Roadmaps, Log)

3. Сборка на основе k-меров

Для сборки генома на основе k-меров используется программа **velvetg:**

```
velvetg velvet_31
```

velvet_31 — директория с k-мерами, подготовленными velveth
По умолчанию velvetg использует настройки:
одиночные чтения (-short) были заданы на этапе velveth
автоматическая оценка покрытия и длины контигов
выходной файл contigs.fa создаётся в той же директории

Из файла получаем: >NODE_1_length_11615_cov_28.861988 (первая строка)
NODE_1 — идентификатор контига
length_11615 — длина контига: 11 615 нуклеотидов
cov_28.861988 — среднее покрытие контига: 28.86…

Извлечение длин и покрытия:

```
grep '>' velvet_31/contigs.fa | awk -F'[ _]' '{print $4, $6}'
> contigs_lengths_cov.txt
```

$4 — число после length_
$6 — число после cov_

В файле:
11615 28.861988
9010 29.288679
5928 28.058199
3473 29.700546
10513 31.405594 и т.д.

Ищем три самых длинных с помощью скрипта **analyze_contigs.py**
Запуск в командной строке: python3 analyze_contigs.py velvet_31/contigs.fa
Результат сохранен в report.txt
Сводка по сборке:
N50: 8600 bp
медианное покрытие: 26.95×

три самых длинных контига:
>NODE_56_length_19821_cov_29.475859 | длина: 19821 | покрытие: 29.48
>NODE_34_length_18714_cov_29.922678 | длина: 18714 | покрытие: 29.92
>NODE_40_length_16436_cov_30.793623 | длина: 16436 | покрытие: 30.79

аномальные контиги (покрытие >5× медианы или <0.2× медианы):
>NODE_41_length_949_cov_266.472076 | длина: 949 | покрытие: 266.47
>NODE_49_length_622_cov_281.516083 | длина: 622 | покрытие: 281.52
>NODE_48_length_429_cov_248.417252 | длина: 429 | покрытие: 248.42
>NODE_116_length_373_cov_294.319031 | длина: 373 | покрытие: 294.32
>NODE_89_length_285_cov_264.084198 | длина: 285 | покрытие: 264.08
>NODE_105_length_216_cov_290.912048 | длина: 216 | покрытие: 290.91
>NODE_103_length_177_cov_295.135590 | длина: 177 | покрытие: 295.14
>NODE_243_length_115_cov_5.313044 | длина: 115 | покрытие: 5.31
>NODE_72_length_111_cov_289.324310 | длина: 111 | покрытие: 289.32
>NODE_212_length_109_cov_4.642202 | длина: 109 | покрытие: 4.64
>NODE_129_length_106_cov_332.877350 | длина: 106 | покрытие: 332.88
>NODE_307_length_93_cov_5.150537 | длина: 93 | покрытие: 5.15
>NODE_283_length_69_cov_3.956522 | длина: 69 | покрытие: 3.96
>NODE_110_length_66_cov_142.651520 | длина: 66 | покрытие: 142.65
>NODE_151_length_64_cov_175.265625 | длина: 64 | покрытие: 175.27
>NODE_18_length_60_cov_412.100006 | длина: 60 | покрытие: 412.10
>NODE_97_length_53_cov_405.245270 | длина: 53 | покрытие: 405.25
>NODE_289_length_37_cov_4.675676 | длина: 37 | покрытие: 4.68
>NODE_143_length_31_cov_3.064516 | длина: 31 | покрытие: 3.06
>NODE_313_length_31_cov_3.806452 | длина: 31 | покрытие: 3.81
>NODE_324_length_31_cov_4.000000 | длина: 31 | покрытие: 4.00
>NODE_330_length_31_cov_5.387097 | длина: 31 | покрытие: 5.39
>NODE_333_length_31_cov_1.709677 | длина: 31 | покрытие: 1.71

4. Анализ

Buchnera aphidicola (Aphis glycines) strain BAg, complete genome - полный геном
хромосома GenBank: **CP009253.1** — это полный хромосомный геном Buchnera aphidicola из сои (Aphis glycines)
*в ENA и RefSeq он тоже доступен под тем же accession-номером

Спойлер: если я получу хотя бы 2.5 за это задание я буду на седьмом небе от счастья)))))))))

В BLAST вставляю fasta последовательности и CP009253

Контиг NODE_56_length_19821_cov_29.475859
Стандартный запуск BLASTN (с megablast) не выявил значимых совпадений. Это связано с тем, что контиг характеризуется крайне высокой AT-насыщенностью и низкой сложностью последовательности, в результате чего большая часть контига была замаскирована фильтром низкой сложности.
Пробуем убрать Low complexity regions - значимых локальных выравниваний обнаружено не было. Вероятно, данный контиг соответствует межгенному или штамм-специфичному участку генома, либо деградированному региону, отсутствующему или сильно дивергировавшему у выбранного банковского штамма.
Делаю тогда в **blastn.** Word size 7, Low complexity off, Mask for lookup table only off

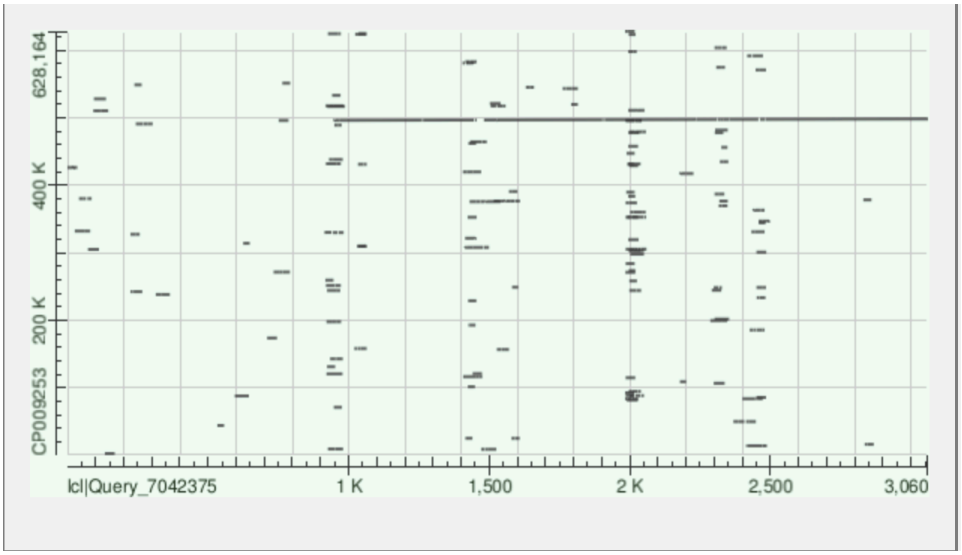

Рис 1

Координаты на хромосоме: 496111 – 498187
Длина выравнивания: ~2 100 bp
Процент идентичности: ~74%
Число гэпов: ~60 / 2100 (~3%)
E-value: 0.0
Strand: Plus / Plus
Карта локального сходства показывает отсутствие протяжённой диагонали и наличие большого числа коротких совпадений, распределённых по всему геному Buchnera aphidicola. Это указывает на то, что контиг NODE_56 содержит многочисленные низкосложные AT-богатые участки, которые находят гомологичные фрагменты в разных частях генома. При этом наблюдается один основной высокозначимый локальный участок сходства (~2.1 кб, E-value = 0.0), соответствующий реальному гомологичному региону, тогда как остальные совпадения являются вторичными и обусловлены повторяющимися последовательностями.

Контиг NODE_34_length_18714_cov_29.922678
Координаты на хромосоме: 8599 – 11103
Длина выравнивания: 2 525 bp
Процент идентичности: 78%
Число гэпов: 50 / 2525 (~1%)
E-value: 0.0
Strand: Plus / Plus

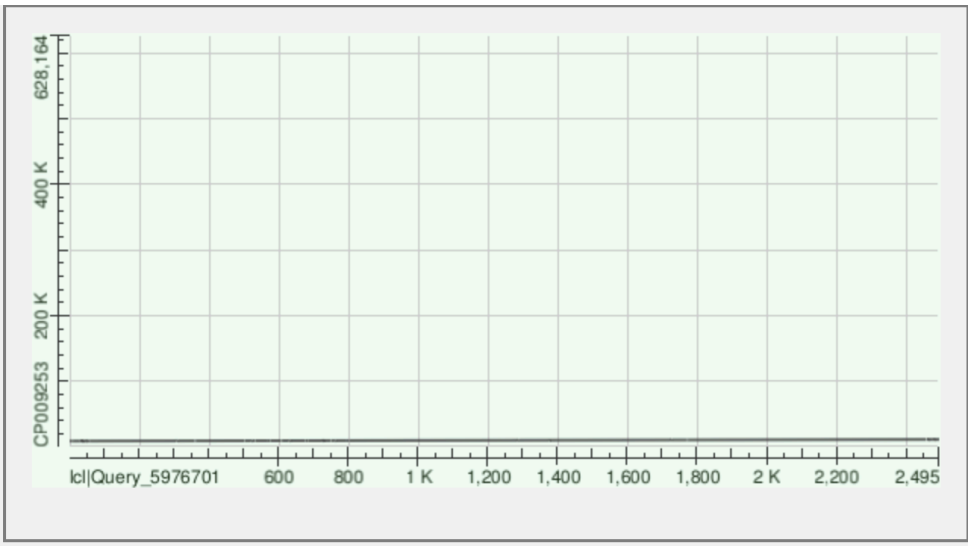

Рис.2

Карта демонстрирует одно основное непрерывное локальное выравнивание. Разрывов выравнивания не наблюдается, что указывает на то, что контиг ложится на геном единым блоком, без перестроек и инверсий. Направление выравнивания совпадает (plus/plus), что свидетельствует об одинаковой ориентации последовательностей.Отсутствие дополнительных диагоналей или разрозненных участков сходства указывает на то, что анализируемый контиг не является повторяющейся последовательностью и не имеет значимых гомологий в других частях хромосомы.

Контиг NODE_40_length_16436_cov_30.793623
Координаты на хромосоме: 471193 – 474242
Длина выравнивания: 3 091 bp
Процент идентичности: 78%
Число гэпов: 74 / 3091 (~2%)
E-value: 0.0
Strand: Plus / Minus

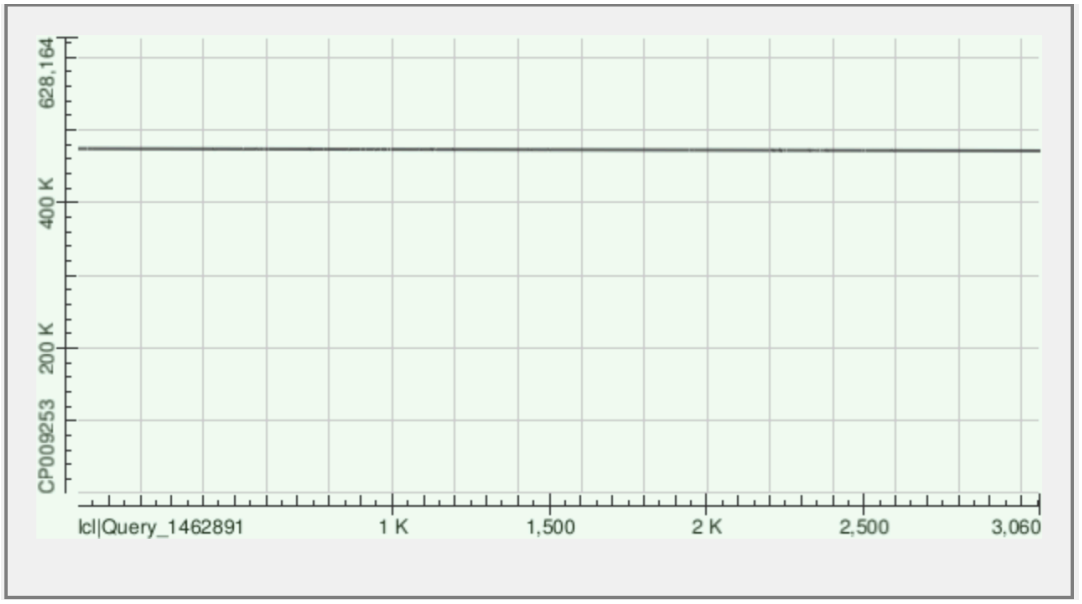

Рис.3

Карта локального сходства демонстрирует одно основное протяжённое локальное выравнивание. Контиг выравнивается с геномом в обратной ориентации (plus/minus), что указывает на инвертированное расположение данного фрагмента относительно выбранного эталонного генома.

Разрывов основного выравнивания не наблюдается, что свидетельствует о том, что контиг ложится на геном единым блоком, без фрагментации. Отсутствие дополнительных диагоналей или разрозненных участков

сходства указывает на то, что анализируемый контиг не является повторяющейся последовательностью и соответствует уникальному участку хромосомы. Наблюдаемые однонуклеотидные различия и гэпы отражают штаммовые различия между исследуемым объектом и банковской последовательностью.