Отчет по практикуму 14
Сборка de novo
Сорокин Илья Андреевич

Этап 1. Подготовка чтений программой Trimmomatic.

Код доступа проекта по секвенированию бактерии *Buchnera aphidicola* str. Tuc7 - **SRR4240380**.

Сначала нужно удалить остатки адаптеров. Для этого использовалась следующая команда:

```
TrimmomaticSE SRR4240380.fastq.gz trimmedreads
ILLUMINACLIP:adapters.fasta:2:7:7
```

В файле **adapters.fasta** находятся последовательности адаптеров, используемые Illumina. Для создания этого файла была использована следующая команда:

```
cat *.fa > adapters.fasta
```

***fa.** - фаста-файлы с последовательностями адаптеров.

По предпоследней строчки выдачи программы TrimmomaticSE можно понять, что 1.88% чтений представляют собой остатки адаптеров:

```
Input reads: 5217318 Surviving: 5119144 (98.12%) Dropped: 98174
(1.88%)
```

Далее нужно удалить с правых концов чтений нуклеотиды с качеством ниже 20 и оставить последовательности с длиной не менее 32 нуклеотидов. Для этого была применена следующая команда:

```
TrimmomaticSE trimmedreads trimmedqualityreads TRAILING:20 MINLEN:32
```

Пояснения:

`TRAILING:20` - удаление нуклеотидов с качеством <20 с конца чтений;

`MINLEN:40` - оставить только чтения длиной не менее 40 нуклеотидов.

По последней строчки выдачи этой программы можно понять, что изначально было **5119144** чтений, а **253785** чтений (4.96%) было удалено:

```
Input reads: 5119144 Surviving: 4865359 (95.04%) Dropped: 253785 (4.96%)
```

Сравним размеры файлов до и после триммирования.

До триммирования размер файла `trimmedreads` составляет **516M**, после триммирования размер файла `trimmedqualityreads` составляет **490M**.

Этап 2. Сборка генома de novo.

Для сборки генома с помощью программы была создана директория velveth, куда будут сохраняться все последующие файлы.

Сначала была использована команда velveth для создания k-меров из чтений:

```
velveth velvet 31 -fastq -short trimmedqualityreads
```

Пояснения:

`velvet` - название выходной директории с файлами выдачи;
`31` - длина k-мера;
`-fastq` - формат входного файла.

Далее с помощью программы `velvetg` запустим сборку генома по папке `velvet`:

```
velvetg velvet
```

Параметр N50 - **12042**. Максимальный контиг имеет длину **25915**.

Чтобы найти длины 3 самых длинных контигов, был составлен конвейер в BASH для файла stats.txt:

```
grep '^>' contigs.fa | tr '_' '\t' | sort -k4 -nr | tr '\t' '_' | head -3
```

Выдача конвейера:

```
>NODE_3_length_25915_cov_27.418676
>NODE_20_length_23850_cov_24.763815
>NODE_23_length_23807_cov_25.725922
```

Таким образом, контиги с номерами **2**, **20**, **23** имеют наибольшую длину.

Медиана покрытий составляет **16**.

Есть контиги, покрытия которых отличаются от медианы как минимум в 5 раз:

>NODE_11_length_2106_cov_126.008545
>NODE_56_length_934_cov_130.479660
>NODE_110_length_80_cov_2.700000

Можно заметить, что эти контиги имеют намного меньшую длину, чем контиги с более низким покрытием.

Этап 3. Анализ

Посмотрим, как 3 самых длинных контига ложатся на хромосому *Buchnera aphidicola str. JF99* (GenBank ID - **CP002302.1**). Самые длиные контиги:

>NODE_3_length_25915_cov_27.418676
>NODE_20_length_23850_cov_24.763815
>NODE_23_length_23807_cov_25.725922

Запустим megablast для каждого контига относительно хромосомы *Buchnera aphidicola str. JF99.*

1) Контиг 3

Контиг ложится 2 участками на начало [1-11373] и конец [627163-641716] хромосомы. Скорее всего, такая картина могла возникнуть, когда последовательности имеют разные точки начала прочтений. Внизу показан файл выдачи megablast Hit-table (text). На рис. 1-2 изображено расположение этого контига на хромосоме и dot-plot выравнивания соответственно.

```
# blastn
# Iteration: 0
# Query: CP002302.1 Buchnera aphidicola str. JF99 (Acyrthosiphon pisum), complete genome
# RID: KUTKM0JV114
# Database: n/a
# Fields: query acc.ver, subject acc.ver, % identity, alignment length, mismatches, gap opens, q. start, q. end, s. start, s. end, evalue, bit score
# 2 hits found
CP002302.1	Query_2447411	99.711	14557	38	3	627163	641716	1	14556	0.0	26646
CP002302.1	Query_2447411	99.798	11374	22	1	1	11373	14572	25945	0.0	20875
```

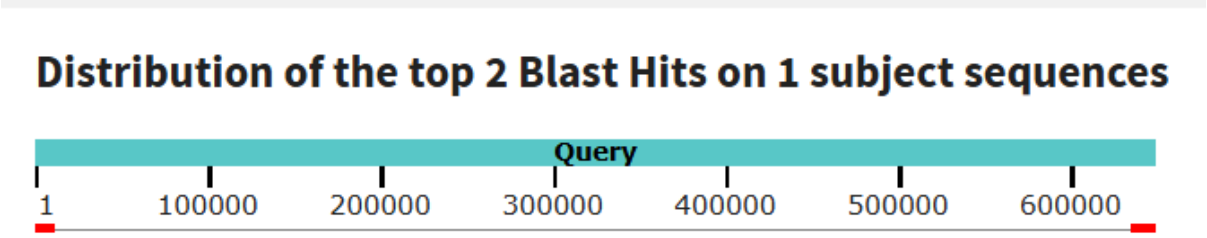

Рис. 1. Расположение контига 3 на хромосоме CP002302.1.

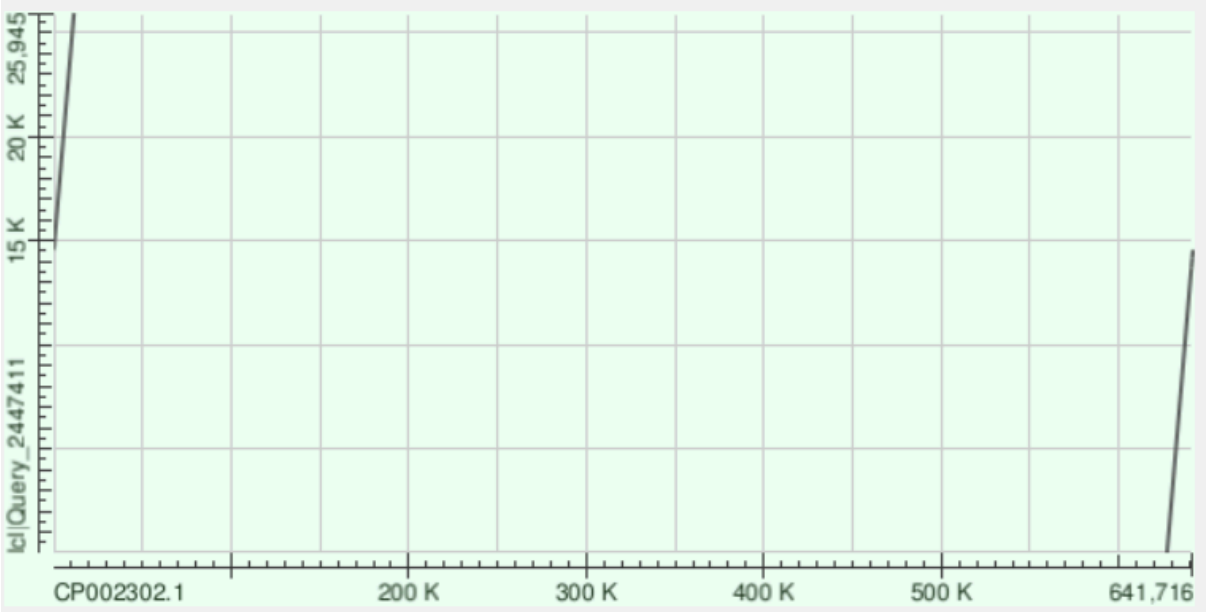

Рис. 2. Dot-plot выравнивания контига 3 с хромосомой CP002302.1.

2) Контиг 20

Координаты участка хромосомы: [232481-256360]. Число однонуклеотидных замен - 42, число гэпов - 2. Практически весь контиг выровнялся на геном. На рис. 3-4 изображено расположение этого контига на хромосоме и dot-plot выравнивания соответственно. Данный контиг соответствует минус-цепи референса: начало контига выравнивается на конец комплементарного участка генома.

```
# blastn
# Iteration: 0
# Query: CP002302.1 Buchnera aphidicola str. JF99 (Acyrthosiphon pisum), complete genome
# RID: KUU6WM0A114
# Database: n/a
# Fields: query acc.ver, subject acc.ver, % identity, alignment length, mismatches, gap opens, q. start, q. end, s. start, s. end, evalue, bit score
# 1 hits found
CP002302.1	Query_6040083	99.816	23880	42	2	232483	256360	23880	1	0.0	43853
```

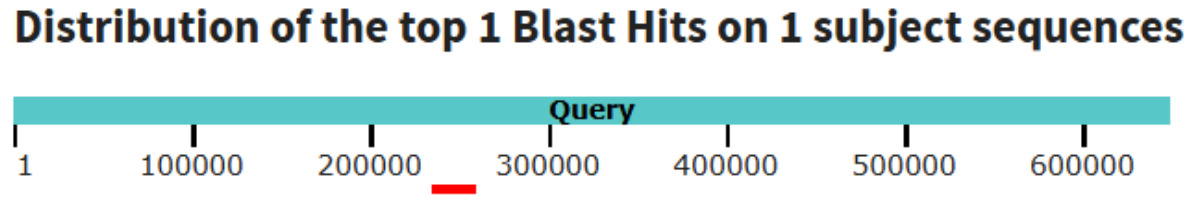

Рис. 3. Расположение контига 20 на хромосоме CP002302.1.

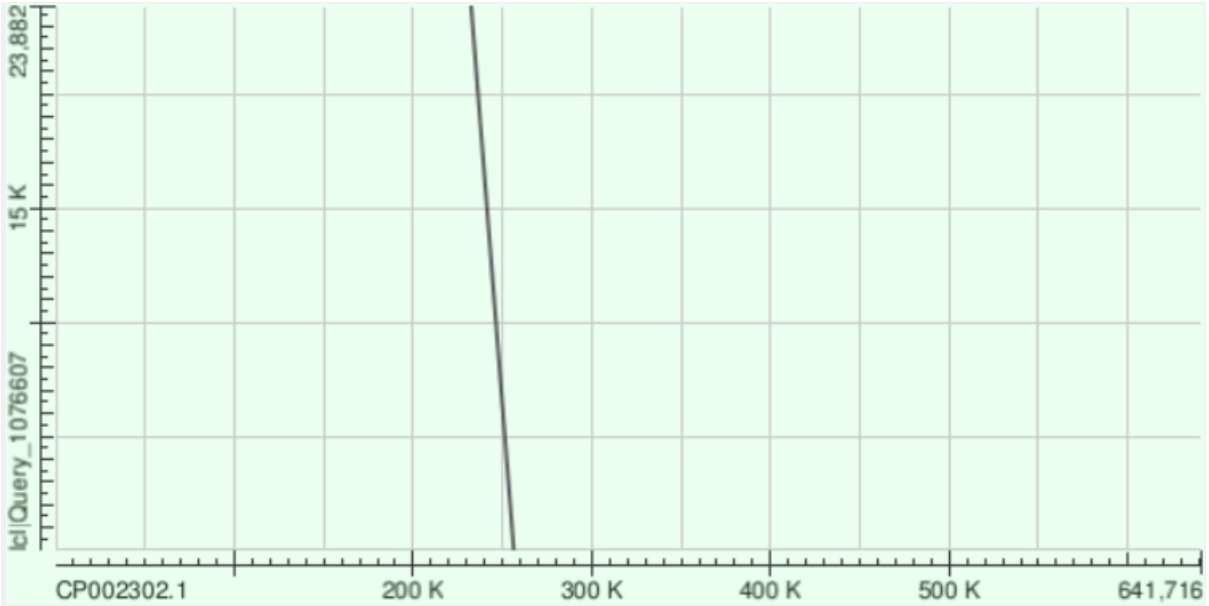

Рис. 4. Dot-plot выравнивания контига 20 с хромосомой CP002302.1.

3) Контиг 23

Координаты участка хромосомы: [584697-608534]. Число однонуклеотидных замен - 60, число гэпов - 1. Практически весь контиг выровнялся на геном. На рис. 5-6 изображено расположение этого контига на хромосоме и dot-plot выравнивания соответственно.

```
# blastn
# Iteration: 0
# Query: CP002302.1 Buchnera aphidicola str. JF99 (Acyrthosiphon pisum), complete genome
# RID: KNY9DV97114
# Database: n/a
# Fields: query acc.ver, subject acc.ver, % identity, alignment length, mismatches, gap opens, q. start, q. end, s. start, s. end, evalue, bit score
# 1 hits found
CP002302.1	Query_4757379	99.744	23838	60	1	584697	608534	1	23837	0.0	43681
```

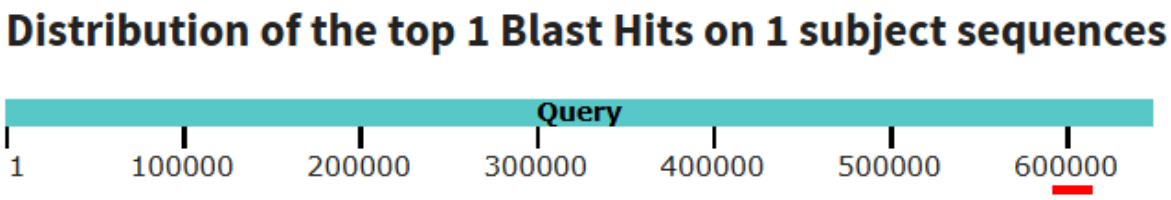

Рис. 5. Расположение контига 23 на хромосоме CP002302.1.

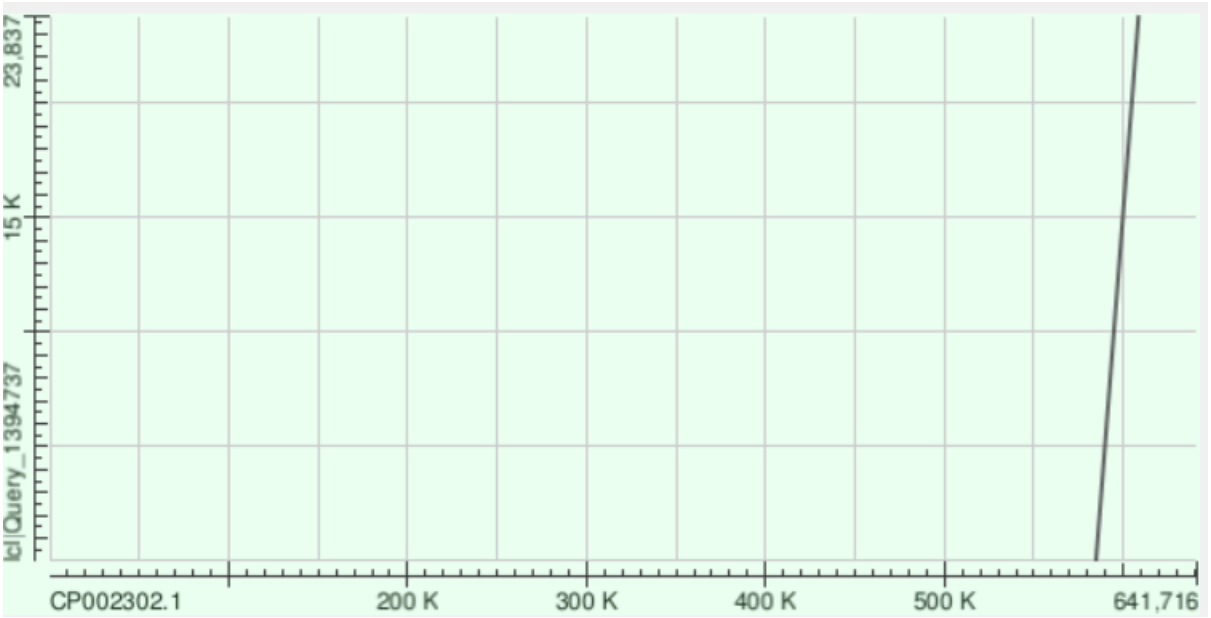

Рис. 6. Dot-plot выравнивания контига 23 с хромосомой CP002302.1.