Практикум 15 Сборка генома de novo

1. Получение сборки

Для начала архив с чтениями (код доступа **SRR4240379**), полученными по секвенированию бактерии *Buchnera aphidicola* str. Tuc7 был скачан в рабочую директорию /mnt/scratch/NGS/olga.kiseleva/pr14 с помощью команды :

```
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR424/009/SRR4240379/SRR4240379.fastq.gz
```

2. Подготовка программой Trimmomatic

a. Удаление адаптеров

Далее были файлы с последовательностями адаптеров из общей папки скопированы в рабочую директорию:

```
cp ../../adapters/*.fa .
```

После этого последовательности были объединены в один файл:

```
cat *.fa > all_adapters.fasta
```

b. Удаление возможных остатков адаптеров

Для триммирования была использована команда TrimmomaticSE,так как прочтения здесь одноконцевые:

```
TrimmomaticSE -phred33 SRR4240379.fastq.gz SRR4240379_trimmed_1.fastq.gz ILLUMINACLIP:all_adapters.fasta:2:7:7
```

Кроме того, была получена информация о количестве чтений, поданных на вход и количестве оставшихся после триммирования:

- Подано на вход: 7400155
- Сохранено с изменениями: 7269852 (98.24%)
- Исключено: 130303 (1.76%)

c. Фильтрация чтений по качеству и длине

Теперь удалим чтения с качеством <20 и длиной <32 тоже с помощью команды TrimmomaticSE:

```
TrimmomaticSE -phred33 SRR4240379_trimmed_1.fastq SRR4240379_trimmed_2.fastq  TRAILING:20 MINLEN:32
```

После триммирования с заданными параметрами получилось чтений:

- Подано на вход: 7269852
- Сохранено с изменениями: 6974267 (95.93%)
- Исключено: 295585 (4.07%)

3. Использование программы velveth

Подготовка k-меров (подстрок длины k=31)

```
velveth velveth 31 -fastq -short SRR4240379_trimmed_2.fastq
```

velveth — создаёт таблицу хэшей k-меров для полученных чтений SRR4240379_trimmed_2.fastq в директории velveth

-fastq — формат входных файлов FASTQ

-short — короткие прочтения (типично для Illumina)

Создана директория с файлами: Log, Roadmaps, Sequences

4. Использование программы velvetg

Программа velvetg выполняет фактическую сборку контигов на основе хэш-таблицы, созданной командой **velveth**

```
velvetg velveth
```

В результате построен граф де Брёйна из k-меров, созданы файлы со сборкой контигов и статистикой, её описывающей

N50 = 25646 (Final graph has 440 nodes, max 49912,total 664650)

a. Поиск самых длинных контигов

Для поиска трех самых длинных контигов и их покрытия был использован файл contigs.fa с аннотацией контигов:

```
grep '^>' contigs.fa | tr '_' '\t' | sort -k 4,4 -r -n | head -n3
```

Информация обобщена в таблице:

Номер контига	Длина	Покрытие
6	49912	35.907238
9	49262	34.772179
5	33085	36.259029

Также найдены три контига с аномально большим покрытием (возможно, являющиеся короткими повторами):

Номер контига	Длина	Покрытие
60	296	181.739868
39	609	177.170776
30	2083	172.516083

5. Сравнение самых длинных контигов с хромосомой Buchnera aphidicola с помощью megablast

a. Анализ 6 контига

В качестве объекта сравнения была выбрана хромосома (RefSeq ID **NZ_CP033006.1**) бактерии Buchnera aphidicola (Macrosiphum euphorbiae) штамм Meu

Контиг выровнялся на 2 крупных участка хромосомы (131k-144k и 144k-181k) и на 1 короткий (641421-641520).
Высокий процент идентичности - 86.40% и прямой ход выравнивания Dot Plot свидетельствуют о комплементарности последовательностей. Поэтому можем говорить о хорошей сборке, ведь выравнивался контиг с последовательностью хромосомы одного вида бактерии.

Результат Dot Plot для 6 контига приведён ниже:

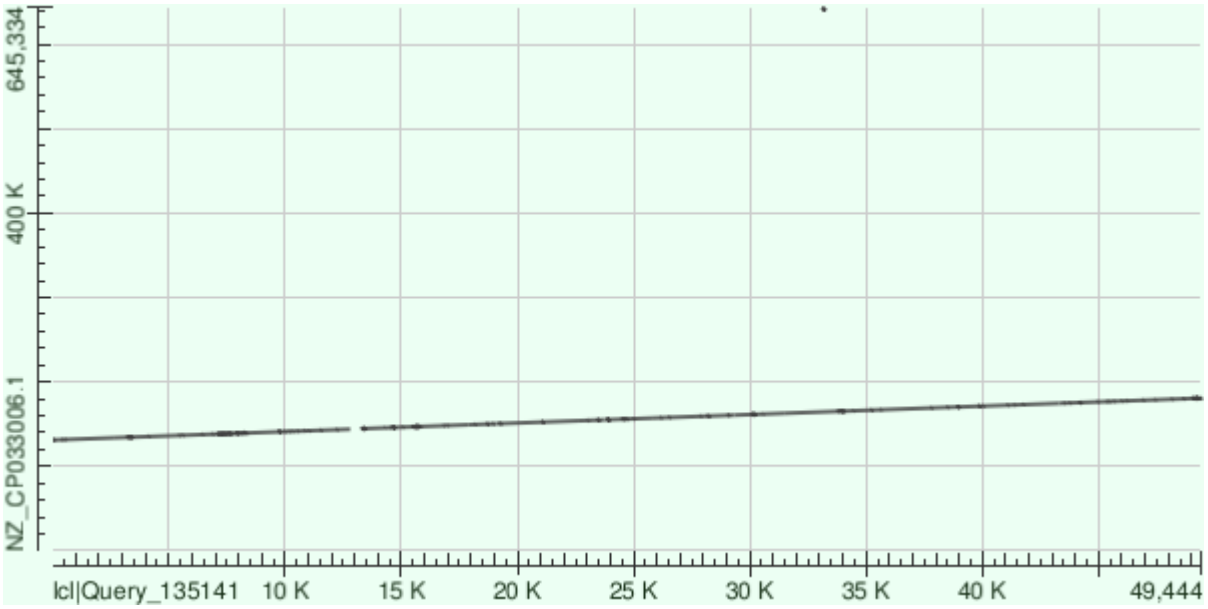

Рис. 1. Dot Plot для 6 контига

```
# blastn
# Iteration: 0
# Query: NODE_6_length_49912_cov_35.907238
# RID: KXB3FSHD114
# 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
# 3 hits found
NODE_6_length_49912_cov_35.907238       NZ_CP033006.1   86.399  36343   4617    255     13282   49444   144733  180929  0.0     39428
NODE_6_length_49912_cov_35.907238       NZ_CP033006.1   84.861  12854   1790    122     26      12811   131310  144075  0.0     12813
NODE_6_length_49912_cov_35.907238       NZ_CP033006.1   78.302  106     14      6       33131   33233   641520  641421  2.41e-08        60.2
```

Рис. 2. Текстовая выдача с результатами megablast для 6 контига

b. Анализ 9 контига

Аналогичная ситуация наблюдается с 9 контигом, который выровнялся на 5 участков хромосомы (они находятся в рамке 507k - 543k). Так же можем говорить и высокой степени комплементарности и хорошей сборке.

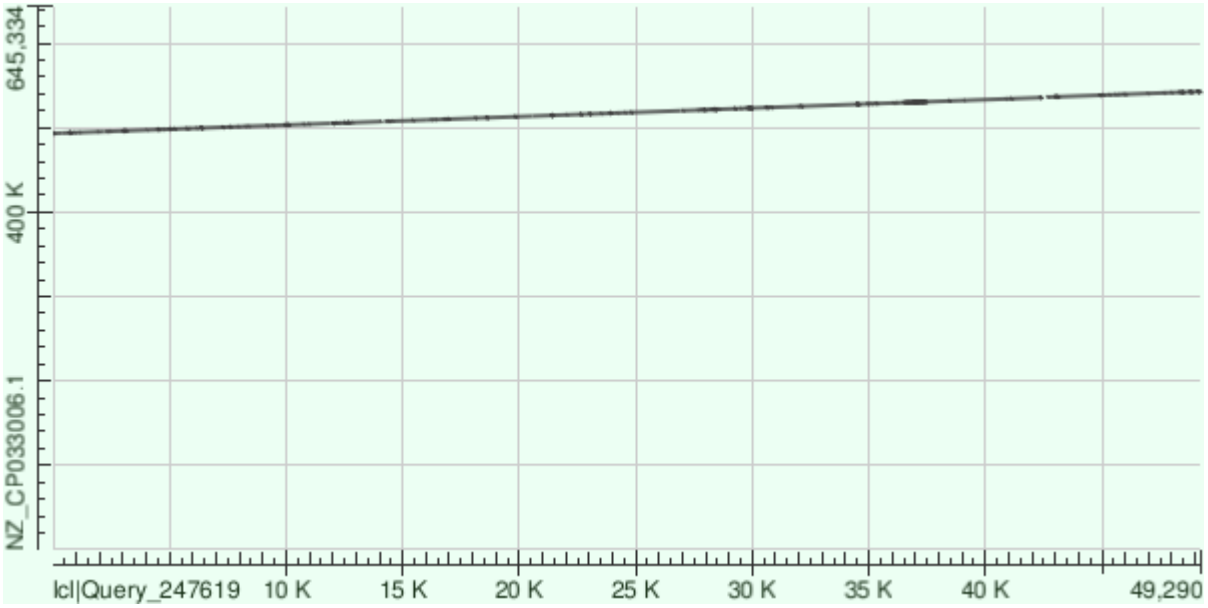

Рис. 3. Dot Plot для 9 контига

```
# Iteration: 0
# Query: NODE_9_length_49262_cov_34.772179
# RID: KXBAS9XV114
# 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
# 5 hits found
NODE_9_length_49262_cov_34.772179       NZ_CP033006.1   85.962  14325   1853    102     14202   28435   507362  521619  0.0     15167
NODE_9_length_49262_cov_34.772179       NZ_CP033006.1   85.503  14099   1826    176     28400   42398   521642  535622  0.0     14512
NODE_9_length_49262_cov_34.772179       NZ_CP033006.1   85.013  14166   1948    133     1       14080   493214  507290  0.0     14236
NODE_9_length_49262_cov_34.772179       NZ_CP033006.1   86.848  5748    695     44      42648   48354   536184  541911  0.0     6370
```

Рис. 4. Текстовая выдача с результатами megablast для 9 контига

c. Анализ 5 контига

В этот раз контиг выровнялся на 4 участка хромосомы в рамке 460k - 477k, но с разрывом в 2 тысячи нуклеотидов, указывающий на неконсервативный участок. Тем не менее можем наблюдать прямой ход выравнивания и хорошую сборку, как и с предыдущими контигами.

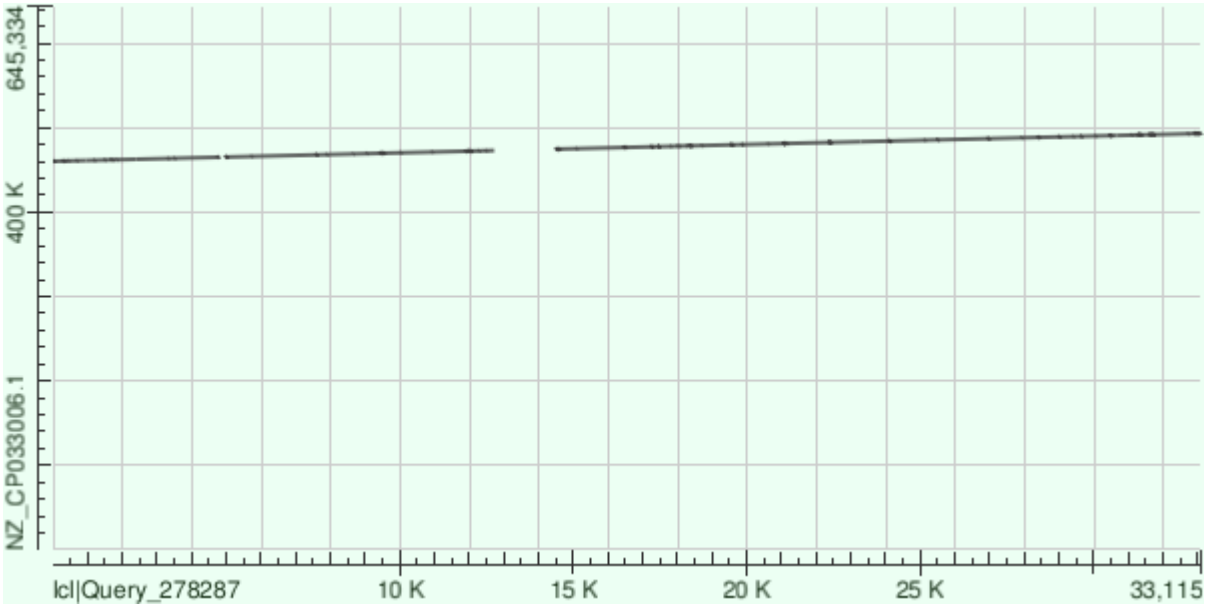

Рис. 5. Dot Plot для 5 контига

```
# blastn
# Iteration: 0
# Query: NODE_5_length_33085_cov_36.259029
# RID: KXBCRPS3114
# 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
# 4 hits found
NODE_5_length_33085_cov_36.259029       NZ_CP033006.1   85.909  15954   2038    151     17244   33115   477417  493242  0.0     16814
NODE_5_length_33085_cov_36.259029       NZ_CP033006.1   86.355  7761    1001    49      4980    12717   465266  472991  0.0     8412
NODE_5_length_33085_cov_36.259029       NZ_CP033006.1   86.086  4808    633     32      1       4787    460246  465038  0.0     5140
NODE_5_length_33085_cov_36.259029       NZ_CP033006.1   88.337  2778    303     19      14500   17269   474648  477412  0.0     3315
```

Рис. 6. Текстовая выдача с результатами megablast для 5 контига