Практикум 15
Сборка генома de novo

Мне достался проект под номером **SRR4240358**, поэтому для скачивания чтений я использовала следующий алгоритм:

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

1. Подготовка чтений программой trimmomatic

a. Удаление адаптеров

Были получены файлы с последовательностями адаптеров при помощи команды:

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

Затем последовательности были объединены в один файл:

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

Для триммирования была использована команда TrimmomaricSE, поскольку прочтения здесь одноконцевые:

```
TrimmomaticSE -phred33 SRR4240358.fastq.gz trimm.fastq.gz ILLUMINACLIP:adapters.fa:2:7:7
```

В результате работы программы был получен файл с очищенными последовательностями trimm.fastq.gz. Также была получена информация о количестве чтений:

- поданных на вход - 10543839
- сохраненных - 10368884 (98.34%)
- остатки адаптеров - 174955 (1.66%)

b. **Фильтрация**

С помощью Trimmomatic были удалены концевые нуклеотиды чтений качеством ниже 20 и чтения длины меньше 32:

```
TrimmomaticSE -phred33 trimm.fastq.gz trimm2.fastq.gz
TRAILING:20 MINLEN:32
```

Результаты триммирования:

- поданных на вход - 10543839
- удалено - 2352447 (22.69%)
- осталось - 8016437 (77.31%)

2. Запуск программы velveth

Сначала с помощью программы velveth были подготовлены k-меры длиной 31:

```
velveth velv 31 -short -fastq.gz trimm2.fastq.gz
```

Далее была произведена сборка на основе этих k-меров с помощью программы velvetg:

```
velvetg velv
```

N50 = 8600

Длины и покрытия трех самых длинных контигов были найдены с помощью конвейера:

```
grep '^>' contigs.fa | sort -k4 -t '_' -n -r | less
```

Самые длинные контиги:

Номер	Длина	Покрытие
56	9821	29.475859
34	18714	29.922678
40	16436	30.793623

Далее нужно было посмотреть, есть ли контиги с аномальным покрытием, это было сделано аналогичной командой, то сортировка по другому столбцу (с -r и без -r).

Контиги с аномально большим покрытием:

Номер	Длина	Покрытие
18	60	412.100006
97	53	405.245270

Контиги с аномально малым покрытием:

Номер	Длина	Покрытие
333	31	1.709677
143	31	3.064516

3. Анализ результатов

На данном этапе три самых длинных контига были сравнены с хромосомой *Buchnera aphidicola* (GenBank/EMBL AC — CP009253) при помощи программы megablast (Рис 1,2,3).

На каждом из графиков разрывы в прямой указывают на неконсервативные участки, разные для двух последовательностей.

a. Контиг 56

В случае с 56 контигом нашлось 3 участка для выравнивания. Контиг некомплементарен последовательности и ложится на участок 496111-514772 на хромосоме (подробную таблицу с данными можно посмотреть ниже).

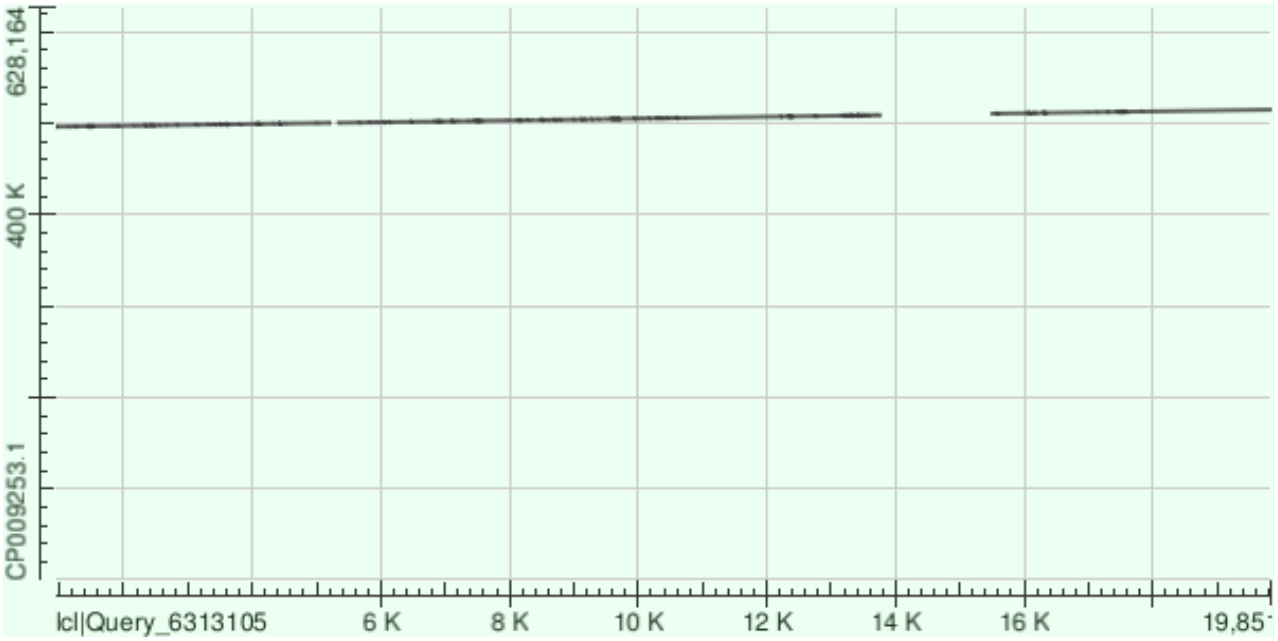

Рис. 1 Выравнивание с контигом 56

Hit table:

blastn
Iteration: 0
Query:
RID: UPMCUHNV114

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

Query_6313105	CP009253.1	75.618	8617	1750	265	5342	13787	500370	508806	0.0	3949
Query_6313105	CP009253.1	81.425	4393	739	57	15478	19851	510438	514772	0.0	3520
Query_6313105	CP009253.1	75.301	4324	914	121	948	5226	496111	500325	0.0	1927

b. **Контиг 34**

34 контиг выровнялся на 6 участков хромосомы, которые находятся в рамке 8599-26764. Контиг некомплементарен (соответствует тому же типу цепи, что и хромосома). Подробную таблицу с данными можно посмотреть ниже.

Рис. 2 Выравнивание с контигом 34

Hit Table:

blastn
Iteration: 0
Query:
RID: UPSP7K38114
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
6 hits found

Query_2696487	CP009253.1	85.405	2220	294	23	9387	11586	17962	20171	0.0	2278
Query_2696487	CP009253.1	77.613	3779	706	104	15025	18744	23067	26764	0.0	2163
Query_2696487	CP009253.1	75.969	3225	689	66	6139	9309	14727	17919	0.0	1583
Query_2696487	CP009253.1	78.297	2525	498	46	1	2495	8599	11103	0.0	1581
Query_2696487	CP009253.1	81.524	1851	291	41	12176	14000	20358	22183	0.0	1476
Query_2696487	CP009253.1	82.008	478	77	8	5505	5979	13994	14465	1.64e-110	398

b. Контиг 40

40 контиг выровнялся на 2 участка хромосомы, находящихся в рамке 462496-474242. Ход прямой на графике показывает, что последовательности комплементарны (подробную таблицу с данными можно посмотреть ниже).

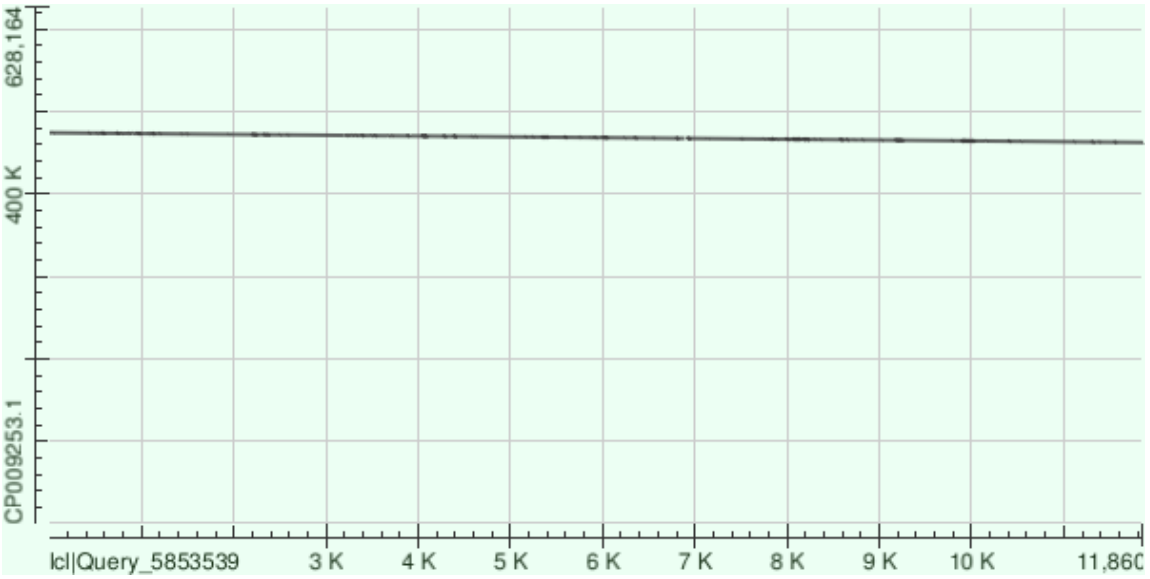

Рис. 3 Выравнивание с контигом 40

Hit Table:

blastn
blastn
Iteration: 0
Query:
RID: UPSUKAFJ114

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

Query_5853539	CP009253.1	76.756	6961	1414	167	3	6889	474242	4674	120.0	3703
Query_5853539	CP009253.1	76.989	5015	992	135	6919	11860	467421	462496	0.0	2719