Тарасов А.М.

Сборка de novo

Код доступа в базе ENA: SRR4240361
Директория для этой работы: /mnt/scratch/NGS/geonosianin/pr14
Перед тем как начать работу, необходимо скачать чтения в директорию на kodomo. Я сделал это при помощи команды *wget*:

wget
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR424/001/SRR4240361/SRR4240361.fastq.gz

Подготовка чтений программой trimmomatic

1. Удалим адаптеры Iilumina. Для этого объединим все файлы с адаптерами в один:

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

Теперь удалим чтения, представляющие из себя адаптеры:

```
TrimmomaticSE -phred33 SRR4240361.fastq.gz SRR4240361_no_adapters.fastq ILLUMINACLIP:adapters.fasta:2:7:7
```

Выдача содержит следующую информацию:

```
Input Reads: 7272621 Surviving: 7238089 (99.53%) Dropped: 34532 (0.47%)
```

То есть, как мы видим, 0,47% являлись адаптерами и были удалены.
Теперь следует избавиться от слишком коротких чтений (длина меньше 32 нуклеотидов), а также от нуклеотидов низкого качества.

```
TrimmomaticSE -phred33 SRR4240361_no_adapters.fastq SRR4240361_trim.fastq.gz TRAILING:20 MINLEN:32
```

Из результатов видим, что было отброшено 5.58% чтений.
Применим команду wc -c и подсчитаем, насколько уменьшился вес.
Исходный файл имел вес в 202179266 байт, итоговые риды имеют вес в 185730284 байт.

2.
Использование опции -help выдало следующую информацию:

```
./velveth directory hash_length {[-file_format][-read_type][-separate|-interleaved] filename1 [filename2 ...]} {...} [options]
```

Запустим программу velveth так, чтобы она на основе моего файла подготовила k-меры длины k=31.

```
velveth task2 31 -fastq.gz -short  SRR4240361_trim.fastq.gz
```

3. Запустим программу velvetg (сборка на основе k-меров):

```
velvetg task2
```

На выход программа выдает файла Contigs.fa, который расположен в директории task2. N50 - 25683. (информация содержится в последней строке выдачи).

Найдем длины у трёх самых длинных контигов и их покрытие:

```
grep '>' contigs.fa | cut -f2,4,6 -d '_' | sort -k2 -t '_' -rn | less
```

Это оказались 6-й контиг (длина: 49238, покрытие: 26.660851), 2-й (длина: 45555, покрытие: 26.450466) и 34-й (длина: 43866, покрытие: 23.514977)

Теперь отсортируем данные по длине покрытия и перенесем результат в файл, чтобы средствами Excel найти медиану:

```
grep '>' contigs.fa | cut -f2,4,6 -d '_' |sort -k3 -t '_' -rn > table.txt
```

Медиана: 11,98

Таким образом, аномально большими можно считать четыре верхних контига в списке. Например, это контиг 78 с покрытием 90.744682 и контиг 91 с покрытием 76.636360.

4. Я скачал нужные контиги в виде текстовых файлов и запустил алгоритм BLAST, сравнив их с хромосомой *Buchnera aphidicola*. Выдачу выравниваний я сохранил в Text, Hit Table (text) и DotPlot.

Контиг 6:

Нашлось 5 участков выравнивания.

50-12790, гэпы: 4%, процент идентичности: 75%
25809-33893, гэпы: 3%, процент идентичности: 77.8%
16429-23828, гэпы: 3%, процент идентичности: 77.7%
34098-38958, гэпы: 2%, процент идентичности: 79.6%
38989-45432, гэпы: 2%, процент идентичности: 76.2%

Выдача:

```
# 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_6_length_49238_cov_26.660851       CP009253.1      74.950  13010   2711    430     50      12790   127825  140555  0.0     5465
NODE_6_length_49238_cov_26.660851       CP009253.1      77.804  8168    1549    191     25809   33893   153752  161738  0.0     4796
NODE_6_length_49238_cov_26.660851       CP009253.1      77.747  7536    1434    178     16429   23828   144368  151796  0.0     4401
NODE_6_length_49238_cov_26.660851       CP009253.1      79.589  4914    891     92      34098   38958   161898  166752  0.0     3415
NODE_6_length_49238_cov_26.660851       CP009253.1      76.216  6517    1391    138     38989   45432   166750  173180  0.0     3301
```

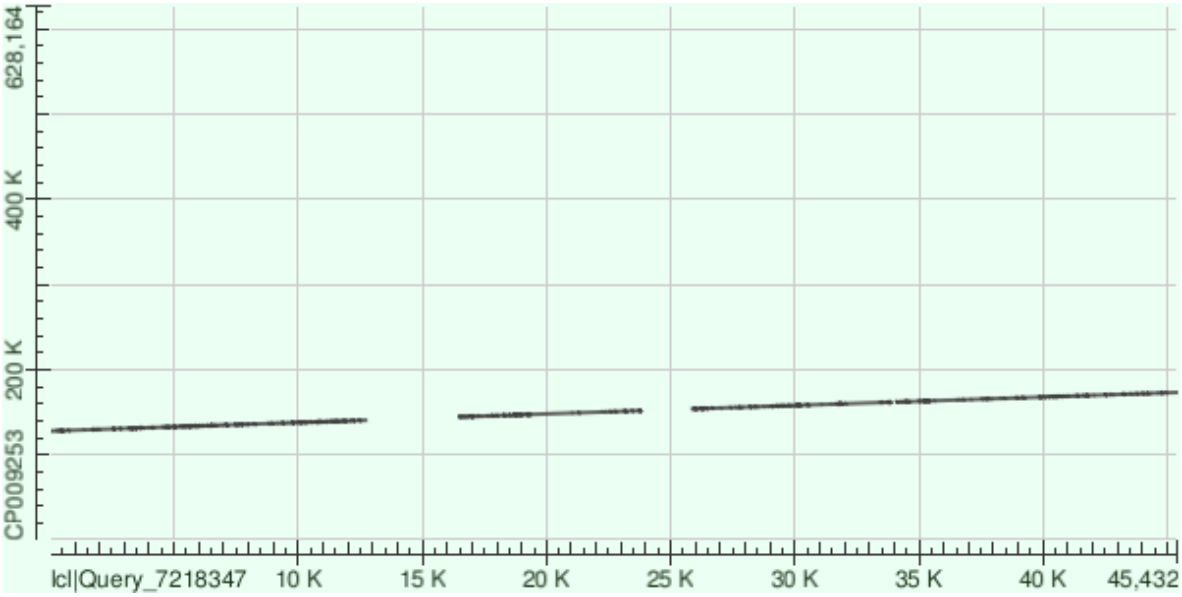

Рис. 1. DotPlot для шестого контига.

Контиг 2:

Ложится на хромосому в девяти местах.
10984-18297, гэпы: 2%, процент идентичности: 77%

18327-23268, гэпы: 3%, процент идентичности: 77%

40383-43410, гэпы: 1%, процент идентичности: 80.3%

5007-10881, гэпы: 4%, процент идентичности: 74.1%

33159-37811, гэпы: 3%, процент идентичности: 75.5%

12-3647, гэпы: 3%, процент идентичности: 76.5%

43540-45215, гэпы: 1%, процент идентичности: 79.1%

4122-4801, гэпы: 2%, процент идентичности: 82.2%

45337-45518, гэпы: 4%, процент идентичности: 88.9%

Выдача:

```
# Fields: query acc.ver, subject acc.ver, % identity, alignment length, mismatches, gap opens, q. start, q. end, s. start, s. end, evalue, bit score
# 9 hits found
NODE_2_length_45555_cov_26.450466       CP009253.1      77.020  7389    1490    168     10984   18297   474667  467412  0.0     4047
NODE_2_length_45555_cov_26.450466       CP009253.1      77.009  5015    991     135     18327   23268   467421  462496  0.0     2724
```

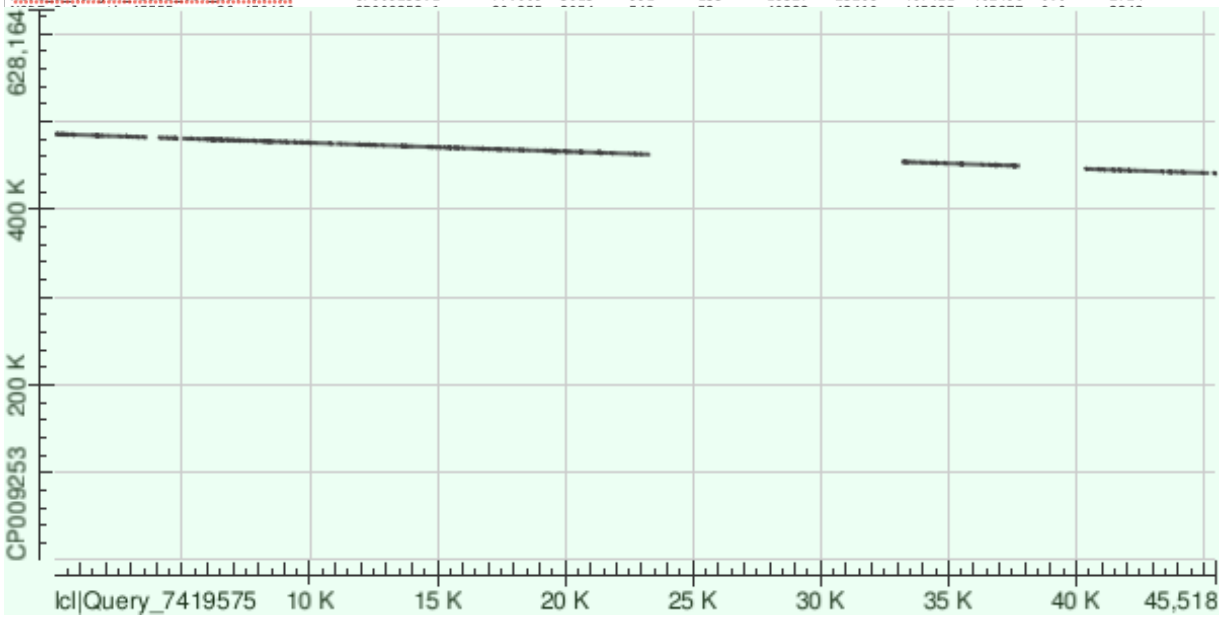

Рис. 2. DotPlot для второго контига.

Контиг 34:
Лёг на хромосому в семи местах.

14198-23677, гэпы: 3%, процент идентичности: 78.8%

23736-31957, гэпы: 5%, процент идентичности: 75.9%

8077-11648, гэпы: 2%, процент идентичности: 77%

37135-40501, гэпы: 2%, процент идентичности: 77.5%

977-5299, гэпы: 4%, процент идентичности: 73.4%

34011-35345, гэпы: 2%, процент идентичности: 76%

32205-33314, гэпы: 4%, процент идентичности: 76.2%

Выдача:

```
# Fields: query acc.ver, subject acc.ver, % identity, alignment length, mismatches, gap opens, q. start, q. end, s. start, s. end, evalue, bit score
# 7 hits found
NODE_34_length_43866_cov_23.514977      CP009253.1      78.789  9660    1688    274     14198   23677   266073  275551  0.0     6154
NODE_34_length_43866_cov_23.514977      CP009253.1      75.881  8396    1596    323     23736   31957   275566  283706  0.0     3890
NODE_34_length_43866_cov_23.514977      CP009253.1      77.080  3617    728     83      8077    11648   260224  263784  0.0     1993
NODE_34_length_43866_cov_23.514977      CP009253.1      77.534  3423    670     73      37135   40501   288181  291560  0.0     1969
NODE_34_length_43866_cov_23.514977      CP009253.1      73.400  4421    981     144     977     5299    253223  257546  0.0     1469
NODE_34_length_43866_cov_23.514977      CP009253.1      75.982  1349    297     25      34011   35345   285200  286535  0.0     671
NODE_34_length_43866_cov_23.514977      CP009253.1      76.237  1132    223     37      32205   33314   283963  285070  1.67e-158       558
```

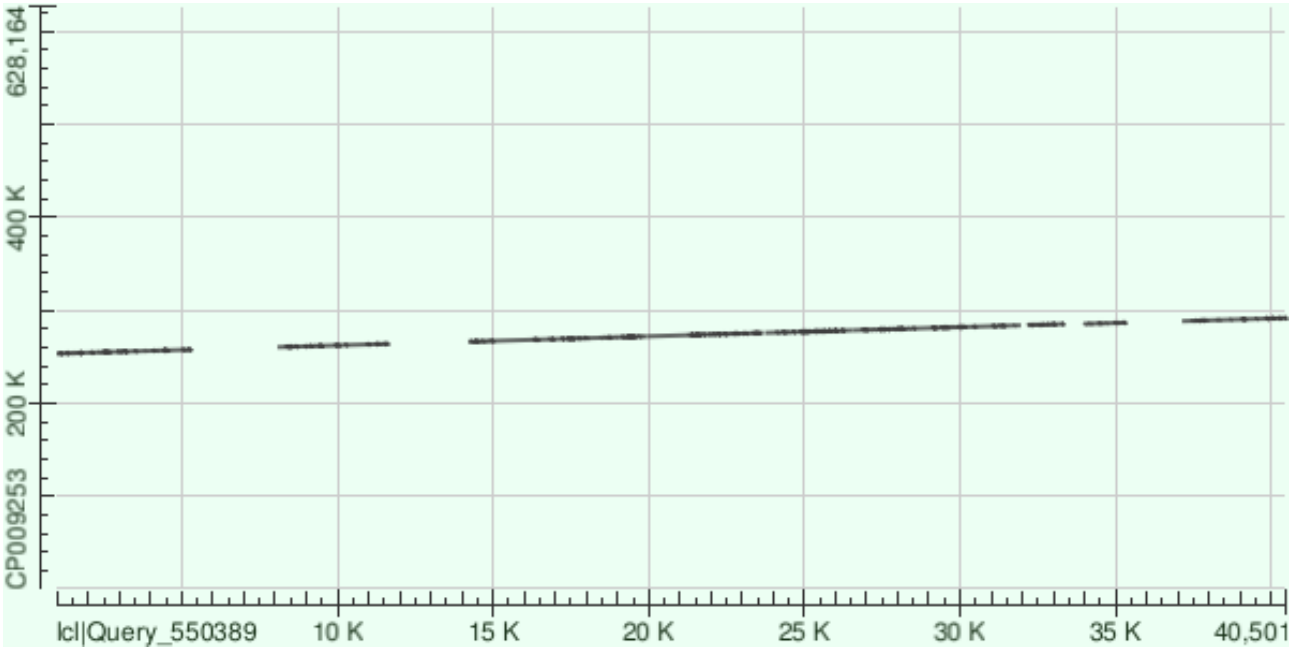

Рис. 3. DotPlot для тридцать четвёртого контига.