Сборка de novo

Код доступа проекта по секвенированию генома бактерии *Buchnera aphidicola* Tuc7: **SRR4240361**

Все дальнейшие команды выполнялись в директории: /mnt/scratch/NGS/iz-mi-al/pr15

Скачиваем чтения:

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

Скаченные чтения являются одиночными.

```
$ ls -lh SRR4240361.fastq.gz
-rw-r--r--. 1 iz-mi-al year-24 193M Dec 14 18:07 SRR4240361.fastq.gz
```

Далее мы проводим их подготовку к дальнейшей работе с помощью триммирования. Но для этого нам также надо получить список адаптеров, которые могут встретиться:

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

Триммируем адаптеры:

```
$ TrimmomaticSE SRR4240361.fastq.gz trim.fq.gz ILLUMINACLIP:adapters.fa:2:7:7
```

Важная информация из результата работы команды:

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

Полученный файл `trim.fq.gz` весит:

```
$ ls -lh trim.fq.gz
-rw-r--r--. 1 iz-mi-al year-24 192M Dec  9 12:11 trim.fq.gz
```

Остатками адаптеров оказалось **0,47%** чтений.

Триммируем по качеству, отрезая нуклеотиды с конца. Порог качества: 20; минимальная длина после триммирования: 32:

```
$ TrimmomaticSE trim.fq.gz res.fq.gz TRAILING:20 MINLEN:32
```

Важная информация из результата работы команды:

```
Input Reads: 7238089 Surviving: 6834335 (94.42%) Dropped: 403754 (5.58%)
```

После этого триммирования было удалено ещё **403 754** чтения.

Полученный файл res.fq.gz весит:

```
$ ls -lh res.fq.gz
-rw-r--r--. 1 iz-mi-al year-24 178M Dec  9 12:14 res.fq.gz
```

Далее необходимо запустить команду `velveth`. Перед этим создаём папку для результатов работы этой команды. Затем запускаем:

```
$ mkdir velvet
$ velveth ./velvet 31 -short -fastq.gz res.fq.gz
```

Параметры для этой команды:

31: k-меры такой длины буду создаваться;

`-short`: говорит о том, что чтения одиночные;

`-fastq.gz`: формат выходных файлов.

На выход получаем 3 файла. Они потребуются для дальнейшей работы `velvetg`.

Далее запускаем команду `velvetg`.

```
$ velvetg velvet/
```

На выход было получено несколько файлов, а также следующая информация:

```
Final graph has 477 nodes and n50 of 25683, max 49238, total 668902, using 0/6834335
reads
```

Отсюда можно получить значение **N50: 25 683**.

Все дальнейшие команды выполнялись в директории: /mnt/scratch/NGS/iz-mi-al/pr15/velvet

Для получения любой информации использовался файл: `contigs.fa`.

Дальше были получены длины (красным) трёх самых длинных контигов, а также их покрытие (зелёным):

```
$ grep '^>' contigs.fa | sort -t '_' -k 4 -nr | head -n 3
>NODE_6_length_49238_cov_26.660851
>NODE_2_length_45555_cov_26.450466
>NODE_34_length_43866_cov_23.514977
```

Дальше необходимо было найти такие контиги, покрытие которых сильно отличается от медианного. Для этого сначала была найдена медиана покрытий:

```
grep "^>" contigs.fa | awk -F'_' '{print $6}' | sort -n | awk '{
    a[NR]=$1
} END {
    if (NR%2==1) print a[(NR+1)/2]
    else print (a[NR/2]+a[NR/2+1])/2
}'
11.975
```
<– медиана

Сильно отличающимися контигами выбрано считать такие, чьё покрытие отличается от медианного больше чем в 5 раз, то есть меньше **2,395** или больше **59,875**.

Были найдены несколько контигов с большим покрытием и 1 с маленьким:

```
# Слишком маленькое
>NODE_391_length_63_cov_2.238095
# Слишком большое
>NODE_78_length_47_cov_90.744682
>NODE_91_length_33_cov_76.636360
>NODE_95_length_31_cov_64.903229
>NODE_185_length_48_cov_62.541668
```

Все эти контиги небольшой длины.

Далее были получены последовательности трёх самых длинных контигов:

```
$ sed -n '/>NODE_6_length_49238_cov_26.660851/,/^>/{/^>/!p}' contigs.fa > node_6.fa
$ sed -n '/>NODE_2_length_45555_cov_26.450466/,/^>/{/^>/!p}' contigs.fa > node_2.fa
$ sed -n '/>NODE_34_length_43866_cov_23.514977/,/^>/{/^>/!p}' contigs.fa > node_34.fa
```

Дальше в NCBI Datasets была найдена бактерия *Buchnera aphidicola*, но другого штамма, а именно: ***Buchnera aphidicola* str. Sg (Schizaphis graminum)**.

Ссылка на страницу в NCBI: https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000007365.1/

Далее каждый из полученных контигов был выронен на геном *Buchnera aphidicola* str. Sg (Schizaphis graminum) с помощью megablast (была изменена длина слова на 16, для большей точности).

>NODE_2

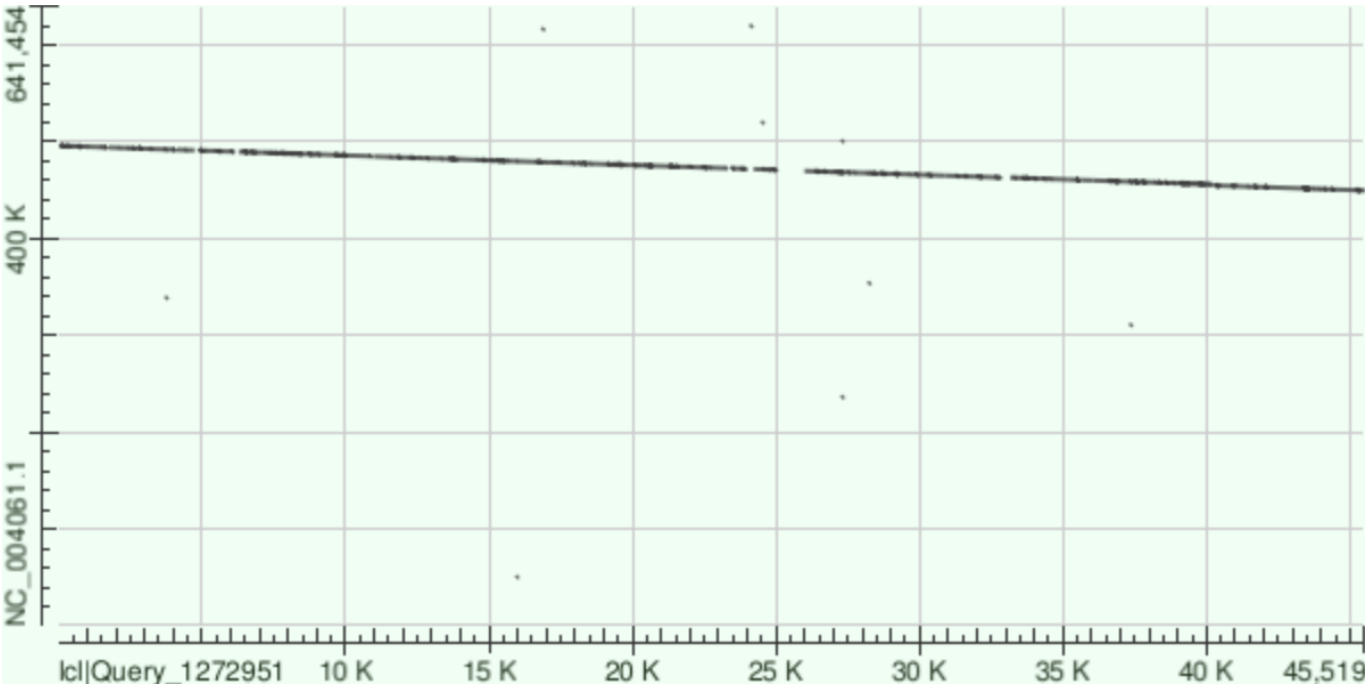

Рисунок 1. Карта локального сходства контига NODE_2

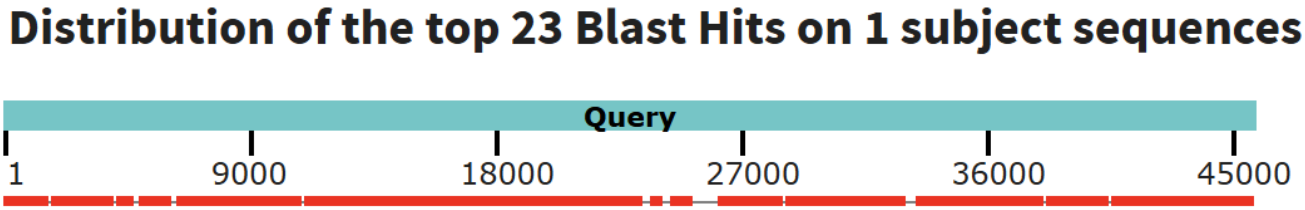

Рисунок 2. График расположения контига NODE_2 на геноме

Данный контиг выровнялся только один раз. На карте локального сходства можно увидеть несколько пробелов. На графике (**рис. 2**) видно 12 инделей. Выравнивание на участке хромосомы **449 737 – 496 030**. Гэпов в сумме получилось: **1 128**. Доля контига, которая выровнялась на геном: **94%**. Число однонуклеотидных замен: **8 264**.

>NODE_6

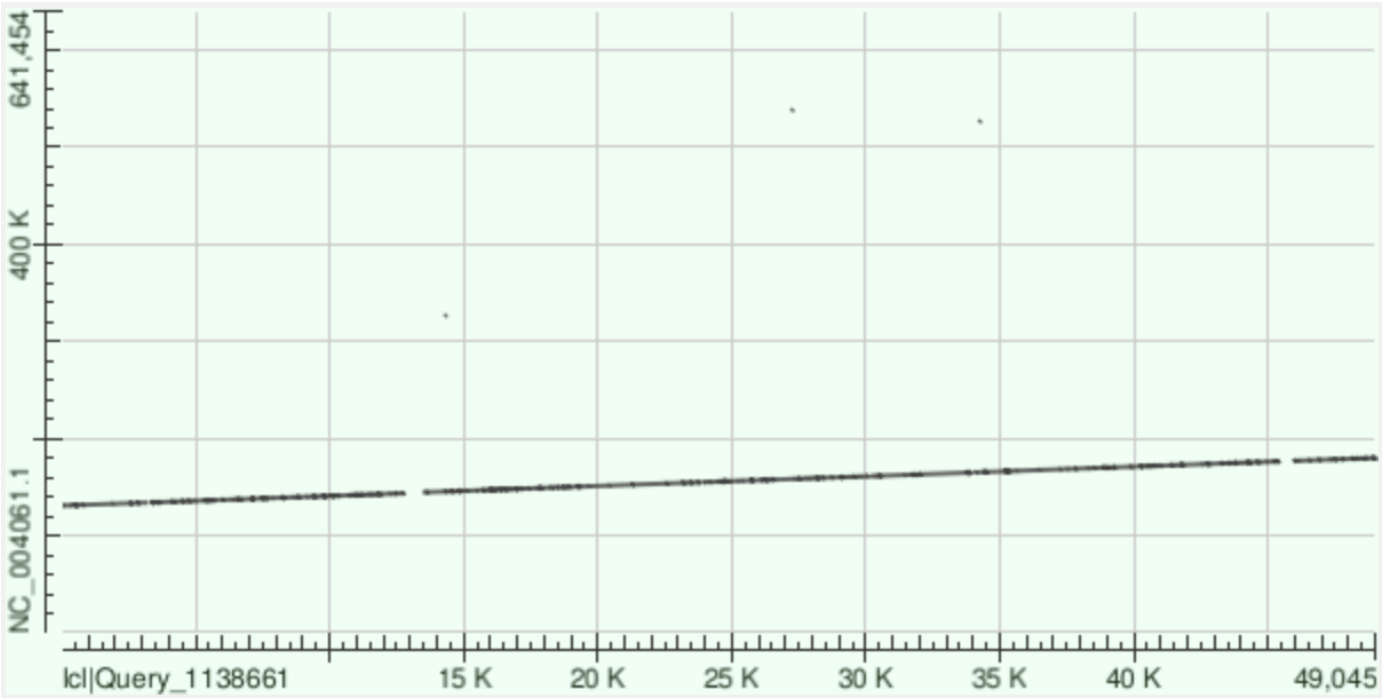

Рисунок 3. Карта локального сходства контига NODE_6

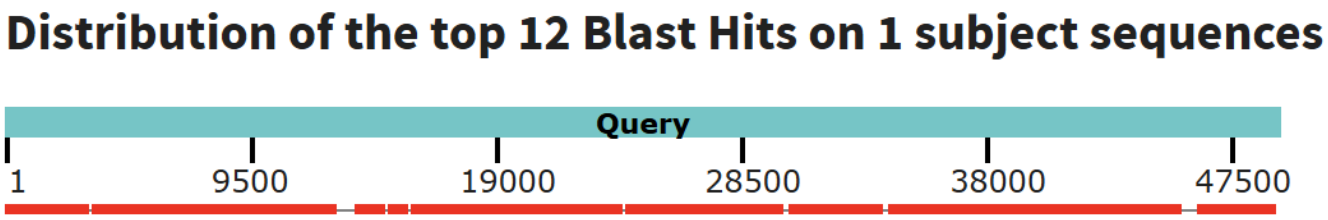

Рисунок 4. График расположения контига NODE_6 на геноме

Данный контиг тоже выровнялся один раз. На графике (**рис. 4**) видно 8 инделей. Выравнивание на участке хромосомы **130 911 – 180 115**. Гэпов в сумме получилось: **991**. Доля контига, которая выровнялась на геном: **95%**. Число однонуклеотидных замен: **9 055**.

>NODE_34

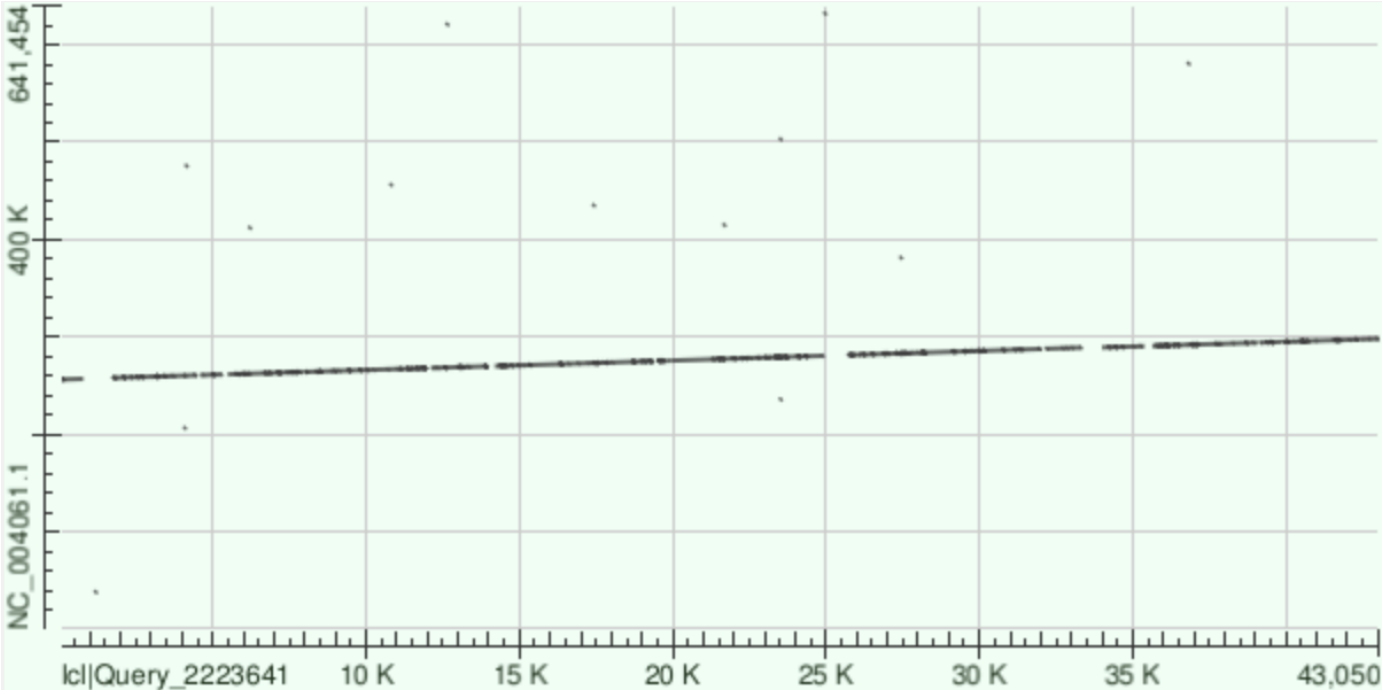

Рисунок 5. Карта локального сходства контига NODE_34

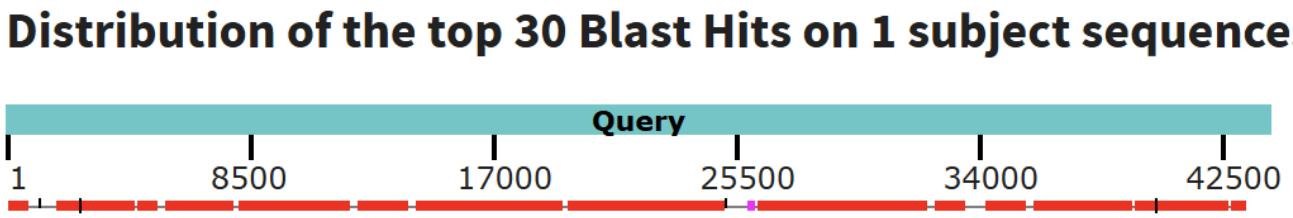

Рисунок 6. График расположения контига NODE_34 на геноме

Данный контиг тоже выровнялся один раз. На графике (**рис. 6**) видно 14 инделей. Выравнивание на участке хромосомы **256 307 – 298 229**. Гэпов в сумме получилось: **1 068**. Доля контига, которая выровнялась на геном: **88%**. Число однонуклеотидных замен: **6 197**.

Все три контига хорошо ложатся на банковский геном, что говорит о довольно точной сборке контигов.