Сборка de novo

I. Подготовка чтений

В данном задании мы будем работать с бактерией *Buchnera aphidicola str. Tuc7*, а именно с ее штамом Acyrthosiphon pisum. Эта бактерия является симбиотической для тли *Acyrthosiphon pisum*.

Первым этапом мы скачали наши чтения с сайта ENA с помощью команды: `wget "ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR424/006/SRR4240356/SRR4240356.fastq.gz"`

После этого нам надо было обрезать адаптеры, оставшиеся после секвенирования у чтений. Мы объединили все варианты адаптеров в один файл:

`cat *.fa > adapters.fa`

где * путь к папке с файлами, содержащими адаптеры

Следующий этап - применение команды Trimmomatic:

`TrimmomaticSE -phred33 SRR4240356.fastq.gz trim_file ILLUMINACLIP:adapters.fa:2:7:7`

где: SE - параметр для одноконцевых чтений,

-phred33 – параметр, указывающей систему кодировки качества,

ILLUMINACLIP:adapters.fa:2:7:7 – удаление адаптеров, список, которых содержится в файле adapters.fa, 2 - допустимые несовпадения в seed, 7 – минимальный балл качества для обрезания в палиндромном случае и обычном.

Справка программы:

- Input Reads: 7 511 529
- Surviving: 7 358 438 (97.96%)
- Dropped: 153 091 (2.04%)
- TrimmomaticSE: Completed successfully

Было отброшено всего 2.04% чтений, это хороший результат, потому что видимо эти чтения представляли из себя просто слипшиеся адаптеры, не представляющую для нас интереса, остальные чтения сохранились.

Следующим этапом мы провели второе триммирование, чтобы отобрать чтения с хорошим качеством:

```
TrimmomaticSE -phred33 trim_file finall_trim.fastq.gz \
 TRAILING:20 MINLEN:32
```

Где trim_file – файл, полученный после предыдущего триммирования,

TRAILING – обрезает нуклеотиды с конца прочтения, если их качество меньше 20,

MINLEN – оставляет только чтения длиной неменьше 32.

Справка по программе:

- Input Reads: 7358438
- Surviving: 7053346 (95.85%)
- Dropped: 305092 (4.15%)
- TrimmomaticSE: Completed successfully

Было отброшено также не очень много чтений, что для нас хорошо. Можно сделать вывод, что секвенирование было качественным и большая часть информации у нас сохранилась

Вес изначального файла: 167M, вес конечного: 155M.

II. Работа с программой velveth

Для создания сборки de novo мы применяем две программы velvet, которые используя граф Де Брейне могут составить нам сборку. Сперва мы применили программу velveth, которая создает список k-меров длиной 31 нуклеотид, встретившихся в наших чтениях.

Код: velveth Assem 31 -short -fastq.gz finall_trim.fastq.gz

Где: Assem – новая директория, куда программа положит полученные файлы,

31 – длина k-меров,

-short – у нас короткие и непарные чтения.

III. Работа с программой velvetg

Следующий этап – это создание самой сборки на основе графа Де Брейне, для этого используем программу velvetg: velvetg Assem/

Результаты:

- N50 – 65 554 base,
- Длина 3 самых длинных контигов: 111 962, 107 488, 80 939 баз
- Покрытие 3 самых длинных контигов: 38.660198, 34.174030, 37.524174
- Длина 3 самых коротких контигов: 31 баз
- Покрытие 3 самых коротких контигов: 3.064516, 4.129032, 5.838710
- Медианное значение покрытие: 17 чтений
- Всего 8 контигов с превышением медианного покрытия больше, чем в 5 раз
    - наибольшее из них 458.4 (его длина 282), покрытие больше медианного в 27 раз!
    - наименьшее 166.2 (его длина 45), больше медианного примерно в 10 раз
- Всего 2 контига с покрытием меньше, чем медианное в 5 раз
    - 3 (с длинной 31 база)
    - 2.6 (с длинной 91 база)

Находили с помощью кода:

```
grep "^>" contigs.fa | tr "_" "\t" | cut -f4,5,6 | sort -n | less
```

Покрытие для самых длинных контигов нормальное, так что можем продолжить их анализировать далее.

IV. Анализ

Следующим этапом нашей работы было сравнение самых длинных контигов, полученных ранее с геномом другого штамма этой бактерии. Для этого будем использовать megablast. Для сравнения я выбрал *Buchnera aphidicola str. Tuc7 (Acyrthosiphon pisum)* (AC: NC_011834).

1) Сравнивался контиг (node_6) с длинной 107 488 баз и покрытием, 34.174030.
   Покрытия у этого контига хромосомы составил 17%, идентичность составила 99.79%, e-value = 0.
   Координаты на хромосоме: 223685 – 331196

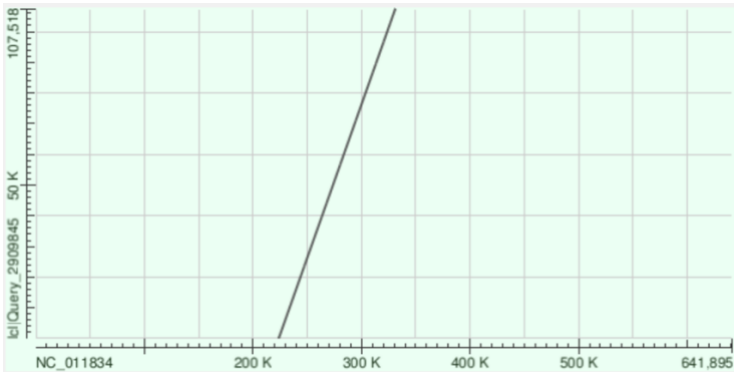

Рис. 1. Карта локального сходства контига (node_6) и NC_011834

Как мы можем заметить, последовательности почти совпадают, присутствует разве что небольшое количество SNP (всего 224) и всего 20 gaps.
Последовательности выровнялись плюс на плюс цепь.

2) Сравнение самого длинного контига (node_8) с NC_011834.
Результат выравнивания почти идентичный, покрытие 17%, идентичность 99.76%
Участок хромосомы: 458067 – 570061

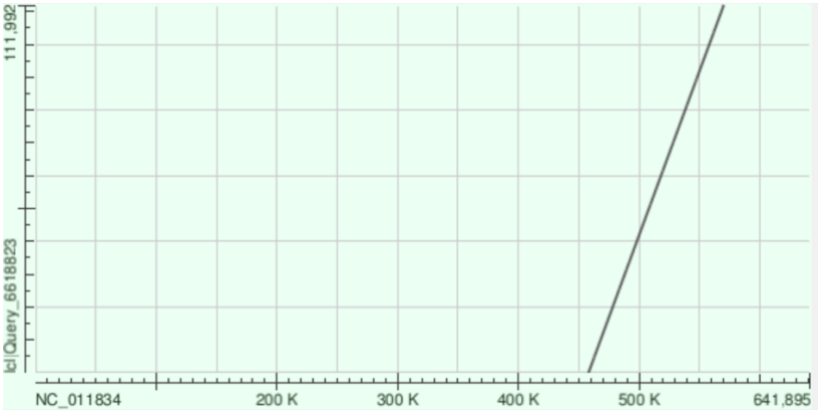

Рис. 2. Карта локального сходства контига (node_8) и NC_011834

Как и в предыдущем пункте видим очень точное наложение, с маленьким кол-вом SNP (272) и всего 13 gaps.
Выравнивание плюс на плюс цепь

3) Сравнение самого короткого из этих трех контигов (node_10) с NC_011834
Как и в двух предыдущих случаях покрытие высокое 13%, но чуть ниже, так как сам контиг короче. Идентичность 99.75%.
Участок хромосомы: 117271– 198241.

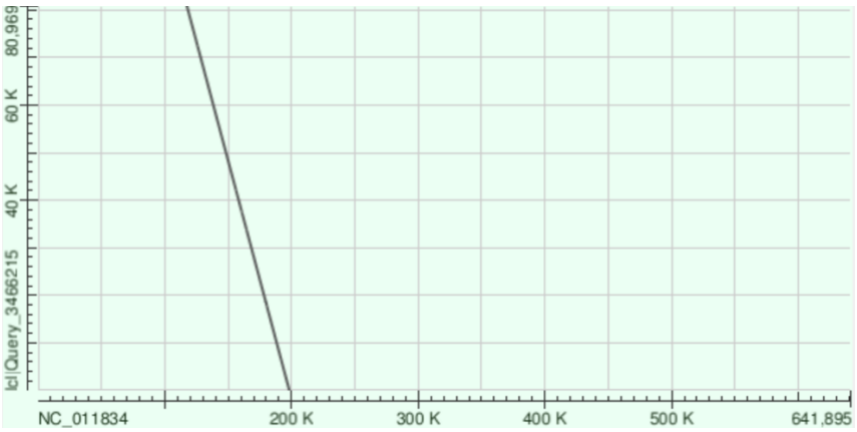

Рис. 3. Карта локального сходства контига (node_10) и NC_011834

Количество SNP также мало (201), количество gaps = 10. Отличие только в том, что выравнивание произошло минус цепь на плюс цепь, поэтому прямая на (Рис. 3.) развернута.

В целом, подводя итог, можно сказать, что наша сборка прошла успешно. Три самых длинных контига, имеют хорошую длину, очень качественно выравниваются на геном близкородственного организма и не перекрываются, а располагаются неподалеку друг от друга. Я думаю, что с такими результатами можно переходить к следующему этапу, и попытаться составить скэфолды из этих контигов.