Практикум 14

ФББ МГУ, 3-й семестр

Выполнил Гославский Лев Николаевич

В этом практикуме мне предстояло поработать с проектом по секвенированию бактерии Buchnera aphidicola str. Tuc7 по коду доступа **SRR4240359**
Чтения там непарные и довольно короткие.
Работа велась в директории /mnt/scratch/NGS/lev.g/pr14

Задание 1: Получение и обработка чтений

Я скачал файл с чтениями при помощи команды:
`wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR424/009/SRR4240359/SRR4240359.fastq.gz`
(использовал свой код доступа)

Теперь надо было удалить остатки адаптеров из чтений. Для этого я объединил все файлы с адаптерами в один при помощи средств bash:
`cat /mnt/scratch/NGS/adapters/*.fa > adap.fasta`

И использовал программу триммоматик для удаления адаптеров:
`TrimmomaticSE SRR4240359.fastq.gz SRR4240359_ad.fastq.gz ILLUMINACLIP:adap.fasta:2:7:7`

Остатками адаптеров оказались 0.41% чтений
Из выдачи: Input Reads: 13557938 Surviving: 13502066 (99.59%) Dropped: 55872 (0.41%)

Далее с правых концов чтений были удалены нуклеотиды с качеством ниже 20.
Остались только чтения, длина которых не меньше 32 нуклеотидов:
`TrimmomaticSE SRR4240359_ad.fastq.gz SRR4240359_tr.fastq.gz TRAILING:20 MINLEN:32`

Удалено было 1317986 (9.76%) чтений, осталось 12184080 (90.24%).
Размер файла до триммирования: 465,9 мб
После триммирования: 403,2 мб

Задание 2: Сборка

С использованием программы velveth были подготовлены k-меры длиной 31:
`velveth . 31 -short -fastq SRR4240359_tr.fastq.gz`

Затем проведена сборка: `velvetg .`
Теперь некая информация о сборке:
N50 = 70607

Чтобы получить информацию о трех самых длинных контигах из файла contigs.fa (он лежит тут: contigs.fa) использовал средства такой конвейер:

```
grep '>' contigs.fa | sort -k4,4 -t '_' -n -r | head -n 3
```

(он берет все строки с информацией о контигах - начинаются с “>”, затем сортирует их по длине с наибольшей до наименьшей, а затем выводит три первых значения)

Самые длинные контиги:
>NODE_11_length_125674_cov_44.550949 (длина = 125674 п.н., покрытие = 44.550949)
>NODE_1_length_108447_cov_42.009186
>NODE_14_length_71403_cov_39.411552

Находим медианное покрытие, чтобы оценить аномальные покрытия:

```
grep '>' contigs.fa | wc -l
```

Выдача: 285
Значит берём покрытие из 143-й строки отсортированного по покрытиям файла.

```
grep '>' contigs.fa | sort -k6,6 -t '_' -n -r | head -n 143 | tail -n 1
```

Медианное покрытие = 5.970588

Теперь находим парочку контигов с аномально большим покрытием (больше 29.85294):

```
grep '>' contigs.fa | sort -k6,6 -t '_' -n -r | head -n 2
```

>NODE_98_length_47_cov_139.489365
>NODE_80_length_40_cov_109.500000

И с аномально малым покрытием (менее 1,1941176):

```
grep '>' contigs.fa | sort -k6,6 -t '_' -n | head -n 2
```

>NODE_609_length_31_cov_2.032258
>NODE_231_length_63_cov_2.190476

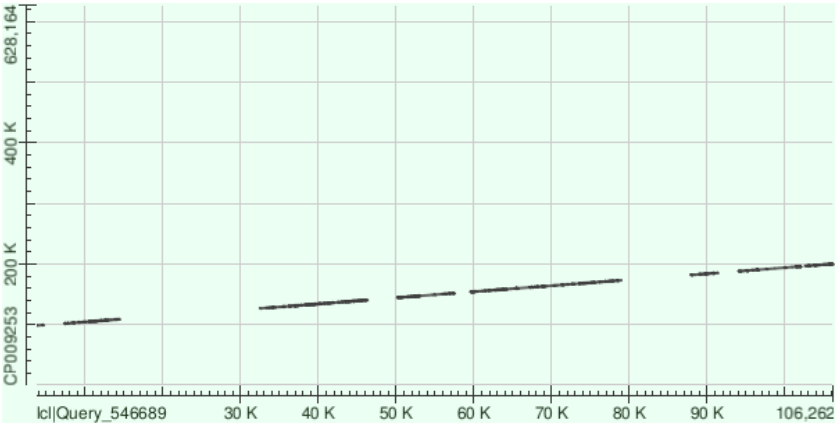
Длина у контигов с аномально большим покрытием и с аномально малым покрытием похожая и довольно маленькая, кстати. Это неудивительно, потому что при небольших длинах отрезков проще (с точки зрения вероятностей) получить аномальные покрытия этих отрезков (контиг).

Задание 3: Анализ

Теперь три самых длинных контига были сравнены с хромосомой *Buchnera aphidicola* (GenBank/EMBL AC — CP009253) алгоритмом megablast на сайте NCBI.
Тут можно посмотреть запросы в blast (недолго, но до проверки хватит):
1 - https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Get&RID=UW2D1SUV114
11 - https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Get&RID=UW1N4CZE114
14 - https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Get&RID=UW2EVM66114
А вообще последовательность контигов в файле contigs.fa

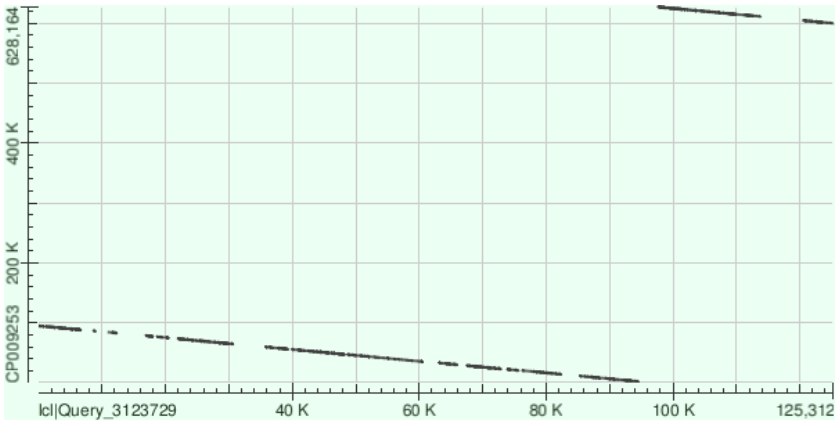
Контиг 1

1-й контиг соответствует участку хромосомы с координатами 98408 – 200246.
Идентичных нуклеотидов в выравнивании в среднем 74.95%. 15 участков контига выровнялись с бактериальной хромосомой. Гэпов от 0% до 7% в разных участках. Контиг соответствует той же цепи, что и хромосома (они не комплементарны друг-другу).



Контиг 11

11-й контиг соответствует участку хромосомы с координатами 2004 – 627104.
Идентичных нуклеотидов в выравнивании в среднем 82.85%. 25 участков контига выровнялись с бактериальной хромосомой. Гэпов от 2% до 7% в разных участках. Контиг соответствует комплементарной по отношению к хромосоме цепи.
Ещё на графике меня смущает этот кусок, который соответствует совершенно другому концу хромосомы…



Контиг 14

14-й контиг соответствует участку хромосомы с координатами 202390 – 273028.
Идентичных нуклеотидов в выравнивании в среднем 80.23%. 14 участков контига выровнялись с бактериальной хромосомой. Гэпов от 1% до 4% в разных участках. Контиг соответствует комплементарной по отношению к хромосоме цепи.