Практикум №14
Сборка de-novo

Звонцов О.А. группа 202

Код доступа проекта по секвенированию бактерии *Buchnera aphidicola:* ***SRR4240378***

1. Подготовка чтений программой trimmomatic

Предварительно нужно очистить чтения от адаптеров. Для одноконцевых чтений воспользовался командой **TrimmomaticSE:**

```
TrimmomaticSE -phred33  SRR4240378.fastq.gz
SRR4240378-cleared.fastq.gz ILLUMINACLIP:adapters.fa:2:7:7
```

- `-phred33` – конвертация качества в Phred-33
- `ILLUMINACLIP:<fastaWithAdaptersEtc>:<seed mismatches>:<palindrome clip threshold>:<simple clip threshold>` – step для обрезки адаптеров Illumina
    1. `<fastaWithAdaptersEtc>` – файл с последовательностями адаптеров (adapters.fa, собрал заранее)
    2. `<seed mismatches>` – максимальное количество несовпадений, при котором последовательности все равно будут считаться полностью совпадающими
    3. `<palindrome clip threshold>` – точность совпадения чтений с адаптерами, чтобы провести выравнивание палиндромных парноконцевых чтений
    4. `<simple clip threshold>` – необходимая точность совпадения адаптера с чтением

В результате было отброшено 81843 чтений (1.85%) из 4338744.

После этого с правых концов чтений были удалены нуклеотиды с качеством ниже 20 (TRAILING:20) и оставлены только чтения, длина которых не меньше 32 нуклеотидов (MINLEN:32):

```
TrimmomaticSE -phred33 SRR4240378-cleared.fastq.gz SRR4240378-trim.fastq.gz
TRAILING:20 MINLEN:32
```

После триммирования удалено 184006 чтений (4.24%)
Вес до очистки: 94476069 байт
После очистки: 87988710 байт

2. Подготовка k-меров

Для работы программ, использующих граф де Брёйна, сначала необходимо подготовить список k-меров, встретившихся в чтениях. Для этого применил программу **`velveth:`**

```
velveth kmers 31 -fastq.gz -short SRR4240378-trim.fastq.gz
```

- `kmers` - директория, с полученными k-мерами
- `31` - длина k-меров
- `-fastq.gz` - формат входного файла
- `-short` - тип чтений (в данном случае короткие и не парные)

3. Сборка на основе k-меров

Для этого использовал команду **velvetg**:

```
velvetg kmers
```

N50=7028

Поиск самых длинных контигов проводился командой:

```
grep '>' contigs.fa | tr '_' '\t' | sort -k4 -n |tail -n3
```

Контиг 8: длина 36746, покрытие 20.017199

Контиг 57: длина 19371, покрытие 20.546642

Контиг 15: длина 16745, покрытие 20.901762

Медианное покрытие: 18.45

С аномально низким покрытием:

- Контиг 166: длина 64, покрытие 2.843750
- Контиг 281: длина 72, покрытие 2.944444
- Контиг 271: длина 31, покрытие 3.225806

С аномально большим покрытием:

- Контиг 19: длина 2106, покрытие 100.555084
- Контиг 81: длина 934, покрытие 102.748390

4. Анализ

Три самых длинных контига были выровнены с помощью megablast на хромосому бактерии *Buchnera aphidicola*

Идеально выровнялся только контиг 15, остальные имеют разрывы. Это можно связать с тем, что количества чтений недостаточно и какие-то нуклеотиды оказались не покрыты. Другое объяснение: поскольку исследовались чтения одного из штаммов, то их последовательность отличается от референсной.

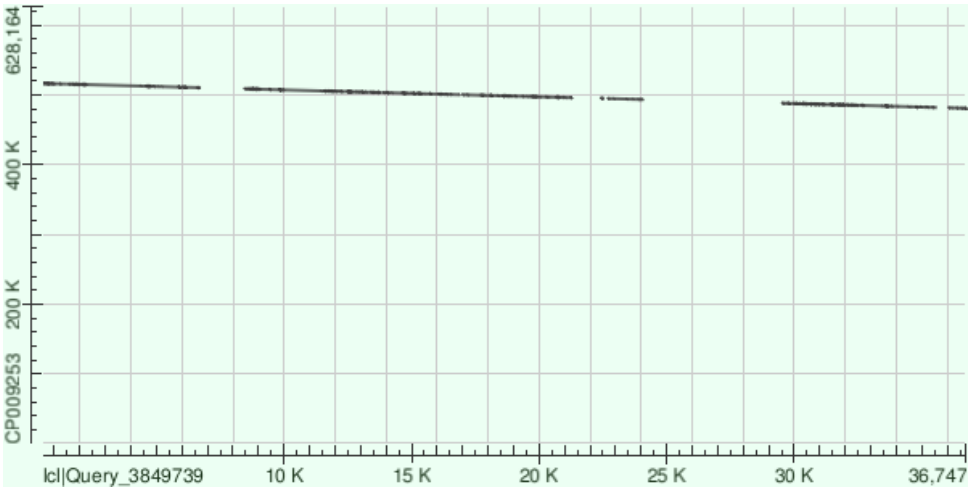

Рис.1 Dot-plot выравнивания контига 8 на хромосому бактерии

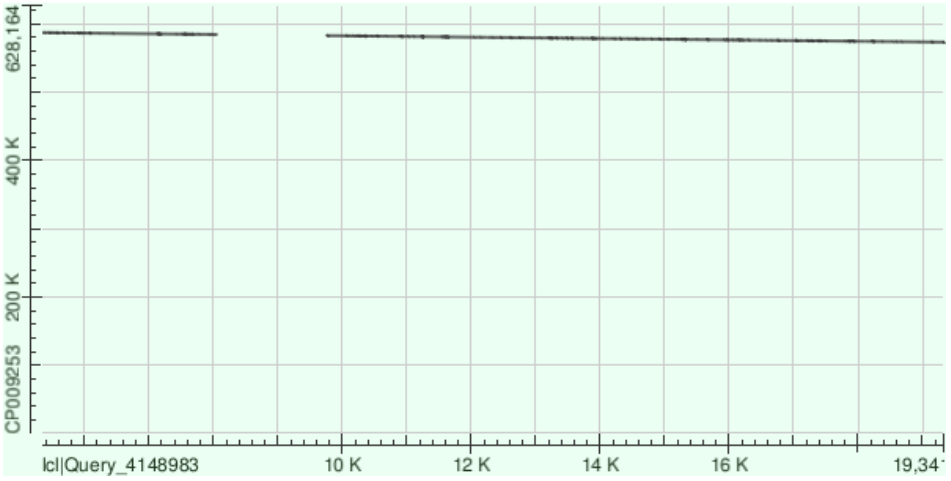

Рис.2 Dot-plot выравнивания контига 57 на хромосому бактерии

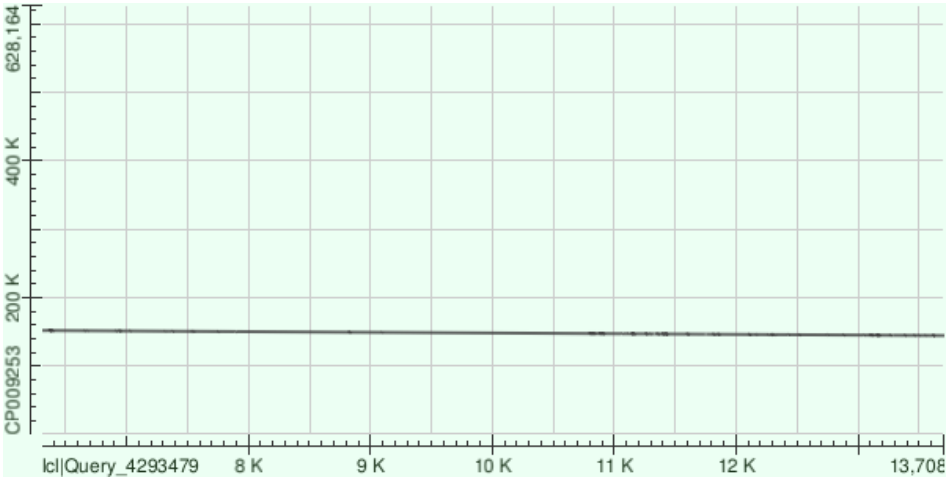

Рис.3 Dot-plot выравнивания контига 57 на хромосому бактерии