Практикум 14
Сборка de novo

Мой код доступа для проекта по секвенированию бактерии *Buchnera aphidicola* str. Tuc7: **SRR4240360**.
Для начала я скачала чтения с помощью **команды**:
`wgetftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR424/000/SRR4240360/SRR4240360.fastq.gz`

1. Подготовка чтений программой trimmomatic.

Сначала я подготовила файл со всеми адаптерами с помощью простой команды **cat.** Далее удалила возможные остатки адаптеров с помощью **команды:**
`TrimmomaticSE -phred33 SRR4240360.fastq.gz noad.fastq.gz ILLUMINACLIP:adapters.fasta:2:7:7`
Выдача:
`Input Reads: 8254632 Surviving: 8212774 (99.49%) Dropped: 41858 (0.51%)`
0.51% последовательностей чтений = остатки адаптеров.
После этого я удалила с правых концов чтений нуклеотиды с качеством ниже 20, оставив только такие чтения, длина которых не меньше 32 нуклеотидов.
Команда:
`TrimmomaticSE -phred33 noad.fastq.gz noad1.fastq.gz TRAILING:20 MINLEN:32`
Выдача:
`Input Reads: 8212774 Surviving: 7915474 (96.38%) Dropped: 297300 (3.62%)`
Удалено чтений: `Dropped: 297300 (3.62%)`
Изначальный размер файла: 194 мб
Итоговый размер: 184 мб

2. Velveth и подготовка k-меров.

Нужно было сделать так, чтобы программа **`velveth`** на основе моего файла подготовила k-меры длины k=31. С помощью опции -help я написала эту **команду:**
`velveth Assem 31 -short -fastq.gz noad1.fastq.gz`

3. Velvetg и сборка на основе k-меров.

Для сборки я воспользовалась **командой:**
`velvetg Assem`
Выдача: (файлы в папке Assem)
`contigs.fa Graph LastGraph Log PreGraph Roadmaps Sequences stats.txt`
N50: 43070
С помощью команды **длину и покрытие** самых длинных контигов:

Команда:

```
less contigs.fa | grep '>'| tr '_' '\t'| sort -k4 -n -r |head -3
```

Выдача:

```
>NODE   1       length  113474  cov     33.525459
>NODE   5       length  83603   cov     33.646065
>NODE   4       length  64155   cov     35.847324
```

Длины трех самых длинных контигов:

*length 64155
*length 83603
*length 113474

Их покрытие:

*35.847324
*33.646065
*33.525459

Контиги с аномально **большим** покрытием я нашла с помощью **команды**:

```
less contigs.fa | grep '>'| tr '_' 't'| sort -k6 -n|tail -3
```

Выдача:

```
>NODE   27      length  31      cov     92.709679
>NODE   140     length  40      cov     99.599998
>NODE   40      length  69      cov     109.391304
```

Контиги с аномально **маленьким** покрытием я нашла с помощью **команды**:

```
less contigs.fa | grep '>'| tr '_' '\t'| sort -k6 -n |head -3
```

Выдача:

```
>NODE   565     length  31      cov     1.612903
>NODE   264     length  33      cov     2.666667
>NODE   358     length  73      cov     2.671233
```

4. Megablast и анализ.

В этой части задания нужно было сравнить программой **megablast** каждый из трёх самых длинных контигов с хромосомой *Buchnera aphidicola* (GenBank/EMBL AC — CP009253).

*** Контиг 1:**

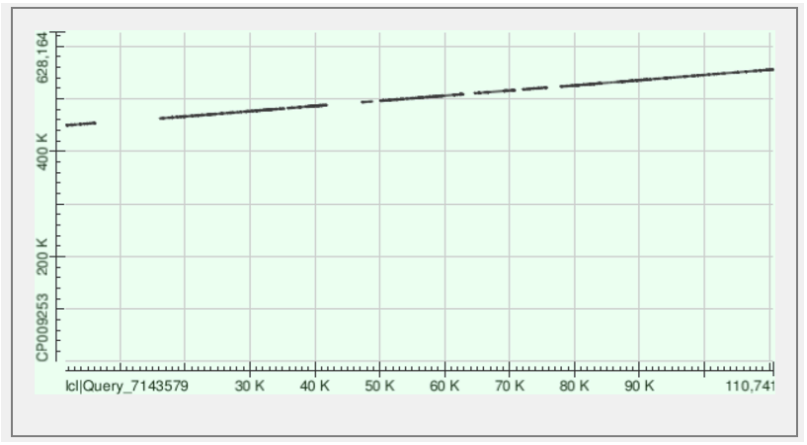

рис.1. Dot plot 1 контиг

По **Dot Plot** видно, что контиг выровнялся 425-555к на хромосоме.

*** Контиг 5:**

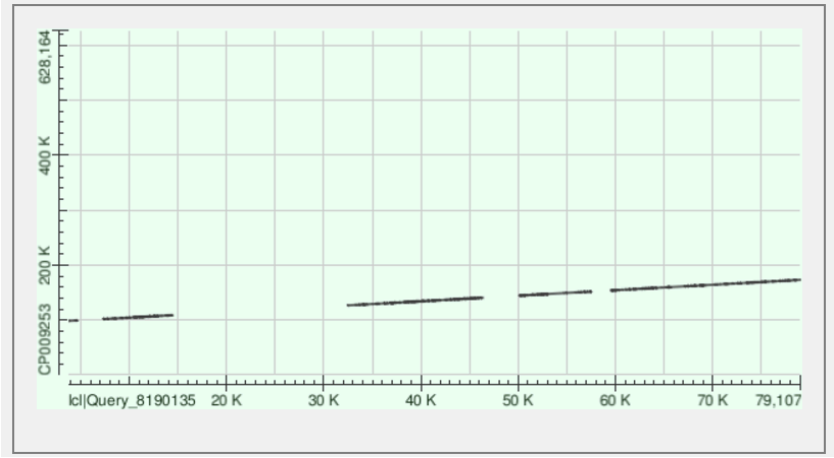

рис.2. Dot plot 5 контиг

Контиг выровнялся на участок 99-175к.

*** Контиг 4:**

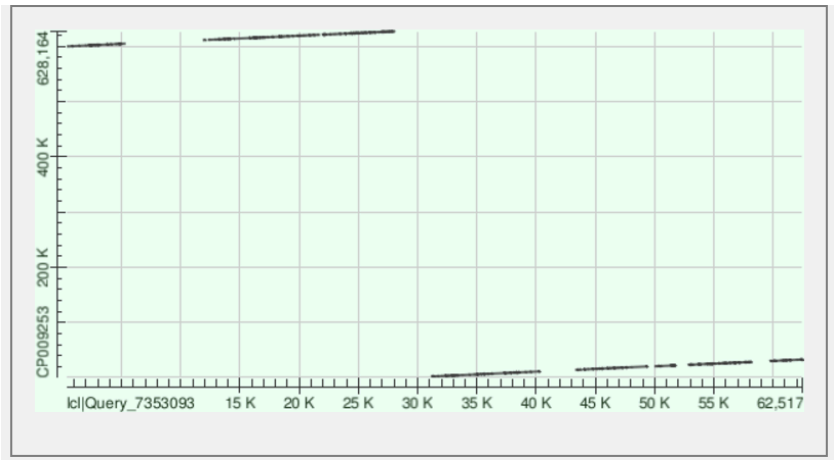

рис.3. Dot plot 4 контиг

Контиг выровнялся на участок 599-31к.