

ПРАКТИКУМ 14. Сборка de novo

Федосова Софья

Мой код доступа проекта по секвенированию бактерии *Buchnera aphidicola str. Tisc7*: SRR4240380.

В директории /mnt/scratch/NGS/ я создала рабочую поддиректорию для этого практикума, а затем перешла в неё. После я скачала чтения. Сделала я это с помощью команды:

```
wget
```

```
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR424/000/SRR4240380/SRR4240380.fastq.gz
```

1) Подготовка чтений программой trimmomatic

Перед выполнением программы trimmomatic я создала файл, который является объединением всех адаптеров:

```
cat /mnt/scratch/NGS/adapters/* > adapters.fasta
```

Далее я удалила все возможные остатки адаптеров при помощи команды:

```
TrimmomaticSE -phred33 SRR4240380.fastq.gz SRR4240380_no_adapters.fastq.gz  
ILLUMINACLIP:adapters.fasta:2:7:7
```

После исполнения команды я получила следующую выдачу:

```
Input Reads: 5217318 Surviving: 5119144 (98.12%) Dropped: 98174 (1.88%)
```

Количество чтений на вход: 5 217 318. Чтений осталось: 5 119 144 (98.12%).

Удалилось: 98 174 (1.88%).

Следующим шагом было удаление чтений, качество которых в конце было меньше 20 и длина которых меньше 32 нуклеотидов. Следующая команда помогла мне в этом:

```
TrimmomaticSE -phred33 SRR4240380_no_adapters.fastq.gz  
SRR4240380_trimmed.fastq.gz TRAILING:20 MINLEN:32
```

Получила такую выдачу:

```
Input Reads: 5119144 Surviving: 4865359 (95.04%) Dropped: 253785 (4.96%)
```

Количество чтений на вход: 5 119 144. Осталось: 4 865 359 (95.04%).

Удалилось: 253 785 (4.96%).

Размер файла с изначальными чтениями: 108 Мб

Размер файла после удаления остатков адаптеров: 106 Мб

Размер итогового файла: 99 Мб

2) Подготовка К-меров

В этом задании необходимо подготовить k-меры длины $k = 31$ на основе моего файла. Разобравшись в команде, я написала следующую команду:

```
velveth kmers 31 -short -fastq.gz SRR4240380_trimmed.fastq.gz
```

Появилась директория kmers, и команда успешно подготовила k-меры.

3) Сборка на основе k-меров

Для сборки на основе k-меров я запустила команду velvetg:

```
velvetg kmers
```

Выдача команды:

```
Final graph has 401 nodes and n50 of 12042, max 25915, total 660284, using  
0/4865359 reads
```

N50: 12 042

Теперь нужно узнать три самых длинных контига и их покрытие. Это можно узнать из файла configs.fa, применив конвейер:

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

Получила:

```
>NODE 3    length 25915  cov  27.418676
>NODE 20   length 23850  cov  24.763815
>NODE 23   length 23807  cov  25.725922
```

Как видно из выдачи конвейера я получила 3 самых длинный контига.

Контиги с аномально большим покрытием:

```
>NODE 75   length 501    cov  86.361275
>NODE 11   length 2106   cov  126.008545
>NODE 56   length 934    cov  130.479660
```

Контиги с аномально маленьким покрытием:

```
>NODE 235  length 62     cov  2.419355
>NODE 110  length 80     cov  2.700000
>NODE 301  length 63     cov  2.888889
```

4) Анализ

В этом задании нужно сравнить программой megablast каждый из трех самых длинных контигов с хромосомой *Buchnera aphidicola* (GenBank/EMBL AC — CP009253).

Контиг 3:

Расположение контига на хромосоме:

Distribution of the top 3 Blast Hits on 1 subject sequences

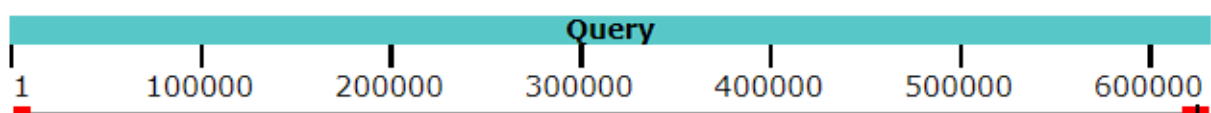


Рис. 1. Расположение контига 3 на хромосоме

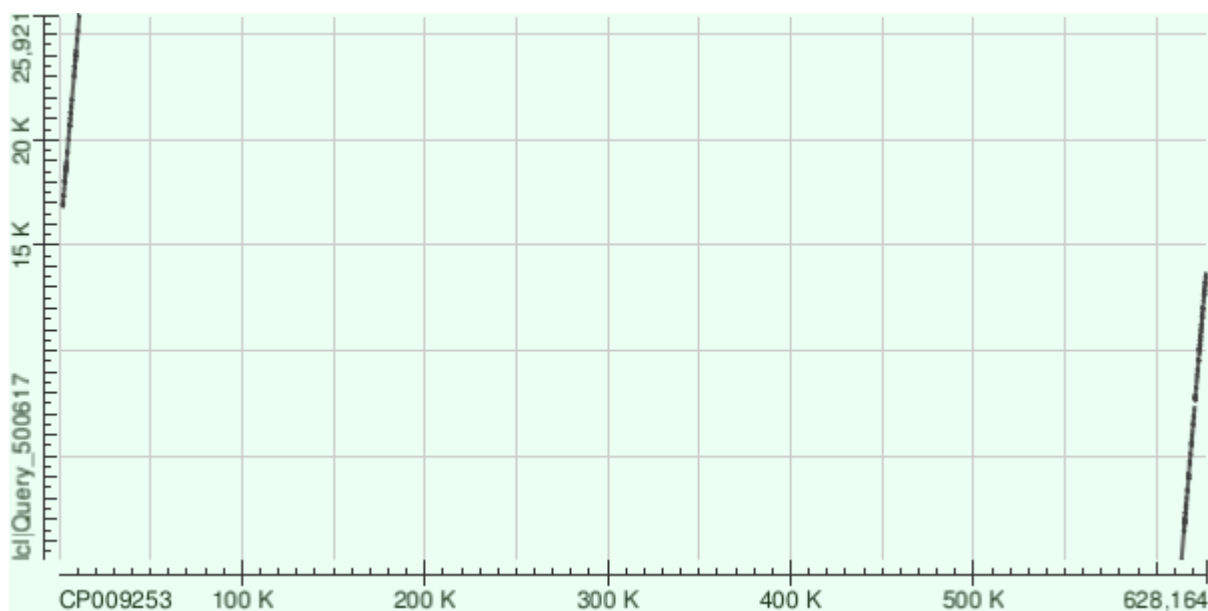


Рис. 2. DotPlot контига 3

По рисункам видно, что контиг выровнялся на участках хромосомы 3-мя кусочками: [2004 - 11103], [613658 - 620923] и [621055 - 627104]

Контиг 20:

Расположение контига на хромосоме:

Distribution of the top 4 Blast Hits on 1 subject sequences

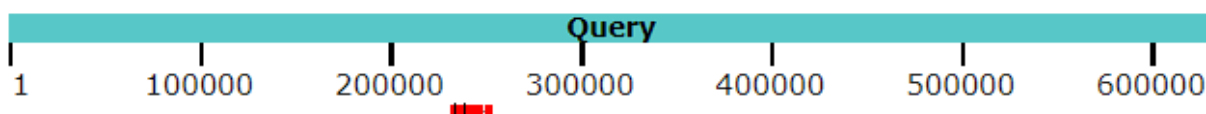


Рис. 3. Расположение контига 20 на хромосоме



Рис. 4. DotPlot контига 20

Таже видно, что контиг 20 выровнялся на участках хромосомы 4-мя кусочками: [236918 - 247596], [232364 - 236859], [229411 - 232057], и [248967 - 252164]

Контиг 23:

Расположение контига на хромосоме:

Distribution of the top 3 Blast Hits on 1 subject sequences

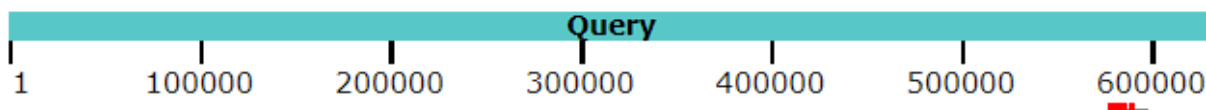


Рис. 5. Расположение контига 23 на хромосоме

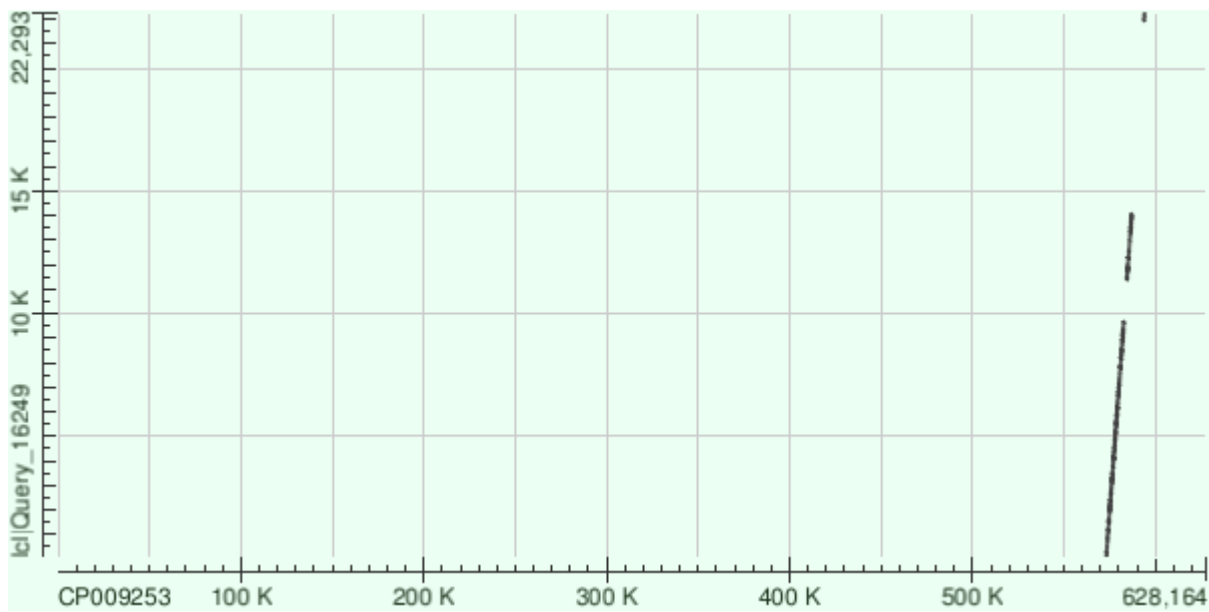


Рис. 6. DotPlot контига 23

По рисункам видно, что контиг выровнялся на участках хромосомы 3-мя кусочками: [573092 - 582686], [584329 - 587055] и [593743 - 594099]