Практикум 14. Сборка генома de novo

Перед работой с архивом чтений, его нужно было скачать. Была использована команда

```
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR424/009/SRR4240379/SRR4240379.fastq.gz
```

1. Подготовка чтений программой trimmomatic

Перед тем, как удалять остатки адаптеров, сначала все адаптеры для Illumina я записала в файл adapters.fa:

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

Далее запустила программу trimmomatic для одиночных чтений:

```
TrimmomaticSE SRR4240379.fastq.gz trim_1.fq.gz ILLUMINACLIP:adapters.fasta:2:7:7
-threads 20 -trimlog trim.log
```

Изначально было 7400155 чтений, размер составил 167M. Удалено 130303 (1.76% от изначального количества) чтений, размер стал равен 165M.

Затем использовала ту же программу, чтобы на правых концах чтений оставить только нуклеотиды с качеством 20 и выше, убрать чтения с длиной меньше 32 нуклеотидов:

```
TrimmomaticSE -phred33 trim_1.fq.gz trim_2.fastq.gz TRAILING:20 MINLEN:32 -threads 20
-trimlog trim_2.log
```

Программа удалила 295585 (4.07%) чтений, размер файла оказался 156M.

2. Программа velveth

Чтобы программа на основе полученного файла подготовила k-меры длины k=31, я использовала такую команду :

```
velveth kmers_velveth 31 -short -fastq.gz trim_2.fastq.gz,
```

где `kmers_velveth` - название папки, создаваемой программой, в которой лежат результаты её работы.

3. Программа velvetg

Запустила программу velvetg:

```
velvetg kmers_velveth
```

N50 = 25646 - из выдачи программы.

Три самых длинных контига и их покрытие нашла в файле stats.txt с помощью команды `sort -r -n -k 2 stats.txt | less`

Контиг	Длина	Покрытие
6	49912	35,91
9	49262	34,78
5	33085	36,26

Таблица 1. параметры трёх самых длинных контига

С помощью `sort -r -n -k 6 stats.txt | less` нашла контиги с аномально большими покрытиями: 474299, 2694, оба были длины 1. Помимо этого был контиг длины 2083 с покрытием 172,52.

4. Анализ

Для дальнейшего анализа командой
`seqretsplit -filter contigs.fa dir/name.format`
получила файлы fasta контигов.

Сравнение программой megablast трёх самых длинных контигов с хромосомой Buchnera aphidicola

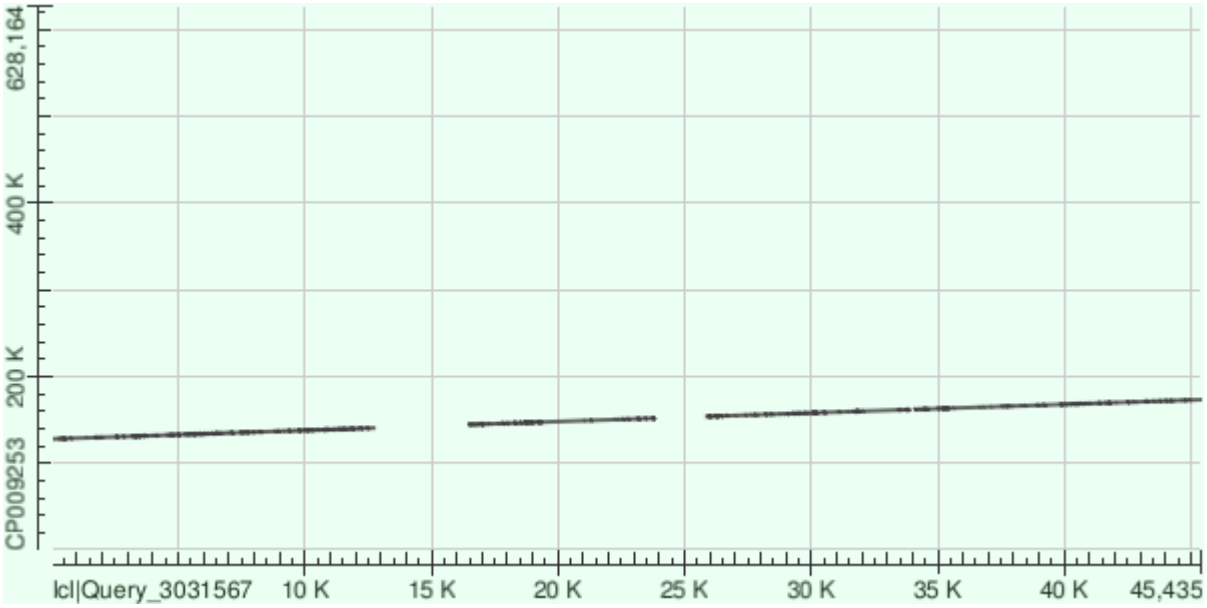

Рис. 1 DotPlot для 6 контига

Участок	Координаты хромосомы	Гэпы	Идентичные нуклеотиды
1	127825-140555	544/13008 (4%)	9741/13008 (75%)
2	153752-161738	266/8169(3%)	6347/8169(78%)
3	144368-151796	243/7536(3%)	5863/7536(78%)
4	161898-166752	108/4912(2%)	3910/4912(80%)
5	166750-173180	159/6517(2%)	4965/6517(76%)

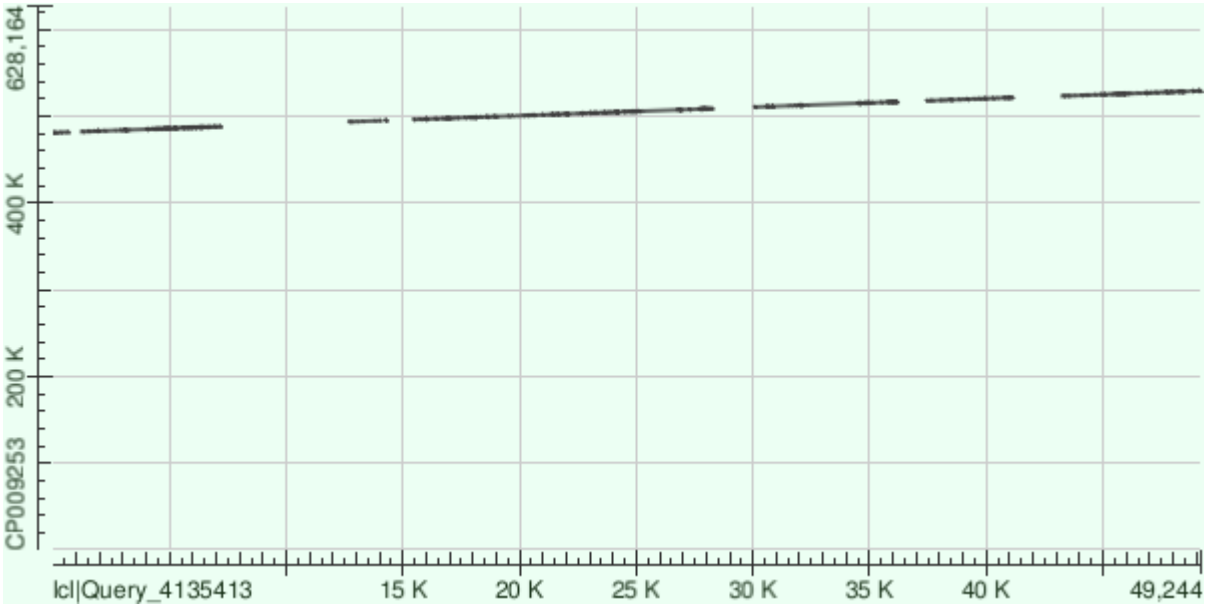

Рис. 2 DotPlot для 9 контига

Участок	Координаты хромосомы	Гэпы	Идентичные нуклеотиды
1	500370-508806	351/8617(4%)	6516/8617(76%)
2	510438-516539	187/6234(2%)	4897/6234(79%)
3	523105-528679	207/5685(3%)	4369/5685(77%)
4	481997-488106	308/6238(4%)	4621/6238(74%)
5	517766-521500	101/3783(2%)	2922/3783(77%)

6	496111-500325	154/4324(3%)	3255/4324(75%)
7	493487-494864	13/1384(0%)	1109/1384(80%)
8	480874-481545	20/686(2%)	564/686(82%)
9	528794-529211	26/425(6%)	357/425(84%)
10	495033-495148	5/120(4%)	108/120(90%)

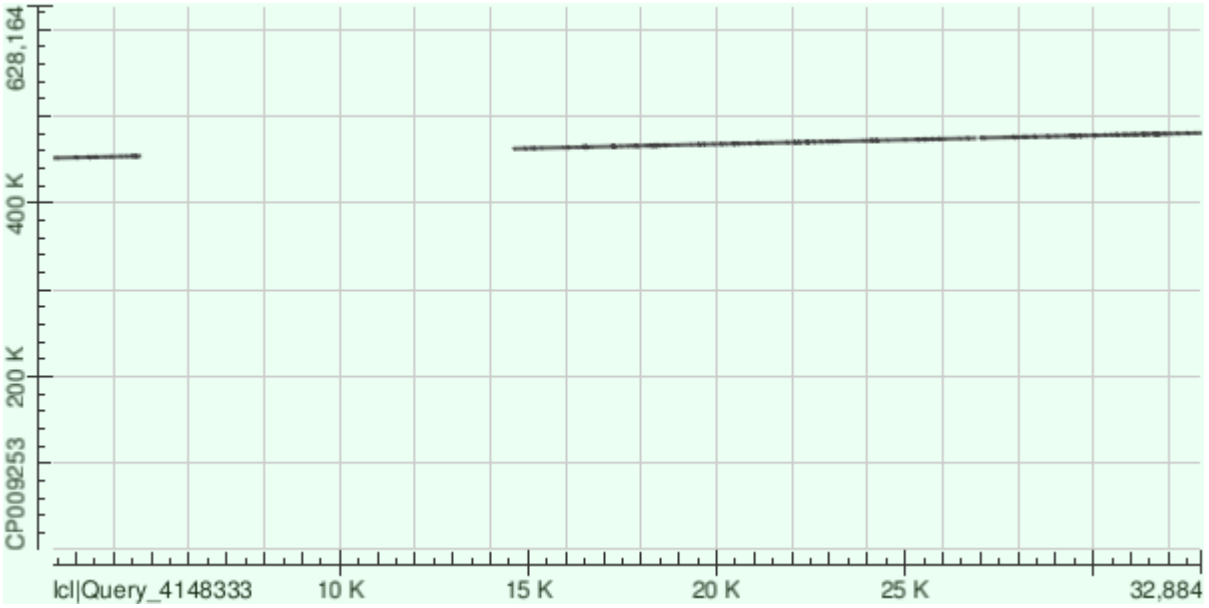

Рис. 3 DotPlot для 5 контига

Участок	Координаты хромосомы	Гэпы	Идентичные нуклеотиды
1	467412-474667	208/7388(2%)	5691/7388(77%)
2	462496-467421	162/5015(3%)	3861/5015(77%)
3	474844-480660	255/5974(4%)	4431/5974(74%)
4	451729-454069	55/2370(2%)	1827/2370(77%)