Практикум 14:Задание по сборке de-novo.

1).Подготовка чтений.

Сперва создадим нужную поддиректорию и перейдём в неё:

```
mkdir /mnt/scratch/NGS/youriy/bacteria
cd /mnt/scratch/NGS/youriy/bacteria
```

Теперь скачаем чтения бактерии с EMBL:

```
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR424/000/SRR4240360/SRR4240360.fastq.gz
```

Создадим единый файл с последовательностями адаптеров:

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

Теперь обрежем адаптеры при помощи TrimmomaticSE:

```
TrimmomaticSE -phred33 SRR4240360.fastq.gz noad.fastq.gz
ILLUMINACLIP:adapters.fasta:2:7:7
```

Получили файл без адапторов -noad.fasta.gz
Также программа в вывела в терминал, что адаптерами оказались 41858 (0.51%) всех чтений. Доделаем начатое и отфильтруем чтения так, чтобы они были не короче 32 нк и не менее 20 по качеству:

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

Отсеялось 297300 (3.62%) чтений. Измерим на сколько файл стал легче:

```
du -b *
```

До триммирования и безадапторов: 201803840. После фильтрации - 192395357.

2).Подготовка k-меров длины 31.

Для этого воспользуемся программой velveth. Но перед этим на всякий случай создадим директорию для вывода.

```
mkdir velout
velveth velout 31 -fastq -short trim.fastq.gz
```

3).Создание сборки на основе k-меров.

Это легко сделать программой velvetg:

```
velvetg velout
```

Результаты:

N50 - 43070

Команда для нахождения самых длинных контигов:

```
cat velout/contigs.fa | grep '>' | cut -d '_' -f2,4,6 | sort -t '_' -k2 -n -r | less
```

Контиги наибольшей длины:

№	Длина	Покрытие
1	113474	33.52546
5	83603	33.64607
4	64155	35.84732

Команды для поиска аномальных контигов:

```
cat velout/contigs.fa | grep '>' | cut -d '_' -f2,4,6 | sort -t '_' -k3 -n -r | less
cat velout/contigs.fa | grep '>' | cut -d '_' -f2,4,6 | sort -t '_' -k3 -n | less
```

№	Длина	Покрытие
40	69	109.39
140	40	99.6

27	31	92.71

4).Выравнивание.

Сначала получим последовательности самых длинных контигов.Я их получил при помощи языка программирования python3:

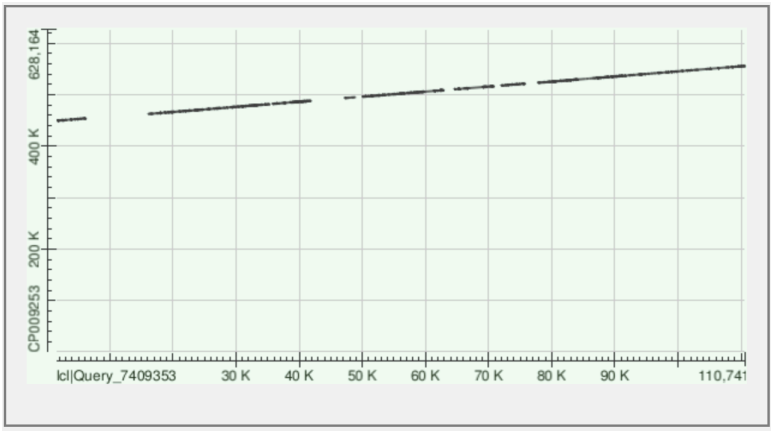
```
B=['1','5','4']
n=0
for i in B:
  with open('contigs.fa', mode='r') as contigs:
    with open(f"contig_{i}", mode='w') as contig:
      for l in contigs:
        if l.split('_')[0] == '>NODE':
          if l.split('_')[1] == i:
            n=1
          else:
            n=0
        if n==1:
          print(l.rstrip(),file=contig)
```

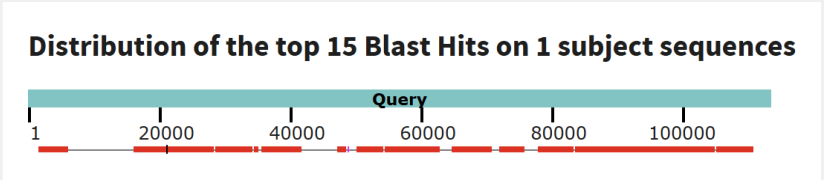
Теперь выровняем последовательности с использованием megablast:

Контиг1:

Первый контиг ложится довольно ровно на + цепь. Выравнивание состоит из 15 частей:

% ident	frag_length	contig_start	contig_end	genome_start	genome_end	E-value	Score
75.465	4732	1574	6226	449411	454069	0.0	2167
77.009	5015	16117	21058	462496	467421	0.0	2724
77.020	7389	21088	28401	467412	474667	0.0	4047
74.125	5971	28504	34378	474844	480660	0.0	2237
82.216	686	34584	35263	480874	481545	1.54e-162	573
74.078	6238	35738	41795	481997	488106	0.0	2278
80.058	1384	47268	48644	493487	494864	0.0	1014
89.167	120	48778	48896	495033	495148	1.43e-33	145
75.249	4323	50062	54339	496111	500325	0.0	1914
75.609	8614	54455	62900	500370	508806	0.0	3949
78.455	6238	64592	70771	510438	516539	0.0	3895
77.261	3782	72037	75766	517766	521500	0.0	2128
76.895	5687	77769	83357	523105	528679	0.0	3029
81.433	21721	83427	104897	528794	550219	0.0	17265
80.866	5655	105104	110741	550361	555905	0.0	4331

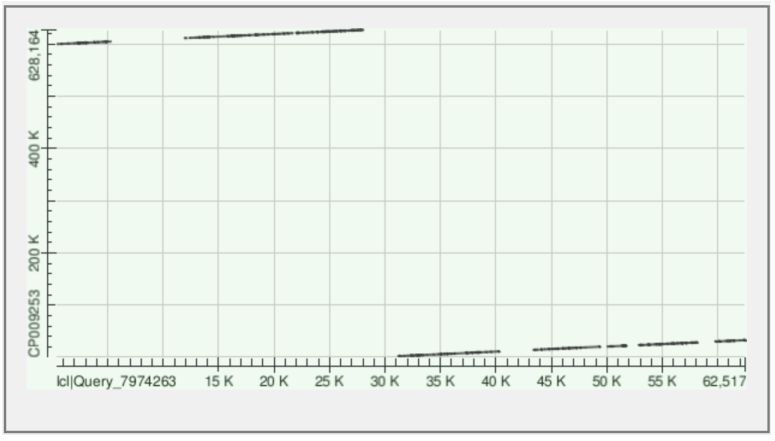


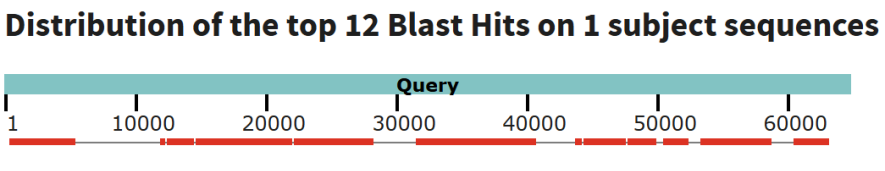


Контиг4:

Четвёртый контиг тоже ровно ложится на хромосому в направлении +(dot plot такой потому что геном кольцевой). Фрагментов - 12:

% ident.	frag_length	contig_start	contig_end	genome_start	genome_end	E-value	Score
78.380	9223	31205	40294	2004	11103	0.0	5749
82.218	478	43304	43778	13994	14465	1.21e-111	403
75.976	3226	43938	47108	14727	17919	0.0	1583
85.253	2231	47186	49396	17962	20182	0.0	2270
81.524	1851	49975	51799	20358	22183	0.0	1476
76.551	5433	52824	58173	23067	28363	0.0	2772
77.422	2777	59781	62517	30013	32745	0.0	1578
78.201	5046	393	5350	599832	604795	0.0	3068
79.461	297	11856	12151	611229	611524	2.83e-53	209
77.900	2086	12283	14349	611633	613671	0.0	1238
79.211	7379	14458	21762	613658	620926	0.0	4959
75.782	6173	21992	28039	621055	627104	0.0	2889





Контиг5:

Пятый контиг, также как и предыдущие, ложится на хромосому ровно в + направлении. Фрагментов - 8.

% ident.	frag_length	contig_start	contig_end	genome_start	genome_end	E-value	Score
81.132	901	3755	4651	98408	99303	0.0	713
76.533	7274	7332	14499	101712	108876	0.0	3777
83.736	1199	32467	33660	126623	127815	0.0	1123
74.950	13010	33725	46465	127825	140555	0.0	5465
77.747	7536	50104	57503	144368	151796	0.0	4401
77.804	8168	59484	67568	153752	161738	0.0	4796
79.589	4914	67773	72633	161898	166752	0.0	3415
76.216	6517	72664	79107	166750	173180	0.0	3301