Практикум 14

Скачать архив с чтениями бактерии Buchnera aphidicola str. Tuc7
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR424/008/SRR4240378/SRR4240378.fastq.gz

1. Подготовка чтений программой trimmomatic

a. Удалить возможные остатки адаптеров

Объединить последовательности адаптеров в один файл:
cat /mnt/scratch/NGS/adapters/*.fa > all_adapters.fa

Для обработки одноконцевых чтений использовался TrimmomaticSE:
TrimmomaticSE -phred33 -threads 10 SRR4240378.fastq.gz nadapters.fastq.qz ILLUMINACLIP:all_adapters.fa:2:7:7

Результат:
Input Reads: 4420587 Surviving: 4338744 (98.15%) Dropped: 81843 (1.85%). То есть 1,85% последовательностей оказалось остатками адаптеров.

b. Удалить с правых концов чтений нуклеотиды с качеством ниже 20, оставить только чтения с длиной не меньше 32 нуклеотидов.

TrimmomaticSE -phred33 -threads 10 nadapters.fastq.qz filt.fastq.qz TRAILING:20 MINLEN:32

Результат:
Input Reads: 4338744 Surviving: 4154738 (95.76%) Dropped: 184006 (4.24%). Было удалено 4,24% последовательностей. Размер nadapters.fastq.qz 437M (до очистки), а файла filt.fastq.qz - 418M (после очистки).

2. Подготовить k-меры длины k=31

velveth velveth 31 -fastq -short filt.fastq.qz

Второе слово - название папки с файлами; 31 - длина k-мера, -fastq - формат, -short - тк чтения короткие и непарные.

3. Сборка на основе k-меров

velvetg velveth
Второе слово указывает на папку с файлами для сборки.

Результат:
Final graph has 369 nodes and n50 of 7028, max 36746, total 657295, using 0/4154738 reads

N50: 7028

Команда для поиска самых больших контигов :
grep '^>NODE' contigs.fa | tr '_' '\t' | cut -f4,5,6 | sort -k1,1 -n -r | lesss

Результат:

Контиг:	Длина	Покрытие
8	36746	20.017199
57	19371	20.546642
15	16745	20.901762

Аномальное покрытие:

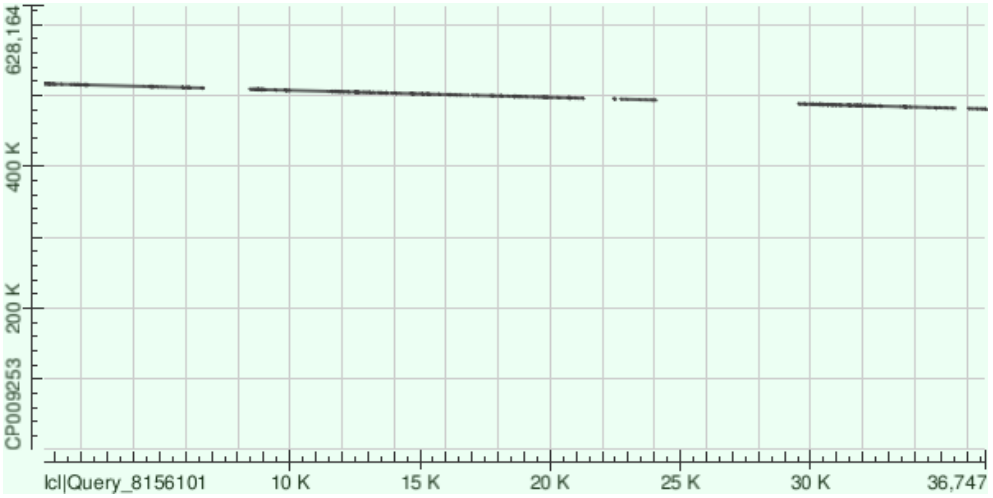
Контиг:	Длина	Покрытие
19	2106	100.555084
81	934	102.748390

4. Анализ

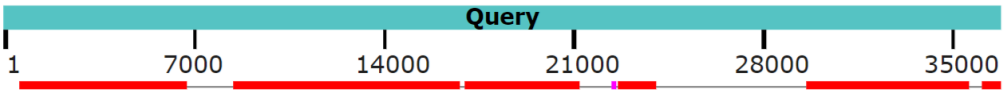
Выполнить сравнение трёх самых длинных контигов с хромосомой Buchnera aphidicola (CP009253) с помощью Megablast. Результаты:

Контиг 8:

- NODE_8_length_36746_cov_20.017199
- Координаты: 508806 - 500370
- На хромосому выровнялось 7 участков
- Открытий гэпов: 802
- Несовпадений: 5495

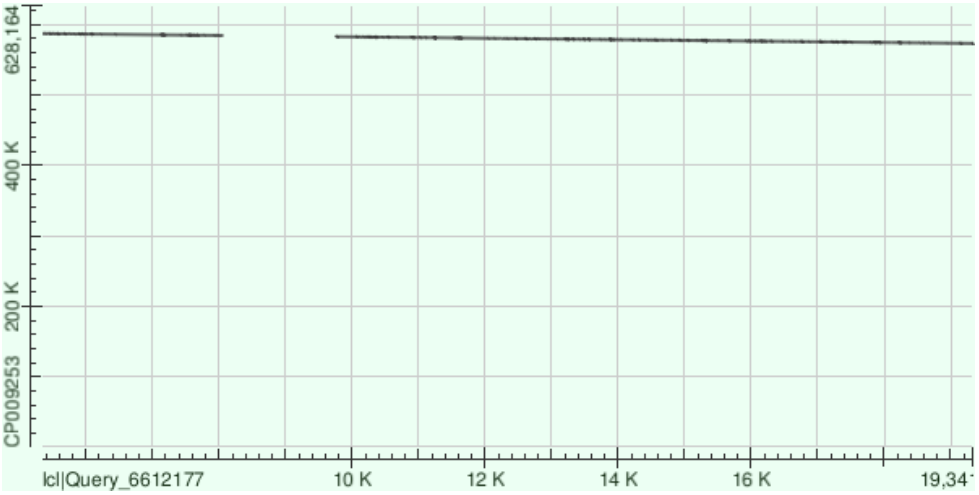


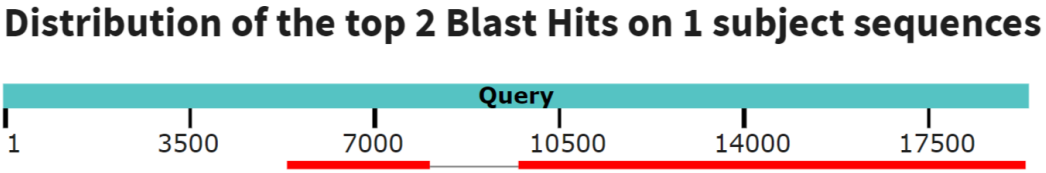
Distribution of the top 7 Blast Hits on 1 subject sequences




```
# blastn
# Iteration: 0
# Query: NODE_8_length_36746_cov_20.017199
# RID: NUNE7AYS114
# Database: n/a
# Fields: query acc.ver, subject acc.ver, % identity, alignment length, mismatches, gap opens, q. start, q. end, s. start, s. end, evalue, bit score
# 7 hits found
NODE_8_length_36746_cov_20.017199	CP009253.1	75.618	8617	1750	265	8431	16876	508806	500370	0.0	3949
NODE_8_length_36746_cov_20.017199	CP009253.1	78.553	6234	1150	140	562	6740	516539	510438	0.0	3932
NODE_8_length_36746_cov_20.017199	CP009253.1	74.078	6238	1309	241	29537	35594	488106	481997	0.0	2278
NODE_8_length_36746_cov_20.017199	CP009253.1	75.278	4324	915	121	16992	21270	500325	496111	0.0	1921
NODE_8_length_36746_cov_20.017199	CP009253.1	80.130	1384	262	13	22688	24064	494864	493487	0.0	1020
NODE_8_length_36746_cov_20.017199	CP009253.1	82.216	686	102	18	36068	36747	481545	480874	4.99e-163	573
NODE_8_length_36746_cov_20.017199	CP009253.1	90.000	120	7	4	22436	22554	495148	495033	9.96e-36	150
```

Контиг 57

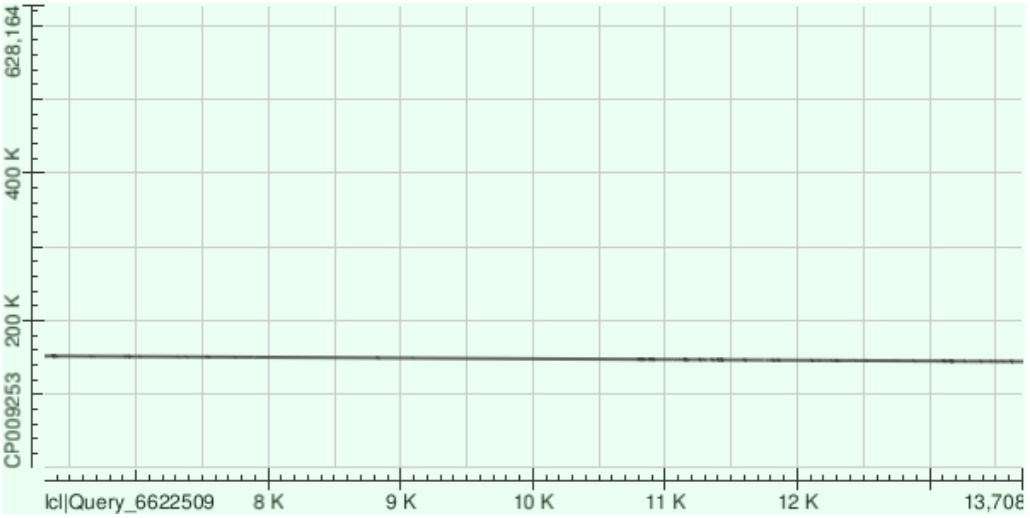
- NODE_57_length_19371_cov_20.546642
- Координаты контига: 587055- 573092
- На хромосому выровнялось 2 участка
- Открытий гэпов: 448
- Несовпадений: 2718
- 

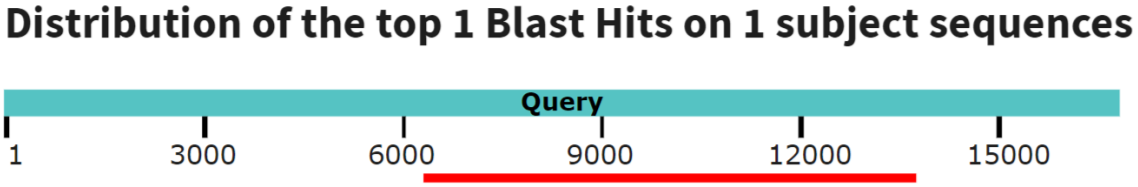
- 

-

```
# blastn
# Iteration: 0
# Query: NODE_57_length_19371_cov_20.546642
# RID: NUNX5NRF114
# Database: n/a
# Fields: query acc.ver, subject acc.ver, % identity, alignment length, mismatches, gap opens, q. start, q. end, s. start, s. end, evalue, bit score
# 2 hits found
NODE_57_length_19371_cov_20.546642   CP009253.1   73.427 9822   2149   362   9754   19341   582686   573092   0.0   3253
NODE_57_length_19371_cov_20.546642   CP009253.1   75.621 2777   569   86   5348   8066   587055   584329   0.0   1279
```

Контиг 15

- NODE_15_length_16745_cov_20.901762
- Координаты контига: 151796- 144368
- На хромосому выровнялся 1 участок
- Открытий гэпов: 178
- Несовпадений: 1430
- 

- 

-

```
# blastn
# Iteration: 0
# Query: NODE_15_length_16745_cov_20.901762
# RID: NUNXTXR1114
# Database: n/a
# Fields: query acc.ver, subject acc.ver, % identity, alignment length, mismatches, gap opens, q. start, q. end, s. start, s. end, evalue, bit score
# 1 hits found
NODE_15_length_16745_cov_20.901762    CP009253.1    77.800 7536    1430    178    6309    13708  151796          144368          0.0       4423
```