

Практикум 15. Сборка de novo

В данном практикуме я проанализировала часть проекта по секвенированию бактерии *Buchnera aphidicola* str. Tuc7. (по коду доступа SRR4240378)

Для начала я скачала архив с чтениями в рабочую директорию /mnt/scratch/NGS/kondratuk.ks/de_novo.
`wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR424/008/SRR4240378/SRR4240378.fastq.gz`

1. Подготовка чтений программой trimmomatic.

Создала файл со всеми адаптерами

```
cat /mnt/scratch/NGS/adapters/* >adapters.fa TrimmomaticSE -phred33
```

Избавилась от чтений, которые были адаптерами :

```
SRR4240378.fastq.gz SRR4240378_deladapt.fastq.gz ILLUMINACLIP:adapters.fa:2:7:7 2> trim.txt
```

Из 4420587 чтений 81843 (1.85%) оказались адаптерами

Далее я убрала слишком короткие и с низким качеством чтения (оставила чтения с длиной не менее 32 и качеством не менее 20)

```
TrimmomaticSE -phred33 SRR4240378_deladapt.fastq.gz SRR4240378_trim.fastq.gz
```

```
ILLUMINACLIP:adapters.fa:2:7:7 TRAILING:20 MINLEN:32 2> trim_cl.txt
```

Из 4338744 чтений 184006 (4.24%) не подошли .

Итого : из 4420587 чтений было убрано 265 849

SRR4240378.fastq.gz - 90MB

SRR4240378_deladapt.fastq.gz - 89MB

SRR4240378_trim.fastq.gz - 84MB

2. Программа velveth - подготовка k-меров

```
velveth velveth/ 31 -short -fastq.gz SRR4240378_trim.fastq.gz
```

3. Программа velvetg - сборка на основе k-меров

```
velvetg velveth/ >velvetg.txt
```

Результаты работы программы будут находиться в папке velveth. Из выдачи команды N50=7028, max = 36746.

Далее я хочу получить более подробную информацию о контигах.

создала файл с информацией о контигах (без их последовательностей).

```
grep '>' contigs.fa > contigs.txt
```

Затем скачала stats.txt и создала таблицу. Отсортировала по убыванию длины и взяла три верхних значения. (использовать stats удобнее - значения совпадают (проверила), а таблица сразу в удобном формате)

ID	lgth	short_cov
8	36746	20.017199
57	19371	20.546642
15	16745	20.901762

Медиана = 17,646867

Аномально большое покрытие - от 88 (таких набралось около 50шт)

ID	lgth	short1_cov
129	1	148170
257	1	968
225	1	923

Аномально малое покрытие - до 3.5 (таких набралось около 40шт)

ID	lgth	short1_cov
231	1	1
347	1	1
356	1	1

4. Анализ сборки

Далее я сравнила программой megablast каждый из трёх самых длинных контигов с хромосомой *Buchnera aphidicola* (GenBank/EMBL AC — CP009253).

`seqretsplit -filter ../contigs.fa dir/name.format` - для удобства разделила контиги на отдельные fasta-файлы, которые хранятся в папке configs.

Мне нужны файлы с именами node_8_length_36746_cov_20.017199, node_57_length_19371_cov_20.546642, node_15_length_16745_cov_20.901762 - в них находятся последовательности в fasta-формате самых длинных контигов с ID 8, 15, 57.

(Из text берется информация о гэпах, из Hit Table (text) берется информация о покрытиях контига)

Descriptions

Graphic Summary

Alignments

Dot Plot

Sequences producing significant alignments

☒ select all 1 sequences selected

Description	Scientific Name
☒ NODE_15_length_16745_cov_20.901762	

Download

Select columns

Show 100

?

FASTA (completesequences)

FASTA (aligned sequences)

Hit Table (text)

Hit Table (CSV)

Text

Descriptions Table (CSV)

XML

ASN.1

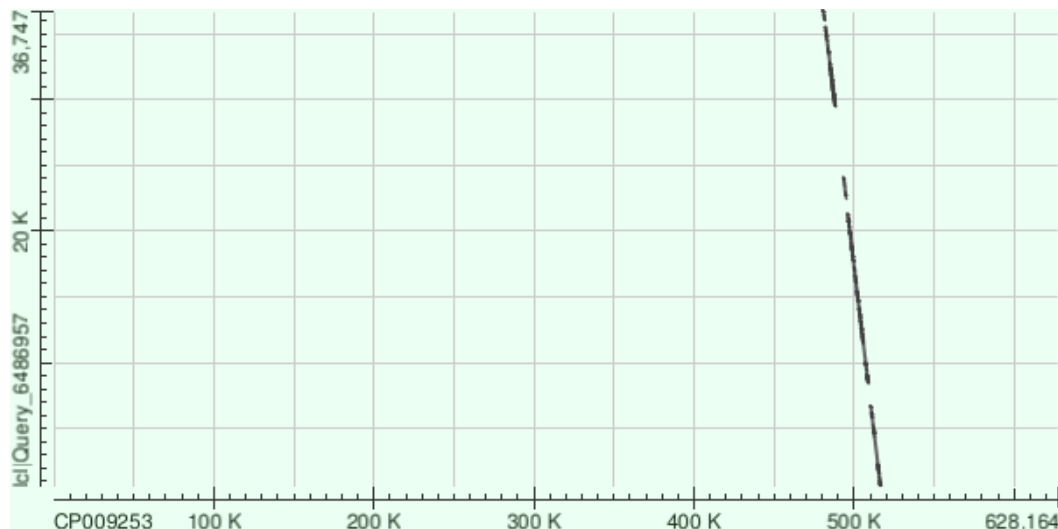
Graphics

MSA Viewer

E value	Per. Ident	Acc. Len	Accession
0.0	77.81%	16775	Query_7037189

node_8_length_36746_cov_20.017199

DotPlot



8-й контиг ложится на банковскую хромосому в 7 местах :

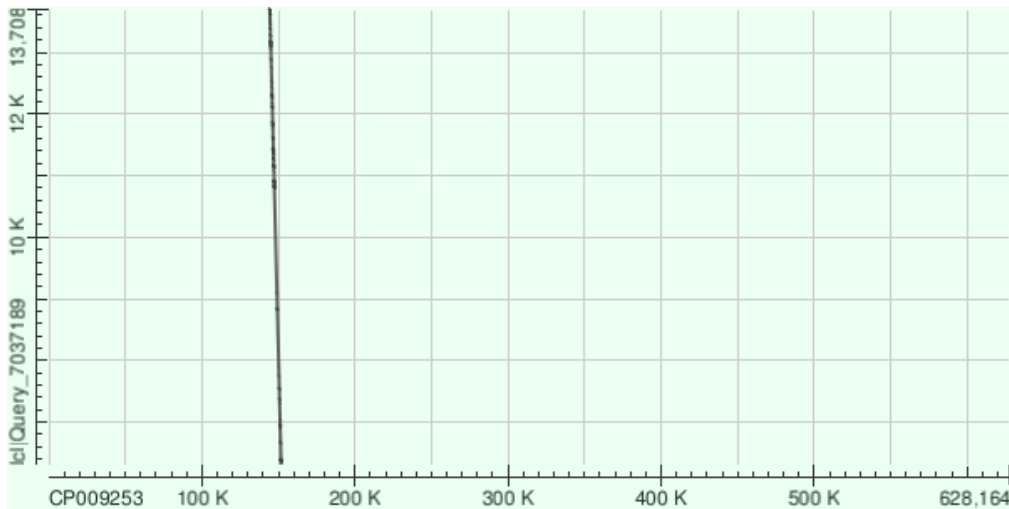
- 500370-508806, 75.6% идентичных нуклеотидов, 4%
- 510441-516539, 78.6% идентичных нуклеотидов, 3%
- 481997-488128, 74.1% идентичных нуклеотидов, 5%
- 496111-500325, 75.3% идентичных нуклеотидов, 3%
- 493487-494864, 80.1% идентичных нуклеотидов, 1%
- 480874-481548, 82.1% идентичных нуклеотидов, 2%
- 495033-495148, 90% идентичных нуклеотидов, 4%

Визуализация выдачи Hit Table (text)

```
# blastn
# Iteration: 0
# Query: CP009253.1 Buchnera aphidicola (Aphis glycines) strain BAg, complete genome
# RID: NFWV7H6M114
# 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
CP009253.1 NODE_8_length_36746_cov_20.017199 75.609 8614 1756 261 500370 508806 16876 8431 0.0 3949
CP009253.1 NODE_8_length_36746_cov_20.017199 78.569 6234 1143 147 510441 516539 6737 562 0.0 3932
CP009253.1 NODE_8_length_36746_cov_20.017199 74.074 6264 1306 249 481997 488128 35594 29517 0.0 2278
CP009253.1 NODE_8_length_36746_cov_20.017199 75.283 4325 913 123 496111 500325 21270 16992 0.0 1921
CP009253.1 NODE_8_length_36746_cov_20.017199 80.144 1385 260 15 493487 494864 24064 22688 0.0 1020
CP009253.1 NODE_8_length_36746_cov_20.017199 82.096 687 107 15 480874 481548 36747 36065 4.99e-163 573
CP009253.1 NODE_8_length_36746_cov_20.017199 90.000 120 7 4 495033 495148 22554 22436 9.96e-36 150
```

node_15_length_16745_cov_20.901762

DotPlot



15-й контиг ложится на банковскую хромосому в 1 месте :

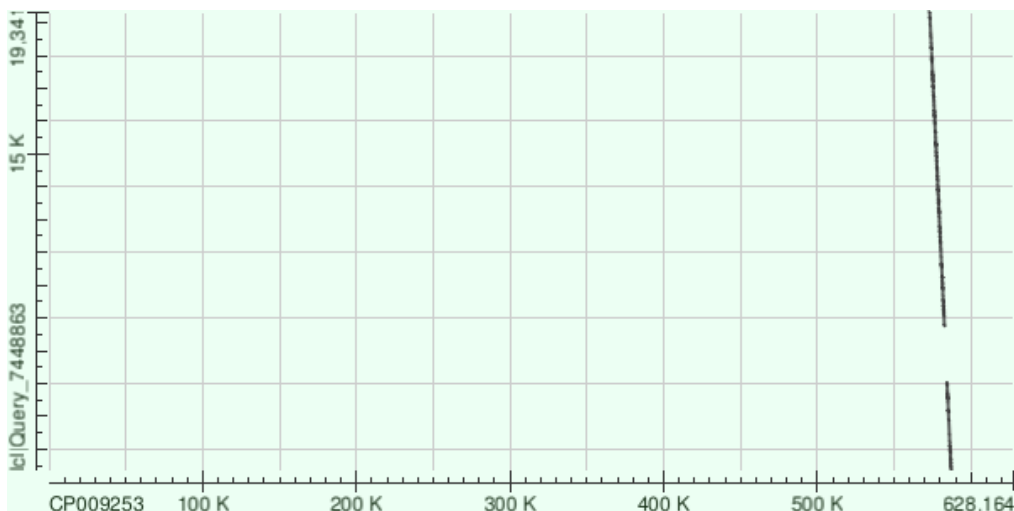
- 144368-151796, 77.8% идентичных нуклеотидов, 3% гэпов

Визуализация выдачи Hit Table (text)

```
# blastn
# Iteration: 0
# Query: CP009253.1 Buchnera aphidicola (Aphis glycines) strain BAg, complete genome
# RID: NFX2KWXR114
# 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
CP009253.1 NODE_15_length_16745_cov_20.901762 77.806 7538 1426 181 144368 151796 13708 6309 0.0 4423
```

node_57_length_19371_cov_20

DotPlot



57-й контиг ложится на банковскую хромосому в 2 местах :

- 573092-582689, 73.4% идентичных нуклеотидов, 4% гэпов
- 584329-587055, 75.6% идентичных нуклеотидов, 3% гэпов

Визуализация выдачи Hit Table (text)

```

# blastn
# Iteration: 0
# Query: CP009253.1 Buchnera aphidicola (Aphis glycines) strain BAg, complete genome
# RID: NFXARUCW114
# Database: n/a
# Fields: query acc.ver, subject acc.ver, % identity, alignment length, mismatches, gap opens, q. start, q. end, s. start, s. end, evaluate, bit score
# 2 hits found
CP009253.1    NODE_57_length_19371_cov_20.546642    73.428 9826    2148    363    573092 582689 19341 9751    0.0    3253
CP009253.1    NODE_57_length_19371_cov_20.546642    75.630 2778    567     88    584329 587055 8066 5348    0.0    1279

```