Практикум 15

Подготовка чтений программой trimmomatic

Я сделала файл adapters.fa, в котором собраны последовательности всех адаптеров. Архив с чтениями был получен командой `wget`.

Далее я удалила возможные остатки адаптеров:

```
TrimmomaticSE –threads 15 –phred33 SRR4240380.fastq.gz
SRR4240380_trimmed.fastq ILLUMINACLIP:adapters.fa:2:7:7

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

Как можно заметить, 1.88% последовательностей чтений оказались остатками адаптеров.

После этого с правых концов чтений были удалены нуклеотиды с качеством ниже 20; я оставила только такие чтения, длина которых не меньше 32 нуклеотидов:

```
TrimmomaticSE –threads 15 –phred33
SRR4240380_trimmed.fastq SRR4240380_final.fastq
TRAILING:20 MINLEN:32

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

Было удалено 253785 чтений (4,96%).
Файл до финальной очистки: 516M
Файл после финальной очистки: 490M

Программа `velveth`

```
velveth ./res 31 –short –fastq SRR4240380_final.fastq
```

На основе моего файла (SRR4240380_final.fastq) команда подготавливает k-меры длины k=31 (максимально возможной при нашей длине чтений).

./res - папка, в которой будут сохранены результаты, 31 - длина k-мера, -short - рассматриваются только короткие и не парные чтения.

```
STDOUT:
4865359 sequences in total.
```

Программа `velvetg`

Теперь сделаем сборку на основе созданных k-меров:

```
velvetg res
```

res- папка с результатами работы программы velteth

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

Таким образом, **N50** = 12042

Самые длинные контиги (из stats.txt):

ID	Длина	Покрытие
3	25915	27.418676
20	23850	24.763816
23	23807	25.725921

Можно найти контиги с **аномально большим** (700650 (ID: 84), 541 (ID: 131), 3856 (ID: 159)) и **аномально маленьким** (1 (ID:393, 304, 397)) **покрытием, все они имеют длину 1.** Однако, для примера, рассмотрю два других контига:

ID	Длина	short1_cov	short1_Ocov
56	934	130.479657	123.157388
228	18	2.722222	2.722222

Анализ BLAST

Программой **megablast** я сравнила каждый из трёх самых длинных контигов (ID: 3, 20, 23) с хромосомой *Buchnera aphidicola* (GenBank/EMBL AC — CP009253). Ниже приведены результаты (сводная таблица и dotplot):

а) для контига 3

Участок	Идентичные нуклеотиды	Гэпы
2004 – 11103	7229/9221 (78%)	252/9221 (2%)
613658 – 620926	5850/7385 (79%)	190/7385 (2%)
621055 – 627104	4678/6170 (76%)	240/6170 (3%)

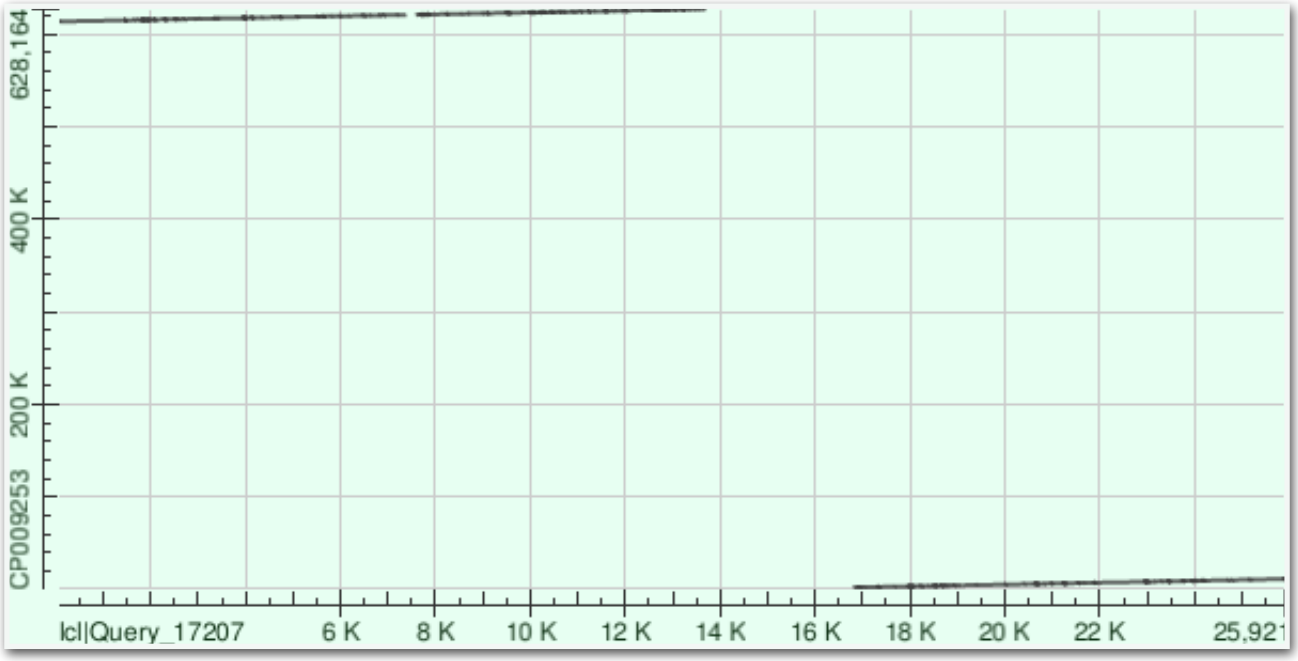

Рис. 1. Выравнивание контига 3 на хромосому *Buchnera aphidicola* (GenBank/EMBL AC - CP009253).

б) для контига 20

Участок	Идентичные нуклеотиды	Гэпы
236918 – 247596	8182/10884 (75%)	391/10884 (3%)
232358 – 236859	3466/4581 (76%)	130/4581 (2%)
229411 – 232057	2156/2685 (80%)	71/2685 (2%)
248967 – 252161	2527/3246 (78%)	94/3246 (2%)

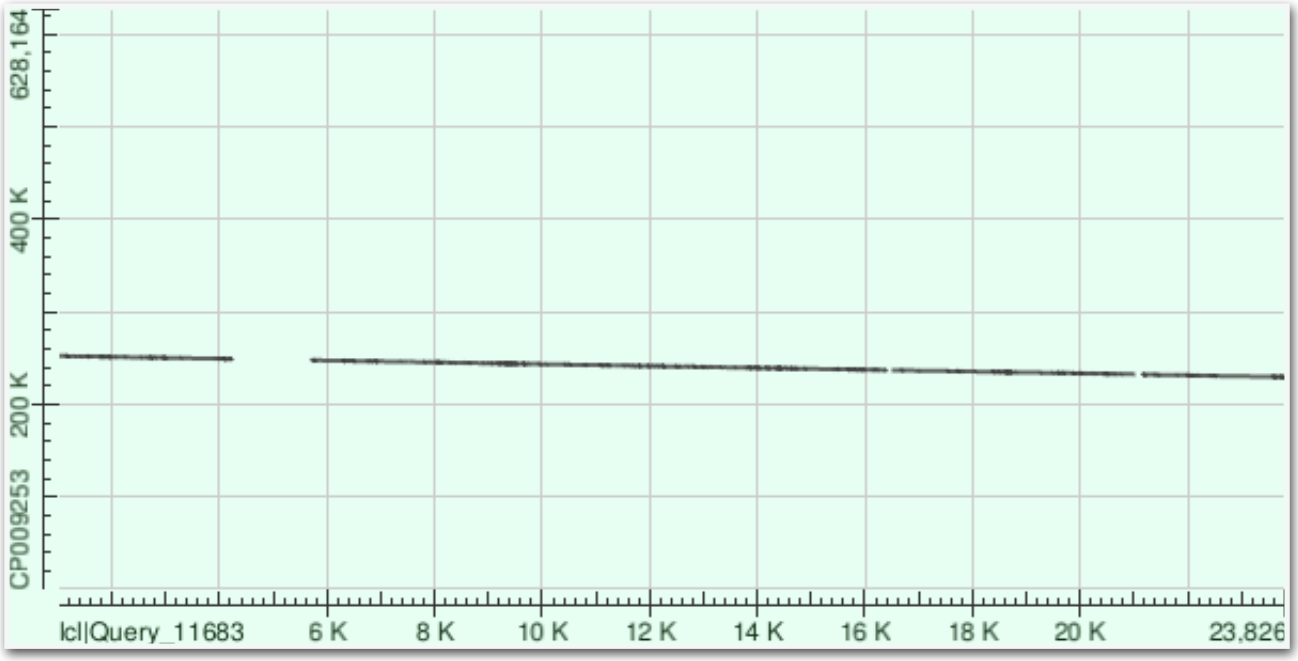

Рис. 2. Выравнивание контига 20 на хромосому *Buchnera aphidicola* (GenBank/EMBL AC - CP009253).

в) для контига 23

Участок	Идентичные нуклеотиды	Гэпы
573092 – 582686	7212/9822 (73%)	461/9822 (4%)
584329 – 587055	2100/2777 (76%)	108/2777 (3%)
593743 – 594099	289/359 (81%)	4/359 (1%)

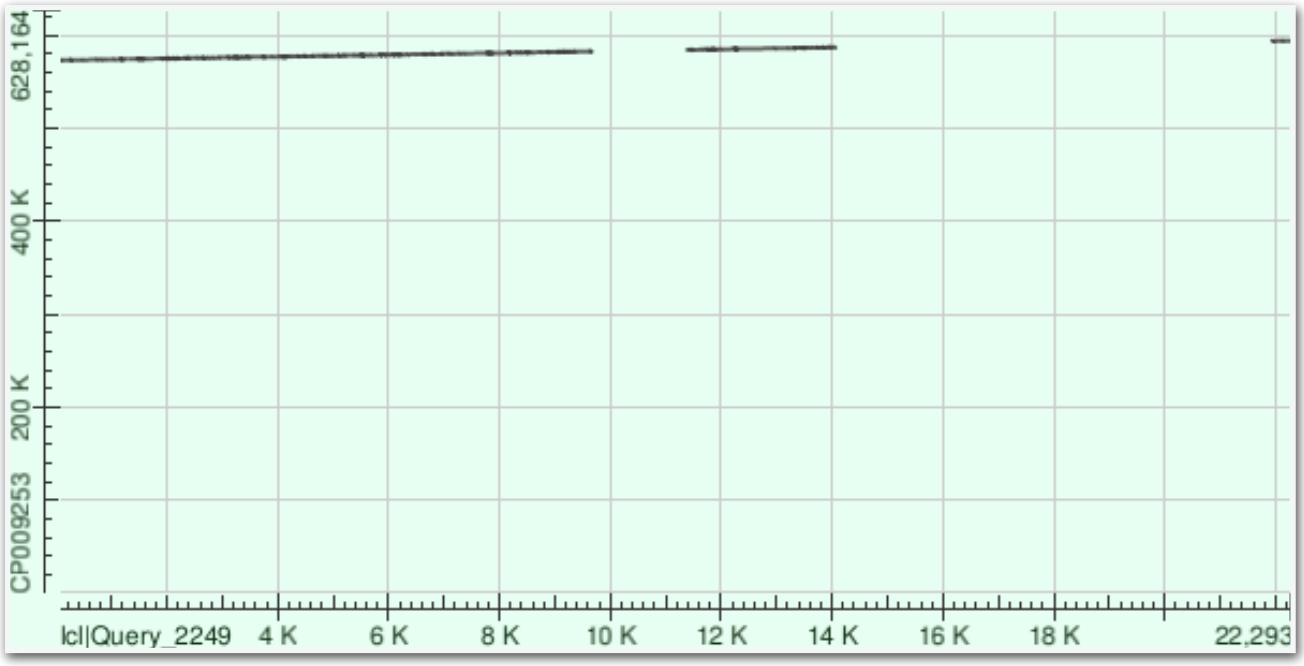

Рис. 3. Выравнивание контига 23 на хромосому *Buchnera aphidicola* (GenBank/EMBL AC - CP009253).