Отчет по практикуму №14.

Сборка de novo.

Выполнила Симоненкова Светлана.

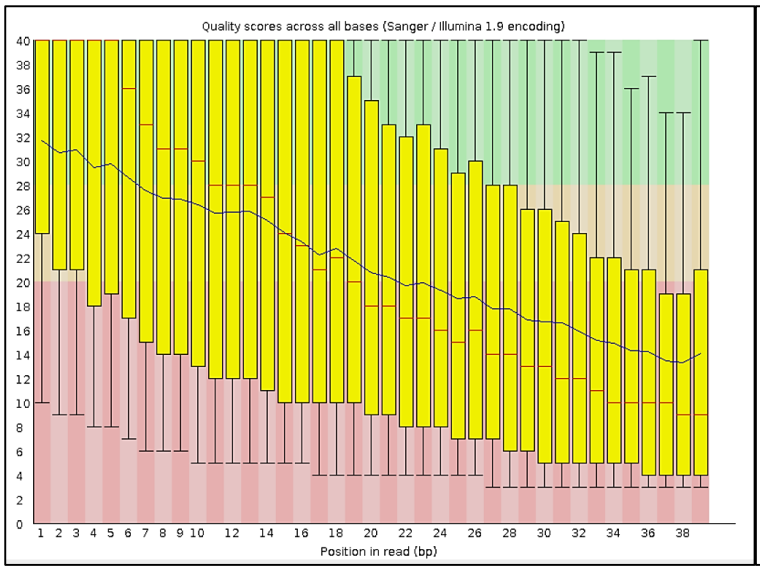
Для de novo сборки генома бактерии *Buchnera aphidicola* str. Tuc7 был предложен набор чтений с АС SRR4240358.

Качество чтений до триммирования.

Исходные данные представляют собой короткие одноконцевые чтения длины 39, полученные по технологии Illumina. С помощью программы `fastqc` оценим качество чтений до этапа подготовки.

Per base sequence quality (рис.1). Качество чтений оставляет желать лучшего: нижние децили по всем позициям достигают значения меньше Q=20 (то есть находятся в красной зоне), медиана опускается ниже «хорошего» уровня (Q=28) с 14 позиции, а среднее значение – уже с 7 позиции, и в целом качество сильно падает к концу чтения.

Per sequence quality scores (рис.2). Среднее качество большинства чтений составляет от Q=15 до Q=28. Если чтения хорошего качества, то пик на данном графике смещен вправо, здесь же наблюдается пик посередине. Это подтверждает невысокое качество чтений.



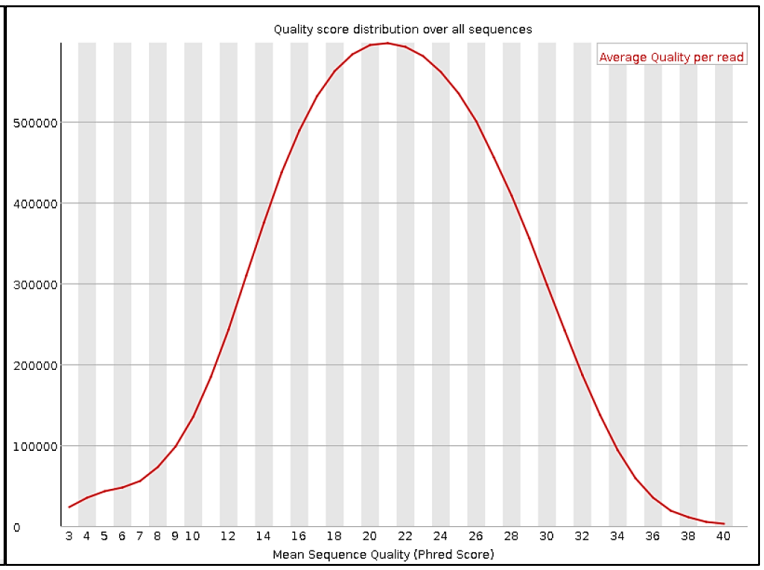


Рис.1. График качества чтений по позициям (Per base sequence quality). Рис.2. График оценки качества чтений (Per sequence quality scores).

Per base sequence content (рис.3). Также программой было отмечено неравномерное процентное содержание нуклеотидов по позициям чтений, что достаточно странно для геномной библиотеки.

Overrepresented sequences (рис.4). Последнее, на что стоит обратить внимание при рассмотрении «сырых» (насколько они сырые в базе данных…) чтений – это перепредставленные чтения. Наблюдается много (более 0,1% от общего числа) поли-А последовательностей, что, предположительно, может быть связано с загрязнением праймерами или адаптерами, хотя показатель содержание адаптеров (Adapter Content, рис.5) не говорит о таком.

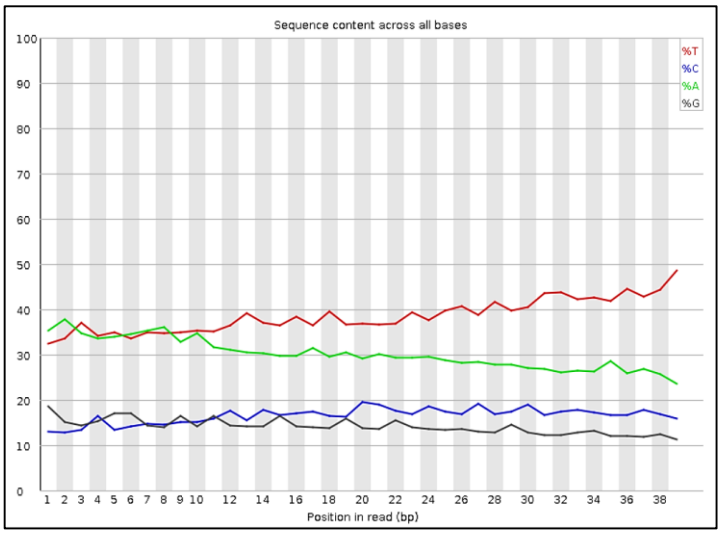


Рис.3. Содержание чтений по нуклеотидам (Per base sequence content).

Sequence	Count	Percentage	Possible Source
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	13575	0.12874817227387483	No Hit

Рис.4. Перепредставленные чтения (Overrepresented sequences).

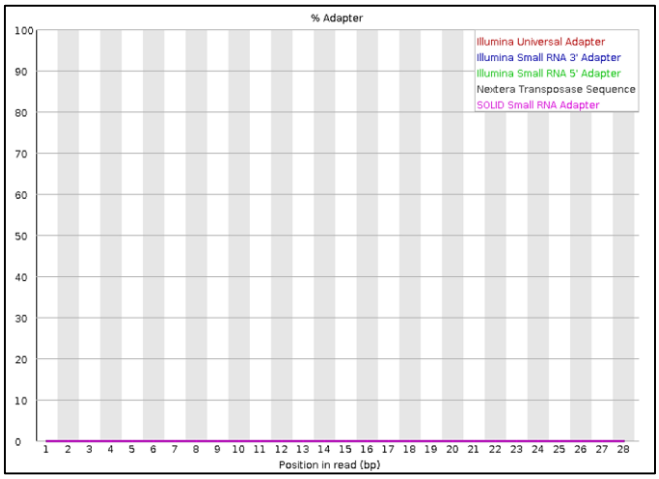


Рис.5. Содержание адаптеров (Adapter Content).

Подготовка чтений программой trimmomatic.

Чтобы удалить возможные остатки адаптеров, для начала удобно объединить все адаптеры в один файл, для чего использовалась команда:

```
cat ../../adapters/* > adapters.fa
```

Собственно, триммирование – удаление адаптеров:

```
TrimmomaticSE -threads 4 SRR4240358.fastq.gz trim_adapters.fq.gz
ILLUMINACLIP:adapters.fa:2:7:7
```

Было удалено 1,66% чтений (174955), то есть данный процент оказался остатками праймеров.

Далее чтения были подвержены фильтрации по длине (не короче 32) и удалению нуклеотидов с качеством ниже 20 с правых концов:

```
TrimmomaticSE -threads 4 trim_adapters.fq.gz trim.fq.gz TRAILING:20 MINLEN:32
```

Было удалено 22.69% чтений (2352447), программа оценила свою работу успешно. В совокупности было удалено 23,97% чтений*, размер файла с сырыми чтениями отличается от файла с фильтрованными чтениями без адаптеров на 27, 65% (с 492799624 до 356526344).

Как выдала `fastqc`, показатели, обсуждавшиеся выше, улучшились незначительно: улучшилось качество по позициям в конце чтения, пик распределения среднего качества по чтению стал едва заметно уже. Вероятно, поскольку все еще проблемными остались поли-А последовательности (хоть их и стало меньше), то и состав чтений по нуклеотидам не сильно улучшился.

* это можно проверить совместным запуском программ и подтвердить фактом, что размеры файлов изменяются одинаково в обоих вариантах запуска.

Создание набора k-меров программой velveth.

Для выше отобранных по длине чтений максимально возможная длина k-мера составляет k=31. Опция –short указывает программе, что на вход подаются короткие непарные чтения.

```
velveth kmers 31 -short -fastq trim.fq.gz
```

В результате работы программы создана папка kmers, содержащая файл Log с информацией о данном запуске программы (далее в этот файл также будет записана информация о запуске velvetg), и два файла для сборки генома программой velvetg.

Сборка генома на основе k-меров программой velvetg.

```
velvetg kmers
```

После запуска программы velvetg помимо прогресса выполнения в терминал также выводится основной результат, такой, как N50, равный в данном случае 8600 (N50=8600). *Топ контигов* характеризуется так:

ID (NODE)	*Длина*	*Покрытие (cov)*
56	19821	29,48
34	18714	29,92
40	16436	30,79

Конвейер для получения самых длинных контигов:

```
grep '^>' contigs.fa | awk -F"_" '{print $4}'| sort -nr | head -n 3
```

Конвейер для получения ID и покрытия тех же контигов:

```
grep -e '19821' -e '18714' -e '16436' contigs.fa
```

Чтобы оценить наличие контигов с аномальным покрытием для начала найдем медианное покрытие:

```
grep '^>' contigs.fa | wc –l
#195 => медиана 98
```

```
grep '^>' contigs.fa | awk -F"_" '{print $6}'| sort -nr | head -n 98 | tail -n 1
#медианное покрытие 26,95
```

Различия в значениях покрытий оценим на глаз (пятикратное различие):

```
grep '^>' contigs.fa | awk -F"_" '{print $6}'| sort -nr | less
```

Аномально большие и малые покрытия соответственно:

ID (NODE)	*Покрытие (cov)*
18	412.100006
97	405.245270
129	332.877350
103	295.135590
116	294.319031
105	290.912048
72	289.324310
49	281.516083
41	266.472076
89	264.084198
48	248.417252
151	175.265625
110	142.651520

ID (NODE)	*Покрытие (cov)*
330	5.387097
243	5.313044
307	5.150537
289	4.675676
212	4.642202
324	4.000000
283	3.956522
313	3.806452
143	3.064516
333	1.709677

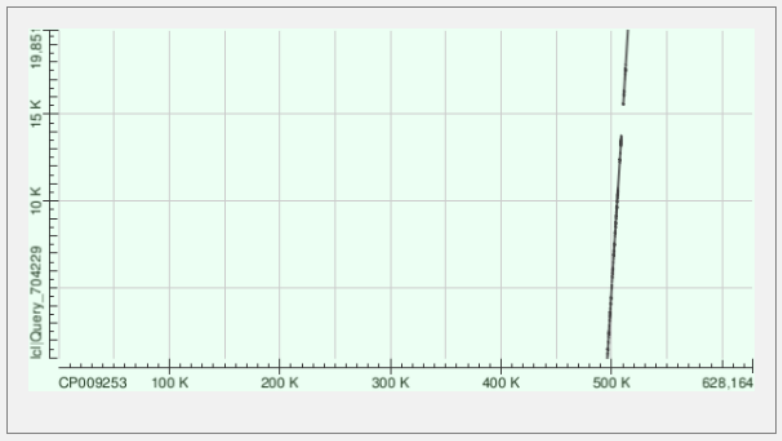
Описание некоторых контигов с аномальными покрытиями: как и предполагалось, два контига с наибольшим покрытием довольно короткие (60 и 53 соответственно). Интересно, что 7 из 10 контигов с минимальными покрытиями в длину совпадают с k-мером (то есть 31). Всего контигов с такой длиной на всю сборку 9 штук. В целом, если контиги имеют ту же длину, что и k-меры, то это свидетельствует о плохой сборке. Но, поскольку их несравнимо мало, и они имеют небольшое покрытие, все более-менее удовлетворительно.

Сравнение программой megablast самых длинных контигов с хромосомой.

В качестве референса используется хромосома *Buchnera aphidicola* (AC CP009253), ее длина составляет 628164. Сравнение представляет собой анализ выравниваний программой megablast трех самых длинных контигов на указанную хромосому, в частности описание карты локального сходства Dot Plot и таких характеристик, как совпадающие позиции и гэпы выравнивания.

На графиках снизу – референсная хромосома, сбоку – анализируемый контиг. В таблицах выравнивания располагаются в порядке уменьшения веса.

1. Выравнивание на хромосому контига 56.



Контиг выравнивается на хромосому в области 500 килобаз. Выравниваются прямые последовательности, можно видеть разрыв – негомологичный участок. Нашлось 3 участка для выравнивания, характеристики выравниваний внесены в таблицу.

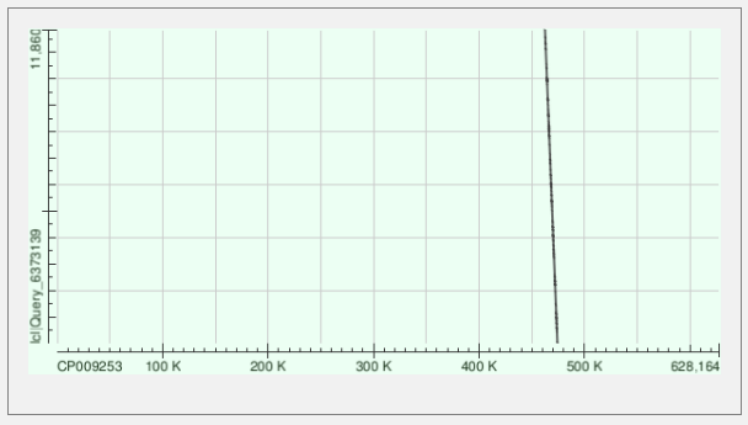
Участок референса	Совпадающие нуклеотиды	Гэпы
500370…508806	6513/8614(76%)	345/8614(4%)
510438…514772	3580/4396(81%)	83/4396(1%)
496111…500325	3257/4325(75%)	156/4325(3%)

2. Выравнивание на хромосому контига 34.

Контиг выравнивает на хромосому в области от начала последовательности до 30 килобаз. Выравниваются прямые последовательности. В данном случае наблюдается больше негомологичных участков (разрывов). Нашлось 6 участков для выравнивания, характеристики выравниваний внесены в таблицу.

Участок референса	Совпадающие нуклеотиды	Гэпы
17962…20171	1896/2220(85%)	30/2220(1%)
23067…26764	2935/3781(78%)	144/3781(3%)
14727…17919	2453/3228(76%)	92/3228(2%)
8599…11103	1982/2530(78%)	60/2530(2%)
20358…22183	1508/1850(82%)	49/1850(2%)
13994…14465	392/478(82%)	9/478(1%)

3. Выравнивание на хромосому контига 40.



Контиг выравнивается на хромосому в области 460-480 килобаз. Выравниваются прямая и обратная последовательности. Карта выглядит так, будто контиг полностью гомологичен хромосоме в данной области, так как нет разрывов. Нашлось всего 2 участка для выравнивания, характеристики выравниваний внесены в таблицу.

Участок референса	Совпадающие нуклеотиды	Гэпы
467412…474242	5344/6962(77%)	206/6962(2%)
462496…467424	3864/5019(77%)	164/5019(3%)