Все задания практикума 15 выполнялись в папке /mnt/scratch/NGS/labanovvlad/pr15

Пр. 15

Триммирование

Был скачан архив с ридами (код доступа SRR4240361):

```
mkdir reads
cd reads
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR424/001/SRR4240361/SRR4240361.fastq.gz
```

Был создан файл, в котором объединены все адаптеры из соответствующих файлов:

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

Далее были удалены возможные остатки адаптеров:

```
TrimmomaticSE -threads 16 -phred33 -trimlog trimlog.txt SRR4240361.fastq.gz
trimmed_no_adapters.fq.gz ILLUMINACLIP:adapters.fa:2:7:7
```

Во входном файле было 7272621 ридов (Input Reads), из них остатками адаптеров оказалось 34532 (0.47%) (Dropped).

Далее с правых концов ридов были удалены нуклеотиды с качеством ниже 20 и были удалены риды, длина которых меньше 32 нуклеотидов:

```
TrimmomaticSE -threads 16 -phred33 -trimlog trimlog.txt trimmed_no_adapters.fq.gz
trimmed_final.fq.gz TRAILING:20 MINLEN:32
```

Во входном файле было 7238089 ридов (`Input Reads`), из них было удалено 403754 (5.58%) (`Dropped`).

До триммирования размер файла был 193 MB, после удаления адаптеров – 192 MB, размер финального файла был 178 MB.

Сборка

С помощью программы `velveth` были подготовлены k-меры длины k=31, затем с помощью программы `velvetg` на основе этих k-меров были собраны контиги:

```
cd ..
mkdir assembly
cd assembly
velveth . 31 -short -fastq.gz ../reads/trimmed_final.fq.gz
velvetg .
```

N50 = 25683

3 самых длинных контига были получены командой:

```
sort -nk2 stats.txt | tail -3
```

Полученные контиги описаны в таблице 1.

Таблица 1. Три самых длинных контига.

ID	Длина	Покрытие
34	43866	23.514977
2	45555	26.450466
6	49238	26.660851

Для поиска контигов с аномальным покрытием все контиги были отсортированы по покрытию:

```
sort -nk6 stats.txt | less
```

Было обнаружено, что покрытие контигов более-менее равномерно нарастает с 1 до 396. Покрытия 4 последних контигов выбиваются из общей картины, эти контиги описаны в таблице 2.

Таблица 2. Контиги с аномальным покрытием.

ID	Длина	Покрытие
162	1	500
161	1	561
309	1	865
62	1	212829

Анализ

3 описанных самых длинных контига были выравнены программой megablast с хромосомой *Buchnera aphidicola* (GenBank/EMBL AC — CP009253). Дот-плоты выравниваний контигов 34, 2, 6 представлены на изображениях 1, 2, 3 соответственно.

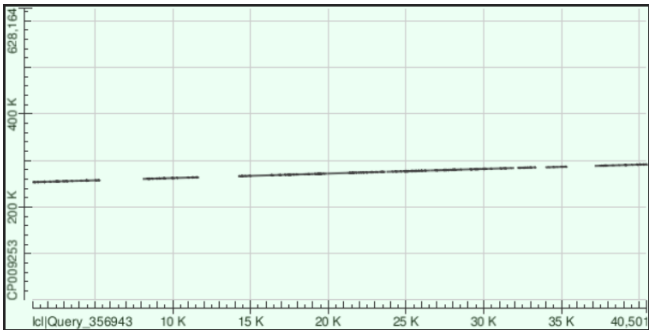

Рис. 1. Выравнивание контига 34.

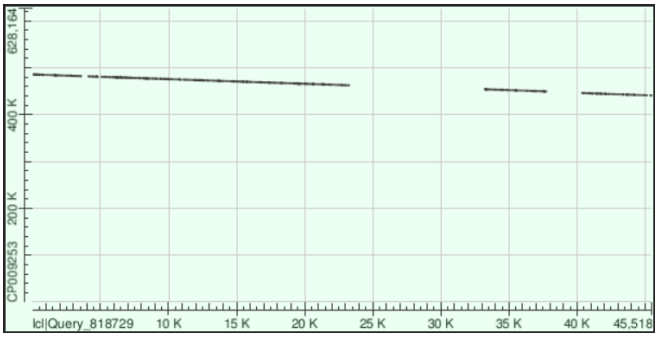

Рис. 2. Выравнивание контига 2.

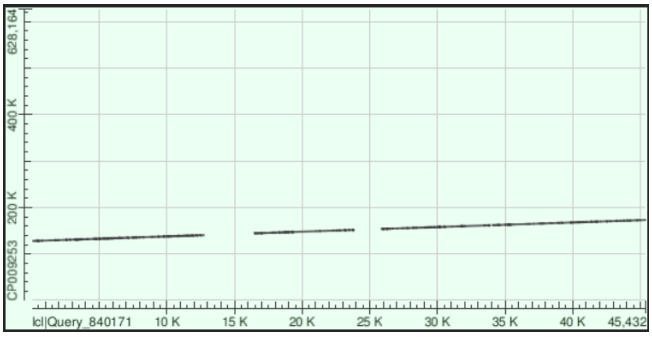

Рис. 3. Выравнивание контига 6.

Можно заметить, что все 3 контига выравнялись не полностью, видны протяженные участки замен. Также важно указать, что контиги 34 и 6 выравнялись на плюс-цепь хромосомы CP009253.1, а контиг 2 выравнялся на минус-цепь этой хромосомы.

Выравнивания контигов 34, 2, 6 подробно описаны в таблицах 3, 4, 5 соответственно.

Таблица 3. Выравнивание контига 34.

Координаты CP009253.1	Координаты контига	Identity [%]	Mismatches	Gaps
253223..257546	977..5299	76.8	981	144
260224..263784	8077..11648	79.3	728	83

266073..275551	14198..23677	81.8	1688	274
275566..283706	23736..31957	80.0	1596	323
283963..285070	32205..33314	79.5	223	37
285200..286535	34011..35345	77.5	297	25
288181..291560	37135..40501	79.8	670	73

Таблица 4. Выравнивание контига 2.

Координаты CP009253.1	**Координаты контига**	**Identity [%]**	**Mismatches**	**Gaps**
441135..442817	43540..45215	80.4	326	24
442877..445895	40383..43410	81.9	542	52
449411..454069	33159..37811	78.0	1009	112
462496..467421	18327..23268	79.6	991	135
467412..474667	10984..18297	79.3	1490	168
474844..480660	5007..10881	77.4	1295	196
480874..481545	4122..4801	84.7	102	18
481997..485679	12..3647	79.3	745	89

Таблица 5. Выравнивание контига 6.

Координаты CP009253.1	**Координаты контига**	**Identity [%]**	**Mismatches**	**Gaps**
127825..140555	50..12790	78.2	2711	430
144368..151796	16429..23828	80.3	1434	178
153752..161738	25809..33893	80.4	1549	191
161898..166752	34098..38958	81.4	891	92
166750..173180	38989..45432	78.1	1391	138