Практикум 14. Сборка de novo

В данном практикуме я работала с проектом **SRR4240359** по секвенированию бактерии *Buchnera aphidicola* str. Tuc7.

Архив с чтениями был скачан с помощью команды:
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR424/009/SRR4240359/SRR4240359.fastq.gz

С помощью программы **trimmomatic** были удалены остатки адаптеров; перед этим все последовательности адаптеров были объединены в один файл **adapters.fa**:

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

Команда для **удаления остатков адаптеров**:

```
TrimmomaticSE -phred33 SRR4240359.fastq.gz
SRR4240359_noadapt.fastq.gz ILLUMINACLIP:adapters.fa:2:7:7
-trimlog trimmomatic10.log
```

Предпоследняя строка выдачи программы была следующей:

```
Input Reads: 13557938 Surviving: 13502066 (99.59%)
Dropped: 55872 (0.41%)
```

Таким образом, было удалено 0,41% чтений.

Далее с помощью trimmomatic с правых концов чтений были ***удалены нуклеотиды с качеством ниже 20, а также удалены чтения с длиной меньше 32 нуклеотидов***:

```
TrimmomaticSE -phred33 SRR4240359_noadapt.fastq.gz
SRR4240359_trim.fastq.gz TRAILING:20 MINLEN:32 -trimlog
trimmomatic20.log
```

Предпоследняя строка выдачи программы была следующей:

```
Input Reads: 13502066 Surviving: 12184080 (90.24%)
Dropped: 1317986 (9.76%)
```

Таким образом, было удалено 9,76% чтений.

Далее с помощью программы **velveth** на основании файла SRR4240359_trim.fastq.gz были подготовлены **k-меры длины 31**:
`velveth velvh 31 -short -fastq.gz SRR4240359_trim.fastq.gz`

На основании содержимого папки velvh, полученного после работы velveth, была запущена программа **velvetg**:
`velvetg velvh`

Последняя строка выдачи программы:
`Final graph has 709 nodes and n50 of 70607, max 125674, total 682378, using 0/12184080 reads`
Из неё можно определить следующие параметры:

- N50 = 70607
- максимальная длина контига (max) = 125674

В результате работы velvetg в папке velvh появился файл stats.txt, далее был проведен его анализ. Ниже представлена информация о **трех самых длинных контигах**, полученная командой:
`sort -n -r -k 2 stats.txt | head`

Номер контига	Длина контига	Покрытие контига
11	125674	44.550949
1	108447	42.009184
14	71403	39.411551

С помощью команды `sort -n -k 6 -r stats.txt | head -n 3` получены 3 примера контигов **с самым большим покрытием**. Ими оказались контиг 111 (длина 1, покрытие 411220.0), контиг 574 (длина 1, покрытие 1395.0) и контиг 138 (длина 6, покрытие 233.0). Меня немного удивило, что контиги с малой длиной имеют такое большое покрытие. Я предполагаю, что эти контиги могут быть геномными повторами или шумом.

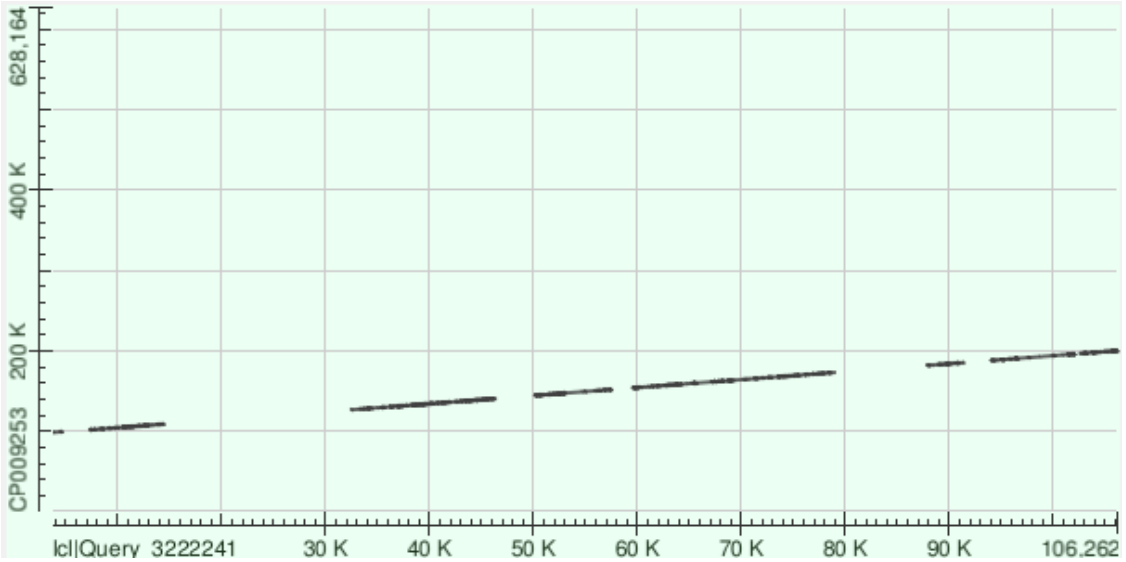
С помощью команды `sort -n -k 6 stats.txt | head -n 4` были получены 3 примера контигов **с самым малым покрытием**. Это контиги 662, 699 и 705, все с покрытием 1,0. Примечательно, что эти контиги имеют также очень малую длину 1 (т.е. 31 нуклеотид).

Последовательностей 3 самых длинных контигов нужно было извлечь в отдельные файлы. Созданная копия файла contigs.fa (contigs_tosplit.fa) была разбита на файлы с контигами:
`seqretsplit -filter contigs.fa dir/name.format`

Из них были взяты для работы три файла:
node_11_length_125674_cov_44.550949.fasta
node_1_length_108447_cov_42.009186.fasta
node_14_length_71403_cov_39.411552.fasta

Было проведено сравнение программой **megablast** каждого из трёх самых длинных контигов с хромосомой *Buchnera aphidicola* (GenBank/EMBL AC — CP009253).

- **Контиг 1**

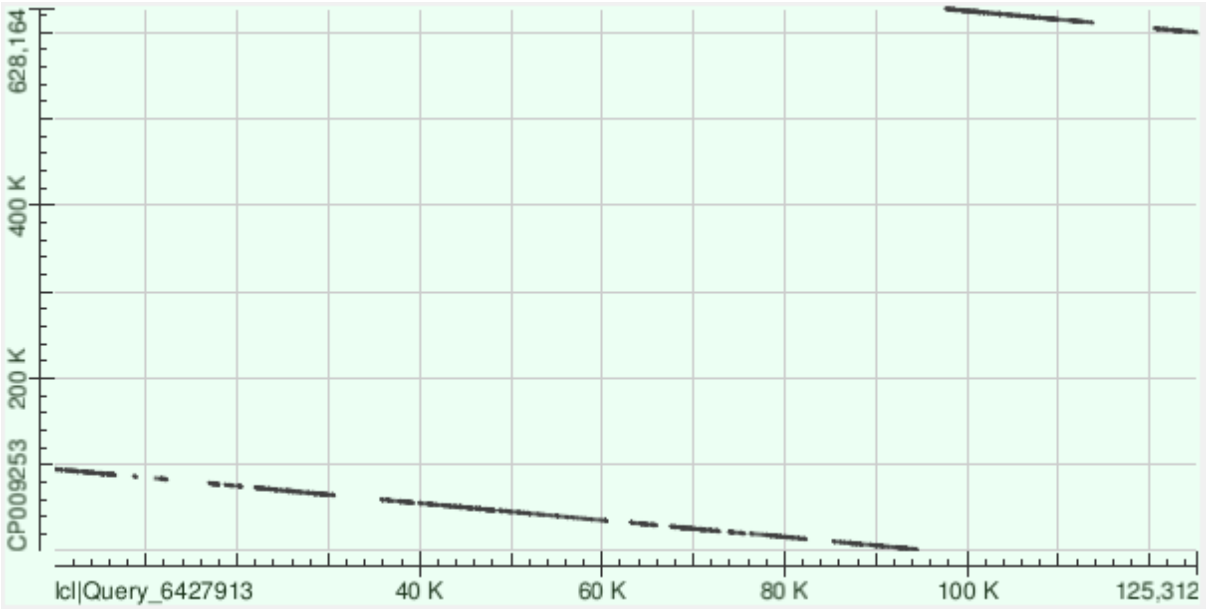


Из приведенной выше карты локального сходства, выданной программой видно, что контиг 1 довольно хорошо “ложится” на геном из GenBank, однако произошло несколько крупных делеций (выпадений участков).

Количество совпадений с референсным геномом - 15, информация об этих совпадениях представлена в таблице ниже:

Координаты участка генома	% совпадения (Identities)	Число гэпов (Gaps) и их %
127825 to 140555	9751/13010(75%)	548/13010(4%)
153752 to 161738	6355/8168(78%)	264/8168(3%)
144368 to 151796	5859/7536(78%)	243/7536(3%)
101712 to 108876	5567/7274(77%)	215/7274(2%)
187938 to 192665	3840/4801(80%)	99/4801(2%)
161898 to 166752	3911/4914(80%)	112/4914(2%)
166750 to 173180	4967/6517(76%)	159/6517(2%)
181712 to 185289	2778/3652(76%)	110/3652(3%)
194042 to 196061	1640/2070(79%)	78/2070(3%)
126623 to 127815	1004/1199(84%)	11/1199(0%)
192777 to 193984	985/1209(81%)	4/1209(0%)
196373 to 198260	1461/1910(76%)	73/1910(3%)
98408 to 99303	731/901(81%)	9/901(0%)
198467 to 199381	724/922(79%)	17/922(1%)
199545 to 200246	551/730(75%)	52/730(7%)

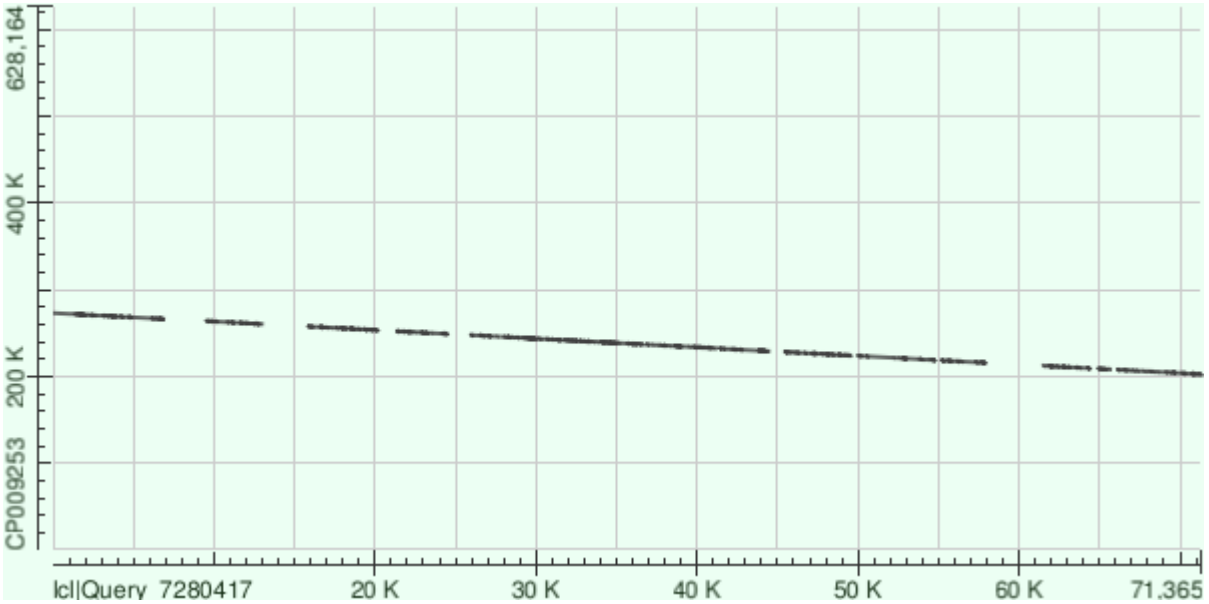
- **Контиг 11**



Из приведенной выше карты локального сходства, выданной программой видно, что контиг 1 довольно хорошо “ложится” на геном из GenBank, однако произошло несколько крупных делеций (выпадений участков). То, что прямая “разорвана” и идет вниз, говорит о том, что контиг записан в обратном направлении и точки начала записей не совпадают. Количество совпадений с референсным геномом - 25, в таблице - примеры первых трех:

Координаты участка генома	% совпадения (Identities)	Число гэпов (Gaps) и их %
35124 to 44693	7981/9633(83%)	130/9633(1%)
2004 to 11103	7229/9223(78%)	256/9223(2%)
613658 to 620926	5845/7379(79%)	184/7379(2%)

- **Контиг 14**



Из приведенной выше карты локального сходства, выданной программой видно, что контиг 1 довольно хорошо “ложится” на геном из GenBank, однако произошло несколько крупных делеций (выпадений участков). То, что прямая “разорвана” и идет вниз, говорит о том, что контиг записан в обратном направлении, однако точки начала записей совпадают.

Количество совпадений с референсным геномом - 14, информация об этих совпадениях представлена в таблице ниже:

Координаты участка генома	% совпадения (Identities)	Число гэпов (Gaps) и их %
266073 to 273028	5664/7060(80%)	197/7060(2%)
236918 to 247596	8178/10884(75%)	389/10884(3%)
202390 to 207661	4183/5329(78%)	137/5329(2%)
219625 to 223720	3342/4130(81%)	61/4130(1%)
224057 to 228137	3218/4178(77%)	163/4178(3%)
232358 to 236859	3468/4583(76%)	134/4583(2%)

228944 to 232057	2499/3165(79%)	95/3165(3%)
260224 to 263784	2788/3617(77%)	101/3617(2%)
248967 to 252161	2523/3245(78%)	92/3245(2%)
215717 to 218384	2145/2713(79%)	72/2713(2%)
209294 to 212243	2302/3007(77%)	104/3007(3%)
253223 to 257546	3245/4421(73%)	195/4421(4%)
208017 to 208904	692/902(77%)	25/902(2%)
218821 to 219491	515/676(76%)	20/676(2%)