

Сборка de novo

Код доступа проекта по секвенированию бактерии *Buchnera aphidicola* str. Tuc7 - **SRR4240359**.

Команда для скачивания чтений:

```
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR424/009/SRR4240359/SRR4240359.fastq.gz
```

1. Подготовка чтений программой trimmomatic

Удаление адаптеров

Я создала файл со всеми адаптерами в своей папке командой:

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

Для удаления возможных остатков адаптеров была использована программа TrimmomaticSE (SE так как чтения одноконцевые).

Команда:

```
TrimmomaticSE -phred33 SRR4240359.fastq.gz SRR4240359_trimm.fastq.gz  
ILLUMINACLIP:adapters.fa:2:7:7
```

Получается: на вход было дано 13 557 938 чтений, удалено 34 532 (0.41%). То есть 0.41% чтений оказались остатками адаптеров.

Обрезание чтений

Для удаления с правых концов чтений нуклеотидов с качеством ниже 20, а также чтений, длина которых меньше 32 нуклеотидов была использована команда:

```
TrimmomaticSE -phred33 SRR4240359_trimm.fastq.gz SRR4240359_trimm2.fastq.gz  
TRAILING:20 MINLEN:32>trimm2.txt
```

На вход было дано 13 502 066 чтений, удалено 1 317 986 (9.76%).

Размеры файлов:

SRR4240361.fastq.gz (скачанный файл) – 465 Мб

Чтения без адаптеров – 463 Мб

Чтения после очистки – 403 Мб

2. Подготовка k-меров

Для получения k-меров была использована следующая команда:

```
velveth k_mers 31 -short -fastq.gz SRR4240359_trimm2.fastq.gz
```

31 - длина k-меров

-short - короткие и непарные чтения

-fastq.gz - формат входного файла

k_mers - папка, в которой окажутся получаемые файлы

3. Сборка de novo

Для сборки de novo из получившихся k-меров была использована программа velvetg.

Команда:

```
velvetg k_mers
```

Из вывода программы в stdout:

N50 = 70607

Чтобы найти длины трёх самых длинных контигов, информация о которых содержится в файле contigs.fa я использовала команду:

```
cat contigs.fa | grep '>' | tr '_' '\t' | sort -k4 -n | tail -3
```

Таким образом, самые длинные контиги **11**, **1** и **14** (см. Табл. 1).

Таблица 1. Характеристика трех самых длинных контигов из сборки.

ID	Длина	Покрытие
11	125 674	44.550949
1	108 447	42.009184
14	71 403	39.408694

Чтобы найти контиги с аномально большим или аномально малым покрытием была использована команда:

```
cat contigs.fa | grep '>' | tr '_' '\t' | sort -k6 -n
```

Контиги с аномально большим покрытием:

ID:**98** – покрытие 139.489365, длина 47

ID:**80** – покрытие 109.500000, длина 40

4. Анализ

Три самых больших контига были картированы на хромосому *Buchnera aphidicola* (AC — CP009253).

Ниже представлены характеристики и карты локального сходства этих выравниваний (см. Табл. 2-4, Рис. 1-3). Во всех DotPlot по оси y расположена последовательность хромосомы *Buchnera aphidicola* (AC - CP009253).

Можно заметить, что выравнивания построились с другой (комплементарной) цепью бактерии относительно контигов.

Контиг 11

megablast выделил 25 гомологичных участков.

Таблица 2. Характеристики выравнивания контига 11 на хромосому *Buchnera aphidicola* (GenBank/EMBL AC — CP009253).

Номер	Координаты участка контига	Координаты участка хромосомы	Длина выравнивания	% идентичности	Число гэпов
1	50907- 60472	44693-35124	9633	82.851	130/9633(1%)
2	85411-94500	11103-2004	9223	78.380	256/9223(2%)
3	103943-111247	620926-613658	7379	79.211	184/7379(2%)
4	39954-48261	55420-47158	8436	76.339	301/8436(3%)
5	24545-30582	70621-64632	6151	76.459	274/6151(4%)
6	120355-125312	604795-599832	5046	78.201	170/5046(3%)
7	97666-103713	627104-621055	6173	75.782	248/6173(4%)
8	67532-72881	28363-23067	5433	76.551	219/5433(4%)
9	1140-6626	93683-88200	5607	75.317	243/5607(4%)
10	76309-78519	20182-17962	2231	85.253	30/2231(1%)
11	35876-39267	59462-56071	3453	78.685	122/3453(3%)
12	78597-81767	17919-14727	3226	75.976	88/3226(2%)
13	63188-65924	32745-30013	2777	77.422	84/2777(3%)
14	73906-75730	22183-20358	1851	81.524	51/1851(2%)
15	48746-50760	46776-44768	2044	79.207	64/2044(3%)
16	111356-113422	613671-611633	2086	77.900	66/2086(3%)
17	21982-24360	73310-70970	2411	73.579	102/2411(4%)
18	10863-12259	84409-83021	1409	77.076	32/1409(2%)
19	2-877	94696-93821	885	79.887	18/885(2%)
20	18742-19686	76468-75528	953	77.230	20/953(2%)
21	16970-18128	78277-77117	1182	73.519	44/1182(3%)
22	81927-82401	14465-13994	478	82.218	9/478(1%)
23	20027-20445	75264-74833	442	76.923	33/442(7%)
24	113554-113849	611524-611229	297	79.461	2/297(0%)
25	8647-8878	86404-86174	234	78.632	5/234(2%)

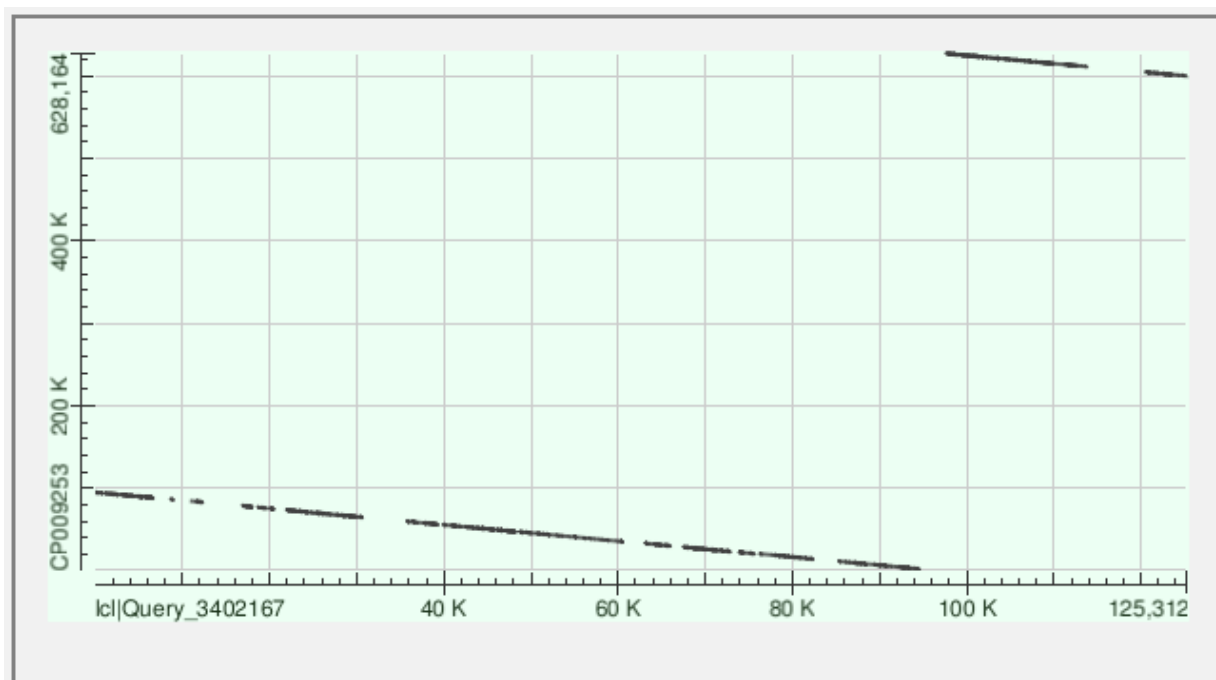


Рис. 1. Dot Plot выравнивания контига 11 на хромосому *Buchnera aphidicola* (GenBank/EMBL AC — CP009253).

Есть 2 линии выравнивая в разных частях Dot Plot, следовательно начала последовательностей не совпали. Также заметно несколько достаточно больших фрагментов, где не было обнаружено значительного сходства (например на участках 31K-37K, 112K-120K).

Контиг 1

megablast выделил 15 гомологичных участков.

Таблица 3. Характеристики выравнивания контига 1 на хромосому *Buchnera aphidicola* (GenBank/EMBL AC — CP009253).

Номер	Координаты участка контига	Координаты участка хромосомы	Длина выравнивания	% идентичности	Число гэпов
1	33726- 46466	127825- 140555	13010	74.950	548/13010(4%)
2	59485-67569	153752-161738	8168	77.804	264/8168(3%)
3	50105-57504	144368-151796	7536	77.747	243/7536(3%)
4	7333-14500	101712-108876	7274	76.533	215/7274(2%)
5	94019-98793	187938-192665	4801	79.983	99/4801(2%)
6	67774-72634	161898-166752	4914	79.589	112/4914(2%)
7	72665-79108	166750-173180	6517	76.216	159/6517(2%)
8	87826-91441	181712-185289	3652	76.068	110/3652(3%)
9	100101-102142	194042-196061	2070	79.227	78/2070(3%)
10	32468-33661	126623-127815	1199	83.736	11/1199(0%)
11	98869-100074	192777-193984	1209	81.472	4/1209(0%)
12	102449-104307	196373-198260	1910	76.492	73/1910(3%)
13	3756-4652	98408-99303	901	81.132	9/901(0%)
14	104513-105424	198467-199381	922	78.525	17/922(1%)
15	105557-106262	199545-200246	730	75.479	52/730(7%)

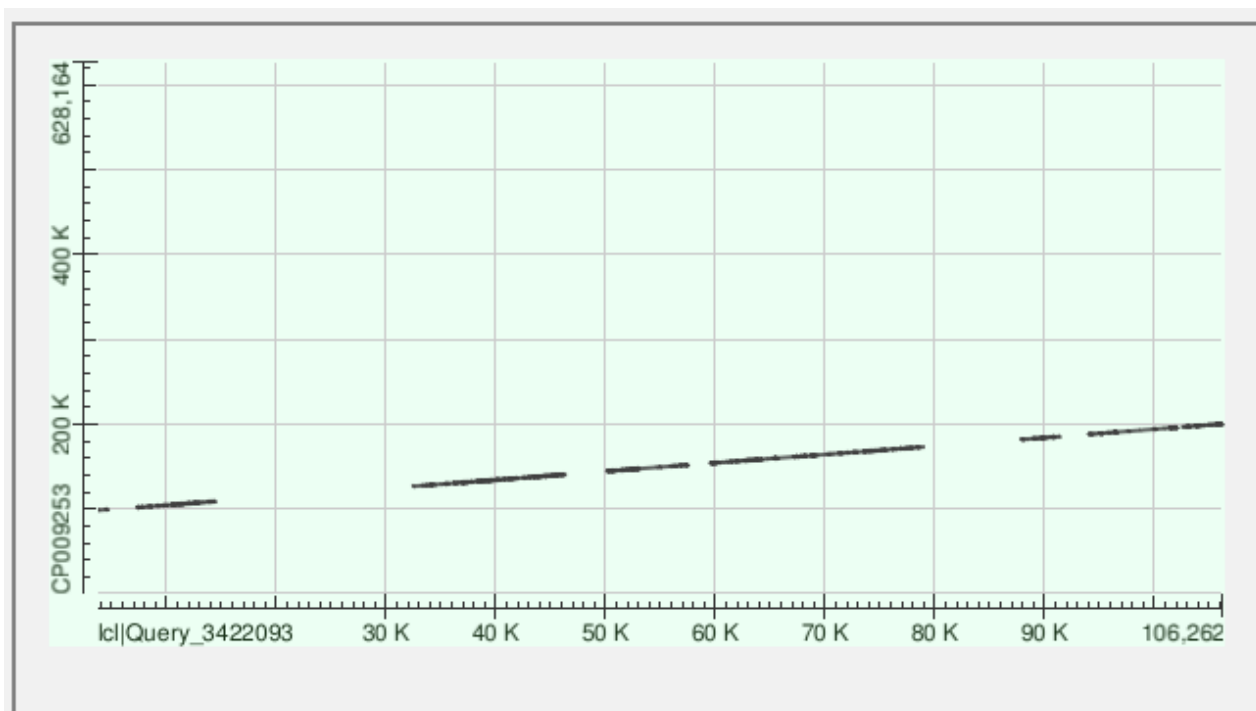


Рис. 2. Dot Plot выравнивания контига 1 на хромосому *Buchnera aphidicola* (GenBank/EMBL AC — CP009253).

В сравнении с контигом 11 здесь в среднем меньше процент идентичности; также есть большие невыровненные фрагменты – самый длинный 15K-32K.

Контиг 14

megablast выделил 14 гомологичных участков.

Номер	Координаты участка контига	Координаты участка хромосомы	Длина выравнивания	% идентичности	Число гэпов
1	1- 6967	273028- 266073	7060	80.227	197/7060(2%)
2	25958-36657	247596-236918	10884	75.138	389/10884(3%)
3	66117-71365	207661-202390	5329	78.495	137/5329(2%)
4	49953-54055	223720-219625	4130	80.920	61/4130(1%)
5	45453-49564	228137-224057	4178	77.022	163/4178(3%)
6	36755-41284	236859-232358	4583	75.671	134/4583(2%)
7	41433-44553	232057-228944	3165	78.957	95/3165(3%)
8	9517-13088	263784-260224	3617	77.080	101/3617(2%)
9	21279-24481	252161-248967	3245	77.750	92/3245(2%)
10	55202-57887	218384-215717	2713	79.064	72/2713(2%)
11	61463-64422	212243-209294	3007	76.555	104/3007(3%)
12	15866-20188	257546-253223	4421	73.400	195/4421(4%)
13	64868-65758	208904-208017	902	76.718	25/902(2%)
14	54103-54763	219491-218821	676	76.183	20/676(2%)

Таблица 4. Характеристики выравнивания контига 14 на хромосому *Buchnera aphidicola* (GenBank/EMBL AC — CP009253).

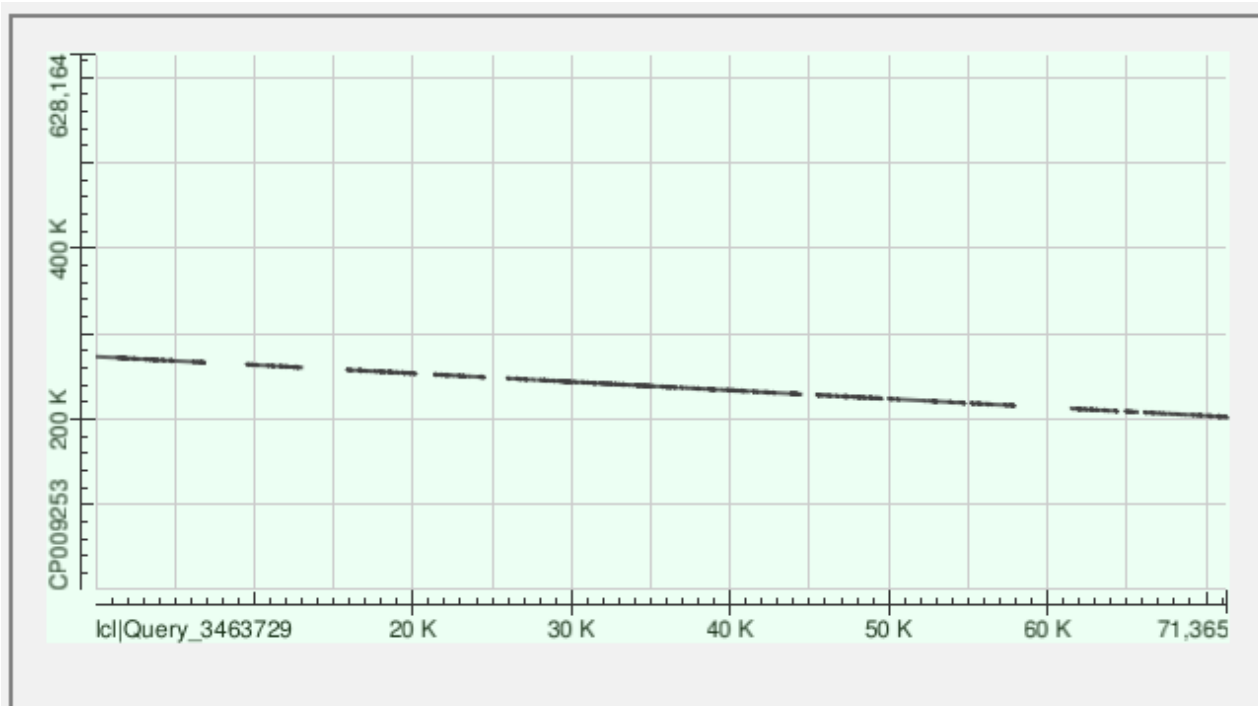


Рис. 3. Dot Plot выравнивания контига 14 на хромосому *Buchnera aphidicola* (GenBank/EMBL AC — CP009253).

В выравнивании с контигом 14 участки, где не было обнаружено сходства уже несколько меньше, хотя все равно заметно их относительно большое количество.