Практикум 14

Выполнила: *Мельникова Ксения (kmelnikova)*

Сборка de novo

Код доступа проекта по секвенированию бактерии *Buchnera aphidicola* str. Tuc7: **SRR4240359**
Скачала чтения в директорию практикума с помощью команды:
`wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR424/009/SRR4240359/SRR4240359.fastq.gz`

I) Подготовка чтений программой trimmomatic
В начале я подготовила файл, в котором объединила все адаптеры:
`cat /mnt/scratch/NGS/adapters/* > adapters.fasta`

Запустила **trimmomatic**, чтобы удалить адаптеры:
```
TrimmomaticSE -phred33 SRR4240359.fastq.gz SRR4240359noadapters.fq.gz
ILLUMINACLIP:adapters.fasta:2:7:7
```

`-phred33` – указывает, что входные файлы используют кодировку качества Phred33
`SRR4240359.fastq.gz` – входной файл с чтениями
`SRR4240359noadapters.fq.gz` – чтения без адаптеров

0.41% последовательностей чтений оказалось остатками адаптеров

После этого я удалила чтения, у которых качество в конце было меньше 20 и длина которых меньше 32 нуклеотидов:
```
TrimmomaticSE -phred33 SRR4240359noadapters.fq.gz SRR4240359_trimm.fq.gz
TRAILING:20 MINLEN:32
```

После триммирования удалено **1317986 (9.76%)** чтений
Размер файла с изначальными чтениями: **465903280 байтов**
Размер файла после очистки: **403228835 байтов**

II) Подготовка k-меров
Алгоритмы, использующие граф де Брёйна, сначала составляют список k-меров, встретившихся в чтениях. Поэтому для сборки de novo нам необходимо их подготовить. Для этого нужна программа **velveth**.

Запустила **velveth** с заданными в задании параметрами:
`velveth kmers 31 -short -fastq.gz SRR4240359_trimm.fq.gz`

`kmers` – директория, в которую попадут k-меры
`31` – длина k-меров
`-short` – параметр для коротких и непарных чтений
`-fastq.gz` – формат входного файла
`SRR4240359_trimm.fq.gz` – входной файл (файл после очистки на предыдущем шаге)

III) Сборка на основе k-меров

Для сборки de novo из получившихся k-меров, я запустила программу **velvetg**:
`velvetg kmers`

`kmers` – директория с k-мерами

N50 = 70607

Из файла contigs.fa узнала три самых длинных контига и их покрытие:
`less contigs.fa | grep '>'| tr '_' '\t'| sort -k4 -n`

Контиг 14) Длина: 71403 нуклеотидов, покрытие: 39.411552
Контиг 1) Длина: 108447 нуклеотидов, покрытие: 42.009186
Контиг 11) Длина: 125674 нуклеотидов, покрытие: 44.550949

Контиги с аномально большим покрытием:
Контиг 80) Длина: 40 нуклеотидов, покрытие: 109.500000
Контиг 98) Длина: 47 нуклеотидов, покрытие: 139.489365
Контиг 126) Длина: 51 нуклеотидов, покрытие: 91.982140

IV) Анализ

Сравнила программой megablast каждый из трёх самых длинных контигов с хромосомой *Buchnera aphidicola* (GenBank/EMBL AC — CP009253). Все контиги “легли” на геном с разрывами, может они плохо собрались или бактерия (чтения которой мы взяли) имеет отличия от референсного генома бактерии этого вида.

*Таблица 1. Характеристика выравнивания **контигов 1** на хромосому Buchnera aphidicola (GenBank/EMBL AC — CP009253)*

Range	Start - Stop	Identities	Gaps
1	127825 - 140555	9751/13010(75%)	548/13010(4%)
2	153752 - 161738	6355/8168(78%)	264/8168(3%)
3	144368 - 151796	5859/7536(78%)	243/7536(3%)
4	101712 - 108876	5567/7274(77%)	215/7274(2%)
5	187938 - 192665	3840/4801(80%)	99/4801(2%)
6	161898 - 166752	3911/4914(80%)	112/4914(2%)

7	166750 - 173180	4967/6517(76%)	159/6517(2%)
8	181712 - 185289	2778/3652(76%)	110/3652(3%)
9	194042 - 196061	1640/2070(79%)	78/2070(3%)
10	126623 - 127815	1004/1199(84%)	11/1199(0%)
11	192777 - 193984	985/1209(81%)	4/1209(0%)
12	196373 - 198260	1461/1910(76%)	73/1910(3%)
13	98408 - 99303	731/901(81%)	9/901(0%)
14	198467 - 199381	724/922(79%)	17/922(1%)
15	199545 - 200246	551/730(75%)	52/730(7%)

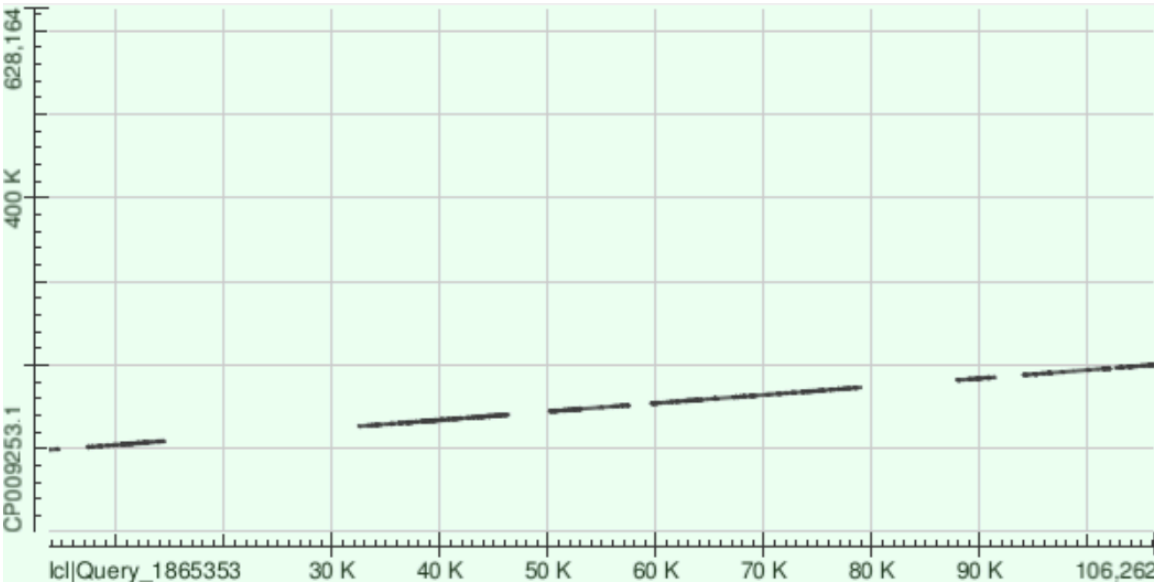

Рис. 1. Dot plot выравнивания контига 1 на хромосому Buchnera aphidicola

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

Range	Start - Stop	Identities	Gaps
1	35124 - 44693	7981/9633(83%)	130/9633(1%)
2	2004 - 11103	7229/9223(78%)	256/9223(2%)
3	613658 - 620926	5845/7379(79%)	184/7379(2%)
4	47158 - 55420	6440/8436(76%)	301/8436(3%)
5	64632 - 70621	4703/6151(76%)	274/6151(4%)

6	599832 - 604795	3946/5046(78%)	170/5046(3%)
7	621055 - 627104	4678/6173(76%)	248/6173(4%)
8	23067 - 28363	4159/5433(77%)	219/5433(4%)
9	88200 - 93683	4223/5607(75%)	243/5607(4%)
10	17962 - 20182	1902/2231(85%)	30/2231(1%)
11	56071 - 59462	2717/3453(79%)	122/3453(3%)
12	14727 - 17919	2451/3226(76%)	88/3226(2%)
13	30013 - 32745	2150/2777(77%)	84/2777(3%)
14	20358 - 22183	1509/1851(82%)	51/1851(2%)
15	44768 - 46776	1619/2044(79%)	64/2044(3%)
16	611633 - 613671	1625/2086(78%)	66/2086(3%)
17	70970 - 73310	1774/2411(74%)	102/2411(4%)
18	83021 - 84409	1086/1409(77%)	32/1409(2%)
19	93821 - 94696	707/885(80%)	18/885(2%)
20	75528 - 76468	736/953(77%)	20/953(2%)
21	77117 - 78277	869/1182(74%)	44/1182(3%)
22	13994 - 14465	393/478(82%)	9/478(1%)
23	74833 - 75264	340/442(77%)	33/442(7%)

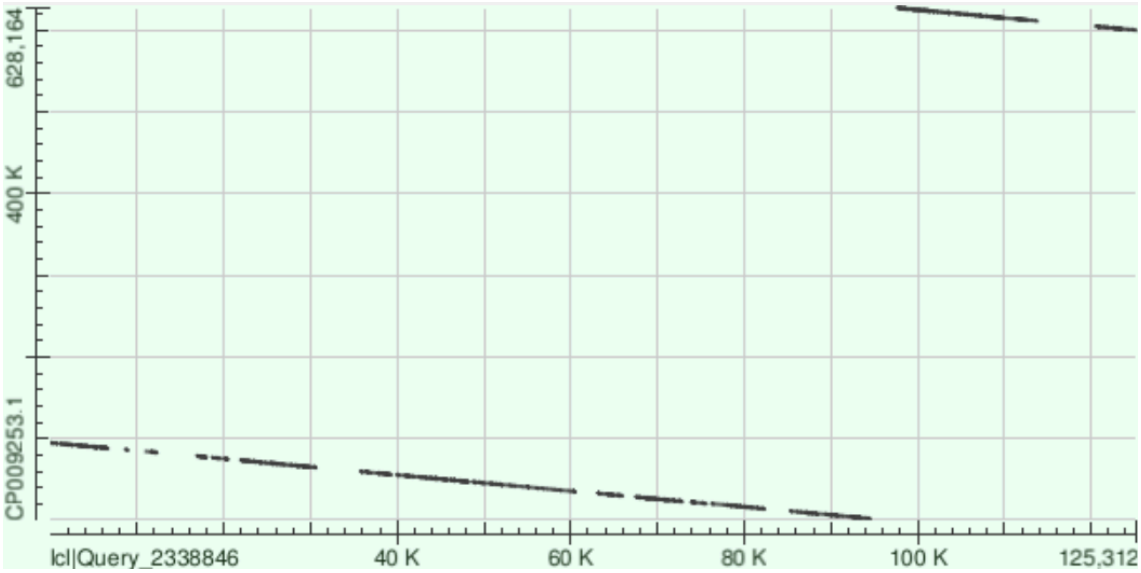

Рис. 1. Dot plot выравнивания контига 11 на хромосому Buchnera aphidicola

*Таблица 3. Характеристика выравнивания **контига 14** на хромосому Buchnera aphidicola (GenBank/EMBL AC — CP009253)*

Range	Start - Stop	Identities	Gaps
1	266073 - 273028	5664/7060(80%)	197/7060(2%)
2	236918 - 247596	8178/10884(75%)	389/10884(3%)
3	202390 - 207661	4183/5329(78%)	137/5329(2%)
4	219625 - 223720	3342/4130(81%)	61/4130(1%)
5	224057 - 228137	3218/4178(77%)	163/4178(3%)
6	232358 - 236859	3468/4583(76%)	134/4583(2%)
7	228944 - 232057	2499/3165(79%)	95/3165(3%)
8	260224 - 263784	2788/3617(77%)	101/3617(2%)
9	248967 - 252161	2523/3245(78%)	92/3245(2%)
10	215717- 218384	2145/2713(79%)	72/2713(2%)
11	209294 - 212243	2302/3007(77%)	104/3007(3%)
12	253223 - 257546	3245/4421(73%)	195/4421(4%)
13	208017 - 208904	692/902(77%)	25/902(2%)
14	218821 - 219491	515/676(76%)	20/676(2%)

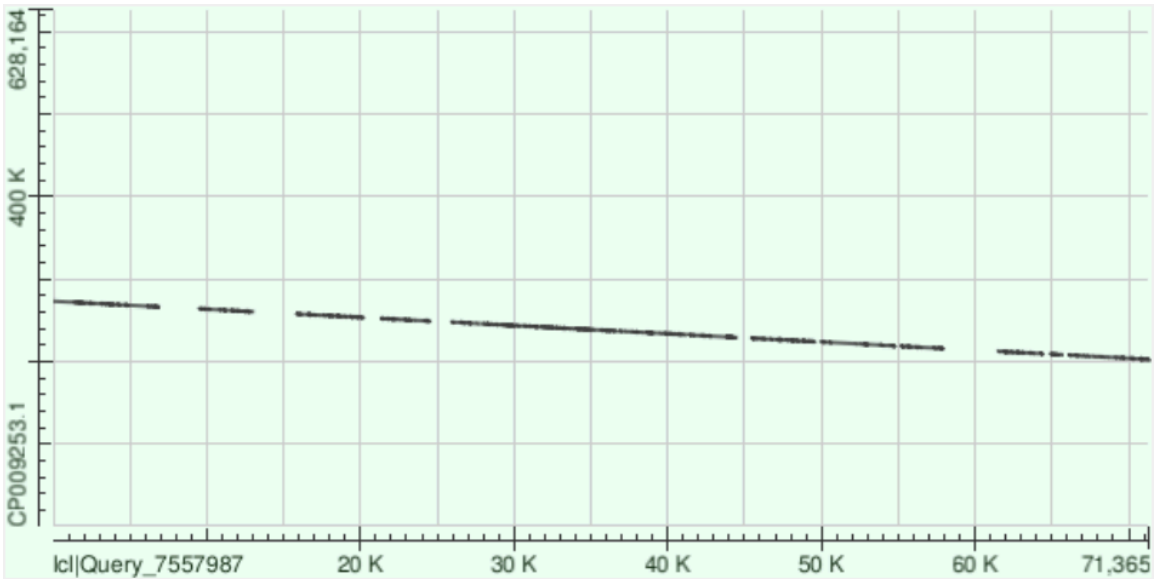

Рис. 1. Dot plot выравнивания контига 14 на хромосому Buchnera aphidicola