Практикум 15
Сборка генома de novo

1. **Получение сборки**
   Для скачивания была использована команда, исполняемая в папке /mnt/scratch/NGS/salimakri/pr15 :

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

2. **Подготовка программой Trimmomatic**
   a. **Удаление адаптеров**
      Для начала, были получены файлы с последовательностями адаптеров из общей папки

```
cp ../../adapters/*.fa .
```

      После этого последовательности были объединены в один файл

```
cat *.fa > all_adapters.fa
```

      Для триммирования была использована команда TrimmomaricSE, так как прочтения здесь одноконцевые.

```
TrimmomaticSE -phred33 -threads 16 SRR4240359.fastq.gz
non_adapters.fastq.qz ILLUMINACLIP:all_adapters.fa:2:7:7
```

      В результате работы TrimmomaricSE получен файл с очищенными последовательностями non_adapters.fastq.qz. Кроме того, была получена информация о количестве чтений поданных на вход и количестве оставшихся после триммирования:
      - Подано на вход: 13557938
      - Исключено: 55872 (0.41%)
      - Сохранено с изменениями: 13502066 (99.59%)

   b. **Фильтрация по качеству и длине**
      Так же с помощью Trimmomatic были удалены концевые нуклеотиды чтений, качество которых ниже 20 и чтения длины меньше 32

```
TrimmomaticSE -phred33 -threads 16 non_adapters.fastq.qz
filtered.fastq.qz TRAILING:20 MINLEN:32
```

      После триммирования с заданными параметрами получилось чтений:
      - Подано на вход: 13502066
      - Исключено: 1317986 (9,76%)
      - Сохранено с изменениями: 12184080 (90,24%)

3. **Использование программы velveth**
   Подготовка k-меров
   Командой была задана длина k-меров = 31

```
velveth velveth 31 -fastq -short filtered.fastq.qz
```

   Создана директория с файлами : Log, Roadmaps, Sequences.

4. **Использование velvetg**
   Для сборки на получившихся k-меров была использована программа velvetg

```
velvetg velveth
```

   В результате были созданы еще несколько файлов сборки и файлы ее описывающие.

   N50 = 70607

   a. **Поиск самых длинных контигов**
      Для поиска трех самых длинных контигов и их покрытия был использован файл contigs.fa, а именно аннотация контигов

```
grep '^>' contigs.fa | tr '_' '\t' | sort -k 3,3 -r -n | head
-n 3| less
```

      Получена следующая информация

Номер контига	Длина	Покрытие
11	125674	44.550949
1	108447	42.009186
14	71403	39.411552

      Также были найдены контиги с аномально большим покрытием:

Номер контига	Длина	Покрытие
98	47	139.489365
80	40	109.500000

5. **Сравнение самых длинных контигов с хромосомой *Buchnera aphidicola* с помощью megablast**
    a. **Контиг 11**
    Посмотреть результат можно ниже.
    Контиг выровнялся на 25 участков хромосомы. При этом на хромосоме эти участки находятся в рамке (11103-621055).

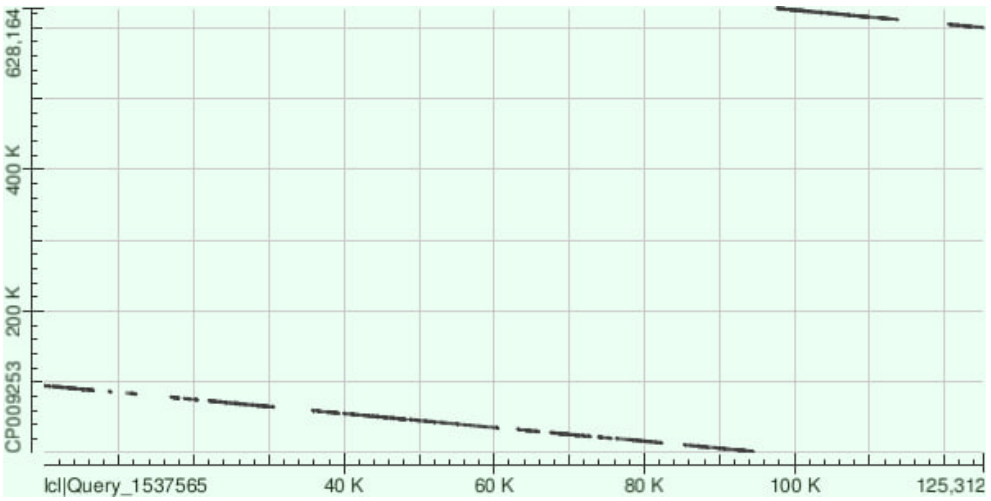

Рис. 1 Dot plot выравнивания контига 11 на геном

Ход прямой на графике показывает, что последовательности комплементарны друг другу. При есть обособленный участок комплементарности. Такая картина возникает, когда последовательности имеют разные точки начала прочтений

    b. **Контиг 1**
    Посмотреть результат можно ниже.
    Контиг выровнялся на 15 участков хромосомы. При этом на хромосоме эти участки находятся в рамке (98408- 200246).

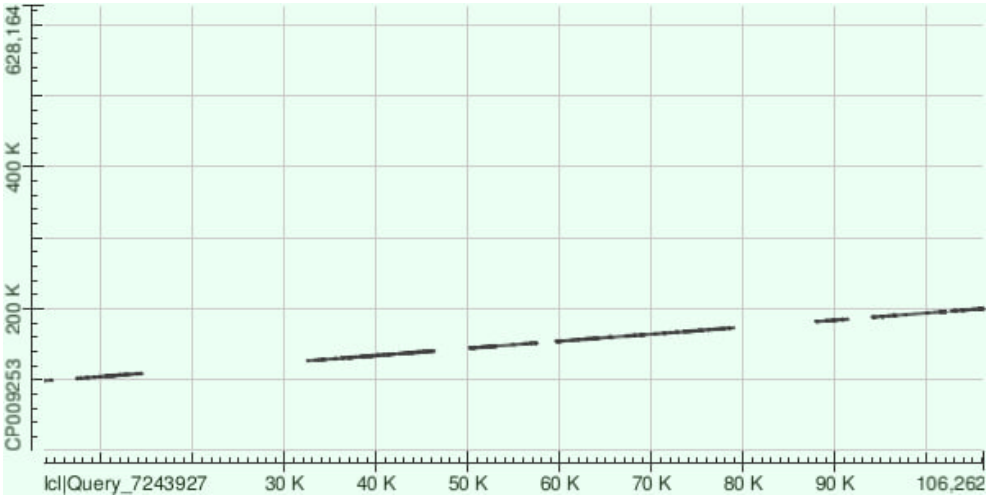

Рис. 2 Dot plot выравнивания контига 1 на геном

Ход прямой в этом случае, в отличие от приведенного выше, указывает на то, что последовательности “совпадают”, т.е. не комплементарны.

c. Контиг 14

Посмотреть результат можно ниже.
Контиг выровнялся на 14 участков хромосомы. При этом на хромосоме эти участки находятся в рамке (207661-266073).

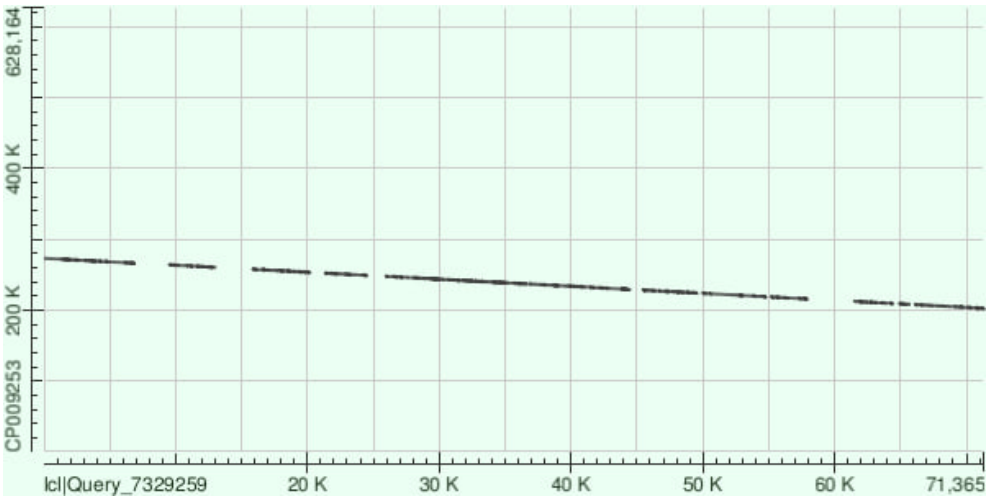

Рис. 3 Dot plot выравнивания контига 14 на геном

Ход прямой на графике показывает, как и для контига 11, что последовательности комплементарны.

На каждом из графиков разрывы в прямой указывают на неконсервативные участки, разные для двух последовательностей

d. Содержимое файлов с результатами megablast

1. Контиг 11

blastn
Iteration: 0
Query: NODE_11_length_125674_cov_44.550949
RID: NR0UPANG114
Database: n/a
Fields: query acc.ver, subject acc.ver, % identity, alignment length, mismatches, gap opens, q. start, q. end, s. start, s. end, evalue, bit score
 # 25 hits found

NODE_11_length_125674_cov_44.550949	CP009253.1	82.851	9633	1522	100	50907	60472	44693	35124	0.0	8517
NODE_11_length_125674_cov_44.550949	CP009253.1	78.380	9223	1738	201	85411	94500	11103	2004	0.0	5749
NODE_11_length_125674_cov_44.550949	CP009253.1	79.211	7379	1350	144	103943	111247	620926	613658	0.0	4959
NODE_11_length_125674_cov_44.550949	CP009253.1	76.339	8436	1695	239	39954	48261	55420	47158	0.0	4242
NODE_11_length_125674_cov_44.550949	CP009253.1	76.459	6151	1174	190	24545	30582	70621	64632	0.0	3085
NODE_11_length_125674_cov_44.550949	CP009253.1	78.201	5046	930	114	120355	125312	604795	599832	0.0	3068
NODE_11_length_125674_cov_44.550949	CP009253.1	75.782	6173	1247	207	97666	103713	627104	621055	0.0	2889
NODE_11_length_125674_cov_44.550949	CP009253.1	76.551	5433	1055	166	67532	72881	28363	23067	0.0	2772
NODE_11_length_125674_cov_44.550949	CP009253.1	75.317	5607	1141	190	1140	6626	93683	88200	0.0	2462
NODE_11_length_125674_cov_44.550949	CP009253.1	85.253	2231	299	23	76309	78519	20182	17962	0.0	2270
NODE_11_length_125674_cov_44.550949	CP009253.1	78.685	3453	614	99	35876	39267	59462	56071	0.0	2187
NODE_11_length_125674_cov_44.550949	CP009253.1	75.976	3226	687	68	78597	81767	17919	14727	0.0	1583
NODE_11_length_125674_cov_44.550949	CP009253.1	77.422	2777	543	71	63188	65924	32745	30013	0.0	1578
NODE_11_length_125674_cov_44.550949	CP009253.1	81.524	1851	291	41	73906	75730	22183	20358	0.0	1476
NODE_11_length_125674_cov_44.550949	CP009253.1	79.207	2044	361	40	48746	50760	46776	44768	0.0	1362
NODE_11_length_125674_cov_44.550949	CP009253.1	77.900	2086	395	42	111356	113422	613671	611633	0.0	1238
NODE_11_length_125674_cov_44.550949	CP009253.1	73.579	2411	535	72	21982	24360	73310	70970	0.0	830
NODE_11_length_125674_cov_44.550949	CP009253.1	77.076	1409	291	28	10863	12259	84409	83021	0.0	784
NODE_11_length_125674_cov_44.550949	CP009253.1	79.887	885	160	14	2	877	94696	93821	2.78e-180	632
NODE_11_length_125674_cov_44.550949	CP009253.1	77.230	953	197	16	18742	19686	76468	75528	1.73e-152	540
NODE_11_length_125674_cov_44.550949	CP009253.1	73.519	1182	269	41	16970	18128	78277	77117	5.09e-113	409
NODE_11_length_125674_cov_44.550949	CP009253.1	82.218	478	76	8	81927	82401	14465	13994	2.37e-111	403
NODE_11_length_125674_cov_44.550949	CP009253.1	76.923	442	69	19	20027	20445	75264	74833	2.56e-56	220
NODE_11_length_125674_cov_44.550949	CP009253.1	79.461	297	59	2	113554	113849	611524	611229	5.54e-53	209
NODE_11_length_125674_cov_44.550949	CP009253.1	78.632	234	45	5	8647	8878	86404	86174	3.41e-35	150

2. Контиг 1

blastn
Iteration: 0
Query: NODE_1_length_108447_cov_42.009186
RID: NR15K397114
Database: n/a
Fields: query acc.ver, subject acc.ver, % identity, alignment length, mismatches, gap opens, q. start, q. end, s. start, s. end, evalue, bit score
 # 15 hits found

NODE_1_length_108447_cov_42.009186	CP009253.1	74.950	13010	2711	430	33726	46466	127825	140555	0.0	5465
NODE_1_length_108447_cov_42.009186	CP009253.1	77.804	8168	1549	191	59485	67569	153752	161738	0.0	4796
NODE_1_length_108447_cov_42.009186	CP009253.1	77.747	7536	1434	178	50105	57504	144368	151796	0.0	4401
NODE_1_length_108447_cov_42.009186	CP009253.1	76.533	7274	1492	188	7333	14500	101712	108876	0.0	3777
NODE_1_length_108447_cov_42.009186	CP009253.1	79.983	4801	862	73	94019	98793	187938	192665	0.0	3450
NODE_1_length_108447_cov_42.009186	CP009253.1	79.589	4914	891	92	67774	72634	161898	166752	0.0	3415
NODE_1_length_108447_cov_42.009186	CP009253.1	76.216	6517	1391	138	72665	79108	166750	173180	0.0	3301
NODE_1_length_108447_cov_42.009186	CP009253.1	76.068	3652	764	81	87826	91441	181712	185289	0.0	1801
NODE_1_length_108447_cov_42.009186	CP009253.1	79.227	2070	352	55	100101	102142	194042	196061	0.0	1369
NODE_1_length_108447_cov_42.009186	CP009253.1	83.736	1199	184	9	32468	33661	126623	127815	0.0	1123
NODE_1_length_108447_cov_42.009186	CP009253.1	81.472	1209	220	4	98869	100074	192777	193984	0.0	989
NODE_1_length_108447_cov_42.009186	CP009253.1	76.492	1910	376	58	102449	104307	196373	198260	0.0	972
NODE_1_length_108447_cov_42.009186	CP009253.1	81.132	901	161	8	3756	4652	98408	99303	0.0	713
NODE_1_length_108447_cov_42.009186	CP009253.1	78.525	922	181	15	104513	105424	198467	199381	1.46e-167	590
NODE_1_length_108447_cov_42.009186	CP009253.1	75.479	730	127	37	105557	106262	199545	200246	4.58e-83	309

3. Контиг 14

blastn
Iteration: 0
Query: NODE_14_length_71403_cov_39.411552
RID: NR18H31U114
Database: n/a
Fields: query acc.ver, subject acc.ver, % identity, alignment length, mismatches, gap opens, q. start, q. end, s. start, s. end, evalue, bit score
 # 14 hits found

NODE_14_length_71403_cov_39.411552	CP009253.1	80.227	7060	1199	156	1	6967	273028	266073	0.0	5121
NODE_14_length_71403_cov_39.411552	CP009253.1	75.138	10884	2317	321	25958	36657	247596	236918	0.0	4748
NODE_14_length_71403_cov_39.411552	CP009253.1	78.495	5329	1009	108	66117	71365	207661	202390	0.0	3365
NODE_14_length_71403_cov_39.411552	CP009253.1	80.920	4130	727	49	49953	54055	223720	219625	0.0	3205
NODE_14_length_71403_cov_39.411552	CP009253.1	77.022	4178	797	133	45453	49564	228137	224057	0.0	2246
NODE_14_length_71403_cov_39.411552	CP009253.1	75.671	4583	981	104	36755	41284	236859	232358	0.0	2163
NODE_14_length_71403_cov_39.411552	CP009253.1	78.957	3165	571	78	41433	44553	232057	228944	0.0	2067
NODE_14_length_71403_cov_39.411552	CP009253.1	77.080	3617	728	83	9517	13088	263784	260224	0.0	1993
NODE_14_length_71403_cov_39.411552	CP009253.1	77.750	3245	630	79	21279	24481	252161	248967	0.0	1908
NODE_14_length_71403_cov_39.411552	CP009253.1	79.064	2713	496	61	55202	57887	218384	215717	0.0	1797
NODE_14_length_71403_cov_39.411552	CP009253.1	76.555	3007	601	75	61463	64422	212243	209294	0.0	1552
NODE_14_length_71403_cov_39.411552	CP009253.1	73.400	4421	981	144	15866	20188	257546			

253223	0.0	1469									
NODE_14_length_71403_cov_39.411552			CP009253.1	76.718	902	185	22	64868	65758	208904	
208017	2.17e-134	479									
NODE_14_length_71403_cov_39.411552			CP009253.1	76.183	676	141	17	54103	54763	219491	
218821	3.85e-92	339									