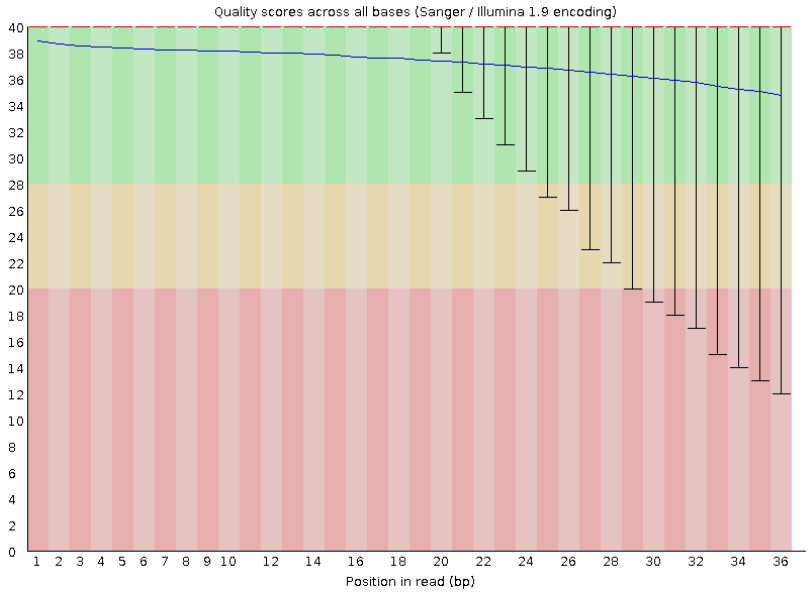
Прак 15. Сборка генома de novo

Загрузим доставшиеся мне прочтения и проанализируем их качество программой fastqc:

fastqc SRR4240360.fastq.gz

Посмотрим в получившийся файл



В целом среднее качество неплохое, но видим, что к концу у большого числа прочтений качество сильно падает, отфильтруем такие риды с помощью trimmomatic.

Для начала создадим общий файл с адаптерами для работы trimmomatic

cat ../../adapters/ >> adapters.fasta*

Проведем тримминг наших прочтений

TrimmomaticSE -threads 4 SRR4240360.fastq.gz output.fq.gz ILLUMINACLIP:adapters.fasta:2:7:7

Результаты:
Input Reads: 8254632 Surviving: 8212774 (99.49%) Dropped: 41858 (0.51%)
Видим, что по итогам фильтрования удалилось 41858 прочтений, они и содержали в себе адаптерные последовательности.

Далее удалим с правых концов чтений нуклеотиды с качеством ниже 20, оставим только такие чтения, длина которых не меньше 32 нуклеотидов.

TrimmomaticSE -threads 4 output.fq.gz output_final.fq.gz TRAILING: 20 MINLEN:32

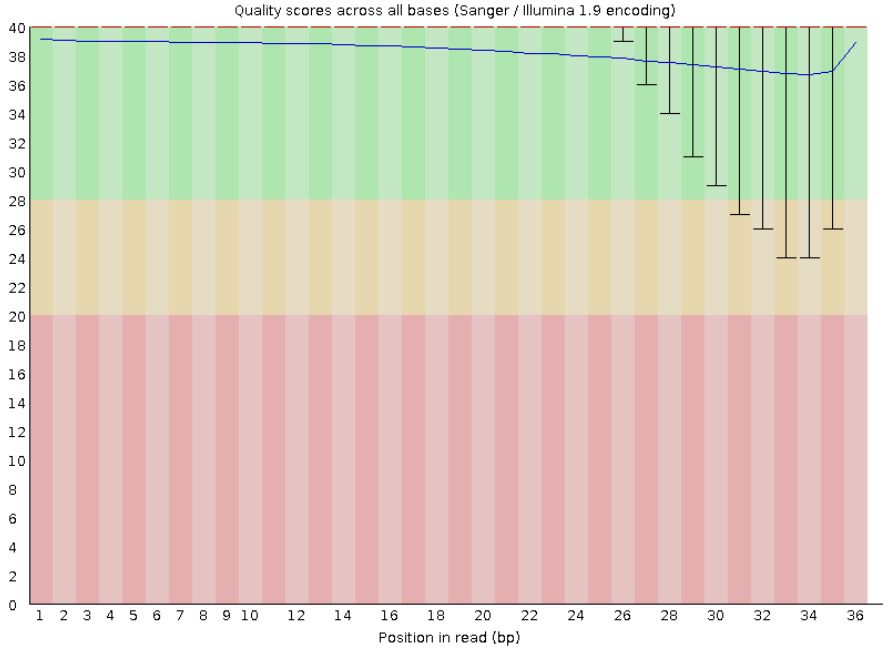
Результаты: Input Reads: 8212774 Surviving: 7915474 (96.38%) Dropped: 297300 (3.62%)

На этом этапе удалилось еще 297,300 прочтений низкого качества. В сумме с предыдущими, всего удалилось 339,158 прочтений (4.1% от исходного числа). Размер файла изменился с 193Мб до 183мБ.

Проведем для получившихся ридов fastqc уже знакомой командой:

fastqc output_final.fq.gz

Посмотрим, что получилось:



Видим, что качество прочтений уже не так сильно проседает к концу, как минимум 90% всех прочтений имеют среднее качество выше 24 (тут я немножко на глаз, но кажется все-таки 24).

Проведем индексирование для последующей сборки генома:

mkdir velveth_31

velveth velveth_31 31 -short -fastq.gz output_final.fq.gz

И наконец, запустим сборку генома:

velvetg velveth_31/

Посмотрим, что получилось.
Всего создалось 603 контига, N50 43070, максимальная длина контига 113474 нуклеотидов, суммарная длина 678075 нуклеотидов.

Быстро с помощью питона найдем 3 самых длинных контига. Это:

a) NODE_1_length_113474_cov_33.525459 (далее контиг 1)
b) NODE_5_length_83603_cov_33.646065 (далее контиг 5)
c) NODE_4_length_64155_cov_35.847324 (далее контиг 4)

В наименовании NODE_1_length_113474_cov_33.525459, 113474 означает длину контига, 33.525459 его покрытие.

Кроме хороших контигов, есть маленькие контиги (~100-400 нуклеотидов) с покрытием примерно в 5 раз ниже чем среднее. Например:

a) NODE_104_length_68_cov_14.147058
b) NODE_105_length_32_cov_2.781250
c) NODE_251_length_62_cov_3.274194

Это просто маленькие контиги, которые плохо собрались и несут в себе не очень много информации.

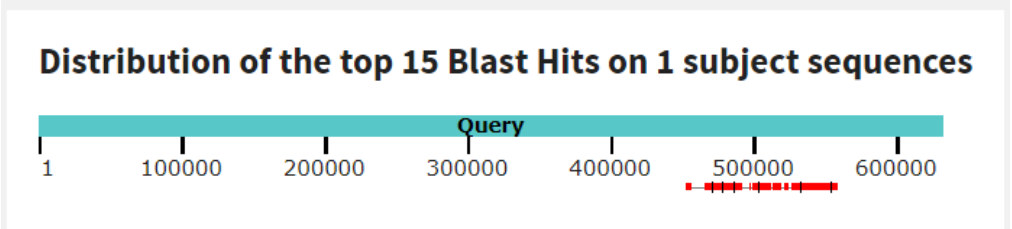
Сравним каждый из трех самых длинных контигов с хромосомой *Buchnera aphidicola* с помощью megablast.

Контиг 1, самый длинный, длина 113474 нуклеотидов, при “бластовании” на геном, начало на геноме: 449411, конец на геноме 550219.
Ложится на геном 15 разными участками. Краткие результаты по картированию собраны в первой таблице:

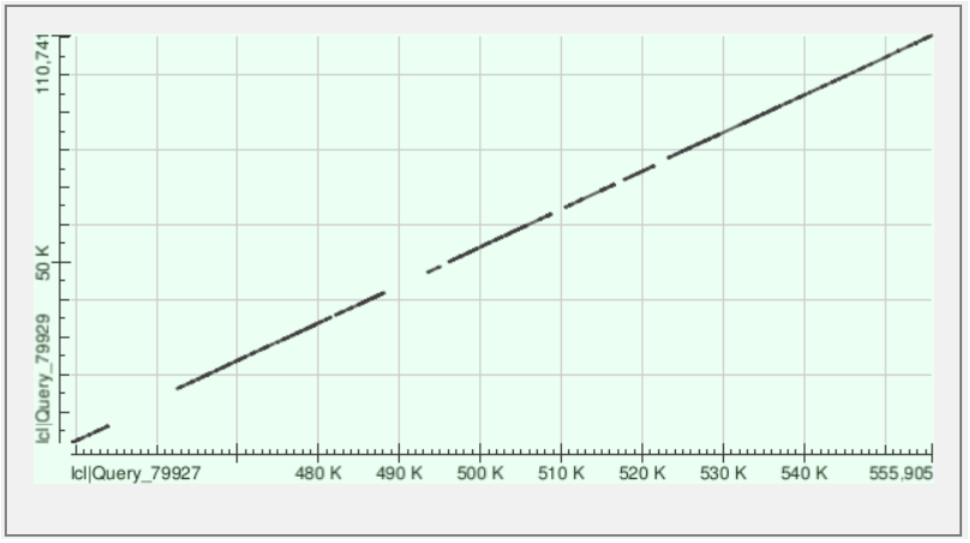
Range №	Start	Stop	Score	E-value	Identities	Gaps	Strand
1	83427	104897	17265 bits	0.0	17687/21720(81%)	543/21720(2%)	Plus/Plus
2	105104	110741	4331 bits	0.0	4572/5654(81%)	125/5654(2%)	Plus/Plus
3	21088	28401	4047 bits	0.0	5690/7388(77%)	206/7388(2%)	Plus/Plus
4	54455	62900	3949 bits	0.0	6510/8611(76%)	339/8611(3%)	Plus/Plus
5	64592	70771	3895 bits	0.0	4896/6240(78%)	198/6240(3%)	Plus/Plus
6	77769	83357	3029 bits	0.0	4371/5685(77%)	206/5685(3%)	Plus/Plus
7	16117	21058	2724 bits	0.0	3863/5016(77%)	164/5016(3%)	Plus/Plus

8	35738	41795	2278 bits	0.0	4625/6242(74%)	316/6242(5%)	Plus/Plus
9	28504	34378	2237 bits	0.0	4432/5977(74%)	262/5977(4%)	Plus/Plus
10	1574	6226	2167 bits	0.0	3574/4735(75%)	158/4735(3%)	Plus/Plus
11	72037	75766	2128 bits	0.0	2922/3782(77%)	99/3782(2%)	Plus/Plus
12	50062	54339	1914 bits	0.0	3254/4324(75%)	155/4324(3%)	Plus/Plus
13	47268	48644	1014 bits	0.0	1109/1385(80%)	15/1385(1%)	Plus/Plus
14	34584	35263	573 bits	2e-162	562/684(82%)	16/684(2%)	Plus/Plus
15	48778	48896	145 bits	1e-33	107/120(89%)	5/120(4%)	Plus/Plus

Визуально контиг ложится на хромосому так:



Дот плот для этого контига:

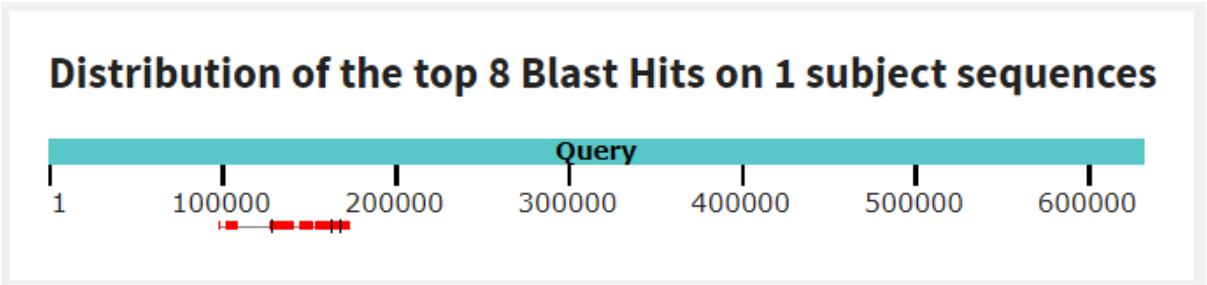


Мы видим, что почти на всем протяжении контиг сильно похож на данный участок референсной хромосомы, делеций, инверсий и инсерций не наблюдается.

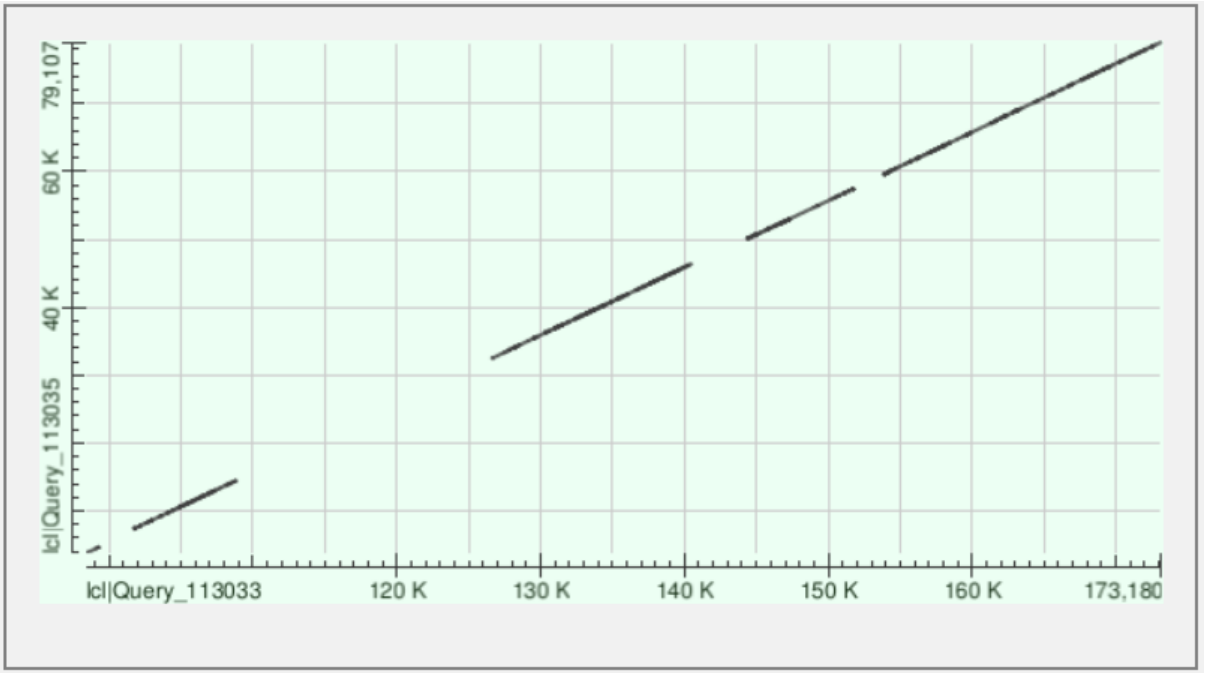
Контиг 5, длина 83603 нуклеотидов, картируется на геном на участок с 98408 по 173180 нуклеотид, при картировании выделяется 8 участков с высокой идентичностью, подробно о каждом из участке вы можете узнать в таблице ниже.

Range №	Start	Stop	Score	E-value	Identities	Gaps	Strand
1	33725	46465	5465 bits	0.0	9755/13014(75%)	556/13014(4%)	Plus/Plus
2	59484	67568	4796 bits	0.0	6359/8172(78%)	272/8172(3%)	Plus/Plus
3	50104	57503	4401 bits	0.0	5861/7538(78%)	247/7538(3%)	Plus/Plus
4	7332	14499	3777 bits	0.0	5568/7275(77%)	217/7275(2%)	Plus/Plus
5	67773	72633	3415 bits	0.0	3909/4912(80%)	108/4912(2%)	Plus/Plus
6	72664	79107	3301 bits	0.0	4964/6514(76%)	153/6514(2%)	Plus/Plus
7	32467	33660	1123 bits	0.0	1004/1199(84%)	11/1199(0%)	Plus/Plus
8	3755	4651	713 bits	0.0	731/901(81%)	9/901(0%)	Plus/Plus

Визуально контиг ложится на хромосому так:



Дот плот выравнивания:



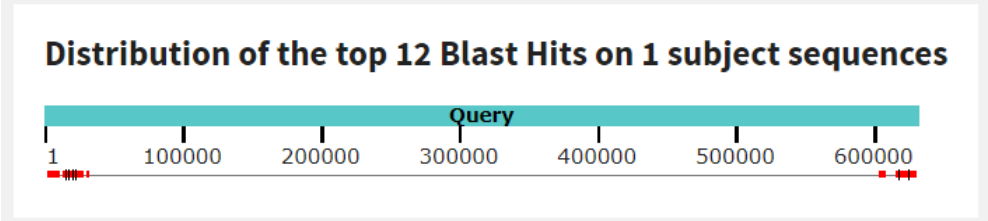
Видим что существует 4 региона с высокой идентичностью, а участки между ними отличаются настолько, что megablast не находит совпадений, однако делеций, инверсий и инсерций не наблюдается.

Контиг 4, третий по длине, длина 64185 нуклеотида, картируется на геном на участок с 599832 по 32745 нуклеотид (проходит через конец хромосомы), при картировании выделяется 12 участков с высокой идентичностью, подробно о каждом из участке вы можете узнать в таблице ниже.

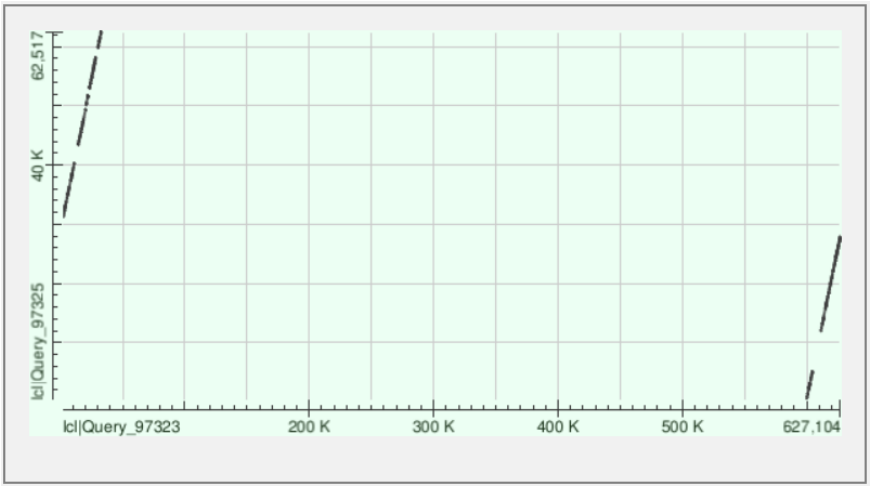
Range №	Start	Stop	Score	E-value	Identities	Gaps	Strand
1	31205	40294	5749 bits	0.0	7222/9216(78%)	242/9216(2%)	Plus/Plus
2	14458	21762	4959 bits	0.0	5843/7377(79%)	180/7377(2%)	Plus/Plus
3	393	5350	3068 bits	0.0	3952/5052(78%)	182/5052(3%)	Plus/Plus
4	21992	28039	2889 bits	0.0	4681/6176(76%)	254/6176(4%)	Plus/Plus
5	52824	58173	2772 bits	0.0	4157/5431(77%)	215/5431(3%)	Plus/Plus

6	47186	49396	2270 bits	0.0	1902/2231(85%)	30/2231(1%)	Plus/Plus
7	43938	47108	1583 bits	0.0	2454/3229(76%)	94/3229(2%)	Plus/Plus
8	59781	62517	1578 bits	0.0	2153/2780(77%)	90/2780(3%)	Plus/Plus
9	49975	51799	1476 bits	0.0	1508/1850(82%)	49/1850(2%)	Plus/Plus
10	12283	14349	1238 bits	0.0	1624/2085(78%)	64/2085(3%)	Plus/Plus
11	43304	43778	403 bits	1e-111	393/478(82%)	9/478(1%)	Plus/Plus
12	11856	12151	209 bits	3e-53	236/297(79%)	2/297(0%)	Plus/Plus

Визуально контиг ложится на хромосому так:



Дот плот выравнивания:



Выглядит немного странно, но это из-за того, что контиг своей серединой проходит через начало референсной хромосомы. Так же как и у других контигов видим, что на большей части контига он почти полностью совпадает с референсной хромосомой, делеции, инверсии и инсерции отсутствуют.

Допы не успел :