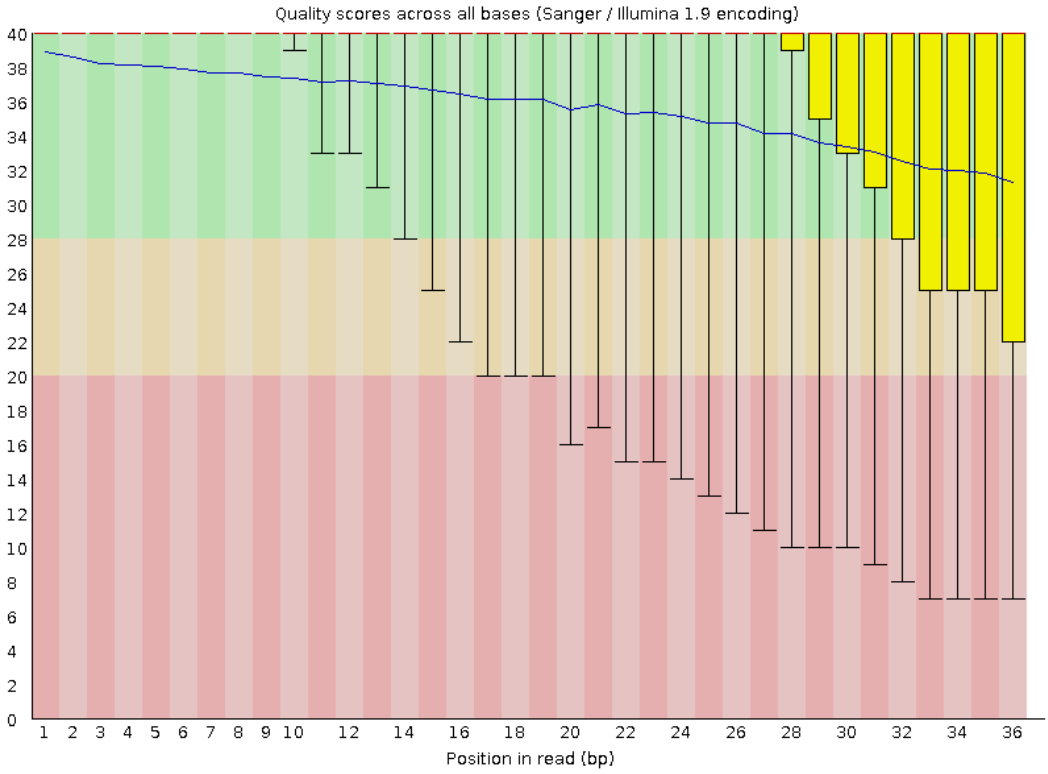
Практикум 15. Сборка генома de novo

1. Подготовка чтений программой trimmomatic

Для начала была произведена оценка качества чтений программой fastqc
`fastqc SRR4240361.fastq.gz`

Качество ридов сильно ухудшается после 32 нуклеотида. Необходим тримминг



До очистки 7 272 621 ридов с единой длиной 36 нуклеотидов
Адаптеры отсутствуют
Есть избыточно представленные последовательности полиA и полиN

Для фильтрации чтений программой trimmomatic создадим вспомогательный файл с возможными адаптерами.
`cat /mnt/scratch/NGS/adapters/* >> adapters.fasta`

С помощью программы trimmomatic удалим адаптеры
`java -jar /usr/share/java/trimmomatic.jar SE -threads 4 SRR4240361.fastq.gz step_one_no_adapters.fq.gz ILLUMINACLIP:adapters.fasta:2:7:7`

Input Reads: 7272621 Surviving: 7238089 (99.53%) Dropped: 34532 (0.47%)
В результаты было удалено менее 0.5% ридов.

Для полученных последовательностей с удаленными адаптерами проведем фильтрацию по качеству прочтений. Удалим с правых концов чтений нуклеотиды с качеством ниже 20 и чтения длина которых меньше 32 нуклеотидов.

```
java -jar /usr/share/java/trimmomatic.jar SE -threads 4 step_one_no_adapters.fq.gz trim_reads.fq.gz TRAILING:20 MINLEN:32
```

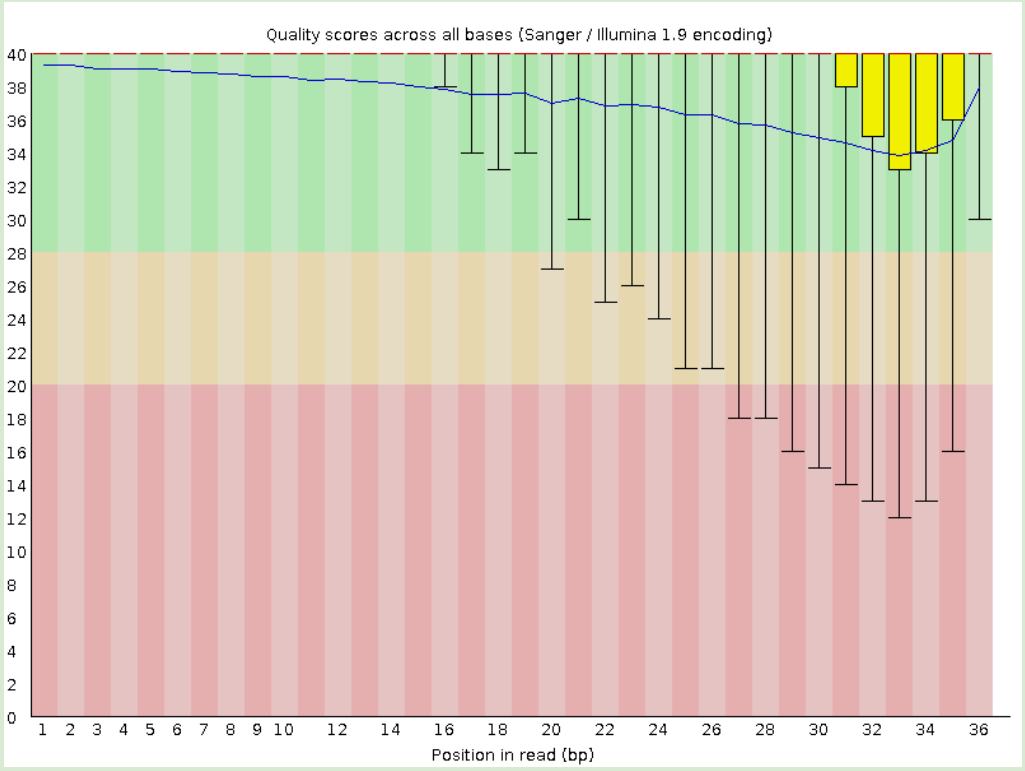
Input Reads: 7238089 Surviving: 6834335 (94.42%) Dropped: 403754 (5.58%)

Суммарно отфильтровано 6% ридов.

Проверим качество чтений после тримминга

```
fastqc trim_reads.fq.gz
```

Качество чтений значительно улучшилось. Медианы уже не попадают в желтую область и усы находятся в красной области начиная с 28 нуклеотида, в отличии от начала, где 16 нуклеотид усами попадал в красную область.



2. Сборка генома

Для работы программы требуется директория, куда будут сохраняться файлы с k-мерами

mkdir velveth_31

velveth velveth_31 31 -fastq.gz trim_reads.fq.gz -short (31 - длина k-мера, -fastq.gz формат входного файла)

Запустим сборку генома
velvetg velveth_31/

N50 25 683
max контиг 49 238

3 максимально длинных контига (length_ДЛИНА_cov_ПОКРЫТИЕ)
>NODE_6_length_49238_cov_26.660851
>NODE_2_length_45555_cov_26.450466
>NODE_34_length_43866_cov_23.514977

Максимальное покрытие 90.744682
Минимальное покрытие 2.238095
Медиана 11.750000

Есть контиги, покрытие которых более чем в 5 раз отличается от медианы, как в большую сторону,
>NODE_78_length_47_cov_90.744682
>NODE_91_length_33_cov_76.636360
>NODE_95_length_31_cov_64.903229

так и в меньшую.
>NODE_391_length_63_cov_2.238095
>NODE_140_length_64_cov_2.906250
>NODE_279_length_75_cov_2.946667

Контиги с большим покрытием имеют более короткую длину чем контиги с меньшим.
Например контиг с медианным покрытием короче чем контиги с самым маленьким покрытием
>NODE_324_length_48_cov_11.750000

3. Анализ

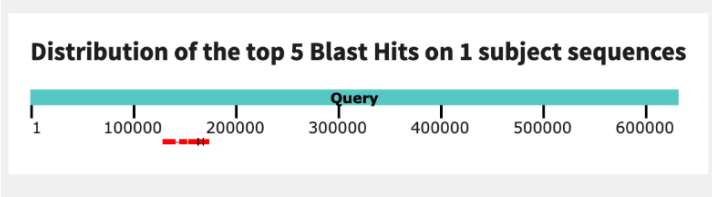
Для 3 самых длинных контигов рассмотрим как они ложатся на хромосому *Buchnera aphidicola* (GenBank/EMBL AC — CP009253). Проведем выравнивание алгоритмом megablast.

>NODE_6_length_49238_cov_26.660851

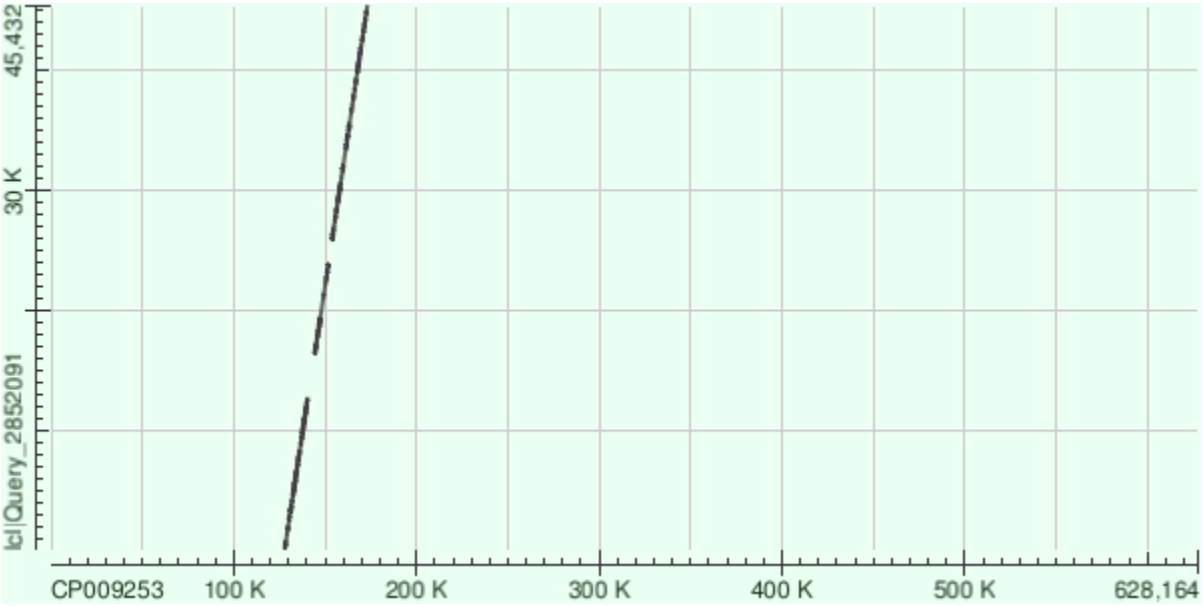
Контиг ложится на участок генома [127825 - 173180] 5 кусочками, соответствует “+” цепи

% идентичности	Длина фрагмента	start_genome	end_genome	start_contig	end_contig	Score	E-value
74.958	13014	127825	140555	50	12790	5465	0.0
77.753	7538	144368	151796	16429	23828	4401	0.0
77.814	8172	153752	161738	25809	33893	4796	0.0
79.581	4912	161898	166752	34098	38958	3415	0.0
76.205	6514	166750	173180	38989	45432	3301	0.0

Расположение контига на хромосоме



Dot plot выравнивания



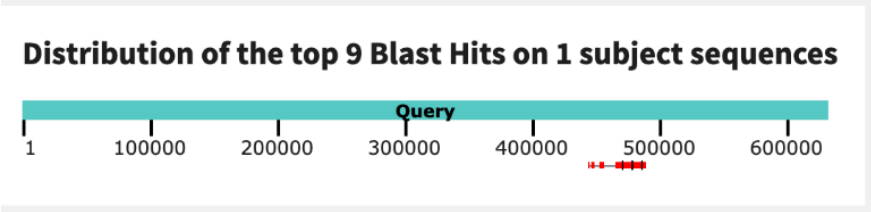
Контиг сильно похож на участок хромосомы, но есть небольшие участки контига, которые не выровнялись с референсом (до 3,5 kb)

>NODE_2_length_45555_cov_26.450466

Контиг ложится на участок генома [440755 - 485679] 9 кусочками. При этом контиг соответствует “-” цепи (начало контига ложится на конец участка комплементарности с геномом)

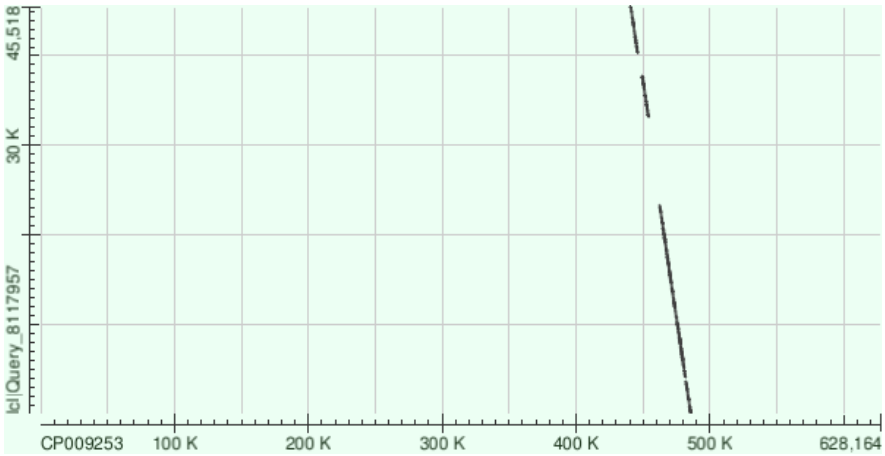
% идентичности	Длина фрагмента	start_genome	end_genome	start_contig	end_contig	Score	E-value
88.947	190	440755	440944	45518	45337	228	5.55e-59
79.044	1694	441135	442817	45215	43540	1134	0.0
80.255	3054	442877	445895	43410	40383	2242	0.0
75.480	4735	449411	454069	37811	33159	2167	0.0
77.007	5019	462496	467424	23268	18324	2724	0.0
77.017	7388	467412	474667	18297	10984	4047	0.0
74.151	5977	474844	480660	10881	5007	2237	0.0
82.096	687	480874	481548	4801	4119	573	6.19e-163
76.531	3724	481997	485679	3647	12	1916	0.0

Расположение контига на хромосоме



Контиг достаточно похож на участок хромосомы, но есть участки контига, которые не выровнялись с референсом (например участок [23268 - 33159] 10 000 bp, который не ложится на хромосому)

Dot plot выравнивания

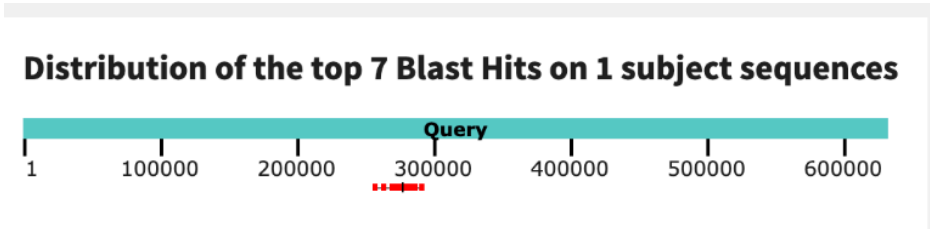


>NODE_34_length_43866_cov_23.514977

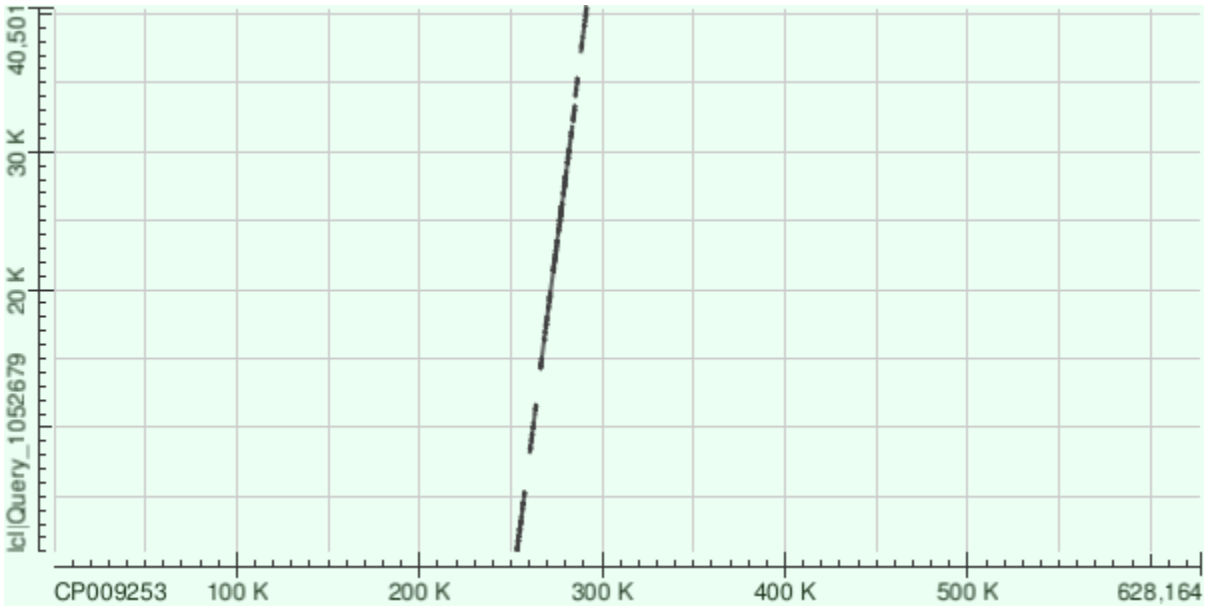
Контиг ложится на участок генома [253223 - 291560] 7 кусочками. Контиг соответствует “+” цепи.

% идентичности	Длина фрагмента	start_genome	end_genome	start_contig	end_contig	Score	E-value
73.436	4427	253223	257546	977	5299	1469	0.0
77.030	3609	260224	263784	8077	11648	1993	0.0
78.782	9657	266073	275551	14198	23677	6154	0.0
75.881	8396	275566	283706	23736	31957	3890	0.0
76.237	1132	283963	285070	32205	33314	558	1.67e-158
75.982	1349	285200	286535	34011	35345	671	0.0
77.508	3419	288181	291560	37135	40501	1969	0.0

Расположение контига на хромосоме



Dot plot выравнивания



Контиг сильно похож на участок хромосомы, но есть небольшие участки контига, которые не выровнялись с референсом : 1000 bp в начале контига и 3000 bp в конце. Так же 2 участка внутри контига длиной 3000 bp которые не выровнялись на геном.