Сборка de novo по архиву с чтениями.

Скачаем архив:

wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR424/000/SRR4240360/SRR4240360.fastq.gz

1) Уберём возможные остатки адаптеров:
TrimmomaticSE SRR4240360.fastq.gz SRR4240360_forward.fastq.gz ILLUMINACLIP:adapters.fasta:2:7:7

Сколько процентов последовательностей чтений оказалось остатками адаптеров: 0.50%

Удалим с правых концов чтений нуклеотиды с качеством ниже 20, оставим только такие чтения, длина которых не меньше 32 нуклеотидов:
TrimmomaticSE SRR4240360_forward.fastq.gz SRR4240360_forward_HQ.fastq.gz TRAILING:20 MINLEN:32

Сколько чтений было удалено: 291607
Размер файла до очистки: 193M
Размер файла после очистки: 184M

2) Подготовим k-меры с помощью velveth (k=31):
velveth velveth_SRR4240360_forward_HQ.fastq.gz 31 -short -fastq.gz SRR4240360_forward_HQ.fastq.gz

3) Создадим сборку с помощью velvetg:
velvetg velveth_SRR4240360_forward_HQ.fastq.gz

N50: 43070
3 самых длинных контига и их покрытия:

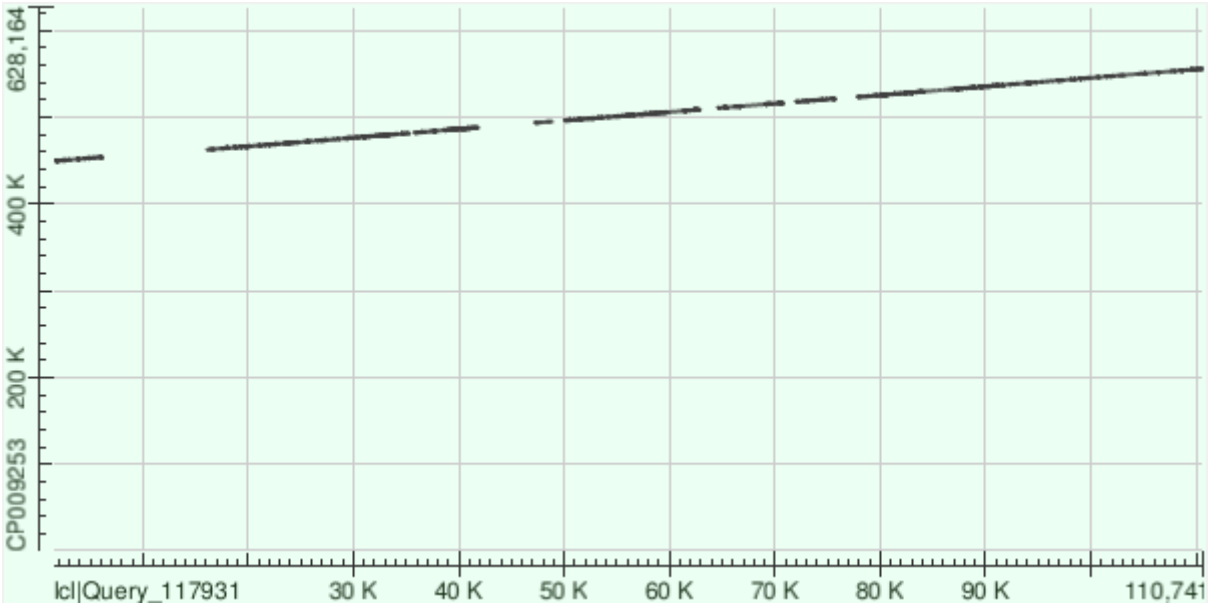
ID	lgth	out	in	long_cov	short1_cov	short1_Ocov
1	113474	0	0	0.000000	33.534396	33.534396
4	64155	0	2	0.000000	35.869924	35.869924
5	91818	0	0	0.000000	33.497430	33.497430

Примеры аномальных контигов:

ID	lgth	out	in	long_cov	short1_cov	short1_Ocov
171	1	4	4	0.000000	134954.000000	134954.000000
204	36	1	1	0.000000	2.694444	2.694444
574	0	3	1	Inf	Inf	Inf

4) Сравним 3 самых длинных контига с хромосомой *Buchnera aphidicola*:
Контиг с ID 1:
Координаты участка хромосомы, соответствующие контигу (выбран участок выравнивания с минимальным E-value и наибольшим покрытием): 528794..550219
Гэпов: 545/21721 (2%)
Идентичных нуклеотидов: 17688/21721 (81%)

Dot Plot 15 участков выравнивания:



Контиг и хромосома имеют одинаковое направление, контиг хорошо выровнялся с хромосомой.

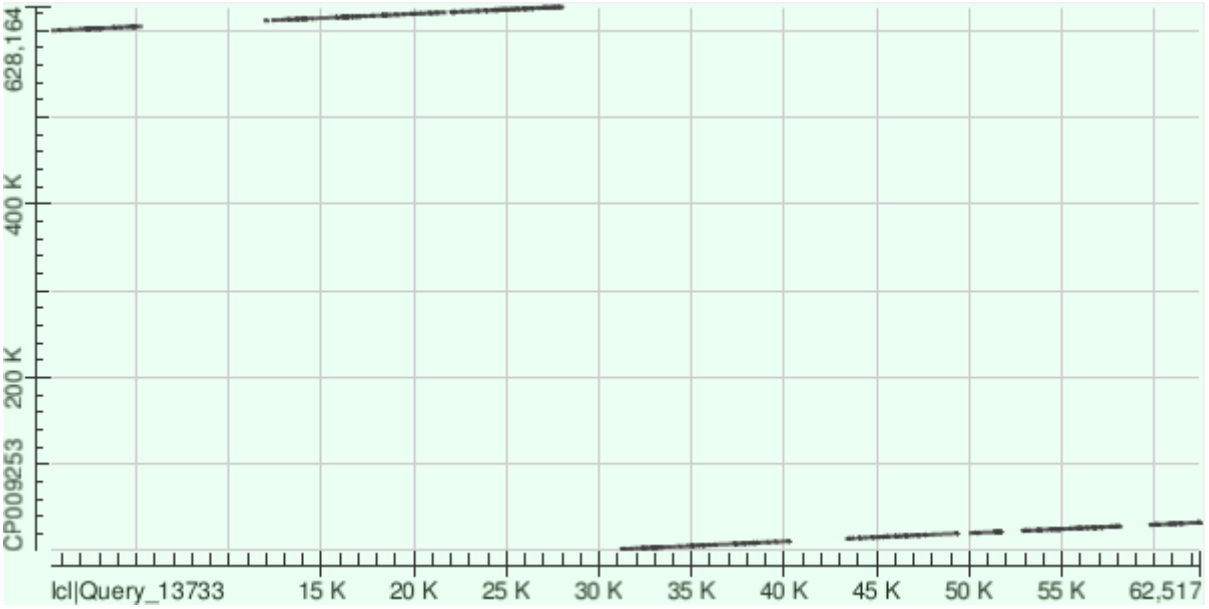
Контиг с ID 4:

Координаты участка хромосомы, соответствующие контигу (выбран участок выравнивания с минимальным E-value и наибольшим покрытием): 2004..11103

Гэпов: 256/9223 (2%)

Идентичных нуклеотидов: 7229/9223 (78%)

Dot Plot 12 участков выравнивания:



Выравнивание получилось кривое, первая половина контига находится в конце хромосомы, а вторая половина в начале.

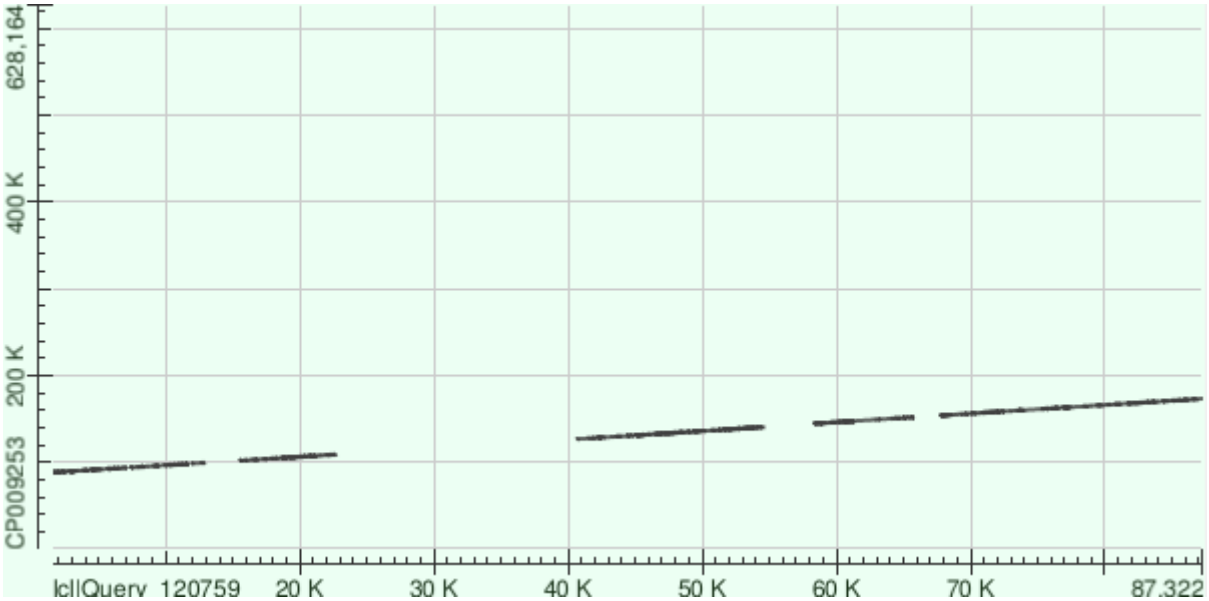
Контиг с ID 5:

Координаты участка хромосомы, соответствующие контигу (выбран участок выравнивания с минимальным E-value и наибольшим покрытием): 127825..140555

Гэпов: 548/13010 (4%)

Идентичных нуклеотидов: 9751/13010 (75%)

Dot Plot 10 участков выравнивания:



Выравнивание прямое, однако начало контига и его остальную часть разделяет очень большой гэп (23k-41k).