Сборка генома де ново на основе чтений SRR4240381.

1.

TrimmomaticSE SRR4240380.fastq.gz output.fastq.gz -threads 4 ILLUMINACLIP:adapters.fasta:2:7:7

Input Reads: 5217318 Surviving: 5119144 (98.12%) Dropped: 98174 (1.88%)

TrimmomaticSE Trailing:20 MINLEN:32 output.fastq.gz output_final.fastq.gz

Input Reads: 5119144 Surviving: 4865359 (95.04%) Dropped: 253785 (4.96%)

Удалено было 253785 (4.96%) чтений. Размер файлов уменьшился с 105Мб до 98,3Мб.

2.

mkdir velveth_31

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

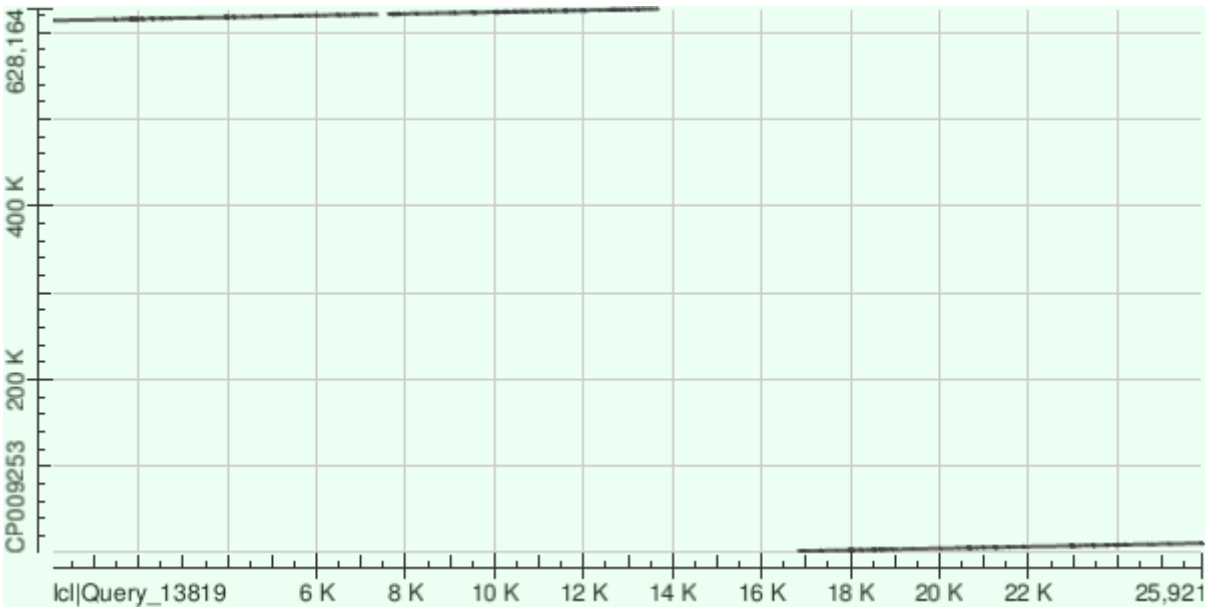
velvetg velveth_31/

Final graph has 401 nodes and n50 of 12042, max 25915, total 660284, using 0/4865359 reads

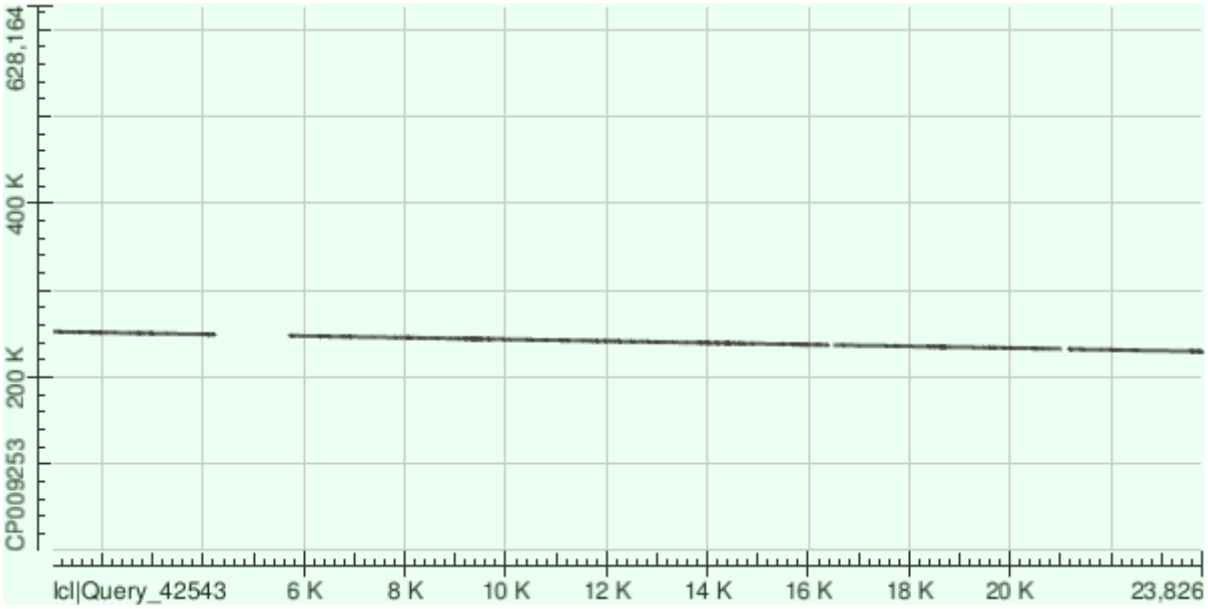
11 контиг имеет покрытие 126% при длине 2106 нт, что может свидетельствовать о том , что он входит в состав повторов.56 контиг имеет покрытие 130% при длине 934, что наталкивает на подобные мысли.

Самые длинные контиги были найдены разделением seqretsplit и поиском самых больших файлов, ими оказались 3, 20 и 23 контиги с длинами 25915, 23850, 23807 и покрытиями 27, 24, 25, соответственно.

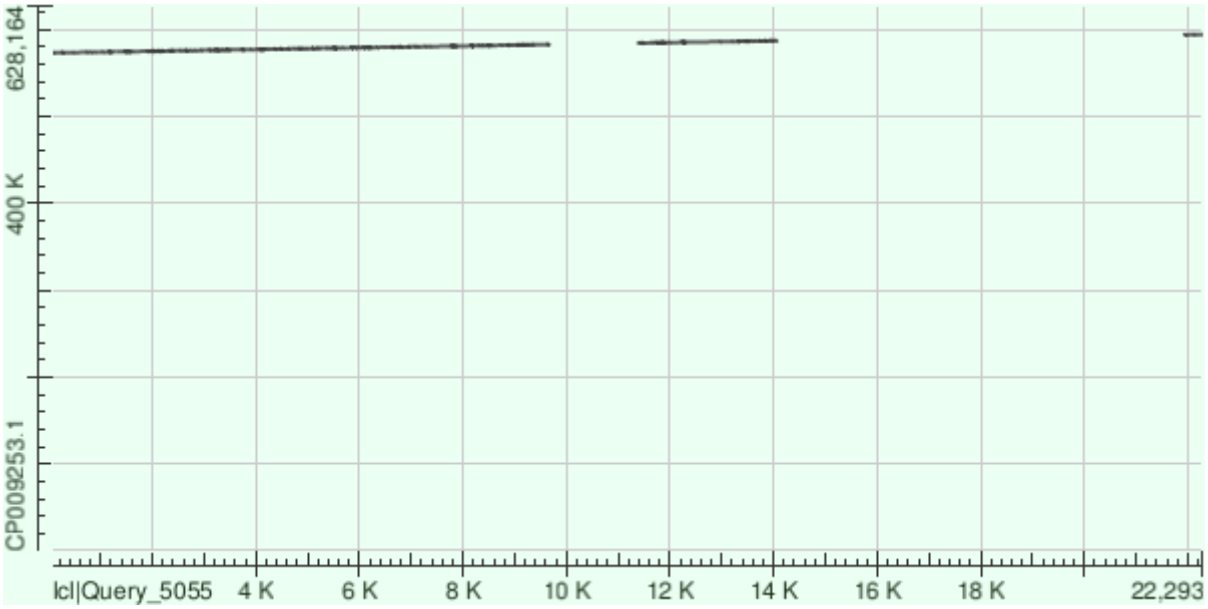
Программой megablast я выровнял три самых длинных контига (3, 20, 23) на хромосому Buchnera aphidicola. 3 контиг лёг на координаты 613658 - 11103 (через начало секвенирования) прямой цепи референса.



Контиг 20 лег на координаты хромосомы 229411 - 252164 с прерыванием 247596 - 248967, он лежит на обратной цепи по отношению к референсу



Контиг 23 лег на координаты хромосомы 573092 - 594099 с прерываниями 582686 - 584329, 587055 - 593743, он лежит на той же цепи, что и референс



Лучше всего выровнялся 3 контиг, так в выравнивании нет больших пропусков, гэпы только небольшие. 20 контиг выровнялся чуть хуже , в выравнивании уже есть промежуток более килобазы. Довольно плохо выровнялся 23 контиг, там помимо небольшого есть ещё 8кб

промежуток, появление которого может быть вызвано плохим качеством прочтений либо большой вариабельностью участка.