cat /mnt/scratch/NGS/adapters/*.fa > all_adapters.fasta – объединение всех адаптеров

TrimmomaticSE SRR4240378.fastq.gz SRR4240378_trimmed.fastq ILLUMINACLIP:all_adapters.fasta:2:7:7 2> trimmomatic1.log – удаление остатков адаптеров.

Сколько процентов последовательностей чтений оказалось остатками адаптеров –
Dropped: 81843 (1.85%)
Затем я удалил с правых концов чтений нуклеотиды с качеством ниже 20, оставил только такие чтения, длина которых не меньше 32 нуклеотидов. Для этого была использована следующая команда:

TrimmomaticSE SRR4240378_trimmed.fastq SRR420378_final.fastq TRAILING:20 MINLEN:32

Dropped: 184006 (4.24%) – в результате удалено.

Размеры файлов:

После удаления последовательностей адаптеров: 437Мб
После чистки: 418Мб.

Для сборки генома с использованием программы velvet, в основе которой лежит алгоритм на графе де Брёйна, первоначальным этапом является подготовка k-меров. Для этого была применена следующая команда:

velveth kmers 31 -fastq -short SRR4240378_final.fastq

Затем была запущена сборка на основе k-меров:

velvetg kmers

Вывод программы показал, что было собрано 208 контигов. N50=7028 bp. Три самых длинных контига (из файла contigs.fa):

>NODE_8_length_36746_cov_20.017199
>NODE_57_length_19371_cov_20.546642
>NODE_15_length_16745_cov_20.901762

Три контига с аномально большим покрытием:

>NODE_81_length_934_cov_102.748390
>NODE_19_length_2106_cov_100.555084
>NODE_88_length_501_cov_68.243515

Три контига с аномально малым покрытием:

>NODE_166_length_64_cov_2.843750
>NODE_281_length_72_cov_2.944444
>NODE_271_length_31_cov_3.225806

Медианное покрытие: 18.4546. Итого, все контиги с аномально малым покрытием меньше медианного более чем в 5 раз. Контиги >NODE_81_length_934_cov_102.748390 >NODE_19_length_2106_cov_100.555084 имеют покрытие больше медианного более чем в 5 раз. Изучим контиги >NODE_166_length_64_cov_2.843750 и >NODE_281_length_72_cov_2.944444.

Их последовательности, соответственно,
GCCGGATCGACAGCCATGTAACGGTCAACTCAGAACTGGCACGGACCAGGGGAATCCGAC
TGTCTAATTAAAACAAAGCATCGCGATGGCCCGG

И
CATTGGTGTGCCTCGAGACGAATTTCTAACCGACGGCCGTCGACCGGGCACCGTTTTTTT
TTCAACTCCGCATCGCCGTACCACTACGCACGGTGATGGTCA

В обоих можно увидеть очень большое количество GC-повторов разного состава, а во втором – еще и относительно длинный участок из T, что можно связать с ошибкой секвенатора.

Затем был произведен анализ с помощью алгоритма BLAST. Я взял штамм Buchnera aphidicola str. Bp. (RefSeq: GCF_000007725.1, AC хромосомы – NC_004545). Затем сравнил программой megablast каждый из трёх самых длинных контигов с этой хромосомой.

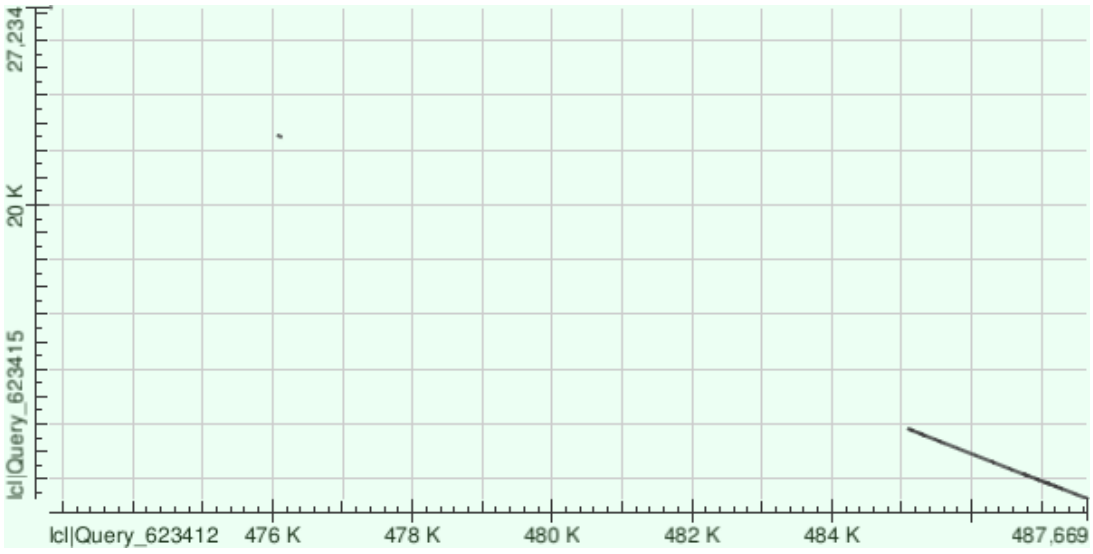

Рис.1 DotPlot для выравнивания контига NODE_8 с хромосомой Buchnera aphidicola str. Bp

Характеристика выравниваний контига 8 на хромосому NC_004545.1:

query acc.ver, subject acc.ver, % identity, alignment length, mismatches, gap opens, q. start, q. end, s. start, s. end, evalue, bit score
NC_004545.1 NODE_8_length_36746_cov_20.017199 74.510 2652 544 111 485079 487669 11839 9259 0.0 1031
NC_004545.1 NODE_8_length_36746_cov_20.017199 94.872 78 4 0 476062 476139 22550 22473 2.13e-27 122
NC_004545.1 NODE_8_length_36746_cov_20.017199 89.362 47 5 0 472806 472852 27234 27188 1.69e-08 60.2

Контиг 8: выравнялся на прямую цепь хромосомы на координатах 472806 - 487669

Рис.2 DotPlot для выравнивания контига NODE_57 с хромосомой Buchnera aphidicola str. Bp

Характеристика выравниваний контига 57 на хромосому NC_004545.1:

query acc.ver, subject acc.ver, % identity, alignment length, mismatches, gap opens, q. start, q. end, s. start, s. end, evalue, bit score
NC_004545.1 NODE_57_length_19371_cov_20.546642 87.200 125 12 4 566447 566570 11235 11114 1.12e-32 139
NC_004545.1 NODE_57_length_19371_cov_20.546642 92.941 85 4 2 566707 566791 11034 10952 1.12e-27 122

Контиг 57: выравнялся на прямую цепь хромосомы на координатах 566447-566791.

Для первых двух контигов я использовал стандартные параметры megablast. Однако с третьим при них не было никаких находок, поэтому я понизил word size до 24 и повысил e-value до 1.

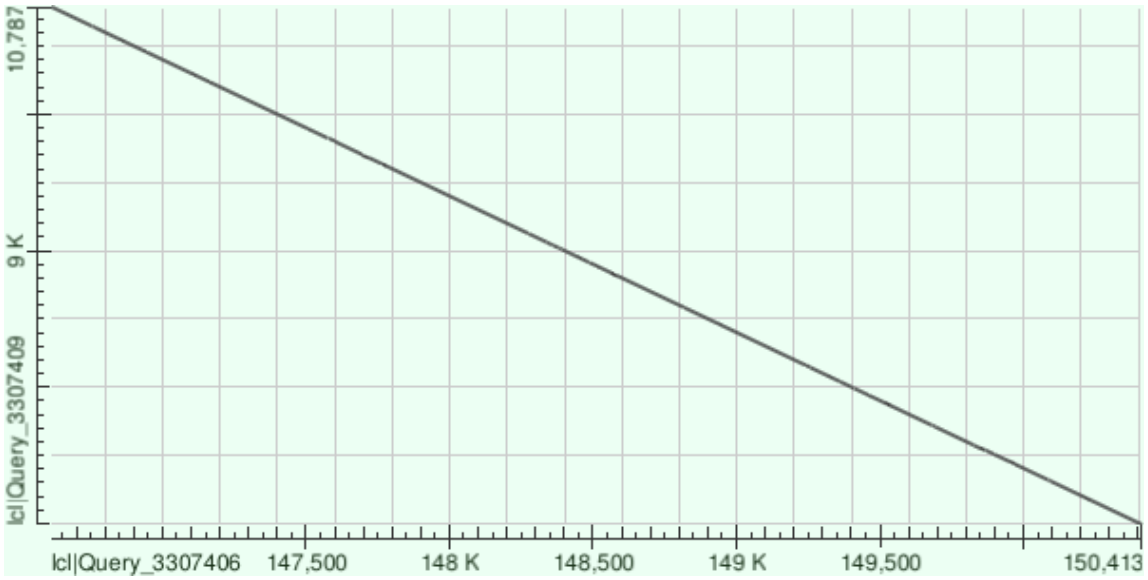

Рис.3 DotPlot для выравнивания контига NODE_15 с хромосомой Buchnera aphidicola str. Bp

Характеристика выравниваваний контига 15 на хромосому NC_004545.1:

query acc.ver, subject acc.ver, % identity, alignment length, mismatches, gap opens, q. start, q. end, s. start, s. end, evalue, bit score
NC_004545.1 NODE_15_length_16745_cov_20.901762 74.746 3845 877 79 146614 150413 10787 6992 0.0 1635

Контиг 15: выравнялся на прямую цепь хромосомы на координатах 146614-150413.

На основании выравнивания наиболее протяжённых контигов можно заключить, что сборка генома *de novo* была проведена в целом вполне успешно, и его структуру удалось восстановить, по крайней мере, частично. Из дотплотов видно, что все контиги, кроме 57, не имеют инверсий, дупликаций и инделей. В дотплоте для контига 57, к сожелению, видны множественные делеции.