Сборка de novo(практикум 14). Домашнее задание

Выполнил Муравлев Артём

1)Скачивание чтений

Чтения(SRR4240360.fastq.gz) генома бактерии Buchnera aphidicola для анализа были скачаны с помощью команды
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR424/001/SRR4240360/SRR4240360.fastq.gz

В директорию на кодомо /mnt/scratch/NGS/muravlev/pr14 , в которой и дальше производилась работа.

2)Удаление адаптеров

Для дальнейшего картирования чтений нужно сначала избавиться от адаптеров в них. Это было сделано с помощью программы trimmomatic, перед этим для удобства все последовательности адаптеров были конкатенированы в один файл

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

TrimmomaticSE -threads 10 SRR4240360.fastq.gz SRR4240360_without_adapters.fastq.gz ILLUMINACLIP:adapters.fasta:2:7:7 2> adlogs.txt

3)Обрезание чтений

Для повышения качества сборки с правых концов чтений были удалены нуклеотиды с качеством ниже 20, также были удалены чтения короче 32 нуклеотидов. Это также было сделано с помощью trimmomatic

TrimmomaticSE -threads 10 SRR4240360_without_adapters.fastq.gz SRR4240360_withoutadap_trimmed.fastq.gz TRAILING:20 MINLEN:32 2> trimmlogs.txt

4)Подготовка k-меров

Для картирования чтений дальше необходимо было подготовить k-меры. Для этого использовалась программа velveth, длина k-меров бралась максимально возможной при данной длине чтений(31)

velveth kmers/ 31 -fastq.gz -short SRR4240360_withoutadap_trimmed.fastq.gz

5)Сборка на основе k-меров

После подготовки k-меров можно было запускать сборку. Это было сделано с помощью программы velvetg

velvetg kmers

Статистика по результатам представлена в файле stats.txt
Контиги представлены в файле contigs.fa

6)Анализ результатов

N50 = 43070(из выдачи программы в stdout)
Всего 678075 контигов
Были найдены три самых крупных контига по длине(результаты в Таблице 1). Для этого приведенной ниже командой были найдены три наибольших длины контига, далее на собственном компьютере с помощью Visual Studio Code были найдены данные контиги по длине(это использовалось также в пункте 7, для копирования последовательностей данных контигов). Длина указывалась в bp.

cut -f2 stats.txt | sort -h | tail -3

ID	Length	Coverage
1	113474	33.525459
4	64155	35.847324
5	83603	33.646065

Таблица 1. Информация о трех самых длинных контигах сборки

С помощью указанной ниже команды было посчитано число контигов сборки(253).

grep '>' contigs.fa | wc -l

Медиана была найдена с помощью команды(116 контиг - медиана)

grep '>' contigs.fa| tr '_' '\t' | sort -k6 -n | head -n 127 | tail -n 1

Результат - медианное покрытие равно 6.355191

С помощью следующей команды была получена отсортированная информация о покрытиях контигов

grep '>' contigs.fa | tr '_' '\t' | sort -k6 -n

От медианы в 5 и более раз отличались значения 1.612903 и от 31.852512 до 109.391304

Три наиболее длинных контига попали в категорию аномалий по покрытию. Случай довольно неожиданный, но, все же, продолжим работу с ними.

7)Выравнивание трех наиболее длинных контигов на хромосому Buchnera aphidicola

С помощью алгоритма megablast на сайте NCBI были выравнены последовательности трех наибольших контигов с хромосомой Buchnera aphidicola(CP009253). Результаты представлены ниже.

7.1) 1-й контиг

На Рис.1 можно ознакомиться с dotplot выравнивания. Видно два крупных и несколько более мелких разрывов в контиге. Контиг(все его выравнявшиеся части) выравнялся на участок хромосомы с координатами 449411 - 555905.

Всего выделено 15 гомологичных участков. Гэпы были в диапазоне 0-4%.

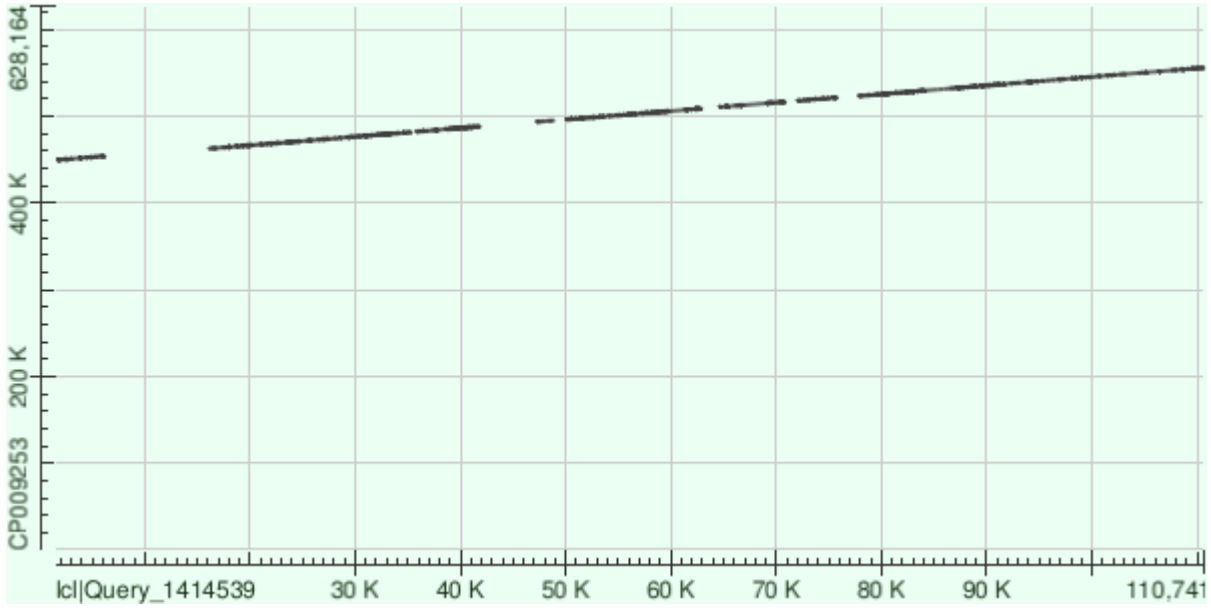

Рис.1 . Выравнивание 1-го контига на CP009253

7.2) 4-й контиг

На Рис.2 можно ознакомиться с dotplot выравнивания. Видно один крупный и несколько более мелких разрывов в контиге. График смещен вверх, то есть была инсерция(в хромосоме)/делеция(в контиге). Также видно, что кольцевую хромосому порезали в области, гомологичной контигу. Контиг(все его выравнявшиеся части) выравнялся на участки хромосомы с координатами 2004 - 32745; 599832 - 627104.

Всего выделено 12 гомологичных участков. Гэпы были в диапазоне 0-4%.

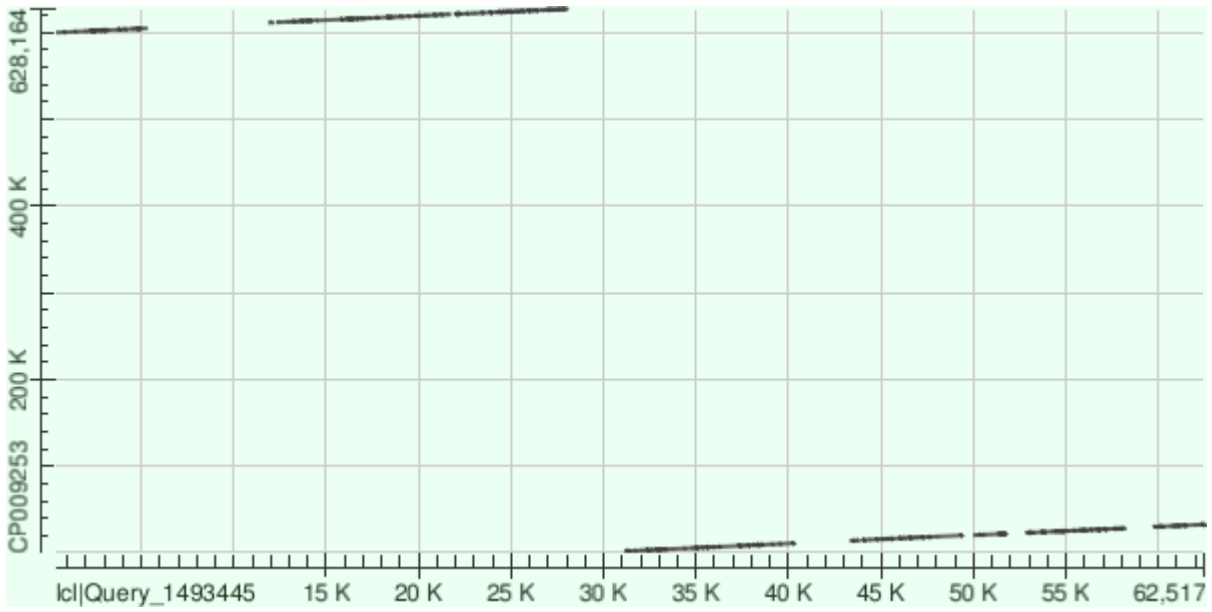

Рис. 2 . Выравнивание 4-го контига на CP009253

7.3) 5-й контиг

На Рис.3 видно один очень крупный и несколько более мелких разрывов в контиге. Контиг(все его выравнявшиеся части) выравнялись на участок хромосомы 98408 -173180.

Всего было выделено 8 гомологичных участков. Гэпы были в диапазоне 0-4%.

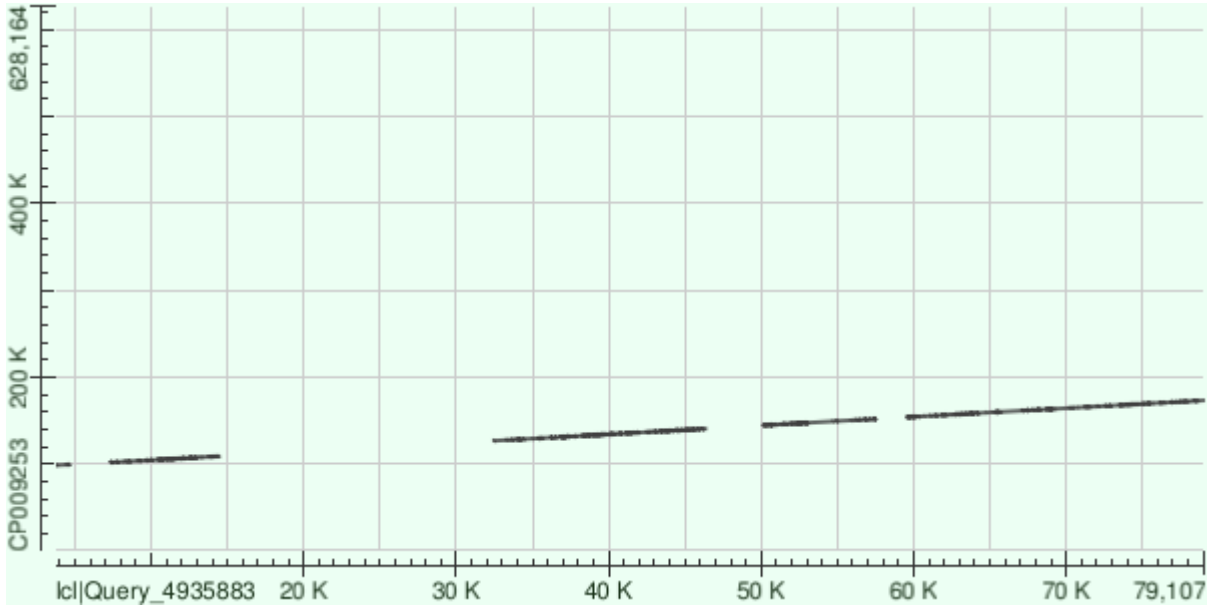

Рис. 3 . Выравнивание 5-го контига на CP009253