Практикум 14

Сборка de novo

Мой код доступа проекта по секвенированию бактерии Buchnera aphidicola str. Tuc7: SRR4240380.

Для начала я создала рабочую поддиректорию «pr14» для этого практикума и скачала туда архив с чтениями с помощью команды:

wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR424/000/SRR4240380/SRR4240380.fastq.gz

1. Подготовка чтений программой trimmomatic

Сначала я подготовила файл со всеми адаптерами с помощью команды cat, а затем удалила возможные остатки адаптеров с помощью команды:

TrimmomaticSE -phred33 SRR4240380.fastq.gz noad.fastq.gz ILLUMINACLIP:adapters.fasta:2:7:7

Выдача:

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

где 5217318 – количество чтений на вход; 5119144 (98.12%) – чтений осталось; 98174 (1.88%)– чтений удалилось.

После этого я удалила с правых концов чтений нуклеотиды с качеством ниже 20, оставив только такие чтения, длина которых не меньше 32 нуклеотидов. Для этого я воспользовалась командой:

TrimmomaticSE -phred33 noad.fastq.gz noad_trimmed.fastq.gz TRAILING:20 MINLEN:32

Выдача:

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

где 5119144 – количество чтений на вход, 4865359 (95.04%) – чтений осталось, 253785 (4.96%) – чтений удалилось.

Размер файла с изначальными чтениями: 108 Мб

Размер итогового файла: 99 Мб

2. Velveth и подготовка К-меров

В этом задании я запустила программу velveth, чтобы она на основе моего файла подготовила k-меры длины k=31 (максимально возможной), с помощью команды:

velveth Assem 31 -short -fastq.gz noad_trimmed.fastq.gz

3. Velvetg (сборка на основе k-меров)

Для сборки на основе k-меров я запустила команду “velvetg Assem”, её выдача:

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

откуда N50: 12 042.

Дальше я перешла в папку Assem, тогда с помощью следующей команды можно узнать три самых длинных контига и их покрытие:

less contigs.fa | grep '>'| tr '_' '\t'| sort -k4 -n -r |head -3

Выдача:

>NODE 3 length 25915 cov 27.418676

>NODE 20 length 23850 cov 24.763815

>NODE 23 length 23807 cov 25.725922

Контиги с аномально большим покрытием я нашла с помощью команды:

less contigs.fa | grep '>'| tr '_' 't'| sort -k6 -n|tail -3

Выдача:

>NODE 75 length 501 cov 86.361275

>NODE 11 length 2106 cov 126.008545

>NODE 56 length 934 cov 130.479660

Контиги с аномально маленьким покрытием я нашла с помощью команды:

less contigs.fa | grep '>'| tr '_' '\t'| sort -k6 -n |head -3

Выдача:

>NODE 235 length 62 cov 2.419355

>NODE 110 length 80 cov 2.700000

>NODE 301 length 63 cov 2.888889

4. Анализ

В этом задании я сравнила программой megablast каждый из трех самых длинных контигов с хромосомой Buchnera aphidicola (GenBank/EMBL AC —CP009253).

Контиг 3:

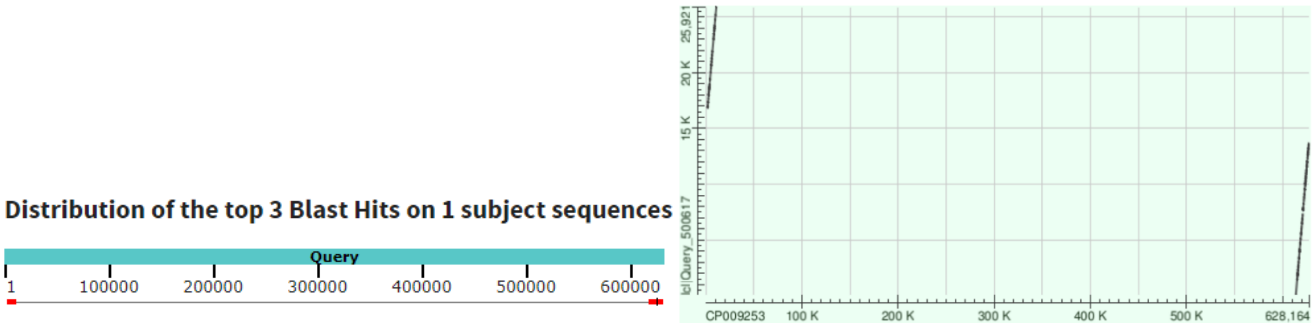

Рис 1. Расположение контига 3 на хромосоме и Dot Plot соответственно

Видно, что контиг выровнялся на участках хромосомы тремя частями: 2к- 11к, 613к – 627к

Контиг 20:

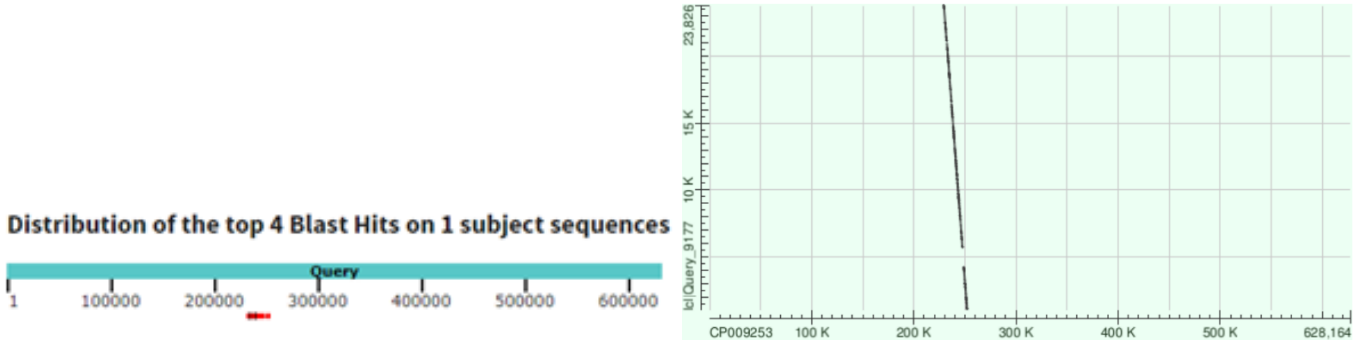

Рис 2. Расположение контига 20 на хромосоме и Dot Plot соответственно

Контиг 20 выровнялся на участках хромосомы четырьмя частями: 236к – 252к

Контиг 23:

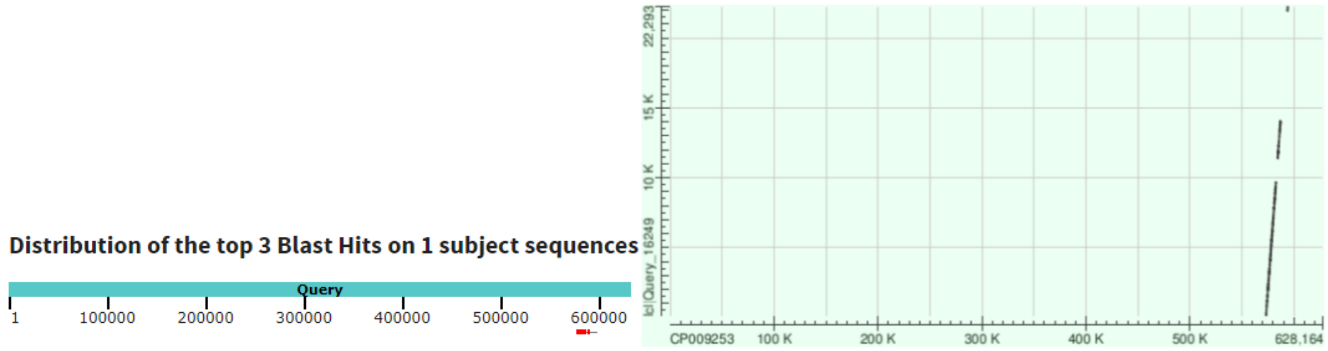

Рис 3. Расположение контига 23 на хромосоме и Dot Plot соответственно

Контиг 23 выровнялся на участках хромосомы тремя частями: 573 к – 594к.