Практикум 14

Кушнарева Дарья 202 ПЗ

Данные: SRR1724089

Работа проводилась в директории /mnt/scratch/NGS/d.kushnareva/pr14
Со страницы https://www.ebi.ac.uk/ena/browser/view/SRR1724089 был скачан файл SRR1724089.fastq.gz.

```
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR172/009/SRR1724089/SRR1724089.fastq.gz
ls -lh
>-rw-r--r--. 1 d.kushnareva year-23 943M Apr 16 13:16 SRR1724089.fastq.gz
zcat SRR1724089.fastq.gz | wc -l
>57804360
```

1. Подготовка чтений программой trimmomatic

Перед тем как подготовить чтения программой trimmomatic соберем все адаптеры лежащие в /mnt/scratch/NGS/adapters/ в один файл adapters.fasta

```
cat /mnt/scratch/NGS/adapters/*.fa > adapters.fasta
```

Далее подготовим чтения и проанализируем исходные и полученные файлы.

```
TrimmomaticSE -phred33 SRR1724089.fastq.gz trimmedSRR1724089.fastq.gz
 ILLUMINACLIP:adapters.fasta:2:7:7 TRAILING:20 MINLEN:32
>Input Reads: 14451090 Surviving: 13951662 (96.54%) Dropped: 499428 (3.46%)
```

SE - опция для одиночных чтений, TRAILING:20 удаляет чтения качеством ниже 20, MINLEN:32 удаляет чтения длиной меньше 32

```
fastqc  SRR1724089.fastq.gz
fastqc  trimmedSRR1724089.fastq.gz
```

Исходный файл (размер: 943M) содержал 14451090 чтений. В полученном файле (размер: 891M) чтений 13951662. Было оставлено 96,54% чтений, и соответственно удалено 3,46%.
Благодоря триммированию чтений улучшилось качество концов, но в целом малое количество удаленных чтений говорит о хорошем качестве изначальных данных.

2. Сборка k-меров длины 31

Для того чтобы собрать k-меры длины 31 воспользуемся программой velveth. Полученные k-меры будут лежать в папке kmere/.

```
velveth kmere 31 -short -fastq.gz trimmedSRR1724089.fastq.gz
```

Опция 31 - длина (это максимально возможная длина при одиночных чтениях длины 101), -short означает что чтения не парные.

3. Сборка на основе k-меров

Для того чтобы сделать сборку на основе k-меров воспользуемся velvetg

```
velvetg kmere
```

N50 = 71, максимальная длина контига - 2293, общая длина контигов - 7845093.

Теперь рассмотрим три самых длинных контига:

Контиг	Длина	Покрытие
NODE_52608_length_2293_cov_15.692542	2293	15.692542
NODE_77357_length_1771_cov_7.328628	1771	7.328628
NODE_77322_length_1595_cov_19.445768	1595	19.445768

Их среднее покрытие 14,16. Значит контигами с аномально высоким покрытием будут контиги с покрытием больше 70,8, а с аномальным низким будут контиги с покрытием меньше 2,83 (считаем что покрытие будет аномальным если отличается в пять раз от среднего покрытия трех самых длинных контигов).
Рассмотрим контиги с аномально высоким покрытием:

Контиг	Длина	Покрытие
NODE_1773_length_160_cov_346.625000	160	346.625000
NODE_2625_length_86_cov_460.476746	86	460.476746
NODE_4980_length_41_cov_428.585358	41	428.585358

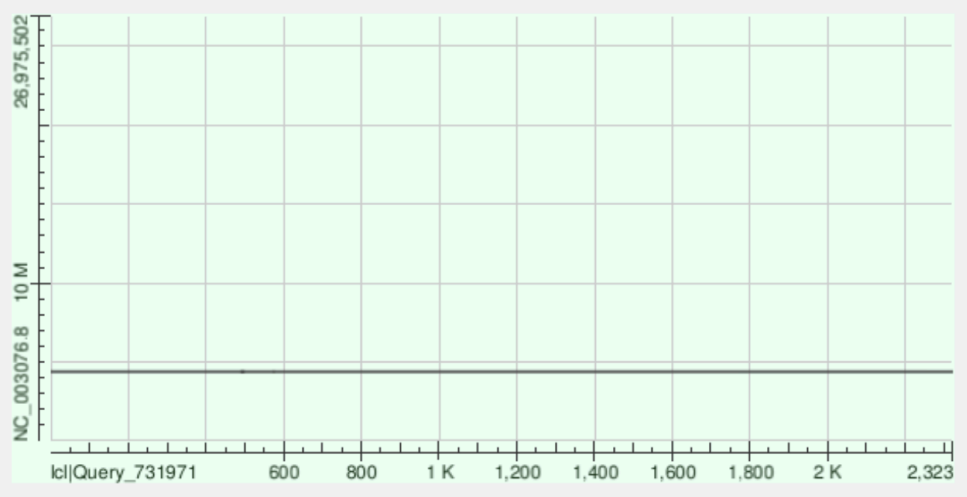
Рассмотрим контиги с аномально низким покрытием:

Контиг	Длина	Покрытие
NODE_1774_length_179_cov_2.067039	179	2.067039
NODE_3034_length_69_cov_2.000000	69	2.000000
NODE_4023_length_71_cov_1.591549	71	1.591549

4. Анализ трех самых длинных контигов

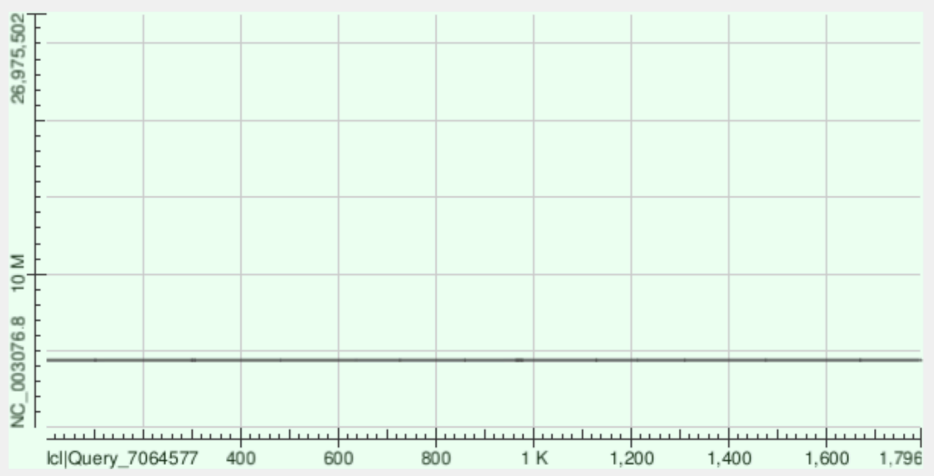
Проанализируем три самых длинных контига, для этого запустим megablast по банку "RefSeq Genome Database", ограниченному на вид *Arabidopsis thaliana*.

a. Контиг NODE_52608
   Длина: 2293; Покрытие: 15.692542
   Результат: выравнивание на пятую хромосому(**NC_003076.8**), 3 участка, все без гэпов, процент идентичности 100%-99%. Контиг почти полностью выровнялся на хромосому
   Координаты выравнивания: 4388653 – 4391169



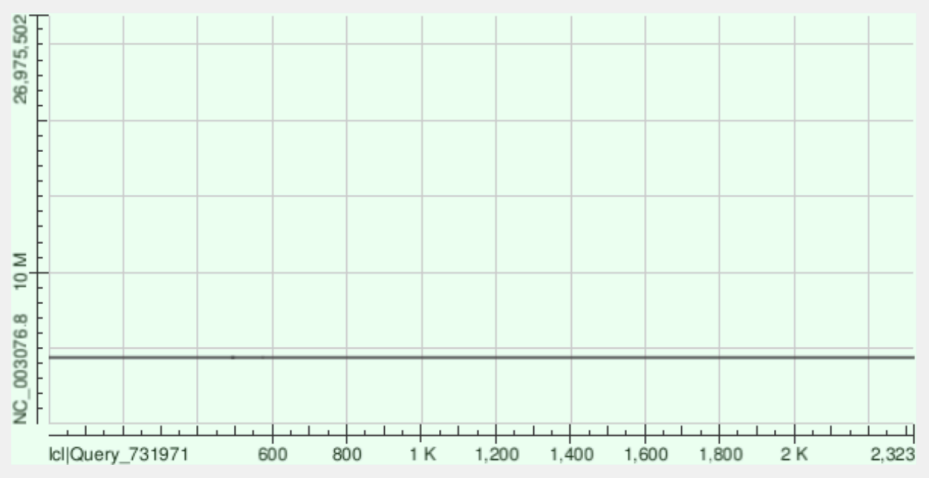
На хромосоме этот участок соответствует гену GUN5, который кодирует Mg-протопорфирин IX хелатазу (субъединицу CHLH). Данный ген определен на основе гомологии. Mg-протопорфирин IX хелатаза играет ключевую роль в биосинтезе хлорофилла, катализируя первую стадию — вставку иона Mg^{+} в протопорфирин IX. Кроме того, ChlH/GUN5 участвует в ретроградном сигналинге (сигнал от хлоропластов к ядру)

b. Контиг NODE_77357
Длина: 1771; Покрытие: 7.328628
Результат: выравнивание на пятую хромосому(**NC_003076.8**), 13 участков, гэпов не более двух, процент идентичности 100%-98%. Контиг почти полностью выровнялся на хромосому
Координаты выравнивания: 4398423 – 4400892



На хромосоме этот участок соответствует гену SVR3, который кодирует белок семейства факторов элонгации. Семейство факторов элонгации это группа белков, обеспечивающих синтез белка, способствуя удлинению пептидной цепи на рибосоме во время трансляции.

c. Контиг NODE_77322
Длина: 1595; Покрытие: 19.445768
Результат: выравнивание на четвертую хромосому(**NC_003075.7**), 2 уастка, оба без гэпов, процент идентичности 100%-99%. Контиг почти полностью выровнялся на хромосому
Координаты выравнивания: 2989843 - 2994784



На хромосоме на этом участке находится локус AT4G05631, который является неаннотированным блок кодирующим участком.