Сборка de novo

1. Подготовка чтений программой trimmomatic.

На первом шаге получения сборки *de novo* из файла с чтениями SRR1724091.fastq.gz, с помощью программы Trimmomatic, были удалены остатки адаптеров:

```
TrimmomaticSE -phred33 -threads 4 SRR1724091.fastq.gz
SRR1724091_trimmed_adapters.fastq.gz
ILLUMINACLIP:adaptersSE_united.fasta:2:7:7
```

Здесь: adaptersSE_united.fasta - файл с последовательностями одноконцевых адаптеров. Всего было удалено 2329(0,02%) чтений из 15263296.

Далее, с правых концов чтений были удалены нуклеотиды с качеством ниже 20 и оставлены только чтения с длиной не менее 32:

```
TrimmomaticSE -phred33 -threads 4 SRR1724091_trimmed_adapters.fastq.
gz SRR1724091_trimmed_20.fastq.gz MINLEN:32 TRAILING:20
```

Удалено 254175(1.67%) чтений, итоговый размер файла - 922M (до триммирования - 949M).

2. Получение k-меров для сборки

На следующем шаге, с помощью программы velveth, были получены k-меры для создания сборки с помощью графа де-Брёйна:

```
velveth kmers 31 -fastq.gz -short SRR1724091_trimmed_20.fastq.gz
```

Непосредственно сборка на основе k-меров производилась с помощью программы velvetg:

```
velvetg kmers
```

Значение N50 в выдаче программы равно 71, всего контигов 312069, что говорит о низком качестве сборки.
Контиги с наибольшей длиной:

```
>NODE_33424_length_2141_cov_14.062120
>NODE_39141_length_1312_cov_42.967224
>NODE_7507_length_1271_cov_8.489378
```

Среднее покрытие в этих наиболее длинных контигов равно 21.84. Всего контигов, покрытие которых сильно отличается от этого числа, 44974 в файле. Вот некоторые из них:

NODE_1895_length_34_cov_637.823547
NODE_1896_length_294_cov_631.581604
NODE_2931_length_93_cov_2.591398
NODE_4255_length_259_cov_3.146718

Как видно, в сборке присутствует множество контигов с самой разной длиной и покрытием. Вместе с низким значением N50 можно предположить, что такая неоднородность сборки обусловлена попаданием в пробу РНК из ядра или цитоплазмы. Особенно на это указывает наличие большого количества контигов с большой длиной и низким покрытием.

3. Анализ полученных контигов

Далее, последовательности наиболее длинных контигов были выровнены на вид *Arabidopsis thaliana* по банку RefSeq Genome Database.

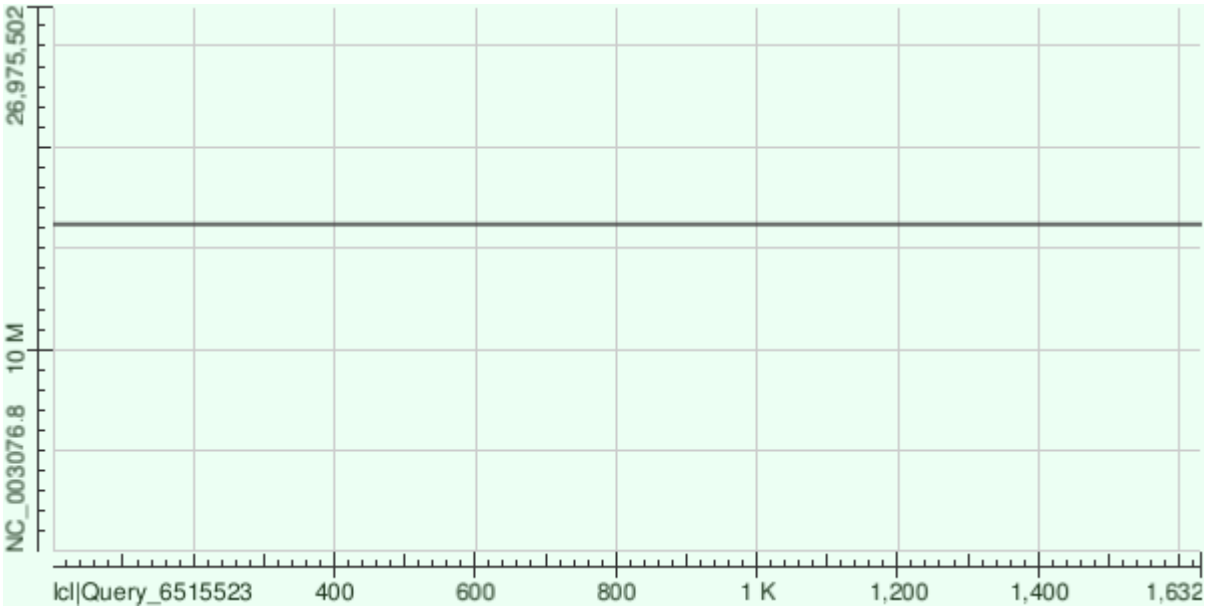

Рис. 1. Карта локального сходства наиболее длинного контига (NODE_33424) на участке 16185780 - 16187411. Карты локального сходства для трех других выравниваний имеют тот же вид.

По итогу выравнивания наиболее длинного с *Arabidopsis thaliana*, получилось 4 выравнивания с 5 хромосомой. Несмотря на наличие 4 выравниваний в результате, на самом деле контиг выровнялся только на одном участке хромосомы 5. Наличие остальных трех выравниваний обусловлено тем, что длина контига превосходит длину участка хромосомы, и поэтому остальные выравнивания соответствуют участкам того же части хромосомы.
Карта локального сходства и параметры выравнивания показывают высокое сходство контига с выровненым участком: идентичность 99%, гэпы 0, score 4019, e-value 0.

На этом участке расположен ген At5g40450, участвующий в процессах онтогенеза растений.

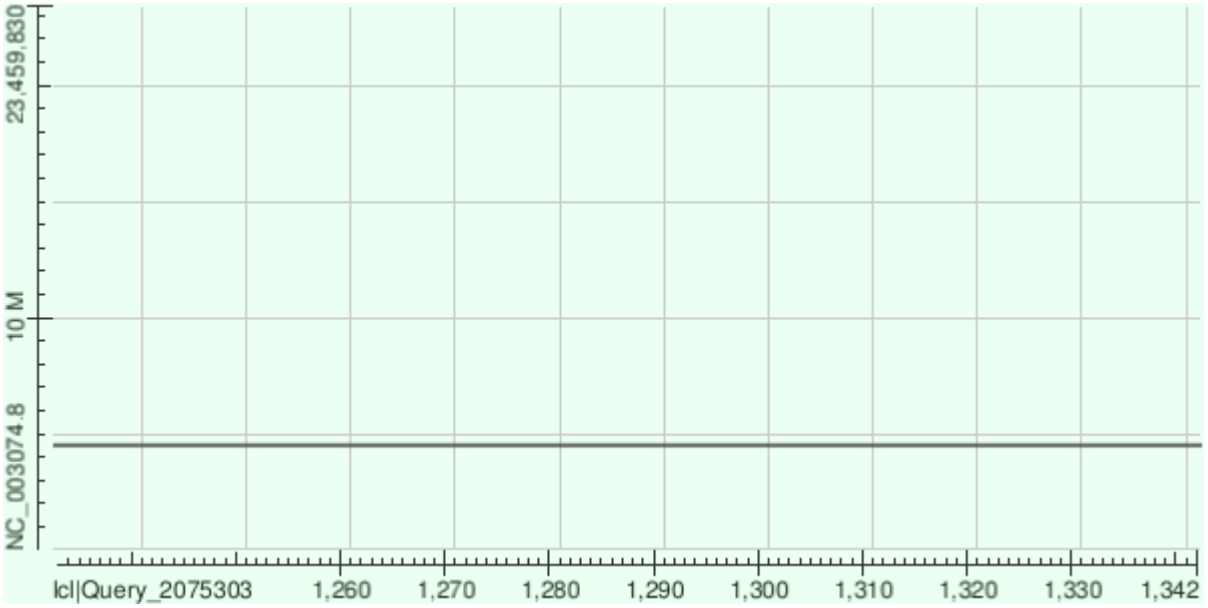

Рис. 2. Карта локального сходства наиболее длинного контига (NODE_39141) на участке 4505686 - 4505795.

Второй по длине контиг выровнялся с участком 3 хромосомы от 4505686 до 4505795. Ген AT3G13740, находящийся на этом участке хлоропластную рибонуклеазу, участвующую в метаболизме РНК.
Параметры выравнивания: score 204, e-value 5e-51, идентичность 100%, гэпы 0.

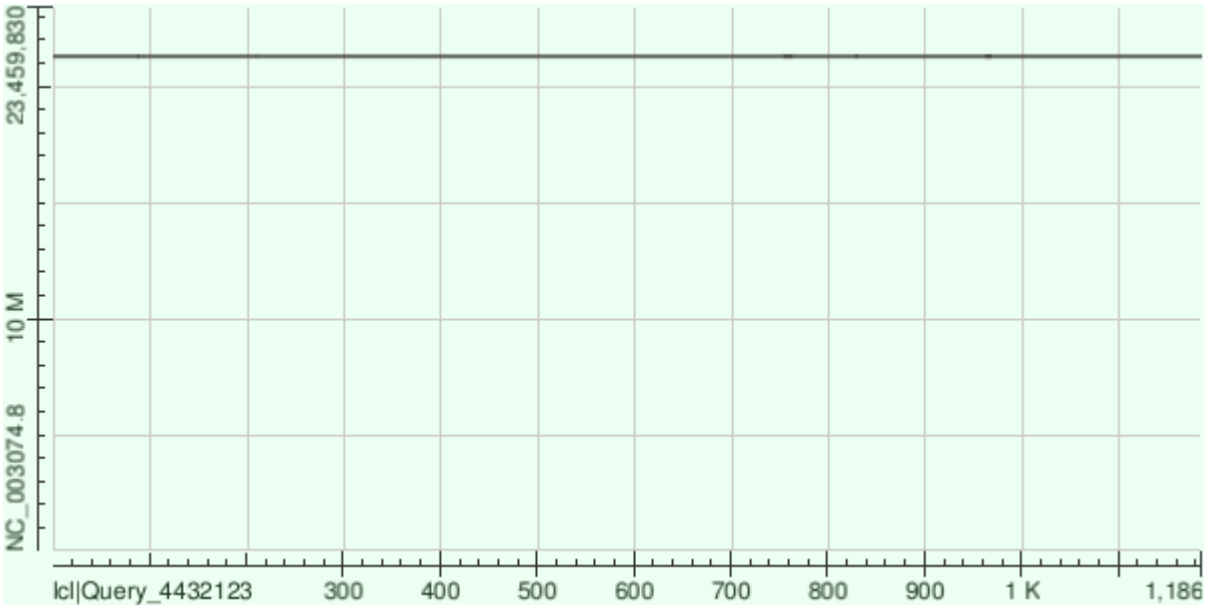

Рис. 2. Карта локального сходства наиболее длинного контига (NODE_7507) на участке 21287601-21293602

Для последнего контига получилось 8 выравниваний, с тем же геном из предыдущего пункта (AT3G13740). Скорее всего , такой результат обусловлен тем, что контиг выровнялся на экзомные участки этого гена (координаты выравнивающих участков совпадают с координатами экзонов в этом гене).
Параметры выравнивания: score 2215, e-value 0, идентичность 100%, гэпы 0.