Сборка de novo

В данном практикуме у нас нет референсного генома для картирования чтений. Задача заключается в де-ново сборке коротких чтений в длинные последовательности (контиги).

Рабочая директория: **/mnt/scratch/NGS/gaponksenia/pr14**

С помощью команды

```
wget
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR424/001/SRR4240361/
SRR4240361.fastq.gz
```

был скачан архив с чтениями бактерии *Buchnera aphidicola* str. Tuc7 **SRR4240361.fastq.gz**

1. Подготовка чтений программой trimmomatic

Объединяем для удобства все адаптеры в один файл командой:

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

Удаляем все возможные остатки адаптеров с помощью команды:

```
TrimmomaticSE -phred33 SRR4240361.fastq.gz
SRR4240361_adapters_removed.fastq.gz
ILLUMINACLIP:combined_adapters.fasta:2:7:7
```

SE - чтения одноконцевые
-phred33 - тик кодирования качества
ILLUMINACLIP - обрезаем адаптеры
2 - допустить 2 ошибки при поиске адаптера
первое 7 - насколько строго искать палиндромные адаптеры
второе 7 - если нашли совпадение большее либо равное 7 - обрезать

Часть логов выполненной команды:

```
Input Reads: 7272621 Surviving: 7238089 (99.53%)
Dropped: 34532 (0.47%)
```

То есть 0,47 % последовательностей чтений оказались остатками адаптеров.

Удаляем с правых концов нуклеотиды низкого качества и отбрасываем короткие чтения с помощью команды:

```
TrimmomaticSE -phred33
SRR4240361_adapters_removed.fastq.gz
SRR4240361_final_trimmed.fastq.gz TRAILING:20
MINLEN:32
```

TRAILING:20 - удаляем нуклеотиды качеством ниже 20
MINLEN:32 - отбрасываем чтение короче 32 нуклеотидов

Часть логов выполненной команды:

```
Input Reads: 7238089 Surviving: 6834335 (94.42%)
Dropped: 403754 (5.58%)
```

То есть 5,58 % последовательностей чтений было удалено.

Размер исходного файла **SRR4240361.fastq.gz** - 193 Мб
Размер файла после удаления адаптеров
SRR4240361_adapters_removed.fastq.gz - 192 Мб
Размер конечного файла **SRR4240361_final_trimmed.fastq.gz** - 178 Мб

2. Подготовка k-меров

Подготовим список k-меров длинной 31 нуклеотид из наших коротких и непарных чтений с помощью команды:

```
velveth output_dir 31 -short -fastq.gz
SRR4240361_final_trimmed.fastq.gz
```

3. Сборка на основе k-меров

Осуществим непосредственно сборку контигов с помощью команды:
`velvetg output_dir`

Результаты сборки (контиги) в файле output_dir/contigs.fa.
Информацию о длине и покрытии каждого контига смотрим в строках заголовка, которые начинаются с символа >.

В логе найдем N50:

```
Final graph has 477 nodes and n50 of 25683, max
49238, total 668902, using 0/6834335 reads
```

N50 сборки (по velvetg) = 25683

С помощью вот такого вот конвейера ищем самые длинные контиги и их покрытие:

```
grep "^>" output_dir/contigs.fa | tr '_' '\t' | cut
-f2,4,6  | sort -k2,2nr | head -n 3
```

Выдача:

```
6    49238    26.660851
2    45555    26.450466
34   43866    23.514977
```

Второй столбец - длина, третий - покрытие.

Конвейером пронумеруем все строки и найдем медианное покрытие:

```
grep "^>" output_dir/contigs.fa | tr '_' '\t' | cut
-f2,4,6  | sort -k3,3nr | nl -ba
```

Медианное покрытие контигов составило ~10.7×.
Аномальными считались покрытие <2.1× или >53.5× (±5 раз от медианы).

Было найдено 6 контикогов с аномально высоким покрытием:

1	78	47	90.744682
2	91	33	76.636360
3	95	31	64.903229
4	185	48	62.541668
5	47	2655	56.362713
6	143	408	56.284313

Такое аномально высокое покрытие, вероятнее всего, связано с повторяющимися или мультикопийными участками генома, которые в процессе сборки были слиты в один контиг, поэтому все риды из нескольких копий попали на одну.

4. Анализ

В NCBI Nucleotide нашли полный геном *Buchnera aphidicola* str. Sg (NC_004061.1)

1) Выравнивание контига NODE_2 и NC_004061.1. Координаты на хромосоме, соответствующие контигу 472743 - 484952. Число однонуклеотидных различий - 8. Число гэпов - 295. Доля контига, выровненная на геном - 82%.
   Dot-plot для контига NODE_2 и хромосомы NC_004061.1 показывает один длинный диагональный участок, соответствующий локальному выравниванию почти по всей длине контига, с небольшими разрывами. Это означает, что контиг ложится на хромосому как один линейный фрагмент без инверсий и крупных перестроек; разрывы соответствуют небольшим гэпам/участкам без выравнивания.

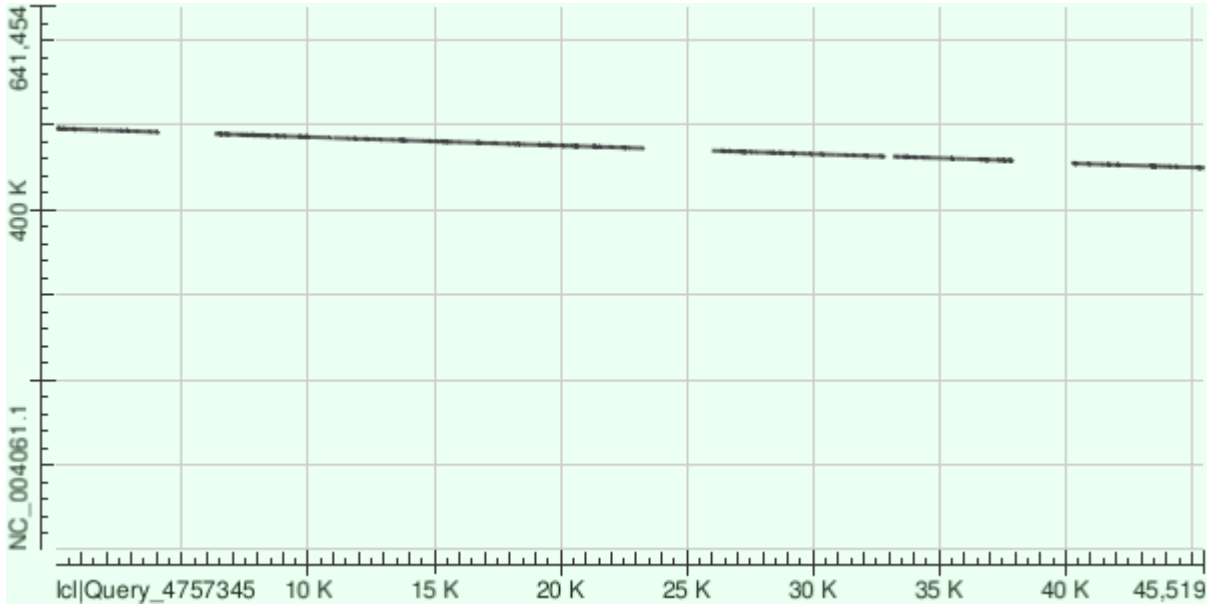

Рис. 1 Карта локального сходства NODE_2 и NC_004061.1

2) Выравнивание контига NODE_6 и NC_004061.1. Координаты на хромосоме, соответствующие контигу 165186-176511. Число однонуклеотидных различий - 6. Число гэпов - 242. Доля контига, выровненная на геном - 75%.
   На карте локального сходства NODE_6 и хромосомы NC_004061.1 виден один почти непрерывный диагональный трек без инверсий.

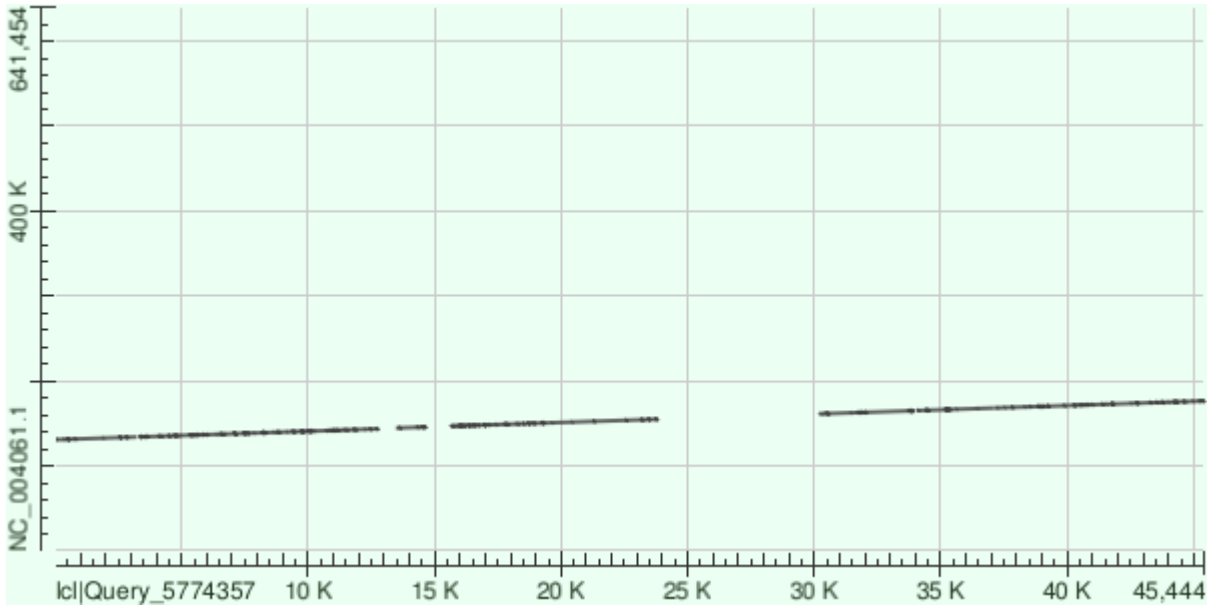

Рис. 2 Карта локального сходства NODE_6 и NC_004061.1

3) Выравнивание контига NODE_34 и NC_004061.1. Координаты на хромосоме, соответствующие контигу 275325 - 280685. Число однонуклеотидных различий - 11. Число гэпов - 172. Доля контига, выровненная на геном - 70%.
   Dot-plot для контига NODE_34 и хромосомы NC_004061.1 показывает один протяжённый диагональный участок без инверсий, слегка разбитый на несколько соседних сегментов. Это означает, что NODE_34 выравнивается на геном как непрерывный линейный фрагмент хромосомы, а разбиение диагонали на куски соответствует небольшим участкам без выравнивания или гэпам в последовательности.

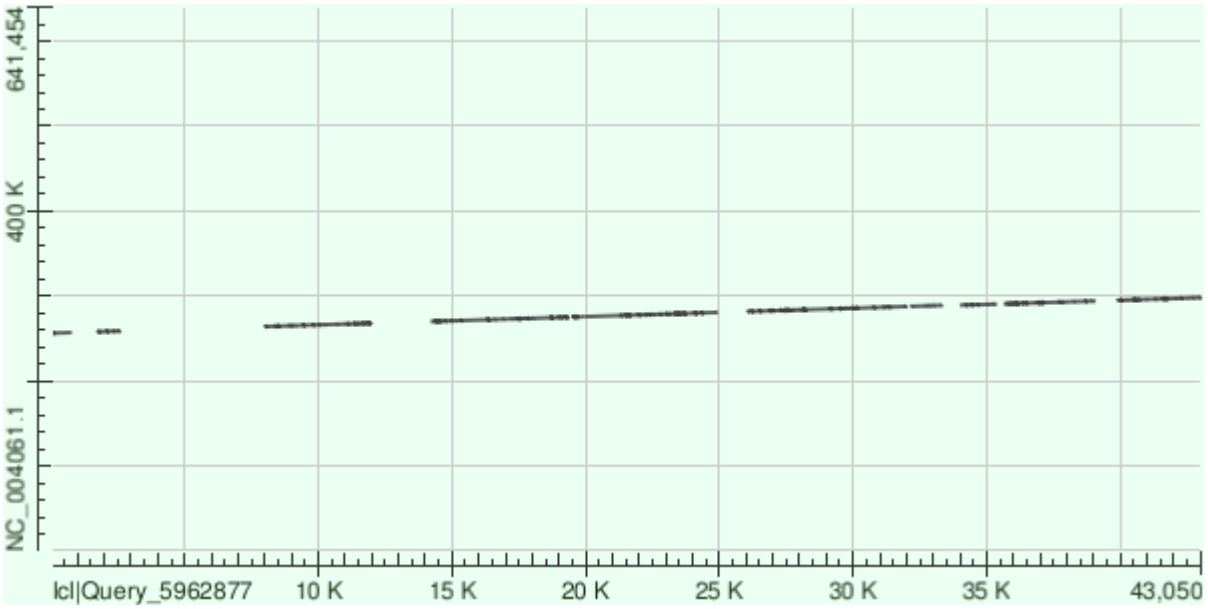

Рис. 3 Карта локального сходства NODE_34 и NC_004061.1

Три самых длинных контига (NODE_6, NODE_2 и NODE_34) хорошо выравниваются на референс-хромосому Buchnera aphidicola NC_004061.1: каждый из них даёт один протяжённый диагональный трек на dot-plot без инверсий, что соответствует линейным участкам хромосомы. Доля контигов, выровненная на геном, составляет ~75–82%, при этом идентичность выравниваний высока (около 79–80%), а число однонуклеотидных различий и гэпов невелико относительно длины выровненных фрагментов. Это говорит о том, что полученная сборка корректно воспроизводит большую часть бактериальной хромосомы