Практикум 15

Сборка de novo

Мой код доступа: SRR4240356

Команда скачивания чтений:

```
wget
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR424/006/SRR4240356/SRR42
40356.fastq.gz
```

1. Подготовка чтений программой trimmomatic

Сначала скопировала файлы адаптеров к себе в папку.

Объединение остатков адаптеров:

```
cat *fa > adapters.fasta
```

Trimmomatic:

```
  java -jar /usr/share/java/trimmomatic.jar SE -threads 10
SRR4240356.fastq.gz cut_adapt.fastq.gz
ILLUMINACLIP:adapters.fasta:2:7:7 2> logs.txt
```

```
ILLUMINACLIP: Using 2 prefix pairs, 17 forward/reverse sequences, 0 forward only
 sequences, 0 reverse only sequences
Quality encoding detected as phred33
Input Reads: 7511529 Surviving: 7358438 (97.96%) Dropped: 153091 (2.04%)
TrimmomaticSE: Completed successfully
```

Рис 1. Содержимое файла log.txt (самый низ)

Удалено 153091 чтений(2,04%), осталось 97,96%

2. Триммирование чтений

```
java -jar /usr/share/java/trimmomatic.jar SE -threads 10
cut_adapt.fastq.gz trimmed.fastq.gz TRAILING:20 MINLEN:32 2>
logs2.txt
```

```
TrimmomaticSE: Started with arguments:
 -threads 10 cut_adapt.fastq.gz trimmed.fastq.gz TRAILING:20 MINLEN:32
Quality encoding detected as phred33
Input Reads: 7358438 Surviving: 7053346 (95.85%) Dropped: 305092 (4.15%)
TrimmomaticSE: Completed successfully
```

Рис 2. содержимое файла logs2.txt

4,15% чтений было удалено, осталось 95,85%

Размеры файлов после очистки (du -h):

SRR4240356.fastq.gz: 167M

cut_adapt.fastq.gz: 164M

trimmed.fastq.gz: 155M

Явно видно, как уменьшались файлы

3. Подготовка K-меров

```
velveth meow/ 31 -fastq.gz -short trimmed.fastq.gz
```

Комментарии:

Входные данные: -fastq.gz - указывает, что файл на вход в формате .fastq.gz, файл с триммированными чтениями

Опции: 31 - длина k-меров (hash_length), -short - короткие и непарные чтения

4. Сборка на основе k-меров

```
velvetg meow/
```

Входные данные: файлы, полученные velveth

Выходные данные: 8 файлов

Из Log найдем:

```
Final graph has 286 nodes and n50 of 65554, max 111962, total 659837, using 0/7
53346 reads
```

N50 = 65554

Max = 111962

Найдем 3 самых длинных контига:

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

```
ID       lgth      out     in      long_cov     short1_cov
10       80939   2       1       0.000000        37.524173
6        107488  0       0       0.000000        34.174029
8        111962  0       1       0.000000        38.660197
```

Теперь проверим, есть ли аномальные покрытия (cut -f6 stats.txt | sort -h):

Да, есть: аномально маленькие от 1 до 5 - много, и аномально большие вплоть до 1134.000000 и 266951.000000

4. Анализ

Достала 3 файла с самыми длинными контигами и положила их в папку de novo:

```
seqretsplit -filter contigs.fa dir/name.format

cp node_10_length_80939_cov_37.524174.fasta ../

cp node_6_length_107488_cov_34.174030.fast ../

cp node_8_length_111962_cov_38.660198.fasta ../
```

Контиг 1 (ID 10)

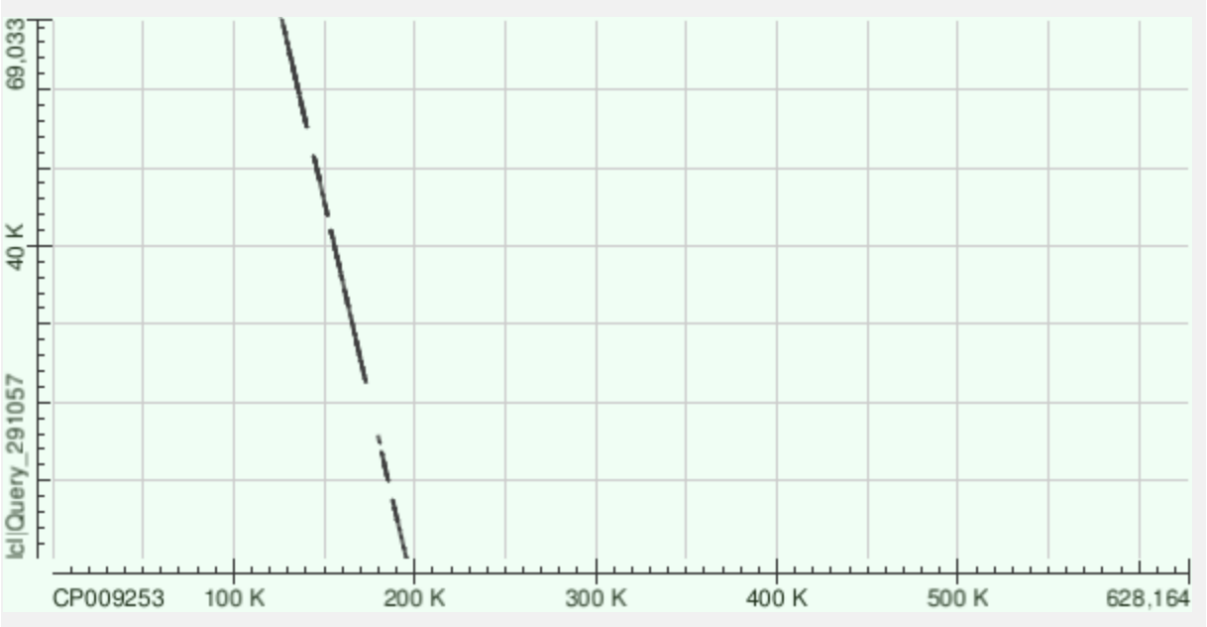

Рис 3. Dotplot 10 контига (заметим крупные делеции и то, что последовательность контига перевернута)

55035 to 67775

Score	Expect	Identities	Gaps	Strand
5421 bits(2935)	0.0	9745/13012(75%)	552/13012(4%)	Plus/Minus

33933 to 42017

Score	Expect	Identities	Gaps	Strand
4741 bits(2567)	0.0	6348/8171(78%)	270/8171(3%)	Plus/Minus

43997 to 51396

Score	Expect	Identities	Gaps	Strand
4423 bits(2395)	0.0	5865/7538(78%)	247/7538(3%)	Plus/Minus

И тд, всего 11 совпадений

Контиг 2 (ID 6)

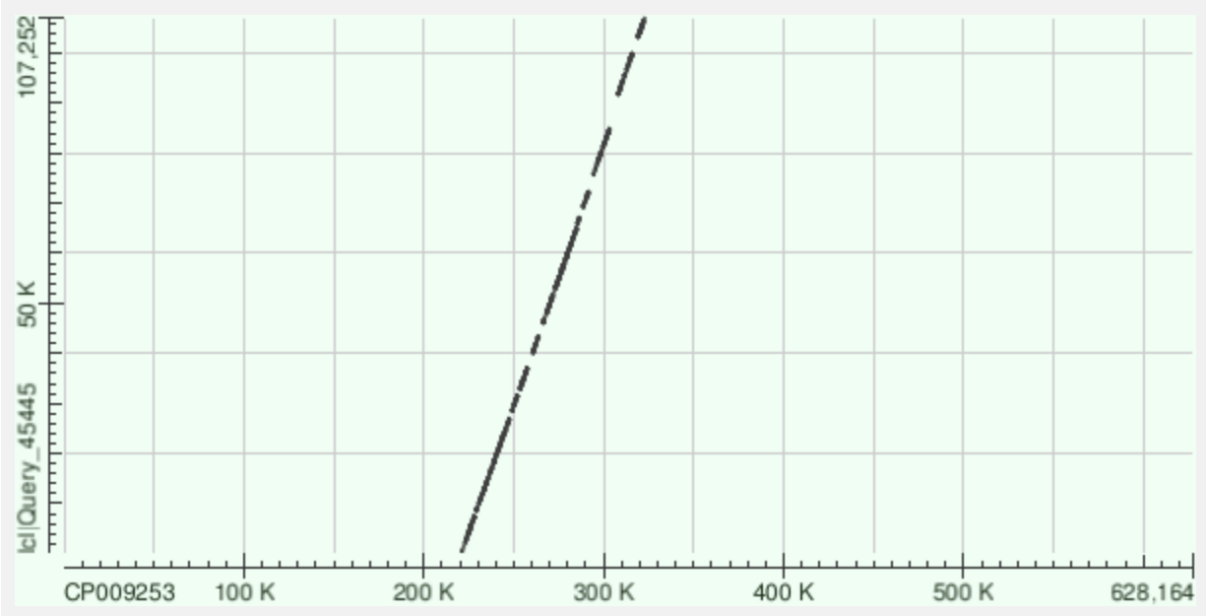

Рис 4. Dotplot 6 контига (заметим делеции и расположение относительно друг друга)

45989 to 55468:

Score	Expect	Identities	Gaps	Strand
6137 bits(3323)	0.0	7606/9658(79%)	357/9658(3%)	Plus/Plus

16292 to 26990

Score	Expect	Identities	Gaps	Strand
4771 bits(2583)	0.0	8180/10882(75%)	386/10882(3%)	Plus/Plus

77556 to 84909

Score	Expect	Identities	Gaps	Strand
3947 bits(2137)	0.0	5705/7438(77%)	204/7438(2%)	Plus/Plus

55527 to 63756

Score	Expect	Identities	Gaps	Strand
3925 bits(2125)	0.0	6376/8396(76%)	421/8396(5%)	Plus/Plus

И так далее (всего 18 совпадений)

Контиг 3 (ID 8)

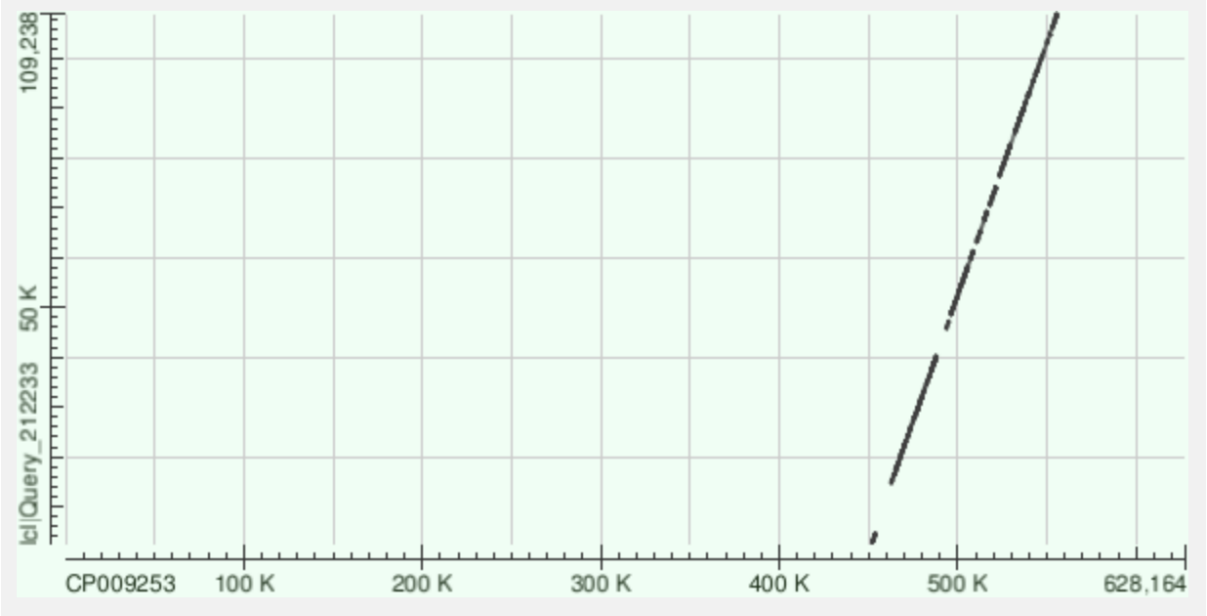

Рис 5. Dotplot 8 контига (заметим делеции и расположение относительно друг друга)

81925 to 103395

Score	Expect	Identities	Gaps	Strand
17304 bits(9370)	0.0	17694/21720(81%)	543/21720(2%)	Plus/Plus

103601 to 109238

Score	Expect	Identities	Gaps	Strand
4325 bits(2342)	0.0	4574/5657(81%)	131/5657(2%)	Plus/Plus

И тд, всего 15 совпадений