Практикум 14

1. **Подготовка чтений программой trimmomatic.**

   Сначала я скачала архив с чтениями с помощью wget:

   wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR424/009/SRR4240379/SRR4240379.fastq.gz

   Создала свой файл, в котором объединила все адаптеры:

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

   Теперь можем удалять остатки адаптеров:

   TrimmomaticSE SRR4240379.fastq.gz cleanSRR4240379.fastq.gz
   ILLUMINACLIP:adapters.fasta:2:7:7 -threads 20 -trimlog tr.log

   Input Reads: 7400155 Surviving: 7269852 (98.24%) Dropped: 130303 (1.76%)
   Видим, что 1.76% составляли остатки адаптеров

   Удаляем с правых концов чтений нуклеотиды с качеством ниже 20 и оставляем только такие чтения, длина которых не меньше 32 нуклеотидов:

   TrimmomaticSE -phred33 cleanSRR4240379.fastq.gz clean2.fastq.gz
   TRAILING:20 MINLEN:32 -threads 20 -trimlog tr2.log

   Input Reads: 7269852 Surviving: 6974267 (95.93%) Dropped: 295585 (4.07%)
   Видим, что 295585 (4.07%) чтений удалилось
   Изначально размер SRR4240379.fastq.gz равен 167M
   Размер же конечного файла clean2.fastq.gz стал равен 156M

   Далее я использовала программу velveth, чтоб она на основе моего файла подготовила k-меры длины k=31

   velveth kmer 31 -short -fastq.gz clean2.fastq.gz

   Используем программу velvetg:

velvetg kmer

N50=25646
Для поиска трех самых длинных контигов испотльзовала:

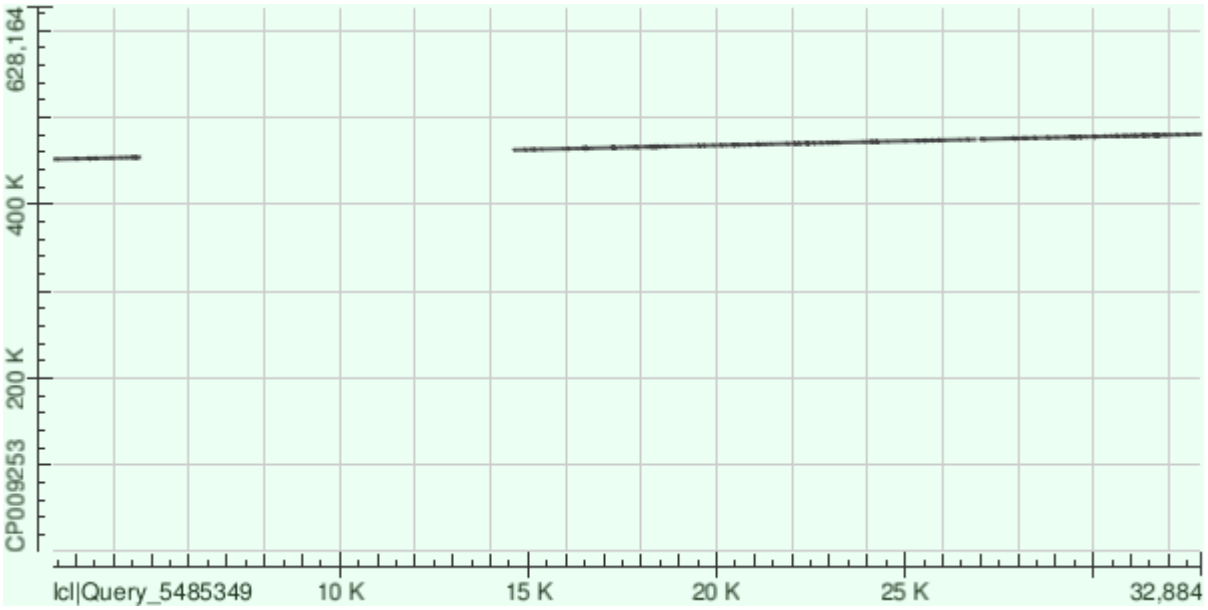
sort -k2 -nr stats.txt | less

Три самых длинных контига: 49912; 49262; 33085
Их покрытия соответственно: 35.91; 34.77; 36.26;

Для поиска аномально больших покрытий использовала команду:

sort -k6 -nr stats.txt | less

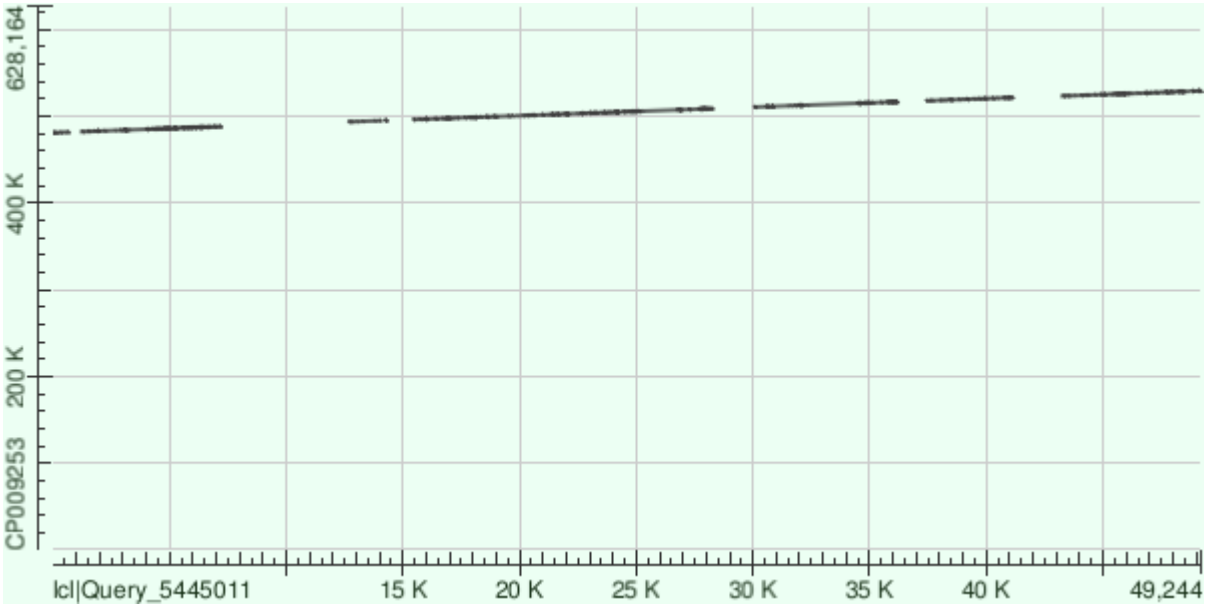
Процент покрытия этих контигов равен 474299; 2694
Длина равна 1



NODE_5_length_33085_cov_36.259029

участок	координаты участка хромосомы	число гэпов	идентичные нуклеотиды

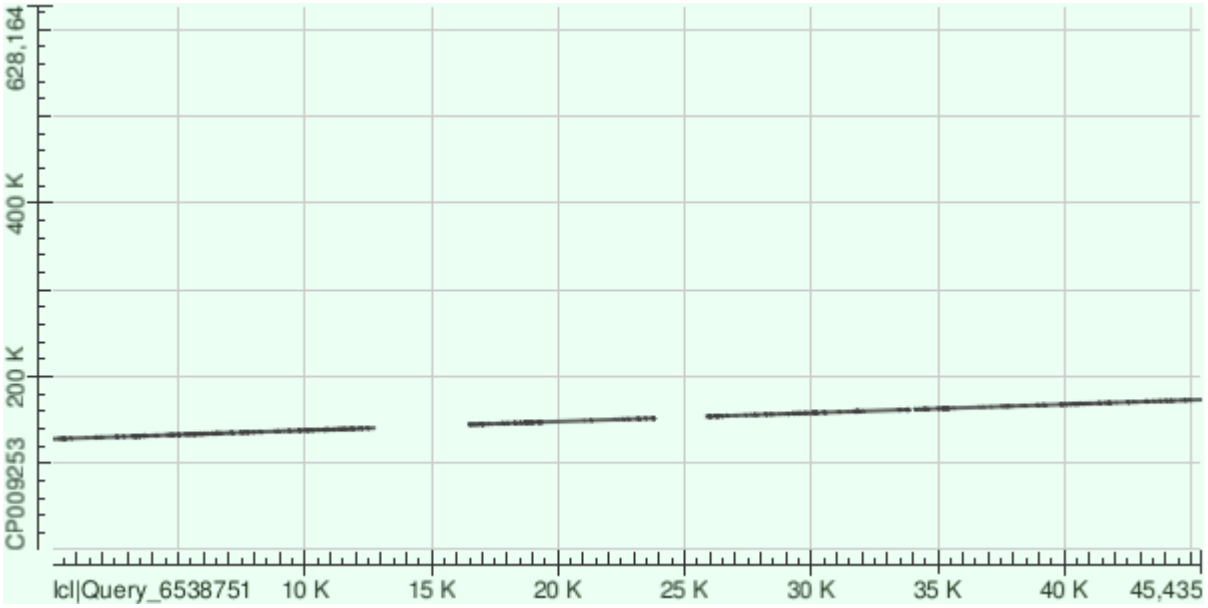
1	467412 to 474667	208/7388(2%)	5691/7388(77%)
2	462496 to 467421	162/5015(3%)	3861/5015(77%)
3	474844 to 480660	255/5974(4%)	4431/5974(74%)
4	451729 to 454069	55/2370(2%)	1827/2370(77%)



NODE_9_length_49262_cov_34.772179

участок	координаты участка хромосомы	число гэпов	идентичные нуклеотиды
10	495033 to 495148	5/120(4%)	108/120(90%)
9	528794 to 529211	26/425(6%)	357/425(84%)
8	480874 to 481545	20/686(2%)	564/686(82%)

7	493487 to 494864	13/1384(0%)	1109/1384(80%)
6	496111 to 500325	154/4324(3%)	3255/4324(75%)
5	517766 to 521500	101/3783(2%)	2922/3783(77%)
4	481997 to 488106	308/6238(4%)	4621/6238(74%)
3	523105 to 528679	207/5685(3%)	4369/5685(77%)
2	510438 to 516539	187/6234(2%)	4897/6234(79%)
1	500370 to 508806	351/8617(4%)	6516/8617(76%)



NODE_6_length_49912_cov_35.907238

участок	координаты участка хромосомы	число гэпов	идентичные нуклеотиды

1	127825 to 140555	544/13008(4%)	9741/13008(75%)
2	153752 to 161738	266/8169(3%)	6347/8169(78%)
3	144368 to 151796	243/7536(3%)	5863/7536(78%)
4	161898 to 166752	108/4912(2%)	3910/4912(80%)
5	166750 to 173180	159/6517(2%)	4965/6517(76%)