Сорокина Тамара, ДЗ 10.

1. Анализ качества сырых прочтений:

conda activate ngs
mkdir -p fastqc_raw_output
*fastqc /projects/mipt_dbmp_biotechnology/genome/illumina_reads_R1_001.fastq *
*/projects/mipt_dbmp_biotechnology/genome/illumina_reads_R2_001.fastq *
-o fastqc_raw_output

2. Trimming: bash-скрипт в файле на гитхабе

а) Удаление адаптеров с использованием файла `adapters.fa` с опцией `ILLUMINACLIP`.

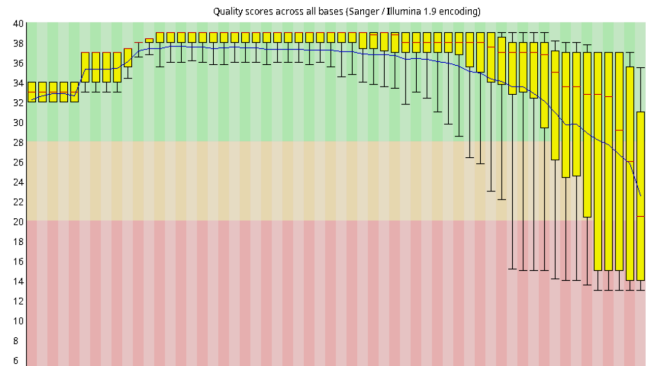
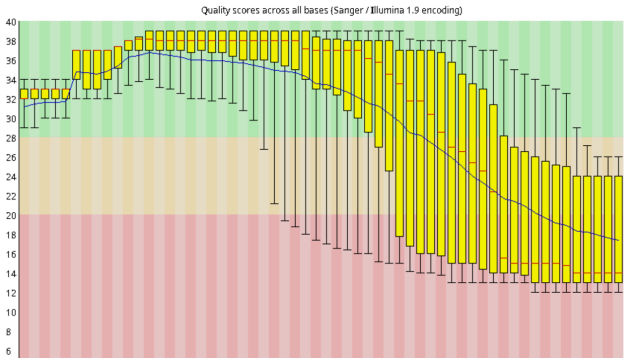
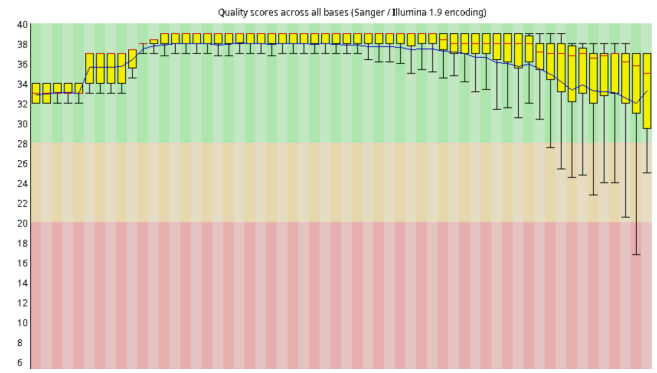
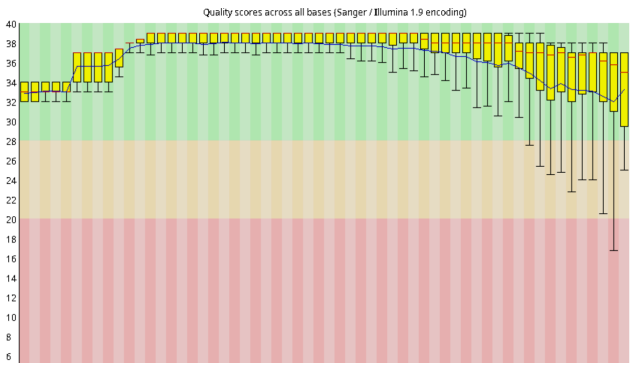
б) Обрезка низкокачественных концов ридов с использованием параметров `LEADING:20` и `TRAILING:20`.

в) Обрезка коротких ридов с использованием параметра `MINLEN:50`.

3. Анализ качества после тримминга:

*fastqc trimmed_output/reads_R1_paired.fastq *
*trimmed_output/reads_R2_paired.fastq *
-o fastqc_trimmed_output

Качество прочтений до и после тримминга представлено в таблице.

	R1	R2
raw		
trimmed		

Так же приведу basic statistics для R1 до и после тримминга:

Basic Statistics

Measure	Value
Filename	illumina_reads_R1_001.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	2051346
Total Bases	514.8 Mbp
Sequences flagged as poor quality	0
Sequence length	251
%GC	43

до

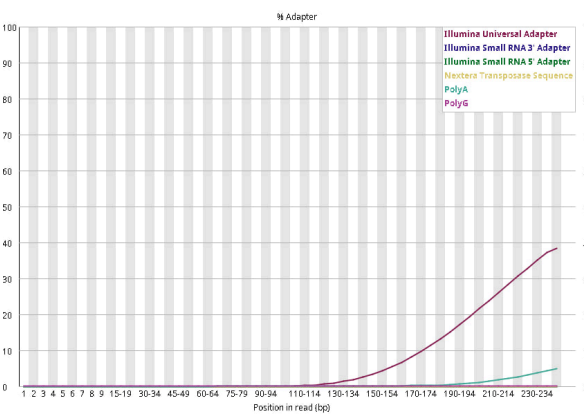
Basic Statistics

Measure	Value
Filename	reads_R1_paired.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1867238
Total Bases	390.2 Mbp
Sequences flagged as poor quality	0
Sequence length	50-251
%GC	42

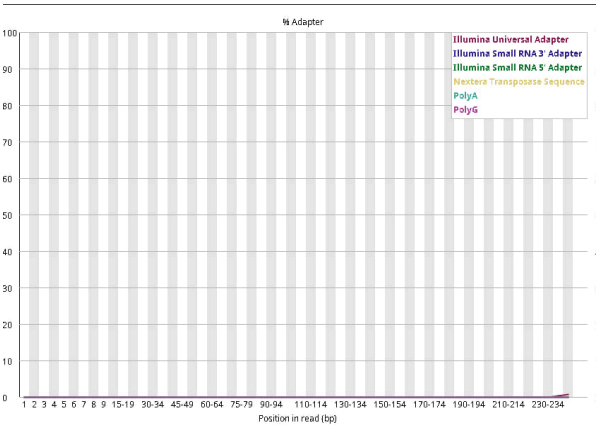
после

Как видно, в сырых файлах на концах качество прочтения сильно падает, и эту проблему отлично решает тримминг. Он обрезал низкокачественные основания с концов ридов, поэтому точность повысилась, а длина сиквенса стала варьировать от 50 до 251, тогда как в сыром файле у всех сиквенсов была длина 251.

Так же тримминг убрал адаптеры. Это видно на графике Adapter Content:



до



после

Таким образом, после тримминга качество ридов значительно улучшилось: убраны адаптеры, увеличено качество концов ридов и удалены низкокачественные риды.