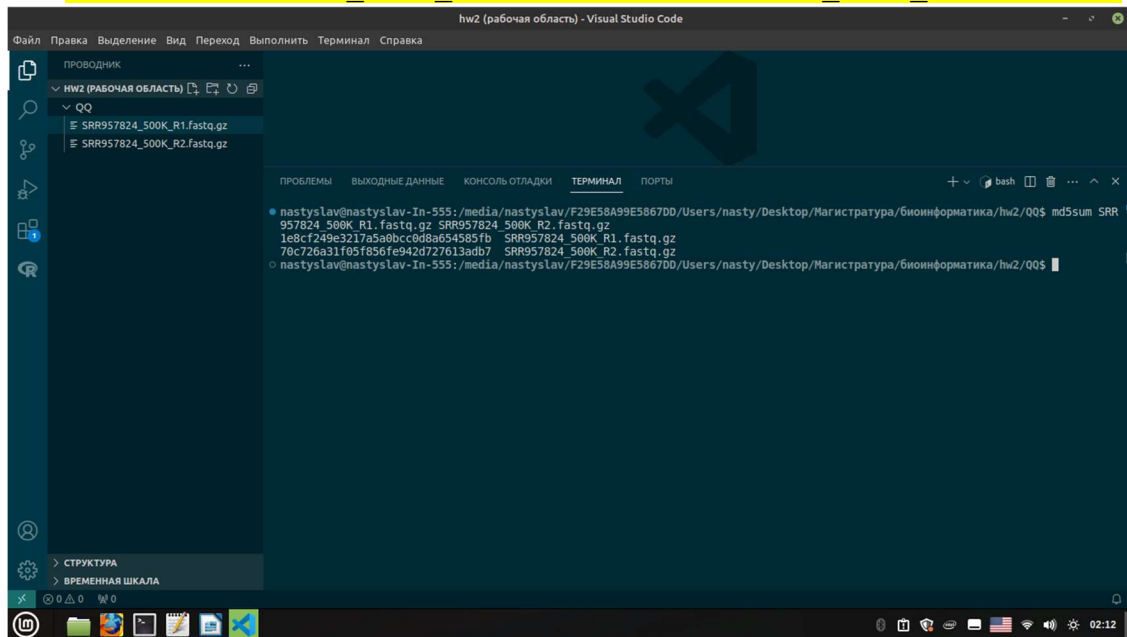


В этом задании будет предложено поработать с датасетом, полученным с Illumina MiSeq при секвенировании ДНК E.coli O157, смертельного штамма кишечной палочки.

Секвенирование парных ридов длиной по 150 п.о. проводилось в Сент-Луисе, США, в 2011 году, при вспышке заболевания.

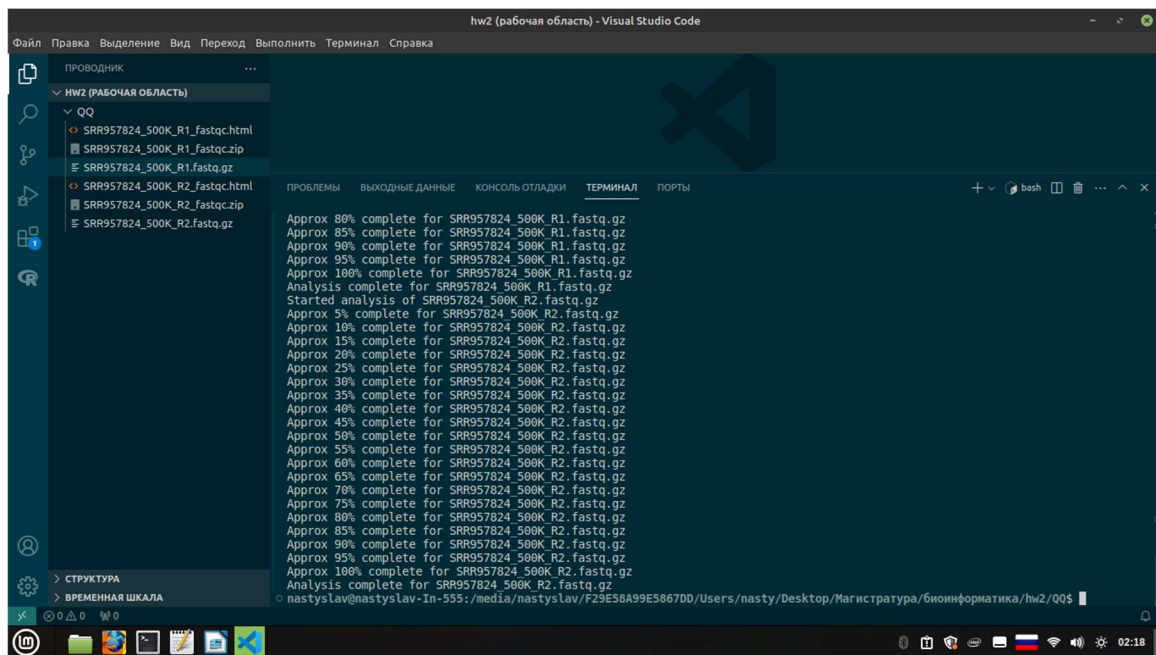
- 1) Загрузите fastq-файл с прямыми прочтениями:
<https://drive.google.com/file/d/1UIZt8EqSEzIjrbY0p75ZADOEzeNofDIC/view?usp=sharing>
- 2) Загрузите обратные прочтения:
<https://drive.google.com/file/d/15ddqwpScJ3iJGMjP7SES3j3ooilkNAf/view?usp=sharing>
- 3) Проверьте, совпадают ли md5 строки этих файлов со следующими:
1e8cf249e3217a5a0bcc0d8a654585fb и
70c726a31f05f856fe942d727613adb7 (это важно, чтобы понять, скачались ли файлы в целости; для проверки используйте md5sum:
`md5sum SRR957824_500K_R1.fastq.gz SRR957824_500K_R2.fastq.gz`).



The screenshot shows a Visual Studio Code interface with a terminal window open. The terminal displays the following command and output:

```
nastyslav@nastyslav-In-555:/media/nastyslav/F29E58A99E5867DD/Users/nasty/Desktop/Магистратура/биоинформатика/hw2/QQ$ md5sum SRR957824_500K_R1.fastq.gz SRR957824_500K_R2.fastq.gz
1e8cf249e3217a5a0bcc0d8a654585fb  SRR957824_500K_R1.fastq.gz
70c726a31f05f856fe942d727613adb7  SRR957824_500K_R2.fastq.gz
```

- 4) Примените FastQC к файлам: `fastqc SRR957824_500K_R1.fastq.gz SRR957824_500K_R2.fastq.gz`



- 5) Проанализируйте полученные отчёты. Какой файл лучше по качеству?
Мне кажется SRR957824_500K_R1_fastqc.html лучше по качеству, т. к. содержимое и качество базовой последовательности намного лучше.
- 6) Скачайте файл с адаптерами, которые чаще всего используются при секвенировании:
<https://drive.google.com/file/d/1pp9rQ1dYMmwKCcdlIGwN9PiNh511SzfA/view?usp=sharing>
- 7) Вы можете применить любой из следующих инструментов для обрезки ридов: fastp, scythe, cutadapt, trimmomatic, а также sickle для устранения низкокачественных концов ридов. Пример команды trimmomatic: `java -jar trimmomatic-[версия программы].jar PE [файл 1] [файл 2] [файл 1, вывод] [файл 1, удалённые риды] [файл 1, вывод] [файл 1, удалённые риды] ILLUMINACLIP:[путь к adapters.fasta]:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36`

```
SRR957824_500K_R2_fastqc.html - hw2 (рабочая область) - Visual Studio Code

PROBLEМЫ 2 Выходные данные КОНСОЛЬ ОТЛАДКИ ТЕРМИНАЛ ПОРТЫ

+89uTnsq0ltbSW1tJawKtraS2tpbW0ltbSWlPa2ktraW1tJbW0lpaS2tpLa2ltbSW1tJawKtraS2tpbW0ltbSWlPa2ktraW
1tJbW0lpaS2tpLa2ltbSW1tJawKtraS2tpbW0ltbSWlPa2ktraW1tJbW0lpaS/VERz/6sT/8w3/86Z/+0Ld/+u/
6It6lBlx4+HP+7zHvuqgPVRd33Xztv03P/dzjz/00BN/
00r0GUVh0MMf6AEV1YDMK/4EDl a3lt+rsw1t+1hw0lnc37nl a3lt+hw1t+3aw6t+ra37+rhw0l+hw0lnc37nl a3lt+rsw1t+1hw0lnc37nl

Q30 bases: 57225034(97.4384%)

Read2 after filtering:
total reads: 416912
total bases: 55852631
Q20 bases: 55352339(99.1043%)
Q30 bases: 53632850(96.0256%)

Filtering result:
reads passed filter: 833824
reads failed due to low quality: 20
reads failed due to too many N: 0
reads failed due to too short: 166156
reads with adapter trimmed: 78303
bases trimmed due to adapters: 1290257
reads corrected by overlap analysis: 4644
bases corrected by overlap analysis: 4795

Duplication rate: 0.0496253%

Insert size peak (evaluated by paired-end reads): 150

JSON report: trimmed/anc.fastp.json
HTML report: trimmed/anc.fastp.html

fastp --detect_adapter_for_pe --overrepresentation analysis --correction --cut_right --thread 2 --html trimmed/anc.fastp.h
tml --json trimmed/anc.fastp.json -i SRR957824_500K_R1.fastq.gz -I SRR957824_500K_R2.fastq.gz -o trimmed/R1.trimmed.fastq.
gz -o trimmed/R2.trimmed.fastq.gz
fastp v0.20.1, time used: 25 seconds
nastyslav@nastyslav-In-555:/media/nastyslav/F29E58A99E5867DD/Users/nasty/Desktop/Магистратура/биоинформатика/hw2/QQ$
```

8) Ещё раз создайте отчёты FastQC. Что изменилось?

```
SRR957824_500K_R2_fastqc.html - hw2 (рабочая область) - Visual Studio Code

PROBLEМЫ 2 Выходные данные КОНСОЛЬ ОТЛАДКИ ТЕРМИНАЛ ПОРТЫ

+89uTnsq0ltbSW1tJawKtraS2tpbW0ltbSWlPa2ktraW1tJbW0lpaS2tpLa2ltbSW1tJawKtraS2tpbW0ltbSWlPa2ktraW
1tJbW0lpaS2tpLa2ltbSW1tJawKtraS2tpbW0ltbSWlPa2ktraW1tJbW0lpaS/VERz/6sT/8w3/86Z/+0Ld/+u/
6It6lBlx4+HP+7zHvuqgPVRd33Xztv03P/dzjz/00BN/
00r0GUVh0MMf6AEV1YDMK/4EDl a3lt+rsw1t+1hw0lnc37nl a3lt+hw1t+3aw6t+ra37+rhw0l+hw0lnc37nl a3lt+rsw1t+1hw0lnc37nl

Approx 75% complete for R1.trimmed.fastq.gz
Approx 80% complete for R1.trimmed.fastq.gz
Approx 85% complete for R1.trimmed.fastq.gz
Approx 90% complete for R1.trimmed.fastq.gz
Approx 95% complete for R1.trimmed.fastq.gz
Analysis complete for R1.trimmed.fastq.gz
c R2.trimmed.fastq.gz
Started analysis of R2.trimmed.fastq.gz
Approx 5% complete for R2.trimmed.fastq.gz
Approx 10% complete for R2.trimmed.fastq.gz
Approx 15% complete for R2.trimmed.fastq.gz
Approx 20% complete for R2.trimmed.fastq.gz
Approx 25% complete for R2.trimmed.fastq.gz
Approx 30% complete for R2.trimmed.fastq.gz
Approx 35% complete for R2.trimmed.fastq.gz
Approx 40% complete for R2.trimmed.fastq.gz
Approx 45% complete for R2.trimmed.fastq.gz
Approx 50% complete for R2.trimmed.fastq.gz
Approx 55% complete for R2.trimmed.fastq.gz
Approx 60% complete for R2.trimmed.fastq.gz
Approx 65% complete for R2.trimmed.fastq.gz
Approx 70% complete for R2.trimmed.fastq.gz
Approx 75% complete for R2.trimmed.fastq.gz
Approx 80% complete for R2.trimmed.fastq.gz
Approx 85% complete for R2.trimmed.fastq.gz
Approx 90% complete for R2.trimmed.fastq.gz
Approx 95% complete for R2.trimmed.fastq.gz
Analysis complete for R2.trimmed.fastq.gz
nastyslav@nastyslav-In-555:/media/nastyslav/F29E58A99E5867DD/Users/nasty/Desktop/Магистратура/биоинформатика/hw2/fastq
$
```

Первый и второй отчет улучшились по качеству, но второй по содержанию базовой последовательности все равно не очень, нужно еще обрезать.

9) Можете сделать один отчёт для проекта – для объединения можно воспользоваться программой MultiQC: перейдите в папку, где содержатся все отчёты FastQC в HTML-формате и наберите команду **multiqc**.

