# Обработка RNA-seq данных и структура GTF

На первом этапе производится проверка качества исходных ридов при помощи FastQC, а затем результаты сводятся в единый отчёт с помощью MultiQC. Для этого используется скрипт `qc.slurm`.  
  
Далее выполняется этап очистки данных: осуществляется обрезка низкокачественных участков и адаптеров при помощи скрипта `fastp.slurm`.  
  
После тримминга проводится повторная оценка качества уже очищенных файлов — снова через FastQC и MultiQC, на этот раз с использованием скрипта `qc\_fastp.slurm`.  
  
При сравнении до- и послеобработанных отчётов можно заметить, что, по большому счёту, изменения минимальны. В моем случае единственное различие заключалось в небольшом смещении распределения длин — оно стало чуть менее равномерным.  
  
Чтобы не усложнять процесс, для дальнейшей работы было решено оставить исходные (необрезанные) риды.  
  
Следующим шагом выполняется выравнивание прочтений на референсный геном человека GRCh38, путь к которому: `/projects/mipt\_dbmp\_biotechnology/GRCh38/`. Используется выравниватель STAR, запуск осуществляется с параметрами, позволяющими сразу получить данные по количеству прочтений. Всё это реализовано через скрипт `star.slurm`.  
  
Чтобы получить аннотационный файл в формате GTF, применяем утилиту `stringtie.sh`, которая формирует файл `transcripts.gtf`.  
  
СТРУКТУРА GTF-ФАЙЛА  
  
Согласно документации, каждый элемент в GTF-файле (например, транскрипт или экзон) представлен отдельной строкой, содержащей 9 колонок:  
  
1. seqname — обозначение хромосомы или контига (например, “1”).  
2. source — название программы, которая сгенерировала эту запись (в нашем случае — StringTie).  
3. feature — тип элемента: может быть transcript или exon.  
4. start — начальная координата на последовательности.  
5. end — конечная координата.  
6. score — численная оценка (часто равна 1000 или отсутствует).  
7. strand — указывает направление (прямая цепь + или обратная -).  
8. frame — фазность кодона (0, 1, 2 либо . — если не применимо).  
9. attributes — дополнительная информация в формате ключ-значение.  
  
Пример транскрипта:  
gene\_id "HM\_RNA\_SEQ.1";  
transcript\_id "HM\_RNA\_SEQ.1.1";  
cov "1.617600";  
FPKM "3.104465";  
TPM "49.656040";  
reference\_id "ENST00000469563";  
ref\_gene\_id "ENSG00000188976";  
ref\_gene\_name "NOC2L";  
  
Пример экзона:  
gene\_id "HM\_RNA\_SEQ.1";  
transcript\_id "HM\_RNA\_SEQ.1.1";  
exon\_number "1";  
cov "1.000000";  
  
КЛЮЧЕВЫЕ МЕТРИКИ ЭКСПРЕССИИ  
  
StringTie вычисляет два основных показателя для анализа уровня экспрессии:  
  
- FPKM (Fragments Per Kilobase per Million) — нормализует число прочтений, учитывая длину транскрипта и общее количество ридов в образце.  
- TPM (Transcripts Per Million) — более универсальная метрика, позволяющая сравнивать экспрессию между генами внутри одного образца, т.к. она нормализует данные по суммарному уровню экспрессии.