

Домашнее задание – Практикум 15

Задача практикума: построить экспрессионный профиль на основании данных секвенирования РНК.

Для сдачи практикума 14 программный конвейер писать не нужно, но если он будет - покажите его при сдаче коллоквиума и получите дополнительные баллы. Вы можете изменить программный конвейер, написанный для сдачи практикумов 11- 13 (у вас теперь ОДНОконцевые чтения и RNA-seq секвенирование).

В любом случае сохраняйте все актуальные команды, выходные файлы, логи и пр., они могут пригодиться на зачете.

1. Описание образца

Найдите ID вашего образца в базе ENCODE (<https://www.encodeproject.org/>). Мы не проходили этот ресурс, но он крайне полезен и интересен своим содержанием.

Укажите:

- a. ID образца РНК-чтений
- b. ссылку на информацию об образце
- c. организм и ткань (если есть)
- d. стратегию секвенирования (тотальная РНК, малые РНК, ...)
- e. парноконцевые или одноконцевые чтения
- f. цепь-специфичность

2. Проверка качества исходных чтений

Проанализируйте качество исходных чтений с помощью программы **fastqc**, как было проделано ранее. В данном случае у нас только один файл с чтениями.

Не удаляйте получившиеся файлы (.html), они могут пригодиться на зачете.

Укажите:

- a. количество чтений
- b. краткий комментарий качества чтений по результатам fastqc (картинка Per base sequence quality)
- c. краткий комментарий о длине ваших чтений по результатам fastqc (картинка Sequence Length Distribution)
- d. (*) краткий комментарий о любых других результатах fastqc

3. Картирование чтений на референс

Напоминаю, что референс (ваша хромосома) мы УЖЕ проиндексировали для hisat2, а чтения решили НЕ триммировать.

Воспользуйтесь командой: **hisat2 -x *your_genome_index* -k 3 -U file.fastq.gz**

Результат сохраните в файл sam. Сохраняйте логи. Изучите лог-файл.

Переведите sam файл в сортированный bam, проиндексируйте. Sam файл можно удалить. Отберите чтения, легшие только на вашу хромосому (см. предыдущие задания). Напоминаю, что gpa-seq сделан для изучения всех мРНК, а у нас только одна хромосома

- a) Сколько чтений закартировалось на вашу хромосому?

4. Поиск экспрессирующихся генов

Вам нужен файл с генной разметкой (см. описание данных).

Обратите внимание, что любые внешние файлы, содержащие какие-либо координаты, должны соответствовать версии референсного генома.

Изучите [файл](#) с генной разметкой.

Опишите кратко, как он устроен.

(*) Сколько на вашей хромосоме аннотировано генов?

Для каждого гена из разметки посчитайте число картированных на этот ген (!!!) чтений с помощью программы [htseq-count](#). Примените опции: -f, -s, -m (любую настройку), -t.

Объясните параметры:

-f

-s

-m

-t

Работает достаточно долго.

Изучите получившийся файл.

Обратите внимание на самый конец (!!!) выходного файла.

- a) Сколько чтений попало в границы генов?
- b) Сколько чтений попало мимо границ генов?

(*) Объясните все строки аннотационного файла, начинающиеся с “_” (они в самом конце файла).

5. Аннотация высоко экспрессируемых генов

Отсортируйте по убыванию количества каунтов (второй столбец) файл с экспрессионным профилем, полученный в предыдущем пункте. Принимайте во внимание только строки, касающиеся аннотированных генов.

- 1) Выведите в явном виде топ 10 самых высоко экспрессируемых генов в вашем случае
- 2) Выберите любой ген (желательно белок-кодирующий) и визуализируйте его с помощью геномного браузера:
 - a. Приведите картинку, на которой изображены экзон-интронные структуры известных для данного гена транскриптов
 - b. Настройте геномный браузер так, чтобы был виден трек консервативности любой версии, дайте свои комментарии
- 3) Воспользуйтесь любым справочным сервисом, который мы проходили (GeneCards, HPA, ...) и кратко опишите функции выбранного гена или его белкового продукта.

Если ничего интересного нет – выберите другой ген.