Домашнее задание №5

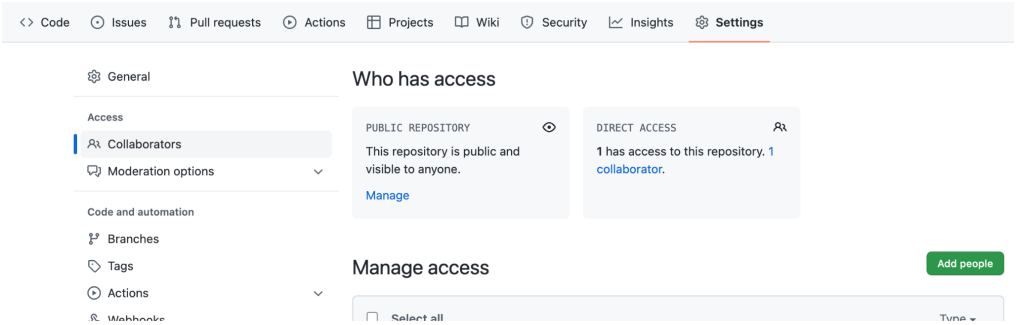
Введение

В данном практическом задании вы научитесь работать с данными scRNA-seq.

Мы анализируем scRNA-seq данные из статьи "Single-cell mapping of the thymic stroma identifies IL-25-producing tuft epithelial cells".

Обязательная часть задания (8 баллов)

1. На сайте github.com создаем **приватный** репозиторий и приводим ссылку на этот репозиторий в общей гугл-таблице (**вкладка HW5**) https://docs.google.com/spreadsheets/d/1bdKBlDMFvhX8xStZUL2YHRTGLpmdiLNU5JcVmy8K-Yw/edit#gid=0
    a. Также необходимо открыть доступ для **efrsw** для будущей проверки (Settings => Collaborators => Add people):



2. **Образец Google Colab** ноутбука: https://colab.research.google.com/drive/1x6gZQfC8iOPw_r-odOVyWTclXCk5mTHO?usp=sharing
3. Получить count-матрицу объединением результатов разных экспериментов https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE103970
4. Провести нормализацию данных (получаем значения TPM).
5. Построить heatmap для экспрессии маркерных генов.
6. Построить визуализацию данных по экспрессии на основе UMAP и PCA.
7. Проанализировать полученные результаты.

Бонусная часть задания (2 балла)

Сравнить уровни экспрессии генов, полученных по scRNA-seq (подгруппа mTEC-IV, где экспрессируется ген Aire) с классическим bulk RNA-seq для целого тимуса мыши из статьи Meredith et al, 2015. По данным bulk RNA-seq имеются 2 реплики - SRR2038194 и SRR2038195. Для сравнения с single cell данными имеет смысл сделать следующее:

1. Провести нормализацию bulk RNA-seq данных (получаем значения TPM).
2. Для каждого гена берем среднее значение TPM по двум репликам bulk RNA-seq
3. Для каждого гена берем среднее значение TPM по всем клеткам scRNA-seq (подгруппа mTEC-IV)
4. Подготовить подвыборку генов для анализа
    a. Можно взять все маркерные гены из основной части задания
    b. 100-200 наиболее высоко-экспрессированных генов как в bulk, так и scRNA (наибольшие значения TPM).
5. Для выбранного набора генов рисуем график с точками, где каждая точка это 1 ген, координаты -- это средняя экспрессия гена в bulk (ось OX) и средняя экспрессия в mTEC-IV (ось OY)
6. Сделать грубый вывод (примерную статистику) сколько генов имеют схожую экспрессию как в bulk, так и в scRNA, для скольки генов разница существенна.
    a. Это можно сделать "на глаз", а можно применить одну из процедур дифференциального анализа (например, DESeq-2). - x2 бонус

Список файлов для сдачи

- В репозитории в файле *README*.md
    - Ссылка на google colab ноутбук.
    - Описание метода нормализации данных.
    - Heatmap для экспрессии маркерных генов.
    - Полученные визуализации UMAP и PCA.
    - Анализ результатов.
    - Результаты выполнения бонусного задания.

Форма отчетности

Github репозиторий, содержащий все полученные результаты.

Последний срок сдачи: среда, 14 декабря до 23:59 (будет отслеживаться по последнему коммиту в репозиторий). Штраф -0.5 балла за каждый день просрочки.

В случае возникновения вопросов обращаться в telegram: @efrsw