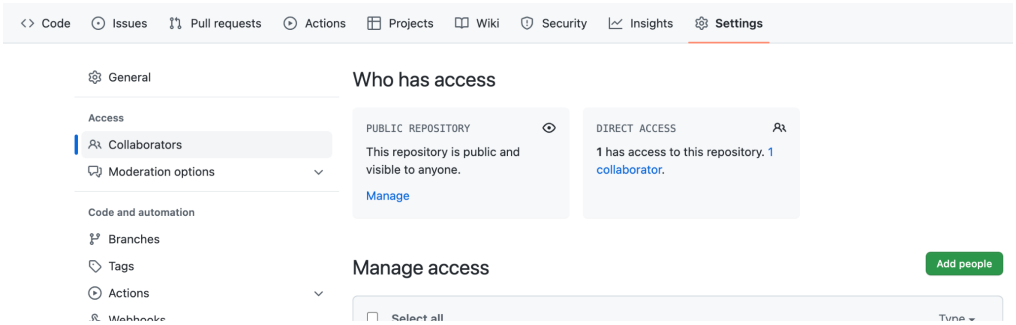
Домашнее задание №4

Введение

Целью данного задания является сравнение RNA-seq данных перепрограммированных и неперепрограммированных (контрольных) мышиных эмбриональных фибробластов (MEFs) и нахождение генов, которые наиболее сильно изменяют свою экспрессию в этом процессе.

Обязательная часть задания (8 баллов)

- На сайте github.com создаем **приватный** репозиторий «hse22_hw4» и приводим ссылку на этот репозиторий в общей гугл-таблице (**вкладка HW4**) https://docs.google.com/spreadsheets/d/1bdKBlDMFvhX8xStZUL2YHRTGLpmdiLNU5JcVmy8K-Yw/edit#gid=0
    - Также необходимо дать доступ ассистентам (**dRabbit-ab**) к репозиторию для будущей проверки (Settings => Collaborators => Add people):
    - 

- Рекомендуется выполнять работу в Google Colab ноутбуках.
    - Если вы будете выполнять работу на сервере или на своем компьютере, необходимо будет также загрузить написанный код на Github
- В данном задании будут проанализированы следующие 6 RNA-seq образцов:
    - Перепрограммированные образцы: SRR3414629, SRR3414630, SRR3414631
    - Контрольные образцы: SRR3414635, SRR3414636, SRR3414637
- Выравнивание RNA-seq чтений на геном мыши:
    - Пример Google Colab ноутбука с примерами запуска только для одного файла (образца). **Вам следует сделать это для всех 6-ти файлов**.
    - https://colab.research.google.com/drive/1TXcuFbEgQXl-4SRh-nx3zwjVlxQoXdt3?usp=sharing
    - Результатом этой части задания будет сводная таблица ALL.counts, где указано кол-во чтений уникально картированных на каждый ген в каждом образце следующего вида:

	c1	c2	c3	r1	r2	r3
ENSMUSG00000000001.4	3431	3504	4031	4489	3919	5700
ENSMUSG00000000003.15	0	0	0	0	0	0
ENSMUSG00000000028.14	148	135	151	345	273	467
ENSMUSG00000000031.15	55525	48225	56062	64504	33248	64991
ENSMUSG00000000037.16	41	44	51	77	68	83
ENSMUSG00000000049.11	10	8	10	4	1	1

Бонусная часть задания 1 (2 балла)

- Поиск генов, которые значимо поменяли свою экспрессию (дифференциально-экспрессированные гены) в результате перепрограммирования с помощью R-пакета DESeq2
    - Пример Google Colab ноутбука на R https://colab.research.google.com/drive/1OAlyTlVFBCllorI9mBfgj0Avg76ZHNJK?usp=sharing
    - Для этой части задания потребуется файл ALL.counts (созданный выше), а также небольшой файл ALL.info с информацией по каждому образцу следующего вида

	id	condition
	<chr>	<chr>
c1	SRR3414635	control
c2	SRR3414636	control
c3	SRR3414637	control
r1	SRR3414629	reprogramming
r2	SRR3414630	reprogramming
r3	SRR3414631	reprogramming

Список файлов для сдачи

- В репозитории в файле *README*.md
    - Ссылки на google colab ноутбуки
    - Скриншоты и статистика из файлов FastQC или multiQC
    - Таблицу со статистикой по каждому из 6-ти образцов:
        - ID образца
        - Тип образца (перепрограммированние или контроль)
        - Общее кол-во исходных чтений
        - Кол-во и процент чтений, которые были успешно откартированы на геном (уникально или нет)
        - Кол-во и процент уникально откартированных чтений
        - Общее кол-во чтений, которые попали на гены
    - Графики из анализа DESeq2 (бонус)
        - MA-plot

        - Тепловая карта, которая показывает, что все контрольные образцы похожи между собой, а перепрограммированные образцы -- между собой
        - Для нескольких генов, которые наиболее значимо поменяли свою экспрессию -- графики со значениями "Normalized counts" в контрольных и перепрограммированных образцах
- В репозитории в папке data
    - Файл ALL.counts -- сводная таблица, где указано кол-во чтений уникально картированных на каждый ген в каждом образце
    - Файл differentially_expressed_genes.txt -- результат работы DESeq2 для всех генов
- В репозитории в папке src – любой другой код, который был использован для выполнения задания

Форма отчетности

Github репозиторий, содержащий все полученные результаты.

Последний срок сдачи: среда, 23 ноября до 23:59 (будет отслеживаться по последнему коммиту в репозиторий).

В случае возникновения вопросов пишите на Telegram ассистента @dbushnev (или в общий чат курса).