Анализ данных single-cell RNA-Seq

# Краткие сведения

**Примерная длительность:** 23 часа

**Количество человек:** 20-25 (4-5 команд), биологи и информатики

**Название проекта:** анализ данных single-cell RNA-Seq

**Краткое описание:** В ходе проекта информатики смогут познакомиться с концепцией single-cellRNA-Seq и потренировать свои навыки в машинном обучении, а биологи смогут ознакомиться с пайплайном CellRanger для анализа single-cell RNA-Seq данных

**Этапы работы:**

1. Распределение ролей в команде
2. Изучение задачи
3. Работа с исходными данными
4. Кластеризация матрицы (экспрессия гена vs клетка)
5. Валидация результатов
6. Финальная презентация (включает биологическую интерпретацию)

**Фабула:**

Участникам соревнования предлагается поучаствовать в имитации реальной деятельности команды, которая работает на биоинформатическом проекте.

Как и в реальной жизни, задачи могут быть поставлены достаточно размыто, эффективность работы команды будет оценена на каждом этапе с помощью объективных критериев (описание ниже) и субъективных (жюри будет выполнять роль заказчика).

# Детальный план

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **#** | **Краткое описание** | **Навыки** | **Результат** | **Оценивание** | **t, чаc** | **Max** | **Комментарии** |
| 1 | Участникам команды предстоит познакомиться друг с другом, выбрать лидера, попытаться распределить роли (например, ведущий разработчик (пишет код), аналитики (работают с публикациями, биолог и математик), тестеровщик). Роли могут быть совмещены. Деление может быть любым, но должно быть обосновано. | Работа в команде | Краткое описанией ролей всех участников команды с обоснованием. |  | 0.5 |  |  |
| 2 | Команды изучают поставленную задачу, материалы по теме, выясняют научную значимость и актуальность. Участникам предстоит выбрать один из предложенных наборов данных (10xgenomics.com/single-cell-gene-expression/datasets/ ) для дальнейшего анализа. | Работа с источниками | Презентация (2 слайда) | 1 - упоминание 5 источников, 1 - док-во актуальности и значимости, 1 - приз зрительских симпатий | 2 | 3 | Дефолтный набор тут - https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.0.1/pbmc8k . Если выберут другой и обоснуют, то очень хорошо. на дефолтном наборе проще получить различимые кластеры (это конечная цель). Задача больше для биологов |
| 3 | Первичный этап анализа подразумевает разделение ридов по клеткам и получение матрицы экспрессии, в которой в столбах находится идентификатор клетки, а в строках ген, на пересечении же находится количество транскриптов в клетке для данного гена. Для этого участникам требуется установить CellRanger 10xgenomics.com/single-cell-gene-expression/software/what-is-cell-ranger и получить из fastq файлов матрицы экспрессии для двух датасетов PBMC(Peripheral blood mononuclear cells): 4тыс. клеток pbmc4k и 8 тыс. клеток pbmc8k. | Работа с кластером, навык работы с сырыми данными | Матрица экспрессии | 1 - выполнение инструкций по работе с кластером, 2 - получение матрицы с помощью стандартной процедуры, 5 - полчение матрицы с помощью другого инструмента, при выигрыше во времени большем, чем статистическая погрешность, 2 - выводы из этой части задания (может быть сравнение инстурментов из интернета, ориентировочное время работы и т.д.) | 6 | 10 |  |
| 4 | Участникам надо кластеризовать данные и определить присутствующие клеточные популяции (для этого очень удобно воспользоваться следующим пакетом в R- http://satijalab.org/seurat/pbmc-tutorial.html). 1. Уменьшить пространство (вариации на тему PCA) 2. Кластеризовать (h-clust, k-means, mixture modules, etc) | Машинное обучение (кластеризация, методы понижения размерности), R, Python (?) | Набор кластеров, которые соответсвуют клеточным популяциям. | 2 - получение результата с помощью стандартного алгоритма, описанного в статье, 3 - теоретическое сравнение методов понижения размерности, обоснование выбора, 4 - использование не описанного в статье метода понижения размерности с обоснованием выбора, 4 - использование не описанного в статье метода кластеризации с обоснованием | 8 | 13 | Матрица на сайте |
| 5 | Участникам предлагается выбрать метод валидации результата, которые наиболее полно соотвествует поставленным в начале работы целям. Рекомендовано использовать советы по ссылке [2]. | Умение оценить результат, Интерпретация результатов | Таблицы с результатами валидации кластеризации | 4 - описание методов валидации (обязательно должно включать сравнение двух типов кластеризации), обоснование выбора наиболее подходящего, 4 - биологическое обоснование полученных результатов, 2 - дополнительные баллы, если результат будет заметно лучше осталньых команд | 4 | 10 |  |
| 6 | Необходимо составить финальную презентацию, которая включает описание всех фаз проекта | Презентация, автомачтиеское добавление результатов в презентацию | Презентация (12-18 слайдов) | 4 - оценка презентации, 2 - использование нетривиальных и эффективных средств визуализации | 2.5 | 6 |  |
| **S** |  |  |  |  | **23** | **44** |  |

# Ссылки

1. <http://satijalab.org/seurat/pbmc3k_tutorial.html> - полный туториал
2. <https://ranalytics.github.io/data-mining/103-Clustering-Quality.html> - кластерная валидация
3. <https://arxiv.org/pdf/1703.09282.pdf> - статья по кластерной валидации
4. <https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.0.1/pbmc8k> - рекомендуемый набор данных
5. <https://academic.oup.com/bioinformatics/article/31/12/1974/214505/Identification-of-cell-types-from-single-cell> статья по рекомендуемому методу кластериации