Курс “Анализ транскриптомных данных”

Описание курса

Курс «Анализ транскриптомных данных» посвящён анализу данных экспрессий генов, полученных при помощи платформ высокопроизводительного секвенирования. В ходе курса будут освещены как вопросы анализа данных bulk RNA-Seq, так и становящихся всё более популярных в последние годы данных scRNA-Seq. Особое внимание будет уделено методам машинного обучения (от GLM до VAE и методов снижения размерности), которые сейчас являются «золотым стандартом» на всех стадиях работы с транскриптомными данными.

Курс состоит из 15 лекций и 15 семинаров. На лекциях основное внимание будет уделено теоретическим основам применяемых методов анализа, а также дискуссиям насчёт областей применимости тех или иных подходов. На семинарах будут рассмотрены конкретные примеры использования различных инструментов, а также рассмотрены некоторые углубленные вопросы из курса. После каждого семинара будет даваться домашнее задание, направленное на закрепление материалов, полученных на занятии.

Типичный слушатель курса — это студент естественно-научной специальности, который хочет овладеть современными методами анализа экспрессионных данных, а также качественно применять их в своей исследовательской работе. Для того, чтобы полностью освоить курс, требуется владение языками Python и R, а также базовое понимание статистики, теории вероятностей и линейной алгебры.

Программа курса

1. **Лекция**: Технологии секвенирования следующего поколения (NGS). Экспериментальные подходы к секвенированию РНК тканей (bulk RNA-Seq). Сходства и различия с микрочиповыми технологиями. Основные базы данных (SRA, GEO). **Семинар**: Базовая работа с прочтениями. SRA-Toolkit, SRA-Explorer, FastQC, MultiQC.
2. **Лекция**: Выравнивания (STAR, HISAT2) и псевдовыравнивания (kallisto, Salmon). EM-алгоритм для оценки представленности транскриптов (RSEM). **Семинар**: «препарирование» EM-алгоритма и его реализация на Python.
3. **Лекция**: Основные распределения, встречающиеся в омиксных данных. Методы нормализации в bulk RNA-Seq: от RPKM и TPM до RLE и TMM. Контроль за дисперсией в данных. **Семинар**: Статистические подходы к определению максимально правдоподобных распределений данных.
4. **Лекция**: Дифференциальная экспрессия, параметрические и непараметрические тесты. Линейные модели и обобщённые линейные модели (GLM). Работа с экспрессиями на уровне транскриптов. tximport и Sleuth. **Семинар**: Написание собственного алгоритма определения дифференциально экспрессированных генов. Работа с пакетами DESeq2 и edgeR.
5. **Лекция**: Системный анализ bulk RNA-Seq: анализ обогащённости (GO Enrichment Analysis), Gene Set Enrichment Analysis (GSEA и ssGSEA). Работа с экспрессионными данными на уровне генных сигнатур. Понятие деконволюции bulk RNA-Seq. **Семинар**: Практическая работа с экспрессионными данными на

уровне генных сигнатур. Сравнение различных подходов к определению клеточного состава bulk RNA-Seq (signature-based vs. deconvolution).

6. **Лекция**: Понятие и необходимость scRNA-Seq. Методы подготовки библиотек scRNA-Seq. Сравнение различных подходов для подготовок библиотек для scRNA-Seq. Batch effect в данных scRNA-Seq. **Семинар**: Работа с базами данных scRNA-Seq. Дискуссия на тему правильного выбора стратегии подготовки библиотек.
7. **Лекция**: Выравнивания и псевдовыравнивания в scRNA-Seq. Контроль качества клеток в scRNA-Seq. Определение и устранение пустых клеток и дублетов. **Семинар**: Собственная реализация алгоритма поиска пустых капель.
8. **Лекция**: Процессинг данных scRNA-Seq: сходства и различия с bulk RNA-Seq. SCTransform, LogNorm, pagoda2 и прочие способы контроля за дисперсией данных. **Семинар**: Собственная реализация алгоритма SCTransform.
9. **Лекция**: Проклятие размерности. Feature selection при помощи регуляризаций. Методы feature selection, принятые в scRNA-Seq: выделение высоко-вариабельных генов и подходы к этому выделению. Методы снижения размерности: PCA, t-SNE, UMAP, ForceAtlas2. Графовое представление данных. **Семинар**: Реализация алгоритма t-SNE и «тюнинг» его функции потерь.
10. **Лекция**: Подходы к устранению батч-эффекта в scRNA-Seq: Harmony, bbkNN, Scanorama, MNN, conos. Анализ методом канонических корреляций (CCA). **Семинар**: Сравнение подходов для устранения батч-эффектов в данных scRNA-Seq.
11. **Лекция**: Использование вариационных аутоэнкодеров для процессинга scRNA-Seq. scVI-tools. **Семинар**: Препарирование scVI, написание собственного вариационного аутоэнкодера на PyTorch и Pyro.
12. **Лекция**: Подходы к кластеризации данных. Иерархическая кластеризация, K-Means, графовые алгоритмы кластеризации (Louvain, Leiden, SNN). Понятие стабильности кластера, бутстрэп. **Семинар**: Реализация алгоритма оценки стабильности кластеров.
13. **Лекция**: Определение траекторий дифференцировки клеток в scRNA-Seq: Monocle2, Monocle3, иные подходы. Обобщённые аддитивные модели (GAM) и их использование для определения генов, которые меняют свою экспрессию по ходу дифференцировки клеток. RNA velocity. **Семинар**: Написание собственного алгоритма определения генов, которые меняют свою экспрессию по ходу псевдо-времени.
14. **Лекция**: Определение типов клеток в scRNA-Seq: автоматическое и мануальное. Поиск взаимодействий между различными типами клеток, CellPhoneDB. **Семинар**: Написание алгоритма автоматического определения типов клеток. Сравнение существующих алгоритмов.
15. **Лекция**: Мультимодальные омики одиночных клеток. Подходы для анализа мультимодальных омик одиночных клеток: MOFA, WNN, totalVI, multiVI. CLR-transformation в омиксных данных. Работа с омиксными данными как с композиционными данными. **Семинар**: Воркшоп по анализу мультимодальных омиксных данных.