DOI: 10.25205/978-5-4437-1843-9-31

ИЕРАРХИЧЕСКАЯ АННОТАЦИЯ КЛЕТОК И ВЫСОКОТОЧНАЯ ИМПУТАЦИЯ МОДАЛЬНОСТЕЙ ПРИ ПОМОЩИ ИНСТРУМЕНТА SCPARADISE *

SCPARADISE: TUNABLE HIGHLY ACCURATE MULTI-LEVEL CELL TYPE ANNOTATION AND MODALITY IMPUTATION

Е. С. Чечехина, В. И. Чечехин

Московский государственный университет им. М. В. Ломоносова

E. S. Chechekhina, V. I. Chechekhin

Lomonosov Moscow State University

✉ voynovaes.pharm@gmail.com

Аннотация

scParadise — интегрированный инструмент для комплексного анализа данных РНК-секвенирования одиночных клеток, включающий модули для многоуровневой автоматической аннотации клеток (scAdam), импутации модальностей (scEve) и оценки качества импутации и аннотации (scNoah). Инструмент обеспечивает высокую точность и гибкость для мультимодальных исследований.

Abstract

scParadise is an integrated tool for comprehensive single-cell RNA sequencing data analysis. It includes modules for multi-level automatic cell type annotation (scAdam), modality imputation (scEve), and quality assessment of annotation and imputation (scNoah). The tool delivers high accuracy and flexibility for multimodal studies.

Современный анализ РНК-секвенирования одиночных клеток связан с необходимостью в короткие сроки осуществлять одновременную аннотацию множества датасетов, содержащих разнообразные клеточные типы. Несмотря на многообразие алгоритмов автоматической аннотации, исследователи сталкиваются с рядом проблем. Ключевыми из них являются невысокая точность существующих алгоритмов, трудности идентификации редких популяций клеток, низкая универсальность методов, а также отсутствие стандартизированных подходов к анализу качества аннотаций. Помимо этого, современные датасеты РНК-секвенирования одиночных клеток часто являются мультимодальными. Такой тип данных помимо экспрессии мРНК в клетках также содержит информацию о представленности поверхностных белков и/или открытости хроматина. Мультимодальные датасеты позволяют точнее аннотировать клетки и анализировать их функциональное состояние. Однако мультимодальные исследования являются дорогостоящими и технически сложными. В связи с этим стали развиваться методы импутации модальностей в данные РНК-секвенирования одиночных клеток. Но все они на сегодняшний день связаны с интеграцией данных с референсными мультимодальными датасетами. Использование интеграционных методов для импутации данных ограничены вычислительными мощностями и доступностью референсных данных. Для преодоления ограничений современных методов автоматической аннотации и импутации данных мы разработали новый инструмент — scParadise, объединяющий в себе три полноценных модуля: scAdam, scEve и scNoah.

Модуль scAdam реализует многоуровневую аннотацию клеточных типов — от основных классов до конкретных функциональных состояний клеток. ScAdam превосходит существующие инструменты, такие как CellTypist, scGPT, Seurat, Azimuth, scANVI, TOSICA и Symphony, по сбалансированной точности и некоторым другим метрикам качества на датасетах мононуклеаров периферической крови, поджелудочной и рака гортани человека [1–6]. Помимо этого, scAdam способен осуществлять многоуровневую аннотацию клеток, что не поддерживается ни одним другим нейросетевым автоматическим аннотатором клеток (CellTypist, scGPT, TOSICA). Также scAdam включает более 30 предобученных моделей для разных тканей и биологических видов, что позволяет легко интегрировать инструмент в широкий спектр биологических задач.

Модуль scEve представляет собой инструмент для импутации уровня представленности поверхностных белков в данные РНК-секвенирования одиночных клеток. ScEve не только превосходит интеграционные методы импутации модальностей (например, Seurat, Azimuth) по метрикам ошибки, но и предоставляет возможность

* Исследование выполнено за счет гранта Российского научного фонда (проект № 25-75-30005).

применения моделей на различных тканях. Так, модель, обученная на мононуклеарах периферической крови человека, успешно выявляет новые функциональные состояния NK-клеток в жировой ткани. Межтканевое применение scEve дает преимущество над интеграционными методами импутации модальностей. Кроме того, модели scEve и scAdam поддерживают различные варианты оптимизации, такие как подбор гиперпараметров и дообучение на новых данных.

Модуль scNoah служит универсальным инструментом для оценки результатов аннотации и импутации модальностей. С его помощью можно проанализировать качество моделей и визуализировать полученные результаты, отображая распределения ошибок для каждой популяции клеток и формируя отчеты с метриками качества аннотации и импутации модальностей.

Таким образом, scParadise представляет собой единый инструмент для анализа данных РНК-секвенирования одиночных клеток. Он расширяет возможности автоматического определения клеточных типов, добавляя иерархическую аннотацию, а также позволяет осуществлять межтканевую импутацию модальностей в данные. ScParadise является пакетом для языка программирования Python и свободно доступен в PyPi для установки. Инструмент сопровождается подробной документацией, что облегчает его интеграцию в существующие биоинформатические пайплайны по анализу данных РНК-секвенирования одиночных клеток. ScParadise опубликован в виде препринта и доступен по ссылке [7].

Литература

1. Domínguez Conde C., Xu C., Jarvis L. B. et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans // Science. 2022. Vol. 376. P. eabl5197.

2. Cui H., Wang C., Maan H. et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI // Nature Methods. 2024. Vol. 21. P. 1–11.

3. Chen J., Xu H., Tao W. et al. Transformer for one stop interpretable cell type annotation // Nature Communications. 2023. Vol. 14. P. 223.

4. Butler A., Darby C., Hao Y. et al. Azimuth: A Shiny App Demonstrating a Query-Reference Mapping Algorithm for Single-Cell Data. 2023.

5. Gayoso A., Lopez R., Xing G. et al. A Python library for probabilistic analysis of single-cell omics data // Nature Biotechnology. 2022. Vol. 40. P. 163–166.

6. Hao Y., Stuart T., Kowalski M. H. et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis // Nature Biotechnology. 2024. Vol. 42. P. 293–304.

7. Chechekhina E., Tkachuk V., Chechekhin V. scParadise: Tunable highly accurate multi-task cell type annotation and surface protein abundance prediction // bioRxiv. 2024. URL: https://www.biorxiv.org/content/10.1101/2024.09.23.614509v1.full.