Секция «Биоинформатика»

Поиск возможностей анализа низкоэкспрессируемых РНК по данным секвенирования РНК единичных клеток

Точилкина Мария Сергеевна
Студент (специалист)
Московский государственный университет имени М.В.Ломоносова, Факультет биоинженерии и биоинформатики, Москва, Россия
E-mail: mtoch@mail.ru

РНК-секвенирование единичных клеток (scRNA-seq) - это технология, позволяющая анализировать экспрессионные профили отдельных клеток. При этом она обладает определёнными недостатками, такими как низкое соотношение сигнал-шум и высокий процент дропаутов. Дропаут - это явление, при котором экспрессия гена, в действительности экспрессирующегося в клетке, равна нулю в данных scRNA-seq. Дропаут возникает чаще при низкой глубине секвенирования или у низкоэкспрессируемых генов. UMI-методы scRNA-seq менее чувствительны по сравнению с full-length подходами и в большей степени подвержены дропауту. В то же время UMI подходы более дешёвые и высокопроизводительные, поэтому более популярны.

Высокий уровень дропаутов может значительно повлиять на результаты в случаях, когда важным параметром является доля клеток, экспрессирующих определённый ген-маркер. Так, например, клеточный тип или восприимчивость к вирусу, часто определяется по экспрессии одного или 2-3 генов-маркеров, которые могут экспрессироваться на низком уровне. Например, подверженность ткани к заражению вирусом SARS-CoV-2 можно оценивать по доле клеток, экспрессирующих гены факторов входа коронавируса в клетку. При этом на scRNA-seq данных эпителия трахеи мыши было показано, что доля клеток, в которых они экспрессируются, значительно занижена в UMI данных по сравнению с full-length данными.

Для борьбы с разреженностью scRNA-seq данных, полученных UMI методами, разработаны биоинформатические подходы - методы импутации и шумопонижения. Наша задача - проверить, насколько эти методы применимы в задачах, где ключевым фактором является доля клеток, в которых экспрессируется один маркер, поскольку обычно при исследовании эффективности методов данная проблема не рассматривается.

В этой работе мы применили 6 методов импутации (scVI, SAVER, MAGIC, kNN-smoothing, ALRA, scImpute) на scRNA-seq данных линии K562. Этот набор данных получен с помощью UMI подхода 10x Genomics, но его отличает высокая глубина секвенирования, поэтому мы предполагаем, что он практически не подвержен дропауту. Мы сгенерировали подвыборки, имитирующие более низкие глубины секвенирования, чтобы полученные наборы данных имели большее количество дропаутов. Далее мы разделили гены на 6 групп, основываясь на том, как изменяется доля экспрессирующих клеток и уровень экспрессии гена при увеличении глубины секвенирования. Одна из полученных групп генов отличалась низкой долей экспрессирующих клеток (∼0-5%), и эта метрика значительно не изменялась при увеличении глубины секвенирования. Значение доли экспрессирующих клеток в случае генов из остальных групп росло, а затем выходило на плато. Полученные подвыборки были обработаны методами импутации. Все методы значительно завысили долю экспрессирующих клеток. При этом применение SAVER, MAGIC и scVI привело к значению доли 100% для практически всех генов, что биологически неверно. Возможно, аккуратный подбор параметров запуска алгоритмов импутации позволит получить более корректные результаты.