БИОХИМИЯ, 2023, том 88, вып. 2, с. 171 – 198

УДК 577.21:575.853

ВОЗМОЖНОСТИ КОМПЛЕКСНОГО АНАЛИЗА ДАННЫХ СЕКВЕНИРОВАНИЯ РНК ЕДИНИЧНЫХ КЛЕТОК

Обзор

© 2023 А.А. Хозяинова[1]*, А.А. Валяева[2], М.С. Арбатский[2], С.В. Исаев[3,4], П.С. Ямщиков[1,5], Е.В. Волчков[6], М.С. Сабиров[7], В.Р. Зайнуллина[1], В.И. Чечехин[2], Р.С. Воробьёв[1], М.Е. Меняйло[1], П.А. Тюрин-Кузьмин[2], Е.В. Денисов[1]

[1] *Научно-исследовательский институт онкологии, Томский национальный исследовательский медицинский центр Российской академии наук, 634050 Томск, Россия; электронная почта: khozyainova@onco.tnimc.ru*

[2] *Московский государственный университет имени М.В. Ломоносова, 119991 Москва, Россия*

[3] *Институт персонализированной медицины, Национальный центр персонализированной медицины эндокринных заболеваний, Национальный медицинский исследовательский центр эндокринологии, 117036 Москва, Россия*

[4] *Московский физико-технический институт, физтех-школа биологической и медицинской физики, 115184 Долгопрудный, Россия*

[5] *Национальный исследовательский Томский государственный университет, 634050 Томск, Россия*

[6] *ФГБУ «НМИЦ ДГОИ им. Дмитрия Рогачева» Минздрава России, 117198 Москва, Россия*

[7] *ФГБУН Институт биологии развития им. Н.К. Кольцова РАН, 119334 Москва, Россия*

Поступила в редакцию 23.09.2022
После доработки 13.12.2022
Принята к публикации 13.12.2022

Секвенирование РНК единичных (отдельных/индивидуальных) клеток (single-cell RNA-sequencing, scRNA-seq) является революционным инструментом для изучения физиологии нормальных и патологически изменённых тканей. Данный подход предоставляет информацию о молекулярных особенностях (генной экспрессии, мутациях, степени открытости хроматина и др.) клеток, открывает возможность для анализа траекторий клеточной дифференцировки/филогении и межклеточных взаимодействий и позволяет обнаруживать новые типы клеток и ранее неизученные процессы. В клиническом аспекте scRNA-seq позволяет проводить более глубокий и детальный анализ молекулярных механизмов развития различных заболеваний и предоставляет основу для разработки новых профилактических, диагностических и терапевтических решений. В данном обзоре описываются различные подходы к анализу данных scRNA-seq, рассматриваются сильные стороны и недостатки биоинформатических инструментов, приводятся рекомендации и примеры их успешного использования и предлагаются потенциальные направления в области их совершенствования. Также подчёркивается необходимость создания новых, в том числе мультиомиксных, протоколов для подготовки библиотек единичных клеток с целью получения более полного и системного представления о каждой клетке.

КЛЮЧЕВЫЕ СЛОВА: секвенирование РНК единичных клеток, клеточный цикл, кластеризация, дифференциальная экспрессия, клеточные типы, траектории развития, межклеточная коммуникация, генные регуляторные сети, вариации числа копий ДНК, однонуклеотидные замены, филогенетика, эпигеномика, пространственная транскриптомика.

DOI: 10.31857/S032097252302001X, **EDN:** QFSJMW

Принятые сокращения: ДЭГ – дифференциально экспрессирующиеся гены; aCGH – микроматричная сравнительная геномная гибридизация; bulk RNA-seq – секвенирование тотальной РНК; CNV – вариации числа копий ДНК; MLPA – амплификация лигированных зондов; SNV – однонуклеотидные замены; scRNA-seq – секвенирование РНК единичных клеток; WGCNA – анализ взвешенных сетей коэкспрессии генов; WGS – полногеномное секвенирование.

* Адресат для корреспонденции.

ВВЕДЕНИЕ

Секвенирование РНК единичных клеток (single-cell RNA-sequencing, scRNA-seq) стало поистине революционным методом, позволившим в значительной степени расширить понимание гетерогенности и динамики транскриптома клеток в различных биологических объектах. Впервые данный метод был применён в 2009 г. для изучения бластомеров мыши на стадии второго деления [1]. Именно тогда было показано, что секвенирование единичных клеток существенно превосходит технологию микрочипов для количественного анализа экспрессии генов. Однако главным ограничением того времени являлась невозможность мультиплексирования образцов, и библиотека каждой клетки создавалась вручную в отдельной пробирке. Однако уже в 2011 г. был разработан первый протокол мультиплексного scRNA-seq [2], а в 2014 г. – первая коммерчески доступная платформа автоматической подготовки библиотек единичных клеток Fluidigm C1 [2]. На настоящий момент доступны различные платформы для выполнения scRNA-seq, среди которых Fluidigm C1/Smart-seq, BD Rhapsody («BD Biosciences», США), Chromium («10x Genomics», США) и другие, которые обеспечивают высокую производительность данного типа анализа [3, 4].

Процесс scRNA-seq схематически представлен на рис. 1. Посредством гомогенизации из исследуемого образца получают суспензию клеток, которые далее разделяются либо физически, например, с помощью сортировки и микроманипуляции, либо посредством баркодирования с использованием олигонуклеотидов в составе планшетов или на основе микрофлюидики и комбинаторики [5, 6]. Образцы крови и клеточных культур подвергают сортировке и микроманипуляции без подготовки суспензии. Полученные клетки используют для подготовки библиотек и последующего секвенирования, данные которого обрабатываются биоинформатически.

Развитие технологий в области scRNA-seq позволило охарактеризовать основные клеточные и молекулярные механизмы, вовлечённые

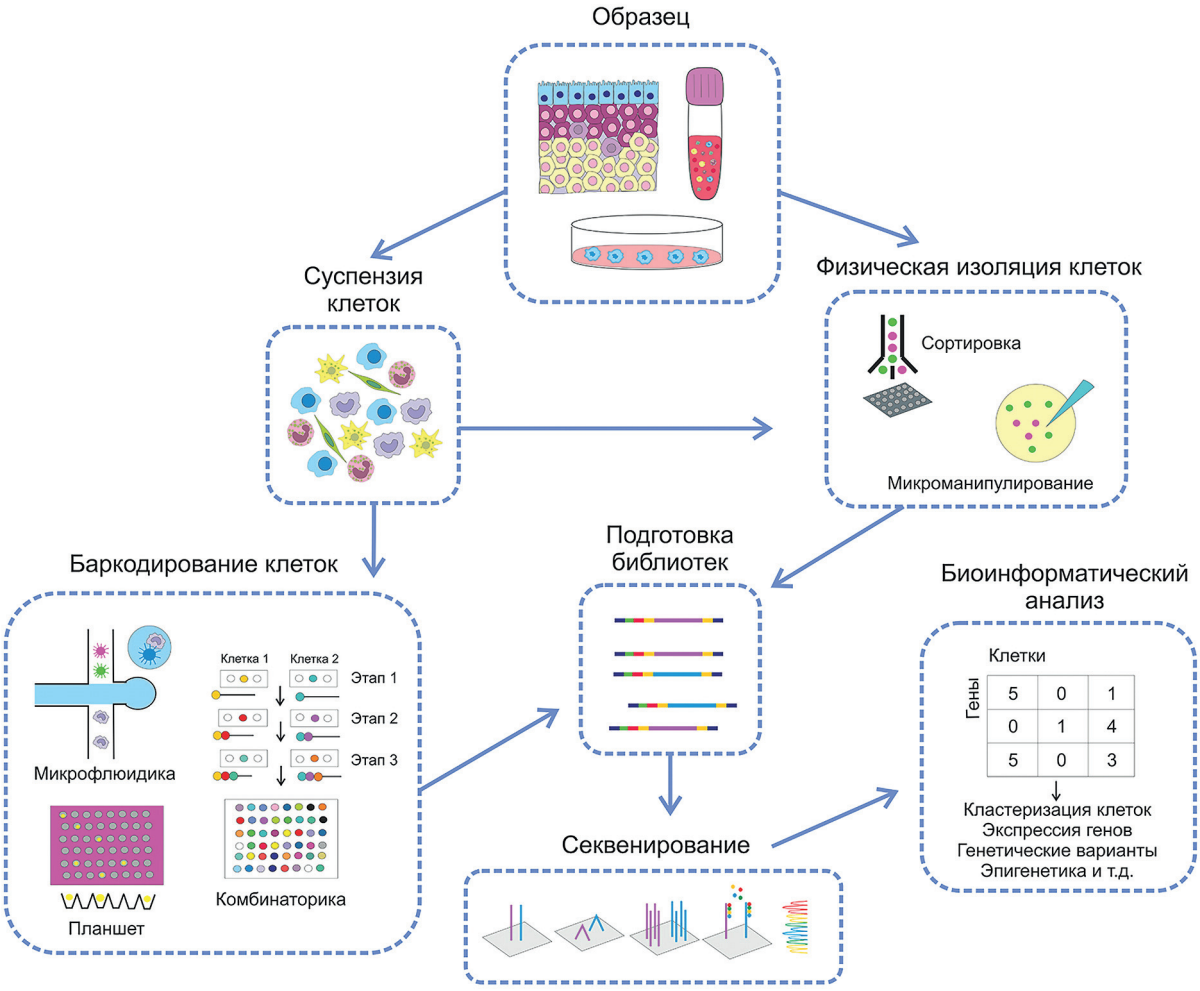

Рис. 1. Процесс scRNA-seq

в развитие сердечно-сосудистых [7], нейродегенеративных [8, 9], онкологических [10] и других заболеваний, определить клеточную таксономию широко используемых модельных организмов, таких как резуховидка Таля [11], дрозофила обыкновенная [12] и данио-рерио [13], и расшифровать гетерогенность клеток иммунной системы в нормальном и патологических состояниях, включая онкологические заболевания [14, 15].

В начале развития эры секвенирования единичных клеток исследователи в большей степени акцентировали внимание на изучении популяционного состава исследуемых образцов, анализе дифференциальной экспрессии, оценке клеточного цикла и идентификации клеточных типов [1, 16]. Хотя уже тогда было ясно, что возможности применения scRNA-seq намного шире, и создание биоинформатических методов, способных интегрировать и преобразовывать полученные нуклеотидные прочтения в информацию о различных состояниях клетки в мультимодальном измерении, было лишь вопросом времени.

В данном обзоре мы описываем прогресс в области обработки данных scRNA-seq и связанные с ним методы анализа для получения информации о клеточном цикле, клеточных

Рис. 2. Подходы биоинформатического анализа данных scRNA-seq

кластерах и типах клеток, дифференциально экспрессирующихся генах (ДЭГ), траекториях развития и скорости РНК (Cell trajectory inference and RNA Velocity), межклеточных взаимодействиях, генетических вариантах, включая вариации числа копий ДНК (CNV) и однонуклеотидные замены (SNV), филогении клеток, доступности хроматина, сайтах связывания транскрипционных факторов и пространственной транскриптомике (рис. 2). Для каждого из мультимодальных приложений подчёркиваются сильные стороны, недостатки, возможные пути их преодоления и биологические примеры использования.

ОЦЕНКА КЛЕТОЧНОГО ЦИКЛА

Секвенирование РНК единичных клеток, в отличие от секвенирования тотальной РНК (bulk RNA-seq), позволяет получить информацию о транскрипционном профиле каждой индивидуальной клетки в исследуемом образце. С помощью scRNA-seq возможно исследование клеточного многообразия тканей, обнаружение ранее неизвестных популяций клеток и изучение биологических процессов на уровне единичных клеток. Однако увеличению разрешения метода сопутствует и повышение доли технического и биологического шума в данных. При этом одним из главных источников биологического шума в данных scRNA-seq оказывается клеточный цикл [17]. При анализе данных scRNA-seq стадия клеточного цикла часто рассматривается как конфаундер, то есть та переменная, которая может исказить биологический эффект, искомый исследователем, будь то различия между клеточными типами или изменения транскрипционных программ при заболеваниях или в процессе лечения. Клетки исследуемого образца могут находиться в различных временных точках и, соответственно, иметь различные профили экспрессии, даже если они являются клетками одного типа. Поэтому на этапе подготовки матрицы генной экспрессии для последующего анализа возможно удаление из данных дисперсии, связанной с клеточным циклом. Эта процедура представляется особо целесообразной, когда исследователь не ожидает увидеть в своих данных активно пролиферирующие клетки, например, при изучении апоптоза. Также удаление генов клеточного цикла или пролиферирующих клеток целесообразно в случае, если большая часть наиболее вариабельных генов представлена генами клеточного цикла, что отрицательно сказывается на определении дифференциальной экспрессии. Однако в некоторых случаях, например, при сравнении субпопуляций делящихся и неделящихся клеток, информация о стадии клеточного цикла может оказаться важной, и данный конфаундер удалять не стоит.

Популярные программные пакеты для анализа данных scRNA-seq, такие как Seurat [18] и Scanpy [19], предлагают считать два параметра [20], основанных на усреднённой экспрессии известных генов-маркеров стадий клеточного цикла S и G2/M [21]. В зависимости от рассчитанных значений параметра клетка может быть проаннотирована как находящаяся на стадии G2/M, G1 или S клеточного цикла. Удаление эффекта, связанного с клеточным циклом, происходит с помощью линейной регрессии, в которой учитываются рассчитанные параметры. Если для последующего анализа необходимо сохранить разделение между субпопуляциями делящихся и покоящихся клеток и одновременно избавиться от различий в стадиях клеточного цикла, то рекомендуется в линейной регрессии использовать не G2/M- и S-параметры, а разницу между ними.

Ещё один метод Cyclone [20] также использует известные гены-маркеры стадий клеточного цикла для оценки и удаления эффектов, связанных с клеточным делением. Однако этот алгоритм построен на сравнении экспрессии пар маркерных генов, поскольку их отношение позволяет определить стадию клеточного цикла, в которой клетка находится.

Другие методы позволяют проводить более подробный анализ клеточного цикла на основе данных scRNA-seq и реконструировать продвижение индивидуальных клеток по клеточному циклу. Некоторые из них, такие как peco [22] и reCAT [23], как и вышеперечисленные методы, используют наборы известных генов, ассоциированных с клеточным циклом. Другие алгоритмы (Cyclum [24], CYCLOPS [25] и CCPE [26]) основаны на обучении без учителя (unsupervised learning/самообучение) и для расчёта «псевдовремени» клеточного цикла учитывают цикличность в экспрессии генов. При этом с помощью этих алгоритмов можно экстрагировать информацию о генах, наиболее ассоциированных с каждой из стадий клеточного цикла.

ОПРЕДЕЛЕНИЕ КЛЕТОЧНЫХ КЛАСТЕРОВ

Типичный анализ данных экспрессии генов в единичных клетках, полученных с помощью scRNA-seq, начинается с определения

клеточного состава исследуемого образца. На этом этапе происходит объединение индивидуальных клеток в транскрипционные кластеры на основе сходства их профилей экспрессии и определение клеточного типа каждого из кластеров по уровням экспрессии специфичных и дифференциально-экспрессирующихся маркерных генов. Для решения этих задач используются алгоритмы кластеризации данных и методы анализа дифференциальной экспрессии генов. Однако эти алгоритмы и методы применимы не только в контексте первоначального анализа данных scRNA-seq, их можно использовать и как самостоятельные инструменты для решения конкретных биологических задач.

Анализ ДЭГ применяется для исследования влияния различных патологических или экспериментальных условий на транскрипционные профили изучаемых популяций клеток. С помощью такого подхода были определены гены и ассоциированные с ними клеточные процессы, активируемые и подавляемые в разных типах клеток при COVID-19 [27, 28], болезни Альцгеймера [29], аутизме [30] и многих других заболеваниях. Кроме того, дифференциальный анализ находит применение в идентификации генов, чья экспрессия меняется при клеточной дифференцировке или другом динамическом процессе, затрагивающем фенотипы популяций клеток. Также данный метод может применяться для отбора генов, экспрессия которых необходима для построения клеточных траекторий.

Применение алгоритмов кластеризации также не ограничивается начальным этапом анализа данных scRNA-seq. Кластеризация клеток как элементов набора данных может быть проведена несколько раз последовательно при биоинформатическом анализе, например, с целью поиска редких популяций клеток [31]. При этом повторная кластеризация клеток может проводиться не по всему изначальному набору экспрессионных профилей, а, например, по генной подписи – определённому набору генов, по которым можно идентифицировать интересующий тип клеток – или по любому другому набору признаков, описывающих клетку.

Аннотация клеток проводится не только для определения их типов, но и для других задач. Например, кластеры клеток можно аннотировать по вирусной нагрузке и их состоянию [32].

Текущий и два последующих раздела настоящего обзора рассказывают о каждой биоинформатической задаче в отдельности: определении кластеров, поиске ДЭГ и идентификации клеточных типов.

Определение клеточных, или транскрипционных кластеров в данных scRNA-seq, осуществляемое с помощью алгоритмов кластеризации, – это одна из популярных задач обучения без учителя. Цель кластеризации заключается в объединении похожих по своим транскрипционным профилям клеток в группы, которые затем можно охарактеризовать как кластеры одного клеточного типа, одной стадии дифференцировки или клеточного цикла. Стоит понимать, что кластеры – это математически определяемые группы клеток, которые действительно могут состоять из клеток одного типа, однако на практике биология зачастую имеет малое отношение к алгоритмам кластеризации.

На настоящий момент существует множество применимых для кластеризации scRNA-seq методов, каждый из которых обладает своими преимуществами и недостатками. Консенсуса о наиболее эффективном методе до сих пор нет [33]. Методы кластеризации данных scRNA-seq основываются на широко применяемых алгоритмах, таких как иерархическая кластеризация, метод k-средних и кластеризация графов. На кластеризацию клеток оказывает влияние не только выбор метода и его параметров, но и увеличение количества измеряемых данных. При экспоненциальном возрастании их количества происходит увеличение размерности, что сопровождается возникновением «проклятья размерности». Многомерность данных приводит к сближению индивидуальных клеток, что вызывает проблемы в определении расстояния между клетками. Наиболее удалённые (различающиеся по экспрессии множества генов) клетки в многомерных данных находятся на таком же расстоянии, что и наиболее близкие (похожие по профилю РНК). Из-за этой проблемы стандартные методы кластеризации не могут разделить отличные по паттерну РНК клетки. Для многомерных данных scRNA-seq для снижения размерности и ускорения вычислений проводят отбор значимых генов и применяют метод главных компонент (principal component analysis, PCA). В качестве значимых генов (признаков) могут быть выбраны высоковариабельные гены (highly variable genes, HVG), чья экспрессионная изменчивость объясняется преимущественно биологическими различиями между анализируемыми клетками, а не техническим шумом [34]. Также существуют методы для моделирования технического шума и отбора значимых генов, например, на основе построенной модели M3Drop [35]. Для кластеризации данных затем используются расстояния между клетками в пространстве пониженной размерности.

Алгоритм иерархической кластеризации подразумевает под собой объединение клеток в растущие кластеры (восходящий, или агломеративный подход) или разбиение кластеров на подгруппы (нисходящий, или дивизионный подход) на основании матрицы расстояний между клетками. Программы, использующие этот алгоритм для идентификации клеточных субпопуляций, например, pcaReduce [36] или CIDR [37], позволяют определять взаимоотношения между кластерами с помощью дендрограммы, но по сравнению с другими методами кластеризации работают медленнее, что может быть критично для больших объёмов данных scRNA-seq.

Метод кластеризации данных scRNA-seq SC3 [38] использует в своей основе алгоритм k-средних, идея которого заключается в итеративном определении центров масс (центроидов) заданного числа кластеров и уточнении границ кластеров. Отличительной особенностью методов, основанных на алгоритме k-средних, является получение кластеров примерно одного размера. В такой ситуации большие субпопуляции клеток будут разбиваться на несколько кластеров, а редкие клетки будут объединены с другими кластерами. Кроме того, результат кластеризации с помощью k-средних во многом зависит от случайной инициализации центроидов и не обязательно представляет собой глобальный минимум.

В случае больших данных scRNA-seq для более оптимального решения проблемы определения клеточных кластеров предлагают методы кластеризации, основанные на поиске кластеров в графах – лувенский метод (Louvain method) и алгоритм Лейдена (Leiden method). Данные методы предварительно строят граф k ближайших соседей (kNN-граф), используя проекцию данных в пространство пониженной размерности, и затем определяют кластеры клеток как группы наиболее плотно связанных друг с другом вершин графа. Количество определяемых кластеров напрямую не задаётся, но на него влияет значение параметра разрешения, а также указанное число k ближайших соседей, используемое для построения графа. Графовый подход для идентификации субпопуляций реализован в программе PhenoGraph [39] и программных пакетах Seurat [18] и Scanpy [19]. Более подробно преимущества и недостатки каждой из групп методов описаны Kiselev et al. [33].

С помощью кластеризации данных scRNA-seq можно обнаружить уже известные типы или состояния клеток с характерной высокой экспрессией ряда генов-маркеров или определить ранее не описанные популяции клеток, существование которых впоследствии подтверждается дополнительными экспериментами, например, иммуноокрашиванием [40, 41]. При этом для поиска редких и малочисленных субпопуляций клеток возможно использование более сложных подходов для нормализации данных [41] и ре-кластеризации отобранных, одного или нескольких наиболее интересных, кластеров [31]. Избавиться от биологического шума в данных scRNA-seq, вызванного контаминацией исследуемой ткани, возможно с помощью удаления кластера нецелевых клеток с последующей ре-кластеризацией данных [40].

АНАЛИЗ ДИФФЕРЕНЦИАЛЬНОЙ ЭКСПРЕССИИ

Анализ дифференциальной экспрессии генов позволяет установить различия между кластерами и определить клеточные типы и их маркеры. Дифференциальная экспрессия позволяет изучить транскрипционную динамику в процессе дифференцировки клеток, развития заболеваний или при воздействии каких-либо экзогенных и/или эндогенных факторов.

Несмотря на то что задачи поиска ДЭГ на основе данных bulk RNA-seq и scRNA-seq формулируются схожим образом, методы их решения различны. Методы анализа дифференциальной экспрессии по данным bulk RNA-seq, такие как DESeq2 [42] и edgeR [43], учитывают сложность в определении дисперсии экспрессии генов на основании малого количества повторностей – биологически различных образцов из одной экспериментальной группы. Такой проблемы в scRNA-seq не существует, так как каждую клетку можно рассматривать как повторность. Однако увеличение количества измерений (клеток) может привести к ложноположительным результатам. Например, в одном из кластеров экспрессия гена может быть повышена статистически значимо, но всего в 1,1 раза. Если уменьшить количество измерений, статистическая значимость может быть полностью нивелирована. В связи с этим следует находить ДЭГ, повышенные до определённого уровня, причём для каждых конкретных данных значение этого порога будет различно. Тем не менее, вышеуказанные методы оказываются применимыми и для анализа данных scRNA-seq. Так, для анализа ДЭГ с помощью DESeq2 и edgeR из данных scRNA-seq используют процедуру взвешивания значений экспрессии для каждой клетки

и каждого гена, которая позволяет преодолеть проблему высокой разреженности данных scRNA-seq ввиду чрезвычайно высокой доли генов с нулевой экспрессией в каждой клетке. Ранее методы, специально разработанные для анализа scRNA-seq, решали эту проблему за счёт использования отрицательной биномиальной модели с избытком нулей (zero-inflated negative binomial, ZINB) [44]. Но сегодня для scRNA-seq рекомендуется использовать отрицательную биномиальную модель без избытка нулей (negative binomial, NB) [45]. Кроме того, при сравнении эффективности методов анализа дифференциальной экспрессии было показано, что при простом дизайне эксперимента (сравнение контроля и «эксперимента» в отсутствии других переменных) лучшие результаты показывает непараметрический тест Манна–Уитни [46].

Однако вышеперечисленные методы не учитывают ряд особенностей, характерных для данных scRNA-seq. Так, распределение экспрессии генов в единичных клетках отличается бимодальностью. Значения экспрессии либо положительны в тех клетках, где соответствующий транскрипт был детектирован, либо равны нулю в клетках, где по некоторой биологической или технической причине экспрессия не была детектирована. Отсутствие экспрессии может объясняться её стохастической природой (биологическая причина) или потерей транскрипта на этапе подготовки кДНК-библиотек (техническая причина). Данная особенность транскриптомики единичных клеток принята во внимание в таких методах, как SCDE [47] и MAST [48]. Метод SCDE использует комбинацию отрицательного биномиального распределения для положительных значений экспрессии и Пуассоновского для «нулевых» генов, в случае которых может присутствовать также фоновый сигнал. Метод SCDE возможно использовать только для поиска ДЭГ между двумя группами клеток (контроль и «эксперимент»). Использование в SCDE других переменных, таких как номер группы образцов (batch-переменная) или временная точка, невозможно, что ограничивает применимость данного метода только для экспериментов с простым дизайном. Метод MAST учитывает более сложные дизайны экспериментов, например, с несколькими вариантами «воздействия», и использует модель преодоления препятствий (hurdle model) для описания экспрессии генов в единичных клетках.

Поиск ДЭГ между субпопуляциями клеток зависит от этапа кластеризации, в ходе которого используется та же самая информация об экспрессии генов в индивидуальных клетках. По этой причине анализ дифференциальной экспрессии, следующий за кластеризацией клеток, приводит к искусственно заниженным значениям статистической значимости (*p*-значение). Данная проблема решается с помощью теста TN (truncated normal), который учитывает уровни экспрессии генов, определяющие границы между клеточными кластерами [49].

Идея о том, что каждая индивидуальная клетка представляет из себя независимый образец, в действительности является серьёзным допущением, поскольку все клетки в образце обычно объединены общим происхождением от одного или нескольких доноров. Поэтому при сравнении субпопуляций клеток, полученных из одного организма, происходит оценка вариабельности транскрипционных профилей не в популяции, а только внутри одного индивидуума. Если в наборе данных присутствует выборка клеток, полученных от нескольких доноров, то для анализа дифференциальной экспрессии можно использовать подход с подсчётом агрегированной (псевдобалк, pseudobulk) экспрессии или обобщённую линейную смешанную модель, в которой донор указан как случайный фактор [50]. Агрегированная экспрессия рассчитывается путём суммирования или усреднения экспрессии гена среди клеток каждого из доноров, в результате чего получается несколько псевдоповторностей, соответствующих независимым донорам. Таким образом, с помощью данного подхода задача анализа ДЭГ на основе данных scRNA-seq сводится к задаче, которая уже успешно решается для bulk RNA-seq.

Однако результаты анализа дифференциальной экспрессии, полученные с использованием bulk RNA-seq и scRNA-seq, могут значительно различаться. В случае scRNA-seq дифференциальная экспрессия детектируется преимущественно среди генов с высоким средним уровнем экспрессии, но средними значениями разницы в экспрессии между сравниваемыми группами (fold change) [51]. В связи с этим с помощью scRNA-seq-подходов оказывается труднее идентифицировать гены с сильным изменением уровня экспрессии в ответ на изучаемое воздействие.

Из-за особенностей протоколов подготовки scRNA-seq-библиотек, а именно использования олиго(dT)-праймеров в подавляющем большинстве методов, в транскриптомах единичных клеток детектируются преимущественно полиаденилированные РНК – мРНК и некоторые длинные некодирующие РНК

(днРНК). Для анализа экспрессии других типов РНК на уровне единичных клеток, например, микроРНК (миРНК), используются специальные методы пробоподготовки образцов [52]. Тем не менее, судить об активности микроРНК в индивидуальных клетках возможно, используя в качестве такой оценки уровень экспрессии предшественников микроРНК, которые могут полиаденилироваться и кэпироваться, тем самым обеспечивая своё попадание в секвенируемый транскриптом.

С помощью анализа дифференциальной экспрессии с использованием scRNA-seq на клеточном уровне были определены причины нарушения регенерации эпителия альвеол лёгких при COVID-19 [27, 53] и охарактеризованы особенности цитокинового шторма и иммунного ответа на вирусную инфекцию, который обеспечивается разными типами иммунных клеток [54]. Выявлены маркеры, вовлечённые в лекарственную резистентность и прогрессирование саркомы Юинга [55]. Оценена функциональная гетерогенность мультипотентных стромальных клеток человека и мыши по характеру экспрессии аденилатциклаз [56]. Изменения экспрессии генов в процессе клональной экспансии и возможного сопутствующего истощения Т-клеток при противоопухолевой иммунотерапии также были изучены на уровне единичных клеток с помощью методов дифференциальной экспрессии [57]. Эти же методы используются для поиска генных сигнатур, то есть наборов предиктивных маркеров, которые могли бы предсказывать ответ на иммунотерапию с использованием ингибиторов контрольных точек иммунного ответа [58].

ИДЕНТИФИКАЦИЯ КЛЕТОЧНЫХ ТИПОВ

Общепринято, что идентификация типов клеток выполняется стандартными гистологическими методами, из которых основным является иммуноокрашивание – связывание антител с белковыми маркерами клеток и последующая визуализация. Альтернативой может быть РНК-секвенирование и последующее биоинформатическое типирование, когда маркеры клеток детектируются на уровне транскриптов. Однако хорошо известно, что наличие мРНК в клетке не всегда коррелирует с продукцией соответствующего белка ввиду обилия посттранскрипционных и посттрансляционных механизмов регуляции [59, 60]. Типирование клеток, основанное на scRNA-seq, может быть автоматическим или ручным.

Автоматическое типирование. Автоматическое типирование осуществляется за счёт сравнения клеток изучаемого образца с известными маркерными генами, информация о которых представлена в различных базах, содержащих данные микрочипов, bulk RNA-seq или scRNA-seq для клеток определённого типа. В случае совпадения профилей экспрессии программа автоматически определяет тип клетки. По такому принципу работает R-пакет SingleR [61], в состав которого входит пакет celldex, содержащий доступ к семи клеточным базам. Для типирования клеток в автоматическом режиме также возможно использовать инструменты ScType [62], scCATCH [63], scSorter [64] и SCINA [65]. В своей работе автоматические аннотаторы могут использовать ранее проведённую кластеризацию или проводить её перерасчёт в соответствии с клеточными типами, которые были обнаружены в образце. Важно отметить, что автоматические аннотаторы способны распознавать ограниченное число клеточных типов ввиду отсутствия данных об экспрессионных профилях множества типов клеток в используемых базах.

Ещё одним способом аннотировать клетки в автоматическом режиме является использование аннотированных образцов других исследовательских групп. В биоинформатике эта методика называется label transferring [66]. Суть этого метода состоит в том, что в исследуемом образце сначала находятся клетки, совпадающие по паттерну экспрессии с образцом-эталоном. После того как найдены якорные клетки (совпадающие между образцами), на определяемый образец переносится информация о типе клетки с образца-эталона. По такому принципу работает веб-сервис Azimuth [67]. На сегодняшний день в веб-сервисе доступны 11 наборов эталонных данных.

Существует и промежуточный вариант автоматического типирования с созданием собственной библиотеки аннотированных образцов именно тех клеток, с которыми работает исследователь. Этот подход позволяет самостоятельно отобрать самые лучшие открытые данные и иметь чёткое представление о дизайне эксперимента, в котором эти образцы были получены.

Ручное типирование. Необходимость ручного типирования прежде всего обусловлена нахождением большей части клеток исследуемого образца в промежуточных, недифференцированных формах. Такие клетки, как правило, не имеют специфических маркеров, характерных для их дифференцированных форм,

и не могут быть проаннотированы системами автоматического типирования. Кроме того, в реальной практике классических генов-маркеров может быть недостаточно для идентификации и дифференцированных форм. В таких случаях тип клеток может быть определён вручную, на основании менее известных или заданных пользователем маркерных генов [68]. Также, по нашему мнению, типирование клеток может быть основано на анализе их вовлечённости в различные биологические процессы, переходных генов или положения клеток исследуемого образца относительно траектории развития.

Типирование по менее известным или заданным пользователем специфическим маркерам в большей мере подходит для определения типа дифференцированных клеток и осуществляется за счёт изучения списка генов каждого кластера, полученного после этапа кластеризации. Исследователь визуально оценивает список высокопредставленных генов на предмет наличия определённых маркеров и при условии их наличия аннотирует клетки к известному типу. Ещё одной возможной реализацией данного подхода является присвоение кластерам клеточного типа на основании заданных пользователем маркеров в Seurat и Scanpy. Маркерные гены могут быть выбраны пользователем на основании литературных данных. Так, использование заранее заданной панели генов позволило выявить типы клеток при сравнительном анализе идентичных регионов мозга высших приматов [69]. Данный вид типирования использовался и при идентификации субпопуляций клеток фолликулярной лимфомы, которые выявляются при прогрессировании и рецидивировании заболевания [70].

Типирование по биологическим процессам основано на выявлении групп генов, участвующих в определённых биохимических процессах, специфичных для некоторых клеток в контексте индуцирующего воздействия. По списку ДЭГ можно определить биологические процессы, которые активны в данном кластере клеток. Для этого очень удобно использовать веб-сервис g:Profiler, который объединяет информацию о ДЭГ кластера и определяет все биологические процессы, сигнальные пути и клеточные компоненты, за которые ответственны белковые продукты этих генов. С помощью данного подхода становится возможным типирование клеток, находящихся в процессе дифференцировки на основании детекции маркеров, ассоциированных с изменением клеточного фенотипа.

Типирование по переходным генам, в отличие от первых двух способов ручного типирования, помимо белок-кодирующих транскриптов, учитывает несплайсированные формы будущих мРНК. Соотношение сплайсированных и несплайсированных форм мРНК позволяет оценить, в каком состоянии находится экспрессия того или иного белка на момент исследования – индуцированном или репрессированном, и выделить те гены, которые являются ключевыми для развития клетки на момент анализа, например, с помощью пакета scVelo [71]. Среди данных генов вручную осуществляется поиск ответственных за переход клетки в дифференцированную форму. Таким образом, исследователь может предположить, предшественником какого типа является исследуемая группа клеток.

Ручное типирование клеток также может осуществляться на основании результатов вывода траектории развития. В большинстве случаев при выводе данной модальности кластеры с отсутствием специфических маркеров находятся между кластерами с наличием таковых. В таком случае можно предположить, что данный кластер является промежуточным и содержит клетки в переходном состоянии между исходной и конечной формами.

ТРАЕКТОРИИ РАЗВИТИЯ И СКОРОСТЬ РНК

Любой вид секвенирования является снимком момента жизни клетки, предоставляющим информацию об интересующей модальности на момент проведения исследования. Библиотека scRNA-seq содержит информацию о транскрипционном профиле нескольких сотен и тысяч клеток, гетерогенность которых в том числе обусловлена динамическим процессом клеточного развития. Методы вывода траекторий развития, также называемые анализом псевдовремени, позволяют упорядочивать клетки исследуемого образца вдоль смоделированной временной траектории на основе сходства их паттернов экспрессии. Результатом построения траектории развития в псевдовремени является графическое изображение всех клеток образца, расположенных друг за другом, начиная с начальной/исходной клетки (root cell) и до конечной или дифференцированной клетки (end cell). С помощью вывода траекторий развития становится возможным изучение интересующего биологического явления, например, путей дифференцировки, клеточного цикла или иммунных реакций, в динамическом контексте.

Впервые для построения траекторий развития был предложен R-пакет Monocle. Monocle сначала использует тест дифференциальной экспрессии для уменьшения количества генов, а затем применяет анализ независимых компонентов для дополнительного уменьшения размерности. Для построения траектории Monocle вычисляет минимальное остовное дерево, а затем находит самый длинный соединённый путь в этом дереве. Ячейки проецируются на ближайшую к ним точку на этом пути [72]. После Monocle было предложено ещё более 50 различных методов, самыми известными из которых стали TSCAN [73] и Slingshot [74]. Методы отличаются друг от друга по многим параметрам: указание начальных и конечных клеток; тип визуализации графа (прямой, линейное псевдовремя, циклическое псевдовремя, вероятность конечного состояния, кластерная оценка, ортогональная проекция и клеточный граф); тип траектории (несвязный и связный граф, цикличный и ациклический граф, древовидный граф) [75]. На сегодняшний день методов стало настолько много, что появилась необходимость создания единой платформы, где можно анализировать свои данные с помощью сразу нескольких методов. Одной из таких платформ является dynverse, объединившая в себе 45 методов построения траекторий развития.

Чтобы улучшить качество выводимых траекторий, в некоторых методах вместо подсчёта экспрессии генов или в дополнение к ним используются дополнительные источники информации, наиболее популярным из которых на сегодняшний день является скорость РНК (RNA velocity) [71]. Идея RNA velocity возникла при изучении данных scRNA-seq, полученных на различных платформах (Smart-seq2, STRT/C1, inDrop и 10x Genomics Chromium). Оказалось, что от 15 до 25% прочтений содержат несплайсированные интронные последовательности, что объясняется наличием поли(А)-участков не только в поли(А)-хвосте, но и в поли(А)-вставке [76]. В связи с таким наблюдением было предложено при анализе данных учитывать как сплайсированные, так и несплайсированные формы мРНК. Под скоростью в данном случае понимается производная по времени от стадии экспрессии гена. Весь процессинг был поделен на три стадии: транскрипция, сплайсинг и деградация. Экспрессия гена констатируется в случае преобладания транскрипции и сплайсинга над деградацией и ингибируется, если деградация преобладает над транскрипцией и сплайсингом. Значение скорости определяет направление вектора каждой клетки в пространстве со сниженной размерностью, так формируется векторное поле, в котором можно видеть направление развития клеток в образце. Учитывая то, что векторное поле накладывается на заранее полученные кластеры клеток, можно предполагать направление дифференцировки или восприятия клетками фактора воздействия.

МЕЖКЛЕТОЧНАЯ КОММУНИКАЦИЯ

Развитие, функционирование, регенерация и гомеостаз тканей и органов обеспечиваются путём межклеточной коммуникации, или межклеточного сигналинга – процесса, происходящего за счёт лиганд-рецепторного взаимодействия различных клеток. В роли лигандов могут выступать цитокины, хемокины, гормоны, факторы роста и нейромедиаторы.

Межклеточный сигналинг принято делить на аутокринный (выделяемый клеткой лиганд взаимодействует с рецептором той же клетки), паракринный (выделяемый клеткой лиганд взаимодействует с рецепторами клеток из той же ткани) и эндокринный (выделяемый клеткой лиганд взаимодействует с рецепторами клеток из других тканей или органов). Отдельно можно выделить межклеточные взаимодействия, то есть физический контакт двух клеток друг с другом. Межклеточные взаимодействия могут быть как участниками межклеточной коммуникации (при так называемом межклеточном распознавании), так и выполнять исключительно структурную функцию.

Изучение межклеточной коммуникации помогает понять механизмы дифференцировки и морфогенеза клеток, этиологию заболеваний [77] и особенности формирования иммунного ответа [78]. Понимание межклеточного сигналинга позволяет разрабатывать новые терапевтические стратегии [79] и прогнозировать тяжесть течения различных заболеваний [80, 81].

Исследования межклеточного сигналинга берут своё начало с определения белок-белковых взаимодействий при помощи двугибридных систем, коиммунопреципитации и иных методов [82]. С их помощью накоплен целый пласт экспериментально подтверждённых лиганд-рецепторных взаимодействий, который, однако, был получен лишь для конкретных типов клеток в конкретных тканях. ScRNA-seq позволяет оценивать уровни экспрессии генов лигандов и рецепторов в тысячах клеток за один эксперимент и не только изучать клеточный состав ткани, но и на системном

уровне оценивать возможные паракринные и аутокринные регуляции.

Анализ межклеточного сигналинга по данным scRNA-seq ставит перед собой задачу понять, коммуницирует ли определённая пара типов клеток A-B по определённому каналу лиганд-рецептор l-r. Простые методы, такие как iTalk [83] и CellTalker, решают эту задачу следующим образом: если ген лиганда l дифференциально активирован в типе клеток A, а ген рецептора r дифференциально активирован в типе клеток B, то такие клетки считаются взаимодействующими. Эти методы интуитивно понятны и легко интерпретируемы, однако они нечувствительны к коммуникациям, которые характерны для большого числа типов клеток ткани.

В более сложных методах вводится понятие силы, или активности коммуникации S, которую оценивают как функцию от средних экспрессий l в A (lA) и r в B (rB) – от их суммы (метод CellPhoneDB [84]) либо от их произведения (SingleCellSignalR [85]). Алгоритм CellCall [86] для оценки активности коммуникации между клетками A и B дополнительно использует информацию об экспрессии регулона *RegB* (набор генов-мишеней транскрипционного фактора, которые коэкспрессируются вместе с транскрипционным фактором), находящегося под регуляцией транскрипционного фактора, который активируется при воздействии на клетку через рецептор r. Отдельно стоит упомянуть случаи, когда рецептор состоит из нескольких субъединиц, кодируемых разными генами. В таком случае за r будет взята либо минимальная экспрессия среди всех субъединиц рецептора (CellPhoneDB), либо их среднее геометрическое (CellCall).

Однако не все клетки, коэкспрессирующие пару лиганд-рецептор, коммуницируют в реальности. Одним из способов преодоления ложноположительных результатов является пермутационный тест (реализован в CellPhoneDB), в ходе которого метки клеточных типов множество раз случайно перемешиваются, а сила коммуникации S считается заново, задавая нулевое распределение, по которому будет рассчитываться *p*-значение для исходного S. Минус такого подхода схож с минусами подходов, основанных на дифференциальной экспрессии: широко распространённые в исследуемом наборе данных коммуникации могут оказаться статистически незначимыми. Иное решение этой проблемы реализовано в SingleCellSignalR: авторы этого алгоритма предлагают считать значимыми все коммуникации с силой выше некоторого установленного ими порогового значения. Другой алгоритм CellCall предполагает, что коммуникация значима, если ожидаемая доля ложных отклонений анализа обогащения набора генов (FDR GSEA) регулона *RegB* меньше 0,05.

Особо следует выделить алгоритм scTensor [87], в котором сначала из данных формируется тензор третьего ранга размерности A × A × L, где A – количество типов клеток, L – количество исследуемых пар лиганд-рецептор, а (a, b, l)-й элемент этого тензора – сила коммуникации клеток A и B при помощи пары лиганд-рецептор l-r. Т.е. тензор состоит из всех попарных сил коммуникации всеми возможными парами лиганд-рецептор. Сила коммуникации в данном методе рассчитывается как простое произведение lA и rB. Сконструированный тензор преобразуется в произведение трёх матриц и нового тензора при помощи неотрицательного разложения Таккера. В результате информация о межклеточном сигналинге описывается сразу для всего набора данных, и это позволяет увидеть более комплексные эффекты, в частности, включающие в себя целые коммуникационные сети. Несмотря на свои достоинства, этот метод не пользуется большой популярностью в первую очередь из-за сложности интерпретации результатов.

Описанные выше подходы к определению межклеточного сигналинга принципиально отличаются в первую очередь гипотезами, которые они тестируют. CellTalker, iTalk и CellPhoneDB позволяют определить сигналинги, уникальные для некоторых типов клеток в исследуемом наборе данных. SingleCellSignalR, CellCall и scTensor позволяют детектировать большее число коммуникаций, в том числе и неспецифичных, однако могут оказаться нечувствительными в случаях, когда сила коммуникации низкая [88]. Кроме того, все перечисленные выше методы говорят только о возможных путях сигналинга, которые необходимо в дальнейшем валидировать экспериментально, и результат работы данных инструментов сильно зависит от базы лиганд-рецепторных взаимодействий. Более значимое подтверждение коммуникации между различными клетками можно получить при помощи бурно развивающихся методов пространственной транскриптомики [89], которые, по всей видимости, позволят точно ответить на множество вопросов о том, каким образом формируется и поддерживается архитектура различных тканей.

ГЕННЫЕ РЕГУЛЯТОРНЫЕ СЕТИ

Регуляция экспрессии генов внутри клетки осуществляется за счёт сложного сочетания процессов синтеза и сплайсинга РНК, а также деградации уже зрелой мРНК. В основном уровень экспрессии генов тесно связан с активностью транскрипции мРНК. Транскрипция, в свою очередь, регулируется за счёт воздействия на клетки различного рода сигналов. Например, гормоны, воздействуя на специфические рецепторы, запускают сигнальные каскады, локализующиеся в основном в цитоплазме клетки. Сигнальные каскады запускают транскрипционные факторы, которые взаимодействуют с сайтами связывания на генах-мишенях. Эти взаимодействия осуществляются в ядре клетки и называются генными регуляторными сетями (gene regulatory networks). Именно генные регуляторные сети осуществляют поддержание клеточного гомеостаза, формирование клеточной гетерогенности, а их нарушение может приводить к развитию различных патологических состояний и утяжелять течение заболеваний [90, 91]. Изучение генных регуляторных сетей улучшает понимание механизмов различных биологических процессов в живых организмах и позволяет разрабатывать новые терапевтические стратегии для борьбы с заболеваниями.

Построение генных регуляторных сетей из данных scRNA-seq может осуществляться на основании регрессионных моделей, корегуляторных взаимодействий и вывода траектории развития.

Подходы на основе регрессии работают с конкретными списками генов и позволяют оценивать связь между регуляторами и генами-мишенями, а также делают вывод об интенсивности этого взаимодействия. Первоначально для построения генных регуляторных сетей на основе регрессии был разработан метод GENIE3 [92]. Этот метод широко используется для построения генных сетей из данных bulk RNA-seq и scRNA-seq. Однако применение GENIE3 невозможно в случаях scRNA-seq, когда количество исследуемых клеток исчисляется тысячами. Данная проблема была успешно решена с помощью градиентного бустинга в методе GRNBoost2 [93]. Тем не менее значительным недостатком анализа регуляторных сетей, выведенных из регрессионного анализа отдельных клеток, является большее количество ложноположительных связей по сравнению с анализом bulk RNA-seq. Использование инструмента SCENIC [94] позволяет преодолеть данный недостаток за счёт отбора связей между регуляторами и генами-мишенями, в которых гены-мишени имеют предполагаемый сайт связывания с соответствующими транскрипционным факторами. При этом транскрипционный фактор вместе с активируемыми генами-мишенями называют регулоном (regulon).

Построение генных сетей, основанных на корегуляторных взаимодействиях, подразумевает подсчёт корреляции экспрессии генов в единичных клетках с помощью коэффициента Пирсона и рангового коэффициента Спирмена и реализуется посредством анализа взвешенных сетей коэкспрессии генов (WGCNA) [95]. Коэкспрессионные модули соотносятся с функциями генов с помощью метода GSEA [96] и баз данных, таких как STRING [97] и HumanNet [98]. Применение WGCNA на данных scRNA-seq позволяет идентифицировать функциональные модули и составляющие их ключевые гены для каждого типа клеток, которые могут быть связаны с конкретным физиологическим или патофизиологическим состоянием [99]. Ключевые гены имеют наибольшее количество корреляционных связей в плане коэкспрессии и в большей степени определяют функциональную принадлежность модулей. Выявление таких генов позволяет обнаруживать, например, факторы, связанные с устойчивостью к химиотерапии [100], или прогностические маркеры [101].

Анализ траекторий развития позволяет рассмотреть данные секвенирования единичных клеток как динамическую систему, что даёт возможность выйти за пределы статической природы транскриптома и получить псевдовремя для последующего построения генных сетей с помощью метода обычных дифференциальных уравнений. Такие сети отражают генные взаимодействия в динамике, т.е. изменение экспрессии генов в течение непрерывного псевдовремени характеризуется функцией, которая включает активирующее или подавляющее влияние других генов в качестве переменных [102]. Данный подход наиболее точно описывает генные взаимодействия в непрерывных процессах, таких как дифференцировка, и реализован в инструменте SCODE [103].

АНАЛИЗ CNV

CNV вносят важный вклад в генетическую изменчивость живых организмов и определяют предрасположенность к различным заболеваниям. К основным критериям определения структурного варианта как CNV относят

повторяемость, числовую изменчивость и «значительную» длину. Несмотря на заданные критерии, границы между типами структурных вариантов формируются по-разному в различных работах, поэтому некоторые CNV соответствуют одновременно нескольким категориям [104]. В настоящее время многие исследователи определяют CNV как несбалансированные хромосомные перестройки – делеции и вставки участков ДНК, размеры которых варьируют от нескольких килобаз до целых хромосом и могут включать мобильные элементы и некодирующие последовательности [105]. Соответственно, в зависимости от размера CNV могут быть фокальными и полнохромосомными. Последние генерируются анеуплоидными клетками с аномальным числом хромосом и ведут к изменению уровня транскрипции большого количества генов. CNV могут быть представлены как нейтральными, так и патогенными формами. Патогенность определяется прямым влиянием CNV на экспрессию генов и/или образованием новых белковых продуктов [106].

Классическими методами для идентификации CNV являются микроматричная сравнительная геномная гибридизация (aCGH), мультиплексная амплификация лигированных зондов (MLPA) и секвенирование следующего поколения (NGS), главным образом полногеномное секвенирование (WGS). Однако aCGH и MLPA ограничены разрешением чипа (связанным с охватом и плотностью флуоресцентных зондов) и неспособны детектировать копий-нейтральные потери гетерозиготности. Стоимость, продолжительность обработки данных и высокие вычислительные требования усложняют проведение анализа CNV с помощью WGS [107].

Существует лишь небольшое количество методов, созданных для идентификации CNV по данным scRNA-seq. Все эти методы основаны на предположении, что дифференциальная экспрессия генов коррелирует с CNV [108]. Метод inferCNV основан на усреднении уровня экспрессии генов и сравнении профиля CNV изучаемого образца с эталонным. Несмотря на то что такой метод с высокой точностью выявляет клональные изменения на уровне плеч хромосом, inferCNV с трудом удается идентифицировать субклональные изменения. Результаты, получаемые с помощью inferCNV, также высокочувствительны к выбору эталонных клеток. Ввиду этого необходима независимая нормализация различных клеточных типов с помощью соответствующих эталонных клеток [109]. На конечные результаты оказывает влияние и отсутствие данных об относительной нормализованной мере соотношения измеренных интенсивностей двух аллелей (BAF), что приводит к повышенному количеству ложноположительных результатов.

В других инструментах для анализа CNV реализован подход объединения генетической и транскрипционной информации. Например, метод HoneyBADGER [109], использующий байесовский подход с интегрированной скрытой марковской моделью, рассчитывает отклонение доли аллелей гетерозиготных вариантов от ожидаемой и определяет регионы CNV. Для защиты от ложноположительных результатов для предсказанных регионов оценивается апостериорная вероятность принадлежности области CNV заданному состоянию. При использовании HoneyBADGER необходимо предварительно определять SNV, так как инструмент подтверждает наличие CNV в регионах-кандидатах на основании моноаллельного характера экспрессии SNV в данных регионах. Другой метод, CaSpER [110], использует многомасштабную декомпозицию для сглаживания сигналов экспрессии и аллельного сдвига (allelic shift), благодаря чему большая часть шума удаляется. Ввиду того, что данный инструмент генерирует профиль сигнала аллельного сдвига из выровненных прочтений, определение SNV не требуется. Однако поскольку сигнал сдвига частоты альтернативного аллеля вычисляется путём объединения всех прочтений, клетки, имеющие большое количество прочтений, могут доминировать в сигнале сдвига и быть основным фактором искажения результата. Перечисленные методы были разработаны для анализа полноразмерных транскриптов, однако были валидированы для данных секвенирования одноконцевых транскриптов [109, 110]. Для последнего был разработан инструмент CopyKAT [111] с интегративным байесовским подходом и иерархической кластеризацией. Данный метод в большей степени подходит для анализа опухолевых клеток, которые часто являются анеуплоидными. Так, данный метод показал свою эффективность в идентификации опухолевых и гибридных клеток среди циркулирующих эпителиальных клеток у больных раком молочной железы [112].

В то время как инструменты по поиску CNV, используемые при WGS, основаны на равномерном покрытии генома прочтениями, при scRNA-seq сигнал концентрируется только на экзонных участках. В этом плане рекомендуется проведение анализа аллельного дисбаланса для понимания корреляции между геномом и транскриптомом. Однако отличие

настоящих генетических вариантов от технических артефактов достаточно осложнено из-за выпадения аллелей, неоднородности и низких показателей глубины секвенирования [113]. Таким образом, на данный момент анализ плоидности ДНК и идентификация анеуплоидии в scRNA-seq даёт более корректные результаты, чем нахождение фокальных CNV.

ИДЕНТИФИКАЦИЯ ОДНОНУКЛЕОТИДНЫХ ЗАМЕН

Как и CNV, однонуклеотидные варианты составляют генетическую изменчивость живых организмов, влияют на протекание биологических процессов и могут выступать в роли генетических факторов предрасположенности к заболеваниям. Идентификация SNV возможна посредством использования различных молекулярно-генетических методов, основными из которых являются полимеразная цепная реакция, микроматричный анализ, секвенирование по Сэнгеру и NGS. Для обнаружения SNV на уровне отдельных клеток классическим методом является секвенирование ДНК. Наиболее информативным и концептуально верным для этой цели является scDNA-seq, реализованный в платформе Tapestri (Mission Bio). Однако анализ SNV также возможно проводить на основе данных scRNA-seq, тем самым получая одновременно информацию и об экспрессии генов. Основным ограничением является анализ SNV только в белок-кодирующих участках (экзонах), поскольку в качестве исходного материала для scRNA-seq чаще всего выступает матричная РНК. Более того, различные паттерны экспрессии генов и альтернативный сплайсинг существенно ограничивают доступную для анализа белок-кодирующую область генома. Другим важным моментом является то, что при анализе экспрессии генов чаще всего применяется короткое одноконцевое секвенирование с 5'- или 3'-конца, что опять же исключает из анализа значительную часть финальной библиотеки. Эта проблема наиболее остро возникает в случае с 3'-секвенированием, где прочтение затрагивает только небольшой участок с поли(А)-хвоста мРНК, и большая часть экзонных последовательностей, наиболее интересных для SNV-анализа, теряется. В случае 5'-секвенирования при достаточной экспрессии интересующего участка экзома проблема частично решается, если в процессе пробоподготовки происходит случайная фрагментация захваченной за поли(А)-хвост мРНК и конверсия образовавшихся фрагментов в цепь кДНК, на основе которых и будет происходить подготовка библиотеки. Другим ограничением при SNV-анализе данных scRNA-seq может быть выпадение одного исследуемого аллеля (allelic dropout), в частности, при использовании технологии масляных капель для изоляции отдельных клеток перед баркодированием и амплификацией целевых молекул, что затрудняет идентификацию гетерозиготных субпопуляций клеток. Таким образом, при планировании исследования SNV на основе данных scRNA-seq необходимо учитывать указанные ограничения и по возможности использовать двуконцевое прочтение при секвенировании с последующей верификацией находок классическими молекулярно-генетическими методами.

Идентификация SNV на основе scRNA-seq в большинстве своём осуществляется методами, разработанными для анализа данных секвенирования ДНК: SAMtools, GATK, CTAT, FreeBayes, MuTect2, Strelka2, VarScan2 и др. Общая схема работы данных алгоритмов заключается в четырёх последующих операциях: картирование на референсный геном, предобработка, идентификация вариантов и фильтр ложноположительных вариантов. Для картирования чаще всего используют алгоритм STAR, рекомендованный GATK Best Practices [114]. Для анализа данных scRNA-seq может дополнительно использоваться инструмент GSNAP, позволяющий работать с короткими и сложнокартируемыми последовательностями [115]. Предобработка предназначена для удаления дубликатов, повторного выравнивания и базовой оценки качества прочтений. Выявление генетических вариантов проводится на основе расхождения нуклеотидных последовательностей с референсом и удаления вариантов с низким качеством или недостаточным покрытием. Хотя MuTect2, Strelka2 и VarScan2 применяются в основном для секвенирования ДНК, а также и bulk RNA-seq, выявленные с помощью данных алгоритмов варианты могут быть соотнесены с кластерами единичных клеток на основе других инструментов, например, VarTrix, с целью вывода связи генотип-фенотип. Стоит отметить, что большинство описанных алгоритмов, за исключением SAMtools [116], в той или иной степени работают на основе GATK. Более детальное сравнение описанных пайплайнов представлено в обзоре Liu et al. [117]. Что касается SAMtools, то в литературе имеется сообщение о применения инструмента Pysam, функционирующего на основе SAMtools, для

детекции вариантов в митохондриальной ДНК методом scRNA-seq [118].

При использовании разных платформ для подготовки библиотек при scRNA-seq необходимо учитывать их разную «пропускную способность», то есть количество клеток, которое можно проанализировать за один запуск и, как следствие, число прочтений на одну клетку. Так, например, для Fluidigm C1 (1000 клеток за запуск) значение глубины секвенирования может достигать 1 миллиона ридов на клетку, а для 10x Genomics Chromium (до 10 000 клеток за запуск) глубина секвенирования в реальной практике редко превышает 10–20 тысяч. Это приводит к тому, что существует вероятность не обнаружить варианты со слабой экспрессией и субклональные SNV. Дальнейшее увеличение количества прочтений может быть малоэффективным в случае малой «сложности» библиотеки и большой скорости «насыщения» секвенирования (sequencing saturation rate). В частности, это связано с тем, что подавляющее число ридов будет картироваться на ограниченную группу сильно представленных транскриптов, а детекция слабо экспрессирующихся генов и вариантов в них потребует сильного увеличения глубины прочтения. Такая ситуация, например, описана для 10x Genomics scRNA-seq мононуклеаров периферической крови, где показатель «насыщения» секвенирования составлял более 90%. Повысить вероятность детекции SNV в таких случаях можно путём анализа дубликатов ПЦР, образующихся в результате многократной амплификации малого числа исходных молекул. В стандартных биоинформатических алгоритмах такие дубликаты удаляются из последующего анализа как источник ложноположительных вариантов. Однако в статье Wilson et al. описан пайплайн scSNV, позволяющий анализировать подобные дубликаты с низким процентом ложноположительных SNV [119]. Суть метода заключается в «слиянии» дублирующих прочтений в длинные молекулы после выравнивания на референс и последующий анализ. При этом риды с низкой «сложностью» и артефакты из неправильно картированных прочтений, являющихся основным источником ложноположительных вариантов, удаляются.

ФИЛОГЕНЕТИКА ЗЛОКАЧЕСТВЕННЫХ НОВООБРАЗОВАНИЙ

Одним из основных признаков онкологических заболеваний является геномная нестабильность [120]. Генетические нарушения, в частности однонуклеотидные замены и аберрации числа копий ДНК, являются драйверами клональной эволюции опухолевых клеток, приводя к формированию клонов и субклонов, устойчивых к противоопухолевому лечению и обладающих высоким потенциалом к метастазированию и рецидивированию. Исследование клонального состава опухолей, особенно в динамике терапии, позволяет не только понять механизмы появления и прогрессирования злокачественных новообразований, но и разработать эффективные методы лечения, в том числе адаптированные под конкретного пациента.

Как правило, для изучения генетической гетерогенности и клональной эволюции опухоли используют bulk DNA-seq. Однако при смешивании ДНК нескольких тысяч или миллионов клеток информация о редких событиях зачастую теряется. Использование scDNA-seq в полной степени способно решить проблему поиска редких вариантов и анализа клональной структуры опухолей из-за баркодирования каждой клетки. Однако на сегодняшний день применение данного метода в значительной степени ограничено. Единственная коммерчески доступная технология scDNA-seq Tapestri позволяет судить лишь о структуре заранее выбранной пользователем или предоставленной производителем панели генов. Подходы, основанные на полногеномной амплификации, страдают от ряда проблем, в числе которых недостаточный процент охвата генома либо систематическая ошибка амплификации, которая может привести к высокой зашумлённости данных [121]. В связи с этим особо привлекательной задачей становится анализ клональной эволюции на основе данных scRNA-seq ввиду возможности совместной оценки генетической и транскрипционной гетерогенности. Однако такой подход для построения филогении опухолей осложнён рядом ограничений, в частности, невозможностью поиска генетических вариантов в нетранскрибируемых регионах, наличием аллель-специфичной экспрессии, низкими показателями глубины scRNA-seq и высоким уровнем шума в полученных данных [117, 122, 123]. В связи с этим биоинформатический анализ данных scRNA-seq для понимания клональной архитектуры опухолей является вызовом и пока реализован только в некоторых инструментах: DENDRO, Cardelino, Trisicell и SASC.

DENDRO позволяет учитывать транскрипционные всплески (transcriptional bursting), выпадение SNV и ошибки секвенирования [124].

С использованием DENDRO была оценена мутационная нагрузка, определены неоантигены для каждого опухолевого субклона и выявлена связь между транскриптомными изменениями и генетической дивергенцией опухолевых клеток [124].

Байесовский метод Cardelino позволяет интегрировать информацию о филогении, построенной на основе bulk или scDNA-seq с данными об аллельных вариантах, полученных с помощью scRNA-seq [125]. Данный подход учитывает стохастические выпадения SNV в транскриптомных данных и систематический аллельный дисбаланс ввиду моноаллельного характера экспрессии или влияния регуляторных факторов. Кроме того, Cardelino может работать только на данных scRNA-seq, предоставляя информацию о субклональной иерархии опухолевых клеток.

Большинство инструментов для вывода филогении основаны на предположении о бесконечных участках (infinite sites assumption), согласно которому каждая мутация возникает не более одного раза и не элиминируется в процессе филогенеза. Использование такой теории значительно упрощает вычислительный процесс и приемлемо для построения филогении нормальных клеток, но не злокачественных, ввиду высокой скорости накопления мутаций, а также их элиминации за счёт возникновения CNV. Авторы инструмента для анализа внутриопухолевой прогрессии SASC отходят от модели совершенной филогении и используют филогенетическую Dollo-k, допускающую элиминацию мутаций на протяжении филогенеза [126]. Использование данной модели приближает выводимое *in silico* филогенетическое древо к реальному. Кроме того, инструмент учитывает различия в частоте ложноотрицательных результатов для каждой мутации ввиду разницы в уровне экспрессии генов.

Для уточнения результатов, повышения производительной мощности существующих инструментов вывода филогении, а также сравнения древ, полученных с помощью различных инструментов и/или из различных наборов данных, был разработан инструмент Trisicell [127]. Trisicell состоит из трёх вычислительных методов: Trisicell-Boost, Trisicell-PartF и Trisicell-Cons. Trisicell-Boost увеличивает производительность и точность других инструментов за счёт многократного отбора случайных подмножеств мутаций, для каждого из которых строится филогенетическое мутационное древо. После этого Trisicell-Boost проводит сравнение различных деревьев для одного образца и осуществляет построение результирующего дерева на основе механизма консенсуса. Затем Trisicell-PartF вычисляет вероятность содержания каждого узла консенсусного древа в исследуемых клетках. Trisicell-Cons, в свою очередь, предназначен для вывода консенсусных филогенетических древ, полученных с помощью различных инструментов и/или из данных scDNA и scRNA-seq. Trisicell-Cons минимизирует количество ветвей двух или более древ, выводя более достоверную историю прогрессирования опухоли.

Стоит отметить, что перечисленные выше инструменты в большей степени применимы для обработки данных секвенирования полноразмерных транскриптов (например, Smart-seq, NuGen Solo и др.), обеспечивающего наиболее равномерное покрытие и относительно низкий уровень шума [128]. Анализ данных секвенирования одноконцевых транскриптов (10x Genomics Chromium, BD Rhapsody и др.) ввиду низкой глубины секвенирования может приводить к ошибкам в идентификации генетических вариантов и, как следствие, построению некорректных филогенетических деревьев. В этом плане секвенирование одноконцевых транскриптов целесообразно комбинировать с bulk DNA-seq или scDNA-seq и проводить совместный биоинформатический анализ, например, с помощью Cardelino или Trisicell.

ЭПИГЕНОМИКА: ДОСТУПНОСТЬ ХРОМАТИНА, ИДЕНТИФИКАЦИЯ САЙТОВ СВЯЗЫВАНИЯ ТРАНСКРИПЦИОННЫХ ФАКТОРОВ

Пространственно-временная динамика экспрессии генов обусловлена работой различных транскрипционных факторов и регулируется со стороны разного рода эпигенетических механизмов. Доступность и транскрипционная активность хроматина на регуляторных элементах генома является одним из ключевых факторов регуляции экспрессии генов. Регуляторные элементы, в частности промоторы и энхансеры, представляют собой последовательности ДНК (от нескольких сотен до тысяч пар оснований), которые состоят из уникальных сайтов связывания для транскрипционных факторов [129]. Сравнительный анализ мотивов связывания на регуляторных элементах в сочетании с информацией об экспрессии транскрипционных факторов даёт возможность пролить свет на механизмы клеточных процессов и причины возникновения различных заболеваний.

Клеточно-специфичная активность промоторов обычно определяется энхансерами. В многоклеточном организме энхансеры в первую очередь ответственны за точный контроль экспрессии генов [130]. Таким образом, изменение экспрессии одного и того же гена в разных типах клеток или в одном и том же типе клеток в разных состояниях может быть объяснено изменениями в этих цис-регуляторных элементах [131].

В последнее время общепринятым подходом для определения открытых участков ДНК стал анализ доступности хроматина для транспозазы с последующим секвенированием (ATAC-seq). Данный подход получил особую популярность из-за относительно простой экспериментальной процедуры, а также небольшого количества исходного материала – от единичных клеток до нескольких сотен [132]. Однако без информации о транскрипционных факторах, которые связывают тот или иной открытый участок хроматина, а также характерного профиля пост-трансляционных модификаций гистонов, фланкирующих доступную ДНК нуклеосом, анализ только доступности хроматина не является достаточным для определения типа регуляторного элемента. Метод иммунопреципитации хроматина (ChIP-seq) с использованием антител к транскрипционным факторам или специфичным пост-трансляционным модификациям гистонов позволил картировать расположение различных регуляторных элементов в активном и репрессированном состояниях [131]. Не так давно был разработан метод single-cell CUT&TAG для изучения полногеномного распределения различных гистоновых модификаций на уровне единичных клеток [133]. Эффективное картирование регуляторных элементов генома достигается за счёт совместного применения вышеуказанных методик. Однако это накладывает серьёзные ограничения, связанные со стоимостью, а также с потребностью в большом количестве биологического материала, так как данные методы нельзя выполнять в одних и тех же клетках одновременно. Помимо этого, данные scATAC-seq и single-cell CUT&TAG по своей природе дискретны, поскольку каждый геномный локус имеет только две копии хроматина, которые можно проанализировать внутри клетки. Данные, полученные с помощью этих методов, немногочисленны (~10^4 прочтений на клетку) и, таким образом, имеют узкий динамический диапазон. Это отличает их от данных scRNA-seq, которые являются более непрерывными, поскольку ген может иметь несколько поддающихся анализу транскриптов в клетке.

Ещё одним свойством функционально активных промоторов и энхансеров является транскрипция. С промоторов транскрибируется РНК соответствующего им гена, а с энхансеров – энхансерная РНК (эРНК). Изучение активности регуляторных элементов показало, что эРНК транскрибируется в ходе контакта энхансера с промотором-мишенью, коррелирует с гистоновыми модификациями активных энхансеров и ассоциирована с активацией транскрипции [129].

Большое количество накопленных к настоящему времени данных RNA-seq и ATAC-seq позволило выявить корреляцию между профилем доступности хроматина и транскрипцией. Современные методы анализа данных позволяют использовать модели машинного обучения с целью предсказания ландшафта хроматина на основе данных RNA-seq. Основанный на регрессионных моделях инструмент BIRD (Big data Regression for predicting DNase I hypersensitivity) использует данные RNA-seq для предсказания открытых участков хроматина. Изначально разработанный для метода DNase-seq в качестве тренировочного набора, он был успешно применён и с использованием данных ATAC-seq. BIRD применим для предсказания открытого хроматина как в данных bulk RNA-seq, так и в scRNA-seq [134]. Полученные результаты можно использовать для анализа обогащения мотивов связывания специфичных транскрипционных факторов на промоторах ДЭГ с помощью широко применяемого набора инструментов MEME, основанных на скрытых марковских моделях [135] и разработанных для пакета ChromVAR в среде R [136]. Однако данный подход неприменим для аналогичного анализа на энхансерах, которыми являются далеко не все дистальные от гена открытые участки хроматина. Один из последних инструментов, разработанных для анализа регуляторных элементов на основе данных scRNA-seq, SCAFE (Single Cell Analysis of Five-prime Ends), позволяет решить эту проблему. В своей основе он использует факт транскрипции с активных регуляторных элементов. Использование 5'-концевого scRNA-seq позволяет идентифицировать сайты старта транскрипции (TSS) как с промоторов (для оценки транскрипции гена), так и с энхансеров (для оценки транскрипции эРНК). Для поиска транскрибирующихся цис-регуляторных элементов SCAFE, как и ранее описанный метод BIRD, использует пакет референсного генома, который содержит информацию о существующих открытых участках хроматина для соответствующего вида, полученных

с помощью ATAC-seq. Идентифицировав TSS, данный инструмент позволяет получить набор активных регуляторных элементов. Помимо анализа мотивов связывания транскрипционных факторов, SCAFE позволяет оценивать изменения динамики транскрипции активных регуляторных элементов в разных состояниях и выявлять котранскрибирующиеся энхансеры и промоторы для предсказания их физического взаимодействия [137]. Такой анализ можно проводить, используя R-пакет Cicero, разработанный для предсказания взаимодействий между цис-регуляторной ДНК на основе данных доступности хроматина [138].

Используя один тип экспериментальных данных, стало возможным получать информацию о природе ДЭГ, оценивая активность регуляторных элементов всех интересующих генов. Серьёзным ограничением является необходимость наличия данных ATAC-seq как части тренировочного набора при обучении модели, что сильно затрудняет использование BIRD и SCAFE при работе с немодельными организмами. В отличие от классических экспериментальных подходов, информация о доступных участках хроматина на основе scRNA-seq формируется только за счёт генов, чьи транскрипты удалось проанализировать. Поэтому невозможно составить полноценной эпигеномной картины для каждой клетки. Однако полученных данных достаточно, чтобы оценить различия между разными клетками в исследуемом образце. Дальнейшее развитие экспериментальных методик для увеличения глубины секвенирования каждой отдельной клетки, а также усовершенствование вычислительных подходов повысит эффективность предсказания динамики доступности хроматина и активности регуляторных элементов на основе данных scRNA-seq.

РЕКОНСТРУКЦИЯ ПРОСТРАНСТВЕННОЙ ТРАНСКРИПТОМИКИ

Пространственное расположение клеток в пределах тканей и органов тесно взаимосвязано с их биологическими функциями. Хотя все клетки имеют один и тот же геном, их морфология и паттерны экспрессии генов сильно различаются в зависимости от принадлежности к типу тканей и месторасположению. Такая клеточная гетерогенность связана как с регуляторными механизмами внутри клеток, так и с влиянием внеклеточного микроокружения. Последнее наиболее ярко выражено при различных злокачественных новообразованиях, где клетки опухолевого микроокружения вносят вклад в клиническое течение и ответ на противоопухолевую терапию [139–141].

ScRNA-seq позволяет определять клеточный состав исследуемых образцов, транскрипционные особенности клеток, траектории их дифференцировки и другие показатели, рассмотренные выше. Однако пространственное расположение клеток в структуре тканей оказывается утерянным ввиду диссоциации образцов во время подготовки кДНК-библиотек и может быть предсказано лишь приблизительно. Алгоритм реконструкции пространственной организации novoSpaRc основан на теории схожести транскрипционного профиля клеток, расположенных в физической близости друг от друга [142], т.е. соседние клетки демонстрируют большее сходство в транскрипционном профиле, чем клетки, находящиеся далеко друг от друга. Однако при реконструкции пространственной организации novoSpaRc использует предопределённую геометрическую форму в качестве эталона, и, таким образом, все расчёты строятся на геометрических особенностях выбранного пространства. Кроме того, экспрессионная схожесть клеток действительно может быть следствием их близости друг к другу, но никак не предопределять её. Другой инструмент, CSOmap, предсказывает координаты каждой клетки в трёхмерном псевдо-пространстве, не ограниченном заданной формой [143]. CSOmap построен на предположении, что пространственное расположение клеток связано с их взаимодействиями по типу лиганд-рецептор. В частности, данный инструмент комбинирует профили экспрессии генов единичных клеток и общедоступную информацию о лиганд-рецепторных взаимодействиях [144, 145] для создания матрицы аффинности, которая переводится в трёхмерное пространство. Такой подход позволяет не только реконструировать пространственную организацию *de novo*, но и оценивать статистическую значимость межклеточных взаимодействий и вклад отдельных пар лиганд-рецептор в формирование таких коммуникаций. К недостаткам данного инструмента можно отнести вариабельность конечных результатов: общедоступная информация о лиганд-рецепторных взаимодействиях может различаться среди доступных источников, тем самым влияя на выводимые данные.

Иммуногистохимическое окрашивание, различные варианты гибридизации *in situ* и экспрессионное профилирование, совмещённое с

лазерной микродиссекцией, не являются идеальными методами для изучения пространственной транскриптомики. Для первых характерна большая площадь захвата, но малый охват транскриптов. Для третьего, наоборот, свойствен большой спектр анализируемых генов, но низкая исследуемая область. В последние годы разработаны различные экспериментальные методы, позволяющие проводить анализ большого количества транскриптов на больших участках ткани [146]. Коммерчески доступными являются Visium («10x Genomics»), GeoMx «NanoString Technologies», Molecular Cartography от «Resolve Biosciences», Stereo-seq от «BGI STOmics» и другие методы пространственной транскриптомики. Однако в данный момент Visium и GeoMx не способны предоставить разрешение на уровне единичной клетки. Размер ячейки с пространственным штрих-кодом на слайде Visium составляет 55 мкм, что создаёт вероятность попадания в одну ячейку нескольких клеток. Технически GeoMx способен проводить захват на уровне единичной клетки, однако высокое отношение шум/сигнал ограничивает эту возможность. Кроме того, применение данных методов ограничивается их высокой стоимостью и, как следствие, относительной недоступностью. В 2021 г. компания «10x Genomics» анонсировала Visium HD – технологию пространственной транскриптомики с разрешением, в 400 раз превосходящим таковое у классического Visium и, соответственно, возможностью анализа на уровне единичной клетки.

ВЫВОДЫ И ДАЛЬНЕЙШИЕ ПЕРСПЕКТИВЫ

Прогресс в области мультиплексирования кДНК-библиотек единичных клеток и в разработке вычислительных методов биоинформатического анализа позволил в значительной степени расширить спектр информации, которую возможно извлечь, используя данные scRNA-seq. Помимо классических для scRNA-seq приложений, таких как определение клеточного цикла, идентификация клеточных кластеров, анализ дифференциальной экспрессии и сигнальных путей и типирование клеток, стало возможным исследовать предопределяющие фенотип генетические и эпигенетические характеристики клеток (CNV/SNV и состояние хроматина), предсказывать направление их дифференцировки, получать информацию о межклеточных взаимодействиях и филогении, в том числе в контексте пространственной организации тканей и органов.

Тем не менее получение информации о CNV/SNV, филогении и доступности хроматина в значительной степени зависит от качества и глубины секвенирования. Наиболее корректным в этом плане решением может быть коммерчески доступный, но трудоёмкий Smart-seq, основанный на захвате полноразмерных транскриптов, или интегративный анализ данных scRNA-seq, bulk DNA-seq и scDNA-seq. Другим потенциальным решением может стать усовершенствование протоколов подготовки библиотек полноразмерных транскриптов. Так, в 2022 г. был представлен протокол FLASH-seq, превосходящий по скорости и чувствительности любой из существующих протоколов scRNA-seq [147]. В основе разработки лежит протокол Smart-seq 2, однако для уменьшения временных затрат и повышения разрешающей способности авторы внесли в него несколько ключевых модификаций: объединили обратную транскрипцию и предварительную амплификацию кДНК; заменили обратную транскриптазу Superscript II на более эффективную Superscript IV; увеличили количество дезоксицитидинтрифосфата для индукции *C*-хвостовой активности SSRTIV и усиления реакции переключения матрицы и разместили рибогуанозин в позиции 3' олигонуклеотида, необходимого для инвазии цепи ДНК и смены матрицы. Кроме того, многообещающей перспективой могут быть платформы для мультиомиксного анализа единичных клеток. Подобные протоколы уже разработаны и предоставляют возможность комбинирования оценки доступности хроматина и транскриптома единичных клеток (sci-CAR [148] и SNARE-seq [149]) и совместного проведения полногеномного секвенирования и профилирования экспрессии генов (DNTR-seq) [150]. Однако данные методы чрезвычайно трудоёмки, дороги в исполнении и характеризуются высоким процентом ложноположительных результатов.

Стоит также уделить внимание нюансам в определении межклеточных взаимодействий на основе анализа пар лиганд-рецептор. Инструменты, позволяющие выводить данную модальность из данных scRNA-seq, опираются на информацию о лиганд-рецепторных взаимодействиях из различных источников. Соответственно, при использовании различных эталонных данных конечный результат будет отличаться. Информация о лиганд-рецепторных взаимодействиях также используется и при *de novo* реконструкции пространственной организации в инструменте CSOmap. Из-за потенциальной вариабельности конечных

результатов данные методы могут дать лишь ряд гипотез, которые необходимо валидировать в других экспериментах, например, с помощью методов пространственной транскриптомики.

В целом, мы ожидаем, что бурное развитие инструментов биоинформатического анализа совместно с усовершенствованием протоколов по подготовке библиотек РНК/ДНК единичных клеток и разработкой платформ для мультиомиксного анализа в значительной степени увеличит качество биомедицинских исследований. Технический прогресс в области технологий на уровне единичных клеток поможет расшифровать клеточную гетерогенность, обусловленную совокупностью конститутивных и функциональных особенностей, что, в свою очередь, позволит расширить понимание биологических процессов в норме и патологии и сформировать принципиально новые подходы к персонализированной терапии заболеваний.

Вклад авторов. А.А. Хозяинова, Е.В. Денисов – концепция обзора; А.А. Хозяинова, А.А. Валяева, М.С. Арбатский, С.В. Исаев, П.С. Ямщиков, Е.В. Волчков, М.С. Сабиров, В.Р. Зайнуллина, В.И. Чечехин, Р.С. Воробьёв, М.Е. Меняйло, П.А. Тюрин-Кузьмин, Е.В. Денисов – сбор информации, анализ публикаций, написание и редактирование разделов обзора.

Финансирование. Работа выполнена при финансовой поддержке Российского научного фонда (грант № 19-75-30016).

Благодарности. Мы благодарим А.А. Щеголеву за графическое сопровождение статьи.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Соблюдение этических норм. Настоящая статья не содержит экспериментов и каких-либо исследований с участием людей или животных в качестве объектов.

СПИСОК ЛИТЕРАТУРЫ

1. Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., Lao, K., and Surani, M. A. (2009) mRNA-Seq whole-transcriptome analysis of a single cell, *Nat. Methods*, **6**, 377-382, doi: 10.1038/nmeth.1315.
2. Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J. B., Lönnerberg, P., and Linnarsson, S. (2011) Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq, *Genome Res.*, **21**, 1160-1167, doi: 10.1101/gr.110882.110.
3. Ke, M., Elshenawy, B., Sheldon, H., Arora, A., and Buffa, F. M. (2022) Single cell RNA-sequencing: A powerful yet still challenging technology to study cellular heterogeneity, *BioEssays*, **44**, 2200084, doi: 10.1002/bies.202200084.
4. Luo, G., Gao, Q., Zhang, S., and Yan, B. (2020) Probing infectious disease by single-cell RNA sequencing: progresses and perspectives, *Comput. Struct. Biotechnol. J.*, **18**, 2962-2971, doi: 10.1016/j.csbj.2020.10.016.
5. Yifan, C., Fan, Y., and Jun, P. (2020) Visualization of cardiovascular development, physiology and disease at the single-cell level: opportunities and future challenges, *J. Mol. Cell. Cardiol.*, **142**, 80-92, doi: 10.1016/j.yjmcc.2020.03.005.
6. Pan, Y., Cao, W., Mu, Y., and Zhu, Q. (2022) Microfluidics facilitates the development of single-cell RNA sequencing, *Biosensors*, **12**, 450, doi: 10.3390/bios12070450.
7. Wehrens, M., de Leeuw, A. E., Wright-Clark, M., Eding, J. E., Boogerd, C. J., Molenaar, B., van der Kraak, P. H., Kuster, D. W., van der Velden, J., and Michels, M. (2022) Single-cell transcriptomics provides insights into hypertrophic cardiomyopathy, *Cell Rep.*, **39**, 110809, doi: 10.1016/j.celrep.2022.110809.
8. Olah, M., Menon, V., Habib, N., Taga, M. F., Ma, Y., Yung, C. J., Cimpean, M., Khairallah, A., Coronas-Samano, G., and Sankowski, R. (2020) Single cell RNA sequencing of human microglia uncovers a subset associated with Alzheimer's disease, *Nat. Commun.*, **11**, 1-18, doi: 10.1038/s41467-020-19737-2.
9. Kamath, T., Abdulraouf, A., Burris, S., Langlieb, J., Gazestani, V., Nadaf, N. M., Balderrama, K., Vanderburg, C., and Macosko, E. Z. (2022) Single-cell genomic profiling of human dopamine neurons identifies a population that selectively degenerates in Parkinson's disease, *Nat. Neurosci.*, **25**, 588-595, doi: 10.1038/s41593-022-01061-1.
10. Zhou, S., Huang, Y.-E., Liu, H., Zhou, X., Yuan, M., Hou, F., Wang, L., and Jiang, W. (2021) Single-cell RNA-seq dissects the intratumoral heterogeneity of triple-negative breast cancer based on gene regulatory networks, *Mol. Ther. Nucleic Acids*, **23**, 682-690, doi: 10.1016/j.omtn.2020.12.018.
11. Zhang, T.-Q., Chen, Y., and Wang, J.-W. (2021) A single-cell analysis of the Arabidopsis vegetative shoot apex, *Dev. Cell*, **56**, 1056-1074.e1058, doi: 10.1016/j.devcel.2021.02.021.
12. Fu, Y., Huang, X., Zhang, P., van de Leemput, J., and Han, Z. (2020) Single-cell RNA sequencing identifies novel cell types in Drosophila blood, *J. Genet. Genomics*, **47**, 175-186, doi: 10.1016/j.jgg.2020.02.004.

13. Jiang, M., Xiao, Y., Weigao, E., Ma, L., Wang, J., Chen, H., Gao, C., Liao, Y., Guo, Q., and Peng, J. (2021) Characterization of the zebrafish cell landscape at single-cell resolution, *Front. Cell Dev. Biol.*, **9**, 743421, doi: 10.3389/fcell.2021.743421.
14. Ho, D. W.-H., Tsui, Y.-M., Chan, L.-K., Sze, K. M.-F., Zhang, X., Cheu, J. W.-S., Chiu, Y.-T., Lee, J. M.-F., Chan, A. C.-Y., and Cheung, E. T.-Y. (2021) Single-cell RNA sequencing shows the immunosuppressive landscape and tumor heterogeneity of HBV-associated hepatocellular carcinoma, *Nat. Commun.*, **12**, 1-14, doi: 10.1038/s41467-021-24010-1.
15. Zhao, J., Zhang, S., Liu, Y., He, X., Qu, M., Xu, G., Wang, H., Huang, M., Pan, J., and Liu, Z. (2020) Single-cell RNA sequencing reveals the heterogeneity of liver-resident immune cells in human, *Cell Discov.*, **6**, 1-19, doi: 10.1038/s41421-020-0157-z.
16. Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., and Martersteck, E. M. (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets, *Cell*, **161**, 1202-1214, doi: 10.1016/j.cell.2015.05.002.
17. Kowalczyk, M. S., Tirosh, I., Heckl, D., Rao, T. N., Dixit, A., Haas, B. J., Schneider, R. K., Wagers, A. J., Ebert, B. L., and Regev, A. (2015) Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells, *Genome Res.*, **25**, 1860-1872, doi: 10.1101/gr.192237.115.
18. Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species, *Nat. Biotechnol.*, **36**, 411-420, doi: 10.1038/nbt.4096.
19. Wolf, F. A., Angerer, P., and Theis, F. J. (2018) SCANPY: large-scale single-cell gene expression data analysis, *Genome Biol.*, **19**, 1-5, doi: 10.1186/s13059-017-1382-0.
20. Scialdone, A., Tanaka, Y., Jawaid, W., Moignard, V., Wilson, N. K., Macaulay, I. C., Marioni, J. C., and Göttgens, B. (2016) Resolving early mesoderm diversification through single-cell expression profiling, *Nature*, **535**, 289-293, doi: 10.1038/nature18633.
21. Tirosh, I., Izar, B., Prakadan, S. M., Wadsworth, M. H., Treacy, D., Trombetta, J. J., Rotem, A., Rodman, C., Lian, C., and Murphy, G. (2016) Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq, *Science*, **352**, 189-196, doi: 10.1126/science.aad0501.
22. Hsiao, C. J., Tung, P., Blischak, J. D., Burnett, J. E., Barr, K. A., Dey, K. K., Stephens, M., and Gilad, Y. (2020) Characterizing and inferring quantitative cell cycle phase in single-cell RNA-seq data analysis, *Genome Res.*, **30**, 611-621, doi: 10.1101/gr.247759.118.
23. Liu, Z., Lou, H., Xie, K., Wang, H., Chen, N., Aparicio, O. M., Zhang, M. Q., Jiang, R., and Chen, T. (2017) Reconstructing cell cycle pseudo time-series via single-cell transcriptome data, *Nat. Commun.*, **8**, 1-9, doi: 10.1038/s41467-017-00039-z.
24. Liang, S., Wang, F., Han, J., and Chen, K. (2020) Latent periodic process inference from single-cell RNA-seq data, *Nat. Commun.*, **11**, 1-8, doi: 10.1038/s41467-020-15295-9.
25. Anafi, R. C., Francey, L. J., Hogenesch, J. B., and Kim, J. (2017) CYCLOPS reveals human transcriptional rhythms in health and disease, *Proc. Natl. Acad. Sci. USA*, **114**, 5312-5317, doi: 10.1073/pnas.1619320114.
26. Liu, J., Yang, M., Zhao, W., and Zhou, X. (2022) CCPE: cell cycle pseudotime estimation for single cell RNA-seq data, *Nucleic Acids Res.*, **50**, 704-716, doi: 10.1093/nar/gkab1236.
27. Melms, J. C., Biermann, J., Huang, H., Wang, Y., Nair, A., Tagore, S., Katsyv, I., Rendeiro, A. F., Amin, A. D., Schapiro, D., et al. (2021) A molecular single-cell lung atlas of lethal COVID-19, *Nature*, **595**, 114-119, doi: 10.1038/s41586-021-03569-1.
28. Delorey, T. M., Ziegler, C. G., Heimberg, G., Normand, R., Yang, Y., Segerstolpe, Å., Abbondanza, D., Fleming, S. J., Subramanian, A., Montoro, D. T., et al. (2021) COVID-19 tissue atlases reveal SARS-CoV-2 pathology and cellular targets, *Nature*, **595**, 107-113, doi: 10.1038/s41586-021-03570-8.
29. Mathys, H., Davila-Velderrain, J., Peng, Z., Gao, F., Mohammadi, S., Young, J. Z., Menon, M., He, L., Abdurrob, F., Jiang, X., et al. (2019) Single-cell transcriptomic analysis of Alzheimer's disease, *Nature*, **570**, 332-337, doi: 10.1038/s41586-019-1195-2.
30. Velmeshev, D., Schirmer, L., Jung, D., Haeussler, M., Perez, Y., Mayer, S., Bhaduri, A., Goyal, N., Rowitch, D. H., and Kriegstein, A. R. (2019) Single-cell genomics identifies cell type-specific molecular changes in autism, *Science*, **364**, 685-689, doi: 10.1126/science.aav8130.
31. Kathiriya, J. J., Brumwell, A. N., Jackson, J. R., Tang, X., and Chapman, H. A. (2020) Distinct airway epithelial stem cells hide among club cells but mobilize to promote alveolar regeneration, *Cell Stem Cell*, **26**, 346-358.e344, doi: 10.1016/j.stem.2019.12.014.
32. Steuerman, Y., Cohen, M., Peshes-Yaloz, N., Valadarsky, L., Cohn, O., David, E., Frishberg, A., Mayo, L., Bacharach, E., Amit, I., and Gat-Viks, I. (2018) Dissection of influenza infection in vivo by single-cell RNA sequencing, *Cell Systems*, **6**, 679-691.e674, doi: 10.1016/j.cels.2018.05.008.
33. Kiselev, V. Y., Andrews, T. S., and Hemberg, M. (2019) Challenges in unsupervised clustering of single-cell RNA-seq data, *Nat. Rev. Genet.*, **20**, 273-282, doi: 10.1038/s41576-018-0088-9.
34. Brennecke, P., Anders, S., Kim, J. K., Kołodziejczyk, A. A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S. A., Marioni, J. C., and Heisler, M. G. (2013) Accounting for technical noise

in single-cell RNA-seq experiments, *Nat. Methods*, **10**, 1093-1095, doi: 10.1038/nmeth.2645.
35. Andrews, T. S., and Hemberg, M. (2019) M3Drop: dropout-based feature selection for scRNASeq, *Bioinformatics*, **35**, 2865-2867, doi: 10.1093/bioinformatics/bty1044.
36. Yau, C. (2016) pcaReduce: hierarchical clustering of single cell transcriptional profiles, *BMC Bioinformatics*, **17**, 1-11, doi: 10.1186/s12859-016-0984-y.
37. Lin, P., Troup, M., and Ho, J. W. (2017) CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data, *Genome Biol.*, **18**, 1-11, doi: 10.1186/s13059-017-1188-0.
38. Kiselev, V. Y., Kirschner, K., Schaub, M. T., Andrews, T., Yiu, A., Chandra, T., Natarajan, K. N., Reik, W., Barahona, M., Green, A. R., and Hemberg, M. (2017) SC3: consensus clustering of single-cell RNA-seq data, *Nat. Methods*, **14**, 483-486, doi: 10.1038/nmeth.4236.
39. Levine, J. H., Simonds, E. F., Bendall, S. C., Davis, K. L., El-ad, D. A., Tadmor, M. D., Litvin, O., Fienberg, H. G., Jager, A., Zunder, E. R., et al. (2015) Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis, *Cell*, **162**, 184-197, doi: 10.1016/j.cell.2015.05.047.
40. Montoro, D. T., Haber, A. L., Biton, M., Vinarsky, V., Lin, B., Birket, S. E., Yuan, F., Chen, S., Leung, H. M., and Villoria, J. (2018) A revised airway epithelial hierarchy includes CFTR-expressing ionocytes, *Nature*, **560**, 319-324, doi: 10.1038/s41586-018-0393-7.
41. Plasschaert, L. W., Žilionis, R., Choo-Wing, R., Savova, V., Knehr, J., Roma, G., Klein, A. M., and Jaffe, A. B. (2018) A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte, *Nature*, **560**, 377-381, doi: 10.1038/s41586-018-0394-6.
42. Love, M. I., Huber, W., and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2, *Genome Biol.*, **15**, 1-21, doi: 10.1186/s13059-014-0550-8.
43. Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data, *Bioinformatics*, **26**, 139-140, doi: 10.1093/bioinformatics/btp616.
44. Van den Berge, K., Perraudeau, F., Soneson, C., Love, M. I., Risso, D., Vert, J.-P., Robinson, M. D., Dudoit, S., and Clement, L. (2018) Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications, *Genome Biol.*, **19**, 1-17, doi: 10.1186/s13059-018-1406-4.
45. Tang, W., Bertaux, F., Thomas, P., Stefanelli, C., Saint, M., Marguerat, S., and Shahrezaei, V. (2020) bayNorm: Bayesian gene expression recovery, imputation and normalization for single-cell RNA-sequencing data, *Bioinformatics*, **36**, 1174-1181, doi: 10.1093/bioinformatics/btz726.
46. Soneson, C., and Robinson, M. D. (2018) Bias, robustness and scalability in single-cell differential expression analysis, *Nat. Methods*, **15**, 255-261, doi: 10.1038/nmeth.4612.
47. Kharchenko, P. V., Silberstein, L., and Scadden, D. T. (2014) Bayesian approach to single-cell differential expression analysis, *Nat. Methods*, **11**, 740-742, doi: 10.1038/nmeth.2967.
48. Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., Slichter, C. K., Miller, H. W., McElrath, M. J., Prlic, M., et al. (2015) MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data, *Genome Biol.*, **16**, 1-13, doi: 10.1186/s13059-015-0844-5.
49. Zhang, J. M., Kamath, G. M., and David, N. T. (2019) Valid post-clustering differential analysis for single-cell RNA-Seq, *Cell Systems*, **9**, 383-392.e386, doi: 10.1016/j.cels.2019.07.012.
50. Zimmerman, K. D., Espeland, M. A., and Langefeld, C. D. (2021) A practical solution to pseudoreplication bias in single-cell studies, *Nat. Commun.*, **12**, 1-9, doi: 10.1038/s41467-021-21038-1.
51. Denninger, J. K., Walker, L. A., Chen, X., Turkoglu, A., Pan, A., Tapp, Z., Senthilvelan, S., Rindani, R., Kokiko-Cochran, O. N., and Bundschuh, R. (2022) Robust transcriptional profiling and identification of differentially expressed genes with low input RNA sequencing of adult hippocampal neural stem and progenitor populations, *Front. Mol. Neurosci.*, **15**, 810722, doi: 10.3389/fnmol.2022.810722.
52. Hücker, S. M., Fehlmann, T., Werno, C., Weidele, K., Lüke, F., Schlenska-Lange, A., Klein, C. A., Keller, A., and Kirsch, S. (2021) Single-cell microRNA sequencing method comparison and application to cell lines and circulating lung tumor cells, *Nat. Commun.*, **12**, 1-13, doi: 10.1038/s41467-021-24611-w.
53. Valyaeva, A. A., Zharikova, A. A., Kasianov, A. S., Vassetzky, Y. S., and Sheval, E. V. (2020) Expression of SARS-CoV-2 entry factors in lung epithelial stem cells and its potential implications for COVID-19, *Sci. Rep.*, **10**, 1-8, doi: 10.1038/s41598-020-74598-5.
54. Liao, M., Liu, Y., Yuan, J., Wen, Y., Xu, G., Zhao, J., Cheng, L., Li, J., Wang, X., Wang, F., et al. (2020) Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19, *Nat. Med.*, **26**, 842-844, doi: 10.1038/s41591-020-0901-9.
55. Yakushov, S., Menyailo, M., Denisov, E., Karlina, I., Zainullina, V., Kirgizov, K., Romantsova, O., Timashev, P., and Ulasov, I. (2022) Identification of factors driving doxorubicin-resistant ewing tumor cells to survival, *Cancers*, **14**, 5498, doi: 10.3390/cancers14225498.
56. Tyurin-Kuzmin, P. A., Karagyaur, M. N., Kulebyakin, K. Y., Dyikanov, D. T., Chechekhin, V. I., Ivanova, A. M., Skryabina, M. N., Arbatskiy, M. S., Sysoeva, V. Y., Kalinina, N. I., and Tkachuk, V. A.

(2020) Functional heterogeneity of protein kinase a activation in multipotent stromal cells, *Int. J. Mol. Sci.*, **21**, 4442, doi: 10.3390/ijms21124442.

57. Bassez, A., Vos, H., Van Dyck, L., Floris, G., Arijs, I., Desmedt, C., Boeckx, B., Vanden Bempt, M., Nevelsteen, I., Lambein, K., et al. (2021) A single-cell map of intratumoral changes during anti-PD1 treatment of patients with breast cancer, *Nat. Med.*, **27**, 820-832, doi: 10.1038/s41591-021-01323-8.
58. Bi, K., He, M. X., Bakouny, Z., Kanodia, A., Napolitano, S., Wu, J., Grimaldi, G., Braun, D. A., Cuoco, M. S., Mayorga, A., et al. (2021) Tumor and immune reprogramming during immunotherapy in advanced renal cell carcinoma, *Cancer Cell*, **39**, 649-661.e645, doi: 10.1016/j.ccell.2021.02.015.
59. Hoernes, T. P., Hüttenhofer, A., and Erlacher, M. D. (2016) mRNA modifications: Dynamic regulators of gene expression? *RNA Biol.*, **13**, 760-765, doi: 10.1080/15476286.2016.1203504.
60. Maier, T., Güell, M., and Serrano, L. (2009) Correlation of mRNA and protein in complex biological samples, *FEBS Lett.*, **583**, 3966-3973, doi: 10.1016/j.febslet.2009.10.036.
61. Aran, D., Looney, A. P., Liu, L., Wu, E., Fong, V., Hsu, A., Chak, S., Naikawadi, R. P., Wolters, P. J., Abate, A. R., et al. (2019) Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage, *Nat. Immunol.*, **20**, 163-172, doi: 10.1038/s41590-018-0276-y.
62. Ianevski, A., Giri, A. K., and Aittokallio, T. (2022) Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data, *Nat. Commun.*, **13**, 1-10, doi: 10.1038/s41467-022-28803-w.
63. Shao, X., Liao, J., Lu, X., Xue, R., Ai, N., and Fan, X. (2020) scCATCH: automatic annotation on cell types of clusters from single-cell RNA sequencing data, *iScience*, **23**, 100882, doi: 10.1016/j.isci.2020.100882.
64. Guo, H., and Li, J. (2021) scSorter: assigning cells to known cell types according to marker genes, *Genome Biol.*, **22**, 1-18, doi: 10.1186/s13059-021-02281-7.
65. Zhang, Z., Luo, D., Zhong, X., Choi, J. H., Ma, Y., Wang, S., Mahrt, E., Guo, W., Stawiski, E. W., Modrusan, Z., Seshagiri, S., Kapur, P., Hon, G. C., Brugarolas, J., and Wang, T. (2019) SCINA: a semi-supervised subtyping algorithm of single cells and bulk samples, *Genes*, **10**, 531, doi: 10.3390/genes10070531.
66. Xu, C., Lopez, R., Mehlman, E., Regier, J., Jordan, M. I., and Yosef, N. (2021) Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models, *Mol. Syst. Biol.*, **17**, e9620, doi: 10.15252/msb.20209620.
67. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck III, W. M., Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zager, M., et al. (2021) Integrated analysis of multimodal single-cell data, *Cell*, **184**, 3573-3587.e3529, doi: 10.1016/j.cell.2021.04.048.
68. Pasquini, G., Arias, J. E. R., Schäfer, P., and Busskamp, V. (2021) Automated methods for cell type annotation on scRNA-seq data, *Computat. Struct. Biotechnol. J.*, **19**, 961-969, doi: 10.1016/j.csbj.2021.01.015.
69. Khrameeva, E., Kurochkin, I., Han, D., Guijarro, P., Kanton, S., Santel, M., Qian, Z., Rong, S., Mazin, P., Sabirov, M., et al. (2020) Single-cell-resolution transcriptome map of human, chimpanzee, bonobo, and macaque brains, *Genome Res.*, **30**, 776-789, doi: 10.1101/gr.256958.119.
70. Han, G., Deng, Q., Marques-Piubelli, M. L., Dai, E., Dang, M., Ma, M. C. J., Li, X., Yang, H., Henderson, J., Kudryashova, O., et al. (2022) Follicular lymphoma microenvironment characteristics associated with tumor cell mutations and MHC class II expression, *Blood Cancer Discov.*, **3**, 428-443, doi: 10.1158/2643-3230.BCD-21-0075.
71. La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M. E., Lönnerberg, P., Furlan, A., et al. (2018) RNA velocity of single cells, *Nature*, **560**, 494-498, doi: 10.1038/s41586-018-0414-6.
72. Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S., and Rinn, J. L. (2014) The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells, *Nat. Biotechnol.*, **32**, 381-386, doi: 10.1038/nbt.2859.
73. Ji, Z., and Ji, H. (2016) TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis, *Nucleic Acids Res.*, **44**, e117-e117, doi: 10.1093/nar/gkw430.
74. Street, K., Risso, D., Fletcher, R. B., Das, D., Ngai, J., Yosef, N., Purdom, E., and Dudoit, S. (2018) Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics, *BMC Genomics*, **19**, 1-16, doi: 10.1186/s12864-018-4772-0.
75. Saelens, W., Cannoodt, R., Todorov, H., and Saeys, Y. (2019) A comparison of single-cell trajectory inference methods, *Nat. Biotechnol.*, **37**, 547-554, doi: 10.1038/s41587-019-0071-9.
76. Nam, D. K., Lee, S., Zhou, G., Cao, X., Wang, C., Clark, T., Chen, J., Rowley, J. D., and Wang, S. M. (2002) Oligo (dT) primer generates a high frequency of truncated cDNAs through internal poly (A) priming during reverse transcription, *Proc. Natl. Acad. Sci. USA*, **99**, 6152-6156, doi: 10.1073/pnas.092140899.
77. Perrimon, N., Pitsouli, C., and Shilo, B. Z. (2012) Signaling mechanisms controlling cell fate and embryonic patterning, *Cold Spring Harb. Perspect. Biol.*, **4**, a005975, doi: 10.1101/cshperspect.a005975.
78. Gaud, G., Lesourne, R., and Love, P. E. (2018) Regulatory mechanisms in T cell receptor signalling, *Nat. Rev. Immunol.*, **18**, 485-497, doi: 10.1038/s41577-018-0020-8.

79. Yeung, T. L., Sheng, J., Leung, C. S., Li, F., Kim, J., Ho, S. Y., Matzuk, M. M., Lu, K. H., Wong, S. T. C., and Mok, S. C. (2019) Systematic identification of druggable epithelial-stromal crosstalk signaling networks in ovarian cancer, *J. Natl. Cancer Institute*, **111**, 272-282, doi: 10.1093/jnci/djy097.
80. Chua, R. L., Lukassen, S., Trump, S., Hennig, B. P., Wendisch, D., Pott, F., Debnath, O., Thürmann, L., Kurth, F., Völker, M. T., Kazmierski, J., Timmermann, B., Twardziok, S., Schneider, S., Machleidt, F., Müller-Redetzky, H., Maier, M., Krannich, A., Schmidt, S., Balzer, F., et al. (2020) COVID-19 severity correlates with airway epithelium-immune cell interactions identified by single-cell analysis, *Nat. Biotechnol.*, **38**, 970-979, doi: 10.1038/s41587-020-0602-4.
81. Yuan, D., Tao, Y., Chen, G., and Shi, T. (2019) Systematic expression analysis of ligand-receptor pairs reveals important cell-to-cell interactions inside glioma, *Cell Commun. Signal.*, **17**, 48, doi: 10.1186/s12964-019-0363-1.
82. Rao, V. S., Srinivas, K., Sujini, G. N., and Kumar, G. N. (2014) Protein-protein interaction detection: methods and analysis, *Int. J. Proteomics*, **2014**, 147648, doi: 10.1155/2014/147648.
83. Wang, Y., Wang, R., Zhang, S., Song, S., Jiang, C., Han, G., Wang, M., Ajani, J., Futreal, A., and Wang, L. (2019) iTALK: an R package to characterize and illustrate intercellular communication, *bioRxiv*, 507871, doi: 10.1101/507871.
84. Efremova, M., Vento-Tormo, M., Teichmann, S. A., and Vento-Tormo, R. (2020) CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes, *Nat. Protocols*, **15**, 1484-1506, doi: 10.1038/s41596-020-0292-x.
85. Cabello-Aguilar, S., Alame, M., Kon-Sun-Tack, F., Fau, C., Lacroix, M., and Colinge, J. (2020) SingleCellSignalR: inference of intercellular networks from single-cell transcriptomics, *Nucleic Acids Res.*, **48**, e55, doi: 10.1093/nar/gkaa183.
86. Zhang, Y., Liu, T., Hu, X., Wang, M., Wang, J., Zou, B., Tan, P., Cui, T., Dou, Y., Ning, L., Huang, Y., Rao, S., Wang, D., and Zhao, X. (2021) CellCall: integrating paired ligand-receptor and transcription factor activities for cell-cell communication, *Nucleic Acids Res.*, **49**, 8520-8534, doi: 10.1093/nar/gkab638.
87. Tsuyuzaki, K., Ishii, M., and Nikaido, I. (2019) Uncovering hypergraphs of cell-cell interaction from single cell RNA-sequencing data, *bioRxiv*, 566182, doi: 10.1101/566182.
88. Armingol, E., Officer, A., Harismendy, O., and Lewis, N. E. (2021) Deciphering cell-cell interactions and communication from gene expression, *Nat. Rev. Genet.*, **22**, 71-88, doi: 10.1038/s41576-020-00292-x.
89. Fischer, D. S., Schaar, A. C., and Theis, F. J. (2021) Learning cell communication from spatial graphs of cells, *bioRxiv*, doi: 10.1101/2021.07.11.451750.
90. Van Dam, S., Võsa, U., van der Graaf, A., Franke, L., and de Magalhães, J. P. (2018) Gene co-expression analysis for functional classification and gene-disease predictions, *Brief. Bioinform.*, **19**, 575-592, doi: 10.1093/bib/bbw139.
91. Rambow, F., Rogiers, A., Marin-Bejar, O., Aibar, S., Femel, J., Dewaele, M., Karras, P., Brown, D., Chang, Y. H., Debiec-Rychter, M., Adriaens, C., Radaelli, E., Wolter, P., Bechter, O., Dummer, R., Levesque, M., Piris, A., Frederick, D. T., Boland, G., Flaherty, K. T., et al. (2018) Toward minimal residual disease-directed therapy in melanoma, *Cell*, **174**, 843-855.e819, doi: 10.1016/j.cell.2018.06.025.
92. Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010) Inferring regulatory networks from expression data using tree-based methods, *PLoS One*, **5**, e12776, doi: 10.1371/journal.pone.0012776.
93. Moerman, T., Aibar Santos, S., Bravo González-Blas, C., Simm, J., Moreau, Y., Aerts, J., and Aerts, S. (2019) GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks, *Bioinformatics*, **35**, 2159-2161, doi: 10.1093/bioinformatics/bty916.
94. Aibar, S., González-Blas, C. B., Moerman, T., Huynh-Thu, V. A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.-C., Geurts, P., and Aerts, J. (2017) SCENIC: single-cell regulatory network inference and clustering, *Nat. Methods*, **14**, 1083-1086, doi: 10.1038/nmeth.4463.
95. Langfelder, P., and Horvath, S. (2008) WGCNA: an R package for weighted correlation network analysis, *BMC Bioinformatics*, **9**, 1-13, doi: 10.1186/1471-2105-9-559.
96. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, *Proc. Natl. Acad. Sci. USA*, **102**, 15545-15550, doi: 10.1073/pnas.0506580102.
97. Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N. T., Morris, J. H., Bork, P., et al. (2019) STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets, *Nucleic Acids Res.*, **47**, D607-D613, doi: 10.1093/nar/gky1131.
98. Kim, C. Y., Baek, S., Cha, J., Yang, S., Kim, E., Marcotte, E. M., Hart, T., and Lee, I. (2022) HumanNet v3: an improved database of human gene networks for disease research, *Nucleic acids Res.*, **50**, D632-D639, doi: 10.1093/nar/gkab1048.

99. Xue, Z., Huang, K., Cai, C., Cai, L., Jiang, C.-Y., Feng, Y., Liu, Z., Zeng, Q., Cheng, L., Sun, Y. E., et al. (2013) Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing, *Nature*, **500**, 593-597, doi: 10.1038/nature12364.
100. Wu, H., Chen, S., Yu, J., Li, Y., Zhang, X.-Y., Yang, L., Zhang, H., Jiang, M., Brunicardi, F. C., Wang, C., and Wu, S. (2018) Single-cell transcriptome analyses reveal molecular signals to intrinsic and acquired paclitaxel resistance in esophageal squamous cancer cells, *Cancer Lett.*, **420**, 156-167, doi: 10.1016/j.canlet.2018.01.059.
101. Lu, J., Chen, Y., Zhang, X., Guo, J., Xu, K., and Li, L. (2022) A novel prognostic model based on single-cell RNA sequencing data for hepatocellular carcinoma, *Cancer Cell Int.*, **22**, 1-12, doi: 10.1186/s12935-022-02469-2.
102. Lee, W.-P., and Tzou, W.-S. (2009) Computational methods for discovering gene networks from expression data, *Brief. Bioinform.*, **10**, 408-423, doi: 10.1093/bib/bbp028.
103. Matsumoto, H., Kiryu, H., Furusawa, C., Ko, M. S., Ko, S. B., Gouda, N., Hayashi, T., and Nikaido, I. (2017) SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation, *Bioinformatics*, **33**, 2314-2321, doi: 10.1093/bioinformatics/btx194.
104. Pös, O., Radvanszky, J., Buglyó, G., Pös, Z., Rusnakova, D., Nagy, B., and Szemes, T. (2021) DNA copy number variation: Main characteristics, evolutionary significance, and pathological aspects, *Biomed. J.*, **44**, 548-559, doi: 10.1016/j.bj.2021.02.003.
105. Lye, Z. N., and Purugganan, M. D. (2019) Copy number variation in domestication, *Trends Plant Sci.*, **24**, 352-365, doi: 10.1016/j.tplants.2019.01.003.
106. Zhao, Y., Carter, R., Natarajan, S., Varn, F. S., Compton, D. A., Gawad, C., Cheng, C., and Godek, K. M. (2019) Single-cell RNA sequencing reveals the impact of chromosomal instability on glioblastoma cancer stem cells, *BMC Med. Genom.*, **12**, 1-16, doi: 10.1186/s12920-019-0532-5.
107. Zhou, B., Ho, S. S., Zhang, X., Pattni, R., Haraksingh, R. R., and Urban, A. E. (2018) Whole-genome sequencing analysis of CNV using low-coverage and paired-end strategies is efficient and outperforms array-based CNV analysis, *J. Med. Genet.*, **55**, 735-743, doi: 10.1136/jmedgenet-2018-105272.
108. Shao, X., Lv, N., Liao, J., Long, J., Xue, R., Ai, N., Xu, D., and Fan, X. (2019) Copy number variation is highly correlated with differential gene expression: a pan-cancer study, *BMC Med. Genet.*, **20**, 1-14, doi: 10.1186/s12881-019-0909-5.
109. Fan, J., Lee, H.-O., Lee, S., Ryu, D.-E., Lee, S., Xue, C., Kim, S. J., Kim, K., Barkas, N., Park, P. J., et al. (2018) Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data, *Gen. Res.*, **28**, 1217-1227, doi: 10.1101/gr.228080.117.
110. Serin Harmanci, A., Harmanci, A. O., and Zhou, X. (2020) CaSpER identifies and visualizes CNV events by integrative analysis of single-cell or bulk RNA-equencing data, *Nat. Commun.*, **11**, 1-16, doi: 10.1038/s41467-019-13779-x.
111. Gao, R., Bai, S., Henderson, Y. C., Lin, Y., Schalck, A., Yan, Y., Kumar, T., Hu, M., Sei, E., Davis, A., et al. (2021) Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes, *Nat. Biotechnol.*, **39**, 599-608, doi: 10.1038/s41587-020-00795-2.
112. Menyailo, M. E., Zainullina, V. R., Tashireva, L. A., Zolotareva, S. Y., Gerashchenko, T. S., Alifanov, V. V., Savelieva, O. E., Grigoryeva, E. S., Tarabanovskaya, N. A., Popova, N. O., Khozyainova, A. A., Choinzonov, E. L., Cherdyntseva, N. V., Perelmuter, V. M., and Denisov, E. V. (2022) Heterogeneity of circulating epithelial cells in breast cancer at single-cell resolution: identifying tumor and hybrid cells, *bioRxiv*, doi: 10.1101/2021.11.24.469962.
113. Müller, S., Liu, S. J., Di Lullo, E., Malatesta, M., Pollen, A. A., Nowakowski, T. J., Kohanbash, G., Aghi, M., Kriegstein, A. R., Lim, D. A., and Diaz, A. (2016) Single-cell sequencing maps gene expression to mutational phylogenies in PDGF-and EGF-driven gliomas, *Mol. Syst. Biol.*, **12**, 889, doi: 10.15252/msb.20166969.
114. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M. A. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data, *Gen. Res.*, **20**, 1297-1303, doi: 10.1101/gr.107524.110.
115. Wu, T. D., Reeder, J., Lawrence, M., Becker, G., and Brauer, M. J. (2016) GMAP and GSNAP for Genomic Sequence Alignment: Enhancements to Speed, Accuracy, and Functionality, in *Statistical Genomics*, Springer, pp. 283-334.
116. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009) The sequence alignment/map format and SAMtools, *Bioinformatics*, **25**, 2078-2079, doi: 10.1093/bioinformatics/btp352.
117. Liu, F., Zhang, Y., Zhang, L., Li, Z., Fang, Q., Gao, R., and Zhang, Z. (2019) Systematic comparative analysis of single-nucleotide variant detection methods from single-cell RNA sequencing data, *Gen. Biol.*, **20**, 1-15, doi: 10.1186/s13059-019-1863-4.
118. Ludwig, L. S., Lareau, C. A., Ulirsch, J. C., Christian, E., Muus, C., Li, L. H., Pelka, K., Ge, W., Oren, Y., Brack, A., et al. (2019) Lineage tracing in humans enabled by mitochondrial mutations and single-cell genomics, *Cell*, **176**, 1325-1339. e1322, doi: 10.1016/j.cell.2019.01.022.

119. Wilson, G. W., Derouet, M., Darling, G. E., and Yeung, J. C. (2021) scSNV: accurate dscRNA-seq SNV co-expression analysis using duplicate tag collapsing, *Gen. Biol.*, **22**, 1-27, doi: 10.1186/s13059-021-02364-5.
120. Yao, Y., and Dai, W. (2014) Genomic instability and cancer, *J. Carcinog. Mutagen.*, **5**, 1000163, doi: 10.4172/2157-2518.1000165.
121. Fu, Y., Zhang, F., Zhang, X., Yin, J., Du, M., Jiang, M., Liu, L., Li, J., Huang, Y., and Wang, J. (2019) High-throughput single-cell whole-genome amplification through centrifugal emulsification and eMDA, *Commun. Biol.*, **2**, 1-10, doi: 10.1038/s42003-019-0401-y.
122. Schnepp, P. M., Chen, M., Keller, E. T., and Zhou, X. (2019) SNV identification from single-cell RNA sequencing data, *Hum. Mol. Genet.*, **28**, 3569-3583, doi: 10.1093/hmg/ddz207.
123. Ramazzotti, D., Angaroni, F., Maspero, D., Ascolani, G., Castiglioni, I., Piazza, R., Antoniotti, M., and Graudenzi, A. (2022) Variant calling from scRNA-seq data allows the assessment of cellular identity in patient-derived cell lines, *Nat. Commun.*, **13**, 1-3, doi: 10.1038/s41467-022-30230-w.
124. Zhou, Z., Xu, B., Minn, A., and Zhang, N. R. (2020) DENDRO: genetic heterogeneity profiling and subclone detection by single-cell RNA sequencing, *Genome Biol.*, **21**, 1-15, doi: 10.1186/s13059-019-1922-x.
125. McCarthy, D. J., Rostom, R., Huang, Y., Kunz, D. J., Danecek, P., Bonder, M. J., Hagai, T., Lyu, R., Wang, W., Gaffney, D. J., Simons, B. D., Stegle, O., and Teichmann, S. A. (2020) Cardelino: computational integration of somatic clonal substructure and single-cell transcriptomes, *Nat. Methods*, **17**, 414-421, doi: 10.1038/s41592-020-0766-3.
126. Ciccolella, S., Ricketts, C., Soto Gomez, M., Patterson, M., Silverbush, D., Bonizzoni, P., Hajirasouliha, I., and Della Vedova, G. (2020) Inferring cancer progression from Single-Cell Sequencing while allowing mutation losses, *Bioinformatics*, **37**, 326-333, doi: 10.1093/bioinformatics/btaa722.
127. Mehrabadi, F. R., Marie, K. L., Pérez-Guijarro, E., Malikić, S., Azer, E. S., Yang, H. H., Kızılkale, C., Gruen, C., Robinson, W., Liu, H., et al. (2021) Profiles of expressed mutations in single cells reveal subclonal expansion patterns and therapeutic impact of intratumor heterogeneity, *bioRxiv*, doi: 10.1101/2021.03.26.437185.
128. Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., Leonhardt, H., Heyn, H., Hellmann, I., and Enard, W. (2017) Comparative analysis of single-cell RNA sequencing methods, *Mol. Cell*, **65**, 631-643.e634, doi: 10.1016/j.molcel.2017.01.023.
129. Kim, T.-K., and Shiekhattar, R. (2015) Architectural and functional commonalities between enhancers and promoters, *Cell*, **162**, 948-959, doi: 10.1016/j.cell.2015.08.008.
130. Shlyueva, D., Stampfel, G., and Stark, A. (2014) Transcriptional enhancers: from properties to genome-wide predictions, *Nat. Rev. Genet.*, **15**, 272-286, doi: 10.1038/nrg3682.
131. Wray, G. A. (2007) The evolutionary significance of cis-regulatory mutations, *Nat. Rev. Genet.*, **8**, 206-216, doi: 10.1038/nrg2063.
132. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., and Greenleaf, W. J. (2013) Transposition of native chromatin for multimodal regulatory analysis and personal epigenomics, *Nat. Methods*, **10**, 1213, doi: 10.1038/nmeth.2688.
133. Wu, S. J., Furlan, S. N., Mihalas, A. B., Kaya-Okur, H. S., Feroze, A. H., Emerson, S. N., Zheng, Y., Carson, K., Cimino, P. J., and Keene, C. D. (2021) Single-cell CUT&Tag analysis of chromatin modifications in differentiation and tumor progression, *Nat. Biotechnol.*, **39**, 819-824, doi: 10.1038/s41587-021-00865-z.
134. Zhou, W., Ji, Z., Fang, W., and Ji, H. (2019) Global prediction of chromatin accessibility using small-cell-number and single-cell RNA-seq, *Nucleic Acids Res.*, **47**, e121-e121, doi: 10.1093/nar/gkz716.
135. Bailey, T. L., Johnson, J., Grant, C. E., and Noble, W. S. (2015) The MEME suite, *Nucleic Acids Res.*, **43**, W39-W49, doi: 10.1093/nar/gkv416.
136. Schep, A. N., Wu, B., Buenrostro, J. D., and Greenleaf, W. J. (2017) chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data, *Nat. Methods*, **14**, 975-978, doi: 10.1038/nmeth.4401.
137. Moody, J., Kouno, T., Suzuki, A., Shibayama, Y., Terao, C., Chang, J.-C., López-Redondo, F., Yip, C. W., Ando, Y., Yamamoto, K., Carninci, P., Shin, J. W., and Hon, C.-C. (2021) Profiling of transcribed cis-regulatory elements in single cells, *bioRxiv*, doi: 10.1101/2021.04.04.438388
138. Pliner, H. A., Packer, J. S., McFaline-Figueroa, J. L., Cusanovich, D. A., Daza, R. M., Aghamirzaie, D., Srivatsan, S., Qiu, X., Jackson, D., and Minkina, A. (2018) Cicero predicts cis-regulatory DNA interactions from single-cell chromatin accessibility data, *Mol. Cell*, **71**, 858-871.e858, doi: 10.1016/j.molcel.2018.06.044.
139. Oczko-Wojciechowska, M., Pfeifer, A., Jarzab, M., Swierniak, M., Rusinek, D., Tyszkiewicz, T., Kowalska, M., Chmielik, E., Zembala-Nozynska, E., Czarniecka, A., et al. (2020) Impact of the tumor microenvironment on the gene expression profile in papillary thyroid cancer, *Pathobiology*, **87**, 143-154, doi: 10.1159/000507223.
140. Pape, J., Magdeldin, T., Stamati, K., Nyga, A., Loizidou, M., Emberton, M., and Cheema, U. (2020) Cancer-associated fibroblasts mediate cancer progression and remodel the tumouroid stroma, *Br. J. Cancer*, **123**, 1178-1190, doi: 10.1038/s41416-020-0973-9.

141. Liu, J., Li, P., Wang, L., Li, M., Ge, Z., Noordam, L., Lieshout, R., Verstegen, M. M., Ma, B., and Su, J. (2021) Cancer-associated fibroblasts provide a stromal niche for liver cancer organoids that confers trophic effects and therapy resistance, *Cell. Mol. Gastroenterol. Hepatol.*, **11**, 407-431, doi: 10.1016/j.jcmgh.2020.09.003.
142. Moriel, N., Senel, E., Friedman, N., Rajewsky, N., Karaiskos, N., and Nitzan, M. (2021) NovoSpaRc: flexible spatial reconstruction of single-cell gene expression with optimal transport, *Nat. Protocols*, **16**, 4177-4200, doi: 10.1038/s41596-021-00573-7.
143. Ren, X., Zhong, G., Zhang, Q., Zhang, L., Sun, Y., and Zhang, Z. (2020) Reconstruction of cell spatial organization from single-cell RNA sequencing data based on ligand-receptor mediated self-assembly, *Cell Res.*, **30**, 763-778, doi: 10.1038/s41422-020-0353-2.
144. Chen, L., and Flies, D. B. (2013) Molecular mechanisms of T cell co-stimulation and co-inhibition, *Nat. Rev. Immunol.*, **13**, 227-242, doi: 10.1038/nri3405.
145. Ramilowski, J. A., Goldberg, T., Harshbarger, J., Kloppmann, E., Lizio, M., Satagopam, V. P., Itoh, M., Kawaji, H., Carninci, P., Rost, B., and Forrest, A. R. (2015) A draft network of ligand-receptor-mediated multicellular signalling in human, *Nat. Commun.*, **6**, 7866, doi: 10.1038/ncomms8866.
146. Moses, L., and Pachter, L. (2022) Museum of spatial transcriptomics, *Nat. Methods*, **19**, 534-546, doi: 10.1038/s41592-022-01409-2.
147. Hahaut, V., Pavlinic, D., Carbone, W., Schuierer, S., Balmer, P., Quinodoz, M., Renner, M., Roma, G., Cowan, C. S., and Picelli, S. (2022) Fast and highly sensitive full-length single-cell RNA sequencing using FLASH-seq, *Nat. Biotechnol.*, **40**, 1447-1451, doi: 10.1038/s41587-022-01312-3.
148. Cao, J., Cusanovich, D. A., Ramani, V., Aghamirzaie, D., Pliner, H. A., Hill, A. J., Daza, R. M., McFaline-Figueroa, J. L., Packer, J. S., Christiansen, L., et al. (2018) Joint profiling of chromatin accessibility and gene expression in thousands of single cells, *Science*, **361**, 1380-1385, doi: 10.1126/science.aau0730.
149. Chen, S., Lake, B. B., and Zhang, K. (2019) High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell, *Nat. Biotechnol.*, **37**, 1452-1457, doi: 10.1038/s41587-019-0290-0.
150. Zachariadis, V., Cheng, H., Andrews, N., and Enge, M. (2020) A highly scalable method for joint whole-genome sequencing and gene-expression profiling of single cells, *Mol Cell*, **80**, 541-553.e545, doi: 10.1016/j.molcel.2020.09.025.

OPPORTUNITIES OF COMPLEX ANALYSIS IN SINGLE-CELL RNA SEQUENCING

Review

A. A. Khozyainova[1]*, A. A. Valyaeva[2], M. S. Arbatsky[2], S. V. Isaev[3,4], P. S. Iamshchikov[1,5], E. V. Volchkov[6], M. S. Sabirov[7], V. R. Zainullina[1], V. I. Chechekhin[2], R. S. Vorobev[1], M. E. Menyailo[1], P. A. Tyurin-Kuzmin[2], and E. V. Denisov[1]

[1] *Cancer Research Institute Tomsk NRMC, 634050 Tomsk, Russia; e-mail: khozyainova@onco.tnimc.ru*

[2] *Lomonosov Moscow State University, 119991 Moscow, Russia*

[3] *Research Institute of Personalized Medicine, National Center for Personalized Medicine of Endocrine Diseases, The National Medical Research Center for Endocrinology, 117036 Moscow, Russia*

[4] *Phystech School of Biological and Medical Physics, Moscow Institute of Physics and Technology (National Research University), 115184 Dolgoprudny, Russia*

[5] *National Research Tomsk State University, 634050 Tomsk, Russia*

[6] *Dmitry Rogachev National Research Center of Pediatric Hematology, Oncology and Immunology, 117198 Moscow, Russia*

[7] *Koltzov Institute of Developmental Biology, 119334 Moscow, Russia*

Single-cell RNA sequencing (scRNA-seq) is a revolutionary tool for studying the physiology of normal and pathologically altered tissues. This approach provides information about the molecular features (gene expression, mutations, chromatin accessibility, etc.) of cells, opens up the possibility to analyze cell differentiation trajectories/phylogeny and cell–cell interactions and allows discovering new cell types and previously unexplored processes. From a clinical point of view, scRNA-seq allows a deeper and more detailed analysis of the molecular mechanisms of various diseases and serves as the basis for the development of new

preventive, diagnostic and therapeutic solutions. This review describes the different approaches to analysis of scRNA-seq data, reviews the strengths and weaknesses of bioinformatic tools, provides recommendations and examples of their successful use and suggests potential directions for improvement. It also emphasizes the need to create new, including multi-omics, protocols for the preparation of DNA/RNA libraries of single cells in order to obtain a more complete and systematic understanding of each cell.

Keywords: single-cell RNA sequencing, cell cycle, clustering, differential expression, cell type, trajectory inference, cell–cell interaction, gene regulatory network, copy number variation, single nucleotide variant, phylogenetics, epigenomics, spatial transcriptomics