Вавиловский журнал генетики и селекции
Vavilov Journal of Genetics and Breeding. 2024;28(8):974-981

doi 10.18699/vjgb-24-104

Методы реконструкции генных регуляторных сетей на основе транскриптомных данных отдельных клеток

М.А. Рыбаков[1, 2], Н.А. Омельянчук[1], Е.В. Землянская[1, 2]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
ezemlyanskaya@bionet.nsc.ru

Аннотация. Генные регуляторные сети – интерпретируемые графовые модели регуляции экспрессии генов – являются важным инструментом для понимания и исследования механизмов, которые клетки реализуют в процессе развития и при ответе на различные внутренние и внешние стимулы. Исторически первый подход для реконструкции генных регуляторных сетей основывался на анализе литературных сведений, в том числе обобщенных в базах данных. В настоящее время основной способ системной реконструкции генных регуляторных сетей – анализ омиксных (в первую очередь транскриптомных) данных; разработан ряд математических подходов для решения этой задачи. Развитие технологий получения омиксных данных для отдельных клеток сделало возможным проведение широкомасштабных молекулярно-генетических исследований с беспрецедентно высоким уровнем разрешения. В частности, появилась возможность реконструировать генные регуляторные сети для отдельных клеточных типов и для различных стадий развития клеток. Однако технические и биологические особенности омиксных данных отдельных клеток требуют специальных программ для решения этой задачи. В обзоре описаны подходы и программы, которые разработаны и используются для построения генных регуляторных сетей по транскриптомным данным отдельных клеток (scRNA-seq). Разбираются преимущества применения транскриптомных данных для отдельных клеток по сравнению с транскриптомами многоклеточных образцов, а также их недостатки в рамках решения задачи реконструкции регуляторных генных сетей. Существенное внимание уделяется повышению точности генных регуляторных сетей, построенных по транскриптомным данным отдельных клеток с помощью привлечения других омиксных данных, в первую очередь данных по сайтам связывания транскрипционных факторов и профилирования районов открытого хроматина (scATAC-seq). Рассматриваются вопросы применимости получаемых сетей в молекулярно-генетических исследованиях, приводятся примеры успешного использования генных регуляторных сетей, реконструированных различными методами с применением омиксных данных отдельных клеток для решения конкретных биологических задач. Обсуждаются перспективные направления развития этой области.

Ключевые слова: регуляторная генная сеть; данные для отдельных клеток; секвенирование РНК; scRNA-seq; scATAC-seq.

Для цитирования: Рыбаков М.А., Омельянчук Н.А., Землянская Е.В. Методы реконструкции генных регуляторных сетей на основе транскриптомных данных отдельных клеток. *Вавиловский журнал генетики и селекции*. 2024; 28(8):974-981. doi 10.18699/vjgb-24-104

Финансирование. Работа выполнена в рамках бюджетного проекта FWNR-2022-0020.

Reconstruction of gene regulatory networks from single cell transcriptomic data

M.A. Rybakov[1, 2], N.A. Omelyanchuk[1], E.V. Zemlyanskaya[1, 2]

[1] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
ezemlyanskaya@bionet.nsc.ru

Abstract. Gene regulatory networks (GRNs) – interpretable graph models of gene expression regulation – are a pivotal tool for understanding and investigating the mechanisms utilized by cells during development and in response to various internal and external stimuli. Historically, the first approach for the GRN reconstruction was based on the analysis of published data (including those summarized in databases). Currently, the primary GRN inference approach is the analysis of omics (mainly transcriptomic) data; a number of mathematical methods have been adapted for that. Obtaining omics data for individual cells has made it possible to conduct large-scale molecular genetic studies with an extremely high resolution. In particular, it has become possible to reconstruct GRNs for individual cell types and for various cell states. However, technical and biological features of single-cell omics data require specific approaches for

GRN inference. This review describes the approaches and programs that are used to reconstruct GRNs from single-cell RNA sequencing (scRNA-seq) data. We consider the advantages of using scRNA-seq data compared to bulk RNA-seq, as well as challenges in GRN interference. We pay specific attention to state-of-the-art methods for GRN reconstruction from single-cell transcriptomes recruiting other omics data, primarily transcription factor binding sites and open chromatin profiles (scATAC-seq), in order to increase inference accuracy. The review also considers the applicability of GRNs reconstructed from single-cell omics data to recover and characterize various biological processes. Future perspectives in this area are discussed.

Key words: gene regulatory network; single-cell data; RNA sequencing; scRNA-seq; scATAC-seq.

For citation: Rybakov M.A., Omelyanchuk N.A., Zemlyanskaya E.V. Reconstruction of gene regulatory networks from single cell transcriptomic data. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2024; 28(8):974-981. doi 10.18699/vjgb-24-104

Введение

Генная сеть – это управляющая формированием определенных признаков группа координированно экспрессирующихся генов, которые взаимодействуют друг с другом через кодируемые ими РНК и белки, а также продукты активности белков (Колчанов и др., 2013). Генные сети – центральный объект системной биологии. С целью более глубокого исследования отдельных аспектов выделяют специализированные типы генных сетей. Особое место среди них занимают генные регуляторные сети (ГРС), которые описывают осуществляемое транскрипционными факторами (ТФ) управление экспрессией генов – ключевой механизм гибкой реализации генетической информации (Huynh-Thu, Sanguinetti, 2019). Генные регуляторные сети визуализируют в виде графа взаимодействий между ТФ и регулируемыми ими генами (рис. 1, *а*) (Badia-i-Mompel et al., 2023). Каждая вершина в ГРС представляет собой ген (причем некоторые из них кодируют ТФ), а каждое ребро соответствует регуляторным отношениям между генами, кодирующими ТФ, и другими генами (эти отношения могут отражать реальные молекулярные взаимодействия ТФ с промоторами генов-мишеней или только их статистическую взаимосвязь). Ребро может иметь знак, указывающий, описывает ли оно активацию или ингибирование транскрипции, а также вес, отражающий, насколько сильно влияние регулятора. Таким образом, ГРС являются моделями логики регуляторных событий между генами в ходе выполнения клеточных программ (Tieri, Castiglione, 2021). Они представляют действенную альтернативу классическому моделированию в дифференциальных уравнениях в тех случаях, когда информация по кинетике процессов недоступна.

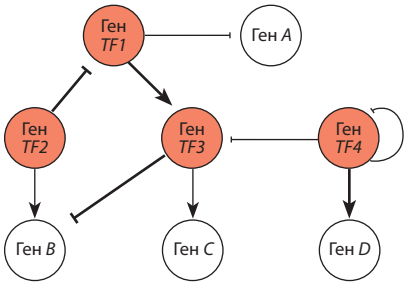
Генные регуляторные сети можно построить на основе информации о ТФ и их генах-мишенях из опубликованных статей или вывести *de novo* из транскриптомных данных (Badia-i-Mompel et al., 2023). Для решения этой задачи широко применялись транскриптомные даные клеточных популяций, получаемые секвенированием РНК (RNA-seq). В них уровень экспрессии каждого гена суммирован по всем клеткам в образце ткани или органа, взятом для секвенирования. Эти данные можно представить в виде так называемой матрицы экспрессии, в которой приводятся значения уровней экспрессии для каждого гена (им соответствуют строки) в различных образцах (им соответствуют столбцы) (см. рис. 1, *б*). Исходя из того, что уровни экспрессии генов, представленные в этих матрицах, являются результатом регуляции, осуществляемой путем связывания ТФ с промоторами генов-мишеней, можно построить математическую модель, которая объясняла бы наблюдаемую экспрессию генов (Mercatelli et al., 2020; Nguyen et al., 2021). На этой предпосылке основано большинство современных методов реконструкции ГРС по транскриптомным данным (Mercatelli et al., 2020). В настоящее время реконструкция ГРС по данным RNA-seq – одно из направлений системной биологии, в рамках которого разработано большое количество методов и компьютерных программ (Nguyen et al., 2019; Mercatelli et al., 2020).

В то же время описанный выше подход имеет слабые стороны. Во-первых, транскриптомные данные не содержат в явном виде сведения о конкретных регуляторных событиях (например, о связывании ТФ с промотором регулируемого им гена), все связи выводятся математически по уровням экспрессии генов. В результате в ГРС могут быть реконструированы несуществующие (ошибочные) связи. Привлечение данных, которые напрямую описывают управление транскрипцией (например, полногеномных профилей открытого хроматина или сайтов связывания ТФ), может значительно улучшить точность построенных ГРС (Sönmezer et al., 2020; Isbel et al., 2022). Во-вторых, данные RNA-seq не учитывают гетерогенность клеточных популяций, тогда как экспрессия генов может кардинально различаться в клетках разного типа. Эта проблема решается секвенированием транскриптомов отдельных клеток (scRNA-seq) (Tang et al., 2009).

Транскриптомные данные отдельных клеток представляют собой матрицу экспрессии, в которой строки соответствуют генам, а столбцы – клеткам (см. рис. 1, *в*), которые могут быть объединены по клеточным типам с помощью специально разработанных подходов (Luecken, Theis, 2019). Данные scRNA-seq открывают возможности исследовать биологические процессы на уровне отдельных типов клеток и новые перспективы для реконструкции и анализа ГРС (Nguyen et al., 2021). ГРС для отдельных типов клеток позволят обнаружить регуляторные контуры, специфичные для их состояний или степеней дифференцировки.

В настоящем обзоре мы рассматриваем методы реконструкции ГРС на основе транскриптомных данных отдельных клеток, подробно останавливаемся на привлечении для этого других омиксных данных, в первую очередь по сайтам связывания ТФ и полногеномным профилям открытого хроматина. Особое внимание уделяется описанию биологических результатов, которых удалось достичь с применением этих подходов.

а



б

Образцы

	s1	s2	s3	s4
Ген *TF1*	3	2	2	3
Ген *TF2*	6	4	5	6
Ген *TF3*	6	3	4	5
Ген *TF4*	3	2	3	4
Ген *A*	2	1	2	2
Ген *B*	2	1	1	3
Ген *C*	1	1	2	1
Ген *D*	1	1	1	0

в

Клетки

	a	b	c	d	e	f	g	h	i	j
Ген *TF1*	1	0	0	0	0	0	1	0	0	1
Ген *TF2*	0	1	0	1	1	0	0	2	1	0
Ген *TF3*	1	0	0	0	0	1	2	0	1	1
Ген *TF4*	1	1	0	1	0	0	0	0	0	0
Ген *A*	0	0	0	1	0	0	0	1	0	0
Ген *B*	0	0	0	1	1	0	0	0	0	0
Ген *C*	0	0	0	0	0	1	0	0	0	0
Ген *D*	0	1	0	0	0	0	0	0	0	0

Рис. 1. Генная регуляторная сеть и транскриптомные данные, на основании которых она может быть построена.

а – визуализация графовой модели ГРС; *б* – матрица экспрессии, построенная по данным RNA-seq для нескольких образцов (s1–s4); *в* – матрица экспрессии, построенная по данным scRNA-seq для одного образца. В вершинах графа (*а*) находятся гены, ребра отображают наличие регуляторной связи, ее направление, тип (активация или ингибирование транскрипции) и величину (чем больше вес ребра, тем сильнее влияние регулятора на транскрипцию). Красные вершины графа соответствуют генам, кодирующим ТФ, белые – другим генам. В ГРС ребра исходят только из вершин, которые соответствуют генам, кодирующим ТФ. На панели (*в*) разными цветами обозначены разные типы клеток.

Транскриптомы отдельных клеток как источник информации для ГРС

Помимо того, что транскриптомные данные отдельных клеток открывают возможность реконструировать ГРС для индивидуальных типов клеток, они имеют и другие преимущества по сравнению с транскриптомами клеточных популяций для реконструкции ГРС. Поскольку количество взаимодействий в ГРС, как правило, достаточно велико, для их качественной реконструкции необходимо использовать большее число транскриптомных профилей (столбцы в матрице экспрессии, см. рис. 1). Это не всегда достижимо для данных RNA-seq (см. рис. 1, *б*) (Altay, 2012), в то время как данные scRNA-seq содержат представительный (от нескольких сотен до нескольких тысяч) набор транскриптомов (см. рис. 1, *в*) (Luecken, Theis, 2019).

Генные регуляторные сети призваны описывать динамику регуляции экспрессии генов в различных биологических процессах, включая дифференцировку клеток и их реакции на различные внутренние и внешние стимулы. Для их наиболее точной реконструкции по данным RNA-seq требуется временной ряд образцов. В отличие от RNA-seq, данные scRNAseq, полученные в результате секвенирования одного образца, могут содержать информацию об изменении экспрессии генов во времени, если клетки в образце участвуют в одном и том же непрерывном биологическом процессе (например, дифференцировке) и находятся на разных его стадиях (Saelens et al., 2019; Hou et al., 2023). В таком случае вычислительное размещение клеток вдоль псевдовременной траектории (где порядок клеток определяется расстоянием между их транскриптомами) позволяет в хорошем приближении реконструировать динамику экспрессии генов в ходе процесса.

Однако не следует забывать, что в некоторых образцах клетки могут находиться в статичном состоянии или участвовать в многочисленных независимых процессах, что делает невозможной реконструкцию имеющих биологический смысл псевдовременных траекторий (Pratapa et al., 2020). Поэтому при выборе метода для реконструкции ГРС крайне важно определить, присутствует ли псевдовременная информация в наборе транскриптомов отдельных клеток, поскольку некоторые методы разработаны исключительно для данных с клеточной динамикой, в то время как другие подходят только для статических данных. Существуют также методы, применимые к обоим типам данных.

Одновременно транскриптомные профили отдельных клеток имеют свои особенности, которые затрудняют их анализ в целом и реконструкцию ГРС в частности (Wagner et al., 2016; Nguyen et al., 2021). К ним относятся кратковременная активация или низкий уровень экспрессии некоторых генов, изменения экспрессии генов в зависимости от стадии клеточного цикла и другие факторы. Широкое применение технологии scRNA-seq в биологии вызвало разработку большого числа алгоритмов для анализа генерируемых ею данных, по разному решающих эти проблемы.

Реконструкция ГРС по данным scRNA-seq

В этом разделе мы описываем основные категории популярных алгоритмов, используемых для построения ГРС по данным scRNA-seq: анализ корреляции и взаимной информации, регрессию, байесовские и логические сети, математическое моделирование на основе дифференциальных уравнений (рис. 2). Нужно отметить, что в ряде исследований, где проводилась оценка эффективности соответствующих инструментов с использованием как смоделированных, так и реальных данных scRNA-seq, ни один метод не был признан универсально лучшим (Chen, Mar, 2018; Blencowe et al., 2019; Pratapa et al., 2020). Такая вариативность может быть обусловлена тем, что каждый из этих методов подходит для конкретных типов и источников данных, для которых он был разработан.

Корреляция

Корреляция Пирсона, общепризнанный статистический индекс для расчета ассоциации между двумя переменными, была применена для измерения коэкспрессии генов, кодирующих ТФ, и их потенциальных мишеней в наборах данных RNA-seq и scRNA-seq (Hong et al., 2013; Nguyen et al., 2021). Будучи симметричной в своих аргументах, корреляция не предсказывает направленность регуляторного взаимодействия. Она может определять ассоциации между парами генов, которые не обязательно

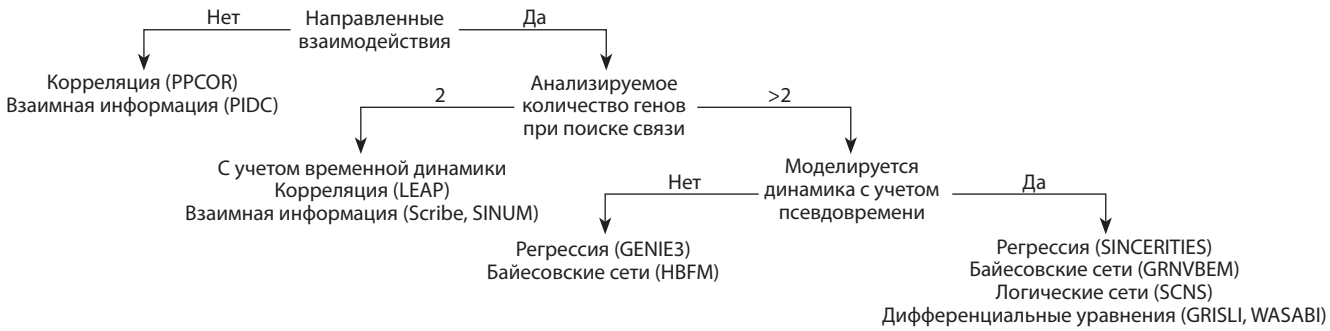


Рис. 2. Основные категории популярных алгоритмов, используемых для построения ГРС по данным scRNA-seq.

имеют прямую регуляторную связь. Такие методы, как PPCOR (Kim, 2015), учитывают влияние других генов, вычисляют получастные корреляции. LEAP (Specht, Li, 2017), алгоритм, разработанный специально для анализа данных отдельных клеток, определяет максимальную корреляцию Пирсона между каждой парой генов в окнах с переменным запаздыванием при условии, что клетки расположены в псевдовременном порядке. Поскольку этот тип корреляции несимметричный, он способен реконструировать направленные генные регуляторные сети. В результате тестирования данной программы на транскриптомах 564 отдельных дендритных клеток мыши LEAP выявил несколько тысяч ранее неизвестных связей между генами (Shalek et al., 2014).

Взаимная информация

Информационно-теоретические подходы используют взаимную информацию, которая измеряет снижение энтропии для одной переменной (например, уровня экспрессии одного гена) с учетом значения другой переменной (уровня экспрессии другого гена) (Chan et al., 2017; Qiu et al., 2020; Chang et al., 2024). Чтобы уменьшить ложные срабатывания, возникающие из-за косвенных связей между двумя генами, такие методы, как PIDC (Chan et al., 2017), используют частичное разложение информации (partial information decomposition, PID) для вычисления пропорционального уникального вклада (proportional unique contribution, PUC) для пары генов, который не может быть объяснен экспрессией третьего гена. Поскольку эта связь симметрична, реконструируемые ребра ненаправленны.

Метод PIDC был успешно применен для реконструкции ГРС по транскриптомам отдельных клеток для трех процессов у мыши: дифференцировки мегакариоцитов и эритроцитов от общего предшественника, раннего эмбриогенеза и эмбрионального гематопоэза. Во всех трех примерах PIDC нашел ранее неизвестные связи, эффективно выделил модули генов разных стадий дифференцировки и предполагаемые взаимодействия генов, осуществляющие переход между стадиями. При системной оценке 12 основанных на разных типах моделирования программ построения ГРС метод PIDC был указан как один из самых эффективных (Pratapa et al., 2020).

Алгоритм Scribe (Qiu et al., 2020) использует псевдовремя для вычисления ограниченной направленной информации (restricted directed information, RDI). Эта величина измеряет взаимную информацию между предшествующим уровнем экспрессии гена, кодирующего ТФ, и текущим уровнем экспрессии гена-мишени, обусловленную экспрессией регулятора ранее в псевдовременном ряду. Поскольку взаимная информация между предшествующим и текущим выражением несимметрична, Scribe может выводить направленные ребра. Этот алгоритм применялся как для верификациии существования отдельных связей в разных генных сетях, так и для реконструкции генной сети раннего эмбриогенеза *Caenorhabditis elegans*, где с его помощью была выведена известная иерархия регулирования транскрипции генов.

Третья программа, SINUM, также оценивающая взаимную информацию между любыми двумя генами и определяющая, являются ли они зависимыми или независимыми в конкретной клетке, была апробирована на различных данных и показала высокую эффективность в определении клеточных типов, их маркерных генов и связей между генами (Chang et al., 2024). Также с помощью программы были установлены изменения в ассоциациях генов в ходе дифференцировки эмбриональных стволовых клеток человека в эндодерму.

Регрессия

Генные регуляторные сети можно реконструировать, моделируя экспрессию каждого гена как функцию уровня экспрессии других генов и применяя методы, основанные на регрессии, для решения полученной системы уравнений (Huynh-Thu et al., 2010; Gao et al., 2017; Moerman et al., 2018). Алгоритм GENIE3 использует метод случайного леса, основанный на ансамбле регрессионных деревьев (Huynh-Thu et al., 2010). Вес ребра от ТФ к гену-мишени возникает из значимости ТФ в прогнозировании экспрессии гена-мишени, усредненной по всем деревьям регрессии в случайном лесу. Алгоритм GENIE3 был разработан и широко применялся для работы на данных массового секвенирования транскриптомов клеточных пупуляций. Программное обеспечение GRNBoost2 улучшает масштабируемость GENIE3, особенно с точки зрения эффективной обработки больших наборов данных отдельных клеток (Moerman et al., 2018). Оба инструмента, GENIE3 и GRNBoost2, продемонстрировали свою эффективность для реконструкции ГРС транскриптомов отдельных клеток, показав хорошее пересечение с реальными биологическими взаимодействиями (Kang et al., 2021).

Алгоритм SINCERITIES был создан специально для транскриптомов отдельных клеток, он решает регрессион-

ную модель, основанную на изменениях распределения уровней экспрессии каждого гена во времени и в псевдовремени (Gao et al., 2017). Алгоритмы GRNBoost2 и SINCERITIES были названы среди самых эффективных при системной оценке 12 основанных на разных типах моделирования программ построения ГРС (Pratapa et al., 2020). Тем не менее недавний сравнительный анализ на разных данных по различным показателям выявил, что GRNBoost2 в целом работает лучше, чем SINCERITIES, и более точно идентифицирует хабы в ГРС (Stock et al., 2024).

Байесовские сети

Еще один способ вывода ГРС заключается в моделировании регуляторных взаимодействий в байесовской сети. Алгоритм GRNVBEM работает с данными, представляющими собой временной ряд, т. е. клетки предварительно должны быть ранжированы по псевдовремени (Sanchez-Castillo et al., 2017). Затем он моделирует кратность изменения экспрессии гена между последовательными временными точками как линейную комбинацию экспрессии регуляторов гена в предшествующей точке в байесовской сети. Реконструкция ГРС раннего эмбриогенеза мыши и клеток почек *Danio rerio* с помощью этого метода позволила выделить хабы и сформировать гипотезы о регуляторах дифференцировки.

Метод HBFM основан на коэкспрессии генов с использованием разреженной иерархической модели байесовского фактора для снижения влияния высокой межклеточной изменчивости и шума в наборах данных отдельных клеток на прогнозируемую сеть (Sekula et al., 2020). Результаты работы программы показали значительное совпадение с известными и предсказанными белок-белковыми взаимодействиями из базы данных STRING.

Логические сети

В то время как ранее представленные методы предсказывают сети, описывающие регуляторные эффекты отдельных ТФ, они не учитывают логические правила, управляющие комбинаторным эффектом нескольких ТФ на экспрессию гена-мишени (Nguyen et al., 2021). Например, регуляторные механизмы могут включать активацию гена только в присутствии нескольких определенных ТФ или, в качестве альтернативы, его ингибирование другим ТФ, независимо от дополнительных факторов. Булевы сети способны характеризовать эти комбинации взаимодействий, представляя активное или неактивное состояние гена как двоичную переменную, дискретизированную с использованием порога экспрессии гена, и объединяя эти состояния с помощью операций AND, OR и NOT для объяснения экспрессии всех генов в системе.

Программа SCNS вычисляет логические правила, объясняющие прогрессию экспрессии генов от одной точки псевдовремени к другой (Woodhouse et al., 2018). На основе применения программы к транскриптомам клеток ранних стадий зародыша человека была построена коровая ГРС преимплантационного развития эмбриона. Алгоритм LogicNet использует вероятностную непрерывную логику для построения булевой сети, в которой экспрессия генов моделируется как непрерывная, а не двоичная переменная между 0 и 1 для построения ГРС с направленными и знаковыми ребрами (Malekpour et al., 2020). С помощью LogicNet были построены ГРС раннего эмбриогенеза мыши.

Дифференциальные уравнения

Наличие псевдовременной информации в данных scRNA-seq позволяет моделировать экспрессию генов с помощью обыкновенных дифференциальных уравнений (ОДУ) (Nguyen et al., 2021). Здесь скорость изменения экспрессии гена-мишени является функцией экспрессии гена, кодирующего регулирующий его ТФ. Решая эту систему уравнений, можно определить регуляторные связи на основе веса каждого ТФ в функции, описывающей изменение экспрессии гена. Алгоритм SCODE делает упрощающее предположение о том, что изменения в экспрессии генов можно определить как линейную комбинацию пространств уменьшенных размерностей для эффективного решения менее сложной системы уравнений с использованием линейной регрессии (Matsumoto et al., 2017). В качестве альтернативы GRISLI оценивает скорость, с которой экспрессия каждого гена меняется в соответствии с динамическим процессом в каждой клетке (Aubin-Frankowski, Vert, 2020). Впоследствии он упрощает систему уравнений, основываясь на предположении о том, что искомая ГРС имеет мало регуляторных ребер относительно числа генов в сети, тем самым сводя задачу к проблеме разреженной регрессии.

Ценным свойством GRISLI является то, что он позволяет клеткам следовать множественным траекториям дифференцировки, в то время как большинство методов допускают лишь линейную, неветвящуюся траекторию. Алгоритм DynGENIE3 применяет подход случайного леса GENIE3 для решения системы ОДУ, в которой изменение экспрессии одного гена определяется как потенциально нелинейная комбинация экспрессии других генов (Huynh-Thu, Geurts, 2018).

Другой класс подходов основан на наблюдении о том, что вариации экспрессии генов от клетки к клетке могут возникать из-за стохастической природы молекулярных регуляторных взаимодействий (Nguyen et al., 2021). Модель кусочно-детерминированного марковского процесса (PDMP) задает ОДУ для экспрессии гена как функцию стохастического двухстадийного марковского процесса, указывающего, активирована ли транскрипция гена, а не напрямую как функцию экспрессии регулирующих ТФ (Herbach et al., 2017).

Для каждого гена функция вероятности, представляющая переходы между активным и неактивным состояниями, включает вес каждого потенциального регулятора. Метод PDMP использует оценку максимального правдоподобия для определения этих весов и, таким образом, выводит ребра ГРС. Алгоритм WASABI реализует альтернативную оценку максимального правдоподобия, основанную на концепции, что наблюдаемому увеличению или уменьшению экспрессии гена должно предшествовать изменение количества регулирующего его активность белка в более раннем временном окне (Bonnaffoux et al., 2019). С применением WASABI для реконструкции ГРС дифференцировки эритроцитов у птиц обнаружены

ее необычные свойства: отсутствие хабов, распределенная структура сети и контроль экспрессии большинства генов непосредственно фактором, вызывающим дифференцировку.

Уточнение ГРС, реконструированных по данным scRNA-seq, путем распознавания сайтов связывания ТФ

Несмотря на широкое использование данных scRNA-seq для вывода ГРС, точность реконструкции реального регуляторного механизма на основании их остается неудовлетворительной (Chen, Mar, 2018; Pratapa et al., 2020). Эта проблема возникает потому, что программы для реконструкции ГРС из данных транскриптома основаны на предположении о том, что выявляемые взаимосвязи между уровнями экспрессии ТФ-кодирующего гена и его потенциальных генов-мишеней подразумевают прямое регулирование транскрипции. Однако наблюдаемые ассоциации могут быть вызваны другими биологическими явлениями или даже случайными причинами. Транскриптомные данные не содержат прямой информации о регуляторных событиях (например, о связывании ТФ с регуляторными районами генов). Таким образом, сложно найти различие между прямым и косвенным регулированием, основываясь исключительно на данных scRNA-seq.

Для решения этих проблем и повышения эффективности вывода ГРС необходимо привлекать дополнительные данные, напрямую характеризующие факторы, вовлеченные в регуляцию транскрипции. Например, геномные последовательности, содержащие регуляторные коды, могут быть использованы для определения потенциальных сайтов связывания ТФ. В этом случае наличие мотива, связывающего ТФ и расположенного в регуляторном районе гена-мишени, свидетельствует в пользу прямой регуляции между ТФ и геном-мишенью.

Для такого способа уточнения ГРС, идентифицированных с помощью алгоритма GENIE3, конвейер SCENIC использует базу данных мотивов связывания ТФ (Aibar et al., 2017). Он включает взаимодействия в сеть только в том случае, если мотивы, описывающие сайты связывания ТФ, обогащены в промоторных областях генов-мишеней. Более поздняя версия pySCENIC использует распараллеливание для повышения эффективности (Van de Sande et al., 2020). Конвейер SCENIC показал свою эффективность в определении типов клеток мозга мыши и человека (включая даже те, что были представлены двумя-шестью клетками), а также стадий развития раковых опухолей, которые более трудно различимы, чем клеточные типы (Aibar et al., 2017; Van de Sande et al., 2020). В дополнение для каждого типа клеток и стадий опухоли был установлен специфический набор ТФ, включая ранее неизвестные онкологические маркеры. Роль части из них в прогрессировании опухоли была подтверждена экспериментально в этих же исследованиях.

Интеграция данных scRNA-seq и scATAC-seq для реконструкции ГРС

ДНК в геноме упакована в нуклеосомы, базовые структурные единицы хроматина, которые препятствуют связыванию ТФ с ДНК, предотвращая транскрипцию генов (Parmar, Padinhateeri, 2020). Активация гена возможна, если только его регуляторный район свободен от нуклеосом. Нуклеосомная упаковка ДНК – это регулируемый процесс, она может различаться в зависимости от условий и клеточных типов. Технология scATAC-seq (single cell Assay for Transposase-Accessible Chromatin using sequencing) позволяет в отдельных клетках идентифицировать открытые участки хроматина, т. е. доступные для связывания ТФ регуляторные районы ДНК (Buenrostro et al., 2015). Таким образом, данные scATAC-seq могут способствовать более точной реконструкции прямой регуляторной связи между ТФ и их генами-мишенями в ГРС.

Было показано, что интеграция данных RNA-seq и ATAC-seq (или других эпигеномных данных) для многоклеточных образцов значительно повышает точность построения ГРС (Qin et al., 2014; Wang et al., 2015; Ackermann et al., 2016). Эта методология также применима к данным секвенирования отдельных клеток. Однако из-за специфичности транскриптомных и эпигеномных профилей по типу клеток и состоянию комбинирование данных RNA-seq с данными ATAC-seq или ChIP-seq, как правило, требует, чтобы оба набора данных были получены из клеток одного и того же типа в идентичных условиях.

На сегодняшний день разработаны технологии, позволяющие проводить одновременное секвенирование транскриптома и эпигенома для одной и той же клетки (Angermueller et al., 2016; Hu et al., 2016; Chen et al., 2019). Альтернативой является интегрированное исследование данных scRNA-seq и scATAC-seq, полученных с различных биологических образцов одинаковой природы. В этом случае дополнительную проблему для реконструкции ГРС составляет поиск соответствия между кластерами клеток, представляющих один и тот же тип, условие или состояние, по двум видам данных секвенирования. Для решения этой задачи разрабатываются методы так называемой диагональной интеграции (Argelaguet et al., 2021).

Поскольку по сравнению с другими методами профилирования эпигенома отдельных клеток scATAC-seq используется наиболее часто, было разработано несколько биоинформатических инструментов для объединения данных scRNA-seq и scATAC-seq при реконструкции ГРС (Loers, Vermeirssen, 2024). ГРС, реконструируемые на основе этих данных, получили специальное название «энхансерные ГРС» (эГРС). Метод STREAM реконструирует эГРС на базе совместно профилированных scRNA-seq и scATAC-seq данных с помощью модели задачи леса Штейнера, гибридного бикластерного конвейера и субмодульной оптимизации для вывода генной сети (Li et al., 2024). Метод STREAM был апробирован на данных отдельных клеток человека из органов с патологиями (болезни Альцгеймера и лимфоцитарной лимфомы) и показал свою эффективность в реконструкции связей ТФ–открытый сайт связывания–ген вдоль псевдовременной траектории и выявлении специфичных для этих болезней регуляций транскрипции генов.

Существуют и программы, которые используют результаты предварительного раздельного анализа данных scRNA-seq и scATAC-seq. Например, scMTNI на вход берет схему дифференцировки клеток, результаты scRNA-seq и

априорные сети на основе данных scATAC-seq для каждого типа клеток (Zhang et al., 2023). Применение scMTNI к данным scRNA-seq и scATAC-seq по репрограммированию клеток у мышей и дифференцировке гемопоэтических клеток у человека позволило построить эГРС как для линейных, так и для разветвленных путей дифференцировки и определить регуляторы и другие компоненты эГРС, специфичные для их разных этапов.

Заключение

Выявление взаимоотношений генов в процессе регуляции их экспрессии – ключ к пониманию механизмов, обеспечивающих реализацию генетической информации в тот или иной фенотипический признак. Построение ГРС по омиксным данным отдельных клеток предоставляет уникальную возможность системно исследовать механизмы клеточной дифференцировки, поскольку теоретически позволяет воссоздавать регуляторные генные сети для отдельных типов клеток и даже на отдельных стадиях их развития. На сегодняшний день разработан целый ряд методов для реконструкции таких ГРС, многие из них доступны пользователям в виде компьютерных программ. Несмотря на перспективность данного подхода, его потенциал далеко не полностью реализован. Не все доступные методы удобны для работы и для интерпретации их результатов.

Актуальной проблемой является также развитие методов верификации реконструируемых ГРС. Вероятно, поэтому использование данных моделей в конкретных биологических исследованиях пока ограничено, и можно говорить лишь об отдельных примерах успешного применения клеточных ГРС для решения биологических задач. Дальнейшее развитие технологий молекулярно-генетического исследования отдельных клеток и компьютерных методов анализа генерируемых ими данных (в частности, с целью реконструкции ГРС и их анализа) позволит существенно сократить разрыв между нашими знаниями о молекулярных детерминантах признаков (в том числе на клеточном уровне) и транскрипционных каскадах, запускаемых под воздействием внешних или внутренних стимулов. Прорывные открытия, сделанные с помощью ГРС, реконструированных на основании омиксных данных для отдельных клеток, возможно, ждут нас в будущем.

Список литературы / References

Колчанов Н.А., Игнатьева Е.В., Подколодная О.А., Лихошвай В.А., Матушкин Ю.Г. Генные сети. *Вавиловский журнал генетики и селекции*. 2013;17(4/2):833-850
[Kolchanov N.A., Ignatieva E.V., Podkolodnaya O.A., Likhoshvai V.A., Matushkin Yu.G. Gene networks. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2013;17(4/2):833-850 (in Russian)]

Ackermann A.M., Wang Z., Schug J., Naji A., Kaestner K.H. Integration of ATAC-seq and RNA-seq identifies human alpha cell and beta cell signature genes. *Mol. Metab*. 2016;5(3):233-244. doi 10.1016/j.molmet.2016.01.002

Aibar S., González-Blas C.B., Moerman T., Huynh-Thu V.A., Imrichova H., Hulselmans G., Rambow F., Marine J., Geurts P., Aerts J., Van Den Oord J., Atak Z.K., Wouters J., Aerts S. SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods*. 2017;14(11):1083-1086. doi 10.1038/nmeth.4463

Altay G. Empirically determining the sample size for large-scale gene network inference algorithms. *IET Syst. Biol*. 2012;6(2):35-43. doi 10.1049/iet-syb.2010.0091

Angermueller C., Clark S.J., Lee H.J., Macaulay I.C., Teng M.J., Hu T.X., Krueger F., Smallwood S.A., Ponting C.P., Voet T., Kelsey G., Stegle O., Reik W. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat. Methods*. 2016;13(3):229-232. doi 10.1038/nmeth.3728

Argelaguet R., Cuomo A.S.E., Stegle O., Marioni J.C. Computational principles and challenges in single-cell data integration. *Nat. Biotechnol*. 2021;39(10):1202-1215. doi 10.1038/s41587-021-00895-7

Aubin-Frankowski P., Vert J. Gene regulation inference from single-cell RNA-seq data with linear differential equations and velocity inference. *Bioinformatics*. 2020;36(18):4774-4780. doi 10.1093/bioinformatics/btaa576

Badia-i-Mompel P., Wessels L., Müller-Dott S., Trimbour R., Flores R.O.R., Argelaguet R., Saez-Rodriguez J. Gene regulatory network inference in the era of single-cell multi-omics. *Nat. Rev. Genet*. 2023;24(11):739-754. doi 10.1038/s41576-023-00618-5

Blencowe M., Arneson D., Ding J., Chen Y.W., Saleem Z., Yang X. Network modeling of single-cell omics data: challenges, opportunities, and progresses. *Emerg. Top. Life Sci*. 2019;3(4):379-398. doi 10.1042/ETLS20180176

Bonnaffoux A., Herbach U., Richard A., Guillemin A., Gonin-Giraud S., Gros P., Gandrillon O. WASABI: a dynamic iterative framework for gene regulatory network inference. *BMC Bioinformatics*. 2019;20(1):220. doi 10.1186/s12859-019-2798-1

Buenrostro J.D., Wu B., Litzenburger U.M., Ruff D., Gonzales M.L., Snyder M.P., Chang H.Y., Greenleaf W.J. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*. 2015;523(7561):486-490. doi 10.1038/nature14590

Chan T.E., Stumpf M.P., Babtie A.C. Gene regulatory network inference from single-cell data using multivariate information measures. *Cell Syst*. 2017;5(3):251-267.e3. doi 10.1016/j.cels.2017.08.014

Chang L., Hao T., Wang W., Lin C. Inference of single-cell network using mutual information for scRNA-seq data analysis. *BMC Bioinformatics*. 2024;25(S2):292. doi 10.1186/s12859-024-05895-3

Chen S., Mar J.C. Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. *BMC Bioinformatics*. 2018;19(1):232. doi 10.1186/s12859-018-2217-z

Chen S., Lake B.B., Zhang K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat. Biotechnol*. 2019;37(12):1452-1457. doi 10.1038/s41587-019-0290-0

Gao N.P., Ud-Dean S.M.M., Gandrillon O., Gunawan R. SINCERITIES: inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. *Bioinformatics*. 2017;34(2):258-266. doi 10.1093/bioinformatics/btx575

Herbach U., Bonnaffoux A., Espinasse T., Gandrillon O. Inferring gene regulatory networks from single-cell data: a mechanistic approach. *BMC Syst. Biol*. 2017;11(1):105. doi 10.1186/s12918-017-0487-0

Hong S., Chen X., Jin L., Xiong M. Canonical correlation analysis for RNA-seq co-expression networks. *Nucleic Acids Res*. 2013;41(8):e95. doi 10.1093/nar/gkt145

Hou W., Ji Z., Chen Z., Wherry E.J., Hicks S.C., Ji H. A statistical framework for differential pseudotime analysis with multiple single-cell RNA-seq samples. *Nat. Commun*. 2023;14(1):7286. doi 10.1038/s41467-023-42841-y

Hu Y., Huang K., An Q., Du G., Hu G., Xue J., Zhu X., Wang C., Xue Z., Fan G. Simultaneous profiling of transcriptome and DNA methylome from a single cell. *Genome Biol*. 2016;17(1):88. doi 10.1186/s13059-016-0950-z

Huynh-Thu V.A., Geurts P. dynGENIE3: dynamical GENIE3 for the inference of gene networks from time series expression data. *Sci. Rep*. 2018;8(1):3384. doi 10.1038/s41598-018-21715-0

Huynh-Thu V.A., Sanguinetti G. Gene regulatory network inference: An introductory survey. *Methods Mol. Biol.* 2019;1883:1-23. doi 10.1007/978-1-4939-8882-2_1

Huynh-Thu V.A., Irrthum A., Wehenkel L., Geurts P. Inferring regulatory networks from expression data using tree-based methods. *PLoS One*. 2010;5(9):e12776. doi 10.1371/journal.pone.0012776

Isbel L., Grand R.S., Schübeler D. Generating specificity in genome regulation through transcription factor sensitivity to chromatin. *Nat. Rev. Genet*. 2022;23(12):728-740. doi 10.1038/s41576-022-00512-6

Kang Y., Thieffry D., Cantini L. Evaluating the reproducibility of single-cell gene regulatory network inference algorithms. *Front. Genet*. 2021;12:617282. doi 10.3389/fgene.2021.617282

Kim S. ppcor: An R package for a fast calculation to semi-partial correlation coefficients. *Commun. Stat. Appl. Methods*. 2015;22(6):665-674. doi 10.5351/CSAM.2015.22.6.665

Li Y., Ma A., Wang Y., Guo Q., Wang C., Fu H., Liu B., Ma Q. Enhancer-driven gene regulatory networks inference from single-cell RNA-seq and ATAC-seq data. *Brief. Bioinform*. 2024;25(5):bbae369. doi 10.1093/bib/bbae369

Loers J.U., Vermeirssen V. A single-cell multimodal view on gene regulatory network inference from transcriptomics and chromatin accessibility data. *Brief. Bioinform*. 2024;25(5):bbae382. doi 10.1093/bib/bbae382

Luecken M.D., Theis F.J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol*. 2019;15(6):e8746. doi 10.15252/msb.20188746

Malekpour S.A., Alizad-Rahvar A.R., Sadeghi M. LogicNet: probabilistic continuous logics in reconstructing gene regulatory networks. *BMC Bioinformatics*. 2020;21(1):318. doi 10.1186/s12859-020-03651-x

Matsumoto H., Kiryu H., Furusawa C., Ko M.S.H., Ko S.B.H., Gouda N., Hayashi T., Nikaido I. SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. *Bioinformatics*. 2017;33(15):2314-2321. doi 10.1093/bioinformatics/btx194

Mercatelli D., Scalambra L., Triboli L., Ray F., Giorgi F.M. Gene regulatory network inference resources: A practical overview. *Biochim. Biophys. Acta Gene Regul. Mech*. 2020;1863(6):194430. doi 10.1016/j.bbagrm.2019.194430

Moerman T., Santos S.A., González-Blas C.B., Simm J., Moreau Y., Aerts J., Aerts S. GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics*. 2018;35(12):2159-2161. doi 10.1093/bioinformatics/bty916

Nguyen H., Shrestha S., Tran D., Shafi A., Draghici S., Nguyen T. A comprehensive survey of tools and software for active subnetwork identification. *Front. Genet*. 2019;10:155. doi 10.3389/fgene.2019.00155

Nguyen H., Tran D., Tran B., Pehlivan B., Nguyen T. A comprehensive survey of regulatory network inference methods using single cell RNA sequencing data. *Brief. Bioinform*. 2021;22(3):bbaa190. doi 10.1093/bib/bbaa190

Parmar J.J., Padinhateeri R. Nucleosome positioning and chromatin organization. *Curr. Opin. Struct. Biol*. 2020;64:111-118. doi 10.1016/j.sbi.2020.06.021

Pratapa A., Jalihal A.P., Law J.N., Bharadwaj A., Murali T.M. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat. Methods*. 2020;17(2):147-154. doi 10.1038/s41592-019-0690-6

Qin J., Hu Y., Xu F., Yalamanchili H.K., Wang J. Inferring gene regulatory networks by integrating ChIP-seq/chip and transcriptome data via LASSO-type regularization methods. *Methods*. 2014;67(3):294-303. doi 10.1016/j.ymeth.2014.03.006

Qiu X., Rahimzamani A., Wang L., Ren B., Mao Q., Durham T., McFaline-Figueroa J.L., Saunders L., Trapnell C., Kannan S. Inferring causal gene regulatory networks from coupled single-cell expression dynamics using scribe. *Cell Syst*. 2020;10(3):265-274. doi 10.1016/j.cels.2020.02.003

Saelens W., Cannoodt R., Todorov H., Saeys Y. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol*. 2019;37(5):547-554. doi 10.1038/s41587-019-0071-9

Sanchez-Castillo M., Blanco D., Tienda-Luna I.M., Carrion M.C., Huang Y. A Bayesian framework for the inference of gene regulatory networks from time and pseudo-time series data. *Bioinformatics*. 2017;34(6):964-970. doi 10.1093/bioinformatics/btx605

Sekula M., Gaskins J., Datta S. A sparse Bayesian factor model for the construction of gene co-expression networks from single-cell RNA sequencing count data. *BMC Bioinformatics*. 2020;21(1):361. doi 10.1186/s12859-020-03707-y

Shalek A.K., Satija R., Shuga J., Trombetta J.J., Gennert D., Lu D., Chen P., Gertner R.S., Gaublomme J.T., Yosef N., Schwartz S., Fowler B., Weaver S., Wang J., Wang X., Ding R., Raychowdhury R., Friedman N., Hacohen N., Park H., May A.P., Regev A. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature*. 2014;510(7505):363-369. doi 10.1038/nature13437

Sönmezer C., Kleinendorst R., Imanci D., Barzaghi G., Villacorta L., Schübeler D., Benes V., Molina N., Krebs A.R. Molecular co-occupancy identifies transcription factor binding cooperativity *in vivo*. *Mol. Cell*. 2020;81(2):255-267. doi 10.1016/j.molcel.2020.11.015

Specht A.T., Li J. LEAP: constructing gene co-expression networks for single-cell RNA-sequencing data using pseudotime ordering. *Bioinformatics*. 2017;33(5):764-766. doi 10.1093/bioinformatics/btw729

Stock M., Popp N., Fiorentino J., Scialdone A. Topological benchmarking of algorithms to infer gene regulatory networks from single-cell RNA-seq data. *Bioinformatics.* 2024;40(5):btae267. doi 10.1093/bioinformatics/btae267

Tang F., Barbacioru C., Wang Y., Nordman E., Lee C., Xu N., Wang X., Bodeau J., Tuch B.B., Siddiqui A., Lao K., Surani M.A. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods*. 2009;6(5):377-382. doi 10.1038/nmeth.1315

Tieri P., Castiglione F. Modeling macrophage differentiation and cellular dynamics. In: Wolkenhauer O. (Ed.). Systems Medicine. Integrative, Qualitative and Computational Approaches. Academic Press, 2021;511-520. doi 10.1016/B978-0-12-801238-3.11644-7

Van de Sande B., Flerin C., Davie K., De Waegeneer M., Hulselmans G., Aibar S., Seurinck R., Saelens W., Cannoodt R., Rouchon Q., Verbeiren T., De Maeyer D., Reumers J., Saeys Y., Aerts S. A scalable SCENIC workflow for single-cell gene regulatory network analysis. *Nat. Protoc*. 2020;15(7):2247-2276. doi 10.1038/s41596-020-0336-2

Wagner A., Regev A., Yosef N. Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol*. 2016;34(11):1145-1160. doi 10.1038/nbt.3711

Wang P., Qin J., Qin Y., Zhu Y., Wang L.Y., Li M.J., Zhang M.Q., Wang J. ChIP-Array 2: integrating multiple omics data to construct gene regulatory networks. *Nucleic Acids Res*. 2015;43(W1):264-269. doi 10.1093/nar/gkv398

Woodhouse S., Piterman N., Wintersteiger C.M., Göttgens B., Fisher J. SCNS: a graphical tool for reconstructing executable regulatory networks from single-cell genomic data. *BMC Syst. Biol.* 2018;12(1):59. doi 10.1186/s12918-018-0581-y

Zhang S., Pyne S., Pietrzak S., Halberg S., McCalla S.G., Siahpirani A.F., Sridharan R., Roy S. Inference of cell type-specific gene regulatory networks on cell lineages from single cell omic datasets. *Nat. Commun*. 2023;14(1):3064. doi 10.1038/s41467-023-38637-9

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.
Поступила в редакцию 28.10.2024. После доработки 21.11.2024. Принята к публикации 22.11.2024.