

Моделирование экспрессии генов в экспериментах секвенирования молекул РНК одиночных клеток

М. К. Чепелева¹, Н. Н. Яцков¹, А. Л. Одинец¹, П. В. Назаров²

¹Белорусский государственный университет, Минск, Беларусь;
e-mail:maryna.chepeleva@gmail.com

²Luxembourg Institute of Health, Strassen, Luxembourg

В работе предложена модель генерации данных экспрессии генов в экспериментах секвенирования молекул РНК одиночных клеток. Модель является универсальной, отличается широкой вариацией генерируемых данных из-за использования различных распределений экспрессии генов, позволяет воспроизводить работу биологических функций. Модель может применяться для валидации алгоритмов анализа данных транскриптомов одиночных клеток, в том числе, алгоритмов деконволюции сигналов генов и поиска биологических функций.

Ключевые слова: моделирование транскриптома, секвенирование одиночных клеток, экспрессия генов, биологические функции.

Введение

Одним из направлений исследований в биоинформатике является анализ влияния биофизических параметров на экспрессию генов [1]. При геномном секвенировании образцов тканей (bulk samples) измеряется средний уровень экспрессии генов множества клеток, что является существенным ограничением [2]. Новые технологии следующего поколения (NGS – next generation sequencing) позволяют извлекать и амплифицировать даже незначительное количество молекул РНК, что обеспечивает секвенирование на уровне отдельных клеток [3]. Секвенирование РНК одиночных клеток (scRNA-seq) – мощная высокопроизводительная технология, затрагивающая клеточно-специфические особенности транскриптомов, что позволяет выявить важную информацию о межклеточной гетерогенности в различных тканях, органах и биосистемах, а так же способствует открытиям новых типов и состояний клеток [4, 5]. Возможности метода осложняются отсутствием унифицированных протоколов проведения экспериментов и необходимостью разработки эффективных алгоритмов анализа больших данных.

Классическим подходом к валидации методов анализа является использование смоделированных синтетических данных. Смоделированные данные позволяют последовательно проверять гипотезы, оценивать правильность реализации тестируемых алгоритмов, выявлять ограничения методов, а также значительно снижать стоимость экспериментов [2]. Из-за внутренней биологической вариативности и технического шума данные секвенирования последовательностей РНК одиночных клеток часто демонстрируют мультимодальные распределения [6]. В основе большинства имитационных моделей данных лежит генерация значений экспрессии каждого гена по заданному распределению. Наиболее широко используемые распределения для гомологичных клеток: Пуассона, Пуассона с нулевым раздутием (zero-inflated), отрицательное биномиальное, отрицательное биномиальное с нулевым раздутием [7]. Параметры распределений задаются вручную либо оцениваются экспериментально. Важным свойством модели является сохранение взаимодействий между генами, в основе которых могут лежать абстракции функционирования генных сетей или биологических функций.

Для использования модели в систематическом сравнительном анализе алгоритмов необходимо покрытие различных технических (глубина секвенирования, количество нулей в данных) и морфологических (клеточные линии и типы клеток)

особенностей данных. Кроме того, немаловажным фактором является гибкость и прозрачность модели.

Цель данной работы – разработка модели для генерации данных секвенирования одиночных клеток, отвечающей условиям универсальности в применении к различным задачам и сохранения генных взаимодействий.

Имитационная модель

На рис. 1 представлена схема генерации данных секвенирования одиночных клеток k разных типов или линий.

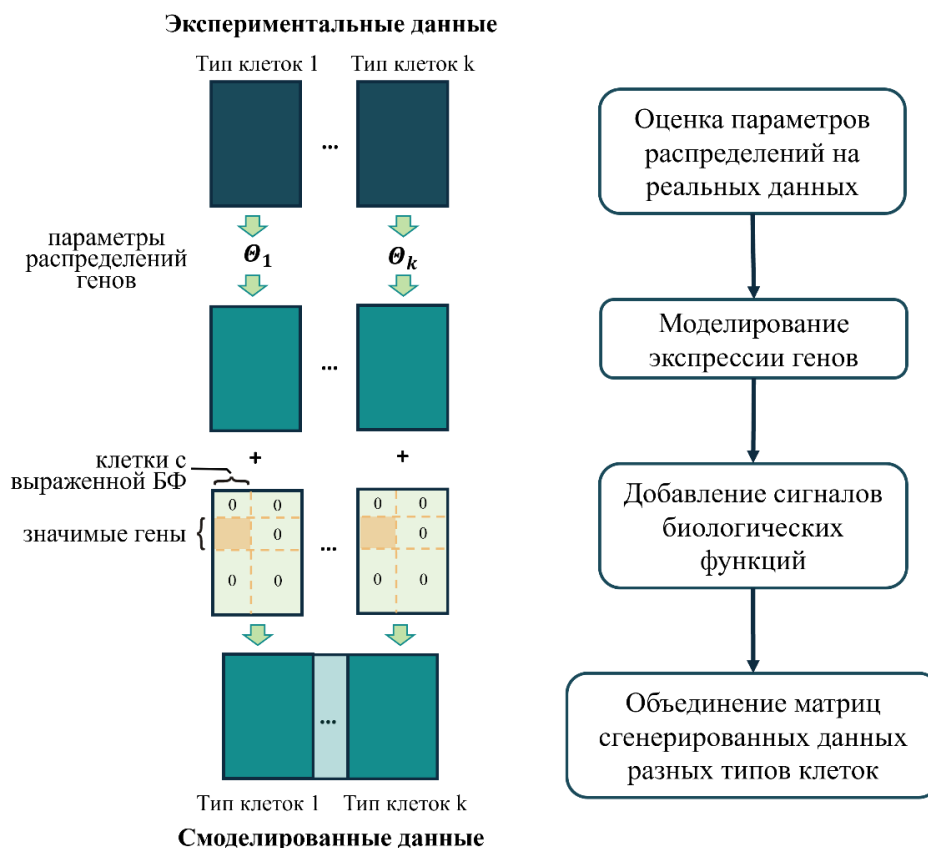


Рис. 1. – Схема алгоритма генерации данных секвенирования одиночных клеток.

Для того, чтобы разработанная модель могла быть адаптированной под различные протоколы экспериментов и удовлетворяла условиям универсальности, параметры модели настраиваются на реальном наборе данных. В процессе подбора параметров, реализованном на базе R-пакета *scDesign2* [8], измеренная экспрессия гена аппроксимируется отрицательным биномиальным распределением с нулевым раздутием ZINB (zero-inflated negative binomial), в частных случаях сводимых к распределениям Пуассона, Пуассона с нулевым раздутием, отрицательному биномиальному [9]. Оцененные параметры используются для моделирования экспрессии гена.

Модель генерирует матрицу экспрессии X размером $n \times m$, где n – количество генов, m – количество клеток. Для моделирования вектора экспрессии X_i для i -го гена длиной m задается распределение с заранее определенными параметрами:

$$X_i \sim \text{ZINB}(p_i, \mu_i, u_i), \quad (1)$$

где p_i – величина раздутия, u_i – параметр дисперсии. Математическое ожидание и дисперсия случайной величины X_i [8]:

$$M[X_i] = \mu_i, \quad (2)$$

$$D[X_i] = \mu_i + \frac{\mu_i^2}{u_i}. \quad (3)$$

Задание генных взаимодействий через биологические функции производится путем добавления дополнительного гауссовского сигнала к n_s значимым генам (обогащающим заданную биофункцию) для m_c образцов или клеток с выраженной биофункцией, то есть исследуемой группы. Остальные клетки – контрольная группа образцов:

$$X' = X + Bf, \quad (4)$$

$$Bf_{lr} = \begin{cases} \mathcal{N}_{lr}(\mu_{Bf}, \sigma_{Bf}^2), & l \leq n_s, r \leq m_c, \\ 0, & l > n_s, r > m_c \end{cases}, \quad (5)$$

где μ_{Bf} и σ_{Bf}^2 – математическое ожидание и дисперсия нормального распределения соответственно, причем $\mu_{Bf} > 0$.

Такое приближение соответствует наложению сигнала биофункции Bf с положительной регуляцией генов, задействованных в данном биологическом процессе.

В случае генерации k типов клеток, итоговая матрица является объединением

$$X' = [X^{(1)'} \dots X^{(k)'}], \quad (6)$$

где $X^{(k)}$ – матрица сгенерированных значений для типа клеток k .

Результаты

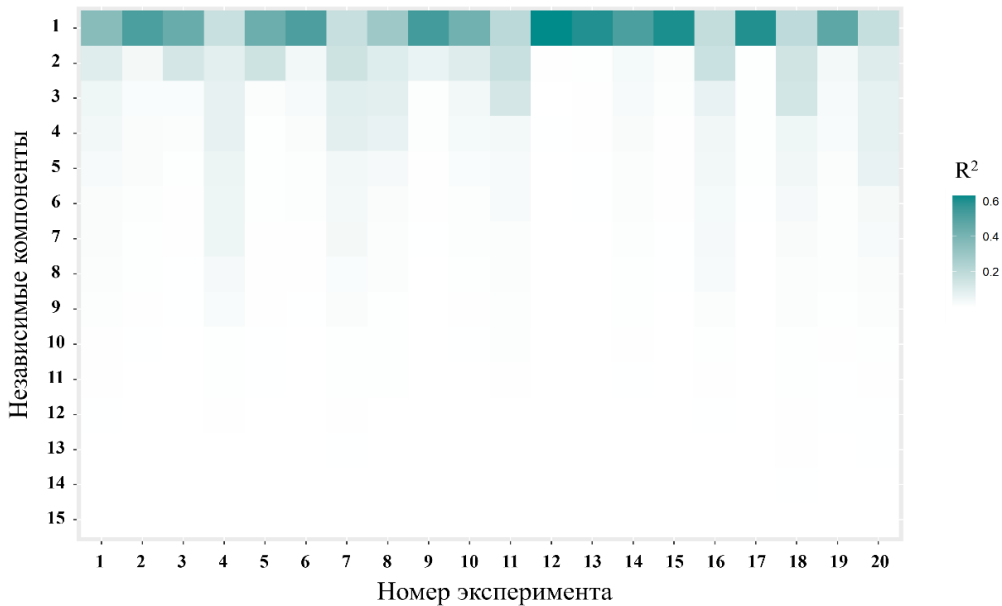


Рис. 2. – Квадрат коэффициента корреляции Пирсона (R^2) между выделенными независимыми компонентами и сигналом биологической функции для 20 экспериментов генерации данных. Компоненты в каждом эксперименте отсортированы по убыванию R^2 .

Разработанная модель позволяет генерировать данные секвенирования одиночных клеток, схожие с реальными наборами данных различных экспериментов за счет настройки параметров на реальных данных. Механизм добавления сигналов

биологической функции был протестирован с помощью деконволюции сгенерированных данных методом независимых компонент. После разложения на $m/3$ компонент была выявлена компонента, скоррелированная с добавочным сигналом биологической функции (рис. 2).

Заключение

Разработана модель генерации данных экспрессии генов в экспериментах секвенирования молекул РНК одиночных клеток. Модель является универсальной, отличается широтой вариации генерируемых данных, что обусловлено использованием различных распределений экспрессии генов, и позволяет воспроизводить работу биологических функций.

Предложенная модель может применяться для тестирования и валидации алгоритмов интеллектуального анализа данных транскриптомов одиночных клеток, в том числе, алгоритмов деконволюции данных и выделения сигналов биологических процессов.

Литература

1. Advances in Bioinformatics. Edit. Singh V., Ajay Kumar A. 2021. P. 8–10.
2. Zappia L., Phipson B., Oshlack A. Splatter: simulation of single-cell RNA sequencing data. *Genome Biol.* 2017. Vol. 18, No. 1. P. 174.
3. 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. Vol. 6, No. 5. P. 377–382.
4. Villani A.-C., Satija R., Reynolds G., Sarkizova S., Shekhar K., Fletcher J., Griesbeck M., Butler A., Zheng S., Lazo S., et al. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science.* 2017. Vol. 356, No. 6335. P. eaah4573.
5. Birnbaum K. D. Power in numbers: single-cell RNA-seq strategies to dissect complex tissues. *Ann Rev Genet.* 2018. Vol. 52. P. 203–221.
6. Bacher R., Kendziorowski C. Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol.* 2016. Vol. 17. P. 63–77.
7. Sarkar A., Stephens M. Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis. *Nat Genet.* 2021. Vol. 53, No. 6. P. 770–777.
8. Sun T., Song D., Li W. V. scDesign2: a transparent simulator that generates high-fidelity single-cell gene expression count data with gene correlations captured. *Genome Biol.* 2021. Vol. 22, No. 1. P. 163–200.
9. Williamson J. M., Lin H.-M., Lyles R. H., Hightower A. W. Power Calculations for ZIP and ZINB Models, *J. data sci.* 2021. Vol. 5, No. 4. P. 519–534.

Simulation of single-cell RNA- sequencing gene expression data

M. K. Chepeleva¹, M. M. Yatskou¹, A. L. Adzinets¹, P. V. Nazarov²

¹ Belarusian State University, Minsk, Belarus; e-mail:maryna.chepeleva@gmail.com

² Luxembourg Institute of Health, Strassen, Luxembourg

Here we propose the simulating model for generating gene expression data for single cell RNA sequencing. The simulator is data-driven and adaptive to various experimental protocols and problems. It uses four gene expression distributions that increases the variation in simulated data and allows adding biological processes signals in data. The model can be used to validate algorithms for single cell transcriptome data analysis, including deconvolution methods and enrichment analysis.

Keywords: transcriptome modeling, single-cell RNA-seq, gene expression, biological processes.