



ПРОБЛЕМАТИКА ДОСТУПНОСТИ БАЗ ОДНОКЛЕТОЧНЫХ ОМИКСНЫХ ДАННЫХ

Мусакаев Руслан

Ташкентский Государственный технический
университет

Кафедра Биотехнологии

Аспирант (PHD)

Email: marasres08121999@mail.ru

Аннотация Стремительное развитие технологий одноклеточной транскриптомики, эпигеномики, протеомики и других направлений одноклеточной омиксики привело к формированию нового класса биомедицинских данных, характеризующихся значительным объёмом, высокой структурной сложностью и большим количеством сопутствующих метаданных. В отличие от традиционных bulk-методов, одноклеточные технологии позволяют исследовать молекулярную неоднородность тканей, выявлять редкие клеточные популяции и анализировать различные состояния отдельных клеток. Одновременно увеличение объёма одноклеточных данных приводит к проблемам их хранения, поиска, стандартизации, интероперабельности и повторного использования. В статье рассматриваются основные проблемы доступности баз одноклеточных омиксных данных, включая фрагментацию репозиториев, неоднородность форматов, недостаточную стандартизацию метаданных, влияние различий экспериментальных протоколов, проблемы воспроизводимости и ограничения, связанные с конфиденциальностью данных человека. Рассматриваются FAIR-принципы как один из основных подходов к повышению доступности и повторного использования научных

данных, а также перспективы развития интегрированных клеточных атласов.

Ключевые слова: *одноклеточная омиксика, single-cell omics, scRNA-seq, биоинформатика, базы данных, метаданные, стандартизация, FAIR-принципы, клеточные атласы, интеграция данных.*

Введение

Развитие технологий анализа отдельных клеток стало одним из наиболее значимых направлений современной биоинформатики, молекулярной биологии и медицины. В отличие от традиционных методов, при которых исследуется усреднённый молекулярный профиль клеточной популяции, одноклеточные технологии позволяют получать информацию о состоянии отдельных клеток. Это делает возможным исследование клеточной гетерогенности и выявление редких клеточных популяций, которые могут быть неразличимы при использовании bulk-методов [1, 2]. Одним из наиболее распространённых методов является секвенирование РНК отдельных клеток — single-cell RNA sequencing (scRNA-seq). Его применение позволило существенно расширить представления о клеточном составе тканей и механизмах развития различных заболеваний. Одновременно развиваются методы одноклеточной эпигеномики, протеомики и пространственной транскриптомики, позволяющие получать комплексную информацию о состоянии отдельных клеток [1, 3]. Увеличение числа одноклеточных исследований сопровождается быстрым ростом объёма генерируемых данных. Один эксперимент может включать десятки тысяч и более клеток, для каждой из которых определяются тысячи молекулярных

признаков. Кроме того, к основным матрицам данных добавляются метаданные об образцах, донорах, клеточных типах, экспериментальных условиях и параметрах обработки. В результате возникает новая проблема: наличие большого количества данных не гарантирует их эффективного использования. Для современной науки важна не только возможность сохранить данные, но и возможность обнаружить их, получить доступ, понять контекст получения, сопоставить с другими исследованиями и повторно использовать. Поэтому проблема доступности баз одноклеточных омиксных данных становится самостоятельной научной и информационной задачей [4].

1. Рост объёма одноклеточных омиксных данных

Одноклеточные эксперименты характеризуются высокой размерностью и структурной сложностью. В случае scRNA-seq результатом исследования обычно является матрица, в которой строки соответствуют генам, а столбцы — отдельным клеткам. Дополнительно могут храниться показатели качества секвенирования, результаты кластеризации, координаты клеток и информация об их принадлежности к определённым клеточным типам [2]. При использовании нескольких омиксных модальностей объём информации существенно возрастает. Например, в одном исследовании могут одновременно анализироваться экспрессия генов, доступность хроматина и поверхностные белки. Интеграция подобных данных требует хранения не только самих матриц, но и информации о соответствии между различными модальностями [5]. Большой объём данных создаёт инфраструктурные проблемы. Сырые файлы секвенирования могут занимать значительное дисковое

пространство, а их передача и повторная обработка требуют соответствующих вычислительных ресурсов. При этом обработанные матрицы, хотя и имеют меньший размер, зависят от применённого алгоритма анализа. Таким образом, развитие одноклеточных технологий приводит не только к росту количества научных данных, но и к необходимости создания специализированной инфраструктуры для их хранения, описания и повторного использования.

2. Фрагментация существующих баз данных

Одной из основных проблем доступности одноклеточных данных является их распределение между различными репозиториями и специализированными информационными системами. Сырые данные секвенирования могут размещаться в крупных международных архивах, тогда как обработанные матрицы и метаданные могут предоставляться через специализированные платформы или как дополнительные материалы к научным публикациям. В результате исследователю нередко приходится использовать несколько информационных ресурсов для получения полного набора сведений об одном эксперименте. Фрагментация данных приводит к нескольким проблемам:

- увеличивается время поиска необходимых наборов данных;
- усложняется сопоставление исследований;
- повышается вероятность потери информации о происхождении данных;
- затрудняется автоматизированная обработка;
- усложняется построение интегрированных клеточных атласов.

Определённую роль в решении этой проблемы играют агрегирующие платформы. Например, CELLxGENE предоставляет инструменты

для поиска и визуализации большого количества одноклеточных наборов данных, а CELLxGENE Census использует унифицированную структуру для представления клеток, генов и связанных с ними метаданных [6, 7]. Особое значение имеет Human Cell Atlas (HCA), представляющий собой международную инициативу по созданию комплексной карты клеток человеческого организма [8]. Развитие подобных проектов демонстрирует переход от отдельных наборов данных к интегрированным информационным ресурсам.

3. Проблема стандартизации форматов

Одноклеточные данные могут распространяться в различных файловых форматах и структурах. В биоинформатике используются, в частности, HDF5, AnnData, Loom, CSV и другие форматы. Само существование нескольких форматов не является критической проблемой. Основная трудность возникает при отсутствии единых правил представления биологической и экспериментальной информации. Например, один и тот же клеточный тип может иметь разные обозначения в различных наборах данных. Клетки могут быть описаны как «T cell», «T lymphocyte» или более специфично — как определённый подтип Т-лимфоцитов. Без использования стандартизированных терминов автоматическое сопоставление таких объектов становится затруднительным. Аналогичные проблемы возникают при описании тканей, заболеваний, экспериментальных условий и других характеристик. Поэтому для обеспечения интероперабельности необходимо использовать стандартизированные онтологии и контролируемые словари [4, 9].

Таким образом, стандартизация должна охватывать не только файловые форматы, но и содержание данных.

4. Недостаточная стандартизация метаданных

Метаданные являются одним из важнейших компонентов одноклеточных баз. Они позволяют определить происхождение данных, характеристики биологического материала, используемый экспериментальный протокол и параметры последующей обработки. Для полноценного повторного использования набора данных желательно иметь информацию о: виде организма; типе ткани; возрасте и поле донора; состоянии здоровья; типе заболевания; способе получения клеток; используемой экспериментальной технологии; платформе секвенирования; количестве исследованных клеток; параметрах контроля качества; методах нормализации и фильтрации; способе определения клеточных типов. Недостаток метаданных существенно ограничивает возможности вторичного анализа. Исследователь может получить матрицу экспрессии, но без информации об условиях эксперимента не сможет корректно оценить причины обнаруженных различий. Проблема стандартизации метаданных особенно важна при интеграции нескольких исследований. Различия в названиях клеточных типов, тканей и экспериментальных условий могут препятствовать автоматическому объединению данных. Для решения этой задачи разрабатываются специализированные стандарты метаданных. В частности, Human Cell Atlas Metadata Standard предназначен для унификации описания данных и повышения их интероперабельности [9].

5. Различия экспериментальных протоколов

Одноклеточные данные зависят не только от биологических характеристик исследуемого материала, но и от технических параметров эксперимента. На результаты могут влиять способ выделения клеток, время до фиксации или секвенирования, используемая платформа, глубина секвенирования и другие параметры [2]. По этой причине два исследования, выполненные на сходных биологических образцах, могут демонстрировать различия, связанные не с биологическими процессами, а с особенностями экспериментальных протоколов. Особенно существенной становится проблема batch effect — эффекта партии. Он представляет собой систематическое различие между группами данных, обусловленное техническими факторами. При объединении нескольких одноклеточных исследований batch effect может приводить к формированию искусственных кластеров и искажать биологическую интерпретацию [10]. Для устранения подобных эффектов используются различные алгоритмы интеграции данных. Однако корректность их применения зависит от полноты метаданных и понимания экспериментального происхождения каждого набора. Следовательно, база данных должна хранить подробную информацию об экспериментальном дизайне, позволяющую оценить возможные источники технической вариабельности.

6. Доступность сырых и обработанных данных

Существенной проблемой является соотношение между сырыми и обработанными данными. Сырые данные позволяют исследователю самостоятельно выполнять весь pipeline анализа и использовать современные методы обработки. Однако их объём может быть значительным, что увеличивает требования к инфраструктуре

хранения и передачи. Обработанные данные удобнее для непосредственного анализа, но зависят от выбранного авторами pipeline. Изменение параметров фильтрации, нормализации или аннотации клеток может привести к получению других результатов. Поэтому наиболее эффективным является сохранение как исходных данных, так и обработанных результатов вместе с информацией о применённых методах. Такой подход повышает воспроизводимость исследования и позволяет повторно анализировать данные с использованием новых биоинформатических инструментов [4, 11]. Крупные международные проекты используют комбинированный подход. В частности, инфраструктура Human Cell Atlas предусматривает хранение обработанных данных и метаданных в открытых ресурсах, при этом доступ к некоторым исходным данным может регулироваться с учётом требований конфиденциальности [8].

7. Проблема воспроизводимости

Доступность набора данных сама по себе не гарантирует воспроизводимость исследования. Для повторения анализа необходимо знать последовательность применённых процедур: фильтрацию клеток, нормализацию, выбор признаков, снижение размерности, кластеризацию, аннотацию клеток и статистические методы [11]. Особое значение имеет сохранение информации о программном обеспечении и версиях используемых инструментов. Изменение алгоритма или его параметров может приводить к отличиям в результатах даже при использовании одного и того же исходного набора данных. Важным аспектом является также клеточная аннотация. Определение клеточного типа может выполняться на основе маркерных генов, автоматизированных

методов или сопоставления с референсными атласами. Если способ аннотации не зафиксирован, последующее сравнение результатов становится затруднительным. Таким образом, современная база одноклеточных данных должна рассматриваться не только как хранилище файлов, но и как система хранения контекста, происхождения и истории обработки данных.

8. Доступность данных и защита конфиденциальности

Особую проблему представляют одноклеточные данные человека. Несмотря на то что матрица экспрессии не всегда содержит прямые идентификаторы личности, исходные геномные данные могут быть связаны с конкретными донорами. По этой причине некоторые наборы данных предоставляются в рамках контролируемого доступа. Исследователю может потребоваться регистрация, подтверждение исследовательских целей или выполнение определённых условий использования. С одной стороны, подобные ограничения необходимы для защиты участников исследований. С другой стороны, они усложняют автоматизированный поиск и массовое использование данных. Следовательно, современная инфраструктура должна обеспечивать баланс между принципом открытости научной информации и требованиями биомедицинской этики и защиты данных.

9. FAIR-принципы как основа повышения доступности данных

Одним из наиболее перспективных подходов к организации научных данных являются FAIR-принципы, сформулированные Wilkinson и соавторами [12].

Аббревиатура FAIR включает четыре основных требования: **Findable** — данные должны быть обнаруживаемыми; **Accessible** —

данные должны быть доступными; **Interoperable** — данные должны быть интероперабельными; **Reusable** — данные должны быть пригодными для повторного использования.

Для одноклеточных данных реализация FAIR-принципов предполагает использование устойчивых идентификаторов, стандартизированных метаданных, машиночитаемых форматов, подробного описания происхождения данных и прозрачных условий их использования. Особенно важным является принцип интероперабельности. Он предполагает возможность использования данных совместно с другими информационными системами без необходимости ручной переработки каждого набора. FAIR-подход позволяет перейти от модели «архива файлов» к модели научной инфраструктуры, в которой данные становятся частью взаимосвязанной экосистемы.

10. Интегрированные клеточные атласы как направление развития

Одним из перспективных направлений развития одноклеточных баз является создание интегрированных клеточных атласов. Human Cell Atlas является одним из наиболее масштабных международных проектов, направленных на систематическое описание клеточного состава человеческого организма [8]. Подобные инициативы позволяют объединять данные множества исследовательских групп и создавать более полные представления о клеточных состояниях. Другим примером является CELLxGENE Census, предоставляющий унифицированный доступ к большим коллекциям одноклеточных данных [7]. Интегрированные атласы позволяют исследователю решать задачи, которые трудно выполнить

на основе одного эксперимента. Например, можно сравнивать характеристики одного клеточного типа между различными тканями, возрастными группами или патологическими состояниями. Однако создание таких ресурсов требует решения ряда технических задач: стандартизации метаданных, устранения batch effects, унификации клеточных аннотаций и контроля качества данных [10].

Заключение

Проблематика доступности баз одноклеточных омиксных данных имеет комплексный характер и включает ряд взаимосвязанных аспектов: хранение больших объёмов информации, фрагментацию репозитория, неоднородность форматов, недостаточную стандартизацию метаданных, различия экспериментальных протоколов, проблемы воспроизводимости и ограничения, связанные с защитой данных человека. Современные одноклеточные исследования генерируют огромные объёмы информации, однако научная ценность этих данных определяется не только их количеством, но и возможностью их повторного использования. Если исследователь не может найти необходимый набор, определить его происхождение, понять структуру метаданных или корректно сопоставить его с другими исследованиями, потенциальная ценность данных существенно снижается. Одним из ключевых направлений решения проблемы является внедрение FAIR-принципов, предусматривающих обнаруживаемость, доступность, интероперабельность и повторное использование данных [12]. Не менее важными являются стандартизация метаданных, развитие общих онтологий и создание унифицированных форматов. Перспективным направлением

является развитие интегрированных клеточных атласов, объединяющих данные множества независимых исследований. Такие системы позволяют перейти от хранения отдельных наборов данных к формированию единой информационной инфраструктуры одноклеточной биологии. Таким образом, дальнейшее развитие одноклеточной омиксики должно сопровождаться параллельным развитием инфраструктуры данных. Только при условии стандартизации, документирования и обеспечения доступности результатов исследований накопленные данные смогут эффективно использоваться для фундаментальных исследований, диагностики заболеваний, разработки новых методов лечения и развития персонализированной медицины.

Список литературы

1. Regev A., Teichmann S. A., Lander E. S. et al. The Human Cell Atlas // **eLife**. — 2017. — Vol. 6. — Art. e27041. — DOI: 10.7554/eLife.27041.
2. Lafzi A., Moutinho C., Picelli S., Heyn H. Tutorial: guidelines for the experimental design of single-cell RNA sequencing studies // **Nature Protocols**. — 2018. — Vol. 13, № 12. — P. 2742–2757. — DOI: 10.1038/s41596-018-0073-y.
3. Hao Y., Hao S., Andersen-Nissen E. et al. Integrated analysis of multimodal single-cell data // **Cell**. — 2021. — Vol. 184, № 13. — P. 3573–3587.e29. — DOI: 10.1016/j.cell.2021.04.048.
4. Zappia L., Theis F. J. Over 1000 tools for single-cell analysis: the single-cell landscape // **Nature Reviews Genetics**. — 2021. — Vol. 22. — P. 145–161. — DOI: 10.1038/s41576-020-00414-z.

5. Stuart T., Satija R. Integrative single-cell analysis // **Nature Reviews Genetics**. — 2019. — Vol. 20. — P. 257–272. — DOI: 10.1038/s41576-019-0093-7.
6. Chan Zuckerberg Initiative. CELLxGENE Discover [Электронный ресурс]. — URL: <https://cellxgene.cziscience.com/> (дата обращения: 11.09.2026).
7. Chan Zuckerberg Initiative. CELLxGENE Census: Data and schema [Электронный ресурс]. — URL: <https://chanzuckerberg.github.io/cellxgene-census/> (дата обращения: 11.09.2026).
8. Human Cell Atlas. Data Portal [Электронный ресурс]. — URL: <https://data.humancellatlas.org/> (дата обращения: 11.09.2026).
9. Human Cell Atlas. HCA Metadata Standard [Электронный ресурс]. — URL: https://ebi-ait.github.io/hca-metadata-community/metadata_standard/metadata-standard.html (дата обращения: 11.09.2026).
10. Tran H. T. N., Ang K. S., Chevrier M. et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data // **Genome Biology**. — 2020. — Vol. 21. — Art. 12. — DOI: 10.1186/s13059-019-1850-9.
11. Heumos L., Schaar A. C., Lance C. et al. Best practices for single-cell analysis across modalities // **Nature Reviews Genetics**. — 2023. — Vol. 24. — P. 550–572. — DOI: 10.1038/s41576-023-00586-w.
12. Wilkinson M. D., Dumontier M., Aalbersberg I. J. et al. The FAIR Guiding Principles for scientific data management and stewardship // **Scientific Data**. — 2016. — Vol. 3. — Art. 160018. — DOI: 10.1038/sdata.2016.18.