



ПРОБЛЕМАТИКА ДОСТУПНОСТИ БАЗ ОДНОКЛЕТОЧНЫХ ОМИКСНЫХ ДАННЫХ ЧЕЛОВЕКА

Мусакаев Руслан

Ташкентский Государственный технический
университет

Аспирант (PHD)

Email: marasres08121999@mail.ru

Аннотация: Стремительное развитие технологий одноклеточного секвенирования привело к накоплению значительных объемов омиксных данных человека, характеризующих молекулярные особенности отдельных клеток. Базы одноклеточных данных играют ключевую роль в хранении, обработке и повторном использовании результатов исследований. Вместе с тем доступность таких данных сопровождается рядом проблем, включая фрагментацию информационных ресурсов, отсутствие единых стандартов метаданных, ограничения доступа, связанные с защитой персональных данных, а также высокие требования к вычислительной инфраструктуре. В статье рассматриваются основные проблемы доступности баз одноклеточных омиксных данных человека, анализируются существующие международные инициативы по их решению и определяются перспективные направления развития биоинформационных платформ. Особое внимание уделяется принципам FAIR, вопросам интероперабельности данных и повышению эффективности международного научного сотрудничества. [1; 4]

Ключевые слова: одноклеточные омические данные, биоинформатика, базы данных, геномика человека, FAIR-принципы, открытые данные, стандартизация данных.

Abstract: *The rapid development of single-cell sequencing technologies has resulted in the accumulation of vast amounts of human omics data that characterize the molecular features of individual cells. Single-cell data repositories play a crucial role in the storage, processing, and reuse of research results. However, the accessibility of these databases is associated with several challenges, including fragmentation of information resources, lack of standardized metadata, restricted access due to personal data protection requirements, and high computational infrastructure demands. This article examines the main issues related to the accessibility of human single-cell omics databases, analyzes existing international initiatives aimed at addressing these challenges, and identifies promising directions for the development of bioinformatics platforms. Particular attention is paid to FAIR principles, data interoperability, and the enhancement of international scientific collaboration. [1; 4]*

Keywords: *single-cell omics data, bioinformatics, databases, human genomics, FAIR principles, open data, data standardization.*

Annotatsiya: *Bir hujayrali sekvenslash texnologiyalarining jadal rivojlanishi inson omiks ma'lumotlarining katta hajmda to'planishiga olib keldi, bu ma'lumotlar alohida hujayralarning molekulyar xususiyatlarini tavsiflaydi. Bir hujayrali ma'lumotlar bazalari tadqiqot natijalarini saqlash, qayta ishlash va ulardan qayta foydalanishda muhim ahamiyatga ega. Shu bilan birga, bunday ma'lumotlarning ochiqligi va foydalanish imkoniyati axborot resurslarining tarqoqligi,*

metama'lumotlarning yagona standartlari mavjud emasligi, shaxsiy ma'lumotlarni himoya qilish bilan bog'liq cheklovlar hamda yuqori hisoblash infratuzilmasi talablariga duch kelmoqda. Mazkur maqolada insonning bir hujayrali omiks ma'lumotlari bazalaridan foydalanishdagi asosiy muammolar ko'rib chiqiladi, ularni hal etishga qaratilgan xalqaro tashabbuslar tahlil qilinadi va bioinformatika platformalarini rivojlantirishning istiqbolli yo'nalishlari aniqlanadi. Tadqiqotda FAIR tamoyillari, ma'lumotlarning interoperabelligi va xalqaro ilmiy hamkorlik samaradorligini oshirish masalalariga alohida e'tibor qaratilgan.

Kalit so'zlar: *bir hujayrali omiks ma'lumotlari, bioinformatika, ma'lumotlar bazalari, inson genomikasi, FAIR tamoyillari, ochiq ma'lumotlar, ma'lumotlarni standartlashtirish. [1; 4]*

Введение

Развитие высокопроизводительных технологий секвенирования нового поколения (Next Generation Sequencing, NGS) существенно изменило современные подходы к изучению биологических систем. Особое значение приобрели методы одноклеточного анализа, позволяющие исследовать молекулярные характеристики отдельных клеток и выявлять клеточную гетерогенность, ранее недоступную при использовании традиционных методов массового секвенирования [1].

Современные технологии single-cell RNA sequencing (scRNA-seq), single-cell ATAC sequencing (scATAC-seq), пространственной транскриптомики и мультимодальных омиксных исследований обеспечивают получение огромных объемов данных. Для хранения и распространения таких данных созданы специализированные базы,

включая Human Cell Atlas, Gene Expression Omnibus, European Genome-phenome Archive и Single Cell Expression Atlas [2; 5–7].

Несмотря на активное развитие биоинформационной инфраструктуры, доступность одноклеточных омиксных данных остается одной из наиболее актуальных проблем современной биоинформатики и цифровой медицины.

1. Особенности одноклеточных омиксных данных

Одноклеточные омиксные данные представляют собой результаты анализа отдельных клеток на различных молекулярных уровнях: геномном, транскриптомном, эпигеномном, протеомном и метаболомном [1].

В отличие от традиционных bulk-исследований, одноклеточные технологии позволяют выявлять редкие клеточные популяции, отслеживать процессы клеточной дифференцировки и исследовать механизмы развития заболеваний на уровне отдельных клеток [2].

Особенностью данных данного типа являются высокая размерность, разреженность матриц экспрессии, значительный уровень технического шума и необходимость сложной предварительной обработки [1; 3]. Указанные особенности оказывают существенное влияние на способы хранения и распространения данных.

2. Основные проблемы доступности баз одноклеточных омиксных данных

2.1 Фрагментация информационных ресурсов

Одной из наиболее значимых проблем является распределение данных между многочисленными репозиториями и специализированными платформами. Различные исследовательские

группы публикуют результаты в независимых хранилищах, использующих собственные стандарты организации информации [1].

В результате исследователям приходится осуществлять поиск данных одновременно в нескольких системах, что увеличивает временные затраты и снижает воспроизводимость исследований [4].

2.2 Недостаточная стандартизация метаданных

Метаданные играют ключевую роль в интерпретации биологических результатов. Они включают сведения о биологическом материале, методах подготовки образцов, используемых платформах секвенирования и параметрах биоинформационного анализа.

Отсутствие унифицированных стандартов описания приводит к тому, что данные различных проектов оказываются несовместимыми между собой [4].

Несогласованность терминологии особенно заметна при аннотации типов клеток, тканей и патологических состояний [2].

2.3 Ограничения доступа, связанные с защитой персональных данных

Одноклеточные данные человека зачастую содержат сведения, потенциально позволяющие идентифицировать пациента посредством генетической информации [7].

По этой причине многие наборы данных размещаются в репозиториях контролируемого доступа. Для получения доступа исследователи обязаны проходить процедуры согласования, включающие подтверждение научных целей исследования, подписание соглашений об использовании данных и соблюдение требований по защите персональной информации [7].

Подобные ограничения необходимы с точки зрения биоэтики, однако они значительно усложняют международное научное сотрудничество.

2.4 Высокие требования к вычислительной инфраструктуре

Размер современных одноклеточных наборов данных может достигать десятков и сотен гигабайт. Для их анализа необходимы высокопроизводительные серверы, облачные вычислительные платформы и специализированное программное обеспечение [1].

Во многих научных организациях, особенно в развивающихся странах, доступ к подобной инфраструктуре ограничен.

2.5 Проблемы качества данных

Доступность данных не гарантирует их пригодность для дальнейшего использования.

Наиболее распространенными проблемами являются неполная аннотация, ошибки классификации клеток, различия в протоколах подготовки образцов, пакетные эффекты (batch effects) и отсутствие информации о процедурах контроля качества [3].

Это приводит к снижению достоверности результатов вторичного анализа и затрудняет воспроизводимость исследований [1].

3. Международные инициативы по повышению доступности данных

В последние годы реализуется ряд международных проектов, направленных на создание единого информационного пространства для хранения и обмена одноклеточными данными.

Особую роль играет Human Cell Atlas, целью которого является создание полного клеточного атласа человеческого организма [2].

Среди ключевых направлений развития можно выделить:

1. внедрение FAIR-принципов (Findable, Accessible, Interoperable, Reusable) [4];
2. развитие облачных платформ хранения данных;
3. унификацию биомедицинских онтологий;
4. автоматизацию аннотации клеточных типов;
5. разработку стандартов обмена данными между репозиториями [1].

Применение данных подходов способствует повышению воспроизводимости исследований и ускоряет внедрение результатов фундаментальной науки в клиническую практику [2; 4].

4. Перспективы совершенствования баз одноклеточных омиксных данных

Для решения существующих проблем представляется целесообразным развитие следующих направлений.

Технологические меры

Создание распределенных облачных инфраструктур, внедрение технологий искусственного интеллекта для автоматической аннотации данных и разработка унифицированных форматов хранения мультимодальных данных [1; 3].

Организационные меры

Расширение международного сотрудничества, формирование единых стандартов публикации данных и создание национальных центров хранения биомедицинской информации [2; 4].

Правовые меры

Совершенствование механизмов деперсонализации данных, гармонизация национальных нормативных требований и разработка международных соглашений по обмену геномной информацией [7].

Заключение

Одноклеточные омиксные технологии являются одним из наиболее динамично развивающихся направлений современной биомедицины. Однако эффективность их использования во многом определяется доступностью соответствующих баз данных. В настоящее время основными проблемами остаются фрагментация информационных ресурсов, недостаточная стандартизация метаданных, ограничения доступа, связанные с защитой персональных данных, высокие требования к вычислительным ресурсам и неоднородное качество данных [1; 3; 7].

Перспективы дальнейшего развития связаны с внедрением FAIR-принципов, развитием международных инициатив по интеграции биоинформационных ресурсов и созданием единых стандартов хранения и обмена данными [2; 4]. Решение указанных задач позволит обеспечить более эффективное использование накопленных биомедицинских данных и ускорить развитие персонализированной медицины.

Список литературы

1. Lähnemann Dennis, Köster Johannes и др. Eleven grand challenges in single-cell data science // *Genome Biology*. 2020. Т. 21. № 1. С. 31.
2. Regev Aviv, Teichmann Sarah A., Lander Eric S. и др. The Human Cell Atlas // *eLife*. 2017. Vol. 6. e27041.
3. Stuart Tim, Butler Andrew и др. Comprehensive Integration of Single-Cell Data // *Cell*. 2019. Vol. 177. № 7. P. 1888–1902.

4. Wilkinson Mark D., Dumontier Michel и др. The FAIR Guiding Principles for scientific data management and stewardship // Scientific Data. 2016. Vol. 3. Article 160018.
5. National Center for Biotechnology Information. Gene Expression Omnibus (GEO): Data Repository for Functional Genomics.
6. European Molecular Biology Laboratory. Single Cell Expression Atlas.
7. European Genome-phenome Archive. Controlled Access Human Genomic Data Repository