На правах рукописи 

Менделевич Ася Владимировна

Статистические вопросы, связанные с техническими и биологическими вариациями, возникающие при аллель-специфическом анализе данных секвенирования

Специальность 1.5.8. —
«математическая биология, биоинформатика»

Автореферат

диссертации на соискание учёной степени
кандидата физико-математических наук

Москва — 2023

Работа выполнена в Сколковском институте науки и технологий

Научный руководитель: доктор биологических наук, профессор
Гельфанд Михаил Сергеевич

Официальные оппоненты: **Лагарькова Мария Андреевна,**
член-корреспондент РАН, доктор биологических наук, профессор РАН,
Федеральное государственное бюджетное учреждение «Федеральный научно-клинический центр физико-химической медицины имени академика Ю.М. Лопухина Федерального медико-биологического агентства»,
генеральный директор

Раменский Василий Евгеньевич,
кандидат физико-математических наук,
Федеральное государственное бюджетное учреждение «Национальный медицинский исследовательский центр терапии и профилактической медицины» Министерства здравоохранения Российской Федерации,
руководитель лаборатории геномной и медицинской биоинформатики

Ведущая организация: Федеральный исследовательский центр «Фундаментальные основы биотехнологии» Российской Академии Наук

Защита состоится 25 сентября 2023 г. в 17:00 на заседании диссертационного совета 24.1.101.01 на базе Федеральное государственное бюджетное учреждение науки Институт проблем передачи информации им. А.А. Харкевича Российской академии наук (ИППИ РАН) по адресу: 127051, г. Москва, Большой Каретный переулок, д.19 стр. 1.

С диссертацией можно ознакомиться в библиотеке ФГБУ науки Института проблем передачи информации им. А.А. Харкевича Российской академии наук (ИППИ РАН), а также на сайте ИППИ РАН.

Автореферат разослан _____ 2023 года.

Ученый секретарь
диссертационного совета 24.1.101.01,
доктор биологических наук Казенников Олег Васильевич

Общая характеристика работы

Актуальность темы. Понимание источников шума в экспериментах необходимо для точного количественного анализа и интерпретации данных. В данных секвенирования существует множество источников вариации. Наибольший интерес представляют биологические источники вариации, включающие в себя генетические и эпигенетические различия внутри тканей и между ними, клональность клеточных популяций или гетерогенность клеток, а также биологический шум, такой как транскрипционные всплески и связанные с ними явления [1]. В то же время, любое экспериментальное измерение имеет сопутствующий шум, накапливающийся из-за обработки экспериментальных и вычислительных данных, выборки и множества других неучтенных факторов. Отделение этого технического шума от биологической вариации имеет фундаментальное значение для понимания природы данных [2—4], что делает использование соответствующих статистических методов основной мерой защиты от ложных открытий. Однако, как подробно описано в разделе обзора литературы, тщательным анализом свойств шума в экспериментах по секвенированию часто пренебрегают, в частности, в случае анализа аллель-специфической экспрессии (ASE). Ярким примером является широко распространенное применение биномиального теста для оценки технического шума в данных высокопроизводительного секвенирования в исследованиях ASE [5], при том, что в тех же работах авторы показывают, что это приводит к существенной недооценке технического шума [6—8]. Также наблюдается выраженное противоречие между желанием максимально полно использовать чрезвычайно дорогие и крупномасштабные наборы данных и ограничениями, заложенными в этих данных. В некоторых случаях авторы прибегают к попыткам обойти эти ограничения, в результате нарушая предположения, лежащие в основе стандартных статистических методов, и не учитывая это должным образом в последующем анализе. Например, в главе 2 приводится пример использования чтений, которые не покрывают ни одного однонуклеотидного полиморфизма (SNP) в аллель-специфическом анализе [9; 10]. В более общем виде, та же проблема изменения распределений относится и к обычным методам нормализации, что говорит о том, что их использование не всегда является корректным.

Целью данной работы является разработка метода для точного количественного анализа дифференциальной аллель-специфической экспрессии.

Для достижения поставленной цели необходимо было решить следующие **задачи**:

1. Изучить то, насколько существующие подходы справляются с задачей оценки аллельного дисбаланса и дифференциальной ASE;
2. Определить количество технических реплик, необходимых для измерения уровня технического шума;
3. Оценить влияние технического шума на воспроизводимость получаемых результатов;
4. Разработать вычислительные инструменты для измерения и учёта технического шума, для проведения точного количественного анализа данных ASE;
5. Применить разработанные методы для изучения эпигенетического митотически стабильного механизма ДНК-метилирования;
6. Разработать экспериментальные протоколы и адаптировать инструменты для проведения анализа данных ASE экономичным и экспериментально масштабируемым способом.

Основные положения, выносимые на защиту:

1. Одной библиотеки РНК-секвенирования недостаточно для надёжной оценки вклада технического шума в наблюдаемый сигнал ASE. Для оценки и учёта технической избыточной дисперсии в количественных и дифференциальных задачах ASE на данных РНК-секвенирования был разработан вычислительный подход, опирающийся на анализ различий в оценках AI между техническими репликами. Метод был реализован в виде R-пакета `Qllelic`.
2. Гены с моноаллельной аутосомной экспрессией (MAE) демонстрируют митотически стабильный выбор аллелей, приводящий к устойчивым транскрипционным различиям между клональными клеточными линиями, при этом механизм MAE, во многих случаях, неизвестен. Использование новой стратегии скрининга с помощью секвенирования позволило обнаружить ключевую роль метилирования ДНК в поддержании MAE. Полногеномный анализ показал, что MAE является частью более общего механизма регуляции генов, и обнаружил ранее недооцененное взаимодействие гене-

тического и эпигенетического контроля аллель-специфической транскрипции. В то время как цис-регуляция определяет общее базовое состояние для всех генетически идентичных клеток, метилирование ДНК выполняет роль аллель-специфического реостата и определяет множество регуляторных состояний, различающихся между клональными клеточными линиями.

3. Применение внешних РНК-контролей в экспериментах с большим количеством образцов, позволяет решить вопрос оценки избыточного шума в аллельном дисбалансе с не меньшей точностью, чем доступна при технической репликации, однако с существенно меньшей стоимостью (около 5-10% против минимум двухкратного увеличения в случае приготовления двух или более библиотек для каждого образца). Новый метод был реализован в виде R-пакета `ControlFreq` и включает в себя функционал работы с техническими репликами, в качестве специального случая.

Научная новизна:

1. Было показано, что, вопреки распространенности соответствующих практик, техническая компонента избыточной дисперсии неотделима от биологического разнообразия без технической репликации или другого технического контроля, что привело к необходимости разработки новых подходов для точной количественной оценки аллель-специфической экспрессии.
2. Более того, было показано, что вопрос "сколько необходимо реплик" менее важен, чем вопрос о том, как должны обрабатываться данные из таких реплик для правильного измерения и учета шума в данных.
3. Был предложен новый экспериментальный дизайн, который позволяет проводить точный количественный анализ данных ASE экономичным и масштабируемым способом.
4. С помощью разработанных методов было показано, что метилирование ДНК является ключевым механизмом для митотически стабильного поддержания моноаллельной аутосомной экспрессии (MAE). Кроме того, были исследованы полногеномные эффекты применения ингибитора метилтрансферазы 5-аза-2'-деоксицитидина (5-аза-dC) на различных клеточных линиях.

Научная и практическая значимость. Полученные в диссертации результаты подтверждают, что корректный учёт технической избыточной

дисперсии позволяет существенно повысить воспроизводимость при работе с аллель-разрешёнными данными РНК-секвенирования и избежать завышенного уровня ложноположительных результатов. Следование предложенным протоколам для экспериментальной и вычислительной обработки данных позволяет достигнуть большей статистической корректности, а в случае РНК-контролей это и не требует существенного увеличения затрат на эксперимент. Разработанный метод может быть полезен во многих транскриптомных исследованиях, и может стимулировать разработку аналогичных протоколов при работе с другими типами данных, таких как длинноридное или одноклеточное РНК-секвенирование, и в смежных областях, таких как эпигенетика и организация хроматина.

Степень достоверности и апробация результатов. Результаты работы были представлены на следующих международных конференциях и научных семинарах:

- ИТиС (Информационные технологии и системы), Иннополис, Россия, 26–30 сентября 2018, постер
- 3я ежегодная Skoltech-MIT конференция (Collaborative Solutions for Next Generation Education, Science and Technology), Москва, Россия, 15–16 октября 2018, постер
- RECOMB (Research in Computational Molecular Biology), Вашингтон, США, 4–8 мая 2019, доклад на RECOMB Genetics: «Accurate estimation of transcriptome-wide differential allelic expression»
- ISMB/ECCB (Intelligent Systems For Molecular Biology / European Conference On Computational Biology), Базель, Швейцария, 21–25 июля 2019, постер
- MCCMB (Moscow Conference on Computational Molecular Biology), Москва, Россия, 27–30 июля 2019, постер
- ИТиС (Информационные технологии и системы), Пермь, Россия, 18–19 сентября 2019, постер
- Семинар программы Variant To Function, Broad Institute, Бостон, США, 22 октября 2019, доклад: «Unexpected variability of allelic imbalance estimates from RNA sequencing»
- Выездной семинар кафедры Генетики Гарвардской Медицинской Школы, Genetics Retreat, Бостон, США, 23–24 февраля 2020, постер

– Семинар кафедры Генетики Гарвардской Медицинской Школы, Data club, Бостон, США, 10 июля 2020, доклад: «Unexpected variability of allelic imbalance estimates from RNA sequencing»
– ISMB/ECCB (Intelligent Systems For Molecular Biology / European Conference On Computational Biology), Лион, Франция, 23–27 июля 2023, (принятый) доклад: «Foreign RNA spike-ins enable accurate allele-specific expression analysis at scale»

Публикации. По теме диссертации опубликовано 3 статьи в рецензируемых международных научных журналах, входящих в основные библиометрические базы данных (PubMed, WoS и Scopus):

1. Replicate sequencing libraries are important for quantification of allelic imbalance / **Asia Mendelevich**, Svetlana Vinogradova, Saumya Gupta, Andrey A. Mironov, Shamil R. Sunyaev, Alexander A. Gimelbrant // *Nature Communications* — 2021 — DOI:10.1038/s41467-021-23544-8
2. RNA sequencing-based screen for reactivation of silenced alleles of autosomal genes / Saumya Gupta, Denis L Lafontaine, Sebastien Vigneau, **Asia Mendelevich**, Svetlana Vinogradova, Kyomi J Igarashi, Andrew Bortvin, Clara F Alves-Pereira, Anwesha Nag, Alexander A Gimelbrant // *G3 Genes|Genomes|Genetics* — 2022 — DOI:10.1093/g3journal/jkab428
3. Foreign RNA spike-ins enable accurate allele-specific expression analysis at scale / **Asia Mendelevich**, Saumya Gupta, Aleksei Pakharev, Athanasios Teodosiadis, Andrey A. Mironov, Alexander A. Gimelbrant // *Bioinformatics (ISMB/ECCB issue)* — 2023 — DOI:10.1093/bioinformatics/btad254

Личный вклад:

Публикация 1. Соискатель сделала основной вклад в формулировку задачи, разработку статистического метода, написание кода, обработку и систематизацию полученных данных, и написании текста статьи.

Публикация 2 (и препринт Gupta et al., 2020). Существенная часть обработки данных была выполнена соискателем: соискатель написала пайплайн для обработки RNA-seq данных, обработки дифференциальной ASE, написала большую часть скриптов для постобработки и отрисовки данных. Соискатель участвовала в планировании эксперимента и интерпретации результатов в работе, включенной в главу 3, а также в написании и редактировании текстов статей Gupta et al., 2020 и Gupta et al., 2021.

Публикация 3. В равной доле с научным руководителем, соискатель спланировала проект и эксперименты. Соискатель разработала статистический метод, внесла основной вклад в написание кода, обработала и систематизировала полученные данные, и внесла основной вклад в написание текста статьи.
Соискатель была основным *автором кода* во всех упомянутых в диссертации (4) репозиториях. Если в тексте диссертации не указано обратного, то соискатель является исключительным автором итоговой версии упомянутых библиотек и программ. Весь код находится в открытом доступе с permissive licenses (MIT / GPLv3) в проекте лаборатории Gimelbrant Lab на GitHub.

Объем и структура работы. Диссертация состоит из введения, четырёх глав, заключения и трёх приложений. Полный объём диссертации составляет **174** страницы с **52** рисунками и **4** таблицами. Список литературы содержит **125** наименований.

Содержание работы

Во **введении** обосновывается актуальность исследований, проводимых в рамках данной диссертационной работы, формулируется цель, ставятся задачи работы, формулируется научная новизна и практическая значимость представляемой работы.

В первой главе рассмотрена история развития области исследования аллельного дисбаланса и представлен обзор научной литературы по известным механизмам поддержания аллельного дисбаланса в транскрипции, исследованиям цис-регуляции, и методам анализа аллель-специфической экспрессии. Отдельное внимание уделяется вопросам измерения технического шума в данных РНК-секвенирования (Рис. 1) и отделения его от биологической вариабельности.

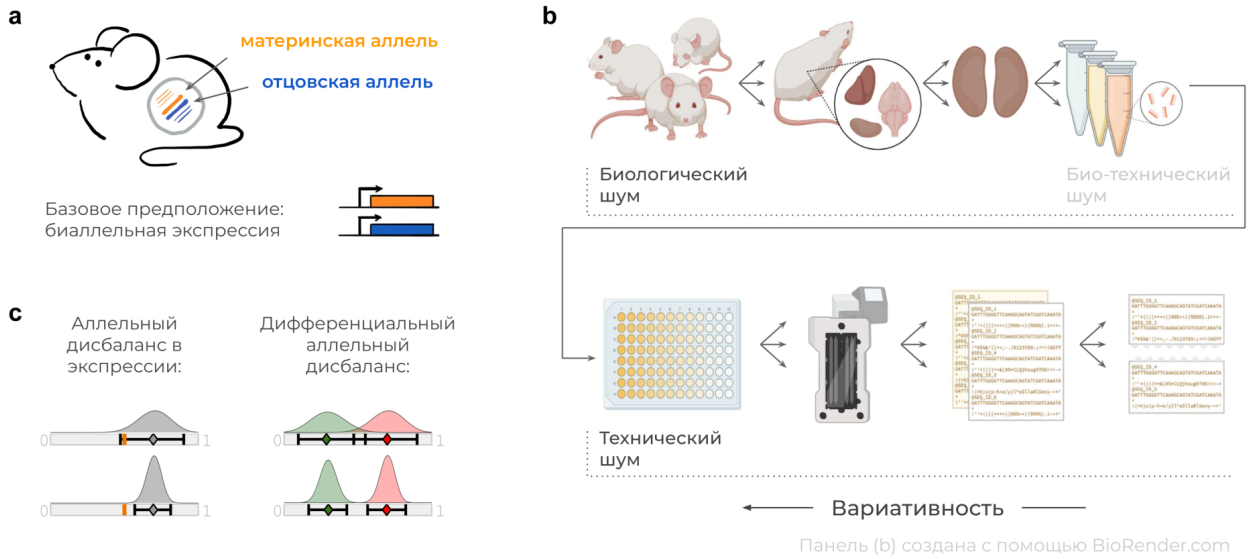

Рис. 1 — **Накопление экспериментального шума в процессе производства данных РНК-секвенирования.** **(a)** Схематическое изображение родительских аллелей и уровня транскрипции. **(b)** Степень различности двух образцов зависит от точки их разделения. **(c)** Две наиболее типичных задачи в области аллель-специфической экспрессии, и схематическое изображение влияния уровня шума на результаты статистических тестов: меньший уровень избыточной дисперсии позволяет видеть более слабый сигнал.

Вторая глава посвящена разработке вычислительного подхода для учёта технического шума при количественной оценке аллельного дисбаланса, основанного на обработке данных технической репликации. Метод `Qllelic` реализован в качестве `R`-пакета и находится в открытом доступе.

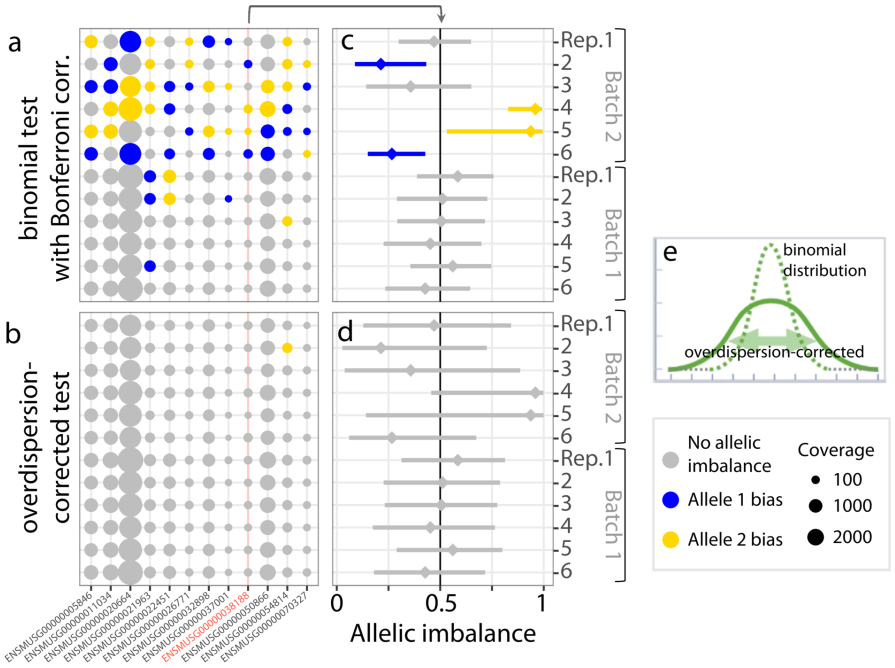

Рис. 2 — **Аллель-специфический сигнал в данных РНК-секвенирования может существенно изменяться под влиянием технического шума. (a)** Показаны все гены с аллельным покрытием >10 и демонстрирующие противоречивый аллельный дисбаланс по результатам биномиального теста; серый — отсутствие AI, жёлтый — значительное преобладание материнской аллели, синий — отцовской. **(b)** Те же данные, что и в (a), но обработанные модифицированным тестом, учитывающим избыточную дисперсию. **(c-d)** Значения AI для одного из генов (*ENSMUSG00000038188*, аллельное покрытие 110 ± 45). Ромб – точечная оценка AI, отрезок — доверительный интервал, определённый с помощью соответствующей модели. **(e)** Схематичное изображение супер-биномиальной дисперсии (непрерывная линия) в сравнении с биномиальной (пунктирная).
Данные: из (Mendelevich et al., 2021), «Batch 1» и «Batch 2» – реплики из эксп. 2 и 3 (SMART-Seq, количество исходной тотальной РНК: 10 нг и 100 нг, в пределах рекомендуемого диапазона).

В данной главе мы сначала демонстрируем наличие эксперимент-специфической избыточной дисперсии в оценках аллель-специфической экспрессии (ASE) и обсуждаем невозможность оценки технического шума из одной технической реплики. Далее мы показываем, что в аллельных данных поли-А РНК-секвенирования избыточная дисперсия имеет мультипликативную природу, и одинакова для разных участков генома и уровней покрытия генов. Это позволяет нам вычислять меру избыточной дисперсии, коэффициент коррекции качества (QCC), сравнивая попарно технические реплики и оценивая, насколько экспериментально наблюдаемые квантили распределений ΔAI отличаются от ожидаемых значений в предположении биномиальной модели (Рис. 4). Также это даёт возможность применять простую коррекцию к биномиальным тестам, используя покрытие, в QCC^2 раз меньшее наблюдаемого:

$$\widetilde{\text{Bin}}\left(\frac{m}{\text{QCC}^2} \;\middle|\; \frac{m+p}{\text{QCC}^2},\ \text{AI}\right),$$

а использование аппроксимирующих пропорциональных тестов позволяет решить проблему нецелых значений. Наконец, мы показываем, что использование поправки на избыточную дисперсию с помощью `Qllelic` существенно повышает воспроизводимость результатов при анализе аллель-специфической экспрессии (Рис. 2) и открывает возможность проводить точный дифференциальный анализ аллельного дисбаланса (AI). Также в этой главе мы исследуем источники эксперимент-специфической избыточной дисперсии в экспериментах РНК-секвенирования.

Данная глава основана на статье (Mendelevich et al., 2021).

В **третьей главе** показывается, что метилирование ДНК является одним из ключевых механизмов поддержания митотически стабильной моноаллельной аутосомной экспрессии (MAE).

С помощью разработанного аппарата `Qllelic`, мы изучаем полногеномное воздействие ингибитора метилтрансферазы, 5-аза-2'-дезоксицитидина (5-aza-dC), на экспрессию в различных клональных линиях. Мы показываем, что подмножество генов демонстрирует сходимость к общему значению AI после воздействия эквитоксичных концентраций 5-aza-dC, что указывает на то, что метилирование ДНК играет роль "аллельного реостата" и определяет непрерывное множество стабильных состояний ASE. В то же время, другое подмножество генов демонстрирует резистенстность к воздействию 5-aza-dC, что согласуется с наличием других источников регуляции транскрипции.

Данная глава основана на статьях (Gupta et al., 2022) и (Gupta et al., 2020, препринт на bioRxiv).

В **четвертой главе** приведено описание разработанного экспериментального протокола и усовершенствованных вычислительных инструментов, позволяющих производить экономный и масштабируемый анализ данных ASE. Представленный экспериментальный метод базируется на добавлении внешней РНК перед приготовлением библиотек образцов, выступающей в роли контрольной компоненты для оценки технического шума и тем избавляющей от необходимости производства технических реплик.

Новый разработанный аппарат `ControlFreq` позволяет оценивать уровень избыточной дисперсии, iQCC, как в случае РНК-контролей, так и в слу-

чае небольшого количества технических реплик образцов. Чтобы избавиться от существовавших ранее ограничений на данные и получить возможность интегрировать информацию из всего набора образцов при подборе значений iQCC, мы реализовали модифицированную модель урны Полья, перейдя от бета-биномиальных распределений, которые могут быть выражены через материнское и отеческое покрытия гена, m и p, аллельный дисбаланс, AI, и избыточную вариацию, $Q = \text{iQCC}^2$:

$$BB(m \mid n = m + p,\ \text{AI},\ \text{Q}) =$$

$$= BB\left(m \mid n = m + p,\ \alpha = \text{AI} \cdot \frac{n - Q}{Q - 1},\ \beta = (1 - \text{AI}) \cdot \frac{n - Q}{Q - 1}\right),$$

к расширенным бета-биномиальным распределениям:

$$eBB(m \mid n = m + p,\ \text{AI},\ Q) =$$

$$= eBB\left(m \mid n = m + p,\ \alpha = \text{AI},\ \beta = 1 - \text{AI},\ d = \frac{Q - 1}{n - Q}\right).$$

Так как данное обобщённое распределение не зависит от одновременного масштабирования параметров, иначе говоря,

$$\forall x > 0 \quad eBB(m \mid n = m + p, \alpha x, \beta x, dx) = eBB(m \mid n = m + p, \alpha, \beta, d),$$

семейство распределений разбивается на три случая: наличие избыточной ($d = 1$) или уменьшенной ($d = -1$) дисперсии по сравнению с биномиальной моделью, и, собственно, сама биномиальная модель ($d = 0$). В случае небольшого числа технических реплик, возможность корректно вычислять нижнюю и верхнюю оценки меры избыточной дисперсии позволяет оценивать iQCC как их геометрическое среднее (Рис. 9).

Мы также продемонстрировали устойчивость метода к варьированию параметров, включая пропорцию контрольной РНК, распределение истинных пропорций ASE, и выбор контрольного образца (или смеси двух линий для симуляции диплоидности). Как и `Qllelic`, R-пакет `ControlFreq` находится в открытом доступе на GitHub.

Данная глава основана на статье (Mendelevich et al., 2023).

В **заключении** приведены основные результаты работы, которые заключаются в следующем:

Разработка метода точной количественной оценки аллельного дисбаланса при наличии технических реплик. Мы продемонстрировали, что данных из одной библиотеки РНК-секвенирования недостаточно для надежной количественной оценки вклада технического шума в наблюдаемый сигнал AI (Рис. 3). Для учёта избыточной дисперсии и точной оценки ASE в данных РНК-секвенирования мы разработали вычислительный подход, опирающийся на попарные сравнения технических реплик (библиотек из одной пробы РНК), и реализовали его в `R`-пакете `Qllelic` (`github.com/gimelbrantlab/Qllelic`). Этот подход концептуально прост: он эквивалентен биномиальному тестированию, однако наблюдаемое покрытие рассматривается меньшим в QCC^2 раз, где QCC — коэффициент коррекции качества, рассчитанный с помощью `Qllelic` (Рис. 4). Проводить такую поправку позволяет наблюдение, что избыточная дисперсия имеет одинаковую мультипликативную природу на каждом участке генома, при анализе данных поли-А РНК-секвенирования. Коррекция на избыточную дисперсию с помощью QCC продемонстрировала существенное снижение количества ложноположительных результатов, возникающих из-за технического шума (Рис. 5ab), в сравнении с широко используемым биномиальным тестом и методами, оценивающими избыточную дисперсию из одной реплики [7; 11—15].

Важно отметить, что использование `Qllelic` позволяет проводить надежный дифференциальный анализ аллель-специфической экспрессии и сравнивать образцы из разных экспериментов, если значения QCC могут быть вычислены для всех участвующих в сравнении образцов (Рис. 5).

Мы также показали, что, наиболее вероятно, наибольший вклад в избыточную дисперсию вносит процесс приготовления библиотек, в то время как сам процесс секвенирования в наших экспериментах оказывал незначительное влияние. Важно, что вычислительная дедупликация прочтений не избавляет образцы от избыточной дисперсии. В дополнение к более очевидным систематическим различиям между протоколами для приготовления библиотек, различия между экспериментами, проведенными по одному и тому же протоколу, могут быть также значительными — поэтому рекомендуется иметь по крайней мере две технических реплики для каждого образца.

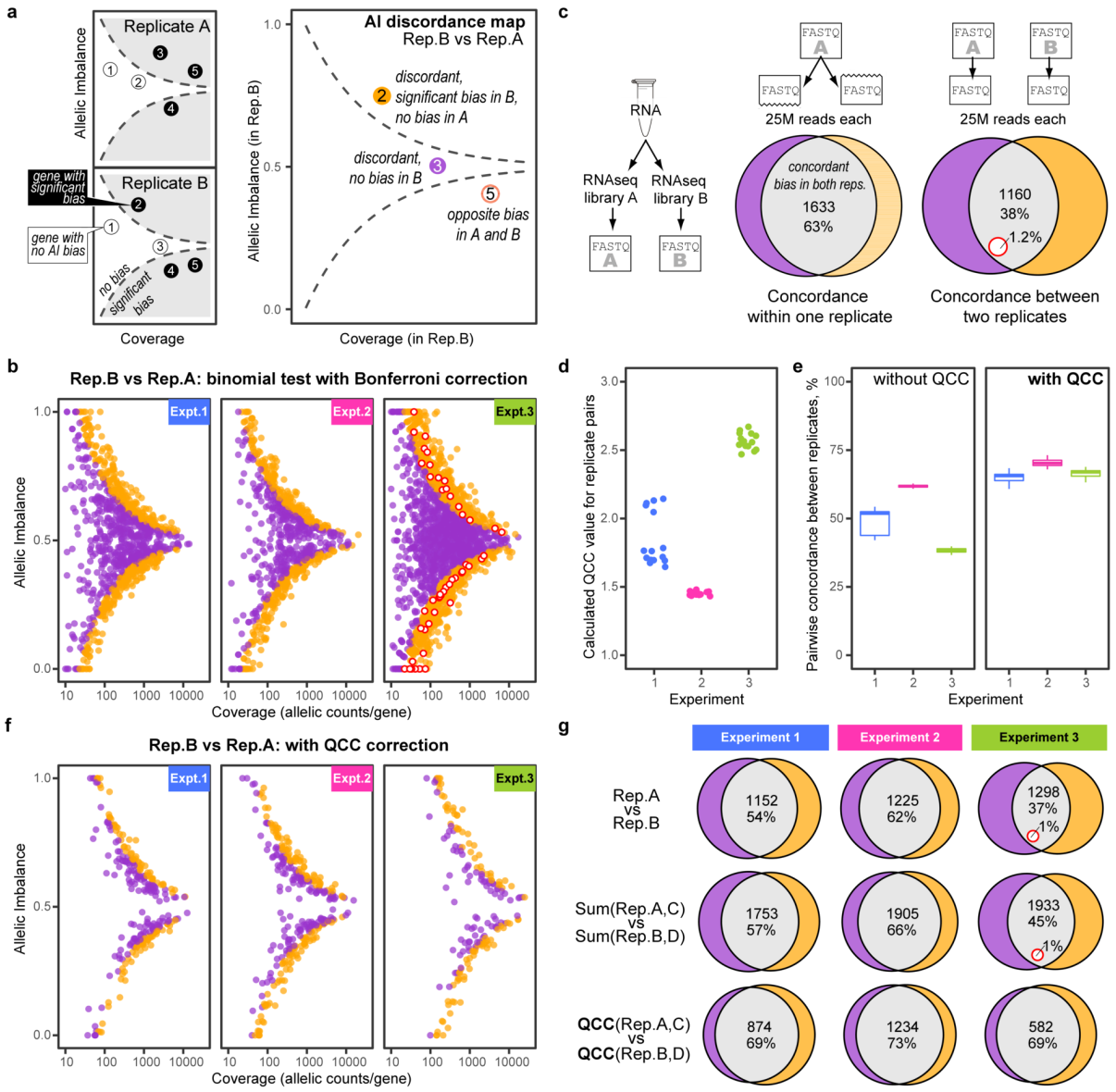

Рис. 3 — **Значения аллельного дисбаланса не совпадают для разных технических реплик и экспериментов РНК-секвенирования.** **(a)** Объяснение графика рассогласования оценок аллельного дисбаланса: оранжевый — нет смещения в реплике A, есть смещение в B; пурпурный — есть смещение в реплике A, нет смещения в B; красный незаполненный круг — смещение в обоих репликах, но в противоположных направлениях. **(b)** Графики рассогласованности AI для репрезентативных пар технических реплик во всех трёх экспериментах. **(c)** Диаграммы Эйлера для генов со значительным смещением аллельной экспрессии, при сравнении двух технических реплик или образцов, полученных выборкой без возвращения из данных одной библиотеки. Данные: реплики 1 и 2 из Эксп. 3. **(d)** Коэффициент коррекции качества (QCC), мера избыточной дисперсии, для всех 15 пар реплик в каждом из Экспериментов 1 (синий), 2 (красный), или 3 (зелёный). Заметим общую консистентность значений QCC внутри экспериментов, и чувствительность к одной реплике-выбросу в Эксп. 1 (соответствует пяти парам-выбросам). **(e)** Доля согласованно смещённых генов [сравните с серой областью в (c)] для всех 15 пар реплик в Эксп. 1–3, до и после коррекции на QCC. **(f)** То же, что и в (b), но гипотеза H_0 тестировалась при помощи пропорционального теста, скорректированного на QCC. **(g)** Применение QCC увеличивает согласованность между репликами и между экспериментами. Данные: из (Mendelevich et al., 2021).

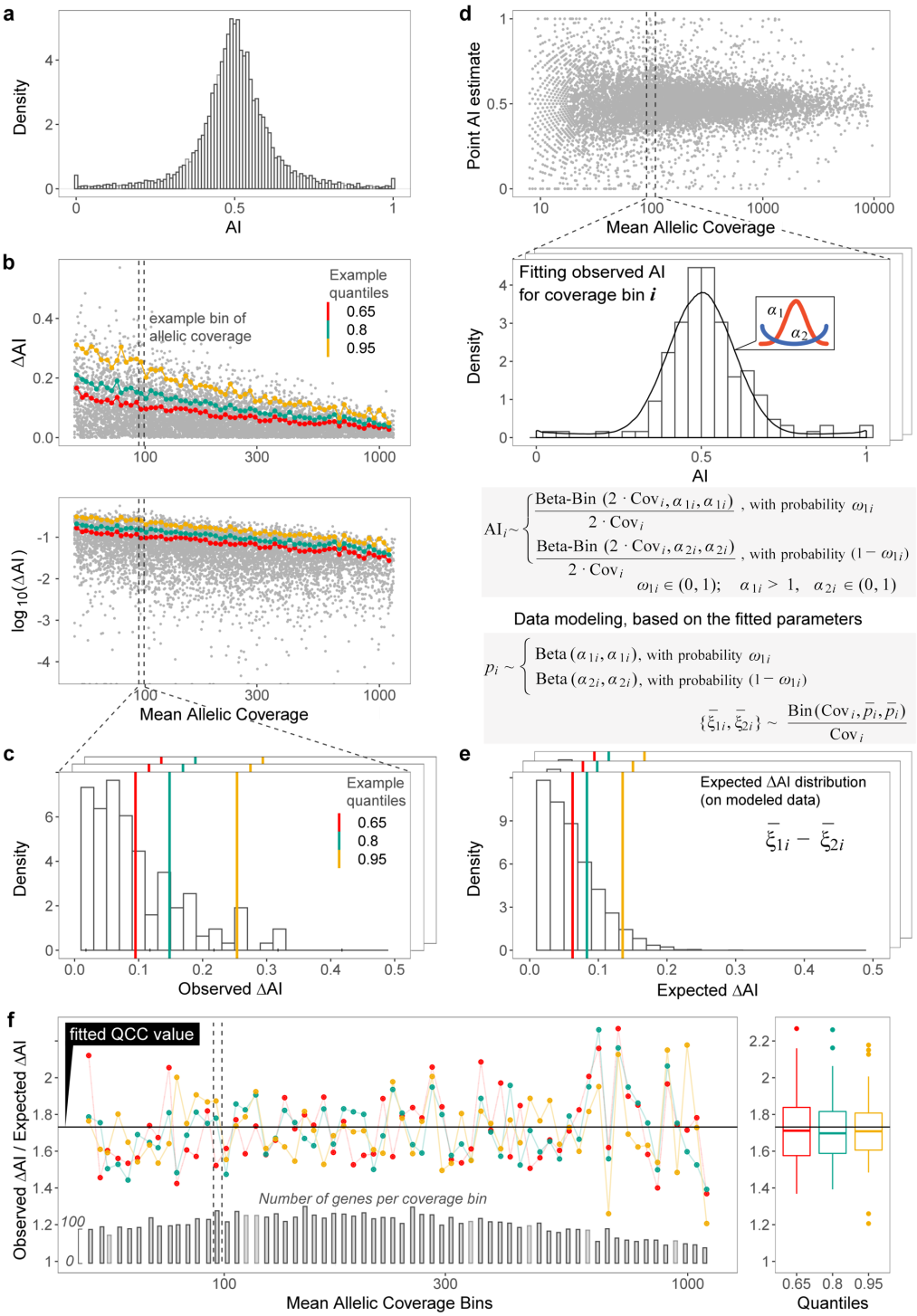

Рис. 4 — **Вывод коэффициента коррекции качества (QCC) из наблюдаемых и смоделированных разностей AI между техническими репликами. (a)** Распределение точечных оценок AI для генов с аллельным покрытием более 10 в шести объединённых репликах (180 млн прочтений суммарно) Эксп. 2. **(b-c)** Вычисление квантилей наблюдаемых распределений абсолютных разностей между аллельными дисбалансами в двух репликах (ΔAI). Гены разбиты на корзины по логарифму покрытия, показан пример корзины. Здесь и далее, показаны три фиксированных квантиля: 0.65 (красным), 0.8 (зелёным) и 0.95 (оранжевым). **(d-e)** Вычисление квантилей ожидаемых распределений ΔAI. **(f)** Частные наблюдаемого и ожидаемого значений для фиксированных квантилей ΔAI. Подобранная константа (черная линия) определяет QCC.

Данные: из (Mendelevich et al., 2021).

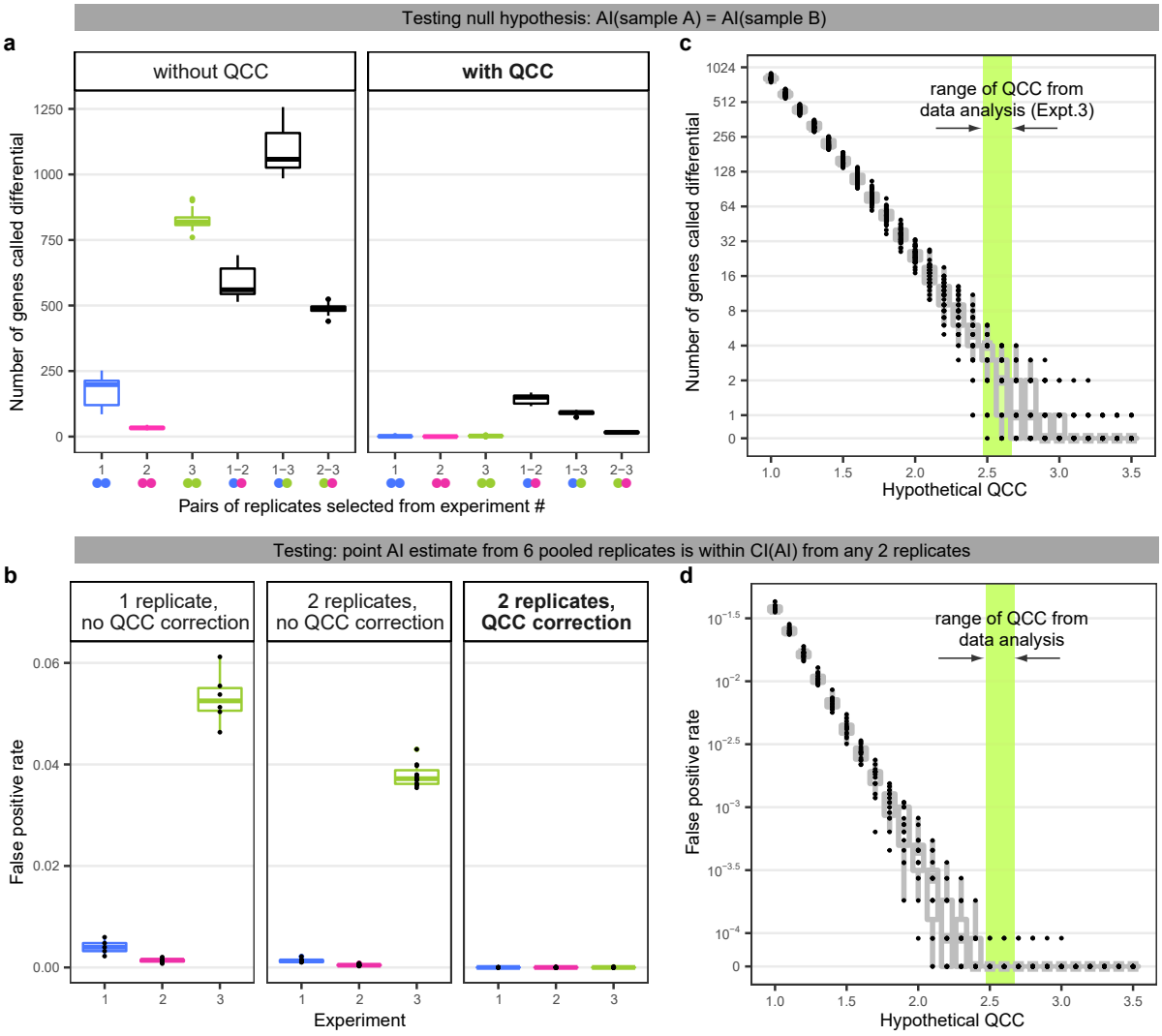

Рис. 5 — **QCC позволяет совершать дифференциальный анализ AI, и находится в прямой зависимости с избыточной дисперсией покрытий. a:** Количество генов с ложноположительной дифференциальной аллельной экспрессией, до (слева) и после коррекции на QCC (справа). **b:** Влияние QCC на долю ложноположительных существенных отличий точечной оценки AI от оценки по всем репликам: “золотой стандарт” оценки AI не лежит в доверительном интервале одной реплики (слева), двух объединённых реплик (посередине), и двух реплик с коррекцией на QCC (справа). **c, d:** Вычисленные значения QCC (все возможные сочетания реплик из Эксп.3, содержатся внутри окрашенных рамок) близки к оптимальному балансу между излишним количеством ложноположительных результатов и потерей сигнала.
Данные: из (Mendelevich et al., 2021).

Моноаллельная аутосомная экспрессия и эпигенетическая цис-регуляция. С помощью метода скрининга секвенированием было показано, что метилирование ДНК является ключевым механизмом, вовлеченным в митотически стабильное поддержание моноаллельной экспрессии в клональных лимфоидных клеточных линиях млекопитающих. Мы предложили простую модель (Рис. 7fg), в которой аллель-специфический регуляторный ландшафт определяется генетической вариацией, в то время как конкретное состояние ASE в популяции клональных клеток зависит от метилирования ДНК. Отметим, что эта модель предполагает наличие специфических цис-регуляторных элементов вблизи затронутых генов. Такие геномные элементы могли бы дать простое объяснение эволюционной консервации MAE генов в человеческих популяциях [16] и между разными организмами, например, человеком и мышью [17—19]. Важно отметить, что деметилирование ДНК влияет не на все MAE гены (Рис. 6ab), что позволяет предположить, что поддержание MAE для некоторых локусов включает другие механизмы в дополнение к (или вместо) метилированию ДНК. Это могло бы объяснить, почему влияние деметилирования ДНК на аллельный дисбаланс не было ранее обнаружено для определённых MAE генов [20; 21].

Используя `Qllelic`, мы оценили влияние полногеномного деметилирования ДНК на транскрипционный аллельный дисбаланс (Рис. 6). В четырех проанализированных клонах мы обнаружили значительные сдвиги AI в более чем 600 аутосомных генах (Рис. 7a-b). Точный количественный анализ данных РНК-секвенирования выявил более сложный ландшафт клонального разнообразия в аллель-специфической регуляции генов, чем подразумевает моноаллельная/биаллельная дихотомия (Рис. 6c, 7c-g). В разных клонах аллельный дисбаланс гена может охватывать диапазон значений от 0 до 1, и эпигенетическая регуляция действует как более тонко настраиваемый механизм контроля, чем переключатель “вкл/выкл”.

Наконец, наши результаты свидетельствуют о том, что ингибиторы ДНК-метилтрансферазы, вероятно, влияют на регуляцию генов у пациентов таким образом, который трудно обнаружить без аллель-специфического анализа. Кроме того, мы отмечаем, что значительные сдвиги в AI часто не связаны с существенным изменением в суммарном покрытии гена, следовательно, тщательный количественный анализ аллель-специфической регуляции генов

в поликлональных и моноклональных популяциях клеток может привести к новым клинически значимым открытиям.

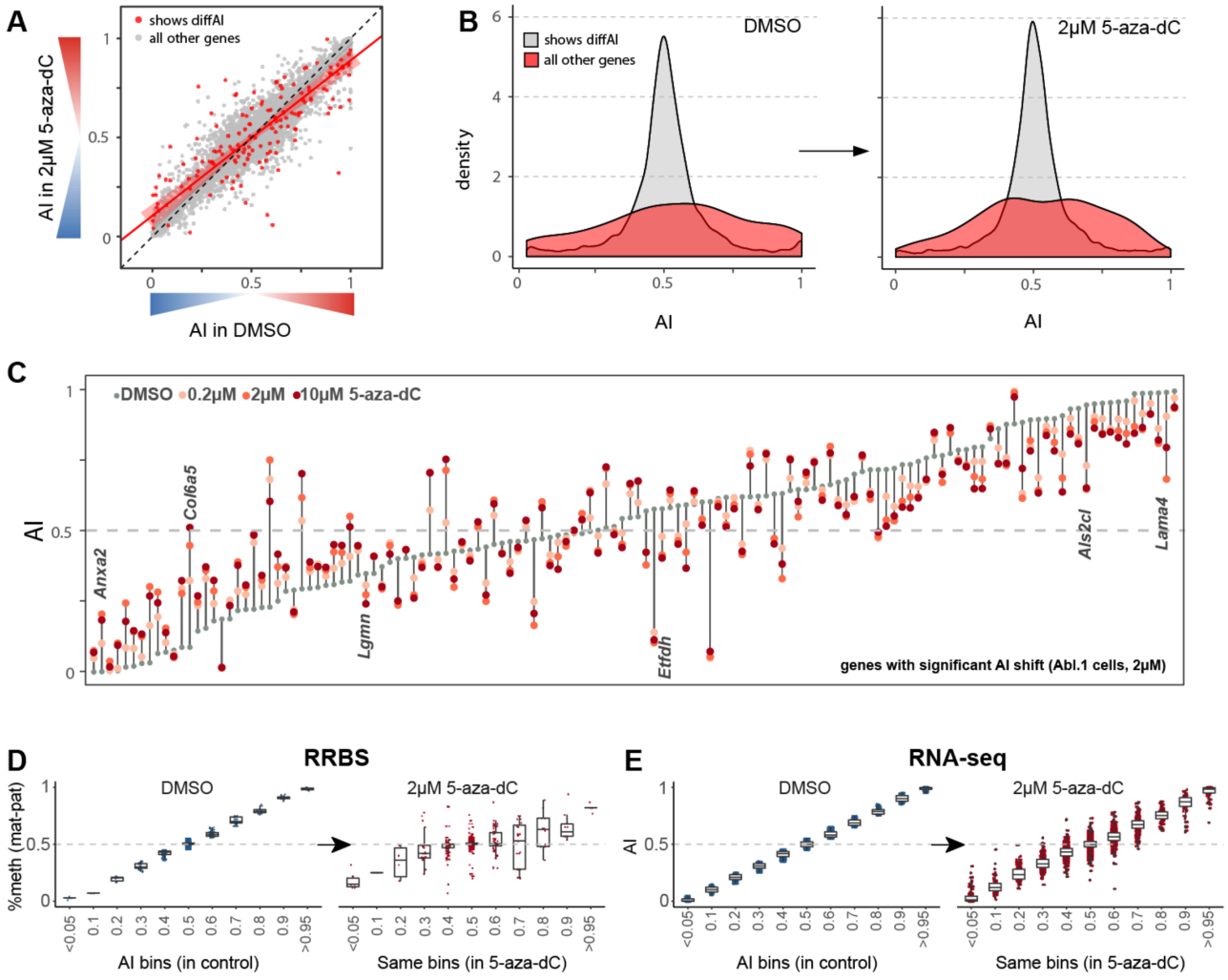

Рис. 6 — **Влияние препарата 5-aza-dC на аллель-специфическую экспрессию в масштабе всего генома. (a)** Полногеномное сопоставление аллельного дисбаланса генов к клетках Abl.1, между контрольными образцами (x) и обработанных 2 μM 5-aza-dC (y). Красные точки – гены, показавшие существенный сдвиг в ASE, серые – остальные. Красная прямая с 95% доверительным интервалом — многомерная линейная регрессия без предикторов. **(b)** Распределения значений AI для генов без существенных изменений (серым) и имеющих дифференциальный AI между контролем и обработкой 2 μM 5-aza-dC (красным), данные как в (a). **(c)** Сдвиги в ASE для генов с дифференциальным AI, данные как в (a,b). Различные цвета кодируют концетрации 5-aza-dC. **(d-e)** Аллель-специфические изменения в ДНК метиломе (RBBS) и транскриптоме (RNA-seq) в клетках Abl.1. Гены были разбиты по группам $X \pm 0.05$ согласно (d) разнице в пропорциях метилированных CpGs на материнской и отцовской аллелях или (e) их значениям AI, в образцах с 1% ДМСО.

Данные: из (Gupta et al., 2021; Gupta et al., 2020).

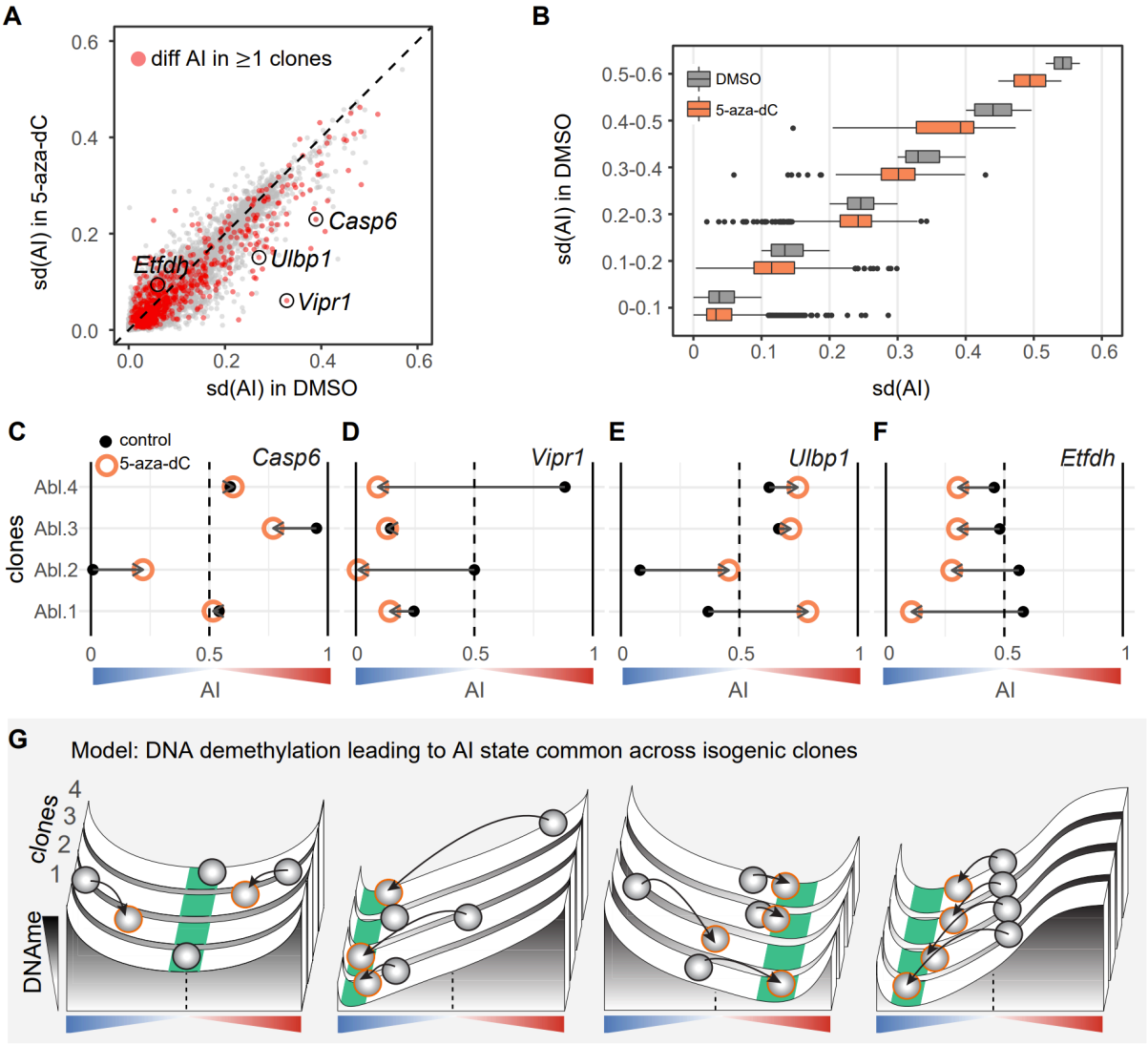

Рис. 7 — **Деметилирование ДНК приводит к большему сходству между клонами в аллель-специфической транскрипции.** **(a)** Сопоставление стандартных отклонений значений AI между четырьмя клонами (Abl.1: 1% ДМСО и 2 μM 5-aza-dC, Abl.2-4: 1% ДМСО и 0.2 μM 5-aza-dC, эквитоксичные дозировки) в контрольном образце (x) после обработки 5-aza-dC (y). Красным обозначены гены с дифференциальной ASE хотя бы в одном из клонов. **(b)** Разница в дисперсии для генов, разделённых на имеющие и не имеющие дифференциальную ASE, как в (a). Гены сгруппированны по значениям дисперсии в контрольных образцах. **(c-f)** Примеры генов с существенными изменениями в ASE и разными типами сдвигов в AI. Чёрные точки соответствуют ДМСО, оранжевые окружности – 5-aza-dC. **(g)** Концептуальная диаграмма контроля ASE через клон-специфичное метилирование ДНК при общем для всех клонов генетическом регуляторном ландшафте. Диаграммы соответствуют примерам из (c-f), чёрный и оранжевый цвета шаров кодируют ту же информацию, зелёная зона – точка сходимости, “общий” аллельный дисбаланс в экспрессии. Данные: из (Gupta et al., 2021; Gupta et al., 2020).

Внешние РНК-контроли позволяют провести точный аллель-специфический анализ в масштабных экспериментах РНК-секвенирования. Мы разработали экспериментальный и вычислительный подход с добавлением внешних РНК-контролей, выполняющих роль технической компоненты в экспериментах с большим количеством образцов. Использование РНК-контролей позволяет не делать технические реплики для каждого образца и требует увеличения стоимости эксперимента всего на $5-10\%$ по сравнению со стандартным методом без технической репликации. Помимо уменьшения дополнительной стоимости эксперимента, новый подход позволяет обойти другие ограничения: не требует сопоставимого общего аллельного покрытия у сравниваемых образцов, не ожидает симметричности в истинном AI, и оценивает избыточную дисперсию индивидуально для каждого образца. Возможность обрабатывать образцы с разным уровнем шума, среди прочего, существенно облегчает поиск статистических выбросов. Всё вышеперечисленное существенно расширяет применимость метода и делает его более устойчивым.

Подход с внешними РНК-контролями также очень устойчив к варьированию параметров эксперимента. Образец для РНК-контроля должен быть одинаковым только внутри одной группы образцов: как только оценки избыточной дисперсии произведены для всех образцов группы, эти образцы могут быть использованы для сравнительного анализа между собой или с образцами из других наборов данных.

Мы установили следующие требования к используемому в качестве РНК-контроля организму и к выбору протокола экспериментальной обработки данных: (1) прочтения с "основного" и контрольного образца должны быть различимы при выравнивании, (2) плотность полиморфизмов в контрольном организме должна быть достаточной для эффективного разрешения прочтений по аллелям, (3) контрольная РНК должна быть способна пройти тот же процесс подготовки библиотеки, что и основной образец. Ограничение на процент контрольной РНК в смеси отсутствует, и необходимый процент определяется тотальным объёмом секвенирования — в случае значительного размера секвенируемой библиотеки, может быть достаточно $5-10\%$ добавленной внешней РНК. Кроме того, нами было установлено, что смеси РНК гомозиготных линий *C.elegans* выполняют роль "синтетического гибрида", не требуя

при этом соблюдения пропорции 1:1. Это существенно упрощает получение образцов для РНК-контролей, так как не требует скрещивания организмов для получения гибридов. Мы полагаем, что возможно использование РНК дрожжей или создание стандартизованного набора синтетических молекул для анализа аллель-специфической экспрессии.

Новый вычислительный метод оценки избыточной дисперсии реализован в виде R-пакета `controlFreq` (github.com/gimelbrantlab/controlFreq), и может быть использован как в случае технической репликации, так и в случае РНК-контролей.

В данной работе мы использовали данные поли-А РНК-секвенирования, однако предложенные методы можно экстраполировать и на другие экспериментальные протоколы (например, создание библиотек с удалением рРНК) и типы данных. Среди очевидных целей для дальнейшего развития метода — данные одноклеточного РНК-секвенирования и длиннoридного РНК-секвенирования. Более того, мы не ограничены только работой с транскриптомными данными. Задачами для разработки похожих методов могут стать изучение открытости хроматина, ДНК метилирования или ДНК-белковых взаимодействий.

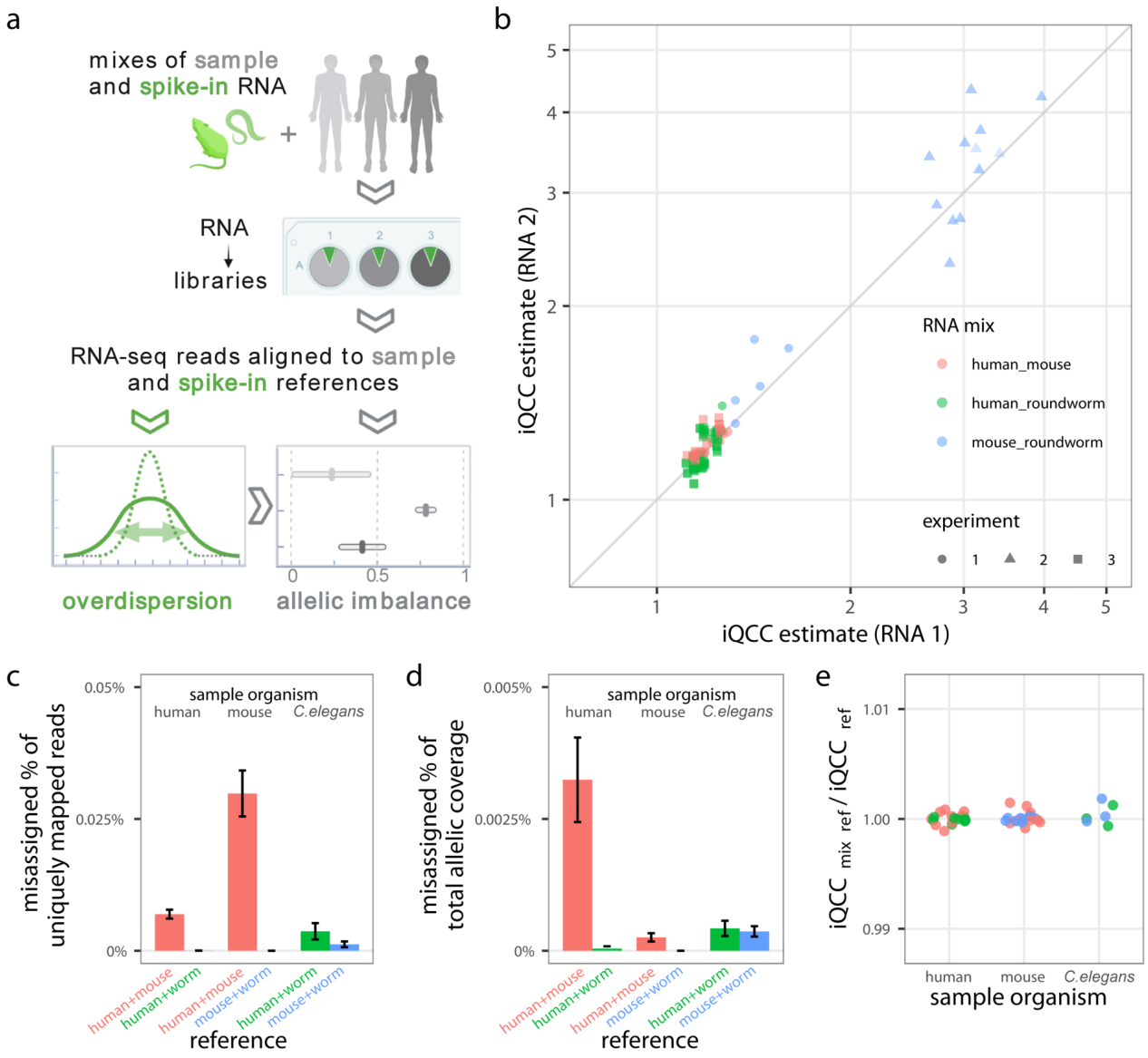

Рис. 8 — **В библиотеке, состоящей из РНК двух различных организмов, избыточные аллельные дисперсии для обоих организмов близки.** **(a)** Диаграмма экспериментальных и вычислительных шагов в алгоритме оценки аллельного дисбаланса. Внешняя РНК (зелёный) добавлена в основной образец (серый) до подготовки библиотеки РНК-секвенирования. **(b)** В рамках одной библиотеки РНК-секвенирования, измеренная избыточная дисперсия iQCC крайне схожа для обоих компонент РНК (коэффициент корреляции Пирсона 0.97). **(c-e)** Оценка степени потери данных из-за выравнивания на неправильный организм прочтений из смешанных библиотеках РНК-секвенирования. Данные: 3 биологических реплики человека, 3 — мыши, 1 — *C.elegans* (и 3 технических реплики на каждую), выравненные на индивидуальные или химерные референсы. Цвет представляет смесь референсов использованную для выравнивания, кодирование пар совпадает с кодированием смесей в (b). **(c-d)** Процент неверно выравненных прочтений среди (c) всех уникально выравненных прочтений, (b) аллель-разрешённых прочтений. **(e)** Сравнение значений iQCC вычисленных при выравнивании на одиночный или смешанный (химерный) референс. Всего было найдено 0 генов с дифференциальным аллельным дисбалансом для каждой возможной пары.

Данные: из (Mendelevich et al., 2023).

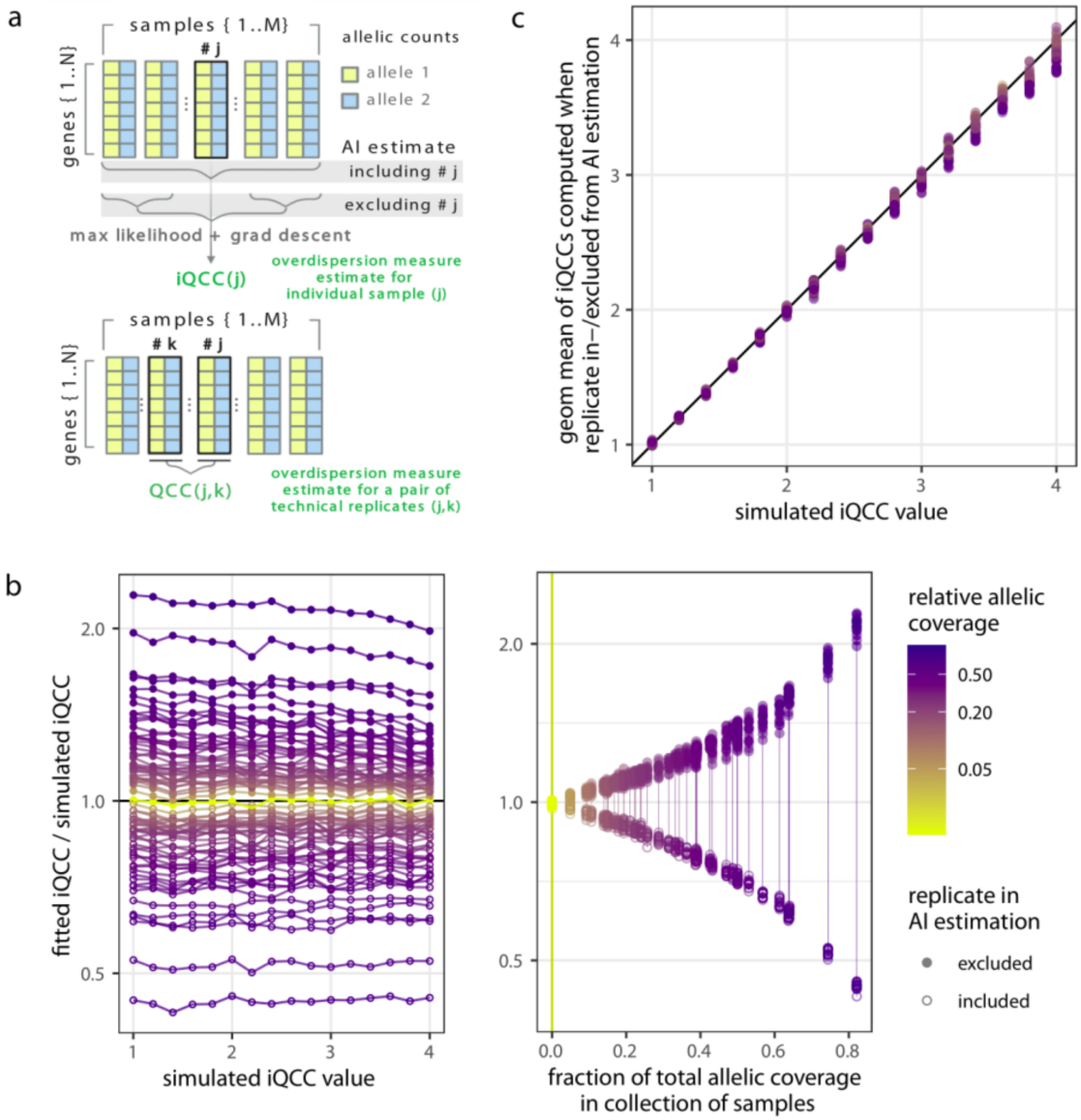

Рис. 9 — **Алгоритм вычисления iQCC принимает широкий диапазон количеств образцов и размеров библиотек. (a)** Схема вычисления iQCC для образца j (`controlFreq`) и вычисления QCC для пары i,j (`Qllelic`). **(b)** Верхняя (без включения образца j в оценку AI) и нижняя (с включением) оценки iQCC из общего набора реплик. Левая панель: вместе с увеличением аллельного покрытия верхняя и нижняя оценки сходятся к действительному значению iQCC (золотое); правая: расширение панели (a), показывает симметрию между верхней и нижней оценками (линии соединяют оценки из одной комбинации реплик). **(c)** Те же данные, что и в (b), геометрические средние нижней и верхней оценок явно коррелируют с симулированными уровнями избыточной дисперсии. Данные (b-c) произведены с помощью симуляций с разными общими покрытиями и уровнями избыточной дисперсии (iQCC были близки для всех реплик внутри одной генерации, что симулирует техническую репликацию или похожие на неё условия экспериментов). Для значений iQCC между 1 (нет избыточной дисперсии) и 4 (высокая избыточная дисперсия), покрытия генов на аллелях для 10 реплик (“библиотек”) были выбраны с использованием предопределённого распределения аллельных дисбалансов и коэффициентов для общих аллельных покрытий; процедура была повторена три раза для каждого набора параметров. Данные: из (Mendelevich et al., 2023).

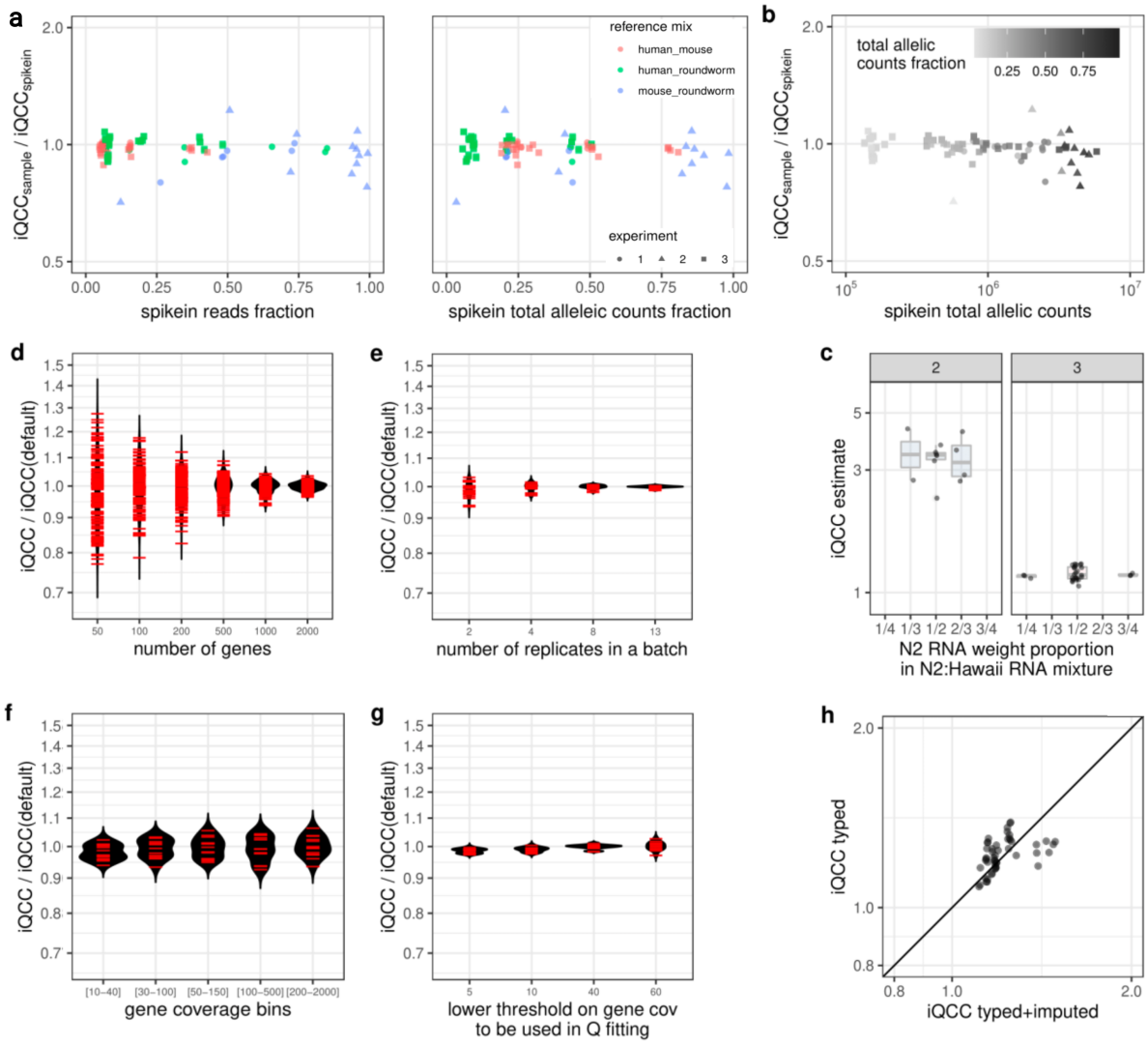

Рис. 10 — **Оценка избыточной дисперсии устойчива к варьированию относительного количества и состава РНК-контролей. (a)** Соотношение значений iQCC для подмножеств библиотеки из организмов 1 и 2 (“основной образец” и spike-in) остается близким к 1 с уменьшением доли контрольной РНК в исходной смеси. **(b)** Соотношение значений iQCC для организмов 1 и 2 при различных общих аллельных покрытиях против доли прочтений из РНК-контроля. Точность оценки iQCC не снижается при низком общем покрытии в десятки тысяч аллельных прочтений. **(c)** Диплоидность в РНК-контроле может быть смоделирована смесью двух генетически далёких линий. Оценки iQCC устойчивы между смесями в различных пропорциях линий N2 и Hawaii (C.elegans). **(d)** Оценки iQCC для случайных выборок из N генов, на 13 образцах (с 10% контроля), выборка производилась 10 раз. **(e)** Оценки iQCC вычислены на различных подмножествах образцов: 11 пар, 3 четвёрок, 3 восьмёрок and 2 подмножества из 13 образцов. **(f)** Оценки iQCC вычислены для разных интервалов аллельных покрытий, на 13 образцах (с 10% контроля), были использованы только те гены, где аллельное покрытие принадлежало интервалу во всех 13 образцах (количество генов в интервалах на рисунке: 611, 610, 399, 496, 232). **(g)** Оценки iQCC сделаны с разным минимальным порогом аллельного покрытия в процессе вычисления iQCC, на 13 образцах (с 10% контроля). **(d-g)** Данные: Эксп. 3, мышиный РНК-контроль. По умолчанию, значения iQCC рассчитаны с порогом аллельного покрытия = 30, с использованием всех 21 образцов, без дополнительных ограничений. **(h)** Значения iQCC вычислены на образцах, содержащих РНК человека, с использованием или отсутствием импутированных вариантов. Данные: из (Mendelevich et al., 2023).

Выводы:

1. Разработан новый вычислительный подход, реализованный в пакете `Qllelic` на языке `R`, для точной оценки ASE в данных РНК-секвенирования, учитывающий избыточную дисперсию с использованием технических реплик. Данный подход существенно снижает количество ложноположительных результатов по сравнению с традиционными методами. Рекомендуется иметь по крайней мере две технических реплики для каждого образца.
2. Использование `Qllelic` позволяет проводить надежный дифференциальный анализ аллель-специфической экспрессии и сравнивать образцы из разных экспериментов при наличии реплик библиотек для вычисления меры избыточной дисперсии, QCC. Избыточная дисперсия в экспериментах поли-А РНК-секвенирования имеет мультипликативную природу и равномерна для разных участков генома.
3. Наибольший вклад в избыточную дисперсию вносит процесс приготовления библиотек, в то время как процесс секвенирования оказывает незначительное влияние. Некоторые вычислительные методы также способны увеличивать уровень шума.
4. Метилирование ДНК является одним из ключевых механизмов для поддержания моноаллельной экспрессии в клональных лимфоидных клеточных линиях млекопитающих. Не все MAE гены подвержены влиянию деметилирования ДНК, что указывает на наличие дополнительных механизмов поддержания MAE.
5. При оценке влияния полногеномного деметилирования ДНК на транскрипционный аллельный дисбаланс с применением `Qllelic`, были выявлены значительные смещения ASE в более чем 600 аутосомных генах. Часть генов продемонстрировала сходимость AI в четырёх клонах к общему значению. Это указывает на то, что ландшафт клонального разнообразия в аллель-специфической регуляции генов сложнее, чем предполагалось ранее.
6. Не все изменения в ASE при воздействии 5-aza-dC оказались связаны с изменением в покрытии гена. Это значит, что ингибиторы ДНК-метилтрансферазы могут влиять на регуляцию генов у пациентов таким образом, который трудно обнаружить без аллель-специфического анализа.

7. Разработан новый экспериментальный и вычислительный подход, `ControlFreq`, основанный на добавлении внешних РНК-контролей в каждый образец перед приготовлением библиотеки. Данный подход не требует значительного увеличения стоимости эксперимента (на 5-10%, в отличии от минимум 2× при технической репликации), что упрощает применение этого метода в масштабных экспериментах. Его применение не влечёт за собой потери в точности анализа, по сравнению с анализом технических реплик.
8. Метод `ControlFreq` свободен от ряда существовавших ранее ограничений, связанных с размером библиотек и распределениями истинного AI. Он позволяет анализировать образцы с разным уровнем избыточной дисперсии, что облегчает поиск статистических выбросов. Использование `ControlFreq` для оценки избыточной дисперсии возможно как в случае РНК-контролей, так и при наличии технических реплик.
9. Подход с внешними РНК-контролями устойчив к варьированию параметров эксперимента. Для точной оценки избыточной дисперсии достаточно сравнительно небольшого общего аллельного покрытия контрольной компоненты в библиотеке, при наличии частых гетерозиготных полиморфизмов в РНК-контроле. Диплоидность в РНК-контроле может быть сымитирована смесью двух генетически далёких линий.

Список литературы

1. *noisyR*: enhancing biological signal in sequencing datasets by characterizing random technical noise / I. Moutsopoulos [и др.] // Nucleic Acids Research. — 2021. — июнь. — т. 49, № 14. — e83—e83. — DOI: 10.1093/nar/gkab433. — URL: https://doi.org/10.1093/nar/gkab433.

2. Interpretable and tractable models of transcriptional noise for the rational design of single-molecule quantification experiments / G. Gorin [и др.] // bioRxiv preprint. — 2021. — DOI: 10.1101/2021.09.06.459173. — URL: https://www.biorxiv.org/content/early/2021/12/26/2021.09.06.459173.

3. Accounting for technical noise in single-cell RNA-seq experiments / P. Brennecke [и др.] // Nature Methods. — 2013. — сент. — т. 10, № 11. — с. 1093—1095. — DOI: 10.1038/nmeth.2645. — URL: https://doi.org/10.1038/nmeth.2645.

4. *Grün D.*, *Kester L.*, *Oudenaarden A. van.* Validation of noise models for single-cell transcriptomics // Nature Methods. — 2014. — апр. — т. 11, № 6. — с. 637—640. — DOI: 10.1038/nmeth.2930. — URL: https://doi.org/10.1038/nmeth.2930.

5. Transcriptome variation in human tissues revealed by long-read sequencing / D. A. Glinos [и др.] // Nature. — 2022. — авг. — т. 608, № 7922. — с. 353—359. — DOI: 10.1038/s41586-022-05035-y. — URL: https://doi.org/10.1038/s41586-022-05035-y.

6. Tools and best practices for data processing in allelic expression analysis / S. E. Castel [и др.] // Genome Biology. — 2015. — сент. — т. 16, № 1. — DOI: 10.1186/s13059-015-0762-6. — URL: https://doi.org/10.1186/s13059-015-0762-6.

7. GeneiASE: Detection of condition-dependent and static allele-specific expression from RNA-seq data without haplotype information / D. Edsgärd [и др.] // Scientific Reports. — 2016. — февр. — т. 6, № 1. — DOI: 10.1038/srep21134. — URL: https://doi.org/10.1038/srep21134.

8. *Jiang Y., Zhang N. R., Li M.* SCALE: modeling allele-specific gene expression by single-cell RNA sequencing // Genome Biology. — 2017. — апр. — т. 18, № 1. — DOI: 10.1186/s13059-017-1200-8. — URL: https://doi.org/10.1186/s13059-017-1200-8.

9. Salmon provides fast and bias-aware quantification of transcript expression / R. Patro [и др.] // Nature Methods. — 2017. — март. — т. 14, № 4. — с. 417—419. — DOI: 10.1038/nmeth.4197. — URL: https://doi.org/10.1038/nmeth.4197.

10. Near-optimal probabilistic RNA-seq quantification / N. L. Bray [и др.] // Nature Biotechnology. — 2016. — апр. — т. 34, № 5. — с. 525—527. — DOI: 10.1038/nbt.3519. — URL: https://doi.org/10.1038/nbt.3519.

11. *GTEx_Consortium.* Genetic effects on gene expression across human tissues // Nature. — 2017. — окт. — т. 550, № 7675. — с. 204—213. — DOI: 10.1038/nature24277. — URL: https://doi.org/10.1038/nature24277.

12. Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins / A. Buil [и др.] // Nature Genetics. — 2014. — дек. — т. 47, № 1. — с. 88—91. — DOI: 10.1038/ng.3162. — URL: https://doi.org/10.1038/ng.3162.

13. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data / J. F. Degner [и др.] // Bioinformatics. — 2009. — окт. — т. 25, № 24. — с. 3207—3212. — DOI: 10.1093/bioinformatics/btp579. — URL: https://doi.org/10.1093/bioinformatics/btp579.

14. dsPIG: a tool to predict imprinted genes from the deep sequencing of whole transcriptomes / H. Li [и др.] // BMC Bioinformatics. — 2012. — т. 13, № 1. — с. 271. — DOI: 10.1186/1471-2105-13-271. — URL: https://doi.org/10.1186/1471-2105-13-271.

15. MBASED: allele-specific expression detection in cancer tissues and cell lines / O. Mayba [и др.] // Genome Biology. — 2014. — авг. — т. 15, № 8. — DOI: 10.1186/s13059-014-0405-3. — URL: https://doi.org/10.1186/s13059-014-0405-3.

16. Genes with monoallelic expression contribute disproportionately to genetic diversity in humans / V. Savova [и др.] // Nature Genetics. — 2016. — янв. — т. 48, № 3. — с. 231—237. — DOI: 10.1038/ng.3493. — URL: https://doi.org/10.1038/ng.3493.

17. *Wang Z.*, *Gerstein M.*, *Snyder M.* RNA-Seq: a revolutionary tool for transcriptomics // Nature Reviews Genetics. — 2009. — янв. — т. 10, № 1. — с. 57—63. — DOI: 10.1038/nrg2484. — URL: https://doi.org/10.1038/nrg2484.

18. Autosomal monoallelic expression in the mouse / L. M. Zwemer [и др.] // Genome Biology. — 2012. — т. 13, № 2. — R10. — DOI: 10.1186/gb-2012-13-2-r10. — URL: https://doi.org/10.1186/gb-2012-13-2-r10.

19. dbMAE: the database of autosomal monoallelic expression / V. Savova [и др.] // Nucleic Acids Research. — 2015. — окт. — т. 44, № D1. — с. D753—D756. — DOI: 10.1093/nar/gkv1106. — URL: https://doi.org/10.1093/nar/gkv1106.

20. Developmental Dynamics and Disease Potential of Random Monoallelic Gene Expression / A.-V. Gendrel [и др.] // Developmental Cell. — 2014. — февр. — т. 28, № 4. — с. 366—380. — DOI: 10.1016/j.devcel.2014.01.016. — URL: https://doi.org/10.1016/j.devcel.2014.01.016.

21. Random Monoallelic Gene Expression Increases upon Embryonic Stem Cell Differentiation / M. A. Eckersley-Maslin [и др.] // Developmental Cell. — 2014. — февр. — т. 28, № 4. — с. 351—365. — DOI: 10.1016/j.devcel.2014.01.017. — URL: https://doi.org/10.1016/j.devcel.2014.01.017.