СКОЛКОВСКИЙ ИНСТИТУТ НАУКИ И ТЕХНОЛОГИЙ

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»

Диссертация на соискание учёной степени
кандидата физико-математических наук

Научный руководитель:
доктор биологических наук, профессор
Гельфанд Михаил Сергеевич

Москва — 2023

Оглавление

Стр.

Введение

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 оно побуждает авторов к попыткам обойти эти ограничения, которые часто приводят к нарушению распределений, лежащих в основе стандартных методов, и зачастую без должного учета в последующем анализе (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 может быть некорректным.

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. Некоторые 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 генетического и эпигенетического контроля аллель-специфической транскрипции. В то время как цис-регуляция определяет общее базовое состояние для всех гене-

тически 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 не отделима 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 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 дости-

гать большей статистической корректности (в случае РНК-контролей, без существенного увеличения затрат на эксперимент). Разработанный метод может быть полезен во многих транскриптомных исследованиях, и потенциально стимулировать разработку аналогичных протоколов при работе с другими типами данных, таких как длиннoридное 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, Feb 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) до-

клад: «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

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.

Глава 1. Обзор литературы

1.1 Механизмы аллельного дисбаланса

Разнообразные генетические и эпигенетические факторы влияют на относительные уровни экспрессии двух копий каждого конкретного гена в клетках диплоидных организмов. Помимо генетической вариации в регуляторных областях, влияющей на аллель-специфическую экспрессию [11; 12], существует по крайней мере три основных типа не-менделевских, эпигенетических явлений, которые контролируют аллель-специфическую экспрессию у млекопитающих. Одно из них — инактивация X-хромосомы [13]: во время развития женских эмбрионов около половины клеток выбирают инактивировать материнскую X-хромосому, а остальные инактивируют отцовскую, что затрагивает большинство генов, сцепленных с X-хромосомой [14—17]. Другим примером является генный импринтинг: такие гены, как IGF2 и H19, экспрессируются либо с отцовской, либо с материнской аллели [18; 19].

Моноаллельная аутосомная экспрессия (MAE), схожим образом с инактивацией X-хромосомы и импринтингом, является митотически стабильным эпигенетическим механизмом, который существенным образом влияет на взаимосвязь генотипа и фенотипа у млекопитающих, контролируя относительную экспрессию двух родительских аллелей. MAE является самым распространенным из этих эпигенетических явлений, влияющим на тысячи аутосомных генов в человеческом геноме, включая множество генов, ассоциированных с раком и неврологическими заболеваниями [20; 21].

Хотя отдельные примеры генов MAE были давно известны (например, гены обонятельных рецепторов [22]), полно-транскриптомные исследования аллель-специфической экспрессии привели к неожиданному выводу: около 4000 аутосомных генов человека могут быть иметь моноаллельную экспрессию [23; 24]. Моноаллельная экспрессия наблюдалась в каждом исследованном типе клеток, включая периферическую кровь и производные клеточные линии, а также в человеческой плаценте, мышиных лимфоидных клетках и фибробластах, и

мышиных эмбриональных стволовых клетках и нейрональных предшественниках (NPCs) [23—29].

Аналогично инактивации X-хромосомы, где развитие опухоли зависит от того, какая аллель гена FOXP3 подавляется [30], стабильные различия в аллель-специфической экспрессии генов MAE могут привести к существенным функциональным различиям между похожими в остальном клетками — например, между B-лимфоцитами, разделенными по их ответу на липополисахариды, в зависимости от того, какая аллель гена *Tlr4* подавляется [31].

Обнаружение большого числа генов MAE за счёт использования их специфической хроматиновой конфигурации позволило получить представление об их эволюции через анализ на уровне больших популяций. Гены MAE вносят существенный вклад в генетическую [32] и транскрипционную [33] вариацию в человеческих популяциях. Моноаллельные аутосомные гены, как группа, подвержены долгосрочному балансирующему отбору [32]. Этот вывод, подтвержденный другими группами [34; 35], предполагает гетерозиготное преимущество для MAE генов. То, что гетерозиготность приводит к повышению приспособленности организма, указывает на то, что биологическая функция MAE связана с созданием гетерогенности в клеточных популяциях.

1.2 Аллельный дисбаланс используется для изучения генной регуляции

Предыдущие исследования продемонстрировали перспективность анализа локусов количественных признаков (QTLs) и аллельного дисбаланса (AI) для понимания генетического обоснования транскрипционной цис-регуляции [36; 37]. С появлением высокопроизводительного секвенирования были разработаны статистические методы для моделирования данных дисбаланса в аллельной экспрессии [38—40]. Было показано, что покрытие является ключевым параметром, определяющим мощность и чувствительность измерения AI [41]. Некоторые группы использовали аллельный дисбаланс для нахождения признаков GWAS (полногеномного поиска ассоциаций) [40; 42; 43].

Аллель-специфический анализ транскрипции также использовался для понимания эпигенетических механизмов регуляции генов, включая Х-инактивацию и импринтинг [19], а также их нарушение при заболеваниях, например, влияние Х-инактивации на развитие рака [44]. Другие исследования были направлены на использование аллель-специфического анализа для понимания хода дифференциации в клональных линиях [45].

Наиболее распространённым объектом для аллель-специфического анализа является РНК. Также аллель-специфический сигнал с геномной ДНК человека или мыши может использоваться для изучения доступности хроматина с помощью картирования участков гиперчувствительности к ДНКазе [46]; пространственной организации Х-хромосомы в ядре с использованием Hi-C [47], времени хромосомной репликации [48], и связывания транскрипционных факторов с использованием ChIP-seq [49].

Высокопроизводительное секвенирование коротких прочтений (Illumina) является наиболее распространенным технологическим подходом для аллель-специфического анализа ДНК или РНК в исследованиях полногеномного масштаба. В то же время, следует отметить существование спектра технических подходов для оценки аллель-специфического сигнала. Все они опираются на наличие участков последовательности, различающихся между отцовской и материнской аллелями, таких как однонуклеотидные полиморфизмы (SNP). Эти методы включают в себя удлинение праймеров [11; 50], генотипирующие чипы [23; 24], таргетированное секвенирование [26; 51], аллель-специфический FISH (Флуоресцентная гибридизация in situ) [52]. Наиболее современные работы начали использовать секвенирование длинных прочтений с помощью технологий третьего поколения [5; 53; 54].

Учитывая распространенность аллель-специфических анализов данных РНК-секвенирования, мы сосредоточили своё внимание на повышении точности оценки сигнала в этом типе эксперимента.

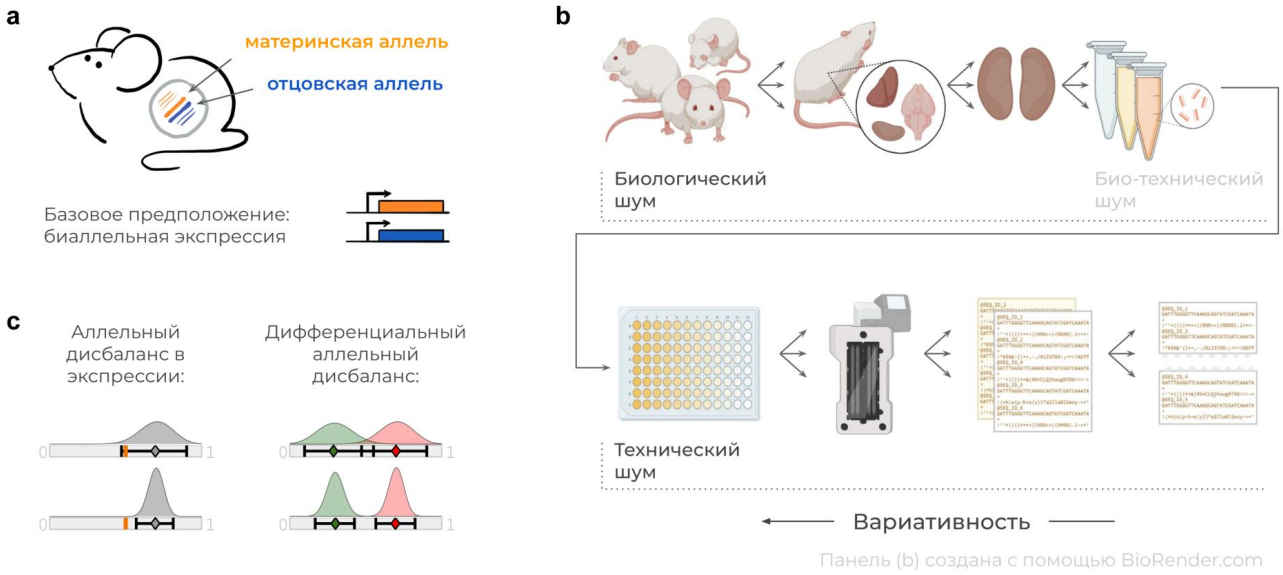

Рисунок 1.1 — Накопление экспериментального шума в процессе производства данных РНК-секвенирования.

(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.

1.3 Измерение шума в данных РНК-секвенирования и одноклеточного РНК-секвенирования

Технический шум. Измерение аллель-специфической экспрессии требует решения проблем, связанных с ошибками измерения. Технический шум в транскрипционных данных — давно известное явление [55], и его присутствие влияет на эффективность улавливания биологического сигнала, если не учитывается должным образом [1; 56]. Необходимость отделения технического шума от биологической вариации особенно актуальна для таких шумных методов, как одноклеточное РНК-секвенирование [4; 57—59].

Анализ технических реплик [55] — один из способов измерения и учёта технического шума в данных секвенирования. Использование ДНК контролей для нормализации уровня дисперсии в аллельном дисбалансе и референсного перекоса представляют другой, принципиально сходный с анализом техниче-

ских реплик подход к решению проблемы оценки шума [60]. Высокая стоимость эксперимента с технической репликацией является сильным стимулом для избегания их использования. Поэтому ряд подходов был направлен на оценку технического шума при сравнении данных внутри одной библиотеки секвенирования (например, с помощью сравнения разных сегментов генома или разных одноклеточных полиморфизмов в одном и том же гене [7; 61], или с помощью сравнения *in silico* образцов [62] и бутстрэпинга [63]).

Техническая репликация в одноклеточном РНК-секвенировании невозможна, поэтому большинство попыток вычисления уровня экспрессии в одноклеточных данных представляют собой байесовские и иерархические модели [64; 65], однако известен также не обобщаемый в практическом смысле способ, основанный на разделении клеток на две [27]. Показано также, что использование 96 стандартов (ERCC) в анализе дифференциальной экспрессии [66] решает проблему контроля за различными размерами библиотек и вариацией числа копий (CNV) в РНК-секвенировании, и плоидностью и дуплетами в одноклеточном РНК-секвенировании [67; 68].

Наиболее распространенным модельным распределением для аллельного шума является биномиальное [5—8], в этой модели не учитывается никакой дополнительный шум (избыточный шум, или «сверхдисперсия»), присутствующий в данных, и хорошо известный в контексте анализа покрытий генов в данных РНК-секвенирования [69—71]. Для учёта избыточной дисперсии в аллель-специфической экспрессии было разработано несколько моделей, основанных на бета-биномиальном распределении [7; 72; 73]. Среди альтернативных вариантов есть также иерархические и байесовские модели [60; 65].

Источники технического шума. Различные биологические вопросы в некоторых случаях представлены схожими статистическими задачами, имеющими похожие решения. Например, как в анализе данных РНК-секвенирования для измерения аллель-специфической экспрессии (ASE), так и в задаче количественной оценки покрытия альтернативно сплайсированных изоформ, наблюдается следующая проблема: большинство прочтений не информативны, поскольку они не включают SNP или границу экзона. В попытке избежать потери большой части данных, многие популярные программные инструменты распределяют не информативные прочтения на основе статистик, вычисленных на

информативных прочтениях. Более того, многие инструменты утверждают, что эту стратегию можно применять как к количественной оценке изоформ, так и к аллель-специфической экспрессии. К таким инструментам относятся байесовская и использующая алгоритм поиска максимального правдоподобия модель `RSEM` [74], графовые алгоритмы `Salmon` [9] и `Kallisto` [70], байесовский подход `ASE-TIGAR` [75], и инструмент на основе бутстрэпа `IsoEM2(IsoDE2)` [63]. Таким образом, технический шум может быть не умышленно увеличен на последних стадиях обработки данных.

Экспериментальные источники избыточного шума в данных РНК-секвенирования не до конца изучены и существенно варьируют между различными протоколами. Одной из самых известных проблем в экспериментах глубокого секвенирования является влияние артефактов ПЦР-амплифицирования [76—78]. И использование баркодирования уникальными молекулярными идентификаторами (UMI) [79] считается самым эффективным способом учёта этих артефактов, однако далеко не все эксперименты сделаны с их использованием. Применения методов дедупликации на этапе вычислительной обработки данных в научной среде является поводом для непрекращающихся дебатов. Артефакты обратной транскрипции являются достаточно изученной, но редко учитываемой в современных исследованиях проблемой [80; 81]. Помимо этого, очевидно, что сэмплирование образца в процессе эксперимента способно вносить избыточность в дисперсию, в том случае, если выбор происходит неслучайным образом, или сложность библиотеки недостаточна и не позволяет рассматривать модель сэмплирования как пуассоновскую. Неравномерность кДНК фрагментации также вносит в свой вклад в неслучайность сэмплирования [82].

Дифференциальный анализ. Дифференциальный анализ данных РНК-секвенирования является одной из основных задач в полнотранскриптомном анализе. Для его проведения разработано множество подходов, каждый из которых обладает собственными ограничениями и областью применения [83—85], но в целом они считаются сопоставимыми. Наиболее часто используемые подходы для работы с дифференциальной экспрессией включают в себя `DeSeq2` [71] и `edgeR` [69], а также надстройку `Voom` [86] над пакетом `limma` [87], который расширяет его применение с микрочипов до РНК-секвенирования. Эти инструменты естественно использовать в анализе аллель-специфической экс-

прессии, если рассматривать родительские аллели как “образцы” для сравнения (rpubs.com/mikelove/ase).

Помимо учета технического шума, упомянутого выше, важным статистическим вопросом является нормализация данных, которая позволяет сравнивать разные наборы данных между собой [88]. Все модели в той или иной степени опираются на предположение, что большинство генов не дифференциально экспрессированы, и имеют своей целью преобразовать данные так, чтобы они принадлежали похожим распределениям. В частности, в `DeSeq2`, `edgeR` и `limma` имплементированы разные методы нормализации: `DESeq` масштабирует покрытие гена на геометрическое среднее среди всех образцов, `EdgeR` использует подход нормализации на покрытие генов фиксированного образца из набора данных (Trimmed Mean of M-values, TMMs), в то время как `limma` основана на квантильном масштабировании.

В то время, как желательность биологический репликации обычно не подвергается сомнению, наиболее распространённые рекомендации заключаются в производстве одной технической реплики на образец.

Также существует несколько инструментов, специально разработанных для дифференциального анализа аллель-специфической экспрессии, например, `ASE-TIGAR` [75], `MBASED` [61], `GeneiASE` [7], причём в последнем отсутствует возможность обработки реплик. В главе 2 приводится сравнение результатов, полученных разными программами.

Глава 2. Реплики библиотек секвенирования играют важную роль в количественной оценке аллельного дисбаланса

2.1 Введение

Точный количественный анализ в данных РНК-секвенирования невозможен без корректного отделения биологической вариабельности от экспериментального шума. Если нет технических реплик, оценка технического шума неизбежно вынуждена полагаться на не всегда надёжные предположения о его природе. Несмотря на это, среди общедоступных наборов данных РНК-секвенирования трудно найти такие, которые содержат технические реплики (т.е. реплики библиотек для каждого образца или даже для части образцов). Более того, в исследованиях ASE *de facto* стандартом является простой биномиальный тест с коррекцией на множественное тестирование [89—92] (иными словами, неявное предположение об отсутствии технического шума). Также существует несколько методов, оценивающих технический шум из одной реплики [7; 61].

Мы задались целью определить, достаточно ли одной технической реплики для оценки вклада технического шума в наблюдаемый сигнал ASE. Здесь мы приводим экспериментальные и теоретические доказательства, что дизайн с одной технической репликой может привести к неучтённой избыточной дисперсии и повышенной доле ошибок в аллель-специфическом анализе РНК-секвенирования.

Для исследования природы технического шума в аллель-специфических данных РНК-секвенирования мы провели эксперимент с большим количеством реплик библиотек из одной и той же РНК, варьируя метод создания библиотеки и исходное количество РНК. Результаты анализа этого набора данных продемонстрировали, что уровень избыточной дисперсии эксперимент-специфичен. Эти наблюдения были подкреплены аналогичными результатами при анализе общедоступных данных.

Мы разработали вычислительный метод, `Qllelic` (github.com/gimelbrantlab/Qllelic), способный оценивать и учитывать технический шум, используя две и более библиотеки РНК-секвенирования. Мы

показали, что его применение существенно улучшает воспроизводимость оценок транскриптомного AI. Мы также продемонстрировали преимущества `Qllelic` в дифференциальном анализе ASE, проведя сравнения с другими широко применяемыми методами для анализа ASE.

Наконец, мы исследовали источники технической избыточной дисперсии в наблюдаемом сигнале транскриптомного AI.

Эта глава основана на публикации:
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

2.2 Результаты

2.2.1 Одной технической реплики РНК-секвенирования недостаточно для оценки технического шума аллельного дисбаланса

Для точного анализа данных РНК-секвенирования, биологический сигнал должен быть отделён от экспериментального шума. Один из очевидных источников технической вариации — это то, что аликвота есть подвыборка биологического образца, которая будет использована для подготовки библиотеки РНК-секвенирования. Эта часть вариации обычно может быть учтена при помощи биномиального распределения [6; 89]. Многие существующие подходы к анализу аллельного дисбаланса также включают в себя дополнительную компоненту шума, избыточную дисперсию сверх биномиальной [61; 72; 73].

При некотором упрощении модели истинного распределения AI, одна техническая реплика может предоставлять достаточное количество данных, в рамках модели. Примером такого упрощения является тримодальное представление распределения истинного AI, где принимаемые значения ограничены пропорцией 50:50 (биаллельная экспрессия) и предельными значениями (только материнская или отеческая, см. Раздел A.1.1). Однако для точного количествен-

ного анализа ASE — например, в задаче дифференциального анализа аллельного дисбаланса — требуются более реалистичные априорные модели распределения AI, чем классификация на смещённый / несмещённый. Естественным усовершенствованием тримодальной модели является непрерывная модель колоколообразного распределения с центром в 50:50 и толстыми хвостами в 0:100 и 100:0 (см. Рис.A.31), которую естественно приблизить смесью двух бета-биномиальных распределений. Добавочный к выборочному, технический шум часто моделируется как бета-компонента в бета-биномиальном распределении [7; 60; 61; 72].

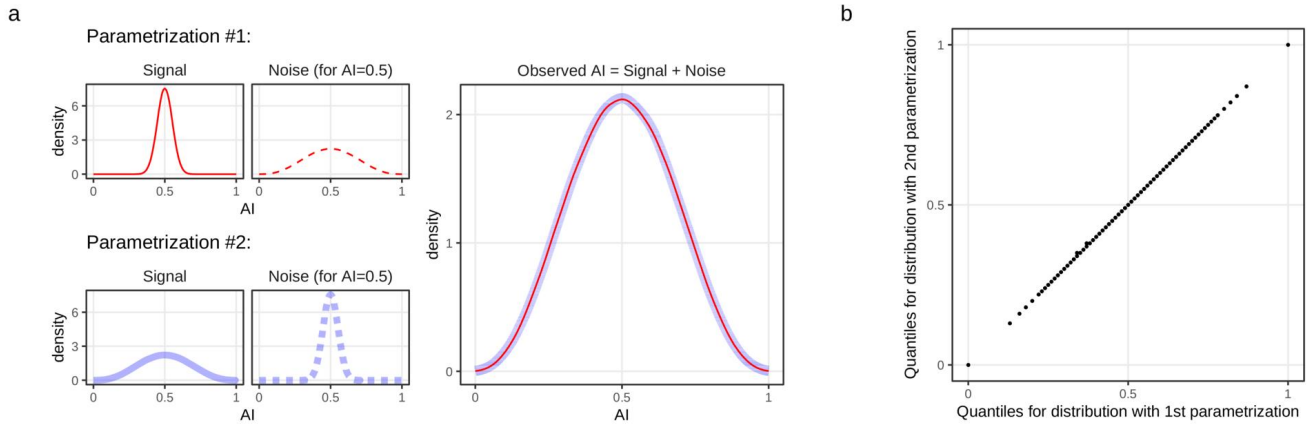

Рисунок 2.1 — Разные комбинации параметров сигнала и шума могут давать неотличимые наблюдаемые распределения AI.

a: Две симулированные параметризации (слева) настоящего сигнала AI (AI_{true}; сплошная линия) и шума (пунктирная линия), которые вместе дают перекрывающиеся наблюдаемые значения AI (справа; красным и синим, соответственно). Наблюдения из этих двух распределений неотличимы тестами Манна-Уитни-Уилкоксона и Колмогорова-Смирнова; см. также Раздел A.1.1. Распределения AI показаны для аллельного покрытия 100. Распределения шума показаны для параметра $AI_{true} = 0.5$. И сигнал, и шум смоделированы как бета-биномиальные распределения; параметры на рисунке: [$\rho_{Signal} = 0.001$, $\rho_{Noise} = 0.1$] и [$\rho_{Signal} = 0.1$, $\rho_{Noise} = 0.001$]; размер симулированной выборки — 500000. Другие уровни покрытия и комбинации значений ρ показаны в Разделе A.1.1. **b:** Квантиль-квантиль график распределений, полученных параметризациями 1 и 2 из панели (a). Значения квантилей были взяты в арифметической прогрессии от 0 до 1 с шагом 0.01.

Даже в рамках этой простой непрерывной модели наблюдаемое распределение значений AI может быть результатом различных параметризаций бета-биномиальных распределений, описывающих истинный сигнал ASE и шум (Рис.2.1a). Анализ симуляций (см. Раздел A.1.1) показывает, что такие параметризации, задающие очень разный уровень шума (Рис.2.1a), невозможно различить с помощью тестов Манна-Уитни-Уилкоксона и Колмогорова-Смирнова

(см. Раздел A.1.1). Здесь шум включает в себя все источники: и технический шум при подготовке библиотеки и измерении, и биологические вариации.

В Разделе A.1.1 мы приводим два других примера классов моделей, и аналитически показываем, что несколько параметризаций могут привести к эквивалентным с точки зрения наблюдений распределениям и в этих случаях. В одном из таких классов и истинный сигнал, и шум распределены нормально. В другом шум моделируется либо бета-биномиальным, либо биномиальным распределением, а истинные значения AI — как смесь трех дельта-функций Дирака или смесь бета-распределений соответственно. Эти примеры убедительно показывают, что данных РНК-секвенирования из одной технической реплики недостаточно для подбора параметризации технического шума и истинного сигнала AI, если только модели не являются очень ограниченными.

В остальной части главы мы описываем и экспериментально проверяем метод учета технического шума и точной оценки AI с использованием данных РНК-секвенирования из двух или более технических реплик.

2.2.2 Использованные данные

В создании библиотек в экспериментах РНК-секвенирования есть два основных параметра: (а) можно использовать различные протоколы; и (б) при использовании одного и того же протокола подготовка библиотек может начинаться с разного количества РНК. Чтобы протестировать влияние этих двух переменных, мы приготовили три набора библиотек поли-А РНК-секвенирования из одной и той же РНК, выделенной из почки мыши. Чтобы увеличить долю аллель-специфических прочтений, мы использовали РНК из гибрида мыши (129S1×Cast/Ei), имеющего геномную плотностью однонуклеотидных полиморфизмов (SNPs) ∼1/118 п.н., что примерно в 10 раз выше, чем у человека.

Каждый набор (“эксперимент”) состоял из шести библиотек, приготовленных параллельно. Библиотеки для эксперимента 1 (“NEBNext (100ng)”) были подготовлены с использованием протокола для большого количества исходной РНК (100 нг тотальной РНК, подробности см. в Методах). В экспериментах 2 и 3 использовались библиотеки, приготовленные с помощью набора SMART-

Seq v4 Ultra Low Input RNA Kit (Clontech) с количеством исходной тотальной РНК в пределах рекомендуемого диапазона — 10 нг и 100 пг, соответственно (“SMARTseq (10ng)” и “SMARTseq (0.1ng)”). Информация об этих данных обобщена в Таблице 2.

Мы также проанализировали опубликованные наборы данных, информация о них представлена в Таблице 2, а более подробная информация об их анализе приведена в Таблице 3 (образцы человека [93]) и Таблице 4 (образцы мыши [29]).

2.2.3 Разброс оценок аллельного дисбаланса в репликах библиотек РНК-секвенирования

Для оценки согласованности оценок AI между техническими репликами мы применили единый статистический тест для всех технических реплик и экспериментов, чтобы построить “карту рассогласованности AI” для пар реплик (Рис.2.2a). Для каждого гена в любой из сравниваемых реплик мы проверили нулевую гипотезу о биаллельной экспрессии (H_0 AI $= 0.5$) с помощью биномиального теста. Гены с отвергнутой H_0 (при $P = 0.05$, после строгой поправки на множественные проверки гипотез Бонферрони [94]) классифицируются как имеющие смещение AI; остальные классифицируются как не имеющие смещения. На диаграмме показаны только гены с противоречивой классификацией в двух репликах, с отдельным выделением генов с противоположным смещением AI. Этот тест, широко используемый в исследованиях ASE [89—92], дает большое количество не согласованных между техническими репликами результатов (рис.2.2b). Стоит заметить, что AI генов, подающих противоречивые сигналы, распределены достаточно широко вокруг границы, определенной биномиальным тестом, что говорит о том, что чистый биномиальный шум не является хорошим приближением экспериментально наблюдаемой дисперсии значений AI (см. также Раздел A.1.2).

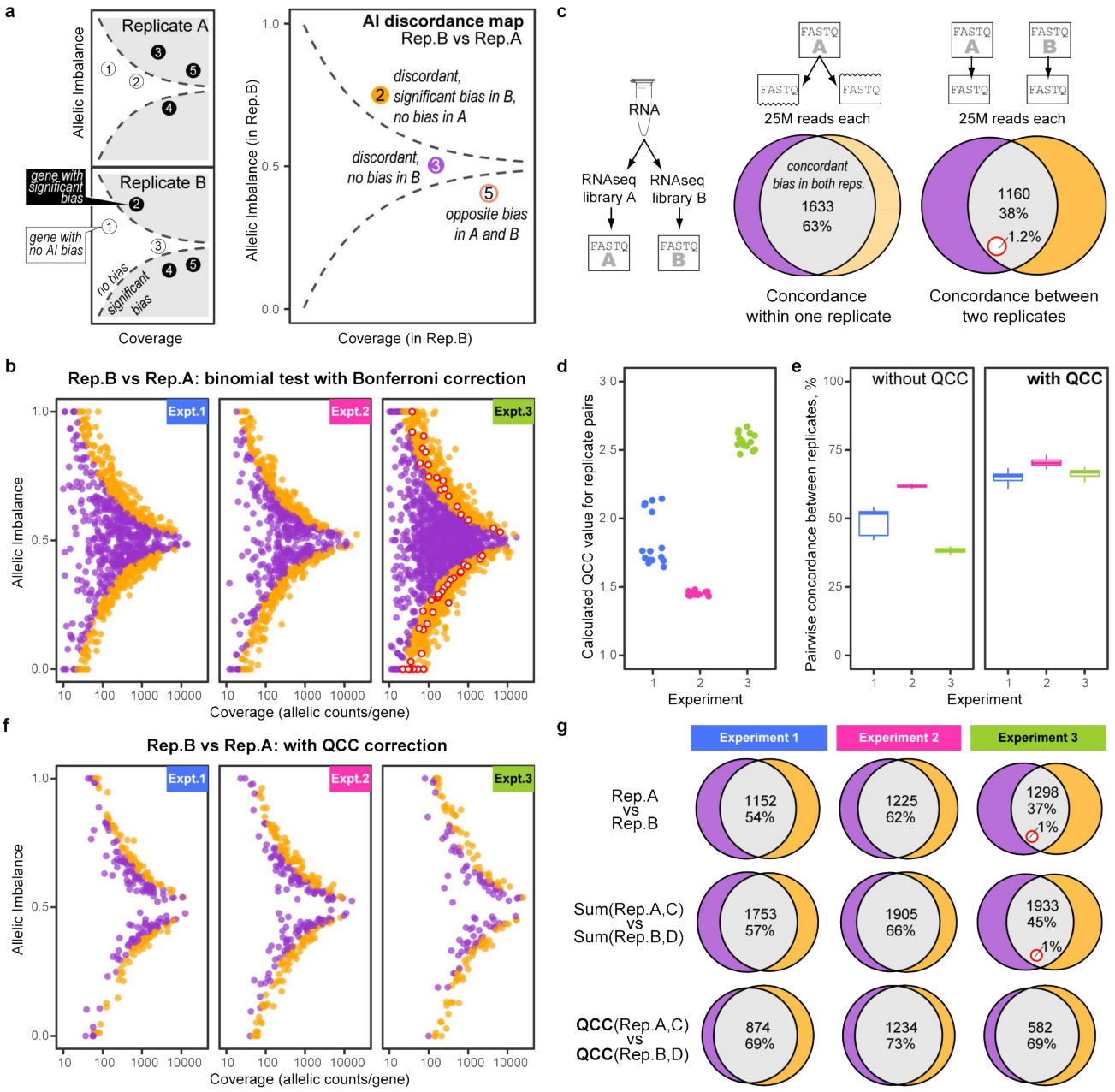

Рисунок 2.2 — Значения аллельного дисбаланса не совпадают для разных технических реплик и экспериментов РНК-секвенирования.

a: Объяснение графика рассогласования оценок аллельного дисбаланса. Типы рассогласования: оранжевый — нет смещения в реплике РНК-секвенирования A, есть смещение в реплике B; пурпурный — есть смещение в реплике A, нет смещения в реплике B; красный незаполненный круг — смещение в обоих репликах, но в противоположных направлениях. Значимое смещение: гипотеза H_0 (AI = 0.5) для гена отвергнута двусторонним биномиальным тестом при уровне $p = 0.05$ с поправкой Бонферрони. **b:** Графики рассогласованности AI для репрезентативных пар технических реплик во всех трёх экспериментах (см. Методы и Табл.2). Цвета те же, что и в (a). Все сравнения здесь произведены на подмножествах из 30 миллионов уникально картированных прочтений для каждой реплики, если не указано обратное. **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. Цвета те же, что и в (a); проценты показывают долю всех рассогласованных генов. Данные: реплики 1 и 2 из Эксперимента 3 (см. Табл.2). **d:** Коэффициент коррекции

качества (QCC), мера избыточной дисперсии, определённая в этой главе, посчитана (см Рис.2.3) 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 Эксперименте 1 (соответствует пяти парам-выбросам). **e:** Доля согласованно смещённых генов [сравните с серой областью в (c)] для всех 15 пар реплик в Экспериментах 1–3. Слева: двусторонний биномиальный тест. Справа: двусторонний пропорциональный тест с коррекцией на QCC. Элементы диаграмм размаха — центральная отметка: медиана; границы прямоугольника: верхние и нижние квартили; усы: 1.5x интерквартильный интервал; точки: выбросы. **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. Цвета те же, что и в (c). Верхний ряд: сравнение двух индивидуальных реплик [реплики 2 и 3; см. Табл.2], выборки в 30 миллионов прочтений; тест H_0 при помощи двустороннего биномиального теста с поправкой Бонферрони. Средняя и нижняя строки: сравнение объединённых пар реплик [реплики 2+4 против 3+5; см. Табл.2], выборки 30M прочтений на реплику. Средняя строка: двусторонний биномиальный тест с поправкой Бонферрони. Нижняя строка: двусторонний пропорциональный тест с коррекций на QCC.

Известно, что в данных РНК-секвенирования существует избыточная дисперсия — более высокая вариабельность сигнала, чем можно было бы ожидать, исходя из предположения о биномиальном шуме [69—71]. Существуют подходы, измеряющие избыточную дисперсию по одной технической реплике, к примеру, сравнением результатов выборок без повторений из одного и того же набора прочтений [62] или бутстрэпинг [69]. В самом деле, выборка без возвращения, естественная симуляция секвенирования двух аликвот из одной библиотеки, должна иметь большую дисперсию, чем биномиальная выборка в пределах одной реплики (Рис.A.19), что позволяет предположить, что эта процедура отражает некоторую избыточную дисперсию. Однако технические реплики внутри одного эксперимента показывают большую дисперсность, чем при выборке без возвращения (Рис.2.2c), что демонстрирует наличие дополнительного шума, не обнаруживаемого при анализе внутри одной библиотеки.

Отметим, что уровень согласованности результатов между репликами существенно различается для разных экспериментов (Рис.2.2b, слева направо), оставаясь схожим для пар реплик внутри одного эксперимента: $49{,}2\% \pm 4.7$ для эксперимента 1 ($52.3\% \pm 1.3$, при исключении реплики-выброса), $61.8\% \pm 0.6$ для эксперимента 2 и $38.3\% \pm 1.0$ для эксперимента 3 (Рис.2.2e, слева). В дополнение к этому, мы проанализировали общедоступные данные, показав наличие и внутреннюю консистентность избыточной дисперсии, вычисленной на парах технических реплик в этих экспериментах [29; 93] (см. Таблицу 3 и 4).

Таким образом, мы можем заключить, что избыточная дисперсия AI, наблюдаемая в данных поли-A РНК-секвенирования, зависит от эксперимента.

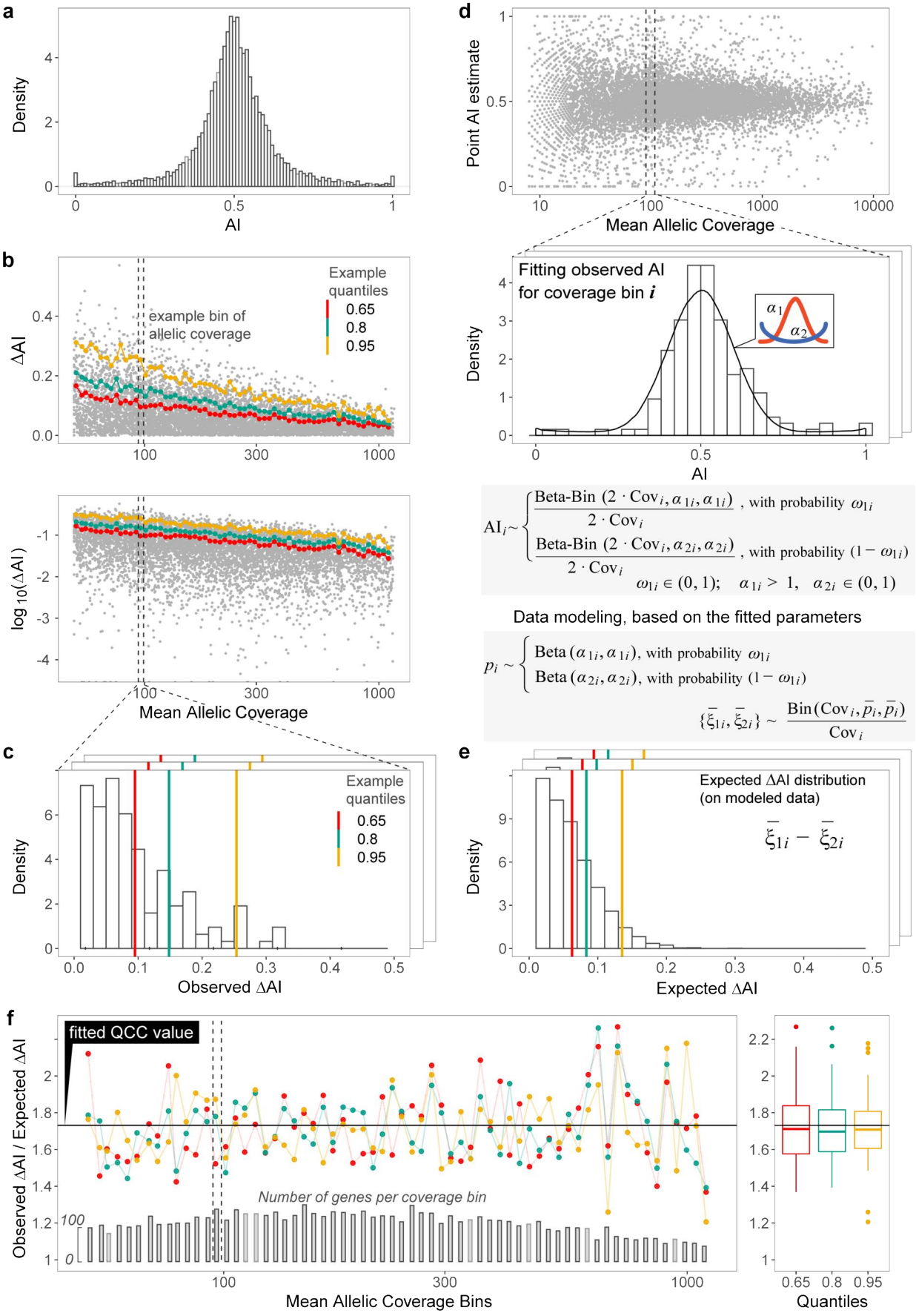

Рисунок 2.3 — Вывод коэффициента коррекции качества (QCC) из наблюдаемых и смоделированных разностей AI между техническими репликами.

a-c: Вычисление квантилей наблюдаемых распределений абсолютных разностей между аллельными дисбалансами в двух репликах (ΔAI). **a:** Распределение точечных оценок AI для генов с аллельным покрытием более 10 в шести объединённых репликах (180 миллионов прочтений суммарно) Эксперимента 2. **b:** После выборки равного количества прочтений из двух технических реплик, для каждого гена посчитан AI. На графике изображена зависимость ΔAI от среднего покрытия гена (сумма покрытий по всем SNP) в линейной (сверху) или логарифмической (снизу) шкале. Гены разбиты на корзины по логарифму покрытия (показан пример корзины). Для каждой корзины посчитаны квантили. Здесь и далее, показаны три фиксированных квантиля: 0.65 (красным), 0.8 (зелёным) и 0.95 (оранжевым). **c:** Распределение наблюдаемых значений ΔAI в корзине. Показаны фиксированные квантили. **d-e:** Вычисление квантилей ожидаемых распределений ΔAI. **d:** Сверху: для каждого гена посчитан AI после объединения прочтений SNP из обоих реплик. Заметим, что мы используем среднее покрытие гена, поэтому корзины содержат одинаковые гены в обоих репликах. Снизу: для каждой корзины покрытия, распределение значений AI подобрано при помощи смеси двух симметричных бета-биномиальных распределений (красная и синяя кривые). **e:** Распределение ожидаемых значений ΔAI в корзине. Чтобы сформировать ожидаемую ΔAI:

- мы генерируем симулированный набор из 5000 генов, с распределением точных значений аллельного дисбаланса (ξ) согласно подобранным параметрам;
- для этих генов, мы симулируем два набора данных реплик, с покрытием, соответствующим корзине, и берём выборку из биномиального распределения;
- наконец, мы вычисляем симулированную ΔAI для каждого гена, и ищем квантили их распределений.

f: Частные наблюдаемого и ожидаемого значений для фиксированных квантилей ΔAI. Подобранная константа (черная линия) определяет QCC. Диаграммы размаха (справа) показывают тренд значений для всех корзин покрытия (слева). Элементы диаграмм размаха — центральная отметка: медиана; границы прямоугольника: верхние и нижние квартили; усы: 1.5x интерквартильный интервал; точки: выбросы.

2.2.4 Оценка избыточной дисперсии AI из наблюдаемых и смоделированных данных

Чтобы оценить избыточную дисперсию между парой реплик библиотек РНК-секвенирования, мы смотрим на то, насколько экспериментально наблюдаемое значение дисперсии отличается от ожидаемого значения в предположении подобранной модели. Вместо того, чтобы подбирать параметры для всех уровней покрытия одновременно при помощи комбинации отрицательного биномиального и бета-биномиального распределений (как это сделано в [39; 60; 61; 72; 73]), мы разделяем диапазон возможных покрытий на интервалы и распределяем гены по получившимся интервалам. Набор генов с покрытиями, лежащими в одном и том же интервале, здесь и далее мы будем называть корзиной генов. Разбиение генов на корзины мы активно используем для дальнейшего анализа (Рис.2.3).

При анализе каждой корзины, мы используем непрерывное распределение значений AI вместо тримодальной классификации ($AI = \{0, 0.5, 1\}$), которая обычно используется в бета-биномиальных моделях избыточной дисперсии аллельного дисбаланса [39; 72; 73]. Чтобы подобрать экспериментально наблюдаемое распределение AI, мы используем смесь бета-распределений, которая должна лучше описывать распределение AI с тяжелыми хвостами (Рис.2.3a). Для измерения экспериментально наблюдаемой дисперсии, мы делаем квантильный анализ распределения значений ΔAI в корзине покрытий, где ΔAI — это разница в значениях AI для гена между двумя репликами (Рис.2.3b).

Для оценки избыточной дисперсии, мы нормализуем наблюдаемые квантили значений на ожидаемые квантили. Заметим, что гены с разными AI имеют разный вклад в общую дисперсию сигнала (см. Раздел A.1.3). Поэтому модель должна рассматривать распределение AI в каждой конкретной корзине.

Для моделирования ожидаемого распределения ΔAI в каждой корзине покрытий и подсчёта соответствующих квантилей мы проводим следующую процедуру. Для конкретной корзины, мы подбираем смесь бета-биномиальных распределений к наблюдаемому распределению AI по генам (Рис.2.3c, сверху). Используя подобранные параметры этой модели, мы симулируем две реплики РНК-секвенирования (Рис.2.3c, посередине). Ожидаемое распределение ΔAI приходит из предположения о биномиальной выборке аллелей в этих двух симулированных репликах. Наконец, мы считаем квантили для ожидаемых распределений ΔAI (Рис.2.3c, снизу), и находим отношение наблюдаемых квантилей к ожидаемым.

Это частное наблюдаемых квантилей ΔAI к ожидаемым оказывается постоянным (с некоторыми случайными флуктуациями). Для двух реплик, показанных на Рис.2.3d, это частное равно 1.73 ± 0.18. Константа зависит от эксперимента. Идеальная Пуассонова выборка соответствует отсутствию избыточной дисперсии и значению частного 1; в данных экспериментов мы ожидаем, что частное больше единицы. Мы называем это подобранное частное, зависящее от эксперимента, коэффициентом коррекции качества (QCC).

2.2.5 Применение QCC повышает согласованность между репликами

Когда для пары реплик посчитано значение QCC, его можно напрямую использовать для коррекции дисперсии сверх биномиальной. Для того, чтобы учесть расширение распределения AI, количество аллельных прочтений делится на QCC^2. Полученные значения получаются нецелыми, поэтому мы использовали пропорциональный тест, который позволяет проводить анализ на нецелых значениях (детали см. в Методах). Все "QCC-скорректированные" тесты, описанные ниже, также включают в себя коррекцию Бонферрони для всех проанализированных генов, чтобы учесть множественное тестирование.

Значение QCC отражает "качество" данных в смысле согласованности результатов тестов на аллельный дисбаланс между репликами. Эксперименты с более низкими долями согласованности давали большие значения QCC (сравните Рис.2.2d с Рис.2.2b и Рис.2.2e). Значения QCC близки для всех пар реплик внутри одного эксперимента (Рис.2.2d), что является сильным свидетельством в пользу того, что QCC является свойством эксперимента, а не конкретной реплики.

Коррекция на QCC ведёт к повышенной согласованности между репликами внутри одного эксперимента и между экспериментами (Рис.2.2e), показывая, что QCC учитывает большую долю избыточной дисперсии, приходящей из эксперимента. Заметим, что одна из шести реплик в Эксперименте 1 похожа на выброс, так как имеет большие значения QCC при сравнении с остальными репликами. После коррекции на QCC, эта разница в согласованности сильно уменьшается.

Заметим, что после коррекции на QCC, рассогласованных решений становится сильно меньше в абсолютном значении, и их количество начинает сходится в разных экспериментах (Рис.2.2f). Более того, рассогласованные решения распределены близко к границе между значимым и не значимым смещением, где значимость определяется скорректированным на QCC статистическим тестом (сравните Рис.2.2f и 2.1b), из чего следует, что наблюдаемые данные лучше подходят под скорректированную модель шума (см. Раздел A.1.2). Коррекция QCC уменьшает количество дисбалансных по результатам тестов генов и

при этом сильно увеличивает согласованность между парами реплик библиотек (Рис.2.2g). В более шумных наборах данных (например, в Эксперименте 3) количество генов, дисбалансных по результатам тестов, уменьшилось сильнее (Рис.2.2g и Рис.A.21c), что ожидаемо при соответственном уменьшении доверия к данным.

Отметим, что мы применяем коррекцию на QCC после объединения всех прочтений из обоих реплик, чтобы полностью использовать имеющиеся данные. При этом важно, что основные улучшения в уровне согласованности происходят именно из-за использования коррекции на QCC, а не простого объединения данных из реплик (Рис.2.2g).

2.2.6 Применение QCC улучшает дифференциальный анализ аллельного дисбаланса

Весь обсуждаемый до этого анализ ставил целью бинарную классификацию генов на имеющие сбалансированную и смещённую аллельную экспрессию, в зависимости от результата тестирования гипотезы H_0, что AI = 0.5. Более тонкий вопрос — количественная оценка аллельного дисбаланса гена (или другого интересующего нас региона). Этот вопрос подразумевает вычисление точечной оценки AI и доверительного интервала (CI) для настоящей пропорции. Учёт избыточной дисперсии образца позволяет проводить более точный дифференциальный анализ AI при сравнении двух или более образцов.

Наши данные предоставляют возможность основательного анализа ложноположительности результатов: мы знаем, что реплики сделаны из одной РНК, и мы ожидаем ноль отвергающих нулевую гипотезу решений после поправки на множественное тестирование, поэтому любые решения об отвержении нулевой гипотезы являются ложноположительными. Когда анализ проводится под биномиальными предположениями (с поправкой Бонферрони на множественное тестирование), сравнения между репликами библиотеки РНК-секвенирования и между экспериментами показывают сотни генов со значимым дифференциальным AI (Рис.2.4a, слева). Похожие результаты получаются при применении некоторых существующих инструментов на тех же данных [7; 61] (Рис.A.22).

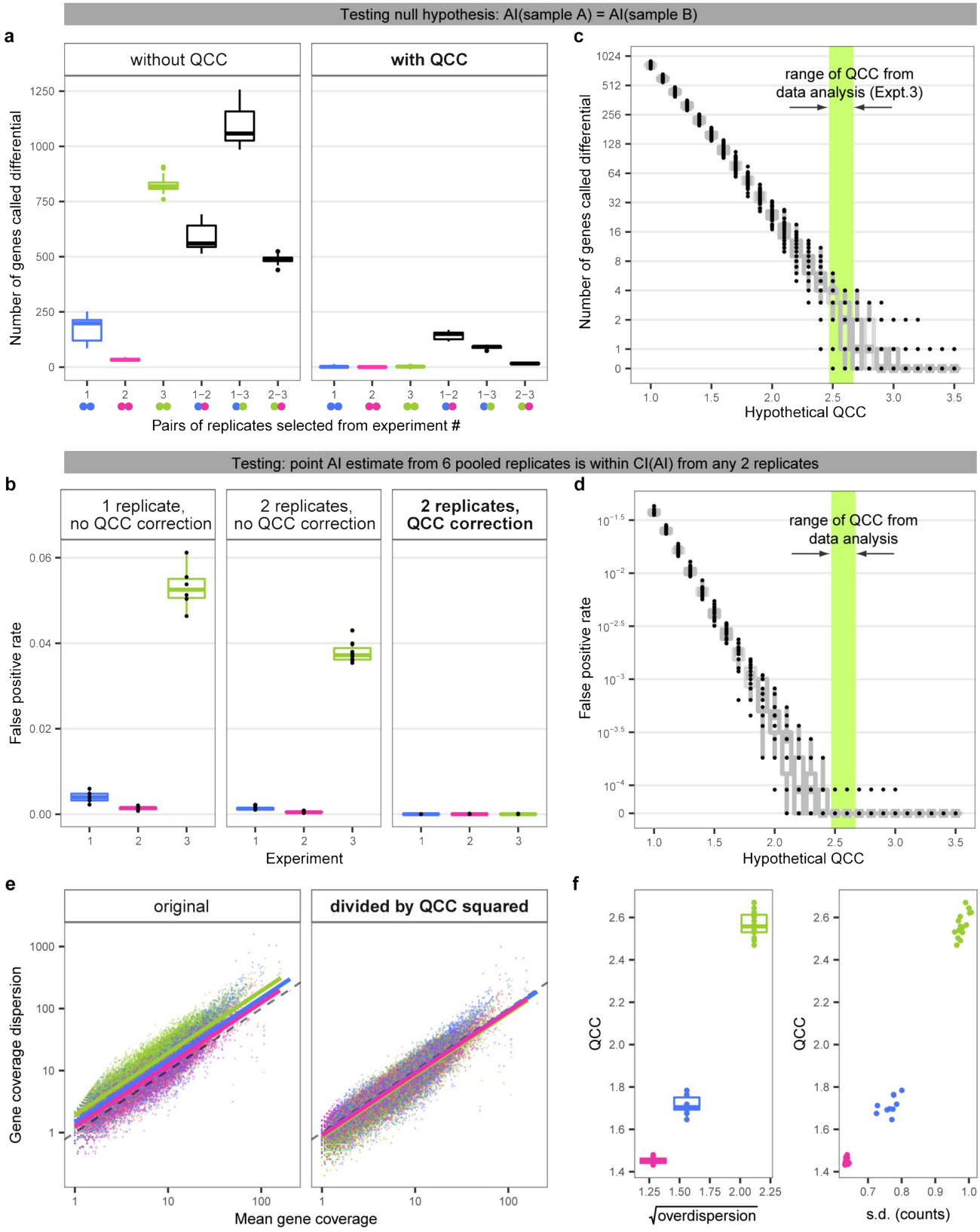

Рисунок 2.4 — QCC позволяет совершать дифференциальный анализ AI, и находится в прямой зависимости с избыточной дисперсией покрытий.

a: Количество генов с ложноположительной дифференциальной аллельной экспрессией, до (слева) и после коррекции на QCC (справа). В каждом из трёх экспериментов были сделаны все 45 сравнений пар. Для сравнений между экспериментами были случайно выбраны 45 из возможных 225 комбинаций. Для каждой

пары были посчитаны значения QCC и идентифицированы гены с значимо разным AI. Эксперименты: 1 (синим), 2 (красным) и 3 (зелёным). Элементы диаграмм размаха как и в Рис.2.2. Здесь и в (b), (c), (d) были использованы двусторонние тесты. **b:** Влияние QCC на долю ложноположительных существенных отличий точечной оценки AI от "золотого стандарта" AI, оцененного по всем репликам: "золотая" оценка AI не лежит в доверительном интервале одной реплики (слева), двух объединённых реплик (посередине), и двух реплик с коррекцией на QCC (справа). Реплика-выброс из Эксперимента 1 была изъята. **c, d:** Вычисленные значения QCC (содержащиеся внутри окрашенных рамок) близки к оптимальному балансу между излишним количеством ложноположительных результатов и потерей сигнала. **c:** Количество ложноположительных результатов (дифференциальный AI в репликах из одной РНК) посчитано для разных потенциальных значений QCC для всех возможных 45 комбинаций двух пар реплик из Эксперимента 3. **d:** Доля ложноположительных генов (как и в (b)) для различных возможных значений QCC и всех возможных пар реплик в Эксперименте 3. **e:** Избыточная дисперсия покрытий генов в разных экспериментах пропорциональна QCC^2. Слева: избыточная дисперсия может быть подобрана как лог-линейная регрессия (сплошные линии) для каждого эксперимента. Они оказываются выше ожидаемой Пуассоновской дисперсии (пунктирная линия). Справа: то же, но после деления избыточной дисперсии на QCC^2. Реплика-выброс из Эксперимента 1 была убрана из (e) и (f). **f:** Корреляция QCC и избыточной дисперсии покрытий генов. Значения QCC такие же, как и в (e). Избыточная дисперсия была посчитана как экспонента смещения лог-линейной регрессии (см. (e)) между средним и дисперсией общего числа прочтений. Слева: для всех реплик в эксперименте; справа: для всех возможных пар реплик.

Коррекция на QCC, напротив, полностью убирает ложноположительные результаты из сравнений внутри отдельно взятого эксперимента (Рис.2.4a, справа). Количество ложноположительных результатов в межэкспериментальных сравнениях резко уменьшилось после коррекции на QCC, но не до нуля (Рис.2.4a). Это говорит о том, что QCC-коррекция может быть использована для сравнения AI между экспериментами, но систематические различия в протоколах могут привести к некоторому количеству ложноположительных результатов.

Мы также задались вопросом, насколько коррекция на QCC лучше представляет распределения шума, чем простая биномиальная модель. Мы посчитали количество случаев, когда точечная оценка AI, полученная из шести объединённых реплик ("золотой стандарт" AI), не лежала внутри доверительного интервала оценки AI, посчитанного одним из трёх следующих способов. Рис.2.4b показывает доли ложноположительных отличий оценок AI от "золотого стандарта" для интервалов, полученных из одной реплики с биномиальными предположениями (слева); из объединённых пар реплик с биномиальными предположениями (посередине); и подправленных на QCC объединённых пар реплик (справа). Ожидается (см. Раздел A.1.4), что после поправки Бонферрони количество ложноположительных результатов станет близким к нулю. Среди протестиро-

ванных подходов, только коррекция на QCC существенно уменьшила долю ложноположительных результатов. Другие инструменты в этом анализе также показали большое количество ложноположительных результатов (Рис.A.22b,c).

После этого мы решили выяснить, приходит ли это снижение количества ложноположительных результатов вместе с излишне консервативной коррекцией шума. Рис.2.4c,d показывает, что вычисленные значения QCC находятся рядом с участком, где доля ложноположительных результатов достигает плато в 0, и не располагаются сильно дальше от этого плато. Мы интерпретируем это как то, что баланс между точностью и полнотой близок к оптимальному (также см. Рис.A.29d). Баланс между количеством ложноположительных результатов и сигнала в дифференциальном анализе AI исследуется в деталях в Разделе A.1.5. В этом Разделе, мы проводим анализ статистической силы тестов и показываем, что потери сигнала, внесённые использованием коррекции на QCC, могут быть компенсированы дополнительным секвенированием. Напротив, внесённые простым, не подправленным на избыточную техническую дисперсию биномиальным тестом ложноположительные результаты не могут быть поправлены дополнительным покрытием.

Мы заключаем, что в дифференциальном анализе AI отсутствие коррекции на избыточную дисперсию склонно приводить к большому количеству ложноположительных результатов, и коррекция на QCC эффективно решает эту проблему. В результате анализа экспериментальных и симулированных данных мы также заключаем, что коррекция на QCC не приводит к излишне консервативному тестированию.

2.2.7 Источники избыточной дисперсии AI: анализ данных

Чтобы найти возможные источники избыточной дисперсии AI, мы рассмотрели этапы эксперимента РНК-секвенирования (Рис.A.23a): (1) шаги от биологического объекта до выделения РНК, включительно; (2) формирование библиотеки РНК-секвенирования; (3) процесс секвенирования библиотеки; (4) обработка данных секвенирования, от выравнивания прочтений до статистического анализа аллельного дисбаланса.

Вклад этапа (1) был исключён исходя из конструкции нашего эксперимента: все 18 реплик библиотек были приготовлены из одной общей РНК почки мыши. Взятие однородных аликвот очищенной РНК может считаться справедливым Пуассоновым процессом выборки (это не так в одноклеточных экспериментах, где возникают дополнительные источники шума, такие как транскрипционные всплески [95; 96]).

Этап (4), анализ данных, включает множество шагов, и мы можем оценить их вклад в избыточную дисперсию AI (Рис.A.23b,c, Рис.A.22). Сперва заметим, что эти шаги, взятые вместе, не дают сильного вклада в изменчивость между экспериментами, так как применение вычислительного протокола на идентичных данных даёт консистентные значения AI и QCC (если не брать в расчёт шум от процедур симуляции).

Заметим, что процедуры, используемые для подсчёта аллельных прочтений, могут влиять на избыточную дисперсию AI. Например, несколько популярных инструментов для анализа данных РНК-секвенирования (включая Kallisto [10], Salmon [9] и RSEM [74]) используют не только прочтения, покрывающие SNP, но позволяют распределять остальные не информативные прочтения между двумя аллелями. Такое присвоение гаплотипов приводит к линейному приросту в покрытии, но квадратичному приросту в вариации, и поэтому к увеличению избыточной дисперсии. Применение одного из таких инструментов [10] показывает систематическое увеличение избыточной дисперсии по сравнению с учётом только прочтений, покрывающих SNP (Рис.A.24).

Мы рассмотрели вклад процедуры вычисления QCC в оценку избыточной дисперсии AI. На симулированных общих аллельных прочтениях с известной избыточной дисперсией, значения QCC были близки к ожидаемым (Рис.A.25; обозначены как i на Рис.A.23c). Это влечёт вывод, что вычисления QCC сами по себе не дают большого вклада в шум. Соответствующий анализ, начинающийся со случайной биномиальной выборки из данных секвенирования одной реплики (ii на Рис.A.23c), должен показывать только избыточную дисперсию, связанную с подсчётом аллельных покрытий и процессом вычисления QCC. Такой анализ выдал значения QCC в диапазоне 1.01–1.04 (Рис.A.26b), то есть практически отсутствующую избыточную дисперсию (QCC ~ 1.0).

Заметим, что при случайном разделении парноконцевых прочтений из одного прогона на две равные части (сравните с Рис.2.2c, посередине), получивши-

еся “полуреплики” не находятся в биномиальном соотношении. В этих сравнениях, (*iii* в Рис.A.23c), значения QCC лежали в диапазоне 1.45–1.48 (Рис.A.26), отражая дисперсию, приходящую из данных одного прогона секвенирования одной библиотеки.

Хорошо задокументировано, что не аллель-специфические прочтения в РНК-секвенировании показывают избыточную дисперсию сверх биномиальной [69—71]. Мы задались вопросом о том, как эта избыточная дисперсия связана с избыточной дисперсией AI, измеренной QCC. Избыточная дисперсия видна во всех трёх экспериментах: подобранный лог-линейный тренд находится выше ожидаемой Пуассоновой дисперсии (Рис.2.4d, слева). Более того, избыточная дисперсия отличалась во всех трёх экспериментах. Когда значения дисперсии для каждого гена были разделены на QCC^2, регрессии для всех экспериментов практически совпали друг с другом и с Пуассоновой ожидаемой дисперсией (Рис.2.4d, справа). Избыточная дисперсия положительно коррелирует со значениями QCC (Рис.2.4e). В симуляциях QCC также сильно коррелировал с установленными значениями избыточной дисперсии (Рис.A.25). Основываясь на этом анализе, мы выдвигаем гипотезу о том, что не аллель-специфическая избыточная дисперсия и аллель-специфическая избыточная дисперсия приходят в основном из одних процессов. Несмотря на то, что эти дисперсии сильно коррелировали в обозначенных примерах, вычисление коэффициента QCC является более устойчивой процедурой, не зависящей от этой корреляции и процедуры обработки данных.

2.2.8 Источники избыточной дисперсии AI: эксперименты

Анализ двух прогонов секвенирования одной библиотеки (*iv* в Рис.A.23c) дал значения QCC, похожие на значения при выборке *in silico* из той же библиотеки (Рис.A.26). Из этого можно сделать вывод, что дополнительный прогон секвенирования даёт результаты, похожие на увеличение количества прочтений в оригинальном прогоне (сравните *ii* и *iv* на Рис.A.23c). Это согласуется с аргументом о том, что процесс секвенирования не даёт большого вклада в вариацию

избыточной дисперсии AI между репликами. Для более определённого заключения необходимо большее число экспериментов.

Коэффициент QCC оказался сильно больше для сравнений между репликами, чем для сравнений между половинами реплик (v на Рис.A.23c), что показывает, что приготовление библиотеки вносит дополнительный шум. Заметим, что это в очередной раз подчёркивает, что анализ с одной репликой не позволяет корректно оценить и исправить избыточную дисперсию AI.

Итого, биологический шум (этап 1) исключён по конструкции эксперимента, этапы 3 и 4 не вносят сильного шума, и главным источником избыточной дисперсии остаётся подготовка библиотеки (этап 2). Это подкрепляется наблюдением о том, что разные процедуры (протоколы в целом или разные начальные количества РНК) в наших Экспериментах 1–3 произвели данные с сильно отличными друг от друга значениями QCC, тогда как технические реплики в рамках каждого эксперимента дали похожие значения (см. Рис.2.2d).

Формирование библиотеки РНК-секвенирования включает в себя множество шагов, от обратной транскрипции и фрагментации кДНК до амплификации фрагментов, и эти шаги могут значительно отличаться между протоколами. Детальный анализ конкретных протоколов выходит за рамки этой главы и этой работы. Однако отметим, что одна широко известная проблема экспериментов высокопроизводительного секвенирования — это влияние артефактов ПЦР амплификации [76—78]. Чтобы оценить влияние артефактов амплификации на избыточную дисперсию AI, мы сравнили результаты анализа до и после удаления прочтений-дубликатов. Дедупликация не уменьшила значения QCC до близких к единице, а в некоторых случаях привела к увеличению QCC (Рис.A.27). Это показывает, что основной вклад в избыточную дисперсию AI дают другие источники. Дедупликация может привести к потере большого количества легитимных данных, и может иметь другие нежелательные эффекты, такие как искажение распределения сигнала биологического образца [76—78]. Поэтому с практической точки зрения дедупликация прочтений имеет ограниченную полезность, а её влияние на избыточную дисперсию AI в целом учтено в анализе QCC. Заметим, что в парноконцевых данных РНК-секвенирования, длина фрагмента кДНК является подобием уникального молекулярного идентификатора (UMI) [97]. Поэтому результаты дедуплицирования в парноконцевых

данных (Рис.A.27d) позволяют предположить, что использование UMI может не устранить всю избыточную дисперсию AI.

В конечном итоге, наблюдения выше говорят о том, что формирование РНК-библиотеки является наиболее вероятным источником избыточной эксперимент-специфической дисперсии аллельного дисбаланса, и техническая вариабельность не может быть объяснена артефактами ПЦР.

2.3 Обсуждение результатов

Мы показали, что если используются сравнительно реалистичные модели сигнала и шума в аллельном дисбалансе, то данные из одной библиотеки РНК-секвенирования недостаточны для надёжного количественного анализа вклада технического шума в наблюдаемый сигнал аллельного дисбаланса (см. Рис.2.1 и Раздел A.1.1). Для изучения природы экспериментального шума, мы произвели 18 библиотек РНК-секвенирования, приготовленных из одной РНК при помощи двух различных протоколов и трёх различных начальных количеств тотальной РНК. Анализ этих данных и существующих наборов данных из открытых источников показал, что уровень технического шума может отличаться между экспериментами в несколько раз. Это демонстрирует, что предположение о существовании единой и универсальной модели шума для всех экспериментов РНК-секвенирования неверно.

Мы разработали вычислительный подход, который сравнивает две и более технических реплики, и реализовали его в пакете `Qllelic` (github.com/gimelbrantlab/Qllelic) на языке `R`. Этот подход концептуально прост; его использование эквивалентно применению биномиальной модели, где количество аллельных прочтений уменьшено на коэффициент QCC, возведённый в квадрат. Коэффициент QCC, коэффициент коррекции качества, может быть оценён из сравнения технических реплик. Несмотря на простоту, подход `Qllelic` позволяет хорошо приблизить наблюдаемые данные (см. Рис.2.2f и Рис.A.28,A.29,A.30,A.31). Применение Qllelic уменьшает доли ложноположительных результатов, которые получаются из-за недоучтённой технической ва-

риации, практически до нуля, при этом сохраняя сигнал аллельного дисбаланса (см. Рис.2.4c,d, Рис.A.29d и Раздел A.1.5).

Этот подход имеет гораздо более хорошие показатели, чем биномиальный тест с поправкой на множественное тестирование (который широко используется в опубликованных работах, например [89—92]) и методы, которые оценивают избыточную дисперсию из данных одной технической реплики [7; 61] (Рис.A.22). Стоит заметить, что подбор коэффициента избыточной дисперсии ρ в бета-биномиальной модели, единого для целого эксперимента и не зависящего от покрытия гена(например [60; 61; 73]), предполагает, что избыточная дисперсия возрастает вместе с покрытием гена (Рис.A.32). Однако в экспериментальных данных мы наблюдаем, что избыточная дисперсия практически постоянна на всех уровнях покрытия генов (см. Рис.2.3d), что согласуется с моделью QCC.

Важно отметить, что использование подхода QCC/`Qllelic` позволяет проводить точный дифференциальный анализ AI. Также он делает сравнение образцов между исследованиями и экспериментами; значения QCC могут быть предподсчитаны, делая дифференциальный анализ AI быстрым и удобным. В то время как наш анализ сфокусирован на генах, избыточная дисперсия AI в эксперименте видна уже на уровне отдельных SNP (Рис.A.33).

Ниже мы описываем разобранные примеры с двумя типичными вопросами в аллель-специфическом анализе.

Случай 1: оценка значений AI в образце. Для дальнейшего аналитического использования, точечные оценки AI должны быть сопровождены доверительными интервалами. Разобранный пример (Раздел A.1.6) демонстрирует процедуру оценки AI, подсчёт избыточной дисперсии при помощи `Qllelic`, с использованием данных двух или более технических реплик РНК-секвенирования, и, наконец, вычисление доверительных интервалов для каждой оценки AI. Эта процедура также включает тестирование нулевой гипотезы о существенности отличия наблюдения от определённого значения AI (например, $H_0 : \mathrm{AI} = 0.5$, $p = 0.05$); поправка Бонферрони на множественное тестирование применяется ко всему списку генов с оценёнными AI и CI(AI). Тестируемое значение AI может быть одним и тем же для всех генов (например, $\mathrm{AI} = 0.5$) или быть для каждого гена своим (например, при сравнении с некоторым “золотым стандартом” в экспрессии).

Если доступно более двух технических реплик, значения QCC сначала подсчитываются попарно (см. Методы). Это помогает идентифицировать реплики-выбросы, если такие есть, для последующего исключения из рассмотрения. Данные реплик объединяются, чтобы определить точечную оценку AI, а среднее всех попарных значений QCC используется как значение QCC для данного образца. Заметим, что в сравнениях могут участвовать реплики или выборки из реплик с сопоставимым суммарным количеством прочтений, ограниченным сверху репликой с наименьшим размером библиотеки.

Случай 2: дифференциальный анализ AI между двумя образцами. Знание точечных оценок AI и избыточной дисперсии в обоих образцах позволяет нам проводить дифференциальный анализ AI. Чтобы сделать дифференциальный тест на определённом уровне доверия, мы используем пропорциональный тест, позволяющий работать с нецелыми значениями (см. Методы; заметим, что непересечение двух доверительных интервалов даёт намного более строгий тест, чем того требует уровень доверия). Разобранный пример сравнивает аллельную экспрессию в двух клональных линиях мыши 129S1×CAST в Разделе A.1.6.

Невозможность оценить избыточный шум при анализе ASE из одной технической реплики вносит существенные трудности в интерпретацию результатов, полученных на некоторых существующих наборах данных. Лишь немногие опубликованные исследования экспериментов РНК-секвенирования включают в себя технические реплики. В исследовании человеческих клеток Geuvadis [93], только 5 образцов из 462 имели технические реплики, с значениями QCC в диапазоне 1.04-1.21. Коррекция на QCC уменьшает количество генов, определённых как имеющие аллельный дисбаланс, до полутора раз (см. Табл.3). Учитывая то, что избыточная дисперсия может варьироваться даже в наборе реплик (см. Эксперимент 1 в Рис.2.2), мы должны быть осторожны при работе с оценками AI, полученных на данных, где значение QCC не может быть восстановлено.

Чтобы проиллюстрировать влияние варьирования значений избыточной дисперсии на интерпретацию существующих наборов данных, рассмотрим данные случайно выбранного индивида из исследования GTEx [89] (Рис. A.34). В исследовании был применён биномиальный тест, который предполагает QCC = 1. В данном случае, применение биномиального теста идентифицирует 121 ген с отвергнутой нулевой гипотезой (AI = 0.5) для ткани печени и 96 для ткани

лёгкого. При отсутствии реплик, действительная величина избыточной дисперсии в этом эксперименте неизвестна. Если бы мы предположили QCC = 2, дисбалансных генов осталось бы 28 и 20 соответственно, при QCC = 3 – 19 и 11. Другими словами, размах вариации, который мы наблюдали при использовании популярных коммерческих наборов для подготовки библиотеки, может влиять даже на порядки количества находок в результатах простого анализа.

Мы выяснили, какие какие вычислительные и экспериментальные шаги способны вносить техническую избыточную дисперсию в данные. Вычислительные процедуры для подсчёта аллельных прочтений могут повлиять на анализ как и без увеличения избыточной дисперсии (например, консистентное предубеждение в пользу референса при картировании), так и с увеличением избыточной дисперсии. Процедура подсчёта (см. Методы), которую мы используем для создания таблиц аллельных покрытий, подаваемых на вход `Qllelic`, борется с предубеждением в пользу референса при помощи картирования на два синтетических родительских псевдогенома с вставленными однонуклеотидными полиморфизмами. Заметим, что предложенный в этой главе вычислительный протокол для создания таблиц аллельных покрытий генов можно дополнительно улучшить, перейдя от суммирования значений покрытий над отдельными SNP в гене (этот подход подразумевает независимость наблюдений, что неверно для организмов с высокой плотностью полиморфизмов) к количеству самих прочтений (см. главу 4 для подробностей).

Экспериментальный анализ выявил, что процесс формирования библиотеки секвенирования вносит основной вклад в избыточную дисперсию, тогда как дополнительный шум, приходящий с этапа секвенирования, может быть несущественен. Интересно отметить, что вычислительная дедупликация прочтений (удаление потенциальных артефактов ПЦР-амплификации, имеющих одинаковые координаты при выравнивании) не устраняет, а иногда и увеличивает избыточную дисперсию (см. Рис.A.27). Несмотря на систематическую разницу между протоколами подготовки библиотеки РНК-секвенирования (Эксперименты 1, 2 и 3), вариации между экспериментами с одним протоколом всё ещё могут быть существенны (см. реплику-выброс в Эксперименте 1 и вариацию QCC на Рис.A.26b,c). Поэтому мы рекомендуем иметь 2 реплики для каждого образца (заметим, что обнаружение выбросов можно оценить с достаточной долей уверенности только при трёх репликах, из-за парности подхода подсчёта QCC).

Некоторые инструменты анализа данных РНК-секвенирования [9; 10; 74] могут присваивать гаплотип не аллель-информативным прочтениям (то есть прочтениям, не покрывающим SNP при выравнивании), распределяя их согласно статистикам, полученных с информативных прочтений. Эта процедура ведёт к большому и разнородному увеличению избыточной дисперсии AI (см. Рис.A.24). Мы рекомендуем проявлять осторожность при использовании подобных процедур.

Использование биологических реплик в исследованиях РНК-секвенирования встречается намного чаще, чем технических, особенно когда доступно много родственных особей (например, мышей). Биологические реплики формально могут быть использованы в вычислении QCC для последующей коррекции тестов; например, в исследовании аллель-специфической экспрессии в клетках мыши [29], применение анализа `Qllelic` к двум образцам с доступными биологическими репликами дало QCC в диапазоне 1.51–1.56 (Табл.4). Эта процедура всё ещё даёт более точный учёт комбинации из биологической и технической вариаций, чем объединение или усреднение данных по биологическим репликам. Однако следует помнить, что свойства биологической вариации существенно отличаются от свойств технического шума (в частности, предположение об униформности биологической вариации для любого участка генома, очевидным образом, неверно), что может привести к ошибочным оценкам избыточной дисперсии. Поэтому технический шум не может быть полноценно и надёжно учтён при работе с биологическими репликами.

2.4 Материалы и методы

2.4.1 Подготовка РНК и библиотек РНК-секвенирования

Тотальная РНК была выделена с помощью Trizol из свежесобранной ткани почек взрослой самки мыши (гибрид 129S1×CAST/Ei), содержавшихся в мышином питомнике Института рака Дана-Фарбер (DFCI), источник животных - Jackson Laboratories (Bar Harbor). Все работы с животными проводились

в соответствии с протоколом DFCI 09-065, одобренным Комитетом по уходу и использованию животных DFCI. Животные содержались в соответствии с Руководством по уходу и использованию лабораторных животных). Целостность РНК оценивали с помощью Bioanalyzer, а ее количественное содержание – с помощью прибора Qubit. Аликвоты общей РНК были использованы для приготовления трёх наборов реплик библиотек, все из которых начинались с выделения поли-А РНК: 6 библиотек с помощью набора NEBNext, с изначальным количеством тотальной РНК – 100 нг; 6 библиотек с помощью набора SMARTseq v4, начиная с 10 нг РНК; и то же самое, с 0.1 нг РНК. Все библиотеки были просеквенированы на приборе HiSeq2500 в центре секвенирования DFCI, в соответствии с инструкциями производителей.

Для создания набора данных, рассмотренного в примере дифференциального анализа AI, клональные клеточные линии Abl.1 и Abl.2 лимфобластоидных клеток Абельсона из гибрида мыши (129S1×CAST/Ei) [24] выращивали в среде RPMI (Gibco), содержащей 15% FBS (Sigma), 1X L-Глутамин (Gibco), 1X Пенициллин/Стрептомицин (Gibco) и 0.1% β-меркаптоэтанол (Sigma). Тотальную РНК выделяли с помощью магнитных частиц, Sera-Mag SpeedBeads (GE Healthcare). Выделенную тотальную РНК обрабатывали ДНКазой RQ1 DNase (Promega). Библиотеки для секвенирования РНК готовили с помощью набора SMARTseq v.4 (Takara), начиная с 10 нг общей РНК для каждой реплики. Секвенирование проводили на приборе HiSeq4000 в Novogene, Inc.

2.4.2 Дополнительные источники данных

Набор данных из работы Geuvadis включает данные РНК-секвенирования лимфобластоидных клеточных линий, полученных от 462 человек из пяти популяций [93]. Файлы FASTQ с парноконцевыми чтениями (2 × 75 bp) для 5 индивидов (HG00117, HG00355, NA06986, NA19095, NA20527), с 7 репликами для каждого образца, были загружены с сайта проекта 1000 Genomes (ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/). Данные аллельных покрытий (обработанные с помощью стандартного вычислительного протокола GTEx) для случайно выбранного образца GTEX-11NUK

из промежуточной фазы проекта GTEx были загружены из dbGaP (ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000424.v7.p2). Мы также использовали данные РНК-секвенирования из клеток-предшественников нейронов мыши (GSE54016) (ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE54016).

2.4.3 Вычислительный протокол получения оценок AI

Инструменты для оценки AI, описанные здесь, разделены на две части по назначению, и реализованы по отдельности. Обработка данных, от выравнивания прочтений до подсчёта аллельных покрытий генов, эволюционировала из функции `ASEReadCounter` в пакете GATK [6]. Соответствующие функции взяли в основу программу на языке `Python`, написанную S. Castel (github.com/secastel/allelecounter), и были потому названы `ASEReadCounter*` (github.com/gimelbrantlab/asereadcounter_star). Вычисления QCC, оценка доверенных интервалов и дифференциальный анализ AI реализованы в методе `Qllelic` (github.com/gimelbrantlab/Qllelic).

Подготовка референса: Для картирования были использованы два родительских "псевдогенома" [98; 99], см. `ASEReadCounter*`. Для гибрида 129S1×CAST/Ei, аллели определены хорошо изученными геномными вариантами родительских линий. Для человеческих образцов из работы Geuvadis [93] были использованы фазированные SNP. Соответствующие аллельные варианты из базы данных однонуклеотидных полиморфизмов dbSNP142 [100] или структурные варианты из фазы 3 проекта 1000 геномов были вставлены в референсный геном (GRCm38.86 или hs37d5, 1000 геномов, фаза 2) – так "родительских" псевдогеномов для дальнейшего анализа (пример см. в Разделе A.1.6). Для каждого организма был создан VCF файл, в котором один аллель рассматривался как референсный (материнский 129S1 или аллель 1), а другой – как альтернативный, для оценки ASE использовались только гетерозиготные сайты.

Подсчёт аллельных прочтений.

Выравнивание: Прочтения библиотеки РНК-секвенирования были выравнены при помощи программы `STAR` (v.2.5.4a) [101] на каждый из двух псевдогеномов, с порогами на качество выравнивания по умолчанию. Были использованы только уникально выравненные прочтения (параметр `-outFilterMultimapNmax 1`).

Присваивание аллели: Прочтения, которые были выравнены только на одну аллель, и прочтения, которые имели лучшее качество выравнивания на одну из аллелей, были определены к соответствующей аллели. Оставшиеся прочтения (не пересекающие гетерозиготные позиции SNP) не были использованы в дальнейшем анализе. Эта часть процедуры взяла в основу идею метода, созданного S.Castel.

Дедупликация прочтений: При необходимости применения согласно схеме анализа, был использован инструмент `Picard` (v.2.8.0; broadinstitute.github.io/picard) `MarkDuplicates`.

Подвыборка библиотеки: Чтобы убедиться, что все выравненные прочтения лежат в похожих распределениях, из BAM файлов, соответствующих одному эксперименту, были сделаны подвыборки одного размера при помощи написанной под эту цель программы на `Bash`, с применением `shuf` для обеспечения случайности.

Аллельные прочтения для индивидуального SNP: Для каждой позиций гетерозиготных полиморфизмов из VCF файла (см. обсуждение пункта “Подготовка референса”) покрытие было посчитано при помощи `samtools mpileup` (v.1.3.1). Результаты были конвертированы в таблицу с количеством аллельный прочтенияй для каждого SNP. Эта процедура основана на `Python`-программе от S.Castel.

Аллельные прочтения для генов: Все экзоны, принадлежащие одному гену, были объединены, согласно аннотации из GTF файла (файлы RefSeq GTF, GRCm38.68 и GRCh37.63, были загружены из

Ensembl, ftp://ftp.ensembl.org/pub/release-68/gtf/ [102]). Регионы, принадлежащие нескольким генам, были исключены. Фазированные прочтения для всех SNPs внутри одного гена были просуммированы:

$$\begin{aligned} M_g &= \sum_{\mathrm{SNP}\in g} M_{\mathrm{SNP}} \\ C_g &= \sum_{\mathrm{SNP}\in g} M_{\mathrm{SNP}} + P_{\mathrm{SNP}} \end{aligned} \tag{2.1}$$

Если не указано иное, для дальнейшего анализа были использованы только гены с аллельным покрытием $\geq$10.

Оценка аллельного дисбаланса: Оценка AI для гена g была получена как пропорция материнских прочтений гена (M_g) к общему количеству аллельных прочтений гена ($C_g = M_g + P_g$):

$$\mathrm{AI}_g = \frac{M_g}{C_g} \tag{2.2}$$

Дополнительные инструменты для подсчёта аллельного дисбаланса. В нашем сравнительном анализе мы использовали три инструмента: `Qllelic` (v0.3.2), `MBASED` (v1.20.0) и `GeneiASE` (v1.0.1). Для единообразности, входные данные для сравнений были предобработаны одинаково для всех инструментов. В случае действительных данных, одни и те же гены были отфильтрованы так, чтобы входные данные удовлетворяли всем требованиям инструментов на количество и покрытия SNP (см. Рис.A.22). Ниже следуют параметры, которые были использованы при обработке инструментами (см. Рис.A.22,A.25b):

1. Анализ одного образца:

   `Qllelic`: `PerformBinTestAIAnalysisForConditionNPoint()`, параметры по умолчанию

   `MBASED`: `runMBASED()`, при `isPhased = TRUE, numSim = 10000`, и др. – по умолчанию

   `GeneiASE`: параметры по умолчанию функции `geneiase -t static`

2. Анализ двух образцов:

   `Qllelic`: `PerformBinTestAIAnalysisForTwoConditions()`, параметры по умолчанию

   `MBASED`: `runMBASED()`, при `isPhased = TRUE, numSim = 10000`, и др. – по умолчанию

   `GeneiASE`: параметры по умолчанию функции `geneiase -t icd`

2.4.4 Вычисление коэффициента коррекции качества для двух реплик

Так как покрытие гена является одним из параметров пропорциональной бета-биномиальной модели аллельного дисбаланса, мы начали со стандартной процедуры разбиения генов на корзины по покрытию для дискретизации нашей модели. Границы корзин были определены как округлённые степени основания $b = 1.05$: $\bar{C} = \{\lceil b^1 \rceil, \lceil b^2 \rceil, \lceil b^3 \rceil, \dots\}$. Заметим, что вычисления QCC несущественно зависят от изменений в размере корзин, см. Рис.A.25. Каждый ген g был отнесён к корзине, соответствующей среднему его прочтений C_{1g} и C_{2g} из двух технических реплик:

$$\forall g : \frac{C_{1g} + C_{2g}}{2} \in B_i = (\bar{C}_{i-1}, \bar{C}_i] \Rightarrow g \in G_i \ , \tag{2.3}$$

и каждая корзина B_i, содержащая множество генов G_i, была обработана отдельно.

Подбор распределения AI как смеси бета-биномиальных распределений: Чтобы подобрать параметры смеси двух пропорциональных бета-биномиальных распределений, представляющие наблюдаемый AI из выбранной реплики в каждой корзине по покрытию B_i:

$$a_i \sim \begin{cases} \frac{\text{Beta-Bin}(2\cdot\widehat{C}_i, \alpha_{1i}, \alpha_{1i})}{2\cdot\widehat{C}_i} \text{ , с вероятностью } \omega_{1i} \\ \frac{\text{Beta-Bin}(2\cdot\widehat{C}_i, \alpha_{2i}, \alpha_{2i})}{2\cdot\widehat{C}_i} \text{ , с вероятностью } \omega_{2i} \end{cases} \tag{2.4}$$

$$\widehat{C}_i = \sqrt{\bar{C}_{i-1} \cdot \bar{C}_i}$$

$$\omega_{1i} + \omega_{2i} = 1$$

$$\alpha_{1i} > 1, \quad \alpha_{2i} \in (0,1) \ ,$$

мы используем алгоритм максимизации ожидания (EM) (см. Рис.2.3d). Наша процедура подбора похожа на процедуру подбора в классической модели смеси Гауссовых распределений [103].

Для процедуры подбора на данных в этой главе, мы использовали порог на общее аллельное покрытие гена (50 для мыши и 30 для человека). Все корзины, которые не удовлетворили требованию на минимум 40 наблюдений (генов) были также убраны из всех этапов процесса подбора QCC.

Начиная со значений $\omega_{1i}^0 = \omega_{2i}^0 = 0.5$, $\alpha_{1i}^0 = 10$, $\alpha_{2i}^0 = \frac{1}{50}$, и вектора наблюдений аллельного дисбаланса $\{\mathrm{AI}_{\theta i}\}_{\theta \in \{1..N_i\}}$, где N_i — количество генов в корзине B_i:

$$x_{ni} = \mathrm{AI}_{ni} \cdot \widehat{C}_i \ , \tag{2.5}$$

мы производили последовательные шаги EM до тех пор, пока разница между параметрами на последовательных шагах не стала меньше пороговой (Рис.A.31).

Шаг E

$$\gamma_{nki}^t = \frac{\omega_{ki}^{t-1}\mathrm{BetaBin}(x_{ni} \mid 2\widehat{C}_i, \alpha_{ki}^{t-1}, \alpha_{ki}^{t-1})}{\sum_{j=\{1,2\}} \omega_{ji}^{t-1}\mathrm{BetaBin}(x_{ni} \mid 2\widehat{C}_i, \alpha_{ji}^{t-1}, \alpha_{ji}^{t-1})} \tag{2.6}$$

для $k \in \{1,2\}$, $n \in \{1,\ldots,N_i\}$, где t — номер шага EM.

Шаг M Так как мы ожидаем, что $\mu = \widehat{C}_i$, и бета-биномиальные распределения симметричны:

$$\omega_{ki}^t = \frac{1}{N_i}\sum_{n=1}^{N_i}\gamma_{nki}^t$$

$$\Sigma_{ki}^t = \frac{\sum_{n=1}^{N_i}\gamma_{nki}^t \cdot (x_{ni} - \widehat{C}_i)^2}{\sum_{n=1}^{N_i}\gamma_{nki}^t}$$

$$\Sigma_{ki}^t = \frac{2\widehat{C}_i \cdot {\alpha_{ki}^t}^2 \cdot (2\alpha_{ki}^t + 2\widehat{C}_i)}{4{\alpha_{ki}^t}^2 \cdot (2\alpha_{ki}^t + 1)} = \frac{4 \cdot \widehat{C}_i \cdot \alpha_{ki}^t + (2\widehat{C}_i)^2}{8 \cdot \alpha_{ki}^t + 4} \Longrightarrow \alpha_{ki}^t = \frac{(2\widehat{C}_i)^2 - 4\Sigma_{ki}^t}{8\Sigma_{ki}^t - 4\widehat{C}_i} \tag{2.7}$$

Симуляция пары реплик Используя подобранную тройку параметров $\{\omega_{1i}, \alpha_{1i}, \alpha_{2i}\}$, в каждой корзине B_i мы создали выборку из взвешенной смеси

двух бета-распределений вероятностей $\{p_{\theta i}\}_{\theta\in\{1..5000\}}$, 5000 “генов”:

$$\{p_{\theta i}\}_{\theta} \sim \begin{cases} \text{Beta}(\alpha_{1i}, \alpha_{1i}) \text{ , с вероятностью } \omega_{1i} \\ \\ \text{Beta}(\alpha_{2i}, \alpha_{2i}) \text{ , с вероятностью } (1-\omega_{1i}) \end{cases} \tag{2.8}$$

Далее, для каждого “гена” мы создали пару AI, распределённых бета-биномиально. Это даёт нам две реплики.

$$\{\xi_{1\theta i}, \xi_{2\theta i}\} \sim \frac{\text{Bin}(\widehat{C}_i, p_{\theta i}, p_{\theta i})}{\widehat{C}_i} \tag{2.9}$$

Разность $\xi_{1\theta i} - \xi_{2\theta i}$ задаёт ожидаемое распределение ΔAI.

Анализ квантилей и значений QCC Чтобы измерить избыточную дисперсию, мы выполнили квантильный анализ между наблюдаемым распределением ΔAI (Рис.2.3c) и ожидаемым распределением ΔAI (Рис.2.3e), внутри корзин по покрытию. Для каждой корзины i и набора квантилей $q \in \{0.2, 0.35, 0.5, 0.65, 0.8, 0.9, 0.95\}$, были посчитаны частные квантилей наблюдаемой ΔAI и квантилей ожидаемой ΔAI: $Q^{\text{obs.}}_{q,i}/Q^{\text{exp.}}_{q,i}$.

Далее к полученным точкам был подобран линейный тренд без наклона. Полученную константу мы называем коэффициентом коррекции качества (QCC), так как он отражает разницу между наблюдением и предположением модели о биномиальной выборке (см Рис.2.3f).

2.4.5 Анализ более, чем двух реплик

Когда для анализа имеется больше двух реплик, количество аллельных прочтений генов и ASE оцениваются из всех $M \geq 3$ реплик в образце вместе, и для коррекции интервалов доверия (CI) используется среднее всех попарных

коэффициентов QCC:

$$\mathrm{QCC} = \frac{\sum_{r_i, r_j \in \{1..M\},\ r_i \neq r_j} \mathrm{QCC}_{r_i r_j}}{\binom{M}{2}}. \tag{2.10}$$

Заметим, что перед выполнением этого шага полезно проверить, есть ли среди реплик выбросы, и при наличии, исключить их из дальнейшего анализа.

2.4.6 Поправка интервалов доверия аллельного дисбаланса

Чтобы оценить доверительный интервал точечной оценки AI, мы использовали функцию пропорционального теста `prop.test` из стандартного `R`-пакета `stats`. На вход тесту мы подавали меньшие в QCC^2 раз покрытия.

Поясним причину выбора этого теста. Мы наблюдаем, что квантили δAI в QCC раз шире, чем те, которые получены из биномиальной модели. Чтобы смоделировать это свойство данных, мы относимся к наблюдениям AI как к пропорциям, которые пришли из биномиального распределения с покрытием, меньшим в QCC^2 раз:

$$\mathrm{AI}_g \sim \frac{\mathrm{QCC}^2 \cdot \mathrm{Bin}(\frac{1}{\mathrm{QCC}^2} \cdot C_g, a_i)}{C_g} = \frac{\mathrm{Bin}(\frac{1}{\mathrm{QCC}^2} \cdot C_g, a_i)}{\frac{1}{\mathrm{QCC}^2} \cdot C_g}. \tag{2.11}$$

В этом приближении, прочтения генов, делённые на QCC^2, в целом не обязаны быть целым. Это ограничивает применимость биномиального теста. Тем не менее, мы всё ещё можем применить пропорциональный тест, который основан на интервалах Уилсона.

Для методологической консистентности, биномиальный тест в наших процедурах реализован как применение пропорционального теста `prop.test` на не скорректированных величинах прочтений генов. Это даёт математически эквивалентный результат.

2.4.7 Дифференциальный количественный анализ аллельного дисбаланса

Более точные оценки доверительных интервалов позволяет проводить нам дифференциальный анализ ASE от конкретных значений AI или между образцами:

- Разница между оценкой AI и конкретным числом считается значимой, если соответствующий интервал доверия не покрывает это значение.
- Для определения дифференциальной ASE между двумя образцами мы используем ту же функцию `prop.test` на аллельных покрытиях, подправленных на QCC^2 по процедуре, описанной выше.

Глава 3. Метилирование ДНК является ключевым механизмом для поддержания моноаллельной экспрессии на аутосомах

Эпигенетические механизмы контролируют аллель-специфическую транскрипцию в тысячах генов млекопитающих. Гены, подверженные аутосомной моноаллельной экспрессии (MAE), демонстрируют митотически стабильный выбор аллелей, что приводит к устойчивым транскрипционным различиям между клональными клеточными линиями. При этом механизм митотического поддержания MAE не изучен. Используя новую стратегию скрининга с помощью секвенирования (Screen-seq), позволяющую оценивать несколько локусов одновременно, мы обнаружили, что наиболее выраженная реактивация локусов MAE происходила в присутствии ингибитора ДНК-метилтрансферазы, 5-аза-2'-дезоксицитидина (5-аза-dC).

Выявление ключевой роли метилирования ДНК в поддержании MAE позволило впервые оценить полногеномные последствия нарушения аллель-специфического регуляторного ландшафта в клетках млекопитающих. Сотни генов продемонстрировали изменения в ASE после обработки препаратом 5-аза-dC. При этом, подмножество локусов MAE оказалось нечувствительным к деметилированию ДНК, что указывает на механистическую неоднородность MAE. В то время как цис-регуляция определяет общее базовое состояние ASE для всех клеток, метилирование ДНК играет роль "аллельного реостата" и определяет множество стабильных регуляторных состояний, различающихся между клональными клеточными линиями.

Наши результаты свидетельствуют о том, что аллель-специфический анализ клональных клеточных популяций может выявить долгосрочные транскрипционные ответы на возмущения, вызванные лекарственными препаратами.

Эта глава основана на публикации:

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

и препринте (*bioRxiv*):
DNA methylation is a key mechanism for maintaining monoallelic expression on autosomes / Saumya Gupta, Denis L. Lafontaine, Sebastien Vigneau, Svetlana Vinogradova, **Asia Mendelevich**, Kyomi J. Igarashi, Andrew Bortvin, Clara F. Alves-Pereira, Kendell Clement, Luca Pinello, Andreas Gnirke, Henry Long, Alexander Gusev, Anwesha Nag, Alexander A. Gimelbrant // *bioRxiv preprint* — 2020 — DOI:10.1101/2020.02.20.954834

3.1 Результаты

3.1.1 Подход скрининга методом секвенирования для поиска изменений в аллель-специфической экспрессии

Для проверки реактивации выключенных аллелей мы изучали сдвиги в аллельном дисбалансе при воздействии лекарств. Чтобы повысить вероятность обнаружения изменения в дисбалансе среди генов с потенциально различной регуляцией, наш подход к скринингу в идеале должен был сочетать возможность анализа нескольких определённых генов, чувствительность к изменениям AI и высокую пропускную способность для обработки множества образцов после воздействия рядом возмущений. Для достижения этих целей мы разработали стратегию скрининга с помощью секвенирования – Screen-seq.

Схема эксперимента Screen-seq представлена на Рис.3.1B. Клетки выращивали и лизировали в 96-луночных планшетах; РНК выделяли с помощью магнитных частиц, а обратную транскрипцию полиА РНК проводили в пристствии случайных праймеров и олиго-dT праймеров с уникальными молекулярными идентификаторами (UMIs)[104; 105]. Это позволило таргетировать два типа полиморфизмов на следующем этапе, мультиплексной ПЦР: полиморфизмы, расположенные близко к 3’ концу, позволяют использовать олиго-dT-UMIs с последующим ген-специфическим праймером, в то время как другие полиморфизмы были таргетированы с помощью двух ген-специфических праймеров в кДНК со случайным праймером. Затем с помощью ПЦР были добавлены барко-

ды, кодирующие планшеты и индивидуальные лунки. Далее материал из всех лунок был объединён, добавлены адаптеры Illumina, и объединенная библиотека была просеквенирована. Наконец, покрытия SNP были присвоены конкретным генам, а баркоды – конкретным лункам с известным возмущением.

Для эксперимента были выбраны 27 однонуклеотидных полиморфизмов и 23 генов, включая 15 клон-специфичных MAE генов, 3 биаллельных, один импринтированный и 4 X-инактивированных локуса (см. Рис. 3.1C). Выбранные MAE гены продемонстрировали абсолютный аллельный перекос (например, *Afap1*, AI = 1) или сильный, но не полный перекос (например, *Dlc1*, AI = 0.1) в Abl.1 клоне, но, при этом, имели противоположное аллельное смещение или были биаллельными в другом клоне, Abl.2 [24; 26].

3.1.2 Выявление возмущений, влияющих на аллель-специфическую экспрессию генов

Известно, что MAE ассоциирована со специфической конфигурацией хроматина, таких как комбинации модификаций гистонов в клетках человека и мыши [26; 28], что позволяет предположить, что механизмы модификации хроматина могут быть вовлечены в поддержание MAE. Поэтому, для оценки влияния на аллельный дисбаланс в выбранных локусах, мы применили 43 препарата, известных своим влиянием на активность ферментов, участвующих в добавлении и удалении меток метилирования и ацетилирования на гистонах и ДНК. Клетки Abl.1 в 96-луночных планшетах подвергались воздействию отдельных препаратов в течение 21 дня в остальном обычных условиях. Каждый препарат применялся в трех концентрациях (1 μM, 10 μM и 20 μM в 1% ДМСО). Контролем служили клетки с добавлением только растворителя (1% ДМСО). Свежая среда (с препаратами или без них) заменялась каждые два дня. На 7, 14 и 21 день аликвоты клеток отбирались для анализа.

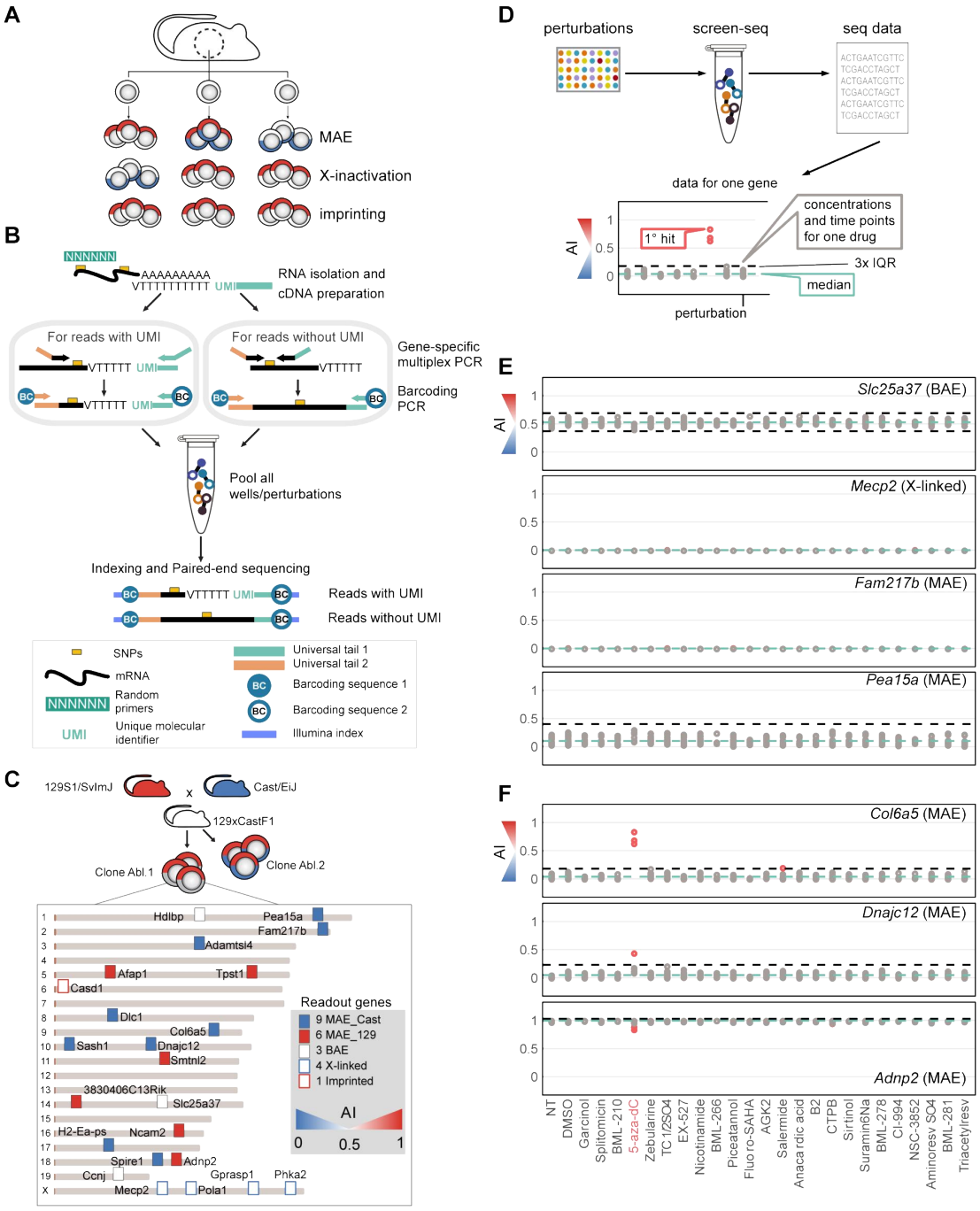

Рисунок 3.1 — Метод скрининга секвенированием позволяет находить возмущения, которые реактивируют выключенные аллели MAE генов.

(A) Разные эпигенетические типы MAE: униформность или клональный мозаицизм. **(B)** Схема методологии Screen-seq (для подробностей, см. Методы). **(C)** 23 гена, проанализированные с помощью Screen-seq, и их расположение на референсном геноме. AI данных генов в клональной популяции Abl.1 отображён цветом. Центромера обозначена коричневой меткой слева. **(D-F)** Результаты эксперимента для репрезентативной подвыборки генов. Каждый из 48 препаратов был использован в трёх концентрациях: 1 μM, 10 μM, и 20 μM в 1% ДМСО. Среда заменялась раз в 2 дня, клетки были собраны на 7, 14 и 21 дни. **(D)** Пояснительная схема отображения экспериментальных результатов, для (E-F). **(E)** Гены, не изменившие AI. **(F)** Гены с существенными изменениями AI под какими-то из воздействий.

В случае 19 из 43 препаратов, даже при наименьшей концентрации, живых клеток не осталось. Таким образом, имеющиеся результаты включают в себя 72 лунки с обработанными клетками (24 оставшихся препарата в трех концентрациях) и 24 лунки с контролем (12 необработанных и 12 обработанных ДМСО). Итого, в этом эксперименте Screen-seq мы оценили 7 776 экспериментальных точек (аллель-специфические измерения 27 SNPs $\times$ 96 лунок $\times$ 3 временные точки). Но при 1000 прочтений на экспериментальную точку, для всего скрининга потребовалось менее 10 млн секвенированных фрагментов.

В качестве кандидатов мы определили условия, существенно меняющие AI (См. Рис.3.1D-F). Некоторые гены (например, *Fam217b* или *Mecp2*) демонстрировали очень консистентный алелльный дисбаланс, вне зависимости от концентрации препарата и временных точек, в то время как для других генов (например, *Pea15a* или *Col6a5*) наблюдалась большая вариабельность.

Как и ожидалось в случае стабильно поддерживаемой алелль-специфической экспрессии, в контрольных (не подвергавшиеся действию препарата) клетках не наблюдались отклонения AI от исходного состояния. Наиболее выраженные отклонения AI (красный на Рис.3.1F) были обнаружены для 3 MAE генов при добавлении ингибитора ДНК-метилтрансферазы, 5-аза-2'-дезоксицитидина (5-аза-dC). Также наблюдались значительные сдвиги AI в отдельных генах после воздействия модуляторов деацетилазы гистонов Salermide и BML-278. Величина наблюдаемых сдвигов варьировала в зависимости от генов и условий (концентрация препарата и время воздействия). Среди тестовых локусов наиболее ярким примером оказался сдвиг в гене *Col6a5* от AI ~ 0.1 в контроле к AI ~ 0.8 за 7 дней в присутствии 1 μM 5-аза-dC (см. Рис.3.1F). Особо примечательно, что после обработки игибитором метилтрансферазы AI не стал ближе к 50:50, как ожидалось, а поменялся из одной крайности в другую. Значительные, но меньшие сдвиги наблюдались после воздействия того же препарата в генах MAE *Adnp2* (от AI $= 1$ до AI $= 0.8$) и *Dnajc12* (AI ~ 0.1 до AI ~ 0.2). В других протестированных генах сдвига AI при обработке 5-аза-dC обнаружено не было.

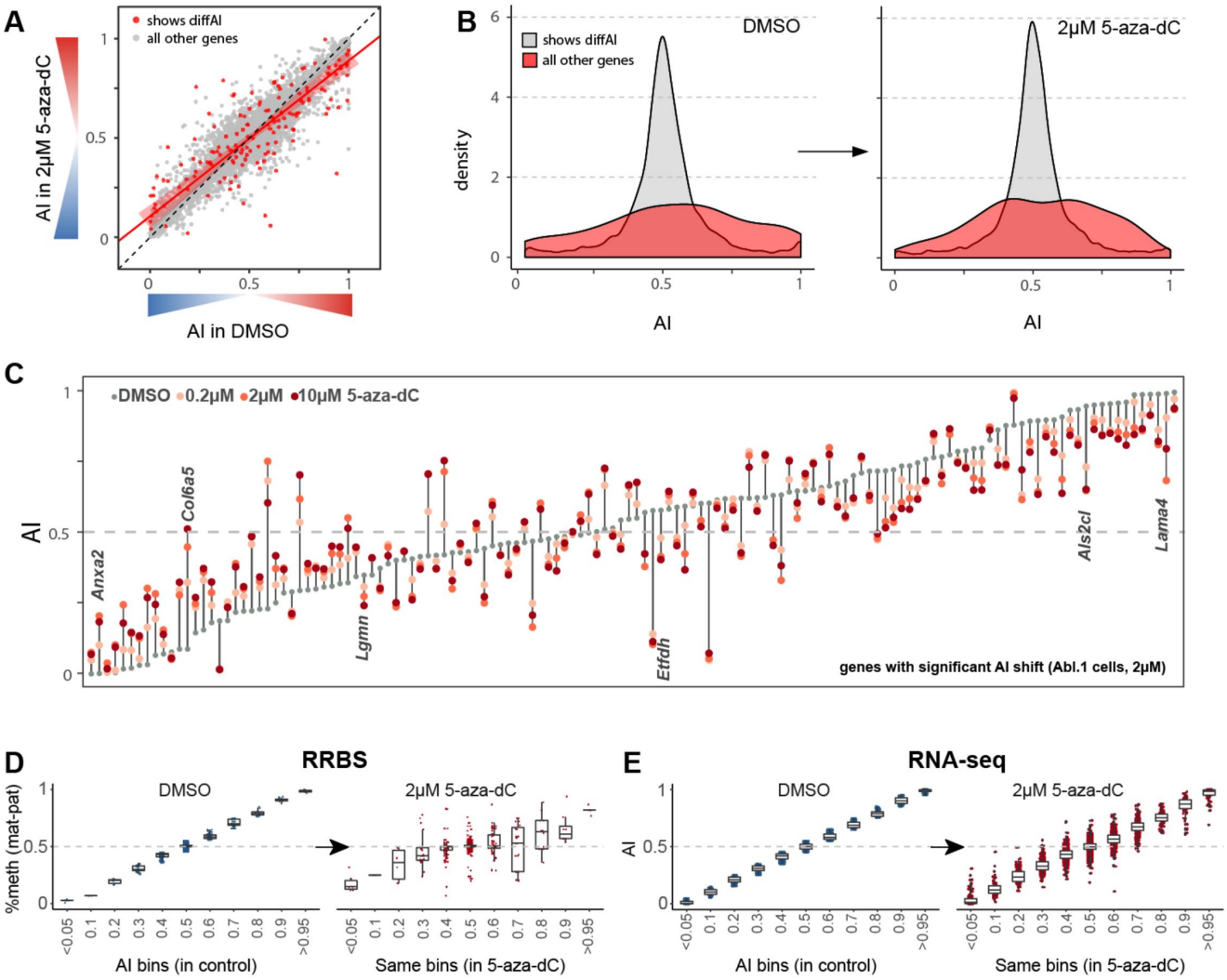

Рисунок 3.2 — Влияние препарата 5-aza-dC на аллель-специфическую экспрессию в масштабе всего генома.

(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 является результатом многомерной линейной регрессии (подробности см. в Сопроводительной Заметке Б.1.1). **(B)** Распределения значений AI для генов без существенных изменений (grey) и имеющих дифференциальный AI между контролем и обработкой 2 μM 5-aza-dC (red), данные те же, что и в (a). **(C)** Сдвиги в экспрессионном AI для генов с диффернциальным AI, что и в панелях (a,b). Серой точкой изображено значение в контрольном образце, цветными – при различных концетрациях 5-aza-dC. Аналогичные графики для линий Abl.2-4 отображены на Рис.Б.3. **(D-E)** Аллель-специфические изменения в ДНК метиломе и транскриптоме в клетках Abl.1. **(D)** Аллель-специфическое ДНК метилирование, RRBS (бисульфитное секвенирование с пониженной репрезентативностью). Слева: гены были разбиты по группам $X \pm 0.05$ согласно разнице в пропорциях метилированных CpGs на материнской и отеческой аллелях в образцах с 1% ДМСО; справа: значения на тех же генах после применения 2 μM 5-aza-dC. **(E)** Алелль-специфическая экспрессия. Слева: гены были разбиты по группам $X \pm 0.05$ согласно их значениям AI, справа: значения на тех же генах после применения 2 μM 5-aza-dC.

3.1.3 Полногеномное влияние деметилирования ДНК на аллель-специфическую экспресиию

Мы оценили глобальное влияние 5-aza-dC на аллель-специфическое метилирование ДНК и транскрипцию, подвергая клетки клона Abl.1 в течение 2 дней воздействию низких (0.2 μM), средних (2 μM) и высоких (10 μM) концентраций 5-aza-dC, с обработкой 1% ДМСО в качестве контроля. Для проведения аллель-специфического анализа в транскрипции, мы применили подход `Qllelic` (см. главу 2 и [106]). Дифференциальный анализ выявил 53, 137 и 135 генов с существенным сдвигом в AI после воздействия низкой, средней и высокой концентраций 5-aza-dC соответственно (Рис.3.2A). Направление сдвигов при различных концентрациях 5-aza-dC совпадало, и большая концентрация обыкновенно соответствовала большему сдвигу (Рис.3.2C). Ни один из известных импринтированных генов не показал значительных изменений в AI в этих условиях, что говорит о более надежном митотическом поддержании импринтинга и X-инактивации, чем MAE.

3.1.4 5aza-dC уменьшает различия между клональными популяциями.

Определяющей особенностью MAE является то, что аллельная предрасположенность генов MAE отличает клональные клеточные линии друг от друга [23; 107]. Поэтому мы задались вопросом, как деметилирование ДНК влияет на гетерогенность аллель-специфической экспрессии между клонами. Мы измерили аллель-специфическую экспрессию в четырех клональных линиях лимфоидных клеток до и после воздействия 5-aza-dC. Помимо клона Abl.1, мы проанализировали клоны Abl.2, Abl.3 и Abl.4 (все эти клетки получены из скрещивания мышей из одних линий, таким образом, генетически они практически идентичны) [24].

Для сравнения последствий деметилирования между клонами мы использовали “эквитоксичные” концентрации 5-aza-dC. Клоны Abl.2, Abl.3 и Abl.4 име-

ли аналогичную выживаемость при 0.2 μM с выживаемостью Abl.1 при 2 μM. После двухдневного воздействия 5-aza-dC в этих концентрациях мертвые клетки удаляли, а на оставшихся живых клетках проводили РНК-секвенирование. В результате анализа 676 генов показали значительные сдвиги AI в одном или нескольких из четырех исследованных клонов: 243 гена в клетках Abl.2, 265 в Abl.3 и 176 в Abl.4.

Только некоторые из генов со значительными различиями в AI между клонами показали значительные изменения AI после деметилирования ДНК (367 из 1767 таких генов). В остальных локусах дисбаланс не изменился после воздействия 5-аза-dC, что позволяет предположить, что в этих локусах в митотическом поддержании различий в AI между клонами участвует дополнительный механизм, отличный от метилирования ДНК. Чтобы исключить возможные случаи потери гетерозиготности (которые проявились бы как специфические для клонов различия, не чувствительные к состоянию метилирования ДНК), мы провели секвенирование экзомов этих клеток, и исключили исключили из сравнения все локусы с выраженным аллельным смещением в геномной ДНК (AI$<$ 0.3 или AI$>$ 0.7).

Примечательно, что после деметилирования общая вариация значений AI по клонам уменьшилась (Рис.3.3A,B): как для 676 генов со значительными изменениями в AI ($p < 10^{-16}$, односторонний парный тест Уилкоксона для sd(ДМСО)$<$sd(5-аза)), так и для всего транскриптома ($p < 10^{-80}$). Часть генов имели тенденцию к сближению в четырех клонах (Рис.3.3C-E) после обработки 5-aza-dC, у некоторых генов вариация между AI после обработки увеличилась, однако был заметен ярко выраженный тренд (рис.3.3F). Наблюдаемая тенденция сближения согласуется с идеей о том, что функциональная роль MAE заключается в поддержании клональной гетерогенности.

Интересно, что в случае сходимости значений AI в разных клонах, предельной точкой не обязательно оказывался баланс 50:50 (например, рис.3.3C), напротив, во многих случаях это сопровождалось смещением от 50:50 к экстремальным значениям AI (Рис.3.3D-F, Рис.3.2C). В соответствии с понятием “схождения к” некоторому устойчивому значению, в клонах, где ген MAE уже находился в устойчивом состоянии AI в контрольных условиях (например, *Casp6* в клонах Abl.1 и Abl.4; Рис.3.3C), не было никаких дальнейших сдвигов AI после деметилирования ДНК.

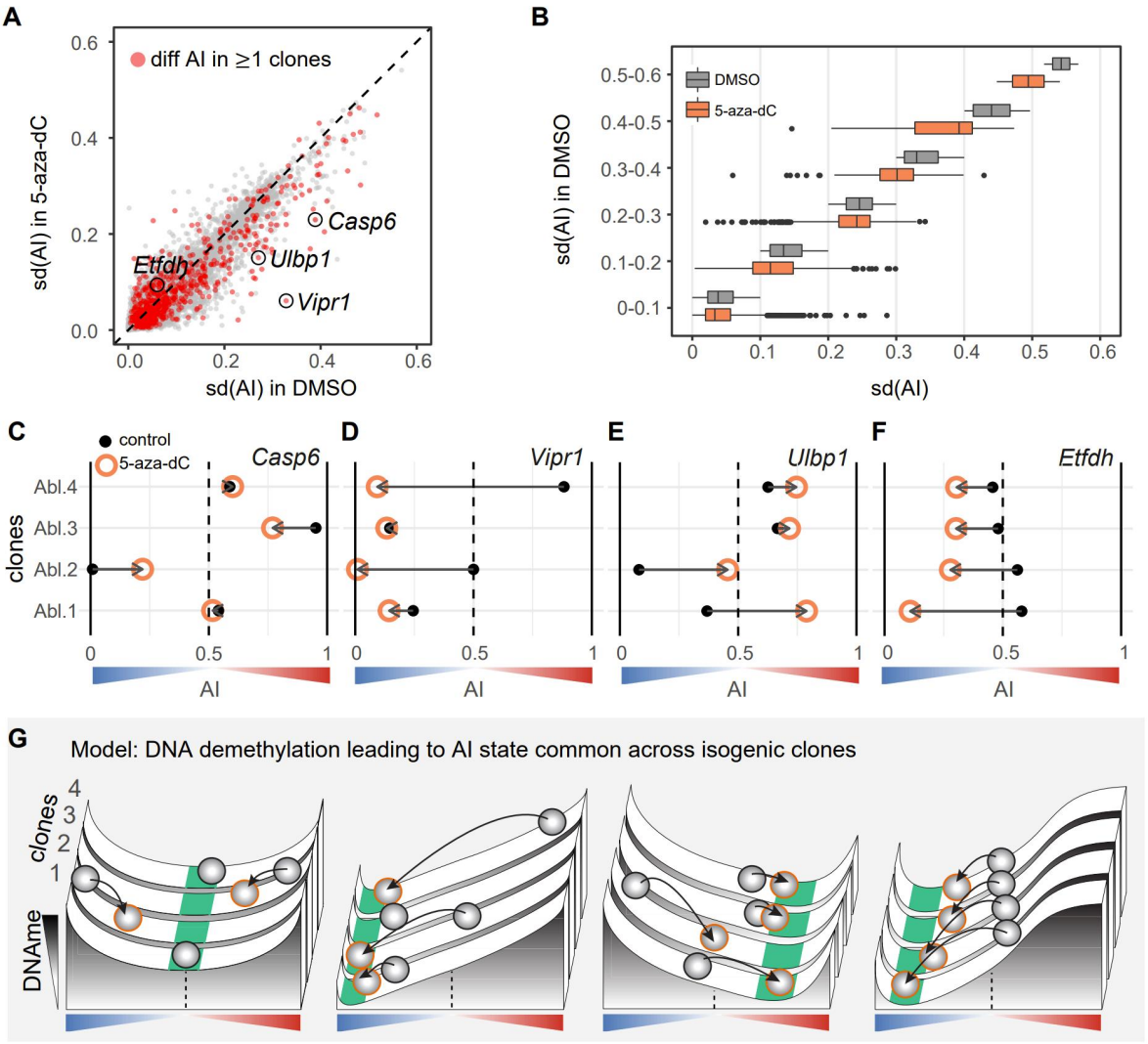

Рисунок 3.3 — Деметилирование ДНК приводит к большему сходству между клонами в аллель-специфической транскрипции.

(A) Сопоставление стандартных отклонений значений AI междё 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 между контрольными образцами и обработанными 5-aza-dC (хотя бы в одном из клонов), серым – остальные. Гены под диагональю $x = y$ демонстрируют меньшую вариабельность после деметилирования. **(B)** Разница в sd для генов, разделённых на имеющие и не имеющие дифференциальную ASE, как в (a). Гены сгруппированны по значениям sd в контрольных образцах, интервалами ширины 0.1. **(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, что и в (C-F), зелёная зона – точка сходимости, “общий” аллельный дисбаланс в экспрессии.

Чтобы объединить наши наблюдения, мы предлагаем спекулятивную модель (Рис.3.3G), в которой AI для многих генов MAE имеют значения в разных клонах и поддерживаются посредством метилирования ДНК цис-регуляторных последовательностей у затронутых генов. В этой модели деметилирование приводит к тому, что эти значения AI сходятся к генетически определенному состоянию, а частичное деметилирование приводит к промежуточному значению.

3.2 Обсуждение результатов

Используя метод скрининга секвенированием, мы установили ключевую роль метилирования ДНК в митотическом поддержании моноаллельной экспрессии в клональных линиях лимфоидных клеток млекопитающих (Примечание: сходные результаты были позже представлены в работе [108].) *Dnmt1*-зависимое поддержание ДНК метилирования предлагает простое объяснение стабильности MAE, поскольку это очень стабильная форма молекулярной памяти: в качестве экстремального примера, метилирование цитозина в геноме *Cryptococcus*, как полагают, поддерживалось в течение миллионов лет в отсутствие метилирования *de novo* [109]. Мы предлагаем простую модель (Рис.3.3G), в которой аллель-специфический регуляторный ландшафт определяется генетическими вариациями (возможно, во взаимодействии с эпигенетическими механизмами), а специфическое состояние клональной популяции клеток зависит от метилирования ДНК. Эта модель предусматривает наличие специфических регуляторных элементов, расположенных близко к затронутым генам. Такие геномные элементы могли бы дать простое объяснение эволюционной сохранности статуса MAE генов в человеческих популяциях [32] и между человеком и мышью [24; 110; 111]. Когда и как происходит метилирование ДНК в этих регуляторных регионах, еще предстоит выяснить.

Деметилирование ДНК повлияло не на все гены MAE, что позволяет предположить, что поддержание MAE для некоторых локусов включает другие механизмы в дополнение (или вместо) метилирования ДНК. Это дает одно из вероятных объяснений предыдущим наблюдениям, что препараты, деметилирующие ДНК, не влияли на аллельный дисбаланс ни в одном из нескольких

рассмотренных генов MAE [28; 29]. В соответствии с идеей о дополнительных механизмах поддержания MAE, активатор SIRT1 BML-278 и ингибитор сиртуина, салермид, оказались другими кандидатами в нашем скрининге, что позволяет дополнительно предположить, что расширенное применение стратегии Screen-seq может выявить больше таких механизмов. Кроме того, недавно было высказано предположение о роли CTCF-опосредованной динамики хроматина в регуляции аллельной транскрипции [112].

Мы оценили полногеномное влияние деметилирования ДНК на аллельный дисбаланс: в четырех проанализированных клонах мы обнаружили значительные сдвиги аллельного дисбаланса в более чем 600 аутосомных генах. Точный количественный анализ данных РНК-секвенирования выявляет более сложный ландшафт митотически стабильного клонального разнообразия в аллель-специфической регуляции генов, чем подразумевается моноаллельной/биаллельной дихотомией. Дополнительно интересно, что многие из сдвигов AI при деметилировании были увеличением аллельного дисбаланса. Также в разных клонах AI гена может принимать значения от 0 до 1, а эпигенетическая регуляция действует как значительно более тонкий механизм управления, чем переключатель “вкл/выкл”. Это расширяет идею о реостатической роли метилирования ДНК, предложенную в качестве регуляции покрытия транскриптов [113].

Значительное влияние препаратов для деметилирования ДНК на аллель-специфическую экспрессию в лимфоцитах имеет особое значение, поскольку и 5-aza-2’-dC, и 5-azaC используются для лечения острого лейкоза и других злокачественных опухолей [114]. Более того, концентрации этих препаратов в наших экспериментах (0.2 - 1.0 μM) сходны с концентрациями, измеренными в плазме пациентов (5-aza-2’-dC при $\sim$60 нг/мл, около 0.25 μM [115]). Таким образом, наши результаты свидетельствуют о том, что ингибиторы ДНМТ, вероятно, влияют на регуляцию генов у пациентов таким образом, который трудно обнаружить без аллель-специфического анализа. Это говорит о том, что количественный анализ аллель-специфической регуляции генов в популяциях поликлональных и моноклональных клеток должен привести к новым клинически значимым открытиям.

3.3 Материалы и методы

3.3.1 Клеточная культура

Клональные линии про-Б клеток мыши Abl.1, Abl.2, Abl.3 и Abl.4 были получены от самок мышей 129S1/SvImJ × Cast/EiJ F1 (иммортализованы с помощью вируса мышиной лейкемии Абельсона и клонированы с помощью сортировки одиночных клеток) и ранее описаны в [24], выращивались в среде RPMI (Gibco), содержащей 15% FBS (Sigma), 1X L-Глутамин (Gibco), 1X Пенициллин/Стрептомицин (Gibco) и 0.1% β-меркаптоэтанол (Sigma).

Среднее расстояние между SNP в геноме этих мышей, за исключением повторов, составляет ∼80 п.н., и почти 80% генов содержит по крайней мере один информативный SNP.

3.3.2 Обработка препаратами

Для первоначального скрининга веществ использовалась библиотека SCREEN-WELL Epigenetics arrayed drug library, приобретенная у Enzo Life sciences (BML-2836). Клон Abl.1 обрабатывали всеми веществами из библиотеки в 96-луночных планшетах при концентрациях 1 μM, 10 μM и 20 μM, чтобы охватить достаточно широкий диапазон концентраций, при которых лекарства потенциально фармакологически активны. Культуры обрабатывали в течение 21 дня, при этом среду меняли каждый второй день. Токсичность клеток проверяли визуально, и 19 из 43 препаратов, в средах с которыми через шесть дней не было обнаружено живых клеток, не рассматривали.

Для экспериментов по полногеномному секвенированию 5×10^6 клеток высевали в среду с концентрацией 5-aza-dC 0.2 μM (низкая), 2 μM (средняя) и 10 μM (высокая). Клетки собирали на 2-й день. На этом этапе большое внимание уделялось жизнеспособности клеток: (a) клетки поддерживались при плотности $1 - 2 \times 10^6$ клеток/мл, при которой лимфоидные клеточные линии

Абельсона показывают идеальный рост (б) секвенировались только живые клетки. Живые клетки отделяли путем градиентного центрифугирования сахарозы (Histopaque-1077, Sigma). РНК выделяли из 2×10^5 живых клеток, а оставшиеся живые клетки промывали 1X PBS и замораживали на сухом льду для последующего выделения геномной ДНК. AI для *Col6a5* был измерен в живых клетках с помощью метода ддПЦР (капельная цифровая ПЦР), и значения AI хорошо согласуются с предыдущими наблюдениями.

3.3.3 Приготовление ДНК и РНК

Для всех клональных линий лимфоидных клеток РНК выделяли по протоколу на основе магнитных частиц с использованием Sera-Mag SpeedBeads (GE Healthcare). Изолированную РНК обрабатывали ДНКазой RQ1 (Promega). Синтез кДНК первой цепи проводили с использованием обратной транскриптазы Episcript RNase H (Epicentre), при этом образцы РНК праймировали случайными гексамерами (NEB) или олиго-дТ праймерами с UMI, как описано ниже. . Обработка ДНК и синтез кДНК проводились в соответствии со спецификациями производителя с минимальными изменениями. Для получения РНК из селезенки мыши, клетки выделяли путем раздавливания всей селезенки задней частью шприца объемом 1 мл в нейлоновом фильтре 40 μM и промывания фильтра 1X PBS (Phosphate-buffered saline, Sigma) для сбора клеток. Клетки из селезенки отжимали и выделяли РНК с помощью реактива Trizol (Invitrogen). Выделение геномной ДНК для проверки чувствительности Screen-seq проводили методом высаливания, а для бисульфитного секвенирования с пониженной репрезентативностью (RRBS) - с помощью набора Sigma GenElute kit (G1N10-1KT). RT-qPCR проводили с использованием iTaq Universal SYBR Green Supermix (BioRad) по протоколу производителя на системе 7900HT Fast Real-Time PCR (Applied Biosystems Inc.).

3.3.4 Скрининг секвенированием

После подготовки библиотек для скрининга секвенированием, как описано выше, их секвенировали в UMass Boston и Center for Cancer Systems Biology (CCSB) на секвенаторах Illumina HiSeq 2500 и MiSeq, соответственно, с использованием четырехцветных наборов реагентов. С конца адаптера P7 было секвенировано 65 п.н. (прочтение 1), включая один из двух баркодов для кодирования лунок планшета (и UMI, где необходимо). Со стороны адаптера P5 секвенировали оставшиеся 135 п.н. (прочтение 2), включая второй баркод для кодирования лунок и ампликон кДНК, содержащий полиморфизм. Кроме того, стандартные баркоды Illumina использовались для различения плашек в объединенной библиотеке, с демультиплексированием перед дальнейшей обработкой. Прочтения выравнивались с помощью программы `bowtie2` [116] на геномную сборку мыши mm10. Полученные BAM файлы обрабатывались с помощью пользовательских скриптов на языке `Perl`, для извлечения аллель-специфических, скорректированных с помощью UMI подсчетов для каждого гена и каждой лунки.

Для выявления препаратов-кандидатов оценки AI анализировались с помощью пользовательских скриптов `R`. Вкратце, для каждого гена бралась медианная оценка AI для всех лекарственных условий и интерквартильный размах (IQR = Q3 - Q1, где Q1 и Q3 - 25-й и 75-й процентили). Наблюдения размером менее 30 были отфильтрованы (наблюдение состоит из количества аллелей для одного гена в одной лунке). Обычно для определения выбросов используются значения $\leq$Q1$-1.5\times$IQR или $\geq$Q3$+1.5\times$IQR, мы использовали более строгий порог $3\times$IQR, чтобы снизить вероятность ложноположительных результатов.

3.3.5 Обработка данных РНК-секвенирования

Библиотеки для РНК-секвенирования готовили из клеток, собранных на 2-й день обработки 5-aza-dC, используя как минимум два технических повтора для одной и той же РНК (5 повторов для клеток Abl.1, обработанных 2 μM

5-aza-dC), с помощью набора SMARTseqv4 (Clonetech), начиная с 10 нг РНК для каждой библиотеки, в соответствии с инструкциями производителя. Подготовка библиотек, контроль качества и секвенирование проводились в центре молекулярной биологии Института рака Дана-Фарбер. Одноконцевые 75 п.н. чтения генерировались с помощью Nextseq 500 (Illumina).

Анализ аллель-специфической экспрессии генов проводили с использованием `ASEReadCounter`* и `Qllelic` версии v0.3.1 (см. главу 2 и [106]). Различия в AI принимались как значимые после учета специфической для эксперимента избыточной дисперсии, оцененной с помощью анализа технических реплик.

Для многомерной линейной регрессии без предикторов (Рис.3.2A) был использован метод описанный в Заметке Б.1.1, с предположением похожей избыточной дисперсии и покрытий в сопоставляемых наборах данных ($\sigma 1 = \sigma 2$).

Экзомное секвенирование было сделано на геномной ДНК всех клонов для нахождения артефактов копийности и потери гетерозиготности. Подготовка библиотеки, контроль качества и секвенирование (50x) были выполнены в LC Sciences (TX, США). Выделение экзома проводилось с помощью SureSelect (Agilent Technologies) в соответствии с протоколом, рекомендованным производителем. Парные чтения длиной 150 п.н. генерировались с помощью секвенатора Hiseq X Ten (Illumina). Гены с общим количеством прочтений < 10 и гены с геномным AI > 0.7 или < 0.3 были исключены из рассмотрения в соответствующих клонов перед сравнением данных РНК-секвенирования между клонами.

Для бисульфитного секвенирования с пониженной репрезентативностью (RRBS) библиотеки генерировали из 50 нг исходной геномной ДНК с использованием системы NuGEN Ovation RRBS Methyl-Seq System (Tecan) в соответствии с рекомендациями производителя. Библиотеки амплифицировали методом ПЦР, 11 циклов. Парные чтения по 100 п.н. генерировались с помощью прибора HiSeq 2500 (Illumina). Чтения выравнивали на геном мыши mm10 с помощью `BSmap3` с флагами `-v 0.05 -s 16 -w 100 -S 1 -p 8 -u`. Пользовательские скрипты, написанные на языка `Perl`, использовались для расчета процента метилирования для CpGs, покрытых 4 или более считываниями в местах расположения известных SNP. Вкратце, файлы VCF, содержащие SNPs между линиями 129S1 и CAST, были отфильтрованы на предмет C $\rightarrow$ T или G $\rightarrow$ A. Для каждого SNP были извлечены RRBS чтения, которые перекрывались с этим SNP, и статус метилирования геномных цитозинов был рассчитан путем

деления числа неконвертированных (метилированных) цитозинов (C) на общее число неконвертированных (C) или конвертированных (T) цитозинов. Статус метилирования всех цитозинов в чтениях, перекрывающих SNP, объединяли по статусу SNP для создания среднего значения метилирования для референсного и альтернативного генотипа.

Глава 4. Внешние РНК-контроли позволяют проводить точный аллель-специфический анализ экспрессии на большом количестве образцов

На результаты анализа аллель-специфической экспрессии сильно влияет технический шум, присутствующий в экспериментах РНК-секвенирования (Рис. 4.1). Ранее мы показали, что технические реплики могут быть использованы для измерения избыточной дисперсии, и предоставили программный метод, R-пакет `Qllelic`, для аккуратной обработки аллель-специфической экспрессии. Точность этого метода высокая, однако он является дорогостоящим из-за необходимости производить две или более реплик каждой библиотеки, что ограничивает его применение в крупномасштабных экспериментах.

В этой главе описывается альтернативный подход для аллель-специфического анализа, требующий небольшого увеличения затрат при большом количестве образцов. Основная идея нового протокола заключается в добавлении аликвоты чужеродной РНК к каждому образцу (например, добавление РНК гетерозиготной мыши к РНК человека), в качестве РНК-контроля, перед началом приготовления библиотеки.

Аналитическая часть нашего подхода опирается на единообразие аликвот РНК во всех библиотеках в наборе данных. Мы экстраполируем нашу ранее описанную идею обработки данных из технических реплик на порции контрольной РНК. Контрольная РНК проходит через все стадии эксперимента вместе с РНК основного образца, накапливая тот же уровень технического шума: мы экспериментально показали, что избыточная дисперсия в аллель-специфических данных РНК-контроля отражает избыточную дисперсию в основном образце. Таким образом, РНК-контроли выступают в качестве стандарта для оценки технического шума в каждой библиотеке. Добавление контрольной РНК к основному образцу в соотношении 1:10, устраняет необходимость в дополнительных библиотеках за счет небольшого увеличения общей глубины секвенирования.

Мы экспериментально продемонстрировали эффективность этого подхода, используя комбинации РНК из видов, легко отличаемых при выравнивании, а именно: мыши, человека и *C.elegans*. Наш новый подход, `controlFreq`, позволяет проводить высокоточный и вычислительно эффективный анализ ал-

лель-специфической экспрессии внутри и между исследованиями произвольного масштаба, при общем увеличении расходов на эксперимент на $\sim 5\%$.

Отметим, что не разрешённые по аллелям РНК-контроли ранее использовались для оценки технического шума при измерении уровня транскрипции в экспериментах многоклеточного РНК-секвенирования [67; 117], а также для оценки фонового шума, групповых эффектов и дуплетов в экспериментах одноклеточного РНК-секвенирования [3; 4; 118]. Однако оценки неаллельной избыточной дисперсии имеют ограниченное применение для аллель-специфического анализа (см главу 2 и [106]).

Реализация метода доступна на GitHub в качестве `R`-пакета `controlFreq` (`github.com/gimelbrantlab/controlFreq`).

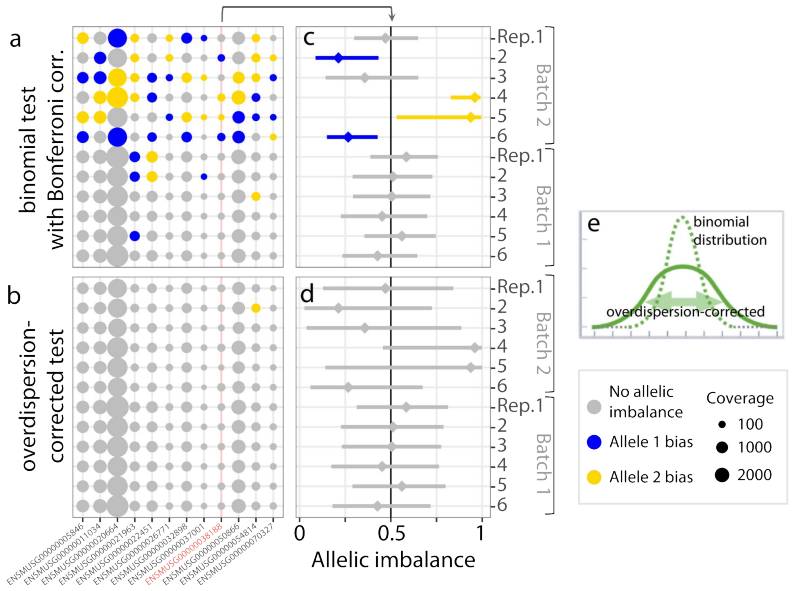

Рисунок 4.1 — Аллель-специфический сигнал в данных РНК-секвенирования может существенно изменяться под влиянием технического шума.

Данные: см. главу 2 и [106], реплики из экспериментов 2 и 3 обозначены здесь как “Batch 1” и “Batch 2”; один и тот же метод подготовки библиотек (SMART-Seq), количество исходной тотальной РНК: 10 нг и 100 нг (в пределах рекомендуемого диапазона). **(a)** Показаны все гены с аллельным покрытием >10 и демонстрирующие противоречивый аллельный дисбаланс по результатам биномиального теста [89]; 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*, выделен красным на панелях (a-b), аллельное покрытие 110 ± 45). Ромб обозначает точечную оценку AI, отрезок — доверительный интервал, определённый с помощью соответствующей модели. **(e)** Схематичное изображение супер-биномиальной дисперсии (непрерывная линия) в сравнении с биномиальной (пунктирная).

Эта глава основана на публикации:

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

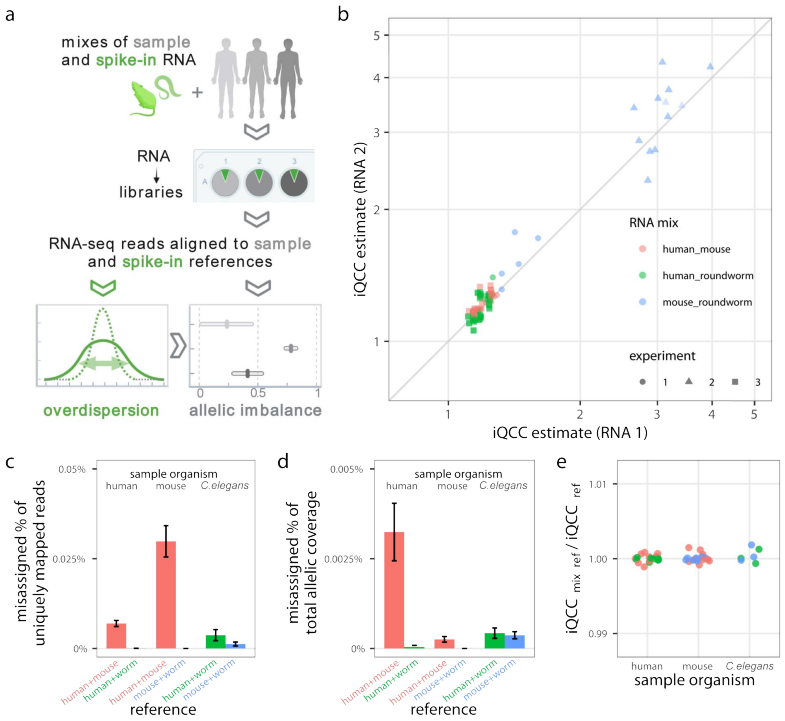

Рисунок 4.2 — В библиотеке, состоящей из РНК двух различных организмов, избыточные аллельные дисперсии для обоих организмов близки.

(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). Для описания данных, см. Таблицу 1. **(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) из эксперимента 1, выравненные на индивидуальные или химерные референсы. Цвет представляет смесь референсов использованную для выравнивания, кодирование пар совпадает с кодированием смесей в (b). **(c)** Процент неверно выравненных прочтений среди всех уникально выравненных прочтений. **(d)** Процент аллель-разрешённых прочтений, выравненных на неверный организм. **(e)** Сравнение значений iQCC вычисленных на выравнивании на одиночный или смешанный (химерный) референс. (Всего было найдено 0 генов с дифференциальным аллельным дисбалансом для каждой возможной пары).

4.1 Материалы и методы

4.1.1 Измерение избыточной дисперсии при помощи расширенной бета-биномиальной модели

В данной секции мы обновляем и дополняем процедуру, описанную в главе 2 и [106]. Напомним, что технические реплики определены как набор отдельных библиотек, приготовленных из одной РНК, поэтому различия между ними отражают технический шум, накапливаемый с этапа приготовления библиотеки. Имея несколько технических реплик, мы можем подсчитать попарные коэффициенты коррекции качества (QCC). Эти коэффициенты можно рассматривать как коэффициенты, расширяющие ожидаемые биномиальные квантили для реплик с похожими размерами библиотек. В главе 2 мы делали попарные сравнения и присваивали среднее значение избыточной дисперсии каждой реплике, исходя из значений QCC для соответствующих пар.

Далее мы изложим новую процедуру, способную проводить аллель-специфический анализ на большом количестве образцов (Рисунок **4.2a**). Она присваивает индивидуальные коэффициенты коррекции качества (iQCC) каждому образцу из набора (Рисунок **4.3a**), и применима как в случае технических реплик, так и в случае внешних РНК-контролей.

Расширенное бета-биномиальное распределение

Бета-биномиальное семейство распределений часто используется для моделирования распределения с дисперсией, избыточной относительно соответствующего биномиального распределения. Один из способов его параметризовать – при помощи вещественных параметров $\alpha \geq 0$ и $\beta \geq 0$. Эти параметры можно интерпретировать как количество шаров различного типа в модели урны Полья. В данной параметризации функция вероятности может быть записана

следующим образом:

$$BB(m \mid n = m + p, \alpha, \beta) = \binom{n}{m} \frac{\prod_{i=0}^{m-1}(\alpha + i) \cdot \prod_{j=0}^{p-1}(\beta + j)}{\prod_{k=0}^{n-1}(\alpha + \beta + k)}.$$

Вариация такого распределения равна

$$\frac{\alpha + \beta + n}{\alpha + \beta + 1} \cdot \frac{n\alpha\beta}{(\alpha + \beta)^2}.$$

Второй сомножитель вариации может быть отождествлён с вариацией биномиального распределения с вероятностью $\mathrm{AI} = \frac{\alpha}{\alpha+\beta}$, поэтому избыточная дисперсия равна $\frac{\alpha+\beta+n}{\alpha+\beta+1}$. Обозначим её параметром Q. Из параметров $0 \leq \mathrm{AI} \leq 1$ и $1 \leq Q \leq n$, мы можем восстановить α и β как $\alpha = \mathrm{AI}\frac{n-Q}{Q-1}$ и $\beta = (1 - \mathrm{AI})\frac{n-Q}{Q-1}$.

Два крайних значения параметра Q — это 1 и n. Когда Q стремится к n, α и β стремятся к 0, а вероятности всех исходов кроме 0 и n стремятся к 0. Предельное распределение равно AI в исходе 0, и равно AI в исходе n. Когда Q приближается к 1, α и β стремятся к $+\infty$, а распределение стремится к биномиальному распределению с вероятностью AI.

Один из недостатков бета-биномиального распределения заключается в отсутствии возможности моделирования недодисперсных распределений, то есть распределений, требующих присвоить параметру Q значение, меньшее 1. Тогда как большинство встречаемых нами образцов дают распределения с избыточной дисперсией, некоторые моделируются точнее при помощи недодисперсных распределений. Оставлять такие распределения в точке $Q = 1$ не имеет вычислительного смысла. Далее мы опишем способ обойти это препятствие при помощи экстраполяции бета-биномиального распределения за границу $Q = 1$, в гипергеометрические распределения. Некоторые из описанных ниже вещей были впервые сделаны в статье [119].

Рассмотрим следующую модифицированную модель урны Полья. В качестве входных параметров модель будет иметь два вещественных числа α и β с тем же смыслом и дополнительный вещественный параметр d. Положим $\alpha \geq 0$ шаров первого типа и $\beta \geq 0$ шаров второго типа в урну. Достанем n шаров из урны, один за другим, как и в обычной модели. Модификация касается только того, как мы восполняем урну после каждого взятия. В бета-биномиальном случае, после взятия шара мы кладём обратно в урну 2 шара такого

же типа, увеличивая количество шаров в урне на один после каждого взятия. В нашей модификации, мы кладём в урну $d+1$ шар вместо двух, увеличивая количество шаров в урне на d после каждого взятия. Функция вероятности итогового расширенного бета-биномиального распределения довольно похожа на исходную функцию вероятности бета-биномиального распределения, за исключением того, что шаг арифметических прогрессий в числителе и знаменателе теперь равен d, а не 1:

$$eBB(m \mid n = m+p, \alpha, \beta, d) = \binom{n}{m} \frac{\prod_{i=0}^{m-1}(\alpha + id) \cdot \prod_{j=0}^{p-1}(\beta + jd)}{\prod_{k=0}^{n-1}(\alpha + \beta + kd)}.$$

Несколько свойств этого распределения следуют сразу из определения.

1. Расширенное распределения не зависит от одновременного масштабирования параметров, то есть при любом $x > 0$

$$eBB(m \mid n = m+p, \alpha x, \beta x, dx) = eBB(m \mid n = m+p, \alpha, \beta, d).$$

2. Если $d = 1$, расширенное распределение совпадает с бета-биномиальным распределением с параметрами α и β.
3. Если $d = 0$, расширенное распределение совпадает с биномиальным распределением $\mathrm{AI} = \frac{\alpha}{\alpha+\beta}$.
4. Гипергеометрическое распределение $HG(m \mid n = m+p, M, P)$ принадлежит этому семейству как

$$eBB(m \mid n = m+p, \alpha = M, \beta = N, d = -1).$$

Иначе говоря, расширенное семейство содержит в себе бета-биномиальное семейство, биномиальное семейство и гипергеометрическое семейство. Из свойства 1 мы заключаем, что расширенное семейство двумерное, а не трёхмерное, и может быть описано при помощи параметров $0 \le \mathrm{AI} \le 1$ и $Q \le n$ аналитическим продолжением формул из бета-биномиального случая. Описание через параметры AI и Q может быть записано следующим образом:

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

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

Биномиальные распределения с параметром $Q = 1$ перестали быть крайним случаем, так как теперь возможно рассматривать регион $Q < 1$. Также мы можем однообразно моделировать недобиномиальную и сверхбиномиальную дисперсию.

Поиск Q с наибольшим правдоподобием

Мы реализовали несколько функций на языке R для точного и быстрого вычисления функции вероятности расширенного бета-биномиального распределения.

Определяющая формула функции вероятности расширенного бета-биномиального распределения не может быть напрямую применена для точных вычислений в модели вещественных чисел с плавающей запятой стандарта IEEE 754. Дело в том, что и числитель, и знаменатель интересующего нас выражения обычно на несколько порядков превышают настоящее значение дроби в абсолютном значении, поэтому прямая реализация будет терять значимые знаки после запятой. Дабы избежать потери точности, вместо вычисления числителя и знаменателя по отдельности, мы составляем попарное сочетание множителей в числителе и множителей в знаменателе, вычисляем частное в каждой паре и перемножаем частные между собой. В приложениях нам будет интересно значение логарифма функции вероятности, поэтому мы можем применить ещё один вычислительный приём для увеличения числа значимых знаков. Вместо прямого вычисления логарифма каждой дроби, мы раскладываем дробь как сумму единицы и остатка, и на остатке используем специальную функцию `log1p`. В целом, мы получаем

$$\begin{gathered}\log eBB(m \mid n, \alpha, \beta, d) = \\ = -\sum_{i=0}^{m-1} \texttt{log1p}\left(\frac{\beta}{\alpha+\beta+id}\right) + \\ + \sum_{j=0}^{p-1} \texttt{log1p}\left(\frac{-(j+1)\alpha + m\beta - md}{(j+1)(\alpha+\beta+(m+j)d)}\right).\end{gathered}$$

Взятие дифференциала от этого выражения не представляет проблем.

Эти вычисления составляют одну из основных частей решения следующей оптимизационной задачи.

Дан вектор аллельных покрытий $m_l, p_l, 0 \leq l \leq g-1$,
и вектор соответствующих оценок аллельных дисбалансов $\mathrm{AI}_l, 0 \leq l \leq g-1$,
Найти Q которое максимизирует правдоподобие

$$L(Q) = \prod_{l=0}^{g-1} eBB(m_l \mid m_l + p_l, \mathrm{AI}_l, Q).$$

Подсчёт избыточной дисперсии при помощи внешних РНК-контролей

Одно из самых важных свойств протокола внешних РНК-контролей — это большое количество образцов, задействованных в вычислении избыточной дисперсии. Из-за этого мы ожидаем, что общая сумма аллельных покрытий в контрольных образцах сильно выше, нежели в каждом из образцов по отдельности. Моделирование аллельных покрытий в таком случае становится затруднительным, так как в общем случае сумма двух бета-биномиальных распределений не является бета-биномиальным распределением. Для того, чтобы сохранить возможность аналитического моделирования, мы сперва допускаем, что коэффициенты избыточной дисперсии в образцах близки друг к другу. В таком случае сумма двух распределений будет хорошо приближаться бета-биномиальным распределением, иначе говоря $m_1 + m_2 \sim eBB(n_1 + n_2, \mathrm{AI}, Q = \mathrm{iQCC}^2)$ когда $Q_1 \simeq Q_2$, и m_i и n_i обозначают материнские и отеческие покрытия соответственно. Образцы с контрольной РНК не удовлетворяют всем критериям технических реплик, однако они приходят из одного источника РНК. В таком случае наилучшая оценка аллельного дисбаланса может быть посчитана исходя

из общего распределения прочтений во всех образцах:

$$\mathrm{AI}_j = \frac{\sum_{i \neq j} \frac{m_i}{Q_i}}{\sum_{i \neq j} \frac{(m_i + p_i)}{Q_i}} \tag{4.1}$$

для данного образца j и набора заданных значений $Q_{i \neq j}$. Эта оценка может быть получена из бета-биномиальной модели с общим покрытием N. Когда N на несколько порядков превосходит n, дисперсией оценки AI_j можно пренебречь, и процедура подбора наилучших $Q_j = \mathrm{iQCC}_j^2$ сойдётся без проблем. Обычно в случае с контролем внешними РНК верно, что $N \gg n$. Однако, когда n и N имеют похожий порядок (например, когда количество образцов мало), оценка может сильно отклоняться от действительных значений AI_j, и значения Q_j будут неизбежно переоценены. На рисунке **4.3a,b** показано, насколько Q_j будет переоценено в зависимости от частного $n/(n+N)$.

Если все значения Q_i одного порядка, одного цикла алгоритма подбора будет достаточно. Его начальными значениями будут $Q_i = 1$, поэтому выражение (4.1) упростится до $\mathrm{AI}_j = \sum_{i \neq j} m_i / \sum_{i \neq j} (m_i + p_i)$.

Иначе, если после первого шага алгоритма обнаружится, что значения Q_i имеют сильный разброс по образцам, мы можем повторять его шаги, перемежая две процедуры: подбор Q и подбор AI. Подбор AI выполняется при помощи формулы (4.1), а подбор Q происходит при помощи градиентного спуска. После каждого шага подбора Q, новые значения сравниваются со старыми, и если расхождение становится меньше порогового, алгоритм останавливается.

Подсчёт избыточной дисперсии в случае технических реплик

В случае, когда количество образцов невелико, мы должны решить проблему переоценки Q_j. Для этого мы объединяем верхнюю оценку с нижней: если мы включаем образец j в оценку аллельного дисбаланса j-го образца, то соответствующие подобранные Q_j будут недооценены. Более того, на симулированных данных мы видим, что нижняя и верхняя оценка iQCC отличаются от действительного значения на *одно и то же отклонение* в логарифмической шкале, см. Рисунок **4.3b**. В случае технической репликации мы используем это

для того, чтобы приближенно оценить iQCC_j при помощи геометрического среднего верхней и нижней оценки (Рисунки **4.3c, B.2**). Этот приём позволяет нам использовать тот же набор процедур и на небольшом количестве образцов, если верно предположение, что образцы имеют похожую избыточную дисперсию.

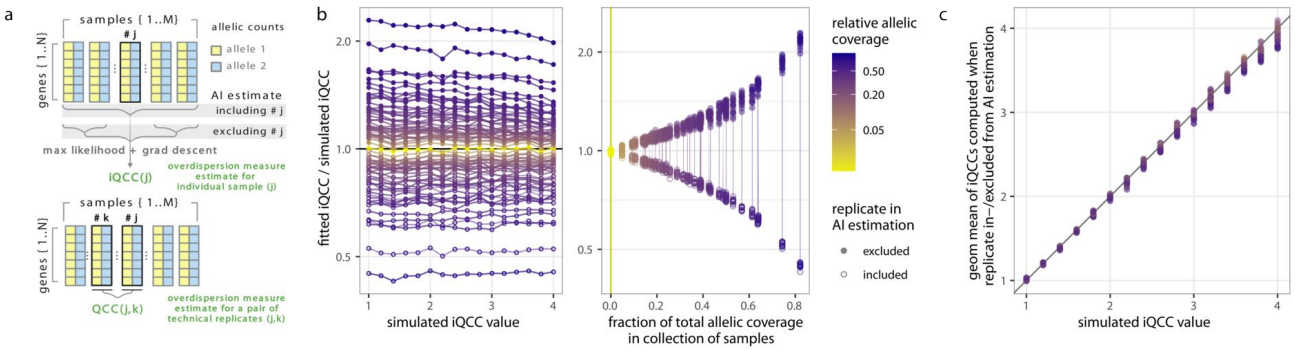

Рисунок 4.3 — Алгоритм вычисления iQCC принимает широкий диапазон количеств образцов и размеров библиотек.

(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 (см. Методы). На левой панели — вместе с увеличением аллельного покрытия верхняя и нижняя оценки сходятся к действительному значению iQCC (золотое); правая панель — 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. **(b-c)** Симуляции данных реплик (все созданные iQCC были близки для всех реплик, что симулирует техническую репликацию или похожие на неё условия экспериментов) с разными общими покрытиями и уровнями избыточной дисперсии. 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.

4.1.2 Данные

Информация о наборах данных, использовавшихся в этой работе, включая источники РНК, протоколы подготовки библиотек, секвенирование и обработку данных, обобщена в Таблице 1.

Таблица 1 — Описание данных, использованных в проекте `controlFreq`.

Эксп. #	1			2		3		
GEO Подсерия	GSE228002			GSE228003		GSE228004		
Образцы	**18 библиотек** со смесями человеческой (H), мышиной (M) and *C.elegans* (W) РНК. H1:M1, H1:W и M1:W (75:25 и 50:50 от общего аллельного покрытия), x3 техн. реплик. W РНК – 1:1 смесь N2 и Hawaii РНК. **21 библиотека** с одновидовой РНК: (H1, H2, H3, M1, M2, M3, W) x3 техн. реплик. H1-3 и M1-3 – биологические реплики (клетки из соседних лунок).			**12 библиотек**: смеси РНК из 3 разных мышей с *C.elegans* РНК-контролями W1, W2 и W3 (N2:Haw РНК смеси, 1:1, 2:1, 1:2, соотв.). **Включая**: 6 библиотек M1 (селезенка + W1, селезенка + W2, печень1 + W1, печень1 + W2, печень2 + W1, печень2 + W3); 4 библиотеки M2 (селезенка + W1, селезенка + W2, печень + W1, печень + W2); 2 библиотеки M3 (печень + W1, печень + W3).		**32 библиотеки**:РНК лимфоцитов из 3 доноров (H1-H3), с M и W (1:1 N2:Haw) РНК-контролями. Включая: множество (H1a, H1b, H2, H3) x (M РНК 10% от образца, M 25%, M 50%) – всего 12 библиотек; (H1a, H1b, H2, H3) x (W 10%, W 25%, W 50%) – 12 библиотек; (H1a, H1b, H2, H3) x 2 техн. реплик без РНК-контролей – 8 библиотек. **9 библиотек**: ((H4, H5, H6, H7) + M 10%) x 2 био реплики, H1c + M 10%; **9 библиотек** – то же с W 10%; **9 библиотек** со смесями N2:Haw РНК (1:1, 3:1, 1:3) x3 техн. реплики.		
Данные	Illumina 150PE			Illumina 150PE		Illumina 151PE		
Lib. Prep	NEBNext Single Cell Low Input RNA			NEBNext Single Cell Low Input RNA		SMART-Seq v4 PLUS		
Организм	Человек	Мышь	*C.elegans*	Мышь	*C.elegans*	Человек	Мышь	*C.elegans*
RNA Prep.*	Agilent Mini	Agilent Mini	Trizol	Trizol	Trizol	Qiagen Mini	Agilent Mini	Trizol
Генотипы**	NA12878	129xCastF1 (Abl.1 клон)	РНК смесь (N2 + Hawaii)	129xCastF1 (органы, био реплики)	РНК смесь (N2 + Hawaii)	Доноры 1-7 ***	129xCastF1 (Abl.1 клон)	РНК смесь (N2 + Hawaii)
Референсы	GRCh38 p13	GRCm38 68	PRJNA13758 WS276	GRCm38 68	PRJNA13758 WS276	GRCh38 p13	GRCm38 68	PRJNA13758 WS276
Варианты	Illumina Pt genomes, 2017	Sanger dbSNP142, 2014	CeNDR, soft-filtered, 2020	Sanger dbSNP142, 2014	CeNDR, soft-filtered, 2020	Imputed, this study	Sanger dbSNP142, 2014	CeNDR, soft-filtered, 2020
Замечания	* RNA prep: Agilent Absolutely RNA miniprep (Agilent Mini), Qiagen RNeasy miniprep (Qiagen Mini). ** Генотипы: 129S1/SvImJ x CAST/EiJ F1 (129XCastF1). *** Доноры: #1 представлен в 3 био репликах, #2-5 - в 2х, #6-7 - в 1ой; M1,M2,M3 и W1,W2,W3 нотации различаются в разных экспериментах, см. описания.							

Биологический материал: Клональная линия про-B клеток мыши Abl.1 [24], полученная из женской особи мыши (гибрид F1 129S1/SvImJ × Cast/EiJ) и клональная линия про-B клеток человека GM12878.DF1 [26] выращивались в среде RPMI (Gibco), содержащей 15% FBS (Sigma), 1X L-Глутамин (Gibco), 1X Пенициллин/Стрептомицин (Gibco) и 0.1% β-меркаптоэтанол (Sigma). Использование человеческих моноцитов периферической крови из деидентифицированных доноров было соответствующим образом согласовано для геномного секвенирования и неограниченного размещения в публичных базах данных. Ли-

нии *C. elegans* N2 и CB4856 (Hawaiian) выращивались при температуре 20°C в чашках Петри со средой для выращивания нематод (NGM) и с Escherichia coli OP50 [120].

Приготовление РНК: Биологические реплики человеческих и мышиных образцов (см. Табл. 1) были получены путем выращивания клеток в отдельных лунках 6-луночного планшета при плотности посева 500 000 на мл (1 500 000 клеток на лунку). Выделение РНК и обрабока ДНКазой для обеих клональных клеточных линий производились с помощью набора Absolutely RNA Microprep Kit (Agilent) в соответствии с инструкцией. Выделение РНК из C.elegans производилось тризольным методом (Invitrogen), с обработкой ДНКзой с помощью набора TURBO DNA-free kit (Ambion). Целостность РНК оценивали с помощью прибора Bioanalyzer и количественно определяли с помощью Qubit RNA HS Assay. Обработанные ДНКазой РНК из линий N2 и Hawaiian были смешаны в пропорциях, описанных в Таблице 1. РНК мышей была получена из цельных тканей взрослых особей мышей (гибрид F1 129S1/SvImJ × Cast/EiJ), содержавшихся в мышином питомнике Института рака Дана-Фарбер (DFCI), источник животных - Jackson Laboratories (Bar Harbor). Все работы с животными проводились в соответствии с протоколом DFCI 09-065, одобренным Комитетом по уходу и использованию животных DFCI. Животные содержались в соответствии с Руководством по уходу и использованию лабораторных животных. Собранные ткани измельчали с помощью ступки и пестика в жидком азоте. Эти измельченные в порошок ткани либо брали непосредственно для выделения РНК с помощью реактива Trizol, либо хранили в жидком азоте для последующего использования.

Приготовление библиотек и секвенирование: В экспериментах 1 и 2 (см. Табл. 1), аликвоты общей РНК использовались для подготовки библиотек с помощью набора NEBNext Single Cell/Low Input RNA Library Prep Kit (NEB). Библиотеки были секвенированы компанией Novogene на платформе Illumina NovaSeq, были сгенерированы парные чтения длиной 150 п.н. В эксперименте 3 (см. Табл. 1), библиотеки были подготовлены с помощью набора SMART-Seq v4 PLUS kit (Takara), а секвенирование с производством парных чтений 151 п.н. проводилось на Illumina NovaSeq в Институте Биомедицинских наук Altius.

Поиск вариантов и импутация: Импутацию генотипов в образцах доноров в эксперименте 3 проводили только на аутосомах. Для разделения файлов по хромосомам использовался `VCFTools`. Файлы генотипов были посчитаны на сервере TopMed (imputation.biodatacatalyst.nhlbi.nih.gov) со следующими параметрами: `imputation=minimac4-1.0.2; phasing=eagle-2.4; panel=TOPMED.vR2; Mode: Quality Control & Imputation`. После исключения 131 562 позиций (131 452 из которых не являлись полиморфизмом), были оставлены 572 955 позиций для дальнейшей импутации. После импутации были оставлены только маркеры со значением R2 более 0.9. *De novo* SNP позиции были аннотированы с помощью базы dbSNP (v151).

4.1.3 Генерация таблиц аллельных покрытий

Подготовка референсных файлов: Чтобы получить *in silico* химерные псевдогеномные референсные файлы, "содержащие" больше одного организма, нами были созданы соответствующие псевдогеномные файлы для каждого из них, а именно — псевдо-референсные FASTA файлы для обоих гаплотипов (см. подробности в главе 2 и [106]), аллельные "F1" VCF файлы (однонуклеотидные полиморфизмы, с одним из гаплотипов в качестве референсной аллели, см. главу 2) и не изменённый референсный GTF файл. Далее, эти файлы были совмещены в один, с соответствующим переименованием хромосом для последующего различения (например, хромосома 1 может стать "1h" в человеке и "1m" в мыши). Полученные файлы были использованы в последующей обработке данных в качестве референса. Индексация *in silico* химерных псевдогеномов производится с помощью `STAR`, с приравниванием `sjdbOverhang` параметра рекомендованному специфичному для данных значению `len(read)-1`.

Предварительная обработка данных: Отметим, что мы пересмотрели рекомендации по созданию таблиц аллельных покрытий, представленные в Главе 2, чтобы избавиться от проблемы потенциальной статистической зависимости значений покрытия на соседних вариантах при работе с организмами с высоким процентом полиморфизмов (см. детали по ссылке:

`github.com/gimelbrantlab/controlFreq`): в этой главе единицей анализа избыточной дисперсии является таблица сумм аллель-распределённых прочтений, покрывающих каждый ген. Чтобы получить такую таблицу, мы сначала выравниваем РНК-прочтения на оба псевдогенома с помощью версии `STAR` (v2.7.9a), которая поддерживает выравнивание с учетом аллелей (`-outSAMattributes vA vG`) при предоставлении предварительно подготовленных FASTA и VCF (подробнее см. раздел выше). Для дальнейшего анализа мы выбираем прочтения, покрывшие хотя бы один SNP для определения аллели, и отсеиваем прочтения, не показывающие консистентности в позициях выравнивания на два гаплотипа или имеющие противоречивые полиморфизмы, сигнализирующие о разных аллелях. Оставшиеся прочтения приписываются к одному из гаплотипов в соответствии с вариантами в позициях SNP и качеством выравнивания. Наконец, мы присваиваем аллельные прочтения генам и рассчитываем покрытия генов с помощью `featureCount` (v2.0.2) (с параметрами `-countReadPairs -B -C`). Для расчета оценок iQCC использовались аутосомные гены.

4.2 Результаты

4.2.1 Смеси РНК из существенно генетически различающихся организмов показывают одинаковую избыточную дисперсию во всех компонентах

Ранее мы показали, что избыточная дисперсия алелльного дисбаланса в экспериментах поли-А РНК-секвенирования ведет себя как мультипликативный параметр к биномиальной дисперсии, одинаковый для всех генов в образце. Мы предположили, что если мы смешаем РНК двух сильно различающихся видов, то избыточная дисперсия будет одинаковой для всех компонент. Мы разработали экспериментальный и вычислительный протокол (оперирующий усовершенствованной мерой избыточной дисперсии, iQCC, индивидуальным коэффициентом коррекции качества, см. Рис. 4.2а), и показали схожесть избыточной дисперсии, измеренной между различных участками генома, включая хромосомы

различного видового происхождения в наших *in-silico* химерах (см. Рис. 4.2b, Рис. B.2, Рис. 4.4).

Примечательно, что данные человека, мыши и круглого червя эффективно картируются на собственный геном в любых своих сочетаниях. Количество уникально выравненных прочтений изменилось на сотые доли процента (см. Рис. B.2c,d) при выравнивании человеческих образцов на смесь мыши и человека, при этом менее 0,03% уникально картированных прочтений было выравнено на неправильный организм (см. Рис. 4.2c-e). Как и ожидалось, при сравнении картирования на объединенный референсный геном и картирования на одновидовый референс не было обнаружено генов, демонстрирующих дифференциальный AI.

4.2.2 Использование одной РНК для многих образцов может выступать заменой технической репликации

Анализ избыточной дисперсии на компонентах контрольной РНК отличается от анализа на технических репликах, поскольку в этом случае мы не связаны строгим предположением об общей избыточной дисперсии между всеми образцами в эксперименте. Чтобы соответствовать измененным критериям, нам пришлось модифицировать и улучшить наш предыдущий метод вычисления избыточной дисперсии, который в настоящее время покрывает оба функционала: либо позволяет пользоваться предположением об одинаковой избыточной дисперсии (случай технических реплик), либо опирается на большое количество реплик (случай РНК-контролей, или spike-in) и использует предположение об одинаковом состоянии аллельных пропорций в изначальной РНК в обоих сценариях (см. Рис. 4.3a, и Методы). Наконец, оценки избыточной дисперсии (iQCC), полученные с помощью текущей процедуры, сильно коррелируют с оценками QCC, вычисленными с помощью предыдущего метода, `Qllelic` (см. Рис. B.1).

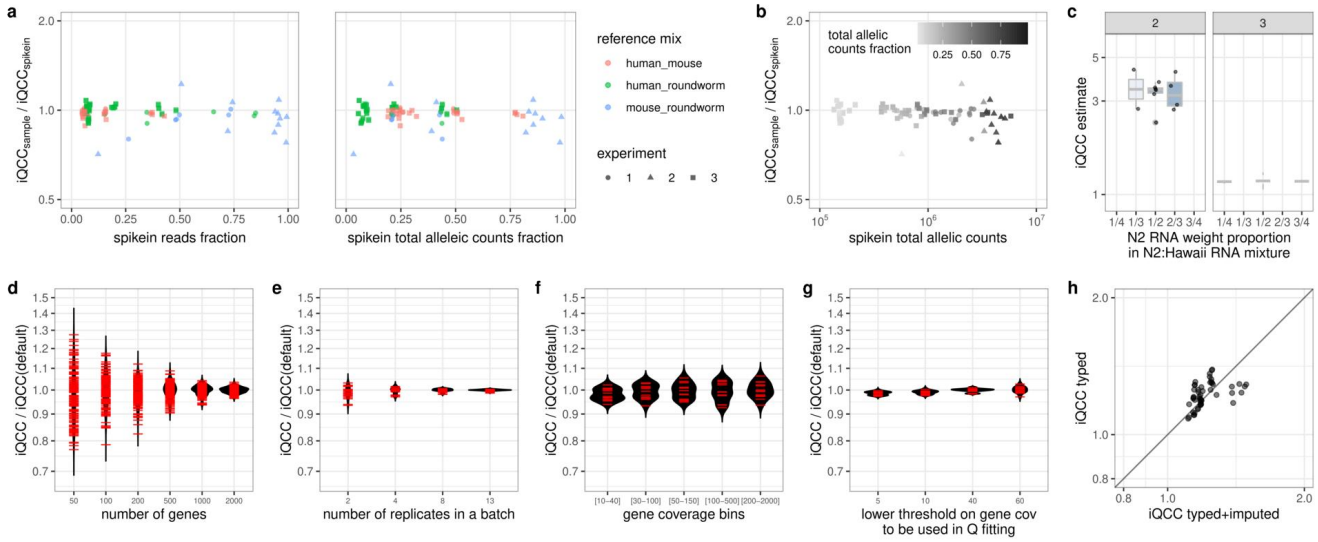

Рисунок 4.4 — Оценка избыточной дисперсии устойчива к варьированию относительного количества и состава РНК-контролей.

(a) Соотношение значений iQCC для подмножеств библиотеки из организмов 1 и 2 (обозначенных как "основной образец" и контроль, spike-in) остается близким к 1 с уменьшением доли контрольной РНК в исходной смеси (информация про образцы доступна в Табл. 1). **(b)** Соотношение значений iQCC для организмов 1 и 2 при различных общих аллельных покрытиях (отражающих размер библиотеки) против доли прочтений из РНК-контроля. Точность 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 репликах (образцы 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: эксперимент 3, мышиный РНК-контроль. По умолчанию, значения iQCC рассчитаны с порогом аллельного покрытия = 30 и с использованием всех 21 образцов с РНК-контролем, без дополнительных ограничений. **(h)** Значения iQCC вычислены на образцах эксперимента #3, содержащих РНК человека, с использованием или отсутствием импутированных вариантов.

4.2.3 Протокол использования РНК-контролей является достаточно гибким и позволяет варьировать параметры

С целью определить ограничения методологии РНК-контролей, мы варьировали и тестировали различные условия. Комбинации различных пар организмов, смешанных в разных пропорциях, привели к сопоставимым измерениям избыточной дисперсии в обоих компонентах (см. Рис. 4.4a). Более то-

го, единственным ограничением является общее аллельное покрытие образца (см. Рис. 4.4b,f), которое очевидным образом зависит от пропорции контрольной РНК, общего количества секвенированных прочтений и плотности полиморфизмов в организме (для человека она составляет порядка 1/1000, а для использованных нами образцов мышей и круглых червей — 1/100 и 1/500 соответственно). Более того, смеси РНК, приготовленные из отдельных родительских штаммов *C.elegans* (N2 и Hawaiian (CB4856)), оказались удобной заменой гетерозиготного гибрида N2xHawaiian (см. Рис. 4.2b и Рис. 4.4a). Мы также показали, что РНК из разных штаммов можно смешивать в разных пропорциях без потери качества оценки избыточной дисперсии (см. Рис. 4.4c и Обсуждение). Это делает такой способ экспериментальной подготовки РНК-контролей намного более простым, чем разведение червей или мышей.

4.3 Обсуждение результатов

Как уже говорилось ранее [106], аллель-специфический анализ эксперимента РНК-секвенирования без технических реплик подвержен фундаментальной неопределенности в отношении вклада технического шума в ложноположительные результаты (см. Рис. 4.1). Эта неопределённость больше при высокой технической избыточной дисперсии — например, iQCC > 4 (см. Рис.4.2b) эквивалентен переоценке уровня аллельного покрытия более чем в 16 раз при использовании биномиального теста. Таким образом, обычные пороговые методы отсечения недостаточно покрытых данных (такие как “не менее 10 прочтений на ген”) могут быть недостаточны и обманчивы.

Новый подход с РНК-контролями, `controlFreq`, точен не менее, чем представленный ранее `Qllelic`, и включает его функциональность. Помимо этого, он не требует увеличения стоимости эксперимента в несколько раз, а подразумевает дополнительное увеличение стоимости эксперимента всего на $\sim 5-10\%$.

Усовершенствование вычислительного протокола включает в себя переход от попарного подсчёта коэффициента коррекции качества, QCC, к расчету индивидуальных коэффициентов коррекции качества (iQCC) для каждого образца с учётом информации со всех образцов с общим контролем. Когда

`controlFreq` используется для анализа экспериментов с небольшим числом реплик (например, 2 или 3), он имеет то же ограничение, что и `Qllelic`, а именно, предположение о сходстве избыточной дисперсии между репликами, однако больше не требует сходства в размерах общих покрытий образцов. При наличии большого числа реплик (в виде РНК-контролей), новый метод не нуждается в сходстве избыточной дисперсии во всех образцах, и также позволяет уверенно детектировать выбросы среди образцов, и, в целом, очень устойчив к колебаниям различных параметров.

Важной практической особенностью подхода с использованием РНК-контролей является то, что контрольная РНК должна быть одинаковой в пределах одного набора образцов. Как только поправки на избыточную дисперсию рассчитаны для партии образцов, их можно корректно сравнивать с образцами из любой другой партии, включая те, в которых используются совершенно другие РНК-контроли (или, более того, любой образец с правильно оцененной избыточной дисперсией на аллельных данных). Это означает, что контрольную РНК можно готовить по мере необходимости и не обязательно из одного организма. Должно быть возможным одновременное использование других контролей в эксперименте: например, синтетические РНК стандарты “секвины” [121] или стандарты ERCC для контроля числа клеток [66] могут быть использованы без помех описанному выше процессу.

Насколько нам удалось установить, единственными требованиями к материалу для РНК-контролей являются (i) наличие в нем достаточной плотности полиморфизмов (чем выше, тем лучше) и (ii) способность пройти тот же процесс подготовки библиотеки, что и основной образец. Так, например, бактериальная РНК не подходит для приготовлении библиотек с использованием выделения поли-А-содержащей мРНК, а РНК *C.elegans* не может использоваться в качестве контроля для образцов млекопитающих при использовании протокола с удалением рРНК. В остальном, материал для РНК-контролей может быть получен любым удобным способом; мы полагаем, что после того, как состояние аллельных пропорций “зафискировано” для всей библиотеки, происхождение компонентов библиотеки не имеет значения. Например, мы успешно использовали “синтетические гибриды” — смеси РНК гомозиготных линий *C.elegans* (см. Рис.4.4c), что существенно проще, чем получение гетерозиготных организмов путём скрещивания. Вполне возможно, что также получится использовать

дрожжевую РНК или разработать набор синтетических молекул для аллель-специфического анализа РНК-секвенирования, аналогично стандартам ERCC [67]. Примечательно, что для определения уровня избыточной дисперсии требуется очень небольшое покрытие: для эффективного использования РНК-контролей было достаточно 125 тысяч в качестве общего аллельного покрытия (см. Рис.4.4ab). Следовательно, в случае контрольного организма с высокой плотностью SNP, требуется сравнительно небольшое общее количество прочтений: как мы обсуждали в главах 2 и 4, избыточная дисперсия одинакова на любом участке генома.

От длины прочтений очевидным образом зависит общее аллельное покрытие образца: чем короче прочтение, тем меньше вероятность покрыть различающий аллели полиморфизм и тем выше шанс быть выровненным на несколько участков. Например, при человеческих парных прочтениях длины 50×2 ожидаемое общее аллельное покрытие будет примерно в 3 раза меньше, чем при прочтениях длины 150×2 (см. Рис.B.3). Необходимое общее покрытие основного образца зависит от целей исследования. Полезным ориентиром является то, что для надежного обнаружения аллельного дисбаланса 80:20 в интересующем гене (с учетом поправки Бонферрони при анализе на 1000 генах), достаточное аллельное покрытие составляет примерно $50 \times (\text{iQCC}^2)$; $110 \times (\text{iQCC}^2)$ – для дисбаланса 70:30, и $420 \times (\text{iQCC}^2)$ – для дисбаланса 60:40.

Отметим снова, что описанные выше эксперименты были проведены с выделения поли-А-содержащей мРНК в процессе подготовки библиотек. Хотя мы ожидаем, что применение этого подхода может быть экстраполировано на множество других экспериментальных методик, в которых важно точное измерение аллельного сигнала, в каждом конкретном случае следует проводить пилотное исследование, чтобы проверить применимость протокола. Ярким примером возможного применения является измерение AI при одноклеточном РНК-секвенировании: это стало бы существенным продвижением, поскольку иначе техническая репликация на уровне одной клетки невозможна [118] (разделение одноклеточной РНК на две реплики чрезвычайно сложно технически [27]). При этом известно, что в одноклеточном РНК-секвенировании уровень избыточной дисперсии ещё выше, а покрытия генов для подавляющего большинства генов очень низкие, что только увеличивает потребность в надёжных методах, позволяющих отличить технический шум от значимых биологических вариаций.

Наконец, аналогичный подход может быть применим и к другим типам данных, помимо уровня экспрессии – таким как метилирование ДНК, ДНК-белковые взаимодействия (например, ChIP-seq), открытость хроматина (например, ATAC-seq, HiC).

Заключение

Основные результаты работы 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. 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`. 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, в сравнение с широко используемым биномиальным тестом [89—92] и методами, оценивающими избыточную дисперсию из одной реплики [7; 61].

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

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.

Моноаллельная аутосомная экспрессия и эпигенетическая цис-регуляция. Подход с использованием QCC-коррекции был применен в двух опубликованных исследованиях, посвященных дифференциальной ASE в клональных культурах. Одно из них включало анализ популяций иммунных клеток, развившихся *in vivo* из отдельных гемопоэтических стволовых клеток [45].

Другое исследование [51; 122] составляет основу главы 3. 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 (см. Рис.3.3g), 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 [32] и между разными организмами, например, человеком и мышью [24; 110; 111]. Важно отметить, что деметилирование ДНК влияет не на все MAE гены, 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 [28; 29].

Используя `Qllelic`, мы оценили влияние полногеномного деметилирования ДНК на транскрипционный аллельный дисбаланс. В четырех проанализированных клонах мы обнаружили значительные сдвиги AI в более чем 600 аутосомных генах. Точный количественный анализ данных РНК-секвенирования выявил более сложный ландшафт клонального разнообразия в аллель-специфической регуляции генов, чем подразумевает моноаллельная/биаллельная дихотомия. 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.

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 ор-

ганизме 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.

Выводы

1. Разработан новый вычислительный подход, реализованный в пакете на языке `R`, `Qllelic`, 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-секвеинрования 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 механизмом 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 делает возможным 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.

Благодарности

Я писала диссертацию под чутким руководством профессора **Михаила Сергеевича Гельфанда**, профессора **Александра Александровича Гимельбранта** и профессора **Андрея Александровича Миронова**. Я благодарю их за неиссякаемую поддержку, бесчисленное количество плодотворных бесед, и советы по научной работе и научной жизни, важность которых для меня трудно переоценить.

Также хочу выразить признательность профессору Шамилю Сюняеву за выделенное время и ценные замечания и идеи, касающиеся написания статьи. Эта работа также была бы невозможна без участия членов лаборатории А.А. Гимельбранта, особенно Светланы Виноградовой и Сомии Гупты, с которыми у нас сложилось наиболее тесное сотрудничество. Я искренне благодарна Сколтеху, Институту рака Дана-Фарбер и Институту биомедицинских наук Альтиус за предоставленные возможности в разные периоды моей аспирантуры. Вычислительно тяжёлая часть диссертации была выполнена на кластерах МГУ ("Макарыч"), Гарвардской Медицинской школы ("O2") и Альтиуса, и я признательна людям, поддерживающим их вычислительные инфраструктуры.

И, конечно же, моя благодарность направлена моему любящему мужу, семье и друзьям (особенно – Зое, Вале и Алисе) за их помощь и непоколебимую поддержку в тяжёлое для меня время. Отдельное спасибо кошке Басе за охрану моего психологического благополучия.

Список сокращений и условных обозначений

AI	Аллельный дисбаланс, определённый как пропорция одной определённой аллели
ASE	Аллель-специфическая экспрессия
CNV	Вариация числа копий
eBB	Расширенное бета-биномиальное распределение [глава 4][123]
GWAS	Полногеномный поиск ассоциаций
MAE	Моноаллельная аутосомная экспрессия [23]
QCC	Коэффициент коррекции качества, мера избыточной дисперсии [глава 2][106]
iQCC	Индивидуальный коэффициент коррекции качества, мера избыточной дисперсии [глава 4][123]
Q	iQCC^2
QTL	локус количественных признаков
SNP	Single Nucleotide Polymorphism
XCI	X-chromosome inactivation

Словарь терминов

Аллельное покрытие гена — Покрытие гена, учитывающее только информативные прочтения, которые пересекают позиции гетерозиготных полиморфизмов

Корзина генов — Набор всех генов в образце с покрытиями, лежащими в заданном интервале покрытий.

Общее аллельное покрытие образца — Сумма аллельных покрытий генов в образце

Относительное общее аллельное покрытие образца — Пропорция общего аллельного покрытия выбранного образца среди суммы общих покрытий образцов в наборе данных

РНК-контроль, или Spike-in — Внешняя РНК отличимого выравниванием от образца организма, добавляемая в серию образцов перед приготовлением библиотек

Избыточная дисперсия, или Свехрдисперсия — Избыточный шум относительно принятой модели (например, биномиальной)

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. Detection of regulatory variation in mouse genes / C. R. Cowles [и др.] // Nature Genetics. — 2002. — окт. — т. 32, № 3. — с. 432—437. — DOI: 10.1038/ng992. — URL: https://doi.org/10.1038/ng992.

12. Allelic Variation in Human Gene Expression / H. Yan [и др.] // Science. — 2002. — авг. — т. 297, № 5584. — с. 1143—1143. — DOI: 10.1126/science.1072545. — URL: https://doi.org/10.1126/science.1072545.

13. *Galupa R.*, *Heard E.* X-Chromosome Inactivation: A Crossroads Between Chromosome Architecture and Gene Regulation // Annual Review of Genetics. — 2018. — нояб. — т. 52, № 1. — с. 535—566. — DOI: 10.1146/annurev-genet-120116-024611. — URL: https://doi.org/10.1146/annurev-genet-120116-024611.

14. *Berletch J. B.*, *Yang F.*, *Disteche C. M.* Escape from X inactivation in mice and humans // Genome Biology. — 2010. — т. 11, № 6. — с. 213. — DOI: 10.1186/gb-2010-11-6-213. — URL: https://doi.org/10.1186/gb-2010-11-6-213.

15. *Carrel L.*, *Willard H. F.* X-inactivation profile reveals extensive variability in X-linked gene expression in females // Nature. — 2005. — март. — т. 434, № 7031. — с. 400—404. — DOI: 10.1038/nature03479. — URL: https://doi.org/10.1038/nature03479.

16. *Lyon M. F.* Gene Action in the X-chromosome of the Mouse (Mus musculus L.) // Nature. — 1961. — апр. — т. 190, № 4773. — с. 372—373. — DOI: 10.1038/190372a0. — URL: https://doi.org/10.1038/190372a0.

17. Global survey of escape from X inactivation by RNA-sequencing in mouse / F. Yang [и др.] // Genome Research. — 2010. — апр. — т. 20, № 5. — с. 614—622. — DOI: 10.1101/gr.103200.109. — URL: https://doi.org/10.1101/gr.103200.109.

18. *Glaser R. L.* The imprinted gene and parent-of-origin effect database now includes parental origin of de novo mutations // Nucleic Acids Research. — 2006. — янв. — т. 34, № 90001. — с. D29—D31. — DOI: 10.1093/nar/gkj101. — URL: https://doi.org/10.1093/nar/gkj101.

19. Genomic Imprinting and Physiological Processes in Mammals / V. Tucci [и др.] // Cell. — 2019. — февр. — т. 176, № 5. — с. 952—965. — DOI: 10.1016/j.cell.2019.01.043. — URL: https://doi.org/10.1016/j.cell.2019.01.043.

20. *Chess A.* Monoallelic Gene Expression in Mammals // Annual Review of Genetics. — 2016. — нояб. — т. 50, № 1. — с. 317—327. — DOI: 10.1146/annurev-genet-120215-035120. — URL: https://doi.org/10.1146/annurev-genet-120215-035120.

21. *Khamlichi A. A.*, *Feil R.* Parallels between Mammalian Mechanisms of Monoallelic Gene Expression // Trends in Genetics. — 2018. — дек. — т. 34, № 12. — с. 954—971. — DOI: 10.1016/j.tig.2018.08.005. — URL: https://doi.org/10.1016/j.tig.2018.08.005.

22. Allelic inactivation regulates olfactory receptor gene expression / A. Chess [и др.] // Cell. — 1994. — сент. — т. 78, № 5. — с. 823—834. — DOI: 10.1016/s0092-8674(94)90562-2. — URL: https://doi.org/10.1016/s0092-8674(94)90562-2.

23. Widespread Monoallelic Expression on Human Autosomes / A. Gimelbrant [и др.] // Science. — 2007. — нояб. — т. 318, № 5853. — с. 1136—1140. — DOI: 10.1126/science.1148910. — URL: https://doi.org/10.1126/science.1148910.

24. 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.

25. Stochastic Choice of Allelic Expression in Human Neural Stem Cells / A. R. Jeffries [и др.] // Stem Cells. — 2012. — авг. — т. 30, № 9. — с. 1938—1947. — DOI: 10.1002/stem.1155. — URL: https://doi.org/10.1002/stem.1155.

26. Chromatin signature of widespread monoallelic expression / A. Nag [и др.] // eLife. — 2013. — дек. — т. 2. — DOI: 10.7554/elife.01256. — URL: https://doi.org/10.7554/elife.01256.

27. Single-Cell RNA-Seq Reveals Dynamic, Random Monoallelic Gene Expression in Mammalian Cells / Q. Deng [и др.] // Science. — 2014. — янв. — т. 343, № 6167. — с. 193—196. — DOI: 10.1126/science.1245316. — URL: https://doi.org/10.1126/science.1245316.

28. 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.

29. 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.

30. FOXP3 Is an X-Linked Breast Cancer Suppressor Gene and an Important Repressor of the HER-2/ErbB2 Oncogene / T. Zuo [и др.] // Cell. — 2007. — июнь. — т. 129, № 7. — с. 1275—1286. — DOI: 10.1016/j.cell.2007.04.034. — URL: https://doi.org/10.1016/j.cell.2007.04.034.

31. Monoallelic expression of the murine gene encoding Toll-like receptor 4 / J. P. Pereira [и др.] // Nature Immunology. — 2003. — март. — т. 4, № 5. — с. 464—470. — DOI: 10.1038/ni917. — URL: https://doi.org/10.1038/ni917.

32. 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.

33. Risk alleles of genes with monoallelic expression are enriched in gain-of-function variants and depleted in loss-of-function variants for neurodevelopmental disorders / V. Savova [и др.] // Molecular Psychiatry. —

2017. — март. — т. 22, № 12. — с. 1785—1794. — DOI: 10.1038/mp.2017.13. — URL: https://doi.org/10.1038/mp.2017.13.

34. Signatures of Long-Term Balancing Selection in Human Genomes / B. D. Bitarello [и др.] // Genome Biology and Evolution / под ред. P. Majumder. — 2018. — март. — т. 10, № 3. — с. 939—955. — DOI: 10.1093/gbe/evy054. — URL: https://doi.org/10.1093/gbe/evy054.

35. Frequent monoallelic or skewed expression for developmental genes in CNS-derived cells and evidence for balancing selection / S. Branciamore [и др.] // Proceedings of the National Academy of Sciences. — 2018. — окт. — т. 115, № 44. — DOI: 10.1073/pnas.1808652115. — URL: https://doi.org/10.1073/pnas.1808652115.

36. *Nica A. C.*, *Dermitzakis E. T.* Expression quantitative trait loci: present and future // Philosophical Transactions of the Royal Society B: Biological Sciences. — 2013. — июнь. — т. 368, № 1620. — с. 20120362. — DOI: 10.1098/rstb.2012.0362. — URL: https://doi.org/10.1098/rstb.2012.0362.

37. *Wittkopp P. J.*, *Kalay G.* Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence // Nature Reviews Genetics. — 2011. — дек. — т. 13, № 1. — с. 59—69. — DOI: 10.1038/nrg3095. — URL: https://doi.org/10.1038/nrg3095.

38. Complex genetic dependencies among growth and neurological phenotypes in healthy children: Towards deciphering developmental mechanisms / L. Uechi [и др.] // PLOS ONE / под ред. A. Palsson. — 2020. — дек. — т. 15, № 12. — e0242684. — DOI: 10.1371/journal.pone.0242684. — URL: https://doi.org/10.1371/journal.pone.0242684.

39. *Kumasaka N.*, *Knights A. J.*, *Gaffney D. J.* Fine-mapping cellular QTLs with RASQUAL and ATAC-seq // Nature Genetics. — 2015. — дек. — т. 48, № 2. — с. 206—213. — DOI: 10.1038/ng.3467. — URL: https://doi.org/10.1038/ng.3467.

40. Leveraging allelic imbalance to refine fine-mapping for eQTL studies / J. Zou [и др.] // PLOS Genetics / под ред. X. Wen. — 2019. — дек. — т. 15, № 12. — e1008481. — DOI: 10.1371/journal.pgen.1008481. — URL: https://doi.org/10.1371/journal.pgen.1008481.

41. Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo / M. T. Maurano [и др.] // Nature Genetics. — 2015. — окт. — т. 47, № 12. — с. 1393—1401. — DOI: 10.1038/ng.3432. — URL: https://doi.org/10.1038/ng.3432.

42. Allele-specific NKX2-5 binding underlies multiple genetic associations with human electrocardiographic traits / P. Benaglio [и др.] // Nature Genetics. — 2019. — сент. — т. 51, № 10. — с. 1506—1517. — DOI: 10.1038/s41588-019-0499-3. — URL: https://doi.org/10.1038/s41588-019-0499-3.

43. Multiple Hepatic Regulatory Variants at the GALNT2 GWAS Locus Associated with High-Density Lipoprotein Cholesterol / T. S. Roman [и др.] // The American Journal of Human Genetics. — 2015. — дек. — т. 97, № 6. — с. 801—815. — DOI: 10.1016/j.ajhg.2015.10.016. — URL: https://doi.org/10.1016/j.ajhg.2015.10.016.

44. Tumor-suppressor genes that escape from X-inactivation contribute to cancer sex bias / A. Dunford [и др.] // Nature Genetics. — 2016. — нояб. — т. 49, № 1. — с. 10—16. — DOI: 10.1038/ng.3726. — URL: https://doi.org/10.1038/ng.3726.

45. In Vivo Clonal Analysis Reveals Random Monoallelic Expression in Lymphocytes That Traces Back to Hematopoietic Stem Cells / N. Kubasova [и др.] // Frontiers in Cell and Developmental Biology. — 2022. — авг. — т. 10. — DOI: 10.3389/fcell.2022.827774. — URL: https://doi.org/10.3389/fcell.2022.827774.

46. Global reference mapping of human transcription factor footprints / J. Vierstra [и др.] // Nature. — 2020. — июль. — т. 583, № 7818. — с. 729—736. — DOI: 10.1038/s41586-020-2528-x. — URL: https://doi.org/10.1038/s41586-020-2528-x.

47. Structural organization of the inactive X chromosome in the mouse / L. Giorgetti [и др.] // Nature. — 2016. — июль. — т. 535, № 7613. — с. 575—579. — DOI: 10.1038/nature18589. — URL: https://doi.org/10.1038/nature18589.

48. Chromosomal coordination and differential structure of asynchronous replicating regions / B. Blumenfeld [и др.] // Nature Communications. —

2021. — февр. — т. 12, № 1. — DOI: 10.1038/s41467-021-21348-4. — URL: https://doi.org/10.1038/s41467-021-21348-4.

49. Genetic determinants of chromatin reveal prostate cancer risk mediated by context-dependent gene regulation / S. C. Baca [и др.] // Nature Genetics. — 2022. — сент. — т. 54, № 9. — с. 1364—1375. — DOI: 10.1038/s41588-022-01168-y. — URL: https://doi.org/10.1038/s41588-022-01168-y.

50. Coordination of the random asynchronous replication of autosomal loci / N. Singh [и др.] // Nature Genetics. — 2003. — февр. — т. 33, № 3. — с. 339—341. — DOI: 10.1038/ng1102. — URL: https://doi.org/10.1038/ng1102.

51. RNA sequencing-based screen for reactivation of silenced alleles of autosomal genes / S. Gupta [и др.] // G3 Genes|Genomes|Genetics / под ред. J. Prendergast. — 2021. — дек. — т. 12, № 2. — DOI: 10.1093/g3journal/jkab428. — URL: https://doi.org/10.1093/g3journal/jkab428.

52. Allele-specific RNA imaging shows that allelic imbalances can arise in tissues through transcriptional bursting / O. Symmons [и др.] // PLOS Genetics / под ред. G. S. Barsh. — 2019. — янв. — т. 15, № 1. — e1007874. — DOI: 10.1371/journal.pgen.1007874. — URL: https://doi.org/10.1371/journal.pgen.1007874.

53. Single-molecule regulatory architectures captured by chromatin fiber sequencing / A. B. Stergachis [и др.] // Science. — 2020. — июнь. — т. 368, № 6498. — с. 1449—1454. — DOI: 10.1126/science.aaz1646. — URL: https://doi.org/10.1126/science.aaz1646.

54. DiMeLo-seq: a long-read, single-molecule method for mapping protein–DNA interactions genome wide / N. Altemose [и др.] // Nature Methods. — 2022. — апр. — т. 19, № 6. — с. 711—723. — DOI: 10.1038/s41592-022-01475-6. — URL: https://doi.org/10.1038/s41592-022-01475-6.

55. RNA-seq: technical variability and sampling / L. M. McIntyre [и др.] // BMC Genomics. — 2011. — июнь. — т. 12, № 1. — DOI: 10.1186/1471-2164-12-293. — URL: https://doi.org/10.1186/1471-2164-12-293.

56. *Varabyou A.*, *Salzberg S. L.*, *Pertea M.* Effects of transcriptional noise on estimates of gene and transcript expression in RNA sequencing experiments //

Genome Research. — 2020. — дек. — т. 31, № 2. — с. 301—308. — DOI: 10.1101/gr.266213.120. — URL: https://doi.org/10.1101/gr.266213.120.

57. *Jiang P.* Quality Control of Single-Cell RNA-seq // Methods in Molecular Biology. — Springer New York, 2019. — с. 1—9. — DOI: 10.1007/978-1-4939-9057-3_1. — URL: https://doi.org/10.1007/978-1-4939-9057-3_1.

58. *Kim B., Lee E., Kim J. K.* Analysis of Technical and Biological Variability in Single-Cell RNA Sequencing // Methods in Molecular Biology. — Springer New York, 2019. — с. 25—43. — DOI: 10.1007/978-1-4939-9057-3_3. — URL: https://doi.org/10.1007/978-1-4939-9057-3_3.

59. The Technology and Biology of Single-Cell RNA Sequencing / A. A. Kolodziejczyk [и др.] // Molecular Cell. — 2015. — май. — т. 58, № 4. — с. 610—620. — DOI: 10.1016/j.molcel.2015.04.005. — URL: https://doi.org/10.1016/j.molcel.2015.04.005.

60. A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data / D. A. Skelly [et al.] // Genome Research. — 2011. — Aug. — Vol. 21, no. 10. — P. 1728–1737. — DOI: 10.1101/gr.119784.110. — URL: https://doi.org/10.1101/gr.119784.110.

61. 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.

62. A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals / J. Chen [и др.] // Nature Communications. — 2016. — апр. — т. 7, № 1. — DOI: 10.1038/ncomms11101. — URL: https://doi.org/10.1038/ncomms11101.

63. Fast bootstrapping-based estimation of confidence intervals of expression levels and differential expression from RNA-Seq data / I. Mandric [и др.] // Bioinformatics / под ред. Z. Bar-Joseph. — 2017. — июнь. — т. 33, № 20. — с. 3302—3304. — DOI: 10.1093/bioinformatics/btx365. — URL: https://doi.org/10.1093/bioinformatics/btx365.

64. *Kharchenko P. V.*, *Silberstein L.*, *Scadden D. T.* Bayesian approach to single-cell differential expression analysis // Nature Methods. — 2014. — май. — т. 11, № 7. — с. 740—742. — DOI: 10.1038/nmeth.2967. — URL: https://doi.org/10.1038/nmeth.2967.

65. *Choi K.*, *Raghupathy N.*, *Churchill G. A.* A Bayesian mixture model for the analysis of allelic expression in single cells // Nature Communications. — 2019. — нояб. — т. 10, № 1. — DOI: 10.1038/s41467-019-13099-0. — URL: https://doi.org/10.1038/s41467-019-13099-0.

66. Assessing technical performance in differential gene expression experiments with external spike-in RNA control ratio mixtures / S. A. Munro [и др.] // Nature Communications. — 2014. — сент. — т. 5, № 1. — DOI: 10.1038/ncomms6125. — URL: https://doi.org/10.1038/ncomms6125.

67. Synthetic spike-in standards for RNA-seq experiments / L. Jiang [и др.] // Genome Research. — 2011. — авг. — т. 21, № 9. — с. 1543—1551. — DOI: 10.1101/gr.121095.111. — URL: https://doi.org/10.1101/gr.121095.111.

68. Single-cell RNA-seq analysis reveals ploidy-dependent and cell-specific transcriptome changes in Arabidopsis female gametophytes / Q. Song [и др.] // Genome Biology. — 2020. — июль. — т. 21, № 1. — DOI: 10.1186/s13059-020-02094-0. — URL: https://doi.org/10.1186/s13059-020-02094-0.

69. *Robinson M. D.*, *McCarthy D. J.*, *Smyth G. K.* edgeR: a Bioconductor package for differential expression analysis of digital gene expression data // Bioinformatics. — 2009. — нояб. — т. 26, № 1. — с. 139—140. — DOI: 10.1093/bioinformatics/btp616. — URL: https://doi.org/10.1093/bioinformatics/btp616.

70. Differential analysis of gene regulation at transcript resolution with RNA-seq / C. Trapnell [и др.] // Nature Biotechnology. — 2012. — дек. — т. 31, № 1. — с. 46—53. — DOI: 10.1038/nbt.2450. — URL: https://doi.org/10.1038/nbt.2450.

71. *Love M. I.*, *Huber W.*, *Anders S.* Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2 // Genome Biology. — 2014. — дек. — т. 15, № 12. — DOI: 10.1186/s13059-014-0550-8. — URL: https://doi.org/10.1186/s13059-014-0550-8.

72. QuASAR: quantitative allele-specific analysis of reads / C. T. Harvey [и др.] // Bioinformatics. — 2014. — дек. — т. 31, № 8. — с. 1235—1242. — DOI: 10.1093/bioinformatics/btu802. — URL: https://doi.org/10.1093/bioinformatics/btu802.

73. WASP: allele-specific software for robust molecular quantitative trait locus discovery / B. van de Geijn [и др.] // Nature Methods. — 2015. — сент. — т. 12, № 11. — с. 1061—1063. — DOI: 10.1038/nmeth.3582. — URL: https://doi.org/10.1038/nmeth.3582.

74. *Li B.*, *Dewey C. N.* RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome // BMC Bioinformatics. — 2011. — авг. — т. 12, № 1. — DOI: 10.1186/1471-2105-12-323. — URL: https://doi.org/10.1186/1471-2105-12-323.

75. A Bayesian approach for estimating allele-specific expression from RNA-Seq data with diploid genomes / N. Nariai [и др.] // BMC Genomics. — 2016. — янв. — т. 17, S1. — DOI: 10.1186/s12864-015-2295-5. — URL: https://doi.org/10.1186/s12864-015-2295-5.

76. Effect of method of deduplication on estimation of differential gene expression using RNA-seq / A. V. Klepikova [и др.] // PeerJ. — 2017. — март. — т. 5. — e3091. — DOI: 10.7717/peerj.3091. — URL: https://doi.org/10.7717/peerj.3091.

77. *Marx V.* How to deduplicate PCR // Nature Methods. — 2017. — апр. — т. 14, № 5. — с. 473—476. — DOI: 10.1038/nmeth.4268. — URL: https://doi.org/10.1038/nmeth.4268.

78. Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches / M. T. W. Ebbert [и др.] // BMC Bioinformatics. — 2016. — июль. — т. 17, S7. — DOI: 10.1186/s12859-016-1097-3. — URL: https://doi.org/10.1186/s12859-016-1097-3.

79. Elimination of PCR duplicates in RNA-seq and small RNA-seq using unique molecular identifiers / Y. Fu [и др.] // BMC Genomics. — 2018. — июль. — т. 19, № 1. — DOI: 10.1186/s12864-018-4933-1. — URL: https://doi.org/10.1186/s12864-018-4933-1.

80. Antisense artifacts in transcriptome microarray experiments are resolved by actinomycin D / F. Perocchi [и др.] // Nucleic Acids Research. — 2007. — сент. — т. 35, № 19. — e128. — DOI: 10.1093/nar/gkm683. — URL: https://doi.org/10.1093/nar/gkm683.

81. Actinomycin D Inhibition of DNA Strand Transfer Reactions Catalyzed by HIV-1 Reverse Transcriptase and Nucleocapsid Protein / W. R. Davis [и др.] // Biochemistry. — 1998. — сент. — т. 37, № 40. — с. 14213—14221. — DOI: 10.1021/bi9814890. — URL: https://doi.org/10.1021/bi9814890.

82. Molecular indexing enables quantitative targeted RNA sequencing and reveals poor efficiencies in standard library preparations / G. K. Fu [и др.] // Proceedings of the National Academy of Sciences. — 2014. — янв. — т. 111, № 5. — с. 1891—1896. — DOI: 10.1073/pnas.1323732111. — URL: https://doi.org/10.1073/pnas.1323732111.

83. Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis / L. A. Corchete [и др.] // Scientific Reports. — 2020. — нояб. — т. 10, № 1. — DOI: 10.1038/s41598-020-76881-x. — URL: https://doi.org/10.1038/s41598-020-76881-x.

84. *Thawng C. N.*, *Smith G. B.* A transcriptome software comparison for the analyses of treatments expected to give subtle gene expression responses // BMC Genomics. — 2022. — июнь. — т. 23, № 1. — DOI: 10.1186/s12864-022-08673-8. — URL: https://doi.org/10.1186/s12864-022-08673-8.

85. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis / M.-A. Dillies [и др.] // Briefings in Bioinformatics. — 2012. — сент. — т. 14, № 6. — с. 671—683. — DOI: 10.1093/bib/bbs046. — URL: https://doi.org/10.1093/bib/bbs046.

86. voom: precision weights unlock linear model analysis tools for RNA-seq read counts / C. W. Law [и др.] // Genome Biology. — 2014. — т. 15, № 2. — R29. — DOI: 10.1186/gb-2014-15-2-r29. — URL: https://doi.org/10.1186/gb-2014-15-2-r29.

87. limma powers differential expression analyses for RNA-sequencing and microarray studies / M. E. Ritchie [и др.] // Nucleic Acids Research. — 2015. — янв. — т. 43, № 7. — e47—e47. — DOI: 10.1093/nar/gkv007. — URL: https://doi.org/10.1093/nar/gkv007.

88. *Evans C.*, *Hardin J.*, *Stoebel D. M.* Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions // Briefings in Bioinformatics. — 2017. — февр. — т. 19, № 5. — с. 776—792. — DOI: 10.1093/bib/bbx008. — URL: https://doi.org/10.1093/bib/bbx008.

89. *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.

90. 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.

91. 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.

92. 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.

93. Transcriptome and genome sequencing uncovers functional variation in humans / T. Lappalainen [и др.] // Nature. — 2013. — сент. — т. 501, № 7468. — с. 506—511. — DOI: 10.1038/nature12531. — URL: https://doi.org/10.1038/nature12531.

94. *Bonferroni C. E.* Statistical class theory and calculation of probability // Publications of High R Institute of Economic and Commercial Sciences of Florence. — 1936.

95. *Battich N.*, *Stoeger T.*, *Pelkmans L.* Control of Transcript Variability in Single Mammalian Cells // Cell. — 2015. — дек. — т. 163, № 7. — с. 1596—1610. — DOI: 10.1016/j.cell.2015.11.018. — URL: https://doi.org/10.1016/j.cell.2015.11.018.

96. Imaging individual mRNA molecules using multiple singly labeled probes / A. Raj [и др.] // Nature Methods. — 2008. — сент. — т. 5, № 10. — с. 877—879. — DOI: 10.1038/nmeth.1253. — URL: https://doi.org/10.1038/nmeth.1253.

97. Counting absolute numbers of molecules using unique molecular identifiers / T. Kivioja [и др.] // Nature Methods. — 2011. — нояб. — т. 9, № 1. — с. 72—74. — DOI: 10.1038/nmeth.1778. — URL: https://doi.org/10.1038/nmeth.1778.

98. A Novel Statistical Approach for Jointly Analyzing RNA-Seq Data from F1 Reciprocal Crosses and Inbred Lines / F. Zou [и др.] // Genetics. — 2014. — февр. — т. 197, № 1. — с. 389—399. — DOI: 10.1534/genetics.113.160119. — URL: https://doi.org/10.1534/genetics.113.160119.

99. Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance / J. J. Crowley [и др.] // Nature Genetics. — 2015. — март. — т. 47, № 4. — с. 353—360. — DOI: 10.1038/ng.3222. — URL: https://doi.org/10.1038/ng.3222.

100. *Sherry S. T.* dbSNP: the NCBI database of genetic variation // Nucleic Acids Research. — 2001. — янв. — т. 29, № 1. — с. 308—311. — DOI: 10.1093/nar/29.1.308. — URL: https://doi.org/10.1093/nar/29.1.308.

101. STAR: ultrafast universal RNA-seq aligner / A. Dobin [и др.] // Bioinformatics. — 2012. — окт. — т. 29, № 1. — с. 15—21. — DOI: 10.1093/bioinformatics/bts635. — URL: https://doi.org/10.1093/bioinformatics/bts635.

102. Ensembl 2018 / D. R. Zerbino [и др.] // Nucleic Acids Research. — 2017. — нояб. — т. 46, № D1. — с. D754—D761. — DOI: 10.1093/nar/gkx1098. — URL: https://doi.org/10.1093/nar/gkx1098.

103. *Bishop C. M.* Pattern Recognition and Machine Learning (Information Science and Statistics). — Berlin, Heidelberg : Springer-Verlag, 2006. — ISBN 0387310738.

104. Counting absolute numbers of molecules using unique molecular identifiers / T. Kivioja [и др.] // Nature Methods. — 2011. — нояб. — т. 9, № 1. — с. 72—74. — DOI: 10.1038/nmeth.1778. — URL: https://doi.org/10.1038/nmeth.1778.

105. Quantitative single-cell RNA-seq with unique molecular identifiers / S. Islam [и др.] // Nature Methods. — 2013. — дек. — т. 11, № 2. — с. 163—166. — DOI: 10.1038/nmeth.2772. — URL: https://doi.org/10.1038/nmeth.2772.

106. Replicate sequencing libraries are important for quantification of allelic imbalance / A. Mendelevich [и др.] // Nature Communications. — 2021. — июнь. — т. 12, № 1. — DOI: 10.1038/s41467-021-23544-8. — URL: https://doi.org/10.1038/s41467-021-23544-8.

107. *Bix M.*, *Locksley R. M.* Independent and Epigenetic Regulation of the Interleukin-4 Alleles in CD4+ T Cells // Science. — 1998. — авг. — т. 281, № 5381. — с. 1352—1354. — DOI: 10.1126/science.281.5381.1352. — URL: https://doi.org/10.1126/science.281.5381.1352.

108. Locus specific epigenetic modalities of random allelic expression imbalance / L. Marion-Poll [и др.] // Nature Communications. — 2021. — сент. — т. 12, № 1. — DOI: 10.1038/s41467-021-25630-3. — URL: https://doi.org/10.1038/s41467-021-25630-3.

109. Evolutionary Persistence of DNA Methylation for Millions of Years after Ancient Loss of a De Novo Methyltransferase / S. Catania [и др.] // Cell. — 2020. — янв. — т. 180, № 2. — 263—277.e20. — DOI: 10.1016/j.cell.2019.12.012. — URL: https://doi.org/10.1016/j.cell.2019.12.012.

110. *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.

111. 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.

112. CTCF-Mediated Genome Architecture Regulates the Dosage of Mitotically Stable Mono-allelic Expression of Autosomal Genes / K. R. Chandradoss [и др.] // Cell Reports. — 2020. — окт. — т. 33, № 4. — с. 108302. — DOI: 10.1016/j.celrep.2020.108302. — URL: https://doi.org/10.1016/j.celrep.2020.108302.

113. Methylation-Sensitive Expression of a DNA Demethylase Gene Serves As an Epigenetic Rheostat / B. P. Williams [и др.] // PLOS Genetics / под ред. O. M. Scheid. — 2015. — март. — т. 11, № 3. — e1005142. — DOI: 10.1371/journal.pgen.1005142. — URL: https://doi.org/10.1371/journal.pgen.1005142.

114. *Christman J. K.* 5-Azacytidine and 5-aza-2′-deoxycytidine as inhibitors of DNA methylation: mechanistic studies and their implications for cancer therapy // Oncogene. — 2002. — авг. — т. 21, № 35. — с. 5483—5495. — DOI: 10.1038/sj.onc.1205699. — URL: https://doi.org/10.1038/sj.onc.1205699.

115. *Karahoca M., Momparler R. L.* Pharmacokinetic and pharmacodynamic analysis of 5-aza-2'-deoxycytidine (decitabine) in the design of its dose-schedule for cancer therapy // Clinical Epigenetics. — 2013. — февр. — т. 5, № 1. — DOI: 10.1186/1868-7083-5-3. — URL: https://doi.org/10.1186/1868-7083-5-3.

116. *Langmead B., Salzberg S. L.* Fast gapped-read alignment with Bowtie 2 // Nature Methods. — 2012. — март. — т. 9, № 4. — с. 357—359. — DOI: 10.1038/nmeth.1923. — URL: https://doi.org/10.1038/nmeth.1923.

117. Revisiting Global Gene Expression Analysis / J. Lovén [и др.] // Cell. — 2012. — окт. — т. 151, № 3. — с. 476—482. — DOI: 10.1016/j.cell.2012.10.012. — URL: https://doi.org/10.1016/j.cell.2012.10.012.

118. Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression / J. K. Kim [и др.] // Nature Communications. — 2015. — окт. — т. 6, № 1. — DOI: 10.1038/ncomms9687. — URL: https://doi.org/10.1038/ncomms9687.

119. *Prentice R. L.* Binary Regression Using an Extended Beta-Binomial Distribution, with Discussion of Correlation Induced by Covariate Measurement Errors // Journal of the American Statistical Association. — 1986. — июнь. — т. 81, № 394. — с. 321—327. — DOI: 10.1080/01621459.1986.10478275. — URL: https://doi.org/10.1080/01621459.1986.10478275.

120. *Brenner S.* The genetics of Caenorhabditis elegans // Genetics. — 1974. — май. — т. 77, № 1. — с. 71—94. — DOI: 10.1093/genetics/77.1.71. — URL: https://doi.org/10.1093/genetics/77.1.71.

121. Spliced synthetic genes as internal controls in RNA sequencing experiments / S. A. Hardwick [и др.] // Nature Methods. — 2016. — авг. — т. 13, № 9. — с. 792—798. — DOI: 10.1038/nmeth.3958. — URL: https://doi.org/10.1038/nmeth.3958.

122. DNA methylation is a key mechanism for maintaining monoallelic expression on autosomes / S. Gupta [и др.] // bioRxiv preprint. — 2020. — февр. — DOI: 10.1101/2020.02.20.954834. — URL: https://doi.org/10.1101/2020.02.20.954834.

123. Foreign RNA spike-ins enable accurate allele-specific expression analysis at scale / A. Mendelevich [и др.] // Bioinformatics. — 2023. — апр. — DOI: 10.1093/bioinformatics/btad254. — URL: https://doi.org/10.1093/bioinformatics/btad254.

124. *Reinhart A.* Statistics Done Wrong. — San Francisco, CA : No Starch Press, 01/2015. — URL: https://www.statisticsdonewrong.com/.

125. The human transcriptome across tissues and individuals / M. Melé [и др.] // Science. — 2015. — т. 348, № 6235. — с. 660—665. — DOI: 10.1126/science.aaa0355. — eprint: https://www.science.org/doi/pdf/10.1126/science.aaa0355. — URL: https://www.science.org/doi/abs/10.1126/science.aaa0355.

Список рисунков

Список таблиц

Приложение А

К главе 2, «Реплики библиотек секвенирования играют важную роль в количественной оценке аллельного дисбаланса»

А.1 Сопроводительные заметки к главе 2

А.1.1 Достаточно ли одной технической реплики для отделения сигнала от шума в аллельном дисбалансе?

При некоторых специальных условиях, одна техническая реплика может предоставить достаточное количество данных для отделения сигнала от шума в аллельном дисбалансе.

Один из таких сценариев возникает, если мы знаем точное распределение шума. Например, если шум возникает только в результате сэмплирования, тогда вклад шума в наблюдаемый сигнал можно было бы смоделировать универсально биномиальной моделью, зависящей только от покрытия и аллельной пропорции. Однако сообществом признано, что дисперсия шума аллельного дисбаланса превышает ожидаемую биномиальную дисперсию, и, таким образом, шум затруднительно задать биномиальной моделью [7; 60; 61; 72].

С другой стороны, если вероятностное распределение биологических данных принадлежит специфическому классу распределений, то мы могли бы разделить сигнал от шума в наблюдениях, основываясь лишь на одной технической реплике. Пример такого гипотетического сценария — распределение настоящего аллельного дисбаланса только в трёх точках: биаллельная экспрессия 1:1 и два крайних случая моноаллельной экспрессии 1:0 и 0:1. (детали см. в секции А.1.2).

Однако везде, кроме описанных экстремальных сценариев, мы не можем гарантированно отделить сигнал от шума в наблюдениях, основываясь на одной технической реплике, даже если нам известны классы распределений настоящего сигнала и шума. Далее мы подробнее рассмотрим несколько случаев.

Случай нормально распределённого сигнала и шума

Чтобы продемонстрировать неразличимость параметров сигнала и шума в некоторых случаях, возьмём переменную $\bar{x} \sim \mathcal{N}(\bar{\mu}, \bar{\sigma}^2)$, и примем нормально распределённую ошибку измерения $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$. Тогда

$$\bar{x} \sim \mathcal{N}(\bar{\mu}, \bar{\sigma}^2)$$

$$x \sim \mathcal{N}(\bar{x}, \sigma_\varepsilon^2) \Leftrightarrow x \sim \mathcal{N}(\bar{\mu},\ \bar{\sigma}^2 + \sigma_\varepsilon^2)$$

Возьмём образец данных в таком сценарии. Матожидание $\bar{\mu}$ можно оценить как среднее значение наблюдаемой величины в данных, где точность этой оценки будет зависеть от размера выборки. Но можем ли мы восстановить значение σ_ε^2 из дисперсии выборки или любой другой статистической оценки выборки?

Плотности вероятности нормальных распределений совпадают тогда и только тогда, когда оба параметра μ и σ совпадают. Поэтому единственная информация, которая может быть получена из выборки данных — это сумма $\bar{\sigma}^2 + \sigma_\varepsilon^2$. Размерность решений уравнения $\bar{\sigma}_1^2 + \sigma_{\varepsilon 1}^2 = \bar{\sigma}_2^2 + \sigma_{\varepsilon 2}^2$ равна единице, поэтому мы не можем однозначно определить σ_ε (Рисунок **A.1**).

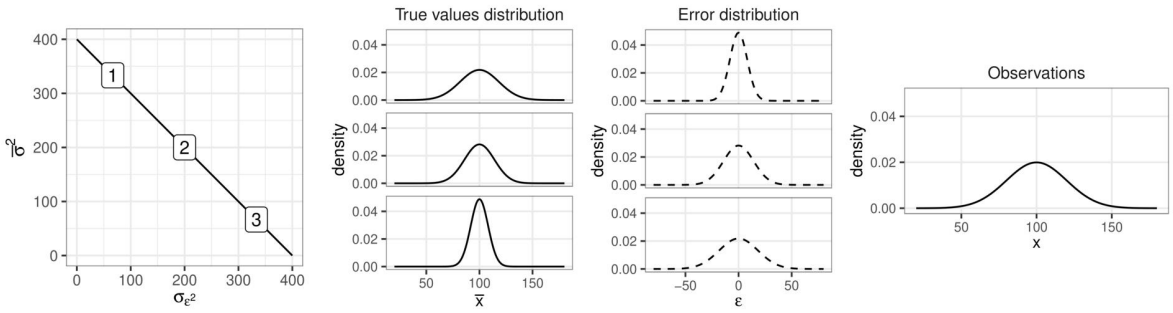

Рисунок A.1 — Различные параметризации сигнала и шума ведущие к одному распределению наблюдаемых.

Ситуация разительно меняется, если нам доступно второе измерение. Мы можем рассмотреть рассмотреть распределение величины $\Delta x = x_1 - x_2$:

$$x_1, x_2 \sim \mathcal{N}(\bar{x}, \sigma_\varepsilon^2)$$

$$x_1 - x_2 \sim \mathcal{N}(0, 2 \cdot \sigma_\varepsilon^2)$$

Параметр σ_ε становится однозначно вычислим.

Сигнал: тримодальное дельта- или бета-распределение, шум: бета-биномиальное или биномиальное распределение

Одна из часто используемых в аллель-специфическом анализе моделей основана на тримодальном предположении о настоящем значении аллельного дисбаланса. В этом предположении, или экспрессия биаллельна, или полностью моноаллельна. Вариации в данных рассматриваются как следствие бета-биномиального шума [72; 73].

Например, мы можем попробовать подобрать наблюдаемую выборку при помощи смеси двух симметричных бета-биномиальных распределений (Рисунок **A.2**), предполагая, что биаллельная экспрессия даёт выпуклое бета-биномиальное распределение ($\alpha_1 = \beta_1 > 1$), а моноаллельная – вогнутое бета-биномиальное распределение ($1 > \alpha_2 = \beta_2 > 0$).

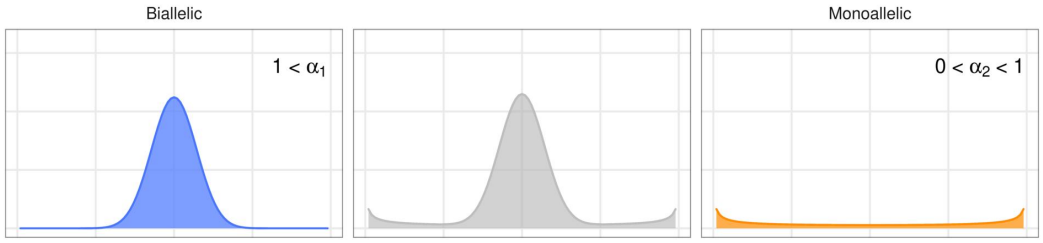

Рисунок A.2 — Схематичное представление распределения генов согласно подбору смеси выпуклого и вогнутого бета-биномиального распределения.

Тогда эта ситуация по построению неотличима от случая, когда мы, наоборот, рассматриваем взятие выборки без избыточной дисперсии (то есть биномиальный шум), но настоящие значения аллельного дисбаланса приходят из смеси двух бета-распределений с соответствующими α_1 и α_2 (Рисунок **A.3**).

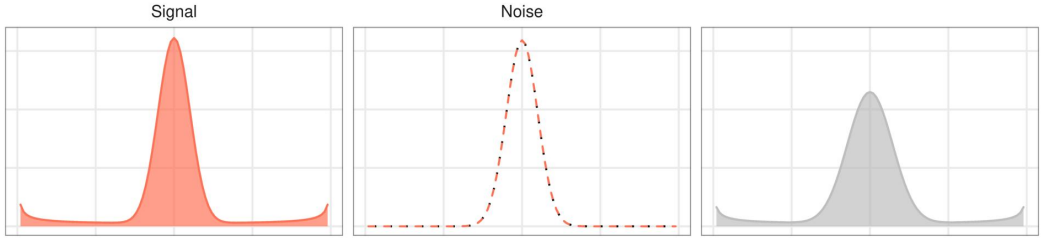

Рисунок A.3 — Вторая параметризация, дающая то же распределение наблюдаемой величины: сигнал с бета-распределением и биномиальный шум.

Заметим, что если параметры α_1 и α_2 зафиксированы, то распределения, построенные выше, сойдутся в разные финальные распределения при $n \to \infty$, а именно к тримодальному распределению Дирака и бета-распределению, и поэтому потенциально две ситуации можно отличить на генах с разным покрытием.

Однако, мы наблюдаем, что избыточная дисперсия остаётся постоянной при любом покрытии (см. рисунки **2.3 и A.32**). Чтобы отразить это, мы можем рассмотреть параметры, задающие форму распределения, как функции n, чтобы сделать пределы похожими на предел первой модели, то есть на тримодальное распределение Дирака (Рисунок **A.4**). Например:

$$\alpha_1(n) = \alpha_{01} \cdot \ln(n)$$

$$\alpha_2(n) = \frac{\alpha_{02}}{\ln(n)}$$

для некоторых констант α_{01} и α_{02}.

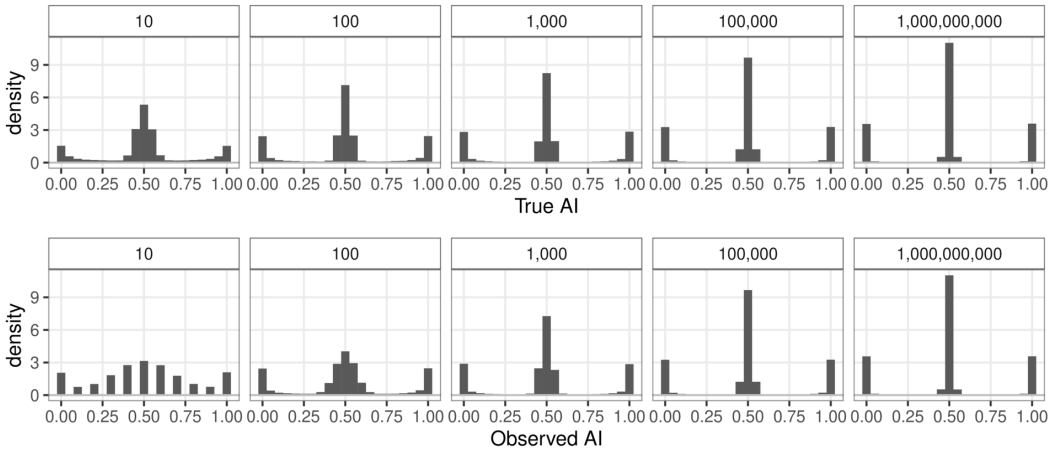

Рисунок A.4 — Пример с бета-смесью и соответствующей бета-биномиальной смесью, которые сходятся к одному тримодальному распределению при увеличении покрытия.

Значение параметров на рисунке: $w_1 = 0.6$, $\alpha_1 = 20$, $\alpha_2 = 19/20$, коррекция: $\log_2(\text{cov})$.

Когда α_i зависит от n, эти два бета-биномиальных распределения становятся неотличимы.

Случай бета-биномиального сигнала и шума

Пусть и настоящие значения аллельного дисбаланса, и шум распределены согласно бета-биномиальному распределению. Предположение о бета-биномиальном шуме часто встречается в литературе [7; 60; 61; 72], и этот класс распределений даёт разумные формы, которые хорошо годятся для моделирования значений аллельного дисбаланса.

Тогда для какого-то покрытия C и пары параметров избыточной дисперсии ρ_1 и ρ_2:

$$\mathrm{AI}_{\mathrm{obs}} \sim \text{Beta-Bin}(C, p = \mathrm{AI}_{\mathrm{true}}, \rho_2)$$

где

$$\mathrm{AI}_{\mathrm{true}} \sim \text{Beta-Bin}(C, p = 0.5, \rho_1)$$

Соотношение между сигналом и шумом не является симметричным, и “дополняющие” распределения не равны:

$$\text{Beta-Bin}(C, \text{Beta-Bin}(C, 0.5, \rho_1), \rho_2) \neq \text{Beta-Bin}(C, \text{Beta-Bin}(C, 0.5, \rho_2), \rho_1)$$

для $\rho_1 \neq \rho_2$ (Рисунок **A.5**). Мы будем использовать эти пары дополняющих распределений для иллюстрации идеи того, что абсолютно разные параметризации (которые не могут быть рассмотрены как равномерно близкие) могут в итоге дать похожие наблюдения.

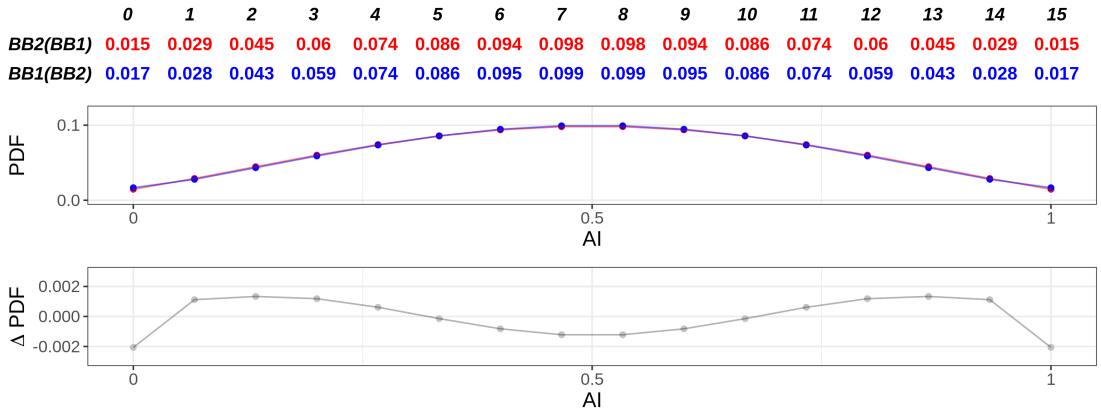

Рисунок A.5 — Пример неравенства двух дополняющих распределений.

Несмотря на то, что функции плотности распределения выглядят похоже, (*верхний график*), они не равны (*таблица* и *нижний график*). Построенные распределения: $\text{Beta-Bin}(C, \text{Beta-Bin}(C, 0.5, \rho_1), \rho_2)$ (обозначены *BB2(BB1)*) и $\text{Beta-Bin}(C, \text{Beta-Bin}(C, 0.5, \rho_2), \rho_1)$ (обозначены *BB1(BB2)*): $n = 15$, $\rho_1 = 0.01$, $\rho_2 = 0.1$.

Важно отметить, что в целом образцы из дополняющих распределений неотличимы (см. функции плотности вероятности и квантиль-квантиль графики для дополняющих распределений на рисунке **A.6**). Распределения p-уровней значимости по тесту Манна-Уитни-Уилкоксона неотличимы от равномерных распределений (Рисунок **A.7a**), когда распределения p-уровней значимости по тесту Колмогорова-Смирнова похожи для распределений с дополняющей параметризацией (Рисунок **A.7b**).

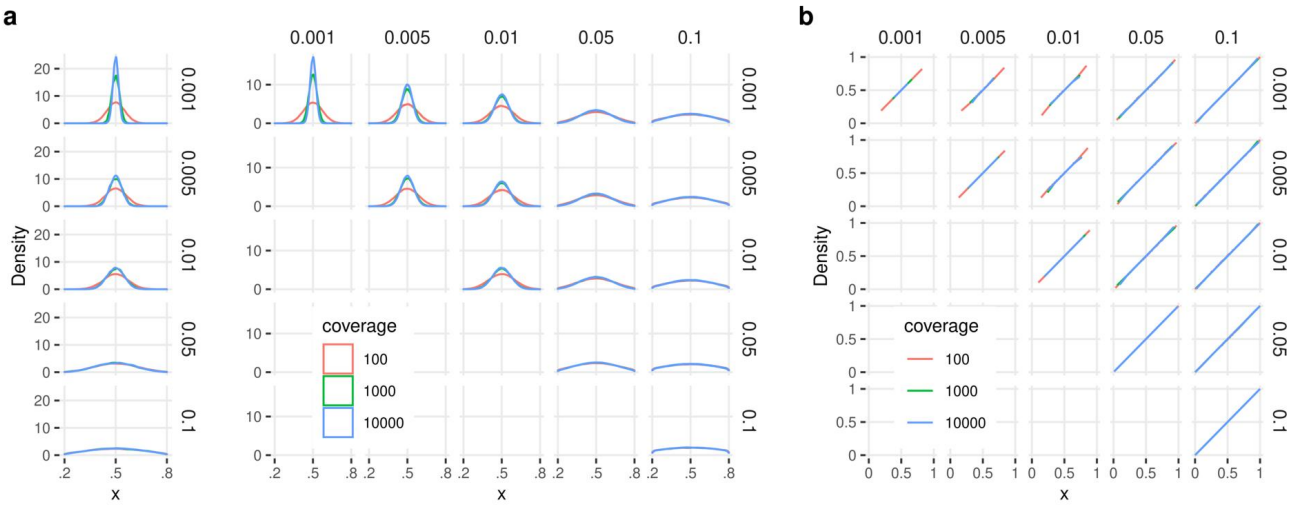

Рисунок A.6 — Пример дополняющих распределений ($\rho_{11} = \rho_{22}$ и $\rho_{21} = \rho_{12}$) для разных пар ρ_1, ρ_2 и уровней покрытия.

(a) Графики плотности вероятности, **(b)** квантиль-квантиль графики между дополняющими распределениями.

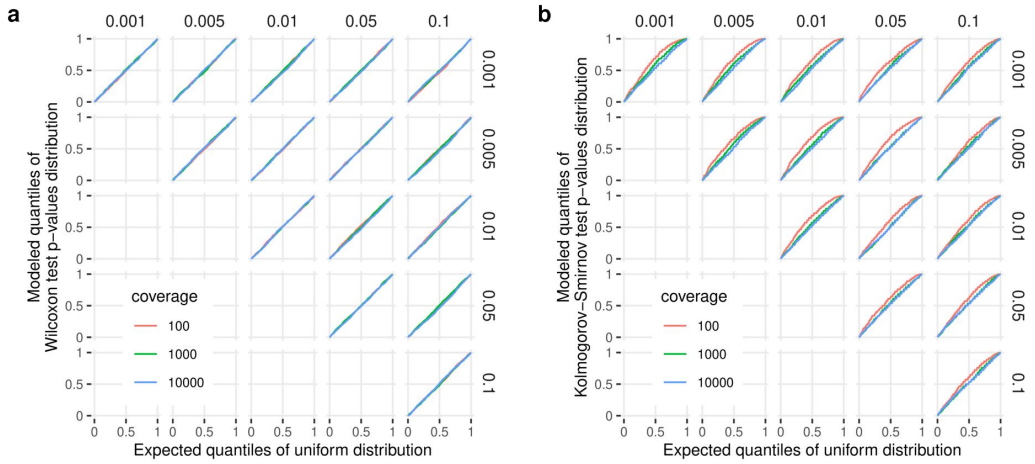

Рисунок A.7 — Квантиль-квантиль график для равномерного распределения и распределений p-уровня значимости на дополняющих распределениях ($\rho_{11} = \rho_{22}$ и $\rho_{21} = \rho_{12}$).

(a) p-уровни значимости по двустороннему тесту Манна-Уитни-Уилкоксона **(b)** p-уровни значимости по тесту Колмогорова-Смирнова

Критика использования фиксированного распределения для моделирования аллельного дисбаланса

Все подходы к анализу аллельного дисбаланса в данных РНК-секвенирования начинаются с предположения о том, что некоторое фиксированное распределение из семейства является хорошим общим приближением к данным из всего эксперимента. Однако простой мысленный эксперимент показывает, что распределение аллельного дисбаланса может перестать принадлежать конкретному семейству; более общо, предположение о том, что какое-либо конкретное распределение предоставляет хорошую модель для всех генов не является состоятельным.

Для того, чтобы проиллюстрировать первый аргумент, рассмотрим возмущение (например, введение лекарства), которая изменяет аллельный дисбаланс в большой доле генов. После возмущения гены будут поделены на те, которые принадлежат "старому" распределению, и гены, которые будут отвечать на воздействие и, таким образом, принадлежать новому распределению. Для примера такого воздействия допустим, что половина генов отвечает уменьшением дисбаланса, их AI становится наполовину ближе к 1:1. Если начальное распределение в этом примере было бета-распределением, новая смесь распределений не является бета-распределением.

Это рассуждение показывает, что никакое конкретное распределение не может служить универсальной моделью для всех экспериментов. Поэтому вместо рассмотрения обобщённого распределения аллельного дисбаланса по всем покрытиям, мы рассматриваем каждый случай отдельно, с целью отражения локального состояния данных. Нам всё ещё необходимо априорное распределение для того, чтобы моделировать реплики без избыточной дисперсии, поэтому мы использовали бета-распределение в каждом случае для локальной оценки поведения аллельного дисбаланса.

A.1.2 Учёт избыточной дисперсии ведёт к ожидаемому бимодальному распределению значений аллельного дисбаланса в рассогласованных результатах.

С какой вероятностью биномиальный тест (Рисунок **A.8a**) покажет разные результаты на двух наблюдаемых материнских покрытиях M_1 и M_2 для данной пропорции a (Рисунок **A.8b**)?

Рассмотрим эту вероятность как функцию конкретного покрытия гена N и соответствующих границ C_1 и C_2 биномиального теста BT_{QCC} с $H_0 : p = 0.5$, на прочтениях, подправленных на QCC. (Рисунок **A.8c**). Тогда вероятность рассогласованных результатов теста BT_{QCC} Верно/Ложно на двух технических репликах для настоящей пропорции $a \in (0,1)$ равна:

$$f_{N,QCC}(a) = P(\mathrm{BT}_{QCC}(M_1) \neq \mathrm{BT}_{QCC}(M_2)) =$$

$$= 2 \cdot (P(M \leq C_1|a) \cdot P(M \in (C_1, C_2)|a) + P(M \geq C_2|a) \cdot P(M \in (C_1, C_2)|a)) =$$

$$= 2 \cdot P(M \in (C_1, C_2)|a) \cdot (P(M \leq C_1|a) + P(M \geq C_2|a)) =$$

$$= 2 \cdot \int_{C_1}^{C_2} \mathrm{Bin}_{QCC}(x; N, a)dx \cdot \left(\int_0^{C_1} \mathrm{Bin}_{QCC}(x; N, a)dx + \int_{C_2}^{N} \mathrm{Bin}_{QCC}(x; N, a)dx \right)$$

Взяв дискретизированные распределения $U_N(a)$ настоящих значений аллельного дисбаланса (Рисунок **A.8d**), мы можем получить распределение аллельных дисбалансов генов с рассогласованными результатами теста BT_{QCC} на 2 технических репликах, как произведение $U_N(a)$ и $f_{N,QCC}(a)$ (Рисунок **A.8e**).

Заметим, что если тест использует распределение, которое хорошо описывает данные, то значения аллельного дисбаланса с рассогласованными результатами будут распределены вокруг границ теста (Рисунок **A.10**). Напротив, отсутствие учёта избыточной дисперсии влечёт намного более широкие, иногда даже унимодальные распределения, какие мы и наблюдаем в наших данных (Рисунок **A.9**, также **Fig.2.2b,f**).

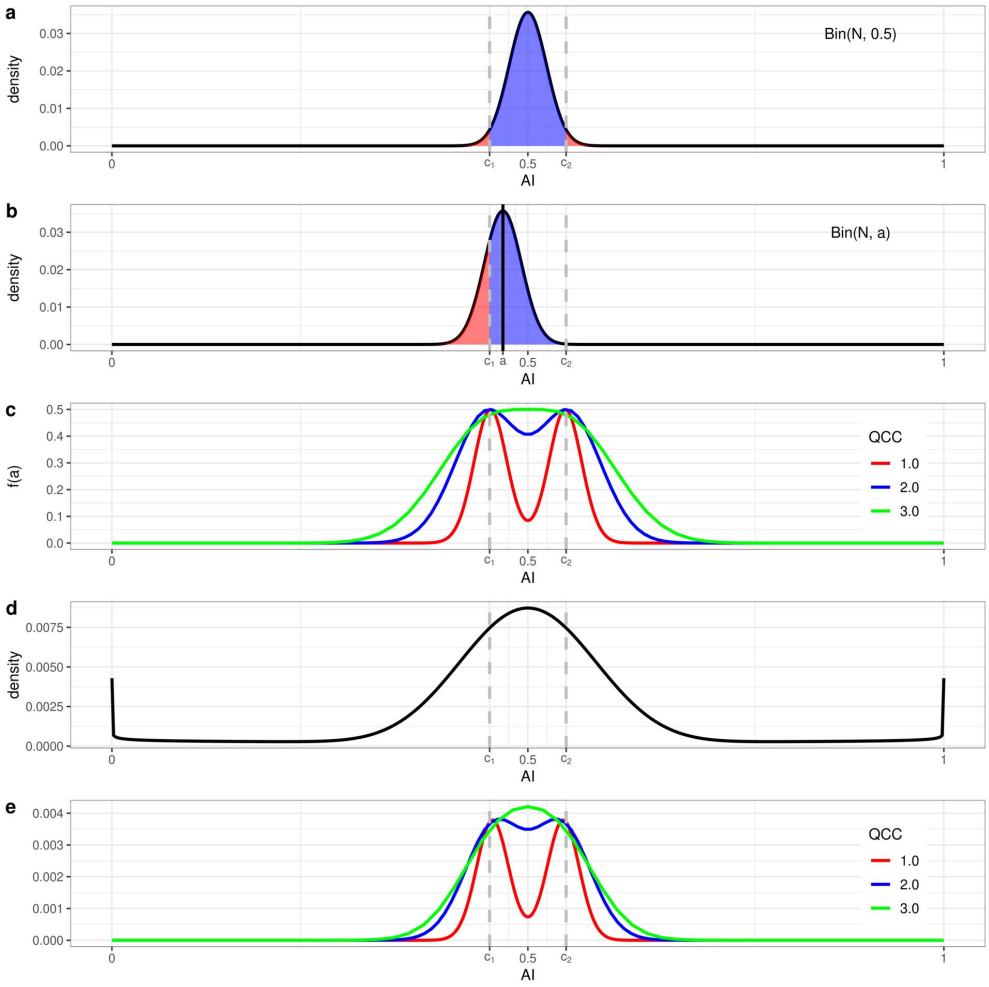

Рисунок А.8 — Биномиальный тест на двух наблюдениях

(a) Биномиальное распределение и границы биномиального теста с пропорцией 0.5, для $N = 500$ и уровня доверия 0.95; **(b)** Биномиальное распределение наблюдаемого аллельного дисбаланса для настоящего дисбаланса $a = 0.47$, покрашено согласно результатам биномиального теста; **(c)** Вероятность получить рассогласованные результаты теста для двух наблюдений при разных QCC (напомним, что $QCC = 1$ соответствует биномиальному распределению и поэтому представляет случай, когда тест подходит данным); **(d)** Распределение настоящего аллельного дисбаланса; **(e)** Распределение настоящего аллельного дисбаланса в генах которые могут быть отмечены другой категорией при помощи биномиального теста для 2 технических реплик. При $QCC = 1$ распределение бимодально, и моды расположены около границ биномиального теста.

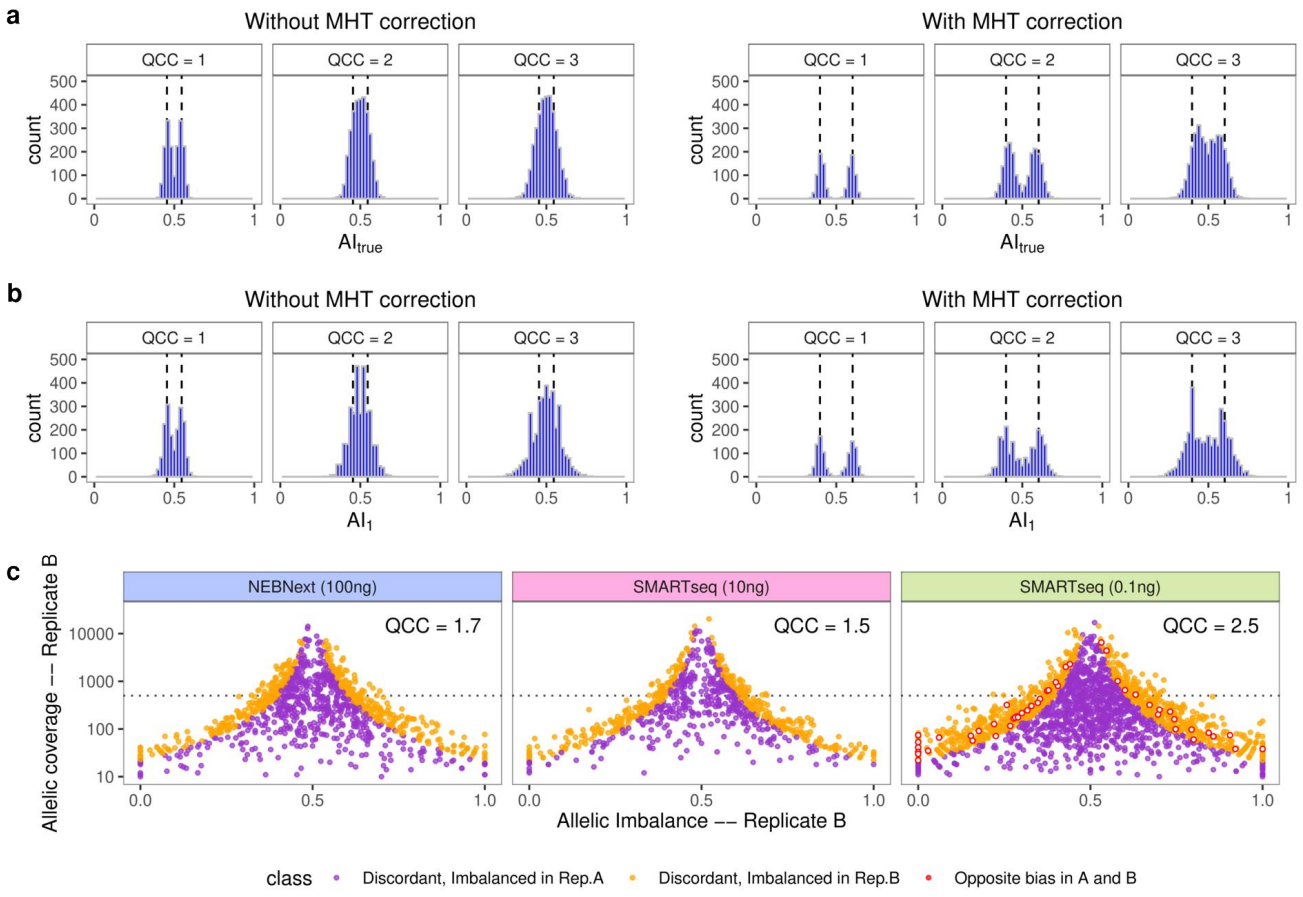

Рисунок A.9 — Тест нулевой гипотезы аллельного дисбаланса без коррекции на избыточную дисперсию.

(a)(b) Симулированные данные. Распределения **(a)** настоящих значений AI_{true} аллельного дисбаланса **(b)** наблюдаемых значений аллельного дисбаланса из одной реплики (AI_1) для генов с рассогласованными результатами дисбаланса в двух симулированных репликах. Симулированные данные: 10000 генов с покрытием 500 с настоящими значениями аллельного дисбаланса распределёнными как 0.85:0.15 смесь бета-распределений с $\alpha_1 = \beta_1 = 20$ и $\alpha_1 = \beta_1 = 0.8$; 95% уровень доверия; наличие поправки Бонферрони обозначено в подписях. Заметим, что разница в формах распределений аллельного дисбаланса отражает разницу в уровне избыточной дисперсии данных, как описано выше (Рисунок **A.8e**). **(c)** График рассогласования AI для экспериментальных данных (те же данные, что и в Рисунке **2.2b**). Панели **a** и **b** (с коррекцией) представляют графики плотности вдоль пунктирной линии на покрытии 500.

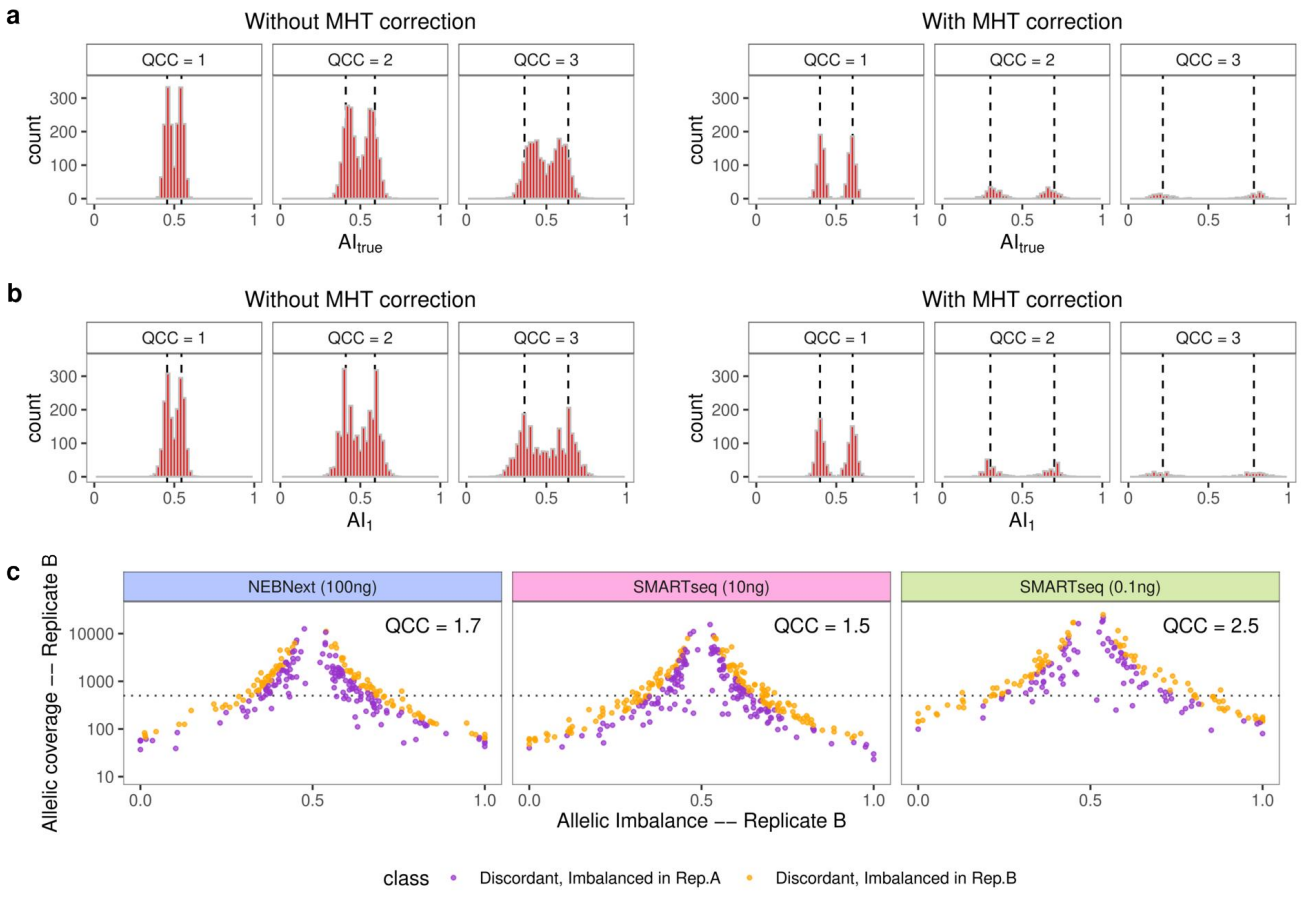

Рисунок A.10 — Тот же анализ, что и на Рисунке **A.9**, но с учётом избыточной дисперсии аллельного дисбаланса.

Заметим, что оба распределения AI_{true} **(a)** и AI_1 **(b)** намного более явно являются распределениями вокруг границ теста. Тот же эффект наблюдается на данных из реплик **(c)** когда мы применяем QCC коррекцию (также см. **2.2**).

A.1.3 Гены с различными аллельными дисбалансами имеют разное влияние на общую дисперсию сигнала.

Рассмотрим гены в конкретном интервале покрытий. Если значения аллельного дисбаланса для двух данных реплик, $x_1 = \{x_{1i}\}$ и $x_2 = \{x_{2i}\}$, принадлежат одному распределению, похожему на биномиальное, и соответствующие аллельные пропорции $a = \{a_i\}$ принадлежат симметричному распределению, тогда имея ввиду, что $var(x_{1i} - x_{2i}) = var(x_{1i}) + var(x_{2i})$, мы предполагаем, что $var(x_{1i} - x_{2i}) \sim a_i(1 - a_i)$ для любого гена i.

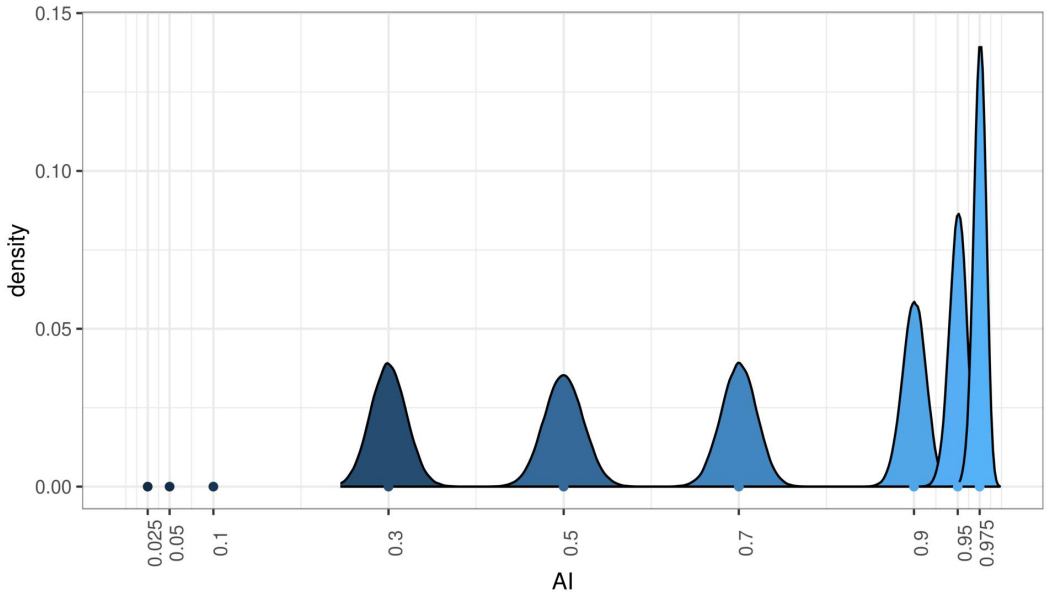

Рисунок A.11 — Вариация наблюдений аллельного дисбаланса имеет различия по интервалу подлежащего AI (покрытие гена 500).

Выражение $a_i(1 - a_i)$ достигает максимума при $a_i = 0.5$ и минимума при 0 и 1 (Рисунок **A.11**).

Для $X = \{x_{1i} - x_{2i}\}_i$ мы имеем $var(X) = E((X - \mu)^2) = E((X - 0)^2) = \frac{1}{n}\sum_{i=1}^{n}(x_{1i} - x_{2i})^2$, что зависит от распределения подлежащих аллельных пропорций $a = \{a_i\}$.

Так как истинные AI принадлежат некоторому сложному распределению, и гены с разными истинными AI имеют разное влияние на общую вариацию сигнала, распределение AI должно быть учтено при попытке сделать какие-либо выводы о распределении разностей между наблюдениями AI в технических репликах.

Поэтому мы делаем вывод, что необходимо учитывать распределение истинного аллельного дисбаланса при оценке разности в AI между техническими репликами. Предположение о едином простом распределении (например, тримодальном) здесь не является состоятельным. Мы учитываем наблюдаемое распределение в бета-биномиальной модели (см. Рисунок **2.3d**).

A.1.4 Мы ожидаем около нуля генов с ложноположительным отличием AI, оценённого из двух реплик, от AI, оценённого из шести.

Здесь мы определяем ложноположительность как событие, когда точечные оценки аллельного дисбаланса для гена из шести реплик находятся вне доверительного интервала для того же гена, основанного на двух из шести реплик (после поправки Бонферрони).

Для конкретного гена и соответствующих подлежащих AI p мы ожидаем, что материнские прочтения в n репликах будут распределены согласно тому же распределению $m_{i\in\{1..n\}} \sim \mathrm{Bin}(C, p)$ (здесь мы рассматриваем случай $\mathrm{QCC} = 1$, но он легко может быть обобщён). Тогда для $\mathrm{CI}(\mathrm{AI}_2 = \frac{m_1+m_2}{2\cdot C})$ посчитанного на первых двух репликах, вероятность того, что подлежащий AI p попадёт вне $\mathrm{CI}(\mathrm{AI}_2)$ (который по построению определён уровнем важности, например, 0.05) меньше, чем вероятность того, что оценка AI на k репликах кроме первых двух реплик $\frac{\sum_3^{k+2} m_i}{k\cdot C} \sim \frac{\mathrm{Bin}(k\cdot C,p)}{k\cdot C}$ попадёт вне интервала, но больше, чем вероятность того, что оценка AI, посчитанная на всех k репликах $\frac{\sum_1^{k} m_i}{k\cdot C} \sim \frac{\mathrm{Bin}(k\cdot C,p)}{k\cdot C}$) попадёт вне интервала. Заметим, что последнее не независимо от $\frac{m_1+m_2}{2\cdot C}$.

В нашем случае:

$$P\left(\frac{\sum_1^6 m_i}{6\cdot C} \notin \mathrm{CI}\left(\frac{m_1+m_2}{2\cdot C}\right)\right) \leq P\left(p \notin \mathrm{CI}\left(\frac{m_1+m_2}{2\cdot C}\right)\right) \leq$$

$$\leq P\left(\frac{\sum_3^8 m_i}{6\cdot C} \notin \mathrm{CI}\left(\frac{m_1+m_2}{2\cdot C}\right)\right)$$

Симуляции, подтверждающие эти аргументы, представлены в рисунках **A.12** и **A.13**.

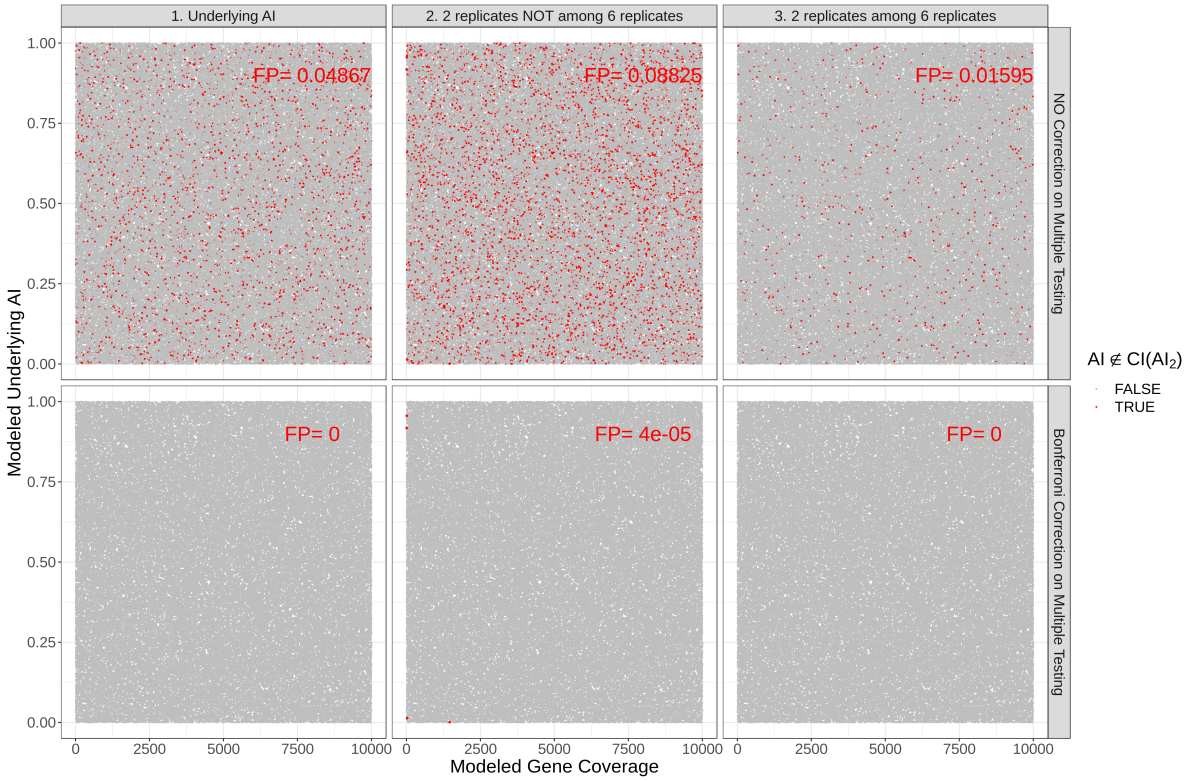

Рисунок A.12 — Количество ложноположительных результатов для смоделированных 100000 генов с разным покрытием и подлежащим AI.

Точка тестировалась на принадлежность $CI(AI_2)$ вычисленному на двух случайно выбранных репликах. Описание точки слева направо: подлежащий AI, оценка AI из 4 реплик кроме двух начальных, оценка AI из всех 6 реплик. Доля ложноположительных результатов приведена до и после коррекции на множественное тестирование.

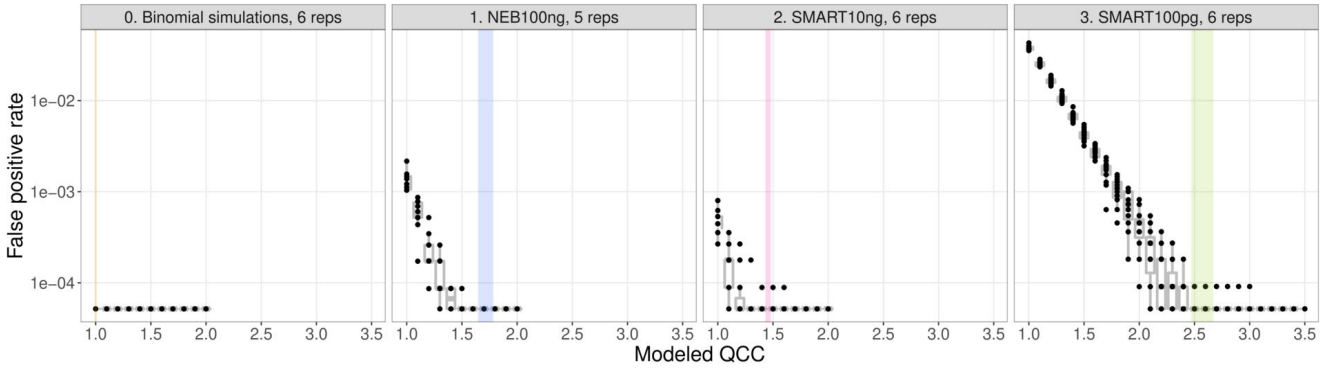

Рисунок A.13 — Доля ложноположительных результатов для биномиальной симуляции ($QCC = 1$) и различных экспериментов при различных QCC.

Интервал настоящего QCC покрашен.

A.1.5 Статистическая сила теста, поправленного на QCC.

Мы показали, что коррекция на QCC сильно уменьшает долю ложноположительных (FP) результатов. Естественно спросить себя: как такая коррекция сказывается на балансе между FP ошибками и потерей сигнала?

Заметим, что дать определение действительно положительным результатам на практике труднее, чем ложноположительным: можно мыслить себе эту трудность в терминах точек в рисунке. Если рисунок доступен нам в неограниченно высоком разрешении, то любые две различные точки в рисунке можно отличить. Однако любое конкретное разрешение ставит ограничение на нашу способность отделять точки друг от друга. В этой аналогии разрешение рисунка можно сравнить со сложностью библиотеки, а величину покрытия – с увеличением, при помощи которого мы смотрим на рисунок. Другими словами, то, можем ли мы отличить два значения AI друг от друга — это свойство данных, и анализ статистической силы теста даёт нам оценку на покрытие, необходимое для вычисления интересующих нас свойств в данных.

1. Как меру величины сигнала, мы будем использовать статистическую силу, частый способ оценить нужное количество образцов для определения действительной разницы на интересующей нас шкале с заданным уровнем доверительности [124]. Рассмотрим задачу разделения аллельного дисбаланса в двух разных образцах: если мы рассматриваем ген такой, что ΔAI между образцами 1 и 2 равно некоторому конкретному значению, и через AI_0 мы обозначаем среднее этих значений, какова вероятность того, что мы не обнаружим значимой разницы в AI, если нам даны покрытие гена, техническая избыточная дисперсия (одинаковая для обоих образцов), с требуемым уровнем значимости (Рисунок **A.14**)?

Заметим, что сравнение AI одного образца с заданным значением являет собой формально похожую задачу, но формулировка с двумя образцами имеет намного биологически более релевантную интерпретацию.

Статистическая сила зависит от аллельного покрытия и 4 дополнительных параметров, задающих конкретную альтернативную гипотезу H_1:

– основное AI (AI_0) и разница в AI (ΔAI), которые определяют AI_1 и AI_2 как $AI_0 \pm \frac{\Delta AI}{2}$, используется в симуляции данных;
– параметр избыточной дисперсии $\mathcal{O}$ данных, используется в симуляции данных;
– QCC используется в тесте (Биномиальный $\Leftrightarrow$ QCC $= 1$; тест, откорректированный на QCC $\Leftrightarrow$ QCC $= \sqrt{\mathcal{O}}$).

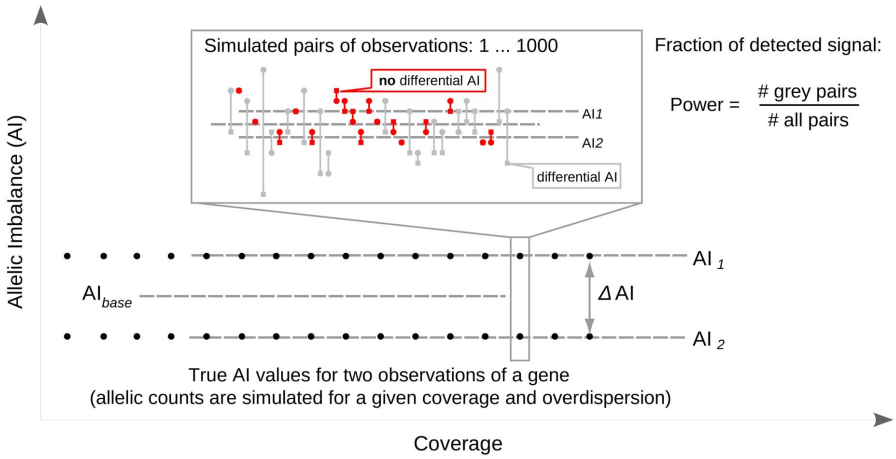

Рисунок A.14 — Схематическое изображение статистической силы (доля обнаруженного сигнала для фиксированного уровня разрешения).

Чтобы оценить значения силы $\mathcal{P}(\text{cov}_i \mid \mathcal{O}, \text{QCC}, \text{AI}_0, \Delta\text{AI})$ в этом анализе, мы сформируем 1000 пар AI $\{\text{AI}_{ijk}\}_{k \in 1...1000}$ из распределения $\text{Bin}(\frac{\text{cov}_i}{\mathcal{O}}, \text{AI}_j)$ (для $j \in \{1,2\}$) для разных уровней покрытия cov_i, и возьмём процент событий, когда дифференциальный тест возвращает положительный результат для AI_{i1k} и AI_{i2k} (дифференциальный на выбранном уровне значимости, здесь: 0.95).

Снова заметим, что статистическая сила описывает в полной мере характер действительно позитивных решений теста, так как она зависит от фиксированного AI_0 и ΔAI, и поэтому может быть использована только для оценки необходимого покрытия при данном уровне точности.

2. Мера количества FP (ложноположительных) событий определена как обычно, как дополнительная к статистической силе: как много идентичных точек классифицированы как имеющие значительную разницу в AI. Чтобы проиллюстрировать эффект недооценки избыточной дисперсии на обнаружении сигнала, мы сравним наш метод с биномиальным тестом. Рисунок **A.15** показывает, что доля FP остаётся примерно одинаковой вне зависимости от покрытия, и

избыточная дисперсия ведёт к увеличению доли FP при использовании биномиального теста (то есть недостаточная коррекция при предположении отсутствия технической избыточной дисперсии). Откорректированный по QCC тест контролирует дисперсию сверх биномиальной.

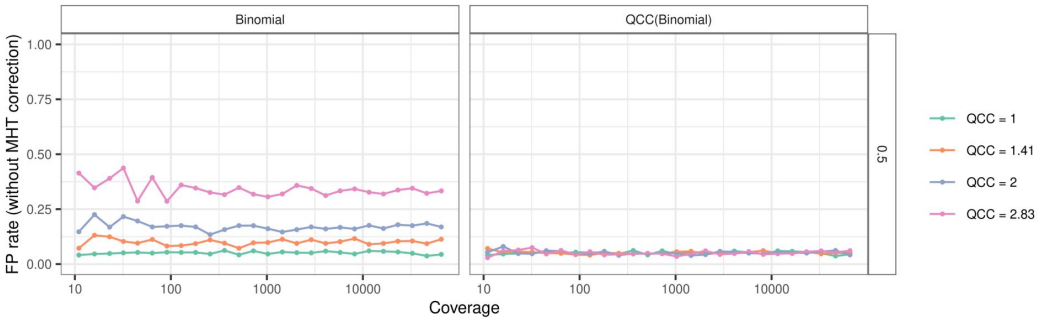

Рисунок А.15 — Доля ложноположительных результатов с простым биномиальным тестом и с откорректированным по QCC.

Показано для данных, сформированных с разной избыточной дисперсией (показано цветом), разным покрытием и фиксированным аллельным дисбалансом AI = 0.5.

3. Построив эти измерения, мы можем оценить их зависимость от покрытия и избыточной дисперсии (Рисунок **А.16**).

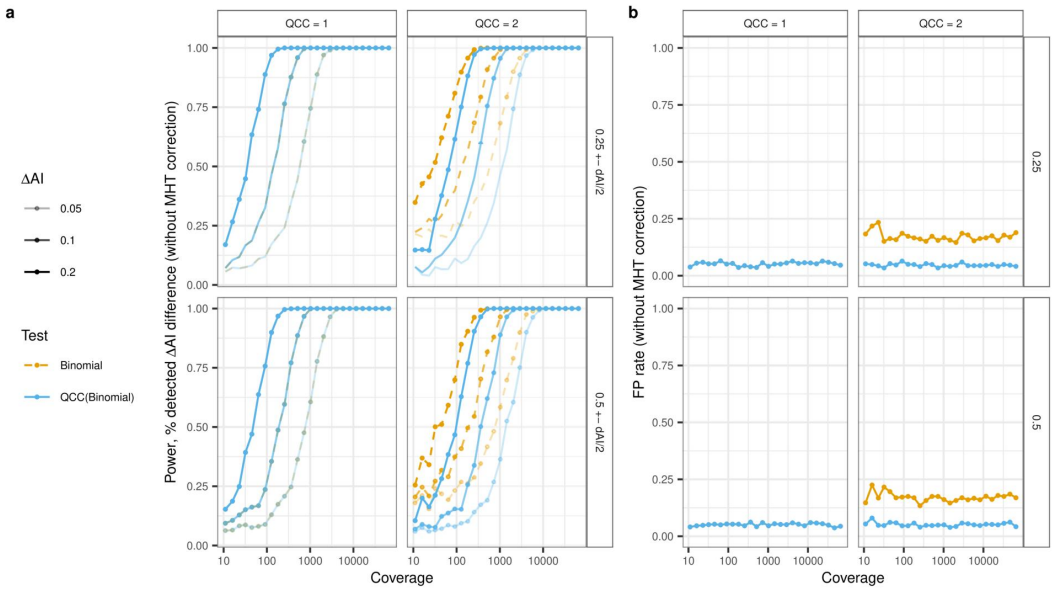

Рисунок А.16 — Статистическая сила и доля ложноположительных решений дифференциальных тестов с биномиальным и QCC-откорректированным предположениями.

(a) Статистическая сила **(b)** доля FP. Показано для данных, сформированных с разной избыточной дисперсией (колонки, QCC = 1 и QCC = 2), с разным уровнем покрытия и двумя аллельными пропорциями (строки, AI = 0.25 и AI = 0.5). Когда QCC = 1, результаты дифференциальных тестов совпадают.

(a) Прозрачность отражает разницу ΔAI между настоящими значениями AI в двух сформированных образцах.

В отсутствии технической избыточной дисперсии, биномиальный тест и QCC-откорректированный тест совпадают. Когда в данных присутствует избыточная дисперсия, и при низком покрытии, тест, который берёт недооцененную избыточную дисперсию, показывает большую силу, нежели QCC-откорректированный тест (Рисунок **A.16.a**), что неизбежно следует вместе с повышенной долей FP (Рисунок **A.16.b**). Заметим, что сила дифференциального теста с биномиальным предположением не приближается к нулю при низком покрытии (Рисунок **A.16.a**). Разница в силах биномиального и QCC-откорректированного тестов улетучивается при увеличении в покрытии, и для обоих тестов сила неотличима от 1 начиная с некоторого уровня покрытия. Очевидно, что большие ΔAI различаются лучше на любом конкретном уровне покрытия, а достаточно маленькие ΔAI не могут быть обнаружены при малом покрытии (Рисунок **A.17**).

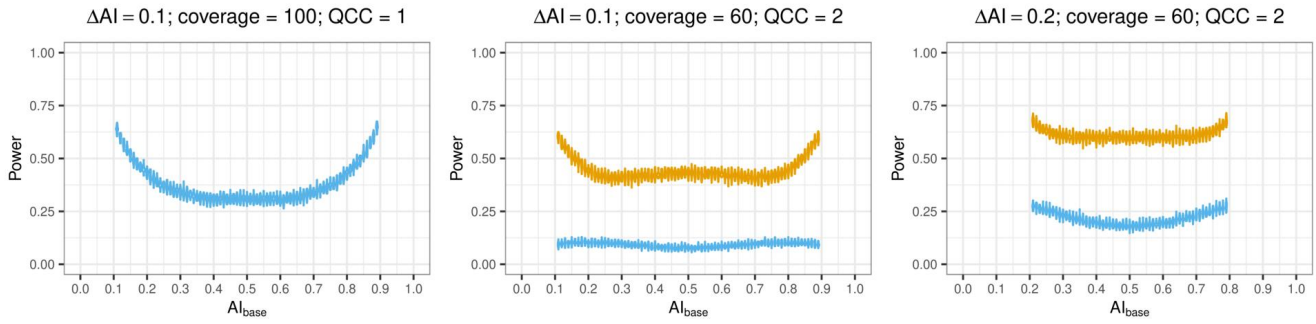

Рисунок A.17 — Статистическая сила дифференциальных тестов с биномиальными (оранжевым) и QCC-откорректированными (синим) предположениями.

Показано для данных, сформированных на основе нескольких множеств параметров: коэффициент избыточной дисперсии (QCC), покрытие и разница в настоящих значениях AI (ΔAI). Каждая диаграмма-“скрипка” соответствует сравнению между образцами с настоящими значениями AI $\mathrm{AI}_{\mathrm{base}} - \frac{\Delta \mathrm{AI}}{2}$ и $\mathrm{AI}_{\mathrm{base}} + \frac{\Delta \mathrm{AI}}{2}$. Когда $\mathrm{QCC} = 1$, результаты дифференциальных тестов совпадают.

4. Важное практическое заключение — QCC-откорректированный анализ резко сокращает долю FP, и при этом практически не теряет сигнала начиная с некоторого порога покрытия. “Стоимость”, которую мы платим статистической силой, уменьшается при увеличении покрытия, и в некоторый момент пропада-

ет (Рисунок A.18). При необходимости, покрытие может быть увеличено путём дополнительного секвенирования. Напротив, доля FP у биномиального теста всегда будет ненулевой, вне зависимости от покрытия.

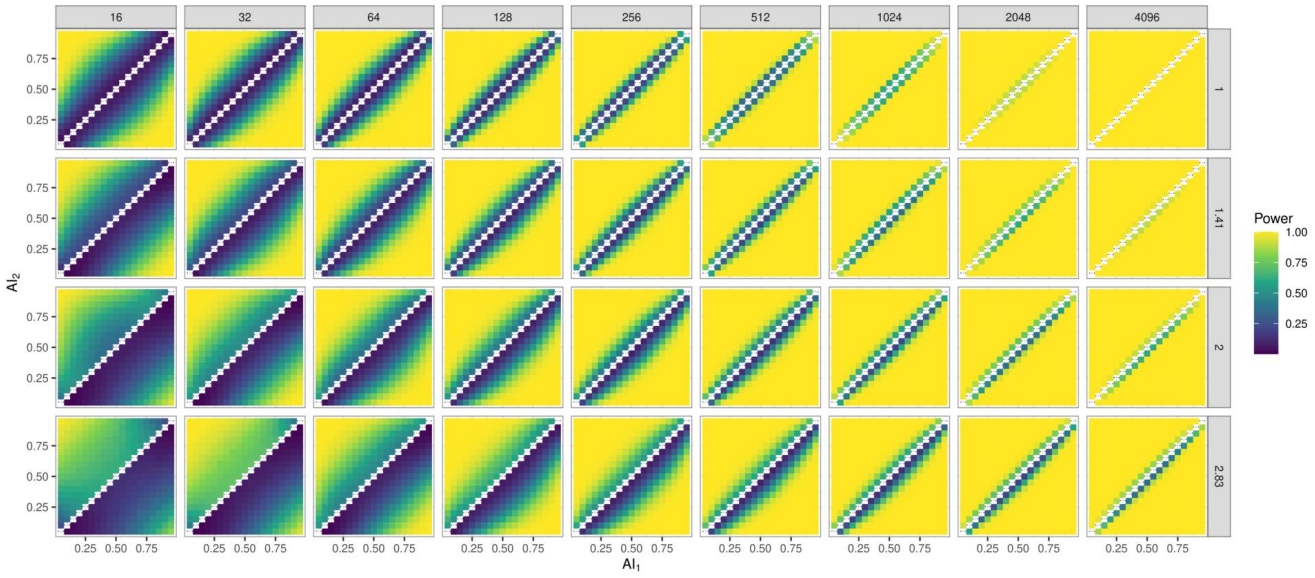

Рисунок A.18 — Тепловая карта статистической силы дифференциальных тестов с биномиальными (верхний левый треугольник) и QCC-откорректированными (нижний правый треугольник) предположениями.

Показано для данных, сформированных с несколькими множествами параметров: коэффициент избыточной дисперсии (строки, QCC), уровень покрытия (колонки) и настоящие значения AI (оси, AI_1 и AI_2) использованы в генерации образцов для сравнения.

A.1.6 Инструкция по вычислению QCC, начиная с fastq.

fastq → покрытия:

https://github.com/gimelbrantlab/ASEReadCounter_star/wiki/2.-Allelic-Counts-Table-Creation

покрытия → QCC:

https://github.com/gimelbrantlab/Qllelic/wiki/Use-case-1:-One-biological-sample

A.1.7 Инструкция по проведению дифференциального анализа AI для двух образцов.

https://github.com/gimelbrantlab/Qllelic/wiki/Use-case-2:-Differential-AI-analysis

A.2 Сопроводительные рисунки к главе 2

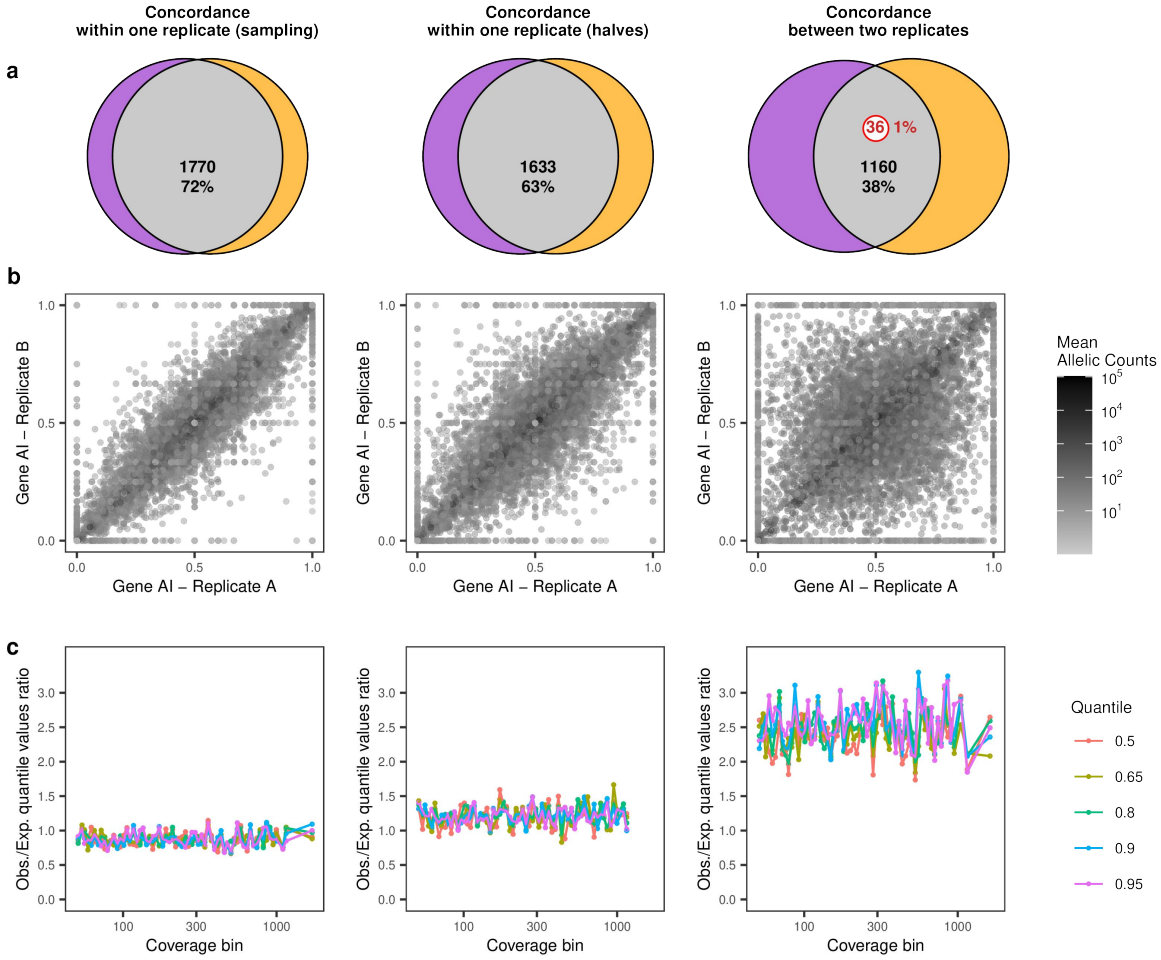

Рисунок A.19 — Разные способы выборки раскрывают разные доли технической избыточной дисперсии AI.

Слева направо: разные процедуры выборки приводят к множествам A и B в 25 миллионов прочтений, каждое взято из реплик Эксперимента 3. **Слева:** Множества A и B независимо подвыбраны из 52 миллионов прочтений в одной и той же реплике (A и B находятся в биномиальном взаимоотношении друг с другом); **Посередине:** Множества A и B получены разбиением 50 миллионов прочтений из одной библиотеки РНК-секвенирования вполовину. **Справа:** Множества A и B выбраны из двух технических реплик. **a:** Диаграммы Эйлера согласованности “аллельно дисбалансных” генов (H_0: AI = 0.5 отвергнуто биномиальным тестом; $p = 0.05$ с поправкой Бонферрони). Тот же анализ, что и в Рис.2.2c. **b:** Сравнение значений AI генов в данных, изображённых на панели **a**. **c:** Доли наблюдаемых и ожидаемых значений ΔAI для пяти данных квантилей (перечислены справа) по различным корзинам покрытия для тех же данных (сравните с Рис.2.3).

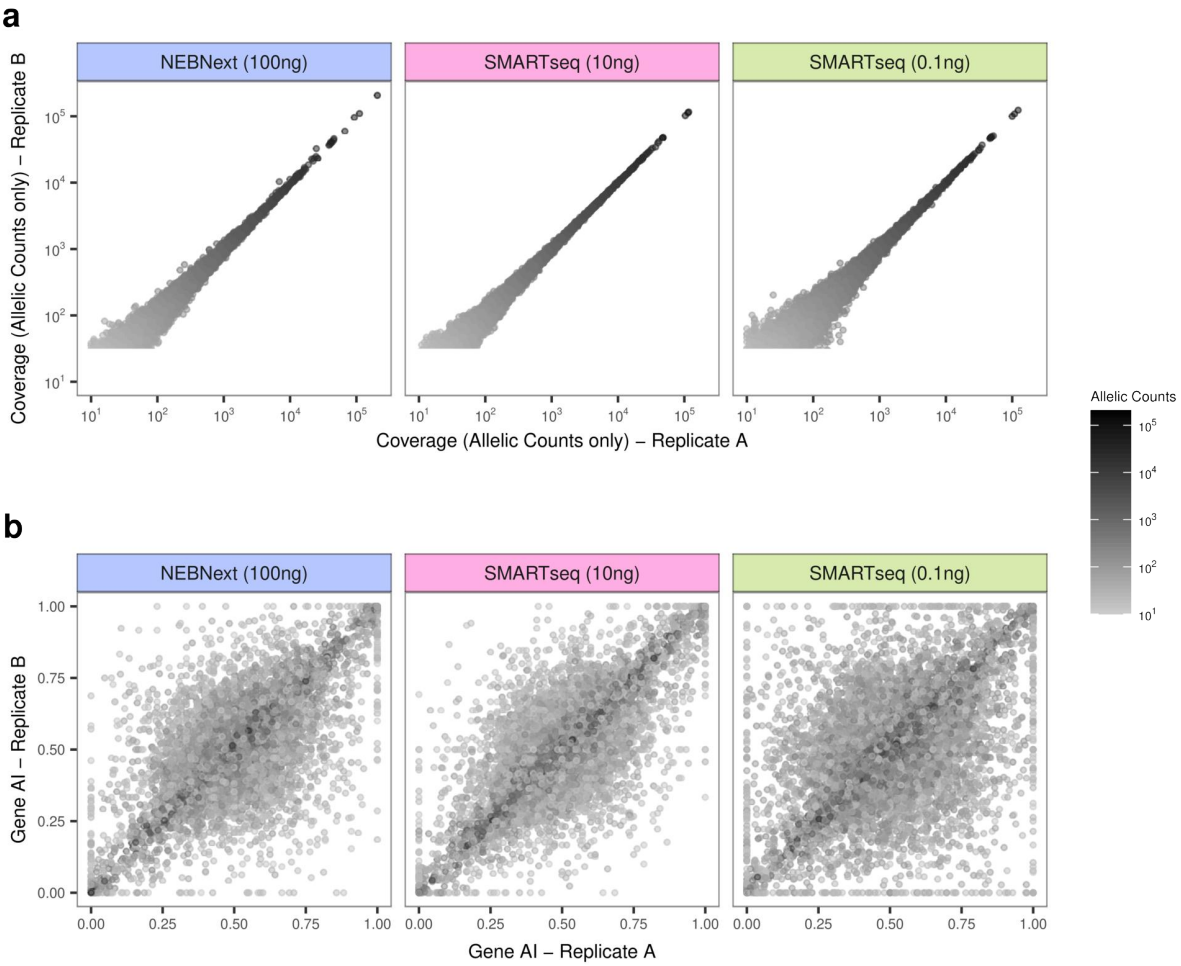

Рисунок A.20 — Сравнение согласованности между техническими репликами по аллельному покрытию и AI.

Сравнение двух реплик библиотек РНК-секвенирования, приготовленных из одной и той же РНК из почки мыши 129xCastF1. Слева направо: Эксперимент 1 [NEBNext(100ng)], Эксперимент 2 [SMARTseq(10ng)], и Эксперимент 3 [SMARTseq(0.1ng)]. Аллельные покрытия отражены шкалой серого; показаны только гены с аллельным покрытием > 10. **a:** Сравнение покрытий генов (только аллельные покрытия); **b:** Сравнение значений AI [(материнское аллельное покрытие)/(общее аллельное покрытие)].

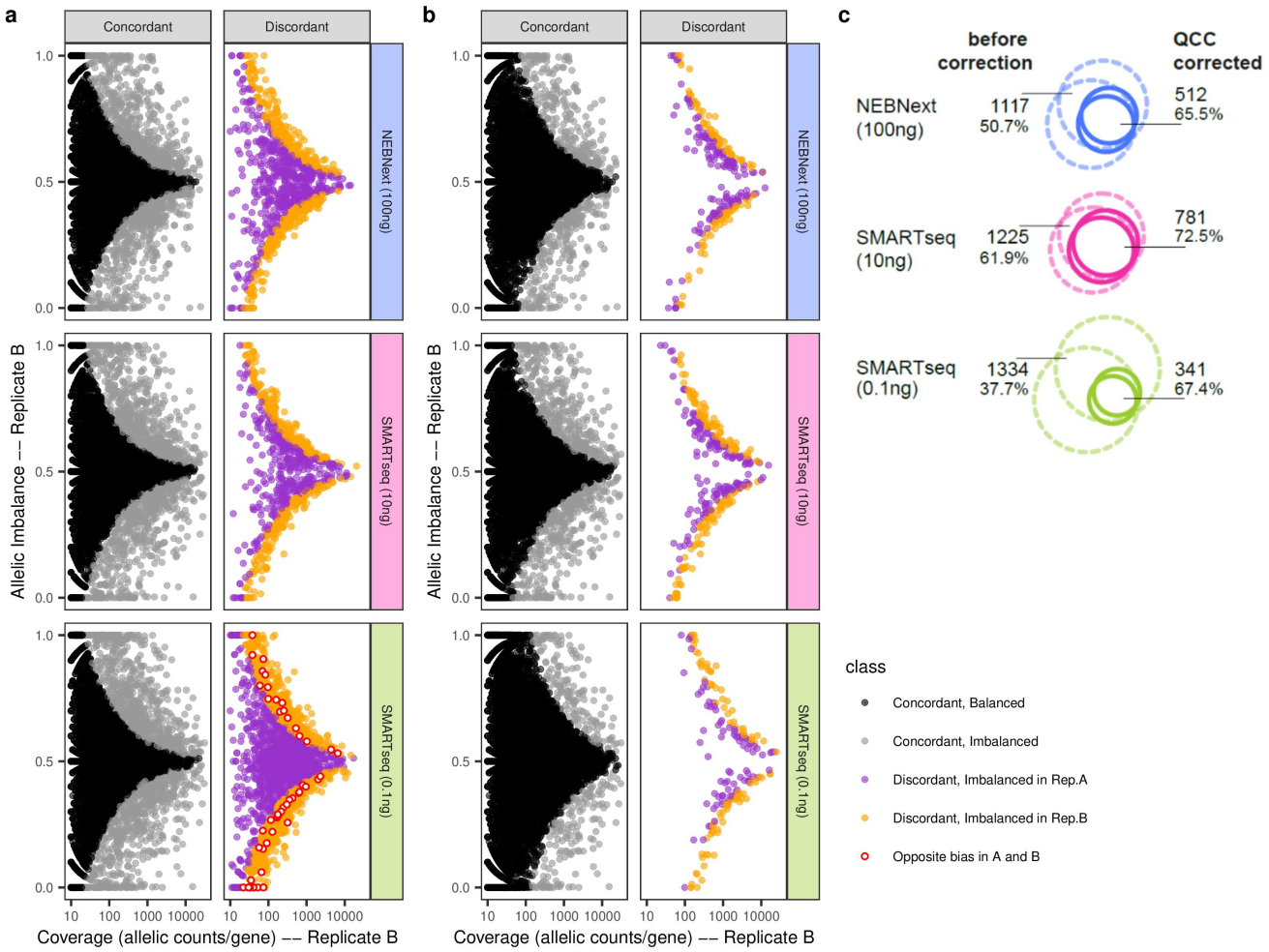

Рисунок A.21 — Эффект коррекции на QCC на согласованность между результатами тестов на смещённость аллельной экспрессии.

Применены или биномиальные тесты, или QCC-подправленные биномиальные тесты (H_0: AI = 0.5 отвергнуто; $p = 0.05$ с коррекцией Бонферрони). Рассмотрены только гены с общим аллельным покрытием более 10. **a-b:** Графики согласованности (серый/чёрный) и рассогласованности (цвета описаны в легенде; также см. Рис.2.2a между двумя техническими репликами. **a:** Биномиальный тест. Графики рассогласованности такие же, как и в Рис.2.2b. **b:** QCC-подправленный биномиальный тест. Графики рассогласованности такие же, как и в Рис.2.2f. **c:** Согласованность между генами классифицированными как аллельно дисбалансные между парами реплик в библиотеке РНК-секвенирования (то есть отражает все точки на панелях **a,b**, кроме чёрных, которые согласованно сбалансированны). Заметим, что все 12 кругов этой диаграммы Эйлера имеют согласованные масштабы. Пунктирная линия: использован биномиальный тест; сплошная линия: те же реплики, используя QCC-подправленный биномиальный тест. Сверху вниз: Эксперимент 1 [NEBNext (100ng)], Эксперимент 2 [SMARTseq (10ng)], Эксперимент 3 [SMARTseq (0.1ng)]. Показаны абсолютные значения перекрывающихся генов и % объединения, покрытый пересечением (% генов с согласованной дисбалансностью среди генов, названных дисбалансными хотя бы одной репликой).

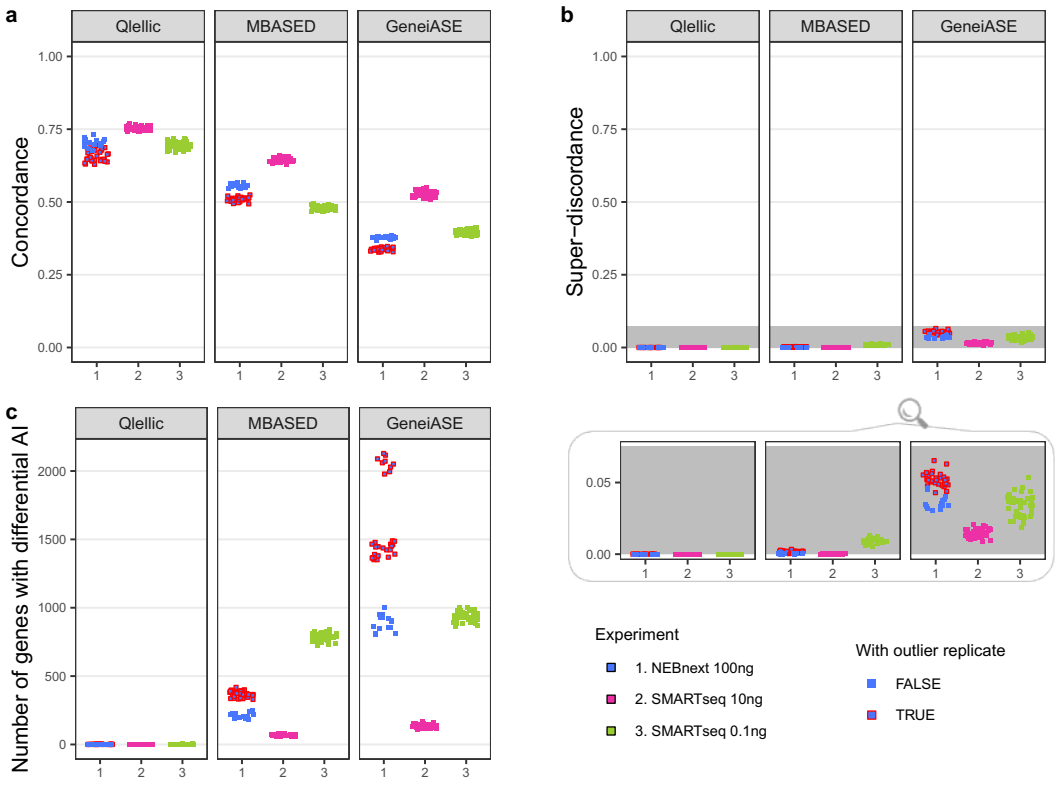

Рисунок А.22 — Согласованность между репликами при использовании различных инструментов для анализа аллель-специфической экспрессии.

a: Согласованность списков генов с значимым аллельным дисбалансом (H_0: $AI = 0.5$ отвергнуто) между всеми парами объединённых реплик. Также см. Рис.2.2e. **b:** Пропорция сверх-рассогласованных генов — таких, что каждая сторона сравнения утверждает о наличии значимого отклонения, но к разным аллелям. Также см. Рис.2.2b,f,g. Снизу: увеличение этого графика. **c:** Количество генов с значимым дифференциальным аллельным дисбалансом (H_0: $\mathrm{AI}_A = \mathrm{AI}_B$ отвергнуто) — заметим, что все тесты имеют дело с техническими репликами, поэтому такие результаты являются ложноположительными. Для MBASED, взято объединение асимметричных решений (A против B даёт результат, отличный от B против A). Также см. Рис.2.4a. Единица сравнения — некоторая комбинация двух индивидуальных реплик (например, [30M прочтений из реплики 1 вместе с 30M прочтениями из реплики 2] в сравнении с [30M прочтений из реплики 3 вместе с 30M прочтениями из реплики 4]). Показаны все 45 возможных попарных сравнений между такими комбинациями. Одни и те же прочтения и гены были использованы как вход для каждого из инструментов. Таблица количеств SNP была создана при помощи вычислительного протокола ASEReadCounter*. Для Qllelic, данные были обработаны согласно описанию соответствующей главы. Для MBASED и GeneiASE, прочтения из двух реплик были сложены вместе, так как эти инструменты не предлагают других способов работы с репликами. Одни и те же наборы генов были использованы после фильтрации (в каждом из 45 сравнений) используя требования и ограничения всех инструментов вместе: после суммирования прочтений в паре реплик, ген должен иметь менее чем 23 SNP с покрытиями для каждой из сторон сравнения, не обязательно одинаковые (MBASED: не может обработать гены с хотя бы 23 покрытыми SNP); гены должны иметь покрытие хотя бы 1 в каждой реплике, и хотя бы 8 в каждой объединённой паре реплик. **a-c:** Легенда применима ко всему рисунку. Все комбинации, которые включают реплику-выброс [#1 в эксперименте NEBnext (100ng)] обведены красным.

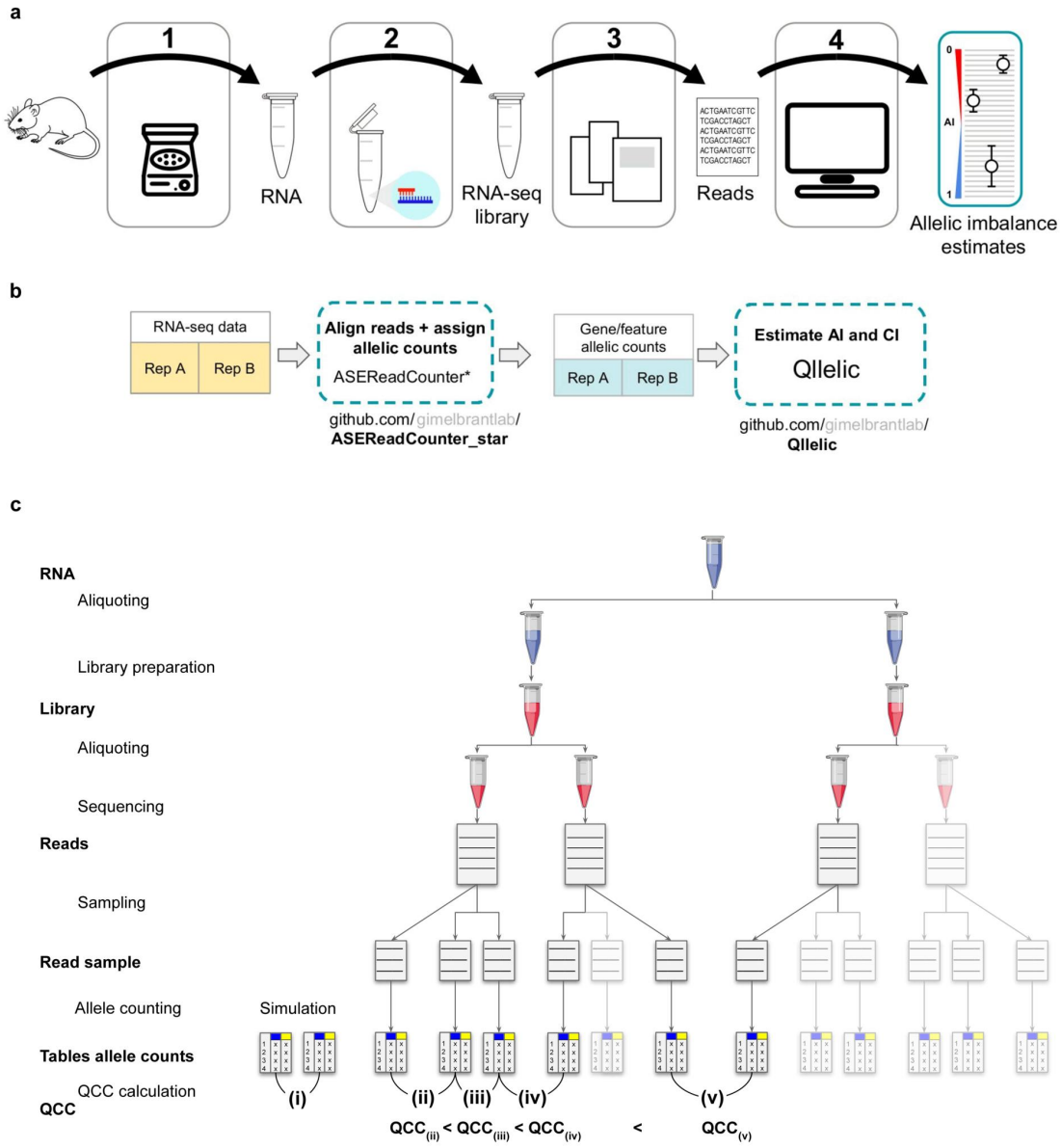

Рисунок А.23 — Основные этапы эксперимента РНК-секвенирования и последующего анализа данных.

a: Шаги эксперимента РНК-секвенирования. 1 — изоляция РНК; 2 — подготовка библиотеки; 3 — секвенирование, дающее прочтения (которое может быть парным в PE секвенировании). Отдельные прогоны секвенирования на одной библиотеке приводят к физической подвыборке библиотеки; 4 — анализ данных. **b:** Обзор анализа данных при помощи ASEReadCounter* и Qllelic. Заметим, что для подсчёта аллельных прочтений могут быть использованы и инструменты, отличные от ASEReadCounter* (например, см. Рис.А.24). **c:** Как наборы данных были подготовлены для анализа на источник избыточной дисперсии в обработке данных. Сравнения (i)–(v) описаны в деталях в секции результатов данной главы.

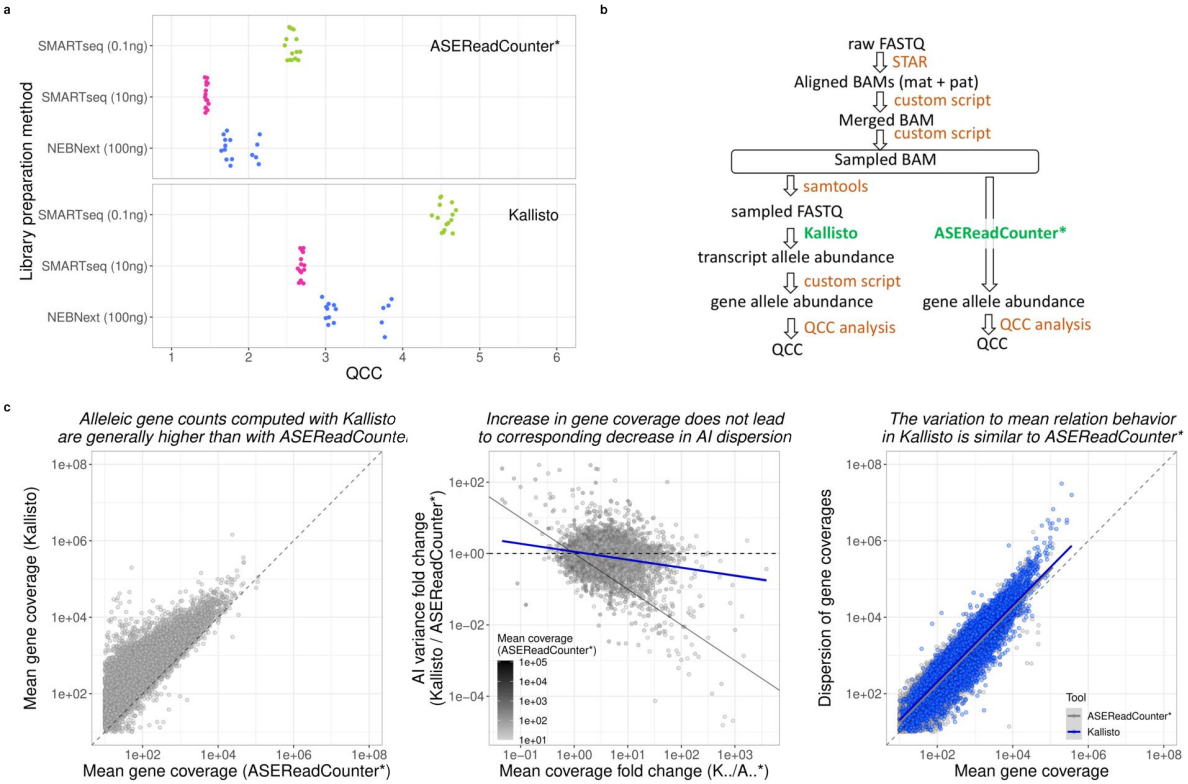

Рисунок А.24 — Присвоение аллель-неинформативных прочтений гаплотипам ведёт к увеличению избыточной дисперсии AI.

a: Значения QCC посчитаны для аллельных покрытий генов полученых из одного набора выравненных прочтений способом по умолчанию (ASEReadCounter*, учтены только информативные прочтения) и инструментом, который распределяет неинформативные прочтения гаплотипам (Kallisto). **b:** Схема обработки данных с помощью ASEReadCounter* и Kallisto. **c:** Источник повышенной избыточной дисперсии при использовании Kallisto для подсчёта аллельных покрытий. *Слева* — Kallisto возвращает более высокие покрытия (из-за вклада учтённых неинформативных прочтений). *Посередине* — несмотря на повышенные значения аллельных покрытий в результатах Kallisto, серьёзного уменьшения дисперсии AI не наблюдается (чёрная сплошная линия — ожидаемое взаимоотношение между относительным увеличением дисперсии AI и относительным увеличением в среднем покрытии генов; пунктирная линия — ожидаемые, если бы дисперсия AI была независима от среднего покрытия; синяя линия — линейный подбор наблюдаемого соотношения). *Справа* — График взаимоотношений между покрытием генов и дисперсией AI имеет похожий вид у Kallisto (синим) и ASEReadCounter* (серым), однако результаты Kallisto имеют большую избыточную дисперсию. Также сравните с Рис.2.4e. Данные:Для этого рисунка анализ был произведен на 30M выборочных прочтениях из каждой из шести реплик в наборе данных SMARTseq (10 ng).

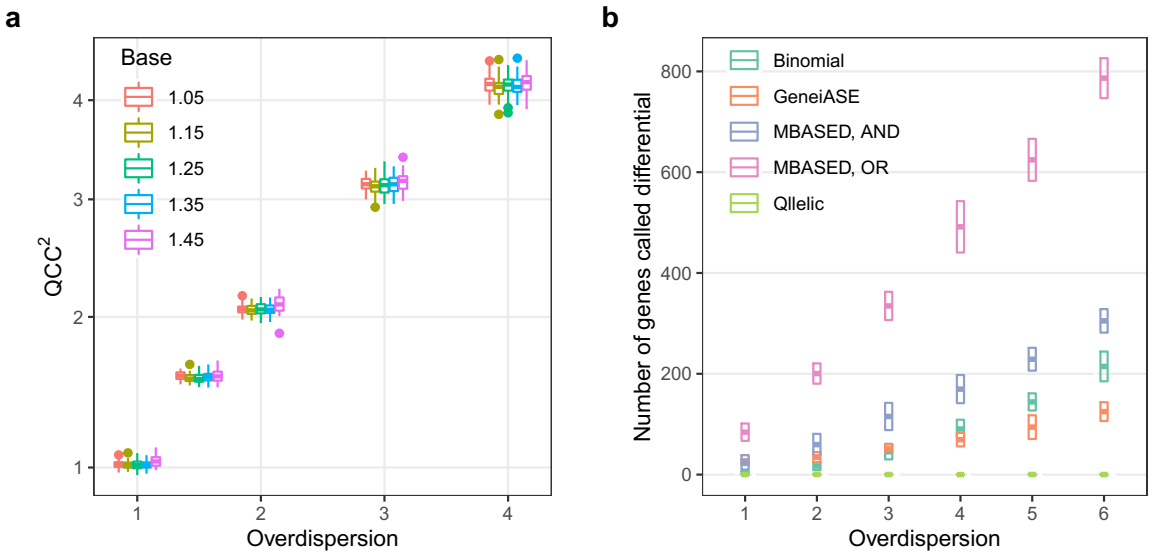

Рисунок A.25 — QCC отражает избыточную дисперсию AI, заложенную в симулированных данных.

a: Размеры корзин покрытий генов не меняют значимо оценки QCC. Значения QCC были посчитаны для симулированных данных с заданными значениями избыточной дисперсии AI (горизонтальная ось), для нескольких размеров корзин покрытий (показаны цветом и значением экспоненциального основания корзин). Заметим, что посчитанная избыточная дисперсия (QCC) почти идеально коррелирует со значениями, определёнными симуляцией. **b:** Точность инструментов Qllelic, MBASED, GeneiASE и биномиального теста на симулированных репликах с заданными значениями избыточной дисперсии AI: количество ложноположительных результатов, то есть генов, которые были классифицированы как имеющие значимо разные аллельные дисбалансы (отвергнуто H_0: $\mathrm{AI}_A = \mathrm{AI}_B$) при сравнении симулированных технических реплик. Из-за асимметрии вывода MBASED (итоговые значения p-уровней зависят от порядка указания входных образцов), результаты для каждой пары реплик были или объединены (OR), или пересечены (AND) в этом анализе. Заметим, что Qllelic не была дана предопределённая избыточная дисперсия; напротив, QCC был оценен исходя только из симулированных данных. **a,b:** Элементы диаграмм размаха — центральная отметка: медиана; границы прямоугольника: верхние и нижние квартили; усы: 1.5x интерквартильный интервал; точки: выбросы.
Описание симуляций см. в статье [123]

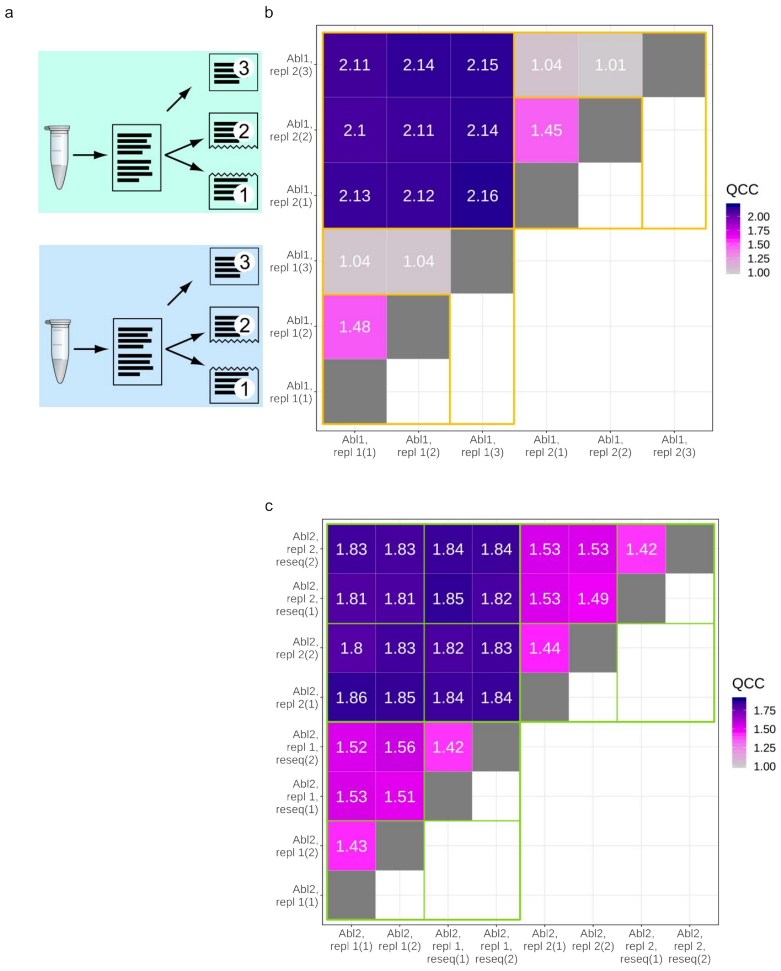

Рисунок A.26 — Источники избыточной дисперсии AI: влияние in-silico выборок и повторных прогонов секвенирования (физическая подвыборка библиотеки).

a: Схема выборки. “1” и “2” — два непересекающихся подмножества прочтений [см. определение в A.22c] из одного прогона секвенирования конкретной реплики после выравнивания и присваивания прочтений гаплотипам. “3” — независимая выборка (такого же размера, что “1” и “2”) из того же набора прочтений. Эта выборка находится в биномиальном соотношении с каждой из выборок “1” и “2”, и может пересекаться с ними. **b:** При сравнении внутри реплики, in-silico выборка без замещения даёт больший вклад в избыточную дисперсию AI, чем биномиальная выборка. Анализ попарных QCC был произведён на данных Abl.1, реплики 1 и 2. Подвыборки “1” и “2” являются непересекающимися подмножествами (15136606 фрагментов в каждой) данного прогона секвенирования библиотеки; “3” является случайным подмножеством 15136606 фрагментов из того же прогона секвенирования. Оранжевые прямоугольники подчёркивают сравнения между “1” и “2” и между ними и “3”. **c:** Повторные прогоны секвенирования той же библиотеки (“физическая подвыборка”) имеют малый эффект на избыточную дисперсию AI. Анализ попарных QCC был произведён на данных Abl.2, реплики 1 и 2. Повторные прогоны секвенирования аннотированы. Зелёные прямоугольники подчёркивают сравнения внутри и между повторными секвенированиями.

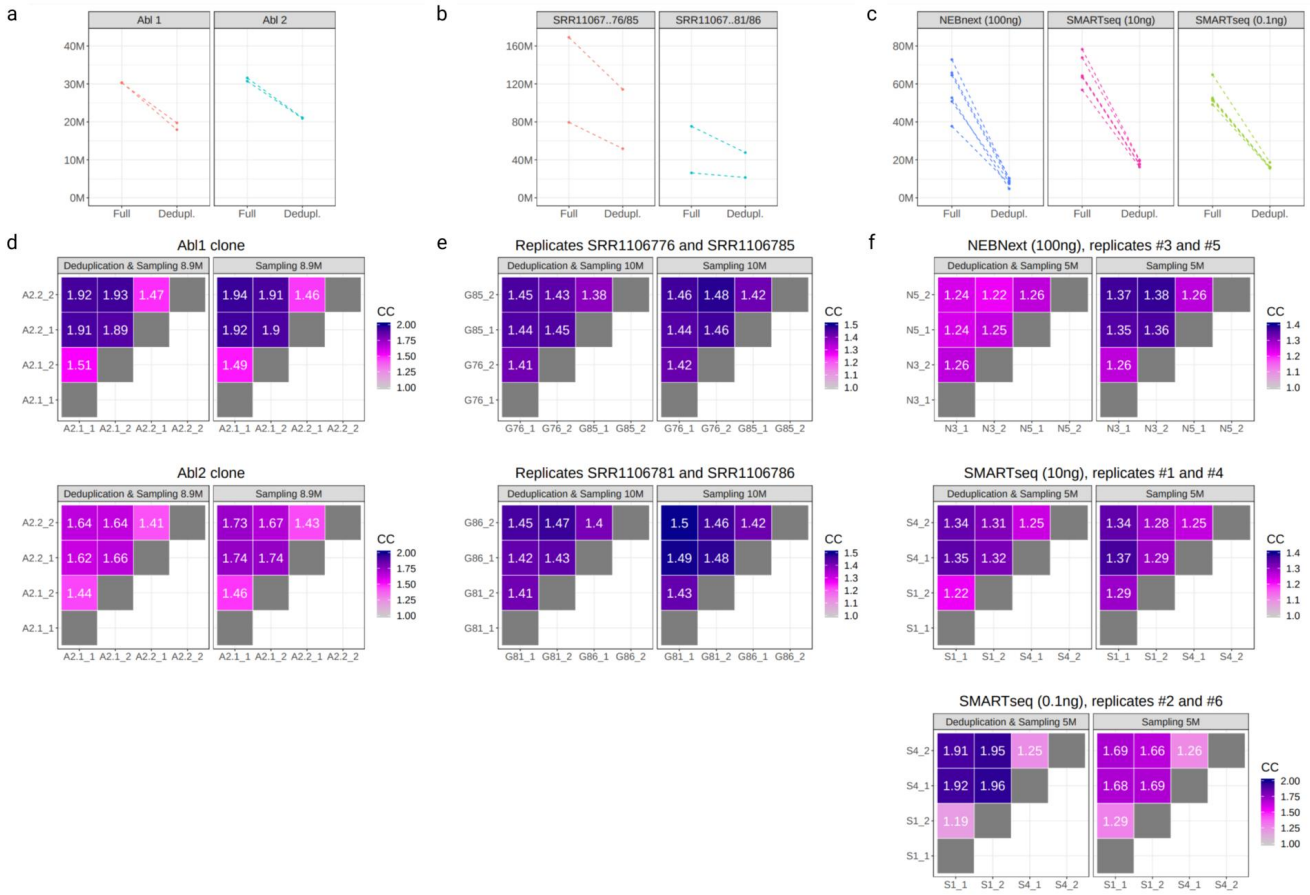

Рисунок A.27 — Источники избыточной дисперсии AI: влияние дедупликации.

a, b, c: Число фрагментов, остающихся после дедупликации (используя Picard MarkDuplicates). **a** — данные PE150 из Abl.1 и Abl.2 **b** — данные PE75 из NPC cells (набор данных Gendrel). **c** — данные SE75 для библиотеки РНК-секвенирования почек в Экспериментах 1-3. **d, e, f:** QCC после дедупликации всё ещё выше при сравнении двух реплик, чем при сравнении двух половин одной реплики. **d** — QCC до и после дедупликации: данные Abl.1 и Abl.2. **e -** QCC до и после дедупликации NPC клеток (набор данных Gendrel). **f -** QCC до и после дедупликации для случайно выбранных пар реплик из данных РНК-секвенирования почек. Заметим, что QCC может стать больше после дедупликации.

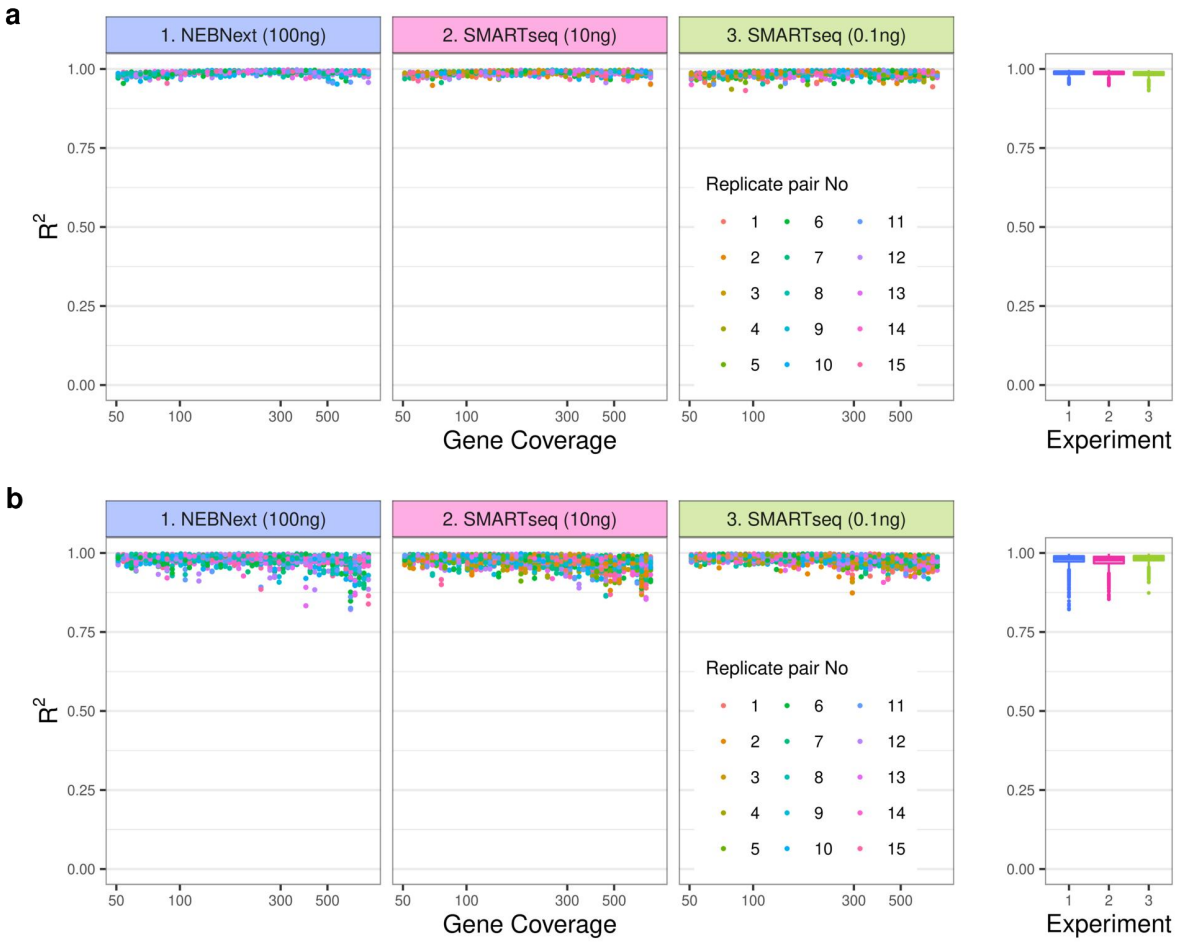

Рисунок А.28 — Качество подбора — R^2 для наблюдаемых и ожидаемых квантилей.

Статистика R^2 была посчитана для линейной регрессии между наблюдаемыми и ожидаемыми значениями квантилей (от 0.025 до 0.975 с шагом 0.025), используя функцию `summary.lm` из пакета `stats` в языке R. Каждая точка показывает корзину покрытий и одну из 15 пар реплик, проанализированных вместе (реплика-выброс в Эксперименте 1 была опущена в рамках этого анализа). **a:** Качество подбора QCC при помощи линейной регрессии, на значениях квантилей для наблюдаемых и смоделированных распределений ΔAI (детали см. в Рис.2.3f). **b:** Качество подбора AI при помощи модели смеси бета-биномиальных распределений, на значениях квантилей для наблюдаемых и смоделированных распределений AI (детали см. в Рис.2.3d). Для каждой пары реплик, были сделаны сравнения симулированных реплик с действительными репликами. **a,b:** Элементы диаграмм размаха — центральная отметка: медиана; границы прямоугольника: верхние и нижние квартили; усы: 1.5x интерквартильный интервал; точки: выбросы.

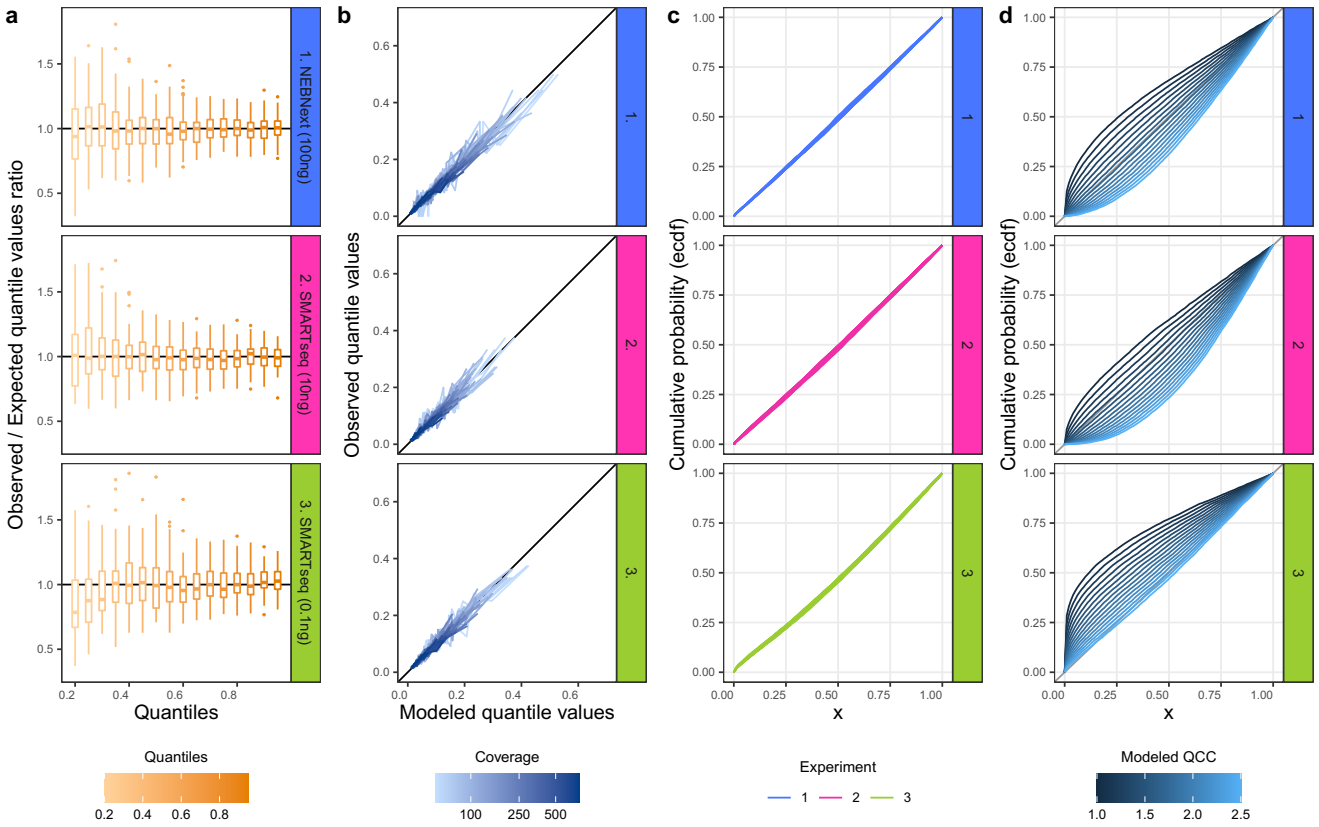

Рисунок A.29 — Качество подбора — квантиль-квантиль графики.

a, b: После подсчёта QCC, оценки AI для двух симулированных реплик (см. Рис.2.3d) были перегенерированы, используя подобранные бета-параметры распределения AI и добавляя шум биномиальной выборки согласно QCC (покрытие гена в QCC^2 раз меньше). Значения квантилей (от 0.20 до 0.95 с шагом 0.05) были сравнены у распределений ΔAI смоделированных и наблюдаемых данных. **a:** Диаграммы размаха пропорций значений квантилей между наблюдаемыми и смоделированными данными. Заметим, что данные распределены вокруг 1, что является признаком того, что скорректированная модель хорошо описывает данные. Каждая диаграмма размаха показывает пропорции квантилей для разных корзин покрытий, на примере пары реплик (реплики #3 против реплики #4 в каждой эксперименте). **b:** Квантиль-квантиль графики для наблюдаемых и смоделированных распределений, на примере пары реплик (#3 против #4). Корзины покрытий обозначены цветом. **c:** Эмпирическая кумулятивная функция распределения (ecdf) для p-уровней дифференциальных тестов во всех 45 множествах двух пар реплик в каждом эксперименте (15 множествах в эксперименте NEBnext, так как реплика-выброс #1 была опущена), проделанных с вычисленными соответственно QCC. Заметим, что ecdf близка к диагонали $x = y$, как и ожидалось от подходящего данным теста. **d:** Эмпирическая кумулятивная функция распределения p-уровней дифференциальных тестов между двумя парами реплик (#2&#3) против (#4&#5), для QCC, смоделированных от 1 до 2.5 (с шагом 0.1). Кривые *ниже* диагонали представляют излишне консервативные тесты, кривые *над* диагональю отражают недооценённую дисперсию. Заметим, что линия, наиболее близкая к диагонали, соответствует значению QCC, определённому в анализе Qllelic (1.67 и 1.72 для пар в Эксперименте 1; 1.45 и 1.47 для Эксперимента 2; 2.62 и 2.53 для Эксперимента 3).

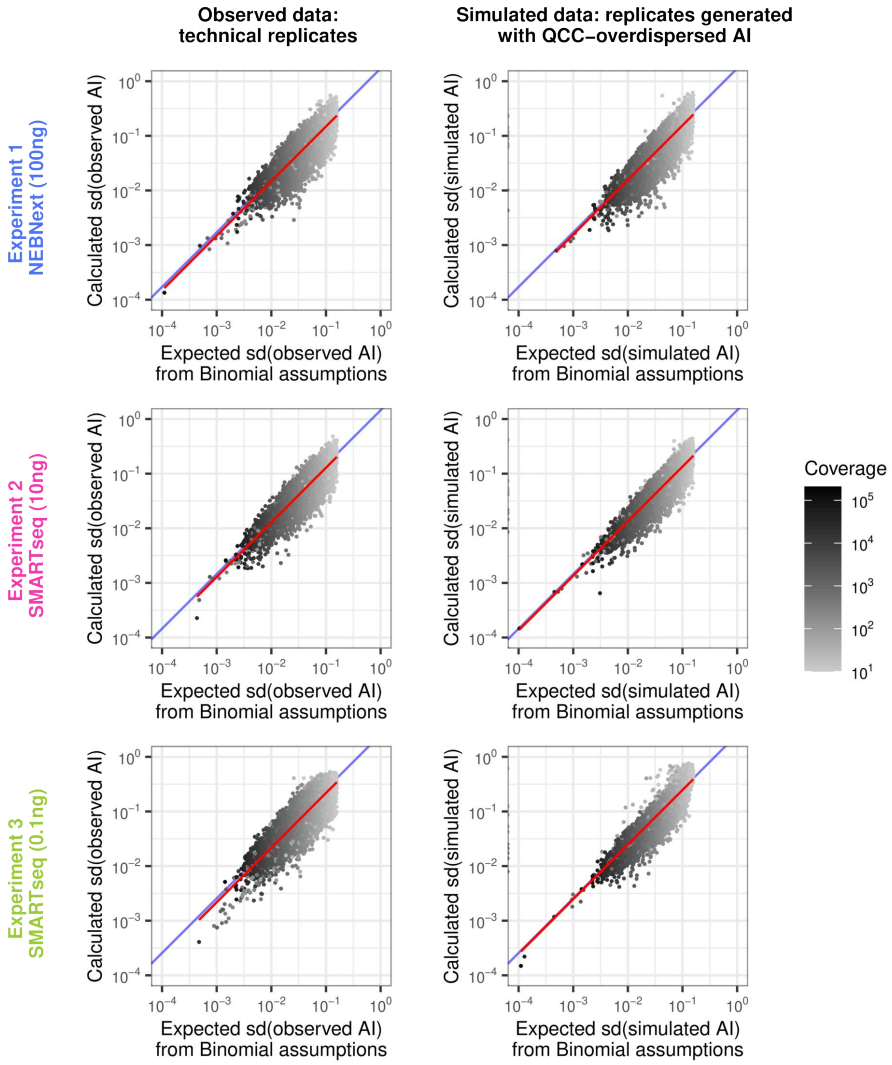

Рисунок А.30 — Качество подбора — на уровне единичного гена, дисперсия AI соответствует ожиданиям.

Регрессия между посчитанной $sd(AI)$ и ожидаемой $sd(AI)$ дисперсиями на уровне гена (красная линия) близко к тренду QCC (синия линия) и в реальных (*слева*), и в симулированных данных (*справа*). *Сверху вниз* — эксперименты 1–3; для симуляций были использованы значения AI и покрытий из экспериментальных данных. Вычисленные значения $sd(AI)$ были получены на 5 репликах (Эксперимент 1, с опущенной репликой-выбросом) или всех 6 репликах в Экспериментах 2 and 3. Ожидаемые значения $sd(AI)$ были оценены исходя из биномиальных предположений: $\sqrt{\frac{\text{AI}\cdot(1-\text{AI})}{\text{coverage}}}$. Точечные оценки AI и покрытия были получены из всех участвующих реплик. Для симуляции, средние наблюдаемые значения покрытия генов были взяты как параметры λ для Пуассоновского распределения, использованного для симуляции покрытия для каждой из реплик в каждом эксперименте (то есть 5 или 6). Значения AI были взяты из точечных оценок на наблюдаемых данных, и материнские аллельные покрытия были выбраны из биномиального распределения со значениями n, равными покрытиям, делённым на экспериментально полученный QCC в квадрате.

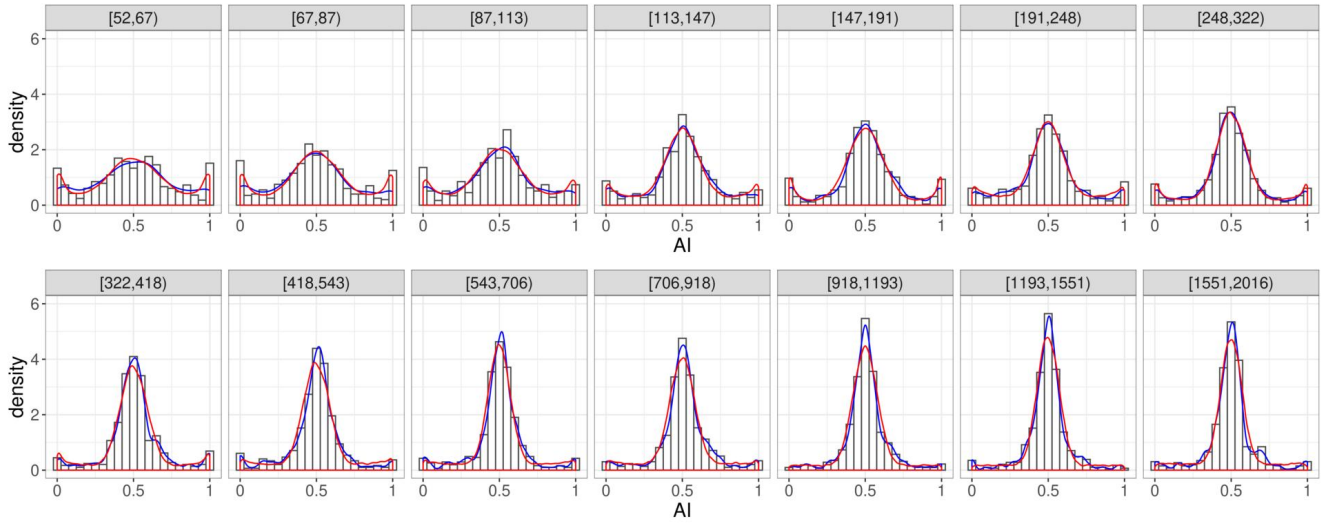

Рисунок A.31 — Максимизация ожидания при подборе аллельного дисбаланса при помощи смеси бета-биномиальных распределений.

Наблюдаемое распределение AI в экспериментальных данных в корзинах покрытий генов (согласно пометкам в графиках). Заметим, что сглаженный график плотности гистограммы (синяя линия) близок к подобранному распределению (красная линия). Данные: не дедуплицированные данные РНК-секвенирования для клона Abl.1, из которой подвыбраны 30 273 212 PE150 прочтений в каждой из двух технических реплик; основание экспоненциальной разбивки на корзины равно 1.3.

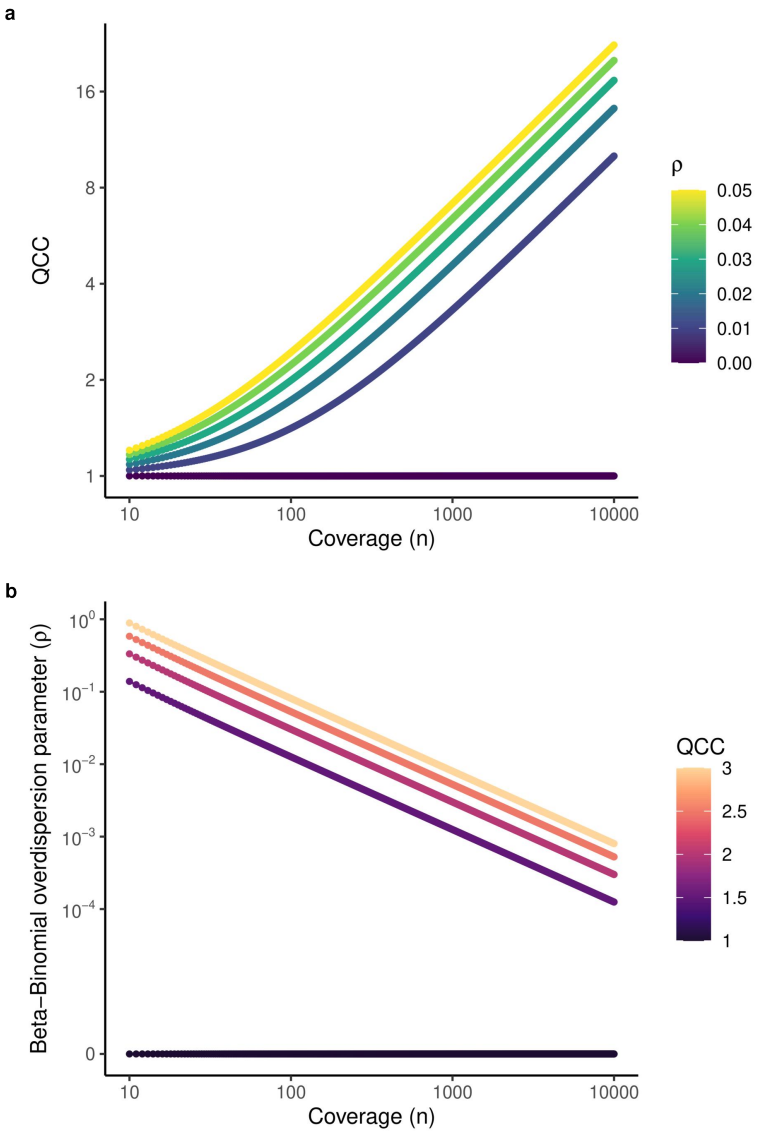

Рисунок A.32 — Соотношение между QCC и учётом избыточной дисперсии AI при помощи бета-биномиального распределения с эксперимент-специфическим параметром избыточной дисперсии ρ.

a: Значение QCC, которое соответствует данному параметру ρ и даёт ту же ширину распределения AI на конкретном покрытии. **b:** Параметр избыточной дисперсии ρ, который соответствует данному значению QCC и даёт ту же ширину распределения AI на конкретном покрытии. Заметим, что использование $\rho = 0.03$ ([61]) приводит к недооценке избыточной дисперсии AI для плохо покрытых генов и переоценке избыточной дисперсии AI для хорошо покрытых генов.

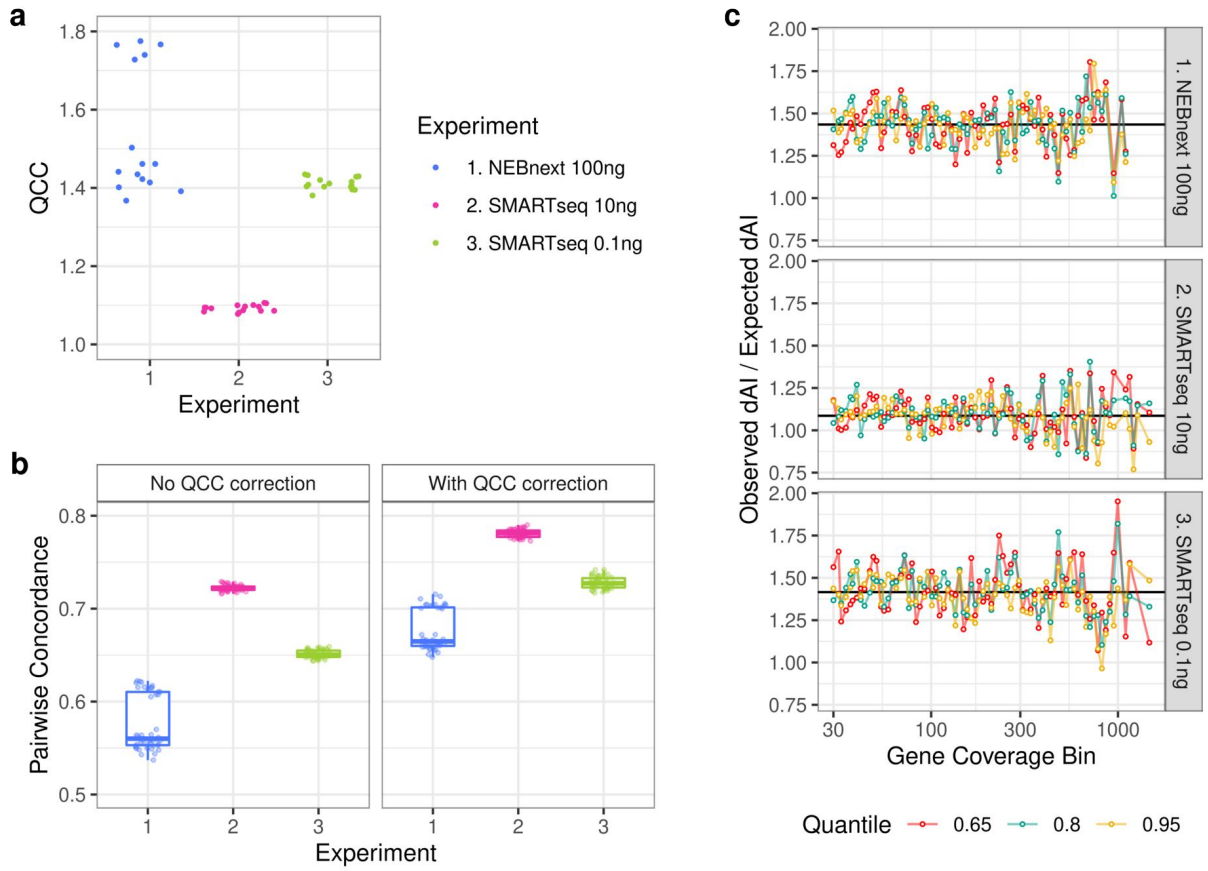

Рисунок A.33 — Избыточная дисперсия AI остаётся эксперимент-специфической при подсчёте на покрытиях индивидуальных SNP.

a: Значения QCC, посчитанные для всех возможных пар реплик для всех трёх экспериментов (сравните с Рис.2.2d). **b:** Согласованность теста на дисбалансность во всех возможных парах реплик, H_0 : AI = 0.5 отвергнуто биномиальным (*слева*) или QCC-откорректированным биномиальным (*справа*) тестом; уровень достоверности равен 0.95, применена поправка Бонферрони (сравните с Рис.2.2e). Элементы диаграмм размаха — центральная отметка: медиана; границы прямоугольника: верхние и нижние квартили; усы: 1.5x интерквартильный интервал; точки: выбросы. **c:** Частное наблюдаемых к ожидаемым значениям ΔAI для трёх фиксированных квантилей по различным корзинам покрытий для двух выбранных реплик из каждого из трёх экспериментов (сравните с Рис.2.3f).

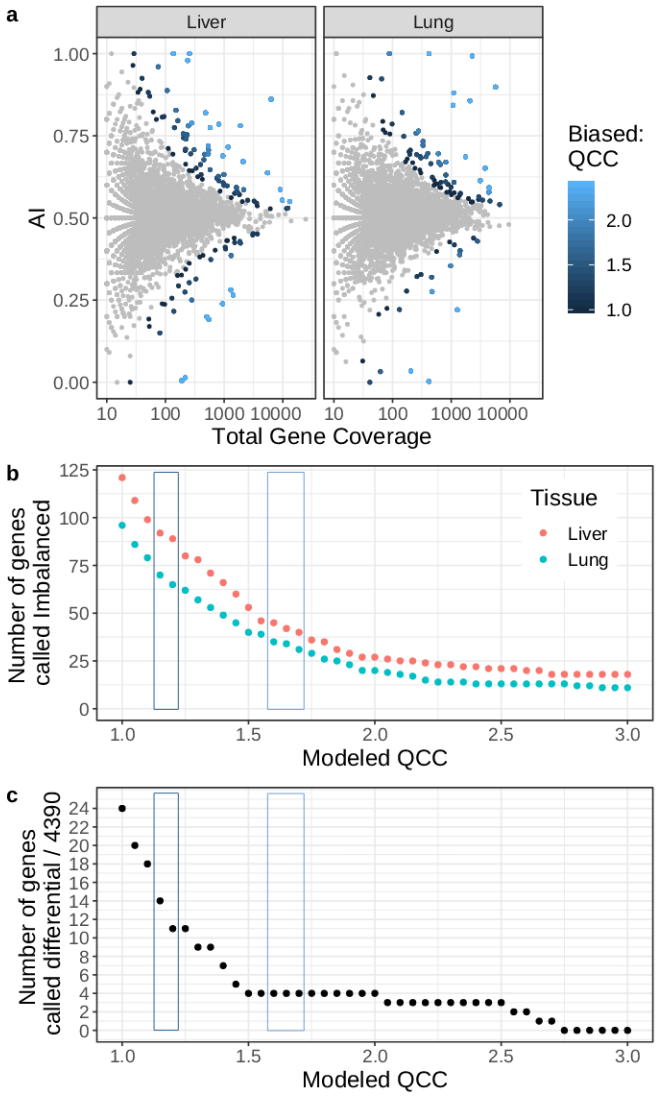

Рисунок A.34 — Влияние значений QCC на анализ аллель-специфической экспрессии в взятом для примера наборе данных GTEx.

Чтобы проиллюстрировать влияние избыточной дисперсии на анализ дисбаланса аллельной экспрессии в наборе данных без реплик, мы рассмотрели данные РНК-секвенирования для двух человеческих тканей, печени и лёгких, из одного и того же произвольно выбранного индивида GTEX-11NUK из проекта GTEx (репликация отсутствует). Аллельные прочтения для этих образцов предоставлены в исходных файлах статьи [106] (файл `SourceData.zip`). **a:** Расположение значимо дисбалансных генов (тестирование H_0 : AI = 0.5) при различных опробованных значениях QCC от 1.0 (избыточная дисперсия отсутствует) до 3.0. **b:** Тот же анализ, что и в **a**, показывающий количество дисбалансных генов для двух тканей. **c:** Дифференциальный анализ дисбаланса аллельной экспрессии между двумя тканями после коррекции при помощи различных значений QCC (равное для двух образцов). Покрашенные прямоугольники соответствуют диапазону QCC, наблюдаемому в данных Geuvadis (1.15-1.2) и наших данных, эксперимента NEBnext (1.6-1.7).

A.3 Сопроводительные таблицы к главе 2

Таблица 2 — Наборы данных РНК-секвенирования, анализировавшиеся в проекте `Qllelic`.

Образцы	Репл.	# фрагментов	Тип секв.	GEO ID	Ссылка
F1 129S1/CAST почка, Эксп. 1 (NEBnext 100ng)	1	63850091	SE 75	GSE143310	[106]
	2	77092011			
	3	87531947			
	4	45547354			
	5	79260529			
	6	61713808			
F1 129S1/CAST почка, Эксп. 2 (SMARTseq 10ng)	1	85084237	SE 75		
	2	73162672			
	3	65565213			
	4	74019699			
	5	73733135			
	6	90229986			
F1 129S1/CAST почка, Эксп. 3 (SMARTseq 0.1ng)	1	57996564	SE 75		
	2	58816137			
	3	72585208			
	4	57195273			
	5	54874918			
	6	57807668			
мышь, клон Abl.1	1	34228003	PE 150		
	2	33297046			
мышь, клон Abl.2	1	33116795	PE 150		
	2	33929680			
мышь, клон Abl.2 вторая итерация секв-я	1	37381472	PE 150		
	2	36658803			
Данные из других исследований					
Человеческие данные РНК-секвенирования.					
Geuvadis study – см Таблицу 3					[93]
GTEx study: печень лёгкое		 GTEX-11NUK-1226-SM-5P9GM GTEX-11NUK-0826-SM-5HL4U			[125]
Мышиные данные РНК-секвенирования.					
Нейрональных клетки-предшественники (GSE54016) – см Таблицу 4					[29]

Таблица 3 — Анализ технических реплик РНК-секвенирования в человеческих клеточных линиях.

образец	технические реплики	# картир. фрагментов в образце (PE 75)	# генов с покрытием >8	# генов AI! = 0.5 (QCC = 1)	QCC	# генов AI! = 0.5 (с учётом QCC) в попарных сравнениях
HG00117	ERR205004	24,185,758	4609	269, 272, 272,	1.104, 1.124, 1.104,	225, 236, 236,
	ERR204894	29,013,132	4489	268, 275, 263,	1.152, 1.133, 1.113,	219, 225, 225,
	ERR204909	22,442,567	4617	285, 264, 277,	1.119, 1.093, 1.108,	234, 235, 228,
	ERR204950	28,202,246	4651	266, 269, 276,	1.096, 1.14, 1.148,	237, 216, 230,
	ERR204975	15,440,366	4351	269, 278, 272,	1.163, 1.172, 1.132,	216, 225, 224,
	ERR205006	26,668,943	4478	265, 273, 272,	1.138, 1.143, 1.155,	213, 227, 210,
	ERR204879	21,722,554	4605	280, 262, 271	1.124, 1.152, 1.162	226, 211, 214
			4542.9 ± 107.51	**271.3 ± 5.85**	**1.13 ± 0.023**	**224.4 ± 8.61**
HG00355	ERR204824	26,593,833	4793	212, 203, 205,	1.093, 1.109, 1.082,	189, 160, 180,
	ERR204831	28,088,219	4755	200, 199, 208,	1.088, 1.106, 1.173,	184, 170, 156,
	ERR204846	28,142,959	4737	204, 205, 194,	1.126, 1.13, 1.101,	167, 172, 173,
	ERR204854	24,273,505	4756	206, 205, 199,	1.13, 1.142, 1.115,	172, 175, 158,
	ERR204901	23,717,053	4760	195, 196, 202,	1.088, 1.148, 1.169,	172, 155, 150,
	ERR204953	29,204,264	4726	204, 201, 203,	1.076, 1.126, 1.163,	187, 160, 164,
	ERR204972	18,434,828	4503	200, 205, 198	1.09, 1.144, 1.137	171, 166, 170
			4718.6 ± 97.34	**202.0 ± 4.42**	**1.12 ± 0.029**	**169.1 ± 10.49**
NA06986	ERR204855	30,501,691	4654	295, 300, 304,	1.125, 1.143, 1.154,	250, 255, 236,
	ERR204860	22,698,694	4633	310, 308, 315,	1.141, 1.112, 1.083,	241, 261, 267,
	ERR204863	27,287,384	4546	301, 299, 298,	1.113, 1.158, 1.165,	257, 230, 232,
	ERR204929	18,328,969	4494	308, 320, 313,	1.107, 1.108, 1.101,	259, 272, 266,
	ERR204955	24,106,890	4616	307, 322, 316,	1.129, 1.058, 1.045,	253, 297, 290,
	ERR204968	23,724,945	4641	304, 322, 313,	1.161, 1.142, 1.104,	241, 256, 264,
	ERR205005	23,033,870	4631	316, 310, 332	1.148, 1.121, 1.098	243, 265, 280
			4602.1 ± 59.19	**310.1 ± 9.35**	**1.12 ± 0.032**	**257.9 ± 17.77**
NA19095	ERR204843	28,887,221	5782	267, 278, 255,	1.157, 1.159, 1.208,	211, 211, 196,
	ERR204858	28,260,083	5991	262, 258, 274,	1.151, 1.185, 1.192,	204, 200, 183,
	ERR204861	18,124,458	5719	255, 262, 263,	1.098, 1.094, 1.099,	219, 226, 224,
	ERR204868	29,178,865	5901	256, 271, 261,	1.111, 1.077, 1.117,	223, 234, 217,
	ERR204891	22,350,859	5984	262, 246, 253,	1.102, 1.087, 1.085,	218, 214, 225,
	ERR204930	19,336,064	6018	264, 258, 271,	1.112, 1.118, 1.122,	219, 213, 214,
	ERR205009	25,050333	5965	255, 279, 255	1.142, 1.103, 1.125	201, 228, 215
			5908.6 ± 115.22	**262.1 ± 8.61**	**1.13 ± 0.037**	**214.0 ± 12.01**
NA20527	ERR204830	32,519,775	4766	213, 211, 209,	1.097, 1.063, 1.077,	184, 189, 184,
	ERR204874	25,882,966	4849	220, 204, 206,	1.118, 1.15, 1.068,	178, 166, 188,
	ERR204908*	12,598,312	4728	208, 212, 211,	1.11, 1.11, 1.141,	178, 187, 168,
	ERR204934	24,903,624	4825	204, 208, 213,	1.079, 1.112, 1.123,	190, 183, 187,
	ERR204965	19,808,430	4862	222, 210, 209,	1.176, 1.132, 1.133,	168, 173, 176,
	ERR204978	33,101,465	4836	205, 207, 219,	1.12, 1.188, 1.107,	182, 159, 186,
	ERR204993	27,981,881	4855	205, 212, 201	1.177, 1.107, 1.095	155, 179, 178
			4817.29 ± 50.73	**210.0 ± 5.44**	**1.12 ± 0.034**	**178.0 ± 9.92**

* - образец с минимальных общим покрытием; для однообразного анализа, все образцы были сэмплированы до этого числа прочтений.

Технические реплики были доступны для пяти образцов из работы [93], с 7 библиотеками сделанными для каждой приготовленной РНК. Избыточная дисперсия и другие метрики AI показаны для всех возможных парных сравнений внутри каждого набора реплик.

Таблица 4 — Избыточная дисперсия и другие свойства для данных РНК-секвенирования нейрональных клеток-предшественников мыши.

образец	биологические реплики	# картир. фрагментов в образце (PE 100)	# генов с покрытием >8	# генов AI! = 0.5 (QCC = 1)	QCC	# генов AI! = 0.5 (с учётом QCC) в попарных сравнениях
SRS529152	SRR1106776	79,524,208	12531	3338	1.51	1995
SRS529162	SRR1106785	169,065,784	12661			
SRS529159	SRR1106781	75,319,044	12383	3104	1.56	1699
SRS529163	SRR1106786*	26,302,221	12373			

* - образец с минимальных общим покрытием; для однообразного анализа, все образцы были сэмплированы до этого числа прочтений.

2 биологических реплики были доступны для двух образцов в [29]. Избыточная дисперсия и другие метрики AI показаны для пар реплик.

Приложение Б

К главе 3, «Метилирование ДНК является ключевым механизмом для поддержания моноаллельной экспрессии на аутосомах»

Б.1 Сопроводительные заметки к главе 3

Б.1.1 Многомерная линейная регрессия без предикторов

Метод разработан А. А. Мироновым

Мотивация

Линейная регрессия — это часто используемый метод для описания взаимоотношений между несколькими векторами измерений. Самый частый двумерный случай представляет из себя анализ независимой величины X и зависимой величины Y ($Y \sim X$). Иначе говоря, X является предиктором Y. Например, этот анализ выполняется функцией `lm()` из пакета `stats` на языке `R`, и выбор независимой переменной для этого анализа влияет на результаты (Рисунок **Б.1a,b**).

Когда переменные независимы друг от друга (нет предиктора), этот подход не применим. Чтобы решить задачу подбора регрессии без предикторов, мы опишем и будем использовать метод основанный на сингулярном разложении матрицы (SVD), который имеет своей целью подобрать наилучший линейный тренд, описывающий данные. Этот метод относится ко всем размерностям равнозначно, с учётом выбора дисперсии нормального шума по каждой размерности. Мы также расширили этот метод добавлением доверительной полосы (Рисунок **Б.1c**).

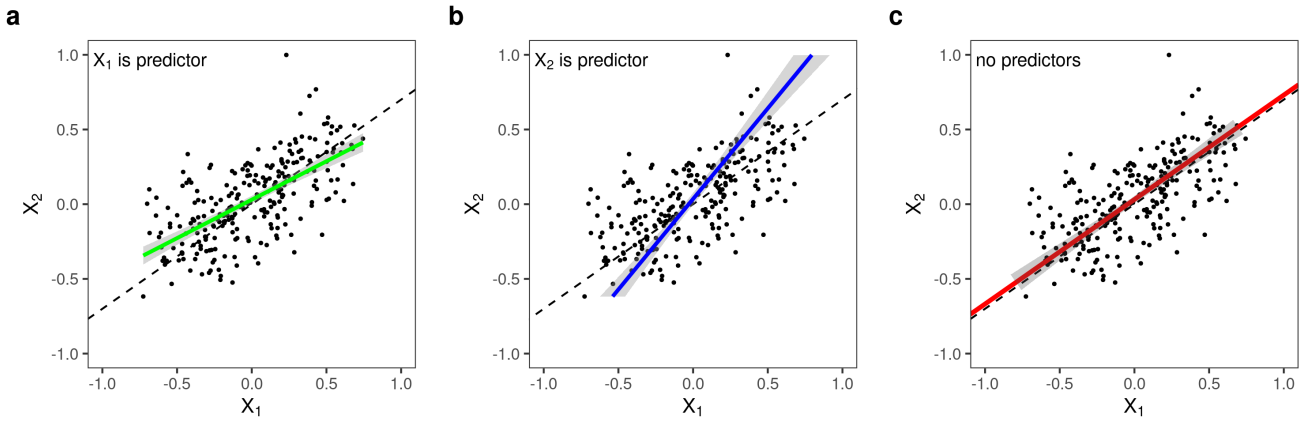

Рисунок Б.1 — Пример подбора линейного тренда для двумерных данных

(a) линейная регрессия при помощи `lm()`; $X_2 \sim X_1$. **(b)** то же, что (a), $X_1 \sim X_2$. **(c)** линейная регрессия без предикторов, вместе с доверительной полосой. **(a-c)** В рисунках использованы одни и те же данные: 250 точек были выбраны случайно со средним отношением координат 10:7. Тренд обозначен пунктирной чёрной линией. К тренду был добавлен нормальный шум с параметрами $\mathrm{sd}(X_1) = 0.15$ и $\mathrm{sd}(X_2) = 0.2$.

Многомерный подбор линейного тренда

Постановка задачи Пусть у нас есть некоторое множество *наблюдаемых* векторов $\{x_d\}$ размерности D. Например, это двумерные вектора. Мы хотим в некотором смысле симметрично и оптимально провести прямую линию, которая лучше всего описывает данные (Рисунок Б.1):

$$x_{d,i} = \tilde{x}_{d,i} + \xi_d; \quad \tilde{x}_{d,i} = a_d \cdot t_i + b_d$$

или в векторной форме:

$$\boldsymbol{x}_i = \tilde{\boldsymbol{x}}_i + \boldsymbol{\xi}; \quad \tilde{\boldsymbol{x}}_i = \boldsymbol{a} \cdot t_i + \boldsymbol{b}$$

i — порядковый номер вектора в наборе данных. Здесь мы предположили две вещи. Во-первых, наблюдаемые значения $\{x_d\}$ есть "истинные" значения $\tilde{x}_{d,i}$ плюс случайный шум $\{\xi_d\}$; "истинные" значения $\tilde{x}_{d,i}$ линейно параметризованы некоторым параметром t. Более того мы можем предполагать, что шум также

может зависеть от параметра t:

$$\boldsymbol{x}_i = \boldsymbol{a} \cdot t_i + \boldsymbol{b} + \boldsymbol{\xi}$$

Задача заключается в том, чтобы найти оптимальные значения параметров $\boldsymbol{a}, \boldsymbol{b}$, определяющих тренд. Будем оптимизировать правдоподобие. Здесь и далее индекс d относится к номеру размерности, а индексы i, k – к номеру наблюдения.

Можно также после определения тренда $\boldsymbol{a}, \boldsymbol{b}$ поставить задачу поиска наиболее вероятных значений $\tilde{\boldsymbol{x}}_i$, которые, как бы, были на самом деле.

Правдоподобие Для подсчета правдоподобия допустим, что параметры $\boldsymbol{a}, \boldsymbol{b}$ нам известны. Тогда вычислим вероятность того, что некоторый вектор порожден некоторым значением параметра t (Рисунок Б.2).

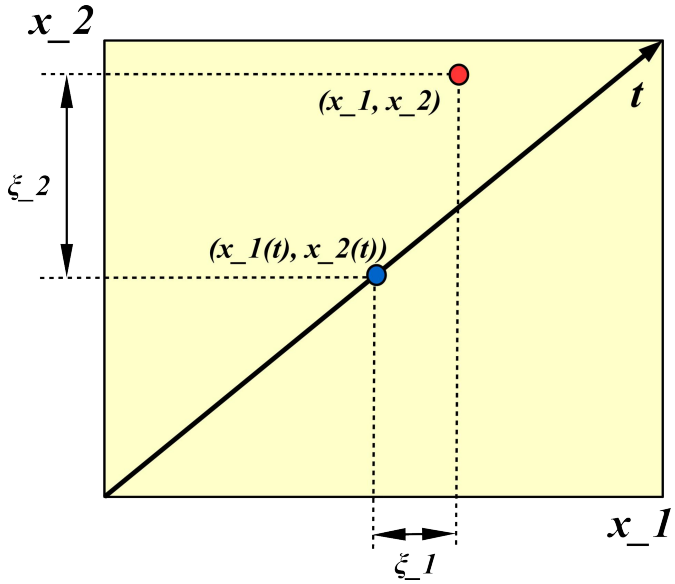

Рисунок Б.2 — Вектор ($\boldsymbol{x}$) порожден некоторым значением параметра t

Вероятность такого порождения определяется формулой

$$\Pr(\boldsymbol{x} \,|\, t) = \prod_d \Pr(\boldsymbol{x} \,|\, t) \qquad \text{(Б.1)}$$

здесь мы считаем, что по всем координатам шумы независимы. Вероятность наблюдаемых значений определим как

$$\Pr(x_d|t) = f_d(x_d - (a_d \cdot t + b_d), t)$$

где $f_d(x,t)$ – плотность распределения шума $\xi_d(t)$ при значении параметра t. Тогда уравнение (Б.1) перепишется в виде:

$$\Pr(\boldsymbol{x}|t) = \prod_d \Pr(x_d|t) = \prod f_d(x_d - (a_d \cdot t + b_d),t) \prod f_d(x_d - \hat{x_d},t) \tag{Б.2}$$

Чтобы получить правдоподобие $\Pr(x)$ надо проинтегрировать уравнение (Б.2) по t:

$$\Pr(\boldsymbol{x}) = \int_{-\infty}^{+\infty} \prod f_d(x_d - \hat{x_d}, x_d)dt \tag{Б.3}$$

Правдоподобие всей серии наблюдений

$$\begin{aligned} \Pr(\boldsymbol{x}_i|\boldsymbol{a},\boldsymbol{b}) &= \prod_i \int_{-\infty}^{+\infty} \prod_d f_d(x_{d,i} - \hat{x_d}, t)dt \\ &= \prod_i \int_{-\infty}^{+\infty} S_i(\boldsymbol{a},\boldsymbol{b},t)dt \\ S_i(\boldsymbol{a},\boldsymbol{b},t) &= \prod_d f_d(x_{d,i} - (a_d \cdot t + b_d), t) \end{aligned} \tag{Б.4}$$

Оптимизация Параметр t определен с точностью до линейного преобразования. Изменение масштаба t приведет к соответствующему компенсаторному изменению значений a_d. Смещение t также приведет к изменению параметров b_d. Поэтому надо ввести дополнительные ограничения. Поскольку линейное преобразование t имеет две степени свободы, то ограничений должно быть тоже 2, скажем

$$\sum_d (a_d)^2 = (\boldsymbol{a},\boldsymbol{a}) = 1; \quad \sum_d b_d = 0; \tag{Б.5}$$

Дальше оптимизацию можно вести методом неопределенных множителей Лагранжа.

Нормальный шум Пусть шум имеет нормальное распределение с постоянным стандартным отклонением σ_d и нулевым средним:

$$f_d(x) = \frac{1}{\sigma_d\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma_d^2}\right)$$

Тогда выражение $S_i(\boldsymbol{a},\boldsymbol{b},t)$ примет вид:

$$S_i(\boldsymbol{a},\boldsymbol{b},t) = Z \exp\left(-\frac{1}{2}\sum_d \frac{(x_{d,i} - a_d t - b_d)^2}{\sigma_d^2}\right) \tag{Б.6}$$

где

$$Z = \frac{1}{(2\pi)^{\frac{n}{2}} \prod_d \sigma_d}$$

Перемасштабируем все параметры на σ_d:

$$\begin{aligned} \alpha_d &= \frac{a_d}{\sigma_d};\ \ \beta_d = \frac{b_d}{\sigma_d} \\ z_{d,i} &= \frac{x_{d,i}}{\sigma_d};\ \ y_{d,i} = \frac{x_{d,i} - b_d}{\sigma_d} = z_{d,i} - \beta_d \end{aligned} \tag{Б.7}$$

В векторной форме

$$\begin{aligned} \boldsymbol{\alpha} &= (\boldsymbol{\sigma^{-1}})\boldsymbol{a};\ \ \boldsymbol{\beta} = (\boldsymbol{\sigma^{-1}})\boldsymbol{b} \\ \boldsymbol{z}_i &= (\boldsymbol{\sigma^{-1}})\boldsymbol{x}_i;\ \ \boldsymbol{y}_i = (\boldsymbol{\sigma^{-1}})(\boldsymbol{x}_i - \boldsymbol{b}) = \boldsymbol{z}_i - \boldsymbol{\beta} \end{aligned}$$

Где $(\boldsymbol{\sigma^{-1}})$ – диагональная матрица: $(\boldsymbol{\sigma^{-1}})_{cd} = \frac{1}{\sigma_d}\delta(cd)$ Получаем выражения для и S_i:

$$S_i(\boldsymbol{a},\boldsymbol{b},t) = Z \exp\left(-\frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{\alpha}t)^2\right) = Z \exp\left(Q_i(t)\right)$$

Чтобы удобно интегрировать, выражение под экпонентой будем рассматривать как квадратичную форму относительно t:

$$\begin{aligned} Q_i(t) &= -\frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{\alpha}t)^2 = -A \cdot t^2 + B \cdot t + C; \\ A &= \frac{1}{2}\boldsymbol{\alpha}^2;\ B = (\boldsymbol{y}_i,\boldsymbol{\alpha});\ C = -\frac{1}{2}\boldsymbol{y}_i^2 \end{aligned} \tag{Б.8}$$

Тогда вычислим интеграл:

$$\begin{aligned}\int\limits_{-\infty}^{+\infty} S_i(t)dt &= \int\limits_{-\infty}^{+\infty} Z\exp\left(Q_i(t)\right)dt \\ &= Z\int\limits_{-\infty}^{+\infty}\exp\left(-A\cdot t^2+B\cdot t+C\right)^2 dt \\ &= Z\sqrt{\frac{\pi}{A}}\exp\left(\frac{B^2}{4A}+C\right) \\ &= Z\sqrt{\frac{2\pi}{\boldsymbol{\alpha}_d^2}}\exp\left(\frac{1}{2}\left(\frac{(\boldsymbol{y}_i,\boldsymbol{\alpha})^2}{\boldsymbol{\alpha}^2}-\boldsymbol{y}_i^2\right)\right)\end{aligned} \tag{Б.9}$$

Окончательно для правдоподобия получаем:

$$\Pr(\boldsymbol{x}_i|\boldsymbol{a},\boldsymbol{b}) = Z^n\sqrt{\frac{2\pi}{\boldsymbol{\alpha}^2}}\prod_i\exp\left(\frac{1}{2}\left(\frac{(\boldsymbol{y}_i\boldsymbol{\alpha})^2}{\boldsymbol{\alpha}^2}-\boldsymbol{y}_i^2\right)\right) \tag{Б.10}$$

В качестве ограничений на параметры выберем:

$$\boldsymbol{\alpha}^2 = 1;\quad \sum_d \beta_d = 0 \tag{Б.11}$$

Для удобства оптимизации прологарифмируем (Б.10) с учетом ограничений (Б.11):

$$\begin{aligned}L &= \ln Z^n\sqrt{2\pi}+\frac{1}{2}\left(\sum_i(\boldsymbol{y}_i\boldsymbol{\alpha})^2-\sum_i\boldsymbol{y}_i^2\right) \\ &= G+\frac{1}{2}\left(\sum_i B_i^2-\sum_i\boldsymbol{y}_i^2\right)\end{aligned}$$

Неопределенные множители Лагранжа дают функцию для оптимизации:

$$\Phi(\alpha,\beta) = L-\lambda(\boldsymbol{\alpha}^2-1)-\mu(\sum_d\beta_d)$$

Производная по β_p:

$$\frac{\partial\Phi(\alpha,\beta)}{\partial\beta_p} = \sum_i B_i \frac{\partial B_i}{\partial\beta_p} - \mu = \alpha_p \left(\boldsymbol{\alpha}, \sum_i (\boldsymbol{z}_i - \boldsymbol{\beta})\right) - \mu$$
$$= \alpha_p \sum_{d,i} \alpha_d \cdot \left(\sum_i z_{d,i} - N \cdot \beta_d\right) - \mu = 0$$

Уравнение очевидно удовлетворяется при

$$\boldsymbol{\beta} = \frac{\sum_i \boldsymbol{z}_i}{N}; \quad \mu = 0; \quad \hat{\boldsymbol{b}} = \frac{\sum_i \boldsymbol{x}_i}{N} \tag{Б.12}$$

т.е. параметры b_d – просто средние значения компонент наблюдений.

Производная по α_p:

$$\frac{\partial\Phi(\alpha,\beta)}{\partial\alpha_p} = \sum_i B_i \frac{\partial B_i}{\partial\alpha_p} - 2\lambda\alpha_p = \sum_i y_{p,i} \sum_d y_{d,i}\alpha_d - 2\lambda\alpha_p = 0$$

Уравнение является задачей на собственные значения, а ее решение – собственный вектор матрицы M:

$$M\alpha = \lambda\alpha; \quad M_{pd} = \sum_i y_{p,i} y_{d,i} = Y \times Y^T$$

Для поиска минимума выбираем первый собственный вектор, т.е. первую главную компоненту в координатах, масштабированных на стандартные отклонения шумов.

Вернемся к исходным координатам. Обозначим

$$\boldsymbol{X} = \sum_i \boldsymbol{x}_i, \boldsymbol{X'} = \sum_i (\boldsymbol{x}_i - \overline{\boldsymbol{x}})$$

где $\overline{\boldsymbol{x}}$ – среднее значение $\boldsymbol{x}_i$. Тогда матрица M перепишется в виде

$$M = Y \times Y^T = ((\boldsymbol{\sigma^{-1}})\boldsymbol{X'} \times ((\boldsymbol{\sigma^{-1}})\boldsymbol{X'})^T = (\boldsymbol{\sigma^{-1}})\boldsymbol{X'} \times X'^T(\boldsymbol{\sigma^{-1}});$$

Получаем задачу на собственные значения

$$(\boldsymbol{\sigma^{-1}})X' \times X'^T(\boldsymbol{\sigma^{-1}})\boldsymbol{\alpha} = \lambda\boldsymbol{\alpha} \tag{Б.13}$$

Здесь использовано, что $(\boldsymbol{\sigma^{-1}})$ — диагональная матрица, поэтому $(\boldsymbol{\sigma^{-1}})^T = (\boldsymbol{\sigma^{-1}})$. Заметим, что выражение $(\boldsymbol{\sigma^{-1}})X'$ означает, что просто все исходные значения надо просто разделить на соответствующую σ_d. Решив задачу на собственные значения (Б.13), получим вектор $\boldsymbol{\alpha}$. Далее пересчитываем его в исходные масштабы (умножаем на соответствующую σ_d), окончательно получаем:

$$\hat{\boldsymbol{a}} = (\boldsymbol{\sigma})\boldsymbol{\alpha} \tag{Б.14}$$

Если наблюдаемые распределены в интервале [0,1] В этом случае возникают проблемы с сильным перекосом шума при крайних значениях наблюдаемых. Мы с помощью обратного логистического преобразования отобразить все на пространство $[-\infty, +\infty]$

$$x'_{d,i} = \ln \frac{x_{d,i}}{1 - x_{d,i}}; \quad t' = \ln \frac{t}{1-t}$$

В этом случае проблема краевых значений исчезает и можно предположить нормальность шума для преобразованных данных. Далее, в этом пространстве построить аппроксимацию, а затем вернуться к исходным координатам. В исходных координатах линейность сохранится только если $a_i = 1;\ b_i = 0$, однако при не слишком сильных отклонениях параметров линейность сохраняется.

Восстановление "истинных" значений $\hat{x}$ Пусть значения x_d порождены некоторым значением t

$$\boldsymbol{x} = \boldsymbol{a} \cdot t + \boldsymbol{b} + \boldsymbol{\xi} = \hat{\boldsymbol{x}} + \boldsymbol{\xi}$$

Мы хотим узнать, каким значением t они порождены, и какие "истинные" значения им соответствуют. Можно написать апостериорную вероятность для t:

$$\begin{aligned} \Pr(t|\boldsymbol{x}) \sim \Pr(t)\Pr(\boldsymbol{x}|t) &= \Pr(t)\prod_d f_d(x_d - \hat{x}_d) \\ &= \Pr(t)\prod_d f_d(x_d - a_d \cdot t - b_d) \end{aligned} \tag{Б.15}$$

где f_d – плотность распределения шума. Два замечания про уравнение (Б.15). Во-первых, это уравнение дает *распределение* для ожидаемых t, порождающих

(единственное) наблюдаемое значение $\hat{x}$. Во-вторых, нам необходимо априорное распределение для t. В Байесовской парадигме для оценки значения $\hat{t}$ применяют различные подходы:

$$\hat{t}_L = \arg\max_t \prod_d f_d(x_d - a_d \cdot t + b_d) \tag{Б.16}$$

$$\hat{t}_E = \frac{\int_t t \Pr(t) \prod_d f_d(x_d - a_d \cdot t + b_d) dt}{\int_t \Pr(t) \prod_d f_d(x_d - a_d \cdot t + b_d) dt} \tag{Б.17}$$

$$\hat{t}_{MAP} = \arg\max_t \Pr(t) \prod_d f_d(x_d - a_d \cdot t + b_d) \tag{Б.18}$$

- $\hat{t}_L$ (Б.16) – оценка максимального правдоподобия. Не зависит от априорного распределения $\Pr(t)$
- $\hat{t}_E$ (Б.17) – оценка математического ожидания для переменной t при условии наблюдения: $E(t|x_d)$. Зависит от априорного распределения $\Pr(t)$
- $\hat{t}_{MAP}$ (Б.18) – оценка максимальной апостериорной вероятности $\Pr(t|x_d)$. Зависит от априорного распределения $\Pr(t)$

Сделаем MAP-оценку. Пусть априорное распределение t – нормальное с математическим ожиданием 0 и ошибкой σ_t и шум тоже нормальный. Для данной точки апостериорная вероятность будет:

$$\begin{aligned}
\Pr(t|\boldsymbol{x}) &\sim \exp\left(-\frac{t^2}{2\sigma_t^2} - \frac{1}{2}(\boldsymbol{\sigma^{-1}}(\boldsymbol{x} - \boldsymbol{a} \cdot t - \boldsymbol{b}))^2\right) \\
&= \exp\left(-\frac{t^2}{2\sigma_t^2} - \frac{(\boldsymbol{y} - \boldsymbol{\alpha} \cdot t)^2}{2}\right) \\
&= \exp - \left(\frac{t^2}{2} \cdot \left(\frac{1}{\sigma_t^2} + \boldsymbol{\alpha}^2\right) - t \cdot (\boldsymbol{y}, \boldsymbol{\alpha}) + \frac{1}{2}\boldsymbol{y}^2\right)
\end{aligned} \tag{Б.19}$$

Принимая во внимание, что $\boldsymbol{\alpha}^2 = 1$, получаем оптимальное значение t:

$$\hat{t} = \frac{\boldsymbol{y}\boldsymbol{\alpha}}{1 + \frac{1}{\sigma_t}}$$

Полагая $\sigma_t \gg 1$, окончательно для $\hat{\boldsymbol{y}}$ получаем:

$$\hat{\boldsymbol{y}} = \boldsymbol{\alpha} \cdot \boldsymbol{y}\boldsymbol{\alpha}$$

Иными словами, оптимальный вектор $\hat{\boldsymbol{y}}$ является просто проекцией $\boldsymbol{y}$ на направление $\boldsymbol{\alpha}$. Эта оценка совпадает с L-оценкой (максимизация правдоподобия). Возвращаясь в исходные координаты, получим:

$$\begin{aligned} \hat{\boldsymbol{x}}_i &= \boldsymbol{a}(\boldsymbol{\sigma}^{-1}(\boldsymbol{x_i} - \overline{\boldsymbol{x}}), \boldsymbol{\sigma}^{-1}\boldsymbol{a}) + \boldsymbol{b} \\ \hat{t}_i &= (\boldsymbol{a}, \boldsymbol{y}_i) \end{aligned} \tag{Б.20}$$

Несколько замечаний про $\boldsymbol{\sigma}$

– Обычно параметры $\boldsymbol{\sigma}$ заранее не известны. Тогда возникает вопрос – как строить модель, ведь формулы для модели явно содержат ссылку на $\boldsymbol{\sigma}$. Если мы уверены, что по разным направлениям среднеквадратичные отклонения не сильно различаются, то можно просто построить модели в предположении $\sigma_d = 1$, поскольку для построения модели не важны сами значения $\boldsymbol{\sigma}$, а важно их соотношение. В этом легко убедиться, просто умножив все $\boldsymbol{\sigma}$ на положительную константу. По предсказанию $\hat{\boldsymbol{x}}_i$ можно сделать апостериорную оценку для $\boldsymbol{\sigma}$:

$$\hat{\boldsymbol{\sigma}}^2 = \frac{1}{n-2}\sum_i (\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i)^2 \tag{Б.21}$$

– Если мы положим одну или несколько $\sigma_k = 0$, то мы получим обычную (многомерную) регрессию. Направления, соответствующие нулевым σ отвечают предикторам. Правда, предложенные формулы не позволяют обнулять σ, но можно положить, например $\sigma_k = 10^{-7}$.

– Так как этот анализ полагается на пропорции значений $\boldsymbol{\sigma}$ для двух образцов, когда мы работает с аллельным дисбалансом, мы предлагаем использовать $\boldsymbol{\sigma}_1 = 1$ для одного из образцов, и вычислить относительное значение $\boldsymbol{\sigma}_2$ для второго образца используя среднюю разницу в покрытии и разницу в QCC:

$$\boldsymbol{\sigma}_2 = \frac{\mathrm{QCC}_2}{\mathrm{QCC}_1} \cdot \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} \frac{C_{i1}}{C_{i2}}},$$

где n обозначает количество генов в анализе, а C_{ij} — аллельное покрытие i-го гена в j-м образце.

Доверительные интервалы на коэффициенты регрессии Поскольку наблюдения независимы, то по каждому направлению мы имеем линейную регрессию $\hat{x}_{d,i} = a_d \cdot t_i + b_d$. Заменяя t_i на оценку $\hat{t}_i$, получим стандартную линейную регрессию с определенной независимой переменной (предиктором):

$$\hat{x}_{d,i} = a_d \cdot \hat{t}_i + b_d$$

В предположении нормального шума, можно использовать стандартные оценки для $\hat{\sigma}_{a_d}, \hat{\sigma}_{b_d}$:

$$\begin{aligned} \hat{\sigma}_a &= \frac{\hat{\sigma}}{\sum(\hat{t}_i - \bar{t})^2} \\ \hat{\sigma}_b &= \hat{\sigma}_a \sqrt{\frac{1}{n}\sum \hat{t}_i} \end{aligned} \tag{Б.22}$$

где $\hat{\sigma}_{b_d}$ оценки стандартного отклонения (Б.21).

Доверительные интервалы определим используя распределение Стьюдента с $n-2$ степенями свободы:

$$\begin{aligned} a &\in [a \pm q_{n-2}(\gamma) \cdot \hat{\sigma}_a] \\ b &\in [b \pm q_{n-2}(\gamma) \cdot \hat{\sigma}_b] \end{aligned} \tag{Б.23}$$

где γ — уровень доверительного интервала (вероятность); $q_{n-2}(\gamma)$ — γ-квантиль распределения Стьюдента.

Реализация алгоритма в R

Ниже мы реализуем алгоритм для двумерных данных на языке R.

Листинг Б.1 Листинг процедуры нахождения наилучшего линейного тренда и доверительной полосы

```
# 2D linear regression without predictors:
#
# x -- set of input vectors (rows)
# sigma -- vector of sigmas
#
# return -- list:
# a -- vector of coefficients 'a' in formula  x=a*t+b
```

```
#                                   (dim=number of rows in x)
# b -- vector of coefficients 'b' in formula  x=a*t+b
#                                   (dim=number of rows in x)
# xx -- predicted values            (matrix)
# sdx -- vector of stanard deviatios for (x-xx)
#                                   (dim=number of rows in x)
# t -- vector of predictor values 't'
#                                   (dim=number of observations=number of columns
    )
# sda -- vector of sd for the coefficient 'a'
# sdb -- vector of sd for the coefficient 'b'
#
linFit <- function(x, sigma=NULL){
  dm = dim(x)
  d  = dm[1]                        #============== number of dimensions
  n  = dm[2]                        #============== number of observations
  if(is.null(sigma)){
    sigma = rep(1, d)
  }
  b       = apply(x,1,mean)         #============== calc means
  y       = (x-b) / sigma           #============== rescale
  aa      = y %*% t(y)              #============== Matrix
  c       = eigen(aa)
  alpha   = c$vectors[,1]           #============== direction in y-space
  a       = alpha * sigma           #============== direction in native space
  t       = as.vector(alpha %*% y)  #============== projection in y-space
  xx      = b+a %o% t               #============== Predicted values

  sdx     = apply(x-xx, 1, sd)      #============== sd for difference input-
      predicted

  #============================== calculation sd for coefficients
  rs      = x - xx                  #============== residuals
  rs2     = rs * rs
  ssr     = apply(rs2, 1, sum)
  seps    = sqrt(ssr / (n-2))       #============== sd for residuals
  et      = mean(t)
  rt      = t - et                  #============== deviations of t

  da      = seps / sqrt(sum(rt*rt)) #============== sd for a
  db      = da * sqrt(sum(t*t) / n) #============== sd for b

  res = list(a = a, b = b, xx = xx, sdx = sdx, t = t, sda = da, sdb = db)
  return(res)
}

# Confidence band for linear regression without predictors (linFit):
#
```

```
linFit_compCI <- function(res, CI) {
  xx = res$xx    # xx -- predicted values (matrix)
  a  = res$a     # a -- vector of coefficients 'a' in formula  x=a*t+b
                 # (dim=number of rows in x); slope=a[2]/a[1]
  b  = res$b     # b -- vector of coefficients 'b' in formula  x=a*t+b
                 # (dim=number of rows in x)
  da = res$sda   # sda -- vector of sd for the coefficient 'a'
  db = res$sdb   # sdb -- vector of sd for the coefficient 'b'
  t  = res$t     # t -- vector of predictor values 't'
                 # (dim=number of observations=number of columns)
  n  = length(t)

  if(CI > 0){
    # t-test for conf. interval:
    qq = qt(CI, n-1)
    # delta a and delta b:
    da1 = qq * da[1]
    da2 = qq * da[2]
    db1 = qq * db[1]
    db2 = qq * db[2]
    # central corners:
    x0 = b[1] - db1
    y0 = b[2] + db2
    x1 = b[1] + db1
    y1 = b[2] - db2
    # polygon points:
    xp = c(x0)
    yp = c(y0)
    tm = max(t)
    xp = c(xp, x0 + (a[1]-da1) * tm)
    yp = c(yp, y0 + (a[2]+da2) * tm)
    xp = c(xp, x1 + (a[1]+da1) * tm)
    yp = c(yp, y1 + (a[2]-da2) * tm)
    xp = c(xp, x1)
    yp = c(yp, y1)
    tm = min(t)
    xp = c(xp, x1 + (a[1]-da1) * tm)
    yp = c(yp, y1 + (a[2]+da2) * tm)
    xp = c(xp, x0 + (a[1]+da1) * tm)
    yp = c(yp, y0 + (a[2]-da2) * tm)
    xp = c(xp, x0)
    yp = c(yp, y0)
  }

  dfCI = data.frame(x = xp, y = yp)
  return(dfCI)
}
```

Б.2 Сопроводительные рисунки к главе 3

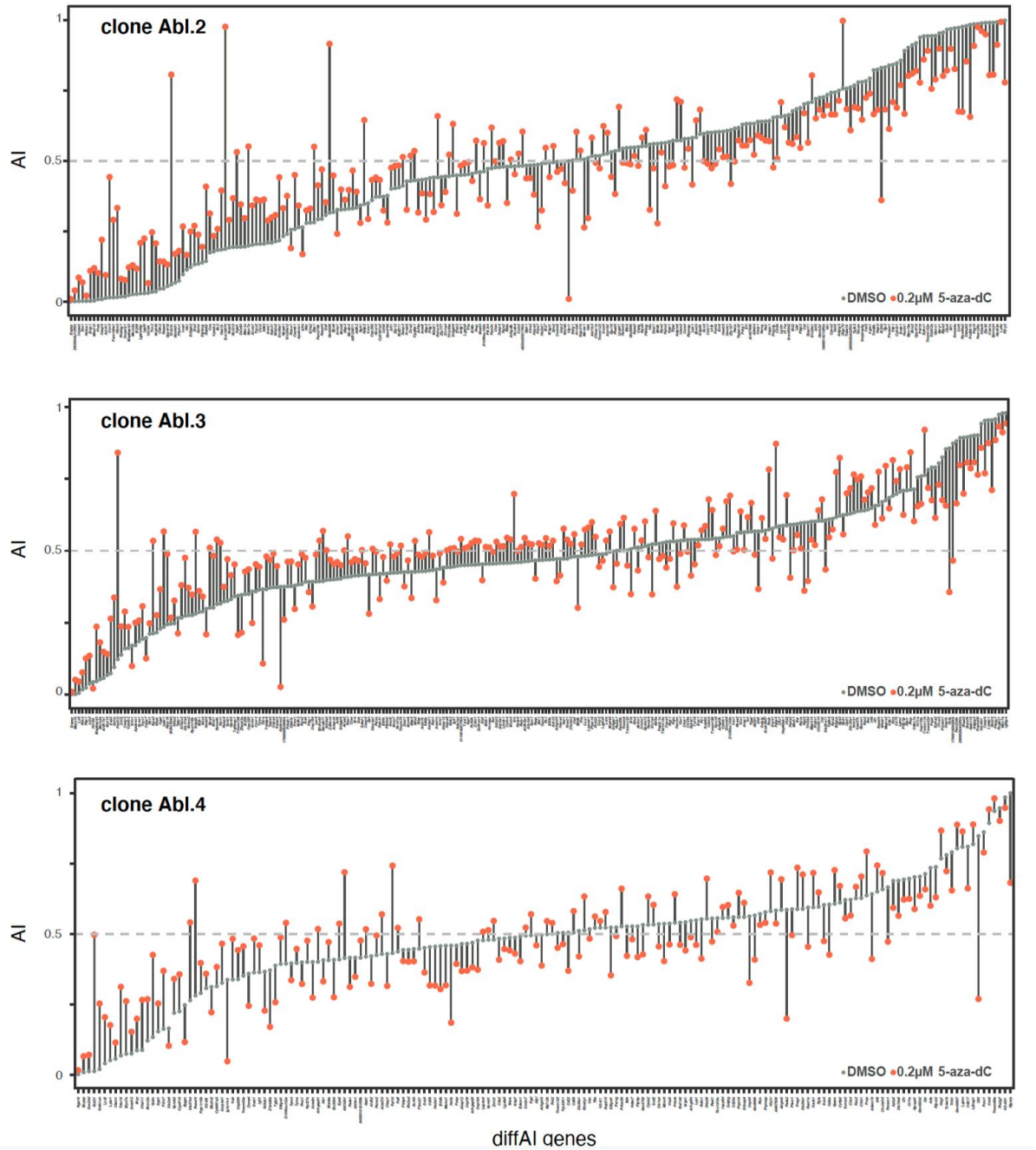

Рисунок Б.3 — Аутосомные гены, имеющие существенный дифференциальный аллельный дисбаланс (diffAI) между образцами, обработанными DMSO и 0.2μM 5-aza-dC, в клонах Abl.2, Abl.3 и Abl.4.

Серые кружки отображают AI в контрольном образце, а красные – после обработки 5-aza-dC. Гены отсортированы по уровню AI в контроле (1% DMSO).

Приложение B

К главе 4, «Внешние РНК-контроли позволяют проводить точный аллель-специфический анализ экспрессии на большом количестве образцов»

B.1 Сопроводительные заметки к главе 4

B.1.1 Расширенные методы

Расширенное бета-биномиальное распределение

Мы должны проверить, является ли определённое нами расширенное бета-биномиальное распределение корректно определённым вероятностным распределением с суммой вероятностей 1. Любой параметр $1 < Q < n$ определяет корректное вероятностное распределение, бета-биномиальное распределение. Так как для любых n,m,p формула, определяющая $eBB(m \mid n = m + p, \mathrm{AI}, Q)$, является аналитической функцией от AI и Q, то сумма всех вероятностей этого распределения будет равна 1.

Также мы можем напрямую проверить, какие значения Q гарантируют, что $eBB(m \mid n, \mathrm{AI}, Q) \geq 0$ для всех $0 \leq m \leq n$. В терминах α, β и d это эквивалентно следующим неравенствам:

$$\alpha + (n-1)d \geq 0,$$
$$\beta + (n-1)d \geq 0,$$
$$\alpha + \beta + (n-1)d > 0.$$

В терминах модели урны Полья, неравенства соответствуют необходимости взять шар одного из типов большее число раз, чем того позволяет существующее количество шаров. В урне, соответствующей гипергеометрическому распределению, такой проблемы нет: любому такому сценарию соответствует

вероятность 0. В расширенном бета-биномиальном распределении формула может не давать 0, так как α и β могут не быть кратны d.

Мы можем применить небольшую модификацию к распределению для того, чтобы лучше имитировать поведение гипергеометрического распределения. Возьмём отрицательное d такое, что неравенство $\alpha + \beta + (n-1)d > 0$ всё ещё выполнено. Тогда формулы $eBB(m \mid n, \alpha, \beta, d)$ должны давать хотя бы одно положительное значение на отрезке $0 \leq m \leq n$ (можно показать, что для m, наиближайшего к $\frac{n\alpha}{\alpha+\beta}$, значение точно будет положительным). С другой стороны, формулы могут дать отрицательное значение для m ближе к 0 или n, если одно из неравенств $\alpha + (m-1)d > 0$ и $\beta + (n-m-1)d > 0$ перестанет выполняться. Тогда мы можем заменить значения распределения в этих точках на 0. Напоследок, поделим все значения на константу таким образом, чтобы сумма всех значений снова стала равна 1.

Таким образом, мы можем расширить область определения распределения вплоть до точки $d = -\frac{\alpha+\beta}{n-1}$. В терминах Q, теперь распределение определено вплоть до $Q > -\frac{1}{n-2}$. Отрицательные значения Q, на первый взгляд, противоречат первоначальному смыслу Q — значению избыточной дисперсии. Это видимое противоречие разрешается просто. В тот момент, когда мы переопределили некоторые значения распределения на ноль, Q перестаёт быть действительной мерой избыточной дисперсии.

Программное вычисление функции вероятности расширенного бета-биномиального распределения

Здесь мы явно опишем сочетание множителей числителя и знаменателя на пары. Без ограничений общности, положим, что $\alpha \geq \beta$. Тогда сочетание

имеет следующий вид:

$$eBB(m \mid n, \alpha, \beta, d) = \binom{n}{m} \frac{\prod_{i=0}^{m-1}(\alpha + id)\prod_{j=0}^{p-1}(\beta + jd)}{\prod_{k=0}^{n-1}(\alpha+\beta+kd)} =$$
$$= \frac{\prod_{i=0}^{m-1}\frac{\alpha+id}{i+1}\prod_{j=0}^{p-1}\frac{\beta+jd}{j+1}}{\prod_{k=0}^{n-1}\frac{\alpha+\beta+kd}{k+1}} = \frac{\prod_{i=0}^{m-1}\frac{\alpha+id}{i+1}}{\prod_{k=0}^{m-1}\frac{\alpha+\beta+kd}{k+1}} \cdot \frac{\prod_{j=0}^{p-1}\frac{\beta+jd}{j+1}}{\prod_{k=m}^{n-1}\frac{\alpha+\beta+kd}{k+1}} =$$
$$= \prod_{i=0}^{m-1}\frac{\alpha+id}{\alpha+\beta+id} \cdot \prod_{j=0}^{p-1}\frac{m+j+1}{j+1}\frac{\beta+jd}{\alpha+\beta+(m+j)d}.$$

Теперь выражение является произведением n дробей, где абсолютное значение каждой из них значительно ближе к 1.

Чтобы вычислить производные этого выражения, не требуется специальных трюков. Финальный ответ выглядит так:

$$\frac{d\log eBB(m \mid n, \mathrm{AI}, Q)}{dQ} =$$
$$= \sum_{i=0}^{m-1}\frac{(1-\mathrm{AI})(n-1)i}{\big((i-\mathrm{AI})Q + (\mathrm{AI}n - i)\big)\big((i-1)Q + (n-1)\big)} +$$
$$+ \sum_{j=0}^{p-1}\frac{(n-1)\big(\mathrm{AI}j - (1-\mathrm{AI})m\big)}{\big((j+\mathrm{AI}-1)Q + ((1-\mathrm{AI})n - j)\big)\big((m+j-1)Q + (p-j)\big)}$$

Вариации градиентного спуска для поиска оптимального Q

Наиболее важный аспект для вычисления функции правдоподобия — это то, что функция определена только от некоторого отрицательного $Q = Q_{min}$ до $Q = Q_{max} = \min_l m_l + p_l$.

Для начала, мы должны выбрать способ перевода производной в значение шага градиентного спуска. Для этого до начала градиентного спуска мы оцениваем возможный диапазон значений производной через её вычисление в 5 точках, распределённым по отрезку определения. Мы берём медиану 5 точек как оценку медианного значения производной на отрезке. Далее в процедуре мы нормализуем каждое значение производной на эту медиану.

Чтобы уменьшить вероятность прыжков влево и вправо вокруг локального максимума, и вероятность медленного прохождения "плато" с маленьким абсолютным значением градиента, мы вводим переменную `streak`. Значение каждого шага умножается на $1.1^{\texttt{streak}}$. Когда производная в точке сонаправлена предыдущему шагу, мы увеличиваем `streak` на один. Когда производная противонаправлена предыдущему шагу, мы уменьшаем `streak` на один. Также мы приравниваем `streak` к нулю перед тем, как тренд `streak` изменяется.

Две граничные точки отрезка определения функции отличаются по своим свойствам. Граница Q_{min} возникает, так как предел $\log L(Q)$ равен $-\infty$ когда Q стремится к Q_{min}. Из-за этого производная функции правдоподобности также становится всё более отрицательной в точках, близких в Q_{min}. Мы могли бы предотвратить это при помощи замены переменной, однако функционально эквивалентное поведение можно достичь просто путём дополнительного домножения каждого шага спуска на $Q - Q_{min}$.

Верхняя граница Q_{max} имеет другую суть. Пределы $\log L(Q)$ и $\frac{d \log L(Q)}{dQ}$ существуют и являются конечными когда Q стремится к Q_{max}. Глобальный максимум функции правдоподобности может быть достигнут в точке $Q = Q_{max}$. В таком случае, мы принимаем решение исходя из биоинформатической сути задачи. Мы фильтруем все гены с таким покрытием, и повторяем процедуру снова. Если в какой-то момент генов не останется, мы возвращаем $Q = +\infty$.

B.2 Сопроводительные рисунки к главе 4

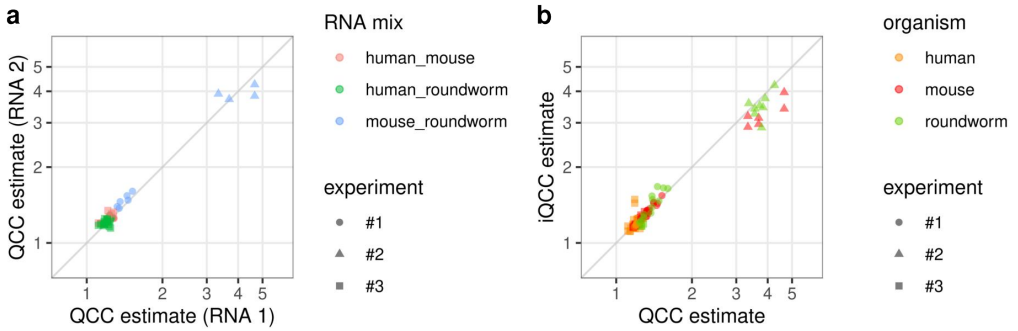

Рисунок B.1 — Независимо полученные оценки QCC и iQCC хорошо коррелируют.

(a) Корреляция оценок QCC для двух компонент смесей РНК (коэффициент корреляции Пирсона = 0.98). **(b)** Корреляция оценок iQCC и QCC (посчитанных с помощью `Qllelic`, коэффициент корреляции Пирсона = 0.98). **(a-b)** Данные, использованные для этого рисунка: все смешанные образцы из экспериментов 1-3. Коэффициенты QCC были посчитаны для всех пар реплик с разницей в общем аллельном покрытии $\leq 15\%$, и итоговое значение для каждой реплики было определено как геометрическое среднее между всеми оценками, включавшими эту реплику при подсчёте.

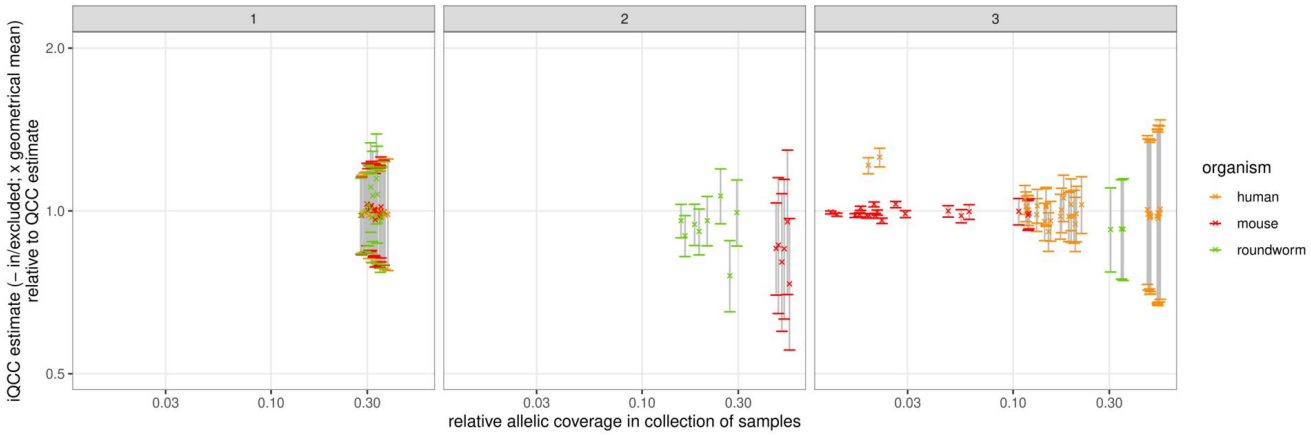

Рисунок B.2 — Сходимость верхней и нижней оценок iQCC и их геометрического среднего к оценке QCC с уменьшением относительного общего аллельного покрытия.

Значения QCC посчитаны с помощью `Qllelic` на смешанных образцах из экспериментов 1-3 (данные те же, что и на Рис.B.1) .

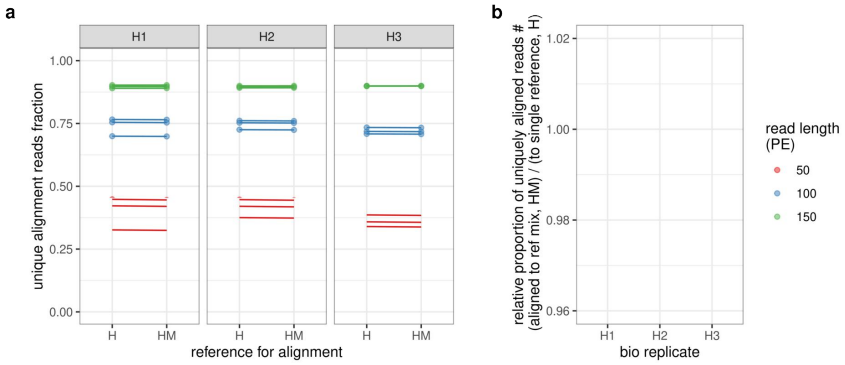

Рисунок В.3 — Зависимость доли уникальных выравниваний от длины прочтений и референсного генома.

(a) Доля уникально выровненных с помощью `STAR` прочтений. Большая часть потерь происходит за счёт коротких выравниваний: до 6%, 25% и 60%, соответственно. В то же время, % прочтений, выровненных в слишком большое количество мест, составляет максимум 7% для любого из образцов или длины прочтения. **(b)** Отношение количества картированных прочтений между выравниваниями на смешанный (человек и мышь) и на человеческий референсный геном. **(a-b)** Для этого рисунка были использованы человеческие образцы из эксперимента 1. Парные прочтения длиной 150×2 были подрезаны до 100×2 и 50×2 с помощью `trimmomatic HEADCROP`, и далее обработаны обычным образом.