

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ
ФЕДЕРАЦИИ

ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ОБРАЗОВАНИЯ
"МОСКОВСКИЙ ФИЗИКО-ТЕХНИЧЕСКИЙ ИНСТИТУТ
(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ИНСТИТУТ)"

ФИЗТЕХ-ШКОЛА ПРИКЛАДНОЙ МАТЕМАТИКИ И
ИНФОРМАТИКИ
КАФЕДРА ДИСКРЕТНОЙ МАТЕМАТИКИ

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА ПО
НАПРАВЛЕНИЮ 01.03.02
ПРИКЛАДНАЯ МАТЕМАТИКА И ИНФОРМАТИКА
НА ТЕМУ:
**Вариационный байесовский вывод в графических моделях в
задаче восстановления клональной структуры опухоли**

Студент _____ Иванов В.В.
Научный руководитель к.ф-м.н. _____ Юнхуа Хуань.
Зав. кафедрой д.ф-м.н., профессор _____ Райгородский А.М.

МОСКВА, 2020

1 Аннотация

В данной дипломной работе предложена графическая байесовская модель машинного обучения XClone. Её задача — восстановление клонального состава опухоли по данным ДНК-секвенирования одиночных клеток. XClone — статистическая модель опухолевого образца, оптимальные параметры которой подбираются посредством вариационного байесовского вывода. По этим параметрам можно восстановить структурные мутации на каждой из хромосом в каждой из клеток образца, что позволяет проследить эволюцию опухоли.

В работе описана формальная постановка задачи и приведена реализация алгоритма на языке программирования Python. Актуальность задачи подтверждается тем, что в последние годы несколько статей схожей тематики — SiCloneFit[1], InferCNV, CaSpER[2], CHISEL[3] — было опубликовано в высокоимпактных научных журналах, но среди них не было полных аналогов. Практическую ценность работы подтверждает то, что задача пришла из клинической практики немецких врачей-онкологов. Научная новизна работы заключается в том, что, несмотря на популярность темы, у XClone пока есть всего один прямой конкурент — алгоритм CHISEL, — от которого XClone выгодно отличается как производительностью, так и тем, что допускает естественное обобщение на несколько модальностей, каждая из которых уточняет диагноз: scRNA-seq, scATAC-seq, митохондриальные геномы, соматические мутации.

Ведётся активная работа над тем, чтобы сделать XClone одним из первых алгоритмов поиска аллель-зависимых структурных вариаций в данных РНК-секвенирования одиночных клеток. С этой целью намечена коллаборация с учёными из университета Гонконга и из Стенфорда с аprobацией алгоритма XClone на их экспериментальных данных.

Содержание

1	Аннотация	1
2	Обозначения, сокращения, основные определения	4
3	Введение	13
4	Материалы и методы	18
4.1	Вероятностная модель ДНК-секвенирования	18
4.2	Алгоритмы предобработки данных	20
4.2.1	Извлечение данных из ВАМ-файлов	21
4.2.2	Статистическое фазирование гаплотипов	22
4.2.3	Подходы к сегментации генома	23
4.2.4	Исправление ошибок смены цепи	26
4.3	XClone-V1: только BAF-модуль	33
4.3.1	Структура модели	34
4.3.1.1	Параметры:	34
4.3.1.2	Визуализация:	37
4.3.1.3	Статистическая модель	38
4.3.2	Поиск оптимальных параметров	40
4.3.3	Поиск наиболее вероятной перестановки меток . . .	43
4.4	XClone-V2: BAF- и RDR-модули	46
4.4.1	Структура BAF-модуля	47
4.4.2	Структура RDR-модуля	48
4.4.3	Генеративная модель данных	49
4.4.4	Вариационный байесовский вывод	50
4.4.4.1	Общее описание метода	50
4.4.4.2	VB в алгоритме XClone	55
4.4.4.3	Вывод ELBO для BAF-модуля	56

4.4.4.4	Вывод ELBO для RDR-модуля	59
4.4.5	Известные недостатки и планы по их исправлению .	60
4.4.5.1	Избыточная сложность исходной модели .	60
4.4.5.2	Численное интегрирование в оценке обос- нованности актуальной версии RDR-модуля	66
4.4.5.3	Слишком большая разница масштабов от- дельных слагаемых в ELBO	67
4.4.5.4	Концептуальная невозможность детектиро- вания WGD	69
4.4.5.5	Отсутствие масштабирования клеток в RDR- модуле	72
4.4.5.6	Необходимость вручную задавать ожидае- мое число клональных линий в образце .	73
4.4.5.7	Отсутствие явной связи с деревом онкогенеза	73
4.5	Использованные данные	73
5	Полученные результаты	76
5.1	Синтетические данные	76
5.2	Реальные данные: STP, scDNA-seq	80
5.3	Реальные данные: STP, scRNA-seq	90
6	Заключение. План дальнейших исследований	97
7	Благодарности	101
Список использованных источников		103

2 Обозначения, сокращения, основные определения

В силу междисциплинарного характера данной дипломной работы, автор считал уместным определить все понятия из биологии, без которых понимание работы будет затруднено или невозможно, не вдаваясь по возможности в технические детали. Для терминов, не имеющих устоявшегося перевода на русский язык, были использованы принятые в научном сообществе транслитерации.

Определение 2.1 (Центральная догма молекулярной биологии).

Наблюдаемая в природе закономерность передачи генетической информации: она распространяется от нуклеиновых кислот к белкам, вначале от ДНК к РНК в процессе **транскрипции**, а затем от РНК к белкам в процессе **трансляции**. Правило было впервые сформулировано Фрэнсисом Криком в 1958 году.

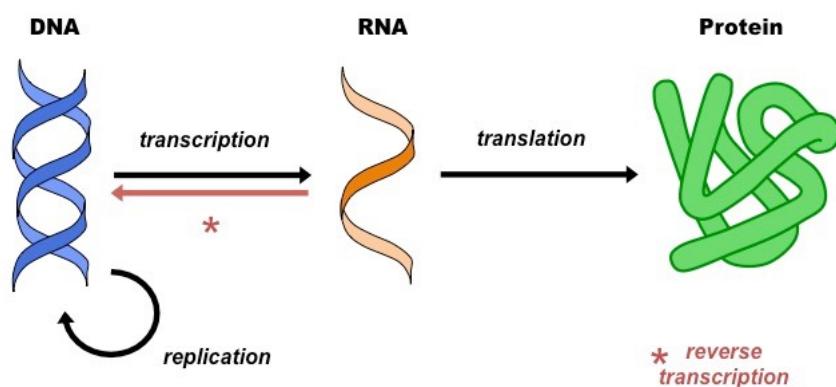


Рис. 2.1: Центральная догма молекулярной биологии

В упрощённом понимании, в процессе транскрипции участок ДНК преобразуется в т.н. **пре-матричную РНК (пре-мРНК)**, которая после **сплайсинга** — вырезания **инtronов**, участков, не кодирующих белковые последовательности, — превращается в **матричную РНК (мРНК)**,

которая транслируется в белковую последовательность в рибосомах.

Определение 2.2 (*Геном, экзом, транскриптом*).

1. **Геном** — генетический материал клетки, нуклеотидная последовательность ДНК организма.
2. **Экзом** — набор **экзонов** организма — участков генома, кодирующих белковые последовательности.
3. **Транскриптом** — совокупность всех транскриптов, синтезируемых одной клеткой или группой клеток, включая мРНК и некодирующие РНК. Представляет собой ту часть экзома, которая преобразуется в белки в момент наблюдения, и зависит от типа клетки, стадии клеточного цикла, условий внешней среды и т.д.

Определение 2.3 (*Референсный геном*). Секвенированный, собранный и проаннотированный консенсусный геном организма того же вида, к которому относится анализируемый образец.

Определение 2.4 (*Секвенирование*). Процесс определения первичной последовательности нуклеиновых кислот в клетке — ДНК или РНК. Прибор, осуществляющий секвенирование, называют **секвенатором**.

Определение 2.5 (*Прочтение (рид)*). Короткий нуклеотидный фрагмент, распознанный секвенатором после ПЦР. В научном сообществе чаще используется транслитерация **рид** от английского *sequencing read*. Набор ридов, извлечённых из образца, является основным конечным продуктом секвенирования.

Определение 2.6 (*NGS*). Next Generation Sequencing — общее название современных методов секвенирования, позволяющих, в отличие от исторических предшественников, получать полный геном, экзом или транскриптом в ходе одного эксперимента.

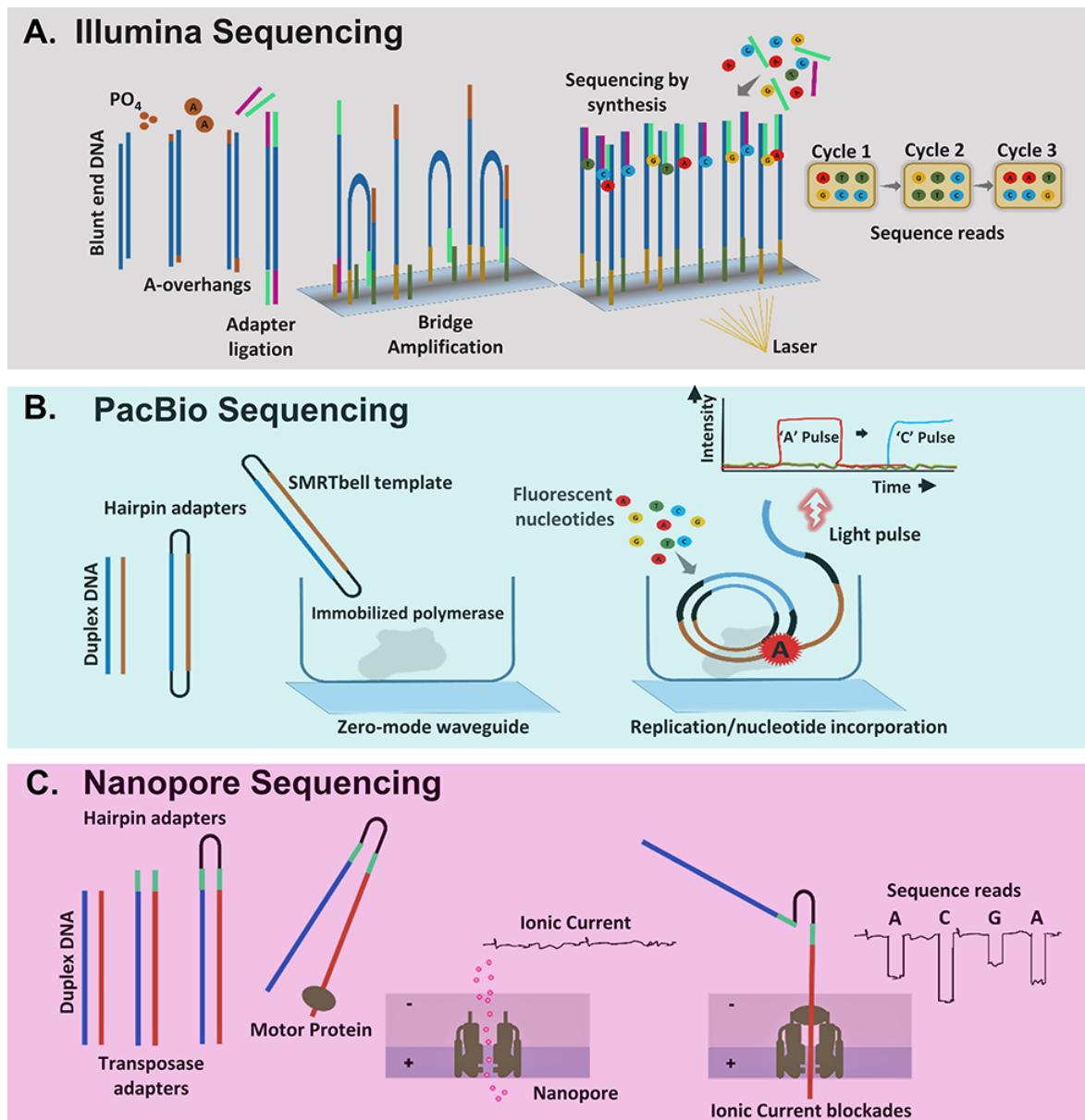


Рис. 2.2: Основные технологии высокопроизводительного секвенирования [23]. В данной работе использованы данные, полученные по протоколам Illumina (1) и Oxford Nanopore (3).

Несмотря на то, что для обработки данных секвенирования не нужно понимать все тонкости работы секвенаторов, важно понимать, как эти технологии соотносятся. Длина ридов, полученных по технологии Oxford Nanopore, заметно больше — десятки кб против 50-300 б у Illumina, — что компенсируется меньшей пропускной способностью и значительно

более высокой ценой. Кроме того, нужно понимать, что при подготовке ДНК к эксперименту последняя термически или химически дробится на малые фрагменты, которые затем амплифицируются — многократно копируются при помощи специальных ферментов. То, насколько хорошо расщепляются отдельные участки ДНК, зависит от их нуклеотидного состава, и это нужно учитывать при обработке результатов.

Определение 2.7 (*Bulk-секвенирование*). Разновидность секвенирования, при которой генетический материал предварительно выделяется из клеток ткани посредством их разрушения. Сигнал отдельных клеток при этом агрегируется, полученные данные отражают состояние ткани "в среднем".

Определение 2.8 (*Секвенирование одиночных клеток*). Совокупность новых методов секвенирования, позволяющих извлекать нуклеотидные последовательности из каждой из клеток образца в отдельности.

Определение 2.9 (*Матрица прочтений*). При фиксированной сегментации генома, матрица прочтений представляет собой целочисленную матрицу, где для каждой клетки подсчитано, сколько рядов выравниваются на последовательности в пределах каждого интервала из сегментации.

Определение 2.10 (*10X Genomics*). На момент написания данного текста, основной производитель технологий и программного обеспечения в нише секвенирования одиночных клеток. Представленная в работе статистическая модель проектировалась совместимой с ПО и форматом данных от 10X Genomics.

Определение 2.11 (*ОНП (снип)*). Однонуклеотидный полиморфизм — позиция в геноме, на которой в статистически значимыхолях популяции встречаются несколько различных вариантов нуклеотидов. Могут

существенно влиять на фенотип, в том числе быть причиной патологий. В сообществе более принята траслитерация **снп** от английского *SNP — single nucleotide polymorphism*.

Определение 2.12 (*Гетеро- и гомозиготные ОНП*). ОНП называется **гомозиготным**, если в родительских хромосомных наборах на соответствующей позиции находится один и тот же нуклеотид, и **гетерозиготным** в противном случае.

Определение 2.13 (*Гаплотипирование ОНП*). Общее название набора методов для определения **гаплотипов** в геноме — непрерывных участков ДНК, содержащих полиморфизмы, обычно наследуемые вместе. В англоязычной литературе это называют *SNP phasing*.

Определение 2.14 (*Ген*). Последовательность ДНК, составляющие сегменты которой не обязательно должны быть физически смежными. Эта последовательность ДНК содержит информацию об одном или нескольких продуктах в виде белка или РНК. Продукты гена функционируют в составе генетических регуляторных сетей, результат работы которых реализуется на уровне фенотипа.

Определение 2.15 (*Аллель*). Вариант фрагмента ДНК, встречающийся в статистически значимой доле популяции. Частные случаи — ОНП, варианты генов.

Определение 2.16 (*Аллельный дисбаланс*). Ситуация, когда один из аллелей доминирует над остальными — например, экспрессируется сильнее, более представлен в данных секвенирования и т.д.

Определение 2.17 (*CNV — структурные вариации генома*). *CNV — copy number variation* — масштабные структурные модификации генома, такие как:

- **Loss events:**

Делеция — удаление фрагмента;

- **Gain events:**

Дупликация — удвоение фрагмента (может происходить более одного раза и порождать больше двух копий);

Удвоение генома — удвоение числа хромосомных наборов;

- **Инверсия** — обращение непрерывного подотрезка;

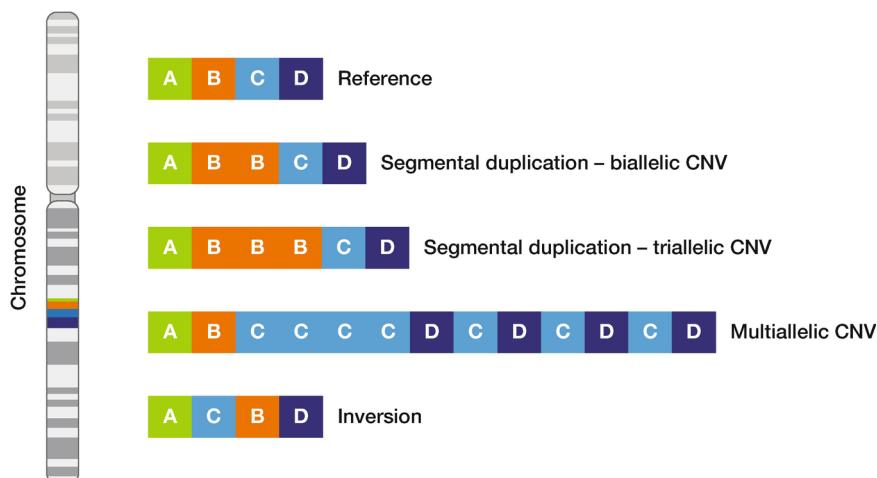
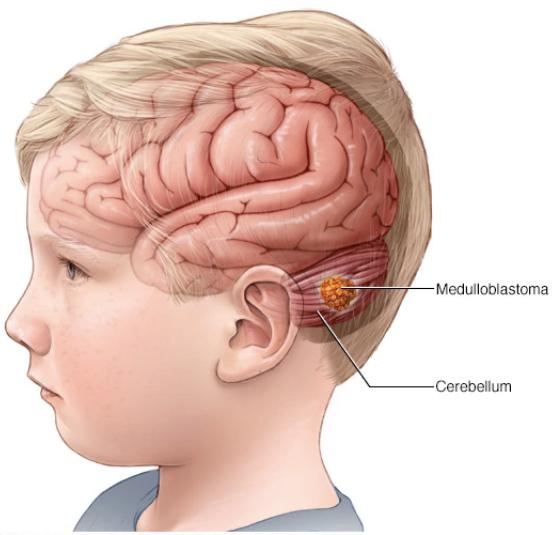


Рис. 2.3: Основные типы структурных вариаций

Разновидности структурных вариантов на этом не исчерпываются, но в контексте данной работы наибольший интерес представляет число копий крупных фрагментов генома. Такого рода структурные вариации характерны для опухолевых клеток.

Определение 2.18 (*ASCNV – аллель-специфические CNV*). CNV, для которых известна аллельная информация. В контексте данной работы обычно имеются в виду пары вида (c_m, c_p) , где c_m это число копий некоторого участка генома на материнской копии хромосомы, а c_p , соответственно, на отцовской.

Определение 2.19 (Медуллобластома). Самый распространённый тип педиатрической опухоли мозга. Поражает мозжечок.



© MAYO FOUNDATION FOR MEDICAL EDUCATION AND RESEARCH. ALL RIGHTS RESERVED.

Определение 2.20 (Хромотрипсис). Мутационный процесс, в ходе которого тысячи локальных структурных вариаций случаются в небольших фрагментах генома, локализованных в одной или нескольких хромосомах. Играет важную роль в онкогенезе в отдельных типах рака и в появлении некоторых врождённых заболеваний.

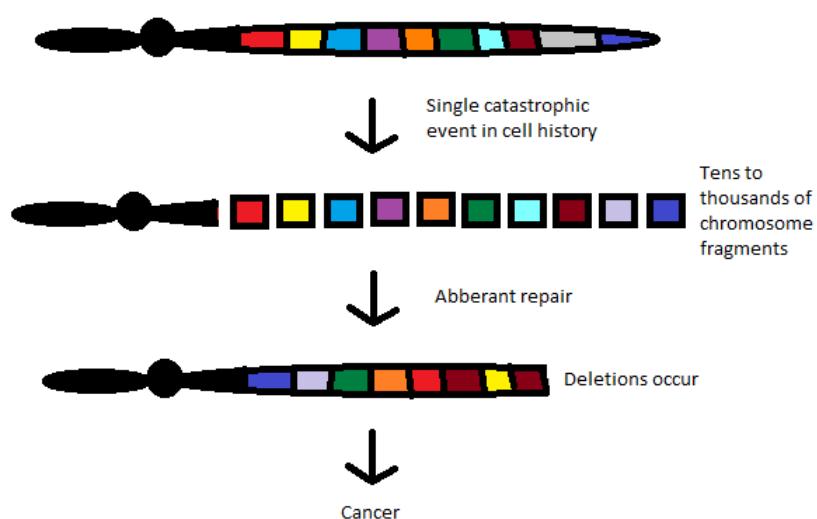


Рис. 2.4: Хромотрипсис

Определение 2.21 (Клональная линия). Класс эквивалентности клеток по отношению схожести генного материала в них. Конкретное определение этой "схожести" зависит от контекста. Для того, чтобы определить понятие клональной линии в онкологии, вначале нужно дать определение **клонального события** — масштабной наследуемой мутации. К клональным событиям относят, к примеру, крупные структурные вариации, хромотрипсис, дупликацию генома, а также короткие нуклеотидные замены, которые качественно меняют фенотип клетки.

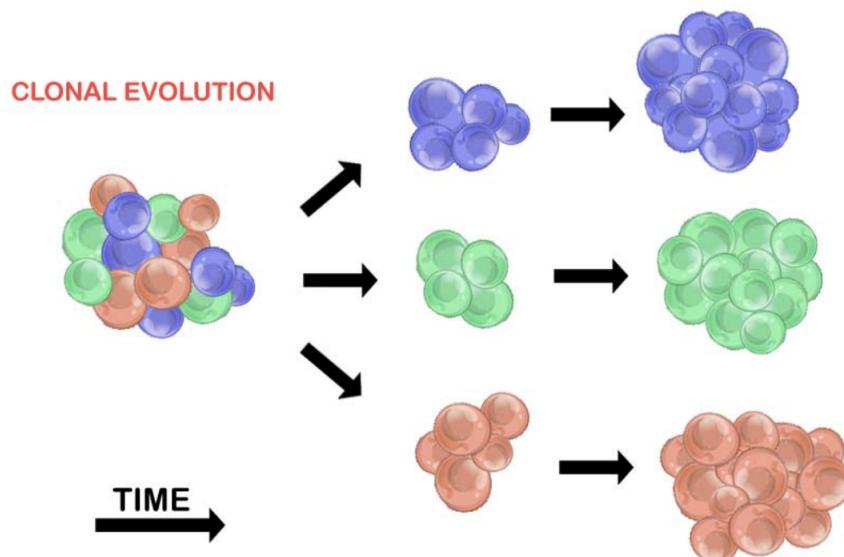


Рис. 2.5: Клональная эволюция опухоли: клетки первичной опухоли порождают т.н клональные линии. Клетки в пределах линии генетически однородны, но клетки двух любых различных линий существенно различны.

Клональные события позволяют задать псевдовремя на стадиях онкогенеза. Именно **псевдовремя**, т.к. клональное событие может произойти только в одной из двух произвольно выбранных генетически идентичных клеток, в связи с чем их потомки будут качественно отличаться друг от друга. Эту концепцию можно визуализировать посредством де-

рева онкогенеза, в корне которого находятся здоровые клетки, а каждое ветвление соответствует клональному событию. Тогда клональные линии можно определить как классы эквивалентности клеток, которые возникнут естественным образом, если геному каждой из них сопоставить вершину дерева онкогенеза.

Определение 2.22 (BAF). BAF – B-Allele Frequency – нормализованная мера дисбаланса аллелей числа ридов, выравнивающихся на гетерозиготные аллели A и B. BAF, равный 0 или 1, означает полное отсутствие одного из двух аллелей, т.е. фенотип AA или BB, а BAF означает сбалансированное присутствие обоих аллелей, т.е. фенотип AB или BA.

Определение 2.23 (RDR). Отношение числа ридов, однозначно выравнивающихся на конкретный участок генома, к ожидаемой глубине покрытия этого участка. Используется для определения числа loss и gain events в заданных сегментах генома.

3 Введение

Секвенирование одиночных клеток на протяжении последних нескольких лет было одной из самых горячих тем в науке: single-cell технологии дважды получали звание "method of the year" по версии Nature Methods[24, 25], а 10X Genomics, основная компания-производитель оборудования для single-cell секвенирования, заняла 69 строчку в рейтинге 500 наиболее быстрорастущих компаний США¹. Возможность изучать биологические процессы в тканях на уровне отдельных клеток привела к прорывным открытиям во многих областях науки[26, 27], особенно в персонализированной онкологии². В частности, в задаче восстановления клональной структуры опухоли — определения групп клеток, имеющих схожий набор индуцированных генетических мутаций. Понимать состав опухоли критически важно для подбора лечения, особенно в высокоинвазивных раках с высокой частотой мутаций.

¹<https://www.inc.com/inc5000/2019/top-private-companies-2019-inc5000.html>

²<https://www.nature.com/articles/d42473-019-00310-5>

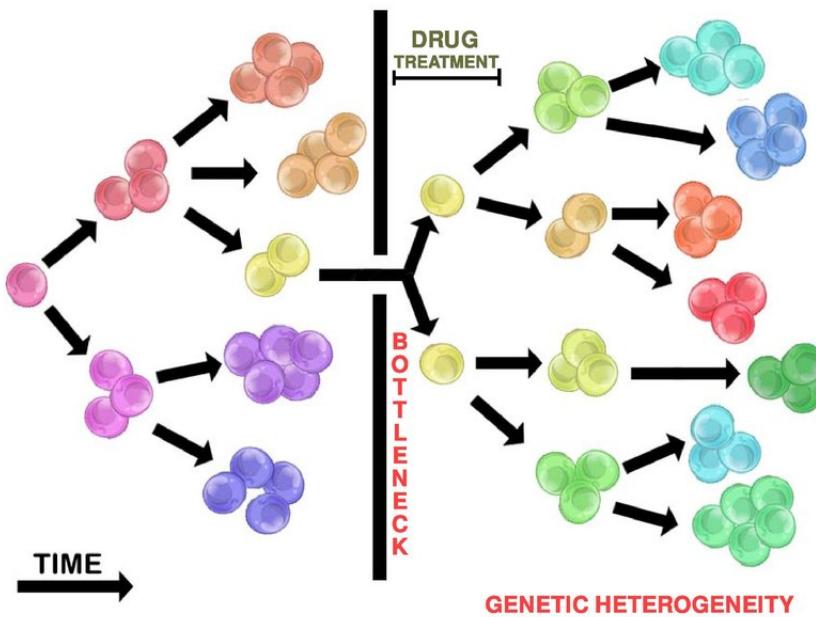


Рис. 3.1: При подборе терапии нужно учитывать клональный состав опухоли. Лечение может убивать некоторые клональные линии, но не действовать на остальные, из-за чего случается рецидив.

В данной дипломной работе рассматривается байесовский подход к этой задаче, которую называют одной из 11 главных задач вычислительной биологии одиночных клеток[41]. В работе предложены две версии алгоритма байесовского машинного обучения XClone — графической модели опухоли. В задачи XClone входит не только анализ клонального состава образца, но и поиск аллель-специфических структурных мутаций, по которым можно проанализировать эволюцию опухоли. Алгоритм XClone развивает идеи, заложенные доктором Янхуа Хуань в статьях "Cardelino"[10] и "Vireo"[15].

Текущая версия XClone состоит из двух основных частей, связанных расстановкой клональных меток.

Первая часть получила название "RDR-модуль где RDR расшифровывается как "read depth ratio отношение наблюдаемой доли числа прочтений

ний внутри фиксированного сегмента генома к ожидаемой. В теоретических моделях ДНК-секвенирования, где предполагается равномерное распределение прочтений по геному, RDR должно стремиться к реальному числу копий этого сегмента пополам. Это основная величина, на которую опираются алгоритмы поиска структурных вариаций генома, в том числе CellRanger от 10X Genomics и CHISEL[3]. В процессе работы над текстом ВКР стало ясно, что изначальная модель RDR-модуля не совсем точна. Это позволило разработать третью версию XClone, о которой идёт речь в заключительном разделе ВКР. Она будет реализована в ходе дальнейшей научной работы.

Вторая часть носит название "BAF-модуль где BAF означает "B-allele frequency". Эта часть модели позволяет не только понять, сколько копий сегмента образовалось в процессе онкогенеза, но и сколько из этих копий лежит на каждой конкретной хромосоме. Важность этой информации была обоснована во многих статьях по онкологии. [28, 29, 30, 31]. К примеру, такая утрата гетерозиготности, при которой один аллель утерян, а второй дуплицирован, из-за чего суммарное число копий остаётся равным двум, характерна для многих видов рака [29, 32, 33, 34]. Аллель-специфические структурные вариации также важны для детектирования дупликации генома [35] и для уточнения момента в эволюции опухоли, когда дупликация произошла[28, 35, 36]. Несмотря на это, в более ранних методах считалось, что данные single-cell секвенирования слишком разреженные для их определения [37, 38]. Существующие методы, кроме CHISEL[3], способны определять только число копий посредством анализа RDR. Стандартной техникой определения аллельного дисбаланса по данным bulk-секвенирования является анализ частоты гетерозиготных аллелей. Ранее считалось, что по данным single-cell секвенирования эти частоты надёжно определить нельзя, но и XClone, и CHISEL демонстри-

рут, что это возможно.

По состоянию на июнь 2020 года, опубликован только один непосредственный аналог XClone — алгоритм CHISEL^[3], разработанный в лаборатории Бена Рафаэля в Принстонском университете. CHISEL был выложен на онлайн-архив предпубликаций bioRxiv уже после того, как была начата работа над XClone. Тем не менее, несмотря на то, что оба метода дают сравнимые результаты в задаче поиска аллель-специфических структурных вариаций, XClone выгодно отличается от CHISEL тем, что в его статистическую модель можно естественным образом интегрировать другие типы данных, такие как данные РНК-секвенирования, данные о соматических мутациях, данные о митохондриальной ДНК, в то время как алгоритм CHISEL решает узкую задачу и никаким тривиальным образом не обобщается. Авторы верят, что именно в этой гибкости и масштабируемости заключается научная новизна XClone.

Метод был опробован на реальных данных, извлечённых из медуллобластомы — детской опухоли мозга, — полученных по протоколам компаний 10X Genomics³ и Smart-Seq⁴, но концептуально он не привязан к конкретной платформе и может быть адаптирован к другим форматам входных данных.

На концептуальном уровне, XClone также можно использовать для интеграции геномных и транскриптомных данных, но на данный момент, в силу низкого разрешения данных РНК-секвенирования одиночных клеток, эту гипотезу не удается убедительно доказать. Кроме того, в процессе работы над алгоритмом, к сожалению, стало ясно, что опухоль в образцах медуллобластомы однородная: в ней выжила всего одна клonalная линия. Тем не менее, авторы ведут переговоры с научной группой

³<https://www.10xgenomics.com/products/single-cell-cnv/>

⁴<https://www.illumina.com/science/sequencing-method-explorer/kits-and-arrays/smart-seq2.html>

пой из Стэнфорда, располагающей более удачными образцами опухоли желудка. По предварительным сведениям, их образцы содержат порядка 10 отчётливо различных клональных линий, а глубина покрытия должна позволить получить приемлемое отношение сигнала к шуму.

Специфика предметной области определила не совсем обычную для статей по машинному обучению структуру экспериментов в данной ВКР: каждая опухоль индивидуальна, а секвенирование стоит больших денег, потому как таковой обучающей выборки не было. Даже генерация синтетических данных это отдельная нетривиальная задача, которая до сих пор не решена на должном уровне. Более того, в момент начала работы метод не имел аналогов, потому статистическая модель приобрела свой нынешний вид методом проб и ошибок.

Даже сравнить методы и доказать, что один из них лучше другого, на данный момент возможно только на качественном уровне, т.к. по состоянию на 2020 год просто нет реальных данных, в которых в каждой из клеток были бы достоверно известны все структурные вариации генома. Масштабные же вариации хорошо предсказывает как XClone, так и CHISEL. Тем не менее, ведётся активная разработка более продвинутой генерации синтетических клеток. Как только она будет завершена, оба метода будут подвергнуты тщательному сравнительному анализу.

Но несмотря на все эти трудности всего за полгода был получен значительный прогресс. Есть основания рассчитывать на публикацию улучшенной версии алгоритма XClone в высокоимпактных журналах: алгоритм Cardelino[10], на котором была основана первая версия XClone, в марте 2020 года был опубликован в Nature Methods.

4 Материалы и методы

4.1 Вероятностная модель ДНК-секвенирования

При работе с данными секвенирования часто возникает задача оценить матожидание числа прочтений по заданному участку генома. В случае DNA-seq, эта величина описывается простой вероятностной моделью

$$X_i \sim \text{Poisson}(S p_i Q(g_i) m_i)$$

Рис. 4.1: Вероятностная модель числа прочтений X_i в сегменте i в DNA-seq. S — *scale factor*, p_i — число копий сегмента, $Q(g_i)$ — влияние *GC-состав*, m_i — *bin mappability*

Ниже приведены определения этих факторов:

Определение 4.1 (*Scale factor*). Число ридов, которые фрагмент ДНК порождает при секвенировании. Эта величина зависит как от **сложности библиотеки** (матожидание числа различных молекул, которые могут получиться в ходе ПЦР), так и от **глубины секвенирования** (точное значение зависит от технологии, но неформально стоит понимать как число ридов на единицу длины; увеличение амплификации повышает глубину покрытия, но и увеличивает затраты).

Определение 4.2 (*Bin mappability*). Bin mappability неформально следует понимать как долю k -меров из заданного диапазона, которые однозначно выравниваются на этот же диапазон, где k подчиняется Пуассоновской модели данных секвенирования. Если диапазон состоит из повторов одного короткого участка, то его mappability будет низкой, так как однозначно выравниваться будут только риды длиной больше половины от размера этого диапазона, вероятность которых будет мала.

При заданной сегментации, эту величину можно с заданной точностью посчитать аналитически, но обычно для этого используют метод Монте-Карло.

Определение 4.3 (GC-состав). Доля гуанина (G) и цитозина (C) среди нуклеотидов последовательности. В комплементарной GC-паре три водородных связи вместо двух как у AT-пар, потому последовательности с высоким содержанием G и C более устойчивы к нагреву, а потому реже расщепляются на фрагменты, достаточно короткие для амплификации при ПЦР. Аналогично, если GC-состав очень мал, то велик шанс, что при нагреве последовательность распадётся на слишком маленькие части, к которым уже нельзя будет присоединить праймер. Как следствие, покрытие последовательностей со слишком большим или слишком маленьким GC-составом в среднем ниже.

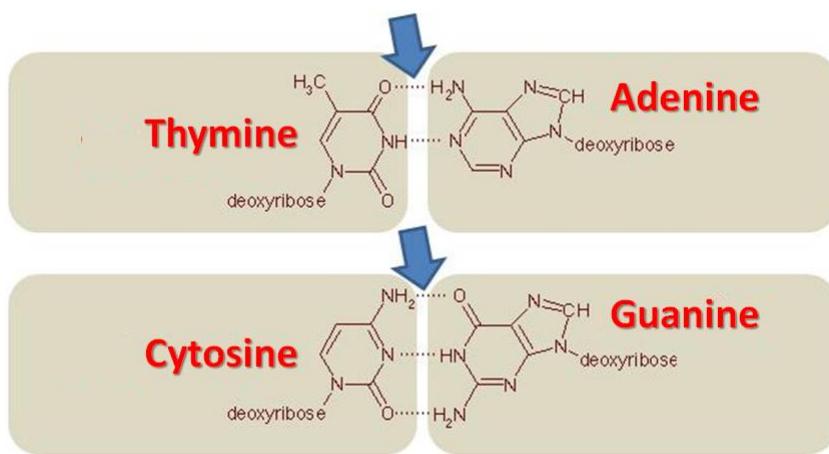


Рис. 4.2: Комлементарные пары: аденин-тимин и гуанин-цитозин

В случае RNA-seq простой модели, к сожалению, быть не может. Дело в том, что геном статичен. Факторов, которые могут повлиять на распределение ридов в DNA-seq, не так много: это либо структурные вариации, либо особенности нуклеотидной последовательности как строчки, без привязки к её биологическому смыслу. Картина экспрессии же

постоянно меняется. На неё влияет и клеточный цикл, и окружающая среда, и патологии отдельных компонент клетки. Многочисленные регуляторные механизмы не позволяют моделировать экспрессию генов по отдельности: уровни экспрессии часто коррелируют, а иногда и зависят друг от друга нелинейно (т.н. **синергия генов**). Хуже того, в науке хорошо изучено такое явление как **эпистаз** — мутации в одном гене могут приводить к качественным изменениям фенотипа, выходящим далеко за пределы непосредственных функций этого гена. В современной науке существует множество моделей транскрипции, принимающих во внимание многие из этих факторов, но их содержательный обзор выходит далеко за рамки данной работы.

4.2 Алгоритмы предобработки данных

Профессия вычислительного биолога подразумевает рутинную обработку больших гетерогенных данных, особенно что касается single-cell технологий. В связи с этим был реализован протокол предобработки данных секвенирования, основные шаги которого разобраны в данном разделе.

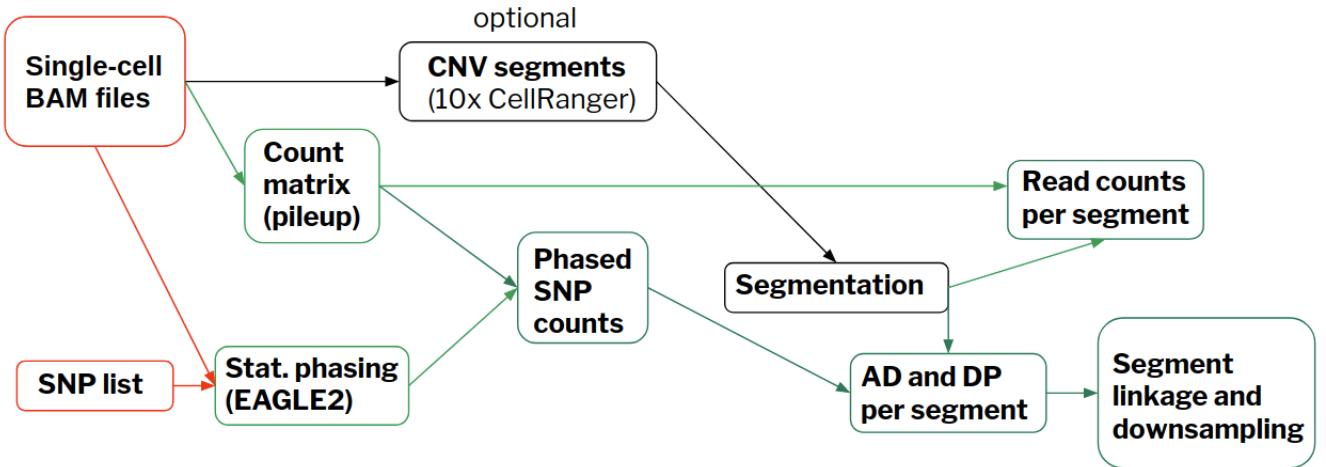


Рис. 4.3: Граф протокола предобработки данных для алгоритма XClone.

Красным обозначены входные данные, чёрным — опциональные шаги, зелёным — реализованные стадии.

4.2.1 Извлечение данных из ВАМ-файлов

ВАМ — *binary SAM* — *binary sequence alignment/map format* — общепринятый формат сжатого хранения данных секвенирования, с подробностями которого можно ознакомиться в оригинальной публикации [4]. ВАМ-файл, полученный по протоколам 10x Genomics, занимает до нескольких терабайтов дискового пространства, потому эффективное извлечение информации из ВАМ-файлов это нетривиальная инженерная задача. Главные входные файлы XClone — матрицы прочтений. Таких матриц требуется три:

- матрица RD всех прочтений достаточного качества;
- матрица DP всех прочтений, накрывающих хоть один ОНП в пределах сегментов;
- матрица AD всех прочтений, накрывающих хоть альтернативный аллель ОНП в пределах сегментов.

Для получения матрицы RD из данных scRNA-seq был использован протокол **count** из **CellRanger**. Для всех остальных матриц во всех остальных случаях был использован **CellsNP**⁵.

4.2.2 Статистическое фазирование гаплотипов

Гаплотипирование — определение того, от какого родителя унаследован каждый аллель в геноме — одна из ключевых задач генетики человека. Сложность её решения обусловлена контекстом, в котором она возникает в современных исследованиях, когда секвенируются порядка $2 \cdot 10^4 - 10^6$ позиций в геномах тысяч человек. Если прочтения короткие и не накрывают много позиций одновременно, то нужно секвенировать обоих родителей каждого участника эксперимента, что непрактично и не всегда возможно. Следовательно, нужно разрабатывать статистические методы гаплотипирования. Они основаны на наблюдении, что некоторые группы аллелей часто наследуются совместно. Это явление называется **неравновесной сцепленностью**. Если прогаплотипировано достаточное количество представителей популяции, то можно построить приближённые таблицы сцепленности и гаплотипировать новые образцы методом максимизации правдоподобия.

На момент написания этого текста, стандартом статистического гаплотипирования считается алгоритм **EAGLE2**[5]⁶. Этот алгоритм основан на скрытых марковских моделях и использует 32,470 образца из базы данных **Haplotype Reference Consortium**[6].

Алгоритм EAGLE2 обладает существенным недостатком: его метки имеют только локальный смысл. В пределах окна в 20-50 килобаз любые два ОНП с одинаковой наследуются совместно, но при сдвиге окна смысл

⁵<https://github.com/single-cell-genetics/cellSNP>

⁶<https://data.broadinstitute.org/alkesgroup/Eagle/>

меток может спонтанно поменяться на противоположный, это так называемая **ошибка смены цепи**. Т.е. два ОНП с разных концов хромосомы, помеченные одной меткой, могут быть унаследованы от разных родителей. Из-за этого в матрицах прочтений размывается сигнал аллельного дисбаланса: чтобы сделать данные менее разреженными, прочтения соседних небольших сегментов суммируются, в том числе и аллель-специфичные. Ясно, что если среди двух соседних сегментов с одинаковой меткой один полностью унаследован от отца, а второй — от матери, то при сложении их аллель-специфичные сигналы скомпенсируют друг друга. Это, в свою очередь, приводит к неправильному предсказанию аллель-специфичных структурных вариаций и неправильной кластеризации клеток. Авторы EAGLE2 в переписке явно дали понять, что в общем случае детектировать и исправлять такого рода ошибки их подход не позволяет. Но в контексте модели XClone удалось разработать статистический метод, показавший хорошие результаты при устраниении ошибок смены цепи. Его подробное описание можно найти в одноимённом разделе.

4.2.3 Подходы к сегментации генома

Одной из основных задач XClone является предсказание **ASCNV** — аллель-специфических структурных вариаций генома. Это происходит в несколько этапов: (1) вначале производится сегментация генома с одновременным подсчётом матриц прочтений, (2) затем глубина покрытия сегментов сравнивается с эталонной для подсчёта RDR, (3) откуда получается оценка общего числа копий, (4) которая затем уточняется при помощи сигналов аллельного дисбаланса. Тем не менее, на точность предсказания влияют ещё и технические факторы, фигурирующие в вероятностной модели числа прочтений. Наиболее существенным факто-

ром является bin mappability.

Подсчёт bin mappability — задача чисто техническая и довольно утомительная, т.к. она подразумевает проведение симуляций процесса секвенирования по какому-то конкретному протоколу. Кроме того, она давно считается решённой, а потому не представляет особого научного интереса. В связи с этим, для отфильтровывания участков низкого качества используется готовое решение — **CellRanger DNA**, алгоритм⁷ от 10X Genomics, поставщика оборудования для single-cell секвенирования в научной группе автора. Этот алгоритм разбивает геном на сегменты длиной в 20кб, после чего отфильтровывает те, для которых bin mappability меньше, чем 70% (не более 10-15% при использовании референсного генома GRCh37). CellRanger DNA сам по себе является алгоритмом поиска CNV. Тем не менее, он размечает максимально возможную часть генома каждой из клеток, в том числе участки без структурных вариантов. Благодаря этому можно гарантировать, что все участки генома, пригодные для надёжного определения ASCNV, войдут в итоговую сегментацию.

Найденные участки накрывают некоторое подмножество референсного генома, которое затем подразделяется на сегменты размера 20-50 киlobаз, в пределах которых вероятность ошибки смены цепи невелика, а потому сигнал аллельного дисбаланса статистически достоверный.

⁷https://support.10xgenomics.com/single-cell-dna/software/pipelines/latest/algorithms/cnv_calling

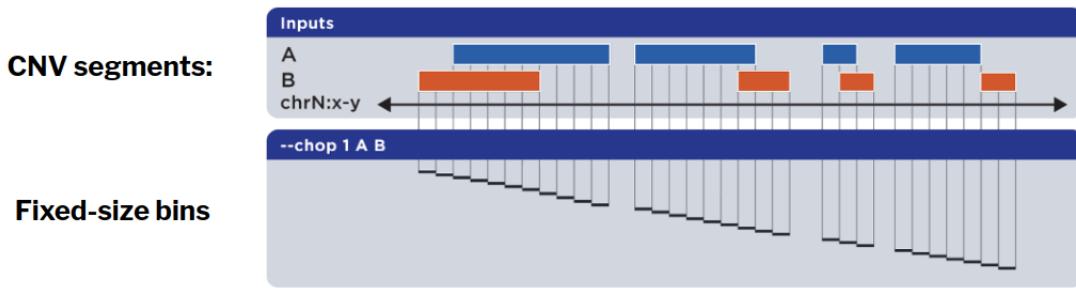


Рис. 4.4: Иллюстрация алгоритма сегментирования генома. Длина индивидуальных сегментов задаётся заранее и выбирается из диапазона 20-50кб. Каждые k подряд идущих фрагментов затем объединяются в блоки. Соответствующие подматрицы прочтений при этом суммируются с одновременной коррекцией ошибок смены цепи.

В силу того, что структурные вариации обычно охватывают участки генома размеров хотя бы в несколько мегабаз, перед началом предсказания уместно агрегировать подряд идущие сегменты в блоки фиксированного размера (обычно 1-5 мегабаз), чтобы получить менее шумные BAF и RDR. Тем не менее, наивно агрегировать содержимое сегментов внутри блока — просуммировать числа прочтений — не получится, т.к. можно потерять аллель-специфический сигнал из-за ошибок смены цепи. В связи с этим был разработан алгоритм суммирования с коррекцией ошибок, который разобран в следующем разделе.

Такой подход к сегментации генома используется в заключительной версии XClone. Тем не менее, изначально большие надежды возлагались на более продвинутый метод, основанный на данных секвенирования длинными прочтениями по технологии Oxford Nanopore. Если прочтения достаточно длинные, они могут накрывать сразу несколько ОНП. Благодаря этому их гаплотипы определяются однозначно: просто из нуклеотидной последовательности рида понятно, какие именно аллели ле-

жат на одной хромосоме. Если покрытие генома достаточно хорошее, то длинные прочтения будут накладываться друг на друга, за счёт чего можно получить достаточно длинные гаплотипы. В силу того, что в первой версии модели ASCNV считались известными, как и клональные линии клеток в DNA-seq образце, для получения итоговой сегментации диапазоны структурных вариаций, обнаруженных CellRanger DNA, пересекались с гаплотипами, полученными по данным Oxford Nanopore.

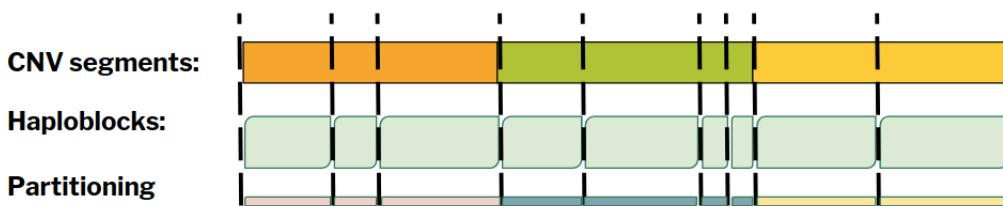


Рис. 4.5: Сегментирование генома в первоначальной версии XClone. Гаплотипические блоки, найденные по данным Oxford Nanopore DNA-seq, пересекаются с диапазонами структурных вариаций, найденных CellRanger DNA.

Тем не менее, такой подход оказался хорош лишь в теории. На практике, найденные таким способом гаплотипические блоки оказывались слишком короткими, порядка нескольких килобаз. Кроме того, распределение их длин имело достаточно большую дисперсию, что доставляло неудобства как при реализации, так и при интерпретации результатов: чем меньше сегмент, тем более зашумленный от него исходит сигнал. Когда таких сегментов много, модель переобучалась на шум в них, что приводило к неадекватному предсказанию меток классов.

4.2.4 Исправление ошибок смены цепи

Поскольку одной из главных задач XClone является предсказание *аллель-специфичных* структурных вариаций в геноме, матрицы AD и DP аллель-

специфичных прочтений должны отражать биологию аллельного дисбаланса в клетках образца. Для этого нужно понимать, к какому гаплотипу принадлежит каждый ОНП. В разделе про статистическое гаплотипирование ОНП был сделан акцент на том, что существующие алгоритмы гарантируют только локальную корректность: при использовании алгоритма EAGLE2, следует ожидать, что при разбиении хромосомы на непересекающиеся окна длины 20-50 килобаз все гетерозиготные ОНП в пределах одного окна будут иметь одинаковый гаплотип, если это на самом деле так. Тем не менее, гаплотипы соседних сегментов с точки зрения алгоритма могут не совпадать даже тогда, когда на самом деле должны. К этому приводят так называемые **ошибки смены цепи** (*switching error*) — спонтанная и неявная замена гаплотипических меток на противоположные внутри алгоритма. Классификацию ошибок смены цепи можно найти в статье [7], цитата из которой приведена ниже:

*"Phasing accuracy is typically measured by counting the number of '**switches**' between known maternal and paternal haplotypes that should not occur if individual maternal and paternal chromosomal nucleotide sequence content has been accurately characterized. If an inconsistency is identified, then it is called a '**switch error**.' These switch errors manifest themselves as induced and false recombination events in the inferred haplotypes compared with the true haplotypes. To identify **switch errors**, the phase of each site is compared with upstream neighboring phased sites. The switch error rate (SER) is defined as the number of switch errors divided by the number of opportunities for switch errors. Switch errors were further classified into three categories: **long**, **point**, and **undetermined**. A long switch appears as a large-scale pseudo recombination event; that is, there are no other switches in the local neighborhood around the long switch (e.g., no other switches within three consecutive heterozygous sites). On the contrary, a small-scale switch error*

appearing as two neighboring switch errors is considered as a point switch (e.g., two switches within three consecutive heterozygous sites, with the pair of switches counted as a point switch). The remaining switches are considered undetermined (e.g., only two sites phased in a small phasing block, so the switch error could not be classified into long or point)."

Тем не менее, разбиение генома на фрагменты по 20-50 килобаз непрактично: в силу разреженности данных, это даёт слабый и зашумленный сигнал аллельного дисбаланса. В связи с этим был разработан метод, одновременно решающий обе описанные проблемы. На первом шаге алгоритма происходит разбиение генома на непересекающиеся сплошные сегменты длины L . Затем каждые N подряд идущих сегментов объединяются в блок длины NL . В пределах блока переключения моделируются бернуlliевскими случайными величинами, по одной на каждый сегмент. Параметры этих распределений, в свою очередь, выводятся **EM-алгоритмом**. После исправления ошибок, прочтения сегментов внутри блока суммируются, что даёт более стабильный сигнал. Эта идея была сформулирована в [3], но технические детали были осознанно исключены авторами CHISEL из препринта.

Прежде чем приступать к рассмотрению метода, сформулируем необходимые определения:

Определение 4.4 (EM-алгоритм).

EM-алгоритм (от английского "*EM*" — "*Expectation Maximization*") — метод поиска оценок максимального правдоподобия (ОМП) или оценок апостериорного максимума (ОАП) параметров статистических моделей,

содержащих скрытые переменные.

Алгоритм 1: ЕМ-алгоритм в общем виде

Результат: Θ^* , $p(\mathbf{Z} \mid \mathbf{X}, \Theta^*)$

$t = 0$;

$\Theta^{(0)}$ инициализируется случайно;

до тех пор, пока $Q(\Theta^{(t+1)} \mid \Theta^{(t+1)}) - Q(\Theta^{(t)} \mid \Theta^{(t)}) > \varepsilon$

выполнять

$\mathcal{L}(\Theta^{(t)}; \mathbf{Z}, \mathbf{X}) := p(\mathbf{X}, \mathbf{Z} \mid \Theta^{(t)})$;
$Q(\Theta \mid \Theta^{(t)}) := \mathbb{E}_{\mathbf{Z} \mid \mathbf{X}, \Theta^{(t)}} \log \mathcal{L}(\Theta; \mathbf{Z}, \mathbf{X})$ // Е-шаг
$\Theta^{(t+1)} := \arg \max_{\Theta} Q(\Theta \mid \Theta^{(t)})$ // М-шаг
$t = t + 1$

конец

$\Theta^* := \Theta^{(t)}$

Здесь \mathbf{Z} — дискретные скрытые переменные, Θ — параметры статистической модели, \mathbf{X} — выборка, $\varepsilon > 0$, p — функция плотности. Каждая итерация алгоритма состоит из двух основных шагов:

1. **Е-шаг**, на котором устраняется явная зависимость от скрытых переменных посредством взятия матожидания логарифма совместной функции правдоподобия по условному распределению $\mathbf{Z} \mid \mathbf{X}, \Theta^{(t)}$;
2. **М-шаг**, на котором параметры нового апостериорного распределения $\Theta^{(t+1)}$ выбираются таким образом, чтобы максимизировать $Q(\Theta, \Theta^{(t)})$ — функцию правдоподобия "в среднем".

С теоретическим обоснованием и формальным доказательством корректности ЕМ-алгоритма можно ознакомиться в ([8], стр. 363-365). В контексте решаемой задачи $\mathbf{X}, \mathbf{Z}, \Theta$ имеют следующий смысл:

- $\mathbf{Z} = \{z_1, \dots, z_N\}$ — независимые в совокупности индикаторы кор-

ректности гаплотипов сегментов

$$\forall i : z_i \sim \text{Bern}(p_i)$$

$$\forall q \in \{0, 1\}^N : p(\mathbf{Z} = q \mid p_1, \dots, p_n) = \prod_{i=1}^N p(z_i = q_i \mid p_i) = \prod_{i=1}^N p_i^{q_i} (1 - p_i)^{1-q_i}$$

Если $z_i = 1$, то будем говорить, что сегмент i имеет корректный гаплотип, иначе — инвертированный. Эти обозначения имеют смысл только в пределах одного блока, в соседних блоках метки могут иметь противоположный смысл. Из этого наблюдения становится ясно, что алгоритм не решает проблему переключения полностью, но уменьшает число ошибок за счёт агрегации сегментов в блоки.

- Обозначим через M число клеток образца, тогда $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_M)$, $X_c := (\mathbf{a}_c, \mathbf{b}_c)$ — вектора прочтений для каждой из клеток, по компоненте на сегмент. $\mathbf{a}_c = (a_{c,1}, \dots, a_{c,N})$ — число прочтений аллеля А (альтернативный аллель), $\mathbf{b}_c = (b_{c,1}, \dots, b_{c,N})$ — аллеля Б (референсный аллель).
- $\forall c \in \overline{1, M} : \mathbf{r}_c := \mathbf{a}_c + \mathbf{b}_c$ — вектора прочтений обоих аллелей вместе.
- $\Theta = (\theta_1, \dots, \theta_M; p_1, \dots, p_N)$, где θ_c — пропорция ридов гаплотипа 1 в блоке в клетке c . Алгоритм предполагает, что пропорция гаплотипа 1 одинакова во всех сегментах внутри блока с точностью до переключения.

В этих обозначениях можно сформулировать и доказать следующее утверждение:

Утверждение 4.1. Правила пересчёта параметров апостериорного рас-

пределения на M-шаге EM-алгоритма имеют вид:

$$\begin{aligned} p_i^{(t+1)} &= \frac{p_i^{(t)} \prod_{c=1}^M (\theta_c^{(t)})^{a_{c,i}} (1 - \theta_c^{(t)})^{b_{c,i}}}{p_i^{(t)} \prod_{c=1}^M (\theta_c^{(t)})^{a_{c,i}} (1 - \theta_c^{(t)})^{b_{c,i}} + (1 - p_i^{(t)}) \prod_{c=1}^M (\theta_c^{(t)})^{b_{c,i}} (1 - \theta_c^{(t)})^{a_{c,i}}} \\ \theta_c^{(t+1)} &= \frac{\sum_{i=1}^N a_{i,c} \gamma_{i,1}^{(t)} + b_{i,c} \gamma_{i,0}^{(t)}}{\sum_{i=1}^N r_{i,c}} \end{aligned} \quad (4.1)$$

где $\forall j \in \{0, 1\} : \gamma_{i,j}^{(t)} := P(z_i = j \mid \mathbf{X}, \Theta^{(t)})$.

Доказательство. Вектора прочтений в клетках независимы в совокупности, потому:

$$P(\mathbf{X} \mid \mathbf{Z}, \Theta) = \prod_{c=1}^M p(\mathbf{X}_c \mid \mathbf{Z}, \Theta) = \prod_{c=1}^N \theta_c^{\hat{a}_c(\mathbf{Z})} (1 - \theta_c)^{\hat{b}_c(\mathbf{Z})}$$

Где

$$\begin{cases} \hat{a}_c(\mathbf{Z}) := \sum_{i=1}^N [z_i a_{c,i} + (1 - z_i) b_{c,i}], \\ \hat{b}_c(\mathbf{Z}) := \sum_{i=1}^N [(1 - z_i) a_{c,i} + z_i b_{c,i}], \\ c \in \overline{1, M} \end{cases}$$

Тогда функция правдоподобия и её логарифм принимают вид

$$\begin{aligned} \mathcal{L}(\Theta; \mathbf{X}, \mathbf{Z}) &= p(\mathbf{X}, \mathbf{Z} \mid \Theta) = p(\mathbf{X} \mid \mathbf{Z}, \Theta) p(\mathbf{Z} \mid \Theta) \\ l(\Theta; \mathbf{X}, \mathbf{Z}) &= \log \mathcal{L}(\Theta; \mathbf{X}, \mathbf{Z}) = \\ &= \log \prod_{\mathbf{q} \in \{0,1\}^N} \left(\prod_{c=1}^M \theta_c^{\hat{a}_c(\mathbf{q})} (1 - \theta_c)^{\hat{b}_c(\mathbf{q})} \prod_{i=1}^N p_i^{q_i} (1 - p_i)^{1-q_i} \right)^{\mathbb{I}\{\mathbf{Z}=\mathbf{q}\}} = \\ &= \sum_{\mathbf{q} \in \{0,1\}^N} \mathbb{I}\{\mathbf{Z} = \mathbf{q}\} \left(\sum_{c=1}^M \sum_{i=1}^N \hat{a}_{c,i}(\mathbf{q}) \log \theta_c + \hat{b}_{c,i}(\mathbf{q}) \log(1 - \theta_c) \right) + \\ &\quad + \sum_{\mathbf{q} \in \{0,1\}^N} \mathbb{I}\{\mathbf{Z} = \mathbf{q}\} \left(\sum_{i=1}^N q_i \log p_i + (1 - q_i) \log(1 - p_i) \right) \end{aligned}$$

Изменением порядка суммирования можно показать, что каждая из этих двух сумм распадается на N сумм поменьше, по одной на каждую из

скрытых переменных. В следствие этого и того, что компоненты случайноговектора \mathbf{Z} независимы в совокупности, шаги EM-алгоритма имеют вид:

E-шаг:

$$\begin{aligned}
 p(\mathbf{Z} \mid \mathbf{X}, \Theta^{(t)}) &\propto p(\mathbf{X} \mid \mathbf{Z}, \Theta^{(t)})p(\mathbf{Z} \mid \Theta^{(t)}) \implies \\
 \implies \mathbb{E}_{\mathbf{Z} \mid \mathbf{X}, \Theta^{(t)}} l(\Theta; \mathbf{Z}, \mathbf{X}) &= \sum_{i=1}^N \mathbb{E}_{\mathbf{z}_i \mid \mathbf{X}_i, \Theta^{(t)}} \log \mathcal{L}(\Theta; \mathbf{z}_i, \mathbf{X}_i) = \\
 = \sum_{i=1}^N \sum_{q_i=0}^1 p(\mathbf{z}_i = q_i \mid \mathbf{X}_i, \Theta^{(t)}) &\left(\sum_{c=1}^M \left[\hat{a}_{c,i}(q_i) \log \theta_c + \hat{b}_{c,i}(q_i) \log(1 - \theta_c) \right] + \right. \\
 &\left. + \log p(\mathbf{z}_i = q_i \mid \Theta) \right) = \\
 = \sum_{i=1}^N &\left[\gamma_{i,1}^{(t)} \left(\sum_{c=1}^M [a_{c,i} \log \theta_c + b_{c,i} \log(1 - \theta_c)] + \log p_i \right) + \right. \\
 &\left. + \gamma_{i,0}^{(t)} \left(\sum_{c=1}^M [b_{c,i} \log \theta_c + a_{c,i} \log(1 - \theta_c)] + \log(1 - p_i) \right) \right] = Q(\Theta \mid \Theta^{(t)})
 \end{aligned}$$

M-шаг:

$$\begin{aligned}
 p_i^{(t+1)} = \arg \max_{p_i} Q(\Theta \mid \Theta^{(t)}) &\iff \frac{\gamma_{i,1}^{(t)}}{p_i^{(t+1)}} - \frac{\gamma_{i,0}^{(t)}}{1 - p_i^{(t+1)}} = 0 \iff p_i^{(t+1)} = \gamma_{i,1}^{(t)} \\
 \theta_c^{(t+1)} = \arg \max_{\theta_c} Q(\Theta \mid \Theta^{(t)}) &\iff \frac{\sum_{i=1}^N \gamma_{i,1}^{(t)} a_{c,i} + \gamma_{i,0}^{(t)} b_{c,i}}{\theta_c^{(t+1)}} - \frac{\sum_{i=1}^N \gamma_{i,1}^{(t)} b_{c,i} + \gamma_{i,0}^{(t)} a_{c,i}}{1 - \theta_c^{(t+1)}} = 0 \\
 &\iff \theta_c^{(t+1)} = \frac{\sum_{i=1}^N \gamma_{i,1}^{(t)} a_{c,i} + \gamma_{i,0}^{(t)} b_{c,i}}{\sum_{i=1}^N a_{c,i} + b_{c,i}}
 \end{aligned}$$

Где необходимое условие локального экстремума является также достаточным в силу выпуклости функции $Q(\Theta \mid \Theta^{(t)})$ ([8], стр. 363-364). \square

После того, как значения z_1, \dots, z_N были определены по данным RNA-seq, их же можно использовать для предобработки данных RNA-seq, полученных из образцов тканей того же пациента. Это даёт возможность интеграции двух модальностей в единую статистическую модель.

Стоит отметить, что на практике $p_i^{(t+1)}$ следует считать по эквивалентной, но уже численно устойчивой формуле:

$$p_i^{(t+1)} = \left(1 + \exp \left[\log(1 - p_i^{(t)}) - \log(p_i^{(t)}) + \sum_{c=1}^M \Delta_{c,i} (\log(\theta_c^{(t)}) - \log(1 - \theta_c^{(t)})) \right] \right)^{-1}$$

Где $\Delta_{c,i} := b_{c,i} - a_{c,i}$, а показатель экспоненты стоит искусственно приводить к диапазону $[-C; C]$ для некоторого $C > 0$ (авторами было выбрано $C = 100$). В противном случае $\prod_{c=1}^M (\theta_c^{(t)})^{a_{c,i}} (1 - \theta_c^{(t)})^{b_{c,i}}$ может представлять собой произведение тысяч или даже миллионов очень маленьких величин в больших степенях. Стандартной реализации чисел с плавающей запятой двойной точности недостаточно для хранения результатов промежуточных вычислений при использовании наивной формулы.

4.3 XClone-V1: только ВАФ-модуль

Исходная идея заключалась в том, чтобы определить клональные линии по ДНК-образцу и затем сопоставлять им из клетки РНК-образца, используя аллельный дисбаланс. Основные шаги алгоритма имели вид:

1. По заданной сегментации генома алгоритмом CellRanger оценивалось число копий каждого участка;
2. Потом по scDNA-seq оценивался аллельный дисбаланс в пределах каждого сегмента;
3. ВАФ-вектора конкатенировались с CNV-векторами, после чего клональные линии определялись как вершины на каком-то уровне дерева иерархической кластеризации;
4. ВАФ сегмента использовался для уточнения числа клонального копий на копиях хромосомы: из конфигураций $\{(i,k, c_{i,k} - m_{i,k})\}_{m_{i,k}=0}^{c_{i,k}}$,

где $c_{i,k}$ это общее число копий участка i в клonalной линии k , а $m_{i,k}$ — число копий на материнской хромосоме, подбиралась та, для которой отклонение $|\frac{m_{i,k}}{c_{i,k}-m_{i,k}} - \text{BAF}(i, k)|$ от BAF минимально. При этом BAF-вектор клonalной линии определялся как усреднённые BAF-вектора клеток этой линии.

5. Аналогично, подсчитывался аллельный дисбаланс в клетках из scRNA-seq образца. Затем клетке присваивалась метка той клonalной линии, ASCNV-вектор которой наилучшим образом объяснял бы наблюдаемые BAF-ы сегментов (в смысле максимизации правдоподобия в бета-биномиальной модели, где вероятности задавались долей копий сегмента на материнской копии хромосомы).

Подход хорошо зарекомендовал себя на синтетических данных, но столкнулся с теми же проблемами, что и алгоритм CaSpER[9], который вышел незадолго после начала работы над XClone. Практика показала, что надёжно оценить BAF по данным scRNA-seq, полученных по протоколам от 10X Genomics, крайне трудно в силу их разреженности. В силу того, что других данных с достаточно богатой клonalной структурой на тот момент не было, от этой архитектуры пришлось отказаться, а вместо переноса меток с геномных данных на транскриптомные сосредоточиться на поиске ASCNV в данных.

4.3.1 Структура модели

Структура XClone-V1 вдохновлялась графической моделью «Cardelino»[10] и имеет с ней много общего.

4.3.1.1 Параметры:

1. Константы:

- M^G, M^E — число клетках в ДНК- и РНК-образцах соотв-но.
- K — предполагаемое число клональных линий в образце.
- N — количество сегментов в выбранной сегментации генома (она одна и та же в обоих образцах).
- T_{\max} — максимальное допустимое число копий сегмента (всё, что больше этого значения, округляется к нему). Значение по умолчанию — 6.
- τ — набор допустимых конфигураций аллель-специфического числа копий. Значение по умолчанию —

$$\{(1, 0), (0, 1), (2, 0), (1, 1), (1, 2), \dots, (T_{\max}, 0), \dots, (0, T_{\max})\}$$

Случай $(0, 0)$, когда рассматриваемого сегмента нет на обеих копиях хромосомы, не рассматривается. Это логично, ведь из-за разреженности данных неясно, в самом ли деле произошла двойная делеция, или просто не хватает данных.

2. Прочие заранее известные величины:

- D^G, D^E — матрицы числа прочтений. Имеют размерности (N, M^G) и (N, M^E) соотв-но, где число прочтений внутри сегмента это сумма числа прочтений гетерозиготных ОНП внутри сегмента.
- A^G, A^E — то же, но для прочтений материнского аллеля.
- $\mathbf{f} = (f_1, \dots, f_K)$ — K -мерный вектор с предсказанными частотами каждой из клональных линий ($\sum_{k=1}^K f_k = 1$). По умолчанию распределение равномерное: $\forall k : f_k = \frac{1}{K}$.
- \mathbf{T} — матрица клональных ASCNV. Имеет размерность (N, K) , $\mathbf{T}_{i,k}$ — конфигурация сегмента i в клональной линии k . Способ

задания этой матрицы: берётся матрица числа копий, предсказанных CellRanger, а затем уточняется с помощью BAF-сигнала в данных. Число копий сегмента определяется как округлённое матожидание копий содержащихся в нём CNV-событий с точки зрения алгоритма CellRanger.

- \mathbf{I}^G — клональные линии клеток ДНК-образца. Определяются посредством максимизации правдоподобия в биномиальной модели: назначается метка той клональной линии, BAF-профиль которой больше всего похож на то, что видно в данных.

3. Скрытые переменные:

- Θ — ожидаемая доля прочтений материнских аллелей в зависимости от ASCNV. $\Theta_{i,t}$ — BAF сегмента i в конфигурации t . Если бы в модели были только ДНК-данные, то не было бы никакой нужды в сегмент-специфических частотах. Тем не менее, BAF-сигнал того же сегмента в той же клетке, но в РНК-данных может сильно отличаться от BAF-сигнала в ДНК-данных. Сегмент-специфические параметры позволяют подобрать компромиссное значение, которое достаточно хорошо описывает и ДНК-данные, и РНК-данные. Диапазон допустимого отклонения от теоретического значения контролируется заданием априорного распределения.
- $\mathbf{I}^E \in [K]^M$ — клональные линии клеток РНК-образца.

4. Сокращения:

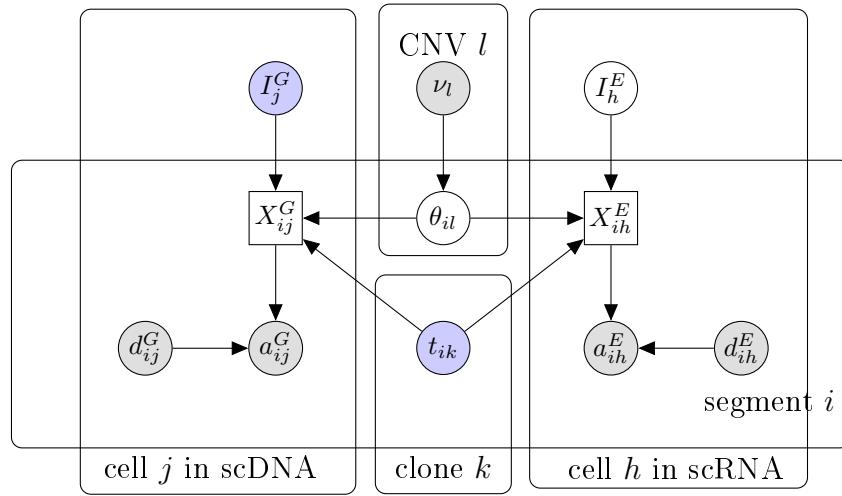
- $\mathbf{H}^G, \mathbf{H}^E$ — ASCNV-конфигурация сегментов в клетках при текущей расстановке клональных меток:

$$\mathbf{H}_{i,j}^G := \mathbf{T}_{i,I_j^G}, \quad \mathbf{H}_{i,j}^E := \mathbf{T}_{i,I_j^E}$$

- $\mathbf{X}^G, \mathbf{X}^E$ — ожидаемая доля прочтений материнских аллелей при текущей расстановке клональных меток:

$$\mathbf{X}_{i,j}^G := \Theta_{i,H_{i,j}^G}, \quad \mathbf{X}_{i,j}^E := \Theta_{i,H_{i,j}^E}$$

4.3.1.2 Визуализация: В этих обозначениях модель XClone-V1 можно изобразить так:



Это т.н. plate notation — часто используемый способ краткой записи графических байесовских моделей с повторяющимися фрагментами. Ориентированное ребро означает, что вершина-исток задаёт распределение вершины-стока. Серый цвет кодирует известные данные. Фиолетовый цвет — авторская вольность, он кодирует скрытые величины, предказанные по данным на этапе предобработки. Белый цвет кодирует скрытые переменные, распределения на которых выводятся в ходе обучения модели. Величины, принадлежащие одному прямоугольнику, образуют смысловой блок, который повторяется для каждого из объектов (например, клетки).

4.3.1.3 Статистическая модель

$$\begin{aligned} \mathcal{L}(\Theta, \mathbf{I}^E) &= \left(\prod_{j=1}^{M^G} \sum_{k^G=1}^K \text{P}(\mathbf{A}_j^G \mid \mathbf{D}_j^G, \mathbf{I}_j^G = k^G, \Theta) \right) \times \\ &\quad \times \left(\prod_{j=1}^{M^E} \sum_{k^E=1}^K \text{P}(\mathbf{A}_j \mid \mathbf{D}_j^E, \mathbf{I}_j^E = k^E, \Theta) \cdot \text{P}(\mathbf{I}_j^E = k^E \mid \mathbf{f}) \right) \end{aligned} \quad (4.2)$$

где отдельные сомножители записываются так:

- Апостериорное распределение на клональных метках:

$$\begin{aligned} \text{P}(\mathbf{I}_j^E = k_0 \mid \mathbf{A}_j, \mathbf{D}_j, \mathbf{f}, \Theta) &= \\ &= \frac{\text{P}(\mathbf{A}_j \mid \mathbf{D}_j, \mathbf{I}_j^E = k_0, \Theta) \text{P}(\mathbf{I}_j^E = k_0 \mid \mathbf{f})}{\sum_{k=1}^K \text{P}(\mathbf{A}_j \mid \mathbf{D}_j, \mathbf{I}_j^E = k, \Theta) \text{P}(\mathbf{I}_j^E = k \mid \mathbf{f})} \end{aligned} \quad (4.3)$$

- Правдоподобие в BAF-модели: Клетки можно считать независимыми в совокупности, потому правдоподобие факторизуется по ним:

$$\begin{aligned} \text{P}(\mathbf{A}_j^G \mid \mathbf{D}_j^G, \mathbf{I}_j^G = k^G, \Theta^E) &= \prod_{i=1}^N \text{P}(\mathbf{A}_{i,j}^G \mid \mathbf{D}_{i,j}^G, \mathbf{X}_{i,j}^G) \\ \text{P}(\mathbf{A}_j^E \mid \mathbf{D}_j^E, \mathbf{I}_j^E = k^E, \Theta) &= \prod_{i=1}^N \text{P}(\mathbf{A}_{i,j}^E \mid \mathbf{D}_{i,j}^E, \mathbf{X}_{i,j}^E) \end{aligned} \quad (4.4)$$

Если блоки тоже считать независимыми в совокупности, то правдоподобие клетки факторизуется по ним, а правдоподобие отдельного блока подчиняется биномиальному закону:

$$\begin{aligned} \text{P}(\mathbf{A}_{i,j}^G \mid \mathbf{D}_{i,j}^G, \Theta) &= \text{Binom}(\mathbf{A}_{i,j}^G \mid \mathbf{D}_{i,j}^G, \mathbf{X}_{i,j}^G) \\ \text{P}(\mathbf{A}_{i,j}^E \mid \mathbf{D}_{i,j}^E, \Theta) &= \text{Binom}(\mathbf{A}_{i,j}^E \mid \mathbf{D}_{i,j}^E, \mathbf{X}_{i,j}^E) \end{aligned} \quad (4.5)$$

На параметрах Θ задано априорное распределение ν . Итоговая BAF-модель подчиняется бета-биномиальному закону. По формуле Байеса отсюда получаем следующую пропорциональность для апостериорного рас-

пределения:

$$\begin{aligned} P(\Theta | \mathbf{A}, \mathbf{D}, \mathbf{f}, \boldsymbol{\nu}) &\propto P(\Theta | \boldsymbol{\nu}) \times \mathcal{L}(\Theta) = \\ &= \prod_{l=1}^N \prod_{t \in \tau} \text{Beta}(\alpha_{l,t}, \beta_{l,t}) \times \mathcal{L}(\Theta) \end{aligned} \quad (4.6)$$

Где параметры $\alpha_{l,t}, \beta_{l,t}$ задаются таким образом, что мода $\text{Beta}(\alpha_t, \beta_t)$, задаваемая выражением $(\alpha_{l,t} - 1)/(\alpha_{l,t} + \beta_{l,t} - 2)$, равняется $t_0/(t_0 + t_1)$, ведь в теории доля прочтений материнских аллелей внутри сегмента l должна мало отличаться от доли копий сегмента на материнской хромосоме. Опуская для краткости индексы, можно формализовать задачу так:

$$\frac{\alpha - 1}{\alpha + \beta - 2} = \frac{t_0}{t_0 + t_1}, \quad \alpha \geq 1, \beta \geq 1$$

Разберём случаи:

- Если $t_0 = 0$, то очевидно, что $\alpha = 1$ и любое $\beta > 1$ подходит (для определённости, можно положить его равным $1 + \varepsilon$ для какого-нибудь $\varepsilon > 0$).
- Если $t_0 > 0$, то:

$$\begin{aligned} (\alpha - 1)(t_0 + t_1) &= (\alpha + \beta - 2)t_0 \\ t_1\alpha - (t_0 + t_1) + 2t_0 &= t_0\beta \\ \beta &= \frac{t_1}{t_0}\alpha + \left(1 - \frac{t_1}{t_0}\right) \end{aligned} \quad (4.7)$$

При $\alpha = C$ отсюда получаем, что:

$$\beta = (C - 1)\frac{t_1}{t_0} + 1$$

Можно взять любое $C > 1$, причём чем эта константа больше, тем более сильные априорные предположения и тем меньше допустимое отклонение от них.

4.3.2 Поиск оптимальных параметров

Для поиска оценки апостериорного максимума для параметров модели используется семплирование по Гиббсу. Это частный случай алгоритма Метрополиса-Гастингса, основанное на знании условного распределения каждого параметра модели при фиксированных остальных. В силу того, что этот метод работает довольно медленно, от него было решено отказаться в пользу вариационного байесовского вывода в более поздних версиях XClone, потому подробный обзор метода в общем случае в данной работе не приводится. Заинтересованный читатель может ознакомиться с ним в классической книге Бишопа[11]. Важным практическим соображением является то, что семплирование по Гиббсу, в отличие от вариационного байесовского вывода, в пределе обязательно даёт оценку максимума а постериори, но явных оценок на скорость сходимости обоих методов нет, а на практике семплирование по Гиббсу работает гораздо медленнее.

Для того, использовать семплирование по Гиббсу, нужно знать распределение для каждого из параметров при фиксированных остальных. Для **клональных меток** эти распределения имеют простой вид:

Утверждение 4.2. При фиксированных остальных параметрах модели верна формула

$$\begin{aligned} P(\mathbf{I}_j^E = k \mid \mathbf{I}_{-j}^E, \mathbf{A}^E, \mathbf{D}^E, \mathbf{f}, \Theta) &\propto \\ &\propto P(\mathbf{A}_j^E \mid \mathbf{D}_j^E, \mathbf{I}_j^E = k, \Theta) P(\mathbf{I}_j^E = k \mid \mathbf{f}) \end{aligned} \tag{4.8}$$

Где $P(\mathbf{I}_j^E = k \mid \mathbf{f})$ это константа, а $P(\mathbf{A}_j^E \mid \mathbf{D}_j^E, \mathbf{I}_j^E = k, \Theta)$ это просто правдоподобие в бета-биномиальной BAF-модели.

Доказательство. Согласно формуле Байеса,

$$\begin{aligned} \text{P}(\mathbf{I}_j^E = k \mid \mathbf{I}_{-j}^E, \mathbf{A}^E, \mathbf{D}^E, \mathbf{f}, \Theta) &= \\ &= \frac{\text{P}(\mathbf{I}_{-j}^E, \mathbf{A}^E \mid \mathbf{D}^E, \mathbf{I}_j^E = k, \mathbf{f}, \Theta) \text{P}(\mathbf{I}_j^E = k \mid \mathbf{f})}{\sum_{c=1}^K \text{P}(\mathbf{I}_{-j}^E, \mathbf{A}^E \mid \mathbf{D}^E, \mathbf{I}_j^E = c, \mathbf{f}, \Theta) \text{P}(\mathbf{I}_j^E = c \mid \mathbf{f})} \end{aligned}$$

Клетки независимы в совокупности по предположению, потому числитель распадается в произведение по ним. Сомножители, не относящиеся к клетке j , в числителе и знаменателе сократятся, и в итоге получится

$$\begin{aligned} \text{P}(\mathbf{I}_j^E = k \mid \mathbf{I}_{-j}^E, \mathbf{A}^E, \mathbf{D}^E, \mathbf{f}, \Theta) &= \\ &= \frac{\text{P}(\mathbf{A}_j^E \mid \mathbf{D}_j^E, \mathbf{I}_j^E = k, \mathbf{f}) \text{P}(\mathbf{I}_j^E = k \mid \mathbf{f})}{\sum_{c=1}^K \text{P}(\mathbf{A}_j^E \mid \mathbf{D}_j^E, \mathbf{I}_j^E = c, \mathbf{f}, \Theta) \text{P}(\mathbf{I}_j^E = c \mid \mathbf{f})} \end{aligned}$$

□

Формулы для **аллельных частот** тоже, в целом, естественные:

Утверждение 4.3. При фиксированных остальных параметрах модели верна формула

$$\begin{aligned} \Theta_{l,t} \mid \mathbf{I}^G, \mathbf{I}^E &\sim \text{Beta}(\alpha_{l,t} + u_{l,t}, \beta_{l,t} + v_{l,t}) \\ u_{l,t} &= \sum_{j^G=1}^{M^G} \mathbf{A}_{l,j^G}^G \cdot \mathbb{I}\left\{\mathbf{H}_{l,j^G}^G = t\right\} + \sum_{j^E=1}^{M^E} \mathbf{A}_{l,j^E}^E \cdot \mathbb{I}\left\{\mathbf{H}_{l,j^E}^E = t\right\} \quad (4.9) \\ v_{l,t} &= \sum_{j^G=1}^{M^G} \mathbf{B}_{l,j^G}^G \cdot \mathbb{I}\left\{\mathbf{H}_{l,j^G}^G = t\right\} + \sum_{j^E=1}^{M^E} \mathbf{B}_{l,j^E}^E \cdot \mathbb{I}\left\{\mathbf{H}_{l,j^E}^E = t\right\} \end{aligned}$$

где $\mathbf{B}_{l,j} := \mathbf{D}_{l,j} - \mathbf{A}_{l,j}$.

Доказательство. Для того, чтобы получить формулы для конкретного $\Theta_{i,t}$ при известных остальных параметрах, нужно расписать апостериорное распределение всех аллельных частот, предполагая их все неизвест-

ными:

$$\begin{aligned}
 & P(\Theta | \mathbf{A}, \mathbf{D}, \mathbf{I}^G, \mathbf{I}^E, \mathbf{f}, \boldsymbol{\nu}) \propto \\
 & \propto \left\{ \prod_{l=1}^N \prod_{t \in \tau} \text{Beta}(\alpha_{l,t}, \beta_{l,t}) \right\} \left[\prod_{j^G=1}^{M^G} P(\mathbf{A}_{j^G}^G | \mathbf{D}_{j^G}^G, \mathbf{I}_{j^G}^G, \Theta) \right] \left[\prod_{j^E=1}^{M^E} P(\mathbf{A}_{j^E}^E | \mathbf{D}_{j^E}^E, \mathbf{I}_{j^E}^E, \Theta) \right] = \\
 & = \prod_{l=1}^N \prod_{t \in \tau} \text{Beta}(\alpha_{l,t}, \beta_{l,t}) \left[\prod_{j^G=1}^{M^G} \text{Binom}(\mathbf{A}_{j^G}^G | \mathbf{D}_{j^G}^G, \Theta_{i,t}) \right] \left[\prod_{j=1}^{M^E} \text{Binom}(\mathbf{A}_{j^E}^E | \mathbf{D}_{j^E}^E, \Theta_{i,t}) \right] = \\
 & = \prod_{l=1}^N \prod_{t \in \tau} \left[\text{Beta}(\alpha_{l,t}, \beta_{l,t}) \prod_{j^G=1}^{M^G} \prod_{j^E=1}^{M^E} \left(\text{Binom}(\mathbf{A}_{l,j^G}^G | \mathbf{D}_{l,j^G}^G, \mathbf{X}_{l,j^G}^G)^{\mathbb{I}\{\mathbf{H}_{l,j^G}^G=t\}} \times \right. \right. \right. \\
 & \quad \left. \left. \left. \times \text{Binom}(\mathbf{A}_{l,j^E}^E | \mathbf{D}_{l,j^E}^E, \mathbf{X}_{l,j^E}^E)^{\mathbb{I}\{\mathbf{H}_{l,j^E}^E=t\}} \right) \right]
 \end{aligned}$$

Отсюда, расписав формулы выше, можно получить параметры апостериорного распределения. Это простое, но утомительное техническое рассуждение, потому остаток доказательства опущен для краткости изложения. \square

Поиск оптимальных параметров модели устроен следующим образом:

Алгоритм 2: Семплирование по Гиббсу в модели XClone-V1

Результат: \mathbf{I}_*^E, Θ_* — приближение ОМА

Проинициализировать \mathbf{I}^E случайным образом.;

до тех пор, пока $\mathcal{L}(\Theta, \mathbf{I}^E)$ **растёт выполнять**

цикл $j = 1$ до M^E с шагом 1 выполнять

$C_j := c_j \sim P(\mathbf{I}_j^E = k \mid \mathbf{I}_{-j}^E, \mathbf{A}^E, \mathbf{D}^E, \mathbf{f}, \Theta);$

конец

цикл $l = 1$ до N с шагом 1 выполнять

для каждого $t \in \tau$ выполнять

$\Phi_{l,t} := \varphi_{l,t} \sim P(\Theta_{l,t} \mid \mathbf{I}^G, \mathbf{I}^E, \dots);$

конец

конец

если $\mathcal{L}(\Phi, C) > \mathcal{L}(\Theta, \mathbf{I}^E)$ тогда

$\Theta = \Phi;$

$\mathbf{I}^E = \mathbf{C};$

конец

конец

$\mathbf{I}_*^E := \mathbf{I}^E;$

$\Theta_* := \Theta;$

4.3.3 Поиск наиболее вероятной перестановки меток

При валидации модели на синтетических данных ключевым предсказываемым объектом было распределение вероятностей на K клональных метках — матрица

$$\mathbf{P} \in \mathbb{R}_+^{M \times K}, \forall i \in [M] : \sum_{j=1}^K p_{i,j} = 1$$

Тем не менее, модель предсказывала метки с точностью до неизвестной перестановки. Если предсказание точное, т.е. что \mathbf{P} с точностью до пе-

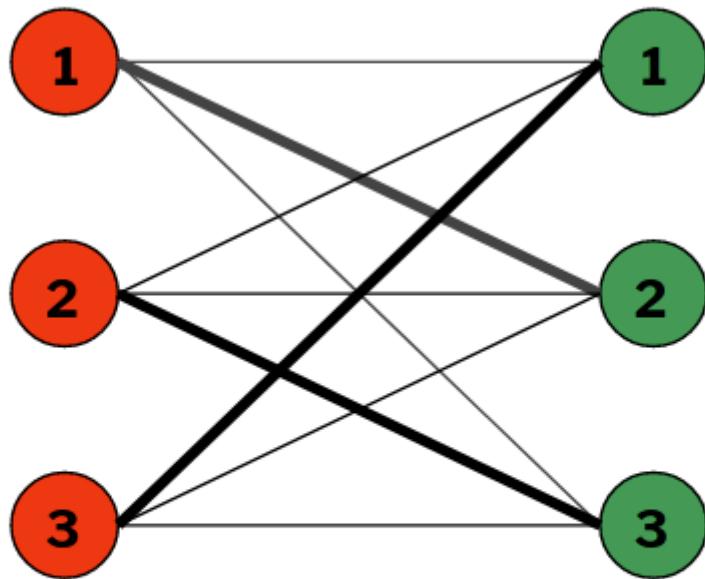
рестановки совпадает с истинной, которая задаётся бинарной матрицей \mathbf{Q} , то сама перестановка определяется легко: биекция между столбцами \mathbf{P} и \mathbf{Q} тривиально строится за $O(K(K + M))$. Но на практике модель ошибается: либо не может восстановить метку, либо не успевает это сделать за отведённое число итераций. Определения качества предсказания в таком случае — нетривиальная задача. Перебор всех возможных перестановок столбцов, коих $K!$, и выбор той, на которой достигается минимальное расстояние между столбцами матриц \mathbf{P} и \mathbf{Q} , реализуем при малых K . Тем не менее, сверхполиномиальная асимптотика не позволяет масштабировать алгоритм: проверить работоспособность модели было бы невозможно даже при $M = 10^4$ и $K = 10$ наивный алгоритм потребовал бы порядка $O(K!M)$ операций, т.е. порядка 4×10^{10} . При этом в открытом доступе опубликованы образцы с $M = 1.3 \times 10^{68}$ ⁸, и типичное количество клеток в данных от 10X Genomics от года к году монотонно увеличивается.

Для решения этой проблемы был разработан полиномиальный алгоритм, находящий оптимальную перестановку за $O(K^4)$ операций и $O(K^2)$ дополнительной памяти. Алгоритм основан на сведении к задаче поиска в двудольном графе совершенного паросочетания минимального веса. А именно: строится взвешенный полный двудольный граф $K_{K \times K}$, столбцам \mathbf{P} сопоставляются вершины левой доли, v_1, \dots, v_K , столбцам \mathbf{Q} — вершины правой доли, u_1, \dots, u_K , а ребру (v_i, u_j) — вес, равный $d(P_i, Q_j)$, где P_i, Q_j — соотв. столбцы, а d — метрика (значение по умолчанию — l_1 -норма). То, что совершенному паросочетанию в таком графе соответствует именно $\arg \min_{\sigma \in S_K} \sum_{j=1}^K d(P_j, Q_{\sigma(j)})$, очевидно по построению. Задача поиска совершенного паросочетания минимального веса в двудольном

⁸<https://www.10xgenomics.com/blog/our-13-million-single-cell-dataset-is-ready-to-download>

графе решается — это т.н. **задача о назначениях**, одна из фундаментальных задач комбинаторной оптимизации. Для её решения применяется так называем "венгерский алгоритм опубликованный в 1955 году американским математиком Гарольдом Куном[12].

Предсказанные классы



Истинные классы

$\pi = [2, 3, 1]$ — перестановка меток,
построенная по совершенному
паросочетанию минимального веса

Рис. 4.6: Иллюстрация сведения задачи восстановления наиболее вероятной перестановки меток классов к поиску совершенного паросочетания минимального веса в двудольном графе. Веса — расстояния между столбцами матриц P и Q , матриц предсказанных и истинных вероятностей клonalных меток.

4.4 XClone-V2: BAF- и RDR-модули

Текущая версия алгоритма XClone использует BAF и RDR для того, чтобы восстановить клональную структуру опухоли и определить профили ASCNV для каждого клона. Она лишена сразу нескольких недостатков прошлой версии:

- Не зависит от внешних алгоритмов детектирования CNV по данным single-cell секвенирования;
- Для поиска оценки апостериорного максимума вместо медленного семплирования по Гиббсу используется эффективно автоматизированный вариационный байесовский вывод на GPU;

От пользователя требуется только предоставить BAM-файлы образцов и набор геномных позиций, использованных при секвенировании. Вся предобработка данных и дальнейшее обучение модели реализовано в виде отдельных скриптов с консольным интерфейсом, для использования которых достаточно задать пути к данным.

Сам алгоритм принимает на вход три матрицы:

- \mathbf{RD} — матрица числа прочтений;
- \mathbf{DP} — матрица числа прочтений, которые выравниваются на последовательности, содержащие хотя бы один ОНП;
- \mathbf{AD} — матрица числа тех прочтений из \mathbf{DP} , которые выравниваются на материнский аллель;

Все три матрицы принадлежат $\mathbb{N}^{N \times M}$, где N это число блоков после предобработки из раздела 4.2.4, а M — число клеток образца. Все три матрицы должны быть подсчитаны по данным scDNA-seq.

Также на вход алгоритма подаётся K — ожидаемое число клональных линий в образце — и τ , набор допустимых ASCNV $\{c_t\}_{t=1}^T$, где $c_t := (c_{t,m}, c_{t,p})$, где $c_{t,m}$ — число копий на материнской хромосоме, а $c_{t,p}$ — соответственно, на отцовской. По умолчанию для τ содержит следующие 16 состояний:

$$\begin{aligned} & ([0,0], [0,1], [0,2], [0,3], \\ & [1,0], [1,1], [1,2], [1,3], \\ & [2,0], [2,1], [2,2], [2,3], \\ & [3,0], [3,1], [3,2], [3,3]) \end{aligned}$$

BAF- и RDR-модули связаны скрытыми переменными \mathbf{Z}, \mathbf{Y} :

- $z_{j,k} := I\{\text{клетка } j \text{ принадлежит клональной линии } k\}$;
- $y_{i,k,t} := I\{\text{блок } i \text{ в клональной линии } k \text{ находится в состоянии } c_t\}$;

на которых заданы априорные вероятности $\boldsymbol{\pi}, \mathbf{U}$:

$$\begin{aligned} p(z_{j,k} = 1 \mid \boldsymbol{\pi}) &= \text{Multinom}(1; K, \boldsymbol{\pi}_j) = \pi_{j,k} \\ p(y_{i,k,t} = 1 \mid \mathbf{U}) &= \text{Multinom}(1; T, \mathbf{u}_{i,k}) = u_{i,k,t} \end{aligned} \tag{4.10}$$

4.4.1 Структура ВАФ-модуля

Если блок i в клетке j находится в состоянии c_t , то $a_{i,j}$ подчиняется биномиальной модели с параметром $\theta_{i,t} \sim \text{Beta}(\alpha_t, \beta_t)$, где α_t, β_t — параметры априорного распределения. Они одни и те же для всех блоков в состоянии c_t и полагаются равными $\alpha_t = (c_{t,1} + 0.01)$ и $\beta_t = (c_{t,2} + 0.01)$. В формульной записи:

$$p(a_{i,j} \mid d_{i,j}, \theta_{i,t}) = \text{Binom}(a_{i,j} \mid d_{i,j}, \theta_{i,t}). \tag{4.11}$$

где $a_{i,j}, d_{i,j}$ — элементы матриц \mathbf{AD}, \mathbf{DP} на позиции (i, j) . В предположении, что числа прочтений в соседних блоках независимы, функция правдоподобия записывается следующим образом:

$$p(\mathbf{AD}, \mathbf{DP} \mid \mathbf{Z}, \mathbf{Y}, \boldsymbol{\theta}) = \prod_{i=1}^N \prod_{j=1}^M \prod_{k=1}^K \prod_{t=1}^T p(a_{i,j} \mid d_{i,j}, \theta_{i,t})^{z_{j,k} \times y_{i,k,t}} \quad (4.12)$$

4.4.2 Структура RDR-модуля

Для каждой клональной линии k модель числа прочтений задаётся мультиномиальным распределением. Каждому блоку i в этой модели назначается вероятность $f_{i,k}$. Она задаётся формулой

$$f_{i,k} := \frac{\sum_{t=1}^T m_i \exp[\gamma_t] P(y_{i,k,t} = 1)}{\sum_{b=1}^N \sum_{t=1}^T m_b \exp[\gamma_t] P(y_{b,k,t} = 1)} \quad (4.13)$$

Где m_i — доля прочтений, попадающих в блок i в здоровой клетке при стремлении глубины покрытия к бесконечности (или какая-то оценка этой величины), $\exp\{\gamma_t\}$ — амплификация, вызванная наличием $c_{t,m} + c_{t,p}$ копий блока в геноме (здесь возведение в степень нужно для того, чтобы гарантировать неотрицательность). Эта модель согласуется с теорией для данных scDNA-seq, т.к. после коррекции технических факторов для ДНК-секвенирования правомерны предположения о том, что прочтения распределяются по геному равномерно, а систематические отклонения от этого правила могут быть вызваны только loss- или gain-событиями (см. модель ДНК-секвенирования из раздела 4.1).

Тогда функция правдоподобия запишется как

$$\begin{aligned} \mathbf{r}_j &:= (r_{1,j}, \dots, r_{N,j}) \\ p(\mathbf{r}_j | \mathbf{f}_k) &:= \text{Multinom}\left(\mathbf{r}_j \left| \sum_{i=1}^N r_{i,j}, \mathbf{f}_k\right.\right) \\ p(\mathbf{RD} | \mathbf{Z}, \mathbf{Y}, \mathbf{m}, \boldsymbol{\gamma}) &:= \prod_{j=1}^M \prod_{k=1}^K p\left(\mathbf{r}_j \left| \sum_{i=1}^N r_{i,j}, \mathbf{f}_k\right.\right)^{z_{j,k}} \end{aligned} \quad (4.14)$$

Если \mathbf{m} заранее неизвестно, то его можно смоделировать посредством задания априорного распределения $\text{Dirichlet}(\boldsymbol{\omega})$. Аналогично для $\boldsymbol{\gamma}$ в качестве априорного распределения задаётся $\text{GP}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, где под GP подразумевается дискретный гауссовский процесс с $\mu_t := \log((c_{t,m} + c_{t,p})/2 + \varepsilon)$ для достаточного малого положительного ε , а матрица ковариаций $\boldsymbol{\Sigma}$ задаётся RBF-ядром $K(\mathbf{c}_t, \mathbf{c}_{t'}) := l_1 \exp\{-l_2[(c_{t,m} - c_{t',m})^2 + (c_{t,p} - c_{t',p})^2]\}$, где l_1, l_2 — гиперпараметры модели. В формульной записи:

$$\begin{aligned} \mathbf{m} | \boldsymbol{\omega} &\sim \text{Dirichlet}(\boldsymbol{\omega}) \\ \boldsymbol{\gamma} | \boldsymbol{\mu}, \boldsymbol{\Sigma} &\sim \text{GP}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \end{aligned} \quad (4.15)$$

4.4.3 Генеративная модель данных

XClone предполагает следующий процесс генерации клетки j с векторами прочтений $\mathbf{r}_j, \mathbf{d}_j$:

1. Сгенерировать метку клональной линии: $z_j \sim \text{Multinom}(1; K, \boldsymbol{\pi})$.
2. Сгенерировать ASCNV для каждого сегмента $y_{i,z_j} \sim \text{Multinom}(1; T, \mathbf{u}_{i,z_j})$.
3. Сгенерировать параметры аллельного дисбаланса для каждого сегмента: $\theta_i \sim \text{Beta}(\alpha_{y_{i,z_j}}, \beta_{y_{i,z_j}})$.
4. Сгенерировать распределение прочтений, накрывающих материнские аллели гетерозиготных сайтов, по сегментам: $a_{i,j} \sim \text{Binom}(d_{i,j}, \theta_i)$

Запись этой версии XClone в plate notation не приведена, т.к. в силу асимметрии ASE- и RDR-модулей такая форма записи менее информативна.

4.4.4 Вариационный байесовский вывод

4.4.4.1 Общее описание метода

Пусть $\mathbf{X} = (X_1, \dots, X_n)$ — выборка из распределения с плотностью $p_\theta(\mathbf{X})$ ($p_\theta(\mathbf{X}) = \prod_{i=1}^n p_\theta(X_i)$), где θ — набор параметров распределения. Пусть q — априорное распределение на Θ — множестве допустимых параметров. Тогда $f(\theta, \mathbf{X}) := p_\theta(\mathbf{X})q(\theta)$ — плотность совместного распределения на $\Theta \times \mathcal{X}$. Тогда апостериорная плотность на параметрах выражается как

$$q(\theta \mid \mathbf{X}) := \frac{f(\theta, \mathbf{X})}{p(\mathbf{X})} = \frac{f(\theta, \mathbf{X})}{\int_{\Theta} f(t, \mathbf{X}) dt} = \frac{f(\theta, \mathbf{X})}{\int_{\Theta} p_t(\mathbf{X}) q(t) dt}$$

Величину $p(\mathbf{X})$ в знаменателе называют **обоснованностью**. В общем случае она не выражается в аналитических функциях.

Пусть стоит задача найти оценку апостериорного максимума (ОАМ) — $\arg \max_{\Theta} q(\theta \mid \mathbf{X})$, — а также распределение на всех параметрах из θ . В силу того, обоснованность $p(\mathbf{X})$ это константа, можно максимизировать не $q(\theta \mid \mathbf{X})$, а $f(\theta, \mathbf{X})$. Но совместная плотность как функция может быть устроена произвольно сложно. Уже при счётом Θ поиск её максимума это нетривиальная задача. Тем не менее, задача поиска ОАМ допускает альтернативную формулировку в форме задачи непрерывной оптимизации.

Определение 4.5 (KL-дивергенция). Пусть p, q — плотности распределений P, Q над одним вероятностным пространством (Ω, \mathcal{F}, P) . Тогда KL-дивергенцией $KL(P \parallel Q)$ называют величину $\int_{\Omega} p(x) \log \frac{p(x)}{q(x)} dx$.

Утверждение 4.4. Пусть $\mathbb{F}_{\mathcal{X}}$ — пространство абсолютно непрерывных распределений над вероятностным пространством $\mathcal{X} = (\Omega, \mathcal{F}, P)$, тогда

1. $\forall P, Q \in \mathbb{F}_{\mathcal{X}} : KL(P \parallel Q) \geq 0$, причём $KL(P \parallel Q) = 0 \iff P \stackrel{\text{П.ВС.}}{=} Q$.
2. $\forall P_1, P_2, Q \in \mathbb{F}_{\mathcal{X}} : KL(P_1 \parallel P_2) \leq KL(P_1 \parallel Q) + KL(Q \parallel P_2)$

Кроме того, для произвольного \mathcal{X} могут найтись $P, Q \in \mathbb{F}_{\mathcal{X}}$, такие что $KL(P \parallel Q) \neq KL(Q \parallel P)$, что не позволяет считать KL-дивергенцию метрикой на $\mathbb{F}_{\mathcal{X}}$.

Вариационный байесовский вывод (далее — VB) основан на следующей идеи: заменить апостериорное распределение $Q(\theta \mid \mathbf{X})$, которое может быть устроено сколь угодно сложно, на удобную для оптимизации аппроксимацию $Q^*(\theta)$ из семейства распределений \mathcal{Q} . Тогда есть следующий алгоритм приближённого поиска ОМА:

Алгоритм 3: Вариационный байесовский вывод, общий вид

Lined Результат: θ_0^*, Q_0^* — приближение ОМА и соотв.

аппроксимация

$$\begin{aligned} Q_0^* &:= \arg \min_{Q^* \in \mathcal{Q}} KL(Q^* \parallel Q(\cdot \mid \mathbf{X})); \\ \theta_0^* &:= \arg \max_{\theta \in \Theta} Q_0^*(\theta) \end{aligned}$$

Этот метод центральный для алгоритма XClone, потому в этом разделе приведен его подробный обзор.

Определение 4.6 (ELBO). В обозначениях данного раздела, величину

$$\mathcal{L}(Q^*) := \int_{\Theta} q^*(\theta) \log \frac{f(\theta, \mathbf{X})}{q^*(\theta)} d\theta$$

называют ELBO — evidence lower bound. На русский язык это можно перевести как «нижняя оценка логарифма обоснованности», но устоявшегося перевода в сообществе нет, потому обычно используют сокращение ELBO.

Утверждение 4.5. Максимизация ELBO по $Q^* \in \mathcal{Q}$ эквивалентна минимизации $KL(Q^* \parallel Q_{\mathbf{X}})$, где $Q_{\mathbf{X}}(\theta)Q(\cdot \mid \mathbf{X})$.

Доказательство. Распишем $KL(Q^* \parallel Q_{\mathbf{X}})$:

$$\begin{aligned}
 KL(Q^* \parallel Q_{\mathbf{X}}) &= \mathbb{E}_{\varphi \sim Q^*} \log \frac{q^*(\varphi)}{q_{\mathbf{X}}(\varphi)} = \\
 &= -\mathbb{E}_{\varphi \sim Q^*} \log \frac{q_{\mathbf{X}}(\varphi)}{q^*(\varphi)} = \\
 &= -\mathbb{E}_{\varphi \sim Q^*} \log \frac{f(\varphi, \mathbf{X})}{p(\mathbf{X})q^*(\varphi)} = \\
 &= -\mathbb{E}_{\varphi \sim Q^*} \log \frac{f(\varphi, \mathbf{X})}{q^*(\varphi)} + \mathbb{E}_{\varphi \sim Q^*} \log p(\mathbf{X}) = \\
 &= -\mathcal{L}(Q^*) + \log p(\mathbf{X})
 \end{aligned} \tag{4.16}$$

Отсюда получаем, что $\log p(\mathbf{X}) = L(Q^*) + KL(Q^* \parallel Q_{\mathbf{X}})$. Логарифм обоснованности не зависит от θ , потому максимизация $\mathcal{L}(Q^*)$ эквивалентна минимизации $KL(Q^* \parallel Q_{\mathbf{X}})$, а потому и решению задачи вариационного байесовского вывода.

ELBO можно получить и иначе, расписав $\log p(\mathbf{X})$:

$$\begin{aligned}
 \log p(\mathbf{X}) &= \log \int_{\Theta} f(\theta, \mathbf{X}) d\theta \\
 &= \log \int_{\Theta} f(\theta, \mathbf{X}) \frac{q^*(\theta)}{q^*(\theta)} d\theta = \\
 &= \log \mathbb{E}_{\varphi \sim Q^*} \frac{f(\varphi, \mathbf{X})}{q^*(\varphi)} = \\
 &= \left[\text{Неравенство Йенсена} \right] \geqslant \\
 &\geqslant \mathbb{E}_{\varphi \sim Q^*} \log \frac{f(\varphi, \mathbf{X})}{q^*(\varphi)} = \mathcal{L}(Q^*)
 \end{aligned} \tag{4.17}$$

Такое доказательство не проясняет связь с VB, но полезно в контексте других приложений и помогает понять, почему эта величина имеет именно такое название. \square

Замечание 4.1. В силу того, что KL-дивергенция не симметрична, вид оптимальной аппроксимации может противоречить интуитивному представлению о том, как она должна быть устроена. Для лучшего понимания полезно рассмотреть несколько примеров из [8] (стр. 734), кото-

рые иллюстрируют разницу между $KL(Q^* \parallel Q_X)$ — *forward KL* — и $KL(Q_X \parallel Q^*)$ — *reverse KL*.

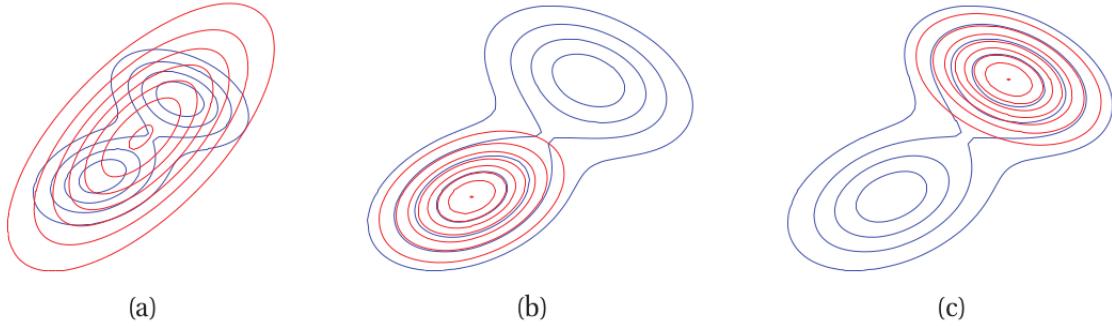


Рис. 4.7: Аппроксимация бимодального распределения Q (синие линии) двумерным гауссовским распределением Q^* (красные линии). (а) — результат минимизации $KL(Q \parallel Q^*)$, (б)-(с) — варианты оптимального приближения при минимизации $KL(Q^* \parallel Q)$

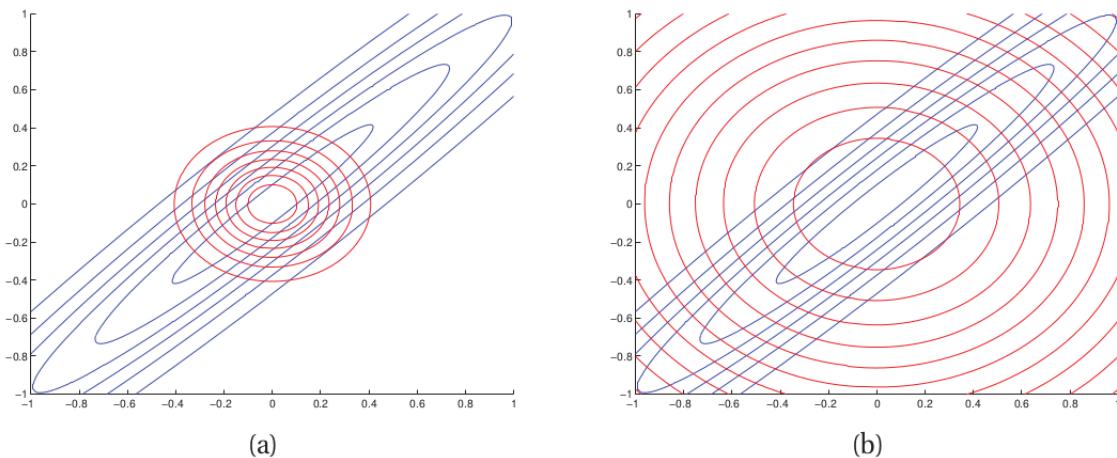


Рис. 4.8: Аппроксимация распределения Q симметричным двумерным гауссовским распределением Q^* . (а) — результат минимизации $KL(Q^* \parallel Q)$, (б) — $KL(Q \parallel Q^*)$

Видно, что при минимизации reverse-KL локальные максимумы аппроксимации Q^* скорее всего будут совпадать с таковыми у Q . При этом мо-

жет быть так, что Q^* будет хорошо приближать Q только на каком-то множестве, которому распределение Q сопоставляет большую вероятность, потому её называют **zero-forcing**. Минимизация же forward KL будет стремиться хорошо объяснить всё распределение Q в целом, а не только отдельные, важные фрагменты, её называют **zero-avoiding**. В обоих случаях наблюдаемое поведение связано с тем, какая плотность оказывается в знаменателе отношения под логарифмом.

Указанная в предыдущем замечании особенность **очень важная**. Она показывает, что VB, в отличие от MCMC, может даже в пределе не давать возможности семплировать из истинного распределения. Тем не менее, VB существенно проще вычислительно, так как это оптимизационная задача, для решения которой есть хорошо разработанные методы стохастической и распределённой оптимизации. Более того, часто одного только VB бывает достаточно. Вопрос о том, когда использовать VB, а когда MCMC, хорошо резюмирует приведенная цитата из [13]:

«Thus, variational inference is suited to large data sets and scenarios where we want to quickly explore many models; MCMC is suited to smaller data sets and scenarios where we happily pay a heavier computational cost for more precise samples. For example, we might use MCMC in a setting where we spent 20 years collecting a small but expensive data set, where we are confident that our model is appropriate, and where we require precise inferences. We might use variational inference when fitting a probabilistic model of text to one billion text documents and where the inferences will be used to serve search results to a large population of users. In this scenario, we can use distributed computation and stochastic optimization to scale and speed up inference, and we can easily explore many different models of the data.»

Замечание 4.2. ELBO можно переписать как $\mathcal{L}(Q^*) = \mathbb{E}_{\varphi \sim Q^*} f(\varphi, \mathbf{X}) - H(Q^*)$, где H — энтропия. Если $Q^* = Q_{\mathbf{X}}$, то левое слагаемое это в

точности величина, которая максимизируется на M-шаге EM-алгоритма. Это сходство неслучайно и поясняется в [13]: «*Unlike variational inference, EM assumes the expectation over posterior distribution of latent variables is computable and uses it in otherwise difficult parameter estimation problems. Unlike EM, variational inference does not estimate fixed model parameters — it is often used in a Bayesian setting where classical parameters are treated as latent variables. Variational inference applies to models where we cannot compute the exact conditional of the latent variables.*»

Чаще всего \mathcal{Q} выбирают таким, чтобы распределения из него факторизуются по параметрам, т.е. что

$$q^*(\theta) = \prod_{i=1}^m q_i^*(\theta_i)$$

где m это число параметров. Это т.н. **mean field approximation** — метод, вдохновлённый т.н. моделью Изинга из статистической физики. В [13] (стр. 9-10) описан эффективный алгоритм оптимизации маржинальных плотностей q_i^* . Тем не менее, в модели XClone он не используется, т.к. максимизация ELBO происходит автоматически за счёт использования примитивов из Tensorflow.Distributions[14]. При этом $\mathcal{L}(Q^*)$ оптимизируется неявно, посредством максимизации правой части равенства $\mathcal{L}(Q^*) = KL(Q^* \parallel Q) + \log p(\mathbf{X})$, т.к. использование сопряжённых распределений позволяет расписать её в пригодном для покоординатной оптимизации виде.

4.4.4.2 VB в алгоритме XClone

Апостериорные распределения на параметрах модели XClone имеют вид

$$\begin{aligned} & p(\mathbf{Z}, \mathbf{Y}, \boldsymbol{\theta}, \mathbf{m}, \boldsymbol{\gamma} \mid \mathbf{AD}, \mathbf{DP}, \mathbf{RD}) \propto \\ & \propto p(\mathbf{Z}, \mathbf{Y}, \boldsymbol{\theta}, \mathbf{m}, \boldsymbol{\gamma}) p(\mathbf{AD}, \mathbf{DP}, \mathbf{RD} \mid \mathbf{Z}, \mathbf{Y}, \boldsymbol{\theta}, \mathbf{m}, \boldsymbol{\gamma}) = \\ & = p(\mathbf{Z}, \mathbf{Y}, \boldsymbol{\theta}, \mathbf{m}, \boldsymbol{\gamma}) \cdot \underline{p(\mathbf{AD}, \mathbf{DP} \mid \mathbf{Z}, \mathbf{Y}, \boldsymbol{\theta})} \cdot \underline{p(\mathbf{RD} \mid \mathbf{Z}, \mathbf{Y}, \mathbf{m}, \boldsymbol{\gamma})} \end{aligned} \quad (4.18)$$

где функция правдоподобия распадается в произведение по BAF- и RDR-модулям, т.к. матрицу \mathbf{RD} можно по смыслу считать независящей от матриц \mathbf{AD}, \mathbf{DP} по смыслу. Строго говоря, между \mathbf{DP} и \mathbf{RD} есть положительная корреляция, но она не добавляет новой информации в модель, потому ею было решено пренебречь. Скрытые переменные из $\Omega := \{\mathbf{Z}, \mathbf{Y}, \mathbf{m}, \boldsymbol{\gamma}\}$ независимы в совокупности. Апостериорное распределение $p(\Omega | \mathbf{AD}, \mathbf{DP}, \mathbf{RD})$, как следствие, факторизуется по скрытым переменным. В силу того, что в модели всюду используются сопряжённые распределения, апостериорное распределение имеет тот же вид, что и априорное. Тем не менее, явно вычислить его параметры затруднительно, проще выразить VB через примитивы Tensorflow.Distributions[14] и найти ОАМ стохастическим градиентным спуском в пространстве параметров.

Стоит отметить, что ELBO в стат.модели XClone по явной формуле (через матожидание) оптимизировать сложнее, чем его представление через разность логарифма обоснованности и KL-дивергенции.

$$\text{L}(q) = \int q(\Omega) \log p(\mathbf{AD}, \mathbf{DP}, \mathbf{RD} | \Omega) d\Omega - \text{KL}(q(\Omega) \| p(\Omega)) \quad (4.19)$$

Логарифм обоснованности распадается в сумму по BAF- и RDR-модулям.

$$\begin{aligned} & \int q(\Omega) \log p(\mathbf{AD}, \mathbf{DP}, \mathbf{RD} | \Omega) d\Omega = \\ &= \underbrace{\int q(\mathbf{Z}, \mathbf{Y}, \boldsymbol{\theta}) \log p(\mathbf{AD}, \mathbf{DP} | \mathbf{Z}, \mathbf{Y}, \boldsymbol{\theta}) d\mathbf{Z} d\mathbf{Y} d\boldsymbol{\theta}}_{+} + \\ & \underbrace{\int q(\mathbf{Z}, \mathbf{Y}, \mathbf{m}, \boldsymbol{\gamma}) \log p(\mathbf{RD} | \mathbf{Z}, \mathbf{Y}, \mathbf{m}, \boldsymbol{\gamma}) d\mathbf{m} d\boldsymbol{\gamma}}_{(4.20)} \end{aligned}$$

4.4.4.3 Вывод ELBO для BAF-модуля

Утверждение 4.6. Лог-боснованность BAF-модуля можно расписать как:

$$\begin{aligned} \mathbb{E}_{\mathbf{Z}, \mathbf{Y}, \boldsymbol{\theta}} \log p(\mathbf{AD}, \mathbf{DP} \mid \mathbf{Z}, \mathbf{Y}, \boldsymbol{\theta}) &= \\ &= \int q(\mathbf{Z}, \mathbf{Y}, \boldsymbol{\theta}) \log p(\mathbf{AD}, \mathbf{DP} \mid \mathbf{Z}, \mathbf{Y}, \boldsymbol{\theta}) d\mathbf{Z} d\mathbf{Y} d\boldsymbol{\theta} \\ &= \sum_{i=1}^N \sum_{j=1}^M \sum_{k=1}^K \sum_{t=1}^T \left\{ \tilde{\pi}_{j,k} \tilde{u}_{i,k,t} [w_{i,j} + a_{i,j} \psi(\tilde{\alpha}_t) + b_{i,j} \psi(\tilde{\beta}_t) - d_{i,j} \psi(\tilde{\alpha}_t + \tilde{\beta}_t)] \right\} \end{aligned} \quad (4.21)$$

где $b_{i,j} := d_{i,j} - a_{i,j}$, $\tilde{\cdot}$ указывает, что имеется в виду аппроксимация истинной апостериорной вероятности, $w_{i,j} := \log \binom{d_{i,j}}{a_{i,j}}$, $\psi(\cdot)$ — дигамма-функция, $\psi(z) := \frac{\Gamma'(z)}{\Gamma(z)} = -\gamma + \int_0^1 \left(\frac{1-t^z}{1-t} \right) dt$, где Γ это гамма-функция Эйлера, а γ это константа Эйлера-Маскерони, $\gamma := \lim_{n \rightarrow \infty} (-\ln n + \sum_{k=1}^n \frac{1}{k})$.

Доказательство. BAF-модуль эквивалентен модели [15] и имеет аналогичное доказательство.

$$\begin{aligned} &\int q(\mathbf{Z}, \mathbf{Y}, \boldsymbol{\theta}) \log p(\mathbf{AD}, \mathbf{DP} \mid \mathbf{Z}, \mathbf{Y}, \boldsymbol{\theta}) d\mathbf{Z} d\mathbf{Y} d\boldsymbol{\theta} \\ &= \mathbb{E}_{\mathbf{Z}, \mathbf{Y}, \boldsymbol{\theta}} [\log p(\mathbf{AD}, \mathbf{DP} \mid \mathbf{Z}, \mathbf{Y}, \boldsymbol{\theta})] \\ &= \mathbb{E}_{\mathbf{Z}, \mathbf{Y}, \boldsymbol{\theta}} \left[\log \prod_{i=1}^N \prod_{j=1}^M \prod_{k=1}^K \prod_{t=1}^T p(a_{i,j} \mid d_{i,j}, \theta_t)^{z_{j,k} \times y_{i,k,t}} \right] \\ &= \mathbb{E}_{\mathbf{Z}, \mathbf{Y}, \boldsymbol{\theta}} \sum_{i=1}^N \sum_{j=1}^M \sum_{k=1}^K \sum_{t=1}^T [z_{j,k} y_{i,k,t} \log \text{Binom}(a_{i,j} \mid d_{i,j}, \theta_t)] \\ &= \sum_{i=1}^N \sum_{j=1}^M \sum_{k=1}^K \sum_{t=1}^T \mathbb{E}_{\mathbf{Z}, \mathbf{Y}, \boldsymbol{\theta}} [z_{j,k} y_{i,k,t} \log \text{Binom}(a_{i,j} \mid d_{i,j}, \theta_t)] \\ &= \sum_{i=1}^N \sum_{j=1}^M \sum_{k=1}^K \sum_{t=1}^T \mathbb{E}_{\mathbf{Z}} [z_{j,k}] \mathbb{E}_{\mathbf{Y}} [y_{i,k,t}] \mathbb{E}_{\boldsymbol{\theta}} [\log \text{Binom}(a_{i,j} \mid d_{i,j}, \theta_t)] \\ &= \sum_{i=1}^N \sum_{j=1}^M \sum_{k=1}^K \sum_{t=1}^T \left\{ \tilde{\pi}_{j,k} \tilde{u}_{i,k,t} \mathbb{E}_{\boldsymbol{\theta}} [\log \text{Binom}(a_{i,j} \mid d_{i,j}, \theta_t)] \right\} \\ &= \sum_{i=1}^N \sum_{j=1}^M \sum_{k=1}^K \sum_{t=1}^T \left\{ \tilde{\pi}_{j,k} \tilde{u}_{i,k,t} [w_{i,j} + a_{i,j} \psi(\tilde{\alpha}_t) + b_{i,j} \psi(\tilde{\beta}_t) - d_{i,j} \psi(\tilde{\alpha}_t + \tilde{\beta}_t)] \right\} \end{aligned}$$

Где матожидание логарифма плотности бета-биномиального распределения

ния, имеет следующие представление, задействующее дигамма-функцию:

$$\begin{aligned}
 \mathbb{E}_{\alpha,\beta}[\log q(a \mid d, \theta)] &= \\
 &= \mathbb{E}_{\alpha,\beta}[\log \text{Binom}(a; d, \theta)] = \\
 &= \mathbb{E}_{\alpha,\beta} \left[\log \binom{a}{d} + a \log \theta + (d-a) \log (1-\theta) \right] = \\
 &= \log \binom{a}{d} + a \mathbb{E}_\theta[\log \theta] + (d-a) \mathbb{E}_{\beta,\alpha}[\log (1-\theta)] = \\
 &= \log \binom{a}{d} + a(\psi(\alpha) - \psi(\alpha+\beta)) + (d-a)(\psi(\beta) - \psi(\alpha+\beta)) = \\
 &= \log \binom{a}{d} + a\psi(\alpha) + (d-a)\psi(\beta) - d\psi(\alpha+\beta)
 \end{aligned}$$

Где тождество $\mathbb{E}_{\alpha,\beta}[\log \theta] = \psi(\alpha) - \psi(\alpha+\beta)$ доказывается так:

$$\begin{aligned}
 \mathbb{E}_{\alpha,\beta} \log \theta &= \int_0^1 \ln x \text{Beta}(x; \alpha, \beta) dx = \\
 &= \int_0^1 \ln x \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\text{B}(\alpha, \beta)} dx = \\
 &= \frac{1}{\text{B}(\alpha, \beta)} \int_0^1 \frac{\partial x^{\alpha-1}(1-x)^{\beta-1}}{\partial \alpha} dx = \\
 &= \frac{1}{\text{B}(\alpha, \beta)} \frac{\partial}{\partial \alpha} \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx = \\
 &= \frac{1}{\text{B}(\alpha, \beta)} \frac{\partial \text{Beta}(\alpha, \beta)}{\partial \alpha} dx = \\
 &= \frac{\partial \ln \text{B}(\alpha, \beta)}{\partial \alpha} = \\
 &= \frac{d \ln \Gamma(\alpha)}{d\alpha} - \frac{\partial \ln \Gamma(\alpha+\beta)}{\partial \alpha} = \\
 &= \psi(\alpha) - \psi(\alpha+\beta)
 \end{aligned}$$

□

Замечание 4.3. Отдельный интерес представляет подход к вычислению величины $w_{i,j} = \log \binom{d_{i,j}}{a_{i,j}}$. Дело в том, что числа с плавающей точкой в реализации большинства языков программирования имеют ограниченную точность, которая не позволяет отдельно вычислить биномиальный

коэффициент, а потом взять от него логарифм. Даже в языке Python, в котором реализована длинная арифметика в целых числах, взятие логарифма может привести к переполнению типа при больших значениях $d_{i,j}$ и $a_{i,j} \simeq \frac{1}{2}d_{i,j}$.

Безусловно, можно вычислять $\log \binom{d_{i,j}}{a_{i,j}}$ как

$$\sum_{s=0}^{a_{i,j}-1} \log(d_{i,j} - s) - \sum_{s=0}^{a_{i,j}} \log s$$

Тем не менее, при большой разнице между $d_{i,j}$ и $a_{i,j}$ при таком подходе ошибки представления логарифмов накапливаются, что может привести к заметному расхождению подсчитанной величины с истинным значением.

Этих недостатков лишена красавая аппроксимация величины $\log n!$, полученная логарифмированием асимптотической формулы для $n!$, полученной великим математиком Сринивасой Рамануджаном[16] и уточняющей формулу Стирлинга.

$$\log n! \simeq n \log n - n + \frac{\log(n(1 + 4n(1 + 2n)))}{6} + \frac{\log \pi}{2}$$

Именно эта формула используется при вычислении $w_{i,j}$ в алгоритме XClone.

4.4.4.4 Вывод ELBO для RDR-модуля

Утверждение 4.7. Лог-обоснованность RDR-модуля можно записать как

$$\begin{aligned} & \mathbb{E}_{\mathbf{Z}, \mathbf{Y}, \mathbf{m}, \boldsymbol{\gamma}} \log p(\mathbf{RD} \mid \mathbf{Z}, \mathbf{Y}, \mathbf{m}, \boldsymbol{\gamma}) = \\ &= \int q(\mathbf{Z}, \mathbf{Y}, \mathbf{m}, \boldsymbol{\gamma}) \log p(\mathbf{RD} \mid \mathbf{Z}, \mathbf{Y}, \mathbf{m}, \boldsymbol{\gamma}) d\mathbf{m}, d\boldsymbol{\gamma} = \\ &= \sum_{i=1}^N \sum_{j=1}^M \sum_{k=1}^K \tilde{\pi}_{j,k} \cdot r_{ij} \cdot \mathbb{E}_{\mathbf{m}, \boldsymbol{\gamma}} \log \tilde{f}_{i,k} + C \end{aligned} \tag{4.22}$$

где $\tilde{\cdot}$ указывает, что имеется в виду аппроксимация истинной апостери-

орной вероятности, а

$$\tilde{f}_{i,k} \mid \mathbf{m}, \boldsymbol{\gamma} := \frac{\sum_{t=1}^T m_i \exp[\gamma_t] \tilde{u}_{i,k,t}}{\sum_{b=1}^N \sum_{t=1}^T m_b \exp[\gamma_t] \tilde{u}_{b,k,t}}$$

Доказательство.

$$\begin{aligned} & \int q(\mathbf{Z}, \mathbf{Y}, \mathbf{m}, \boldsymbol{\gamma}) \log p(\mathbf{RD} \mid \mathbf{Z}, \mathbf{Y}, \mathbf{m}, \boldsymbol{\gamma}) d\mathbf{m}, d\boldsymbol{\gamma} = \\ &= \mathbb{E}_{\mathbf{Z}, \mathbf{Y}, \mathbf{m}, \boldsymbol{\gamma}} \log p(\mathbf{RD} \mid \mathbf{Z}, \mathbf{Y}, \mathbf{m}, \boldsymbol{\gamma}) = \\ &= \text{см. формулу 4.14} = \\ &= \mathbb{E}_{\mathbf{Z}, \mathbf{Y}, \mathbf{m}, \boldsymbol{\gamma}} \left[\sum_{j=1}^M \sum_{k=1}^K \log p \left(\mathbf{r}_j \mid \sum_{i=1}^N r_{i,j}, \mathbf{f}_k \right)^{z_{j,k}} \right] = \\ &= \sum_{j=1}^M \sum_{k=1}^K \mathbb{E}_{\mathbf{Z}, \mathbf{Y}, \mathbf{m}, \boldsymbol{\gamma}} \left[z_{j,k} \cdot \log \left(r_j! \prod_{i=1}^N \frac{\tilde{f}_{i,k}^{r_{i,j}}}{r_{i,j}!} \right) \right] = \\ &= \text{обозначим сумму всех констант через } C = \\ &= \sum_{i=1}^N \sum_{j=1}^M \sum_{k=1}^K \mathbb{E}_{\mathbf{Z}, \mathbf{Y}, \mathbf{m}, \boldsymbol{\gamma}} \left[z_{j,k} \cdot r_{ij} \log \tilde{f}_{i,k} \right] + C = \\ &= \sum_{i=1}^N \sum_{j=1}^M \sum_{k=1}^K \tilde{\pi}_{j,k} \cdot r_{ij} \cdot \mathbb{E}_{\mathbf{m}, \boldsymbol{\gamma}} \log \tilde{f}_{i,k} + C \end{aligned}$$

□

4.4.5 Известные недостатки и планы по их исправлению

4.4.5.1 Избыточная сложность исходной модели

Эксперименты показали, что модель можно упростить:

- В коде XClone \mathbf{m} фиксировано и выводится из теоретических соображений о равномерном распределении прочтений по геному. Т.е. $m_i = \frac{l_i}{\sum_{b=1}^N l_b}$ это достаточно хорошая оценка матожидания доли прочтений в сегменте i , где l_i это длина i -го блока в сегментации.

- Далее, среднее значение γ тоже фиксировано, $\mathbb{E}\gamma_t := \log \frac{c_{t,m} + c_{m,2}}{2} - \frac{e^{l_1^2}}{2}$, благодаря чему $\mathbb{E} \exp(\gamma_t) = \frac{c_{t,m} + c_{m,2}}{2}$, где l_1 это гиперпараметр ковариационной функции гауссовского процесса γ . По изначальной задумке, по которой XClone можно было применять и для scRNA-seq данных, переменное γ было нужно для того, чтобы учесть не связанные с числом копий причины аллельного-дисбаланса в транскриптомных данных.

После того, как стало ясно, что транскриптомные данные этой же моделью обработать не получится из-за шумного BAF-сигнала и более сложной природы RDR-сигнала (в силу разреженности данных), потребность в переменном γ отпала. Более того, в силу того, что соотв. KL-слагаемое давало несущественный вклад в функцию ошибки, распределение на γ буквально за несколько итераций вырождалось и теряло содержательный смысл.

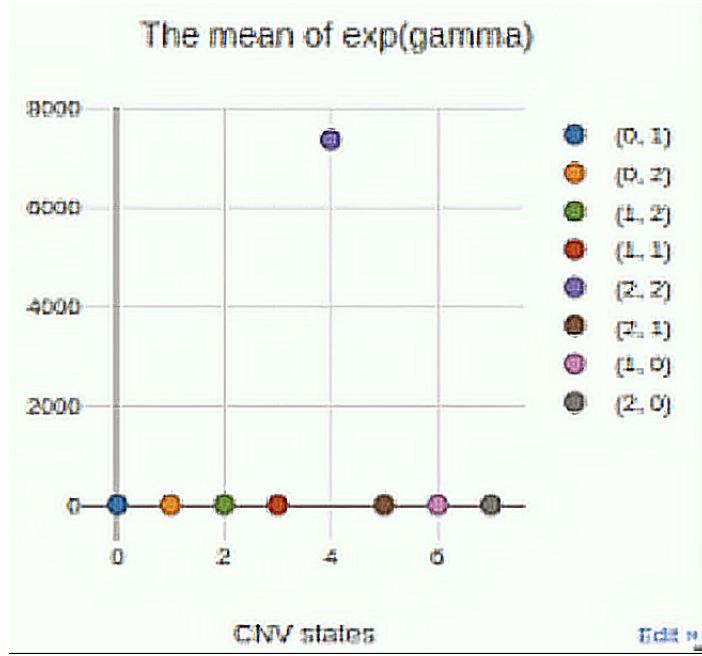


Рис. 4.9: Вырождение апостериорного распределения. Ось Ох — координаты, ось Оу — матожидание вдоль компоненты. До того, как было принято решение зафиксировать γ_t , распределение на γ буквально за пару итераций ВВ вырождалось в одномерное.

Такое поведение можно было подавить умножением соотв. KL-слагаемого на большой множитель (порядка 10^6), но такое преобразование было лишено смысла.

- Аналогично, связь между BAF- и RDR-модулями через \mathbf{Z} и \mathbf{Y} оказалась слишком слабой. Из-за этого получалась абсурдная ситуация, когда BAF-модуль выводил правильные θ , но они были никак не согласованы с ASCNV, которые выводил RDR-модуль. Как следствие было решено избавиться от первой размерности: $\forall i, j \forall t : \theta_{i,t} = \theta_{j,t}$, $\mathbb{E}\theta_{i,t} = \frac{c_{t,m}}{c_{t,m} + c_{t,p}}$, где $=$ означает равенство по распределению, и зафиксировать параметры α_t и β_t .

В самом деле, блок-зависимые BAF-ы имели смысл для транскриптомных данных, где картина экспрессии зависела от состояния клет-

ки в момент секвенирования, что искажало сигнал аллельного дисбаланса. После того, как стало ясно, что VB будет применяться только к геномным данным, потребность в блок-специфичных BAF-ах отпала сама собой.

Более того, невооружённым взглядом видно, что BAF-модуль гораздо проще, чем RDR-модуль. Более того, гетерозиготные ОНП составляют малую долю от всего генома, и примерно половины прочтений теряется при переходе от \mathbf{RD} к \mathbf{DP} . Как следствие, обоснованность RDR-модуля была на несколько порядков больше обоснованности BAF-модуля, да и оптимизировать её было значительно трудней. BAF-модуль быстро переобучался, буквально за несколько десятков итераций, после чего мешал VB находить оптимальные параметры RDR-модуля.

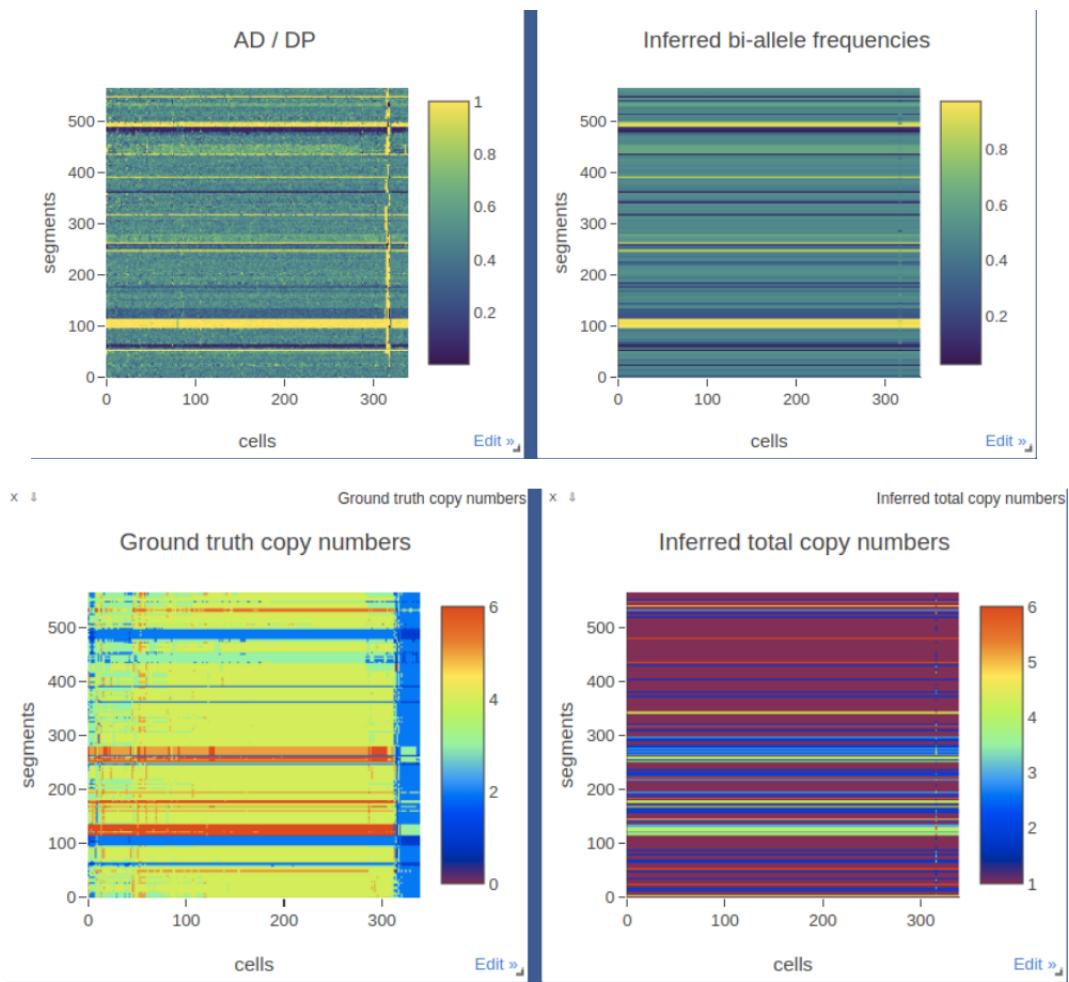


Рис. 4.10: Несогласованность найденных VB параметров при запуске на реальных данных при использовании блок-зависимых BAF-ов (на верхних тепловых картах). Слева — ожидаемые значения, справа — предсказанные. Видно, как BAF-модуль переобучается, а RDR-модуль хоть и улавливает какие-то закономерности, но, в целом, предсказывает что-то не то (даже по модулю того, что XClone не умеет детектировать whole-genome duplication).

Но как только частоты аллелей в блоках стали, по сути, определяться исключительно расстановкой наиболее вероятных Y и Z , BAF-модуль перестал переобучаться на шум в данных и изменения параметров модулей стали согласованными.

- В ходе критического переосмысления алгоритма, сопровождавшего

написание текста данной дипломной работы, стали очевидны проблемы с определением ключевой для RDR-модуля величины

$$f_{i,k} = \frac{\sum_{t=1}^T m_i \exp[\gamma_t] u_{i,k,t}}{\sum_{b=1}^N \sum_{t=1}^T m_b \exp[\gamma_t] u_{b,k,t}}$$

Сама форма знаменателя делает величины $\{f_{1,k}, \dots, f_{N,k}\}$ попарно зависимыми, хотя из модели ДНК-секвенирования из раздела 4.1 это никак не следует. Более того, число прочтений в каком-то сегменте по смыслу не может зависеть от числа прочтений на других хромосомах если речь идёт о ДНК, а не РНК, где между уровнями экспрессии участков генома есть сложные регуляторные зависимости. Эта формула не лишена и многих других проблем: к примеру, сейчас, когда стало понятно, что \boldsymbol{m} и $\boldsymbol{\gamma}$ можно и нужно зафиксировать, непонятно, зачем брать им матожидание по распределению γ . Это существенно замедляет обучение модели, хотя из теоретических соображений ясно, что при достаточно большой глубине покрытия (которая имеет место в данных ДНК-секвенирования) отклонения RDR от $\frac{c_{t,m} + c_{t,p}}{2}$ должны быть пренебрежимо малы. Это наблюдение подтверждается экспериментами. Как следствие, была предложена идея сделать RDR-модуль в некотором смысле "симметричным" BAF-модулю, а именно: моделировать каждый блок b сегментации отдельно от остальных и описывать его RDR отрицательным биномиальным распределением, параметр p которого в конкретной клональной линии тоже должен получаться амплификацией параметра p_0 из статистической модели нормальных клеток. Сами коэффициенты амплификации можно не моделировать гауссовским процессом, а просто положить равными $\frac{c_{t,m} + c_{t,p}}{2}$. Лог-обоснованность такой модели будет выражаться через матожидание по апостериорному распределению \mathbf{Y} , и это гораздо более интуитивный подход, чем тот, что использует-

ся сейчас. Тем не менее, все эти недостатки были обнаружены уже в процессе написания текста ВКР, потому альтернативный RDR-модуль в ней не представлен, но обязательно будет реализован при дальнейшей разработке метода.

4.4.5.2 Численное интегрирование в оценке обоснованности актуальной версии RDR-модуля Аналитическую формулу для обоснованности RDR-модуля, увы, получить не удалось. Даже если фиксировать \mathbf{m} , неясно, как устроено отношение сумм зависимых логнормальных распределений в $\tilde{f}_{i,k}$. В [17] показано, что даже распределение числителя и знаменателя в отдельности имеет сложную структуру и не выражается через элементарные операции над стандартными распределениями. Как следствие, в коде алгоритма RDR-обоснованность приближается интегрированием по Монте-Карло.

Это существенный недостаток XClone в текущей редакции, т.к. метод Монте-Карло в общем случае имеет корневую сходимость⁹, и для получения точности хотя бы до третьего знака приходится использовать порядка 10^6 точек, и это при фиксированном \mathbf{m} . В противном случае, пришлось бы семплировать ещё больше точек, чтобы сделать поправку на размерность пространства признаков (от нескольких сотен до нескольких тысяч).

Более того, в данных ДНК-секвенирования и нет особых причин вводить распределение на γ . Положить фактор амплификации γ_t равным $(c_{t,m} + c_{t,p})/2$ оказалось достаточно для того, чтобы заметно улучшить качество алгоритма как на синтетических, так и на реальных данных. Это в итоге и сподвигло пересмотреть устройство RDR-модуля, о чём

⁹Конспект лекции в MIT: https://ocw.mit.edu/courses/mechanical-engineering/2-086-numerical-computation-for-mechanical-engineers-fall-2014/nutshells-guis/MIT2_086F14_Monte_Carlo.pdf

подробно написано в заключительном разделе данной ВКР.

4.4.5.3 Слишком большая разница масштабов отдельных слагаемых в ELBO

Вариационный байесовский вывод в модели XClone основан на следующем представлении ELBO:

$$L(q) = \int q(\Omega) \log p(\mathbf{AD}, \mathbf{DP}, \mathbf{RD} \mid \Omega) d\Omega - KL(q(\Omega) \parallel p(\Omega)) \quad (4.23)$$

При этом логарифм обоснованности распадается в сумму по BAF- и RDR-модулям:

$$\begin{aligned} & \int q(\Omega) \log p(\mathbf{AD}, \mathbf{DP}, \mathbf{RD} \mid \Omega) d\Omega = \\ &= \underbrace{\int q(\mathbf{Z}, \mathbf{Y}, \boldsymbol{\theta}) \log p(\mathbf{AD}, \mathbf{DP} \mid \mathbf{Z}, \mathbf{Y}, \boldsymbol{\theta}) d\mathbf{Z} d\mathbf{Y} d\boldsymbol{\theta}}_{+} + \\ & \quad \underbrace{\int q(\mathbf{Z}, \mathbf{Y}, \mathbf{m}, \boldsymbol{\gamma}) \log p(\mathbf{RD} \mid \mathbf{Z}, \mathbf{Y}, \mathbf{m}, \boldsymbol{\gamma}) d\mathbf{m} d\boldsymbol{\gamma}}_{(4.24)} \end{aligned}$$

а KL-дивергенция распадается на сумму по скрытым переменным:

$$\begin{aligned} KL(q(\Omega) \parallel p(\Omega)) &= KL(q(\mathbf{Z}) \parallel p(\mathbf{Z})) + KL(q(\mathbf{Y}) \parallel p(\mathbf{Y})) + \\ & \quad + KL(q(\mathbf{m}) \parallel p(\mathbf{m})) + KL(q(\boldsymbol{\gamma}) \parallel p(\boldsymbol{\gamma})) \end{aligned} \quad (4.25)$$

При этом оптимальные параметры моделищаются градиентным методом (Adam[18], если быть точным). Метод не может знать о содержательном смысле параметров. Алгоритм будет в первую очередь оптимизировать те параметры, которые дают наибольший вклад в ELBO. На практике соотношение между компонентами ELBO выглядит так:

```
Total loss: 2619717888.0
Total KL: 2641.35986328125.
Observed BAF-CNV discrepancy: 1.2505922317504883
    - CNV state assignment KL: 2268.0888671875
    - Clonal label assignment KL: 373.2406005859375
    - ASE KL: 0.03028840757906437
    - RDR KL: 0.0
Total log-likelihood: -2619451136.0
    - ASE log-likelihood: -2037201.625
    - RDR log-likelihood: -2617330176.0
```

Рис. 4.11: Величины различных слагаемых в разложении ELBO, логи модели в процессе обучения. Видно, что лог-обоснованность RDR-модуля по модулю на три порядка больше, чем аналогичная величина для BAF-модуля. Величины компонент KL-дивергенции пренебрежимо малы по сравнению с модулем лог-обоснованности модели.

Очевидный дисбаланс в величинах слагаемых приводит к тому, что алгоритм оптимизации практически не обращает внимание на KL-дивергенцию, если априорные предположения о распределениях не слишком сильные (читай, если априорные распределения не близки к вырожденным). Это приводит к тому, что апостериорные аспределения на параметрах иногда вырождаются только потому, что это почему-то позволяет сильно уменьшить лог-обоснованность одного из модулей. В силу зашумленности реальных данных, такое вырождение может происходить случайно. Это существенно влияет на интерпретируемость результатов.

Кроме того, лог-обоснованность RDR-модуля доминирует над всеми остальными слагаемыми. Это не так существенно, когда \boldsymbol{m} и $\boldsymbol{\gamma}$ фиксированы, а величины $\theta_{i,t}$ в BAF-модуле не являются блок-специфичными, но может быть проблемой в будущем, при добавлении новых модулей в модель (скажем, для учёта соматических мутаций или митохондриальных геномов). Тем не менее, такой дисбаланс вызван техническими фактами: элементы матрицы \mathbf{RD} обычно заметно больше соотв. элементов

матриц \mathbf{AD} и \mathbf{DP} , но при этом «больше» не означает «важнее».

Для борьбы с этим эффектом кажется уместным либо приводить элементы \mathbf{RD} и \mathbf{DP} к одному масштабу некоторым целочисленным преобразованием, либо ввести веса для компонент ELBO, хоть это и сугубо практический трюк, который нарушает логическую стройность теоретических выкладок. Также можно попробовать заменить градиентный метод Adam на квазиньютоновский L-BFGS[19]. Тем не менее, планирование эксперимента, который бы показал, что один из методов лучше другого, это отдельная нетривиальная задача, к которой пока непонятно, как подступиться.

4.4.5.4 Концептуальная невозможность детектирования WGD

Коммерческий алгоритм CellRanger не способен надёжно определить WGD — мутацию, при которой каждая хромосома в клетке удваивается, — в чём представители компании честно признаются в официальной документации¹⁰.

¹⁰Раздел "Determining absolute copy numbers" по ссылке <https://support.10xgenomics.com/single-cell-dna/software/pipelines/latest/interpret/overview>

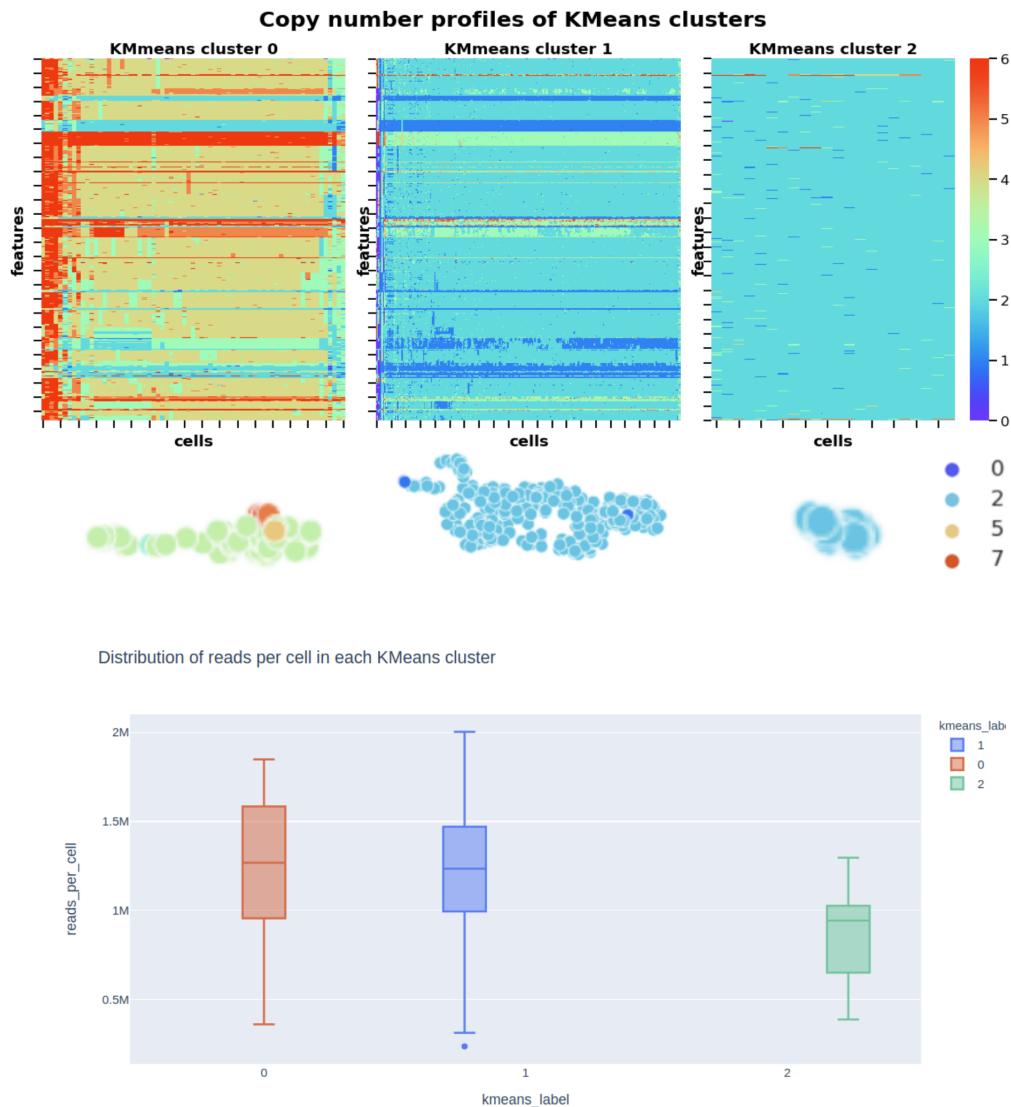


Рис. 4.12: Три основных клональных линии в образцах медуллобластомы с точки зрения алгоритма CellRanger. 0 — тетраплоидные опухолевые клетки, 1 — диплоидные опухолевые клетки, 2 — здоровые клетки. Тепловые карты показывают предсказанное CellRanger число копий. Ящики с усами показывают распределение числа прочтений в клетках каждой клональной линии.

Из тепловых карт следует, что медиана класса 0 должна быть в два раза выше примерно одинаковых медиан классов 1 и 2. Тем не менее, видно, что это не так: ящики для классов 0 и 1 практически совпадают. Более

того, биологи, работавшие с этой опухолью, говорят, что есть основания полагать, что большая часть опухолевых клеток образца тетраплоидная.

Алгоритм XClone, как и CellRanger, не замечает кратное увеличение глубины покрытия во всём геноме сразу, которое имеет место при WGD. Это связано с тем, что RDR-модуль использует мультиномиальную модель. Если в формуле

$$f_{i,k} = \frac{\sum_{t=1}^T m_i \exp[\gamma_t] u_{i,k,t}}{\sum_{b=1}^N \sum_{t=1}^T m_b \exp[\gamma_t] u_{b,k,t}}$$

заменить каждое $\exp(\gamma_t)$ на $t \exp(\gamma_t)$ для любого $t \in \mathbb{N} \setminus \{0\}$, то величина $f_{i,k}$ не изменится, т.к. t в числителе и знаменателе сократится. Описанная в разделе 4.4.5.1 модель, в которой блоки описываются отрицательным биномиальным распределением, лишена этой проблемы. Похожий подход хорошо зарекомендовал себя в алгоритме CHISEL[3], потому есть основания полагать, что он сработает и здесь, но экспериментально это пока что проверено не было.

STP-Nuclei. CHISEL CNV heatmap (additional evidence)

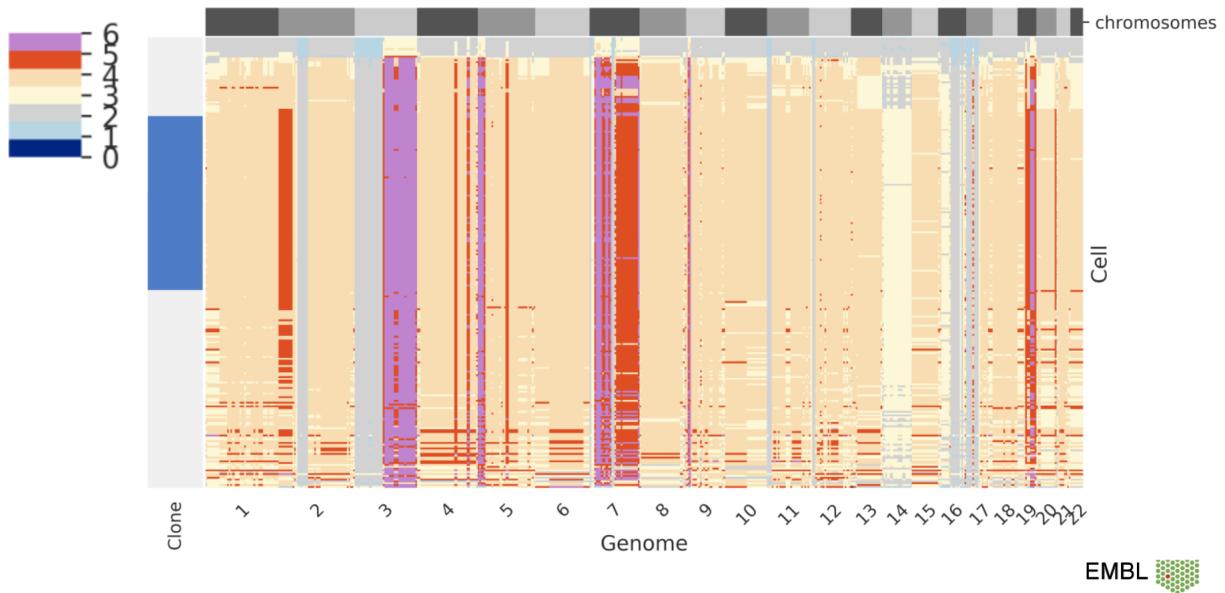


Рис. 4.13: Тепловая карта числа копий в блоках длины 5 мегабаз с точки зрения алгоритма CHISEL. Сверху — полоса нормальных клеток. Остальные клетки тетраплоидные, что согласуется с экспериментальными данными лучше, чем результаты алгоритма CellRanger.

4.4.5.5 Отсутствие масштабирования клеток в RDR-модуле

В силу того, как устроены формулы для лог-обоснованности модулей XClone, каждая клетка образца вносит тем больший вклад, чем больше её глубина покрытия. Тем не менее, здесь больше не значит лучше, т.к. отличие может быть следствием чисто технических причин.

Следовало бы приводить вектора прочтений клеток к одному масштабу. Это стандартная практика, которая используется, например, в scVI[20] — нейросетевом методе обработки данных single-cell секвенирования, тоже основанном на вариационном байесовском выводе, а также в т.н. latent factor mixed models — линейных моделях со скрытыми переменными, которые часто используют в генетике (например, LIMIX[21]).

4.4.5.6 Необходимость вручную задавать ожидаемое число клonalных линий в образце

В данный момент K — сколько клonalных линий искать в образце — это гиперпараметр модели. Это не проблема, когда у пользователя есть обоснованные гипотезы о составе опухоли, но так бывает не всегда. Было бы лучше, если бы этот параметр определялся автоматически с помощью байесовского метода "Tree-structured Parzen Estimator (TPE)"[22] из библиотеки hyperopt¹¹.

4.4.5.7 Отсутствие явной связи с деревом онкогенеза

В данный момент явным образом не используется то, что онкогенез описывается деревом, где каждое ветвление соответствует важной наследуемой мутации. Если бы алгоритм XClone моделировал эту структуру явно, т.е. обрабатывал каждый уровень дерева по отдельности: сначала отделял опухолевые клетки от нормальных, потом находил грубые клональные линии, потом рекурсивно уточнял каждую из них — это бы повысило интерпретируемость результатов.

4.5 Использованные данные

Работа над XClone была мотивирована наличием данных секвенирования образцов медуллобластомы — детской опухоли мозга, — недавно извлечённых хирургическим путём в одной из клиник Штутгарта. Из предварительного анализа данных bulk-секвенирования было известно, что у этого пациента эволюция опухоли сопровождалась масштабными структурными вариациями. Были обнаружены делеции и дупликации больших участков генома, а также хромотприсис на хромосоме 7. Кроме того, у пациента случился рецидив и был повод предполагать, что на

¹¹<http://jaberg.github.io/hyperopt/>

какие-то из клональных линий лечение не действует. Сам пациент при этом перенёс ещё одну операцию, потому в распоряжении биологов были две опухоли: **primary** (до лечения) и **relapse** (после).

Для детекции структурных вариаций нужны данные ДНК-секвенирования, которые были только для первичной опухоли, поэтому данные из relapse-опухоли в данной работе не рассматриваются. Более того, от секвенировать relapse-опухоль уже нельзя, потому что она уже полностью израсходована.

Для первичной опухоли доступны следующие образцы:

- **STP-Nuclei** — *Stuttgart Primary Tumour, single-Nuclei-seq* — ДНК- и РНК-образец, полученные по технологии single nucleus sequencing. В них секвенировалась не вся клетка целиком, а только извлечённое из неё ядро. Такой подход часто используют при секвенировании нейронных клеток, т.к. выделить их ядра, как оказалось, проще, чем разделить ткань на клетки (ведь нейрон имеет множество отростков-дendritov, из-за которых клетки крепко "держатся" друг за друга). В ДНК-образце 404 ядра, а в РНК-образце — порядка 6000 (в зависимости от выбранного порога качества). Именно этот образец был выбран для валидации XClone-V2, потому с его подробным описанием можно ознакомиться в разделе "Результаты".
- **STP-G&T** — *Stuttgart Primary Tumour, Genome and Transcriptome* — геном и транскриптом, извлечённые из одних и тех же клеток. Секвенирование по такой технологии требует высокого профессионализма. Следствие этого — высокая цена и малая пропускная способность метода: в образцах всего 96 клеток. Тем не менее, они гораздо менее разреженные, чем в образце STP-Nuclei. Наличие таких данных в перспективе могло бы стать конкурентным преимуществом.

ством. Если метод обнаруживает корректную клональная структуру в ДНК-данных, то можно попробовать сопоставить клетки РНК-образца найденным клональным линиям и измерить точность предсказания. Это уникальная ситуация, т.к. клональные линии клеток из обычных scRNA-seq образцов заранее неизвестны и можно лишь строить гипотезы о корректности результатов предсказания. После того, как стало ясно, что стабильный ВАF-сигнал из РНК-данных извлечь не получается, эти данные отошли на второй план, но ещё сыграют важную роль при валидации XClone-V3. При этом естественное желание слить ДНК-данные из этого образца с ДНК-данными из STP-Nuclei удовлетворить нельзя, т.к. для секвенирования были использованы разные платформы, а потому распределения прочтений в клетках сильно различаются.

- **STP-PDX** — *Stuttgart Patient Tumour, Patient-Derived Xenograft* — ДНК- и РНК-данные, полученные из т.н. mouse model — опухоль пересаживается в организм лабораторной мыши или крысы с искусственно вызванным иммунодефицитом. Это позволяет опухоли дюрасти до нужного для экспериментов размера вне тела пациента. Эти данные пока не были использованы, т.к. ДНК-образец в них хуже, чем в STP-Nuclei, а объединить два образца в один нельзя, т.к. STP-PDX тоже были получены при помощи секвенирования на разных платформах.
- Bulk-секвенирование длинными прочтениями — данные ДНК-секвенирования, полученные с помощью платформы Oxford Nanopore. Были попытки использовать их для уточнения статистического фазирования гаплотипов, но в итоге был сделан выбор в пользу более простого подхода.

5 Полученные результаты

5.1 Синтетические данные

Обе версии алгоритма XClone хорошо показали себя на синтетических данных, полученных в соответствии с генеративными моделями. Все эксперименты воспроизводимы, т.к. при запуске обязательно нужно указать **random seed** — строку, хэш от которой инициализирует генераторы псевдослучайных чисел в коде.

Приведенные графики соответствуют валидации XClone-V1 на синтетических данных. Эксперимент был запущен с такими параметрами:

- И в ДНК-, и в РНК-модуле по 100 клеток. Это имитация протокола G&T[39]: геном и транскриптом будто бы извлекается из одних и тех же клеток;
- Вектора \mathbf{D}_j^G и \mathbf{D}_j^E берутся из реальных данных;
- 7 клональных линий, существенно отличающихся друг от друга картины аллельного дисбаланса — векторами \mathbf{X}_k ;
- \mathbf{A}_j^G — вектора числа прочтений материнских аллелей — результат поэлементного перемножения \mathbf{D}_j^G и \mathbf{X}_k . Аналогично для \mathbf{A}_j^E .
- 10^4 итераций семплирования по Гиббсу. Для стабилизации распределений на метках обычно хватало нескольких тысяч итераций, так что десяти тысяч хватало с большим запасом.

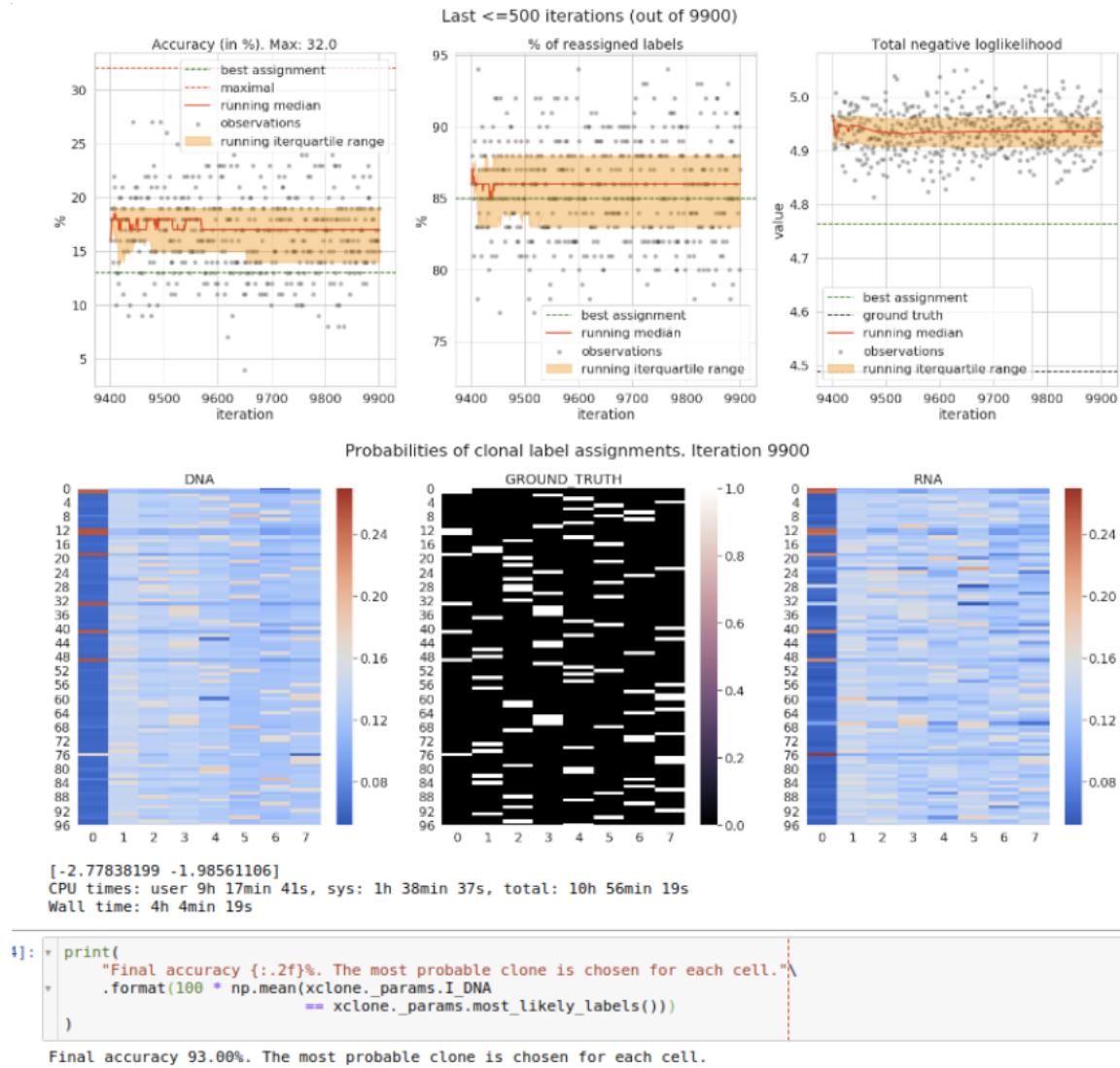


Рис. 5.1: Рукописная визуализация процесса обучения XClone-V1, последние 500 итераций. Строкам тепловых карт в нижнем ряду соответствуют клетки, столбцам — клональные линии, а ячейке на позиции (j, k) — вероятность того, что клетка j принадлежит линии k . Центральная карта показывает истинные клональные линии, левая — вероятности, подсчитанные по scDNA-seq, правая — по scRNA-seq. Верхние тепловые карты позволяют, как распределения из нижних тепловых карт менялись в ходе обучения. Левая карта показывает точность предсказания, если клональная линия назначается случайно в соответствии с текущим распределением вероятностей на метках. Центральная карта показывает, какой процент меток при таком подходе отличается от расстановки, полученной 100 итераций назад. Правая карта показывает, как менялся минус логарифм функции правдоподобия от итерации к итерации.

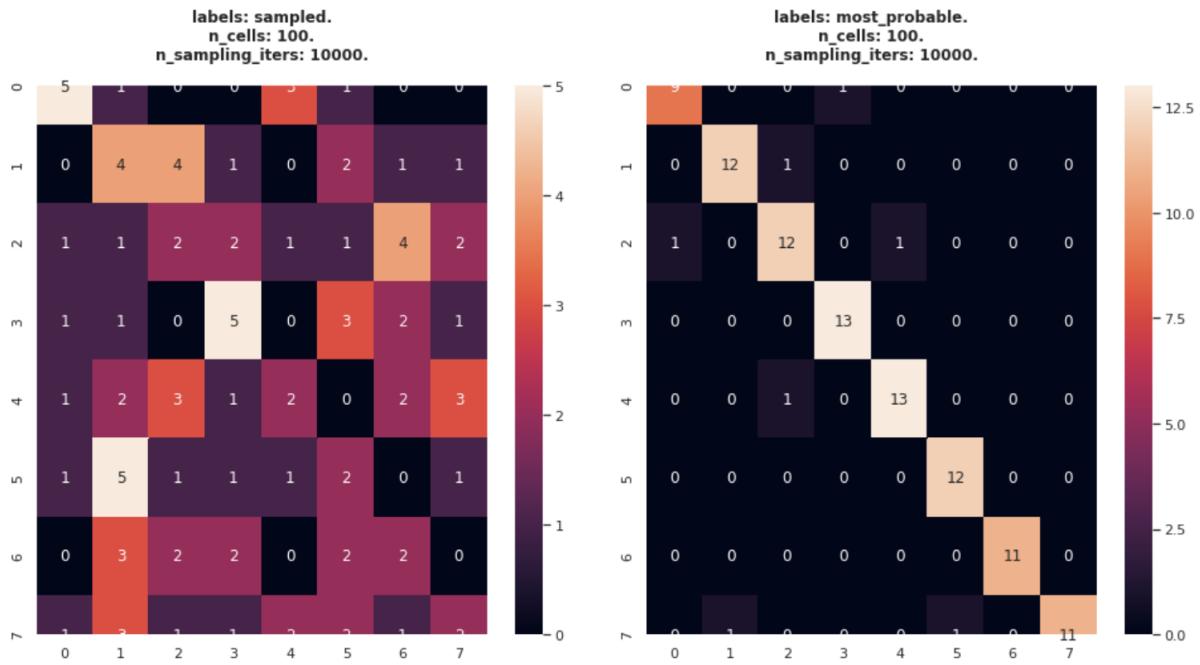


Рис. 5.2: Матрицы ошибок (confusion matrix) для предсказанных клonalных линий. Слева — метки, полученные семплированием, справа — моды соответствующих распределений.

По графикам видно, что алгоритм находит правильные метки, но не очень уверен в них. С ростом глубины покрытия при генерации клеток эта проблема становилась менее выраженной, но и синтетические данные при этом всё меньше напоминали реальные. Симуляция показала, что для надёжной работы алгоритма нужны были РНК-данные гораздо более высокого качества, чем в образцах от DKFZ. Кроме того, итоговая точность сильно зависела от качества ВАФ-сигнала, который в РНК-образцах раз за разом оказывался слишком зашумленным. Именно в связи с этим и была разработана XClone-V2, где во главе угла не ВАФ, а RDR.

Эксперименты для валидации XClone-V2 устроены аналогично, разве что добавляется информация о числе всех прочтений R_j . Кроме того, теперь нужно генерировать ASCNV. Доля затронутых ими блоков сегментации

в каждом из клонов задаётся пользователем, равно как и частоты отдельных аллель-специфических конфигураций ($c_{t,m}, c_{t,p}$).

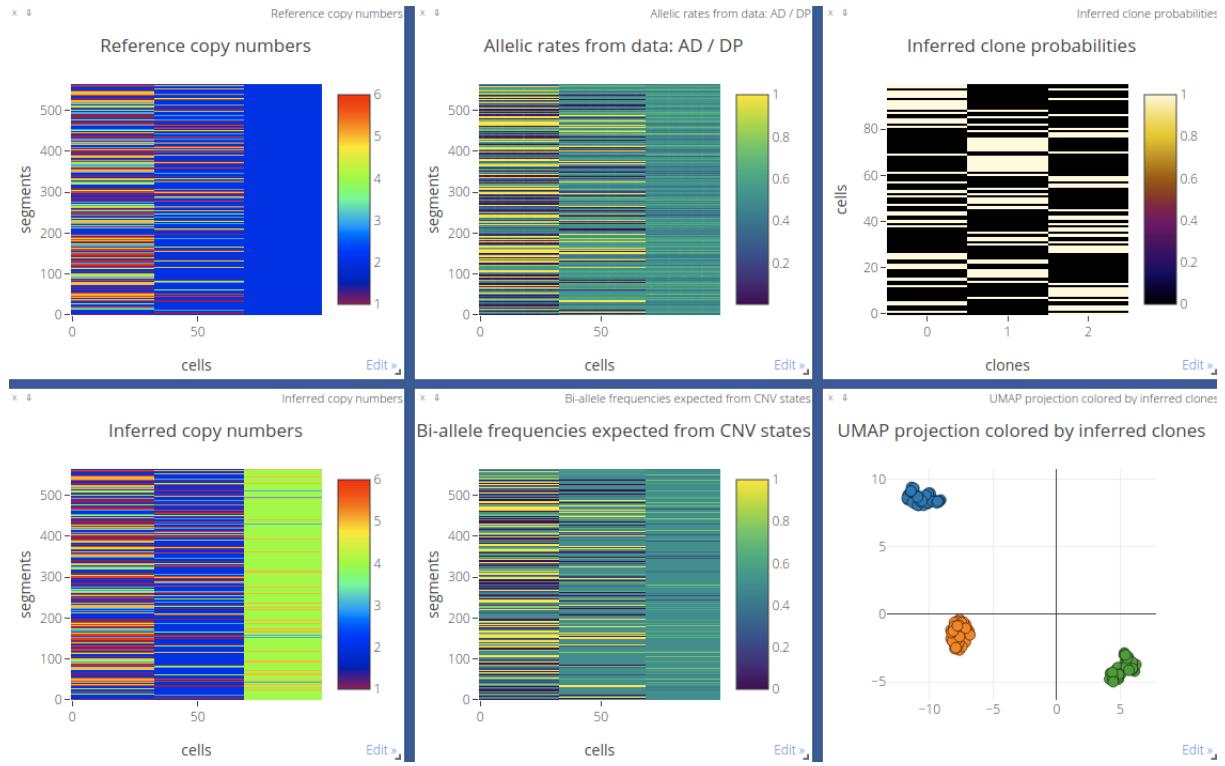


Рис. 5.3: Визуализация процесса обучения XClone-V2, выполненная в Visdom от Facebook Research. Тепловые карты в верхней строке, слева направо: истинные клональные RDR- и BAF-сигналы, а также вероятности принадлежности клеток клональным линиям (все сошлись к истинным). Нижняя строка, слева направо: предсказанное число копий в каждом из клонов, предсказанный аллельный дисбаланс и проекция клеток конкатенации этих двух векторов на 2D с помощью алгоритма UMAP, где каждый цвет соответствует клональной линии.

Скачок качества очевиден: корректное распределение на метках клональных линий удаётся получить всего за пару десятков итераций VB, чего в XClone-V1 не удавалось достичь даже за 10000 итераций семплирования по Гиббсу.

При этом бросается в глаза, что модель сильно ошибается на нормаль-

ных клетках: она предсказывает WGD там, где его нет. Но ранее упоминалось, что мультиномиальная модель RDR-модуля в XClone-V2 не позволяет детектировать WGD, потому на стадии байесовского вывода такие ошибки неизбежны. Их можно попытаться скорректировать на стадии пост-обработки, об этом можно прочесть в заключительном разделе ВКР. В остальном же и BAF-, и RDR-модуль предсказывают корректные величины.

Кроме того, справедливо будет заметить, что пока что эксперименты не позволяют оценить качество модели как классификатора. В идеальных условиях незашумленных BAF- и RDR-сигналов в данных классифицировать клетки просто: видно, что клональные линии линейно разделимы при проекции на 2D. Алгоритм KMeans в таком случае тоже даёт 100% точность. Поэтому симуляции сейчас проверяют не то, как классифицируются клетки, а как модель восстанавливает ASCNV. Сейчас это не первоочередная задача, но в последующих версиях модели будут реализованы и более продвинутые подходы к генерации данных.

5.2 Реальные данные: STP, scDNA-seq

Перед тем, как перейти к рассмотрению результатов, полученных XClone-V2 на данных STP-Nuclei, стоит взглянуть на данные поближе. Дизайн XClone-V2 был вдохновлён двумя важными наблюдениями, полученными из реальных данных.

Во первых, разработанный метод коррекции ошибок смены цепи показал, что по данным scDNA-seq можно получить надёжный BAF-сигнал. На данный момент ни один опубликованный метод, кроме CHISEL[3], не может этим похвастать.

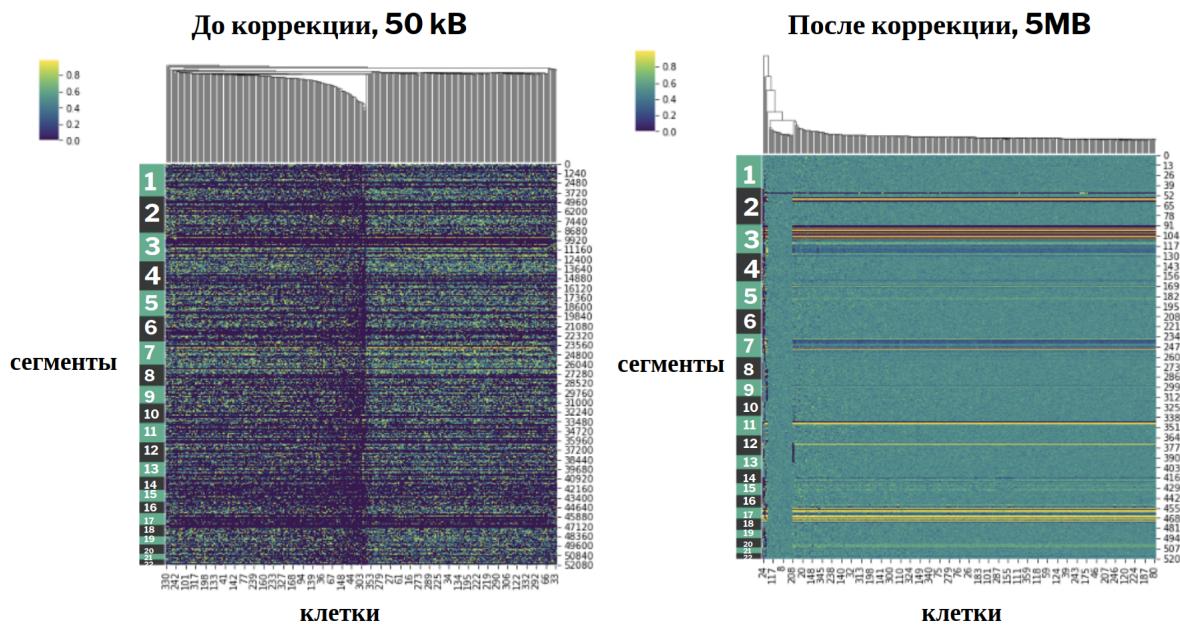


Рис. 5.4: Коррекция ошибок смены цепи на примере ДНК-образца медуллобластомы, STP-Nuclei. На рисунке изображены тепловые карты долей аллеля из «материнского» гаплотипа, числа слева обозначают номер хромосомы.

Картина аллельного дисбаланса до коррекции практически не прослеживается, после — становится очевидна. Видно, что при текущем подходе к сегментации теряются сложные структурные вариации в духе хромотрипсиса на хромосоме 7. Это известный недостаток, который будет устранён в следующих версиях XClone: при объединении первичных сегментов длиной 50кБ будет играть роль их взаимное подобие.

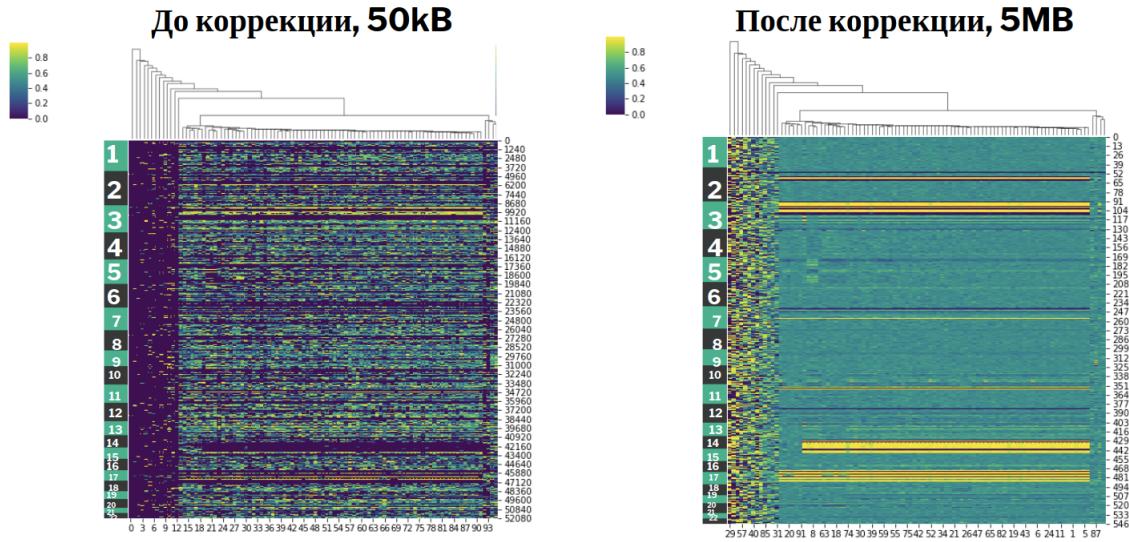


Рис. 5.5: Коррекция ошибок смены цепи на примере ДНК-образца медуллобластомы, STP-G&T. Суть та же, что и на предыдущей иллюстрации, но в этом образце глубина покрытия клеток в среднем выше, а потому аллельный сигнал проступает и в масштабе 50 килобаз.

Видно, что в образце STP-G&T гораздо отчётливее проступает большая делеция q-плеча 14-й хромосомы и р-плеча 15-й. Причём в части клеток этой делеции нет, и это может быть субклон. В STP-Nuclei весь сигнал, приходящий с 14 хромосомы, сильно размыт. Занимательно, что он прослеживается в транскриптомных данных, как это будет показано в следующем разделе. Но при этом в STP-G&T довольно существенная доля мёртвых клеток, а сам образец в два с половиной раза меньше, потому эксперименты было решено проводить на STP-Nuclei. А два там клона или три уже было не так важно. Эта опухоль в обоих случаях слишком однородная для того, чтобы использовать её при последующей публикации статьи, а переговоры по получению более интересных образцов пока ещё не завершены.

Полученная картина аллельного дисбаланса в образце согласовывалась с априорными знаниями врачей об этой опухоли. Это позволило перейти

к следующему шагу: проверке пришедшей от биологов гипотезы о том, что изменение глубины покрытия отдельных сегментов пропорционально, в первую очередь, числу копий этих участков, а остальными факторами можно пренебречь. В матрице числа прочтений обнаруживается та же структура: однородная опухоль с небольшой примесью здоровых клеток.

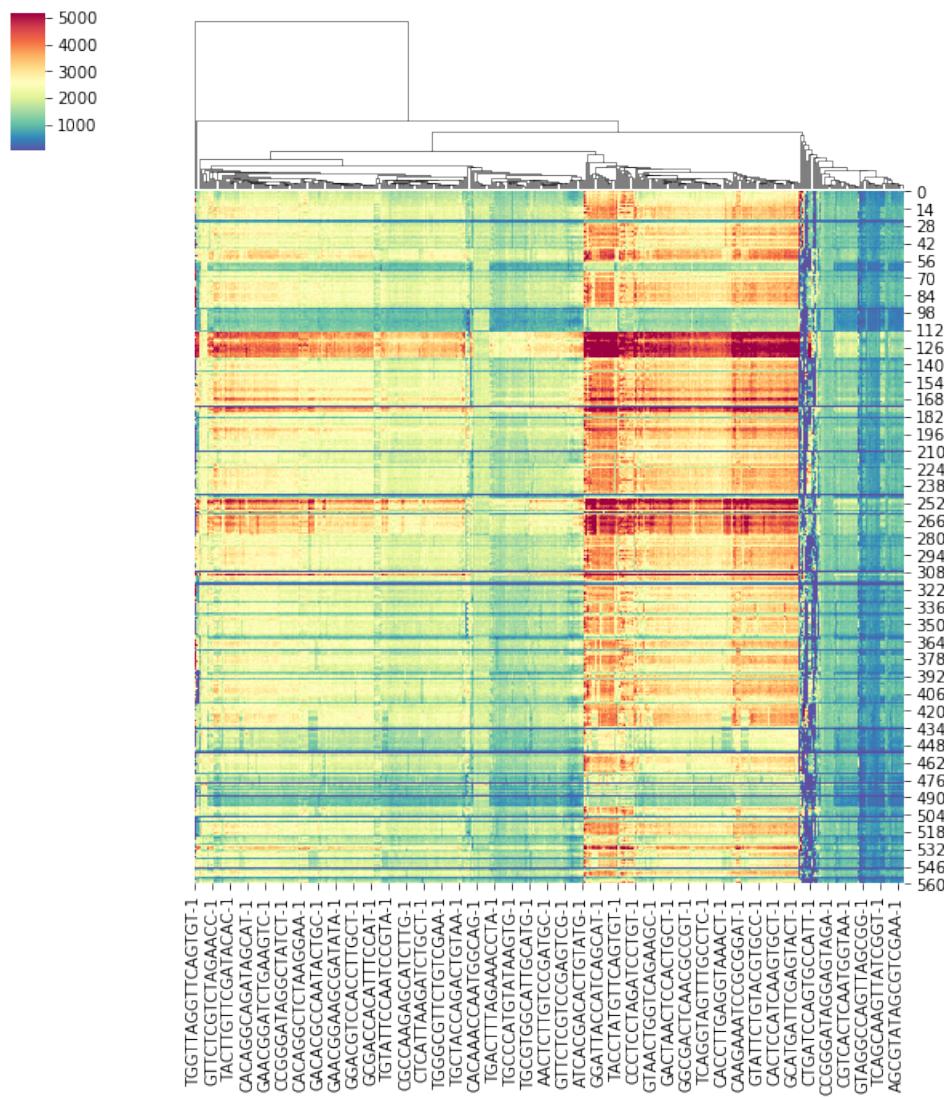


Рис. 5.6: Матрица числа прочтений. STP-Nuclei, scDNA-seq. Строки — сегменты длины в 5 мегабаз, столбцы — клетки.

Видно, что среднее количество прочтений сильно меняется от клетки к клетке: есть как плохие (возможно, мёртвые) клетки (на тепловой кар-

те справа), так и очень хорошие (оранжево-красная полоса на тепловой карте справа). Тем не менее, непохоже, что клетки с малой глубиной покрытия это отдельная клональная линия: распределение прочтений по сегментам там такое же, как и в остальных клетках. Также непохоже, что наблюдаемая субпопуляция клеток с большой глубиной покрытия это отдельная клональная линия, в которой произошло на 1 WGD больше: глубина больше, но не в степень двойки раз (50000 против 40000 это в пределах погрешности).

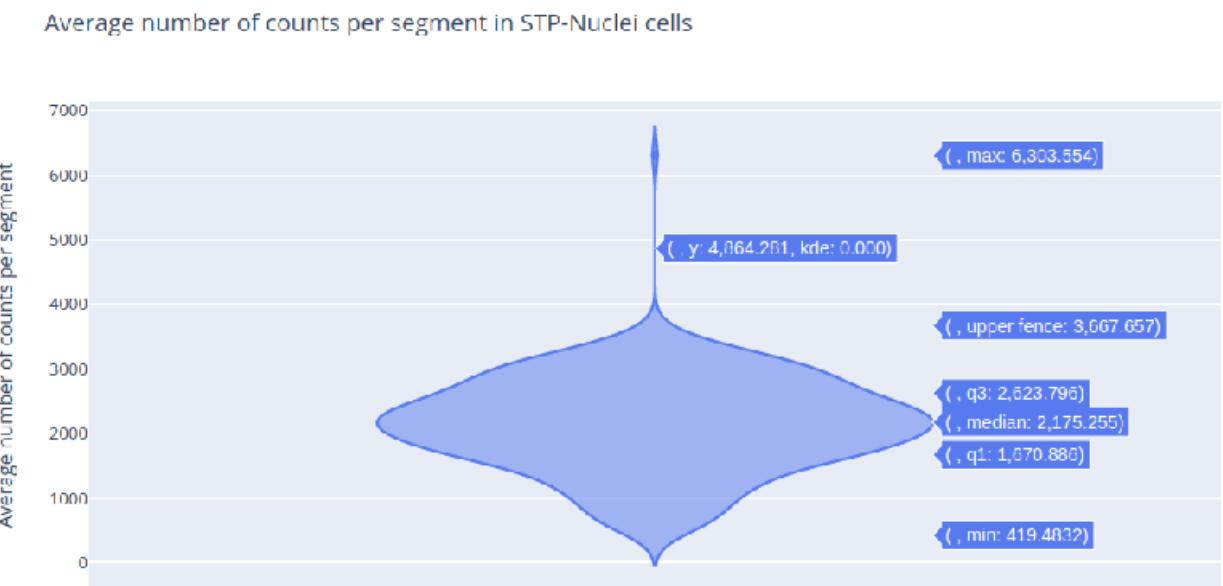


Рис. 5.7: Распределение среднего числа прочтений в сегменте. STP-Nuclei, scDNA-seq.

Гипотеза, что по распределению прочтений по геному можно точно предсказать число копий, можно проверить, сравнив наблюдаемое в клетке i распределение прочтений по сегментам с ожидаемым —

$$\left\{ \frac{m_i c_{i,j} / 2}{\sum_{b=1}^N m_b c_{b,j} / 2} \right\}_{i=1}^N$$

где N — число сегментов, m_i — ожидаемая доля прочтений в сегменте i в нормальных клетках, а $c_{i,j}$ — предсказанное число копий (нужно

поделить на 2, т.к. в норме геном диплоидный).

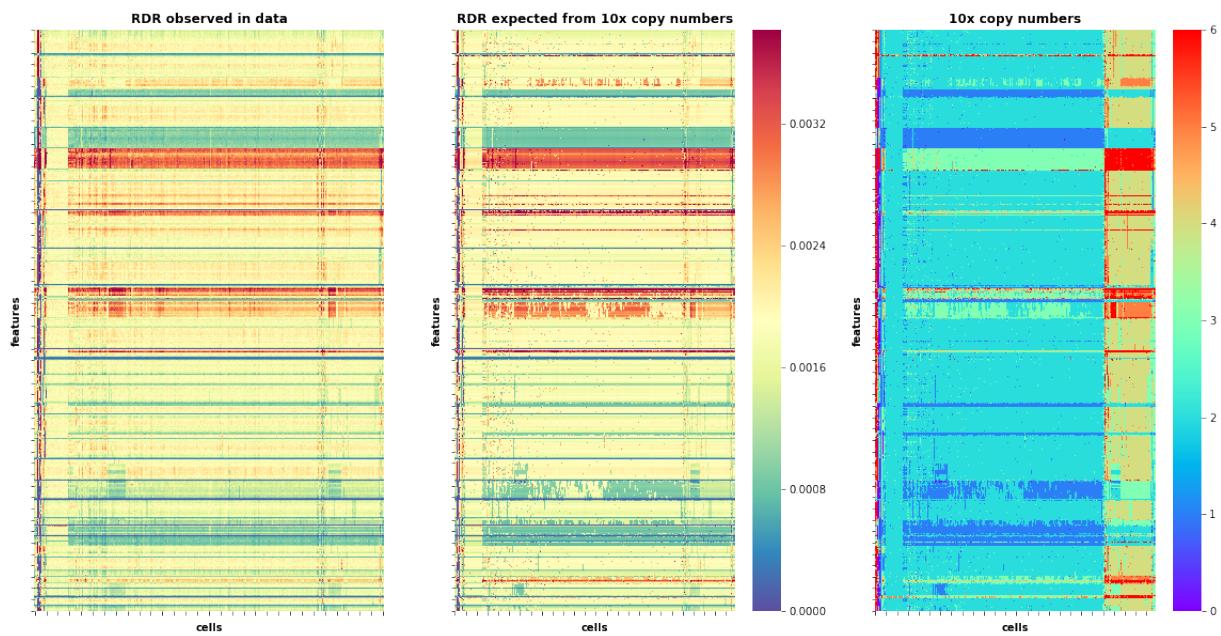


Рис. 5.8: Распределение прочтений по геному. STP-Nuclei, scDNA-seq. Первые две тепловые карты показывают наблюдаемую долю прочтений и ожидаемую по предсказанному алгоритмом CellRanger числу копий, которое само по себе изображено на третьей тепловой карте.

Видно, что распределения, в целом, похожи. Более детальный анализ каждой из хромосом в отдельности подтверждает первое впечатление:

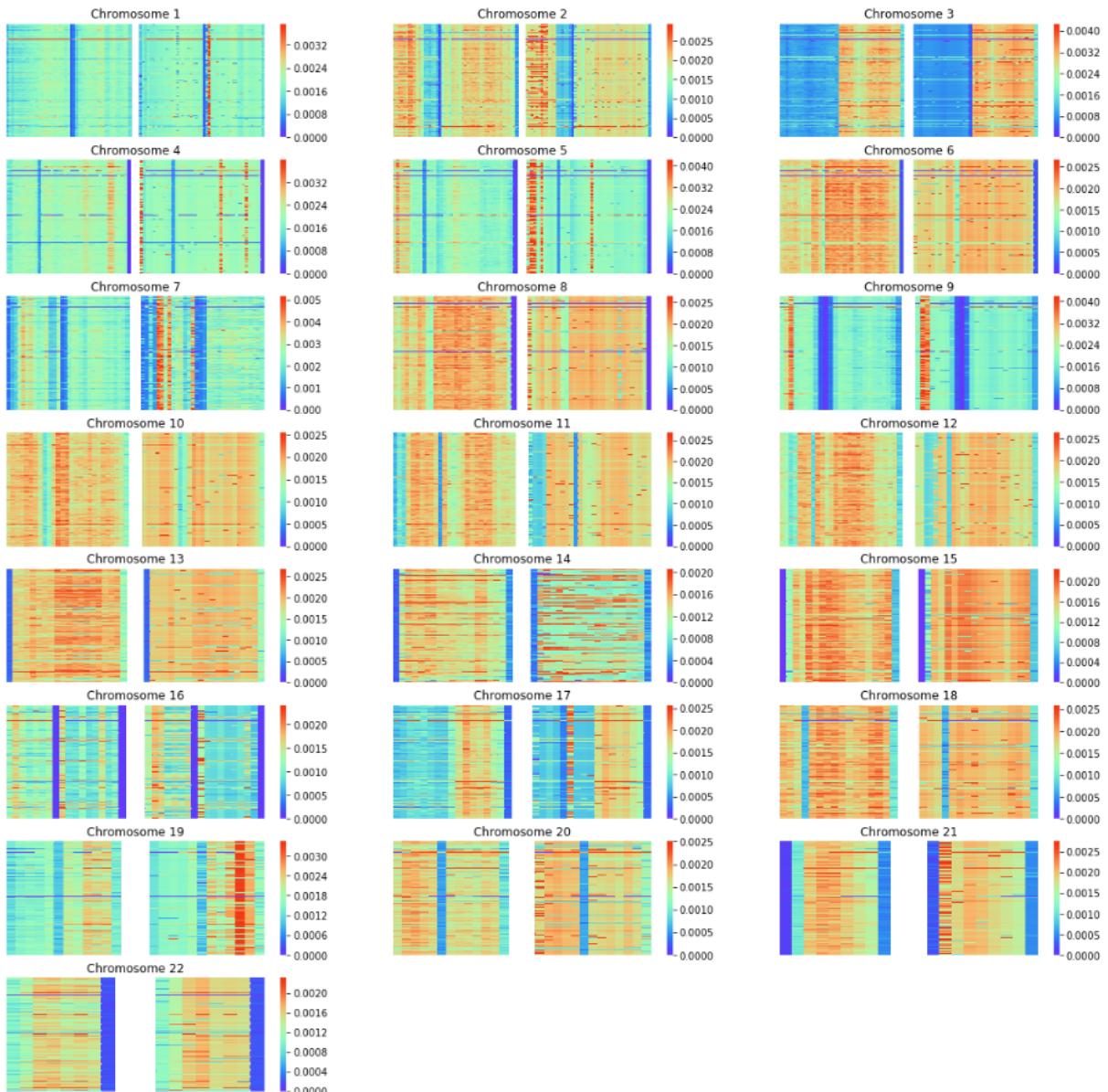


Рис. 5.9: Распределение прочтений по каждой из хросомом в отдельности. STP-Nuclei, scDNA-seq. Слева — доли прочтений, подсчитанные по реальным данным, справа — доли прочтений, ожидаемые на основе числа копий, предсказанных CellRanger.

Как уже отмечалось ранее, есть основания полагать, что диплоидные с точки зрения CellRanger клетки опухоли всё-таки тетраплоидные: глубина покрытия в этой группе клеток меньше нестатзначимо.

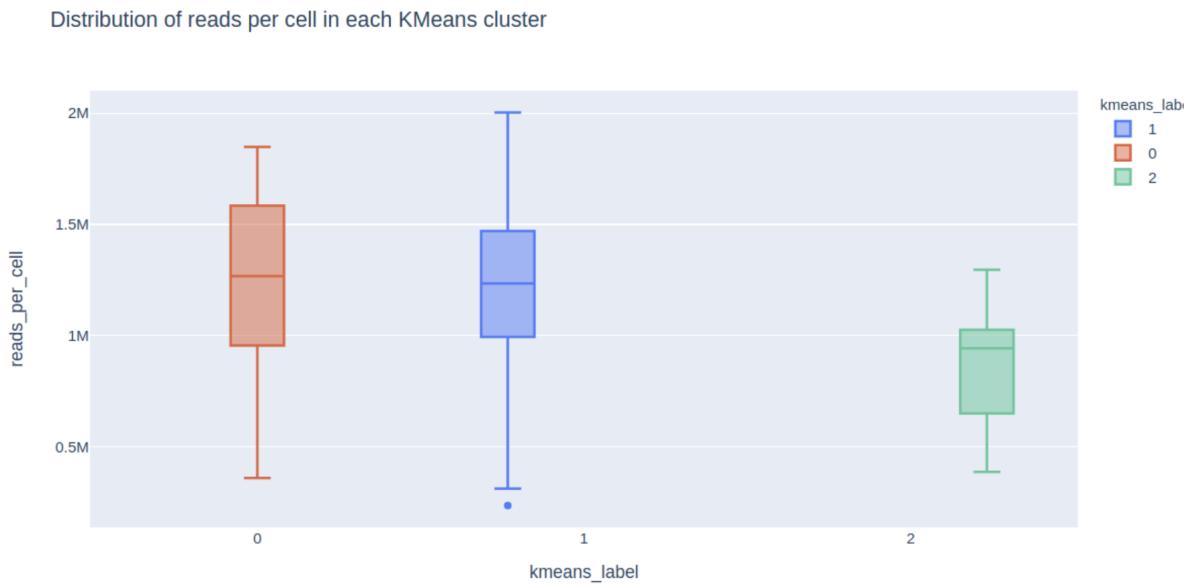


Рис. 5.10: Распределения глубины прочтений в клетках клональных линий, полученных иерархической кластеризацией мутационных профилей, предсказанных CellRanger. STP-Nuclei, scDNA-seq. 0 — тетраплоидные» клеток опухоли, 1 — «диплоидные» клетки опухоли, 2 — нормальные клетки. Видно, что глубина покрытия у опухолевых клеток значительно выше, чем у нормальных, но между собой два класса опухолевых клеток почти не отличаются.

Наблюдения хорошо согласовывались с теоретической моделью XClone-V2. Запуск на данных STP-Nuclei дал следующий результат:

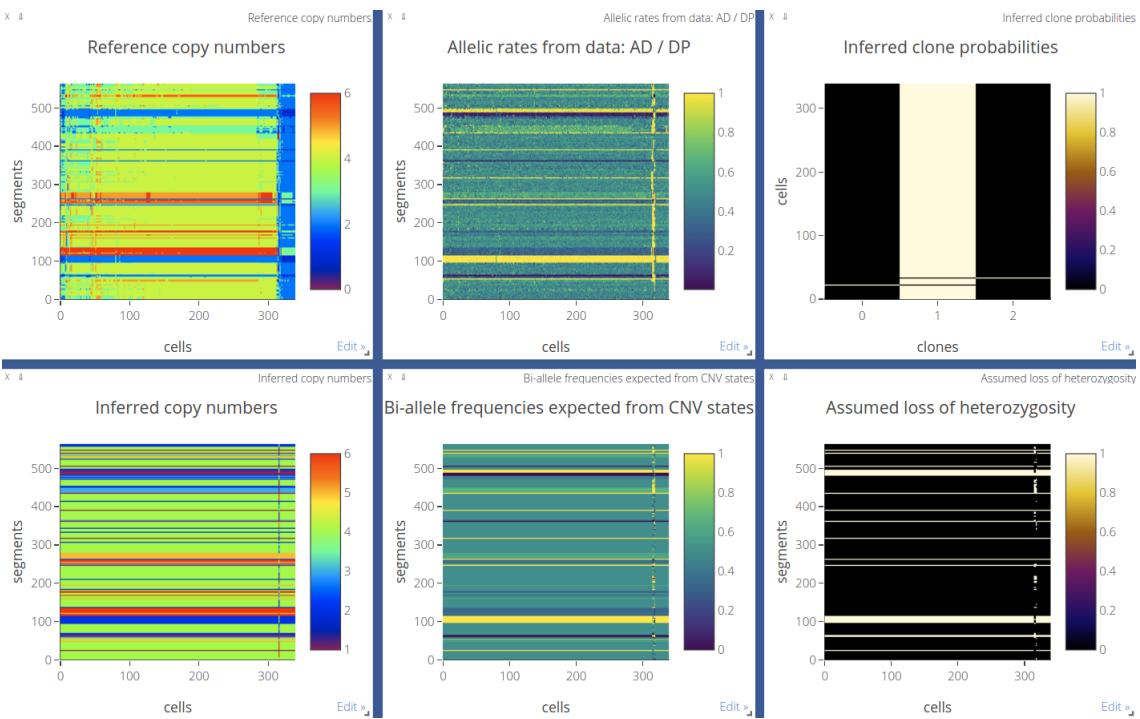


Рис. 5.11: Результаты запуска XClone-V2 (только опухолевые клетки). STP-Nuclei, scDNA. Верхняя строка, слева направо: число копий, предсказанное CHISEL; аллельный дисбаланс в данных; вероятности клональных линий (строки — клетки, столбцы — метки). Нижняя строка, слева направо: предсказанное число копий; предсказанный аллельный дисбаланс (определяется ASCNV); маска утраты гетерозиготности (по предсказанным ASCNV)

Как и ожидалось, алгоритм находит только одну клональную линию. Тем не менее, аллель-зависимые структурные вариации в ней метод находит правдоподобные.

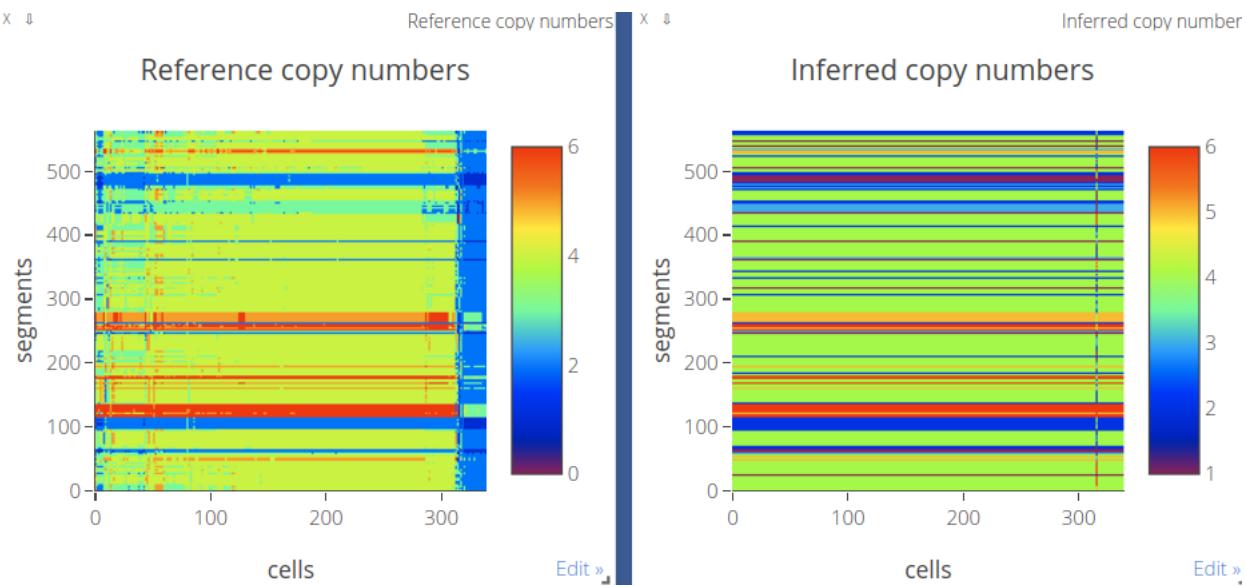


Рис. 5.12: Сравнение числа копий, предсказанного CHISEL и XClone-V2. STP-Nuclei, scDNA. С точностью до WGD, структурные вариации согласуются.

Учитывая, что результаты CHISEL и CellRanger на этих данных особо не отличаются, можно заключить, что три метода дают согласованный результат. Безусловно, это не полноценное сравнение трёх методов, оно только предстоит и будет обязательно проведено, когда:

- Будут исправлены известные недостатки XClone-V2;
- Будет найден образец с клональной структурой, достоверно известной до мелочей.

Обзорных статей по сравнению методов детекции CNV по данным single-cell секвенирования, в целом, мало, т.к. это относительно новая и пока мало изученная задача, но перед публикацией такой анализ обязательно будет проведен.

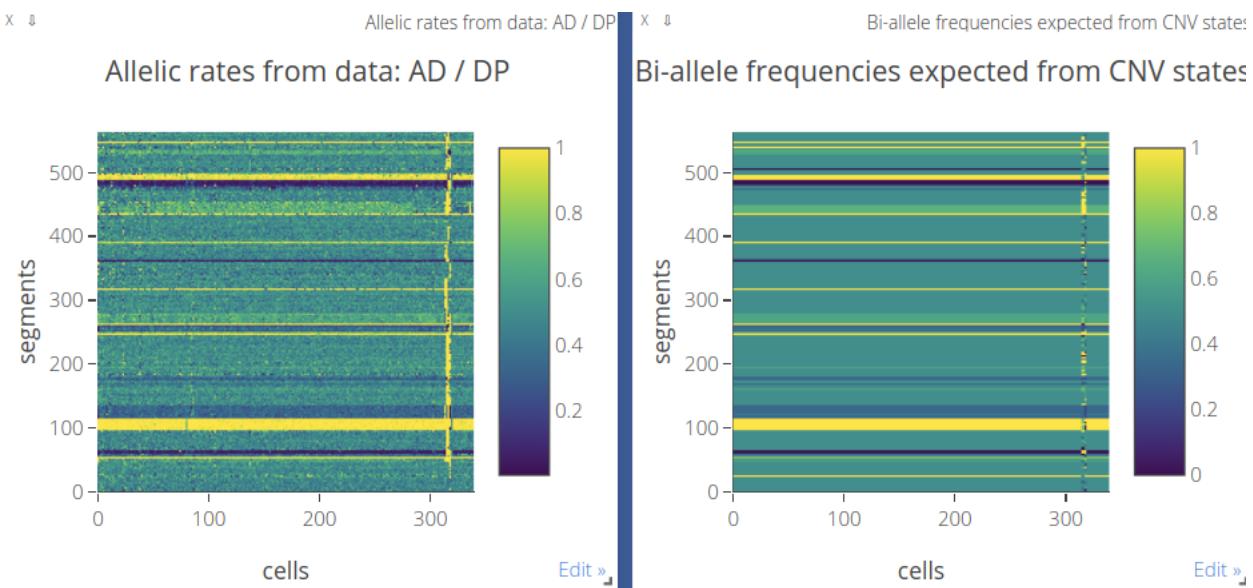


Рис. 5.13: Сравнение числа копий, предсказанного CHISEL и XClone-V2. STP-Nuclei, scDNA. С точностью до WGD, структурные вариации согласуются.

Предсказанные ASCNV с высокой точностью воспроизводят наблюдаемую картину аллельного дисбаланса в образце. Вместе с результатами, полученными на синтетических данных, это демонстрирует, что направление исследований перспективное, а алгоритм, в целом, работает. Увы, опухоль оказалась не настолько неоднородной, как авторы долгое время предполагали, но это стало понятно довольно поздно, и будет исправлено использованием других данных на следующих итерациях разработки метода.

5.3 Реальные данные: STP, scRNA-seq

Как уже упоминалось ранее, изначально XClone задумывался как метод совместного анализа нескольких омик. В планах было увязать друг с другом хотя бы геномные и транскриптомные данные: определять клonalную структуру опухоли по scDNA-seq, а потом классифицировать

по клональным линиям клетки последовательных scRNA-seq образцов и следить за эволюцией опухоли в динамике. Это позволило бы лучше понимать, как опухоль реагирует на лечение. Если бы метод работал надёжно, то он мог бы стать ценным инструментом в руках врачей-онкологов.

Но интеграция нескольких single-cell омик неспроста считается одной из 11 главных задач анализа single-cell данных [41]. Первые результаты выглядели вдохновляюще:

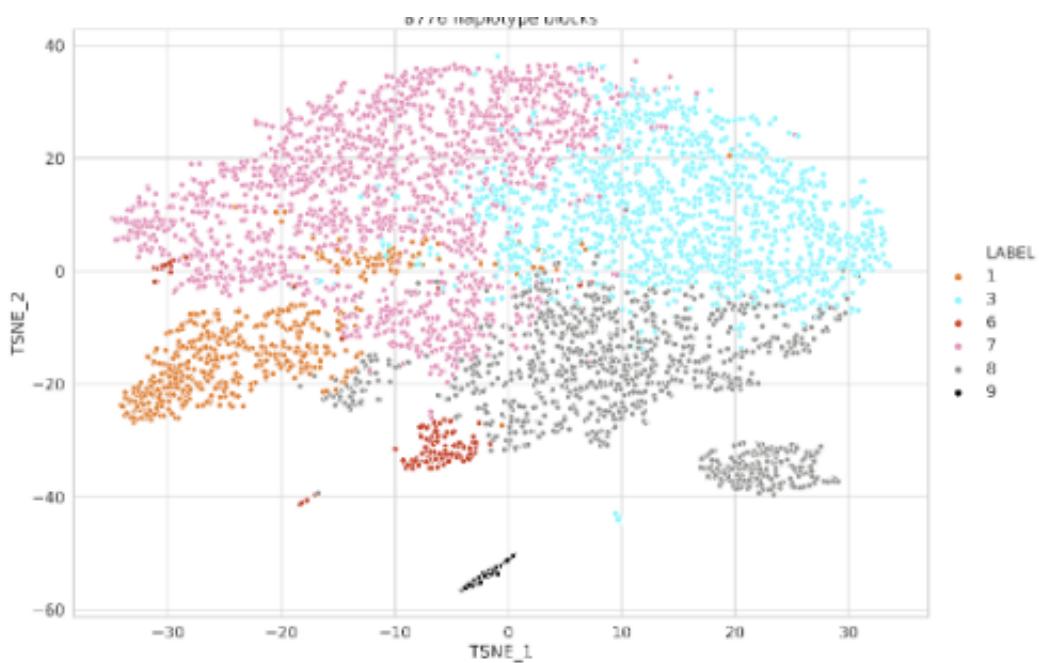


Рис. 5.14: Клетки РНК-образца, STP-Nuclei. Цвет соответствует клональной линии, предсказанной XClone-V1. Большой кластер точек по центру — опухолевые клетки, кластер снизу справа — нормальные клетки, природу малых кластеров в нижней части рисунка установить не удалось.

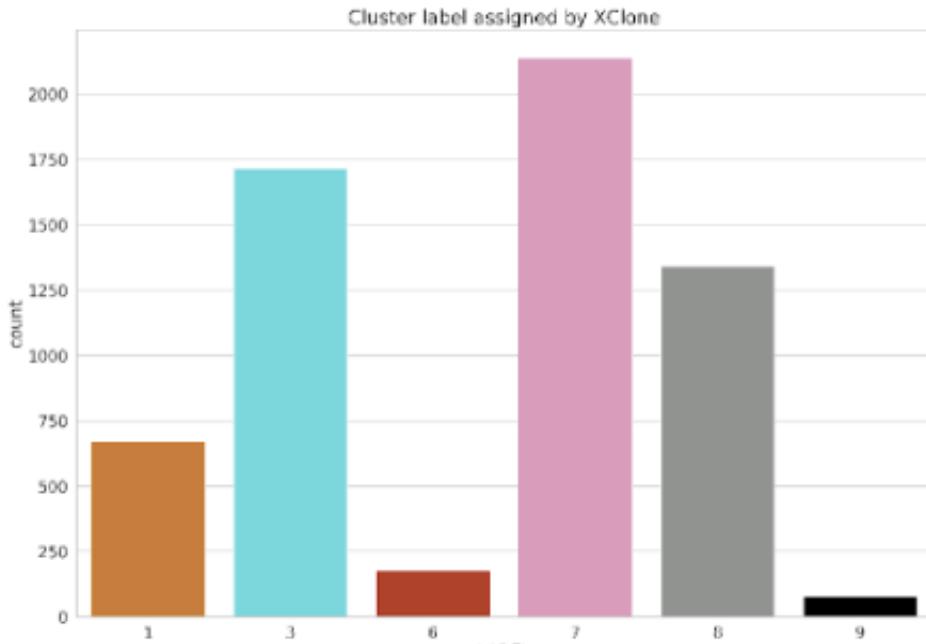


Рис. 5.15: Клетки РНК-образца, STP-Nuclei. Гистограммы предсказанных классов на предыдущем рисунке.

Да, часть опухолевых клеток была размечена как нормальные, но изначально это списали на то, что в данных могла быть примесь других клеточных типов, которые по профилю транскрипции находятся где-то между нормальными и опухолевыми. Тем не менее, систематический перебор random seed показал, что пропорции предсказанных классов, равно как и их количество и размеры, меняются в слишком больших пределах. Проще говоря, транскриптомных данных было слишком мало, чтобы получить достаточно хороший BAF-сигнал.

Предпринималось много попыток сгруппировать гетерозиготные ОНП так, чтобы суммарный сигнал был достаточно сильным для увереной классификации. Тем не менее, все они упирались в проблемы статистического фазирования гаплотипов, причём как статистического, так и основанного на длинных прочтениях, полученных по технологии Oxford Nanopore. Тем не менее, работа продолжалась. Большие надежды возла-

гались на разработанный метод коррекции ошибок смены цепи, который зарекомендовал себя на данных scDNA-seq образцов STP-Nuclei и STP-G&T.

Тем не менее, scRNA-seq данные STP-Nuclei оставались слишком разреженными даже в масштабе десятков мегабаз, когда предсказание числа копий уже теряло содержательный смысл:

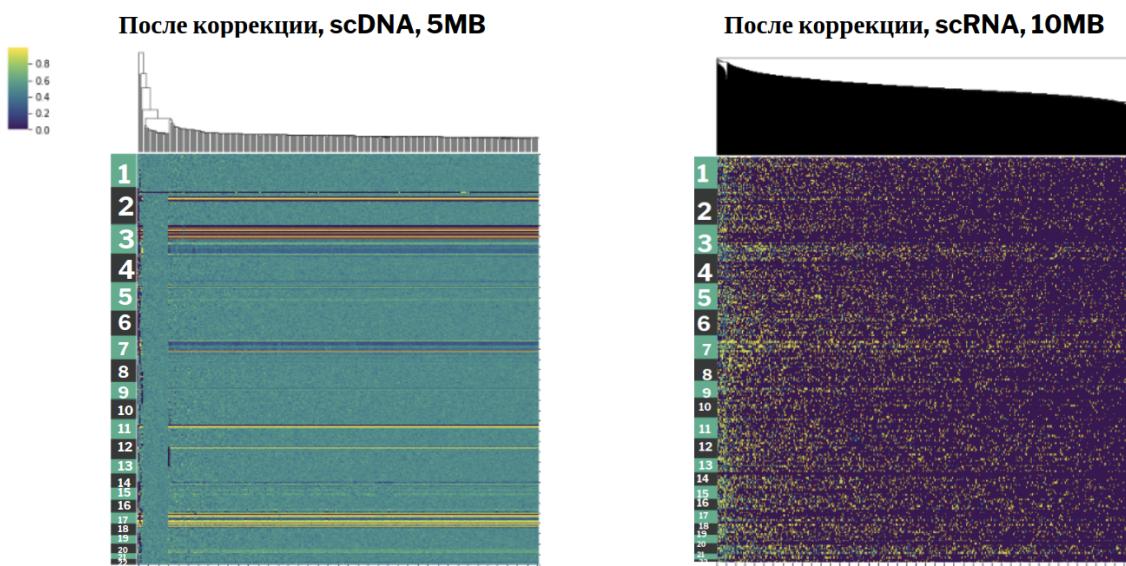


Рис. 5.16: Попытка коррекция ошибок смены цепи на примере ДНК- и РНК-образцов медуллобластомы, STP-Nuclei. Здесь фиолетовый цвет означает "нет данных". Даже в масштабе 10 мегабаз структура аллельного дисбаланса в данных РНК-секвенирования практически не просматривается.

Несмотря на то, что РНК-данные из STP-G&T гораздо менее разреженные, чем РНК-данные из STP-Nuclei, структура аллельного дисбаланса в них тоже не просматривалась:

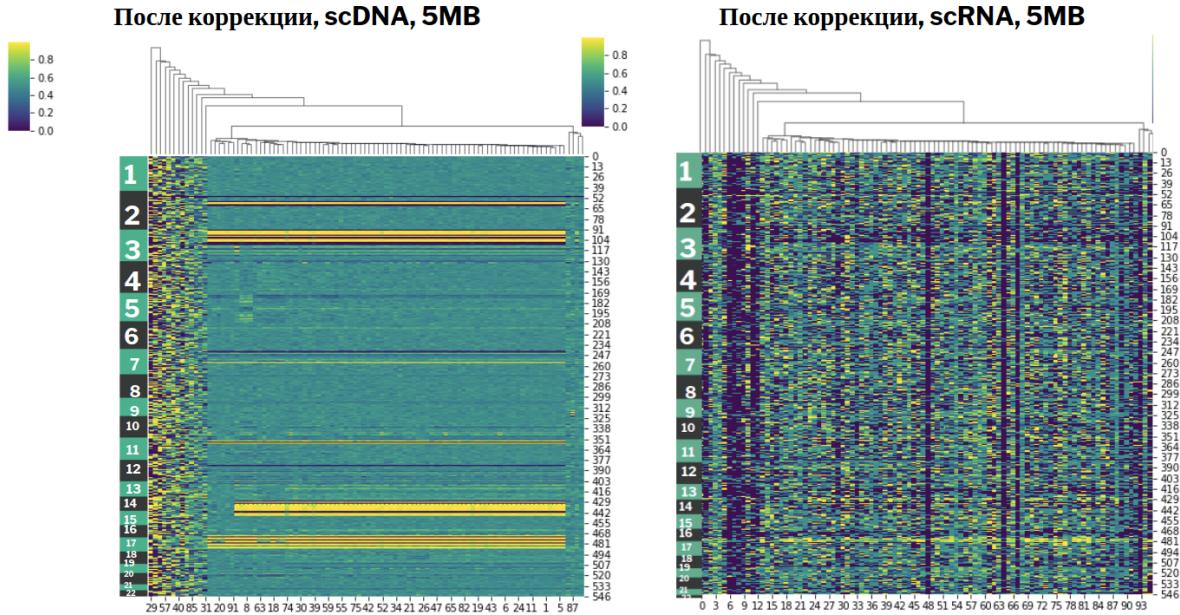


Рис. 5.17: Попытка коррекция ошибок смены цепи на примере ДНК- и РНК-образцов медуллобластомы, STP-G&T. Порядок клеток одинаков на обеих тепловых картах.

С похожими проблемами столкнулись и авторы алгоритма CaSpER[9]. Ожидать, что много научных групп сможет позволить себе глубокое scRNA-секвенирование не приходится. Как следствие, было принято решение сосредоточиться на получении RDR-сигнала из РНК-данных. Перспективность этой гипотезы подтверждается картиной loss/gain-событий, которую предсказывает алгоритм InferCNV:

STP-Nuclei - with merged neuronal cell reference

- malignant_7 = Immune Cells
- malignant_9 = Purkinje Cells
- malignant_8 = Endothelial Cells
- malignant_6 = Cells in cell division
- malignant_4 = Neuronal development Cells
- malignant_2 = Cell cycle & repair Cells

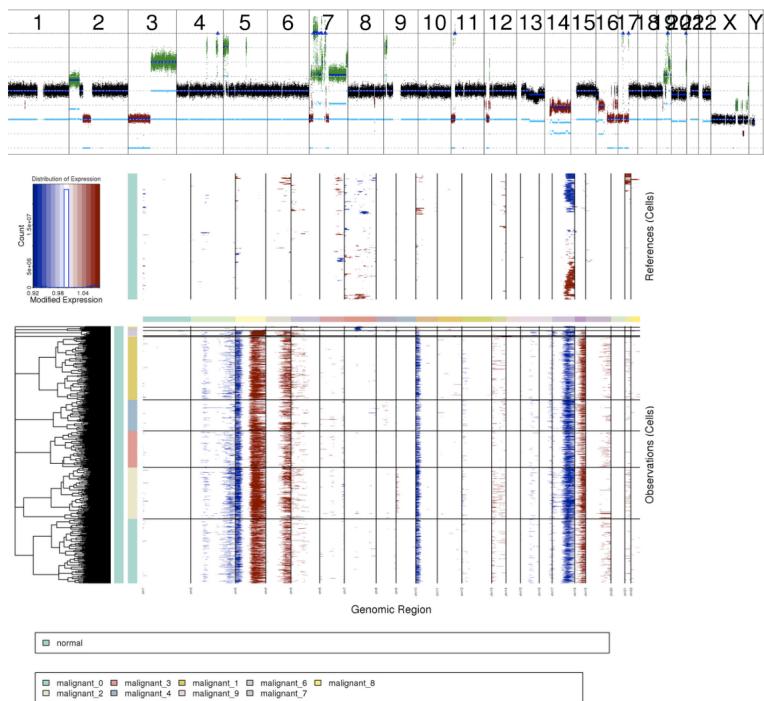


Рис. 5.18: Loss-/gain- события, предсказанные алгоритмом InferCNV по scRNA-seq образцу STP-Nuclei. Видны основные клональные события: делеция р-плеча и амплификация q-плеча хромосомы 3, хромотрипсис на хромосоме 7, делеции на хромосомах 14 и 16-17. Несмотря на то, что в образце 5751 клетка, опухоль всё ещё выглядит однородной, что подтверждает наблюдения, полученные по ДНК-образцу этой же опухоли.

Да, алгоритм InferCNV не совершенен, предсказывает только loss/gain-метку, но не число копий, и сильно зависит от качестве эталонного профиля экспрессии в нормальных клетках. В онкологической практике нормальную ткань часто не секвенируют, а базы данных, к которой можно было бы обратиться, нет. Точнее, пока нет: в проекте «Human Cell Atlas»[42] уже не первый год ведётся активная работа по сбору и систематизации данных обо всех типах клеток человека. Тем не менее, single-cell секвенирования это молодая технология, атлас не завершён даже для нормальных клеток, что уже говорить о всевозможных патологиях. Кроме того, уровни экспрессии сильно зависят от множества факторов:

стадия клеточного цикла, состояние окружающей среды, возраст клетки, состояние ткани... Это т.н. batch effect, и без поправок на него невозможно сравнивать результаты, полученные по данным из разных образцов. Устранение влияния технических факторов тоже считается одной из 11 главных задач анализа single-cell данных[41].

При всём при этом метод дарит надежду, что при известной клональной структуре, предсказанной по ДНК-образцу той же опухоли, можно будет разметить и клетки РНК-образца, используя RDR-сигнал, подсчитанный при аналогичной многостадийной предобработке данных. Достаточно было бы просто выбрать ту клональную линию, RDR-профиль которой наиболее похож (в смысле максимизации правдоподобия в модели, полученной при адаптации RDR-модуля из XClone). В момент написания этого текста ведётся активная работа в данном направлении.

6 Заключение. План дальнейших исследований

В данной работе представлен новый подход к решению двух важных задач вычислительной онкологии:

- Восстановления клональной структуры опухоли;
- Предсказания аллель-специфических структурных вариаций.

Изначально была и третья, главная задача — перенос клональной структуры с ДНК-образца на существенно более дешёвые РНК-образцы. Достичь в ней успеха не позволила разреженность РНК-данных. В связи с этим было решено сосредоточиться на ДНК-данных и решении первых двух задач.

Метод получил временное название XClone и представляет собой графическую байесовскую модель опухоли. XClone опирается на два важных типа биологических сигналов в данных секвенирования:

- BAF — аллельный дисбаланс, насколько чаще прочтения выравниваются на материнский аллель;
- RDR — амплификация доли прочтений, насколько чаще прочтения выравниваются на участок генома чем это предсказывает вероятностная модель секвенирования.

Разработанные алгоритмы предобработки данных показали, что оба этих сигнала можно надёжно извлекать из данных ДНК-секвенирования одиночных клеток, несмотря на их разреженность. В случае BAF-сигнала это долгое время считалось нерешённой задачей. Был предложен метод агрегирования гетерозиготных ОНП с сохранением тренда аллельных

частот на основе ЕМ-алгоритма, который показывает хорошие результаты на реальных данных. Этот метод представляет отдельный интерес, т.к. сам по себе имеет научную новизну.

В данной работе представлены две последовательные версии этого алгоритма, XClone-V1 и XClone-V2. Их различия резюмирует следующая таблица:

Версия	XClone-V1	XClone-V2
Главная задача	Перенос клональной структуры ДНК-образца на РНК-образец	Поиск аллель-зависимых структурных вариаций и восстановление по ним клональной структуры ДНК-образца
Метод вывода	Семплирование по Гиббсу	Вариационный байесовский вывод
Реализация	Pure Python + numpy, scipy, numba	Tensorflow 2.0
Поддержка GPU	Нет	Да

Оба метода показали хорошее качество на синтетических данных. В силу разреженности реальных РНК-данных, от XClone-V1 пришлось отказаться, в связи с чем большая часть результатов, представленных в данной работе, получена уже с помощью XClone-V2. Было продемонстрировано, что XClone-V2 может с высокой долей достоверности воспроизвести аллель-специфические структурные вариации по данным, извёчённым из реальной опухоли — медуллобластомы, детской опухоли мозга — на коммерческой платформе 10X Genomics, основных производителей

оборудования для секвенирования одиночных клеток.

На данный момент, у XClone-V2 есть только один непосредственный конкурент — алгоритм CHISEL[3], опубликованный в ноябре 2019 года учёными из лаборатории Бена Рафаэля, Принстон. Результаты CHISEL и XClone-V2 согласуются, при этом XClone-V2 работает гораздо быстрее, порядка 10 минут против нескольких часов CHISEL на CPU, а на GPU ещё в несколько раз быстрее (CHISEL не поддерживает GPU вовсе). Кроме того, вариационный байесовский вывод в XClone-V2 происходит автоматически, т.к. его удалось выразить в синтаксисе Tensorflow.Probability. Благодаря этому в XClone-V2, в отличие от CHISEL, можно легко совершенствовать статистическую модель и добавлять в неё поддержку новых модальностей (scRNA-seq, scATAC-seq, соматические мутации, митохондриальная ДНК), чтобы использовать все имеющиеся под рукой данные для уточнения диагноза. Такой подход позволяет в перспективе разработать линейку специализированных методов под разные нужды учёных и врачей.

Экспериментально были выяснены и недостатки модели XClone-V2, подробно описанные в конце главы "Материалы и методы". Их анализ позволил разработать проект новой, улучшенной версии алгоритма — XClone-V3, в которой RDR-составляющая заменена с мультиномиальной модели, которую трудно оптимизировать и в которой неявно создаются лишние зависимости между признаками, на модели, симметричной BAF-модулю и использующей отрицательное биномиальное распределение для моделирования числа прочтений в сегментах генома опухолевых клеток. Это стандартный подход в моделировании данных ДНК-секвенирования одиночных клеток, и авторы убеждены, что он будет давать лучшие результаты.

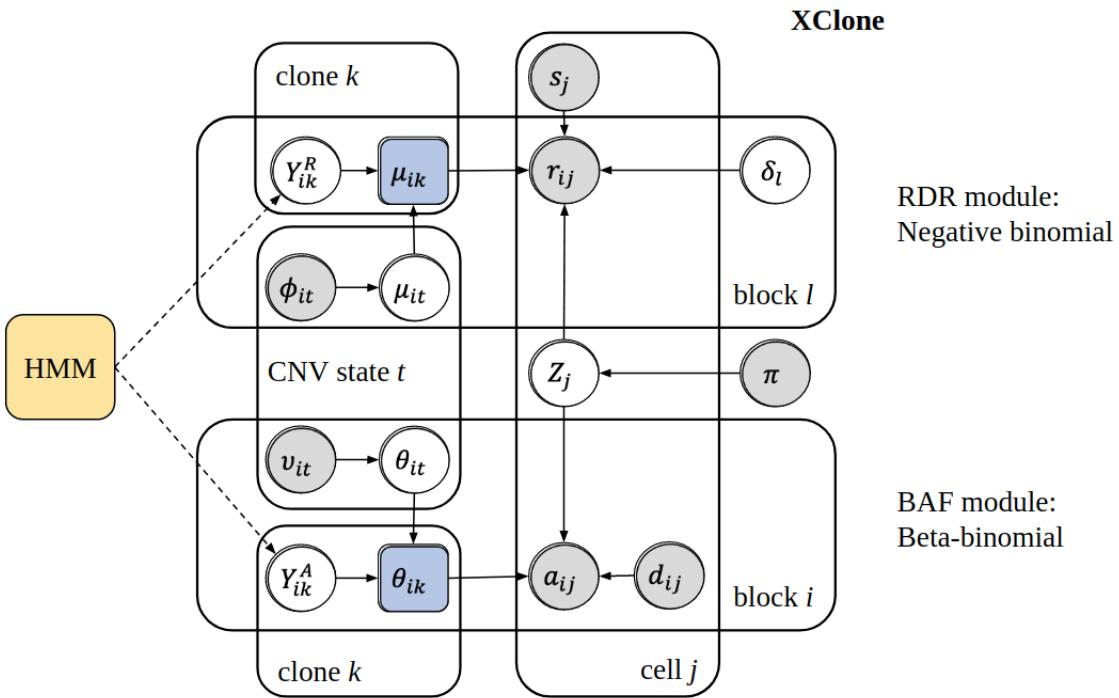


Рис. 6.1: Третья версия алгоритма XClone в plate notation.

Кроме того, продвижения в задаче определения структурных вариаций по данным scRNA-seq, а именно алгоритм InferCNV, вдохновили авторов вернуться к решению первоначальной задачи — переносу клональной структуры с ДНК-образца на РНК-образцы, — в этот раз на основе преобразованных RDR-сигналов, аналогичных тем, что используются в методе InferCNV. Увы, имеющиеся данные, в силу однородности опухоли, ограничивают дальнейшую разработку метода. В связи с этим данный момент идут переговоры с лабораторией Кристины Кёртис, Стэнфорд, с целью валидации алгоритма XClone-V3 на образцах опухоли желудка с значительно более богатой клональной структурой (порядка 10 клональных линий) и дальнейшей коллaborации. Если XClone-V3 покажет хорошие результаты, это будет первый алгоритм предсказания аллель-специфических структурных вариаций по данным scRNA-seq, что позволит опубликовать метод в высокоимпактном журнале.

7 Благодарности

Данная работа посвящена приложениям байесовских методов машинного обучения к актуальной задаче вычислительной онкологии — задаче восстановления клональной структуры опухоли по данным высокоприводительного секвенирования одиночных клеток. Такой выбор темы отражает научные интересы автора — статистическое моделирование на больших медицинских данных, — сформировавшиеся за время работы над проектами по вычислительной системной биологии под руководством к.ф.-м.н. Юрия Львовича Притыкина¹². Автор благодарен ему за время, уделённое на протяжении двух лет работы под его руководством, и за возможность уже на младших курсах приобщиться к высоким стандартам современной академической науки.

Работа над дипломом велась под руководством Оливера Штегле¹³, профессора Heidelberg University¹⁴ (Университет Хайдельберга, Германия) а также действующих и бывших сотрудников его научных групп в DKFZ¹⁵ (Немецкий Центр Онкологических Исследований), EMBL Heidelberg¹⁶ (Европейская Лаборатория Молекулярной Биологии) и HKU¹⁷ (Университет Гонконга). Формальным научным руководителем следует считать к.ф.-м.н. Yuanhua Huang¹⁸, заведующего группой вычислительной биологии в Университете Гонконга. Автор признателен ему за профессионализм и тщательность, с которой он на протяжении многих месяцев руководил разработкой метода. Кроме того, автор выражает личную

¹²<https://scholar.google.com/citations?user=Arx56RkJBrYC&hl=en>

¹³<https://scholar.google.com/citations?user=CISXZ4IAAAAJ&hl=en>

¹⁴<https://www.uni-heidelberg.de/en>

¹⁵<https://www.dkfz.de/en/index.html>

¹⁶<https://www.embl.de/>

¹⁷<https://www.hku.hk/>

¹⁸<https://www.sbmss.hku.hk/staff/yuanhua-huang>

благодарность профессору Штегле и к.ф.-м.н. Ханне Сьюзак, Николе Казирахи, Родриго Гонсало Парра, а также Д. О. Бредихину и В. А. Огородникову за ценные замечания и советы, придавшие многим аспектам метода завершённый, логически стройный вид.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

- [1] Hamim Zafar и др. «SiCloneFit: Bayesian inference of population structure, genotype, and phylogeny of tumor clones from single-cell genome sequencing data». B: 29 (нояб. 2019), с. 1847–1859. DOI: 10.1101/gr.243121.118. URL: <https://doi.org/10.1101/gr.243121.118>.
- [2] Serin A. Harmanci, A.O. Harmanci и X. Zhou. «CaSpER identifies and visualizes CNV events by integrative analysis of single-cell or bulk RNA-sequencing data.» B: *Nature Communications* 89 (нояб. 2020). DOI: 10.1038/s41467-019-13779-x. URL: <https://doi.org/10.1038/s41467-019-13779-x>.
- [3] S. Zaccaria и B.J. Raphael. «Characterizing the allele- and haplotype-specific copy number landscape of cancer genomes at single-cell resolution with CHISEL». B: *bioRxiv* (нояб. 2019). DOI: 10.1101/837195. URL: <https://doi.org/10.1101/837195>.
- [4] H. Li и др. «The Sequence Alignment/Map format and SAMtools». B: *Bioinformatics* (авг. 2009), с. 2087–2079. DOI: 10.1093/bioinformatics/btp352. URL: <https://doi.org/10.1093/bioinformatics/btp352>.
- [5] P.R. Loh и др. «Reference-based phasing using the Haplotype Reference Consortium panel». B: *Nature Genetics* (48 окт. 2016), с. 1443–1448. DOI: 10.1038/ng.3679. URL: <https://doi.org/10.1038/ng.3679>.
- [6] Haplotype Reference Consortium. «A reference panel of 64,976 haplotypes for genotype imputation». B: *Nature Genetics* (48 авг. 2016), с. 1279–1283. DOI: 10.1038/ng.3643. URL: <https://doi.org/10.1038/ng.3643>.
- [7] Y. Choi и др. «Comparison of phasing strategies for whole human genomes». B: *PLOS GENETICS* (апр. 2018). DOI: 10.1371/journal.

- p_{gen}. 1007308. URL: <https://doi.org/10.1371/journal.pgen.1007308>.
- [8] K.P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012. ISBN: 978-0-262-01802-9.
- [9] A. Serin Harmancı, A. O. Harmancı и X. Zhou. «CaSpER identifies and visualizes CNV events by integrative analysis of single-cell or bulk RNA-sequencing data». B: *Nat Commun* 11.1 (янв. 2020), c. 89.
- [10] D.J. McCarthy, R. Rostom и Huang Y. «Cardelino: computational integration of somatic clonal substructure and single-cell transcriptomes.» B: *Nature Methods* 17 (май 2020), c. 414—421. DOI: 10.1038/s41592-020-0766-3. URL: <https://doi.org/10.1038/s41592-020-0766-3>.
- [11] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.
- [12] H. W. Kuhn. «The Hungarian method for the assignment problem». B: *Naval Research Logistics Quarterly* 2.1-2 (1955), c. 83—97. DOI: 10.1002/nav.3800020109. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/nav.3800020109>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/nav.3800020109>.
- [13] David M. Blei, Alp Kucukelbir и Jon D. McAuliffe. «Variational Inference: A Review for Statisticians». B: *Journal of the American Statistical Association* 112.518 (февр. 2017), c. 859—877. ISSN: 1537-274X. DOI: 10.1080/01621459.2017.1285773. URL: <http://dx.doi.org/10.1080/01621459.2017.1285773>.
- [14] Joshua V. Dillon и др. «TensorFlow Distributions». B: *CoRR* abs/1711.10604 (2017). arXiv: 1711.10604. URL: <http://arxiv.org/abs/1711.10604>.

- [15] Huang Yuanhua, J. McCarthy Davis и Oliver Stegle. «Vireo: Bayesian demultiplexing of pooled single-cell RNA-seq data without genotype reference». B: *Genome Biology* 20 (дек. 2019). DOI: 10.1186/s13059-019-1865-2. URL: <https://doi.org/10.1186/s13059-019-1865-2>.
- [16] S. Ramanujan. *The Lost Notebook and other Unpublished Papers*. Под ред. S. S. Raghavan S. and Rangachari. Narosa, New Dehli, 1987.
- [17] Daniel Dufresne. «The log-normal approximation in financial and other computations». B: *Advances in Applied Probability* 36 (3 июль 2004), c. 747–773. DOI: 10.1239/aap/1093962232. URL: <https://doi.org/10.1239/aap/1093962232>.
- [18] Diederik P. Kingma и Jimmy Ba. «Adam: A Method for Stochastic Optimization». B: (2014). cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015. URL: <http://arxiv.org/abs/1412.6980>.
- [19] Dong C. Liu и Jorge Nocedal. «On the limited memory BFGS method for large scale optimization.» B: *Math. Program.* 45.1-3 (1989), c. 503–528. URL: <http://dblp.uni-trier.de/db/journals/mp/mp45.html#LiuN89>.
- [20] Romain Lopez и др. «Deep generative modeling for single-cell transcriptomics». B: *Nature Methods* 15.12 (дек. 2018), c. 1053–1058. ISSN: 1548-7105. DOI: 10.1038/s41592-018-0229-2. URL: <https://doi.org/10.1038/s41592-018-0229-2>.
- [21] Christoph Lippert и др. «LIMIX: genetic analysis of multiple traits». B: *bioRxiv* (2014). DOI: 10.1101/003905. eprint: <https://www.biorxiv.org/content/early/2014/05/22/003905.full.pdf>. URL: <https://www.biorxiv.org/content/early/2014/05/22/003905>.

- [22] James S. Bergstra и др. *Algorithms for Hyper-Parameter Optimization*. Под ред. J. Shawe-Taylor и др. Curran Associates, Inc., 2011, с. 2546—2554. URL: <http://papers.nips.cc/paper/4443-algorithms-for-hyper-parameter-optimization.pdf>.
- [23] Richa Bharti и Dominik Grimm. «Current challenges and best-practice protocols for microbiome analysis». В: *Briefings in bioinformatics* (дек. 2019). DOI: 10.1093/bib/bbz155.
- [24] «Method of the Year 2013». В: *Nature Methods* (11 янв. 2014). DOI: 10.1038/nmeth.2801. URL: <https://doi.org/10.1038/nmeth.2801>.
- [25] «Method of the Year 2019: Single-cell multimodal omics». В: *Nature Methods* (17 янв. 2020). DOI: 10.1038/s41592-019-0703-5. URL: <https://doi.org/10.1038/s41592-019-0703-5>.
- [26] Daniel C. Koboldt и др. «The Next-Generation Sequencing Revolution and Its Impact on Genomics». В: *Cell* 155 (1 сент. 2013). DOI: 10.1016/j.cell.2013.09.006. URL: <https://doi.org/10.1016/j.cell.2013.09.006>.
- [27] «Linnarsson, S., Teichmann, S.A. Single-cell genomics: coming of age.» В: *Genome Biology* 97 (17 2016). DOI: 10.1186/s13059-016-0960-x. URL: <https://doi.org/10.1186/s13059-016-0960-x>.
- [28] Travis I. Zack и др. «Pan-cancer patterns of somatic copy number alteration». В: *Nature Genetics* 45.10 (окт. 2013), с. 1134—1140. ISSN: 1546-1718. DOI: 10.1038/ng.2760. URL: <https://doi.org/10.1038/ng.2760>.
- [29] Erin D. Pleasance, R. Keira Cheetham и Philip J. Stephens. «A comprehensive catalogue of somatic mutations from a human cancer genome». В: *Nature* 463.7278 (янв. 2010), с. 191—196. ISSN: 1476-4687. DOI: 10.1038/nature08658. URL: <https://doi.org/10.1038/nature08658>.

- [30] Nicola Waddell и др. «Whole genomes redefine the mutational landscape of pancreatic cancer». B: *Nature* 518.7540 (февр. 2015), с. 495—501. ISSN: 1476-4687. DOI: 10 . 1038 / nature14169. URL: <https://doi.org/10.1038/nature14169>.
- [31] Stefan C. Dentro и др. «Portraits of genetic intra-tumour heterogeneity and subclonal selection across cancer types». B: *bioRxiv* (2018). DOI: 10 . 1101 / 312041. eprint: <https://www.biorxiv.org/content/early/2018/07/13/312041.full.pdf>. URL: <https://www.biorxiv.org/content/early/2018/07/13/312041>.
- [32] Sam Thiagalingam и др. «Mechanisms underlying losses of heterozygosity in human colorectal cancers». B: *Proceedings of the National Academy of Sciences* 98.5 (2001), с. 2698—2702. ISSN: 0027-8424. DOI: 10 . 1073 / pnas . 051625398. eprint: <https://www.pnas.org/content/98/5/2698.full.pdf>. URL: <https://www.pnas.org/content/98/5/2698>.
- [33] Jacqueline A. Langdon и др. «Combined genome-wide allelotyping and copy number analysis identify frequent genetic losses without copy number reduction in medulloblastoma». B: *Genes, Chromosomes and Cancer* 45.1 (2006), с. 47—60. DOI: 10 . 1002 / gcc . 20262. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/gcc.20262>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/gcc.20262>.
- [34] Pablo Lapunzina и David Monk. «The consequences of uniparental disomy and copy number neutral loss-of-heterozygosity during human development and cancer». B: *Biology of the Cell* 103.7 (2011), с. 303—317. DOI: 10 . 1042 / BC20110013. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1042/BC20110013>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1042/BC20110013>.

- [35] Scott L. Carter и др. «Absolute quantification of somatic DNA alterations in human cancer». B: *Nature Biotechnology* 30.5 (май 2012), с. 413—421. ISSN: 1546-1696. DOI: 10.1038/nbt.2203. URL: <https://doi.org/10.1038/nbt.2203>.
- [36] Moritz Gerstung и др. «The evolutionary history of 2,658 cancers». B: *Nature* 578.7793 (февр. 2020), с. 122—128. ISSN: 1476-4687. DOI: 10.1038/s41586-019-1907-7. URL: <https://doi.org/10.1038/s41586-019-1907-7>.
- [37] Hans Zahn и др. «Scalable whole-genome single-cell library preparation without preamplification». B: *Nature Methods* 14.2 (февр. 2017), с. 167—173. ISSN: 1548-7105. DOI: 10.1038/nmeth.4140. URL: <https://doi.org/10.1038/nmeth.4140>.
- [38] Emma Laks и др. «Resource: Scalable whole genome sequencing of 40,000 single cells identifies stochastic aneuploidies, genome replication states and clonal repertoires». B: *bioRxiv* (2018). DOI: 10.1101/411058. eprint: <https://www.biorxiv.org/content/early/2018/09/13/411058.full.pdf>. URL: <https://www.biorxiv.org/content/early/2018/09/13/411058>.
- [39] Iain C. Macaulay, Wilfried Haerty и Parveen Kumar. «G&T-seq: parallel sequencing of single-cell genomes and transcriptomes». B: *Nature Methods* 12.6 (июнь 2015), с. 519—522. ISSN: 1548-7105. DOI: 10.1038/nmeth.3370. URL: <https://doi.org/10.1038/nmeth.3370>.
- [40] J. Fan и др. «Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data». B: *Genome Res.* 28.8 (авг. 2018), с. 1217—1227.
- [41] David Laehnemann, Johannes Koester и Ewa Szczurek. «Eleven grand challenges in single-cell data science». B: *Genome Biology* 21.1 (февр.

- 2020), с. 31. ISSN: 1474-760X. DOI: 10.1186/s13059-020-1926-6.
URL: <https://doi.org/10.1186/s13059-020-1926-6>.
- [42] Aviv Regev и др. «The Human Cell Atlas». eng. В: *eLife* 6 (дек. 2017). e27041[PII], e27041. ISSN: 2050-084X. DOI: 10.7554/eLife.27041.
URL: <https://doi.org/10.7554/eLife.27041>.