

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. _____ Yuanhua Huang.

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 клетках 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[16], InferCNV¹, Casper[2], CHISEL[15] — 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 него пока есть всего один прямой конкурент — алгоритм CHISEL, — от которого XClone выгодно отличается тем, что допускает естественное обобщение на случай нескольких модальностей. Концептуально XClone может поддерживать не только геномные, но и транскриптомные данные, а также информацию о соматических мутациях и митохондриальной ДНК в клетках образца. Каждая из этих модальностей имеет клиническую ценность и позволяет лучше понимать эволюцию опухоли.

¹<https://github.com/broadinstitute/inferCNV/wiki>

Содержание

1	Аннотация	1
2	Обозначения, сокращения, основные определения	3
3	Введение	11
4	Материалы и методы	14
4.1	Вероятностная модель числа прочтений	14
4.2	Алгоритмы предобработки данных	16
4.2.1	Извлечение данных из BAM-файлов	16
4.2.2	Статистическое фазирование гаплотипов	17
4.2.3	Подходы к сегментации генома	18
4.2.4	Исправление ошибок смены цепи	21
4.3	Использованные данные	28
4.4	Первоначальная версия XClone: только ASE-модуль	28
4.4.1	Plate notation	28
4.4.2	Семплирование по Гиббсу	28
4.4.3	Предложенная модель, её недостатки	28
4.4.4	Поиск наиболее вероятной перестановки меток	33
4.5	Заключительная версия XClone: ASE- и RDR-модули	35
4.5.1	Вариационный байесовский вывод	35
4.5.2	Структура ASE-модуля	35
4.5.3	Структура RDR-модуля	41
4.5.4	Известные недостатки и планы по их исправлению	41
5	Благодарности	42
	Список использованных источников	43

2 Обозначения, сокращения, основные определения

В силу междисциплинарного характера данной дипломной работы, автор счел 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 Френсисом Криком в 1958 году.

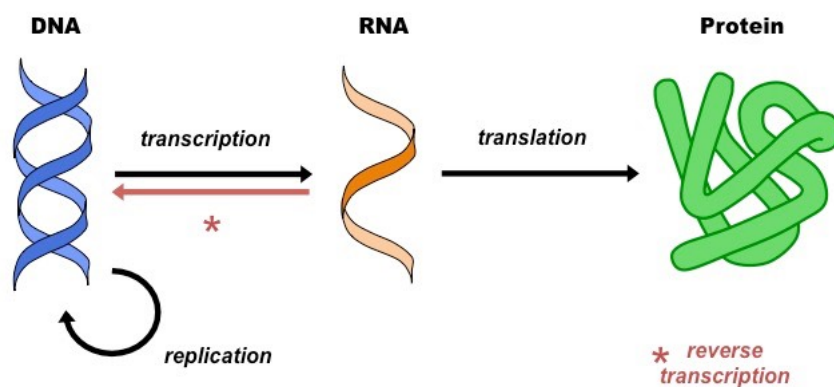


Рис. 2.1: Центральная догма молекулярной биологии

В упрощённом понимании, в процессе транскрипции участок ДНК преобразуется в т.н. **пре-матричную РНК (пре-мРНК)**, которая после **сплайсинга** — вырезания **интронов**, 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 экзона, 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 (rid)*). Короткий нуклеотидный фрагмент, распознанный секвенатором после ПЦР. В научном сообществе чаще используется транслитерация **rid** 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.

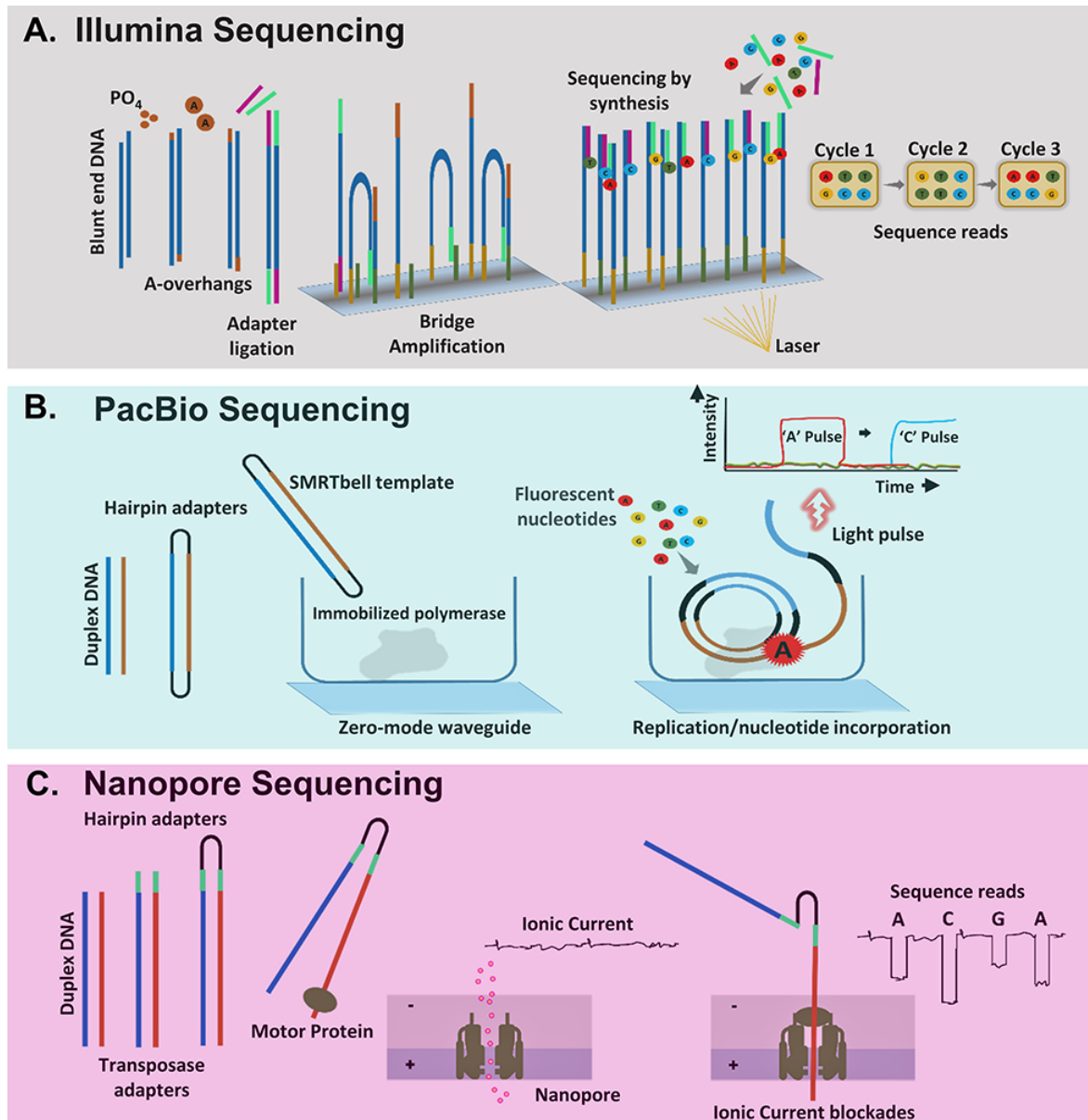


Рис. 2.2: Основные технологии высокопроизводительного секвенирования [1]. В данной работе использованы данные, полученные по протоколам Illumina (1) и Oxford Nanopore (3).

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 50-300б 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 (*Секвенирование одиночных клеток*). Совокупность новых методов секвенирования, позволяющих извлекать нуклеотидные последовательности из каждой из клеток образца в отдельности.

Определение 2.8 (*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 2.9 (*ОНП (снип)*). Однонуклеотидный полиморфизм — позиция в геноме, на которой в статистически значимых долях популяции встречаются несколько различных вариантов нуклеотидов. Могут существенно влиять на фенотип, в том числе быть причиной патологий. В сообществе более принята транслитерация **снип** от английского *SNP* — *single nucleotide polymorphism*.

Определение 2.10 (*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 2.11 (*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 2.12 (*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 2.13 (*Аллель*). Вариант фрагмента ДНК, встречающийся в статистически значимой доле популяции. Частные случаи — ОНП, варианты генов.

Определение 2.14 (*Аллельный дисбаланс*). Ситуация, когда один из аллелей доминирует над остальными — например, экспрессируется сильнее, более представлен в данных секвенирования и т.д.

Определение 2.15 (*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;

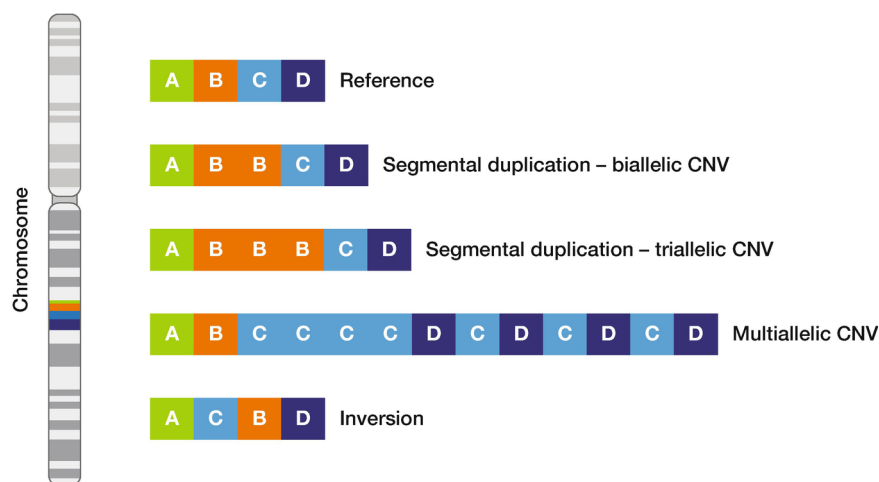
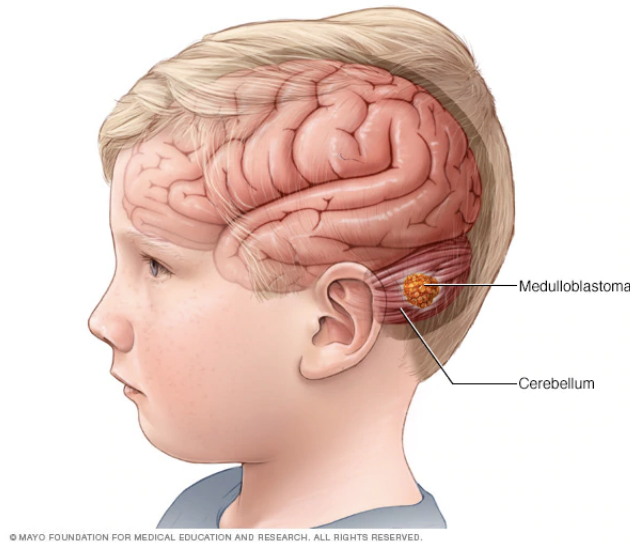


Рис. 2.3: Основные типы структурных вариаций

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 2.16 (*Медуллобластома*). Самый распространённый тип педиатрической опухоли мозга. Поражает мозжечок.



Определение 2.17 (*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.

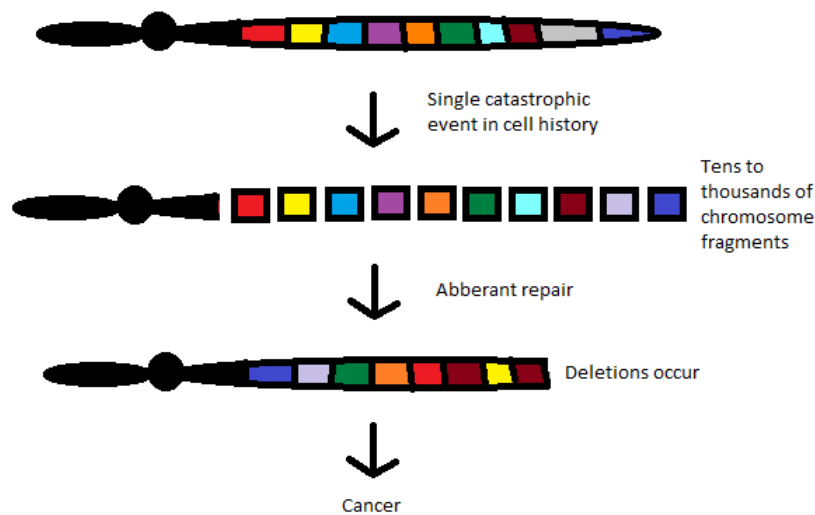


Рис. 2.4: Хромотрипсис

Определение 2.18 (*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.

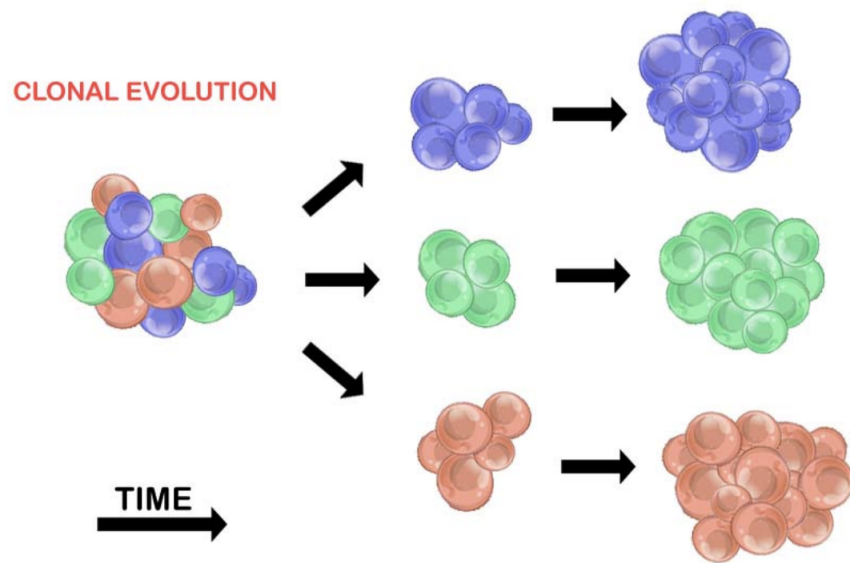


Рис. 2.5: Клональная эволюция опухоли: клетки первичной опухоли порождают т.н клональные линии. Клетки в пределах линии генетически однородны, но клетки двух любых различных линий существенно различны.

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 **дерева 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 2.19 (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[11][12], а 10X Genomics, основная компания-производитель оборудования для single-cell секвенирования, заняла 69 строчку в рейтинге 500 наиболее быстрорастущих компаний США². Возможность изучать биологические процессы в тканях на уровне отдельных клеток привела к прорывным открытиям во многих областях науки[5][8], особенно в персонализированной онкологии³. 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.

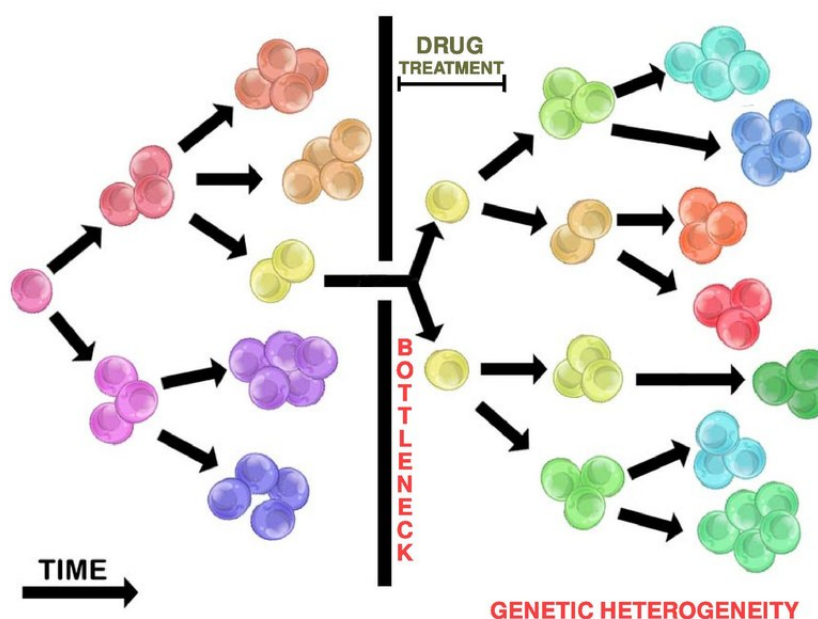


Рис. 3.1: При подборе терапии нужно учитывать клональный состав опухоли. Лечение может убивать некоторые клональные линии, но не действовать на остальные, из-за чего случается рецидив.

²<https://www.inc.com/inc5000/2019/top-private-companies-2019-inc5000.html>

³<https://www.nature.com/articles/d42473-019-00310-5>

В данной дипломной работе рассматривается байесовский подход к этой задаче. В ней предложен алгоритм байесовского машинного обучения XClone — графическая статистическая модель опухолевых образцов. В её задачи входит не только анализ клонального состава образца, но и поиск аллель-специфических структурных мутаций, по которым можно пронаблюдать эволюцию опухоли. Алгоритм XClone развивает идеи, заложенные доктором Янхуа Хуань в статьях "Cardelino"[10] и "Vireo"[14]. 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, хорошие результаты в этом направлении были получены только на синтетических данных. Метод был опробован на реальных данных, извлечённых из медуллобластомы — детской опухоли мозга, — полученных по протоколам компаний 10X Genomics и Smart-Seq, но концептуально он не привязан к конкретной платформе и может быть адаптирован к другим форматам входных данных.

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 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[15].

Вторая часть носит название "ASE-модуль где ASE означает "allele-specific expression". 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. 10,33–35. For example, copy-neutral loss of heterozygosity (LOH) – where one allele is lost

and the other duplicated so the total copy number remains 2 – is common in many cancers 33,36–41. Allele-specific copy numbers have also been shown to be essential for accurate inference of whole-genome duplications (WGDs)24,42 and timing WGDs relative to other CNAs10,24,43. Despite the demonstrated importance of allele-specific copy numbers, previous single-cell sequencing studies have assumed that low-coverage data is too shallow to obtain allele-specific information from single cells7,8,44–46. Existing methods for identifying CNAs from single-cell sequencing data6–8,45–50 are limited to the inference of total copy number, which indicates only the sum of copy numbers at each locus, by analyzing differences between the observed and expected number of sequencing reads aligned to a locus, or the read-depth ratio (RDR). The signal to detect allele-specific copy numbers is the B-allele frequency (BAF), or relative proportions of reads from the two alleles of a genomic region; however, standard methods to calculate the BAF from individual germline heterozygous SNPs do not work with extremely low coverage sequencing data.

По состоянию на июнь 2020 года, опубликован только один непосредственный аналог XClone — алгоритм CHISEL[15], 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.

4 Материалы и методы

4.1 Вероятностная модель числа прочтений

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.

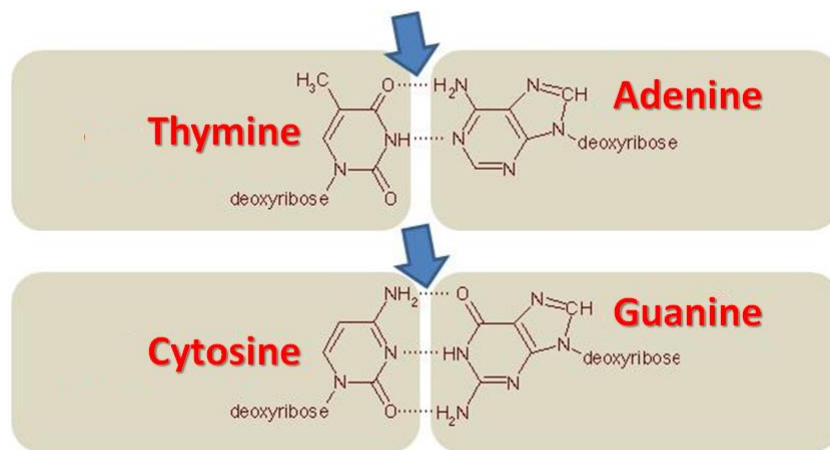


Рис. 4.2: Комплементарные пары: аденин-тимин и гуанин-цитозин

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 строки, 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.

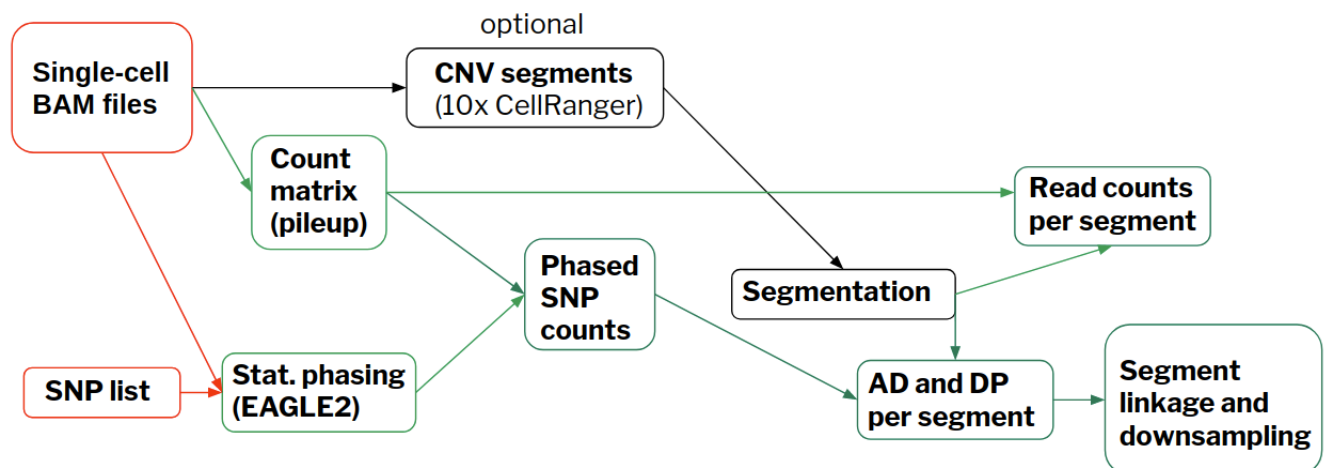


Рис. 4.3: Граф протокола предобработки данных для алгоритма XClone. Красным обозначены входные данные, чёрным — опциональные шаги, зелёным — реализованные стадии.

4.2.1 Извлечение данных из BAM-файлов

BAM — *binary SAM — binary sequence alignment/map format* — общепринятый формат сжатого хранения данных секвенирования, с подробностями которого можно ознакомиться в оригинальной публикации [7]. BAM-файл, полученный по протоколам 10x Genomics, занимает до нескольких терабайтов дискового пространства, потому эффективное извлече-

ние информации из BAM-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**⁴.

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**[9]⁵. Этот алгоритм основан

⁴<https://github.com/single-cell-genetics/cellSNP>

⁵<https://data.broadinstitute.org/alkesgroup/Eagle/>

на скрытых марковских моделях и использует 32,470 образца из базы данных **Haplotype Reference Consortium**[4].

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 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 устранении 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 ве-

роятностной модели числа прочтений. Наиболее существенным фактором 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**, алгоритм⁶ 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 покрывают некоторое подмножество референсного генома, которое затем подразделяется на сегменты размера 20-50 килобаз, в пределах которых вероятность ошибки смены цепи невелика, а потому сигнал аллельного дисбаланса статистически достоверный.

⁶https://support.10xgenomics.com/single-cell-dna/software/pipelines/latest/algorithms/cnv_calling

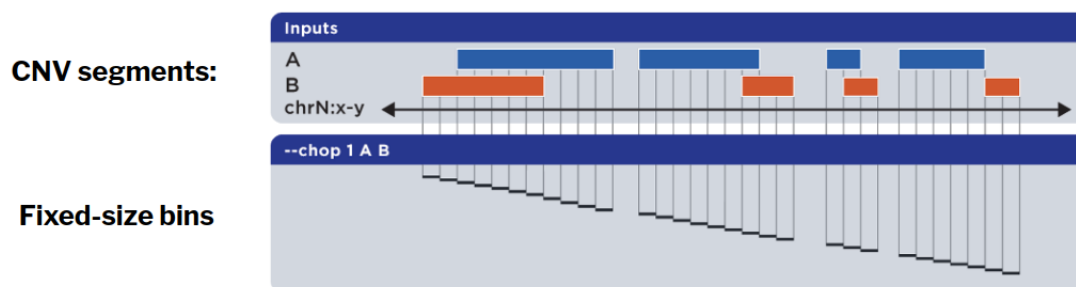


Рис. 4.4: Иллюстрация алгоритма сегментирования генома. Длина индивидуальных сегментов задаётся заранее и выбирается из диапазона 20-50кб. Каждые k подряд идущих фрагментов затем объединяются в блоки. Соответствующие подматрицы прочтений при этом суммируются с одновременной коррекцией ошибок смены цепи.

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 покрывать сразу несколько ОНП. Благодаря этому их гаплотипы определяются однозначно: просто из нуклеотидной последовательности ряда понятно, какие именно аллели лежат 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.

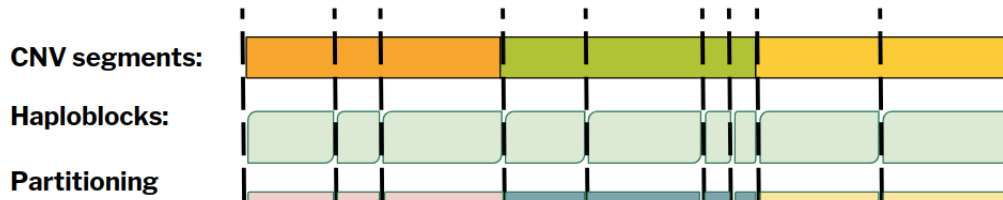


Рис. 4.5: Сегментирование генома в первоначальной версии XClone. Гаплотипические блоки, найденные по данным Oxford Nanopore DNA-seq, пересекаются с диапазонами структурных вариаций, найденных CellRanger DNA.

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 [3], 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 проблем. На первом шаге алгоритма происходит разбиение генома на непересекающиеся сплошные сегменты длины L . Затем каждые N подряд идущих сегментов объеди-

няются в блок длины NL . В пределах блока переключения моделируются бернуллиевскими случайными величинами, по одной на каждый сегмент. Параметры этих распределений, в свою очередь, выводятся **ЕМ-алгоритмом**. После исправления ошибок, прочтения сегментов внутри блока суммируются, что даёт более стабильный сигнал. Эта идея была сформулирована в [15], но технические детали были осознанно исключены авторами CHISEL из препринта.

Прежде чем приступать к рассмотрению метода, сформулируем необходимые определения:

Определение 4.4 (*ЕМ-алгоритм*).

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

Algorithm 1: ЕМ-алгоритм в общем виде

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

$t = 0$;

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

while $Q(\Theta^{(t+1)} \mid \Theta^{(t+1)}) - Q(\Theta^{(t)} \mid \Theta^{(t)}) > \varepsilon$ **do**

$\mathcal{L}(\Theta^{(t)}; \mathbf{Z}, \mathbf{X}) := p(\mathbf{X}, \mathbf{Z} \mid \Theta^{(t)})$;

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

$\Theta^{(t+1)} := \arg \max_{\Theta} Q(\Theta \mid \Theta^{(t)})$ // М-шаг

$t = t + 1$

end

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

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 математического ожидания логарифма совместной функции правдоподобия по условному распределению $\mathbf{Z} \mid \mathbf{X}, \Theta^{(t)}$;
2. **М-шаг**, на котором параметры нового апостериорного распределе-

ния $\Theta^{(t+1)}$ 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 ([13], 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 корректности 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 распределения на М-шаге ЕМ-алгоритма имеют вид:

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

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

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

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

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

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

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 случайного вектора \mathbf{Z} независимы в совокупности, шаги ЕМ-алгоритма имеют вид:

Е-шаг:

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

М-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 $Q(\Theta \mid \Theta^{(t)})$ ([13], стр. 363-364). \square

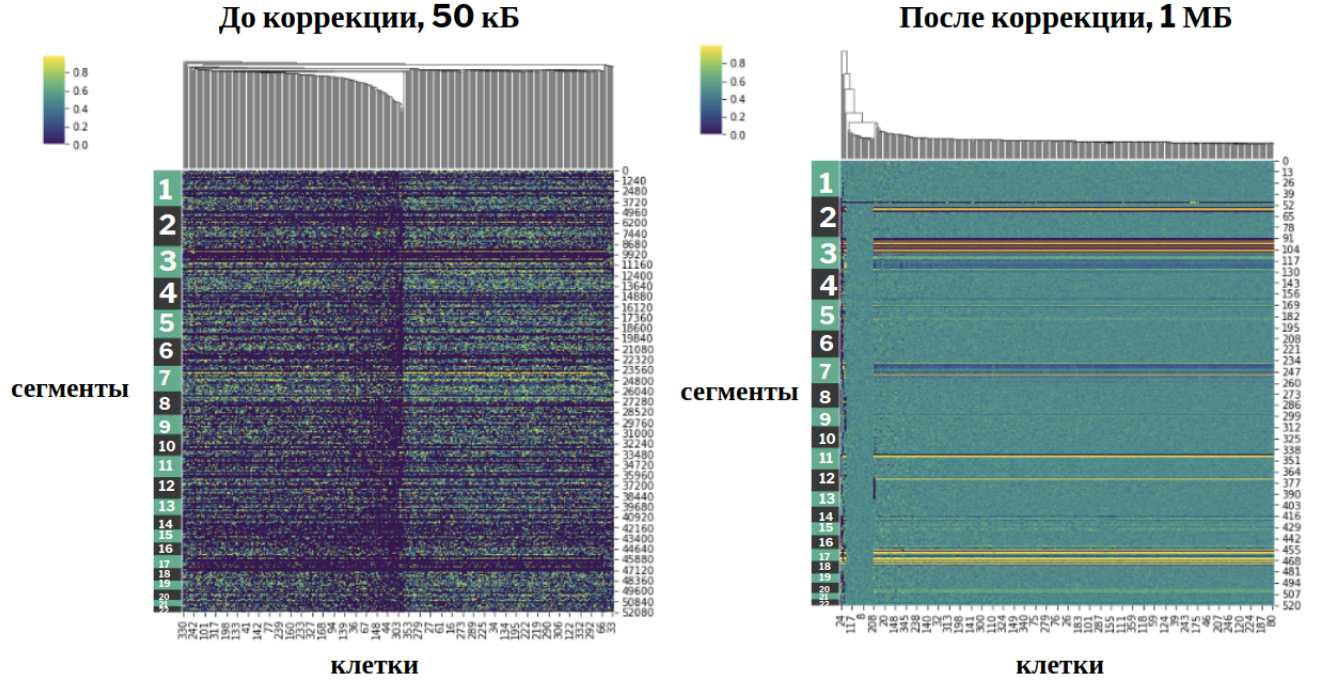


Рис. 4.6: Коррекция ошибок смены цепи на примере ДНК-образца STR-Nuclei. На рисунке изображены тепловые карты долей аллеля из «материнского» гаплотипа, числа слева обозначают номер хромосомы. Картина аллельного дисбаланса до коррекции практически не прослеживается, после — становится очевидна.

После того, как значения z_1, \dots, z_N были определены по данным DNA-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 Используемые данные

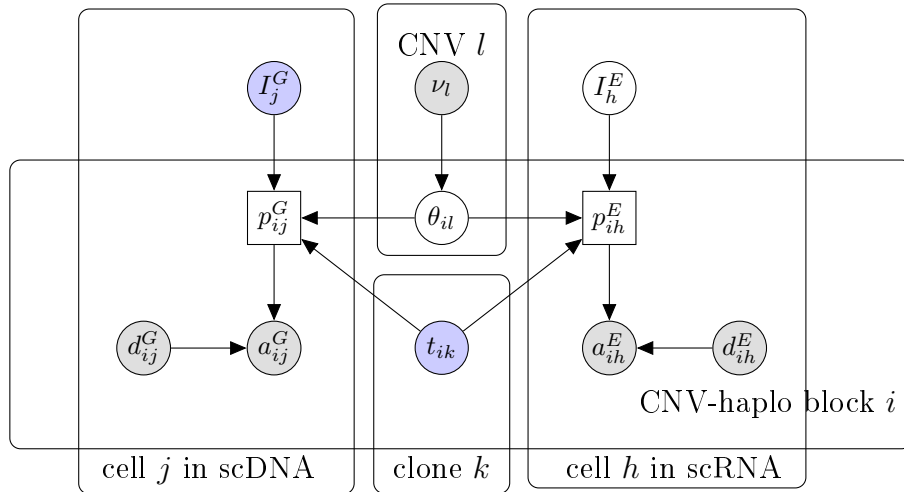
4.4 Первоначальная версия XClone: только ASE-модуль

4.4.1 Plate notation

4.4.2 Семплирование по Гиббсу

4.4.3 Предложенная модель, её недостатки

scDNA + CNV and scRNA, different samples



In this model, clonal assignment \mathbf{I}^G of cells in scDNA sample is assumed fixed. For each clone k , for each same-CNV block i the clonal CNV state $\mathbf{T}_{i,k}$ is defined to be the most frequent CNV state in that position across all cells assigned to clone k . Clonal assignment \mathbf{I}^E in scRNA sample is learned.

This model learns the clonal structure in scRNA sample using ASE profiles from scDNA.

Basic definitions

1. Constants:

- M^G, M^E — number of cells in scDNA and scRNA samples respectively.
- K — estimated number of clones in the sample.
- N — number of same-CNV blocks in scDNA sample (also used in scRNA sample).
- T_{\max} — maximal possible CNV number. User-defined with the default of 5.
- τ — set of possible CNV configurations:

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

The case of zero CNV number should be treated with care if we can't say for sure whether the part of the chromosome (both arms) is deleted).

2. Other known quantities:

- $\mathbf{D}^G, \mathbf{D}^E$ — total read counts. To get a count of the block, one simply adds up the counts of the variants within the block. Here we assume that variants are far enough from each other, so that almost no reads overlap two variants at the same time. Otherwise, adding things up wouldn't make sense.
- $\mathbf{A}^G, \mathbf{A}^E$ — same for allele-specific counts.
- $\mathbf{f} = (f_1, \dots, f_K)$ — K -dimensional vector of estimated clonal fractions ($\sum_{i=1}^K f_i = 1$)
- \mathbf{T} — CNV states of blocks ($\mathbf{T}_{i,k}$ is a CNV state of a block i in clone k).

3. Inferred quantities:

- Θ — allelic rates of variants located within blocks of fixed CNV status. $\Theta_{i,t}$ is an allelic rate of a block i with a CNV status t . Set of blocks for each clone is unique. Blocks are ordered as tuples of the form (*block start, block length*).

- $\mathbf{I}^E \in [K]^M$ — cell-to-clone assignment in scRNA sample.

4. Some notational conventions:

- Capitalized letter without superscript (like Θ) denotes the information for both samples.
- $\mathbf{H}^G, \mathbf{H}^E$ — CNV status of the blocks in accordance with the current label assignment:

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

- $\mathbf{X}^G, \mathbf{X}^E$ — a shortcut to simplify the notation: $\mathbf{X}_{i,j}^{G|E}$ is an allelic rate of a block i in cell j based of the current cell-to-clone label assignment.

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

Generative model formulation

- Cell-to-clone assignment posterior:

$$P(\mathbf{I}_j^E = k_0 \mid \mathbf{A}_j, \mathbf{D}_j, \mathbf{f}, \Theta) = \frac{P(\mathbf{A}_j \mid \mathbf{D}_j, \mathbf{I}_j^E = k_0, \Theta)P(\mathbf{I}_j^E = k_0 \mid \mathbf{f})}{\sum_{k=1}^K P(\mathbf{A}_j \mid \mathbf{D}_j, \mathbf{I}_j^E = k, \Theta)P(\mathbf{I}_j^E = k \mid \mathbf{f})} \quad (2)$$

- ASE model:

$$\begin{aligned} P(\mathbf{A}_{i,j}^G \mid \mathbf{D}_{i,j}^G, \Theta) &= \text{Binom}(\mathbf{A}_{i,j}^G \mid \mathbf{D}_{i,j}^G, \mathbf{X}_{i,j}^G) \\ P(\mathbf{A}_{i,j}^E \mid \mathbf{D}_{i,j}^E, \Theta) &= \text{Binom}(\mathbf{A}_{i,j}^E \mid \mathbf{D}_{i,j}^E, \mathbf{X}_{i,j}^E) \end{aligned} \quad (3)$$

- ASE likelihood (both terms factorize over variants):

$$\begin{aligned} P(\mathbf{A}_j^G \mid \mathbf{D}_j^G, \mathbf{I}_j^G = k^G, \Theta^E) &= \prod_{i=1}^N \text{Binom}(\mathbf{A}_j^G \mid \mathbf{D}_j^G, \Theta_{i,k^G}) \\ P(\mathbf{A}_j^E \mid \mathbf{D}_j^E, \mathbf{I}_j^E = k^E, \Theta) &= \prod_{i=1}^N \text{Binom}(\mathbf{A}_j^E \mid \mathbf{D}_j^E, \Theta_{i,k^E}) \end{aligned} \quad (4)$$

- **Allelic rate likelihood:**

$$\begin{aligned} \mathcal{L}(\Theta) = & \left(\prod_{j=1}^{M^G} \sum_{k^G=1}^K P(\mathbf{A}_j^G \mid \mathbf{D}_j^G, \mathbf{I}_j^G = k^G, \Theta) \right) \times \\ & \times \left(\prod_{j=1}^{M^E} \sum_{k^E=1}^K P(\mathbf{A}_j \mid \mathbf{D}_j^E, \mathbf{I}_j^E = k^E, \Theta) \cdot P(\mathbf{I}_j^E = k^E \mid \mathbf{f}) \right) \end{aligned} \quad (5)$$

To view the clonal assignment in a Bayesian way, we introduce informative prior ν for Θ .

Using that «posterior \propto prior \times likelihood», we obtain:

$$\begin{aligned} P(\Theta \mid \mathbf{A}, \mathbf{D}, \mathbf{f}, \nu) & \propto P(\Theta \mid \nu) \times \mathcal{L}(\Theta) = \\ & = \prod_{l=1}^N \prod_{t \in \tau} \text{Beta}(\alpha_{l,\tau}, \beta_{l,\tau}) \times \mathcal{L}(\Theta) \end{aligned} \quad (6)$$

Parameters α_t, β_t are selected in such a way that the mode of $\text{Beta}(\alpha_t, \beta_t)$ equals to $1/t$ ⁷.

Selecting a prior for Θ CNV state of t hides a plethora of possible configurations: it can mean " k copies of maternal chromosome and $t - k$ copies of paternal" for any $k \in \{0, \dots, t\}$, all the variants are possible. But they are not equally possible: some are more supported by evidence than the rest. During initialization, for each block in each clone we should find the (k, t) -configuration (k_0, t) , such that k_0/t is as close to the observed ASE ratio as possible. Then we choose values (α, β) such that the mode of $\text{Beta}(\alpha, \beta)$, given by $(\alpha - 1)/(\alpha + \beta - 2)$, equals to k_0/t . That means, we must solve the following problem:

$$\frac{\alpha - 1}{\alpha + \beta - 2} = \frac{k_0}{t}, \alpha \geq 1, \beta \geq 1$$

Let's derive the solution. If $k_0 = 0$, it is clear that $\alpha = 1$, while any $\beta > 1$

⁷because if we assume that allelic rates only depend on the CNV status t then those rates could be computed as $1/t$

works⁸. Otherwise:

$$\begin{aligned}
 (\alpha - 1)t &= (\alpha + \beta - 2)k_0 \\
 k_0\beta &= (t - k_0)\alpha - t + 2k_0 \\
 \beta &= \left(\frac{t}{k_0} - 1\right)\alpha - \left(\frac{t}{k_0} - 2\right) \\
 \implies \alpha &= 1 + \frac{t - 2k_0}{t - k_0}, \beta = 1
 \end{aligned} \tag{7}$$

As β is linearly dependent from α , any increase in α will pull β up, "sharpening" the shape of the distribution and making it more biased, thereby we decided to choose the minimal feasible α .

Inference (Gibbs sampler) To use a Gibbs sampler, we define conditional probability distribution for each scalar random variable:

Cell-to-clone label assignment:

$$P(\mathbf{I}_j^E = k \mid \mathbf{I}_{-j}^E, \mathbf{A}^E, \mathbf{D}^E, \mathbf{f}, \boldsymbol{\Theta}) \propto P(\mathbf{A}_j^E \mid \mathbf{D}_j^E, \mathbf{I}_j^E = k, \boldsymbol{\Theta}) \cdot P(\mathbf{I}_j^E = k \mid \mathbf{f}) \tag{8}$$

Allelic rates: Assuming fixed assignment, let's expand the joint likelihood equation:

$$\begin{aligned}
 P(\boldsymbol{\Theta} \mid \mathbf{A}, \mathbf{D}, \mathbf{I}^G, \mathbf{I}^E, \mathbf{f}, \nu) &\propto \\
 &\propto \left\{ \prod_{l=1}^N \prod_{t \in \tau} \text{Beta}(\alpha_{T_{l,t}}, \beta_{T_{l,t}}) \right\} \left[\prod_{j^G=1}^{M^G} P(\mathbf{A}_{j^G}^G \mid \mathbf{D}_{j^G}^G, \mathbf{I}_{j^G}^G, \boldsymbol{\Theta}) \right] \left[\prod_{j^E=1}^{M^E} P(\mathbf{A}_{j^E}^E \mid \mathbf{D}_{j^E}^E, \mathbf{I}_{j^E}^E, \boldsymbol{\Theta}) \right] \\
 &= \prod_{l=1}^N \prod_{t \in \tau} \text{Beta}(\alpha_{T_{l,t}}, \beta_{T_{l,t}}) \left[\prod_{j^G=1}^{M^G} \text{Binom}(\mathbf{A}_{j^G}^G \mid \mathbf{D}_{j^G}^G, \boldsymbol{\Theta}_{i,t}) \right] \left[\prod_{j^E=1}^{M^E} \text{Binom}(\mathbf{A}_{j^E}^E \mid \mathbf{D}_{j^E}^E, \boldsymbol{\Theta}_{i,t}) \right] \\
 &= \prod_{l=1}^N \prod_{t \in \tau} \left[\text{Beta}(\alpha_{T_{l,t}}, \beta_{T_{l,t}}) \prod_{j^G=1}^{M^G} \prod_{j^E=1}^{M^E} \left(\text{Binom}(\mathbf{A}_{l,j^G}^G \mid \mathbf{D}_{l,j^G}^G, \mathbf{X}_{l,j^G}^G)^{\mathbb{I}\{\mathbf{H}_{l,j^G}^G=t\}} \times \right. \right. \\
 &\quad \left. \left. \times \text{Binom}(\mathbf{A}_{l,j^E}^E \mid \mathbf{D}_{l,j^E}^E, \mathbf{X}_{l,j^E}^E)^{\mathbb{I}\{\mathbf{H}_{l,j^E}^E=t\}} \right) \right]
 \end{aligned} \tag{9}$$

From here we derive update rules for individual allelic rates:

$$\boldsymbol{\Theta}_{l,t} \mid \mathbf{I}^G, \mathbf{I}^E \sim \text{Beta}(\alpha_{T_{l,t}} + u_{l,t}, \beta_{T_{l,t}} + v_{l,t}) \tag{10}$$

⁸Nevertheless, it is not clear which one to choose. As we try to reduce prior bias, let's set it to be equal $1 + \varepsilon$ for some reasonable $\varepsilon > 0$

where

$$\begin{aligned}
 u_{l,t} &= \sum_{j^G=1}^M \mathbf{A}_{l,j^G}^G \cdot \mathbb{I} \left\{ \mathbf{H}_{l,j^G}^G = t \right\} + \sum_{j^E=1}^M \mathbf{A}_{l,j^E}^E \cdot \mathbb{I} \left\{ \mathbf{H}_{l,j^E}^E = t \right\} \\
 v_{l,t} &= \sum_{j^G=1}^{M^G} (\mathbf{D}_{l,j^G}^G - \mathbf{A}_{l,j^G}^G) \cdot \mathbb{I} \left\{ \mathbf{H}_{l,j^G}^G = t \right\} + \sum_{j^E=1}^M (\mathbf{D}_{l,j^E}^E - \mathbf{A}_{l,j^E}^E) \cdot \mathbb{I} \left\{ \mathbf{H}_{l,j^E}^E = t \right\}
 \end{aligned} \tag{11}$$

4.4.4 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 перестановки 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 $M = 1.3 \times 10^6$ ⁹, и типичное

⁹<https://www.10xgenomics.com/blog/our-13-million-single-cell-dataset-is-ready-to-download>

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 $\arg \min_{\sigma \in S_K} \sum_{j=1}^K d(P_j, Q_j)$, очевидно по построению. Задача поиска совершенного паросочетания минимального веса в двудольном графе решается — это т.н. **задача о назначениях**, одна из фундаментальных задач комбинаторной оптимизации. Для её решения применяется так называемый "венгерский алгоритм опубликованный в 1955 году американским математиком Гарольдом Куном[6].



Рис. 4.7: 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 клональных меток.

4.5 Заключительная версия XClone: ASE- и RDR-модули

4.5.1 Вариационный байесовский вывод

4.5.2 Структура ASE-модуля

In the xclone model, we aim to cluster cells in scRNA-seq data by its copy number variation (CNV) states across many CNV blocks. The similarity between two CNV states are measured at both allelic fraction and expression

count levels, which reflects both the maternal and paternal copy numbers $c_t = [c_{t,1}, c_{t,2}]$ for CNV state t . The key novelty of xclone is that it allows to integrate additional information on CNV states and the according parameters of allelic fraction and expression count, for example extracting from scDNA-seq data.

Due to different magnitudes of expressions and allelic bias across genes even with the same CNV block, we define the model at per gene basis. If a CNV block contains multiple genes, we will introduce separate functions to link CNV states for these neighbouring genes.

As most scRNA-seq data has very low coverage, especially for the droplet based protocols, it is hard to accurately estimate the allelic fraction from expression for most individual heterozygous variants. Therefore, we aggregate multiple variants across one gene by using statistical phasing with haplotype reference. This may increase the power of allelic fraction estimation, though we need to avoid multiple counting when one read covering several nearby SNPs. Once again, multiple genes within one CNV block could be further aggregated, through a separate linkage function.

Allelic fraction module By aggregating multiple SNPs for each gene, we could have observations on N genes across M cells in scRNA-seq data. For gene i in cell j , we denote the read (or UMI) count for alternative allele as $a_{i,j}$ and the total depth (i.e., for both alternative and reference alleles) as $d_{i,j}$. Assuming there are T CNV states and for each state t the maternal and paternal copy number vector $c_t = [c_{t,1}, c_{t,2}]$. If gene i in cell j is in CNV state t , we assume the allelic expression follows a binomial distribution with CNV state specific parameter θ_t , as follows,

$$p(a_{i,j}|d_{i,j}, \theta_t) = \text{Binom}(a_{i,j}|d_{i,j}, \theta_{i,t}). \quad (12)$$

Let A and D respectively denote the gene-by-cell count matrices for the alternative allele and the total read depths, and each cell comes from one of the K clones. If cell j is from clone k , the cell identity variable $z_{j,k} = 1$ otherwise 0. Also, if gene i in clone k is in CNV state t , the corresponding CNV state variable $y_{i,k,t} = 1$ otherwise 0. Their according matrix or tensor Z and Y are unknown variables. Now, we could define the likelihood for allelic

fraction as follows, TODO

$$p(A, D|Z, Y, \boldsymbol{\theta}) = \prod_{i=1}^N \prod_{j=1}^M \prod_{k=1}^K \prod_{t=1}^T p(a_{i,j}|d_{i,j}, \theta_{i,t})^{z_{j,k} \times y_{i,k,t}} \quad (13)$$

Here, we introduce uniform **Multinomial** distribution for Z and Y and **Beta** distribution for $\boldsymbol{\theta}$ with default hyper-parameters $\alpha_t = (c_{t,1} + 0.01)$ and $\beta_t = (c_{t,2} + 0.01)$ as follows,

$$\begin{aligned} p(z_{j,k} = 1|\boldsymbol{\pi}) &= \text{Multinom}(1; \boldsymbol{\pi}_j) = \pi_{j,k} \\ p(y_{i,k,t} = 1|U) &= \text{Multinom}(1; \mathbf{u}_{i,k}) = u_{i,k,t} \\ p(\theta_{i,t}|\alpha_t, \beta_t) &= \text{Beta}(\theta_{i,t}|\alpha_t, \beta_t) \end{aligned} \quad (14)$$

For a start point, we only consider $T = 16$ CNV states:

$\{[0,0], [0,1], [0,2], [0,3], [1,0], [1,1], [1,2], [1,3], [2,0], [2,1], [2,2], [2,3], [3,0], [3,1], [3,2], [3,3]\}$

Expression count module For a particular cell j , we assume that the expression counts for N genes follows a **Multinomial** distribution with total counts as $d_j = \sum_{i=1}^N d_{i,j}$ and probability vector as $\mathbf{f}_j = \{f_{1,j}, \dots, f_{N,j}\}$, hence $\sum_{i=1}^N f_{i,j} = 1$. In the case of diploid genome, $f_{i,j} = m_i$.

When CNVs exist in the consensual genome for clone k , the amplification factor $\boldsymbol{\gamma}$ needs to take into account, and the probability to observe the read counts for N genes can be expressed as

$$\begin{aligned} \mathbf{d}_j &:= [d_{1,j}, \dots, d_{N,j}] \\ p(\mathbf{d}_j|d_j, \mathbf{f}_j) &:= \text{Multinom}(\mathbf{d}_j|d_j, \mathbf{f}_j); \\ f_{i,k} &:= \frac{\sum_{t=1}^T m_i \exp[\gamma_t] P(y_{i,k,t} = 1)}{\sum_{b=1}^N \sum_{t=1}^T m_b \exp[\gamma_t] P(y_{b,k,t} = 1)} \end{aligned} \quad (15)$$

As can be seen that the diploid expression profile \mathbf{m} is shared by all cells, the likelihood for M cells jointly can be written as follows,

$$p(D|Z, Y, \mathbf{m}, \boldsymbol{\gamma}) := \prod_{j=1}^M \prod_{k=1}^K p(\mathbf{d}_j|d_j, \mathbf{f}_k)^{z_{j,k}} \quad (16)$$

Here, we introduce **Dirichlet**($\boldsymbol{\omega}$) as the prior distribution for \mathbf{m} , and Gaussian process **GP**($\boldsymbol{\mu}, \Sigma$) as the prior distribution for $\boldsymbol{\gamma}$. The GP prior has fixed hyper-parameters $\mu_t = \log((c_{t,1} + c_{t,2})/2 + \varepsilon)$ for CNV state t , where ε is some

reasonably small offset, and the covariance matrix Σ is defined by the radial basis function kernel $K(\mathbf{c}_t, \mathbf{c}_{t'}) = l_1 \exp\{-l_2[(c_{t,1} - c_{t',1})^2 + (c_{t,2} - c_{t',2})^2]\}$, as follows,

$$\begin{aligned} \mathbf{m}|\boldsymbol{\omega} &\sim \text{Dirichlet}(\boldsymbol{\omega}) \\ \boldsymbol{\gamma}|\boldsymbol{\mu}, \Sigma &\sim \mathcal{N}(\boldsymbol{\mu}, \Sigma) \end{aligned} \quad (17)$$

Posterior and inference Now, by combining the allelic fraction and expression count likelihoods and their specific prior distributions, we could have the posterior distribution as follows,

$$\begin{aligned} p(Z, Y, \boldsymbol{\theta}, \mathbf{m}, \boldsymbol{\gamma}|A, D) &\propto p(Z, Y, \boldsymbol{\theta}, \mathbf{m}, \boldsymbol{\gamma}) p(A, D|Z, Y, \boldsymbol{\theta}, \mathbf{m}, \boldsymbol{\gamma}) \\ &= p(Z, Y, \boldsymbol{\theta}, \mathbf{m}, \boldsymbol{\gamma}) p(D|Z, Y, \mathbf{m}, \boldsymbol{\gamma}) p(A, D|Z, Y, \boldsymbol{\theta}) \end{aligned} \quad (18)$$

where the prior distribution is independent for all five variables $\Omega = \{Z, Y, \boldsymbol{\theta}, \mathbf{m}, \boldsymbol{\gamma}\}$, hence can be multiplied from Eq (14) and Eq (17). Similarly, the likelihood term can be multiplied by the two likelihood terms in Eq(13) and Eq(15).

Here, we introduce a variational Bayes to infer the posterior of the five unknown variables $p(\Omega|A, D)$. Then the inference problem becomes an optimisation problem for minimising the evidence lower bound (ELBO), as follows

$$\mathcal{L}(q) = -\mathcal{KL}(q(\Omega)||p(\Omega)) + \int q(\Omega) \log p(A, D|\Omega) d\Omega \quad (19)$$

where the variational distribution $q(\Omega)$ is in the same form as its prior distribution Eq (14) and Eq (17), while we aim to optimize their parameters to achieve the highest ELBO. The details of the derivations can be found in page 463 in Bishop, PRML 2006 or Vireo paper (TODO: introduce it here).

For the right part, we can split it into the allelic fraction and expression count modules, namely

$$\begin{aligned} \int q(\Omega) \log p(A, D|\Omega) d\Omega &= \int q(Z, Y, \boldsymbol{\theta}) \log p(A, D|Z, Y, \boldsymbol{\theta}) dZ dY d\boldsymbol{\theta} + \\ &\quad \int q(Z, Y, \mathbf{m}, \boldsymbol{\gamma}) \log p(D|Z, Y, \mathbf{m}, \boldsymbol{\gamma}) d\mathbf{m} d\boldsymbol{\gamma} \end{aligned} \quad (20)$$

The allelic fraction module could further analytically written as follows,

$$\begin{aligned}\mathbb{E}_{Z,Y,\theta}[\log p(A,D|Z,Y,\theta)] &= \int q(Z,Y,\theta) \log p(A,D|Z,Y,\theta) dZ dY d\theta \\ &= \sum_{i=1}^N \sum_{j=1}^M \sum_{k=1}^K \sum_{t=1}^T \left\{ \tilde{\pi}_{j,k} \tilde{u}_{i,k,t} [w_{i,j} + a_{i,j} \varphi(\tilde{\alpha}) + b_{i,j} \varphi(\tilde{\beta}) - d_{i,j} \varphi(\tilde{\alpha} + \tilde{\beta})] \right\}\end{aligned}\quad (21)$$

where $w_{i,j} = \log \binom{d_{i,j}}{a_{i,j}}$ is the logarithm of binomial coefficient, $\varphi(\cdot)$ is the digamma function, and $\tilde{\cdot}$ is the parameters for variational posterior distributions.

This part is identical to Vireo model.

Supplementary Materials Derivation of likelihood equation (3.9, ASE module) The allelic fraction module could be analytically derived as follows,

$$\begin{aligned}& \int q(Z,Y,\theta) \log p(A,D|Z,Y,\theta) dZ dY d\theta \\ &= \mathbb{E}_{Z,Y,\theta}[\log p(A,D|Z,Y,\theta)] \\ &= \mathbb{E}_{Z,Y,\theta}[\log \prod_{i=1}^N \prod_{j=1}^M \prod_{k=1}^K \prod_{t=1}^T p(a_{i,j}|d_{i,j}, \theta_t)^{z_{j,k} \times y_{i,k,t}}] \\ &= \mathbb{E}_{Z,Y,\theta} \sum_{i=1}^N \sum_{j=1}^M \sum_{k=1}^K \sum_{t=1}^T [z_{j,k} y_{i,k,t} \log \text{binom}(a_{i,j}|d_{i,j}, \theta_t)] \\ &= \sum_{i=1}^N \sum_{j=1}^M \sum_{k=1}^K \sum_{t=1}^T \mathbb{E}_{Z,Y,\theta} [z_{j,k} y_{i,k,t} \log \text{binom}(a_{i,j}|d_{i,j}, \theta_t)] \\ &= \sum_{i=1}^N \sum_{j=1}^M \sum_{k=1}^K \sum_{t=1}^T \mathbb{E}_Z [z_{j,k}] \mathbb{E}_Y [y_{i,k,t}] \mathbb{E}_\theta [\log \text{binom}(a_{i,j}|d_{i,j}, \theta_t)] \\ &= \sum_{i=1}^N \sum_{j=1}^M \sum_{k=1}^K \sum_{t=1}^T \left\{ \tilde{\pi}_{j,k} \tilde{u}_{i,k,t} \mathbb{E}_\theta [\log \text{binom}(a_{i,j}|d_{i,j}, \theta_t)] \right\} \\ &= \sum_{i=1}^N \sum_{j=1}^M \sum_{k=1}^K \sum_{t=1}^T \left\{ \tilde{\pi}_{j,k} \tilde{u}_{i,k,t} [w_{i,j} + a_{i,j} \varphi(\tilde{\alpha}_t) + b_{i,j} \varphi(\tilde{\beta}_t) - d_{i,j} \varphi(\tilde{\alpha}_t + \tilde{\beta}_t)] \right\}\end{aligned}\quad (22)$$

The integral on **binom** distribution can be derived by the use of digamma

function, as follows.

$$\begin{aligned}
& \mathbb{E}_\theta[\log p(a|d, \theta)] \\
&= \mathbb{E}_\theta[\log \text{binom}(a; d, \theta)] \\
&= \mathbb{E}_\theta[\log \binom{a}{d} + a \log \theta + (d - a) \log (1 - \theta)] \\
&= \log \binom{a}{d} + a \mathbb{E}_\theta[\log \theta] + (d - a) \mathbb{E}_\theta[\log (1 - \theta)] \\
&= \log \binom{a}{d} + a \mathbb{E}_\theta[\log \theta] + (d - a) \mathbb{E}_{1-\theta}[\log (1 - \theta)] \\
&= \log \binom{a}{d} + a(\varphi(\alpha) - \varphi(\alpha + \beta)) + (d - a)(\varphi(\beta) - \varphi(\alpha + \beta)) \\
&= \log \binom{a}{d} + a\varphi(\alpha) + (d - a)\varphi(\beta) - d\varphi(\alpha + \beta)
\end{aligned} \tag{23}$$

where $\theta \sim \text{Beta}(\alpha, \beta)$ and $1 - \theta \sim \text{Beta}(\beta, \alpha)$. By using digamma function $\varphi(\cdot)$, one could easily calculate the logarithm of the geometric mean $\mathbb{E}[\log \theta] = \varphi(\alpha) - \varphi(\alpha + \beta)$ (see derivation on wikipedia for Beta distribution).

Derivation of likelihood equation (3.9, RDR module)

$$\begin{aligned}
& \int q(Z, Y, \mathbf{m}, \gamma) \log p(D|Z, Y, \mathbf{m}, \gamma) d\mathbf{m}, d\gamma = \\
&= \mathbb{E}_{Z, Y, \mathbf{m}, \gamma} \log p(D|Z, Y, \mathbf{m}, \gamma) = \\
&= (\text{Using eq. 16}) = \\
&= \mathbb{E}_{Z, Y, \mathbf{m}, \gamma} \left[\sum_{j=1}^M \sum_{k=1}^K \log p(\mathbf{d}_j | d_j, \mathbf{f}_k)^{z_{j,k}} \right] = \\
&= \sum_{j=1}^M \sum_{k=1}^K \mathbb{E}_{Z, Y, \mathbf{m}, \gamma} \left[z_{j,k} \cdot \log \left(d_j! \prod_{i=1}^N \frac{\tilde{f}_{i,k}^{d_{i,j}}}{d_{i,j}!} \right) \right] \propto \\
&\propto (\text{Dropping out constant terms}) \propto \\
&\propto \sum_{i=1}^N \sum_{j=1}^M \sum_{k=1}^K \mathbb{E}_{Z, Y, \mathbf{m}, \gamma} \left[z_{j,k} \cdot d_{i,j} \log \tilde{f}_{i,k} \right] = \\
&= \sum_{i=1}^N \sum_{j=1}^M \sum_{k=1}^K \tilde{\pi}_{j,k} \cdot d_{i,j} \cdot \mathbb{E}_{\mathbf{m}, \gamma} \log \tilde{f}_{i,k}
\end{aligned} \tag{24}$$

4.5.3 Структура RDR-модуля

4.5.4 Известные недостатки и планы по их исправлению

5 Благодарности

Данная работа посвящена приложениям байесовских методов машинного обучения к актуальной задаче вычислительной онкологии — задаче восстановления клональной структуры опухоли по данным высокопроизводительного 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 Притыкина¹⁰. 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 Штегле¹¹, профессора Heidelberg University¹² (Университет Хайдельберга, Германия) а также действующих и бывших сотрудников его научных групп в DKFZ¹³ (Немецкий Центр Онкологических Исследований), EMBL Heidelberg¹⁴ (Европейская Лаборатория Молекулярной Биологии) и HKU¹⁵ (Университет Гонконга). Формальным научным руководителем следует считать к.ф.-м.н. Yuanhua Huang¹⁶, 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 благодарность профессору Штегле и к.ф.-м.н. Ханне Сюзак, Николе Казирахи, Родриго Гонцало Парра, а также Д. О. Бредихину и В. А. Огородникову за ценные замечания и советы, придавшие многим аспектам метода завершённый, логически стройный вид.

¹⁰<https://scholar.google.com/citations?user=Arx56RkJBrYC&hl=en>

¹¹<https://scholar.google.com/citations?user=ClSXZ4IAAAAJ&hl=en>

¹²<https://www.uni-heidelberg.de/en>

¹³<https://www.dkfz.de/en/index.html>

¹⁴<https://www.embl.de/>

¹⁵<https://www.hku.hk/>

¹⁶<https://www.sbms.hku.hk/staff/yuanhua-huang>

Список использованных источников

- [1] 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.
- [2] «CaSpER identifies and visualizes CNV events by integrative analysis of single-cell or bulk RNA-sequencing data.» В: *Nature Communications* 89 (нояб. 2020). DOI: 10.1038/s41467-019-13779-x. URL: <https://doi.org/10.1038/s41467-019-13779-x>.
- [3] Y. Choi и др. «Comparison of phasing strategies for whole human genomes». В: *PLOS GENETICS* (апр. 2018). DOI: 10.1371/journal.pgen.1007308. URL: <https://doi.org/10.1371/journal.pgen.1007308>.
- [4] Haplotype Reference Consortium. «A reference panel of 64,976 haplotypes for genotype imputation». В: *Nature Genetics* (48 авг. 2016), с. 1279—1283. DOI: 10.1038/ng.3643. URL: <https://doi.org/10.1038/ng.3643>.
- [5] 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>.
- [6] H. W. Kuhn. «The Hungarian method for the assignment problem». В: *Naval Research Logistics Quarterly* 2.1-2 (1955), с. 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>.
- [7] H. Li и др. «The Sequence Alignment/Map format and SAMtools». В: *Bioinformatics* (авг. 2009), с. 2087—2079. DOI: 10.1093/bioinformatics/btp352. URL: <https://doi.org/10.1093/bioinformatics/btp352>.

-
- [8] «Linnarsson, S., Teichmann, S.A. Single-cell genomics: coming of age.» B: *Genome Biology* 97 (17 2016). DOI: 10.1186/s13059-016-0960-x. URL: <https://doi.org/10.1186/s13059-016-0960-x>.
- [9] 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>.
- [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), с. 414—421. DOI: 10.1038/s41592-020-0766-3. URL: <https://doi.org/10.1038/s41592-020-0766-3>.
- [11] «Method of the Year 2013». B: *Nature Methods* (11 янв. 2014). DOI: 10.1038/nmeth.2801. URL: <https://doi.org/10.1038/nmeth.2801>.
- [12] «Method of the Year 2019: Single-cell multimodal omics». B: *Nature Methods* (17 янв. 2020). DOI: 10.1038/s41592-019-0703-5. URL: <https://doi.org/10.1038/s41592-019-0703-5>.
- [13] K.P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012. ISBN: 978-0-262-01802-9.
- [14] 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>.
- [15] 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>.
- [16] 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>.