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

Содержание

1 Аннотация **1**

2 Материалы и методы **3**

2.1 Вероятностная модель числа прочтений 3

2.2 Алгоритмы предобработки данных 5

2.2.1 Извлечение данных из BAM-файлов 5

2.2.2 Статистическое фазирование гаплотипов 6

2.2.3 Подходы к сегментации генома 7

2.2.4 Исправление ошибок смены цепи 10

2.3 Использованные данные 17

2.4 Первоначальная версия XClone: только ASE-модуль 17

2.4.1 Plate notation 17

2.4.2 Семплирование по Гиббсу 17

2.4.3 Предложенная модель, её недостатки 17

2.4.4 Поиск наиболее вероятной перестановки меток . . . 22

2.5 Заключительная версия XClone: ASE- и RDR-модули . . . 24

2.5.1 Вариационный байесовский вывод 24

2.5.2 Структура ASE-модуля 24

2.5.3 Структура RDR-модуля 30

2.5.4 Известные недостатки и планы по их исправлению . 30

Список использованных источников **31**

2 Материалы и методы

2.1 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. 2.1: Вероятностая 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 2.1 (*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 2.2 (*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 2.3 (*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.

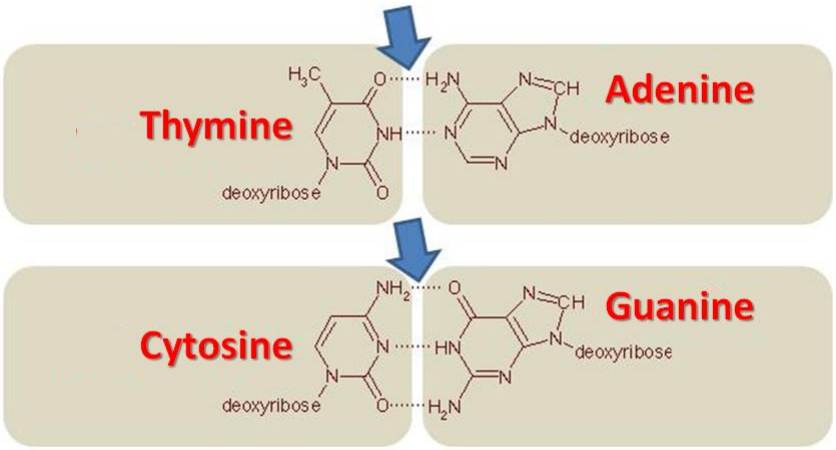


Рис. 2.2: Комлементарные пары: аденин-тимин и гуанин-цитозин

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 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.

2.2 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.

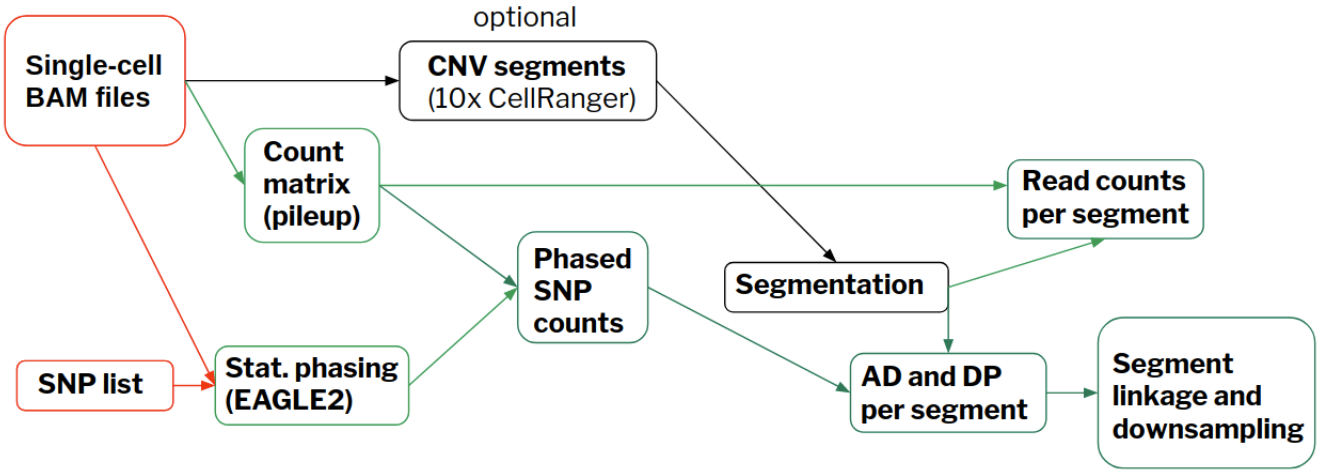


Рис. 2.3: Граф протокола предобработки данных для алгоритма XClone. Красным обозначены входные данные, чёрным — опциональные шаги, зелёным — реализованные стадии.

2.2.1 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 [5]. 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**[1].

2.2.2 Статистическое фазирование гаплотипов

Гаплотипирование — определение того, от какого родителя унаследован каждый алллель 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**[6][2]. Этот алгоритм основан

[1]https://github.com/single-cell-genetics/cellSNP

[2]https://data.broadinstitute.org/alkesgroup/Eagle/

на скрытых марковских моделях и использует 32,470 образца из базы данных **Haplotype Reference Consortium**[3].

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, помеченые 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.

2.2.3 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**, алгоритм[3] от 10X Genomics, поставщика оборудования для single-cell севенирования 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 килобаз, в пределах которых вероятность ошибки смены цепи невелика, а потому сигнал аллельного дисбаланса статистически достоверный.

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

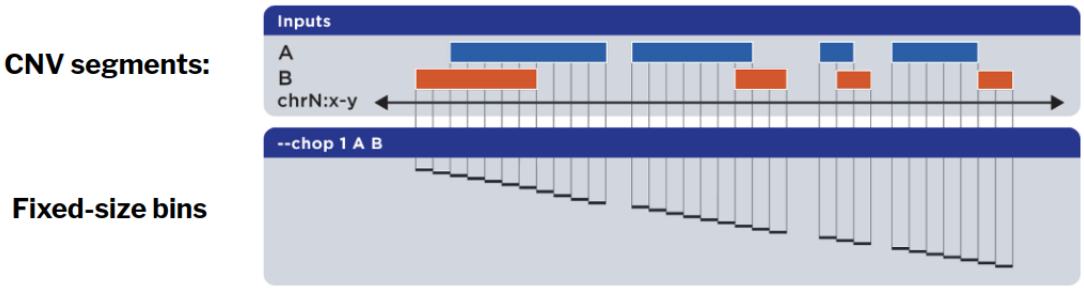


Рис. 2.4: Иллюстрация алгоритма сегментирования генома. Длина индивидуальных сегментов задаётся заранее и выбирается из диапазона 20-50кб. Каждые k подряд идущих фрагментов затем объединяются в блоки. Соответствующие подматрицы прочтений при этом суммируются с одновременной коррекцией ошибок смены цепи.

В силу того, что структурные вариации обычно охватывают участки генома размеров хотя бы в несколько мегабаз, перед началом предсказания уместно аггрегировать подряд идущие сегменты в блоки фиксированного размера (обычно 1-5 мегабаз), чтобы получить менее шумные BAF и RDR. Тем не менее, наивно аггрегировать 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 нуклеиотидной последовательности рида 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.

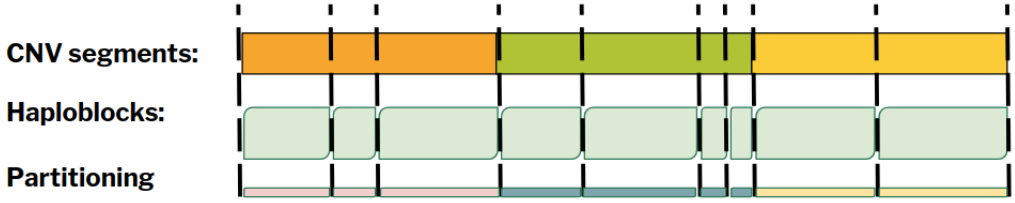


Рис. 2.5: Сегментирование генома в первоначальной версии XClone. Гаплотипические блоки, найденные по данным Oxford Nanopore DNA-seq, пересекаются с диапазонами структурных вариаций, найденных CellRanger DNA.

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.

2.2.4 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 [2], 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 **EM-алгоритмом**. После исправления ошибок, прочтения сегментов внутри блока суммируются, что даёт более стабильный сигнал. Эта идея была сформулирована в [8], но технические детали были осознанно исключены авторами CHISEL из препринта.

Прежде чем приступать к рассмотрению метода, сформулируем необходимые определения:

Определение 2.4 (*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.

Algorithm 1: EM-алгоритм в общем виде

Result: $\boldsymbol{\Theta}^*$, $p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\Theta}^*)$
$t = 0$;
$\boldsymbol{\Theta}^{(0)}$ инициализируется случайно;
while $Q(\boldsymbol{\Theta}^{(t+1)} \mid \boldsymbol{\Theta}^{(t+1)}) - Q(\boldsymbol{\Theta}^{(t)} \mid \boldsymbol{\Theta}^{(t)}) > \varepsilon$ **do**
  $\mathcal{L}(\boldsymbol{\Theta}^{(t)}; \boldsymbol{Z}, \boldsymbol{X}) := p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\Theta}^{(t)})$;
  $Q(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{(t)}) := \mathbb{E}_{\boldsymbol{Z}\mid\boldsymbol{X},\boldsymbol{\Theta}^{(t)}} \log \mathcal{L}(\boldsymbol{\Theta}; \boldsymbol{Z}, \boldsymbol{X})$ `// E-шаг`
  $\boldsymbol{\Theta}^{(t+1)} := \arg\max_{\boldsymbol{\Theta}} Q(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{(t)})$ `// M-шаг`
  $t = t + 1$
end
$\boldsymbol{\Theta}^* := \boldsymbol{\Theta}^{(t)}$

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

1. **E-шаг**, на котором устраняется явная зависимость от скрытых переменных посредством взятия матожидания логарифма совместной функции правдоподобия по условному распределению $\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\Theta}^{(t)}$;

2. **M-шаг**, на котором параметры нового апостериорного распределе-

ния $\boldsymbol{\Theta}^{(t+1)}$ выбираются таким образом, чтобы максимизировать $Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(t)})$ — функцию правдоподобия "в среднем".

С теоретическим обоснованием и формальным доказательством корректности EM-алгоритма можно ознакомиться в ([7], стр. 363-365). В контексте решаемой задачи $\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\Theta}$ имеют следующий смысл:

- $\boldsymbol{Z} = \{z_1, \ldots, z_N\}$ — независимые в совокупности индикаторы корректности гаплотипов сегментов

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

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

  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 аггрегации сегментов в блоки.

- Обозначим через M число клеток образца, тогда $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_M)$, $X_c := (\boldsymbol{a}_c, \boldsymbol{b}_c)$ — вектора прочтений для каждой из клеток, по компоненте на сегмент. $\boldsymbol{a}_c = (a_{c,1}, \ldots, a_{c,N})$ — число прочтений аллеля А (альтернативный аллель), $\boldsymbol{b}_c = (b_{c,1}, \ldots, b_{c,N})$ — аллеля Б (референсный аллель).

- $\forall c \in \overline{1, M} : \boldsymbol{r}_c := \boldsymbol{a}_c + \boldsymbol{b}_c$ — вектора прочтений обоих аллелей вместе.

- $\boldsymbol{\Theta} = (\theta_1, \ldots, \theta_M; p_1, \ldots, p_N)$, 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 2.1. Правила пересчёта параметров апостериорного распределения на M-шаге EM-алгоритма имеют вид:

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

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

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

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

Где

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

Тогда функция правдоподобия и её логарифм принимают вид

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

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 $\boldsymbol{Z}$ независимы в совокупности, шаги EM-алгоритма имеют вид:

E-шаг:

$$p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\Theta}^{(t)}) \propto p(\boldsymbol{X} \mid \boldsymbol{Z}, \boldsymbol{\Theta}^{(t)}) p(\boldsymbol{Z} \mid \boldsymbol{\Theta}^{(t)}) \implies$$

$$\implies \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\Theta}^{(t)}} l(\boldsymbol{\Theta}; \boldsymbol{Z}, \boldsymbol{X}) = \sum_{i=1}^{N} \mathbb{E}_{\boldsymbol{z}_i|\boldsymbol{X}_i,\boldsymbol{\Theta}^{(t)}} \log \mathcal{L}(\boldsymbol{\Theta}; \boldsymbol{z}_i, \boldsymbol{X}_i) =$$

$$= \sum_{i=1}^{N} \sum_{q_i=0}^{1} p(\boldsymbol{z}_i = q_i \mid \boldsymbol{X}_i, \boldsymbol{\Theta}^{(t)}) \left(\begin{array}{l} \sum\limits_{c=1}^{M} \left[\widehat{a}_{c,i}(q_i) \log \theta_c + \widehat{b}_{c,i}(q_i) \log(1-\theta_c) \right] + \\ + \log p(\boldsymbol{z}_i = q_i \mid \boldsymbol{\Theta}) \end{array} \right) =$$

$$= \sum_{i=1}^{N} \left[\begin{array}{l} \gamma_{i,1}^{(t)} \left(\sum\limits_{c=1}^{M} \left[a_{c,i} \log \theta_c + b_{c,i} \log(1-\theta_c) \right] + \log p_i \right) + \\ + \gamma_{i,0}^{(t)} \left(\sum\limits_{c=1}^{M} \left[b_{c,i} \log \theta_c + a_{c,i} \log(1-\theta_c) \right] + \log(1-p_i) \right) \end{array} \right] = Q(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{(t)})$$

M-шаг:

$$p_i^{(t+1)} = \arg\max_{p_i} Q(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{(t)}) \iff \frac{\gamma_{i,1}^{(t)}}{p_i^{(t+1)}} - \frac{\gamma_{i,0}^{(t)}}{1 - p_i^{(t+1)}} = 0 \iff p_i^{(t+1)} = \gamma_{i,1}^{(t)}$$

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

$$\iff \theta_c^{(t+1)} = \frac{\sum_{i=1}^{N} \gamma_{i,1}^{(t)} a_{c,i} + \gamma_{i,0}^{(t)} b_{c,i}}{\sum_{i=1}^{N} a_{c,i} + b_{c,i}}$$

Где необходимое условие локального экстремума является также достаточным в силу выпуклости функции $Q(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{(t)})$ ([7], стр. 363-364). $\square$

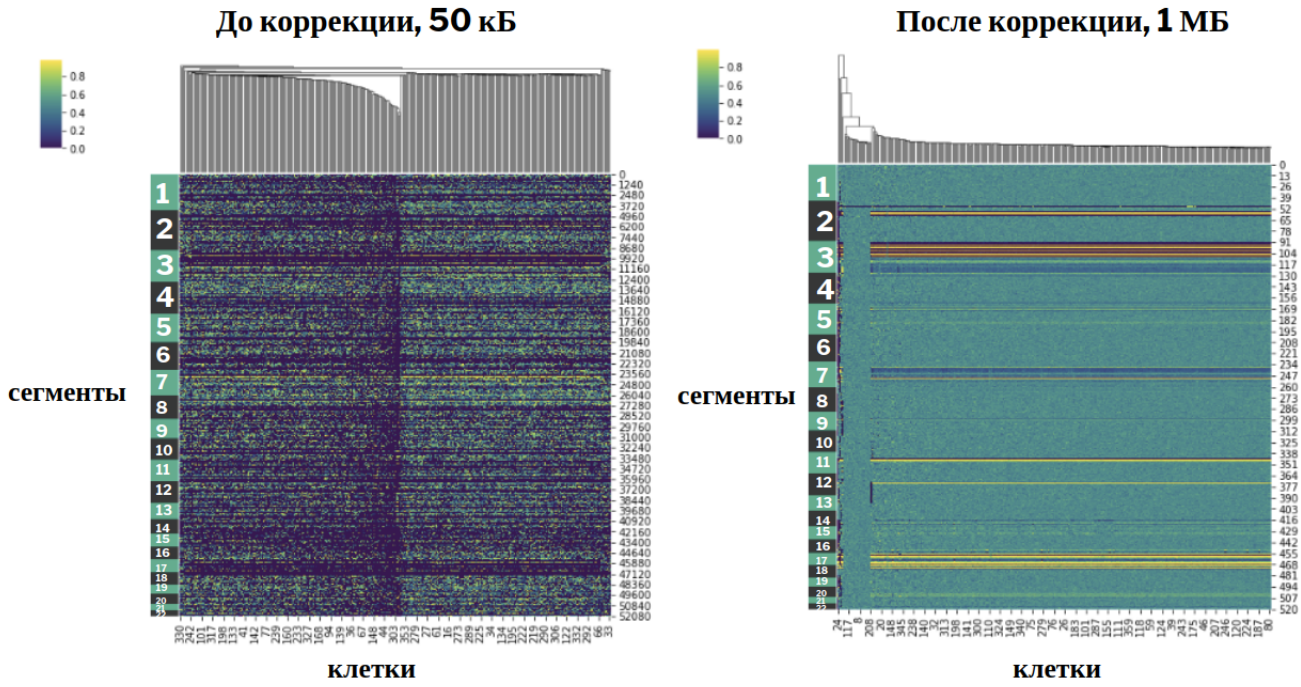


Рис. 2.6: Коррекция ошибок смены цепи на примере ДНК-образца STP-Nuclei. На рисунке изображены тепловые карты долей аллеля из «материнского» гаплотипа, числа слева обозначают номер хромосомы. Картина аллельного дисбаланса до коррекции практически не прослеживается, после — становится очевидна.

После того, как значения $z_1, \ldots, z_N$ 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 эквивалетной, 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.

2.3 Использованные данные

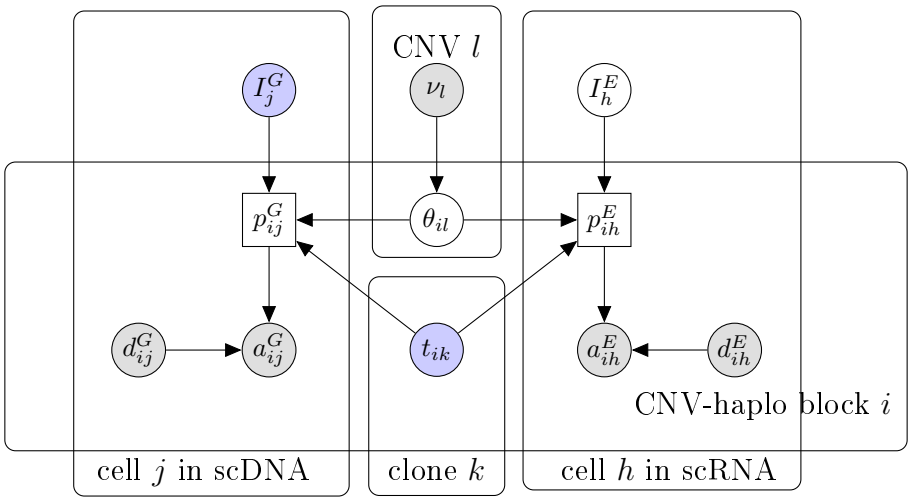
2.4 Первоначальная версия XClone: только ASE-модуль

2.4.1 Plate notation

2.4.2 Семплирование по Гиббсу

2.4.3 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:

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

  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.
- $\mathbf{f} = (f_1, \ldots, f_K)$ — 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:**

- $\mathbf{\Theta}$ — allelic rates of variants located within blocks of fixed CNV status. $\mathbf{\Theta_{i,t}}$ 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 $\mathbf{\Theta}$) denotes the information for both samples.
- $\mathbf{H^G}, \mathbf{H^E}$ — CNV status of the blocks in accordance with the current label assignment:

$$\mathbf{H^G_{i,j}} := \mathbf{T_{i,I^G_j}}, \quad \mathbf{H^E_{i,j}} := \mathbf{T_{i,I^E_j}}$$

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

$$\mathbf{X^G_{i,j}} := \mathbf{\Theta_{i,H^G_{i,j}}}, \quad \mathbf{X^E_{i,j}} := \mathbf{\Theta_{i,H^E_{i,j}}}$$

Generative model formulation

- **Cell-to-clone assignment posterior:**

$$\mathrm{P}(\mathbf{I^E_j} = k_0 \mid \mathbf{A_j}, \mathbf{D_j}, \mathbf{f}, \mathbf{\Theta}) = \frac{\mathrm{P}(\mathbf{A_j} \mid \mathbf{D_j}, \mathbf{I^E_j} = k_0, \mathbf{\Theta})\mathrm{P}(\mathbf{I^E_j} = k_0 \mid \mathbf{f})}{\sum_{k=1}^{K} \mathrm{P}(\mathbf{A_j} \mid \mathbf{D_j}, \mathbf{I^E_j} = k, \mathbf{\Theta})\mathrm{P}(\mathbf{I^E_j} = k \mid \mathbf{f})} \tag{2}$$

- **ASE model:**

$$\begin{aligned} \mathrm{P}(\mathbf{A^G_{i,j}} \mid \mathbf{D^G_{i,j}}, \mathbf{\Theta}) &= \mathrm{Binom}(\mathbf{A^G_{i,j}} \mid \mathbf{D^G_{i,j}}, \mathbf{X^G_{i,j}}) \\ \mathrm{P}(\mathbf{A^E_{i,j}} \mid \mathbf{D^E_{i,j}}, \mathbf{\Theta}) &= \mathrm{Binom}(\mathbf{A^E_{i,j}} \mid \mathbf{D^E_{i,j}}, \mathbf{X^E_{i,j}}) \end{aligned} \tag{3}$$

- **ASE likelihood (both terms factorize over variants):**

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

- **Allelic rate likelihood:**

$$\mathcal{L}(\mathbf{\Theta}) = \left(\prod_{j=1}^{M^G} \sum_{k^G=1}^{K} \mathrm{P}(\mathbf{A_j^G} \mid \mathbf{D_j^G}, \mathbf{I_j^G} = k^G, \mathbf{\Theta}) \right) \times \\ \times \left(\prod_{j=1}^{M^E} \sum_{k^E=1}^{K} \mathrm{P}(\mathbf{A_j} \mid \mathbf{D_j^E}, \mathbf{I_j^E} = k^E, \mathbf{\Theta}) \cdot \mathrm{P}(\mathbf{I_j^E} = k^E \mid \mathbf{f}) \right) \tag{5}$$

To view the clonal assignment in a Bayesian way, we introduce informative prior ν for $\mathbf{\Theta}$.

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

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

Parameters α_t, β_t are selected in such a way that the mode of $\mathrm{Beta}(\alpha_t, \beta_t)$ equals to $1/t$ [4].

Selecting a prior for $\mathbf{\Theta}$ 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 $k \in \{0, \ldots, t\}$, 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 $\mathrm{Beta}(\alpha, \beta)$, given by $(\alpha-1)/(\alpha+\beta-2)$, equals to k_0/t. That means, we must solve the following problem:

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

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

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

works[5]. 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:

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

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

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

From here we derive update rules for individual allelic rates:

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

[5]Nevertheless, 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

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

2.4.4 Поиск наиболее вероятной перестановки меток

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

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

Тем не менее, модель предсказывала метки с точностью до неизвестной перестановки. Если предсказание точное, т.е. что $\boldsymbol{P}$ с точностью до перестановки совпадает с истинной, которая задаётся бинарной матрицей $\boldsymbol{Q}$, то сама перестановка определяется легко: биекция между столбцами $\boldsymbol{P}$ и $\boldsymbol{Q}$ 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 $\boldsymbol{P}$ и $\boldsymbol{Q}$, 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 $M = 1.3 \times 10^{66}$, и типичное

[6]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 $\boldsymbol{P}$ сопоставляются вершины левой доли, $v_1, \ldots, v_K$, столбцам $\boldsymbol{Q}$ — вершины правой доли, $u_1, \ldots, u_K$, 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 кобминаторной оптимизации. Для её решения применяется так называем "венгерский алгоритм опубликованный в 1955 году американским математиком Гарольдом Куном[4].

Рис. 2.7: Иллюстрация сведения задачи восстановления наиболее вероятной перестановки меток классов к поиску совершенного паросочетания минимального веса в двудольном графе. Веса — расстояния между столбцами матриц $\boldsymbol{P}$ и $\boldsymbol{Q}$, матриц предсказанных и истинных вероятностей клональных меток.

2.5 Заключительная версия XClone: ASE- и RDR-модули

2.5.1 Вариационный байесовский вывод

2.5.2 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,

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

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

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

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,

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

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

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 $\boldsymbol{f}_j = \{f_{1,j},...,f_{N,j}\}$, 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

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

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

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

Here, we introduce $\texttt{Dirichlet}(\boldsymbol{\omega})$ as the prior distribution for $\boldsymbol{m}$, and Gaussian process $\texttt{GP}(\boldsymbol{\mu}, \Sigma)$ 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 $K(\boldsymbol{c}_t, \boldsymbol{c}_{t\prime}) = l_1 \exp\{-l_2[(c_{t,1} - c_{t\prime,1})^2 + (c_{t,2} - c_{t\prime,2})^2]\}$, as follows,

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

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,

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

where the prior distribution is independent for all five variables $\Omega = \{Z, Y, \boldsymbol{\theta}, \boldsymbol{m}, \boldsymbol{\gamma}\}$, 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 five000 unknown variables $p(\Omega|A,D)$. Then the inference problem becomes an optimisation problem for minimising the evidence lower bounder (ELBO), as follows

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

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 achive 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

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

The allelic fraction module could further analytically written as follows,

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

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,

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

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

function, as follows.

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

where $\theta \sim \texttt{Beta}(\alpha,\beta)$ and $1-\theta \sim \texttt{Beta}(\beta,\alpha)$. 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)

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

2.5.3 Структура RDR-модуля

2.5.4 Известные недостатки и планы по их исправлению

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] 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.

[3] 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.

[4] 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.

[5] 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.

[6] P.R. Loh и др. «Reference-based phasing using the Haplotype Reference Consortium panel». В: *Nature Genetics* (48 окт. 2016), с. 1443—1448. DOI: 10.1038/ng.3679. URL: https://doi.org/10.1038/ng.3679.

[7] K.P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012. ISBN: 978-0-262-01802-9.

[8] S. Zaccaria и B.J. Raphael. «Characterizing the allele- and haplotype-specific copy number landscape of cancer genomes at single-cell resolution with CHISEL». В: *bioRxiv* (нояб. 2019). DOI: 10.1101/837195. URL: https://doi.org/10.1101/837195.