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

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.

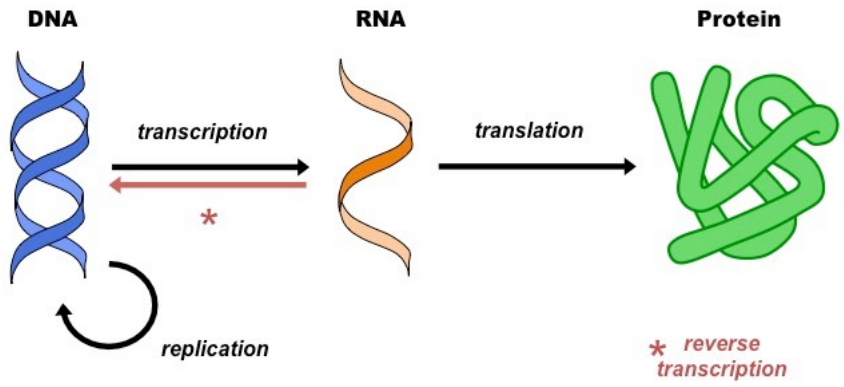

Рис. 2.1: Центральная догма молекулярной биологии

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 экзома, которая преобразуется в белки в момент наблюдения, и зависит от типа клетки, стадии клеточного цикла, условий внешней среды и т.д.

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

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

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

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

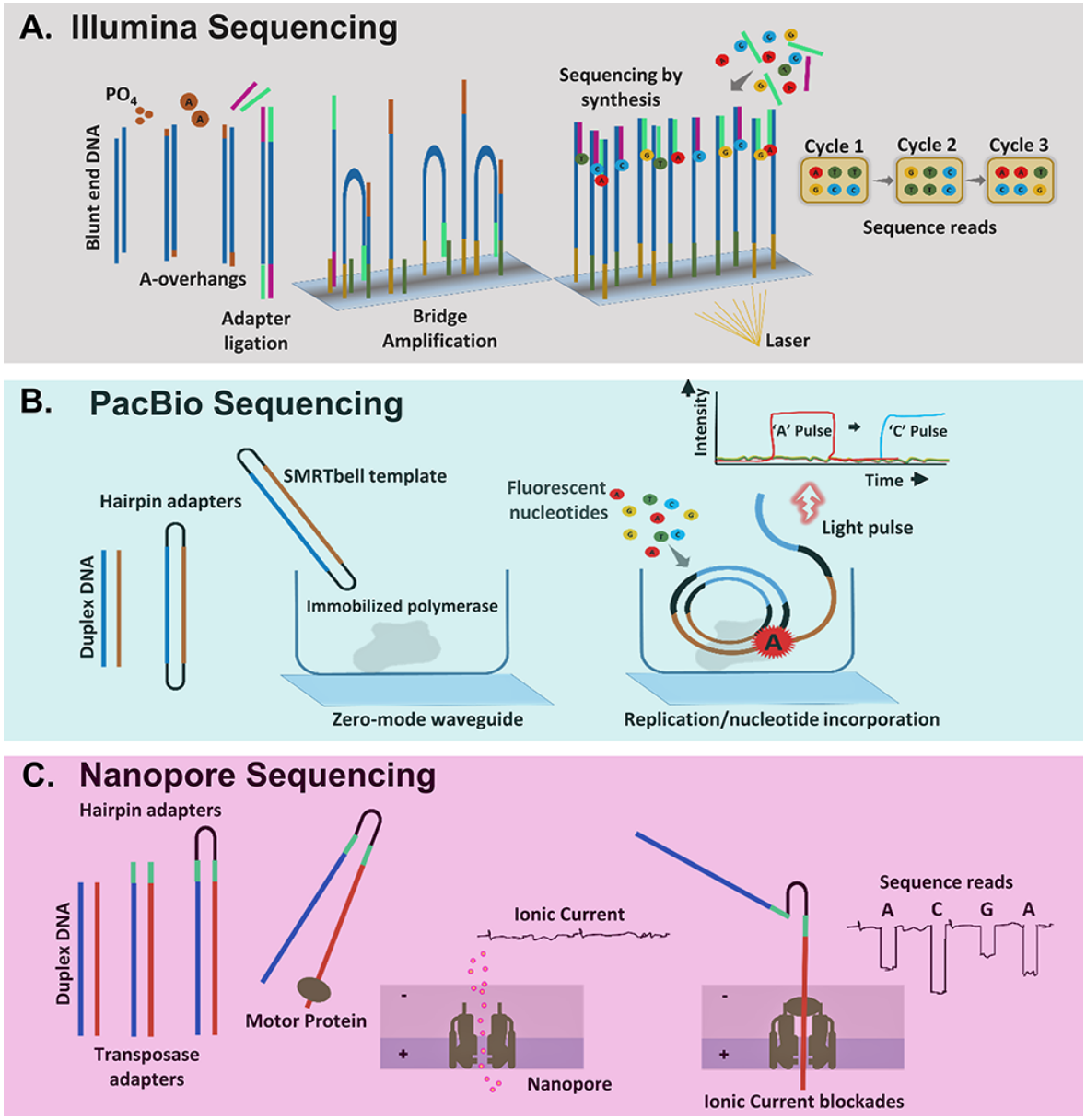

Рис. 2.2: Основные технологии секвенирования [1].

В данной работе использованы протоколы Illumina и Oxford Nanopore. Сквенаторы Illumina вначале дробят входную последовательность нуклеотидов на небольшие фрагменты, которые затем **амплифицируются** (многократно воспроизводятся) мащиной в ходе **ПЦР** — полимеразной цепной реакции. Так называемое **секвенирование нанопорами** — метод, не требующий применения ПЦР. В контексте данной работы важно, что длина ридов, полученных по этой технологии, заметно больше, что компенсируется меньшей пропускной способностью и более высокой ценой.

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 траслитерация ***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;

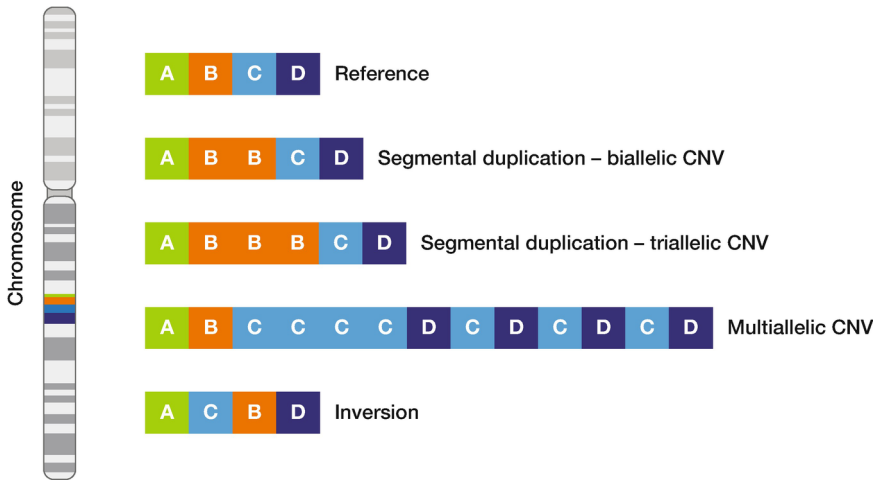

Рис. 2.3: Основные типы структурных вариаций

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.

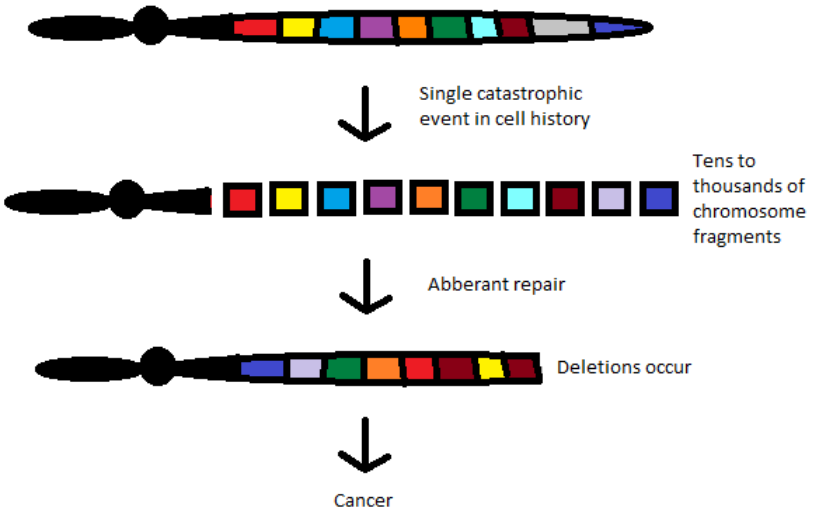

Рис. 2.4: Хромотрипсис

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. Клональные события позволяют задать псевдовремя на стадиях онкогенеза. Именно **псевдо**время, 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 (*Сегментация генома*)**.** Разбиение генома на непересекающиеся подотрезки.

Определение 2.20 (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.

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. 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.

[5] 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.

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

[7] 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.