ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ОБРАЗОВАНИЯ

“САНКТ-ПЕТЕРБУРГСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ”

Факультет *Информационных технологий и программирования*

Направление (специальность) *Прикладная математика и информатика*

Квалификация (степень) *Магистр прикладной математики и информатики*

Кафедра *Компьютерных технологий* Группа *6538*

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

на тему

Построение модели для определения вероятностей ошибок в данных секвенирования в зависимости от контекста

Автор магистерской диссертации Ливарский Р. Ю. (подпись)
(Фамилия, И., О.)

Научный руководитель Фильченков А. А. (подпись)
(Фамилия, И., О.)

Руководитель магистерской программы Васильев В. Н. (подпись)
(Фамилия, И., О.)

К з а щ и т е д о п у с т и т ь
Зав. кафедрой Васильев В. Н. (подпись)
(Фамилия, И., О.)

“____”_____20____г.

Санкт-Петербург, 2015 г.

Содержание

Введение

Процесс нахождения аминокислотной или нуклеотидной последовательности биополимеров (нуклеиновых кислот и белков) известен как секвенирование. В результате в символьном представлении получают формальное описание их первичной структуры в виде последовательности нуклеотидов. Знание генома организмов требуется для многих задач современной биологии и медицины. Генной инженерии, исследованию наследственных заболеваний и многим другим задачам нужен быстрый, дешевый и качественный способ безошибочного определения измененных участков ДНК.

Технологии секвенирования нового поколения позволяют прочитать единовременно несколько участков генома, что является главным отличием от предыдущих технологий. Секвенаторы на основе новых технологий стали значительно эффективнее и гораздо дешевле своих предшественников. В настоящее время производительность некоторых секвенаторов превосходит сотни миллиардов нуклеотидов за рабочий цикл, что дает возможность таким приборам лишь за несколько дней получить геном человека.

Наличие ошибок в определении правильного нуклеотида в чтении ДНК является общей проблемой для всех технологий секвенирования. К сожалению, секвенаторы на основе последних технологий, как показывает практика, более подвержены ошибкам, чем предыдущие. Последствия таких ошибок зависят от конкретного применения, начиная от незначительных информационных неприятностей, заканчивая крупными проблемами, влияющими на биологические выводы. Поэтому требуются надежные показатели качества данных.

Современные секвенаторы вместе с прочитанным нуклеотидом предоставляют показатель качества, отображающий вероятность его ошибочного определения, говоря другими словами, уверенность секвенатора в

своем выборе. Такой показатель основан на физических характеристиках процесса. К сожалению, эти показатели не всегда описывают настоящее состояние дел. Такие показатели особенно важны для поиска мутаций. Тем не менее, в конкретных случаях, может быть получено большое количество неверно распознанных мутаций. Нужно уметь различать позиции с настоящими мутациями и позиции с ошибками секвенирования.

В настоящее время различают несколько видов ошибок секвенирования. Одним из них является вид систематических ошибок, возникающих вследствие наличия определенной последовательности нуклеотидов перед ними. Такие ошибки, заисимые от контекста, имеют свойство скапливаться в некоторой геномной позиции ввиду того что, большинство чтений, соответствующие данной позиции, будут содержать одинаковые «плохие» фрагменты. Такое ошибочное скопление пагубно влияет на алгоритмы поиска мутаций. По факту имеем большое содержание в реальности несуществующего нуклеотида с высоким качеством, предоставленным секвенатором, в определенной геномной позиции, которое логично объяснить только мутацией.

Идея метода, разработанного в данной работе, заключается в построении неявной статистической модели распределения систематических ошибок секвенирования в засимости от контекста. Она позволяет оценить показатель качества нуклеотидов, возникщих благодаря ошибкам, показать насколько нельзя им доверять в конкретной геномной позиции. Метод повышает эффективность алгоритмов поиска мутаций, которые в большинстве своем основаны на анализе показателей качества, благодаря замене заявленных качеств на вычисленные, отражающие всю суть возникшей ситуации с ними.

Глава 1. Обзор предметной области

В данной главе будет дан обзор основных элементов предметной области. Будет сформулирована задача определения вероятности ошибки на основе её контекста, будут описаны проблемы распознавания ошибок секвенирования от мутаций и существующие способы для их различия.

1.1. Ген и геном

Ген — структурная и функциональная единица наследственности, участок молекулы ДНК, где закодирована информация о синтезе одной полипептидной цепи с определенной аминокислотной последовательностью. Гены на молекулярном уровне осуществляют контроль над всеми клеточными процессами и обеспечивают белковый синтез. Ген и признак организма не связаны простым соотношением. Требуется много генов для развития одного сложного признака (обоняние, например). В то же время на формирование нескольких признаков может влиять всего лишь один ген. Различные варианты одного признака (например, цвет волос) обуславливаются несколькими формами существования (аллелями) соотвествующего гена.

Геном — совокупность наследственного материала (генов), сосредоточенного в гаплоидном наборе хромосом некоторого живого организма. Развитие и функционирование всех известных живых существ определяется их геномом. Кроме того, геном определяет то, каким может получиться потомство того или иного существа. Различия между разными видами, как и между разными особями одного вида, возникают в результате того, что у них различные гены.

1.2. Структура ДНК

Нуклеотид — сложное вещество, состоящее из азотистого основания, сахара (дезоксирибозы) и фосфатной группы. Являются основой для построения ДНК: цепи ДНК состоят из последовательности нуклеотидов. В ДНК бывает четыре вида оснований: аденин, гуанин, тимин, цитозин.

ДНК (Дезоксирибонуклеиновая кислота) — молекула, осуществляющая хранение генетической информации [1]. Структура молекулы представляется в виде двойной спирали. Нуклеотиды в каждой из двух цепей соединяются фосфодиэфирными связями. Сами эти цепи держатся вместе за счет водородных связей, возникающих между азотистыми основаниями, которые находятся друг напротив друга в цепях. Водородные связи возникают согласно принципу комплементарности: гуанин связывается только с цитозином, аденин — с тимином . Из этого следует что, зная последовательность нуклеотидов в одной цепи, можно абсолютно точно установить последовательность в другой. Эти последовательности обратно комплементарны друг другу. В каждой из них есть свое направление, и эти направления противоположны друг другу.

1.3. Генетический код

Генетический код — способ записи структуры белков в молекулах нуклеиновых кислот в виде последовательности нуклеотидов. Каждый белок состоит из последовательности аминокислот [2]. Существует 20 основных аминокислот, которые могут входить в состав белков. Из четырех возможных нуклеотидов можно составить 64 различных «слов» длины три. Такие трёхбуквенные слова называется кодоны или триплеты. 61 кодон из 64 возможных кодирует определенные аминокислоты. Некоторые аминокислоты кодируются несколькими кодонами, так как число кодирующих кодонов больше количества основных аминокислот. Это свойство называется вырожденностью кода. Другое свойство, при котором один кодон вхо-

дит в состав только одной аминокислоты, называется специфичностью или однозначностью кода. Вдобавок, код не перекрывается, считывается последовательно в одной направлении. И, наконец, он единый для большинства живых существ. Это самое удивительное свойство — универсальность. Расшифровка кода была осуществлена в 1961-1965 гг. и является одним из ярких достижений биологии.

1.4. Репликация

Репликация — процесс синтеза дочерней ДНК на матрице родительской ДНК [3]. Во время клеточного деления связи между цепями ДНК разрушаются, и нити разделяются. Потом на каждой из них строится комплементарная дочерняя цепь. В результате в обеих клетках получается по одной полной копии молекулы ДНК. Идентичность требуется, чтобы полностью передать всю информацию, которая заключена в геноме клетки. На самом деле ДНК удваивается не последовательно, а отдельными участками — репликонами. Они синтезируется независимо. Благодаря этому, репликация ДНК выполняется параллельно в нескольких местах, обеспечивая высокую скорость копирования.

1.5. Размножение и передача генов

Сложные организмы размножаются половым путем. Обычно клетки таких организмов содержат диплоидный набор хромосом: по одному набору хромосом от каждого из родителей. При половом размножении такие организмы производят гаметы — половые клетки, обладающие гаплоидным набором хромосом. Такие клетки создаются в результате мейоза. Мейоз — вид клеточного деления, при котором из диплоидных клеток образуются гаплоидные гаметы. В профазе мейоза происходит кроссинговер — перекомбинация, обмен частями гомологичных хромосом. Поэтому гаплоидный набор половой клетки является комбинацией геномов, которые получены

от каждого из родителей особи. В зиготе происходит объединение таких клеток, благодаря чему число хромосом вновь становится диплоидным. В итоге каждая такая клетка, полученная в результате слияния двух половых клеток родителей, содержит по одному «смешанному» геному от каждого из родителей. Таким образом поддерживается генетическое разнообразие.

1.6. Мутация

Мутация — наследуемое преобразование генотипа, которое может вызвать изменение свойств и признаков живого организма. Мутации могут случаться естественно, а могут возникать под воздействием внешней среды, индуцироваться различными факторами — мутагенами.

Можно разделить мутации по следующим признакам:

- по масштабу воздействия (ген, хромосома, геном);
- по месту происхождения (соматические или половые клетки);
- по проявлению (доминантные или рецессивные);
- по влиянию (полезные или вредные (летальные);
- по причине возникновения (индуцируемые или спонтанные).

Точечные (генные) мутация связана с изменением участка ДНК, чаще всего одного гена. Причинами могут быть удаление, вставка или замена пары нуклеотидов. Замены классифицируют на:

- транзиции — пурин замещается на пурин (например. A на G), либо пиримидин на пиримидин (T на C);
- трансверсии — пурин (A или G) замещается на пиримидин (T или C), либо наоборот.

Транзиции происходят в два раза чаще, чем трансверсии, благодаря схожести по химическим свойствам заменяемых нуклеотидов. Такой вид замен не сильно вляет на структуру ДНК и на функционирование и работоспособность всего организма в конечном итоге.

Хромосомная мутация (аберрация) заключается в нарушении структуры хромосом в связи с удвоением (дупликация), потерей (делеция), перемещением (транслокацией) их отдельных частей. Могут возникать как в одной хромосоме, так и между гомологичными и негомологичными.

Геномная мутация состоит в изменении числа хромосом. В результате неверного клеточного деления в хромосомном наборе может присутствовать лишняя хромосома, или, наоборот, отсутствовать.

В большинстве случаев мутации, ухудшающие работу клетки, приводят к её уничтожению. В стабильной среде обитания организмы содержат почти оптимальный генотип, поэтому большинство мутаций вызывают нарушение функций организма и приводят к смерти. Поэтому такие вредные мутации не распространяются дальше по поколениям. В редких случаях мутация может оказаться положительной и стать причиной появления полезных признаков. Такие особи получают преимущество и оказываются более приспособленными к условиям внешней среды и, как следствие, более живучими. Таким образом, мутации являются частью естественного отбора.

Для защиты от влияния вредной внешней среды и сопутствующих мутаций организм и клетки, в частности, вырабатывают механизмы противоборства им. Благодаря полученному иммунитету, такие клетки могут выживать под сильным воздействием мутагенов (например, воздействие ультрафиолетового излучения или высокой температуры).

1.7. Секвенаторы последнего поколения

Благодаря разнообразным улучшениям, многочисленным модификациям и увеличению производительности вычислительной техники были разработаны технологии секвенирования, которые основываются на единовременном «прочтении» (распараллеливании) нескольких участков генома. В связи с этим секвенаторы на основе таких технологий стали намного эффективнее по скорости и стоимости [4], чем предыдущие. Производитель-

ность таких приборов настолько высока, что, например, лишь за несколько дней позволяет получить геном человека.

На рынке существует достаточно большое количество секвенаторов, основанных на различных принципах определения нуклеотидов. Поэтому все они различаются в скорости и стоимости секвенирования, стоимости самого прибора, сопутствующих типах ошибок и их количестве. В статьях [5, 6] подробно описаны многие характеристики. Часть характеристик приведены в таблице 1.1.

Платформа	Illumina MiSeq	Illumina GAIIx	Illumina HiSeq 2000	Ion Torrent PGM	PacBio RS	454 GS FLX
Принцип	SBS (sequncing-by-synthesis)	SBS	SBS	Ионный полупроводник	Real-time	Пиросеквенирование
Стоимость прибора	$ 128k	$ 256k	$ 654k	$ 80k	$ 695k	$ 500k
Время работы за цикл	27 ч.	10 д.	11 д.	2 ч.	2 ч.	23 ч.
Длина чтений	до 150 нукл.	до 150 нукл.	до 150 нукл.	∼200 нукл.	∼2500 нукл.	700 нукл.
Количество нуклиотидов за цикл	1.5-2 Гн	30 Гн	600 Гн	20-1000 Мн	100 Мн	700 Мн
Стоимость 1Г нукл.	$ 502	$ 148	$ 41	$ 1000	$ 2000	$ 7000
Доля ошибок	0.8 %	0.76%	0.26%	1.71 %	12.86 %	0.49 %
Преимущ. тип ошибок	Замена	Замена	Замена	Вставка/ удаление	Вставка/ удаление	Вставка/ удаление

Таблица 1.1: Сравнительная характеристика известных секвенаторов.

1.8. Ошибки секвенирования

Во всех платформах последнего поколения для считывания нуклеотидных последовательностей (в сравнении с предыдущими аналогами) за более высокую скорость и низкую стоимость обработки данных приходится расплачиваться более частыми ошибками секвенирования [7, 8]. Но кроме обычных случайных несовпадений существуют еще и систематические ис-

точники ошибок. Поскольку любая ошибка в определении нуклеотида может быть расценена как мутация, исправление как можно большего количества ошибок любого типа является жизненно важным. Нужны хорошие статистические методы, которые бы различали высокую вероятность ошибки и гетерозиготный генотип в позициях с низким покрытием в чтениях. В некоторых геномных позициях во многих чтениях случаются скопления несовпадений с референсным нуклеотидом. Известно, что ошибки происходят ближе к концам чтений [8—10] и зависят от окружающих её мотивов последовательности. Например, ошибки склонны к появлению перед «GG» или после некоторого количества «GGC» [10]. Было обнаружено [11], что есть геномные позиции, в которых ошибки случаются намного чаще, чем это может быть объяснено описанными эффектами. Такие ошибки имеют систематический характер появления. Было замечено, что в местах скопления систематических ошибок, как правило, ошибки появляются только с одной стороны секвенирования (прямая или обратная). Эта тенденция была замечена в [12], где направление возникновения ошибок использовалось для разделения мнимых мутаций и гетерозиготных позиций (настоящих мутаций). Возможным объяснением этому может служить, что в процессе секвенирования некоторых мотивов ДНК (различных в противоположных направлениях) возрастает вероятность ошибки в правильном определении нуклеотида. Это согласуется с известным наложением спектра поглощения «G» и «T» каналов, определяемых одним лазером в *Illumina* [13]. В дальнейшем, систематические, зависимые от последовательности ошибки будут называться контекстными ошибками, а мотивы последовательности, которые индуцируют ошибки — контекстами. Для отличия контекстных ошибок от настоящих мутаций будет использоваться дисбаланс количества ошибок в разных направлениях: поскольку ошибки вызываются предшествующими мотивами, а не последующими, ошибка должна присутствовать с одной стороны и отсутствовать с другой. Благодаря предыдущим работам [10, 11], обнаружение и отбор позиций с таким дисбалансом стали обычным этапом

в обработке данных. *The Genome Analysis Toolkit (GATK)* [14, 15], для примера, вычисляет ко всем предполагаемым мутациям *p-value*, полученное из теста для определения независимости несовпадений от направления чтений (точный тест Фишера). Такой метод, однако, требует достаточного покрытия чтениями [16]: если покрытие слабое, статистическая сила для такого определения слаба. В работе [16] было предложено объединять позиции и ошибки в них на основе их геномного контекста. При проверке на существенный дисбаланс для вычисления *p-value* используется собранная статистика по всем позициям с таким контекстом, а не только ошибки в текущей единственной позиции. Такой «общий взгляд» на все позиции компенсирует общее низкое покрытие и держит статистическую силу на высоком уровне. К сожалению, такой метод позволяет только выделять наиболее склонные к ошибкам контексты относительно других, но не определять вероятность ошибки после них. На практике было бы полезнее иметь некую количественную характеристику (на сколько после каждого контекста всё плохо).

1.9. Форматы хранения генетической информации

Данные секвенатора нужно сохранять в удобном, простом, логичном формате. Вместе с прочитанными нуклеотидами секвенатор может выдавать дополнительные характеристики к ним. Например, показатели качества прочтения нуклеотидов. Всё это надо правильно сохранять для удобства обработки.

1.9.1. Формат FASTA

FASTA — наиболее распространенный формат представления последовательностей нуклеотидов [17]. Формат FASTA представляет собой обычным текстовый файл. Опишем кратко формат:

- Первая строка начинается с символа «>» или «;». Она содержит

идентификаторы библиотеки чтений и самого чтения;

- Следующие строки, которые начинаются опять же с «;» или «>», содержат комментарии.
- Далее, в оставшихся строках идет сама последовательность нуклеотидов в виде латинских букв из алфавита $\{A, C, G, T, U, N\}$, которые соответствует аденину, цитозину, гуанину, тимину и урацилу. Символ N обозначает неизвестный нуклеотид.

1.9.2. Формат FASTQ

Формат FASTA не предусматривал хранение дополнительной информации к каждому чтению. В FASTQ это поправили. Этот формат в выходных данных секвенатора, так как вместе с самой последовательностью предоставляет вероятность ошибки в выборе каждого элемента последовательности [18]. Опишем формат:

- Для хранения каждого чтения отведено по четыре строки;
- Первая строка начинается с символа «@». Описывает то же, что и в FASTA;
- Вторая строка описывает нуклеотидную последовательность в том же виде как и в FASTA;
- Третья строка служит разделителем и содержит «+»;
- Четвертая строка содержит для каждого нуклеотида второй строки некоторый символ, соответствующий его показателю качества.

Для расчета показателя качества *Phred* используется следующая формула [19]:

$$Q = -10 \cdot log_{10}p,$$

где p — вероятность ошибки в определении нуклеотида в некоторой позиции.

Теперь чтобы получить $ASCII$-код символа, нужно округлить Q до ближайщего целого и прибавить константу. Константа чаще всего равна 33 или 64 в зависимости от секвенатора.

1.10. Формулировка задачи

По данному набору чтений в формате FASTQ требуется определить фрагменты, которые склонны к образованию ошибок секвенирования после них. Также требуется вычислить соответствующую вероятность возникновения такой ошибки. Полученные показатели качества должны повышать эффективность обнаружения мутаций в настоящем геноме, представленным этими чтениями.

1.11. Проблемы задачи

В идеальном случае при полном совпадении референсного генома и настоящего генома, представленного в виде набора его чтений, все несовпадения, встречаемые в чтениях, являются ошибками секвенирования. Но на практике существуют мутации, которые обуславливают их различие. Такие мутации происходят в среднем в одной из тысячи позиций. В этом случае не каждое несовпадение является ошибкой, и, наоборот, не каждое совпадение верным. Суммарная погрешность будет зависеть от количества чтений, покрывающих каждую мутацию. Если мутации не учитывать, заметно снизится качество нуклеотидов в таких геномных позициях, что отрицательно повлияет на их обнаружение. Проблема состоит в разделении верных нуклеотидов и ошибочных. Но не всегда наличие большого числа разных нуклеотидов свидетельствует о систематической ошибке, в каком-то из них. Задача усложняется в диплоидном случае, при котором в одной геномной позиции могут быть одновременно два различных верных нуклеотида, так называемый гетерозиготный генотип.

1.12. Выводы по главе 1

Даны общие определения предметной области. Сформулирована основная задача настоящей работы. Описаны проблемы отличия ошибки секвенирования от мутации и способы её решения.

Глава 2. Описание предлагаемого метода

Однонуклеотидный полиформизм (ОНП) является наиболее популярным видом мутации [12, 20]. В диплоидном организме различаются два типа ОНП. Гомозиготный ОНП отличается от референсного генома в обеих аллелях, в то время как гетерезиготный ОНП только в одной аллеле. Зависимая от контекста ошибка может быть принята за гетерозиготный ОНП, когда некоторая часть чтений отличается от референса. Для того, чтобы уменьшить количество таких случаев, был разработан предлагаемый метод.

2.1. Основная идея метода

Основная идея заключается в следующем. Так как молекула ДНК состоит из двух цепочек, секвенатор может получать чтения с любой из них. Поэтому чтение, которое получено с прямого генома, имеет прямое направление, а которое получено с обратного генома, соответственно, имеет обратное направление. В итоге получается, что некоторую геномную позицию покрывают чтения, идущие в противоположных направлениях. Поэтому одной и той же геномной позиции в чтениях, идущих в противоположных направлениях, предшествуют разные контексты. Тогда в ситуации, когда некоторый контекст склонен к ошибке, в чтениях с противоположной стороны находится другой контекст, и ошибок не возникает. Вероятность наличия плохих контекстов в одной геномной позиции с двух сторон крайне мала, и такие случаи опускаются. Ошибка после контекста не зависит от геномной позиции и возникает с некоторой вероятностью всегда во всех позициях, где этот контекст появляется. Поэтому, было бы правильным собрать статистику для всех контекстов по всем их позициям и оценить насколько больше ошибок возникает с одной стороны, чем с другой. Стоит заметить, что при наличии в позиции мутации ошибки возникают с обе-

их сторон в примерно одинаковом количестве, и они не будут влиять на общую картину.

Предлагаемый метод можно разбить на следующие части:

1. построение таблицы сопряжённости для каждого геномного мотива;
2. отбор наиболее склонных к ошибкам мотивов на основе построенных таблиц;
3. выбор наиболее однозначных (в плане настоящего генотипа) позиций в геноме;
4. подсчет статистики для мотивов чтений по отобранным позициям.

2.2. Построение таблицы сопряжённости

Для данного мотива m определяются все вхождения его в референсный и обратно комплементарный геном. Будем называть геномный интервал, где мотив совпадает с прямым референсом, F-интервалом и последнюю позицию F-интервала F-позицией. Также интервал, где мотив совпадает с обратно комплементарным референсом, R-интервалом и его первую позицию R-позицией. Геномная позиция i всегда ссылается на прямой референс. Будем называть чтение, которое замапилось на прямой референс, F-чтением, на обратное — R-чтением.

Совпадение произошло в позиции i, если чтение является F-чтением и нуклеотид в прямом референсе в позиции i совпал с нуклеотидом в чтении, или если чтение является R-чтением и нуклеотид в прямом референсе в позиции i совпал с комплементарным к соответствующему нуклеотиду в чтении. В других случаях происходит несовпадение. По конвенции нуклеотиды в R-чтениях уже комплементированы. Поэтому пайлап (*pileup*) [21] можно всегда сравнивать с прямым референсом. В одной R-позиции на рис. 2.1, в то время как пайлап показывает $A \to C$ несовпадение во многих R-чтениях, это по факту технически $T \to G$ несовпадение по этой конвенции. Две позиции, показанные на этом гипотетическом примере, постоянно показывали бы смещение ошибок к одному направлению, то есть

контекстно-зависимую ошибку (более точнее, $T \to G$ несовпадение после $GCTGG$). Таблица сопряжённости для мотивов вычисляется следующим образом:

1. инициализировать $a = b = c = d = 0$;
2. для каждой F-позиции мотива m получаем пайлап и увеличиваем a на число совпадений в F-чтениях, b на число несовпадений в F-чтениях, c на число совпадений в R-чтениях и d на число несовпадений в R-чтениях;
3. для каждой R-позиции мотива m получаем пайлап и увеличиваем a на число совпадений в R-чтениях, b на число несовпадений в R-чтениях, c на число совпадений в F-чтениях и d на число несовпадений в F-чтениях.

Для начала нужно подсчитать число совпадений и различий с геномом в каждом из двух направлений для каждой геномной позиции в замапленных чтениях. Псевдокод представлен в листинге 1.

Листинг 1 Подсчет статистики по геномным позициям

```
for всех чтений в наборе do
  for каждой позиции в чтении do
    pos — соответствующая геномная позиция
    ref — референсный нуклеотид
    nucl — текущий нуклеотид в чтении
    if (чтение обратное) then
      if (nucl равен ref) then
        rm[pos] добавить единицу
      else
        rmm[pos] добавить единицу
      end if
    else
      if (nucl равен ref) then
        fm[pos] добавить единицу
      else
        fmm[pos] добавить единицу
      end if
    end if
  end for
end for
stat = (fm, rm, fmm, rmm)
```

После этого этапа известно состояние покрытия каждой геномной позиции. Нас интересуют только хорошо и равномерно покрытые пози-

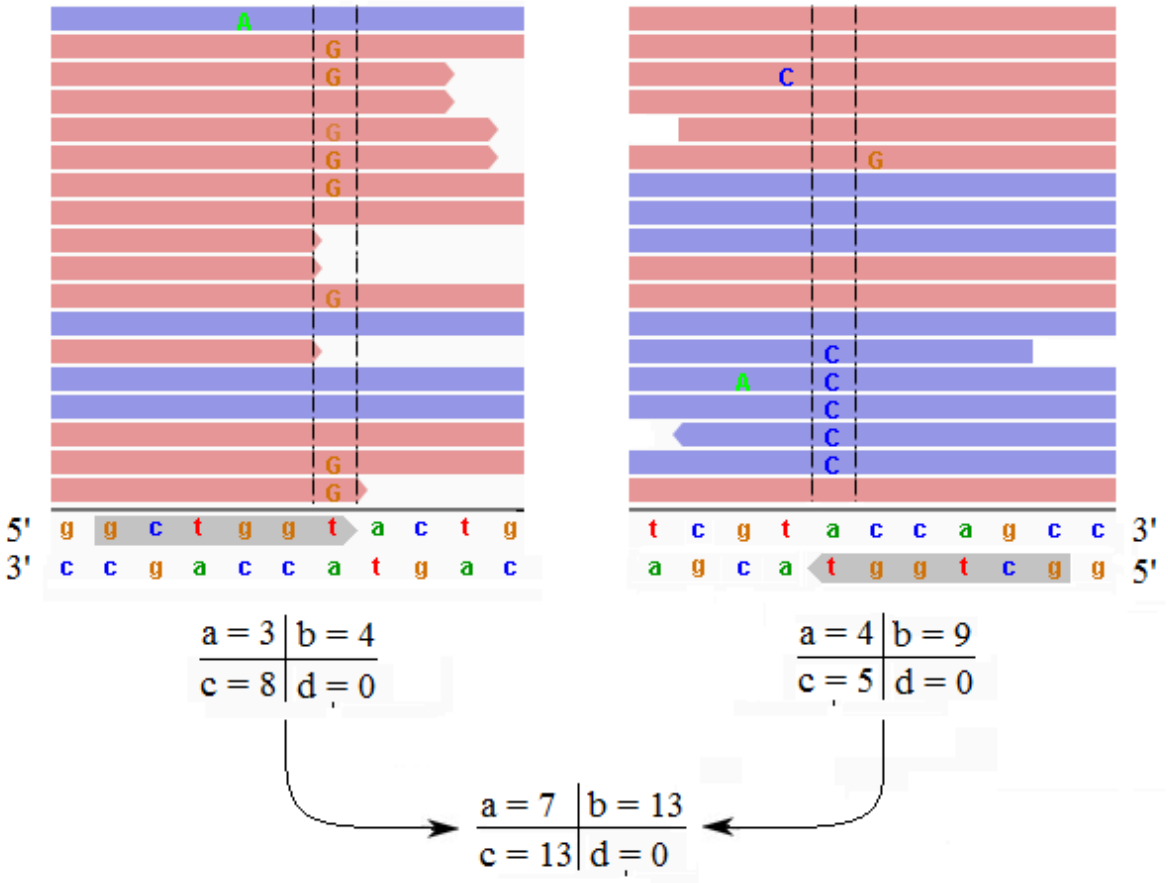

Рис. 2.1: Построение таблицы для мотива GCTGGT. Прямой референс (5' — 3') изображен снизу после чтений, чуть ниже комплементарный ему (3' — 5'). F-чтения обозначены красными стрелками, R-чтения — синими. F-интервал помечен на прямом референсе, R-интервал на обратно комплементарном. Одна F-позиция и одна R-позиция отмечены вертикальными столбцами. Таблица для каждой такой позиции и суммарная итоговая для мотива $GCTGGT$ показаны после референса внизу.

ции. Дело в том, что если в некоторой геномной позиции случится сильно смещенное в одну сторону покрытие и одновременно в ней же мутация, алгоритм неверно расценит такую ситуацию в сторону наличия плохого контекста. Поэтому отбираются позиции, чьи покрытия соотвествуют следующим условиям:

- покрытие с каждой из стороной должно быть больше или равно 3;
- отношение покрытия с большей стороны к меньше стороне не должно превосходить 5.

Далее в листинге 2 приведен псевдокод алгоритма подсчета таблицы сопряжённости для каждого k-мера.

Теперь имеется суммарная статистика для всех вхождений каждо-

Листинг 2 Подсчет статистики для к-мера с помощью статистики для геномных позиций

Вход: k — размер к-мера map — хэш-таблица для каждого к-мера $stat$ — статистика покрытия для геномной позиции

1: **for** всех К-мер, встречающихся в геноме **do**
2: **for** каждой позиции вхождения k-мера в геном **do**
3: $first$ — позиция первого элемента К-мера
4: $last$ — позиция последнего элемента
5: Добавить статистику последней позиции для k-мера $map[kmer] += stat[last]$
6: Добавить статистику первой позиции для обратно комплементарного к-мера $map[rev\_comp(kmer)] += stat[first]$
7: **end for**
8: **end for**

го k-мера, встречающегося в геноме. Заметим, что если существует ошибка, индуцируемая некоторым контекстом, то в чтении после это контекста ошибки будут появлятся чаще, чем в чтениях, идущих в противоположном направлении. Нужно лишь посчитать различие в вероятности возникновения ошибки с разных сторон. И чем она больше, тем хуже контекст.

2.2.1. Точный тест Фишера

Точный тест Фишера — тест статистической значимости взаимосвязи между двумя переменными в таблице сопряженности признаков размерности 2×2 [22].

Точный тест Фишера вычисляет *p-value* из таблицы сопряжённости 2×2 для определенения независимости двух характеристик данных: «направление чтения» и «доля несовпадений», что эквивалентно проверке, что строки имеют одинакое распределение. Если это так, то в данном случае нет основания на наличие смещения ошибок по направлениям.

	Совпадение	**Несовпадение**	**Всего**
Прямое	a	b	f
Обратное	c	d	k
Всего	m	s	n

Таблица 2.1: Таблица сопряженности 2×2.

Для расчета *p-value* таблицы 2.1, тест Фишера предполагает, что все маргинальные итоги заданы и фиксированы. Обозначим маргинальную информацию как $\mathcal{M} = (f,\ k,\ m,\ s,\ n)$, где $n = f + k = m + s$. Дано $\mathcal{M}$ и один элемент таблицы (без потери общности пусть это a), можно вычислить

все остальные элементы, и как раз $(a|\mathcal{M})$ обозначает такое представление таблицы. Вероятность нулевой гипотезы $Pr_{H_0}(a|\mathcal{M})$ наблюдаемой таблицы $(a|\mathcal{M})$:

$$Pr_{H_0}(a|\mathcal{M}) = \frac{\binom{a+b}{a} \cdot \binom{c+d}{c}}{\binom{a+b+c+d}{a+c}}$$

P-value $(a|\mathcal{M})$ — это вероятность наблюдать такую или или более экстремальную таблицу при нулевой гипотезе. Под более экстремальной таблицей подразумевается таблица с меньшей вероятностью, чем данная:

$$p{-}value(a|\mathcal{M}) = \sum_{a' \in E(a,\mathcal{M})} Pr_{H_0}(a'|\mathcal{M}),$$

где «экстремальные» значения a' из множества $E(a, \mathcal{M}) = \{a' : Pr_{H_0}(a'|\mathcal{M}) \leq Pr_{H_0}(a|\mathcal{M})\}$.

Если *p-value* достаточно мало, нулевая гипотеза отклоняется, подразумевая что две строки не были получены из одного распределения. Величину $-log_{10}(p{-}value(a|\mathcal{M}))$ можно рассматривать как количественную меру смещения.

Тест Фишера вычислительно сложен для таблиц с большими значениями и может быть заменен на χ^2-тест в этом случае [23].

2.2.2. Множественное тестирование

При проведении многих статистических тестов ожидаемое количество ложноположительных результатов может быть тоже большим. Существует много стратегий борьбы с такими ситуациями в множественной проверки гипотез. Например, одним из популярных подходов является контроль групповой вероятности ошибки (первого рода) в виде поправки Бонферрони [24].

Пусть $H_1, ..., H_m$ — набор гипотез, а $p_1, ..., p_m$ — их *p-value*, I_0 — неизвестное подмножество I истинных нулевых гипотез. Групповая вероятность ошибки (*family-wise error rate*, $FWER$) — вероятность отклонения как минимум одной гипотезы из I_0 (получения одной ошибки первого рода как минимум). Метод Бонферрони говорит, что для достижения $FWER \leq \alpha$, достаточно отвергать гипотезы с $p_i < \frac{\alpha}{m}$, где m — количество всех гипотез.

В нашем случае *p-value* важных мотивов очень малы по причине большого количества данных (чтений) и, отчасти, нашей стратегии объединения по контекстам. Таким образом, мы можем выбрать метод поправки Бонферрони для уменьшения потери многих важных мотивов. В таком случае поправка Бонферрони будет равна $T = \frac{\alpha}{|S(q)|}$, где α — константа, задающая желаемый уровень ошибок, а $|S(q)| = 4^q$ — мощность множества мотивов длины q, каждый элемент которых принадлежит алфавиту $\Sigma = \{A,\ C,\ G,\ T\}$.

2.2.3. Отбор существенных мотивов

Поскольку нулевой гипотезой являлось независимость направления от числа ошибок по направлениям, нужно наоборот оставить те мотивы, где присутствует дисбаланс, то есть чьи уровни значимости меньше или равны порогу Бонферрони: $p_i \leq T$.

Для всех таких мотивов подсчитывается прямая вероятность ошибки $FER = \frac{b}{a+b}$ и обратная вероятность ошибки $RER = \frac{d}{c+d}$ в обозначениях таблицы. Потом удаляются мотивы, имеющие $RER \geq \epsilon$, то есть мотивы, при которых вероятности ошибки с обеих сторон велики. Для оставшихся мотивов подсчитывается различие в вероятностях ошибок с каждой стороны $ERD = FER - RER$. Удаляются мотивы со слишком маленьким ERD, то есть $ERD < \delta$.

После определения наиболее склонных к ошибкам контекстов нужно оценить вероятность ошибки после них. По причине того, что в настоящем геноме присутствуют естественные мутации, нужно отобрать наиболее

уверенные позиции, где не возникает сомнений, мутация там или нет. Для этого по принципу максимального правдоподобия осуществляется отбор.

2.3. Отбор по принципу консенсуса

По принципу консенсуса, если в данной позиции только один тип нуклеотида, то он и будет выбран. Если два или больше, обычно обращают внимание только на два наиболее часто встречающихся. Можно считать, что позиция покрыта только двумя типами нуклеотидов. Этого можно добиться, если расценивать другие типы как ошибки.

Пусть позиция покрыта n чтениями. Для нуклеотида i-го чтения настоящий нуклеотид B_i и наблюдаемый $\hat{B}_i = \hat{b}^{(i)}$ с вероятностью ошибки ϵ_i. Для определенности будем считать первые k нуклеотидов как b_1 (то есть $\hat{b}^{(1)} = ... = \hat{b}^{(k)} = b_1$) и оставшиеся $n-k$ как b_2 (то есть $\hat{b}^{(k+1)} = ... = \hat{b}^{(n)} = b_2$). $\mathcal{D}$ обозначает наблюдаемые данные, а $\langle \cdot, \cdot \rangle$ — неупорядоченный генотип в этой позиции. Имеем:

$$P(\mathcal{D}|\langle b_2, b_2\rangle) = Pr\{\hat{B}_1 = ... = \hat{B}_k = b_1, \hat{B}_{k+1} = ... = \hat{B}_n = b_2 | \langle b_2, b_2\rangle\}$$
$$= \prod_{i=1}^{k} \epsilon_i \cdot \prod_{j=k+1}^{n} (1-\epsilon_j)$$

$$P(\mathcal{D}|\langle b_1, b_1\rangle) = \prod_{i=1}^{k} (1-\epsilon_i) \cdot \prod_{j=k+1}^{n} \epsilon_j$$

и

$$\begin{aligned}
&P(\mathcal{D}|\langle b_1, b_2\rangle) \\
&= Pr\{\hat{B}_1 = ... = \hat{B}_k = b_1, \hat{B}_{k+1} = ... = \hat{B}_n = b_2 | \langle b_1, b_2\rangle\} \\
&= \sum_{a_1=1}^{2} ... \sum_{a_n=1}^{2} Pr\{\hat{B}_1 = ... = \hat{B}_k = b_1, \hat{B}_{k+1} = ... = \hat{B}_n = b_2 | B_1 = b_{a_1}, ..., B_n = b_{a_n}\} \\
&\quad \cdot Pr\{B_1 = b_{a_1}, B_2 = b_{a_2}, ..., B_n = b_{a_n} | \langle b_1, b_2\rangle\} \\
&= \frac{1}{2^n} \prod_{i=1}^{n} \sum_{a_i=1}^{2} Pr\{\hat{B}_i = \hat{b}^{(i)} | B_i = b_{a_i}\}
\end{aligned}$$

Теперь по формуле Байеса:

$$P(\langle b_1, b_2\rangle|\mathcal{D}) = \frac{P(\mathcal{D}|\langle b_1, b_2\rangle)P(\langle b_1, b_2\rangle)}{P(\mathcal{D})}$$

По принципу максимального правдоподобия предполагаемым генотипом будет тот, который максимизирует эту апостериорную вероятность:

$$\hat{g} = argmax\ P(g|\mathcal{D}) = argmax\ P(\mathcal{D}|g)\ P(g)$$

Априорная вероятность того, что генотип является гетерозиготным $g = \langle b, b'\rangle$, равняется $r = 10^{-3}$ [25]. Тогда вероятность увидеть гомозиготный генотип $P(\langle b, b\rangle) = 1 - 2r$, так как генотип не упорядочен.

Для диплоидного организма в геномной позиции может быть один из десяти возможных генотипов: четыре гомозиготных (AA, CC, GG, TT) и шесть гетерозиготных (AC, AG, AT, CG, CT, GT). Из всех этих генотипов выбираются три наиболее вероятных, чьи нуклеотиды являются двумя доминантными. Например, если в некоторой позиции встречаются A и C чаще других, тогда возможные генотипы будут AA, CC, AC.

В силу того, что показатели качества у нуклеотидов являются не точными и хочется больше гарантий в правильности выбора, нужно сделать «зазор» между наиболее правдоподобным генотипом и вторым после него. Зазор представляет собой отношение логарифмов соответствующих вероятностей, которое не должно быть меньше некоторого «доверительного» значения. В том случае, если два наиболее правдоподобных генотипа находятся «близко» друг другу, нельзя придти к консенсусу в такой позиции. Позиции с правдоподобным гетерозиготным генотипом нужно исключать из доверительных, поскольку в них более велика вероятность мутации, чем ошибки.

В итоге отбираются позиции удовлетворяющие следующим условиям:

- покрытие с каждой из сторон должно быть больше или равно 3;
- отношение покрытия с большей стороны к меньшей стороне не должно превосходить 5;

- с каждой из сторон достигнут консенсус в выборе генотипа, эти генотипы совпадают и являются гомозиготными. В том случае, если с одной из сторон присутствует плохой контекст, такая сторона исключается, и генотип оставшейся стороны должен быть гомозиготным. В случае, если с двух сторон присутствуют плохие контексты, то такая позиция исключается;

2.4. Анализ контекстов чтений

По оставшимся «уверенным» позициям собирается статистика по теперь уже контекстам чтений.

Листинг 3 Подсчет статистики для к-мера с помощью статистики для геномных позиций

Вход: k — размер к-мера map — хэш-таблица для каждого к-мера и качества его последнего элемента

```
1: for всех чтений в наборе do
2:   for всех К-мер, встречающихся в чтении do
3:     first_phred — качество первого элемента К-мера
4:     last_phred — качество последнего элемента
5:     ref — референсный нуклеотид
6:     if (чтение обратное) then
7:       map[⟨reverse_complement(kmer), first_phred⟩][ref] добавить единицу
8:     else
9:       map[⟨kmer, last_phred⟩][ref] добавить единицу
10:    end if
11:  end for
12: end for
```

Теперь для каждого k-мера и показателя качества его последнего элемента имеется число совпадений с каждым типом нуклеотида референсного генома, который соответствовал этому последнему элементу в чтении. В идеале все совпадения должны приходиться на этот самый последний нуклеотид (например, для всех $ACCT$ в чтениях для нуклеотида T в соответствующей позиции в референсе тоже должен быть T, в идеале всегда). Но на практике в реальных данных присутствуют помехи — ошибки секвенирования. Поэтому секвенатор присваивает показатель своей уверенности каждому нуклеотиду. Опять же, в идеале, этот показатель уверенности должен соответствовать числу ошибок, которые были получены статистикой. На деле, даже при подсчете по «уверенным» позициям это не так. Это получается по нескольким причинам:

- качество, данное секвенатором, не соответствует действительности;
- в полученных «уверенных» позициях присутствуют трудноразличимые мутации (даже специальными алгоритмами для их поиска);
- наличие контекстно-зависимых ошибок секвенирования.

Но мутаций ограниченное число, и они случаются независимо, то есть вносят ошибки во все контексты примерно одинаково, тогда как контекстные ошибки влияют только на определенные контексты и значительно портят их статистику совпадений. Именно по этой статистике и определяются наиболее значительные ошибки и их контексты. После всего останется только заменить в наборе чтений показатели качества этих контекстов на соответствующие.

2.5. Выводы по главе 2

Был описан предлагаемый метод для вычисления вероятностей ошибок на основе последовательности нуклеотидов перед ними в чтениях. Описаны возникающие проблемы и способы борьбы с ними.

Глава 3. Результаты работы метода

Предложенный метод был опробован на чтениях генома человека. За референсный геном была взята сборка $GRCh37$ ($hg19$, *Genome Reference Consortium Human Reference 37* ($GCA\_000001405.1$) [26]. В качестве чтений была взяла библиотека парных чтений $SRR622461$ [27], предоставленная 1000 *Genomes* [28] и полученная секвенатором *Illumina HiSeq 2000* [29]. Чистые чтения были замаплены с помощью программы BWA (Burrows-Wheeler Aligner) [30] командами:

$bwa\ index\ -a\ bwtsw\ ucsc.hg19.fasta$

$bwa\ aln\ -q\ 15\ -f\ read1.sai\ ucsc.hg19.fasta\ SRR622461\_1.filt.fastq.gz$

$bwa\ aln\ -q\ 15\ -f\ read2.saßucsc.hg19.fasta\ SRR622461\_2.filt.fastq.gz$

$bwa\ sampe\ -a\ 750\ -f\ fullgenome.sam\ ucsc.hg19.fasta\ read1.sai\ read2.sai$
$SRR622461\_1.filt.fastq.gz\ SRR622461\_2.filt.fastq.gz$

Далее были удалены дубликаты с помощью *Picard MarkDuplicates* [31] и были удалены неуникальные чтения, которые мапятся в разные позиции, с помощью *SamTools* [32]. В таблице 3.1 показано число чтений, замапленных на каждую хромосому, и среднее покрытие каждой геномной позиции.

Номер хромосомы	Длина	Количество соотв. чтений	Ср. покрытие
1	249250621	11624327	4.66
2	243199373	12297590	5.06
3	198022430	10139425	5.12
4	191154276	9819297	5.14
5	180915260	9153074	5.06
6	171115067	8584662	5.02
7	159138663	7946173	4.99
8	146364022	7336225	5.01
9	141213431	5780517	4.09
10	135534747	6904404	5.09
11	135006516	6677618	4.94
12	133851895	6694814	5
13	115169878	4976920	4.32
14	107349540	4516567	4.2
15	102531392	4031897	3.9
16	90354753	4058929	4.49
17	81195210	3761871	4.63
18	78077248	3884870	4.98
19	59128983	2636592	4.46
20	63025520	2981738	4.73

Таблица 3.1: Количество чтений, замапленных на каждую хромосому.

3.1. Геномные контексты

Далее в таблицах 3.2, 3.3, 3.4, 3.5 и 3.6 предоставлены результаты работы первой части алгоритма по поиску наиболее склонных к ошибках контекстов. Были проведены запуски для контекстов длин 3, 4, 5, 6, 8 соответственно. В таблицах показаны наиболее значимые контексты с разностью долей ошибок $ERD \geq 0.001$, $ERD \geq 0.01$ и $ERD \geq 0.1$ в зависимости от длины контекста. Можно заметить, что контексты большей длины увеличивают разницу между вероятностями разносторонних ошибок. Это вызвано тем, что контекстная ошибка более специфична, и малой длины недостаточно для её явного выявления. При увеличении длины увеличивается пространство контекстов для поиска, и точнее охарактеризовывается последовательность. Например, контекст $CGGGT$ является самым «плохим» контекстом длины пять, но из его возможного расширения до длины шесть (контексты $NCGGGT$, где N – любой нуклеотид) только контекст $GCGGGT$ присутствует в таблице контекстов длины шесть 3.5 и также является самым «плохим» среди них. Получается, что остальные три кон-

текста $ACGGGT$, $CCGGGT$, $TCGGGT$ не склонны к ошибкам, хотя они считаются «плохими» для длины пять.

К-мер	Вхождения	Пр. сов.	Пр. несов.	Обр. сов.	Обр. несов.	Пр. ошибка	Обр. ошибка	Разность
GGT	60732854	41985781	42586008	565297	118547	0.013	0.003	0.01
AGT	84204009	67999337	68419092	490541	210975	0.007	0.003	0.004
CGT	13132547	8961784	8991629	51931	32347	0.006	0.004	0.002
TGT	105296533	86717445	86882193	389189	229229	0.004	0.002	0.002

Таблица 3.2: Статистика для контекстов длины три.

К-мер	Вхождения	Пр. сов.	Пр. несов.	Обр. сов.	Обр. несов.	Пр. ошибка	Обр. ошибка	Разность
GGGT	15196794	9460056	9665451	172636	25330	0.018	0.003	0.015
AGGT	20659881	14747122	14999554	215593	41287	0.014	0.003	0.011
CGGT	2626319	1512962	1516950	23800	7207	0.015	0.005	0.01
GAGT	18144417	13463587	13594955	141662	43941	0.01	0.003	0.007
TGGT	22249856	16265641	16404053	153268	44723	0.009	0.003	0.006
AAGT	25792322	22174554	22334986	168525	76986	0.008	0.003	0.005

Таблица 3.3: Статистика для контекстов длины четыре.

К-мер	Вхождения	Пр. сов.	Пр. несов.	Обр. сов.	Обр. несов.	Пр. ошибка	Обр. ошибка	Разность
CGGGT	990114	442364	452276	12915	1721	0.028	0.004	0.024
GGGGT	3886664	2145220	2197336	48798	5921	0.022	0.002	0.02
AGGGT	4459507	2969088	3035824	59436	7893	0.02	0.003	0.017
GAGGT	5713420	3545840	3642448	65807	11105	0.018	0.003	0.015
ACGGT	797855	502430	501395	12170	4555	0.023	0.009	0.014
GCGGT	726432	332825	335080	5298	710	0.016	0.002	0.014
CAGGT	6000798	3864332	3937277	60452	10627	0.015	0.003	0.012
AAGGT	5126210	4185958	4233780	59217	12563	0.014	0.003	0.011
TGGGT	5860509	3903384	3980015	51487	9795	0.013	0.002	0.011
GGAGT	5459860	3349963	3383885	44776	9614	0.013	0.003	0.01
CTGGT	5317600	3611481	3648318	41274	8474	0.011	0.002	0.009
AGAGT	5990699	4834273	4892044	57677	17450	0.011	0.003	0.008

Таблица 3.4: Статистика для контекстов длины пять.

К-мер	Вхождения	Пр. сов.	Пр. несов.	Обр. сов.	Обр. несов.	Пр. ошибка	Обр. ошибка	Разность
GCGGGT	251963	80429	87008	5272	291	0.062	0.003	0.059
AACGGT	150467	117374	116122	7766	3714	0.062	0.03	0.032
GGCGGT	166292	55775	58739	1919	148	0.033	0.003	0.03
CGGGGT	408673	146530	147834	4793	354	0.032	0.002	0.03
GGGGGT	840046	414738	431525	12632	1403	0.029	0.003	0.026
GCAGGT	1247617	741624	770163	21764	1959	0.027	0.002	0.025
GGAGGT	1720857	926138	957869	24750	3039	0.026	0.003	0.023
GGCTGT	1310515	792915	818194	18734	1989	0.023	0.002	0.021
GAGGGT	1030710	636174	658828	15192	1884	0.023	0.003	0.02
CGAGGT	394678	173761	179768	4264	640	0.024	0.004	0.02
CACGTA	262918	205672	205869	9025	4589	0.042	0.022	0.02

Таблица 3.5: Статистика для контекстов длины шесть.

K-мер	Вхождения	Пр. сов.	Пр. несов.	Обр. сов.	Обр. несов.	Пр. ошибка	Обр. ошибка	Разность
TCGAGTCA	9022	6903	7226	2792	612	0.288	0.078	0.21
GAACGGTT	7945	6613	6641	1635	53	0.198	0.008	0.19
TGGCGGGT	34865	6191	7405	1278	29	0.171	0.004	0.167
CGGCGGGT	5779	608	792	112	2	0.156	0.003	0.153
CGGCAGGT	15983	4286	5164	751	25	0.149	0.005	0.144
TGGCAGGT	115629	46330	55014	7477	155	0.139	0.003	0.136
CATTCGAA	11890	15898	15990	4664	1610	0.227	0.091	0.135
GACTCGAG	9676	5754	5700	1744	616	0.233	0.098	0.135
TGGCTGGT	103487	44936	53871	7137	128	0.137	0.002	0.135
CGGCTGGT	16037	4611	5446	700	10	0.132	0.002	0.130
CTCACCGA	12545	12380	12177	1925	62	0.135	0.005	0.130
CGCTTTGG	22270	17049	16607	4614	1609	0.213	0.088	0.125
GGCGGGGT	29042	6195	7084	778	20	0.112	0.003	0.109
AGGCGGGT	84997	23347	27209	2612	94	0.1	0.003	0.097
GTGGCTGT	108175	48167	55280	5042	123	0.095	0.002	0.093

Таблица 3.6: Статистика для контекстов длины восемь.

На рис. 3.1 показана позиция с явным присутствием склонности к ошибкам с одной стороны. Для визуализации использовался *Integrative Genomics Viewer, IGV* [33, 34].

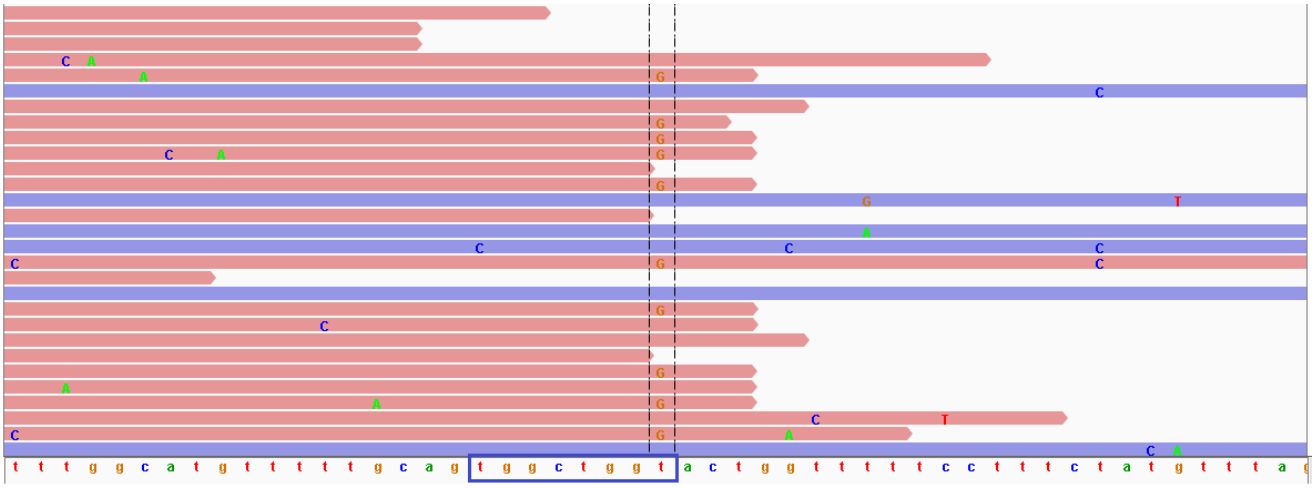

Рис. 3.1: Пример контекстных ошибок при *TGGCTGGT*. Красный цветом обозначены чтения, идущие в прямом направлении, красным — в обратном. Показана часть чтений позиции 7936309 двадцатой хромосомы. Всего в прямом направлении 49 совпадений и 43 ошибки G, в обратном только 106 совпадений.

Для наилучшей точности и в предположении о том, что малая часть контекстов индуцируют ошибки, были отобраны тринадцать наиболее ошибочных контекстов длины восемь с $ERD \geq 0.1$. Эти контексты показаны в таблице 3.7.

Контексты	ERD
TCGAGTCA	0.21
GAACGGTT	0.19
TGGCGGGT	0.167
CGGCGGGT	0.153
CGGCAGGT	0.144
TGGCAGGT	0.136
CATTCGAA	0.135
GACTCGAG	0.135
TGGCTGGT	0.135
CGGCTGGT	0.130
CTCACCGA	0.130
CGCTTTGG	0.125
GGCGGGGT	0.109

Таблица 3.7: Выбранные контексты с разностью вероятности ошибки с разных сторон.

Суммарное число вхождений всех этих контекстов не превышает и десятой доли процента для каждой хромосомы. Но в абсолютном значении это количество потенциальных мест на присутствие большого скопления ошибок секвенирования, и оно достаточно большое, чтобы повысить эффективность алгоритмов поиска мутаций в конечном итоге. Подробее показано в таблице 3.8.

Номер хромосомы	Длина	Вхождения	%
1	249250621	99988	0.04
2	243199373	98676	0.04
3	198022430	77766	0.039
4	191154276	69658	0.036
5	180915260	69917	0.038
6	171115067	66212	0.038
7	159138663	66578	0.041
8	146364022	58370	0.039
9	141213431	52373	0.037
10	135534747	57691	0.042
11	135006516	56392	0.041
12	133851895	55077	0.041
13	115169878	35968	0.031
14	107349540	37631	0.035
15	102531392	37064	0.036
16	90354753	41000	0.045
17	81195210	41942	0.051
18	78077248	30040	0.038
19	59128983	33961	0.057
20	63025520	29563	0.047

Таблица 3.8: Количество вхождений выбранных контекстов в каждую хромосому.

3.2. Выбор геномных позиций

Далее были отобраны удовлетворяющие нашим условия геномные позиции, на основе которых будет собираться статистика для контекстов чтений. В среднем было отобрано по 17 % у каждой хромосомы.

Номер хромосомы	Длина	Позиции	%
1	249250621	42379457	17
2	243199373	47137187	19.38
3	198022430	39654565	20.03
4	191154276	38920699	20.36
5	180915260	35810456	19.79
6	171115067	33371937	19.5
7	159138663	29581306	18.59
8	146364022	28299277	19.33
9	141213431	21491835	15.22
10	135534747	24405500	18
11	135006516	24955988	18.49
12	133851895	25381024	18.96
13	115169878	19601846	17.02
14	107349540	17217889	16.04
15	102531392	14970590	14.6
16	90354753	13176982	14.58
17	81195210	12211382	15.04
18	78077248	14936698	19.13
19	59128983	7432700	12.57
20	63025520	10427796	16.55

Таблица 3.9: Количество отобранных позиций для каждой хромосомы.

3.3. Контексты чтений

Учитывая отобранные позиции в геноме, была собрана статистика для каждого контекста длины восемь и качества его последнего нуклеотида. На рис. 3.2 заметно, что при качестве от 15 до 40 секвенатор сильно занижает оригинальные качества, а от 2 до 15 немного завышает. Особое внимание надо уделить качеству 2, которое, возможно, секвенатор присвает к очень зашумленным данным. Оно совсем не соответствует действительности, на практике ситуация гораздо лучше, чем оценивает её секвенатор.

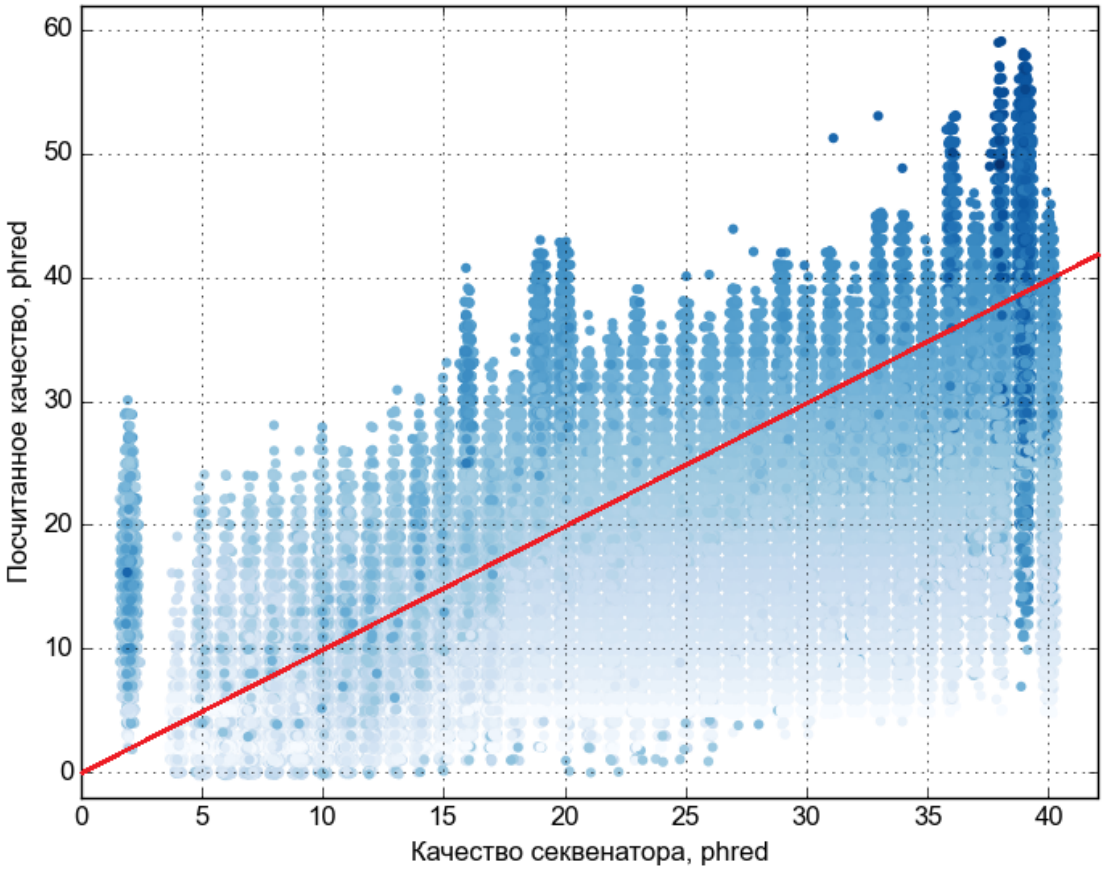

Рис. 3.2: График соответствия показателей качества нуклеотидов, присвоенных секвенатором, и посчитанных в зависимости от контекста. Точки — контексты. Интенсивность отражает количество вхождений контекста во все чтения. Красной линией изображено ожидаемое равенство присвоенных и посчитанных показателей. Ниже — переоценные, выше — недооцененные.

Рассмотрим геномный контекст $TGGCTGGGT$, в последнем элементе которого склонны скапливаться ошибки. Найдем контекст в чтениях, который вызывает эти ошибки. Возможно четыре соответствующих ему контекста: $TGGCTGGGT$, $TGGCTGGGG$, $TGGCTGGGA$, $TGGCTGGGC$. Посмотрев их статистику, легко сказать, что только у $TGGCTGGGG$ заявленные качества совсем не соответствуют посчитанным (рис. 3.3).

На рис. 3.4 приведена диаграмма распределения геномных нуклеотидов в позициях, которые соотвествуют последнему элементу контекста $TGGCTGGG$ в чтениях. Заметно, что для всех промежутков высока вероятность ошибки замены T на G, то есть секвенатор независимо от присвоен-

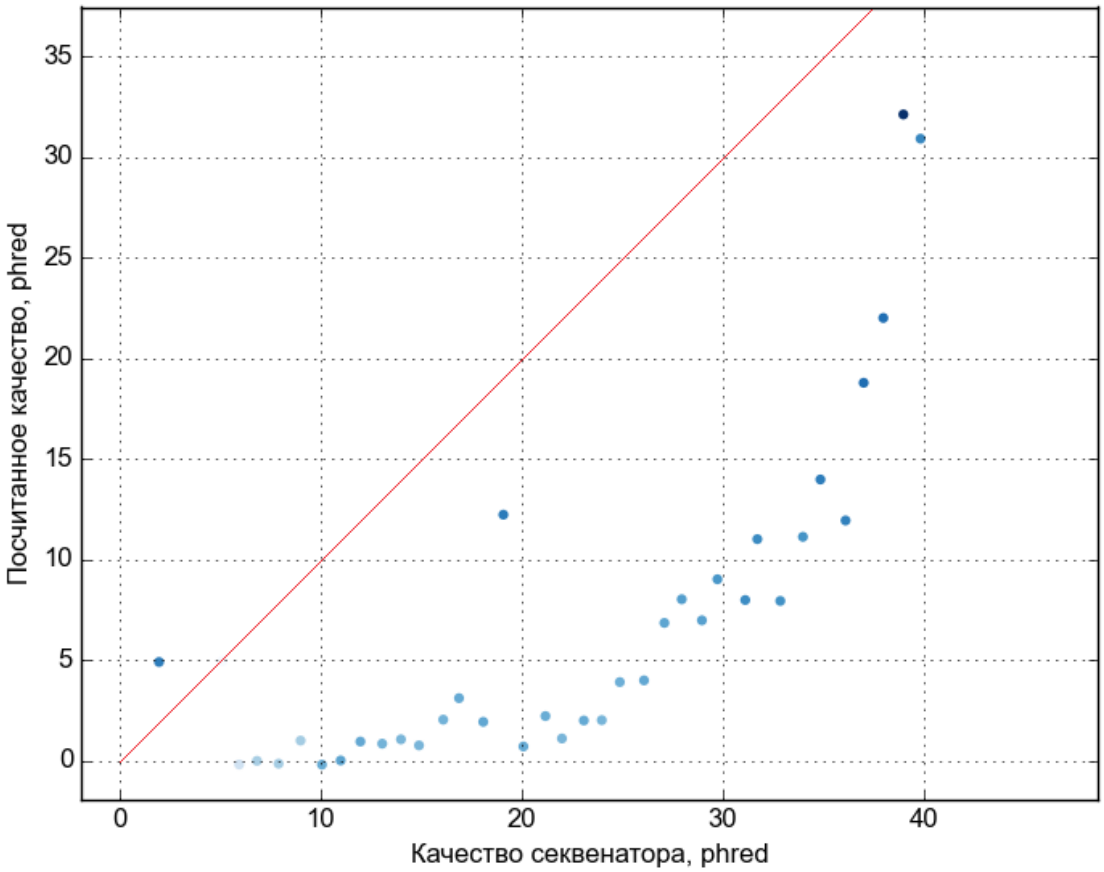

Рис. 3.3: График распределения качества нуклеотидов после контекста чтения $TGGCTGGG$. Все точки находятся намного ниже диагонали.

ного качества после последовательности нуклеотидов $TGGCTGG$ неверно прочитывает настоящий нуклеотид T в сторону G, тем самым вызывая скопления ошибок, и как следствие, ложный гетерозиготный генотип в соответствующей позиции. В таблице 3.10 показано подробное распределение нуклеотидов после этого контекста $TGGCTGGG$ для каждого показателя качества последнего элемента G.

3.4. Замена качества в чтениях и анализ эффекта

Замена (перекалибровка) показателей качества, присвоенных секвенатором, должна повышать эффективность поиска ОНП. Показателем эф-

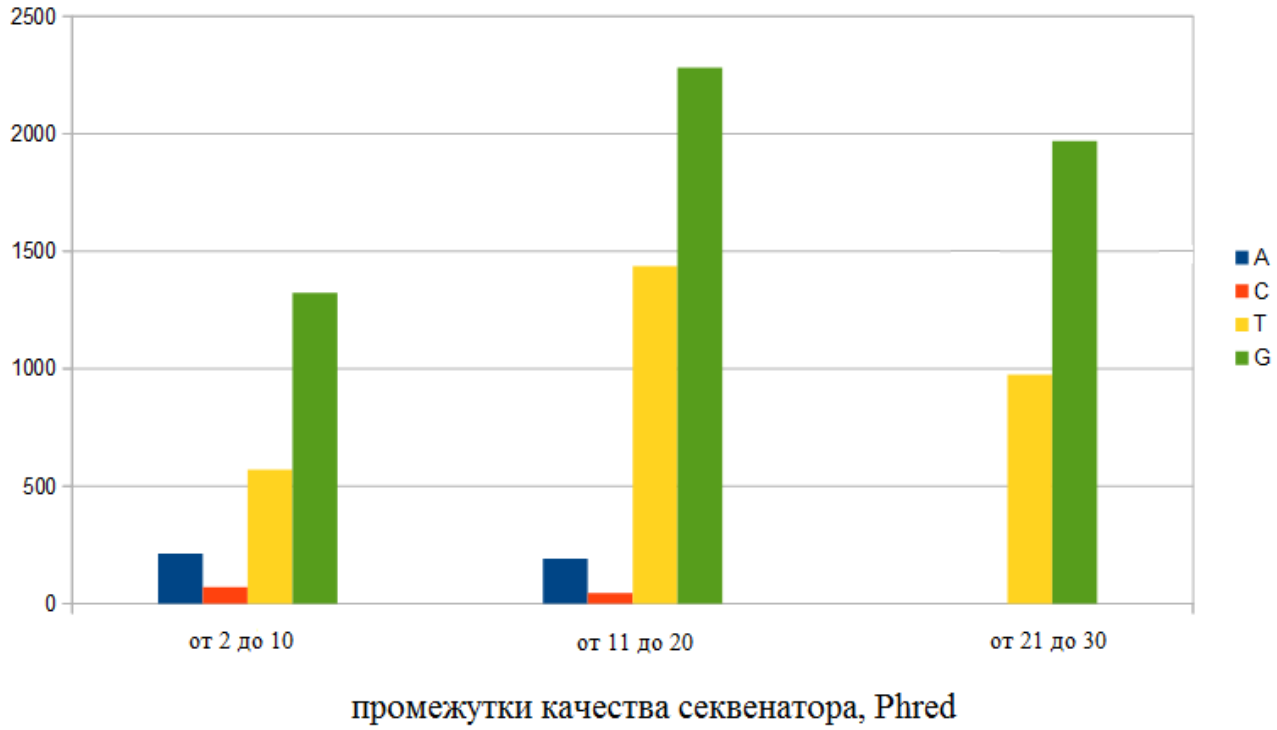

Рис. 3.4: Диаграмма распределения референсных нуклеотидов, соответствующих последнему нуклеотиду G контекста $TGGCTGGG$.

фективности чаще всего является отношение ложноположительных мутаций ко всем найденным, так называемая частота ложных обнаружений (*False Discovery Rate, FDR*). Для валидации был выбран размеченный набор мутаций двадцатой хромосомы, который был специально подготовлен консорциумом *Genome in a Bottle* [35], организованный Национальным институтом стандартов и технологий (*National Institute of Standards and Technology, NIST*) [36], версии 2.18 [37]. Этот набор был создан с помощью объединения 14 различных библиотек чтений, полученных на пяти различных технологиях секвенирования, и удалением участков с неразрешимыми различиями или с наличием признаков дисбаланса, описанного в предыдущей главе. Для сравнения были выбраны следующие данные:

- оригинальный набор чтений 20-ой хромосомы;
- набор чтений с заменой только тех качеств, которые лучше оригинальных;
- набор чтений с заменой всех качеств нуклеотидов на рассчитанные;

Phred-качество секвенатора	Вероятность ошибки по Phred	A	C	T	G	Вероятность ошибки	Вероятность T
2.0	0.631	138	49	330	1301	0.2846	0.1815
7.0	0.1995	6	1	32	1	0.9524	0.8
8.0	0.1585	13	3	26	3	0.9149	0.5778
9.0	0.1259	8	2	30	5	0.8723	0.6667
10.0	0.1	47	15	147	12	0.9417	0.6652
11.0	0.0794	73	16	224	15	0.9515	0.6829
12.0	0.0631	57	18	157	39	0.8535	0.5793
13.0	0.0501	34	3	100	23	0.8519	0.625
14.0	0.0398	10	5	91	34	0.7535	0.65
15.0	0.0316	5	1	114	21	0.8462	0.8085
16.0	0.0251	5	1	110	51	0.6923	0.6587
17.0	0.02	2	0	135	112	0.5498	0.5422
18.0	0.0158	2	0	188	91	0.6749	0.669
19.0	0.0126	3	0	123	1849	0.0642	0.0623
20.0	0.01	0	1	195	46	0.8074	0.8058
21.0	0.0079	0	0	126	93	0.5747	0.5753
22.0	0.0063	0	0	117	25	0.8194	0.8239
23.0	0.005	0	0	162	82	0.6626	0.6639
24.0	0.004	0	0	88	62	0.5855	0.5867
25.0	0.0032	0	0	83	106	0.4398	0.4392
26.0	0.0025	0	0	113	158	0.4176	0.417
27.0	0.002	0	0	68	242	0.2212	0.2194
28.0	0.0016	0	0	66	368	0.1537	0.1521
29.0	0.0013	0	0	70	302	0.1898	0.1882
30.0	0.001	0	0	81	531	0.1336	0.1324
31.0	0.0008	0	0	136	797	0.1465	0.1458
32.0	0.0006	0	0	84	947	0.0823	0.0815
33.0	0.0005	0	0	113	652	0.1486	0.1477
34.0	0.0004	0	0	61	640	0.0882	0.087
35.0	0.0003	0	0	68	1724	0.0385	0.0379
36.0	0.0003	0	0	94	1415	0.0629	0.0623
37.0	0.0002	0	0	44	3392	0.0131	0.0128
38.0	0.0002	0	0	17	3112	0.0057	0.0054
39.0	0.0001	0	0	25	41742	0.0006	0.0006
40.0	0.0001	0	0	1	1183	0.0008	0.0008

Таблица 3.10: Распределение геномных нуклеотидов при контексте чтений *TGGCTGGG*.

- набор чтений с просто заменой всех качеств на максимальные, которые может присвоить секвенатор.

Для поиска возможных мутаций был выбран метод *Unified Genotyper* [38], входящий в набор инструментов по анализу генома (*The Genome Analysis Toolkit, GATK*), разработанных *Broad Institute* [39]. Результаты эксперимента приведены в таблице 3.11. Точность показывает долю по-настоящему верных мутаций относительно всего количества кандидатов и вычисляется как $\frac{TP}{TP+FP}$, полнота — долю действительно верных среди всех известных

верных в тестовой выборке и равна $\frac{TP}{TP+FN}$. F-мера является гармоническим средним между полнотой и точностью и равна $2 \cdot \frac{Precision \times Recall}{Precision+Recall}$. Значения меры находятся в промежутке между нулем и единицей; чем больше значение, тем лучше.

По результатам самое большое количество истинно-положительных мутаций достигается при замене всех качеств нуклеотидов на максимально возможное для секвенатора, то есть на $phred$ равным сорока для секвенаторов $Illumina$. По сути все нуклеотиды в чтениях становятся равнозначными и решающую роль начинает играть их количество (соотношение). В таком случае небольшое наличие одинаковых нуклеотидов, которые отличаются от референсного, заставляет воспринимать их как мутацию. Естественно, будет заявлено больше подозрительных позиций, некоторая часть которых действительно будет являться настоящими мутациями, что и показал эксперимент. Но вместе с ними будет и больше ложно-положительных (мнимых) мутаций, что крайне нежелательно. В итоге такая замена имеет наибольшую полноту и наименьшую точность. В то же время F_1-мера тоже максимальна, но в нашем случае точность и полнота не равнозначны. Важнее точность или аналогичная ей величина $FDR = 1-$ точность.

Замена только на улучшающие показатели качества, то есть на такие показатели, которые больше оригинальных, увеличивает как точность, так и полноту. Такая замена добавляет некоторое количество настоящих мутаций, не увеличивая, а в данном случае даже уменьшая количество мнимых. Это показывает, что метод правильно расчитывает улучшающие показатели качества (как минимум) и позволяет немного, но повысить содержание настоящих мутаций.

Замена всех показателей (как улучшающих, так и ухудшающих) дает еще большую полноту, сохраняя точность, чем просто замена на улучшающие. Такой способ имеет чуть меньшую F_1-меру, чем при замене на константу, но гораздо большую точность. Это показывает, что замена показателей качества согласно предложенному методу эффективнее, чем урав-

нивание (исключение) их. А так же доказывает, что показатели качества важны для эффективного поиска мутаций, оригинальные показатели секвенатора не соответствуют действительности и их можно улучшить, тем самым повысив эффективность поиска.

Данные	Всего	TP	FP	FN	TN	Точность, %	Полнота, %	F1-мера	FDR
Ориг.	102240	61633	614	13649	26344	99.01	81.87	0.8963	0.99
Замена на лучшее	102240	62568	**612**	12714	26346	99.03	83.11	0.9038	0.97
Замена всех	102240	63641	621	11641	26337	**99.03**	84.54	0.9121	**0.97**
Замена на константу	102240	**63902**	697	11380	26261	98.92	**84.88**	**0.9137**	1.08

Таблица 3.11: Результаты эксперимента.

3.5. Выводы по главе 3

Приведены результаты основных частей метода. Приведен пример, что контекстно-зависимые ошибки секвенирования действительно имеют место в реальных данных секвенатора. Предложенный метод успешно находит такие контексты и учитывает их при подсчете статистики для расчета вероятностей ошибок. Было показано, что показатели качества нуклеотидов важны для эффективного поиска мутаций, и замена их согласно построенной статистической модели позволяет повысить эту эффективность.

Заключение

Современные секвенаторы совершают ошибки в определении правильного нуклеотида и присновенные ими показатели качества не соответствуют действительности. В настоящей работе был предложен метод определения вероятности ошибки секвенирования в зависимости от её контекста. Был реализован соответствующий алгоритм и опробован на человеческом геноме. По результатам экспериментов было показано, что метод повышает эффективность алгоритма поиска мутаций путём замены показателей качества нуклеотидов в чтениях на рассчитанные на основе их контекста.

Список литературы

1. *Watson J. D.*, *Crick F. H.* Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. // Nature. Vol.171. No.4356. Pp. 737–8.
2. *Nirenberg M.*, *Leder P.*, *Bernfield M.*, *Brimacombe R.*, *Trupin J.*, *Rottman F.*, *O'Neal C.* RNA Codewords and Protein Synthesis, VII. On the General Nature of the RNA Code / Proceedings of the National Academy of Sciences of the United States of America. Vol. 53. No.5. Pp. 1161–1168.
3. *Meselson M.*, *Stahl F. W.* The replication of DNA in ESCHERICHIA COLI / Proceedings of the National Academy of Sciences of the United States of America. Vol. 44. No.7. 1958. Pp. 671–82.
4. *Nielsen R.* Genomics: In search of rare human variants // Nature. 2010. Vol.467. №7319. С. 1050—1.
5. *Quail M. A.*, *Smith M.*, *Coupland P.*, *Otto T. D.*, *Harris S. R.*, *Connor T. R.*, *Bertoni A.*, *Swerdlow H. P.*, *Gu Y.* A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers // BMC Genomics. 2012. Vol.13. P. 341.
6. *Gilles A.*, *Meglécz E.*, *Pech N.*, *Ferreira S.*, *Malausa T.*, *Martin J.-F.* Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing // BMC Genomics. 2011. Vol.12. P. 245.
7. *Hoff K.* The effect of sequencing errors on metagenomic gene prediction // BMC Genomics. 2009. 10. P. 520.
8. *Dohm J. C.*, *Lottaz C*, *Borodina T*, *Himmelbauer H.* Substantial biases in ultra-short read data sets from high-throughput DNA sequencing // Nucleic Acids Research. 2008. Vol.36. No.16 e105.
9. *Taub M*, *Bravo H*, *Irizarry R.* Overcoming bias and systematic errors in next generation sequencing data // Genome Medicine. 2010. Vol.2. P. 87.
10. *Nakamura K*, *Oshima T*, *Morimoto T*, *Ikeda S*, *Yoshikawa H*, *Shiwa Y*, *Ishikawa S*, *Linak M. C.*, *Hirai A*, *Takahashi H*, *Altaf-Ul-Amin M*, *Ogasawara N*, *Kanaya S.* Sequence-specific error profile of Illumina sequencers // Nucleic Acids Research. 2011. Vol.39. No.13, e90.
11. *Meacham F*, *Boffelli D*, *Dhahbi J*, *Martin D*, *Singer M*, *Pachter L.* Identification and correction of systematic error in high-throughput sequence data // BMC Bioinformatics. 2011. Vol.12. P. 451.
12. 1000 Genomes Project Consortium: A map of human genome variation from population-scale sequencing // Nature 2010. Vol.467. No.7319. Pp. 1061–1073.
13. *Dohm J. C.*, *Lottaz C.*, *Borodina T.*, *Himmelbauer H.* Substantial biases in ultra-short read data sets from high-throughput DNA sequencing // Nucleic Acids Research. 2008. Vol.36. No.16 e105.
14. *DePristo M. A.*, *Banks E*, *Poplin R*, *Garimella K. V.*, *Maguire J. R.*, *Hartl C*, *Philippakis A. A.*, *Angel G*, *Rivas M. A.*, *Hanna M*, *McKenna A*, *Fennell T. J.*, *Kernytsky A. M.*, *Sivachenko A. Y.*, *Cibulskis K*, *Gabriel S. B.*, *Altshuler D*, *Daly M. J.* A framework for variation discovery and genotyping using next-generation DNA sequencing data // Nature Genetics. 2011. Vol.43. No.5. Pp. 491–8.
15. *McKenna A*, *Hanna M*, *Banks E*, *Sivachenko A*, *Cibulskis K*, *Kernytsky A*, *Garimella K*, *Altshuler D*, *Gabriel S*, *Daly M*, *DePristo M. A.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data // Genome Research. 2010. Vol.20. No.9. Pp. 1297–1303.
16. *Allhoff M.*, *Schönhuth A.*, *Martin M.*, *Costa I. G.*, *Rahmann S.*, *Marschall T.* Discovering motifs that induce sequencing errors // BMC Bioinformatics. 2013. Vol.14 Suppl. 5.
17. http://www.ncbi.nlm.nih.gov/BLAST/blastcgihelp.shtml.
18. *Cock P. J. A.*, *Fields C. J.*, *Goto N.*, *Heuer M. L.*, *Rice P. M.* The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants // Nucleic Acids Research. Vol.38. No.6. 1767–1771.
19. *Ewing B.*, *Green P.* Base-calling of automated sequencer traces using phred. II. Error probabilities. // Genome Research. Vol.8. No.3. Pp. 186–94.

20. *Webb T.* SNPs: can genetic variants control cancer susceptibility? // J Natl Cancer Inst. 2002. Vol.94. No.7. Pp. 476–8.
21. Pileup Format. http://samtools.sourceforge.net/pileup.shtml.
22. *Agresti A.* A Survey of Exact Inference for Contingency Tables // Statistical Science. 1992. Vol.7. No.1. Pp. 131–153.
23. *Greenwood P. E., Nikulin M. S.* A Guide to Chi-Squared Testing. Wiley, 1996.
24. *Dunn O. J.* Multiple Comparisons Among Means // Journal of the American Statistical Association. 1961. Vol.56. No.293. Pp. 52–64.
25. *Li H, Ruan J, Durbin R.* Mapping short DNA sequencing reads and calling variants using mapping quality scores // Genome Research. 2008. Vol.18. No.11. Pp. 1851–8.
26. GRCh37. http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/.
27. Библиотека чтений SRR622461. ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/phase3/data/NA12878/sequence_read/.
28. The 1000 Genomes Project. http://www.1000genomes.org/.
29. HiSeq 2000 Sequencing System. http://www.illumina.com/documents/products/datasheets/datasheet_hiseq2000.pdf.
30. *Li H, Durbin R.* Fast and accurate short read alignment with Burrows-Wheeler transform // Bioinformatics. 2009. Vol.25. No.14. Pp. 1754–60.
31. A set of tools (in Java) for working with next generation sequencing data in the BAM. http://broadinstitute.github.io/picard/.
32. *Li H, Handsaker B, Wysoker A, Fennell T, N Homer J. R., Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup.* The Sequence alignment/map (SAM) format and SAMtools // Bioinformatics. Vol.25. No.16. Pp. 2078–9.
33. *Robinson J. T., Thorvaldsdóttir H., Winckler W., Guttman M., Lander E. S., Getz G., Mesirov J. P.* Integrative Genomics Viewer // Nature Biotechnology. Vol.29. Pp. 24–26.
34. *Thorvaldsdóttir H., Robinson J. T., Mesirov J. P.* Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration // Briefings in Bioinformatics. Vol.14. Pp. 178–192.
35. Genome in a Bottle Consortium. https://sites.stanford.edu/abms/giab.
36. National Institute of Standards and Technology. http://www.nist.gov/.
37. NIST Genome in a Bottle NA12878 v2.18. ftp://ftp-trace.ncbi.nih.gov/giab/ftp/release/NA12878_HG001/NISTv2.18/.
38. GATK UnifiedGenotyper. https://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_gatk_tools_walkers_genotyper_UnifiedGenotyper.php.
39. Broad Institute. https://www.broadinstitute.org/.