Университет ИТМО

Факультет информационных технологий и программирования
Кафедра компьютерных технологий

Муравьёв Сергей Борисович

Разработка метода оценки качества сборки генома на основе принципа максимального правдоподобия

Научный руководитель: кандидат технических наук, доцент кафедры
Компьютерных технологий
Ф. Н. Царев

Санкт-Петербург
2014

Содержание

Введение

Задача корректной оценки качества сборки геномной последовательности [1] является важной частью биоинформатики. Задача состоит в том, чтобы по набору чтений и сборке геномной последовательности определить качество сборки, не имея референсного генома – уже известного генома данного организма. Основными проблемами в этом процессе являются наличие большого числа ошибок в исходных данных, большой объем входных данных, исчисляющийся сотнями гигабайт, а также отсутствие универсальной метрики оценки качества сборки.

Последние достижения в области секвенирования следующего поколения резко снизили стоимость секвенирования. С развитием сборщиков появилась возможность использовать большие объёмы секвенируемой информации. Благодаря технологии секвенирования методом дробовика стало появляться всё больше геномов различных организмов, от маленьких бактерий до млекопитающих. Несмотря на это, некоторые генетические последовательности получаются напрямую из материи, содержащей в себе несколько организмов, при помощи single-cell секвенирования и метагеномного секвенирования.

При сборке конкретного организма возникают ошибки, обусловленные шумами в информации, большим объёмом данных и особенностями сборщика. В случае с метагеномной сборкой возникают дополнительные трудности: глубина покрытия чтениями распределена неравномерано, неоднозначность в анализе повторяющихся областей в случае метагеномной сборки усугубляется. Было разработано несколько инструментов для обнаружения ошибок при неметагеномной сборке. Однако такие средства используют референсную или близкую к референсной сборку организма. В случае отсутствия референсной сборки подобные инструменты предоставляют косвенные оценки качества сборки генома, такие как *N50*, глубина

покрытия, оценка расстояния между парными чтениями. Такие метрики предоставляют информацию о производительности сборщика, но не обеспечивают внутренних прямых измерений качества сборки.

В данной работе было разработано улучшение одной из статистических моделей для оценки качества сборки генома в условиях отсутствия референсной сборки. Данная модель представляет из себя формулу на основе формулы Байеса, которая вычисляет вероятность того, что сборка правильная, то есть соответствует референсному геному.

Глава 1. Обзор предметной области

1.1. Основные понятия

Биоинформатика — это наука на стыке двух дисциплин: биологии и информатики [2]. Многие задачи биологии требуют обработки колоссального объема данных, что и привело к возникновению дисциплины. Одной из важных задач биоинформатики является задача корректной оценки собранной геномной последовательности.

1.1.1. Строение ДНК

Геном — это совокупность информации, передаваемой всеми живыми существами по наследству [3]. Геном большинства живых организмов состоит из молекул *ДНК* — дезоксирибонуклеиновой кислоты [4]. ДНК — это полимерная молекула, представляющая из себя две закрученных в спирали цепочки, состоящие из соединённых в последовательность *нуклеотидов* [5]. Нуклеотиды, входящие в ДНК, разделяют по азотистым основаниям на четыре группы: *аденин* (A), *цитозин* (C), *гуанин*(G), и *тимин* (T)[5]. Цепи ДНК соединены между собой по *принципу комплементарности*: аденин с тимином, цитозин с гуанином.

Геном в биоинформатике представляется одной из двух комплементарных цепей молекулы ДНК. Строка, состоящая из символов A, G, C и T, соответвующих типам нуклеотидов,является наиболее удобным представлением генома в биоинформатике.

1.1.2. Секвенирование генома

Для определения линейной последовательности нуклеотидов в молекуле ДНК геном подвергают секвенированию. Одним из популярных методов секвенирования является *метод дробовика* [6]. Метод состоит в выделении из молекулы ДНК коротких участков (порядка нескольких сотен

последовательных нуклеотидов), после чего происходит посимвольное считывание концов выделенных участков. Таким образом, получаются парные чтения (mate-pairs). В силу множества различных факторов при прочтении отдельных нуклеотидов могут быть допущены ошибки. Также неизвестно точное расстояние между чтениями, известно лишь распределение длин фрагментов.

1.1.3. Сборка генома

Традиционно процесс сборки генома состоит из трех этапов: исправление ошибок в парных чтениях, сборка *контигов*, длинных последовательных частей генома, и сборка *скэффолдов*, наборов упорядоченных ориентированных контигов с оценками расстояния между соседними контигами. Таким образом, задачей сборки скэффолдов является построение вышеупомянутых наборов по множеству контигов и библиотекам парных чтений. Про библиотеки парных чтений известны математическое ожидание и стандартное отклонение длин фрагментов, из которых были получены парные чтения.

1.2. Постановка задачи

В последнее время появилось множество сборщиков генома. Возникла проблема сравнения качества их работы. В прошлом это делалось при помощи таких популярных метрик как *N50/N90*, *NG50/NG90*, длина наибольшего контига или скэффолда [7]. Хотя исследования показали, что простые метрики коррелируют с качеством сборки, используемые в настоящее время метрики являются грубыми и не обеспечивают полной информации о результате сборки [8]. Например сборка, состоящая из просто склеенных конец-в-конец чтений, имеет очень большой *N50*, но, очевидно, плохое качество сборки.

Требуется разработать метрику, оценивающую качество сборки на основе принципа максимального правдоподобия.

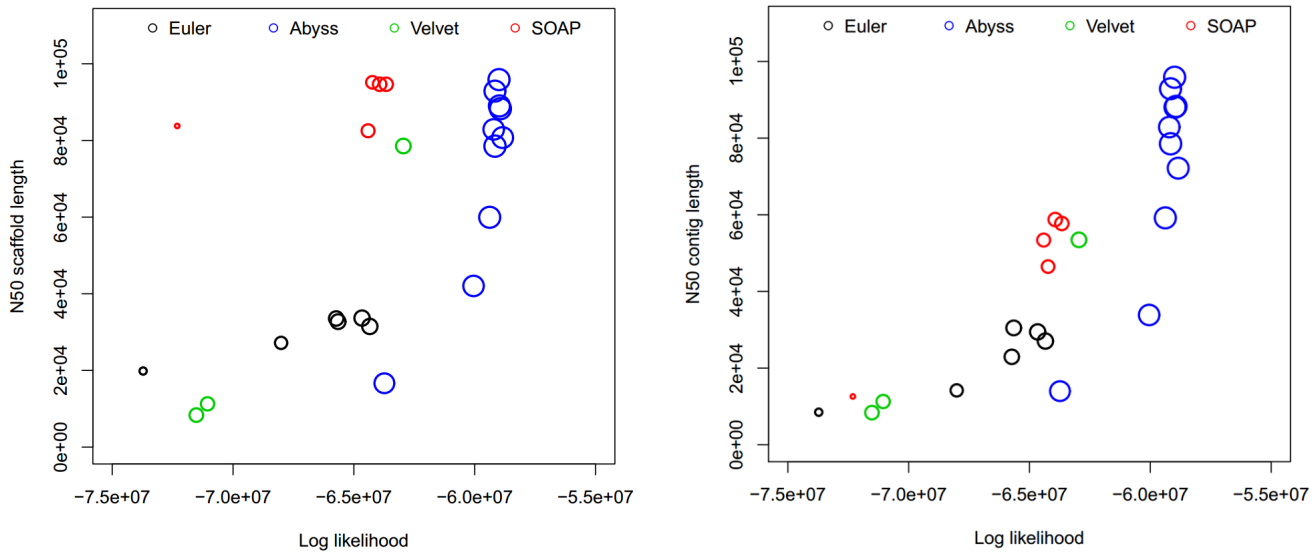

Рис. 1.1. Графики корреляции N50 и loglikelihood, посчитанным с помощью CGAL, для *E.coli*. Окружности соответствуют сборкам с различными длинами k-меров. Размер окружности соответствует схожести с эталонной сборкой.

1.3. Обзор существующих методов и их проблем

За последние два года было разработано несколько новых способов качественной оценки сборки геномной последовательности на основе принципа максимального правдоподобия. Их общая идея заключается в получении *loglikelihood* – логарифма вероятности того, что сборка является верной при наличии заданного набора чтений.

1.3.1. CGAL

Данный метод [9] является первым методом оценки на основе принципа максимального правдоподобия. В его основе лежит слудующая формула:

$$l(A,R) = \ln \prod_{i=1}^{n} p(r_i|A) \approx \sum_{i=1}^{N} \ln \sum j = 1_i^M p_F(l_{i,j}) p_S(s_{i,j}) p_E(r_i|a_{i,j}, e_{i,j})$$

- R – множество чтений;
- A – сборка;

- M_i – количество возможных "соответствий"в сборке чтения r_i;
- $l_{i,j}$, $s_{i,j}$, $a_{i,j}$ и $e_{i,j}$ – соответственно длина чтения, позиция чтения, подпоследовательность сборки и ошибки для "соответствия"j для чтения i.

Проблема данного метода заключается в предположении, что чтения распределены равномерно по всей длине генома. Это является существенным недостатком, поскольку в реальной ситуации чтения распределены чаще всего неравномерно.

1.3.2. de Novo

В основе данного метода [10] лежит формула Байеса:

$$P(A|R) = \frac{P(R|A)P(A)}{P(R)}$$

A – событие, при котором сборка является эталонной геномной последовательностью R – событие, при котором исследуется определённый набор чтений. $P(A)$ и $P(R)$ являются константами.

$$P(R|A) = \prod_{r \in R} P(r|A)$$

Задача сводится к получению оценки $P(r|A)$ с использованием динамического программирования. Данный метод достаточно точно вычисляет вероятности ошибок для больших наборов чтений. Однако существенными недостатками являются большой объём потребляемой памяти ($O(n^2)$, где n – длина сборки) и сложность в реализации данного алгоритма. Кроме того, на сегодняшний день нет ни одной работающей реализации данного подхода.

1.3.3. ALE

На данный момент этот метод [11] является наиболее совершенным. В его основе также лежит формула Байеса, как и в *de Novo*.

$$P(A|R) = \frac{P(R|A)P(A)}{Z}$$

Z – константа. $P(A)$ описывает качество сборки в отсутствие какой-либо информации о чтениях. $P(R|A)$ оценивается по следующей формуле:

$$P(R|A) = P_{placement}(R|A)P_{insert}(R|A)P_{depth}(R|A)$$

- $P_{placement}$ оценивает, насколько содержание чтений совпадает со сборкой
- $P_{insert}(r|A)$ – оценивает, насколько априорные расстояния между парными чтениями (insert length) совпадают с получившимися в результате сборки
- $P_{depth}(r|A)$ – оценивает, насколько априорная глубина покрытия в каждой позиции совпадает с получившейся в результате сборки на основе GC-контента.

Глубина покрытия – это количество чтений, которое покрыло данную позицию в сборке.

GC-контент – это процентный состав суммы всех нуклеотидов, являющихся гуанином(G) или цитозином(C) по отношению к длине исследуемого участка генома.

Более подробно о принципах работы и проблемах, возникающих при использовании данного метода будет изложено в следующей главе.

1.4. Выводы по главе 1

В условиях быстро появляющихся новых сборщиков, возникла задача качественного сравнения их между собой, поэтому разработка метода оценки качества сборки геномной последовательности является важной задачей биоинформатики. Существующие популярные метрики, такие

как *N50/N90*, не дают полной информации о результате сборки, поэтому был предложен новый подход, заключающийся в вычислении *loglikelihood* – логарифма вероятности того, что сборка является верной при заданном наборе чтений. В последнее время было предложено несколько способов получения *loglikelihood*, однако они не лишены своих недостатков. Наиболее совершенным инструментом оценки на сегодняшний день является ALE. В данной работе представлен способ оценки качества на базе ALE.

Глава 2. Описание используемого подхода

2.1. Особенности ALE

2.1.1. Описание метода

Как уже было изложено выше, ALE оценивает качество сборки при помощи следующей формулы:

$$P(A|R) = \frac{P_{placement}(R|A)P_{insert}(R|A)P_{depth}(R|A)P(A)}{Z}$$

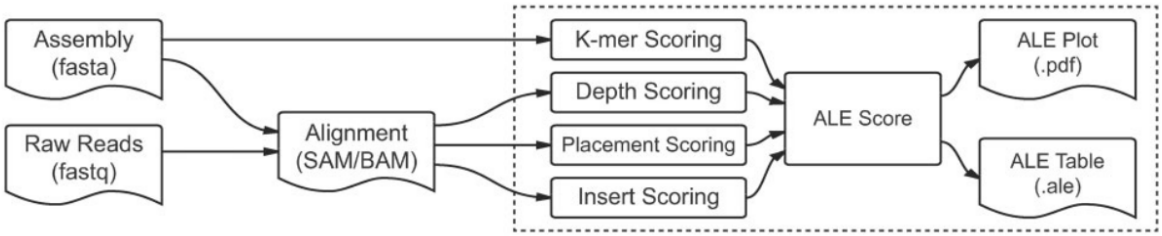

Рис. 2.1. Общая схема работы ALE

2.1.1.1. Нормализационная константа Z

Z является константой, вычисляется как:

$$Z = \sum_{A'} P(R|A')P(A')$$

Данную величину вычисляют приближённо, поскольку явно её вычислить невозможно из-за слишком большого множества сборок A' (4^L, где L – длина сборки).

2.1.1.2. Оценка сборки при отсутствии чтений

$$P(A) = ZP_{kmer}(A)$$

$P_{kmer}(A) = \Pi_{i \in K} f_i^{n_i}$, где K – множество уникальных k-меров, n_i – количество раз, когда k-мер i встречается в текущем контиге в сборке. f_i – частота появление k-мера i в контиге: $f_i = n_i / \sum_{j \in K} n_j$.

2.1.1.3. Корректность чтений

$P_{placement}(R|A)$ – оценивает насколько чтения соответствуют сборке, выражается следующим образом:

$$P_{placement}(R|A) = \prod_{r_i \in R} P_{placement}(r_i|A) = \prod_{r_i \in R} P_{matches}(r_i|A) P_{orientation}(r_i|A)$$

$P_{matches}(r_i|A)$ оценивает, насколько содержимое чтения соответствует тому участку сборки, на который было произведено картирование данного чтения. Каждый нуклеотид j, считанный при помощи секвенатора, имеет качество считывания Q_j, $Q_j \subseteq [0; 1]$, тогда $P_{matches}(r_i|A) = \Pi_{j \in r_i} P_{base_j|A}$, где $P_{base_j|A} = Q_j$ когда нуклеотид j совпадает со сборкой и $P_{base_j|A} = (1 - Q_j)/4$ в противном случае. Если в сборке встречается неизвестный нуклеотид N, то считаем, что $P_{base_j|A} = 1/4$ Если инструмент картирования сопоставил чтение более чем в одно место, ALE выбирает позицию, у которой $P_{placement}(R|A)$ наибольший.

$P_{orientation}(r_i|A)$ оценивает корректность ориентации в случае парных чтений. Величина вычисляется эмпирически.

2.1.1.4. Расстояние между парными чтениями

$P_{insert}(R|A)$ оценивает расстояние между парными чтениями, вычисляется как $P_{insert}(R|A) = \Pi_{r_i \in R} P_{insert}(r_i|A)$. $P_{insert}(r_i|A) = Normal(L_i; \mu, \sigma^2)$, где L_i – расстояние между парными чтениями r_i, а параметры μ и σ^2 подбираются на основе информации о расстояниях между всеми парными чтениями.

2.1.2. Оценка глубины покрытия

Рассмотрим более подробно вычисление $P_{depth}(R|A)$. Эта величина описывает, насколько глубина покрытия в каждой позиции, соответствует

глубине, которую мы бы ожидали увидеть, если бы картирование производилось на эталонную сборку.

$$P_{depth}(R|A) = \prod_i P_{depth}(d_i|A)$$

d_i – глубина в позиции i. Предполагается, что глубины распределены по Пуассону с центром, вычисленным из независимого гамма-распределения с центром в ожидаемой глубине в данной позиции и GC-контентом в качестве второго параметра. Рассмотрим это утверждение более подробно.

Сначала рассчитываются средние глубины для каждого из 100 множеств GC-контента: 0-1, 1-2 ... 99-100%. Пусть X_i – это средний GC-контент по всем чтениям, которые покрывают позицию i. $\mu_{depth}(X_i)$ – средняя глубина для того множества GC-контентов, которому соответствует X_i. Минимальное значение $\mu_{depth}(X_i)$ устанавливается равным 10. В итоге для каждой позиции i в сборке оценка глубины будет производиться по следующей формуле:

$$P_{depth}(d_i|A, X_i) = \int_0^\infty Poisson(d_i, Y_i)Gamma(Y_i, max(10, \mu_{depth}(X_i), 1))dY_i =$$
$$= NegBinom(d_i, \mu_{depth}(X_i), 0.5)$$

Подытоживая вышесказанное, получаем, что глубины распределены по отрицательному биномиальному распределению $NB(r, p)$ [12] с параметрами $r = \mu_{depth}(X_i)$ и $p = 0.5$.

2.1.3. Проблема оценки глубины покрытия

В ходе выполнения данной работы было выявлено, что глубины покрытий распределены не по отрицательному биномиальному распределению. Для того, чтобы убедиться в этом для каждого из 100 множеств GC-контента было построено эмпирическое распределение глубин покрытий. В итоге получилась таблица $100 \times maxdepth$, где $maxdepth$ – это максимальное значение глубины покрытия позиции чтениями в данной сборке.

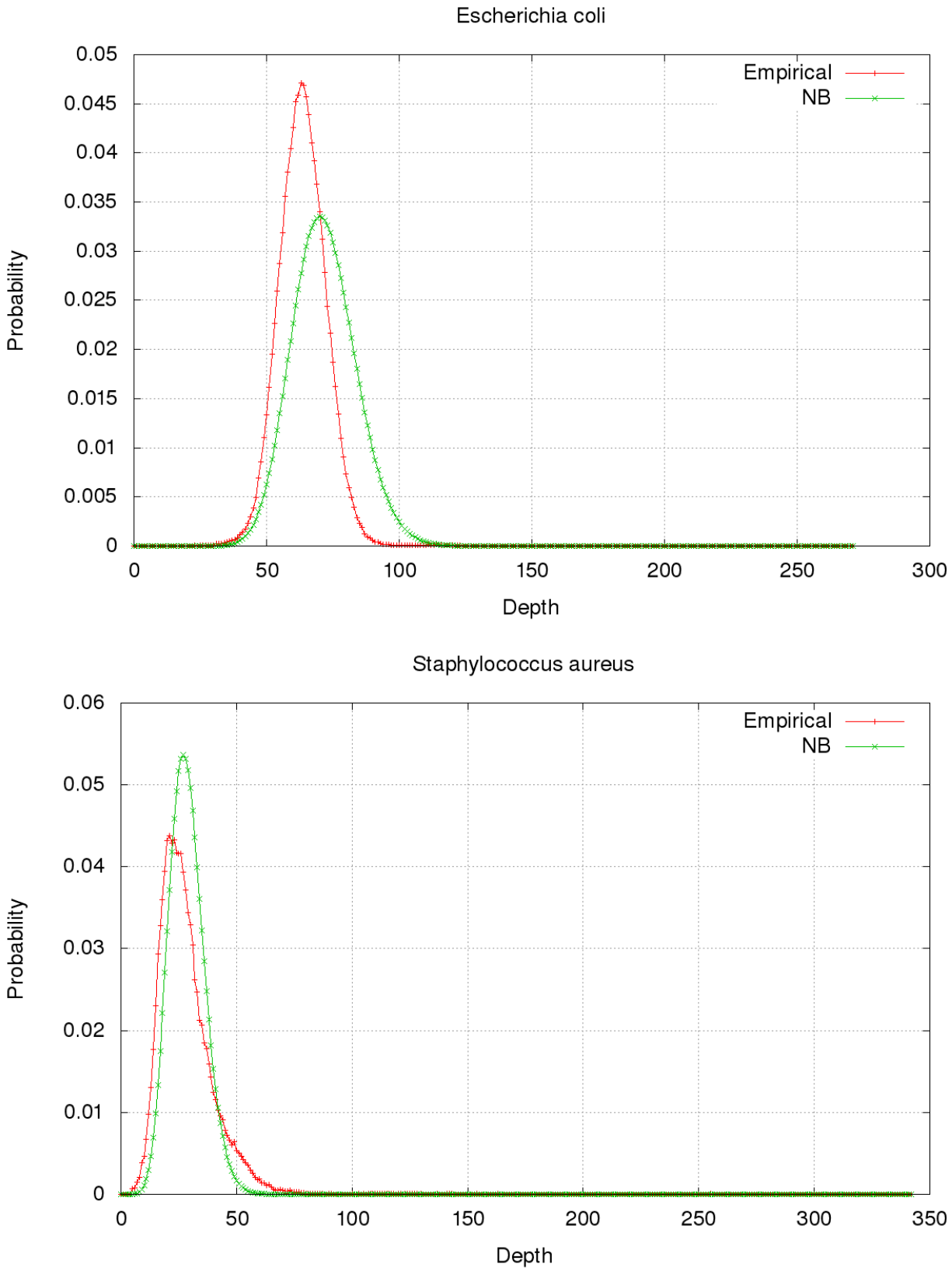

Рис. 2.2. Графики распределения глубин покрытий для E.coli и для Staphylococcus aureus, посчитанные эмпирически и при помощи отрицательного биномиального распределения.

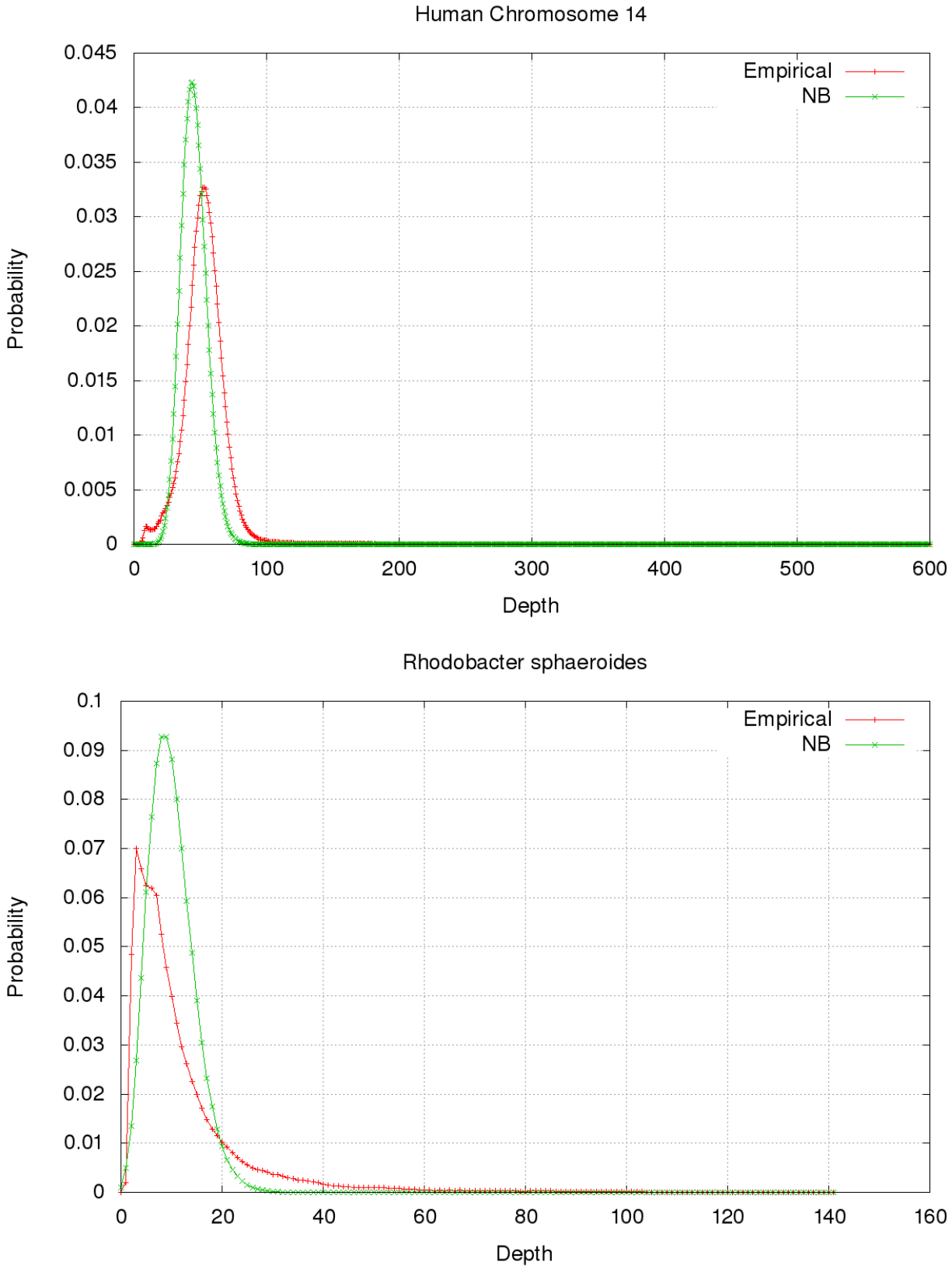

Рис. 2.3. Графики распределения глубин покрытий для 14-й хромосомы человека и для Rhodobacter sphaeroides, посчитанные эмпирически и при помощи отрицательного биномиального распределения.

Далее было выбрано такое множество GC-контента, которому соответствует максимальное число позиций в сборке, и построены графики сравнения эмпирического и отрицательного биномиального распределения, представленные на графиках 2.2 и 2.3.

Также были проведены статистические тесты, которые показали, что реальное распределение (эмпирическое) глубины покрытия не соответствует теоретическому (отрицательному биномиальному). В частности на *Staphylococcus aureus* после проверки критерия χ^2 [13], было выявлено, что гипотеза о том, что глубины распределены по отрицательному биномиальному распределению неверна: p-value $= 2 \times 10^{-5}$ [14]. Приближение на основе принципа максимального правдоподобия [15] также показало отрицательный результат: логарифм вероятности H_0 – события, при котором глубина покрытия распределена по отрицательному биномиальному распределению, $ln(p(H_0)) \approx -5$.

2.2. Улучшение оценки P_{depth}

2.2.1. Общая идея

Предпринимались попытки приближения различными распределениями и комбинациями распределений, однако они не принесли результатов. p-value и логарифм вероятности соответствия двух распределений друг другу, описанный выше, либо не удавалось значительно улучшить, либо в случае видимого улучшения на сборке одного организма метод не работал на сборках других организмов.

В связи с этим, для оценки глубины покрытия в каждой позиции в данной работе предлагается использовать эмпирическое распределение. Оно заведомо точнее отрицательного биномиального, поскольку учитывает реальное распределение при заданном наборе чтений.

2.2.2. Выбор GC-контента

GC-контент в каждой позиции вычисляется как процентное содержание гуанина и цитозина по всем отрезкам длины n, где n – это средняя длина чтения. Аналогично ALE будем выделять 100 множеств GC-контентов: 0-1, 1-2 ... 99-100%. Такой выбор связан с тем, что у современных секвенаторов Illumina и Ion Torrent средняя длина чтений около 100 нуклеотидов [16, 17]. Были проведены опыты с разбиением на меньшее и большее число множеств. В первом случае существенно падает точность оценки, во втором сильно возрастает время работы оценщика, при этом значительного выигрыша в точности нет.

2.2.3. Учёт ошибок

Эксперименты показали, что в большинстве случаев, нуклеотиды, не покрытые чтениями, являются ошибками сборки геномной последовательности, поэтому предлагается при подсчёте эмпирического рапределения не учитывать позиции в геноме с нулевыми глубинами покрытия. Эффективность такого подхода будет рассмотрена в следующей главе.

2.2.4. Анализ производительности

В отличие от ALE, метод не требует предварительного вычисления средней глубины покрытия для заданного отрезка GC-контента $\mu_{depth}(X_i)$. Для хранения вероятностей глубин в каждой позиции требуется таблица T размером $100 \times maxdepth$, где $maxdepth$ – это максимальное значение глубины покрытия в данной сборке. Нахождение вероятности производится производится обращением к соответствующей ячейке таблицы $T[gc_i][depth_i]$, где $depth_i$ и gc_i – это соответственно глубина в позиции i и множество, к которому принадлежит GC-контент в позиции i. Это гораздо быстрее ALE, который высчитывает функцию вероятности для отрицательного биномиального распределения каждый раз заново для каждой позиции.

2.3. Выводы по главе 2

В данной главе был рассмотрен принцип работы ALE. Был выявлен его главный недостаток: оценка глубины покрытия на основе отрицательного биномиального распределения. Проведено сравнение с реальным распределением глубин покрытий, и обнаружено несоответствие на основе различных статистик. Для оценки глубины покрытия в каждой позиции в сборке было предложено использовать эмпирическое распределение без учёта непокрытых чтениями позиций, с глубиной $depth = 0$.

Глава 3. Применение используемого метода и результаты работы на различных тестах

3.1. Ошибки сборки

Ошибки, содержащиеся в сборке геномной последовательности содержится несколько типов ошибок:

- ошибки вставки/удаления участков генома
- замещения некоторых нуклеотидов
- разворот или копия участка последовательсти нуклеотидов

3.2. Корректность сравнения

В качестве результата программа, реализующая ALE, выдаёт $\ln(P(A|R))$, который называется *ALE-score*. Аналогично $\ln(P_{placement}(R|A))$, $\ln(P_{insert}(R|A)/Z)$ и $\ln(P_{depth}(R|A)/Z)$ именуются *ALE-placement-score*, *ALE-insert-score* и *ALE-depth-score* соответственно. Инструмент визуалицации сборок IGV [18] отображает *score* каждого типа в каждой позиции, поэтому для корректности сравнения результатов вероятность глубины покрытия, посчитанная при помощи предложенного метода, как и в ALE, была отнормирована по Z, а затем от полученного результата был взят натуральный логарифм. Будем называть эту величину *depth-score*.

3.3. Синтетические тесты

Чтобы проверить способность предложенного метода к обнаружению всех типов ошибок, были сгенерированы синтетические чтения из генома *Escherichia coli K12 Substrain DH10B* [19]. Затем было внесено несколько

тысяч ошибок различных типов, при этом было предварительно зафиксировано, где будут допущены сгенерированные ошибки. Далее было проведено картирование чтений на сборку с ошибками, результатом которого был SAM-файл. Имея чтения и SAM-файл, был запущен ALE и реализация метода, представленного в данной работе. При помощи IGV было проведено дальнейшее сравнение.

3.3.1. Выявление ошибок

В ходе выполнения синтетических тестах было внесено:

- 10 ошибок замещения по 1000 нулеотидов: из них все ошибки были найдены.

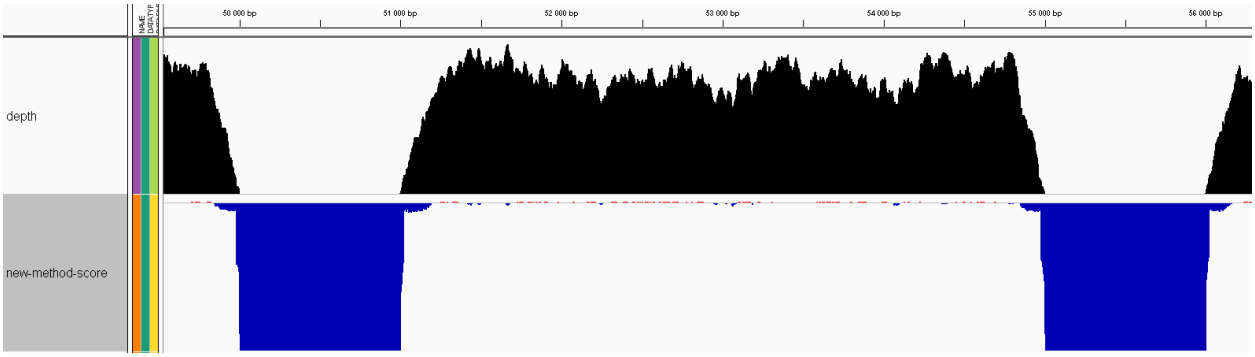

Рис. 3.1. E.coli, позиции 49526-56435 в IGV. Анализ ошибок замещения в позициях 50000-51000 и 55000-56000.

- 1000 ошибок вставки и удаления по 100 нуклеотидов. Все ошибки вставки были найдены. Ошибки удаления слабо отражаются распределением глубин, резкое падение значения глубины покрытия наблюдается только в той позиции, откуда участок был удалён. Резкое падение значения *depth-score* было зафиксировано как раз в таких позициях.

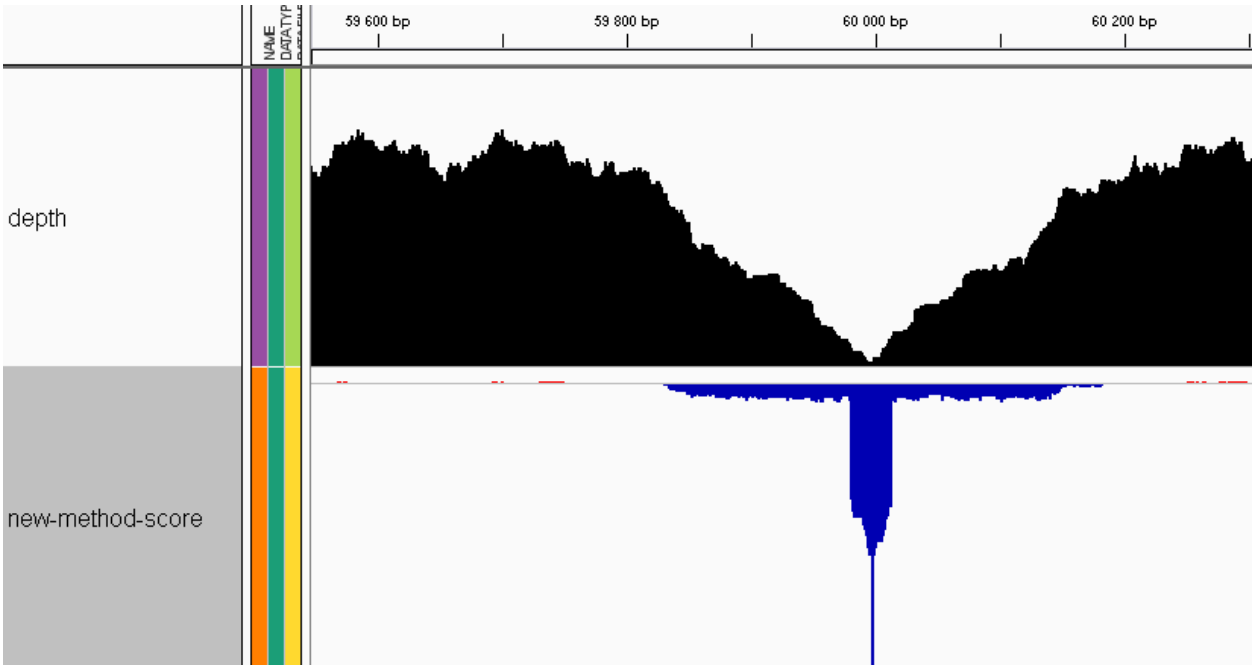

Рис. 3.2. E.coli, позиции 59546-60300 в IGV. Анализ ошибок удаления в позициях 60000-60100.

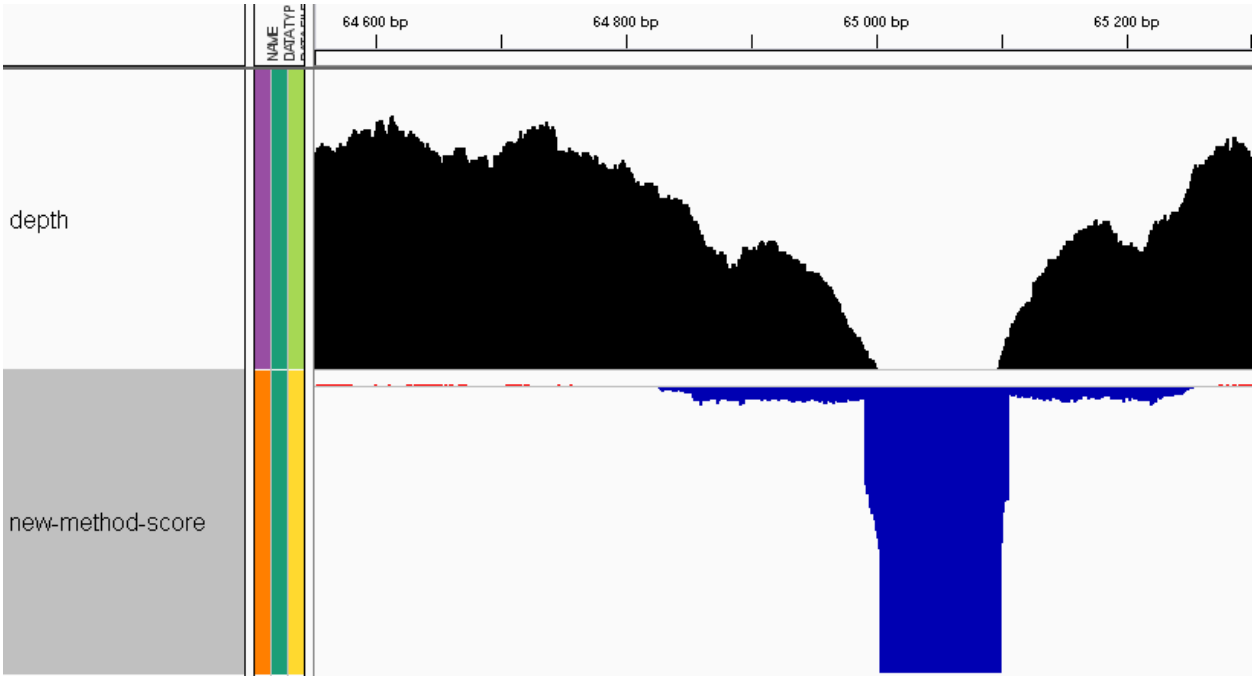

Рис. 3.3. E.coli, позиции 64551-65300 в IGV. Анализ ошибок вставки в позициях 65000-65100.

- 1000 ошибок копирования участков генома длиной 100 нуклеотидов. Ошибки копирования также слабо отражаются распределением глубин покрытий, наблюдается резкое падение глубины в позиции, где заканчивается скопированный участок, *depth-score* так же резко па-

дает в данных позициях.

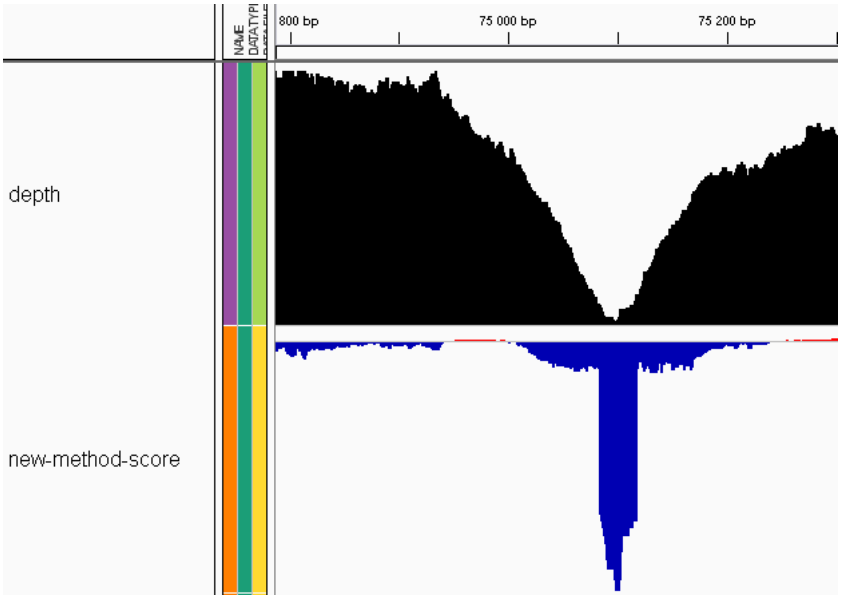

Рис. 3.4. E.coli, позиции 74787-75300 в IGV. Анализ ошибок копирования в позициях 75000-75100.

- 1000 ошибок разворота участков генома длиной 100 нуклеотидов. Все такие ошибки были найдены.

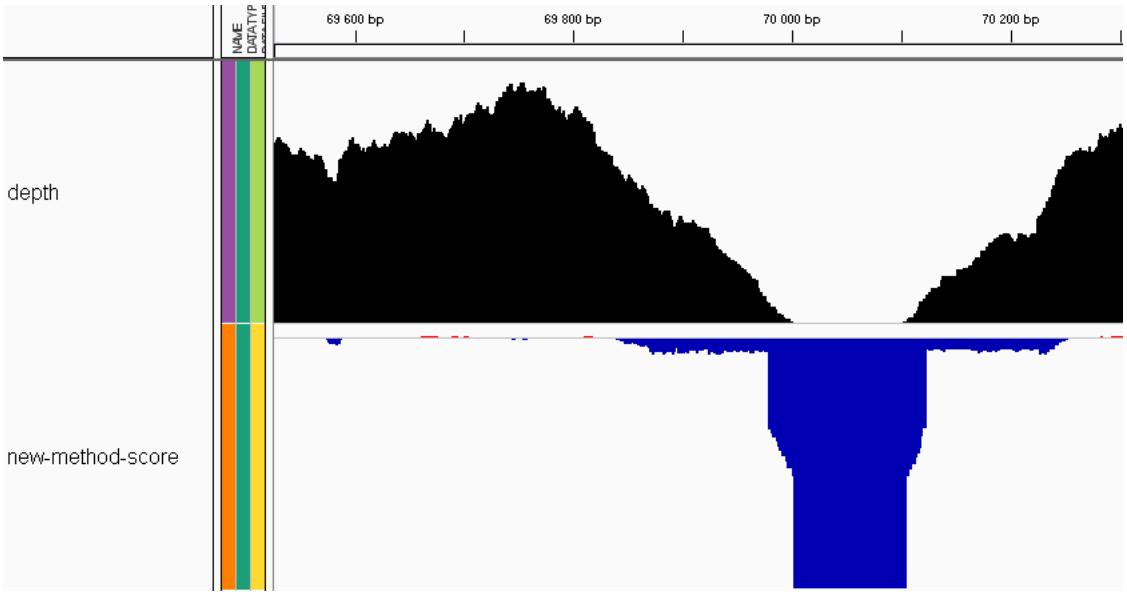

Рис. 3.5. E.coli, позиции 69526-70300 в IGV. Анализ ошибок разворота в позициях 70000-70100.

3.3.2. Ложные срабатывания

Во время выполнения экспериментов на синтетической сборке и последующего сравнения у ALE был выявлен существенный недостаток: при

резком возрастании глубины покрытия на определённых участках ALE определяет эти участки как ошибочные (*ale-depth-score* устанавливается на нижней границе -120.0, т.е. считает, что глубина в данной позиции неверная). При этом установлено, что данные участки не являются повторами. Метод предложенный в данной работе избегает подобных ситуаций. *depth-score* в таких позицияю существенно снижается, однако не определяет глубины в этих позициях как ошибочные.

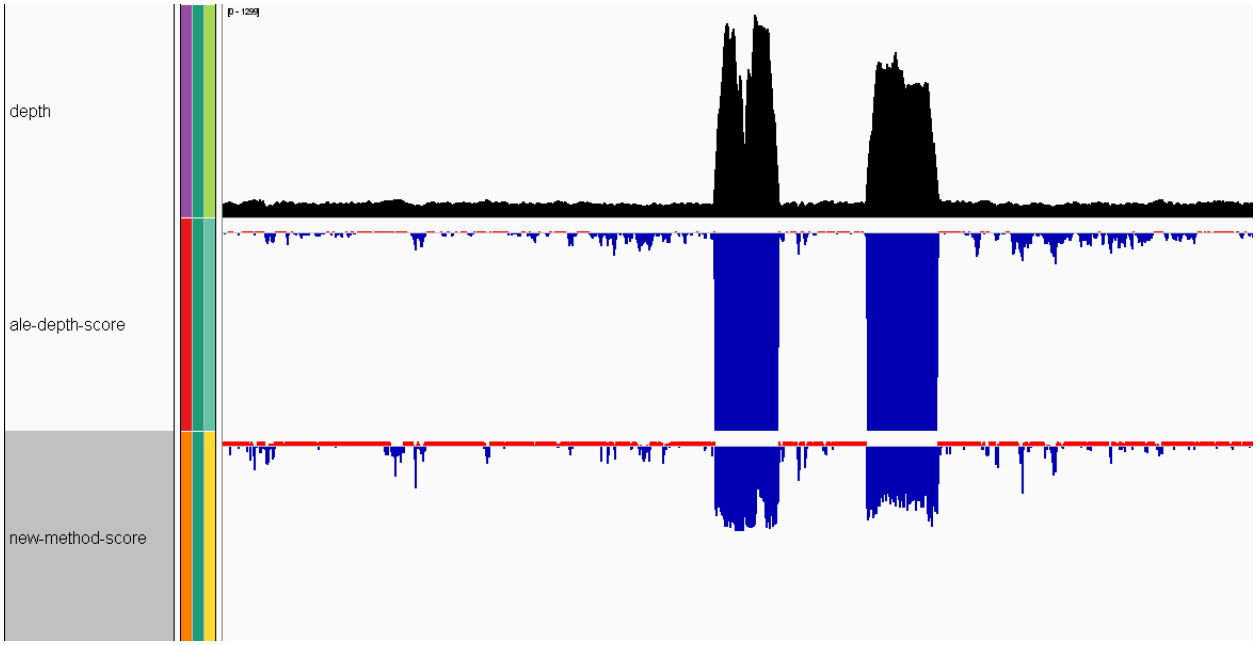

Рис. 3.6. E.coli позиции 2147177–2178816 в IGV. Сравнение ALE и нового метода на участках с резким возрастанием глубины покрытия.

3.3.3. Выводы

Результаты тестов на искусственной сборке с ошибками, показали, что предложенный метод может выявлять все типы ошибок, внесённые в сборку. Более того, в ходе выполнения экспериментов было установлено, что оценка качества изменяется обратно пропорционально количеству внесённых ошибок разных типов: чем больше ошибок вносится в сборку, тем меньше становится *depth-score*. Это говорит о том, что предложенный метод корректно отображает качество сборки.

3.4. Тесты на реальных данных

Для тестирования на реальных данных были рассмотрены несколько сборок *Staphylococcus aureus*, *Rhodobacter sphaeroides* и *Homo sapies Chromosome 14* из базы геномных сборок *GAGE*[20] а также эталонные сборки *Staphylococcus aureus NC_010079* [21], *Rhodobacter sphaeroides CP000143* [22] и *Homo sapies Chromosome 14 NC_000014.8* [23] соответственно. Далее, при помощи стандартных инструментов картирования было проведено сравнение каждой сборки с соответствующим эталоном с целью выявления позиций, где допущены в сборке.

Далее на каждой сборке был запущен ALE и предложенный в данной работе метод.

3.4.1. Выявление ошибок

Ошибки определялись по пороговому значению, ниже которого опускается *depth-score* и *ale-depth-score*. Далее определялись позиции, в которых произошло падение того или иного *srore* и сопоставлялись с данными об ошибках, полученных при помощи инструментов картирования.

Известное количество ошибок в таблицах – это достоверное количество ошибок вставки/удаления, копирования и замещения длинной более пяти нуклеотидов [24]. Столбец true negative – это число ошибок, которые реально существуют в сборке и были найдены тем или иным методом. Столбец false negative – это число ошибок, которых реально не существует в сборке и метод зафиксировал ошибку в неправильном месте.

3.4.1.1. *S. aureus*

Таблица 3.1. Анализ работы ALE и предложенного метода на сбоках бактерии *S. aureus*

Сборщик	Известное количество ошибок	ALE		Предложенный метод	
		true negative	false negative	true negative	false negative
ABySS	19	19	3	19	0
Allpaths-LG	20	18	4	19	0
SGA	10	9	2	10	0
Velvet	42	40	4	41	1
MSR-CA	34	32	6	33	0

3.4.1.2. *R. sphaeroides*

Таблица 3.2. Анализ работы ALE и предложенного метода на сбоках *R. sphaeroides*

Сборщик	Известное количество ошибок	ALE		Предложенный метод	
		true negative	false negative	true negative	false negative
ABySS	76	75	2	76	0
Allpaths-LG	49	45	5	48	0
SGA	12	12	1	12	0
Velvet	47	45	4	47	1
MSR-CA	52	48	6	50	0

3.4.1.3. *H. sapies Chr. 14*

Таблица 3.3. Анализ работы ALE и предложенного метода на сбоках *H. sapies Chr. 14*

Сборщик	Известное количество ошибок	ALE		Предложенный метод	
		true negative	false negative	true negative	false negative
ABySS	704	691	15	701	10
Allpaths-LG	2760	2584	352	2700	20
SGA	981	913	27	975	0
Velvet	4910	4567	560	4879	107
MSR-CA	5550	5326	427	5516	41

3.4.2. Выводы

Результаты показали, что предложенный метод точнее ALE, обнаруживает больше реально существующих ошибок, чем ALE, а также совершает меньше ложных срабатываний.

3.5. Выводы по главе 3

В данной главе был протестирован предложенный в работе метод на синтетических и на реальных данных.

Тесты на синтетических данных выявили, что предложенный метод может выявлять все типы ошибок, внесённые в сборку. Кроме того оценка качества сборки корректна, послокольку меняется пропорционально количеству внесённых в сборку ошибок.

Тесты на реальных данных показали, что предложенный метод работает точнее ALE: находит больше реально существующих в сборке ошибок, показывает меньше ложных срабатываний.

Заключение

Было разработано и реализовано улучшение одного из методов оценки качества сборки генома на основе принципа максимального правдоподобия. Входными данными являются чтения и сборка генома, которую требуется оценить. Результатом является общая оценка сборки, которая представляет из себя логарифм вероятности того, что сборка является правильной геномной последовательностью при заданном наборе чтений. Также вычисляется оценка качества для каждой позиции, которая разбивается на три составляющих: оценка содержимого чтений, оценка расстояния между парными чтениями и оценка глубины покрытия.

В работе было предложено оценивать глубину покрытия на базе эмпирического распределения, поскольку предложенная ранее оценка на базе отрицательного биномиального распределения не точна и была более затратна по вычислениям. Строится сто эмпирических распределений. Каждое распределение вычисляется для позиций с определённым GC-контентом – средним процентом содержания гуанина и цитозина по всем отрезкам длины n покрывающим определённую позицию, где n – это средняя длина чтения. Для более эффективного обнаружения ошибок во время подсчёта распределений было предложено не учитывать не покрытые чтениями позиции.

Предложенный улучшенный метод был успешно реализован. Тестирование проводилось на синтетически сгенерированных чтениях и сборке *Escherichia coli*, а также на реальных чтениях и сборках *Staphylococcus aureus*, *Rhodobacter sphaeroides* и *Homo sapies Chromosome 14* из базы данных *GAGE*. Анализ результатов сравнения явно указывает на то, что улучшенный метод превосходит в точности существующие метрики по количеству правильно найденных ошибок в сборке и по количеству найденных неправильно.

Предложенный в работе улучшенный метод рекомендуется приме-

нять для оценки результатов работы как существующих, так и перспективных сборщиков.

Список литературы

1. Сборка генома: мифы и реальность. http://genome.ifmo.ru/ru/node/46.
2. Bioinformatics – Wikipedia. http://en.wikipedia.org/wiki/Bioinformatics.
3. Talking glossary of genetic terms: genome. National Human Genome Research Institute. http://www.genome.gov/Glossary/.
4. *Alberts B.*, *Johson A.*, *Lewis J.*, *Raff M.*, *Roberts K.* Molecular Biology of the Cell. Fourth Edition. New York: Garland Science, 2002. ISBN: 0815332181. http://www.ncbi.nlm.nih.gov/books/NBK21054/.
5. *Watson J.*, *Crick F.* Molecular structure of nucleic acids; a structure of deoxyribose nucleic acid // Nature. 1953. №171.
6. *S. A.* Shotgun DNA sequencing using cloned DNase I-generated fragments // Nucleic Acids Res. 1981. №9(13). С. 3015–3027.
7. N50 Statistics – Wikipedia. http://en.wikipedia.org/wiki/Bioinformatics.
8. *Salzberg S.*, *Phillippy A.*, *Zimin A.*, *Puiu D.*, *Magoc T.*, *Koren S.*, *Treangen T.*, *Schatz M.*, *Delcher A.*, *Roberts M.*, *Marcais G.*, *Pop M.*, *Yorke J.* GAGE: a critical evaluation of genome assemblies and assembly algorithms // Genome Research. 2012. http://www.ncbi.nlm.nih.gov/pubmed/22147368.
9. *Atif R.*, *Lior P.* CGAL: computing genome assembly likelihoods // Genome Biology. 2013. №1. http://genomebiology.com/2013/14/1/R8/abstract.
10. *Mohammadreza G.*, *Hill C. M.*, *Irina A.* De novo likelihoodbased measures for comparing genome assemblies // BMC Research Notes. 2013. №6. http://www.biomedcentral.com/1756-0500/6/334.
11. *Clark S. C.*, *Egan R.*, *Frazier P. I.* ALE: a generic assembly likelihood evaluation framework for assessing the accuracy of genome // Bioinformatics. 2013. DOI: 10.1093/bioinformatics/bts723. http://www.ncbi.nlm.nih.gov/pubmed/23303509.
12. Negative binomial distribution – Wikipedia. http://en.wikipedia.org/wiki/Negative_binomial_distribution.
13. *Чернова Н.* Математическая статистика. Новосибирск, 2007. ISBN: 9785943565236. http://www.nsu.ru/mmf/tvims/chernova/ms/ms_nsu07.pdf.
14. p-value – Wikipedia. http://en.wikipedia.org/wiki/P-value.
15. Maximum-likelihood Fitting of Univariate Distributions. http://stat.ethz.ch/R-manual/R-patched/library/MASS/html/fitdistr.html.
16. Ion semiconductor sequencing – Wikipedia. http://en.wikipedia.org/wiki/Ion_semiconductor_sequencing.
17. Sequencing Platform Comparison. http://www.illumina.com/systems/sequencing-platform-comparison.ilmn.
18. Integrative genome viewer. https://www.broadinstitute.org/igv/home.
19. *Durfee T.*, *Nelson R.*, *Blattner F. R.* The complete genome sequence of Escherichia coli DH10B: insights into the biology of a laboratory workhorse // Journal of Bacteriology. 2008. DOI: 10.1128/JB.01695-07. http://www.ncbi.nlm.nih.gov/pubmed/18245285.
20. Genome assembly gold-standart evaluations. http://gage.cbcb.umd.edu/data/index.html.
21. Staphylococcus aureus subsp. aureus $USA300_TCH1516chromosome, completegenome$. http://www.ncbi.nlm.nih.gov/nuccore/161508266.
22. Rhodobacter sphaeroides 2.4.1 chromosome 1, complete sequence. http://www.ncbi.nlm.nih.gov/nuccore/CP000143.2.
23. Homo sapiens chromosome 14, GRCh37.p13 Primary Assembly. http://www.ncbi.nlm.nih.gov/nuccore/224589805.
24. GAGE. Final assembly analysis. http://gage.cbcb.umd.edu/results/index.html.