БИООРГАНИЧЕСКАЯ ХИМИЯ, 2024, том 50, № 3, с. 231–255

ОБЗОРНАЯ СТАТЬЯ

УДК 575.112

БИОИНФОРМАТИЧЕСКИЕ ПОДХОДЫ ДЛЯ ДЕТЕКЦИИ ГИБРИДНЫХ ГЕНОВ И ПРОДУКТОВ *ТРАНС*-СПЛАЙСИНГА[1]

© 2024 г. И. Ю. Мусатов*, **, #, М. И. Сорокин, А. А. Буздин*, ***, ******

** ФГАОУ ВО "Московский физико-технический институт (национальный исследовательский университет)", Россия, 141701 Долгопрудный, Институтский пер., 9*

*** Институт персонализированной онкологии и персонализированного здравоохранения ФГАОУ ВО Первого МГМУ им. И.М. Сеченова Минздрава России (Сеченовский университет), Россия, 119048 Москва, ул. Трубецкая, 8/2*

**** ФГБУН "Институт биоорганической химии им. академиков М.М. Шемякина и Ю.А. Овчинникова" РАН, Россия, 117997 Москва, ул. Миклухо-Маклая, 16/10*

***** ГНЦ РФ ФГБУ "Национальный медицинский исследовательский центр эндокринологии" Минздрава России, Россия, 117292 Москва, ул. Дм. Ульянова, 11*

Поступила в редакцию 23.11.2023 г.
После доработки 11.12.2023 г.
Принята к публикации 14.12.2023 г.

Гибридные гены и транскрипты могут являться маркерами и быть причинами развития опухоли за счет дополнительной функциональности получающихся генных продуктов. Современные алгоритмы и методы высокопроизводительного секвенирования, позволяющие детектировать гибридные гены, развиваются как взаимно дополняющие ключи к разгадке вопроса возникновения и диагностики опухоли, а также к фундаментальному вопросу возникновения гибридов и их влияния на молекулярные процессы. Разработаны десятки алгоритмов для детектирования гибридных генов, отличающиеся скоростью, чувствительностью и специфичностью, а также ориентированные на определенный дизайн эксперимента. В зависимости от длины прочтения (50–300 п.н. – короткие прочтения, 5000–100 000 п.н. – длинные прочтения) существуют три типа алгоритмов: работающие только с короткими, только с длинными прочтениями или совмещающие возможности обоих подходов. Кроме того, сами программы разделяют на: 1) программы-картировщики на геном/транскриптом, которые осуществляют поиск прочтений по ряду признаков, например, поиск таких прочтений, которые картируются одновременно на экзоны разных генов и, тем самым, подтверждают наличие гибрида в образце (STAR-Fusion, Arriba); 2) программы-сборщики генома/транскриптома *de novo* с последующим отбором гибридных транскриптов (Fusion-Bloom); 3) программы, использующие "псевдовыравнивание" (Kallisto&Pizzly), когда реального картирования не происходит, а идет сравнение предварительно вычисленного индекса для подпоследовательности транскрипта с индексом, вычисляемым для подпоследовательности конкретного прочтения. В данном обзоре рассмотрены основные классы имеющихся программных инструментов для детектирования гибридных генов, приведены характеристики этих программ, их достоинства и недостатки. Наиболее ресурсоемкими и медленными на сегодняшний момент по-прежнему остаются алгоритмы, осуществляющие сборку генома, их опережают алгоритмы-картировщики. Наиболее быстрыми и сберегающими компьютерные ресурсы являются алгоритмы, осуществляющие псевдовыравнивание, что снижает качество выравнивания в целом.

Ключевые слова: РНК-секвенирование, гибридные гены, гибридные транскрипты, опухоль, фиксированные формалином парафинизированные образцы ткани, псевдовыравнивание, сборка генома de novo, сборка транскриптома de novo, транс-сплайсинг

DOI: 10.31857/S0132342324030033, **EDN:** OAIWBS

[1] Дополнительная информация для этой статьи доступна по doi 10.31857/S0132342324030033 для авторизованных пользователей.

Сокращения: ММР – максимальный отображаемый префикс; ХМЛ – хронический миелоидный лейкоз; STAR – алгоритм выравнивания сплайсированных транскриптов на эталонный геном/транскриптом; SNP – однонуклеотидный полиморфизм.

Автор для связи: (тел.: +7 (985) 147-97-18; эл. почта: musatov.mailbox@yandex.ru).

СОДЕРЖАНИЕ

1. ВВЕДЕНИЕ

Гибридными (химерными) генами называют гены, образовавшиеся в результате слияния двух ранее независимых генов или их частей. Такие гибриды могут образоваться в результате хромосомных перестроек или *цис*/*транс*-сплайсинга. *Цис*-сплайсинг близлежащих генов (*cis*-SAGe) возникает в результате игнорирования ДНК-зависимой РНК-полимеразой стоп-кодона транскрипции; вместо этого последовательности ДНК двух соседних генов считываются и транскрибируются в гибридный транскрипт мРНК [1–3].

Транс-сплайсинг генов – молекулярный процесс, в результате которого транскрипты двух разных генов соединяются в один. *Транс*-сплайсинг хорошо задокументирован у низших эукариот, однако в ряде исследований такие события были обнаружены и в опухолях человека [3–5].

Еще один процесс, приводящий к образованию гибридов, – соматические хромосомные перестройки (транслокации, инверсии и делеции). Хромосомная транслокация была первым известным механизмом, приведшим к возникновению гибридного гена [6], а именно гена *BCR::ABL1*, известного как ген “филадельфийской хромосомы”, который образуется в результате несбалансированной соматической транслокации участков хромосом 9 и 22 [t(9;22)(q34;q11.2)] [7, 8]. Наличие этого гена приводит к возникновению в клетке постоянно активной тирозинкиназы, что вызывает постоянную пролиферацию такой клетки и ее невосприимчивость к сигналам апоптоза. Эти процессы ускоряют мутагенез и позволяют опухолевым клеткам стать устойчивыми к воздействию лекарственных препаратов [9, 10].

В результате такая перестройка приводит к возникновению ХМЛ (хронического миелолейкоза) и встречается в 90% случаев данного вида миелолейкоза. Такие гибриды оказывают влияние на развитие рака, например, за счет нарушения регуляции некоторого молекулярного пути. В таком случае продукт некоторого гена *A* получает дополнительную регулирующую активность, не свойственную данному гену *A*. Со времени обнаружения первого гибридного гена в ХМЛ было открыто множество гибридов, встречающихся в других типах рака [11–18]. Наличие гибридов может быть, как непосредственным признаком наличия болезни, так и ее причиной, поэтому были созданы препараты, нацеленные на ряд хорошо известных гибридов в различных типах рака [11].

Ввиду того, что гибридные гены могут быть как маркерами, так и потенциальными целями противораковых препаратов, поиск гибридов представляет собой актуальную задачу. Высокопроизводительное секвенирование трансформировало сферу геномики рака, позволив секвенировать целые геномы и транскриптомы злокачественных опухолей человека по сравнительно небольшой стоимости, что привело к экспоненциальному возрастанию объема обрабатываемых данных и разработке различных программных инструментов и алгоритмов для поиска гибридных генов. В настоящем обзоре рассмотрены различные биоинформатические подходы для детекции гибридных генов, успешно применяемые в последние несколько лет, их преимущества, недостатки и перспективы развития данной области исследований.

2. ОСНОВНЫЕ ПОДХОДЫ ДЛЯ ДЕТЕКЦИИ ГИБРИДНЫХ ГЕНОВ

На сегодняшний день разработано большое количество алгоритмов и инструментов, позволяющих детектировать гибридные гены в больших массивах данных, генерируемых методами высокопроизводительного секвенирования. Существующие алгоритмы, дающие наиболее эффективную прогностическую оценку транскриптов как кандидатов в новые ранее не открытые гибриды, часто используют данные секвенирования как генома, так и транскриптома для образца. Такие алгоритмы называются метаалгоритмы.

Поиск гибридов и продуктов *транс*-сплайсинга в таких алгоритмах реализован двумя способами:

1) выравниванием прочтения на эталонную последовательность, например, чтобы идентифицировать последовательности, которые одновременно картируются в несколько мест (multimapping reads) [19–21] (рис. 1);

2) сборкой прочтений в более длинные транскрипты *de novo* за счет перекрывающихся последовательностей длиной *k* нуклеотидов (*k*-меры) (рис. 2). Такая сборка может осуществляться либо из *k*-меров прочтений (рис. 3), либо для этого могут быть использованы транскрипты эталонного транскриптома. Тогда из транскриптов эталонного транскриптома составляется структура из подпоследовательностей нуклеотидов длиной *k*, именуемая граф де Брюйна из *k*-меров транскриптома/генома [21, 22]. Если *k*-мер имеет сходную подпоследовательность с $(k + 1)$-мером, т.е. является префиксом или суффиксом $(k + 1)$-мера, то такой *k*-мер будет иметь связь в графе де Брюйна с $(k + 1)$-мером эталонного транскриптома (см. Дополнительные материалы к статье). В результате сборки получится набор транскриптов, который будет либо соответствовать эталонным транскриптам из эталонного транскриптома, либо будет собран гибридный транскрипт, не присутствующий в эталонном транскриптоме.

Таким образом, *k*-меры, которые имеют префикс и суффикс, соответствующие разным генам, при сравнении с эталонным транскриптомом идентифицируют как потенциальные гибриды – такие *k*-меры (собранные алгоритмом в гипотетические транскрипты) могут соответствовать хромосомным перестройкам.

Собранная длинная последовательность из перекрывающихся подпоследовательностей прочтений называется контигом. Такие контиги затем сравниваются с числом транскрибируемых генов с учетом альтернативного сплайсинга, что позволяет оценить информативность сборки.

Подход, использующий в качестве первого шага выравнивание, более чувствителен с точки зрения числа детектируемых гибридов и продуктов *транс*-сплайсинга.

В то же время сборка прочтений в более длинные транскрипты *de novo* может быть полезна исследователю для поиска новых изоформ гибрида [23], не указанных в различных базах данных, или таких изоформ, для которых:

1) отсутствует аннотация или эталонная последовательность;

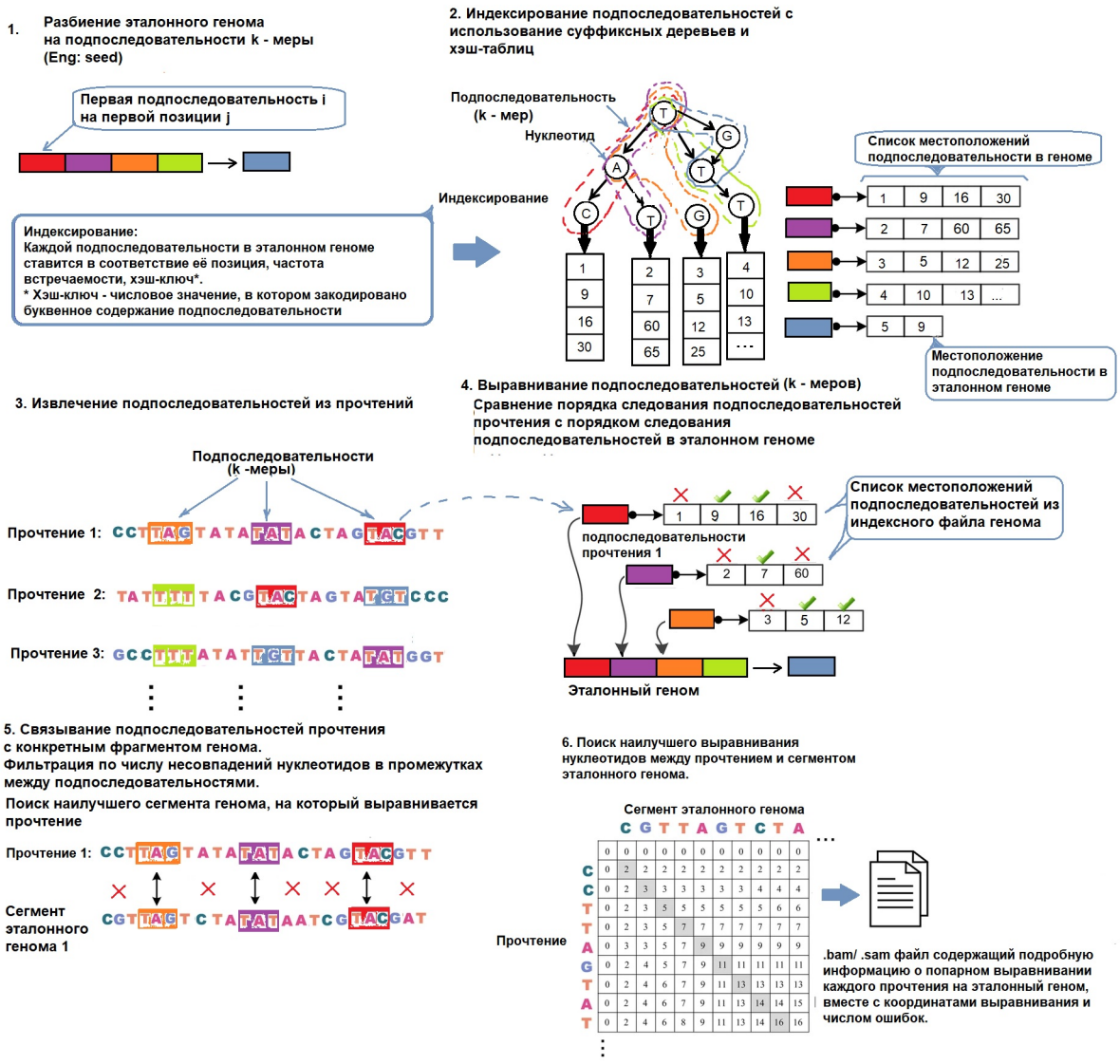

Рис. 1. Основные этапы работы алгоритмов картирования прочтений на геном (транскриптом). Рисунок подготовлен и переработан на базе материалов статьи Alser et al. [63].

2) имеется аннотация для обоих экзонов в отдельности, при этом ни один транскрипт, объединяющий данные экзоны, не аннотирован;

3) в геноме имеются недавно возникшие гены-паралоги, которые производят две изоформы, различающиеся в последовательности [23].

Один из “золотых стандартов”, использующих предварительное выравнивание прочтений с последующей фильтрацией прочтений, – программа картирования STAR с соответствующей ей надстройкой STAR-Fusion [21, 24].

3. ПРОГРАММЫ ДЛЯ ДЕТЕКЦИИ ГИБРИДНЫХ ГЕНОВ, ИСПОЛЬЗУЮЩИЕ ВЫРАВНИВАНИЕ НА ГЕНОМ/ТРАНСКРИПТОМ. STAR И STAR-FUSION

3.1. Программа STAR

Программа STAR (Spliced Transcripts Alignment to the Reference) была разработана для выравнивания несмежных последовательностей непосредственно с эталонным геномом. Алгоритм STAR состоит из нескольких основных этапов: этапа

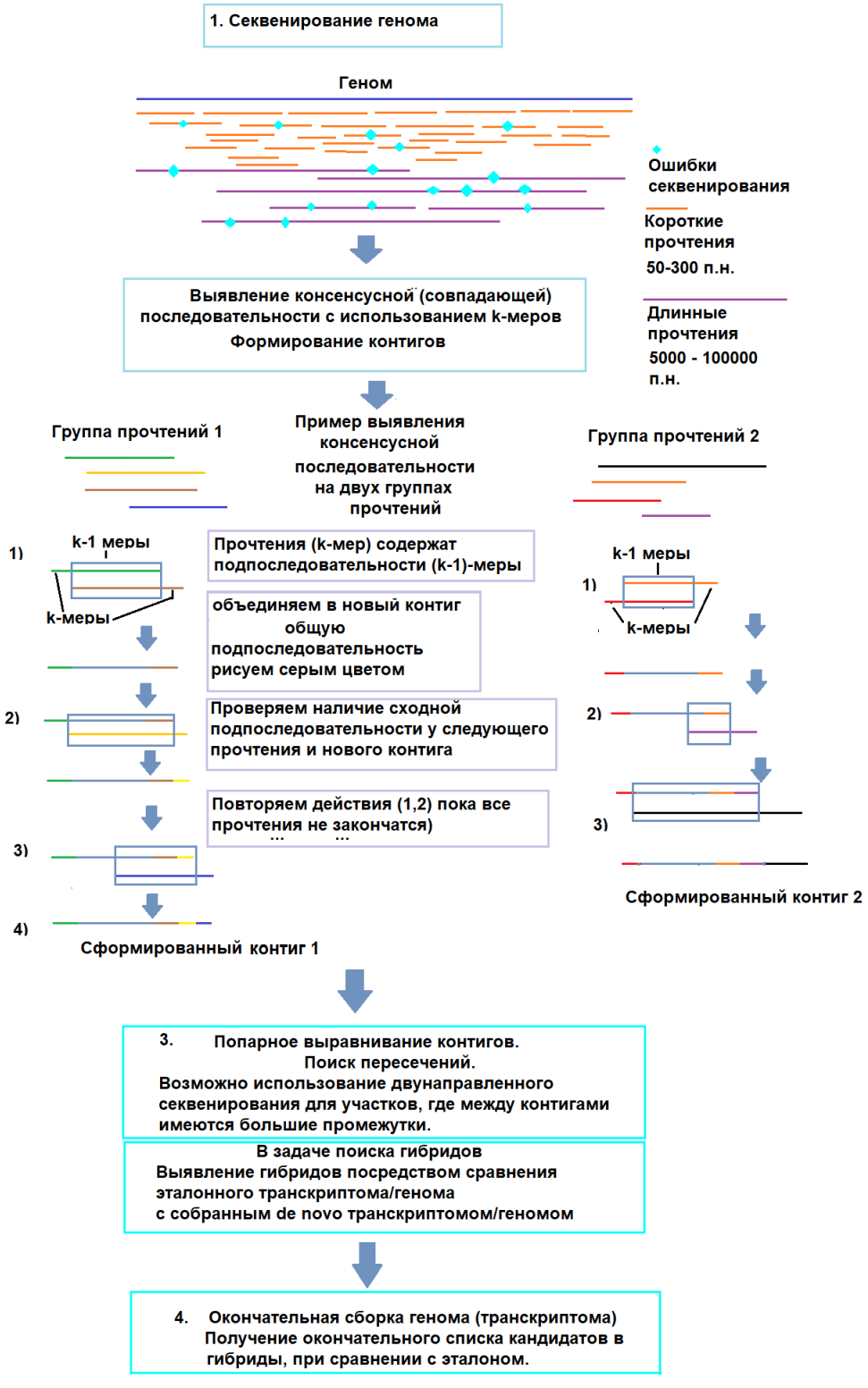

Рис. 2. Основные этапы работы алгоритмов-сборщиков генома (транскриптома) *de novo*.

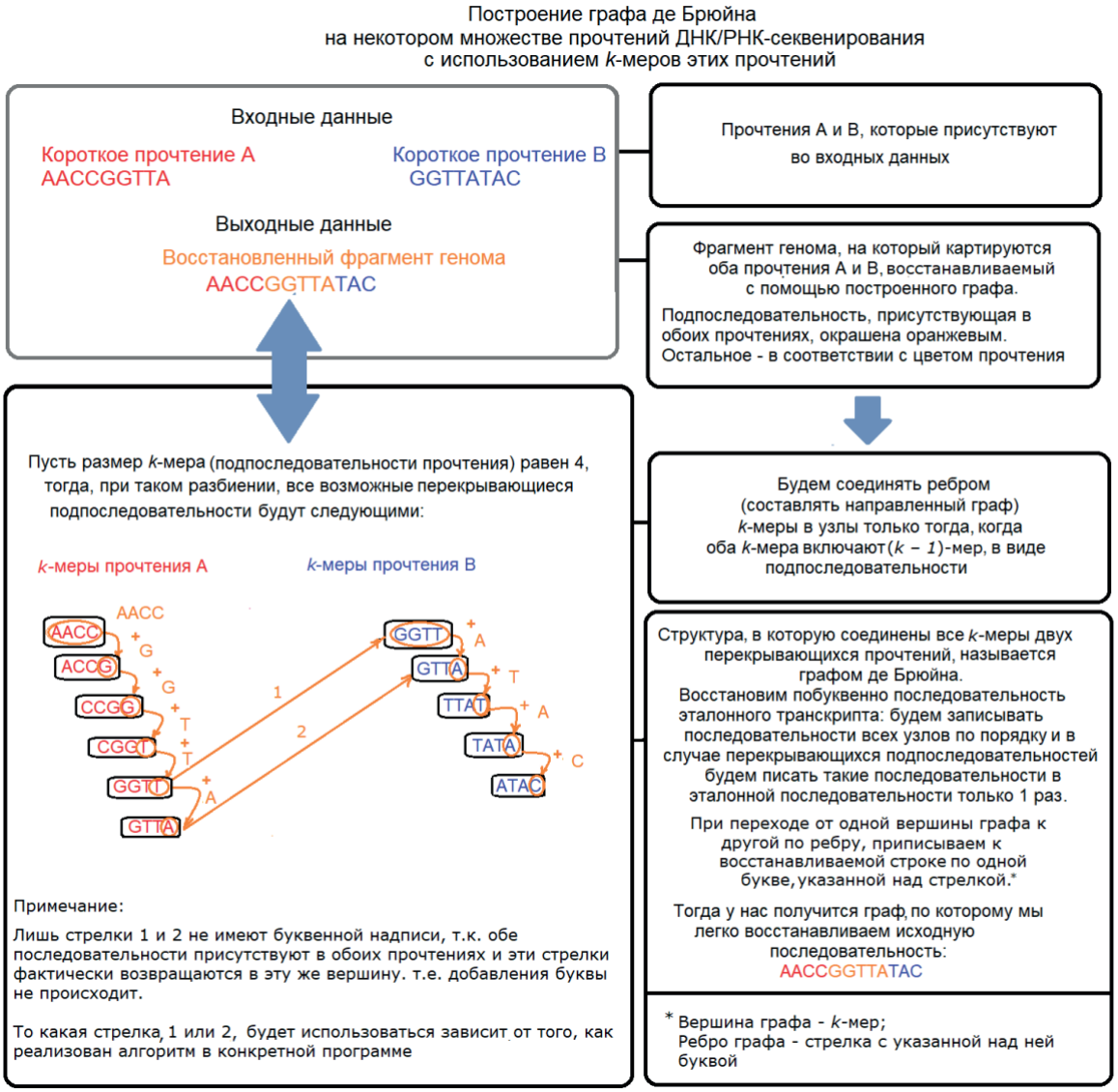

Рис. 3. Построение графа де Брюйна на некотором множестве прочтений ДНК/РНК-секвенирования с использованием *k*-меров (подпоследовательностей) таких прочтений.

поиска начального максимального по длине картируемого отображения – максимального отображаемого префикса (MMP – maximum mappable prefix) – и этапов кластеризации, сшивания и оценки [19, 21].

STAR будет искать самую длинную подпоследовательность прочтения, которая соответствует одному или нескольким местам в эталонном геноме. Эти длинные совпадающие подпоследовательности называются максимальными отображаемыми префиксами (MMP).

Алгоритм картирует прочтение на геном, и если какая-то часть последовательности прочтения не картируется (т.е. во время поиска MMP выравнивание не проходит до конца прочтения), то прочтение разбивается на картированный и некартированный фрагменты. Положение, где прочтение будет разбито на две части, будет в таком случае гипотетической точкой разрыва (breakpoint), а найденный MMP относится к конкретному гену в геноме. Далее для некартированной части прочтения (суффикса) будет произведена аналогичная

процедура [21]. Первый этап – поиск ММР – реализован как поиск в бинарной строке для несжатых суффиксных массивов, что выполняется быстрее (в отличие от стандартного поиска полноразмерного совпадения) и работает как для самого прочтения, так и для дополняющей его обратно-комплементарной последовательности.

Таким образом, можно найти не только последовательности, которые при прочих равных и так успешно выравнивались на геном, но и гибридные прочтения, состоящие из экзонов двух разных генов.

Если в прочтении имеются несовпадения, то берется следующий нуклеотид (буква), и если число ошибок при увеличении размера суффикса последовательно растет, то расширенная часть прочтения обрезается.

На втором этапе происходит окончательное выравнивание последовательности с учетом того, какой ММР был выбран (так называемый "якорный ММР" с наилучшим выравниванием). Далее последовательность ММР сшивается с оставшимися выровненными частями прочтений, причем так, чтобы все выравнивания попали в так называемое окно выравнивания. Такое окно формируется следующим образом: первоначально геном разбивается на фрагменты одинаковой длины, начальные координаты которых совпадают с координатами якорных ММР. Затем все якорные ММР, которые находятся на расстоянии i букв между друг другом (такое расстояние в i букв указывается пользователем в качестве размера интрона), составят окно выравнивания. После этого начинается сшивание фрагментов в соответствии с наилучшими значениями выравнивания для оставшихся ММР с помощью алгоритма динамического программирования, при этом алгоритм допускает произвольное число несовпадений, но только одну делецию или одну инсерцию в ходе сшивания ММР.

Если выравнивание происходит в пределах одного геномного окна и не охватывает прочтение целиком, STAR попытается найти два или более окон, охватывающих всю последовательность, в результате чего образуется химерное выравнивание с различными частями прочтения, сопоставленными с дистальными геномными локусами или разными хромосомами [21].

При сшивании двух ММР используется схема оценки локального выравнивания с учетом штрафов за несовпадения в последовательности, делеции и инсерции, а также оценки за верное выравнивание, задаваемые пользователем и изменяемые в зависимости от длины прочтения. Это позволяет также выравнивать прочтения, полученные методами секвенирования третьего поколения – РНК-секвенирования длинными прочтениями [19, 25].

3.2. STAR-Fusion

STAR-Fusion – надстройка над основной программой картирования STAR-Aligner и компонент инструмента Trinity Cancer Transcriptome Analysis Toolkit (CTAT), которая в ходе картирования прочтений на геном возвращает наиболее вероятные химерные прочтения. Фильтрация результатов выдачи картировщика STAR-Aligner проводится этой надстройкой в два этапа: базовая и продвинутая фильтрация [24].

На первом этапе исключаются все прочтения, которые попадают на участки, имеющие попарное сходство нуклеотидной последовательности, посредством поиска по базе эталонных транскриптов комплементарной ДНК, проиндексированных алгоритмом BLASTN [26]. Также исключаются дубликаты парноконцевых прочтений.

После этого отбираются химерные гибриды так, чтобы направление транскрипции для экзонов, составляющих гибрид, совпадало, и точка разрыва поддерживалась как минимум двумя прочтениями. Если точка разрыва соответствует сайтам сплайсинга экзонов, то как минимум одно из прочтений должно поддерживать данную точку разрыва. В случае же несоответствия точки разрыва эталонным сайтам сплайсинга экзонов требуются как минимум три прочтения для того, чтобы подтвердить такую точку разрыва. При этом если точка разрыва не соответствует эталонным сайтам сплайсинга, то для прочтений, перекрывающих точку разрыва, требуется картирование как минимум 25 нуклеотидов по обе стороны от точки разрыва [24].

4. ФИЛЬТР ПАРАЛОГОВ

На более глубоком уровне фильтрации анализируются уже индивидуальные характеристики каждого гибрида. Так, например, из выдачи исклю-

чаются прочтения, которые картируются на гибриды, состоящие из генов, являющихся паралогами друг друга. Также исключаются случаи, если два гибрида содержат гены, являющиеся паралогами друг друга, причем в большинстве своем поддерживается только один из паралогов [21].

Пусть, например, в гибридах "*Ген_1_Ген_2*" и "*Ген_1_Ген_3*" *Ген_2* и *Ген_3* являются паралогами друг друга; и пусть прочтений и прочих свидетельств, поддерживающих "*Ген_1_Ген_2*", больше, чем прочтений, поддерживающих "*Ген_1_Ген_3*". Тогда в качестве финального результата будет выбран тот гибрид, который имеет большее количество прочтений, относящихся к данному гибриду.

5. ФИЛЬТР ПРОЧТЕНИЙ, КАРТИРУЮЩИХСЯ ОДНОВРЕМЕННО В НЕСКОЛЬКО МЕСТ

Если имеется несколько гибридных генов, состоящих из некоторого определенного гена и других генов с разным числом прочтений, относящихся к каждому гибриду, т.е. пусть существуют гибриды "*Ген_1_Ген_2*", "*Ген_1_Ген_3*", "*Ген_1_Ген_4*", и уровень поддержки (число прочтений, относящихся к данному гибриду) одного из них в 20 раз превышает уровень поддержки остальных. Тогда останется только гибрид с максимальным уровнем поддержки, а прочие гибриды будут исключены из рассмотрения. При этом если у *Гена_1* имеется >10 вариантов генов-партнеров, образующих гибрид с этим геном, то все гибриды, включающие *Ген А*, исключаются из рассмотрения [21].

Вследствие высокого полиморфизма, гомологии и ошибок секвенирования некоторые прочтения невозможно однозначно сопоставить с местом их происхождения в эталонном геноме. Поэтому из рассмотрения исключаются гибридные гены, имеющие в своем составе высокополиморфные гены комплекса гистосовместимости или митохондриальные гены, а также гибриды, присутствующие в нормальных тканях [21, 27].

Кроме того, для каждого потенциального кандидата в гибридные гены проводится оценка на основе экспрессии и глубины секвенирования. Кандидаты, отобранные как потенциальные гибриды и имеющие менее чем одно прочтение на 10 млн прочтений, отбрасываются как кандидаты с недостаточным уровнем поддержки [21].

6. STAR И ARRIBA

Набор фильтров Arriba основан на программе STAR-Aligner и представляет собой его дополнение (по аналогии со STAR-Fusion). Arriba позволяет использовать результаты выравнивания РНК-секвенирования (RNASeq) прочтений с помощью программы STAR и предложенные ей химерные выравнивания для дальнейшего анализа и отбора прочтений по собственным критериям Arriba.

Авторы обращают внимание, что программа STAR нацелена на поиск двух типов прочтений, подтверждающих гибридные гены:

1) прочтений, картированных на точку слияния/разрыва двух генов (spanning reads);

2) разделенных прочтений (split reads), которые картируются какой-то частью на один из генов [28].

В дополнение к прочтениям, которые находит STAR и предоставляет в качестве кандидатов в гибридные гены, Arriba также нацелена на поиск слияний, возникающих в результате фокальных делеций. При возникновении фокальных делеций 5′-конец вышестоящего гена сближается с 3′-концом нижележащего гена. STAR осуществляет выравнивание, при котором гибридные гены были бы соединены путем сплайсинга, строго по границам гена. Arriba же проверяет выдаваемый STAR список и дополняет его такими прочтениями, которые выходят за границы аннотированного гена, чтобы не пропустить гибриды, возникшие из-за фокальной делеции [28].

В дополнение к фильтрам, имеющимся в программе STAR-Aligner, в Arriba применяются дополнительные фильтры, отбрасывающие следующие прочтения:

1) прочтения, содержащие большое количество одинаковых нуклеотидов;

2) тандемные повторы;

3) прочтения, содержащие большое количество несовпадений между прочтением и эталоном.

Также отбрасываются прочтения из списка прочтений, соответствующих транскриптам, часто встречающимся в доброкачественных опухолях.

Список транскриптов, часто встречающихся в доброкачественных опухолях, был получен посредством машинного обучения на образцах РНК-секвенирования из пяти различных международных проектов:

1) Атласа белков человека (Human Protein Atlas), цель которого – выяснение происхождения всех белков человека с использованием современных методов [29];

2) Illumina Human BodyMap2, главная задача которого – определение профиля транскрипции для 16 типов тканей человека с использованием высокопроизводительного секвенирования [30, 31];

3) проекта ENCODE (Encyclopaedia of DNA Elements) [32], исследующего функциональные элементы генома;

4) проекта Roadmap Epigenomics [33], цель которого – создание эталонной карты эпигенетических меток;

5) проекта NCT/DKTK MASTER [34], нацеленного на исследование образцов редких типов рака, а также злокачественных образований у пациентов, у которых рак был обнаружен в молодом возрасте.

Вдобавок к вышеописанным фильтрам ложноположительных результатов Arriba имеет фильтры ложноотрицательных результатов. Такие фильтры возвращают в общий перечень кандидатов в гибридные гены кандидаты и соответствующие им прочтения:

1) прочтения, указанные в пользовательском списке известных или часто встречающихся гибридов;

2) прочтения, картированные на известные сайты сплайсинга, если такие прочтения были первоначально отброшены [21, 28].

Также Arriba имеет собственную статистическую модель, применяемую для фильтрации кандидатов в гибридные гены [28].

Здесь и далее для уточнения терминологии дадим следующее определение: будем называть точкой разрыва (breakpoint) место в нуклеотидной последовательности и его координаты, в котором произошел разрыв хромосомы, например, в результате транслокации, с последующим встраиванием участка другого гена (слиянием), в результате чего получился гибридный ген, состоящий из двух различных генов.

Разработчики Arriba отмечают, что уровень шума (число ложноположительных кандидатов в гибриды) коррелирует с глубиной секвенирования, расстояниями между точками разрыва, отношением числа интронов к числу экзонов, а также отношением между числом инверсий и прочтений с одинаковой нуклеотидной последовательностью. Коррелирующие значения влияют на базовый уровень шума, что, в свою очередь снижает точность предсказания, поэтому величина их значения выступает как штраф, увеличивая потенциальную ошибку. В модели, используемой программой Arriba, ожидаемый уровень (*e_value*) шума оценивается как произведение следующих множителей [28]:

$$e_{value} = base\_level\_background\_noise \times depth\_penalty \times \times distance\_penalty \times inversions\_to\_duplications\_ratio \times \times intron\_to\_exons\_ratio, \quad (1)$$

где *e_value* – ожидаемый уровень шума; *base_level_background_noise* – уровень базового фонового шума; *depth_penalty* – штраф за глубину секвенирования; *distance_penalty* – штраф за расстояния между точками разрыва; *inversions_to_duplications_ratio* – отношение между числом инверсий и прочтений с одинаковой нуклеотидной последовательностью; *intron_to_exons_ratio* – отношение числа интронов к числу экзонов.

В данном инструменте предполагается, что отношение числа прочтений, подтверждающих гибрид (сигнал), к общему числу прочтений (шум) распределено полиномиально. В итоговой выдаче результатов сообщается только о тех кандидатах, которые содержат число прочтений выше уровня шума. Кандидаты в гибриды, имеющие уровень *e_value* (ожидаемого фонового шума) меньше установленного пользователем порога (или значения по умолчанию), будут отобраны в качестве итогового результата.

Отметим, что авторы утверждают, что предположения, на основе которых был сделан вывод о том, что данная характеристика имеет полиномиальную зависимость, основаны на эмпирических свидетельствах [28], предполагая наличие нелинейной зависимости между числом прочтений, подтверждающих гибрид, и общим числом прочтений.

Уровень базового фонового шума для каждого гена рассчитывается по формуле (2) [28]:

$$base\_level\_background\_noise = \frac{total\_candidates\_of\_gene}{sum\_of\_exon\_lengths\_of\_gene} \times \times (supporting\_reads - SHIFT_{noise})^{SLOPE_{noise}} \times INTERCEPT_{noise}, \quad (2)$$

где *base_level_background_noise* – уровень базового фонового шума; *total_candidates_of_gene* – общее число кандидатов в гибридные гены для данного гена; *sum_of_exon_lengths_of_gene* – сумма длин экзонов данного гена; *supporting_reads* – число прочтений, поддерживающих гибридный ген; $SHIFT_{noise}$ – смещение$_{шум}$ = –0.73; $SLOPE_{noise}$ – наклон$_{шум}$ = –2.28; $INTERCEPT_{noise}$ – пересечение с осью $y_{шум} = 10^{-1.75}$. Причем смещение$_{шум}$, наклон$_{шум}$, пересе-

чение с осью $y_{шум}$ – эмпирически определенные константы, которые были получены на образцах секвенирования РНК из когорты NCT/DKTK MASTER и оказались достаточно стабильными в различных наборах данных [28], а именно:

$$depth\_penalty = SLOPE_{depth} \times (SLOPE\_MODIFIER)^{supporting\_reads} \times \times mapped\_read, \quad (3)$$

где *depth_penalty* – штраф за глубину секвенирования; $SLOPE_{depth}$ – наклон$_{глубина}$ = 2×10^{-11}; *SLOPE_MODIFIER* – модификатор наклона = 0.02; *mapped_reads* – число картированных прочтений.

Отметим, что штраф за расстояние между точками разрыва *distance_penalty* (4) применяется для точек, находящихся на расстоянии менее $4 \times \times 10^5$ п.н. [28]. Штраф увеличивается с уменьшением расстояния между точками разрыва.

$$distance\_penalty = (distance)^{SLOPE_{distance}} \times \times INTERCEPT_{distance}, \quad (4)$$

где *distance_penalty* – штраф за расстояния между точками разрыва; *distance* – расстояние между точками разрыва; $SLOPE_{distance}$ – наклон$_{расстояние}$; $INTERCEPT_{distance}$ – пересечение с осью $y_{расстояние}$.

В зависимости от того, находятся ли точки разрыва ближе друг к другу или на расстоянии >400 п.н., используются разные коэффициенты:

1) если расстояние между точками разрыва <400 нуклеотидов, то $SLOPE_{distance} = -4.58$ и $INTERCEPT_{distance} = 8.27 \times 10^{10}$;

2) если расстояние между точками разрыва >400 нуклеотидов, то $SLOPE_{distance} = -1.53$ и $INTERCEPT_{distance} = 3.73 \times 10^8$.

Библиотеки, ориентированные согласно направлению цепи (stranded RNAseq), могут иметь дубликаты [28]. Чтобы учесть влияние протокола подготовки образцов РНК-секвенирования (standed RNAseq/non-stranded RNAseq), Arriba считает соотношение между инверсиями и дубликатами. И за наличие инвертированных прочтений и прочтений с повторяющимися последовательностями в конечную оценку шума для кандидата в гибридные гены входит штраф (формулы (5) и (6) [28]), называемый соотношением между инверсиями и прочтениями с повторяющимися последовательностями. Такое соотношение рассчитывается по следующим формулам:

1) если для данного кандидата в гибридный ген в прочтениях присутствует инверсия, то для расчета коэффициента используют формулу:

$$inversions\_to\_duplications\_ratio = \frac{total\_inversions}{total\_candidates}, \quad (5)$$

где *inversions_to_duplications_ratio* – соотношение между всеми прочтениями, содержащими инверсии, и общим числом прочтений для данного кандидата в гибридный ген; *total_inversions* – число всех инвертированных прочтений для данного кандидата в гибридный ген; *total_candidates* – число всех прочтений для данного кандидата в гибридный ген;

2) если для данного кандидата в гибридный ген присутствуют прочтения, содержащие повторяющиеся последовательности, то коэффициент рассчитывают по формуле:

$$inversions\_to\_duplications\_ratio = \frac{total\_duplicates}{total\_candidates}, \quad (6)$$

где *inversions_to_duplications_ratio* – соотношение между прочтениями с повторяющейся последовательностью и общим числом прочтений для данного кандидата в гибридный ген; *total_duplicates* – число всех прочтений с повторяющейся последовательностью для данного гена; *total_candidates* – число всех прочтений для данного гена.

Аналогичным образом кандидату в гибриды присваивается штраф (формула (7) [28]) в зависимости от местонахождения точки разрыва: в интроне, в экзоне или в месте сплайсинга. Такой штраф называется соотношением между числом интронов и числом экзонов, и в зависимости от положения точки разрыва данный коэффициент рассчитывают следующим образом [28, 35]:

1) $intron\_to\_exon\_ratio = \frac{total\_intronic\_candidates}{total\_candidates}$, (7)

если точка разрыва (breakpoint) находится в интроне;

2) $intron\_to\_exon\_ratio = \frac{total\_exonic\_candidates}{total\_candidates}$, (7)

если точка разрыва (breakpoint) находится в экзоне;

3) $intron\_to\_exon\_ratio = \frac{total\_spliced\_candidates}{total\_candidates}$, (7)

если точка разрыва (breakpoint) находится в сайте сплайсинга.

В формулах (7) *total_intronic_candidates* – число всех прочтений, содержащих точку разрыва в интроне; *total_exonic_candidates* – число всех прочтений, содержащих точку разрыва в экзоне; *total_spliced_candidates* – число всех прочтений, содержащих точку разрыва в месте сплайсинга; *total_candidates* – число всех прочтений для данного гена.

7. ПРОГРАММЫ ДЛЯ ДЕТЕКЦИИ ГИБРИДНЫХ ГЕНОВ, ИСПОЛЬЗУЮЩИЕ СБОРКУ ТРАНСКРИПТОМА *DE NOVO*. FUSION-BLOOM И RNA-BLOOM

Fusion-Bloom – программа, основанная на сборке транскриптома *de novo*, использует алгоритм RNA-Bloom и программу для поиска различных структурных вариантов PAVFinder для обнаружения гибридных генов [36]. Отметим, что алгоритм RNA-Bloom был задуман как алгоритм сборки прочтений секвенирования РНК одной клетки (single-cell RNAseq). Выполнение данной программы состоит из нескольких этапов [37]. На первом этапе Прочтения РНК-секвенирования собирают в более длинные последовательности (контиги) за счет наборов перекрывающихся сегментов РНК с использованием алгоритма RNA-Bloom. Собранные таким образом последовательности получаются разной длины, поскольку перекрывающиеся фрагменты имеют разную длину и область перекрытия. Такие контиги фильтруют по длине так, чтобы длина контига была больше длины первого квартиля всей сборки, т.е. для конкретного собираемого транскриптома строится распределение длин контигов. Далее такое распределение разбивают на квартили и проверяют, чтобы длина собранного контига превышала длину контига, соответствующего первому квартилю распределения для всей сборки. На следующем этапе используется несколько алгоритмов выравнивания, чтобы установить в дальнейшем прочтения, поддерживающие гибридные гены. Во-первых, контиги выравниваются на геном с помощью алгоритма GMAP, во-вторых, эти же контиги выравниваются на эталонные транскрипты алгоритмом BWA MEM. Отметим, что эталонные транскрипты были получены с использованием скрипта “extract_transcript_sequence.py” входящего в пакет программ PAVFinder (Post Assembly Variant Finder) [36, 38]. Картирование на геном с последующим картированием на эталонные транскрипты позволяет отделить контиги, которые выравниваются уникально и неуникально на определенные гены и транскрипты как на геномном, так и на транскриптомном уровнях, и, таким образом, выяснить, какие контиги являются гибридами. После этого прочтения РНК-секвенирования выравниваются на отобранные гибридные контиги с помощью алгоритма Minimap2 – для оценки уровня экспрессии предполагаемых гибридных генов и последующей фильтрации гибридных контигов. Затем, руководствуясь результатами трех выравниваний (прочтений на контиги, контигов на геном и контигов на транскриптом), алгоритм PAVFinder составляет список предполагаемых гибридных генов и поддерживающих их прочтений и возвращает пользователю список предполагаемых гибридов в формате BEDPE [36].

Формат BEDPE содержит следующую информацию:

1) координаты генов 1 и 2, составляющих гибрид;

2) ориентация генов;

3) число экзонов для генов 1 и 2;

4) идентификатор соответствующего контига в сборке, на который картирован гибрид;

5) информацию о сдвиге рамки считывания в гибриде;

6) число прочтений, картированных непосредственно на точку разрыва;

7) число прочтений, картированных на тот же гибрид, но располагающихся по флангам относительно точки разрыва.

Более подробно ознакомиться с данным форматом можно, используя статью авторов алгоритма Fusion-Bloom, а также спецификацию данного формата [36, 39, 40]. На сегодняшний день, по словам авторов, Fusion-Bloom представляет собой один из наиболее удачных инструментов для поиска химерных генов, использующих технологию сборки транскриптома *de novo*.

8. ПРОГРАММЫ, ИСПОЛЬЗУЮЩИЕ ДЛЯ ДЕТЕКЦИИ ГИБРИДНЫХ ГЕНОВ ПСЕВДОВЫРАВНИВАНИЕ. KALLISTO И PIZZLY

Kallisto – инструмент для оценки сырых данных РНК-секвенирования, осуществляющий процедуру псевдовыравнивания, чтобы подсчитать число выровненных прочтений без проведения выравнивания. Основной подход такого выравнивания – использовать *k*-мер транскрипта (часть транскрипта, состоящего из *k*-букв) и с применением хэш-таблицы сравнивать, насколько *k*-мер, соответствующий данному транскрипту, схож с прочтением [41]. Кроме того, транскрипты сопоставляются с прочтениями, как правило, на основе некоторого статистического порога, поэтому авторы

поменяли постановку вопроса: не где в транскрипте картировано прочтение, а какие транскрипты могли породить такое прочтение.

Рассмотрим, как происходит процедура псевдовыравнивания. Сравнение прочтений секвенирования с транскриптами выполняется с использованием графа де Брюйна. Такой граф строится на основе *k*-меров транскриптома, т.е. последовательностей длиной *k* букв, выбранных из транскрипта. Каждый такой *k*-мер связан с набором транскриптов, для которого этот *k*-мер является подпоследовательностью, а набор соответствующих транскриптов – "классом эквивалентности".

После построения графа де Брюйна создается хэш-таблица, содержащая соответствие между *k*-мерами и участками *de novo* собранного транскриптома (контигами), более точно – *k*-мер прочтения, имеющий хэш-ключ, наиболее близкий по значению хэш-ключу *k*-мера контига транскриптома.

Затем для каждого прочтения РНК-секвенирования берется его подпоследовательность из *k*-букв, называемая *k*-мером прочтения, и далее, чтобы найти транскрипты, на которые можно картировать прочтение, алгоритм выбирает пересечение по хэш-таблице для всех *k*-меров прочтения и всех соответствующих им *k*-меров транскриптов (рис. 1).

Таким образом, процедура псевдовыравнивания заключается в том, чтобы обойти по эйлеровому пути граф де Брюйна, содержащий наборы транскриптов, составляющих в сумме весь транскриптом, и сравнить хэш-ключи *k*-меров таких транскриптов с хэш-ключами *k*-меров прочтений, что осуществляется с помощью EM-алгоритма (максимизации ожидаемого сходства) и позволяет осуществлять сверхбыстрый анализ [41].

Таким образом, каждому прочтению будут соответствовать некоторые множества транскриптов, содержащих этот *k*-мер, которые называют классами эквивалентности. Обычному прочтению, которое возникло в ходе секвенирования обычного транскрипта, будет соответствовать хотя бы один транскрипт, следовательно, пересечение прочтения с классами эквивалентности не будет пустым.

Отсюда следует, что прочтению, которое картируется на стыке двух экзонов, не будет соответствовать ни один транскрипт, и его пересечение с классами эквивалентности будет пустым. При должных настройках Kallisto в режиме поиска гибридных генов находит либо прочтения, которые имеют пустое пересечение с классом эквивалентности, либо такие прочтения, которые можно разбить на две части, где каждая часть будет относиться к своему классу эквивалентности, и при этом полное прочтение нельзя будет отнести ни к одному из классов эквивалентности.

В случае парноконцевого секвенирования (paired-end sequencing) отбираются случаи, когда каждое из пары прочтений имеет непустое пересечение с некоторым классом эквивалентности, но при этом вместе парноконцевые прочтения не имеют общего класса эквивалентности.

На выходе Kallisto выдает список прочтений с соответствующими кандидатами в гибридные гены. Набор фильтров Pizzly, в свою очередь, принимает на вход такой список и, уже придерживаясь геномной аннотации, анализирует список гибридных генов для повышения специфичности метода детекции [42]. Pizzly удаляет прочтения, которые картируются на транскрипты в нескольких геномных локациях, кроме того, удаляются парноконцевые прочтения, которые поддерживают транскрипты одного и того же гена, содержащие однонуклеотидный полиморфизм (SNP) [42].

9. ПРОГРАММЫ ДЛЯ ДЕТЕКЦИИ ГИБРИДНЫХ ГЕНОВ В ДАННЫХ СЕКВЕНИРОВАНИЯ ДЛИННЫМИ ПРОЧТЕНИЯМИ. JAFFA И JAFFAL

Программа JAFFA специализируется на коротких прочтениях. JAFFA сравнивает и обрабатывает транскриптомные данные, полученные от некоторой раковой линии клеток, и данные эталонного транскриптома, где эталонными считаются транскрипты, полученные из GENCODE [43] (проекта, цель которого – аннотирование всех генов в геномах человека и мыши; является частью более глобального проекта ENCODE (Encyclopaedia of DNA Elements). Эта программа имеет три режима для поиска гибридных генов:

1) режим сборки – короткие прочтения собираются в более длинные транскрипты на этапе, предшествующем поиску гибридов;

2) прямой режим – режим, при котором алгоритм использует прочтения для картирования на эталонный транскриптом непосредственно, с последующим отбором гибридных транскриптов;

3) гибридный режим – режим, в котором происходит как сборка транскриптов, так и выявляются прочтения, не соответствующие ни эталонному транскриптому, ни сборке.

Необходимый режим можно выбрать на основании длины прочтения. По умолчанию JAFFA предполагает использование прочтений не короче 60 п.н. и не длиннее 99 п.н., а потому требует, чтобы как минимум 30 нуклеотидов выравнивались по обе стороны от точки разрыва, и использует BLAT в качестве картировщика, что соответствует гибридному режиму работы [44, 45].

Для прочтений короче 60 п.н. JAFFA использует режим сборки транскриптов *de novo*, а для прочтений длиннее 100 п.н. и более – прямой режим. Сборка транскриптов *de novo* осуществляется с помощью программных пакетов Oases [46, 47] и Velvet [46, 47], реализующих непосредственную сборку прочтений в контиги.

Вне зависимости от используемого режима работы прочтения выравниваются на геном. Затем прочтения, выравнивающиеся на интронные или межгенные участки, удаляются первыми для улучшения производительности. После этого, в зависимости от длины прочтений, подразумевающих выбранный оптимальный режим, происходит либо сборка транскриптов *de novo* и получаются собранные транскрипты и оставшиеся прочтения, либо прочтения используются непосредственно.

В любом случае – как при наличии сборки *de novo*, так и при ее отсутствии – следующим шагом происходит отбор прочтений, которые выравниваются на несколько генов, и выполняется подсчет числа прочтений, поддерживающих данную точку разрыва. Затем уже потенциальные кандидаты в гибридные гены выравниваются еще раз на геном – для определения координаты точки разрыва. Далее следует этап фильтрации и классификации кандидатов на основе следующих метрик: числа поддерживающих прочтений, выравнивания точек разрыва по границам экзонов, а также величины геномного интервала между собранными транскриптами. В результате фильтрации программа JAFFA предоставляет список кандидатов в гибриды вместе с их последовательностью.

JAFFAL, в свою очередь, является развитием программы JAFFA для длинных прочтений, генерируемых платформами секвенирования третьего поколения PacBio [48] и ONT [49, 50]. JAFFAL написан на языке bpipe, придуманном специально для работы с биоинформатическими программными инструментами [43, 51, 52]. JAFFAL использует прямой режим, который был ранее описан в JAFFA, и состоит из следующих шагов:

1) прочтения выравнивают посредством алгоритма Minimap2 [53] на эталонный транскриптом (Gencode 22, hg38) [54];

2) прочтения, картированные алгоритмом одновременно на экзоны разных генов, отбираются для дальнейшего анализа;

3) отобранные прочтения повторно выравниваются алгоритмом Minimap2 на эталонный геном hg38. Те прочтения, которые не выравниваются на несколько генов одновременно, удаляются. Этот шаг значительно уменьшает объем анализируемых данных и сокращает время работы;

4) затем кандидаты в гибридные гены повторно выравниваются на геном – для определения координаты точки разрыва. Основной критерий – это выравнивание точек разрыва на соответствующие границы экзонов.

В данной программе считается, что точки разрыва (breakpoints) выравниваются по границам экзонов, если границы экзонов находятся не далее 20 п.н. от исходных точек разрыва выравнивания.

Инсерции и делеции, точки разрыва в теле экзона приводят к тому, что многие прочтения не удовлетворяют условию совпадения координат точки разрыва между прочтениями. Такие прочтения сгруппированы по месту картирования в геноме. В таком случае для каждого кластера назначается такая точка разрыва, которая сохраняет границы экзонов, либо такая точка разрыва, которая имеет самое большое число выровненных прочтений. При этом кластеризация осуществляется посредством перебора всех точек разрыва, не принадлежащих экзонам, начиная с тех точек разрыва, которые имеют наименьшее число выровненных прочтений.

Группа прочтений, соответствующих точке разрыва 1, будет переназначена на ближайшую точку разрыва 2, связанную со второй группой прочтений, если первая группа прочтений находится не далее 50 п.н. При этом если в пределах 50 п.н. не обнаружено другой точки разрыва, будет назначена текущая точка разрыва;

5) в итоге точки разрыва, как и в JAFFA, разбиваются по классам – в соответствии с рангом:

– высокая достоверность – точки разрыва поддерживаются двумя и более прочтениями с такими точками разрыва, при этом точки разрыва совпадают с границами экзонов;

– низкая достоверность – точки разрыва поддерживаются двумя или более прочтениями, но точки разрыва не совпадают с границами экзонов;

– потенциальный *транс*-сплайсинг – точки разрыва поддерживаются единственным прочтением, при этом точки разрыва совпадают с границами экзонов.

Стоит отметить, что в образцах РНК-секвенирования здоровых тканей наблюдаются многочисленные события потенциального *транс*-сплайсинга – такие события должны быть отфильтрованы [43, 55]. Случается, что о некоторых настоящих гибридах JAFFAL сообщает как о потенциальном *транс*-сплайсинге, однако авторы оправдывают это либо низким уровнем экспрессии или низкой долей раковых клеток в тканях солидной опухоли, называемой “чистотой опухоли” (“tumor purity”) [51, 56]. Доля раковых клеток в тканях влияет на последующее РНК-секвенирование и соответствующий анализ. Существующие на данный момент методы нормализации и коррекции не устраняют данную проблему [51, 56].

Также отфильтровываются гибриды, в которых участвует митохондриальная ДНК. Если же точки разрыва находятся в пределах 200 т.п.н., где гены транскрибируются в том же порядке, что и эталонный геном, то прочтения, поддерживающие такие точки разрыва, также отфильтровываются по умолчанию [51, 57].

Для каждой точки разрыва, которая проходит фильтрацию, JAFFAL сообщает следующую информацию:

– задействованные гены;

– геномные координаты;

– число прочтений, подтверждающих слияние;

– соответствующий ранг;

– совпадает ли гибрид с рамкой считывания относительно имеющейся точки разрыва, наблюдалась ли такая точка разрыва ранее в базах данных геномных перестроек Мительмана [51].

Одна из основных особенностей JAFFAL – присваивание прочтению ранга, соответствующего рангу точки разрыва и с учетом сохранения открытой рамки считывания (ORF) для данной точки разрыва [51, 57].

10. LONG GF

Long GF был одним из первых программных инструментов, предназначенных для поиска гибридных транскриптов в транскриптомных или экзомных данных, полученных методом секвенирования длинными прочтениями. Он написан на языке C++ и использует по умолчанию алгоритм Minimap2 [53] для картирования прочтений на транскриптом. Как и некоторые картировщики, обрабатывающие короткие прочтения, Long GF картирует длинные прочтения на эталонный транскриптом, после чего отбирает такие длинные прочтения, которые картируются на несколько генов так, чтобы прочтение, картированное на несколько генов, имело при этом достаточное перекрытие как с первым, так и со вторым геном [58], т.е. ищет прочтения, картирующиеся в несколько мест, и порог перекрытия в таком случае устанавливается пользователем.

При этом в качестве входных данных требуется BAM-файл, содержащий выравненные длинные прочтения (по умолчанию LongGF использует Minimap2 и файл аннотации GTF, содержащий информацию о генах). На выходе же пользователь получает файл с прочтениями, отсортированными по приоритету, которые LongGF соотносит с потенциальными гибридами и сортирует такие гибриды по числу прочтений, поддерживающих гибриды [57, 58].

11. FUSIONSEEKER

Одна из недавно выпущенных программ FusionSeeker [59] осуществляет поиск гибридов в данных секвенирования длинными прочтениями. На вход данная программа принимает сортированный файл в формате .bam, который был предварительно выровнен картирующим алгоритмом, например, Minimap2 [53], на геном человека (Genome Reference Consortium Human Build 38 (GRCh38) – эталонная сборка генома версии 38) с версией аннотации № 104 проекта Ensembl (научного проекта, цель которого – поддержка и создание геномных баз данных 50 видов позвоночных, включая человека). После обработки предоставленных данных программа возвращает список кандидатов в гибридные гены и список

de novo реконструированных транскриптов (консенсусных последовательностей).

Файл в формате .bam содержит два раздела: раздел “Заголовок” и раздел “Выравнивания”. Раздел “Заголовок” содержит информацию о названии образца, длине образца и методе выравнивания. Раздел “Выравнивания” включает информацию о названии прочтения, хромосоме, начальной координате выравнивания, качестве выравнивания, последовательности прочтения, качестве прочтения и пользовательские теги.

На первом этапе FusionSeeker проверяет файл выравнивания, полученный на входе, и выявляет случаи, когда прочтение выравнивается на число генов большее либо равное двум. В таких случаях каждому прочтению ставятся в соответствие гены, на которые выравнивается такое прочтение [59].

FusionSeeker сообщает о гипотетическом гибриде в том случае, если соблюдаются следующие условия:

1) две точки разрыва для прочтения относятся к двум разным генам (*Ген_1* и *Ген_2*);

2) длина выравнивания для обоих генов превышает 100 п.н. для каждого из генов;

3) длина перекрытия между двумя выравниваниями сегментов прочтения, картированных на каждый из генов в отдельности, менее 100 п.н.;

4) координаты *Гена_1* и *Гена_2* не перекрываются в файле аннотации (.gtf);

5) *Ген_1* не является антисмысловой последовательностью *Гена_2*.

Затем прочтения кластеризуют так, чтобы каждому гибриду, состоящему из пары или большего числа генов, соответствовал список прочтений, которые картируются на такой гибрид. После того как гипотетическому гибриду (паре генов) поставлены в соответствие прочтения, содержащие сегменты, картированные на такие гены, прочтения кластеризуют на основе точек разрыва. При этом допустимое расстояние между точками разрыва в прочтениях, подтверждающих гибрид, составляет не более 20 п.н. для прочтений HiFi (PacBio High Fidelity reads – прочтения с высоким уровнем достоверности 99.9%, точность совпадения нуклеотидов при секвенировании) и не более 40 п.н. для остальных прочтений [48]. В таком случае точкой разрыва назначается точка, представляющая собой среднее значение среди всех прочтений, поддерживающих гибрид. Эта процедура осуществляется алгоритмом DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [60].

После этого прочтения фильтруются посредством установки нижней границы для числа прочтений, картирующихся на гибрид, чтобы отсеять прочтения, содержащие большое число ошибок или ошибочно выровненных на гибрид.

По умолчанию минимальное число прочтений рассчитывается как Nmin = Ncan/50000 + 3, где Ncan – общее количество кандидатов в гибридные гены, обнаруженных в данном наборе данных. Отметим, что автор не уточняет причины выбора константы 50000 [59].

После фильтрации FusionSeeker осуществляет множественное выравнивание последовательностей прочтений РНК-секвенирования, содержащих кандидаты в гибридный ген, чтобы сгенерировать *de novo* консенсусную последовательность транскрипта с помощью алгоритма bsalign [61]. Затем консенсусные последовательности собираются в отдельный FASTA-файл, где каждый гибрид имеет свой ID. Далее консенсусные последовательности (транскрипты, собранные *de novo*) выравниваются повторно на геном для уточнения координат точки разрыва. После этого координата точки разрыва, рассчитанная как среднее значение, заменяется на координату точки разрыва, полученную при картировании реконструированного транскрипта *de novo* – для всех прочтений, поддерживающих данный гибрид.

12. ДОСТОИНСТВА И НЕДОСТАТКИ АКТУАЛЬНЫХ ПОДХОДОВ К ДЕТЕКЦИИ И ГИБРИДНЫХ ГЕНОВ И ТРАНСКРИПТОВ. СРАВНИТЕЛЬНЫЕ ИССЛЕДОВАНИЯ СУЩЕСТВУЮЩИХ ПРОГРАММ ДЛЯ ПОИСКА ГИБРИДНЫХ ГЕНОВ И ИХ ПРОДУКТОВ

Для алгоритмов, описанных выше, проводились сравнительные исследования с точки зрения ресурсоемкости, использованного компьютерного времени, чувствительности и специфичности методов.

Касательно коротких прочтений, размер которых находится в пределах 50–100 п.н., то в следующем исследовании от 2021 г. [28] указывалось, что на сегодняшний день наиболее точным среди алгоритмов картировщиков, выявляющих ранее обнаруженные гибриды, является Arriba. Arriba

совмещает в себе высокий уровень чувствительности и специфичности, а вторым по данным показателям ожидаемо стал STAR-Fusion [24]. Более ранее исследование 2019 г. указывало на то, что, наоборот, наибольшую эффективность имеет STAR-Fusion [21]. Однако оба алгоритма основаны на картирующей программе STAR и отличаются скорее набором фильтров, которые могут играть как положительную роль, повышая точность исследования, так и отрицательную – сокращая чувствительность. Причем, поскольку истинный набор гибридов в транскриптоме опухолевой ткани определить бывает весьма сложно [21], то, в некотором смысле, успех будет зависеть и от качества предоставленных данных и самого образца [21].

STAR-Fusion – один из первых вариантов программ для поиска гибридов, основанных на картирующей программе STAR. Оба алгоритма наиболее быстрые (согласно оценкам авторов исследования [28]), однако стоит отметить, что данный алгоритм потребляет наибольший объем памяти среди всех описанных алгоритмов за счет размера индексных файлов, созданных программой STAR [19].

Отметим также другой подход, позволяющий осуществлять сверхбыстрое сопоставление между прочтениями РНК-секвенирования и эталонным транскриптомом, основанный на псевдовыравнивании.

Программы, использующие данный подход, такие как Kallisto и Pizzly [41, 42], имеют как свои преимущества (в виде высокой скорости и небольшого уровня потребления памяти), так и свои недостатки (например, они не осуществляют локального выравнивания прочтений, как это делает большинство алгоритмов-картировщиков) [21, 41]. Также алгоритмы, основанные на псевдовыравнивании, имеют меньшую точность с точки зрения выявления гибридов [21, 41].

Kallisto и Pizzly пробуют искать максимальную совпадающую с эталоном подпоследовательность, чтобы локализовать местоположение прочтения в эталонном транскриптоме, а далее с помощью графа де Брюйна, построенного заранее из последовательностей эталонного транскриптома, – точнее локализовать ген или транскрипт, на который может выравниваться такое прочтение. Локальное выравнивание в таком случае не осуществляется, как это происходит в алгоритмах-картировщиках, осуществляющих помимо глобального еще и локальное выравнивание [62, 63]. Такой информации достаточно для оценки уровней экспрессии генов в образце. Например, возможно оценить уровень экспрессии хорошо известных генов, у которых уровень экспрессии выше среднего, как, например, у ряда белок-кодирующих генов [63].

Появление технологий секвенирования третьего поколения [48–50], генерирующих прочтения длиной в десятки тысяч пар нуклеотидов (так называемые “длинные” прочтения), потребовало создания новых алгоритмов, позволяющих картировать такие прочтения. Кроме того, это позволило изучать новые изоформы (structural variants) транскриптов, которые ранее невозможно было детектировать [57, 64, 65]. Однако на сегодняшний день технологии длинного секвенирования и алгоритмы, осуществляющие картирование таких прочтений, имеют меньшее геномное покрытие и все еще более высокий уровень ошибок по сравнению с секвенированием короткими прочтениями. В настоящее время наибольшую чувствительность, согласно существующим оценкам, имеет программа FusionSeeker [59], однако более высокую точность показал алгоритм LongGF, как и сравнимую по уровню чувствительность.

Алгоритмы, созданные для работы с длинными прочтениями, вдобавок к гибридам, имеющим строение, похожее на то, что выявлялось ранее, находят, например, гибриды, содержащие большую интронную вставку [56]. Некоторые из существующих программ по выравниванию длинных прочтений разбивают прочтение на несколько подпоследовательностей – сегментов, например, длиной 250 п.н., и выравнивают каждый сегмент в отдельности, другие используют хэширование (см. Дополнительные материалы) для последовательности [63].

На сегодняшний день возможности программ, выявляющих гибриды, ограничены подходами, заложенными в модели поиска и прогноза форм и путей образования гибридов, узкой направленностью по отношению к форматам используемых данных и источниками информации, которые они используют. Так, например, программы, сначала выравнивающие прочтения на эталонную последовательность для идентификации последовательностей, картирующихся в несколько мест, как правило, имеют высокую чувствительность, однако могут быть недоста-

точно специфичными [66]. Поэтому исследователи вынуждены создавать различные типы фильтров, основанные как на объективных критериях (таких как, например, дубликаты прочтений или прочтения, картированные на паралоги генов), так и на вероятностной эвристике разного рода, такой как частота встречаемости однонуклеотидных полиморфизмов (SNP), уровень экспрессии транскрипта, число прочтений, поддерживающих данную точку разрыва/слияния, и прочие характеристики, которые хотя и ориентированы на экспериментальные данные, но отличаются в зависимости от инструмента и образца [24, 57, 67]. Кроме того, алгоритмы-картировщики, как правило, наиболее ресурсоемкие.

Программы, собирающие прочтения в более длинные транскрипты *de novo* за счет ориентации на предполагаемые продукты экзонов с последующей идентификацией химерных РНК-транскриптов, соответствующих хромосомным перестройкам, могут находить различные новые изоформы гибридных генов. Однако такие программы порождают много искусственных химерных прочтений, которые бывает сложно выявить. Кроме того, этот подход также весьма ресурсоемкий [67].

Программы, использующие псевдовыравнивание на транскриптом с целью идентификации химерных транскриптов, – как правило, самые быстрые алгоритмы, но при этом наименее точные и более других основаны на вероятностном подходе [42].

Все программы различаются по объемам используемого времени и оперативной памяти компьютера, по скорости вычислений и имеют значительные различия как с точки зрения “прогностической эффективности”, так и с точки зрения сопоставимости результатов работы алгоритма. Это происходит ввиду различных моделей и источников используемой информации, а также специфичной работы самой программы и ее алгоритма. Существующие на данный момент инструменты, кроме прочего, производят ограниченное количество совпадающих между самими программами экспериментально валидированных результатов. Тем не менее с важными с точки зрения поиска гибридов характеристиками, такими как чувствительность и специфичность алгоритмов, для всех описанных в статье алгоритмов, можно ознакомиться в табл. 1–4.

13. ПЕРСПЕКТИВЫ РАЗВИТИЯ ПРОГРАММ И МЕТОДОВ ПОИСКА ГИБРИДНЫХ ГЕНОВ И ИХ ПРОДУКТОВ

На сегодняшний день все большее распространение получают программы, сочетающие в себе результаты секвенирования как с помощью длинных, так и с помощью коротких прочтений. С одной стороны, это обосновано фундаментальным интересом в отношении наличия любых других форм гибридных генов и их транскриптов, а также необходимостью изучения молекулярных взаимодействий, которые такие мультисегментные формы транскриптов, содержащие некоторую вставку (например, интронную), способны порождать; с другой стороны – повышением качества данных, которые получаются в результате секвенирования [57]. Однако существуют и более узкие задачи, которые также имеют важное приложение, например, в клинической практике, где онкологические образцы часто представлены в виде фиксированных формалином парафинизированных тканей [68], в которых длинное секвенирование (5000–100000 п.н.) может быть не оправдано ввиду коротких фрагментов РНК (не более 300 п.н. на фрагмент), содержащихся в самой ткани после такой обработки [69].

В связи с ежегодным увеличением вычислительной мощности компьютеров, авторам настоящего обзора также видится, что развитие методов картирования продолжится в сторону уменьшения ресурсоемкости, а также увеличения чувствительности и специфичности для усиления прогностической ценности имеющихся моделей [70].

14. ОБСУЖДЕНИЕ ВОЗМОЖНОСТЕЙ И ОГРАНИЧЕНИЙ СУЩЕСТВУЮЩИХ ПРОГРАММ ДЛЯ ПОИСКА ГИБРИДОВ И ИХ ПРОДУКТОВ

Стоит отметить, что алгоритмы, используемые для поиска гибридов, также ориентируются на протокол проведения эксперимента, т.е. на то, какое секвенирование выбрано (одноконцевое или парноконцевое), какой длины прочтения рассматриваются (от десятков до тысяч пар оснований), специфично ли такое секвенирование относительно направления цепи (strand-specific) или не специфично. С помощью секвенирования длинными прочтениями можно находить сложные формы мультисегментных гибридов, например, гибридов, содержащих интронную вставку. В то же время короткие прочтения часто имеют более высокую чувствительность по сравнению

Таблица 1. Сравнение чувствительности (%) программ, использующих короткие прочтения РНК-секвенирования, на синтетических наборах данных для инструментов Fusion-Bloom, JAFFA, STAR-Fusion и pizzly в зависимости от концентрации гибрида в образце

Образцы	Молярность, –log10 пкмоль	Чувствительность*, %			
		Fusion-Bloom	JAFFA	STAR-Fusion	pizzly
SRR1659951	3.47	100	89	100	100
SRR1659961	4.17	100	89	100	100
SRR1659953	4.87	100	89	100	100
SRR1659963	5.57	100	89	100	100
SRR1659959	5.87	100	89	100	100
SRR1659964	6.17	100	89	100	100
SRR1659965	6.57	100	89	100	100
SRR1659957	6.87	100	89	100	78
SRR1659958	7.17	100	89	100	89
SRR1548811	8.57	100	78	100	78

Примечание: таблица сравнения приведена в сокращенном виде из дополнительных материалов статьи Chiu et al. [36], в которой сравнение чувствительности алгоритмов проводилось на синтетических образцах из статьи Tembe et al. [75], в которых присутствовали уже хорошо изученные девять гибридов (EWS-ATF1, TMPRSS2-ETV1, EWS-FLI1, NTRK3-ETV6, CD74-ROS1, HOOK3-RET, EML4-ALK, AKAP9-BRAF, BRD4-NUT) в разной концентрации в образце.
* Чувствительность = (TP/9) × 100, где TP – число верно найденных гибридов, 9 – общее число гибридов в каждом образце [36].

Таблица 2. Сравнение чувствительности программ, использующих короткие прочтения РНК-секвенирования для поиска гибридов в наборах данных реальных 56 раковых транскриптомов* для инструментов Arriba, STAR-Fusion, JAFFA, Pizzly

Название инструмента	Чувствительность, %	Специфичность, %	Ссылки
Arriba	80	63	[28]
STAR-Fusion	85	78	[24]
JAFFA-Assembly**	65	47	[43]
JAFFA-Direct**	78	35	[43]
JAFFA-Hybrid**	60	20	[43]
Pizzly	48	44	[42]

Примечание: расчеты в таблице сравнения сделаны на основе материалов статьи Haas et al. [21]. Чувствительность = (TP/(TP+FN)) × 100; Специфичность = (TP/(TP+FP)) × 100; где TP – число гибридов, найденное всеми шестью программами; FP – число гибридов, не найденное другими программами, но найденное текущей программой; FN – число гибридов, не найденное текущей программой, но найденное остальными программами из набора (TP).
* Поскольку истинный набор гибридов в транскриптоме раковых клеток определить на данный момент нельзя, в статье используется принцип согласованности программ: гибрид считается истинным, если он подтверждается вышеуказанными программами одновременно [21].
** JAFFA-Assembly, JAFFA-Direct, JAFFA-Hybrid – три различных режима поиска гибридов программы JAFFA.

Таблица 3. Сравнение чувствительности программ, использующих длинные прочтения РНК-секвенирования, на синтетических наборах данных для инструментов JAFFAL, FusionSeeker и LongGF в зависимости от уровня экспрессии гипотетического гибрида в образце

Название инструмента	Номер образца*	Тип прочтений						Ссылки
		PacBio IsoSeq			ONT			
		Уровень экспрессии			Уровень экспрессии			
		высокий (100×)	средний (50×)	низкий (10×)	высокий (100×)	средний (50×)	низкий (10×)	
		Чувствительность, %						
FusionSeeker	Образец 1	96.00	97.96	90.20	98.00	100.0	100.0	[59]
	Образец 2	100.0	94.12	88.64	100.0	98.04	95.45	
	Образец 3	100.0	100.0	91.84	100.0	100.0	100.0	
	Среднее	98.67	97.36	90.22	99.33	99.35	98.48	
JAFFAL	Образец 1	54.00	55.10	33.33	62.00	57.14	35.29	[51]
	Образец 2	49.09	56.86	43.18	52.73	62.75	45.45	
	Образец 3	57.45	57.41	53.06	61.70	61.11	53.06	
	Среднее	53.51	56.46	43.19	58.81	60.33	44.60	
LongGF	Образец 1	82.00	85.71	64.71	84.00	87.76	70.59	[58]
	Образец 2	81.82	86.27	90.91	81.82	90.20	90.91	
	Образец 3	80.85	81.48	87.76	80.85	81.48	91.84	
	Среднее	81.56	84.49	81.12	82.22	86.48	84.44	

Примечание: сравнение чувствительности приведено из дополнительных материалов статьи Chen et al. [59]. Алгоритм картирования прочтений во всех трех случаях – Minimap2 [53], для PacBio Isoseq гибриды смоделированы с помощью Badread (v0.2.0) [76]; для ONT (Nanopore) – с помощью pbsim [77].
* Образец здесь указан в качестве повторности: для каждого образца генерировался свой набор гибридов по следующему принципу: общее число искусственных гибридов составляло 150, причем 100 имели точку разрыва в экзоне, а 50 – в интроне. Каждому гибриду был случайно присвоен разный уровень экспрессии из трех возможных: 10×, 50× или 100×. Чувствительность (%) = (TP/150) × 100, где TP – число верно найденных гибридов, 150 – общее число сгенерированных истинных гибридов.

Таблица 4. Сравнение точности программ, использующих длинные прочтения РНК-секвенирования, на синтетических наборах данных для инструментов JAFFAL, FusionSeeker и LongGF в зависимости от уровня экспрессии гипотетического гибрида в образце

Название инструмента	Номер образца*	Тип прочтений						Ссылки
		PacBio IsoSeq			ONT			
		уровень экспрессии			уровень экспрессии			
		высокий (100×)	средний (50×)	низкий (10×)	высокий (100×)	средний (50×)	низкий (10×)	
		точность, %						
FusionSeeker	Образец 1	94.56	65.31	29.25	92.81	61.44	31.37	[59]
	Образец 2	93.29	61.74	31.54	88.41	58.54	29.88	
	Образец 3	94.63	63.76	30.87	91.82	61.64	30.19	
	Среднее	94.16	63.60	30.56	91.02	60.54	30.48	
JAFFAL	Образец 1	82.56	79.07	3.49	82.80	76.34	6.45	[51]
	Образец 2	78.13	70.83	7.29	75.00	67.59	7.41	
	Образец 3	87.50	75.00	12.50	88.89	75.76	13.13	
	Среднее	82.73	74.97	7.76	82.23	73.23	9.00	
LongGF	Образец 1	96.67	78.33	18.33	96.03	76.19	19.84	[58]
	Образец 2	94.85	70.59	24.26	94.24	70.50	23.74	
	Образец 3	96.90	75.97	20.93	95.49	75.19	20.30	
	Среднее	96.14	74.96	21.18	95.26	73.96	21.29	

Примечание: таблица сравнения точности приведена по данным статьи Chen et al. [59]. Алгоритм картирования прочтений во всех трех случаях – Minimap2 [53]. Для PacBio Isoseq гибриды смоделированы с помощью Badread (v0.2.0) [76], для ONT (Nanopore) – с помощью pbsim [77].
* Образец указан в качестве повторности: для каждого образца генерировался свой набор гибридов по следующему принципу: общее число искусственных гибридов составляло 150, причем 100 имели точку разрыва в экзоне, а 50 – в интроне. Каждому гибриду был случайно присвоен разный уровень экспрессии из трех возможных: 10×, 50× или 100×. Точность (%) = (TP/TP + FP) × 100, где TP – число верно найденных гибридов, FP – число ложноположительных гибридов.

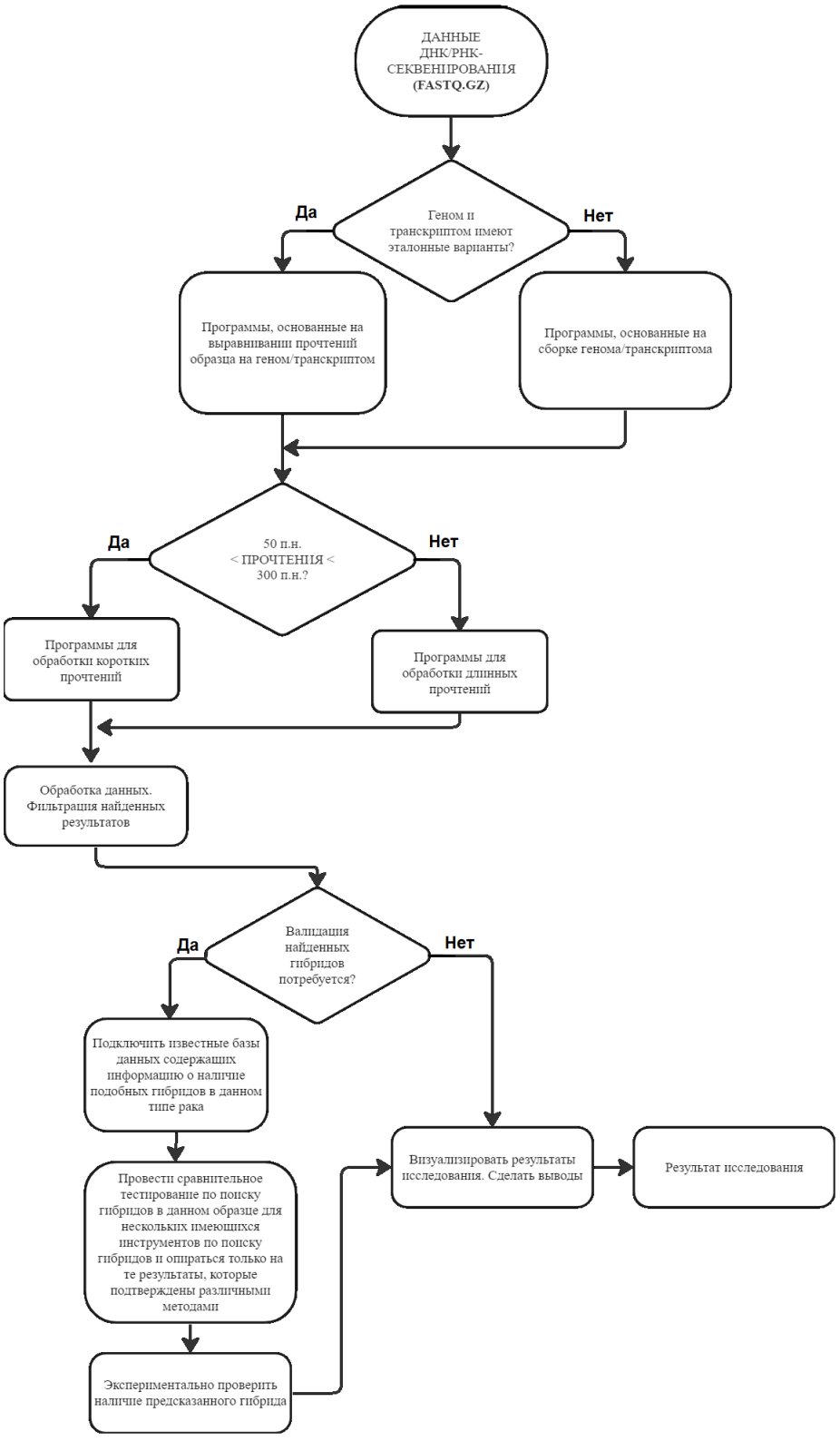

Схема 1. Общая схема действий при исследовании образца ДНК/РНК-секвенирования на наличие гибридов.

с секвенированием длинными прочтениями [57]. Для удобства использования имеющихся знаний об алгоритмах и составления плана обработки различных наборов данных секвенирования с целью поиска новых или уже известных гибридов авторы настоящего обзора подготовили схему 1.

Проблема создания эффективного поиска гибридов алгоритмом заключается в том, что качество работы алгоритма фактически оценивается по тому, насколько хорошо такой алгоритм выявляет уже ранее найденные экспериментально гибриды. Это фактически ограничивает "прогностическую" способность алгоритма и сводит его работу к выявлению ранее незамеченных гибридов, формирование которых потенциально подчиняется определенным закономерностям, детектируемым компьютерно [71]. При этом алгоритмы в больших массивах данных могут находить ранее не известные гибриды, которые валидируются исключительно экспериментально, что является дополнительным ограничением компьютерных методов. Тем не менее, даже несмотря на такие сильные ограничения с точки зрения исследования, компьютерные методы не ограничены в гипотезах о закономерностях, которые выявляются с их помощью. Поэтому задача биоинформатики в этом смысле заключается в создании таких алгоритмов, которые бы эффективно связывали и устанавливали закономерности формирования таких гибридов, которые невозможно детектировать без компьютера. Кроме того, необходимо с использованием различных гипотез и моделей выстроить такую систему, которая позволила бы эффективнее ставить экспериментальные цели и видеть исследование на шаг вперед.

Более того, поскольку поиск гибридов осуществлялся в разные годы разными методами, образцы хранились по-разному, и сами эксперименты проводились разными людьми, часто возникает проблема сравнения данных, полученных тем или иным способом. Для этого на сегодняшний момент разрабатываются так называемые "гармонизаторы" данных, т.е. программы, позволяющие при определенной обработке сравнивать данные, полученные разными способами [72–74].

15. ЗАКЛЮЧЕНИЕ

На сегодняшний день наиболее быстрыми и эффективными из существующих программ для поиска гибридных генов являются программы-картировщики, такие как STAR-Fusion и Arriba. Другой интересный подход использует программы-сборщики транскриптома, такие как Fusion-Bloom, которые способны эффективно определять изоформы транскриптов, слабо представленные в референсной ДНК: например, последовательности бактериальных и вирусных геномов. Наконец, наиболее быстрыми, но вместе с тем наименее точными являются программы, основанные на псевдовыравнивании (Kallisto, Pizzly), которые при должной фильтрации могут обеспечить эффективное выявление гибридов. С учетом развития технологий секвенирования для анализа образцов нуклеиновых кислот высокого качества перспективными выглядят алгоритмы, работающие с длинными прочтениями, такие как LONG GF и FusionSeeker, при этом на текущий момент первый имеет наибольшую точность, а второй – наибольшую чувствительность.

ФОНДОВАЯ ПОДДЕРЖКА

Статья подготовлена на основании результатов, полученных в ходе реализации Соглашения о предоставлении гранта в форме субсидий из федерального бюджета на осуществление государственной поддержки создания и развития научных центров мирового уровня, выполняющих исследования и разработки по приоритетам научно-технологического развития от 20 апреля 2022 г. № 075-15-2022-310.

СОБЛЮДЕНИЕ ЭТИЧЕСКИХ СТАНДАРТОВ

Настоящая статья не содержит описания каких-либо исследований с участием людей или животных в качестве объектов исследования.

КОНФЛИКТ ИНТЕРЕСОВ

Авторы заявляют об отсутствии конфликта интересов.

СПИСОК ЛИТЕРАТУРЫ

1. *Barresi V., Cosentini I., Scuderi C., Napoli S., Di Bella V., Spampinato G., Condorelli D.F.* // Int. J. Mol. Sci. 2019. V. 20. P. E5252. https://doi.org/10.3390/ijms20215252
2. *Friedrich S., Sonnhammer E.L.L.* // BMC Med. Genomics. 2020. V. 13. P. 110., https://doi.org/10.1186/s12920-020-00738-5
3. *Sun Y., Li H.* // Genes (Basel). 2022. V. 13. P. 741. https://doi.org/10.3390/genes13050741
4. *Li Z., Qin F., Li H.* // Curr. Opin. Genet. Dev. 2018. V. 48. P. 36–43. https://doi.org/10.1016/j.gde.2017.10.002

5. *Xie Z., Babiceanu M., Kumar S., Jia Y., Qin F., Barr F.G., Li H.* // Proc. Natl. Acad. Sci. USA. 2016. V. 113. P. 13126–13131. https://doi.org/10.1073/pnas.1612734113
6. *Shtivelman E., Lifshitz B., Gale R.P., Canaani E.* // Nature. 1985. V. 315. P. 550–554. https://doi.org/10.1038/315550a0
7. *Pagani I.S., Dang P., Kommers I.O., Goyne J.M., Nicola M., Saunders V.A., Braley, J., White D.L., Yeung D.T., Branford S., Hughes T.P., Ross D.M.* // Haematologica. 2018. V. 103. P. 2026–2032. https://doi.org/10.3324/haematol.2018.189787
8. *Zhou T., Medeiros L.J., Hu S.* // Curr. Hematol. Malig. Rep. 2018. V. 13. P. 435–445. https://doi.org/10.1007/s11899-018-0474-6
9. *Mertens F., Johansson B., Fioretos T., Mitelman F.* // Nat. Rev. Cancer. 2015. V. 15. P. 371–381. https://doi.org/10.1038/nrc3947
10. *Sorokin M., Rabushko E., Rozenberg J.M., Mohammad T., Seryakov A., Sekacheva M., Buzdin A.* // Ther. Adv. Med. Oncol. 2022. V. 14. P. 108. https://doi.org/10.1177/17588359221144108
11. *Salokas K., Dashi G., Varjosalo M.* // Cancers (Basel). 2023. V. 15. P. 3678. https://doi.org/10.3390/cancers15143678
12. *Stransky N., Cerami E., Schalm S., Kim J.L., Lengauer C.* // Nat. Commun. 2014. V. 5. P. 4846. https://doi.org/10.1038/ncomms5846
13. *Salokas K., Weldatsadik R.G., Varjosalo M.* // Sci. Rep. 2020. V. 10. P. 14169. https://doi.org/10.1038/s41598-020-71040-8
14. *Chu Y.-H.* // Surg. Pathol. Clin. 2023. V. 16. P. 57–73. https://doi.org/10.1016/j.path.2022.09.007
15. *Nagy Z., Jeselsohn R.* // Front. Oncol. 2022. V. 12. P. 1037531. https://doi.org/10.3389/fonc.2022.1037531
16. *Apfelbaum A.A., Wrenn E.D., Lawlor E.R.* // Front. Oncol. 2022. V. 12. P. 1044707. https://doi.org/10.3389/fonc.2022.1044707
17. *Bowling G.C., Rands M.G., Dobi A., Eldhose B.* // Mol. Cancer Ther. 2023. V. 22. P. 168–178. https://doi.org/10.1158/1535-7163.MCT-22-0527
18. *Shen Z., Qiu B., Li L., Yang B., Li G.* // Front. Oncol. 2022. V. 12. P. 1033484. https://doi.org/10.3389/fonc.2022.1033484
19. *Dobin A., Davis C.A., Schlesinger F., Drenkow J., Zaleski C., Jha S., Batut P., Chaisson M., Gingeras T.R.* // Bioinformatics. 2013. V. 29. P. 15–21. https://doi.org/10.1093/bioinformatics/bts635
20. *Петров С.Н., Урошлев Л.А., Касьянов А.С., Макеев В.Ю.* // Мол. биофизика. 2018. Т. 63. С. 421–429.
21. *Haas B.J., Dobin A., Li B., Stransky N., Pochet N., Regev A.* // Genome Biol. 2019. V. 20. P. 213. https://doi.org/10.1186/s13059-019-1842-22
22. *Nurk S., Bankevich A., Antipov D., Gurevich A.A., Korobeynikov A., Lapidus A., Prjibelski A.D., Pyshkin A., Sirotkin A., Sirotkin Y., Stepanauskas R., Clingenpeel S.R., Woyke T., McLean J.S., Lasken R., Tesler G., Alekseyev M.A., Pevzner P.A.* // J. Comput. Biol. 2013. V. 20. P. 714–737. https://doi.org/10.1089/cmb.2013.0084
23. *Benoit-Pilven C., Marchet C., Chautard E., Lima L., Lambert M.-P., Sacomoto G., Rey A., Cologne A., Terrone S., Dulaurier L., Claude J.-B., Bourgeois C.F., Auboeuf D., Lacroix V.* // Sci. Rep. 2018. V. 8. P. 4307. https://doi.org/10.1038/s41598-018-21770-7
24. *Haas B., Dobin A., Stransky N., Li B., Yang X., Tickle T., Bankapur A., Ganote C., Doak T., Pochet N., Sun J., Wu C., Gingeras T., Regev A.* // BioRxiv. 2017. P. 120295. https://doi.org/10.1101/120295
25. *Križanovic K., Echchiki A., Roux J., Šikic M.* // Bioinformatics. 2018. V. 34. P. 748–754. https://doi.org/10.1093/bioinformatics/btx668
26. *Chen Y., Ye W., Zhang Y., Xu Y.* // Nucleic Acids Res. 2015. V. 43. P. 7762–7768., https://doi.org/10.1093/nar/gkv784
27. *Conesa A., Madrigal P., Tarazona S., Gomez-Cabrero D., Cervera A., McPherson A., Szcześniak M.W., Gaffney D.J., Elo L.L., Zhang X., Mortazavi A.* // Genome Biol. 2016. V. 17. P. 13. https://doi.org/10.1186/s13059-016-0881-8
28. *Uhrig S., Ellermann J., Walther T., Burkhardt P., Fröhlich M., Hutter B., Toprak U.H., Neumann O., Stenzinger A., Scholl C., Fröhling S., Brors B.* // Genome Res. 2021. V. 31. P. 448–460. https://doi.org/10.1101/gr.257246.119
29. *Uhlén M., Fagerberg L., Hallström B.M., Lindskog C., Oksvold P., Mardinoglu A., Sivertsson Å., Kampf C., Sjöstedt E., Asplund A., Olsson I., Edlund K., Lundberg E., Navani S., Szigyarto C.A., Odeberg J., Djureinovic D., Takanen J.O., Hober S., Alm T., Pontén F.* // Science. 2015. V. 347. P. 1260419. https://doi.org/10.1126/science.1260419
30. *Barbosa-Morais N.L., Irimia M., Pan Q., Xiong H.Y., Gueroussov S., Lee L.J., Slobodeniuc V., Kutter C., Watt S., Colak R., Kim T., Misquitta-Ali C.M., Wilson M.D., Kim P.M., Odom D.T., Frey B.J., Blencowe B.J.* // Science. 2012. V. 338. P. 1587–1593. https://doi.org/10.1126/science.1230612

31. Expression Atlas. RNA-Seq of human individual tissues and mixture of 16 tissues (Illumina Body Map). https://www.ebi.ac.uk/gxa/experiments/E-MTAB-513/Results
32. ENCODE Project Consortium // A User's Guide to the Encyclopedia of DNA Elements (ENCODE) // PLoS Biol. 2011. V. 9. P. e1001046. https://doi.org/10.1371/journal.pbio.1001046
33. *Roadmap Epigenomics Consortium, Kundaje A., Meuleman W., Ernst J., Bilenky M., Yen A., Heravi-Moussavi A., Kheradpour P., Zhang Z., Wang J., Ziller M.J., Amin V., Whitaker J.W., Schultz M.D., Ward L.D., Sarkar A., Quon G., Sandstrom R.S., Eaton M.L., Wu Y.-C., Kellis M.* // Nature. 2015. V. 518. P. 317–330. https://doi.org/10.1038/nature14248
34. *Jahn A., Rump A., Widmann T.J., Heining C., Horak P., Hutter B., Paramasivam N., Uhrig S., Gieldon L., Drukewitz S., Kübler A., Bermudez M., Hackmann K., Porrmann J., Wagner J., Arlt M., Franke M., Fischer J., Kowalzyk Z., William D., Klink B.* // Ann. Oncol. 2022. V. 33. P. 1186–1199. https://doi.org/10.1016/j.annonc.2022.07.008
35. Arriba. Documentation: workflow, internal algorithm, visualization. https://arriba.readthedocs.io/en/latest/visualization/
36. *Chiu R., Nip K.M., Birol I.* // Bioinformatics. 2020. V. 36. P. 2256–2257. https://doi.org/10.1093/bioinformatics/btz902
37. *Nip K.M., Chiu R., Yang C., Chu J., Mohamadi H., Warren R.L., Birol I.* // BioRxiv. 2019. P. 701607. https://doi.org/10.1101/701607
38. PAVFinder – Post Assembly Variants Finder (Github). https://github.com/bcgsc/pavfinder
39. *Quinlan A.R., Hall I.M.* // Bioinformatics. 2010. V. 26. P. 841–842. https://doi.org/10.1093/bioinformatics/btq033
40. *Aaron R. Quinlan, Ira M.* // Hall. Bedtools 2.31.0 // BEDTools_documentation. BEDPE Format. 2010. https://bedtools.readthedocs.io/en/latest/content/general-usage.html#bedpe-format
41. *Bray N.L., Pimentel H., Melsted P., Pachter L.* // Nat. Biotechnol. 2016. V. 34. P. 525–527. https://doi.org/10.1038/nbt.3519
42. *Melsted P., Hateley S., Joseph I.C., Pimentel H., Bray N., Pachter L.* // bioRxiv. 2017. P. 166322. https://doi.org/10.1101/166322
43. *Frankish A., Diekhans M., Jungreis I., Lagarde J., Loveland J.E., Mudge J.M., Sisu C., Wright J.C., Armstrong J., Barnes I., Berry A., Bignell A., Boix C., Carbonell Sala S., Cunningham F., Di Domenico T., Donaldson S., Fiddes I.T., García Girón C., Gonzalez J.M., Flicek P.* // Nucleic Acids Res. 2021. V. 49. P. D916–D923. https://doi.org/10.1093/nar/gkaa1087
44. *Davidson N.M., Majewski I.J., Oshlack A.* // Genome Med. 2015. V. 7. P. 43. https://doi.org/10.1186/s13073-015-0167-x
45. *Kent W.J.* // Genome Res. 2002. V. 12. P. 656–664. https://doi.org/10.1101/gr.229202
46. *Schulz M.H., Zerbino D.R., Vingron M., Birney E.* // Bioinformatics. 2012. V. 28. P. 1086–1092. https://doi.org/10.1093/bioinformatics/bts094
47. *Zerbino D.R., Birney E.* // Genome Res. 2008. V. 18. P. 821–829. https://doi.org/10.1101/gr.074492.107
48. *Hon T., Mars K., Young G., Tsai Y.-C., Karalius J.W., Landolin J.M., Maurer N., Kudrna D., Hardigan M.A., Steiner C.C., Knapp S.J., Ware D., Shapiro B., Peluso P., Rank D.R.* // Sci. Data. 2020. V. 7. P. 399. https://doi.org/10.1038/s41597-020-00743-4
49. *Logsdon G.A., Vollger M.R., Eichler E.E.* // Nat. Rev. Genet. 2020. V. 21. P. 597–614. https://doi.org/10.1038/s41576-020-0236-x
50. *Kasianowicz J.J., Brandin E., Branton D., Deamer D.W.* // Proc. Natl. Acad. Sci. USA. 1996. V. 93. P. 13770–13773. https://doi.org/10.1073/pnas.93.24.13770
51. *Davidson N.M., Chen Y., Sadras T., Ryland G.L., Blombery P., Ekert P.G., Göke J., Oshlack A.* // Genome Biol. 2022. V. 23. P. 10. https://doi.org/10.1186/s13059-021-02588-5
52. *Sadedin S.P., Pope B., Oshlack A.* // Bioinformatics. 2012. V. 28. P. 1525–1526. https://doi.org/10.1093/bioinformatics/bts167
53. *Li H.* // Bioinformatics. 2018. V. 34. P. 3094–3100. https://doi.org/10.1093/bioinformatics/bty191
54. *Harrow J., Frankish A., Gonzalez J.M., Tapanari E., Diekhans M., Kokocinski F., Aken B.L., Barrell D., Zadissa A., Searle S., Barnes I., Bignell A., Boychenko V., Hunt T., Kay M., Mukherjee G., Rajan J., Despacio-Reyes G., Saunders G., Steward C., Hubbard T.J.* // Genome Res. 2012. V. 22. P. 1760–1774. https://doi.org/10.1101/gr.135350.111
55. *Lei Q., Li C., Zuo Z., Huang C., Cheng H., Zhou R.* // Genome Biol. Evol. 2016. V. 8. P. 562–577. https://doi.org/10.1093/gbe/evw025
56. *Molania R., Foroutan M., Gagnon-Bartsch J.A., Gandolfo L.C., Jain A., Sinha A., Olshansky G., Dobrovic A., Papenfuss A.T., Speed T.P.* // Nat. Biotechnol. 2023. V. 41. P. 82–95. https://doi.org/10.1038/s41587-022-01440-w

57. *Dorney R., Dhungel B.P., Rasko J.E.J., Hebbard L., Schmitz U.* // Brief. Bioinformatics. 2023. V. 24. https://doi.org/10.1093/bib/bbac519
58. *Liu Q., Hu Y., Stucky A., Fang L., Zhong J.F., Wang K.* // BMC Genomics. 2020. V. 21. P. 793. https://doi.org/10.1186/s12864-020-07207-4
59. *Chen Y., Wang Y., Chen W., Tan Z., Song Y., Human Genome Structural Variation Consortium, Chen H., Chong Z.* // Cancer Res. 2023. V. 83. P. 28–33. https://doi.org/10.1158/0008-5472.CAN-22-1628
60. *Ester M., Kriegel H.-P., Sander J., Xu X.A.* // KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. 1996. P. 226–231. https://dl.acm.org/doi/10.5555/3001460.3001507
61. GitHub – ruanjue/bsalign: Banded Striped DNA Sequence Alignment. https://github.com/ruanjue/bsalign
62. Illumina Online Support Service – RNAseq Analysis Methods – STAR. https://support.illumina.com/help/BS_App_RNA-Seq_Alignment_OLH_1000000006112/Content/Source/Informatics/STAR_RNAseq.htm
63. *Alser M., Rotman J., Deshpande D., Taraszka K., Shi H., Baykal P.I., Yang H.T., Xue V., Knyazev S., Singer B.D., Balliu B., Koslicki D., Skums P., Zelikovsky A., Alkan C., Mutlu O., Mangul S.* // Genome Biol. 2021. V. 22. P. 249. https://doi.org/10.1186/s13059-021-02443-7
64. *Jain M., Koren S., Miga K.H., Quick J., Rand A.C., Sasani T.A., Tyson J.R., Beggs A.D., Dilthey A.T., Fiddes I.T., Malla S., Marriott H., Nieto T., O'Grady J., Olsen H.E., Pedersen B.S., Rhie A., Richardson H., Quinlan A.R., Snutch T.P., Loose M.* // Nat. Biotechnol. 2018. V. 36. P. 338–345. https://doi.org/10.1038/nbt.4060
65. *Merker J.D., Wenger A.M., Sneddon T., Grove M., Zappala Z., Fresard L., Waggott D., Utiramerur S., Hou Y., Smith K.S., Montgomery S.B., Wheeler M., Buchan J.G., Lambert C.C., Eng K.S., Hickey L., Korlach J., Ford J., Ashley E.A.* // Genet. Med. 2018. V. 20. P. 159–163. https://doi.org/10.1038/gim.2017.86
66. *Carrara M., Beccuti M., Lazzarato F., Cavallo F., Cordero F., Donatelli S., Calogero R.A.* // Biomed Res. Int. 2013. V. 2013. P. 340620. https://doi.org/10.1155/2013/340620
67. *Kumar S., Razzaq S.K., Vo A.D., Gautam M., Li H.* // Wiley Interdiscip. Rev. RNA. 2016. V. 7. P. 811–823. https://doi.org/10.1002/wrna.1382
68. *Suntsova M., Gaifullin N., Allina D., Reshetun A., Li X., Mendeleeva L., Surin V., Sergeeva A., Spirin P., Prassolov V., Morgan A., Garazha A., Sorokin M., Buzdin A.* // Sci. Data. 2019. V. 6. P. 36. https://doi.org/10.1038/s41597-019-0043-4
69. *Yi Q.-Q., Yang R., Shi J.-F., Zeng N.-Y., Liang D.-Y., Sha S., Chang Q.* // J. Int. Med. Res. 2020. V. 48. P. 1259. https://doi.org/10.1177/0300060520931259
70. *Langmead B., Salzberg S.L.* // Nat. Methods. 2012. V. 9. P. 357–359. https://doi.org/10.1038/nmeth.1923
71. *Rabushko E., Sorokin M., Suntsova M., Seryakov A.P., Kuzmin D.V., Poddubskaya E., Buzdin A.A.* // Biomedicines. 2022. V. 10. P. 1866. https://doi.org/10.3390/biomedicines10081866
72. The Harmonizome 3.0: Integrated Knowledge about Genes and Proteins. https://maayanlab.cloud/Harmonizome/about
73. *Rouillard A.D., Gundersen G.W., Fernandez N.F., Wang Z., Monteiro C.D., McDermott M.G., Ma'ayan A.* // Database (Oxford). 2016. V. 2016. P. baw100. https://doi.org/10.1093/database/baw100
74. *Borisov N., Buzdin A.* // Biomedicines. 2022. V. 10. P. 2318. https://doi.org/10.3390/biomedicines10092318
75. *Tembe W.D., Pond S.J., Legendre C., Chuang H.Y., Liang W.S., Kim N.E., Montel V., Wong S., McDaniel T.K., Craig D.W., Carpten J.D.* // BMC Genomics. 2014. V. 15. P. 824. https://doi.org/10.1186/1471-2164-15-824
76. *Wick R.R.* // J. Open Source Software. 2019. V. 4. P. 1316. https://doi.org/10.21105/joss.01316
77. *Yukiteru O., Kiyoshi A., Michiaki H.* // Bioinformatics. 2013. V. 29. P. 119–121. https://doi.org/10.1093/bioinformatics/bts649

Bioinformatic Approaches for Detection of Fusion Genes and *Trans*-Splicing Products

I. Yu. Musatov*, **, #, M. I. Sorokin, and A. A. Buzdin*, ***, ******

 # Phone: +7 (985) 147-97-18; e-mail: musatov.mailbox@yandex.ru

** Moscow Institute of Physics and Technology, Institutskiy per. 9, Dolgoprudniy, 141701 Russia*
*** Institute for Personalized Oncology of World-Class Research Center "Digital Biodesign and Personalized Healthcare", Federal State Autonomous Educational Institution of Higher Education I.M. Sechenov First Moscow State Medical University of the Ministry of Health of the Russian Federation (Sechenov University), ul. Trubetskaya 8/2, Moscow, 119048 Russia*
**** Shemyakin–Ovchinnikov Institute of Bioorganic Chemistry of the Russian Academy of Sciences, ul. Miklukho-Maklaya 16/10, Moscow, 117997 Russia*
***** Endocrinology Research Centre, ul. Dm. Ulyanova 11, Moscow, 117292 Russia*

Chimeric genes and transcripts can be biological markers as well as the reasons for tumor progression and development. Modern algorithms and high-throughput sequencing are the complementary clues to the question of the tumor origin and cancer detection as well as to the fundamental question of chimeric genes origin and their influence on molecular processes of the cell. A wide-range of algorithms for chimeric genes detection was developed, with various differences in computing speed, sensitivity, specificity, and focus on the experimental design. There exist three main types of bioinformatic approaches, which act according to the sequencing read length. Algorithms, which focus on short-read high-throughput sequencing (about 50–300 bp of read length) or long-read sequencing (about 5000–100000 bp of read length) exclusively or algorithms, which combine the results of both short and long-read sequencing. These algorithms are further subdivided into: 1) mapping-first approaches (STAR-Fusion, Arriba), which map reads to the genome or transcriptome directly and search the reads supporting the fused gene or transcript; 2) assembly-first approaches (Fusion-Bloom), which assemble the genome or transcriptome from the overlapping reads, and then compare the results to the reference transcriptome or genome to find transcripts or genes not present in the reference and therefore raising questions; 3) pseudoalignment approaches, which do not make local alignment, but just search for the closest transcript subsequence to the reads seed, following the precomputed index for all reference transcripts and provides the results. This article describes the main classes of available software tools for chimeric gene detection, provides the characteristics of these programs, their advantages and disadvantages. To date the most resource intensive and slowest are still assembly-first algorithms. Mapping-first approaches are quite fast and rather accurate at fusion detection, still the fastest and resource-saving are the pseudoalignment algorithms, but, worth noting, that the quick search is carried out at the expense of chimeras search quality decrease.

Keywords: RNAseq, chimeric genes, chimeric transcripts, tumor, FFPE samples, pseudoalignment, genome de novo assembly, transcriptome de novo assembly, trans-splicing