ГЕНЕТИКА, 2023, том 59, № 6, с. 687–707

МАТЕМАТИЧЕСКИЕ МОДЕЛИ И МЕТОДЫ

УДК 575.224.23

СРАВНИТЕЛЬНЫЙ АНАЛИЗ ИНСТРУМЕНТОВ ДЛЯ ПОИСКА СТРУКТУРНЫХ ВАРИАНТОВ НА ДАННЫХ КОРОТКОРИДОВОГО ПОЛНОГЕНОМНОГО СЕКВЕНИРОВАНИЯ

© 2023 г. А. А. Мкртчян[1, *, #], К. С. Грамматикати[1, #], П. Г. Казакова[1, #], С. И. Митрофанов[1, #], П. Ю. Земский[1], А. А. Ивашечкин[1], М. Н. Пилипенко[1], Д. В. Светличный[1], А. П. Сергеев[1], Е. А. Снигирь[1], Л. В. Фролова[1], Т. А. Шпакова[1], В. С. Юдин[1], А. А. Кескинов[1], С. М. Юдин[1], В. И. Скворцова[2]

[1]Федеральное государственное бюджетное учреждение Центр стратегического планирования и управления медико-биологическими рисками здоровью Федерального медико-биологического агенства России, Москва, 119121 Россия

[2]Федеральное медико-биологическое агентство России, Москва, 123182 Россия

**e-mail: AMkrtchyan@cspfmba.ru*

Поступила в редакцию 25.10.2022 г.
После доработки 16.01.2023 г.
Принята к публикации 19.01.2023 г.

Проанализирована работа трех биоинформатических инструментов (Manta, Smoove, Delly) для поиска структурных вариантов на данных полногеномного секвенирования с использованием четырех разных алгоритмов выравнивания: DRAGEN, GDC DNA-SeqAlignment Workflow, GDC DNA-SeqAlignmentWorkflow + GDC DNA-SeqCo-cleaning Workflow, NovoAlign, разных длин “сырых” прочтений: 2 × 150 и 2 × 250 пн, разных показателей среднего покрытия генома. Полученные показания соотнесены с эталонными результатами команды GIAB. Проведена проверка выявленных структурных вариантов с помощью секвенирования по методу Сэнгера. Выявили, что структурные варианты типа делеция и инсерция лучше всего определяются с помощью инструмента Manta. Мы получили 89–96% точности и 59–70% чувствительности для проанализированных делеций, а также 96–99% точности и 15–36% чувствительности для инсерций. Smoove и Delly продемонстрировали менее точные и чувствительные результаты (Smoove: 91–95% точности и 8–54% чувствительности для делеций; Delly: 78–87% точности и 31–66% чувствительности для делеций, 99–100% точности и 1–13% чувствительности для инсерций). Показали, что использование комбинации из двух и даже трех инструментов не повышает процента подтвержденных структурных вариантов для типов делеция, инсерция, дупликация. Проведенный анализ выявил, что с увеличением среднего значения покрытия генома увеличиваются показатели точности и чувствительности работы биоинформатических инструментов для поиска структурных вариантов, а увеличение длины прочтений с 150 до 250 пн влияет в разной степени на точность и чувствительность работы отдельных инструментов. Также сделали вывод, что точность работы инструментов для поиска структурных вариантов варьируется в зависимости от диапазона размеров структурных вариантов. Например, Manta лучше находит делеции в диапазоне от 200 и более пн, Delly – от 1000 до 10000 пн, Smoove – от 200 до 10000 пн.

Ключевые слова: структурные варианты, полногеномное секвенирование, секвенирование на коротких прочтениях, выравнивание, GIAB.

DOI: 10.31857/S0016675823060115, **EDN:** SUEMFU

Однонуклеотидные полиморфизмы (SNP) являются наиболее изученными генетическими вариантами. Клиническая значимость именно однонуклеотидных генетических вариантов чаще всего интерпретируется в рамках работы различных герминальных и соматических пайплайнов. Первоначально считалось, что именно SNP вносят наиболее существенный вклад в генетическое разнообразие людей, обеспечивая геномную вариативность между двумя индивидуумами около 0.1%. Однако сейчас установлено, что структурные варианты в геноме человека также дают значительный вклад в генетическое разнообразие, увеличивая геномную вариативность до 1.5% [1, 2].

Структурными вариантами называются генетические перестройки длиной не менее 50 пн [3]. Структурные варианты делятся на простые (делеции, дупликации, инсерции, инверсии, трансло-

[#] Вклад этих авторов в работу равнозначный.

кации) и комплексные, к которым относятся, например, хромоплексия и хромотрипсис.

Выделяют отдельный подтип структурных вариантов – вариации числа копий (CNV), при котором количество копий некоторого участка ДНК варьирует в геномах разных индивидуумов [4]. CNV образуются в основном за счет делеций и дупликаций участков ДНК размером более 1000 пн [5–8].

По другому основанию структурные варианты делятся на сбалансированные и несбалансированные. Несбалансированные представляют собой увеличение или уменьшение количества фрагментов ДНК. Сбалансированные структурные варианты возникают в результате генетической перестройки без изменения количества копий участков ДНК, что характерно для инверсий и транслокаций [9].

Научные исследования в области молекулярной биологии и медицинской генетики показали, что структурные варианты играют значимую роль в развитии ряда фенотипических нарушений [10, 11] и таких тяжелых заболеваний, как синдром Паркинсона [12], склерозирующая пневмоцитома [13], гематологические опухоли [14], рак груди и яичников [15]. Кроме того, многие заболевания, в том числе шизофрения [16] и спорадический рак простаты [17], ассоциированы с наличием в геноме CNV [18]: некоторые нейродегенеративные заболевания, например повышенный риск развития аутоиммунных заболеваний (системная красная волчанка [19], болезнь Крона [20], псориаз [21] и др.) также связан с наличием вариаций числа копий в геноме человека.

Установление перечисленных выше ассоциаций и зависимостей привело исследователей к выводу о том, что понимание патогенеза многих трудноизлечимых заболеваний невозможно без изучения и генотипирования структурных вариантов.

В настоящее время исследования генома ведутся в основном на базе более доступного секвенирования с помощью коротких прочтений, хотя технологии секвенирования с помощью длинных прочтений позволяют более детально изучать структурные варианты.

Данные секвенирования используются в первую очередь для выявления новых и идентификации существующих однонуклеотидных и структурных вариантов биоинформатическими методами. Этот процесс осложняется тем, что многие структурные варианты встречаются в тех регионах генома, которые богаты повторами. При этом точность идентификации вариантов во многом зависит от выравнивания прочтений, качество которого может снижаться в таких сложных регионах, как участки с тандемными повторами, теломерные участки хромосом и др. [5]. Соответственно глубокое изучение структурных вариантов невозможно без совершенствования биоинформатических инструментов, предназначенных для поиска структурных вариантов и внедрения их в пайплайны.

Команда GIAB опубликовала эталонный файл для выявления структурных вариантов [22]. Этот файл в формате VCF был получен для образца HG002 на основе данных секвенирования с помощью четырех технологий при использовании более 30 биоинформатических инструментов. Используя этот материал, мы смогли оценить полученные нами данные о качестве работы трех биоинформатических инструментов: Manta, Smoove и Delly. Кроме того, мы провели проверку с помощью секвенирования по методу Сэнгера нескольких структурных вариантов (разных типов и длин) на биоматериале образцов от двух доноров.

МАТЕРИАЛЫ И МЕТОДЫ

Сбор биоматериала

Для данного исследования взяты образцы цельной крови от двух доноров (не HG002), подписавших добровольное информированное согласие. Забор биоматериала осуществлялся в соответствии с ГОСТ Р53079.4-2008. Транспортировка биоматериала велась при постоянном контроле температурного режима. Образцы прошли проверки качества, включая проверку на отсутствие признаков гемолиза и хилеза. Биоматериал подвергался заморозке не более одного раза.

Выделение ДНК, подготовка геномных библиотек и секвенирование образцов с применением технологии Illumina

Выделение ДНК из образцов цельной крови проводили с помощью набора MagAttract HMW DNA Kit (Qiagen, Германия). Протокол выделения ДНК автоматизирован на станции Tecan Freedom EVO (Tecan, Швейцария). Измерение концентрации и чистоты выделенной ДНК осуществляли с помощью микропланшетного ридера Tecan Infinite 200 PRO F Nano+ (Tecan).

Геномные библиотеки для секвенирования готовились с использованием набора Nextera DNA Flex (Illumina, США) в соответствии с рекомендациями производителя. Каждый образец в проточной кювете был помечен с помощью индексов IDT-ILMN Nextera DNA UD (Illumina) для исключения возможности перекрестной контаминации.

Измерение концентрации геномных библиотек проводили на микропланшетном ридере Tecan Infinite 200 PRO F Nano+. Размер вставки геномных библиотек измеряли на системе Agilent 4200 TapeStation System (Agilent, США) с использованием набора Agilent DNA 1000 Kit (Agilent).

Библиотечные пулы включали помимо двух целевых образцов 22 образца из других исследований, так как одна ячейка S4 используемого секвенатора вмещает 24 образца, объединение которых осуществлялось с использованием автоматизированной станции Tecan Freedom EVO (Tecan). Размер вставки для каждого образца составил 323 и 333 пн.

Полногеномное секвенирование проводилось с использованием секвенатора Illumina NovaSeq 6000 (Illumina) и комплекта реагентов NovaSeq 6000 S4 (300 циклов) (Illumina) для парноконцевых прочтений 2 × 150 пн.

Секвенирование образцов по методу Сэнгера

Выделение геномной ДНК проводили из 200 мкл замороженной цельной крови с использованием набора MagAttract HMW DNA Kit в соответствии с протоколом производителя. Элюция проводилась в 100 мкл воды, свободной от ДНКаз.

Концентрацию образцов измеряли с помощью набора QuantiFluor dsDNA System (Promega, США) на флуориметре Quantus (Promega). Измерение чистоты выделенной ДНК проводилось с использованием спектрофотометра NanoDrop8000 (Thermo Scientific, США). Отбирались образцы ДНК, имеющие соотношения длин волны A260/280 в пределах 1.7–1.9 и A230/260 в пределах 1.8–2.2.

Для секвенирования по методу Сэнгера была сформирована выборка из 120 локусов с различными типами структурных вариантов. В выборку вошли 60 делеций (50%), 28 инсерций (23.3%), 23 дупликации (19.2%) и 9 инверсий (7.5%).

С помощью программного обеспечения Primer-BLAST [23] удалось подобрать праймеры к 17 делециям, 12 инсерциям, 12 дупликациям и двум инверсиям. Праймеры подобраны отдельно к прямой и обратной цепи ДНК.

Для синтеза праймеров использовались олигонуклеотидные синтезаторы OligoMaker X12 (OligoMaker ApS, Дания) и Polygen (Polygen GmbH, Германия).

Для наработки ампликонов проводили ПЦР со следующим температурным циклом: начальная денатурация 95°C – 5 мин; 35 циклов 95°C – 15 с, 60°C – 30 с, 72°C – 30 с; финальная элонгация 72°C – 5 мин; хранение при 4°C – неограниченное время.

Для проведения ПЦР отбиралось 50–100 нг ДНК. После завершения ПЦР проводилась очистка ампликонов с помощью магнитных частиц Sample Purification Beads (Beckman Coulter, США) по следующей схеме:

– к 25 мкл образца добавляется 25 мкл частиц;

– смесь перемешивается и инкубируется в течение 10 мин при комнатной температуре;

– смесь ставится на магнитный штатив в ожидании просветления;

– после просветления супернатант удаляется;

– осадок дважды промывается 200 мкл 80%-ного этилового спирта;

– промытый осадок сушится на воздухе в течение 3–5 мин и элюируется 25 мкл Resuspension Buffer (Illumina);

– элюированный осадок инкубируется при комнатной температуре 2 мин, ставится на магнитный штатив;

– очищенные ампликоны переносятся в новые пробирки в объеме 20 мкл.

Контроль очистки проводили с помощью Agilent 4200 TapeStation System и набора реагентов D1000 Reagents (Agilent, США) согласно инструкции производителя.

Секвенирующая и терминирующая реакции проводились с использованием набора реагентов BigDye Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems, США) согласно рекомендациям производителя. В каждую реакционную пробирку добавлялось 50–100 нг ДНК-матрицы образца.

Продукты реакции анализировали методом капиллярного электрофореза в автоматическом секвенаторе SeqStudio Genetic Analyzer (Applied Biosystems). При последующем анализе учитывали только те варианты, для которых были получены прямые и обратные прочтения.

Биоинформатическая обработка данных полногеномного секвенирования

Для контроля качества секвенирования в целом использовалась программа Illumina Sequencing Analysis Viewer v2.4.7 [24].

На первом этапе обработки данных полногеномного секвенирования осуществлялась демультипликация, при которой исходная выдача секвенатора Illumina NovaSeq 6000 конвертировалась из формата BCL в формат FASTQ с помощью программного обеспечения bcl2fastq v2.20 [25]. Для контроля качества прочтений применялся биоинформатический инструмент FastQC v0.11.9 [26]. Качество анализируемых образцов провере-

но по следующим метрикам: показатель равномерности распределения нуклеотидов в прочтениях; G/C состав; показатель Q30.

На втором этапе биоинформатической обработки проводилось выравнивание прочтений на референсный геном с помощью программно-аппаратной платформы Illumina DRAGEN Bio-IT Platform v07.021.510.3.5.7 [27]. В качестве референсного генома человека использовалась последовательность GRCh38.d1.vd1 [28]. После выравнивания оценивалось среднее значение покрытия. У всех анализируемых образцов среднее покрытие по геному составило не менее 30×.

Затем в файлах выравнивания осуществляли поиск структурных вариантов с помощью трех различных инструментов: Manta v1.6.0 [29], Smoove v0.2.6 [30] и Delly v0.9.1 [31].

Так как эталонные VCF- и BED-файлы GIAB сделаны для версии генома GRCh37 [32], а данный анализ проводился с выравниванием на версию генома GRCh38, то для того, чтобы отфильтровать полученные VCF-файлы по подтвержденным регионам эталонного BED-файла, его координаты конвертировались с GRCh37 на GRCh38 с помощью инструмента CrossMap [33].

Для проверки генетических вариантов секвенированием по методу Сэнгера отбирались делеции разной длины, а также дупликации и инверсии длиной до 500 пн, которые были найдены всеми тремя инструментами или двумя из них. Кроме того, отбирались инсерции длиной до 500 пн, которые были найдены хотя бы одним из инструментов Manta и Delly.

Биоинформатическая обработка данных HG002 и сравнение с эталоном

Для валидации качества работы инструментов, предназначенных для поиска структурных вариантов, с сайта NCBI были скачаны эталонные файлы: `HG002_SVs_Tier1_v0.6.VCF.GZ`, `HG002_SVs_Tier1_v0.6.VCF.GZ.TBI` и `HG002_SVs_Tier1_v0.6.BED` (https://ftp.ncbi.nih.gov/giab/ftp/data/AshkenazimTrio/analysis/NIST_SVs_Integration_v0.6/), а также прочтения образца HG002, полученные с помощью секвенатора Illumina HiSeq (2 × 250 пн) в формате FASTQ.GZ: https://ftp.ncbi.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG002_NA24385_son/NIST_Illumina_2x250bps/reads/, 68 файлов общим объемом 163 Гб, предполагаемое среднее покрытие – 68x. Размер вставки составил 350 пн.

Также были скачаны часть прочтений образца HG002, полученная с помощью секвенатора Illumina HiSeq (2 × 150 пн) в формате FASTQ.GZ: https://ftp.ncbi.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG002_NA24385_son/NIST_HiSeq_HG002_Homogeneity-10953946/HG002_HiSeq300x_fastq/, три первые по алфавиту директории, которые далее по тексту будем обозначать d1, d2, d3, общим объемом 174 ГБ, предполагаемое среднее покрытие для каждой директории – 25x.

Затем были подготовлены директории с объединенными прочтениями: d1 + d2 с предполагаемым средним покрытием 50x и d1 + d2 + d3 с предполагаемым средним покрытием 75x.

Дополнительно был скачан BAM-файл с выравниванием прочтений образца HG002, полученных с секвенатора Illumina HiSeq (2 × 250 пн), инструментом NovoAlign на 37-ю версию референсного генома человека (hs37d5): https://ftp.ncbi.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG002_NA24385_son/NIST_Illumina_2x250bps/novoalign_bams/HG002.hs37d5.2x250.bam, 122 ГБ, среднее покрытие 72x.

Размер вставки для данных прочтений составил 550 пн.

Для скачанных прочтений проведено выравнивание на референсный геном GRCh37 тремя способами: с помощью программно-аппаратной платформы DRAGEN, с помощью пайплайна GDC DNA-Seq Alignment Workflow (DNA-Seq A), а также с помощью пайплайна GDC DNA-Seq Alignment + Co-cleaning Workflow (DNA-Seq A + C) [34], в котором для выравнивания используется программа BWA [35].

Для скачанного BAM-файла выравнивания, выполненного инструментом NovoAlign, проведена сортировка и индексация с помощью инструмента samtools [36].

Во всех файлах с выравниваниями осуществлен поиск структурных вариантов с помощью инструментов Manta v1.6.0, Smoove v0.2.6 и Delly v0.9.1. Затем проведена дополнительная обработка результирующего VCF-файла. Для файлов Manta это была конвертация структурных вариантов типа BND в INV с помощью скрипта разработчиков данного инструмента, а также дополнительная аннотация инструментом Duphold v0.2.3 [37] с последующей фильтрацией в соответствии с рекомендациями разработчика (Приложение 1). Кроме того, для всех файлов проведена фильтрация с помощью bcftools 1.15 [36] (Приложение 1).

Таблица 1. Структурные варианты в эталонном файле HG002_SVs_Tier1_v0.6.vcf.gz

Количество вариантов										
до фильтрации, ед.	после фильтрации PASS									
	всего		в том числе типа:							
	ед.	%	DEL		INS		DUP		INV	
			ед.	%	ед.	%	ед.	%	ед.	%
74012	12745	100	5464	43	4462	35	2819	22	0	0

Примечание. DEL – делеции, INS – инсерции, DUP – дупликации, INV – инверсии.

При этом для Smoove по умолчанию брались варианты без фильтрации по полю FILTER, так как инструмент не проводит соответствующей фильтрации.

Сводные VCF-файлы готовили с помощью инструмента SURVIVOR v1.0.7 [38]. Для сравнения всех полученных VCF-файлов с эталонным VCF-файлом использовался инструмент witty.er v0.3.4 [39], конфигурационный файл для которого скачан из репозитория инструмента на ресурсе GitHub – https://github.com/Illumina/witty.er/blob/master/docs/examples/SV/config.json.

Команды запуска инструментов bcftools, SURVIVOR и witty.er представлены в Приложении 1.

Биоинформатическая обработка данных секвенирования по методу Сэнгера

Файлы прочтений прямой и обратной последовательностей в формате AB1 сконвертированы в формат FASTQ с помощью программы Unipro UGENE [40]. Далее эти прочтения выравнены на референсный геном GRCh38.d1.vd1 с помощью инструмента BWA, а полученные выравнивания проанализированы в программе IGV [41], предназначенной для просмотра генетических данных. К выравниванию применялось окрашивание soft-clipped-прочтений, т.е. таких, у которых часть прочтения выравнивается на референс, а часть нет, и при этом прочтение остается в выравнивании. Кроме того, использовалось окрашивание прочтений по последовательностям. Визуализация варианта помогала сделать вывод о наличии или отсутствии структурного варианта в образце.

РЕЗУЛЬТАТЫ

Валидация работы инструментов поиска структурных вариантов

Командой GIAB опубликован эталонный набор данных структурных вариантов в формате VCF для версии генома GRCh37, а также BED-файл для образца HG002 на основе данных секвенирования с помощью четырех разных технологий. Для оценки качества работы биоинформатических инструментов, предназначенных для поиска структурных вариантов, с сайта NCBI скачаны указанные выше эталонные файлы. Эти файлы использовались для сравнения с полученными данными. В табл. 1 представлена статистика по вариантам эталонного VCF-файла Tier1_v0.6.

Как следует из табл. 1, в эталонном файле GIAB после фильтрации по фильтру “PASS” и по типу вариантов “не равно BND” (breakends) остается немногим более 17% структурных вариантов, большинство из которых (около 43%) составляют делеции, тогда как инверсии отсутствуют.

Эталонный BED-файл Tier1_v0.6 включает в себя 34830 регионов на всех хромосомах, за исключением митохондриальной.

Для прочтений образца HG002 нами были запущены три пайплайна выравнивания: DRAGEN, GDC’s DNA-Seq Alignment Workflow и GDC’s DNA-Seq Alignment + Co-cleaning Workflow. В качестве входных данных для пайплайнов выступали прочтения разной длины и глубины секвенирования, скачанные с сайта NCBI по протоколу FTP. Файлы с выравниваниями сортировались и индексировались с помощью инструмента samtools.

На сортированных BAM-файлах были запущены три биоинформатических инструмента: Manta, Delly и Smoove. В случае с Manta для VCF-файлов были применены аннотация программой Duphold и фильтрация на основе аннотированных метрик.

В табл. 2 представлена характеристика трех программ, детектирующих различные типы структурных вариантов, на основе данных секвенирования короткими прочтениями (2 × 150).

Таблица 2. Программы, используемые для поиска структурных вариантов

Название, версия и дата последнего обновления программы	Используемые программой методы выявления структурных вариантов	Определяемые типы структурных вариантов				
		DEL	INS	DUP	INV	TRA
Manta v1.6.0, 07.2019	read pairs + split reads + assembly	+	+	+	+	+
Delly v0.9.1, 12.2021	read pairs + split reads	+	+	+	+	+
Smoove v0.2.6, 10.2021	read pairs + split reads + read depth	+	–	+	+	+

Примечание. TRA – транслокации.

На рис. 1 представлены 16 диаграмм, на которых изображены результаты работы биоинформатических инструментов Manta, Smoove, Delly, а также результаты совместной работы инструментов Manta и Smoove, Manta и Delly, Delly и Smoove, Manta, Smoove и Delly. Каждому из инструментов, а также их парам и тройке присвоен свой цвет. Результаты работы каждого из инструментов по отдельности, а также их совместной работы обозначены прямоугольниками разных цветов. Высота каждого из цветных прямоугольников соответствует количеству найденных структурных вариантов, которые показаны в следующей разбивке: до фильтрации (Total) и после фильтрации (PASS), в том числе типов DEL (делеции), INS (инсерции), DUP (дупликации) и INV (инверсии). В надписях над каждой диаграммой указан использованный способ выравнивания, длина прочтений (в пн) и среднее покрытие. Рис. 1 показывает, что Delly нашел больше вариантов во всех случаях, но более половины не прошли фильтрации (PASS) при выборе пайплайнов DNA-Seq A и DNA-Seq A + C. При этом количество вариантов, найденных Delly, сильно росло с ростом покрытия (в 1.5 раза), а также при выборе пайплайнов выравнивания DNA-Seq A и DNA-Seq A + C (более чем в 2 раза). Smoove нашел меньше вариантов в сравнении с Delly и Manta во всех случаях. Количество вариантов, найденных Smoove и Manta, также увеличивалось с увеличением покрытия, но не так сильно, как для Delly (в 1.1 раз). При выборе пайплайнов выравнивания DNA-Seq A и DNA-Seq A + C количество вариантов незначительно увеличивалось для Smoove (в 1.3 раза), а для Manta почти не изменялось. Использование пайплайна выравнивания NovoAlign практически сводит к нулю число структурных вариантов типа DEL, находимых инструментом Smoove.

На рис. 2 и 3 представлены сопоставимые количественные оценки качества работы разных биоинформатических инструментов по выявлению структурных вариантов типа DEL и INS соответственно. Показано влияние выравнивания разными пайплайнами с входными данными с разным предполагаемым покрытием. Оценка качества работы инструментов проводилась посредством расчета и анализа трех показателей: точности, чувствительности и Fscore (суммарный показатель точности и чувствительности).

Как видно из рис. 2 и табл. 3, показатель чувствительности возрастает на 3–7% для инструментов Manta, Delly и Smoove при увеличении покрытия с 25x до 50x, а при увеличении покрытия с 50x до 75x почти не меняется. Показатель точности для инструментов Manta, Smoove и Delly падает при изменении покрытия с 25x до 75x. Таким образом, увеличение покрытия с 25x до 75x приводит к некоторому повышению качества работы двух из трех биоинформатических инструментов, предназначенных для поиска структурных вариантов типа DEL, по показателю Fscore. Так, показатель Fscore для инструмента Smoove увеличивается с 65–66 до 67–68%, а для инструмента Manta – с 73–74 до 78–79%. Для инструмента Delly такой зависимости не выявлено.

Увеличение длины прочтений со 150 до 250 пн по-разному влияет на качество работы разных инструментов при разных пайплайнах выравнивания. Для инструмента Delly показатель Fscore для вариантов типа DEL увеличивается с 64.5 (в среднем) до 67 при пайплайне DRAGEN (точность и чувствительность увеличиваются на 2–3% и до 72 при пайплайне DNA-Seq Alignment (точность увеличивается на 3%, чувствительность на 11%). А при пайплайне DNA-Seq Alignment + + Co-cleaning этот показатель повышается с 65.8 (в среднем) до 74% (точность повышается на 4%,

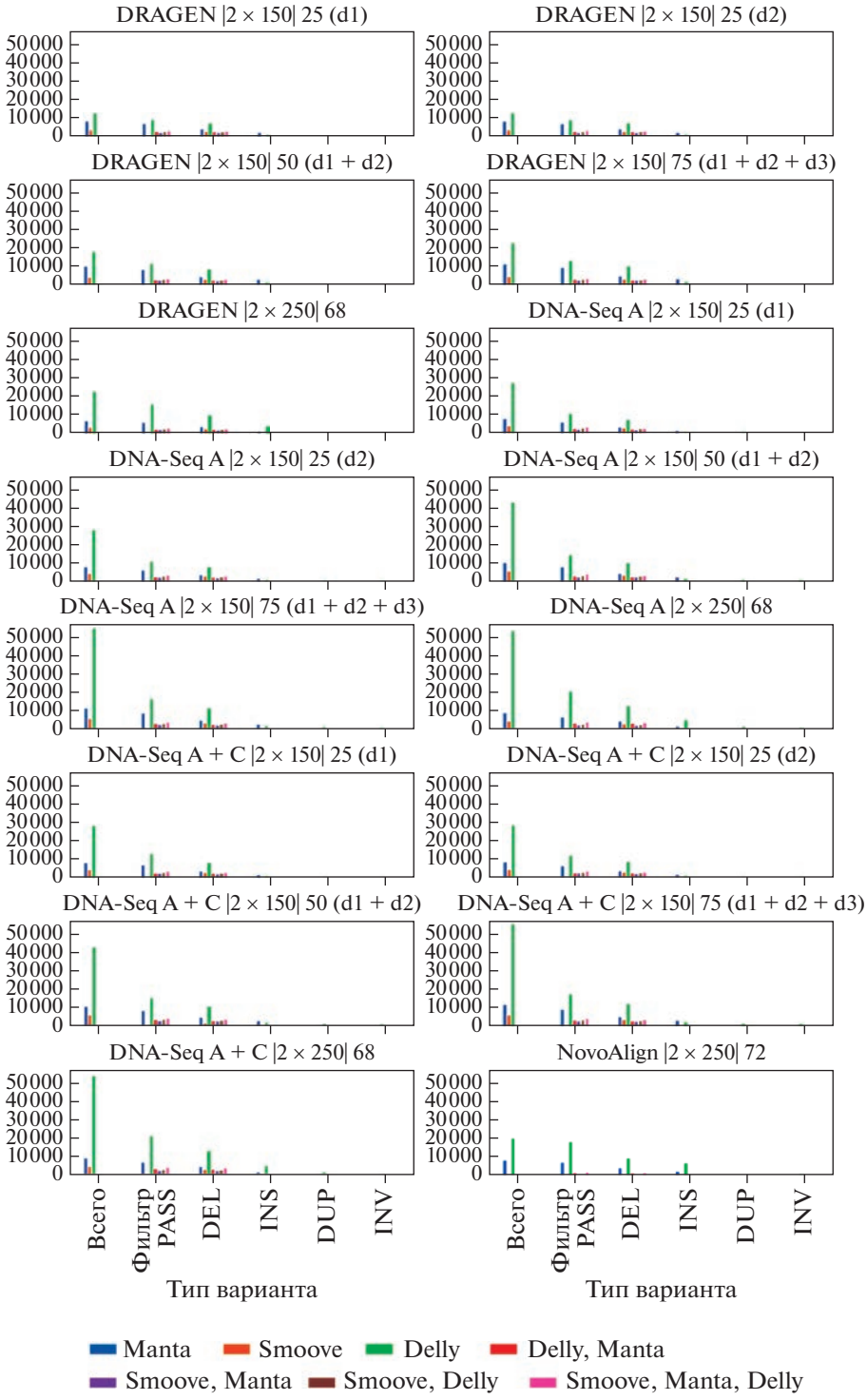


Рис. 1. Структурные варианты, найденные с помощью разных биоинформатических инструментов на результатах выравнивания разными пайплайнами.

чувствительность на 10%). Для инструмента Smoove показатель Fscore при увеличении длины прочтений на 100 пн практически не меняется при всех трех пайплайнах (соответственно точность и чувствительность также практически не менются). В то же время для инструмента Manta этот показатель при пайплайнах DRAGEN и DNA-Seq Alignment + Co-cleaning изменяется не-

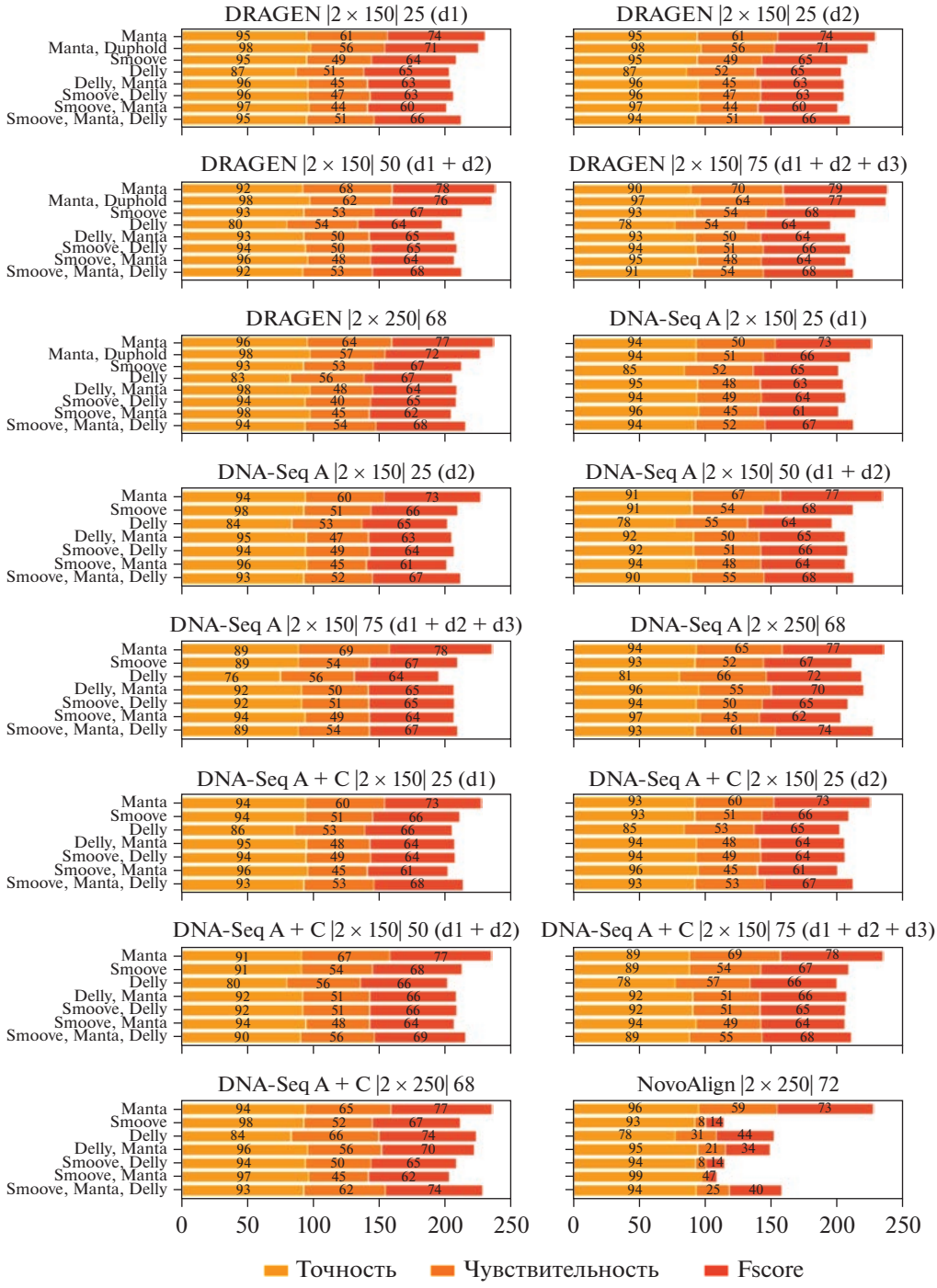


Рис. 2. Оценка качества работы разных биоинформатических инструментов по выявлению структурных вариантов типа DEL.

значительно, а при пайплайне DNA-Seq Alignment он увеличивается с 72.2 (в среднем) до 77% (точность увеличивается на 3–4%, а чувствительность уменьшается на 2–4% при всех трех пайплайнах). Фактически ни длина прочтений, ни глубина секвенирования, ни используемый для выравнивания пайплайн не дают существенного вклада в изменение показателя Fscore, дающего сводную оценку качества детектирования структурных вариантов типа DEL.

Что касается самих биоинформатических инструментов для выявления структурных вариантов, то, судя по всем трем показателям: точность,

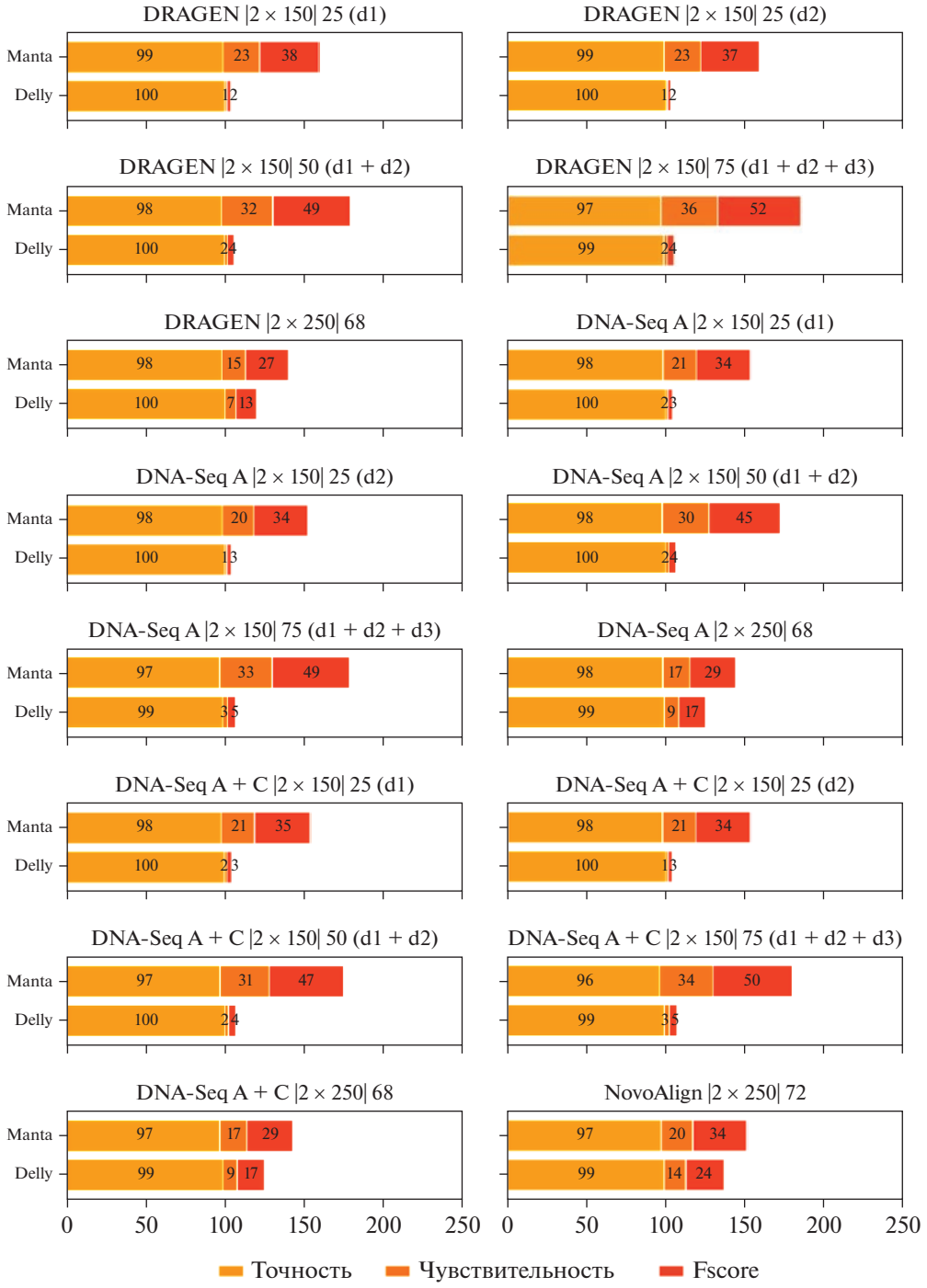


Рис. 3. Оценка качества работы разных биоинформатических инструментов по выявлению структурных вариантов типа INS.

чувствительность и Fscore, наиболее предпочтительным инструментом является Manta, который дает самые высокие показатели Fscore по сравнению с другими инструментами и их комбинациями при любом пайплайне выравнивания, включая NovoAlign, при любой глубине секвенирования и при любой из двух длин прочтений.

Сравнение результатов поиска делеций (см. рис. 2) и инсерций (см. рис. 3) демонстрирует, что поиск структурных вариантов типа делеция характеризуется более низкими показателями точности и более высокими показателями чувствительности по сравнению с поиском инсерций. При этом биоинформатический инструмент Manta детектирует

Таблица 3. Сравнение качества работы биоинформатических инструментов с помощью показателя Fscore на примере поиска структурных вариантов типа DEL

№	Пайплайн выравнивания	Длина прочтения (пн), глубина секвенирования	Показатель Fscore для биоинформатического инструмента							
			Manta	Manta с аннот. Duphold	Smoove	Delly	Manta + Smoove	Manta + Delly	Delly + Smoove	Manta + + Smoove + Delly
1	DRAGEN	2 × 150, 25x (d1)	74	71	64	65	60	63	63	66
		2 × 150, 25x (d2)	74	71	65	65	60	63	63	66
		2 × 150, 50x (d1 + d2)	78	76	67	64	64	65	65	68
		2 × 150, 75x (d1 + d2 + d3)	79	77	68	64	64	64	66	68
	Средняя для 2 × 150		*76.2*	*73.8*	*66.0*	*64.5*	*62.0*	*63.8*	*64.25*	*67.0*
		2 × 250, 68x	77	72	67	67	62	64	65	68
2	DNA-Seq Alignment	2 × 150, 25x (d1)	73	–	66	65	61	63	64	67
		2 × 150, 25x (d2)	73	–	66	65	61	63	64	67
		2 × 150, 50x (d1 + d2)	77	–	68	64	64	65	66	68
		2 × 150, 75x (d1 + d2 + d3)	78	–	67	64	64	65	65	67
	Средняя для 2 × 150		*72.2*	–	*66.8*	*64.5*	*62.5*	*64.0*	*64.75*	*67.2*
		2 × 250, 68x	77	–	67	72	62	70	65	74
3	DNA-Seq Alignment + Co-Cleaning	2 × 150, 25x (d1)	73	–	66	66	61	64	64	68
		2 × 150, 25x (d2)	73	–	66	65	61	64	64	67
		2 × 150, 50x (d1 + d2)	77	–	68	66	64	66	66	69
		2 × 150, 75x (d1 + d2 + d3)	78	–	67	66	64	66	65	68
	Средняя для 2 × 150		*75.2*	–	*66.8*	*65.8*	*62.5*	*65.0*	*64.75*	*68.0*
		2 × 250, 68x	77	–	67	74	62	70	65	74
4	NovoAlign	2 × 250, 72x	73	–	14	44	7	34	14	40

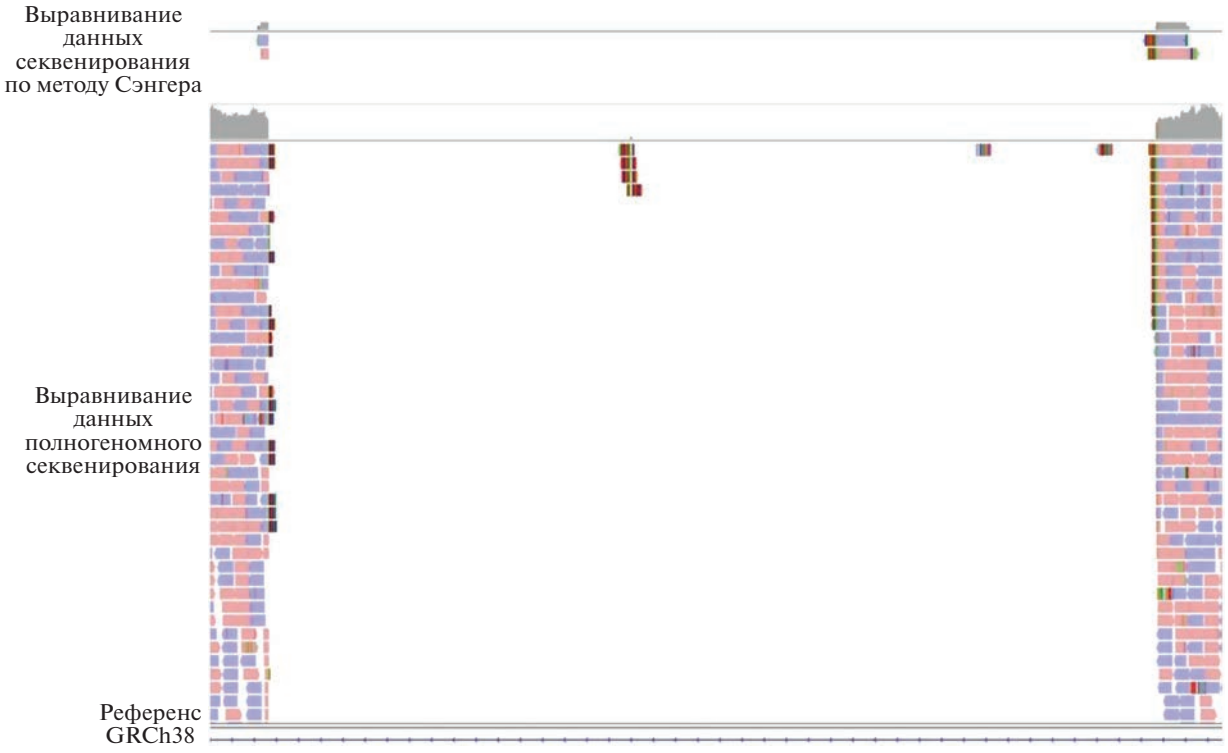


Рис. 4. Подтверждение наличия делеции chr5:12810907–12820414 (9507 пн). Здесь и далее розовая и синяя окраски прочтений обозначают выравнивание на обратную и прямую цепи ДНК соответственно.

структурные варианты лучше, чем инструменты Delly и Smoove, а также их комбинация. Как оказалось, файлы, полученные от совместной работы нескольких программ, фактически не увеличивают точности поиска и еще значительно снижают чувствительность. Фильтрация duphold, сочетаемая с инструментом Manta, несколько увеличивает точность нахождения делеций в выравниваниях, сделанных пайплайном DRAGEN, но уменьшает чувствительность и Fscore.

В ходе данного эксперимента точность поиска дупликаций и инверсий оценить не представилось возможным, поскольку формат записи таких типов структурных вариантов в эталонном файле отличается от формата записи в результирующих файлах анализируемых инструментов, однако проверка этих типов вариантов проводилась секвенированием по методу Сэнгера.

В Приложениях 2–4 показано, что точность работы биоинформатических инструментов для поиска структурных вариантов варьирует в зависимости от длины этих вариантов. Например, Manta лучше находит делеции длиной 200 и более пн, Smoove – от 200 до 10000 пн, Delly – длиной от 1000 до 10000 пн.

Валидация структурных вариантов

После полногеномного секвенирования был проведен поиск структурных вариантов в двух отобранных образцах. Найденные варианты проверялись с помощью секвенирования по методу Сэнгера. Для проверки были случайно отобраны делеции (размером от 50 до 45000 пн), инсерции, дупликации и инверсии (размером от 50 до 500 пн), которые были обнаружены инструментами Manta, Smoove и Delly, а также их комбинациями: Manta + Smoove, Manta + Delly, Smoove + Delly, Manta + Smoove + Delly.

Прочтения, полученные в результате секвенирования по методу Сэнгера, выравниваются по краям от делеции, без захвата участка собственно делеции.

После выравнивания прочтений (FASTQ) BAM-файл загружался в браузер IGV, с помощью которого визуально подтверждалось наличие или отсутствие структурного варианта в данной позиции. На рис. 4–6 показаны примеры подтвержденных делеции, инсерции и дупликации соответственно.

В табл. 4 представлена информация по количеству отобранных секвенированных и подтвержденных структурных вариантов.

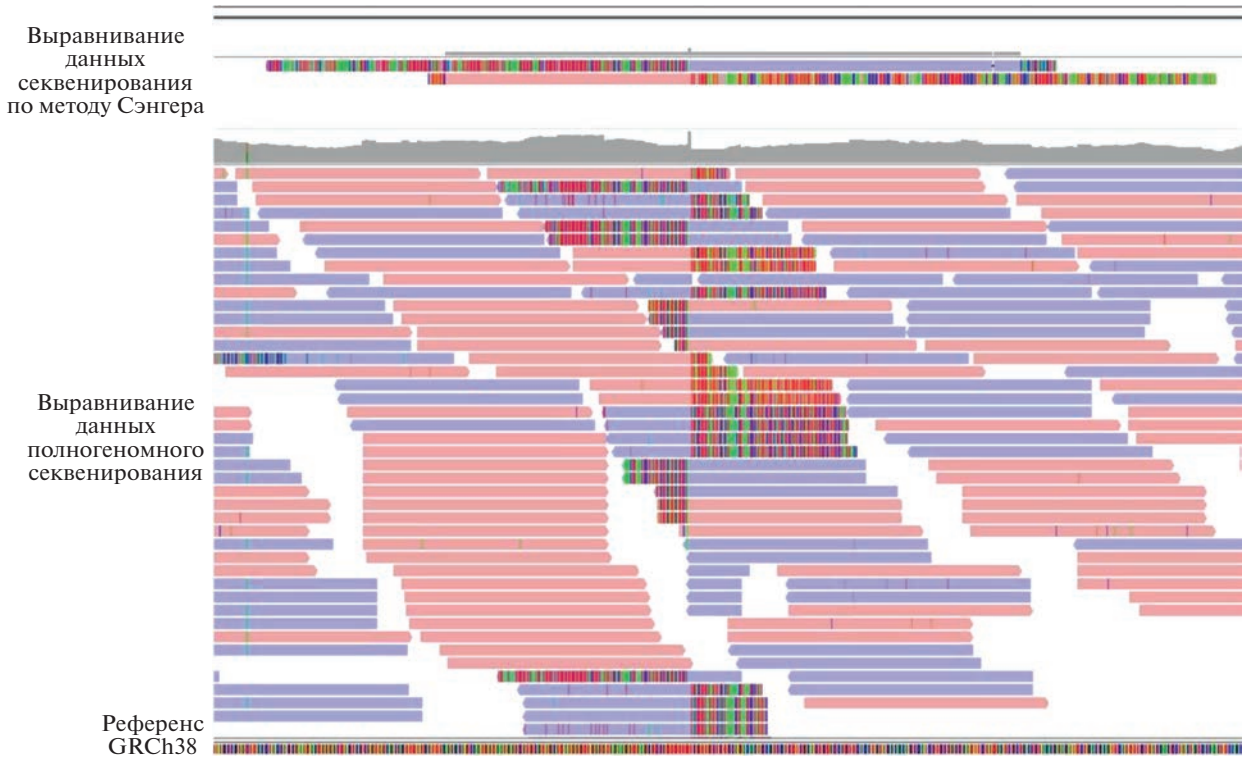


Рис. 5. Подтверждение наличия инсерции chr20:6665192 (67 пн). Здесь и далее яркие разноцветные участки по краям инсерции являются soft-clipped нуклеотидами, они не выравниваются на референсный геном, но остаются в выравнивании.

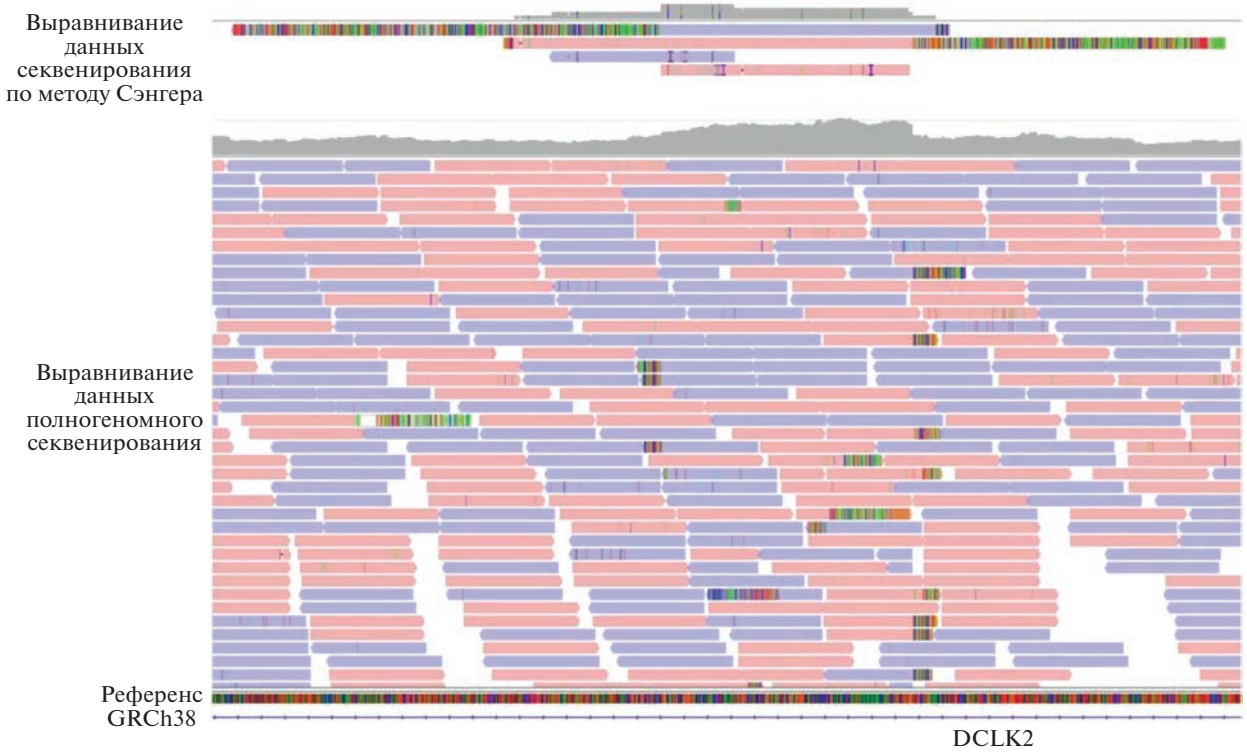


Рис. 6. Подтверждение наличия дупликации chr4:150090160–150090487 (327 пн).

Таблица 4. Сопоставление подтвержденных и секвенированных структурных вариантов

Ход эксперимента	Структурные варианты, ед.				Всего, ед.
	DEL	INS	DUP	INV	
Отобрано вариантов	60	28	23	9	120
из них секвенировано по Сэнгеру	17	12	12	2	43
из них подтвердились	11	7	6	0	24
доля подтвердившихся от секвенированных	*65%*	*58%*	*50%*	–	*56%*
не удалось секвенировать по методу Сэнгера	43	16	11	7	77
доля несеквенированных от отобранных	*72%*	*57%*	*48%*	*78%*	*64%*

В ходе данного эксперимента не все структурные варианты были секвенированы, поскольку для некоторых из них не произошло специфического связывания праймеров с матрицей, либо было невозможно подобрать праймеры из-за сложности региона (большое количество повторов во фланкирующем регионе), либо рассчитанная температура отжига праймеров не обеспечивала высокую специфичность ПЦР. Доля вариантов, которые не удалось секвенировать по методу Сэнгера, составила 78% для инверсий, 72% для делеций, 57% для инсерций и 48% для дупликаций.

Как следует из табл. 4, для 17 биоинформатически выявленных делеций, которые были проверены с помощью секвенирования по методу Сэнгера, подтвердились 11, т.е. 65%. Для инсерций аналогичная доля подтвердившихся вариантов среди всех секвенированных составила 58%. Для дупликаций эта доля составила 50%. В то же время из двух инверсий, выявленных биоинформатически и проверенных секвенированием по методу Сэнгера, не подтвердилось ни одной.

Таким образом, молекулярно-генетическая проверка показала, что делеции подтверждаются лучше, чем инсерции, дупликации и тем более инверсии. Варианты, которые были найдены при совместном использовании двух и трех биоинформатических инструментов, как оказалось, тоже подтверждаются далеко не всегда.

В табл. 5 представлены результирующие данные по вариантам, прошедшим молекулярно-генетическую проверку.

Как видно из табл. 5, из отобранных в рамках данной работы двух инверсий не подтвердилась ни одна. Следовательно, инверсии, найденные с помощью биоинформатических инструментов Manta и Smoove, требуют дополнительной проверки.

Табл. 5 и Приложение 4 также свидетельствуют о том, что структурные варианты типа DEL, INS и DUP целесообразно выявлять с помощью инструмента Manta v1.6.0 – использование двух или трех инструментов одновременно не увеличило показателя Fscore.

ОБСУЖДЕНИЕ

С развитием технологий секвенирования, со снижением стоимости и ресурсоемкости последующего анализа растет количество данных, по которым можно изучать ассоциации генетических вариантов, в том числе и структурных, с развитием целого ряда заболеваний. В настоящее время создано более 30 инструментов для поиска структурных вариантов и со временем это число будет только расти.

К сожалению, среди такого большого числа инструментов нет универсального, который бы имел высокие показатели точности и чувствительности для всех типов структурных вариантов, подходил бы для массовых вычислений, результирующие файлы которого были бы совместимы с другими инструментами для поиска структурных вариантов и инструментами объединения и создания файла с вариантами, найденными комбинацией двух и более инструментов. Поэтому в зависимости от цели исследования используемые инструменты и их режимы запуска могут различаться между собой.

Для настоящего исследования мы выбрали три биоинформатических инструмента (Manta, Smoove, Delly), которые подходят для массовой обработки данных. Вторым важным фактором, повлиявшим на выбор именно этих трех инструментов, стало то, что их результирующие файлы имеют аналогичный формат записи информации о структурных вариантах, включая аннотации. Соответ-

Таблица 5. Результаты проверки структурных вариантов с помощью секвенирования по методу Сэнгера

№	Тип SV	Длина SV, пн	Инструмент поиска структурных вариантов		
			Manta v1.6.0	Smoove v0.2.8	Delly v0.9.1
Делеции					
1	DEL	3356	chr1:115686865	chr1:115686865	chr1:115686865
2	DEL	1482	н.о.	chr1:17349677	chr1:17349677
3	DEL	319	chr1:240453690	chr1:240453707	chr1:240453690
4	DEL	68	chr2:1208249	chr2:1208263	chr2:1208249
5	DEL	414	chr2:1354296	chr2:1354296	н.о.
6	DEL	43366	chr3:130044537–130087902	chr3:130044539–130087902	chr3:130044537–130087903
7	DEL	3000	chr4:185520481	chr4:185520483	chr4:185520481
8	DEL	9507	chr5:12810905–12820411	chr5:12810907–12820411	chr5:12810905–12820412
9	DEL	474	chr11:677345	chr11:677345	chr11:677345
10	DEL	62	chr11:21767543	chr11:21767543	chr11:21767543
11	DEL	11030	chr12:12378536–12389566	chr12:12378535–12389566	chr12:12378537–12389567
12	DEL	316	chr12:1754845	chr12:1754845	chr12:1754845
13	DEL	877	chr17:17327060–17327936	chr17:17327067–17327919	chr17:17327060–17327937
14	DEL	316	chr18:79739850	chr18:79739850	н.о.
15	DEL	3133	н.о.	chr19:605301	н.о.
16	DEL	73	chr20:64212162	н.о.	chr20:64212162
17	DEL	2482	chrX:77865888–77868369	chrX:77865918–77868355	chrX:77865920–77868356
Результат*			11-2-2-0–0.85-1.00-0.92	10-0-4-1–0.71-0.91-0.80	9-1-3-2–0.75-0.82-0.78
Manta + Smoove			10-2-2-1–0.83-0.91-0.90		
Smoove + Delly				8-1-3-3–0.72-0.73-0.73	
Manta + Delly			9-2-2-2–0.82-0.82-0.82		
Manta + Delly + Smoove			8-2-2-3–0.80-0.73-0.76		
Инсерции					
18	INS	317	chr1:4228390	н.о.	н.о.
19	INS	67	chr4:136314242	н.о.	chr4:136314242
20	INS	57	chr5:134232927	н.о.	chr5:134232927
21	INS	135	chr8:1912540	н.о.	н.о.
22	INS	54	chr8:102867903	н.о.	chr8:102867903
23	INS	316	chr10:9432883	н.о.	н.о.
24	INS	56	chr13:92723353	н.о.	chr13:92723353
25	INS	80	chr15:53110125	н.о.	н.о.

Таблица 5. Окончание

№	Тип SV	Длина SV, пн	Инструмент поиска структурных вариантов		
			Manta v1.6.0	Smoove v0.2.8	Delly v0.9.1
26	INS	52	chr16:839939	н.о.	chr16:839939
27	INS	69	chr19:49867632	н.о.	н.о.
28	INS	66	chr20:6665192	н.о.	chr20:6665192
29	INS	60	chrX:145338072	н.о.	chrX:145338072
Результат*			7-0-3-0–0.70-1.00-0.82	–	4-2-1-3–0.80-0.57-0.67
Manta + Delly			4-2-1-3–0.80-0.57-0.67		
Дупликации					
30	DUP	234	н.о.	chr2:239723885	н.о.
31	DUP	372	chr3:990665	chr3:990665	chr3:990665
32	DUP	329	chr4:150090160	chr4:150090160	chr4:150090160
33	DUP	189	chr6:1051543	chr6:1051543	н.о.
34	DUP	189	chr7:40915751	chr7:40915751	н.о.
35	DUP	247	chr9:2814058	chr9:2814058	chr9:2814058
36	DUP	205	chr14:23900201	chr14:23900201	chr14:23900201
37	DUP	196	chr14:82276270	н.о.	chr14:82276270
38	DUP	162	chr15:101390179	chr15:101390179	н.о.
39	DUP	54	chr16:919548	н.о.	н.о.
40	DUP	196	н.о.	chr18:9809064	chr18:9809064
41	DUP	144	chr20:10323922	н.о.	н.о.
Результат*			5-0-4-1–0.56-0.83-0.67	4-1-3-2–0.57-0.67-0.62	3-3-1-3–0.75-0.50-0.60
Manta + Smoove			3-1-3-3–0.50-0.50-0.50		
Smoove + Delly				2-3-1-4–0.67-0.33-0.55	
Manta + Delly			3-3-1-3–0.75-0.50-0.60		
Manta + Delly + Smoove			2-3-1-4–0.67-0.33-0.44		
Инверсии					
42	INV	75	chr10:12825172	chr10:12825172	н.о.
43	INV	11	chr11:104240549	chr11:104240549	н.о.
Результат*			–	–	–

Примечание. Красный цвет – вариант не подтвердился, зеленый – вариант подтвердился, оранжевый – последовательности выравниваются таким образом, что нельзя достоверно сказать подтверждается вариант или нет (например, одна из цепей не выравнивается). н.о. – не обнаружено.

* Числа даны в следующем порядке: TP (истинно положительные)-TN (истинно отрицательные)-FP (ложно положительные)-FN (ложноотрицательные)–точность-чувствительность-Fscore; варианты с сомнительными результатами проверки не учитывались.

ственно файлы от этих трех инструментов удобно сопоставлять и анализировать совместно в разных комбинациях.

Нами проведена оценка качества работы инструментов Manta, Smoove и Delly посредством проверки найденных ими вариантов с помощью секвенирования по методу Сэнгера. Рассматривались структурные варианты, найденные отдельно каждым инструментом, а также совместно двумя и тремя инструментами. Мы оценили влияние длины входных прочтений и покрытия в целом, а также влияние разных алгоритмов этапа выравнивания и показали, что с увеличением покрытия образца растет достоверность идентификации структурных вариантов биоинформатическими инструментами. В то же время длина коротких прочтений, при равном покрытии, почти не влияет на качество работы инструментов Manta, Smoove и Delly. Ограничением данного исследования стал тот факт, что влияние длины фрагментов и insert size на точность, чувствительность, Fscore нами не изучалось.

Также нам удалось показать, что выбор алгоритма выравнивания оказывает влияние на качество работы инструментов. Например, для инструмента Smoove выравнивание с помощью NovoAlign уменьшает число найденных вариантов почти в 7 раз по сравнению с другими алгоритмами выравнивания.

В результате настоящего исследования мы сделали вывод о том, что каждый из трех изученных инструментов эффективнее работает в определенном диапазоне размеров структурных вариантов. Исходя из этого, мы предполагаем, что изучение структурных вариантов, имеющих разные длины, целесообразно проводить используя совместно два и более инструмента.

Проверка данных, полученных в ходе нашего исследования с помощью секвенирования по методу Сэнгера, показала, что в действительности точность работы всех трех биоинформатических инструментов оказалась намного ниже, чем предполагалось. Фактически молекулярно-генетический метод не подтвердил наличие многих структурных вариантов, детектированных инструментами Manta, Smoove и Delly. При этом молекулярно-генетическая проверка показала, что в среднем, без учета спорных вариантов, делеции и инсерции биоинформатически определяются надежнее, чем инверсии и дупликации.

Таким образом, наше исследование позволило дать следующие рекомендации:

– для анализа данных WGS, популяционных исследований, поиска ассоциаций с фенотипом, расчета частоты встречаемости вариантов в выборке целесообразно совместно использовать хотя бы два биоинформатических инструмента для поиска структурных вариантов;

– при клинической интерпретации генетических данных наличие структурного варианта необходимо подтверждать с помощью альтернативных молекулярно-генетических методов.

Исследование не имело спонсорской поддержки.

Все процедуры, выполненные в исследовании с участием людей, соответствуют этическим стандартам институционального и/или национального комитета по исследовательской этике и Хельсинкской декларации 1964 г. и ее последующим изменениям или сопоставимым нормам этики.

От каждого из включенных в исследование участников было получено информированное добровольное согласие.

Авторы заявляют, что у них нет конфликта интересов.

ПРИЛОЖЕНИЕ 1

Некоторые команды запуска инструментов, описанных в Материалах и методах:

1) команда запуска инструмента bcftools для фильтрации структурных вариантов: `bcftools view -i 'FILTER="PASS" && INFO/SVTYPE != "BND"`;

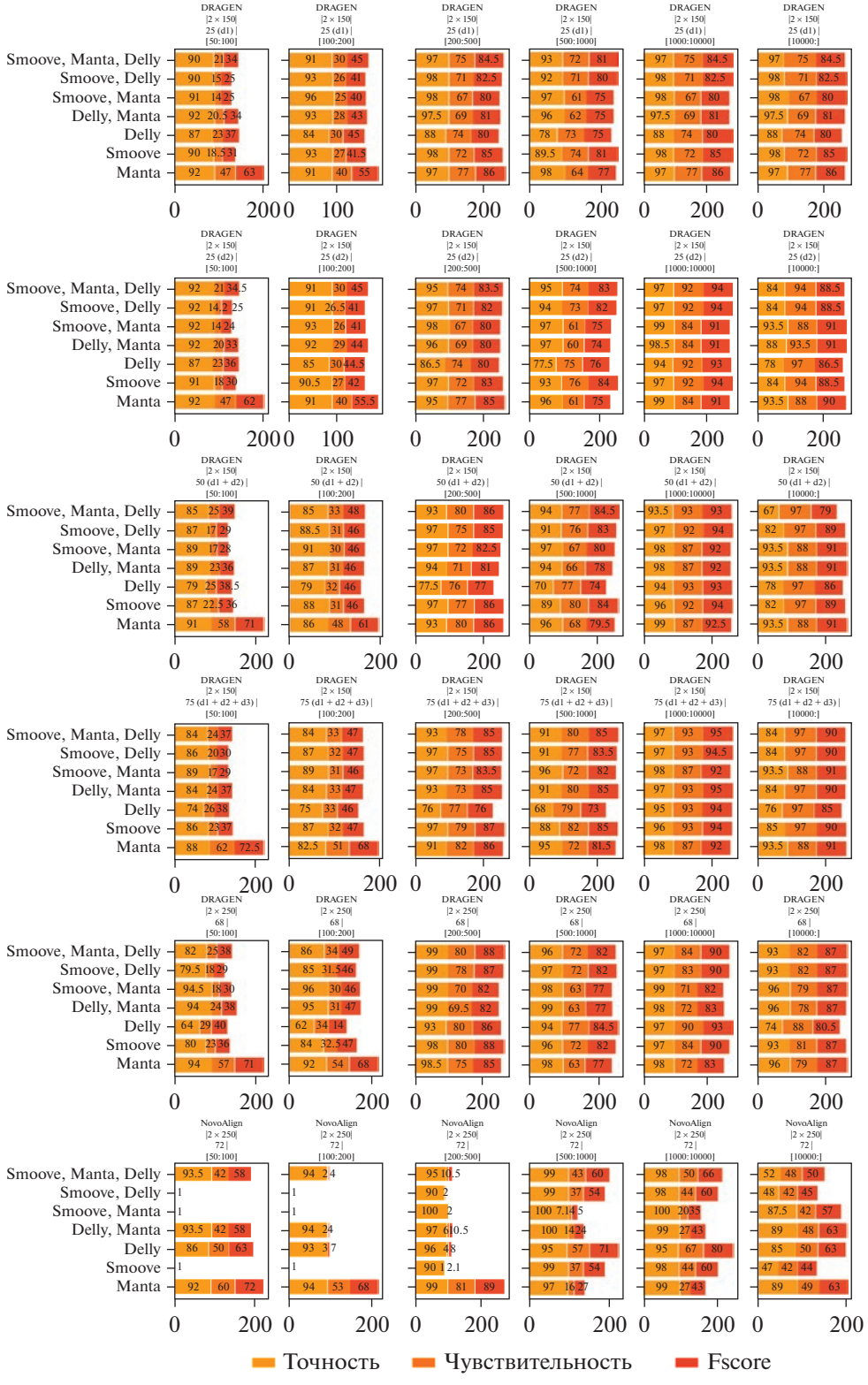
2) команда запуска инструмента bcftools для фильтрации структурных вариантов, аннотированных duphold: `bcftools view -i '(SVTYPE == "DEL" & FMT/DHFFC[0] < 0.7) | (SVTYPE == "DUP" & FMT/DHBFC[0] > 1.3)' $svvcf`;

3) команда запуска инструмента SURVIVOR:

`SURVIVOR merge <list_of_VCF_files_for_merging> 1000 2 1 1 0 50 <merged.vcf>`);

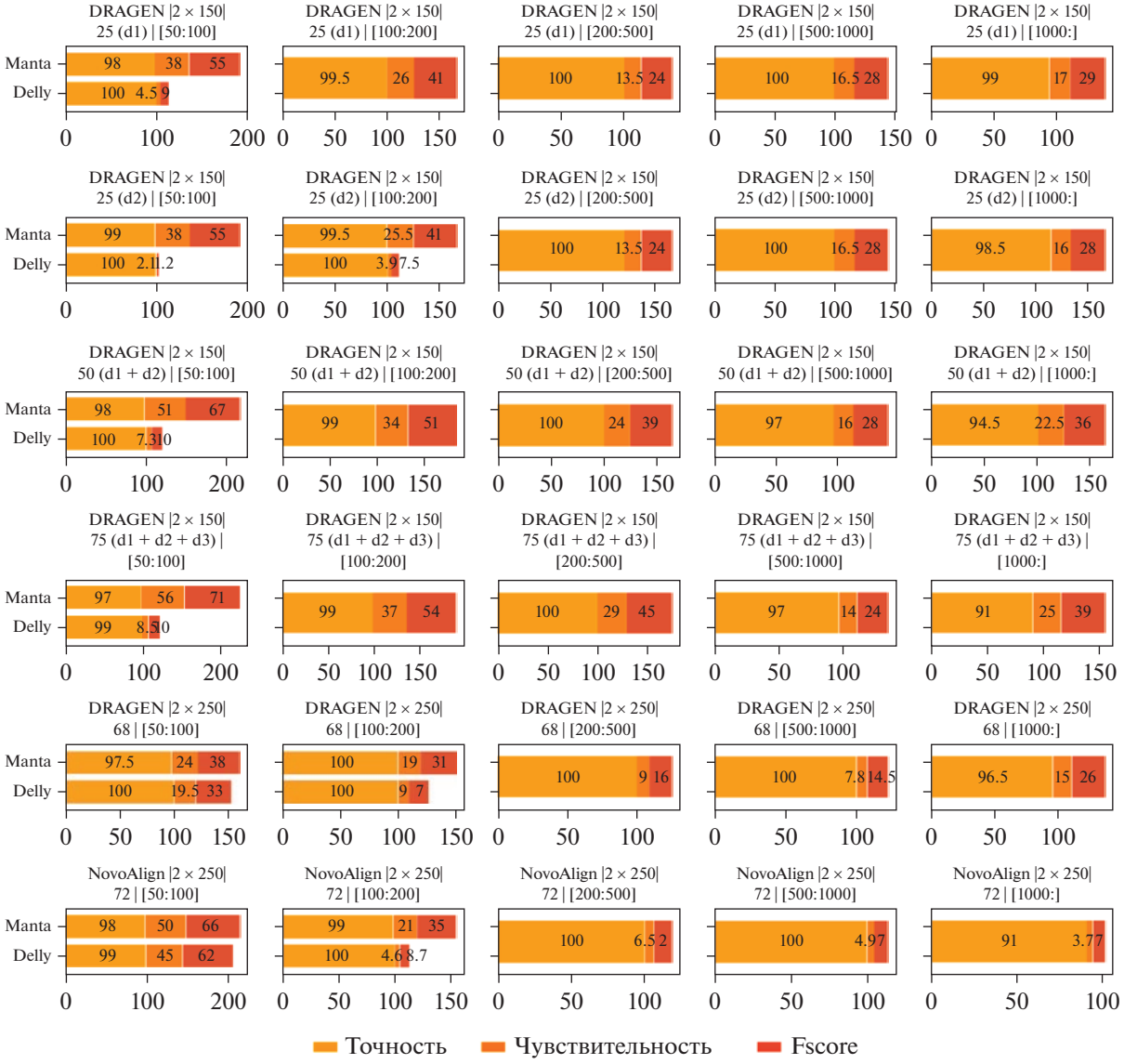
4) команда запуска инструмента Witty.er: `wittyer -i <vcf_to_validate.vcf> -t HG002_SVs_Tier1_v0.6.vcf –em SimpleCounting –configFile config.json –includeBed HG002_SVs_Tier1_v0.6.bed -o <dir_with_comarisson_results_from_wittyer>`.

ПРИЛОЖЕНИЕ 2



Результаты оценки чувствительности и точности работы инструментов для поиска структурных вариантов для файлов выравнивания, полученных с помощью DRAGEN на прочтениях разных длин и покрытия для разных диапазонов длин вариантов, для структурных вариантов типа делеция.

ПРИЛОЖЕНИЕ 3



Результаты оценки чувствительности и точности работы инструментов для поиска структурных вариантов для файлов выравнивания, полученных с помощью DRAGEN на прочтениях разных длин и покрытия для разных диапазонов длин вариантов, для структурных вариантов типа инсерция.

ПРИЛОЖЕНИЕ 4

Сравнение качества работы биоинформатических инструментов с помощью показателя Fscore на примере поиска структурных вариантов типа DEL на разных диапазонах длин

№	Диапазон длин, пн	Пайплайн выравнивания	Длина прочтения (пн), глубина секвенирования	Показатель Fscore для биоинформатического инструмента						
				Manta	Smoove	Delly	Manta + + Smoove	Manta + + Delly	Delly + + Smoove	Manta + + Smoove + + Delly
1	50 : 100	DRAGEN	2 × 150, 25x (d1)	63	31	37	25	34	25	34
			2 × 150, 25x (d2)	62	30	36	24	33	25	34.5
			2 × 150, 50x (d1 + d2)	71	36	38.5	28	36	29	39
			2 × 150, 75x (d1 + d2 + d3)	72.5	37	38	29	37	30	37
		Средняя для 2 × 150		*67.1*	*33.5*	*37.4*	*26.5*	*35.0*	*27.3*	*36.1*
			2 × 250, 68x	71	36	40	30	38	29	38
		NovoAlign	2 × 250, 72x	72	0	63	1	58	0	58
2	100 : 200	DRAGEN	2 × 150, 25x (d1)	55	41.5	45	40	43	41	45
			2 × 150, 25x (d2)	55.5	42	44.5	41	44	41	45
			2 × 150, 50x (d1 + d2)	61	46	46	45	46	46	48
			2 × 150, 75x (d1 + d2 + d3)	63	47	46	46	47	47	47
		Средняя для 2 × 150		*58.6*	*44.1*	*45.4*	*43.0*	*45.0*	*43.8*	*46.3*
			2 × 250, 68x	68	47	44	46	47	46	49
		NovoAlign	2 × 250, 72x	68	0	3.7	1	4	0	4
3	200 : 500	DRAGEN	2 × 150, 25x (d1)	86	83	80	80	81	82.5	84.5
			2 × 150, 25x (d2)	85	83	80	80	80	82	83.5
			2 × 150, 50x (d1 + d2)	86	86	77	82.5	81	85	86
			2 × 150, 75x (d1 + d2 + d3)	86	87	76	83.5	85	85	85
		Средняя для 2 × 150		*85.8*	*84.8*	*78.3*	*81.5*	*81.8*	*83.6*	*84.8*
			2 × 250, 68x	85	88	86	82	82	87	88
		NovoAlign	2 × 250, 72x	89	2.1	9	2	10.5	2	10.5
	500 : 1000	DRAGEN	2 × 150, 25x (d1)	77	81	75	75	75	80	81
			2 × 150, 25x (d2)	75	84	76	75	74	82	83
			2 × 150, 50x (d1 + d2)	79.5	84	74	80	78	83	84.5
			2 × 150, 75x (d1 + d2 + d3)	81.5	85	73	82	85	83.5	85
		Средняя для 2 × 150		*78.3*	*83.5*	*74.5*	*78.0*	*78.0*	*82.1*	*83.4*
			2 × 250, 68x	77	82	84.5	77	77	82	82
		NovoAlign	2 × 250, 72x	27	54	71	14	24	54	60
	1000 : 10000	DRAGEN	2 × 150, 25x (d1)	91	94	94	91	91	94	94
			2 × 150, 25x (d2)	91	94	93	91	91	94	94
			2 × 150, 50x (d1 + d2)	92.5	94	93	92	92	94	93
			2 × 150, 75x (d1 + d2 + d3)	92	94	94	92	95	94.5	95
		Средняя для 2 × 150		*91.6*	*94.0*	*93.5*	*91.5*	*92.3*	*94.1*	*94.0*
			2 × 250, 68x	83	90	93	82	83	90	90
		NovoAlign	2 × 250, 72x	43	60	80	35	43	60	66
	10000:	DRAGEN	2 × 150, 25x (d1)	89	90	85	90	89	90	89
			2 × 150, 25x (d2)	90	88.5	86.5	91	91	88.5	88.5
			2 × 150, 50x (d1 + d2)	91	89	86	91	91	89	79
			2 × 150, 75x (d1 + d2 + d3)	91	90	85	91	90	90	90
		Средняя для 2 × 150		*90.3*	*89.4*	*85.6*	*90.8*	*90.3*	*89.4*	*86.6*
			2 × 250, 68x	87	87	80.5	87	87	87	87
4		NovoAlign	2 × 250, 72x	63	44	63	57	63	45	50

СПИСОК ЛИТЕРАТУРЫ

1. *Pang A.W., MacDonald J.R., Pinto D. et al.* Towards a comprehensive structural variation map of an individual human genome // Genome Biol. 2010. V. 11. № 5. P. R52. https://doi.org/10.1186/gb-2010-11-5-r52
2. *The International HapMap Consortium*. The international hapmap project // Nature. 2003. P. 789–796. https://doi.org/10.1038/nature02168
3. *Sudmant P.H., Rausch T., Gardner E.J. et al.* An integrated map of structural variation in 2,504 human genomes: 7571 // Nature. 2015. V. 526. № 7571. P. 75–81. https://doi.org/10.1038/nature15394
4. *Pös O., Radvanszky J., Buglyo G. et al.* DNA copy number variation: Main characteristics, evolutionary significance, and pathological aspects // Biomed. J. 2021. V. 44. № 5. P. 548–559. https://doi.org/10.1016/j.bj.2021.02.003
5. *Alkan C., Coe B.P., Eichler E.E.* Genome structural variation discovery and genotyping // Nat. Rev. Genet. 2011. V. 12. № 5. P. 363–367. https://doi.org/10.1038%2Fnrg2958
6. *Mahmoud M., Gobet N., Cruz-Davalos D.I. et al.* Structural variant calling: the long and the short of it // Genome Biol. 2019. V. 20. № 1. P. 246. https://doi.org/10.1186/s13059-019-1828-7
7. *Carvalho C.M., Lupski J.R.* Mechanisms underlying structural variant formation in genomic disorders // Nat. Rev. Genet. 2016. V. 17. № 4. P. 224–238. https://doi.org/10.1038/nrg.2015.25
8. *Sedlazeck F.J., Lee H., Darby C.A. et al.* Piercing the dark matter: bioinformatics of long-range sequencing and mapping // Nat. Rev. Genet. 2018. V. 19. № 6. P. 329–346. https://doi.org/10.1038/s41576-018-0003-4
9. *Collins R.L., Brand H., Karczewski K.J. et al.* A structural variation reference for medical and population genetics // Nature. 2021. V. 581. P. 444–451. https://doi.org/10.1038/s41586-020-2287-8
10. *Weischenfeldt J., Symmons O., Spitz F. et al.* Phenotypic impact of genomic structural variation: Insights from and for human disease // Nat. Rev. Genet. 2013. V. 14. № 2. P. 125–138. https://doi.org/10.1038/nrg3373
11. *Stankiewicz P., Lupski J.* Structural variation in the human genome and its role in disease // Annu. Rev. Med. 2010. V. 61. P. 437–455. https://doi.org/10.1146/annurev-med-100708-204735
12. *Schüle B., McFarland K.N., Lee K. et al.* Parkinson's disease associated with pure ATXN10 repeat expansion // NPJ Parkinson's Disease. 2017. V. 3. P. 27. https://doi.org/10.1038/s41531-017-0029-x
13. *Yeh Y.C., Ho H.L., Wu Y.C. et al.* AKT1 internal tandem duplications and point mutations are the genetic hallmarks of sclerosing pneumocytoma // Mod. Pathol. 2020. V. 33. № 3. P. 391–403. https://doi.org/10.1038/s41379-019-0357-y
14. *Schütte J., Reusch J., Khandanpour C. et al.* Structural variants as a basis for targeted therapies in hematological malignancies // Front. Oncol. 2019. V. 9. P. 839. https://doi.org/10.3389%2Ffonc.2019.00839
15. *Ewing A., Meynert A., Churchman M. et al.* Structural variants at the BRCA1/2 loci are a common source of homologous repair deficiency in high-grade serous ovarian carcinoma // Clin. Cancer Res. V. 27. № 11. P. 3201–3214. https://doi.org/10.1158/1078-0432.ccr-20-4068
16. *Malhotra D., Sebat J.* CNVs: Harbingers of a rare variant revolution in psychiatric genetics // Cell. 2012. V. 148. № 6. P. 1223–1241. https://doi.org/10.1016%2Fj.cell.2012.02.039
17. *Huse K., Taudien S., Groth M. et al.* Genetic variants of the copy number polymorphic β-defensin locus are associated with sporadic prostate cancer // Tumor Biol. 2008. V. 29. № 2. P. 83–92. https://doi.org/10.1159/000135688
18. *Wellcome Trust Case Control Consortium*. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls // Nature. 2010. V. 464. P. 713–720. https://doi.org/10.1038/nature08979
19. *Fanciulli M., Norsworthy P.J., Petretto E. et al.* FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity // Nat. Genet. 2007. V. 39. № 6. P. 721–723. https://doi.org/10.1038/ng2046
20. *Fellermann K., Stange D.E., Schaeffeler E. et al.* A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to crohn disease of the colon // Am. J. Hum. Genet. 2006. V. 79. № 3. P. 439–448. https://doi.org/10.1086/505915
21. *Hollox E.J., Huffmeier U., Zeeuwen P.L. et al.* Psoriasis is associated with increased β-defensin genomic copy number // Nat. Genet. 2008. V. 40. № 1. P. 23–25. https://doi.org/10.1038/ng.2007.48
22. *Zook J.M., Hansen N.F., Olson N.D. et al.* A robust benchmark for detection of germline large deletions and insertions // Nat. Biotechnol. 2020. V. 38. P. 1347–1355. https://doi.org/10.1038/s41587-020-0538-8
23. *Ye J., Coulouris G., Zaretskaya I. et al.* Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction // BMC Bioinformatics. 2012. V. 13. № 1. P. 134. https://doi.org/10.1186/1471-2105-13-134
24. Sequencing analysis viewer [Electronic resource] // Sequencing analysis viewer support. URL: https://support.illumina.com/sequencing/sequencing_software/sequencing_analysis_viewer_sav.html (accessed: 15.09.2022).
25. bcl2fastq [Electronic resource] // bcl2fastq. URL: https://emea.support.illumina.com/sequencing/sequencing_software/bcl2fastq-conversion-software/downloads.html (accessed: 13.05.2022).
26. FastQC [Electronic resource] // Babraham bioinformatics FastQC a quality control tool for high throughput sequence data. URL: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed: 13.05.2022).
27. Illumina DRAGEN Bio-IT Platform [Electronic resource] // Illumina DRAGEN Bio-IT platform | variant calling & secondary genomic analysis software tool. URL: https://www.illumina.com/products/by-type/informatics-products/dragen-bio-it-platform.html (accessed: 13.05.2022).
28. *Schneider V.A., Graves-Lindsay T., Howe K. et al.* Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference

assembly // Genome Res. 2017. V. 27. № 5. P. 849–864.
https://doi.org/10.1101/gr.213611.116

29. *Chen X., Schulz-Trieglaff O., Shaw R. et al.* Manta: Rapid detection of structural variants and indels for germline and cancer sequencing applications // Bioinformatics. 2016. V. 32. № 8. P. 1220–1222.
https://doi.org/10.1093/bioinformatics/btv710
30. Smoove [Electronic resource] // *Brent S.* Smoove: structural-variant calling and genotyping with existing tools. URL: https://github.com/brentp/smoove (accessed: 13.07.2022).
31. *Rausch T., Zichner T., Schlattl A. et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis // Bioinformatics. 2012. V. 28. № 18. P. i333–i339.
https://doi.org/10.1093/bioinformatics/bts378
32. NCBI [Electronic resource] // GRCh37 hg19 genome assembly NCBI. URL: https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.13/ (accessed: 15.09.2022).
33. *Zhao H., Sun Z., Wang J. et al.* CrossMap: A versatile tool for coordinate conversion between genome assemblies // Bioinformatics. 2014. V. 30. № 7. P. 1006–1007.
https://doi.org/10.1093/bioinformatics/btt730
34. GDC [Electronic resource] // GDC viewer docs. URL: https://docs.gdc.cancer.gov/Data_Dictionary/viewer/#?view=table-definition-view&id=alignment_cocleaning_workflow (accessed: 13.07.2022).
35. *Li H., Durbin R.* Fast and accurate long-read alignment with Burrows–Wheeler transform // Bioinformatics. 2010. V. 26. № 5. P. 589-595.
https://doi.org/10.1093/bioinformatics/btp698
36. *Danecek P., Bonfield J.K., Liddle J. et al.* Twelve years of samtools and bcftools // GigaScience. 2021. V. 10. № 2.
https://doi.org/10.1093/gigascience/giab008
37. *Pedersen B.S., Quinlan A.R.* Duphold: scalable, depth-based annotation and curation of high-confidence structural variant calls // GigaScience. 2019. V. 8. № 4.
https://doi.org/10.1093/gigascience/giz040
38. *Jeffares D.C., Jolly C., Hoti M. et al.* Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast // Nat. Commun. 2017. V. 8. P. 14061.
https://doi.org/10.1038/ncomms14061
39. Witty.er [Electronic resource] // *Wan Y., Ho K.* Witty.er. URL: https://github.com/Illumina/witty.er (accessed: 13.07.2022).
40. Unipro UGENE [Electronic resource] // Unipro UGENE: a unified bioinformatics toolkit. Bioinformatics. Oxford academic. URL: https://academic.oup.com/bioinformatics/article/28/8/1166/195474?login=false (accessed: 13.07.2022).
41. *Robinson J.T., Thorvaldsdottir H., Winckler W. et al.* Integrative genomics viewer // Nat. Biotechnol. 2011. V. 29. № 1. P. 24–26.
https://doi.org/10.1038/nbt.1754

Comparative Analysis of Structural Variant Callers on the Short-Read Whole-Genome Sequencing Data

A. A. Mkrtchian[a], *, K. S. Grammatikati[a], P. G. Kazakova[a], S. I. Mitrofanov[a], P. U. Zemsky[a], A. A. Ivashechkin[a], M. N. Pilipenko[a], D. V. Svetlichny[a], A. P. Sergeev[a], E. A. Snigir[a], L. V. Frolova[a], T. A. Shpakova[a], V. S. Yudin[a], A. A. Keskinov[a], S. M. Yudin[a], and V. I. Skvortsova[b]

[a] *Federal State Budgetary Institution "Centre for Strategic Planning and Management of Biomedical Health Risks" of the Federal Medical Biological Agency, Moscow, 119121 Russia*

[b] *The Federal Medical Biological Agency of Russia, Moscow, 123182 Russia*

**e-mail: AMkrtchyan@cspfmba.ru*

In this study three structural variant callers (Manta, Smoove, Delly) were analysed on the whole-genome sequencing data using four different alignment algorithms: DRAGEN, GDC DNA-Seq Alignment Workflow, GDC DNA-Seq Alignment Workflow + GDC DNA-Seq Co-Cleaning Workflow, NovoAlign, different lengths of raw reads: 2 × 150 bp and 2 × 250 bp, different mean genome coverage values. Results were compared to etalon results of GIAB team. Structural variants validation was hold also with Sanger sequencing. Structural variants deletions and insertions as it turned out were best determined with Manta tool. We've got 89–96% of accuracy and 59–70% of sensitivity for analysed deletions, and also 96–99% of accuracy and 15–36% of sensitivity for insertions. Smoove and Delly showed less accurate and sensitive results (Smoove: 91–95% of accuracy and 8–54% of sensitivity for deletions, Delly: 78–87% of accuracy and 31–66% of sensitivity for deletions, 99–100% of accuracy and 1–13% of sensitivity for insertions). Simultaneous using of two or even three structural variant callers didn't give a rise of accuracy and sensitivity for deletions. Analysis showed that accuracy and sensitivity of structural variant callers rise with the rising of mean genome coverage value, increasing of reads length from 150 to 250 bp influence in to varying degrees on the accuracy and sensitivity of individual tools. Another inference of this study was that accuracy of structural variants callers vary depends on structural variants size range. For example, Manta finds better deletions in the range from 200 and more bp, Delly – from 1000 to 10000 bp, Smoove – from 200 to 10000 bp.

Keywords: structural variants, WGS, short-read sequencing, alignment, GIAB.