НАУЧНАЯ СТАТЬЯ | SCIENTIFIC ARTICLE
УДК 577.21
https://doi.org/10.35825/2587-5728-2023-7-4-366-383
https://elibrary.ru/jvpqyq

Современные биоинформационные решения, используемые для анализа генетических данных

Я.А. Кибирев, А.В. Кузнецовский, С.Г. Исупов, И.В. Дармов

Филиал федерального государственного бюджетного учреждения «48 Центральный научно-исследовательский институт» Министерства обороны Российской Федерации (г. Киров), 610000, Российская Федерация, г. Киров, Октябрьский проспект, д. 119
e-mail: 23527@mil.ru

Поступила 11.11.2023 г. Принята к публикации 27.12.2023 г.

Эффективное противодействие биологическим угрозам как природного, так и техногенного характера требует наличия средств и методов быстрой и достоверной идентификации микроорганизмов и всестороннего изучения их основных биологических свойств. За последнее десятилетие арсенал отечественных микробиологов пополнили многочисленные методы анализа геномов патогенов, в первую очередь, основанные на секвенировании нуклеиновых кислот. *Цель работы* – выявить возможности современного технического и методического арсенала, применяемого для углубленного молекулярно-генетического изучения микроорганизмов, в том числе биоинформационных решений, используемых для анализа генетических данных. *Источниковая база исследования* – англоязычная научная литература, доступная через сеть «Интернет», документация биоинформационного программного обеспечения. *Метод исследования* – анализ научных источников от общего к частному. Рассматривали особенности платформ для секвенирования, основные этапы анализа генетической информации, актуальные биоинформационные утилиты, их взаимодействие и организацию в единый рабочий процесс. *Результаты и обсуждение*. Производительность современных генетических анализаторов позволяет проводить полную расшифровку бактериального генома в течение одних суток, включая время, требуемое для подготовки пробы к исследованию. Ключевым фактором, во многом определяющим эффективность применяемых молекулярно-генетических средств, является знание и грамотное применение соответствующего программного обеспечения. К основным этапам стандартного биоинформационного анализа первичных генетических данных относятся оценка качества секвенирования, предварительная обработка данных, их картирование на референсный геном или сборка генома *de novo*, аннотирование генома, типирование и выявление значимых генетических детерминант (устойчивости к антибактериальным препаратам, факторов патогенности и т.д.), филогенетический анализ. Для каждого из этапов разработаны биоинформационные утилиты, отличающиеся реализованными в них алгоритмами анализа. *Заключение*. С учетом специфики деятельности подразделений войск РХБ защиты ВС РФ, из числа известных программных продуктов наибольший интерес представляют утилиты с открытым исходным кодом, не требующие для своей работы доступа к удаленным ресурсам.

Ключевые слова: *биоинформатика; генетический анализ; идентификация; микроорганизмы; нуклеиновые кислоты; программное обеспечение; секвенирование.*

Для цитирования: *Кибирев Я.А., Кузнецовский А.В., Исупов С.Г., Дармов И.В. Современные биоинформационные решения, используемые для анализа генетических данных. Вестник войск РХБ защиты. 2023;7(4):366–383. EDN:jvpqyq.*
https://doi.org/10.35825/2587-5728-2023-7-4-366-383

Modern Bioinformatics Solutions Used for Genetic Data Analysis

Ya.A. Kibirev, A.V. Kuznetsovskiy, S.G. Isupov, I.V. Darmov

Branch Office of the Federal State Budgetary Establishment «48 Central Scientific Research Institute» of the Ministry of Defence of the Russian Federation (Kirov), Oktyabrsky Avenue 119, Kirov 610000, Russian Federation
e-mail: 23527@mil.ru

Received November 11, 2023. Accepted December 27, 2023
Effective counteraction to biological threats, both natural and man-made, requires the availability of means and methods for rapid and reliable microorganism identification and a comprehensive study of their basic biological properties. Over the past decade, the arsenal of domestic microbiologists has been supplemented by numerous methods for analyzing the genomes of pathogens, primarily based on nucleic acid sequencing. *The purpose of this work* is to provide the reader with information about capabilities of modern technical and methodological arsenal used for in-depth molecular genetic study of microorganisms, including bioinformatics solutions used for the genetic data analysis. *The source base for this research* is English-language scientific literature available via the Internet, bioinformation software documentation. *The research method* is an analysis of scientific sources from the general to the specific. We considered the features of sequencing platforms, the main stages of genetic information analysis, current bioinformation utilities, their interaction and organization into a single workflow. *Results and discussion*. The performance of modern genetic analyzers allows for complete decoding of the bacterial genome within one day, including the time required to prepare the sample for research. The key factor that largely determines the effectiveness of the genetic analysis methods used is the competent use of the necessary bioinformatics software utilities. Standard stages of primary genetic data analysis are assessment of the quality control, data preprocessing, mapping to a reference genome or *de novo* genome assembly, genome annotation, typing and identification of significant genetic determinants (resistance to antibacterial drugs, pathogenicity factors, etc.), phylogenetic analysis. For each stage bioinformation utilities have been developed, differing in implemented analysis algorithms. *Conclusion*. Open source utilities that do not require access to remote resources for their operation are of greatest interest due to activities specifics of NBC protection corps units.

Keywords: bioinformatics; genetic analysis; identification; microorganisms; nucleic acids; sequencing; software.

For citation: Kibirev Ya.A., Kuznetsovskiy A.V., Isupov S.G., Darmov I.V. Modern Bioinformatics Solutions Used for Genetic Data Analysis. Journal of NBC Protection Corps. 2023;7(4):366–383. EDN:jvpqyq. https://doi.org/10.35825/2587-5728-2023-7-4-366-383

Обеспечение биологической безопасности является одной из ключевых составляющих общей системы безопасности любого государства, обеспечивающей его стабильное развитие. Несмотря на все усилия, предпринимаемые как на региональном, так и на международном уровне, ни одной из стран, независимо от уровня развития, не удалось достичь состояния устойчивого биологического благополучия.

В последние десятилетия в мире отмечается стойкая тенденция к активизации природных очагов, распространению инфекционных заболеваний на новые для них географические территории, страны и континенты. Серьезное беспокойство специалистов также вызывают участившиеся случаи выявления ранее не встречавшихся патогенов – возбудителей атипичных вариантов заболеваний либо новых нозологических форм. В качестве наиболее ярких примеров, имевших серьезные социально-экономические последствия, можно привести вспышки, вызванные вирусами SARS-CoV (2002 г.) и MERS-CoV (2012 г.), лихорадок Зика (2015 г.) и Чикунгунья (2014 г.), эпидемию «свиного» гриппа H1N1 (2009 г.), вызванную кишечной палочкой *Escherichia coli* O104:H4 инфекцию, сопровождающуюся гемолитико-уремическим синдромом (2011 г.). К этому же списку, безусловно, относится и начавшаяся в конце 2019 г. пандемия новой коронавирусной инфекции, унесшей, по официальным данным ВОЗ, жизни почти 7 млн человек[1] [1].

Вместе с ростом нестабильности и напряженности в международных отношениях

[1] WHO Coronavirus (COVID-19) Dashboard. URL: https://covid19.who.int (дата обращения: 28.10.2023).

растут и риски использования против России биологического оружия. Исследования по созданию такого оружия, в том числе разработке новых его видов, проводятся во многих зарубежных странах при полной государственной поддержке и финансировании[2].

Одним из ключевых элементов в системе по предупреждению и реагированию на угрозы биологической безопасности страны, применяемым для проведения эпидмониторинга, а также установления причин вспышек инфекционных заболеваний, включая выявление возможных признаков их искусственного происхождения, организации и проведения адекватных противоэпидемических мероприятий, разработки перспективных средств диагностики, специфической профилактики и лечения заболеваний, является эффективная лабораторная диагностика с применением самых современных методов и технологий изучения патогенных микроорганизмов. Данное утверждение неоднократно доказывалось на практике: так, опыт противодействия пандемии новой коронавирусной инфекции продемонстрировал возможности средств и методов углубленного молекулярно-генетического исследования патогенов, позволивших в кратчайшие сроки разработать медицинские изделия для *in vitro* диагностики заболевания [2, 3].

Цель работы – выявить возможности современного биоинформационного программного обеспечения, используемого при анализе генетических данных, на примере решения задач углубленного молекулярно-генетического изучения патогенных микроорганизмов.

Источниковая база исследования – англоязычная научная литература, доступная через сеть «Интернет», документация биоинформационного программного обеспечения.

Метод исследования – анализ научных источников от общего к частному. Рассматривали особенности платформ для секвенирования, основные этапы анализа генетической информации, актуальные биоинформационные утилиты, их взаимодействие и организацию в единый рабочий процесс.

Основная часть

Разработка и применение средств обработки, хранения и анализа информации неразрывно связана с научной деятельностью человека. Начиная с середины XX в., с момента интенсификации работ по изучению строения и функций биополимеров, в первую очередь, белков, и для исследователей-биологов и биохимиков очевидной стала необходимость эффективных инструментов для работы с постоянно растущими массивами биоданных.

С конца 1970-х гг., после внедрения в лабораторную практику метода определения последовательности нуклеиновых кислот (секвенирования), разработанного группой английских биохимиков под руководством Фредерика Сэнгера [4], перед исследователями встала задача обработки нового вида информации – генетической.

Период конца XX–начала XXI вв. стал следующим своеобразным переломным моментом в становлении биоинформатики как самостоятельной научной дисциплины. Во многом этому способствовало широкое распространение полностью автоматических приборов для секвенирования нуклеиновых кислот «по Сэнгеру», генерирующих генетические данные в количествах, не позволяющих производить их обработку в «ручном» режиме [5]. В это же время существенно выросли вычислительные возможности, а также емкость машинных носителей информации персональных компьютеров, что сделало работу с биоинформацией доступной не только для крупных профильных учреждений, но и для частных исследователей, студентов и – фактически – любого желающего. До настоящего времени, несмотря на все последующие разработки, секвенаторы первого поколения остаются чрезвычайно востребованными в решении задач, требующих установления последовательности относительно небольших (несколько сотен нуклеотидов) фрагментов нуклеиновых кислот[3].

Очередным значимым событием, подстегнувшим развитие биоинформатики, стало изобретение методов высокопроизводительного секвенирования нуклеиновых кислот – так называемое второе поколение технологий секвенирования. Их реализация на практике сделала реальностью выполнение полного секвенирования небольших – вирусных, бактериальных – геномов за один или несколько запусков оборудования. Недосягаемым ли-

[2] Материалы брифингов начальника войск радиационной, химической и биологической защиты Вооруженных Сил Российской Федерации генерал-лейтенанта Кириллова И.А. Телеграм-канал «Проект РХБЗ». URL: https://t.me/projectRHBZ (дата обращения: 28.10.2023).

[3] New Genetic Analyzer Brings Advanced Capabilities to Sanger Sequencing and Fragment Analysis. https://www.businesswire.com/news/home/20220208005374/en/New-Genetic-Analyzer-Brings-Advanced-Capabilities-to-Sanger-Sequencing-and-Fragment-Analysis (дата обращения: 28.10.2023).

дером в указанной области стала и остается до настоящего времени технология, предложенная английской компанией «Solexa» (приобретена в 2007 г. американской корпорацией «Illumina»)[4] [6, 7].

Вторым ключевым игроком на рынке секвенирования «нового поколения» (Next Generation Sequensing, NGS) стала компания «Ion Torrent Systems Inc.» (США) (в настоящее время входит в состав концерна «Thermo Fisher Scientific Inc.») со своими генетическими анализаторами «Proton», «Personal Genome Machine» и др. Разработанная компанией технология ионного полупроводникового секвенирования обеспечивала относительно невысокую стоимость и заметно большую, чем у продуктов «Illumina»/«Solexa», скорость анализа. Существенным недостатком таких секвенаторов, так до конца и не преодоленным по настоящее время, является высокая частота ошибок, особенно при прочтении последовательностей, богатых гомополимерными участками (повторами одинаковых нуклеотидов)[5] [8].

Новые решения в области секвенирования нуклеиновых кислот характеризовались не только высочайшей производительностью, но и позволили значительно снизить стоимость генетических исследований (в расчете на единицу объема получаемых данных), только за период 2007–2008 гг. – более чем в 100 раз (рисунок 1).

Обратной стороной доступности полногеномного секвенирования стало кратное увеличение объемов получаемой биоинформации. Например, наиболее производительные модели выпускаемых компанией «Illumina» NGS-секвенаторов способны генерировать до 16 Тб первичных генетических данных за один запуск (48 ч)[6].

Следующим шагом в развитии технологии секвенирования нуклеиновых кислот, ее третьим поколением стали решения, обеспечившие драматическое увеличение средней длины прочтения.

Секвенаторы второго поколения, как «Illumina», так и «Ion Torrent», в силу особенностей, заложенных в их конструкции методических и технических решений, генерируют первичные данные в виде массива относительно коротких – несколько сотен нуклеотидов – фрагментов последовательностей («короткие риды», от английского «read» – «прочтение»).

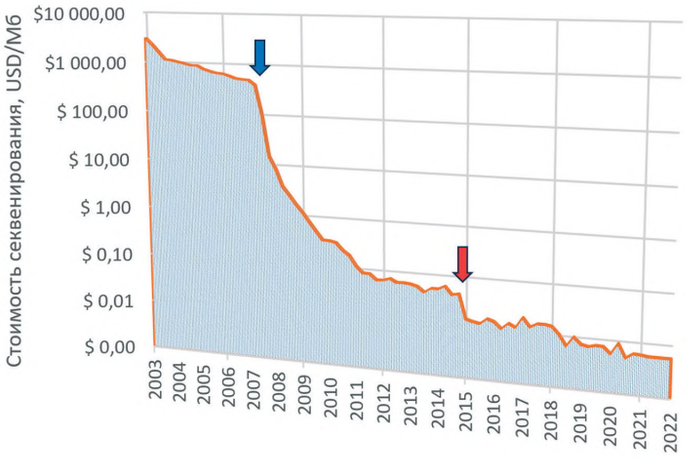

Рисунок 1 – *Динамика изменения стоимости секвенирования ДНК (по данным Национального института изучения генома человека, США). Хорошо виден эффект от внедрения в практику методов секвенирования второго (стрелка синего цвета) и третьего (стрелка красного цвета) поколения (по данным Национального института изучения генома человека, США; URL: https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data, дата обращения: 28.10.2023)*

[4] Illumina sequencing platforms. URL: https://www.illumina.com/systems/sequencing-platforms.html (дата обращения: 28.10.2023).

[5] Next-Generation Sequencing (NGS). URL: https://www.thermofisher.com/ru/ru/home/life-science/sequencing/next-generation-sequencing.html (дата обращения: 28.10.2023).

[6] NovaSeq X Series. URL: https://www.illumina.com/systems/sequencing-platforms/novaseq-x-plus.html (дата обращения: 28.10.2023).

Первой компанией, выпустившей на рынок секвенатор, преодолевший психологически важный барьер длины прочтения в 10 тыс. нуклеотидов, стала в 2010 г. американская компания Pacific Biosciences of California, Inc. (PacBio) со своим прибором «PacBio RS» [9]. Технология «мономолекулярного секвенирования в реальном времени» (SMRT, Singe Molecule sequencing in Real Time) не требовала предварительной амплификации ДНК в анализируемом образце, а длина чтения потенциально упрощала последующую сборку геномов. В более поздних моделях, благодаря совершенствованию технической составляющей, оптимизации реагентной базы и программных алгоритмов, удалось значительно снизить частоту ошибок и повысить производительность анализа. Также компанией разработана аппаратная платформа «Onso», реализующая короткие прочтения по собственной оригинальной методике[7].

В отличие от решений PacBio, приборы компании Oxford Nanopore Technologies, в частности, ее первый выпущенный серийно в 2014 г. секвенатор «MinION», будучи чрезвычайно компактными и относительно недорогими, позволяли – в идеальных условиях – достигать длины прочтения, измеряющейся сотнями тысяч нуклеотидов [10]. Немаловажно и то, что разработанная компанией технология нанопорового секвенирования позволяла достаточно легко масштабировать оборудование, обеспечивая потребности максимально широкого круга потенциальных потребителей[8].

Повсеместное использование NGS привело и к кратному росту нагрузки на онлайн-сервисы, осуществляющие хранение и анализ биоинформационных данных. Так, в настоящее время количество уникальных записей в двух ведущих общедоступных базах GenBank и WGS Project превышает 200 млн и 2 млрд соответственно, а их суммарная «протяженность» составляет более 2 трлн и 20 трлн нуклеотидов соответственно. При этом объем информации, депонированной в этих базах, удваивается в среднем каждые 18 месяцев (рисунок 2).

Необходимо понимать, что информация, депонированная в таких общедоступных базах – это только своеобразная вершина айсберга. В процессе своей повседневной деятельности специалисты вынуждены обрабатывать колоссальные массивы данных, значительно превышающие приведенные выше значения. Так, по самым осторожным оценкам, только исследования генома человека, ставшие за последнее десятилетие вполне рутинными для многих научных и медицинских учреждений, генерируют ежегодно более 40 эксабайт (40×10^{18} байт) генетических данных. Для понимания масштаба информационного «цунами» можно отметить, что все видеофайлы, загружаемые пользователями за год на сервера самого популярного видеохостинга YouTube, суммарно не превышают 2 эксабайт[9].

Как бы то ни было, широкое применение на практике NGS-решений явилось одним из основных стимулов развития аппаратных

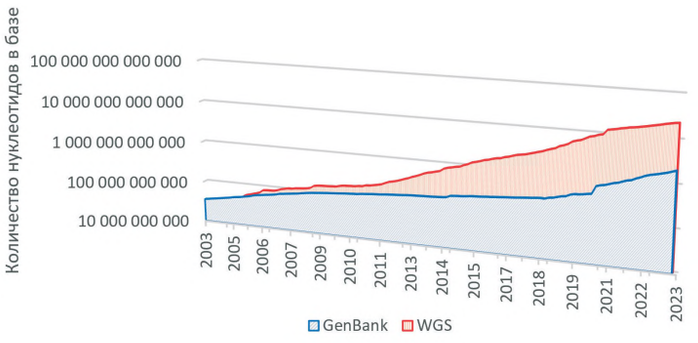

Рисунок 2 – *Динамика накопления генетических данных в базах GenBank и WGS Project (по статистике проектов GenBank и WGS Project; URL: https://www.ncbi.nlm.nih.gov/genbank/statistics, дата обращения: 28.10.2023)*

[7] Sequencing systems – PacBio. URL: https://www.pacb.com/sequencing-systems (дата обращения: 28.10.2023).

[8] Oxford Nanopore flow cells and devices. URL: https://nanoporetech.com/products/sequence (дата обращения: 28.10.2023).

[9] Big Data among Big Data: Genome Data. URL: https://3billion.io/blog/big-data-among-big-data-genome-data (дата обращения: 28.10.2023).

и программных средств обработки получаемой информации, окончательно сформировав представление о биоинформатике как самостоятельной научной дисциплине, применяющей компьютерные методы и математические подходы для получения, анализа, хранения, организации и визуализации различных биологических данных, в первую очередь, связанных со структурой, функцией и эволюцией биологических макромолекул – белков и нуклеиновых кислот.

Огромный потенциал биоинформатики вывел ее из стен академических научных учреждений в практическое поле, сделав незаменимым инструментом при решении самых разноплановых задач в таких областях деятельности человека, как информатика и инженерия, математика и статистика, физика и химия, биология и биотехнология, генетика и геномика, медицина и фармацевтика, ветеринария и сельское хозяйство и многие другие.

Безусловно, все многообразие арсенала современной биоинформатики невозможно осветить в рамках одной статьи. В этой связи, а также учитывая общность основных принципов и подходов к анализу различных типов биоданных, было решено продемонстрировать возможности современных биоинформационных инструментов на примере решения задач анализа генетической информации о патогенных микроорганизмах в целях их идентификации и лабораторной диагностики вызываемых заболеваний[10] [11–17].

В общем виде, стандартный протокол генетического исследования пробы с применением методов NGS включает в себя всего три этапа: подготовка образца к анализу, непосредственно секвенирование и обработка полученных данных. Именно третий этап является областью интереса биоинформатики. За кажущейся простотой названия скрывается целый комплекс задействуемых математических алгоритмов и компьютерных программ (рисунок 3).

Первый обязательный этап процесса обработки NGS-данных – контроль качества первичных результатов секвенирования. Вероятность правильной идентификации каждого нуклеотида в анализируемой последовательности в силу технических ограничений используемого оборудования всегда меньше единицы. Численным выражением «правильности» установления основания является показатель качества Q, связанный с вероятностью данного события P формулой Q = -10lgP. Типичные значения Q находятся в диапазоне от 0 до 40, хорошим для технологии Illumina считается качество выше 30, что соответствует вероятности ошибки не более 0,001. В наиболее распространенном файловом формате представления первичных результатов секвенирования fastq показатель качества Q для каждого нуклеотида указывается в виде ASCII-символа в соответствии с одним из стандартов кодировки, поддерживаемых производителем оборудования (рисунок 4).

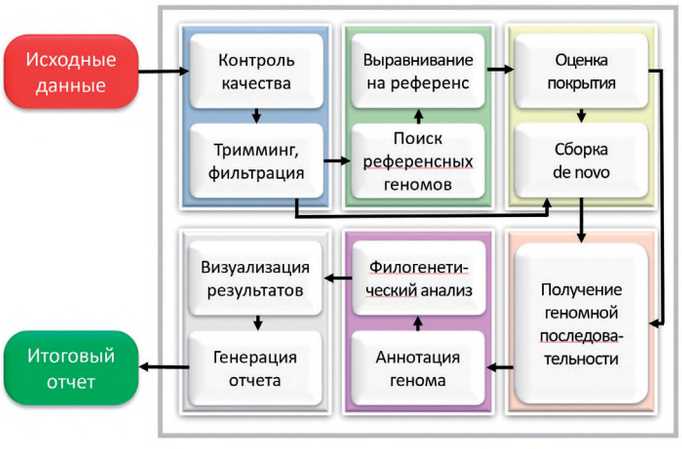

Рисунок 3 – *Биоинформационный «черный ящик» рабочего процесса анализа NGS-данных (рисунок авторов)*

[10] В настоящей статье не рассматривается базовое программное обеспечение, поставляемое вместе с генетическими анализаторами. Также за рамками статьи оставлены коммерческие программные продукты для анализа генетической информации в силу их ограниченной доступности, закрытости исходных файлов и невозможности контролировать и ограничивать передачу обрабатываемой информации на удаленные сервера. По этой же причине не описаны открытые онлайн-сервисы типа Galaxy, требующие для своей работы выгрузки всех первичных данных в глобальную сеть.

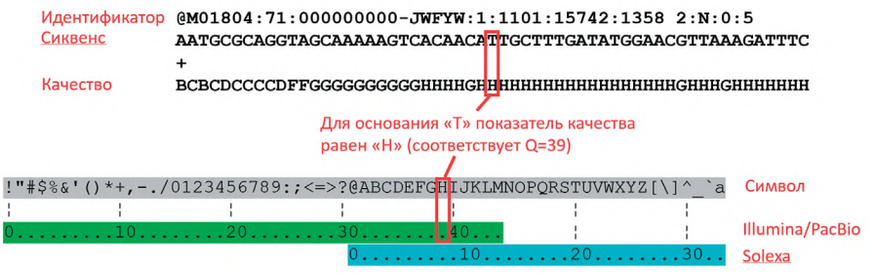

Рисунок 4 – *Кодирование качества секвенирования в fastq-файле (рисунок авторов). Качество для каждого нуклеотида в файле формата fasta (фрагмент файла приведен в верхней части рисунка) записывается ASCII-символом (серая шкала). Наиболее распространены две кодировки качества секвенирования: применяемая в системах «Solexa» и ранних версиях «Illumina» (кодируется символом ASCII, соответствующим численному значению качества, увеличенному на 64, зеленая шкала), применяемая в поздних версиях «Illumina», «PacBio» и «Nanopore» (кодируется символом ASCII, соответствующим численному значению качества, увеличенному на 33, голубая шкала). Например, выделенный красной рамкой нуклеотид «T» имеет качество «H», что по системе «+33» кодирует значение качества 39 (соответствует вероятности ошибки P, равной =0,00013)*

Для каждой первичной последовательности с использованием специализированной утилиты FastQC[11] или любого из ее аналогов рассчитывается обобщенный показатель качества; полученные результаты позволяют, в том числе визуально, оценить пригодность первичных данных к дальнейшему анализу и провести более гибкую настройку утилит, применяемых на следующих этапах (рисунок 5).

С учетом результатов оценки качества далее выполняется предварительная обработка первичного массива данных путем удаления (тримминг) «технических» последовательностей, оставшихся после пробоподготовки образца (праймеры, адаптеры), исключение из дальнейшего анализа последовательностей с низким качеством прочтения, дефектных по длине, объединение прямых и обратных парных последовательностей и т.д.

Наиболее популярными продуктами, используемыми для тримминга, являются утилиты Trimmomatic[12], Fastp[13] и Cutadapt[14]. Разработаны и активно применяются и другие программные решения, в том числе узкоспециализированные. Так, для фильтрации ридов по качеству может быть использована утилита Prinseqlite[15], для соединения парных прочтений – Pear (Pair-End AssembleR)[16], отдельные частные задачи возможно решить с применением программы Seqtk[17].

Крайне желательно на данном этапе провести дополнительную фильтрацию массива первичных данных путем удаления последовательностей, заведомо не представляющих интерес в рамках проводимого исследования [18–22].

Например, является доказанным фактом, что любой клинический образец содержит в себе фрагменты нуклеиновых кислот человека (пациента, в некоторых случаях – еще и исследователя). Соответственно, часть первичных NGS-данных также будет представлять собой последовательности человеческой ДНК, при этом их число в общем пуле генетической информации может значительно превышать количество фрагментов бактериального ге-

[11] Babraham Bioinformatics – FastQC A quality control tool for high throughput sequence data. URL: https://www. bioinformatics.babraham.ac.uk/projects/fastqc (дата обращения: 28.10.2023).

[12] USADELLAB.org – Trimmomatic: A flexible read trimming tool for Illumina NGS data. URL: http://www.usadellab.org/cms/?page=trimmomatic (дата обращения: 28.10.2023).

[13] GitHub – OpenGene/fastp: An ultra-fast all-in-one FASTQ preprocessor. URL: https://github.com/opengene/fastp (дата обращения: 28.10.2023).

[14] Cutadapt – Cutadapt 4.6 documentation. URL: https://cutadapt.readthedocs.io/en/stable (дата обращения: 28.10.2023).

[15] GitHub – alces-software/packager-base: prinseqlite. URL: https://github.com/alces-software/packager-base/blob/master/apps/prinseqlite (дата обращения: 28.10.2023).

[16] GitHub – tseemann/PEAR: Pair-End AssembeR. URL: https://github.com/tseemann/PEAR (дата обращения: 28.10.2023).

[17] GitHub – lh3/seqtk: Toolkit for processing sequences in FASTA/Q formats. URL: https://github.com/lh3/seqtk (дата обращения: 28.10.2023).

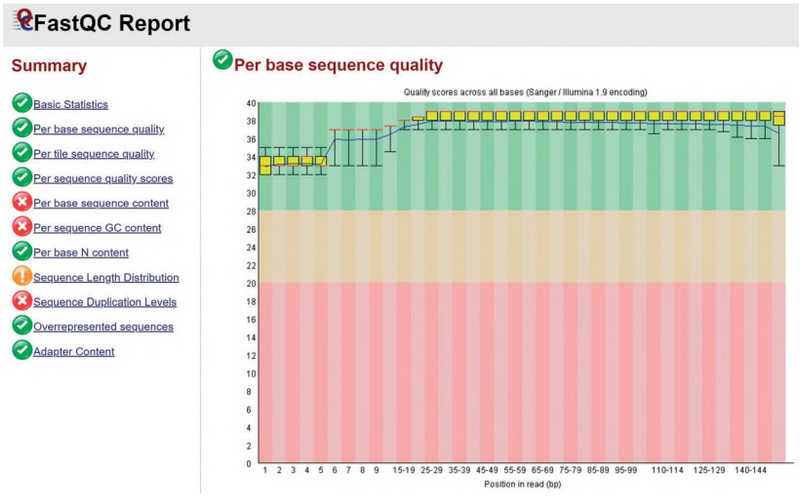

Рисунок 5 – *Фрагмент отчета оценки качества первичных данных утилитой FastQC (данные авторов). Несмотря на высокое качество секвенирования в целом (прочтения расположены в «зеленой зоне», соответствующей качеству выше 20, по всему диапазону длин), анализ выявил существенные отклонения от среднестатистических значений по таким показателям, как частота встречаемости оснований «Per base sequence content», однородность содержания G+C «Per sequence GC content», а также равномерность частоты повторения прочтений «Sequence Duplication Levels»*

нома. Особо остро указанная проблема стоит в отношении образцов, заведомо содержащих большое количество клеточного материала человека (биоптаты, гистопрепараты, цельная кровь и др.) [18, 19].

Очевидно, что для целей генетического изучения патогенного микроорганизма, предположительно находящегося в анализируемой пробе, ДНК пациента представляет собой информационный «балласт», обработка которого потребует значительно больших временных и вычислительных затрат, и, более того, может стать причиной получения ошибочных результатов на следующих этапах. Серьезность ситуации подтверждается, в частности, результатами оценки качества депонированных бактериальных сиквенсов, показавшими, что более половины из них имеет признаки контаминации последовательностями генома человека [18].

Особое опасение исследователей вызывают неиллюзорные риски, связанные с контаминацией референсных геномов, используемых в качестве образца сравнения при сборке и анализе NGS-данных в последующих исследованиях. Более того, большая часть используемых для фильтрации первичной информации программ также опираются в своей работе на референсные геномы. Следствием этого может стать лавинообразная генерация ошибочных данных с плохо прогнозируемыми последствиями [18].

В идеальном случае, когда имеются сведения о таксономической принадлежности изучаемого микроорганизма, возможно провести так называемую «положительную» фильтрацию, выделив из пула первичных данных только потенциально интересующие последовательности.

На практике же исследователь значительно чаще сталкивается с образцами, в отношении которых отсутствует достоверная информация о виде (роде, семействе и т.д.) патогена. В таких случаях применяются приемы «отрицательной» фильтрации данных путем исключения из анализа нуклеотидных последовательностей, заведомо не относящихся к области интереса.

Наиболее простой вариант такой фильтрации – использование универсальных программ-картировщиков, например, BWA (Burrow-Wheeler Aligner)[18], bowtie2[19], BBMap[20] и SNAP (Scalable Nucleotide Alignment

[18] Burrows-Wheeler Aligner. URL: https://bio-bwa.sourceforge.net (дата обращения: 28.10.2023).

[19] GitHub – BenLangmead/bowtie2: A fast and sensitive gapped read aligner. URL: https://github.com/benlangmead/bowtie2 (дата обращения 28.10.2023).

[20] BBTools – DOE Joint Genome Institute. URL: https://jgi.doe.gov/data-and-tools/software-tools/bbtools (дата обращения: 28.10.2023).

Program)[21], в отношении референсного генома человека.

Более эффективным является применение специализированных программ для контроля генетической контаминации. Большая часть таких утилит в своей работе использует собственные референсные базы маркерных детерминант, например, генов рибосомальной РНК или рибосомных белков, с которыми и проводится сравнение анализируемых данных с их последующей фильтрацией. В качестве примера таких программ можно привести CheckM[22], EukCC[23], Forty-Two и некоторые другие [18].

Менее распространены референс-независимые утилиты, производящие «сортировку» первичных данных на основании определения таких параметров, как содержание «G+C», частоты встречаемости коротких, длиной 4-6 нуклеотидов, последовательностей и некоторых других характеристик, значения которых отличаются у геномов разных таксонов. Из числа программ данной группы одними из наиболее популярных являются BlobTools[24] и PhylOligo[25].

Необходимо понимать, что ни одна из описанных программ не способна обеспечить полное удаление генетического «балласта». Именно по этой причине настоятельно рекомендуется сочетать в применяемом рабочем процессе одновременное использование нескольких утилит, например, упомянутых выше Bowtie2 и SNAP [19].

На следующем этапе исследований осуществляют сборку подготовленных первичных данных в единую геномную последовательность.

Методически более простыми являются случаи, когда имеется возможность использовать ранее собранный референсный геном изучаемого вида. В такой ситуации применяют те же программы-картировщики, что и для фильтрации данных (BWA, Bowtie2, BBMap и др.).

Принцип работы всех таких программ, которых за последние годы было разработано несколько десятков, заключен в последовательном поиске для каждого из ридов наиболее похожего участка в референсном геноме с учетом допустимых алгоритмом и пользовательскими настройками несовпадений (нуклеотидных замен, инсерций, делеций) [23]. Результаты картирования записываются в отдельный текстовый файл в формате SAM (Sequence Alignment Map, карта выравнивания сиквенса), содержащий, в числе прочей информации, точную локализацию каждого рида и его последовательность, а также данные о качестве сопоставления с референсом [24].

При необходимости, качество покрытия референсного генома в целом, его кратность и равномерность можно оценить визуально, с применением специализированных утилит типа Tablet[26] (рисунок 6).

Далее применяют утилиты для поиска так называемых вариантов – отличий генома исследуемого образца от референсного, например программу GATK (Genome Analysis ToolKit)[27]. В ходе такого анализа первоначально отбираются все возможные варианты, далее осуществляется их фильтрация по заданным оператором критериям достоверности (частоты встречаемости в данных секвенирования, для исключения влияния случайных ошибок), после чего на основании полученных данных генерируется новый полногеномный сиквенс.

Правильный выбор референсного образца во многом определяет результат проведенной работы. Очевидно, что чем дальше в эволюционном и филогенетическом плане будет отстоять от него изучаемый микроорганизм, тем более различны будут их геномы и меньше вероятность получить полное покрытие. В таких случаях исследователь может провести повторное выравнивание на другой референсный геном либо перейти к алгоритму сборки *de novo*.

Кроме того, для бактериальных патогенов необходимо дополнительно учитывать, что их клетки, помимо хромосомы, могут также

[21] GitHub – amplab/snap: Scalable Nucleotide Alignment Program. URL: https://github.com/amplab/snap (дата обращения: 28.10.2023).

[22] GitHub – Ecogenomics/CheckM: Assess the quality of microbial genomes recovered from isolates, single cells, and metagenomes. URL: https://github.com/Ecogenomics/CheckM (дата обращения: 28.10.2023).

[23] GitHub – EBI-Metagenomics/EukCC: Tool to estimate genome quality of microbial eukaryotes. URL: https://github.com/ebi-metagenomics/eukcc (дата обращения: 28.10.2023).

[24] GitHub – DRL/blobtools: Modular command-line solution for visualisation, quality control and taxonomic partitioning of genome datasets. URL: https://github.com/drl/blobtools (дата обращения: 28.10.2023).

[25] Releases – itsmeludo/PhylOligo - GitHub. https://github.com/itsmeludo/PhylOligo/releases (дата обращения: 28.10.2023).

[26] Tablet: Information & Computational Sciences. URL: https://ics.hutton.ac.uk/tablet (дата обращения: 28.10.2023).

[27] GATK. URL: https://gatk.broadinstitute.org/hc/en-us (дата обращения: 28.10.2023).

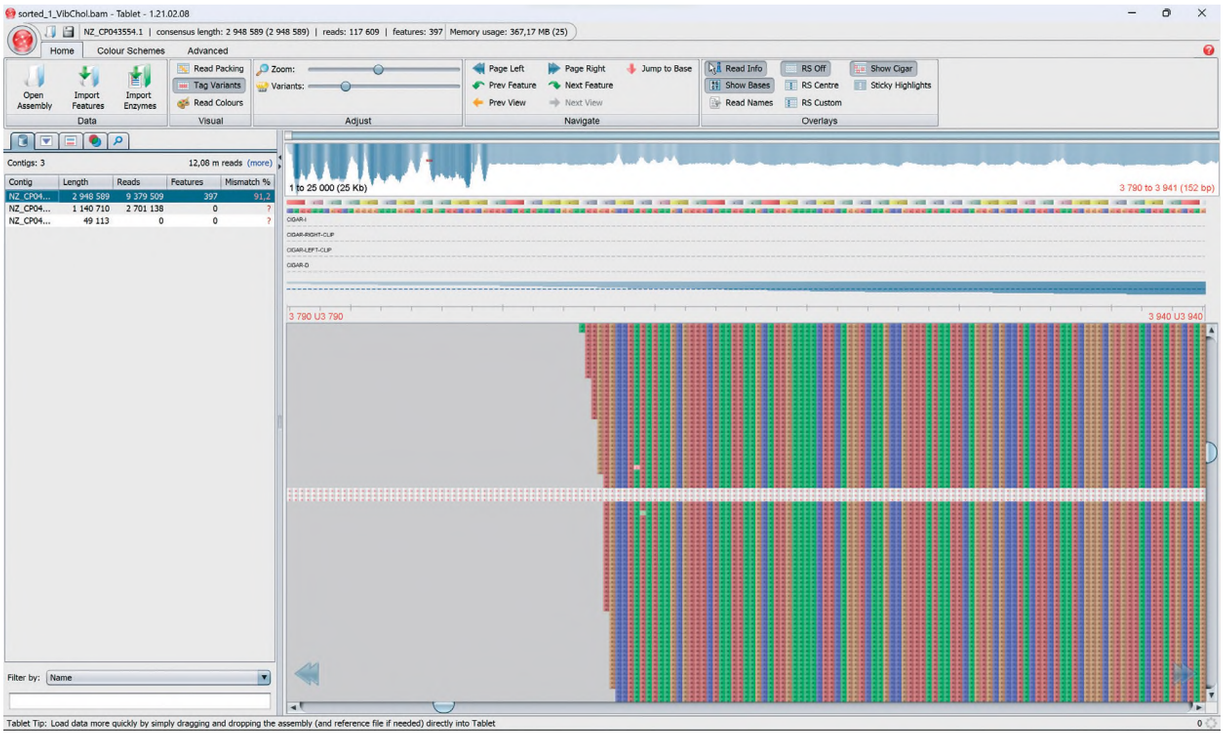

Рисунок 6 – *Визуализация покрытия референсного генома в программе Tablet (данные авторов). Каждая горизонтальная строка соответствует одному прочтению, картированному на данный участок референсного генома. Для удобства восприятия основания обозначены разными цветами*

нести одну или несколько плазмид, способных содержать генетические детерминанты факторов патогенности, устойчивости к антибактериальным препаратам и другие структуры, установление нуклеотидной последовательности которых имеет важное значение для всестороннего изучения микроорганизма [25, 26]. В зависимости от сложности задачи, помимо стандартных утилит, могут быть применены и специализированные – для поиска, сборки и аннотирования плазмидной ДНК и других мобильных генетических элементов, например, PlasmidID[28], PlasmidTron[29], plasmidSPAdes и metaplasmidSPAdes из набора утилит SPAdes[30] и некоторые другие.

Особое место в деятельности специалиста-биоинформатика занимает сборка генома изучаемого микроогранизма *de novo*, то есть без использования референса [7, 13–16]. В основе такой сборки лежит использование специальных программ – геномных ассемблеров, осуществляющих последовательную группировку и структурирование исходных результатов секвенирования (рисунок 7).

На первом этапе такой сборки частично перекрывающиеся риды собирают в более крупные последовательности – контиги (contig, от англ. contiguous – прилегающий, смежный). Далее, если имеется информация о взаимном расположении контигов в геноме, проводится их сборка в скаффолды (scaffold) – последовательность из нескольких неперекрывающихся контигов, разделенных промежутком известной длины. Это становится возможным, например, в случаях, когда в разные контиги попадают прямой и обратный риды одной пары, относительное положение в геноме и расстояние между которыми известно и связано с особенностями используемой платформы секвенирования [13, 14].

[28] GitHub – BU-ISCIII/plasmidID: PlasmidID is a mapping-based, assembly-assisted plasmid identification tool that analyzes and gives graphic solution for plasmid identification. URL: https://github.com/bu-isciii/plasmidid (дата обращения: 28.10.2023).

[29] GitHub – sanger-pathogens/plasmidtron: Assembling the cause of phenotypes and genotypes from NGS data. URL: https://github.com/sanger-pathogens/plasmidtron (дата обращения: 28.10.2023).

[30] GitHub – ablab/spades: SPAdes Genome Assembler. URL: https://github.com/ablab/spades (дата обращения: 28.10.2023).

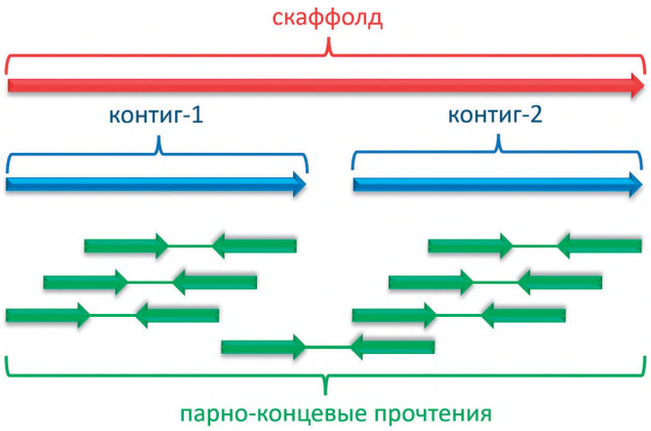

Рисунок 7 – *Иллюстрация взаимосвязи прочтений, контигов и скаффолдов (рисунок авторов). Связанные зеленые стрелки обозначают пары прямого и обратного прочтений, частично перекрывающиеся пары объединяются в общую последовательность, называемую контиг (стрелки голубого цвета). Благодаря тому, что имеется пара прочтений, попавших в разные контиги, их можно объединить в один скаффолд (стрелка красного цвета)*

Пожалуй, одним из наиболее популярных в мире геномных «сборщиков» является разработанный специалистами лаборатории «Центр алгоритмической биотехнологии» Санкт-Петербургского государственного университета пакет утилит SPAdes. Используя в своей работе оригинальную реализацию метода графов Де Брёйна, совместно с алгоритмами коррекции неравномерности покрытия, ошибок и артефактов секвенирования, SPAdes неизменно показывает высокие результаты в большинстве сравнительных испытаний по показателям, характеризующим эффективность сборки протяженных контигов и скаффолдов, а также геномов в целом[31] [27].

Среди других известных геномных ассемблеров, широко применяемых на практике для сборки малых (бактериальных, вирусных) геномов, можно упомянуть такие программы, как Velvet[32], INNUca[33], Canu[34], MEGAHIT[35], IDBA-UD[36] и некоторые другие. Каждая из них имеет свою «специализацию» в зависимости от реализованных в ней алгоритмов. Так, утилита Velvet более эффективно собирает контиги из коротких прочтений, в то время как Canu более оптимизирована для работы с длинными ридами.

По окончании сборки производят оценку ее качества с использованием встроенных утилит либо внешних программ типа QUAST[37], по результатам которой может быть принято решение о проведении повторного анализа, с другими параметрами [27].

Окончательным итогом работы по сборке становится файл в формате multi-fasta, содержащий последовательности всех контигов (скаффолдов) с их уникальными идентификаторами. Для заполнения промежутков контигов и окончательной сборки генома можно применять специализированные программные решения [28], однако их результа-

[31] SPAdes – Center for Algorithmic Biotechnology. URL: https://cab.spbu.ru/software/spades (дата обращения: 28.10.2023).

[32] GitHub – dzerbino/velvet: Short read de novo assembler using de Bruijn graphs. URL: https://github.com/dzerbino/velvet (дата обращения: 28.10.2023).

[33] GitHub – theInnuendoProject/INNUca: INNUENDO quality control of reads, de novo assembly and contigs quality assessment, and possible contamination search. URL: https://github.com/innuendocon/innuca (дата обращения: 28.10.2023).

[34] GitHub – marbl/canu: A single molecule sequence assembler for genomes large and small. URL: https://github.com/marbl/canu (дата обращения: 28.10.2023).

[35] GitHub – voutcn/megahit: Ultra-fast and memory-efficient (meta-)genome assembler. URL: https://github.com/voutcn/megahit (дата обращения: 28.10.2023).

[36] IDBA-Bioinfomatics Research Group of Hong Kong University. URL: https://i.cs.hku.hk/~alse/hkubrg/projects/idba_ud (дата обращения: 28.10.2023).

[37] GitHub – ablab/quast: Genome assembly evaluation tool. URL: https://github.com/ablab/quast (дата обращения: 28.10.2023).

тивность зачастую невысока. Более эффективно, хотя и ресурсозатратно, проведение повторного секвенирования, в частности, с использованием праймеров, непосредственно фланкирующих промежутки[38].

На основании полученного генома возможно установление таксономической принадлежности изучаемого патогена с высокой достоверностью. Для этих целей применяют так называемые таксономические классификаторы, например, kraken2[39], MIDAS[40], CLARK[41], Centrifuge[42] и некоторые другие. В основе алгоритма работы таких программ лежит сравнение каждой нуклеотидной последовательности анализируемого образца с выбранными локальными и/или глобальными базами референсных геномов (доступны базы геномов архей, прокариот, простейших, грибов, растений, человека, а также специализированные базы – плазмиды, последовательности линкеров, адаптеров, праймеров, 16S РНК и т.д.); для каждой последовательности создается описание, в котором, в случае обнаружения соответствия с референсом, указывается совпавший таксон [29]. Полученные результаты могут быть визуализированы с использованием специализированных утилит типа Pavian (рисунок 8)[43].

К числу ключевых этапов биоинформатического анализа генетической информации относится также аннотирование полученной сборки генома, в том числе поиск генетических маркеров, имеющих важное клиническое (установление диагноза, назначение этиотропной терапии), эпидемиологическое (установление первичного источника в очаге, заключение об эпидемической значимости штамма, прогнозирование последствий распространения заболевания и т.д.) или научное (филогенетический анализ, изучение

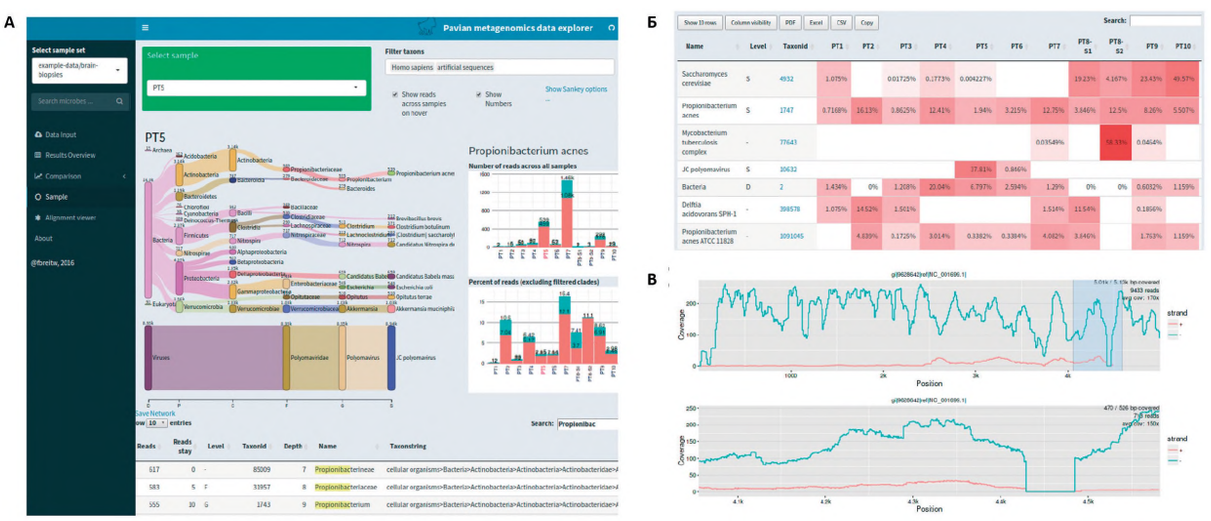

Рисунок 8 – *Визуализация результатов работы таксономического классификатора утилитой Pavian. А – Распределение по таксонам (диаграмма Сэнкей); Б – распределение по таксонам («тепловая карта»); В – покрытие референсного генома»*
(GitHub – fbreitwieser/pavian: Interactive analysis of metagenomics data. URL: https://github.com/fbreitwieser/pavian; дата обращения: 28.10.2023)

[38] Для решения абсолютного большинства практических задач не требуется получения единой геномной последовательности, практически все утилиты для аннотирования, поиска детерминант патогенности и устойчивости к антибиотикам, для мультилокусного сиквенс-типирования *in silico* и т.п. способны эффективно работать с набором контигов, отдельные – в том числе и с исходными прочтениями.

[39] GitHub – DerrickWood/kraken2: The second version of the Kraken taxonomic sequence classification system. URL: https://github.com/derrickwood/kraken2 (дата обращения: 28.10.2023).

[40] GitHub – snayfach/MIDAS: An integrated pipeline for estimating strain-level genomic variation from metagenomic data. URL: https://github.com/snayfach/midas (дата обращения: 28.10.2023).

[41] CLARK: Overview. URL: http://clark.cs.ucr.edu (дата обращения: 28.10.2023).

[42] GitHub – DaehwanKimLab/centrifuge: Classifier for metagenomic sequences. URL: https://github.com/daehwankimlab/centrifuge (дата обращения: 28.10.2023).

[43] Применение классификаторов, помимо прямого назначения, позволяет также оценить наличие в полученной геномной сборке генетической контаминацией ДНК человека и/или посторонней микрофлоры.

эволюционных механизмов, коллекционная деятельность) значение.

К настоящему времени разработаны сотни разнообразных утилит для проведения такого анализа, сведения об основных из которых приведены в таблице 1.

Одним из заключительных опциональных этапов изучения генома патогена является филогенетический анализ – установление связей и степени родства различных организмов, имеющих общего предка, на основании сравнительного анализа их нуклеиновых кислот.

Среди многочисленных программных решений для построения так называемого филогенетического дерева наибольшей популярностью пользуются такие инструменты, как RAxML[44], Gubbins[45], ClonalFrameML[46] и PHYLOViZ 2.0[47]. Результаты их работы могут быть визуализированы в удобном для последующего анализа виде (рисунок 9).

Как видно из представленных данных, подавляющее большинство программных решений для биоинформатического анализа генетических данных представляет собой не имеющие графического интерфейса утилиты, изначально написанные авторами для использования в среде Linux-подобных операционных систем. Такой подход во многом оправдан. Во-первых, необходимо учитывать, что биоинформатика – это своеобразный бэкенд современной науки, классический

Таблица 1 – *Утилиты для аннотирования и анализа геномов (составлена авторами)*

Функциональная группа	Наименование утилиты	URL адрес
Пангеномные и генные аннотаторы	Roary	https://github.com/sanger-pathogens/Roary
	Prokka	http://github.com/tseemann/prokka
	RAST	http://rast.nmpdr.org/
	PGAP	https://www.ncbi.nlm.nih.gov/genome
	Artemis	http://sanger-pathogens.github.io/Artemis/
	MicrobeAnnotator	https://github.com/cruizperez/MicrobeAnnotator
	Glimmer	http://ccb.jhu.edu/software/glimmer
	Prodigal	http://github.com/hyattpd/prodigal
Поиск детерминант устойчивости к антибактериальным препаратам	RGI-CARD	http://www.card.mcmaster.ca/analyze/rgi
	ResFinder	https://cge.cbs.dtu.dk/services/ResFinder/
	ariba	https://github.com/sanger-pathogens/ariba
	ABRicate	https://github.com/tseemann/abricate
	AMRplusplus	https://github.com/meglab-metagenomics/amrplusplus_v2
	DeepARG	https://github.com/gaarangoa/deeparg2.0
	fARGene	https://github.com/fannyhb/fargene
	StarAMR	https://github.com/phac-nml/staramr
Поиск и анализ однонуклеотидных вариантов	SNVPhyl	http://snvphyl.readthedocs.io/en/latest
	metaSNV	https://github.com/metasnv-tool/metaSNV
	biohansel	https://github.com/phac-nml/biohansel
	SNP-sites	https://github.com/sanger-pathogens/snp-sites
	Snippy	http://github.com/tseemann/snippy
Мультилокусное сиквенс-типирование	MLST_check	https://github.com/sanger-pathogens/mlst_check
	MLST	https://github.com/tseemann/mlst
	meta MLST	https://github.com/SegataLab/metamlst
	SRST2	http://github.com/katholt/srst2
Поиск открытых рамок считывания	ORFipy	https://github.com/urmi-21/orfipy
	ORF_finder	https://github.com/averissimo/orf_finder
	ORFFinder	https://github.com/Chokyotager/ORFFinder

[44] The Exelixis Lab. URL: https://cme.h-its.org/exelixis/web/software/raxml/index.html (дата обращения: 28.10.2023).

[45] GitHub – nickjcroucher/gubbins: Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. URL: https://github.com/nickjcroucher/gubbins (дата обращения: 28.10.2023).

[46] GitHub – xavierdidelot/ClonalFrameML: ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes. URL: https://github.com/xavierdidelot/ClonalFrameML (дата обращения: 28.10.2023).

[47] PHYLOViZ. URL: https://www.phyloviz.net (дата обращения: 28.10.2023).

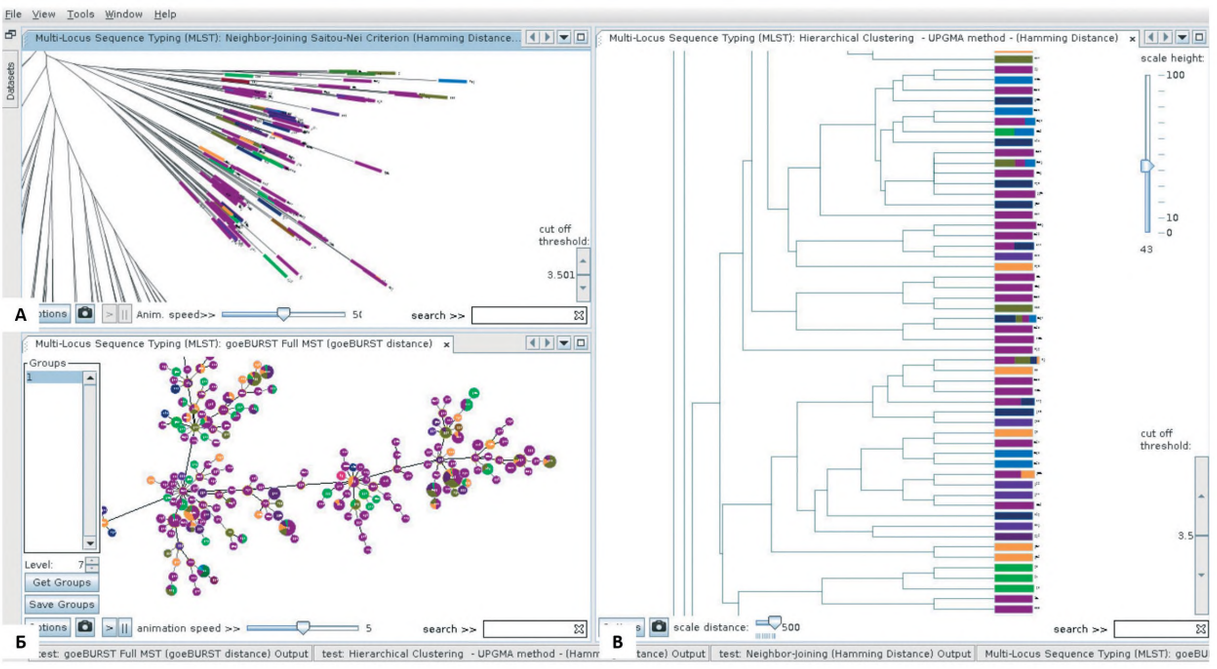

Рисунок 9 – *Визуализация результатов филогенетического анализа программой PHYLOViZ 2.0 с применением различных алгоритмов: Саитоу-Нэи (А), goeBURST (Б) и UPGMA (В) [30]*

«черный ящик», ведь конечного пользователя интересует именно результат, а не процесс. Кроме того, отказ от визуальных эффектов позволяет значительно экономить вычислительные ресурсы и сократить время анализа информации. Во-вторых, такой «модульный» характер программного обеспечения позволяет с учетом всех особенностей решаемой задачи, сформировать для нее так называемый пайплайн (от англ. pipeline – трубопровод, туннель) – последовательность операций обработки информации, обеспечивающих наибольшую эффективность анализа. В-третьих, это способствует объединению научного биоинформационного сообщества, обеспечивает возможность совместной работы над программными инструментами в целях их оптимизации и устранения ошибок, сохраняя при этом определенный – и достаточно высокий – порог вхождения.

Безусловно, существуют биоинформатические программы с привычным «оконным» интерфейсом, предназначенные для использования в среде Windows, однако большая их часть является коммерческими продуктами со всеми следующими из этого ограничениями либо общедоступными, но с крайне ограниченным функционалом. Одним из немногочисленных исключений является Ugene – бесплатная кросс-платформенная программа с открытым исходным кодом, разрабатываемая специалистами новосибирской компании «Юнипро»[48].

Благодаря интеграции десятков биоинформационных утилит, Ugene предоставляет удобный интуитивно понятный доступ к решению самых различных задач, включая обработку первичных данных секвенирования, полученную с использованием основных существующих на рынке платформ, контроль качества и тримминг, сборку по референсным геномам и *de novo*, выравнивание последовательностей, аннотирование, картирование, визуализация, ПЦР и клонирование *in silico*, построение и редактирование филогенетических деревьев и т.д. Дополнительно расширяет возможности программы встроенный конструктор задач, позволяющий создавать собственные несложные пайплайны (рисунок 10) [31].

Несмотря на то, что возможности программы ограничены числом адаптированных на текущий момент модулей расширения, а также их несколько меньшую скорость работы по сравнению с консольными вариантами соответствующих утилит, Ugene представляет несомненный интерес в силу своей открытости, совместимости и наличия графической оболочки, облегчающей освоение программы.

[48] Unipro UGENE – Integrated Bioinformatics Tools. URL: http://ugene.net (дата обращения: 28.10.2023).

Рисунок 10 – *Графический интерфейс программы Unipro UGENE (данные авторов). А – выравнивание прочтений на референсный геном; Б – работа с кольцевой формой генома; В – поиск открытых рамок считывания; Г – режим конструирования задач*

Заключение

Проведенный анализ показал, что современному исследователю доступен обширный арсенал средств и методов изучения генетических особенностей патогенных микроорганизмов, что является важным элементом при осуществлении эпидемиологических расследований вспышек инфекционных заболеваний, в том числе верификации результатов установления их возможного искусственного характера, а также для проведения любых других исследований особенностей организации и функционирования бактериальных и вирусных геномов.

Особое место среди молекулярно-генетических методов занимает секвенирование нуклеиновых кислот. Ключевым фактором, во многом определяющим эффективность его использования на практике, является знание и грамотное применение соответствующих биоинформационных утилит.

Среди многочисленных программных продуктов, предназначенных для оценки качества секвенирования, предварительной обработки данных, их картирования на референсный геном, сборки генома *de novo*, его аннотирования, типирования и выявления значимых генетических детерминант (устойчивости к антибактериальным препаратам, факторов патогенности и т.д.), проведения филогенетического анализа и некоторых других научных и прикладных задач, с учетом специфики деятельности подразделений войск РХБ защиты ВС РФ наибольший интерес представляют утилиты с открытым исходным кодом, не требующие для своей работы доступа к удаленным ресурсам.

Список источников/References

1. Morens DM, Fauci AS. Emerging pandemic diseases: how we got to COVID-19. *Cell*. 2020;182(5):1077–92. https://doi.org/10.1016/j.cell.2020.08.021

2. Smit M, Marinosci A, Agoritsas T, Calmy A. Prophylaxis for COVID-19: a systematic review. *Clin Microbiol Infect*. 2021;27(4):532–7. https://doi.org/10.1016/j.cmi.2021.01.013

3. Graña C, Ghosn L, Evrenoglou T, Jarde A, Minozzi S, Bergman H, et al. Efficacy and safety of COVID-19 vaccines. *Cochrane Database Syst Rev*. 2022;12(12):CD015477.
https://doi.org/10.1002/14651858.CD015477

4. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, et al. Nucleotide sequence of bacteriophage φX174 DNA. *Nature*. 1977;265(5596):687–95.
https://doi.org/10.1038/265687a0

5. Watts D, MacBeath JRE. Automated fluorescent DNA sequencing on the ABI PRISM 310 Genetic Analyzer. In: *DNA Sequencing Protocols. Methods in Molecular Biology, vol 167*. Graham CA, Hill AJM, Eds. Humana Press; 2001.
https://doi.org/10.1385/1-59259-113-2:153

6. Shendure J, Ji H. Next-generation DNA sequencing. *Nat Biotechnol*. 2008;26:1135-45.
https://doi.org/10.1038/nbt1486

7. Hernandez D, François P, Farinelli L, Osterås M, Schrenzel J. De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome Res*. 2008;18(5):802-9.
https://doi.org/10.1101/gr.072033.107

8. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, et al. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*. 2012;13:341.
https://doi.org/10.1186/1471-2164-13-341

9. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-time DNA sequencing from single polymerase molecules. *Science*. 2009;323(5910):133–38.
https://doi.org/10.1126/science.1162986

10. Arumugam K, Bessarab I, Liu X, Natarajan G, Drautz-Moses DI, Wuertz S, et al. Improving recovery of member genomes from enrichment reactor microbial communities using MinION–based long read metagenomics. *bioRxiv*. 2018:465328.
https://doi.org/10.1101/465328

11. Maljkovic Berry I, Melendrez MC, Bishop-Lilly KA, Rutvisuttinunt W, Pollett S, Talundzic E, et al. Next generation sequencing and bioinformatics methodologies for infectious disease research and public health: approaches, applications, and considerations for development of laboratory capacity. *J Infect Dis*. 2020;221(Suppl 3):S292–S307.
https://doi.org/10.1093/infdis/jiz286

12. Besser J, Carleton HA, Gerner-Smidt P, Lindsey RL, Trees E. Next-generation sequencing technologies and their application to the study and control of bacterial infections. *Clin Microbiol Infect*. 2018;24(4):335–41.
https://doi.org/10.1016/j.cmi.2017.10.013

13. Robinson JM, Pasternak Z, Mason CE, Elhaik E. Forensic applications of microbiomics: a review. *Front Microbiol*. 2021;11:608101.
https://doi.org/10.3389/fmicb.2020.608101

14. Allali I, Arnold JW, Roach J, Cadenas MB, Butz N, Hassan HM, et al. A comparison of sequencing platforms and bioinformatics pipelines for compositional analysis of the gut microbiome. *BMC Microbiol*. 2017;17(1):194.
https://doi.org/10.1186/s12866-017-1101-8

15. Chaudhari HG, Prajapati S, Wardah ZH, Raol G, Prajapati V, Patel R, et al. Decoding the microbial universe with metagenomics: a brief insight. *Front Genet*. 2023;14:1119740.
https://doi.org/10.3389/fgene.2023.1119740

16. Vincent AT, Derome N, Boyle B, Culley AI, Charette SJ. Next-generation sequencing (NGS) in the microbiological world: How to make the most of your money. *J Microbiol Methods*. 2017;138:60–71.
https://doi.org/10.1016/j.mimet.2016.02.016

17. Lema NK, Gemeda MT, Woldesemayat AA. Recent advances in metagenomic approaches, applications, and challenge. *Curr Microbiol*. 2023;80(11):347.
https://doi.org/10.1007/s00284-023-03451-5

18. Cornet L, Baurain D. Contamination detection in genomic data: more is not enough. *Genome Biol*. 2022;23:60.
https://doi.org/10.1186/s13059-022-02619-9

19. Bush SJ, Connor TR, Peto TEA, Crook DW, Walker AS. Evaluation of methods for detecting human reads in microbial sequencing datasets. *Microb Genom*. 2020;6(7):mgen000393.
https://doi.org/10.1099/mgen.0.000393

20. Salzberg SL, Breitwieser FP, Kumar A, Hao H, Burger P, Rodriguez FJ, et al. Next-generation sequencing in neuropathologic diagnosis of infections of the nervous system. *Neurol Neuroimmunol Neuroinflamm*. 2016;3(4):e251.
https://doi.org/10.1212/NXI.0000000000000251

21. Brennan C, Salido RA, Belda-Ferre P, Bryant M, Cowart C, Tiu MD, et al. Maximizing the potential of high-throughput next-generation sequencing through precise normalization based on read count distribution. *mSystems*. 2023;8(4):e0000623.
https://doi.org/10.1128/msystems.00006-23

22. Portik DM, Brown CT, Pierce-Ward NT. Evaluation of taxonomic classification and profiling methods for long-read shotgun metagenomic sequencing datasets. *BMC Bioinformatics*. 2022;23(1):541. https://doi.org/10.1186/s12859-022-05103-0

23. Reinert K, Langmead B, Weese D, Evers DJ. Alignment of next-generation sequencing reads. *Annu Rev Genomics Hum Genet*. 2015;16:133-51.
https://doi.org/10.1146/annurev-genom-090413-025358

24. Liu Y, Shen X, Gong Y, Liu Y, Song B, Zeng X. Sequence Alignment/Map format: a comprehensive review of approaches and applications. *Brief Bioinform*. 2023;24(5):bbad320.
https://doi.org/10.1093/bib/bbad320

25. Antipov D, Raiko M, Lapidus A, Pevzner PA. Plasmid detection and assembly in genomic and metagenomic data sets. *Genome Res*. 2019;29(6):961-8.
https://doi.org/10.1101/gr.241299.118

26. Gupta SK, Raza S, Unno T. Comparison of de-novo assembly tools for plasmid metagenome analysis. *Genes Genomics*. 2019;41(9):1077–83.
https://doi.org/10.1007/s13258-019-00839-1

27. Gurevich A, Saveliev V, Vyahhi N, Tesler G, QUAST: quality assessment tool for genome assemblies. *Bioinformatics*. 2013;8(29):1072–5.
https://doi.org/10.1093/bioinformatics/btt086

28. Huang B, Wei G, Wang B, Ju F, Zhong Y, Shi Z, et al. Filling gaps of genome scaffolds via probabilistic searching optical maps against assembly graph. *BMC Bioinformatics*. 2021;22(1):533.
https://doi.org/10.1186/s12859-021-04448-2

29. Lu J, Rincon N, Wood DE, Breitwieser FP, Pockrandt C, Langmead B, et al. Metagenome analysis using the Kraken software suite. *Nat Protoc*. 2022;17(12):2815–39.
https://doi.org/10.1038/s41596-022-00738-y

30. Nascimento M, Sousa A, Ramirez M, Francisco AP, Carriço JA, Vaz C. PHYLOViZ 2.0: providing scalable data integration and visualization for multiple phylogenetic inference methods. *Bioinformatics*. 2017;33(1):128–9.
https://doi.org/10.1093/bioinformatics/btw582

31. Rose R, Golosova O, Sukhomlinov D, Tiunov A, Prosperi M. Flexible design of multiple metagenomics classification pipelines with UGENE. *Bioinformatics*. 2018;11(35):1963–5.
https://doi.org/10.1093/bioinformatics/bty901

Вклад авторов / Authors' contributions

Все авторы подтверждают соответствие своего авторства критериям ICMJE. Наибольший вклад распределен следующим образом: **Я.А. Кибирев** – сбор и анализ данных научной литературы, написание текста рукописи; **А.В. Кузнецовский** – формирование концепции статьи, критический пересмотр и коррекция текста рукописи, окончательное утверждение рукописи для публикации; **С.Г. Исупов** – переработка текста рукописи; **И.В. Дармов** – анализ данных научной литературы и коррекция текста рукописи / All authors confirm that they meet the International Committee of Medical Journal Editors (ICMJE) criteria for authorship. The most significant contribution were as follows: **Ya.A. Kibirev** – collection and analysis of scientific literature data, drafting the manuscript; **A.V. Kuznetsovskiy** – formation of the concept of the article, critical revision and correction of the text of the manuscript, final approval of the manuscript for publication; **S.G. Isupov** – revision the manuscript; **I.V. Darmov** – analysis of scientific literature data and correction of the manuscript.

Информация о конфликте интересов / Conflict of interest statement

Авторы заявляют, что исследования проводились при отсутствии любых коммерческих или финансовых отношений, которые могли бы быть истолкованы как потенциальный конфликт интересов / The authors declare that the research was conducted in the absence of any commercial or financial relationship that could be construed as a potential conflict of interest.

Сведения о рецензировании / Peer review information

Статья прошла двустороннее анонимное «слепое» рецензирование двумя рецензентами, специалистами в данной области. Рецензии находятся в редакции журнала и в РИНЦе / The article has been double-blind peer reviewed by two experts in the respective field. Peer reviews are available from the Editorial Board and from Russian Science Citation Index database.

Финансирование / Funding

Филиал федерального государственного бюджетного учреждения «48 Центральный научно-исследовательский институт» (г. Киров) Министерства обороны Российской Федерации / Branch Office of the Federal State Budgetary Establishment «48 Central Scientific Research Institute» of the Ministry of Defence of the Russian Federation (Kirov).

Об авторах / Authors

Филиал федерального государственного бюджетного учреждения «48 Центральный научно-исследовательский институт» Министерства обороны Российской Федерации, 610000, Российская Федерация, г. Киров, Октябрьский проспект, д. 119.

Кибирев Ярослав Александрович. Начальник отдела, канд. биол. наук.

Кузнецовский Андрей Владимирович. Начальник отдела планирования НИР – заместитель начальника филиала по НИР, канд. биол. наук.

Исупов Сергей Геннадьевич. Заместитель начальника отдела, канд. мед. наук.

Дармов Илья Владимирович. Главный научный сотрудник управления, доктор мед. наук, профессор.

Контактная информация для всех авторов: 23527@mil.ru

Контактное лицо: Кибирев Ярослав Александрович; 23527@mil.ru

Branch Office of the Federal State Budgetary Establishment «48 Central Scientific Research Institute» of the Ministry of Defence of the Russian Federation, Oktyabrsky Avenue 119, Kirov 610000, Russian Federation.

Yaroslav A. Kibirev. Chief of the Department. Cand. Sci. (Biol.).

Andrey V. Kuznetsovskiy. Deputy Chief of the Branch Office. Cand. Sci. (Biol.).

Sergey G. Isupov. Deputy Chief of the Department. Cand. Sci. (Med.).

Ilya V. Darmov. Leading Researcher. Dr. Sci. (Med.), Professor.

Contact information for all authors: 23527@mil.ru

Contact person: Yaroslav A. Kibirev; 23527@mil.ru