РАЗРАБОТКА ВЕБ-СЕРВИСА ДЛЯ СБОРКИ И АНАЛИЗА ГЕНОМОВ МИКРООРГАНИЗМОВ

И. П. Баженов

Белорусский государственный университет, г. Минск;
ivan.bazhenov333@gmail.com;
науч. рук. – Р. С. Сергеев, науч. сотр. ГНУ «ОИПИ НАН Беларуси»

Приложения на основе веб-сервисов являются удобным инструментом для представления результатов практически значимых исследований широкому кругу пользователей. В настоящей работе рассматривается реализация веб-приложения, позволяющего выполнять биоинформатический анализ геномных данных возбудителя туберкулеза легких с целью установления типа лекарственной устойчивости.

Ключевые слова: веб-технологии; биоинформатика; анализ геномных данных.

ВВЕДЕНИЕ

Биоинформатика представляет собой относительно молодую междисциплинарную область исследований и активно развивающуюся индустрию. В частности, в связи со стремительным падением стоимости чтения нуклеиновых кислот следует ожидать более глубокого проникновения технологий секвенирования в повседневную диагностическую практику. Однако для продвижения результатов, достигнутых в этой и смежных областях, пользователи должны иметь удобную возможность выполнять анализ геномных данных без необходимости обладать навыками программирования. Приложения, созданные на основе современных веб-технологий, хорошо зарекомендовали себя для решения похожих задач благодаря простоте использования и минимальному набору требований, позволяя организовать взаимодействие с пользователем лишь средствами веб-браузера.

Целью настоящей работы являлась реализация веб-сервиса для биоинформатического анализа полногеномных данных возбудителя туберкулеза легких, установления типа лекарственной устойчивости и филогенетической принадлежности.

ДАННЫЕ К ЗАДАЧЕ

Для тестирования приложения использовались наборы полных геномов клинических изолятов Mycobacterium tuberculosis, которые были выделены у пациентов из Беларуси [1] и по итогам секвенирования помещены в международную базу данных NCBI (https://www.ncbi.nlm.nih.gov/) в виде проекта с идентификатором

PRJNA200335. В качестве референсного был выбран геном штамма H37Rv, имеющий идентификатор NC_000962.3. Анализ мутаций выполнялся с использованием информации, собранной по результатам обзора литературы о проведенных в этой области научных исследованиях.

ОБРАБОТКА ГЕНОМНЫХ ДАННЫХ

Процедура анализа геномных данных была реализована в виде пайплайна – цепочки операций, производимых с данными, в которой входные данные каждой последующей операции являются результатом выполнения предыдущих шагов. Для удобства работы было сконфигурировано две версии пайплайна: полная и сокращенная.

В первом случае от пользователя требуется ввести SRA-идентификатор архива (например, SRR1180924), который однозначно определяет геном в базе данных NCBI. Для сборки геномов, выполнения запроса вариантов и преобразования данных между различными форматами использовалось стороннее свободно распространяемое программное обеспечение:

- SRA Toolkit
- BWA
- SAMtools
- Pilon
- VCFtools

В случае сокращенной цепочки операций пользователю предлагается пропустить шаги картирования геномных прочтений и запроса вариантов (variant calling), загрузив уже готовый файл в формате VCF (variant call format) со списком мутаций, идентифицированных относительно заложенного в систему референсного генома.

Выгрузка данных. Каждый набор геномных прочтений, помещенный в международную базу NCBI, хранится в формате SRA (sequence read archive) и имеет уникальный буквенно-цифровой идентификатор. Программный пакет SRA Toolkit (https://github.com/ncbi/sra-tools) предоставляет возможность программного доступа к информации в базе NCBI и позволяет конвертировать выгруженные данные в другие форматы.

Картирование коротких прочтений. Программный пакет BWA [2] используется для установления соответствий между позициями коротких прочтений относительно референсного генома с учетом того, что данные могут содержать ошибки. Пакет BWA включает три модуля, реализующие заложенные в нем алгоритмы: BWA-BackTrack, BWA-SW и BWA-MEM. В основе алгоритмов пакета BWA лежат преобразование

Барроуза-Уилера, алгоритмы работы с суффиксными массивами и алгоритм выравнивания Смита—Ватермана.

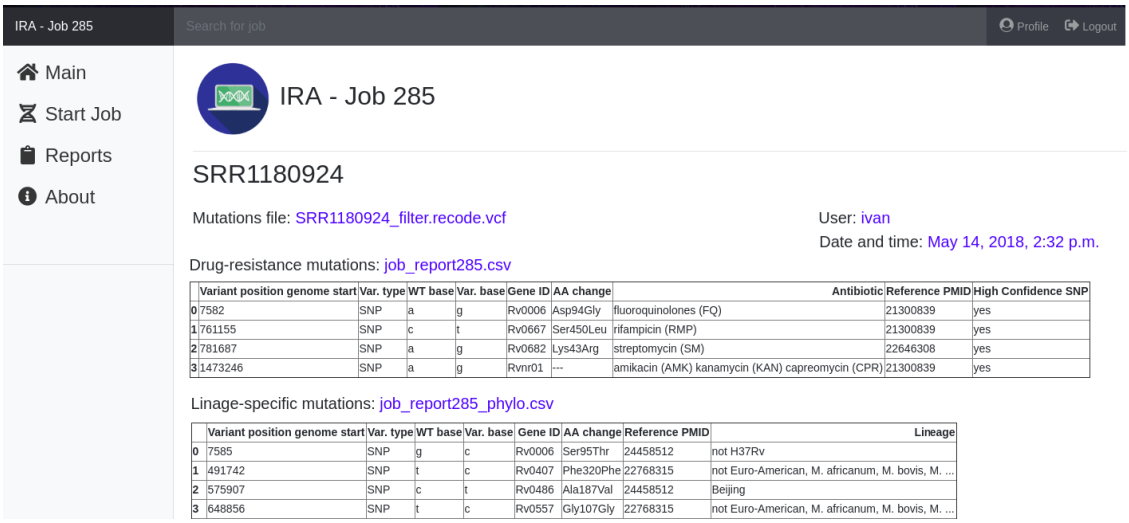
Запрос вариантов. По результатам картирования коротких прочтений на референсный геном можно восстановить последовательность ДНК исходного образца. Однако при отображении всех коротких прочтений на референсную последовательность возможны ситуации, когда вследствие ошибок секвенирования выравненные друг относительно друга прочтения могут иметь разные нуклеотиды в одной и той же позиции, либо значения нуклеотидов в какой-либо позиции могут оказаться неизвестны. В базовом случае можно считать, что значение нуклеотида в такой позиции совпадает с наиболее часто встречающимся значением среди прочтений, которые покрывают данную позицию. Для выполнения запроса вариантов с учетом возможных ошибок секвенирования используется программа Pilon [3]. Основываясь на выравнивании коротких прочтений, это позволяет точно и полно восстановить исходный бактериальный геном, а также получить файл в формате VCF, в котором перечислены отличия исходного генома от референсного.

Диагностика лекарственной устойчивости. Используя полученный VCF-файл можно выполнить поиск мутаций, связанных с лекарственной устойчивостью к некоторым известным противомикробным препаратам, а также идентифицировать мутации, определяющие филогенетическую линию исследуемого штамма. Наборы мутаций, ассоциированных с лекарственной устойчивостью, были подготовлены на основе анализа литературы и систем молекулярно-генетической диагностики, таких как GenoType MTBDRsl/MTBDRplus (HAIN-тест) [4] и GeneXpert MTB/RIF [5].

РЕАЛИЗАЦИЯ

Веб-приложение было реализовано на языке Python с использованием фреймворка Django. Разработка отдельных элементов системы была выполнена в виде скриптов на языке Bash для функционирования под управлением операционной системы Linux. Помимо непосредственной обработки геномных данных в приложении была реализована возможность ведения очереди задач и учетных записей пользователей.

Для динамического изменения отдельных элементов пользовательского интерфейса была применена технология AJAX и использована библиотека jQuery языка JavaScript. Элементы интерфейса веб-приложения, включая типовой отчет с результатами анализа пользовательских данных, представлены на Рисунке 1.



	Variant position genome start	Var. type	WT base	Var. base	Gene ID	AA change	Antibiotic	Reference PMID	High Confidence SNP
0	7582	SNP	a	g	Rv0006	Asp94Gly	fluoroquinolones (FQ)	21300839	yes
1	761155	SNP	c	t	Rv0667	Ser450Leu	rifampicin (RMP)	21300839	yes
2	781687	SNP	a	g	Rv0682	Lys43Arg	streptomycin (SM)	22646308	yes
3	1473246	SNP	a	g	Rvnr01	---	amikacin (AMK) kanamycin (KAN) capreomycin (CPR)	21300839	yes

	Variant position genome start	Var. type	WT base	Var. base	Gene ID	AA change	Reference PMID	Lineage
0	7585	SNP	g	c	Rv0006	Ser95Thr	24458512	not H37Rv
1	491742	SNP	t	c	Rv0407	Phe320Phe	22768315	not Euro-American, M. africanum, M. bovis, M. ...
2	575907	SNP	c	t	Rv0486	Ala187Val	24458512	Beijing
3	648856	SNP	t	c	Rv0557	Gly107Gly	22768315	not Euro-American, M. africanum, M. bovis, M. ...

Рис. 1. Пример пользовательского отчета с результатами анализа геномных данных.

ЗАКЛЮЧЕНИЕ

В результате выполненной работы было разработано веб-приложение, которое реализует операции, необходимые для сборки геномов возбудителя туберкулеза из наборов коротких прочтений и их последующего анализа с целью выявление геномных маркеров лекарственной устойчивости. Учитывая расширяемость приложения, подобный подход может использоваться и для анализа других микроорганизмов. Программное обеспечение, создаваемое на основе веб-технологий, является удобным и мощным инструментом для реализации биоинформатических алгоритмов и предоставления полученных с их помощью результатов анализа данных широкому кругу пользователей.

Библиографические ссылки

1. *Кириченко В. В. и др.* // Медицинская панорама. 2015. № 9. С. 75–78.
2. *Li H. and Durbin R.* // Bioinformatics. 2009. Vol. 25. P. 1754–1760.
3. *Walker B. J. et al.* // PLoS ONE. 2014. Vol. 9(11). e112963.
4. GenoType MTBDRplus [Электронный ресурс] // Hain Lifesciences: [сайт]. [2018]. URL: https://www.hain-lifescience.de/en/products/microbiology/mycobacteria.html
5. *Boehme C. C. et al.* // N. Engl. J. Med. 2010. Vol. 363. P. 1005–1015.