РАЗРАБОТКА ПРОГРАММНОГО ОКРУЖЕНИЯ НА ОСНОВЕ DOCKER ДЛЯ ОБРАБОТКИ ДАННЫХ СЕКВЕНИРОВАНИЯ

Н.В. Воронова-Барте, В.Б. Степаненко, Е.Р. Яхницкая, С.С. Левыкина

УО «Белорусский государственный университет», Минск, Беларусь
E-mail: vikastepanenko97@gmail.com

Разработан новый подход к анализу геномных данных через контейнерную технологию. Созданное программное окружение обеспечивает гибкую среду для обработки данных секвенирования в Docker-контейнерах, упрощая и автоматизируя процесс анализа для исследователей геномики и биоинформатики.

Ключевые слова: биоинформатика; контейнеризация; геномика.

ВВЕДЕНИЕ

С развитием высокопроизводительных методов секвенирования объем данных, генерируемых при исследованиях, значительно вырос, что повлекло за собой необходимость в разработке эффективных методов обработки и анализа геномных последовательностей. Docker представляет собой мощный инструмент для создания гибких и масштабируемых сред для анализа геномных данных. В данной работе представлен новый подход к анализу геномных данных с использованием конвейера, который включает в себя широкий спектр биоинформатических инструментов для генерации, обработки и анализа геномных последовательностей.

МАТЕРИАЛЫ И МЕТОДЫ

В работе использовались программные инструменты SPAdes, Bowtie, BWA, Samtools и Bcftools [1-5], которые были объединены в конвейер для дальнейшего создания Docker-контейнера. Также использовался язык bash для написания скриптов, решающих локальные задачи, такие как автоматизация процессов сборки Docker-образа и управление зависимостями.

Docker-контейнер представляет собой стандартизированный, изолированный и портативный пакет программного обеспечения, включающий в себя все необходимые для запуска приложений, включая код, среду выполнения, системные инструменты, библиотеки и настройки [6]. Созданный контейнер размещен в облачный репозитории - Docker Hub. Это позволяет любому пользователю получить доступ к контейнеру и использовать его на своей системе без необходимости внесения изменений в среду выполнения.

Инструменты были подобраны специально для изучения геномов эндосимбионтов, таких как выбранный тестовый объект *Buchnera aphidicola* (GCF_003099975.1).

РЕЗУЛЬТАТЫ

Схема работы конвейера изображена на рисунке.

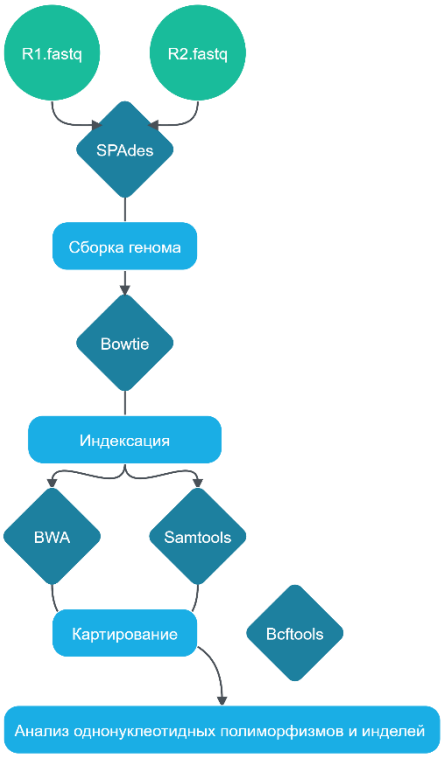

Схема работы созданного конвейера

Генерация исходного файла, содержащего собранные геномные последовательности, происходит в формате FASTA. Для этой цели используется программное обеспечение SPAdes, которое является известным в области сборки геномов и характеризуется высокой степенью эффективности [1]. Для сборки используются два файла в формате FASTQ, обозначаемые как R1 и R2. Эти файлы содержат парные прочтения, полученные из секвенирования фрагментов ДНК. Затем полученный новый файл FASTA, содержащий собранные последовательности, подвергается индексации с использованием инструмента Bowtie.

Процесс индексации позволяет эффективно представить геномные последовательности в виде структуры данных, которая оптимизирует последующее картирование геномных ридов. Такой подход способствует

ускорению и улучшению процесса анализа геномных данных, обеспечивая более быстрый доступ к информации и более точное выравнивание геномных последовательностей на референсный геном [2].

После процесса индексации следует непосредственно картирование данных на указанный референс с применением инструмента BWA (Burrows-Wheeler Aligner). Далее полученные выравнивания подвергаются обработке при помощи инструмента Samtools. Этот процесс включает в себя несколько этапов, включая сортировку выравниваний по их позиции в геноме, фильтрацию для исключения низкокачественных выравниваний или артефактов, а также индексацию выравниваний для обеспечения быстрого доступа к информации о выравниваниях в последующих анализах [7].

Обработка выравниваний с использованием инструмента Samtools создает основу для последующего анализа геномных вариантов. Точные и надежные выравнивания являются необходимым предварительным условием для корректного геномного анализа.

Последним этапом работы созданного конвейера является непосредственно проведение SNP-анализа (анализа однонуклеотидных полиморфизмов) и идентификация инделей (делеций и инсерций), что имеет важное значение для понимания генетической изменчивости и ее связи с фенотипическими характеристиками исследуемых организмов. Для этого используются инструменты Samtools и Bcftools, которые позволяют обнаруживать точные местоположения геномных вариаций, определяя отличия в последовательности нуклеотидов между исследуемыми образцами и референсным геномом [8].

Однонуклеотидные полиморфизмы и инделы могут быть ключевыми маркерами для идентификации генетических особенностей, связанных с адаптацией к окружающей среде, эволюционным процессам и различным заболеваниям.

Так, при анализе генома *Buchnera aphidicola* с использованием созданного контейнера было найдено 38 инделей, из которых: 1 индел длиной в 11 нуклеотидов (с делецией), 2 инделя длиной в 4 нуклеотида (с делецией), 3 инделя длиной в 3 нуклеотида (с делецией), 5 инделей длиной в 2 нуклеотида (с делецией), 14 инделей длиной в 1 нуклеотид (с делецией), 5 инделей длиной в 1 нуклеотид (с инсерцией), 4 инделя длиной в 2 нуклеотида (с инсерцией), 1 индел длиной в 4 нуклеотида (с инсерцией), 1 индел длиной в 6 нуклеотидов (с инсерцией), 1 индел длинойв 8 нуклеотидов (с инсерцией), 1 индел длиной в 12 нуклеотидов (с инсерцией).

Также в изучаемом геноме найдено 26 457 однонуклеотидных полиморфизмов. Обнаружено, что наиболее часто встречающиеся типы замен – A>T, T>A, A>G и T>C, в то время как наименее часто встречающиеся – C>G, G>C и G>T.

ЗАКЛЮЧЕНИЕ

Был разработан конвейер, включающий пять программных инструментов: SPAdes, Bowtie, BWA, Samtools и Bcftools. Для удобства его использования был создан Docker-контейнер, который позволяет значительного сократить время, затрачиваемое исследователями на поиск и установку соответствующих программ, выполнение отдельных геномных анализов, а также нивелирует проблему переноса настроенных приложений от одного пользовательня к другому, что делает его эффективным инструментом для исследований полных геномов эндосимбионтов.

В результате в геноме *Buchnera aphidicola* было найдено 26 457 однонуклеотидных полиморфизмов. Также обнаружено, что из 38 инделей25 были делециями длиной от 1 до 11 нуклеотидов, 13 – инсерциями от1 до 12 нуклеотидов.

БИБЛИОГРАФИЧЕСКИЕ ССЫЛКИ

1. Spades assembly - commandline - ABRPI-Training [Электронный ресурс]. URL: https://sepsis-omics.github.io/tutorials/modules/spades_cmdline/ (дата обращения: 23.03.2024).
2. Bowtie: Tutorial [Электронный ресурс]. URL: https://bowtie-bio.sourceforge.net/tutorial.shtml (дата обращения: 23.03.2024).
3. Учебное пособие по BWA-MEM [Электронный ресурс]. URL: https://docs.tinybio.cloud/docs/bwa-mem-tutorial (дата обращения: 23.03.2024).
4. Учебное пособие по Samtools. [Электронный ресурс]. URL: https://docs.tinybio.cloud/docs/samtools-overview (дата обращения: 23.03.2024).
5. BCFtools [Электронный ресурс]. URL: https://samtools.github.io/bcftools/bcftools.html (дата обращения: 23.03.2024).
6. Docker: Lightweight Linux Containers for Consistent Development and Deployment [Электронный ресурс]. URL: https://www.linuxjournal.com/content/docker-light-weight-linux-containers-consistent-development-and-deployment (дата обращения: 23.03.2024).
7. Картирование на референсный геном | Учебный сайт Александра Злобина. [Электронный ресурс]. URL: https://kodomo.fbb.msu.ru/~alexander.zlobin/terms/third/mapping/ (дата обращения: 23.03.2024).
8. Однонуклеотидные полиморфизмы, индели и сборка | Учебный сайт Александра Злобина. [Электронный ресурс]. URL: https://kodomo.fbb.msu.ru/~alexander.zlobin/terms/third/velvet/ (дата обращения: 23.03.2024).