НОВОЕ ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ ДЛЯ ОБРАБОТКИ ДАННЫХ ПОЛНЫХ ГЕНОМОВ МИКОБАКТЕРИЙ ТУБЕРКУЛЕЗА

М. В. СПРИНДЖУК, Л. П. ТИТОВ, О. М. ЗАЛУЦКАЯ, А. Е. СКРЯГИН, А. М. СКРЯГИНА, Р. С. СЕРГЕЕВ

Объединенный институт проблем информатики Академии наук Беларуси, г. Минск, Беларусь

В статье представлено описание разработанного программного обеспечения, предназначенного для обработки данных полных геномов микобактерий туберкулеза человека.

Ключевые слова: туберкулез, геномика, программное обеспечение

Для цитирования: Спринджук М. В., Титов Л. П., Залуцкая О. М., Скрягин А. Е., Скрягина А. М., Сергеев Р. С. Новое программное обеспечение для обработки данных полных геномов микобактерий туберкулеза // Туберкулёз и болезни лёгких. – 2017. – Т. 95, № 6. – С. 41-44.

DOI: 10.21292/2075-1230-2017-95-6-41-44

Туберкулез представляет собой серьезную проблему современной мировой эпидемиологии и экономики [12, 16]. С девяностых годов прошлого столетия для изучения геномов микобактерий туберкулеза (МБТ) используется полногеномное секвенирование [9-11, 13, 14].

Для обработки данных полных геномов в течение последних двух десятилетий разрабатывается соответствующее программное обеспечение. Наиболее развитые коммерческие программные комплексы – DNA Star Laser Gene [3, 5-7], Partek [8, 15], Next Gene, Genomics Workbench. Наиболее популярное открытое бесплатное программное обеспечение – R Bioconductor, BioPython, BioRuby, BioJava, Galaxy [17, 18].

Цель работы: разработка программного обеспечения для обработки данных полных геномов МБТ с удобным понятным пользователю интерфейсом.

Материалы и методы

Использовался язык программирования Python 2.7, для интерфейса была отобрана PyGTK. Утилиты Linux вызывались командами языка Shell/Bash. Для вызова скриптов Linux был применен модуль Executor.

Результаты исследования

Программный продукт был разработан для обработки и изучения данных полных геномов мультирезистентных МБТ, являющихся возбудителями туберкулеза в Белоруссии. Функциональная основа программы – бесплатные инструменты для биоинформатики:

1) FASTQ-dump из SRA-Toolkit (https://github.com/ncbi/sra-tools/wiki/HowTo:-Binary-Installation);
2) BWA для картирования/выравнивания полных геномов (https://sourceforge.net/projects/bio-bwa/files/);
3) SAMTools для конвертации файлов из BAM формата в BAM (http://samtools.sourceforge.net/);
4) VarScan для запроса вариантов с бинарного файла результата картирования (http://varscan.sourceforge.net).

Программа VarScan генерирует список одиночных полиморфизмов, составляющих мутационный профиль вводимого генома. Параметр — vcf позволяет получить на выходе файл формата запроса вариантов, который, правда, требует серьезных навыков практического программирования и биоинформатики, поддается аннотации для получения удобного отчета со списком имен генов, в которых был вычислен мутационный профиль. Для этого шага используются программы SNPEff [4], VCF Annotator, Mannovar, VCF Anno [19].

Алгоритм работы разработанного программного обеспечения и его интерфейс представлены на рис. 1 и 2.

Рис. 1. Блок-схема функций разработанного программного обеспечения
Fig. 1. Functional diagram of the developed software

Разработанный программный продукт имеет удобный понятный пользователю интерфейс и позволяет выполнять следующие задачи практической геномики:

1) загружать файлы FASTQ по идентификатору SRA;
2) выполнять индексирование референсного генома и выравнивание загруженных данных на этот геном;
3) конвертировать полученный на предыдущем этапе обработки данных результат в формат BAM, а также сортировать и индексировать этот файл;
4) генерировать список одиночных полиморфизмов (SNPs).

Заключение

Данное программное обеспечение уже используется для вычисления мутационного профиля образцов возбудителя туберкулеза в Белоруссии в РНПЦ эпидемиологии и микробиологии г. Минска в рамках текущих научных проектов. Система может быть модифицирована для использования при изучении геномов практически любых бактерий млекопитающих и человека.

Код программного обеспечения доступен бесплатно по ссылке https://github.com/MatveySprindzuk/GenomicsSoftware.

Для сотрудничества по вопросам биоинформатики, консультаций и запроса расширенной версии программного обеспечения, пожалуйста, обращайтесь к автору msprindzhuk@mail.ru, +375 29 567 10 73

Конфликт интересов. Авторы заявляют об отсутствии у них конфликта интересов.

Авторы выражают благодарность доцентам В. А. Горбунову и А. П. Кончицу за помощь в научной работе. Исследование выполнялось при поддержке грантов научных проектов CRDF, ОИПИ НАН Беларуси, РНПЦ микробиологии и эпидемиологии г. Минска.

Conflict of Interests. The authors state that they have no conflict of interests.

The authors express their deepest gratitude to V.A. Gorbunov and A.P. Konchitsa, Associate Professors, for their assistance in this research. The research was supported by grants from CRDF, United Institute of Informatics Problems of Belarus, the Republican Research and Practical Center for Epidemiology and Microbiology of Minsk.
REFERENCES