УДК 579.873.21:579.25].083 **DOI 10.21292/2075-1230-2017-95-6-41-44**

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 микобактерии 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

NEW SOFTWARE FOR PROCESSING DATA OF ENTIRE GENOME OF MYCOBACTERIUM TUBERCULOSIS

M. V. SPRINDZHUK, L. P. TITOV, O. M. ZALUTSKAYA, A. E. SKRYAGIN, A. M. SKRYAGINA, R. S. SERGEEV

United Institute of Informatics Problems, Minsk, Belarus

The article describes the software developed for processing of data of entire genomes of human tuberculous mycobacteria.

Key words: tuberculosis, genomics, software

For citations: Sprindzhuk M.V., Titov L.P., Zalutskaya O.M., Skryagin A.E., Skryagina A.M., Sergeev R.S. New software for processing data of entire genome of *Mycobacterium tuberculosis*. *Tuberculosis and Lung Diseases,* 2017, Vol. 95, no. 6, P. 41-44. (In Russ.) 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].

За последнее десятилетие появились сотни бесплатных программ-инструментов для обработки данных, обладающих разной вычислительной эффективностью и имеющих различное качество кода и документации. Однако специализированного программного обеспечения с хорошим понятным интерфейсом для обработки данных геномов МБТ немного [1, 2, 20], в том числе рассчитанного на врачей, микробиологов, биологов, профессионалов биомедицинских наук, имеющих знания пользователя персонального компьютера, но не имеющих навыков профессионального программирования, работы с командной строкой и настройкой непопулярных операционных систем, которыми являются на сегодняшний день Linux и его варианты.

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. (http://www.pygtk.org/). Утилиты Linux вызывались командами языка Shell/Bash. Для вызова скриптов Linux был применен модуль Executor (https://pypi.python.org/pypi/executor).

Результаты исследования

Программный продукт был разработан для обработки и изучения данных полных геномов мультирезистентных МБТ, являющихся возбудителем 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 SAM 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], VCFAnnotator, Mannovar, VCFAnno [19].

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

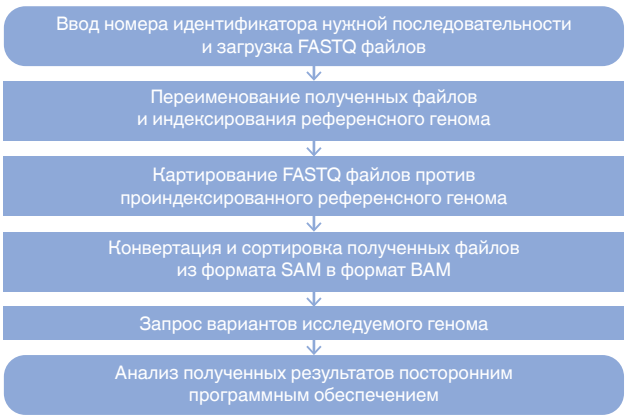

Рис. 1. *Блок-схема функций разработанного программного обеспечения*

Fig. 1. *Functional diagram of the developed software*

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

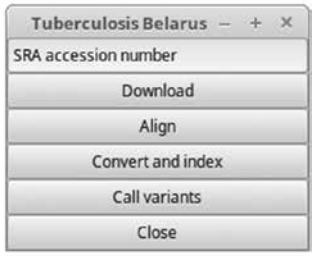

Рис. 2. *Интерфейс программного обеспечения*

Fig. 2. *Software interface*

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 эпидемологии 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.

ЛИТЕРАТУРА

1. Bradley P., Gordon N. C., Walker T. M., Dunn L., Heys S. et al. Corrigendum: Rapid antibiotic-resistance predictions from genome sequence data for *Staphylococcus aureus* and *Mycobacterium tuberculosis* // Nat. Commun. – 2016. – Vol. 7. – P. 11465.
2. Bradley P., Gordon N. C., Walker T. M., Dunn L., Heys S. et al. Rapid antibiotic-resistance predictions from genome sequence data for *Staphylococcus aureus* and *Mycobacterium tuberculosis* // Nat. Commun. – 2015. – Vol. 6. – P. 10063.
3. Burland T. G. DNASTAR's Lasergene sequence analysis software // Methods Mol. Biol. – 2000. – Vol. 132. – P. 71-91.
4. Cingolani P., Platts A., Wang le L., Coon M., Nguyen T. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3 // Fly (Austin). – 2012. – Vol. 6, № 2. – P. 80-92.
5. Clewley J. P. GENEMAN of LASERGENE // Methods Mol. Biol. – 1997. – Vol. 70. – P. 189-196.
6. Clewley J. P. Macintosh sequence analysis software. DNAStar's LaserGene // Mol. Biotechnol. – 1995. – Vol. 3, № 3. – P. 221-224.
7. Clewley J. P., Arnold C. MEGALIGN. The multiple alignment module of LASERGENE // Methods Mol. Biol. – 1997. – Vol. 70. – P. 119-129.
8. Downey T. Analysis of a multifactor microarray study using Partek genomics solution // Methods Enzymol. – 2006. – Vol. 411. – P. 256-70.
9. Farhat M. R., Sultana R., Iartchouk O., Bozeman S., Galagan J. et al. Genetic determinants of drug resistance in *Mycobacterium tuberculosis* and their diagnostic value // Am. J. Respir. Crit. Care Med. – 2016. – P. 652-657.
10. Ford C., Yusim K., Ioerger T., Feng S., Chase M. et al. *Mycobacterium tuberculosis* – heterogeneity revealed through whole genome sequencing // Tuberculosis (Edinb). – 2012. – Vol. 92, № 3. – P. 194-201.
11. Ford C. B., Lin P. L., Chase M. R., Shah R. R., Iartchouk O. et al. Use of whole genome sequencing to estimate the mutation rate of *Mycobacterium tuberculosis* during latent infection // Nat. Genet. – 2011. – Vol. 43, № 5. – P. 482-486.
12. Gaifer Z., Babiker A., Rizavi D. Epidemiology of drug-resistant tuberculosis in a tertiary care center in Oman, 2006-2015 // Oman. Med. J. – 2017. – Vol. 32, № 1. – P. 36-40.
13. Gilbert G. L., Sintchenko V. The use of mycobacterial interspersed repetitive unit typing and whole genome sequencing to inform tuberculosis prevention and control activities // N S W Public Health Bull. – 2013. – Vol. 24, № 1. – P. 10-14.
14. Glynn J. R., Guerra-Assuncao J. A., Houben R. M., Sichali L., Mzembe T. et al. Whole genome sequencing shows a low proportion of tuberculosis disease is attributable to known close contacts in rural malawi // PLoS One. – 2015. – Vol. 10, № 7. – P. e0132840.
15. Grayson B. L., Aune T. M. A comparison of genomic copy number calls by Partek Genomics Suite, Genotyping Console and Birdsuite algorithms to quantitative PCR // BioData Min. – 2011. – Vol. 4. – P. 8.
16. Guthrie J. L., Gardy J. L. A brief primer on genomic epidemiology: lessons learned from *Mycobacterium tuberculosis* // Ann. N Y Acad Sci. – 2017. – Vol. 1388, № 1. – P. 59-77.
17. Hiltemann S., Mei H., de Hollander M., Palli I., van der Spek P. et al. CGtag: complete genomics toolkit and annotation in a cloud-based Galaxy // Gigascience. – 2014. – Vol. 3, № 1. – P. 1.
18. Madduri R. K., Sulakhe D., Lacinski L., Liu B., Rodriguez A. et al. Experiences Building Globus Genomics: A Next-Generation Sequencing Analysis Service using Galaxy, Globus, and Amazon Web Services // Concurr Comput. – 2014. – Vol. 26, № 13. – P. 2266-2279.
19. Pedersen B. S., Layer R. M., Quinlan A. R. Vcfanno: fast, flexible annotation of genetic variants // Genome Biol. – 2016. – Vol. 17, № 1. – P. 118.
20. Steiner A., Stucki D., Coscolla M., Borrell S., Gagneux S. KvarQ: targeted and direct variant calling from fastq reads of bacterial genomes // BMC Genomics. – 2014. – Vol. 15. – P. 881.

REFERENCES

1. Bradley P., Gordon N.C., Walker T.M., Dunn L., Heys S. et al. Corrigendum: Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and *Mycobacterium tuberculosis*. *Nat. Commun.*, 2016, vol. 7, pp. 11465.
2. Bradley P., Gordon N.C., Walker T.M., Dunn L., Heys S. et al. Rapid antibiotic-resistance predictions from genome sequence data for *Staphylococcus aureus* and *Mycobacterium tuberculosis*. *Nat. Commun.*, 2015, vol. 6, pp. 10063.
3. Burland T.G. DNASTAR's Lasergene sequence analysis software. *Methods Mol. Biol.*, 2000, vol. 132, pp. 71-91.
4. Cingolani P., Platts A., Wang le L., Coon M., Nguyen T. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)*, 2012, vol. 6, no. 2, pp. 80-92.
5. Clewley J.P. GENEMAN of LASERGENE. *Methods Mol. Biol.*, 1997, vol. 70, pp. 189-196.
6. Clewley J.P. Macintosh sequence analysis software. DNAStar's LaserGene. *Mol. Biotechnol.*, 1995, vol. 3, no. 3, pp. 221-224.
7. Clewley J.P., Arnold C. MEGALIGN. The multiple alignment module of LASERGENE. *Methods Mol. Biol.*, 1997, vol. 70, pp. 119-29.
8. Downey T. Analysis of a multifactor microarray study using Partek genomics solution. *Methods Enzymol.*, 2006, vol. 411, pp. 256-70.
9. Farhat M.R., Sultana R., Iartchouk O., Bozeman S., Galagan J. et al. Genetic Determinants of Drug Resistance in *Mycobacterium tuberculosis* and Their Diagnostic Value. *Am. J. Respir. Crit. Care Med.*, 2016, pp. 652-657.
10. Ford C., Yusim K., Ioerger T., Feng S., Chase M. et al. Mycobacterium tuberculosis – heterogeneity revealed through whole genome sequencing. *Tuberculosis (Edinb.)*, 2012, vol. 92, no. 3, pp. 194-201.
11. Ford C.B., Lin P.L., Chase M.R., Shah R.R., Iartchouk O. et al. Use of whole genome sequencing to estimate the mutation rate of *Mycobacterium tuberculosis* during latent infection. *Nat. Genet.*, 2011, vol. 43, no. 5, pp. 482-486.
12. Gaifer Z., Babiker A., Rizavi D. Epidemiology of Drug-resistant Tuberculosis in a Tertiary Care Center in Oman, 2006-2015. *Oman. Med. J.,* 2017, vol. 32, no. 1, pp. 36-40.
13. Gilbert G.L., Sintchenko V. The use of mycobacterial interspersed repetitive unit typing and whole genome sequencing to inform tuberculosis prevention and control activities. *N S W Public Health Bull.*, 2013, vol. 24, no. 1, pp. 10-14.
14. Glynn J.R., Guerra-Assuncao J.A., Houben R.M., Sichali L., Mzembe T. et al. Whole genome sequencing shows a low proportion of tuberculosis disease is attributable to known close contacts in rural malawi. *PLoS One*, 2015, vol. 10, no. 7, pp. e0132840.
15. Grayson B.L., Aune T.M. A comparison of genomic copy number calls by Partek Genomics Suite, Genotyping Console and Birdsuite algorithms to quantitative PCR. *BioData Min.*, 2011, vol. 4, pp. 8.
16. Guthrie J.L., Gardy J.L. A brief primer on genomic epidemiology: lessons learned from *Mycobacterium tuberculosis*. *Ann. N Y Acad Sci.*, 2017, vol. 1388, no. 1, pp. 59-77.
17. Hiltemann S., Mei H., de Hollander M., Palli I., van der Spek P. et al. CGtag: complete genomics toolkit and annotation in a cloud-based Galaxy. *Gigascience*, 2014, vol. 3, no. 1, pp. 1.
18. Madduri R.K., Sulakhe D., Lacinski L., Liu B., Rodriguez A. et al. Experiences Building Globus Genomics: A Next-Generation Sequencing Analysis Service using Galaxy, Globus, and Amazon Web Services. *Concurr Comput.*, 2014, vol. 26, no. 13, pp. 2266-2279.
19. Pedersen B.S., Layer R.M., Quinlan A.R. Vcfanno: fast, flexible annotation of genetic variants. *Genome Biol.*, 2016, vol. 17, no. 1, pp. 118.
20. Steiner A., Stucki D., Coscolla M., Borrell S., Gagneux S. KvarQ: targeted and direct variant calling from fastq reads of bacterial genomes. *BMC Genomics*, 2014, vol. 15, pp. 881.

ДЛЯ КОРРЕСПОНДЕНЦИИ:

Объединенный институт проблем информатики
Национальной академии наук Беларуси,
220012, Республика Беларусь, г. Минск, ул. Сурганова, д. 6.

Спринджук Матвей Владимирович
научный сотрудник,
разработчик программного обеспечения.
Тел.: +375 29 567 10 73.
E-mail: msprindzhuk@mail.ru

Сергеев Роман Сергеевич
научный сотрудник, математик.
Тел.: +375 29 557 41 67.
E-mail: roma.sergeev@gmail.com

РНПЦ фтизиатрии и пульмонологии,
220035, Республика Беларусь, г. Минск,
Долгиновский тракт, д. 157.

Скрягина Алена Михайловна
доктор медицинских наук, профессор,
заместитель директора.
Тел.: +375 29 184 07 83.
E-mail: alena.skrahina@gmail.com

Залуцкая Оксана Михайловна
врач-бактериолог.
Тел.: +375 29 325 36 06.
E-mail: akasana@inbox.ru

Титов Леонид Петрович
РНПЦ эпидемиологии и микробиологии,
доктор медицинских наук, профессор,
член-корреспондент НАН Беларуси, академик РАМН,
заведующий лабораторией иммунологии и микробиологии.
220114, Республика Беларусь, г . Минск, ул. Филимонова, д. 23.
Тел.: +375 29 626 58 66.
E-mail: leonidtitov@tut.by

Скрягин Александр Егорович
Белорусский государственный медицинский университет,
кандидат медицинских наук, доцент кафедры
анестезиологии и реаниматологии, врач-реаниматолог.
220116, Республика Беларусь, г. Минск, пр. Дзержинского, д. 83.
Тел.: +375 29 679 98 71.
E-mail: aliaksandr_skrahin@tut.by

Поступила 20.04.2017

FOR CORRESPONDENCE:

United Institute of Informatics Problems,
National Academy of Sciences of Belarus,
6, Surganova St., Minsk, Belarus, 220012

Matvey V. Sprindzhuk
Researcher,
Developer of the software.
Phone: +375 29 567 10 73.
E-mail: msprindzhuk@mail.ru

Roman S. Sergeev
Researcher, Mathematician.
Phone: +375 29 557 41 67.
E-mail: roma.sergeev@gmail.com

Republic Scientific and Practical Center
of Pulmonology and Tuberculosis,
157, Dolginovsky Road, Minsk, Belarus Republic, 220035

Alena M. Skryagina
Doctor of Medical Sciences,
Professor, Deputy Director.
Phone: +375 29 184 07 83.
E-mail: alena.skrahina@gmail.com

Oksana M. Zalutskaya
Bacteriologist.
Phone: +375 29 325 36 06.
E-mail: akasana@inbox.ru

Leonid P. Titov
Republican Research and Practical Center for Epidemiology
and Microbiology, Doctor of Medical Sciences, Professor,
Correspondent Member of NAS of Belarus, Academician of
RAMS, Head of Immunological and Microbiological Laboratory
23, Filimonova St., Minsk, Belarus Republic, 220114
Phone: +375 29 626 58 66.
E-mail: leonidtitov@tut.by

Aleksander E. Skryagin
Belorussian State Medical University, Candidate of Medical
Sciences, Associate Professor of Anesthesiology and Intensive
Care Department, Intensive Care Physician.
83, Dzerzhinskogo Ave., Belarus Republic, 220116
Phone: +375 29 679 98 71.
E-mail: aliaksandr_skrahin@tut.by

Submitted as of 20.04.2017