УДК 519.7:614.4.

ПРОГРАММНО-ВЫЧИСЛИТЕЛЬНЫЙ КОМПЛЕКС «ОКУНЬ-2» ДЛЯ ОЦЕНКИ МУТАЦИОННОГО ПРОФИЛЯ ГЕНОВ РЕЗИСТЕНТНОСТИ И ВИРУЛЕНТНОСТИ СЕКВЕНИРОВАННЫХ ГЕНОМОВ МИКОБАКТЕРИИ ТУБЕРКУЛЕЗА

М.В. СПРИНДЖУК[1], Л.П. ТИТОВ[2], В.В. СЛИЗЕНЬ[3], А.Е. СКРЯГИН[4], Е.М. СКРЯГИНА[4], О.М. ЗАЛУЦКАЯ[4], А.П. КОНЧИЦ[5]

[1]*Объединенный институт проблем информатики НАН Беларуси, Республика Беларусь*

[2]*РНПЦ эпидемиологии и микробиологии, Республика Беларусь*

[3]*Белорусский государственный медицинский университет, Республика Беларусь*

[4]*РНПЦ фтизиатрии и пульмонологии, Республика Беларусь*

[5]*Институт леса НАН Беларуси, Республика Беларусь*

Поступила в редакцию 29 марта 2018

Аннотация. Приводится описание нового программно-вычислительного комплекса, предназначенного для обработки данных полных геномов микобактерии туберкулеза человека с целью получения информации о профиле резистентности и вирулентности туберкулеза.

Ключевые слова: туберкулез, геномика, полногеномное секвенирование, антибиотики, устойчивость к противотуберкулезным лекарственным средствам, биоинформатика.

Abstract. Authors describe a new software and computer complex designed and developed for the processing of the data of the whole genomes of mycobacterium tuberculosis with the purpose of obtaining information about the profile of tuberculosis resistance and virulence.

Keywords: tuberculosis, genomics, whole genome sequencing, antibiotics, resistance to anti-tuberculosis drugs, bioinformatics.

Doklady BGUIR. 2018, Vol. 116, No. 6, pp. 40-45
«Okun-2» software-computing complex for calculating the mutational profile of samples of the sequenced whole genomes of mycobacteria tuberculosis
M.V. Sprindzuk, L.P. Titov, V.V. Slizen, A.E. Skryahin, E.M. Skryahina, O.M. Zalutskaya, A.P. Konchits

Введение

Несмотря на значительные достижения в области клинической медицины, эпидемиологии и микробиологии, проблема туберкулеза остается весьма актуальной и в XXI веке. Согласно данным ВОЗ, ежегодно в мире заболевает около 10,5 млн человек и около 3 миллионов умирает. В популяции возбудителя туберкулеза – Mycobacterium tuberculosis в последние несколько десятилетий произошло ряд существенных изменений: а) появились и широко распространились генетические варианты множественно- и экстремально резистентных к противотуберкулезным препаратам бактерий; б) расширился спектр мутаций в генах, определяющих резистентность и вирулентность микроба;

в) клональное распространение генетических вариантов поражающих человека микобактерий на географических территориях. Одновременно быстрыми темпами происходит развитие молекулярно-генетических технологий, основанных на анализе полимеразной цепной реакции ДНК, мультилокусмном сиквенс-типировании, секвенировании полных геномов, и их внедрение в лабораторную практику с целью ускорения диагностики, определения спектра мутаций резистентности и молекулярных маркеров эпиданализа. К настоящему времени в международных базах данных накопилось значительное количество завершенных и фрагментов незавершенных секвенированных геномов микобактерий туберкулеза, для анализа которых требуются сложные биоинформационные программы и соответствующие специалисты по биоинформатике и смежным дисциплинам.

Вместе с тем разработка качественного программного обеспечения для практических задач современной микробиологии и смежных дисциплин − актуальная тема кибернетики и прикладной математики. Достижения полногеномного секвенирования и развитие его технических аппаратных средств требует разработки и внедрения нового программного обеспечения, способного оптимизировать труд научных сотрудников, работников лабораторий и клинических специалистов. Изучение особо опасных микробов, каким является туберкулезная палочка Коха, очевидно, доминирует по тематическому приоритету научной деятельности [1].

Цели и задачи исследования

Целью работы было разработать и внедрить новый программно-вычислительный комплекс, предназначенный для обработки данных полных геномов микобактерии туберкулеза человека с целью получения информации о профиле резистентности и вирулентности туберкулеза. Для достижения цели были изучены аналоги, подобраны модули программы, был спроектирован алгоритм необходимой обработки данных, разработан программный интерфейс, написан и протестирован программный код, написаны элементы документации программного обеспечения.

Материалы и методы. Обсуждение результатов

Для реализации алгоритма обработки геномных данных были выбраны языки программирования Python 2.7 и Linux Shell. Для создания интерфейса программного обеспечения был отобран пакет Python GTK (см. рис. 1).

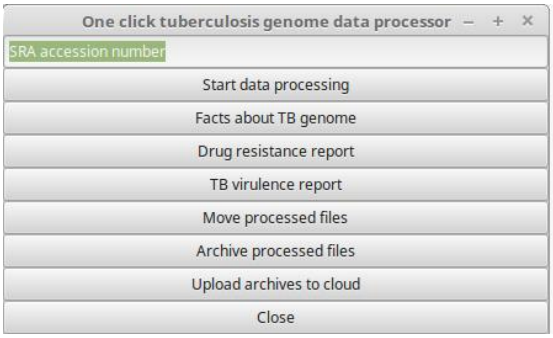

Рис. 1. Интерфейс разработанного программного обеспечения

Известны многоцелевые программно-вычислительные комплексы для задач биоинформатики: Ugene, NextGene Softgenetics, DNAStar, CLC Genomics Workbench, Galaxy, требующие инвестирования больших объемов времени на настройку, изучение функциональности и т. п. Существуют специальные базы геномических данных микобактерии туберкулеза и сервисы для их обработки [2, 3, 5, 7, 8, 11]. Имеется целый ряд программного обеспечения, специально разработанного для изучения антибиотикорезистентности патогенных микроорганизмов [4, 6, 9, 10].

Авторами разработан программный комплекс с возможностью исследовать как полный мутационный профиль исследуемого геномного образца, так и фокусироваться на отдельных группах генов. Как компоненты-модули были использованы следующие программные инструменты и библиотеки кодов (см. рис. 2, табл. 1): SRA-tools (1, 2), BWA (3), SAM-

Tools (4), Pilon (5), VT и RTG (6), BedTools и SNPEff (7), Mega.py (8). Текущая версия программного обеспечения способна вычислять мутации в 40 генах резистентности и 20 – вирулентности (см. табл. 2 и 3).

Рис. 2. Алгоритм разработанного программного обеспечения

Таблица 1. **Описание данных вычислений и результатов работы программного обеспечения**

0001.fastq, 0002.fastq	Автоматически переименованные загруженные по SRA идентификатору исходные файлы
aligned_sam_result.sam	Первый результат картирования загруженного генома против ссылочного генома-эталона
unsorted_bam.bam	Результат конвертации формата SAM в формат BAM
file.sorted.bam	Отсортированный BAM файл
pilon_output.pilon.vcf	Первый файл запроса вариантов, результат работы Pilon
PilonOutputFileFinal.vcf	Первый результат предобработки сырого файла запроса вариантов
PilonOutputFileFinalBetter.vcf	Второй результат предобработки/нормализации/ декомпозиции сырого файла запроса вариантов
DecomposedPilonReducedresult.vcf	VCF файл после обработки VT
PilonSNPEffOutputStats.html	Основной отчет работы программы SNPEff
PilonSNPEffOutputStats.genes.txt	Список проаннотированных генов мутаций
PilonAnnotatedSNPeffResults.vcf	Проаннотированный SNPEff VCF файл, вышедший из Pilon
reducedPilonOutputFileFinal.vcf.log	Лог-файл ReducVCF
pilon_output.pilon.fasta	FASTA файл с мутациями, результат работы Pilon
SRAIdFromPythonInput.txt	Текстовый файл, содержащий SRA идентификаторы, введенные пользователем в форму интерфейса программного обеспечения
converted2bedvcf.bed	VCF файл, конвертированный в формат BED
TuberculosisSampleVirulenceReport!!!	Отчет о мутациях генов, ответственных за вирулентность (txt, текстовый файл)
TuberculosisSampleDrugResistanceReport.csv	Отчет о мутациях генов, ответственных за вирулентность (эксель файл, значения, разделенные запятой)
TuberculosisSampleVirulenceReport.csv	Отчет о мутациях генов, ответственных за вирулентность (эксель файл, значения, разделенные запятой)
RTG_VCF_Stats_Results.txt	Файл с результатами анализа фактов о геноме

Таблица 2. **Отобранные гены вирулентности микобактерии туберкулеза человека**

Хромосома	Начало	Конец	Идентификатор гена
AL123456	2726193	2726780	ahpC
AL123456	593871	594779	cmaA2
AL123456	3274072	3274902	drrC
AL123456	3243697	3245448	fadD26
AL123456	3283335	3285077	fadD28
AL123456	3983125	3984144	fadE28
AL123456	2487615	2489051	glnA1
AL123456	2278498	2278932	hspX
AL123456	557527	558813	icl11
AL123456	3023565	3024257	ideR
AL123456	2153889	2156111	katG
AL123456	199895	200935	mce1B
AL123456	736298	737203	mmaA4
AL123456	2630537	2632075	plcA
AL123456	2628781	2630319	plcB
AL123456	2627172	2628698	plcC
AL123456	1986854	1987696	plcD
AL123456	4161815	4162258	Rv3718c
AL123456	1364413	1365186	sigE
AL123456	3598901	3599551	sigH

Таблица 3. **Отобранные гены резистентности микобактерии туберкулеза человека**

Хромосома	Начало	Конец	Идентификатор гена
AL123456	2520743	2522164	accD6
AL123456	2726193	2726780	ahpC
AL123456	3153039	3154631	efpA
AL123456	4243233	4246517	embA
AL123456	4246514	4249810	embB
AL123456	4246514	4249810	embB
AL123456	4239863	4243147	embC
AL123456	1416181	1417347	embR
AL123456	4326004	4327473	ethA
AL123456	2516787	2517695	fabD
AL123456	3505363	3506769	fadE24
AL123456	156578	157600	fbpC
AL123456	408634	409173	furA
AL123456	4407528	4408202	gid
AL123456	7302	9818	gyrA
AL123456	5123	7267	gyrB
AL123456	1674202	1675011	inhA
AL123456	409362	410801	iniB
AL123456	412757	414238	iniC
AL123456	2518115	2519365	kasA
AL123456	2153889	2156111	katG
AL123456	1673440	1674183	mabA
AL123456	2101651	2103042	ndh
AL123456	4007331	4008182	nhoA
AL123456	2725571	2726087	oxyR
AL123456	2288681	2289241	pncA
AL123456	398658	399524	rmlA
AL123456	3646895	3647809	rmlD
AL123456	759807	763325	rpoB
AL123456	781560	781934	rpsL
AL123456	408634	409173	Rv0340
AL123456	408634	409173	Rv0340
AL123456	1792400	1793740	Rv1592c
AL123456	2006636	2006947	Rv1772
AL123456	3489506	3490375	Rv3124
AL123456	3490476	3491651	Rv3125c
AL123456	3491808	3492122	Rv3126
AL123456	3073680	3074471	thyA
AL123456	1917940	1918746	tlyA

Заключение

Разработан новый программно-вычислительный комплекс, предназначенный для обработки данных полных геномов микобактерии туберкулеза человека с целью получения информации о профиле резистентности и вирулентности туберкулеза. Программное обеспечение может быть адаптировано для обработки данных практически любого микроба. Код программного обеспечения доступен для интересующихся исследователей по письменному запросу у авторов.

Авторы заявляют об отсутствии конфликта интересов. Исследование выполнялось при поддержке CRDF, ОИПИ НАН Беларуси, БГМУ, РНПЦ эпидемиологии и микробиологии Министерства здравоохранения Республики Беларусь.

Список литературы / References

1. Transmission Electron Microscopy of XDR Mycobacterium tuberculosis Isolates Grown on High Dose of Ofloxacin / M. Arjomandzadegan [et al.] // Sci Pharm. 2017. № 1. P. 3–10.
2. Computational databases, pathway and cheminformatics tools for tuberculosis drug discovery / S. Ekins [et al.] // Trends Microbiol. 2011. № 2. P. 65–74.
3. Bioinformatics tools and databases for whole genome sequence analysis of Mycobacterium tuberculosis / K. Faksri [et al.] // Infect Genet Evol. 2016. № 1. P. 359–368.
4. PhyResSE: a Web Tool Delineating Mycobacterium tuberculosis Antibiotic Resistance and Lineage from Whole-Genome Sequencing Data / S. Feuerriegel [et al.] // J Clin Microbiol. 2015. № 6. P. 1908–1914.
5. Unipro UGENE NGS pipelines and components for variant calling, RNA-seq and ChIP-seq data analyses / O. Golosova [et al.] // PeerJ. 2014. № 2. P. e644.
6. CASTB (the comprehensive analysis server for the Mycobacterium tuberculosis complex): A publicly accessible web server for epidemiological analyses, drug-resistance prediction and phylogenetic comparison of clinical isolates / H. Iwai [et al.] // Tuberculosis (Edinb). 2015. № 6. P. 843–844.
7. Jain N.C. Information retrieval of tuberculosis literature in e-databases // Indian J Tuberc. 2014. № 3. P. 186–188.
8. Shared bioinformatics databases within the Unipro UGENE platform / I.V. Protsyuk [et al.] // J Integr Bioinform. 2015. № 1. P. 257.
9. Mycobacterium tuberculosis resistance prediction and lineage classification from genome sequencing: comparison of automated analysis tools / V. Schleusener [et al.] // Scientific Reports. 2017. № 4. P. 46327.
10. KvarQ: targeted and direct variant calling from fastq reads of bacterial genomes / A. Steiner [et al.] // BMC Genomics. 2014. № 15. P. 881.
11. ExpertDiscovery and UGENE integrated system for intelligent analysis of regulatory regions of genes / Y.Y. Vaskin [et al.] // In Silico Biol. 2011. № 3–4. P. 97–108.

Сведения об авторах

Спринджук М.В., научный сотрудник Объединенного института проблем информатики НАН Беларуси.

Титов Л.П., д.м.н., профессор, член-корреспондент НАН Беларуси, заведующий лабораторией РНПЦ эпидемиологии и микробиологии.

Слизень В.В., к.м.н., доцент, доцент Белорусского государственного медицинского университета.

Скрягин А.Е., к.м.н., доцент, врач-фтизиатр, анестезиолог-реаниматолог РНПЦ фтизиатрии и пульмонологии.

Скрягина Е.М., д.м.н., профессор, заместитель директора РНПЦ фтизиатрии и пульмонологии.

Information about the authors

Sprindzuk M.V., researcher of United institute of informatics problems of National academy of sciences of Belarus.

Titov L.P., D.Sci, professor, corresponding member of NAS of Belarus, head of laboratory of Republican scientific and practical centre for epidemiology and microbiology.

Slizen V.V., PhD, associate professor, associate professor of Belarusian state medical university.

Skryahin A.E., PhD, associate professor, phthisiatrician, intensive care physician and anesthesiologist of RSPC for pulmonology and pulmonology.

Skriahina E.M., D.Sci, professor, deputy director of RSPC for pulmonology and pulmonology.

Залуцкая О.М., врач-бактериолог Республиканской референс-лаборатории РНПЦ фтизиатрии и пульмонологии.

Кончиц А.П., к.б.н., ведущий научный сотрудник Института леса НАН Беларуси.

Адрес для корреспонденции

220012, Республика Беларусь
г. Минск, ул. Сурганова, 6,
Объединенный институт
проблем информатики НАН Беларуси
тел. +375-33-682-57-55;
e-mail: bioinformatics_bel@yahoo.com;
Спринджук Матвей Владимирович

Zalutskaya A.M., bacteriologist of Republican reference laboratory of RSPC for pulmonology and pulmonology.

Konchits A.P., PhD, leader researcher of Forestry institute of NAS of Belarus.

Address for correspondence

220012, Republic of Belarus
Minsk, Surganova st., 6,
United institute of informatics problems
of National academy of sciences of Belarus
tel. +375-33-682-57-55;
e-mail: bioinformatics_bel@yahoo.com
Sprindzuk Matvey Vladimirovich