http://doi.org/10.35596/2522-9613-2022-28-3-73-81

Оригинальная статья
Original paper

УДК 578.834.1, 5, 575.112

ПРОФИЛИРОВАНИЕ ГЕНОМНЫХ ДАННЫХ КОРОНАВИРУСА ЧЕЛОВЕКА, ПОЛУЧЕННЫХ ОТ ПАЦИЕНТОВ В БЕЛАРУСИ

М. В. СПРИНДЖУК, А. С. ВЛАДЫКО, Л. П. ТИТОВ, В. И. БЕРНИК

Институт математики Национальной академии наук Беларуси»
(г. Минск, Республика Беларусь)

Поступила в редакцию 22.04.2022

Аннотация. Новая коронавирусная инфекция стала причиной смерти и заболеваний миллионов людей и животных. Пандемия показала недостатки систем здравоохранения даже самых экономически развитых государств. Геномика и биоинформатика предоставляют возможность получать, изучать и анализировать геномные тексты микробов (в частности – коронавирусов). В статье представлены результаты анализа геномов SARS-CoV-2 от пациентов в Беларуси и (для сравнения) в России. Выполнено геномное профилирование с целью определения и статистического анализа кластеров и линий передачи новой коронавирусной инфекции, в соответствии с предложенными классификациями кластеров и штаммов COVID-19. Приводятся сведения по оценке качества исходных данных, выполнена и графически представлена визуализации полученных результатов. Доминирующие в Беларуси и России клады-кластеры это B.1 («Базельский кластер») и B.1.1. Оба имеют европейско-британское распространение.

Ключевые слова: коронавирус, пандемия, статистика, системы медицинского назначения, медицинская кибернетика, геномика, биоинформатика, прикладная математика, программное обеспечение, полногеномное секвенирование

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Для цитирования. Спринджук М. В., Владыко А. С., Титов Л. П., Берник В. И. Профилирование геномных данных коронавируса человека, полученных от пациентов в Беларуси. *Цифровая трансформация*. 2022; 18(3): 73-81.

PROFILING HUMAN CORONAVIRUS GENOMIC DATA OBTAINED FROM PATIENTS IN BELARUS

MATVEY V. SPRINDZUK, ALEXANDER S. VLADYKO,
LEONID P. TITOV, VASSILII I. BERNIK

Institute of Mathematics of the National Academy of Sciences of Belarus
(Minsk, Republic of Belarus)

Submitted 22.04.2022

Abstract. The new coronavirus infection has caused the death and injury of millions of people and animals. The pandemic has shown the shortcomings of the health care systems of even the most economically developed countries. Genomics and bioinformatics provide an opportunity to obtain, study and analyze the genomic texts of microbes, coronaviruses in particular. The article presents the results of the analysis of SARS-CoV-2 genomes from patients in Belarus and (for comparison) in Russia. Genomic profiling was performed to identify and statistically analyze clusters and lines of transmission of the new coronavirus infection, in accordance with the proposed classifications of COVID-19 clades. The information on the assessment of the quality of the initial data are reported, the visualization of the results obtained is made and graphically presented. The dominant clades-clusters in Belarus and Russia are B.1 ("Basel cluster") and B.1.1. Both have European-British geographical distribution.

Keywords: coronavirus, pandemic, statistics, medical systems, medical cybernetics, genomics, bioinformatics, applied mathematics, software, whole genome sequencing.

Conflict of interests. The authors declare no conflict of interests.

For citation. Sprindzuk M. V., Vladyko A. S., Titov L. P., Bernik V. I. Profiling Human Coronavirus Genomic Data Obtained from Patients in Belarus. *Digital Transformation*. 2022; 18(3): 73-81.

Введение

Вспышка новой коронавирусной инфекции COVID-19 началась в середине декабря 2019 г. в Китае, в городе Ухань, и распространилась на многие города Китая, Юго-Восточной Азии, а также по всему миру. За период с декабря 2019 г. по март 2022 г. в мире умерло 5 873 066 больных с подтвержденным диагнозом новой коронавирусной инфекции. За это же время в Беларуси умерло 6 313 пациента, в России – 343 173, на Украине – 103 824 [https://www.worldometers.info/coronavirus/]. Основным источником коронавирусной инфекции является больной человек, в том числе находящийся в инкубационном периоде заболевания.

Современные технологии высокопроизводительного секвенирования позволяют выделить и амплифицировать сравнительно самый крупный геном РНК коронавируса [1–6]. Анализ полученных биоинформационных данных дает возможность идентифицировать и классифицировать элементарные структуры генома для сравнительного анализа с известными геномами различных микроорганизмов, растений и млекопитающих. С усовершенствованиями и доступностью компьютерной техники появились новые возможности эффективно изучать биоинформационные данные, под которыми понимают геномный текст и соответствующие метаданные. В широком смысле данного термина под такими данными подразумевается сигнальная информация от биологических объектов, которая может иметь не только текстовый формат, но и представление в виде изображений, звука, видео и т.п.

Методика проведения эксперимента, материалы и методы

В течение 2021 г. авторами была выполнена загрузка SARS-CoV-2 геномов из общедоступной базы данных GISAID (Global Initiative on sharing all influenza data = Глобальная инициатива по обмену всеми данными о гриппе) [https://www.epicov.org/epi3/frontend#275474] четыре раза, в последний раз 04.11.2021 г. На этот момент в базе данных находилось 88 геномов, но только 46 были полными и имели достаточно высокое покрытие. Для предобработки полученных данных с целью оценки качества и состава генетического материала и идентификации происхождения изолятов мы отобрали программное обеспечение Pangolin и Nextclade (рис. 1–8). Для визуализации было отобрано и применено статистическое программное обеспечение общего назначение JMP SAS 10 (временная испытательная версия) [7, 8]. Pangolin [9, 10] был разработан для реализации динамической номенклатуры линий и кластеров-кладов передачи SARS-CoV-2, известной как номенклатура Pango. Это позволяет пользователю компьютерной программы исследовать последовательности геномов SARS-CoV-2 путем идентификации вероятной линии происхождения (линия Pango) вируса. Pangolin присваивает распознанную линию и имя кластера по принципу, опубликованному в первоисточнике A. Rambaut др., 2020 (рис. 9) [11]. Программное обеспечение Pangolin доступно как инструмент командной строки Linux и веб-приложение. Инструмент командной строки имеет открытый исходный код, доступный под лицензией GNU General Public License v3.0.

Nextclade [12] – это инструмент, который определяет различия между загруженными пользователем геномными текстами и эталонной последовательностью и использует эти различия для идентификации, распознавания и присвоения линий передачи и кластеров-кладов, а также сообщает о потенциальных проблемах качества последовательностей в представляемых данных. Целесообразно использовать этот инструмент для анализа последовательностей перед их загрузкой в базу данных или для отбора по качеству для последующих вычислительных экспериментов. Исходный код приложения и алгоритмов является открытым и доступен на GitHub. Руководство пользователя доступно по адресу docs.nextstrain.org/projects/nextclade. Nextclade был первоначально разработан во время пандемии COVID-19, в первую очередь ориентирован на SARS-CoV-2, но может анализировать и другие данные.

Доминирующие в Беларуси клады-кластеры B.1 («Базельский кластер») и B.1.1 имеют европейско-британское распространение, имеются публикации об аналогичном нашему исследованию в Швейцарии, где общая частота распространения B.1 составила 68,2 % [13]. Информации о специфических клинических и лабораторных особенностях этого кластера на сегодняшний день недостаточно. Анализ данных выполнялся авторами до появления штамма Omicron B.1.529.

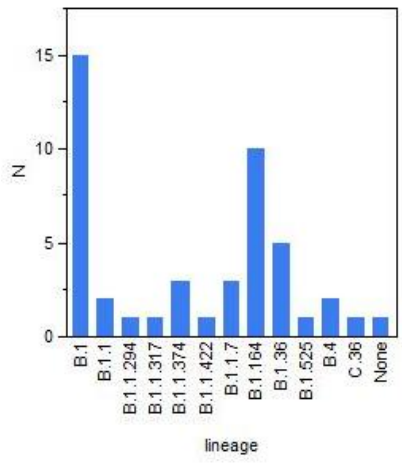

Рис. 1. Столбчатая диаграмма числового представления результатов классификации геномов SARS-CoV-2, полученных от пациентов на территории Беларуси
Fig. 1. Bar chart of numerical representation of the results of the classification of SARS-CoV-2 genomes obtained from patients in Belarus

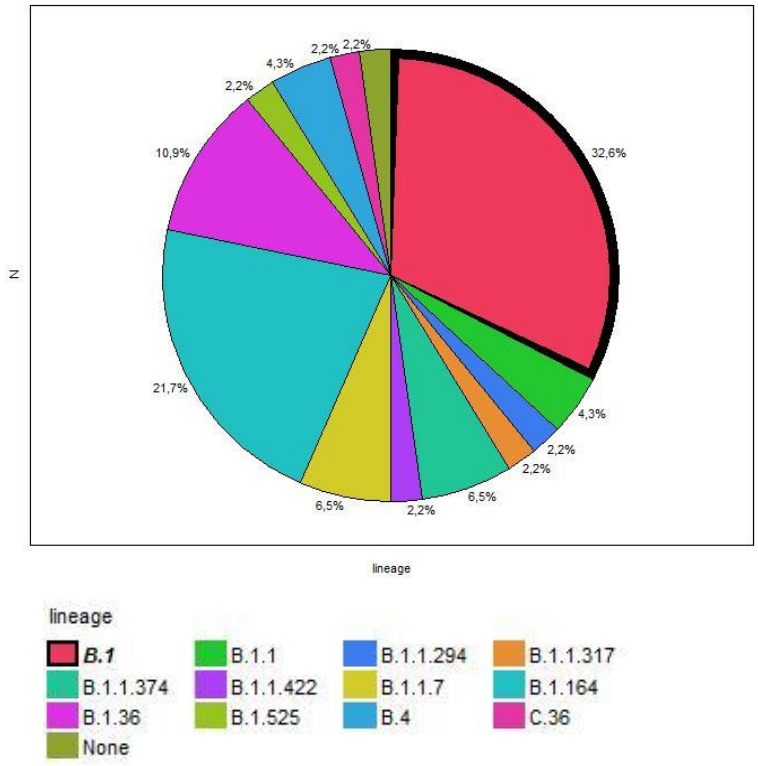

Рис. 2. Круговая диаграмма процентного представления результатов классификации геномов SARS-CoV-2, полученных от пациентов на территории Беларуси
Fig. 2. Pie chart of the percentage representation of the results of the classification of SARS-CoV-2 genomes obtained from patients in Belarus

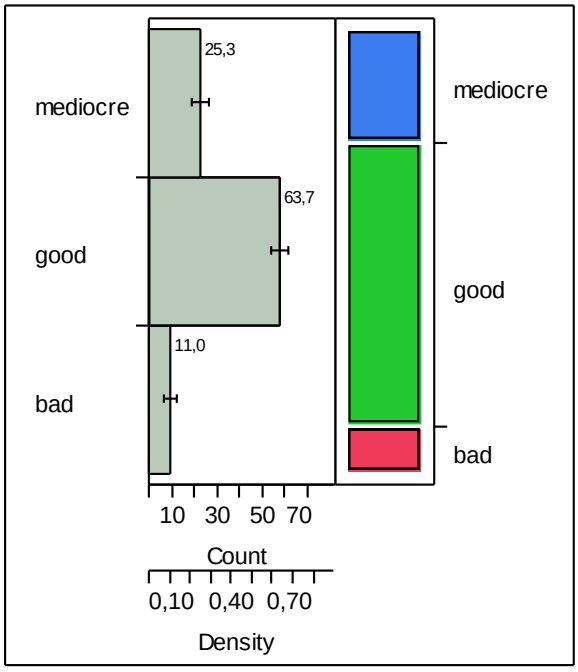

Рис. 3. Столбчатая диаграмма числового представления результатов оценки качества всех геномных текстов SARS-CoV-2, полученных от пациентов и животных на территории Беларуси (88 образцов)
Fig. 3. Bar chart of numerical representation of the results of quality assessment of all SARS-CoV-2 genomic texts obtained from patients and animals in Belarus (88 samples)

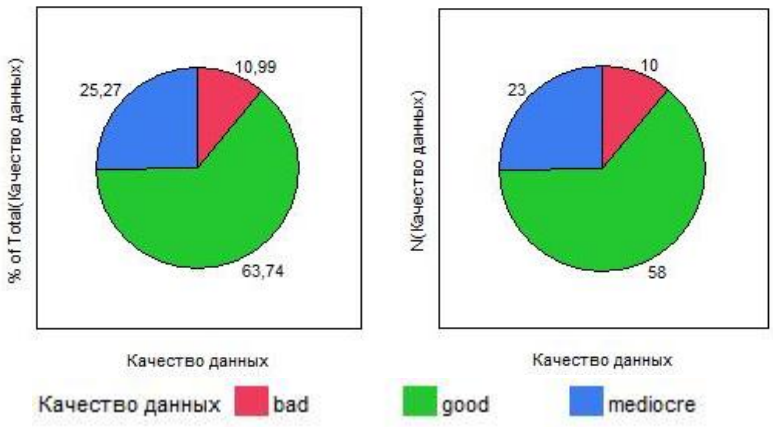

Рис. 4. Круговая диаграмма числового (слева) и процентного (справа) представления результатов оценки качества всех геномных текстов SARS-CoV-2, полученных от пациентов и животных на территории Беларуси (88 образцов)

Fig. 4. Pie chart of count (left) and percentage (right) representation of the results of quality assessment of all SARS-CoV-2 genomic texts obtained from patients and animals in Belarus (88 samples)

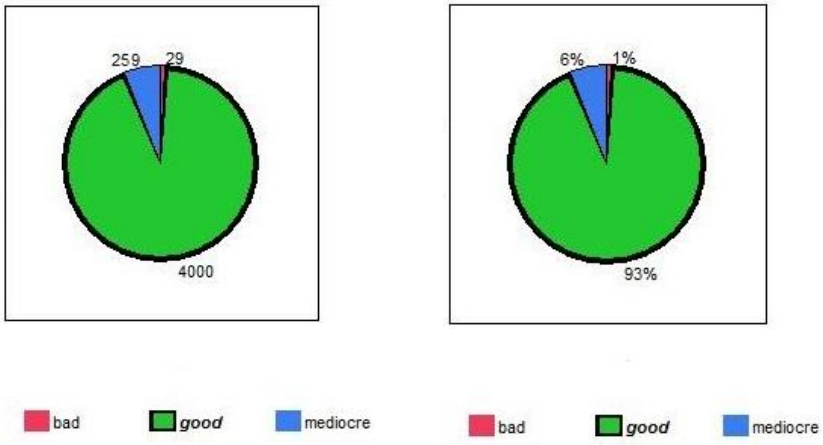

Рис. 5. Круговая диаграмма числового (слева) и процентного (справа) представления результатов оценки качества всех геномных текстов SARS-CoV-2, полученных от пациентов на территории России (4085 образцов)

Fig. 5. Pie chart of the count (left) and percentage (right) presentation of the quality assessment results for all SARS-CoV-2 genomic texts obtained from patients in Russia (4085 samples)

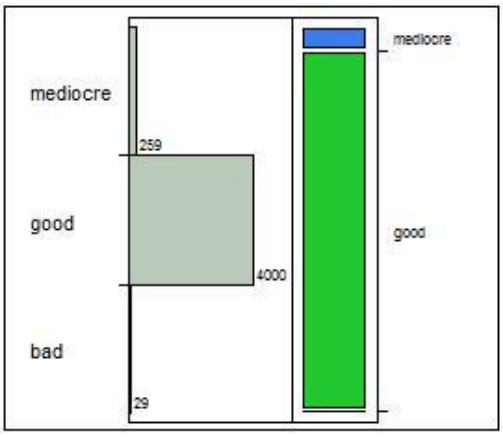

Рис. 6. Столбчатая диаграмма численного представления результатов оценки качества всех геномных текстов SARS-CoV-2, полученных от пациентов и на территории России (4085 образцов)

Fig. 6. Bar chart of numerical representation of the results of quality assessment of all SARS-CoV-2 genomic texts obtained from patients in Russia (4085 samples)

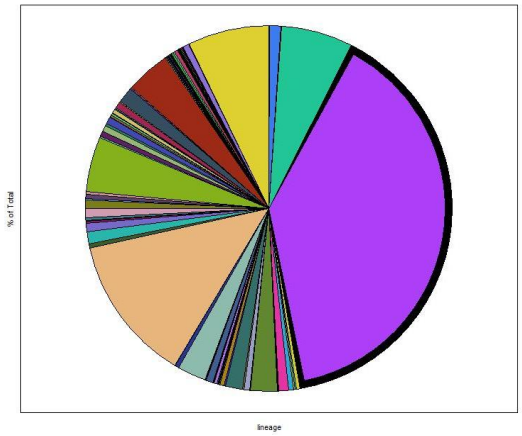

Рис. 7. Круговая диаграмма процентного представления результатов классификации всех геномных текстов SARS-CoV-2, полученных от пациентов и на территории России (4085 образцов)
Числа процентов не приводятся по причине большого количеств секторов
Fig. 7. Pie chart of the percentage presentation of the classification results for all SARS-CoV-2 genomic texts obtained from patients and in Russia (4085 samples)
Percentage numbers are not given due to the large number of sectors

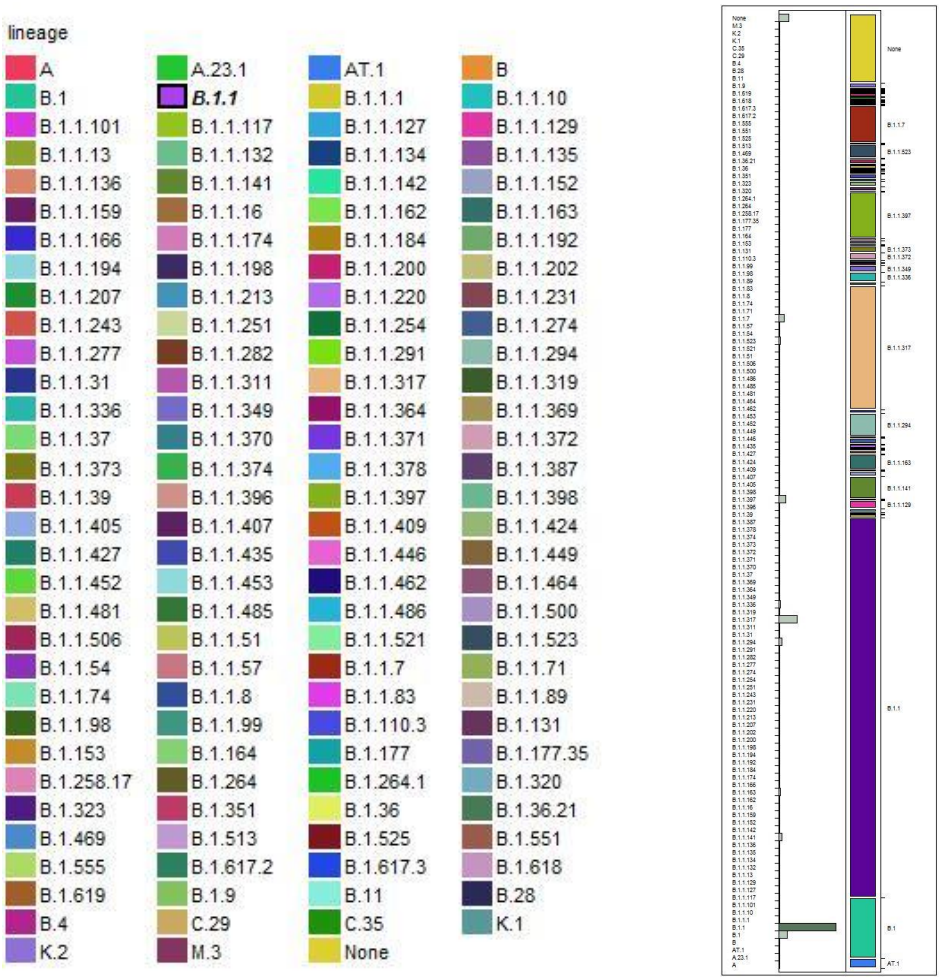

Рис. 8. Пояснение к круговой диаграмме процентного представления результатов классификации всех геномных текстов SARS-CoV-2, полученных от пациентов на территории России (4085 образцов)
Жирным курсивом и фиолетовым цветом показан доминирующий кластер B.1.1
Fig. 8. Explanation of the pie chart of the percentage presentation of the classification results for all SARS-CoV-2 genomic texts obtained from patients in Russia (4085 samples).
Bold italics and purple show the dominant cluster B.1.1

Линии по Rambaut и др.	Примечания к Rambaut и др.	Клады Nextstrain	Клады GISAID	Значимые штаммы или мутации
A.1–A.6		19B	S	
B.3–B.7, B.9, B.10, B.13–B.16		19A	L	
			O	
B.2			V	
B.1	B.1.5–B.1.72	20A	G	Линия B.1 по Rambaut и др. включает в себя штаммы с мутацией D614G{{{1}}}
	B.1.9, B.1.13, B.1.22, B.1.26, B.1.37		GH	
	B.1.3–B.1.66	20C		Включает Штамм 501.V2
	B.1.1	20B	GR	Включает Штамм 202012/01, штаммы B.1.1.207 и B.1.1.284
	B.1.177	20A.EU1	GV	

Рис. 9. Соответствие различных классификаций линий-кластеров SARS-CoV-2 [https://ru.wikipedia.org/wiki/Штаммы_SARS-CoV-2]
Fig. 9. Correspondence of different classifications of SARS-CoV-2 cluster lineages

Заключение

Осуществлены сбор и профилирование геномных данных коронавируса, полученных от пациентов в Беларуси. Для сравнения были загружены и обработаны российские данные. Выполнено распознавание линий передачи коронавирусной инфекции, а также оценка качества геномных текстов для отбора необходимых данных для дальнейших вычислительных экспериментов.

Результаты были визуализированы и разнопланово представлены. Апробированное программное обеспечение можно будет использовать как модули в разрабатываемой автоматизированной системе анализа биоинформационных данных геномной природы.

Список литературы / References

1. Woo, P. C. Coronavirus genomics and bioinformatics analysis / P. C. Woo [et al.] // Viruses. – 2010. – No. 2(8). – P. 1804-1820.
2. Abro, S. H. Bioinformatics and evolutionary insight on the spike glycoprotein gene of QX-like and Massachusetts strains of infectious bronchitis virus / S. H. Abro [et al.] // Virol J. – 2012. – No. 9. – P. 211.
3. Feng, D. Bioinformatics analysis of the factors controlling type I IFN gene expression in autoimmune disease and virus-induced immunity / D. Feng, B. J. Barnes // Front Immunol. – 2013. – No. 4. – P. 291.
4. Brinkmann, A. Proficiency Testing of Virus Diagnostics Based on Bioinformatics Analysis of Simulated In Silico High-Throughput Sequencing Data Sets / A. Brinkmann [et al.] // J Clin Microbiol. – 2019. – No. 57 (8). – DOI: 10.1128/JCM.-19.
5. Kellam, P. Virus bioinformatics: databases and recent applications / P. Kellam, M. M. Alba // Appl. Bioinformatics. – 2002. – No. 1(1). – P. 37–42.

6. Xing, J. F. Sequence analysis for genes encoding nucleoprotein and envelope protein of a new human coronavirus NL63 identified from a pediatric patient in Beijing by bioinformatics / J. F. Xing [et al.] // Bing Du Xue Bao. – 2007. – No. 23 (4). – P. 245–251.
7. Jones, B. JMP statistical discovery software / B. Jones, J. Sall // Wiley Interdisciplinary Reviews: Computational Statistics. – 2011. – No. 3(3). – P. 188–194.
8. Sall, J. JMP start statistics: a guide to statistics and data analysis using JMP / J. Sall [et al.]. – Sas Institute, 2017.
9. O'Toole, Á. Assignment of epidemiological lineages in an emerging pandemic using the Pangolin tool / Á. O'Toole [et al.] // Virus Evolution. – 2021. – No. 7(2). – Article ID: veab064.
10. Pipes, L. Assessing uncertainty in the rooting of the SARS-CoV-2 phylogeny / L. Pipes // Molecular biology and evolution. – 2021. – No. 38(4). – P. 1537–1543.
11. Rambaut, A. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology / A. Rambaut [et al.] // Nat Microbiol. – 2020. – No. 5 (11). – P. 1403–1407.
12. Aksamentov, I. Nextclade: clade assignment, mutation calling and quality control for viral genomes / I. Aksamentov [et al.] // Journal of Open SourceSoftware. – 2021. – No. 6(65). – P. 3773.
13. Stange, M. SARS-CoV-2 outbreak in a tri-national urban area is dominated by a B.1 lineage variant linked to a mass gathering event / M. Stange [et al.] // PLoS Pathogens. – 2021. – No. 17(3). – Article ID: e1009374.

Вклад авторов

Спринджук М.В. осуществил постановку задачи для проведения исследования, выполнил сбор и анализ данных, подготовил рукопись статьи.

Владыко А.С., Титов Л.П. участвовали в постановки целей и задач исследования, консультировали Спринджука М.В. по вопросам вирусологии и иммунологии, вычитывали рукопись статьи.

Берник В.И. участвовал в создании концепции исследования и редактировании рукописи.

Authors contribution

Sprindzuk M.V. carried out the formulation of the problem for the study, performed data analysis, prepared the manuscript of the article.

Vladyko A.S. and Titov L.P. participated in setting the goals and objectives of the study, advised Sprindzuk M.V. on questions of virology and immunology, proofread the manuscript.

Bernik V.I. participated in the creation of the concept of the study and editing the manuscript.

Работа выполнена при поддержке Белорусского республиканского фонда фундаментальных исследований в рамках проектов:

Ф21МН-001 «Математическое моделирование передачи и распространения COVID-19 инфекции на основе систем дифференциальных уравнений и алгоритмов обработки данных с применением технологии машинного обучения», № ГР 20213518 от 27.09.2021;

М21КОВИД-026 «Ретроспективный анализ клинического и иммунологического статуса групп COVID-19 пациентов с сопутствующим туберкулезом и ВИЧ инфекцией по данным РНПЦ Пульмонологии и фтизиатрии г. Минска», № ГР 20210456 от 31.03.2021;

М21COVID-001 «Разработка и скрининг мукозной вакцины против COVID-19 на основе векторной платформы кишечного аденовируса», № ГР 20210889 от 26.04.2021.

Сведения об авторах

Спринджук М. В., к. т. н., старший научный сотрудник Объединенного института проблем информатики НАН Беларуси.

Владыко А. С., д. м. н., профессор, главный научный сотрудник РНПЦ эпидемиологии и микробиологии.

Титов Л. П., д. м. н., профессор, академик НАН Беларуси, заведующий лабораторией экспериментальной иммунологии РНПЦ эпидемиологии и микробиологии.

Берник В. И., д. ф.-м. н., профессор, главный научный сотрудник отдела теории чисел государственного научного учреждения «Институт математики Национальной академии наук Беларуси».

Information about the authors

Sprindzuk M. V., Cand. of Sci., Senior Computer Scientist of the United Institute of Informatics Problems of NAS of Belarus.

Vladyko A. S., Dr. of Sci. (Med.), Professor, Principal Researcher of the Republican Scientific and Practical Center for Epidemiology and Microbiology.

Titov L. P., Dr. of Sci. (Med.), Professor, Academic of NASB, Head of the Laboratory of Experimental Immunology of RRPC for Epidemiology and Microbiology.

Bernik V. I., Dr. of Sci. (Phys. and Math.), Professor, Principal Researcher at the Department of Number Theory of the State Scientific Institution "Institute of Mathematics of the National Academy of Sciences of Belarus".

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

Address for correspondence

220012, Republic of Belarus,
Minsk, Surganova St. 6,
United Institute of Informatics Problems,
Belarus National Academy of Sciences;
tel. +375 33 682-57-55;
e-mail: bioinformatics_bel@yahoo.com;
Sprindzuk Matvey Vladimirovich