УДК 004.056

АВТОМАТИЗИРОВАННЫЙ КОНВЕЙЕР АНАЛИЗА ДАННЫХ ГЕНОМОВ КОРОНАВИРУСА

М.В. СПРИНДЖУК, В.И. БЕРНИК, Н.И. КАЛОША, А.С. ВЛАДЫКО, Б. УЛЗИЙБАТ, Б. БАТГЭРЕЛ, академик Л.П. ТИТОВ, Д.А. КЛИМУК, Е.М. СКРЯГИНА, А.Е. СКРЯГИН, Т.Н. ГЛИНСКАЯ

Государственное учреждение Институт математики Национальной академии наук Беларуси
Минск, Республика Беларусь

Аннотация. Разработан алгоритм-конвейер для анализа коронавирусных контигов. Осуществлена серия вычислительных экспериментов для изучения геномов нового коронавируса человека, направленных на оценку качества, профилирование, аннотирование геномных данных, полученных от пациентов Беларуси. Выявлены доминирующие на сегодняшний день линии передачи коронавируса *Дельта* и *Омикрон.* Выполненные исследования соответствуют мировым тенденциям исследования коронавируса и сфокусированы на изучении вирусной эволюции в современных условиях. Результаты способствуют распространению научной информации об этиологии, патогенезе, эпидемиологии и лечению коронавирусной инфекции и других опасных респираторных инфекций.

Ключевые слова: системы медицинского назначения, медицинская кибернетика, анализ геномов, пандемия, коронавирусная инфекция, программное обеспечение, автоматизация обработки данных, отображение информации, большие данные

AUTOMATED PIPELINE FOR ANALYSIS OF CORONAVIRUS GENOME DATA

SPRINDZUK M.V., BERNIK V.I., KALOSHA N.I., VLADYKO A.S., ULZIBAT B., BATGEREL B., TITOV L.P., KLIMUK D.A., SKRIAHINA E.M., SKRIAHIN A.E., GLINSKAYA T.N.

State Institution Institute of Mathematics of the National Academy of Sciences of Belarus Minsk, Belarus

Abstract. The data processing pipeline has been developed for the analysis of coronavirus contigs. A series of computational experiments were carried out to investigate the genomes of the new human coronavirus aimed at assessing the quality, profiling, and annotating genomic data obtained from patients in Belarus. The currently dominant lines of transmission of the coronavirus – the Delta and the Omicron clades – have been identified. The research is in line with the global trends in coronavirus-related research and is focused on studying viral evolution in the current conditions. The results contribute to dissemination of scientific information on etiology, pathogenesis, epidemiology and treatment of coronavirus infection and other dangerous respiratory infections.

Keywords: medical systems, medical cybernetics, genome analysis, pandemic, coronavirus infection, software, data processing automation, data visualization, big data

Введение

Новый коронавирус человека (SARS-CoV-2) погубил жизни более 6 000 000 человек, стал причиной инвалидизации, временной и стойкой потери трудоспособности сотен миллионов людей по всей планете. Учитывая продолжающиеся новые войны, экологические бедствия, в том числе лесные пожары и ураганы, реально ожидать появление новых вирусных пандемических инфекций.

Актуальность исследований в области биоинформатики, биофизики, кибернетики, молекулярной эпидемиологии и прикладной математики вирусов [1] также обусловлена широким распространением вирусов иммунодефицита человека и обезьян, герпеса, цитомегаловирусной инфекции и вирусных гепатитов, обнаружением новых онкогенных вирусов, а также с открытием причастности вирусных частиц к этиологии ряда малоизученных заболеваний и аутоиммунной патологии.

Новая коронавирусная инфекция стала причиной смерти и заболеваний миллионов не только людей, но и животных. Геномика и биоинформатика предоставляют возможность получать, изучать и анализировать геномные тексты микробов, в частности коронавирусов. Тема вирусного филогенеза также исключительно актуальна и сегодня рассматривается на страницах множества современных научных статей и монографий [2, 3].

Вспышка новой коронавирусной инфекции COVID-19 началась в середине декабря 2019 года в Китае, в городе Ухань и распространилась на многие города Китая, Юго-Восточной Азии, а также по всему миру.

Методика проведения эксперимента

В течение интервала времени 2019-2022 гг. нами был 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*) пять раз, в последний раз 11.09.2022 года. Изначально в базе данных находилось 88 геномов, к осени 2022 года – уже 526 геномов.

Как модули разработанного конвейера при помощи технологии платформы Galaxy (рисунок 1), для обработки данных с целью оценки качества и состава генетического материала и идентификации происхождения изолятов мы отобрали программные модули Pangolin и Nextclade, для аннотирования – Prokka [4] и BARRNAP, MAFFT [5] и ClustalW [6] – для множественного выравнивания, которое необходимо для филогенетического анализа.

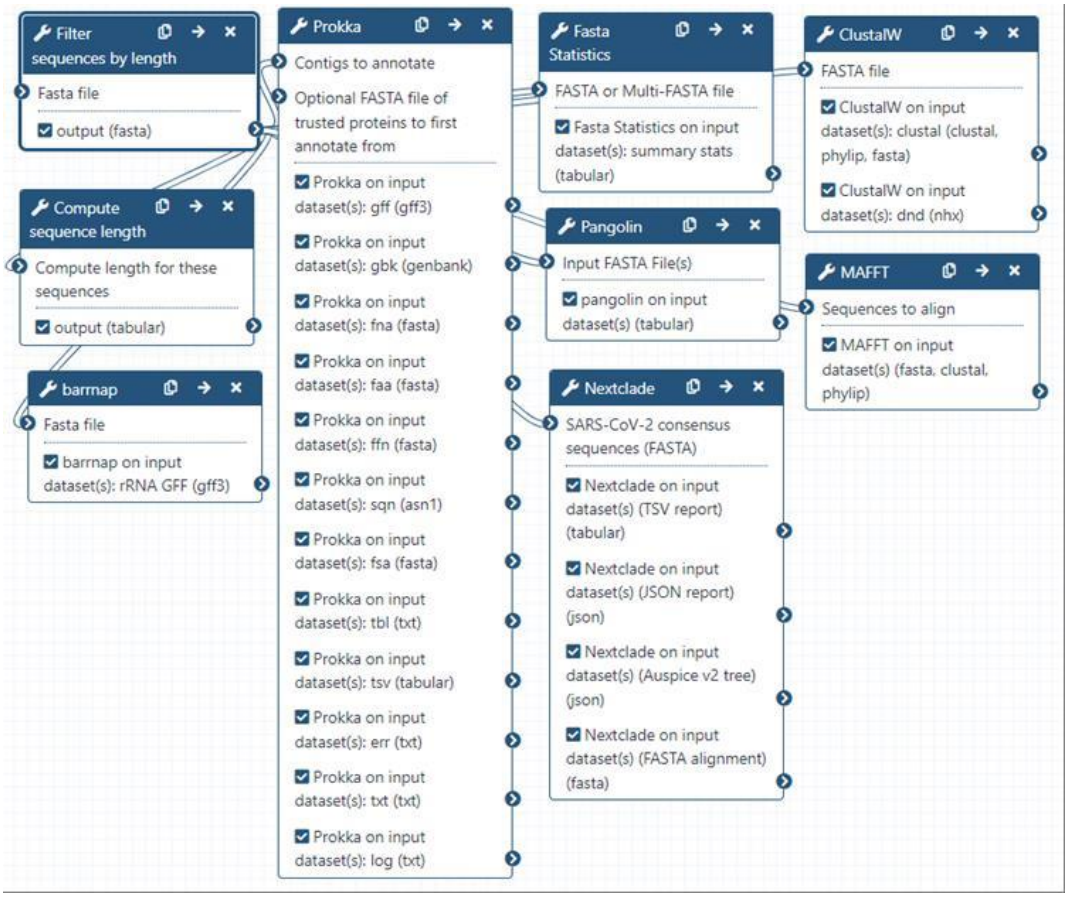

Рис. 1. Разработанный на платформе Galaxy конвейер обработки контигов коронавируса

Pangolin [7, 8] был разработан для реализации динамической номенклатуры линий и кластеров-кладов передачи SARS-CoV-2, известной как номенклатура Pango (рисунки 2-3). Pangolin присваивает распознанную линию и имя кластера по принципу опубликованному A. Rambaut др., 2020 [14].

Nextclade [9] – 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. Руководство пользователя доступно по адресу docs.nextstrain.org/projects/nextclade.

Также нами было изучено и использовано следующее программное обеспечение и веб-сервисы:

1. Программный комплекс для статистических вычислений JMP SAS 7 [10, 11];
2. Веб-сервис CoVizu;
3. Веб-сервис Центр ресурсов для биоинформатики бактерий и вирусов (Bacterial and Viral Bioinformatics Resource Center (BV-BRC);
4. Веб-сервис Nextstrain (рисунок 4).

Основная исходная выборка коронавирусных геномных текстов состояла из 526 полных геномов различного качества и происхождения (несколько образцов происходило от животных: домашних кошек, собак и диких норок), полученных в Беларуси с декабря 2019 по май 2022 года (рисунки 2-4).

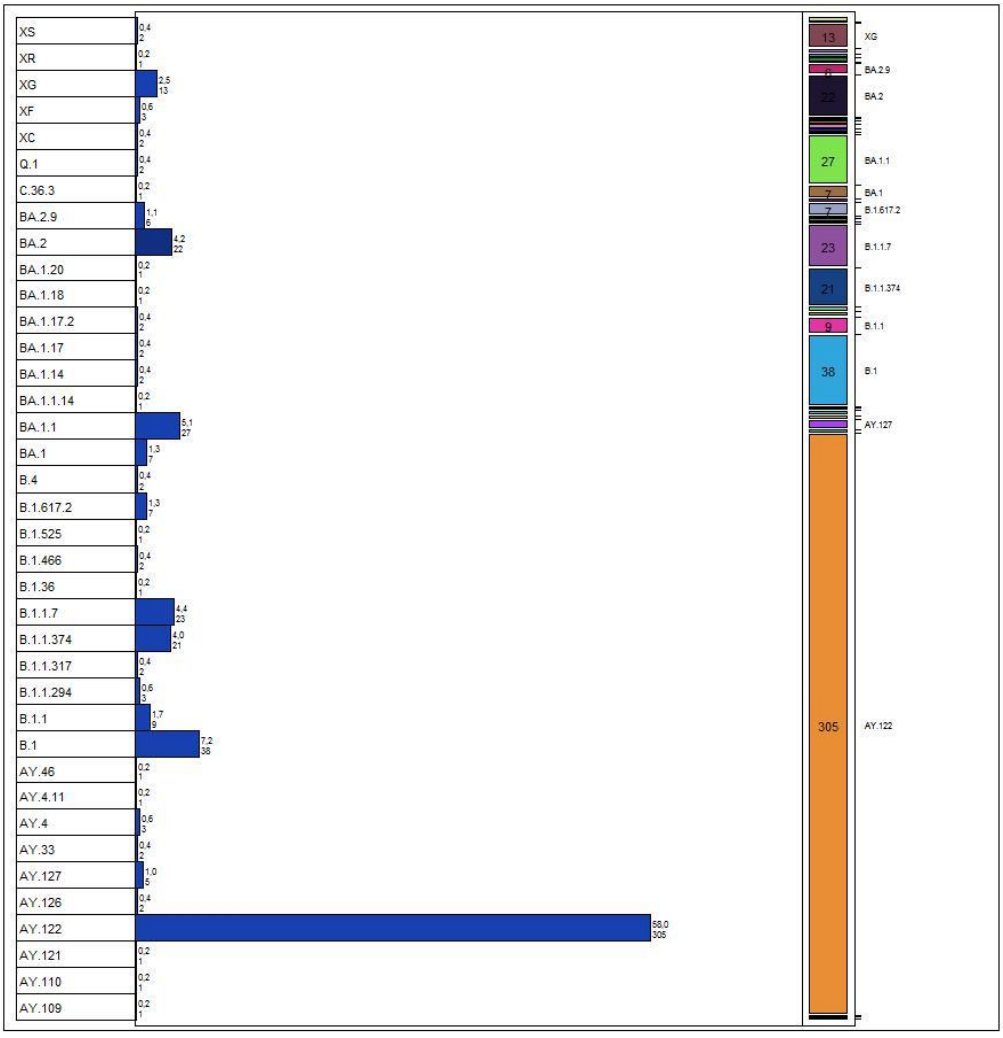

Рис. 2. Столбчатая диаграмма, представляющая геномный профиль 526 образцов белорусского коронавируса. Здесь и далее напротив столбика приводится процентное и числовое количественное представление каждой идентифицированной линий передачи

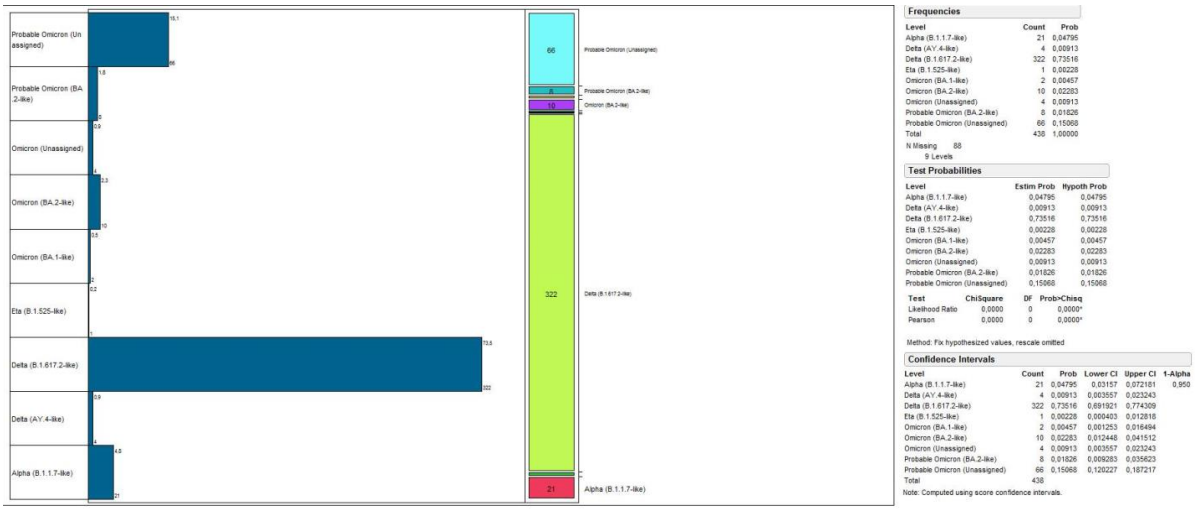

Рис. 3. Столбчатая диаграмма представления результатов классификации кладов геномов SARS-CoV-2, полученных от пациентов на территории Беларуси

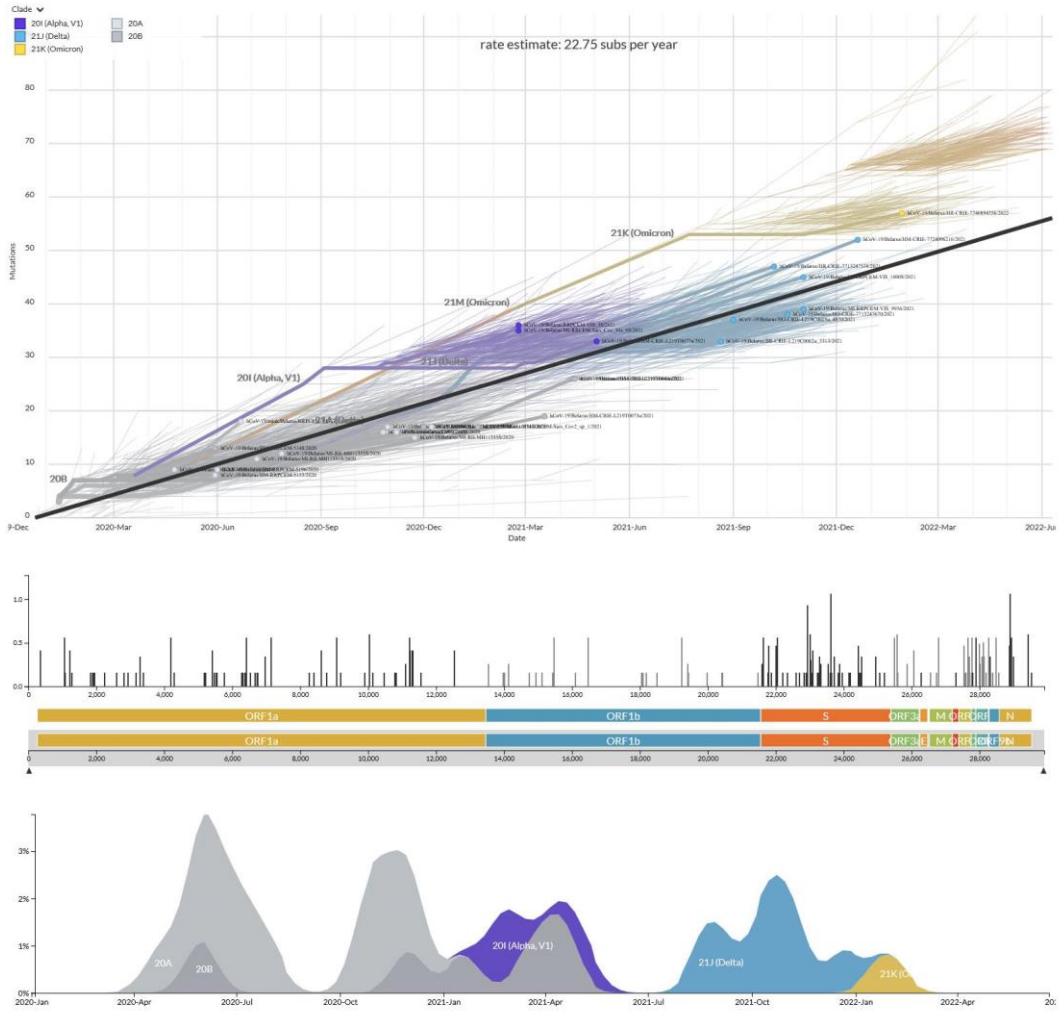

Рис. 4. Диаграмма представления динамического распространения кладов геномов SARS-CoV-2, полученных от пациентов на территории Беларуси

Результаты и их обсуждение

Был разработан и протестирован конвейер для анализа геномов SARS-CoV-2, полученныхот пациентов в Беларуси и (для сравнения) в России. Выполнено геномное профилирования с целью определения и статистического анализа кластеров и линий передачи новой коронавирусной инфекции, в соответствии с предложенными классификациями кластеров и штаммов COVID-19. Также были получены сведения по оценке качества исходных данных, выполнена и графически представлена визуализации полученных результатов. Нами наблюдалась эволюция коронавируса от изначально доминирующего в Беларуси и России клада-кластера B.1 («Базельский кластер») и B.1.1. до Дельта и Омикрон кластеров, коорые преобладают в Беларуси на сегодняшний день (рисунки 2-4). Изначально (2019-2020 гг.) доминирующие в Беларуси клады-кластеры B.1 («Базельский кластер») и B.1.1 имеют европейско-британское распространение, имеются публикации об аналогичному нашему исследованию в Швейцарии, где общая частота распространения B.1 составила 68, 2% [12]. Информации о специфических клинических и лабораторных особенностях этого кластера на сегодняшний день недостаточно. Уже вы 2022 году нами было выявлено повсеместное в Беларуси и России распространение Дельта-штамма и текущая тенденция доминирования Omicron B.1.529. По данным классификации Pango в последней выборке данных значительно доминировал клад Дельта (более 90%), но в настоящее время наблюдается тенденция к доминирования модификаций вариантов линии передачи Омикрон, что согласуется с недавними литературными данными из России [13].

Заключение

Выполненные исследования вписываются в мировые тенденции исследования коронавируса и направлены на изучение вирусной эволюции в современных условиях. Результаты способствуют распространению научной информации об этиологии, патогенезе, эпидемиологии и лечению коронавирусной инфекции и других опасных респираторных инфекций. Разработанный автоматизированный конвейер будет в дальнейшем усовершенствован и его можно будет использовать для изучения просеквенированных приборами и технологиями Illumina, Oxford Nanopore, PacBio, Ion Torrent геномов различных микробов.

Результаты наших исследований по теме геномики коронавируса уже нашли практическое применение в РНПЦ Эпидемиологии и Микробиологии г. Минска при Минздраве Беларуси. Разработанные алгоритмы и автоматизированные конвейеры будут использоваться в качестве практического инструмента для новых научных исследований.

Работа выполнена при поддержке проектов БРФФИ:

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

«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;

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

Список литературы

1. Frishman D., Marz M. Virus Bioinformatics: CRC Press; 2021. 280 p.
2. Nain Z., Rana H.K., Liò P., Islam S.M.S., Summers M.A., Moni M.A. Pathogenetic profiling of COVID-19 and SARS-like viruses. Briefings in Bioinformatics. 2021;22(2):1175-96.
3. Фомченко Н., Воропаев Е. Филогенетическине аспекты молекулярной биологии в медицине (обзор литературы). Проблемы здоровья и экологии. 2014(3 (41)):30-5.
4. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014 Jul 15;30(14):2068-9.

5. Katoh K., Standley D.M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013 Apr;30(4):772-80.
6. Sievers F., Higgins D.G. Clustal Omega, accurate alignment of very large numbers of sequences. Multiple sequence alignment methods: Springer; 2014. p. 105-16.
7. O'Toole Á., Scher E., Underwood A., Jackson B., Hill V., McCrone J.T., Colquhoun R., Ruis C., Abu-Dahab K., Taylor B. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. Virus Evolution. 2021;7(2):veab064.
8. Pipes L., Wang H., Huelsenbeck J.P., Nielsen R. Assessing uncertainty in the rooting of the SARS-CoV-2 phylogeny. Molecular biology and evolution. 2021;38(4):1537-43.
9. Aksamentov I., Roemer C., Hodcroft E.B., Neher R.A. Nextclade: clade assignment, mutation calling and quality control for viral genomes. Journal of Open SourceSoftware. 2021;6(65):3773.
10. Jones B., Sall J. JMP statistical discovery software. Wiley Interdisciplinary Reviews: Computational Statistics. 2011;3(3):188-94.
11. Sall J., Stephens M.L., Lehman A., Loring S. JMP start statistics: a guide to statistics and data analysis using JMP: Sas Institute; 2017.
12. Stange M., Mari A., Roloff T., Seth-Smith H.M.B., Schweitzer M., Brunner M., Leuzinger K., Søgaard K.K., Gensch A., Tschudin-Sutter S., Fuchs S., Bielicki J., Pargger H., Siegemund M., Nickel C.H., Bingisser R., Osthoff M., Bassetti S., Schneider-Sliwa R., Battegay M., Hirsch H.H., Egli A. SARS-CoV-2 outbreak in a tri-national urban area is dominated by a B.1 lineage variant linked to a mass gathering event. PLoS Pathogens. 2021;17(3):e1009374.
13. Антонец Д., Старчевская М., Колосова Н., Суслопаров И., Даниленко А., Боднев С., Швалов А., Трегубчак Т., Рыжиков А., Пьянков О. Предварительный анализ генетической изменчивости изолятов вируса SARS-CoV-2, относящихся к варианту Омикрон, циркулирующих на территории Российской Федерации: COVID-19-PREPRINTS. MICROBE. RU.