Вавиловский журнал генетики и селекции. 2023;27(7):839-850
DOI 10.18699/VJGB-23-97

Перевод на английский язык https://vavilov.elpub.ru/jour

О пространстве вариантов генетических последовательностей SARS-CoV-2

А.Ю. Пальянов[1, 2, 3] ✉, Н.В. Пальянова[2]

[1] Институт систем информатики им. А.П. Ершова Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Научно-исследовательский институт вирусологии, Федеральный исследовательский центр фундаментальной и трансляционной медицины, Новосибирск, Россия
[3] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
✉ palyanov@iis.nsk.su

Аннотация. Пандемия коронавирусной инфекции, вызванная вирусом SARS-CoV-2, которой человечество противостояло с использованием новейших достижений науки, оставила после себя в том числе обширные генетические данные. Ежедневно начиная с конца 2019 г. в мире собирались образцы геномов вируса, что предоставляет возможность детально проследить его эволюцию с момента возникновения до настоящего времени. Накопленная статистика результатов экспресс-тестирования показала, что число подтвержденных случаев заражения SARS-CoV-2 составило не менее 767.5 млн (9.5 % нынешнего населения Земли без учета бессимптомников), а число секвенированных геномов вируса – более 15.7 млн (что составляет чуть более 2 % от общего числа заразившихся). Эти новые данные потенциально несут в себе информацию о механизмах изменчивости и распространения вируса, его взаимодействия с иммунной системой человека, об основных параметрах, характеризующих механизмы развития пандемии, и многое другое. В этой статье мы анализируем пространство возможных вариантов генетических последовательностей SARS-CoV-2 как с математической точки зрения, так и с учетом биологических ограничений, присущих этой системе (основанных на общебиологических знаниях и учитывающих особенности данного конкретного вируса). Для этого мы разработали программное обеспечение, способное загружать и анализировать нуклеотидные последовательности SARS-CoV-2 в формате FASTA, определять позиции 5′ и 3′ UTR, число и расположение неидентифицированных нуклеотидов (“N”), осуществлять выравнивание относительно референсной последовательности посредством вызова предназначенных для этого программ, определять мутации, делеции и вставки, а также рассчитывать различные характеристики геномов вирусов с заданным шагом по времени (дни, недели, месяцы и т. д.). Полученные данные свидетельствуют о том, что, несмотря на кажущееся математическое многообразие возможных вариантов изменения вируса во времени, коридор эволюционной траектории, которым прошел коронавирус, представляется достаточно узким. Это дает основание полагать, что он в некоторой степени детерминирован, что позволяет надеяться на возможность моделирования эволюции коронавируса.
Ключевые слова: коронавирус; SARS-CoV-2; геном; пространство вариантов; эволюция; изменчивость.

Для цитирования: Пальянов А.Ю., Пальянова Н.В. О пространстве вариантов генетических последовательностей SARS-CoV-2. *Вавиловский журнал генетики и селекции*. 2023;27(7):839-850. DOI 10.18699/VJGB-23-97

On the space of SARS-CoV-2 genetic sequence variants

A.Yu. Palyanov[1, 2, 3] ✉, N.V. Palyanova[2]

[1] A.P. Ershov Institute of Informatics Systems of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Research Institute of Virology, Federal Research Center of Fundamental and Translational Medicine of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[3] Novosibirsk State University, Novosibirsk, Russia
✉ palyanov@iis.nsk.su

Abstract. The coronavirus pandemic caused by the SARS-CoV-2 virus, which humanity resisted using the latest advances in science, left behind, among other things, extensive genetic data. Every day since the end of 2019, samples of the virus genomes have been collected around the world, which makes it possible to trace its evolution in detail from its emergence to the present. The accumulated statistics of testing results showed that the number of confirmed cases of SARS-CoV-2 infection was at least 767.5 million (9.5 % of the current world population, excluding asymptomatic people), and the number of sequenced virus genomes is more than 15.7 million (which is over 2 % of the total number of infected people). These new data potentially contain information about the mechanisms of the variability and spread of the virus, its interaction with the human immune system, the main parameters characterizing the mechanisms of the development of a pandemic, and much more. In this article, we analyze the space of possible variants of SARS-CoV-2 genetic sequences both from a mathematical point of view and taking into account the biological limitations inherent in this system, known both from general biological knowledge and from the consideration of the characteristics of this particular virus. We have developed software capable of loading and analyzing

SARS-CoV-2 nucleotide sequences in FASTA format, determining the 5′ and 3′ UTR positions, the number and location of unidentified nucleotides ("N"), performing alignment with the reference sequence by calling the program designed for this, determining mutations, deletions and insertions, as well as calculating various characteristics of virus genomes with a given time step (days, weeks, months, etc.). The data obtained indicate that, despite the apparent mathematical diversity of possible options for changing the virus over time, the corridor of the evolutionary trajectory that the coronavirus has passed through seems to be quite narrow. Thus it can be assumed that it is determined to some extent, which allows us to hope for a possibility of modeling the evolution of the coronavirus.
Key words: coronavirus; SARS-CoV-2; genome; space of variants; evolution; variability.

For citation: Palyanov A.Yu., Palyanova N.V. On the space of SARS-CoV-2 genetic sequence variants. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(7):839-850. DOI 10.18699/VJGB-23-97

Введение

Возможность компьютерного моделирования эволюции, жизненного цикла и размножения простейшего биологического организма с детализацией до генного уровня стала бы научным прорывом, однако это по-прежнему находится далеко за пределами возможностей современных суперкомпьютеров. Процесс естественного отбора наиболее приспособленных особей происходит с учетом огромного количества факторов как внешней, так и внутренней среды. Особенности организма реализуются через наборы особенностей белков, а влияние изменений каждого белка на приспособленность оценить достаточно трудно в связи с необходимостью учитывать все возникающие изменения взаимодействий этого белка со всеми факторами среды и другими белками, число которых весьма значительно.

В компьютерных моделях эволюционирующих объектов, как правило, внесение изменений в геном потомков осуществляется не явным образом (посредством воспроизведения молекулярных механизмов), а лишь имитируется посредством описания алгоритмов внесения изменений в копию генома предков. Однако и сами механизмы внесения мутаций и горизонтального переноса генов являются субъектами эволюции, и среди возможных изменений, не приводящих к гибели или стерильности особи, встречаются и те, что влияют на скорость и точность репликации генома. Благодаря этому возникает внутривидовая конкуренция, в результате которой, например, для SARS-CoV-2 с момента его появления и до настоящего времени длительность инкубационного периода, напрямую связанная со скоростью репликации вируса, постоянно снижается (Malone et al., 2022).

По сравнению с клеточными формами жизни вирусы представляются значительно более удобными объектами для изучения и компьютерного моделирования эволюции благодаря достаточно простому устройству и значительно меньшему геному при широком спектре взаимодействий с внешней средой и организмом хозяина. До появления технологий быстрого секвенирования геномов эволюцию вирусов можно было рассматривать лишь в рамках моделей «паразит–хозяин», описывающих статистические, но не молекулярные особенности их взаимодействия. С начала пандемии SARS-CoV-2 число подтвержденных случаев заражения SARS-CoV-2 составило не менее 767.5 млн (9.5 % нынешнего населения Земли без учета бессимптомников) (Palyanova et al., 2022). Мировым научным сообществом было получено более 15.7 млн вариантов геномов данного коронавируса (включая дату взятия образца и географическое расположение места его получения), предоставляющих беспрецедентно обширные данные о его эволюции, в таком количестве не имеющиеся ни для какого из других вирусов.

На основе этих данных может быть рассчитана динамика распространения и изменения вируса не только в физическом пространстве и времени, но и в многомерном пространстве возможных жизнеспособных вариантов вирусных геномов. Метрика такого пространства определяется минимальным числом единичных изменений (мутация, делеция или вставка), необходимых для преобразования одного генома в другой (расстояние Левенштейна, или «редакционное расстояние»). При этом вирус изменяется, в том числе в ответ на вакцинацию и формирование иммунитета у переболевших. А значит, изменяется как геном вируса, так и его «фенотипические» проявления при взаимодействии с организмом носителя, т. е. одновременно происходят два процесса: изменение (распространение) множества (облака) точек, представляющих популяцию вируса в тот или иной момент времени в пространстве возможных последовательностей РНК, и изменение самого ландшафта этой многомерной поверхности «функции приспособленности» вируса. Каждая точка в пространстве возможных состояний соответствует определенной нуклеотидной последовательности, более или менее отличающейся от исходного референсного генома (с которого все началось в конце 2019 г. (Wu et al., 2020)) некоторым количеством изменений – мутаций, делеций и вставок.

Между парами точек, каждая из которых соответствует жизнеспособной последовательности, если одна из них получилась из другой вследствие изменений, произошедших с вирусом с момента попадания в организм носителя до появления вирионов следующего поколения (как правило, до этого проходит далеко не один цикл репликации генома вируса), могут и должны существовать переходы. Большинство возможных изменений, возникающих при репликации (у каждого экземпляра вирусной последовательности – свои собственные), приведут к его нежизнеспособности (особенно делеции или вставки, длина которых не кратна трем, т. е. такие, что приведут к сдвигу рамки считывания при трансляции). Однако некоторые изменения могут оставить приспособленность вируса на прежнем уровне или даже улучшить ее – например, повысив скорость синтеза новых вирусных частиц или увеличив их количество, производимое в единицу времени (что увеличит их преимущество перед остальными вариантами, находящимися в это же время в организме, т. е. возникает внутривидовая конкуренция). Под функцией

приспособленности можно понимать число экземпляров вирусной последовательности, существующей в человеческой популяции в данный момент времени (с нормировкой на общее число экземпляров вируса в ней или без таковой).

Таким образом, формируется (проявляется) ландшафт «поверхности» (многомерной) функции приспособленности, который может иметь более или менее обширные «долины», соответствующие множеству сходных последовательностей (появившихся в результате небольших изменений варианта, впервые попавшего в эту долину), «горные хребты» или «плато», разграничивающие «долины» (все точки которых соответствуют нежизнеспособным последовательностям), «перевалы», по которым можно перемещаться между «долинами». Области нежизнеспособных последовательностей соответствуют тем из них, которые, к примеру, не могут создавать свои копии из-за повреждения гена, кодирующего РНК-зависимую РНК-полимеразу (RdRp), осуществляющую репликацию вирусной РНК, либо тем, у которых изменения в структуре соответствующих белков не позволяют вирусу сформировать белковую оболочку, а также в силу множества других разнообразных причин. Также, надо полагать, имеются «долины», для которых ни один из принадлежащих им вариантов последовательностей еще не был реализован, однако в которые все-таки возможно попасть – например, в результате возникновения жизнеспособного рекомбинантного штамма, получившегося при сочетании двух более-менее различных вариантов геномов. Возможно, именно этим путем и возник изначальный вариант коронавируса SARS-CoV-2.

В настоящее время существуют две основные базы данных, предоставляющие пользователям онлайн-доступ к генетическим последовательностям SARS-CoV-2. Крупнейшей из них является созданная в 2006 г. GISAID (Global Initiative on Sharing All Influenza Data – глобальная инициатива по обмену всеми данными о гриппе, https://gisaid.org) (Khare et al., 2021), которая с момента появления коронавируса SARS-CoV-2 в конце 2019 г. стала также репозиторием для накопления секвенированных вариантов этого вируса, полученных лабораториями по всему миру. В июле 2023 г. в ней насчитывалось более 15.7 млн записей. Вторая база данных, NCBI SARS-CoV-2 Data Hub (Sayers et al., 2022) (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?VirusLineage_ss=taxid:2697049), содержит более 7.7 млн образцов геномов SARS-CoV-2. Столь беспрецедентно обширные и детальные данные прежде не были доступны человечеству, поэтому необходимо извлечь как можно больше полезной информации и знаний на основе их всестороннего анализа. Наша работа представляет собой лишь первые шаги на этом пути, и многое еще предстоит сделать.

Высокой значимостью для научного сообщества исследователей вирусных геномов обладает также проект Nextstrain/Nextclade (https://clades.nextstrain.org/) (Aksamentov et al., 2021), предоставляющий онлайн-инструменты для анализа и визуализации генетических данных по различным вирусам, включая SARS-CoV-2. Функциональные возможности Nextstrain выгодно выделяются на общем фоне и включают в том числе графическое отображение карты генома загруженных последовательностей с изображением мутаций, делеций, вставок, неопределенных нуклеотидов (“N”) и ряда других особенностей каждой последовательности, например принадлежность к классу реассортантных (рекомбинантных) вариантов.

Описание пространства вариантов генетических последовательностей SARS-CoV-2 принципиально включает в себя: те, которые мы уже можем наблюдать и изучать благодаря обширному секвенированию; варианты из реального пространства вариантов, которые уже реализовались, однако не попали в поле зрения исследователей; остальные возможные варианты, которые могли бы реализоваться в будущем и представляют особый интерес, поскольку являются потенциально опасными для человечества и было бы хорошо заранее быть готовыми к их возможному появлению (экспресс-тесты для их выявления, вакцины и т. д.).

Рассмотрим теперь наиболее важные характеристики SARS-CoV-2 как системы, основой существования которой является саморепликация в клетках носителя, и которые будут важны в будущем при создании его эволюционного симулятора. Они включают скорость репликации генома (600–700 нт/с, самая большая среди известных скоростей работы вирусных РНК-полимераз) (Shannon et al., 2020), время репликации вирусной РНК ($\frac{3 \cdot 10^4 \text{ нт}}{600 \text{ нт/с}} = 50$ с), время воспроизводства вируса целиком (7–24 ч) (Grebennikov et al., 2021) и частоту возникновения ошибок ($1.3 \cdot 10^{-6} \pm 0.2 \cdot 10^{-6}$ на позицию за инфекционный цикл заражения клетки, т. е. от входа вируса в нее до выхода новых вирионов наружу) (Amicone et al., 2022). При этом скорость эволюционных изменений в геноме SARS-CoV-2 оценивается как как $8.9 \cdot 10^{-4}$ замен в год в каждой позиции (Sonnleitner et al., 2022), что могло бы привести в среднем к 93 заменам за 3.5 года. Это хорошо соотносится с тем, что один из наиболее далеких от референсной последовательности вариантов, относящийся к линии «Омикрон», полученный 20.06.2023, имеет 103 замены (максимальное число мутаций среди вариантов, см. таблицу). У вариантов «Альфа» и «Бета» отличия от первоначальной референсной последовательности составляют более 30 точечных мутаций и более 17 делеций. У вариантов, возникших позднее, отличий больше. Также заметно, что в процессе эволюции вируса увеличивается число делеций, достигая 59 шт. в одной из современных ветвей «Омикрона».

Как уже было сказано, коронавирус SARS-CoV-2 обладает самой быстрой РНК-полимеразой. При этом она имеет еще и один из самых низких для РНК-вирусов показателей частоты возникновения мутаций в процессе репликации, что необходимо ввиду его достаточно большого генома. Это достигается благодаря экзонуклеазе (nsp14-ExoN), корректирующей ошибки, которая встречается только у вирусов с большими геномами (коронавирусов и торовирусов) (Campagnola et al., 2022).

Важными параметрами являются также минимальная инфекционная доза (количество вирионов, необходимое для заражения), составляющая около 100 частиц (Karimzadeh et al., 2021), репродуктивное число (1.8–3.2) (Xu et al., 2021), количество вирусных частиц, которые переносит больной во время пика инфекции ($(1–100) \cdot 10^9$ шт.) и

Наиболее поздние представители различных ветвей филогенетического дерева коронавируса SARS-CoV-2 (https://nextstrain.org/ncov/open/global/all-time)

Имя вируса	Дата получения образца	Идентификатор последовательности в БД	Код классификатора Pangolin	Клада, вариант	Число мутаций	Число делеций	Длина генома
hCoV-19/Wuhan/ WIV04/2019 (референсный геном в GISAID)	30.12.2019	EPI_ISL_402124	B	19A	0	0	29891
Wuhan-Hu-1 (референсный геном в Genbank)	12.2019	NC_045512.2	B	19A	0	0	29903
hCoV-19/Tunisia/ S-1180/2021	29.10.2021	EPI_ISL_11333927	B.1.1.7	20I (Alpha, V1)	37	19	29758
hCoV-19/Madagascar/LA2M-112753/2021	16.01.2021	EPI_ISL_7722749	B.1.351.2	20H (Beta, V2)	31	18	29818
PHL/COVID-74517/2021	01.07.2021	OL629469	B.1.351	20H (Beta, V2)	32	9	29854
hCoV-19/Brazil/AM-IMTSP-CD24003/2021	10.08.2021	EPI_ISL_14800432	P.1.4	20J (Gamma, V3)	42	9	29772
LAO/LOMWRU-0461/2021	24.11.2021	OQ028273	P.1	20J (Gamma, V3)	32	18	29699
hCoV-19/Australia/WA11930/2023	28.02.2023	EPI_ISL_17187319	XBC.1.4	21I (Delta) XBC	77	36	29308
hCoV-19/Yunnan/ YNCDC-1019/2023	23.05.2023	EPI_ISL_17778593	DY.1	22B (Omicron)	89	59	29806
hCoV-19/Japan/TKYmbc38047/2023	06.06.2023	EPI_ISL_17941095	XBB.2.3.11	22F (XBB.2.3)	99	56	29726
hCoV-19/Heilong-jiang/HLJCDC-1665/2023	20.06.2023	EPI_ISL_17850574	XBB.1.5	23A (Omicron) (XBB.1.5)	103	56	29781

Примечание. Представители некоторых ветвей (в основном различных вариантов «Омикрона») продолжают встречаться среди секвенированных последовательностей недавно заболевших людей, а некоторые перестали обнаруживаться вовсе («Альфа», «Бета», «Гамма», «Дельта» и др.). Референсные последовательности в обеих базах различаются только длиной поли-А участка, расположенного в самом в конце, а в остальном совпадают.

число вирионов, в среднем содержащихся в зараженной клетке (10^5 шт.) (Sender et al., 2021), а также другие эпидемиологические характеристики. Вирусные частицы обнаруживаются во многих тканях и органах, от легких до мозга, однако выйдут наружу и могут быть переданы следующим носителям только те, что присутствуют в дыхательных путях или кишечнике. Все остальные вирионы не оставят «потомков», что ощутимо сужает эволюционный коридор. В работах (Day et al., 2020; Markov et al., 2023) рассмотрен ряд важных вопросов, касающихся эпидемиологии и эволюции вируса SARS-CoV-2, включая механизм возникновения рекомбинантных штаммов.

Материалы и методы

Наиболее рациональным с точки зрения как скорости обработки данных, так и обеспечения ничем не ограниченных возможностей (которые при необходимости можно расширять) при их анализе, на наш взгляд, является работа с исходными fasta-файлами с помощью программного комплекса, сочетающего наши собственные разработки со сторонними библиотеками и программами. К настоящему времени реализован прототип, включающий минимально необходимые функциональные возможности. Для разработки использовался язык программирования C++, среда разработки – Microsoft Visual Studio Community 2019. Аппаратное обеспечение – ПК на базе процессора Intel Core i7-10700K, 3.8 ГГц, 8 ядер, 16 Гб оперативной памяти.

Использованные в нашей работе методы в основном могут быть отнесены к следующим двум категориям:

– теоретические оценки и численные расчеты некоторых важных характеристик рассматриваемой системы;
– анализ доступных генетических данных с помощью собственного прикладного программного обеспечения и с помощью существующих программных средств.

Полногеномные генетические последовательности SARS-CoV-2. Базы данных GISAID и Genbank предоставляют посредством веб-интерфейса некоторый набор функциональных возможностей для изучения свойств содержащихся в них последовательностей, однако они недостаточно гибкие для осуществления анализа, который необходим для исследований пространства генетических вариантов SARS-CoV-2, являющихся целью настоящей работы. Для GISAID также существует API (Application Programming Interface – программный интерфейс приложения) (Wirth, Duchene, 2022), реализованный на языке R, однако и для этого способа имеются существенные ограничения (включая скорость работы при значительных объемах обрабатываемой информации) по сравнению с прямым доступом к генетическим последовательностям, хранящимся в виде fasta-файлов на локальной рабочей станции. GISAID существенно ограничивает закачку со своего сайта: за одну загрузку система позволяет скачать не более 2000 последовательностей, что полностью исключает возможность загрузки значимого объема данных «вручную». В NCBI SARS-CoV-2 Data Hub подобных ограничений нет.

Для анализа уже реализовавшихся генетических вариантов SARS-CoV-2 были использованы полногеномные последовательности из баз данных GISAID (https://gisaid.org/) и NCBI Virus SARS-CoV-2 Data Hub (https://www.ncbi.nlm.nih.gov/labs/virus/). Последовательности из Genbank за 2019–2020 гг. были скачаны на локальную вычислительную станцию и проанализированы с помощью

разработанного нами программного обеспечения ParSeq. Последовательности из GISAID, ввиду ограничений на скачивание, мы не загружали, воспользовавшись вместо этого доступом по API для получения лишь некоторых их характеристик (например, полной длины последовательности; при этом возможность определения длины транслируемой части или позиций ее начала и конца не поддерживается).

Для расчета редакционного расстояния между парами коронавирусных последовательностей, включая отдельный расчет количества мутаций, делеций и вставок, использовался веб-ресурс Nextstrain/Nextclade (https://clades.nextstrain.org).

Результаты

Оценка количества реализовавшихся и потенциальных генетических вариантов SARS-CoV-2

Начнем с рассмотрения пространства генетических последовательностей как такового с математической точки зрения в самом общем случае. Любая пара последовательностей характеризуется мерой различия между ними, называемой расстоянием Левенштейна, или редакционным расстоянием – минимальным количеством точечных (одиночных) замен (мутаций, делеций, вставок), которые необходимо сделать в первой последовательности, чтобы преобразовать ее во вторую. Каждый элемент множества последовательностей заданной длины L будет удален от пустой последовательности (Ø) ровно на L. Число вариантов нуклеотидных последовательностей длиной L составляет 4^L. Число возможных точечных мутаций для последовательности длиной L равно $3{\cdot}L$ (нуклеотид в каждой позиции может быть заменен на любой из трех других). Также возможно $3{\cdot}L$ одиночных делеций и $3{\cdot}(L{+}1)$ одиночных вставок. В результате всех возможных одиночных делеций для всех возможных последовательностей длиной L получается множество всех возможных последовательностей длиной $(L{-}1)$, с числом вариантов, равным $4^{(L-1)}$. А в результате всех возможных одиночных вставок – множество всех возможных последовательностей длиной $(L{+}1)$, с числом вариантов $4^{(L+1)}$.

Рассмотрим все возможные варианты нуклеотидных последовательностей длиной $L = 2$ (рис. 1). Последовательности длиной 2 нуклеотида – простой случай, однако даже для него уже необходим гиперкуб в четырехмерном пространстве (тессеракт с 16 вершинами). Для более сложного случая, $L = 4$, аналогичным образом может быть использован 6-мерный гиперкуб (гексеракт) с 64 вершинами, изображение которого вместе с подписями последовательностей и ребер будет перенасыщено деталями и затруднительно для восприятия. Однако его можно в некоторой степени отобразить на двумерной плоскости с помощью одного из вариантов кодов Грея (Mütze, 2023), теория которых тесно связана с гиперкубами, а именно 2D кода, который нам удалось подобрать для данного случая (рис. 2).

Привычная метрика, определяемая как корень из суммы квадратов разностей декартовых координат, в этом случае, как видно, не подходит.

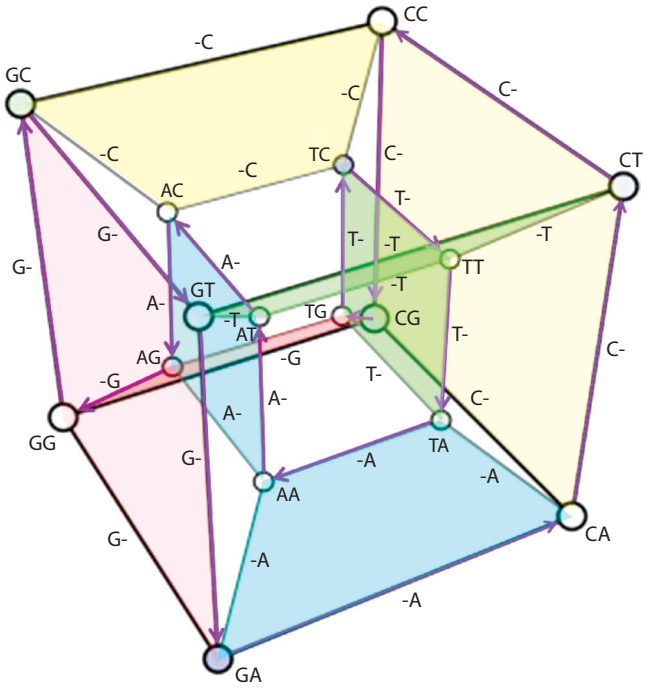

Рис. 1. Пространство вариантов нуклеотидных последовательностей длиной 2, представленное в виде гиперкуба.

Показан один из множества гамильтоновых циклов на гиперкубе (фиолетовые стрелки) – замкнутый путь, проходящий через каждую вершину ровно один раз. Каждый переход соответствует единичному изменению (мутации, делеции или вставке). Также имеются гиперплоскости, которые можно сопоставить подпоследовательностям меньшей длины, получающимся в данном случае, при $L = 2$, посредством делеций слева (-A,-T,-G,-C) и справа (A-,T-,G-,C-), которые для этого простого случая получаются одними и теми же.

Число всех возможных последовательностей такой же длины, как и длина референсного генома SARS-CoV-2, $L = 29\,903$, составляет астрономическое число $4^{29\,903}$, или $\approx 2.511{\cdot}10^{18\,003}$. В этом пространстве вариантов множество последовательностей, соответствующих реализованным вариантам генома SARS-CoV-2, составляет лишь малую часть – точку, соответствующую исходной референсной последовательности, и ее небольшую окрестность, ограниченную на данный момент расстоянием от референсной последовательности до наиболее современного штамма «Омикрона». Можно оценить количество возможных вариантов последовательностей в пределах этой дистанции. Для референсной последовательности с $L = 29\,903$ число ее различных вариаций с одной одиночной мутацией равно $3{\cdot}L$, с двумя мутациями – $(3{\cdot}L)^2 - 3{\cdot}L = 3{\cdot}L{\cdot}(3{\cdot}L{-}1)$ (вычитаем из всех возможных вариантов все те случаи, когда вторая мутация произойдет в той же позиции, что и первая, и тогда получится одна из уже существующих последовательностей – референсная или отличающаяся от нее на 1). Аналогично для третьей мутации: $(3{\cdot}L)^3 - ((3{\cdot}L)^2 - 3{\cdot}L)$, и так далее. Для $L = 29\,903$ получаем, что количество всех вариантов последовательностей с числом мутаций от 0 до n (относительно референсной последовательности) для $n = 103$ составляет $1.387{\cdot}10^{510}$, аналогично для $L = 29\,847$ (56 делеций) – $1.108{\cdot}10^{510}$. Суммируя по всем длинам от 29 903 до 29 847, получаем $7.190{\cdot}10^{511}$.

A = 00	T = 01	G = 10	C = 11		AA	GA	CA	TA	TT	AT	GT	CT	CC	TC	AC	GC	GG	CG	TG	AG
1	0	0	0	0	AA	AAAA	GAAA	CAAA	TAAA	TTAA	ATAA	GTAA	CTAA	CCAA	TCAA	ACAA	GCAA	GGAA	CGAA	TGAA	AGAA
2	1	0	0	0	GA	AAGA	GAGA	CAGA	TAGA	TTGA	ATGA	GTGA	CTGA	CCGA	TCGA	ACGA	GCGA	GGGA	CGGA	TGGA	AGGA
3	1	1	0	0	CA	AACA	GACA	CACA	TACA	TTCA	ATCA	GTCA	CTCA	CCCA	TCCA	ACCA	GCCA	GGCA	CGCA	TGCA	AGCA
4	0	1	0	0	TA	AATA	GATA	CATA	TATA	TTTA	ATTA	GTTA	CTTA	CCTA	TCTA	ACTA	GCTA	GGTA	CGTA	TGTA	AGTA
5	0	1	0	1	TT	AATT	GATT	CATT	TATT	TTTT	ATTT	GTTT	CTTT	CCTT	TCTT	ACTT	GCTT	GGTT	CGTT	TGTT	AGTT
6	0	0	0	1	AT	AAAT	GAAT	CAAT	TAAT	TTAT	ATAT	GTAT	CTAT	CCAT	TCAT	ACAT	GCAT	GGAT	CGAT	TGAT	AGAT
7	1	0	0	1	GT	AAGT	GAGT	CAGT	TAGT	TTGT	ATGT	GTGT	CTGT	CCGT	TCGT	ACGT	GCGT	GGGT	CGGT	TGGT	AGGT
8	1	1	0	1	CT	AACT	GACT	CACT	TACT	TTCT	ATCT	GTCT	CTCT	CCCT	TCCT	ACCT	GCCT	GGCT	CGCT	TGCT	AGCT
9	1	1	1	1	CC	AACC	GACC	CACC	TACC	TTCC	ATCC	GTCC	CTCC	CCCC	TCCC	ACCC	GCCC	GGCC	CGCC	TGCC	AGCC
10	0	1	1	1	TC	AATC	GATC	CATC	TATC	TTTC	ATTC	GTTC	CTTC	CCTC	TCTC	ACTC	GCTC	GGTC	CGTC	TGTC	AGTC
11	0	0	1	1	AC	AAAC	GAAC	CAAC	TAAC	TTAC	ATAC	GTAC	CTAC	CCAC	TCAC	ACAC	GCAC	GGAC	CGAC	TGAC	AGAC
12	1	0	1	1	GC	AAGC	GAGC	CAGC	TAGC	TTGC	ATGC	GTGC	CTGC	CCGC	TCGC	ACGC	GCGC	GGGC	CGGC	TGGC	AGGC
13	1	0	1	0	GG	AAGG	GAGG	CAGG	TAGG	TTGG	ATGG	GTGG	CTGG	CCGG	TCGG	ACGG	GCGG	GGGG	CGGG	TGGG	AGGG
14	1	1	1	0	CG	AACG	GACG	CACG	TACG	TTCG	ATCG	GTCG	CTCG	CCCG	TCCG	ACCG	GCCG	GGCG	CGCG	TGCG	AGCG
15	0	1	1	0	TG	AATG	GATG	CATG	TATG	TTTG	ATTG	GTTG	CTTG	CCTG	TCTG	ACTG	GCTG	GGTG	CGTG	TGTG	AGTG
16	0	0	1	0	AG	AAAG	GAAG	CAAG	TAAG	TTAG	ATAG	GTAG	CTAG	CCAG	TCAG	ACAG	GCAG	GGAG	CGAG	TGAG	AGAG

Рис. 2. Множество вариантов нуклеотидных последовательностей длиной 4, изображенное на плоскости с использованием 2D кодов Грея.

Верхний край таблицы стыкуется с нижним, а левый – с правым, т. е. можно отобразить это множество на поверхность тора. Тогда при движении как по горизонтали, так и по вертикали (в системе координат таблицы), в соответствии со свойствами кодов Грея, каждая пара соседних последовательностей будет отличаться ровно на одну точечную замену.

Последовательности с синонимичными однонуклеотидными мутациями, не приводящими к замене аминокислоты, тоже являются частью полного пространства вариантов. Однако реальное число вариантов в контексте рассмотрения структуры и функций белков, транслируемых с вирусной РНК, существенно меньше из-за вырожденности генетического кода (20 аминокислот кодируются 61 триплетом РНК, т. е. в среднем имеем 3.05 триплета, кодирующего одну и ту же аминокислоту). Также учтем, что белки кодирует не весь геном SARS-CoV-2, 771 из 29903 нт является некодирующим. В результате зависимость, пропорциональная $(3L)^n$, преобразуется в $\approx((L-771)+(3\cdot771))^n$, и, таким образом, скорректированное число вариантов белковых последовательностей может быть оценено как $1.02\cdot10^{465}$. Если же предположить, что когда-нибудь число мутаций превысит вышеупомянутые 103 шт. в 10–11 раз, то последовательность, скорее всего, все еще будет коронавирусной, но уже будет принадлежать другому виду. К примеру, коронавирус летучих мышей RaTG13, ближайший сосед SARS-CoV-2 в пространстве вариантов генетических последовательностей, отличается от него на 1135 точечных мутаций.

Попробуем взглянуть на множество испробованных природой вариантов с биологической точки зрения. Вирус попадает в организм (как правило, воздушно-капельным путем), оказывается в легких и проникает в клетку, где рибосома носителя начинает синтезировать вирусные белки в соответствии с последовательностью нуклеотидов в геноме SARS-CoV-2. Среди этих белков – вирусная РНК-полимераза (RdRp), которая начинает репликацию вирусной РНК. Поначалу, когда в клетке находится одна вирусная РНК и одна RdRp, вероятность их встречи чрезвычайно мала, но затем, по мере накопления тех и других молекул в клетке, она начинает стремительно расти. В итоге концентрация достигает уровня, достаточного для осуществления сборки новых вирионов, и когда их количество в клетке достигает примерно 10^5 шт., клетка разрушается, и эти вирионы начинают заражать как соседние клетки, так и все прочие – если часть вирионов попадет в кровоток и будет разнесена по организму. Учитывая, что количество вирусных частиц, переносимых больным во время пика инфекции, может достигать 10^{11} шт. (Sender et al., 2021), разделим это значение на среднее число вирионов в зараженной клетке и получим 10^6 зараженных клеток в организме. У человека примерно $3\cdot10^{13}$ клеток, т. е. заражено оказывается менее 10^{-4} %.

Частота возникновения ошибок при репликации SARS-CoV-2 составляет, согласно (Amicone et al., 2022), $1.3\cdot10^{-6}\pm0.2\cdot10^{-6}$ замен на позицию за один инфекционный цикл заражения клетки, а по другим данным – $(1–2)\cdot10^{-6}$ на позицию (за цикл репликации) (Markov et al., 2023), т. е. в среднем примерно $1.4\cdot10^{-6}$. С учетом длины последовательности получаем вероятность возникновения одной мутации на всю последовательность за цикл репликации ≈0.04. Даже если все зараженные клетки в организме в какой-то момент будут содержать один и тот же вариант вирусной РНК, то спустя один цикл репликации в организме могут оказаться все возможные варианты одиночных замен (3·29903 шт.) относительно исходной вирусной РНК (до начала этого цикла). Таких будет около 4 %, большинство из которых нежизнеспособны, а 96 % окажутся точными копиями реплицированной последовательности. Какова при этом вероятность возникновения жизнеспособной несинонимичной мутации (изменяющей не только последовательность РНК вируса, но и аминокислотную последовательность одного из его белков), к тому же превосходящей предшественника по приспособленности? Этот вопрос остается открытым, однако искомая вероятность определенно будет весьма незначительной. В подавляющем большинстве случаев

все экземпляры вируса, распространяемые заболевшим во внешнюю среду, являются идентичными, и лишь изредка в одном организме встречаются одновременно два варианта. Каким же образом новые мутантные варианты не просто появляются, но и достаточно быстро вытесняют своих предшественников в масштабах планеты?

Учитывая, что соотношение 4 % : 96 % с каждым последующим циклом репликации будет изменяться в сторону уменьшения доли мутантных последовательностей («эффект основателя» (Ruan et al., 2020)) до полного их вытеснения, возможными сценариями возникновения и распространения мутантных вариантов SARS-CoV-2 представляются следующие довольно маловероятные события.

(а) Иммунитета от SARS-CoV-2 у организма нет, он с ним еще не встречался. В клетку попал единственный экземпляр вирусной РНК, при первом же цикле репликации в нем возникла мутация, и она оказалась жизнеспособной (такое может произойти при заражении коронавирусом хотя и с малой, но ненулевой вероятностью). Тогда все новые вирионы, синтезированные этой клеткой, будут носителями данной мутации, и если она заметно увеличивает их приспособленность, то есть шансы, что в итоге они вытеснят исходный вариант.

(б) У организма уже есть иммунитет от SARS-CoV-2. В него попадает одновременно два варианта вирионов SARS-CoV-2 – доминирующий в популяции и новый, мутантный (возникший по механизму из (а) или рекомбинант). Иммунная система уничтожает знакомый ей «старый» вариант, а новый остается незамеченным, и в результате размножается и передается дальше именно он.

Вероятности возникновения двух этих вариантов еще предстоит оценить, однако и без того видно, что коридор возможных вариантов, по которому прошла эволюция, оказался достаточно узким. Противоположность этой картине представляет, например, вирус гриппа, отличительной особенностью и основой выживания которого является высокая изменчивость, обусловленная механизмами антигенного дрейфа и антигенного сдвига (Kim et al., 2018).

Мы оцениваем моделирование эволюции SARS-CoV-2 как возможное, поскольку, несмотря на большое количество вариантов, которые уже должны были реализоваться и которые могли бы реализоваться с точки зрения математики (теории вероятности) и биологии, в реальности была реализована лишь малая их часть, и мы наблюдаем лишь малую часть возможного пространства вариантов.

Разработка программного обеспечения ParSeq

Для анализа генетических последовательностей SARS-CoV-2 нами было разработано программное обеспечение, названное ParSeq (**Par**ser of **Seq**uences) – парсер и анализатор нуклеотидных последовательностей в формате FASTA, которую мы также использовали при анализе последовательностей SARS-CoV-2 в регионах Сибирского федерального округа (Palyanova et al., 2023). Ниже описаны его основные, уже реализованные на данный момент функциональные возможности.

- Загрузка списка fasta-файлов для анализа и последовательное чтение каждого из них, включая разбор заголовка (содержащего поля Accession ID, Length, Pangolin, Nuc. Completeness, Collection Date, Geo Location, Country) и загрузку нуклеотидной последовательности.
- Первичный анализ нуклеотидной последовательности, включая расчет ее длины, содержания нуклеотидов A, U(T), G, C и неидентифицированных нуклеотидов, обозначаемых буквой “N”. Также в некоторых последовательностях иногда встречаются следующие буквы расширенного алфавита (https://www.bioinformatics.org/sms/iupac.html):

R	Y	S	W	K
A ‖ G	C ‖ T	G ‖ C	A ‖ T	G ‖ T
M	B	D	H	Y
A ‖ C	C ‖ G ‖ T	A ‖ G ‖ T	A ‖ C ‖ T	A ‖ C ‖ G

- Определение позиций начала и конца кодирующей части последовательности. В случае референсной последовательности ее полная длина составляет 29903 нт, длина некодирующего 5′ UTR участка – 265 нт, некодирующего 3′ UTR участка – 229 нт. Для этого используется следующий довольно очевидный алгоритм: в случае 5′ UTR движемся вдоль последовательности от ее начала до 500-го нуклеотида (для удобства выбрано «круглое» значение, при котором 265 находится примерно посередине) окном длиной 17 и считаем число совпадений нуклеотидов в этом окне с фрагментом референсной последовательности той же длины из интервала 266–282 (266 – позиция старта трансляции в референсном геноме). Если из 17 совпадают 14 и более, то считаем, что позиция определена (значения подобраны как достаточные для корректной работы в подавляющем большинстве случаев при малой длине окна, чтобы избежать лишних вычислений). В случае 3′ UTR аналогично: движемся окном длиной 17 от позиции L–500 до конца анализируемой последовательности, сравнивая его содержимое с 17 нуклеотидами, которыми оканчивается кодирующий участок референсной последовательности. Критерий тот же – 14 или более совпадений.
- Расчет длин некодирующих 5′ UTR и 3′ UTR, а также находящегося между ними кодирующего участка, составляющего подавляющую часть генома вирусной последовательности (98.35 % его длины в случае референсной последовательности).
- Расчет распределений этих значений для любой выборки последовательностей геномов SARS-CoV-2 (например, в пределах указанного интервала времени для даты получения образца, который был секвенирован, или содержащих не более заданного числа NNN, или и то и другое одновременно, и т. п.).
- Расчет числа вариантов последовательностей в пределах интервала длин, присутствующих в базах данных.

Результаты, полученные с помощью ParSeq

С помощью разработанного нами программного обеспечения был осуществлен анализ нуклеотидных последовательностей SARS-CoV-2, доступных пользователям по

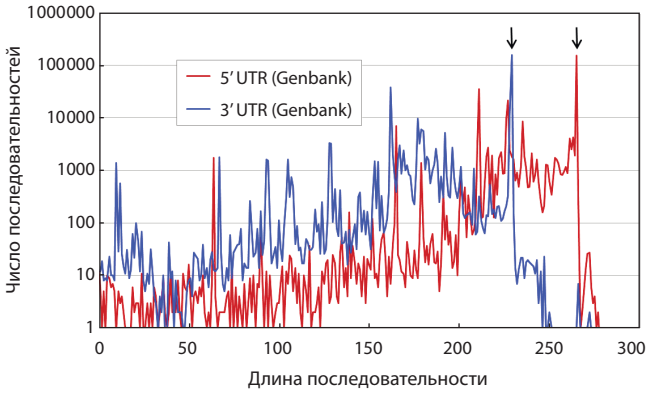

Рис. 3. Распределение длин 5′ UTR и 3′ UTR для последовательностей из базы данных Genbank за период с момента появления SARS-CoV-2 в конце 2019 г. до конца 2020 г.

Длины 5′ UTR и 3′ UTR в референсном геноме SARS-CoV-2 составляют 265 и 229 нт соответственно. Пиковые значения обеих кривых соответствуют именно этим длинам.

всему миру благодаря проектам Genbank и GISAID, их базам данных и веб-ресурсам. В результате были установлены следующие факты.

1. Расчет распределения генетических последовательностей по их полным длинам (5′ UTR + кодирующая последовательность + 3′ UTR) среди последовательностей, имеющих длину ≥ 28000, выявил, что для данных из Genbank (за период с 01.12.2019 по 31.12.2022) минимальное значение длины полной последовательности составило 28784, а максимальное – 29985. Распределение практически полностью расположено левее длины референсной последовательности, составляющей 29903. Таким образом, разница между референсным и минимальным значением длины составила 1119. Скорее всего, такие слишком короткие или слишком длинные последовательности соответствуют данным низкого качества, с ошибками сборки генома, поскольку они плохо соотносятся с данными таблицы, согласно которым максимальная разница между длиной референсной и какой-либо другой последовательности составляет около 159 (103 мутации + 56 делеций). Более того, при таком различии эта последовательность, скорее всего, принадлежала бы уже другому виду вирусов, так как похожую разницу имеют референсная последовательность SARS-CoV-2 и коронавирус летучих мышей RaTG13 (GenBank MN996532.2, collection_date=24-Jul-2013). По данным (Li et al., 2023), они отличаются на 96.2 %, т. е. на 1136 одиночных мутаций (достаточно равномерно рассредоточенных по последовательности). Расчет расстояния между этими же последовательностями, произведенный с помощью веб-сервиса Nextstrain, показал разницу в 1135 одиночных мутаций, а также 20 делеций (в кодирующей части RaTG13 относительно референсной последовательности SARS-CoV-2). Полная длина генома RaTG13 составляет 29855, т. е. число делеций относительно SARS-CoV-2 составляет 48.

Поскольку разница между полной длиной референсного генома SARS-CoV-2 и остальными содержащимися в базе данных последовательностями для некоторых из них существенно превышает число различий (точечных мутаций, делеций и вставок) между референсным геномом SARS-CoV-2 и наиболее отличным от него современным вариантом «Омикрон» (см. таблицу, последняя строка), мы решили исследовать распределение как полных длин геномов, так и их кодирующих и некодирующих участков (рис. 3 и 4). Как видно на рис. 3, участки 5′ UTR и 3′ UTR, встречающиеся в базах данных, обладают длинами от нуля до референсных значений, а также в малом количестве случаев незначительно превосходят их. Последователь-

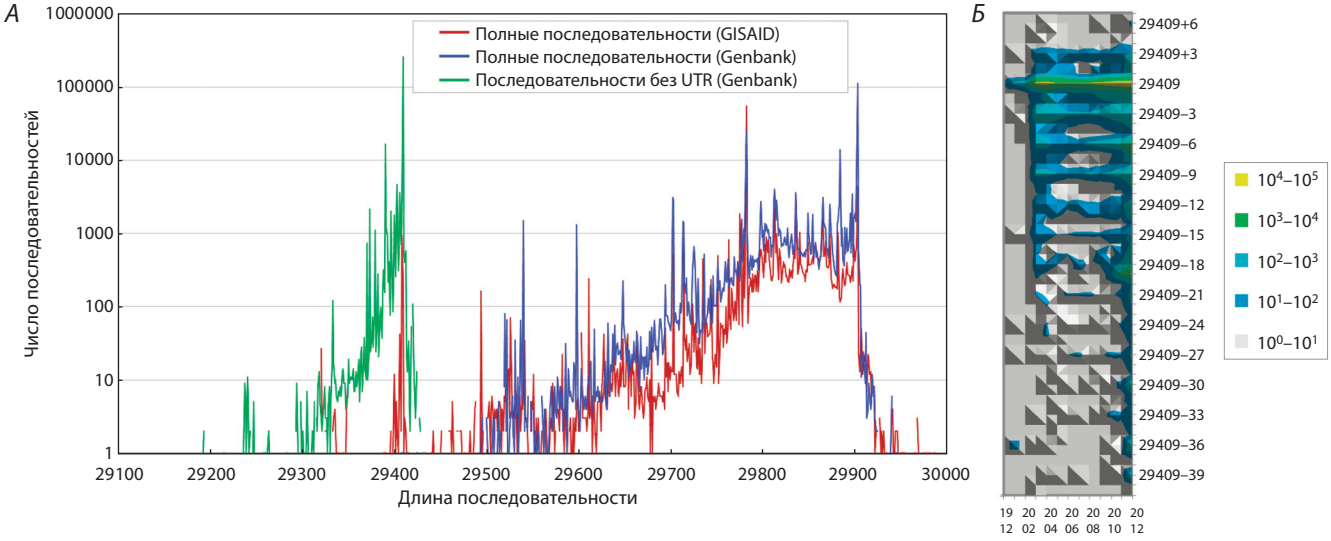

Рис. 4. *А* – распределение полных геномов (GISAID, Genbank) и геномов без UTR (Genbank) за период с момента появления SARS-CoV-2 до конца 2020 г. Длина полного референсного генома SARS-CoV-2 составляет 29903 нт, его же без UTR – 29409 нт. Пиковые значения всех трех кривых соответствуют этим длинам. *Б* – изменение распределения длин геномов без UTR (Genbank) за 2019–2020 гг. по месяцам. По горизонтали – год и месяц, по вертикали – длина генома без UTR, цвета соответствуют числу последовательностей (логарифмическая шкала).

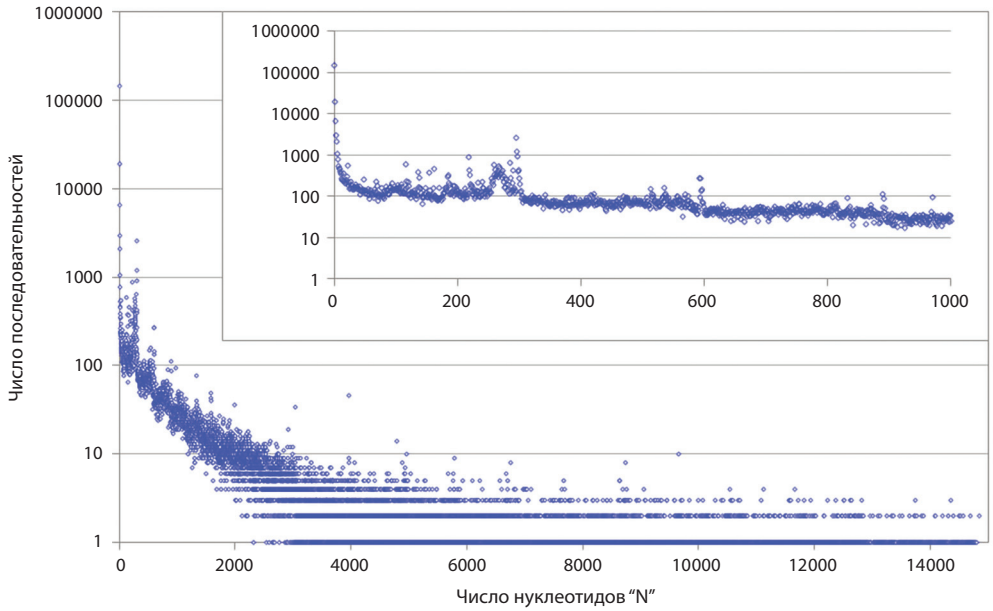

Рис. 5. Соотношение числа неидентифицированных или частично идентифицированных нуклеотидов в транслируемой части геномов SARS-CoV-2, содержащихся в Genbank, полученных в период с конца 2019 г. (начало пандемии) до конца 2020 г.

Врезка содержит часть того же графика, что и на основной картинке, но для области от 0 до 1000 по горизонтали.

ности, длины 5′ UTR и 3′ UTR которых совпадают с референсными, составляют 49.7 и 51.2 % от их полного числа соответственно. Последовательности, длины 5′ UTR и 3′ UTR которых отличаются от референсных не более чем на 10 нт, составляют 55.9 и 55.7 % от их полного числа соответственно.

Из рис. 4, *А* также видно, что основным источником наблюдаемого разброса значений полных длин геномов SARS-CoV-2 действительно были нетранслируемые участки – 5′ и 3′ UTR. Если рассматривать только кодирующую часть, то разброс значительно сокращается: 84.9 % всех последовательностей имеют такую же длину кодирующей части, как и референсный геном, а 90.7 % – длину кодирующей части, отличающуюся от таковой у референсного генома не более чем на 10 нт. Впрочем, среди геномов, длина кодирующей части которых отличается от таковой у референсной последовательности (29409), преобладают те, у которых это отличие кратно трем, – для предотвращения сдвига рамки считывания при трансляции, что обычно приводит к нежизнеспособности (см. рис. 4, *Б*). Таким образом, большинство вирусных последовательностей представляются биологически осмысленными.

Видно, что распределения, полученные на основе данных о полных геномах из GISAID (посредством программных запросов с использованием API) и Genbank (посредством анализа скачанных последовательностей с помощью разработанных нами программных средств) имеют достаточно высокое сходство – вероятно, по причине того, что большинство последовательностей содержатся в обеих базах данных (см. рис. 4). Вопрос о том, сколько последовательностей, отличающихся по длине от референсной, действительно имеют делеции или вставки, а сколько имеют эти отличия из-за ошибок секвенирования и сборки геномов, остается открытым.

2. При изучении генетических последовательностей, представляющих геномы различных вариантов вируса, изменяющиеся с течением времени, часто возникает необходимость сравнения их между собой. Даже если у пары рассматриваемых последовательностей одинаковые длины кодирующих участков, возможность вычислить величину различия между ними (число точечных мутаций) будет зависеть от того, присутствуют ли в этих последовательностях неопределенные нуклеотиды, обычно обозначаемые "N", или другие буквы, помимо стандартных A, T(U), G и C. Используя разработанное нами программное обеспечение и геномы вариантов SARS-CoV-2, полученные в 2019–2020 гг. (содержащиеся в базе данных Genbank), мы рассчитали соотношение между числом последовательностей и числом неидентифицированных или частично идентифицированных нуклеотидов в них (рис. 5).

На большей части графика число последовательностей экспоненциально уменьшается с ростом числа неидентифицированных нуклеотидов, хотя и встречаются участки с некоторыми особенностями. Число последовательностей, в которых все нуклеотиды определены, составляет 47.8 %, а число последовательностей, где неопределенными являются менее 10 нуклеотидов – 58.9 %. Таким образом, для анализа эволюционных изменений, происходящих с вирусом, остается весьма значительная доля от общего числа последовательностей, хранящихся в базе данных.

Обсуждение

Мы осуществили ряд оценок, расчетов и компьютерных вычислений, в том числе с помощью разработанных нами

программных средств, для улучшения понимания того, каким является пространство вариантов генетических последовательностей коронавируса SARS-CoV-2, каковы его основные свойства и особенности, связанные с достаточно длинной геномной последовательностью вируса и низкой для РНК-вирусов вероятностью возникновения мутаций.

Из-за относительно большой длины генома SARS-CoV-2 количество его жизнеспособных вариантов значительно превышает таковое для вирусов, обладающих в несколько раз более коротким геномом. Попробуем определить некоторые другие ориентиры в пространстве вариантов генетических последовательностей. SARS-CoV-2 относится к одноцепочечным РНК(+) вирусам (Modrow et al., 2013). Один из самых маленьких оцРНК(+) вирусов человека – это астровирус 1-го типа (длина генома 6771 нуклеотид) (Lewis et al., 1994). Еще меньшим геномом среди вирусов этого класса обладает нодавирус креветки (*Penaeus vannamei nodavirus*) (Chen et al., 2019) – 4294 нуклеотида. Полное число вариантов различных последовательностей для этих двух длин составляет $3.533 \cdot 10^{4076}$ и $1.760 \cdot 10^{2585}$ соответственно.

Если расширить пространство поиска вирусов с самым малым геномом до ДНК-вирусов, то среди рекордсменов обнаружится цирковирус свиней первого типа, *Porcine circovirus* 1 (PCV1) (Cao et al., 2018), размер генома которого составляет всего 1757–1759 пар нуклеотидов (в 17 раз меньше, чем у SARS-CoV-2). Число возможных вариантов последовательности такой длины составляет $6.597 \cdot 10^{1057}$. Это по-прежнему очень далеко от числа вариантов, которые были потенциально доступны коронавирусу SARS-CoV-2 за период его существования (3.5 года), – $7.985 \cdot 10^{511}$. Весьма близким числом всех возможных вариантов последовательностей, равным $5.636 \cdot 10^{511}$, обладал бы геном размером 850 нт. Впрочем, существуют инфекционные агенты на основе одноцепочечной кольцевой РНК с еще меньшими длинами последовательностей (от 246 до 467 нт) – вироиды (Katsarou et al., 2015). Их РНК не защищена какой-либо оболочкой и не кодирует белков.

Число всех потенциально возможных нуклеотидных последовательностей, которые были бы идентифицированы как варианты SARS-CoV-2, на много порядков превышает число как тех вариантов этого вируса, которые уже были обнаружены и секвенированы, так и тех, которые были опробованы в ходе эволюции, но оказались нежизнеспособными. Рассмотрим коронавирус летучих мышей RaTG13 ($L = 29855$), являющийся ближайшим известным соседом SARS-CoV-2 (но при этом уже другим вирусом) в пространстве вариантов генетических последовательностей, отличающийся от него на 1135 точечных мутаций. Число вариантов последовательностей, отличающихся от референсного генома SARS-CoV-2 не более чем на 1135 мутаций, составляет $\approx 2.943 \cdot 10^{5621}$, что более чем на 1000 порядков превышает полное количество возможных вариантов последовательностей длиной как 4294 ($1.76 \cdot 10^{2585}$), так и 6771 ($3.53 \cdot 10^{4076}$), т. е. может содержать внутри себя объемы информации, которых хватит на огромное количество различных небольших вирусов.

Глобальное филогенетическое древо коронавируса показывает, что вирус с течением времени подвержен изменениям (возможно, вынужденным) – видимо, в связи с тем, что на него действует давление естественного отбора. Еще одной причиной изменений является внутривидовая конкуренция; например, варианты с более быстрыми РНК-полимеразами вытесняют варианты с более медленными (поскольку их число растет быстрее) и тем самым уменьшают инкубационный период вируса, а менее летальные штаммы позволяют вирусу дольше и шире распространяться (носитель не умирает, а является распространителем вируса на протяжении почти всего периода заболевания; легко перенося болезнь, человек остается социально активным и заражает больше других людей в своем окружении). В отличие от упомянутых выше вироидов, изменения в геноме настоящих вирусов, в том числе SARS-CoV-2, могут оказывать различный эффект на внутривидовую конкуренцию в зависимости от функций белков, закодированных в геноме. Этот вопрос остался за рамками данной работы, однако в последующих публикациях мы планируем уделить ему должное внимание.

Кроме того, влияние оказывает и формирование у человечества иммунитета к этому вирусу. Вероятно, имеются и другие механизмы. При этом все эти изменения должны происходить не в ущерб функциональным возможностям вируса. Таким образом, получается, что пространство доступных коронавирусу SARS-CoV-2 вариантов является довольно узким, а траектории его развития, возможно, в некоторой степени детерминированы. И действительно, было показано, что геном SARS-CoV-2 имеет гораздо более низкую частоту мутаций и генетическое разнообразие по сравнению с вирусом SARS-CoV, вызвавшим вспышку атипичной пневмонии в 2002–2003 гг. (Jia et al., 2020; Zhou et al., 2020; Никонова и др., 2021). Так, например, для S-белка значения d_N и d_S для SARS-CoV-2 оказались приблизительно в 12 и 7 раз ниже, чем для SARS-CoV (d_N – доля последовательностей в выборке геномов, в которых присутствуют несинонимичные мутации в определенном гене, d_S – аналогично доля синонимичных мутаций). Для более консервативных генов ORF1a и ORF1b отношения частот мутаций

$$(d_N^{\text{SARS-CoV-2}}/d_N^{\text{SARS-CoV}},\ d_S^{\text{SARS-CoV-2}}/d_S^{\text{SARS-CoV}})$$

имеют меньшую величину, но значения для SARS-CoV-2 ниже, чем для SARS-CoV (лежат в пределах интервала от $\frac{1}{4.8}$ до $\frac{1}{1.5}$). Гипотеза о частичной детерминированности траекторий развития вируса состоит в том, что если бы развитие пандемии SARS-CoV-2, с самого ее начала в декабре 2019 г., в силу случайных факторов пошло бы несколько иначе, то, несмотря на это, раньше или позже, в том же порядке или в ином пространство жизнеспособных вариантов, «посещенных» вирусом, все равно оказалось бы примерно таким же. Вышесказанное позволяет предположить возможность создания эволюционного симулятора на основе анализа траекторий изменения вируса с течением времени, что входит в наши планы на будущее в рамках работы по данной тематике.

Заключение

Изучение пространства вариантов генетических последовательностей – важный этап в разработке подходов к моделированию эволюции вирусов и других организмов. Для построения новой, существенно более реалистичной модели эволюции вируса, способной рассчитывать потенциально возможные, но еще не реализованные в природе варианты новых геномов, чтобы заблаговременно противостоять их появлению, необходимо ответить на такие вопросы как: «Какова вероятность возникновения рекомбинантных вариантов вируса и имеются ли предпочитаемые позиции, по которым происходит обмен частями генома?», «Можем ли мы предположить или рассчитать, какой вариант реализуется, а какой окажется нежизнеспособным?», «Можно ли было предсказать “Дельту” или “Омикрон”?» и, наконец, «Если бы удалось создать реалистичную модель эволюции SARS-CoV-2 и несколько раз рассчитать процесс с самого начала, от исходной референсной последовательности, каждый раз он протекал бы по-разному и приводил к существенно различающимся результатам, или все происходило бы примерно одинаково с небольшими вариациями?»

Список литературы / References

Никонова А.А., Файзулоев Е.Б., Грачева А.В., Исаков И.Ю., Зверев В.В. Генетическое разнообразие и эволюция биологических свойств коронавируса SARS-CoV-2 в условиях глобального распространения. *Acta Naturae*. 2021;13(3):77-89. DOI 10.32607/actanaturae.11337
[Nikonova A.A., Faizuloev E.B., Gracheva A.V., Isakov I.Yu., Zverev V.V. Genetic diversity and evolution of the biological features of the pandemic SARS-CoV-2. *Acta Naturae*. 2021;13(3):77-89. DOI 10.32607/actanaturae.11337]

Aksamentov I., Roemer C., Hodcroft E.B., Neher R.A. Nextclade: clade assignment, mutation calling and quality control for viral genomes. *J. Open Source Software*. 2021;6(67):3773. DOI 10.21105/joss.03773

Amicone M., Borges V., Alves M.J., Isidro J., Zé-Zé L., Duarte S., Vieira L., Guiomar R., Gomes J.P., Gordo I. Mutation rate of SARS-CoV-2 and emergence of mutators during experimental evolution. *Evol. Med. Public Health*. 2022;10(1):142-155. DOI 10.1093/emph/eoac010

Campagnola G., Govindarajan V., Pelletier A., Canard B., Peersen O.B. The SARS-CoV nsp12 polymerase active site is tuned for large-genome replication. *J. Virol.* 2022;96(16):e0067122. DOI 10.1128/jvi.00671-22

Cao L., Sun W., Lu H., Tian M., Xie C., Zhao G., Han J., Wang W., Zheng M., Du R., Jin N., Qian A. Genetic variation analysis of PCV1 strains isolated from Guangxi Province of China in 2015. *BMC Vet. Res.* 2018;14(1):43. DOI 10.1186/s12917-018-1345-z

Chen N.C., Yoshimura M., Miyazaki N., Guan H.-H., Chuankhayan P., Lin C.-C., Chen S.-K., Lin P.-J., Huang Y.-C., Iwasaki K., Nakagawa A., Chan S.I., Chen C.J. The atomic structures of shrimp nodaviruses reveal new dimeric spike structures and particle polymorphism. *Commun. Biol.* 2019;2:72. DOI 10.1038/s42003-019-0311-z

Day T., Gandon S., Lion S., Otto S.P. On the evolutionary epidemiology of SARS-CoV-2. *Curr. Biol.* 2020;30(15):R849-R857. DOI 10.1016/j.cub.2020.06.031

Grebennikov D., Kholodareva E., Sazonov I., Karsonova A., Meyerhans A., Bocharov G. Intracellular life cycle kinetics of SARS-CoV-2 predicted using mathematical modelling. *Viruses*. 2021;13(9):1735. DOI 10.3390/v13091735

Jia Y., Shen G., Nguyen S., Zhang Y., Huang K., Ho H., Hor W., Yang C., Bruning J.B., Li C., Wang W. Analysis of the mutation dynamics of SARS-CoV-2 reveals the spread history and emergence of RBD mutant with lower ACE2 binding affinity. *bioRxiv.* 2020. DOI 10.1101/2020.04.09.034942

Karimzadeh S., Raj B., Nguyen T.H. Review of infective dose, routes of transmission and outcome of COVID-19 caused by the SARS-COV-2: comparison with other respiratory viruses. *Epidemiol. Infect.* 2021;149:e96. DOI 10.1017/S0950268821000790

Katsarou K., Rao A.L.N., Tsagris M., Kalantidis K. Infectious long non-coding RNAs. *Biochimie*. 2015;117:37-47. DOI 10.1016/j.biochi.2015.05.005

Khare S., Gurry C., Freitas L., Schultz M.B., Bach G., Diallo A., Akite N., Ho J., Lee R.T., Yeo W., Curation Team GC, Maurer-Stroh S. GISAID’s role in pandemic response. *China CDC Weekly*. 2021;3(49):1049-1051. DOI 10.46234/ccdcw2021.255

Kim H., Webster R.G., Webby R.J. Influenza virus: dealing with a drifting and shifting pathogen. *Viral Immunol.* 2018;31(2):174-183. DOI 10.1089/vim.2017.0141

Lewis T.L., Greenberg H.B., Herrmann J.E., Smith L.S., Matsui S.M. Analysis of astrovirus serotype 1 RNA, identification of the viral RNA-dependent RNA polymerase motif, and expression of a viral structural protein. *J. Virol.* 1994;68(1):77-83. DOI 10.1128/JVI.68.1.77-83.1994

Li P., Hu J., Liu Y., Ou X., Mu Z., Lu X., Zan F., Cao M., Tan L., Dong S., Zhou Y., Lu J., Jin Q., Wang J., Wu Z., Zhang Y., Qian Z. Effect of polymorphism in *Rhinolophus affinis* ACE2 on entry of SARS-CoV-2 related bat coronaviruses. *PLoS Pathog.* 2023;19(1):e1011116. DOI 10.1371/journal.ppat.1011116

Malone B., Urakova N., Snijder E.J., Campbell E.A. Structures and functions of coronavirus replication-transcription complexes and their relevance for SARS-CoV-2 drug design. *Nat. Rev. Mol. Cell Biol*. 2022;23(1):21-39. DOI 10.1038/s41580-021-00432-z

Markov P.V., Ghafari M., Beer M., Lythgoe K., Simmonds P., Stilianakis N.I., Katzourakis A. The evolution of SARS-CoV-2. *Nat. Rev. Microbiol.* 2023;21(6):361-379. DOI 10.1038/s41579-023-00878-2

Modrow S., Falke D., Truyen U., Schätzl H. Viruses with single-stranded, positive-sense RNA genomes. In: Molecular Virology. Berlin: Springer, 2013;185-349. DOI 10.1007/978-3-642-20718-1_14

Mütze T. Combinatorial Gray codes – an updated survey. *Electron. J. Comb.* 2023;30(3):DS26. DOI 10.37236/11023

Palyanova N., Sobolev I., Alekseev A., Glushenko A., Kazachkova E., Markhaev A., Kononova Y., Gulyaeva M., Adamenko L., Kurskaya O., Bi Y., Xin Y., Sharshov K., Shestopalov A. Genomic and epidemiological features of COVID-19 in the Novosibirsk region during the beginning of the pandemic. *Viruses.* 2022;14(9):2036. DOI 10.3390/v14092036

Palyanova N.V., Sobolev I.A., Palyanov A.Y., Kurskaya O.G., Komissarov A.B., Danilenko D.M., Fadeev A.V., Shestopalov A.M. The development of the SARS-CoV-2 epidemic in different regions of Siberia in the 2020–2022 period. *Viruses*. 2023;15:2014. DOI 10.3390/v15102014

Ruan Y., Luo Z., Tang X., Li G., Wen H., He X., Lu X., Lu J., Wu C.I. On the founder effect in COVID-19 outbreaks: how many infected travelers may have started them all? *Natl. Sci. Rev.* 2020;8(1):nwaa246. DOI 10.1093/nsr/nwaa246

Sayers E.W., Bolton E.E., Brister J.R., Canese K., Chan J., Comeau D.C., Connor R., Funk K., Kelly C., Kim S., Madej T., Marchler-Bauer A., Lanczycki C., Lathrop S., Lu Z., Thibaud-Nissen F., Murphy T., Phan L., Skripchenko Y., Tse T., Wang J., Williams R., Trawick B.W., Pruitt K.D., Sherry S.T. Database resources of the national center for biotechnology information. *Nucleic Acids Res*. 2022;50(D1):D20-D26. DOI 10.1093/nar/gkab1112

Sender R., Bar-On Y.M., Gleizer S., Bernshtein B., Flamholz A., Phillips R., Milo R. The total number and mass of SARS-CoV-2 virions. *Proc. Natl. Acad. Sci. USA.* 2021;118(25):e2024815118. DOI 10.1073/pnas.2024815118

Shannon A., Selisko B., Le N.T., Huchting J., Touret F., Piorkowski G., Fattorini V., Ferron F., Decroly E., Meier C., Coutard B., Peersen O.,

Canard B. Rapid incorporation of Favipiravir by the fast and permissive viral RNA polymerase complex results in SARS-CoV-2 lethal mutagenesis. *Nat. Commun*. 2020;11(1):4682. DOI 10.1038/s41467-020-18463-z

Sonnleitner S.T., Prelog M., Sonnleitner S., Hinterbichler E., Halbfurter H., Kopecky D.B.C., Almanzar G., Koblmüller S., Sturmbauer C., Feist L., Horres R., Posch W., Walder G. Cumulative SARS-CoV-2 mutations and corresponding changes in immunity in an immunocompromised patient indicate viral evolution within the host. *Nat. Commun.* 2022;13(1):2560. DOI 10.1038/s41467-022-30163-4

Wirth W., Duchene S. GISAIDR: programmatically interact with the GISAID databases. *Zenodo*. 2022. DOI 10.5281/zenodo.6474693

Wu F., Zhao S., Yu B., Chen Y.M., Wang W., Song Z.G., Hu Y., Tao Z.W., Tian J.H., Pei Y.Y., Yuan M.L., Zhang Y.L., Dai F.H., Liu Y., Wang Q.M., Zheng J.J., Xu L., Holmes E.C., Zhang Y.Z. A new coronavirus associated with human respiratory disease in China. *Nature*. 2020;579(7798):265-269. DOI 10.1038/s41586-020-2008-3

Xu H., Zhang Y., Yuan M., Ma L., Liu M., Gan H., Liu W., Lum G.G.A., Tao F. Basic reproduction number of the 2019 novel coronavirus disease in the major endemic areas of China: a latent profile analysis. *Front. Public Health*. 2021;9:575315. DOI 10.3389/fpubh.2021.575315

Zhou P., Yang X.-L., Wang X.-G., Hu B., Zhang L., Zhang W., Si H.R., Zhu Y., Li B., Huang C.L., Chen H.D., Chen J., Luo Y., Guo H., Jiang R.D., Liu M.Q., Chen Y., Shen X.R., Wang X., Zheng X.S., Zhao K., Chen Q.J., Deng F., Liu L.L., Yan B., Zhan F.X., Wang Y.Y., Xiao G.F., Shi Z.L. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*. 2020;579(7798):270-273. DOI 10.1038/s41586-020-2012-7

ORCID ID

A.Yu. Palyanov orcid.org/0000-0003-1108-1486
N.V. Palyanova orcid.org/0000-0002-1783-5798

Финансирование. Исследование выполнено за счет гранта Российского научного фонда, проект № 23-64-00005.

Благодарности. Мы выражаем признательность всем, кто получил и сделал общедоступными генетические последовательности вариантов SARS-CoV-2, которые были задействованы при проведении некоторых расчетов в ходе нашего исследования, – авторам, лабораториям, ответственным за получение образцов, лабораториям, осуществившим секвенирование и ввод метаданных, а также размещение этой информации в базах данных GISAID и Genbank. Мы также благодарны авторам проекта Nextclade, предоставляющего онлайн-инструменты для анализа и визуализации генетических последовательностей различных вирусов.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 16.07.2023. После доработки 14.09.2023. Принята к публикации 18.09.2023.