Филогенетическая реконструкция начальных этапов распространения вируса SARS-CoV-2 на Евроазиатском и Американском континентах посредством анализа геномных данных (краткое сообщение)

Ю. С. Букин[1,4], А. Н. Бондарюк[1,2], С. В. Балахонов[2], Ю. П. Джиоев[3], В. И. Злобин[3]*

[1] Лимнологический институт Сибирского отделения Российской Академии наук, Иркутск, Россия

[2] Научно-исследовательский противочумный институт Сибири и Дальнего востока, Иркутск, Россия

[3] Иркутский государственный медицинский университет, Иркутск, Россия

[4] Иркутский государственный университет, Иркутск, Россия

*email: bukinyura@mail.ru

Резюме: Проанализированы 252 полных генома вируса *SARS-CoV-2* первой волны (декабря 2019 - июль 2020 г.) пандемии COVID-19 из 21 страны мира, включая Россию, посредством Байесовского филогенетического метода с молекулярными часами. Используемая нами методика показала, что первые заболевшие COVID-19 в человеческой популяции появились в период с июля по ноябрь 2019 г. в Китае. Распространение *SARS-CoV-2* из Китая по всем регионам мира произошло с декабря 2019 по начало февраля 2020 года. Появление вируса в России датируется второй половиной января 2020 года. Скорость эволюции кодирующей части генома *SARS-CoV-2* равная в среднем 7.3×10^{-4} (5.95×10^{-4} – 8.68×10^{-4}) нуклеотидных замен на сайт в год сопоставима со скоростями накопления замен в геномах других человеческих РНК-содержащих вирусах (*Measles morbillivirus, Rubella virus, Enterovirus C*).

Ключевые слова: SARS-CoV-2, COVID-19, геномика, Байесовский филогенетический анализ, молекулярные часы.

В конце 2019 года в г. Ухань, Китайской Народной Республики была зафиксирована массовая вспышка острой респираторной инфекции с осложнениями и большим количеством летальных исходов. Анализ генетического материала, взятого от пациентов в г. Ухань Китая за период с 23 – 26 декабря 2019 г., показал [1, 2], что заболевание вызывается новым, неизвестным науке, РНК-содержащим вирусом. Первый

полный расшифрованный геном нового вируса был депонирован в базу данных GenBank уже 5 января 2020 года (длина генома ≈ 29903, длина кодирующей части ≈ 29260 нуклеотидов). Генетически вирус оказался близок к подроду *Sarbecovirus*, рода *Betacoronavirus* семейства *Coronaviridae*. Вирусы этой группы вызывают тяжелый острый респираторный синдром (ТОРС - SARS) человека и животных. Новый вирус получил название SARS-CoV-2, а вызываемое им заболевание - COVID-19 [3]. По структуре генома вирус SARS-CoV-2 оказался близок выявленному ранее коронавирусу *SARS-CoV* – этиологическому агенту эпидемии атипичной пневмонии 2002-2003 годов [2, 4] и *MERS-CoV*, вызвавшему вспышку ближневосточного респираторного синдрома в 2013 году в Саудовской Аравии [2, 5].

Заболевание COVID-19 за короткий период распространилось из Китая практически на все страны мира. Сообщения в СМИ о первых заболевших COVID-19 за пределами Китая (другие страны Азии, Европы, Америки и Австралии) стали массовыми во второй половине января 2020 года.

В России первые заболевшие были зарегистрированы 31 января 2020 (два гражданина КНР в Читинской и Тюменской областях). Массовая регистрация новых случаев COVID-19 в России началась с 7 марта 2020 года.

На сегодня [6] расшифровано большое количество полных геномов различных штаммов *SARS-CoV-2*, изолированных во время глобальной пандемии. Полученные данные депонированы в открытом доступе в базы “GISAID” (https://gisaid.org) и “GenBank” (https://www.ncbi.nlm.nih.gov/).

Методы геномного филогенетического анализа с использованием байесовского подхода позволяют реконструировать эволюционную историю эпидемий вирусов с использованием концепции молекулярных часов, датируя древо на основе времени изоляции штаммов. Обычно считается, что метод молекулярных часов очень приблизителен даже в масштабах многих миллионов лет, однако вирусы мутируют с большой скоростью. Благодаря этому может быть достигнута высокая точность датировок эволюционных событий до нескольких месяцев и даже дней.

Сложность эпидемиологических исследований COVID-19 заключается в массовом распространении бессимптомных и слабо выраженных форм заболевания. Поэтому ранние этапы распространения инфекции в мире, или в какой либо стране, могут быть скрыты от глаз эпидемиологов. Применение методов байесовской филогении, как оказалось, позволяет выявить закономерности эволюции и распространения вируса, которые невозможно установить путем анализа эпидемиологических данных.

Целью нашего исследования была реконструкция датированного филогенетического древа байесовским филогенетическим методом с молекулярными часами для идентификации даты появления первых заболевших вирусом SARS-CoV-2 в человеческой популяции, выявления региона (страны) происхождения инфекции, определения даты появления и распространения инфекции в России и в мире, фиксации скорости накопления нуклеотидных замен в геноме вируса.

Исходные данные для анализа первой волны пандемии COVID-19 (с декабря 2019 по июль 2020 г.) включали от 10 до 16 геномов вируса *SARS-CoV-2* из 20 разных стран мира (Китай, Южная Корея, Таиланд, Япония, США, Мексика, Польша, Вьетнам, Франция, Испания, Египет, Израиль, Греция, Казахстан, Италия, Украина, Великобритания, Бразилия) и 20 геномов из России. Расшифрованные геномы были взяты из баз “GISAID” и “GenBank”. Мы выбрали страны с учетом возможности охвата всех регионов мира и наличия значительного пассажиропотока с Россией. В данные по Китаю включены все 4 генома, выделенные в декабре 2019 г. В анализе использовали только кодирующие части геномов, так как некодирующие участки геномов из базы “GISAID” содержали большое количество неидентифицированных нуклеотидов и не были пригодны для анализа.

Предварительный анализ в программе IQ-TREE показал [7], что эволюционная модель с разделением кодирующей части генома на 1+2 и 3 позиции кодона имеет значительное информативное преимущество перед анализом без разделения на позиции кодонов. Весь дальнейший эволюционный анализ выполнялся по рекомендуемой в этом случае модели эволюции ДНК HKY+I+G [8] с разделением на 1+2 и 3 позиции кодона.

Филогенетический анализ с молекулярными часами путем датировки древа временем изоляции штаммов *SARS-CoV-2* проводился байесовским филогенетическим методом в программе “BEAST v. 2.6.2” [9].

В пакете “MODEL-SELECTION” для “BEAST v. 2.6.2” с помощью метода “Path sampling” [11] было проведено тестирование с целью выяснения возможности использования времени изоляции штаммов для датирования филогенетического древа. В результате было установлено, что наибольший вес имеет эволюционная реконструкция, датированная временем изоляции штаммов с экспоненциальным ростом эффективной численности популяции вируса (ростом числа болеющих) и расслабленными молекулярными часами (разная скорость эволюции в разных филогенетических линиях вируса). Реконструированное по этой модели древо представлено на рисунке 1.

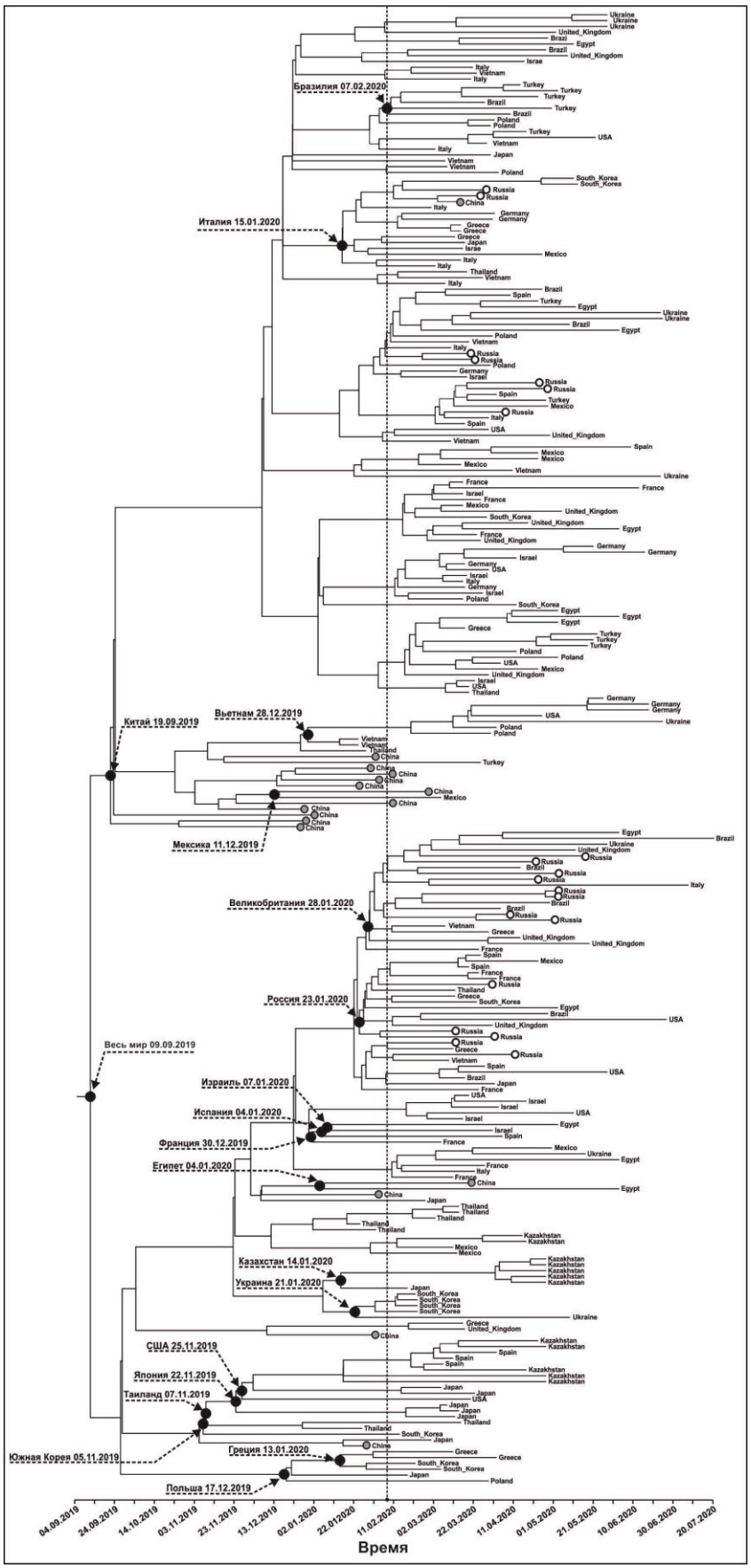

Рис. 1. Датированное филогенетическое древо вируса *SARS-CoV-2*, построенное на основе кодирующей части полных геномов штаммов, изолированных в период с декабря 2019 до июля 2020 года. Белыми кругами отмечены штаммы, изолированные в России, серыми кругами отмечены штаммы, изолированные в Китае. Прочие штаммы отмечены названиями стран. В узлах показаны датировки появления первых филогенетических линий *SARS-CoV-2* в каждой из 21 исследованной страны.

Анализ показал (рис. 1), что первые заболевшие COVID-19 в популяции человека появились в период с 21.07.2019 по 27.10.2019 г. Наиболее ранняя филогенетическая линия вируса *SARS-CoV-2* принадлежит Китаю (19.09.2020 г.). Наши результаты потвердели первоначальные выводы некоторых экспертов [1 -3], что впервые вирус, видимо, стал поражать людей, перейдя от животных к человеку в середине 2019 г. в Китае. Там же проходил начальный этап эпидемии. С декабря 2019 г. по начало февраля 2020 года вирус *SARS-CoV-2* распространился по всему миру, заболевание COVID-19 приняло характер пандемии. В России наиболее ранняя филогенетическая линия *SARS-CoV-2* датируется 23 января 2020 г.

В базальной части филогенетического древа (с 09.09.2019 по 15.02.2020) можно выделить несколько кластеров (рис. 1). Однако апостериорные поддержки этих кластеров были ≤ 0.75. Поэтому ни о каких закономерностях в топологии древа и наличия устойчивых генотипов вируса *SARS-CoV-2* в этот период времени говорить нельзя.

Таблица 1. Оценки скоростей эволюции в кодирующей части геномов вируса *SARS-CoV-2*

Тип замен	Скорость эволюции - нуклеотидных замен на сайт в год (плюс 95% доверительный интервал)	Среднее количество замен за месяц на кодирующую часть (плюс 95% доверительный интервал)
1+2 позиции кодона	6.02×10^{-4} 4.98×10^{-4}– 7.16×10^{-4}	0.98 0.81 – 1.16
3 позиция кодона	8.59×10^{-4} 6.92×10^{-4} – 10.2×10^{-4}	0.7 0.56 – 0.83
1+2+3 позиции кодона	7.31×10^{-4} 5.95×10^{-4} – 8.68×10^{-4}	1.68 1.37 – 1.99

Оцененные значения скоростей накопления замен в кодирующей части геномов *SARS-CoV-2* представлены в таблице 1. Мутации в 1+2 позициях кодона, как правило, приводят к изменениям аминокислот. Поэтому линии SARS-CoV-2 в месяц будут накапливать в среднем 0.81 – 1.16 аминокислотных замен. Скорость мутационного процесса у *SARS-CoV-2* меньше или сопоставима со скоростями эволюции таких РНК-содержащих вирусов как: вирус кори (*Measles morbillivirus*) [11], вирус краснухи (*Rubella virus*) [12], вирус полиомиелита (*Enterovirus C*) [13].

К сожалению, небольшая для вирусов скорость эволюции и отсутствие высоких поддержек топологии филогенетического древа в области базальных узлов не позволяют

достоверно определить пути миграции и перехода SARS-CoV-2 из одной страны в другую в первой волне пандемии COVID-19 только по геномным данным без дополнительной информации о перемещении конкретных заболевших.

Линия *SARS-CoV-2*, циркулирующая сегодня в глобальном масштабе в популяции человека, отделена от общего предка с блажащим родственником вирусом *Bat coronavirus RaTG13* рода *Betacoronavirus* ~ 2% (96% сходства в полном геноме) нуклеотидных замен [14]. Исходя из рассчитанных нами скоростей эволюции, отделение вируса *SARS-CoV-2* от ближайшего общего предка с другими короновирусами произошло в пределах от 33 до 23 лет назад (конец 20 века).

Расчеты в программе “BEAST v. 2.6.2” при тестировании филогенетических гипотез и реконструкции древа потребовали привлечения очень больших вычислительных ресурсов обеспеченных доступом к суперкомпьютеру “Академик Матросов” Иркутского суперкомпьютерного центра СО РАН. Проведенный анализ потребовал около двух месяцев расчетов на пяти вычислительных узлах кластера, каждый из которых оснащен двумя 18 ядерными 36 поточными процессорами Intel Xeon E5-2695 v4 «Broadwell».

Проведенный нами анализ выборки геномных данных *SARS-CoV-2* позволяет сделать следующие выводы: 1) как самостоятельная эволюционная линия вирус *SARS-CoV-2* появился в природе в конце 20 века; 2) первые заражения вирусом людей от животных произошли в Китае в середине 2019 года; 3) распространите вируса из Китая, практически по всем странам мира, произошло в период с конца декабря 2019 года по начало февраля 2020 г.; 4) Российская филогенетическая линии *SARS-CoV-2* датируется концом января 2020 г.; 5) скорость эволюции кодирующей части генома *SARS-CoV-2* сопоставима с другими человеческими РНК-содержащими вирусами (*Measles morbillivirus, Rubella virus, Enterovirus C*), после заболевания, которыми формируется устойчивый иммунитет и (или) разработаны высокоэффективные вакцины.

Используемые в работе геномы с номерами “GISAID” и “GenBank”, xml файлы для программы “BEAST v. 2.6.2” и предельные значения функция правдоподобия “Path sampling” анализа для различных вариантов эволюционных реконструкций доступны по ссылке: https://github.com/barnsys/SARS-CoV-2_genome_data.

Список литературы

1. Wu F., et al. // Nature. 2020. V. 579. №. 7798. P. 265-269. DOI https://doi.org/10.1038/s41586-020-2008-3
2. Andersen K. G., , et al. // Nature medicine. 2020. V. 26. № 4. P. 450-452. DOI https://doi.org/10.1038/s41591-020-0820-9
3. Velavan T. P., Meyer C. G. // Tropical medicine & international health. 2020. V. *25*. № 3. P. 278. DOI https://doi.org/10.1111/tmi.13383
4. He R., et al. // Biochemical and biophysical research communications. 2004. V. 316. № 2. P. 476-483. DOI https://doi.org/10.1016/j.bbrc.2004.02.074
5. van Boheemen S., et al. // MBio. 2020. V. 3. № 6. e00473-12. DOI https://doi.org/10.1128/mBio.00473-12
6. Li J., et al. // Scientific reports. 2020 V. 10. № 1. 17492. DOI https://doi.org/10.1038/s41598-020-74656-y
7. Nguyen L. T., et al. // Molecular biology and evolution. 2015. V. *32*. № 1. P. 268-274. DOI https://doi.org/10.1093/molbev/msu300
8. Shapiro B., Rambaut A., Drummond A. J. // Molecular biology and evolution. 2006. V. *23*. № 1. P. 7-9. DOI https://doi.org/10.1093/molbev/msj021
9. Bouckaert R., et al. (2014). *PLoS Comput Biol*, 2014. V. *10*. № 4. e1003537. DOI https://doi.org/10.1371/journal.pcbi.1003537
10. Baele G., et al. // Molecular biology and evolution. 2012. V. 29. № 9. P. 2157-2167. DOI https://doi.org/10.1093/molbev/mss084
11. Furuse Y., Suzuki A., Oshitani H. // Virology journal. 2010. V. 7. № 1. P. 1-4. DOI https://doi.org/10.1186/1743-422X-7-52
12. Zhu Z., Cui A., Wang H. // Journal of clinical microbiology. 2012. V. 50. № 2. P. 353-363. DOI https://doi.org/0.1128/JCM.01264-11
13. Smura T., et al. // PloS one. 2014 V. 9. № 4. e94579. DOI https://doi.org/10.1371/journal.pone.0094579
14. Andersen K. G., Rambaut A., Lipkin W. I., Holmes E. C., Garry R. F. // Nature medicine. 2020. V. *26*. *№* 4. P. 450-452.

Источник финансирования. Работа поддержана бюджетной темой Лимнологического института СО РАН № 0345–2016–0004 (АААА-А16-116122110060-9).

Благодарности. Авторы выражают благодарность Иркутскому суперкомпьютерному центру СО РАН и его администратору Ивану Сидорову за предоставление доступа к высокопроизводительному кластеру “Академик Матросов” для проведения вычислений, академику РАН Михаилу Александровичу Грачеву за ценные комментарии при написании работы.

Информация о конфликте интересов. Авторы заявляют об отсутствии конфликта интересов.

Phylogenetic reconstruction of the initial stages of the spread of the SARS-CoV-2 virus in the Eurasian and American continents by analyzing genomic data

Yu. S. Bukin[1,4], A. N. Bondaryuk[1,2], S. V. Balakhonov[2], Y. P. Dzhioev[3], V. I. Zlobin[3]*

[1] Limnological Institute Siberian Branch of the Russian Academy of Sciences, Irkutsk, Russia

[2] Irkutsk Antiplague Research Institute of Siberia and Far East, Irkutsk, Russia

[3] Irkutsk State Medical University, Irkutsk, Russi

[4] Irkutsk State University, Irkutsk, Russia

*email: bukinyura@mail.ru

Presented by Academician of the RAS M. A. Grachev

Abstract: 252 complete genomes of the SARS-CoV-2 isolated during the first wave (December 2019 - July 2020) of the global COVID-19 pandemic from 21 countries of the world, including Russia, were analyzed using the Bayesian phylogenetic method with a molecular clock. Results showed that the first cases of COVID-19 in the human population appeared in the period between July and November 2019 in China. The spread of SARS-CoV-2 from China toward all regions of the world occurred from December 2019 to early February 2020. The appearance of the virus in Russia dates back to the second half of January 2020. The rate of evolution of the coding part of the SARS-CoV-2 genome equal to 7.3×10^{-4} (5.95×10^{-4} - 8.68×10^{-4}) nucleotide substitutions per site per year is comparable to the rates of accumulation of substitutions in genomes of other human RNA viruses (*Measles morbillivirus, Rubella virus, Enterovirus C*).

Keywords: SARS-CoV-2, COVID-19, genomics, Bayesian phylogenetic analysis, molecular clock.