Федеральное государственное автономное
образовательное учреждение
высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»
Институт фундаментальной биологии и биотехнологии
Кафедра геномики и биоинформатики

УТВЕРЖДАЮ
Заведующий кафедрой
_____ И.Е. Ямских

« _____» _____20 ___ г.

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

«Сравнение методов построения классификации геномов по частотам триплетов и на основе выравнивания на примере геномов семейства *Coronaviridae*»

06.04.01. «Биология»

06.04.01.06 Геномика и биоинформатика

Руководитель _____ профессор, д-р физ.-мат. наук Садовский М.Г.
подпись, дата должность, ученая степень инициалы, фамилия

Студент _____ Кириченко А.Д.
подпись, дата инициалы, фамилия

Красноярск 2021

РЕФЕРАТ

Магистерская диссертация по теме «Сравнение методов построения классификации геномов по частотам триплетов и на основе выравнивания на примере геномов семейства *Coronaviridae*» содержит 78 страниц текстового документа, 2 приложения, 13 иллюстраций, 1 таблицу и 158 использованных источников.

Ключевые слова:

КОРОНАВИРУСЫ, ЭВОЛЮЦИЯ, ФИЛОГЕНЕТИЧЕСКИЙ АНАЛИЗ, ВЫРАВНИВАНИЕ, МЕТОДЫ СВОБОДНЫЕ ОТ ВЫРАВНИВАНИЯ, МЕТОД ДИНАМИЧЕСКИХ ЯДЕР, МЕТОД УПРУГИХ КАРТ.

Объект исследования – геномы коронавирусов.

Цель работы – сравнение двух методов выявления филогенетического отношения на примере семейства геномов коронавирусов: метода, основанного на выравнивании и метода классификации геномов по частотам триплетов.

В результате проделанной работы было отобрано 69 геномов коронавирусов из двух общедоступных баз данных и проанализированы филогенетические отношения между ними методом полногеномного выравнивания и построения филогении на его основе, а также методом классификации геномов по частотам триплетов. Топологии деревьев, полученных методом максимального правдоподобия и с применением байесовского подхода, идентичны и имели умеренную и высокую поддержку ветвей. При помощи метода динамических ядер с последовательно возрастающим числом классов построено иерархическое дерево, последний слой которого представляет собой классы тесно связанных между собой геномов, точно совпадающих с кластерами геномов, выделяемых на филогенетическом дереве. Метод упругих карт выявляет нелинейные связи между объектами (геномами), дополнительные к линейным связям, выявляемых, например, методом динамических ядер. Кластеры, выделенные методом упругих карт, также хорошо совпадают, как с классами последнего слоя слоистого графа, полученного методом динамических ядер, так и с классическим выравниванием. Проведенные вычисления показали высокую эффективность каждого из этих методов. Оба метода также показали хорошее расхождение геномов по таксономическому признаку, но по характеру вызываемых заболеваний соответствие меньше.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Коронавирусы — это большое семейство одноцепочечных РНК-вирусов, которые поражают многие виды животных, включая человека, вызывая респираторные, желудочно-кишечные, печеночные и неврологические заболевания (Weiss et al., 2011) Это семейство является самым крупным из известных семейств РНК-вирусов и делится на четыре рода: альфа-, бета-, гамма- и дельта-коронавирусы (Yang et al., 2015).

Альфа- и бетакоронавирусы заражают только млекопитающих, обычно вызывая респираторные заболевания у людей и гастроэнтериты у животных. Гамма- и дельтакоронавирусы заражают птиц, однако некоторые из них могут также заражать и млекопитающих (Woo et al., 2012). Альфа- и бетакоронавирусы могут вызывать тяжелые болезни домашнего скота. Эти вирусы включают вирус свиного трансмиссивного гастроэнтерита (Brian et al., 2005), вирус кишечной диареи свиней (PEDV) (Lin et al., 2016), коронавирус, вызывающий синдром острой диареи свиней (SADS-CoV) (Zhou et al., 2018) и некоторые другие.

К настоящему времени идентифицировано 7 коронавирусов человека (HCoVs): альфа-коронавирусы HCoVs-NL63 и HCoVs-229E, а также бета-коронавирусы HCoVs-OC43, HCoVs-HKU1, бета-коронавирусы, вызывающие острые респираторные синдромы (SARS-CoV и SARS-CoV-2 и ближневосточный респираторный синдром (MERS-CoV)) (Zaki et.al., 2012; Cheng et.al., 2020). При заражении SARS-CoV-2, SARS-CoV и MERS-CoV развивается тяжелый респираторный синдром у людей; остальные четыре коронавируса человека вызывают легкие заболевания верхних дыхательных путей у иммунокомпетентных хозяев. При этом некоторые из них могут вызывать тяжелое течение инфекции у младенцев, маленьких детей и пожилых людей. Современные представления о происхождении высоко патогенных штаммов коронавирусов человека позволяют утверждать, что в качестве первичного носителя выступали животные. Считается, что у SARS-CoV,

MERS-CoV, HCoV-NL63 и HCoV-229E такими носителями выступали летучие мыши, в то время как HCoV-OC43 и HKU1, вероятно, имели грызунов в качестве таких хозяев (Su et al., 2016; Forni et.al., 2020).

Пандемия, вызванная SARS-CoV-2, резко усилила интерес к данному семейству в целом. Недавние исследования показали, что геном этого вируса на 96 % идентичен геномам коронавирусов летучих мышей, что может свидетельствовать о том, что летучие мыши являются наиболее вероятным первичным хозяином SARS-CoV-2 (Zhou et al., 2020; Ji et.al., 2020). Ранее были высказаны различные предположения о том, кто является промежуточным хозяином данного вируса. В этот список попали змеи (Ji et al., 2020), ящеры (Lam et al., 2020) и норки (Cheng et.al., 2020), однако промежуточных хозяев может быть и несколько. Таким образом, идентификация источника вируса поможет контролировать его распространение.

Все вирусы этого семейства являются достаточно близкими для осуществления полногеномного выравнивания и построения филогении на его основе. Однако данный метод классификации является весьма трудоёмким в реализации, что делает задачу поиска альтернативных методов актуальной. Кроме того, такие альтернативные методы могут выявлять связи между геномами, которые не выявляются выравниванием. Так, например, сравнение полных геномов митохондрий животных по частотам триплетного состава выявляет очень сильную связь между классами, выделяемыми в пространстве частот классификацией без учителя (методом динамических ядер) и традиционной систематикой животных (Sadovsky et al., 2015), а также между функцией кодируемого гена и таксономией (Putinseva et al., 2019; Fedotovskaya et.al., 2020; Sadovsky et.al., 2020).

Целью настоящей работы является сравнение двух методов выявления филогенетического отношения на примере семейства геномов коронавирусов: метода, основанного на выравнивании и метода классификации геномов по частотам триплетов. Здесь следует подчеркнуть, что целью нашей работы является не установление полной эквивалентности этих двух методов, а

выявление содержательных различий в выявляемых ими филогенетических связях.

Для достижения этой цели были поставлены следующие задачи:

1. Провести филогенетический анализ геномов коронавируса на основе их полногеномного выравнивания;
2. Выполнить методами динамических ядер и упругих карт кластерный анализ частотных словарей триплетов, построенных для изучаемых геномов;
3. Сравнить полученные результаты для этих двух методов.

Глава 1. Обзор литературы

1.1. Структура коронавирусов

Коронавирусы принадлежат к подсемейству *Coronavirinae* семейства *Coronaviridae* порядка *Nidovirales*, это подсемейство включает в себя четыре рода: *Alphacoronavirus*, *Betacoronavirus*, *Gammacoronavirus* и *Deltacoronavirus*. Геном коронавирусов представляет собой одноцепочечную РНК длиной около 30 т.п.н., что позволяет считать их геномы самыми крупными из известных среди РНК-вирусов, потому как средний размер генома РНК-вирусов 10 т.п.н.

Геном коронавирусов упакован внутри спирального капсида, образованного нуклеокапсидным белком (N) и дополнительно окруженного оболочкой. С вирусной оболочкой связаны, по крайней мере, три структурных белка: мембранный белок (M) и белок оболочки (E), участвующие в сборке вируса, тогда как спайковый белок (S) обеспечивает проникновение вируса в клетки хозяев. Некоторые коронавирусы также кодируют связанный с оболочкой белок гемагглютинин-эстеразы (HE) (Chen et al., 2020).

Геном типичного коронавируса состоит не менее, чем из шести обязательных генов. На 5'-конце генома находится лидерная последовательность, которая играет важную роль в экспрессии генов коронавирусов во время прерывистой субгеномной репликации. Вблизи 5'-кэп-структуры находится ген репликазы, который состоит из двух перекрывающихся открытых рамок считывания, ORF 1a и 1b, которые занимают около 2/3 всего генома. Между ORF1a и ORF1b существует сдвиг рамки считывания -1, что приводит к продукции двух полипротеинов: pp1a и pp1ab. Затем полипротеины процессируются двумя или тремя вирусными доменами протеиназы с образованием мембраносвязанного комплекса репликаза-транскриптаза (Brockway et al., 2003). После протеолитического процессинга полипептид ORF 1ab со сдвигом рамки считывания генерирует 15–16 неструктурных белков (nsp), многие из которых участвуют либо в синтезе РНК, либо в протеолитическом процессинге, необходимом для репликации

вируса: nsp1 – nsp11, кодируются в ORF 1a, а nsp12–16, кодируются в ORF1b (Ziebuhr et al., 2000).

После ORF1b расположены четыре открытых рамки считывания, которые кодируют общий для всех коронавирусов набор структурных белков: спайковый белок (S), мембраны (M), оболочки (E) и нуклеокапсида (N). Белок S опосредует прикрепление вируса к специфическим клеточным рецепторам и слияние между оболочкой и плазматической мембраной и является основным индуктором нейтрализующих вирус антител. Белок оболочки (E) играет важную роль в сборке вирусной оболочки, но не является существенным для размножения вируса. Мембранный белок (M), наиболее распространенный структурный компонент, представляет собой гликопротеин III типа, состоящий из короткого аминоконцевого эктодомена, трансмембранного домена и длинного внутреннего домена с карбоксильным концом. Белок нуклеокапсида (N) представляет собой фосфопротеин, который помимо своей функции в вирионе также модулирует синтез вирусной РНК (Brian et al., 2005).

Помимо этих четырех основных структурных белков, разные коронавирусы кодируют специальные структурные и вспомогательные белки, такие как белок гемагглютинин-эстераза (HE), белок 3a/b, белок 4a/b и другие, они не нужны при репликации, но, видимо, играют важную роль в патогенезе. Их количество, нуклеотидная последовательность и порядок могут заметно различаться у разных коронавирусов. Функция дополнительных белков в большинстве случаев неизвестна. Однако они действительно играют важную роль во взаимодействиях вируса-хозяина, поскольку обычно сохраняются во время естественного инфицирования, а их потеря приводит к снижению вирулентности (Chen et al., 2020; Van Boheemen et al., 2012; Lu et al.,2015).

На рисунке 1 приведены филогенетическое дерево и структуры геномов типичных представителей коронавирусов (Chen et al., 2020).

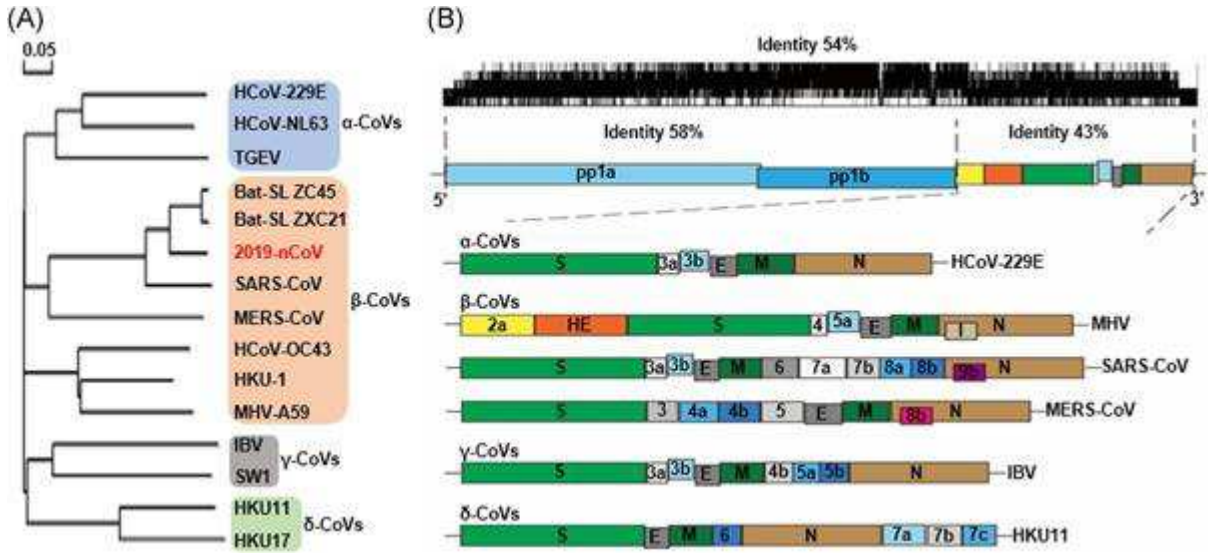

Рисунок 1 – А. Филогенетическое дерево репрезентативных видов коронавирусов. В. Структура геномов четырех родов коронавирусов. Последовательности pp1a и pp1b представляют собой два длинных полипептида, которые процессируются в 16 неструктурных белков. S, E, M и N обозначают четыре структурных белка: спайковый, оболочки, мембраны и нуклеокапсида, а также HE − гемагглютинин-эстераза (Chen et al., 2020)

Исходя из репрезентативной выборки, представленной на рисунке 1, выравнивание геномных последовательностей коронавирусов показывает 58% идентичности в области, кодирующей неструктурные белки, и 43% идентичности в области, кодирующей структурные белки среди различных коронавирусов. Идентичность на уровне всего генома 54%. То есть, предполагается, что область, кодирующая неструктурные белки более консервативна, а область, кодирующая структурные белки более разнообразна и нуждается в адаптации к новым хозяевам(Chen et al., 2020).

1.1.1. Альфакоронавирусы

Репрезентативные альфакоронавирусы включают коронавирусы человека NL63 (HCoV-NL63) и 229E ((HCoV-229E), коронавирус трансмиссивного гастроэнтерита свиней (TGEV), вирус эпидимической диареи свиней (PEDV), респираторный коронавирус свиней (PRCV) и вирус инфекционного перитонита кошек (FIPV).

Помимо обязательного набора генов у них так же зачастую во второй трети геноме есть дополнительные гены (Dye et al., 2005; Haijema et al., 2007; Tekes et al., 2008). Коронавирусы человека HCoV-NL63 и PEDV, кодируют один дополнительный белок между генами S и E. Однако другие альфакоронавирусы, такие как человеческий HCoV-229E, TGEV и FIPV, содержат два, три и пять дополнительных белков, соответственно. Так, например, у альфакаронавирусов в области между генами, кодирующими S и E белки, находится ORF3, кодирующая вспомогательные белки 3a, 3b и 3c (Tekes et al., 2016). Например, у кошачьего (FIPV) и собачьего (CCoV) коронавирусов имеется все 3 белка, в то время как у коронавируса трансмиссивного гастроэнтерита свиней присутствуют только белки 3a и 3b, как и у коронавирусов человека NL63 и 229E. При этом анализ геномных последовательностей указывает на тот факт, что у всех вышеперечисленных вирусов 3a белки являются гомологичными, в то время как ген, кодирующий 3c кошачьего коронавируса является гомологом для 3c коронавируса, поражающего собак и гомологом 3b трансмиссивного гастроэнтерита свиней (Narayanan et al., 2007).

Также отмечается, что разные представители альфакоронавирусов содержат различное количество дополнительных генов, расположенных уже после гена, кодирующего белок нуклеокапсида (Haijema et al., 2004).

1.1.2. Бетакоронавирусы

Типичные бетакоронавирусы включают SARS-CoV, SARS-CoV-2, MERS-CoV, коронавирус летучих мышей HKU4, коронавирус гепатита мышей (MHV), коронавирус крупного рогатого скота (BCoV), коронавирус человека OC43 и др. Они также содержат разное количество дополнительных белков: каждый из коронавирусов MHV, HCoV-OC43 и MERS-CoV имеет пять дополнительных белков, а в SARS-CoV идентифицировано не менее восьми (Li et al., 2014, Chafekar et al., 2018).

Как известно, основной ответ клеток млекопитающих на вирусную инфекцию — это активация врожденного иммунного ответа посредством продукции интерферона I типа. Исследования с использованием временной сверхэкспрессии дополнительных белков ORF4a, ORF4b и ORF5 показали, что они ингибируют как индукцию интерферона I типа (Yang et al., 2013; Niemeyer et al., 2013; Matthews et al., 2014), так и пути передачи сигналов NF-kappaB, который является универсальным фактором транскрипции, контролирующим экспрессию генов иммунного ответа (Matthews et al., 2014).

Белок HE (гликопротеин гемагглютинин-эстеразы) входит в структуру только бетакоронавирусов. Часть белка HE на поверхности клетки-хозяина связывается с сиаловой кислотой и, вероятно, способствует начальной адсорбции вируса мембраной. Эстераза отщепляет ацетильные группы от сиаловой кислоты. Гены HE коронавирусов имеют гомологичные последовательности с гликопротеином HE вируса гриппа и способны отображать раннюю рекомбинацию между двумя вирусами (Klausegger et al., 1999).

1.1.3. Гаммакоронавирусы

Типичный гаммакоронавирус — коронавирус птичьего инфекционного бронхита (IBV). Он имеет два дополнительных гена, называющиеся 3 и 5, которые в итоге транслируются в четыре дополнительных белка. Ген 3 имеет три открытых рамки считывания, кодирующие белки 3a, 3b и 3c, который является белком оболочки (E). Ген 5 транслирует белки 5a и 5b. Их функции пока остаются неизвестными (Cavanagh et al., 2007).

1.1.4. Дельтакоронавирусы

Типичный представитель — дельтакоронавирус свиней (PdCoV). Подобно другим дельтакоронавирусам, PDCoV кодирует два специфических дополнительных гена, а именно NS6, расположенный между генами M и N, и NS7, который является альтернативной рамкой считывания гена N. Из

недавних исследований известно, что NS7 не требуется для эффективной репликации в клетках, но вирусы, лишенные функционального NS6, демонстрирует заметно более низкую эффективность репликации (Zhang et al., 2019).

1.2. Эволюция коронавирусов

Первый обнаруженный коронавирус, вирус инфекционного бронхита (IBV), был выделен из куриных эмбрионов в 1937 году (Beaudette et al., 1937). За ним последовало открытие вируса гепатита мышей (MHV) и коронавирусов других млекопитающих в 1940-х годах (Cheever et al., 1949; Doyle et.al., 1946). Первыми обнаруженными коронавирусами человека являются 229E (HCoV-229E) и OC43 (HCoV-OC43), которые были обнаружены в 1960-х годах (Tyrrell et.al., 1966; Hamre et al., 1966). HCoV-229E — это альфакоронавирус, возникший у летучих мышей и передавшийся человеку через альпак (Corman et al., 2015), а OC43 — бетакоронавирус, передавшийся человеку от грызунов через крупнорогатый скот (Corman et al., 2018). После эпидемии атипичной пневмонии 2002–2003 гг. интерес к коронавирусам возобновился и позволил открыть еще два вируса, поражающих человека, — альфакоронавирус HCoV-NL63 и бетакоронавирус HCoV-HKU1, полученные от летучих мышей и грызунов, соответственно (Tao et al., 2017). Все эти четыре вируса обычно вызывают легкие респираторные симптомы у иммунокомпетентных пациентов.

SARS-CoV и MERS-CoV — это два неродственных бетакоронавируса, которые произошли от летучих мышей и передаются людям от диких плотоядных животных и верблюдов, соответственно. В отличие от других HCoV, эти два вируса проявляли повышенную вирулентность, вызывая тяжелую пневмонию и даже смерть пораженных людей, при этом уровень смертности составлял около 10% и 30%, соответственно (Guarner et al., 2020).

Возникает вопрос, как коронавирусы переходят межвидовой барьер? В результате транскрипции вирусной РНК синтезируются как геномные (полноразмерные транскрипты), так и субгеномные (усеченные) РНК.

Субгеномные РНК служат мРНК для структурных и вспомогательных генов, которые расположены ниже полипротеинов репликазы. Все субгеномные плюс-цепи РНК образуют набор вложенных РНК, что является отличительной чертой порядка *Nidovirales* (Sethna et al., 1991).

Коронавирусы характеризуются исключительной генетической пластичностью и быстро развиваются, изменяя свой антигенный профиль, тропизм тканей или круг хозяев с помощью двух различных механизмов. Вирусная репликаза (РНК-зависимая РНК-полимераза) может работать с ошибками, поэтому включение неправильных нуклеотидов в каждом цикле репликации и последующее накопление мутаций в вирусном геноме приводит к дифференциации вирусного потомства от родительского штамма. Этот механизм хорошо известен для вирусов гриппа и ответственен, за так называемый, антигенный дрейф, который может вызывать постепенную адаптацию поверхностей белков вируса к рецепторам клеток новых видов животных, повышая приспособленность вируса. К тому же, особый механизм репликации коронавирусов способствует рекомбинации из-за присутствия субгеномных РНК. В случае заражения более чем одним штаммом коронавируса, РНК-полимераза может перескакивать с РНК одного штамма на РНК другого штамма, синтезируя гибридную РНК, содержащую последовательности от обоих вирусов. Рекомбинация может происходить не только с геномными последовательностями других коронавирусов (гомологичная рекомбинация), но также с РНК разных вирусов и других организмов (гетерологичная рекомбинация) (Banner et al., 1991; Lai et al., 1996; Zeng et al., 2008).

Рекомбинация — это альтернативный механизм, который позволяет коронавирусам приобретать новые биологические свойства с точки зрения вирулентности, диапазона хозяев и тропизма тканей. Штаммы коронавируса, которые являются непатогенными или низкопатогенными для исходного хозяина, могут повышать свою патогенность у тех же видов или адаптироваться

к разным видам, распространятся в новом хозяине с исключительной скоростью (Banner et al., 1991).

1.2.1. Коронавирусы птиц

С момента появления SARS-CoV в 2002 году возрос интерес к коронавирусам у других видов, включая птиц. До этого информация о коронавирусах птиц была ограничена в основном исследованием трех родов отряда *Galliformes*. Были изучены вирус инфекционного бронхита (IBV) у дикой банкивской курицы (*Gallus gallus*), а также коронавирусы TCoV и PhCoV у индейки (род *Meleagris*) и фазана (род *Phasianidae*), соответственно (Decaro et al., 2020).

Этот изменилось после обнаружения нескольких новых коронавирусов с высоким генетическим разнообразием у разных видов птиц. Все эти вирусы, а также IBV-подобные коронавирусы, обнаруженные у других птиц, включая пингвинов, голубей, павлина, попугаев, были отнесены к одному и тому же вирусному виду, известному как птичий коронавирус (ACoV), принадлежащий к роду гаммакоронавирус (Cavanagh et al., 2002; Liu et al., 2005; Hughes et al., 2009; Circella et al., 2007; Torres et al., 2013; Domanska-Blicharz et al., 2014; Decaro et al., 2020).

Исторически все коронавирусы птиц относились к роду гаммакоронавирусов и, в свою очередь, все коронавирусы, принадлежащие к этому роду, были идентифицированы только у птиц. Однако это предположение было опровергнуто свидетельством принадлежности к роду гаммакоронавирусов коронавируса белухи, впервые обнаруженного в 2008 году (*Beluga whale coronavirus SW1*) (Mihindukulasuriya et al., 2008), и трех новых коронавирусов: BuCoV HKU11 у птиц отряда воробьинообразные (*Passeriformes),* ThCoV HKU12 у дроздов (*Turdidae*) и MuCoV HKU13 у мунии (*Lonchura punctulate*), которые филогенетически не группировались с существующими коронавирусами, идентифицированными у птиц. Эти три

вируса отличались от известных гаммакоронавирусов, образуя уникальный кластер в филогенетическом дереве, который послужил основой для создания рода дельтакоронавирус (Woo et al., 2009; Chu et al., 2011; Torres et al., 2016). Эти вирусы объединяются в кластер с ранее неклассифицированными коронавирусами, обнаруженными у различных азиатских хищников, например, у азиатской леопардовой кошки (*Prionailurus bengalensis*) и енотовидной собаки (*Nyctereutes procyonoides*) (Dong et al., 2007).

1.2.2. Коронавирусы летучих мышей

Первые доказательства, что летучие мыши также могут переносить коронавирусы, были опубликованы в 2005 году (Poon et al., 2005). На сегодняшний день у летучих мышей идентифицировано более 200 новых коронавирусов (Chen et al., 2014). Лишь небольшая часть этих коронавирусов была официально признана ICTV (Международный комитет по таксономии вирусов); многие другие все еще ожидают официального определения.

Близкородственные коронавирусы могут быть обнаружены у одних и тех же видов летучих мышей, живущих в местах, разделенных тысячами милями (Drexler et al., 2010), а разные виды или роды коронавирусов могут быть обнаружены у разных видов летучих мышей, живущих на одних и тех же местах ночевок. Также было показано, что некоторые коронавирусы являются видоспецифичными, однако, у летучих мышей, коронавирусы, в отличие от других вирусов и бактерий, не вызывают явных заболеваний, в том числе и при направленном заражении (Mühldorfer et al., 2011). Этот феномен, по-видимому, связан с особенностями их иммунной системы (Ahn et al., 2019; Brook et al., 2020).

Основываясь на имеющихся геномных данных, было показано, что, хотя птицы представляют собой резервуар для коронавирусов, принадлежащих к родам гамма- и дельтакоронавирусов, летучие мыши являются естественным резервуаром для альфа- и бетакоронавирусов. Коронавирусы с высокой

частотой выявлялись у летучих мышей на всех континентах, причем альфакоронавирусы более распространены, чем бетакоронавирусы (Wong et al., 2019).

Есть предположения, что многие человеческие вирусы взяли начало от коронавирусов летучих мышей. Вирусный штамм BtKYNL63-9b, идентифицированный в 2010 г. у летучих мышей *Triaenops afer* из Кении являются родственным группе NL63 *(NL63-related bat coronavirus strain BtKYNL63-9b),* в которую входит коронавирус человека NL63 (Tao et al., 2017). В связи с этим было предположено, что два вируса человека HCoV-229E и HCoV-NL63, принадлежащих к альфакоронавирусам и вызывающих легкие респираторные симптомы у иммунокомпетентных людей, также произошли от летучих мышей.

В качестве прямого предка HCoV-229E признается альфакоронавирус альпак, который, в свою очередь, происходит от вирусов связанных с 229E-CoV, выявленных у подковогубых летучих мышей (Corman et al., 2015). HCoV-NL63, вероятно, является рекомбинантным вирусом, происходящим из отдаленно родственных 229E-CoV, связанных с подковогубыми летучими мышами, и коронавирусом, связанным с летучими мышами *Triaenops afer* (Tao et al., 2017). Белок S HCoV-NL63 более тесно связан с белком S 229E-CoV, тогда как остальная часть генома идентична штамму BtKYNL63-9b коронавируса летучих мышей (Tao et al., 2017). Анализ генома респираторного коронавируса альпаки показал, что этот вирус был тесно связан с альфа-коронавирусом HCoV-229E (Crossley et al., 2012). Совсем недавно для HCoV-229E были обнаружены близкородственные виды вирусов у африканских подковогубых летучих мышей. Интересно, что и у вирусов летучих мышей, и у вирусов альпаки обнаружен интактный вспомогательный ген ORF8, расположенный на 3'-конце генома, в то время как HCoV-229E сохранил только консервативный участок последовательности, регулирующей транскрипцию, предшествующий остаткам этой ORF, что указывает на его потерю после

приобретения человеком коронавируса 229E. Следовательно, HCoV-229E, вероятно, является потомком коронавируса альпаки (Corman et al., 2015).

В 2002 году, в начале эпидемии атипичной пневмонии, почти все первые пациенты были в контакте с животными на рынке в провинции Гуандун до того, как заболели. После того, как SARS-CoV был идентифицирован, его РНК и/или специфические антитела были обнаружены у маскированных пальмовых циветт (*Paguma larvata*) и у заводчиков животных на рынке. Однако более поздние исследования о пойманных в дикой природе циветтах показали, что штаммы SARS-CoV, обнаруженные в циветтах с рынка, передавались им от других диких животных (Tu et al., 2004; Kan et al., 2005). Впоследствии новые коронавирусы, связанные с SARS-CoV человека (SARS-rCoVs), были обнаружены у подковообразных летучих мышей (род *Rhinolophus*) в Китае и Гонконге (Lau et al., 2005; Li et al., 2005). Эти SARS-rCoVs продемонстрировали идентичность геномных последовательностей на 88–90% между собой и на 87–92% идентичности с изолятами SARS-CoV человека или циветты. Также SARS-rCoV были обнаружены у *Rhinolophus spp*. летучих мышей из других регионов Китая (Tang et al., 2006; Woo et al., 2006).

Эти данные свидетельствуют о том, что летучие мыши могут быть естественными хозяевами SARS-CoV, а дикие плотоядные животные были лишь промежуточными хозяевами. Хотя эти SARS-CoV-подобные вирусы продемонстрировали высокую идентичность последовательности с SARS-CoV, было продемонстрировано, что они неспособны связываться с рецептором ангиотензинпревращающего фермента II (ACE2) клеток человека в результате делеций в их S белке (Ren et al., 2008). Кроме того, теория происхождения SARS-CoV от летучих мышей не получила убедительной поддержки из-за невозможности прямого выделения этого вируса от летучих мышей. Таким образом, учитывая, что прямых предшественников SARS-CoV у летучих мышей обнаружено не было и что рекомбинация РНК является топливом для эволюции коронавирусов, было высказано предположение, что SARS-CoV возник в результате рекомбинации SARS-CoV-подобных вирусов летучих

мышей. Эта гипотеза была выдвинута после доказательства наличия единственной пещеры летучих мышей в Юньнани, Китай, с очень высоким разнообразием коронавирусов в этой популяции, которые содержали все необходимые для формирования SARS-CoV генетические элементы (Ge et al., 2013).

Анализ рекомбинации также убедительно подтвердил гипотезу о том, что SARS-CoV штамм SZ3 циветт возник в результате рекомбинации двух существующих штаммов летучих мышей, WIV16 и Rf4092 (Hu et al., 2017). Наиболее частые точки нарушения рекомбинации были внутри гена S и ORF8, который кодирует вспомогательный белок. Эти гены также участвовали в важнейших путях адаптации SARS-CoV от летучих мышей к диким хищникам, от диких хищников к людям и от человека к человеку (Cui et al., 2019). Было показано, что WIV16 обладает способностью связываться с рецептором ACE2 человека, циветты и летучей мыши (Ge et al., 2013). Выделение в культуре клеток штамма SARS-CoV с высокой степенью родства в сочетании с доказательством наличия функционального белка S, способного использовать тот же рецептор ACE2, предоставило надежные и убедительные доказательства происхождения SARS-CoV от летучих мышей.

Напротив, коронавирус летучей мыши, тесно связанный с MERS-CoV человека, не был обнаружен. Действительно, геномные последовательности MERS-CoV человека обладают лишь приблизительно 65–80% нуклеотидной идентичностью с таковыми у других представителей подрода *Merbecovirus* от разных летучих мышей. Вместо этого человеческие MERS-CoV были почти идентичны MERS-CoV, выявленным у одногорбых верблюдов (*Camelus dromedaries*).В последнее время анализ геномной последовательности показал, что коронавирусы, теперь принадлежащие к виду *MERS-rCoV,* были обнаружены у нескольких видов летучих мышей из двух семейств *Vespertilionidae* и *Nycteridae* (Lelli et al., 2013; Corman et al., 2014). Однако ни один из этих MERS-rCoV не является прямым предшественником MERS-CoV, поскольку их S-белки существенно отличаются от белков человеческого

вируса. Ближайшим родственником MERS-CoV человека и верблюда является штамм MERS-rCoV Neoromicia/5038, выделенный из летучих мышей *Neoromicia capensis* в Южной Африке (Geldenhuys et al., 2018). Короткая последовательность (около 200 нуклеотидов) вирусной РНК, идентичная таковой у MERS-CoV, была также обнаружена у летучих мышей *Taphozous perforates* в Саудовской Аравии (Memish et al., 2013).

SARS-CoV-2 сравнивался с существующими видам сотен известных вирусов, в основном изолированных от летучих мышей. Важно отметить, было подтверждено, что SARS-CoV-2 использует рецептор ACE2 через рецептор-связывающий домен (RBD) белка S (Hoffmann et al., 2020; Zhou et al., 2020). Скорее всего, SARS-CoV-2 также пошел от летучих мышей. Согласно доступным к настоящему времени последовательностям генома, наиболее близким вирусом (96,2% идентичности нуклеотидной последовательности) с SARS-CoV-2 является штамм BatCoV RaTG13, идентифицированный от летучей мыши *Rhinolophus affinis*, из провинции Юньнань, Китай (Tang et al., 2020). Рецептор-связывающий спайковый белок SARS-CoV-2 сильно отличается от других коронавирусов (с менее чем 75% идентичностью нуклеотидной последовательности со всеми ранее описанными SARS-rCoV, за исключением 93,1% нуклеотидной идентичности с BatCoV RaTG13) (Zhou et al., 2020). Хотя SARS-CoV-2 использует рецептор ACE2, пять из шести критических аминокислотных остатков в RBD различались между SARS-CoV-2 и SARS-CoV; вместо этого те же остатки были идентичны остаткам SARS-rCoV панголина, и, в свою очередь, только один из этих остатков был идентичен остаткам BatCoV RaTG13 (Tang et al., 2020), хотя последний показывает наивысшую идентичность нуклеотидной последовательности с SARS-CoV-2 по всему геному. Таким образом, было логично предположить, что область RBD SARS-CoV-2 могла возникнуть в результате недавнего события рекомбинации у ящеров или что SARS-CoV-2 и SARS-rCoV у панголинов представляют собой результат случайной эволюции (Tang et al., 2020; Lam et al., 2020). В целом, еще предстоит решить, понадобился ли SARS-CoV-2 промежуточный хозяин,

прежде чем он смог заразить людей, как это было в случае SARS-CoV и других HCoV.

1.2.3. Коронавирусы свиней

Наиболее известными коронавирусам свиней считаются вирус трансмиссивного гастроэнтерита свиней (TGEV), вирус эпидемической диареи свиней (PEDV) и SADS-CoV, относящиеся к роду альфакоронавирусов, бетакоронавирус гемагглютинирующего энцефаломиелита свиней (PHEV) и один дельтакоронавирус свиней (PDCoV).

TGEV был впервые описан в Великобритании в 1950-х годах и представляет собой самый старый из известных коронавирусов свиней. TGEV тесно связан с собачьим коронавирусом (CCoV) и кошачьим коронавирусом (FCoV), образуя с этими коронавирусами уникальный вид, называемый *Alphacoronavirus-1* (Lorusso et al., 2008).

PEDV был обнаружен в популяции свиней в 1970-х годах, вероятно, как следствие вторичного распространения от летучих мышей. PEDV более тесно связан с коронавирусом летучих мышей *Scotophilus bat coronavirus 512,* чем с другими известными альфа-коронавирусами, включая TGEV и альфа-коронавирусы человека (HCoV-229E и HCoV-NL63). Следовательно, PEDV и *Scotophilus bat coronavirus 512*, вероятно, имеют общего эволюционного предка, и межвидовая передача вируса могла происходить между летучими мышами и свиньями (Banerjee et al., 2019).

SADV-CoV, теперь называемый кишечным альфакоронавирусом свиней (SeACoV), также произошел от летучих мышей и имеет 86% идентичность последовательности с альфакоронавирусом летучих мышей HKU2-CoV (Zhou et al., 2018).

Энцефаломиелит поросят (PHEV), который был впервые описан в 1957 году в Онтарио, Канада, его эволюционная история тесно связана с двумя другими близкородственными бетакоронавирусами, HCoV-OC43 и самым

старым из известных − коронавирус кропнорогатого скота BCoV, с которым PHEV может иметь общих предков (Vijgen et al., 2006) и включен в один и тот же вирусный вид *Betacoronavirus-1* (Corman et al., 2018). Скорее всего, HCoV-OC43 и PHEV произошли от бетакоронавируса грызунов в результате предварительной адаптации к BCoV (Corman et al., 2018).

Дельтакоронавирус PDCoV был обнаружен в 2012 году в Гонконге во время молекулярного эпиднадзора за коронавирусами у птиц и млекопитающих. Наиболее близкий родственник PDCoV был идентифицирован у перепела − дельтакоронавирус UAE-HKU30, и было предложено, что этот вирус был рекомбинантным между двумя другими дельтакоронавирусами птиц, коронавирусом воробья HKU15 и коронавирусом соловья HKU11. Все эти дельтакоронавирусы теперь принадлежат к одному виду Coronavirus HKU15 (Lau et al., 2018).

Было обнаружено, что свиньи подвержены экспериментальному заражению бетакоронавирусом MERS-CoV (Vergara-Alert et al., 2017), а РНК SARS-CoV была обнаружена у свиней и диких кабанов (Chen W et al., 2005; Wang et al., 2005) Напротив, недавнее экспериментальное заражение продемонстрировало, что свиньи не восприимчивы к SARS-CoV-2 (Shi et al., 2020).

1.2.4. Коронавирусы грызунов

По аналогии с летучими мышами, грызуны играют значительную роль в эволюции коронавирусов, в частности тех, которые принадлежат к подроду *Embecovirus* рода бетакоронавирус. На протяжении десятилетий только один вид коронавируса − мышиный коронавирус был связан с грызунами. Прототип вируса, который был назван вирусом гепатита мышей (MHV), был впервые выделен у мышей в 1949 г. (Cheever et al., 1949). Вариант MHV у крыс был идентифицирова в 1970 году (Parker et al., 1970). Коронавирус крыс (RCoV) вызывает эпидемии респираторных заболеваний в колониях лабораторных

крыс. Двумя прототипами штамма RCoV являются вирус сиалодакриоаденита (SDAV) и коронавирус крыс Паркера (PRC) (Bhatt et al., 1972; Parker et al., 1970).

Роль грызунов в эволюции коронавирусов, принадлежащих к эмбековирусам, была недавно подчеркнута посредством открытия нового бета-коронавируса у норвежских крыс (*Rattus norvegicus*) в Китае. Этот вирус образует отдельный вид под названием *China Rattus coronavirus HKU24* (ChRCoV HKU24). Этот вирус филогенетически отличался от MHV и HCoV-HKU1 и обладал характеристиками генома, которые были промежуточными между вирусом крупнорогатого скота (BCoV) и MHV. Следовательно, ChRCoV HKU24 мог послужить основой для происхождения BCoV, а грызуны, вероятно, являются важным резервуаром для предков подрода *Embecovirus* (Lau et al., 2015).

Репликация SARS-CoV была изучена на мышах, сирийских золотых и китайских хомяках. Была создана модель заражения SARS-CoV на старых мышах (Gretebeck et al., 2015), для этого были разработаны трансгенные мыши, экспрессирующие ACE2 человека. Некоторые животные модели были протестированы и проанализированы на геномном и протеомном уровне для изучения патогенеза SARS-CoV. Следовательно, есть основания полагать, что такие модели будут работать и для SARS-CoV-2. Напротив, исследования показали, что мыши, морские свинки и хомяки не восприимчивы к инфекции MERS-CoV главным образом потому, что их гомологичные молекулы DPP4 (дипептидилпептидаза-4) не функционируют как рецепторы для проникновения MERS-CoV (Cockrell et al., 2014). Первая модель инфекции MERS на мышах, о которой сообщалось в 2014 году, включала трансдукцию животных рекомбинантным аденовирусом 5, кодирующим молекулы DPP4 человека (hDPP4), интраназально, что привело к репликации MERS-CoV в легких (Song et al., 2019).

1.2.5. Коронавирусы жвачных животных

Самым старым известным коронавирусом жвачных животных является коронавирус крупнорогатого скота (BCoV), который также является прототипом вида *Betacoronavirus-1*. BCoV является образцом того, как коронавирусы могут преодолевать межвидовые барьеры, создавая отдельные вирусные линии, поражающие дыхательные и/или кишечные тракты человека (HCoV-OC43), свиней (PHEV), лошадей (коронавирус лошадей, ECoV), и собак (респираторный коронавирус собак, CRCoV).

У одногорбых верблюдов, восприимчивых к инфекции MERS-CoV, развиваются бессимптомные инфекции или легкие заболевания верхних дыхательных путей, поэтому они считаются естественными хозяевами MERS-CoV (Hemida et al., 2017). Некоторые высказываются о том, что инфицированные верблюды многократно заносили вирус в человеческую популяцию (Hemida et al., 2017). Хотя другие исследования исключили восприимчивость других домашних жвачных животных к MERS-CoV (Reusken et al., 2013; Adney et al., 2016), недавнее исследование обнаружило специфические антитела и РНК в сыворотке и носовых секретах, домашних жвачных животных выращены в Африке, включая овец, коз и крупный рогатый скот (Kandeil et al., 2019).

1.2.6. Коронавирусы хищников

Некоторые домашние и дикие плотоядные животные также восприимчивы к инфекции SARS-CoV. Хотя потенциальными естественными резервуарами являются подковообразные летучие мыши, было обнаружено, что штаммы SARS-подобных коронавирусов широко распространены у маскированных пальмовых циветт (*Paguma larvata*) и енотовидных собак, которые предположительно являются его промежуточными хозяевами (Guan et al., 2003). Анализ последовательности вируса SARS-CoV маскированных пальмовых циветт показал, что они были высоко гомологичны SARS-CoV

человека с нуклеотидной идентичностью более 99,6%, что указывает на то, что вирус не циркулировал в популяции маскированных пальмовых циветт в течение длительного времени (Shi et al., 2008).

Среди плотоядных животных SARS-CoV-2 способен заражать кошек, хорьков и, в меньшей степени, собак (Shi et al., 2020).

Резюмируя, коронавирусы известны ветеринарии на протяжении многих десятилетий; некоторые из этих вирусов, такие как IBV, TGEV, BCoV и другие могут вызывать заболевания, которые имеют большое влияние на сельскохозяйственную промышленность. Другие коронавирусы, а именно FIPV, FRSCV и MHV, вызывают тяжелые заболевания у домашних (кошки, хорьки) или лабораторных (мыши) животных. Коронавирусы животных показывают, как коронавирусы эволюционируют посредством накопления точечных мутаций и гомологичной (и гетерологичной) рекомбинации, генерируя различные генотипы и патотипы. Эволюция коронавирусов может привести к изменению диапазона хозяев с одного вида животных на другой или с животных на людей. Первое событие хорошо задокументировано в ветеринарии, поскольку множество вирусов происходит от IBV и BCoV, которые адаптировались к различным видам животных. Однако наиболее интересным сценарием является скачок и дальнейшая адаптация коронавирусов животных к человеку. Появляется все больше свидетельств того, что все известные в настоящее время человеческие коронавирусы происходят от коронавирусов животных, причем коронавирусы летучих мышей или грызунов являются наиболее вероятными предками. В большинстве случаев предполагается, что другие млекопитающие служат промежуточными хозяевами до окончательной адаптации к человеку, то есть альпаки и крупнорогатый скот для низкопатогенных HCoV-229E и HCoV-OC43 соответственно, а также дикие хищники и одногорбые верблюды для высокопатогенных SARS-CoV и MERS-CoV, соответственно. Два других вируса HCoV-NL63 и HCoV-HKU1, вероятно, произошли от летучих мышей и грызунов соответственно, но в настоящее время неизвестно, требовалось ли для

этой передачи промежуточное млекопитающее-хозяин. Происхождение SARS-CoV-2 должно быть зоонозным, поскольку у летучих мышей были обнаружены последовательности, с высокой степенью идентичности, но окончательный промежуточный хозяин до сих пор не идентифицирован.

1.3. Методы филогенетического анализа

1.3.1. Методы выравнивания

Выравнивание последовательностей ДНК является предварительным условием практически для всех сравнительных геномных анализов, включая идентификацию консервативных мотивов последовательностей, оценку эволюционного расхождения между последовательностями и исторических взаимосвязей между генами и видами. Ошибки в выравнивании последовательностей оказывают значительное негативное влияние на последующий вывод о дивергенции последовательностей и о филогенетических деревьях.

Есть два типа выравнивания: глобальное и локальное. В случае глобального выравнивания две или более последовательности выравниваются от начала и до конца. Это подходит для последовательностей генов, кодирующих белки, для коротких участков геномной последовательности и в целом для очень схожих последовательностей примерно одной длины. Для более длинной сложной геномной ДНК необходимо учитывать перестройки, вставки и делеции в последовательностях, что требует создания локальных выравниваний. Локальное выравнивание отличается от глобального выравнивания тем, что первое фокусируется на общих областях с высоким сходством, игнорируя области, которые не демонстрируют высокую гомологию последовательностей между последовательностями (Brudno et al., 2003).

В данной работе важным шагом для получения филогенетических деревьев является построение множественного глобального выравнивания последовательностей (MSA). Большинство методов, обычно используемых в

настоящее время, основано на прогрессивном выравнивании. Этот популярный алгоритм MSA представляет собой простую агломерационную процедуру. Последовательности сначала попарно сравниваются друг с другом, чтобы заполнить матрицу расстояний, содержащую процент идентичности. Затем к этой матрице расстояний применяется алгоритм кластеризации (UPGMA или NJ) для создания корневого направляющего дерева. Алгоритм агломерации следует топологии дерева и проходит путь от листа к корню, выравнивая попарно каждую пару последовательностей связанную с каждым встреченным узлом. Процедура может применяться с использованием любого алгоритма, способного выровнять две последовательности или два выравнивания. В большинстве пакетов это алгоритм Нидлмана-Вунша (Needleman and Wunsch, 1970) или алгоритм Витерби (Durbin et al., 1998).

ClustalW (Thompson et al., 1994) часто считается прообразом прогрессивного выравнивания. Главной его проблемой считается наследование ошибки первого шага, то есть в случае, если возникла какая-то ошибка в самом начале при первом попарном выравнивании, это ошибка сохраниться и при добавлении последовательностей для выравнивания, потому как оптимальное выравнивание для новой группы последовательностей не ищется, что в конечном итоге может привести к совершенно неправильному результату. Это хорошо известная проблема, обычно решается с помощью итеративной стратегии. В итеративной схеме группы последовательностей перестраиваются определенное количество раз, используя либо случайные разбиения, либо разбиения, предложенные направляющим деревом. Самые сложные итеративные стратегии (Muscle и PRRP), включают два вложенных итерационных цикла: внутренний, в котором выравнивание оптимизируется по отношению к направляющему дереву, и внешний, в котором текущее MSA используется для повторной оценки направляющего дерева. Процедура продолжается до тех пор, пока выравнивание и направляющее дерево не сойдутся. Было показано, что эти итерации почти всегда улучшают точность MSA (Wallace et al., 2005).

Программные инструменты для выравнивания последовательностей, такие как BLAST (Altschul et al., 1990) и CLUSTAL (Thompson et al., 1994), являются наиболее широко используемыми методами биоинформатики. Хотя подходы на основе выравнивания обычно остаются эталоном, методы на основе MSA сложно использовать на очень больших наборах данных, которые доступны сегодня (Chan et al., 2013). Кроме того, методы, основанные на выравнивании, оказались неточными в случаях с низкой идентичностью последовательностей, например, регуляторные последовательности генов (Kantorovitz et al., 2007; Ivan et al., 2008) и отдаленно родственные гомологи белков (Zielezinski et al., 2017; Vinga et al., 2004). Более того, алгоритмы выравнивания предполагают, что линейный порядок гомологии сохраняется в сравниваемых последовательностях, поэтому эти алгоритмы не могут применяться напрямую, если есть перестройки последовательностей (например, рекомбинация и замены белковых доменов (Terrapon et al., 2014) или горизонтальный перенос (Cong et al., 2016), в случае, когда обрабатываются крупные наборы данных о последовательностях, например, для полногеномной филогенетики (Ondov et al., 2016)). Кроме того, выравнивание двух длинных последовательностей ДНК (длиной в миллионы нуклеотидов) на практике невозможно. Поэтому в качестве альтернативы выравниванию последовательностей было разработано множество так называемых подходов к анализу последовательностей без выравнивания (AF) (Zielezinski et al., 2017).

Был разработан широкий спектр методов без выравнивания, направленный на сравнение последовательностей. Эти подходы включают методы, основанные на подсчете слов или *k*-мер, длине общих подстрок, микро-выравниваниях, преобразования Фурье, теории информации и системы повторяющихся функций (Zielezinski et al., 2017). В настоящее время наиболее широко используемый подход AF − методы основанные на *k*-мер (Luczak et al., 2019). Эти методы очень разнообразны и обеспечивают множество статистических показателей, которые реализуются в разных программных инструментах (Lu et al., 2017; Chan et al., 2014).

1.3.2. Методы построения филогенетических деревьев

Филогенетика – это область биологии, изучающая эволюционные взаимоотношения между биологическими объектами (видами, индивидуумами или генами). Филогенетический анализ применяется для выявления взаимоотношений между таксонами (данными, объектами), а на основе его результатов строится кладограмма или дерево.

Филогенетическое дерево – схема, отражающая эволюционные взаимосвязи между различными видами или другими объектами, имеющими общего предка. Филогенетическое дерево состоит из листьев, узлов и корня (максимум один). Листья — это конечные вершины, т.е. те, в которые входят ровно по одному ребру; каждый лист отображает определенный объект. Каждый узел представляет эволюционное событие: разделение предкового вида на два или более, которые в дальнейшем эволюционировали независимо. Корень представляет общего предка всех рассматриваемых объектов. Ветви — рёбра филогенетического дерева; взаимное расположение ветвей называется топологией.

Важным этапом перед построением дерева выбрать оптимальную модель эволюции нуклеотидов. Модели эволюции нуклеотидов описывают скорость изменения фиксированных мутаций среди последовательностей и составляют основу эволюционного анализа генетических данных на молекулярном уровне.

Первая и самая простая модель, имитирующая процесс замены нуклеотидов в ДНК, была описана Джуксом и Кантором (JC) в 1969 году (Jukes and Cantor, 1969). Эта модель предполагает равные частоты нуклеотидов и равные частоты замены нуклеотидов (мутации). Однако изменения между основаниями с одинаковой химической структурой (транзиции) более распространены, чем изменения между основаниями с разной химической структурой (трансверсии), потому что замена аналогичной структуры более

вероятна с точки зрения молекулярной эволюции. Более того, генетический код допускает больше транзиций, чем трансверсий без замены аминокислы (Kimura, 1980; Collins and Jukes, 1994). Руководствуясь этими доказательствами, Кимура в 1980 году (Kimura, 1980) представил модель с двумя параметрами (K80), где скорости замены нуклеотидов различаются между транзициями и трансверсиями. Точно так же Фельзенштейн в 1981 году (F81) (Felsenstein, 1981) расширил модель JC, включив в нее разные частоты нуклеотидов, которые также могут возникать как следствие физико-химических свойств нуклеотидов и действия естественного отбора. Позже был разработан ряд моделей, включающих расширения этих исходных моделей (например, HKY (Hasegawa, 1985) и SYM (Zharkikh, 1994)). Следуя этой тенденции, самая сложная модель − General time reversible (GTR)(Tavaré, 1986), включает разные скорости для каждого типа замены и разные частоты нуклеотидов. Кроме того, доля неизменных сайтов (+ I) (Shoemaker and Fitch, 1989) и/или скорость вариации между сайтами (+ G) (Yang, 1994) могут быть включены в любую модель.

Стационарные, обратимые и гомогенные модели замещения ДНК, полученные из всех возможных комбинаций (одинаковые/разные скорости замены нуклеотидов, равные/неравные частоты нуклеотидов, с/без +G и/или +I; более 1600 моделей) уже определены и в настоящее время внедрены в некоторые филогенетические программы (Darriba et al., 2012; Arenas and Posada, 2014). Однако, несмотря на большое количество доступных моделей замещения ДНК, модель GTR + G + I обычно лучше соответствует реальным данным, чем другие (более простые) альтернативные модели (Sumner et al., 2012), предполагая, что эволюционный процесс очень сложен. Но выбор самой сложной модели замещения может и улучшить реальные данные (Jayaswal et al., 2011).

Чтобы улучшить соответствие реальным данным, современные тенденции в развитии моделей замещения включают необратимые (асимметричные) и нестационарные (нуклеотидный состав может меняться со

временем) матрицы (Jayaswal et al., 2011; Boussau and Gouy, 2006) или даже учитывать взаимодействия между соседями (Lunter and Hein, 2004), которые могут привести к более точным филогенетическим выводам (Boussau and Gouy, 2006; Kaehler et al., 2015). Однако эти модели еще не реализованы в наиболее популярных пакетах филогенетического программного обеспечения из-за их сложности.

После того как определена модель замен нуклеотидов можно переходить к построению деревьев. Есть 2 типа методов построения деревьев: дистанционные (UPGMA, NJ) и дискретные (MP,ML и байесовский подход).

Самыми популярными дистанционными методами являются UPGMA (метод невзвешенного попарного среднего) и NJ (метод присоединения ближайшего соседа) методы.

UPGMA − это простой агломеративный метод иерархической кластеризации, то есть дерево собирается от листьев к корню. Это самый простой и быстрый метод построения корневого ультраметрического филогенетического дерева. Однако главный недостаток метода − предположение об одинаковой скорости эволюции для всех линий. Это означает, что скорость мутаций в этих линиях постоянна во времени. Кроме того, следуя этому методу, все ветви дерева имеют одинаковую длину. Поскольку трудно иметь одинаковую частоту мутаций для всех ветвей, в действительности метод UPGMA чаще генерирует ненадежные топологии деревьев. Кроме того, метод UPGMA начинается с матрицы попарных расстояний. Первоначально предполагается, что каждый вид представляет собой отдельный кластер. Затем он объединяет два ближайших кластера с наименьшим значением расстояния в матрице расстояний. После он пересчитывает расстояние пары в узле, взяв среднее значение. Затем алгоритм повторяет процесс, пока все виды не будут объединены в один кластер (Sokal and Michener, 1958).

Метод присоединения ближайшего соседа (Neighbor-joining; NJ) – это кластерный метод, не требующий, чтобы данные были ультраметрическими.

Для осуществления данного метода нужна матрица расстояний, задающая расстояния между каждой парой объектов. Алгоритм начинает с того, что строит полностью неразрешенное дерево (все ветви отходят от одного центрального узла). Два объекта объединяются в один узел, если расстояние между ними минимально. После рассчитывается расстояние до нового узла и расстояние от этого нового узла до каждого из оставшихся объектов. После матрица перестраивается, объекты, вошедшие в новый узел, удаляются, и добавляется новая вершина, то есть расстояние от этого нового узла до всех объектов. Эти шаги повторяются до тех пор, пока не станут известны длины всех ветвей. Метод кластеризации, используемый этим алгоритмом, сильно отличается от UPGMA, поскольку он не пытается кластеризовать наиболее тесно связанные объекты, скорее, он минимизирует длину всех внутренних ветвей и, следовательно, длину всего дерева (Saitou and Nei, 1987).

Метод максимальной экономии (Maximum parsimony; MP) – это дискретный метод, основная идея которого состоит в том, что предпочтительно самое простое объяснение данных, поскольку оно требует наименьшего количества предположений. По этому критерию MP дерево − это дерево с наименьшим количеством замен/эволюционных изменений для всех последовательностей, происходящих от общего предка. Для каждого сайта выравнивания оцениваются все возможные деревья, и присваивается оценка, основанная на количестве эволюционных изменений, необходимых для получения наблюдаемых изменений последовательности. Таким образом, лучшим деревом является то, которое минимизирует общее количество мутаций на всех участках. Однако этот метод дает мало информации о длине ветвей и сильно страдает от притяжения длинных ветвей, то есть длинные ветви искусственно связываются из-за накопления неоднородных сходств, даже если они вовсе не филогенетически связаны. К тому же MP метод дает более одного дерева с одинаковым счетом (Mount, 2008).

Метод максимального правдоподобия (Maximum likelihood; ML) аналогичен методу MP в том, что он исследует различные топологии деревьев и

оценивает относительную поддержку путем суммирования по всем позициям последовательности. Алгоритмы ML ищут дерево, которое максимизирует вероятность наблюдения состояний символов (нуклеотидов или аминокислот), учитывая топологию дерева и модель эволюции. Для конкретного дерева вычисление вероятности включает суммирование всех возможных нуклеотидных (или аминокислотных) состояний в предковых (внутренних) узлах. Методы численной оптимизации используются для нахождения комбинации длин ветвей и эволюционных параметров, которая максимизирует вероятность. В зависимости от алгоритма поиска по этому критерию производится поиск вероятности ряда топологий деревьев, и дерево, дающее наибольшую вероятность, выбирается как лучшее дерево (Schmidt and von Haeseler, 2009).

Стоит также сказать о методе проверки деревьев, полученных данными методами − бутстрэппинг. Бутстрэп (англ. bootstrap) − метод проверки достоверности топологии дерева, основанный на многократном повторении произвольной выборки с повторениями из половины столбцов выравнивания и последующем построении дерева (на половине исходных данных).

При проведении бутстрэпа, каждой ветви присваивается значение – поддержка − количество случаев присутствия ветви. Причем значения поддержки можно отразить как на исходном дереве, так и на консенсусном дереве, построенном из непротиворечащих ветвей с максимальной поддержкой (Schmidt and von Haeseler, 2009).

Байесовский подход в филогенетике – это вероятностный метод, который также использует критерий оптимальности, но он концептуально очень отличаются от MP и ML тем, что не пытается найти только одно лучшее дерево. Байесовский метод также использует концепцию вероятности, но, ориентируясь на распределение вероятностей деревьев, он ищет набор правдоподобных деревьев или гипотез для данных. Это заданное распределение деревьев по своей сути содержит доверительную оценку любого эволюционного отношения. Этот метод позволяет учитывать

филогенетическую неопределенность, использовать априорную информацию и сложные модели эволюции.

Применение байесовского подхода в филогенетике состоит в следующем. Всё множество допустимых филогенетических деревьев описывается дискретными параметрами (топология деревьев) и непрерывными параметрами (длины ветвей деревьев и параметры эволюционной модели замен). Все параметры задаются вручную в соответствии с набором данных. Применение метода статистических испытаний (который также называется методом Монте-Карло) на Марковских цепях позволяет получить приближенные значения апостериорных вероятностей и уменьшить вычислительную сложность алгоритма поиска наиболее вероятного дерева по критерию максимума апостериорной вероятности (Schmidt and von Haeseler, 2009).

1.4. Некоторые методы анализа последовательностей без выравнивания

1.4.1. Методы, основанные на частотных словарях

Обоснование этих методов простое: похожие последовательности имеют похожие слова/*k*-меры (подпоследовательности длины *k*), а математические операции с вхожими словами дают хорошую относительную меру несходства последовательностей. Этот метод также тесно связан с идеей геномных сигнатур, которые были впервые введены для динуклеотидного состава (например, содержания GC) (Deschavanne et al., 1999) и в дальнейшем распространены на более длинные слова.

Этот процесс можно разбить на три основных этапа. Во-первых, сравниваемые последовательности должны быть разбиты на наборы уникальных слов заданной длины. Второй шаг – преобразовать каждую последовательность в массив чисел (вектор), например, путем подсчета количества раз, когда каждое конкретное слово появляется в последовательностях. Последний шаг включает количественную оценку

несходства между последовательностями посредством применения функции расстояния к векторам, представляющим последовательность. Эта разница очень часто вычисляется евклидовым расстоянием, хотя может применяться любая метрика. Чем выше значение несходства, тем более отдалены последовательности; таким образом, две идентичные последовательности приведут к расстоянию, равному нулю (Vinga, 2007).

Алгоритмы без выравнивания на основе слов бывают разными, с методологическими вариациями на каждом из трех основных шагов. На первом этапе можно попробовать разные длины слова - важно выбрать слова, которые вряд ли обычно встречаются в последовательности (чем короче слово, тем более вероятно, что оно появится в последовательности случайным образом). В анализе нуклеотидных последовательностей *k* можно установить равным 8–10 символам для генов или РНК (Höhl M, 2006), 9–14 оснований для общих филогенетических анализов (Sims, 2009) и до 25 оснований при сравнении изолятов одного вида бактерий (Bernard, 2017). Как показывает практика, меньшие *k*-меры должны использоваться, когда последовательности явно различны (например, они не связаны), тогда как более длинные *k*-меры могут использоваться для очень похожих последовательностей (Sims, 2009). Второй шаг (отображение последовательностей, как векторы) является наиболее настраиваемым; вместо использования векторов количества слов или частот слов есть много других способов создания векторов, которые варьируются от методов взвешивания до нормализации и кластеризации (Bonham-Carter et al., 2014). Кроме того, поскольку методы на основе слов работают с векторами, их математическая элегантность позволяет использовать более 40 функций, помимо евклидова расстояния, таких как коэффициент корреляции Пирсона (Vinga and Almeida J, 2003), расстояние Манхэттена и другие.

1.4.2. Методы, основанные на теории информации

Методы, основанные на теории информации, распознают и вычисляют объем информации, совместно используемой двумя анализируемыми биологическими последовательностями. Нуклеотидные и аминокислотные последовательности в конечном итоге представляют собой цепочки символов, и их цифровая организация естественным образом интерпретируется с помощью инструментов теории информации, таких как сложность и энтропия.

Например, колмогоровскую сложность последовательности можно измерить длиной ее кратчайшего описания. Соответственно, последовательность AAAAAAAAAA может быть описана несколькими словами (10 повторений A), тогда как CGTGATGT, по-видимому, не имеет более простого описания, чем определение нуклеотида за нуклеотидом (1 C, затем 1 G и так далее). Интуитивно более длинные описания последовательностей указывают на большую сложность. Сложность чаще всего аппроксимируется с помощью общих алгоритмов сжатия (например, реализованных в программах zip или gzip), где длина сжатой последовательности дает оценку ее сложности (т.е. более сложная строка будет менее сжимаемой) (Li and Vitányi, 2008.). Расчет расстояния между последовательностями с использованием сложности относительно прост. Эта процедура берет сравниваемые последовательности (x =ATGTGTG и y =CATGTG) и объединяет их для создания одной более длинной последовательности (xy = ATGTGTGCATGTG). Если x и y в точности совпадают, то сложность (сжатая длина) xy будет очень близка к сложности отдельных x или y. Однако, если x и y различны, то сложность xy будет иметь тенденцию к совокупной сложности x и y (Zielezinski et al., 2017).

Применение теории информации в области анализа и сравнения последовательностей резко расширилось в последние годы, начиная от глобального до локального анализа генома (Vinga, 2014).

Глава 2. Материалы и методы

2.1. Генетический материал

Было отобрано 69 последовательностей геномов коронавирусов всех 4 родов: 19 альфакоронавирусов (2 поражающих человека — NL63 и 229E и 8 коронавирусов, поражающих летучих мышей); 24 бетакоронавирусов (5 поражающих человека — SARS-CoV-2, OC43, MERS-CoV, HKU1 и SARS-CoV; 2 коронавируса, поражающих панголинов и 3 коронавируса, поражающих летучих мышей); 6 гаммакоронавирусов; 10 дельтакоронавирусов, а также 10 неклассифицированных. Большая часть последовательностей была взята из GenBank (Sayers et al., 2019), однако последовательности 3 геномов коронавирусов, обнаруженных в июне в Пекине и последовательности вируса RaTG13, были взяты из базы данных Gisaid (Shu and McCauley, 2017).

Набор последовательностей, использованных в работе, был сформирован таким образом, чтобы основные клады коронавирусов были представлены равномерно полногеномными последовательностями, содержащими не более 5% непрочитанных участков. Полногеномные нуклеотидные последовательности *Coronaviridae*, использованные в настоящей работе, представлены в приложении А.

2.2. Полногеномное выравнивание и построение филогении

Выравнивание последовательностей проводили с помощью MAFFT версии 7 (Katoh et. al., 2002). Был выбран алгоритм итеративного уточнения L-INS-i, поскольку он универсален и подходит для последовательностей, содержащих большое количество протяженных инделов.

Поиск оптимальной модели молекулярной эволюции и построение филогенетического дерева методом максимального правдоподобия выполнены с использованием программного обеспечения IQtree версии 1.6.12 (Nguyen et.

al., 2015**)**. Была выбрана модель молекулярной эволюции GTR (General Time Reversible) с учетом инвариантных сайтов и гамма-распределения (GTR+I+G).

Построение филогенетического дерева с использованием байесовского подхода проводилось в MrBayes версии 3.2.7a (Ronquist et. al., 2012). Использовались параметры модели lset nst=6 rates=invgamma, что соответствует модели GTR+I+G, выбранной программой IQtree. По умолчанию MrBayes выполняет два одновременных, полностью независимых анализа, построенных на основе различных случайных деревьев (nruns=2). И каждое из этих 2 деревьев перестраивается параллельно и независимо друг от друга в 8 (nchains=8) разных деревьев, используя различные эвристики для поиска оптимального дерева. Одновременное выполнение нескольких анализов позволяет MrBayes провести диагностику конвергенции в процессе вычисления. Каждый запуск начинается со случайно выбранного дерева. На ранних этапах вычисления выбираются 2 дерева, которые могут довольно сильно отличаться друг от друга, но при достижении конвергенции две выборки деревьев должны быть очень похожи.

Построение филогенетического дерева методом минимальной парсимонии осуществлялось при помощи R-пакета phangorn версии 2.5.5 с учетом выбранной модели (Schliep, 2011).

Деревья визуализировали с помощью программы FigTree версии 1.4.4 (http://tree.bio.ed.ac.uk/software/figtree).

2.3. Построение частотных словарей

Каждая нуклеотидная последовательность является символьной последовательностью из четырёхбуквенного алфавита $\aleph$ = {A, C, G, T} длины *N*. Для каждого исследуемого генома строился частотный словарь толщиной 3. Частотный словарь толщины 3 это список всех троек $\omega = \nu_1\nu_2\nu_3$ идущих подряд нуклеотидов с указанием частот этих троек. Всего существует 64 таких триплета. Частота f_ω данного триплета — это отношение числа копий n_ω

данного триплета к общему числу всех возможных триплетов N, где N — сумма всех n_ω:

$$f_\omega = \frac{n_\omega}{N}.$$

Рамка считывания при построении словаря W_3 смещалась на один нуклеотид, при этом словарь W_3 задает отображение генома в 64-мерном пространстве. Два генома считаются близкими, если соответствующие им точки в 64-мерном пространстве близки в смысле Евклидовой метрики.

2.4. Классификация методом динамических ядер

Метод динамических ядер (*k*-means) — наиболее простой и широко используемый итеративный алгоритм классификации, разделяющий множество данных на *k* классов, расположенных на возможно больших расстояниях друг от друга.

Алгоритм метода сводится к тому, что множество точек, представленных в *n*-мерном пространстве (в нашем случае 63-мерном), разделяется на *k* количество классов, при этом каждая точка попадает в какой-то один определенный класс. Точки группируются по показателю близости, например, по Евклидову расстоянию. Затем определяется динамическое ядро — среднее арифметическое значение частот каждого из триплетов точек для определения центра класса, после этого рассчитывается расстояние от точки до каждого центра кластеров. Если точка, первоначально принадлежавшая классу K1, при пересчете оказывается ближе к центру класса K2, то ее принадлежность к классу меняется. Положения центров вновь образованных классов пересчитываются, и весь алгоритм повторяется до тех пор, пока не прекратится переход точек между классами.

Для реализации метода динамических ядер (*k*-means) использовалось ПО ViDaExpert (Зиновьев, 2000). В данной работе не проводилось никаких

специальных исследований, направленных на поиск оптимального числа кластеров, поэтому мы перебирали значения *k* равными от 2 до 5.

2.5. Кластеризация методом упругих карт

Упругая карта применяется для нелинейного сокращения размерности и визуализации многомерных данных. Суть метода: в многомерном пространстве данных (в нашем случае 63-мерном; из анализа исключался триплет CGA, в виду его наименьшего вклада в различимость геномов — для него наблюдался минимум стандартного отклонения по всему ансамблю геномов) располагается упругая пластина, которая деформируется таким образом, чтобы приблизить имеющиеся точки данных и при этом пытается быть не слишком изогнутой и растянутой. Данные проецируются на эту пластину (определяются ортогональные проекции) и они отображаются на ней, как на карте.

Более строго метод заключается в следующем. На первом шаге находятся первые две главные компоненты и на них, как на осях, строится плоскость. Затем на эту плоскость проектируются все точки данных, и определяется минимальный квадрат, содержащий все точки. Квадрат делится на определённое число меньших квадратов (16 в нашем случае). На втором шаге каждая точка данных соединяется математической пружиной с узлом решётки, наиболее близко находящимся к проекции. Затем жёсткая плоскость (точнее, та её часть, которая соответствует большему квадрату) заменяется на упругую мембрану, упругие свойства которой однородны. На третьем шаге вся система отпускается, и пружины сокращаются (либо растягиваются под действием мембраны и соседних пружин), а мембрана деформируется. При этом деформация мембраны и её финальное состояние определяются минимумом суммарной энергии деформации. На четвёртом шаге положение каждой точки на деформированной карте переопределяется: находится новая ортогональная проекция (точка на деформированной карте, наиболее близко находящаяся к оригиналу). Наконец, все математические пружины удаляются,

и деформированная мембрана возвращается в исходное плоское состояние; при этом образы точек также меняют своё положение на упругой карте. Такое преобразование называется переходом во внутренние координаты.

Также нужно отметить, что упругость карты выбирается «вручную». Чем более упруга карта, тем более гладкую модель она представляет (при больших значениях коэффициента упругости узлы карты практически находятся в одной плоскости, и это плоскость главных компонент).

Для исследуемых геномов строилось распределение точек, соответствующих частотным словарям в пространстве частот. Определение кластера на упругой карте, представленной во внутренних координатах, проводилось по локальной плотности. Для этого каждая точка на упругой карте (напомним — образ исходной точки данных) снабжалась колоколообразной функцией; понятно, что выбор функций такого вида весьма широк, но мы использовали Гауссовскую функцию. Затем значения всех функций для каждой точки суммировалось по всем точкам, и значение суммарной функции и определяло локальную плотность. При отображении этой функции мы использовали схему с 15 уровнями значения функции локальной плотности; кластером считалась область с локальной плотностью, превышающей 9 снизу уровень.

Также на данные карты можно нанести информационные слои раскрасок, отображающих разнообразную информацию, например, локальную плотность. Для реализации данного метода также использовалось ПО ViDaExpert (Зиновьев, 2000).

Глава 3. Результаты

3.1. Филогенетический анализ

Для установления филогенетических связей были построены деревья тремя дискретными методами. Полученные филогенетические деревья были проассоциированы с двумя признаками: таксономической принадлежностью генома (Рис. 2А, 3А, 4А) и характером вызываемого заболевания (Рис. 2Б, 3Б, 4Б).

В результате анализа было получено филогенетическое дерево методом максимального правдоподобия с полностью разрешенной топологией, однако далеко не все узлы (~ 4,3%) имеют высокую или умеренную поддержку (Рис. 2А и 2Б).

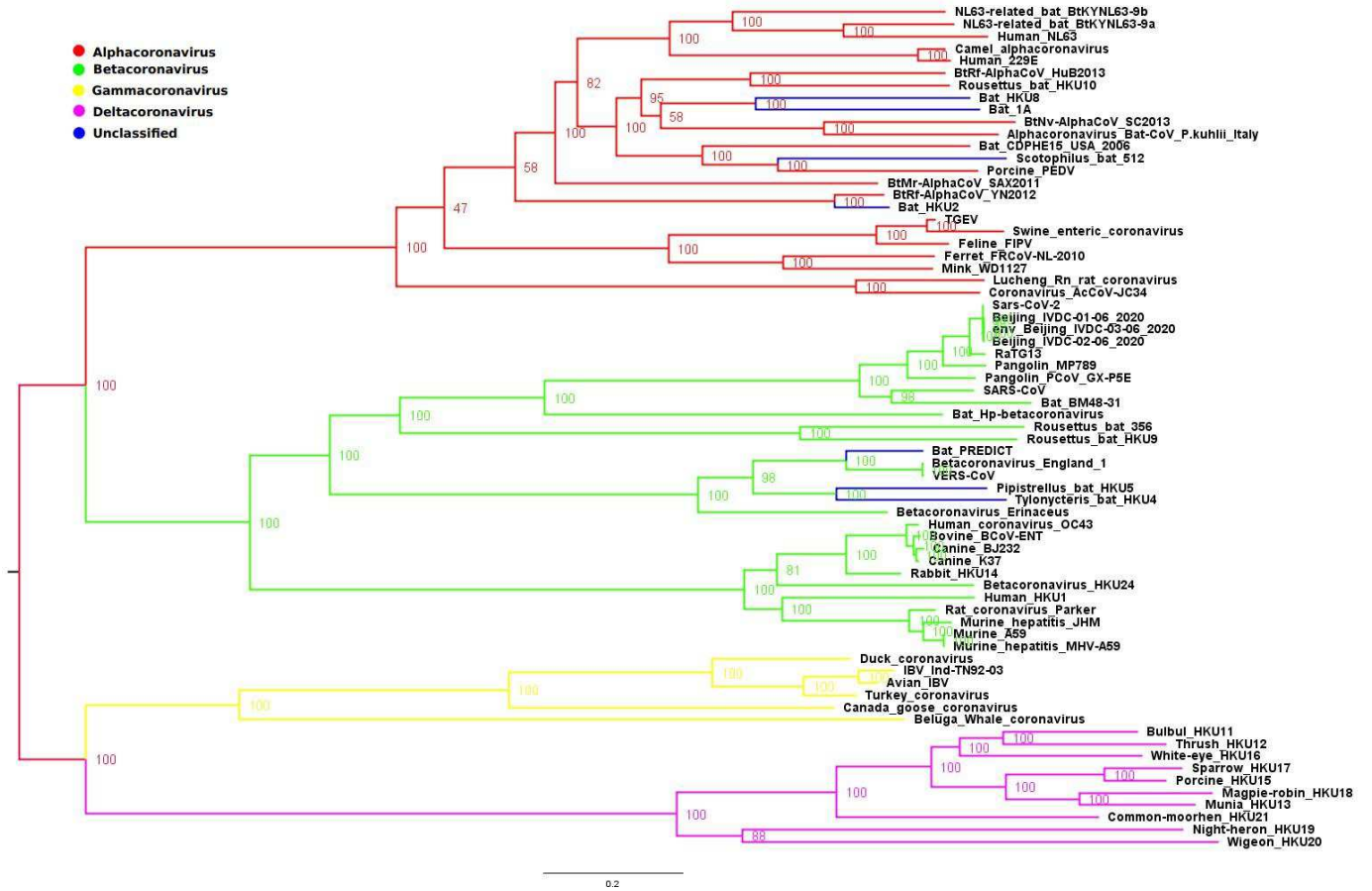

Рисунок 2А – Филогенетическое дерево, построенное методом максимального правдоподобия по отобранным 69 полногеномным последовательностям *Coronaviridae*. В качестве переменной использовалась принадлежность к тому или иному таксону *Coronaviridae*

Изъято 18 страниц

ЗАКЛЮЧЕНИЕ

Главной целью данной работы являлось сравнение метода, основанного на полногеномном выравнивании и метода классификации геномов по частотам триплетов на примере семейства геномов коронавирусов. Метод полногеномного выравнивания обладает рядом существенных недостатков. Первым из них является произвольный выбор штрафных функций для допустимых ошибок; вторым важным недостатком является расходимость результатов при использовании различных алгоритмов, а также высокая вычислительная сложность. Попытки построения методов, не использующих идею выравнивания, предпринимаются достаточно давно. Широкому использованию таких методов свободных от выравнивания препятствует отсутствие разумной системы сопоставления результатов сравнения, полученных разными методами. Во многом это обусловлено трудностью подбора соответствующего генетического материала.

Геномы коронавирусов представляют собой очень удачный объект для сравнительного анализа методов, основанных на полногеномном выравнивании и свободных от него подходов. В настоящей работе, на примере 69 геномов коронавирусов, было проведено сравнение классического выравнивания с последующим построением филогении, классического метода классификации без учителя (*k*-means) и современного метода нелинейной статистики — метода упругих карт.

В ходе данной работы было выполнено множественное полногеномное выравнивание для отобранных геномов коронавирусов и построены филогенетические деревья методами максимального правдоподобия, минимальной парсимонии и с применением байесовского подхода. Топологии деревьев, построенных методом максимального правдоподобия и с применением байесовского подхода в филогенетике, идентичны. Дерево, построенное методом минимальной парсимонии, отличалось по топологии, и

имело низкий уровень поддержки в узлах, в которых случились перестройки в сравнении с деревьями, полученными двумя другими методами. Выявлено соответствие филогении и таксономического положения изучаемых вирусов, однако связь между характером болезней и реконструированной эволюционной историей вирусов была незначительной.

Классификация частотных словарей триплетов методом динамических ядер с последовательно возрастающим числом классов позволила построить иерархическое дерево, последний слой которого представляет собой классы тесно связанных между собой геномов, которые с хорошей точностью совпадают с кластерами, выделяемыми по филогении.

Кластеризация методом упругих карт выявила нелинейные связи между геномами. Кластеры, выделенные методом упругих карт, также хорошо совпадают, как с классами последнего слоя слоистого графа, полученных методом динамических ядер, так и с результатами классического выравнивания.

Расстояния между центрами и радиусы кластеров подтверждают надежность полученной кластеризации и дают представление об истинных взаимоотношениях между геномами в естественных координатах.

Эти результаты показывают, что методы сравнения, не основанные на выравнивании не менее эффективны, более того у таких методов есть значительное преимущество в вычислительной сложности.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Weiss, S. R. Coronavirus pathogenesis / S. R Weiss, J. L. Leibowitz // Adv. Virus Res. – 2011. – V. 81. – P. 85–164.

2. Yang, D. The structure and functions of coronavirus genomic 3’ and 5’ ends / D. Yang, J. L. Leibowitz // Virus Res. – 2015. – V. 206 – P. 120–133.

3. Woo, P. C. Discovery of seven novel mammalian and avian coronaviruses in the genus deltacoronavirus supports bat coronaviruses as the gene source of alphacoronavirus and betacoronavirus and avian coronaviruses as the gene source of gammacoronavirus and deltacoronavirus / P. C. Woo, S. K. Lau, C. S. Lam, C. C. Lau, A. K. Tsang, M. Wang, B. J. Zheng, K. H. Chan, K. Y. Yuen // J. Virol. – 2012. – V. 86. – P. 3995–4008.

4. Brian, D. A. Coronavirus genome structure and replication / D. A. Brian, R. S. Baric // Curr. Top. Microbiol. Immunol. – 2005. – V. 287. – P. 1–30.

5. Lin, C. M. Evolution, antigenicity and pathogenicity of global porcine epidemic diarrhea virus strains / C. M. Lin, L. J. Saif, D. Marthaler, Q. Wang // Virus Res. – 2016. – V. 226. – P. 20–39.

6. Zhou, P. Fatal swine acute diarrhoea syndrome caused by an HKU2-related coronavirus of bat origin / P. Zhou, H. Fan, T. Lan, X. L. Yang et al. // Nature. – 2018. – V. 556. – P. 255–258.

7. Zaki, A. M. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia / A. M. Zaki, S. V. Boheemen, T. M. Bestebroer, A. D. M. E. Osterhaus, R. A. M. Fouchier // N Engl J Med. – 2012. – V. 367. – P. 1814–1820.

8. Cheng, Z. J. 2019 Novel coronavirus: where we are and what we know / Z. J. Cheng, J. Shan // Infection. – 2020. – V. 48. – P. 155–163.

9. Su, S. Epidemiology, genetic recombination, and pathogenesis of coronaviruses / S. Su, G. Wong, W. Shi, J. Liu et al. // Trends Microbiol. – 2016. – V. 24. – P. 490–502.

10. Forni, D. Molecular evolution of human coronavirus genomes / D. Forni, R. Cagliani, M. Clerici, M. Sironi // Trends Microbiol. – 2017. – V. 25. – P. 35–48.

11. Zhou, P. A pneumonia outbreak associated with a new coronavirus of probable bat origin / P. Zhou, X. L. Yang, X. G. Wang, B. Hu, L. Zhang // Nature. – 2020. – V. 579. – P. 270–273.

12. Ji, W. Homologous recombination within the spike glycoprotein of the newly identified coronavirus may boost cross-species transmission from snake to human / W. Ji, W. Wang, X. Zhao, J. Zai, X. Li // J Med Virol. – 2020. – V. 92. – P. 433–440.

13. Lam, T. Identification of 2019-nCoV related coronaviruses in Malayan pangolins / T. Lam, Y. W. Zhang, M. Shum, H. Zhu et al. // Nature. – 2020. – V. 583. – P. 282–285.

14. Sayers, E. W. GenBank / E. W. Sayers, M. Cavanaugh, K. Clark, J. Ostell, K. D. Pruitt, I. Karsch- Mizrachi // Nucleic Acids Res. – 2019. – V. 47. – P. 94–99.

15. Shu, Y. GISAID: global initiative on sharing all influenza data—from vision to reality / Y. Shu, J. McCauley // Euro Surveill. –2017. – V. 22. – P. 1–3.

16. Зиновьев, А. Ю. Визуализация многомерных данных. / А.Ю. Зиновьев. – Красноярск : Изд-во КГТУ, 2000. – 180 с.

17. Sadovsky, M.G. Genome structure of organelles strongly relates to taxonomy of bearers /M. Sadovsky, Yu. Putinseva, A. Chernyshova, V. Fedotova // LNBI. –2015. – V.9043, Part II. – P. 481– 490.

18. Sadovsky, M.G. Function vs. taxonomy: the case of fungi mitochondria ATP\synthase genes/ M.G., Sadovsky, Yu.A.Putintseva, T.O.Shpagina, V.D.Fedotovskaya, A.I.Kolesnikova // LNBI. – 2019. – V. 11465. – P. 335–345.

19. Fedotovskaya, V. Function vs. Taxonomy: Further Reading from Fungal Mitochondrial ATP Synthases / V. Fedotovskaya, Yu. Putintseva, T. Shpagina, A. Kolesnikova // LNBI. – 2020. – V.12108. – P. 438–444.

20. Sadovsky M.G. COMPARATIVE ANALYSIS OF TRIPLET COMPOSITION OF COMMON MITOCHONDRIAL AND CHLOROPLAST GENES OF THE SAME SPECIES / Sadovsky MG, Fedotovskaia V. // bioRxiv. – 2020 Jan.

21. Chen Y. Emerging coronaviruses: Genome structure, replication, and pathogenesis / Chen Y., Liu Q., Guo D. // J Med Virol. – 2020. – V. 92. – P. 418–423.

22. Song Z. From SARS to MERS, Thrusting Coronaviruses into the Spotlight / Song Z., Xu Y., Bao L., Zhang L., et al. // Viruses. – 2019. – V. 11. – 59 p.

23. Brockway, S.M. Characterization of the expression, intracellular localization, and replication complex association of the putative mouse hepatitis virus RNA-dependent RNA polymerase / S. M. Brockway, C. T. Clay, X. T. Lu // J. Virol. – 2003. – V. 77. – P. 10515–10527.

24. Ziebuhr, J. Virus-encoded proteinases and proteolytic processing in the Nidovirales / J. Ziebuhr, E. J. Snijder, A. E. Gorbalenya // J. Gen. Virol. – 2000. – V. 81. – P. 853–879.

25. Dye, C. Genomic RNA sequence of Feline coronavirus strain FIPV WSU-79/ C. Dye, S. G. Siddel // J. Gen. Virol. – 2005. – V. 86. – P. 2249–2253.

26. Haijema, B. J. Feline coronaviruses: a tale of two-faced types / B. J. Haijema, P. J. Rottier, R. J. de Groot // Coronaviruses: molecular and cellular biology. – United Kingdom : Caister Academic Press, 2007.

27. Tekes, G. Genome organization and reverse genetic analysis of a type I feline coronavirus / G. Tekes, R. Hofmann-Lehmann, I. Stallkamp, V. Thiel, H. J. Thiel // J Virol. – 2008. – V. 82. – P. 1851–1859.

28. Tekes, G. Feline Coronaviruses: Pathogenesis of Feline Infectious Peritonitis / G. Tekes, H. J. Thiel // Adv Virus Res. – 2016. – V. 96. – P. 193–218.

29. Narayanan, K. Coronavirus accessory proteins / K. Narayanan, C. Huang, S. Makino // Nidoviruses. – 2007. – P. 235–244.

30. Haijema, B. J. Live, attenuated coronavirus vaccines through the directed deletion of group-specific genes provide protection against feline infectious peritonitis / B. J. Haijema, H. Volders, P. J. Rottier // J Virol. – 2004. V. 78. – P. 3863–3871.

31. Van Boheemen, S. Genomic characterization of a newly discovered coronavirus associated with acute respiratory distress syndrome in humans / S. Van Boheemen, M. de Graaf, C. Lauber, T. M. Bestebroer et al. // MBio. – 2012. – V. 3. – e00473-12.

32. Lu, R. Complete Genome Sequence of Middle East Respiratory Syndrome Coronavirus (MERS-CoV) from the First Imported MERS-CoV Case in China / R. Lu, Y. Wang, W. Wang, K. Nie et al. // Genome Announc. – 2015. – V. 3, № 4. – e00818-15.

33. Chafekar, A. MERS-CoV: Understanding the Latest Human Coronavirus Threat. / A. Chafekar, B. C. Fielding // Viruses. – 2018. – V. 10, № 2. – 93 p.

34. Yang, Y. The structural and accessory proteins M, ORF 4a, ORF 4b, and ORF 5 of Middle East respiratory syndrome coronavirus (MERS-CoV) are potent interferon antagonists / Y. Yang, L. Zhang, H. Geng, Y. Deng et al.// Protein Cell. – 2013. – V. 4. – P. 951– 961.

35. Niemeyer, D. Middle East respiratory syndrome coronavirus accessory protein 4a is a type I interferon antagonist / D. Niemeyer, T. Zillinger, D. Muth, F. Zielecki et al. // J. Virol. – 2013. – V. 87. – P. 12489– 12495.

36. Matthews, K. L. The ORF4b-encoded accessory proteins of Middle East respiratory syndrome coronavirus and two related bat coronaviruses localize to the nucleus and inhibit innate immune signaling / K. L. Matthews, C. M. Coleman, Y. van der Meer, E. J. Snijder, M. B. Frieman // J. Gen. Virol. – 2014. – V. 95. – P. 874–882.

37. Beaudette, F. R. Cultivation of the virus of infectious bronchitis/ F. R. Beaudette, C. B. Hudson // J Am Vet Med Assoc. – 1937. – V. 90.P. – 51–58.

38. Cheever, F. S. A murine virus (JHM) causing disseminated encephalomyelitis with extensive destruction of myelin / F. S. Cheever, J. B. Daniels, A. M. Pappenheimer, O. T. Bailey // J Exp Med. – 1949. – V. 90. – P. 181–194.

39. Doyle, L. P. A transmissible gastroenteritis virus in pigs / L. P. Doyle, L. M. Hutchings // J Am Vet Assoc. – 1946. – V. 108. – P. 257–259.

40. Tyrrell, D. A. Cultivation of viruses from a high proportion of patients with colds / D. A. Tyrrell, M. L. Bynoe // Lancet. – 1966. – V. 1. – P. 76–77.

41. Hamre, D. A new virus isolated from the human respiratory tract / D. Hamre, J. J. Procknow // Proc Soc Exp Biol Med. – 1966. – V. 121. – P. 190–193.

42. Corman, V. M. Evidence for an ancestral association of human coronavirus 229E with bats / V. M. Corman, H. J. Baldwin, A. F. Tateno, R. M. Zerbinati et al. // J. Virol. – 2015. – V. 89. – P. 11858–11870.

43. Corman, V. M. Hosts and sources of endemic human coronaviruses / V. M. Corman, D. Muth, D. Niemeyer, C. Drosten // Adv. Virus Res. – 2018. – V. 100. – P. 163–188.

44. Tao, Y. Surveillance of bat coronaviruses in Kenya identifies relatives of human coronaviruses NL63 and 229E and their recombination history / Y. Tao, M. Shi, C. Chommanard, K. Queen et al. // J. Virol. – 2017. – V. 91. – e01953-16.

45. Guarner, J. Three emerging coronaviruses in two decades / J. Guarner // Am. J. Clin. Pathol. – 2020. – V. 153. – P. 420–421.

46. Li, D. X. Accessory proteins of SARS-CoV and other coronaviruses / D. X. Li, T. S. Fung, K. K. L. Chong et al. // Antiviral Res. – 2014. – V. 109. – P. 97–109.

47. Cavanagh, D. Coronavirus avian infectious bronchitis virus / D. Cavanagh // Vet Res. – 2007. – V. 38. – P. 281–297.

48. Zhang, M. Genetic manipulation of porcine deltacoronavirus reveals insights into NS6 and NS7 functions: a novel strategy for vaccine design / M. Zhang, W. Li, P. Zhou et al. // Emerg Microbes Infect. – 2019. – V. 09. – P. 20–31.

49. Klausegger, A. Identification of a coronavirus hemagglutinin-esterase with a substrate specificity different from those of influenza C virus and bovine coronavirus / A. Klausegger, B. Strobl, G. Regl et al. //J Virol.– 1999. – V. 73. – P. 3737–3743.

50. Sethna, P. B. Minus-strand copies of replicating coronavirus mRNAs contain antileaders / P. B. Sethna, M. A. Hofmann, D. A. Brian // J Virol.–1991. –V. 65. – P. 320–325.

51. Banner, L. R. Random nature of coronavirus RNA recombination in the absence of selection pressure / L. R. Banner, M. M. Lai //Virology.– 1991. – V. 185. – P. 441–445.

52. Lai, M. M. C. Recombination in large RNA viruses: coronaviruses / M. M. C. Lai // Semin. Virol. – 1996. – V. 7. – P. 381–388.

53. Zeng, Q. Structure of coronavirus hemagglutinin-esterase offers insight in corona and influenza virus evolution / Q. Zeng, M. A. Langereis, A. L. W. van Vliet, E. G. Huizinga, R. J. de Groot // Proc. Natl. Acad. Sci. – 2008. – V. 105. – P. 9065–9069.

54. Decaro, N. A. Novel human coronavirus (SARS-CoV-2): A lesson from animal coronaviruses / N. Decaro, A. Lorusso // Vet Microbiol. – 2020. – V. 244. – 108693.

55. Cavanagh, D. Coronaviruses from pheasants (*Phasianus colchicus*) are genetically closely related to coronaviruses of domestic fowl (Infectious Bronchitis Virus) and turkeys / D. Cavanag, K. Mawditt, B. Welchmand Dde, P. Britton, R. E. Gough // Avian Pathol. – 2002. – V. 31. – P. 81–93.

56. Circella, E. Coronavirus associated with an enteric syndrome on a quail farm / E. Circella, A. Camarda, V. Martella, G. Bruni et al. // Avian Pathol. – 2007. – V. 36. – P. 251–258.

57. Domanska-Blicharz, K. Detection and molecular characterization of infectious bronchitis-like viruses in wild bird populations / K. Domanska-Blicharz, A. Jacukowicz, A. Lisowska, K. Wyrostek, Z. Minta // Avian Pathol. –2014. – V. 43. – P. 406–413.

58. Torres, C. A. An avian coronavirus in quail with respiratory and reproductive signs / C. A. Torres, L. Y. B. Villarreal, G. G. R. Ayres, L. Richtzenhain, P. E. Brandão // Avian Dis. – 2013. – V. 57. – P. 295–299.

59. Hughes, L. A. Genetically diverse coronaviruses in wild bird populations of northern England / L. A. Hughes, C. Savage, C. Naylor, M. Bennett, J. Chantrey, R. Jones // Emerg Infect. Dis. – 2009. – V. 15. – P. 1091–1094.

60. Liu, S. Isolation of avian infectious bronchitis coronavirus from domestic peafowl *Pavo cristatus* and teal Anas / S. Liu, J. Chen, J. Chen, X. Kong et al. // J. Gen. Virol. – 2005. – V. 86. – P. 719–725.

61. Mihindukulasuriya, K. A. Identification of a novel coronavirus from a beluga whale by using a panviral microarray / K. A. Mihindukulasuriya, G. Wu, J. St Leger, R. W. Nordhausen, D. Wang // J. Virol. – 2008. – V. 82. – P. 5084–5088.

62. Woo, P. C. Comparative analysis of complete genome sequences of three avian coronaviruses reveals a novel group 3c coronavirus / P. C. Woo, S. K. Lau, C. S. Lam, K. Lai et al. // J. Virol. – 2009. – V. 83. – P. 908–917.

63. Chu, D. K. Avian coronavirus in wild aquatic birds / D. K. Chu, C. Y. Leung, M. Gilbert, P. H. Joyner et al. // J. Virol. – 2011. – V. 85. – P. 12815–12820.

64. Torres, C. A. Gammacoronavirus and deltacoronavirus in Quail / C. A. Torres, A. Hora, P. O. Tonietti, S. A. Taniwaki et al. // Avian Dis. – 2016. – V. 60. – P. 656–661.

65. Dong, B. Q. Detection of a novel and highly divergent coronavirus from Asian leopard cats and Chinese ferret badgers in southern China / B. Q. Dong, W. Liu, X. H. Fan, D. Vijaykrishna et al. // J. Virol. – 2007. – V. 81. – P. 6920–6926.

66. Poon, L. L. Identification of a novel coronavirus in bats / L. L. Poon, D. K. Chu, K. H. Chan, O. K. Wong et al. // J. Virol. – 2005. – V. 79. – P. 2001–2009.

67. Chen, L. DBatVir: The Database of Bat-associated Viruses / L. Chen, B. Liu, J. Yang, Q. Jin // Database (Oxford). – 2014. – 2014:bau021.

68. Banerjee, A. Bats and coronaviruses. Viruses / A. Banerjee, K. Kulcsar, V. Misra, M. Frieman, K. Mossman // Viruses. – 2019. – V. 11, № 1 – P. 41.

69. Drexler, J. F. Genomic characterization of severe acute respiratory syndrome- related coronavirus in European bats and classification of coronaviruses based on partial RNA- dependent RNA polymerase gene sequences / J. F. Drexler, F. Gloza-Rausch, J. Glende, V. M. Corman et al. // J. Virol. – 2010. – V. 84. – P. 11336–11349.

70. Mühldorfer, K. Diseases and causes of death in European bats: dynamics in disease susceptibility and infection rates / K. Mühldorfer, S. Speck, A. Kurth, R. Lesnik et al. // PLoS One. – 2011. – V. 6, № 12. – e29773.

71. Ahn, M. Dampened NLRP3-mediated inflammation in bats and implications for a special viral reservoir host / M. Ahn, D. E. Anderson, Q. Zhang, C. W. Tan et al. // Nat. Microbiol. – 2019. – V. 4. – P. 789–799.

72. Brook, C. E. Accelerated viral dynamics in bat cell lines, with implications for zoonotic emergence / C. E. Brook, M. Boots, K. Chandran, A. P. Dobson et al. // Elife. – 2020. – V. 9. – e48401.

73. Crossley, B. M. Identification and characterization of a novel alpaca respiratory coronavirus most closely related to the human coronavirus 229E / B. M. Crossley, R. E. Mock, S. A. Callison, S. K. Hietala // Viruses. – 2012. – V. 4. – P. 3689–3700.

74. Tu, C. Antibodies to SARS coronavirus in civets / C. Tu, G. Crameri, X. Kong, J. Chen et al. // Emerg. Infect. Dis. – 2004. – V. 10. – P. 2244–2248.

75. Kan, B. Molecular evolution analysis and geographic investigation of severe acute respiratory syndrome coronavirus-like virus in palm civets at an animal market and on farms / B. Kan, M. Wang, H. Jing, H. Xu et al. // J. Virol. – 2005. – V. 79. – P. 11892–11900.

76. Lau, S. K. Severe acute respiratory syndrome coronavirus-like virus in Chinese horseshoe bats / S. K. Lau, P. C. Woo, K. S. Li, Y. Huang et al. // Proc. Natl. Acad. Sci. U. S. A. – 2005. – V. 102. – P. 14040–14045.

77. Tang, X. C. Prevalence and genetic diversity of coronaviruses in bats from China / X. C. Tang, J. X. Zhang, S. Y. Zhang, P. Wang et al. // J. Virol. – 2006. – V. 80. – P. 7481–7490.

78. Woo, P. C. Molecular diversity of coronaviruses in bats / P. C. Woo, S. K. Lau, K. S. Li, R. W. Poon et al. // Virology. – 2006. – V. 351. – P. 180–187.

79. Ren, W. Difference in receptor usage between severe acute respiratory syndrome (SARS) coronavirus and SARS-like coronavirus of bat origin / W. Ren, X. Qu, W. Li, Z. Han et al. // J. Virol. – 2008. – V. 82. – P. 1899–1907.

80. Ge, X. Y. Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor / X. Y. Ge, J. L. Li, X. L. Yang, A. A. Chmura et al. // Nature. – 2013. – V. 503. – P. 535–538.

81. Hu, B. Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus / B. Hu, L. P. Zeng, X. L. Yang, X. Y. Ge et al. // PLoS Pathog. – 2017. – V. 13.

82. Cui, J. Origin and evolution of pathogenic coronaviruses / J. Cui, F. Li, Z. L. Shi // Nat. Rev. Microbiol. – 2019. – V. 17. – P. 181–192.

83. Lelli, D. Detection of coronaviruses in bats of various species in Italy / D. Lelli, A. Papetti, C. Sabelli, E. Rosti et al. // Viruses. – 2013. –V. 5. – P. 2679–2689.

84. Corman, V. M. Rooting the phylogenetic tree of Middle East respiratory syndrome coronavirus by characterization of a conspecific virus from an African bat / V. M. Corman, N. L. Ithete, L. R. Richards, M. C. Schoeman et al. // J. Virol. – 2014. – V. 88. – P. 11297–11303.

85. Geldenhuys, M. A metagenomic viral discovery approach identifies potential zoonotic and novel mammalian viruses in *Neoromicia* bats within South Africa / M. Geldenhuys, M. Mortlock, J. Weyer, O. Bezuidt et al. // PLoS One. – 2018. – V. 13, № 3. – e0194527.

86. Memish, Z. A. Middle East respiratory syndrome coronavirus in bats, Saudi Arabia / Z. A. Memish, N. Mishra, K. J. Olival, S. F. Fagbo et al. // Emerg. Infect. Dis. – 2013. – V. 19. – P. 1819–1823.

87. Hoffmann, M. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor / M. Hoffmann, H. Kleine-Weber, S. Schroeder, N. Krüger et al. // Cell. – 2020. – V. 16. – P. 271–280.

88. Tang, X. On the origin and continuing evolution of SARS-CoV-2 / X. Tang, C. Wu, X. Li, Y. Song et al. // Sci. Rev. – 2020. – nwaa036.

89. Lam, T. T. Identifying SARS-CoV-2 related coronaviruses in Malayan pangolins / T. T. Lam, M. H. Shum, H. C. Zhu, Y. G. Tong et al. // Nature. – 2020. – V. 583. – P. 282–285.

90. Lorusso, A. Gain, preservation, and loss of a group 1a coronavirus accessory glycoprotein / A. Lorusso, N. Decaro, P. Schellen, P. J. Rottier et al. // J. Virol. – 2008. – V. 82. – P. 10312–10317.

91. Vijgen, L. Evolutionary history of the closely related group 2 coronaviruses: porcine hemagglutinating encephalomyelitis virus, bovine coronavirus, and human coronavirus OC43 / L. Vijgen, E. Keyaerts, P. Lemey, P. Maes et al. // J. Virol. – 2006. – V. 80. – P. 7270–7274.

92. Lau, S. K. P. Discovery and sequence analysis of four deltacoronaviruses from birds in the Middle East reveal interspecies jumping with recombination as a potential mechanism for avian-to-avian and avian-to-mammalian transmission / S. K. P. Lau, E. Y. M. Wong, C. C. Tsang, S. S. Ahmed et al. // J. Virol. – 2018. – V. 92. – e00265–18.

93. Vergara-Alert, J. Livestock susceptibility to infection with Middle East respiratory syndrome coronavirus / J. Vergara-Alert, J. M. van den Brand, W. Widagdo, M. Muñoz et al. // Emerg. Infect. Dis. – 2017. – V. 23. – P. 232–240.

94. Chen, W. SARS-associated coronavirus transmitted from human to pig. Emerg / W. Chen, M. Yan, L. Yang, B. Ding et al. // Infect. Dis. – 2005. – V. 11. – P. 446–448.

95. Wang, M. Surveillance on severe acute respiratory syndrome associated coronavirus in animals at a live animal market of Guangzhou in 2004/ M. Wang, H. Q. Jing, H. F. Xu, X. G. Jiang et al. // Zhonghua Liu Xing Bing Xue Za Zhi. – 2005. – V. 26. – P. 84–87.

96. Shi, J. Susceptibility of Ferrets, Cats, Dogs, and Different Domestic Animals to SARS-coronavirus-2 / J. Shi, Z. Wen, G. Zhong, H. Yang et al. // Science. – 2020. – V. 29. – P. 1016–1020.

97. Parker, J. C. Rat coronavirus (RCV): a prevalent, naturally occurring pneumotropic virus of rats / J. C. Parker, S. S. Cross, W. P. Rowe // Arch. Gesamte Virusforsch. – 1970. – V. 31. – P. 293–302.

98. Bhatt, P. N. Characterization of the virus of sialodacryoadenitis of rats: a member of the coronavirus group / P. N. Bhatt, D. H. Percy, A. M. Jonas // J. Infect. Dis. – 1972. – V. 126. – P. 23–130.

99. Lau, S. K. Discovery of a novel coronavirus, China Rattus coronavirus HKU24, from Norway rats supports the murine origin of Betacoronavirus 1 and has implications for the ancestor of Betacoronavirus lineage A / S. K. Lau, P. C. Woo, K. S. Li, A. K. Tsang et al. // J. Virol. – 2015. – V. 89. – P. 3076–3092.

100. Gretebeck, L. M. Animal models for SARS and MERS coronaviruses / L. M. Gretebeck, K. Subbarao // Curr. Opin. Virol. – 2015 – V. 13. – P. 123–129.

101. Cockrell, A. S. Mouse dipeptidyl peptidase 4 is not a functional receptor for Middle East respiratory syndrome coronavirus infection / A. S. Cockrell, K. M. Peck, B. L. Yount, S. S. Agnihothram et al. // J. Virol. – 2014. – V. 88. – P. 5195–5199.

102. Hemida, M. G. Dromedary camels and the transmission of Middle East respiratory syndrome coronavirus (MERS-CoV) Transbound / M. G. Hemida, A. Elmoslemany, F. Al-Hizab, A. Alnaeem et al. // Emerg. Dis. – 2017. – V. 64. – P. 344–353.

103. Reusken, C. B. Middle East Respiratory Syndrome coronavirus (MERS-CoV) serology in major livestock species in an affected region in Jordan, June to September 2013 / C. B. Reusken, M. Ababneh, V. S. Raj, B. Meyer et al. // Euro Surveill. – 2013. – V. 18. – P. 206–262.

104. Adney, D. R. Inoculation of goats, sheep, and horses with MERS-CoV does not result in productive viral shedding / D. R. Adney, V. R. Brown, S. M. Porter, H. Bielefeldt-Ohmann et al. // Viruses. – 2016. – V. 8. – E230.

105. Kandeil, A. Middle East respiratory syndrome coronavirus infection in non-camelid domestic mammals / A. Kandeil, M. Gomaa, M. Shehata, A. El-Taweel et al. // Emerg. Microbes Infect. – 2019. – V. 8. – P. 3–108.

106. Guan, Y. Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China / Y. Guan, B. J. Zheng, Y. Q. He, X. L. Liu et al. // Science. – 2003. – V. 302. – P. 276–278.

107. Shi, Z. A review of studies on animal reservoirs of the SARS coronavirus / Z. Shi, Z. Hu // Virus Res. – 2008. – V. 133. – P. 74–87.

108. Brudno, M. LAGAN and Multi-LAGAN: Efficient tools for large-scale multiple alignment of genomic DNA / M. Brudno, C. B. Do, G. M. Cooper, M. F. Kim et al. // Genome Res. – 2003. – V. 13. – P. 721–731.

109. Needleman, S. B. A general method applicable to the search for similarities in the amino acid sequence of two proteins / S. B. Needleman & C. D. Wunsch // J Mol Biol. – 1970. – V. 48. – P. 443–453.

110. Biological Sequence Analysis / R. Durbin et al. – Cambridge: Cambridge University Press, 1998. – 357 p.

111. Thompson, J. D. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice/ J. D. Thompson, D. G. Higgins, T. J. Gibson // Nucleic Acids Res. – 1994. – V. 22. – P. 4673–4680.

112. Wallace, I. M. Evaluation of iterative alignment algorithms for multiple alignment / I. M. Wallace, O. O'Sullivan, D. G. Higgins // Bioinformatics. – 2005. – V. 21. – P. 1408–1414.

113. Altschul, S. F. Basic local alignment search tool / S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman // J Mol Biol. – 1990. – V. 215. – P. 403–410.

114. Chan, C. X. Next-generation phylogenomics / C, X. Chan, M. A. Ragan // Biol Direct. – 2013. – V. 8. – 3 p.

115. Zielezinski, A. Alignment-free sequence comparison: benefits, applications, and tools / A. Zielezinski, S. Vinga, J. Almeida, W. M. Karlowski // Genome Biol. – 2017. – V. 18, № 1. – 186 p.

116. Kantorovitz, M. R. A statistical method for alignment-free comparison of regulatory sequences / M. R. Kantorovitz, G. E. Robinson, S. Sinha // Bioinformatics. – 2007. – V. 23. – i249-55.

117. Ivan, A. Computational discovery of cis-regulatory modules in Drosophila without prior knowledge of motifs / A. Ivan, M. S. Halfon, S. Sinha // Genome Biol. – 2008. – V. 9, № 1. – R22.

118. Vinga, S. Comparative evaluation of word composition distances for the recognition of SCOP relationships / S. Vinga, R. Gouveia-Oliveira, J. S. Almeida // Bioinformatics. – 2004. – V. 20. – P. 206–215.

119. Terrapon, N. Rapid similarity search of proteins using alignments of domain arrangements / N. Terrapon, J. Weiner, S. Grath, A. D. Moore, E. Bornberg-Bauer // Bioinformatics. – 2014. – V. 30. – P. 274–281.

120. Cong, Y. A novel alignment-free method for detection of lateral genetic transfer based on TF-IDF / Y. Cong, Y-B Chan, M. A. Ragan // Sci Rep. – 2016. – V. 6. – P. 303–308.

121. Ondov, B. D. Mash: fast genome and metagenome distance estimation using MinHash / B. D. Ondov, T. J. Treangen, P. Melsted, A. B. Mallonee et al. // Genome Biol. – 2016. – V. 17. – 132 p.

122. Luczak, B. B. A survey and evaluations of histogram-based statistics in alignment-free sequence comparison / B. B. Luczak, B. T. James, H. Z. Girgis // Brief Bioinform. – 2019. – V. 20. – P. 1222–1237.

123. Lu, Y. Y. CAFE: aCcelerated Alignment-FrEe sequence analysis / Y. Y. Lu, K. Tang, J. Ren, J. A. Fuhrman et al. // Nucleic Acids Res. – 2017. – V. 45. – P. 554–559.

124. Chan, C. X. Inferring phylogenies of evolving sequences without multiple sequence alignment / C. X. Chan, G. Bernard, O. Poirion, J. M. Hogan, M. A. Ragan // Sci Rep. – 2014. – V. 4. – 6504.

125. Jukes, T. H. and Cantor, C. R. Evolution of protein molecules /in Mammalian Protein Metabolism, ed H. M. Munro / T. H. Jukes, C. R. Cantor // New York : Academic Press. – 1969. – P. 21–132.

126. Kimura, M. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences / M. Kimura // J. Mol. – 1980. – V. 16, – P. 111–120.

127. Collins, D. W. Rates of transition and transversion in coding sequences since the human-rodent divergence / D. W. Collins, T. H. Jukes // Genomics. – 1994. – V. 20. – P. 386–396.

128. Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach / J. Felsenstein // J. Mol. Evol. –1981. – V.17. – P. 368–376.

129. Hasegawa, M. Dating the human-ape splitting by a molecular clock of mitochondrial DNA / M. Hasegawa, H. Kishino, T. Yano // J. Mol. Evol. –1985. – V. 22. – P. 160–174.

130. Zharkikh, A. Estimation of evolutionary distances between nucleotide sequences / A. Zharkikh // J. Mol. Evol. – 1994. – V. 39. – P. 315–329.

131. Tavaré, S. Some probabilistic and statistical problems in the analysis of DNA sequences / S. Tavaré // Lect. Math. Life. Sci. – 1986. – V. 17. – P. 57–86.

132. Shoemaker, J. S. Evidence from nuclear sequences that invariable sites should be considered when sequence divergence is calculated / J. S. Shoemaker & W. M. Fitch // Mol. Biol. Evol. – 1989. – V. 6. – P. 270–289.

133. Yang, Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods / Z. Yang // J. Mol. Evol. – 1994. – V. 39. – P. 306–314.

134. Darriba, D. jModelTest 2: more models, new heuristics and parallel computing / D. Darriba, G. L. Taboada, R. Doallo, D. Posada // Nat. Methods. – 2012. – V. 9, № 8. – 772 p.

135. Arenas, M. Simulation of genome-wide evolution under heterogeneous substitution models and complex multispecies coalescent Histories / M. Arenas & D. Posada // Mol. Biol. Evol. – 2014. – V. 31. – P. 1295–1301.

136. Sumner, J. G. Is the general time-reversible model bad for molecular phylogenetics? / J. G. Sumner, P. D. Jarvis, J. Fernández-Sánchez, B. T. Kaine et al. // Syst. Biol. – 2012. – V. 61. – P. 1069–1074.

137. Jayaswal V. Two stationary nonhomogeneous Markov models of nucleotide sequence evolution / V. Jayaswal, L. S. Jermiin, L. Poladian, J. Robinson // Syst. Biol.– 2011. – V. 60. – P. 74–86.

138. Boussau, B. Efficient likelihood computations with nonreversible models of evolution / B. Boussau, M. Gouy // Syst. Biol. – 2006. – V. 55. – P. 756–768.

139. Lunter, G. A nucleotide substitution model with nearest-neighbour interactions / G. Lunter, J. Hein // Bioinformatics. – 2004. – V. 20. – P. 216–223.

140. Kaehler, B. D. Genetic distance for a general non-stationary markov substitution process/ B. D. Kaehler, V. B. Yap, R. Zhang, G. A. Huttley //Syst. Biol. – 2015. – V. 64. – P. 281–293.

141. Sokal, R. R. A statistical method for evaluating systematic relationships [J]. / R. R. Sokal, C. D. Michener // Univ. Kans. Sci. Bull. – 1958. – V. 28. – P. 1409–1438.

142. Saitou, N & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees / N. Saitou, M. Nei // Mol Biol Evol. – 1987. – V. 4. – P. 406–425.

143. Mount, D. W. Maximum parsimony method for phylogenetic prediction / D. W. Mount // Cold Spring Harbor Protocols. – 2008. – V. 8, № 4. – pdb-top32.

144. Lemey, P. The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing (2nd ed.) / P. Lemey, M. Salemi, A. Vandamme. – Cambridge : Cambridge University Press, 2009. – 723 p.

145. Deschavanne, P. J. Genomic signature: characterization and classification of species assessed by chaos game representation of sequences/ P. J. Deschavanne, A. Giron, J. Vilain, G. Fagot, B. Fertil // Mol Biol Evol. – 1999. – V. 16. – P. 1391–1399.

146. Vinga, S. Biological sequence analysis by vector-valued functions: revisiting alignment-free methodologies for DNA and protein classification / S. Vinga // In: Pham T. D., Yan H., Crane D.I., editors. Advanced computational methods for biocomputing and bioimaging. New York : Nova Science, 2007. – P. 70–105.

147. Höhl, M. Pattern-based phylogenetic distance estimation and tree reconstruction / M. Höhl, I. Rigoutsos, M. A. Ragan // Evol Bioinform Online. – 2006. – V. 2. – P. 359–375.

148. Sims, G. E. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions / G. E. Sims, S. Jun, G. A. Wu, S. Kim // Proc Natl Acad Sci U S A. – 2009. – V.106, № 8. – P. 2677–2682.

149. Bernard, G. Alignment-free inference of hierarchical and reticulate phylogenomic relationships / G. Bernard, C. X. Chan, Y. Chan, X-Y Chua, Y. Cong, J. M. Hogan et al. // Brief Bioinform. – 2019. – V. 20. – P. 426–435.

150. Bonham-Carter, O. Alignment-free genetic sequence comparisons: a review of recent approaches by word analysis / O. Bonham-Carter, J. Steele, D. Bastola // Brief Bioinform. – 2014. – V. 15. – P. 890–905.

151. Vinga, S. Almeida J. Alignment-free sequence comparison—a review / S. Vinga // Bioinformatics. – 2003. – V. 19. – P. 513–523.

152. Li, M. An introduction to Kolmogorov complexity and its applications / M. Li, P. Vitányi // New York, NY: Springer New York. – 2008. – 448 p.

153. Vinga, S. Information theory applications for biological sequence analysis / S. Vinga // Brief Bioinform. – 2014. – V. 15. – P. 376–389.

154. Katoh, K. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform / K. Katoh, K. Misawa, K. Kuma, T. Miyat // Nucleic Acids Res. – 2002. – V. 30. – P. 3059–3066.

155. Nguyen, L.-T. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum likelihood phylogenies / L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, B. Q. Minh // Mol. Biol. Evol. – 2015. – V. 32. – P. 268–274.

156. Ronquist, F. MRBAYES 3.2: Efficient Bayesian phylogenetic inference and model selection across a large model space / F. Ronquist, M. Teslenko, P. van der Mark, D. L. Ayres et al. // Syst. Biol. – 2012. – V. 61. – P. 539–542.

157. Schliep, K. P. phangorn: phylogenetic analysis in R / K. P. Schliep // Bioinformatics. – 2011. – V. 27. – P. 592–593

158. FigTree [Электронный ресурс] : a graphical viewer of phylogenetic trees designed by the University of Edinburgh. – Режим доступа: http://tree.bio.ed.ac.uk/software/figtree

ПРИЛОЖЕНИЕ А

Таблица А. 1 − Отобранные нуклеотидные последовательности Coronaviridae. Первый столбец «Название генома» окрашен согласно вызываемым заболеваниям: синим – заболевания летучих мышей, красным – респираторные заболевания людей, желтым – кишечные заболевания млекопитающих, голубым – респираторные заболевания млекопитающих, розовым – мышиные гепатиты, зеленым – респираторные заболевания птиц. Третий столбец «Таксономическая принадлженость» окрашен согласно таксономии: красным – альфакоронавирусы, зеленым – бетакоронавирусы, желтым – гаммакоронавирусы, розовым – дельтакоронавирусы.

Название генома	Инвентарный номер	Таксономическая принадлежность	Хозяин вируса
Piglets/(TGEV) PUR46-MAD	NC_038861.1	Альфакоронавирус	Исключительно поросята
Ferret/FRCoV-NL-2010	NC_030292.1	Альфакоронавирус	Хорьки, Норки
Swine/Italy/213306/2009	NC_028806.1	Альфакоронавирус	Свиньи
Mink/WD1127	NC_023760.1	Альфакоронавирус	Норки
Feline/FIPV	NC_002306.3	Альфакоронавирус	Кошки
Porcine/PEDV	NC_003436.1	Альфакоронавирус	Свиньи
Rodents/AcCoV-JC34	NC_034972.1	Альфакоронавирус	Грызуны
Camel/camel/Riyadh/Ry141/2015	NC_028752.1	Альфакоронавирус	Верблюды, люди
Human/NL63	NC_005831.2	Альфакоронавирус	Человек
Human/229E	NC_002645.1	Альфакоронавирус	Человек
Bat/CDPHE15/USA/2006	NC_022103.1	Альфакоронавирус	Летучие мыши
Bat/NL63-BtKYNL63-9b	NC_048216.1	Альфакоронавирус	Летучие мыши
Bat/CoV/P.kuhlii/Italy/3398-19/2015	NC_046964.1	Альфакоронавирус	Летучие мыши
Bat/NL63-BtKYNL63-9a	NC_032107.1	Альфакоронавирус	Летучие Мыши
Bat/BtNv-AlphaCoV/SC2013	NC_028833.1	Альфакоронавирус	Летучие Мыши
Bat/BtRf-AlphaCoV/YN2012	NC_028824.1	Альфакоронавирус	Летучие Мыши
Bat/BtRf-AlphaCoV/HuB2013	NC_028814.1	Альфакоронавирус	Летучие Мыши
Bat/PREDICT/PDF-2180BtMr-AlphaCoV/SAX2011	NC_028811.1	Альфакоронавирус	Летучие Мыши
Lucheng Rn rat/Lucheng-19	NC_032730.1	Альфакоронавирус	Летучая мышь (Домовые гладконосы)
Rousettus bat/HKU10	NC_018871.1	Альфакоронавирус	Летучая мышь (Домовые гладконосы)
Pangolin/MP789	MT121216.1	Бетакоронавирус	Панголины
Pangolin/PCoV_GX-P5E	MT040336.1	Бетакоронавирус	Панголины
Erinaceus/VMC/DEU/2012	NC_039207.1	Бетакоронавирус	Ежи
Norway rats/HKU24	NC_026011.1	Бетакоронавирус	Норвежские мыши

Окончание приложения А.

Название генома	Инвентарный номер	Таксономическая принадлежность	Хозяин вируса
Rabbit/HKU14	NC_017083.1	Бетакоронавирус	Кролики
Rat/PRC	NC_012936.1	Бетакоронавирус	Крысы
Bovine/BCoV	NC_003045.1	Бетакоронавирус	Коровы
Canine/BJ232	KX432213.1	Бетакоронавирус	Собаки
Canine/K37	JX860640.1	Бетакоронавирус	Собаки
Human/SARS-CoV	NC_045512.2	Бетакоронавирус	Человек
Human/OC43	NC_006213.1	Бетакоронавирус	Человек
Human/England 1	NC_038294.1	Бетакоронавирус	Человек
Human/MERS-CoV	NC_019843.3	Бетакоронавирус	Человек
Human/HKU1	NC_006577.2	Бетакоронавирус	Человек
Human/SARS-CoX-2	NC_004718.3	Бетакоронавирус	Человек
Rousettus bat/GCCDC1-356	NC_030886.1	Бетакоронавирус	Летучие мыши
Bat/Hp-B/Zhejiang2013	NC_025217.1	Бетакоронавирус	Летучие мыши
Mouse/MHV-A59 C12	NC_001846.1	Бетакоронавирус	Мыши
Mouse/A59	NC_048217.1	Бетакоронавирус	Мыши
Mouse/JHM	AC_000192.1	Бетакоронавирус	Мыши
Beluga Whale/SW1	NC_010646.1	Гаммакоронавирус	Белуга (в неволе)
Turkey/TCoV	NC_010800.1	Гаммакоронавирус	Индейки
Duck/DdCoV	NC_048214.1	Гаммакоронавирус	Куры и утки
Avian/IBV	NC_048213.1	Гаммакоронавирус	Птицы
Canada goose/Cambridge_Bay_2017	NC_046965.1	Гаммакоронавирус	Гуси
Avian/IBVB	NC_001451.1	Гаммакоронавирус	Птицы
Porcine/HKU15	NC_039208.1	Дельтакоронавирус	Свиньи
Bulbul/HKU11-934	NC_011547.1	Дельтакоронавирус	Бюльбюлевые(птицы)
Sparrow/HKU17	NC_016992.1	Дельтакоронавирус	Воробьи
White-eye/HKU16	NC_016991.1	Дельтакоронавирус	Бюльбюлевые(дрозды)
Common-moorhen/HKU21	NC_016996.1	Дельтакоронавирус	Камышница(птица)
Wigeon/HKU20	NC_016995.1	Дельтакоронавирус	Свиязь (птица)
Night-heron/HKU19	NC_016994.1	Дельтакоронавирус	Кваква(птица)
Magpie-robin/HKU18	NC_016993.1	Дельтакоронавирус	Сорока
Munia/HKU13-3514	NC_011550.1	Дельтакоронавирус	Мунии(птица)
Thrush/HKU12-600	NC_011549.1	Дельтакоронавирус	Дрозды
Bat/PREDICT/PDF-2180	NC_034440.1	unclassified	Летучие мыши
Bat/BM48-31/BGR/2008	NC_014470.1	unclassified	Летучие мыши
Bat/HKU8	NC_010438.1	unclassified	Летучие мыши
Bat/1A	NC_010437.1	unclassified	Летучие мыши
Bat/HKU2	NC_009988.1	unclassified	Летучие мыши
Bat/HKU9-1	NC_009021.1	unclassified	Летучие мыши
Bat/HKU5-1	NC_009020.1	unclassified	Летучие мыши
Bat/HKU4-1	NC_009019.1	unclassified	Летучие мыши
Scotophilus bat/512	NC_009657.1	unclassified	Летучие мыши

ПРИЛОЖЕНИЕ Б

Кластеризация на детальной упругой карте 25 × 25.

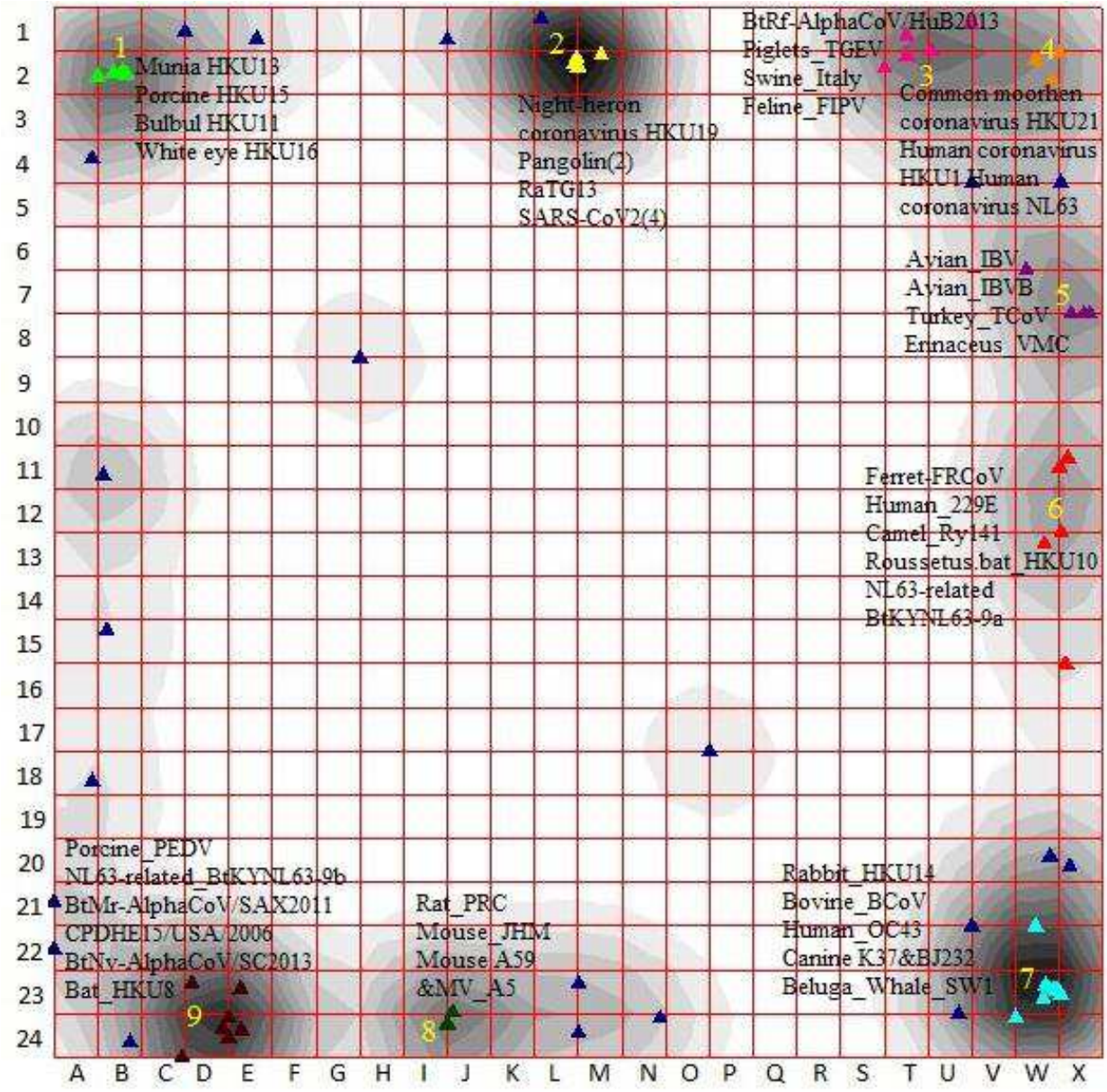

Рисунок Б. 1 − Детальная карта 25×25 во внутренних координатах

Примечание: количество кластеров уменьшилось до 9, при этом монотонности по включению кластеров, полученных с помощью мягкой упругой (16×16) карты, в кластеры, полученные с помощью детальной (25×25) карты, не наблюдается: некоторые из них слились в один, а некоторые распались на новые.

Продолжение приложения Б.

Таблица Б. 1 − Расстояние между центрами и радиусы (R) кластеров (L) на мягкой карте 25×25. Желтым цветом выделены наименьшие расстояния между кластерами

L	2	3	4	5	6	7	8	9	10
1	0,0208	0,0275	0,0524	0,0324	0,0347	0,0426	0,0337	0,0325	0,0175
2		0,0171	0,0460	0,0250	0,0287	0,0385	0,0361	0,0343	0,0233
3			0,0365	0,0157	0,0161	0,0262	0,0259	0,0243	0,0225
4				0,0326	0,0334	0,0207	0,0416	0,0425	0,0466
5					0,0221	0,0201	0,0222	0,0285	0,0267
6						0,0254	0,0269	0,0165	0,0276
7							0,0236	0,0311	0,0353
8								0,0233	0,0243
9									0,0235

Продолжение приложения Б.

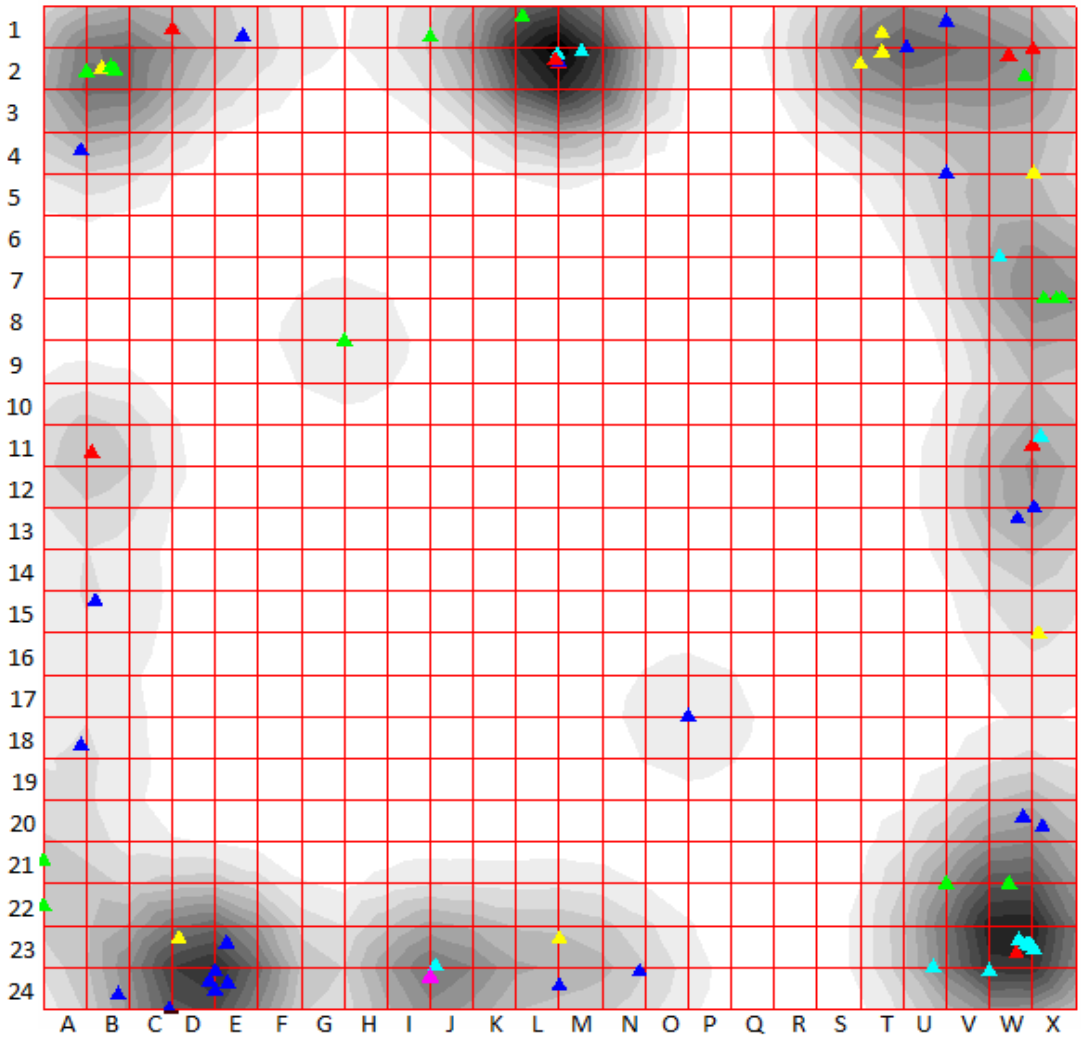

Рисунок Б. 2 − Мягкая карта 16×16 с отображением локальной плотности. Синим обозначаются вирусы, вызывающие болезни летучих мышей, красным — респираторные заболевания людей, голубым — респираторные заболевания млекопитающих, желтым — кишечные заболевания млекопитающих, розовым — гепатиты мышей и зеленым — респираторные заболевания птиц

Окончание приложения Б.

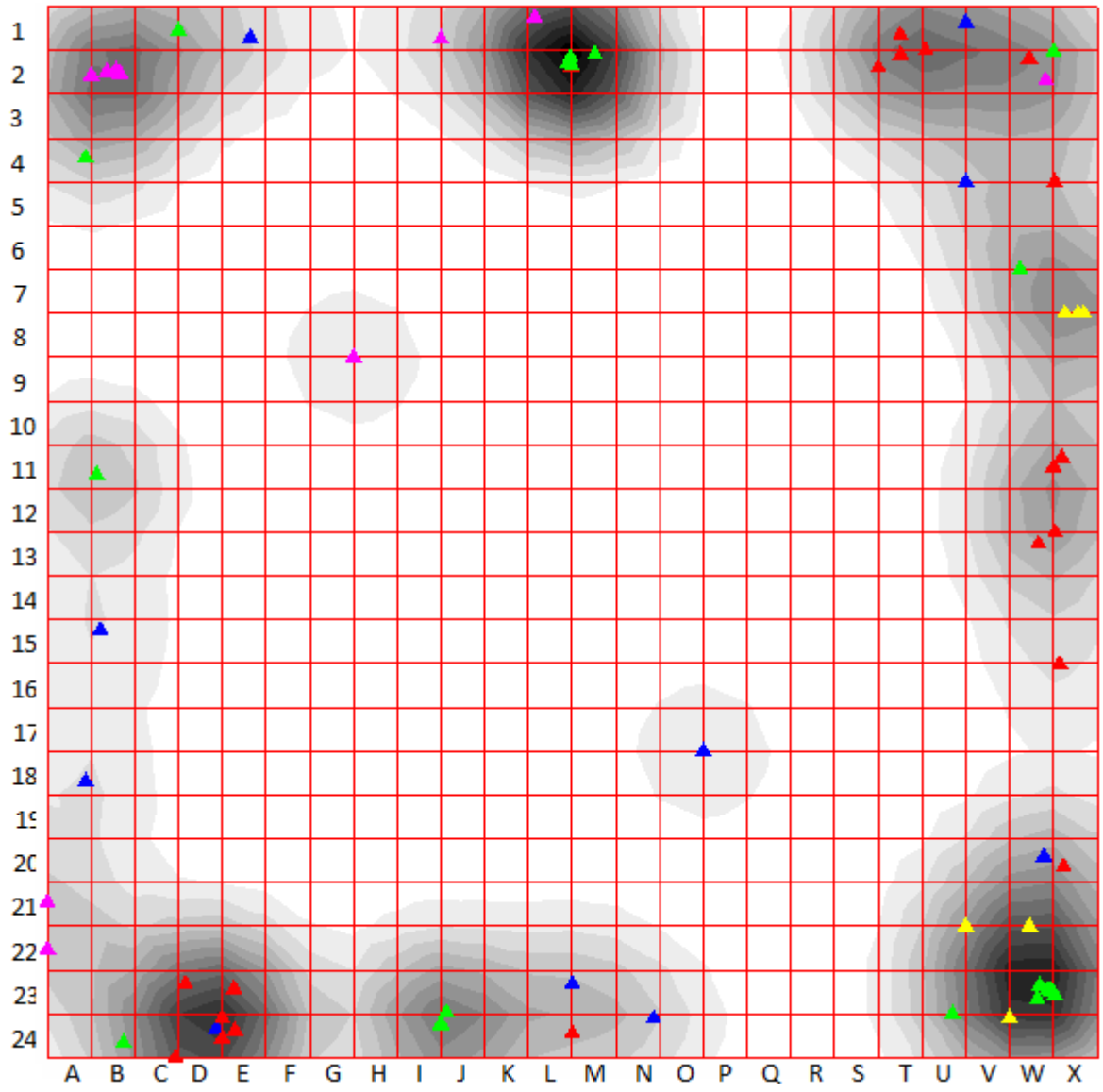

Рисунок Б. 3 − Мягкая карта 25×25 с отображением локальной плотности. Цвет точек обозначает принадлежность геномов к родам: красные — альфакоронавирусы, зеленые — бетакоронавирусы, желтые — гаммакоронавирусы, розовые — дельтакоронавирусы и синие — не классифицированные коронавирусы

Федеральное государственное автономное
образовательное учреждение
высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»
Институт фундаментальной биологии и биотехнологии
Кафедра геномики и биоинформатики

УТВЕРЖДАЮ
Заведующий кафедрой
_____ И.Е. Ямских

«25» июня 2021 г.

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

«Сравнение методов построения классификации геномов по частотам триплетов и на основе выравнивания на примере геномов семейства *Coronaviridae*»

06.04.01. «Биология»

06.04.01.06 Геномика и биоинформатика

Руководитель _____ профессор, д-р физ.-мат. наук Садовский М.Г.
подпись, дата должность, ученая степень инициалы, фамилия

Студент _____ Кириченко А.Д.
подпись, дата инициалы, фамилия

Красноярск 2021