(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
27 May 2022 (27.05.2022)

WIPO | PCT

(10) International Publication Number
WO 2022/108965 A1

(51) International Patent Classification:
G16H 50/80 (2018.01) *G16B 50/00* (2019.01)
G16B 30/00 (2019.01) *G16B 20/00* (2019.01)

(21) International Application Number:
PCT/US2021/059616

(22) International Filing Date:
17 November 2021 (17.11.2021)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
63/198,849 17 November 2020 (17.11.2020) US

(71) Applicant: **THE UNIVERSITY OF CHICAGO** [US/US]; 929 E. 57th Street, Chicago, Illinois 60637 (US).

(72) Inventor: **CHATTOPADHYAY, Ishanu**; 929 E. 57th Street, Chicago, Illinois 60637 (US).

(74) Agent: **BRENNAN, Patrick E.** et al.; Armstrong Teasdale LLP, 7700 Forsyth Boulevard, St. Louis, Missouri 63105-1847 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV,

(54) **Title:** METHODS AND SYSTEMS FOR GENOMIC BASED PREDICTION OF VIRUS MUTATION

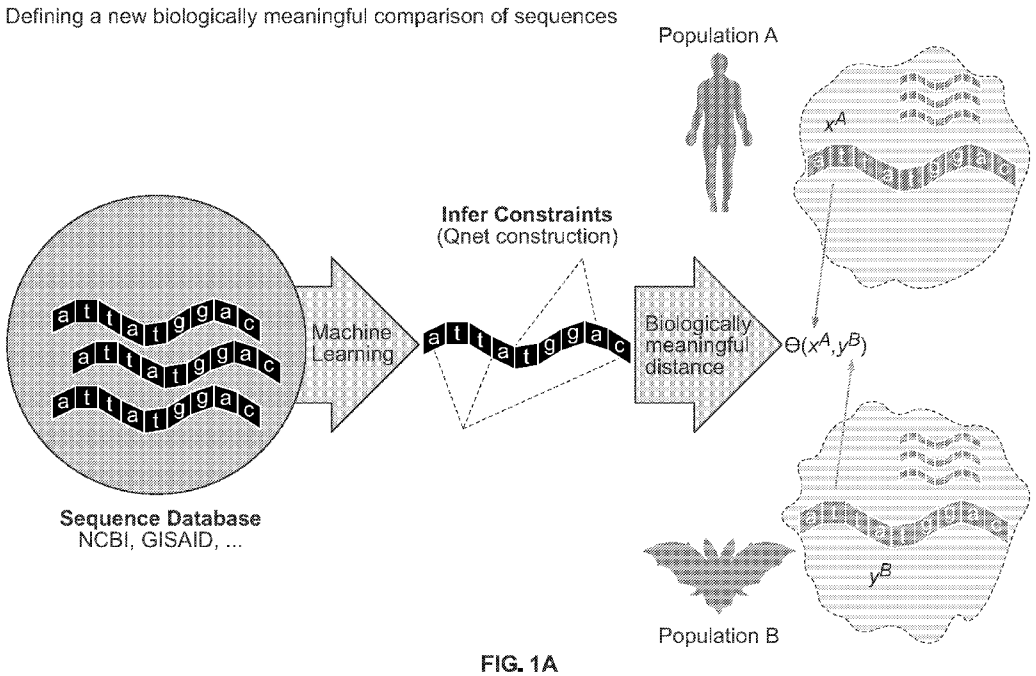

FIG. 1A

(57) **Abstract:** A method includes receiving a first plurality of aligned genomic sequences of a virus from a database. The aligned genomic sequences have a first common background. The method includes calculating a Qnet for each genomic sequence of the first plurality of aligned genomic sequences. The Qnet for each sequence is calculated by calculating a conditional inference tree for each index of the aligned genomic sequences using other indices in the aligned genomic sequences as predictive features, and calculating predictors for indices that were used as predictive features when calculating the conditional inference tree for each index.

WO 2022/108965 A1

[Continued on next page]

MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

METHODS AND SYSTEMS FOR GENOMIC BASED PREDICTION OF VIRUS MUTATION

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims priority to U.S. Provisional Patent Application Serial No. 63/198,849, filed November 17, 2020, the entire disclosure of which is hereby incorporated by reference in its entirety.

STATEMENT REGARDING FEDERALLY SPONSORED RESEARCH & DEVELOPMENT

[0001] This invention was made with government support under HR0011-18-9-0043 awarded by the Department of Defense. The government has certain rights in the invention.

TECHNICAL FIELD

[0002] This disclosure relates generally to the computational prediction of jump-likelihood between viruses, and, more specifically, to systems and methods for predicting dominant circulating strains and inter-species jump risk of viruses.

BACKGROUND

[0003] With estimated mortality rates significantly higher compared to that of the seasonal flu, the COVID-19 (SARS-CoV-2) pandemic of 2020 may be one of the most destructive pandemics of the past century. Improved preparation for the next pandemic is desirable. The ability to predict emergence of novel pathogens, such as

SARS-CoV-2, with an actionable timeline may help to reduce the negative impact of future pandemics. Current surveillance paradigms, while capable of mapping disease ecosystems, are limited in their ability to address such a challenge. Habitat encroachment, climate change, and other ecological factors increase the odds of novel viruses "jumping" from a host species to humans, resulting in pandemics such as that associated with SARS-CoV-2. At least some known efforts aimed at tracking and modeling these effects to date have not been able to successfully quantify future risk of emergence of a specific strain from a specific host species. Existence of viral diversity in hosts such as bats, swines or wild ducks, while important, might not transparently map to emergence risk, and may not address the problem at hand.

[0004] One of the key hurdles in addressing this problem has been the ability to quantitatively assess the risk of emergence from strains that circulate in the wild. Current techniques generally do not have the tools necessary to numerically compute the likelihood of a biological sequence replicating in the wild and spontaneously giving rise to another by random chance. Currently the similarity between two genomic sequences is typically measured by how many mutations it takes to change one sequence to the other, e.g. the number of mutations that make an avian flu strain human-adapted. However, without taking into account the odds of those mutations occurring in the wild, such a measure may not accurately measure the true jump-risk between species. The odds of one sequence mutating to another is a function of not just how many mutations they are apart to begin with, but also how specific mutations incrementally affect fitness. Without taking into account the constraints arising from the need to conserve function, assessing the jump-likelihood is subjective and inaccurate.

[0005] Current surveillance protocols are also generally insufficient for predicting dominant viral strains in seasonal epidemics, such as Influenza A (flu), that will circulate in the human population during any given year. Currently, the World Health Organization (WHO) decides on flu vaccine recommendations primarily from bio-surveillance data collected in the previous flu season and afterward. However, there is no established method to accurately model how the virus is expected to evolved, and it has generally been assumed that such dynamics are either too complex to model or are entirely random.

[0006] The development of computational methods and systems that are able to calculate the precise likelihood of one viral sequence mutating into another would allow for better risk assessment of inter-species jump of novel viruses and better prediction of dominant strains of seasonal viruses.

[0007] This background section is intended to introduce the reader to various aspects of art that may be related to the present disclosure, which are described and/or claimed below. This discussion is believed to be helpful in providing the reader with background information to facilitate a better understanding of the various aspects of the present disclosure. Accordingly, it should be understood that these statements are to be read in this light, and not as admissions of prior art.

BRIEF DESCRIPTION

[0008] One aspect of this disclosure is a method that includes receiving a first plurality of aligned genomic sequences of a virus from a database. The aligned genomic sequences have a first common background. The method includes calculating a

Qnet for each genomic sequence of the first plurality of aligned genomic sequences. The Qnet for each sequence is calculated by calculating a conditional inference tree for each index of the aligned genomic sequences using other indices in the aligned genomic sequences as predictive features, and calculating predictors for indices that were used as predictive features when calculating the conditional inference tree for each index.

[0009] Another aspect of this disclosure is a method that operates on a plurality of aligned genomic sequences of an organism from a database, to calculate a "Qnet" for that organism. The Qnet is a representation of computationally inferred dependencies between non-colocated mutations recorded in the database for the organism selected, as present in the wild around the time of collection of the sequences, subject to the selection pressures in effect at that time. The Qnet is calculated by estimating a conditional inference tree for each index of the aligned genomic sequences using other indices in the aligned genomic sequences as predictive features. The Qnet is thus a collection of predictors, where the target variable for a component prediction of the Qnet shows ups as predictive features for other predictors. Thus, the Qnet is a recursively dependent forest of predictors.

[0010] Various refinements exist of the notion of Qnet in relation to the above-mentioned aspects. Further features may also be incorporated in the above-mentioned aspects as well. These refinements and additional features may exist individually or in any combination. For instance, various features discussed below in relation to any of the illustrated examples may be incorporated into any of the above-described aspects, alone or in any combination.

BRIEF DESCRIPTION OF THE DRAWINGS

[0011] FIG. 1A is a schematic showing how sequence variations are used to derive a new biological metric, q-distance, for comparing differences between mutating sequences.

[0012] FIG. 1B is a schematic showing how q-distance can be used to calculate the likelihood of a jump between strains and to model future emergence risk.

[0013] FIG. 2A is a schematic showing a portion of the recursive forest underlying the Qnet for human Influenza A hemagglutinin (HA) during the 2018-2019 season.

[0014] FIG. 2B is a Qnet dependency graph for SARS-CoV-2 spike protein from the 2019 COVID-19 pandemic.

[0015] FIG. 2C is a Qnet dependency graph for Influenza A HA from the 2009 Swine Flu pandemic.

[0016] FIG. 3A is a bar graph comparing the accuracy of the WHO- and Qnet-predicted dominant strain sequence for Influenza A H1N1 HA in the southern hemisphere.

[0017] FIG. 3B is a bar graph comparing the accuracy of the WHO- and Qnet-predicted dominant strain sequence for Influenza A H1N1 neuraminidase (NA) in the southern hemisphere.

[0018] FIG. 3C is a bar graph comparing the accuracy of the WHO- and Qnet-predicted dominant strain sequence for Influenza A H3N2 HA in the southern hemisphere.

[0019] FIG. 3D is a bar graph comparing the accuracy of the WHO- and Qnet-predicted dominant strain sequence for Influenza A H3N2 NA in the southern hemisphere.

[0020] FIG. 3E is a bar graph comparing the accuracy of the WHO- and Qnet-predicted dominant strain sequence, using a multi-cluster approach, for Influenza A H1N1 NA in the southern hemisphere.

[0021] FIG. 3F is a bar graph comparing the accuracy of the WHO- and Qnet-predicted dominant strain sequence, using a multi-cluster approach, for Influenza A H3N2 NA in the southern hemisphere.

[0022] FIG. 3G is a bar graph comparing the accuracy of the WHO- and Qnet-predicted dominant strain sequence for Influenza A H1N1 HA in the northern hemisphere.

[0023] FIG. 3H is a bar graph comparing the accuracy of the WHO- and Qnet-predicted dominant strain sequence for Influenza A H1N1 NA in the northern hemisphere.

[0024] FIG. 3I is a bar graph comparing the accuracy of the WHO- and Qnet-predicted dominant strain sequence for Influenza A H3N2 HA in the northern hemisphere.

[0025] FIG. 3J is a bar graph comparing the accuracy of the WHO- and Qnet-predicted dominant strain sequence for Influenza A H3N2 NA in the northern hemisphere.

[0026] FIG. 3K is a bar graph comparing the accuracy of the WHO- and Qnet-predicted dominant strain sequence, using a multi-cluster approach, for Influenza A H1N1 NA in the northern hemisphere.

[0027] FIG. 3L is a bar graph comparing the accuracy of the WHO- and Qnet-predicted dominant strain sequence, using a multi-cluster approach, for Influenza A H3N2 NA in the southern hemisphere.

[0028] FIG. 4A is a sequence comparison of the WHO-predicted, Qnet-predicted, and actual dominant H1N1 HA strain for the northern hemisphere in 2018-2019.

[0029] FIG. 4B is a sequence comparison for the WHO-predicted, Qnet-predicted, and actual dominant H1N1 HA strain for the northern hemisphere in 2019-2020.

[0030] FIG. 4C is a sequence comparison for the WHO-predicted, Qnet-predicted, and actual dominant H1N1 HA strain for the northern hemisphere in 2018-2019.

[0031] FIG. 4D is a sequence comparison for the WHO-predicted, Qnet-predicted, and actual dominant H1N1 HA strain for the northern hemisphere in 2016-2017.

[0032] FIG. 4E is a sequence comparison for the WHO-predicted, Qnet-predicted, and actual dominant H1N1 HA strain for the southern hemisphere in 2014-2015.

[0033] FIG. 4F is a sequence comparison for the WHO-predicted, Qnet-predicted, and actual dominant H3N2 HA strain for the northern hemisphere in 2015-2016.

[0034] FIG. 4G is a molecular structure of the HA protein highlighting residues that deviate between WHO- and Qnet-predicted sequences.

[0035] FIG. 5A is a pair of bar graphs comparing the log-likelihood of spontaneous jump from (i) different animal hosts to the SARS-CoV-2 sequences of the 2019 pandemic and (ii) from specific species to their nearest SARS-CoV-2 neighbors.

[0036] FIG. 5B is a map showing the habitats of the top four most frequently occurring species from FIG. 5A (ii).

[0037] FIG. 5C is a graph plotting the log-likelihood of a jump between various sequences and their nearest neighbors against the year of collection.

[0038] FIG. 5D is a map showing the density of habitat overlap from FIG. 5B.

[0039] FIG. 6 is a phylogenetic tree derived from q-distance of various coronaviruses.

[0040] FIG. 7 is a plot showing the distribution around the seasonal dominant strain for various influenza strains.

[0041] FIG. 8A is a phylogenetic tree for various coronaviruses based on q-distance.

[0042] FIG. 8B is a phylogenetic tree for various coronaviruses based on standard edit distance.

[0043] FIG. 9A is a distribution space map comparing the distribution of Influenza HA q-sampled mutations compared to random mutations.

[0044] FIG. 9B is a graph plotting the p-blast scores for q-sampled and random mutants against known sequences.

[0045] FIG. 9C is a graph plotting the distance between mutant quasi-species vs. iteration step for q-distance and edit distance metric.

[0046] FIG. 9D is a graph plotting variance of pairwise distances for q-sampled and random mutants compared to known sequences.

[0047] FIG. 9E is a bar graph plotting the probability per year of BLAST match for q-sampled and random mutants.

[0048] FIG. 10A is a scatter plot showing the membership degree of initial hCOV-19 strains.

[0049] FIG. 10B is a distribution plot of the membership degrees for hCoV-19 strains collected on March 31, 2020 (black) and April 3, 2020 (red).

[0050] FIG. 10C is a bar graph showing the average membership p-values for hCoV-19 strains collected on different dates.

[0051] FIG. 11 is a schematic showing an exemplary method of evaluating jump-likelihood probability between different viral strains.

[0052] FIG. 12 is a schematic of a system for carrying out the method of FIG. 11.

[0053] FIG. 13 is a schematic of a computing device for the system of FIG. 12.

[0054] Although specific features of various embodiments may be shown in some drawings and not in others, this is for convenience only. Any feature of any drawing may be referenced and/or claimed in combination with any feature of any other drawing.

[0055] Unless otherwise indicated, the drawings provided herein are meant to illustrate features of embodiments of the disclosure. These features are believed to be applicable in a wide variety of systems comprising one or more embodiments of the disclosure. As such, the drawings are not meant to include all conventional features known by those of ordinary skill in the art to be required for the practice of the embodiments disclosed herein.

DETAILED DESCRIPTION

[0056] The following detailed description illustrates embodiments of the disclosure by way of example and not by way of limitation. Embodiments of the systems and methods described herein may predict the circulating strain of evolving pathogens, such as Influenza A (flu), with actionable lead time to inform vaccine design. The disclosed embodiments may be expected to predict dominant strains of future seasonal epidemics with more accuracy than the World Health Organization (WHO) recommendations used currently in flu shot compositions. Embodiments of the systems and methods described herein may also calculate the risk of evolving pathogens to jump between species, enable the origin of past pandemics to be traced and future pandemics to

be predicted. While the exemplary embodiments include prediction of the dominant seasonal strain of Influenza A and tracing the species origin of SARS-CoV-2, the described embodiments are in no way meant to be limiting. Embodiments of the systems and methods described herein may be applied to any viral sequences, granted the sequences are similar enough to conduct a sequence alignment and there is sufficient diversity of observed strains.

[0057] Embodiments of the systems and methods described herein provide for the calculation of the likelihood that a viral sequence will mutate into another, leading to emergence of new viral strains either within species (e.g., novel dominant seasonal strains of influenza) or between species (e.g., novel human SARS-CoV-2). Embodiments of the systems and methods described herein provide for building a computational model, or Qnet, for providing these predictions. A suite of customized machine learning algorithms may be used to infer the Qnet from aligned genomic sequences sampled from similar populations, for example, hemagglutinin (HA) from human Influenza A in year 2008, or the spike protein from all bat betacoronaviruses. The Qnet can predict the nucleotide distribution over the base alphabet (the four nucleic acid bases ATGC) at any specific index, conditioned on the nucleotides making up the rest of the sequence of the gene or genome fragment under consideration.

[0058] As described herein, the Qnet can learn to predict the mutational variations at each index of the genomic sequence using other indices as features, ultimately uncovering a recursive dependency structure. Collectively, these inter-dependent predictors represent the constraints that shape evolutionary trajectories driven by selection.

[0059] Embodiments of the systems and methods described herein provide for a Qnet-derived metric for calculating similarity between species, referred to as the q-distance. The q-distance can be defined as the square-root of the Jensen-Shannon (JS) divergence of the conditional distributions from one sequence to another, wherein the conditional distributions are produced by the Qnet, and averaged over the entire sequence. As a function of the q-distance, the bounds on the explicit probability of a spontaneous jump between nearby variants can be computed.

[0060] Embodiments of the systems and methods described herein demonstrably improve strain predictions for Influenza A vaccines compared to historic WHO strain predictions that form the basis for current vaccine recommendations. The recommendations produced by the systems and methods described herein are repeatedly closer to the true dominant strain, illustrating the ability of these systems and methods to correctly predict evolutionary trajectories. High season-to-season genomic variation in the key Influenza capsidic proteins is driven by two opposing influences: the need to conserve function limiting random mutations, and hyper-variability to escape recognition by neutralizing antibodies. Even a single residue change in the surface proteins might dramatically alter recognition characteristics, brought about by unpredictable changes in local or regional properties such as charge, hydropathy, side chain solvent accessibility.

[0061] Embodiments of the systems and methods described herein provide for a Qnet predicted strain (QNT) that is more likely to be closer to the dominant strain of Influenza A than the WHO predicted strain over the past two decades, and almost consistently over the last decade. The Qnet predicted strain is able to predict residues present in the dominant strain that are not predicted by the WHO. These residues are

largely localized within the receptor binding domain (RBD), with >57% occurring within the RBD on average. When the WHO-predicted strain deviates from the Qnet predicted/dominant strain matched residue, the "correct" residue is often replaced in the WHO recommendation with one that has a very different side chain, hydropathy and/or chemical properties, suggesting deviations in recognition characteristics. Because circulating strains are almost always within a few edits of the dominant strain, hosts vaccinated with the Qnet recommendation may be more likely to have season-specific antibodies that are more likely to recognize a larger cross-section of the circulating strains.

[0062] Embodiments of the systems and methods described herein demonstrate that the deviations in the Qnet predicted and WHO predicted strain residues are largely localized in the HA1 subunit of the HA molecular structure with the most frequent deviations occuring around the ≈ 200 loop, the ≈ 220 loop, the ≈180 helix, and the ≈100 helix, in addition to some residues in the HA2 subunit (≈49 and ≈124). The residues most impacted in the HA1 subunit (the globular top of the fusion protein) have been repeatedly implicated in receptor binding interactions. Embodiments of the systems and methods described herein are able to fine tune the future influenza vaccine recommendation over the state of the art, largely by modifying residue recommendations around the RBD and structures affecting recognition dynamics.

[0063] Embodiments of the systems and methods described herein can calculate the likelihood of viral strains collected across disparate host species to give rise to the observed SARS-CoV-2 strains, and offer new insights into the SARS-CoV-2 origin of the 2020 pandemic backed by precise numerical assessments. In the context of the origin problem of the 2020 pandemic, the state of the field regarding SARS-CoV-2 ancestry is

still developing, with emerging consensus on horseshoe bats of Chinese origin as the potential host of the progenitor sequence. This narrative is primarily driven by observed edit-distance and motif similarities to bat coronavirus (RaTG13, accession MN996532.1) detected in *R. affinis* from the Yunan province. However, this consensus does not explain the existence of a polybasic furin cleavage site on the spike protein which is absent in RaTG13 and related betacoronaviruses, but do occur in other human coronaviruses including HKU130.

[0064] Embodiments of the systems and methods described herein provide for a q-distance analysis that demonstrates not only the progenitor host potential of *R. affinis*, but also that a related species *R. sinicus* is a slightly more probable source for SARS-CoV-2. Also, it can be demonstrated that several other closely related horseshoe bats including *R. ferrumequinum* and *R. monoceros,* and other bats such as *T. pachypus*, *V. superans*, and *P .abramus* are also potential progenitor hosts. In addition, rodents such as *R. argentiventer*, *N. confucianus*, and *A. agrarius* have credible potential as hosting a SARS-CoV-2 ancestor.

[0065] In some embodiments, the collection times of animal samples may be plotted against the average lower log-likelihood bound on spontaneous jump to SARS-CoV-2 sequences to demonstrate dependence of the jump probability on collection date. This dependence suggests risk-progression over time. Embodiments of the systems and methods described herein demonstrate that the early risky sequences of SARS-CoV-2 are exclusively from rodents, and the risk elevates through late 2018, with the majority of the hosts switching from rodents to bats to human coronaviruses (OC43 and HKU1). This progression can be further highlighted by a LOWESS regression (local polynomial fit to

the data points), which shows an almost constant gradient of risk elevation over the past decade. Additionally, habitats of the top species that pose this risk can be overlapped, suggesting a normalized habitat distribution consistent with the presumed ground zero of the outbreak (Wuhan, China).

[0066] The quantitative assessments provided by the systems and methods described herein are not enabled by the prior art, and suggest that the evolution of SARS-CoV-2 began in rodents and jumped to bats, with final maturation in humans. The gradual elevation of risk through multiple host species, the overlapping habitats of those species, and the ability to quantify the minimum bounds on jump probability enabled by the present disclosure provide significant utility in preparing for future pandemics.

[0067] Embodiments of the systems and methods described herein provide for a data-driven metric, q-distance, to track subtle deviations in sequences, and quantify jump risk of risky viral pathogens. The systems and methods described herein demonstrate ability to predict future strains of Influenza via subtle variations in a limited set of immunologically important residues, suggesting that the systems and methods provided herein may be useful in preempting and actionably mitigating the next pandemic.

EXAMPLE METHOD FOR BUILDING VIRAL PREDICTION MODEL (QNET) AND CALCULATING Q-DISTANCE

[0068] In an example embodiment, relevant coding sequences can be collected pertaining to key genes implicated in cellular entry from two public databases (NCBI and GISAID, see e.g., TABLE 4 below for number of distinct sequences used). For example, an excess of 30,000 distinct sequences for betacoronaviruses and Influenza A can

be used, focusing on three genes or proteins. For each organism, a network of dependencies between individual mutations can be revealed through subtle variations of the aligned sequences. These dependencies can then be used to define an organism-specific model referred to as the quasi-species network, or Qnet (see e.g., FIG. 1A-FIG. 1B and 2A-FIG. 2C).

[0069] FIG. 1A-FIG. 1B is a series of schematics that provide an overview of the Qnet algorithm capability to quantify risk and rank-order strains. FIG. 1A shows that sequence variations observed in large databases can be used to distill evolutionary constraints on a genomic sequence to induce a biology-aware metric for comparing subtle differences in mutating sequences. This metric (q-distance) adjusts to specific organisms, background populations and selection pressures, and reflects the true likelihood of a spontaneous jump from one sequence to the other. This sequence level metric can be used to compute distances between a sequence and a population, and two populations. FIG. 1B illustrates that bounds on the exact likelihood of a spontaneous jump between strains can be calculated and rank-order strains observed in a diverse set of hosts to accurately model future emergence risk.

[0070] FIG. 2A-FIG. 2C illustrates the Qnet computation scheme. FIG. 2A shows, beginning with aligned sequences, a conditional inference tree can be calculated for index 1274, which involves indices 1064, 1445, 197 as predictive features. These features are automatically selected by the algorithm, as being maximally predictive of the base at 1274. Then, predictors for each of these predictive indices are calculated, e:g: the inference tree computed for index 1064 is shown, which involves index 1314 and 339 as features. Continuing, the predictor for 1314 involves indices 1263, 636 and 21, and that for

1263 involves 1314, 667 and 313. Note that recursive dependencies arise automatically: the predictor for 1263 depends on 1314, and that for 1314 depends on 1263. FIG. 2B and FIG. 2C are Qnet dependency graphs for SARS-CoV-2 spike protein and Influenza A HA respectively, illustrating the distinct patterns of mutational constraints inferred. Both HA in Influenza A and the spike protein in SARS-CoV-2 are implicated in viral entry into host cells, and crucial for host specificity of infections. Additionally, the inferred structures underscore the significantly more complex dependencies in SARS-CoV-2 compared to Influenza A.

[0071] In some embodiments, only sequences of high fitness may be observed, and only a small subset of viable sequences may ever be isolated. For example, a single 10 KB observed sequence represents a single observation in a 10,000 dimensional space; thus, enough data points may not be collected to exhaustively model the set of epistatic dependencies for any realistic genome length. Nevertheless, the systems and methods described herein demonstrate that sufficient sequences have been accumulated to yield meaningful results, at least for some RNA viruses with high mutational rates that reveal enough of the hidden constraints. The ability of the systems and methods described herein to quantitatively contrast sequence similarity addresses key aspects the viral surveillance and prediction problem, allowing for precise comparisons to be made there were not possible before.

[0072] In some embodiments, a suite of customized machine learning algorithms can be employed to infer the Qnet from aligned genomic sequences sampled from similar populations, e.g., HA from Human Influenza A in year 2008, or the spike protein from all bat beta coronaviruses. For the machine learning algorithms, conditional

inference trees may be used to predict each index as a function of the other indices, which were chosen automatically by the inference algorithm while optimizing the best split in the course of the decision tree construction. For example, sequences for the spike (S) protein on betacoronaviruses, which plays a crucial role in host cellular entry, and the Hemagglutinin (HA) and Neuraminidase (NA) for Influenza A (for subtypes H1N1 and H3N2), which are key enablers of cellular entry and exit mechanisms respectively, can be used. The sequences may be obtained from sequence databases, for example, National Center for Biotechnology Information (NCBI) virus and GISAID databases. In some embodiments, a total of 30,204 sequences can be used (see e.g., TABLE 4).

[0073] The Qnet of the present disclosure can predict the nucleotide distribution over the base alphabet (the four nucleic acid bases ATGC) at any specific index, conditioned on the nucleotides making up the rest of the sequence of the gene or genome fragment under consideration. A q-distance can then be defined as the square-root of the Jensen-Shannon (JS) divergence of the conditional distributions from one sequence to another, averaged over the entire sequence. In defining the q-distance, the mutational variations at the individual indices of a genomic sequence are not assumed to be independent (see e.g., FIG. 1A). Irrespective of whether mutations are truly random, since only certain combinations of individual mutations are viable, individual mutations across a genomic sequence replicating in the wild appear constrained, which is what can be explicitly modeled in the systems and methods described herein. The mathematical form of the q-distance metric is not arbitrary; JS divergence is a symmetricised version of the more common Kullbeck Leibler (KL) divergence between distributions, and among different possibilities, the q-distance is the simplest metric such that the likelihood of a

spontaneous jump is provably bounded above and below by simple exponential functions of the q-distance.

[0074] As an example, consider a set of random variables $X = \{X_i\}$ with $i = \in \{1, \ldots N\}$, each taking value from the respective sets $\sum_i$. A sample $x \in \prod_1^N \sum_i$ is an ordered N-tuple, consisting of a realization of each of the variables X_i with the i^{th}entry x_i being the realization of random variable X_i. The notation x_{-i} and $x^{i,\sigma}$ is used to denote:

$$x_{-i} \triangleq x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_N$$
$$x^{i,\sigma} \triangleq x_1, \cdots, x_{i-1}, \sigma, x_{i+1}, \cdots, x_N, \sigma \in \Sigma_i$$

D(S) can be used to denote the set of probability measures on a set *S*, e.g., $D(\sum_i)$ is the set of distributions on $\sum_i$. Note that *X* defines a random field over the index set $\{1, \ldots N\}$. Also, to highlight the biological relevance, the sample x is herein referenced as an amino acid or nucleotide sequence, wherein the entry at each index is identified with the corresponding protein residue or the nucleotide base pair.

[0075] For a random field $X = \{X_i\}$, indexed by $i = \in \{1, \ldots N\}$, the Qnet can be defined to be the set of predictors:

$$\Phi_i : \prod_{j \neq i} \Sigma_j \to \mathscr{D}(\Sigma_i),$$

where for a sequence x, $\Phi_i(x_{-i})$ estimates the distribution of X_i on the set $\sum_i$. Conditional inference trees can then be used as models for predictors, although more general models may also be used.

[0076] Next, the Qnet model can be used to calculate q-distance, a novel, 'biology-aware' metric that calculates the distance between sequences incorporating information about biological context and evolutionary constraints. The q-distance, informed by the dependencies modeled by the inferred Qnets, can adapt to the specific organism, allelic frequencies, and nucleotide variations in the background population. Because the role of epistatic effects in phenotypic change is well-recognized, these effects can be incorporated in a numerically precise manner to compute bounds on the likelihood of specific strains giving rise to target variants.

[0077] The q-distance can be defined as the square-root of the Jensen-Shannon (JS) divergence of the conditional distributions from one sequence to another, obtained from the Qnet model, averaged over the entire sequence. Given two sequences $x, y \in \prod_1^N \sum_i$, such that x, y are drawn from the populations P, Q inducing the $Qnet\Phi^P, \Phi^Q$, respectively, a pseudo-metric $\theta(x, y)$ can be defined as follows:

$$\theta(x, y) \triangleq \mathbb{E}_i \left(\mathbb{J}^{\frac{1}{2}} \left(\Phi_i^P(x_{-i}), \Phi_i^Q(y_{-i}) \right) \right)$$

, where $\mathbb{J}(.,.)$ is the Jensen-Shannon divergence and E_i indicates expectation over the indices. The square-root in the definition of the q-distance arises naturally from the provable bounds, and is dictated by the form of Pinsker's inequality, ensuring that the distances along a path in a constructed phylogeny sum linearly. Therefore in some embodiments, standard algorithms can used for phylogeny construction.

[0078] Importantly, the q-distance defined above is technically a pseudo-metric because distinct sequences can induce the same distributions over each index, and thus evaluate to have a zero distance. This is a desirable feature, because the distance

should not be sensitive to changes that are not biologically relevant: not all sequence variations brought about by substitutions are equally important or likely. Even with no selection pressure, random variations at an index can occur if such variations do not affect the replicative fitness. Under that scenario, the corresponding Φ_i will predict a flat distribution no matter what the input sequence is, thus contributing nothing to the overall distance. Furthermore, even if two strains x, y have the same entry at some index i, the remaining residues might induce different distributions Φ_i based on the remote dependencies, e.g., the entries in $x_{-i}y_{-i}$.

[0079] In some embodiments, the q-distance between two sequences may change if the background populations change, and not the sequences themselves (see e.g., TABLE 2 for examples, where the distance between two specific Influenza A H1N1 Hemagglutinin sequences vary when they are assumed to be collected in different years), and sequences might have a large q-distance and a small edit distance, and vice versa (although on average the two distances tend to be positively correlated, see e.g., TABLE 3). Therefore, in some embodiments, a new Qnet may be constructed whenever the background populations are expected to be substantially different. For example, separate Qnets may be constructed for betacoronavirus S protein sequences isolated from bats, rodents, cattle, non-SARS-CoV-2 human betacoronaviruses, and SARS-CoV-2 strains. As another example, for tracking drift in Influenza A, a seasonal Qnet may be constructed for each subtype and protein that is considered. Because the q-distance assumes aligned sequences of identical length (although gaps arising from alignment are acceptable and are modeled as missing data), the Qnet framework is applicable to closely related sequences, and is well-suited to track subtle changes in evolving viral populations.

[0080] In some embodiments, it may be considered whether sequences come from two different background populations , Q , e.g., if the induced Qnets $\Phi^P \Phi^Q$ are different. For example, if Qnets are constructed for H1N1 Influenza A separately for the collection years 2008 and 2009, then the same exact sequence collected in the respective years might have a non-zero distance between them, reflecting the fact that the background population the sequences arose from are different, inducing possibly different expected mutational tendencies.

[0081] Embodiments of the systems and methods described herein provide for measuring the q-distance between a sequence and a population and between two populations. The q-distance between populations can be defined using the notion of Hausdorff metric between sets:

$$\forall x \in P, y \in Q,$$
$$\theta(x, Q) = \min_{y \in Q} \theta(x, y)$$
$$\theta(P, Q) = \max \left\{ \max_{x \in P} \theta(x, Q), \max_{y \in Q} \theta(y, P) \right\}$$

[0082] Embodiments of the systems and methods described herein also provide for a quantitative test of how well the Qnet represents the data, whether predictors need to be recalculated, and whether there are sufficiently many sequences. In one implementation, an explicit membership test can formulated to quantitatively test these parameters. Given a population P inducing the Qnet Φ^P and a sequence x, the membership probability of x can be calculated as:

$$\omega_x^P \triangleq Pr(x \in P) = \prod_{j=1}^{N} \left(\Phi_j^P(x_{-j})|x_j\right)$$

Note that x_j is the j^{th} entry in x, and is thus an element in the set $\sum_j$. Since the most pertinent case is when $\sum_j$ is a finite set, $\Phi_j^P(x_{-j})\Big|_{x_j}$ is the entry in the probability mass function corresponding to the element of $\sum_j$ which appears at the j^{th} index in sequence x. This calculation can be carried out for a sequence x known to be in the population P as well, which allows for a membership degree ω_x^P to be defined. If X is a random field representing a population P, e.g., X = x is a randomly drawn sequence from P, then the membership degree ω^P is a function of the random variable X:

$$\omega^P(X) \triangleq \prod_{j=1}^{N} \left(\Phi_j^P(X_{-j})|x_j\right)$$

ω^P takes values in the unit interval [0,1], and the probability x is a member of the population P is ω^P(X = x), denoted briefly as ω_x^P or ω_x if P is clear from context. Since ω^P(X) is a random variable, sets of sequences can then be computed that better represent the population P, and ones that are on the fringe. In some embodiments, evaluation can be performed using a pre-specified significance-level if a particular sequence is not from the population P, thus identifying if predictors Φ need to be recomputed, or if the base population needs to be split. In some embodiments, a hypothesis testing scenario can be set up to determine if sequences are indeed from a test population, as follows.

[0083] For example, given a population P, inducing a Qnet Φ^P, and a sequence x, the null hypothesis can be assumed to be $x \notin P$. The null hypothesis can be rejected with a pre-specified significance α, if

$$Pr(\omega^P(X) \geq \omega^P(X = x)) \leq \alpha$$

[0084] In some embodiments, the fraction of newly observed sequences that do not reject the null hypothesis can then be used as an estimate of the species-specific divergence in population characteristics. In an example embodiment, the membership degrees are calculated for the SARS-CoV-2 sequences in the early days of the pandemic, with respect to the constructed Qnet, and illustrated in FIG. 10A-FIG. 10C. The membership degree quantifies the likelihood that a test sequence actually is generated by the inferred model (e.g., the Qnet). In this example, the distribution of membership degrees is demonstrated to be very stable, and exhibits almost no change when more sequences are added (see e.g., FIG. 10B). In addition, as more sequences are collected, the p-value improves (see e.g, FIG. 10C), and stabilizes to about 0.02, demonstrating the validity of the model.

[0085] In some embodiments, the mathematical intuition behind relating the q distance to jump-probability is illustrated by the prediction of a biased outcome when a fair coin is tossed sequentially. With an overwhelming probability, such an experiment with a fair coin should result in roughly equal number of heads and tails. However, "large deviations" can happen, and the probability of such rare events is quantifiable with existing theory. Embodiments of the systems and methods described herein demonstrate that the likelihood of a spontaneous transition of a genomic sequence to a substantially different

variant by random chance may also be similarly bounded, provided the Qnet as an estimated model of the evolutionary constraints.

[0086] In some embodiments, the Qnet framework described herein provides for rigorous computation of the bounds on several quantities of interest. The fundamental bound is on the probability of a spontaneous change of one strain to another, brought about by chance mutations. While any sequence of mutations is equally likely, the "fitness" of the resultant strain, or the probability that it will even result in a viable strain, is not. Thus the necessity of preserving function dictates that not all random changes are viable, and the probability of observing some trajectories through the sequence space are far greater than others. In some embodiments, the Qnet framework allows these constrained dynamics to be explored, as revealed by a sufficiently large set of genomic sequences.

[0087] With the exponentially exploding number of possibilities in the sequence space, it is computationally intractable to exhaustively model these dynamics. Nevertheless, possibilities can be constrained using the patterns distilled by the Qnet construction. As an example, given a sequence x of length N that transitions to a strain $y \in Q$, the following bounds exist at significance level α:

$$\omega_y^Q e^{\frac{\sqrt{8N^2}}{1-\alpha}\theta(x,y)} \geq Pr(x \rightarrow y) \geq \omega_y^Q e^{-\frac{\sqrt{8N^2}}{1-\alpha}\theta(x,y)}$$

, where ω_y^Q is the membership probability of strain y in the target population Q, and $\theta(x,y)$ is the q-distance between x, y. Using Sanov's theorem on large deviations, the probability of spontaneous jump from strain $x \in P$ to strain $y \in Q$, with the possibility $P \neq Q$, is given by:

$$Pr(x \rightarrow y) = \prod_{i=1}^{N} \left(\Phi_i^P(x_{-i})|_{y_i}\right)$$

Writing the factors on the right hand side as:

$$\Phi_i^P(x_{-i})|_{y_i} = \Phi_i^Q(y_{-i})|_{y_i} \left(\frac{\Phi_i^P(x_{-i})|_{y_i}}{\Phi_i^Q(y_{-i})|_{y_i}}\right)$$

It can be noted that $\Phi_i^P(x_{-i}), \Phi_i^Q(y_{-i})$ are distributions in the same index i, hence:

$$|\Phi_i^P(x_{-i})_{y_i} - \Phi_i^Q(y_{-i})_{y_i}| \leq \sum_{y_i \in \Sigma_i} |\Phi_i^P(x_{-i})_{y_i} - \Phi_i^Q(y_{-i})_{y_i}|$$

Using a standard refinement of Pinsker's inequality, and the relationship of Jensen-Shannon divergence with total variation, the following results:

$$\theta_i \geq \frac{1}{8}|\Phi_i^P(x_{-i})_{y_i} - \Phi_i^Q(y_{-i})_{y_i}|^2 \Rightarrow \left|1 - \frac{\Phi_i^Q(y_{-i})_{y_i}}{\Phi_i^P(x_{-i})_{y_i}}\right| \leq \frac{1}{a_0}\sqrt{8\theta_i}$$

, where a_0 is the smallest non-zero probability value of generating the entry at any index. This parameter is related to statistical significance of the bounds. First, the lower bound can be formulated as follows:

$$\log\left(\prod_{i=1}^{N} \frac{\Phi_i^P(x_{-i})|_{y_i}}{\Phi_i^Q(y_{-i})|_{y_i}}\right) = \sum_i \log\left(\frac{\Phi_i^P(x_{-i})|_{y_i}}{\Phi_i^Q(y_{-i})|_{y_i}}\right) \geq \sum_i \left(1 - \frac{\Phi_i^Q(y_{-i})_{y_i}}{\Phi_i^P(x_{-i})_{y_i}}\right) \geq \frac{\sqrt{8}}{a_0}\sum_i \theta_i^{1/2} = -\frac{\sqrt{8}N}{a_0}\theta$$

Similarly, the upper bound may be derived as:

$$\log\left(\prod_{i=1}^{N} \frac{\Phi_i^P(x_{-i})|_{y_i}}{\Phi_i^Q(y_{-i})|_{y_i}}\right) = \sum_i \log\left(\frac{\Phi_i^P(x_{-i})|_{y_i}}{\Phi_i^Q(y_{-i})|_{y_i}}\right) \leq \sum_i \left(\frac{\Phi_i^Q(y_{-i})_{y_i}}{\Phi_i^P(x_{-i})_{y_i}} - 1\right) \leq \frac{\sqrt{8}N}{a_0}\theta$$

Combining the equations, the following can be concluded:

$$\omega_y^Q e^{\frac{\sqrt{8}N}{a_0}\theta} \geqq Pr(x \rightarrow y) \geqq \omega_y^Q e^{-\frac{\sqrt{8}N}{a_0}\theta}$$

[0088] Now, interpreting a_0 as the probability of generating an unlikely event below the desired threshold (e.g., a "failure"), it can be noted that the probability of generating at least one such event is given by $1-(1-a_0)^N$. Hence, if α is the pre-specified significance level, for N > > 1:

$$a_0 \approx (1-\alpha)/N$$

Hence, it can be concluded that at significance level $\geq \alpha$, the bounds are:

$$\omega_y^Q e^{\frac{\sqrt{8}N^2}{1-\alpha}\theta} \geqq Pr(x \rightarrow y) \geqq \omega_y^Q e^{-\frac{\sqrt{8}N^2}{1-\alpha}\theta}$$

This bound can be rewritten in terms of the log-likelihood of the spontaneous jump and constants independent of the initial sequence x as:

$$|\log Pr(x \rightarrow y) - C_0| \leqq C_1\theta$$

, where the constants are given by:

$$C_0 = \log \omega_y^Q$$
$$C_1 = \frac{\sqrt{8}N^2}{1-\alpha}$$

[0089] As a consequence of the bounds defined above, it follows that the lower bound of the likelihood of a jump to a target sequence is higher if the final sequence is more fit in the target population. Note that the membership degree by definition

quantifies the probability of generating a sequence from the inferred Qnet, and since the collection of dominant strains is far more likely when a survey of a population is conducted, it follows that the membership degree is related to the qualitative notion of fitness.

[0090] Conversely, as the fitness of the initial strain (in the neighborhood of $\omega_x^P = 1$) measured by its membership degree falls, the minimum probability of going through a spontaneous jump is higher. This can be demonstrated first by noting that for $x \neq y$:

$$\omega_x^P = 1 \Rightarrow Pr(x|y) = 0$$

, which follows because each term in the product on the right hand side is either zero or one if $\omega_x^P = 1$, and there is at least one zero since $x \neq y$. To demonstrate that the suppression of probability of a jump is not simply true if $\omega_x^P = 1$ but also in the neighborhood, note that:

$$\theta_i \geq \frac{1}{8}\left|\Phi_i^P(x_{-i})_{y_i} - \Phi_i^Q(y_{-i})_{y_i}\right|^2 \Rightarrow \delta\theta_i \geq \frac{1}{4}\left(\Phi_i^P(x_{-i})_{y_i} - \Phi_i^Q(y_{-i})_{y_i}\right)\delta\Phi_i^P(x_{-i})_{y_i}$$

, which implies that in the neighborhood of $\omega_x^P = 1$:

$$\frac{\delta\theta_i}{\delta\Phi_i^P(x_{-i})_{y_i}} \geq \frac{1}{4}\left(1 - \Phi_i^Q(y_{-i})_{y_i}\right) > 0$$

This implies that the distance decreases as the membership degree of x falls, thus lowering the lower bound on the probability of a spontaneous jump. This is not necessarily true if x is not in the neighborhood of $\omega_x^P = 1$ in the first place, and so is of lesser practical interest.

[0091] The ability of the systems and methods described herein to estimate the probability of spontaneous jump between sequences in terms of θ has crucial implications. Embodiments of the systems and methods described herein allow for construction of a new phylogeny that directly relates the probability of jumps rather than the number of mutations between descendants, simulation of realistic trajectories in the sequence space from any given initial strain, and estimation of drift in the sequence space through analysis of the statistical characteristics of the diffusion occurring in the strain space.

EXAMPLE OF PREDICTING DOMINANT SEASONAL STRAINS OF INFLUENZA

[0092] Exemplary embodiments of the systems and methods described herein can be used to predict dominant circulating strains for the seasonal Influenza epidemics. Periodic adjustment of the Influenza vaccine components is necessary to account for antigenic drift. The flu shot is annually prepared at least six months in advance, and comprises a cocktail of historical strains determined by the WHO via global surveillance, hoping to match the circulating strain(s) in the upcoming flu season. A variety of hard-to-model effects hinders this prediction, and has limited vaccine effectiveness in recent years.

[0093] In some embodiments, analyzing the distribution of sequences using a Qnet inferred q-distance allows for seasonal drift to be estimated, which is particularly applicable to Influenza and Influenza-like viruses for which periodic adjustments of vaccine components are necessary to account for antigenic variations. The prediction of the dominant seasonal strain of Influenza is based on the following

assumptions: because the probability of spontaneous jump to a strain further away in the q-distance is exponentially lower, the q-centroid of the strain distribution (the centroid computed in the q-distance metric) observed over a season is expected to move slowly, and will be close to the dominant strain in the next season. Thus, the predicted dominant strain $\hat{x}^{t+1}$ at time $t+1$ as a function of the observed population at time t can be estimated as follows:

$$\hat{x}^{t+1} = \underset{x \epsilon P^t}{arg\min} \sum_{y \in P^t} \theta(x, y)$$

, where P^t is the sequence population at time t. In some embodiments, the unit of time is chosen to reflect the appropriate frequency over which vaccine components are re-assessed. For the exemplary embodiment relevant to Influenza, this is typically one year. In some embodiments, this formulation can be used to test whether the Qnet predicted strain recommendations are closer to the dominant strain in the classical edit distance, when compared against the WHO vaccine recommendation for that season.

[0094] In example, the past two decades of sequence data for Influenza A (H1N1 and H3N2) were tested and the q-distance based prediction demonstrably outperformed WHO recommendations by reducing the distance between the predicted and the dominant strain (see e.g., FIG. 3A-FIG. 3L).

[0095] FIG. 3A-FIG. 3L illustrates the relative out-performance of Qnet predictions against WHO recommendations for H1N1 and H3N2 subtypes for the HA and NA coding sequences for the northern (FIG. 3A-FIG. 3F) and southern (FIG. 3G-FIG. 3L) hemispheres. The negative bars (red) indicate the reduced edit distance between the Qnet predicted sequence and the actual dominant strain that emerged that year (e.g., Qnet

outperforms WHO). The positive bars (black) indicate outperformance of Qnet by WHO. Qnet outperformed WHO for the overwhelming majority of seasons. Note that the recommendations for the northern hemisphere are given in February, while those for the southern hemisphere are given at the end of December the previous year, as the flu season in the south begins a few months early. FIG. 3E-FIG. 3F and FIG. 3K-FIG. 3L show further possible improvement in NA predictions when three recommendations (e.g., multi-cluster) were returned instead of one each year.

[0096] In the example, the dominant strain was identified to be the one that occurs most frequently, computed as the centroid of the strain distribution observed in a given season in the classical sense (number of mutations). For H1N1 HA, the Qnet induced recommendation outperformed the WHO suggestion by >31% on average over the last 19 years, and >81% in the last decade in the northern hemisphere. The gains for NA over the same time periods for H1N1 for the north were >60% and >22% respectively. For the southern hemisphere, the gains for H1N1 over the last decade were >72% for HA, and >50%. The full table of results are shown in TABLE 1.

[0097] As another illustration pertaining to this implementation, FIG. 7 shows the distribution of the number of mutations from the seasonal dominant strain over the years for various strains of Influenza A. The quasispecies that circulates each season for each subtype was tightly distributed around the dominant strain on average.

[0098] FIG. 3A-FIG. 3L also illustrates the relative gains computed for both subtypes and the two hemispheres (since the flu season occupies distinct time periods and may have different dominant strains in the northern and southern hemispheres). In one implementation, additional improvement was demonstrated when multiple strains were

recommended every season for the vaccine cocktail (FIG. 3F-FIG. 3L). The details of the specific strain recommendations made the Qnet approach for two subtypes (H1N1, H3N2), for two genes (HA, NA) and for the northern and the southern hemispheres over the previous 19 years are enumerated in (see e.g. TABLE 2).

[0099] As another illustration pertaining to this implementation, FIG. 4A-FIG. 4G shows the sequence comparisons between WHO predicted, Qnet predicted, and dominant strains of influenza and a molecular model of the influenza HA protein. For the observed dominant strain, the correct Qnet deviations were localized within the receptor-binding domain (RBD), both for H1N1 and H3N2 for HA (see e.g., FIG. 4A). Additionally, by comparing the type, side chain area, and the accessible side chain area, the changes were observed to often have very different properties (see e.g., FIG. 4B-FIG. 4F). FIG. 4G shows the localization of the deviations in the molecular structure of HA, wherein the changes were most frequent in the HA1 subunit (the globular head), and around residues and structures that have been commonly implicated in receptor binding interactions, e.g. the ≈200 loop, the ≈220 loop and the ≈180-helix.

[00100] In at least one implementation, the key factors contributing to successful prediction of the dominant strain in the next season were investigated. A multivariate regression was performed with data diversity, the complexity of the inferred Qnet, and the edit distance of the WHO recommendation from the dominant strain as independent variables. Data diversity was defined as the number of clusters in the input set of sequences, such that any two sequences five or less mutations apart were in the same cluster. Qnet complexity was measured by the number of decision nodes in the component decision trees of the recursive forest. Several plausible structures of the regression equation

were selected, and in each case data diversity had most important and statistically significant contribution.

EXAMPLE OF IDENTIFYING ORIGIN SPECIES OF SARS-COV-2

[00101] Embodiments of the systems and methods described herein provide for the determination of an origin host species or origin reservoir of a virus, e.g., identification of the animal host of a progenitor viral sequence. In at least one implementation, the origin species of SARS-CoV-2 was determined by quantifying the likelihood of different animal species hosting the immediate progenitor. For any novel pathogen, a plausible history of emergence can generally be constructed by estimating similarity of the consensus strain with candidates in suspected animal hosts. However, interpreting a small edit distance as being indicative of a higher chance of a species-jump is problematic, particularly if multiple potential progenitor candidates arise. In contrast, a smaller average q-distance of a novel strain from animal reservoir A vs that from B implies that there is indeed a quantifiably higher probability of a jump from A.

[00102] In some embodiments, the Qnet based phylogenetic analysis provides a significantly more reliable history of the progenitor strain. For a pandemic strain $y \in H$, and an animal strain $x \in P$:

$$\log \frac{1}{\omega_y} Pr(x \to y) \gtrsim -\frac{\sqrt{8N^2}}{1-\alpha}\theta(x,y) \Rightarrow \log \frac{1}{\omega_y} \mathop{\mathbf{E}}_{x \in P} Pr(x \to y) \gtrsim -\frac{\sqrt{8N^2}}{1-\alpha} \mathop{\mathbf{E}}_{x \in P} \theta(x,y)$$

There are constants C, C' such that

$$-\log \mathop{\mathbf{E}}_{x \in P} Pr(x \to y) \lesssim C + C' \mathop{\mathbf{E}}_{x \in P} \theta(x,y)$$

Note that because N is known, C' can be calculated without the knowledge of the pandemic strain y. For the example of the SARS-CoV-2 spike protein, at 95% significance:

$$C' = 3187^2 \times 1/(1-0.95) \times \sqrt{8} = 5.75 \times 10^8$$

[00103] If the pandemic strain is known and it is desired to compare and contrast the likelihood of jump from potential hosts after the emergence event, then C can be explicitly calculated. For the example of SARS-CoV-2, this estimate was calculated as 4,805.4 (see e.g., FIG. 10A), which leads to the following linear relationship between log-likelihood of emergence and the average distance calculated in the Qnet framework:

$$-\log \mathop{\mathbf{E}}_{x \in P} Pr(x \to y) \lessapprox 4.8054 \times 10^3 + 5.75 \times 10^8 \mathop{\mathbf{E}}_{x \in P} \theta(x, y)$$

, thus providing a quantitative ranking of potential progenitor hosts. It follows that for rank-ordering potential hosts, only the average distance $E_{x \in P}\theta(x, y)$ needs to be considered. It also follows from the relative magnitudes of the constants in the case of SARS-CoV-2, that C can be ignored, yielding the approximation:

$$\log \mathop{\mathbf{E}}_{x \in P} Pr(x \to y) \gtrapprox -5.75 \times 10^8 \mathop{\mathbf{E}}_{x \in P} \theta(x, y)$$

[00104] For the example of SARS-CoV-2, the fitness term is approximately five orders of magnitude smaller, which implies the jump probabilities are roughly symmetric. However, this is not required to be true in general. At the same time, it is important to note that in some embodiments the probability of jump from strain x to strain y vs. the reverse is actually asymmetric due to the contribution from the population-specific membership degree.

[00105] The numerical bounds were estimated on the likelihood of the SARS-CoV-2 progenitor arising from specific hosts.

[00106] FIG. 5A-FIG. 5D shows the prediction of animal hosts for likely progenitors of SARS-CoV-2. FIG. 5A (i) shows the average lower bounds on the log-likelihood of jump from different animal hosts to the set of SARS-CoV-2 sequences collected in the early days of the pandemic. FIG. 5A (ii) shows the lower bounds on the log-likelihood of jump from specific species to their respective nearest SARS-CoV-2 neighbors (among sequences collected in the early days of the pandemic). FIG. 5B shows the geographic extent of the habitats of the top four most frequently occurring species among the list shown in FIG. 5A(ii). Also, the location of Wuhan, China, ground zero for COVID19 is shown. FIG. 5C plots the lower bound on log-likelihood of various sequences to their nearest neighbors over the time of collection, suggesting a trend of increasing risk over time, and across hosts, as evidenced by a nearly constant gradient LOWESS fit (black line) with 99% confidence bounds. FIG. 5D shows the normalized footprint of risk-mediating hosts from overlapping the geographic extents of the habitats of all species from the list in FIG. 5A(ii).

[00107] Betacoronavirus sequences from NCBI databases corresponding to different animal hosts were used to estimate the mean q-distance of SARS-CoV-2 sequences to bats, mouse/rodents, cattle (including camels) and pre-existing human strains including SARS-CoV1, OC43 and HKU1 strains (see e.g., FIG. 5A, showing the average log-likelihood of jump from different animal species). No *a priori* restriction was used to hosts geographically bound to South East Asia, and it was demonstrated that this localization arises naturally from the analysis. The results corroborate other studies

suggesting high probability of the progenitor originating from bats. (see e.g., FIG. 5A (i), which shows the average lower bound of the log-likelihood of a spontaneous jump from broad host categories to SARS-CoV-2 strains collected up to early March in 2020).

[00108] A ranked list of related bat species with the highest potential of hosting a SARS-CoV-2 progenitor was also identified (see e.g., FIG. 5A (ii), which shows the minimum likelihood of jump to the nearest SARS-CoV-2 strain for the respective host species). Additionally, a high likelihood of a close ancestor of SARS-CoV-2 existing in rodents was also discovered (see e.g., FIG. 5A).

EXAMPLE OF CONSTRUCTING QNET PHYLOGENETIC TREES (Q-PHYLOGENY)

[00109] In some embodiments, the systems and methods described herein can be used to construct phylogenetic trees based on the Qnet provided metric, q-distance. The majority of algorithms for constructing phylogenies generally require a notion of distance between biological sequences, and the edit distance is the one that is most commonly used. In some embodiments, the Qnet induced distance or q-distance is used to construct phylogenetic trees that are distinct from those obtained using the classical metric of edit distance. In some embodiments, the Qnet induced phylogeny (e.g., Q-phylogeny) is reflective of evolutionary change in a manner that conventional trees are not. As a path is traced in a Q-phylogeny, the probability of the changes represented by that path can be explicitly computed. This probability can be bounded above and below by a function of the total path length, e.g., the sum of the q-distances along the path. As an example, for the path

$$x = x^0 \rightarrow \cdots x^k \rightarrow \cdots x^m = z,$$

$$\frac{\sqrt{8N^2}}{1-\alpha}\Theta \geq \log Pr(x \rightarrow z) - \sum_{i=1}^{m} \log \omega_{x^i} \geq -\frac{\sqrt{8N^2}}{1-\alpha}\Theta, \text{ where } \Theta = \sum_{i=1}^{m} \theta(x^{k-1}, x^k)$$

Considering only the lower bound,

$$\log \frac{Pr(x \rightarrow z)}{\prod_{i=1}^{m} \omega_{x^i}} \geq -\frac{\sqrt{8N^2}}{1-\alpha}\Theta$$

, where $\omega_x i$ is the membership probability in the base population of the strain x^i. Thus, closer phylogenetic distance can be related to explicit probability of spontaneous jump. The definition of the distance function in the Qnet framework allows the summation in the equation, allowing the use of standard tools to construct the phylogenetic tree in some embodiments.

[00110] FIG. 6 illustrates an example of a q-distance induced phylogenetic tree. Importantly, the chronology of SARS-CoV-2 vs. existing betacoronaviruses was automatically preserved, as well as an intriguing clade-hierarchy between bat, rodent and SARS-CoV-2 strains. Some branches of the phylogenetic tree were collapsed, and the numbers in bracket list the magnitude of q-distance within which leaves were collapsed.

[00111] The q-distance induced phylogenetic tree illustrates a previously unknown role of rodents in the SARS-CoV-2 pandemic: SARS-CoV-2 strains and betacoronaviruses from rodents appeared in the same clade nested within the clade comprising betacoronaviruses from bats, rodents and SARS-CoV-2 strains (while the rodent strains were not actually closer than those isolated in bats, see e.g., FIG. 6).

[00112] FIG. 8A-FIG. 8B also illustrates examples of phylogenetic trees derived from q-distance (FIG. 8A) and classical edit distance (FIG. 8B). The numbers within brackets are the distance within which the specific branch is collapsed for visualization. The classical edit distance produced a phylogeny which clearly violates chronological ordering, as the novel coronavirus appears before strains that have been collected years before, including the SARs-1 strains. The new distance using Qnet automatically respected this known ordering.

EXAMPLE OF VALIDATING QNET CONSTRAINTS IN SILICO

[00113] *In silico* corroboration of the Qnet constraints was performed to corroborate that the constraints represented within an inferred Qnet are indeed reflective of the biology at play. The results of simulated mutational perturbations were compared to sequences from databases (for which Qnets were already constructed), and then the NCBI BLAST tool was used to identify if the perturbed sequences match with existing sequences in the databases (and if so, then where and how many matches they produce). FIG. 9A-FIG. 9E illustrates the results comparing such Qnet constrained perturbations against random variations.

[00114] FIG. 9A-FIG. 9E demonstrates the validation of q-distance *in silico* using Influenza A sequences from the NCBI database. FIG. 9A illustrates that the Qnet-induced modeling of evolutionary trajectories initiated from known haemagglutinnin (HA) sequences are distinct from random paths in the strain space. In particular, random trajectories have more variance, and more importantly, diverge to different regions of the landscape compared to Qnet predictions. FIG. 9B-FIG. 9E show that unconstrained Q-sampling produces sequences that maintain a higher degree of similarity to known

sequences, as verified by blasting against known HA sequences, have a smaller rate of growth of variance, and produce matches in closer time frames to the initial sequence. FIG. 9C shows that this is not due to simply restricting the mutational variations, which increase rapidly in both the Qnet and the classical metric.

[00115] In at least one implementation, the systems and methods described herein demonstrate that in contrast to random variations, which rapidly diverge the trajectories, the Qnet constraints produced smaller variance in the trajectories, maintained a high degree of match as trajectories are extended, and produced matches closer in time to the collection time of the initial sequence — suggesting that the Qnet does indeed capture realistic constraints.

[00116] FIG. 11 is an example method of predicting the likelihood that a viral sequence will mutate into another according to the techniques describe above. In the method, aligned genomic viral sequences are acquired from a database. A Qnet model is constructed by calculating the conditional inference tree for each index of the aligned genomic sequences. A q-distance metric is calculated from the conditional distributions produced by the Qnet model. The jump-likelihood probability between different viral strains can then be evaluated using the q-distance metric.

[00117] FIG. 12 is an example system for evaluating the jump-likelihood probability between different viral strains. The system may be used, for example, to perform the method shown in FIG. 11. The system includes a computing device and a database. The computing device is communicatively coupled to the database to receive data from the database. In some embodiments, the computing retrieves aligned sequence data from the database. Moreover, in some embodiments, the database is integrated into

the computing device, while in other embodiments, the database is located remote from the computing device.

[00118] The computing device may include, a general purpose central processing unit (CPU), a microcontroller, a reduced instruction set computer (RISC) processor, an application specific integrated circuit (ASIC), a programmable logic circuit (PLC), and/or any other circuit or processor capable of executing the functions described herein. The methods described herein may be encoded as executable instructions embodied in a computer-readable medium including, without limitation, a storage device and/or a memory device. Such instructions, when executed by a processor, cause the processor to perform at least a portion of the methods described herein.

[00119] FIG. 13 is an example computing device for use as the computing device shown in FIG. 12. The computing device includes a processor, a memory, a media output component, an input device, and a communications interface. Other embodiments include different components, additional components, and/or do not include all components shown in FIG. 13.

[00120] The computing device may include, a general purpose central processing unit (CPU), a microcontroller, a reduced instruction set computer (RISC) processor, an application specific integrated circuit (ASIC), a programmable logic circuit (PLC), and/or any other circuit or processor capable of executing the functions described herein. The methods described herein may be encoded as executable instructions embodied in a computer-readable medium including, without limitation, a storage device and/or a memory device. Such instructions, when executed by a processor, cause the processor to perform at least a portion of the methods described herein.

[00121] The processor is configured for executing instructions. In some embodiments, executable instructions are stored in the memory. The processor may include one or more processing units (e.g., in a multi-core configuration). The term processor, as used herein, refers to central processing units, microprocessors, microcontrollers, reduced instruction set circuits (RISC), application specific integrated circuits (ASIC), logic circuits, and any other circuit or processor capable of executing the functions described herein. The above are examples only, and are thus not intended to limit in any way the definition and/or meaning of the term "processor."

[00122] The media output component is configured for presenting information to the user (e.g., the operator of the system). The media output component is any component capable of conveying information to the user. In some embodiments, the media output component includes an output adapter such as a video adapter and/or an audio adapter. The output adapter is operatively connected to the processor and operatively connectable to an output device such as a display device (e.g., a liquid crystal display (LCD), light emitting diode (LED) display, organic light emitting diode (OLED) display, cathode ray tube (CRT), "electronic ink" display, one or more light emitting diodes (LEDs)) or an audio output device (e.g., a speaker or headphones).

[00123] The computing device includes, or is connected to, the input device for receiving input from the user. The input device is any device that permits the computing device to receive analog and/or digital commands, instructions, or other inputs from the user, including visual, audio, touch, button presses, stylus taps, etc. The input device may include, for example, a variable resistor, an input dial, a keyboard/keypad, a pointing device, a mouse, a stylus, a touch sensitive panel (e.g., a touch pad or a touch

screen), a gyroscope, an accelerometer, a position detector, an audio input device, or any combination thereof. A single component such as a touch screen may function as both an output device of the media output component and the input device.

[00124] The memory stores computer-readable instructions for performance of the techniques described herein. In some embodiments, the memory stores computer-readable instructions for providing a user interface to the user via media output component and, receiving and processing input from input device. The memory may include, but is not limited to, random access memory (RAM) such as dynamic RAM (DRAM) or static RAM (SRAM), read-only memory (ROM), erasable programmable read-only memory (EPROM), electrically erasable programmable read-only memory (EEPROM), and non-volatile RAM (NVRAM). Although illustrated as separate from the processor, in some embodiments the memory is combined with the processor, such as in a microcontroller or microprocessor, but may still be referred to separately. The above memory types are example only, and are thus not limiting as to the types of memory usable for storage of a computer program.

[00125] The communication interface enables the computing device to communicate with remote devices and systems, such as remote databases, remote computing devices, and the like, and may include more than one communication interface for interacting with more than one remote device or system. The communication interfaces may be wired or wireless communications interfaces that permit the computing device to communicate with the remote devices and systems directly or via a network. Wireless communication interfaces may include a radio frequency (RF) transceiver, a Bluetooth® adapter, a Wi-Fi transceiver, a ZigBee® transceiver, a near field communication (NFC)

transceiver, an infrared (IR) transceiver, and/or any other device and communication protocol for wireless communication. (Bluetooth is a registered trademark of Bluetooth Special Interest Group of Kirkland, Washington; ZigBee is a registered trademark of the ZigBee Alliance of San Ramon, California.) Wired communication interfaces may use any suitable wired communication protocol for direct communication including, without limitation, USB, RS232, I2C, SPI, analog, and proprietary I/O protocols. In some embodiments, the wired communication interfaces include a wired network adapter allowing the computing device to be coupled to a network, such as the Internet, a local area network (LAN), a wide area network (WAN), a mesh network, and/or any other network to communicate with remote devices and systems via the network.

[00126] The computer systems discussed herein may include additional, less, or alternate functionality, including that discussed elsewhere herein. The computer systems discussed herein may include or be implemented via computer-executable instructions stored on non-transitory computer-readable media or medium.

[00127] A processor or a processing element may employ artificial intelligence and/or be trained using supervised or unsupervised machine learning, and the machine learning program may employ a neural network, which may be a convolutional neural network, a deep learning neural network, or a combined learning module or program that learns in two or more fields or areas of interest. Machine learning may involve identifying and recognizing patterns in existing data in order to facilitate making predictions for subsequent data. Models may be created based upon example inputs in order to make valid and reliable predictions for novel inputs.

[00128] In some aspects, at least one of a plurality of machine learning methods and algorithms may be applied, which may include but are not limited to: linear or logistic regression, instance-based algorithms, regularization algorithms, decision trees, Bayesian networks, cluster analysis, association rule learning, artificial neural networks, deep learning, dimensionality reduction, and support vector machines. In various aspects, the implemented machine learning methods and algorithms are directed toward at least one of a plurality of categorizations of machine learning, such as supervised learning, unsupervised learning, and reinforcement learning.

[00129] In one aspect, machine learning methods and algorithms are directed toward supervised learning, which involves identifying patterns in existing data to make predictions about subsequently received data. Specifically, machine learning methods and algorithms directed toward supervised learning are "trained" through training data, which includes example inputs and associated example outputs. Based on the training data, the machine learning methods and algorithms may generate a predictive function which maps outputs to inputs and utilize the predictive function to generate machine learning outputs based on data inputs.

[00130] In another aspect, machine learning methods and algorithms are directed toward unsupervised learning, which involves finding meaningful relationships in unorganized data. Unlike supervised learning, unsupervised learning does not involve user-initiated training based on example inputs with associated outputs. Rather, in unsupervised learning, unlabeled data, which may be any combination of data inputs and/or machine learning outputs, is organized according to an algorithm-determined relationship.

[00131] In yet another aspect, machine learning methods and algorithms are directed toward reinforcement learning, which involves optimizing outputs based on feedback from a reward signal. Specifically, machine learning methods and algorithms directed toward reinforcement learning may receive a user-defined reward signal definition, receive a data input, utilize a decision-making model to generate a machine learning output based on the data input, receive a reward signal based on the reward signal definition and the machine learning output, and alter the decision-making model so as to receive a stronger reward signal for subsequently generated machine learning outputs.

[00132] As will be appreciated based upon the foregoing specification, the above-described embodiments of the disclosure may be implemented using computer programming or engineering techniques including computer software, firmware, hardware or any combination or subset thereof. Any such resulting program, having computer-readable code means, may be embodied or provided within one or more computer-readable media, thereby making a computer program product, i.e., an article of manufacture, according to the discussed embodiments of the disclosure. The computer-readable media may be, for example, but is not limited to, a fixed (hard) drive, diskette, optical disk, magnetic tape, semiconductor memory such as read-only memory (ROM), and/or any transmitting/receiving medium, such as the Internet or other communication network or link. The article of manufacture containing the computer code may be made and/or used by executing the code directly from one medium, by copying the code from one medium to another medium, or by transmitting the code over a network.

[00133] These computer programs (also known as programs, software, software applications, "apps", or code) include machine instructions for a programmable processor, and can be implemented in a high-level procedural and/or object-oriented programming language, and/or in assembly/machine language. As used herein, the terms "machine-readable medium" and "computer-readable medium" refer to any computer program product, apparatus and/or device (e.g., magnetic discs, optical disks, memory, Programmable Logic Devices (PLDs)) used to provide machine instructions and/or data to a programmable processor, including a machine-readable medium that receives machine instructions as a machine-readable signal. The "machine-readable medium" and "computer-readable medium," however, do not include transitory signals. The term "machine-readable signal" refers to any signal used to provide machine instructions and/or data to a programmable processor.

[00134] As used herein, a processor may include any programmable system including systems using micro-controllers, reduced instruction set circuits (RISC), application specific integrated circuits (ASICs), logic circuits, and any other circuit or processor capable of executing the functions described herein. The above examples are example only, and are thus not intended to limit in any way the definition and/or meaning of the term "processor."

[00135] As used herein, the terms "software" and "firmware" are interchangeable, and include any computer program stored in memory for execution by a processor, including RAM memory, ROM memory, EPROM memory, EEPROM memory, and non-volatile RAM (NVRAM) memory. The above memory types are example only,

and are thus not limiting as to the types of memory usable for storage of a computer program.

[00136] In another embodiment, a computer program is provided, and the program is embodied on a computer-readable medium. In an example embodiment, the system is executed on a single computer system, without requiring a connection to a server computer. In a further example embodiment, the system is being run in a Windows® environment (Windows is a registered trademark of Microsoft Corporation, Redmond, Washington). In yet another embodiment, the system is run on a mainframe environment and a UNIX® server environment (UNIX is a registered trademark of X/Open Company Limited located in Reading, Berkshire, United Kingdom). In a further embodiment, the system is run on an iOS® environment (iOS is a registered trademark of Cisco Systems, Inc. located in San Jose, CA). In yet a further embodiment, the system is run on a Mac OS® environment (Mac OS is a registered trademark of Apple Inc. located in Cupertino, CA). In still yet a further embodiment, the system is run on Android® OS (Android is a registered trademark of Google, Inc. of Mountain View, CA). In another embodiment, the system is run on Linux® OS (Linux is a registered trademark of Linus Torvalds of Boston, MA). The application is flexible and designed to run in various different environments without compromising any major functionality.

[00137] In some embodiments, the system includes multiple components distributed among a plurality of computer devices. One or more components may be in the form of computer-executable instructions embodied in a computer-readable medium. The systems and processes are not limited to the specific embodiments described herein. In addition, components of each system and each process can be practiced independent and

separate from other components and processes described herein. Each component and process can also be used in combination with other assembly packages and processes. The present embodiments may enhance the functionality and functioning of computers and/or computer systems.

[00138] Any logic flows depicted in the figures do not require the particular order shown, or sequential order, to achieve desirable results. In addition, other steps may be provided, or steps may be eliminated, from the described flows, and other components may be added to, or removed from, the described systems. Accordingly, other embodiments are within the scope of the following claims.

[00139] It will be appreciated that the above embodiments that have been described in particular detail are merely example or possible embodiments, and that there are many other combinations, additions, or alternatives that may be included.

[00140] Also, the particular naming of the components, capitalization of terms, the attributes, data structures, or any other programming or structural aspect is not mandatory or significant, and the mechanisms that implement the disclosure or its features may have different names, formats, or protocols. Further, the system may be implemented via a combination of hardware and software, as described, or entirely in hardware elements. Also, the particular division of functionality between the various system components described herein is merely one example, and not mandatory; functions performed by a single system component may instead be performed by multiple components, and functions performed by multiple components may instead performed by a single component.

[00141] Approximating language, as used herein throughout the specification and claims, may be applied to modify any quantitative representation that could permissibly vary without resulting in a change in the basic function to which it is related. Accordingly, a value modified by a term or terms, such as "about" and "substantially", are not to be limited to the precise value specified. In at least some instances, the approximating language may correspond to the precision of an instrument for measuring the value. Here and throughout the specification and claims, range limitations may be combined and/or interchanged, such ranges are identified and include all the sub-ranges contained therein unless context or language indicates otherwise.

[00142] Various changes, modifications, and alterations in the teachings of the present disclosure may be contemplated by those skilled in the art without departing from the intended spirit and scope thereof. It is intended that the present disclosure encompass such changes and modifications.

[00143] This written description uses examples to describe the disclosure, including the best mode, and also to enable any person skilled in the art to practice the disclosure, including making and using any devices or systems and performing any incorporated methods. The patentable scope of the disclosure is defined by the claims, and may include other examples that occur to those skilled in the art. Such other examples are intended to be within the scope of the claims if they have structural elements that do not differ from the literal language of the claims, or if they include equivalent structural elements with insubstantial differences from the literal languages of the claims.

TABLES

TABLE 1: Out-performance of Qnet recommendations over WHO for Influenza A vaccine composition

subtype	gene	hemisphere	Two decades (% Improvement)	One decade (% Improvement)
H1N1	HA	North	31.75	81.32
H1N1	HA	South	33.71	72.04
H1N1	HA	avg	32.73	76.68
H3N2	HA	North	39.39	41.38
H3N2	HA	South	31.00	28.81
H3N2	HA	avg	35.20	35.10
H1N1	NA	North	22.09	60.00
H1N1	NA	South	10.81	50.79
H1N1	NA	avg	16.45	55.40
H3N2	NA	North	28.38	45.95
H3N2	NA	South	24.69	47.73
H3N2	NA	avg	26.53	46.84

TABLE 2: Qnet induced distance varying for fixed sequence pair when background population changes (rows 1-5), sequences with small edit distance and large q-distance, and the converse (rows 6-9)

	edit dist.	sequence A	sequence B	q-distance	year A*	year B*
1	18	A/Singapore/23J/2007	A/Tennessee/UR06-0294/2007	0.0111	2007	2007
2	18	A/Singapore/23J/2007	A/Tennessee/UR06-0294/2007	0.0094	2008	2008
3	18	A/Singapore/23J/2007	A/Tennessee/UR06-0294/2007	0.0027	2009	2009
4	18	A/Singapore/23J/2007	A/Tennessee/UR06-0294/2007	0.0025	2010	2010
5	18	A/Singapore/23J/2007	A/Tennessee/UR06-0294/2007	0.6163	2007	2010
6	11	A/Nayp/law/M763/2008	A/Singapore/201/2008	0.8852	2008	2008
7	15	A/Cambodia/W0908339/2012	A/Singapore/DMS1233/2012	0.2737	2012	2012
8	126	A/South Dakota/03/2008	A/Singapore/10/2008	0.3034	2008	2008
9	141	A/Jodhpur/3248/2012	A/Cambodia/W0908339/2012	0.2405	2012	2012

*year A and year B correspond to the assumed collection years for sequences A and B respectively for the purpose of this example. Sequence A in row 1 is collected in 2007, but is assumed to be from different years in rows 2-4 to demonstrate the change in q-distance from sequence B, arising only from a change in the background population.

TABLE 3: Correlation between q-distance and edit distance between sequence pairs

phenotypes	correlation
Influenza H1N1 HA	0.76
Influenza H1N1 NA	0.74
Influenza H3N2 HA	0.85
Influenza H3N2 NA	0.79
SARS-CoV-2	0.52

TABLE 4: Number of sequences collected from public databases

Database	Strain	No. of Sequences
NCBI	Influenza H1N1 HA	7,761
NCBI	Influenza H1N1 NA	5,640
NCBI	Influenza H3N2 HA	6,568
GISAID	Influenza H3N2 HA	2,000
NCBI	Influenza H3N2 NA	4,919
GISAID	Influenza H3N2 NA	2,000
NCBI	SARS-CoV-2	24
GISAID	SARS-CoV-2	371
NCBI	betacoronavirus (non-SARS-CoV-2)	921
Total		30,204

TABLE 5: General linear model for evaluating effect of data diversity on Qnet performance

variable name	description
qnet_complexity	Cumulative number of nodes in all predictors in the corresponding Qnet
data_diversity	Number of clusters in set of input sequence where each sequence in a specific cluster is separated by at least 5 mutations from sequences not in the cluster
ldistance_WHO	Deviation of WHO predicted strain from the dominant strain

model:dev ~ qnet_complexity + data_diversity + qnet_complexity * data_diversity + idistance_WHO

Generalized Linear Model Regression Results

Dep. Variable:	dev	No. Observations:	235
Model:	GLM	Df Residuals:	230
Model Family:	Gaussian	Df Model:	4
Link Function:	Identity	Scale:	23.214
Method:	YHLS	Log-Likelihood:	-700.43
Date:	Thu, 11 Jun 2020	Deviance:	5339.2
Time:	16:05:46	Pearson chi2:	5.34e+03
No. Iterations:	3	Covariance Type:	nonrobust

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-0.1116	1.030	-0.108	0.913	-2.128	2.025
qnet_complexity	0.0005	0.000	1.075	0.282	-0.000	0.001
data_diversity	0.3137	0.126	2.501	0.011	0.072	0.560
qnet_complexity:data_diversity	-6.332e-05	5.01e-05	-1.263	0.167	-0.000	3.49e-05
idistance_WHO	-0.0348	0.035	-1.007	0.314	-0.102	0.033

model:dev ~ qnet_complexity + data_diversity + idistance_WHO

Generalized Linear Model Regression Results

Dep. Variable:	dev	No. Observations:	235
Model:	GLM	Df Residuals:	231
Model Family:	Gaussian	Df Model:	3
Link Function:	Identity	Scale:	23.306
Method:	YHLS	Log-Likelihood:	-701.41
Date:	Thu, 11 Jun 2020	Deviance:	5383.6
Time:	16:05:47	Pearson chi2:	5.38e+03
No. Iterations:	3	Covariance Type:	nonrobust

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	1.0841	0.665	1.630	0.103	-0.219	2.387
qnet_complexity	-4.12e-05	0.000	-0.156	0.876	-0.001	0.000
data_diversity	0.1788	0.075	2.392	0.017	0.032	0.325
idistance_WHO	-0.0695	0.024	-2.930	0.003	-0.116	-0.023

TABLE 6: H1N1 HA northern hemisphere

year	WHO recommendation	dominant strain	Onet recommendation	WHO error	Onet error
2001-2002	A/New Caledonia/20/99	A/Canterbury/41/2001	A/Dunedin/2/2006	4	8
2002-2003	A/New Caledonia/20/99	A/Taiwan/587/2002	A/Canterbury/41/2001	3	1
2003-2004	A/New Caledonia/20/99	A/Memphis/5/2003	A/New York/291/2002	5	2
2004-2005	A/New Caledonia/20/99	A/Thailand/Sirikaj-Rama-TT/2004	A/Memphis/5/2003	7	4
2005-2006	A/New Caledonia/20/99	A/Niedersachsen-217/2005	A/Canterbury/106/2004	8	10
2006-2007	A/New Caledonia/20/99	A/India/34980/2006	A/Auckland/819/2005	8	1
2007-2008	A/Solomon Islands/3/2006	A/Norway/1701/2007	A/Auckland/819/2005	8	11
2008-2009	A/Brisbane/59/2007	A/Pennsylvania/02/2008	A/Kentucky/UR06-0476/2007	2	2
2009-2010	A/Brisbane/59/2007	A/Singapore/GN1050/2009	A/Belem/241/2008	119	119
2010-2011	A/California/7/2009	A/England/81220740/2010	A/Singapore/GN1050/2009	5	5
2011-2012	A/California/7/2009	A/Punjab/041/2011	A/England/81220740/2010	7	2
2012-2013	A/California/7/2009	A/British Columbia/601/2012	A/Punjab/041/2011	11	4
2013-2014	A/California/7/2009	A/Moscow/CRIE-32/2013	A/Helsinki/1199/2012	10	2
2014-2015	A/California/7/2009	A/Thailand/CU-C5169/2014	A/Thailand/CU-C5169/2014	12	0
2015-2016	A/California/7/2009	A/Georgia/15/2015	A/Thailand/CU-C5169/2014	14	2
2016-2017	A/California/7/2009	A/Hawaii/21/2016	A/Hawaii/21/2016	16	0
2017-2018	A/Michigan/45/2015	A/Michigan/291/2017	A/Beijing-Huairou/SWL1335/2016	5	4
2018-2019	A/Michigan/45/2015	A/Washington/55/2018	A/Michigan/291/2017	6	1
2019-2020	A/Brisbane/02/2018	A/Kentucky/06/2019	A/Washington/55/2018	5	1
2020-2021	A/Hawaii/70/2019	-1	A/Italy/8451/2019	-1	-1

* Dominant strain is calculated as the one closest to the centroid in the strain space that year in the edit distance metric

TABLE 7: H1N1 NA southern hemisphere

year	WHO recommendation	dominant strain	Qnet recommendation	WHO error	Qnet error
2001-2002	A/New Caledonia/20/99	A/Canterbury/41/2001	A/South Canterbury/50/2000	4	6
2002-2003	A/New Caledonia/20/99	A/Taiwan/567/2002	A/Canterbury/41/2001	3	1
2003-2004	A/New Caledonia/20/99	A/Memphis/5/2003	A/New York/291/2002	3	2
2004-2005	A/New Caledonia/20/99	A/Thailand/Sirinj-Rama-TT/2004	A/Memphis/5/2003	7	4
2005-2006	A/New Caledonia/20/99	A/Niedersachsen/217/2005	A/Canterbury/106/2004	5	10
2006-2007	A/New Caledonia/20/99	A/India/34980/2006	A/Niedersachsen/217/2005	6	2
2007-2008	A/New Caledonia/20/99	A/Norway/1701/2007	A/Thailand/CU88/2008	14	6
2008-2009	A/Solomon Islands/3/2006	A/Pennsylvania/02/2008	A/Kentucky/UR06-0476/2007	9	2
2009-2010	A/Brisbane/59/2007	A/Singapore/ON1060/2009	A/Belem/241/2008	119	119
2010-2011	A/California/7/2009	A/England/01220745/2010	A/Singapore/ON1060/2009	5	1
2011-2012	A/California/7/2009	A/Punjab/041/2011	A/England/01220746/2010	7	2
2012-2013	A/California/7/2009	A/British Columbia/601/2012	A/Punjab/041/2011	15	4
2013-2014	A/California/7/2009	A/Moscow/CRIE-32/2013	A/India/P122045/2012	10	5
2014-2015	A/California/7/2009	A/Thailand/CU-C5169/2014	A/Jiangsuhaisheng/SWL1382/2013	12	4
2015-2016	A/California/7/2009	A/Georgia/15/2015	A/Thailand/CU-C5169/2014	14	2
2016-2017	A/California/7/2009	A/Hawaii/21/2016	A/Georgia/15/2015	16	2
2017-2018	A/Michigan/45/2015	A/Michigan/291/2017	A/Beijing-Huairou/SWL1335/2015	5	4
2018-2019	A/Michigan/45/2015	A/Washington/55/2018	A/Michigan/291/2017	6	1
2019-2020	A/Michigan/45/2015	A/Kentucky/08/2019	A/Washington/55/2018	7	1
2020-2021	A/Brisbane/02/2018	-1	A/Italy/8451/2019	-1	-1

* Dominant strain is calculated as the one closest to the centroid in the strain space that year in the edit distance metric

TABLE 8: H1N1 NA northern hemisphere

year	WHO recommendation	dominant strain	Qnet recommendation	WHO error	Qnet error
2001-2002	A/New Caledonia/20/99	A/New York/447/2001	A/Memphis/15/2000	4	4
2002-2003	A/New Caledonia/20/99	A/Paris/0833/2002	A/New York/447/2001	1	5
2003-2004	A/New Caledonia/20/99	A/Memphis/5/2003	A/New York/291/2002	3	5
2004-2005	A/New Caledonia/20/99	A/Singapore/14/2004	A/Memphis/5/2003	2	3
2005-2006	A/New Caledonia/20/99	A/Memphis/5/2003	A/Memphis/5/2003	3	0
2006-2007	A/New Caledonia/20/99	A/Massachusetts/08/2006	A/Sofia/361/2005	4	2
2007-2008	A/Solomon Islands/3/2006	A/Massachusetts/08/2006	A/Sofia/361/2005	3	2
2008-2009	A/Brisbane/59/2007	A/Brisbane/59/2007	A/Maryland/64/2007	6	3
2009-2010	A/Brisbane/59/2007	A/Thailand/SR08021/2009	A/Thailand/SP08207/2009	97	97
2010-2011	A/California/7/2009	A/Thailand/SR08021/2009	A/Rome/703/2009	2	3
2011-2012	A/California/7/2009	A/Tula/CRIE-GSYu/2011	A/Thailand/SR08021/2009	4	2
2012-2013	A/California/7/2009	A/Tula/CRIE-GSYu/2011	A/Tula/CRIE-GSYu/2011	4	0
2013-2014	A/California/7/2009	A/Jiangsugusu/SWL1824/2013	A/LongYan/SWL33/2013	5	3
2014-2015	A/California/7/2009	A/LongYan/SWL2457/2014	A/Utah/06/2013	9	3
2015-2016	A/California/7/2009	A/Michigan/45/2015	A/Helsinki/6986/2014	14	4
2016-2017	A/California/7/2009	A/Michigan/45/2015	A/Michigan/45/2015	14	0
2017-2018	A/Michigan/45/2015	A/Illinois/37/2017	A/Michigan/45/2015	3	3
2018-2019	A/Michigan/45/2015	A/Kenya/47/2018	A/Kenya/47/2018	4	0
2019-2020	A/Brisbane/02/2018	A/Kenya/47/2018	A/Kenya/47/2018	1	0
2020-2021	A/Hawaii/70/2019	-1	A/Kenya/47/2018	-1	-1

* Dominant strain is calculated as the one closest to the centroid in the strain space that year in the edit distance metric

TABLE 9: H1N1 NA southern hemisphere

year	WHO recommendation	dominant strain	Qnet recommendation	WHO error	Qnet error
2001-2002	A/New Caledonia/20/99	A/New York/447/2001	A/Canterbury/37/2000	4	6
2002-2003	A/New Caledonia/20/99	A/Paris/0533/2002	A/New York/447/2001	1	5
2003-2004	A/New Caledonia/20/99	A/Memphis/5/2003	A/New York/291/2002	3	5
2004-2005	A/New Caledonia/20/99	A/Singapore/14/2004	A/Memphis/5/2003	2	3
2005-2006	A/New Caledonia/20/99	A/Memphis/5/2003	A/Canterbury/106/2004	3	5
2006-2007	A/New Caledonia/20/99	A/Massachusetts/08/2006	A/Sofia/361/2005	4	2
2007-2008	A/New Caledonia/20/99	A/Massachusetts/08/2006	A/Thailand/RMSC-UDN-25/2006	4	8
2008-2009	A/Solomon Islands/3/2006	A/Brisbane/59/2007	A/Tennessee/UR06-0151/2007	15	13
2009-2010	A/Brisbane/59/2007	A/Thailand/SR08021/2009	A/Nebraska/07/2008	87	87
2010-2011	A/California/7/2009	A/Thailand/SR08021/2009	A/Roma/709/2009	2	9
2011-2012	A/California/7/2009	A/Tula/CRIE-GSYu/2011	A/Thailand/SR08021/2009	4	2
2012-2013	A/California/7/2009	A/Tula/CRIE-GSYu/2011	A/Tula/CRIE-GSYu/2011	4	9
2013-2014	A/California/7/2009	A/Jiangsuguusu/SWL1824/2013	A/Oman/SQUH-83/2012	5	4
2014-2015	A/California/7/2009	A/LongYan/SWL2457/2014	A/NanPing/SWL1546/2013	9	6
2015-2016	A/California/7/2009	A/Michigan/45/2015	A/LongYan/SWL2457/2014	14	5
2016-2017	A/California/7/2009	A/Michigan/45/2015	A/Michigan/45/2015	14	9
2017-2018	A/Michigan/45/2015	A/Illinois/37/2017	A/Michigan/45/2015	3	3
2018-2019	A/Michigan/45/2015	A/Kenya/47/2018	A/Kentucky/26/2017	4	2
2019-2020	A/Michigan/45/2015	A/Kenya/47/2018	A/Kenya/47/2018	4	0
2020-2021	A/Brisbane/02/2018	-1	A/Kenya/47/2018	-1	-1

* Dominant strain is calculated as the one closest to the centroid in the strain space that year in the edit distance metric

TABLE 10: H3N2 HA northern hemisphere

year	WHO recommendation	dominant strain	Qnet recommendation	WHO error	Qnet error
2005-2006	A/California/7/2004	A/Denmark/195/2005	A/Taiwan/169/2004	10	2
2006-2007	A/Wisconsin/67/2005	A/New York/5/2006	A/South Australia/22/2005	5	4
2007-2008	A/Wisconsin/67/2005	A/Tennessee/11/2007	A/Colorado/05/2006	8	5
2008-2009	A/Brisbane/10/2007	A/Massachusetts/13/2008	A/Tennessee/11/2007	3	2
2009-2010	A/Brisbane/10/2007	A/Hawaii/14/2009	A/Wenzhou/03/2008	7	8
2010-2011	A/Perth/16/2009	A/Utah/12/2010	A/Hawaii/14/2009	8	7
2011-2012	A/Perth/16/2009	A/Paul/14202/2011	A/Utah/12/2010	4	4
2012-2013	A/Victoria/361/2011	A/Atbara/927/2012	A/Tehran/995/2012	4	3
2013-2014	A/Victoria/361/2011	A/Delaware/01/2013	A/Singapore/H2012.934/2012	4	1
2014-2015	A/Texas/50/2012	A/Hong Kong/4801/2014	A/Nebraska/03/2013	10	9
2015-2016	A/Switzerland/9715293/2013	A/Hong Kong/4801/2014	A/Hong Kong/4801/2014	10	0
2016-2017	A/Hong Kong/4801/2014	A/Hong Kong/4801/2014	A/Hong Kong/4801/2014	0	0
2017-2018	A/Hong Kong/4801/2014	A/Maryland/25/2017	A/New York/03/2016	3	1
2018-2019	A/Singapore/INFIMH-16-0019/2016	A/Vermont/04/2019	A/Ontario/038/2017	8	5
2019-2020	A/Kansas/14/2017	A/Kentucky/27/2019	A/California/7330/2016	16	12
2020-2021	A/Hong Kong/2671/2019	-1	A/Kentucky/27/2019	-1	-1

* Dominant strain is calculated as the one closest to the centroid in the strain space that year in the edit distance metric

TABLE 11: H3N2 HA southern hemisphere

year	WHO recommendation	dominant strain	Qnet recommendation	WHO error	Qnet error
2005-2006	A/Wellington/1/2004	A/Denmark/195/2005	A/Waikato/21/2004	3	3
2006-2007	A/California/7/2004	A/New York/5/2006	A/South Australia/22/2005	12	4
2007-2008	A/Wisconsin/67/2005	A/Tennessee/11/2007	A/New York/323/2006	8	5
2008-2009	A/Brisbane/10/2007	A/Massachusetts/13/2008	A/Tennessee/11/2007	3	2
2009-2010	A/Brisbane/10/2007	A/Hawaii/14/2009	A/Manhean/03/2008	7	6
2010-2011	A/Perth/16/2009	A/Utah/12/2010	A/Hawaii/14/2009	8	7
2011-2012	A/Perth/16/2009	A/Piaui/14202/2011	A/Utah/12/2010	4	4
2012-2013	A/Perth/16/2009	A/Aborz/927/2012	A/Piaui/14202/2011	8	4
2013-2014	A/Victoria/361/2011	A/Delaware/01/2013	A/Callao/IPE00830/2012	4	7
2014-2015	A/Texas/50/2012	A/Hong Kong/4801/2014	A/Delaware/01/2013	10	7
2015-2016	A/Switzerland/9715293/2013	A/Hong Kong/4801/2014	A/Hong Kong/4801/2014	10	9
2016-2017	A/Hong Kong/4801/2014	A/Hong Kong/4801/2014	A/Hong Kong/4801/2014	0	0
2017-2018	A/Hong Kong/4801/2014	A/Maryland/25/2017	A/Ontario/196/2016	3	4
2018-2019	A/Singapore/INFIMH-16-0019/2016	A/Vermont/04/2015	A/Ontario/836/2017	8	5
2019-2020	A/Switzerland/8060/2017	A/Kentucky/27/2019	A/California/7330/2018	13	12
2020-2021	A/South Australia/34/2019	-1	A/Kentucky/27/2019	-1	-1

* Dominant strain is calculated as the one closest to the centroid in the strain space that year in the edit distance metric

TABLE 12: H3N2 NA northern hemisphere

year	WHO recommendation	dominant strain	Qnet recommendation	WHO error	Qnet error
2003-2004	A/Moscow/10/99	A/Denmark/167/2003	A/New York/181/2002	13	3
2004-2005	A/Fujian/411/2002	A/Hyogo/36/2004	A/New York/20/2003	3	16
2005-2006	A/California/7/2004	A/Denmark/203/2005	A/Denmark/203/2005	4	0
2006-2007	A/Wisconsin/67/2005	A/Berlin/32/2006	A/Mexico/InDRE2237/2005	1	1
2007-2008	A/Wisconsin/67/2005	A/Brazil/80/2007	A/Macau/557/2005	9	7
2008-2009	A/Brisbane/10/2007	A/Perth/16/2009	A/Brazil/80/2007	3	2
2009-2010	A/Brisbane/10/2007	A/Perth/16/2009	A/Wisconsin/24/2008	3	1
2010-2011	A/Perth/16/2009	A/California/17/2010	A/New York/79/2009	2	3
2011-2012	A/Perth/16/2009	A/Texas/14/2011	A/Virginia/05/2010	3	2
2012-2013	A/Victoria/361/2011	A/New York/02/2012	A/Singapore/IC2011.493/2011	4	1
2013-2014	A/Victoria/361/2011	A/Michigan/02/2013	A/Idaho/36/2012	3	1
2014-2015	A/Texas/50/2012	A/Tehran/69834/2014	A/Michigan/02/2013	3	1
2015-2016	A/Switzerland/9715293/2013	A/Parma/471/2015	A/Parma/471/2015	3	0
2016-2017	A/Hong Kong/4801/2014	A/North Carolina/62/2016	A/Parma/471/2015	7	2
2017-2018	A/Hong Kong/4801/2014	A/Texas/277/2017	A/Texas/277/2017	8	9
2018-2019	A/Singapore/INFIMH-16-0019/2016	A/Japan/NHRC_FDX70352/2018	A/Netherlands/3530/2017	4	3
2019-2020	A/Kansas/14/2017	A/Washington/9757/2019		3	11
2020-2021	A/Hong Kong/2671/2019	-1	A/Washington/9757/2019	-1	-1

* Dominant strain is calculated as the one closest to the centroid in the strain space that year in the edit distance metric

TABLE 13: H3N2 NA southern hemisphere

year	WHO recommendation	dominant strain	Qnet recommendation	WHO error	Qnet error
2003-2004	A/Moscow/10/99	A/Denmark/107/2003	A/New York/181/2002	10	3
2004-2005	A/Fujian/411/2002	A/Hyogo/36/2004	A/New York/20/2003	3	16
2005-2006	A/Wellington/1/2004	A/Denmark/203/2005	A/Wellington/1/2004	2	2
2006-2007	A/California/7/2004	A/Berlin/32/2006	A/Mexico/InDRE2227/2005	3	1
2007-2008	A/Wisconsin/67/2005	A/Brazil/80/2007	A/Ohio/06/2006	8	10
2008-2009	A/Brisbane/10/2007	A/Perth/16/2009	A/Brazil/80/2007	3	2
2009-2010	A/Brisbane/10/2007	A/Perth/16/2009	A/Wisconsin/24/2008	3	1
2010-2011	A/Perth/16/2009	A/California/17/2010	A/New York/70/2009	2	3
2011-2012	A/Perth/16/2009	A/Texas/14/2011	A/Virginia/05/2010	3	2
2012-2013	A/Perth/16/2009	A/New York/02/2013	A/Texas/14/2011	4	1
2013-2014	A/Victoria/361/2011	A/Michigan/02/2013	A/New York/02/2013	3	3
2014-2015	A/Texas/50/2012	A/Tehran/89634/2014	A/Michigan/02/2013	3	1
2015-2016	A/Switzerland/9715293/2013	A/Parma/471/2015	A/Tehran/89634/2014	3	2
2016-2017	A/Hong Kong/4801/2014	A/North Carolina/62/2016	A/Parma/471/2015	7	2
2017-2018	A/Hong Kong/4801/2014	A/Texas/277/2017	A/Texas/277/2017	8	0
2018-2019	A/Singapore/INFIMH-16-0019/2016	A/Japan/NHRC_FDX70352/2018	A/Texas/277/2017	4	3
2019-2020	A/Switzerland/8060/2017	A/Washington/9757/2019	A/Pennsylvania/317/2018	10	10
2020-2021	A/South Australia/34/2019	-1	A/Washington/9757/2019	-1	-1

* Dominant strain is calculated as the one closest to the centroid in the strain space that year in the edit distance metric

TABLE 14: H1N1 NA southern hemisphere (multi-cluster)

year	WHO recommendation	Qnet error0	Qnet error1	WHO error	Qnet recommendation_0	Qnet recommendation_1
2001-2002	A/New Caledonia/20/99	1	8	4	A/New South Wales/28/2000	A/Canterbury/37/2000
2002-2003	A/New Caledonia/20/99	0	5	1	A/Paris/0833/2002	A/New York/447/2001
2003-2004	A/New Caledonia/20/99	2	8	3	A/Paris/0833/2002	A/Taiwan/141/2002
2004-2005	A/New Caledonia/20/99	3	4	2	A/Memphis/5/2003	A/Hanoi/1004/2003
2005-2006	A/New Caledonia/20/99	0	1	3	A/Memphis/5/2003	A/Massachusetts/08/2006
2006-2007	A/New Caledonia/20/99	2	6	4	A/Sofia/361/2005	A/Wellington/11/2005
2007-2008	A/New Caledonia/20/99	4	8	4	A/New Caledonia/20/99	A/New York/8/2006
2008-2009	A/Solomon Islands/3/2006	13	19	15	A/Tennessee/UR06-0151/2007	A/Ohio/UR06-0178/2007
2009-2010	A/Brisbane/59/2007	88	90	87	A/Sendai/TU66/2008	A/Japan/618/2006
2010-2011	A/California/7/2009	1	6	2	A/South Carolina/WRAIR1645P/2009	A/Wisconsin/629-D00869/2009
2011-2012	A/California/7/2009	1	3	4	A/England/2168/633/2010	A/Hangzhou/178/2010
2012-2013	A/California/7/2009	1	22	4	A/Jashkar-Ola/CRIE-BLP/2011	A/Rio Grande do Sul/578/2011
2013-2014	A/California/7/2009	4	13	5	A/Thailand/MR10580/2012	A/Mexico/NIMEGEN-INER 15/2012
2014-2015	A/California/7/2009	3	7	9	A/Minnesota/62/2013	A/Helsinki/430/2013
2015-2016	A/California/7/2009	4	7	14	A/Helsinki/808M/2014	A/Virginia/NHRC4307S6/2014
2016-2017	A/California/7/2009	0	3	14	A/Michigan/45/2015	A/Colorado/30/2015
2017-2018	A/Michigan/45/2015	3	8	3	A/Michigan/45/2015	A/Arizona/03/2016
2018-2019	A/Michigan/45/2015	0	4	4	A/Kenya/47/2018	A/Michigan/45/2015
2019-2020	A/Michigan/45/2015	0	2	4	A/Kenya/47/2018	A/Colorado/7682/2018
2020-2021	A/Brisbane/02/2018	-1	-1	-1	A/California/NHRC-GID_BOX-IL-0012/2019	A/Indiana/30/2019

* Dominant strain is calculated as the one closest to the centroid in the strain space that year in the edit distance metric

TABLE 15: H3N2 NA southern hemisphere (multi-cluster)

year	WHO recommendation	Qnet error0	Qnet error1	WHO error	Qnet recommendation_0	Qnet recommendation_1
2003-2004	A/Moscow/10/99	4	5	13	A/Auckland/612/2002	A/New York/87/2002
2004-2005	A/Fujian/411/2002	16	18	3	A/New York/20/2003	A/New York/12/2003
2005-2006	A/Wellington/1/2004	1	7	2	A/New York/358/2004	A/Singapore/36/2004
2006-2007	A/California/7/2004	3	8	0	A/Macau/557/2005	A/Hong Kong/HKU58/2005
2007-2008	A/Wisconsin/67/2005	0	10	8	A/Brazil/60/2007	A/Wisconsin/44/2006
2008-2009	A/Brisbane/10/2007	4	19	3	A/Missouri/06/2007	A/Japan/72/2007
2009-2010	A/Brisbane/10/2007	1	7	3	A/Wisconsin/24/2008	A/Mississippi/US07-5042/2008
2010-2011	A/Perth/16/2009	3	8	2	A/New York/70/2009	A/Japan/883/2009
2011-2012	A/Perth/16/2009	2	2	3	A/California/19/2010	A/Virginia/05/2010
2012-2013	A/Perth/16/2009	1	12	4	A/Texas/14/2011	A/Singapore/GP1684/2011
2013-2014	A/Victoria/361/2011	1	5	3	A/Idaho/38/2012	A/Pavia/135/2012
2014-2015	A/Texas/50/2012	1	1	3	A/Nevada/05/2013	A/Michigan/02/2013
2015-2016	A/Switzerland/9715293/2013	0	4	3	A/Parma/471/2015	A/Iran/91244/2014
2016-2017	A/Hong Kong/4801/2014	1	25	7	A/New Jersey/13/2015	A/California/NHRC_BRD41058N/2015
2017-2018	A/Hong Kong/4801/2014	1	4	9	A/Texas/277/2017	A/Victoria/868/2015
2018-2019	A/Singapore/INFIMH-16-0019/2016	2	4	3	A/Netherlands/3530/2017	A/Washington/17/2017
2019-2020	A/Switzerland/8060/2017	4	10	10	A/England/538/2018	A/California/BRD12490N/2018
2020-2021	A/South Australia/34/2019	-1	-1	-1	A/South Australia/34/2019	A/Washington/9757/2019

* Dominant strain is calculated as the one closest to the centroid in the strain space that year in the edit distance metric

TABLE 16: Neighbors at the edge of emergence

accession	country	date	qdistance*	host	log-likelihood bound†
MG197717	China	2015-07-06	0.5934	human|Human coronavirus OC43|	-344880879.6819
MG197719	China	2015-06-04	0.5995	human|Human coronavirus OC43|	-344700037.6025
MG197710	China	2015-06-06	0.6002	human|Human coronavirus OC43|	-346145882.3718
MH940845	Thailand	2017-05-04	0.6017	human|Human coronavirus HKU1|	-346905386.0759
MG197711	China	2015-06-09	0.6035	human|Human coronavirus OC43|	-347017496.9155
MG197716	China	2015-06-06	0.6053	human|Human coronavirus OC43|	-348034981.4953
MG197715	China	2015-05-01	0.6053	human|Human coronavirus OC43|	-348055196.8970
KF294457	China	2012-01-01	0.6058	Rhinolophus monoceros	-348315726.8189
KJ473822	China	2012-01-01	0.6059	Tylonycteris pachypus	-348379385.0294
MK211376	China	2016-09-01	0.6065	Rhinolophus affinis	-348745110.7526
MH002342	China	2013-05-03	0.6065	Pipistrellus bat coronavirus HKU5	-348745431.9254
KJ473816	China	2013-01-01	0.6066	Rhinolophus sinicus	-348778827.4854
KJ473812	China	2013-01-01	0.6066	Rhinolophus ferrumequinum	-348807413.3783
MG772933	China	2017-02-01	0.6066	Rhinolophus sinicus	-348814549.9518
MK211379	China	2016-09-01	0.6067	Rhinolophus affinis	-348846490.8570
MK211375	China	2016-09-01	0.6067	Rhinolophus affinis	-348867989.3104
MK211374	China	2016-09-01	0.6068	Rhinolophus sp.	-348893681.6418
KJ473821	China	2014-05-06	0.6071	Vespertilio superans	-349070080.5700
KF569996	China	2011-01-01	0.6095	Rhinolophus affinis	-350440764.5785
MN611520	China	2018-03-01	0.6095	Pipistrellus abramus	-350462309.6142
KP886809	China	2013-05-23	0.6095	Rhinolophus Ferrumequinum	-350486588.0164
KF886808	China	2013-05-23	0.6095	Rhinolophus Ferrumequinum	-350486588.0164
MN611519	China	2018-03-01	0.6097	Tylonycteris pachypus	-350572885.7797
NC_025217	China	2013-04-29	0.6097	Hipposideros pratti	-350580907.8832
MK211377	China	2016-09-01	0.6105	Rhinolophus affinis	-351127765.1568
KJ473820	China	2013-01-01	0.6118	Pipistrellus abramus	-351796100.8827
MN996532‡	China	2019-07-24	0.6155	Rhinolophus affinis	-353944009.5536
MH002341	China	2014-05-28	0.6167	Pipistrellus bat coronavirus HKU5	-354632651.3696
MH687968	Viet Nam	2014-11-14	0.6174	Rattus argentiventer	-355804271.6441
MH687978	Viet Nam	2015-02-04	0.6183	Rattus argentiventer	-356553631.3715
MH687969	Viet Nam	2014-11-13	0.6184	Rattus argentiventer	-356566733.0149
KF294372	China	2011-01-01	0.6185	Niviventer confucianus	-356664180.7144
MH687974	Viet Nam	2014-11-12	0.6187	Rattus argentiventer	-356792457.2680
MH687973	Viet Nam	2014-11-12	0.6189	Rattus argentiventer	-356892785.0568
MH687972	Viet Nam	2014-11-12	0.6190	Rattus argentiventer	-356956649.5930
KF294370	China	2013-01-01	0.6192	Rattus tanezumi	-357024166.0313
KF294371	China	2013-01-01	0.6192	Rattus losea	-357040365.5036
MH687971	Viet Nam	2014-11-12	0.6194	Rattus argentiventer	-357161570.0562
MH687977	Viet Nam	2015-02-04	0.6199	Rattus argentiventer	-357466591.1490
KF294357	China	2011-01-01	0.6214	Apodemus agrarius	-358298941.1683
KM349744	China	2012-05-17	0.6219	Rattus norvegicus (Norway rat)	-358570433.8439
NC_026011	China	2012-05-17	0.6219	Rattus norvegicus (Norway rat)	-358570433.8433
KM349743	China	2012-05-17	0.6220	Rattus norvegicus (Norway rat)	-358648695.7536

* qdistance: Smaller values implies higher risk
† Likelihood lower bound: Larger values implies higher risk
‡ReTG13

TABLE 17: Numbering conversion to PDM09 and H3 schemes

Query	H1N1pdm	H3	Query	H1N1pdm	H3	Query	H1N1pdm	H3	Query	H1N1pdm	H3
1	-	-	77	60	69	157	140	143	-	-	-
2	-	-	78	61	70	158	141	144	-	-	-
3	-	-	79	62	71	159	142	145	-	-	-
4	-	-	80	63	72	160	143	146	238	221	224
5	-	-	81	64	73	161	144	147	239	222	225
6	-	-	82	65	74	162	145	148	240	223	226
7	-	-	83	66	75	163	146	149	241	224	227
8	-	-	84	67	76	164	147	150	242	225	228
9	-	-	85	68	77	165	148	151	243	226	229
10	-	-	86	69	78	166	149	152	244	227	230
11	-	-	87	70	79	167	150	153	245	228	231
12	-	-	88	71	80	168	151	154	246	229	232
13	-	-	89	72	81	169	152	155	247	230	233
14	-	-	90	73	82	170	153	156	248	231	234
15	-	-	91	74	-	171	154	157	249	232	235
16	-	-	92	75	83	172	155	158	250	233	236
17	-	-	93	76	84	-	-	-	251	234	237
-	-	1	94	77	85	-	-	-	252	235	238
-	-	2	95	78	86	-	-	-	253	236	239
-	-	3	96	79	87	-	-	-	254	237	240
-	-	4	97	80	88	173	156	159	255	238	241
-	-	5	98	81	89	174	157	160	256	239	242
-	-	6	99	82	90	175	158	161	257	240	243
-	-	7	100	83	91	176	159	162	258	241	244
-	-	8	101	84	92	177	160	163	259	242	245
-	-	9	102	85	-	178	161	164	260	243	246
-	-	10	103	86	93	179	162	165	261	244	247
18	1	11	104	87	94	180	163	166	262	245	248
19	2	12	105	88	95	181	164	167	263	246	249
20	3	13	106	89	96	182	165	168	264	247	250
21	4	14	107	90	97	183	166	169	265	248	251
22	5	15	108	91	98	184	167	170	266	249	252
23	6	16	109	92	99	-	-	-	267	250	253
24	7	17	110	93	100	185	168	171	268	251	254
25	8	18	111	94	101	186	169	172	269	252	255
26	9	19	112	95	102	187	170	173	270	253	256
27	10	20	-	-	-		-	-	271	254	257
28	11	21	-	-	-	188	171	174	272	255	258
29	12	22	113	96	103	189	172	175	273	256	259
30	13	23	114	97	104	190	173	176	274	257	260
31	14	24	115	98	105	191	174	177	275	258	261
32	15	25	116	99	106	192	175	178	276	259	262
33	16	26	117	100	107	193	176	179	-	-	-
34	17	27	118	101	108	194	177	180	-	-	-
35	18	28	119	102	109	195	178	181	-	-	-
36	19	29	120	103	110	196	179	182	-	-	-
37	20	30	121	104	111	197	180	183	-	-	-
38	21	31	122	105	112	198	181	184	-	-	-
39	22	32	123	106	113	199	182	185	-	-	-
40	23	33	124	107	114	200	183	186	-	-	-
41	24	34	125	108	115	201	184	187	-	-	-
42	25	35	126	109	116	202	185	188	277	260	-
43	26	36	127	110	117	203	186	189	278	261	263
44	27	37	128	111	118	204	187	190	279	262	264
45	28	38	129	112	119	205	188	191	280	263	265
46	29	39	130	113	120	206	189	192	281	264	266
47	30	40	131	114	121	207	190	193	282	265	267
48	31	41	132	115	122	208	191	194	283	266	268
49	32	42	133	116	123	209	192	195	284	267	269
50	33	43	-	-	-	210	193	196	285	268	270
51	34	44	-	-	-	211	194	197	286	269	271
52	35	45	134	117	124	212	195	198	287	270	272
53	36	46	135	118	125	213	196	199	288	271	273
54	37	47	136	119	-	-	-	-	289	272	274
55	38	48	137	120	-	214	197	200	290	273	275
56	39	49	138	121	-	215	198	201	291	274	276
57	40	50	139	122	126	216	199	202	292	275	277
58	41	51	140	123	127	217	200	203	293	276	278
59	42	52	141	124	128	218	201	204	294	277	279
60	43	53	-	-	-	219	202	205	295	278	280
61	44	54	-	-	-	220	203	206	296	279	281
62	45	-	-	-	-	221	204	207	297	280	282
63	46	55	-	-	-	222	205	208	298	281	283
64	47	56	-	-	-	223	206	209	299	282	284
65	48	57	142	125	129	224	207	210	300	283	285
66	49	58	143	126	130	225	208	211	-	-	-
67	50	59	144	127	131	226	209	212	301	284	286
68	51	60	145	128	132	227	210	213	302	285	287
-	-	-	146	129	133	228	211	214	303	286	288
-	-	-	147	130	-	229	212	215	304	287	289
-	-	-	148	131	134	230	213	216	305	288	290
-	-	-	149	132	135	231	214	217	306	289	291
-	-	-	150	133	136	232	215	218	307	290	292
69	52	61	151	134	137	233	216	219	308	291	293
70	53	62	152	135	138	234	217	220	309	292	294
71	54	63	153	136	139	235	218	221	310	293	295
72	55	64	154	137	140	236	219	222	311	294	296
73	56	65	155	138	141	237	220	223	-	-	-
74	57	66	-	-	-	-	-	-	312	295	297
75	58	67	156	139	142	-	-	-	313	296	298

WHAT IS CLAIMED IS:

1. A method comprising:

receiving a first plurality of aligned genomic sequences of a virus from a database, the aligned genomic sequences having a first common background; and

calculating a Qnet for each genomic sequence of the first plurality of aligned genomic sequences by:

calculating a conditional inference tree for each index of the aligned genomic sequences using other indices in the aligned genomic sequences as predictive features; and

calculating predictors for indices that were used as predictive features when calculating the conditional inference tree for each index.

2. The method of claim 1, wherein the first common background of the first plurality of aligned genomic sequences comprises a common year of collection.

3. The method of claim 1, wherein the first common background of the first plurality of aligned genomic sequences comprises a common species from which the aligned genomic sequences were collected.

4. The method of claim 1, further comprising calculating distances between pairs of sequences of the first plurality of aligned genomic sequences based on the Qnet.

5. The method of claim 4, wherein calculating the distances comprises calculating q-distances as the square root of the Jensen-Shannon divergence of conditional

nucleotide distributions from the Qnet for a sequence to conditional nucleotide distributions from the Qnet for a different sequence.

6. The method of claim 5, further comprising predicting a future dominant strain of the virus based on the calculated q-distances.

7. The method of claim 6, wherein predicting the future dominant strain of the virus comprises determining which sequence of the plurality of aligned genomic sequences has a smallest q-distance from a current dominant strain that is a member of the plurality of aligned genomic sequences.

8. The method of claim 5, further comprising calculating Qnets for a second plurality of aligned genomic sequences of the virus, the second plurality of aligned genomic sequences having a second common background different than the first common background of the first plurality of aligned genomic sequences.

9. The method of claim 8, further comprising calculating q-distances from genomic sequences of the first plurality of aligned genomic sequences to genomic sequences of the second plurality of aligned genomic sequences.

10. The method of claim 9, wherein the first common background comprises a first species, the second common background comprises a second species, and further comprising calculating a probability of the virus jumping from the first species to the second species based on the calculated q-distances from genomic sequences of the first

plurality of aligned genomic sequences to genomic sequences of the second plurality of aligned genomic sequences.

11. A system comprising:

a processor; and

a memory, the memory storing instructions that, when executed by the processor, cause the processor to:

receive a first plurality of aligned genomic sequences of a virus from a database, the aligned genomic sequences having a first common background; and

calculate a Qnet for each genomic sequence of the first plurality of aligned genomic sequences by:

calculating a conditional inference tree for each index of the aligned genomic sequences using other indices in the aligned genomic sequences as predictive features; and

calculating predictors for indices that were used as predictive features when calculating the conditional inference tree for each index.

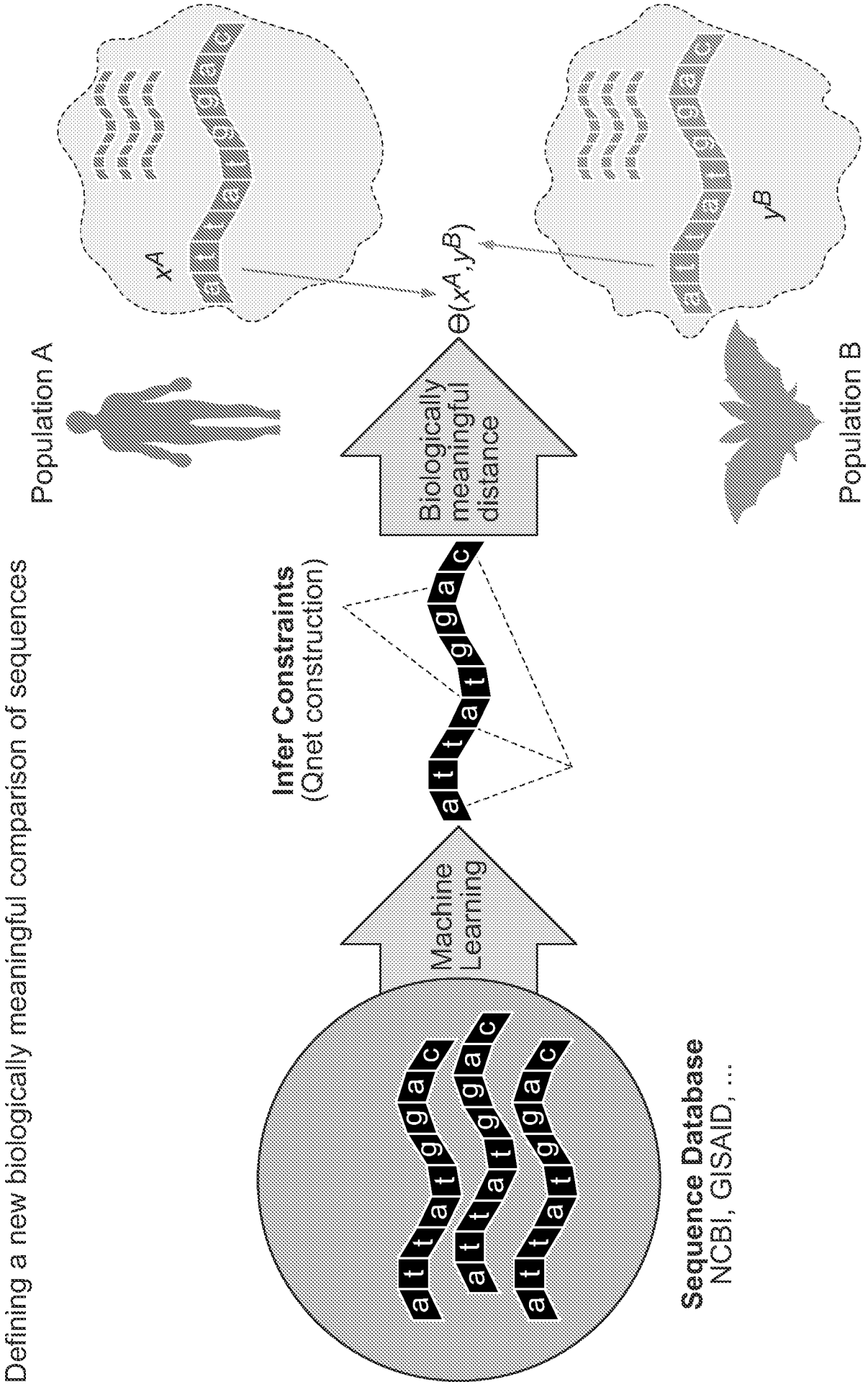
Defining a new biologically meaningful comparison of sequences
Sequence Database
NCBI, GISAID, ...
Machine Learning
Infer Constraints
(Qnet construction)
Biologically meaningful distance
Population A
Population B
x^A
y^B
$\Theta(x^A, y^B)$

FIG. 1A

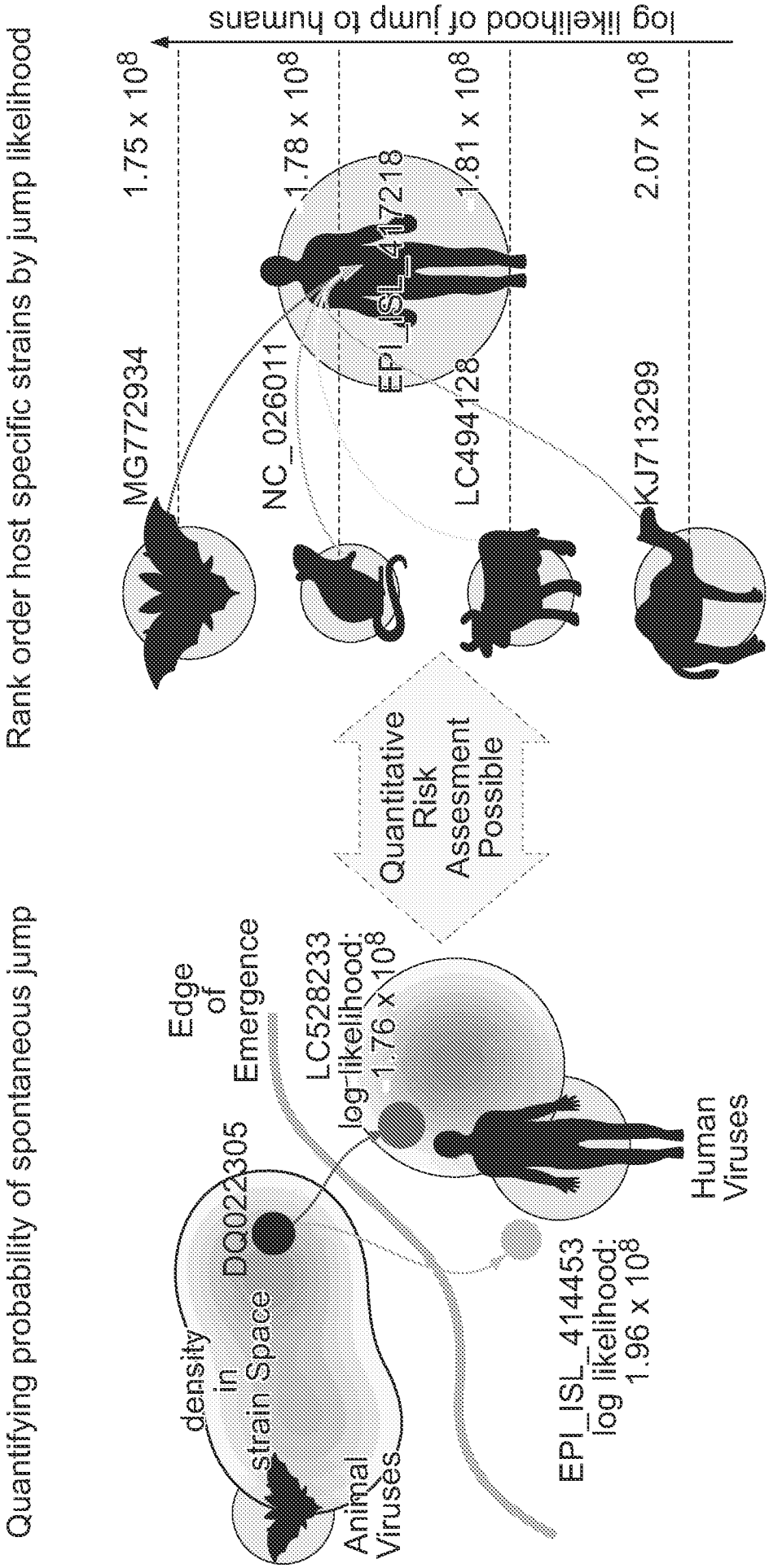
Quantifying probability of spontaneous jump
Rank order host specific strains by jump likelihood
log likelihood of jump to humans
1.75 x 10^8
1.78 x 10^8
1.81 x 10^8
2.07 x 10^8
MG772934
NC_026011
EPI_ISL_417218
LC494128
KJ713299
Quantitative Risk Assesment Possible
Edge of Emergence
LC528233 log-likelihood: 1.76 x 10^8
density in strain Space
DQ022305
Animal Viruses
EPI_ISL_414453 log likelihood: 1.96 x 10^8
Human Viruses

FIG. 1B

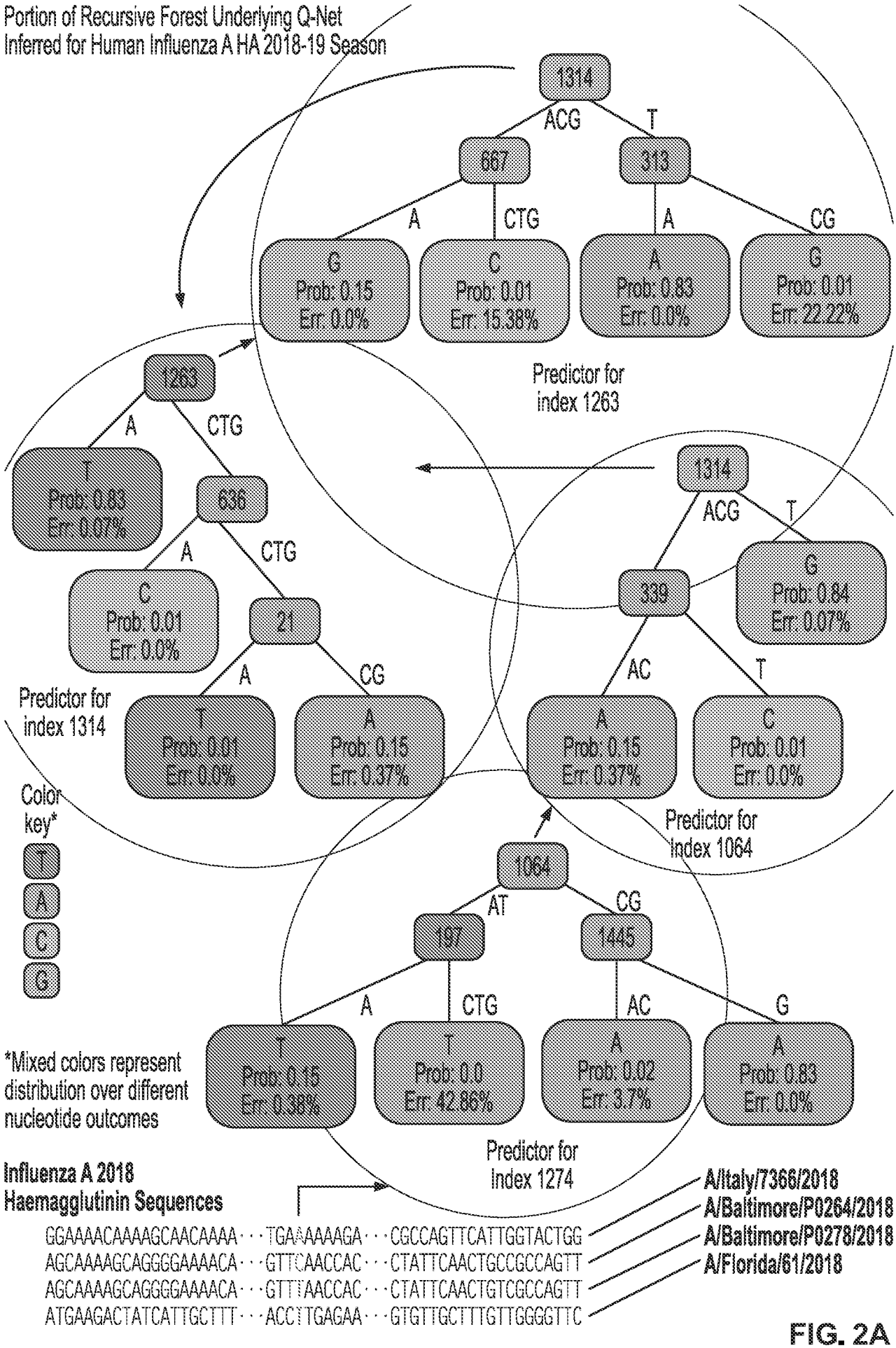
Portion of Recursive Forest Underlying Q-Net
Inferred for Human Influenza A HA 2018-19 Season
1314
ACG
T
667
313
A
CTG
A
CG
G
Prob: 0.15
Err: 0.0%
C
Prob: 0.01
Err: 15.38%
A
Prob: 0.83
Err: 0.0%
G
Prob: 0.01
Err: 22.22%
Predictor for
index 1263
1263
A
CTG
T
Prob: 0.83
Err: 0.07%
636
A
CTG
C
Prob: 0.01
Err: 0.0%
21
A
CG
T
Prob: 0.01
Err: 0.0%
A
Prob: 0.15
Err: 0.37%
Predictor for
index 1314
1314
ACG
T
339
G
Prob: 0.84
Err: 0.07%
AC
T
A
Prob: 0.15
Err: 0.37%
C
Prob: 0.01
Err: 0.0%
Predictor for
Index 1064
Color
key*
T
A
C
G
1064
AT
CG
197
1445
A
CTG
AC
G
T
Prob: 0.15
Err: 0.36%
T
Prob: 0.0
Err: 42.86%
A
Prob: 0.02
Err: 3.7%
A
Prob: 0.83
Err: 0.0%
*Mixed colors represent
distribution over different
nucleotide outcomes
Predictor for
Index 1274
Influenza A 2018
Haemagglutinin Sequences
GGAAAACAAAAGCAACAAAA···TGAAAAAAGA···CGCCAGTTCATTGGTACTGG
AGCAAAAGCAGGGGAAAACA···GTTCAACCAC···CTATTCAACTGCCGCCAGTT
AGCAAAAGCAGGGGAAAACA···GTTTAACCAC···CTATTCAACTGTCGCCAGTT
ATGAAGACTATCATTGCTTT···ACCTTGAGAA···GTGTTGCTTTGTTGGGGTTC
A/Italy/7366/2018
A/Baltimore/P0264/2018
A/Baltimore/P0278/2018
A/Florida/61/2018

FIG. 2A

SARS CoV 2 Spike Jan Mar 20†
(COVID 19 Pandemic 2019)

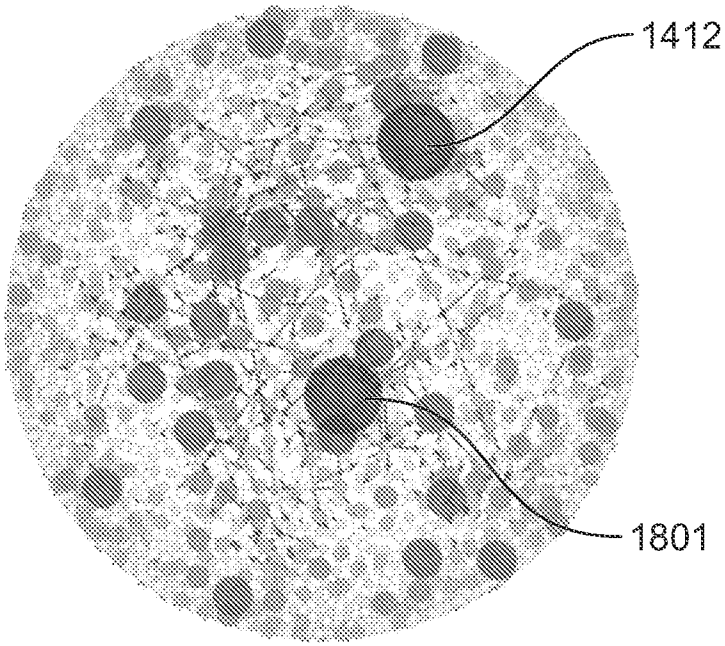

FIG. 2B

Human Influenza A HA 2008 9†
(Coinciding Swine Flu Pandemic 2009)

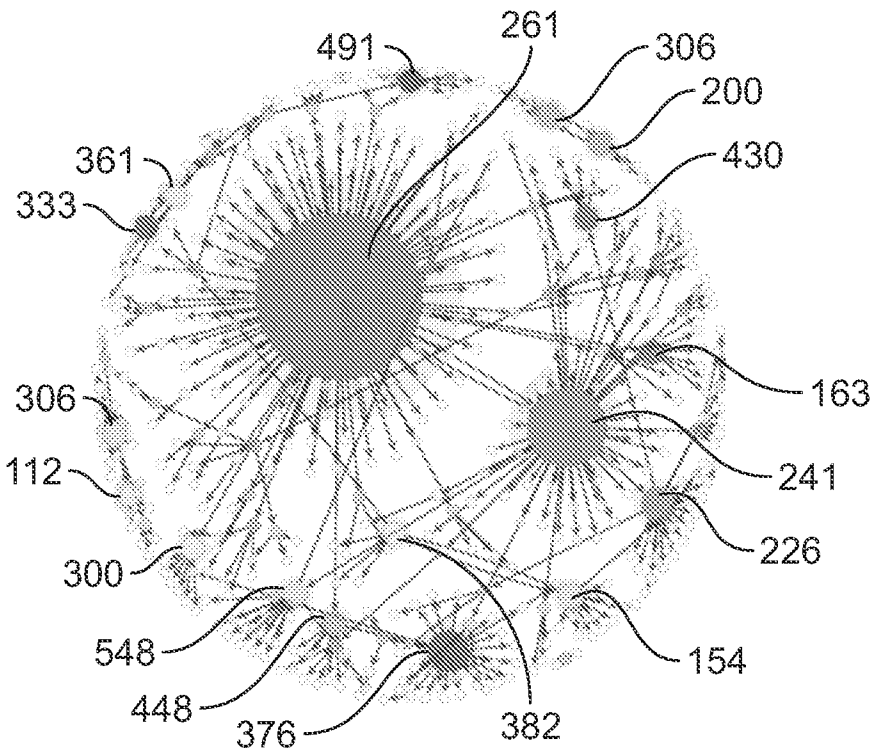

FIG. 2C

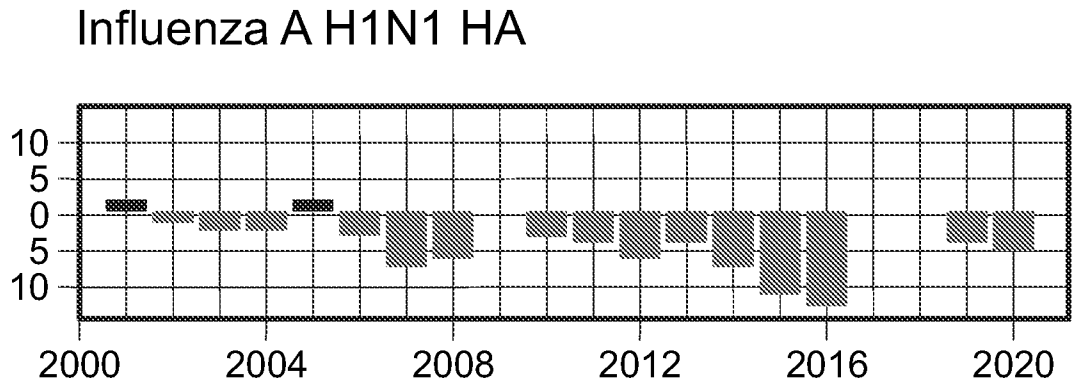

FIG. 3A

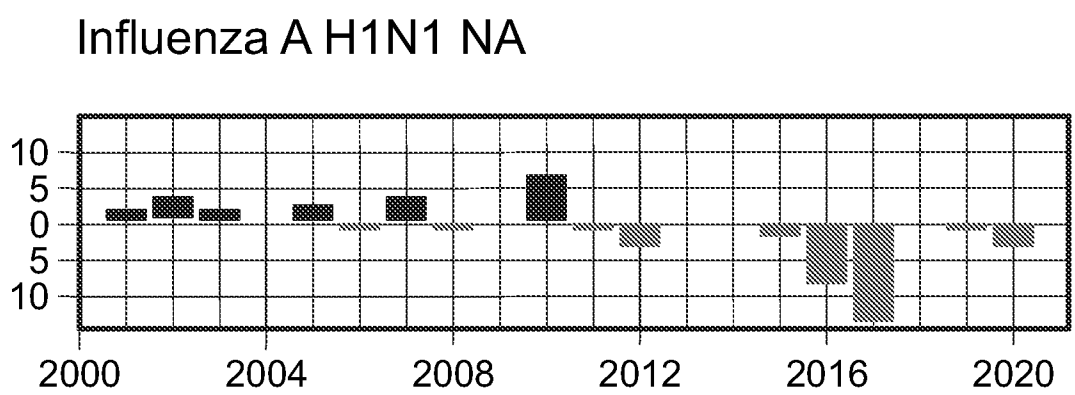

FIG. 3B

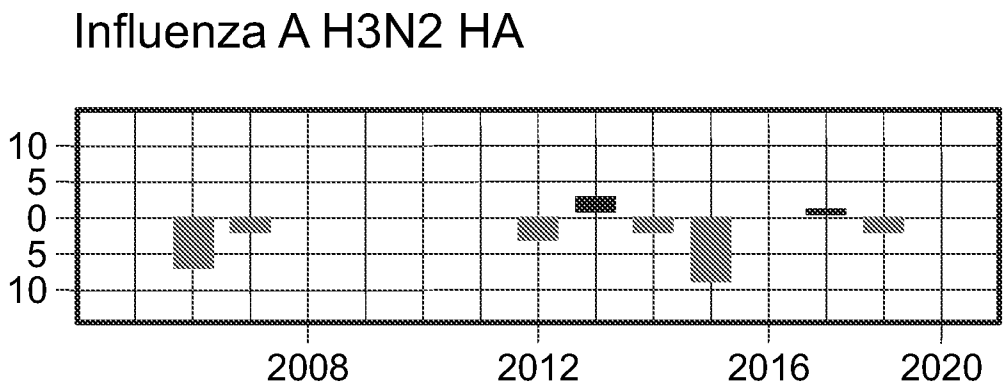

FIG. 3C

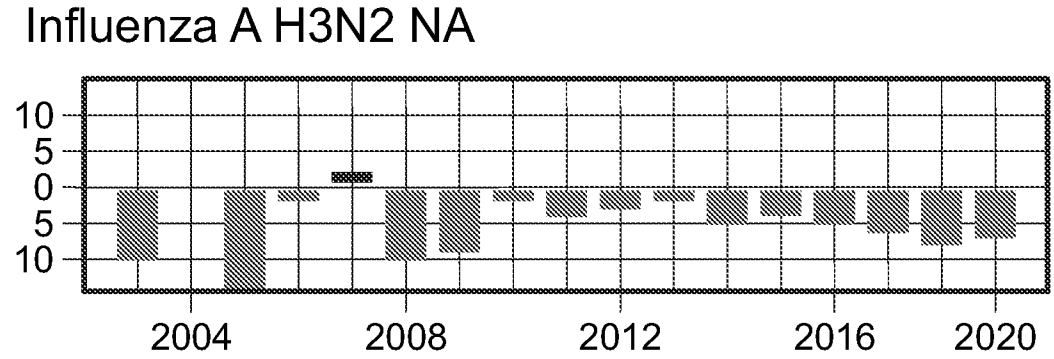

FIG. 3D

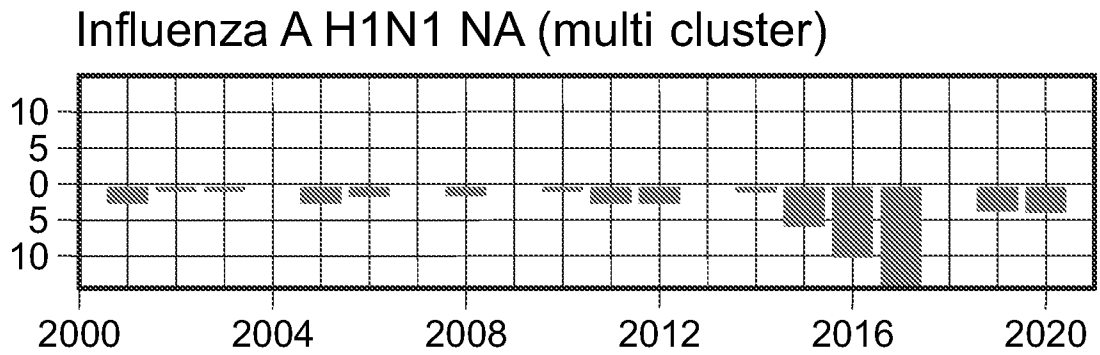

FIG. 3E

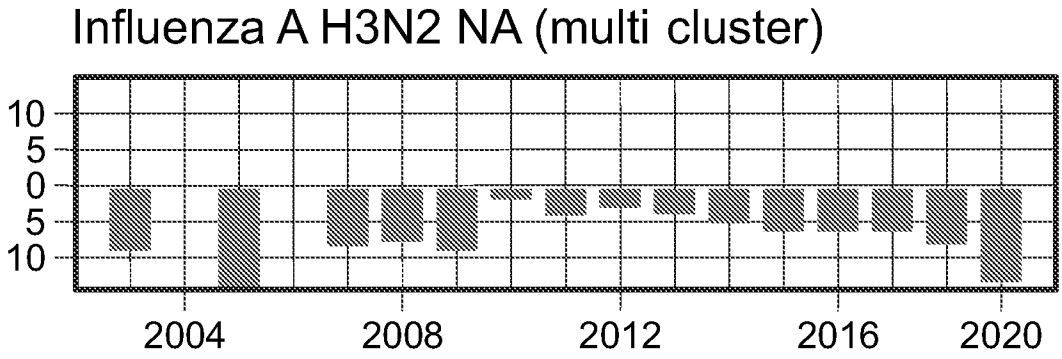

FIG. 3F

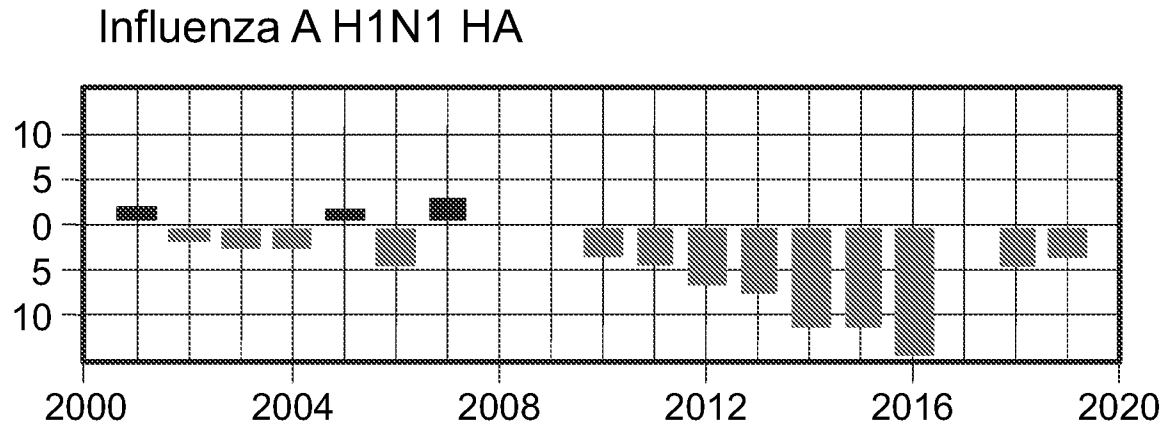

FIG. 3G

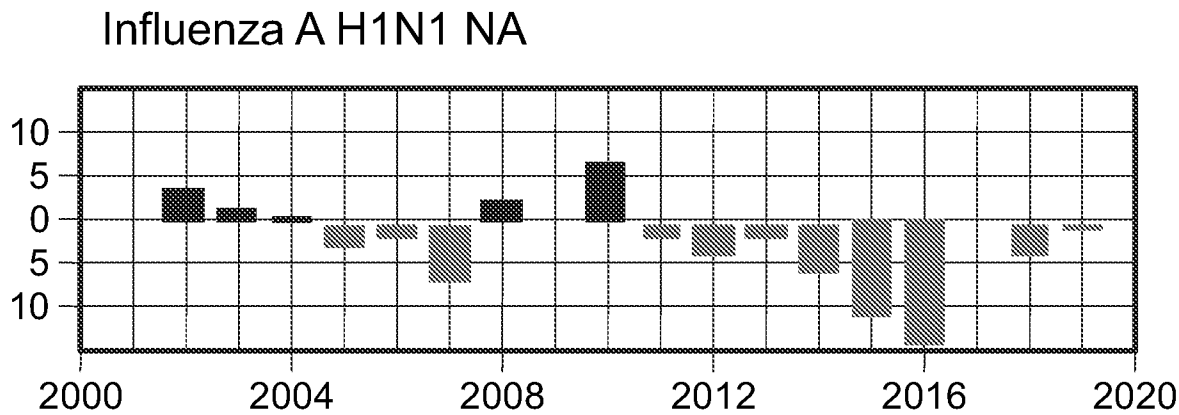

FIG. 3H

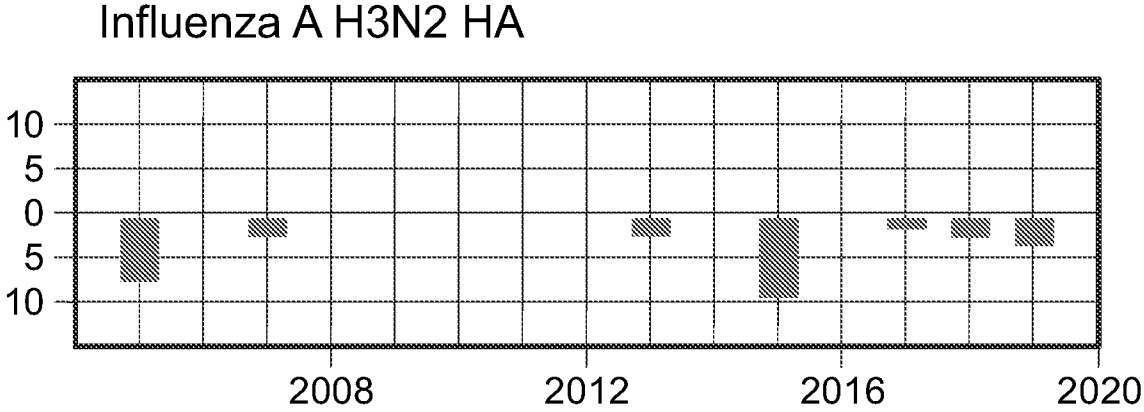

FIG. 3I

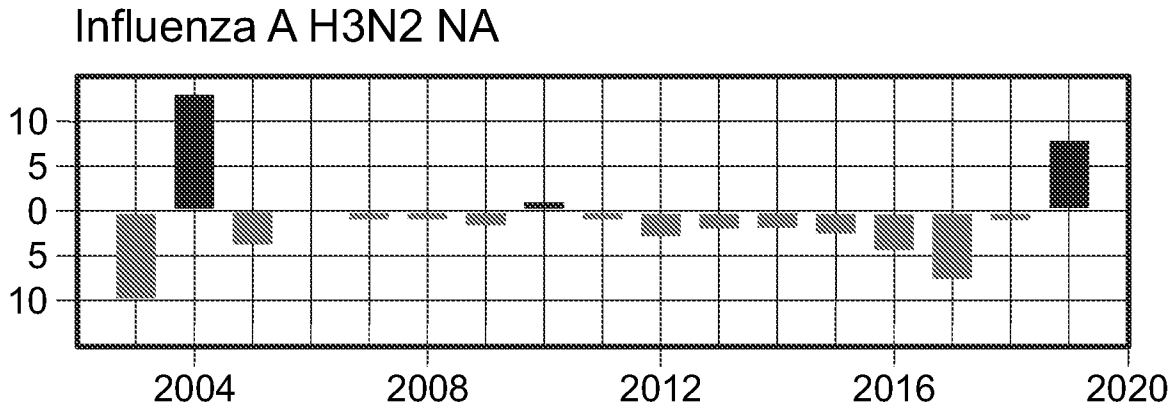

FIG. 3J

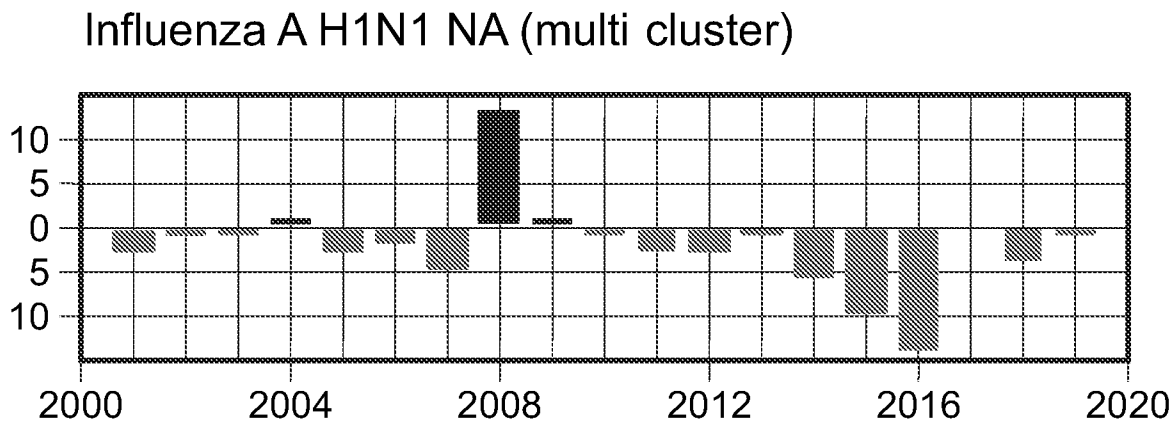

FIG. 3K

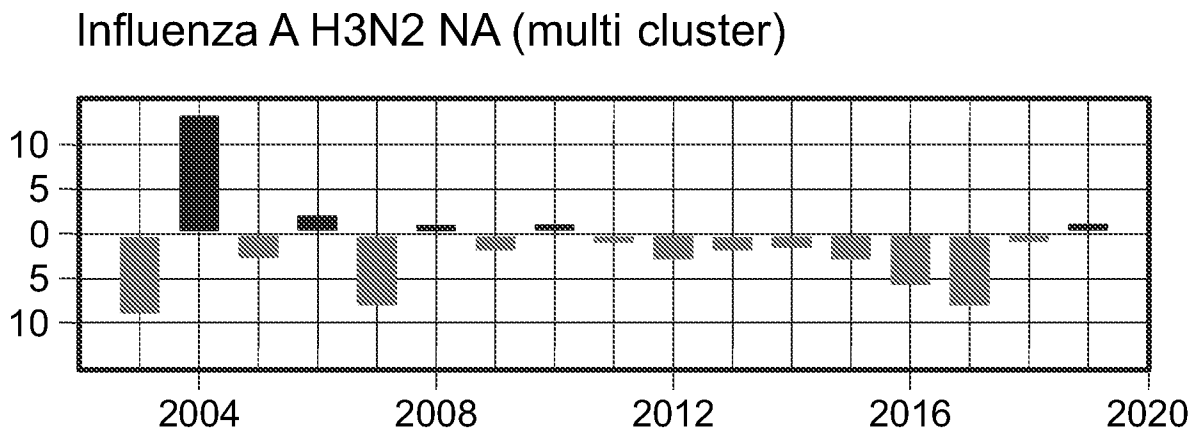

FIG. 3L

9/34

2018-2019 (H1N1 HA Northern Hemisphere)

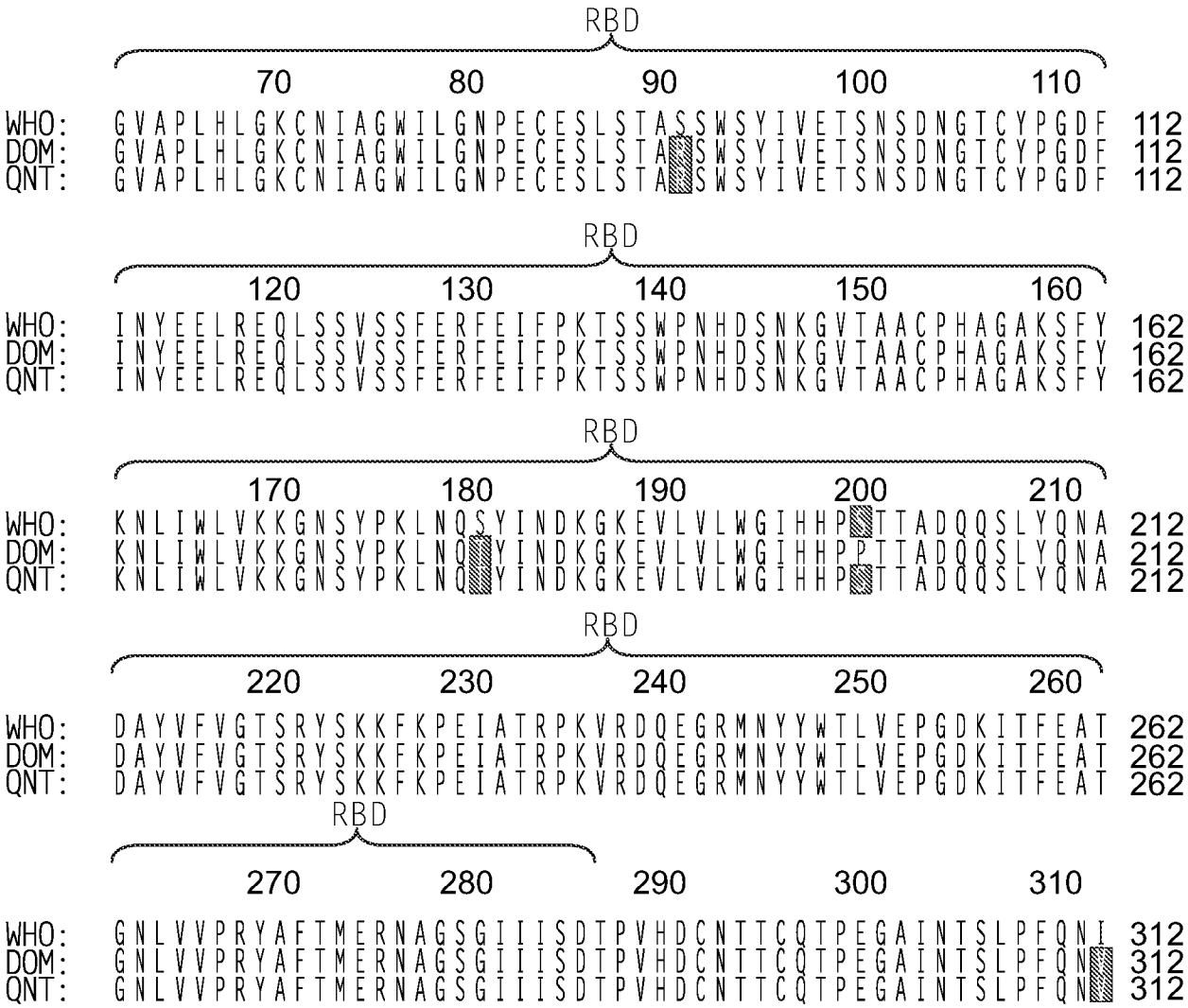

FIG. 4A

2019-2020 (H1N1 HA Northern Hemisphere)

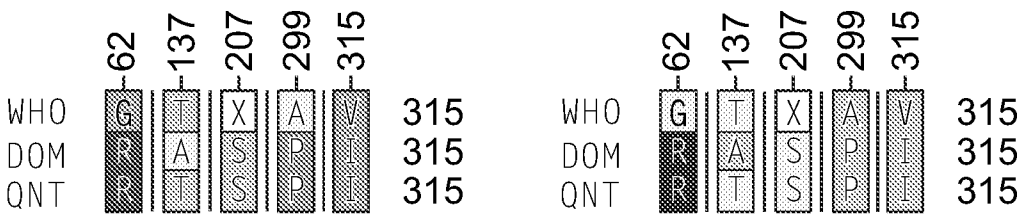

FIG. 4B

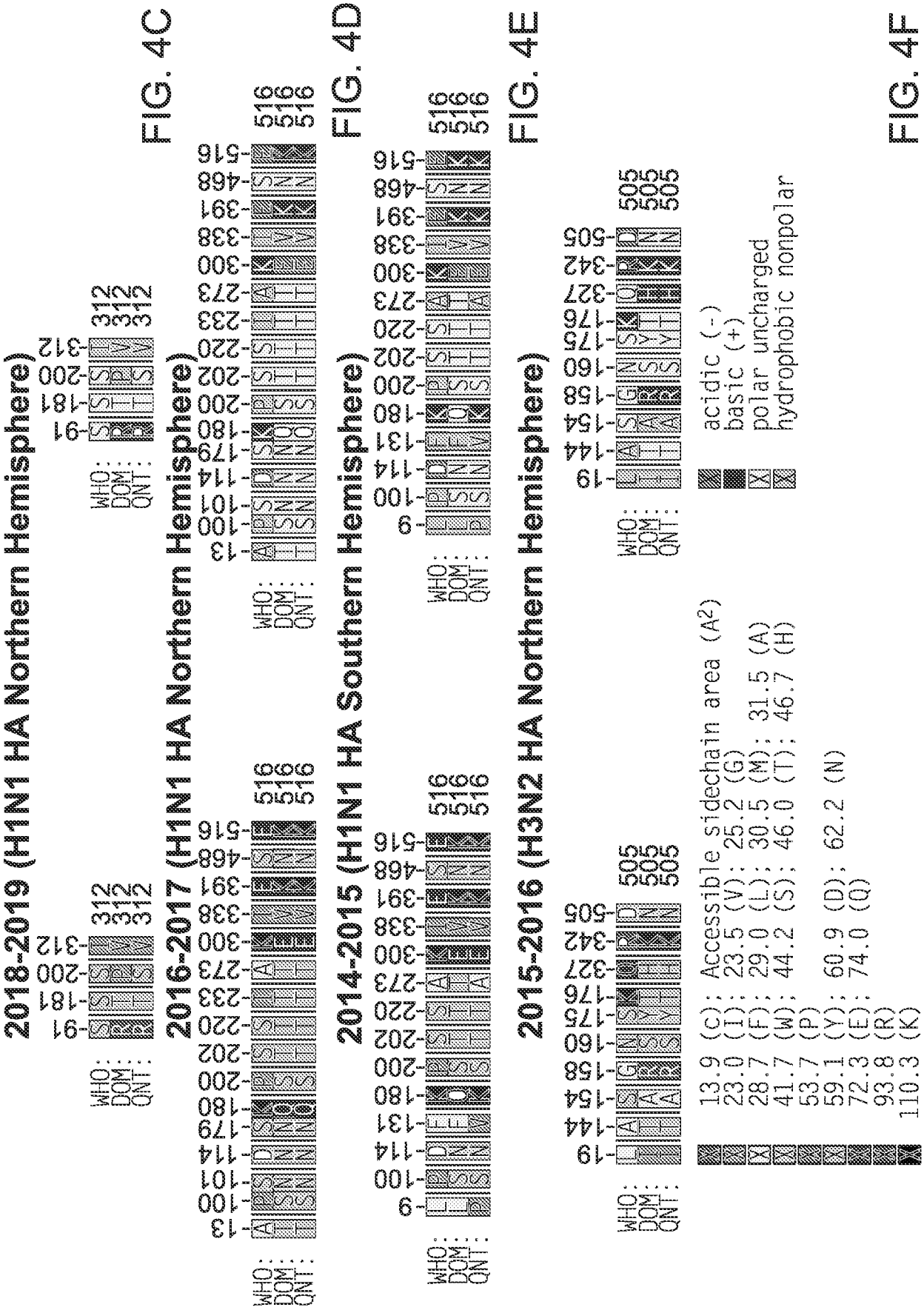

FIG. 4C FIG. 4D FIG. 4E FIG. 4F

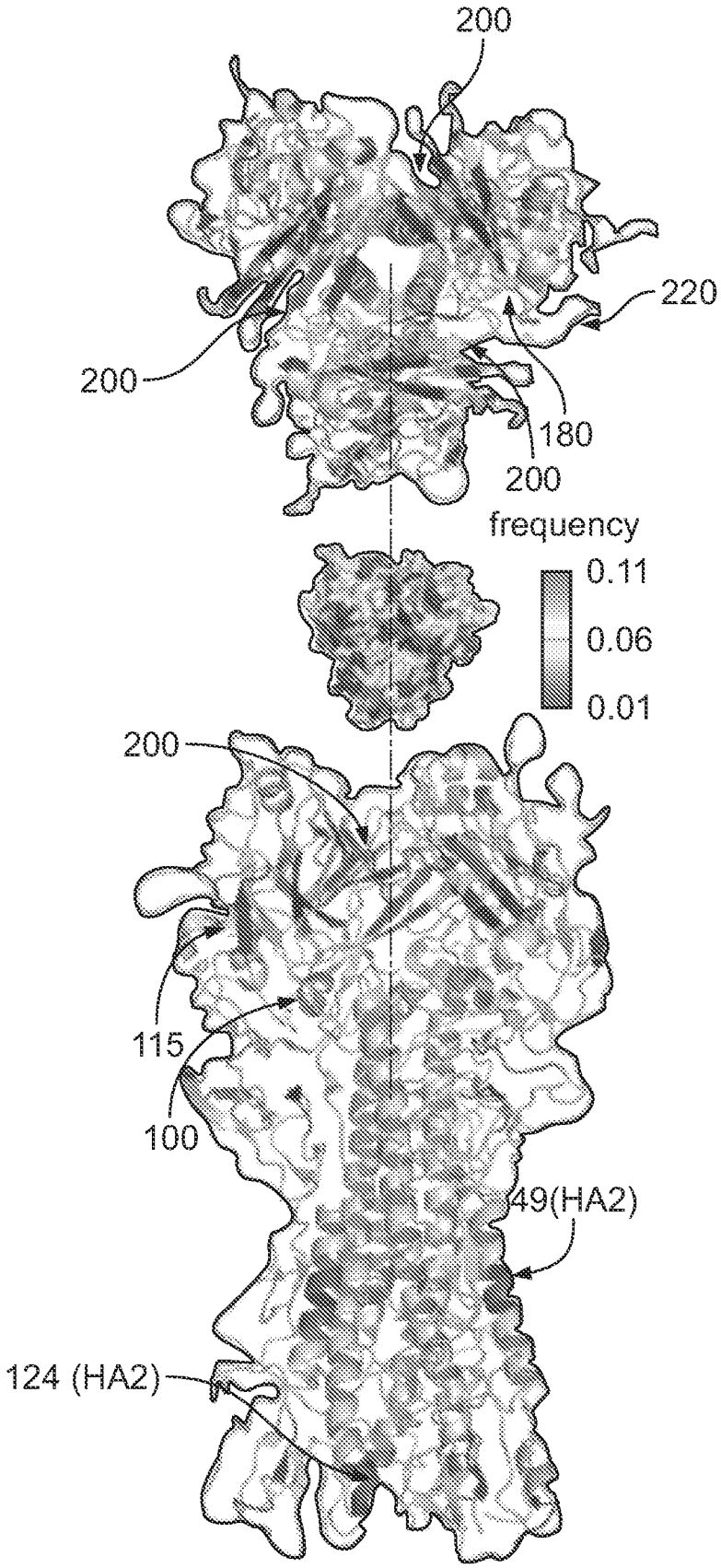
200
220
200
180
200
frequency
0.11
0.06
0.01
200
115
100
49(HA2)
124 (HA2)

FIG. 4G

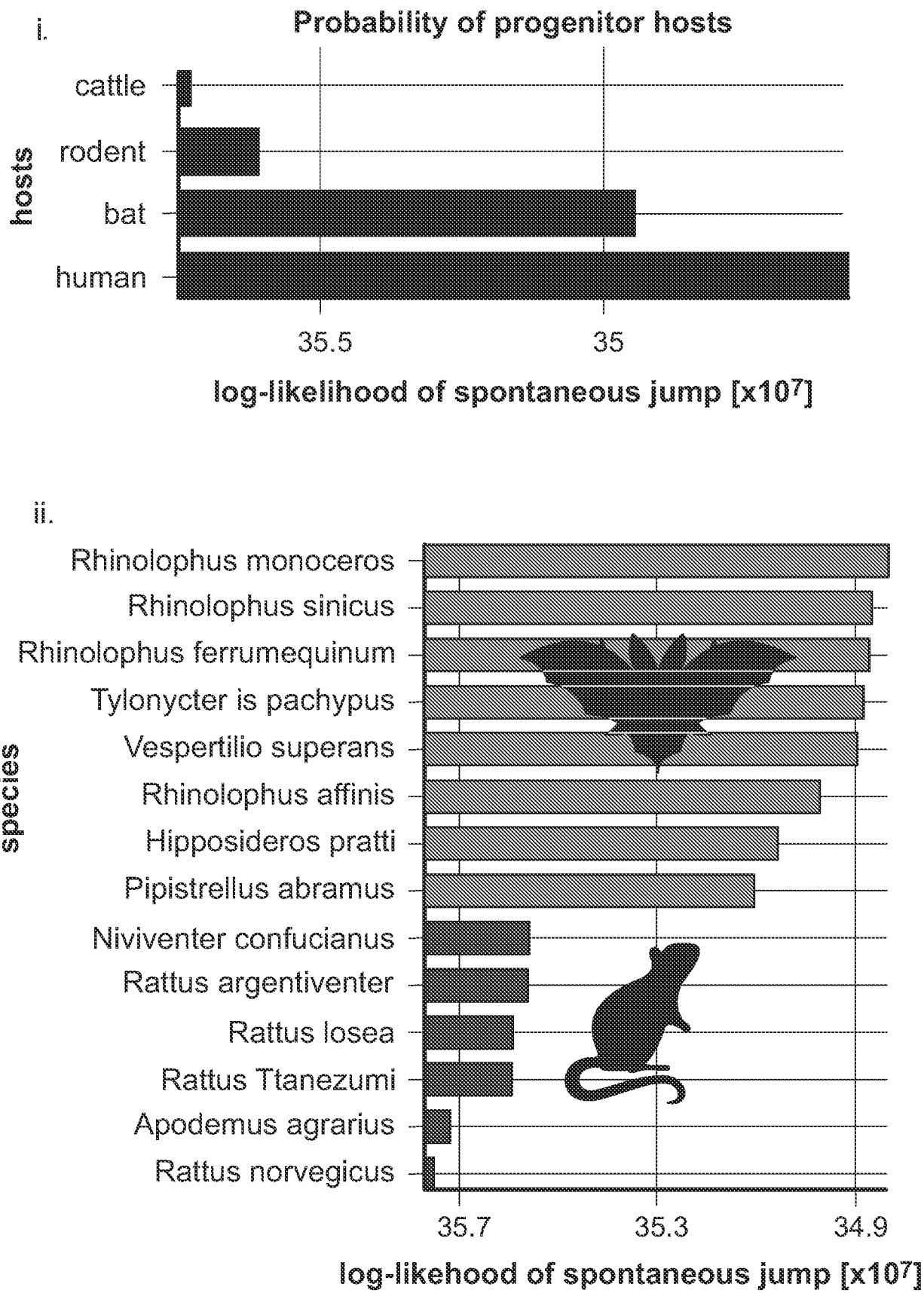
i.
Probability of progenitor hosts
hosts
cattle
rodent
bat
human
35.5
35
log-likelihood of spontaneous jump [x10^7]
ii.
species
Rhinolophus monoceros
Rhinolophus sinicus
Rhinolophus ferrumequinum
Tylonycter is pachypus
Vespertilio superans
Rhinolophus affinis
Hipposideros pratti
Pipistrellus abramus
Niviventer confucianus
Rattus argentiventer
Rattus losea
Rattus Ttanezumi
Apodemus agrarius
Rattus norvegicus
35.7
35.3
34.9
log-likehood of spontaneous jump [x10^7]

FIG. 5A

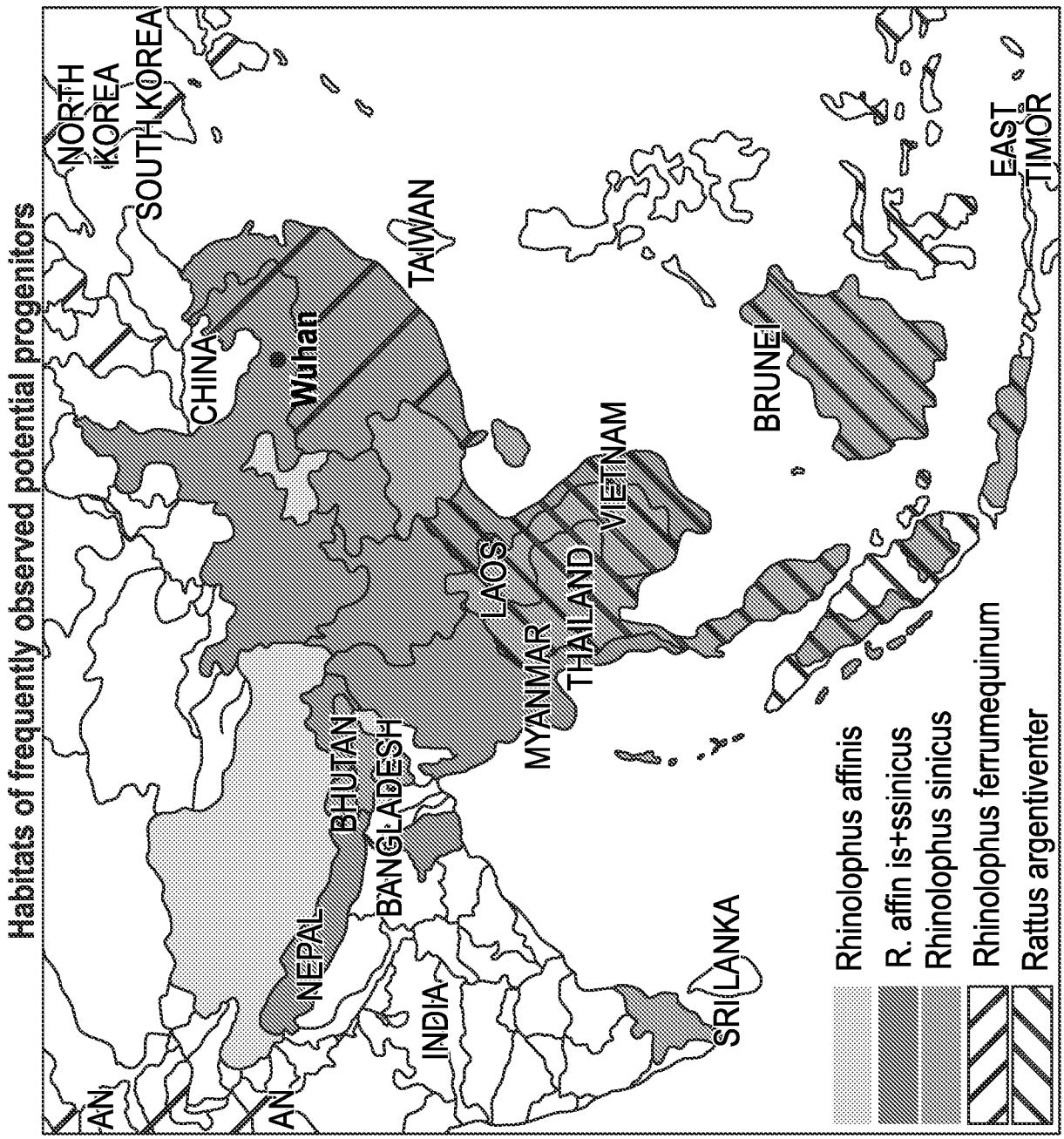
Habitats of frequently observed potential progenitors
NORTH KOREA
SOUTH KOREA
CHINA
Wuhan
TAIWAN
BRUNEI
EAST TIMOR
LAOS
MYANMAR
THAILAND
VIETNAM
BHUTAN
BANGLADESH
NEPAL
INDIA
SRI LANKA
AN
AN
Rhinolophus affinis
R. affin is+ssinicus
Rhinolophus sinicus
Rhinolophus ferrumequinum
Rattus argentiventer

FIG. 5B

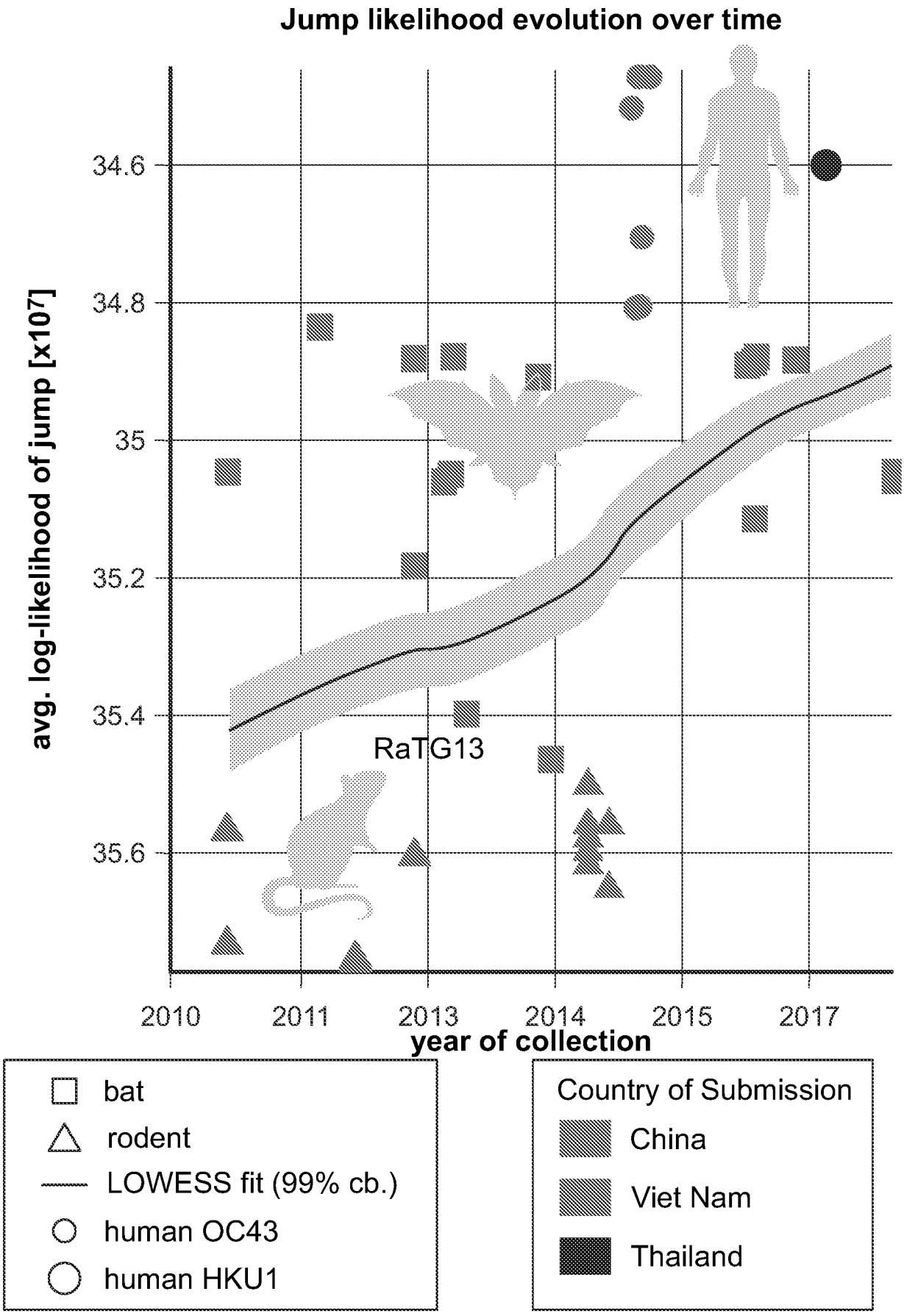
Jump likelihood evolution over time
avg. log-likelihood of jump [x107]
34.6
34.8
35
35.2
35.4
35.6
RaTG13
2010
2011
2013
2014
2015
2017
year of collection
bat
rodent
LOWESS fit (99% cb.)
human OC43
human HKU1
Country of Submission
China
Viet Nam
Thailand

FIG. 5C

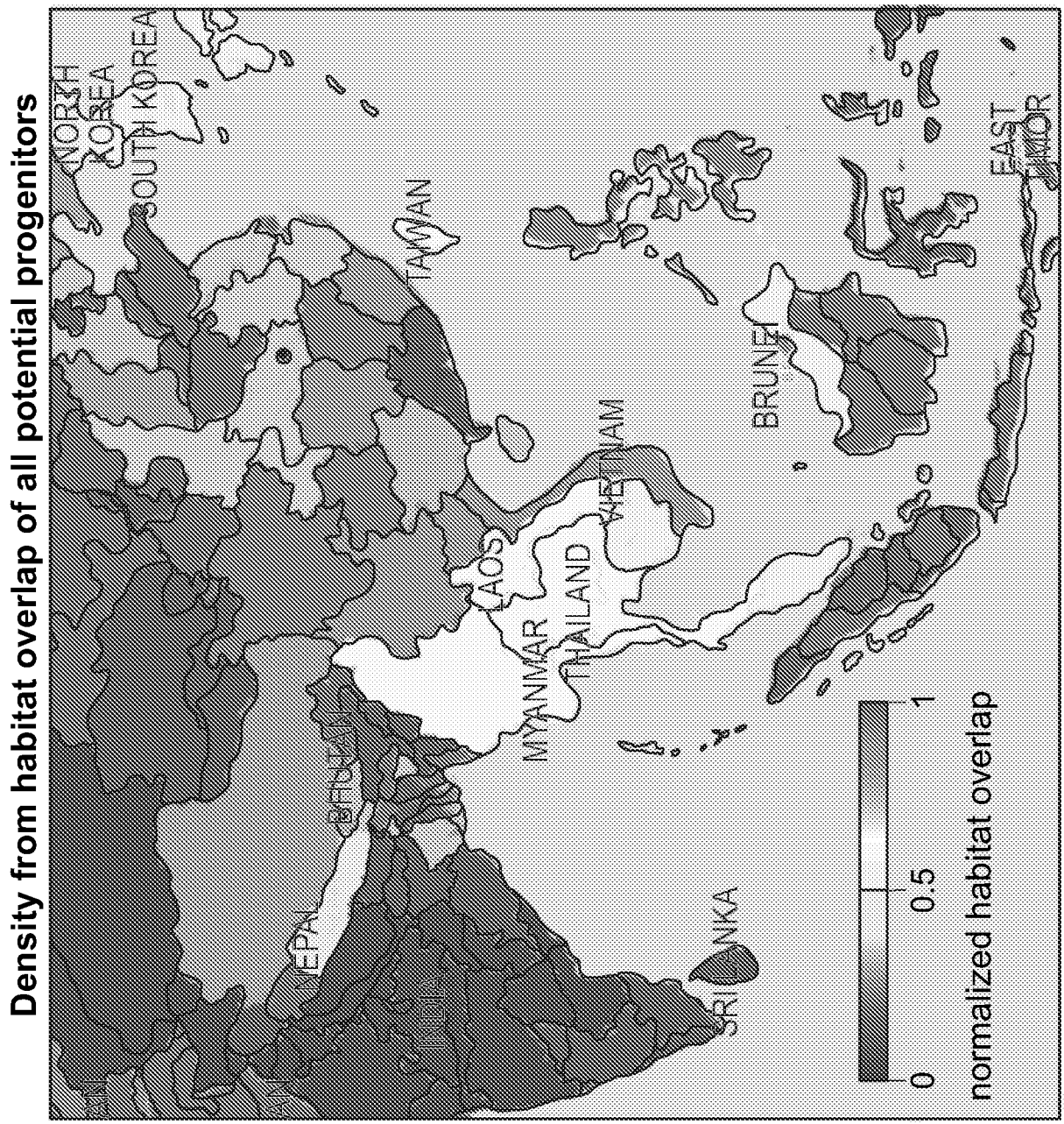
Density from habitat overlap of all potential progenitors
NORTH KOREA
SOUTH KOREA
TAIWAN
EAST TIMOR
BRUNEI
VIETNAM
LAOS
MYANMAR
THAILAND
NEPAL
SRI LANKA
0
0.5
1
normalized habitat overlap

FIG. 5D

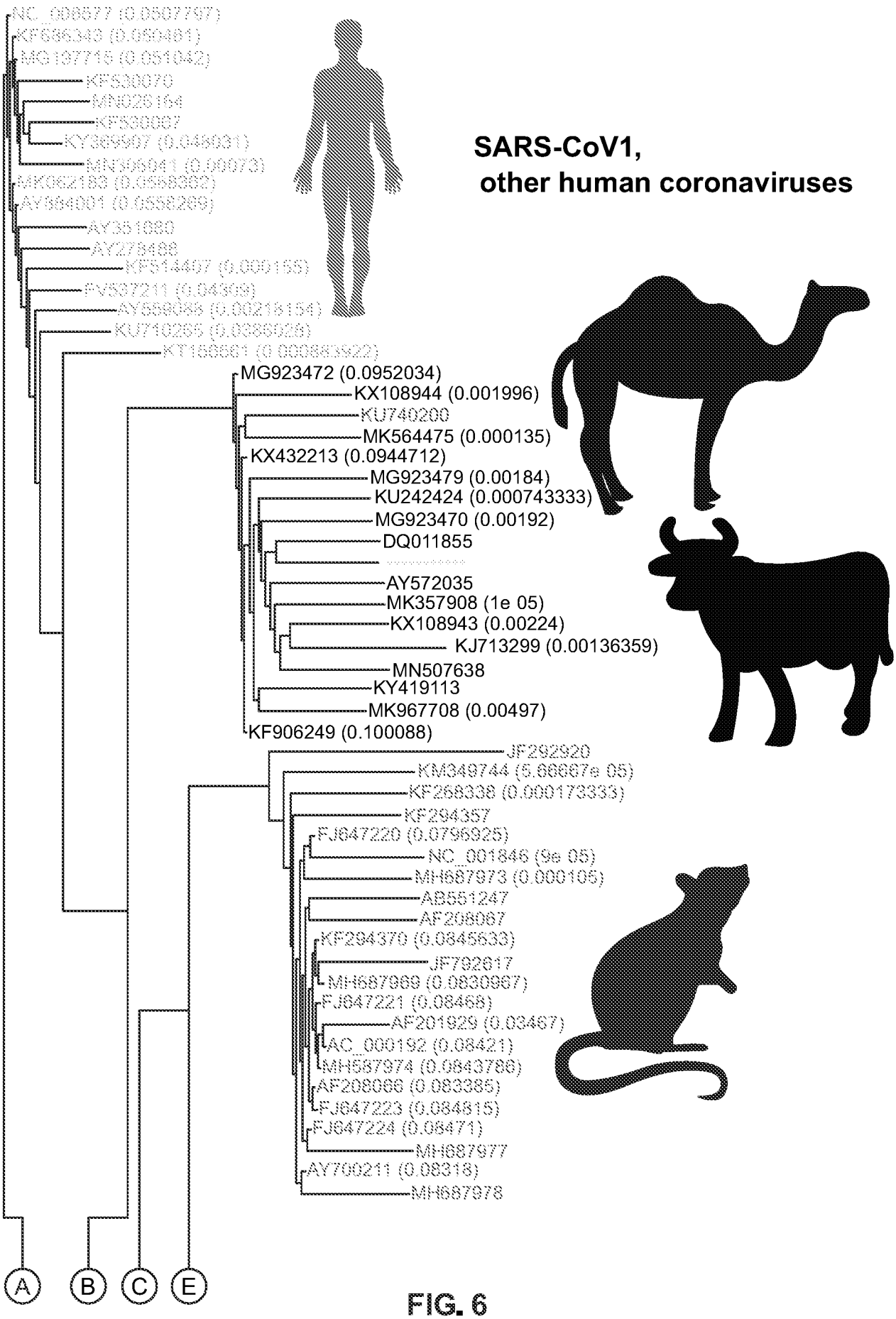
SARS-CoV1,
other human coronaviruses
MG923472 (0.0952034)
KX108944 (0.001996)
KU740200
MK564475 (0.000135)
KX432213 (0.0944712)
MG923479 (0.00184)
KU242424 (0.000743333)
MG923470 (0.00192)
DQ011855
AY572035
MK357908 (1e 05)
KX108943 (0.00224)
KJ713299 (0.00136359)
MN507638
KY419113
MK967708 (0.00497)
KF906249 (0.100088)
JF292920
KM349744 (5.66667e 05)
KF268338 (0.000173333)
KF294357
FJ647220 (0.0796925)
NC_001846 (9e 05)
MH687973 (0.000105)
AB551247
AF208067
KF294370 (0.0845633)
JF792617
MH687969 (0.0830967)
FJ647221 (0.08468)
AF201929 (0.03467)
AC_000192 (0.08421)
MH687974 (0.0843786)
AF208066 (0.083385)
FJ647223 (0.084815)
FJ647224 (0.08471)
MH687977
AY700211 (0.08318)
MH687978
A
B
C
E

FIG. 6

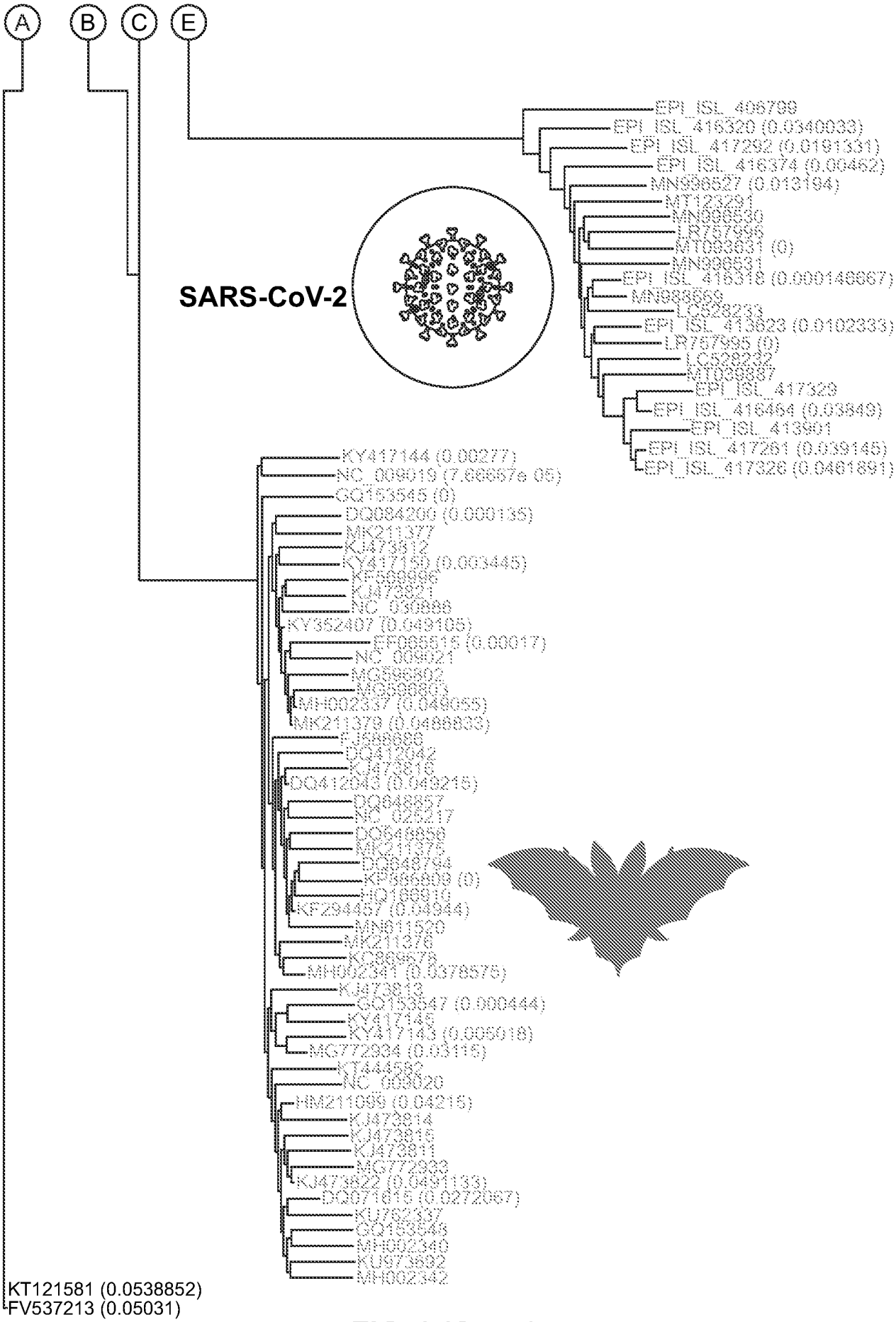
A
B
C
E
SARS-CoV-2
EPI_ISL_406799
EPI_ISL_416320 (0.0340033)
EPI_ISL_417292 (0.0191331)
EPI_ISL_416374 (0.00462)
MN996527 (0.013194)
MT123291
MN996530
LR757996
MT093631 (0)
MN996531
EPI_ISL_416318 (0.000146667)
MN988669
LC528233
EPI_ISL_413523 (0.0102333)
LR757995 (0)
LC528232
MT039887
EPI_ISL_417329
EPI_ISL_416464 (0.03849)
EPI_ISL_413901
EPI_ISL_417261 (0.039145)
EPI_ISL_417326 (0.0461891)
KY417144 (0.00277)
NC_009019 (7.66657e-05)
GQ153545 (0)
DQ084200 (0.000135)
MK211377
KJ473812
KY417150 (0.003445)
KF569996
KJ473821
NC_030886
KY352407 (0.049105)
EF065515 (0.00017)
NC_009021
MG596802
MG596801
MH002337 (0.040055)
MK211379 (0.0488833)
FJ588686
DQ412042
KJ473816
DQ412043 (0.043215)
DQ648857
NC_028217
DQ648858
MK211376
DQ648784
KF569609 (0)
DQ189910
KF294457 (0.04944)
MN611520
MK211378
KC869678
MH002341 (0.0378575)
KJ473813
GQ153547 (0.000444)
KY417145
KY417143 (0.005018)
MG772934 (0.03115)
KT444582
NC_009020
HM211099 (0.04215)
KJ473814
KJ473815
KJ473811
MG772933
KJ473822 (0.0491133)
DQ071615 (0.0272087)
KU762337
GQ153544
MH002340
KU973692
MH002342
KT121581 (0.0538852)
FV537213 (0.05031)

FIG. 6 (Cont.)

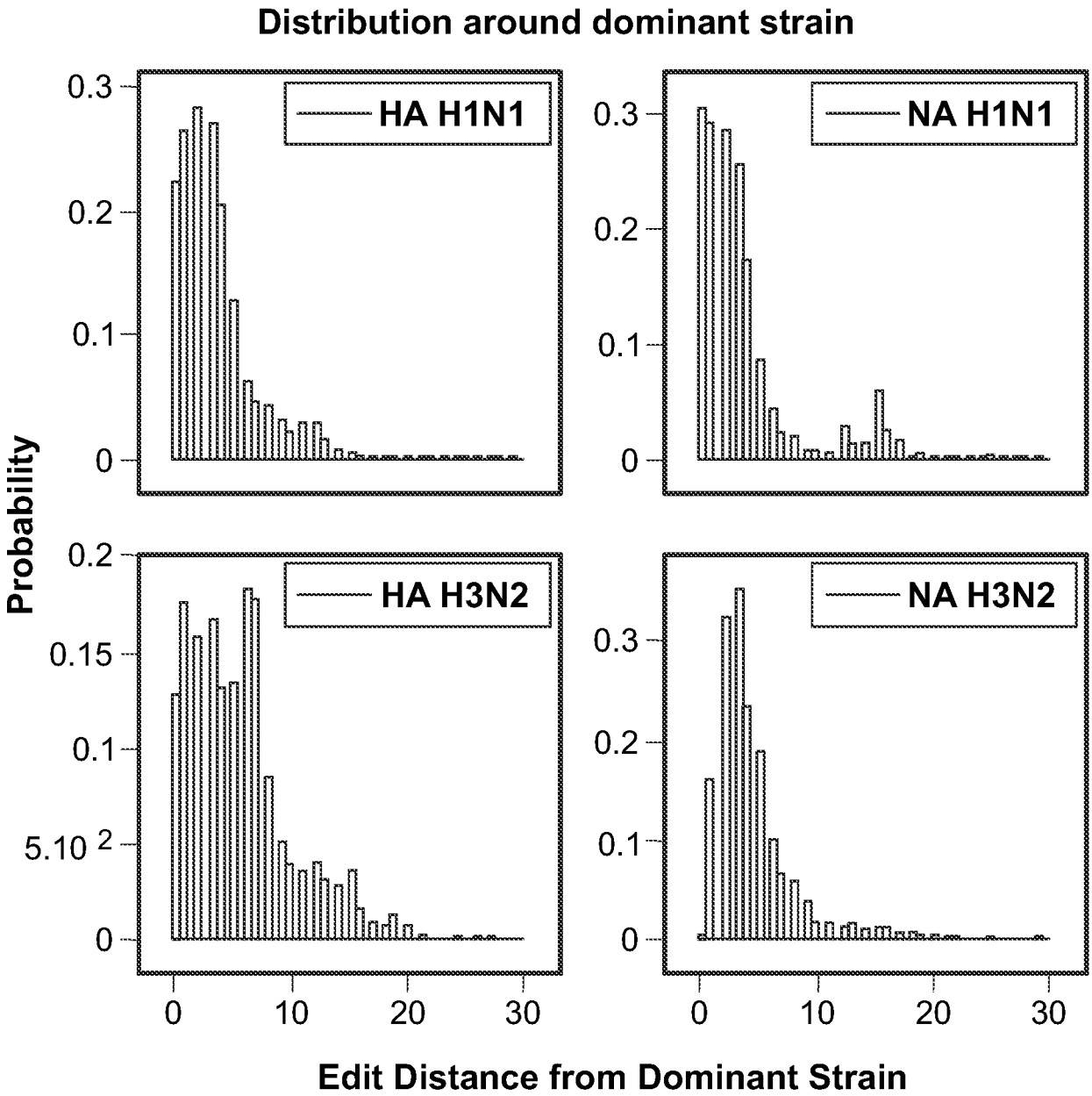
Distribution around dominant strain
HA H1N1
NA H1N1
HA H3N2
NA H3N2
Probability
0.3
0.2
0.1
0
0.15
5.10 2
0
10
20
30
Edit Distance from Dominant Strain

FIG. 7

Phylogenetic tree using q distance

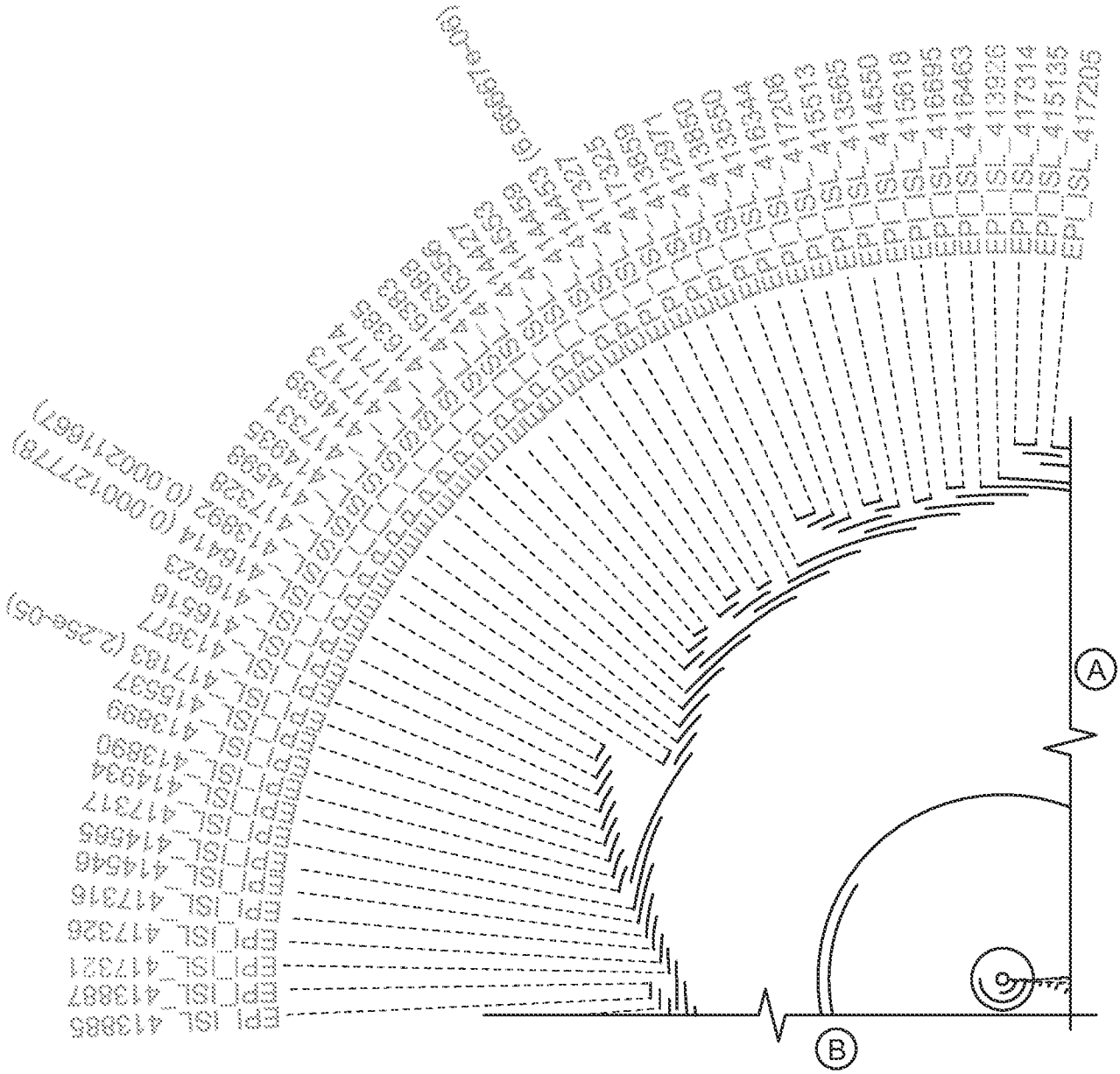

FIG. 8A

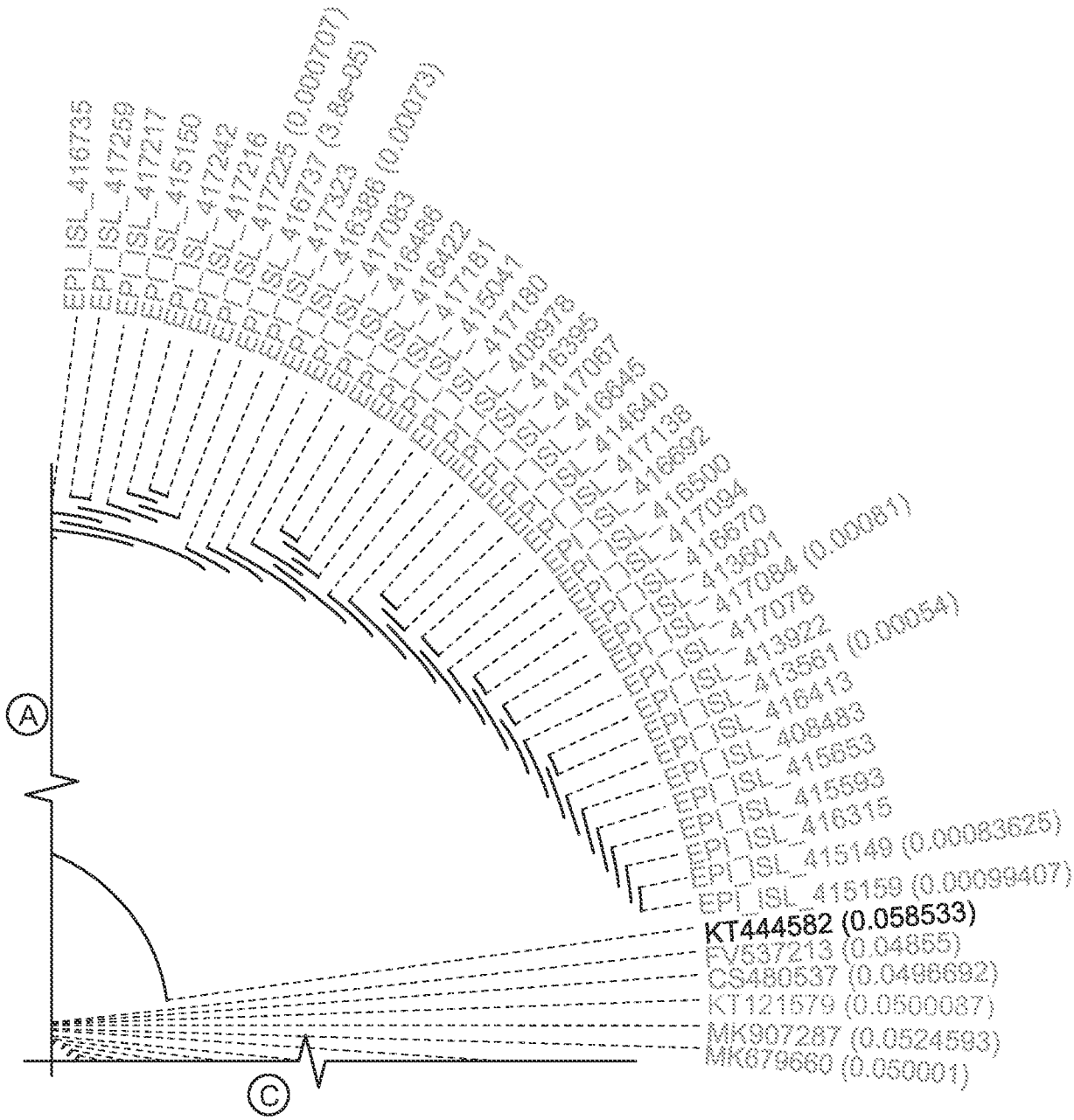
EPI_ISL_416735
EPI_ISL_417269
EPI_ISL_417217
EPI_ISL_415160
EPI_ISL_417242
EPI_ISL_417216
EPI_ISL_417225 (0.000707)
EPI_ISL_416707 (3.8e-05)
EPI_ISL_417323
EPI_ISL_416386 (0.00073)
EPI_ISL_417083
EPI_ISL_416458
EPI_ISL_416422
EPI_ISL_417161
EPI_ISL_416441
EPI_ISL_417160
EPI_ISL_408678
EPI_ISL_416396
EPI_ISL_417067
EPI_ISL_416645
EPI_ISL_414640
EPI_ISL_417138
EPI_ISL_416692
EPI_ISL_416500
EPI_ISL_417094
EPI_ISL_416670
EPI_ISL_413601
EPI_ISL_417084 (0.00081)
EPI_ISL_417078
EPI_ISL_413922
EPI_ISL_413561 (0.00054)
EPI_ISL_416413
EPI_ISL_408483
EPI_ISL_415553
EPI_ISL_415593
EPI_ISL_416315
EPI_ISL_415149 (0.00083625)
EPI_ISL_415159 (0.00099407)
KT444582 (0.058533)
FJ537213 (0.04855)
CS480537 (0.0496692)
KT121579 (0.0500087)
MK907287 (0.0524593)
MK679660 (0.050001)
A
C

FIG. 8A (Cont.)

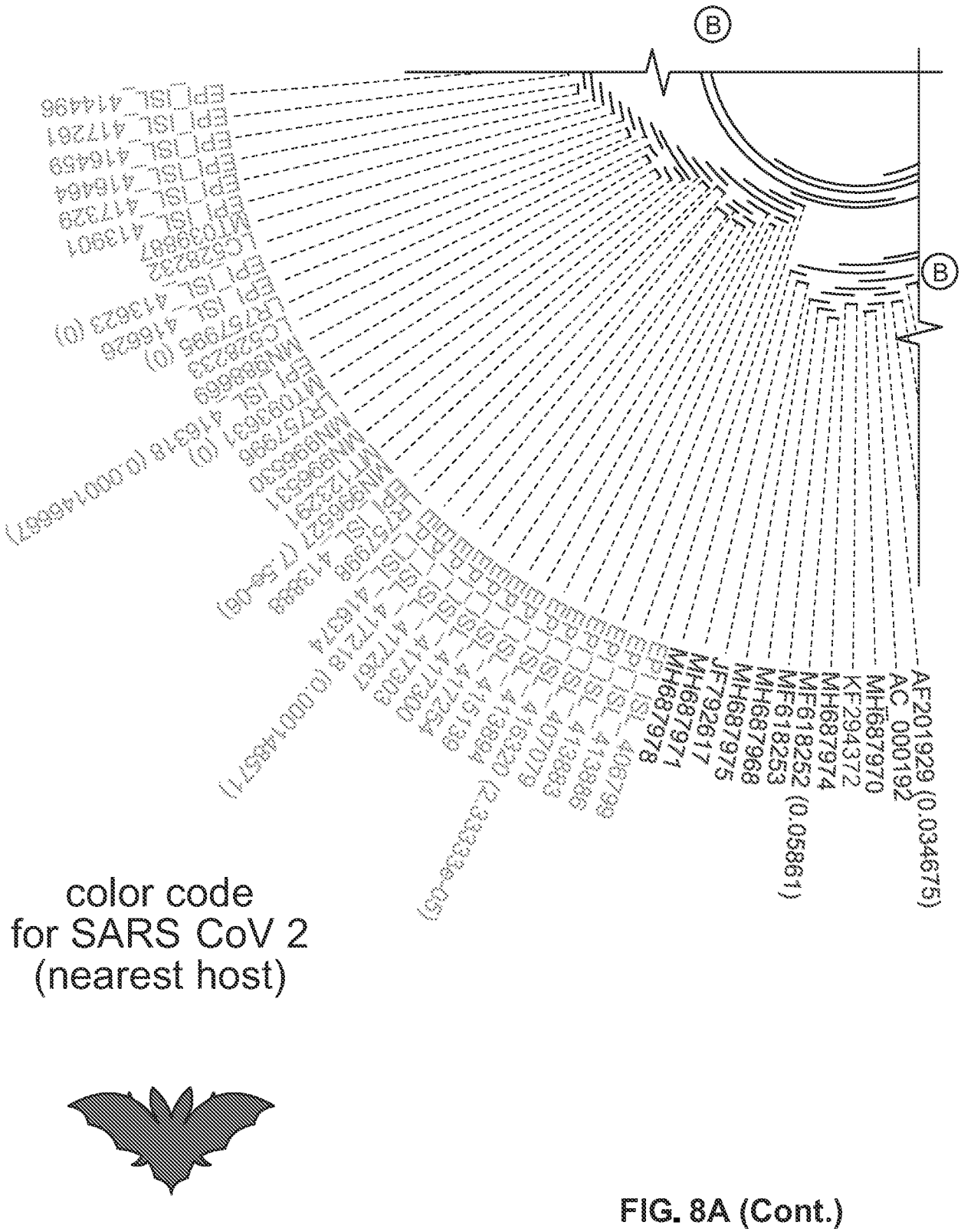
B
B
AF2019629 (0.034675)
AC 000192
MH6887970
KF2943372
MH6887974
MF6182252 (0.058691)
MF6182253
MH6887968
JF7926975
MH6887977
MH6887978
color code
for SARS CoV 2
(nearest host)

FIG. 8A (Cont.)

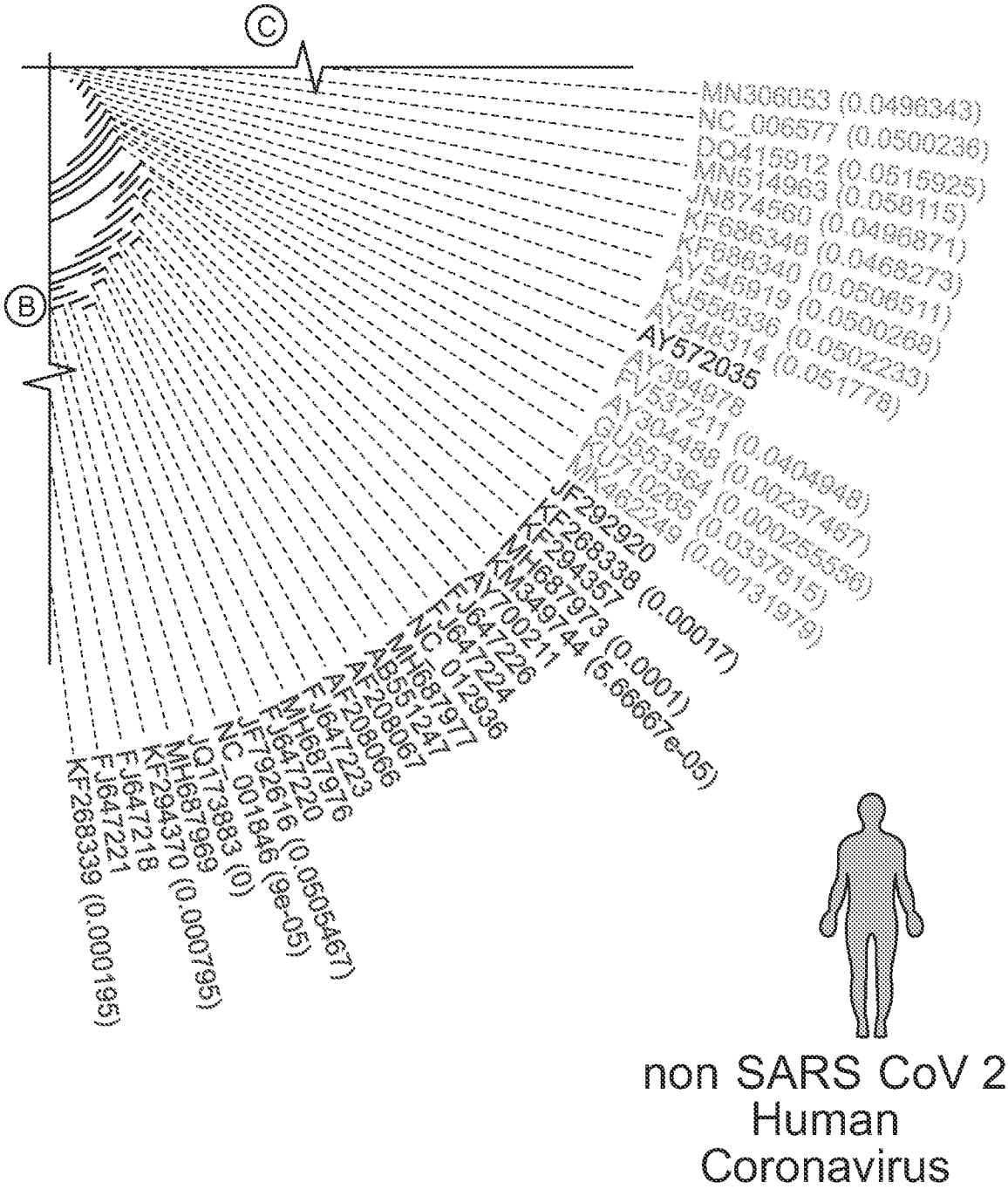
C
B
MN306053 (0.0496343)
NC_006577 (0.0500236)
DQ415912 (0.0515925)
MN514963 (0.058115)
JN874560 (0.0496871)
KF686346 (0.0468273)
KF686340 (0.0506511)
AY545919 (0.0500268)
KJ546336 (0.0502233)
AY345314 (0.051778)
AY572035
AY394978
AY537211 (0.0404948)
AY304488 (0.00237467)
GU553364 (0.000255556)
KU710295 (0.0337915)
MK462249 (0.00131979)
JF292920
KF268338 (0.00017)
KF294357
MH687973 (0.0001)
KM349744 (5.66667e-05)
AY700211
FJ647224
FJ647226
NC_012936
MH687977
AB551247
AF208061
AF208066
FJ647223
MH687976
FJ647220
JF792616 (0.0505467)
NC_001846 (9e-05)
JQ173883 (0)
MH687969
KF294370 (0.000795)
FJ647218
FJ647221
KF268339 (0.000195)
non SARS CoV 2
Human
Coronavirus

FIG. 8A (Cont.)

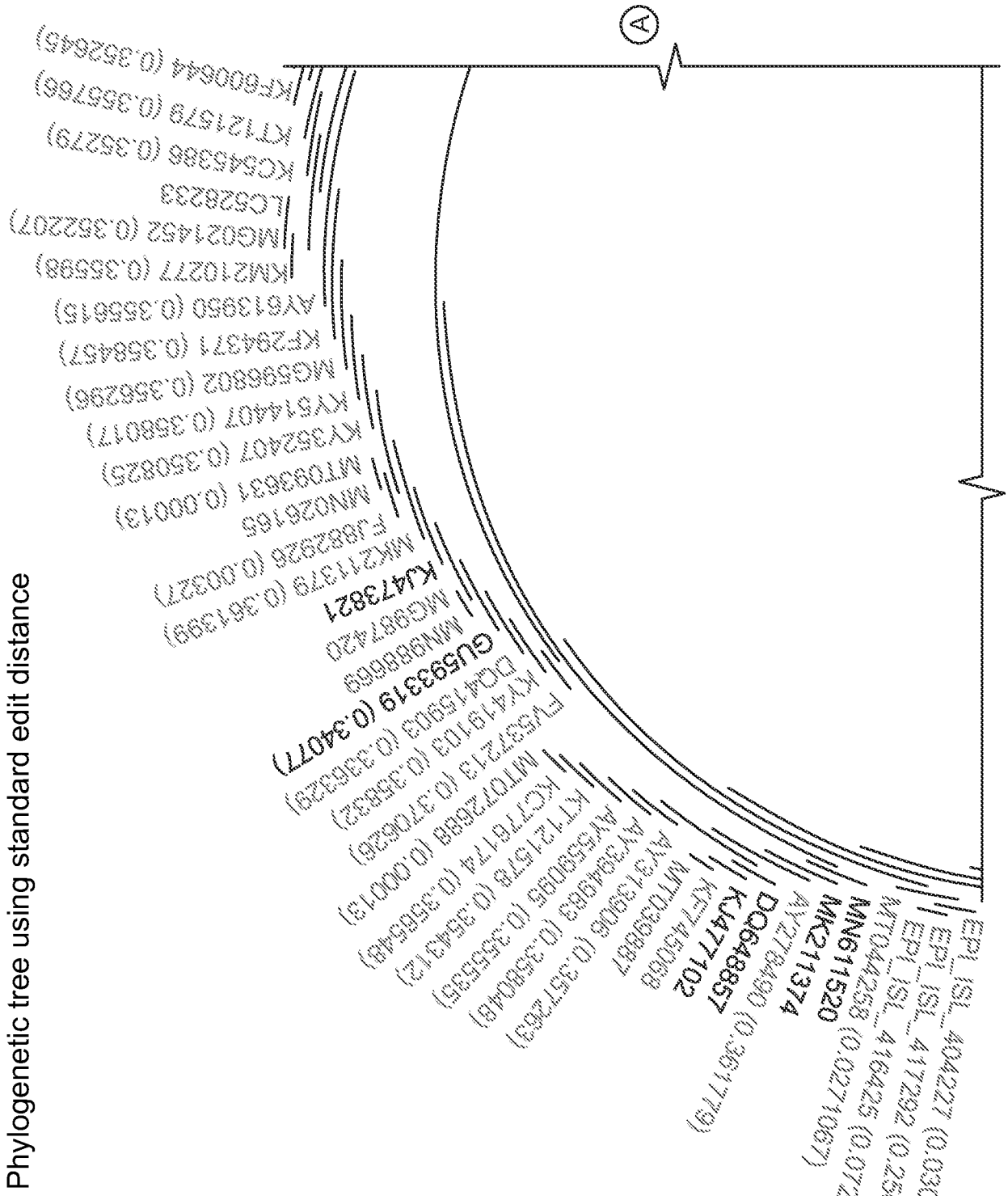
Phylogenetic tree using standard edit distance
A
EpI_ISL_404427 (0.0302229)
EpI_ISL_417292 (0.256076)
EpI_ISL_416425 (0.0724)
MT044258 (0.027067)
MN611520
MK211374
AY278490 (0.36779)
DQ648857
KJ477102
KF745069
MT039887
AY313906 (0.357263)
AY394989 (0.355046)
AY559095 (0.355535)
KT121579 (0.355412)
MT078174 (0.355648)
MC778174 (0.000013)
MT072688 (0.370626)
MT072213 (0.358332)
FJ537213 (0.363329)
KY419103 (0.33629)
DQ415903 (0.33629)
GU593319 (0.3407)
MN988669
MG987420
KJ473821
MK211379 (0.361399)
FJ882926 (0.00327)
MN025165
MT083631 (0.00013)
KY352407 (0.350825)
KY514407 (0.358017)
MG596802 (0.356296)
KF294371 (0.358457)
AY613950 (0.355615)
KM210277 (0.35598)
MG021452 (0.352207)
LC528233
KC546386 (0.35279)
KT121579 (0.356766)
KF600644 (0.352645)

FIG. 8B

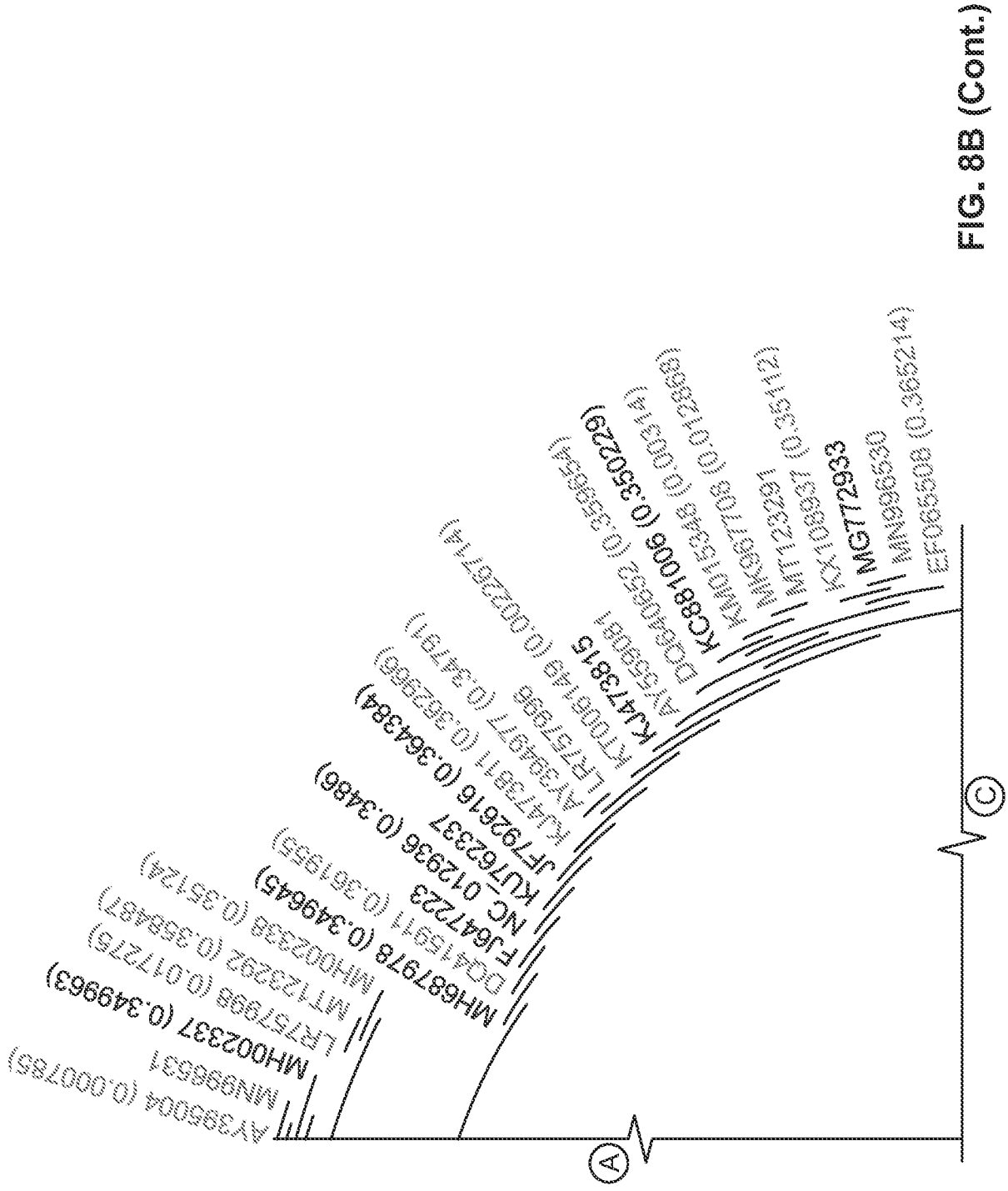
EF065608 (0.36214)
MN896530
MG772833
KX108937 (0.35112)
MT723291
MK067708 (0.01288)
KM015548 (0.00914)
KC881005 (0.35022)
DQ64065 (0.35854)
AY539315
KU473815
KT006149 (0.00228714)
KT757936
LR757936
AY364077 (0.34791)
KU473911 (0.36229B)
JF792616 (0.36438A)
KU762337
MC_012936 (0.3486)
KU764223
FJ647228 (0.34964B)
DQ415911 (0.36195B)
MH002338 (0.35124)
MT123292 (0.35B4B7)
LR757998 (0.01727B)
MH002337 (0.34996B)
MN996531
AY395004 (0.000786)
A
C

FIG. 8B (Cont.)

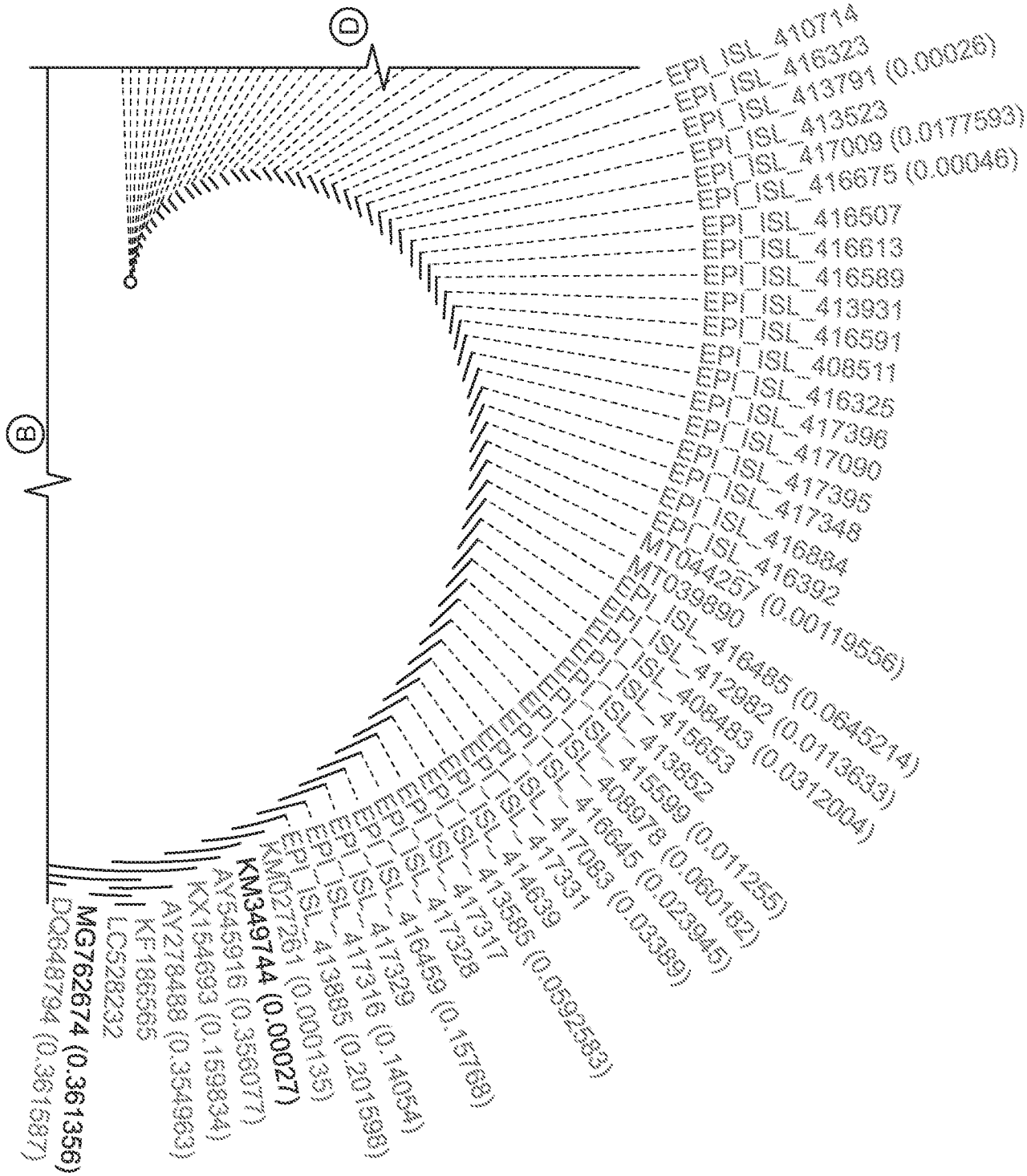
D
B
EPI_ISL_410714
EPI_ISL_416323
EPI_ISL_413791 (0.00026)
EPI_ISL_413523
EPI_ISL_417009 (0.0177593)
EPI_ISL_416675 (0.00046)
EPI_ISL_416507
EPI_ISL_416613
EPI_ISL_416589
EPI_ISL_413931
EPI_ISL_416591
EPI_ISL_408511
EPI_ISL_416325
EPI_ISL_417396
EPI_ISL_417090
EPI_ISL_417395
EPI_ISL_417348
EPI_ISL_416884
EPI_ISL_416392
MT044257 (0.00119556)
MT039890
EPI_ISL_416485 (0.0645214)
EPI_ISL_412982 (0.0113633)
EPI_ISL_408493 (0.0312004)
EPI_ISL_415553
EPI_ISL_413852
EPI_ISL_415599 (0.011255)
EPI_ISL_408978 (0.0601182)
EPI_ISL_416645 (0.023945)
EPI_ISL_417083 (0.03359)
EPI_ISL_417331
EPI_ISL_414639
EPI_ISL_413585 (0.0592583)
EPI_ISL_417317
EPI_ISL_417328
EPI_ISL_416459 (0.15768)
EPI_ISL_417329
EPI_ISL_417316 (0.14054)
EPI_ISL_413885 (0.201599)
EPI_ISL_413261 (0.000135)
KM349744 (0.00027)
KM027261 (0.000135)
AY545916 (0.3580077)
KX154693 (0.1598334)
AY278488 (0.3549093)
KF186565
LC528232
MG762674 (0.361356)
DQ648794 (0.3615987)

FIG. 8B (Cont.)

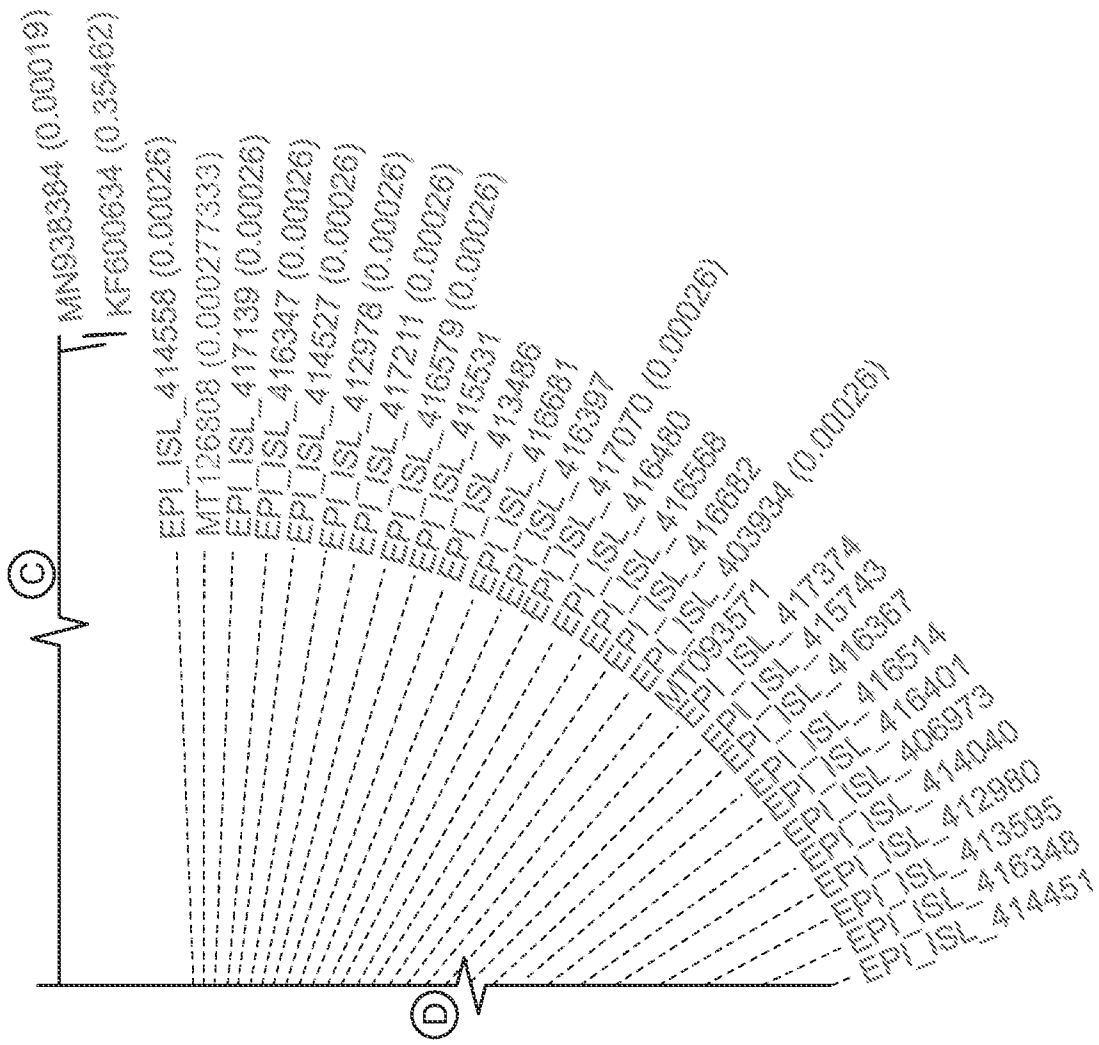
MN938384 (0.00019)
KF600634 (0.35462)
EPI_ISL_414558 (0.00026)
MT126808 (0.00027733)
EPI_ISL_417139 (0.00026)
EPI_ISL_416347 (0.00026)
EPI_ISL_414527 (0.00026)
EPI_ISL_412978 (0.00026)
EPI_ISL_417211 (0.00026)
EPI_ISL_416579 (0.00026)
EPI_ISL_415531
EPI_ISL_413486
EPI_ISL_416681
EPI_ISL_416397
EPI_ISL_417070 (0.00026)
EPI_ISL_416480
EPI_ISL_416568
EPI_ISL_416682
EPI_ISL_403934 (0.00026)
MT093571
EPI_ISL_417374
EPI_ISL_415743
EPI_ISL_416367
EPI_ISL_416514
EPI_ISL_416401
EPI_ISL_406973
EPI_ISL_414040
EPI_ISL_412980
EPI_ISL_413595
EPI_ISL_416348
EPI_ISL_414451
C
D

FIG. 8B (Cont.)

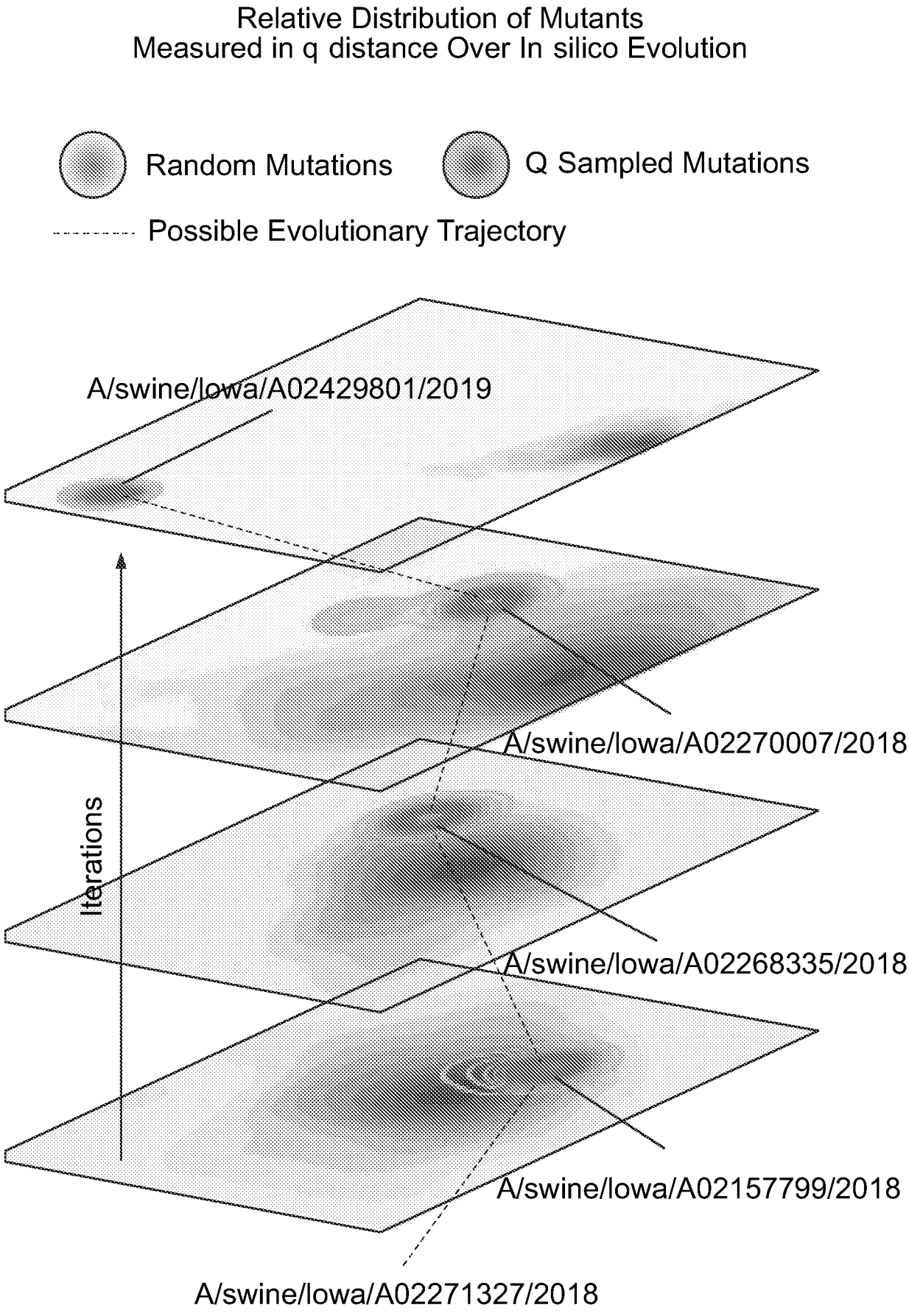
Relative Distribution of Mutants
Measured in q distance Over In silico Evolution
Random Mutations
Q Sampled Mutations
Possible Evolutionary Trajectory
A/swine/lowa/A02429801/2019
A/swine/lowa/A02270007/2018
A/swine/lowa/A02268335/2018
A/swine/lowa/A02157799/2018
A/swine/lowa/A02271327/2018
Iterations

FIG. 9A

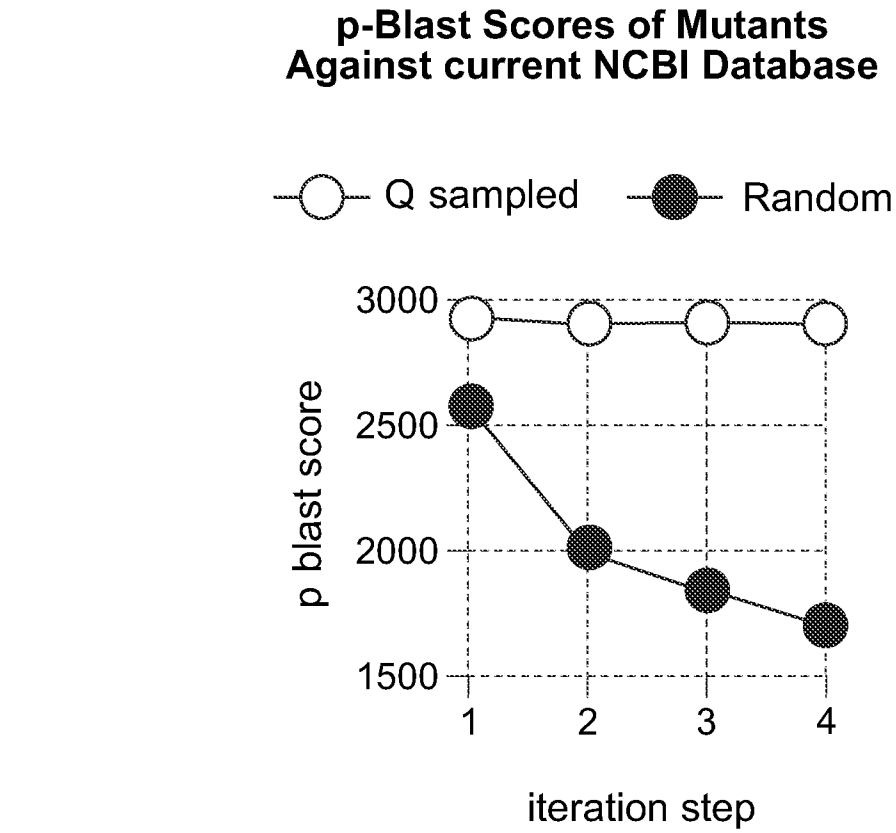
p-Blast Scores of Mutants
Against current NCBI Database
Q sampled
Random
3000
2500
2000
1500
p blast score
1
2
3
4
iteration step

FIG. 9B

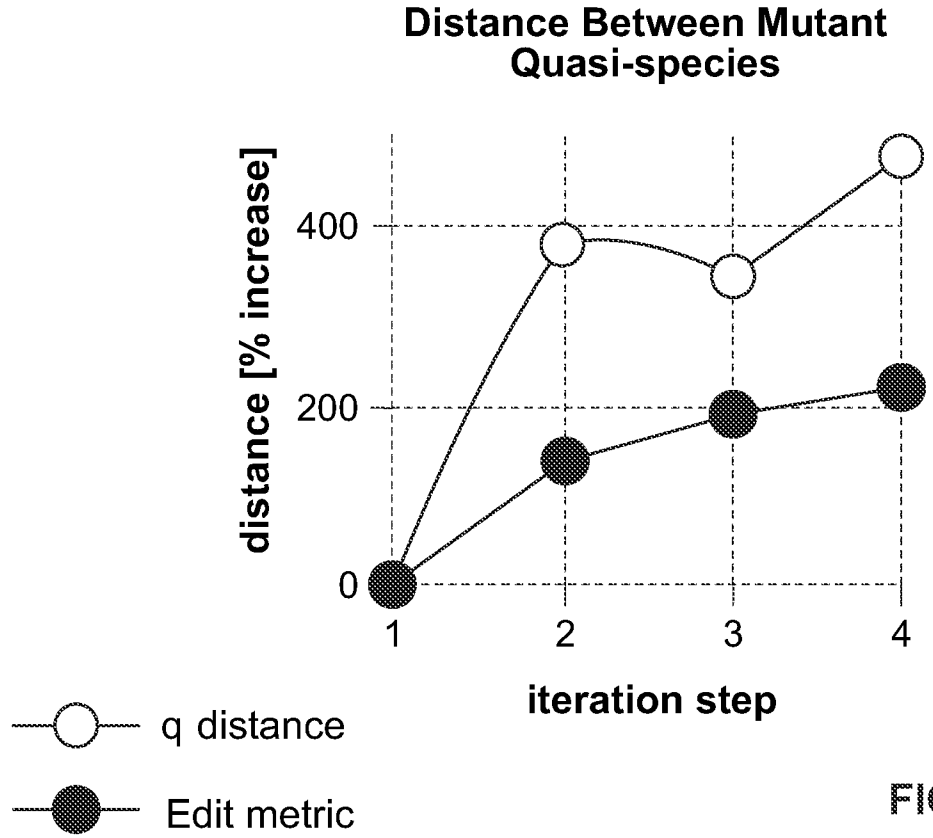
Distance Between Mutant
Quasi-species
distance [% increase]
400
200
0
1
2
3
4
iteration step
q distance
Edit metric

FIG. 9C

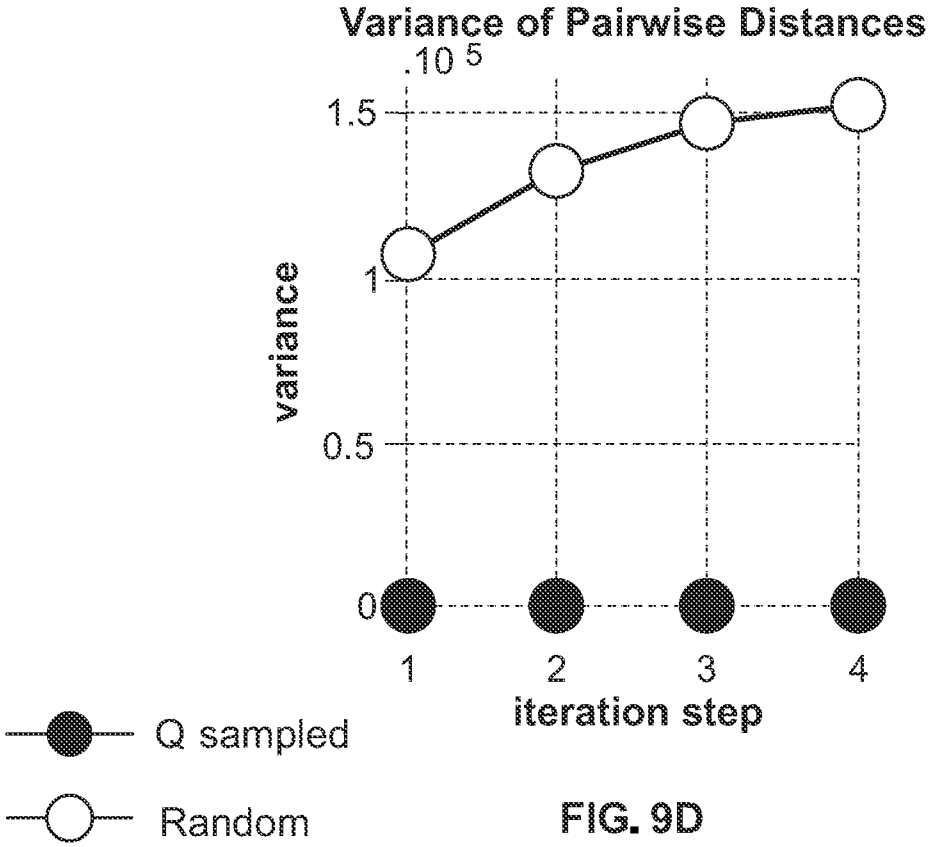
Variance of Pairwise Distances
.10 5
1.5
1
0.5
0
variance
1
2
3
4
iteration step
Q sampled
Random

FIG. 9D

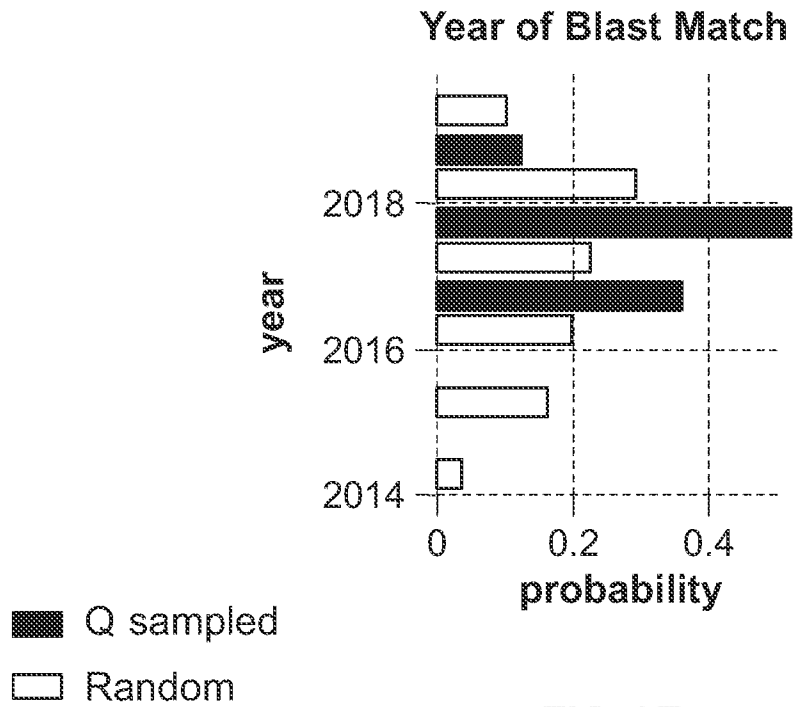
Year of Blast Match
2018
2016
2014
year
0
0.2
0.4
probability
Q sampled
Random

FIG. 9E

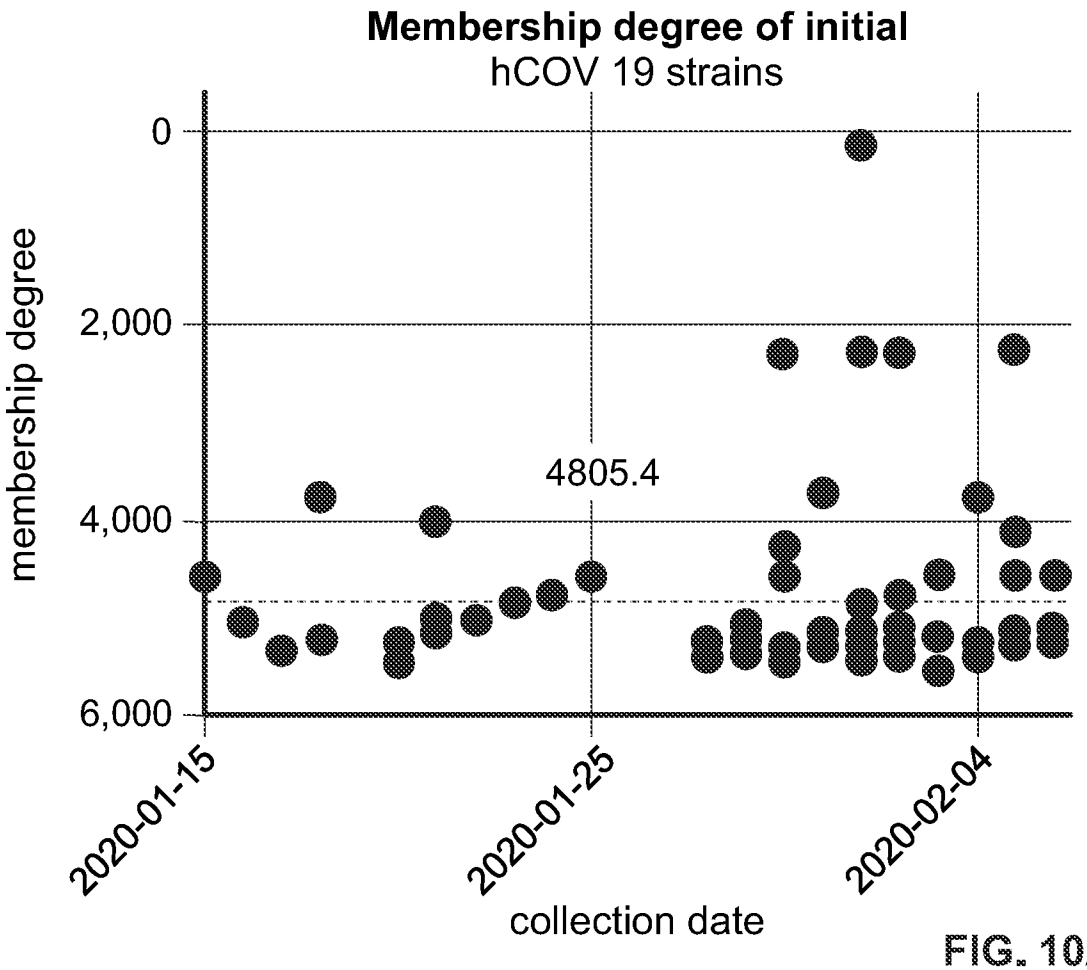

FIG. 10A

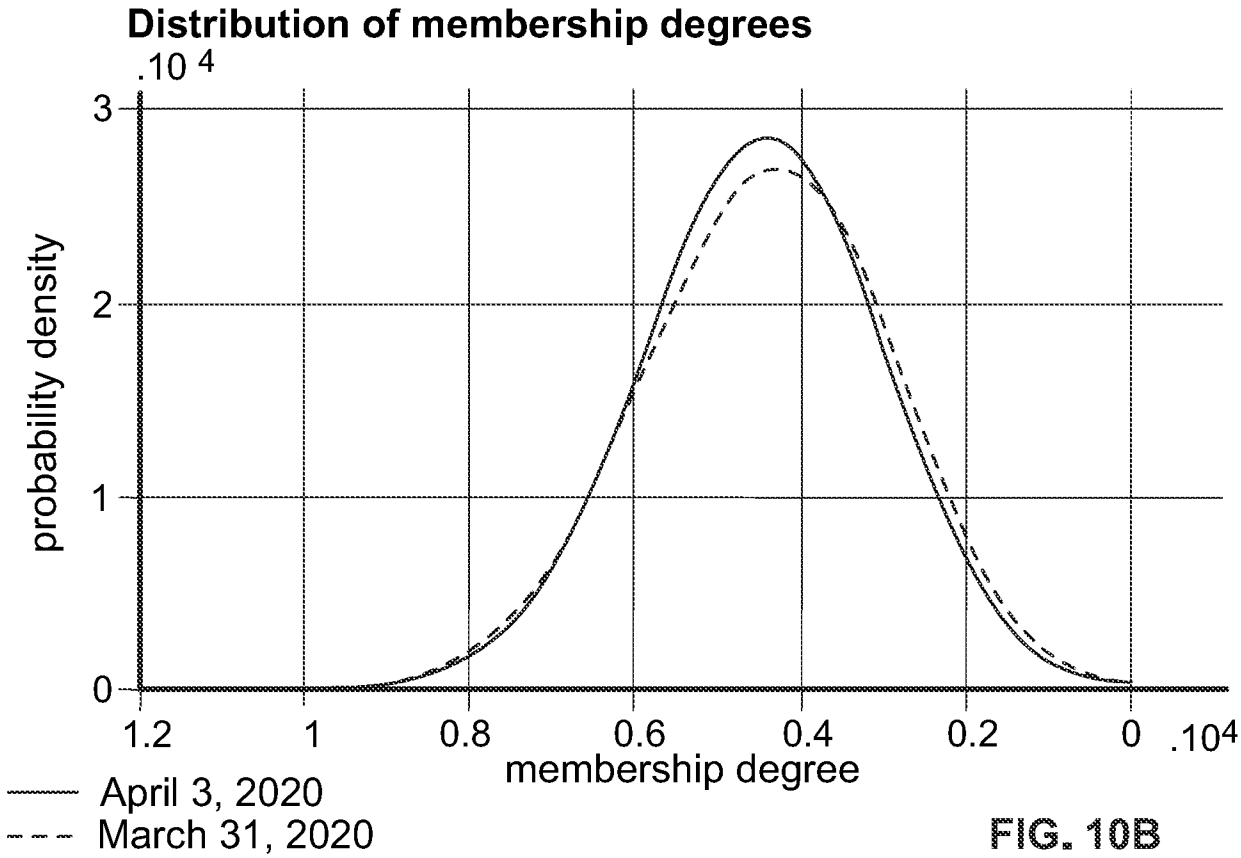

FIG. 10B

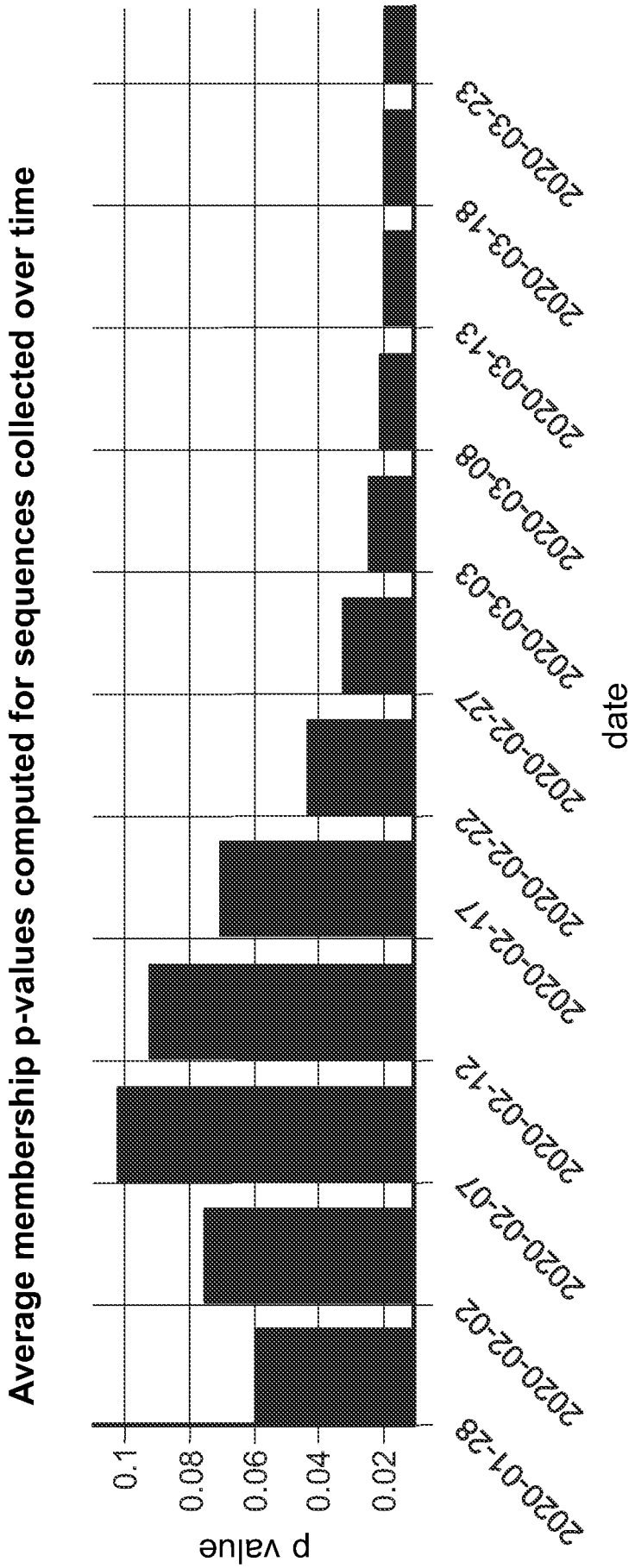
Average membership p-values computed for sequences collected over time
p value
0.1
0.08
0.06
0.04
0.02
2020-01-28
2020-02-02
2020-02-07
2020-02-12
2020-02-17
2020-02-22
2020-02-27
2020-03-03
2020-03-08
2020-03-13
2020-03-18
2020-03-23
date

FIG. 10C

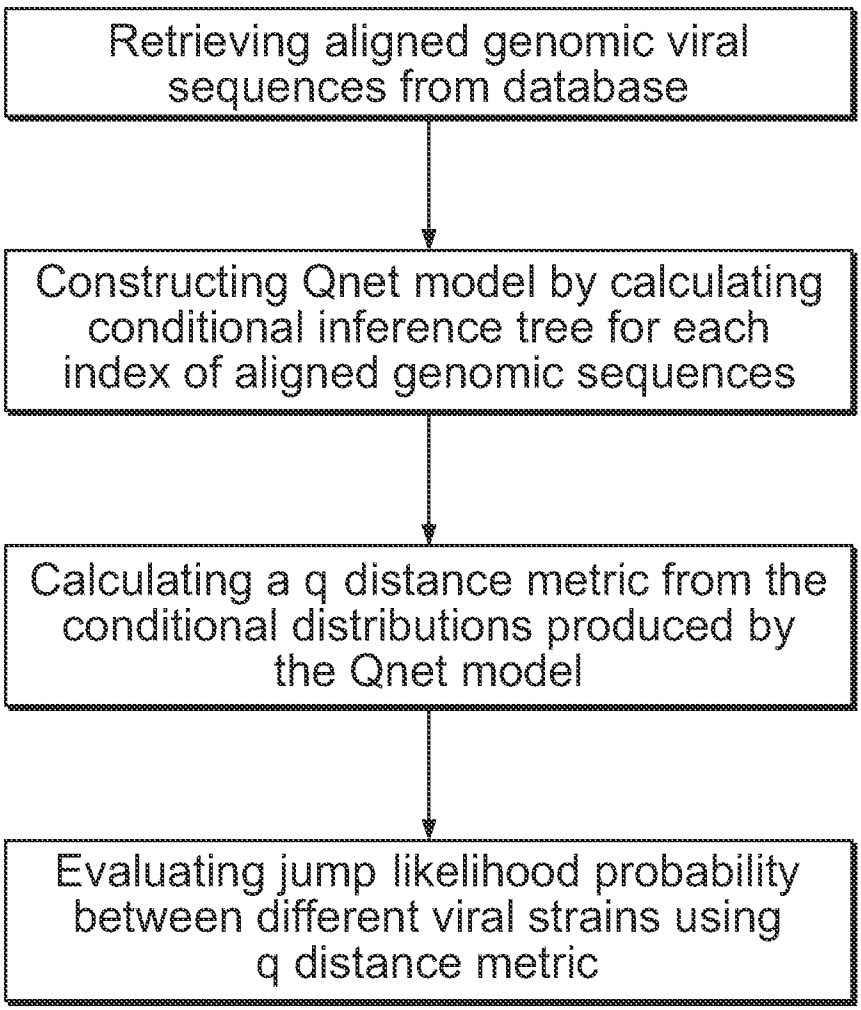
Retrieving aligned genomic viral sequences from database
Constructing Qnet model by calculating conditional inference tree for each index of aligned genomic sequences
Calculating a q distance metric from the conditional distributions produced by the Qnet model
Evaluating jump likelihood probability between different viral strains using q distance metric

FIG. 11

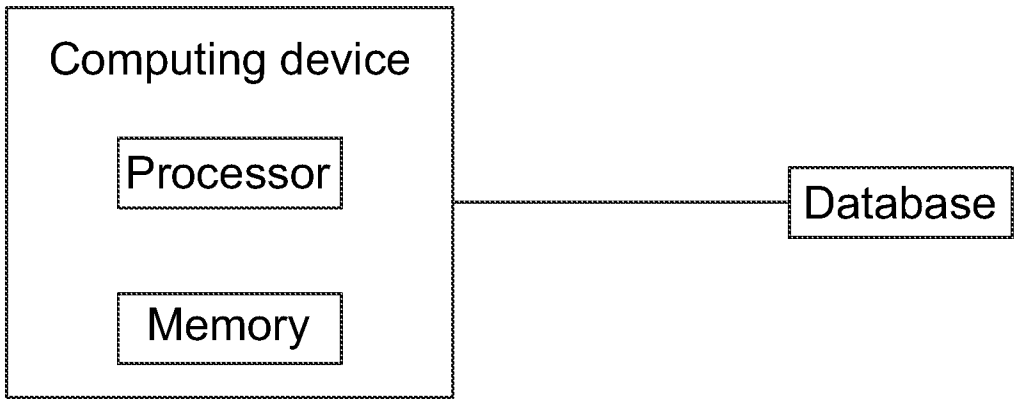

FIG. 12

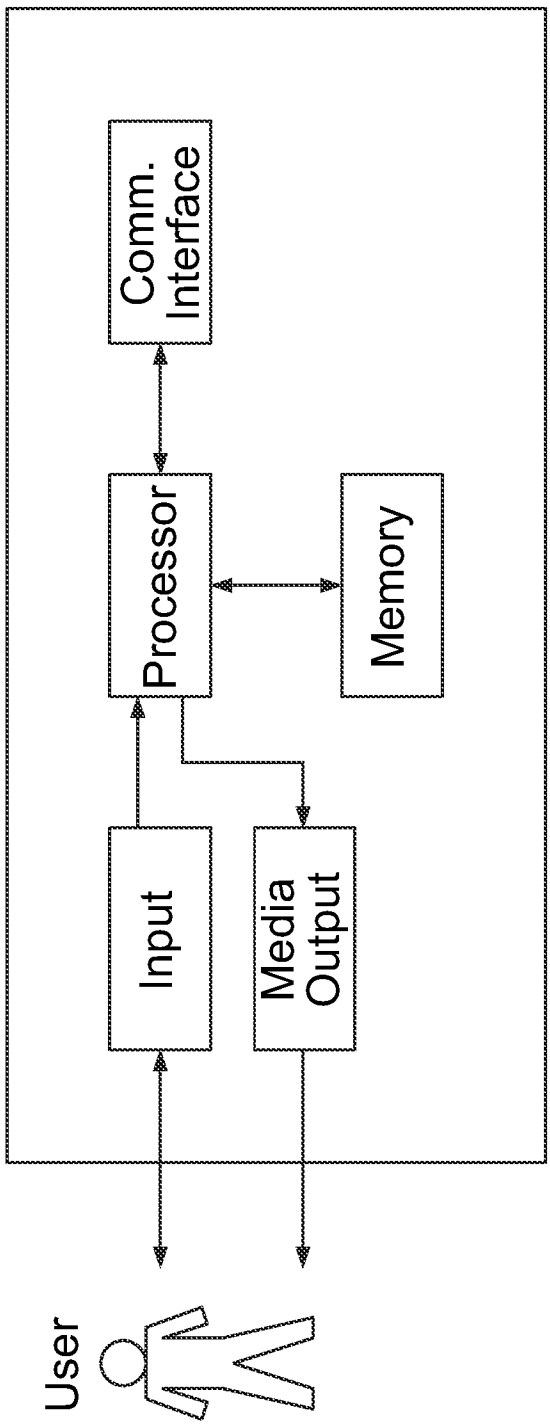

FIG. 13

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US21/59616

A. CLASSIFICATION OF SUBJECT MATTER

IPC - G16H 50/80; G16B 30/00; G16B 50/00; G16B 20/00 (2021.01)

CPC - G16H 50/80; G16B 30/00; G16B 50/00; G16B 20/00; G06N 3/086; C12N 2770/20011

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
See Search History document

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
See Search History document

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
See Search History document

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	(LI, J et al.) Preparing For the Next Pandemic: Learning Wild Mutational Patterns At Scale For Analyzing Sequence Divergence In Novel Pathogens. medRxiv. Epub 20 July 2020, pages 1-14; page 2, 2nd paragraph; page 6, 2nd and 5th paragraphs; page 7, 1st paragraph; page 9, 4th paragraph; page 10, 4th paragraph; page 11, 4th paragraph; figs. 1, 2, 3, 6, S2a; table 1; DOI. 10.1101/2020.07.17.20150004	1-11
A	(HAN, BA et al.) Rodent reservoirs of future zoonotic diseases. Proceeding of the National Academy of Sciences of the USA. 2 June 2015, Epub 18 May 2015, Vol. 112, No. 22; pages 7039-7044; entire document; DOI: 10.1073/pnas.1501598112	1-11

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"D" document cited by the applicant in the international application
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
25 March 2022 (25.03.2022)	APR 07 2022
Name and mailing address of the ISA/US Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-8300	Authorized officer Shane Thomas Telephone No. PCT Helpdesk: 571-272-4300

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US21/59616

Box No. I Nucleotide and/or amino acid sequence(s) (Continuation of item 1.c of the first sheet)

1. With regard to any nucleotide and/or amino acid sequence disclosed in the international application, the international search was carried out on the basis of a sequence listing:

   a. ☐ forming part of the international application as filed:

      ☐ in the form of an Annex C/ST.25 text file.

      ☐ on paper or in the form of an image file.

   b. ☐ furnished together with the international application under PCT Rule 13*ter*.1(a) for the purposes of international search only in the form of an Annex C/ST.25 text file.

   c. ☒ furnished subsequent to the international filing date for the purposes of international search only:

      ☒ in the form of an Annex C/ST.25 text file (Rule 13*ter*.1(a)).

      ☐ on paper or in the form of an image file (Rule 13*ter*.1(b) and Administrative Instructions, Section 713).

2. ☐ In addition, in the case that more than one version or copy of a sequence listing has been filed or furnished, the required statements that the information in the subsequent or additional copies is identical to that forming part of the application as filed or does not go beyond the application as filed, as appropriate, were furnished.

3. Additional comments: