Seasonal influenza circulation patterns and projections for September 2019 to September 2020

Trevor Bedford[1], John Huddleston[1], Barney Potter[1] & Richard A. Neher[2]

[1]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA,
[2]Biozentrum, University of Basel, Basel, Switzerland

September 23, 2019

Abstract

This report details current seasonal influenza circulation patterns as of August 2019 and makes projections up to September 2020 to coincide with selection of the 2020 Southern Hemisphere vaccine strain. This is not meant as a comprehensive report, but is instead intended as particular observations that we've made that may be of relevance. Please also note that observed patterns reflect the GISAID database and may not be entirely representative of underlying dynamics. All analyses are based on the nextflu/nextstrain pipeline [1,2] with continual updates posted to nextstrain.org/flu. **A/H3N2:** A/H3N2 viruses continue to show substantial diversity in HA sequences with a deep split between 3c3.A and 3c2.A1b viruses. The most notable recent developments are the rapid rise of clade A1b/137F – a subclade of A1b/135K – in China and Bangladesh and clade A1b/197R – a subclade of A1b/131K – which dominates the ongoing season in Australia. Our models predict that A1b/137F and A1b/197R will be the dominant clades next year with A1b/197R accounting for most circulation. There is, however, large uncertainty in the true extent of A1b/137F circulation. **A/H1N1pdm:** The S183P substitution has risen to near fixation. The most successful subclade carrying this mutation is 183P-5 which has essentially replaced competing variants. A variant with substitutions 129D/185I is at 60% prevalence globally, while a second variant with substitution 130N is at 50% in North America and ∼10% elsewhere. Substitutions at site 156 to D or K have arisen sporadically and result in loss of recognition by antisera raised against viruses with asparagine at position 156. Despite the large antigenic effect, viruses with mutations at site 156 don't seem to spread. Beyond variants at site 156, little to no antigenic evolution is evident in assays with ferret antisera. **B/Vic:** Antigenically drifted deletion variants at HA1 sites 162, 163 and 164 are now dominating global circulation and have all but taken over. The double deletion variant V1A.1 had previously been circulating at high frequency in the Americas. However, over the course of 2009, the triple deletion variant V1A.3 has increased in frequency globally and is now dominating in all geographic regions. Importantly, V1A.1 and V1A.3 variants appear antigenically distinct by HI assays with 4-8 fold reductions in log2 titer in both directions. **B/Yam**: B/Yam has not circulated in large numbers since the Northern Hemisphere season 2017/2018 and displays relatively little amino acid variation in HA or antigenic diversity. Amino acid variants at sites 229 and 232 have begun to circulate and population is now split between 229D/232D, 229N/232D and 229D/232N variants. These variants show little sign of antigenic difference in HI assays.

Contents

Methods **3**

A/H3N2 . . . . . **5**
Current circulation patterns 5

Antigenic properties 10
Fitness model 11

A/H1N1pdm **13**

B/Vic **18**

B/Yam **22**

Host age distributions **26**

Methods and Notes

Sequence data and subsampling

We base our analysis on sequence data available in GISAID as of Sep 16, 2019. The availability of sequences varies greatly across time and geography and we try to minimize geographical and temporal bias by subsampling the data or analyzing different geographical regions separately when appropriate. While this subsampling reduced geographical biases, it doesn't remove this bias entirely. The most recent months are particularly prone to biases due to variable data deposition schedules across geographic regions.

Phylogenetic analysis

The database contains too many sequences to perform a comprehensive phylogenetic analysis of all available data. We hence subsample the data to 90 sequences per month using the following criteria:

- for each month and each of ten geographic regions, we select 9 viruses (or all available viruses if fewer than 9 are available);
- when the total number of viruses selected by the above criterion in a given month is below 90, we fill the remainder evenly with viruses from geographic regions with more available strains;
- within each month and geographic region, viruses with antigenic data are prioritized.

Parallel evolution, that is repeated occurrence of identical substitutions in different clades of the tree, has become common in A/H3N2 and A/H1N1pdm. Such parallel evolution violates fundamental assumptions of common phylogeny software and can erroneously group distinct clades together if they share too many parallel changes. To avoid such artifacts, we mask sites with rampant parallelism prior to phylogeny inference. Clades are assigned using a collection of "signature" mutations available at the GitHub repository github.com/nextstrain/seasonal-flu for each lineage via the following links:

- config/clades_h3n2_ha.tsv
- config/clades_h1n1pdm_ha.tsv
- config/clades_vic_ha.tsv
- config/clades_yam_ha.tsv

Mutation frequency calculations

In contrast to the phylogenetic analysis, mutation frequencies are based on all available data and calculated separately for each geographic region. To obtain estimates of global mutation frequencies, we average geographic regions weighted by their approximate contribution to the global human population. Specifically, frequencies are calculated as follows:

- amino acid sequences of all isolates within a geographic regions are aligned to a reference sequence;
- for each variable alignment column, we infer a frequency trajectory of the different amino acids at this position using a Brownian motion prior [1];

- in each region, the seasonal pattern of sequence data availability is used as a proxy for seasonal prevalence;
- regional frequencies are then averaged both by the seasonal pattern and their population fraction.

The graphs in the report show frequency trajectories for North- and South America, China, Japan/Korea, Europe, and Oceania and omit Africa, South Asia, West Asia, and Southeast Asia to avoid overloading the graphs although these regions still contribute to globally weighted estimates.

Antigenic analysis

We summarize HI and FRA measurements provided by the WHO CCs in London, Melbourne, Atlanta and Tokyo using our *substitution model* [3] which models log-titers as a sum of effects associated with amino acid differences between the sequences of the test and reference virus. In addition, the model allows for a serum (column) and a virus (row) effect. This model allows to infer titers for virus/serum pairs that have not been antigenically characterized and isolates effects consistently observed across many measurements from the noise inherent in individual measurements.

Persistent analyses

For a better historical record, the analyses available on Sep 23, 2019 have been saved as:

- nextstrain.org/flu/seasonal/h3n2/ha/2y/2019-09
- nextstrain.org/flu/seasonal/h1n1pdm/ha/2y/2019-09
- nextstrain.org/flu/seasonal/vic/ha/2y/2019-09
- nextstrain.org/flu/seasonal/yam/ha/2y/2019-09

A/H3N2

A/H3N2 viruses continue to show substantial diversity in HA sequences with a deep split between 3c3.A and 3c2.A1b viruses. Clades A2/re, A3 and A4 are now rare. 3c3.A had been observed at low levels throughout the last four years in the Western Hemisphere and rose to high frequencies in the 2018/2019 season before falling again in frequency recently. It remains geographically restricted to the Americas and isolated European countries. The substantial antigenic distance between 3c3.A and A1b poses a threat of vaccine mismatch. The most notable recent developments are the rapid rise of clade A1b/137F – a subclade of A1b/135K – in China and Bangladesh and clade A1b/197R – a subclade of A1b/131K – which dominates the ongoing season in Australia. Our models predict that A1b/137F and A1b/197R will be the dominant clades next year with A1b/197R accounting for most circulation. There is, however, large uncertainty in the true extent of A1b/137F circulation.

We base our primary analysis on a set of viruses collected between Sep 2017 and Aug 2019, comprising upwards of 300 viruses per month in the Northern hemisphere winter but fewer counts otherwise (Fig. 1).

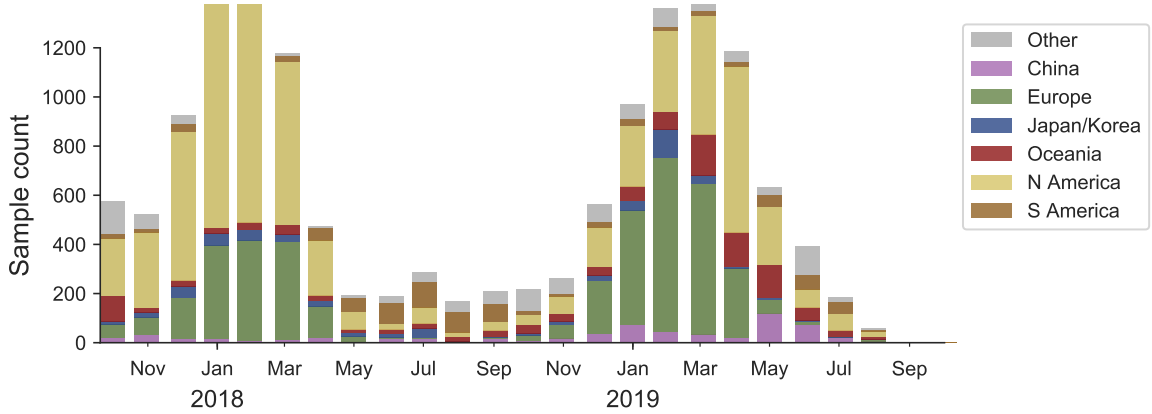

Figure 1. Sample counts through time and across regions. This is a stacked bar plot with the visible height of a color bar corresponding to the sample count from the respective region.

Current circulation patterns

Over the past three years, 3c2.A viruses within A/H3N2 have differentiated into multiple major cocirculating clades with variable geographic distributions (Fig. 2). The past year was dominated by clade A1b. The reassortant clade A2/re that was common in the NH 2017/2018 is now rare. Subclades A1b/131K and A1b/135K comprised the majority of A1b viruses. Over the last 18 months, 3c3.A viruses have increased markedly in the US and Europe and accounted for 60% of isolates in North America and 10-20% in Europe during the last NH winter, see Fig. 4.

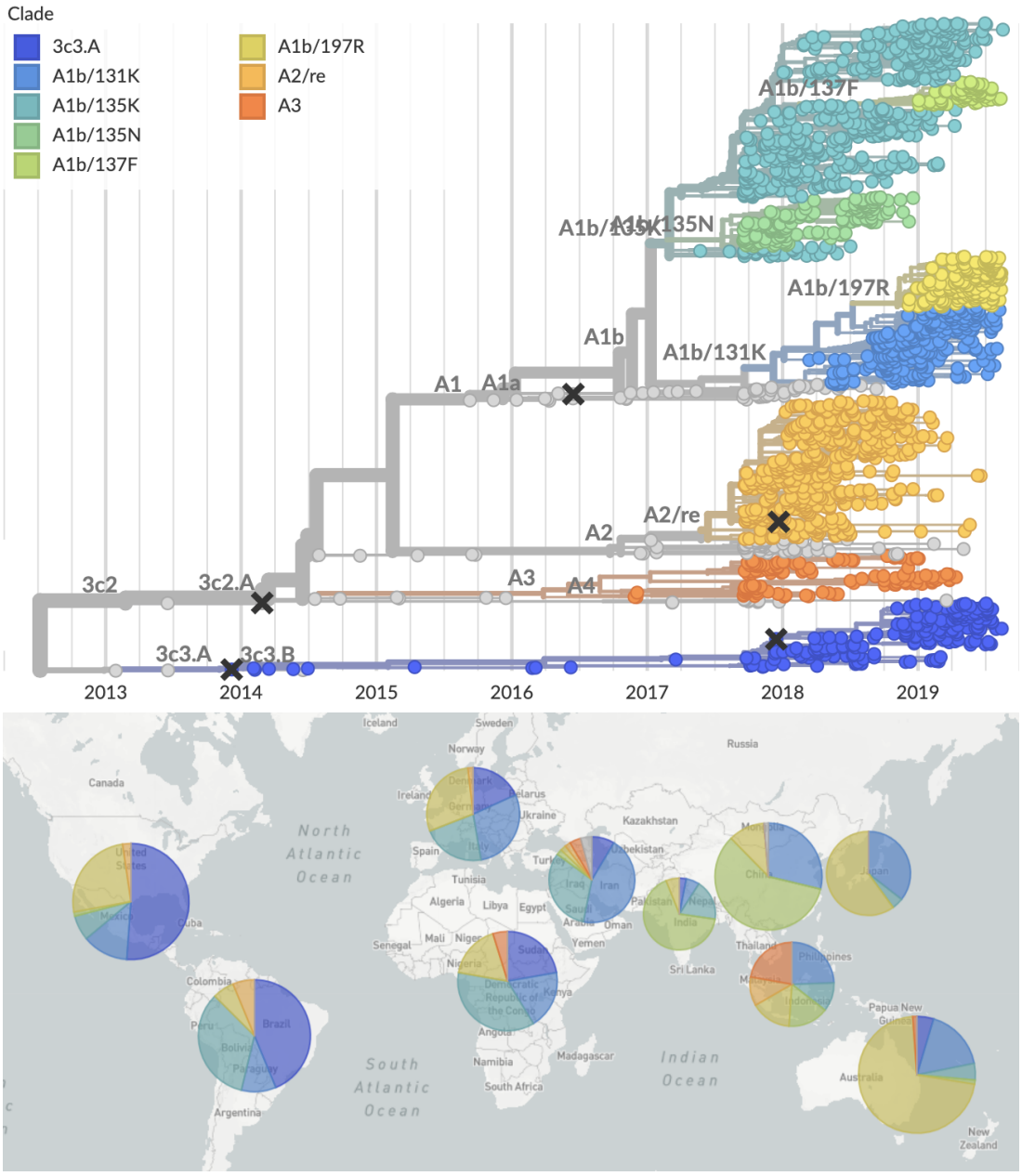

Figure 2. H3N2 phylogeny colored by clade and the corresponding geographic breakdown. The shown geographic distribution is for viruses from 2019 only.

Recently, new subclades within A1b have emerged and risen to appreciable global frequency. These are A1b/197R that has appeared on top of the A1b/131K background and A1b/137F that has appeared on top of the A1b/135K background. A1b/197R viruses have predominated the 2019 SH Oceania season and are now commonly observed in the NH as well. A1b/137F viruses have largely been localized to China and Bangladesh, but are prevalent within these regions. Outside of Bangladesh, very few sequences from South Asia (India in particular) have been deposited in GISAID. There is therefore substantial uncertainty in the estimates of the total circulation of A1b/137F viruses.

Major clades and their associated substitutions include:

- Clade A1b/131K: 142G, T131K
- Clade A1b/135K: 142G, T135K
- Clade A1b/135N: T135N
- Clade A1b/137F: T135K, T128A, S137F, A138S, F193S (subclade within A1b/135K)
- Clade A1b/197R: T131K, Q197R, S219F (subclade within A1b/131K)
- Clade A2/re: T131K, R142K, A212T (reassortant clade)
- Clade 3c3.A: S91N, N144K, F159S, F193S

The frequency trajectories of the these major clades in different geographic regions are shown in Figure 4. In the first half of 2018, clade A2/re and A1b/135K were dominant. Subsequently subclade A1b/131K rose in frequency towards the end of 2018 and now accounts for 50% of global circulation with an increasing trend. Clade A1b/131K and subclade A1b/197R have dominated the anomalously early season in Australia. Clade A1b/135K has had a wide geographic distribution for the last 2 years and stayed at a steady frequency of about 30% globally, albeit with changing regional prevalence. The subclade A1b/137F with additional substitutions A138F and F193S has recently become common in China and Bangladesh and is expanding rapidly. The clade A2/re has not been observed in recent months with the exception of sporadic isolation in South America. Clade 3c3.A continues to be common in the Americas, while having been partly replaced by A1b/131K in North America over the past couple of months. These dynamics of clade frequencies are reflected in the local branching index and other measures of recent clade growth discussed later in the report (see Figure 3).

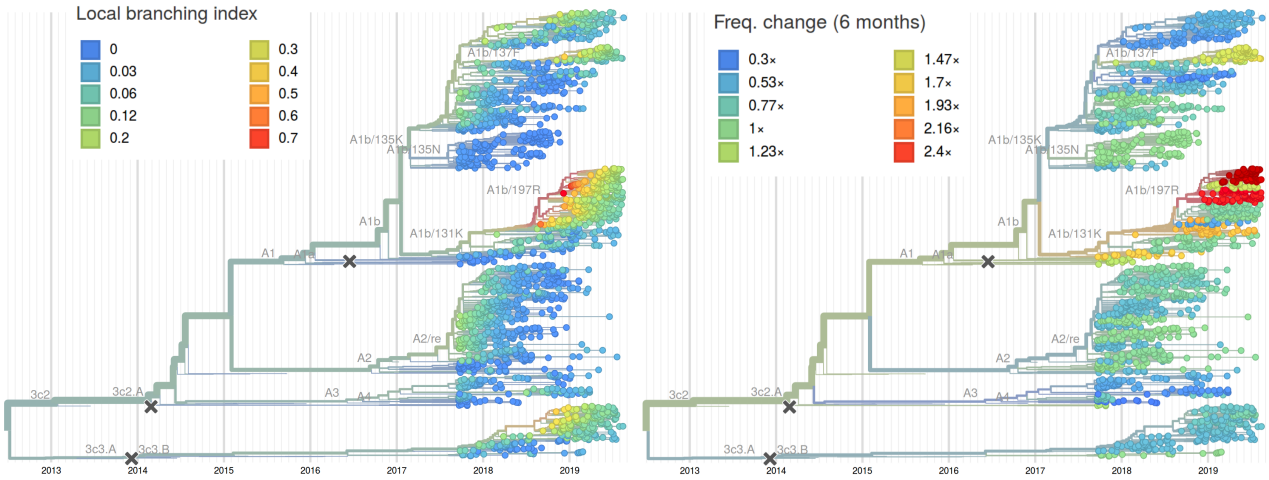

Figure 3. H3N2 phylogenies colored by growth. Left: Colored by LBI [4], a measure of recent clade expansion that is predictive of future success. Right: Colored by recent clade fold-change in frequency. Clade A1b/197R scores highest on both measures. The newly emerging clade A1b/137F scores second highest for recent clade growth, while 3c3.a and A1b/137F are tied for LBI. The LBI averages clade growth over longer time and factors in frequency information, while the growth measure on the right is more sensitive to recent changes and rapid growth of small clades.

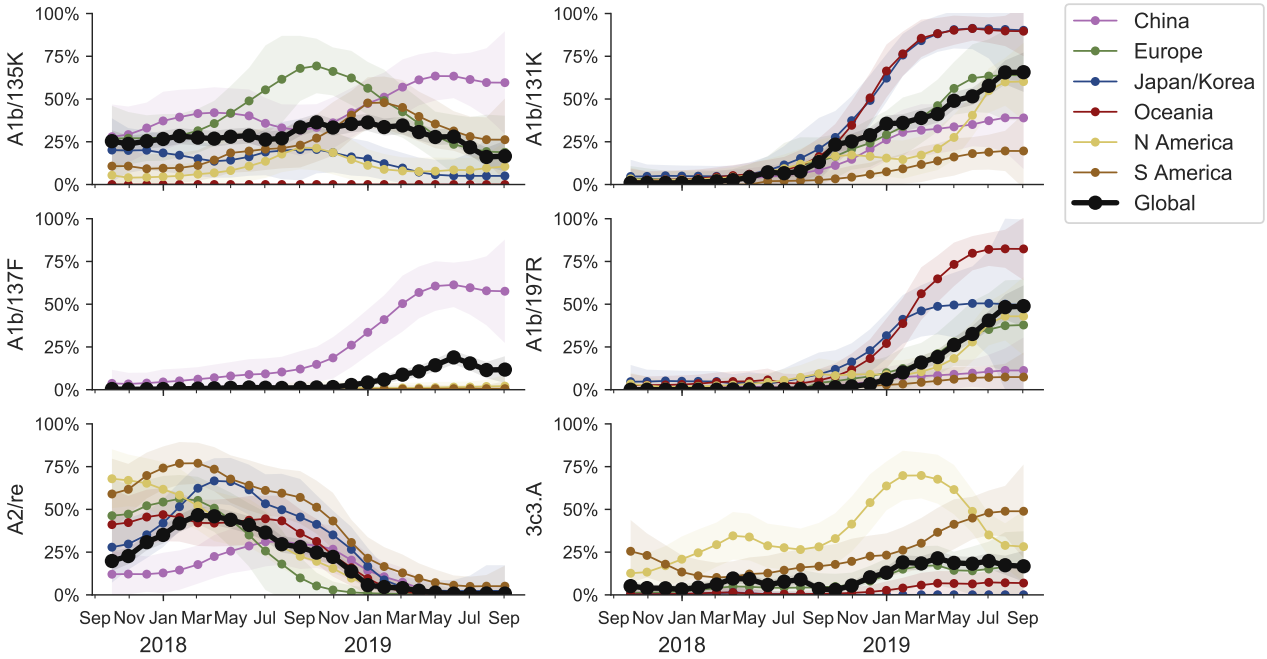

Figure 4. Frequency trajectories of H3N2 clades partitioned by clade then by region. We estimate frequencies of different clades based on sample counts and collection dates of strains included in the phylogeny. We use a Brownian motion process prior to smooth frequencies from month-to-month. Transparent bands show an estimate of the 68% confidence interval based on sample counts.

Frequencies of specific substitutions are shown in Figure 5. We would like to draw attention to the following patterns of convergent evolution and recent emergence:

- **HA1:T128A** arose in a subclade of **A1b/135K** and clade **3c3.A**, but not in A1b/131K.
- **HA1:T131K** was common in the 2017/2018 season as part of **A2/re** and rose again in frequency as part of clade **A1b/131K**
- **HA1:T135K** has arisen multiple times in the recent past, but its current circulation is restricted to clade **A1b/135K**.
- **HA1:137F** recently emerged along with 138S and 193S and rose rapidly in China to levels above 60%. Recent viruses sampled in Bangladesh also fall into this clade, but it has only been observed sporadically elsewhere. **3c3.A** also has A138S.
- **HA1:142G** arose in clades **A1b/135K**, **A1b/131K**, and **3c3.A** and has replaced all other variants at this position.
- **HA1:197R** arose within clade A1b/131K and is dominating the season in Australia. This substitution sits on a 219F background.

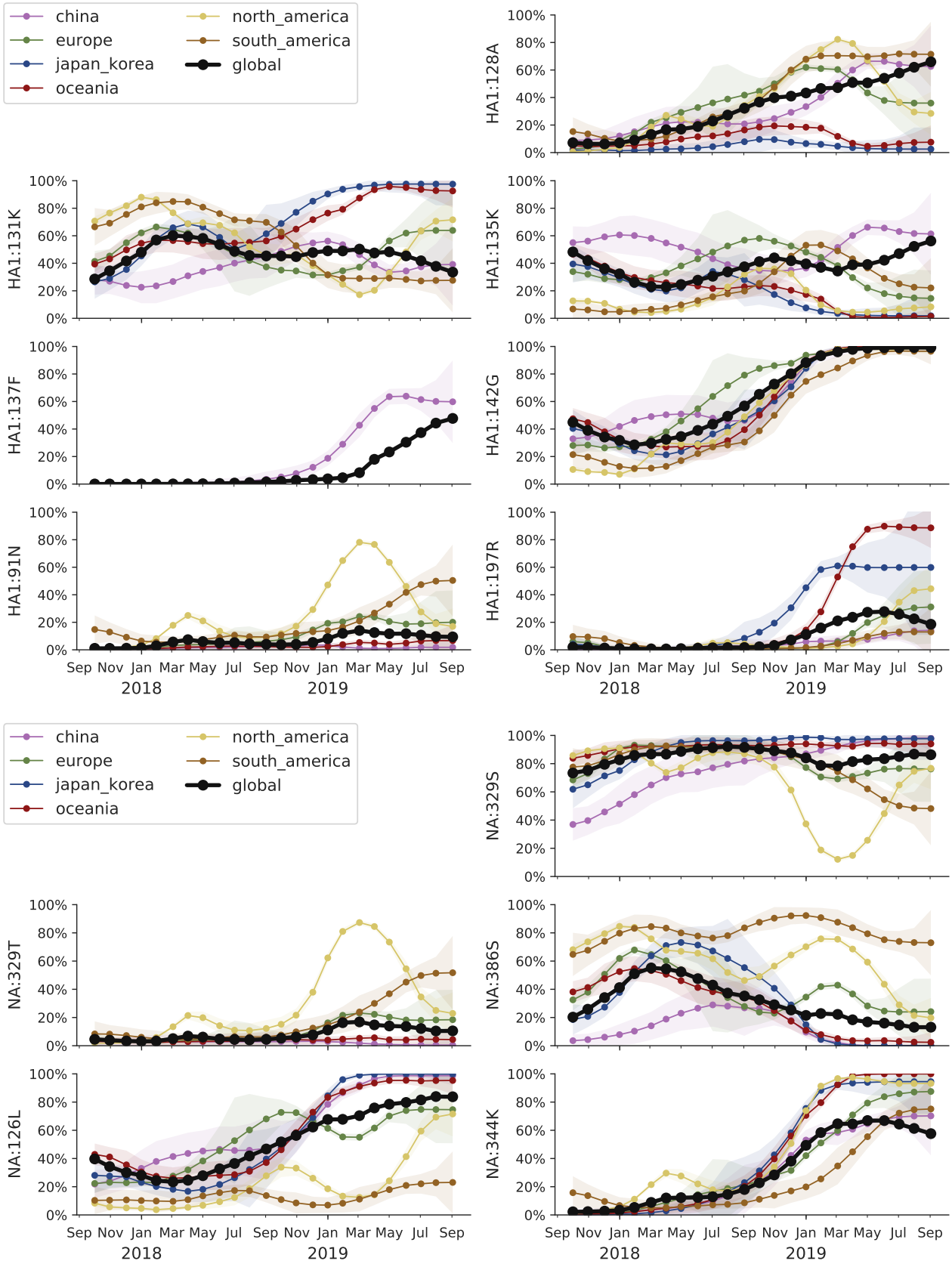

Figure 5. Frequency trajectories of substitutions in H3N2 HA and NA. We estimate frequencies of different amino acid variants based on sample counts and collection dates. These estimates are based on all available data. We use a Brownian motion process prior to smooth frequencies from month-to-month. Transparent bands show an estimate of the 68% confidence interval based on sample counts.

Antigenic properties

We integrated HI and FRA/VN data from the WHO Collaborating Centers in London, Tokyo, Melbourne and Atlanta with molecular evolution of the HA segment. Patterns of recent antigenic evolution are most clearly seen in FRA titers using serum raised against A/NorthCarolina/4/2016 (see Fig. 6). Our titer model suggests a small drop (2-fold) in titers for clade A1b/197R and a moderate drop (2- to 4-fold) for clade A1b/137F. The latter inference, however, is based on a small number of measurements. Overall, little antigenic evolution is detectable by ferret antisera within the 3c2.A clade despite frequent and recurring changes in at epitope sites. Clade 3c3.a, however, is clearly antigenically distinct.

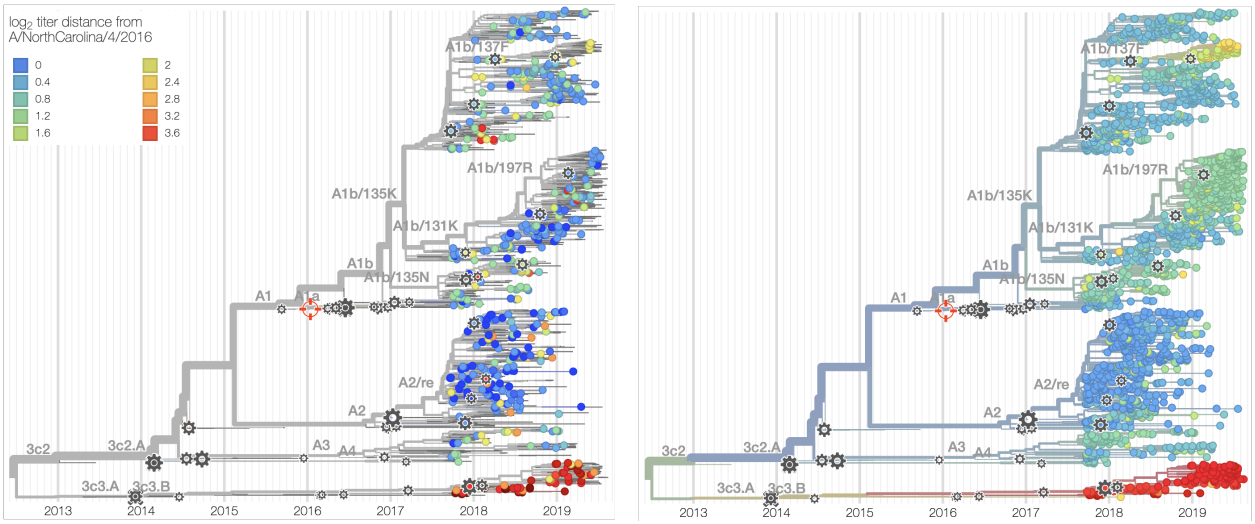

Figure 6. H3N2 phylogeny colored by antigenic distance as measured by FRA to A/NorthCarolina/4/2016. The left panels show the individual measurements, the right panel the model inference. Blue corresponds to good coverage, yellow to a 4-fold drop, red to a 16-fold drop (or higher).

Fitness model

Here, we construct a fitness model to estimate fitness of circulating virus strains and project strain and clade frequencies. We take an approach of measuring an assortment of metrics that may correlate with fitness and then using historical data to weight fitness components in a combined model. We assess the degree to which we can predict changes in strain frequency in one year lookaheads, e.g. how well do projections in Feb 2017 predict strain frequencies in Feb 2018?

Generally, we combine a metrics for antigenic novelty, intrinsic fitness and recent clade growth. More specifically, we include cross-immunity mediated by substitutions at epitope sites, cross-immunity mediated by HI titer differences, substitutions at non-epitope sites as an inverse proxy for intrinsic fitness and recent clade growth as measured by local branching index (LBI). However, in recent years we only observe consistent performance from local branching index with a small improvement from the inclusion of non-epitope sites (Fig. 7). Measurements of antigenic novelty either by epitope mutations or by HI titers have not consistently improved model performance since 2009. Thus, we believe the most robust model is one that focuses on LBI supplemented by non-epitope sites. Here, we present results from an LBI-only model as well as a model that combines LBI and non-epitope mutations.

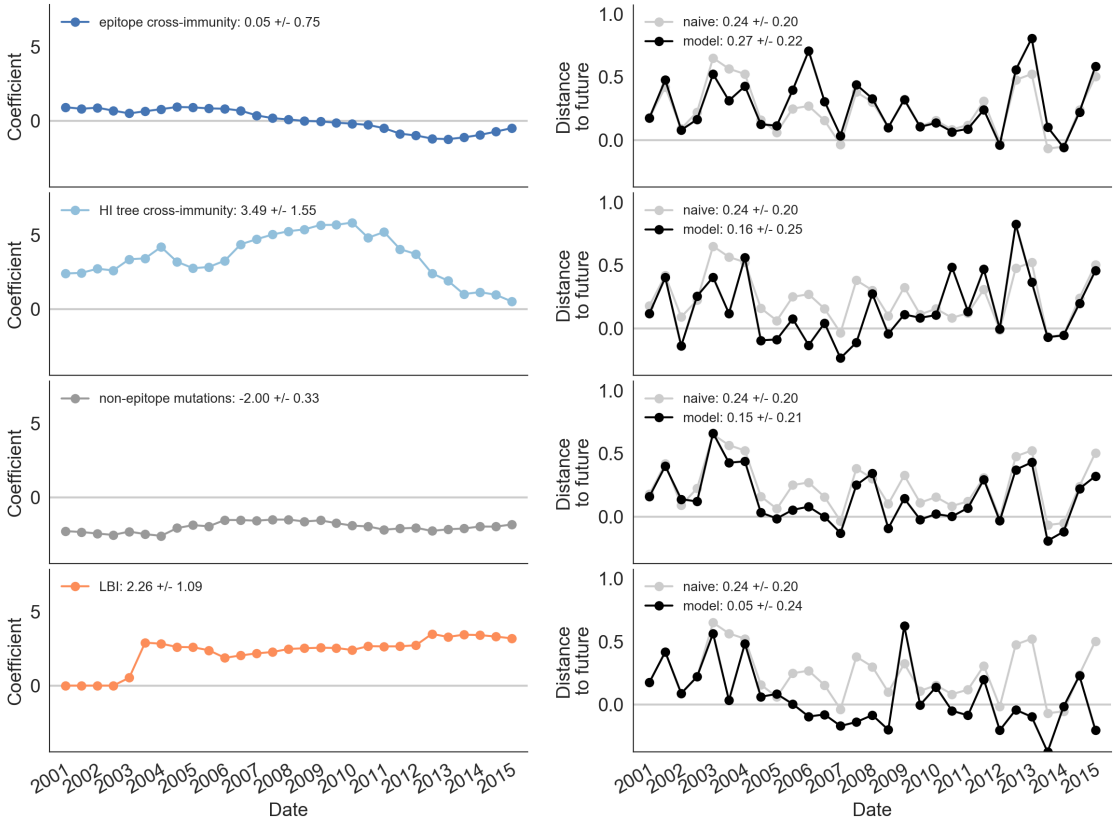

Figure 7. Model performance in previous seasons. Left-hand panels show predictor coefficients; a good predictor will have a consistent non-zero coefficient through time. Right-hand panels show model performance as Hamming distance between predicted strain distribution and retrospectively observed strain distribution; a good model will have black lines (showing model predictions) consistently below gray lines (showing predictions of a 'naive' model).

As illustrated previously (Fig. 3), clade A1b/197R viruses have been increasing in frequency and have the highest values of local branching index. This results in clade A1b/197R and clade A1b/131K having higher fitness than A1b/135K, A1b/137F and 3c3.A viruses. We use these estimated strain fitnesses to project frequencies of individual strains and from these projections forecast resulting clade frequencies (Fig. 8). This shows that both the LBI-only and the LBI with non-epitope sites predictions favor A1b/197R viruses to grow in frequency from ∼32% global frequency to ∼80% global frequency over the next 12 months. This projection has A1b/131K viruses expanding but due completely to the success of A1b/197R. A1b/137F are predicted to persist into the future but at lower frequency than A1b/197R viruses.

It is important to point out that it is likely that clade A1b/197R is better sampled than A1b/137F. Since the LBI is effectively a composite measure of frequency and recent expansion, sampling bias might have led to a too high fitness estimate of A1b/197R and a too low estimate for A1b/137F. Both of these clades will likely persist at high frequencies.

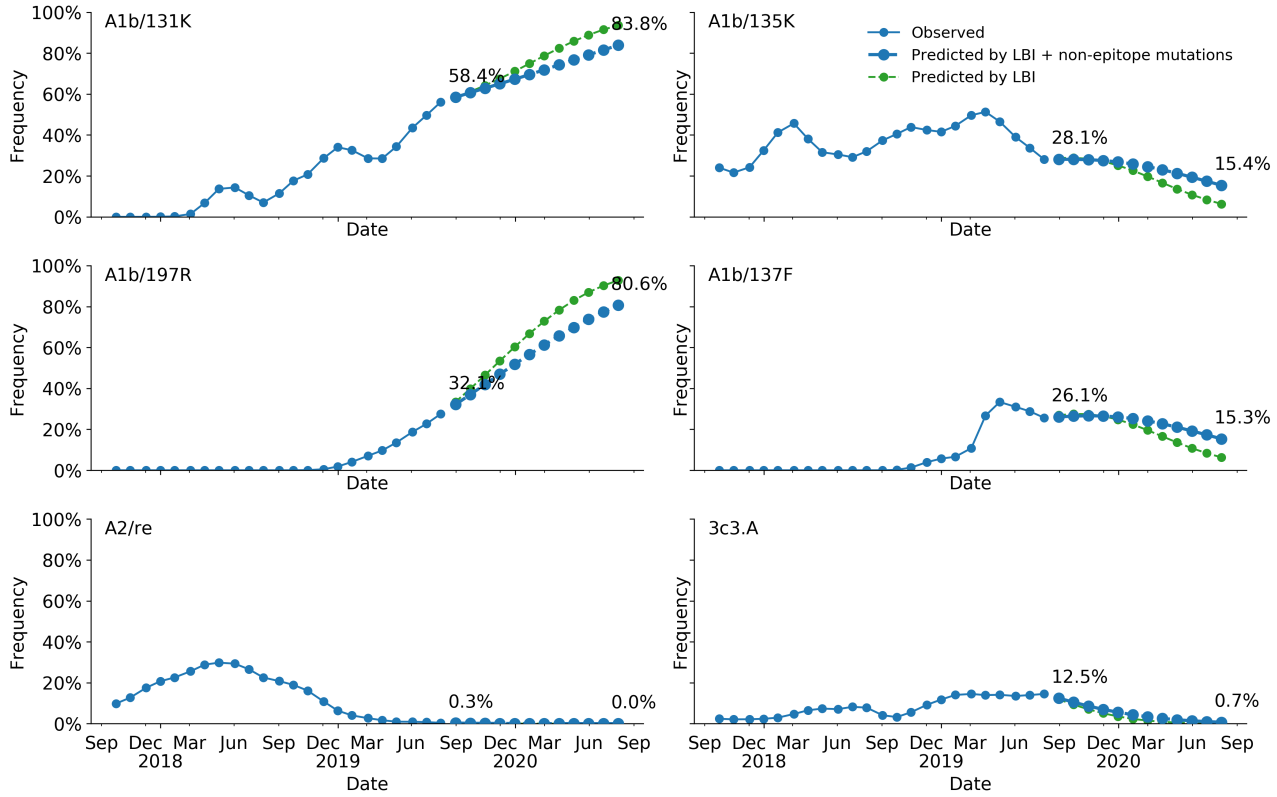

Figure 8. Frequency projection based on fitness model. Having assigned each virus strain with an estimated fitness, it's possible to project clade dynamics. Projections are shown for models using only single predictors as well as for the combined composite model.

A/H1N1pdm

The S183P substitution has risen to near fixation. The most successful subclade carrying this mutation is 183P-5 which has essentially replaced competing variants. This subclade P5, however, is in itself deeply split with a geographically heterogeneous distribution. A variant with substitutions 129D/185I is at 60% prevalence globally, while a second variant with substitution 130N is at 50% in North America and ∼10% elsewhere. Substitutions at site 156 to D or K have arisen sporadically and result in loss of recognition by antisera raised against viruses with asparagine at position 156. Despite the large antigenic effect, viruses with mutations at site 156 don't seem to spread. Beyond variants at site 156, little to no antigenic evolution is evident in assays with ferret antisera.

We base our primary analysis on a set of viruses collected between Jul 2017 and Aug 2019, comprising upwards of 100 viruses per month in almost all months (Fig. 9). We use all available data when estimating frequencies of mutations and weight samples appropriately by regional population size and relative sampling intensity to arrive at a putatively unbiased global frequency estimate. Phylogenetic analyses are based on a representative sample of about 2000 viruses.

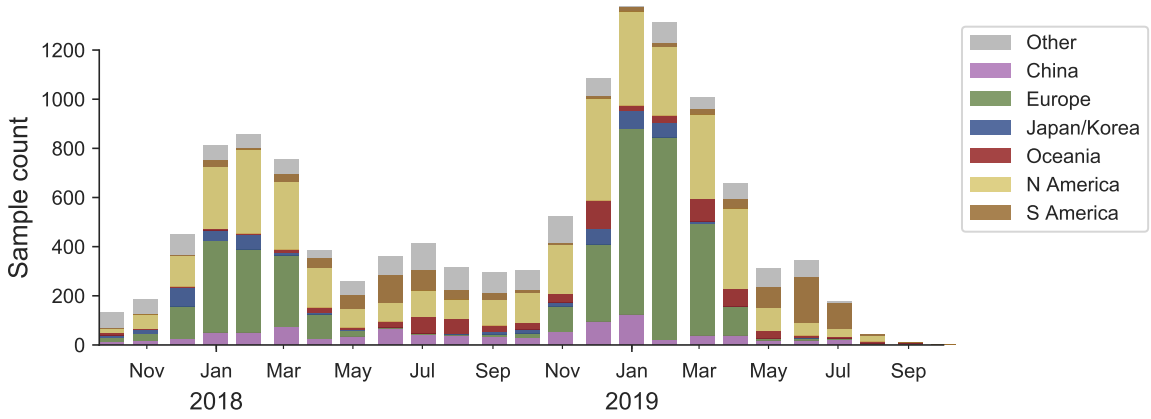

Figure 9. Sample counts through time and across regions. This is a stacked bar plot with the visible height of a color bar corresponding to the sample count from the respective region.

Over the course of the last 2 years, the substitution S183P has arisen multiple times and viruses carrying this substitution have almost completely taken over (Fig. 10). Of the main clades carrying the 183P substitution (labeled P1-P7), clade P5 accounts for 70-100% of recent circulation in different geographic regions (Fig. 12).

This clade P5 is deeply split into a major clade with substitutions N129D and T185I and a minor clade with substitutions K160M, T216K, K130N, H296N. The relative frequencies of these two clades is best analyzed by tracking the frequencies of substitutions 129D and 130N (Fig. 13). The former is dominant in all major geographic regions, while the latter is common, but not dominant, in North America. The antigenically important substitutions N156K and N156D remain at low frequency.

LBI, our measure of clade success, suggests that the major subclade of P5 with substitutions N129D and T185I will continue to dominate.

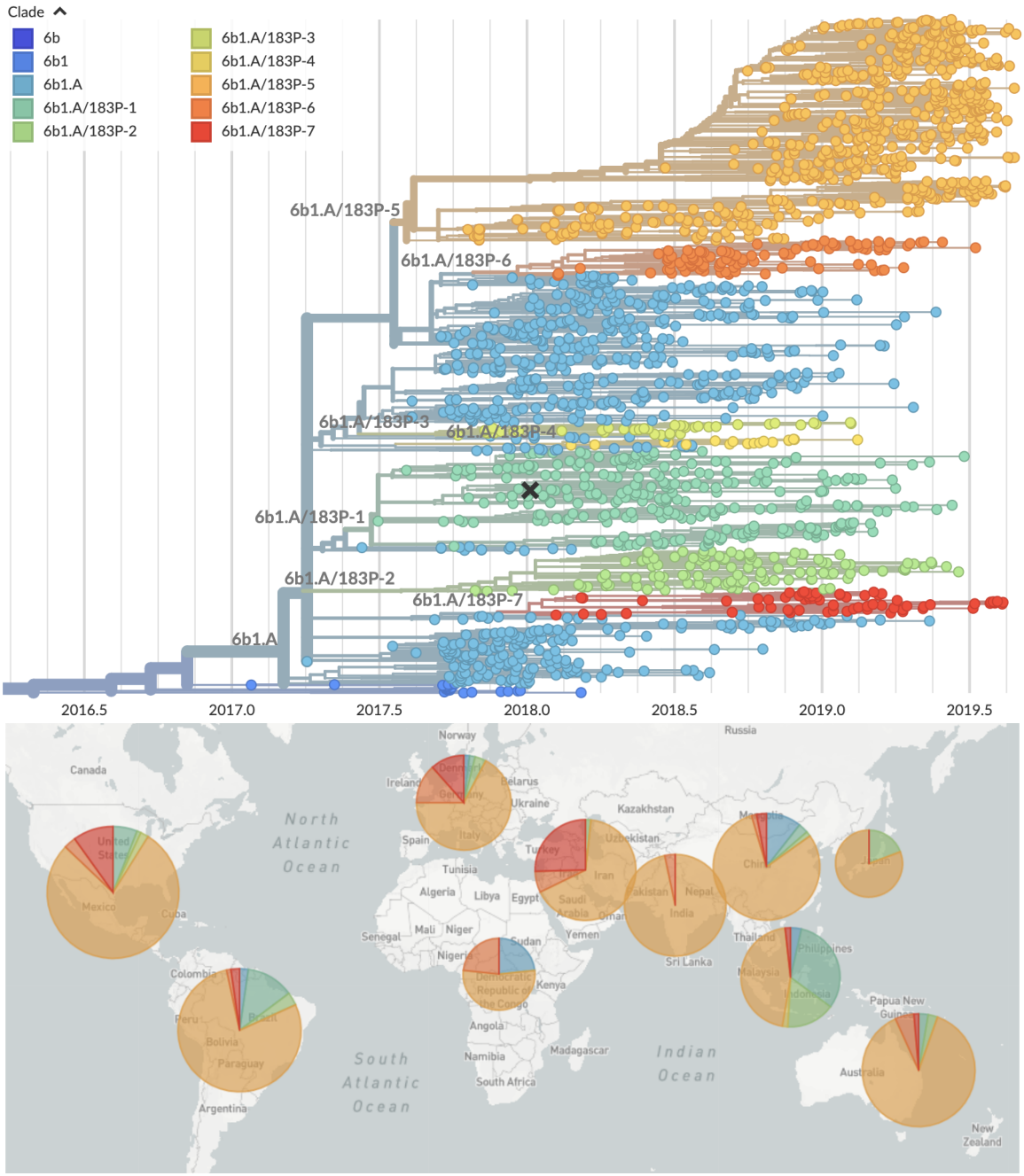

Figure 10. H1N1pdm phylogeny colored by clade and their geographic distribution (2019). This is a time resolved phylogeny.

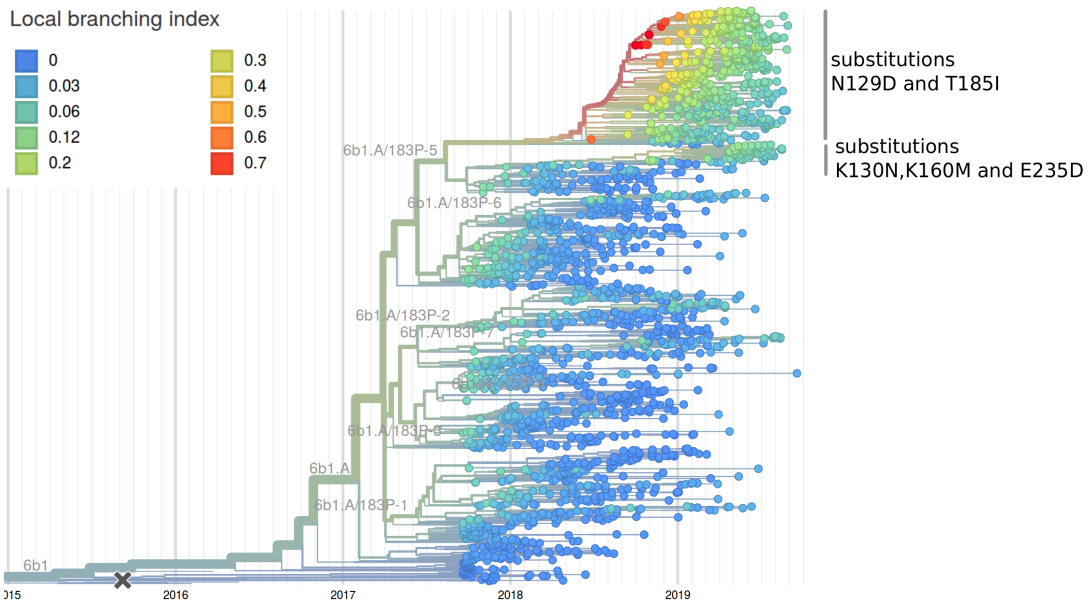

Figure 11. H1N1pdm phylogenies colored by LBI. The LBI accounts for both clade volume and recent expansion and has been predictive of clade success in the past. The larger subclade of P5 has the highest LBI (top), while the smaller subclade has grown rapidly in recent month. The latter might be due to an over-representation of North American sequences (see mutation HA1:130N in Fig. 13).

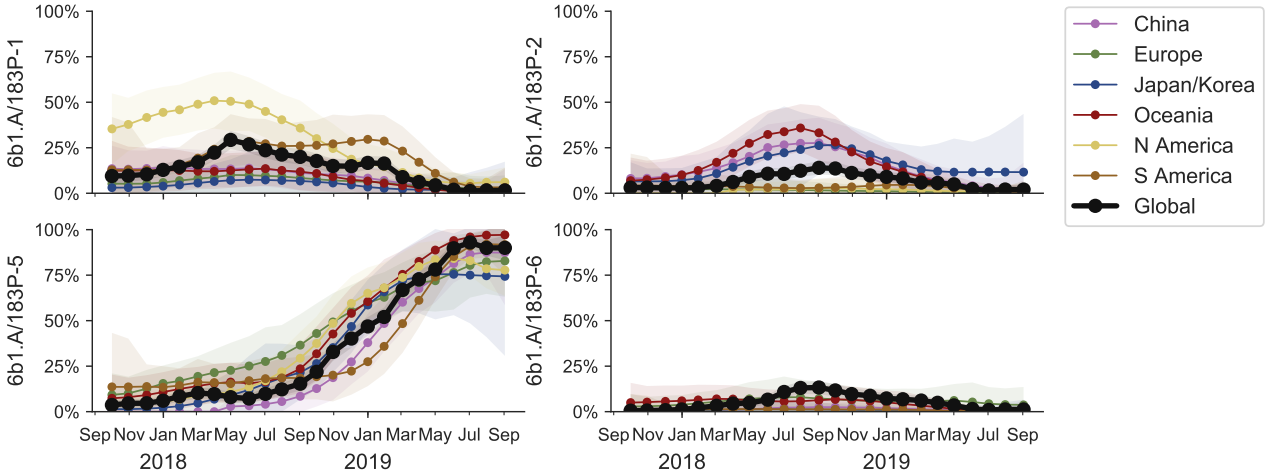

Figure 12. Frequency trajectories of H1N1pdm clades partitioned by clade and then by region. Clade P5 has essentially taken over, but in itself is deeply split with one subclade (substitution K130N) frequent in North America, the major subclade (T185I, N129D) prevalent elsewhere (see Figure 13). We estimate frequencies of different clades based on sample counts and collection dates of strains included in the phylogeny. We use a Brownian motion process prior to smooth frequencies from month-to-month. Transparent bands show an estimate of the 68% confidence interval based on sample counts.

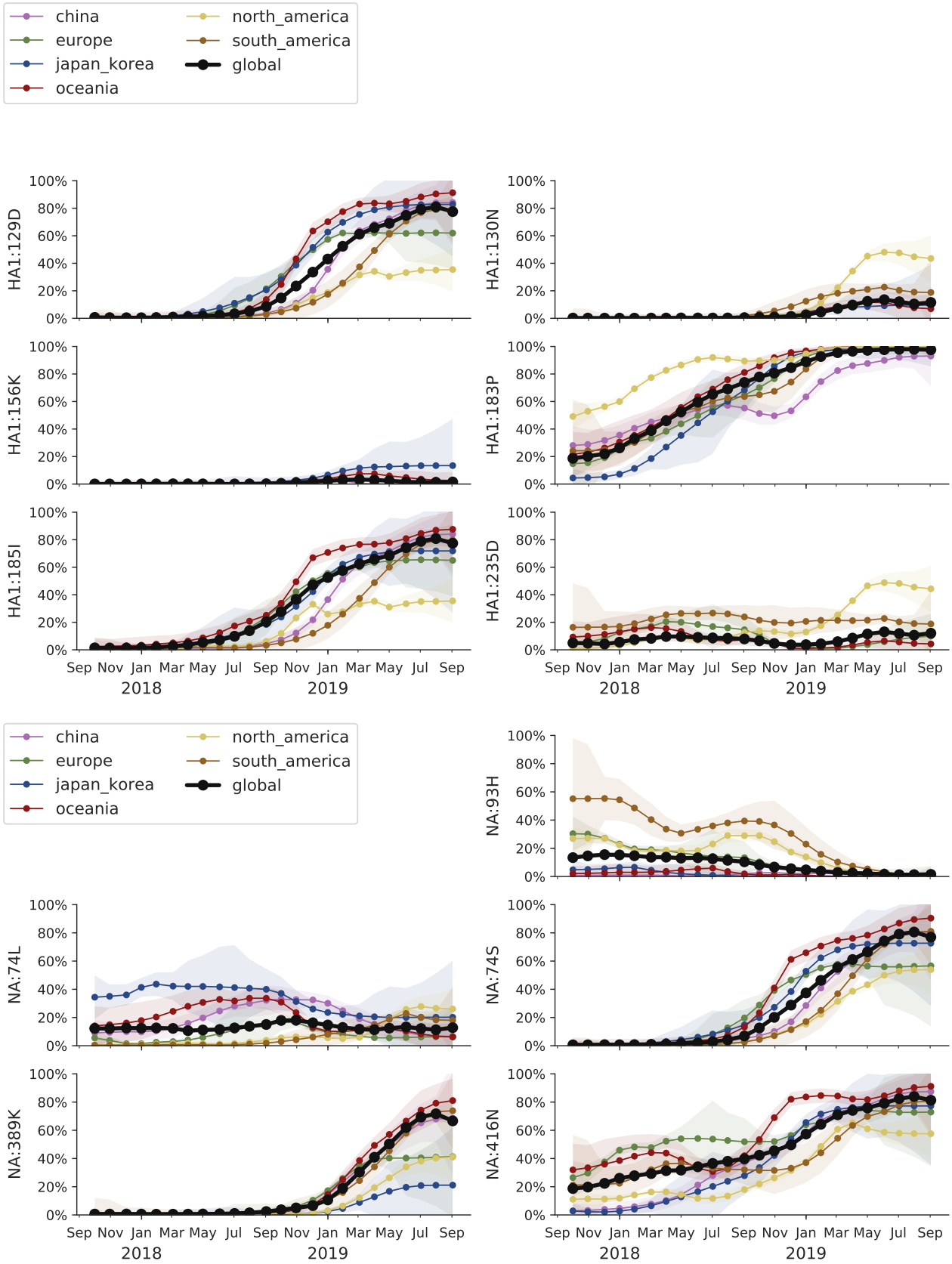

Figure 13. Frequency trajectories of mutations in H1N1pdm HA and NA segments partitioned by region. We estimate frequencies of different amino acid variants based on sample counts and collection dates. These estimates are based on all available data. We use a Brownian motion process prior to smooth frequencies from month-to-month. Transparent bands show an estimate of the 68% confidence interval based on sample counts.

Despite the rapid replacement fixation of the 183P substitution and continued diversification of H1N1pdm viruses at the HA protein level, very little antigenic evolution is apparent in HI assays. The only clear antigenic effect is associated with substitutions N156K and N156D, which result in a 16-fold titer drop. These substitutions have arisen multiple times, but are circulating at low levels and have not increased over the past 6 months. Most of the viruses with the N156K mutation fall into clade P2, most N156D variants are observed in the minor subclade of P5.

B/Vic

Antigenically drifted deletion variants at HA1 sites 162, 163 and 164 are now dominating global circulation and have all but taken over. The double deletion variant V1A.1 had previously been circulating at high frequency in the Americas. However, over the course of 2009, the triple deletion variant V1A.3 has increased in frequency globally and is now dominating in all geographic regions. Importantly, V1A.1 and V1A.3 variants appear antigenically distinct by HI assays with 4-8 fold reductions in log2 titer in both directions.

We base our primary analysis on a set of viruses collected between Aug 2017 and Aug 2019, comprising upwards of 100 viruses per month in Nov 2018 to June 2019 (Fig. 14). Recent months are dominated by data from China and North America.

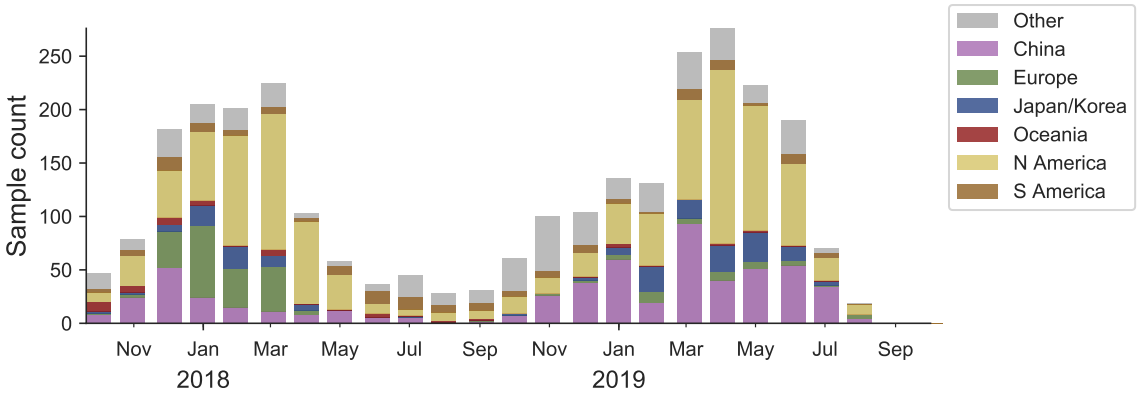

Figure 14. Sample counts through time and across regions. This is a stacked bar plot with the visible height of a color bar corresponding to the sample count from the respective region.

Circulating B/Vic viruses are primarily characterized be the parallel emergence of deletion variants at HA1 sites 162, 163 and 164 (Fig. 15). We label these as clades:

- Clade V1A.1: Deletions at 162 and 163. HA1 I180V. HA1 R152K.
- Clade V1A.2: Deletions at 162, 163 and 164. HA1 I180T, K209N.
- Clade V1A.3: Deletions at 162, 163 and 164. HA1 K136E.

Clades V1A.1 and V1A.3 have been successful with almost all 2019 viruses belonging to one or the other and V1A.3 dominating in most regions (Fig. 15).

The double deletion clade V1A.1 had risen to moderate frequency in 2018, but displayed heterogeneous frequencies across regions (Fig. 16). This clade V1A.1 is now waning in frequency. The triple deletion clade V1A.3 was initially confined to Africa, then but has rapidly in frequency in 2019, first in China and then spreading globally. Within this clade, a subclade with substitution G133R accounts for 60% of global circulation with an increasing trend (Fig. 17).

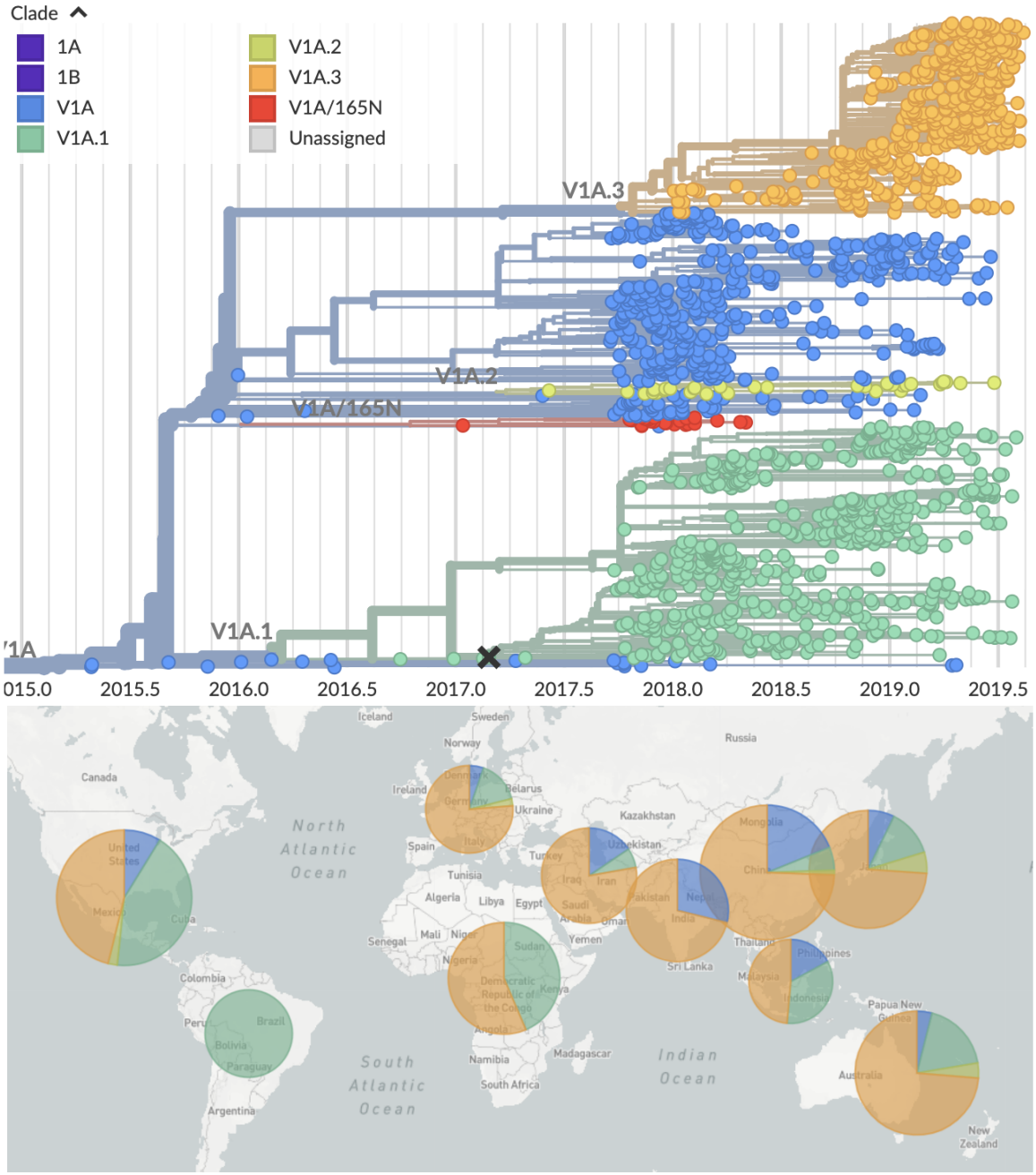

Figure 15. Vic phylogeny colored by clade and their geographic distribution (2019). This is a time resolved phylogeny.

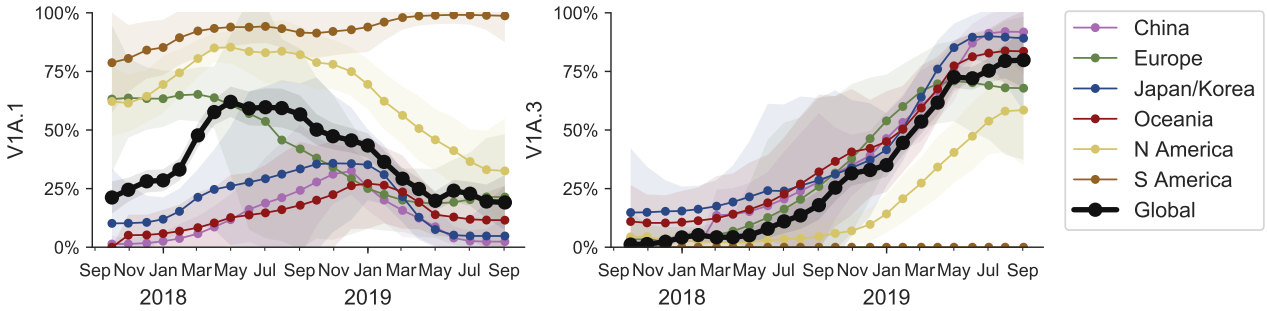

Figure 16. Frequency trajectories of Vic clades partitioned by clade and then by region. We estimate frequencies of different clades based on sample counts and collection dates of strains included in the phylogeny. We use a Brownian motion process prior to smooth frequencies from month-to-month. Transparent bands show an estimate of the 68% confidence interval based on sample counts.

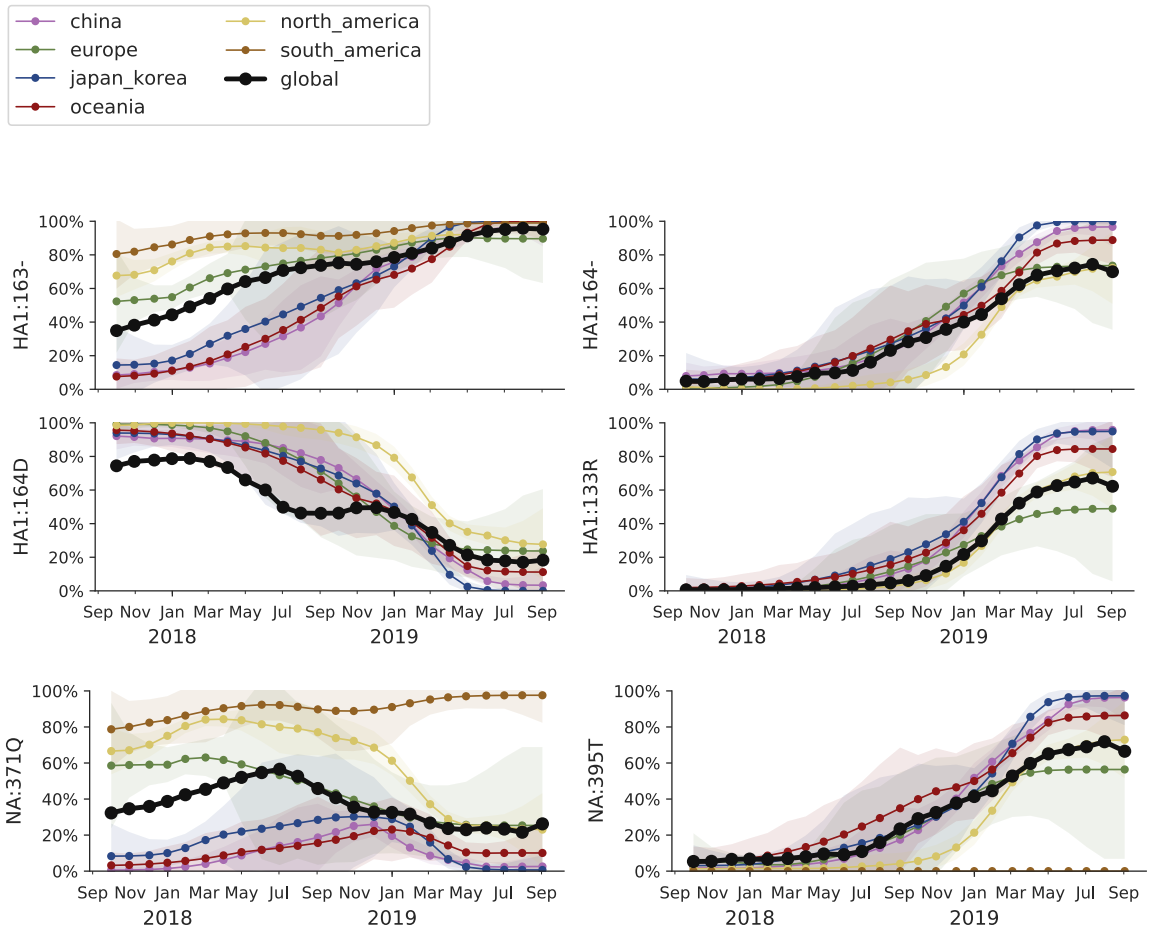

Figure 17. Frequency trajectories of B/Vic mutations partitioned by clade then by region. We estimate frequencies of different amino acid variants based on sample counts and collection dates. These estimates are based on all available data. We use a Brownian motion process prior to smooth frequencies from month-to-month. Transparent bands show an estimate of the 68% confidence interval based on sample counts.

HI titer data using sera raised against cell-passaged B/Colorado/6/2017 (current vaccine strain) and

B/Nigeria/3352/2018 are shown on the HA phylogeny in Figure 18. Viruses from triple deletion V1A.3 clade have about 4- to 16-fold reduced titers against B/Colorado/6/2017. Analogously, viruses from double deletion clade V1A.1 are not well recognized by sera raised against B/Nigeria/3352/2018.

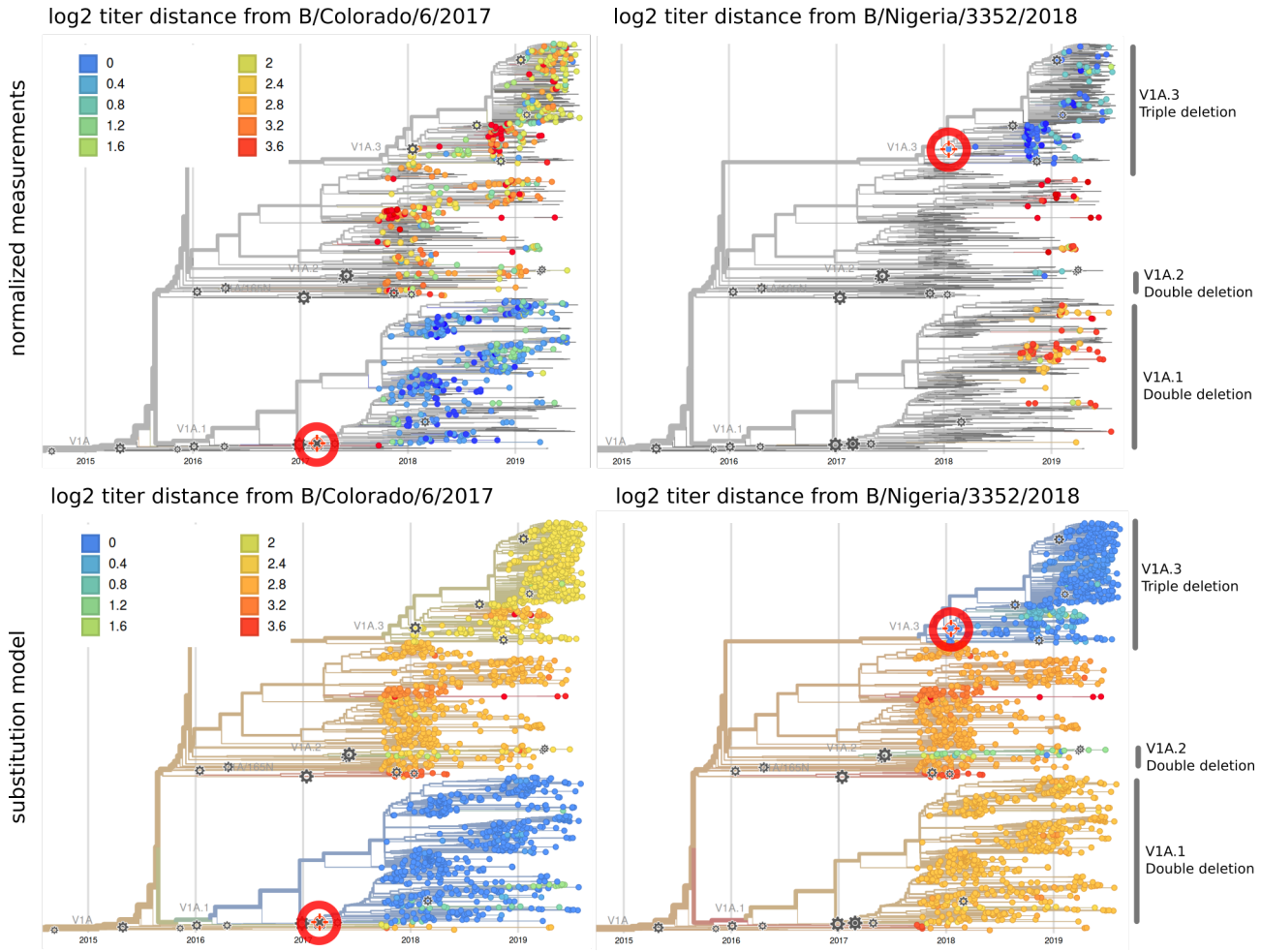

Figure 18. Vic phylogeny colored by antigenic distance from B/Colorado/6/2017 (left, representing V1A.1) and B/Nigeria/3352/2018 (right, representing V1A.3). While B/Colorado/6/2017 covers clade V1A.1 very well; the V1A.3 triple deletion variant (yellow clade) has about 4 to 16-fold reduced titers while isolates without a deletion have 8-fold reduced titers. B/Nigeria/3352/2018 covers both triple deletion variants well, but other viruses show 4- to 16-fold reduced titers.

B/Yam

B/Yam has not circulated in large numbers since the Northern Hemisphere season 2017/2018 and displays relatively little amino acid variation in HA or antigenic diversity. Amino acid variants at sites 229 and 232 have begun to circulate and population is now split between 229D/232D, 229N/232D and 229D/232N variants. These variants show little sign of antigenic difference in HI assays.

We base our primary analysis on a set of viruses collected between Aug 2017 and Jul 2019, comprising between 50 and 1100 viruses per month (Fig. 19), although there are few samples after May 2018. We use all available data when estimating frequencies of mutations and weight samples appropriately by regional population size and relative sampling intensity to arrive at a putatively unbiased global frequency estimate. Phylogenetic analyses are based on a representative sample of about 2000 viruses.

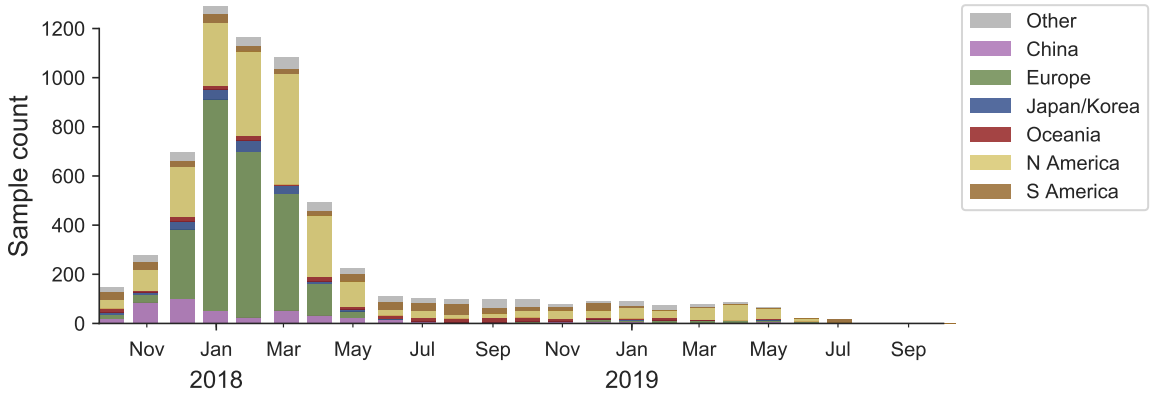

Figure 19. Sample counts through time and across regions. This is a stacked bar plot with the visible height of a color bar corresponding to the sample count from the respective region.

We observe very little variation among the HA segments of B/Yam viruses. Two substitutions, 229N and 232N, have started rising in frequency and are now at about 10-20% globally (Fig. 21).

The NA segment of B/Yam had undergone a series of rapid substitutions between 2014 and 2017 but has recently become more stable (Fig. 21). The mutation 342K rose from low frequency in 2017 close to fixation. Within this clade, the nested mutations 65H and S402P are globally at about 65% frequency. Outside of this clade, A395S is common in North America and Europe (Fig. 21). Site 342 has changed twice in recent years: D342N followed by N342K.

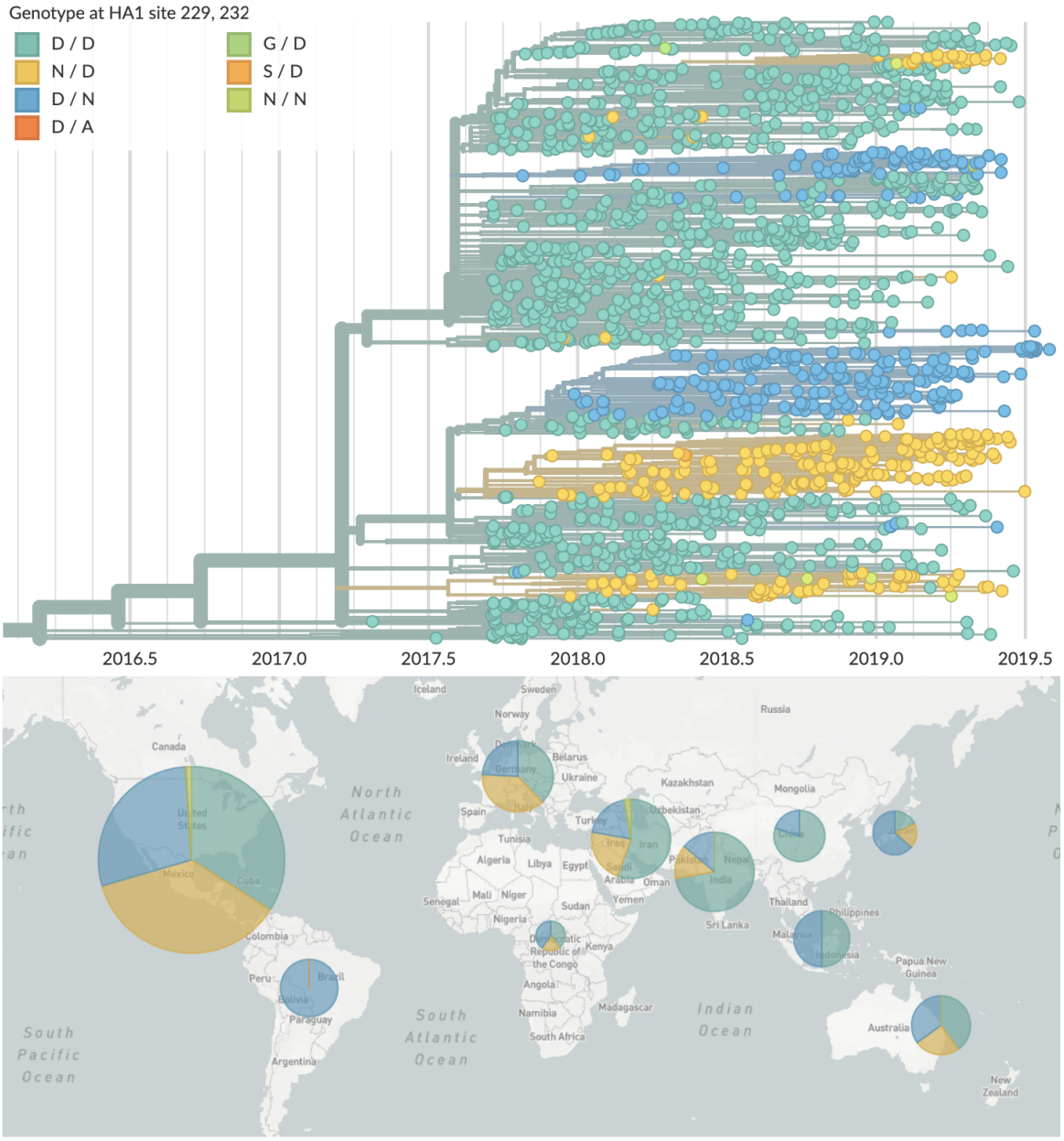

Figure 20. Yam phylogeny colored by mutations at sites 229 and 232 and their geographic distribution (2019). This is a time resolved phylogeny.

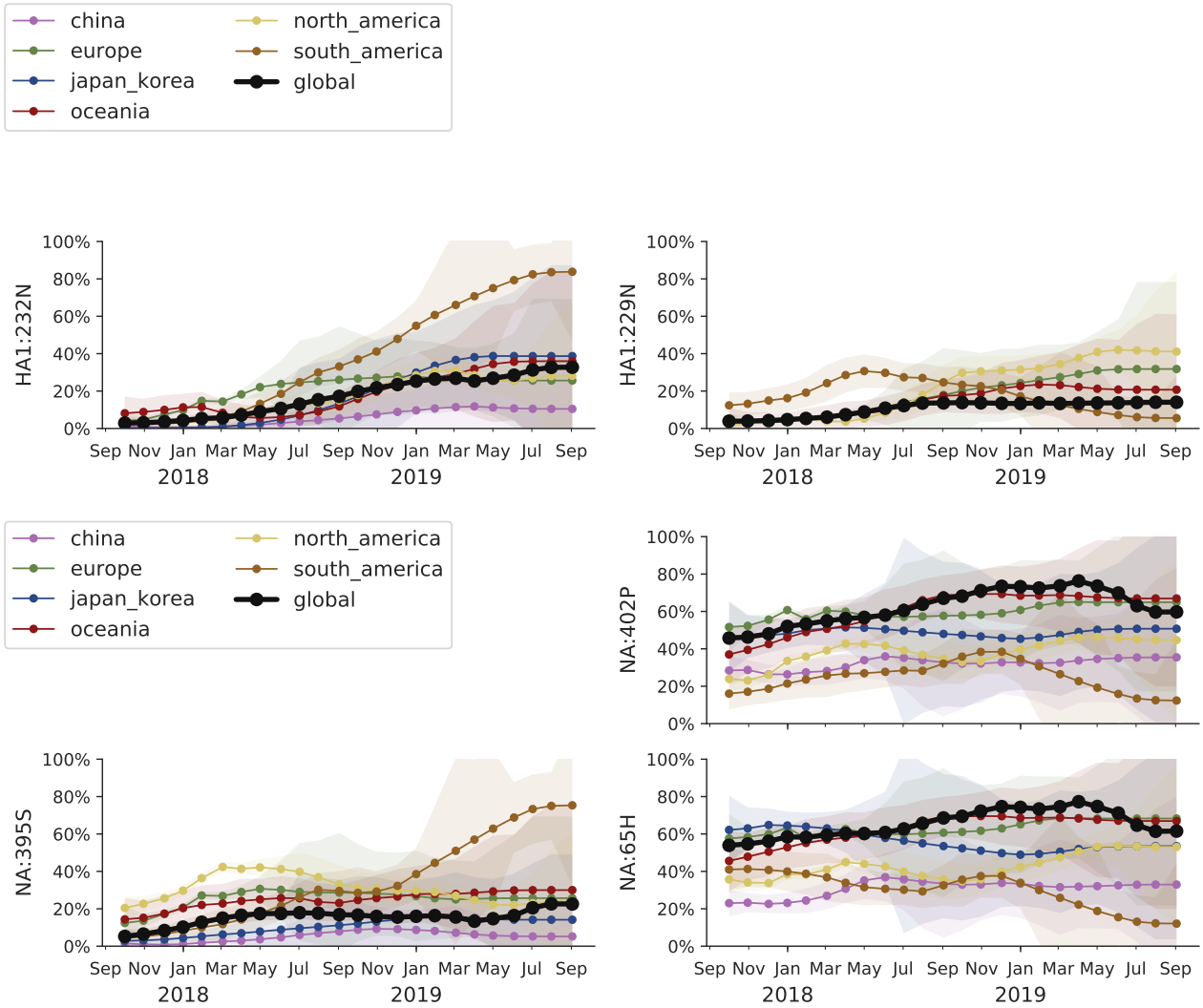

Figure 21. Frequency trajectories of recent mutations in B/Yam viruses. We estimate frequencies of different amino acid variants based on sample counts and collection dates. These estimates are based on all available data. We use a Brownian motion process prior to smooth frequencies from month-to-month. Transparent bands show an estimate of the 68% confidence interval based on sample counts.The almost complete lack of sequence data result in large uncertainty in these frequency estimates.

Acknowledgments

We thank the Influenza Division at the US Centers for Disease Control and Prevention, the Victorian Infectious Diseases Reference Laboratory at the Australian Peter Doherty Institute for Infection and Immunity, the Influenza Virus Research Center at the Japan National Institute of Infectious Diseases, the Crick Worldwide Influenza Centre at the UK Francis Crick Institute for data sharing and feedback. We thank David Wentworth, Rebecca Kondor and Vivien Dugan for insight regarding analysis directions.

References

1. Neher RA, Bedford T (2015) nextflu: real-time tracking of seasonal influenza virus evolution in humans. Bioinformatics 31: 3546–3548.
2. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, et al. (2018) Nextstrain: real-time tracking of pathogen evolution. Bioinformatics 34: 4121–4123.
3. Neher RA, Bedford T, Daniels RS, Russell CA, Shraiman BI (2016) Prediction, dynamics, and visualization of antigenic phenotypes of seasonal influenza viruses. Proc Natl Acad Sci USA 113: E1701–E1709.
4. Neher RA, Russell CA, Shraiman BI (2014) Predicting evolution from the shape of genealogical trees. eLife 3: e03568.

Host age distributions

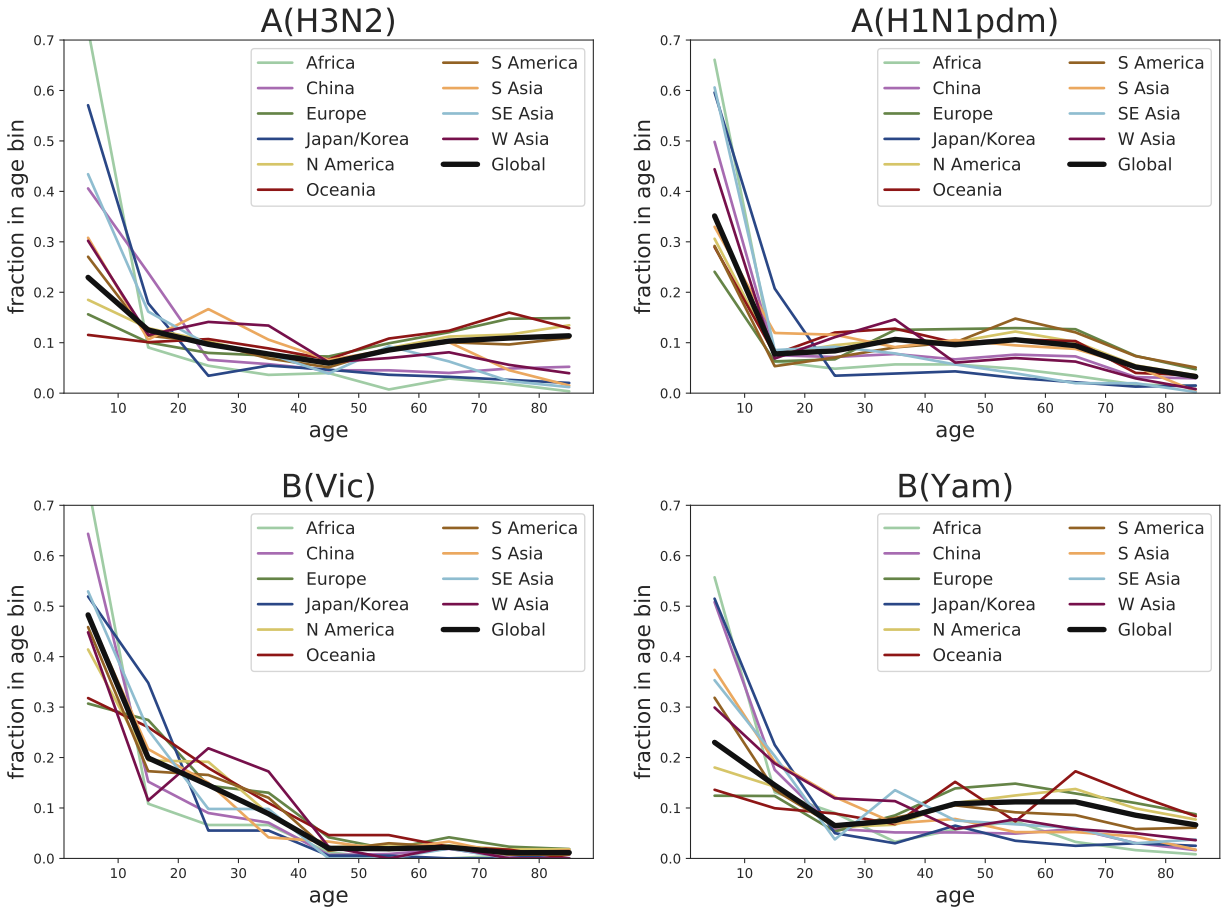

Figure 22. Each panel show the fraction of sequences sampled from hosts in a 10 year age bin for different geographic regions. These distributions are confounded by reporting bias, but the differences between different lineages are big enough that the overall picture is unlikely changed by these biases. As expected, H1N1pdm sequences are more frequently sampled from young children compared to H3N2. The biggest difference is observed between the two influenza B lineages: B/Vic viruses were mostly sampled from young individuals, in sharp contrast to B/Yam which is frequently isolated from patients above 40y of age, in particular in Europe, North America, and Oceania.