Reassortment between influenza B lineages and the emergence of a co-adapted PB1-PB2-HA gene complex

Gytis Dudas[1], **Trevor Bedford**[2], **Samantha Lycett**[1,3] **& Andrew Rambaut**[1,4,5]

[1]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK, [2]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA, [3]Institute of Biodiversity Animal Health and Comparative Medicine, University of Glasgow, Glasgow, UK, [4]Fogarty International Center, National Institutes of Health, Bethesda, MD, USA, [5]Centre for Immunology, Infection and Evolution at the University of Edinburgh, Edinburgh, UK

March 6, 2014

Abstract

Influenza B viruses are increasingly being recognized as major contributors to morbidity attributed to seasonal influenza. Currently circulating influenza B isolates are known to belong to two antigenically distinct lineages referred to as B/Victoria and B/Yamagata. Frequent exchange of genomic segments of these two lineages has been noted in the past, but the observed patterns of reassortment have not been formalized in detail. We investigate inter-lineage reassortments by comparing phylogenetic trees across genomic segments. Our analyses indicate that of the 8 segments of influenza B viruses only PB1, PB2 and HA segments maintained separate Victoria and Yamagata lineages and that currently circulating strains possess PB1, PB2 and HA segments derived entirely from one or the other lineage; other segments have repeatedly reassorted between lineages thereby reducing genetic diversity. We argue that this difference between segments is due to selection against reassortant viruses with mixed lineage PB1, PB2 and HA segments. Given sufficient time and continued recruitment to the reassortment-isolated PB1-PB2-HA gene complex, we expect influenza B viruses to eventually undergo sympatric speciation.

Introduction

Seasonal influenza causes between 250,000 and 500,000 deaths annually and comprises lineages from three virus types (A, B and C) co-circulating in humans, of which influenza A is considered to cause the majority of seasonal morbidity and mortality [1]. However, influenza B viruses are increasingly being recognized as important human pathogens [2]. Following the 2009 A/H1N1 pandemic, influenza B has increased in prevalence and in the 2012/2013 European season as many as 53% of influenza sentinel surveillance samples tested positive for influenza B [3].

Like other members of *Orthomyxoviridae*, influenza B viruses have segmented genomes, which allow viruses co-infecting the same cell to exchange segments, a process known as reassortment. Influenza A viruses are widely considered to be a major threat to human health worldwide due to their ability to cause pandemics in humans via reassortment of circulating human strains with non-human influenza A strains. Although influenza B viruses have been observed to infect seals [4, 5] through a reverse zoonosis, they are thought to primarily infect humans and are thus are unlikely to exhibit pandemics due to the absence of an animal reservoir from which to acquire antigenic novelty. Both influenza A and B evolve antigenically through time in a process known as antigenic drift, in which mutations to the haemagglutinin (HA) protein allow viruses to escape existing human immunity and persist in the human population, leading to recurrent seasonal epidemics [6–8].

Currently circulating influenza B viruses comprise two distinct lineages – Victoria and Yamagata (referred to as Vic and Yam, respectively) – named after strains B/Victoria/2/87 and B/Yamagata/16/88, that are thought to have genetically diverged in HA around 1983 [9]. These two lineages now possess antigenically distinct HA surface glycoproteins [9–13] allowing them to co-circulate in the human population. Phylogenetic analysis of evolutionary rate, selective pressures and reassortment history of influenza B has shown extensive and often complicated patterns of reassortment between all segments of influenza B viruses both between and within the Vic and Yam lineages [14].

Here, we extend previous methods to reveal an evolutionarily intriguing pattern of reassortment in influenza B. In our approach, membership to either the Victoria or Yamagata lineage in the tree of one segment is used to label the individual viruses in the tree of the other segments. By modelling the transition between labels on a phylogenetic tree, reassortment events which result in the replacement of one segment's lineage by another show up as label changes along a branch (Figure 1). We use this method to reconstruct major reassortment events and quantify reassortment dynamics over time in influenza B viruses.

We show that despite extensive reassortment, three of the eight segments – two segments coding for components of the influenza B virus polymerase, PB1 and PB1, and the surface glycoprotein HA – still survive as distinct Victoria and Yamagata lineages, which appear to be co-adapted to the point where virions which do not contain PB1, PB2 or HA segments derived entirely from either the Vic or the Yam lineage have rarely been isolated and only circulate as transient lineages once isolated. In other segments (PA, NP, NA, MP

and NS) a single lineage has introgressed into the opposing background and been fixed in the influenza B population: Yam for PA, NP, NA and MP and Vic for NS. This has occurred through repeated reassortments and subsequent fixation of reassortant genome constellations within the influenza B population.

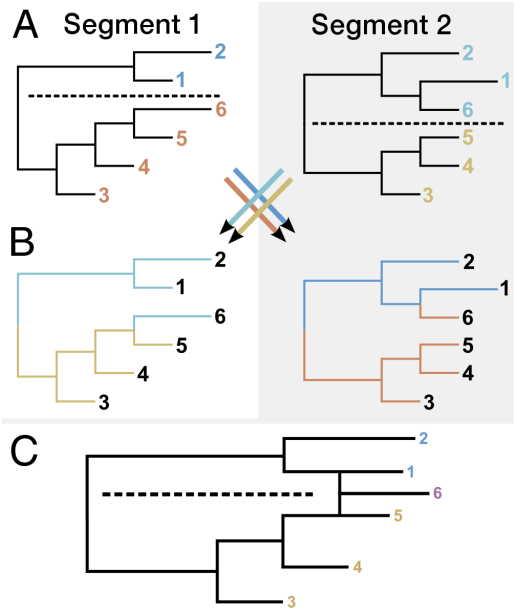

Figure 1. Schematic analysis of reassortment patterns. A) We begin by assigning sequences falling on either side of a specified bifurcation within each segment tree to different lineages, in this case, the Victoria and Yamagata bifurcation that occurred in the early 1980s. B) We then transfer lineage labels from one tree to the same tips in another tree. Transitions between labels along this second tree thus indicate reassortment events that combine lineages falling on different sides of the Vic/Yam bifurcation in the first tree. C) A reassortment graph depiction shows that tip number 6 is determined to be a reassortant based on B).

Results

Analysis of reassortment patterns across Victoria and Yamagata lineages

The differentiation into Vic and Yam lineages can be seen in all segments (Figure 2). Following the split of the two lineages, each segment can be assigned to either Vic or Yam lineage and inter-lineage reassortment events have yielded mixed-lineage genome constellations. On some segments, either the Victoria or Yamagata lineage has become fixed in the influenza B virus population, i.e. became the 'trunk' of the phylogenetic tree of a segment, seen as modern viruses deriving completely from either Victoria or Yamagata lineage (yellow vs purple bars in Figure 2). This pattern is apparent in the PA, NP, NA, MP and NS segments. However, the PB1, PB2 and HA segments of modern viruses are derived from both Victoria and Yamagata lineages. Consistent with fixation of Victoria or Yamagata lineages, the PA, NP, NA, MP and NS segments periodically lose diversity, while

maintenance of parallel Victoria and Yamagata lineages results in continually increasing diversity in segments PB1, PB2 and HA (Figure 3). The PB1, PB2 and HA segments from present-day viruses maintain a common ancestor in ∼1983 and thus accumulate genetic diversity since the split of those segments into Vic and Yam lineages, while other segments often lose diversity with ancestors to present-day viruses appearing between ∼1991 and ∼1999.

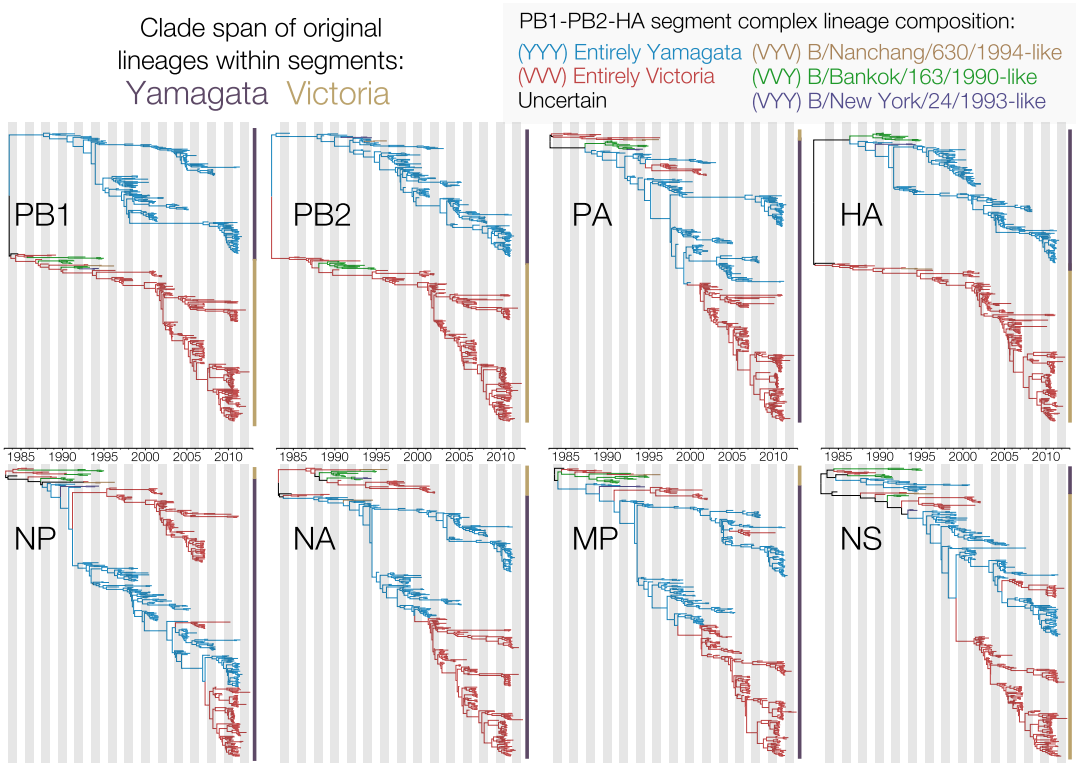

Figure 2. Maximum clade credibility (MCC) trees of all 8 genome segments of influenza B viruses isolated since 1980. Trees are coloured based on inferred PB1-PB2-HA lineage. Vertical bars indicate the original Victoria and Yamagata lineages within each segment. Each tree corresponds to the summarised output from single analyses comprised of 9000 trees sampled from the posterior distribution of trees.

By measuring mean pairwise diversity between branches in each tree that were assigned either a Vic or Yam label in other segments, we look for reductions in between-lineage diversity, which indicate that an inter-lineage reassortment event has taken place (Figure 4). Because we are taking the mean of pairwise comparisons, this method gives a quantitative measure of reassortment-induced loss of diversity between Victoria and Yamagata lineages in two trees, although care should be taken when interpreting the statistic, as it does not correspond to any real TMRCAs in the tree, but instead gives the TMRCA of pairs of viruses. We focus only on PB1, PB2 and HA lineage labels, since all other segments have fixed either the Vic or the Yam lineage. Losses of diversity (represented by more recent mean pairwise TMRCAs between Vic and Yam labels) in Figure 4 indicate that every segment has reassorted with respect to the Victoria and Yamagata lineages of PB1, PB2 and HA segments. However, we also see that PB1, PB2 and HA labels show reciprocal

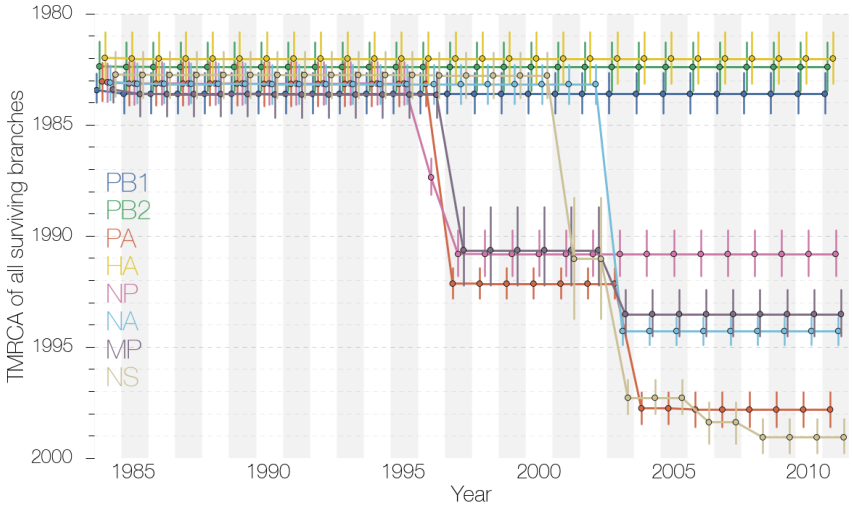

Figure 3. Oldest TMRCA of all surviving branches over time. PA, NP, NA, MP and NS segments of influenza B viruses show periodic losses of diversity, indicating lineage turnover. PB1, PB2 and HA segments, on the other hand, maintain the diversity dating back to the initial split of Vic and Yam lineages. Each point is the mean time of most recent common ancestor (TMRCA) of all surviving lineages existing at each time slice through the tree and vertical lines indicating uncertainty are 95% highest posterior densities (HPDs).

preservation of diversity after 1997. This suggests that after 1997 no reassortment events have taken place between Victoria and Yamagata lineages of PB1, PB2 and HA segments and their lineage labels only 'meet' at the root. We do see reduced diversity between Vic and Yam labels of PB1, PB2 and HA segments in a time period close to the initial split of Vic and Yam lineages (1986–1996). These reductions in diversity represent small clades with reassortant PB1-PB2-HA constellations, which go extinct by 1997. We also observe that the assignment of Vic or Yam lineages of PB1, PB2 and HA segments to branches of other segment trees is very similar and often identical. This results in PB1, PB2 and HA lineage labels switching between Vic and Yam simultaneously in all trees after 1997, suggesting co-reassortment of Vic and Yam lineages of PB1, PB2 and HA segments.

By plotting the ratios of sequences deriving from either the Vic or Yam lineage in each segment (Figure 5) it is evident that losses of diversity in the PA, NP, NA, MP and NS segments are related to the repeated fixation of either the Vic or the Yam lineage. These losses of diversity correspond to fixation of the Vic lineage in NS and fixation of the Yam lineage in PA, NP, NA and MP. Similarly, the lack of reassortment between Vic and Yam lineages and maintenance of diversity of PB1, PB2 and HA can be seen, where the two lineages have been isolated at a ratio close to 50% over long periods of time (Figure 5). On a year-to-year basis, however, the ratios for Vic and Yam lineage PB1, PB2 and HA can fluctuate dramatically consistent with one lineage predominating within a given season, in agreement with surveillance data [15].

We reconstructed reassortment events that were detected by using lineage labels. Figure 6 focuses only on inter-lineage reassortments that have occurred after 1990. We identify 5 major reassortant genome constellations (given in order PB1-PB2-PA-HA-NP-NA-MP-NS

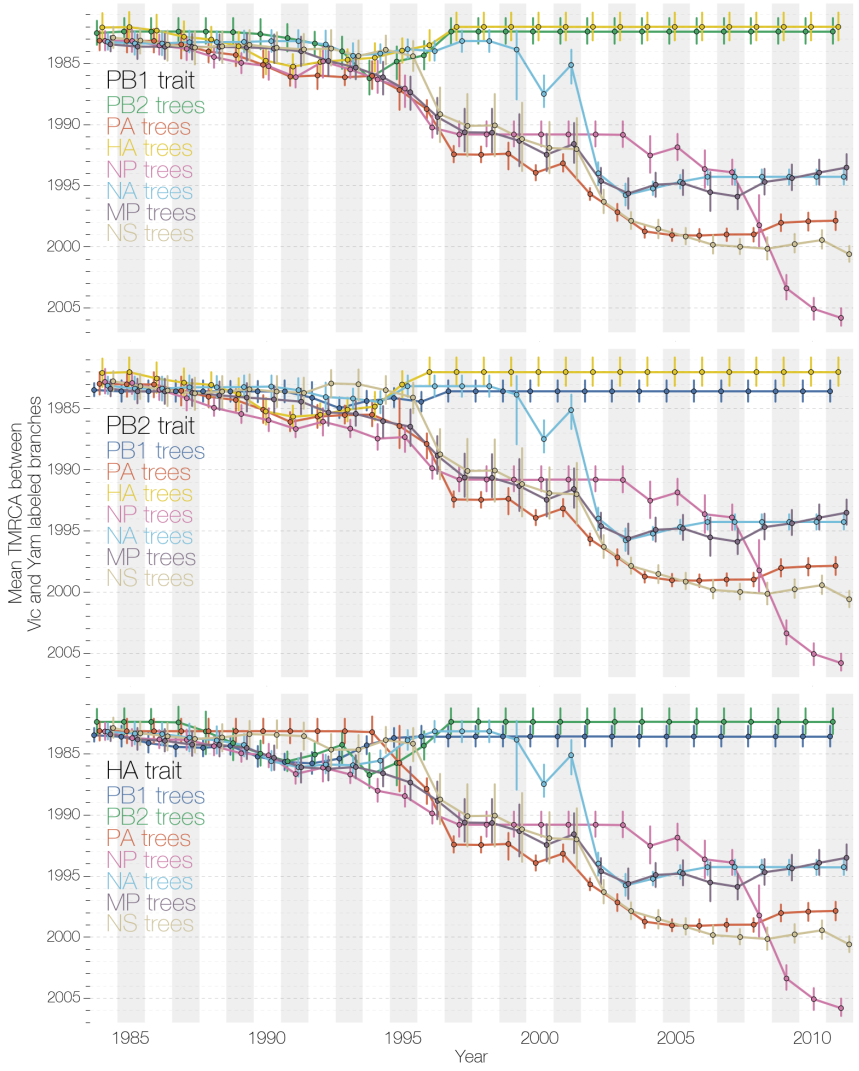

Figure 4. Mean pairwise TMRCA between Vic and Yam branches under PB1, PB2 and HA label sets. PB1, PB2 and HA segment labels indicate that these segments show reciprocal preservation of diversity, which dates back to the split of Vic and Yam lineages. All other segments show increasingly more recent TMRCAs between branches labelled as Vic and Yam in PB1, PB2 and HA label sets. All vertical lines indicating uncertainty are 95% highest posterior densities (HPDs).

with prime (') indicating independently acquired segments) circulating between 1992 and 2011 (Figure 6):

- B/Alaska/12/1996-like (Y-Y-Y-Y-Y-Y-Y-V)
- B/Nanchang/2/1997-like (V-V-Y-V-Y-V-Y-V)
- B/Iowa/03/2002-like (V-V-Y'-V-Y-Y-Y'-V')
- B/California/NHRC0001/2006-like (V-V-Y-V-Y'-Y-Y'-V')

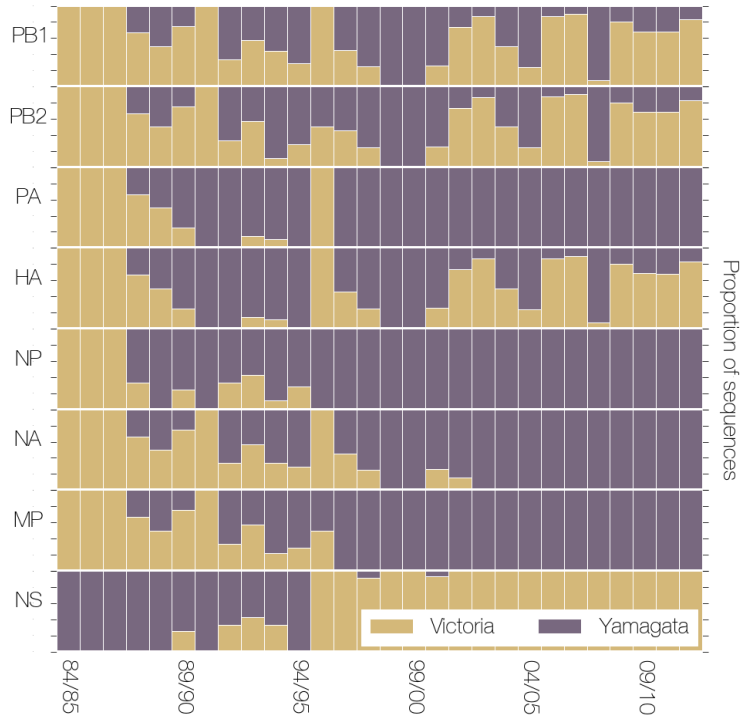

Figure 5. Ratio of lineages in the dataset. The ratio of sequences that derive from the original Victoria clade (yellow) to sequences that derive from the Yamagata clade (purple) in each segment over time. Yamagata lineage PA, NP, NA and MP segments and Victoria lineage NS segment eventually become fixed in the influenza B population. PB1, PB2 and HA segments maintain separate Victoria and Yamagata lineages.

- B/Brisbane/33/2008-like (V-V-Y-V-Y'-Y-Y-V)

Of these 5 constellations 4 (B/Nanchang/2/1997-like, B/Iowa/03/2002-like, B/California/NHRC0001/2006-like and B/Brisbane/33/2008-like) are derived from introgression of Yamagata lineage segments into Victoria lineage PB1-PB2-HA background, with only 1 (B/Alaska/12/1996-like) resulting from introgression of Victoria lineage NS segment into an entirely Yamagata derived background. All 5 inter-lineage reassortment events described here are marked by the preservation of either entirely Victoria or Yamagata derived PB1-PB2-HA segments. Figure 6 also shows that reassorting segments appear to evolve with a considerable degree of autonomy. For example, the NP lineage that entered a largely Victoria lineage derived genome and gave rise to the B/Nanchang/2/1997-like isolates continued circulating until 2010, even though other segments it reassorted with in 1995 – 1996 (PA and MP) went extinct following the reassortments that led to the rise of viruses with B/Iowa/03/2002-like genome constellations. A more extreme example is the NS segment, which in B/Iowa/03/2002-like isolates (and all subsequent Vic PB1-PB2-HA isolates) has originally been derived from the Victoria lineage that had been associated with mostly Yam lineage derived B/Alaska/12/1996-like genomes for a number of years. Overall, we see that successful inter-lineage reassortment events do not break up Vic and Yam PB1-PB2-HA complexes, but any other segments reassorted into one or the other PB1-PB2-HA background evolve with a considerable degree of autonomy.

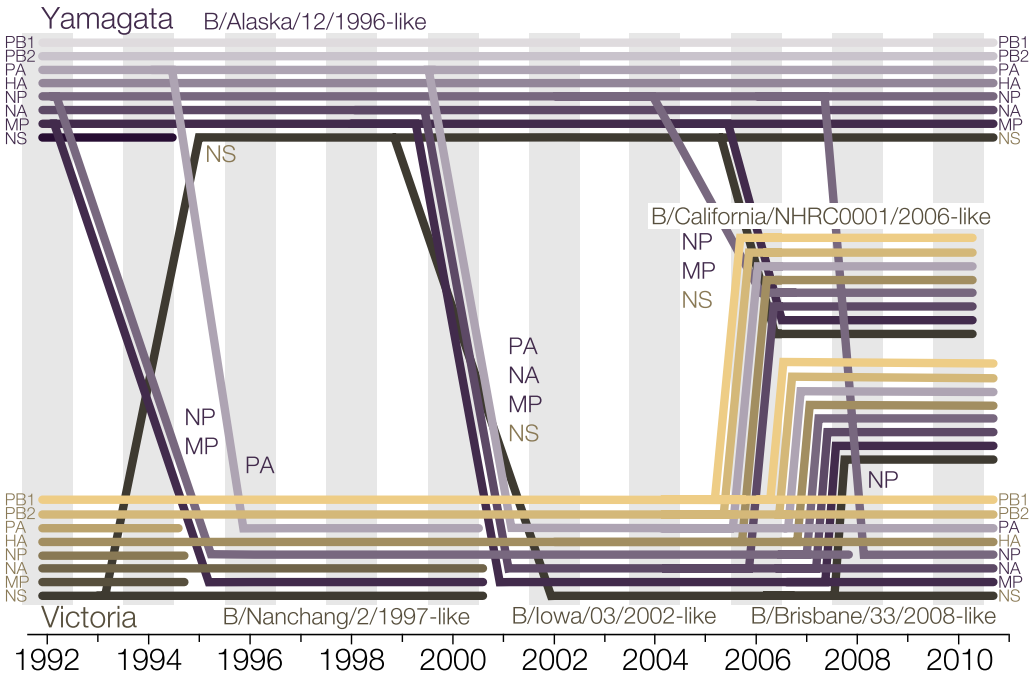

Figure 6. Schematic plot of reconstructed reassortments between Victoria and Yamagata lineage segments of influenza B virus. Lineages that coassort in genomes are represented by 8 parallel lines, with lineages that derive from the original Victoria clade colored yellow/brown and lineages that derive from the original Yamagata clade colored lilac/purple. Inter-lineage reassortment events are indicated by lines entering a different genome. The angle of incoming lineages represents uncertainty in the timing of the event (mean date of the reassortant node and its parent node). Lineage extinction dates are not shown accurately.

The vast majority of influenza B isolates possess either Vic or Yam lineage derived PB1-PB2-HA complexes, although on rare occasions mixed-lineage PB1-PB2-HA constellations emerge. Figure 7 shows the sum of branch lengths which were labelled as having entirely Vic, entirely Yam or mixed-lineage PB1, PB2 and HA segments. Due to lack of reassortment between Vic and Yam lineages of PB1, PB2 and HA (Figure 4) since 1997 all segments have spent significantly longer periods of evolutionary time with either entirely Vic-derived or entirely Yam-derived than with mixed-lineage PB1, PB2 and HA constellations (Figure 7). We have identified 3 instances of mixed-lineage PB1-PB2-HA reassortants from the data with the following PB1-PB2-HA constellations: VVY (B/Bangkok/163/1990-like, 13 sequences isolated 1990 – 5 Jan 1995), VYV (B/Nanchang/630/1994-like, 2 sequences isolated 1994 – 1996) and VYY (B/New York/24/1993-like, 2 sequences isolated 8 Jan 1993 – 1994). An additional instance of reassortants with a YYV PB1-PB2-HA constellation (B/Waikato/6/2005-like, 16 sequences isolated 9 June – 12 November in 2005) was discovered when investigating the larger dataset with only PB1, PB2 and HA sequences. In all cases, PB1-PB2-HA reassortants have not persisted for prolonged periods of time and have not been fixed in the influenza B population. In particular reassortment events

combining PB1 and PB2 segments of different lineages, *e.g.* B/Nanchang/630/1994-like and B/New York/24/1993-like isolates (each represented by two isolates), exhibit poor sampling and short circulation times. In the case of B/Nanchang/630/1994-like viruses this is up to two years. In contrast, reassortments combining PB1+2 and HA of different lineages (B/Bangkok/163/1990-like and B/Waikato/6/2005-like isolates) have more isolates and circulated for much longer periods of time in the past: up to 4 years for B/Bangkok/163/1990-like, although the more recently isolated B/Waikato/6/2005-like viruses circulated for only 5 months.

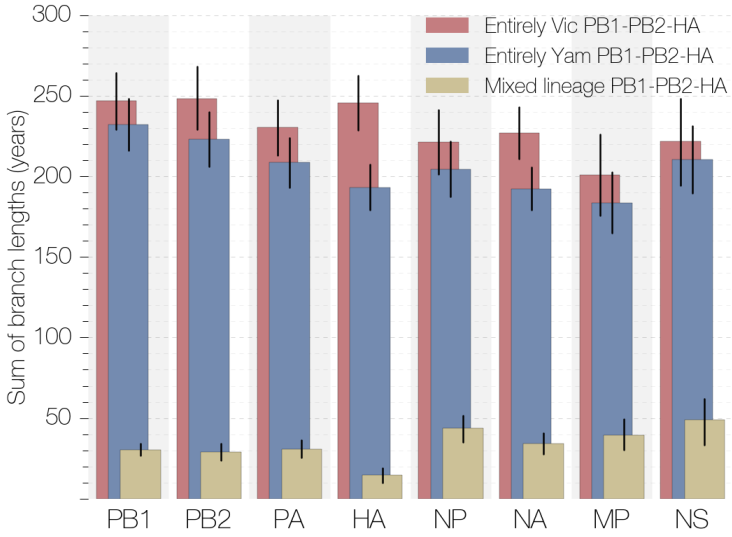

Figure 7. Amount of evolutionary time each segment has spent under different PB1-PB2-HA constellations. All segments have spent significantly more of their history with entirely Vic or entirely Yam-derived PB1-PB2-HA complexes. All horizontal lines indicating uncertainty are 95% highest posterior densities (HPDs).

Analysis of within-lineage reassortment patterns

SPR (subtree prune and regraft) provides a measure of tree-to-tree distance, in which distances between perfectly resolved pairs of trees would be equal to the number of reassortments that have taken place between them [16], as long as each branch experiences, at most, one reassortment. Figure 8 shows approximate SPR distances between all pairs of segment trees after normalization (see Methods). If there are biases in the way segments reassort, so that some segments tend to co-assort more often, we expect to observe a lower reassortment rate between them, which would manifest as small-scale similarities between phylogenetic trees of those segments. In our case we expect SPR distances, which are proportional to the number of reassortment events that have taken place between trees, to reflect the overall (*i.e* both within and between lineages) reassortment rate.

We normalized our SPR distance comparisons to account for differences in tree topology stability over the course of the MCMC chain and also use an approximation of SPR distance [17–19] (see Methods). The 95% highest posterior density (HPD) intervals of normalized approximate SPR distances between pairs of segments encompass most means and

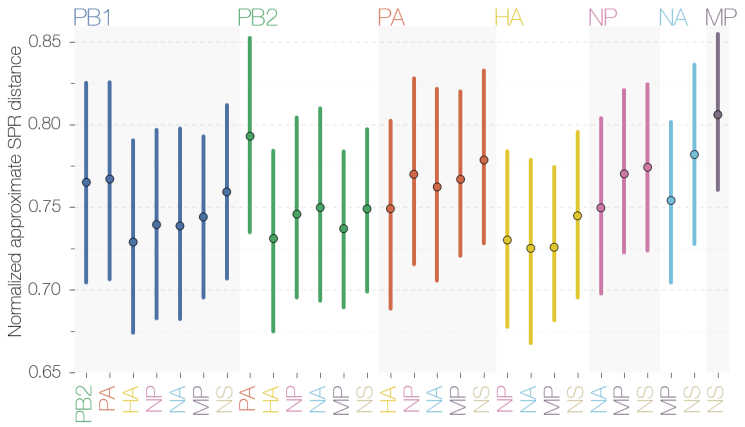

Figure 8. Normalized approximate SPR distances between pairs of segments. Following the normalization procedure approximate SPR distances are similar across all pairwise comparisons. We interpret this as lack of evidence for small-scale topological similarities between trees of all segments, which we expect to arise if any two segments were being co-packaged and co-reassorted. All vertical lines indicating uncertainty are 95% highest posterior densities (HPDs).

occupy a relatively small range, suggesting there is no evidence of differences in the number of reassortments between segments (Figure 8). Reassortment rate given as number of SPR moves per total time in both trees shows similar results (Figure S1). We note, however, that because of phylogenetic uncertainty our estimate of SPR distance might simply lack power. Comparisons between independent analyses of the same segments yield distances that are comparable to distances between different segments (Figures S2 and S3), suggesting that phylogenetic uncertainty is making a considerable contribution to our estimates of approximate SPR distances. Still, we find that independent replicates from the same segment (Figure S3) show lower SPR distances that comparisons between segments (Figure S2), suggesting that phylogenetic noise is not completely overwhelming reassortment signal. In addition, SPR distances themselves can only approximate (and underestimate) the actual numbers of reassortments. Thus we caution against over-interpreting Figure 8. Although there might be concern about using approximate, rather than exact, SPR distances we do estimate exact SPR distances for a limited number of segment pairs - PB1, PB2 and HA - and find that after normalization exact and approximate SPR distances are not significantly different (Figures S4–S6).

Although we do not find evidence of differences in numbers of overall reassortments between segments, we find support for a reassortment 'distance' effect, in which a pair of tips on one segment has a different TMRCA from the same pair of tips on a different segment. Unlike SPR distances, which are counts of the numbers of reassortment events, Δ_{TMRCA} takes time into account, and is most sensitive when only one of the two trees being compared loses diversity via reassortment. Similarly to SPR distances, we normalized our Δ_{TMRCA} comparisons to account for node date stability over the course of the MCMC chain (see Methods). Figure 9 shows normalized mean Δ_{TMRCA} values for all pairs of trees. Most segment pairs show very low values for this statistic with $\Delta_{\mathrm{TMRCA}} \approx 0.1$, indicating that Δ_{TMRCA} measurements between independent analyses of the same segment are up to 10 times smaller than Δ_{TMRCA} values between analyses of different segments.

PB1, PB2 and HA trees, on the other hand, exhibit normalized Δ_{TMRCA} values that are much higher. This shows that TMRCA differences between trees of PB1, PB2 and HA segments which, though noisy, are occasionally very similar to uncertainty in tip-to-tip TMRCAs between replicate analyses of these segments. For example, the distribution of comparison between PB1 and PB2 trees has a mean $\Delta_{\text{TMRCA}} \approx 0.4$, indicating that the sum of Δ_{TMRCA} values from PB1-PB1' and PB2-PB2' comparisons (prime (') indicates an independent MCMC chain) are 40% of twice the Δ_{TMRCA} value from the PB1-PB2 comparison, but this value can range anywhere between ≈0.15 and ≈0.75. Overall, it suggests that PB1, PB2 and HA lineages tend to not reassort among themselves unless both reassorting segments have similar TMRCAs, i.e. we observe isolation by distance. In addition, we see evidence of a similar reassortment 'distance' effect in MP and NA trees in Figure 9.

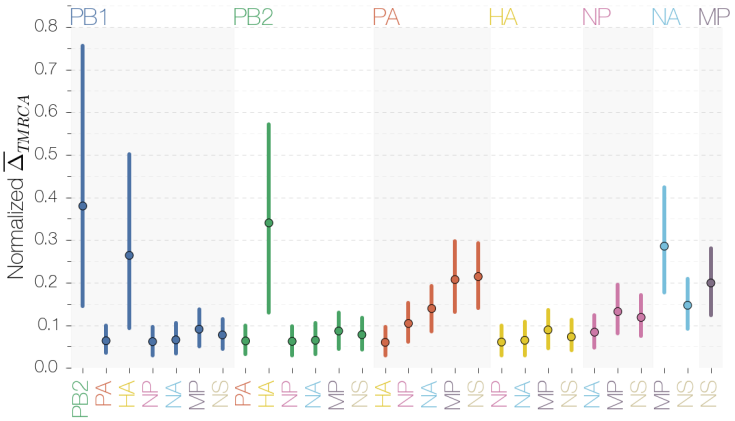

Figure 9. Normalized mean Δ_{TMRCA} statistics between pairs of segments. PB1, PB2 and HA trees exhibit reciprocally highly similar TMRCAs, unlike most other pairwise comparisons. All vertical lines indicating uncertainty are 95% highest posterior densities (HPDs).

Linkage disequilibrium (LD) is a measure of non-random association between pairs of alleles at different polymorphic loci within a population. To estimate LD no other information, except for allele frequencies at polymorphic amino acid sites, is required. Although not an external validation of previous phylogenetic results, we observe significantly greater LD values between PB1, PB2 and HA than between other pairs of segments (Figure 10). This suggests that PB1, PB2 and HA segments possess a considerable number of co-assorting alleles, which upon closer inspection are associated with either Vic or Yam lineage segments. We conclude that Victoria and Yamagata lineages of PB1, PB2 and HA have accumulated lineage-specific amino acid substitutions. Of the amino acid sites that exhibit high LD on PB1, PB2 and HA proteins, there are 5 sites on PB1, 8 on PB2 and 13 on HA proteins which form a network of sites exhibiting high LD (Figures S7 and S8). Figure S7 also shows that the number of variable amino acid sites in PB1 and PB2 proteins is comparable to most other proteins investigated (see Methods for cut-offs used to exclude sites from LD analysis). These sites define the split between Vic and Yam lineages within PB1, PB2 and HA segments. In addition, there are sites on PB1, PB2 and HA proteins which also show high, albeit smaller, LD which correspond to sites which have undergone amino acid replacements some time after the Vic/Yam split.

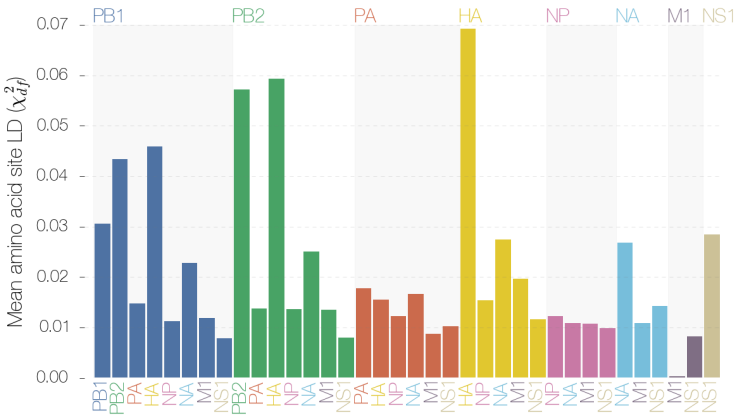

Figure 10. LD comparison between influenza B proteins. Pairwise comparisons of linkage disequilibrium between amino acid sites on influenza B proteins. PB1, PB2 and HA proteins exhibit high mean amino acid LD between themselves. This is evidence of a considerable number of co-assorting alleles within these proteins.

Discussion

Evidence of a co-adapted PB1-PB2-HA gene complex

In this paper we show that the PB1, PB2 and HA segments of influenza B viruses are the only ones that have continuously maintained separate Vic and Yam lineages, while other segments have fixed either Vic or Yam lineages (Figures 2, 5 and 6). Evidence suggests that this is a result of prolonged lack of reassortment between Vic and Yam lineages in PB1, PB2 and HA (Figure 4) which possess co-assorting sequences detectable as high linkage disequilibrium (Figure 10). We propose that this pattern of coassortment is due to the action of selection and not simply biased reassortment. We do not find evidence of bias in reassortment or segment packaging bias between PB1, PB2 and HA segments (Figure 8), but evidence points to ressortments between these segments being disfavored over longer spans of evolutionary time (Figure 9).

When comparing the mean tip-tip TMRCA deviations we find evidence of isolation by distance - when reassortments occur between PB1, PB2 and HA segments those events tend to involve branches that have similar TMRCAs. Other segment pairs, with the exception of MP-NA, seem to show little signal of this and are able to reassort with branches that have been evolving under temporally distant genomic backgrounds. In addition, most strains with mixed-lineage PB1-PB2-HA complexes occurred in the early years of the Vic–Yam split, when the two lineages were presumably much more similar at the nucleotide and amino acid levels. The most recent influenza B viruses with B/Waikato/6/2005-like genome constellations possessed mixed-lineage PB1-PB2-HA complexes and circulated for only 5 months before presumably going extinct.

Causes of co-dependence between PB1-PB2-HA

Previous studies have investigated possible co-dependence patterns between segments of influenza B viruses, by focusing on segments which would be expected to be co-adapted, *e.g.* PB1-PB2-PA and HA-NA [20]. Though it would be easy to explain co-adaptation between these segments by referring to their functional roles *e.g.* PB1-PB2-PA form the polymerase heterotrimer and HA-NA have antagonistic activities, our findings suggest a counter-intuitive relationship between PB1, PB2 and HA segments.

It is clear that PB1-PB2-HA segments of Vic and Yam lineages do not preferentially reassort together: there have been at least 4 sampled mixed-lineage PB1-PB2-HA complex constellations which did not become fixed in the population and our estimates of the number of reassortments do not differ significantly between all segments (Figure 8). The latter is in line with recent experiments in influenza A that have shown that reassortment between segments differing by a single synonymous difference is highly efficient [21]. In the absence of clear functional explanations for why PB1, PB2 and HA should be co-adapted we offer several alternatives.

It is perhaps easiest to explain co-dependence between PB1 and PB2 segments based on their functions as part of the influenza RNA-dependent RNA polymerase (RdRp) heterotrimer. Indeed, trees of PB1 and PB2 segments exhibit high similarity in tip–tip TMRCAs (Figure 9), suggesting highly similar TMRCAs. In addition, PB1–PB2 reassortants are the rarest and least persistent among mixed-lineage PB1-PB2-HA strains and have not been isolated since 1996.

Explaining the co-dependence of PB1+2 and HA segments is more difficult. Many studies have noted a possible link between the HA, NA and PB1 segments of influenza A viruses [22, 23]. A previously used technique for producing vaccine seed strains involved selecting for HA-NA reassortants, which often resulted in PB1-HA-NA reassortants, the second most frequent class of reassortants [22,23]. Recent experiments have suggested that the presence or absence of a 'foreign' PB1 segment can have dramatic effects on HA concentration on the surface of virions and total virion production [24]. Additional evidence for a relationship between PB1 and HA segments in influenza A viruses is given by previous influenza pandemics, which were caused by avian-human influenza A virus reassortants. It has been established that at least for the 1957 and the 1968 influenza pandemics, caused by A/H2N2 and A/H3N2 subtypes, respectively, the viruses responsible were reassortants possessing PB1 and HA segments derived from avian influenza A viruses [25]. However, there have also been reassortant influenza A viruses circulating for prolonged periods of time in humans that did have disparate PB1 and HA segments, *e.g.* H1N2 outbreaks in 2001 [26] and H1N1/09 in 2009 [27].

Another possibility is the action of balancing selection in preserving the diversity in one segment, whilst the other segments hitchhike along. A good candidate for this would be HA, as it is now the sole bearer of substantial antigenic diversity within the influenza B population, as the Vic lineage NA segment went extinct in 2002 (Figures 2, 3, 5 and 6)). We find this scenario unlikely: if PB1+2 segments were hitchhiking with the HA segment stochastically we would expect to see more PB1+2 versus HA reassortants and fixation of Vic or Yam lineage PB1+2 segments in the influenza B population. Our main

dataset comprising 452 complete influenza B virus genomes has one instance of these kinds of reassortants characterised by B/Bangkok/163/1990-like genome constellations. The larger dataset with 1433 PB1, PB2 and HA sequences has an additional example of this: isolates with B/Waikato/6/2005-like PB1-PB2-HA constellations. We note that in both cases the genome constellations of these reassortants were not fixed in the influenza B population and the more recent B/Waikato/6/2005-like reassortants persisted for a much shorter period of time (5 months versus 4 years), suggestive of declining PB1+2/HA reassortant fitness over time.

Previous research has found that avian influenza A virus HA and NA segments, which are the primary vehicles of antigenicity, exhibit vast diversity when compared to 'internal segments' (including PB1 and PB2 segments), which show much more recent TMRCAs and less variability at the amino acid level [28,29]. This is easiest to interpret as frequency-dependent balancing selection acting to preserve antigenic diversity [30]. If this is the case in influenza B viruses, we expect balancing selection to act on HA and indirectly, through some unknown association with HA, on PB1 and PB2 segments.

It is possible that Vic and Yam lineages of PB1, PB2 and HA segments have simply drifted away from each other, without any one segment being the driver of diversity preservation in PB1, PB2 and HA segments. In this case PB1, PB2 and HA segments of Vic and Yam lineage accumulate substitutions that improve their ability to co-operate with segments of the same lineage and worse at interacting segments of a different lineage. This process, termed mutation-driven co-evolution [31], has been suggested to be the cause of hybrid dysfunction in *Saccharomyces* hybrids [32]. It is widely accepted that Victoria lineage HA had been restricted to eastern Asia between 1992 and 2000 [11, 33], offering a potential explanation for why the budding Victoria lineage segments were not homogenized via reassortment with Yamagata lineage segment in the early years of the split between the two lineages. The restriction to eastern Asia would presumably also give sufficient time for PB1, PB2 and HA segments to co-evolve together. Similarly to the previous scenario, however, mutation-driven co-evolution would require an association of some kind between PB1+2 and HA segments to explain the low observed frequency and poor fitness of PB1+2/HA reassortants. In addition, we see no reason why these three particular segments, and not the rest of the genome, would become co-adapted to each other in geographic isolation.

Suggested experiments

The association between Victoria and Yamagata lineage PB1-PB2-HA complexes should be relatively straightforward to test in the lab. Using previously developed plasmid systems [34] it would be possible to create artificial reassortants, combining Vic and Yam lineages of PB1, PB2 and HA segments into mixed-lineage PB1-PB2-HA complexes. We predict that artificially produced viruses with mixed-lineage PB1-PB2-HA complexes will have reduced fitness when compared to viruses with pure-lineage, i.e. entirely Vic or entirely Yam, PB1-PB2-HA complexes. In addition, we expect the relationship between Vic and Yam lineage PB1, PB2 and HA segments to be dependent on date of segment isolation, as viruses with mixed-lineage PB1-PB2-HA complexes isolated earlier should perform better

than viruses with PB1-PB2-HA segments isolated more recently.

Given the existence of B/Waikato/6/2005-like viruses and the rarity of PB1-PB2 reassortants we expect that artificially produced PB1-PB2 reassortants would be much less fit than either PB1-HA or PB2-HA reassortants. We thus expect the following hierarchy of reassortant fitnesses (in order of decreasing fitness): PB1-PB2-HA, PB1-PB2/HA and PB1-HA/PB2 or PB2-HA/PB1, though there is some evidence to suggest that PB1-HA/PB2 might be more fit than PB2-HA/PB1 (Figures S6). The history of reassortments in influenza B viruses (Figure 6) suggests that there are lineage-specific effects too, given the almost universal introgression of Yamagata lineage segments into Victoria PB1-PB2-HA background. However, our analyses do not indicate any obvious differences in synonymous, non-synonymous or nucleotide substitution rates between Vic and Yam PB1-PB2-HA segment complexes or segments associated with either of the two (Figure S9).

However, we also see that epistatic effects might interfere with fitness measurements, if for example non-PB1-PB2-HA segments are also temporally mismatched. Ideally, the co-adaptation would be easier to understand by referring to the structures of PB1 and PB2 proteins, as the link between these would be intuitive. We have identified amino acid sites which are linked between PB1, PB2 and HA proteins of Victoria and Yamagata lineages, but we find very few sites on PB1 and PB2 proteins (Figure S8) that fall within the regions that form contacts within the influenza B polymerase heterotrimer [35], suggesting more subtle roles for sites we have identified.

The future of influenza B viruses

We suggest that the preservation of two PB1-PB2-HA complex lineages is similar to genomic speciation islands, where small numbers of genes resist being homogenized through gene flow [36]. In this context, we see three potential paths of evolution for influenza B viruses. More segments could be recruited into the two currently circulating co-adapted segment complexes (PB1, PB2 and HA segments being the genomic speciation islands), as part of a speciation process, until all circulating influenza B viruses possess genomes with segments firmly associated with either the Vic or Yam lineage PB1-PB2-HA complex which could be referred to as belonging to either 'new Victoria' or 'new Yamagata' lineages. This is the speciation scenario. The two alternatives to this are the '*status quo*' model, where the influenza B genome continues to be homogenized via gene flow with the exception of PB1, PB2 and HA segments and the extinction scenario, whereby one of the two PB1-PB2-HA complexes goes extinct, marking the return of single-strain dynamics in the influenza B virus population.

Sympatric speciation in other systems usually requires strong barriers to introgression, *e.g.* infertility of F1 hybrids can lead to the evolution of prezygotic reproductive isolation otherwise known as reinforcement or the Wallace effect. To what extent this would apply to influenza B viruses remains unknown. If influenza B viruses are undergoing sympatric speciation, it is imperative to determine whether co-infection with Victoria and Yamagata lineages of PB1-PB2-HA segments occur at a sufficiently high frequency and result in considerable losses of fitness to drive the evolution of reassortment isolation mechanisms

(*e.g.* unique packaging signals) or whether co-infection is so rare that speciation occurs via mutation-driven co-evolution. We think that co-dependence between PB1 and PB2 segments can be explained by the fact that they are functionally linked: together with the PA protein they form part of the influenza B virus RNA-dependent RNA polymerase heterotrimer.

It also remains unknown whether reassortment events of the past 20 years (Figure 6) which frequently involved the NP, MP and NS segments from a Yamagata PB1-PB2-HA background reassorting into a Victoria PB1-PB2-HA background are indicative of rare events followed by selective sweeps or stochastic fixation. It's not unfeasible for selective sweeps following reassortments to break down developing co-adaptation of segments, especially if they are not functionally linked and have themselves been reassorted into a new PB1-PB2-HA background recently. This is the second scenario we might expect to occur, whereby relatively frequent reassortments occur between Vic and Yam lineages and are followed by selective sweeps. In this case influenza B viruses would undergo periodic genome homogenization events with the exception of PB1, PB2 and HA segments. Because Vic HA and Yam HA are antigenically dissimilar we think that balancing selection would prevent even strong selective sweeps from driving the opposing PB1-PB2-HA gene complex to extinction. This '*status quo*' model would require strong selective sweeps and/or relatively frequent reassortments, neither of which seem to be lacking in the influenza B virus population.

Given the relatively recent explosion of sequence data available for influenza B, it is difficult to say whether dynamics similar to Victoria and Yamagata lineages have not occurred in influenza B virus genomes before and left no trace through extinction. We find it unlikely that either Victoria or Yamagata lineage PB1-PB2-HA complexes will go extinct stochastically in the near future, as they have co-circulated for prolonged periods at a ratio close to 0.5, suggesting the action of balancing selection (Figure 5). Extinction through depletion of susceptible individuals, such as following influenza pandemics or mass vaccination seem unlikely as well. Both Victoria and Yamagata lineage PB1-PB2-HA complexes survived the admittedly mild influenza pandemic in 2009 and influenza vaccines, which are usually applied to specific subsets of the population, do not produce lifelong immunity.

Conclusion

We have used the Δ_{TMRCA} statistic to determine the degree of similarity in TMRCA dates between two temporally calibrated phylogenies. We believe that patristic distance methods such as this, though themselves far from being new, have considerable power to address a wide variety of problems when combined with temporal phylogenies. One of many useful applications of the Δ_{TMRCA} method would be identifying clades or taxa that are products of reticulate evolution. Δ_{TMRCA} measures between independent analyses of the same alignment ('within-alignment') could be used as a cutoff to detect outlier taxa with greater than expected 'between-alignment' Δ_{TMRCA} values. To develop further, however, the statistical properties of patristic distance methods have to be evaluated in greater detail. In addition, by treating the relative position of each isolate within a

phylogeny of one segment as a label and modelling it on phylogenies of other segments we have also developed a metric similar to 'between population' diversity used in calculating F_{ST}, which is capable of quantifying reticulate evolution-induced loss of diversity between partitions of taxa in two or more phylogenies.

In this paper we apply a novel combination of population genetics and phylogenetic methods to full genome sequences in order to describe and quantify reassortment patterns in influenza B viruses circulating in humans from 1980 to the present day. Our main finding is that in influenza B viruses only PB1, PB2 and HA segments maintain both Victoria and Yamagata lineages which associate with segments of their own lineage, yielding two co-circulating PB1-PB2-HA complexes: one entirely derived from Victoria and one from Yamagata lineage segments. We argue that this is due to selection against viruses with mixed-lineage PB1-PB2-HA complexes. Given sufficient time it should become clear whether PB1-PB2-HA complexes of human influenza B viruses are continuing to resist gene flow whilst the rest of the influenza B virus genome is repeatedly homogenized or whether the two PB1-PB2-HA complexes are recruiting the rest of the genomic segments into co-adapted and co-reassorting segment complexes on their way to sympatric speciation.

Methods

We compiled a dataset of 452 complete influenza B genomes from GISAID [37] dating from 1984 to 2012 (accession numbers and laboratory acknowledgements can be found in supplementary information). The longest protein coding region of each segment was extracted and used for all further analyses. We thus assume that homologous recombination has not taken place and that the evolutionary history of the whole segment can be inferred from the longest coding sequence in the segment. To date there has been little evidence of homologous recombination in influenza viruses [38–40]. The segments of each strain were assigned to either Vic or Yam lineage by making maximum likelihood trees of each segment using PhyML [41] and identifying whether the isolate was more closely related to B/Victoria/2/87 or B/Yamagata/16/88 sequences in that segment, with the exception of the NS segment (B/Victoria/2/87 was a reassortant and possessed a Yam lineage NS [42]), where B/Czechoslovakia/69/1990 was considered as being representative of Victoria lineage. Each strain was thus assigned 8 lineages depending on the combination of lineages from which their genomes were derived, for example all segments except for NS in strain B/Victoria/2/87 belong to Vic lineage and can thus be represented as (V,V,V,V,V,V,V,Y).

We also downloaded all available sequences of influenza B isolates sampled 1984–2013 from GISAID for which PB1, PB2 and HA segments were sequenced. This comprised a dataset of 1433 isolates in total which became available only after the primary analyses were performed using the smaller dataset and had too many sequences to analyze using the methods described later. Neighbor-joining trees [43] of PB1, PB2 and HA segments were made and each sequence assigned to a lineage based on grouping with either B/Victoria/2/87 or B/Yamagata/16/88 sequences, as before. Isolates which were not assigned entirely to either the Vic or the Yam lineage across PB1, PB2 and HA segments were extracted and identified as PB1-PB2-HA reassortants.

Temporally-calibrated phylogenies were recovered for each segment using the Markov chain Monte Carlo (MCMC) methods in the BEAST software package [44]. Here, we modeled the substitution process using the HKY model of nucleotide substitution [45], with separate transition models for each of the 3 codon partitions, and additionally estimate realized synonymous and non-synonymous substitution counts [46]. We used a flexible Bayesian skyride demographic model [47]. We accounted for incomplete sampling dates for 94 sequences (of which 93 had only year and 1 had only year and month of isolation) whereby tip date is estimated as a latent variable in the MCMC integration. A relaxed molecular clock was used, where branch lengths are drawn from a lognormal distribution [48]. We ran 3 independent MCMC chains, each with 200 million states, sampled every 20,000 steps and discarded the first 10% of the MCMC states as burn-in. After assessing convergence of all 3 MCMC chains by visual inspection using Tracer [49], we combined samples across chains to give a total of 27,000 samples from the posterior distribution of trees.

Every sequence was assigned 7 discrete traits in BEAUti corresponding to the lineages of all other segments with which a strain was isolated *e.g.* PB1 tree had PB2, PA, HA, NP, NA, MP and NS as traits and V or Y as trait values. We inferred the ancestral state of lineages in each segment by modelling transitions between these discrete states using

an asymmetric transition matrix [50] with Bayesian stochastic search variable selection (BSSVS) to estimate significant rates. Because the posterior set of trees for a single segment has branches labelled with the inferred lineage in the remaining 7 segments, we can detect inter-lineage reassortments between pairs of segments by observing state transitions, i.e. Yam to Vic or Vic to Yam (Figure 1). In addition, by reconstructing the ancestral state of all other genomic segments jointly we can infer co-reassortment events when more than one trait transition occurs on the same node in a tree.

Measures of diversity

We inferred the diversity of each segment at a single point in time by estimating the date of the most recent common ancestor of all branches at yearly intervals, which places an upper bound on the maximum amount of diversity existing at each time point. A version of this lineage turnover metric has previously been used to investigate the tempo and strength of selection in influenza A viruses during seasonal circulation [51]. In addition, we calculated mean pairwise time of most recent common ancestor (TMRCA) between branches labelled as Vic and Yam for PB1, PB2 and HA. This gave us a measure of how much a particular segment reassorts with respect to Vic and Yam lineages of PB1, PB2 and HA segments. If Vic and Yam lineages of PB1, PB2 and HA segments were to be considered as being separate populations this measure would be equivalent to 'between population' diversity.

We also calculated the total amount of evolutionary time spent by each segment with entirely Vic, entirely Yam or mixed lineage PB1, PB2 and HA segments. We do this by summing the branch lengths in each tree under 3 (2 in the case of PB1, PB2 and HA trees) different lineage combinations of the PB1, PB2 and HA segments: PB1-PB2-HA derived entirely from Yamagata lineage, PB1-PB2-HA entirely derived from Victoria lineage and PB1-PB2-HA derived from a mixture of the two lineages. This gives a measure of how successful, over long periods of time, each particular PB1-PB2-HA constellation has been.

Tree to tree similarities

We express the distance Δ_{TMRCA} between trees belonging to two segments A and B by comparing TMRCA differences between pairs of tips in the two trees. In comparing of trees of segments A and B we calculate

$$\Delta_{\mathrm{TMRCA}}(A_i, B_i) = \frac{f(A_i, A_i') + f(B_i, B_i')}{2\, f(A_i, B_i)}, \tag{1}$$

where $f(A_i, B_i) = \frac{1}{n}\sum_{j=1}^{n} g(A_{ij}, B_{ij})$ and $g(A_{ij}, B_{ij})$ is the absolute difference in TMRCA of a pair of tips j, where the pair is drawn from the ith posterior sample of tree A and the ith posterior sample of tree B. Additionally, $f(A_i, A_i')$ is calculated from the ith posterior sample of tree A and ith posterior sample of an independent analysis of tree A (which we refer to as A'), to control for variability in tree topology stability over the course of the MCMC chain caused by differences in alignment lengths used to produce the trees. We had 3 replicate analyses of each segment and in order to calculate $f(A_i, A_i')$ we used

analyses numbered 1, 2 and 3 as A and analyses numbered 2, 3 and 1 as A', in that order. We subsampled our combined posterior distribution of trees to give a total of 2700 trees on which to analyze Δ_{TMRCA}.

Calculating $\Delta_{\mathrm{TMRCA}}(A_i, B_i)$ for each MCMC state provides us with a posterior distribution of this statistic allowing specific hypotheses regarding similarities between the trees of different segments to be tested. Our approach exploits the branch scaling used by BEAST [44], since the trees are scaled in absolute time and insensitive to variation in nucleotide substitution rates between segments, allowing for direct comparisons between TMRCAs in different trees. Our Δ_{TMRCA} statistic is an extension of patristic distance methods and has previously been used to tackle a wide variety of problems, as phylogenetic distance in predicting viral titer in *Drosophila* infected with viruses from closely related species [52] and to assess temporal incongruence in a phylogenetic tree of amphibian species induced by using different calibrations [53].

Subtree prune and regraft (SPR) distances between phylogenetic trees are an approximate measure of the numbers of reassortment or recombination events. Exact SPR distances are difficult to compute, as they depend on the SPR distance itself and are impractical to compute for posterior distributions of trees except for the most similar trees. We calculated approximate SPR distances [17–19] to quantify the numbers of reassortments that have taken place between all pairs of segments. Approximate SPR distances d_{SPR} were normalized using the procedure described above, where $f(A_i, A'_i)$, $f(B_i, B'_i)$ and $f(A_i, B_i)$ are approximate SPR distances between ith posterior samples from segments A, B and independent analyses thereof (A' and B').

Linkage disequilibrium across the influenza B genome

We estimated linkage disequilibrium (LD) between amino acid sites across the longest proteins encoded by each segment of the influenza B virus genome. To quantify LD we adapt the χ^2_{df} statistic from [54]:

$$\chi^2_{df} = \frac{\chi^2}{N\,(k-1)\,(m-1)}, \tag{2}$$

where χ^2 is calculated from a classical contingency table, N is the number of haplotypes and $(k-1)(m-1)$ are the degrees of freedom. This statistic is equal to the widely used r^2 LD statistic at biallelic loci, but also quantifies LD when there are more than two alleles per locus [55]. LD was estimated only at loci where each nucleotide or amino acid allele was present in at least two isolates. We ignored gaps in the alignment and did not consider them as polymorphisms. We also calculated mean LD for all pairs of segments to quantify the overall association between them. Linkage disequilibrium was not estimated for the larger PB1, PB2 and HA dataset, because it lacked sequences from the rest of the influenza B genome and thus would have failed to put LD between PB1, PB2 and HA proteins into a genome-wide perspective.

Acknowledgements

GD was supported by a NERC studentship D76739X. TB was supported by a Newton International Fellowship from the Royal Society. The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement no. 278433-PREDEMICS and ERC Grant agreement no. 260864. AR and SL acknowledge the support of the Wellcome Trust (grant no. 092807).

References

1. World Health Organization (2009) Influenza Fact sheet (Available at http://www.who.int/mediacentre/factsheets/fs211/en/).
2. Paul Glezen W, Schmier JK, Kuehn CM, Ryan KJ, Oxford J (2013) The burden of influenza b: A structured literature review. American Journal of Public Health 103:e43–e51.
3. Broberg E, Beauté J, Snacken R (2013) Fortnightly influenza surveillance review, 9th May., (European Centre for Disease Prevention and Control, Stockholm), Technical report.
4. Osterhaus ADME, Rimmelzwaan GF, Martina BEE, Bestebroer TM, Fouchier RaM (2000) Influenza b virus in seals. Science 288:1051–1053.
5. Bodewes R, et al. (2013) Recurring influenza b virus infections in seals. Emerging Infectious Diseases 19:511–512.
6. Burnet SFM (1955) Principles of animal virology (Academic Press).
7. Hay AJ, Gregory V, Douglas AR, Lin YP (2001) The evolution of human influenza viruses. Philosophical Transactions of the Royal Society of London. Series B 356:1861–1870 PMID: 11779385 PMCID: PMC1088562.
8. Bedford T, et al. (2014) Integrating influenza antigenic dynamics with molecular evolution. eLife 3.
9. Rota PA, et al. (1990) Cocirculation of two distinct evolutionary lineages of influenza type b virus since 1983. Virology 175:59–68.
10. Kanegae Y, et al. (1990) Evolutionary pattern of the hemagglutinin gene of influenza b viruses isolated in japan: cocirculating lineages in the same epidemic season. Journal of Virology 64:2860–2865.
11. Nerome R, et al. (1998) Evolutionary characteristics of influenza b virus since its first isolation in 1940: dynamic circulation of deletion and insertion mechanism. Archives of Virology 143:1569–1583.
12. Nakagawa N, et al. (2002) Emergence of an influenza b virus with antigenic change. Journal of Clinical Microbiology 40:3068–3070.

13. Ansaldi F, et al. (2003) Molecular characterization of influenza b viruses circulating in northern italy during the 2001–2002 epidemic season. Journal of Medical Virology 70:463–469.

14. Chen R, Holmes EC (2008) The evolutionary dynamics of human influenza b virus. Journal of Molecular Evolution 66:655–663.

15. Reed C, Meltzer MI, Finelli L, Fiore A (2012) Public health impact of including two lineages of influenza b in a quadrivalent seasonal influenza vaccine. Vaccine 30:1993–1998.

16. Svinti V, Cotton JA, McInerney JO (2013) New approaches for unravelling reassortment pathways. BMC Evolutionary Biology 13:1 PMID: 23279962.

17. Whidden C, Zeh N (2009) in Algorithms in Bioinformatics, Lecture Notes in Computer Science, eds Salzberg SL, Warnow T (Springer Berlin Heidelberg) No. 5724, pp 390–402.

18. Whidden C, Beiko RG, Zeh N (2010) in Experimental Algorithms, Lecture Notes in Computer Science, ed Festa P (Springer Berlin Heidelberg) No. 6049, pp 141–153.

19. Whidden C, Beiko RG, Zeh N (2013) Fixed-parameter algorithms for maximum agreement forests. SIAM Journal on Computing 42:1431–1466.

20. McCullers JA, Saito T, Iverson AR (2004) Multiple genotypes of influenza b virus circulated between 1979 and 2003. Journal of Virology 78:12817–12828 PMID: 15542634.

21. Marshall N, Priyamvada L, Ende Z, Steel J, Lowen AC (2013) Influenza virus reassortment occurs with high frequency in the absence of segment mismatch. PLoS Pathog 9:e1003421.

22. Bergeron C, Valette M, Lina B, Ottmann M (2010) Genetic content of influenza H3N2 vaccine seeds. PLoS Currents 2:RRN1165.

23. Fulvini AA, et al. (2011) Gene constellation of influenza a virus reassortants with high growth phenotype prepared as seed candidates for vaccine production. PLoS ONE 6:e20823.

24. Cobbin JCA, Verity EE, Gilbertson BP, Rockman SP, Brown LE (2013) The source of the PB1 gene in influenza vaccine reassortants selectively alters the hemagglutinin content of the resulting seed virus. Journal of Virology 87:5577–5585 PMID: 23468502.

25. Kawaoka Y, Krauss S, Webster RG (1989) Avian-to-human transmission of the PB1 gene of influenza a viruses in the 1957 and 1968 pandemics. Journal of Virology 63:4603–4608 PMID: 2795713 PMCID: PMC251093.

26. Gregory V, et al. (2002) Emergence of influenza a {H1N2} reassortant viruses in the human population during 2001. Virology 300:1 – 7.

27. Smith GJD, et al. (2009) Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza a epidemic. Nature 459:1122–1125.

28. Chen R, Holmes EC (2006) Avian influenza virus exhibits rapid evolutionary dynamics. Molecular Biology and Evolution 23:2336–2341 PMID: 16945980.

29. Obenauer JC, et al. (2006) Large-scale sequence analysis of avian influenza isolates. Science 311:1576–1580 PMID: 16439620.

30. Worobey M, Han GZ, Rambaut A (2014) A synchronized global sweep of the internal genes of modern avian influenza virus. Nature advance online publication.

31. Presgraves DC (2010) The molecular evolutionary basis of species formation. Nature Reviews Genetics 11:175–180.

32. Lee HY, et al. (2008) Incompatibility of nuclear and mitochondrial genomes causes hybrid sterility between two yeast species. Cell 135:1065–1073.

33. Shaw MW, et al. (2002) Reappearance and global spread of variants of influenza B/Victoria/2/87 lineage viruses in the 2000–2001 and 2001–2002 seasons. Virology 303:1–8.

34. Hoffmann E, et al. (2002) Rescue of influenza b virus from eight plasmids. Proceedings of the National Academy of Sciences 99:11411–11416 PMID: 12172012.

35. Sugiyama K, et al. (2009) Structural insight into the essential PB1–PB2 subunit contact of the influenza virus RNA polymerase. The EMBO Journal 28:1803–1811.

36. Turner TL, Hahn MW, Nuzhdin SV (2005) Genomic islands of speciation in anopheles gambiae. PLoS Biol 3:e285.

37. Bogner P, Capua I, Lipman DJ, Cox NJ, et al. (2006) A global initiative on sharing avian flu data. Nature 442:981–981.

38. Chare ER, Gould EA, Holmes EC (2003) Phylogenetic analysis reveals a low rate of homologous recombination in negative-sense RNA viruses. Journal of General Virology 84:2691–2703 PMID: 13679603.

39. Boni MF, Zhou Y, Taubenberger JK, Holmes EC (2008) Homologous recombination is very rare or absent in human influenza a virus. Journal of Virology 82:4807–4811 PMID: 18353939.

40. Han GZ, Boni MF, Li SS (2010) No observed effect of homologous recombination on influenza c virus evolution. Virology Journal 7:227 PMID: 20840780.

41. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic Biology 52:696–704 PMID: 14530136.

42. Lindstrom SE, et al. (1999) Comparative analysis of evolutionary mechanisms of the hemagglutinin and three internal protein genes of influenza b virus: Multiple cocirculating lineages and frequent reassortment of the NP, m, and NS genes. Journal of Virology 73:4413–4426.

43. Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular biology and evolution 4:406–425 PMID: 3447015.

44. Drummond AJ, Suchard MA, Xie D, Rambaut A (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. Molecular Biology and Evolution.

45. Hasegawa M, Kishino H, Yano Ta (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. Journal of Molecular Evolution 22:160–174.

46. O'Brien JD, Minin VN, Suchard MA (2009) Learning to count: Robust estimates for labeled distances between molecular sequences. Molecular Biology and Evolution 26:801–814 PMID: 19131426.

47. Minin VN, Bloomquist EW, Suchard MA (2008) Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. Molecular Biology and Evolution 25:1459–1471.

48. Drummond AJ, Ho SYW, Phillips MJ, Rambaut A (2006) Relaxed phylogenetics and dating with confidence. PLoS Biol 4:e88.

49. Rambaut, A. and Suchard, M. and Drummond, A. (2009) Tracer v1.5 (Available at http://tree.bio.ed.ac.uk/software/tracer/).

50. Lemey P, Rambaut A, Drummond AJ, Suchard MA (2009) Bayesian phylogeography finds its roots. PLoS Comput Biol 5:e1000520.

51. Bedford T, Cobey S, Pascual M (2011) Strength and tempo of selection revealed in viral gene genealogies. BMC Evolutionary Biology 11:220 PMID: 21787390.

52. Longdon B, Hadfield JD, Webster CL, Obbard DJ, Jiggins FM (2011) Host phylogeny determines viral persistence and replication in novel hosts. PLoS Pathog 7:e1002260.

53. Ruane S, Pyron RA, Burbrink FT (2011) Phylogenetic relationships of the cretaceous frog beelzebufo from madagascar and the placement of fossil constraints based on temporal and phylogenetic evidence. Journal of Evolutionary Biology 24:274–285.

54. Hedrick PW, Thomson G (1986) A two-locus neutrality test: Applications to humans, e. coli and lodgepole pine. Genetics 112:135–156 PMID: 3510942.

55. Zhao H, Nettleton D, Soller M, Dekkers JCM (2005) Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of linkage disequilibrium between markers and QTL. Genetics Research 86:77–87.

56. Nath ST, Nayak DP (1990) Function of two discrete regions is required for nuclear localization of polymerase basic protein 1 of A/WSN/33 influenza virus (h1 n1). Molecular and Cellular Biology 10:4139–4145 PMID: 2196448.

57. Fodor E, Smith M (2004) The PA subunit is required for efficient nuclear accumulation of the PB1 subunit of the influenza a virus RNA polymerase complex. Journal of Virology 78:9144–9153 PMID: 15308710.

58. Guilligay D, et al. (2008) The structural basis for cap binding by influenza virus polymerase subunit PB2. Nature Structural & Molecular Biology 15:500–506.

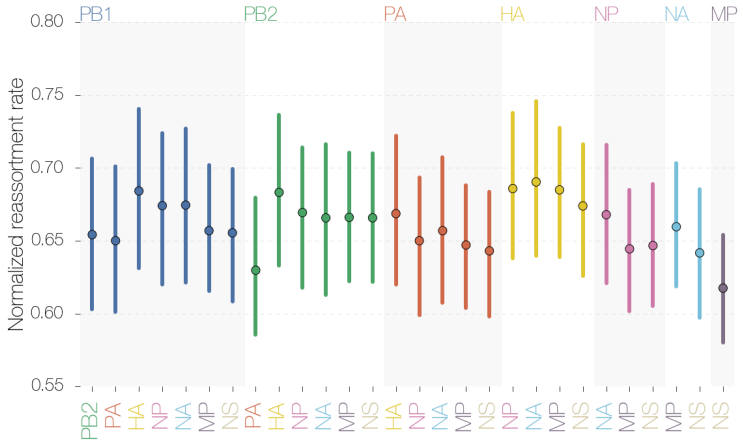

Figure S1. Normalized reassortment rate Reassortment rate is calculated as approximate number of SPR moves per sum of total time in both trees.

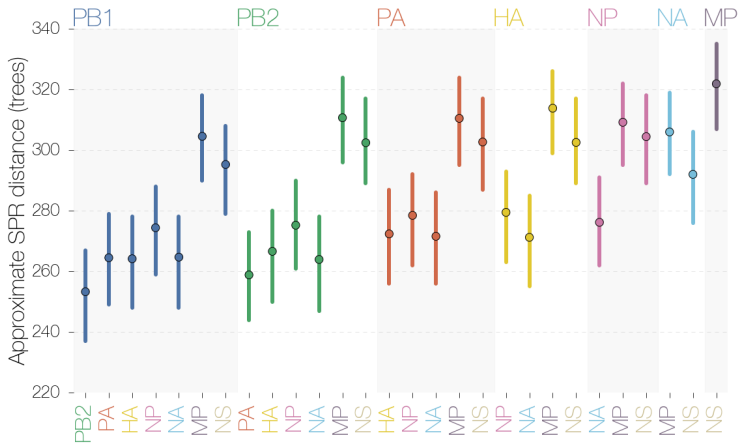

Figure S2. Approximate SPR distances between all pairs of trees of segments. There is a visible trend where comparisons between shorter segments have larger SPR distances, consistent with decreasing tree topology stability over the course of MCMC for shorter segments.

Figure S3. Approximate SPR distances between replicate trees of each segment. Approximate SPR distances between replicates of MP and NS trees are much higher (≈260) than any other segments, suggesting greater variability in tree topology over the course of MCMC. SPR distances between replicates of most other segments are ≈200.

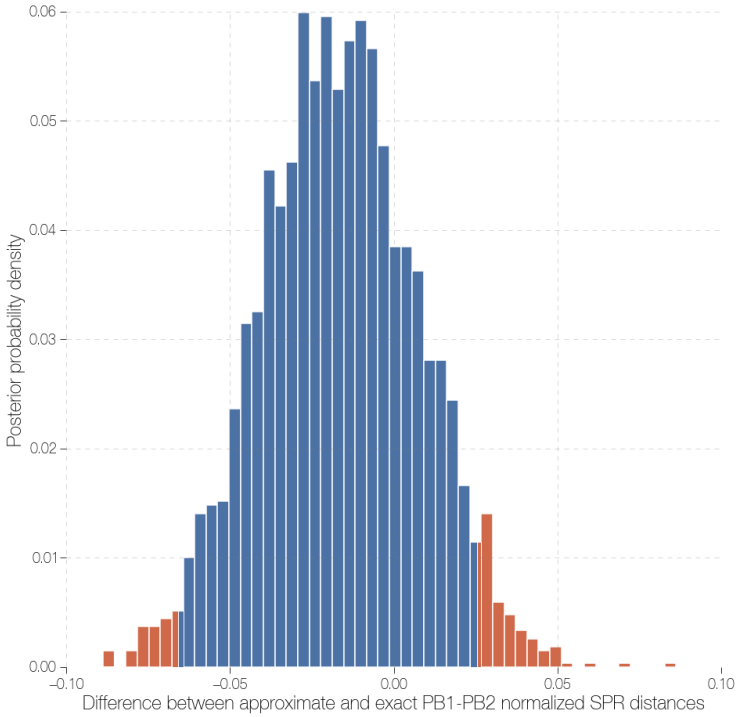

Figure S4. Distribution of differences between exact and approximate PB1-PB2 SPR distances after normalization. 95% HPD interval (blue) overlaps zero, suggesting no evidence of differences between approximate and exact SPR distances following normalization.

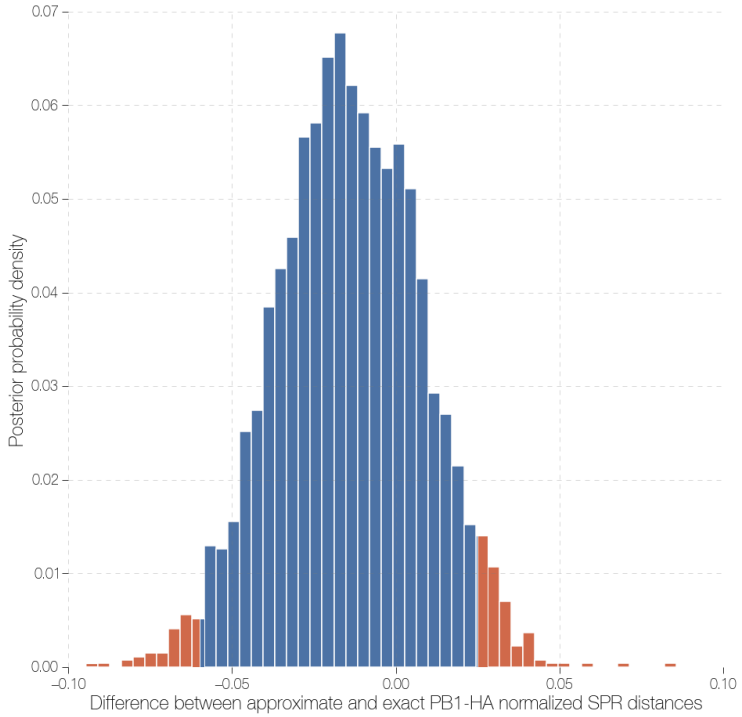

Figure S5. Distribution of differences between exact and approximate PB1-HA SPR distances after normalization. 95% HPD interval (blue) overlaps zero, suggesting no evidence of differences between approximate and exact SPR distances following normalization.

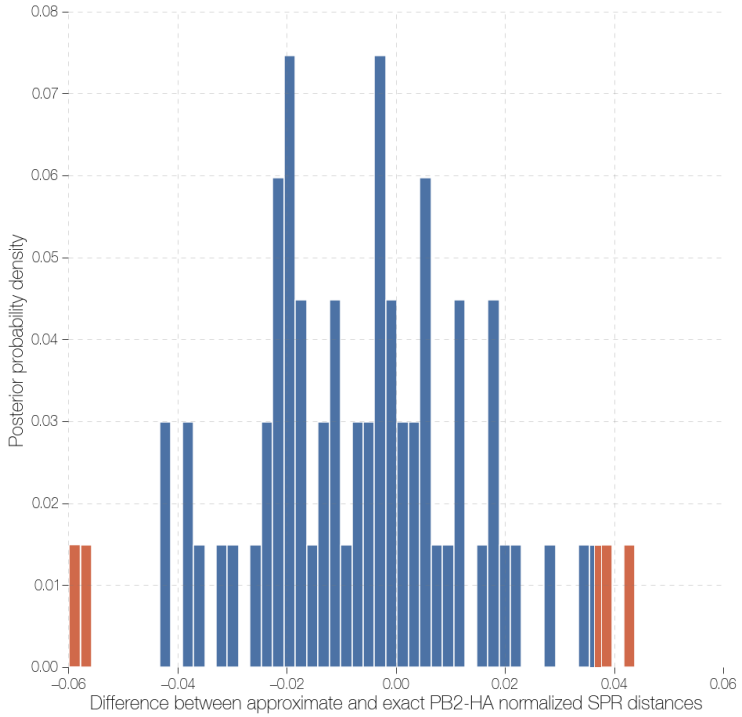

Figure S6. Distribution of differences between exact and approximate PB2-HA SPR distances after normalization. 95% HPD interval (blue) overlaps zero, suggesting no evidence of differences between approximate and exact SPR distances following normalization. Due to excessively long computation time of exact SPR distances between PB2 and HA trees few comparisons were made.

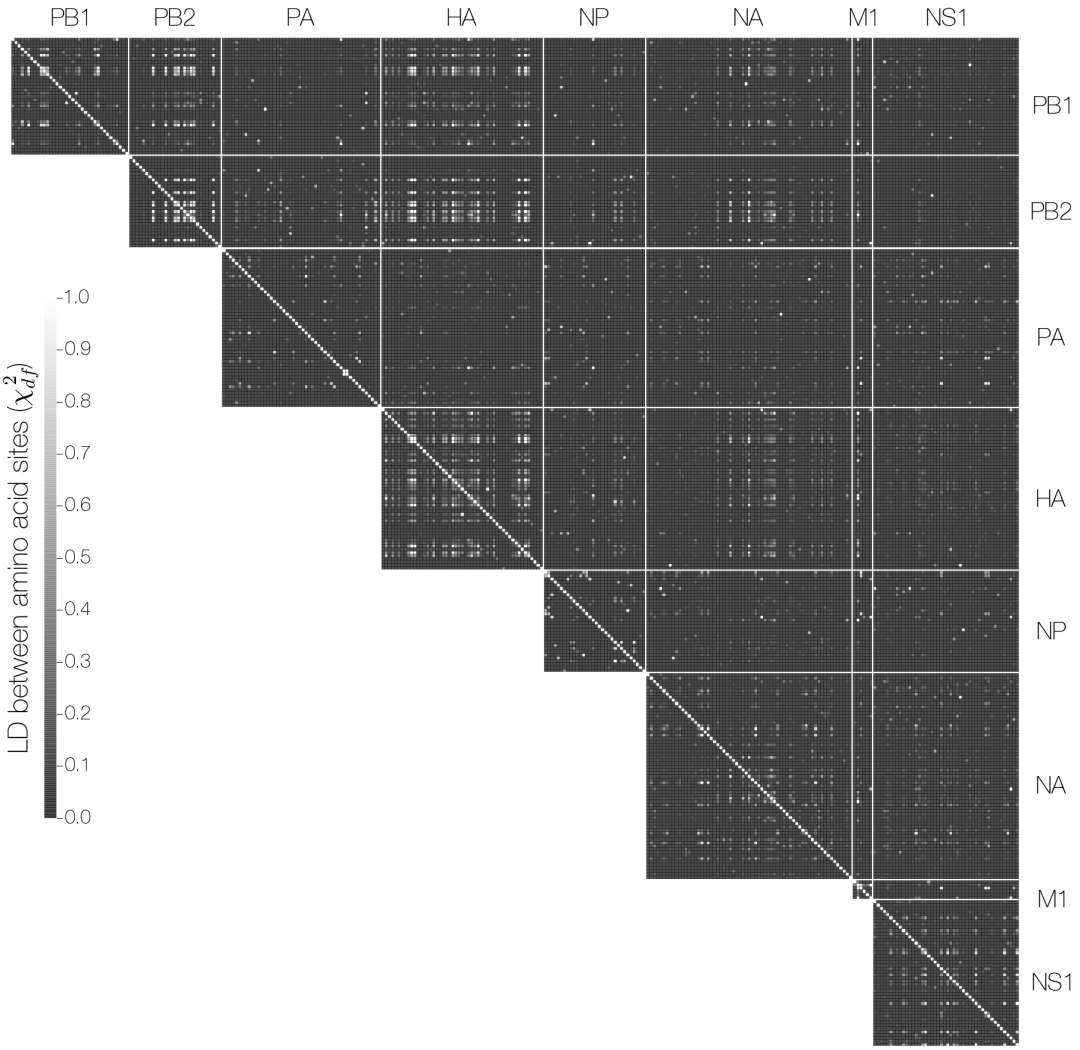

Figure S7. Heatmap of genome-wide linkage disequilibrium (χ^2_{df}) between polymorphic amino acid sites. Patterns of LD across the genome suggest a network of reciprocally linked amino acid sites on PB1, PB2, HA and, to some extent NA, proteins. Proximity of sites on heatmaps might not correspond to proximity of sites within proteins.

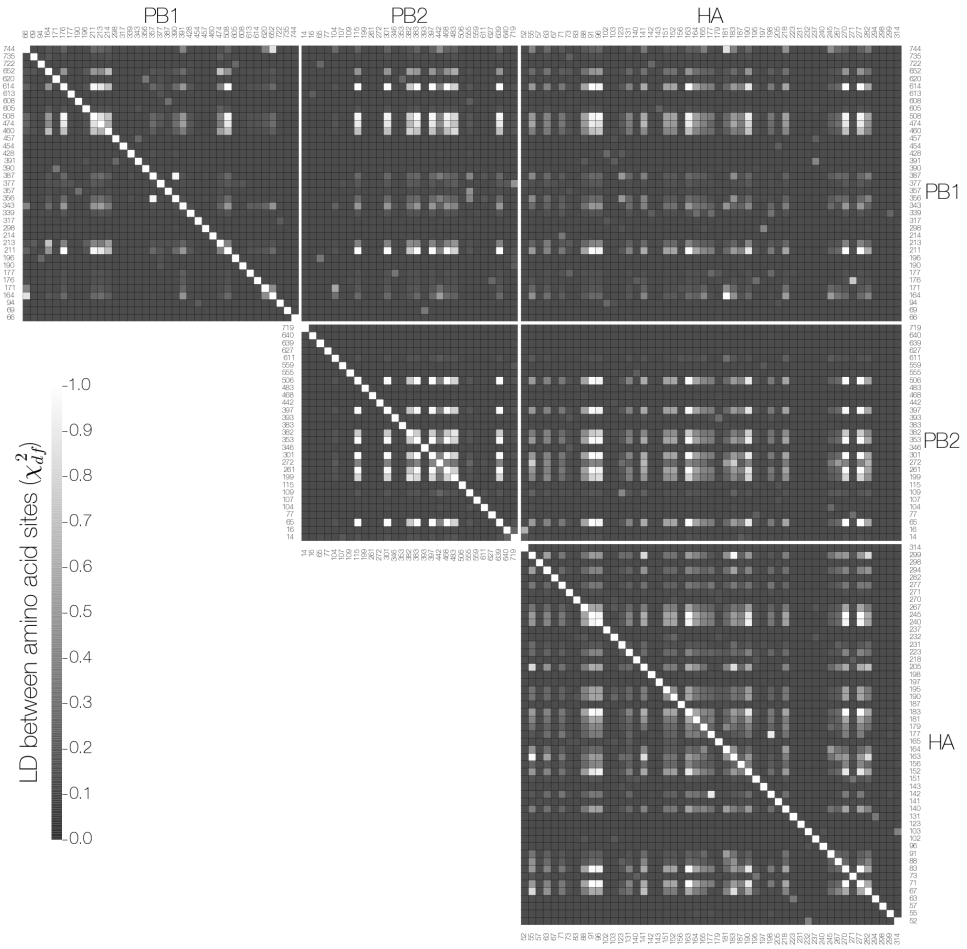

Figure S8. Heatmap of linkage disequilibrium (χ^2_{df}) between amino acid sites on PB1, PB2 and HA proteins. Numbers next to each row and column correspond to amino acid site number within a given protein starting from methionine. Amino acid sites exhibiting reciprocally high LD between PB1, PB2 and HA proteins are: 176, 211, 213, 214, 508 (PB1), 115, 301, 382, 383, 397, 468, 483, 639 (PB2) and 86, 91, 96, 141, 152, 163, 164, 183, 190, 218, 270, 277, 282 (HA). Sites 211, 213 and 214 on the PB1 protein are very close to each other and the stretch of sequence around these residues contains many positively charged amino acids (lysine and arginine). Multiple nuclear localization signals (NLSs) are predicted to occur around this region and sites 211, 213 and 214 are either predicted to be near the end of a mono-partite NLS or the beginning of a bi-partite NLS. Previous research [56] suggests that in the influenza A PB1 protein residue 211 (homologous to influenza B PB1 residue 211) is the last residue of a bi-partite NLS. Almost all Yamagata lineage isolates possess arginine (R) residue at PB1 positions 211 and 214 and a serine (S) residue at position 213, whereas Victoria lineage isolates have lysine (K) at positions 211 and 214 and threonine (T) at position 213. It remains to be seen whether these sites significantly affect the nuclear import efficiency of the PB1 protein of either lineage. Though the PB1 protein is known to accumulate in the nucleus on its own, efficient import into the nucleus requires the presence of the PA protein [57]. Similarly, sites 382 and 397 on the PB2 protein are close to residues 377, 406 and 408 which are homologous to sites in influenza A that are responsible for mRNA cap-binding [58].

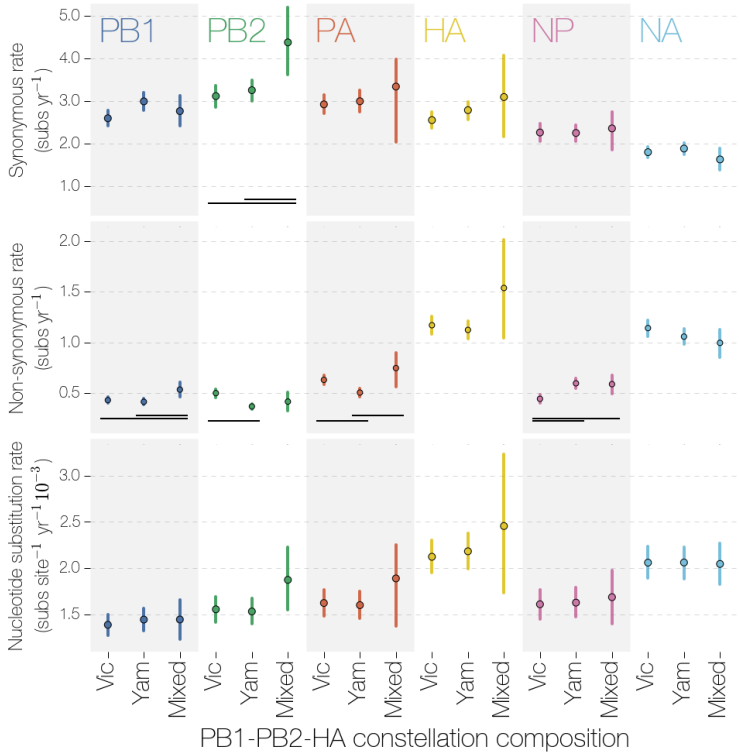

Figure S9. Synonymous, non-synonymous and nucleotide substitution rates in segments under different PB1-PB2-HA complexes. Evolutionary rate dissimilarities under Vic and Yam PB1-PB2-HA complexes are not systematic and appear negligible. Synonymous and non-synonymous rates were calculated by dividing the sum of all substitutions of a given class by the total amount of evolutionary time under each PB1-PB2-HA constellation. Nucleotide rates were calculated by multiplying the inferred nucleotide substitution rate on each branch by the branch length, then dividing this by the total amount of evolutionary time under each PB1-PB2-HA constellation. Vertical bars indicating uncertainty are 95% HPDs, black bars indicate 95% HPDs that do not overlap.

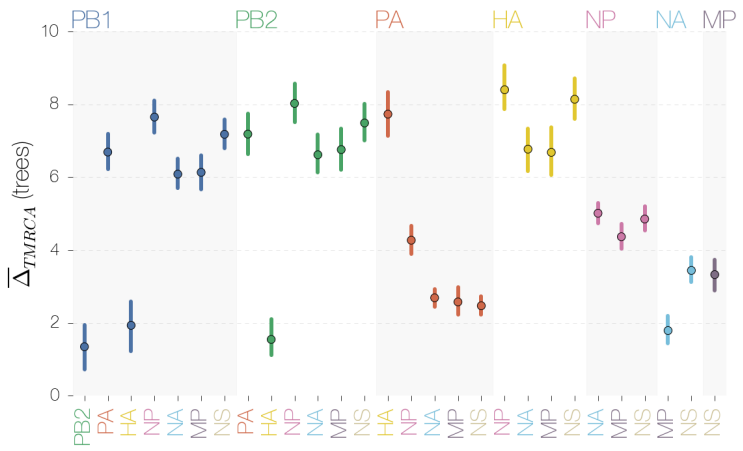

Figure S10. Mean Δ_{TMRCA} between all pairs of trees of segments. Mean Δ_{TMRCA} between trees of segments reveal that tip pairs in PB1, PB2 and HA trees have very similar TMRCAs. The upper tail of the 95% HPD (HPDs are represented as vertical lines) interval of mean Δ_{TMRCA} values for PB1-PB2-HA and MP-NA trees do not exceed 3 years.

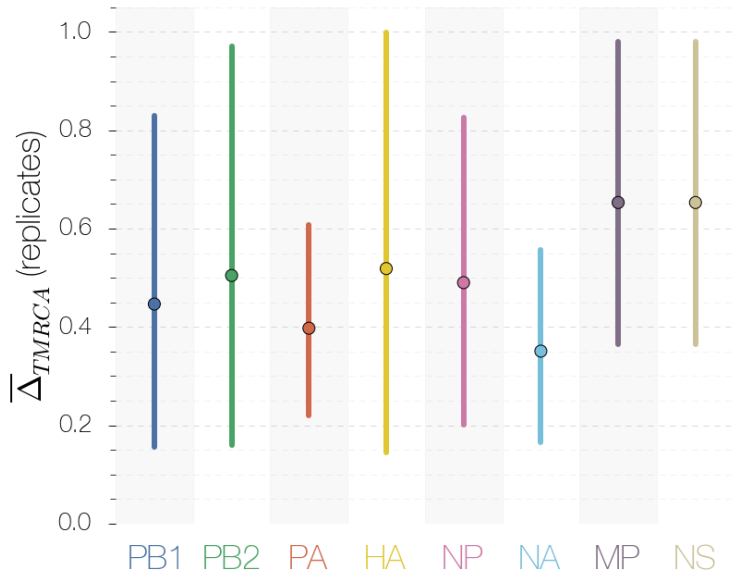

Figure S11. Mean Δ_{TMRCA} between replicate trees of each segment. Mean Δ_{TMRCA} values between independent analyses of each segment show that mean Δ_{TMRCA} values rarely exceed 1 year.