Supplemental information:
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

August 29, 2014

Confirmation of primary findings

We sought to confirm our findings through measurement of linkage disequilibrium (LD), a measure of non-random association between polymorphic loci within a population. We estimated LD directly from haplotype frequencies at polymorphic amino acid sites (see Methods) in the secondary dataset, thereby avoiding phylogenetic reconstruction or Vic/Yam lineage assignment. We observe greater amino acid LD values between PB1, PB2 and HA than between other pairs of segments (Figure S1) in a large secondary dataset. This suggests that PB1, PB2 and HA segments possess a considerable number of co-assorting non-synonymous 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 4 sites on PB1, 4 on PB2 and 4 on HA proteins which form a network of sites exhibiting high LD (Figures S3 and S2). These sites define the split between Vic and Yam lineages within PB1, PB2 and HA segments. In addition, there are sites on PB1, PB2, HA and NA 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.

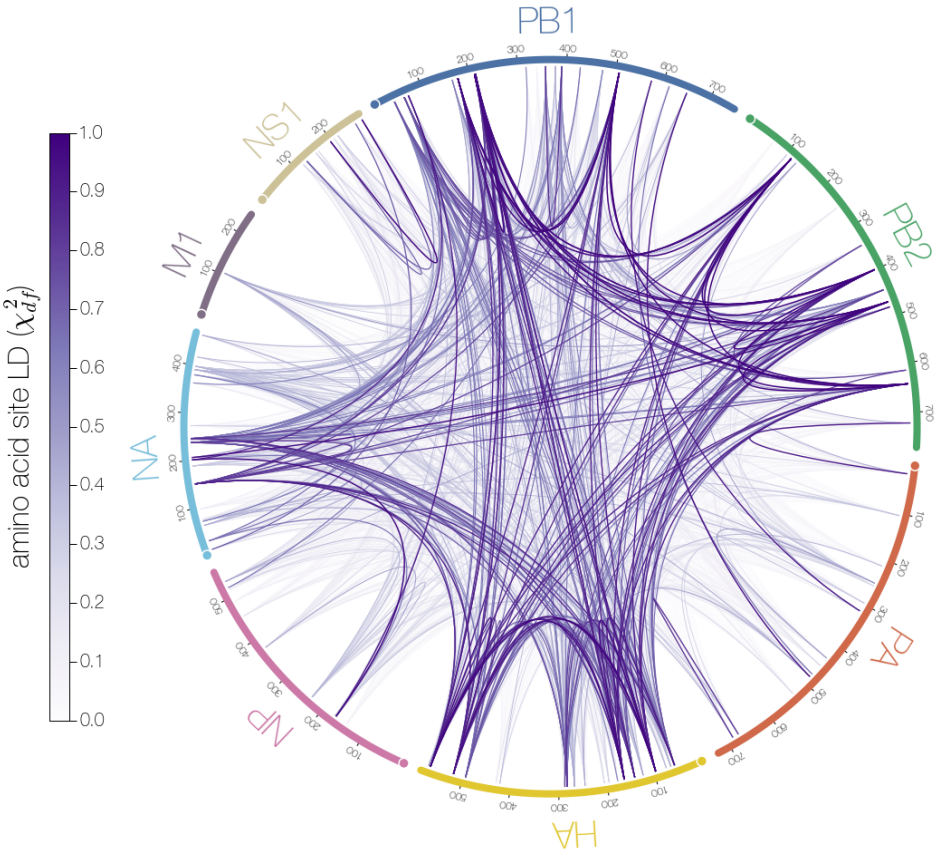

Figure S1. LD comparison between influenza B proteins. Pairwise comparisons of linkage disequilibrium between amino acid sites on influenza B proteins in the secondary dataset. Many polymorphic amino acid sites on PB1, PB2 and HA proteins exhibit high LD between themselves, followed by the NA protein. This is evidence of a considerable number of co-assorting alleles within these proteins.

Analysis of within-lineage reassortment patterns

Subtree prune and regraft (SPR) distances between phylogenetic trees are an approximate measure of the numbers of reassortment or recombination events (Svinti et al., 2013). 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 (Whidden and Zeh, 2009; Whidden et al., 2010, 2013) to quantify the numbers of reassortments that have taken place between all pairs of segments. Normalized approximate SPR distances,d_{SPR}, were recovered using (see Methods):

$$d_{\mathrm{SPR}}(A_i, B_i) = \frac{f(A_i, A'_i) + f(B_i, B'_i)}{2\, f(A_i, B_i)}, \tag{1}$$

where $f(A_i, A'_i)$, $f(B_i, B'_i)$ 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'). Figure S4 shows approximate SPR distances between all pairs of segment trees after normaliza-

Figure S2. 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, 508 (PB1), 115, 399, 470, 642 (PB2) and 96, 163, 520 and 570 (HA). Sites 211 and 213 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 and 213 are either predicted to be near the end of a mono-partite NLS or the beginning of a bi-partite NLS. Previous research (Nath and Nayak, 1990) 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 position 211 and a serine (S) residue at position 213, whereas Victoria lineage isolates have lysine (K) at position 211 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 (Fodor and Smith, 2004). Similarly, site 399 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 (Guilligay et al., 2008).

Figure S3. 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.

tion. 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.

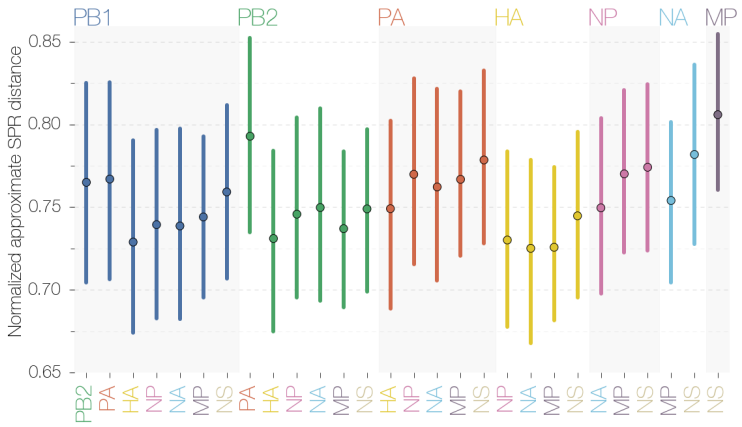

Figure S4. Normalized approximate SPR distances between pairs of segments. Following the normalization procedure approximate SPR distances are similar across all pairwise comparisons. We interpret this as lack of evidence for small-scale topological similarities between trees of all segments, which we 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).

The 95% highest posterior density (HPD) intervals of normalized approximate SPR distances between pairs of segments encompass most means and occupy a relatively small range, suggesting there is no evidence of differences in the number of reassortments between segments (Figure S4). Reassortment rate given as number of SPR moves per total time in both trees shows similar results (Figure S5). This 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 (Marshall et al., 2013). 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 S6 and S7), 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 S7) show lower SPR distances that comparisons between segments (Figure S6), 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 S4. 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 S11–S13).

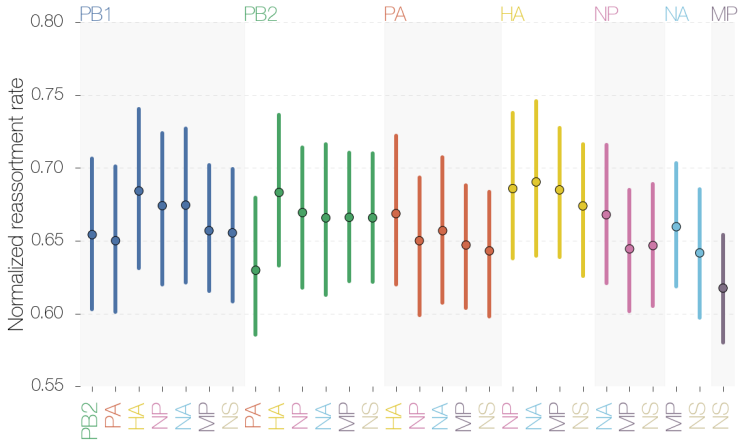

Figure S5. Normalized reassortment rate Reassortment rate is calculated as approximate number of SPR moves per sum of total time in both trees.

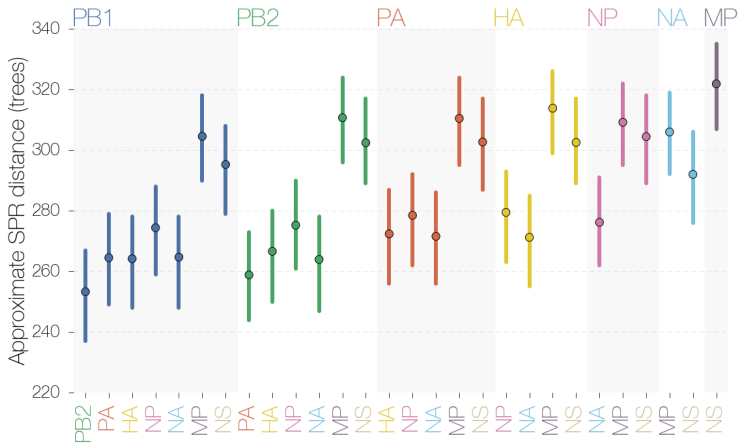

Figure S6. Approximate SPR distances between all pairs of trees of segments. There is a visible trend where comparisons between shorter segments have larger SPR distances, consistent with decreasing tree topology stability over the course of MCMC for shorter segments.

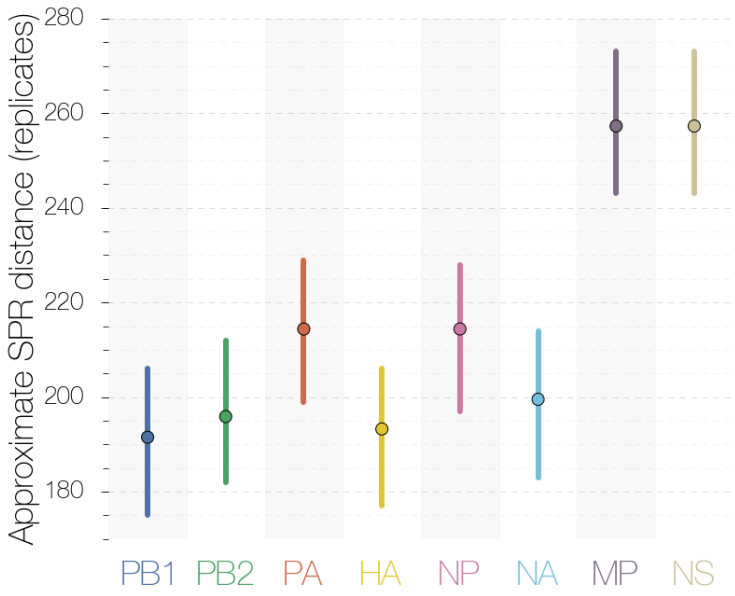

Figure S7. Approximate SPR distances between replicate trees of each segment. Approximate SPR distances between replicates of MP and NS trees are much higher (≈260) than any other segments, suggesting greater variability in tree topology over the course of MCMC. SPR distances between replicates of most other segments are ≈200.

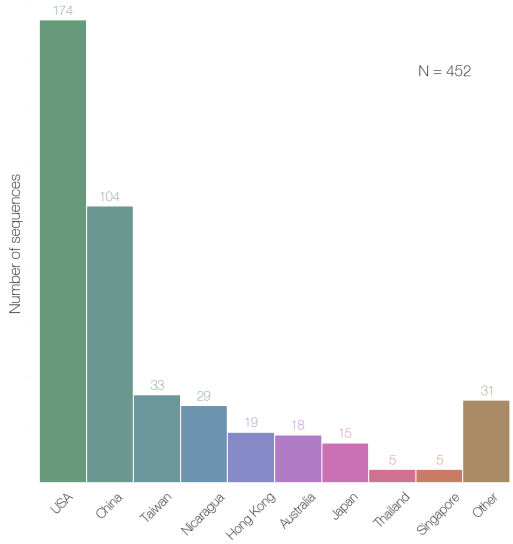

Figure S8. Geographic distribution of sequences in the primary dataset. Sequences were assigned to the "other" category if there were less than 5 sequences from that country. Most of the genomes in the primary dataset were sampled in the USA.

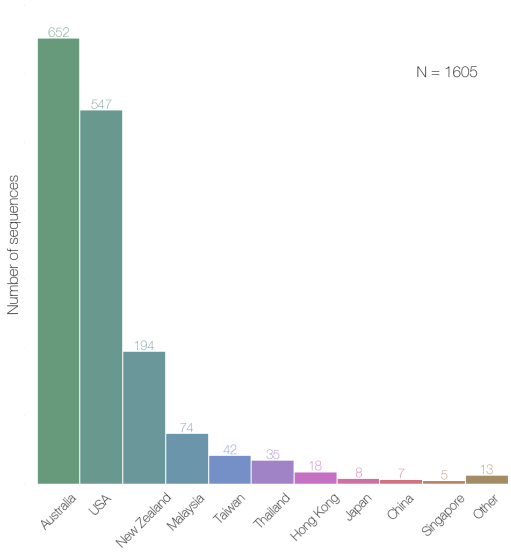

Figure S9. Geographic distribution of sequences in the secondary dataset. Sequences were assigned to the "other" category if there were less than 5 sequences from that country. Most of the genomes in the secondary dataset were sampled in Australia.

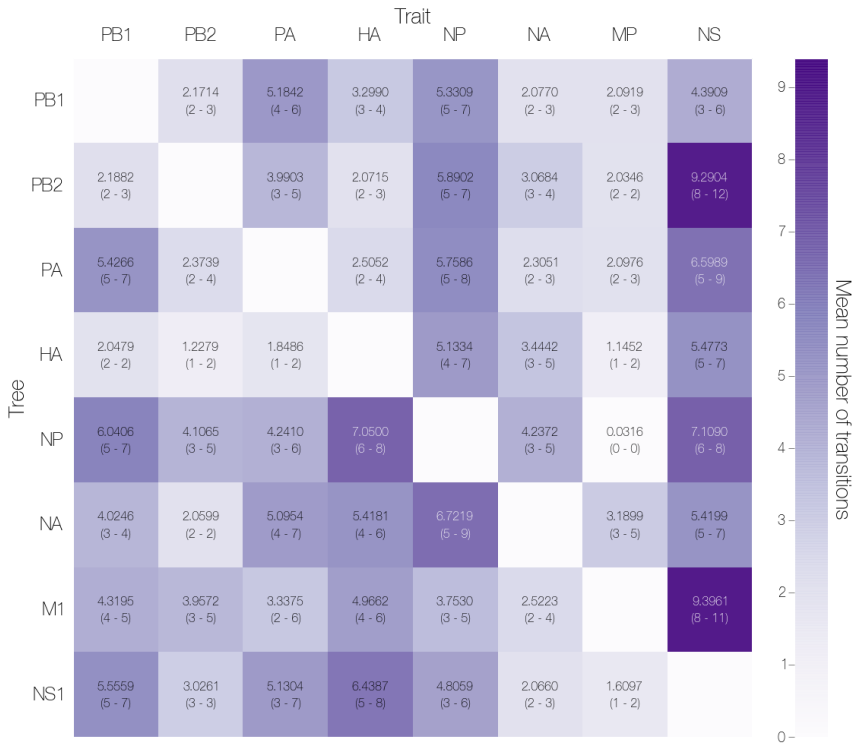

Figure S10. Numbers of trait transitions in each tree. The numbers shown are the mean inferred number of trait transitions (minus one to account for the initial Vic–Yam split) in a given tree and trait combination. The numbers in brackets correspond to 95% highest posterior density intervals. Transitions may not be independent when more than one segment reassorts at the same time.

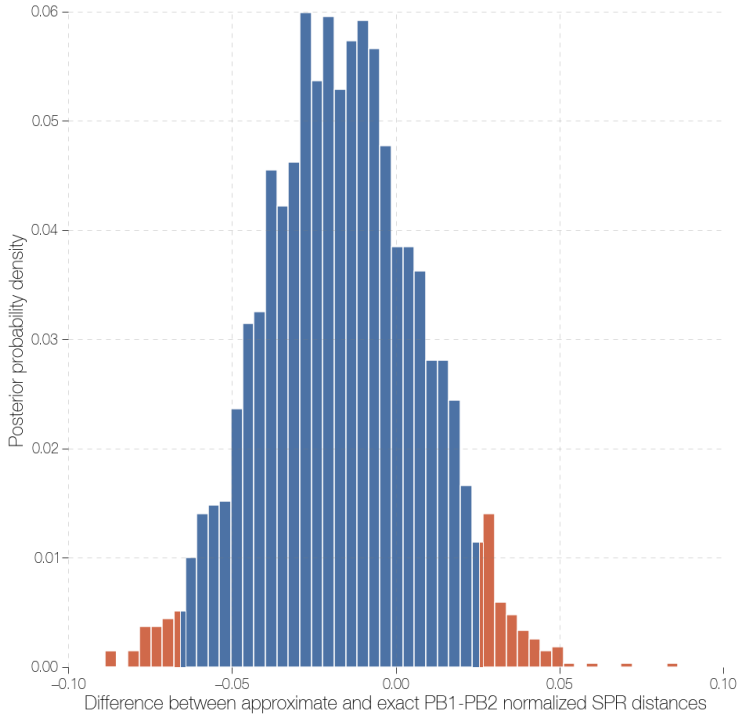

Figure S11. Distribution of differences between exact and approximate PB1-PB2 SPR distances after normalization. 95% HPD interval (blue) overlaps zero, suggesting no evidence of differences between approximate and exact SPR distances following normalization.

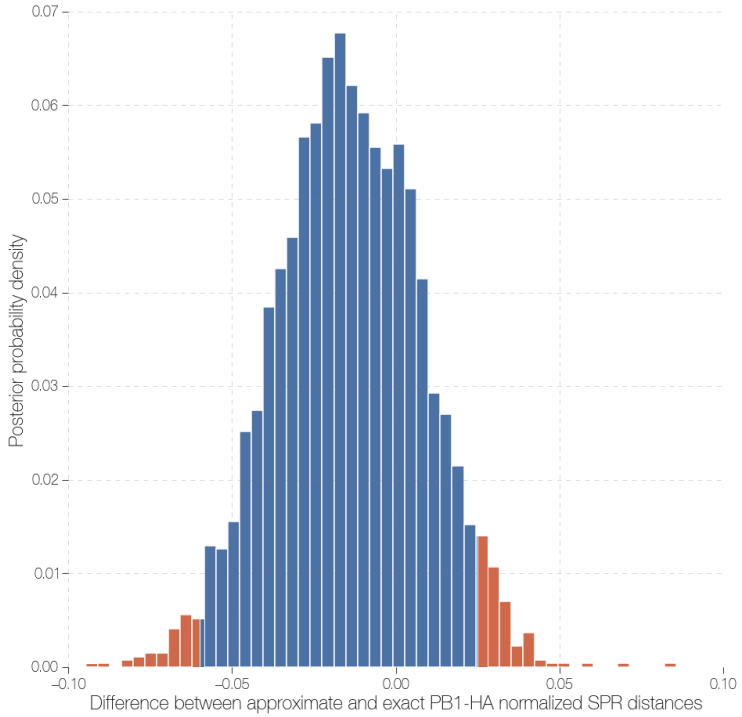

Figure S12. Distribution of differences between exact and approximate PB1-HA SPR distances after normalization. 95% HPD interval (blue) overlaps zero, suggesting no evidence of differences between approximate and exact SPR distances following normalization.

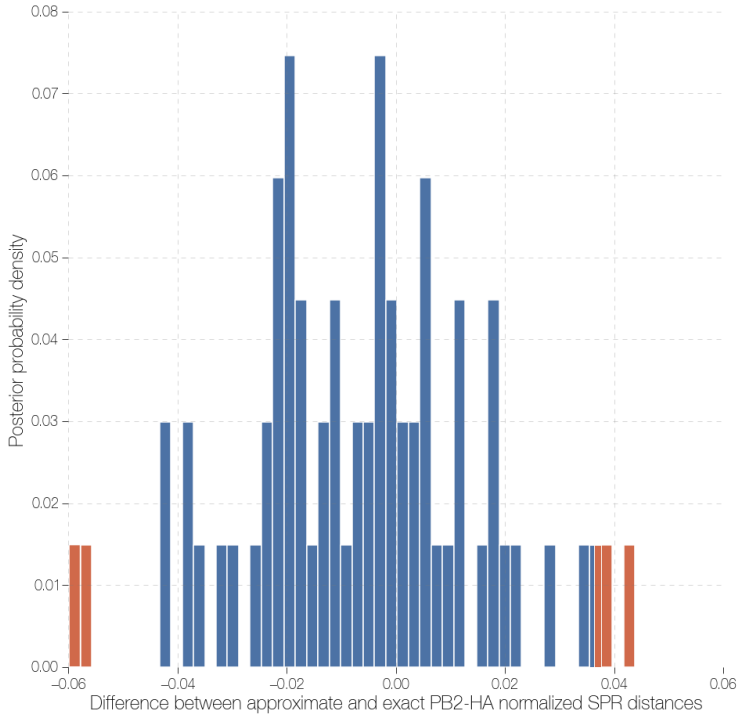

Figure S13. Distribution of differences between exact and approximate PB2-HA SPR distances after normalization. 95% HPD interval (blue) overlaps zero, suggesting no evidence of differences between approximate and exact SPR distances following normalization. Due to excessively long computation time of exact SPR distances between PB2 and HA trees few comparisons were made.

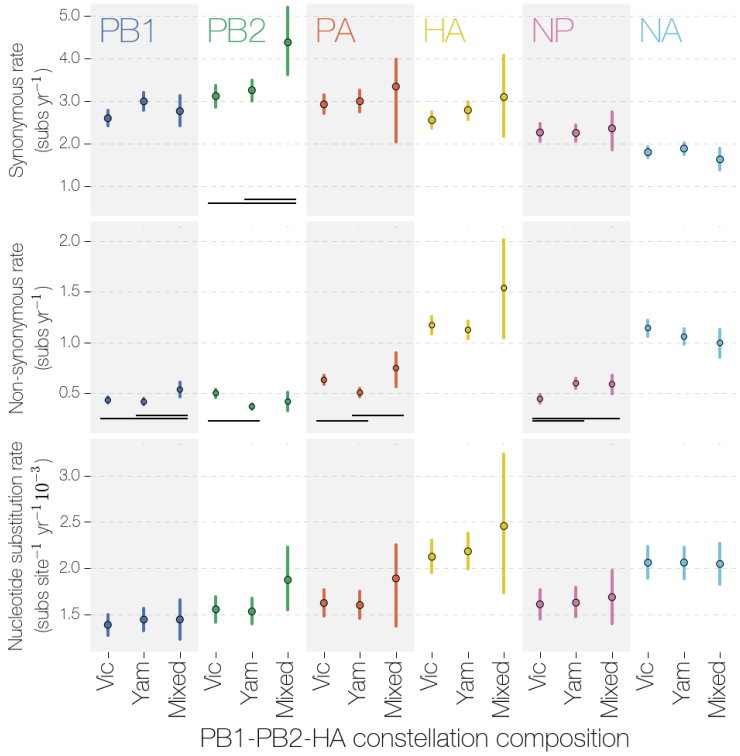

Figure S14. Synonymous, non-synonymous and nucleotide substitution rates in segments under different PB1-PB2-HA complexes. Evolutionary rate dissimilarities under Vic and Yam PB1-PB2-HA complexes are not systematic and appear negligible. Synonymous and non-synonymous rates were calculated by dividing 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.

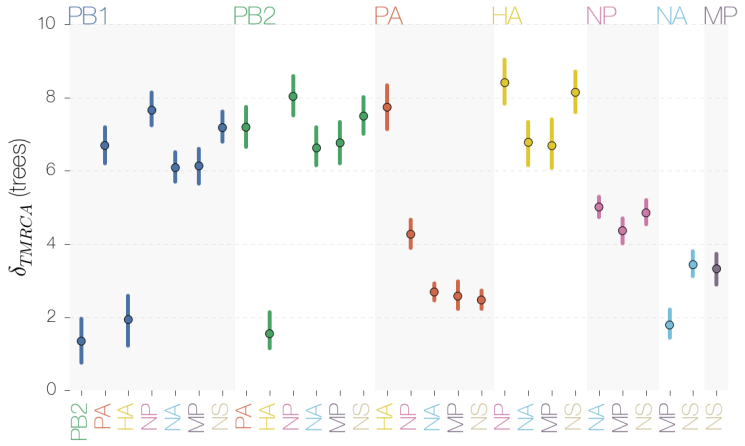

Figure S15. δ_{TMRCA} between all pairs of trees of segments. δ_{TMRCA} between trees of segments reveal that tip pairs in PB1, PB2 and HA trees have very similar TMRCAs. The upper tail of the 95% HPD (HPDs are represented as vertical lines) interval of δ_{TMRCA} values for PB1-PB2-HA and MP-NA trees do not exceed 3 years.

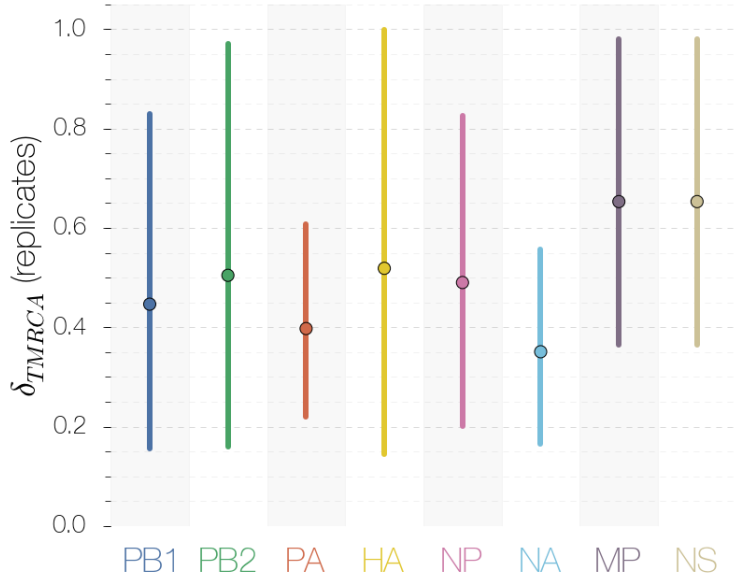

Figure S16. δ_{TMRCA} between replicate trees of each segment. δ_{TMRCA} values between independent analyses of each segment show that mean δ_{TMRCA} values rarely exceed 1 year.

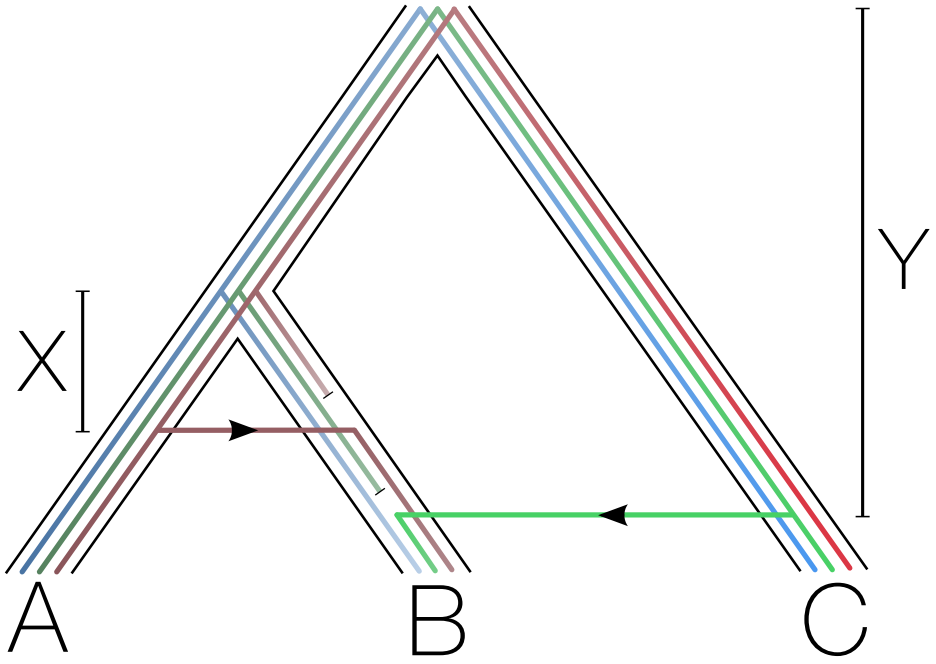

Figure S17. Calculating δ_{TMRCA} from a species tree perspective. Consider an organism that has diverged into 3 taxa (A, B, C) with a genome comprised of 3 segments (blue, green and red). Due to reassortments taxa A and B share a slightly more recent TMRCA in the red segment, likewise for taxa B and C in the green segment. By comparing differences in TMRCAs between taxa A-B, A-C and B-C in blue, red and green segments we would find that the red segment has a lower 'reassortment distance' (X) than the green segment (Y). In the absence of reassortment we expect every segment in the genome to have the same tree, *i.e.* the tree of every segment should recapitulate the 'virus' tree (analogous to 'species' trees in diploid population genetics), including the dates of nodes. Due to population bottlenecks influenza viruses go through each year we expect strains isolated at any given time to have descended from a single recent virus genome. This descent from a single genome should therefore be reflected in the TMRCAs of all segments, the only exception being reassortment, which will dramatically alter the TMRCAs of the reassorted segment tree with respect to the background onto which it is reassorting.

References

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.

Guilligay D, Tarendeau F, Resa-Infante P, et al. (11 co-authors). 2008. The structural basis for cap binding by influenza virus polymerase subunit PB2. Nature Structural & Molecular Biology. 15:500–506.

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.

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.

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

Whidden C, Beiko RG, Zeh N. 2010. Fast FPT algorithms for computing rooted agreement forests: Theory and experiments. In: Festa P, editor, Experimental Algorithms, Springer Berlin Heidelberg, number 6049 in Lecture Notes in Computer Science, pp. 141–153.

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

Whidden C, Zeh N. 2009. A unifying view on approximation and FPT of agreement forests. In: Salzberg SL, Warnow T, editors, Algorithms in Bioinformatics, Springer Berlin Heidelberg, number 5724 in Lecture Notes in Computer Science, pp. 390–402.