Appendix S1. More details of the genome sequencing procedure and the indices used to measure gene flow.

Viral Genome Sequencing. All samples were first sequenced using the traditional PCR/Sanger high-throughput sequencing pipeline at the JCVI (Dugan *et al.* 2008). After sequencing, the sequences were trimmed to remove primer sequence as well as low quality residues, and segments were assembled individually using the small genome assembler Elvira (http://elvira.sourceforge.net/). Sequencing and assembly difficulties associated with the high level of variability in the sample set (e.g. due to the mixed infection of multiple influenza subtypes) led to the processing of all samples using the next generation sequencing pipeline at JCVI that includes the 454/Roche GS-FLX and the Illumina Genome Analyzer II. Viral RNA was first reverse transcribed and amplified by multi-segment RT-PCR (M-RTPCR) (Zhou *et al.* 2009). The cDNA for each sample was primed with barcoded random hexamers using the SISPA protocol (Djikeng *et al.* 2008). One library was prepared for sequencing on the 454/Roche GS-FLX platform using Titanium chemistry while the other was made into a library for sequencing on the Illumina Genome Analyzer II.

The sequence reads from the GS-FLX data were sorted by barcode, trimmed, and searched by TBLASTX against custom nucleotide databases of full-length influenza A segments downloaded from GenBank to filter out both chimeric influenza sequences and non-influenza sequences amplified during the random hexamer-primed amplification. The filtered GS-FLX reads were then binned by segment, and *de novo* assembled using CLC Bio's clc_novo_assemble program. Because of the short read length of the sequences obtained from the Illumina Genome Analyzer II, these were not subjected to the TBLASTX filtering step. Both GS-FLX and Illumina reads were then mapped to the selected reference influenza A virus segments using the clc_ref_assemble_long program. At loci where both GS-FLX and Illumina sequence data agreed on a variation compared to the reference sequence, the latter was updated to reflect the difference. A final mapping of all next generation sequences to the updated reference sequences was then performed. All sequences generated here have been submitted to GenBank and assigned accession numbers (Table S1).

Phylogenetic Analysis

Panoramic phylogenies were estimated using a rapid hill-climbing search method and the GTRGAMMA nucleotide substitution model implemented in RAxML v7.04 (Stamatakis 2006). The tree search was initiated with a random maximum parsimony tree. 200 independent tree searches with different random maximum parsimony starting trees were performed to obtain the phylogeny with the highest likelihood score. Major lineages of North American wild bird avian influenza viruses were identified and extracted for further analyses.

For more accurate phylogenetic analysis of NA-WB-AIV lineages, we used the heuristic search method implemented in PhyML v2.4.5 (Guindon *et al*. 2009). This allowed us to determine the maximum likelihood phylogenetic tree by optimizing the tree topology and branch lengths on the sequence data. The tree search was initiated using a BIONJ tree and then employed nearest-neighbor interchange branch-swapping. All phylogenetic analyses utilized the General Time Reversible (GTR) model of nucleotide substitution which allows variable rates of substitution between the A, T, C and G nucleotides, a proportion of invariable sites (I), and four classes of rate variation among nucleotide sites ($+\Gamma_4$). The RAxML and PhyML trees were extremely similar in topology.

Pseudo-replicates for each data set were generated by resampling the columns of sequence alignment with replacement, and then used in bootstrap analyses to determine phylogenetic robustness. Because of the very large size of the ‘panoramic’ data set (>5,000 sequences), the neighbor-joining clustering method (as implemented in the PAUP* package – Swofford *et al*. 2003) was used to infer bootstrap trees in this case. In the case of the North American wild bird AIV lineage, where more phylogenetic accuracy is required, we used the maximum likelihood bootstrap method implemented in PhyML. More background information about the types of phylogenetic analyses performed here can be found in Lam *et al*. (2010).

All AIV gene segments were subject to the independent phylogenetic and phylogeographic analyses, with the exception that the only HA and NA gene segments analyzed were from subtypes H3, H4, N6, and N8. The other subtypes had HA and NA data sets of insufficient size (<200 sequences or 8 sampling localities) for meaningful phylogeographic analysis.

Measuring Gene Flow Between Populations

Three summary statistics – (i) the Fixation index (F_{ST}), (ii) modified Slatkin-Maddison's *s* (σ), and (iii) the rate of state transition (*q*) – were used to determine the level of gene flow between two localities from either the sample of virus sequences in the two localities or from the inferred phylogenetic trees.

(i) F_{ST}. F_{ST} is derived from the *F*-statistics developed by Sewall Wright to study population structure (Wright 1942). It was later modified and generalized for genetic sequence data sampled in different populations (Nei 1982; Lynch & Crease 1990). Our study employed a F_{ST} similar to that generalized by Hudson et al. (1992) using the GTR substitution model to estimate pairwise genetic distances (δ_{xy}) between two nucleotide sequences (*x* and *y*). F_{ST} between the populations sampled at geographical states *i* and *j* is defined as,

$$F_{ST} = 1 - \frac{H_w}{H_b}$$

H_w is the average within-population variation in states *i* and *j*, which is defined as,

$$H_w = \frac{\left(\frac{2}{n_i(n_i-1)} \sum_{ix<iy} \delta_{ix,iy}\right) + \left(\frac{2}{n_j(n_j-1)} \sum_{jx<jy} \delta_{jx,jy}\right)}{2}$$

where n_i and n_j are the total numbers of sequences sampled from the populations at geographical states *i* and *j*, respectively. *ix* and *iy* are the x^{th} and y^{th} sequences in the population sampled in geographical state *i*, while *jx* and *jy* are the x^{th} and y^{th} sequences in the population sampled in geographical state *j*.

H_b is the between-population variation of state *i* and *j*, which is defined as,

$$H_b = \frac{\sum_{ix,jy} \delta_{ix,jy}}{n_i n_j}$$

This is the average pairwise comparison of sequences sampled in the populations of geographical states *i* and *j*. Extreme F_{ST} estimates of 0 and 1 indicate no and complete population subdivision, respectively. F_{ST} was calculated for each pair of geographical states in the NA-WB-AIV data sets, and was plotted against the spatial distance (km) between each pair of geographical states (Fig. S2). Geographical states with too few sampled sequences ($n \leq 5$ or less than 1% of all sequences) were excluded. The PAUP* program (Swofford 2003) was used to estimate the GTR+I+Γ_4 genetic distance between sequences. A PERL script implementing the calculation of F_{ST} from the genetic distances is available upon request.

(ii) Modified Slatkin-Maddison's *s*. Slatkin and Maddison proposed and demonstrated that the minimum number of geographical state change (*s*) (between two geographical states) observed in the phylogeny provides an estimate of the level of gene flow between these populations (Slatkin & Maddison 1989; Hudson *et al.* 1992),

$$s \approx Nm$$

where *N* is the population size and *m* is the migration rate between the populations.

In situations where the sampling frequencies are highly variable among localities, the number of sequences sampled in two geographical locations is the limiting factor for the minimum number of the geographical state changes (*s*) that can be observed in the phylogeny. To overcome this problem, we used a modified Slatkin-Maddison's *s*, denoted 'σ', to infer the level gene flow between two geographical states (e.g. *A* and *B*). This is defined as,

$$\sigma = \frac{s}{s_P}$$

where *s* is number of changes between state *A* and state *B* in the observed phylogeny, and s_P is the number of changes between state *A* and state *B* occurring in the phylogeny simulated assuming panmixis. In other words, s_P determines the number of state changes that could be observed with the current phylogenetic structure and number of samples available in different geographical states. This scaling turned Slatkin-Maddison's *s* into a measure of the relative level of gene flow compared to the scenario of the completely unrestricted gene flow (i.e. panmixis). Extreme σ estimates of 0 and 1 (they may possibly but rarely exceed 1) indicate no gene flow and panmixis, respectively.

The values of *s* between different geographical states were estimated from each maximum likelihood gene phylogeny using the Fitch parsimony method (Fitch 1971) (implemented in the JAVA BEAST library). Polytomies were resolved randomly 100 times and average *s* values were obtained. To obtain s_P, each pair of geographical states was randomly shuffled at the tree tips for 1,000 iterations, and ancestral geographical states were again reconstructed using parsimony. This generated a distribution of s_P under panmixis, and the average *s* is divided by the mean s_P to give σ. To account for topological uncertainty, 200 bootstrap maximum likelihood trees were analyzed in the manner described above, and the upper and lower 95% estimates were taken as the resultant uncertainty of σ. A JAVA computer program implementing this procedure is available upon request.

To study rates of gene flow within and between migratory flyways, transitions in geographical state, e.g. Minnesota ↔ Alaska, (estimated from the phylogeny using the parsimony method described above) were categorized as transitions in flyways, e.g. MF ↔ PF, depending on which flyway the geographical state belongs to (e.g. Minnesota belongs to Mississippi Flyway (MF), Alaska belongs to Pacific Flyway (PF), see Table S2). Hence, *s* and s_P were counted for each type of transition among four flyways (i.e. MF↔PF, MF↔MF, MF↔CF, MF↔AF, PF↔PF, PF↔CF, PF↔AF, CF↔CF, CF↔AF and AF↔AF).

(iii) Rate of state transition (*q*) by Pagel's maximum likelihood method. Pagel (1994) developed a maximum likelihood method to study trait evolution (as discrete characters) along the phylogenetic tree (Pagel 1994). We employed this method and treated the geographical states as the trait at every tip. Formally, this method maximizes

$$L = f(D|T, M)$$

where *L* is the likelihood of the data set *D* (i.e. the trait states of the taxa) given the phylogeny *T,* and the model of evolution, *M*, of the trait states. In the analysis of gene flow between each pair of different geographical states, we estimated the parameters for the state transition rates (*q*) in the model. In total, there were 120 reversible transition rate parameters (e.g. $q_{\mathrm{MN}\leftrightarrow\mathrm{TX}}$, $q_{\mathrm{ND}\leftrightarrow\mathrm{OR}}$, $q_{\mathrm{SD}\leftrightarrow\mathrm{WA}}$) of interest given the 16 geographical states in the study. The huge number of parameters made the model statistically untraceable. To reduce the large number of parameters, we simplified the state transition model by reducing the rate parameters for two states to rate parameters for two flyways. This constituted the flyway-specific rate model (FRM) of AIV gene flow. In the FRM there were 10 reversible transition rate parameters for four flyways (and 6 parameters for the three flyway model (3-FRM), which combines MF and CF): $q_{\mathrm{CF}\leftrightarrow\mathrm{AF}}$, $q_{\mathrm{CF}\leftrightarrow\mathrm{PF}}$, $q_{\mathrm{CF}\leftrightarrow\mathrm{MF}}$, $q_{\mathrm{CF}\leftrightarrow\mathrm{CF}}$, $q_{\mathrm{AF}\leftrightarrow\mathrm{MF}}$, $q_{\mathrm{AF}\leftrightarrow\mathrm{PF}}$, $q_{\mathrm{AF}\leftrightarrow\mathrm{AF}}$, $q_{\mathrm{PF}\leftrightarrow\mathrm{PF}}$, $q_{\mathrm{PF}\leftrightarrow\mathrm{MF}}$, $q_{\mathrm{MF}\leftrightarrow\mathrm{MF}}$. These parameters were estimated from the phylogeny with the highest likelihood score, using the discrete maximum likelihood method described by Pagel (1994) as implemented in the APE library (Paradis *et al.* 2004) running in the R package version 2.11.1. Various starting values of transition rates were attempted (20, 10, 5, 1, 0.1 and 0.01), and the estimates with the best likelihoods were kept. The flyway-specific rate estimates are shown in Table S3.

References

Djikeng A., Halpin R., Kuzmickas R., Depasse J., Feldblyum J., Sengamalay N., Afonso C., Zhang X., Anderson N.G., Ghedin E. & Spiro D.J. (2008) Viral genome sequencing by random priming methods. *BMC Genomics*, 9, 5

Dugan V.G., Chen R., Spiro D.J., Sengamalay N., Zaborsky J., Ghedin E., Nolting J., Swayne D.E., Runstadler J.A., Happ G.M., Senne D.A., Wang R., Slemons R.D., Holmes E.C. & Taubenberger J.K. (2008) The evolutionary genetics and emergence of avian influenza viruses in wild birds. *PLoS Pathog*, 4, e1000076

Fitch W.M. (1971) Toward defining the course of evolution: minimum change for a specified tree topology. *Syst Zool*, 20, 406-16

Hudson R.R., Slatkin M. & Maddison W.P. (1992) Estimation of levels of gene flow from DNA sequence data. *Genetics*, 132, 583-9

Lam T.T., Hon C.C. & Tang J.W. (2010) Use of phylogenetics in the molecular epidemiology and evolutionary studies of viral infections. *Critical Rev Clin Lab Sci*, 47, 5-49

Lynch M. & Crease T.J. (1990) The analysis of population survey data on DNA sequence variation. *Mol Biol Evol*, 7, 377-94

Nei M. (1982) Evolution of human races at the gene level. *Prog Clin Biol Res*, 103, 167-81

Pagel M. (1994) Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proc R Soc Lond B*, 255, 37-45

Paradis E., Claude J. & Strimmer K. (2004) APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20, 289-90

Slatkin M. & Maddison W.P. (1989) A cladistic measure of gene flow inferred from the phylogenies of alleles. *Genetics*, 123, 603-13

Swofford D.L. (2003) PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). In. Sinauer Associates, Sunderland, Massachusetts.

Wright S. (1942) Isolation by distance. *Genetics*, 28, 114–38

Zhou B., Donnelly M.E., Scholes D.T., St George K., Hatta M., Kawaoka Y. & Wentworth D.E. (2009) Single-reaction genomic amplification accelerates sequencing and vaccine production for classical and swine origin human influenza a viruses. *J Virol*, 83, 10309-13.