Published in final edited form as:
Health Place. 2012 September ; 18(5): 1122–1131. doi:10.1016/j.healthplace.2012.04.009.

Population-environment drivers of H5N1 avian influenza molecular change in Vietnam

Margaret A. Carrel[a], **Michael Emch**[b], **Tung Nguyen**[c], **R. Todd Jobe**[b], and **Xiu-Feng Wan**[d]

Margaret A. Carrel: margaret-carrel@uiowa.edu; Michael Emch: emch@email.unc.edu; Tung Nguyen: nguyentungncvd@hotmail.com; R. Todd Jobe: toddjobe@unc.edu; Xiu-Feng Wan: wan@cvm.msstate.edu

[a]Department of Geography, University of Iowa, Iowa City, IA 52242 USA

[b]Department of Geography, University of North Carolina-Chapel Hill, Chapel Hill, NC 27599 USA

[c]National Center for Veterinary Diagnostics, Hanoi, Vietnam

[d]Department of Basic Sciences, College of Veterinary Medicine, Mississippi State University, Mississippi State, MS 39762 USA

Abstract

This study identifies population and environment drivers of genetic change in H5N1 avian influenza viruses (AIV) in Vietnam using a landscape genetics approach. While prior work has examined how combinations of local-level environmental variables influence H5N1 occurrence, this research expands the analysis to the complex genetic characteristics of H5N1 viruses. A dataset of 125 highly pathogenic H5N1 AIV isolated in Vietnam from 2003–2007 is used to explore which population and environment variables are correlated with increased genetic change among viruses. Results from non-parametric multidimensional scaling and regression analyses indicate that variables relating to both the environmental and social ecology of humans and birds in Vietnam interact to affect the genetic character of viruses. These findings suggest that it is a combination of suitable environments for species mixing, the presence of high numbers of potential hosts, and in particular the temporal characteristics of viral occurrence, that drive genetic change among H5N1 AIV in Vietnam.

Keywords

landscape genetics; H5N1 avian influenza; Vietnam; disease ecology

BACKGROUND

Highly pathogenic H5N1 avian influenza virus has persisted at endemic levels in poultry and human populations in Vietnam and other Asian countries since 2003. The continuous incidence and evolution of H5N1 influenza viruses is likely to be driven by complex and dynamic interactions between birds and people and the social and natural environments in which they circulate. While there exist some research efforts into which combinations of population and environment variables are related to the spatiotemporal patterns of H5N1 incidence, and a multitude of phylogeographic studies explore the molecular evolution of viruses in space and time, there has been little attention paid to how population and environment interactions affect avian influenza molecular evolution (Martin, et al, 2011, Gilbert, et al, 2008, Pfeiffer, et al, 2007, Liang, et al, 2010, Pfeiffer, et al, 2009, Wallace and

Corresponding author: Margaret Carrel, Department of Geography, University of Iowa, Iowa City, IA 52242 USA; margaret-carrel@uiowa.edu; T: 319-335-0154, F: 319-335-2725.

Fitch, 2008, Janies, et al, 2007, Wallace, et al, 2007). Prior research into H5N1 avian influenza in Vietnam has indicated a general north to south movement of viruses, with genetic diversity occurring likely as the result of isolation by distance (Carrel, et al, 2010,Wan, et al, 2008). We sought to further explore these patterns of genetic diversity in Vietnamese H5N1 viral isolates in order to understand what elements in the social and natural environments of the Vietnamese landscape were driving this genetic variation.

The emerging field of landscape genetics focuses on the interactions between evolutionary outcomes and environmental features in the belief that spatial variation in genetics indicates underlying landscape processes (Manel, et al, 2003, Storfer, et al, 2007, Balkenhol, et al, 2009, Guillot, et al, 2005). While primarily employed by biologists and ecologists exploring the genetics of plants and animal populations, there is a growing recognition that the theory and methods of landscape genetics can be used in the investigation of drivers to disease diffusion of human pathogens (Archie, et al, 2009, Criscione, et al, 2010). By combining analytic tools from landscape ecology with genetic analysis, the varying effects of environmental and population characteristics on H5N1 genetic change can be assessed.

Informing this exploration of population and environment drivers of avian influenza evolution is theory from disease ecology. The disease ecology framework that is part of the field of medical geography posits that disease outcomes are the result of complex interactions between people and their environments, and that to understand disease you must examine both the physical (environmental) and social aspects of human lives (Mayer and Meade, 1994, Mayer, 2000, Meade, 1977). Applying this theory, developed to study disease in humans, to the evolution of avian influenza viruses is appropriate, given that H5N1 avian influenza is an anthropozoonotic pathogen and that the majority of infected birds in Vietnam are living as domesticated animals in environments highly mediated by their human owners. Understanding molecular change in H5N1 avian influenza viruses as the outcome of interacting environmental and social pressures facilitates the generation of a dataset of hypothesized drivers of molecular change that are then analyzed using landscape genetics methodology (Figure 1).

DATA & METHODS

The dataset used to explore potential population and environment drivers of H5N1 avian influenza genetic change consists of 125 highly pathogenic H5N1 viruses isolated in Vietnam between 2003 and 2007 (Figure 2). Viruses were either collected by the National Centre for Veterinary Diagnostics (NCVD) of Hanoi, Vietnam or publicly available in GenBank. Each of the isolates used in the analysis had a full or nearly full genetic sequence available, as well as information regarding the province and year in which it was observed. The majority of the viruses in the dataset (110) were detected in domestic poultry such as chickens and ducks. The remaining 15 viruses were found in species such as geese and quail, as well as in environmental sampling of places where poultry live, such as soil. The NCVD collaborates with the regional offices of the Vietnamese Department of Animal Health to detect H5N1 outbreaks in backyard poultry flocks, commercial farms and live bird markets (Wan, et al, 2008).

Phylogenetic analysis of the H5N1 viruses in the dataset indicates that they share a single genetic lineage, descendant from a potential progenitor virus found in Hong Kong in 2002 (A/duck/HongKong/821/2002(H5N1)). This lineage, known as HK821-like, could result from a single introduction of the virus into Vietnam, though exactly how the introduction took place remains unknown (Wan, et al, 2008).

The most likely source of the introduction was overland trade in poultry or poultry products at Vietnam’s northern border with China (Wang, et al, 2008). The genetic distance for each

Vietnamese virus from the progenitor Hong Kong virus was calculated using PATRISTIC methods. Under the PATRISTIC framework, the degree of genetic difference between two viruses is determined by the length of the branches connecting them in a phylogenetic tree (Fourment and Gibbs, 2006). Longer branches result in higher genetic distances and indicate greater degrees of genetic change. Influenza viruses are comprised of eight gene segments which encode ten or eleven proteins, depending on the strain: hemagglutinin (HA), neuraminidase (NA), matrix proteins (MP) M1 and M2, nonstructural proteins NS1 and NS2, a nucleoprotein (NP), three polymerases (PA, PB1, and PB2), and PB1-F2. Each of these gene segments can mutate independently of the others, so eight total genetic distance measures were calculated for each of the 125 viruses (Carrel, et al, in press).

Using a geographic information system (GIS), each virus was assigned the latitude and longitude of the geographic center (centroid) of the province in which it was found. Viruses were located in 28 of Vietnam's 63 provinces (Figure 2). Then, also using the GIS, the geographic distance in kilometers was calculated between the province centroid and the centroid of Hong Kong. Temporal distance in years was calculated simply as the number of years between the progenitor virus (2002) and each of the viruses in our dataset (2003 to 2007).

Population-environment dataset creation

In addition, population and environment variables believed to be potential drivers of genetic change under a disease ecology framework (as outlined in Figure 1) were calculated for each province with an H5N1 viral occurrence (Table 1). Measures of these hypothesized drivers were gathered from the General Statistics Office of Vietnam and from several other online data sources, including: NASA's Shuttle Radar Topography Mission (SRTM30), the University of Maryland's Global Landcover Classification Facility (GLCF) and Columbia University's Center for International Earth Science Information Network (CIESIN) (Center for International Earth Science Information Network (CIESIN), 2010, General Statistics Office of Vietnam, 2010, Hansen, et al, 1998, Shuttle Radar Topography Mission (SRTM), 2009).

The circulation of the human population of Vietnam could influence genetic variation of H5N1 viruses via the movement of poultry between farm and market or the movement of poultry products across the country. Larger human populations also increase the odds of interaction between people and birds, and increase the probability of viruses being transferred across space. Four variables, human population density, passenger traffic, and road and water freight, were included to test these associations.

Measures of the number of rural residents in each province acts as a proxy for the number of people engaged in agriculture that makes use of an integrated farming system including Vuon (agricultural plots), Ao (ponds), and Chuong (caged birds). In this system the droppings of poultry are used in farming fish and to fertilize crops, while the birds themselves are used to consume insect pests in fields (Cristalli and Capua, 2007). The number of urban residents in a province indicates regions of high population circulation, with people moving between cities and rural regions, as well as areas of concentration of live bird markets selling rural-raised poultry to city-based consumers. Measures of income, high school education and medical professionals in a province allow for the testing of hypotheses that socioeconomic status, hygiene practices, knowledge of influenza contamination and spread, access to vaccination and veterinary care, and access to human health care can act as drivers of molecular evolution, by influencing whether humans permit the virus to persist and spread through home environments. Socioeconomic status and high school education are closely linked variables, and can also influence viral evolution via the likelihood of a person to report sick poultry or cull sick flocks.

The number of susceptible hosts, as measured by provincial poultry density, and the number of potential intermediary hosts, as measured by pig density, could act to drive molecular change via increased chances of infection or viral exchange. Finally, spaces in which susceptible and infected hosts can exchange viruses, or spaces where humans can come into contact with and subsequently spread viruses, include water surfaces of varying types, including aquaculture ponds, wet rice agricultural plots, and lakes, ponds or streams (classified as water surface per province). Areas of low elevation are more likely to host wet rice agricultural land, and to have more spaces of species interaction, and paddy areas with higher rice yields can indicate double or triple cropping and thus more time per year covered in water. All human and environment variables, as generated from a disease ecology perspective, were tested for their relationship to H5N1 viral evolution.

Because viruses were geocoded at the province level, the population and environment variables were also scaled to the province. Viral genetic characteristics were then associated with these population and environment variables on the basis of their province of isolation and their year of isolation (for those population and environmental variables where annual-specific numbers were available, such as high school graduation rate). Thus, for every virus there were eight genetic distance measures, temporal and geographic distance from the progenitor virus, and eighteen hypothesized population and environment independent variables.

Ordination analysis

Non-metric multidimensional scaling (NMDS) is one of several ordination techniques that can be used to visualize and explore the underlying structure of multiple dependent variables, and to further relate these structures to independent predictor variables. In NMDS, the object is to find a configuration for *n* points (the 125 viruses) in multidimensional space such that the space between points closely corresponds to the observed dissimilarities measured in *p* elements (the 8 genetic distances calculated as branch lengths). Using an *n* by *p* input matrix (125 by 8), a symmetrical *n* by *n* matrix of all pairwise distances is calculated, in this case with a Euclidean distance measure. Each pairwise distance summarizes the amount of difference between the 125 viruses across all eight genetic measures.

The exact configuration of the points in the final ordination is the result of an iterative process. Distances among the *n* points in the initial configuration are regressed against the original distances in the *n* by *n* matrix using a non-parametric approach fitted by least-squares. A perfect ordination of the points would exhibit an exact match of the ordinated points on the regression line. Stress, or goodness of fit, measures how well the distances between ordinated points correspond to the distances calculated from in the original *n* by *n* matrix. Stress is most commonly calculated as:

$$\sqrt{\sum_{h,i}(d_{hi}-\hat{d}_{hi})^2/\sum_{h,i}d_{hi}^2}$$

where d_{hi} is the ordinated distance between two samples and $\hat{d_{hi}}$ is the distance predicted from the regression (Kruskal, 1964a, Steyvers, 2006). Ordinated points are then moved by small amounts to decrease stress and increase the fit against a re-calculated regression line. This process continues until no further movement of the ordinated *n* points results in a reduction in stress.

The dimensionality of NMDS is an expression of the axes of variation within the data. The optimum number of dimensions used in the NMDS is chosen to minimize stress without

compromising the utility of the scaling process. Too few dimensions will mask variation in the dataset, while too many will split one axis of variation across multiple axes. By plotting stress against dimensions, the point at which adding axes of variation does little to reduce stress is an indication of how many dimensions should be used in the analysis (Kruskal, 1964a). Stress values of 20% and above indicate a poor fit for the data, 10% indicate a fair fit, 5% are good and anything less than 2.5% is excellent (with 0% being a perfect match between the observed dissimilarities and the ordinated dissimilarities) (Kruskal, 1964b).

Once the H5N1 viruses were ordinated according to their genetic characteristics, each of the eighteen population and environment variables, as well as the geographic and temporal distance variables, were associated with the ordination. Each variable is aligned in the viral ordination space in the direction of its most rapid change and where its correlation with the ordination configuration is maximal. A goodness of fit statistic (the squared correlation coefficient, or R^2) is calculated via permutation analysis. Arrow lengths for each population and environment variable indicate goodness of fit scores (i.e. longer arrows for higher R^2). These scores were plotted and variables with values greater than 0.10 were taken to be the most important drivers of genetic change among H5N1 viruses. This cutoff was chosen as an initial way to cull the number of independent predictors of H5N1 genetic change that the study would focus on.

Cluster analysis & linear regression

Clustering techniques were used to assign the ordinated viruses into like groups. The number of clusters that the ordinated points were divided into was chosen to optimize the similarity of points within the cluster and maximize the difference of points between clusters. The Partana ratio measures the within-cluster to among-cluster similarity of classifications, while the silhouette width is a measurement of the mean similarity of each object to the other objects in its cluster, compared to its mean similarity to the most similar cluster. The NMDS ordination was thus classified into the number of clusters at which both the Partana ratio and the silhouette width increased. Subsequently, variation in the variables with scores greater than 0.10 was assessed across the clusters. The relative importance of each of these variables in assigning viruses to clusters was then assessed as an indication of how each differentiation across the range of the predictor variable values corresponded to differentiation in cluster assignments. In other words, which variables seemed to be most associated with the division of viruses into clusters?

While fitting the environmental variables onto the ordination and then examining the influence of the environmental variables across clusters indicates the strength of relationships, it does not indicate the direction of relationships. Linear regression was used to explore the direction and statistical significance of the relationship between the NMDS variables and genetic outcomes. A three-dimensional NMDS was found to be the optimal data configuration, so each of the three NMDS axes scores for the 125 viruses comprised the outcome variable, and predictor variables included in the initial model were all those with $R^2>.10$. Variables were then discarded if they exhibited high multicollinearity with other variables (as indicated by Variance Inflation Factor (VIF) scores of 6 or greater. The choice of which variables to retain among those that exhibited multicollinearity was made based upon improved R^2. Non-significant variables were then discarded iteratively and if their exclusion from the model resulted in improved model fit, as measured by the Akaike Information Criterion (AIC) and the Log-Likelihood Ratio (LLR), they were excluded from the final model.

NMDS, fitting the population and environment variables to the ordination and clustering analysis was carried out in R2.9.2 using the *labdsv*, *vegan*, *MASS*, *optpart*, and

randomForest packages (R Development Core Team, 2011). Regression was conducted in SAS®9.1.3(SAS Institute Inc., 2008).

RESULTS

For the H5N1 influenza genetic dataset, at three dimensions stress is minimized (3.8%) without adding more unnecessary dimensions (Figure 3).

A plot of the observed *n* by *n* matrix versus the ordinated differences (Figure 4) indicates that the final three-dimensional NMDS ordination well-represents the measured genetic distances in the viral dataset.

Fitting the population-environment variables to the ordination

The first two dimensions in the final ordination appear in Figure 5A, and all eighteen population and environment variables hypothesized to relate to genetic differentiation among H5N1 avian flu viruses, along with indicators of the temporal and genetic relationships to the progenitor virus, are fitted into the ordination in Figure 5B. Many of the population-environment variables have the same alignment in the ordinated space (clustered to the right side of the chart), while temporal distance from the progenitor virus (Temporal Distance), the amount of geographic distance between viruses and the progenitor virus (Geographic Distance), and the amount of surface devoted to aquaculture in a province (Aquaculture) have their own distinct axes through the ordinated space.

Plotting the R^2 calculated for each of the population-environment variables indicated that only five had scores of greater than 0.2 (Figure 6). These five variables are temporal distance, geographic distance, aquaculture surface, population density and high school graduation rate. Three other variables had R^2 of greater than 0.15: pig population density, poultry population density and road freight. Five variables had R^2 of greater than 0.10: percent water surface in a province, percent built surface, paddy yield, average income of the lowest quintile income individuals and rural population size.

Plotting only these thirteen variables onto the ordination indicates their differing strengths and relationships to the scaled genetic distances (Figure 7). As mentioned above, temporal distance, geographic distance and aquaculture surface have long axes of differentiation through the lower scores on the first dimension and the higher scores on the second dimension. The other ten variables have closely adjoining directions of change through the ordination, towards the higher scores on the first dimension and the zero range on the second dimension. The axes for high school graduation rate and population density are the longest, reflecting their larger R^2.

Cluster assignment

Each of the 125 ordinated points was assigned to a cluster to examine how the influence of these thirteen independent variables with R^2 greater than 0.10 differed within the dataset. Seven clusters were defined, based upon the number at which both the Partana ratio and the silhouette width was maximized (Figure 8). Examining the cluster patterns (Figure 9) we see that one virus has also been assigned to its own cluster based on its distance from all other viruses in ordination space across all three dimensions. Examining the genetic characteristics of this virus indicates that it is a duck isolate and that it has the highest genetic distance from the progenitor Hong Kong virus on the HA and PB1 gene segments. Box plots are used to display how each of the thirteen independent variables differ among clusters (Figure 10).

From the boxplots, it appears that viruses within each of the clusters (grouped according to their place in the ordination) have different interactions with the thirteen population-environment and geographic and temporal variables. The cluster assignments closely follow the temporal characteristics of the viruses, as seen in the Temporal Distance boxplot, wherein five of the seven clusters have viruses all isolated in the same year, and the remaining three clusters include viruses isolated within a year of one-another. The single virus in Cluster 3 is associated with high geographic and temporal distance, it was isolated in a southern province in 2007, but low population density and poultry and pig populations, as well as low high school graduation. Cluster 2, in contrast, has viruses in provinces with high human, pig and poultry populations and very high rates of high school graduation. Some independent variables show high divergence across cluster assignments (population density, rural population) while others have similar values across clusters (aquaculture, low income quintile).

Figure 11 provides a visual representation of how each variable relates to the overall similarity of viruses included in each cluster. Greater decreases in the Gini index indicate that splitting ordination points according to that variable results in better in-cluster similarity and between-cluster dissimilarity. Thus, splitting ordination points according to the temporal distance variable most improves the cluster assignments, while differentiating points by human, population density, paddy yield or road freight characteristics has less effect. This would indicate that temporal distance has the greatest association with viral ordination, while the road freight variable has much less so. These Gini measures were used to assess the reliability of the R^2 found when fitting the environmental variables. Overall, there is general agreement between those variables with higher R^2 when fit to the ordination plot and those variables with high Gini measures when the ordinated points are clustered.

Regression results

All thirteen variables with R^2 of >0.10, regardless of their importance indicated in Figure 11, were included in the initial regression. Based upon VIF scores of more than 6, the pig and poultry population variables were iteratively removed, and neither had significant interactions with the dependent variables and their exclusion improved model fit so they were dropped in the final model. Four other variables were removed from the final model, based upon non-significant interactions with all three outcome variables and decreased AIC and increased LLR with their elimination. The final models included seven predictor variables (Table 2).

Dimension scores for each virus indicate the amount of genetic difference across all eight gene segments. Thus, a positive relationship between predictor and outcome variables indicates increased difference in ordination scores, while a negative direction indicates closer ordination scores (i.e. viruses that are more similar genetically so more similar in ordination space). As the amount of land devoted to aquaculture in a province decreases, so does the genetic differentiation among viruses. As high school graduation in a province increases, genetic differentiation increases. Population density, on the second NMDS dimension, is significantly and positively related to genetic difference. Temporal distance is a statistically significant, with the largest coefficients, predictor of genetic difference across all three dimensions, though the direction of the relationship varies from dimension to dimension. In the first and third dimensions, increased amount of surface water is significantly associated with increased genetic differentiation, and on the first dimension the income of the poorest populations in a province are positively associated with genetic distance.

DISCUSSION

This paper describes a new approach for understanding spatio-temporal distributions of influenza by investigating the fundamental population and environmental drivers of viral evolution. Differentiation among the eight gene segments of H5N1 avian influenza viruses is most associated with a combination of seven population-environment variables and temporal characteristics: the amount of aquaculture in a province, the high school graduation percentage in a province, the population density of a province, the surface covered in water in a province, the socioeconomic status of provincial residents and the amount temporal distance between viruses and the Hong Kong progenitor virus. These population and environment characteristics associated with genetic differentiation are similar to those associated with H5N1 incidence, although poultry population and wet-rice agriculture were not as significant in our analysis as in incidence studies (Martin, et al, 2011, Gilbert, et al, 2008, Paul, et al, 2010).

As the amount of land devoted to aquaculture decreases, genetic differentiation increases. This result is surprising, given that aquaculture practices and areas have been previously found to be associated with H5N1 incidence and perpetuation of the virus in the environment (Pfeiffer, et al, 2007, Cristalli and Capua, 2007). However, the amount of water surface area in a province was positively associated (with very large coefficients) with genetic differentiation, capturing the fact that surface water sites are zones of interaction among infected and uninfected poultry, where the water surface provides a medium for fecal-oral transmission of the virus (Brown, et al, 2007).

As high school graduation increases, genetic differentiation increases: high school graduation in this case is a proxy for general education levels. Education has been shown to have an effect on hygiene behaviors in households and their knowledge about how influenza is transmitted (Dinh, et al, 2006, Thorson, et al, 2006). In areas where education levels are higher, H5N1 viruses not only are incident but are also more genetically different. High school graduation rates may also be taken as a proxy for socioeconomic status, with richer provinces having higher graduation levels. We observed that as the income of the poorest population quintile in provinces increased, in other words as the socioeconomic status of the poor went up, so too did genetic differentiation. This, combined with the finding about high school graduation status, is counter-intuitive. One would expect that poorer provinces with lower high school graduation rates would be associated with greater amounts of H5N1 genetic diversity if farmers could not afford to treat their flocks or improperly treated them. But we speculate that these findings are capturing the fact that wealthier and better-educated farmers actually raise more poultry and wealthier local populations have greater demand for duck and chicken products (meat, eggs, etc.). These findings are also perhaps capturing the wealthier and better-educated populations in Vietnam’s cities, the regions around which are associated with outbreaks of H5N1 influenza.

This speculation is supported by findings that as population density in a province increases, so does genetic differentiation: more people means more opportunities for mixing and movement of viruses in the landscape as people travel among farms and from homes to markets and back again. Areas of high human population density are also correlated with high domestic poultry population densities, and have previously been associated with H5N1 risk (Gilbert, et al, 2008, Pfeiffer, et al, 2007, Paul, et al, 2010).

Temporal distance from the progenitor virus was the single strongest predictor of genetic differentiation in the NMDS, the clustering algorithm and the regression analyses. As time goes by, genetic change among viruses increases on one of the three NMDS dimensions, but decreases on the other two. This is representative of the process by which viruses gradually

evolve, such that some viruses isolated in the same year are genetically very different while others isolated years apart are genetically very similar. While genetic change can be very rapid, certain genetic sequences can also remain established in viral populations for long periods of time. The role of temporal distance as uncovered by this analysis highlights the fact that landscape-level processes of genetic change are always going to be mediated by the effects of time, and that any consideration of population-environmental drivers of pathogenic evolution that do not control for this influence will be seriously flawed.

Several variables were dropped from the final model, and several were not included in regression modeling because their relationship to the viral ordination was weak. Pig and poultry population densities, though they had high R^2, had relatively low impact on decreasing the Gini index when ordinated points were clustered. Their exclusion from the final model improved how well the ordinated genetic distance measures were predicted. This result is surprising, given the hypothesized effect that increased numbers of susceptible hosts or intermediate hosts for H5N1 avian influenza would increase genetic differentiation. It is possible that the inclusion of human population density can capture this effect, however, given that domestic poultry populations are associated with the presence of human farmers. Road freight, as well as the other circulation variables of passenger traffic and water freight, were not important drivers of molecular differentiation. Similarly, the measures of rural versus urban populations were not significantly associated with the viral ordinations. This suggests that, while overall population density is important, these indicators of cities and population movement are less correlated with viral diversity. It is also noteworthy that the only space of species mixing that was found to influence genetic differentiation was that of aquaculture and percent water surface, that the area in a province devoted to wet rice agriculture (and paddy yield and elevation), was not linked to high levels of genetic change. Finally, the number of medical professionals in a province, a proxy for access to care, evidenced no significant relationship.

This ecological analysis was conducted at the provincial level, based upon the availability of spatial data for each of the 125 viral isolates. Such a scalar limitation is not uncommon in the newly emerging field of landscape genetic analysis of human pathogens. The scale at which human pathogen data is collected and released is often crude from a geographic perspective, and necessitates the aggregation of cases to the centroids of spatial units and the aggregation of predictor variables to a similar scale. If the specific latitude/longitude of the 125 H5N1 cases used in this study was known, then greater understanding of specific landscape-level drivers of genetic change could be examined, with potentially different results from those presented in this paper. The modifiable areal unit problem (MAUP) is one that continually challenges researchers seeking to understand the landscape drivers of pathogenic evolution. It is hoped that just as there is increasing availability of high-resolution spatial data there will be increasing collection and distribution of geographically precise disease data.

The strong and significant relationships that were found between the ordinated viral distances and the population environment datasets indicate that areas with high population densities, non-specifically rural or urban, and relatively high income and education levels, as well as environmental sites where avian species can readily exchange viruses, are areas where genetic differences among viruses are the greatest. These social and environmental variables are mediated, however, by the dominant influence of time on molecular evolution. Examining genetic differentiation rather than simply incidence is important, given the potential for viruses to develop the ability to jump species barriers and increase pathogenicity as they evolve. Additionally, using a disease ecology perspective to frame the study allows the findings to be informed by theories about the ways in which human interactions with avian populations in natural and social environments can affect evolution

of H5N1 viruses, and, when combined with landscape genetics, can potentially be extended to study the evolution of other anthropozoonotic pathogens.

Acknowledgments

This work was supported by National Science Foundation (NSF) Award BCS-0717688, as well as by the NSF Graduate Research Fellowship Program and the NSF IGERT Program at the Carolina Population Center at the University of North Carolina at Chapel Hill. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Archie EA, Luikart G, Ezenwa VO. Infecting epidemiology with genetics: a new frontier in disease ecology. Trends in Ecology & Evolution. 2009; 24:21–30.10.1016/j.tree.2008.08.008 [PubMed: 19027985]

Balkenhol N, Gugerli F, Cushman SA, Waits LP, Coulon A, Arntzen JW, Holderegger R, Wagner HH. Identifying future research needs in landscape genetics: where to from here? Landscape Ecology. 2009; 24:455–463.

Brown JD, Swayne DE, Cooper RJ, Burns RE, Stallknecht DE. Persistence of H5 and H7 avian influenza viruses in water. Avian Diseases. 2007; 51:285–289. [PubMed: 17494568]

Carrel MA, Emch M, Jobe RT, Moody A, Wan XF. Spatiotemporal structure of molecular evolution of H5N1 highly pathogenic avian influenza viruses in Vietnam. PloS One. 2010; 5:e8631.10.1371/journal.pone.0008631 [PubMed: 20072619]

Carrel MA, Wan XF, Nguyen T, Emch M. Genetic Variation of Highly Pathogenic H5N1 Avian Influenza Viruses in Vietnam Shows Both Species-Specific and Spatiotemporal Associations. Avian Diseases. 2011; 55:659–666. [PubMed: 22312987]

Center for International Earth Science Information Network (CIESIN). Gridded Population of the World, version 3 (GPWv3). 2010.

Criscione CD, Anderson JD, Sudimack D, Subedi J, Upadhayay RP, Jha B, Williams KD, Williams-Blangero S, Anderson TJC. Landscape Genetics Reveals Focal Transmission of a Human Macroparasite. PLoS Neglected Tropical Diseases. 2010; 4:308–332.

Cristalli A, Capua I. Practical problems in controlling H5N1 high pathogenicity avian influenza at village level in Vietnam and introduction of biosecurity measures. Avian Diseases. 2007; 51:461–462. [PubMed: 17494607]

Dinh PN, Long HT, Tien NT, Hien NT, Mai le TQ, Phong le H, Tuan le V, Van Tan H, Nguyen NB, Van Tu P, Phuong NT. World Health Organization/Global Outbreak. Alert and Response Network Avian Influenza Investigation Team in Vietnam. Risk factors for human infection with avian influenza A H5N1, Vietnam, 2004. Emerging infectious diseases. 2006; 12:1841–1847. [PubMed: 17326934]

Fourment M, Gibbs M. PATRISTIC: a program for calculating patristic distances and graphically comparing the components of genetic change. BMC Evolutionary Biology. 2006; 6:1. [PubMed: 16388682]

General Statistics Office of Vietnam. Statistical Data. Government of Vietnam; 2010.

Gilbert M, Xiao X, Pfeiffer DU, Epprecht M, Boles S, Czarnecki C, Chaitaweesub P, Kalpravidh W, Minh PQ, Otte MJ, Martin V, Slingenbergh J. Mapping H5N1 highly pathogenic avian influenza risk in Southeast Asia. Proceedings of the National Academy of Sciences of the United States of America. 2008; 105:4769–4774.10.1073/pnas.0710581105 [PubMed: 18362346]

Guillot G, Estoup A, Mortier F, Cosson JF. A Spatial Statistical Model for Landscape Genetics. Genetics. 2005; 170:1261–1280.10.1534/genetics.104.033803 [PubMed: 15520263]

Hansen, M.; DeFries, R.; Townshend, JRG.; Sohlberg, R. UMD Global Land Cover Classification, 1 Kilometer, 1.0. Department of Geography, University of Maryland; College Park, Maryland: 1998. p. 1981-1994.

Janies D, Hill AW, Guralnick R, Habib F, Waltari E, Wheeler WC. Genomic analysis and geographic visualization of the spread of avian influenza (H5N1). Systematic Biology. 2007; 56:321–329.10.1080/10635150701266848 [PubMed: 17464886]

Kruskal JB. Nonmetric multidimensional scaling: a numerical method. Psychometrika. 1964a; 29:115–129.

Kruskal JB. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika. 1964b; 29:1–27.

Liang L, Xu B, Chen Y, Liu Y, Cao W, Fang L, Feng L, Goodchild MF, Gong P, Li W. Combining Spatial-Temporal and Phylogenetic Analysis Approaches for Improved Understanding on Global H5N1 Transmission. PloS One. 2010; 5:e13575. [PubMed: 21042591]

Manel S, Schwartz MK, Luikart G, Taberlet P. Landscape genetics: combining landscape ecology and population genetics. Trends in Ecology & Evolution. 2003; 18:189–197.10.1016/S0169-5347(03)00008-9

Martin V, Pfeiffer DU, Zhou X, Xiao X, Prosser DJ, Guo F, Gilbert M. Spatial distribution and risk factors of highly pathogenic avian influenza (HPAI) H5N1 in China. PLoS Pathogens. 2011; 7:e1001308.10.1371/journal.ppat.1001308 [PubMed: 21408202]

Mayer JD. Geography, ecology and emerging infectious diseases. Social science & medicine. 2000; 50:937–952. [PubMed: 10714918]

Mayer JD, Meade MS. A reformed medical geography reconsidered. The Professional Geographer. 1994; 46:103–106.

Meade MS. Medical Geography as Human Ecology: The Dimension of Population Movement. The Geographical Review. 1977; 67:379–393.

Paul M, Tavornpanich S, Abrial D, Gasqui P, Charras-Garrido M, Thanapongtharm W, Xiao X, Gilbert M, Roger F, Ducrot C. Anthropogenic factors and the risk of highly pathogenic avian influenza H5N1: prospects from a spatial-based model. Veterinary research. 2010:41.

Pfeiffer DU, Minh PQ, Martin V, Epprecht M, Otte MJ. An analysis of the spatial and temporal patterns of highly pathogenic avian influenza occurrence in Vietnam using national surveillance data. The Veterinary Journal. 2007; 174:302–309.10.1016/j.tvjl.2007.05.010 [PubMed: 17604193]

Pfeiffer J, Pantin-Jackwood M, To TL, Nguyen T, Suarez DL. Phylogenetic and biological characterization of highly pathogenic H5N1 avian influenza viruses (Vietnam 2005) in chickens and ducks. Virus research. 2009; 142:108–120.10.1016/j.virusres.2009.01.019 [PubMed: 19428743]

R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; Vienna, Austria: 2011.

SAS Institute Inc. SAS®9.1.3 for Windows. Cary, North Carolina: 2008.

Shuttle Radar Topography Mission (SRTM). Global 30 Arc-Second Elevation Dataset (GTOPO30). 2009.

Steyvers, M. Encyclopedia of Cognitive Science. 2006. Multidimensional Scaling.

Storfer A, Murphy MA, Evans JS, Goldberg CS, Robinson S, Spear SF, Dezzani R, Delmelle E, Vierling L, Waits LP. Putting the “landscape” in landscape genetics. Heredity. 2007; 98:128–142.10.1038/sj.hdy.6800917 [PubMed: 17080024]

Thorson A, Petzold M, Nguyen TK, Ekdahl K. Is exposure to sick or dead poultry associated with flulike illness?: a population-based study from a rural area in Vietnam with outbreaks of highly pathogenic avian influenza. Archives of Internal Medicine. 2006; 166:119–123.10.1001/archinte.166.1.119 [PubMed: 16401820]

Wallace RG, Fitch WM. Influenza A H5N1 immigration is filtered out at some international borders. PLoS ONE. 2008; 3:e1697.10.1371/journal.pone.0001697 [PubMed: 18301773]

Wallace RG, Hodac H, Lathrop RH, Fitch WM. A statistical phylogeography of influenza A H5N1. Proceedings of the National Academy of Sciences of the United States of America. 2007; 104:4473–4478.10.1073/pnas.0700435104 [PubMed: 17360548]

Wan XF, Nguyen T, Davis CT, Smith CB, Zhao ZM, Carrel M, Inui K, Do HT, Mai DT, Jadhao S, Balish A, Shu B, Luo F, Emch M, Matsuoka Y, Lindstrom SE, Cox NJ, Nguyen CV, Klimov A, Donis RO. Evolution of highly pathogenic H5N1 avian influenza viruses in Vietnam between 2001 and 2007. PLoS ONE. 2008; 3:e3462.10.1371/journal.pone.0003462 [PubMed: 18941631]

Wang J, Vijaykrishna D, Duan L, Bahl J, Zhang JX, Webster RG, Peiris JS, Chen H, Smith GJ, Guan Y. Identification of the progenitors of Indonesian and Vietnamese avian influenza A (H5N1)

viruses from southern China. Journal of virology. 2008; 82:3405–3414.10.1128/JVI.02468-07 [PubMed: 18216109]

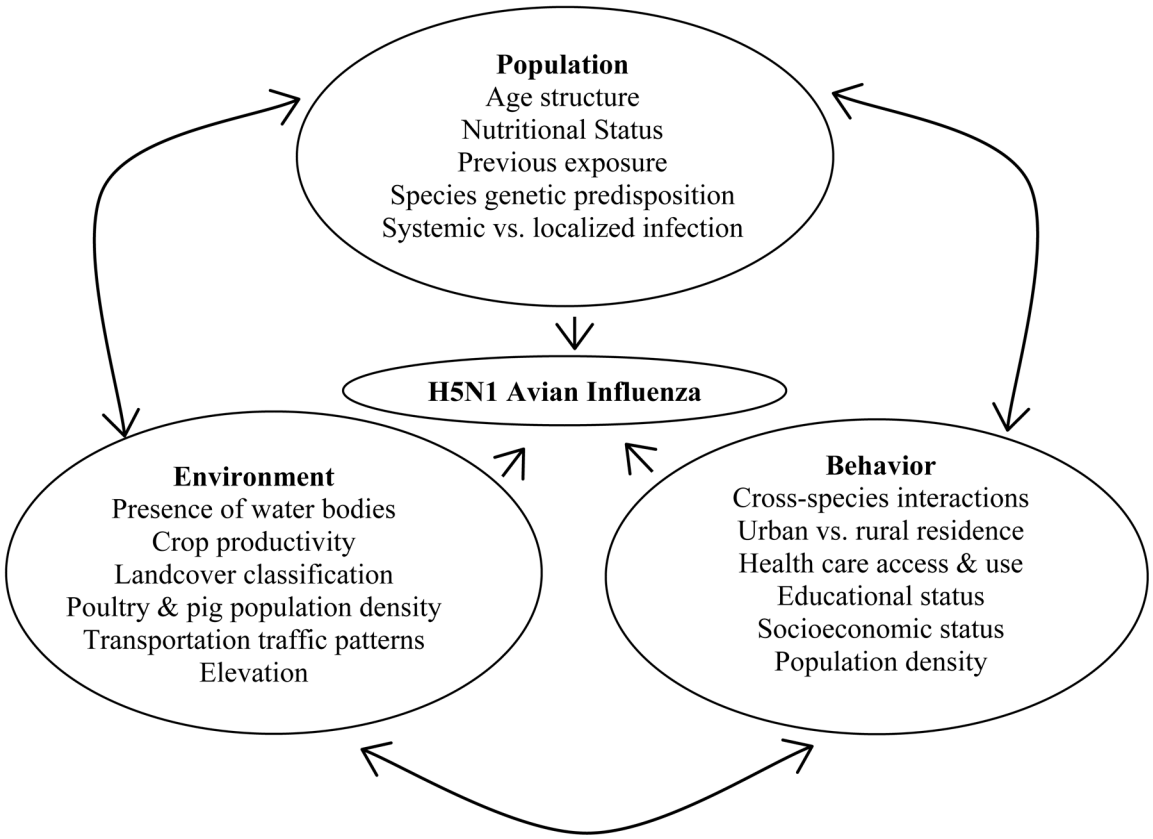

Figure 1.
Framework describing the disease ecology of H5N1 avian influenza in Vietnam's domestic poultry.

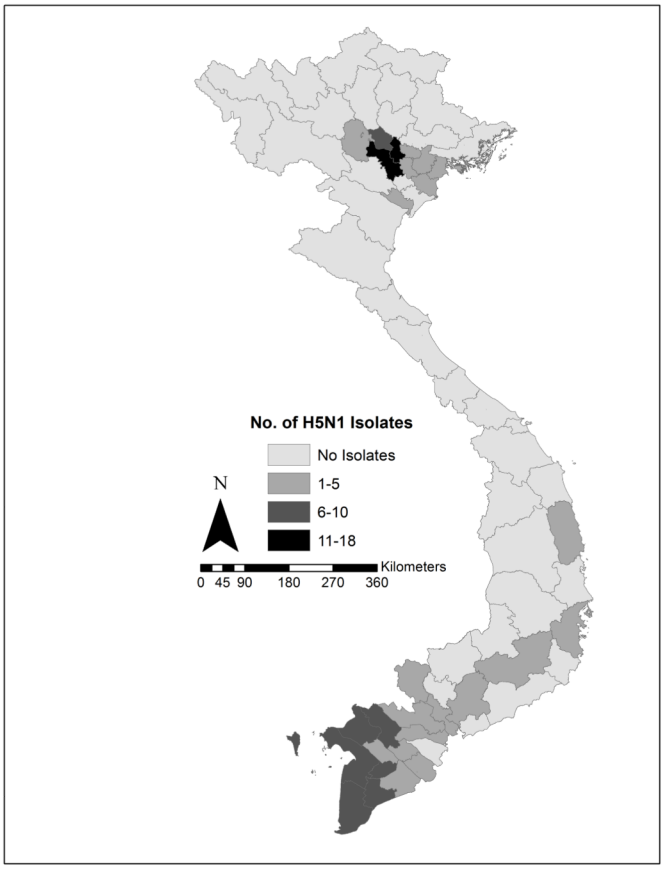

Figure 2.
Provincial boundaries of Vietnam and geographic locations of H5N1 viral isolation.

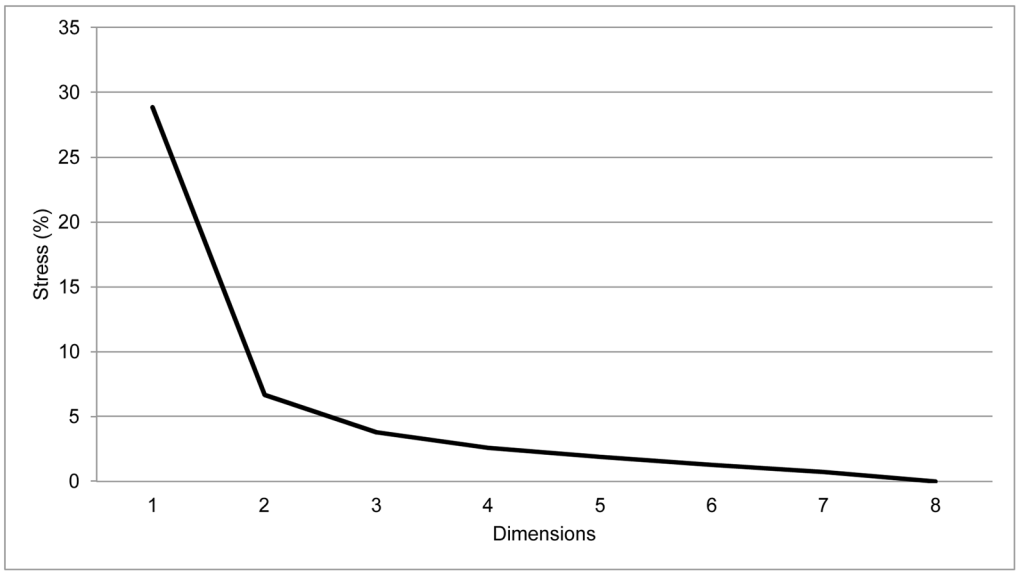

Figure 3.
Measures of stress versus dimensionality. Above three dimensions, reductions in stress are minimal.

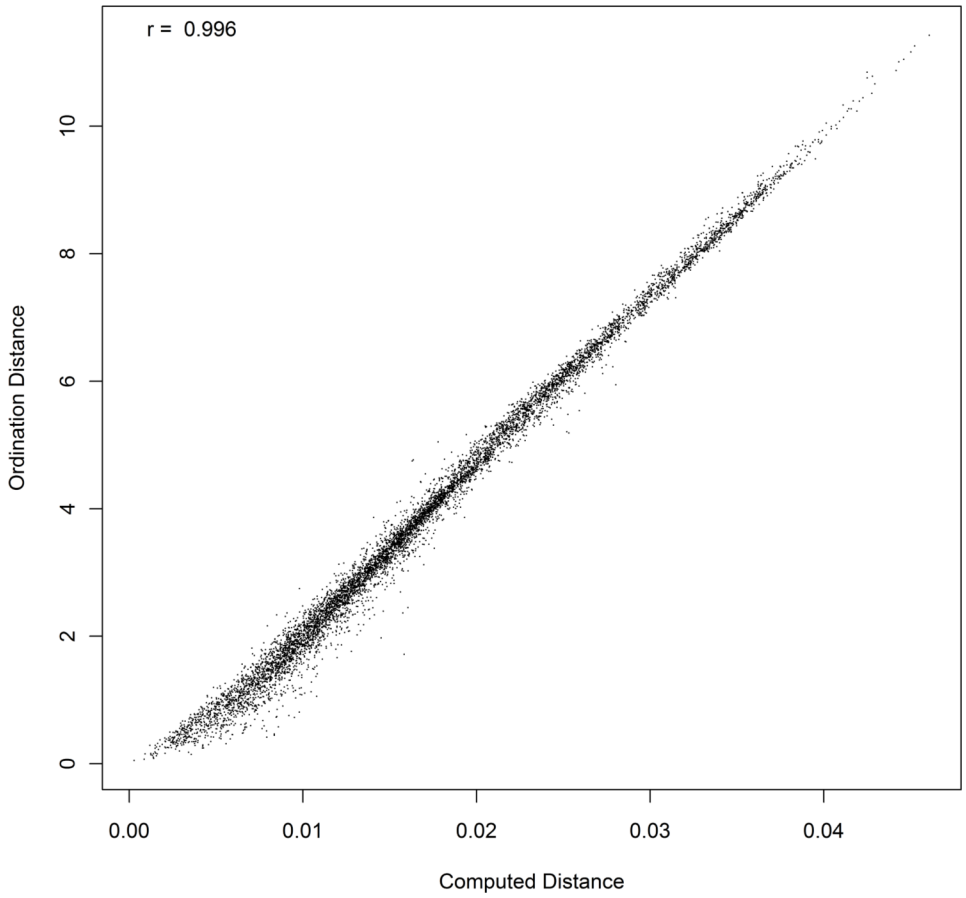

Figure 4.
Distances calculated from the observed data (x-axis) versus the ordinated distance (y-axis).

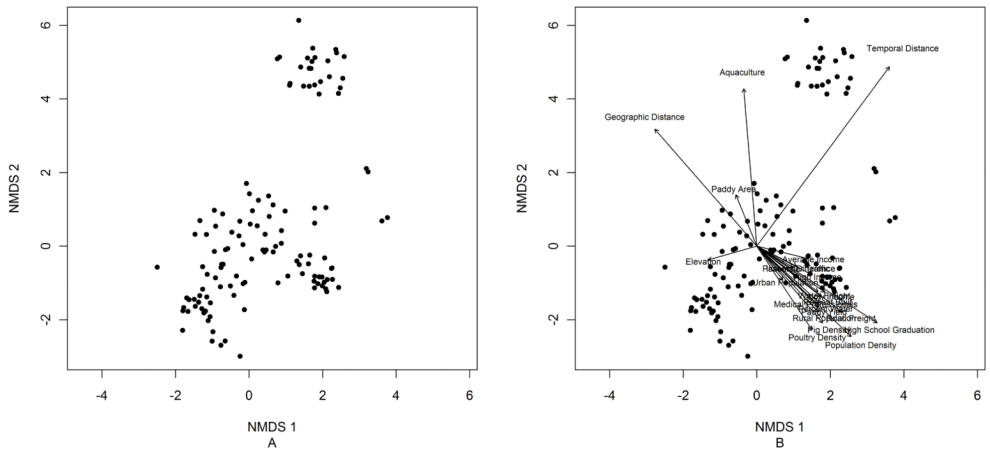

Figure 5.

A- Plot of the 125 viral points within the first two ordinated dimensions. B- All of the hypothesized population-environment drivers of genetic change arrayed over the scaled genetic measures, showing each variable's axis of differentiation through the 3-dimensional space. Longer axes of differentiation indicate greater association with the genetic distances arrayed in the 3-dimensional space.

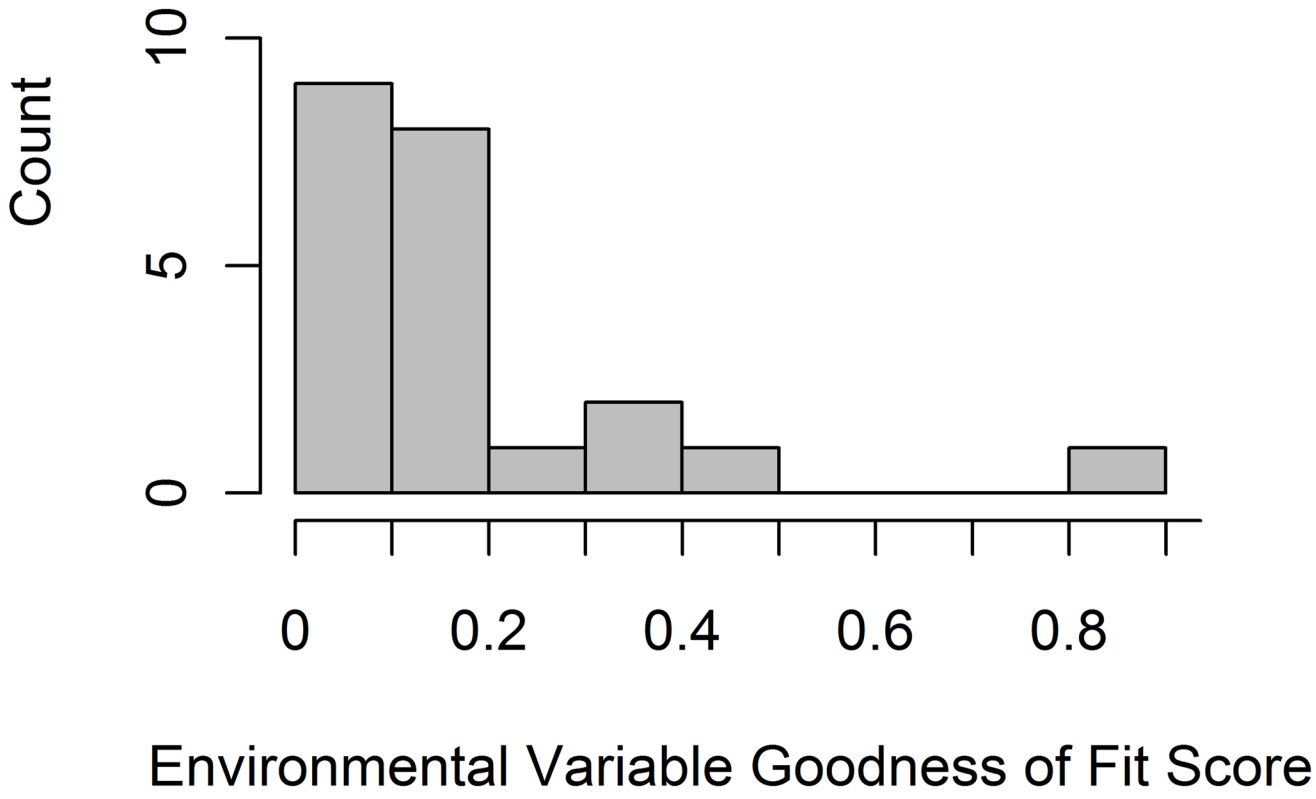

Figure 6.
Histogram of goodness of fit scores for population-environment variables. Only thirteen variables exhibit scores greater than 0.10, indicating a high level of correlation with the scaled genetic data.

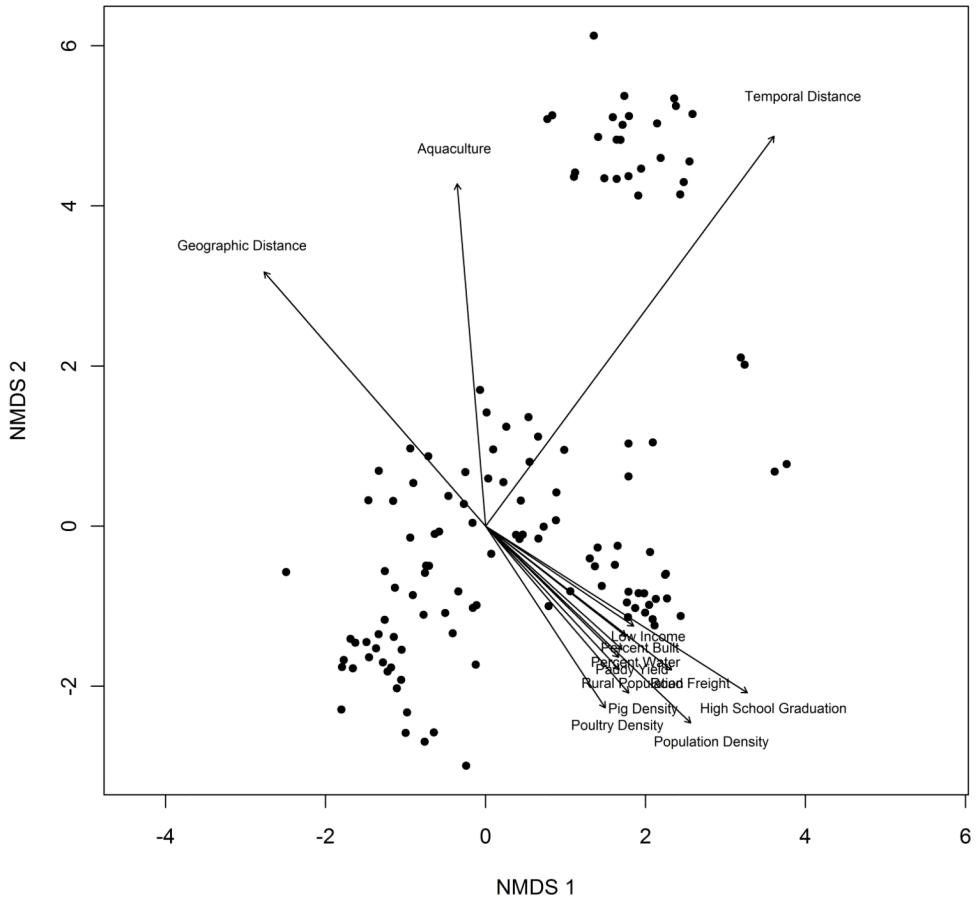

Figure 7.

The axes of differentiation for thirteen independent variables with goodness of fit scores of 0.10 or above: aquaculture, population density, high school graduation, poultry, pigs, road freight, percent water, percent built, paddy yield, rural population, income of poorest quintile, geographic distance and temporal distance.

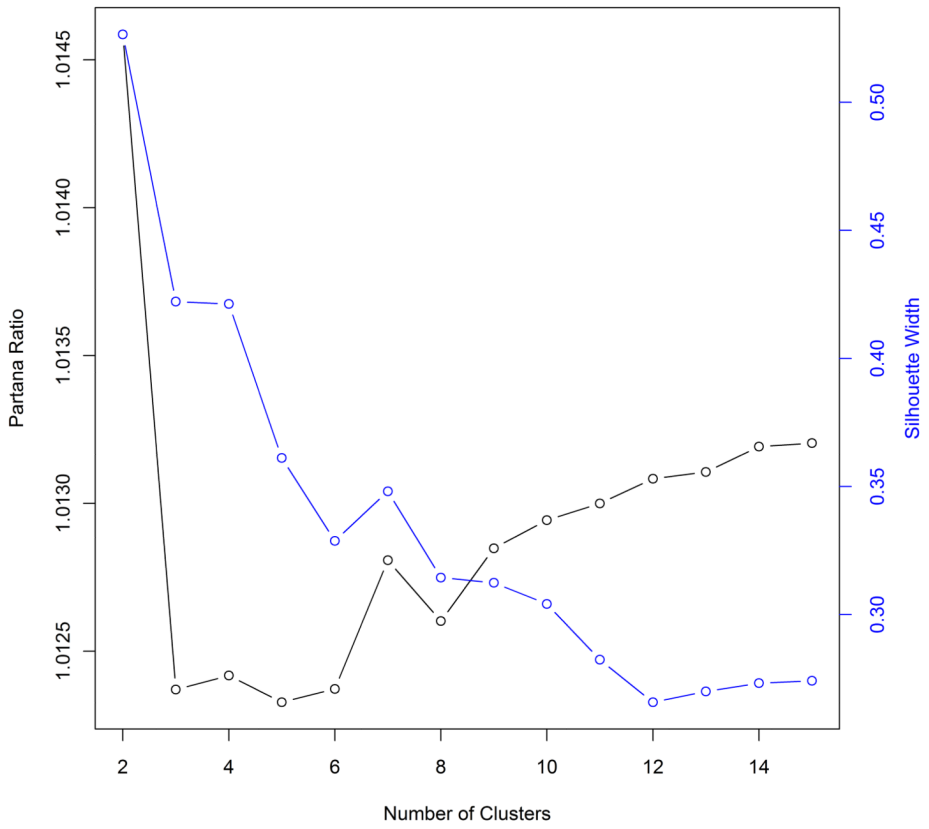

Figure 8.
The optimum number of clusters maximizes both Partana ratio and silhouette width. Seven clusters best describe the scaled genetic data.

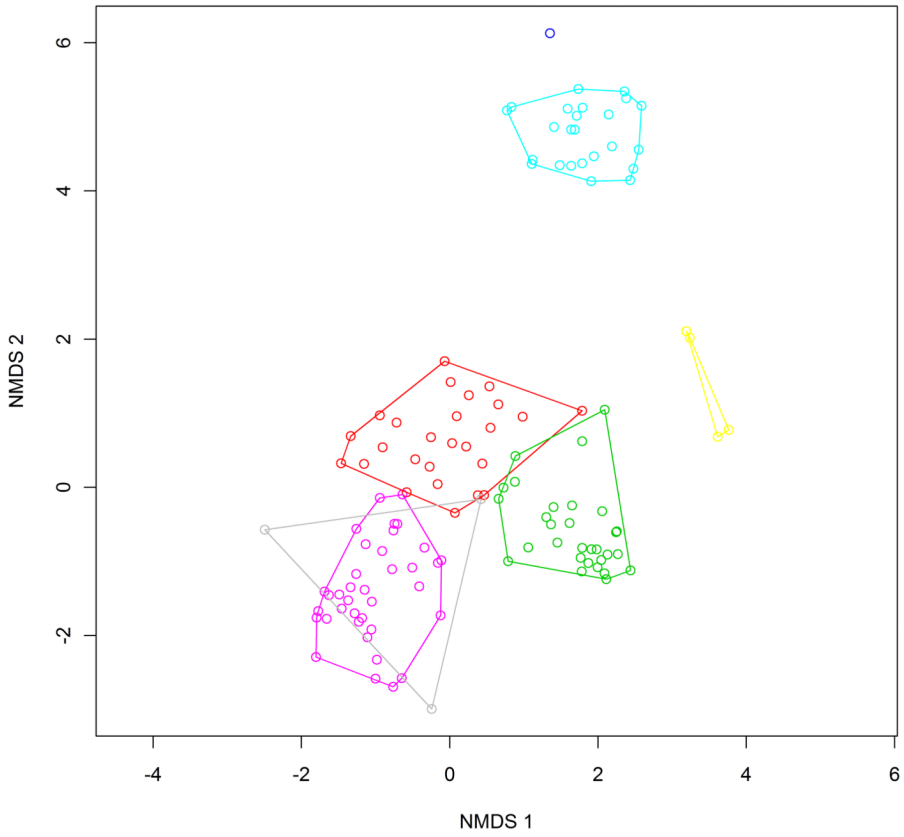

Figure 9.
NMDS results (the first two dimensions, out of three) charted according to cluster assignment.

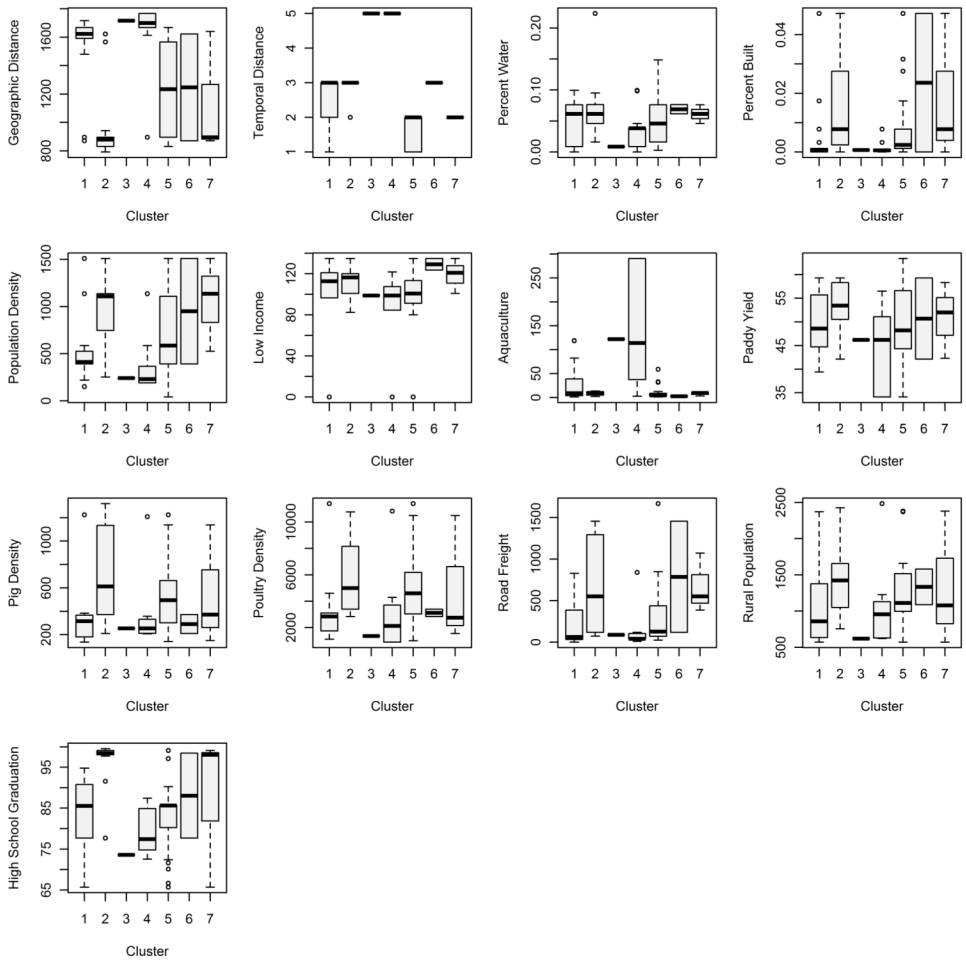

Figure 10.
Distributions of the thirteen independent variables within each of the 7 clusters.

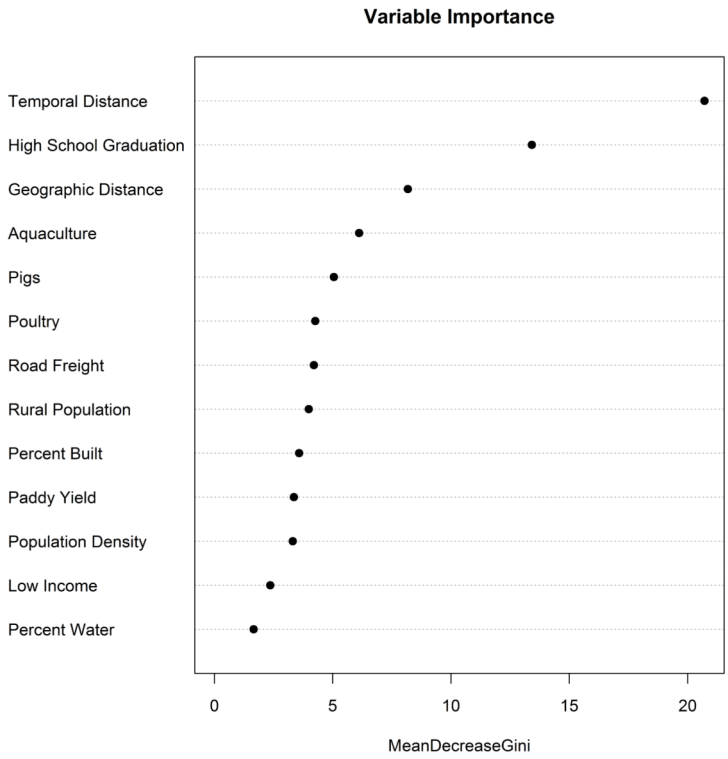

Figure 11.
Varying importance of the thirteen variables in relation to the genetic differentiation of viruses. This is an indication of the strength but not the direction or statistical significance of relationships.

Table 1

Population and environment variables included in the analysis.

Column1	Variable	Measure (per province)	Dates	Source
Population				
	Population density	median persons/square kilometer	2005	‡‡
	Passenger traffic	million persons/road kilometer	2003–2007	**
	Waterway freight traffic	million tons/kilometer	2003–2007	**
	Roadway freight traffic	million tons/kilometer	2003–2007	**
	Rural population	thousand persons	2003–2007	**
	Urban population	thousand persons average monthly income, income	2003–2007	**
	Income indicators	inequality	1999	**
	High school graduates	percent graduates	2003–2007	**
	Medical professionals	total persons	2003–2007	**
Environment				
	Poultry	thousand head	2003–2007	**
	Pigs	thousand head	2003–2007	**
	Planted area of rice paddy	thousand hectares	2003–2007	**
	Yield of rice paddy	quintal per hectare	2003–2007	**
	Water surface for aquaculture	thousand hectares	2003–2007	**
	Water surface area	percent	1981–1994 composite	§§
	Urban/built surface area	percent	1981–1994 composite	§§
	Elevation	median kilometers above sea level	2000	††

** General Statistics Office of Vietnam,

‡‡ CIESIN,

§§ GLCF (UMD)

†† SRTM30 (NASA)

Table 2

Regression results showing the influence of six population-environment independent variables on viral NMDS loading scores.

	V1			V2			V3		
	Coefficient	**t-Statistic**	**p-Value**	**Coefficient**	**t-Statistic**	**p-Value**	**Coefficient**	**t-Statistic**	**p-Value**
Intercept	−8.93636	−4.98	<0.0001	−13.83283	−7.03	<0.0001	−0.15703	−0.08	0.9336
Aquaculture	−0.00962	−3.04	0.0029	0.00357	1.03	0.3055	−0.00678	−2.05	0.0431
High School Graduation	0.13265	6.16	<0.0001	−0.00125	−0.05	0.9578	0.05299	2.35	0.0205
Population Density	−0.00051	−0.91	0.3667	0.00267	4.31	<0.0001	−0.00076	−1.29	0.1984
Temporal Distance	−1.28554	−7.52	<0.0001	4.43109	23.63	<0.0001	−1.14879	0.18	<0.0001
Percent Water	9.46783	2.26	0.0255	−7.1416	−1.56	0.1222	10.04685	2.29	0.0236
Low Income	0.01858	3.26	0.0015	0.00435	0.70	0.4884	−0.00454	−0.76	0.4488

R-squared for each model was: V1=0.70, V2=0.88, V3=0.49.