SPACE-TIME DIFFERENTIATION OF DRIVERS OF AND BARRIERS TO H5N1 AVIAN INFLUENZA EVOLUTION IN VIETNAM

Margaret A. Carrel

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements of the degree of Doctor of Philosophy in the Department of Geography.

Chapel Hill
2011

Approved by:

Michael Emch

Melinda Meade

Aaron Moody

Xiu-Feng Wan

Thomas Whitmore

ABSTRACT

Margaret Carrel: Space-time differentiation of drivers of and barriers to H5N1 avian influenza evolution in Vietnam
(Under the direction of Michael Emch)

The emergence and re-emergence of human pathogens resistant to traditional medical treatment will present a challenge to the international public health community in the coming decades. Geography is uniquely positioned to examine the progressive evolution of pathogens across space and through time, and to link molecular change to interactions between population and environmental drivers. The widespread outbreak of H5N1 avian influenza across Asia in 2003, and its continued circulation within both poultry and human populations, presents an opportunity for the integration of traditional disease ecology with the emergent field of landscape genetics. Combining spatial statistical methods with genetic analytic techniques, geographic space is used to explore genetic evolution of H5N1 highly pathogenic avian influenza viruses (HPAIV) at the sub-national scale in Vietnam.

This dissertation investigates the following topics: differences in genetic characteristics by species of isolation, location and timing of barriers to gene flow, and population-environment characteristics associated with increased viral evolution in Vietnam from 2003 to 2007. A variety of methods are used, including cluster analysis, multidimensional scaling, analysis of variance, and linear regression. Results indicate that genetic differentiation of these viruses varies significantly according to both their host species and the isolation time, but has a complex relationship with the geographic location of

virus isolation. The effect of geographic space, and underlying landscape differentiation, does not appear to create boundaries to gene exchange across Vietnam. Taking these indicators of the influence of species, temporal characteristics and geographic space into account, the drivers of molecular evolution of H5N1 HPAIV in Vietnam are as predicted by a disease ecology framework, a combination of both population and environmental characteristics.

These findings indicate that there are significant spatial and temporal effects on the evolution of H5N1 HPAIVs, and that local-level conditions can affect viral genetic evolution. Given that areas of rapid genetic evolution are more likely to produce a highly pathogenic virus capable of sustained human-to-human transmission, further exploration of spatial variation in molecular change is needed.

ACKNOWLEDGEMENTS

To my committee members, thanks for reading all these pages and making them better. To Mike, thanks for being a super duper advisor, and not making me cry very often.

To my mom and dad, who inspired me to get a PhD in the first place, you make the academic lifestyle look good.

And to my husband, for pouring the vodka tonics and calming the hysteria.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

ABBREVIATIONS

AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
CIESIN	Center for International Earth Science Information Network
GIS	Geographic Information System
GLCF	Global Landcover Classification Facility
HA	Hemagglutinin
HPAIV	Highly Pathogenic Avian Influenza Virus
IBD	Isolation by Distance
IDW	Inverse Distance Weighted
LLR	Log Likelihood Ratio
MANOVA	Multiple Analysis of Variance
MP	Matrix Protein
NA	Neuraminidase
NCVD	National Centre for Veterinary Diagnostics
NMDS	Non-Metric Multidimensional Scaling
NP	Nucleoprotein
NS	Nonstructural Proteins
OIE	World Organization for Animal Health
PA	Polymerase
PB1	Polymerase
PB2	Polymerase
PCA	Principal Components Analysis

RNA	Ribonucleic Acid
SRTM	Shuttle Radar Topography Mission
VAC	Vuon (agricultural plots) Ao (ponds) Chuong (caged birds)
VIF	Variance Inflation Factor
WHO	World Health Organization

CHAPTER 1

INTRODUCTION

H5N1 avian influenza was first identified in domestic geese in southern China in 1996 (1, 2). There was little alarm at the emergence of this new influenza type, however, until the 1997 outbreak of H5N1 in Hong Kong Special Administrative Region (SAR) (3, 4). The Hong Kong outbreak was associated with high mortality in domestic poultry and humans, and represented the first time an avian influenza strain was able to cross the species barrier to humans without an intermediate mammalian host and cause death in human hosts (2, 4). From 1997 to 2002, H5N1 HPAIVs were circulating in domestic poultry populations, but no laboratory confirmed human infections took place (5). Then, from late 2003 to early 2004, nearly simultaneous H5N1 outbreaks in both domestic poultry and people occurred in nine East and Southeast Asian countries: China, Cambodia, Indonesia, Japan, Korea, Lao People's Democratic Republic, Malaysia, Thailand and Vietnam (6, 7). Accompanying the expansion of H5N1's geographic range was an increase in genetic diversity, with the emergence of regionally distinct sub-lineages (7-9). To date, H5N1 viruses in poultry have been detected in over fifty countries across the Eastern hemisphere, spreading outwards from East Asia to the Middle East, Europe and Africa (Figure 1.1) (10, 11). Human cases reported to the World Health Organization (WHO) numbered over 500, as of February 2011, with a case mortality rate of sixty percent (12).

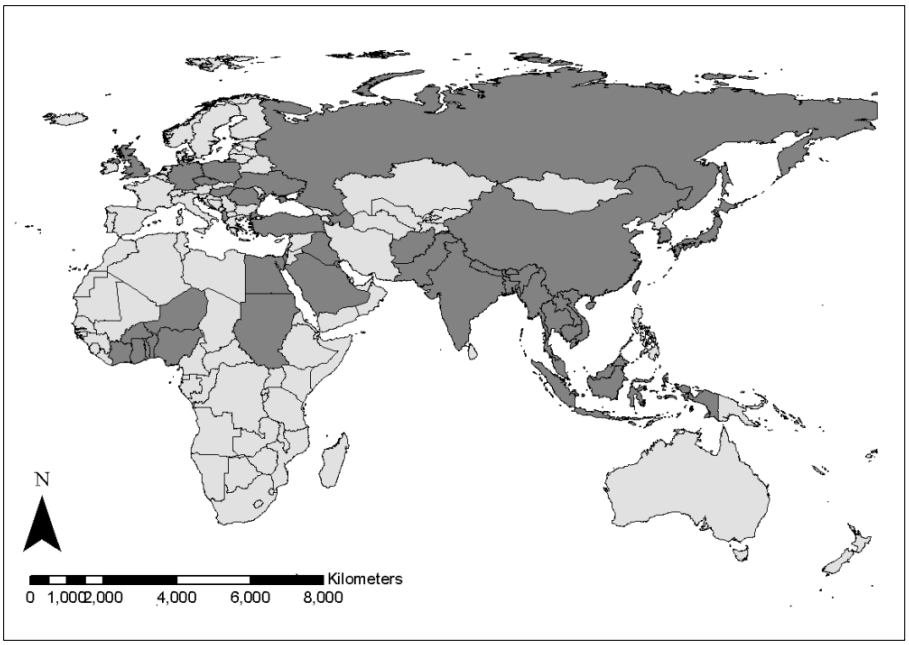

Figure 1.1: Geographic extent of H5N1 avian influenza in poultry (shaded), from 2003 to 2011. Adapted from data gathered by the World Organization for Animal Health (OIE) (11).

While H5N1 viruses are now found in countries across the Eastern hemisphere, novel genotypes continue to emerge from Southeast Asia, and this area remains the focus of intense public health surveillance and influenza research (7-9, 13). Southeast Asia is also the location of the majority of reported human H5N1 cases, with, as of March 2011, 334 of the 534 total worldwide cases reported in Vietnam, Thailand, Myanmar, Cambodia, Lao PDR and Indonesia (12). Some direct human to human transmission events have also occurred in this area, as well as in Pakistan (14, 15) As H5N1 incidence and mutation continues, the chances increase that a strain will develop the necessary attributes for efficient replication within and transmission between people. By examining the geography of viral evolution in Vietnam, this study aims to illuminate the processes by which local-level environments facilitate or inhibit such genetic change.

The purpose of this research is to examine how local-level population and environment variables interact in space and time to drive evolution of H5N1 influenza viruses in Vietnam. Prior to examining *why* evolution is taking place (i.e., what's driving it?), I explore *where* and *when* viral evolution is occurring. Therefore, the specific study questions this project addresses are:

1.) Do species-specific differences in genetic clustering exist in Vietnam? Specifically, do strains isolated from chickens differ from duck isolates in how space and time interact with viral evolution?

2.) Do barriers to evolution of H5N1 influenza viruses exist in Vietnam? Do boundaries to gene flow vary temporally?

3.) What population and environment characteristics are associated with increased genetic change? Are these drivers mediated by time and space?

These study questions have been chosen to integrate spatial and molecular analysis as well as to address major deficits in the literature. First, few of the studies considering H5N1 genetics have included spatial theory or analytic methods, using geography instead as a descriptive element. Genetic space, illustrated by means of phylogenetic trees, is a more common metric of distance than geographic space and is used by geneticists to explain evolutionary relationships among viral isolates. Secondly, those studies that do incorporate geography into H5N1 analysis typically use a simple presence/absence outcome variable (16-18). This overlooks the fact that viruses are continuously evolving and changing, and that perhaps certain environmental settings may sustain H5N1 viral incidence but not encourage rapid viral adaptation. Thirdly, these studies have assumed that the relationship between environmental drivers and H5N1 incidence are static across space and time. It stands to

reason, however, that just as H5N1 evolution varies in space and time, so too will the influences of population and environmental factors on mutation.

Despite the efforts of researchers and public health officials worldwide, there remain questions about how and why H5N1 virus moves and evolves in space and time. Influenza surveillance in domestic poultry in Southeast Asia informs us that H5N1 virus continues to circulate in areas of prior incidence as well as expand its geographic range (19). Legal and illegal poultry trade, combined with migration of wild birds, is the most likely explanation for how H5N1 crosses borders and moves within countries (6, 20). Exploring where and when barriers to genetic mixing exist can illuminate how H5N1 moves and evolves. Additionally, this study explains what combinations of environmental and population variables are associated with high rates of genetic change in H5N1 viruses. It also helps to understand whether there are distinct differences in how H5N1 circulates and evolves in the two most common domestic poultry species in Asia, chickens and ducks. If H5N1 viruses shows species specific patterns, then this could suggest varying surveillance and prevention measures by which species is predominant in an area. Ultimately, this study exemplifies how spatial analytical tools, including spatial statistics and geographic information systems (GIS), can be used in exploratory landscape genetic analysis.

Background: The Influenza Virus & Measuring Genetic Change

Influenza viruses belong to the family *Orthomyxoviridae* and are classified into three types: A, B, and C. Influenza A viruses infect a range of species, including mammals and birds, whereas influenza B and C viruses are found only in mammals. Influenza A viruses are composed of eight ribonucleic acid (RNA) gene segments, which encode ten or eleven proteins, depending on the strain: hemagglutinin (HA), neuraminidase (NA), matrix proteins

M1 and M2, nonstructural proteins NS1 and NS2, a nucleoprotein (NP), and three or four polymerases (PA, PB1, PB1-F2, and PB2) (21-23). The subtypes of Influenza A viruses are determined according to the surface glycoproteins, HA and NA, of which there are sixteen HA and nine NA subtypes. Among influenza's eight gene segments, HA of H5N1 HPAIVs remains genetically associated with the original H5N1 goose isolates, A/goose/Guangdong/1/1996(H5N1). One or more of the remaining seven gene segments of H5N1 AIVs were adopted from other subtypes of influenza A viruses, especially low pathogenic avian influenza viruses. Since its first identification from geese, H5N1 viruses have undergone active reassortments: at least 21 genetic reassortants have been identified (13).

H5N1 viruses are further classified into clades according to their HA segment. Clades are defined by the WHO as sharing a common progenitor virus and having higher genetic similarity within group than without. Currently there are 10 first-order clades, 0-9, and multiple second- and third-order clades which are named hierarchically, e.g., Clade 2.3.4 (24). H5N1 viruses may be further classified according to genotype, a grouping based upon genetic characterization of all eight viral gene segments (8, 25-27). Thus, genotypes do not overlap completely with clade designations.

Mutation of genetic sequences in influenza viruses occurs via two processes: antigenic drift and antigenic shift. Antigenic drift refers to gradual changes in the surface proteins HA and NA of influenza viruses and it can lead to the changes of antigenic profiles of influenza viruses. Antigenic shift, in contrast, represents the unpredictable and rapid change in viral sequences that takes place when a single cell is infected simultaneously with two or more influenza viruses. In cases of antigenic shift, entire segments of RNA are

exchanged between viruses, resulting in strains that have rapidly acquired new genetic code. Antigenic drift has been shown in facilitate the emergence of 1957 H2N2, 1968 H3N2, and 2009 H1N1 pandemic influenza viruses (28). Should a H5N1 virus that is highly virulent come into contact with a virus that is able to efficiently transfer among mammals, it is possible that a new strain will emerge that can cause pandemic H5N1 influenza.

Tracking the evolution of avian influenza viruses is vital to prevention and control measures (26). Multiple methodologies for measuring genetic change between viruses are in use, typically combining sequence alignment analysis with phylogenetic tree construction (5, 9, 10, 29-32). The strength of these methods varies according to several factors, including the completeness of viral sequences and the validity of the sequences chosen as seeds in the phylogenetic analysis (26, 27). Distances between viruses within a phylogenetic tree can be evaluated via a unit-less patristic distance, wherein the amount of genetic change between viruses is equal to the sum of the branches of the tree connecting them (33).

Of influenza's ten proteins, four, HA, NA, PB2 and NS1, have been found to relate to the virulence of the virus and the range of hosts it can infect in the following ways (21, 22, 34). The HA protein codes for spikes of hemagglutinin on the exterior of the virus, which attach and fuse the virus to host cells. The preferred host cell receptors differ among human and avian influenza proteins, a preference which is in part responsible for the low number of human H5N1 cases. Avian influenza viruses bind to sialic acid (SA) alpha2,3 receptors while human viruses bind to SA alpha2,6 receptors (35, 36). A mutation, however, at one of several positions in HA's genetic sequence leads to avian influenza viruses that can preferentially bind to SA alpha2,6 receptors in human respiratory cells (37, 38).

Mutation in the PB2 protein also helps avian influenza viruses to overcome the barrier between bird and mammal infection (34). A change in the amino acid present at position 627, from glutamic acid to lysine, in the PB2 sequence was shown to transform non-lethal viruses H5N1 avian influenza viruses into highly lethal viruses causing systemic infection in mice (34, 39). Viruses isolated at Qinghai Lake in northeastern China had lysine at position 627, as did viruses isolated from humans in Indonesia and Vietnam, indicating that some strains of H5N1 are closer to being able to efficiently transmit to and among mammals (21, 32).

Influenza's NA protein encodes neuraminidase, an enzyme found on the surface of influenza viruses. This enzyme works to open the host cell and efficiently release the viral replicates after infection, and also cleaves SA in mucous that act as decoys for the surfaces of targeted epithelial cells (38). A deletion of twenty amino acids in the stalk region of the NA protein is thought to allow viruses from wild aquatic birds to better adapt to replication in land-based domestic poultry (i.e., chickens) (22, 38). Viruses with NA stalk deletions are highly virulent in poultry and have been found in humans (34).

The NS1 protein boosts H5N1 virulence by affecting host immune responses in two ways: blocking interferons and boosting proinflammatory cytokines (21, 34, 40). Interferon responses are set off when a cell senses viral infection, and result in inhibited viral replication through multiple mechanisms, including RNA degradation and editing and translation inhibition (41). The interferon-antagonistic properties of NS1 result from its ability to hide the production of viral RNA in infected cells from immune system sensors (40, 41). Simultaneously, NS1 increases the production of cytokines; this deregulation of cytokine

manufacture encourages systemic infection in organs outside of the respiratory and gastrointestinal tracts (21, 42).

The characteristics of and mutations within the HA, NA, PB2 and NS1 proteins are understood to be strongly related to the virulence and host range of H5N1 avian influenza viruses. The influence and importance of H5N1's other proteins are less well-known, but could also play a critical role in the ability of viruses to reproduce and transmit within humans, as pathogenesis is a systematic effect, and is the result of interactions among all eight segments (41). As H5N1 avian influenza viruses continue to evolve, the likelihood of a strain developing the ability to efficiently infect and transmit among humans increases. Thus, investigating how influenza viral evolution varies by species and across space and time in Vietnam, as well as what population and environment variables are associated with increased rates of genetic change, can illuminate local-level environments that are more likely to produce a strain with pandemic capabilities.

H5N1 Influenza in Vietnam

H5N1 avian influenza was first detected in Vietnam in 2003. Vietnam is second only to Indonesia in the number of human cases and deaths reported to the WHO, and is similarly prominent in the number of infected poultry reported to the World Organization for Animal Health (OIE) (11, 12). According to some estimates, over 30 million chickens and 14 million ducks were infected in Vietnam between 2004 and 2005 (43). While the precise origins of H5N1 in Vietnam remain unknown, evidence points to its introduction into the north across the shared border with China via trade, legal or illegal, in live poultry or poultry products (44-46). By 2004, H5N1 moved the length of the country, and was reported in

poultry and people in southern Vietnam. How the virus was transmitted inside of Vietnam is also undetermined, though domestic poultry trade is the likeliest explanation.

Several factors are responsible for the emergence and persistence of H5N1 avian influenza viruses in Vietnam. Chief among them is the shared border with southern China, the putative epicenter of influenza viruses (9, 47, 48). Vietnam also shares with China a cultural preference for purchasing live or freshly killed poultry from live bird markets, which act as sites of transmission of viruses among birds and humans (31). Backyard poultry flocks, of ducks, chickens or geese, are also common in Vietnam (22, 49). By some estimates, eighty percent of poor Vietnamese raise their own poultry (50, 51). The livelihoods of many rural Vietnamese are based on the VAC ecosystem, revolving around Vuon (agricultural plots), Ao (ponds), and Chuong (caged birds) (50). In VAC systems, 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.

Multiple sites of interaction that can contribute to H5N1 persistence and transmission exist in the VAC system. H5N1 avian influenza infections in birds are located in both the respiratory and gastrointestinal tracts, and the virus is transmitted through aerosol transmission and also shed in feces and saliva. Rural Vietnamese live in close physical proximity to their backyard flocks, and potentially poor hygiene combined with interaction with infected birds can lead to human infection via fecal-oral contamination (52). Domestic poultry can interact with wild birds when free-grazing in fields or in water bodies, and consumption of water supplies contaminated with infected saliva or feces links domestic bird infection to wild bird infection (35, 49). Finally, the influenza virus can survive in soil, so the use of bird feces as a fertilizer could cause H5N1 infection to spread within or between

flocks (22). There is regional variation in the VAC system. Northern Vietnamese rear chickens in greater numbers, while southern Vietnamese rear more ducks (16, 49). This regional variation carries over into H5N1 influenza viruses: northern viruses are more closely linked to Chinese viruses, while those in the south are closer genetically to Thai and Cambodian viruses (30, 45). Viral incidence and evolution thus appears to be associated with the trade and cultural patterns around the Red River and Mekong River basins. (32, 49)

While some cross-border regional patterns in Vietnamese H5N1 viral characteristics exist, the movement of viruses within the country is highly structured. Emergence and movement of H5N1 avian influenza in Vietnam appears to follow a distinctive pattern: new viral variants first appear in the north, move south within a few months, and then persist in the south while new variants again appear in the north (25). From 2003 to today, multiple sublineages of H5N1 have been detected in Vietnam. Only clade 1 viruses were detected in Vietnam until 2005, when clade 2.3.2 viruses were identified in northern Vietnam. In 2007, clade 2.3.4 viruses were also isolated in northern Vietnam, while clade 1 viruses remained predominant in the south. Phylogenetic analysis indicates that these 2.3.2 and 2.3.4 viruses are closely related to strains subsequently isolated in southern Vietnam (32, 45).

In response to the rapid and widespread outbreaks of H5N1 in humans and birds from 2003-2004, the government of Vietnam instituted a widespread vaccine campaign in 2005. Millions of birds were vaccinated against H5N1 influenza by July of that year, and the campaign appeared successful: no new poultry outbreaks were reported until December (43, 53). Given the rapid re-emergence of H5N1 in poultry in 2006, however, it is likely that the virus continued to circulate at low levels in unvaccinated poultry or wild waterfowl during

that time. No new human cases were reported to the WHO during 2006, though human H5N1 reappeared in Vietnam in 2007 and continues to today (12).

Population & Environmental Associations with H5N1 Avian Influenza

Previous research into population and environmental variables associated with H5N1 avian influenza can be divided into three categories, examining viral survival, viral hosts or sites of transmission.

Viral Survival

Influenza viruses can be transmitted both through direct and indirect contact (54). In direct contact infections, the influenza virus is aerosolized, carried in microscopic droplets of secretions from infected respiratory tracts. In indirect contact infections, viruses are located on surfaces or in materials, such as water supplies, soil, bedding and cages, as fomites. Both indirect and direct contact can lead to bird-to-bird, human-to-human and bird-to-human infections, although to date the majority of human infections have been the result of direct contact with infected poultry (52, 55). Whatever the transmission route, in order for H5N1 avian influenza to transmit between infected and susceptible hosts the virus itself must remain viably infective while outside of a host cell (56).

Laboratory tests indicate that H5N1 avian influenza viruses can survive in water sources for extended periods of time, though not for as long as wild-type viruses, which can remain viable for over 190 days (57). Viral survival in water is highly affected by the temperature, pH and salinity: persistence is inversely proportional to both temperature and salinity. The effect of these factors varies by virus strain, however, making it difficult to predict with any accuracy the survivability of H5N1 viruses in aquatic environments.

The seasonal peak of isolation of H5N1 virus in China is during October-March, the winter months (22, 31). Human influenza infections are also highly seasonal in temperate zones, and slightly seasonal in tropical regions, suggesting that there may be environmental drivers of infection. Environmental effects may be mediated through cycles in host resistance to infection (e.g., fluctuations in melatonin or Vitamin D), seasonal changes in host behavior (e.g., gathering of populations in close contact in winter), or through variation in viral survivability. New research into how influenza viruses survive in aerosolized form indicates that survival is highly negatively correlated not only with temperature and relative humidity, but with absolute humidity (58). A simple inverse relationship between viral survivability and absolute humidity exists, as humidity falls viral survival increases (59). Thus, the dryness of indoor environments in temperate winters leads to conditions that are conducive to viral survivability.

Viral Hosts

Wild birds, predominantly those of the orders Anseriformes (including ducks, geese, and swans) and Charaderiiformes (including gulls and waders), provide the natural reservoir for avian influenza viruses (28, 60). Influenza infections in wild waterbirds are typically minor. H5N1 avian influenza viruses have caused massive die-offs of waterfowl in China, however, and infected birds have shown signs of systemic infections (35, 56, 61). Prior to 2002, H5N1 infection in domestic ducks was usually asymptomatic or expressed as only minor clinical symptoms (35). Asymptomatic ducks could still shed infective viruses in their feces or saliva, however, for up to two weeks after infection (9, 22). Domestic ducks are sometimes referred to as "Trojan horses" of H5N1 infection, unnoticeably sustaining circulation of H5N1 in poultry (22, 61). Since 2002, however, some strains of H5N1 have

emerged that are highly pathogenic in ducks, producing severe systemic infection and lesions in multiple organs (56, 62-64).

In contrast to ducks, chickens experience high (often 100%) mortality after infection with H5N1 viruses, and have symptomology of infection in multiple organs (31, 38). In recent years, H5N1 infections in chickens have become more virulent, indicated by less time between infection and death (64). While H5N1 infections in chickens are stronger, viruses shed in chicken feces remain infective for only one day, as compared to a week for duck feces (21).

Pigs represent a possible intermediary host for avian influenza viruses in humans. Swine respiratory tracts carry surface cell receptors for both avian and human influenza viruses (65). So far, however, pigs have appeared to be only slightly susceptible to H5N1 avian influenza, evidencing only minor symptomology upon infection (56, 65). Presence of pig farming in regions thus may not be as critical an environmental factor influencing H5N1 incidence compared to other types of influenza, though this remains to be seen. The possibility remains, however, that pigs may act as reservoirs of infection or intermediate hosts between bird flocks, so swine density will be considered in this study.

Sites of Transmission

Areas with environments and populations (i.e. chickens) conducive to H5N1 viral survival and transmission include live bird markets, rural areas where households live in close proximity to poultry and wet-rice agriculture, and commercial poultry production sites. Live bird markets are common in Vietnam and other Southeast Asian countries where there exists a cultural preference for fresh poultry products. Viruses have been isolated from birds and surfaces in live bird markets in Vietnam and China, indicating that these are spaces

where infection can and does take place (8, 31). Markets can also act as a source of infection for humans residing in urban or peri-urban areas who have no other bird contact (66). While human infections can often be traced to market visits, in the case of infected birds, it is impossible to tell whether they were infected before arriving at the market, or whether they became infected while there. Birds infected at markets may transmit the disease back to the farm and to susceptible flocks if they remain unsold and return home with farmers at the end of market day. The movement of poultry to and from markets, as well as transport of fomite-infected poultry products, such as bedding, fertilizer and litter, facilitates the transmission of infection across space and time (22, 67)

High numbers of rural Vietnamese rear their own poultry in backyard pens. Environmental specimens gathered in Cambodia, which has a similar backyard poultry production system to Vietnam, detected H5N1 RNA in half of the household ponds tested, as well as in aquatic plants, mud and dry soil (68). Statistical modeling indicates that environmental transmission allows influenza to persist in small populations where the disease would otherwise vanish (69). Backyard poultry practices have also been implicated in the outbreaks of H5N1 influenza in western Africa (10).

The emergence of highly pathogenic H5N1 influenza is sometimes linked by scholars to the intensification of commercial poultry production in Asia (22). To date, however, H5N1 has been only infrequently reported in commercial poultry workers (44). Evidence from Thailand, furthermore, indicates that H5N1 is not a serious concern in commercial production, due to high levels of biosecurity measures maintained at these facilities, including keeping poultry inside and vaccination programs (70).

Research into environmental characteristics associated with H5N1 avian influenza illustrates that areas with more potential sites of transmission and more potential hosts are also associated with higher viral incidence. In Thailand, areas of H5N1 incidence are highly associated with duck and human population densities, low elevation and wet-rice agricultural production (16, 70-72). H5N1 risk in Thailand is also linked to proximity to highways and urban centers (18). In Vietnam, H5N1 outbreaks have been correlated to human population density, percentage of land surface devoted to aquaculture and rice paddies, and chicken and duck populations (49). Human infections with flu-like illness (potentially H5N1 but not definitely) in Vietnam are linked to the presence of sick birds in backyard flocks, low socioeconomic status and both young and old age (51).

The literature thus suggest that population and environment characteristics jointly influence H5N1 incidence. Connections between such features and genetic characteristics of viruses are less established, however. This work explores the possible presence of population and environmental drivers to molecular evolution of H5N1 influenza viruses.

Theoretical Framework

This research is guided by two bodies of theoretical and scientific work. The first, *disease ecology*, is situated in the human-environment tradition of geography, and considers interactions between populations, behaviors and environments that influence the occurrence of diseases in specific places at specific times. The second, *landscape genetics*, emerged from ecology studies and is concerned with examining how variation in genetic characteristics is related to variation in environmental features.

Disease Ecology

The foundational idea of disease ecology is that human life is a process, a continual interaction between the internal and external environments (73). Disease ecology emerged in the second half of the 20th century, in reaction to the belief in medicine that infectious diseases were a thing of the past, and that curing disease was simply a matter of prescribing the right medication. Disease ecologists realized that diseases do not exist independently of environments or hosts, so "for adequate health maintenance, a vision broader than symptomology is necessary." (74) To understand a disease, you must understand both the *person* and the *place* in which the infection occurs (74-76). Disease ecology thus concerns itself with "the ways human behavior, in its cultural and socioeconomic context, interacts with environmental conditions to produce or prevent disease." (76). These interactions are not static, however, but are dynamic and responsive to disturbance. Changes to behavior or environment, via climate change, agricultural simplification, migration, etc., can have positive or negative effects on disease experiences, either magnifying or minimizing risk and exposure. Disease ecologists do not view humans as passive members of the disease system, however, but recognize that humans can change their behaviors or modify their environments in reaction to changes elsewhere in the system (77). Thus, disease ecology is inherently focused on integrating both the physical (environmental) and social aspects of human lives into an understanding of ill-health (78).

Although recognizing that human disease is the outcome of a complex and dynamic interaction between the internal and external environments of an individual or a population seems relatively straightforward, conceptualizing and understanding these interactions can be difficult. One way of doing so is to view disease at the intersection of three types of

variables, *population, environment* and *behavior* (79). Population variables in this framework are those that affect individuals' responses to disease as biological beings, such as nutritional and immunological status, age, etc. The environment category encompasses all aspects of the built, natural and social environments that can affect disease outcomes. Behavior factors include both observable aspects of actions and culture, such as social organization, house type, diet, etc., as well as less tangible variables like perceptions of risk. Disease outcomes are the result of place and time specific interactions among these variables.

While disease ecology is primarily concerned with human health, the theories and framework outlined above can be applied to examining disease in animals, particularly zoonotic diseases that transmit to humans. It is particularly appropriate when considering disease in domesticated animals, which not only have their environments and behaviors modified or affected by human desires, but in some cases their basic biology as well. An outline of how the "triangle of human ecology" looks when applied to domesticated poultry in Vietnam is shown in Figure 1.2. In this model, H5N1 avian influenza occurrence is influenced by the interactions in space and time between bird and human populations in areas of varying environments and behaviors. The factors included are based on prior research (as outlined in sections 2.2 and 2.3) suggesting they are related to H5N1 occurrence in poultry populations.

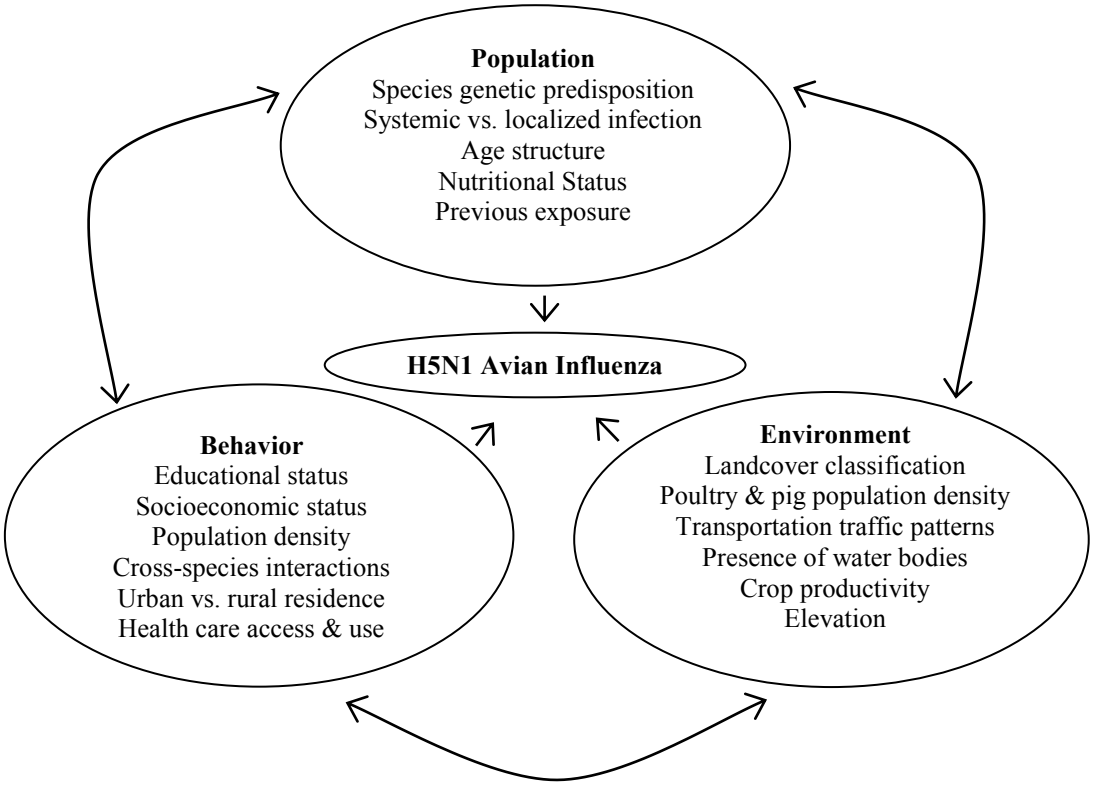

Figure 1.2: Framework describing the disease ecology of H5N1 avian influenza in Vietnam's domestic poultry.

Factors of poultry populations that may lead to H5N1 incidence and evolution include genetic predispositions, previous exposure to H5N1 viruses, and nutritional status. Chickens and geese are more susceptible to systemic H5N1 infection, and are more likely to experience mortality, than are ducks and wild waterfowl. Prior exposure to another H5N1 virus can give poultry immunity to currently circulating viruses, and vaccinated poultry are resistant to infection. Flocks of young ducks and chickens are less likely to have been previously exposed or vaccinated, so may also be more susceptible to infections.

Aspects of the environment that are believed to interact with H5N1 incidence and evolution include wet-rice agriculture and aquaculture, because of the potential for mixing of infected and non-infected poultry in these aquatic environments. The circulation of humans and birds across transportation networks allows for greater potential spread of H5N1 viruses

across the landscape. Poultry trade within and between countries is believed to help drive the circulation of viruses, and the degree of poultry trade is in turn driven by poultry population densities. Behavioral factors associated with H5N1 include residential population density, and rural versus urban residence. Availability and utilization of medical services can also influence H5N1 incidence, as does education level and socioeconomic status, since less educated and poorer individuals perhaps have less knowledge about transmission pathways, lower hygiene levels and less access to or seeking of health care.

Landscape Genetics

The emerging interdisciplinary field of landscape genetics is based on the idea that exploring spatial variation in genetics can illuminate how organisms move through the landscape. Landscape genetics combines theory and methods from population genetics and landscape ecology in order to explore interactions between evolutionary processes and environmental features (80-84). Landscape genetic studies differ from those of biogeography and phylogeography in that they operate at finer spatial and temporal scales, and are made possible by the convergence in the past decade of publicly available, high-resolution molecular and geospatial datasets.

There are two steps to a landscape genetics analysis. The first is to identify patterns of genetic variation in space, the second is to correlate those patterns with landscape or environmental features (80). The aim of the field is to not just identify *pattern*, but to go a step further and explore *process*. The strength of landscape genetics is that, unlike more abstract methods such as calculation of ecological connectivity from a habitat map, it quantifies processes in real landscapes (85). Identifying the real-world factors that drive

evolutionary processes enables scientists to model and predict evolution of genetic diversity (82).

Precisely what constitutes a landscape genetics analysis is somewhat ill-defined, as after Manel et al. (80) coined the phrase it was applied to any study that incorporated landscape features or geographical coordinates into an analysis of genetic variation. Storfer et al. (83) endeavor to bring greater clarity and order to the emergent field by outlining five research categories within landscape genetics:

1.) Quantify the influence of landscape variables and their configuration on genetic variation.
2.) Identify landscape barriers (or channels of) to gene flow.
3.) Identify source and sink dynamics and movement corridors.
4.) Understand the spatial and temporal scales of an ecological process.
5.) Test species-specific ecological hypotheses.

Multiple methodologies exist to explicitly incorporate landscape data into genetic analyses, including matrix correlations, point pattern analysis, tests of spatial autocorrelation, assignment tests and multivariate regression (80, 83). Landscape genetics draws upon landscape ecology methods for analysis, while using data from population genetics as the outcome variable of interest. Merging these two types of data, however, requires that attention be paid to the temporal and spatial scale at which they are collected, as well as what sampling scheme is used for assessing genetic variation among individuals (84).

Landscape genetic studies have, so far, been confined primarily to the study of plants and animals rather than pathogens or people. Ecologists and biologists have used landscape genetics to infer colonization patterns, disease response, habitat restriction and extinction events of mammals, reptiles, trees and insects, issues which are of particular importance to

conservation biologists and managers. In the past years, however, there has been recognition by disease ecologists that landscape genetic techniques can be used to explore drivers of disease spread and parasite transmission as they relate to human illness (86). As yet, little work has applied landscape genetic techniques to anthropozoonotic (infecting both humans and animals) pathogens. Such application needs to be informed by knowledge of how such pathogens are influenced by human and animal interactions with environments, and how such interactions vary in space and time, in order for research findings to be valid or valuable for public health efforts. The strength of disease ecology as an integrative science, drawing upon knowledge about human-environment interactions, lends itself to the application of landscape genetic techniques, developed for plant and animal studies, to the study of pathogenic evolution.

The application of landscape genetic study methods to pathogens also necessitates recognition of potential sampling problems (83, 84). Ideally, sampling is conducted according to a statistically or spatially sound schematic. My outcome variable, however, is an opportunistic rather than a planned sample, since H5N1 avian influenza surveillance and sampling in Vietnamese poultry is carried out unevenly in space and time, dependent on sick birds being reported to animal health officials. Areas not reporting cases (and genetic information) are not necessarily absent of the virus, it might be unreported or unrecognized. Previous landscape genetics studies have used similarly opportunistic sampling schemes, relying on genetic testing of small populations in non-randomized locations (87)(88, 89). I must be aware, however, of how the spatial and temporal configuration of my samples, as well as the small sample size, influences the results of my analyses and the ecological conclusions that I draw (84, 90).

For H5N1 avian influenza, a temporal mismatch between the viral genetic data and the population-environment data use for analysis must also be taken into consideration. The genetic characteristics of H5N1 avian influenza viruses are the result of more than a decade of evolution, and result from both historic and contemporary processes. The population-environment data these genetic changes are being correlated with, however, represent a snapshot in time of population density, landcover, etc. Those landscape features, taken from one or several points in time, will be used to explain genetic patterns that have potentially evolved since 1996, or before. By limiting my study to viruses isolated in specific years and in one country, and by using population and environment datasets drawn from the same time span, I hope to avoid some of the confounding historic influences that are present in landscape genetic studies.

Conclusion

To evaluate my study questions concerning how genetic evolution of H5N1 viruses takes place in space and time in Vietnam, I must comprehensively integrate genetic, environmental and population variables to explain how genetic characteristics of viruses differ by species, where and when barriers slow H5N1 viral evolution, and how population and environmental characteristics interact to influence molecular change. Central to this study is the idea that human modification of natural environments for purposes of poultry production creates places in space and time that either positively or negatively influence the spread and evolution of avian diseases, including avian influenza. It is only through the unique application of landscape genetics methods informed by disease ecology theory to the study of an anthropozoonotic pathogen that allows me to assess the validity of this concept.

References

1. Guo Y, Xu X & Wan X (1998) Genetic characterization of an avian influenza A (H5N1) virus isolated from a sick goose in China. *Zhonghua Shi Yan He Lin Chuang Bing Du Xue Za Zhi* 12: 322-325.

2. Xu X, Subbarao, Cox NJ & Guo Y (1999) Genetic characterization of the pathogenic influenza A/Goose/Guangdong/1/96 (H5N1) virus: Similarity of its hemagglutinin gene to those of H5N1 viruses from the 1997 outbreaks in Hong Kong. *Virology* 261: 15-19.

3. Subbarao K, *et al* (1998) Characterization of an avian influenza A (H5N1) virus isolated from a child with a fatal respiratory illness. *Science* 279: 393.

4. Claas ECJ, *et al* (1998) Human influenza A H5N1 virus related to a highly pathogenic avian influenza virus. *Lancet (1803-1803)* 351: 472.

5. Guan Y, *et al* (2002) H5N1 influenza viruses isolated from geese in southeastern china: Evidence for genetic reassortment and interspecies transmission to ducks. *Virology* 292: 16-23.

6. Sims LD, *et al* (2005) Origin and evolution of highly pathogenic H5N1 avian influenza in Asia. *Vet Rec* 157: 159.

7. Li KS, *et al* (2004) Genesis of a highly pathogenic and potentially pandemic H5N1 influenza virus in Eastern Asia. *Nature* 430: 209-213.

8. Duan L, *et al* (2008) The development and genetic diversity of H5N1 influenza virus in China, 1996-2006. *Virology* 380: 243-254.

9. Chen H, *et al* (2006) Establishment of multiple sublineages of H5N1 influenza virus in Asia: Implications for pandemic control. *Proc Natl Acad Sci U S A* 103: 2845-2850.

10. Ducatez MF, *et al* (2007) Molecular and antigenic evolution and geographical spread of H5N1 highly pathogenic avian influenza viruses in Western Africa. *J Gen Virol* 88: 2297-2306.

11. World Organization for Animal Health (OIE) (2011) Outbreaks of highly pathogenic avian influenza (subtype H5N1) in poultry, from 2003 to 17 February 2011.

12. World Health Organization (2011) Cumulative number of confirmed human cases of avian influenza A (H5N1) reported to WHO.

13. Zhao ZM, Shortridge KF, Garcia M, Guan Y & Wan XF (2008) Genotypic diversity of H5N1 highly pathogenic avian influenza viruses. *J Gen Virol* 89: 2182-2193.

14. World Health Organization (2008) Avian influenza-situation in Pakistan-update 2.2009.

15. World Health Organization (2006) Avian influenza - situation in Indonesia - update 16.2011.

16. Gilbert M*, et al* (2008) Mapping H5N1 highly pathogenic avian influenza risk in Southeast Asia. *Proc Natl Acad Sci U S A* 105: 4769-4774.

17. Martin V*, et al* (2011) Spatial distribution and risk factors of highly pathogenic avian influenza (HPAI) H5N1 in china. *PLoS Pathog* 7: e1001308.

18. Paul M*, et al* (2010) Anthropogenic factors and the risk of highly pathogenic avian influenza H5N1: Prospects from a spatial-based model. *Vet Res* 41:

19. Van Kerkhove MD*, et al* (2009) Poultry movement networks in Cambodia: Implications for surveillance and control of highly pathogenic avian influenza (HPAI/H5N1). *Vaccine* 27: 6345-6352.

20. Kilpatrick AM*, et al* (2006) Predicting the global spread of H5N1 avian influenza. *Proceedings of the National Academy of Sciences* 103: 19368-19373.

21. Gambotto A, Barratt-Boyes SM, de Jong MD, Neumann G & Kawaoka Y (2008) Human infection with highly pathogenic H5N1 influenza virus. *Lancet* 371: 1464-1475.

22. Peiris JS, de Jong MD & Guan Y (2007) Avian influenza virus (H5N1): A threat to human health. *Clin Microbiol Rev* 20: 243-267.

23. Chen W*, et al* (2001) A novel influenza A virus mitochondrial protein that induces cell death. *Nat Med* 7: 1306-1312.

24. WHO, OIE, FAO H, N, and Evolution Working Group (2008) Toward a unified nomenclature system for highly pathogenic avian influenza virus (H5N1). *Emerging Infectious Diseases* 14: e1.

25. Wan XF*, et al* (2008) Evolution of highly pathogenic H5N1 avian influenza viruses in Vietnam between 2001 and 2007. *PLoS ONE* 3: e3462.

26. Wan XF, Chen G, Luo F, Emch M & Donis R (2007) A quantitative genotype algorithm reflecting H5N1 avian influenza niches. *Bioinformatics* 23: 2368.

27. Wan XF*, et al* (2007) Computational identification of reassortments in avian influenza viruses. *Avian Dis* 51: 434-439.

28. Webster RG, Bean WJ, Gorman OT, Chambers TM & Kawaoka Y (1992) Evolution and ecology of influenza A viruses. *Microbiol Rev* 56: 152-179.

29. Janies D*, et al* (2007) Genomic analysis and geographic visualization of the spread of avian influenza (H5N1). *Syst Biol* 56: 321-329.

30. Wallace RG, Hodac H, Lathrop RH & Fitch WM (2007) A statistical phylogeography of influenza A H5N1. *Proc Natl Acad Sci U S A* 104: 4473-4478.

31. Nguyen DC*, et al* (2005) Isolation and characterization of avian influenza viruses, including highly pathogenic H5N1, from poultry in live bird markets in Hanoi, Vietnam, in 2001. *J Virol* 79: 4201-4212.

32. Smith GJ*, et al* (2006) Evolution and adaptation of H5N1 influenza virus in avian and human hosts in Indonesia and Vietnam. *Virology* 350: 258-268.

33. Fourment M & Gibbs M (2006) PATRISTIC: A program for calculating patristic distances and graphically comparing the components of genetic change. *BMC Evolutionary Biology* 6: 1.

34. Neumann G & Kawaoka Y (2006) Host range restriction and pathogenicity in the context of influenza pandemic. *Emerg Infect Dis* 12: 881-886.

35. Swayne DE (2007) Understanding the complex pathobiology of high pathogenicity avian influenza viruses in birds. *Avian Dis* 51: 242-249.

36. Matrosovich M*, et al* (1997) Avian influenza A viruses differ from human viruses by recognition of sialyloligosaccharides and gangliosides and by a higher conservation of the HA receptor-binding site. *Virology* 233: 224-234.

37. Stevens J*, et al* (2006) Structure and receptor specificity of the hemagglutinin from an H5N1 influenza virus. *Science* 312: 404-410.

38. Muramoto Y*, et al* (2006) Molecular characterization of the hemagglutinin and neuraminidase genes of H5N1 influenza A viruses isolated from poultry in Vietnam from 2004 to 2005. *J Vet Med Sci* 68: 527-531.

39. Shinya K*, et al* (2004) PB2 amino acid at position 627 affects replicative efficiency, but not cell tropism, of Hong Kong H5N1 influenza A viruses in mice. *Virology* 320: 258-266.

40. Geiss GK*, et al* (2002) Cellular transcriptional profiling in influenza A virus-infected lung epithelial cells: The role of the nonstructural NS1 protein in the evasion of the host innate defense and its potential contribution to pandemic influenza. *Proc Natl Acad Sci U S A* 99: 10736-10741.

41. Garcia-Sastre A (2006) Antiviral response in pandemic influenza viruses. *Emerg Infect Dis* 12: 44-47.

42. Cheung C*, et al* (2002) Induction of proinflammatory cytokines in human macrophages by influenza A (H5N1) viruses: A mechanism for the unusual severity of human disease?. *The Lancet* 360: 1831-1837.

43. Pfeiffer J, Pantin-Jackwood M, To TL, Nguyen T & Suarez DL (2009) Phylogenetic and biological characterization of highly pathogenic H5N1 avian influenza viruses (Vietnam 2005) in chickens and ducks. *Virus Res* 142: 108-120.

44. Dudley JP (2008) Public health and epidemiological considerations for avian influenza risk mapping and risk assessment. *Ecology & Society* 13: 21[online].

45. Dung Nguyen T*, et al* (2008) Multiple sublineages of influenza A virus (H5N1), Vietnam, 2005-2007. *Emerg Infect Dis* 14: 632-636.

46. Nguyen T*, et al* (2009) Characterization of a highly pathogenic avian influenza H5N1 virus sublineage in poultry seized at ports of entry into Vietnam. *Virology* 387: 250-256.

47. Wallace RG & Fitch WM (2008) Influenza A H5N1 immigration is filtered out at some international borders. *PLoS ONE* 3: e1697.

48. Shortridge KF & Stuart-Harris CH (1982) An influenza epicentre?. *Lancet* 2: 812-813.

49. Pfeiffer DU, Minh PQ, Martin V, Epprecht M & Otte MJ (2007) An analysis of the spatial and temporal patterns of highly pathogenic avian influenza occurrence in Vietnam using national surveillance data. *The Veterinary Journal* 174: 302-309.

50. Cristalli A & Capua I (2007) Practical problems in controlling H5N1 high pathogenicity avian influenza at village level in Vietnam and introduction of biosecurity measures. *Avian Dis* 51: 461-462.

51. Thorson A, Petzold M, Nguyen TK & Ekdahl K (2006) 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. *Arch Intern Med* 166: 119-123.

52. Dinh PN*, et al* (2006) Risk factors for human infection with avian influenza A H5N1, Vietnam, 2004. *Emerg Infect Dis* 12: 1841-1847.

53. Salomon R & Webster RG (2009) The influenza virus enigma. *Cell* 136: 402-410.

54. Rabinowitz P, Perdue M & Mumford E (2009) Contact variables for exposure to avian influenza H5N1 virus at the human-animal interface. *Zoonoses and Public Health* 9999:

55. Hayden F & Croisier A (2005) Transmission of avian influenza viruses to and between humans. *J Infect Dis* 192: 1311-1314.

56. Webster RG*, et al* (2007) Changing epidemiology and ecology of highly pathogenic avian H5N1 influenza viruses. *Avian Dis* 51: 269-272.

57. Brown JD, Swayne DE, Cooper RJ, Burns RE & Stallknecht DE (2007) Persistence of H5 and H7 avian influenza viruses in water. *Avian Dis* 51: 285-289.

58. Lipsitch M & Viboud C (2009) Influenza seasonality: Lifting the fog. *Proceedings of the National Academy of Sciences* 106: 3645-3646.

59. Shaman J & Kohn M (2009) Absolute humidity modulates influenza survival, transmission, and seasonality. *Proc Natl Acad Sci U S A* 106: 3243-3248.

60. Happold JR, Brunhart I, Schwermer H & Stärk KDC (2008) Surveillance of H5 avian influenza virus in wild birds found dead. *Avian Dis* 52: 100-105.

61. Kim JK, Negovetich NJ, Forrest HL & Webster RG (2009) Ducks: The "Trojan horses" of H5N1 influenza

62. Brown JD, Stallknecht DE, Beck JR, Suarez DL & Swayne DE (2006) Susceptibility of North American ducks and gulls to H5N1 highly pathogenic avian influenza viruses. *Emerg Infect Dis* 12: 1663-1670.

63. Pantin-Jackwood MJ & Swayne DE (2007) Pathobiology of Asian highly pathogenic avian influenza H5N1 virus infections in ducks. *Avian Dis* 51: 250-259.

64. Swayne DE & Pantin-Jackwood M (2006) Pathogenicity of avian influenza viruses in poultry. *Dev Biol (Basel)* 124: 61-67.

65. Lipatov AS*, et al* (2008) Domestic pigs have low susceptibility to H5N1 highly pathogenic avian influenza viruses. *PLoS Pathog* 4: e1000102.

66. Wang M*, et al* (2006) Food markets with live birds as source of avian influenza. *Emerg Infect Dis* 12: 1773-1775.

67. Skeik N & Jabr FI (2008) Influenza viruses and the evolution of avian influenza virus H5N1. *Int J Infect Dis* 12: 233-238.

68. Vong S, Ly S, Mardy S, Holl D & Buchy P (2008) Environmental contamination during influenza A virus (H5N1) outbreaks, Cambodia, 2006. *Emerging Infectious Diseases* 14: 1303.

69. Breban R, Drake JM, Stallknecht DE & Rohani P (2009) The role of environmental transmission in recurrent avian influenza epidemics. *PLoS Computational Biology* 5:

70. Tiensin T*, et al* (2007) Geographic and temporal distribution of highly pathogenic avian influenza A virus (H5N1) in Thailand, 2004-2005: An overview. *Avian Dis* 51: 182-188.

71. Gilbert M*, et al* (2006) Free-grazing ducks and highly pathogenic avian influenza, Thailand. *Emerg Infect Dis* 12: 227-234.

72. Tiensin T*, et al* (2009) Ecologic risk factor investigation of clusters of avian influenza A (H5N1) virus infection in Thailand. *J Infect Dis* 199: 1735-1743.

73. Dubos RJ (1987) *Mirage of health: utopias, progress and biological change,* (Rutgers University Press, New Brunswick), pp 282.

74. Hunter JM (1974) in *The Challenge of Medical Geography: Studies in Geography No. 6,* Ed Hunter JM (Department of Geography, University of North Carolina, Chapel Hill),

75. May JM (1958) *The Ecology of Human Disease: Studies in Medical Geography,* (MD Publications, New York).

76. Meade MS & Emch M (2010) *Medical geography,* (Guilford Press, New York).

77. Mayer JD (2000) Geography, ecology and emerging infectious diseases. *Soc Sci Med* 50: 937-952.

78. Mayer JD & Meade MS (1994) A reformed medical geography reconsidered. *The Professional Geographer* 46: 103-106.

79. Meade MS (1977) Medical geography as human ecology: The dimension of population movement. *The Geographical Review* 67: 379-393.

80. Manel S, Schwartz MK, Luikart G & Taberlet P (2003) Landscape genetics: Combining landscape ecology and population genetics. *Trends in Ecology & Evolution* 18: 189-197.

81. Holderegger R, Kamm U & Gugerli F (2006) Adaptive vs. neutral genetic diversity: Implications for landscape genetics. *Landscape Ecol* 21: 797-807.

82. Guillot G, Estoup A, Mortier F & Cosson JF (2005) A spatial statistical model for landscape genetics. *Genetics* 170: 1261-1280.

83. Storfer A*, et al* (2007) Putting the "landscape" in landscape genetics. *Heredity* 98: 128-142.

84. Balkenhol N*, et al* (2009) Identifying future research needs in landscape genetics: Where to from here?. *Landscape Ecol* 24: 455-463.

85. Holderegger R & Wagner HH (2006) A brief guide to landscape genetics. *Landscape Ecol* 21: 793-796.

86. Archie EA, Luikart G & Ezenwa VO (2009) Infecting epidemiology with genetics: A new frontier in disease ecology. *Trends in Ecology & Evolution* 24: 21-30.

87. Lampert KP, Rand AS, Mueller UG & Ryan MJ (2003) Fine-scale genetic pattern and evidence for sex-biased dispersal in the Túngara frog, *Physalaemus pustulosus*. *Mol Ecol* 12: 3325-3334.

88. Piertney SB, Maccoll ADC, Bacon PJ & Dallas JF (1998) Local genetic structure in red grouse (*Lagopus lagopus scoticus*): Evidence from microsatellite DNA markers. *Mol Ecol* 7: 1645-1654.

89. Prugnolle F, De Meeus T, Durand P, Sire C & Theron A (2002) Sex-specific genetic structure in *Schistosoma mansoni*: Evolutionary and epidemiological implications. *Mol Ecol* 11: 1231-1238.

90. Schwartz MK & McKelvey KS (2009) Why sampling scheme matters: The effect of sampling scheme on landscape genetic results. *Conserv Genet* 10: 441-452.

CHAPTER 2

H5N1 AVIAN INFLUENZA GENETIC VARIATION IN VIETNAM SHOWS BOTH SPECIES-SPECIFIC AND SPATIOTEMPORAL ASSOCIATIONS

Abstract:

Domestic poultry act as a reservoir for persistent H5N1 endemicity in Vietnam, and the circulation of poultry flocks across farms and to market is thought to drive the spatial movement and evolution of avian influenza viruses. Using a dataset of highly pathogenic H5N1 avian influenza viruses collected in domestic poultry in Vietnam from 2003 to 2007, I explore potential differences in genetic characteristics according to species of isolation and the spatio-temporal characteristics of the viruses. Through clustering algorithms and analysis of variance, I find that H5N1 viruses in Vietnam show distinct patterns of genetic differentiation according to the year in which they were isolated. There also exist differences in the amount of genetic change that chicken viruses experience as compared to duck viruses, with duck viruses showing higher rates of molecular evolution on all eight of influenza's gene segments. These findings suggest that molecular evolution of avian influenza viruses is continuous through time but also mediated by the species in which the viruses occur, information which has implications for prevention efforts.

Background

Highly pathogenic H5N1 avian influenza was first detected in Vietnam in 2001, and the country was part of a larger pandemic emergence of H5N1 across Southeast Asia in 2003 (1). Since 2003, Vietnam has remained one of the countries hardest hit by H5N1 avian influenza, with continuing poultry and human infection and mortality (2, 3). Vietnam is also a site of molecular evolution of H5N1 viruses, since 2003 there have been at least four novel types that have emerged in Vietnamese H5N1 isolates (4). Part of the reason for H5N1's persistence in Vietnam is socio-environmental: among Vietnamese there is a preference for live or freshly killed poultry, and a large percentage of rural Vietnamese rear their own backyard poultry flocks (5-9). Large numbers of susceptible birds, combined with the circulation of birds and people from farms to markets and a shared border with China, the source of many H5N1 viruses, drive the ongoing H5N1 epidemic (10-12).

Backyard poultry flocks in Vietnam are composed primarily of chickens and aquatic poultry such as ducks and geese. According to a 2003 livestock census, of Vietnam's 261 million domestic poultry, 73.5% were chickens and 26.3% were aquatic birds (8). Transmission of H5N1 within such domestic poultry is thought to provide the major mechanism by which avian influenza viruses remain endemic in Vietnam (10).

The epidemiology of H5N1 differs, however, between chickens and ducks. Prior to 2002, H5N1 infection in domestic ducks was usually asyptomatic or expressed as only minor clinical symptoms (13). Asymptomatic ducks could still shed infective viruses in their feces or saliva, however, for up to two weeks after infection (7, 10). Domestic ducks are sometimes referred to as "Trojan horses" of H5N1 infection, unnoticeably sustaining circulation of H5N1 in poultry (7, 14). Since 2002, however, novel genotypes of H5N1

viruses have emerged and been predominant in both domestic and wild birds that are highly pathogenic in ducks, producing severe systemic infection and lesions in multiple organs (15-18).

In contrast to ducks, chickens experience high (often 100%) mortality after infection with H5N1 viruses, and have symptomology of infection in multiple organs (5, 19). In recent years, H5N1 infections in chickens have become more virulent, indicated by less time between infection and death (17). While H5N1 infections in chickens are stronger, viruses shed in chicken feces remain infective for only one day, as compared to a week for duck feces (20). Ducks thus appear to have greater potential to contaminate water and other environmental surfaces with infective saliva and feces in terms of both duration of viral shedding and duration of shed viral infectivity.

Given these divergent epidemiologies and ecologies, I sought to answer whether the genetic characteristics of H5N1 avian influenza viruses were related to the species in which those viruses occurred. Specifically, does the amount of genetic change that has taken place between Vietnamese viruses and their ancestral virus vary according to species of isolation? Are there differences in the effect of species when considering the spatial and temporal characteristics of viral isolates? I used cluster analysis and multiple analysis of variance (MANOVA) to explore these questions.

Data & Methods

Data for the study consisted of 110 highly pathogenic H5N1 avian flu viruses isolated in Vietnam between 2003 and 2007 (Figure 1). Viruses were either publicly available in online repositories or collected by the National Centre for Veterinary Diagnostics (NCVD) of Hanoi, Vietnam. For each of the 110 isolates there is a complete genetic sequence as well as

the year of isolation and the province in which the virus was found. All viruses belong to a single genetic lineage originating in Hong Kong in 2001 which phylogenetic analysis indicates had a single introduction event into Vietnam (4, 21, 22). While the collection sites of the publicly available viruses are unknown, the NCVD collaborates with the Vietnamese Department of Animal Health's regional offices to detect H5N1 outbreaks in backyard poultry flocks, commercial farms and live bird markets.

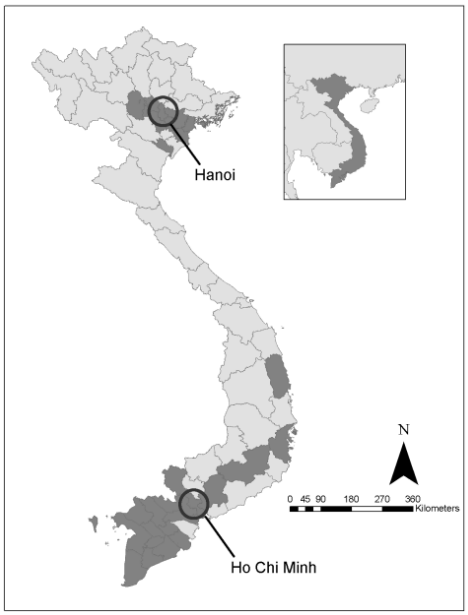

Figure 2.1: Distribution of 110 chicken and duck H5N1 viruses in Vietnam. Darkened provinces indicate locations of virus isolation. The locations of Vietnam's two largest cities, Hanoi and Ho Chi Minh City, are also shown.

Using the common ancestral virus (A/Duck/HongKong/821/2002) as the point of reference, geographic, temporal and genetic distance measures were created for the 125 virus dataset. Viruses were geocoded to the latitude and longitude of the centroid of the province of isolation using a geographic information system (GIS) in order to calculate the distance between the province centroid and Hong Kong's centroid. Temporal distance between the

Vietnamese viruses and the progenitor virus was calculated as the number of years since the isolation of the Hong Kong virus in 2002. Eight genetic distance measures, one for each of the influenza viruses eight gene segments, was calculated using PATRISTIC. A patristic distance is the length of the branches connecting two nodes of a phylogenetic tree, and indicates the amount of genetic change that exists between those two nodes (23).

In addition to the distance measures, each virus was assigned a species designation. The species of isolation was determined by the universal virus identification code (e.g. "Ck/VN/19/03" is a virus isolated in a Vietnamese chicken in 2003). Viruses isolated in chickens were assigned "1", ducks assigned "2". Within the dataset, 53 viruses were isolated in chickens and 57 in ducks.

Cluster analysis using a model clustering technique was carried out using the *mclust* package in R. Cluster analysis was used to assess whether viruses grouped according to genetic characteristics would reflect the species designations of the viruses. In model clustering, the eight genetic distance measures are used to partition the 110 viruses such that within-cluster likeness and between- cluster difference is maximized. The model cluster algorithm investigates a variety of shape and size constraints for the clusters, including equal and unequal volume, equal and unequal shape, and spherical, diagonal or ellipsoidal orientation, and returns indications of how well each of these ten models fit the dataset across different numbers of clusters. The number of clusters that viruses will be divided into, the shape of the clusters and the optimum cluster assignment is determined by the highest Bayesian Information Criterion (BIC). The BIC is calculated given the log-likelihood, the dimensionality of the data (8) and the number of mixture components (125), and varies greatly according to the model type and number of clusters.

Once viruses were assigned to clusters based on their genetic characteristics, those cluster assignments were mapped in the GIS according to geographic location of viral isolation. Maps of clusters were stratified both by year of viral incidence and species of isolation. This allowed us to assess whether the cluster assignments generated in the model clustering algorithm expressed spatial, temporal or species-specific patterns.

To ascertain whether the results I observed in the cluster analysis were statistically significant, MANOVA was performed in R to simultaneously assess the degree of variation present in the eight genetic distance measures attributable to three potential sources: species type (coded as a factor rather than a categorical variable to avoid the statistical implication that 2 ("duck") is numerically more valuable than 1 ("chicken")), geographic distance and temporal distance. Interaction plots were generated to show differences in genetic distances according to species and temporal distance for the eight gene segments.

Results

In partitioning the 110 viruses into clusters, the BIC score of 8127.612 indicated that an eight-cluster partitioning of the data with varying cluster volume and shape and orientation along the coordinate axes provided the best fit (Figure 2). The cluster assignment of all 110 viruses across the eight genetic segments is shown in Figure 3, with the clear grouping of viruses near to one another in genetic space assigned to the same cluster.

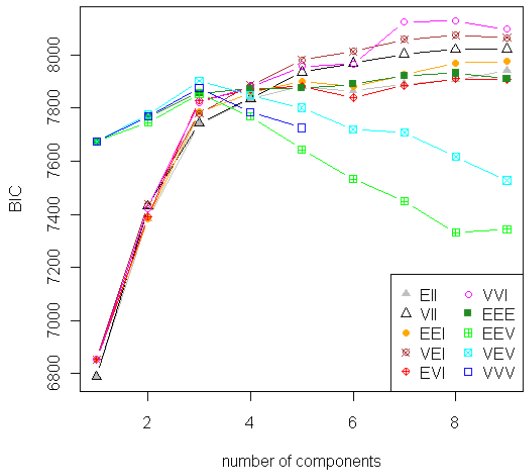

Figure 2.2: Bayesian Information Criterion (BIC) scores for ten types of clustering algorithms with variable numbers of clusters. The BIC is calculated given the log-likelihood, the dimensionality of the data and the number of mixture components. BIC scores indicate that a VVI (diagonal varying volume and shape) model with eight clusters, best describes the dataset.

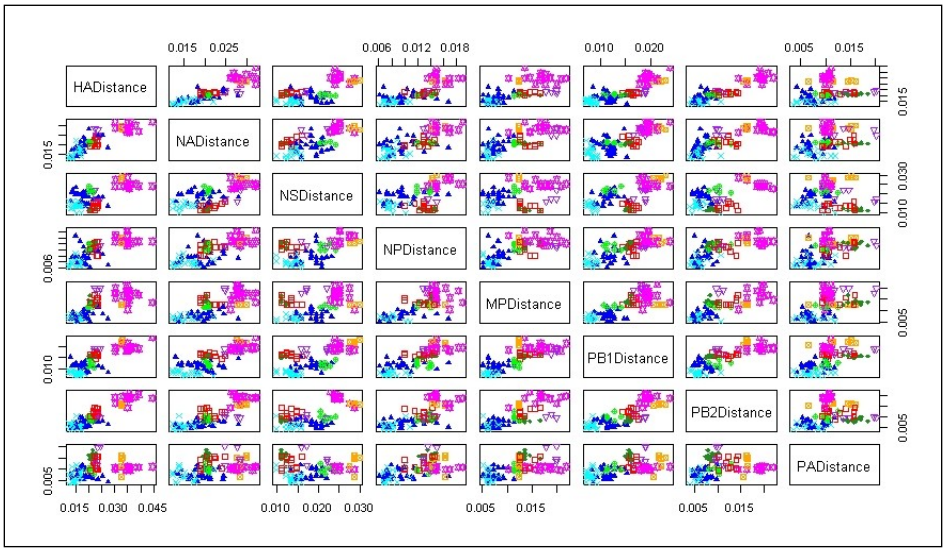

Figure 2.3: Eight genetic distance variables with individual viral cluster assignments shown. The chart is symmetrical along the diagonal, showing the cluster divisions for pairs of gene segments (e.g. HA versus NA). The clean division of viruses into eight clusters across eight distance variables is seen in the tight groupings of symbols.

When the cluster assignments were mapped according to the province of viral isolation, distinct temporal patterning was observed (Figure 4). In 2003, all viruses were assigned to cluster 1 & 2, and are only in northern Vietnam. In 2004, viruses were assigned to clusters 1, 2 and 3 in both northern and southern Vietnam. In 2005, there is a great

diversity of cluster assignments: northern Vietnamese isolates are assigned to clusters 3, 5 and 6, while southern Vietnamese isolates are assigned to clusters 1, 4, 5 and 6. In 2007, viruses are grouped into clusters 7 and 8, and are located primarily in southern Vietnam (although this spatial patterning is likely an artifact of the dataset itself).

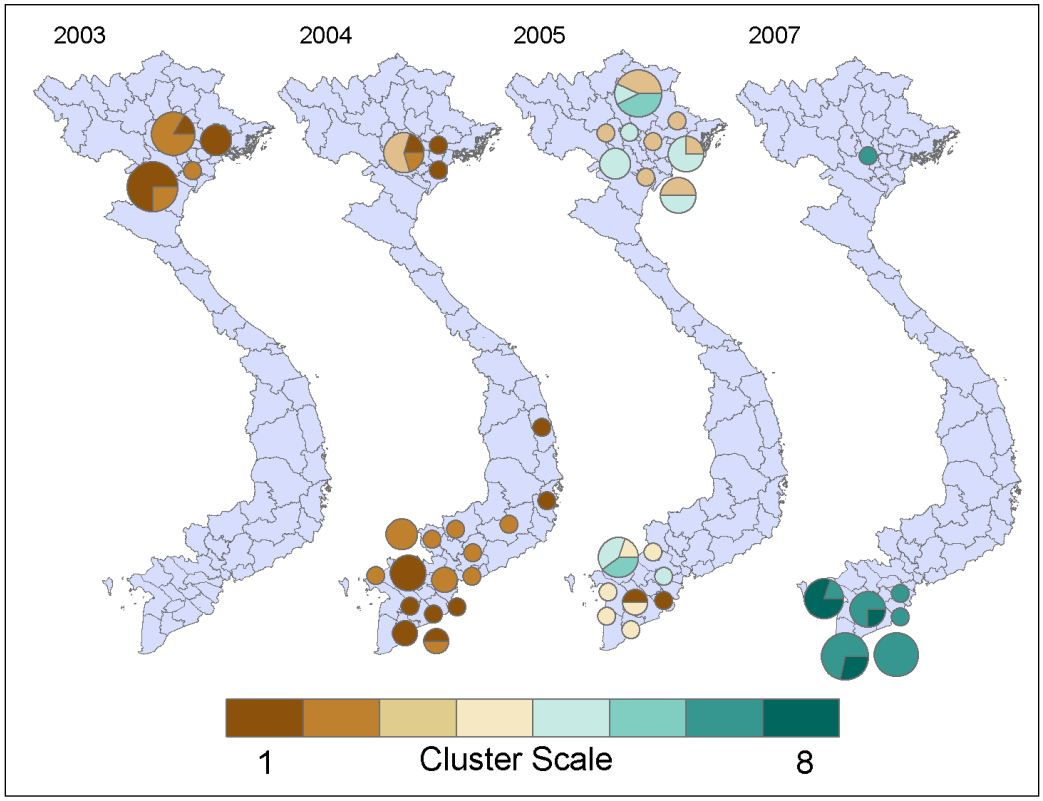

Figure 2.4: Cluster assignments according to province of isolation. Color indicates which cluster each virus was assigned, size of pie chart indicates the number of viruses located in that province (scale varies from 1 to 5).

The genetic characteristics of each viral cluster are closely aligned with the temporal pattern described above. The average genetic distance for each of the eight gene segments in each of the eight clusters is shown in Table 1. Clusters 1 and 2, observed primarily in 2003 and 2004, represent the viral isolates with the lowest genetic distance (.09882 and .07976 respectively), indicating they are closest genetically to the Hong Kong progenitor virus. Viruses taking place in 2005 were assigned to clusters 3 to 6, which have medium-scale average genetic distances (from .11995 to .14602). Viruses isolated in 2007 are furthest

away from the progenitor virus in both time and in genetic space, with total average distances of .17028 for cluster 7 and .16839 for cluster 8. Cluster 7 is composed of viruses with highest genetic distances for the PB2, HA and NA gene segments, while cluster 8 has the highest genetic distance for the PB1, NP and NS gene segments.

Cluster	PB2 Dist	PB1 Dist	PA Dist	HA Dist	NP Dist	NA Dist	MP Dist	NS Dist	Total
1	0.00673	0.01146	0.00779	0.01817	0.00956	0.01915	0.00774	0.01822	0.09882
2	0.00589	0.00814	0.00707	0.01509	0.00772	0.01582	0.00654	0.01348	0.07976
3	0.01185	0.01625	0.01245	0.02235	0.01240	0.02102	0.01364	0.01301	0.12297
4	0.01014	0.01211	0.01004	0.02058	0.01287	0.02061	0.01173	0.02187	0.11995
5	0.00858	0.01571	0.01599	0.02126	0.01393	0.02057	0.01372	0.01140	0.12116
6	0.01009	0.01828	0.01605	0.02304	0.01357	0.02686	0.01925	0.01890	0.14602
7	0.01905	0.01953	0.01050	0.03518	0.01540	0.02894	0.01678	0.02491	0.17028
8	0.01617	0.02184	0.01291	0.03257	0.01516	0.02875	0.01243	0.02856	0.16839

Table 2.1: Average genetic distances for the viruses assigned to each cluster. The sum of the eight average genetic distances indicates which clusters are groupings of viruses with low genetic distance (e.g. Clusters 1 & 2) versus those that are clusters of viruses with high genetic distance (Clusters 7 & 8).

Small geographic differences in the spatial patterns of cluster assignments can be seen. In 2004 the genetic characteristics of isolates are similar between northern and southern Vietnam (clusters 1 and 3), while in 2005 there is divergence in the genetic characteristics of northern and southern isolates. As described previously, cluster 3 is found only in northern Vietnam in 2005 and cluster 4 is found only in southern Vietnam in that year. The genetic distances associated with these two clusters, however, are quite similar. Clusters 5 and 6 are distributed across the country.

When cluster assignments are mapped according to the species of viral isolation, some slight differences between the genetic characteristics of chicken and duck viruses are observable. Cluster assignments for chicken H5N1 isolates are mapped in Figure 5, while duck H5N1 isolates are shown in Figure 6. In 2004, chicken isolates are grouped into clusters 1 and 2, while duck isolates are also grouped into cluster 3. In 2005, chicken viruses

fall into clusters 3, 4, 5 and 6. Duck viruses in 2005 are assigned to clusters 1, 3, 4, 5 and 6. In 2007, both duck and chicken viruses are assigned to cluster 7, the group with the second highest amount of genetic change from the progenitor virus. Only duck viruses in southern Vietnam are found in the highest cluster of genetic distances.

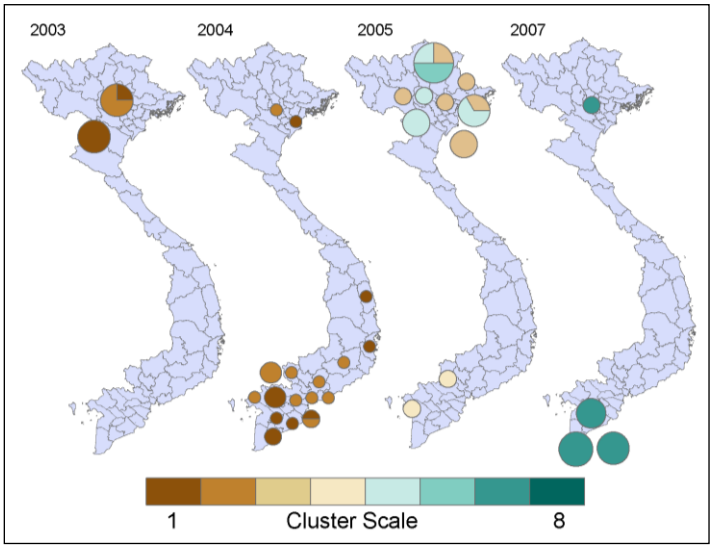

Figure 2.5: H5N1 viruses isolated in chickens, according to cluster assignment and province of isolation. Color indicates which cluster each virus was assigned, size of pie chart indicates the number of viruses located in that province.

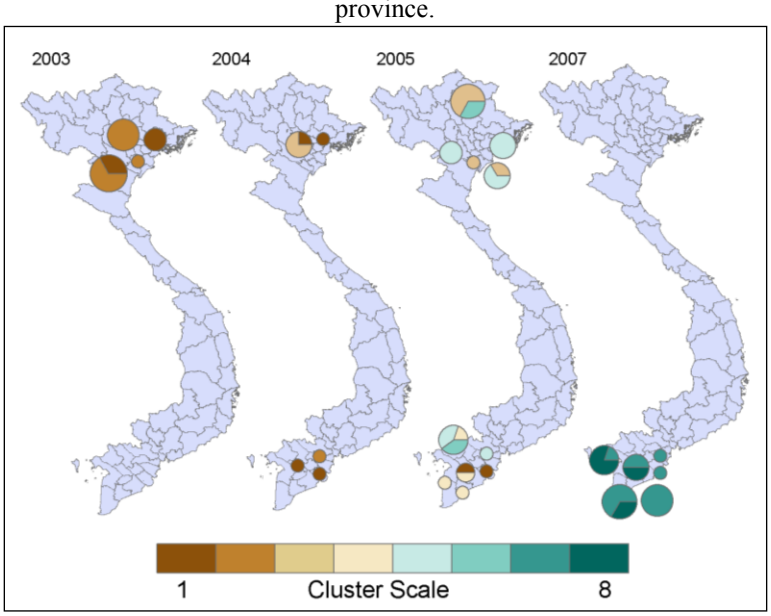

Figure 2.6: H5N1 viruses isolated in ducks, according to cluster assignment and province of isolation. Color indicates which cluster each virus was assigned, size of pie chart indicates the number of viruses located in that province.

In the overall MANOVA model (see Table 2), there is statistically significant variation in mean genetic distance across all 8 gene segments according to the species of isolation, as well as according to geographic and temporal distance.

	df	Wilks λ	approx F	Pr(>F)	
Species	1	0.345	22.52	< 2.2E-16	***
Temporal Distance	1	0.078	139.432	< 2.2E-16	***
Geographic Distance	1	0.477	13.037	1.59E-12	***
Species:Temporal Distance	1	0.827	2.484	0.01715	.
Species:Geographic Distance	1	0.909	1.186	0.31557	
Temporal Distance: Geographic Distance	1	0.386	18.915	< 2.2E-16	***
Species:Temporal Distance:Geographic Distance	1	0.934	0.841	0.56861	
Residuals	102				

Significance values: 0 '***' .0001 '**' .01 '*' 05 '.'

Table 2.2: Summary of the MANOVA using eight genetic distance measures as the dependent variable and species, temporal distance and geographic distance as explanatory variables.

Individual ANOVA results for each of the eight gene segments (Table 3) show that there exist significant differences in genetic distance according to species designation. Temporal distance is also a strong axis of variation, with statistically significant differences in genetic distances when stratified by year of incidence (as measured by temporal distance). The F statistics, an indication of how strongly the null hypothesis is rejected, are much higher for the temporal distance model than for the species model. Four of the gene segments also exhibited significant differences in genetic distance according to the amount of geographic distance from the progenitor virus in Hong Kong.

Model	PB2 Distance	PB1 Distance	PA Distance	HA Distance	NP Distance	NA Distance	MP Distance	NS Distance
Species	57.3278	78.8278	3.652	91.2428	35.5424	41.6047	22.756	28.1479
Temporal Distance	491.564	288.7553	25.4699	627.6387	210.4701	344.6868	137.6578	96.5648
Geographic Distance	2.7497	46.2216	35.2827	0.2566	21.2953	1.7582	17.1428	22.7102
Species:Temporal Distance	0.1829	1.0706	0.6516	0.0424	0.1126	2.7419	2.2531	4.5904
Species:Geographic Distance	2.5844	1.5705	0.5217	0.8907	0.0009	2.0448	0.6665	1.7636

Temporal Distance: Geographic Distance	24.1043	3.3532	15.2456	26.5468	2.3255	0.8508	3.6359	47.4919
Species:Temporal Distance:Geographic Distance	0.038	0.8617	0.0376	0.1031	0.5557	0.0466	3.3536	1.178

Table 2.3: Summary of F statistics for eight individual ANOVA models. Shading represents statistical significance. Dark shading equals p-values of <.01, light shading equals p-values of <.1, no shading indicates a lack of statistical significance.

Interaction plots (Figure 7) of mean genetic distances according to species type and temporal distance demonstrate how the genetic profiles of the 110 H5N1 viruses vary according to both gene segment and species of isolation. The scale of genetic distance for the HA and NA gene segments are larger than for the other gene segments, indicating greater genetic change on those two sections of the influenza genetic code between the progenitor virus and the Vietnam H5N1 dataset.

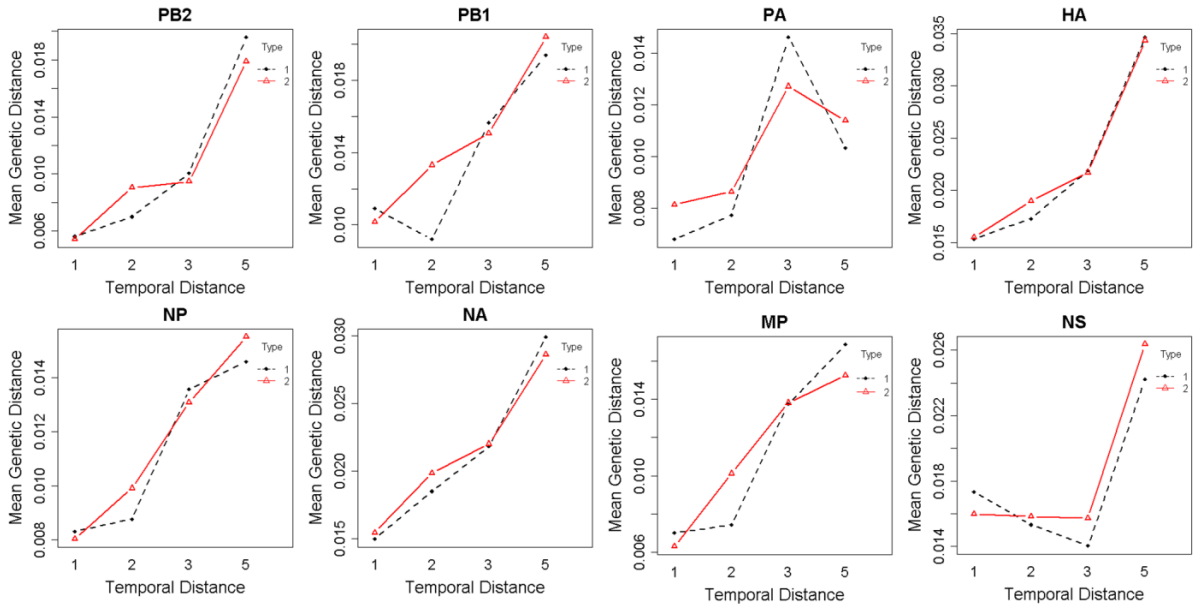

Figure 2.7: Interaction plots for mean genetic distance according to temporal distance and species type. Type 1 viruses are chicken isolates (shown in black), type 2 viruses are duck isolates (shown in red).

Variation in genetic distance for chickens versus ducks is seen in the divergence of the Type 1 and Type 2 lines, this divergence occurs in different years for each gene segment. The HA gene segment, for instance, appears to have less variation according to species than does the PB1 or the MP gene segment (average genetic distances for each gene segment

according to species are shown in Table 4). The strong influence of time is also shown in the interaction plots, with mean genetic distance spiking as time between the isolation of the progenitor virus and year of isolation in Vietnam increases.

	PB2 Dist	PB1 Dist	PA Dist	HA Dist	NP Dist	NA Dist	MP Dist	NS Dist	Total
Chicken	0.06231	0.08903	0.07414	0.15683	0.07288	0.13373	0.07669	0.09722	0.76284
Duck	0.08734	0.12440	0.09004	0.18726	0.09993	0.18056	0.10223	0.15207	1.02384

Table 2.4: Average genetic distance on each gene segment stratified by species, with total average genetic distance among chicken and duck H5N1 viruses.

Discussion

Mapping the cluster assignments of the 110 H5N1 isolates indicates the strong influence of time on genetic characteristics of the viruses. Clear temporal patterns in the cluster assignments can be seen, with viruses at low genetic distances from the progenitor virus grouped together in 2003 and 2004, viruses at further genetic distances assigned to clusters in 2005 and the viruses with the greatest genetic distance from the 2002 Hong Kong virus detected in 2007 and assigned to clusters 7 and 8. The strength with which the temporal progression of the clusters mirrors their genetic characteristics, but with much weaker spatial links to cluster patterns, suggests that genetic change in the dataset is primarily a factor of time, not space.

The results from the MANOVA and the interaction plots further indicate the differences in genetic characteristics of viruses across the length of the study period. For seven of the eight gene segments, 2007 (represented as 5 years of temporal distance from the progenitor virus) has the highest mean genetic distance among viruses of both species.

Some differences in cluster assignments between chickens and ducks were observed in the cluster analysis, although they were not as robust as the dissimilar epidemiology of H5N1 in the two species would have suggested. However, only duck viruses were assigned

to cluster 8, the cluster containing those viruses with the greatest genetic distances from the progenitor virus. The average genetic distances for each gene segment seen in Table 4 also indicate that across all eight segments of the H5N1 genetic code the duck viruses isolated in Vietnam had greater amounts of genetic change. While the variation in genetic distances between species is weaker than between years of isolation, it is still significantly different across all eight gene segments and in the summary model. This is further confirmed in the interaction plots, where chicken and duck viruses in the first year of the study period, 2003, have generally similar genetic distance means, but diverge in the following years.

The differing epidemiology of H5N1 viral infections in chickens and ducks appears to also express itself in the genetic characteristics of the viruses isolated in each of those species. As chickens typically experience higher rates of mortality and faster courses of infection, there is perhaps less time for them to experience co-infection with multiple viral strains or to transmit the virus to other chickens in their flock. In contrast, the longer duration of infection in asymptomatic but virus-shedding ducks could allow for greater viral mixing, thus reassortment and mutation, and account for the greater genetic distance seen in duck isolates in Vietnam. This could also be the result of greater H5N1 vaccine efficacy in chickens, such that viruses were more likely to be isolated in ducks than in chickens in the latter years of the dataset.

Previous research in Thailand implicated free-ranging duck populations as a driver of H5N1 incidence and suggested that new restrictions on the housing and grazing of ducks would decrease H5N1 outbreaks in that country (24, 25). Perhaps free-ranging backyard duck populations in Vietnam are driving not only H5N1 incidence in Vietnam but also viral evolution. If so, application of Thailand's laws on duck husbandry to Vietnam could reduce

not only the persistence of H5N1 avian influenza but the molecular progression of viruses as well.

References

1. Sims LD*, et al* (2005) Origin and evolution of highly pathogenic H5N1 avian influenza in Asia. *Vet Rec* 157: 159.

2. World Organization for Animal Health (OIE) (2011) Outbreaks of highly pathogenic avian influenza (subtype H5N1) in poultry, from 2003 to 17 February 2011.

3. World Health Organization (2011) Cumulative number of confirmed human cases of avian influenza A (H5N1) reported to WHO.

4. Wan XF*, et al* (2008) Evolution of highly pathogenic H5N1 avian influenza viruses in Vietnam between 2001 and 2007. *PLoS ONE* 3: e3462.

5. Nguyen DC*, et al* (2005) Isolation and characterization of avian influenza viruses, including highly pathogenic H5N1, from poultry in live bird markets in Hanoi, Vietnam, in 2001. *J Virol* 79: 4201-4212.

6. Thorson A, Petzold M, Nguyen TK & Ekdahl K (2006) 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. *Arch Intern Med* 166: 119-123.

7. Peiris JS, de Jong MD & Guan Y (2007) Avian influenza virus (H5N1): A threat to human health. *Clin Microbiol Rev* 20: 243-267.

8. Pfeiffer DU, Minh PQ, Martin V, Epprecht M & Otte MJ (2007) An analysis of the spatial and temporal patterns of highly pathogenic avian influenza occurrence in Vietnam using national surveillance data. *The Veterinary Journal* 174: 302-309.

9. Cristalli A & Capua I (2007) Practical problems in controlling H5N1 high pathogenicity avian influenza at village level in Vietnam and introduction of biosecurity measures. *Avian Dis* 51: 461-462.

10. Chen H*, et al* (2006) Establishment of multiple sublineages of H5N1 influenza virus in Asia: Implications for pandemic control. *Proc Natl Acad Sci U S A* 103: 2845-2850.

11. Wallace RG & Fitch WM (2008) Influenza A H5N1 immigration is filtered out at some international borders. *PLoS ONE* 3: e1697.

12. Shortridge KF & Stuart-Harris CH (1982) An influenza epicentre?. *Lancet* 2: 812-813.

13. Swayne DE (2007) Understanding the complex pathobiology of high pathogenicity avian influenza viruses in birds. *Avian Dis* 51: 242-249.

14. Kim JK, Negovetich NJ, Forrest HL & Webster RG (2009) Ducks: The "Trojan horses" of H5N1 influenza. *Influenza and Other Respiratory Viruses* 3: 121

15. Brown JD, Stallknecht DE, Beck JR, Suarez DL & Swayne DE (2006) Susceptibility of North American ducks and gulls to H5N1 highly pathogenic avian influenza viruses. *Emerg Infect Dis* 12: 1663-1670.

16. Pantin-Jackwood MJ & Swayne DE (2007) Pathobiology of Asian highly pathogenic avian influenza H5N1 virus infections in ducks. *Avian Dis* 51: 250-259.

17. Swayne DE & Pantin-Jackwood M (2006) Pathogenicity of avian influenza viruses in poultry. *Dev Biol (Basel)* 124: 61-67.

18. Webster RG, *et al* (2007) Changing epidemiology and ecology of highly pathogenic avian H5N1 influenza viruses. *Avian Dis* 51: 269-272.

19. Muramoto Y, *et al* (2006) Molecular characterization of the hemagglutinin and neuraminidase genes of H5N1 influenza A viruses isolated from poultry in Vietnam from 2004 to 2005. *J Vet Med Sci* 68: 527-531.

20. Gambotto A, Barratt-Boyes SM, de Jong MD, Neumann G & Kawaoka Y (2008) Human infection with highly pathogenic H5N1 influenza virus. *Lancet* 371: 1464-1475.

21. Dung Nguyen T, *et al* (2008) Multiple sublineages of influenza A virus (H5N1), Vietnam, 2005-2007. *Emerg Infect Dis* 14: 632-636.

22. Smith GJ, *et al* (2006) Evolution and adaptation of H5N1 influenza virus in avian and human hosts in Indonesia and Vietnam. *Virology* 350: 258-268.

23. Fourment M & Gibbs M (2006) PATRISTIC: A program for calculating patristic distances and graphically comparing the components of genetic change. *BMC Evolutionary Biology* 6: 1.

24. Tiensin T, *et al* (2007) Geographic and temporal distribution of highly pathogenic avian influenza A virus (H5N1) in Thailand, 2004-2005: An overview. *Avian Dis* 51: 182-188.

25. Gilbert M, *et al* (2006) Free-grazing ducks and highly pathogenic avian influenza, Thailand. *Emerg Infect Dis* 12: 227-234.

CHAPTER 3

LINKING GENETIC DIFFERENCES TO GEOGRAPHIC PATTERNS

The previous chapter provides evidence for differential genetic characteristics among duck and chicken H5N1 avian influenza viruses. Viruses isolated in ducks have greater average genetic distance from a progenitor Hong Kong virus than do viruses isolated in chickens, and this difference is significant. A strong relationship between time and genetics was also observed, the genetic characteristics of viruses varied significantly by the year in which they were isolated. Viruses, regardless of the species in which they were isolated, that were isolated in later years had greater genetic distance from the progenitor virus than those viruses isolated soon after the progenitor virus. Cluster assignments for viruses also appeared to be more associated with the year of viral incidence than species of incidence. These results suggest that the passage of time is strongly associated with the genetic change of viruses, and that rates of genetic evolution may differ between ducks and chickens.

The interaction between geographic distance and genetics was less clear than that of time and species. There appears to be significant variation in genetic characteristics according to geographic distance for some gene segments but not for others, indicating that perhaps some gene segments respond differently to landscape variation. Additionally, it does not appear that geography is an important factor in differentiating between clusters of

genetically similar viruses, although southern Vietnam is the location of clusters with the highest amount of genetic difference from a progenitor Hong Kong virus. This southern clustering could have more to do with those viruses being isolated in ducks, however, than their geographic location.

The first step in landscape genetics research is to explore and describe patterns of genetic characteristics across a landscape. The strong variation in genetic characteristics of viruses by the year in which they were isolated, and the uncertain relationship between geography and genetics, as described in Chapter 2 suggested further exploration of potential differences between viruses isolated in different years and in different regions of Vietnam. The presence of barriers or boundaries to gene flow is a primary tool for examining the effect of landscape on genetic outcomes, and whether the impact of landscape characteristics on gene flow varies in time. The following chapter explores the potential existence of barriers to genetic exchange among avian influenza viruses in Vietnam, examining whether variation in landscape characteristics is reflected in genetic patterns. This approach and these methodologies for assessing the presence of boundaries to gene flow have not previously been applied to the study of H5N1 avian influenza genetics.

CHAPTER 4

H5N1 AVIAN INFLUENZA VIRUSES ENCOUNTER FEW BARRIERS TO GENE FLOW IN VIETNAM

Abstract

Locating areas where genetic change is inhibited can illuminate underlying processes that drive evolution of pathogens. The persistence of H5N1 avian influenza in Vietnam since 2003, and the continuous molecular evolution of Vietnamese avian influenza viruses, indicates that local environmental factors are supportive not only of incidence but also of viral adaptation. I sought to explore whether gene flow is constant across Vietnam, or whether there exist boundary areas where gene flow exhibits discontinuity. Using a dataset of 125 H5N1 highly pathogenic avian influenza viruses, principle components analysis and wombling analysis are used to indicate the location, magnitude and statistical significance of genetic boundaries. Results show that a small number of geographically minor boundaries to gene flow in H5N1 avian influenza viruses exist in Vietnam, but that overall there is little division in genetic exchange from north to south. This suggests that changes in population and environment characteristics from one region to another do not act as barriers to viral incidence or evolution, and that H5N1 avian influenza is able to spread relatively unimpeded across the country.

Background

Avian influenza viruses can undergo dramatic changes in genetic sequence across short temporal and spatial distances. This ability for rapid molecular evolution enables influenza to adapt to overcome host immunological responses, to better survive in the environment, and to more easily transmit from infected to susceptible host (1-3). Tracking the evolution of avian influenza viruses is vital to prevention and control measures, and detecting areas where gene flow is blocked can illuminate how viral evolution varies across a landscape. Gene flow is the transfer or movement of genetic characteristics between populations, in this case from one H5N1 viruses to another.

Highly pathogenic H5N1 avian influenza has been continuously circulating in both human and poultry populations in Vietnam since 2001. Multiple cultural and environmental factors, such as maintenance of backyard poultry flocks and the popularity of live bird markets, allow H5N1 to persist despite government vaccination campaigns and other containment efforts (4-6). Vietnam is also a site of regular molecular evolution of H5N1 viruses, several new types unique to Vietnam have emerged since 2003 (7).

Genetic diversity is reliant upon limited gene flow among populations and either selection pressures in local-level environments or random chance (i.e. genetic shift among viruses concurrently infecting a host) (8). Identifying boundaries to gene flow, common in ecology and in population genetics, is useful to generate hypotheses about underlying processes that drive viral incidence and evolution (9). Difference boundaries, hereafter referred to as either barriers or boundaries, are located in spatial zones of rapid genetic change, and indicate the presence of either a sharp environmental transition and/or a low rate of viral exchange (10).

Previous studies of H5N1 avian influenza suggest that political borders can act as boundaries to gene flow, and that not all viral strains are able to cross from China into Vietnam (11). The presence or absence of similar filtering boundaries within Vietnam itself has not been established, however, leaving unanswered questions about how H5N1 influenza gene flow varies across the country. Given the distinct regional patterning of human and bird populations, with high population densities in the Red River and Mekong River deltas and the major cities of Hanoi and Ho Chi Minh City, and divergence in environmental characteristics along the north-south extent of Vietnam, I sought to explore whether there exist significant boundaries to H5N1 avian influenza genetic exchange and what population-environment characteristics could underlie the spatial and temporal attributes of those boundaries.

Data & Methods

To investigate the potential presence of barriers to gene flow, I used a dataset consisting of 125 highly pathogenic H5N1 avian influenza viruses isolated across Vietnam from 2003 to 2007 (Figure 1). These viruses were primarily found in domestic poultry, such as chickens and ducks, but were also sampled from other bird species, such as geese and quail, and from environmental samples of soil and water. For each virus there exists data on the province where the virus was found (representing 28 of Vietnam's 63 provinces) and the year in which the virus was sampled, as well as a full-length or nearly full-length genetic sequence. Phylogenetic analyses of the 125 viruses indicate that each of their eight gene segments belong to a single viral lineage, descendant from A/Duck/HongKong/821/2002 (7).

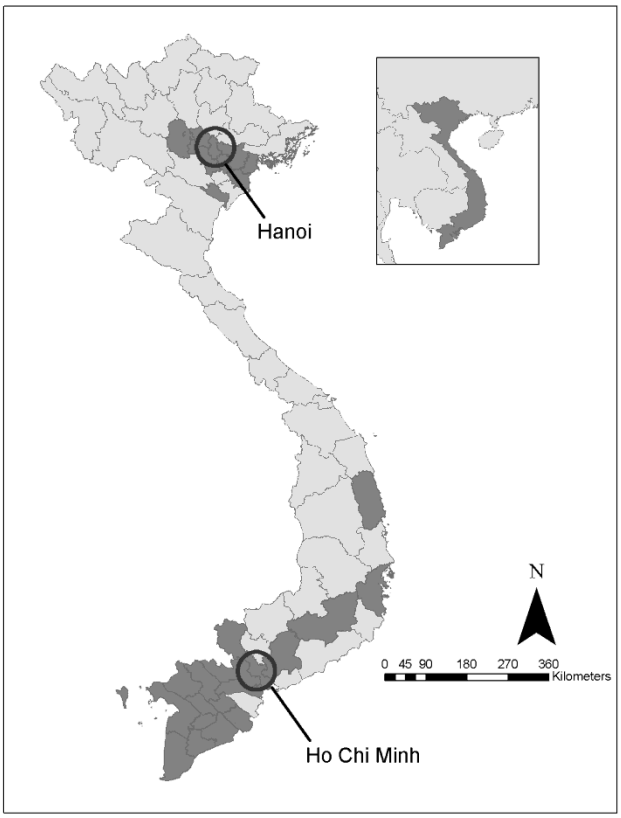

Figure 4.1: Provincial locations of isolation for 125 Vietnamese H5N1 avian influenza viruses. The 28 provinces where avian influenza samples were collected are darkened, and primarily represent the areas around Hanoi and Ho Chi Minh City, and the southern Mekong River delta.

Two genetic distance measures were generated from the dataset via a maximum likelihood phylogenetic tree constructed using nucleotide sequences. First, patristic genetic distance for each of the eight gene segments in the 125 viruses from the ancestral Hong Kong virus (a "straight line" genealogical distance) was calculated using PATRISTIC. Patristic distance indicates a total amount of genetic change that exists between genetic sequences by summing the lengths of branches in the phylogenetic tree between two viruses. Secondly, a 125x125 matrix of genetic distances between viruses in the Vietnamese dataset was generated, also using PATRISTIC (12). Thus, for each virus in the dataset we have not only its molecular distance from the progenitor virus but also the distance from every other virus in the dataset.

The eight straight line genetic distances were used as variables in a principal components analysis (PCA). PCA is used to explore the underlying structure of a dataset, and works by converting the observed values on a set number of possibly correlated variables into a new set of values for uncorrelated principal component variables. This data transformation results in a first principal component (or factor) that accounts for as much variance in the observed dataset as possible, a second principal component that is uncorrelated with the first and accounts for as much remaining variability as possible, and so on. The number of principal components that exist is always less than or equal to the original number of variables in the dataset. Each data observation is assigned a score for each factor detected.

PCA on the 125 observations of eight genetic distance measures was conducted in SPSS using a correlation method, varimax rotation, returning only principal components with Eigenvalues greater than 1. Factor scores for each of the 125 viruses were graphed according to the region in which each virus was isolated (northern Vietnam versus southern Vietnam) and the year of incidence. This allowed for exploration of whether sharp distinctions in PCA scores could be observed across these categories, indicating the presence of potential regional or temporal genetic dissimilarity and, by extension, the existence of genetic discontinuity due to boundaries.

Previous studies of genetic discontinuity have indicated that areas of sharp change in a smoothed map of PCA factor scores can signal the presence of boundaries to gene flow (13, 14). Factor scores for each of the 125 viruses from the first two principal components were mapped across Vietnam using a geographic information system (GIS) and an inverse distance weighted (IDW) interpolation method. IDW is a more appropriate interpolation method than

kriging for this dataset given the irregular spacing of the data. Viruses were grouped according to the province of isolation, and then a province-level factor score for each principal component was calculated by averaging across all viruses in each province. These average factor scores were then assigned to the geographic center of the province in the GIS. IDW interpolation methods were then used to generate a smoothed map of average factor scores across Vietnam, enabling observation of areas of steep changes in PCA scores.

Wombling is one of the original methods used for delineation of barriers to gene flow. Womble (1951) proposed that when examining genetic cline across geographic space, boundaries to genetic exchange exist where the cline exhibits steep gradients (15). Detecting areas of steep slopes of genetic change can illuminate places where gene flow is inhibited by environmental or other features. In wombling, there is no constraint on the position, orientation or shape of detected boundaries (8).

The geographic scale at which the influenza data was collected led to several different viruses being assigned to the same geographic location at the center of the province of isolation. In the PCA analysis the scores for viruses taking place in the same province were averaged before being mapped. In contrast, before conducting the wombling boundary analysis, the average genetic distance for each province on all eight gene segments was calculated by collapsing the 125x125 distance matrices into 28x28 distance matrices. Each of the eight 28x28 distance matrices was then associated with the 28 latitude and longitude coordinates of the province centroids as calculated in the GIS. Collapsing the matrices was preferred over randomly assigning viruses to geographic locations within the provinces as to avoid falsely detecting genetic boundaries between viruses located in the same provinces.

In wombling, magnitude indicates the average absolute slope of surfaces at a point, and direction indicates the orientation of the slope (10). A triangulation wombling procedure was implemented in PASSaGE to account for the irregular spatial sampling of the influenza datapoints (16). Under triangulation wombling, potential boundary elements are located at the center of Delaunay triangles created among the observed datapoints. For each boundary element, the magnitude is calculated as

$$m = \sqrt{\left[\frac{\partial f(x,y)}{\partial x}\right]^2 + \left[\frac{\partial f(x,y)}{\partial y}\right]^2}$$

where $f(x,y) = ax + by + c$ and constants a, b and c are calculated from

$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} x_A & y_A & 1 \\ x_B & y_B & 1 \\ x_C & y_C & 1 \end{bmatrix}^{-1} \begin{bmatrix} Z_A \\ Z_B \\ Z_C \end{bmatrix}.$$

The direction of the slope of change is calculated at each boundary element as

$$\theta = arctan\left(\frac{\partial f(x,y)/\partial x}{\partial f(x,y)/\partial y}\right).$$

Once the surface of genetic change, with magnitude and direction calculated at each boundary element, is created, neighboring boundary elements are linked into boundaries if they satisfied three criteria. The first is that they fall into the top ten percent of all calculated magnitudes. The second is that the directions of the two boundary elements must be less than 90^0 different, so that boundary elements of opposing gradients are not connected (Figure 2). Finally, boundary elements are connected into boundaries only if the angle of their connection is more than 30^0 different than the direction of the slope. So-called "singleton" boundaries consist of barrier elements that have magnitudes of change large enough to meet the first criteria, but that are not located next to other boundary elements that

meet the criteria. The statistical significance of all detected boundaries was evaluated based on 999 random permutations of the input data 28x28 matrix. Boundaries were mapped onto province centroids in the dataset to visualize their spatial patterns.

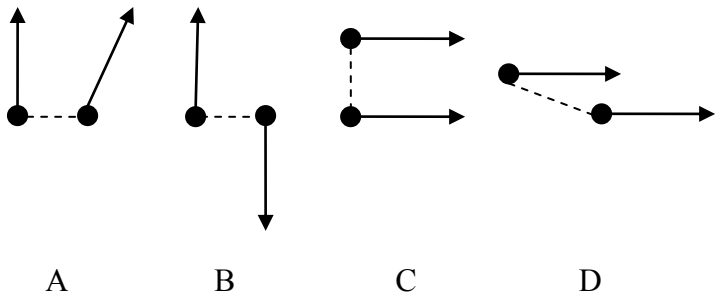

Figure 4.2: Two criteria for boundary element connections. In drawing A, two boundary elements (represented by circles) may be connected into a boundary because the directions of their associated slopes are less than 90^0 different. The boundary elements in drawing B have opposing directions and will not be combined into a boundary. The boundary elements in drawing C will be connected into a boundary because the bearing of the connection is greater than 30^0 different than the direction of the slope. Boundary elements in drawing D have similar bearing for both the boundary connection and the direction of change, so connecting them would not accurately reflect the presence of a boundary. Adapted from (16).

Given the temporal range of the dataset, spanning five years of viral isolation, the potential existed for year-specific genetic boundaries to be obscured in an analysis of the overall dataset. To detect the presence of such boundaries, the 125 viruses were divided into three separate temporally contiguous groups: those that occurred in 2003-2004 (55 viruses), 2004-2005 isolates (82 viruses) and 2005-2007 isolates (70 viruses). Eight genetic distance matrices (one for each gene segment) were then generated for these viral groupings, again with the genetic distances for isolates from the same province being averaged. Wombling was then performed on all eight genetic matrices for each of the three groupings, and the results were mapped onto the province centroids in the dataset.

Results

Two principal components with eigenvalues of greater than 1 were detected in the analysis of the eight straight-line influenza genetic distances. The first component accounted for 47.18% of the variance in the dataset, the second accounted for 37.29%, for a total explanation of 84.47% of the variance. When the PCA scores for these two factors were graphed for all 125 viruses based on the location of incidence (in either northern or southern Vietnam) some regional distinctions in PCA scores were observed (Figure 3). Northern and southern viral PCA scores were mixed across the upper and lower left quadrants of the graph (negative factor 1 scores, both positive and negative factor 2 scores). A cluster of southern H5N1 viruses is seen at the positive end of the factor 1 scale, this cluster also includes one northern virus.

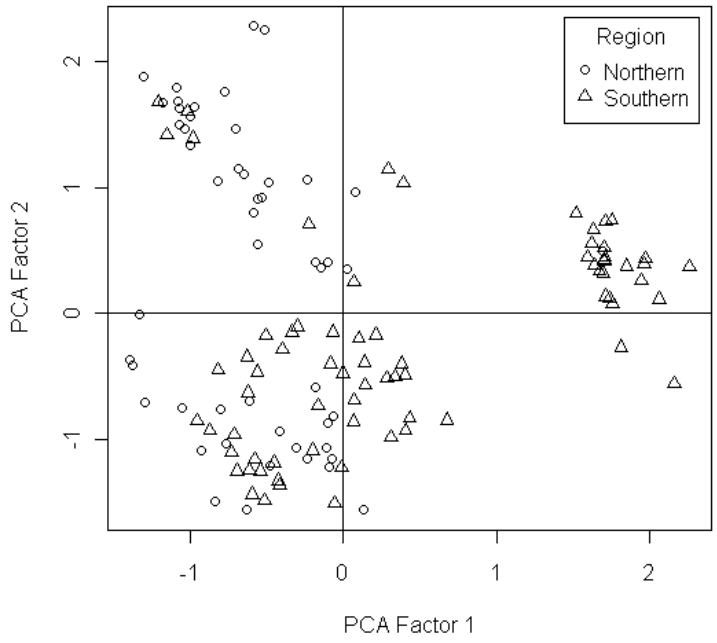

Figure 4.3: PCA scores for the first two factors, plotted according to region of viral incidence.

Smoothing the PCA scores across a map of Vietnam (Figure 4) further indicates the extension of southern viruses across the full range of factor 1 and the position of both northern and southern viruses across the range of factor 2. Areas of both high and low factor

1 PCA scores are seen in southern Vietnam, indicating potential barrier zones for gene flow. Both positive and negative PCA factor 2 scores are seen in northern and southern Vietnam, also indicating the possibility of barriers.

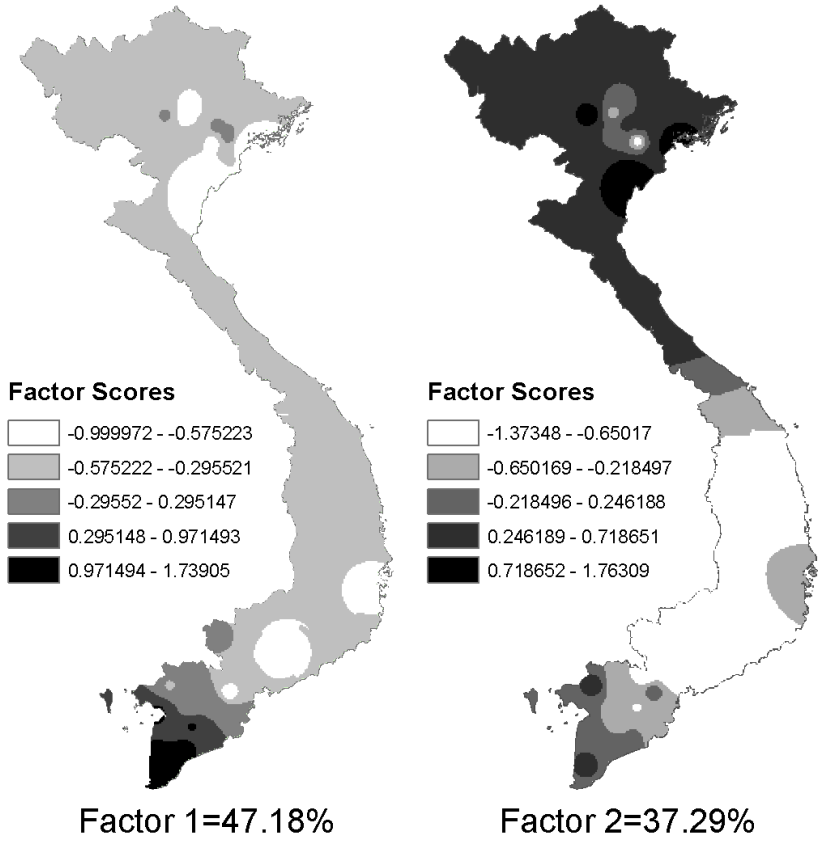

Figure 4.4: Interpolated factor scores for the first and second principal components. Average factor scores were created for provinces with more than one H5N1 influenza virus.

Stratifying the PCA scores by year of viral incidence indicated that the temporal characteristics of the viruses might also be strongly connected to the genetic characteristics of the viruses (Figure 5). Viruses from 2003 and 2004 exhibit primarily negative PCA scores for both factors, while 2005 viruses cluster in the negative factor 1 and positive factor 2 quadrant. Viruses isolated in 2007 are located solely in the cluster of highly positive factor 1 scores.

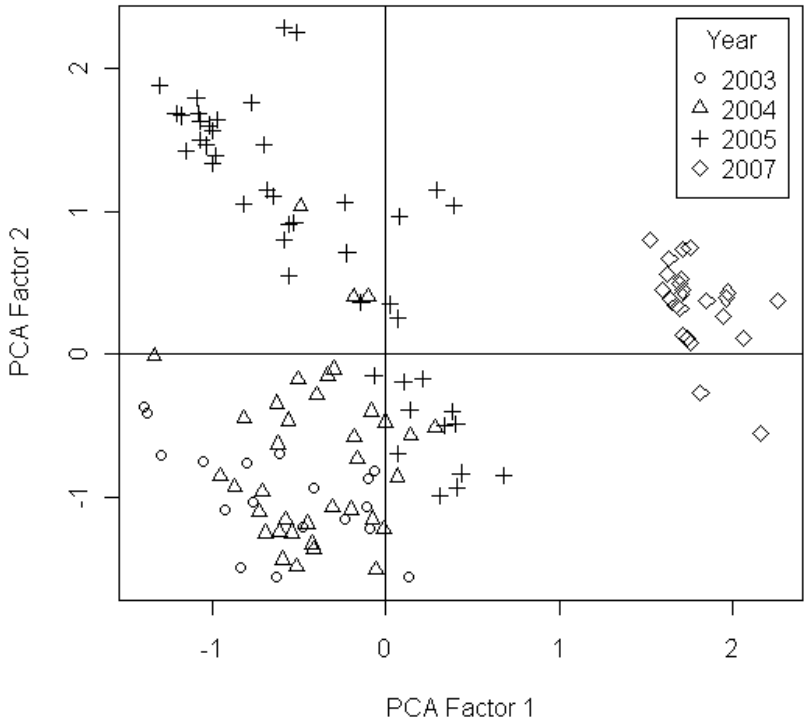

Figure 4.5: PCA factor scores according to year of viral isolation.

Results from womble analyses of the eight 28 by 28 genetic distance matrices detected multiple boundaries to gene flow on each gene segment (Table 1). Of these genetic boundaries, only those detected for the PB1 and the NS gene segments are statistically significant in the randomization trials.

	PB2	PB1	PA	HA	NP	NA	MP	NS
No. of barriers	3	2	4	3	4	3	3	2
P-value	0.447	0.094	1	0.45	1	0.428	0.412	0.082

Table 4.1: Number of barriers detected among viral distances on each gene segment and their associated significance based on a randomization test.

The location, magnitude and direction of the PB1 and NS boundary elements and boundaries are mapped in Figure 6. A single genetic boundary connects three high-magnitude (high gradient of genetic change) PB1 boundary elements in a small area of

northern Vietnam, with directions of change to the north. A singleton boundary element is located in the southwestern portion of the study area, with a genetic slope towards the west. NS boundaries are found in both northern and southern Vietnam. A northern NS boundary consists of a singleton boundary element with a large magnitude and is in a southwestern direction away from the locations of viral incidence. Three southern NS genetic boundary elements, with gradient directions towards the west, are connected into a single boundary.

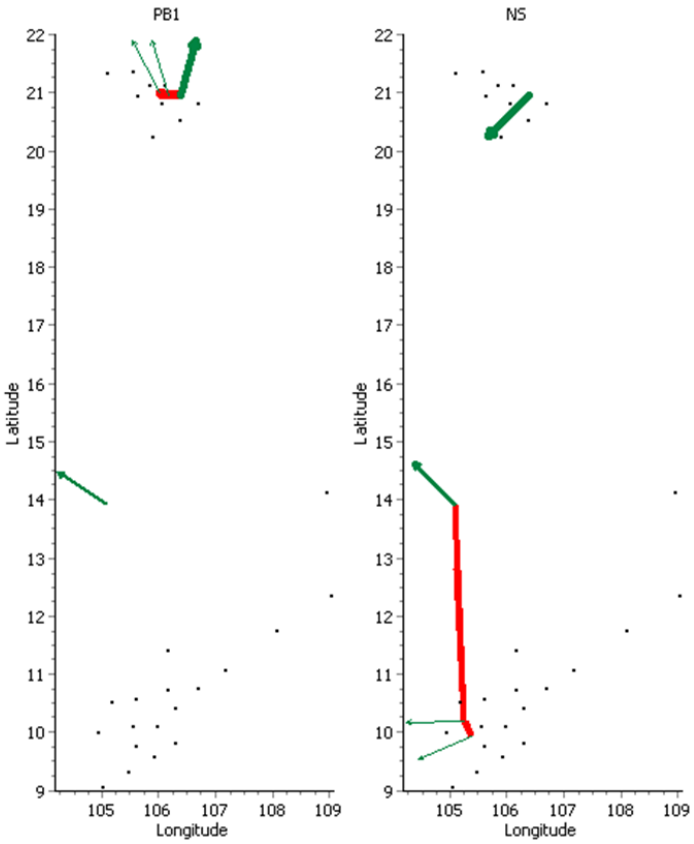

Figure 4.6: Statistically significant boundaries detected in the overall 2003-2007 dataset. Black circles indicate province centroids. Boundary elements are represented in green, thickness indicates degree of magnitude and arrows indicate the direction of the gradient. Boundary connections are represented in red.

Boundary detection for viruses stratified by year of incidence (into three temporally contiguous sets), found statistically significant barriers to gene flow only in 2003-2004 viruses on the PB1 and NP gene segments (Table 2). While boundaries were detected in

other years and on all other gene segments, randomization analysis indicated that they were not unexpected.

No. of barriers	PB2	PB1	PA	HA	NP	NA	MP	NS
2003-2004	3	2	3	3	2	3	2	2
2004-2005	4	4	4	4	4	4	4	4
2005-2007	2	2	2	2	2	2	2	2

Table 4.2: Number of barriers detected for each year category according to gene segment. Only boundaries for PB1 and NP gene segments in 2003-2004 were statistically significant at the $p<.10$ level, all others were insignificant based on randomization trials.

The location, direction and magnitude of these significant barriers can be seen in Figure 7. Nearly identical boundaries were detected for PB1 and NP, located in northern Vietnam and connecting boundary elements with high magnitude and directed towards the northeast (the PB1 boundary also connects to a boundary element directed towards the southwest). A singleton boundary element with a high magnitude was detected in southern Vietnam among 2003-2004 NP genetic distances, and is directed from southwest to northeast.

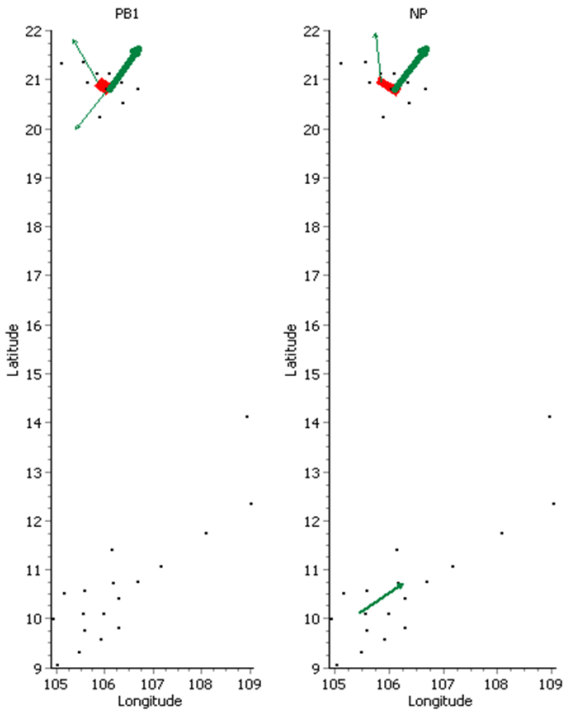

Figure 4.7: Statistically significant barriers detected in 2003-2004 viruses. Province centroids are shown in black, individual boundary elements in green and connected boundaries in red. Magnitude is indicated by thickness, while direction of genetic gradient is indicated by an arrow.

Despite the lack of statistical significance for the majority of detected boundary elements and connected boundaries, mapping the patterns of gene flow across years and for all eight gene segments highlights the fact that the majority of high-magnitude boundary elements were located in northern Vietnam (Figure 8). The direction of gene flow along these northern boundary elements is almost exclusively to the north and northwest, indicating that rates of genetic difference drop from south to north in northern Vietnam. Boundary elements in southern Vietnam were detected only in the 2003-2004 and 2004-2005 datasets, and there are no overall patterns to the direction of these magnitudes as are observed in the northern boundary elements. Additionally, no boundaries or boundary elements were found that divided the gene flow between northern and southern viruses.

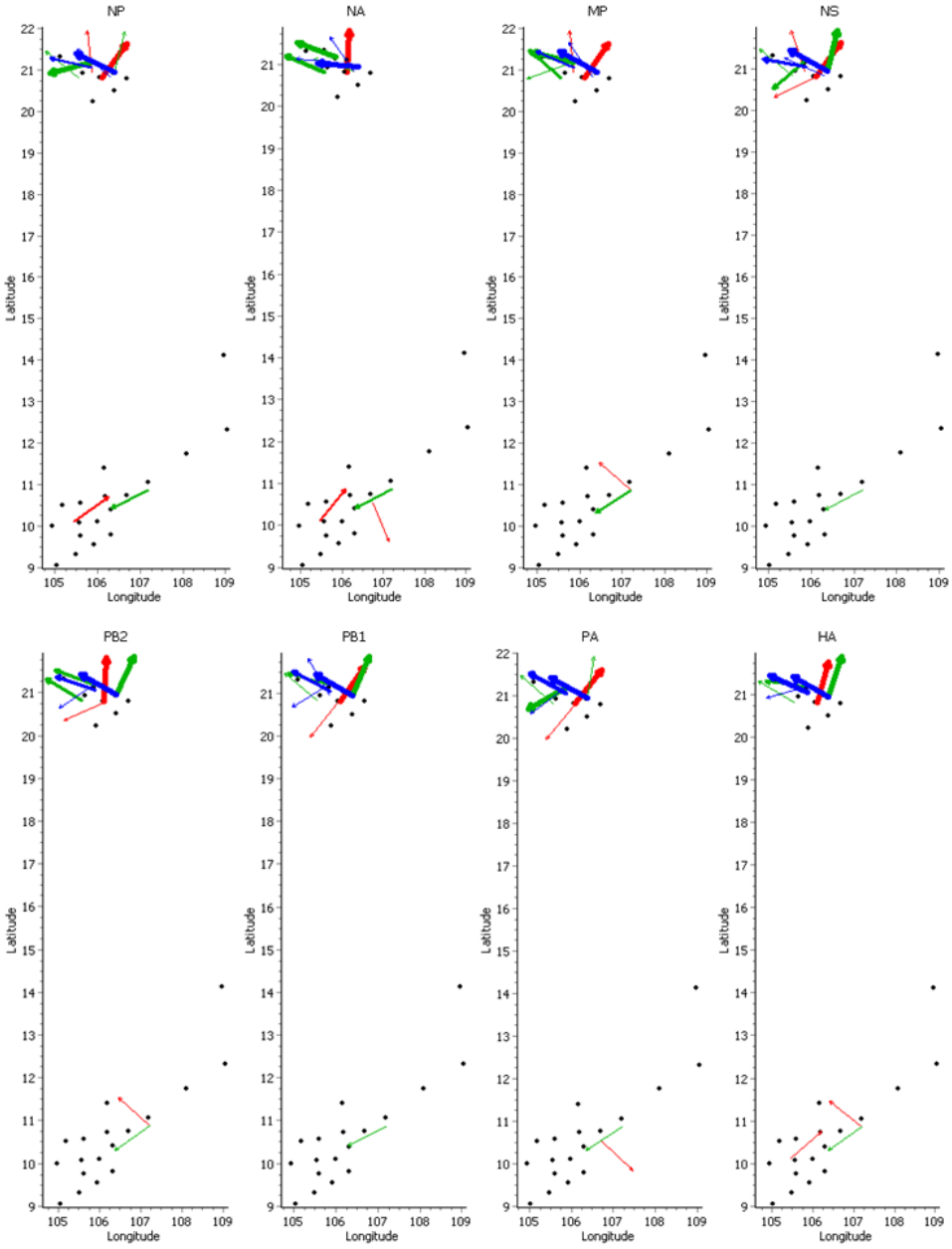

Figure 4.8: Boundary elements detected across the eight gene segments and within the three year categories. Province centroids are shown in black, 2003-2004 results are in red, 2004-2005 in green and 2005-2007 in blue. Arrows indicate the direction of change, and thickness indicates the magnitude.

Discussion

This work represents the first application of genetic boundary analysis to the study of H5N1 avian influenza viruses. Areas of genetic discontinuity, where there are barriers to gene flow, exist among H5N1 avian influenza viruses in Vietnam, but these genetic

boundaries vary temporally and by gene segment. This variation indicates that few, if any, static population or environment characteristics are responsible for genetic discontinuities. Rather, viral evolution in Vietnam is driven by the amount of time that viruses have had to mix and mutate, and viruses appear to be able to spread across the landscape with few impediments.

Results from the PCA indicate that genetic characteristics of H5N1 viruses are heavily influenced by the year in which they occur, such that viruses isolated in 2003 and 2004 have very different PCA scores than those isolated in 2007. Viruses isolated in southern Vietnam had greater diversity in PCA scores than did those isolated in northern Vietnam, and smoothed maps of the PCA scores indicated that, for factor 1, genetic boundaries were more likely to be found in southern Vietnam than northern. Mapping factor 2 scores suggested that boundaries could also be located in northern Vietnam.

Few statistically significant barriers were detected in the womble analyses, either on the overall dataset or the temporally stratified datasets. The location of the significant boundary on the NS gene segment in the overall womble analysis seems to indicate that there is a barrier to gene flow between southern Vietnam and the west, i.e. Cambodia, with genetic diversity falling from east to west. While we did not explicitly test for barriers between Vietnamese and Cambodian isolates, these results suggest further study and also confirm previous work examining gene flow across borders (11). The significant boundaries for the all-years PB1 gene segment, the 2003-2004 PB1 gene segment and the 2003-2004 NP gene segment were located in northern Vietnam, dividing isolates in the provinces surrounding Hanoi, with gene flow decreasing from south to north. This area was also the site of the majority of boundary elements detected in the non-significant womble analyses, rather than

the south, as the PCA factor 1 map would have suggested. These results suggest that northern Vietnam, more so than southern, is the site of small geographic areas with high rates of genetic change.

The gene segments for which boundaries were detected, PB1, NP and NS, are all internally located in influenza viruses. Each of these three gene segments encode for different characteristics of influenza infections. Mutations on the PB1 gene segment could be responsible for viruses in ducks being pathogenic versus non-pathogenic viruses, and both the PB1 and NP affect replication strengths of viruses (17, 18). The NP gene is also linked to high pathogenicity of H5N1 viruses in chickens, and NP mutations may influence the adaptation of viruses from ducks to chicken populations (18, 19). The NS protein is believed to boost H5N1 virulence by affecting host immune responses in two ways: blocking interferons, thus hiding infection from immune system sensors, and boosting proinflammatory cytokines, encouraging systemic rather than localized infection (20-22). Locations of genetic discontinuity only on these internal segments, rather than the surface HA and NA gene segments, indicate that perhaps the location of gene segments within viruses affect the likelihood of boundaries to gene flow occurring.

A general pattern of H5N1 viral dispersal has been suggested for Vietnam. In northern Vietnam, viruses begin to interact and mutate, before gradually spreading southward (7). Viruses in southern Vietnam then begin to evolve and adapt to local conditions and diverge from the northern seed populations. This north to south movement is likely driven by the movement of people and birds between the major population centers around Hanoi and Ho Chi Minh City. This general model is supported by our findings, that Hanoi and the surrounding province are the site of numerous boundaries to gene flow, and also high rates of

genetic distance, and that southern Vietnam is also a site for divergent PCA scores and significant boundaries and boundary elements, but that central Vietnam is not a barrier zone for gene flow. While we had few samples from central Vietnam, womble analysis would still have detected rapid rates of genetic change from northern to central or central to southern regions, or would have indicated that central Vietnam, with its relatively lower population density, acts as a regional sink for genetic change.

The lack of widespread or temporally continuous boundaries in our analyses is consistent with previous research on the genetic character of H5N1 avian influenza in Vietnam. Earlier work has suggested the presence of an isolation by distance (IBD) model of genetic dispersal among avian influenza isolates in Vietnam (23). Under IBD models genetic change is expected to change gradually across geographic space rather than rapidly as in the presence of barriers to gene flow (14). While genetic characteristics of viruses do vary across Vietnam, it is not in such a way as to suggest the driving force of population or environmental boundaries.

References

1. Muramoto Y*, et al* (2006) Pathogenicity of H5N1 influenza A viruses isolated in Vietnam between late 2003 and 2005. *J Vet Med Sci* 68: 735-737.

2. Neumann G & Kawaoka Y (2006) Host range restriction and pathogenicity in the context of influenza pandemic. *Emerg Infect Dis* 12: 881-886.

3. Smith GJ*, et al* (2006) Evolution and adaptation of H5N1 influenza virus in avian and human hosts in Indonesia and Vietnam. *Virology* 350: 258-268.

4. Nguyen DC*, et al* (2005) Isolation and characterization of avian influenza viruses, including highly pathogenic H5N1, from poultry in live bird markets in Hanoi, Vietnam, in 2001. *J Virol* 79: 4201-4212.

5. Chen H*, et al* (2006) Establishment of multiple sublineages of H5N1 influenza virus in Asia: Implications for pandemic control. *Proc Natl Acad Sci U S A* 103: 2845-2850.

6. Cristalli A & Capua I (2007) Practical problems in controlling H5N1 high pathogenicity avian influenza at village level in Vietnam and introduction of biosecurity measures. *Avian Dis* 51: 461-462.

7. Wan XF*, et al* (2008) Evolution of highly pathogenic H5N1 avian influenza viruses in Vietnam between 2001 and 2007. *PLoS ONE* 3: e3462.

8. Barbujani G & Sokal RR (1990) Zones of sharp genetic change in Europe are also linguistic boundaries. *Proc Natl Acad Sci U S A* 87: 1816.

9. Jacquez GM, Maruca S & Fortin MJ (2000) From fields to objects: A review of geographic boundary analysis. *Journal of Geographical Systems* 2: 221-241.

10. Barbujani G, Oden NL & Sokal RR (1989) Detecting regions of abrupt change in maps of biological variables. *Syst Biol* 38: 376.

11. Wallace RG & Fitch WM (2008) Influenza A H5N1 immigration is filtered out at some international borders. *PLoS ONE* 3: e1697.

12. Fourment M & Gibbs M (2006) PATRISTIC: A program for calculating patristic distances and graphically comparing the components of genetic change. *BMC Evolutionary Biology* 6: 1.

13. Brierley AS, Thorpe JP, Pierce GJ, Clarke MR & Boyle PR (1995) Genetic variation in the neritic squid *Loligo forbesi* (myopsida: *Loliginidae*) in the northeast Atlantic Ocean. *Mar Biol* 122: 79-86.

14. Piertney SB, Maccoll ADC, Bacon PJ & Dallas JF (1998) Local genetic structure in red grouse (*Lagopus lagopus scoticus*): Evidence from microsatellite DNA markers. *Mol Ecol* 7: 1645-1654.

15. Womble WH (1951) Differential systematics. *Science* 114: 315-322.

16. Rosenberg MS & Anderson CD PASSaGE: Pattern analysis, spatial statistics and geographic exegesis. Version 2. *Methods in Ecology and Evolution*

17. Hulse-Post D*, et al* (2007) Molecular changes in the polymerase genes (PA and PB1) associated with high pathogenicity of H5N1 influenza virus in mallard ducks. *J Virol* 81: 8515.

18. Wasilenko JL*, et al* (2008) NP, PB1, and PB2 viral genes contribute to altered replication of H5N1 avian influenza viruses in chickens. *J Virol* 82: 4544.

19. Tada T*, et al* (2011) NP body domain and PB2 contribute to increased virulence of H5N1 highly pathogenic avian influenza viruses in chickens. *J Virol* 85: 1834.

20. Gambotto A, Barratt-Boyes SM, de Jong MD, Neumann G & Kawaoka Y (2008) Human infection with highly pathogenic H5N1 influenza virus. *Lancet* 371: 1464-1475.

21. Neumann G & Kawaoka Y (2006) Host range restriction and pathogenicity in the context of influenza pandemic. *Emerg Infect Dis* 12: 881-886.

22. Geiss GK*, et al* (2002) Cellular transcriptional profiling in influenza A virus-infected lung epithelial cells: The role of the nonstructural NS1 protein in the evasion of the host innate defense and its potential contribution to pandemic influenza. *Proc Natl Acad Sci U S A* 99: 10736-10741.

23. Carrel MA, Emch M, Jobe RT, Moody A & Wan XF (2010) Spatiotemporal structure of molecular evolution of H5N1 highly pathogenic avian influenza viruses in Vietnam. *PLoS One* 5: e8631.

CHAPTER 5
THE ROLE OF POPULATION AND ENVIRONMENT IN GENETIC CHANGE

The second step in a landscape genetics study is to link previously described patterns to some underlying process driving those patterns. In Chapter 2, the statistically significant difference in genetic characteristics of H5N1 viruses according to both the species of isolation and the year of isolation was described, with ducks viruses and viruses isolated in later years exhibiting greater genetic change. Chapter 4 results indicate that temporal differences in genetics are not paired with strong geographic differences in genetics, as evidenced by the general absence of barriers to gene flow between northern and southern Vietnam. The lack of genetic boundaries suggests that H5N1 in Vietnam can be analyzed across the entire country without being first partitioned into genetically distinct regions.

The following chapter uses the theory of human-environment interactions developed by disease ecologists (as described in Chapter 1), and the methodologies of landscape genetics to examine the possible influences of province-level population and environment drivers on genetic characteristics of H5N1 viruses. The variables discussed in the following chapter were included because prior research suggests they influence avian influenza occurrence in Southeast Asia, or because they showed significant relationships to genetic outcomes in earlier chapters.

Landscape genetic studies may assist in the generation of hypotheses about population and environment drivers of molecular evolution of many human pathogens,

including H5N1 avian influenza. To date, no studies have examined the relationship between anthropogenic and environmental characteristics and H5N1 genetic outcomes. While the drivers of avian influenza *incidence* have been previously explored, combining disease ecology theory with landscape genetics methodology has the potential to reveal how landscapes influence avian influenza *evolution*.

CHAPTER 6

POPULATION-ENVIRONMENT DRIVERS OF H5N1 AVIAN INFLUENZA MOLECULAR CHANGE IN VIETNAM

Abstract:

A disease ecology perspective posits that health outcomes are the result of complex and dynamic interactions between people and their environments. I extend this framework to the occurrence of a human pathogen, H5N1 avian influenza, as it occurs in non-human hosts, and seek to understand how province-level population and environment characteristics influence the evolution of avian influenza viruses in Vietnam. While prior work has examined what combinations of local-level environmental variables influence H5N1 occurrence, this research expands the analysis to the actual genetic characteristics of H5N1 viruses. Using a dataset of 125 highly pathogenic H5N1 avian influenza viruses isolated in Vietnam from 2003-2007, I explore which population and environment variables are correlated with increased genetic change among H5N1 avian influenza 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 drive genetic change 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 low education levels, an indicator of low socio-economic status of the population, as well as geographic and

temporal characteristics of viral occurrence, that drive genetic change among avian influenza viruses.

Background

Highly pathogenic H5N1 avian influenza has persisted at pandemic levels in poultry and human populations in Vietnam and other Asian countries since 2003. The continuous incidence and evolution of H5N1 influenza viruses is driven by complex and dynamic interactions between birds and people and the social and natural environments in which they circulate. While there exists a great deal of research 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.

The emergent field of landscape genetics focuses on exploring interactions between evolutionary outcomes and environmental features in the belief that spatial variation in genetics indicates underlying landscape processes (1-4). 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 spread of human pathogens (5, 6). 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 within 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 (7-9). 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 allows me to generate a dataset of hypothesized drivers of molecular change that I then analyze using landscape genetics methodology.

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. Viruses were either collected by the National Centre for Veterinary Diagnostics (NCVD) of Hanoi, Vietnam or publicly available in online repositories. Each of the isolates used in the analysis had a full genetic sequence available, as well as information regarding the province and year in which it was observe. 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. While the collection sites of the publicly available viruses are unknown, 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 (10).

Phylogenetic analysis of the H5N1 viruses in the dataset indicates that they share a single genetic lineage, descendant from a progenitor virus found in Hong Kong in 2002 (A/Duck/HongKong/821/2002). This lineage, known as HK821-like, is believed to result

from a single introduction of the virus into Vietnam, though exactly how the introduction took place remains unknown (11, 12). The most likely source of the introduction was overland trade in poultry or poultry products at Vietnam's northern border with China (13).

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 connected them in a phylogenetic tree (14). 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), and three or four polymerases (PA, PB1, PB1-F2, and PB2) (15-17). Each of these gene segments can mutate independently of the others, so we calculated eight total genetic distance measures for each of the 125 viruses.

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 1). Then, also in 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).

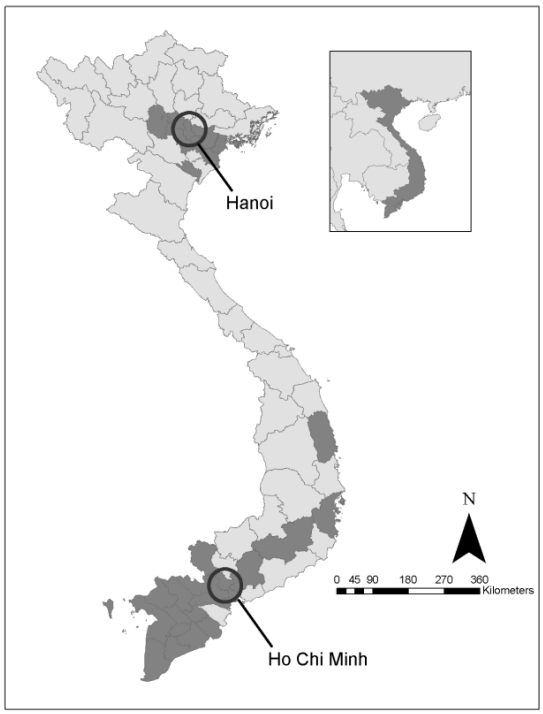

Figure 6.1: Location of viral isolation among the H5N1 avian influenza dataset. Darkened provinces indicate sites where viruses were collected.

In addition, several population and environment variables believed to be potential drivers of genetic change under a disease ecology framework (Chapter 1: Figure 2) were calculated for each province with an H5N1 viral occurrence (see Table 1). 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, 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 Vuon (agricultural plots), Ao (ponds), and Chuong (caged birds), wherein 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 (18). 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, can 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.

Variable	Measure (per province)	Dates	Source
Population			
Population density	median persons/square kilometer	2005	‡‡
High school graduates	percent graduates	2003-2007	**
Rural population	thousand persons	2003-2007	**
Urban population	thousand persons	2003-2007	**
Medical professionals	total persons	2003-2007	**
Passenger traffic	million persons/road kilometer	2003-2007	**
Income indicators	average monthly income, income inequality	1999	**
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	§§
Waterway freight traffic	million tons/kilometer	2003-2007	**
Roadway freight traffic	million tons/kilometer	2003-2007	**
Elevation	median kilometers above sea level	2000	††

**General Statistics Office of Vietnam, ‡‡ CIESIN, §§ GLCF (UMD) †† SRTM30 (NASA)

Table 6.1: Population and environment variables included in the analysis.

These measures were gathered from the General Statistics Office of Vietnam and from 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) (19-22). Due to the coarse nature of the avian influenza dataset, and the geocoding of viruses to the provincial centroid, the population and environment variables were all scaled to the level of 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.

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). 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 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:

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

where d_{hi} is the ordinated distance between two samples and $\hat{d}_{hi}$ is the distance predicted from the regression (23, 24). 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 (25).

Once the H5N1 viruses were ordinated according to their genetic characteristics, each of the eighteen population and environment variables was associated with the ordination. Each variable is aligned in the 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 .15 were taken to be the most important drivers of genetic change among H5N1 viruses.

Clustering techniques were used to assign the ordinated viruses into like groups, and then variation in the variables with scores greater than .15 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. To explore the direction and statistical significance of the relationship between the NMDS variables and genetic outcomes, I used linear regression. 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>.15 as well as a binary variable that indicated whether the virus was isolated in a duck (1) or another species (0). The inclusion of this variable was based on previous work (Chapter 2) that suggests duck isolates are associated with greater genetic change than are chicken isolates. Variables were then discarded if their relationship to the axes scores outcome variables were non-significant or if they exhibited high multicollinearity with other variables (indicated by Variance Inflation Factor (VIF) of 6 or greater) and if their removal from the model improved model fit, as measured by the Akaike Information Criterion (AIC) and the Log-Likelihood Ratio (LLR).

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 (26). Regression was conducted in SAS®9.1.3(27).

Results

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) (25). At three dimensions stress is minimized (3.8%) without adding more unnecessary dimensions (Figure 2).

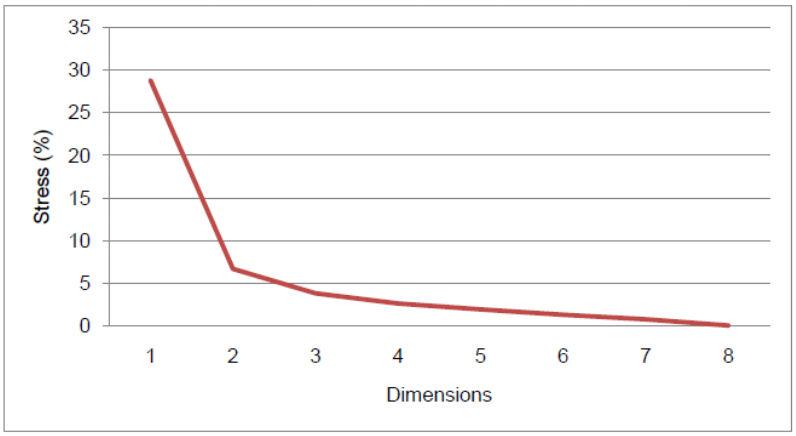

Figure 6.2: Measures of stress according versus dimensionality. Above three dimensions, reductions in stress are minimal.

A regression plot of the observed *n* by *n* matrix versus the ordinated differences (Figure 3) indicates that the final three-dimensional NMDS ordination well-represents the measured genetic distances in the viral dataset.

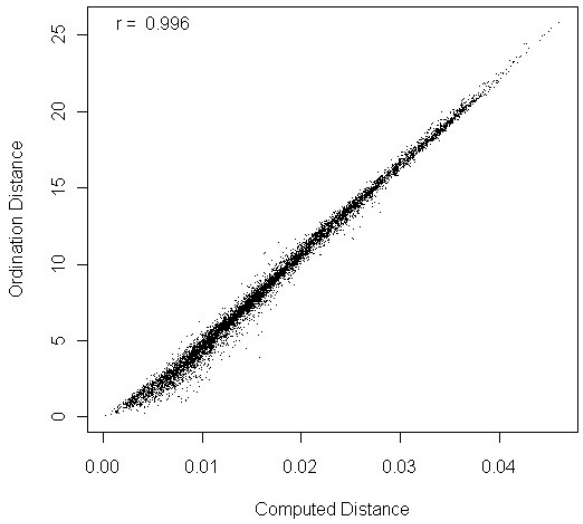

Figure 6.3: Distances calculated from the observed data (x-axis) versus the ordinated distance (y-axis).

The first two dimensions in the final ordination appear in Figure 4A, 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 4B. Many of the population-environment variables have the same alignment in the ordinated space (clustered to the right), 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.

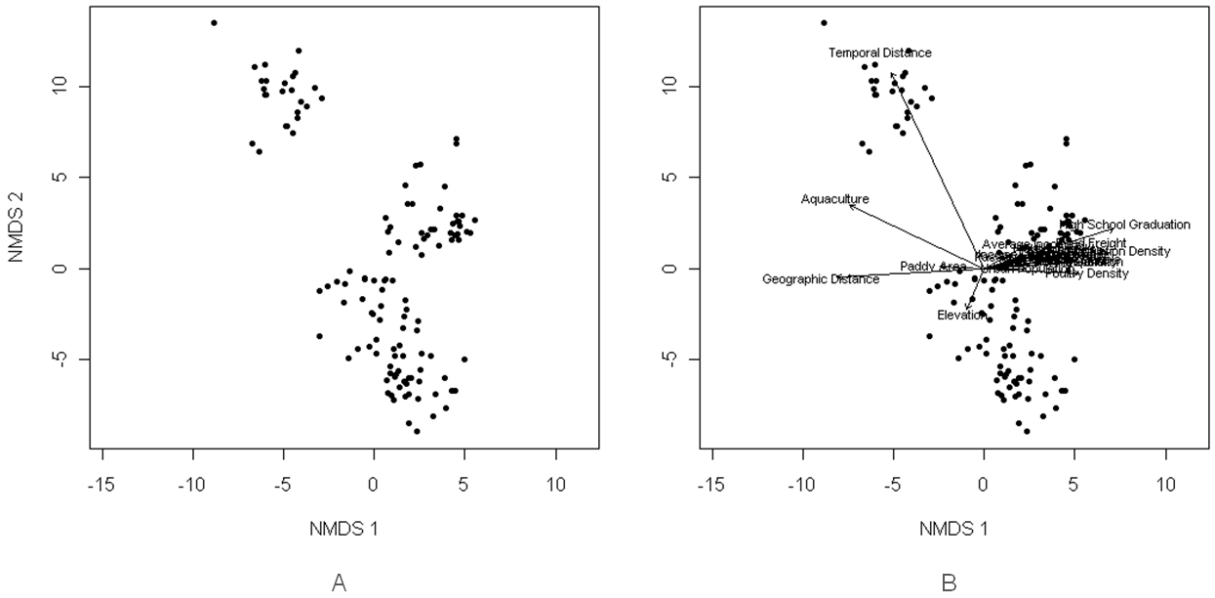

Figure 6.4: A- Plot of the 125 viral points within the first two ordinated dimensions. **B-** All of the hypothesized population-environment drivers of genetic change arrayed over the scaled genetic measures, showing each variable's axis of differentiation through the 3-dimensional space. Longer axes of differentiation indicate greater association with the genetic distances arrayed in the 3-dimensional space.

Plotting the R^2 calculated for each of the population-environment variables indicated that only five had scores of greater than .2 (Figure 5). 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 .15: pig population density, poultry population density and road freight.

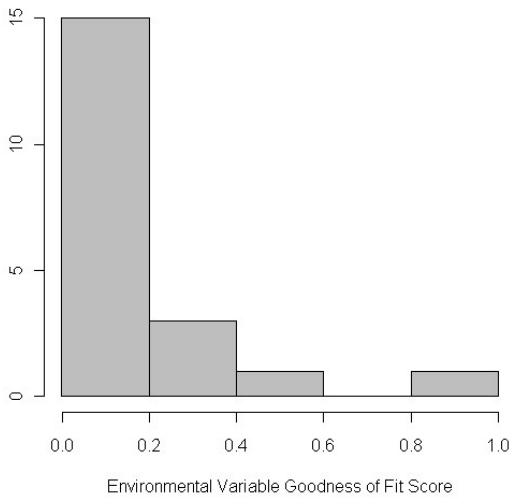

Figure 6.5: Histogram of goodness of fit scores for population-environment variables. Only eight variables exhibit scores greater than .15, indicating a high level of correlation with the scaled genetic data.

Plotting only these eight variables onto the ordination indicates their differing strengths and relationships to the scaled genetic distances (Figure 6). 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 five variables, high school graduation rate (High School Graduation), population density (Population Density), road freight (Road Freight), and pig (Pig Density) and poultry (Poultry Density) populations, 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.

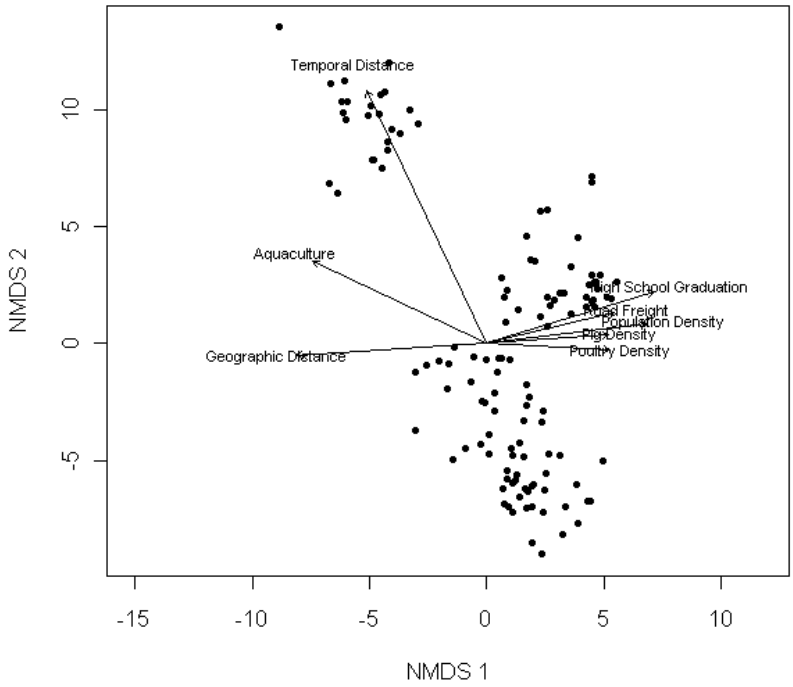

Figure 6.6: The axes of differentiation for eight independent variables with goodness of fit scores of .15 or above: aquaculture, population density, high school graduation, poultry, pigs, road freight, geographic distance and temporal distance.

Each of the 125 ordinated points was then assigned to a cluster to examine how the influence of these eight independent variables with R^2 greater than .15 differed within the dataset. 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 (Figure 7).

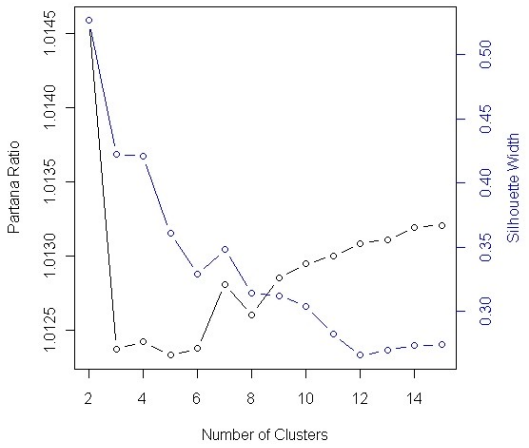

Figure 6.7: The optimum number of clusters maximizes both Partana ratio and silhouette width. Seven clusters best describe the scaled genetic data.

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. I chose to classify the NMDS ordination into seven clusters, the number at which both the Partana ratio and the silhouette width increased.

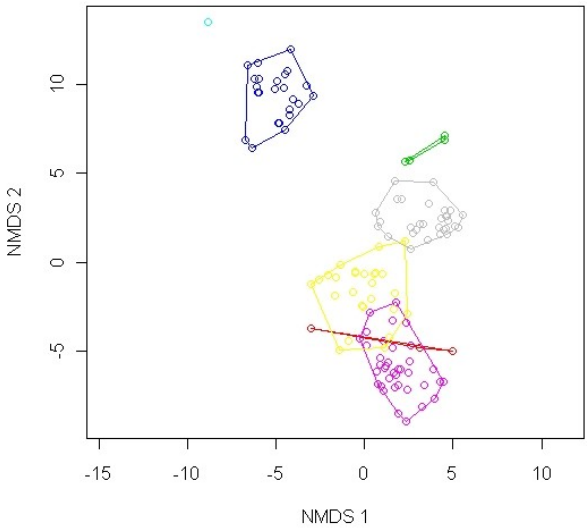

Figure 6.8: NMDS results (the first two dimensions, out of three) charted according to cluster assignment. There is one virus that is in its own cluster (cluster 4), seen in aqua in the upper left of the graphic.

Examining the cluster patterns in the first and second dimensions of the ordinated space shows that some clusters appear to better fit the first and second dimensions, while others are based more on the position of ordinated points in the third dimension. Cluster 1, seen in red, has a large span across dimension 1, seeming to cross over points assigned to the yellow and fuchsia clusters, but has points ordinated close to one another in the second and third dimensions. One virus has also been assigned to its own cluster, Cluster 4, based on its distance in ordination space across all three dimensions from all the other viruses. 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 eight independent variables differ among clusters (Figure 9).

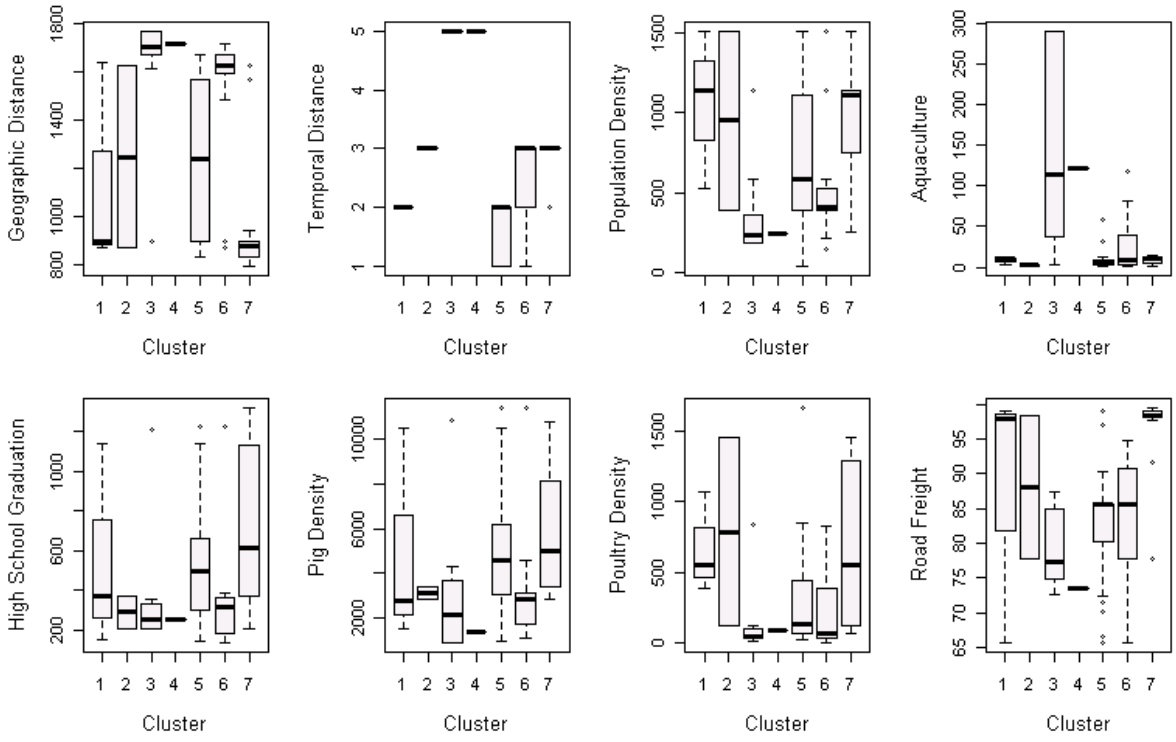

Figure 6.9: Distributions of the eight independent variables within each of the 7 clusters.

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 eight 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 four 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 4 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 7, 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) while others have similar values across clusters (aquaculture).

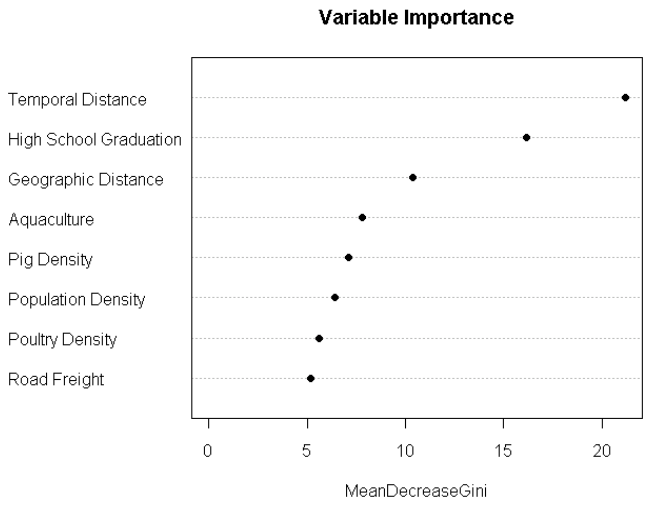

Figure 6.10: Varying importance of the eight variables in relation to the genetic differentiation of viruses. This is an indication of the strength but not the direction of relationships.

Figure 10 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, pig and poultry population characteristics and 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.

All eight variables with R^2 of >.15, regardless of their importance indicated in Figure 10, were included in the initial regression, as was the dichotomous duck variable. Based upon VIF scores of more than 6, the pig and poultry population variables were iteratively removed, neither had significant interactions with the dependent variables. Road freight was also removed from the final model, it had non-significant interactions with all three outcome variables and the AIC decreased and the LLR increased when it was eliminated. This removal was also supported by Road Freight's smallest decrease in the Gini measure when clustering the ordinated points. The final models included six predictor variables (Table 2).

Variable	Dimension 1	Dimension 2	Dimension 3
Intercept	9.41222	-1.00992	-4.08156
Aquaculture	0.00615	0.00887	0.00553
High School Graduation	-0.07981	-0.10855	-0.02383
Population Density	-0.00072	0.00082	0.00122
Geographic Distance	0.002	-0.00151	-0.00038
Temporal Distance	-1.61269	4.11726	2.09455
Duck Virus	0.19251	0.70104	0.52509

Table 6.2: Regression results showing the influence of five population-environment independent variables on viral NMDS loading scores. Shading represents statistical significance. Dark shading equals p-values of <.01, light shading equals p-values of <.1, no shading is for p-values >.1.

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 increases, so does the genetic differentiation among viruses. As high school graduation in a province decreases, genetic differentiation increases. Population density, on the third 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 dimension, increased geographic distance is significantly associated with increased genetic differentiation, and on the second dimension duck viruses have significantly higher genetic difference.

Discussion

Differentiation among the eight gene segments of H5N1 avian influenza viruses is most associated with a combination of five population-environment variables and spatio-temporal characteristics: the amount of aquaculture in a province, the high school graduation percentage in a province, the population density of a province and the amount of geographic and 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 (28-30).

As the amount of land devoted to aquaculture increases, genetic differentiation also increases. Aquaculture spaces can provide a site for interaction among infected and uninfected poultry, where the water surface provides a medium for fecal-oral transmission of the virus (31). Aquaculture practices and areas have been previously found to be associated with H5N1 incidence and perpetuation of the virus in the environment (18, 32). As high school graduation decreases, 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 (33, 34). Thus, in areas where education levels are low, 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. Although the income indicators included in the analysis were not strongly associated with genetic differentiation, and were dropped in the final model, the importance of high school graduation is likely capturing some influence of socioeconomic status. 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 (28-30, 32). Viruses isolated in ducks were also associated with greater genetic differentiation than those found in chickens or other species. Ducks, sometimes called the "Trojan horses" of H5N1 avian influenza because they can experience asymptomatic infections, have been positively associated with the spatial and temporal patterns of H5N1 incidence (29, 35-36).

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 two of the three NMDS dimension, but decreases on the third. This is likely 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 relationship between viral change and geographic space is not as robust as that with temporal distance, although geographic distance was a significant predictor of genetic differentiation in one of the three NMDS dimensions. The direction of the geographic distance and genetic differentiation relationships was also variant across the three NMDS dimensions, suggesting that viruses close in geographic space can be either quite similar or quite different genetically. A constant seeding of viruses from northern Vietnam to southern Vietnam and an isolation by distance model is most likely responsible for this mixed relationship between geographic distance and genetic difference (10, 37).

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, that the area in a province devoted to wet rice agriculture (and paddy yield and elevation) or the area of a province covered in less specific water sources, was not linked to high levels of genetic change. Finally, no indicators of socioeconomic status or hygiene knowledge or access to care, other than high school graduation rates, showed a significant relationship.

These 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 low 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 influences of time and space 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.

References

1. Manel S, Schwartz MK, Luikart G & Taberlet P (2003) Landscape genetics: Combining landscape ecology and population genetics. *Trends in Ecology & Evolution* 18: 189-197.

2. Storfer A*, et al* (2007) Putting the "landscape" in landscape genetics. *Heredity* 98: 128-142.

3. Balkenhol N*, et al* (2009) Identifying future research needs in landscape genetics: Where to from here?. *Landscape Ecol* 24: 455-463.

4. Guillot G, Estoup A, Mortier F & Cosson JF (2005) A spatial statistical model for landscape genetics. *Genetics* 170: 1261-1280.

5. Archie EA, Luikart G & Ezenwa VO (2009) Infecting epidemiology with genetics: A new frontier in disease ecology. *Trends in Ecology & Evolution* 24: 21-30.

6. Criscione CD*, et al* (2010) Landscape genetics reveals focal transmission of a human macroparasite. *PLoS Neglected Tropical Diseases* 4: 308-332.

7. Mayer JD & Meade MS (1994) A reformed medical geography reconsidered. *The Professional Geographer* 46: 103-106.

8. Mayer JD (2000) Geography, ecology and emerging infectious diseases. *Soc Sci Med* 50: 937-952.

9. Meade MS (1977) Medical geography as human ecology: The dimension of population movement. *The Geographical Review* 67: 379-393.

10. Wan XF*, et al* (2008) Evolution of highly pathogenic H5N1 avian influenza viruses in Vietnam between 2001 and 2007. *PLoS ONE* 3: e3462.

11. Smith GJ*, et al* (2006) Evolution and adaptation of H5N1 influenza virus in avian and human hosts in Indonesia and Vietnam. *Virology* 350: 258-268.

12. Dung Nguyen T*, et al* (2008) Multiple sublineages of influenza A virus (H5N1), Vietnam, 2005-2007. *Emerg Infect Dis* 14: 632-636.

13. Wang J*, et al* (2008) Identification of the progenitors of Indonesian and Vietnamese avian influenza A (H5N1) viruses from southern china. *J Virol* 82: 3405-3414.

14. Fourment M & Gibbs M (2006) PATRISTIC: A program for calculating patristic distances and graphically comparing the components of genetic change. *BMC Evolutionary Biology* 6: 1.

15. Gambotto A, Barratt-Boyes SM, de Jong MD, Neumann G & Kawaoka Y (2008) Human infection with highly pathogenic H5N1 influenza virus. *Lancet* 371: 1464-1475.

16. Peiris JS, de Jong MD & Guan Y (2007) Avian influenza virus (H5N1): A threat to human health. *Clin Microbiol Rev* 20: 243-267.

17. Chen W*, et al* (2001) A novel influenza A virus mitochondrial protein that induces cell death. *Nat Med* 7: 1306-1312.

18. Cristalli A & Capua I (2007) Practical problems in controlling H5N1 high pathogenicity avian influenza at village level in Vietnam and introduction of biosecurity measures. *Avian Dis* 51: 461-462.

19. Center for International Earth Science Information Network (CIESIN) (2010) Gridded population of the world, version 3 (GPWv3). Columbia University.

20. General Statistics Office of Vietnam (2010) Statistical data. Government of Vietnam.

21. Hansen M, DeFries R, Townshend JRG & Sohlberg R (1998) UMD global land cover classification, 1 kilometer, 1.0, Department of Geography, University of Maryland, College Park, Maryland, 1981-1994.

22. Shuttle Radar Topography Mission (SRTM) (2009) Global 30 arc-second elevation dataset (GTOPO30). National Aeronautic& Space Adminstration (NASA).

23. Kruskal JB (1964) Nonmetric multidimensional scaling: A numerical method. *Psychometrika* 29: 115-129.

24. Steyvers M (2006) Multidimensional scaling. *Encyclopedia of Cognitive Science.*

25. Kruskal JB (1964) Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* 29: 1-27.

26. R Development Core Team (2011) R: A language and environment for statistical computing.

27. SAS Institute Inc. (2008) SAS®9.1.3 for Windows. Cary, North Carolina.

28. Gilbert M*, et al* (2008) Mapping H5N1 highly pathogenic avian influenza risk in Southeast Asia. *Proc Natl Acad Sci U S A* 105: 4769-4774.

29. Paul M*, et al* (2010) Anthropogenic factors and the risk of highly pathogenic avian influenza H5N1: Prospects from a spatial-based model. *Vet Res* 41:

30. Martin V*, et al* (2011) Spatial distribution and risk factors of highly pathogenic avian influenza (HPAI) H5N1 in China. *PLoS Pathog* 7: e1001308.

31. Brown JD, Swayne DE, Cooper RJ, Burns RE & Stallknecht DE (2007) Persistence of H5 and H7 avian influenza viruses in water. *Avian Dis* 51: 285-289.

32. Pfeiffer DU, Minh PQ, Martin V, Epprecht M & Otte MJ (2007) An analysis of the spatial and temporal patterns of highly pathogenic avian influenza occurrence in Vietnam using national surveillance data. *The Veterinary Journal* 174: 302-309.

33. Dinh PN, *et al* (2006) Risk factors for human infection with avian influenza A H5N1, Vietnam, 2004. *Emerg Infect Dis* 12: 1841-1847.

34. Thorson A, Petzold M, Nguyen TK & Ekdahl K (2006) 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. *Arch Intern Med* 166: 119-123.

35. Gilbert M, *et al* (2006) Free-grazing ducks and highly pathogenic avian influenza, Thailand. *Emerg Infect Dis* 12: 227-234.

36. Kim JK, Negovetich NJ, Forrest HL & Webster RG (2009) Ducks: The "Trojan horses" of H5N1 influenza

37. Carrel MA, Emch M, Jobe RT, Moody A & Wan XF (2010) Spatiotemporal structure of molecular evolution of H5N1 highly pathogenic avian influenza viruses in Vietnam. *PLoS One* 5: e8631.

CHAPTER 7
CONCLUSION: THE LANDSCAPE GENETICS ECOLOGY OF H5N1 AVIAN INFLUENZA IN VIETNAM

The majority of public health and medical geography studies of infectious disease treat the diseases themselves as static outcomes. The reality, however, is that pathogens are constantly evolving to better outwit immune responses. This is particularly true for rapidly mutable RNA viruses such as influenza A viruses and HIV. Additionally, there is increasing concern over the development of drug resistant pathogens, such as amantadine-resistance and oseltamivir-resistance influenza A viruses, and extensively drug resistant tuberculosis (XTB) and Methicillin-resistant *Staphylococcus aureus* (MRSA). Going forward, combining medical geography and landscape ecology via the complementary frameworks of disease ecology and landscape genetics can enhance the contributions that medical geographers make to the study of infectious diseases.

A disease ecology perspective in medical geography views infection or ill-health as the result of some mal-adaptation between humans and their environment. Human health is a spectrum, ranging from poor fitness and response to insult (poorly adapted) to high fitness and impervious to the insults of the environment. Landscape ecologists understand organisms as existing in dynamic and fluid states of equilibrium or disequilibrium with the environment, and believe that spatial patterns of organisms in the landscape can reveal underlying processes of behavior, fitness and function. Linking medical geography and landscape ecology thus provides a framework for understanding that the spatial patterns of

disease within a landscape can reveal the processes, i.e. interactions between people and their environment, that allow the disease to persist. Adding a genetic component to this pairing of medical geography and landscape ecology, *landscape genetics informed by disease ecology*, adds a further layer of complexity, enabling an examination of both the *etiology and ecology of pathogenic evolution*. Not only does the disease exist in a place (or host), it has a particular genetic sequence that can indicate pathogenic response to environmental pressures.

For diseases of known etiologies, disease ecology provides insights into how variation in population-environment factors that are hypothesized to increase or decrease transmission is linked to patterns of disease incidence or absence. For diseases of unknown etiology, disease ecology studies, such as cluster analysis or multi-dimensional scaling, can provide insight into what variables in an individuals' social or natural environment cause or correlate with disease. Landscape genetics uses similar theory and methods to deduce what population-environment variables influence the genetic similarity or dissimilarity of populations. For organisms with known ecologies, landscape genetics can speak to the size of habitats needed to sustain life, the effect of changing environments on population genetics and patterns of migration. For organisms with unknown or unclear ecologies, such as influenza viruses, landscape genetics provides a way to test hypotheses about how genetic evolution varies across landscapes, and what variables in the environment are conducive to viral survival and reproduction.

Disease ecology acknowledges that the holistic population-environment system that results in ill-health is part of a larger system. There are few, if any, closed systems in nature, and recognition that forces outside of those under consideration may have an impact on outcomes is necessary. The same holds true in landscape genetics: populations and their

genetic variability are delimited in space, but with the understanding that most boundaries imposed in the analysis are either permeable or nonexistent in the natural environment. Both fields, then, recognize that false-bounding of some sort occurs in analysis, and that these false-boundings must be acknowledged as artifacts of investigation, not facts.

In addition to a shared understanding of open and holistic systems driving processes and patterns, both disease ecology and landscape genetics are attentive to the importance of space and time in affecting relationships among landscape features. In both fields, the effects of spatial autocorrelation in producing outcomes is recognized, and this spatial autocorrelation is either controlled for in analysis or utilized as an indication of underlying patterns. The temporal aspect of landscape genetic and disease ecology studies is recognized in both fields to have an important modifying effect. Not only do disease systems and population ecosystems exist in open systems, they are also dynamic and changing. Snapshots in time of these systems can give insight into processes that operate over a variety of timescales, or processes that may be radically changed by disturbance events.

The research presented here indicates the utility of a landscape genetics/disease ecology approach. Chapters 2 and 4 comprise the first part of landscape genetics research, exploring patterns of genetic variation across a landscape. The second part of landscape genetics research, linking pattern with process, is where disease ecology helps to inform hypotheses and test theories about both the etiology and ecology of pathogenic evolution. Prior research into the differing epidemiologies of chickens and ducks infected with H5N1 avian influenza viruses was supported by findings that indicate observable differences in the genetic attributes of strains isolated in those two species. Even though there is variation in landscape features across the extent of Vietnam, this does not appear to act as a barrier to

genetic exchange among viruses, given the overall lack of boundaries to gene flow found in Chapter 4. These findings reveal that species, time and, to a lesser extent, space were important variables to consider when exploring genetic variation. They were combined with the elements contained in the hypothesized triangle of human ecology shown in Chapter 1 to understand the processes at work in Vietnam that drive H5N1 evolution. As hypothesized, characteristics of both the social and natural environments influence molecular change, even when controlling for the influences of space, time and species on the ecology of H5N1 genetics.

The emergence, re-emergence and persistence of infectious diseases has been linked by geographers to changing relationships between humans and environments, via processes of urbanization, climate change, trade and mobility (1). Not only are patterns of disease influenced by these processes, so too are the characteristics of pathogens themselves. As organisms responding to evolutionary pressures, pathogens undergo natural selection, responding to changes in host behavior, environmental conditions and other forces (2). The relationships described in Chapter 6 indicate that H5N1 influenza viruses have increased amounts of genetic change in places with certain population and environment characteristics. Increased surface devoted to aquaculture represents increased spaces and places for viral exchange amount domestic poultry populations. Increased viral exchange provides influenza viruses with greater opportunities for mutation. High human population densities, when combined with low educational levels, expose H5N1 viruses to environments where chances for viral movement, whether in infected birds or as fomites on contaminated surfaces such as shoes and cages, are increased and knowledge about influenza transmission or availability of resources to inhibit transmission are decreased. While variables such as poultry density and

urban population density were not found to be strongly associated with viral change, it is the holistic system of population movement and trade in birds and availability of avian hosts, along with other numerous factors, that drives H5N1 incidence and evolution in Vietnam.

While influenza viruses have two methods of genetic change, shift and drift, the molecular evolution analyzed in this research is that of drift. The dataset of Vietnamese H5N1 viruses consists of a stable lineage, without large amounts of genetic material exchanged with other ancestral lines of viral evolution. The population and environment characteristics found to be associated with this drift may differ from those that would be associated with antigenic shifts. Future research should and will examine what landscape patterns of shift look like, as well as what environments are associated with specific mutations that impact viral characteristics, such as host range restrictions. This drift research, however, demonstrate that H5N1 viruses can continue to be not only incident in Vietnam but gradually changing in places and spaces of bird and human interactions. Public health implications resulting from these are that limiting the movement of duck populations into areas of aquaculture, and their subsequent sharing of viruses, as well as increasing access to resources and knowledge about hygiene and viral transmission can slow or even stop viral incidence and evolution.

The accuracy, and utility to public health, of results from landscape genetics/disease ecology analysis is highly dependent upon the spatial accuracy of both the measured outcome and the hypothesized drivers. In the case of this H5N1 avian influenza research, the highly detailed, micro-scale genetic data was not matched by micro-scale measures of where the virus occurred, which limited the ability to associate the genetic data with fine-scale population and environment characteristics. It is also possible, and indeed highly probable,

that the forces operating to produce each molecular combination observed are operating at different spatial and temporal scales. The presence of aquaculture water surfaces where birds mix and share infections, for instance, likely influences genetic change at a smaller spatial scale than does low educational status and high population density. Some type of spatial aggregation must always be present in a landscape genetic/disease ecology study, however, since a viral strain isolated at a specific geographic location is reflective of multiple interacting processes which may or may not be operating at that precise point. While current contributions of landscape genetics studies informed by disease ecology may be limited by the geographic scale at which genetic data is collected and disseminated, scalar mismatches do not completely negate the utility of conducting such research at whatever spatial scale is possible.

The dataset used for this research consisted of a relatively small number of H5N1 viral isolates spread across a wide geographic and temporal range. Viral isolates were located almost exclusively in the provinces around Hanoi and Ho Chi Minh City. The lack of isolates from central Vietnam is either due to a lack of reporting or to limited H5N1 incidence in that region, or a combination of the two. Whatever the reason, the geographic distribution of the data makes it difficult to speculate about whether the research findings are applicable to central Vietnam, or to other provinces in northern and southern Vietnam that did not report cases of H5N1. In the case of a passively collected dataset, rather than active and complete surveillance at a small spatial scale across the entire country, absence of H5N1 reporting may not, and likely does not, actually reflect the absence of H5N1 viruses from those areas. Ideally, combined landscape ecology/landscape genetics H5N1 research would be conducted with true presence/absence data and with genetic information geocoded to

exact coordinates where pathogens were detected. For studies of human diseases other than H5N1 in poultry hosts, however, paired genetic data and spatial data will generate complex issues surrounding confidentiality.

In addition to providing new understanding of the relationship between H5N1 molecular evolution and landscape patterns, the theory and methods used in this research have broad application to the study of infectious diseases. Molecular changes in infectious disease agents, such as malarial parasites or HIV, can have serious implications for vaccine development and preventative behaviors. Because many infectious diseases are mediated by social or environmental variables, examining how the variation in these population and environmental drivers correlate with molecular variation can provide new insight into how multiple types of pathogens evolve in space and time. Combining disease ecology and landscape genetics can thus facilitate exploration of both the etiology and ecologies of human pathogenic evolution.

References

1. Mayer JD (2000) Geography, ecology and emerging infectious diseases. *Soc Sci Med* 50: 937-952.

2. Ewald PW (2000) *Plague time : how stealth infections cause cancers, heart disease, and other deadly ailments,* (Free Press, New York), pp 282.