Geographic and host factors shape the evolution of a newly recognized subgroup within the U genogroup of the fish rhabdovirus IHNV

Allison Black

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2015

Committee:

Gael Kurath

Karen Edwards

Program Authorized to Offer Degree:

School of Public Health - Epidemiology

University of Washington

Abstract

Geographic and host factors shape the evolution of a newly recognized subgroup within the U genogroup of the fish rhabdovirus IHNV

Allison Black

Chair of the Supervisory Committee:
Gael Kurath, PhD
Department of Global Health
School of Aquatic and Fishery Sciences

Infectious hematopoietic necrosis virus (IHNV) is an aquatic rhabdovirus of Pacific salmonids that causes frequent epidemics. We analyzed data on U genogroup IHNV detections from 1971 to 2013 in Washington, Oregon, and Idaho. Using Bayesian coalescent analysis we discovered two previously unrecognized subgroups: UC and UP. Descriptive epidemiological analysis showed that both subgroups displayed host and geographic specificity. UC viruses were detected predominantly in Chinook salmon and steelhead trout in the Columbia River Basin and UP viruses were detected predominantly in sockeye salmon in Washington coastal watersheds. Statistical analysis indicated that viral populations were structured by host species of detection and by geographic range. Our findings indicate that the UC subgroup likely resulted from an adaptation of U genogroup IHNV to Chinook salmon in the Columbia River Basin due to a human-caused shift in host abundance. This work demonstrates anthropogenic influence as a selection driver of viral genetic diversity.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I offer my gratitude to members of the Kurath laboratory and academic community at the Western Fisheries Research Center who reminded me how rewarding working in a non-human system can be. I owe particular thanks to Dr. G. Kurath for her enthusiasm and thoroughness, which supported my scientific development while not compromising the enjoyment of this time in graduate school. Many thanks to Dr. R. Breyta for the significant time she spent training me at the bench, providing context that immeasurably helped my project, and always being willing to get in the weeds and work with me on the inevitable problems that arise during data collection and analysis. I also thank Dr. A. Kell for friendship, guidance, and at times, commiseration.

This work would not have been possible without the ongoing participation of fish health agencies that send us viral field isolates. Their support of this surveillance program extends to funding support I received from the US Fish and Wildlife Service FONS program. I also thank the National Science Foundation Graduate Research Fellowship Program for their support.

I am grateful to Dr. K. Edwards for awarding me the flexibility to pursue an exciting yet non-traditional project, and for reading my thesis. I also thank Dr. T. Bedford for sharing his expertise so openly and providing clear explanations to my litany of questions.

Special thanks goes to my family and to my partner. To my mom and dad, thank you for supporting me both emotionally and financially, and for your unconditional love and support. To J&J. C., thank you for welcoming me into the family and ensuring that I was always well fed. To S.C. thank you for weathering all of the ups and downs beside me. I can't imagine having done it without you.

Chapter 1. INTRODUCTION

1.1 HUMAN HEALTH AND WELLBEING AND THE AQUATIC ENVIRONMENT

Much of epidemiological research has focused on understanding disease incidence, distribution, and control for human pathogens and terrestrial animal pathogens. However, we live on a planet that is 71% water. Humans use aquatic environments as recreational resources and as valuable sources of both wild and farmed food. In contrast to the abundance and importance of aquatic environments, relatively few epidemiologic research programs have focused on aquatic animal health (Peeler & Taylor, 2011).

Aquatic environments support human health and wellbeing through the provision of food (through aquaculture and wild capture fisheries), potable water, irrigation, waste disposal, and also through their use as recreational resources (Peeler & Taylor, 2011). However they also feature high pathogen load and diversity. With viruses being the most abundant life form in the ocean, an estimated 10^{23} viral infections occur every second in the marine environment (Suttle, 2007). As our use of and impact on these environments shifts, we may facilitate changes in disease dynamics and disease emergence events.

While not as extensively discussed as terrestrial zoonotic emergence, pathogens maintained in aquatic environments can cause severe disease in humans. Aquatic bacteria are primarily responsible for human infections derived from fish contact (Lowry & Smith, 2007). For example, *Vibrio vulnificus* is the most common cause of fish-transmitted Vibrio infections in humans (Oliver, 2005). If contracted through a puncture wound, the disease in humans presents with necrotizing fasciitis and swelling of the wound area. If ingested (typically through consumption of shellfish), septicemic infections have mortality rates between 50 and 60%

(Oliver, 2005). In addition to *Vibrio vulnificus*, a variety of other bacterial pathogens maintained in aquatic hosts can cause human disease including: *Aeromonas hydrophila*, *Edwardsiella tarda*, *Erysipelothrix rhusiopathiae*, *Mycobacterium marinum*, *Salmonellosis*, and other *Vibrio spp* bacteria (New South Wales DPI Fish Health Unit, 2015).

1.2 Health effects of agricultural production and landscape management

Anthropogenic pressures and their impact on disease dynamics have been demonstrated in terrestrial systems. Intensification of terrestrial livestock production has been linked to a variety of major human disease emergence events. For instance, outbreaks of Nipah virus encephalitis in Malaysia and Singapore are likely attributable to agricultural practices facilitating disease emergence. Repeated introduction of Nipah virus from flying foxes into livestock pig populations likely created within-farm persistence dynamics that facilitated disease transmission to humans (Pulliam et al., 2012). Avian influenza provides another example of how pathogen emergence is linked to agricultural intensification. Intensive poultry farming practices create large, high-density populations of hosts, which upon pathogen introduction can facilitate viral adaptation and amplification (Gilbert et al., 2007; Graham et al., 2008; Kapan et al., 2006). These reactions increase the probability that a pandemic influenza variant may emerge (Jones et al., 2013). Disease emergence has also resulted due to landscape management of aquatic environments. For example, the damming of rivers and the construction of irrigation canals has increased the density of mosquito breeding sites, leading to greater frequencies of Rift Valley fever outbreaks (Pepin, Bouloy, Bird, Kemp, & Paweska, 2010).

Agricultural intensification is also occurring within aquaculture. We now farm over 230 different aquatic animal species (Subasinghe, 2005), and of the 110 million tons of food that was

supplied from aquatic environments in 2006, 47% was derived from aquaculture (F.A.O, 2008). Consumer demand and technological advances have largely driven the development of this industry, which has grown 10% per annum between 1970 and 2008 (F.A.O, 2008).

An increasing body of literature suggests that agricultural intensification in terrestrial systems may be a driver of zoonotic disease emergence in humans. While perhaps not as often considered, zoonotic pathogen emergence from aquatic animals does lead to human infections. Determining whether the impact of humans on aquatic environments through aquaculture and landscape management may also lead to increased human disease is thus warranted. In addition to the possible zoonotic emergence of aquatic pathogens in humans, aquatic animal disease has the ability to impact human health and wellbeing by threatening food security, recreational resources, and overall ecosystem function. Therefore, development of methods and further studies of aquatic infectious disease epidemiology are also warranted.

1.3 Infectious hematopoietic necrosis virus and Pacific salmonid fish

We present here an epidemiological study of infectious hematopoietic necrosis virus (IHNV), which is the most important viral pathogen occurring in trout and salmon in North America. Using sequences from a partial region of the glycoprotein gene (303nt midG region), phylogenetic analysis indicates three distinct subgroups of IHNV within North America, which are named for their geographic range: U (Upper) genogroup, M (middle) genogroup, and the L (lower) genogroup (Kurath et al., 2003). In the Pacific Northwest, U and M genogroup viruses are the most significant pathogen of Pacific salmon and trout respectively. Infection with IHNV causes acute systemic disease (Wolf, 1988), which in juveniles can result in mortality rates as high as 90% (Bootland & Leong, 1999; Groberg, 1983a, 1983b; LaPatra, Parsons, Jones, & McRoberts, 1993; LaPatra, Turner, Lauda, Jones, & Walker, 1993). Epidemics of IHN disease

were first described in cultured sockeye salmon juveniles in the Columbia River Basin in the early 1950s (Rucker, Whipple, Parvin, & Evans, 1953). However extensive mortality also occurred in juvenile hatchery sockeye at other sites in Washington (Rucker et al., 1953) and Oregon (W. H. Wingfield, Fryer, & Pilcher, 1969) and in juvenile Chinook salmon in California hatcheries (Ross, Pelnar, & Rucker, 1960; Wingfield, Nims, & Fryer, 1970). Presumably these epidemics originated due to the practice of feeding young cultured fish unpasteurized *O. nerka* viscera (Watson, Guenther, & Rucker, 1954; Wolf, 1988), but this practice was abandoned by the late 1960s.

IHNV is a single-stranded negative-sense RNA virus belonging to the *Rhabdoviridae* family. Within the rhabdoviral family, IHNV belongs to the genus *Novirhabdovirus*, which infect aquatic finfish hosts. Like other novirhabdoviruses, IHNV has a linear genome with ~11,000nt containing six genes ordered 3'–N–P–M–G–NV–L– 5'. These represent the nucleocapsid, phosphoprotein, matrix protein, glycoprotein, non-virion protein and RNA-dependent RNA polyermase protein genes respectively (Kurath, Ahern, Pearson, & Leong, 1985; Morzunov, Winton, & Nichol, 1995).

1.4 PACIFIC SALMON

Within the Pacific Northwest salmon are significant cultural icons and an important food source. There are 6 species of Pacific salmon: Chinook salmon (*Oncorhynchus tshawytscha*), sockeye salmon (anadromous form of *O. nerka*), steelhead trout (anadromous form of *O. mykiss*), coho salmon (*O. kisutch*), pink salmon (*O. gorbuscha*), and chum salmon (*O. keta*) (see Table 1.1 below for reference and discussion of anadromy). All species of Pacific salmon can be hosts of IHNV, however the majority of infections and the most important disease interactions

occur with the first three species: Chinook salmon, sockeye salmon, and steelhead trout. Thus these three hosts are the main hosts considered in my thesis work.

Table 1.1 List of Latin names, common names, and life history for the three major host species that are the focus of this work.

Host Species	Common Name	Life History
O. tshawytscha	Chinook salmon	Anadromous *
O. nerka	sockeye salmon	Anadromous
O. nerka	kokanee salmon	Freshwater only
O. mykiss	steelhead trout	Anadromous
O. mykiss	rainbow trout	Freshwater only

*Anadromous indicates that this fish will live both in freshwater and in saltwater. Generally, anadromous fishes will spend roughly the first year of their lives in freshwater, migrate to the ocean where they will mature for several years, and then return to freshwater to spawn. Fishes that are non-anadromous never migrate to the ocean, but rather spend their entire lives in freshwater.

1.5 Abundance of different Pacific salmonids by geography

Knowledge of host abundance is important for contextualizing disease incidence. In the calculation of fish abundances we considered the average number of juvenile fish released annually from hatcheries over the study time period of 1971 to 2013. Importantly, these estimates considered only the numbers of fish that were released from hatcheries, and therefore the numbers of matured adults that returned to the hatchery or numbers of wild fish are not included in these estimates. Additionally, we considered a binary geographic designation: the Columbia River Basin and watersheds that drain into coastal waters. While the Columbia River Basin is a contiguous watershed, coastal watersheds were divided into further subregions to measure host abundances due to differing data reporting practices. These subregions include Puget Sound, the Washington Coast, and the Oregon Coast. All fish abundance data were

extracted from the Regional Mark Information System, which is publicly available at http://www.rmis.org/rmis_login.php?action=Login&system=cwt.

Chinook salmon are the most heavily cultured species of salmonid fish in both the coastal watersheds geography and in the Columbia River Basin. Over the study time period (1971 – 2013), on average roughly 95 million juvenile Chinook salmon were released from hatcheries in the Columbia River Basin every year. Chinook culture is extensive in coastal watersheds as well, where on average around 53 million juvenile Chinook salmon are released into Puget Sound, 11 million are released off of the Washington Coast, and around 3 million are released off of the Oregon Coast. For all geographic areas except the Oregon Coast, Chinook salmon are cultured at higher abundances than other species of salmon.

Sockeye salmon (anadromous *O. nerka*) are cultured in the Columbia River Basin and in all coastal watersheds except those along the Oregon Coast. However, in contrast to Chinook salmon culture, sockeye salmon are cultured in far lower numbers. The greatest numbers of juvenile sockeye salmon are released into Puget Sound, where on average close to 9 million sockeye juveniles are released annually. Hatcheries along the Washington Coast culture the next greatest numbers of sockeye salmon, releasing a little over 500,000 juveniles annually. Finally, the sockeye salmon are cultured in the Columbia River Basin, although in low numbers. Annually, hatcheries in the Columbia River Basin release around 200,000 juvenile sockeye salmon.

Steelhead salmon (anadromous *O. mykiss*) are highly prized sport fishes and are widely cultured in both the Columbia River Basin and in all three major regions of the coastal watersheds geography. The greatest numbers of steelhead trout are cultured in the Columbia River Basin, with on average 10 million juvenile steelhead trout released annually. The next

greatest numbers of steelhead trout are cultured on the Oregon Coast; on average 6 million steelhead juveniles were released annually over the study time period. Steelhead trout are also cultured on the Washington Coast and in Puget Sound. In each of these two regions around 2 million juvenile steelhead trout were released annually over the study time period.

1.6 Salmonid habitat in the Pacific Northwest: comparison of the Columbia River Basin and coastal watersheds

Spatially, we consider two major geographic ranges in the Pacific Northwest: the coastal waters of Oregon and Washington, and the Columbia River Basin. When indicating the Columbia River Basin we refer to the entire watershed drainage area for the Columbia River and its tributaries. In total, this watershed basin covers roughly 258,000 square miles, an area slightly bigger than France. The river basin spans 7 states (Washington, Oregon, Idaho, Montana and parts of Wyoming and Nevada) and one province in Canada (British Columbia). Despite the extension of the Columbia River Basin into small areas of other states, the study area under consideration for this project included only Washington, Oregon, and Idaho.

The watershed is highly managed, with 14 hydroelectric dams on the Columbia River mainstem alone, and further numbers of dams on tributary rivers. Two of these mainstem dams, the Grand Coulee dam and the Chief Joseph dam, are the largest dams in the United States. Prior to major dam construction, the Columbia River Basin was one of the top rivers for salmon migration in the world. However, the impassability of some dams to fish, such as the Chief Joseph dam, has since greatly reduced accessible salmon habitat for anadromous fish. Construction of the Grand Coulee dam in 1941 blocked access to over 500 miles of the upper Columbia River (excluding tributaries), and by 2000, 50% of the Snake River, one of the main

tributaries to the Columbia River, was no longer accessible to fish. Before dam development an estimated 10-16 million adult salmon and steelhead trout migrated up through the Columbia River Basin. In 1995, only 672,100 adult fish returned through the Columbia River Basin (Washington Department of Fish and Wildlife & Oregon Department of Fish and Wildlife, 2002). While habitat restoration has occurred in the Columbia River Basin and fish returns are rising, the system remains impacted.

There is also a high demand on water in the Columbia River Basin for agricultural production. The Grand Coulee dam was originally built to provide water for irrigation. While now also used for hydroelectric energy production, the dam project provides irrigation water for 670,000 acres of Central Washington (US Department of the Interior Bureau of Reclamation, 2014). During the main growing season, 4.7 million acre-feet, or roughly 1.5E12 gallons of water, are pulled from the Columbia River mainstem for irrigation purposes (Washington State Department of Ecology, 2015).

In contrast, coastal watersheds are much smaller, simpler, and far less developed. We define coastal watersheds as rivers and river systems other than the Columbia River Basin that drain directly into the Pacific Ocean, the Salish Sea or into Puget Sound. These rivers feature significantly less human impact from hydroelectric dams and from irrigation draw down than the Columbia River Basin. Additionally, focus has been directed towards habitat restoration in coastal watersheds, as exemplified by the Elwha River Restoration project that saw both the Elwha dam and the Glines Canyon dam removed from the Elwha River in 2014 (US National Park Service, 2015).

While the Columbia River Basin and coastal watersheds are distinctly different habitats, they both are productive habitats with populations of wild and hatchery salmonid fish. Both the

Columbia River Basin and coastal watersheds have extensive federally-run, state-run, and tribal-run salmon hatchery programs. Within Washington alone there are 12 federal hatcheries, 83 state-run hatcheries, and 51 tribal hatcheries across Puget Sound, the Washington coast, and the Columbia River Basin. These hatchery programs serve to rear fish for commercial fisheries, for recreational fishing, and to help conserve naturally-spawning depressed fish stocks. Hatchery programs release millions of fish (further discussion on fish abundances below) with 75% of salmon caught in the Puget Sound and 90% of salmon caught in the Columbia River Basin originating from hatchery programs. (Washington Department of Fish and Wildlife, 2015).

1.7 Host and geographic ranges of North American IHNV

1.7.1 *Known associations between IHNV genogroups and host disease*

Within North America, there are three distinct genogroups of IHNV, the U genogroup, the M genogroup, and the L genogroup. Notably, different salmonid species show different susceptibility to IHNV, and this varies by genogroup. The U genogroup of IHNV likely represents the descendants of an ancestral virus that was associated primarily with sockeye salmon (Kurath et al., 2003). U genogroup viruses are extremely virulent in *O. nerka* (both sockeye salmon and kokanee), causing 69 -100% mortality in controlled laboratory experiments where live fish are exposed to virus. In contrast, M genogroup viruses appear to be less virulent in *O. nerka*, with laboratory experiments demonstrating between 0% and 20% cumulative percent mortality of *O.nerka* exposed to M viruses (Garver, Batts, & Kurath, 2006; Kurath and Peñaranda, unpublished data).

While M genogroup viruses demonstrate low virulence in *O. nerka*, they are much more virulent in rainbow and steelhead trout. In controlled laboratory experiments rainbow trout

experienced between 25% and 85% cumulative percent mortality when challenged with M type viruses. However when challenged with U type viruses, cumulative percent mortality was lower (between 5% and 41%) (Garver et al., 2006). In the field, M genogroup viruses were responsible for an observed emergence with mortality in steelhead trout hatcheries along the Washington coast from 2007 to 2011 (Breyta et al., 2013).

Finally, L genogroup IHNV appears to cause only low to intermediate mortality in both *O. nerka* and in *O. mykiss* (Garver et al., 2006), and L genotype viruses are detected predominantly in Chinook salmon (*O. tshawytscha*) (Kelley, Bendorf, Yun, Kurath, & Hedrick, 2007). L viruses have caused devastating epidemics in Chinook salmon in California hatcheries (Bendorf et al., 2007; Kelley et al., 2007). Experimental challenges of Chinook salmon with L viruses have also demonstrated 30 - 80% mortality (Bendorf, 2010; Hernandez & Kurath, unpublished data).

1.7.2 *Known associations between genogroup and geographic range*

Within North America, IHNV is detected and considered endemic down the west coast from Alaska to northern California and inland to southern Idaho (Wolf, 1988). However, each of the three North American genogroups (U, M, and L) have specific geographic ranges.

The U genogroup encompasses the largest geographic range. In their study of IHNV phylogeography in North America, Kurath et al. (2003) detected U genogroup virus from sites that were 3700km apart. Although historical observations indicate that the U genogroup was originally endemic in Alaskan sockeye salmon, U type viruses have since become widespread into British Columbia, Washington, and Oregon (Amend & Wood, 1972; Grischkowsky & Amend, 1976; Mulcahy et al., 1980), possibly due to historical practices of translocating salmon

or use of unpasteurized salmon by-products in fish feed (Burgner, 1991; Roppel, 1982; Wolf, 1988).

M genogroup viruses have a smaller range and are detected predominantly in the Columbia River Basin. While U and M type viruses overlap through many parts of the Columbia River Basin, only M viruses are found in the Hagerman Valley, a section of the Snake River that features intensive rainbow trout farming (Kurath et al., 2003). Outside of the Columbia River Basin, M viruses have rarely been detected on the Washington and Oregon Coasts, with the exception of one notable M virus event that occurred between 2007 and 2011, affecting 7 different fish culture facilities in 4 discrete watersheds (Breyta et al., 2013).

Finally, the L genogroup is detected primarily in California and on the Oregon Coast south of Cape Blanco (Kurath et al., 2003). Cape Blanco is a biogeographical transition zone for aquatic organisms and divides salmonid migratory patterns depending upon whether fish originate at watersheds north or south of the cape. This division in host migratory patterns also divides the viral population, and L viruses have not been detected north of Cape Blanco.

1.8 PREVIOUS IHNV REGIONAL EPIDEMIOLOGICAL STUDIES

Given the disease impacts of IHNV, epidemiological studies to investigate IHNV genetic diversity, geographic range, host tropism, and disease emergence events have been performed with some frequency. These studies often focus on highly localized study sites, specific regions, or on specific genogroups of IHNV. Many of these studies have focused on the molecular epidemiology of IHNV, namely distinguishing certain circulating viruses from others. While genetically distinct viruses may not be functionally different from each other, observing how genetically identical viruses and genetically distinct viruses circulate allows us to observe transmission dynamics and landscape patterns in viral incidence. Incidence records for genetic

types also allow investigation of patterns of viral evolution temporally, geographically, and in different hosts (Emmenegger, Meyers, Burton, & Kurath, 2000b).

A long-term goal of our group at the Western Fisheries Research Center (WFRC) has been to investigate the evolution and epidemiology of IHNV in the Pacific Northwest. Epidemiologic studies completed through this research program have generally focused on investigating the genetic heterogeneity of IHNV and incidence patterns in localized regions. For instance Anderson et al. (2000) used ribonuclease protection fingerprinting assays (RPA) on 42 isolates of IHNV to investigate transmission dynamics and genetic heterogeneity of IHNV in the Deschutes River watershed of Oregon. From these isolates 16 different haplotypes were observed, and the results indicated that epidemics that had occurred in kokanee salmon (landlocked *O. nerka*) in the area between 1991 and 1995 were due to a newly introduced virus that had been detected in wild adult spawning kokanee in 1988. This new virus type was subsequently transmitted to hatchery fish, demonstrating disease interactions between wild and cultured fish populations.

In 2000(b), Emmenegger et al. published a thorough molecular epidemiological account of IHNV in Alaska, finding that Alaskan IHNV was entirely U genogroup virus, that the maximum nucleotide diversity was 1.99%, and that Alaskan and British Columbian IHNV likely have a common ancestral lineage. Similar to the low genetic diversity of IHNV observed in Alaska, analysis of isolates of IHNV from coastal Washington by both RPA and genetic sequencing indicated low genetic diversity in IHNV occurring in Puget Sound and Washington coastal watersheds (Emmenegger & Kurath, 2002).

In contrast to the low diversity of IHNV in previous studies, IHNV isolates taken from trout farms in the Hagerman Valley, Idaho, were six times as diverse as IHNV along the

Washington and Northern Pacific coast (Ryan M Troyer, Lapatra, & Kurath, 2000). A sampling of 84 isolates from four trout farms within a 12-mile stretch of the Snake River indicated 46 different haplotypes by RPA, with sequence analysis indicating that all isolates came from the M genogroup and had a maximum nucleotide diversity of 7.6% (Ryan M Troyer et al., 2000). Further analysis of isolates of IHNV from the Columbia River Basin excluding the Hagerman valley indicated a higher level of genetic diversity than in IHNV in Alaska, but lower levels of genetic diversity than seen in the Hagerman Valley (Garver, Troyer, & Kurath, 2003). Indeed, analysis of the 303nt midG region of 120 virus isolates from the Columbia River Basin included both U and M genogroup viruses and indicated that viruses detected in the Columbia River Basin were 3 times more diverse than isolates from Alaska, but that they were 2 times less diverse than isolates from the Hagerman Valley (Garver et al., 2003).

While M genogroup is predominantly found in rainbow trout farmed in the Hagerman Valley (Troyer & Kurath, 2003) and steelhead trout in the Columbia River Basin (Garver et al., 2003), it was also the subject of the most recent IHNV epidemiologic study along the Washington Coast. Genetic surveillance indicated an emergence of M genogroup virus on the coast between 2007 and 2011 in coastal steelhead trout. Genetic sequencing of the 303nt midG region of 283 isolates taken over the 4 year time period indicated that two waves of viral emergence occurred and that adult steelhead trout from the Columbia River Basin were likely responsible for trafficking of M genogroup viruses out into coastal waters (Breyta et al., 2013).

1.9 IHNV GENETIC TYPING AND 'EVENT CODING'

As a technical assistance program, the Western Fisheries Research Center (WFRC) provides data from genetic typing of IHNV field isolates to fish health management agencies to help inform their management decisions. Genetic typing is performed on a specific region of the

G gene of the virus, which encodes the viral glycoprotein. Viral glycoproteins bind to host cell receptors to facilitate viral entry into the cell (Carter & Saunders, 2013) and are often under selection due to host immune pressures. Thus these regions of the viral genome tend to be more variable than other genomic regions.

When genotyping IHNV field isolates, we sequence a 303nt long portion of the IHNV G gene, referred to as the **midG** region of the genome. Each unique midG sequence is given a **unique sequence identifier (USD)**, which follows the format of an arbitrary three-digit number and the genogroup identifier, eg. mGXXXU/M/L. Sequences with unique USDs are considered different genotypes of IHNV. Different genotypes are not assumed to be functionally different from each other. Rather, midG genotypes are used as identifiers that allow us to conduct phylogenetic analyses and track population-scale patterns of viral incidence both spatially and by host species. Throughout this thesis 'genotype' and 'type' refer to midG sequences with a specific midG USD.

Some unique genotypes are detected only once or a handful of times, others are responsible for many occurrences of IHNV infection. To distinguish the latter kind of genotypes we refer to 'dominant' genotypes. Genotypes are considered dominant when they are detected at numerous different events at multiple different sampling sites or across wide temporal ranges. For example, the most dominant genotype that our study features is mG001U. This genotype was detected originally in 1973, and has been detected consistently up until 2011 at numerous sampling sites.

Units of surveillance in the WFRC IHNV technical assistance program are virus field isolates. As a passive surveillance program, we receive isolates from participating fish health agencies within the Pacific Northwest. Some agencies tend to submit far higher numbers of

isolates than others, or may submit more samples from certain occurrences of infection than from others. While we cannot correct for bias in terms of infection occurrence that is not sampled, we do correct for over representation of certain IHNV detections through the use of **'event coding'** (originally described in Breyta et al., 2013). The principle behind event coding is to reduce the bias created by unequal representation of virus detection events by different numbers of isolates. Some events have only one or two representative isolates, whereas another event may have as many as 30 isolates. The scale of this bias may mask certain real epidemiological patterns, and may artificially create others, since more densely sampled events may artificially elevate the incidence of a certain genotype, or specific fish populations may appear artificially affected. To assess the degree of bias that is introduced by using numbers of isolates as the unit of analysis, we do most analyses by *both* events and by isolates.

Events are represented by one isolate only and represent detection of a positive fish population. An event is recorded if field isolates demonstrate that a unique fish population is positive, or if a different genotype of IHNV is found in a population that is already positive for IHNV. To determine whether a population is unique, the following factors are considered: sampling year, sampling location, fish species, fish age (adult, yearling or juvenile), and for Chinook salmon, the fish run (either Fall or Spring). Any population that differs from another population in regards to one of the above factors is considered a unique population. When multiple genotypes are detected in a single positive population, multiple events are recorded as to represent the number of distinct genotypes detected. (Breyta et al, 2013).

Chapter 2. METHODS

2.1 IHNV SURVEILLANCE AND SUBMISSION OF ISOLATES

As a reference lab for infectious hematopoietic necrosis virus (IHNV), our lab has been involved in the collection and archiving of field isolates of IHNV since the mid 1960s. The IHNV genetic surveillance program is a passive surveillance initiative that began in the mid-1990s that collects viral field isolates from fish health agency laboratories.

At Pacific Northwest hatcheries, each fish stock being reared is assessed for IHNV annually by testing 60 to 100 returning adult fish (Thoesen, 1994). These adult fish are nearly always asymptomatic. Asymptomatic juveniles and yearlings may also be tested, however the majority of surveillance in juvenile and yearling populations occurs when symptomatic disease is present (fish are either dead or moribund). Tissues from infected fish are collected at hatcheries and sent to laboratories at fish health agencies where virus is grown in cell culture. Generally kidney and spleen tissue or ovarian fluid are used (Thoesen, 1994) and tissues may come from a single fish or from multiple fish (up to 5 fish may be pooled). Infectious virus is cultured on Chinook salmon embryo (CHSE-214) or epithelioma papulosum cyprinid (EPC) cell lines (Lannan, Winton, & Fryer, 1984). After usually 2 to 3 passages the tissue culture supernatant containing infectious virus is collected and frozen. Isolates are transported to our lab at the Western Fisheries Research Center (WFRC) on ice either by mail or by hand delivery. Epidemiological case data corresponding to the isolates is either mailed along with the isolates or is emailed directly to the lab.

Upon arrival at the WFRC lab the isolates are frozen at -80 degrees Celsius and are held there until we can extract RNA and genotype the isolate. The accompanying epidemiological

case data are also checked for errors and omissions. Upon receipt at WFRC, each isolate is assigned a 'Technical Assistance' or TA number. These numbers include the year the isolate was received at WFRC or was genotyped as part of the program, and an arbitrary three-digit number indicating the order in which the isolate was received (example TA-14-001).

While all received field isolates are assigned a TA number, isolates may not be genotyped in the order received. Since the IHNV genetic surveillance program serves to provide epidemiological information to fish health agencies and hatcheries managers, isolates from IHNV outbreaks or isolates from locations where IHNV has not been detected previously are prioritized for genetic typing. The sequence data generated and certain epidemiological data for over 2400 IHNV field isolates are maintained in a database that is publicly available online at http://gis.nacse.org/ihnv/.

2.2 RNA EXTRACTION AND SEQUENCING

The IHNV genetic surveillance program genotypes viral isolates according to a 303 nt portion of the IHNV glycoprotein (G) gene, hereafter referred to as the midG region. The midG region corresponds to nt 686 to 988 as numbered in the full length IHNV G gene sequence of GenBank accession number U50401 and has starting sequence GATTCC and end sequence CTT. As described in Breyta et al. (2013), TriReagent (Sigma) is used to extract viral genomic RNA from either 200µL or 500µL of virus culture supernatant, depending on the available volume of the received field isolate. Virus RNA is precipitated with tRNA (Sigma) to aid in RNA pellet visibility. Reverse Transcription Polymerase Chain Reaction (RT-PCR) is used to create and amplify a cDNA copy of a 550bp fragment containing the midG sequence as described by Emmenegger et al. (2000). RT-PCR is performed in a 50µL reaction using avian myeloblastoma

virus (AMV) reverse transcriptase, *Taq* polymerase, IHNV-specific primers (Table 2.1) and 30 cycles of amplification.

A subsample of 10 µL of RT-PCR products and the positive and negative controls are run on a 1% agarose gel. The negative control (containing no RNA template) ensures that the samples have not been contaminated, and the positive control (previously extracted and amplified viral RNA) serves to ensure that RT-PCR was successful in the case that isolates currently undergoing sequencing do not amplify. The remaining 40µL of PCR products are purified using the StrataPrep PCR purification kit (Agilent Technologies) and re-suspended in water yielding 50µL of each sample. Then, 0.5µL of the purified PCR products serve as templates for synthesis of fluorescently-labeled DNA through BigDye terminator cycle sequencing (Applied Biosystems) at one quarter strength of manufacturer specifications and using 2pmol/uL dilutions of HPLC-purified versions of the same forward and reverse primers as used in the RT-PCR reactions (Integrated DNA Technology). The labeled-DNA products are purified using Sephadex in Centri-Sep columns (Princeton Separations) and analyzed on an AB1-PRISM 310 genetic analyzer (Applied Biosystems). A PTC-100 thermocycler unit (Bio-Rad) was used for all PCR reactions.

Table 2.1: PCR primer sequences as described originally by Emmenegger et al. (2000). These same primers are used both in the reverse transcription PCR and in the sequencing PCR reactions.

Sense Primer: midG 517+	5' - AGAGATCCCTACACCAGAGAC - 3′
Antisense Primer: midG 1209-	5′ - GGTGGTGTTGTTTCCGTGCAA - 3′

2.3 SEQUENCE ANALYSIS

Sequence files representing the sense and antisense sequences were edited and analyzed using Sequencher v4.9 (Gene Codes Corp). In rare cases, heterogeneous sites were observed. Heterogeneous sites occur when, in both the positive and negative sense sequences, fluorescent peaks representing two different bases are both present at a single nucleotide site. This event may be attributable to the pooling of tissues from different fish infected by different IHNV genotypes, or by the presence of multiple genotypes within a single infected host. Heterogeneous sequences are labeled according to the multiple genotypes that they have present. Thus under our naming scheme heterogeneous isolates receive two midG unique sequence identifiers (USDs) indicating both genotypes found within the one sample (eg. mG001UmG174U). In rare cases a single isolate may contain a genotype that is heterogeneous at multiple sites. In such cases the individual sequence types in the mixture cannot be defined without subcloning, so these sequence types are labeled as mGXXXHH, to indicate the presence of multiple heterogeneous sites.

2.4 U GENOGROUP ISOLATES DATASET CREATION

The entire midG sequencing database at the Western Fisheries Research Center currently contains sequences and case information for over 2800 field isolates from sites in Alaska, British Columbia, Washington, Oregon, Idaho, Montana, and California. The dataset includes isolates that belong to each of the three major genogroups of IHNV in North America, U, M, and L. To create the U genogroup dataset used in this analysis, all M and L records were removed. Since our lab is the main reference center for IHNV genetic surveillance for Washington, Oregon, and

Idaho, the study sample area was limited to these three states. While the range of U genogroup extends into British Columbia and Alaska, these regions were not included in our U genogroup dataset since we are not the primary genetic surveillance lab for these regions, and therefore our records are less consistent for British Columbia and Alaskan U genogroup IHNV detections. Records for which the sequence type was not known, the host species was not known, or where the year of isolation was not known were removed from the dataset since these data are integral to the analyses performed on the data. U genogroup isolates isolated from January 1st, 1971 and through to December 31st 2013 were included in the dataset. While there are a few records of U genogroup detections back to 1966, their provenance and epidemiological case data cannot be completely verified for accuracy, and therefore they were excluded from the final dataset. The final U genogroup dataset included 1219 IHNV field isolates.

2.5 U GENOGROUP EVENTS DATASET CREATION

The full U genogroup isolates dataset was analyzed to determine which detection events might be represented by multiple viral isolates. As discussed in the introduction, specific criteria were employed to describe what kind of detection counts as a discrete event. Isolates were given an alphanumeric code indicating whether the cohort population was asymptomatic adult fish, epidemic juvenile fish or moribund fish. For the events dataset, only one isolate per event cohort was retained, thus ensuring that IHNV detection events with multiple isolates were not overly represented due to greater sampling intensity. This process resulted in a dataset of 619 U genogroup events that occurred between 1971 and 2013.

2.6 ASSESSING EVOLUTIONARY MODELS

Models of evolution allow us to correct for evolutionary distance that may not be observed simply by looking at pairwise nucleotide differences. Under a simple model, we would expect that as time since taxa divergence increases, the number of differences between the taxa per site should also increase, most simplistically as a linear function. However, some distances may not be observable by looking at nucleotide differences between the two sequences, since a site could mutate to a different base, and then mutate back to the original base, a process called back mutation. Thus, when differences per site are plotted against time since divergence, the function is not linear but instead plateaus at 0.75 differences per site over increasing time since divergence. This relationship occurs since even once a sequence has changed so completely that it is no longer related to the original sequence, we expect that 0.25 of the sites will by chance remain the same as in the original sequence. This expectation comes from the probability that by chance the same base could occur in both sequences out of the 4 possible base choices (A, C, T or G) (see Felsenstein, 2004: p155-159).

Over sufficiently long time spans, back mutation is expected to occur and an accurate representation of the evolutionary distance between two sequences cannot be inferred simply from the observable number of nucleotide differences between two sequences. Under such conditions it is necessary to employ an evolutionary model that can correct pairwise distances between sequences for distance that accrues yet is not observed.

Initially we used JModelTest version 2.7.1 (Darriba, Taboada, Doallo, & Posada, 2012) to determine which models of evolution best fit the U genogroup midG sequence data. The model search assesses 88 different models of nucleotide substitution, including models with uneven base frequencies and models with invariant sites. Model fit was assessed using the

corrected Bayesian information criteria (Schwarz, 1978). JModelTest found that the U genogroup midG sequence data was best fit by a Tamura-Nei model (Tamura & Nei, 1993) with equal base frequencies and rate variation between sites.

Despite the results of JModelTest and our knowledge that RNA viruses can potentially have fast evolutionary rates, we were unsure whether enough time had passed within the study time frame for significant amounts of base back mutation to occur. We thus compared genetic distances measured by the number of nucleotide differences between sequences to genetic distances inferred under different models of evolution. Distance correction was performed under three different models: a Jukes-Cantor model (Jukes & Cantor, 1969) with gamma distributed rates, a Tamura-Nei (Tamura & Nei, 1993) model with gamma distributed rates (as recommended by JModelTest), and a Maximum Composite Likelihood method (Tamura, Nei, & Kumar, 2004) with gamma distributed rates. All three models used a gamma shape parameter of 4. This parameter determines the shape of the distribution from which site-specific rates are drawn.

2.7 ASSESSING MOLECULAR CLOCKS

In MEGA version 6.06 (Tamura, Stecher, Peterson, Filipski, & Kumar, 2013) a maximum likelihood tree was produced utilizing the full dataset of 619 U genogroup events between 1971 and 2013. A Tamura-Nei (Tamura & Nei, 1993) evolutionary model was employed, and rates were gamma distributed with 4 categories as recommended by JModelTest. The resulting tree was then imported into Path-O-Gen version 1.4 (Rambaut, 2010) to assess the clockliness of U genogroup IHNV midG sequences. Linear regression and R^2 values were used to assess the fit of a constant global clock. Path-O-Gen infers genetic distances between the inferred root of the tree and the sampled tips of the tree. These distances were exported from

Path-O-Gen and were plotted against sampling dates in R v3.0.2 (R Core Team, 2013). Linear and non-linear regression was performed in R on data for all U events, UC events, and UP events.

2.8 Bayesian coalescent analysis

FASTA format text files of sequences under analysis were loaded into BEAUti 1.8.0 (Drummond, Suchard, Xie, & Rambaut, 2012). Tree topology should not change when multiple taxa of the same sequence type are or are not included in the analysis, so for visual ease tree topologies were inferred using single representatives of each midG USD for all U (n=158), M (n=139), and L (n=21) genotypes. The selected sequences represent all U, M, and L genotypes detected before and including in 2013 that our lab currently has records for (Breyta & Kurath, unpublished data). Because this dataset included all known taxa, including U genotypes detected in Canada and Alaska and earlier than the study time frame, the total number of U genotypes included in the phylogeny (n=158) is greater than the number of U genogtypes that were detected within our study area between 1971 and 2013 (n=114). Tip dates for each taxon were set as the year that the genotype was first isolated (referred to as 'first detections'). No further resolution by month or day of isolation was included.

Both simple low parameter models and more complex high parameter models were tested. The U, M, and L first detections data were analyzed using: 1) a strict molecular clock and a constant population size demographic model (Kingman, 1982), 2) a strict molecular clock and GMRF Skyride demographic model (Minin, Bloomquist, & Suchard, 2008) and 3) a relaxed lognormal clock (Drummond et al., 2006) with a GMRF Skyride demographic model. The lognormal relaxed clock and the GMRF Skyride demographic model are more flexible priors, as

described below. Log files from the three differently parameterized runs were analyzed using the AICM (Baele et al., 2012) in Tracer version 1.6 to determine the best fitting model.

Bayesian models for the estimation of demographic history from nucleotide sequence data have the advantage that they can co-estimate tree topology, population history dynamics, and parameters used in the nucleotide substitution model simultaneously (Drummond, Rambaut, Shapiro, & Pybus, 2005). The Bayesian GMRF Skyride model implemented in BEAST assumes that population sizes between adjacent coalescent intervals will be correlated. Differences in population sizes between neighboring coalescent intervals are penalized, either through a simple penalty or a penalty that accounts for the relative length of the coalescent interval (with longer intervals being penalized less). This system of penalization introduces an assumption that over time a population will change gradually as opposed to abruptly (Simon Y W Ho & Shapiro, 2011).

Summarizing clock priors, a strict clock assumes that the rate of evolution is constant across all lineages. Relaxed clock priors allow for rate variation across different branches of the tree. The relaxed uncorrelated lognormal clock model in BEAST assumes that each branch has an independent rate of evolution. The rate for each branch is drawn from an underlying lognormal distribution (Drummond et al., 2006). In our analysis, under both the strict clock model and the relaxed lognormal clock model, the evolutionary rate prior is extrapolated from a gamma distribution encompassing the typical range of error rates exhibited by RNA-dependent RNA polymerases (Biek, Drummond, & Poss, 2006; Drummond et al., 2006; Drummond & Suchard, 2010).

Estimates of the time to the most recent common ancestor (tMRCA) and mean substitution rates were inferred using BEAST suite version 1.8.1 (Drummond et al., 2012) run

through the CIPRES science gateway (Miller, Pfeiffer, & Schwartz, 2010). Analysis for each model used 50 million Markov chain Monte Carlo (MCMC) iterations with a 10% burn-in. Trace files for the runs were checked visually in Tracer to ensure MCMC chain convergence and good mixing. Additionally, estimates were only taken from runs with effective sample sizes greater than 200. Statistical uncertainty is illustrated by 95% highest probability density (HPD) intervals around the point estimate.

The maximum clade credibility tree was inferred using TreeAnnotator v1.8.0 from 10,000 trees with a 20% (2000 trees) burn-in. Using the maximum clade credibility tree, figures were created in FigTree v1.4.0. Posterior probability values represent the node support for the grouping of all lineages that descend from that node. Posterior support is shown as a decimal and ranges between 0 and 1 where 0 indicates no support for the grouping and 1 indicates complete support for the grouping off of the node. Generally, scores greater than or equal to 0.7 are considered strong support for a distinct lineage descending of off the supported node.

2.9 Calculation of F statistics

Two different potential drivers of population structure were assessed: geography and host species. To test for population structure due to geography, data on U genogroup detections were split according to a binary geographic designation. Detections (either isolates or events) were considered to be part of the ‘Columbia River Basin’ population if they had occurred at: (1) sites on the main stem of the Columbia River, Clearwater River, or Snake River, or (2) sites on any tributary to the Columbia River, Clearwater River, or Snake River. Detections occurring in coastal watersheds (draining into the Pacific Ocean or Puget Sound) or at any site on a direct tributary to coastal waters were considered part of the ‘coastal watersheds’ population.

To test for host species as a driver of population structure all detections that occurred in either *O. tshawytscha* (Chinook salmon) or in *O. nerka* (sockeye and kokanee salmon) were selected from U genogroup data. F statistics based on both host species populations and geographic populations were calculated using both the U genogroup isolates dataset and the U genogroup events dataset.

To prepare the data for processing, midG sequences for each included sequence were transformed into FASTA format, where the title included a single letter label designating the group that the sequence belonged to (either Columbia River Basin or coastal watersheds, or *O. tshawytscha* or *O. nerka*). Population structure was assessed using tests for compartmentalization in HyPhy version 2.2.3 (Kosakovsky Pond, Frost, & Muse, 2005). Pairwise distances were estimated using maximum likelihood distance estimation and using Jukes-Cantor (Jukes & Cantor, 1969) as the evolutionary model underpinning the substitution matrix. Branch lengths were estimated using fixed rates across the entire branch. Using these pairwise distances, HyPhy determines the mean subpopulation diversity (π_S), the mean interpopulation diversity (π_B), and the mean total genetic diversity (π_T). These nucleotide diversities represent the mean number of nucleotide differences per site that would occur between two sequences drawn at random from the population (Li, 1997). The Hudson, Boos, and Kaplan (1992) method for detecting genetic differentiation due to geographic subdivision was used to calculate F_{ST} and is given by the formula: $(\pi_T - \pi_S)/ \pi_T$, whereby $\pi_T - \pi_S = \pi_B$. The sampling characteristics of the estimators were determined using bootstrapping with 1000 replicates, and the probability of observing the point estimate or higher was determined using permutation tests with 1000 replicates.

Chapter 3. PHYLOGENETIC INFERENCE OF A SUBGROUP WITHIN THE U GENOGROUP

3.1 INTRODUCTION

Both the M genogroup and L genogroup of IHNV demonstrate further phylogenetic structure within the genogroup. For instance the L genogroup contains two strongly supported and distinct lineages, LI and LII, whose detection correlates with geographic range of detection (Kelley et al., 2007). Within the M genogroup there are 4 strongly supported sub-lineages: MA, MB, MC, and MD (R M Troyer, LaPatra, & Kurath, 2000). In contrast, previous phylogenetic analysis of the U genogroup demonstrated that the U genogroup was homogenous, lacking any further strongly supported lineages within the genogroup (Kurath et al., 2003). That analysis included 39 representative midG sequence types for the U genogroup. Since then, the IHNV technical assistance program for genotyping field isolates has grown, and the current dataset under analysis contains 114 unique genotypes representing 1216 isolates and 619 events of U genogroup IHNV. With increased data and better methods, we wanted to investigate whether further substructure within the U genogroup was developing.

We postulated that substructure might be developing within the U genogroup due to dynamics we were observing in the U genogroup case data. With increased data available on U genogroup IHNV events, epidemiological dynamics were more apparent in the infection incidence data. Some sequence types were detected in excess of 30 separate events, allowing us to see general patterns in the geographic range and host tropism of certain IHNV genotypes. Such was the case for two broadly detected U genotypes: mG001U and mG002U. Originally detected in 1973, mG001U was responsible for the largest number of U genogroup IHNV events.

It had a long temporal span, and was most recently detected in 2011. Genotype mG002U was also detected at many different sites and was detected between 1984 and 2012 (Table 3.1). As data on these two genotypes accumulated, we noticed apparent spatial resolution in the detection sites of these two genotypes. Although broadly found within the Columbia River Basin, we noticed that of our 342 isolates of mG001U, only 8 (2.3%) had ever come from sampling sites outside of the Columbia River Basin (Table 3.1). In contrast, mG002U also displayed a broad geographic range, however for sampling sites *outside* of the Columbia River Basin. Of our 73 isolates of mG002U, only 5 (6.8%) were isolated at sampling sites within the Columbia River Basin. This finding is especially interesting given the expansive range of mG002U outside of the Columbia River Basin; mG002U is detected at coastal sampling sites in Washington, and also beyond the range of our dataset in British Columbia and Alaska.

With further surveillance and genotype data, this pattern of geographic separation between genotypes circulating in the Columbia River Basin and in coastal watersheds continues to hold. This spatial resolution indicates that viral populations to a certain extent are not mixing, which theoretically could drive the development of separate phylogenetic lineages within the U genogroup. Thus we hypothesized that an updated phylogenetic analysis of the U genogroup would possibly demonstrate sub-lineages correlating with detection primarily in the Columbia River Basin or in coastal watersheds.

Table 3.1: Patterns in geography of detection for isolates of mG001U and mG002U. Detections sites are either considered to be in the Columbia River Basin (CRB) or outside of the Columbia River Basin. While the primary dataset is restricted to Washington, Oregon, and Idaho, U genogroup IHNV is detected frequently north of Washington, thus here we also include detections of U genogroup IHNV in British Columbia and Alaska.

Dominant Genotypes (year range)	Total numbers of Sites	Total numbers of Isolates	Isolates from CRB sites	Isolates from outside CRB
mG001U (1973 – 2011)	65	342	334 (97.7%)	8 (2.3%)
mG002U (1984 – 2012)	37	73	5 (6.8%)	68 (93.2%)

3.2 DOES INCREASED DATA PROVIDE INCREASED PHYLOGENETIC RESOLUTION?

Previous phylogenetic analysis of IHNV in North America, including analysis of the U genogroup, used neighbor-joining distance trees (Kurath et al., 2003). Neighbor-joining is a clustering algorithm where tips with the least amount of evolutionary distance between them are joined, iterating through taxa groups until all taxa are finally joined. Trees constructed using the neighbor-joining algorithm will recover the true tree topology assuming that the evolutionary distances between taxa accurately reflect the true evolutionary trajectory (Felsenstein, 2004, pp.166-167).

The current dataset features far greater numbers of isolates and sequence types than previous analyses. For instance, the Kurath et al. (2003) phylogeographic analysis of North American IHNV includes 39 U genogroup sequence types. The dataset currently under analysis contains 114 unique U genogroup sequence types. To test whether further phylogenetic structure within the U genogroup was not observed due to insufficient data, we analyzed the current 114 U sequence types, along with M and L genotypes, using Neighbor-Joining trees as used in Kurath et al. (2003).

Despite including more taxa within the U genogroup, the current Neighbor-Joining tree did not reveal any subgroup differentiation within the U genogroup (not shown). Rather, the U genogroup continued to appear homogenous. This finding also held when a U genogroup IHNV phylogeny was constructed using a Maximum Likelihood approach (Figure 3.1).

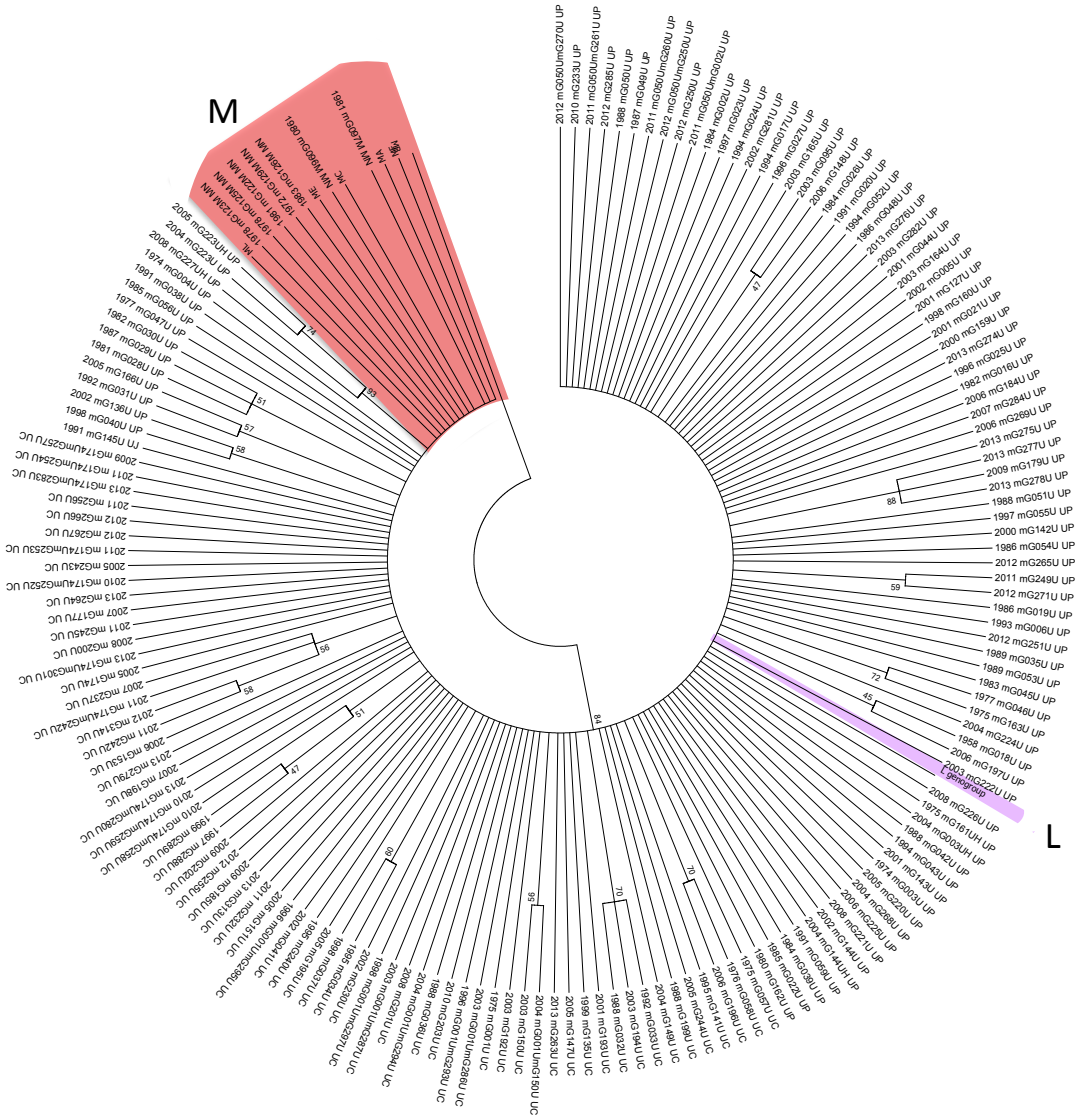


Figure 3.1: Maximum Likelihood phylogenetic tree of U genogroup taxa. Consensus of the topology was tested through 1000 bootstrap replicates, and numbers at nodes indicate bootstrap support on that node. Nodes with less than 45% bootstrap support have been collapsed. M and L sub-lineages do not show known subgroup structure because they have been collapsed to limit their space occupied on the tree.

3.3 BAYESIAN COALESCENT PHYLOGENETIC ANALYSIS

Increasingly, Bayesian coalescent methods of phylogenetic inference have been adopted for inferring phylogenies and population dynamics from viral sequence data. Coalescent methods are generally considered more appropriate for this kind of polymorphic data for a number of reasons. Firstly, phylogenetic methods, such as Neighbor-Joining, model relationships of species descent by creating a genealogy in which each species is usually represented by one sequence. The genealogy is then equated with the species tree, which is generally accurate in most scenarios. However, for complicated demographic scenarios shaped by selection, migration, recombination, and other processes, different genes may estimate different genealogies, thereby inhibiting inference of a single species tree from the genealogy (Rosenberg & Nordborg, 2002). Additionally, a degree of randomness of evolutionary forces shapes observed polymorphic data, and therefore methods that model evolution as a stochastic process, such as the coalescent, are preferable to heuristic phylogenetic models (Rosenberg & Nordborg, 2002).

Using Bayesian coalescent evolutionary analysis implemented in BEAST 1.8.1 (Drummond et al., 2012), we investigated the phylogenies on an updated dataset including all known genotypes of U, M, and L genogroups of IHNV in North America. Since our study of U genogroup IHNV was limited geographically and temporally, there are some U genotypes included in the phylogeny that were not detected within our study of U genogroup in the Pacific Northwest. As seen in previous trees, both M and L genogroups continue to demonstrate further phylogenetic substructure. In contrast to previous trees, within the coalescent tree a distinct lineage within the U genogroup is apparent (Figure 3.2). Rather than a demonstrable bifurcation with strong support for two separate continuations of the lineage, the phylogenetic tree shows the U genogroup as a generally homogenous group with one well-supported subgroup. At its most

basal node this U sub-lineage has a posterior node support of 0.70, demonstrating strong support that genotypes of this lineage group together separately from other genotypes within the U genogroup. The next most basal node within the newly recognized U lineage has even stronger support with posterior node support of 0.77. This node groups a subset of 64 genotypes together, and the slightly more basal node also includes both mG057U and mG058U, relatively old genotypes that were first detected in 1975 and 1976 respectively. Importantly, mG057U was only ever responsible for two events, and mG058U for one event.

To distinguish subgroups from the greater U genogroup we refer to the new sub-lineage as UC, since the genotypes that group within this lineage are detected primarily in the Columbia River Basin. We refer to all other U genotypes that do not fall within the UC lineage as UP genotypes, since they are detected primarily in Pacific coastal waters. In total there are 66 UC genotypes and 92 UP genotypes represented on the phylogeny.

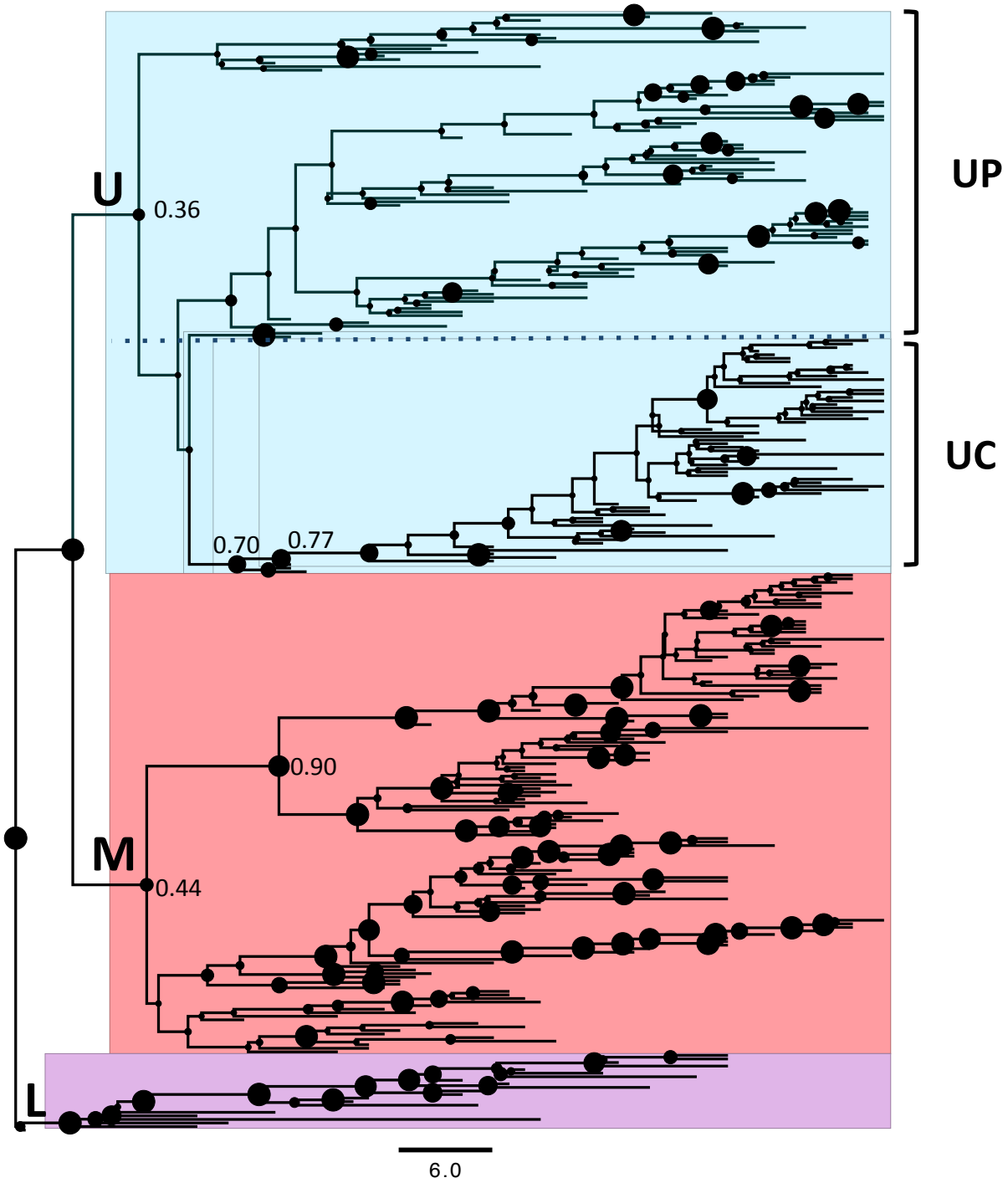


Figure 3.2: Bayesian coalescent tree of the U, M, and L genogroups of IHNV. Node shapes are scaled to show increasing posterior support as circle size increases. Numerical estimates of the posterior support are shown on some nodes for reference, and are shown specifically for the newly detected lineage within the U genogroup. Branch length represents number of nucleotide substitutions per site per year.

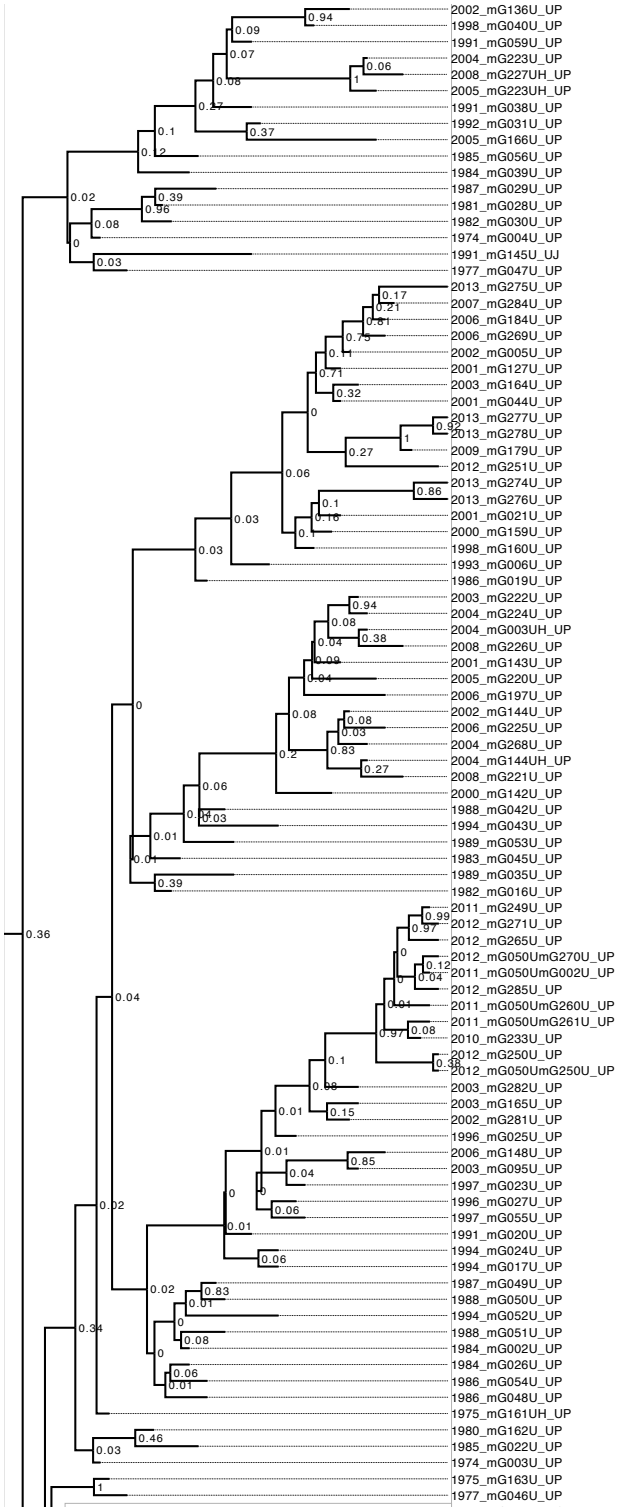


Figure 3.3: Subset of greater U, M, and L tree representing all 92 UP taxa (above dotted line in Fig 3.2). The name of each genotype is indicated, and the posterior support for each node is also provided.

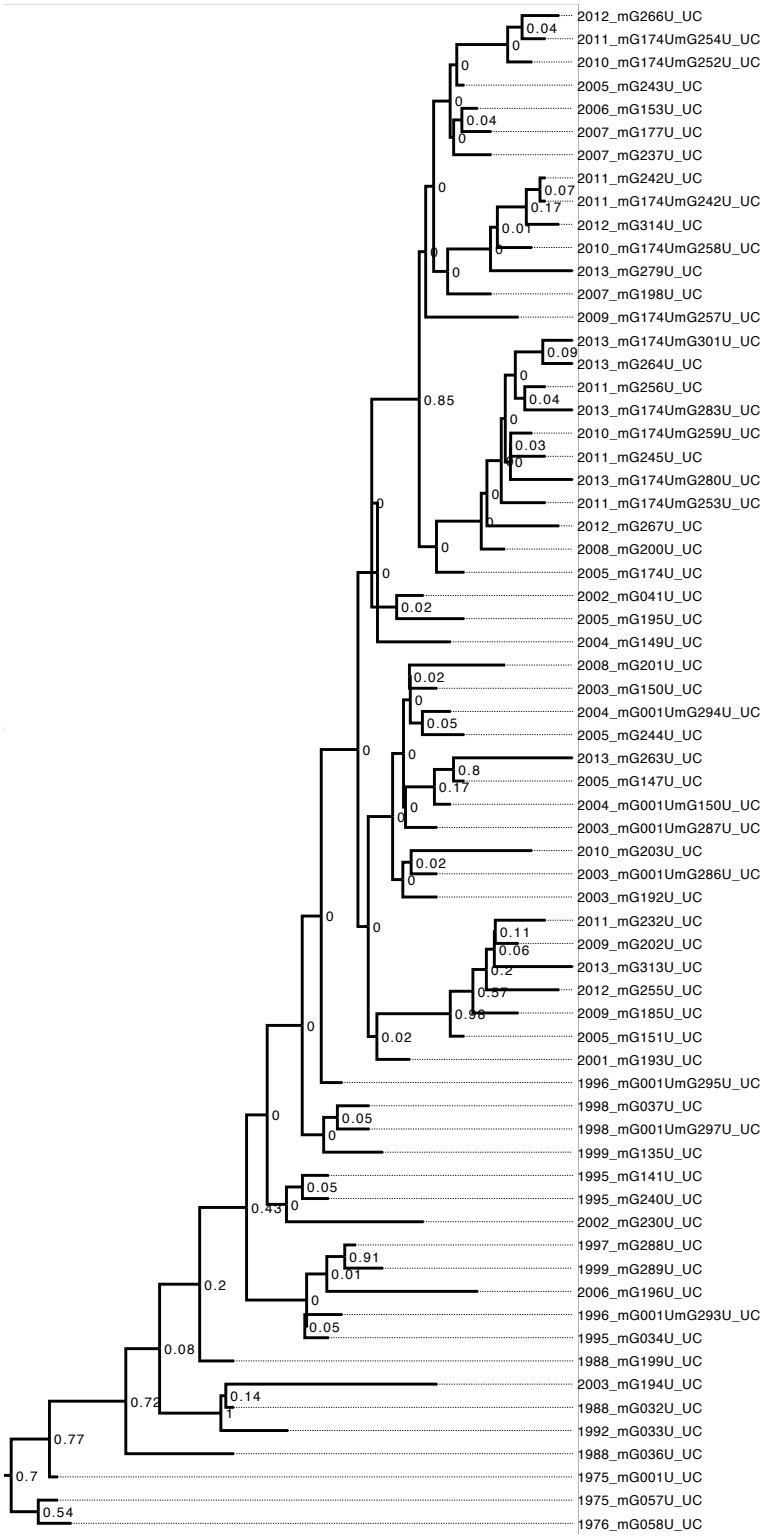


Figure 3.4: Subset of greater U, M, and L tree representing all 66 UC taxa (below dotted line in Fig 3.2). The name of each genotype is indicated, and the posterior support for each node is also provided.

3.4 PARAMETER INFERENCE FROM COALESCENT TREES INCLUDING ESTIMATES OF TMRCA

Generally, when coalescent models for phylogenetic inference are used, the goal of the analysis is not to investigate tree topology, but rather to infer the population level dynamics that give rise to patterns of coalescence in the data. Conceptually, the coalescent is the process whereby sampled lineages (haplotypes or genotypes) are considered to randomly 'choose' their parental haplotypes from all possible haplotypes in the previous generation. When two lineages randomly choose the same parent, then those two lineages coalesce, and this process iterates until the final two lineages coalesce. The rate at which these coalescent events occur is influenced by the size of the population. With greater population sizes there are more parental haplotypes to pick from, which lowers the rate at which child lineages pick the same parent (the rate of coalescence) (Rosenberg & Nordborg, 2002). The rate of coalescence also changes as the number of lineages choosing their parents changes. With more lineages picking parents, the rate of coalescence increases, and with fewer lineages picking parents, the rate of coalescence decreases (Rosenberg & Nordborg, 2002).

Considering viral population dynamics, reconstruction of the coalescent genealogy facilitates inference of the timing that a pathogen was introduced into the host population and how rapidly evolution is occurring within the hosts. Often, when using coalescent methods to infer these kinds of dynamics, the tree topology is considered simply a nuisance parameter (Rosenberg & Nordborg, 2002). In this regard, we use coalescent methods in a slightly non-traditional way since we are interested in both viral population dynamics and in the detection of strongly supported sub-lineages hypothesized by observational patterns in the data.

Coalescent analysis relies on an assumption that the sequences used in the analysis represent a random sample of haplotypes in the greater population, and that haplotype frequencies in the analysis therefore represent the haplotype frequencies of the population. Thus in contrast with most phylogenetic analyses, where one representative sequence per taxon is used, identical sequences should be included in the analysis at the same frequency as that haplotype occurs in the greater population. To use only unique taxa artificially portrays the underlying population as highly diverse. This signal occurs since a population would necessarily be exceedingly diverse for a random sampling of the population to return only unique haplotypes.

Despite the inherent bias in performing coalescent analysis on datasets including only unique genetic sequence types, researchers have published coalescent analyses of IHNV using unique sequences only, as in He, Ding, He, Yan, & Teng (2013). Although estimates of population parameters from the analysis are biased, we also conducted a coalescent analysis on unique taxa to determine whether others' results are replicable with different samples of U genogroup IHNV. We stress however that to estimate parameters such as evolutionary rate and tMRCA accurately, coalescent analyses should be conducted on data that represents as close to a random sample of the underlying viral population as possible, such as the events data. We have been pursuing a coalescent analysis of all U genogroup events, however this analysis has not yet been successful. Therefore, the coalescent analysis we present below is our best current analysis, but additional analyses will be pursued in the future.

A coalescent tree using U, M, and L genotypes was constructed under multiple demographic models and clock models (see Methods for an in depth discussion). The tree reconstructed using a lognormal relaxed clock (Drummond et al., 2006) and GMRF Skyride

(Minin et al., 2008) as the demographic prior was determined to be the best fitting model as assessed through AICM (Baele et al., 2012). From this model, the estimated tMRCA for the U, M, and L genogroups is 55.8 years (95% HPD: 55.0 – 57.1 years). Given that the most recent taxa were isolated in 2013, this places the root of the tree for all North American IHNV in 1957 (95% HPD between roughly 1956 and 1958). The mean rate of evolution across the entire tree (U, M, and L genogroups) was estimated at 7.79E-04 substitutions per site per year (95%HPD: 6.27E-04 – 9.32E-04).

While similar to He et al.'s (2013) findings that U and L likely emerged around 1964, and M a little later in 1970, we stress that these estimates of tMRCA are not accurate. Since only unique taxa are used, the underlying population appears artificially diverse. Extreme diversity of haplotypes would imply that the population under consideration is very large, and thus time to coalescence would likely be systematically underestimated (as population size increases the rate of coalescence decreases). However, as the number of lineages choosing parents increases, the rate of coalescence increases (Rosenberg & Nordborg, 2002). Thus it is challenging to predict how estimates of coalescent rates and tMRCA may change when datasets that mimic a random sample of the underlying population are analyzed, and further research in this direction will be prioritized.

3.5 Genetic diversity of the U genogroup and UC and UP subgroups

To test whether observed genetic distances are accurate measures of total pairwise genetic distance we compared genetic distances under different evolutionary models to raw numbers of nucleotide differences. We found that the mean pairwise distance for all U genogroup events was exactly the same between the raw distances and distances calculated under

three different evolutionary models: Jukes-Cantor (1969), Tamura-Nei (1993), and Maximum Composite Likelihood (2004) (Table 3.2). For all evolutionary models, rate heterogeneity between nucleotide sites was modeled as a gamma distribution with a shape parameter of 4. The only differences observed between the raw distances and among the three different models of evolution was in the maximum pairwise genetic distances. The raw maximum genetic pairwise distance was the lowest at 0.043 differences per site, Jukes-Cantor (1969) produced the next lowest estimate at 0.045 differences per site, and both the Tamura-Nei (1993) model and the Maximum Composite Likelihood model (2004) estimated the maximum pairwise distance as 0.046 differences per site.

Table 3.2: Pairwise genetic distances calculated as the raw number of nucleotide differences and as corrected under different models of evolution. Pairwise distances represent the raw or corrected number of nucleotide differences between two sequences divided by the midG sequence length (303nt).

ALL U - EVENTS	Maximum Pairwise Distance	Minimum Pairwise Distance	Mean Pairwise Distance
Raw distances	0.043	0	0.010
JC69 + gamma(4)	0.045	0	0.010
TN93 + gamma(4)	0.046	0	0.010
MCL + gamma(4)	0.046	0	0.010

Given how similar the genetic pairwise distance estimates are, analysis of nucleotide diversity within the whole U genogroup and the UC and UP subgroups was performed using the unadjusted raw number of nucleotide differences. Under certain circumstances observed genetic distances must be corrected for back mutation, in which a nucleotide change occurs but changes back to the original base. Back mutation results in genetic distances that cannot be observed by

comparing nucleotide differences. Under such circumstances, evolutionary models are used to estimate unobserved genetic distance. However use of these models relies on the assumption that the investigator knows which model of evolution best fits their data. Since this assumption may often be violated, and correction under evolutionary models here does not appear to produce different estimates of genetic distance between U genotypes, we analyzed our data using simply the observable number of nucleotide differences between two sequences.

We characterized the intrapopulation nucleotide diversity of all U genogroup events, events caused by UC genotypes, and events caused by UP genotypes using the minimum, maximum, and mean pairwise distances between genotypes responsible for each event (Table 3.3). Since there are 619 events attributable to 114 distinct sequence types, multiple events were attributable to the same sequence type, and thus the minimum evolutionary distance between events was zero for each dataset. We also calculated these distances using all genotyped isolates from each event. The isolates dataset had some overrepresentation of single events, thus making the groups appear more homogenous. This artificial attenuation of the mean pairwise genetic distance can be seen when comparing the mean distances between events and isolates data, and is included to illustrate how much artificial attenuation may occur through use of isolates data (Table 3.3).

Using the events data, all U genotypes had a maximum pairwise diversity of 13 nucleotides (4.29%) and a mean pairwise diversity of 2.92 nucleotides (0.96%). Comparison of the UC genotypes to the UP genotypes showed that UP genotypes were more diverse than UC genotypes. UP genotypes had a higher mean pairwise genetic distance and a higher maximum pairwise distance (Table 3.3). The mean pairwise distance between events caused by UP genotypes was 2.78 nucleotides (0.91%) compared to UC events, which had a mean pairwise

genetic distance of 1.25 (0.41%) nucleotides. Events caused by UP genotypes also showed a higher maximum pairwise distance with 12 (3.96%) nucleotide differences. Looking solely at events caused by UC genotypes, the maximum nucleotide difference was 6 (1.98%) nucleotides.

Overall, the UP subgroup appeared to be approximately two-fold more diverse than UC, and almost as diverse as the whole U genogroup.

Table 3.3: Pairwise genetic distances between sequences representing events or all virus isolates in datasets containing U genotypes, UC genotypes only, and UP genotypes only. Genetic distances are given by the numbers of nucleotide differences between two sequences, and numbers in italics represent the numbers of nucleotide differences divided by the total sequence length (303 nucleotides). For mean nt distances, this latter measure is an indicator of intrapopulation genetic diversity known as π (Nei and Li, 1979).

	Minimum Distance	Mean Distance		Maximum Distance	
ALL U - EVENTS	0	2.92	*(9.64E-03)*	13	*(0.0429)*
ALL U - ISOLATES	0	2.77	*(9.15E-03)*	13	*(0.0429)*
UC - EVENTS	0	1.25	*(4.14E-03)*	6	*(0.0198)*
UC - ISOLATES	0	1.22	*(4.02E-03)*	6	*(0.0198)*
UP - EVENTS	0	2.78	*(9.17E-03)*	12	*(0.0396)*
UP - ISOLATES	0	2.33	*(7.69E-03)*	12	*(0.0396)*

3.6 Molecular clock models of the U genogroup and UC and UP subgroups

Molecular clocks allow us to use genetic data to model the rate of evolution across lineages. In their simplest form evolutionary rates are approximated by linear relationships between genetic distance and time, and are considered to be constant across all lineages within a phylogeny. In reality rates may be heterogeneous across lineages, or rates may not be well approximated by a linear relationship (Ho & Duchêne, 2014). Notably, two important parameters can be inferred from molecular-clock plots: the time to the most recent common ancestor (the X-intercept of the regression line) and the rate of evolution (the slope of the regression line). For these estimates to be accurate, the clock model must be a good fit for the data. Model fit can be assessed through analysis of summary measures of the model such as R^2 for linear regression models that approximate constant clocks.

We plotted genetic distances as a function of sampling time to investigate the evolutionary rates across the U genogroup as a whole and also across the UC and UP subgroups separately. Genetic distance was measured as the root-to-tip distance for each taxa (tree tip) taken from maximum likelihood trees run for all U events, UC events only, and UP events only (trees not shown). The divergence between the root of the tree and each tree tip measures genetic distances as the number nucleotide mutations per site accumulated over the entire root-to-tip branch length. The genetic distances were plotted against the year the sample was isolated. A linear regression line was plotted to approximate a constant rate of divergence across the group under analysis. *Lowess* models were also included to show possible non-linearity of the data. Lowess (locally weighted scatterplot smoothing) is a non-parametric regression technique that fits a low-order polynomial model to each point in the scatterplot using weighted least squares.

Data closer to the point under consideration are weighted more heavily (Cleveland & Devlin, 1988). Thus the resulting regression line does not follow a specified model, but rather infers one from the data.

Under a constant molecular clock, indicated by the linear regression line, analysis of all 619 U genotype events yielded a rate of divergence of 2.75E-04 mutations per site per year (Figure 3.5). The linear regression between root-to-tip divergence (genetic distance) and sampling year is significantly different from zero ($p<0.0001$). However, U events demonstrate a very weak linear relationship between genetic distance and year of isolation ($R^2 = 0.195$), and therefore the divergence rate is likely better approximated under a non-linear model. Since the linear model is a poor fit for the data, parameters estimable from the regression line should be interpreted carefully. For instance, the time to the most recent common ancestor (tMRCA) of the U genogroup can be inferred from the X intercept of the regression line. Namely, when the divergence between the root and the tip is zero, then you have reached the root. In Figure 3.3, the X intercept indicates the sampling year for the root, thus allowing inference of root age as around 1959 (X-intercept is 1958.7). However, the X intercept can change under different regression models and, as can be seen by comparison with the lowess line in Figure 3.5, the estimate of tMRCA under a linear model is likely artificially recent. Thus inference of tMRCA from Figure 3.5 is cautioned against since a linear model does not approximate our data well.

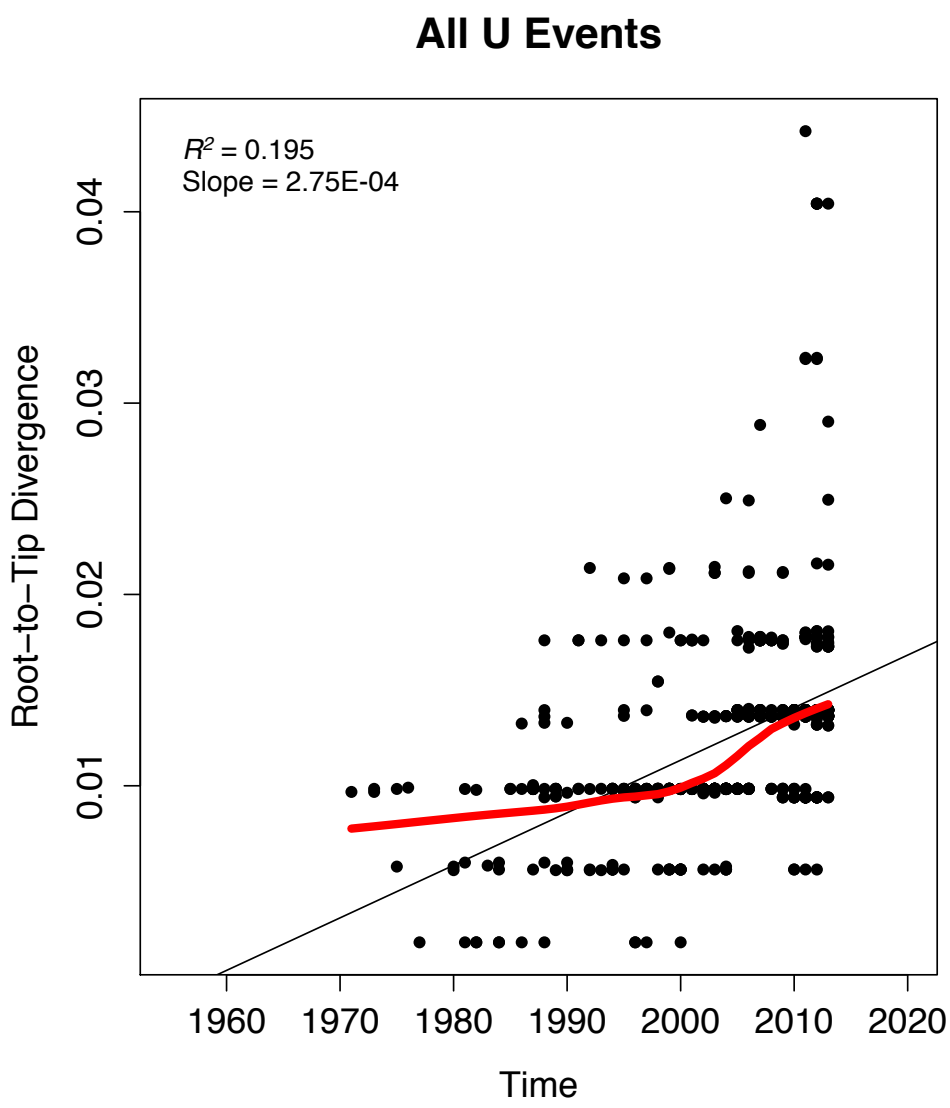


Figure 3.5: Plot of root-to-tip divergence versus sampling time for all U genogroup events data covering 42 years. The correlation coefficient indicates the amount of variation in root-to-tip divergence that is explained by time and equals 0.441, R^2 =0.195, and the slope of the line is 2.75E-04, representing the rate of evolution. The X-intercept is 1958.7. A lowess line is indicated in red, showing the non-linearity of the data.

We also investigated how divergence rates differed between the two major U subgroups, UP and UC. The divergence rate of UC events was 1.58E-04 nucleotide mutations per site per year, slightly lower than the divergence rate observed for all U events together events (Figure 3.6). While still a weak relationship, the divergence rate of UC events was slightly better approximated by the linear model (R^2 = 0.254) than the divergence rate of all U events. For UC events, the linear relationship between genetic distance and year of isolation is significantly different than zero ($p<0.0001$). In contrast, analysis of UP genotypes demonstrated a negligible rate of divergence (R^2 = 0.000318) representing a linear relationship between genetic distance and year of isolation that was not significantly different from 0 ($p=0.832$) (Figure 3.6).

As described above with all U events, the X-intercepts that allow the inference of tMRCA should be interpreted extremely cautiously since the fit of the linear regression lines is exceptionally poor. For instance, given that the slope of the linear regression line for UP events is in fact negative, inference of tMRCA yields a date in the future, in year 3710 (X-intercept = 3709.85). For UC events, the X-intercept under a linear model equals 1974.1. While a more plausible root age, we continue to caution against acceptance of this as an accurate inference of root age given the relatively poor fit of the linear model to the observed data.

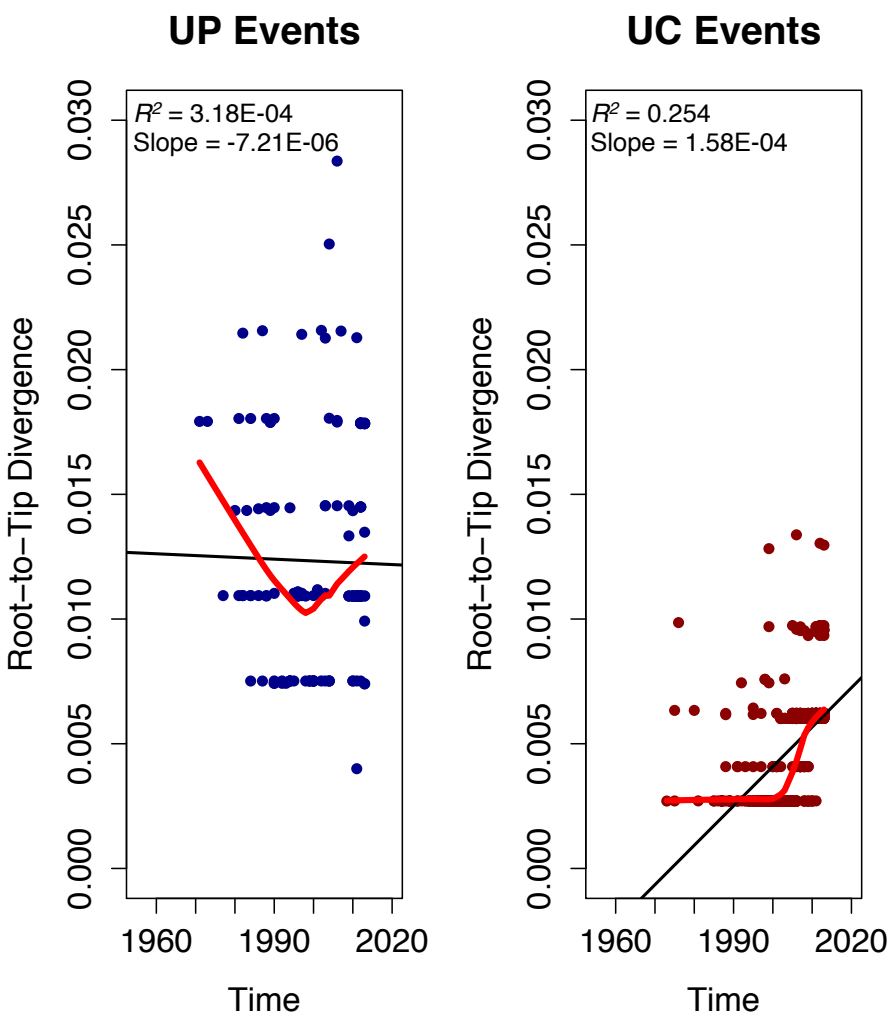


Figure 3.6: Plot of root-to-tip divergence versus sampling time for all UC genotype events and UP genotype events. Data covers 42 years for UP events and 40 years for UC events. For UP events the correlation coefficient equals -0.0178, R^2 = 0.000318, and the slope of the line is -7.21E-06, representing a negligible rate of evolution. For UC events the correlation coefficient equals 0.504, R^2 = 0.254, and the slope of the line is 1.58E-04. Linear regression lines are indicated in black, and lowess lines are indicated in red.

Given the poor fit of the linear model, we also explored a non-linear regression model to see if model fit would be improved. For all U events the data appeared to possibly follow an exponential distribution. To test whether an exponential function might fit the data better, we log-transformed the values of root-to-tip divergence and performed linear regression on the log-transformed data. For all U events linear regression on the log-transformed data indicated an R^2 of 0.2716. While the regression line fits the log-transformed better than the untransformed data (R^2 equaled 0.195 for the untransformed data), the low R^2 continues to indicate that a different or perhaps higher order function is necessary to model the evolutionary rate of the U events data. Again, we stress that the X-intercept should not be interpreted as an accurate estimate of the true age of the tree root until a function better fits the data.

3.7 Geographic characteristics of UC and UP

We initially hypothesized that there should be some amount of phylogenetic substructure within the U genogroup since we observed geographic structure in the detection of different genotypes of U genogroup IHNV. When mapped according to their subgroup, UC genotype viruses tended to occur predominantly in the Columbia River Basin and UP genotype viruses were detected primarily in coastal watersheds (Figure 3.7). While events due to UP genotypes did occur within the Columbia River Basin, and vice versa with UC events in coastal watersheds, these geographic exceptions occur relatively less frequently than detections within the typical geographic range. The principle exception to this rule is the detection of UC genotype viruses on the Oregon coast. While the Oregon coast has only had UC events, these represent a very small number of coastal IHNV events and thus the general rule that coastal detections are predominantly of UP still holds (see circle scaling in figure 3.7).

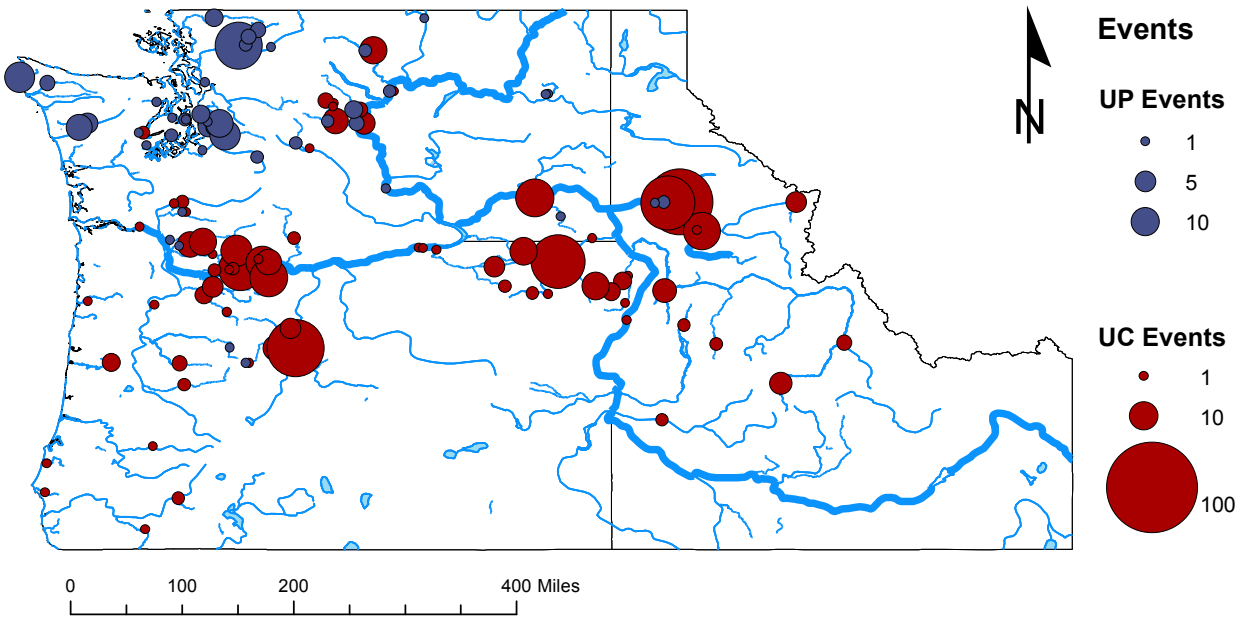


Figure 3.7: Geographic distribution of UC and UP subgroup events over the study area between 1971 and 2013. Circles are scaled to represent the number of detections of either UC or UP genotypes that occurred at sampling sites.

Within the phylogeny we also characterized genotypes according to their primary geography of detection. While genotypes may have been detected in at least one event in both the Columbia River Basin and in a coastal watershed, branches are colored to demonstrate the principal geography of detection, not the sole geography of detection (Figure 3.8). All UC genotypes except for two, mG199U and mG135U, have been detected primarily in the Columbia River Basin. Both of these genotypes have only been detected during one event, and therefore their detection in coastal watersheds has not been observably sustained.

While UP genotypes are largely detected in coastal watersheds, a relatively greater number of UP viruses have been detected primarily in the Columbia River Basin than UC genotypes in coastal watersheds (Figure 3.8). Eleven different UP genotypes were detected

predominantly in the Columbia River Basin. Nine of these eleven UP genotypes were only ever isolated from a single event, and therefore their detection in the Columbia River Basin was neither geographically nor temporally broad. One of the UP types, mG274U was detected at two events in the Columbia River Basin, and at no events within coastal watersheds. Finally, mG028U is a UP genotype virus that was detected at 2 events in the Columbia River Basin and at 2 events in coastal watersheds.

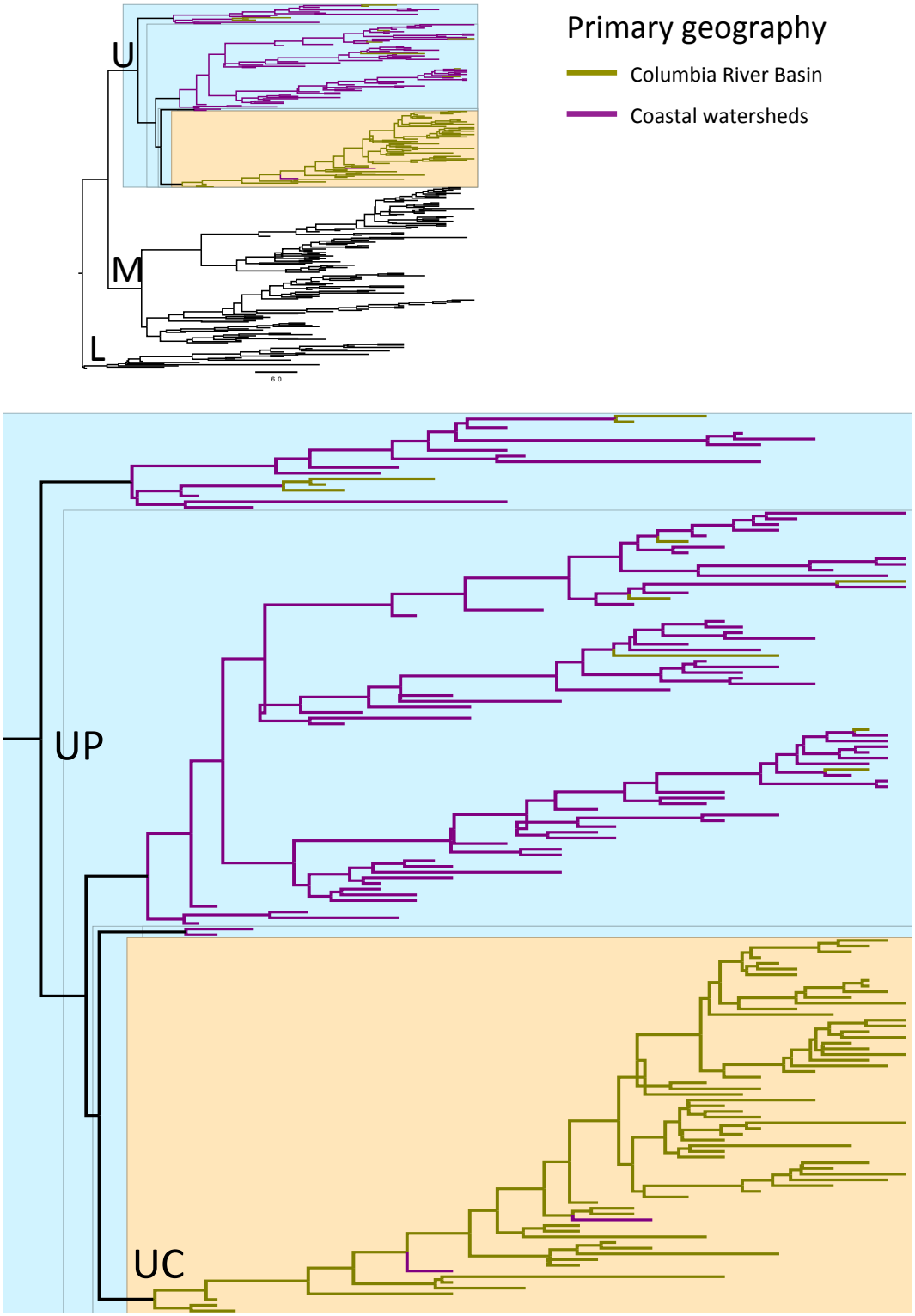


Figure 3.8: Bayesian coalescent phylogeny of U genogroup IHNV. Branches are colored to reflect the geography where that sequence type is primarily detected. This however does not mean that the genotype is solely detected in that geography.

3.8 HOST SPECIFICITY OF UC AND UP SUBGROUPS

To assess whether the phylogeny was structured by host species, we colored branches based on the host species that a genotype infects the majority of the time (Figure 3.9). If the phylogeny is structured by host species then branch colors will be consistent within lineages, and if the phylogeny is not structured by host species, then branch colors will be intermingled within the phylogeny.

The greatest host structuring within the phylogeny was the detection of UP genotype viruses primarily in *O. nerka* and isolation of UC genotype viruses primarily from *O tshawytscha* (Figure 3.9). Within the UP lineage, 15 out of 92 UP genotypes were detected primarily in species other than *O. nerka*. Eight UP genotypes were isolated primarily from *O. tshawytscha*, however these genotypes were only ever detected between one and three times. One UP genotype was detected primarily in *O. mykiss* (mG050UmG261U, detected only once), and one was isolated once from *O. nerka* and twice from *O. tshawytscha* (mG250U). Four UP genotypes were detected predominantly as single isolates in host species that demonstrate minimal IHNV infection, either chum salmon or coho salmon (Figure 3.9).

Within the UC lineage, the majority of genotypes infect primarily *O. tshawytscha*. There were 23 UC genotypes that were primarily isolated from different host species. Fifteen of those UC genotypes were detected primarily in *O. mykiss* (representing 1 to 2 isolates) and 2 UC genoytpes were detected as single isolates in *O. nerka* (mG194U and mG196U). Two genotypes were detected as single isolates once in *O. tshawytscha* and once *O. mykiss* (mG174UmG242U and mG193U). Three genotypes were detected equally in *O. nerka* and *O. mykiss* (mG032U, mG033U, and mG057U). Both mG033U and mG057U were only detected once in each species, however mG032U was detected 11 times in steelhead trout and 11 times in kokanee salmon.

Only one UC genotype (mG185U) was detected primarily in a non-dominant species (in this case, in coho salmon).

Given the abundance of steelhead trout within the Columbia River Basin, we postulated that UC subgroup IHNV, which is found primarily in the Columbia River Basin, could potentially evolve a host tropism to *O. mykiss* separately from the evolution of a host tropism to Chinook salmon. Within the phylogeny, such an event would likely appear as a cluster of genotypes that are found predominantly in steelhead trout. From our analysis the UC lineage did not demonstrate clusters of genotypes found predominantly in *O. mykiss*, therefore an adaptation of UC subgroup IHNV to steelhead trout was not supported.

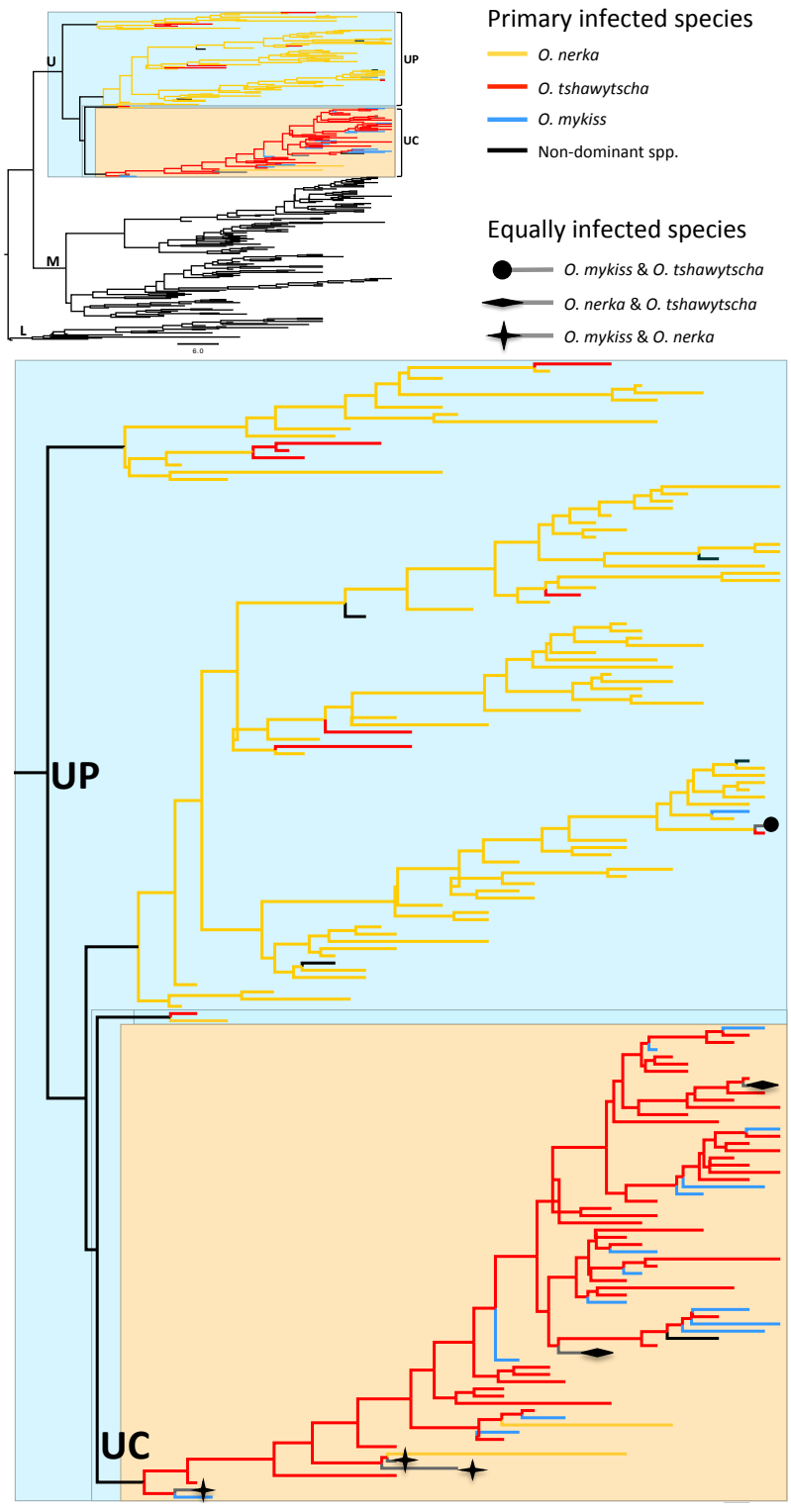


Figure 3.9: Bayesian coalescent phylogeny of U genogroup IHNV. Branches are colored to reflect the fish species from which that genotype is primarily isolates. This however does not necessarily mean that the genotype has been solely isolated from that fish species alone.

3.9 Overlap of geographic range exceptions and host specificities

Analysis of geographic exceptions allows us to explore whether viral genotypes appear specific to a host species simply because that is the predominant host in a genotype's geographic range or whether a genotype is demonstrating a tropism towards a specific host. Of the 11 UP genotypes that were detected primarily in the Columbia River Basin, 6 were detected in the two primary host species of the Columbia River Basin. Five were detected primarily as single isolates in *O. tshawytscha* (mG028U, mG029U, mG030U, mG136U and mG159U), and one was detected as a single isolate in *O. mykiss* (mG050UmG261U). Three genotypes were detected as one or two isolates in sockeye salmon (mG040U, mG127U, and mG274U) and one genotype was detected as a single isolate in kokanee salmon (mG197U). Finally, mG249U was detected as a single isolate in coho salmon in the Columbia River Basin. In contrast, both of the UC genotypes that were detected primarily outside of the Columbia River Basin were detected as single isolates in *O. tshawytscha*.

Chapter 4. DESCRIPTIVE EPIDEMIOLOGY OF U GENOGROUP IHNV BETWEEN 1971 AND 2013

4.1 INTRODUCTION

Descriptive epidemiological studies provide an initial picture of how a disease is distributed by person, place, and time, revealing important questions that are later explored analytically (Koepsell & Weiss, 2003). While genetic typing of isolates has been ongoing, and epidemiological analysis of U genogroup IHNV in the Columbia River Basin (Kyle A Garver, Troyer, & Kurath, 2003), Alaska, and coastal Washington has been performed (Emmenegger, Meyers, Burton, & Kurath, 2000; Emmenegger & Kurath, 2002), these studies have focused on smaller regional areas and are now 10-15 years old. To update our current knowledge of IHNV in the Pacific Northwest, we described the distribution of U genogroup IHNV temporally, geographically, and by host species for all of Washington, Oregon, and Idaho.

IHNV-positive field isolates isolated between January 1st, 1971 and December 31st, 2013 within Washington, Oregon, and Idaho were included in this study. We report results both by numbers of isolates and by numbers of events. As a passive surveillance program, we receive IHNV-positive viral field isolates from collaborating federal, state, and tribal fish health agencies. Thus analysis of results by isolate numbers simply reports data for all isolates obtained. As is typical with passive surveillance programs, sampling densities are heterogeneous and affected by differences in the frequency and quality of agency reporting. Additionally, discrete disease detections may be sampled heterogeneously, resulting in highly variable numbers of isolates per disease event. To control for this second source of bias, the dataset was coded by

IHNV events, as originally described by Breyta et al. (2013) and also discussed in Chapter 2 (Methods).

For descriptions of the dataset by isolates, all genotyped isolates of U genogroup IHNV detected between 1971 and 2013 within Washington, Oregon, and Idaho were used (n=1216). In contrast, describing the datasets by events yielded a smaller, de-biased dataset (n=619). However, while the events dataset was de-biased in terms of overrepresentation of certain events due to denser sampling, certain sampling biases still exist within the events data. For instance, participating fish health agencies may be more inclined to submit samples that seem novel, either detected in a new location or at a different point in time or in a fish population that is somehow unusual or atypical.

In addition to total virus detection events we also considered disease events. IHNV-positive fish may be either asymptomatic or symptomatic, with symptom manifestation generally only occurring in juvenile fish (Bootland & Leong, 1999). While standardized IHNV diagnostic procedures dictate that 60 to 100 spawning adult fish are screened for IHNV upon return to the hatchery (Thoesen, 1994), juveniles generally are not tested unless IHNV infection is suspected, usually due to presentation of symptomatic IHN disease. Thus, as a proxy for establishing counts of IHNV infection resulting in IHN disease, we considered all events in juvenile fish (younger than yearlings) to be disease events. Despite being young fish, we excluded yearlings because they represent a different fish cohort in two important ways. Firstly, yearling salmonids have better developed immune systems than juveniles (Bootland & Leong, 1999), and secondly, juveniles and yearlings tend to be housed separately at hatcheries, and may not be present at the hatchery at the same time.

This analysis described IHNV isolates and events for 114 different sampling sites from Washington, Oregon, and Idaho (Figure 4.1). When describing geographic range we considered a binary definition; sampling sites either occurred within the Columbia River Basin or along coastal watersheds, namely watersheds that drain directly into Puget Sound, the Salish Sea, or the Pacific Ocean. Within the Columbia River Basin there are 77 sampling sites, and there are 37 sampling sites along coastal watersheds.

In total, we analyzed a dataset containing 1216 U genogroup viral field isolates representing 619 U genogroup IHNV events between 1971 and 2013 within the three-state study area. As illustrated in Figure 4.2, the distribution of isolates is left-skewed with a rise in the number of isolates collected occurring in the mid 1990s and again sharply after 2010. When the dataset was analyzed by events the distribution remained left-skewed, however the rise in events after 2010 did not appear as sharp as when analyzed by isolates.

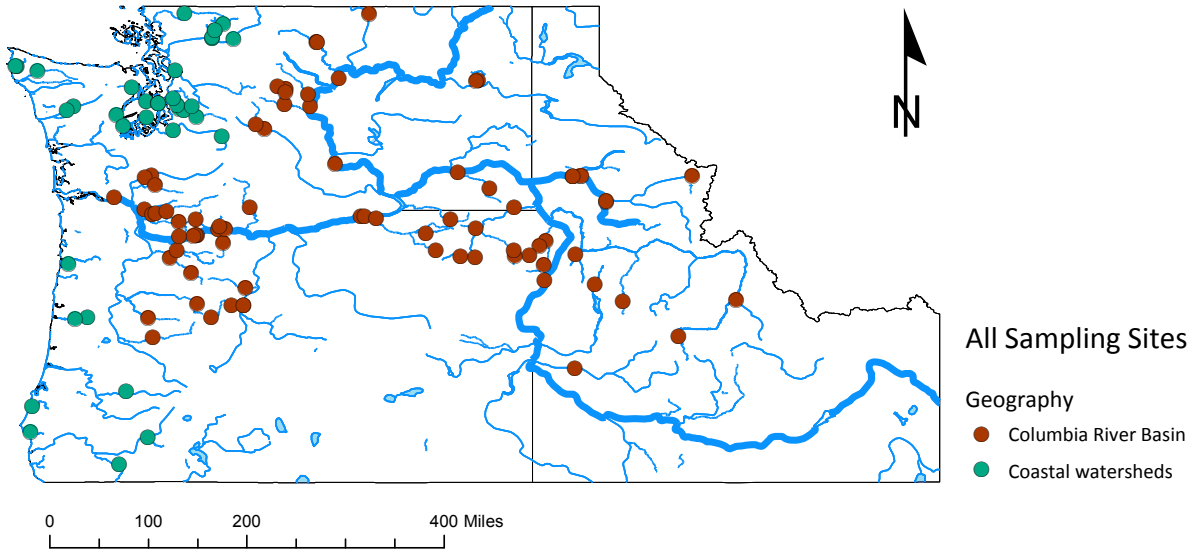


Figure 4.1: Illustration of geographic designations of sampling sites. We considered a binary geographic designation, in which sites were on watersheds that drain directly into coastal waters, or sites were on the Columbia River or one of its tributaries (within the Columbia River Basin).

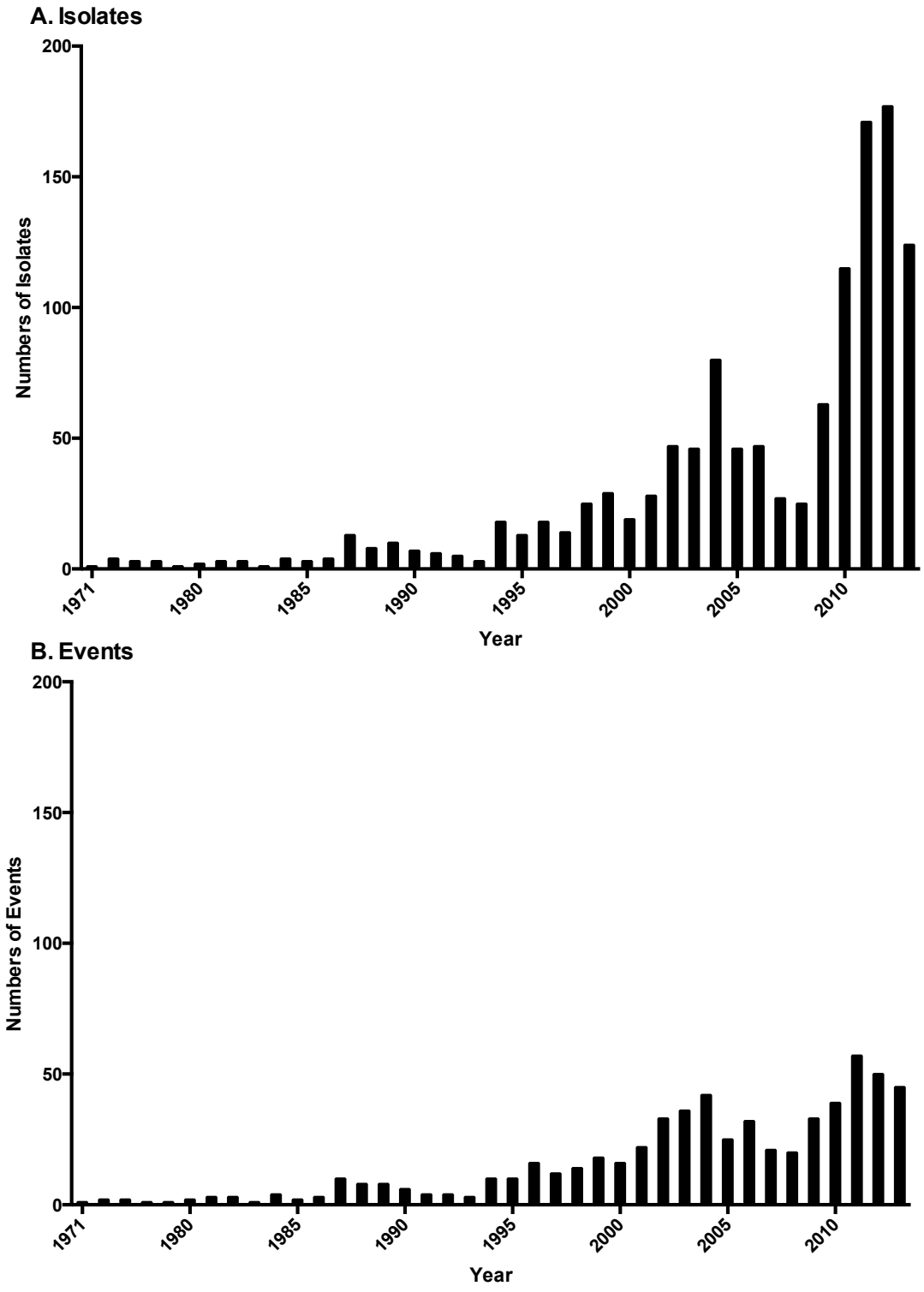


Figure 4.2: Temporal distribution of U genogroup IHNV detections. **A)** Distribution of all U genogroup *isolates* that were genotyped between 1971 and 2013 (n=1216). **B)** Distribution of U genogroup *events* that occurred between 1971 and 2013 (n=619).

4.2 IHNV ISOLATE AND EVENT DISTRIBUTION BY GEOGRAPHY

Of the 1216 submitted isolates, 998 (82.1%) came from sampling sites in the Columbia River Basin (Table 4.1). While the number of events these isolates represent was somewhat lower, the majority of our recorded events (78.8%) occurred at sites within the Columbia River Basin as well (Table 4.2).

Notably, detections of the two phylogenetic U subgroups, UC and UP, correlated with our binary geographic designation. Looking at IHNV events in Table 4.2, 96.8% of all UC events (n=475) occurred at sites within the Columbia River Basin, and only 3.2% of UC events occurred in coastal watersheds. The difference was not quite as marked for UP events; nevertheless of all events caused by a UP subtype virus (n=144), 80.6% of the events occurred at sites along coastal watersheds (Table 4.2).

Table 4.1: Description of IHNV U genogroup *isolates* in Washington, Oregon, and Idaho between 1971 and 2013 by geographic location of detection, by host species, and by age of fish. Tallies of isolates from yearlings and fry were grouped with juveniles.

Timeline: 1971-2013	**All U isolates (n=1216)**		**UC isolates (n=971)**		**UP isolates (n=245)**	
GEOGRAPHY						
Coastal watersheds (37 sites)	218	(17.9%)	18	(1.9%)	200	(81.6%)
Columbia River Basin (77 sites)	998	(82.1%)	953	(98.1%)	45	(18.4%)
HOST						
Sockeye Salmon	177	(14.6%)	16	(1.6%)	161	(65.7%)
Kokanee Salmon	38	(3.1%)	26	(2.7%)	12	(4.9%)
Chinook Salmon	654	(53.8%)	630	(64.9%)	24	(9.8%)
Steelhead Trout	287	(23.6%)	271	(27.9%)	16	(6.5%)
Rainbow Trout	12	(1.0%)	8	(0.8%)	4	(1.6%)
Coho Salmon	24	(2.0%)	16	(1.6%)	8	(3.3%)
Chum Salmon	16	(1.3%)	4	(0.4%)	12	(4.9%)
Atlantic Salmon	8	(0.7%)	N/A		8	(3.3%)
LIFESTAGE						
Juveniles	248	(20.4%)	184	(18.9%)	64	(26.1%)
Adults	958	(78.8%)	783	(80.6%)	175	(71.4%)
Unknown Age	10	(0.8%)	4	(0.4%)	6	(2.4%)

Table 4.2: Description of IHNV U genogroup *events* in Washington, Oregon, and Idaho between 1971 and 2013 by geographic location of detection, by host species, and by age of fish. Tallies of isolates from yearlings and fry were grouped with juveniles.

Timeline: 1971-2013	**All U Events (n=619)**		**UC Events (n=475)**		**UP Events (n=144)**	
GEOGRAPHY						
Coastal watersheds (37 sites)	131	(21.2%)	15	(3.2%)	116	(80.6%)
Columbia River Basin (77 sites)	488	(78.8%)	460	(96.8%)	28	(19.4%)
HOST						
Sockeye Salmon	92	(14.9%)	9	(1.9%)	83	(57.6%)
Kokanee Salmon	31	(5.0%)	22	(4.6%)	9	(6.3%)
Chinook Salmon	295	(47.7%)	282	(59.4%)	13	(9.0%)
Steelhead Trout	153	(24.7%)	141	(29.7%)	12	(8.3%)
Rainbow Trout	10	(1.6%)	6	(1.3%)	4	(2.8%)
Coho Salmon	20	(3.2%)	12	(2.5%)	8	(5.6%)
Chum Salmon	14	(2.3%)	3	(0.6%)	11	(7.6%)
Atlantic Salmon	4	(0.6%)	N/A		4	(2.8%)
LIFESTAGE						
Juveniles	119	(19.2%)	86	(18.1%)	33	(22.9%)
Adults	491	(79.3%)	386	(81.3%)	105	(72.9%)
Unknown Age	9	(1.5%)	3	(0.6%)	6	(4.2%)

4.3 IHNV ISOLATE AND EVENT DISTRIBUTION BY GEOGRAPHY OVER TIME

As described previously, we considered two geographic ranges, the Columbia River Basin and coastal watersheds. Looking at the densities of isolate and event occurrence over time, both the numbers of sequenced isolates received and the number of discrete events rose slightly in the mid 1990s and then rose markedly after 2000 (Figure 4.3A and B). Far greater numbers of isolates came from locations within the Columbia River Basin, and this trend held when the data were corrected to represent the number of IHNV events that occurred during the study period (Figure 4.3B and Figure 4.4). Notably, the corrected event data still showed an increase in sampling density as time progressed during the study period, an increase that appeared to be driven largely by detections in the Columbia River Basin (Figure 4.4).

The distribution of isolates and events in the Columbia River Basin appeared to be bimodal, with one peak centered around 2004 and another peak centered around 2011 (Figure 4.3A and B). In contrast, the distribution of isolates and events from sites along coastal watersheds illustrated a relatively constant density until 2011, after which the numbers of isolates and events rose steeply (Figure 4.3A and B).

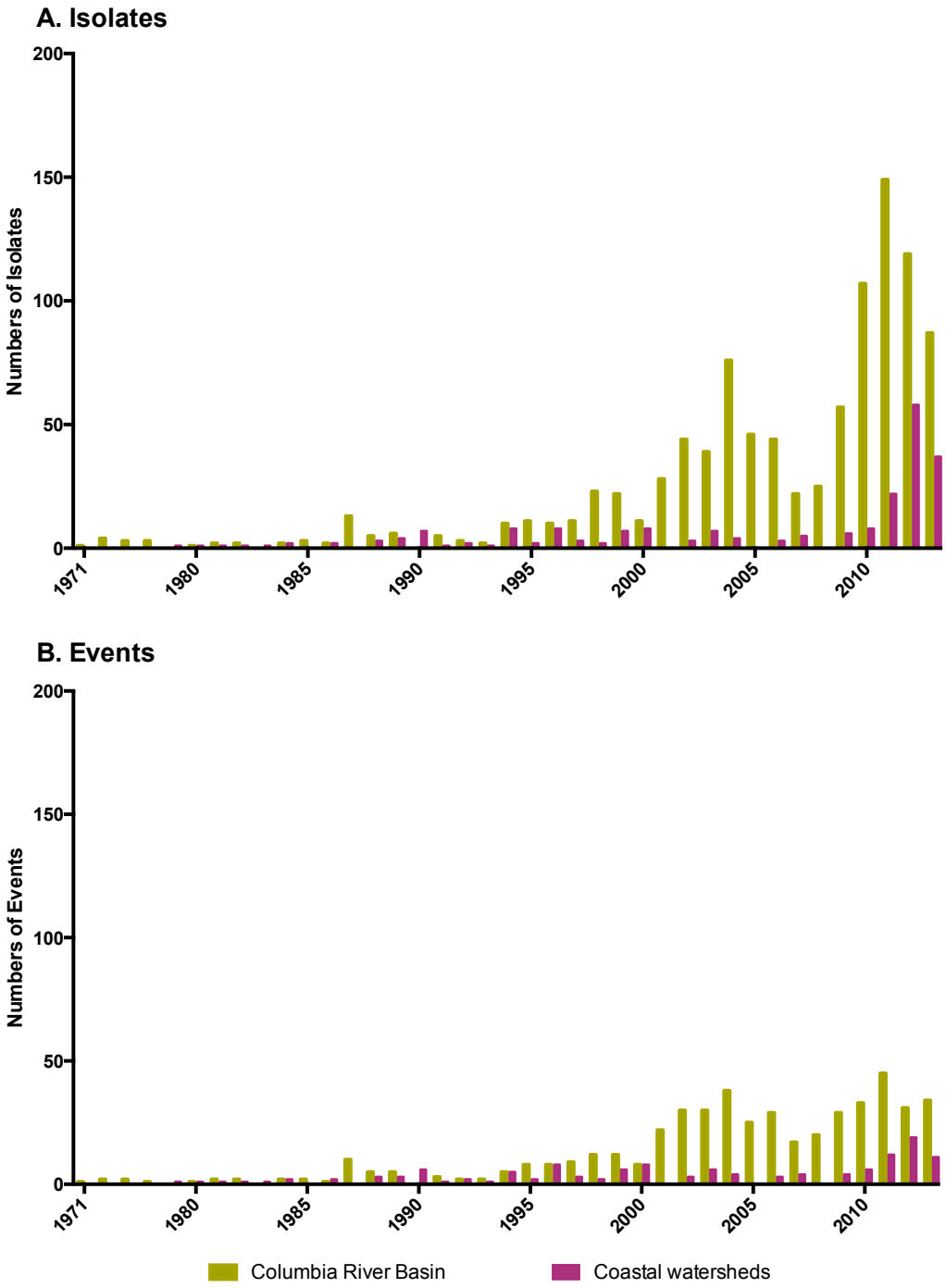


Figure 4.3: **A)** Distribution of U genogroup IHNV *isolates* that were genotyped between 1971 and 2013, coded by the geographic range in which they were detected (either within the Columbia River Basin or in coastal watersheds) (n=1216). **B)** Distribution of U genogroup IHNV *events* that occurred between 1971 and 2013, either in the Columbia River Basin or in coastal watersheds (n=619).

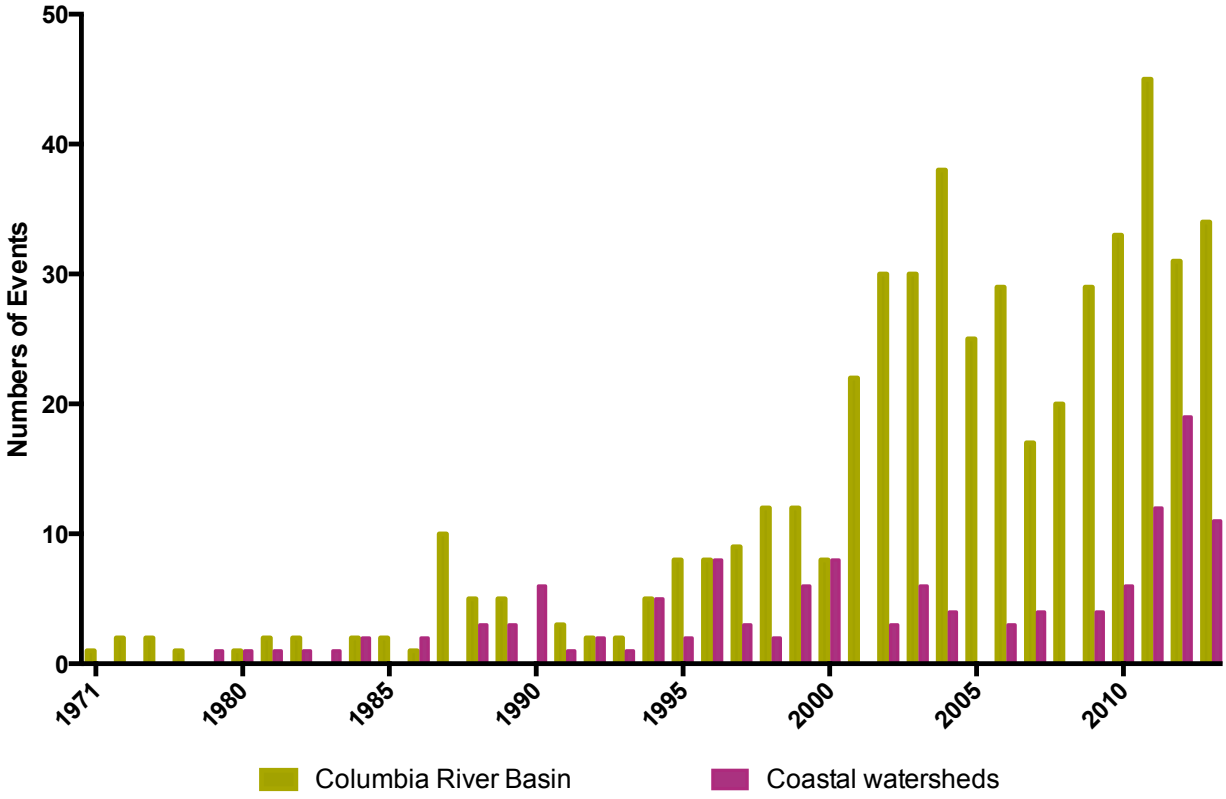


Figure 4.4: Frequency of U genogroup IHNV events from 1971 to 2013 in the Columbia River Basin and in coastal watersheds. This figure displays the same data as in Figure 4.2B, however the axes are scaled more appropriately to the data to facilitate interpretation. Events are considered unique if the detection of U genogroup IHNV occurs at a different site, occurs in a different year, has a different sequence type, or occurs in a different fish cohort (see further explanations in text). There are 619 events in total.

4.4 IHNV ISOLATE AND EVENT DISTRIBUTION BY HOST SPECIES

During the study time period, IHNV was detected in most species of Pacific salmonids, although events were most abundant in sockeye salmon (anadromous *O. nerka*), Chinook salmon (*O. tshawytscha*), and in steelhead trout (anadromous *O. mykiss*) (Table 4.2). When considering all U genogroup events, events were most numerous in Chinook salmon (n=295) followed by steelhead trout (n=153) and sockeye salmon (n=95).

Different host species were generally infected by different U subgroups. UP subgroup events occurred 57.6% of the time in sockeye salmon (Table 4.2). Although UP viruses were detected in 7 other salmonid species, fewer than 10% of UP detections occurred in each of these other hosts, indicating that there was no secondary dominant host for UP subgroup viruses. In contrast, UC subgroup events were generally detected in Chinook salmon (59.4% of events) or in steelhead trout (29.7% of events). Less than 5% of UC events were detected in each of the other host species, therefore no tertiary dominant host appeared to exist for UC subgroup viruses (Table 4.2).

Additionally, different host species showed different rates of IHNV detection when stratified by geography of detection. Overall the majority of IHNV events occurred in Chinook salmon (n=295) and in *O. mykiss* (n=163). However only 2.7% of Chinook salmon events and 11.0% of steelhead trout and rainbow trout (*O. mykiss*) events occurred along coastal watersheds (Table 4.3). In contrast, of the 123 events that occurred in *O. nerka* (either sockeye salmon or kokanee salmon), 68.3% of detections occurred at sites in coastal watersheds (Table 4.3).

Table 4.3: Numbers of events and isolates of U genogroup IHNV in *O. nerka*, *O. tshawytscha*, and *O. mykiss* stratified on geography of detection. We considered two geographic partitions only, the Columbia River Basin and coastal watersheds. Both sockeye salmon and steelhead trout have non-anadromous counterparts, kokanee salmon and rainbow trout respectively. Kokanee salmon and sockeye salmon are the same species, as are steelhead and rainbow trout, however kokanee salmon and rainbow trout do not migrate out to sea.

***O. nerka* (Sockeye salmon and Kokanee)**	**Events (n=123)**	**Isolates (n=215)**
Columbia River Basin	39 (31.7%)	57 (26.5%)
Coastal watersheds	84 (68.3%)	158 (73.5%)
***O. tshawytscha* (Chinook salmon)**	**Events (n=295)**	**Isolates (n=654)**
Columbia River Basin	287 (97.3%)	645 (98.6%)
Coastal watersheds	8 (2.7%)	9 (1.4%)
***O. mykiss* (Steelhead and Rainbow trout)**	**Events (n=163)**	**Isolates (n=299)**
Columbia River Basin	145 (89.0%)	274 (91.6%)
Coastal watersheds	18 (11.0%)	25 (8.4%)

Table 4.4: Events of U genogroup IHNV in the Columbia River Basin (CRB) from 1971 to 2013, and the number of isolates that were genotyped from those events, stratified by host species. Where a species has two life history types but the common name is different, the Latin name of the species is indicated.

Host Species	**CRB Events (n=488)**		**CRB Isolates (n=998)**	
Chinook Salmon	287	(58.8%)	645	(64.6%)
Kokanee (*O. nerka*)	25	(5.1%)	29	(2.9%)
Sockeye Salmon (*O. nerka*)	14	(2.9%)	28	(2.8%)
Steelhead Trout (*O. mykiss*)	138	(28.3%)	265	(26.6%)
Rainbow Trout (*O. mykiss*)	7	(1.4%)	9	(0.9%)
Coho Salmon	14	(2.9%)	18	(0.9%)
Chum Salmon	3	(0.6%)	4	(0.4%)

Table 4.5: Events of U genogroup IHNV in coastal watersheds from 1971 to 2013, and the number of isolates that were genotyped from those events, stratified by host species. Where species are genetically identical but the common name is different, the Latin name of the species is indicated.

Host Species	**Coastal Events (n=131)**		**Coastal Isolates (n=218)**	
Chinook Salmon	8	(6.1%)	9	(4.1%)
Sockeye Salmon (*O. nerka*)	78	(59.5%)	149	(68.3%)
Kokanee (*O. Nerka*)	6	(4.6%)	9	(4.1%)
Steelhead Trout (*O. mykiss*)	15	(11.5%)	22	(10.1%)
Rainbow Trout (*O. mykiss*)	3	(2.3%)	3	(1.4%)
Coho Salmon	6	(4.6%)	6	(2.8%)
Chum Salmon	11	(8.4%)	12	(5.5%)
Atlantic Salmon	4	(3.1%)	8	(3.7%)

Exploring temporal relationships within the dataset, the numbers of isolates obtained from the three dominant host species (*O. tshawytscha*, *O.mykiss*, and *O. nerka*) began to rise in the mid-1990s, and then rose markedly in the early 2000s. While not as pronounced as the rise in the number of isolates, the numbers of observed events also increased according to this same temporal pattern (Figure 4.5A and B). While there were fewer numbers of events than numbers of isolates, the relative proportion of detections in each of the three hosts remained similar. The greatest number of isolates and events occurred in *O. tshawytscha*, followed by *O. mykiss* and *O. nerka* (Figure 4.5).

The distribution of events (Figure 4.5B and 4.6) indicated that the numbers of events that occurred in *O. tshawytscha* prior to the early 2000s was relatively similar to the numbers of events that occurred in the two other primary host species for IHNV except for a peak in Chinook events in 1987, 1988, and 1989. However, after 2000 the numbers of events in Chinook outpaced the numbers of events in *O. nerka* and in *O.mykiss*. Excluding the early peak from 1987-1989, the distribution of events in Chinook appeared roughly bimodal with a peak from 2001 – 2006 (centered at 2004) and a peak from 2009 – 2013 (centered at 2011). This bimodality did not appear to occur in *O. nerka* events nor in *O. mykiss* events (Figure 4.5B and 4.6).

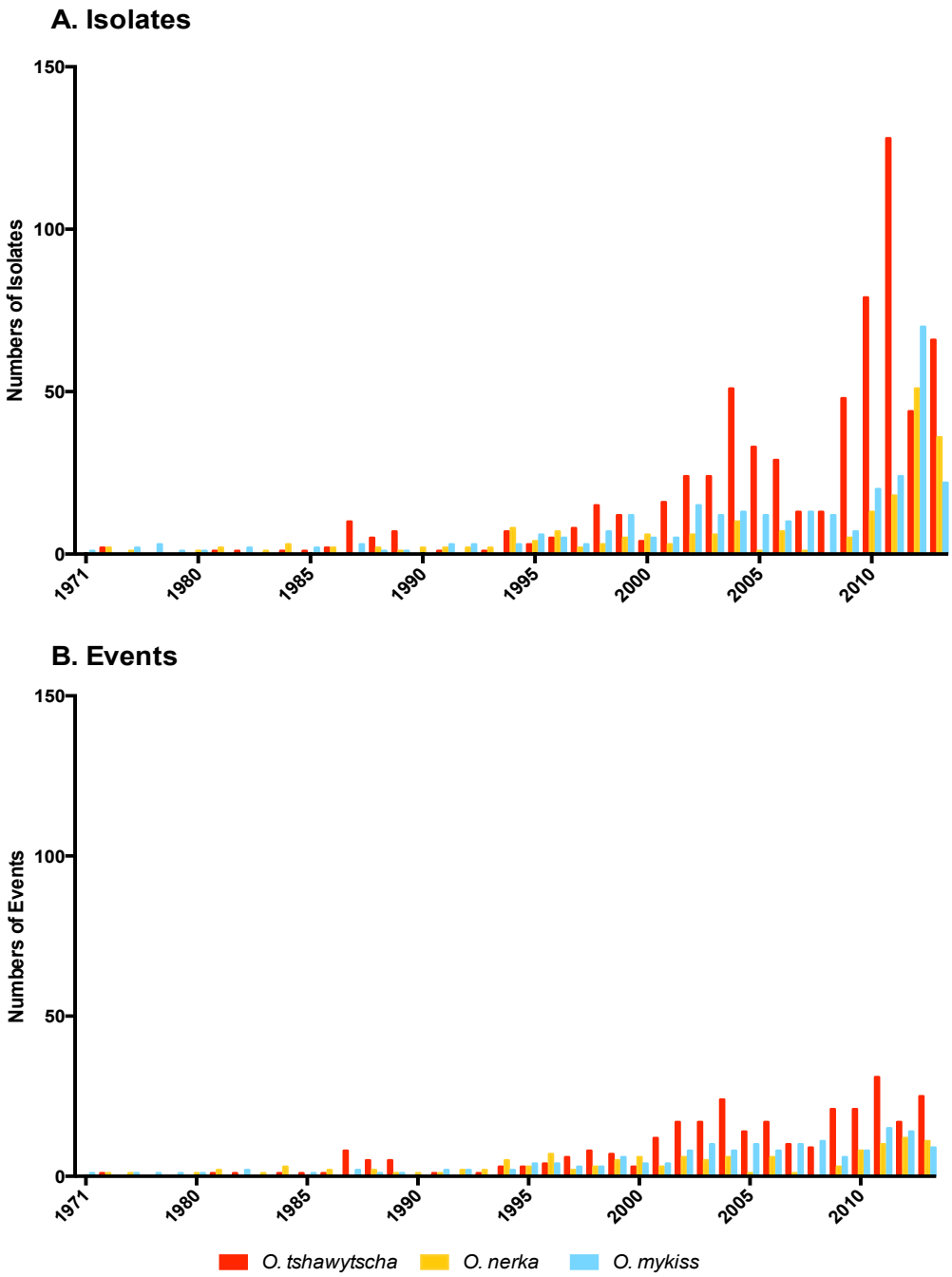


Figure 4.5: A) Distribution of U genogroup IHNV isolates that were genotyped between 1971 and 2013, coded by host species (n=1168). Only the three dominant host species are included. B) Distribution of U genogroup IHNV events that occurred between 1971 and 2013, coded by host species for the three dominant host species (n=581). Event coding seeks to normalize the numbers of isolates that are sequenced from each detection event (see description in text).

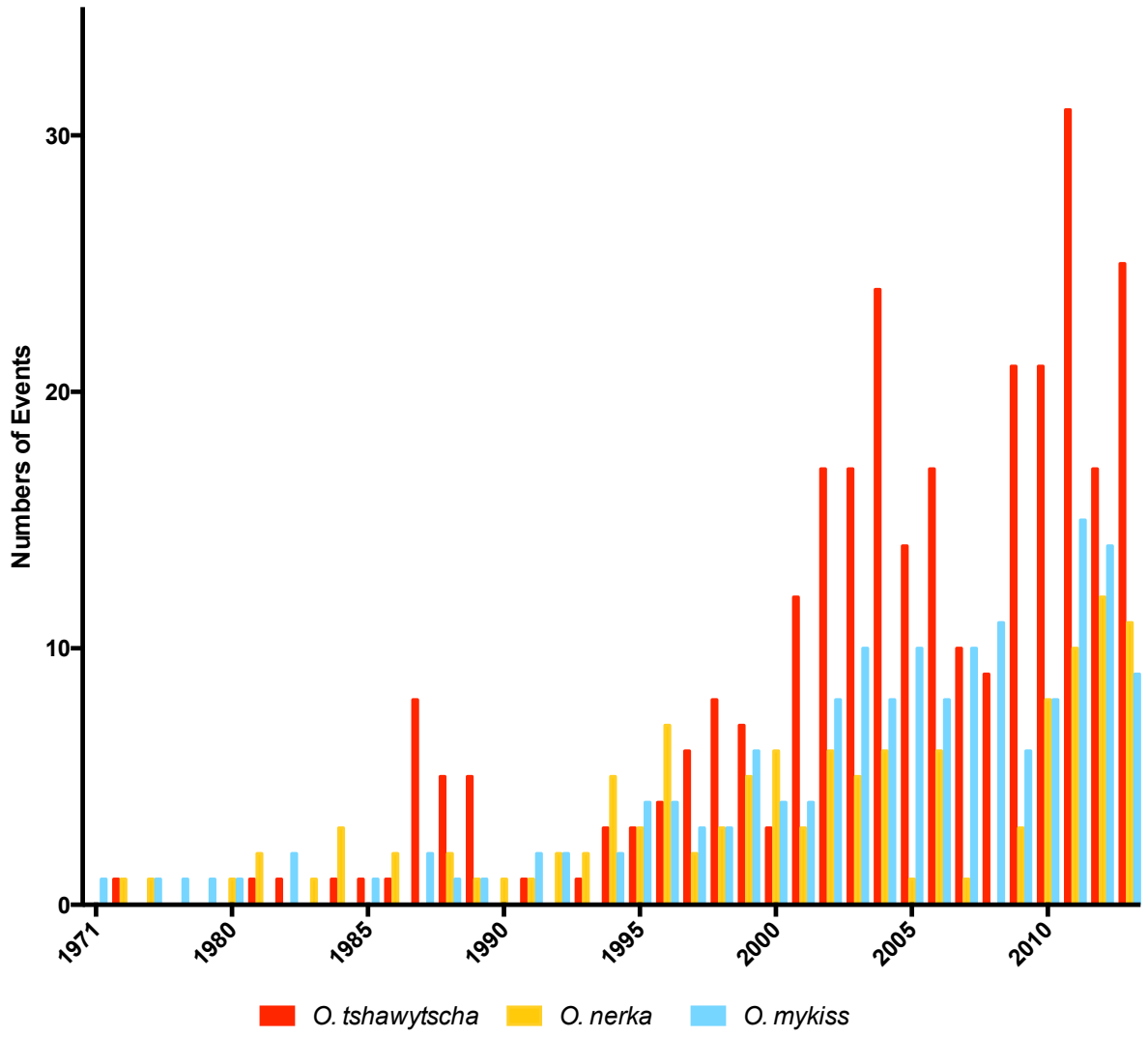


Figure 4.6: Distribution of U genogroup IHNV events between 1971 and 2013 in the three dominant host species (n=581). This figure displays the same data as in Figure 4.5B, however the axes are scaled more appropriately to the data to facilitate interpretation. Events are considered unique if the detection of U genogroup IHNV occurs at a different site, occurs in a different year, has a different sequence type, or occurs in a different fish cohort (see further explanations in Chapter 2).

4.5 EXCEPTIONS TO GEOGRAPHIC RANGE OF UC AND UP SUBTYPES

While the dominant pattern in IHNV events was the aforementioned detection of UC subtypes in the Columbia River Basin and UP subtypes in coastal watersheds, there were exceptions to this rule, with UC subtypes occasionally detected in coastal watersheds, and UP subtypes detected in the Columbia River Basin. Interestingly, there was some degree of spatial trend among the geographic exceptions. When UC subtypes were detected in coastal watersheds, they were often detected in Oregon's coastal watersheds south of the drainage of the Columbia River Basin to the Pacific Ocean (Figure 4.7). When UP subtypes were detected within the Columbia River Basin, they were often found fairly far into the river system in Central and Eastern Washington and Northern Idaho (Figure 4.7).

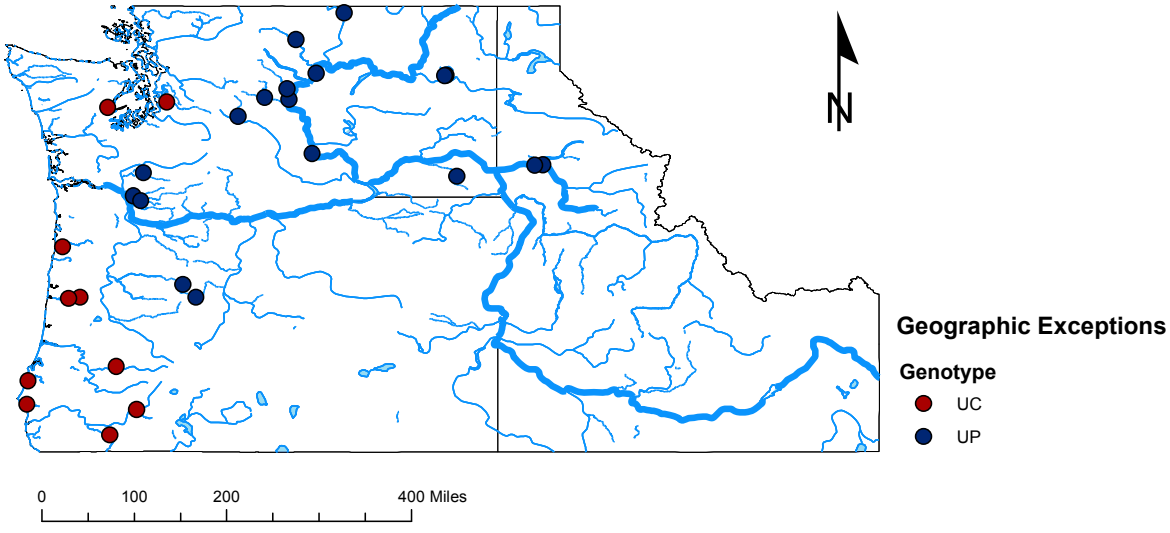


Figure 4.7: Locations of geographic exception events caused UC subtypes along coastal watersheds (n=15) and UP subtypes within the Columbia River Basin (n=28). Certain sampling sites have had more than one event of a geographic exception, however since these numbers are generally quite low, circles are not scaled to numbers of events.

When detected in coastal watersheds, UC subgroup viruses still appeared to have some specificity for Chinook salmon (*O. tshawytscha*) and steelhead and rainbow trout (*O. mykiss*). Of the 15 events where UC genotypes were detected in coastal watersheds, 14 events (93.3%) occurred in Chinook salmon or *O. mykiss*. While UC viruses still appeared to infect the same hosts as they would when detected in their typical geography, this trend was not as strong for UP subtypes. Of the 28 events where UP subtypes were detected within the Columbia River Basin, only 9 events (32.1%) occurred in sockeye salmon or kokanee salmon (both *O. nerka*). Rather, UP events in the Columbia River Basin appeared to occur in the dominant host species of the Columbia River Basin, with 17 of the 28 events (60.7%) of UP geographic exceptions occurring in either Chinook salmon or *O. mykiss* (Table 4.6).

Table 4.6: Description of geographical exceptions in UP and UC subgroup ranges by host species of fish.

Host Species	**UP events in Columbia River Basin (n=28)**	
Chinook salmon (*O. tshawytscha*)	11	(39.3%)
O. nerka	9	(32.1%)
O. mykiss	6	(21.4%)
Coho salmon	2	(7.1%)
Host Species	**UC events in coastal watersheds (n=15)**	
Chinook salmon (*O. tshawytscha*)	6	(40%)
O.nerka	1	(6.7%)
O. mykiss	8	(53.3%)

Table 4.7: Temporal description of UC subtype geographic exception detections in coastal watersheds by events and by isolates. Generally, UC subtype viruses are detected in the Columbia River Basin, and thus these represent exceptions to the general geographic structuring rule described in the introduction.

Year	Events (n=15)	Isolates (n=18)	Genotype(s) Detected
1988	1	1	mG199U
1991	1	1	mG032U
1996	1	1	mG001U
1999	3	4	mG001U, mG135U
2000	2	2	mG001U
2002	1	1	mG001U
2003	1	1	mG001U
2007	3	4	mG032U, mG147U
2013	2	3	mG174U

Table 4.8: Temporal description of UP subtype geographic exception detections in the Columbia River Basin by events and by isolates. Generally, UP subtype viruses are detected in coastal watersheds, and thus these represent exceptions to the general geographic structuring rule described in the introduction.

Year	Events (n=28)	Isolates (n=45)	Genotype(s) Detected
1971	1	1	mG018U
1973	1	2	mG018U
1981	1	1	mG028U
1982	2	2	mG003U, mG030U
1984	2	2	mG002U, mG003U
1987	2	2	mG002U, mG029U
1988	2	2	mG028U, mG050U
2001	1	1	mG127U
2002	1	1	mG136U
2004	1	1	mG040U
2006	1	1	mG197U
2009	1	1	mG050U
2010	2	2	mG050U
2011	5	13	mG050U, mG050U/261U, mG159U, mG249U
2012	3	11	mG050U, mG271U, mG271UHH
2013	2	2	mG274U

4.6 HYPOTHESIS THAT THE UC SUB-LINEAGE MAY REPRESENT AN ADAPTATION OF U GENOGROUP IHNV TO CHINOOK SALMON

As described in Chapter 3, the U genogroup phylogeny demonstrated two subgroups, designated UC and UP, with UC representing a distinct lineage within the U genogroup (Figure 3.2). Viruses within the UP subgroup were isolated primarily from *O. nerka*, whereas viruses within the UC sub-lineage were isolated primarily from Chinook salmon (*O. tshawytscha*) (Figure 3.9).

In congruence with previous findings that U genogroup IHNV is primarily associated with sockeye salmon (Kurath et al., 2003), our descriptive epidemiologic analysis indicated that roughly 58% of UP events occurred in sockeye salmon. This finding is notable given the far greater abundance of Chinook salmon to sockeye salmon in the coastal watersheds where UP viruses are primarily detected. In contrast, only 2% of UC events occurred in sockeye salmon. Rather, the majority of UC events occurred in Chinook salmon (59% of UC events) or in *O. mykiss* (close to 30%). This finding led us to hypothesize that the development of the UC sub-lineage may represent an adaptation of U genogroup IHNV to Chinook salmon. While UC viruses were still detected frequently in *O. mykiss*, the phylogenetic tree did not indicate specific clusters of genotypes that were detected predominantly in *O. mykiss*, and therefore we did not think that a parallel adaptation event was also occurring in *O. mykiss*.

4.7 Distribution of disease events and asymptomatic events by host species and subtype

We explored trends in the numbers of non-disease events and disease events attributable to either UC or UP viruses in the three dominant host species. Events in juvenile fish were used as a proxy for disease events since juveniles are generally only sampled if they show symptomatic IHN disease. Within our dataset, both UC and UP viruses caused non-disease and disease events, however UC viruses were responsible for a greater proportion of both event types in Chinook salmon. For instance, of all 39 known disease events in Chinook salmon, 38 (97.4%) of these events were attributable to UC viruses (Table 4.9). For known non-disease events, of the 254 total events in Chinook, 243 (95.7%) non-disease events were due to infection with UC subgroup IHNV. A similar trend was seen in *O. mykiss*, where 22 out of 24 (91.7%) of known U genogroup disease events and 123 out of 135 (91.1%) known non-disease events were attributable to UC subgroup viruses (Table 4.9).

In contrast, both known non-disease and known disease events in *O. nerka* were caused mainly by UP subgroup viruses. Of the 25 known disease events, 21 (84%) were caused by UP viruses, and of the known non-disease events, 69 out of 96 (71.9%) were attributable to UP viruses (Table 4.9). Finally, for each of the three host species there were some events for which it was unknown whether the detection was asymptomatic or symptomatic, however these numbers of events were comparatively low (Table 4.9).

Table 4.9: Numbers of UC and UP subtype events within the three dominant host species stratified by the type of event: non-disease event (asymptomatic detection), disease event, or the status of the event was unknown. U genogroup events in juveniles were used as a proxy for disease events.

***O. nerka* (Sockeye salmon and Kokanee)**	**UC (n=31)**	**UP (n=92)**
Non-Disease Events	27	69
Disease Events	4	21
Unknown Event Status	0	2
***O. tshawytscha* (Chinook salmon)**	**UC (n=282)**	**UP (n=13)**
Non-Disease Events	243	11
Disease Events	38	1
Unknown Event Status	1	1
***O. mykiss* (Steelhead and Rainbow trout)**	**UC (n=147)**	**UP (n=16)**
Non-Disease Events	123	12
Disease Events	22	2
Unknown Event Status	2	2

4.8 Distribution of virus detection and disease events amongst individual sampling locations

In our descriptive analysis of virus distribution amongst sampling sites, we considered two metrics: the total number of events that a site experienced, and the number of disease events at a site. The total number of events represents disease events, non-disease events, and events for which disease status was not known. Asymptomatic (non-disease) events are detected through screening of adult fish returning to hatcheries, a practice that occurs regularly at all sampling sites. Symptomatic disease occurs almost exclusively in juveniles and juvenile screening generally occurs only when symptoms of IHNV have manifested. Asymptomatic juveniles may

be sampled if an outbreak has occurred in fish being reared proximally to asymptomatic juveniles, and are therefore still indicative that symptomatic disease is occurring at that sampling site. IHN-disease can be devastating in hatchery populations, causing epizootics with up to 90% mortality (Groberg, 1983a, 1983b; LaPatra, Parsons, et al., 1993). Thus understanding risk factors for disease events separately from non-disease events remains a critical component of disease control. To this end, we also assessed sites with elevated numbers of IHNV disease events as a first step in exploring risk factors for IHNV epizootics in juvenile fish.

Over the 43 year time period of these datasets, the majority of sampling sites, both within the Columbia River Basin and along coastal watersheds, experienced relatively few IHNV events, and often those events were asymptomatic infections (Figure 4.8: Columbia River Basin sites, and Figure 4.9: coastal watershed sites). To understand which sampling sites appeared to suffer from an unusually high burden of U genogroup IHNV, we considered both the total number of events that occurred at a site and the number of disease events a site had experienced as well. Within the Columbia River Basin, we found four hatcheries that appeared to demonstrate elevated levels of both total events and disease events: Dworshak National Fish Hatchery, Round Butte Hatchery, Nez Perce Tribal Fish Hatchery, and Lookingglass Hatchery. Each of these hatcheries had between 35 and 55 total events, and between 8 and 14 disease events (Table 4.10).

From amongst the coastal watershed sites we found two hatchery systems that demonstrated elevated numbers of both disease and non-disease events: Baker system and Cedar system. Both Baker and Cedar are referred to as hatchery ‘systems’ because they each include multiple proximal sites where fish are sampled. These sampling sites are combined here due to their physical proximity and because all sites within the system are co-managed. Baker system

experienced 36 events in total, of which 8 were disease events. Cedar system experienced 20 events in total, and also had 8 disease events within the study time period (Table 4.11).

Table 4.10: Sampling sites within the Columbia River Basin with a high burden of U genogroup IHNV. High burden sites were designated based on both the total numbers of events that occurred, and based on the number of disease events that occurred.

Sampling Site	Total Events	Disease Events
Dworshak NFH, ID	55	14
Round Butte Hatchery, OR	40	9
Nez Perce TFH, ID	36	8
Lookingglass Hatchery, OR	35	9

Table 4.11: Sampling sites within the coastal watersheds geography with a high burden of U genogroup IHNV. High burden sites were designated based on both the total numbers of events that occurred, and based on the number of disease events that occurred.

Sampling Site	Total Events	Disease Events
Baker System, WA	36	8
Cedar System, WA	20	8

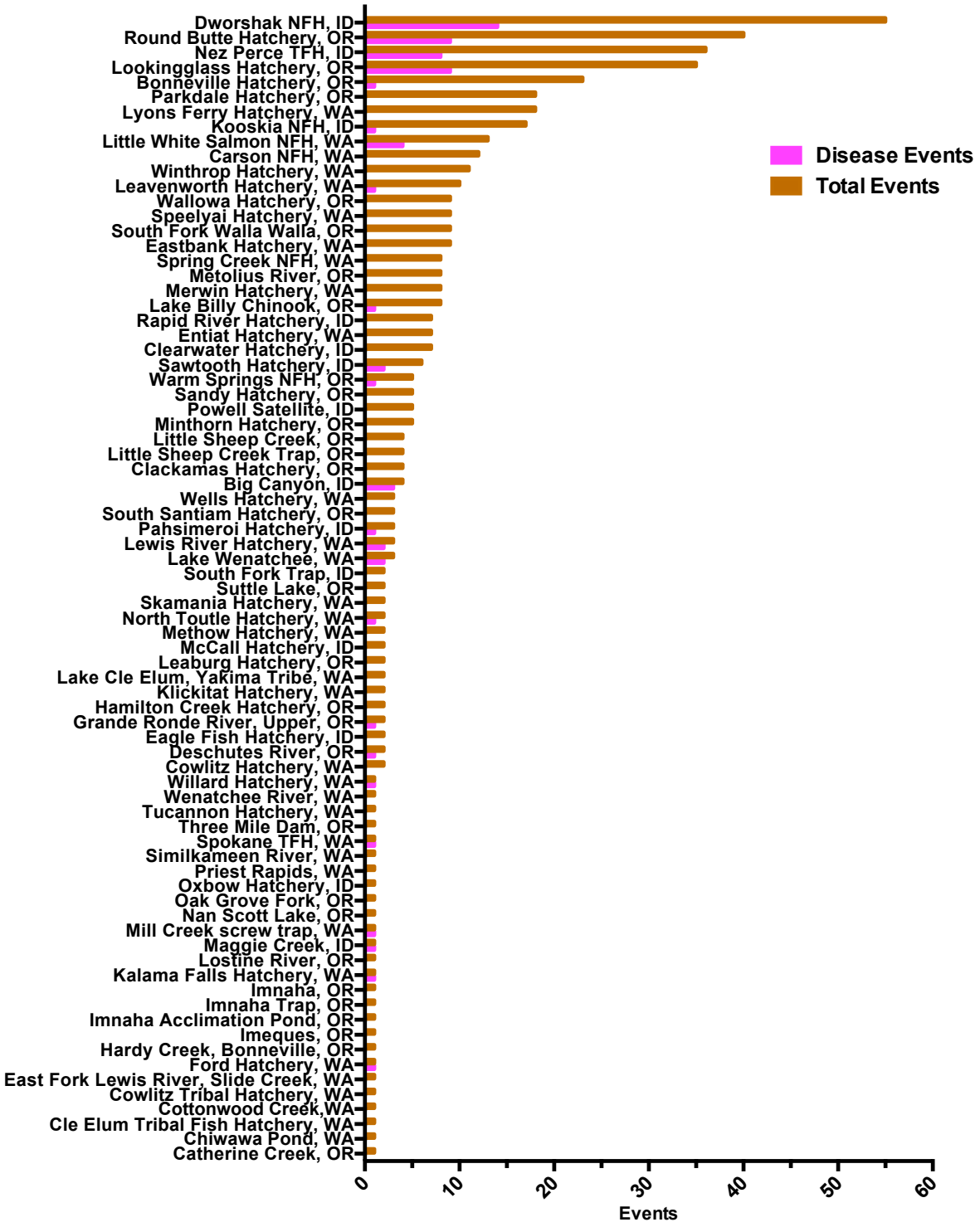


Figure 4.8: Total events and disease events which occurred at 77 sites within the Columbia River Basin from 1971 – 2013. Disease events were defined as any event that occurred in juvenile fish, a proxy for disease since juveniles generally are not sampled unless symptoms of infection are present.

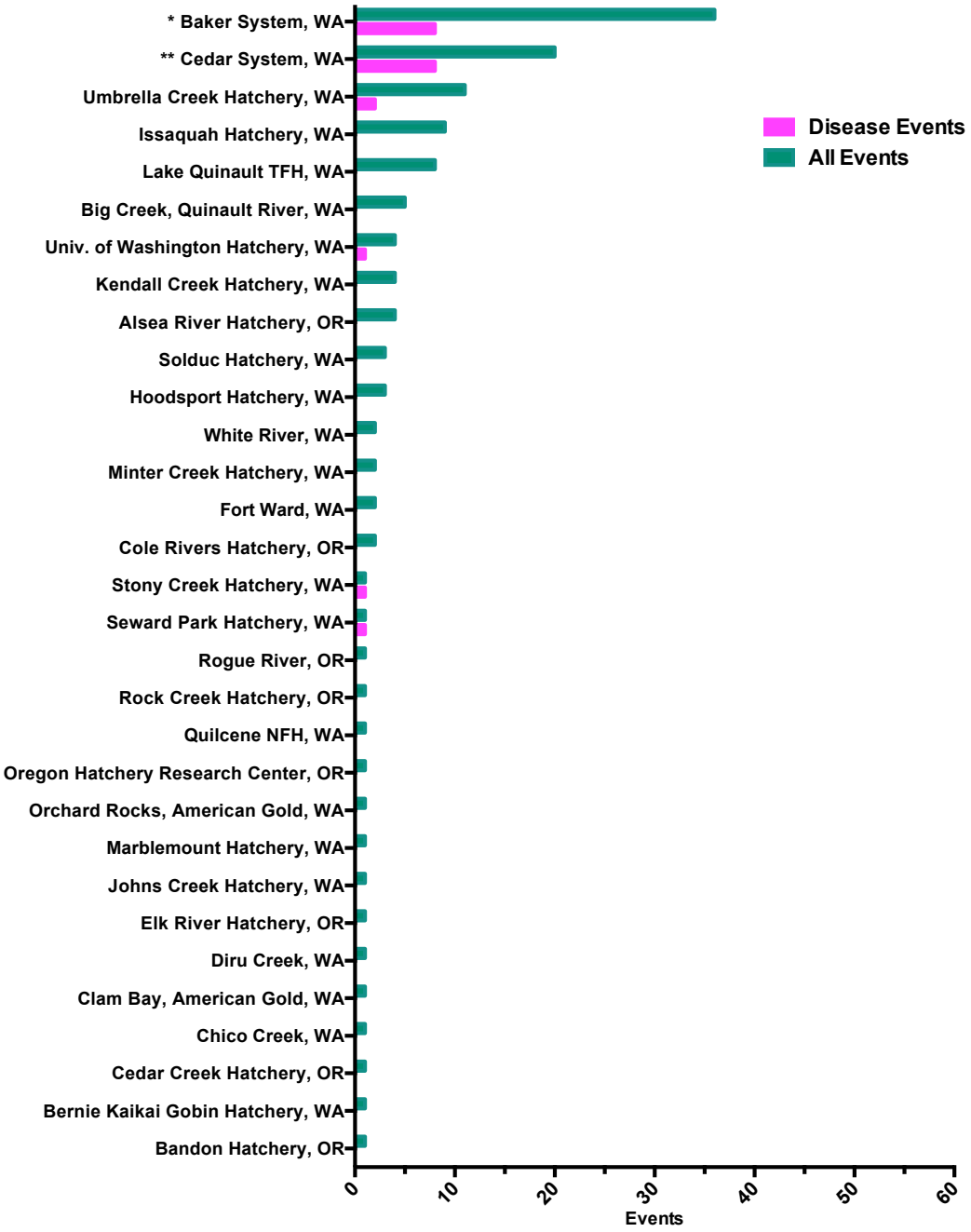


Figure 4.9: Total events and disease events which occurred at 37 sites in coastal watersheds. Disease events were defined as any event that occurred in juvenile fish, a proxy for disease since juveniles generally are not sampled unless symptoms of infection are present. * Baker System is comprised of 4 proximal sites that are managed together: Baker Lake Hatchery, Channel Creek, Channel Creek Natural Spawning Site, and Lake Shannon. ** Cedar System is comprised of 3 proximal sites that are managed together: Cedar Landsburg Hatchery, Cedar River, and Cedar River fry trap.

4.9 DISTRIBUTION OF VIRUS DETECTIONS WITHIN VIRAL GENOTYPES

Within our study there were 114 different genotypes responsible for 619 events. Certain viral genotypes were detected with great frequency and over broad geographic and temporal ranges. Other genotypes were only ever detected once. For instance, of the 114 genotypes of U genogroup IHNV represented in the dataset, 80 types (70.2%) were detected only once (Figure 4.10). However, the most frequently detected genotype, mG001U, was responsible for over 205 events, spanning from 1973 through to 2011 (Table 4.12). Given the range in numbers of events that were attributable to different genotypes, we characterized certain genotypes as 'dominant' U genotypes. Dominant genotypes were defined as those genotypes that caused numerous events at multiple different sampling sites over multiple years. Developing a definition for what constitutes a dominant genotype allows us to monitor new genotypes that may become greater threats in the future, and also allows us to determine whether genotype displacement events are occurring.

Table 4.12: Dominant genotypes of U genogroup IHNV detected between 1971 and 2013. A genotype was considered dominant if it was responsible for 10 or more events. Year ranges designate the first year of detection and the last year of detection.

Dominant Genotypes (year range)	Numbers of Events	Numbers of Isolates
mG001U (1973 – 2011)	205	332
mG151U (2005 – 2013)	65	131
mG174U (2005 – 2013)	60	275
mG147U (2005 – 2013)	33	55
mG050U (1988 – 2013)	32	97
mG032U (1988 – 2009)	27	32
mG002U (1984 – 2012)	26	30
mG003U (1977 – 2000)	12	12

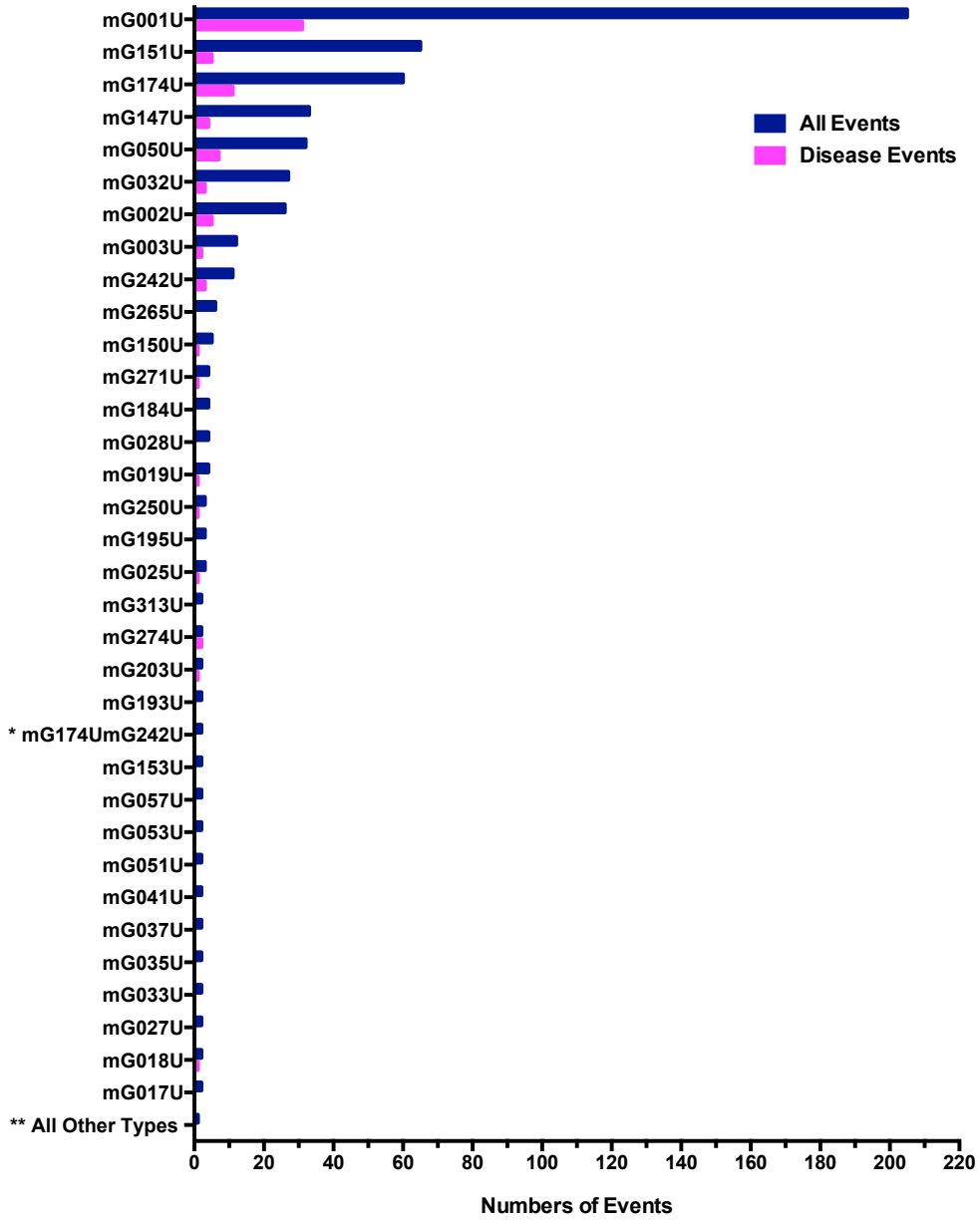


Figure 4.10: Numbers of events caused by 114 unique genotypes of U genogroup IHNV between 1971 and 2013. Any event within a juvenile population was considered a disease event (see justification in text). *mG174UmG242U is a heterogeneous detection of two known sequence types within the same field isolate sample. Since genotyping cannot differentiate between multiple types in the sample due to fish pooling or due to co-infection of a single fish, a heterogeneity of known sequence types is treated separately from events caused by a single type. **The eighty other genotypes not shown were responsible for one event only. Of these eighty types there were 9 that caused a disease event: mG282U, mG194U, mG192U, mG148U, mG095U, mG001UHH, mG001U/293U, mG001U/295U, and mG174U/301U.

Within the current dataset we found that there were eight dominant genotypes: 5 UC subgroup genotypes (mG001U, mG151U, mG174U, mG147U, and mG032U) and 3 UP subgroup viruses (mG050U, mG003U, and mG002U) (Table 4.12). Five of these genotypes were detected at some sampling location over time ranges of greater than 20 years, while three genotypes (mG151U, mG174U, and mG147U) were only detected over the past eight years (Figure 4.11). All of the dominant genotypes except for mG151U were detected outside of their typical geographic range at some point in time (Figure 4.11). Interestingly, mG050U, a UP subgroup virus, was first detected in the Columbia River Basin, and has been detected in both the Columbia River Basin geography and coastal watersheds geography over all years of its detection. Similarly, mG002U, another UP subgroup genotype, was also first detected in the Columbia River Basin. However mG002U has only been detected in coastal watersheds since the late 1980s.

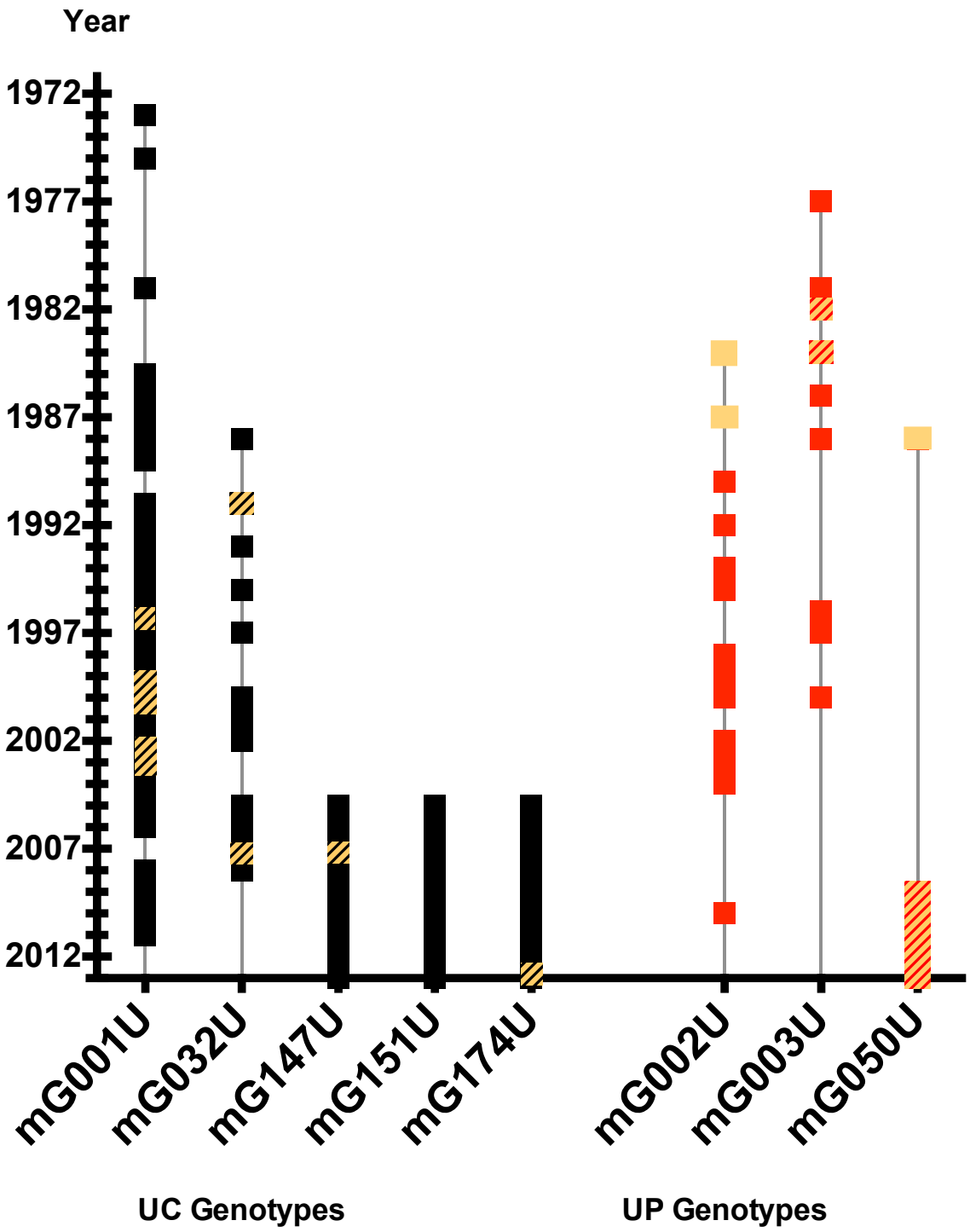


Figure 4.11: Temporal relationships for U genogroup IHNV events for dominant genotypes. Dominant UC genotypes are shown in black, and dominant UP genotypes are shown in red. Hashing in tan indicates occurrence of events both in typical and atypical geographies, and solid tan indicates that the genotype only caused events in an atypical geography.

Chapter 5. F STATISTICS TO INFER DRIVERS OF POPULATION STRUCTURE

5.1 INTRODUCTION

The phylogenies of U genogroup IHNV presented in Chapter 3 demonstrate a subgroup, the UC subgroup, which represents a distinct lineage within the greater U genogroup. Descriptive epidemiological analysis in Chapter 4 has shown that this new lineage shows specific detection patterns by geography and by host species. The separate ranges and apparent host specificity of UC and UP subgroup viruses have led us to hypothesize that geography and host may structure the population of U genogroup viruses.

Originally introduced by Wright (1951), F statistics are tools to quantify the degree of genetic differentiation occurring within and between subpopulations (Holsinger & Weir, 2009). Although originally formalized with regard to allelic frequencies, F statistics have since been extended for use on molecular sequence data. Whereas "haplotype statistics" consider all unique haplotypes whether they vary by few or many nucleotides, "sequence statistics" account for the genetic distance between haplotypes as well (Hudson et al., 1992). Thus, when using F statistics on molecular sequence data, we generally consider nucleotide diversity (π) in place of heterozygosity (Nei, 1982). Such is the case for Nei's (1982) statistic γ_{ST}, which measures the difference between the total nucleotide diversity of the entire population and the mean nucleotide diversity of the subpopulations divided by the total nucleotide diversity, as shown below.

$$\gamma_{ST} = \frac{\pi_T - \overline{\pi_S}}{\pi_T}$$

Nucleotide diversity (π) is estimated through pairwise genetic distances between two sequences. This distance may be raw - calculated as simply the number of nucleotide differences

between the two sequences - or may be corrected through the use of evolutionary models that seek to estimate evolutionary distance not directly observable in the sequence data. Such correction is especially important when considering long evolutionary time periods, since the probability of back mutation increases (Felsenstein, 2004). As discussed in chapter 2, we found that, for U genogroup sequence data, correction under different evolutionary models did not produce different estimates of pairwise genetic distances (Table 3.2). We also found that raw genetic distances (simply the number of nucleotide differences divided by the sequence length) were highly similar to genetic distances corrected under evolutionary models (Table 3.2).

The Hudson, Boos, and Kaplan (1992) test for detecting geographic subdivision uses Nei's (1982) estimation of γ_{ST} to assess sequence variation between subpopulations. Their method is improved over previous methods by its compatibility with the null hypothesis that subpopulations are not genetically distinct. Previous statistics for testing genetic differentiation were unable to describe the distribution under the null hypothesis and thereby were unable to estimate the significance of the point estimates (Hudson, Boos and Kaplan, 1992). In contrast, Hudson, Boos, and Kaplan's (1992) test uses permutation to create a distribution of test statistics derived from samples where sequences are randomly allotted to subpopulations. If there is in fact no population subdivision, then the populations that the sequences are assigned to does not matter, and probability of the initially observed test statistic occurring under the null distribution will be high.

We considered subpopulations defined by two characteristics: geography of detection and host species of detection. We continued to define geography in a dichotomous manner, with detections either occurring within the Columbia River Basin or along coastal watersheds. While IHNV detections within this dataset occurred in a variety of different species of Pacific

salmonids, we hypothesized that genetic differentiation is due to an adaptation from *O. nerka,* an ancestral host of U genogroup IHNV (Kurath et al., 2003), to *O. tshawytscha*. Thus when testing for population structure due to host species, we considered only detections in *O. nerka* and in *O. tshawytscha*.

Importantly, we tested whether groups defined by a specific factor drive the genetic differentiation of a subgroup. Therefore, when populations were defined according to geography of detection, we allotted events to a subpopulation according solely to the geographic range of detection. Thus geographic exception UP subtype viruses were included in the Columbia River Basin population because they were detected within Columbia River Basin, and vice versa for UC viruses detected in coastal watersheds. Likewise, when testing for population structure due to host species, the populations were defined as all detections in *O. tshawytscha* or all detections in *O. nerka*. All detections in both anadromous and non-anadromous *O. nerka* were included since these hosts are the same species.

5.2 TESTING FOR POPULATION STRUCTURE DUE TO GEOGRAPHY OF DETECTION

Given that UC viruses were detected primarily in the Columbia River Basin and UP viruses were detected primarily in coastal watersheds, we considered that geography might play a role in creating distinct populations of IHNV. We tested this hypothesis using both the isolates dataset and the events dataset. The isolates data could potentially make subpopulations appear too similar since they contain multiple representatives of the same genotype at the same location, thereby diluting the genetic diversity within a subpopulation. Despite the potential bias inherent in testing for population structure by isolates, we included this data to determine whether the population structure we observed is attributable mainly to the attenuation of within population diversity.

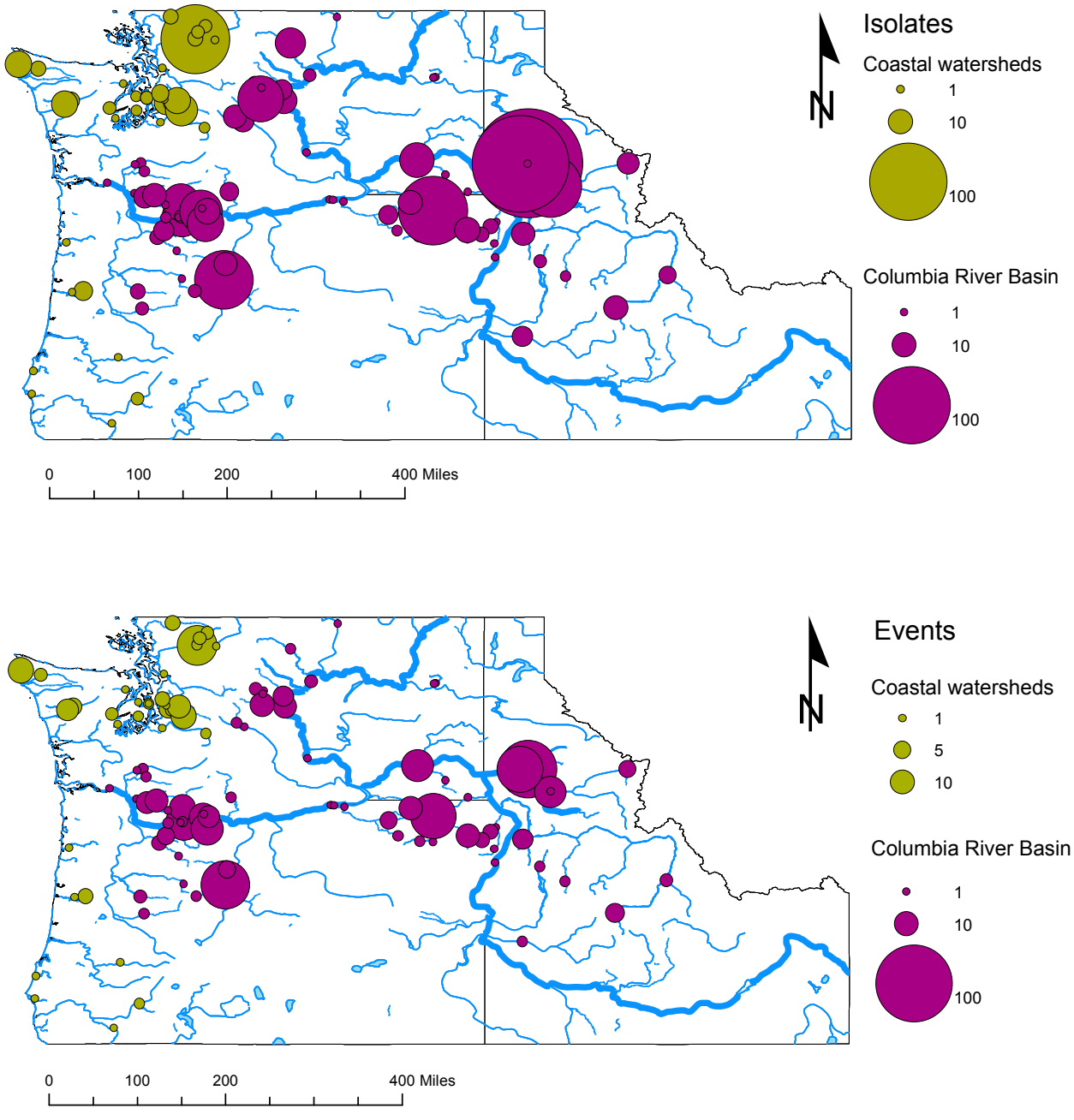


Figure 5.1: Geographic representation of U genogroup detections by isolates and by events in the Columbia River Basin and in coastal watersheds during the study period of 1971 to 2013. Circle size is proportional to the number of times U genogroup IHNV was detected at each site.

All collected isolates and recorded events within the our dataset fell into either the Columbia River Basin population or the coastal watersheds population, thus all sequenced isolates or observed events were included in the analysis of population structure by geography. Using the isolates data, 82.1% of the isolates constituted the Columbia River Basin population, and 17.9% of the isolates formed the coastal watersheds population. In total, across the 1216 isolates with 303nt long sequences, there were 94 distinct sites. By events, all 619 events were included within the analysis, and 488 (78.8%) of events occurred in the Columbia River Basin and 131 events (21.2%) occurred along coastal watersheds.

As can be seen in Figure 5.1, both representatives of the coastal watersheds population and the Columbia River Basin population came from diverse sampling sites within their geographic range. Some sampling sites provided more isolates, and/or had more IHNV events than other sites. This heterogeneity can be seen in the scaling of the circles that represent the isolate count or the event count at each sampling site.

Using the Hudson, Boos, and Kaplan (1992) measure of F_{ST}, we found that F_{ST} equals 0.399, (95%CI: 0.367 – 0.431, $p<0.001$) when we used geography as a driver and when we used all isolates (Table 5.1). When analyzing population structure by geography using the events data we found that F_{ST} equals 0.379 (95%CI: 0.334 – 0.422, $p<0.001$). Both estimates provided good evidence that U genogroup IHNV was structured by the dichotomous geographic designation, although we saw that the F_{ST} estimate by events was slightly lower. This lower estimate was likely attributable to the removal of bias from the within population genetic distance. The removal of this bias was also reflected in the estimates of the mean subpopulation diversity and the metapopulation diversity. As could be expected, the mean subpopulation diversity (the average diversity of each subgroup) and the metapopulation diversity (the genetic diversity of all

included molecular sequences) increased when the analysis was performed by events (Table 5.1). Additionally, the 95% confidence interval around the estimate of F_{ST} widened slightly when the analysis was performed by events; this likely occurred due to the smaller number of sequences included in the analysis, which yielded a slightly less precise estimate of F_{ST}.

Table 5.1: Estimates of nucleotide diversity and F_{ST} for populations defined by geography of detection. Estimates are derived by isolates and by events.

Geography by isolates: *Columbia River Basin and Coastal watersheds*	
Mean interpopulation diversity (π_B)	0.0181
Mean subpopulation diversity ($\overline{\pi_S}$)	0.00556
Metapopulation diversity (π_T)	0.00925
F_{ST} (95% CI), p-value	0.399 (0.367 - 0.431), p<0.001
Geography by events: *Columbia River Basin and Coastal watersheds*	
Mean interpopulation diversity (π_B)	0.0171
Mean subpopulation diversity ($\overline{\pi_S}$)	0.00606
Metapopulation diversity (π_T)	0.00975
F_{ST} (95% CI), p-value	0.379 (0.334 - 0.422), p<0.001

5.3 TESTING FOR POPULATION STRUCTURE DUE TO HOST SPECIES OF DETECTION

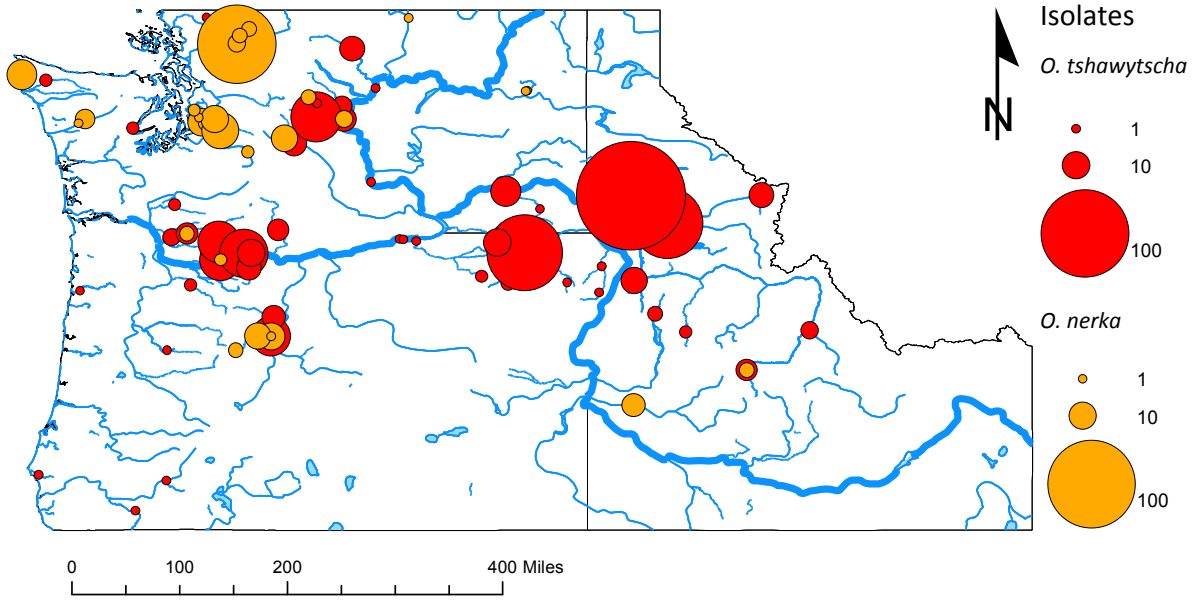
Given that UP subtype viruses were detected primarily in sockeye salmon (anadromous *O. nerka*) and UC subtypes were detected primarily in Chinook salmon (*O. tshawytscha*), we postulated that viral adaptation to different hosts might also structure the U genogroup population. While UC subtype viruses were often detected in *O. mykiss* as well, the phylogeny did not show structure indicating a distinction between UC genotypes detected primarily in Chinook salmon or primarily in *O. mykiss* (see Figure 3.7). Therefore, we thought it was unlikely that UC viruses represent an adaptation to both Chinook salmon and *O. mykiss* separately. If this were the case, we would have seen distinct lineages within UC detected primarily in *O. mykiss*, and other lineages of UC viruses detected primarily in Chinook salmon. Additionally, tests for population structure indicated negligible genetic distances between genotypes responsible for events in *O. mykiss* and events in *O. tshawytscha* (F_{ST} = 0.047, 95%CI: 0.015 – 0.087, probability of observing a higher value of F_{ST} by chance is 0.001).

As performed for geographic range, we tested for population structure defined by host species for both isolates and events. Again, use of the isolates data likely attenuated within population variation due to inclusion of multiple sequences representing the same event (which could represent multiple sequences of the same genotype from the same host). Given that only one genotype was kept per event, analysis of population structure by events was less biased. However both estimates were included to demonstrate how much of the finding was possibly attributable to attenuation of within population variation.

In contrast to the geographic designation of populations, which included all isolates or events, only a subset of isolates and events occurred in either Chinook salmon or in *O. nerka*. Thus the numbers of sequences included in the analysis by host was slightly lower. For example,

while there were sequences for 1216 isolates, only 869 (71.5%) isolates came from either Chinook salmon or *O. nerka*. For all isolates included in the by-host analysis 654 (75.3%) sequences came from detections in Chinook salmon and 215 (24.7%) sequences came from detections in *O. nerka*. Across all 869 sequences included in the by-host analysis, there were 80 distinct sites. By events, 295 (70.6%) sequences represented events that occurred in Chinook salmon, and 123 (29.4%) sequences came from events in *O. nerka*. Thus only 418 of the 619 total events were included in the by-host analysis.

Notably, subpopulations defined by host did not correlate perfectly with subpopulations defined by geography. Detections in Chinook salmon occurred in both the Columbia River Basin and in coastal watersheds, as did detections in *O. nerka* (Figure 5.2). While both host species of detection and geography of detection overlapped, if the populations defined by host and defined by geography had been too similar it would not have been possible to determine whether population structure was attributable to geography as a driver, or to host specificity as a driver.



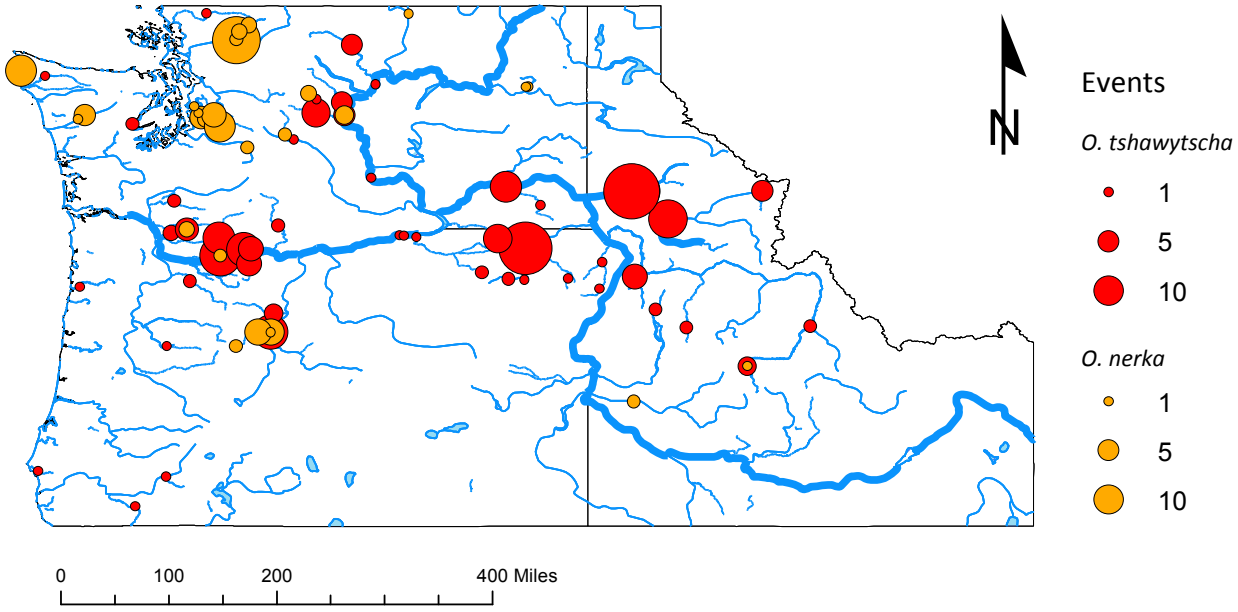


Figure 5.2: Geographic representation of U genogroup detections by isolates and by events in *O. tshawytscha* and *O. nerka* during the study period of 1971 to 2013. Circle size is proportional to the number of times U genogroup IHNV was detected at each site.

When the populations were defined by host species, and using isolates, we found that the Hudson, Boos, and Kaplan (1992) measure of F_{ST} was 0.455 (95%CI: 0.422 – 0.488, p<0.001) (Table 5.2). When looking at the same populations by events, F_{ST} equaled 0.406 (95% CI: 0.358 – 0.458, p<0.001). As before, the point estimate of F_{ST} was slightly lower and the confidence interval was slightly wider when events data were used rather than isolates data. The decrease in the estimate of F_{ST} was likely due to the removal of bias by using events, as the estimates of both the mean subpopulation diversity and the metapopulation diversity increased (Table 5.2). Even using the slightly lower estimate of F_{ST} by events, host species differentiation between Chinook salmon and *O. nerka* appeared to potentially be an even stronger driver of population structure than geography, although further statistical analysis must be performed to determine whether these two estimates differ significantly from each other.

Table 5.2: Estimates of nucleotide diversity and F_{ST} for populations defined by host species. Estimates are derived by isolates and by events.

Host species by isolates: *O. tshawytscha and O. nerka*	
Mean interpopulation diversity (π_B)	0.0172
Mean subpopulation diversity ($\overline{\pi_S}$)	0.00528
Metapopulation diversity (π_T)	0.00971
F_{ST} (95% CI), p-value	0.455 (0.422 - 0.488), p<0.001
Host species by events: *O. tshawytscha and O. nerka*	
Mean interpopulation diversity (π_B)	0.0161
Mean subpopulation diversity ($\overline{\pi_S}$)	0.00608
Metapopulation diversity (π_T)	0.0102
F_{ST} (95% CI), p-value	0.406 (0.358 - 0.458), p<0.001

Table 5.3: Estimates of intrapopulation nucleotide diversities (π) for the subpopulations under analysis in F_{ST} calculations for population subdivision by geography. Reported intrapopulation nucleotide diversities for subpopulations were calculated in MEGA version 6.06. Estimates from HyPhy are the same as estimates reported in Table 5.1. F_{ST} calculated with the unweighted $\overline{\pi_S}$ uses the metapopulation diversity (π_T) calculated in HyPhy. Although estimates of F_{ST} using an unweighted mean are lower than HyPhy estimates of F_{ST}, these are presented for informational purposes only and do not supersede estimates of F_{ST} from HyPhy.

GEOGRAPHY		Mean within subpopulation diversity (π_S)	Unweighted mean subpopulation diversity ($\overline{\pi_S}$)	HyPhy estimated mean subpopulation diversity ($\overline{\pi_S}$)	F_{ST} using unweighted $\overline{\pi_S}$	F_{ST} using HyPhy $\overline{\pi_S}$
Events	Columbia River Basin	0.006	0.0085	0.00606	0.128	0.379
	Coastal Watersheds	0.011				
Isolates	Columbia River Basin	0.005	0.007	0.00556	0.243	0.399
	Coastal Watersheds	0.009				

Table 5.4: Estimates of intrapopulation nucleotide diversities (π) for the subpopulations under analysis in F_{ST} calculations for population subdivision by host species. Reported intrapopulation nucleotide diversities for subpopulations were calculated in MEGA version 6.06. Estimates from HyPhy are the same as estimates reported in Table 5.2. F_{ST} calculated with the unweighted $\overline{\pi_S}$ uses the metapopulation diversity (π_T) calculated in HyPhy. Although estimates of F_{ST} using an unweighted mean are lower than HyPhy estimates of F_{ST}, these are presented for informational purposes only and do not supersede estimates of F_{ST} from HyPhy.

HOST		Mean within subpopulation diversity (π_S)	Unweighted mean subpopulation diversity ($\overline{\pi_S}$)	HyPhy estimated mean subpopulation diversity ($\overline{\pi_S}$)	F_{ST} using unweighted $\overline{\pi_S}$	F_{ST} using HyPhy $\overline{\pi_S}$
Events	*O. tshawystcha*	0.005	0.009	0.00608	0.118	0.406
	O. nerka	0.013				
Isolates	*O. tshawystcha*	0.005	0.008	0.00528	0.176	0.455
	O. nerka	0.011				

Chapter 6. DISCUSSION

6.1 RISE IN THE NUMBERS OF ISOLATES/EVENTS OVER THE STUDY PERIOD

As presented in Chapter 4, the numbers of both isolates and events collected as part of the WFRC IHNV genetic typing program have increased over the study time period, most sharply after 2000. Because this program is a passive surveillance program, the left-skew of these data could be attributable to either changes in reporting practices, with an increase in isolate submission more recently, or could represent an actual change in the number of virus detections that are occurring in the study area.

Based on our records it appears that the rise in U genogroup events over time is primarily due to increased reporting of events as more hatcheries participate in the IHNV genetic surveillance program. The success of Emmenegger et al. (2000; 2002), Garver et al. (2003) and Troyer et al.'s (2000; 2003) use of IHNV midG genotyping to uncover important epidemiological dynamics demonstrated the utility of genetic typing for hatchery management. Thus participation in the program increased after 2000 with isolates being sent in real-time as disease and virus-positive samples were detected. Before 2000, fish health agencies generally only sent current or archival samples that had been saved, a much smaller number of isolates than are collected through general screening practices.

While participation in the IHNV genetic typing program has increased, this greater participation could still be driven by a greater abundance of virus events occurring in the study capture area. Thus while more sources submit isolates, there could also be more positive IHNV detections to send isolates from. To investigate this question, other researchers in our laboratory

have been developing a diagnostics database to investigate whether the proportion of IHNV-tested fish that are positive for IHNV has changed over time.

6.2 Comparison of event distribution by host and by geography

Analysis of the sampling densities of both U genogroup isolates and U genogroup events indicated a bimodal distribution in isolates and events that occurred in the Columbia River Basin, with an initial peak in 2004, and a second higher peak in 2011. A bimodal distribution was not apparent for isolates and events that occurred in coastal watersheds. Rather coastal watersheds appeared generally constant with a slight left-skew.

The bimodal distribution of U genogroup detections (both isolates and events) in the Chinook salmon (*O. tshawytscha*) aligns well with the bimodal distribution of all virus detections in the Columbia River Basin, also indicating a peak in 2004 and a second higher peak in 2011. Detections in *O. mykiss*, the other dominant species of the Columbia River Basin, appeared generally constant with a slight left-skew after 2000. This finding indicates that the temporal patterns of U genogroup detection in the Columbia River Basin appear to be driven largely by detections in Chinook salmon.

6.3 Differences in U genogroup IHNV incidence by geography

The difference in the number of virus-positive sites in the Columbia River Basin (n=77) compared to coastal watersheds (n=37) does not appear to be driven by relative fish abundance. Comparing total numbers of Chinook salmon, steelhead trout, and sockeye salmon that are cultured annually within a geographic region, the Columbia River Basin cultures only slightly higher total numbers of fish. Annually around 105 million juvenile fish are released from hatcheries in the Columbia River Basin and around 86 million juvenile fish are released from hatcheries in coastal Washington, coastal Oregon, and Puget Sound, resulting in a ratio of 1.2 to 1 between abundance of fish cultured in the Columbia River Basin versus cultured fish abundance in coastal watersheds. Yet twice as many virus-positive sampling sites occur in the Columbia River Basin than in coastal watersheds, and when using numbers of events from our dataset, 3.7 events occurred in the Columbia River Basin for every single event in coastal watersheds.

Since the relative fish abundance of the Columbia River Basin and coastal watersheds does not appear to correlate with the increased number of virus-positive sites in the Columbia River Basin, other differences between these two geographies likely affect virus incidence to some degree. For instance, differences in ecology and human-impact may drive part of the disparity in virus-positive sampling sites between the Columbia River Basin and coastal watersheds. As discussed in Chapter 1, the Columbia River Basin is a highly managed system that is impacted more by agriculture and damming than many of the coastal watershed rivers. Therefore it is possible that the environmental differences in these two geographies could drive differences in observed IHNV events.

6.4 INCLUSION OF GREATER NUMBERS OF SEQUENCES DOES NOT GREATLY CHANGE ESTIMATES OF U GENOGROUP GENETIC DIVERSITY

Previous regional epidemiological studies of IHNV have often reported the mean intrapopulation nucleotide diversity (π) of the sequences under study. For instance, Emmenegger and Kurath (2002) studied U genogroup IHNV detected on the Washington coast, finding a mean intrapopulation nucleotide diversity of 0.007 (2.12 nucleotides different over a 303nt sequence length) and a maximum nucleotide diversity of 0.02 (6.06 nts) in their 61 sequenced isolates. These estimates are highly similar to the nucleotide diversity of 42 isolates of IHNV from Alaska where mean intrapopulation nucleotide diversity was 0.006 (1.82nts) and the maximum nucleotide diversity was 0.0199 (6.06nts) (Emmenegger et al, 2000). Finally, in a general phylogeographic analysis of IHNV in North America, Kurath et al. (2003) found that U genogroup isolates included in their study (n=180) had a mean intrapopulation diversity of 0.0088 (2.7 nts) and a maximum nucleotide diversity of 0.0297 (9nts).

Since these previous studies were published we have collected large amounts of additional sequencing data for U genogroup IHNV. To see how much estimates of U genogroup diversity would change when far greater numbers of sequences were included in the analysis we estimated intrapopulation mean and maximum nucleotide diversities for U genogroup isolates and events. While mean intrapopulation nucleotide diversity by isolates is biased towards less diversity by inclusion of multiple sequences from the same event, we include it since the previous analyses of U genogroup IHNV genetic diversity were all performed using viral isolates rather than events. Within our current study, the mean intrapopulation nucleotide diversity of the isolates data was 0.0092 (2.77 nts), which is highly similar to Kurath et al.'s (2003) estimate of the mean intrapopulation diversity of the U genogroup. When using the less biased events data

we found that all U genogroup events exhibited a slightly higher mean intrapopulation diversity of 0.0096 (2.92 nucleotides) as expected, yet this less biased estimate still remains very similar to estimates of U genogroup genetic diversity from previous studies. Within our study, the U genogroup had a maximum nucleotide diversity of 0.0429 (13 nucleotides).

Exploring the diversity of the newly resolved U subgroups, the UP subgroup accounted for the majority of this diversity, with a mean intrapopulation diversity of 0.00917 (2.78nts) and a maximum intrapopulation diversity of 0.0396 (12nts). In contrast, UC events demonstrated roughly 2-fold lower diversity; the mean intrapopulation nucleotide diversity of UC events was 0.00414 (1.25nts) and the maximum nucleotide diversity was 0.0198 (6nts).

Our analysis, which contained more than 400 additional U genogroup sequences from previous analyses, did not show markedly different levels of nucleotide diversity from prior studies of U genogroup IHNV. These data provide evidence that sampling saturation was previously reached in older studies of U genogroup IHNV genetic diversity as we did not observe an increase in the diversity of the viral population with increased sampling. Thus we likely have an accurate estimate of the mean intrapopulation nucleotide diversity not only for our sample but also for the U genogroup viral population circulating in the field.

Additionally, descriptive epidemiological analysis indicates that some sequence types are detected over periods of multiple decades (see Figure 4.11). These findings are particularly interesting given the high mutation rates of RNA viruses. RNA viruses are measurably evolving pathogens, meaning that new mutations can be detected in genetic samples taken at different time points (Biek, Pybus, Lloyd-Smith, & Didelot, 2015). While IHNV certainly fulfills the criteria for a measurably evolving pathogen, it is then unusual that samples with no mutational changes would be detected over relatively long time scales. While the midG sequence is a small

region of the viral genome, as a portion of the sequence encoding the viral glycoprotein it is more variable than other portions of the genome (Nichol, Rowe, & Winton, 1995). Therefore the consistency with which some genotypes are detected over time, and the similar estimates of viral population diversity to older studies, provide further evidence that U genogroup IHNV may have reached a fitness peak, whereby purifying selection maintains some currently circulating genotypes (Kurath et al., 2003).

6.5 Both host and geography play roles in shaping U genogroup evolution

Generally, separate viral lineages evolve due to genetic isolation of viral populations. A variety of mechanisms may be responsible for driving this isolation, and determining which mechanisms contribute most to population structure can have important implications for disease management. Previous studies have demonstrated that evolution of North American IHNV has been correlated with two major factors: host species adaptation and geographic structuring of populations. For example, evolution of the M genogroup of IHNV likely occurred due to a host jump from sockeye salmon to rainbow trout and adaptation to replication at water temperatures of 15 degrees Celsius (Amend & Smith, 1975; Kurath et al., 2003). Alternatively, the division of the L genogroup appears to have been facilitated through geographic structuring alone. Indeed, both LI and LII affect primarily Chinook salmon but correlate strongly with a geographic division between detection in the California Central Valley and detection in coastal watersheds along the northern California coast (Kelley et al., 2007).

As supported by statistical tests for population structure, U genogroup IHNV appeared to be structured by both geography and by host species (see Chapter 5). To further investigate the

role of geography, we compared the patterns we saw in the U genogroup data to those seen by Kelley et al. (2007) in Californian IHNV, a clear example of geographic subdivision of the viral population. Analogously, Kelley et al.'s (2007) IHNV population was structured by a geographic division between a basin geography (the California Central Valley) and a coastal geography (northern California coastal watersheds). Their findings indicate that LI is predominantly detected out in coastal watersheds and LII is detected within the basin. This division of viral populations detected in large contiguous watersheds from viral populations detected in smaller coastal watersheds mirrors our findings that UP subtype viruses occur more frequently in coastal watersheds of the Pacific Northwest and UC subtype detections occur predominantly in the Columbia River Basin.

Despite the highly analogous correlation between geographic division and phylogenetic differentiation in both Kelley et al.'s (2007) study and our study, there are a number of key differences between LI and LII and UP and UC. Firstly, since 1988 the LI and LII subgroups have only been detected in their typical ranges (LI in coastal watersheds and LII in the California Central Valley) despite increasing sampling efforts through time (Kelley et al., 2007). In contrast, while the majority of UP detections happen in coastal watersheds of the Pacific Northwest, and UC detections occur predominantly within the Columbia River Basin, we see occasional occurrences of geographic exceptions to this rule. Within our study 15 UC events occurred in coastal watersheds and 28 UP events occurred within the Columbia River Basin.

Another key difference between the basin/coastal geography of the California L range and the U range in the Pacific Northwest is host species composition. In California, the majority of L detections occur in *O. tshawytscha* (Chinook salmon), in both coastal watersheds and in the California Central Valley. Thus Chinook salmon serve as the primary host for both LI and LII

subtypes. The homogeneity of host species in California L IHNV detections stands in contrast to the patterns we see with U genogroup in the Pacific Northwest. In coastal watersheds the predominant host for IHNV is sockeye salmon (*O. nerka*), whereas infections in the Columbia River Basin occur mainly in Chinook salmon (*O. tshawytscha*) and steelhead trout (*O. mykiss*). Thus, while geographic separation has been demonstrated to solely drive population subdivision with L genogroup in California, U genogroup IHNV was also structured by host species of detection, as supported by statistical tests of population subdivision by host species (see Table 5.2).

6.6 Differentiating the roles of geography and host species in structuring the viral population.

U genogroup IHNV in the Pacific Northwest appears to be genetically structured by at least two drivers: host species and geography. However determining which of these two drivers contributes more to the genetic isolation of viral populations is challenging since host species and geography of detection are somewhat related. For instance sockeye salmon culture is more prevalent in the coastal watersheds geography and Chinook salmon and steelhead salmon are cultured in greatest numbers in the Columbia River Basin.

However while host species and geography are related, abundance data on cultured salmonids demonstrate that these two drivers are not perfectly correlated. Chinook salmon are cultured in much higher abundance to any other salmonid species in the Columbia River Basin. They are also cultured in much higher abundance to other fish species in Puget Sound and on the Washington coast. On the Oregon coast Chinook salmon and steelhead salmon are cultured at similar frequencies, and no sockeye salmon are cultured. Thus the dominant species of U

genogroup IHNV detection in coastal watersheds, sockeye salmon, is *not* the most abundant species in the coastal watersheds range. This finding indicates a role both for host species and geography as drivers of viral population structure. Since the primary host of infection in coastal watersheds is not in fact the most abundant host in the coastal watersheds geography, then IHNV in this geography (UP subgroup viruses) likely has a specificity for sockeye salmon. However, we have also observed coastal U virus infections in Chinook salmon and steelhead trout, generally as geographic exceptions with UC subgroup viruses. Therefore, the general paucity of infections in Chinook salmon and steelhead trout on the coast is not attributable to the lack of any virus that could infect them, but rather that the UC subgroup that appears to be specific to Chinook salmon and also causes infection in *O. mykiss* does not occur as frequently in coastal watersheds.

We postulate that the development of the UC geographic range is determined by a unique selection pressure to adapt to Chinook salmon within the Columbia River Basin. Historically, sockeye salmon were the primary salmonid species of the Columbia River Basin. After construction of the Grand Coulee dam, conservation of sockeye salmon in the Columbia River Basin was supported through sockeye salmon hatchery programs. However, explosive epidemics of IHNV within these cultured sockeye populations caused hatcheries to stop culturing sockeye salmon and switch to culturing Chinook salmon and steelhead trout, species that did not appear to be as susceptible to IHNV disease. Overtime, abundance of sockeye salmon in the Columbia River Basin decreased markedly and the basin became a predominantly Chinook salmon watershed. This human-caused change in host species composition away from the ancestral host of U genogroup IHNV to a new and less susceptible host likely provided a selection pressure for U genogroup IHNV adaptation to Chinook salmon, resulting in the development of the UC sub-

lineage which demonstrates greater specificity for Chinook salmon. This same shift was likely not seen in coastal watersheds since coastal hatchery programs continued to culture sockeye salmon in high numbers. Thus even as greater numbers of Chinook salmon and steelhead trout were reared on the coast, the selection pressure for U adaptation to Chinook salmon was likely precluded on the coast by the continued presence of sockeye salmon.

6.7 What are the possible drivers of geographic exceptions?

Observationally, geographic exception detections (detections of UP viruses in the Columbia River Basin and UC viruses in coastal watersheds) did not distribute randomly around the study area. Rather, the majority of coastal UC events (12 out of 15) occurred in Oregon coastal watersheds, and a large number of UP events in the Columbia River Basin (19 out of 28 events) occurred in the Upper Columbia River in central and eastern Washington.

The Oregon coast is notable for both its paucity of IHNV events and for being the only coastal region within our study area where UC viruses were detected in excess of UP viruses. Indeed, over our study time period the Oregon coast only experienced 12 IHNV events, all of which were caused by UC genotype viruses. Since the Oregon coast is more affected by UC events than other coastal geographic regions, we explored differences between the Oregon coast and the Washington coast and Puget Sound that may influence the sporadic occurrence of UC viruses in coastal watersheds.

Despite the apparent specificity of UC genotype viruses for Chinook salmon, geographic exception detection of UC genotypes in coastal watersheds does not appear to be attributable to Chinook salmon abundance. Of all the coastal watershed geographic regions, hatcheries on the Oregon coast culture the least number of Chinook salmon, roughly 3 million fish annually. In

comparison over 50 million Chinook salmon are cultured in Puget Sound and around 11 million Chinook salmon are cultured on the Washington coast annually. Despite a higher abundance of Chinook salmon, only 3 UC geographic exception events occurred in Puget Sound, and none occurred on the Washington coast. Additionally, of the four events that occurred in Chinook salmon in Puget Sound or on the Washington coast, 2 events were due to UC genotype viruses and 2 events were due to UP genotype viruses.

Rather, it seems more likely that the detection of UC geographic exceptions is determined by the presence or absence of sockeye salmon within a region. The Oregon coast is unique within the coastal watersheds geography since it does not culture sockeye salmon, the predominant host for UP viruses that cause frequent events in coastal watersheds. In contrast, sockeye salmon are cultured both in Puget Sound and on the Washington coast. Thus UC detections in coastal Oregon may be attributable to the lack of UP viruses in the region.

When exploring UP geographic exceptions, it seems unlikely that the abundance of *O. nerka*, the primary host for UP detections in coastal watersheds, drives the observed spatial pattern of geographic exception UP events in the Columbia River Basin. Most UP events in the Columbia River Basin occurred in Chinook salmon or steelhead trout, with only 32% of UP geographic exceptions occurring in *O. nerka*. Rather, it appears as though UP geographic exceptions often occur in areas that are: 1) proximal over land to marine areas where UP viruses are predominantly detected and 2) also represent areas of the Columbia River Basin where M genogroup viruses are not detected. Importantly, these two features point to two possible drivers of UP geographic exception events. Firstly, we postulate that overland transmission of UP viruses may occur since UP exception events are not detected along the entire migratory length of the Columbia River mainstem (despite hatchery presence), but are instead focused near the

mouth of the Columbia River (overland proximity to the Washington coast) and then much farther upriver in the Upper Columbia (overland proximity to Puget Sound) (refer to Figure 4.6). Hypothetically, this type of pattern could arise if UP viruses were being trafficked overland rather than via migratory fish. Virus could presumably be trafficked through movement of infected fish by birds or animals, or potentially by humans moving virus-positive fish (e.g. after being caught during recreational fishing) between geographies. To test this hypothesis, approaches comparing the correlation between genetic distance and physical distances using different metrics (river distance, road distance, and great circle distances) may help in the future to inform key transmission routes.

An alternative hypothesis is that we see UP geographic exception detections in these parts of the Columbia River Basin specifically because M genogroup viruses are not present in these areas. Interestingly, this second hypothetical mechanism of influence on geographic exception events is more similar to the patterns we see in UC geographic exceptions, whereby UC detections in the coastal watersheds geography occurred in areas that lack UP virus detections.

6.8 Geographic population structure of U genogroup IHNV indicates some degree of transmission when host populations are structured

In Chapter 5, we used F statistics to compare the intrapopulation diversity and the total diversity for virus populations structured by a binary geographic designation. The less biased F_{ST} estimate for events structured by geography of 0.379 (95%CI: 0.334 – 0.422, $p<0.001$) indicated that geography is a driver of virus population structure. This finding has important implications for the timing of IHNV transmission.

Historically it has been proposed that the majority of IHNV transmission occurs during the freshwater portions of the salmonid host life cycle, at spawning and after hatching (Bootland & Leong, 1999). However previous molecular epidemiology studies that found identical genotypes of U virus in North America and eastern Russia have suggested that IHNV transmission occurs in the marine environment (Rudakova, Kurath, & Bochkova, 2007) while Asian and North American salmonids co-mingle in the Alaskan gyre and the Bearing Sea gyre (Healy, 1991 pg. 366, Fig. 28). The finding here that IHNV populations are structured by geography supports the original hypothesis that at least a large component of viral transmission is occurring while fish populations are geographically structured, and thus in freshwater. This means that at least some portion of transmission occurs while fish are young and co-mingling in their freshwater habitats or swimming downriver, or when adult fish are returning to spawn. Taken in conjunction with Rudakova et al.'s (2007) findings, the accumulating body of knowledge appears to demonstrate that some transmission occurs both in marine and freshwater environments. In the future, mathematical modeling approaches could be used to determine the contribution of marine and freshwater transmission to the observed epidemiological patterns. Such approaches could also potentially be used to investigate whether virus transmission appears to predominantly occur in an upstream direction (correlating with transmission by returning adult populations) or in a downstream direction (correlating with transmission by out-migrating juvenile populations).

6.9 Strength of the UC subgroup in comparison to subgroups within the M Genogroup

We have defined here a new lineage within the U genogroup, the UC sub-lineage. Although well supported in our Bayesian coalescent phylogeny (posterior support = 0.7), the UC subgroup is less supported than sub-lineages in other IHNV genogroups. For instance, sub-lineages within the M genogroup have higher posterior support values (e.g. posterior support for MB is 0.89, posterior support for MC is 1.0) (Breyta et al., 2013). We postulate that the UC sub-lineage does not show as strong posterior support as IHNV sub-lineages within the M genogroup due to differences in host dynamics. M genogroup viruses are found predominantly in rainbow trout in the Hagerman Valley (Troyer & Kurath, 2003; Troyer, Lapatra, & Kurath, 2000). These fish do not migrate, and since they are a predominantly cultured population, subgroups of hosts are highly fragmented among separate ponds and separate farms. These fish are also reared at a constant high density. These characteristics of the host species life history then also impact viral populations, and the high degree of isolation of different viral subpopulations, along with a constant replenishment of susceptible hosts, allows strong viral sub-lineages in the M genogroup to develop.

In contrast, there is much more fluidity in the host populations for U genogroup viruses (Chinook salmon, steelhead trout, and sockeye salmon). Each of these host species is migratory, thus reducing the isolation of fish populations greatly, especially during fish maturation in the marine environment. Given that some degree of IHNV transmission is likely occurring in the marine environment when fish populations co-mingle (Rudakova et al., 2007), it seems understandable that a signal of genetic isolation of the viral population would be weaker for a sub-lineage within the U genogroup in comparison to the M genogroup.

6.10 Conclusion

In conclusion, our analysis indicates that the U genogroup of IHNV has two newly defined subgroups, UC and UP. These two subgroups correlate both with geography of detection and with host species. UC viruses are mainly detected in the Columbia River Basin and in Chinook salmon and steelhead trout. UP viruses are detected primarily in coastal watersheds of the Washington coast and Puget Sound and are most commonly isolated from sockeye salmon.

Despite having a general geographic range, these U subgroups are also occasionally detected outside their typical geographies. When UC viruses occur outside of the Columbia River Basin they are principally detected in Chinook salmon and steelhead trout in coastal watersheds of Oregon. When UP viruses are detected within the Columbia River Basin they are most commonly detected in *O. mykiss* and Chinook salmon, either at the mouth of the Columbia River or deeper into the basin in central Washington and northern Idaho.

We speculate that the development of the UC sublineage has occurred as a result of U genogroup IHNV adaptation to Chinook salmon within the Columbia River Basin. Importantly, this adaptation was likely facilitated through the human-mediated turnover of the watershed from supporting mainly sockeye salmon to supporting predominantly Chinook salmon.

Finally, given that watershed geography appears to divide the UC and the UP viral populations, host populations must to some extent be geographically structured during infection transmission. This finding indicates that a meaningful proportion of IHNV transmission occurs in freshwater, either in juvenile hosts migrating out to the ocean or in adult fish returning to spawn.

Chapter 7. RELEVANCE TO HUMAN HEALTH

It is highly unlikely that IHNV would develop the ability to infect humans directly (either through fish consumption or handling) since the replication temperature of IHNV is significantly lower than human body temperature (Bootland & Leong, 1999). Thus IHNV poses no direct infection risk to human beings. However, IHNV can impact human health indirectly through a variety of pathways, and study of this virus as a model system may prepare us for zoonotic infections that can infect humans.

Firstly, conservation hatchery programs within the Pacific Northwest seek to protect Pacific salmon since they are cultural icons, a key food source, and animals with important economic and recreational value. The health of Pacific salmon is indirectly related to human wellbeing through the maintenance of secure food systems and the functionality of ecosystems for recreational and environmental purposes. Additionally, salmon culture supports rural communities of the Pacific Northwest economically. As an endemic disease that can cause major epidemics in high-density juvenile populations, fish loss due to IHNV infection can result in significant economic and conservation impacts that may affect human wellbeing.

Secondly, the study of IHNV epidemiological dynamics has value for human health as a model system for understanding transmission dynamics of similar pathogens. IHNV is a single-stranded negative-sense RNA virus, thus falling within the *Mononegavirales* order. Importantly, other pathogens within the *Mononegavirales* order, for example measles virus and mumps virus, cause significant disease in humans. Additionally, several members of this viral order, such as Ebola virus and Nipah virus, are zoonotic infections. Despite being an aquatic pathogen of a non-human host, transmission of IHNV within salmonid hosts may be analogous to transmission of other negative-sense RNA viruses among humans. For instance, the interface of domesticated

fish in high-density populations with less dense wild fish populations is comparable to heterogeneity in human contacts between high-density populations such as schools versus numbers of contacts in neighborhoods.

Finally, transmission of an aquatic pathogen in water is analogous to the transmission of an airborne pathogen on land. Both of these transmission mechanisms allow infection transmission through a medium (either water or air) without the need for direct contact between a susceptible and an infected host or through indirect contact with contaminated fomites. Importantly, given that IHNV is a model system in a non-human animal, hypotheses about evolution of virulence, transmission mechanisms, or other infection dynamics that may be applicable to human pathogens can be tested *in vivo* in a controlled laboratory setting. This ability to experimentally test hypotheses derived from observational epidemiological data is a key benefit to work in a non-human pathogen system, and may greatly help our understanding of pathogens that do directly impact human health via human infections.

BIBLIOGRAPHY

Amend, D. F., & Smith, L. (1975). Pathophysiology of infectious hematopoietic necrosis virus disease in rainbow trout: hematological and blood chemical changes in moribund fish. *Infection and Immunity*, *11*(1), 171–179.

Amend, D. F., & Wood, J. F. (1972). Survey for infectious hematopoietic necrosis (IHN) virus in Washington salmon. *Progressive Fish-Culturist*, *34*(3), 143–147.

Anderson, E., Engelking, H., Emmenegger, E., & Kurath, G. (2000). Molecular Epidemiology Reveals Emergence of a Virulent Infectious Hematopoietic Necrosis (IHN) Virus Strain in Wild Salmon and Its Transmission to Hatchery Fish. *Journal of Aquatic Animal Health*, *12*(2), 85–99. http://doi.org/10.1577/1548-8667(200006)012<0085:MEREOA>2.0.CO;2 To

Baele, G., Lemey, P., Bedford, T., Rambaut, A., Suchard, M. A, & Alekseyenko, A. V. (2012). Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Molecular Biology and Evolution*, *29*(9), 2157–67. http://doi.org/10.1093/molbev/mss084

Bendorf, A. (2010). *Phylogenetic Analysis and Virulence Characteristics for Chinook Salmon, Rainbow Trout, and Steelhead of L Genogroup Infectious Hematopoietic Necrosis Virus (IHNV) from the Feather River, California*. University of California Davis.

Bendorf, C. M., Kelley, G. O., Yun, S. C., Kurath, G., Andree, K. B., & Hedrick, R. P. (2007). Genetic diversity of infectious hematopoietic necrosis virus from Feather River and Lake Oroville, California, and virulence of selected isolates for Chinook salmon and rainbow trout. *Journal of Aquatic Animal Health*, *19*(4), 254–269.

Biek, R., Drummond, A. J., & Poss, M. (2006). A virus reveals population structure and recent demographic history of its carnivore host. *Science (New York, N.Y.)*, *311*(5760), 538–541. http://doi.org/10.1126/science.1121360

Biek, R., Pybus, O. G., Lloyd-Smith, J. O., & Didelot, X. (2015). Measurably evolving pathogens in the genomic era. *Trends in Ecology & Evolution*. http://doi.org/10.1016/j.tree.2015.03.009

Bootland, L., & Leong, J. (1999). Infectious hematopoietic necrosis virus. In P. Woo & D. Bruno (Eds.), *Fish Diseases and Disorders* (pp. 57–112). Wallingford UK: CAB International.

Breyta, R., Jones, A., Stewart, B., Brunson, R., Thomas, J., Kerwin, J., … Kurath, G. (2013). Emergence of MD type infectious hematopoietic necrosis virus in Washington State coastal steelhead trout. *Diseases of Aquatic Organisms*, *104*(3), 179–195. http://doi.org/10.3354/dao02596

Burgner, R. (1991). Life History of Sockeye Salmon. In C. Groot & L. Margolis (Eds.), *Pacific Salmon Life Histories* (pp. 1–118). Vancouver, BC: UBC Press.

Carter, J., & Saunders, V. (2013). *Virology: Principles and Applications* (2nd ed.). Hoboken, NJ: John Wiley & Sons, Inc.

Cleveland, W., & Devlin, S. (1988). Locally Weighted Regression : An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association*, *83*(403), 596–610. Retrieved from http://www.jstor.org/stable/2289282

Darriba, D., Taboada, G. L., Doallo, R., & Posada, D. (2012). jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods*, *9*(8), 772–772. http://doi.org/10.1038/nmeth.2109

Drummond, A. J., Ho, S. Y. W., Phillips, M. J., & Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. *PLoS Biology*, *4*(5), 699–710. http://doi.org/10.1371/journal.pbio.0040088

Drummond, A. J., Rambaut, A., Shapiro, B., & Pybus, O. G. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular Biology and Evolution*, *22*(5), 1185–92. http://doi.org/10.1093/molbev/msi103

Drummond, A. J., & Suchard, M. A. (2010). Bayesian random local clocks, or one rate to rule them all. *BMC Biology*, *8*, 114. http://doi.org/10.1186/1741-7007-8-114

Drummond, A. J., Suchard, M. A., Xie, D., & Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*, *29*(8), 1969–1973. http://doi.org/10.1093/molbev/mss075

Emmenegger, E. J., Meyers, T. R., Burton, T. O., & Kurath, G. (2000). Genetic diversity and epidemiology of infectious hematopoietic necrosis virus in Alaska. *Diseases of Aquatic Organisms*, *40*(3), 163–76. http://doi.org/10.3354/dao040163

Emmenegger, E., & Kurath, G. (2002). Genetic Characterization of Infectious Hematopoietic Necrosis Virus of Coastal Salmonid Stocks in Washington State. *Journal of Aquatic Animal Health*, *14*(1), 25–34. http://doi.org/10.1577/1548-8667(2002)014<0025:GCOIHN>2.0.CO;2

Felsenstein, J. (2004). *Inferring Phylogenies*. Sunderland: Sinauer Associates, Inc.

Food and Agriculture Organization of the United Nations. (2008). *World Fisheries and Aquaculture*. *Aquaculture* (Vol. 35). Retrieved from ftp://ftp.fao.org/docrep/fao/011/i0250e/i0250e.pdf

Garver, K. A., Batts, W. N., & Kurath, G. (2006). Virulence comparisons of infectious hematopoietic necrosis virus U and M genogroups in sockeye salmon and rainbow trout. *Journal of Aquatic Animal Health*, *18*(4), 232–243. http://doi.org/10.1577/H05-038.1

Garver, K. A., Troyer, R. M., & Kurath, G. (2003). Two distinct phylogenetic clades of infectious hematopoietic necrosis virus overlap within the Columbia River basin. *Diseases of Aquatic Organisms*, *55*(3), 187–203.

Gilbert, M., Xiangmin, X., Chaitaweesub, P., Kalpravidh, W., Premashthira, S., Boles, S., & Slingenbergh, J. (2007). Avian influenza, domestic ducks and rice agriculture in Thailand. *Agriculture, Ecosystems & Environment*, (119), 409–415. http://doi.org/10.1016/j.biotechadv.2011.08.021.Secreted

Graham, J. P., Leibler, J. H., Price, L. B., Otte, J. M., Pfeiffer, D. U., Tiensin, T., & Silbergeld, E. K. (2008). The Animal-Human Interface and Infectious Disease in Industrial Food Animal Production: Rethinking Biosecurity and Biocontainment. *Public Health Reports*, *123*(3), 282–299.

Grischkowsky, R. S., & Amend, D. F. (1976). Infectious hematopoietic necrosis virus, prevalence in certain Alaskan sockeye salmon, Oncorhynchus nerka. *J Fish Res Board Can*, *33*, 186–188.

Groberg, W. J. (1983a). Priority research needs concerning fish viruses prevalent among Columbia River Basin salmonids. In J. C. Leong & T. Y. Barila (Eds.), *Proceedings of a Workshop on Viral Diseases of Salmonid Fishes in the Columbia River Basin, Special Publication* (pp. 159–167). Portland, OR: Bonneville Power Administration.

Groberg, W. J. (1983b). The status of viral fish diseases in the Columbia River Basin. In J. C. Leong & T. Barila (Eds.), *Proceedings of a Workshop on Viral Diseases of Salmonid Fishes in the Columbia River Basin, Special Publication*. Portland, OR: Bonneville Power Administration.

He, M., Ding, N.-Z., He, C.-Q., Yan, X.-C., & Teng, C.-B. (2013). Dating the divergence of the infectious hematopoietic necrosis virus. *Infection, Genetics and Evolution : Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases*, *18*, 145–50. http://doi.org/10.1016/j.meegid.2013.05.014

Healy, M. C. (1991). Life History of Chinook Salmon. In C. Groot & L. Margolis (Eds.), *Pacific Salmon Life Histories* (pp. 311–394). Vancouver, BC.

Ho, S. Y. W., & Duchêne, S. (2014). Molecular-clock methods for estimating evolutionary rates and timescales. *Molecular Ecology*, 5947–5965. http://doi.org/10.1111/mec.12953

Ho, S. Y. W., & Shapiro, B. (2011). Skyline-plot methods for estimating demographic history from nucleotide sequences. *Molecular Ecology Resources*, *11*(3), 423–434. http://doi.org/10.1111/j.1755-0998.2011.02988.x

Holsinger, K. E., & Weir, B. S. (2009). Genetics in geographically structured populations: defining, estimating and interpreting F(ST). *Nature Reviews. Genetics*, *10*(9), 639–50. http://doi.org/10.1038/nrg2611

Hudson, R., Boos, D., & Kaplan, N. (1992). A statistical test for detecting geographic subdivision. *Molecular Biology and Evolution*, *9*(1), 138–151. Retrieved from http://mbe.oxfordjournals.org/content/9/1/138.short

Jones, B. A, Grace, D., Kock, R., Alonso, S., Rushton, J., Said, M. Y., … Pfeiffer, D. U. (2013). Zoonosis emergence linked to agricultural intensification and environmental change. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(21), 8399–404. http://doi.org/10.1073/pnas.1208059110

Jukes, T., & Cantor, C. (1969). Evolution of protein molecules. In H. Munro (Ed.), *Mammalian Protein Metabolism* (pp. 21–132). New York: Academic Press.

Kapan, D. D., Bennett, S. N., Ellis, B. N., Fox, J., Lewis, N. D., Spencer, J. H., … Wilcox, B. A. (2006). Avian influenza (H5N1) and the evolutionary and social ecology of infectious disease emergence. *EcoHealth*, *3*(3), 187–194. http://doi.org/10.1007/s10393-006-0044-6

Kelley, G. O., Bendorf, C. M., Yun, S. C., Kurath, G., & Hedrick, R. P. (2007). Genotypes and phylogeographical relationships of infectious hematopoietic necrosis virus in California, USA. *Diseases of Aquatic Organisms*, *77*(1), 29–40.

Kingman, J. (1982). The coalescent. *Stochastic Processes and Their Applications*, *13*, 235–248.

Koepsell, T., & Weiss, N. (2003). *Epidemiologic methods: studying the occurence of illness*. New York: Oxford University Press.

Kosakovsky Pond, S. L., Frost, S. D. W., & Muse, S. V. (2005). HyPhy: Hypothesis testing using phylogenies. *Bioinformatics*, *21*(5), 676–679. http://doi.org/10.1093/bioinformatics/bti079

Kurath, G., Ahern, K. G., Pearson, G. D., & Leong, J. C. (1985). Molecular cloning of the six mRNA species of infectious hematopoietic necrosis virus, a fish rhabdovirus, and gene order determination by R-loop mapping. *Journal of Virology*, *53*(2), 469–476.

Kurath, G., Garver, K. A., Troyer, R. M., Emmenegger, E. J., Einer-Jensen, K., & Anderson, E. D. (2003). Phylogeography of infectious haematopoietic necrosis virus in North America. *The Journal of General Virology*, *84*(Pt 4), 803–814.

Lannan, C. N., Winton, J. R., & Fryer, J. L. (1984). Fish cell lines: establishment and characterization of nine cell lines from salmonids. *In Vitro*, *20*(9), 671–676. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=6542066

LaPatra, S. E., Parsons, J. E., Jones, G. R., & McRoberts, W. O. (1993). Early life stage survival and susceptibility of brook trout, coho salmon, rainbow trout, and their reciprocal hybrids to infectious hematopoietic necrosis virus. *Journal of Aquatic Animal Health*, *5*(4), 270–274.

LaPatra, S. E., Turner, T., Lauda, K. A., Jones, G. R., & Walker, S. (1993). Characterization of the humoral response of rainbow trout to infectious hematopoietic necrosis virus. *Journal of Aquatic Animal Health*, *5*(3), 165–171.

Li, W.-H. (1997). *Molecular Evolution*. Sunderland, MA: Sinauer Associates, Inc.

Lowry, T., & Smith, S. A. (2007). Aquatic zoonoses associated with food, bait, ornamental, and tropical fish. *Journal of the American Veterinary Medical Association*, *231*(6), 876–880. http://doi.org/10.2460/javma.231.6.876

Miller, M. A., Pfeiffer, W., & Schwartz, T. (2010). Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In *2010 Gateway Computing Environments Workshop, GCE 2010*. http://doi.org/10.1109/GCE.2010.5676129

Minin, V. N., Bloomquist, E. W., & Suchard, M. A. (2008). Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Molecular Biology and Evolution*, *25*(7), 1459–1471. http://doi.org/10.1093/molbev/msn090

Morzunov, S. P., Winton, J. R., & Nichol, S. T. (1995). The complete genome structure and phylogenetic relationship of infectious hematopoietic necrosis virus. *Virus Research*, *38*(2-3), 175–192.

Mulcahy, D. M., Tebbit, G. L., Groberg, W. J., McMichael, J. S., Winton, J. R., Hedrick, R. P., … Fryer, J. L. (1980). *The occurrence and distribution of salmonid viruses in Oregon*. Technical paper No. 5504, Oregon Agricultural Experiment Station, Corvallis, OR, ORESU-T-80-004.

Nei, M. (1982). Evolution of human races at gene level. In B. Bonne-Tamir (Ed.), *Human Genetics, Part A: The Unfolding Genome* (pp. 167–181). New York: Alan R Liss Inc.

New South Wales DPI Fish Health Unit. (2015). Aquatic animal disease and human health. Retrieved May 17, 2015, from http://www.dpi.nsw.gov.au/fisheries/pests-diseases/animal-health/fish-diseases-and-human-health

Nichol, S. T., Rowe, J. E., & Winton, J. R. (1995). Molecular epizootiology and evolution of the glycoprotein and non-virion protein genes of infectious hematopoietic necrosis virus, a fish rhabdovirus. *Virus Research*, *38*(2-3), 159–173.

Oliver, J. D. (2005). Wound infections caused by Vibrio vulnificus and other marine bacteria. *Epidemiology and Infection*, *133*(3), 383–391. http://doi.org/10.1017/S0950268805003894

Peeler, E. J., & Taylor, N. G. (2011). The application of epidemiology in aquatic animal health - opportunities and challenges. *Veterinary Research*, *42*(1), 94. http://doi.org/10.1186/1297-9716-42-94

Pepin, M., Bouloy, M., Bird, B. H., Kemp, A., & Paweska, J. (2010). Rift Valley fever virus (Bunyaviridae: Phlebovirus): An update on pathogenesis, molecular epidemiology, vectors, diagnostics and prevention. *Veterinary Research*, *41*(6). http://doi.org/10.1051/vetres/2010033

Pulliam, J. R. C., Epstein, J. H., Dushoff, J., Rahman, S. A., Bunning, M., Jamaluddin, A. A., … Daszak, P. (2012). Agricultural intensification, priming for persistence and the emergence of Nipah virus: a lethal bat-borne zoonosis. *Journal of The Royal Society Interface*, *9*(66), 89–101. http://doi.org/10.1098/rsif.2011.0223

R Core Team. (2013). R: A language and environment for statistical computing. In R Foundation for Statistical Computing (Ed.), . Vienna, Austria. Retrieved from http://www.r-project.org/

Rambaut, A. (2010). Path-O-Gen v1.4. *available from* http://tree.bio.ed.ac.uk/software/pathogen/

Roppel, P. (1982). Alaska Salmon Hatcheries 1891-1959. In *Alaska Historical Commission Studies no. 20*. Library of Congress #82-600591.

Rosenberg, N. A., & Nordborg, M. (2002). Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Reviews. Genetics*, *3*(5), 380–390. http://doi.org/10.1038/nrg795

Ross, A. J., Pelnar, J., & Rucker, R. R. (1960). A virus-like disease of chinook salmon. *Trans Am Fish Soc*, *89*, 160–163.

Rucker, R. R., Whipple, W. J., Parvin, J. R., & Evans, C. A. (1953). A contagious disease of sockeye salmon possibly of virus origin. *US Fish Wild Serv Fish Bull*, *54*, 35–46.

Rudakova, S. L., Kurath, G., & Bochkova, E. V. (2007). Occurrence and genetic typing of infectious hematopoietic necrosis virus in Kamchatka, Russia. *Diseases of Aquatic Organisms*, *75*(1), 1–11.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Subasinghe, R. P. (2005). Epidemiological approach to aquatic animal health management: Opportunities and challenges for developing countries to increase aquatic production through aquaculture. *Preventive Veterinary Medicine*, *67*(2-3 SPEC. ISS.), 117–124. http://doi.org/10.1016/j.prevetmed.2004.11.004

Suttle, C. A. (2007). Marine viruses--major players in the global ecosystem. *Nature Reviews. Microbiology*, *5*(10), 801–812. http://doi.org/10.1038/nrmicro1750

Tamura, K., & Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution*, *10*(3), 512–526. http://doi.org/10.1093/molbev/msl149

Tamura, K., Nei, M., & Kumar, S. (2004). Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(30), 11030–11035. http://doi.org/10.1073/pnas.0404206101

Tamura, K., Stecher, G., Peterson, D., Filipski, A., & Kumar, S. (2013). MEGA6: Molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution*, *30*(12), 2725–2729. http://doi.org/10.1093/molbev/mst197

Thoesen, J. (Ed.). (1994). *Suggested procedures for the detection and identification of certain finfish and shellfish pathogens* (4th ed.). Bethesda, MD: Fish Health Section, American Fisheries Society.

Troyer, R. M., & Kurath, G. (2003). Molecular epidemiology of infectious hematopoietic necrosis virus reveals complex virus traffic and evolution within southern Idaho aquaculture. *Diseases of Aquatic Organisms*, *55*(3), 175–185.

Troyer, R. M., LaPatra, S. E., & Kurath, G. (2000). Genetic analyses reveal unusually high diversity of infectious haematopoietic necrosis virus in rainbow trout aquaculture. *The Journal of General Virology*, *81*(Pt 12), 2823–2832.

US Department of the Interior Bureau of Reclamation. (2014, August). Columbia Basin Project. Retrieved from http://www.usbr.gov/pn/grandcoulee/cbp/

US National Park Service. (2015). Elwha River Restoration. Retrieved May 9, 2015, from http://www.nps.gov/olym/learn/nature/elwha-ecosystem-restoration.htm

Washington Department of Fish and Wildlife. (2015). Salmon Hatcheries Overview. Retrieved May 9, 2015, from http://wdfw.wa.gov/hatcheries/overview.html

Washington Department of Fish and Wildlife, & Oregon Department of Fish and Wildlife. (2002). *Status Report: Columbia River Fish Runs and Fisheries 1938 - 2000*.

Washington State Department of Ecology. (2015). Columbia River Facts and Maps. Retrieved May 9, 2015, from http://www.ecy.wa.gov/Programs/wr/cwp/cwpfactmap.html

Watson, S. W., Guenther, R. W., & Rucker, R. R. (1954). *A virus disease of sockeye salmon: Interim report* (Vol. 36; 138). U.S. Fish and Wildlife Service Spec. Sci. Rep.

Wingfield, W. H., Fryer, J. L., & Pilcher, K. S. (1969). Properties of the sockeye salmon virus (Oregon strain). *Proc Soc Exp Biol Med*, *130*, 1055–1059.

Wingfield, W., Nims, L., & Fryer, J. (1970). Species specificity of the sockeye salmon virus (Oregon strain) and its cytopathic effect in salmonid cell lines. In S. F. Snieszko (Ed.), *A Symposium on Diseases of Fish and Shellfishes* (pp. 319–326). Washington D.C.: American Fisheries Society.

Wolf, K. (1988). Fish viruses and fish viral diseases. Ithaca, New York, USA: Cornell University Press.

Wright, S. (1951). The Genetical Structure of Populations. *Annals of Eugenics*, *15*, 323–354.