Bridging Genomics and Preparedness: Quality Control Metrics and Analysis for Emerging and Circulating Avian Influenza in 2024-2025

By Christie Rose Woodside

B.A. in Computational Biology, May 2022, Colby College

A Thesis submitted to

The Faculty of
The Columbian College of Arts and Sciences
of The George Washington University
in partial fulfillment of the requirements
for the degree of Master of Science

May 18, 2025

Thesis directed by

Raja Mazumder
Professor of Biochemistry and Molecular Medicine

Vahan Simonyan
Adjunct Professor of Biochemistry and Molecular Medicine

Acknowledgments

The author wishes to acknowledge the following people: Emily Pennington for her help and knowledge of the Nearest Neighbor Pipeline and Sankey Diagrams. Raja Mazumder for his guidance, supervision, and mentorship throughout this thesis and program. My family, Lily, Sharry, and Michael, for supporting me and my dreams of becoming a scientist. William Schubert for being my rock throughout this whole program. Without all of you, this thesis and dream would not have been accomplished.

Abstract

The ongoing spread of highly pathogenic avian influenza (HPAI) H5N1 across a growing range of mammalian hosts emphasizes the urgent need for high-quality genomic data. This study presents a comprehensive quality control (QC) analysis of over 3,000 H5N1 genome assemblies collected from diverse host organisms across the United States, with a particular focus on cattle, poultry, and domestic animals. Leveraging a cloud-based computational pipeline from HIVE, sequence data were evaluated using standardized QC metrics, including GC content and Phred quality scores, to assess data integrity and identify potential sources of sequencing error or contamination. Regional differences in sequencing quality, such as elevated GC content and variable Phred scores in samples from California cattle, highlight inconsistencies in upstream sample handling and laboratory protocols. In addition to QC assessments, Sankey diagrams were used to visualize potential clonal diversity and segment-specific alignment across nearest neighbor pipeline-determined reference genomes. The results of this work provide a method to evaluate H5N1 sequences and address current gaps in influenza genomic data. These findings contribute to enhanced biosurveillance capacity and potential diagnostic uses and offer a scalable framework for evaluating future zoonotic threats.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

Chapter 1.1: Introduction

In recent years, the global landscape of infectious diseases has changed dramatically, with zoonotic pathogens posing a heightened threat to public health. The COVID-19 pandemic demonstrated the devastating impacts of zoonotic spillovers, making it clear that understanding these pathogens is important. The emergence of zoonotic diseases, where pathogens are transmitted from animals to humans, has gained renewed attention in the wake of several high-profile outbreaks since 2024. With constant alerts from the news and scientific journals on current and emerging health threats, it may seem that zoonotic pathogens are more prevalent today than ever. Around 75% of emerging infectious diseases are zoonotic and cause approximately one billion cases of illness in humans and millions of deaths each year, yet the heightened awareness of them appears to be more prevalent today[1,2]. It's not simply that these pathogens are rapidly mutating or adapting more efficiently to humans; rather, our habits, resource demands, and land-use practices disrupt delicate ecosystems, increasing our exposure to the microbial world[3].

Intensive farming is a breeding ground for zoonotic pathogens because of the high density of animals, genetic proximity, and increased immunodeficiency. Commercial farming can significantly increase the risk of spread, amplification, and mutation of pathogens within farming facilities, even with biosecurity efforts involved to reduce pathogen introduction[1]. There are currently nine prominent statutes established by

government agencies that address biosecurity in addition to multiple organizations that supply resources, such as the United States Department of Agriculture-Animal and Plant Health Inspection Service (USDA-APHIS) and One Health[4]. Regardless of efforts to mitigate the transmission of pathogens, antimicrobial resistance, and zoonoses persist. As of March 25, 2024, the U.S. Department of Agriculture announced that the H5N1 Avian Influenza clade 2.3.4.4b was found in dairy cattle on commercial dairy farms. As of April 8, 2025, there are 1,000 dairy herds, 168,257,048 poultry, 12,706 wild birds affected, and 70 human cases from exposure to infected animals. While there is currently no evidence of person-to-person transmission, the situation remains a significant cause for concern[5].

However, despite growing awareness of zoonotic risks, many of these pathogens remain insufficiently characterized, often due to limitations in genomic quality standards. Regulatory-grade genomes, essential for accurate detection, tracking, and preparedness, are still a challenge for many circulating and emerging zoonotic pathogens. The High-performance Integrated Virtual Environment (HIVE) is a storage and computing environment to handle next-generation sequencing data and metadata information[6]. This platform performs NGS analysis to produce high-quality QC metrics of the selected sequences. The core QC tables (biosampleMeta, assemblyQC, and ngsQC) contain the HIVE QC analysis results of the genomic sequences and their BioSample metadata. The addition of these pathogens will aim to enhance surveillance efforts and potentially support regulatory and diagnostic applications.

These efforts to sequence and provide quality-control metrics for zoonotic pathogens are vital for building robust surveillance systems and enhancing our preparedness against future outbreaks. As publicly available resources, public health agencies can better anticipate and respond to zoonotic diseases, helping to protect both human and animal health in an ever-evolving world.

Chapter 1.2: Influenza A Virus

The Influenza A virus belongs to the Orthomyxoviridae family and is an enveloped, segmented, single-stranded RNA virus with a negative sense, composed of 8 gene segments. Each gene segment encodes for 10 structural and at least 9 nonstructural/regulatory proteins used in viral replication and immune evasion[7]. The two most vital proteins for viral infection are hemagglutinin (HA) and neuraminidase (NA), which are two surface glycoproteins available for antibody (Ab) binding[8]. HA and NA surface proteins are used to classify Influenza A subtypes, such as H5N1 or H1N1. There are 18 known hemagglutinin (HA) subtypes and 11 neuraminidase (NA) subtypes identified in nature. Current influenza vaccines produce antibodies to respond to viral HA, yet during a natural infection, the body can also produce antibodies to target the viral NA[9].

Recombination in influenza A viruses occurs rapidly in vivo, facilitating the process of antigenic shift, where novel hemagglutinin (HA) and neuraminidase (NA) genes from the extensive animal reservoir are introduced into the human influenza A virome[10]. Antigenic shift occurs when multiple influenza viruses simultaneously infect a

host cell, leading to the reassortment of their genetic material, particularly the segments encoding hemagglutinin (HA) and neuraminidase (NA) proteins. This process generates novel hybrid viruses with surface antigens that differ significantly from those of previously circulating strains. As a result, the new viruses can evade existing immunity within the human population, increasing the potential for widespread infection and pandemic outbreaks.

These genetic shifts that formulate highly virulent strains of influenza are responsible for multiple global pandemics. The 1918 Spanish Flu (H1N1) is known to be one of the most deadly pandemics, with an estimated 500 million people or one-third of the world's population becoming infected with this virus[11]. Others include the 1957 Asian Flu (H2N2), the 1968 Hong Kong Flu (H3N2), and, more recently, the 2009 Swine Flu (H1N1), which was estimated to have caused around 151,700-575,400 deaths worldwide within its first year[12].

Given the devastating impact of past pandemics, global health organizations remain vigilant to prevent another widespread crisis. Protective measures, including the development and distribution of vaccines, have been critical in limiting the potential for such pandemics. However, in recent months, a new clade of the highly pathogenic avian influenza (HPAI) H5N1 strain has been on the rise, raising concerns about future outbreaks in humans and the ongoing need for proactive public health strategies.

Section 1: Influenza A H5N1 Virus

Subtype H5N1, a highly pathogenic avian influenza virus (HPAI), is a member of the influenza A virus family, which primarily infects birds but has zoonotic potential. First identified in 1996 in geese in Guangdong Province, China, H5N1 has since emerged as a significant public health concern due to its ability to cause severe disease symptoms in humans and high mortality rates among poultry. Unlike seasonal influenza viruses, H5N1 has a case mortality rate of approximately 60% in reported human cases, making it one of the most lethal zoonotic pathogens to date[13].

Although H5N1, a relatively recent strain of Influenza A, has been responsible for several significant outbreaks since its emergence in 1996, it has predominantly affected countries in Asia. While human-to-human transmission remained limited during these outbreaks, the mortality rate has been strikingly high. Notably, the 2003-2004 outbreaks in Vietnam, Thailand, and Indonesia resulted in 23 confirmed human cases, 18 of which were fatal[14,15]. Similarly, the 2006 outbreak in Indonesia saw 54 confirmed cases, with 41 fatalities, further highlighting the severity of the virus and its potential for causing widespread harm despite limited human-to-human transmission [16].

These past H5N1 outbreaks demonstrate the likelihood of severe and fatal consequences if a new H5N1 pandemic were to emerge. These outbreaks, predominantly in Asia, have demonstrated the virus's ability to cause significant human mortality, despite limited human-to-human transmission. Exuberantly high fatality rates, such as the 78% mortality in some outbreaks, highlight the risks that HPAI H5N1 poses to public health. Although human transmission is limited and has not been seen in the current

outbreak, the virus's ability to mutate and reassort means that the next strain could become more transmissible, increasing the potential for a global pandemic with each growing vector of transmission. These past experiences emphasize the importance of investing in preventative measures, including vaccines and surveillance programs, to mitigate the risk of a new pandemic.

Section 2: H5N1 2024 Outbreak

The status of the H5N1 outbreak as of April 10, 2025, is that the United States has seen no human-to-human transmission, but there have been reported cases of infected dairy cattle or birds to humans dating back to 2024. Findings by the NIH and CDC study suggest that the route of infection from cattle is most likely transmitted via respiratory droplets to dairy farmers[17] , and the sample virus from the study only caused mild symptoms in the person it was isolated from. Human infections from infected birds most often occur when there is close or lengthy contact without PPE or when the person has been in contact with the birds' saliva, mucus, or feces[18,19]. The results of a CDC study and reports from other human lab-confirmed cases show only mild respiratory symptoms or conjunctivitis, but there is worry that these cases could become more severe.

Section 3: Economic Impact of HPAI in the United States

The ongoing HPAI outbreak continues to exert significant economic pressure on both farmers and consumers across the United States. Since its emergence in early 2022, the H5N1 virus has led to the culling or death of approximately 168 million birds, drastically reducing the nation's poultry supply and disrupting the egg market[20]. As of

March 2025, the retail price of eggs has risen to $6.23 a dozen, marking a nearly 6% increase from the previous month and more than double the price from March 2024, according to the Bureau of Labor Statistics[21]. This price escalation has been attributed to the diminished egg-laying hen population, which stood at about 285 million in March 2025 compared to 315 million before the outbreak[22]. In response to the crisis, the U.S. Department of Agriculture (USDA) announced on February 26, 2025, that it is planning to invest $1 billion in research aimed at curbing the spread of H5N1 and stabilizing the egg market[23]. As Brooke Rollins, the U.S. Secretary of Agriculture, outlined, "$500 million for biosecurity measures, $400 million in financial relief for affected farmers, and $100 million for vaccine research, action to reduce regulatory burdens, and exploring temporary import options." This additional funding is an attempt to mitigate the impact of H5N1 on the agricultural economy and egg prices, but ultimately, investing in strengthening biosecurity measures and controlling the spread will be the most effective in reducing its impact.

Despite these efforts, the economic toll remains. By November 2024, the outbreak had already cost the U.S. taxpayers approximately 1.4 billion dollars, primarily in indemnity and compensation payments to farmers for their culled flocks[24]. In addition to this sum, consumers bore an estimated $1.41 billion burden in 2024 due to elevated egg prices, with expectations of continued financial strain through 2025 as industry workers attempt to replenish their poultry stock and diminish the spread[25].

Furthermore, the outbreak has compelled the United States to import eggs from international sources to increase the limited supply, relying on egg imports from countries like Brazil and several European countries. Brazilian egg imports have surged,

becoming a critical source to alleviate the shortage, but will most likely be affected by the current tariffs in place[26]. Additionally, increased imports from Europe have provided some relief, particularly during peak demand periods such as Easter, reflecting a shift towards international suppliers to stabilize the U.S.'s consumer markets[27].

The economic consequences of the ongoing H5N1 outbreak illustrate how zoonotic pathogens impact more than just public health; they deeply disrupt agricultural economies and consumer markets. As demonstrated by the soaring egg prices, widespread culling, and significant government expenditures to alleviate the burden, zoonotic diseases know no bounds, directly affecting the livelihoods, businesses, and ultimately, the finances of the average consumer. Effective disease management, aided by better agricultural practices and biosecurity measures, is therefore essential, not only to safeguard human and animal health but also to alleviate far-reaching economic repercussions.

Section 4: Vectors of Transmission

Zoonotic spillovers play a critical role in influenza ecology due to the virus's broad host range and frequent jump between species[28]. Wild aquatic birds serve as the natural reservoir for all the influenza subtypes, and avian or swine influenza can occasionally infect humans directly or via an intermediate host[28]. Transmission occurs either through direct animal-to-human contact or via viral reassortment within what is termed a "mixing vessel" species, or organisms that are capable of harboring both avian and human influenza viruses. Historically, swine are the primary mixing vessel given that they express receptors compatible with both avian and human influenza viruses in their

respiratory tract. A recent study indicates that cattle also exhibit this receptor[29], potentially facilitating the current transmission of H5N1 from dairy cattle to humans. Such cross-species events with these "mixing vessel" species are significant sources of novel influenza strains with pandemic potential.

Ducks play a central role in the transmission dynamics of avian influenza, particularly H5N1, serving as the main natural reservoir capable of infecting both domestic poultry and mammals[30]. The virus strains are often carried by ducks asymptomatically, shedding high viral loads from their respiratory and gastrointestinal tract with few or no disease signs[31]. Although ducks, specifically mallards, can shed high levels of H5N1 with no signs of illness, therefore facilitating a stealthy transmission, in contrast, infected chickens will suffer a 100% mortality[32]. Findings from an experimental infection study suggest that migratory duck species can act as long-distance vectors of H5N1, dispersing the virus along migratory flyways[33]. Consequently, wild birds act as both a natural reservoir (harboring influenza diversity) and as mobile vectors, disseminating H5N1 over long distances to new regions and initiating new outbreaks in poultry and wildlife populations.

Historically, cattle were considered resistant to influenza A/B/C viruses and had not been prominently involved in influenza outbreaks. Until recently, the 2.3.4.4b clade has caused frequent spillovers into diverse wild and domestic animals, such as geese, foxes, and skunks, leaving cattle herds unaffected[34]. However, in March of 2024, the first confirmed case of H5N1 emerged in a Texas dairy cow, subsequently spreading amongst other herds in Texas and Kansas, eventually affecting the broader dairy industry. Infected

cattle showed signs of reduced feed intake, respiratory distress, diarrhea, and a sudden drop in milk yield with abnormal milk[35,36]. Pathological exams revealed that H5N1 infects the bovine mammary gland, targeting epithelial cells in the lining of the udder, with the virus consistently detected in milk samples from infected cows[36]. In one significant case, healthy cows transferred from an infected farm to a distant farm initiated new infections, indicating potential asymptomatic carriers that can spread H5N1 between herds[36]. Although the initial infections were likely contracted from wild birds, infected cows subsequently transmitted H5N1 to other species on the premises, including cats and raccoons, who were found dead and tested positive for the virus. A recent study published in Nature confirmed through clinical, pathological, and molecular findings across nine affected dairy farms that the 2.3.4.4b clade had adapted to replicate in cows and direct transmission among cattle[36]. Current research efforts are heavily focused on how the virus invades cattle herds, how it efficiently spreads in herd settings, and what biosecurity strategies can mitigate this novel transmission route[37]. Ultimately, the emergence of H5N1 within cattle herds is an indicator of the necessity of better surveillance across atypical hosts.

In summary, the growing prevalence of zoonotic pathogens, such as H5N1, emphasizes the critical importance of enhanced genomic surveillance, biosecurity measures, and high-quality genomic data to better understand and reduce emerging health threats. The expansion and industrialization of agricultural practices, combined with ecosystem disruption, have significantly increased the risk of zoonotic spillovers, as evidenced by the ongoing H5N1 avian influenza outbreak. Effectively studying and

managing these pathogens requires rigorous genomic characterization and robust-quality control frameworks, such as those provided by the HIVE platform. By providing publicly accessible, high-quality pathogen data, this study can help enhance genomic surveillance capabilities, potential vaccine development, and diagnostic use cases, which in turn will strengthen global preparedness, minimizing the threat zoonotic diseases pose to human and animal health, as well as economic stability.

Chapter 2: Methods

To investigate the prevalence and genomic characteristics of zoonotic pathogens, this research employed a comprehensive workflow combining next-generation sequencing (NGS) analysis, quality control (QC) metrics, and database integration. The primary aim was to generate high-quality genomes with a variety of QC metrics for H5N1 to support surveillance and public health preparedness. Data processing and analysis were conducted using the High-performance Integrated Virtual Environment (HIVE) platform, and the results were curated. Other resources, such as GISAID and NCBI, were utilized in selecting the pathogens and providing genomic metadata information. The following sections detail the steps involved in pathogen selection, genomic sequencing, quality control, and data integration.

The United States Department of Agriculture-Animal and Plant Health Inspection Service (USDA-APHIS) uploaded an NCBI BioProject 1102327 in April 2024 containing assemblies and sequence data for H5N1 subtype clade 2.3.4.4b. A total of 3,528 genomic assemblies containing SRA Read files and BioSample information were selected from this BioProject from February 27, 2025, and prior. As of March 1, 2025, peridomestic animals such as Grackle, Domestic Cats, Canada Goose, Raccoon, Turkey, Goose, Chicken, and Snow Goose were added to the BioProject, which mainly housed cattle. These genomic accession IDs were downloaded from NCBI and saved into a TSV file, which was then used to input the information into HIVE for QC. Many genomic assemblies not found within this BioProject were uploaded to GISAID, while their associated BioSample and SRA information was somehow reported into BioProject

1102327. In this case, BioSample and SRA IDs were computed together along with the GISAID genome, which was manually curated and uploaded into HIVE for further QC.

The USDA Wildlife Services uploaded another NCBI BioProject, 1122849, to provide surveillance of peridomestic animals on USDA-confirmed premises that have been affected by H5N1 clade 2.3.4.4b genotype B3.13. This BioProject is related to the APHIS BioProject PRJNA1102327 directly. As of January 20, 2025, Assembly IDs, BioSample IDs, and SRA IDs for House Mice, House Sparrows, Raccoon, American Robin, and Eurasian Collared Dove have been uploaded to this BioProject. Given that this BioProject is an extension of the APHIS H5N1 in dairy cattle, and this thesis seeks to provide high-quality genomes of circulating pathogens, I thought it was necessary to include these genomes of peridomestic animals in the data table results.

GISAID (a Global Initiative on Sharing Avian Influenza Data)[38] is a data-sharing platform, containing multiple databases, aimed at sharing published and 'unpublished' influenza data. GISAID has now expanded to include databases for Pox (EpiPox), Respiratory Syncytial Virus (RSV) (EpiRSV), and COVID-19 (EpiCoV) genomes[39]. At the beginning of this study, 24 genome assemblies of H5N1 clade 2.3.4.4b were collected for dairy cattle through the EpiFlu database. The selection criteria outlined that the genomes must include all 8 genomic sequences and have been found and collected in the United States. The 24 genomes were selected because of the mentioned selection criteria and because their BioSample ID and reads could be found in NCBI. In addition to sequences, GISAID provides tracking of variants, frequency dashboards, articles, and

news updates. These tools were utilized to identify circulating pathogens and to prioritize specific pathogens for detailed analysis.

GISAID publishes a monthly update titled "H5N1 Bird Flu continues to take its toll in the United States" directly on its front page[40]. The update provides a comprehensive overview of the current infection trends and statistical counts. This update includes subsampled phylogenetic trees specifically for HA and NA gene segments, focusing on the current H5N1 samples. The March 2, 2025, summary was reviewed to select representative samples from each phylogenetic tree for inclusion in the nearest neighbor pipeline. Samples were selected based on the following criteria: 1) the sample is positioned at the terminal node (leaf) of the phylogenetic tree, 2) it is marked as "Newly Added", and 3) the sequencing reads are publicly available on NCBI. These selected samples were fashioned into a dataset and used to illustrate how the nearest neighbor pipeline in HIVE can be used to investigate current H5N1 samples to identify any mutations or other notable genomic differences.

The High-performance Integrated Virtual Platform (HIVE)[41] is a high-throughput, cloud-based distributed computing infrastructure designed for the storage, analysis, and management of genomic and biological data. It contains a multitude of tools that can be applied to Next-Generation Sequencing (NGS) analysis, such as sequence alignment, sequence profiling tools, metagenomics analyzers, phylogenetic tree-building tools using NGS data, clone discovery algorithms, and recombination analysis[41]. The QC Pipeline in HIVE is an automated tool that features a one-click pipeline approach. The pipeline takes

in the SRA read files, Genome Assembly ID or uploaded genome file, and BioSample ID, and calls the necessary tools for QC analysis. These tools are HIVE-Heptagon[42], HIVE-Hexagon[43], and the QC Post-Alignment Quality Control tool. Results are provided in individual JSON format files corresponding to their QC analysis: qcAll(assemblyQC), BioSample (biosampleMeta), and ngsAll (ngsQC).

HIVE-Hexagon is a high-performance DNA sequence alignment tool developed to efficiently process the extensive data generated by NGS technologies. HIVE-Heptagon is a sequence profiling tool integrated into the HIVE environment that specializes in base-calling and single-nucleotide polymorphism (SNP) detection, simultaneously generating comprehensive quality assessments and noise profiles. The computational output of these two tools plays a critical role in generating the QC attributes associated with the ngsQC and assemblyQC data tables. The Hexagon and Heptagon tools are integrated within the automated QC pipeline. Heptagon and Hexagon were utilized as standalone tools in an alternate version of HIVE to provide Sankey diagrams for Figures 3, 4, and 5.

CensuScope[44] is a tool designed by HIVE Lab that will quickly identify all the various organisms present in the metagenomic NGS dataset. It then generates a detailed report that shows the classification, such as species or broader taxonomic groups, of the NGS data that was input. This tool was utilized to analyze a range of recent H5N1 cattle samples from GISAID. The organisms that were found to be the most similar taxonomically to the samples were selected as reference genomes for downstream

analysis in Heptagon and Hexagon. All influenza A organisms that were present within the CensuScope output for the sample were used for the analysis.

Multiple Python scripts were created to convert each of the JSON files from the HIVE computations into TSV-formatted tables. They then proceed to call the NCBI Entrez API through its research command-line tool to retrieve metadata for the remaining missing values of the table[45]. These scripts and other information can be found at the GitHub[46].

To visualize quality control metrics from H5N1 surveillance samples, I used ObservableHQ[47], a JavaScript-based data notebook platform. I uploaded a TSV file named *aphis_data_for_figures.tsv,* which contained the QC data that was downloaded and reformatted, to the web platform and focused on metrics such as GC content and Phred scores to assess sequence quality. Using Observable's built-in support for D3.js, I created interactive graphs to compare quality metrics across states and host species, which can be seen in Figures 1 and 2. The platform's reproducible notebook format enabled efficient exploration of the data and clear presentation of trends in sample quality of the H5N1 samples.

Chapter 2.1: Data Table Construction

The quality control (QC) attributes of the H5N1 samples are organized into three datasets: *H5N1_assemblyQC*, *H5N1_ngsQC*, and *H5N1_biosampleMeta*. Each dataset

adheres to the current QC data dictionary, which defines the column headers, QC metrics, and other relevant attributes.

These three datasets were run through a Python script to generate an Excel file containing only the QC metrics needed for producing Figures 1 and 2. The metrics selected are ‘assembly_gc_content’, ‘phred_average’, and ‘average_ngs_gc_content’, and their values represent the average across all segments for each individual flu assembly. For instance, ‘assembly_gc_content’ reflects the mean GC content calculated from the total number of segments of a given influenza sample. When selecting the SRR reads to be used in the column ‘average_ngs_gc_content’, only reads that used Illumina technology were selected.

After generating the complete Excel sheet for use in Observable using the script *aphis_data_for_figures.py*, a thorough cleanup was necessary. The "host" column pulls directly from the BioSample metadata of each sample, which varies widely in naming conventions. For example, several host names appeared in all caps, such as *Cattle*, *Duck*, *Chicken*, *Turkey*, *Hawk*, *Goose*, *Cat*, *Rock Pigeon*, *Western Gull*, *Goat*, *Tiger*, *Mountain Lion*, *Serval*, *Lynx*, *Bobcat*, *Lion*, and *Fox*. These were standardized using Excel’s Find and Replace tool, converting them to properly capitalized common nouns for consistency.

To further streamline the dataset and reduce redundancy in categorical values, similar host organisms were grouped under a single name. For instance, *domestic-cat* and *feline* were both renamed to *Cat*, as they represented only a few entries. Similarly, *red fox* was simplified to *Fox*, and *Eurasian Collared Dove* and *White-winged Dove* were both

categorized as *Dove*, due to their low representation. *Red-Tailed Hawk* was shortened to *Hawk*, and *Mallard* was grouped under *Duck*, given the small number of records for each specific name.

Chapter 3: Results

As the scientific community continues to monitor the spread of H5N1 from avian species to mammals and other animals, the necessity for preparedness and effective surveillance becomes increasingly apparent. NCBI and other data repositories constantly receive large amounts of data, some of which is not of great quality, making it a challenge to sort through and find the most reliable information. In the APHIS BioProject that was used as the sole data source for this project, the small number of inconsistencies in the data did not go unnoticed. The quality control (QC) metrics generated through the High-performance Integrated Virtual Environment (HIVE) pipeline offer critical insights into the genomic integrity and suitability of the selected H5N1 sequences for regulatory and surveillance applications. These QC attributes provide essential information for accurately interpreting genomic data, ensuring reliability in downstream analyses, and confirming adherence to quality standards. However, beyond these standard QC metrics, further comparative insights are provided by the Sankey diagrams, which visualize the sequence alignment and relationships between different genomic assemblies. By examining these Sankey diagrams, we can effectively identify genomic similarities and divergences between samples, illustrating potential transmission pathways and evolutionary patterns that traditional QC metrics alone may not fully capture. Thus, the integration of traditional QC attributes with visual analytical tools such as Sankey diagrams enhances our understanding of the genomic dynamics of emerging zoonotic pathogens like H5N1.

Chapter 3.1: Sample Distribution

To contextualize the scope of the host infection for avian influenza, I compiled and analyzed the distribution of assembled H5N1 genomes and reads by host organism, which can be seen in Table 1. Cattle and poultry (Chicken and Turkey) represent the bulk of the samples, consistent with the known reservoirs and the primary focus of agricultural surveillance. However, a notable number of sequences were derived from a range of wild and domestic mammals and birds, which was explained earlier in the methods section. The wide host diversity highlights the virus's zoonotic potential and suggests possible cross-species transmission events. The low-frequency hosts may reflect rare spillover events, passive surveillance, or early detection in the species before spread.

Host 1	Samples 1	Host 2	Samples 2
Cattle	2776	Alpaca	3
Chicken	369	House-Mouse	2
Cat	113	Blackbird	2
Turkey	82	American Robin	2
Tiger	28	Dove	2
Goat	19	Sparrow	2
Serval	18	Goose	2
Fox	15	Western Gull	2
Mountain Lion	15	Red Tailed Hawk	2
Duck	11	Great Horned Owl	1
Bobcat	8	Peregrine Falcon	1
Lynx	8	Western Sandpiper	1
Canada Goose	6	Common Raven	1
Skunk	6	American Wigeon	1
Snow Goose	5	Mute Swan	1
Lion	5	White-Faced Ibis	1
Rock Pigeon	5	Hawk	1
Grackle	4	Bald Eagle	1
Raccoon	3	House Sparrow	1
American Crow	3	**Total**	**3,528**

Table 1. *Distribution of assembled H5N1 genomes across host organisms from APHIS BioProject's 1102327 and 1122849*. This table summarizes the number of H5N1 genome assemblies QC'd grouped by host species, highlighting the breadth of host infection. The wide host diversity underscores the virus's zoonotic potential and highlights the current concern for cross-species transmission events.

Chapter 3.2: QC Results and Comparison

In this study, I conducted a QC analysis on as many H5N1 surveillance samples as possible to assess the general sequence quality and provide metrics for high-quality genomes that have been published from this past year, available to researchers. GC content and Phred score values are quality indicators in genomics and provide different, but quality insights into the biological validity and technical reliability of the sequence data.

Chapter 3.3: GC Content and Phred Score Analysis Across the Host Samples

GC content is characteristic of each organism across its genome and is a great technical quality indicator. Deviations from the expected GC content value can suggest sequencing errors, contamination, or sample degradation. Since all the samples used in the figures were sequenced with Illumina technology, this minimizes the risk of cross-platform sequencing bias while still allowing for the detection of potential contamination or other data quality issues. This consistency in the data enhances confidence in using GC content to screen for anomalies and evaluate overall sequencing integrity. Any notable outliers in GC content across the samples may therefore point to issues requiring further investigation and deeper analysis of the sequence itself.

A wide range of host organisms are represented within the APHIS BioProject, including chickens, cattle, geese, and even lions and tigers, as shown in Table 1. Figure 1 a) displays the assembly GC content (as a percentage) for each H5N1 sample, grouped by the host organism, excluding cattle samples, which is analyzed further in the results. A

threshold level of 43.6% (red line) represents the known GC content value for the reference genome, Influenza A Puerto Rico 1934 H1N1. All data points fall above the threshold, indicating that H5N1 flu subtypes may exhibit higher GC content compared to H1N1 subtypes. However, this observation alone does not imply improved quality or functional significance. Most GC content values shown in the figure cluster between 45.45% and 46.5%, which aligns with previously reported values in recent literature[48,49]. However, several outliers fall outside this range, with some approaching 50% and others closer to 44%. Elevated GC content could indicate contamination from a GC-rich organism, mislabeling of samples, bad trimming, or adapter contamination. For example, among the chicken samples, seven sequences cluster close to 50%, very unlikely relevant given that 362 samples are tightly grouped around 46%. These warrant a further investigation into their GC content if necessary. Conversely, the unusually low GC content outliers may suggest degraded sequence quality, over-trimming, misassembly, or missing sequence regions. Cat samples also show a concerning variation, where two points are seen at 44% and five near 49.7%, while the majority (105 samples) fall within the expected range of 45.5-46%. These large deviations in GC content raise concerns about data integrity and suggest that these sequences may not accurately represent the real viral genomic material or could potentially be biologically invalid. A deeper investigation of these outliers would need to be conducted before they are considered significant.

Figure 1 a) Genomic assembly GC Content (%) of H5N1 sequences grouped by Host Organism. Each point represents a single assembled H5N1 genome (average GC content of the total segments). All data points remain well above the reference genome GC content threshold (red line). Most GC values cluster between 45.45% and 46.5%, consistent with previously reported values for H5N1[48,49]. Outliers falling above or below this range may indicate contamination, misassembly, or sequencing artifacts.

Notable deviations are observed in chicken, cat, and turkey samples, where several assemblies approach 50% or drop below 44%, raising concerns about data quality and requiring further investigation.

b) Average genomic assembly Phred quality scores of H5N1 samples grouped by host organism. Each point represents a single assembled H5N1 genome. All samples exceed a Phred score of Q30, indicating high base call accuracy (≥99.9%) suitable for downstream analysis. Despite consistent use of Illumina sequencing technology, the data reveal a bimodal distribution: one group ranges from ~32.5 to 33.25, and the other from ~35.25 to 36.4. This pattern suggests potential batch effects or variability in sequencing protocols across APHIS reporting laboratories. Lower Phred scores, primarily affecting chicken, cat, and turkey samples, may reflect poorer library preparation, degraded RNA, or field/post-mortem sample collection. In contrast, goat and skunk samples exhibit exceptionally high Phred scores (≥36), indicating well-prepared libraries and strict QC measures.

Figure 1 b), displays the average assembly Phred scores by host organism, revealing a notable bimodal distribution. All samples exceed a Phred score above a Q30, indicating a base call accuracy of 99.9%, the standard threshold for high-quality sequencing, making them suitable for any downstream analysis. The samples in this figure were all sequenced using Illumina technology, ensuring platform consistency in the data. There appear to be two distinct groups: one between 32.5-32.25 and another between 35.25-36.4. While the lower groups are still acceptable within quality standards, they hover near the threshold for high-confidence data, suggesting potential quality concerns. These differences could be due to poor library preparation or degraded RNA, particularly given the biosurveillance context; samples could have been collected in the field or post-mortem. There is also a possibility of lab-to-lab variability across the various APHIS reporting laboratories, which could contribute to the observed distribution. Notably, samples from chickens, cats, and turkeys mostly fall into the lower Phred score range, indicating consistent quality limitations within those batches. In contrast, goat samples fall within the 35.75-36.6 range, and skunk samples reach as high as 36.8-37.0, showing excellent sequencing quality in these two host sequences. The higher scores likely result from well-prepped libraries, strict QC pipelines, and effective adapter trimming or error correction. However, even though the data appears to form two distinct clusters, this separation alone may not be biologically meaningful and could simply reflect technical or procedural differences. Overall, the most likely cause for the bimodal distribution is batch effects or sequence collection and protocol variability between submitting laboratories.

Chapter 3.4: GC Content and Phred Score Analysis of Cattle H5N1 Samples

The APHIS BioProject PRJNA1102327 was established as a biosurveillance initiative to track the emergence and spread of H5N1, particularly within cattle populations, by depositing sequence data into NCBI. Many of the sequences originate from states with confirmed outbreaks, reflecting areas of heightened surveillance. To evaluate the quality and consistency of these sequences, I analyzed both the GC content and average Phred quality scores of H5N1 assemblies from cattle samples, grouped by the state of sample collection. This comparison allows for an assessment of sequencing quality and sample consistency by geographic region, as well as insights into the distribution of the QC values among the affected cattle populations.

To explore potential regional differences in sequence quality, Figure 2 a) presents the average genomic assembly GC content (%) of H5N1 samples derived from cattle, grouped by state. Same as Figure 1 a), a threshold value of 43.5% is shown as a red line and most of the data samples fall within the expected 45.5% - 46.5% GC content range, consistent with the established literature. However, a notable cluster of samples from California, approximately 70 data points in total, exhibit elevated GC content between 49.5%–50%, which may indicate a systemic issue. While any isolated outliers are easy to ignore and dismiss as noise, a large group of samples deviating from the standard suggests a pattern of error. Possible explanations could be batch effects, cross-sample contamination of the samples, differences in submitting lab protocols, or misassembly due to incorrect reference mapping. Misannotation or host mislabeling seems less likely

in this case, given the heightened surveillance and investigation surrounding these cattle herds, which likely involved careful and detailed sample documentation. These kinds of systematic anomalies could affect how accurately the virus is represented in the data and may lead to misleading results in things like phylogenetic trees or outbreak tracking. Since H5N1 in cattle raises concerns about potential spread to humans, the sequence data must be reliable. Overall, these samples should be further evaluated to determine the cause of their unusually high GC content and potentially excluded from downstream analysis if deemed unreliable.

a)


Genomic Assembly GC Content % →
50.0
49.5
49.0
48.5
48.0
47.5
47.0
46.5
46.0
45.5
45.0
44.5
44.0
43.5
CA CO IA ID KS MI MN MT NC NM OH SD TX UT WY
State Abbreviation

b)

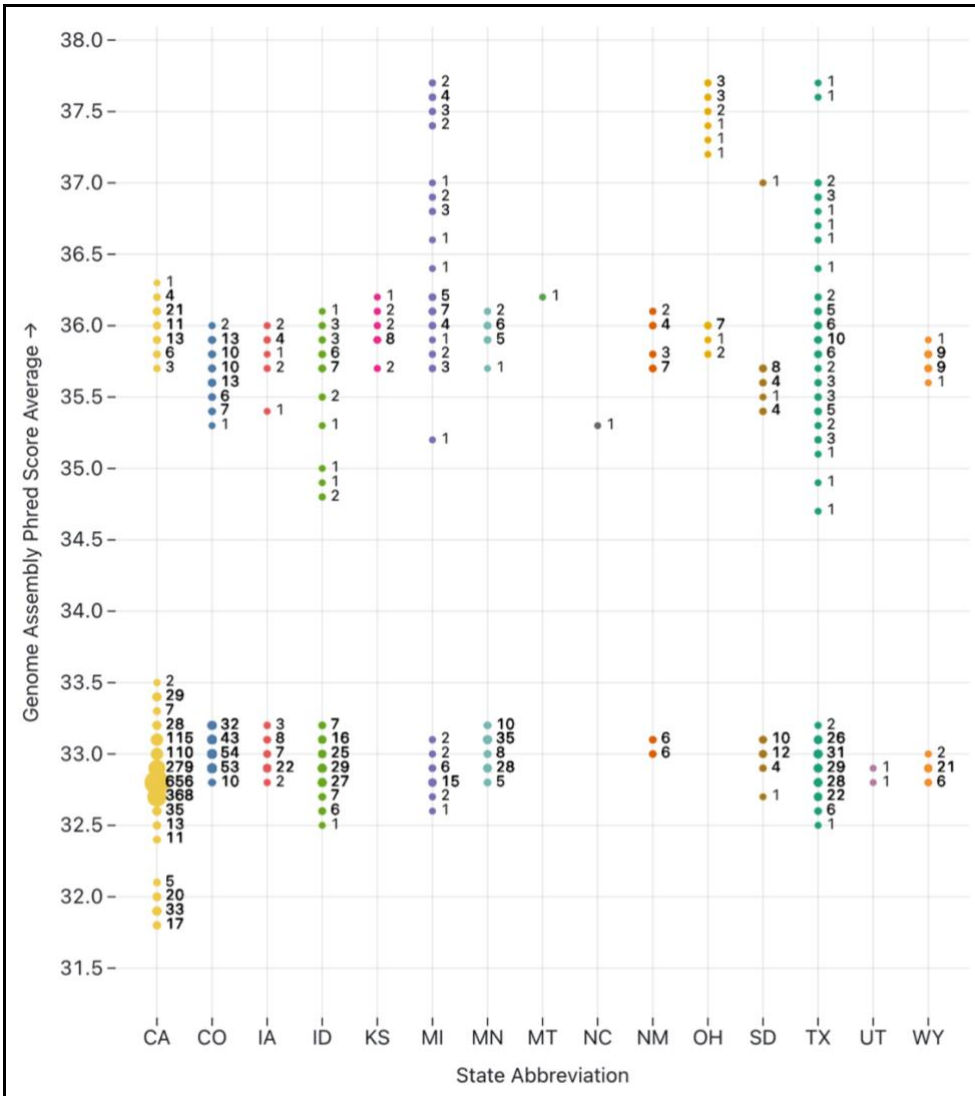

Figure 2 a) *Genomic assembly GC Content average (%) for H5N1 cattle samples, grouped by state.* Each data point represents a single H5N1 genome assembly from a cattle sample (average GC content of the total segments), colored and categorized by state location. All data points remain well above the reference genome GC content threshold (red line). Most samples fall within the expected GC content range of 45.5%–46.5%, consistent with typical avian influenza genomes. However, a distinct cluster of approximately 70 samples from California shows elevated GC content between 49.5%–50%, which may indicate a systemic issue such as batch effects, submitting lab protocol variability, contamination, or misassembly. While other states also show occasional outliers, the concentration in California suggests a pattern that warrants further investigation.

b) *Average genomic assembly Phred quality scores of H5N1 cattle samples, grouped by state.* Each point represents a single genome assembly. A clear bimodal distribution is observed, with one group ranging from ~32–33.5 and

another from ~35–37.75, despite all samples being generated on Illumina platforms. This may reflect differences in sequencing models, sample preparation, or laboratory protocols. California displays a unique trimodal pattern, possibly due to contributions from multiple submitters within the state. Texas also shows a wide spread of scores, while states like Ohio and Missouri cluster tightly at the higher end, suggesting more standardized, high-quality inputs. These patterns highlight potential batch effects and emphasize the importance of consistent sequencing practices and detailed metadata reporting in biosurveillance efforts.

Figure 2 b) illustrates a clear bimodal distribution in genome assembly Phred score averages, with one group ranging between ~32-33.5 and the other ranging between ~ 35-37.75. All samples exceed a Phred Score of 30 or higher, indicating a 99.9% base call accuracy, with some nearing 38. Despite all samples being sequenced on Illumina platforms, this display of distinct groups may reflect differences in sequence models (MiSeq vs. NovaSeq vs. iSeq 100), sample preparation, and submission protocols across the APHIS laboratories. California cattle samples appear to show three groups: ~ 32.0, ~33, and ~36, respectively. This trimodal distribution could be due to data contributions from multiple sources or labs within the state. Unfortunately, due to limited BioSample metadata, it is difficult to determine which APHIS labs submitted these samples, how they were collected, or any other contextual information that might explain the observed clustering. Texas exhibits a particularly widespread range with a range of ~32.5-37+, which may be biologically relevant. While all samples were processed through the same HIVE QC pipeline, ruling out post-sequencing workflow differences, the variation in the data could stem from differences in viral load, RNA quality, or sample condition (live Cow versus post-mortem). In contrast, cattle samples from Ohio and Missouri cluster tightly near the higher end (~35.7-37.5), suggesting consistently high-quality input material. This may reflect a coordinated response during the H5N1 surge in February 2025, when rigorous and standardized protocols were likely prioritized. These patterns underscore the importance of consistent sequencing practices and metadata reporting when interpreting genomic surveillance data. Understanding the sources of variation,

whether technical or biological, is crucial to providing accurate conclusions about avian influenza spread and sample quality.

Chapter 3.5: Explanation of Nearest Neighbors and HIVE Pipeline

To further investigate the H5N1 samples, a nearest-neighbor pipeline in HIVE was used to identify an appropriate set of reference genomes, which were then used to generate Sankey diagrams that visualize both sequence alignment and coverage. A set of recently curated H5N1 samples from GISAID was selected from the website's 'In Focus' HPAI update. Details regarding the GISAID dataset curation are provided in the methods section. Raw sequencing reads from these samples were uploaded to the HIVE platform, where a suite of quality control graphs was generated for each read file. For data pre-processing, the read quality was assessed using two key visualizations: *Quality Position Count* (Figure 3 A) and *Quality Length Count* (Figure 3 B). Reads were considered high quality if they achieved a Phred score of Q30 or higher and were at least 150 nucleotides in length. Reference genomes for alignment were selected using CensuScope, which identifies the most taxonomically similar organism present within the NGS data. The current reference genome for H5N1, *A/Goose/Guangdong/1/96(H5N1)*, was not identified among the nearest neighbors (NN) in the CensuScope results due to limitations in the SlimNT database used at the time of the analysis. As a result it was added alongside the other Influenza A organisms present, which included: *Influenza A virus (A/Puerto Rico/8/1934(H1N1))*, Influenza *A virus (A/Korea/426/1968(H2N2)), Influenza A virus (A/California/07/2009(H1N1)), Influenza A virus (A/Shanghai/02/2013(H7N9)),* and *Influenza A virus (A/New York/392/2004(H3N2))*. For this study, alignments were

focused on two key Influenza A genomic segments: HA and NA. The two individual segments from each nearest neighbor were downloaded and combined into a comprehensive genome; *HA_genomes_censuscope.fasta* and *NA_genomes_censuscope.fasta*. The reads were aligned to this set of nearest neighbors using the HIVE-Hexagon sequence alignment tool. Alignment parameters used for HIVE-Hexagon are detailed in the referenced publication[50] as well as in the Appendix. Reads that aligned to the reference genomes are shown in a pie chart, color-coded by reference genome (Figure 3 C). Because multiple nearest neighbor references were included in the genome file, a multiple sequence alignment was performed using MAFFT[51] to generate a comparative consensus. The MAFFT file was subsequently used in the HIVE-Hexahedron step, enabling visualization and subpopulation analysis. HIVE-Hexahedron presents these results as interactive Sankey diagrams (Figure 3 D), which simultaneously show the depth of coverage per position, alignment coverage, and subpopulation structure across the genome.

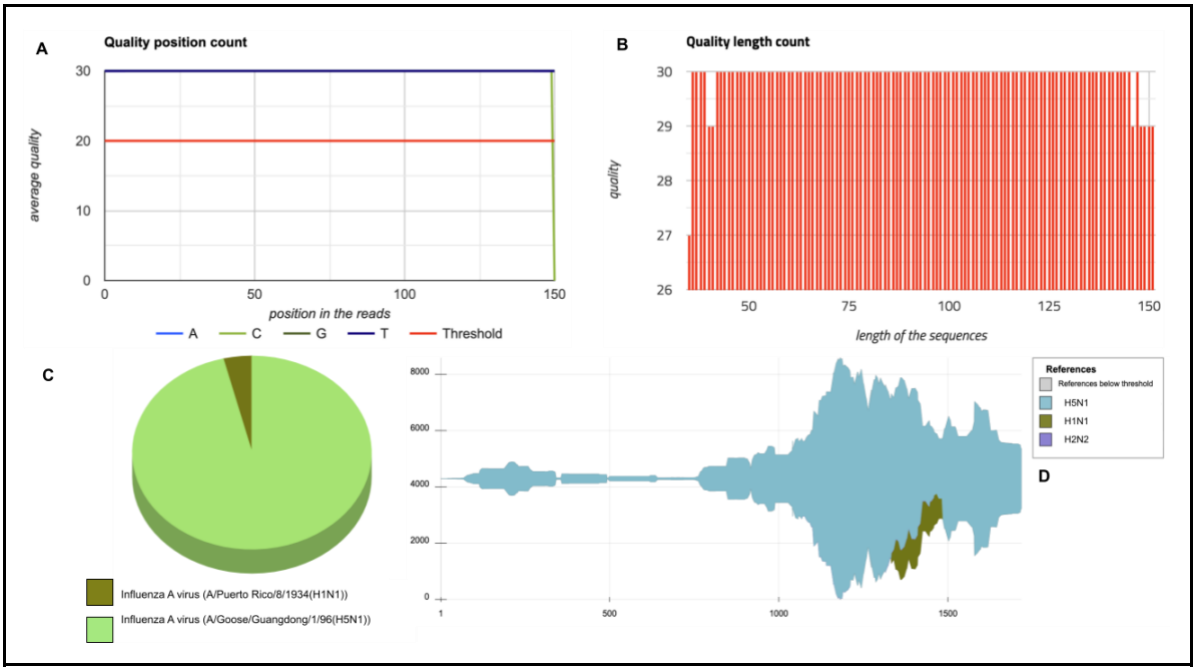

Figure 3. *Explanatory panel figure of HIVE Pipeline.* The general pipeline for nearest neighbor pipeline demonstrated on H5N1 read *SRR32512945*, sample name *(A/dairy_cow/USA/004632-003/2025)*. a) Quality Position Count plot showing average Phred scores at each base position in the sequencing reads, separated by nucleotide (A, C, G, T). All base positions remain well above the Q20 threshold (red line) and appear to be at Q30, indicating high-quality reads across their full length. b) Quality Length Count plot illustrating average Phred quality scores relative to read length. Consistently high values above Q30 are observed across the entire 150 nucleotide read length. c) Pie chart summarizing read alignment proportions to reference genomes identified using CensuScope. Most reads aligned to Influenza A virus (A/Goose/Guangdong/1/96(H5N1)), with a smaller proportion mapping to Influenza A virus (A/Puerto Rico/8/1934(H1N1)), reflecting the dataset's curation focus on H5N1.
d) HIVE-Hexahedron Sankey Diagram showing coverage depth and subtypes across a concatenated composite of HA segment genomes from each reference genome. Only three out of six reference genomes (or rather subtypes) showed successful read alignment, represented by the three colors displayed. This visualization enables simultaneous assessment of variant frequency, depth of alignment, and genomic segment coverage.

Chapter 3.6: Exploration of H5N1 Samples Using the Nearest Neighbor Pipeline

The HIVE pipeline described above enables the identification and confirmation of possible mutations, detection, and analysis of sub clonal populations[50], and the generation of high-quality assemblies from NGS reads, the latter of which will be detailed further in

the Future Case section. The curated GISAID dataset includes a broader collection of newly added Influenza A (H5N1) samples derived from subsampled phylogenetic trees for the HA and NA segments specifically. For illustrative purposes, two representative samples, *SRR32415278* and *SRR32512945*, were selected from this dataset and mapped to their respective phylogenetic trees for both HA and NA. To evaluate alignment quality and population structure, raw reads were analyzed against their respective segment genome files using HIVE-Hexagon and population analysis was conducted using HIVE-Hexahedron.

Figure 4 displays the resulting Sankey diagrams for each sample against the HA segment genome. The x-axis represents the nucleotide position along the HA segment, while the y-axis indicates sequencing depth at that nucleotide position. Both samples exhibit substantial alignment to the *A/Goose/Guangdong/1/96(H5N1)* reference genome (light blue), with the peak coverage observed between positions 1040 to 1473. However, *SRR32512945* displays a significantly higher depth of coverage, peaking at 8,463 reads compared to *SRR32415278*, which peaks at 1,942. In both Sankey diagrams, the alignment to the *A/Puerto Rico/8/1934(H1N1)* reference (olive green) is restricted to just nucleotide positions 1333-1485, suggesting only partial homology to this region. Regions with low coverage appear as flattened areas rather than distinct peaks, as seen between position ~500-762, indicating that the sample reads did not align to these reference genomes at these positions.

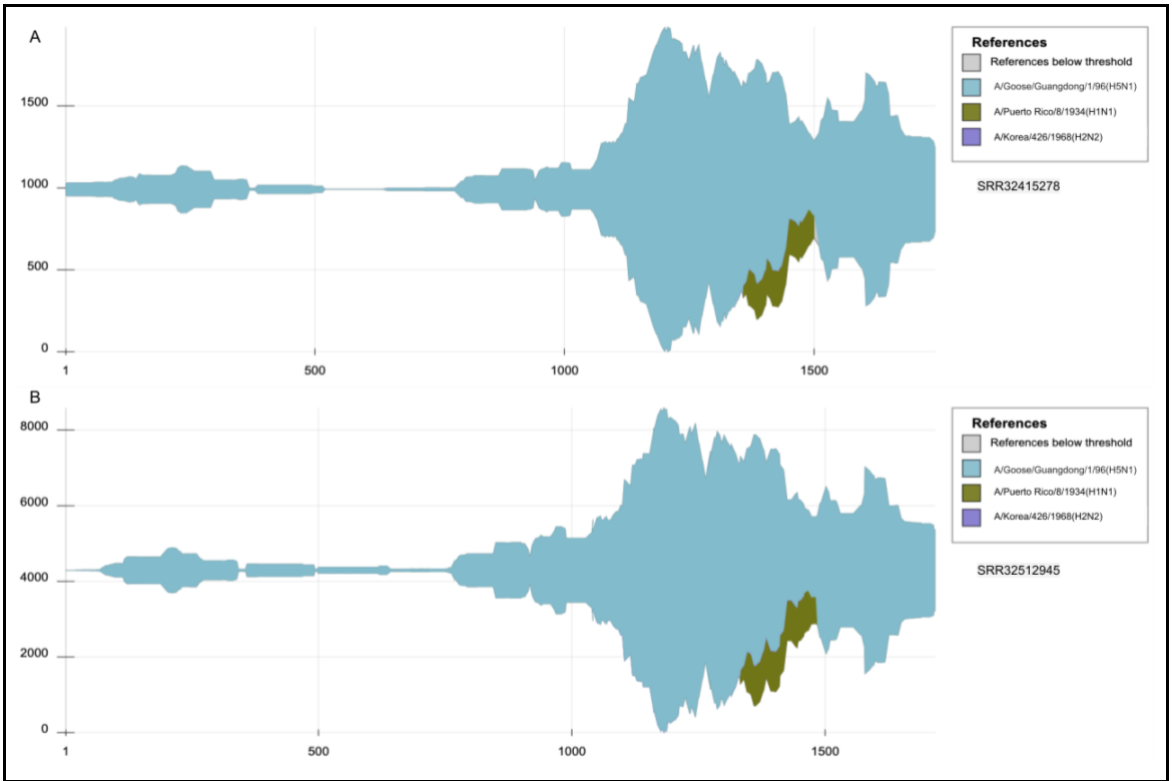

Figure 4. *HA segment genome Clonal Analysis using Sankey Diagrams.* Parameters were set to 500 length, 50 coverage, and 5 support[50]. Sankey diagrams generated using HIVE-Hexahedron depict the clonal structure and coverage depth of sequencing reads aligned to the HA segment genome for two representative H5N1 samples: SRR32415278 and SRR32512945. The x-axis indicates nucleotide position across the HA segment, while the y-axis represents sequencing depth. A) SRR32415278 shows moderate alignment to A/Goose/Guangdong/1/96(H5N1) (light blue), with a peak coverage of 1,942 reads between nucleotide positions 1040–1473. B) SRR32512945 displays higher coverage depth across the same region, peaking at 8,463 reads between nucleotide positions 1040–1473. In both diagrams, limited alignment to A/Puerto Rico/8/1934(H1N1) (olive green) is observed between positions 1333–1485, indicating partial homology to this segment. These visualizations provide insight into clonal representation and segment-specific alignment quality, highlighting substantial differences in sequencing depth between samples.

Next, the same samples were aligned to the NA genomes, shown in Figure 5 A and B. In both samples, substantial alignment to the *A/Goose/Guangdong/1/96(H5N1)* reference genome, now represented in olive green. Notably, sample *SRR32512945* aligned to the *A/California/07/2009(H1N1)* genome across most of the segment, with

only three small regions of unalignment: positions 0-86, 1067-1198, and 1417-1443. In contrast, the sample showed more limited alignment to this reference, with coverage only appearing between positions 531-1101 and 1213-1416, indicating a lack of representation. Alignment to the *A/Puerto Rico/8/1934(H1N1)* genome was inconsistent for both samples, with significant regions of missing coverage. Comparatively, to the HA genome analysis, *SRR32512945* again demonstrated a significantly higher depth of coverage, peaking at 5,279 reads compared to maximum reads of 1,309 for *SRR32415278*. Scattered regions of low coverage across the diagram may result from deletions and/or large insertions at these sites, as HIVE-Hexahedron filters out reads with substantial indels near the ends, resulting in a low depth of coverage at these regions[50].

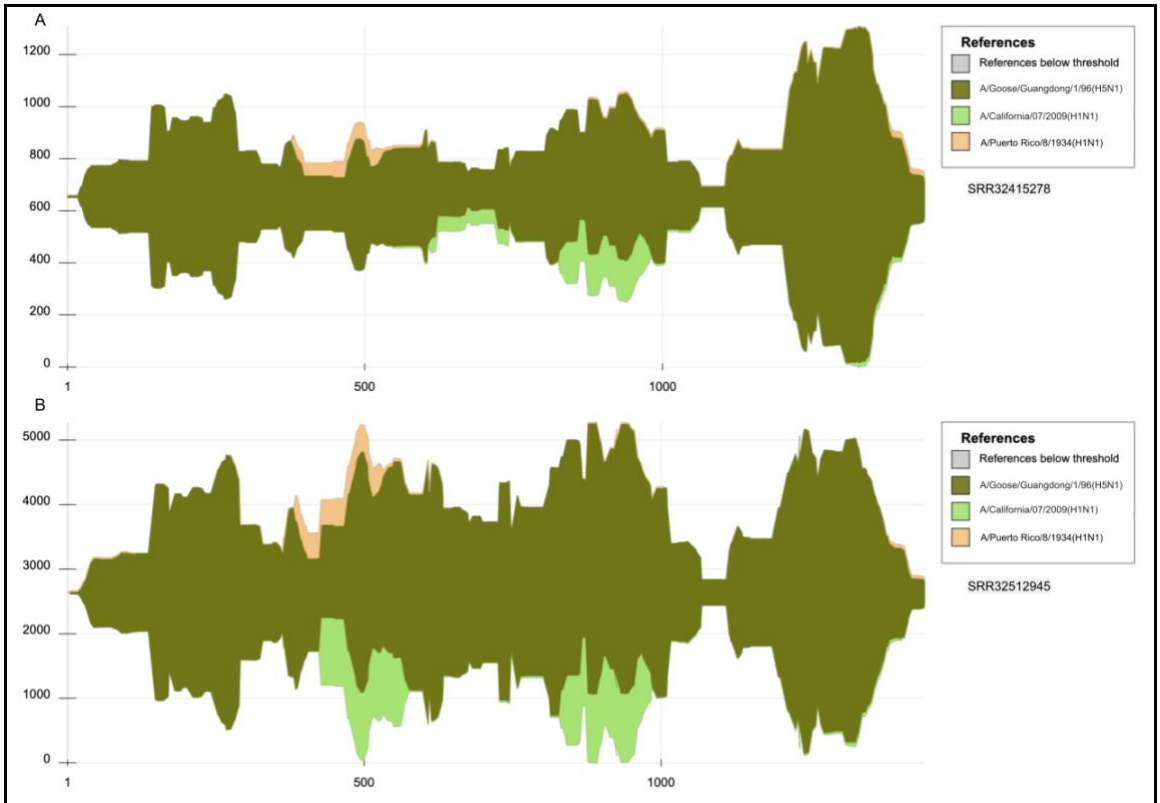

Figure 5. *NA segment genome Clonal Analysis using Sankey Diagrams.* Parameters were set to 500 length, 50 coverage, and 5 support[50]. Sankey diagrams generated via HIVE-Hexahedron display the alignment of sequencing reads from H5N1 samples

SRR32415278 and SRR32512945 to the neuraminidase (NA) segment reference genomes. A) SRR32415278 shows substantial alignment to A/Goose/Guangdong/1/96(H5N1) (olive green), with limited and fragmented alignment to A/California/07/2009(H1N1) and A/Puerto Rico/8/1934(H1N1). B) SRR32512945 exhibits broader and more continuous alignment to the same references, particularly A/California/07/2009(H1N1), with only three narrow regions lacking coverage (positions 0–86, 1067–1198, and 1417–1443). Peak depth of coverage for SRR32512945 reached 5,279 reads, significantly higher than SRR32415278's peak of 1,309. Scattered areas of low coverage in both samples may be attributed to large insertions or deletions, as HIVE-Hexahedron filters out reads with substantial end indels, resulting in decreased coverage in these regions[50].

This nearest neighbor pipeline enables the generation of Sankey Diagrams for evaluating the clonal structure and sequencing depth across the reference genomes. As demonstrated in all four Sankey diagrams, several regions exhibit a lower depth of coverage observed. To observe a greater depth of coverage to the alignment, there is a need for improved references. While the alignment to *A/Goose/Guangdong/1/96(H5N1)* was consistent, partial or fragmented alignment to other H1N1 references suggests the presence of genomic divergence or underrepresented subtypes. The results highlight how Sankey diagrams, when used in conjunction with the nearest neighbor alignment approach, can help identify coverage gaps and suggest the need for a more comprehensive or subtype-specific reference genome, particularly for emerging Influenza A strains such as H5N1.

Chapter 4: Discussion

The ongoing highly pathogenic avian influenza (HPAI) H5N1 outbreaks have revealed critical vulnerabilities in existing surveillance and agricultural systems. The virus's rapid spread across a growing range of mammalian hosts has raised concerns about the pandemic potential of current zoonotic pathogens. While current data suggests limited human infections and is primarily linked to direct contact with animals, the increasing number of cross-species transmission events is deeply concerning. As demonstrated by the COVID-19 pandemic, the gap between sporadic zoonotic spillover and sustained human-to-human transmission can close rapidly, often outpacing public health responses. Despite the ongoing surveillance efforts and agricultural biosecurity measures, our ability to detect, respond to, and prepare for emerging threats remains hindered by inconsistent genomic quality standards, incomplete metadata, and the absence of unified reporting frameworks. The limited availability of high-quality genomic sequences restricts our capacity to detect early transmissions and prepare vaccines to combat the potential spread.

By integrating sequencing data from public repositories, curating relevant BioSample metadata, and applying standardized QC analyses, this study contributes a set of high-quality genomic resources that further enhance the utility of the H5N1 surveillance data.

Chapter 4.1: Sequence Quality Across Hosts

Across the majority of the host organisms, the sequences exhibited GC content values within the expected range of 45.5% to 46.5%, which is consistent with prior literature[48,49]. However, there were a few notable outliers, particularly among chicken and cat samples, that ranged from 44% to nearly 50%. This suggests potential issues such as contamination, degraded RNA, or misassembly. Similarly, although the Phred scores exceeded the Q30 quality threshold across all samples, a bimodal distribution was seen. The lower scores are more common amongst the chicken and cat samples, potentially reflecting inconsistent library preparation or degraded field samples. Overall, the majority of selected H5N1 sequences demonstrated moderate to high quality, and with the addition of standardized QC metrics, they could potentially be used for regulatory use and vaccine diagnostic applications.

For the H5N1 sequences derived from infected cattle, the majority exhibited GC content within the expected range of 45.5% to 46.5%, still consistent with known values for avian influenza viruses. However, there was an interesting observation in a subset of samples from California, where approximately seventy assemblies were clustered at an elevated GC content of 49.5 to 50%. This deviation from the group raises concerns about potential misassembly or contamination and suggests a possible batch effect linked to a specific submitting laboratory in California. In terms of the Phred score quality, all cattle-derived sequences exceeded the Q30 threshold, indicating reliable base call accuracy suitable for downstream analysis. However, the data revealed a bimodal distribution, and in some states, a trimodal distribution, of the Phred scores. The California samples

displayed three distinct clusters, suggesting variability in sample preparation, varying sequencing protocols, or data submission practices. In contrast, states like Missouri and Ohio showed tightly clustered, high-quality Phred scores groups, indicating consistent sequencing inputs and standardized workflows. These patterns underscore the importance of rigorous QC assessment and metadata transparency, especially when they could be used for regulatory or diagnostic purposes.

Chapter 4.2: Regional Differences in Genomic Data

Significant regional differences in genomic quality were observed, particularly in cattle-derived H5N1 sequences from California, which exhibited unusually high GC content (49.5–50%) and a trimodal distribution of Phred quality scores. These observed differences are most likely the result of batch effects, sample contamination, or inconsistent sequencing protocols across different submitting laboratories. This contrasted with the sequences from Missouri and Ohio, which demonstrated high-quality scores amongst their tight clusters, suggesting standardized and reliable workflows. This regional variability stresses that there are critical gaps in biosurveillance and/or inconsistent communication between submitting labs. Even when centralized QC pipelines like HIVE are applied, the quality of the upstream sample collection, sequencing, and processing ultimately determines the results. Systemic anomalies in California, which is currently the state most affected by this virus outbreak, reduce confidence in sequence integrity. Without detailed and standardized metadata accompanying the genomic submissions, it becomes difficult to trace and address the sources of these discrepancies. With these instances alone, this presents a serious concern

for the use of such data in public health applications. The results of the QC analysis suggest reliable data, but moving forward, efforts to harmonize the biosurveillance protocols across states, in addition to extensive BioSample metadata, will be necessary to ensure reliable genomic data. That said, most sequences analyzed across all regions exhibited high Phred scores and QC content within the expected range, reflecting overall strong data quality fit for a multitude of use cases.

Chapter 4.3: Visualization of Genomic Relationships (Sankey Diagrams)

Visualizing the genomic data allows for meaningful insight beyond what is represented in the QC metrics. The Sankey diagrams that were created using HIVE-Hexahedron helped clarify genomic relationships between the Influenza A samples and the HA and NA reference genomes by showing where the reads aligned, the depth of coverage, and the range of alignment. The diagrams made it easy to evaluate the alignment quality between the selected H5N1 samples and their nearest neighbor reference genomes. The diagrams highlighted both the well-covered regions and areas with incomplete or fragmented alignment, which could indicate mutations or genetic differences. Although not all the reference genomes were mapped to the reads, it helped show which references were most appropriate to fit the sample.

While more diagrams across additional samples could have been included in the analysis, the ones selected were representative of the currently sampled genomes, and adding more would likely have shown diminishing returns unless unique patterns or known mutations were present. Across the board, most reads aligned best to the H5N1-specific reference Influenza A *(A/Goose/Guandong/1/96(H5N1))* rather than the common

and standard Influenza A reference *(A/Puerto Rico/8/1934(H1N1))*. Still, some regions lacked coverage entirely, suggesting that even the current H5N1 reference genome may not fully represent the circulating strains and that an updated or more diverse selection of references may be needed.

Overall, the Sankey diagrams were a helpful tool not just for explaining the pipeline functionality but also for exploring sequence diversity, identifying alignment gaps, and guiding future improvements in reference genome selection.

Chapter 4.4: Comparison with Existing Literature

Most H5N1 genome assemblies analyzed exhibited high sequence quality, with GC content values falling within the expected range of 45.5%-46.5% and Phred scores above Q30, indicating reliable base call accuracy suitable for downstream applications. However, regional differences in sequencing quality were observed, where California displayed broader variability and signs of systemic differences, while others, such as Missouri and Ohio, produced consistently high-quality data. The HIVE platform, combined with the nearest neighbor pipeline, enabled a comprehensive QC and visual analysis, illustrating how clonal structure and alignment depth can be leveraged to investigate the HA and NA segments of H5N1 samples. These findings reinforce and build upon existing studies, offering insight into the genomic quality and surveillance of currently circulating H5N1.

To avoid future outbreaks that could have severe implications for both animal and human health, there is a pressing need for enhanced biosurveillance strategies and improved agricultural practices. More specifically, this study emphasizes the critical need for high-quality sequencing data to support rapid detection, track viral evolution, and inform public health responses to minimize the spread. Until recently, Influenza A infections in cattle have been rarely documented[52], resulting in a lack of genomic data, particularly raw read data, from this species and other hosts. Research on HPAI in avian species such as ducks and poultry is more extensive than in cattle[53], but is still limited overall, which has restricted the availability of usable genomic data. However, a recent preprint documents the interstate spread of HPAI H5N1 in U.S. dairy cattle, confirming a major host jump and emphasizing the urgency for better surveillance across mammalian reservoirs[54].

In addition to the QC analysis outlined in the results, a key deliverable of this study is the generation of an extensive set of regulatory-grade Influenza A H5N1 sequences that can support current biosurveillance efforts and diagnostic applications. The Influenza Research Database[55], a U.S. NIAID-sponsored database designed to support influenza virus research, has yet to incorporate recent H5N1 subtype data, despite its urgent relevance. The availability of these H5N1 sequences with robust QC metrics fills this gap and complements existing resources provided by the Bacterial and Viral Bioinformatics Resource Center (BV-BRC), offering researchers a complete and more actionable dataset for this evolving pathogen.

GISAID is a widely recognized platform for sharing HPAI genomic data and has since expanded to include SARS-CoV-2, RSV, and other viral pathogens. It was originally established to facilitate rapid access to H5N1 sequences, particularly considering reluctance from some countries to deposit data in publicly available archives like NCBI and EMBL[56,57]. While data availability has improved in recent years, GISAID continues to serve as a primary source for HPAI assemblies. However, the absence of raw sequencing reads on the platform limits the ability to perform in-depth quality assessments using additional metrics. By integrating assemblies from GISAID with raw reads from NCBI, the findings of this study contribute to and extend GISAID's mission of advancing global influenza surveillance.

Unlike prior studies, which primarily focus on phylogenetic or epidemiological tracking, this study uniquely integrates high-resolution QC metrics with genomic surveillance tools such as HIVE-Heptagon and HIVE-Hexagon. These platforms enable deeper interrogation of subclonal populations and segment-specific coverage, which is rarely applied to H5N1 studies at scale. While similar tools have been successfully implemented for quasispecies detection in other viral contexts, such as HIV[50,58], their application to influenza segment analysis, particularly within the context of regulatory-grade sequencing, remains limited. This work demonstrates how such pipelines can be adapted to address the complexities of H5N1 genomic data and serve as a model for Influenza A or broader pathogen analysis.

Collectively, this study addresses current gaps in H5N1 genomic surveillance but also shows how incorporating QC frameworks can be built into pathogen monitoring or diagnostic pipelines. By coupling a pipeline used for analyzing regulatory-grade sequencing with tools for visualization and subclonal analysis, this work offers a clear, repeatable approach for generating high-quality genomic data. These contributions are especially important now, as H5N1 continues to spread across new hosts, affecting the agricultural industry. Having reliable genomes with detailed QC metrics will be critical for future outbreak response, vaccine development, and improving overall preparedness.

Chapter 4.5: Limitations and Future Work

One of the main limitations of this study is the limited availability of the Influenza A genomic sequences that are paired with raw sequencing reads, which are essential inputs into the HIVE QC pipeline and performing the quality control analysis. While some assemblies are available through platforms like GISAID and NCBI, many lack the accompanying reads needed to assess sequencing accuracy, coverage depth, or detect potential artifacts. Due to data access restrictions and publication limitations, this study could not include H5N1 genomes isolated from human hosts, which further narrows the scope of cross-species comparison. In many cases, human-derived sequences either do not have public raw read data or are subject to privacy and data-sharing policies that restrict their use in public analyses. Having these as potential reference genomes from human hosts would not only strengthen the dataset but also allow for direct comparisons across host species to explore potential genomic differences from one mammal to another. Applying the same QC framework and analysis pipeline to human-derived

sequences could offer insights into cross-species transmission and help refine public health strategies. As a result of the restriction of human samples, the dataset was limited to non-human hosts, primarily poultry, cattle, and peridomestic animals.

Fortunately, a BioProject was identified that contained all necessary components, raw sequencing reads, assembled genomes, and BioSample metadata, allowing for the QC assessment of over 3,000 genomic samples. While a few low-quality sequences were flagged during the analysis, the majority met high-quality thresholds. However, integrating ~3,000 genomes into a reference database would be impractical and redundant. To address this, our group is developing a scoring algorithm[59] to identify a representative subset of sequences from the samples that capture the diversity and quality of the full dataset. The algorithm begins by calculating the pairwise similarities between genomes using the k-mer overlap approach to generate a distance matrix, which is then visualized as a graph. By applying a similarity threshold, the graph is pruned to retain only highly similar connections, displaying a clearer subnetwork structure. Next, the method selects a minimal subset of genomes such that every genome is either in the subset or directly connected to one. Among the possible subsets, the algorithm will then choose the subset that maximizes predefined QC metrics, either individually or through a composite multi-QC score derived using a principal component analysis (PCA). This algorithmic approach helps ensure that the representative set captures the broadest and most relevant QC metrics from the original dataset.

Currently, a complementary pipeline is under development to enable assemblies to be generated from raw sequencing reads in cases where assemblies are not available. The process begins by identifying the candidate sequences to construct a comprehensive pangenome reference, followed by clustering and selecting representative sequences at a defined 95% similarity threshold. The representatives will serve as the basis for generating a multiple sequence alignment, which is in turn used to construct clonal templates using HIVE-Hexahedron. Sequencing reads are then aligned to these templates, and high-confidence assemblies are computed through consensus calling and variant resolution. However, for the cattle-derived H5N1 samples used in this study, the available data consisted of pre-assembled genomes, but the corresponding raw reads were the ones missing. Tools and pipelines capable of generating synthetic reads will be needed to address these cases. An extension of the complementary pipeline could be adapted to meet this need.

Given the continual collection of HPAI samples, the growing demand for high-quality H5N1 genomic data, and the ongoing development of both the representative selection algorithm and the synthetic assembly pipeline, this study lays the foundation for numerous future directions. These efforts will support a more refined approach to genomic curation, including the ability to downsample large datasets into high-quality representative sequences, which is especially helpful in the case of Influenza data. Together, these tools can fix and expand upon the current limitations of this study to provide a stronger and more comprehensive dataset of Influenza A data.

Chapter 5: Conclusions

To address the critical gaps in HPAI genomic surveillance and the limited availability of high-quality H5N1 genomic data, this study provided a comprehensive dataset of high-quality sequences with a robust set of QC metrics that can be applied in research for both biosurveillance and diagnostic use cases. By systematically evaluating over 3,000 genomic assemblies across a diverse set of hosts, this study revealed substantial variability in sequencing quality, notably regional discrepancies suggestive of laboratory-specific batch effects or inconsistent protocols. The rigorous QC framework developed and applied significantly enhances the reliability of the genomes, where they will be publicly shared, thereby strengthening global preparedness against zoonotic threats.

Despite limitations, such as restricted access to human-derived sequences and poorly detailed metadata, the methodologies established provide a replicable approach for ensuring genomic data integrity and utility. Future work should prioritize a set of representative samples for the dataset, incorporating human genomic sequences when and if possible, and expanding these QC pipelines to other emerging and circulating zoonotic pathogens. Ultimately, the availability of this genomic data and its GC metrics will be indispensable for informed public health responses, accurate tracking of viral evolution, and potential vaccine development to help safeguard both human and animal health against future pandemics.

References

1 Espinosa, R., Tago, D. & Treich, N. Infectious Diseases and Meat Production. *Environ Resour Econ (Dordr)* **76**, 1019-1044 (2020). https://doi.org:10.1007/s10640-020-00484-3

2 Karesh, W. B. *et al.* Ecology of zoonoses: natural and unnatural histories. *Lancet* **380**, 1936-1945 (2012). https://doi.org:10.1016/S0140-6736(12)61678-X

3 Marie, V. & Gordon, M. L. The (Re-)Emergence and Spread of Viral Zoonotic Disease: A Perfect Storm of Human Ingenuity and Stupidity. *Viruses* **15** (2023). https://doi.org:10.3390/v15081638

4 Morris, G., Ehlers, S., Aaltonen, P. M., Sheldon, E. & Johnson, A. Review of livestock biosecurity resources and trainings: Local, state, federal, and international organizations. *Journal of Biosafety and Biosecurity* **5**, 162-169 (2023). https://doi.org:10.1016/j.jobb.2023.12.001

5 (CDC), C. f. D. C. a. P. H5 Bird Flu: Current Situation. (2024).

6 Simonyan, V. *et al.* High-performance integrated virtual environment (HIVE): a robust infrastructure for next-generation sequence data analysis. *Database (Oxford)* **2016** (2016). https://doi.org:10.1093/database/baw022

7 Muramoto, Y., Noda, T., Kawakami, E., Akkina, R. & Kawaoka, Y. Identification of novel influenza A virus proteins translated from PA mRNA. *J Virol* **87**, 2455-2462 (2013). https://doi.org:10.1128/JVI.02656-12

8 Kosik, I. & Yewdell, J. W. Influenza Hemagglutinin and Neuraminidase: Yin(-)Yang Proteins Coevolving to Thwart Immunity. *Viruses* **11** (2019). https://doi.org:10.3390/v11040346

9 Chen, Y. Q. *et al.* Influenza Infection in Humans Induces Broadly Cross-Reactive and Protective Neuraminidase-Reactive Antibodies. *Cell* **173**, 417-429 e410 (2018). https://doi.org:10.1016/j.cell.2018.03.030

10 Ince, W. L., Gueye-Mbaye, A., Bennink, J. R. & Yewdell, J. W. Reassortment complements spontaneous mutation in influenza A virus NP and M1 genes to accelerate adaptation to a new host. *J Virol* **87**, 4330-4338 (2013). https://doi.org:10.1128/JVI.02749-12

11 (CDC), C. f. D. C. a. P. *History of 1918 Flu Pandemic*, <https://archive.cdc.gov/www_cdc_gov/flu/pandemic-resources/1918-commemoration/1918-pandemic-history.htm> (2018).

12 (CDC), C. f. D. C. a. P. *2009 H1N1 Pandemic (H1N1pdm09 virus)*, <https://archive.cdc.gov/www_cdc_gov/flu/pandemic-resources/2009-h1n1-pandemic.html> (2019).

13 Poovorawan, Y., Pyungporn, S., Prachayangprecha, S. & Makkoch, J. Global alert to avian influenza virus infection: from H5N1 to H7N9. *Pathog Glob Health* **107**, 217-223 (2013). https://doi.org:10.1179/2047773213Y.0000000103

14 Conly, J. M. & Johnston, B. L. Avian influenza - The next pandemic? *Can J Infect Dis Med Microbiol* **15**, 252-254 (2004). https://doi.org:10.1155/2004/121394

15 (CDC), C. f. D. C. a. P. *Outbreaks of Avian Influenza A (H5N1) in Asia and Interim Recommendations for Evaluation and Reporting of Suspected Cases --- United States, 2004*, <https://www.cdc.gov/mmwr/preview/mmwrhtml/mm5305a1.htm#:~:text=During%20December%202003%2D%2DFebruary,authorities%20in%20Thailand%20and%20Vietnam.> (2004).

16 Sedyaningsih, E. R. *et al.* Epidemiology of cases of H5N1 virus infection in Indonesia, July 2005-June 2006. *J Infect Dis* **196**, 522-527 (2007). https://doi.org:10.1086/519692

17 Health, N. I. o. Study of H5N1 from an infected person. (2024).

18 Garg, S. *et al.* Highly Pathogenic Avian Influenza A(H5N1) Virus Infections in Humans. *N Engl J Med* **392**, 843-854 (2025). https://doi.org:10.1056/NEJMoa2414610

19 (CDC), C. f. D. C. a. P. *Current Situation: Bird Flu in Humans*, <https://www.cdc.gov/bird-flu/situation-summary/inhumans.html#:~:text=Human%20infections%20with%20bird%20flu,mucous%20and%20feces%20have%20touched.> (2024).

20 (CDC), C. f. D. C. a. P. USDA Reported H5N1 Bird Flu Detections in Poultry. (2025).

21 Statistics, U. S. B. o. L. (2025).

22 (NASS), N. A. S. S. Chickens and Eggs 03/21/2025 Report. (2025).

23 (USDA), U. S. D. o. A. (2025).

24 Archives, F. R. N. (2024).

25 James L. Mitchell, J. M. T., and Trey Malone. The Economic Impact of HPAI on U.S. Egg Consumers: Estimating a $1.41 Billion Loss in Consumer Surplus. (Dale Bumpers College of Agriculture, Food & Life Sciences and University of Arkansas System Division of Agriculture, 2025).

26 Douglas, T. P. a. L. in *Reuters* (Reuters Business, 2025).

27 Sopke, V. G. a. K. in *Associated Press* (Time 2025).

28 Abdelwhab, E. M. & Mettenleiter, T. C. Zoonotic Animal Influenza Virus and Potential Mixing Vessel Hosts. *Viruses* **15** (2023). https://doi.org:10.3390/v15040980

29 Kristensen, C., Jensen, H. E., Trebbien, R., Webby, R. J. & Larsen, L. E. The avian and human influenza A virus receptors sialic acid (SA)-α2,3 and SA-α2,6 are widely expressed in the bovine mammary gland. (2024). https://doi.org:10.1101/2024.05.03.592326

30 Kim, J. K., Negovetich, N. J., Forrest, H. L. & Webster, R. G. Ducks: the "Trojan horses" of H5N1 influenza. *Influenza Other Respir Viruses* **3**, 121-128 (2009). https://doi.org:10.1111/j.1750-2659.2009.00084.x

31 Evseev, D. & Magor, K. E. Innate Immune Responses to Avian Influenza Viruses in Ducks and Chickens. *Vet Sci* **6** (2019). https://doi.org:10.3390/vetsci6010005

32 Forrest, H. L., Kim, J. K. & Webster, R. G. Virus shedding and potential for interspecies waterborne transmission of highly pathogenic H5N1 influenza virus in sparrows and chickens. *J Virol* **84**, 3718-3720 (2010). https://doi.org:10.1128/JVI.02017-09

33 Keawcharoen, J. *et al.* Wild ducks as long-distance vectors of highly pathogenic avian influenza virus (H5N1). *Emerg Infect Dis* **14**, 600-607 (2008). https://doi.org:10.3201/eid1404.071016

34 Oguzie, J. U. *et al.* Avian Influenza A(H5N1) Virus among Dairy Cattle, Texas, USA. *Emerg Infect Dis* **30**, 1425-1429 (2024). https://doi.org:10.3201/eid3007.240717

35 Halwe, N. J. *et al.* Outcome of H5N1 clade 2.3.4.4b virus infection in calves and lactating cows. *bioRxiv* (2024). https://doi.org:10.1101/2024.08.09.607272

36 Caserta, L. C. *et al.* Spillover of highly pathogenic avian influenza H5N1 virus to dairy cattle. *Nature* **634**, 669-676 (2024). https://doi.org:10.1038/s41586-024-07849-4

37 Campbell, A. J., Brizuela, K. & Lakdawala, S. S. mGem: Transmission and exposure risks of dairy cow H5N1 influenza virus. *mBio* **16**, e0294424 (2025). https://doi.org:10.1128/mbio.02944-24

38 Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill* **22** (2017). https://doi.org:10.2807/1560-7917.ES.2017.22.13.30494

39 Khare, S. *et al.* GISAID's Role in Pandemic Response. *China CDC Wkly* **3**, 1049-1051 (2021). https://doi.org:10.46234/ccdcw2021.255

40 GISAID. *H5N1 Bird Flu continues to take its toll in the United States*, <https://gisaid.org/resources/gisaid-in-the-news/highly-pathogenic-avian-influenza-outbreak-in-the-united-states/#c5130> (2025).

41 Simonyan, V. & Mazumder, R. High-Performance Integrated Virtual Environment (HIVE) Tools and Applications for Big Data Analysis. *Genes (Basel)* **5**, 957-981 (2014). https://doi.org:10.3390/genes5040957

42 Simonyan, V. *et al.* HIVE-heptagon: A sensible variant-calling algorithm with post-alignment quality controls. *Genomics* **109**, 131-140 (2017). https://doi.org:10.1016/j.ygeno.2017.01.002

43 Santana-Quintero, L., Dingerdissen, H., Thierry-Mieg, J., Mazumder, R. & Simonyan, V. HIVE-hexagon: high-performance, parallelized sequence alignment for next-generation sequencing data analysis. *PLoS One* **9**, e99033 (2014). https://doi.org:10.1371/journal.pone.0099033

44 CensuScope v. 2.0.0 (GitHub, 2024).

45 Kans, J. (NIH NLM, 2013).

46 Lab, H. (2016).

47 ObservableHQ (2017).

48 Landazabal-Castillo, S., Suarez-Aguero, D., Alva-Alvarez, L., Mamani-Zapana, E. & Mayta-Huatuco, E. Highly pathogenic avian influenza A virus subtype H5N1 (clade 2.3.4.4b) isolated from a natural protected area in Peru. *Microbiol Resour Announc* **13**, e0041724 (2024). https://doi.org:10.1128/mra.00417-24

49 Tabynov, K. *et al.* Detection and genomic characterization of an avian influenza virus A/mute swan/Mangystau/1-S24R-2/2024 (H5N1; clade 2.3.4.4b) strain isolated from the lung of a dead swan in Kazakhstan. *Microbiol Resour Announc* **13**, e0026024 (2024). https://doi.org:10.1128/mra.00260-24

50 Hora, B. *et al.* Streamlined Subpopulation, Subtype, and Recombination Analysis of HIV-1 Half-Genome Sequences Generated by High-Throughput Sequencing. *mSphere* **5** (2020). https://doi.org:10.1128/mSphere.00551-20

51 Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**, 3059-3066 (2002). https://doi.org:10.1093/nar/gkf436

52 Sreenivasan, C. C., Thomas, M., Kaushik, R. S., Wang, D. & Li, F. Influenza A in Bovine Species: A Narrative Literature Review. *Viruses* **11** (2019). https://doi.org:10.3390/v11060561

53 Kaplan, B. S. & Webby, R. J. The avian and mammalian host range of highly pathogenic avian H5N1 influenza. *Virus Res* **178**, 3-11 (2013). https://doi.org:10.1016/j.virusres.2013.09.004

54 Nguyen, T.-Q. *et al.* Emergence and interstate spread of highly pathogenic avian influenza A(H5N1) in dairy cattle. *bioRxiv* (2024). https://doi.org:10.1101/2024.05.01.591751

55 Zhang, Y. *et al.* Influenza Research Database: An integrated bioinformatics resource for influenza virus research. *Nucleic Acids Res* **45**, D466-D474 (2017). https://doi.org:10.1093/nar/gkw857

56 Bogner, P., Capua, I., Lipman, D. J. & Cox, N. J. A global initiative on sharing avian flu data. *Nature* **442**, 981-981 (2006). https://doi.org:10.1038/442981a

57 Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob Chall* **1**, 33-46 (2017). https://doi.org:10.1002/gch2.1018

58 Pennington, E. *A Bioinformatic Pipeline for the Detection of Drug Resistant Mutations for Bacterial and Viral Pathogens* Master of Science thesis, George Washington University, (2023).

59 Grady, S. K. *et al.* A graph theoretical approach to experimental prioritization in genome-scale investigations. *Mamm Genome* **35**, 724-733 (2024). https://doi.org:10.1007/s00335-024-10066-z

Appendix

HIVE-Hexagon Parameters for the Nearest Neighbor Pipeline.

Keep the default parameters EXCEPT:

- Minimum match:
  - Length: *90* and Unit: *Percentage (%)*
- Matches to keep: *All equally best alternative matches*
- Mismatches:
  - Percent allowed *45*
  - Computed on: *minimum match length*

Advanced Parameters:

- K-mer Extension Minimal Length Percent: *22*
- Width of Intelligent Diagonal: *60*
- Seed k-mer: *11 letters*
- Low complexity filter:
  - Reference masking: Filter NN and low quality: *Filter Ns only*
- Resolve conflicts:
  - Conflict resolution method: *Markovnikov rule*
  - Score for conflict resolution: *Number of reads*
- Trim:
  - Maximum Mismatch: *10%*
  - Window size: *20*

Filters for Sankey Diagrams in Hexahedron.

- Length: *500*
- Coverage: *50*
- Support: *5*