Final Report for Syslab

Exploring the Spread and Evolution of Avian Influenza through Bioinformatic Analyses

Jasmine Ma

June 2, 2025

Yilmaz - Period 3

Table of Contents

Abstract

To investigate the spread and evolution of avian influenza, I constructed a time-calibrated maximum likelihood phylogenetic tree and generated a geospatial case distribution map from 2018–2023. I applied Freyja, which was originally used for mixed SARS-CoV-2 wastewater analysis, to avian influenza environmental samples, revealing evolving viral lineages at low concentrations. These findings indicate potential reassortment and zoonotic spillover risks. Using GISAID sequences and EMPRES-i outbreak data, I visualized regional spread via migration routes and environmental spots. This approach highlights the potential of environmental surveillance for early detection, offering a cost-effective method for monitoring viral evolution and public health threats.

I. Introduction

Zoonotic spillover, the transfer of pathogens from animals to humans, has been responsible for some of the most significant public health threats in recent history, including the SARS-CoV-2 pandemic and repeated avian influenza outbreaks [3][4]. These events often stem from direct contact with infected animals or fomites and are intensified by the high risk of viral reassortment within animal populations, which can generate new and potentially pandemic-causing strains [2].

This project was motivated by the growing need for earlier and more scalable detection of such outbreaks. Traditional surveillance systems are often reactive and limited in scope. In contrast, environmental and wastewater sampling offer a more affordable and representative method for tracking viral evolution and community transmission before the first clinical cases emerge [1][5]. While widely applied to

SARS-CoV-2, I identified a gap in tools adapted for other viruses with spillover potential, such as avian influenza.

To address this, I explored if Freyja, a computational tool originally developed to detect SARS-CoV-2 lineages in mixed wastewater samples, could be adapted to analyze avian influenza variants in environmental data [9]. This application is novel because it expands Freyja's utility beyond COVID-19, demonstrating its potential to detect low-concentration lineages of avian influenza. By integrating phylogenetic analysis and geospatial mapping, my approach provides an early-warning system that could be used to monitor viral evolution along migratory routes and prevent future spillover events.

II. Background

Avian influenza is known to circulate widely in wild and domestic birds, and its ability to jump species barriers poses a constant threat of zoonotic spillover [2][4]. Existing surveillance methods include traditional field-based sampling and molecular diagnostics such as qPCR, which work well for known pathogens in clinical contexts [5][6]. However, these approaches are geographically constrained, require physical access to animal populations, and often miss asymptomatic or environmental transmission [7]. Environmental surveillance has shown promise as a less invasive, more population-representative approach to tracking viral spread [1][8]. These techniques capture signals from multiple species simultaneously and can detect outbreaks before clinical cases are reported, making them valuable for public health interventions.

To enhance surveillance of AIVs, this study integrates multiple bioinformatics tools. Freyja is a computational tool designed for deconvoluting mixed viral populations in wastewater, providing lineage abundance data even at low concentrations [9]. MAFFT is used for multiple sequence alignment, and IQTree constructs the phylogenetic tree. TimeTree calibrates the tree using temporal metadata to highlight mutation patterns over time [11][12]. Visualizations were produced using Python packages like GeoPandas, Matplotlib, Cartopy, and Contextily [16].

This study draws on two major datasets. First, 1,689 genomic sequences were obtained from the GISAID database, which includes virus subtype, host species, sampling date, and location metadata [13][14]. Second, the EMPRES-i platform from the FAO provides outbreak reports and geographic case distribution data and can be used to contextualize evolutionary relationships [15]. Below are snapshots of the dataset.

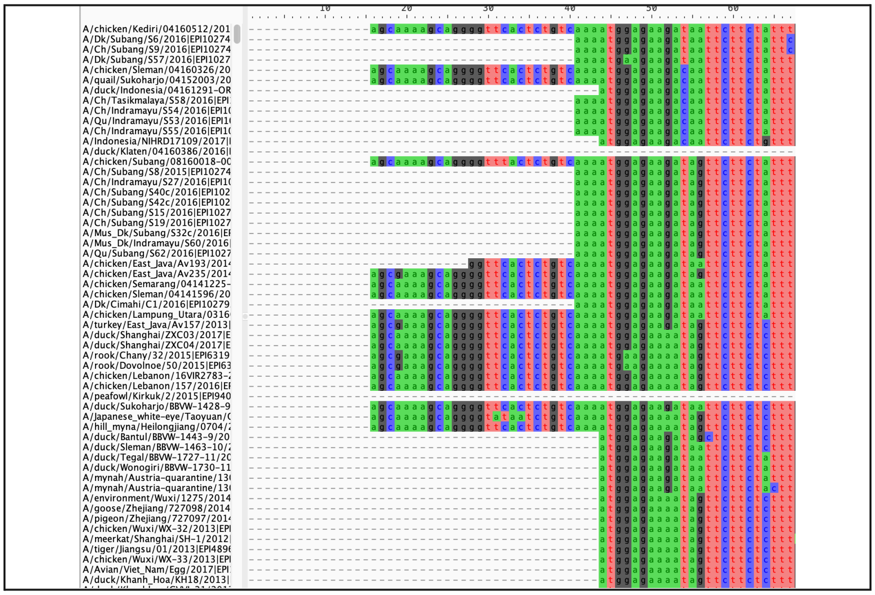

10
20
30
40
50
60
A/chicken/Kediri/04160512/201
A/Dk/Subang/S6/2016|EPI10274
A/Ch/Subang/S9/2016|EPI10274
A/Dk/Subang/S57/2016|EPI1027
A/chicken/Sleman/04160326/20
A/quail/Sukoharjo/04152003/20
A/duck/Indonesia/04161291-OR
A/Ch/Tasikmalaya/S58/2016|EPI
A/Ch/Indramayu/S54/2016|EPI1
A/Qu/Indramayu/S53/2016|EPI1
A/Ch/Indramayu/S55/2016|EPI1
A/Indonesia/NIHRD17109/2017|
A/duck/Klaten/04160386/2016|
A/chicken/Subang/08160018-00
A/Ch/Subang/S8/2015|EPI10274
A/Ch/Indramayu/S27/2016|EPI1
A/Ch/Subang/S40c/2016|EPI102
A/Ch/Subang/S42c/2016|EPI102
A/Ch/Subang/S15/2016|EPI1027
A/Ch/Subang/S19/2016|EPI1027
A/Mus_Dk/Subang/S32c/2016|EI
A/Mus_Dk/Indramayu/S60/2016|
A/Qu/Subang/S62/2016|EPI1027
A/chicken/East_Java/Av193/2014
A/chicken/East_Java/Av235/2014
A/chicken/Semarang/04141225-
A/chicken/Sleman/04141596/20
A/Dk/Cimahi/C1/2016|EPI10279
A/chicken/Lampung_Utara/0316
A/turkey/East_Java/Av157/2013|
A/duck/Shanghai/ZXC03/2017|E
A/duck/Shanghai/ZXC04/2017|E
A/rook/Chany/32/2015|EPI6319
A/rook/Dovolnoe/50/2015|EPI63
A/chicken/Lebanon/16VIR2783-
A/chicken/Lebanon/157/2016|EI
A/peafowl/Kirkuk/2/2015|EPI940
A/duck/Sukoharjo/BBVW-1428-9
A/Japanese_white-eye/Taoyuan/
A/hill_myna/Heilongjiang/0704/
A/duck/Bantul/BBVW-1443-9/20
A/duck/Sleman/BBVW-1463-10/
A/duck/Tegal/BBVW-1727-11/20
A/duck/Wonogiri/BBVW-1730-11
A/mynah/Austria-quarantine/13
A/mynah/Austria-quarantine/13
A/environment/Wuxi/1275/2014
A/goose/Zhejiang/727098/2014
A/pigeon/Zhejiang/727097/2014
A/chicken/Wuxi/WX-32/2013|EPI
A/meerkat/Shanghai/SH-1/2012|
A/tiger/Jiangsu/01/2013|EPI489
A/chicken/Wuxi/WX-33/2013|EPI
A/Avian/Viet_Nam/Egg/2017|EPI
A/duck/Khanh_Hoa/KH18/2013|

Figure 1. Aligned sequences using MAFFT

III. Applications:

Early Outbreak Detection: Environmental and wastewater-based surveillance using tools like Freyja enables detection of viral evolution before clinical symptoms emerge, which can allow for faster public health responses to avian influenza and other zoonotic threats.

Wildlife and Agricultural Management: By tracking viruses across farms, wetlands, and migratory hotspots, this method can inform targeted testing zones and guide wildlife monitoring.

Public Health Preparedness: Phylogenetic and geospatial analyses help predict where and when novel strains might emerge, enabling the allocation of vaccines, antivirals, or containment resources to high-risk areas.

Adaptability to Other Pathogens: Although used here for avian influenza, this pipeline can be adapted to analyze other emerging diseases (e.g., monkeypox, coronaviruses), making it a versatile tool for ongoing biosurveillance.

IV. Methods:

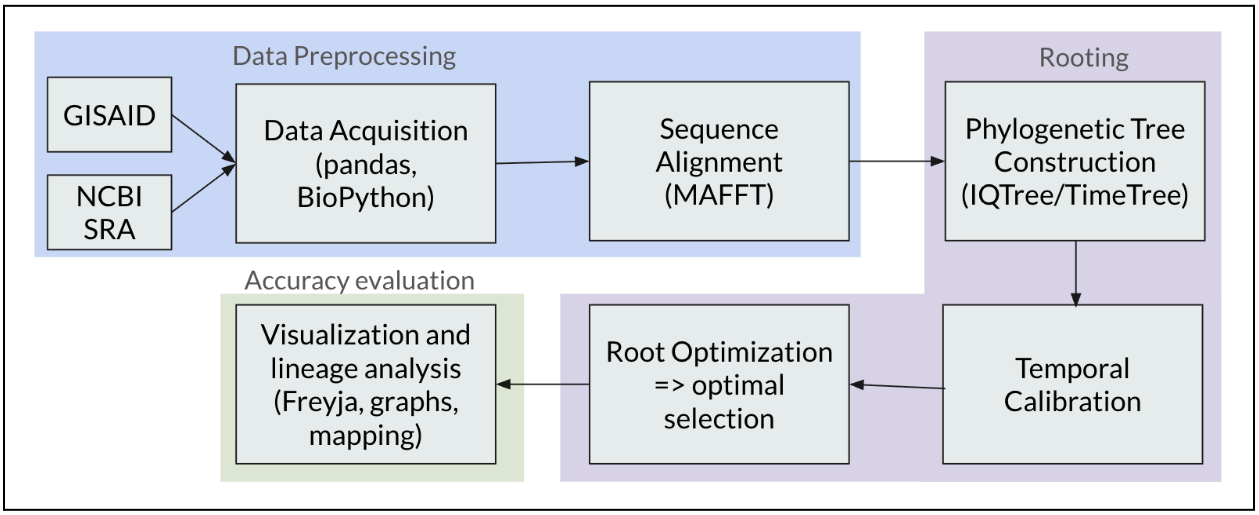

Data Preprocessing
GISAID
NCBI SRA
Data Acquisition (pandas, BioPython)
Sequence Alignment (MAFFT)
Rooting
Phylogenetic Tree Construction (IQTree/TimeTree)
Temporal Calibration
Root Optimization => optimal selection
Accuracy evaluation
Visualization and lineage analysis (Freyja, graphs, mapping)

Figure 2. Systems architecture diagram

This project analyzes avian influenza evolution and spread using genomic sequences from GISAID [13][14] and outbreak metadata from EMPRES-i [15]. Inputs include 1,689 viral sequences with metadata (date, location, host) and environmental sampling data. Outputs include a time-calibrated phylogenetic tree, lineage prevalence profiles, and geospatial outbreak maps. Sequences are first aligned using MAFFT [12], then processed through IQTree to generate a maximum likelihood tree. TimeTree [11] is used to integrate temporal metadata and calibrate the tree. Root optimization is applied using both temporal calibration and outgroup comparison due to high sequence divergence [10]. Variant deconvolution is conducted using Freyja [9], which analyzes mixed environmental samples to estimate lineage abundances, even at low concentrations.

Geospatial visualization is performed using GeoPandas, Matplotlib, Cartopy, and Contextily [16] to map variant spread and mutation clusters. The approach was motivated by the need for early detection of zoonotic spillover and the limited scalability of traditional farm-level testing [7]. Previous tools like qPCR and clinical sequencing were too narrow in scope or failed to capture mixed populations in environmental contexts [6]. Early attempts using IQTree’s default rooting produced inaccurate results, leading to the integration of TimeTree. Sediment samples also introduced noise, so cleaner environmental sources (e.g., fecal) were prioritized [1]. If given more time, I would automate the pipeline, integrate migratory and climate data, and expand to other subtypes or zoonotic pathogens. If I could redo part of the project, I would apply stricter filtering to environmental samples and separate sampling based on source type. Tools used include MAFFT, IQTree, TimeTree, and Freyja, along with open-source Python libraries. Dataset details and tool references are included in the citation list.

V. **Results**:

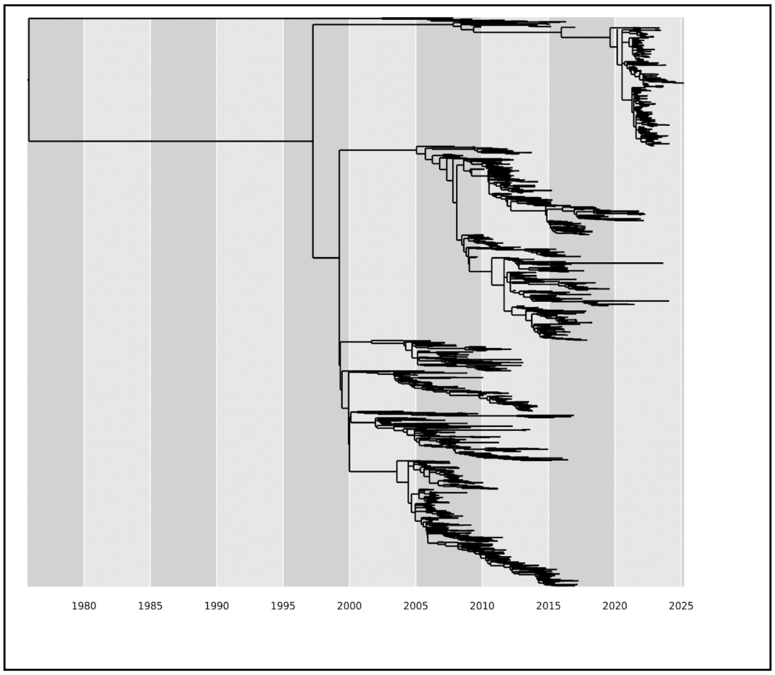


Figure 3. A time-calibrated maximum likelihood phylogenetic tree was constructed using TimeTree based on 1,689 influenza sequence samples.

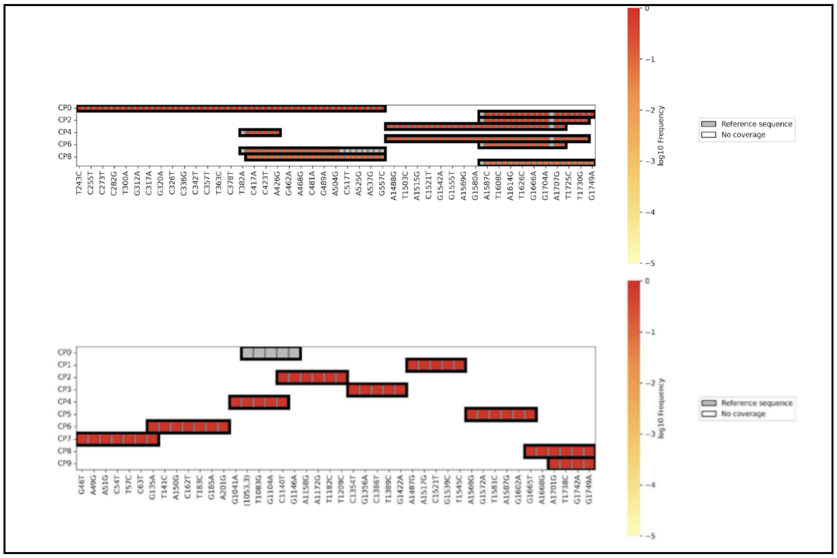
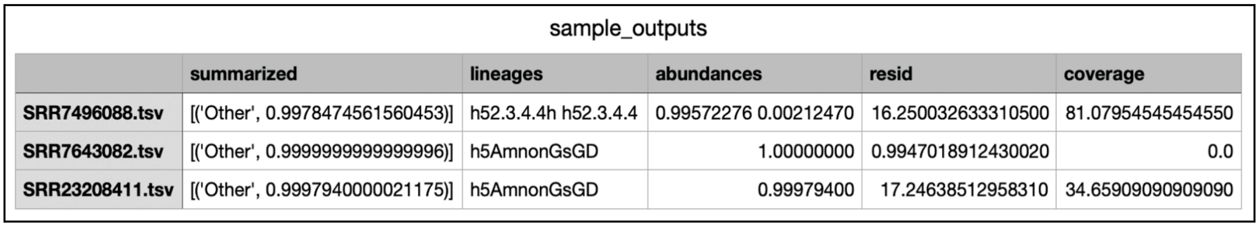


Figure 4. Covariant analysis for SRR7496088 and SRR23208411



sample_outputs

	summarized	lineages	abundances	resid	coverage
SRR7496088.tsv	[('Other', 0.9978474561560453)]	h52.3.4.4h h52.3.4.4	0.99572276 0.00212470	16.250032633310500	81.07954545454550
SRR7643082.tsv	[('Other', 0.9999999999999996)]	h5AmnonGsGD	1.00000000	0.9947018912430020	0.0
SRR23208411.tsv	[('Other', 0.9997940000021175)]	h5AmnonGsGD	0.99979400	17.24638512958310	34.65909090909090

Figure 5. Freyja output after demixing and aggregating 3 mixed environmental avian influenza samples for lineage prevalence.

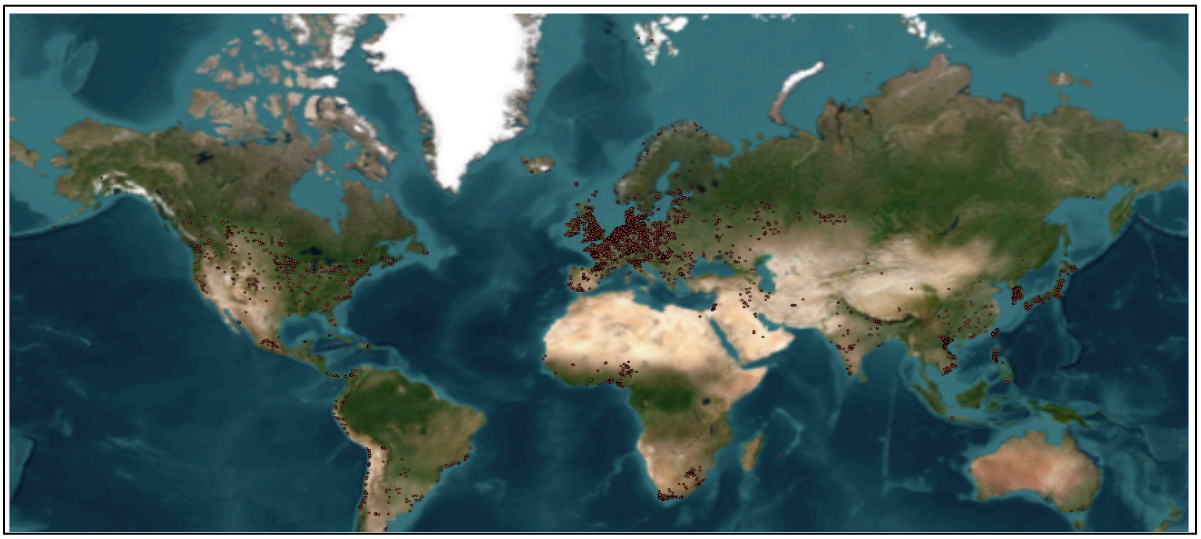


Figure 6. A geospatial map showing avian influenza case distribution between August 2018 to August 2023 was constructed with the Esri World Map basemap layer and plotted using Matplotlib.

Freyja’s covariant and mixed lineage analyses showed a high prevalence of mutated avian influenza lineages in environmental samples, including sediment and fecal sources, confirming its adaptability to non-SARS-CoV-2 pathogens (Figure 4). Multiple samples contained evidence of mixed lineages, indicating potential genetic reassortment and heightened zoonotic spillover risk, particularly in shared environments like farms or water reservoirs that can be seen through mapping (Figure 6).

Phylogenetic trees built using IQTree and temporally calibrated with TimeTree revealed distinct evolutionary clusters and branching patterns, suggesting ongoing mutation and divergence among strains (Figure 3).

Geospatial analyses using EMPRES-i outbreak data and environmental sample metadata further revealed that many clusters aligned with bird migratory routes, highlighting movement-driven spread. Environmental sampling also highlighted a distinction in data quality: sediment-collected samples often introduced more noise, whereas fecal and waterborne samples yielded clearer barcode reads with more patterns recognized (Figure 5). Overall, these results support the use of environmental sequencing and Freyja-based variant analysis as scalable, cost-effective tools for detecting emerging avian influenza threats before clinical outbreaks are reported.

VI. **Limitations**:

Processing covariant data was time-intensive and required manual analysis of each sample, limiting scalability. The workflow lacked automation, and the environmental samples themselves posed challenges. In mixed-source environments like sediment or wastewater, viral genetic material often degraded or became difficult to isolate, resulting in incomplete or noisy reads. This reduced the clarity of barcode-based detection methods. Additionally, data availability was a constraint. Many wastewater surveillance programs have been defunded in the post-COVID era, limiting access to up-to-date environmental sequencing data. High-quality covariant datasets are still rare and difficult to obtain, especially for avian influenza. Finally, while phylogenetic rooting and temporal calibration improved overall tree accuracy, the high variability between avian influenza strains made it difficult to generate a single, fully resolved evolutionary history.

VII. **Conclusion**:

This project demonstrates that phylogenetic and geospatial analyses offer a powerful way to monitor the evolution and spread of avian influenza. Phylogenetic trees constructed from aligned influenza sequences revealed clear evolutionary relationships, allowing us to trace mutation trends over time. The integration of Freyja's mixed-sample analyses allowed the identification of lineages, suggesting ongoing viral reassortment and potential zoonotic spillover risks. Geospatial mapping further strengthened these findings by visualizing the spread of viral strains along known migratory bird routes, revealing geographic hotspots and transmission patterns. Together, these tools form a cost-effective and scalable surveillance approach capable of detecting emerging strains before clinical outbreaks occur. This method has important public health implications, as sampling along frequent migratory corridors can help prevent cross-species transmission and guide targeted interventions to prevent future epidemics.

VIII. **Future Work**:

Future work will focus on expanding the scale, accessibility, and predictive power of the surveillance pipeline. A primary goal is to develop and publicly deploy a Bash script on GitHub that automates the entire workflow. Additionally, working with my internship team, I hope to study more about how Freyja can be used for different types of environmental data.

Additionally, I hope to research how we can integrate climate and seasonal data with viral spread patterns. As climate change alters migratory bird behavior, habitat stability, and the range of evolving species, it will likely influence when and where new strains of avian influenza emerge. Mapping these ecological shifts alongside mutation patterns could help forecast future spillover zones with greater accuracy. In addition, higher-resolution geospatial analysis can be used to detect micro-level outbreak clusters in vulnerable ecosystems. This method could also be adapted for surveillance of other zoonotic diseases, including emerging coronaviruses and monkeypox.

IX. Materials:

Datasets:

- GISAID (Global Initiative on Sharing All Influenza Data) provided 1,689 avian influenza genome sequences with metadata including host species, collection date, subtype, and location [13][14].
- EMPRES-i (Emergency Prevention System for Transboundary Animal and Plant Pests and Diseases) from the FAO offered outbreak coordinates [15].

Sequence Analysis Tools:

- MAFFT was used for multiple sequence alignment of influenza genomes [12].
- IQTree generated maximum likelihood phylogenetic trees.
- TimeTree performed temporal calibration of phylogenetic trees based on sample collection dates [11].

Variant and Lineage Analysis:

- Freyja for mixed lineage analysis [9].
- iVar was also used for variant calling where applicable.

Geospatial Visualization:

- Python libraries GeoPandas, Matplotlib, Cartopy, and Contextily were used to visualize spatial distributions and create outbreak maps [16].
- ArcGIS Online Basemaps were used for geographic context [16].

References:

Link to Presentation Video

Link to GitHub

[1] Himsworth, C. G., et al. (2020). Targeted Resequencing of Wetland Sediment as a Tool for Avian Influenza Virus Surveillance. Journal of Wildlife Diseases, 56(2), 397–408.

[2] Joseph, U., et al. (2017). The ecology and adaptive evolution of influenza A interspecies transmission. Influenza and Other Respiratory Viruses, 11(1), 74–84. https://doi.org/10.1111/irv.12412

[3] Garry, R. F. (2022). The evidence remains clear: SARS-CoV-2 emerged via the wildlife trade. PNAS, 119(47), e2214427119. https://doi.org/10.1073/pnas.2214427119

[4] EFSA et al. (2024). Avian Influenza Update. European Food Safety Authority.

[5] Hoye, B. J., et al. (2010). Surveillance of wild birds for avian influenza virus. Emerging Infectious Diseases, 16(12), 1827–1834. https://doi.org/10.3201/eid1612.100589

[6] Navarro, A., et al. (2021). SARS-CoV-2 detection in wastewater using multiplex qPCR. Science of the Total Environment, 797, 148890. https://doi.org/10.1016/j.scitotenv.2021.148890

[7] Schreiber, M. (2024). Lack of bird flu testing may be hiding true spread of virus on US farms. The Guardian. https://www.theguardian.com

[8] Ladyzhets, B. (2023). A valuable early-warning system for disease outbreaks could be shut down. Scientific American. https://www.scientificamerican.com

[9] Karthikeyan, S., et al. (2022). Wastewater sequencing reveals early cryptic SARS-CoV-2 variant transmission. Nature, 609, 101–108. https://doi.org/10.1038/s41586-022-05049-6

[10] Kinene, T., et al. (2016). Rooting trees, methods for. Encyclopedia of Evolutionary Biology, 489–493. https://doi.org/10.1016/B978-0-12-800049-6.00215-8

[11] Sagulenko, P., et al. (2018). TreeTime: Maximum-likelihood phylodynamic analysis. Virus Evolution, 4(1). https://doi.org/10.1093/ve/vex042

[12] Donoghue, P. C., & Yang, Z. (2016). The evolution of methods for establishing evolutionary timescales. Philosophical Transactions B, 371(1699). https://doi.org/10.1098/rstb.2016.0020

Datasets Used

[13] Shu, Y., & McCauley, J. (2017). GISAID: Global initiative on sharing all influenza data. Euro Surveillance, 22(13), 30494. https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494

[14] Khare, S., et al. (2021). GISAID's role in pandemic response. China CDC Weekly, 3(49), 1049–1051. https://doi.org/10.46234/ccdcw2021.255

[15] Food and Agriculture Organization of the United Nations. (n.d.). Avian Influenza Diseases Map. EMPRES-i. https://empres-i.apps.fao.org/diseases

[16] ArcGIS Online Basemaps [Map]. (2010).