國立臺灣大學醫學院免疫學研究所

博士論文

Graduate Institute of Immunology

College of Medicine

National Taiwan University

Doctoral Dissertation

# 研發生物資訊軟體以整合分析新興病毒的演變與流行趨勢

# Developing the Integrated Suites of Bioinformatic Software to Analyze the Evolutionary Variations of Emerging Viruses and their Epidemic Trends

楊沁儒

Chin-Rur Yang

指導教授：顧家綺 博士

Advisor: Chia-Chi Ku, Ph.D.

中華民國 112 年 6 月

Jun. 2023

doi:10.6342/NTU202301433

# 國立臺灣大學博士學位論文
# 口試委員會審定書

**PhD DISSERTATION ACCEPTANCE CERTIFICATE**
**NATIONAL TAIWAN UNIVERSITY**

## 研發生物資訊軟體以整合分析新興病毒的演變與流行趨勢

## Developing the Integrated Suites of Bioinformatic Software to Analyze the Evolutionary Variations of Emerging Viruses and their Epidemic Trends

本論文係楊沁儒（學號 D05449002）在國立臺灣大學醫學院免疫學研究所完成之博士學位論文，於民國 112 年 6 月 13 日承下列考試委員審查通過及口試及格，特此證明。

The undersigned, appointed by the Graduate Institute of Immunology on 13 June 2023 have examined a PhD dissertation entitled above presented by CHIN-RUR YANG (student ID D05449002) candidate and hereby certify that it is worthy of acceptance.

口試委員 Oral examination committee:

顧家綺
（指導教授 Advisor）

李建國

張淑媛　　詹大千

系主任/所長 Director: 徐立中

doi:10.6342/NTU202301433

# 誌謝

在就讀博士班學習的這七年時光中，時光荏苒也將走到了論文完成及畢業的盡頭，在這段期間中雖然在求學問的道路上十分煎熬，但是在過程中的挫折與突破的點滴回憶將會使我未來更有自信，面對挑戰也無所畏懼，將我的理想寄託於探索未知的未來。

本論文得以完成，首先要感謝的是在本研究中使用所有提供序列的研究單位及先進，因為有你們的努力得以讓小小的我可以參考這些龐大數量的病毒序列，而從混沌的大海中分析出些許可能的脈絡，也希望自己可以向各位加以學習奮進。

感謝我的指導教授顧教授 家綺，因為您的細心指導與教誨讓我可以從各種角度深入思考，並學習到許多自己不足的事物及增進各種能力，最重要的是學習到溝通以及處理事情的倫理與道義，培養出我無論在分析數據實驗分析及判讀上，或是從對於科學事物觀察的敏銳及從邏輯茫茫大海中，找出一條屬於對的方向之光明路，雖然在許多時候很抱歉都造成你很困擾，真的十分感謝您對我的包容及關懷，讓我不會陷於實驗邏輯漩渦中而失去信心，若是感到不安時也能一掃陰霾，堅強的往科學真理之路邁進。

在論文口試期間，承蒙口試委員李教授 建國、張教授 淑媛、黃教授 立民、劉教授 力瑜和詹教授 大千及我的指導老師顧教授 家綺，對於研究的方向的鼓勵與指正研究疏漏處，觀念啟迪與發想也使我獲益良多，感謝口試委員費心審閱，並惠賜諸多建議，使本論文架構更臻完整及更具說服力，在此謹深致謝忱。

在研究所學習中，真的十分感謝在免疫所及所外所有指導過我的老師，包括朱教授 清良、繆教授 希椿及金教授 傳春，還有過去所有教導過我的老師，族繁不及備載，因為你們知識的分享及教誨，讓我得以獲益良多，真的很感謝各位老師費盡心思努力的關懷指導。

在學習過程中，感謝在實驗室中幫助過我的學長姐、同儕及學弟妹，林宇瑞、張毅軒、于主念、陳建甫、簡柏雅、蔡維倫、林承右、賴證升、洪家萱、王珮儒等的激勵及學習上的激盪，讓我得以更加的成長，使我從各位的身上學到了努力及堅持的精神，十分謝謝你們。

在生活上十分感謝我的父母及所有家人的陪伴與支持，因為你們的細心照顧使我沒有後顧之憂，使我在科學領域上專心學習及研究，對於我的關懷體諒，我都會謹記於心，努力去做好自己，逐步去實踐未來。

最後謹以 本文 獻於世界上每一位研究的科學家　　楊沁儒謹誌於

國立臺灣大學醫學院免疫學研究所

中華民國 112 年 6 月

doi:10.6342/NTU202301433

# 中文摘要

新興呼吸道病毒是公共衛生重要議題，它們的高突變率更突顯了監測病毒全基因體序列的必要性，現今許多線上分析工具並不適用於全基因體序列，本論文研究將從克服這些限制為目標開發新的軟體工具，深入分析了禽流感和新冠病毒。

關於禽流感全基因體序列分析部分，先以「流感病毒序列轉換」(FluConvert) 自動處理原始序列數據，並按照病毒命名法 (ABCD 類型/宿主/區域/菌株/年份/HxNy 亞型) 重新排列病毒片段，序列對齊後轉譯為胺基酸序列。隨後「流感病毒序列溯源」(IniFlu) 軟體，彙整了這些具有顯著特徵的胺基酸序列，並根據研究目標分群，檢視不同分群中重要的病毒共有序列。分析結果獲得了除了 HA 還有其他 10 種病毒蛋白中共有 247 個與 H5N2 的高致病性具有相關的胺基酸點位變異，大部分的變異點位尚未被報導。在這套創新的軟體和方法的基礎上，我們繼續分析了 2021 年 4 月至 9 月間台灣爆發新冠肺炎流行的 Alpha 變種病毒株，從病毒基因指紋釐清不同的傳播鏈以及出現和防控主要流行病毒株的流行病學條件。以上二個研究成果說明了本基因序列分析軟體可以成功快速地分析不同病毒株全基因體，同時識別這些多基因共有特徵以進行綜合研究。

總之，這項研究為全面的病毒全基因序列分析提供了一站式平台，可以同時分析整個病毒全基因體，並輕鬆與其他重要資訊整合，以取得具有獨特特徵之病毒序列，未來仍需要努力建立實際實驗證據來驗證分析。然而本研究中所研發軟體分析甚至將此分析法應用於其他快速傳播、具致病力及有全球流行潛力病原，即早偵測具健康威脅新興病毒，找出演變關鍵，協助科學研究進展與成功防控。

關鍵字：基因序列分析平台、流感病毒、新冠病毒、病毒資訊學、風險評估、大流行

doi:10.6342/NTU202301433

# Abstract

Respiratory viruses with high mutation rates have become a significant public health concern, highlighting the need for monitoring complete viral sequences. While online sequence analysis tools exist, they cannot often analyze the entire genomic sequence, creating a gap that requires developing new software tools. This dissertation analyzes two emerging viruses, avian influenza viruses, and SARS-CoV-2.

In the first part of the research work, I developed the analysis software packages to analyze whole AIV genome sequences comprehensively. The FluConvert software automatically processes raw sequence data, organizing viral segments based on virus nomenclature (ABCD Type/Host/Region/Strain/Year/HxNy Subtype) and aligning distinct genes, and translating them into protein sequences. Subsequently, the IniFlu software integrates protein sequences with significant characteristics, allowing for classification based on study objectives and examination of consensus sequences in different subgroups. This innovative approach has led to identifying 247 polygenic consensus signatures associated with highly pathogenic AIV (HPAIV) across HA and ten other proteins, most of which have not been reported in the literature. Our pioneering software and methods enable rapid analysis of diverse strains' genomes while identifying polygenic consensus signatures for integrated investigations.

doi:10.6342/NTU202301433

The second part of the study focused on understanding the dynamic changes of SARS-CoV-2 Alpha variant strains in responses to various control measures during the outbreak in Taiwan from late April to September 2021. The goal was to delineate the epidemiological circumstances that allowed these strains to become predominant. The findings provided valuable insights into the emergence and control of a dominant viral strain during an outbreak.

In conclusion, the study offered an integrated platform for comprehensive viral genome sequence analysis. It allows for simultaneous estimation of the complete viral genome while easily integrating other significant information to extract characteristics-specific viral sequences. Future experimental validation is required to support the analysis. Applying this integrated analysis method to other pathogens with rapid spread, high pathogenicity, and pandemic potential will provide insightful information for the early detection of emerging or health-threatening dominant viruses. Results from the study will contribute to scientific progress and early disease prevention and control success.

Keywords: Sequence analysis platform, Influenza virus, SARS-CoV-2, Viroinformatics, Risk assessment, Pandemic

doi:10.6342/NTU202301433

# Contents

doi:10.6342/NTU202301433

# Contents of Tables

doi:10.6342/NTU202301433

# Contents of Figures

doi:10.6342/NTU202301433

# Chapter 1
# Introduction

doi:10.6342/NTU202301433

Over the past three decades, emerging viruses have cumulatively evolved, producing an increasing clade of RNA viruses. For example, influenza and coronaviruses pose significant threats to the environment and humanity through their transmission and genetic reassortment among animals and human populations. As of April 2023, the novel H5N1 and H5N6 viruses, with a case mortality rate of over 50%, have caused 957 human cases and 507 fatalities (Jiang et al., 2017). Moreover, SARS-CoV-2 initiated the most severe global pandemic in history, infecting 767 million infected people and causing 6.94 million human deaths by June 2023 (Ensheng Dong et al., 2020).

Advancements in sequencing technology have revolutionized the generation of viral sequences, enabling real-time acquisition of such data. The public-domain databases, including National Center for Biotechnology Information Virus (NCBI) and Global Initiative on Sharing All Influenza Data (GISAID), have compiled vast amounts of information on viral sequences, which have accumulated over 11 million and 16 million viral sequences, respectively, as of 2023. To effectively harness and exploit this wealth of information, viroinformatics, a subfield of bioinformatics, has emerged as a crucial discipline for managing and analyzing these datasets. By employing viroinformatics, researchers can gain valuable insights into emerging viruses' genomic changes and evolutionary patterns. This knowledge is pivotal in ensuring public health and well-being

doi:10.6342/NTU202301433

through enhanced surveillance and monitoring of viral outbreaks.

Although there are a few tools accessible to the general public for analyzing viral sequence information, the majority of them are designed to detect variations at a specific position or specific positions within the viral genome. For instance, when examining the influenza virus, which consists of eight segmented genes, these tools can only analyze one gene at a time. Such restriction prevents simultaneous analysis of all genetic variations together with their corresponding vital immunological and epidemiological information. In my thesis research, my goal is to establish a comprehensive suite of analysis software capable to analyze the entire viral genome. This software integrates data from virology, immunology, clinical medicine, and epidemiology. By doing so, it will enable us to explore the dynamic changes of emerging viruses and their correlation with epidemic patterns. In this section, I will provide a detailed literature review introducing the concepts of viral informatics, the evolution of sequencing methods, the unique features of public databases, and an overview of existing analysis tools.

## 1.1 Viroinformatics

Viroinformatics, also known as viral bioinformatics or viral informatics, is an integrated field that utilizes computational techniques to analyze viral genomic sequences in order to identify and characterize viruses. Given that novel viruses periodically emerge,

doi:10.6342/NTU202301433

and pathogens can mutate rapidly, especially RNA viruses, a comprehensive analysis of the spatiotemporal changes of viral genomes and their interactions with humans and other host species can provide insights into the epidemical trends of viral infections. In addition to compiling information on the geographical distribution, target host range, the duration of a particular virus infection, a genome-wide analysis of viral strains collected over time and from different locations can identify novel mutations in viral genomes that emerged after the use of vaccines and antiviral drugs. For example, regular virus surveillance led to the identification of the 2009 H1N1 influenza pandemic (H1N1pdm09) (Dawood et al., 2012), which was derived from a zoonotic H3N2 virus originating from pig farms in Mexico (Mena et al., 2016). Additionally, the H274Y mutation in the NA protein, which is associated with resistance to the antiviral drug "Oseltamivir”, was first found in seasonal H1N1 (Baz et al., 2010; Hurt et al., 2009) and was fixed in H1N1 predominant strains since 2007, culminating in the H1N1pdm09 global pandemic (Bloom et al., 2010).

The viroinformatics approach can also track the evolution of avian influenza viruses isolated from various avian reservoirs (i.e., poultry birds, waterfowls, etc.) and their migration routes (Lee et al., 2017; Lee et al., 2015; Yang et al., 2017). Results from such analyses are particularly important in assessing the potential threat of human infections and deaths. For instance, the mutation of E627K in the PB2 of the avian H5N1 influenza

doi:10.6342/NTU202301433

virus enables it to evade the antiviral innate mechanism by avoiding recognition of 5'ppp-RNA via the innate sensor RIG-I (Weber et al., 2015). The same mechanism can be observed in highly pathogenic avian influenza viruses such as H5N6 and H7N9, which pose a significant zoonotic threat to humans (Peng et al., 2018; Zhu et al., 2015).

Therefore, using viral sequences to analyze the prevalence of emerging infectious diseases will provide a more accurate virus risk assessment for predicting and preventing future pandemics. Monitoring the mutations in the virus that result in resistance against antiviral drugs or a decrease in vaccine effectiveness will provide more insight and direction for research than conventional virological and serological methods.

### 1.2 Generation of viral sequences information

As emerging viruses mutate rapidly, analyzing genetic sequences is the most direct and optimal way to understand viral dynamic changes. Advancements in virus sampling methods and sequencing technologies over the past 30 years have enabled real-time detection of viruses and identification of specific virus strains.

#### 1.2.1 Sanger sequencing

The earliest sequencing method is Sanger sequencing, which involves extracting single-stranded DNA/RNA templates from the sample and adding four types of dideoxynucleotides (ddATP, ddCTP, ddGTP, ddTTP) with radioactive or fluorescent

doi:10.6342/NTU202301433

labels separately. Gene fragments of different lengths are obtained by electrophoresis or signal detection, resulting in clear reads of gene segments between 300~1000 bp (Sanger et al., 1977; Tucker et al., 2009). Sanger sequencing was the gold-standard protocol for completing the Human Genome Project from 1990-2003 and has been extended to other genetic studies (Collins et al., 2003). Although Sanger sequencing has high specificity and clear reads for the target gene segment, it requires good primer design, a large volume of samples, and takes a considerable amount of time to perform multiple rounds of sequencing to resolve the whole genome or multiple genes (Tucker et al., 2009). Especially when sequencing small viral populations in specimens, it requires collecting sufficient virus samples from patients' specimens for sequencing. Further amplification through cell cultures or experimental animals may introduce additional mutations in the virus genome. Therefore, the development of next-generation sequencing (NGS) technology has been crucial in reducing sample usage and obtaining high-throughput whole-genome sequencing information.

### 1.2.2 Next-generation sequencing

#### 1.2.2.1 Library preparation

The breakthrough of NGS methods lies in its high throughput ability to surpass previous Sanger sequencing technological barriers (Tucker et al., 2009). The critical

doi:10.6342/NTU202301433

technical drivers are library preparation, novel nucleic acid detection sequencing platforms, and computational capability advances. In library preparation, the first step is using chemical or physical methods (e.g., ultrasonic fragmentation) to break down the entire genomic sequence into small fragments in 150-400 bp (Head et al., 2014). After breaking down the genomic sequence into small fragments, specific sequencing primers are added to the fragments, and they are barcoded to produce longer contiguous sequences called "contigs." These contigs allow for sample identification. Microfluidics technology is then used to achieve precise sequencing of even minute samples, vastly improving the sequencing throughput of these fragments (Ma et al., 2017). Finally, these barcode-tagged fragments are assembled through computer algorithms to construct the entire genome sequences.

#### 1.2.2.2 Emulsion PCR and pyrosequencing

Aside from technical breakthroughs in library preparation, the establishment of the platform, and innovations in sequencing principles have played a critical role in advancing NGS technology. For instance, the Roche 454 GS FLX sequencing developed the first commercially available platform to complete automated sequencing in 2005, using the novel emulsion PCR and the Pyrosequencing method, which do not rely on the Sanger sequencing principle (M. Margulies et al., 2005). Emulsion PCR (em-PCR), a

doi:10.6342/NTU202301433

bead-based PCR method, is performed by PCR amplifying sample fragments with beads conjugated with barcoded oligonucleotide probes and adaptors containing complemented sequences of the target fragments. As different barcodes in different beads correspond to different sequences, beads collected in microwells enable the differentiation of various sequences (Mardis, 2008; Metzker, 2010). The principle of pyrosequencing is based on detecting pyrophosphate (PPi) released during DNA synthesis. At the beginning of the PCR reaction, DNA polymerase links one dNTP to the sample template and releases PPi. ATP sulfurylase converts PPi to adenosine triphosphate (ATP) and adenosine phosphosulfate (APS), providing ATP energy to produce visible light in the luciferase-catalyzed reaction. Finally, dNTP and ATP are degraded by apyrase, and detection of the light signal is used to obtain the sequence of each nucleotide. These fragments produced are then combined to complete the sequencing process (Marcel Margulies et al., 2005; Nyren et al., 1993).

### 1.2.2.3 Ion semiconductor sequencing

In 2010, a technique similar to Pyrosequencing was further developed into the Life Technologies Ion Torrent semiconductor sequencing, which uses semiconductor chips to detect hydrogen ions generated during DNA polymerization in the PCR process (Merriman & Rothberg, 2012). This sequencing method converted substances released

doi:10.6342/NTU202301433

during DNA synthesis into electrical signals, such as pH changes. However, it faces limitations in sequencing signal conversion when dealing with longer tandem repeats (e.g., TATATA) or homopolymer repeats of the same nucleotide (e.g., AAAAAA), as it cannot determine the exact number of nucleotides, which is a restriction not present in Sanger sequencing (Balzer et al., 2013; Scheible et al., 2014). As a result, recent advances in NGS technology have focused on improving the Sanger sequencing-based approach.

### 1.2.2.4 Illumina dye sequencing

One well-known example of improving Sanger sequencing is Illumina dye sequencing, which employs bridge amplification and four-color distinct fluorescent ddNTPs for a highly refined sequencing method (Canard & Sarfati, 1994; Guo et al., 2008; Meyer & Kircher, 2010). The bridge amplification concept is designed to create different barcodes and 5'- and 3'- end adapters for sample fragments. Then, the oligonucleotides of 5'- and 3'-adapter complementary sequences are coated in the microfluidic channel at close distances. When the sample fragments attach, they cause the entire fragments to bend according to both ends and are firmly fixed in microfluidics (Kim et al., 2013; Ma et al., 2017). Following the principle of Sanger sequencing, cluster amplification is performed by polymerase reaction. The improvement is that the four-color fluorescent ddNTPs distinguish which nucleotide has been attached, and the fluorescence intensity

doi:10.6342/NTU202301433

represents the number of identical nucleotide attachments in tandem repeat situations during sequencing (Guo et al., 2008).

Illumina dye sequencing has become one of the most critical NGS sequencing methods today in various fields of genetics. However, limitations in certain sequencing blind spots, such as short reads of a few hundred base pairs, can compromise accuracy and the ability to assemble complete genes when sequencing regions have more tandem repeats (AT or GC-rich) (Chen et al., 2013). Furthermore, amplifying these short-read sequences multiple times (e.g., PCR) can result in sequencing errors. The cost and equipment mobility associated with processing, assembling, and debugging short-read sequences in NGS is currently a challenge that needs to be overcome.

### 1.2.3 Third-generation sequencing

In recent years there has been a rise in third-generation sequencing technology (TGS), which aims to achieve long-read sequencing in real-time by observing the signal generated with a single nucleotide passing through a single polymerase molecule or nanopore (Flusberg et al., 2010; Wang et al., 2021). For example, Pacific Biosciences single-molecule real-time sequencing (PacBio SMRT sequencing) places a single DNA polymerase in a pore of a zero-mode waveguide (ZMW) of size 20 zeptoliters ($10^{-21}$ liters) (Garoli et al., 2019). The fluorescent-labeled nucleotides pass through the DNA

doi:10.6342/NTU202301433

polymerase at a millisecond rate and generate fluorescent signals that distinguish between different nucleotides to complete the whole long read. This method can produce 10,000 to 30,000 base pair reads (Ardui et al., 2018). However, this sequencing method still presents challenges of high manufacturing costs and large equipment size.

Over the past three decades, nanopore sequencing methods have made sequencing real-time, affordable, and portable (Deamer et al., 2016). Oxford Nanopore sequencing, developed in 2014, enables direct sequencing of DNA or RNA by nanopores of transmembrane proteins embedded in biopolymer films without the fragmentation and PCR method. During electrophoresis, distinct nucleotides pass through transmembrane proteins on the biopolymer films causing structural changes that facilitate sequencing by detecting minuscule electrical signals generated by these structural changes (Jain et al., 2016). The rapid passage of nucleotides through the transmembrane proteins, at a rate of 250 to 450 bases per second and without sample fragmentation and PCR, saves significant time during sample processing and sequence assembly (Wang et al., 2021). The nanopore sequencer's compact, cell phone-sized design, makes it highly portable, enabling applications such as rapid, real-time pathogen analysis in outbreak areas. It played a critical role during the 2014 Ebola outbreak in Africa (Hoenen et al., 2016) and over 25% of the SARS-CoV-2 sequences in public databases worldwide were uploaded using this

doi:10.6342/NTU202301433

technology (Hourdel et al., 2020; Rios et al., 2021). While Oxford Nanopore sequencing allows for whole-length gene sequencing at one time, the method's error rate is concomitantly higher. Comparisons with reference sequences are required during alignment, and the presence of point mutations, such as deletions, may be indiscernible due to interference from background signals (Delahaye & Nicolas, 2021; Sahlin & Medvedev, 2021).

In summary, the NGS and TGS methods have provided the speed and depth for unmet needs in Sanger sequencing analyses, but several unresolved issues remain. For example, the discrepancies in sequencing results need to be addressed when the same sample is sequenced by different methods. Moreover, unbiased and automated analyses are required to identify novel and significant mutations within viral genomes, which will facilitate the detection of consequential emerging viral strains. These are ongoing efforts to develop advanced sequencing approaches in the field of viroinformatics.

### 1.3 Public databases for viral sequences

Advancements in sequencing technology have enabled the rapid generation of vast numbers of viral sequences. Various public-domain databases store sequence information and associated epidemiological information to facilitate further research. Major databases containing influenza virus nucleotide sequences and epidemiological data

doi:10.6342/NTU202301433

include the NCBI Influenza Virus Database (NCBI-IVD) (Bao et al., 2008), GISAID EpiFlu database (Shu & McCauley, 2017), and Bacterial and Viral Bioinformatics Resource Center (BV-BRC, formerly called Influenza Research Database) (Zhang et al., 2017). Access to NCBI-IVD can be achieved through GenBank accession numbers, BLAST searches of viral sequences, published literature containing sequences isolated from human cases, and comprehensive raw data involving the eight segments of full-length viral sequences, known as genome sets (Bao et al., 2008). Genome sets effectively integrate segmented influenza virus genes and facilitate convenient searching, but manual construction is time-consuming and may not reflect real-time updates due to varying upload times and accession number organization.

H5 avian influenza viruses in Asia exhibit faster evolution, wider viral diversity, and greater inter-species transmission than those in Europe and America Continents (Dhingra et al., 2016). Thus, the GISAID-EpiFlu database was created primarily to collect virus information for avian influenza viruses. Although AIV sequences GISAID-EpiFlu are not as complete as those in the NCBI-IRD, their real-time properties of sequences from the GISAID database make them useful in tracking AIV evolution as it occurs.

The BV-BRC database provides various analysis tools for sequence comparison and monitoring variations at specific amino acid residues. It also includes exclusive

doi:10.6342/NTU202301433

information from animal surveillance, identifies sequence features in variant types, generates immune epitope data, and even includes 3D protein structures. With the emergence of COVID-19 disease, these databases have expanded to include SARS-CoV-2 viral sequences. For example, the GISAID-EpiCoV database has collected more than 14 million strains of SARS-CoV2 (Shu & McCauley, 2017). Other databases, such as NCBI-Virus (Hatcher et al., 2017), BV-BRC (Pickett et al., 2012), COVID-19 Genomics UK Consortium (COG-UK), and other government organizations, regularly release virus sequences to the public for tracking changes in SARS-CoV-2 variants. Developing computational tools for analyzing viral sequences of interest retrieved from public domain databases and integrating epidemiological, clinical, and medical information is essential for better understanding virus-host interactions.

### 1.4 Viral sequence analysis tools

Viral sequence analysis involves aligning, annotating, and comparing viral sequences in sequence datasets to identify the emerging strain with transmission potential or infection risks. The ability to handle sequencing alignments is crucial in this process. Pairwise sequence alignment is commonly used to identify variations between newly isolated and reference virus strains. Phylogenetic trees constructed from aligned sequences are also useful in visualizing various virus evolutions (Higgins & Sharp, 1988).

doi:10.6342/NTU202301433

However, dealing with a large number of viral sequences can be computationally intensive and time-consuming. Multiple sequence alignment (MSA) methods, such as Basic Local Alignment Search Tool (BLAST) (Johnson et al., 2008), Clustal Omega (Sievers et al., 2011), and Multiple Alignment using Fast Fourier Transform (MAFFT) (Katoh & Standley, 2013) have made significant advances in recent years. These methods are particularly useful as they can use partial sequences as initial seeds, which allows them to derive an optimal formula that saves computational costs from multiple partial seed sequences using dynamic programming. Subsequently, the resulting optimal formula is applied to the whole sequence alignment, enabling the determination of a consensus sequence. Consensus sequences represent the most common nucleotide or amino acid at each position in a genome set. By identifying the most prevalent nucleotide or amino acid at each position, consensus sequences can reveal the evolutionary trends of viral selection and enable a detailed analysis of conserved sequences, such as motifs, and mutations in viral genes or proteins. Conserved sequences often indicate important functional or structural elements of viral genomes and can provide insight into potential drug targets or vaccine candidates. Additionally, consensus sequences can be used to compare viral sequences across different datasets or databases and facilitate the identification of emerging strains or changes in viral diversity.

doi:10.6342/NTU202301433

Annotation of viral sequences based on known viral characteristics is crucial for understanding viral properties and predicting their potential impact on public health. For example, the BV-BRC database uses the influenza virus sequence feature variant type (Flu-SFVT) method to annotate influenza virus strains based on literature review. This method analyzes the NS1 protein sequences of influenza virus strains and categorizes them into different Flu-SFVT groups based on their  amino acid mutations and host range restriction documented in past literature (Noronha et al., 2012). FluPhenotype is another tool that records IAV amino acid signatures associated with human adaptation, enhanced virulence, and drug resistance reported in the literature, and can map genetic sequences accordingly. By inputting the viral genome or amino acid sequences into FluPhenotype, researchers can obtain predictions related to IAV HA subtypes, viral hosts, and antigenic characteristics (Lu et al., 2020).

Integrating viral sequence data with epidemiological information allows for the monitoring and tracking of viral evolution of viruses and potential risk to animal and human health. Nextstrain is an example of a tool that utilizes viral sequences collected by the GISAID database to create phylogenetic trees for viral genes. These trees can be used for virological surveillance and spatiotemporal analysis to identify single amino acid mutations. Nextstrain is also capable of grouping sequences based on time and location

doi:10.6342/NTU202301433

incorporating them into real-time maps that reveal the dynamic trends of viral transmission (Hadfield et al., 2018). Other web tools like CoVerage shows the phylogenetic dynamics of SARS-CoV-2 lineages (E. Dong et al., 2020), and CoVizu uses the real-time visualization of percentage changes at specific residues through SARS-CoV-2 genomic variations (Ferreira et al., 2021). Additionally, with continuous mutation of SARS-CoV-2, Outbreak.info Research Library based on GISAID sequence data offers a searchable platform to explore new SARS-CoV-2 variants (Tsueng et al., 2022).

In addition to analyzing viral gene sequences, more focus is being given to studying amino acid sequences for structural purposes. Various methods, to determine such as mutual information (MI) (Martin et al., 2005) or sequence correlation from a protein sequence (Goh et al., 2000), have been employed to determine the co-evolution of amino acid variations. For instance, an MI-based State transition network (STN) was generated in a study on the potential co-evolution of influenza virus and its pandemic propensity. By analyzing over 4,000 H3N2 hemagglutinin (HA) sequences from 1968 to 2008 and integrating phylogenetic trees and hemagglutination inhibition (HI) assays, the STN was able to delineate antigenic maps based on HA mutation residues and identify binding regions (Xia et al., 2009). Another study integrated MI and structural analysis to compare H5N1 and H3N2 and identified new HA co-mutated residues (Kasson & Pande, 2009).

doi:10.6342/NTU202301433

A recent study used modified MI methods to examine Polymerase Basic protein 2 (PB2) fragment and monitor PB2_627 amino acid mutation from Glutamine (E) to Lysine (K), a variation known to be associated with high pathogenicity in mammals. The findings suggested that PB2_451 co-evolved with PB2_627 and this correlation constitutes a critical species-associated amino acid residue for influenza virus replication, pathogenicity, and virulence (Gong et al., 2012).

Taken together, the development of analysis tools in viroinformatics has greatly expanded the capacity to handle large data sets, providing critical insights into viral evolution, transmission, and virulence. By identifying key amino acid residues and mutations that contribute to pathogenicity and public health, these tools offer new directions in immunological and epidemiological studies when combined with other data sources such as epidemiological and clinical data.

### 1.5 Rationale and approaches to develop integrated software for viral sequence analysis

My thesis research focuses on the notably accelerated rate of mutation in RNA viruses, which poses a significant threat to both animals and humans. For instance, the worldwide spread of clade 2.3.4 H5 avian influenza viruses (AIVs) and their reassortment with various NA proteins give rise to different subtypes. H5 AIVs are grouped into different clades, including 0, 2.3.4, 2.3.4.4a-f AIVs (Antigua et al., 2019). We know that

doi:10.6342/NTU202301433

multiple mutations can exist across gene segments of influenza viruses, and a specific mutation might impact genomic stability over time (Arai et al., 2018). H5 clade 2.3.4.4 AIVs have a higher amino acid mutation rate than clade 0, and the H5 clade 2.3.4.4 AIVs in Asia have evolved faster, exhibiting higher viral diversity, greater inter-species transmission, and a broader host range than those in Europe and the Americas (Neumann et al., 2010). As of April 24, 2023, H5N1 has led to a total of 874 human cases and 458 fatalities (WHO/GIP, 2020), indicating that this subtype has the potential to infect humans in the future. Given H5 avian influenza's capability for inter-species transmission, infecting both mammals and humans, numerous studies have predominantly focused on the emerging subtype. Recently, the emergence of new clade 2.3.4.4 H5N6 from February 2014 to June 2023 has resulted in 83 human cases and 49 deaths in China, garnering significant public health attention (Jiang et al., 2017). The WHO has warned that the infection of humans with H5 AIVs suggests a pandemic potential for H5 AIVs.

In addition, the incidence and severity of SARS-CoV-2 have far exceeded those of influenza viruses. Therefore, it is crucial to develop advanced and robust tools for analyzing viral sequences that can integrate information from different databases, including epidemiological and clinical data, to provide a comprehensive understanding of viral transmission and pathogenicity. The knowledge can help develop practical tools

doi:10.6342/NTU202301433

for controlling and preventing future pandemics.

The first part of my thesis research aims to develop an innovative and integrated software suite to analyze the entire genome of influenza virus sequences and identify novel signatures that are correlated with host-specific residues, pathogenicity, and other epidemiological characteristics that can increase the risk of a pandemic. The approach involves offering automated packages that can efficiently rearrange sequence data based on standard viral nomenclature (WHO, 1980) and translate nucleotide sequences into three potential polypeptides from 0, +1, and +2 open reading frames (ORF) following simultaneous multiple sequence alignments. The software suite that I have developed can combine sequence information across different databases and integrate viral genetic information with clinical and epidemiologic surveillance data (Yang et al., 2020). We have demonstrated the effectiveness of the new programs by analyzing the highly pathogenic avian influenza virus H5N2, which is defined by the presence of the hallmark amino acid motif (XRRKRR) at the cleavage site between the HA1 and HA2 domains, associated with the viral virulence and mammalian infections (Alexander, 2000). We have identified at least 11 additional evidence-based amino acid substitutions across different gene segments of H5N2 avian influenza viruses that could contribute to viral virulence and mammalian infections (Yang et al., 2020).

The second part of my study focuses on the analysis of SARS-CoV-2 sequences,

doi:10.6342/NTU202301433

utilizing the software developed for AIVs with some modifications. Although Taiwan did not experience widespread community outbreaks of SARS-CoV2 until mid-April 2021, multiple waves of pandemic have occurred globally since 2020. To investigate the possible source of infection and identify epidemiological conditions that facilitated viral spread in the community, we collected 101 strains with whole genome sequences. Our analysis revealed that a predominant strain of the SARS-CoV2 lineage B.1.1.7 (Alpha) variant was predominantly transmitted during the early phase of the outbreak. Its disappearance was correlated with the implementation of multiple layers of disease control measures. Through my research, I have demonstrated that our software can effectively retrieve and analyze viral sequence information from public domain databases, enabling efficient monitoring of dynamic viral shifts and the emergence of novel viral variants with pandemic potential.

doi:10.6342/NTU202301433

# Chapter 2
# Methodology

doi:10.6342/NTU202301433

The viral sequences utilized in the avian influenza or SARS CoV-2 study were obtained from publicly accessible databases. Upon retrieval, these sequences underwent a series of processing steps to ensure data quality. This included the removal of incomplete, duplicated, or erroneous data for a thorough quality check. Subsequently, annotation, alignment, and translation into amino acid sequences were performed, and the resulting dataset was organized and prepared for further analysis. The workflow of sequence analysis was illustrated in Figure I.

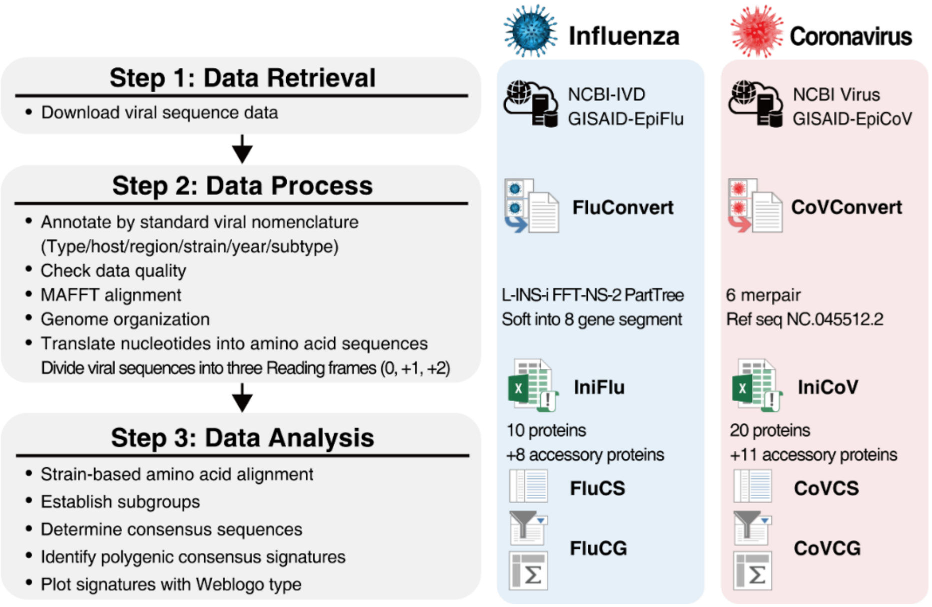

Influenza
Coronavirus
Step 1: Data Retrieval
• Download viral sequence data
NCBI-IVD
GISAID-EpiFlu
NCBI Virus
GISAID-EpiCoV
Step 2: Data Process
• Annotate by standard viral nomenclature (Type/host/region/strain/year/subtype)
• Check data quality
• MAFFT alignment
• Genome organization
• Translate nucleotides into amino acid sequences
Divide viral sequences into three Reading frames (0, +1, +2)
FluConvert
CoVConvert
L-INS-i FFT-NS-2 PartTree
Soft into 8 gene segment
6 merpair
Ref seq NC.045512.2
Step 3: Data Analysis
• Strain-based amino acid alignment
• Establish subgroups
• Determine consensus sequences
• Identify polygenic consensus signatures
• Plot signatures with Weblogo type
IniFlu
10 proteins
+8 accessory proteins
IniCoV
20 proteins
+11 accessory proteins
FluCS
CoVCS
FluCG
CoVCG

**Figure I. Workflow of data analysis** The stepwise processes performed by our developed software (bold text in the colored box on the right) to identify novel signatures of emerging viruses with increasing risk are described as follows. **Step 1:** Viral sequences are obtained from the public-domain databases. **Step 2**: The program automatically annotates and validates the quality of viral sequences, uses different algorithms of MAFFT alignment based on length or the number of sequences, and organizes according to the different viral genomes for subsequent translation into amino acid sequences. **Step 3:** These modules perform strain-based alignments of viral amino sequences, regroup viral strains with epidemiological significance, and compute a consensus sequence for each subgroup. Subsequently, the subgroup-specific unique polygenic amino acid signatures can be simultaneously identified.

doi:10.6342/NTU202301433

### 2.1 Data sources and file format

We downloaded viral sequence data from the GISAID and NCBI databases. To achieve a universal collection of sequences from different databases, we retrieved the FASTA file format, which is a text filetype commonly used in the field of bioinformatics to preserve multiple nucleotides or amino acid sequences, each preceded by a one-line description (also known as a header) that begins with a ">" symbol. By processing the header in viral sequences according to the standard viral nomenclature (ABCD Type/Host/Region/Strain/Year/HxNy Subtype), the header can facilitate the linkage of the virus sequence to the strain's unique identification in different databases (e.g., CY009444_A_human_PuertoRico_8_1934_H1N1_human). Therefore, the customized header definition can be used to change the header with the standard viral nomenclature, such as NCBI defined the sequences as “>[accession] [strain] [segment] [serotype] [host]” and GISAID as “Isolate IDEPI Isolate name Segment number HxNy host.” Acquiring these FASTA sequence files through customized header definitions will facilitate subsequent automated processing of sequence data and enable more accurate analysis.

### 2.2 Data processing

To organize the information in a better format for further analysis, I developed a suite of integrated software based on the Microsoft Windows operating platform for non-programming background users in handling vast amounts of sequence data. To this end,

doi:10.6342/NTU202301433

users only need to put the FASTA file in the input folder and run the program, which automatically selects the appropriate algorithms to process the sequences, following these four steps: arrangement, validation of data quality, alignment, and algorithm selection, translation. The suite of programs' essence lies in its code with the batch scripts languages, such as the batch scripts languages of shell and PowerShell, enabling the textual files of sequence data can be efficiently consolidated and converted into analyzable tables. Detailed information on program download, installation, and usage instructions can be found in the published paper (Yang et al., 2020). Below section described the methods in detail.

### 2.2.1 Sequence arrangement

Influenza viruses possess eight segments of single-stranded RNA (ssRNA) in the genome. They can be sequenced and deposited to the database individually. Therefore, retrieved sequences were saved to eight separate files based on the gene segment and subsequently combined into a single genome based on the  the strain name information associated with each gene segment. As the results, each strain was placed in the order of the standard viral nomenclature, including the type, host, region, strain, year, and subtype within parentheses (WHO, 1980), to normalize the virus name for next step of examination and validation of virus strain name.

doi:10.6342/NTU202301433

### 2.2.2 Validation of data quality

The sequence data validation process can be divided into two parts. One was to ensure that the sequence names were fully complied with the standard viral nomenclature. Any sequence that failed to contain all  six items of standard viral nomenclature or with duplicative name was removed. We also generated an "excluding list" which contained previously identified erroneous sequences, duplicative sequences, sequences with inaccurate information, and those labeled as "retracted sequences" in the databases. All rearranged sequences were inspected to delete those which matched any items in the "excluding list" to ensure data integrity. The second part of data validation was to remove sequences with errors aroused from any of the conditions: (1) nucleotide sequences containing interspersed amino acid sequences or other erroneous textual data, (2) sequences longer than the expected lengths of the genome template, and (3) presence of redundant and meaningless deletions or residues (denoted as 'n', 'x' or '-'). Following these two parts of validation, the resultant high-precision dataset not only preserves the original virus sequences and corresponding information, but it also reduces the probability of sequence processing being erroneous aborted due to meaningless sequences that overflow the value of the original virus template, thereby enhancing the success rate of sequence processing.

doi:10.6342/NTU202301433

### 2.2.3 Sequence alignment and algorithm selection

Traditional pairwise alignment methods have been commonly used to compare two sequences. However, it would become a time-consuming process ($L^N$) when confronted with longer whole-genome sequences (length '$L$') or the need to compare thousands of sequences (number of sequences '$N$'). This can limit the analysis capability in terms of sequence length and quantity. With the advent of multiple sequence alignment (MSA) algorithms, it significantly saves computing time. To handle the challenge of large-scale whole-genome sequence analysis, we implemented the increasingly popular MAFFT multiple sequence alignment program (version 7.52) for nucleotide and amino acid sequences (Katoh & Standley, 2013). MAFFT incorporates the dynamic programming methods (Needleman & Wunsch, 1970), the progressive alignment methods (Feng & Doolittle, 1987) and the iterative refinement methods (Berger & Munson, 1991), coupled with Fourier transformations, to calculate and reduce the dimensionality of the sequence matrix, thereby resulting in the computing time to approximate the sequence length $L$. Depending on the sequence length and quantity, it automatically selects the best fitting algorithm [i.e., L-INS-i (accurate) for aligning <∼200 viral strains/files; FFT-NS-2 (fast) for aligning <∼30,000 viral strains/files to achieve maximal efficiency; and PartTree (fast) for aligning > ∼30,000 viral strains/files] (Katoh & Toh, 2007). These algorithms detect

doi:10.6342/NTU202301433

the sequences and adjust the MSA methods correspondingly, optimizing time and accuracy to present an efficient sequence alignment.

### 2.2.4 Sequence translation

As viruses exploit different Open Reading Frames (ORFs) and mRNA alternative splicing to translate various proteins and accessory proteins, it is crucial to consider the possibilities of different ORFs when translate nucleotide sequences into amino acid polypeptides. Therefore, the aligned sequences were translated into three possible polypeptides from ORF 0, +1, and +2 using the program employing the EMBOSS Transeq version 6.5 code library (Rice et al., 2000).These translated polypeptides were converted to comma-delimited (csv) text files to establish tables for subsequent analysis.

## 2.3 Sequence data analysis

To actualize a platform for the visual analysis of sequences, we proffer a graphical user interface (GUI) for non-programming users, providing a consolidated platform for automated sequence organization and analysis tools. Here, I developed a suite of integrated programs based on the Microsoft Excel spreadsheet software through the scripts language of Visual Basic for Applications (VBA), providing a user-friendly GUI to analyze the sequence data. The advantage lies in incorporating various modular plugins (also known as add-ins) into Excel through the VBA language, allowing data to regroup

doi:10.6342/NTU202301433

and analyze sequences with a simple click. Next, I will present the data input, grouping, and analysis tools for the sequences.

### 2.3.1 Selection of open reading frames and alternative splicing

The translated three polypeptides (ORF 0, +1, and +2) of each gene segment were compared to all possible currently known amino acid sequences that can form proteins, alternatively spliced isoforms, and accessory proteins of influenza viruses. Because that each protein's N-terminus and C-terminus possess different short-conserved sequences (CS) (4~6 aa) from the region of the transcription-regulatory sequence (Kim et al., 2020; Lai, 1990), scanning these CS on the three translated polypeptides allows for determining the amino acid sequence to which protein. Some proteins are formed through mRNA alternative splicing selected between two translated polypeptides. By identifying specific sequence positions for splicing, these sequences can be combined to generate accessory proteins. Finally, scanning these processed amino acid sequences, our program assigned sequential numbering to the residue positions starting from the first methionine. Since they were shown in Excel spreadsheet with the virus-specific genome template, the protein sequences can be easily visualized.

### 2.3.2 Strain-based alignment

Using standard viral nomenclature can simplify the grouping and analysis of protein

doi:10.6342/NTU202301433

sequence data, making it easier to visualize epidemiological and virological information associated with virus sequences. Our designed software allows inputting a list of strain names labeled with standard viral nomenclature. Whether segmented RNA viruses such as influenza viruses or positive-strand RNA viruses such as SARS-CoV-2, standard viral nomenclature can link different genes and provide more items for grouping selection. Given that the filled amino acid sequences based on virus-specific genome templates already carried the six items of standard viral nomenclature (type, host, region, strain, year, and subtype), each amino acid residue of the virus sequence can be compared with the list of strain names, the sequence is filled in the templates and arranged following the list as one strain. After the strain-based alignment process, an aligned sequence matrix is formed with the standard viral nomenclature as the leading identifier, and following each amino acid is filled into the table for grouping purposes.

### 2.3.3 Grouping of viral sequences

The sequence matrix constructed through the strain-based alignment which incorporated the six items of the standard viral nomenclature and sequence into an Excel spreadsheet allowed flexibility in analyzing sequence data by group with re-alignment of the sequences. Furthermore, using intersection filtering of multiple grids allowed for multiple subgroupings and linking epidemiological information with amino acid residues.

doi:10.6342/NTU202301433

This process established groups for subsequent comparison, consensus sequence calculation, and identifying unique polygenic consensus signatures.

### 2.3.4 Determining the consensus sequence

We employed a calculation method to determine the most frequent amino acid in each residue of a specific protein, thereby to identify the most indicative sequence in the grouping and the "consensus signature" can be then generated (Yang et al., 2020).

### 2.3.5 Identification and annotation of polygenic consensus signatures

Since mutations can occur in multiple genes across the viral genome, the identification of polygenic consensus through the software analysis can monitor the viral temporal dynamic changes and epidemiological significance. The most representative (i.e., most frequent) amino acid sequence at each position of the whole genome can be shown through various grouping, subsequently allowing for the derivation and differentiation of consensus sequences for each subgroup.

doi:10.6342/NTU202301433

# Chapter 3

# Cited References

## Introduction and Methodology

## Cited References

Alexander, D. J. (2000). A review of avian influenza in different bird species. Vet Microbiol, 74(1-2), 3-13. https://doi.org/10.1016/s0378-1135(00)00160-7

Antigua, K. J. C., Choi, W. S., Baek, Y. H., & Song, M. S. (2019). The Emergence and Decennary Distribution of Clade 2.3.4.4 HPAI H5Nx. Microorganisms, 7(6). https://doi.org/10.3390/microorganisms7060156

Arai, Y., Kawashita, N., Hotta, K., Hoang, P. V. M., Nguyen, H. L. K., Nguyen, T. C., Vuong, C. D., Le, T. T., Le, M. T. Q., Soda, K., Ibrahim, M. S., Daidoji, T., Takagi, T., Shioda, T., Nakaya, T., Ito, T., Hasebe, F., & Watanabe, Y. (2018). Multiple polymerase gene mutations for human adaptation occurring in Asian H5N1 influenza virus clinical isolates. Scientific Reports, 8(1), 13066. https://doi.org/10.1038/s41598-018-31397-3

Ardui, S., Ameur, A., Vermeesch, J. R., & Hestand, M. S. (2018). Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. Nucleic Acids Res, 46(5), 2159-2168. https://doi.org/10.1093/nar/gky066

Balzer, S., Malde, K., Grohme, M. A., & Jonassen, I. (2013). Filtering duplicate reads from 454 pyrosequencing data. Bioinformatics, 29(7), 830-836. https://doi.org/10.1093/bioinformatics/btt047

Bao, Y., Bolotov, P., Dernovoy, D., Kiryutin, B., Zaslavsky, L., Tatusova, T., Ostell, J., & Lipman, D. (2008). The influenza virus resource at the National Center for Biotechnology Information. J Virol, 82(2), 596-601. https://doi.org/10.1128/jvi.02005-07

Baz, M., Abed, Y., Simon, P., Hamelin, M. E., & Boivin, G. (2010). Effect of the neuraminidase mutation H274Y conferring resistance to oseltamivir on the replicative capacity and virulence of old and recent human influenza A(H1N1) viruses. J Infect Dis, 201(5), 740-745. https://doi.org/10.1086/650464

Berger, M. P., & Munson, P. J. (1991). A novel randomized iterative strategy for aligning multiple protein sequences. Comput Appl Biosci, 7(4), 479-484. https://doi.org/10.1093/bioinformatics/7.4.479

Bloom, J. D., Gong, L. I., & Baltimore, D. (2010). Permissive secondary mutations enable the evolution of influenza oseltamivir resistance. Science, 328(5983), 1272-1275. https://doi.org/10.1126/science.1187816

Borges, V., Pinheiro, M., Pechirra, P., Guiomar, R., & Gomes, J. P. (2018). INSaFLU: an automated open web-based bioinformatics suite "from-reads" for influenza whole-genome-sequencing-based surveillance. Genome Medicine, 10(1), 46. https://doi.org/10.1186/s13073-018-0555-0

Brister, J. R., Ako-Adjei, D., Bao, Y., & Blinkova, O. (2015). NCBI viral genomes resource. Nucleic Acids Res, 43(Database issue), D571-577. https://doi.org/10.1093/nar/gku1207

doi:10.6342/NTU202301433

Canard, B., & Sarfati, R. S. (1994). DNA polymerase fluorescent substrates with reversible 3′-tags. Gene, 148(1), 1-6. https://doi.org/https://doi.org/10.1016/0378-1119(94)90226-7

Chen, Y. C., Liu, T., Yu, C. H., Chiang, T. Y., & Hwang, C. C. (2013). Effects of GC bias in next-generation-sequencing data on de novo genome assembly. PLoS One, 8(4), e62856. https://doi.org/10.1371/journal.pone.0062856

Collins, F. S., Morgan, M., & Patrinos, A. (2003). The Human Genome Project: lessons from large-scale biology. Science, 300(5617), 286-290. https://doi.org/10.1126/science.1084564

Dawood, F. S., Iuliano, A. D., Reed, C., Meltzer, M. I., Shay, D. K., Cheng, P. Y., Bandaranayake, D., Breiman, R. F., Brooks, W. A., Buchy, P., Feikin, D. R., Fowler, K. B., Gordon, A., Hien, N. T., Horby, P., Huang, Q. S., Katz, M. A., Krishnan, A., Lal, R., . . . Widdowson, M. A. (2012). Estimated global mortality associated with the first 12 months of 2009 pandemic influenza A H1N1 virus circulation: a modelling study [Article]. Lancet Infectious Diseases, 12(9), 687-695. https://doi.org/10.1016/s1473-3099(12)70121-4

Deamer, D., Akeson, M., & Branton, D. (2016). Three decades of nanopore sequencing. Nature Biotechnology, 34(5), 518-524. https://doi.org/10.1038/nbt.3423

Delahaye, C., & Nicolas, J. (2021). Sequencing DNA with nanopores: Troubles and biases. PLoS One, 16(10), e0257521. https://doi.org/10.1371/journal.pone.0257521

Dhingra, M. S., Artois, J., Robinson, T. P., Linard, C., Chaiban, C., Xenarios, I., Engler, R., Liechti, R., Kuznetsov, D., Xiao, X., Dobschuetz, S. V., Claes, F., Newman, S. H., Dauphin, G., & Gilbert, M. (2016). Global mapping of highly pathogenic avian influenza H5N1 and H5Nx clade 2.3.4.4 viruses with spatial cross-validation. eLife, 5, e19571. https://doi.org/10.7554/eLife.19571

Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. The Lancet Infectious Diseases, 20(5), 533-534. https://doi.org/https://doi.org/10.1016/S1473-3099(20)30120-1

Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis, 20(5), 533-534. https://doi.org/10.1016/s1473-3099(20)30120-1

Feng, D. F., & Doolittle, R. F. (1987). Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J Mol Evol, 25(4), 351-360. https://doi.org/10.1007/bf02603120

Ferreira, R.-C., Wong, E., Gugan, G., Wade, K., Liu, M., Baena, L. M., Chato, C., Lu, B., Olabode, A. S., & Poon, A. F. Y. (2021). CoVizu: Rapid analysis and visualization of the global diversity of SARS-CoV-2 genomes. Virus Evolution, 7(2). https://doi.org/10.1093/ve/veab092

Flusberg, B. A., Webster, D. R., Lee, J. H., Travers, K. J., Olivares, E. C., Clark, T. A., Korlach, J., & Turner, S. W. (2010). Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat Methods, 7(6), 461-465. https://doi.org/10.1038/nmeth.1459

Garoli, D., Yamazaki, H., Maccaferri, N., & Wanunu, M. (2019). Plasmonic Nanopores for Single-Molecule Detection and Manipulation: Toward Sequencing Applications. Nano Lett, 19(11), 7553-7562. https://doi.org/10.1021/acs.nanolett.9b02759

Goh, C. S., Bogan, A. A., Joachimiak, M., Walther, D., & Cohen, F. E. (2000). Co-evolution of proteins with their interaction partners. Journal of molecular biology, 299(2), 283-293. https://doi.org/10.1006/jmbi.2000.3732

Gong, Y.-N., Chen, G.-W., & Suchard, M. A. (2012). A novel empirical mutual information approach to identify co-evolving amino acid positions of influenza A viruses. Computational Biology and Chemistry, 39, 20-28. https://doi.org/https://doi.org/10.1016/j.compbiolchem.2012.06.004

Guo, J., Xu, N., Li, Z., Zhang, S., Wu, J., Kim, D. H., Sano Marma, M., Meng, Q., Cao, H., Li, X., Shi, S., Yu, L., Kalachikov, S., Russo, J. J., Turro, N. J., & Ju, J. (2008). Four-color DNA sequencing with 3'-O-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. Proc Natl Acad Sci U S A, 105(27), 9145-9150. https://doi.org/10.1073/pnas.0804023105

Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., & Neher, R. A. (2018). Nextstrain: real-time tracking of pathogen evolution. Bioinformatics, 34(23), 4121-4123. https://doi.org/10.1093/bioinformatics/bty407

Hatcher, E. L., Zhdanov, S. A., Bao, Y., Blinkova, O., Nawrocki, E. P., Ostapchuck, Y., Schäffer, A. A., & Brister, J. R. (2017). Virus Variation Resource - improved response to emergent viral outbreaks. Nucleic Acids Res, 45(D1), D482-d490. https://doi.org/10.1093/nar/gkw1065

Head, S. R., Komori, H. K., LaMere, S. A., Whisenant, T., Van Nieuwerburgh, F., Salomon, D. R., & Ordoukhanian, P. (2014). Library construction for next-generation sequencing: overviews and challenges. Biotechniques, 56(2), 61-64, 66, 68, passim. https://doi.org/10.2144/000114133

Higgins, D. G., & Sharp, P. M. (1988). CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene, 73(1), 237-244. https://doi.org/https://doi.org/10.1016/0378-1119(88)90330-7

Hoenen, T., Groseth, A., Rosenke, K., Fischer, R. J., Hoenen, A., Judson, S. D., Martellaro, C., Falzarano, D., Marzi, A., Squires, R. B., Wollenberg, K. R., de Wit, E., Prescott, J., Safronetz, D., van Doremalen, N., Bushmaker, T., Feldmann, F., McNally, K., Bolay, F. K., . . . Feldmann, H. (2016). Nanopore Sequencing as a Rapidly Deployable Ebola Outbreak Tool. Emerg Infect Dis, 22(2), 331-334. https://doi.org/10.3201/eid2202.151796

Hourdel, V., Kwasiborski, A., Balière, C., Matheus, S., Batéjat, C. F., Manuguerra, J. C., Vanhomwegen, J., & Caro, V. (2020). Rapid Genomic Characterization of SARS-CoV-2 by Direct Amplicon-Based Sequencing Through Comparison of MinION and Illumina iSeq100(TM) System. Front Microbiol, 11, 571328. https://doi.org/10.3389/fmicb.2020.571328

Hurt, A. C., Holien, J. K., Parker, M. W., & Barr, I. G. (2009). Oseltamivir resistance and the H274Y neuraminidase mutation in seasonal, pandemic and highly pathogenic influenza viruses. Drugs, 69(18), 2523-2531. https://doi.org/10.2165/11531450-000000000-00000

Jain, M., Olsen, H. E., Paten, B., & Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biology, 17(1), 239. https://doi.org/10.1186/s13059-016-1103-0

Jiang, H., Wu, P., Uyeki, T. M., He, J., Deng, Z., Xu, W., Lv, Q., Zhang, J., Wu, Y., Tsang, T. K., Kang, M., Zheng, J., Wang, L., Yang, B., Qin, Y., Feng, L., Fang, V. J., Gao, G. F., Leung, G. M., . . . Cowling, B. J. (2017). Preliminary Epidemiologic Assessment of Human Infections With Highly Pathogenic Avian Influenza A(H5N6) Virus, China. Clin Infect Dis, 65(3), 383-388. https://doi.org/10.1093/cid/cix334

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., & Madden, T. L. (2008). NCBI BLAST: a better web interface. Nucleic Acids Research, 36(suppl_2), W5-W9. https://doi.org/10.1093/nar/gkn201

Kasson, P. M., & Pande, V. S. (2009). Combining mutual information with structural analysis to screen for functionally important residues in influenza hemagglutinin. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, 492-503. https://doi.org/10.1142/9789812836939_0047

Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution, 30(4), 772-780. https://doi.org/10.1093/molbev/mst010

Katoh, K., & Toh, H. (2007). PartTree: an algorithm to build an approximate tree from a large number of unaligned sequences. Bioinformatics, 23(3), 372-374. https://doi.org/10.1093/bioinformatics/btl592

Kim, D., Lee, J. Y., Yang, J. S., Kim, J. W., Kim, V. N., & Chang, H. (2020). The Architecture of SARS-CoV-2 Transcriptome. Cell, 181(4), 914-921.e910. https://doi.org/10.1016/j.cell.2020.04.011

Kim, H., Jebrail, M. J., Sinha, A., Bent, Z. W., Solberg, O. D., Williams, K. P., Langevin, S. A., Renzi, R. F., Van De Vreugde, J. L., Meagher, R. J., Schoeniger, J. S., Lane, T. W., Branda, S. S., Bartsch, M. S., & Patel, K. D. (2013). A microfluidic DNA library preparation platform for next-generation sequencing. PLoS One, 8(7), e68988. https://doi.org/10.1371/journal.pone.0068988

Kruczkiewicz, Peter, Nguyen, Hai Hoang, & Lung, Oliver. (2022). CFIA-NCFAD/nf-flu v3.1.0 (3.1.0). Zenodo. https://doi.org/10.5281/zenodo.7011213

Lai, M. M. C. (1990). CORONAVIRUS: ORGANIZATION, REPLICATION AND EXPRESSION OF GENOME. Annual Review of Microbiology, 44(1), 303-303. https://doi.org/10.1146/annurev.mi.44.100190.001511

Lee, D.-H., Bertran, K., Kwon, J.-H., & Swayne, D. E. (2017). Evolution, global spread, and pathogenicity of highly pathogenic avian influenza H5Nx clade 2.3.4.4. J Vet Sci, 18(S1), 269-280. https://doi.org/10.4142/jvs.2017.18.S1.269

Lee, D. H., Torchetti, M. K., Winker, K., Ip, H. S., Song, C. S., & Swayne, D. E. (2015). Intercontinental Spread of Asian-Origin H5N8 to North America through Beringia by Migratory Birds. Journal of Virology, 89(12), 6521-6524. https://doi.org/10.1128/jvi.00728-15

Lu, C., Cai, Z., Zou, Y., Zhang, Z., Chen, W., Deng, L., Du, X., Wu, A., Yang, L., Wang, D., Shu, Y., Jiang, T., & Peng, Y. (2020). FluPhenotype-a one-stop platform for early warnings of the influenza A virus. Bioinformatics, 36(10), 3251-3253. https://doi.org/10.1093/bioinformatics/btaa083

Ma, S., Murphy, T. W., & Lu, C. (2017). Microfluidics for genome-wide studies involving next generation sequencing. Biomicrofluidics, 11(2), 021501. https://doi.org/10.1063/1.4978426

Mardis, E. R. (2008). Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet, 9, 387-402. https://doi.org/10.1146/annurev.genom.9.081307.164359

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., . . . Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature, 437(7057), 376-380. https://doi.org/10.1038/nature03959

Martin, L. C., Gloor, G. B., Dunn, S. D., & Wahl, L. M. (2005). Using information theory to search for co-evolving residues in proteins. Bioinformatics, 21(22), 4116-4124. https://doi.org/10.1093/bioinformatics/bti671

Mena, I., Nelson, M. I., Quezada-Monroy, F., Dutta, J., Cortes-Fernández, R., Lara-Puente, J. H., Castro-Peralta, F., Cunha, L. F., Trovão, N. S., Lozano-Dubernard, B., Rambaut, A., van Bakel, H., & García-Sastre, A. (2016). Origins of the 2009 H1N1 influenza pandemic in swine in Mexico. eLife, 5, e16777. https://doi.org/10.7554/eLife.16777

Merriman, B., & Rothberg, J. M. (2012). Progress in ion torrent semiconductor chip based sequencing. Electrophoresis, 33(23), 3397-3417. https://doi.org/10.1002/elps.201200424

Metzker, M. L. (2010). Sequencing technologies — the next generation. Nature Reviews Genetics, 11(1), 31-46. https://doi.org/10.1038/nrg2626

Meyer, M., & Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb Protoc, 2010(6), pdb.prot5448. https://doi.org/10.1101/pdb.prot5448

doi:10.6342/NTU202301433

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol, 48(3), 443-453. https://doi.org/10.1016/0022-2836(70)90057-4

Neumann, G., Chen, H., Gao, G. F., Shu, Y., & Kawaoka, Y. (2010). H5N1 influenza viruses: outbreaks and biological properties. Cell research, 20(1), 51-61. https://doi.org/10.1038/cr.2009.124

Noronha, J. M., Liu, M., Squires, R. B., Pickett, B. E., Hale, B. G., Air, G. M., Galloway, S. E., Takimoto, T., Schmolke, M., Hunt, V., Klem, E., García-Sastre, A., McGee, M., & Scheuermann, R. H. (2012). Influenza virus sequence feature variant type analysis: evidence of a role for NS1 in influenza virus host range restriction. J Virol, 86(10), 5857-5866. https://doi.org/10.1128/jvi.06901-11

Nyren, P., Pettersson, B., & Uhlen, M. (1993). Solid Phase DNA Minisequencing by an Enzymatic Luminometric Inorganic Pyrophosphate Detection Assay. Analytical Biochemistry, 208(1), 171-175. https://doi.org/https://doi.org/10.1006/abio.1993.1024

Olson, R. D., Assaf, R., Brettin, T., Conrad, N., Cucinell, C., Davis, James J., Dempsey, Donald M., Dickerman, A., Dietrich, Emily M., Kenyon, Ronald W., Kuscuoglu, M., Lefkowitz, Elliot J., Lu, J., Machi, D., Macken, C., Mao, C., Niewiadomska, A., Nguyen, M., Olsen, Gary J., . . . Stevens, Rick L. (2022). Introducing the Bacterial and Viral Bioinformatics Resource Center (BV-BRC): a resource combining PATRIC, IRD and ViPR. Nucleic Acids Research, 51(D1), D678-D689. https://doi.org/10.1093/nar/gkac1003

Peng, X., Liu, F., Wu, H., Peng, X., Xu, Y., Wang, L., Chen, B., Sun, T., Yang, F., Ji, S., & Wu, N. (2018). Amino Acid Substitutions HA A150V, PA A343T, and PB2 E627K Increase the Virulence of H5N6 Influenza Virus in Mice. Front Microbiol, 9, 453. https://doi.org/10.3389/fmicb.2018.00453

Pickett, B. E., Sadat, E. L., Zhang, Y., Noronha, J. M., Squires, R. B., Hunt, V., Liu, M., Kumar, S., Zaremba, S., Gu, Z., Zhou, L., Larson, C. N., Dietrich, J., Klem, E. B., & Scheuermann, R. H. (2012). ViPR: an open bioinformatics database and analysis resource for virology research. Nucleic Acids Res, 40(Database issue), D593-598. https://doi.org/10.1093/nar/gkr859

Rice, P., Longden, I., & Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics, 16(6), 276-277. https://doi.org/10.1016/S0168-9525(00)02024-2

Rios, G., Lacoux, C., Leclercq, V., Diamant, A., Lebrigand, K., Lazuka, A., Soyeux, E., Lacroix, S., Fassy, J., Couesnon, A., Thiery, R., Mari, B., Pradier, C., Waldmann, R., & Barbry, P. (2021). Monitoring SARS-CoV-2 variants alterations in Nice neighborhoods by wastewater nanopore sequencing. Lancet Reg Health Eur, 10, 100202. https://doi.org/10.1016/j.lanepe.2021.100202

doi:10.6342/NTU202301433

Sahlin, K., & Medvedev, P. (2021). Error correction enables use of Oxford Nanopore technology for reference-free transcriptome analysis. Nature Communications, 12(1), 2. https://doi.org/10.1038/s41467-020-20340-8

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences, 74(12), 5463-5467. https://doi.org/doi:10.1073/pnas.74.12.5463

Scheible, M., Loreille, O., Just, R., & Irwin, J. (2014). Short tandem repeat typing on the 454 platform: strategies and considerations for targeted sequencing of common forensic markers. Forensic Sci Int Genet, 12, 107-119. https://doi.org/10.1016/j.fsigen.2014.04.010

Shu, Y., & McCauley, J. (2017). GISAID: Global initiative on sharing all influenza data - from vision to reality. Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin, 22(13), 30494. https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J., Thompson, J. D., & Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol, 7, 539. https://doi.org/10.1038/msb.2011.75

Stokel-Walker, C., & Van Noorden, R. (2023). What ChatGPT and generative AI mean for science. Nature, 614(7947), 214-216. https://doi.org/10.1038/d41586-023-00340-6

Tsueng, G., Mullen, J. L., Alkuzweny, M., Cano, M., Rush, B., Haag, E., Curators, O., Lin, J., Welzel, D. J., Zhou, X., Qian, Z., Latif, A. A., Hufbauer, E., Zeller, M., Andersen, K. G., Wu, C., Su, A. I., Gangavarapu, K., & Hughes, L. D. (2022). Outbreak.info Research Library: A standardized, searchable platform to discover and explore COVID-19 resources. bioRxiv. https://doi.org/10.1101/2022.01.20.477133

Tucker, T., Marra, M., & Friedman, J. M. (2009). Massively Parallel Sequencing: The Next Big Thing in Genetic Medicine. The American Journal of Human Genetics, 85(2), 142-154. https://doi.org/https://doi.org/10.1016/j.ajhg.2009.06.022

Wang, Y., Zhao, Y., Bollas, A., Wang, Y., & Au, K. F. (2021). Nanopore sequencing technology, bioinformatics and applications. Nature Biotechnology, 39(11), 1348-1365. https://doi.org/10.1038/s41587-021-01108-x

Weber, M., Sediri, H., Felgenhauer, U., Binzen, I., Bänfer, S., Jacob, R., Brunotte, L., García-Sastre, A., Schmid-Burgk, Jonathan L., Schmidt, T., Hornung, V., Kochs, G., Schwemmle, M., Klenk, H.-D., & Weber, F. (2015). Influenza Virus Adaptation PB2-627K Modulates Nucleocapsid Inhibition by the Pathogen Sensor RIG-I. Cell Host & Microbe, 17(3), 309-319. https://doi.org/https://doi.org/10.1016/j.chom.2015.01.005

WHO. (1980). A revision of the system of nomenclature for influenza viruses: a WHO Memorandum. Bulletin of the World Health Organization, 58(4), 585-591. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2395936/

doi:10.6342/NTU202301433

WHO/GIP. (2020). Cumulative number of confirmed human cases of avian influenza A(H5N1) reported to WHO.

Xia, Z., Jin, G., Zhu, J., & Zhou, R. (2009). Using a mutual information-based site transition network to map the genetic evolution of influenza A/H3N2 virus. Bioinformatics, 25(18), 2309-2317. https://doi.org/10.1093/bioinformatics/btp423

Yang, C. R., King, C. C., Liu, L. D., & Ku, C. C. (2020). FluConvert and IniFlu: a suite of integrated software to identify novel signatures of emerging influenza viruses with increasing risk. BMC Bioinformatics, 21(1), 316. https://doi.org/10.1186/s12859-020-03650-y

Yang, L., Zhu, W., Li, X., Bo, H., Zhang, Y., Zou, S., Gao, R., Dong, J., Zhao, X., Chen, W., Dong, L., Zou, X., Xing, Y., Wang, D., & Shu, Y. (2017). Genesis and Dissemination of Highly Pathogenic H5N6 Avian Influenza Viruses. J Virol, 91(5). https://doi.org/10.1128/jvi.02199-16

Zhang, Y., Aevermann, B. D., Anderson, T. K., Burke, D. F., Dauphin, G., Gu, Z., He, S., Kumar, S., Larsen, C. N., Lee, A. J., Li, X., Macken, C., Mahaffey, C., Pickett, B. E., Reardon, B., Smith, T., Stewart, L., Suloway, C., Sun, G., . . . Scheuermann, R. H. (2017). Influenza Research Database: An integrated bioinformatics resource for influenza virus research. Nucleic Acids Res, 45(D1), D466-d474. https://doi.org/10.1093/nar/gkw857

Zhu, W., Li, L., Yan, Z., Gan, T., Li, L., Chen, R., Chen, R., Zheng, Z., Hong, W., Wang, J., Smith, D. K., Guan, Y., Zhu, H., & Shu, Y. (2015). Dual E627K and D701N mutations in the PB2 protein of A(H7N9) influenza virus increased its virulence in mammalian models. Scientific Reports, 5(1), 14170. https://doi.org/10.1038/srep14170

# Chapter 4

# PART I

## FluConvert and IniFlu: a suite of integrated software to identify novel signatures of emerging influenza viruses with increasing risk

BMC Bioinformatics 21, 316 (2020)
Refer the reprint to Appendix I

doi:10.6342/NTU202301433

# Chapter 5

# PART II

## The Emergence and Successful Elimination of SARS-CoV-2 Dominant Strains with Increasing Epidemic Potential in Taiwan's 2021 Outbreak

Submitted

doi:10.6342/NTU202301433

### 5.1 Abstract

Taiwan’s experience with SARS-CoV in 2003 guided its development of strategies to defend against SARS-CoV-2 in 2020, which enabled the successful control of COVID-19 cases from 2020 through March 2021. However, in late-April 2021, the imported Alpha variant began to cause COVID-19 outbreaks at an exceptional rate in Taiwan. In this study, we aimed to determine what epidemiological conditions enabled the SARS-CoV-2 Alpha variant strains to become dominant and decline later during a surge in the Outbreak. In conjunction with contact-tracing investigations, we used our bioinformatics software, CoVConvert and IniCoV, to analyze whole-genome sequences of 101 Taiwan Alpha strains. Univariate and multivariable regression analyses revealed the factors associated with viral dominance. Univariate analysis showed the dominant Alpha strains were preferentially selected in the surge’s epicenter (p = 0.0024) through intensive human-to-human contact and maintained their dominance for 1.5 months until the Zero-COVID Policy was implemented. Multivariable regression found that the epidemic periods (p = 0.007) and epicenter (p = 0.001) were two significant factors associated with the community-spread dominant viruses. The dominant strains emerged at the outbreak’s epicenter with frequent human-to-human contact and low vaccination coverage. The Level 3 Restrictions and Zero-COVID policy successfully controlled the outbreak in the

community without city lockdowns. Our integrated method can identify the epidemiological conditions for emerging dominant virus with increasing epidemiological potential and support decision makers in rapidly containing outbreaks using public health measures that target fast-spreading virus strains.

## 5.2 Introduction

SARS-CoV-2, an emerging virus has caused over 624 million COVID-19 cases and nearly 6.56 million deaths worldwide by October 18, 2022 (E. Dong et al., 2020). The Taiwan Centers for Disease Control (Taiwan CDC) quickly responded to the SARS-CoV-2 pandemic with early border control measures on December 31, 2019, drawing upon lessons from the SARS-CoV outbreaks in 2003 (Hsueh & Yang, 2005). Although three incidences of limited community spread occurred from 2019 to mid-April 2021, Taiwan did not experience any large COVID-19 outbreaks.

Continuous mutations in the SARS-CoV-2 viral genomes have evolved different lineages with higher transmissibility and increased host fitness. Among the variants of concern (VOCs), the Alpha variant (B.1.1.7 lineage) with the highest relative fitness (Obermeyer et al., 2022) has exacerbated pandemic concerns since its initial detection in the UK in September 2020 (Davies et al., 2021). In December 2020, the Alpha variant was imported into Taiwan for the first time. After a few controllable waves, the re-

doi:10.6342/NTU202301433

introduction of the Alpha variant began to cause new COVID-19 cases at an unprecedented rate in late April, driving 14,311 total indigenous cases. However, within 100 days of implementing Level 3 Restrictions, Taiwan reached zero indigenous cases on August 22, 2021. Contact-tracing investigations confirmed several cluster cases before the surge in 2021. Three key questions thus arose: Were the different strains of SARS-CoV-2 Alpha variants from various early-outbreak transmission chains associated with igniting the community outbreak? What epidemiological factors facilitated the fast spread or blocking Alpha strains in the community? What lessons have we learned from how Taiwan controlled this outbreak, to help other countries quickly contain fast-spreading variants?

## 5.3 Methodology

### 5.3.1 Study design

We analyzed 16,132 laboratory-confirmed SARS-CoV-2-positive cases from January 11, 2020 to September 4, 2021 in Taiwan, then focused on 14,636 cases (14,311 indigenous cases) from the 2021 outbreak (April 16 – September 4). As the majority of outbreak cases (86.27%, 12,346/14,311) occurred in Taipei, New Taipei, and Taoyuan cities, the spatiotemporal distributions of cases in these cities across four different time periods were plotted using Microsoft Power BI. To search for possible viral sequence differences that launched this outbreak, we combined whole-genome sequences of 101

doi:10.6342/NTU202301433

Taiwan SARS-CoV-2 Alpha variants (Table 1). During the onset dates from December 9, 2020 to August 31, 2021 when the viral sequences were collected, they included 12 imported strains before the outbreak (T0, pre-outbreak) and 12 strains from the beginning of this outbreak (T1a, April 16, 2021 – May 7, 2021; early-outbreak). Each confirmed case containing comprehensive contact-tracing through joint efforts from local departments of health (DOH) and Taiwan CDC. The integrated information was helpful to investigate the early transmission chains that might be associated with subsequent community spread (81 strains, May 7, 2021 – August 31, 2021). The 81 indigenous strains involved three time periods based on public health interventions: T1 (April 16 – May 14; pre-Level 3 Restrictions), T2 (May 15 – June 22; post-Level 3 Restrictions, but pre-Zero-COVID Policy), and T3 (June 23 – August 31; post-Zero-COVID Policy) were analyzed to look for whether a dominant virus strain was persistently spreading in the community. Finally, we applied univariate and multivariable analyses to search for factors attributed to the appearance of the dominant virus strains (Figure 1).

### 5.3.2 Study populations of SARS-CoV-2-positive cases in Taiwan

All the laboratory-confirmed SARS-CoV-2-positive cases in 2021 were tested using real-time RT-PCR on patients suspected of or exhibiting COVID-19 clinical symptoms. We plotted an overall epidemic curve of total imported and indigenous SARS-CoV-2-

doi:10.6342/NTU202301433

positive cases from January 1, 2020 to September 4, 2021. According to information released from local DOH and confirmed by Taiwan CDC, we categorized infection sources for indigenous cases into five major risk groups (Yen et al., 2021) (imported-aircraft-associated, healthcare-associated, community-associated, ship-associated, and unidentified sources).

### 5.3.3 SARS-CoV-2 genome sequence alignment and mutation analyses

The 308 whole-genome sequences of SARS-CoV-2 in Taiwan (January 11, 2020 – August 31, 2021) were retrieved from NCBI-Virus and GISAID-EpiCoV databases. We used our in-house developed analytical tools, CoVConvert and IniCoV, to process and analyze these SARS-CoV-2 sequences (Yang et al., 2020). CoVConvert rearranged the sequences of the 101 Taiwan Alpha variants to ensure data quality, then aligned and translated them into three polypeptides from three reading frames. Next, IniCoV automatically divided the translated polypeptides into 31 proteins for each viral strain, combined them with individual epidemiological information, and subsequently compared these 101 strains with the Alpha variants' reference strain (UK-MILK-ACF9CC, referred to as “UK-Alpha-ref-strain”) to analyze any residue differences among these strains involving three groups: (1) the 12 imported strains before the outbreak (T0), (2) the initial 12 strains from early-outbreak (T1a), and (3) the remaining 77 strains (T1b, T2, T3).

doi:10.6342/NTU202301433

### 5.3.4 CoVConvert: a tool to process Coronavirus sequences

CoVConvert (Coronavirus viral sequences converter for genome organization) performed virus strains' names, checked the quality of downloaded sequences and achieved multiple alignments based on the Wuhan-Hu-1 reference nucleotide sequence of SARS-CoV-2 (NC045512.2). Next, data entries with erroneous or incorrect sequences that failed to align were excluded. Last, all qualified and well-aligned DNA sequences were translated into three possible polypeptides from 0, + 1, and +2 reading frames to determine one complete full-length viral peptide using CoVConvert.

### 5.3.5 IniCoV: A Coronavirus information viewer and analyzer

IniCoV (Coronavirus viral information viewer and analyzer for finding out initial source), a program composed of various modules to automatically analyze viral sequencing data in combination with epidemiological information (e.g., viral type, host, region, strain, year, and viral variants or lineages) involving the following two modules: The CoVCS (Coronavirus Cross-Segment alignment) module was used to align amino acid sequences based on SARS-CoV-2 nomenclature and subsequently divide the translated polypeptides into 31 proteins. CoVCS-processed viral genetic information can easily be used for determining the sequence and genome organization based on a particular residue. In the first step, CoVCS generate the genome organization worksheet template based on the Wuhan-1 reference sequence and define the 31 proteins' residue

doi:10.6342/NTU202301433

position. Secondly, CoVCS select one or two CoVConvert generated amino acid sequences to scan the short-conserved sequences (CS) (4~6 aa) from the region of the transcription-regulatory sequence (Kim et al., 2020). Each protein of 31 proteins has unique CS at the front of the N-terminal, and behind the C-terminal, CoVCS organize these amino acid sequences in specific position based on worksheet template. Finally, CoVCS rearrange these organized sequences encoded by strain name and provides the flexibility to group these sequences by name or residues. The CoVCG (Coronavirus Comparative Grouping) module was designed to automatically deduce amino acid sequences from the collected SARS-CoV-2 strains grouped by the question of interest. In short, CoVCG first generated consensus sequences from each subgroup and determined the most representative (i.e., most frequent) amino acid at each position through computing. Unique amino acid residues differentially presented between different subgroups in the whole genome of SARS-CoV-2 computed by CoVCG were re-examined, verified based on the CoVCG-generated substitution table, and visualized. To visualize the population of sequences, we wrote the web-based tool based on d3.js (Data-Driven Documents) JavaScript program language library (Bostock et al., 2011) and presented all substitutions of amino acids at each position as similarly to weblogo. The largest letters presented dominant and small letters presented minor sequences shown in one plot that

doi:10.6342/NTU202301433

can easily summarize sequences population.

### 5.3.6 Contact-tracing investigations and transmissibility analysis of the early-outbreak cases in the Taiwan's 2021 outbreak

To measure viral transmissibility, we applied the epidemiological contact-tracing investigations to compare the effective reproductive numbers over time (Rt) of those early Alpha variant cases. The range and mean ± standard deviation (SD) values of Rt were calculated for three groups: cases before the outbreak, airport-associated cases (pilots, hotel staff) in the early-outbreak, and community-associated cases. Significant differences among the three groups were tested using one-way ANOVA.

### 5.3.7 Univariate and multivariable regression analyses of factors associated with SARS-CoV-2 strains' dominance in the outbreak

To understand which significant factors were associated with the dominant SARS-CoV-2 strains, we used four R packages to examine the 81 Taiwan indigenous strains for univariate analysis. The nine factors included: (1) epidemic periods, (2) epicenter, (3) vaccination coverage, (4) public transport ridership, (5) numbers of daily cases, (6) population size, (7) population density, (8) age, and (9) gender. Factors 3-8 were separated into “high” and “low" groups based on the median. We used Fisher's exact test to assess all factors between subgroups and obtained the crude odds ratios (cOR) with 95% confidence intervals (CIs). All statistically significant factors ($p < 0.05$) were checked with correlations by calculating variance inflation factors (VIF) before running the multivariable regression. The best-fitting model was selected from the candidate models generated from the stepwise (backward/forward) search method by choosing the lowest Akaike information criterion (AIC) value. We also reported the adjusted ORs (aOR) with 95% CIs and p-values from the best-fitting model to present the factors associated with the dominant indigenous Alpha strains.

## 5.4 Results

### 5.4.1 Characteristics in SARS-CoV-2-positive cases before and after Taiwan's 2021 outbreak

In Taiwan, from 1st week of 2020 through the 36th week of 2021, a total of 16,132 cases of SARS-CoV-2 laboratory-confirmed cases were documented, of which 1,486 were imported cases and 14,646 were indigenous cases. From the 1,486 imported cases, originating from Wuhan strains led to an increase in cases during the 4th to 6th weeks of 2020, with an average of three to eight cases per week. The highest number of 125 cases was documented in the 12th week in mid-March 2020 when Taiwanese students returned from Europe and the USA. However, with the government implementation of strict border controls on March 19, 2020, the number of cases declined rapidly. Thereafter, there were two modest peaks in imported cases of the SARS-CoV-2 Alpha variant, first during the post-holiday period from the 48th week of 2020 to the 1st week of 2021, and then again

doi:10.6342/NTU202301433

during the spring break period from the 16th to the 19th week of 2021 (Figure 2A). For indigenous cases, Two minor outbreaks during the 4th to 15th weeks of 2020, as well as the 51st week of 2020 to the 7th week of 2021. These were primarily attributed to the Wuhan strains (114 cases) and Epsilon variants (19 cases), respectively. No further instances of indigenous cases emerged from the 8th to 16th weeks of 2021. Up until April 16, 2021 (the 17th week), the initial cases were confined to airport staff such as pilots and hotel staff. Subsequently, a massive outbreak took place, reaching a peak of 3,363 cases in mid-May of 2021 (the 20th week) (Figure 2A). The origin of the 114 prior-indigenous cases prior to the outbreak in 2021 (spanning from January 22, 2020 to April 15, 2021) was diverse and encompassed various risk groups, with community and unidentified sources accounting for 16.7% (19/114) of the total. Conversely, during the outbreak from April 16, 2021 to September 4, 2021, community and unidentified sources were found to account for a significant increase of 98.3% (14,067/14,311) of all cases, as demonstrated by a statistical significance of $p < 0.0001$ (Figure 2B).

### 5.4.2 Characterization of Taiwan's 2021 outbreak

Before the outbreak (T0 period), the 12 imported cases were reported but no indigenous cases (Figure 3A). However, starting on April 16, the emergence of sporadic clusters of SARS-CoV-2-positive cases associated with the airport and quarantine hotel

doi:10.6342/NTU202301433

triggered a widespread outbreak (Figure 3B, T1a). The weekly mean number and monthly incidence rate of SARS-CoV-2-positive cases increased rapidly in Taipei, New Taipei, and Taoyuan cities. The mean incidence rate (per 100,000 population) was 10.91 ± 19.6 in T1b and peaked at 98.6 ± 120.31 in T2. With the largest population, Taipei City experienced the highest case incidence between May 7 and May 14. The daily total of cases in these three cities peaked at 495 on May 15, prompting the cities of Taipei and New Taipei to implement Level 3 restrictions. By early June, daily case counts in both cities had dropped below 100 cases. Taipei City adopted an enhanced zero-COVID policy on June 23, reducing daily cases to just ten by July 10 (Figure 4). The mean incidence rate dropped to 12.33 ± 13.19 (Figure 3B, T3). It took 100 days from the peak on May 15 to reach zero indigenous cases on August 22 in all cities, without lockdowns. Spatiotemporal analysis of diffusion patterns over time revealed that the Wanhua District in Taipei City had the highest incidence rates throughout the entire outbreak. Its six neighboring districts, which held the second and third highest incidence rates from April 16 to June 22, were regarded as the epicenter of the outbreak (Table 2). Subsequently, the virus rapidly spread from the epicenter to other districts with more substantial populations and higher population densities (Figure 3B, T2 and T3).

doi:10.6342/NTU202301433

### 5.4.3 Contact-tracing investigations to search for dynamic sequence changes

Given the highly genome divergence of the Alpha variants and the rapid community spread of the virus during the 2021 Taiwan outbreak, it is imperative to understand the potential transmission routes that led to such a widespread outbreak in a matter of weeks. Contact-tracing investigations identified six transmission chains before the occurrence of notable cluster cases initiated by ID-1363 that triggered the community outbreak from April 16 to May 7 (Figure 5 and Table 3).

### 5.4.4 Integrating whole-genome sequence analyses

Whole-genome sequencing combined with contact tracing information could determine which of the early transmission chains might have contributed to the subsequent spread of the virus during the outbreak. This study examined whole-genome sequences from twelve viral strains imported during the pre-outbreak period (T0) and twelve strains isolated during the early periods of the outbreak (T1a) and found genome divergence. The result indicated that the sequences of the Alpha strains isolated in T0 were highly varied when compared to the UK-Alpha-ref-strain, and were also dissimilar from those imported during T1a. All of the early viral strains from T1a were found to contain mutations in PLpro (C5144T and C5812T), nsp8 (C12253T), RdRp (C15895T), Helicase (G17615A), and ORF8 (C28957T) compared to the UK-Alpha-ref-strain. However, those isolated from the early-outbreak transmission chains 1 (ID-1091), 2 (ID-

doi:10.6342/NTU202301433

1145), 4 (ID-1078 and ID-1079), 5 (ID-1154, ID-1183, ID-1187), and 6 (ID-1102, ID-1137) each possessed additional nucleotide variations present throughout the genome (Figure 6). Notably, the sequences of ID-1186 isolated from chain 2 were identical to those of ID-3445 and ID-1263. Contact tracing investigations indicated that ID-3445 was a co-worker of the index case, ID-1363, at the teahouse, leading to the conclude that ID-3445 and ID-1263 represented the earliest strains of community transmission in Wanhua District. (Figure 6). Despite ID-3445 and ID-1263 sharing overlapping locations (Wanhua District) in their visiting history, no epidemiological linkage was found between ID-1186 and ID-3445 or ID-1263. Additionally, 60% (3/5) of the indigenous strains isolated during T1b and 28.57% (16/56) of those isolated during T2 were found to be identical to the ID-3445/1263/1186 strain. None of the remaining 42 strains (5+56-3-16 = 42) were identical to any other strains isolated from the early-outbreak clusters (Tables 4 and 5). These findings indicate that even though several transmissions occurred in the early periods of the outbreak, only the strain associated with ID-1186 was identical with those of ID-3445 and ID-1263, which were linked to the 2021 community outbreaks (Figure 7). Interestingly, no strains analogous to the ID-3445/1263/1186 strain were detected after the implementation of the enhanced Zero-COVID Policy (Table 4, T3 period). Furthermore, the whole-genome sequences of all 14 indigenous viral strains isolated

doi:10.6342/NTU202301433

during T3 were disparate, and no new dominant strain emerged. Our analysis did not find a second dominant strain throughout the outbreak (Figure 8).

### 5.4.5 Epidemiological factors associated with viral strain dominance in the 2021 outbreak

The univariate analysis aimed to understand the factors that correlated with the prominence of the ID-3445/1263/1186 strains. Five significant correlations were determined: (1) the epidemic period [1.744 (0.369-7.924), p = 0.0097], (2) the epicenter [0.208 (0.063-0.638), p = 0.0024], (3) vaccination coverage [0.336 (0.1-1.023), p = 0.0479], (4) population size [0.219 (0.057-0.789), p = 0.011], and (5) population density [0.273 (0.086-0.831), p = 0.018]. Of these significant factors, the epidemic period exhibited the highest value of crude OR, indicating that the ID-3445 strain was already prevalent during T1. Our multivariable analysis demonstrated that the epidemic period and the epicenter were the two factors significantly linked with the dominance of the ID-3445/1263/1186 strains during the 2021 outbreak [adjust OR (95% CI), p = 0.007; 0.145 (0.044-0.474), p = 0.001, respectively]. These results suggest that the dominant strain was selected in the epicenter during the early period of the outbreak (T1 period) (Table 7).

doi:10.6342/NTU202301433

### 5.5 Discussion

The fast-mutating and increasingly transmissible SARS-CoV-2 has created unprecedented public health challenges. However, Taiwan successfully halted local SARS-CoV-2 transmission through its rapid response combining strict border control, firm adherence to using facemasks and hand hygiene, and a bundle strategy to minimize nosocomial infection (Yen et al., 2021). Alpha variants, which dominated in Europe and the USA in early 2021 (Liu et al., 2022; Liu et al., 2021; Tai et al., 2022), finally sparked a large outbreak in Taiwan in mid-May 2021 (Akhmetzhanov et al., 2022). This study integrated analyses of whole-genome viral sequences with contact-tracing, spatio-temporal analyses, individual-based effective reproductive numbers, and public health policies, to deliver four major findings (Figure 8). First, the Alpha variants introduced to Taiwan were highly diverse. Second, we identified an epicenter Wanhua District in Taipei City, where a convenient transportation hub and many leisure activities facilitated human contact and viral transmission, driving cases in dense, highly populated neighboring districts, and igniting Taiwan's large 2021 outbreak. Third, one imported SARS-CoV-2 Alpha variant strain from early-outbreak chains was preferentially selected at the epicenter and became dominant in the early epidemic period. The predominant strain extended to the middle period and remained detectable for at least 1.5 months. This was

doi:10.6342/NTU202301433

the only dominant strain throughout the entire outbreak, but it declined after Level 3 Restrictions were implemented, and disappeared following the Zero-COVID Policy without city lockdowns (Dyer, 2022; Normile, 2022). Fourth, multivariable regression supported the finding that the early epidemic period and epicenter were significantly associated with emergence of the predominant community-spread viruses. These results indicate the importance of viral genomic surveillance alongside epidemics, and its usefulness in evaluating public health policies.

Given genomic surveillance's application in control outbreaks (Chen et al., 2022; Gong et al., 2020; Gu et al., 2022; Wilkinson et al., 2021), we linked whole-genome sequencing in Taiwan with epidemiological attributes and discovered that early transmission chains substantially facilitated the mid-April to early May community surge. Therefore, outbreak-associated viral dominance must consider specific epidemiological characteristics (Sutton et al., 2022), including high population density, transportation hubs, and teahouses in the epicenter where patrons mingled without masks, as preludes to this outbreak.

Investigating relationships between epidemiological factors and the emergence, rise, and decline of dominant strains is essential for containing outbreaks quickly. In fact, Alpha variants that entered Taiwan before the outbreak had high viral genome divergence.

doi:10.6342/NTU202301433

However, after ongoing transmission, virus selection occurred under special epidemiological conditions (Sutton et al., 2022), like airport-associated cases and community-related clusters (Gu et al., 2022). Once the case number sharply rose, indicating the selection-advantageous dominant virus strain was continuously spreading, viral diversity plummeted. As with other VOCs (Obermeyer et al., 2022), it took time, 2-3 weeks, for community-derived strain to emerge, which became dominant strain with more homogeneous genome. Control policies can shape trends in the virus population during this crucial time window. Our data showed 39 days after the Level 3 Restrictions implementation and 61 days following the Zero-COVID Policy's rollout (Table 4), the dominant community-spread viruses were successfully eliminated without lockdowns (Akhmetzhanov et al., 2022). No new dominant strain appeared throughout the entire outbreak. Therefore, dominant strains with selection advantages must be eliminated quickly before epidemics expand.

SARS-CoV-2 has continuously evolved worldwide. When the Alpha variant overtook the Wuhan strain (Obermeyer et al., 2022), it indicated the need to find factors associated with viral dominance. Our multivariable analysis again demonstrated that turning points in the early epidemic period and epicenter supported the emergence of dominant community-spread viruses. This conclusion aligns with our findings on an

doi:10.6342/NTU202301433

adaptive mutant in the H1N1pdm09 virus carrying HA2-E374K, which was imported to Taiwan and extended viral survival in a densely populated Taipei City before vaccination rollouts (Kao et al., 2012). Our dengue research discovered that clustering dengue cases with higher transmission intensity helped select a virus strain that caused more severe dengue hemorrhagic fever cases in southern Taiwan, where *Aedes aegypti* mosquitoes are assumed to play important roles in viral selection (Bennett et al., 2003; Wen et al., 2010). These specific epidemiological conditions, including human clustering cases, eating/dining without wearing masks, frequent human-to-human contact in entertainment settings (e.g., teahouses), and the combination of low vaccination coverage and/or SARS-CoV-2 infection helped Alpha strains with a selective advantage through natural selection (prior to immune selection) become dominant and drive a rapid surge in cases. As these mutants continue evolving, their residues for viral replication, transmissibility, immune antagonism (Cheng et al., 2021; Jian et al., 2021; Pan et al., 2021; Verghese et al., 2021), and their epidemic or pandemic potential merit monitoring (Subissi et al., 2022).

This study has four major limitations. First, most cases were reported from passive surveillance. Second, we obtained viral sequences retrospectively from databases without random sampling on epidemiological attributes. Many strains lacked full-length sequences or complete epidemiological information, resulting in a small sample size and

doi:10.6342/NTU202301433

potential selection bias. As we did not have multiple samples from each patient, our results may not fully reflect reality (Li et al., 2022; Tonkin-Hill et al., 2021). The reproductive numbers of each early transmission chain may be underestimated due to asymptomatic/mild infections. Hence, how early transmission chains and viral selection mechanisms (e.g., for increasing viral infectivity or replication) of dominant strains contributed to community clusters remains unclear. Third, individual-based pre-existing comorbidities, vaccination history, past infection, compliance with preventative behavior (Yen et al., 2021), and other potential influencers of viral dynamics were not collected to protect personal privacy. Fourth, although all 81 indigenous viruses in this outbreak carried Spike-M1237I and Helicase-R460K (Table 6), we still do not know how or whether these mutations might increase viral transmissibility and epidemic severity. However, compiling epidemiological linkages within the same transmission cluster and viral sequences, can offer a better picture of early transmission chains. In conclusion, Alpha strains in Taiwan started from imported cases with genomic diversity. A dominant strain emerged under conditions involving human gatherings leading to case clusters from the airport to the quarantine hotel, transportation hubs, and teahouses in the epicenter. Four prerequisites for dominant strains that possibly emerged in the community include: (1) high frequency of human-to-human contact at hotels without early detection of

doi:10.6342/NTU202301433

positive cases, or low compliance with home quarantine, facilitating viral selection without notice, (2) close contacts without adequate protection at teahouses (e.g., removing masks while dining/drinking/chatting), which may have helped viruses gain selection advantages to increase transmissibility (higher Rt values), (3) highly mobile individuals carrying virus from the epicenter outward, and (4) lack of effective population-based control policies against continuous transmission, like the initial absence of rapid community screening for SARS-CoV-2-positive cases, low vaccination coverage (1.3% and 0.7% for the 1st dose of the COVID-19 vaccine in Taipei City and New Taipei City as of May 15, 2021, respectively). Importantly, rigorous individual and population-level prevention policies on May 15, successfully eliminated the spread of the dominant strains. No new viral lineage composition occurred during the 100 days of the 2021 Taiwan outbreak. Future research on VOCs should focus on an integrated approach to timely monitoring of whole-genomic and amino acid changes of novel variants with growing transmissibility, pathogenicity, and fatality, as well as spatio-temporal data analysis to detect dominant strains early on. Our results demonstrate that predominant virus strains with increasing epidemic/pandemic potential at both the micro- and macro-levels are naturally selected by epidemiological conditions even before mass-vaccination (Ko et al., 2018). Moreover, our software and integrated analyses can be applied to timely

doi:10.6342/NTU202301433

monitoring of trends in full-length viral dynamics, searching for dominant strains of any emerging pathogens across entire epidemic, and obtaining the viruses with striking increases in case numbers in the epicenter, as well as evaluating the effectiveness of public health policies. Even after mass vaccination and anti-viral drug development, international collaboration will be imperative to preventing future pandemics.

doi:10.6342/NTU202301433

## 5.6 References

Akhmetzhanov, A. R., Cheng, H. Y., Linton, N. M., Ponce, L., Jian, S. W., & Lin, H. H. (2022). Transmission Dynamics and Effectiveness of Control Measures during COVID-19 Surge, Taiwan, April-August 2021. Emerg Infect Dis, 28(10), 2051-2059. https://doi.org/10.3201/eid2810.220456

Bennett, S. N., Holmes, E. C., Chirivella, M., Rodriguez, D. M., Beltran, M., Vorndam, V., Gubler, D. J., & McMillan, W. O. (2003). Selection-driven evolution of emergent dengue virus. Mol Biol Evol, 20(10), 1650-1658. https://doi.org/10.1093/molbev/msg182

Chen, Z., Azman, A. S., Chen, X., Zou, J., Tian, Y., Sun, R., Xu, X., Wu, Y., Lu, W., Ge, S., Zhao, Z., Yang, J., Leung, D. T., Domman, D. B., & Yu, H. (2022). Global landscape of SARS-CoV-2 genomic surveillance and data sharing. Nat Genet, 54(4), 499-507. https://doi.org/10.1038/s41588-022-01033-y

Cheng, Y. W., Chao, T. L., Li, C. L., Wang, S. H., Kao, H. C., Tsai, Y. M., Wang, H. Y., Hsieh, C. L., Lin, Y. Y., Chen, P. J., Chang, S. Y., & Yeh, S. H. (2021). D614G Substitution of SARS-CoV-2 Spike Protein Increases Syncytium Formation and Virus Titer via Enhanced Furin-Mediated Spike Cleavage. mBio, 12(4), e0058721. https://doi.org/10.1128/mBio.00587-21

Davies, N. G., Abbott, S., Barnard, R. C., Jarvis, C. I., Kucharski, A. J., Munday, J. D., Pearson, C. A. B., Russell, T. W., Tully, D. C., Washburne, A. D., Wenseleers, T., Gimma, A., Waites, W., Wong, K. L. M., van Zandvoort, K., Silverman, J. D., Diaz-Ordaz, K., Keogh, R., Eggo, R. M., et al. (2021). Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. Science, 372(6538). https://doi.org/10.1126/science.abg3055

Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis, 20(5), 533-534. https://doi.org/10.1016/s1473-3099(20)30120-1

Dyer, O. (2022). Covid-19: Lockdowns spread in China as omicron tests "zero covid" strategy. BMJ, 376, o859. https://doi.org/10.1136/bmj.o859

Gong, Y. N., Tsao, K. C., Hsiao, M. J., Huang, C. G., Huang, P. N., Huang, P. W., Lee, K. M., Liu, Y. C., Yang, S. L., Kuo, R. L., Chen, K. F., Liu, Y. C., Huang, S. Y., Huang, H. I., Liu, M. T., Yang, J. R., Chiu, C. H., Yang, C. T., Chen, G. W., & Shih, S. R. (2020). SARS-CoV-2 genomic surveillance in Taiwan revealed novel ORF8-deletion mutant and clade possibly associated with infections in Middle East. Emerg Microbes Infect, 9(1), 1457-1466. https://doi.org/10.1080/22221751.2020.1782271

Gu, H., Cheng, S. S. M., Krishnan, P., Ng, D. Y. M., Chang, L. D. J., Liu, G. Y. Z., Cheuk, S. S. Y., Hui, M. M. Y., Fan, M. C. Y., Wan, J. H. L., Lau, L. H. K., Chu, D. K. W., Dhanasekaran, V., Peiris, M., & Poon, L. L. M. (2022). Monitoring International Travelers Arriving in Hong Kong for Genomic Surveillance of SARS-CoV-2. Emerg Infect Dis, 28(1), 247-250. https://doi.org/10.3201/eid2801.211804

doi:10.6342/NTU202301433

Hsueh, P. R., & Yang, P. C. (2005). Severe acute respiratory syndrome epidemic in Taiwan, 2003. J Microbiol Immunol Infect, 38(2), 82-88.

Jian, M. J., Chung, H. Y., Chang, C. K., Lin, J. C., Yeh, K. M., Chen, C. W., Li, S. Y., Hsieh, S. S., Liu, M. T., Yang, J. R., Tang, S. H., Perng, C. L., Chang, F. Y., & Shang, H. S. (2021). Clinical Comparison of Three Sample-to-Answer Systems for Detecting SARS-CoV-2 in B.1.1.7 Lineage Emergence. Infect Drug Resist, 14, 3255-3261. https://doi.org/10.2147/idr.S328327

Kao, C. L., Chan, T. C., Tsai, C. H., Chu, K. Y., Chuang, S. F., Lee, C. C., Li, Z. R., Wu, K. W., Chang, L. Y., Shen, Y. H., Huang, L. M., Lee, P. I., Yang, C., Compans, R., Rouse, B. T., & King, C. C. (2012). Emerged HA and NA mutants of the pandemic influenza H1N1 viruses with increasing epidemiological significance in Taipei and Kaohsiung, Taiwan, 2009-10. PLoS One, 7(2), e31162. https://doi.org/10.1371/journal.pone.0031162

Ko, H. Y., Li, Y. T., Chao, D. Y., Chang, Y. C., Li, Z. T., Wang, M., Kao, C. L., Wen, T. H., Shu, P. Y., Chang, G. J., & King, C. C. (2018). Inter- and intra-host sequence diversity reveal the emergence of viral variants during an overwintering epidemic caused by dengue virus serotype 2 in southern Taiwan. PLoS Negl Trop Dis, 12(10), e0006827. https://doi.org/10.1371/journal.pntd.0006827

Lamarca, A. P., de Almeida, L. G. P., Francisco, R. D. S., Jr., Lima, L. F. A., Scortecci, K. C., Perez, V. P., Brustolini, O. J., Sousa, E. S. S., Secco, D. A., Santos, A. M. G., Albuquerque, G. R., Mariano, A. P. M., Maciel, B. M., Gerber, A. L., Guimarães, A. P. C., Nascimento, P. R., Neto, F. P. F., Gadelha, S. R., Porto, L. C., . . . Vasconcelos, A. T. R. (2021). Genomic surveillance of SARS-CoV-2 tracks early interstate transmission of P.1 lineage and diversification within P.2 clade in Brazil. PLoS Negl Trop Dis, 15(10), e0009835. https://doi.org/10.1371/journal.pntd.0009835

Li, J., Du, P., Yang, L., Zhang, J., Song, C., Chen, D., Song, Y., Ding, N., Hua, M., Han, K., Song, R., Xie, W., Chen, Z., Wang, X., Liu, J., Xu, Y., Gao, G., Wang, Q., Pu, L., . . . Chen, C. (2022). Two-step fitness selection for intra-host variations in SARS-CoV-2. Cell Rep, 38(2), 110205. https://doi.org/10.1016/j.celrep.2021.110205

Liu, L. T., Tsai, J. J., Chang, K., Chen, C. H., Lin, P. C., Tsai, C. Y., Tsai, Y. Y., Hsu, M. C., Chuang, W. L., Chang, J. M., Hwang, S. J., & Chong, I. W. (2022). Identification and Analysis of SARS-CoV-2 Alpha Variants in the Largest Taiwan COVID-19 Outbreak in 2021. Front Med (Lausanne), 9, 869818. https://doi.org/10.3389/fmed.2022.869818

Liu, M. T., Mu, J. J., Lin, Y. C., & Yang, J. R. (2021). Taiwan CDC MOHW110-CDC-C-315-124116 Report: Pathogen Analysis of Pneumonic Encephalitis and Unexplained Infectious Diseases [In Chinese].

Normile, D. (2022). China refuses to end harsh lockdowns. Science, 376(6591), 333-334. https://doi.org/10.1126/science.abq6109

Obermeyer, F., Jankowiak, M., Barkas, N., Schaffner, S. F., Pyle, J. D., Yurkovetskiy, L.,

doi:10.6342/NTU202301433

Bosso, M., Park, D. J., Babadi, M., MacInnis, B. L., Luban, J., Sabeti, P. C., & Lemieux, J. E. (2022). Analysis of 6.4 million SARS-CoV-2 genomes identifies mutations associated with fitness. Science, 376(6599), 1327-1332. https://doi.org/doi:10.1126/science.abm1208

Pan, Y. H., Liao, M. Y., Chien, Y. W., Ho, T. S., Ko, H. Y., Yang, C. R., Chang, S. F., Yu, C. Y., Lin, S. Y., Shih, P. W., Shu, P. Y., Chao, D. Y., Pan, C. Y., Chen, H. M., Perng, G. C., Ku, C. C., & King, C. C. (2021). Use of seroprevalence to guide dengue vaccination plans for older adults in a dengue non-endemic country. PLoS Negl Trop Dis, 15(4), e0009312. https://doi.org/10.1371/journal.pntd.0009312

Robishaw, J. D., Alter, S. M., Solano, J. J., Shih, R. D., DeMets, D. L., Maki, D. G., & Hennekens, C. H. (2021). Genomic surveillance to combat COVID-19: challenges and opportunities. Lancet Microbe, 2(9), e481-e484. https://doi.org/10.1016/s2666-5247(21)00121-x

Subissi, L., von Gottberg, A., Thukral, L., Worp, N., Oude Munnink, B. B., Rathore, S., Abu-Raddad, L. J., Aguilera, X., Alm, E., Archer, B. N., Attar Cohen, H., Barakat, A., Barclay, W. S., Bhiman, J. N., Caly, L., Chand, M., Chen, M., Cullinane, A., de Oliveira, T., . . . Agrawal, A. (2022). An early warning system for emerging SARS-CoV-2 variants. Nat Med, 28(6), 1110-1115. https://doi.org/10.1038/s41591-022-01836-w

Sutton, M., Radniecki, T. S., Kaya, D., Alegre, D., Geniza, M., Girard, A. M., Carter, K., Dasenko, M., Sanders, J. L., Cieslak, P. R., Kelly, C., & Tyler, B. M. (2022). Detection of SARS-CoV-2 B.1.351 (Beta) Variant through Wastewater Surveillance before Case Detection in a Community, Oregon, USA. Emerg Infect Dis, 28(6), 1101-1109. https://doi.org/10.3201/eid2806.211821

Tai, J.-H., Low, Y. K., Lin, H.-F., Wang, T.-Y., Lin, Y.-Y., Foster, C., Lai, Y.-Y., Yeh, S.-H., Chang, S.-Y., Chen, P.-J., & Wang, H.-Y. (2022). Spatial and temporal origin of the third SARS-COV-2 Outbreak in Taiwan. bioRxiv, 2022.2007.2004.498645. https://doi.org/10.1101/2022.07.04.498645

Tonkin-Hill, G., Martincorena, I., Amato, R., Lawson, A. R. J., Gerstung, M., Johnston, I., Jackson, D. K., Park, N., Lensing, S. V., Quail, M. A., Gonçalves, S., Ariani, C., Spencer Chapman, M., Hamilton, W. L., Meredith, L. W., Hall, G., Jahun, A. S., Chaudhry, Y., Hosmillo, M., . . . Kwiatkowski, D. (2021). Patterns of within-host genetic diversity in SARS-CoV-2. Elife, 10. https://doi.org/10.7554/eLife.66857

Verghese, M., Jiang, B., Iwai, N., Mar, M., Sahoo, M. K., Yamamoto, F., Mfuh, K. O., Miller, J., Wang, H., Zehnder, J., & Pinsky, B. A. (2021). A SARS-CoV-2 Variant with L452R and E484Q Neutralization Resistance Mutations. J Clin Microbiol, 59(7), e0074121. https://doi.org/10.1128/jcm.00741-21

Wen, T. H., Lin, N. H., Chao, D. Y., Hwang, K. P., Kan, C. C., Lin, K. C., Wu, J. T., Huang, S. Y., Fan, I. C., & King, C. C. (2010). Spatial-temporal patterns of dengue in areas at risk of dengue hemorrhagic fever in Kaohsiung, Taiwan, 2002. Int J Infect Dis, 14(4), e334-343. https://doi.org/10.1016/j.ijid.2009.06.006

Wilkinson, E., Giovanetti, M., Tegally, H., San, J. E., Lessells, R., Cuadros, D., Martin, D. P., Rasmussen, D. A., Zekri, A. N., Sangare, A. K., Ouedraogo, A. S., Sesay, A. K., Priscilla, A., Kemi, A. S., Olubusuyi, A. M., Oluwapelumi, A. O. O., Hammami, A., Amuri, A. A., Sayed, A., . . . de Oliveira, T. (2021). A year of genomic surveillance reveals how the SARS-CoV-2 pandemic unfolded in Africa. Science, 374(6566), 423-431. https://doi.org/10.1126/science.abj4336

Yang, C. R., King, C. C., Liu, L. D., & Ku, C. C. (2020). FluConvert and IniFlu: a suite of integrated software to identify novel signatures of emerging influenza viruses with increasing risk. BMC Bioinformatics, 21(1), 316. https://doi.org/10.1186/s12859-020-03650-y

Yen, M. Y., Yen, Y. F., Chen, S. Y., Lee, T. I., Huang, K. H., Chan, T. C., Tung, T. H., Hsu, L. Y., Chiu, T. Y., Hsueh, P. R., & King, C. C. (2021). Learning from the past: Taiwan's responses to COVID-19 versus SARS. Int J Infect Dis, 110, 469-478. https://doi.org/10.1016/j.ijid.2021.06.002

doi:10.6342/NTU202301433

# Tables

**Table 1. List of 101 SARS-CoV-2 genome sequences and important epidemiological information used in this study in Taiwan**

| Case ID | Strain Name | Onset date | Travel history | Identical to ID-3445 | GISAID Accession (EPI_ISL) |
|---|---|---|---|---|---|
| **Imported cases (N = 20)** | | | | | |
| 783 | cgmh-cgu-44 | 2020/12/9 | PHL | - | 956325 |
| 799 | ntu52 | 2020/12/26 | GBR | - | 1041958 |
| 792 | 792 | 2020/12/27 | GBR | - | 1381386 |
| 804 | ntu49 | 2020/12/28 | GBR | - | 1010728 |
| 837 | ntu54 | 2020/12/29 | GBR | - | 1039160 |
| 958 | cgmh-cgu-58 | 2021/2/26 | USA | - | 2249597 |
| 1048 | cgmh-cgu-61 | 2021/3/23 | PHL | - | 2250151 |
| 1065 | ntu62 | 2021/3/28 | PHL | - | 1667475 |
| 1050 | ntu61 | 2021/3/29 | EGY | - | 1667474 |
| 1047 | cgmh-cgu-60 | 2021/4/2 | IDN | - | 2249836 |
| 1081 | cgmh-cgu-63 | 2021/4/3 | IDN | - | 2250184 |
| 1059 | kmuh-3 | 2021/4/9 | JPN | - | 5395633 |
| 1091 | ntu63 | 2021/4/16 | USA | - | 13566006 |
| 1079 | 1079 | 2021/4/17 | USA | - | 2455264 |
| 1078 | 1078 | 2021/4/18 | USA | - | 2455327 |
| 1102 | ntu64 | 2021/4/24 | USA | - | - |
| 1144 | ntu65 | 2021/4/28 | UZB | - | - |
| 1154 | ntu67 | 2021/5/2 | USA | - | 13618360 |
| 1183 | tsgh-43 | 2021/5/6 | USA | - | 2693006 |
| 2018 | cgmh-cgu-64 | 2021/5/14 | HTI | - | 2544700 |
| **T1 Epicenter (N = 9)** | | | | | |
| 1145 | tsgh-42 | 2021/4/28 | NWT | - | 2693005 |
| 1137 | tsgh-44 | 2021/4/30 | TPE | - | 4096803 |
| 3445 | 3445 | 2021/5/5 | TPE | Yes | 2455329 |
| 1187 | ntu66 | 2021/5/6 | TPE | No | 13618344 |
| 1263 | ntu68 | 2021/5/7 | TPE | Yes | 13578728 |
| 1266 | ntu69 | 2021/5/9 | NWT | No | 13578729 |
| 1265 | ntu70 | 2021/5/9 | NWT | No | 13578730 |
| 1290 | ntu71 | 2021/5/10 | TPE | No | 13578731 |
| 2262 | 2262 | 2021/5/14 | TPE | Yes | 2455330 |
| 1145 | tsgh-42 | 2021/4/28 | NWT | - | 2693005 |
| **T1 Other cities (N = 2)** | | | | | |
| 1186 | cgmh-cgu-73 | 2021/5/7 | TAO | Yes | 2544709 |
| 2150 | kmuh-4 | 2021/5/9 | KHH | Yes | 7016374 |
| **T2 epicenter (N = 30)** | | | | | |
| 1419 | ntu72 | 2021/5/15 | TPE | No | 13578732 |
| 1373 | ntu73 | 2021/5/15 | TPE | Yes | 13578733 |
| 1354 | ntu74 | 2021/5/15 | TPE | No | 13618345 |
| 1359 | ntu75 | 2021/5/15 | TPE | No | 13578734 |
| 1356 | ntu76 | 2021/5/15 | TPE | No | 13578345 |
| 1357 | ntu77 | 2021/5/15 | TPE | Yes | 13618347 |
| 1355 | ntu78 | 2021/5/15 | TPE | No | 13578735 |
| 1360 | ntu79 | 2021/5/15 | TPE | No | 13578736 |
| 1358 | ntu80 | 2021/5/15 | TPE | No | 13578737 |
| 5703 | 5703 | 2021/5/21 | TPE | No | 3000790 |

**Table 1. List of 101 SARS-CoV-2 genome sequences and important epidemiological information used in this study in Taiwan *(continued)***

| Case ID | Strain Name | Onset date | Travel history | Identical to ID-3445 | GISAID Accession (EPI_ISL) |
|---|---|---|---|---|---|
| **T2 epicenter (N = 30)** | | | | | |
| 7955 | 7955 | 2021/5/26 | TPE | No | 3040151 |
| 9098 | 9098 | 2021/5/29 | NWT | Yes | 3040149 |
| 10747 | 10747 | 2021/6/2 | TPE | No | 3000409 |
| 12049 | ntu94 | 2021/6/11 | TPE | No | 11333413 |
| 10179 | ntu91 | 2021/6/12 | TPE | No | 11333514 |
| 13112 | ntu95 | 2021/6/12 | TPE | No | 11333432 |
| 13375 | 13375 | 2021/6/14 | TPE | No | 3001055 |
| 13435 | 13435 | 2021/6/14 | TPE | No | 3040140 |
| 13564 | 13564 | 2021/6/15 | TPE | No | 3001368 |
| 11612 | ntu81 | 2021/6/16 | TPE | No | 13578738 |
| 13137 | ntu104 | 2021/6/16 | TPE | No | 11333509 |
| 13386 | ntu82 | 2021/6/17 | TPE | No | 13578739 |
| 13103 | ntu83 | 2021/6/17 | TPE | No | 13618348 |
| 13318 | ntu84 | 2021/6/17 | TPE | No | 13578740 |
| 10480 | ntu85 | 2021/6/17 | TPE | No | 13578741 |
| 13387 | ntu88 | 2021/6/18 | TPE | Yes | 11333411 |
| 13850 | ntu107 | 2021/6/19 | TPE | Yes | 11333511 |
| 14035 | ntu98 | 2021/6/20 | TPE | No | 11333516 |
| 14168 | ntu105 | 2021/6/20 | TPE | No | 11333510 |
| 14181 | ntu108 | 2021/6/21 | TPE | No | 11333512 |
| **T2 other cities (N = 26)** | | | | | |
| 3461 | kmuh-5 | 2021/5/16 | KHH | No | 7016459 |
| 4742 | kmuh-6 | 2021/5/16 | KHH | No | 7016494 |
| - | cgmh-cgu-65 | 2021/5/18 | TAO | Yes | 2544701 |
| - | cgmh-cgu-66 | 2021/5/18 | TAO | Yes | 2544702 |
| - | cgmh-cgu-79 | 2021/5/18 | TAO | No | 5160472 |
| - | cgmh-cgu-68 | 2021/5/19 | TAO | Yes | 2544704 |
| - | cgmh-cgu-67 | 2021/5/20 | TAO | No | 2544703 |
| - | cgmh-cgu-70 | 2021/5/20 | TAO | Yes | 2544706 |
| - | cgmh-cgu-76 | 2021/5/20 | TAO | Yes | 2544712 |
| - | cgmh-cgu-69 | 2021/5/21 | TAO | No | 2544705 |
| - | cgmh-cgu-78 | 2021/5/22 | TAO | Yes | 2544714 |
| - | cgmh-cgu-77 | 2021/5/23 | TAO | No | 2544713 |
| - | cgmh-cgu-75 | 2021/5/26 | TAO | No | 2544711 |
| - | cgmh-cgu-72 | 2021/5/27 | TAO | No | 2544708 |
| - | cgmh-cgu-74 | 2021/5/29 | TAO | No | 2544710 |
| 10321 | 10321 | 2021/6/1 | MIA | Yes | 3040148 |
| 11042 | 11042 | 2021/6/3 | TAO | Yes | 3040145 |
| 11103 | 11103 | 2021/6/3 | CYQ | Yes | 3040147 |
| 11102 | 11102 | 2021/6/3 | TNN | No | 3040152 |
| 11310 | 11310 | 2021/6/4 | MIA | Yes | 3040146 |
| 11282 | tsgh-46 | 2021/6/4 | KEE | No | 4096807 |
| 12288 | 12288 | 2021/6/8 | TAO | Yes | 3040144 |
| 12857 | 12857 | 2021/6/10 | KEE | Yes | 3001841 |
| 12699 | 12699 | 2021/6/10 | KEE | No | 3002178 |
| 12828 | 12828 | 2021/6/10 | TAO | No | 3040141 |
| 14222 | 14222 | 2021/6/20 | KEE | Yes | 3040143 |

**Table 1. List of 101 SARS-CoV-2 genome sequences and important epidemiological information used in this study in Taiwan *(continued)***

| Case ID | Strain Name | Onset date | Travel history | Identical to ID-3445 | GISAID Accession (EPI_ISL) |
|---|---|---|---|---|---|
| **T3 epicenter (N = 11)** | | | | | |
| 14422 | ntu101 | 2021/6/23 | TPE | No | 11333507 |
| 14516 | ntu102 | 2021/6/23 | TPE | No | 11333508 |
| 14166 | ntu106 | 2021/6/23 | TPE | No | 11362237 |
| 14518 | ntu116 | 2021/6/27 | TPE | No | 11333513 |
| 14879 | ntu113 | 2021/6/28 | TPE | No | 11362240 |
| 14495 | ntu103 | 2021/7/2 | TPE | No | 11333517 |
| 15062 | ntu111 | 2021/7/5 | TPE | No | 11362238 |
| 15226 | ntu117 | 2021/7/6 | TPE | No | 11362241 |
| 15774 | tsgh-45 | 2021/7/28 | TPE | No | 4096805 |
| 15702 | ntu123 | 2021/7/29 | TPE | No | 11362244 |
| 16121 | ntu124 | 2021/8/31 | TPE | No | 11333351 |
| **T3 other cities (N = 3)** | | | | | |
| 14491 | kmuh-7 | 2021/6/23 | KHH | No | 7016498 |
| 14454 | 14454 | 2021/6/26 | MIA | No | 3040142 |
| - | cgmh-cgu-85 | 2021/7/24 | TAO | No | 5160564 |

EGY: Egypt, GBR: United Kingdom, HTI: Haiti, IDN: Indonesia, JPN: Japan, PHL: Philippines, UZB: Uzbekistan.

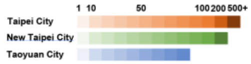
CYQ: Chiayi City, KEE: Keelung City, KHH: Kaohsiung City, MIA: Miaoli County, NWT: New Taipei City, TAO: Taoyuan City, TNN: Tainan City, TPE: Taipei City.

NTU: National Taiwan University, Taiwan CDC: Taiwan Centers for Disease Control, TSGH: Tri-Service General Hospital, CGMH-CGU: Chang Gung Memorial Hospital (University), KMUH: Kaohsiung Medical University Chung-Ho Memorial Hospital.

We used 101 available Taiwan whole-genome sequences of SARS-CoV-2 for analysis. Imported or Indigenous cases were defined through joint epidemiological investigation efforts from local Health Bureaus and Taiwan CDC. A case that had travel history was defined as an imported case.

doi:10.6342/NTU202301433

**Table 2. The district-specific incidence rates of the SARS-CoV-2-positive cases, population sizes, and population densities in the three affected cities by the four time periods during Taiwan 2021 large outbreak**



| Districts in the three outbreak affected cities | Incidence Rates (per 100K) | | | | Population Size | Population Density | Area |
|---|---|---|---|---|---|---|---|
| | T1a | T1b | T2 | T3 | (May 2021) | (Size/Area) | ($km^2$) |
| | 4/16~5/6 | 5/7~5/14 | 5/15~6/22 | 6/23~7/31 | | | |
| **Wanhua District, Taipei City** | 9.42 | 129.16 | 810.25 | 86.05 | 180,396 | 20,378.66 | 8.85 |
| **Zhonghe District, New Taipei City*** | 1.46 | 17.09 | 241.03 | 12.70 | 409,649 | 20,336.03 | 20.14 |
| **Banqiao District, New Taipei City*** | 0.54 | 24.09 | 209.69 | 14.75 | 556,175 | 24,038.03 | 23.14 |
| **Datong District, Taipei City*** | 0.81 | 22.75 | 204.00 | 14.63 | 123,085 | 21,664.17 | 5.68 |
| **Yonghe District, New Taipei City*** | 2.74 | 15.09 | 199.56 | 14.20 | 218,652 | 38,267.35 | 5.71 |
| **Sanchong District, New Taipei City*** | 2.87 | 15.89 | 181.12 | 23.20 | 383,805 | 23,521.79 | 16.32 |
| **Zhongzheng District, Taipei City*** | 1.30 | 12.98 | 154.57 | 15.60 | 154,098 | 20,257.13 | 7.61 |
| Tucheng District, New Taipei City | 1.26 | 13.44 | 179.35 | 10.50 | 238,114 | 8,055.88 | 29.56 |
| Wugu District, New Taipei City | | 17.81 | 150.19 | 24.48 | 89,842 | 2,576.99 | 34.86 |
| Luzhou District, New Taipei City | 2.47 | 12.35 | 124.04 | 10.38 | 202,410 | 27,223.57 | 7.44 |
| Shilin District, Taipei City | | 3.63 | 126.55 | 17.09 | 275,204 | 4,412.57 | 62.37 |
| Shiding District, New Taipei City | | 13.33 | 133.32 | | 7,503 | 51.98 | 144.35 |
| Xinzhuang District, New Taipei City | | 8.98 | 122.48 | 11.11 | 422,978 | 21,429.30 | 19.74 |
| Wenshan District, Taipei City | 1.13 | 7.52 | 116.26 | 13.17 | 265,885 | 8,438.38 | 31.51 |
| Xinyi District, Taipei City | 0.94 | 8.49 | 107.64 | 17.01 | 211,920 | 18,908.43 | 11.21 |
| Taishan District, New Taipei City | 1.28 | 7.69 | 106.41 | 16.67 | 78,010 | 4,071.44 | 19.16 |
| Zhongshan District, Taipei City | 0.45 | 5.43 | 103.71 | 19.03 | 220,944 | 16,148.40 | 13.68 |
| Shenkeng District, New Taipei City | | 8.41 | 88.35 | 29.46 | 23,774 | 1,155.27 | 20.58 |
| Xindian District, New Taipei City | 0.33 | 9.26 | 103.19 | 11.57 | 302,503 | 2,516.13 | 120.23 |
| Nangang District, Taipei City | 1.70 | 9.35 | 99.53 | 5.95 | 117,606 | 5,384.30 | 21.84 |
| Shulin District, New Taipei City | 0.55 | 8.75 | 89.70 | 5.47 | 182,849 | 5,519.34 | 33.13 |
| Xizhi District, New Taipei City | 0.49 | 2.92 | 90.41 | 7.29 | 205,812 | 2,889.18 | 71.24 |
| Daan District, Taipei City | 0.67 | 8.03 | 79.08 | 10.06 | 298,891 | 26,307.59 | 11.36 |
| Beitou District, Taipei City | | 4.83 | 79.81 | 9.67 | 248,237 | 4,368.71 | 56.82 |
| Jinshan District, New Taipei City | | 4.77 | 85.86 | | 20,977 | 426.25 | 49.21 |
| Songshan District, Taipei City | | 8.08 | 73.23 | 9.09 | 198,120 | 21,331.21 | 9.29 |
| Guishan District, Taoyuan City | 0.61 | 4.84 | 75.60 | 6.04 | 165,261 | 2,294.73 | 72.02 |
| Bali District, New Taipei City | | 5.03 | 77.96 | 2.51 | 39,734 | 1,006.09 | 39.49 |
| Tamsui District, New Taipei City | 1.09 | 7.06 | 64.59 | 8.14 | 184,240 | 2,607.54 | 70.66 |
| Sanxia District, New Taipei City | | 5.14 | 65.99 | 5.14 | 116,708 | 609.60 | 191.45 |
| Linkou District, New Taipei City | | 2.44 | 63.52 | 5.70 | 122,792 | 2,267.55 | 54.15 |
| Neihu District, Taipei City | 0.36 | 4.99 | 48.53 | 8.57 | 280,318 | 8,876.81 | 31.58 |
| Taoyuan District, Taoyuan City | 0.22 | 2.84 | 38.39 | 5.89 | 458,376 | 13,169.98 | 34.80 |
| Daxi District, Taoyuan City | | | 31.48 | 10.49 | 95,276 | 906.35 | 105.12 |
| Shimen District, New Taipei City | | | 35.29 | | 11,353 | 221.46 | 51.26 |
| Yingge District, New Taipei City | | 1.14 | 31.86 | 1.14 | 87,850 | 4,158.62 | 21.12 |
| Ruifang District, New Taipei City | | | 33.48 | | 38,839 | 549.09 | 70.73 |
| Bade District, Taoyuan City | | 2.39 | 24.36 | 6.69 | 209,290 | 6,208.34 | 33.71 |
| Wulai District, New Taipei City | | | | 31.67 | 6,315 | 19.66 | 321.13 |
| Sanzhi District, New Taipei City | | | 31.15 | | 22,487 | 340.76 | 65.99 |
| Luzhu District, Taoyuan City | 2.40 | 2.40 | 19.19 | 4.20 | 166,744 | 2,208.46 | 75.50 |
| Wanli District, New Taipei City | | 13.92 | 13.92 | | 21,555 | 340.11 | 63.38 |
| Zhongli District, Taoyuan City | 0.71 | 1.18 | 18.69 | 5.44 | 422,582 | 5,522.50 | 76.52 |
| Longtan District, Taoyuan City | | 0.80 | 15.27 | 8.04 | 124,368 | 1,653.08 | 75.23 |
| Dayuan District, Taoyuan City | 1.07 | 1.07 | 15.99 | 5.33 | 93,814 | 1,073.48 | 87.39 |
| Yangmei District, Taoyuan City | | 0.57 | 14.21 | 0.57 | 175,836 | 1,972.96 | 89.12 |
| Pinglin District, New Taipei City | | | 14.95 | | 6,688 | 39.15 | 170.84 |
| Pingzhen District, Taoyuan City | | 2.19 | 10.50 | 2.19 | 228,594 | 4,786.99 | 47.75 |
| Guanyin District, Taoyuan City | | | 11.56 | 1.44 | 69,211 | 786.66 | 87.98 |
| Xinwu District, Taoyuan City | | | 4.06 | 2.03 | 49,218 | 578.92 | 85.02 |

Rankings of top district-specific incidence rates of SARS-CoV-2 in the three affected cities (Taipei, New Taipei, and Taoyuan cities) during the large 2021 outbreak (April 16 – July 31, 2021) in Taiwan. The Wanhua District had a 3.28-5.42-fold higher incidence than the next highest-ranking district.

* Distances to the center of the six districts in the three affected cities close to the Wanhua District were 2 km, 2.5 km, 3.5 km, 3.5 km, 3.9 km, and 4.1 km in the Zhongzheng, Yonghe, Zhonghe, Datong, Banqiao, and Sanchong Districts respectively.

doi:10.6342/NTU202301433

**Table 3. List of 24 SARS-CoV-2 Alpha variant cases from three risk-clusters in the onset of 2021 large outbreak in Taiwan (from December 9, 2020 to May 16, 2021)**

| Case ID | Onset date | Im/ Id | Loc. | Cluster | Age | Sex | Helicase R460K | Spike M1237I | Epi-linkage (Case ID) | Rt$^{+}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 783 | Dec. 9, 2020 | Im | PHL | - | 27 | M | R | M | NA | 0 |
| 799 | Dec. 26, 2020 | Im | GBR | - | 75 | M | R | M | NA | 0 |
| 792 | Dec. 27, 2020 | Im | GBR | - | 20 | M | R | M | NA | 0 |
| 804 | Dec. 28, 2020 | Im | GBR | - | 37 | M | K | M | NA | 0 |
| 837 | Dec. 29, 2020 | Im | GBR | - | 32 | M | K | M | NA | 0 |
| 958 | Feb. 26, 2021 | Im | USA | - | 52 | M | R | M | NA | 0 |
| 1048 | Mar. 23, 2021 | Im | PHL | - | 63 | M | K | M | NA | 0 |
| 1065 | Mar. 28, 2021 | Im | PHL | - | 32 | M | K | M | NA | 0 |
| 1050 | Mar. 29, 2021 | Im | EGY | - | 20 | M | K | M | NA | 0 |
| 1047 | Mar. 29, 2021 | Im | IDN | - | 23 | M | K | M | NA | 0 |
| 1081 | Mar. 10, 2021 | Im | IDN | - | 41 | M | K | M | NA | 0 |
| 1059 | Apr. 9, 2021 | Im | JPN | - | 24 | M | K | M | NA | 0 |
| 1091 | Apr. 16, 2021 | Im | USA | pilot | 52 | M | K | I | 1090, 1111, 1146 | 3 |
| 1105 | Apr. 19, 2021 | Im | USA | pilot | 46 | M | NA | NA | 1199, 1200 | 2 |
| 1078 | Apr. 18, 2021 | Im | USA | pilot | 52 | M | K | I | 1121 | 1 |
| 1153 | May. 1, 2021 | Im | USA | pilot | 37 | M | K | I | 1183, 1187 | 2 |
| 1102 | Apr. 24, 2021 | Im | USA | pilot | 38 | M | K | M | 1133, 1137 | 2 |
| 1120 | Apr. 17, 2021 | Id | NWT | hotel staff | 48 | M | NA | NA | 1127, 1128, 1129, 1145 | 4 |
| 1363 | May. 2, 2021 | Id | TPE | community | 62 | M | NA | NA | **Earliest Wanhua case and transmitted to 3445** | 1 |
| 3445 | May. 5, 2021 | Id | TPE | community | 53 | F | K | I | 4008, 4009, 4010, 4216, 4305 | 5 |
| 1203 | May. 7, 2021 | Id | TPE | community | 64 | M | NA | I | 1218, 1219, 1223, 1224, 1225, 1226, 1227, 1228, 1229, 1230, 1245, 1246, 1248, 1250, 1251, 1253, 1255, 1256, 1257 | 19 |
| 1257 | May. 9, 2021 | Id | TAO | 1203's family | 47 | M | NA | NA | 1275, 1276, 2140 | 3 |
| 3037 | May. 9, 2021 | Id | PIF | community (Wanhua travel history) | 65 | M | NA | NA | 3869, 4225, 4742, 4743 | 4 |
| 4742 | May. 16, 2021 | Id | KHH | 3037's family | 56 | M | K | I | 4741, 4743, 4744, 4826 | 4 |

Im: Imported, Id: Indigenous, Loc.: location, EGY: Egypt, GBR: United Kingdom, IDN: Indonesia, JPN: Japan, PHL: Philippines, KHH: Kaohsiung City, MIA: Miaoli County, NWT: New Taipei City, PIF: Pingtung County, TAO: Taoyuan City, TNN: Tainan City, TPE: Taipei City.

*ID-1363, 3445, and 1203 had visited the same tea house in the Wanhua District.

The mean ± SD of Rt (Reproductive number over time values) values: the five pilots (onset dates from 16 April to 1 May 2021) associated clusters was 2 ± 0.71 (range 1-3), one hotel-staff (onset date on 17 April 2021) associated cluster was 4, and six earlier community-associated clusters (onset dates for the first case of each cluster ranged from 2 May to 16 May 2021) was 6 ± 6.51 (range 1-19), $p$ = 0.007 (One-way ANOVA)

**Table 4. The percentages of descendants of SARS-CoV-2 Alpha variant strains in the three time periods with their nucleotides were identical to the 9 strains from the cases with onset dates in the T1a period (April 16 to May 6)**

| Case ID | T1b period<br>May 7-May 14<br>(N = 5) | T2 period<br>May 15- June 22<br>(N = 56) | T3 period<br>June 23- August 31<br>(N = 14) |
|---|---|---|---|
| 1091 | 0 | 0 | 0 |
| 1079/1078 | 0 | 0 | 0 |
| 1102 | 0 | 0 | 0 |
| 1145 | 0 | 0 | 0 |
| 1137 | 0 | 0 | 0 |
| 1154 | 0 | 0 | 0 |
| 3445/1263/1186 | 3 (60%) | 16 (28.57%) | 0 |
| 1183 | 0 | 0 | 0 |
| 1187 | 0 | 0 | 0 |

doi:10.6342/NTU202301433

**Table 5. The number of nucleotide variations of the 81 indigenous SARS-CoV-2 Alpha variant strains in Taiwan compared to the Alpha reference strains (UK-MILK-ACF9CC) in the three time periods**

| | No. of strains | No. of SNV<br>(mean ± SD) | % SNV<br>(mean ± SD) | P value<br>(Period vs. all) |
|---|---|---|---|---|
| T1 period<br>April 16-May 14 | 11 | 12.36 ± 4.18 | 0.0413 ± 0.014 | 0.4954 |
| T2 period<br>May 15- June 22 | 56 | 11.29 ± 1.44 | 0.0377 ± 0.0048 | 0.135 |
| T3 period<br>June 23- August 31 | 14 | 13.43 ± 2.31 | 0.0449 ± 0.0077 | 0.0154* |

SNV: single nucleotide variation

*P* value: Student's t-test; *: <0.05.

doi:10.6342/NTU202301433

**Table 6. Mutation prevalence percentages of the 101 Taiwan Alpha variant strains compared to those of the Alpha variant reference strain (UK-MILK-ACF9CC)**

| Residue | Ref | | No. of strains | | Mutation Prevalence (%) | No. of strains |
|---|---|---|---|---|---|---|
| Hel_460 | R | 3.96% | 4 | K | **96.04%** | **97** |
| S_1237 | M | 13.86% | 14 | I | **86.14%** | **87** |
| nsp6_260 | L | 89.11% | 90 | F | **10.89%** | **11** |
| M_82 | I | 94.06% | 95 | S/T | **5.94%** | **6** |
| nsp1_170 | T | 95.05% | 96 | I | **4.95%** | **5** |
| N_135 | T | 95.05% | 96 | I | **4.95%** | **5** |
| nsp2_169 | L | 97.03% | 98 | F | **2.97%** | **3** |
| N_398 | A | 97.03% | 98 | V | **2.97%** | **3** |
| nsp4_17 | F | 98.02% | 99 | L | **1.98%** | **2** |
| 3CLpro_160 | C | 98.02% | 99 | F | **1.98%** | **2** |
| nsp8_141 | T | 98.02% | 99 | M | **1.98%** | **2** |
| nsp9_83 | P | 98.02% | 99 | L | **1.98%** | **2** |
| RdRp_671 | G | 98.02% | 99 | S | **1.98%** | **2** |
| nsp15_185 | V | 98.02% | 99 | I | **1.98%** | **2** |
| S_69 | - | 98.02% | 99 | H | **1.98%** | **2** |
| S_70 | - | 98.02% | 99 | V | **1.98%** | **2** |
| S_144 | - | 98.02% | 99 | Y | **1.98%** | **2** |
| ORF3a_15 | L | 98.02% | 99 | F | **1.98%** | **2** |
| ORF7a_96 | L | 98.02% | 99 | F | **1.98%** | **2** |
| ORF8_27 | X | 98.02% | 99 | Q | **1.98%** | **2** |
| ORF8_68 | K | 98.02% | 99 | - | **1.98%** | **2** |

By comparing the 101 Taiwan strains with the WHO's reference Alpha variants (UK-MILK-ACF9CC), we were able to observe 141 amino acid changes during the outbreak. We further calculated the mutation prevalence (in **Boldface text with black shadow**) of these amino acid changes during the outbreak and found that the top prevalence percentages of the two amino acid mutations were 96.04% (97/101) for Helicase-R460K and 86.14% (87/101) for Spike-M1237I presence during the outbreak. Blocks in different colors represent 31 different proteins.

**Table 7. Univariate analysis and Multivariable regression analysis of the factors associated with the frequency of SARS-CoV-2 genome sequences identical to those of the dominant strains from cluster cases of ID-3445/1186/1263 with their onset dates from December 9, 2020 to October 31, 2021**

| | | | **Univariate analysis** | | **Multivariable regression analysis** | |
|---|---|---|---|---|---|---|
| **Factors** | **Seq** | **Identical to dominant 3445/1186/1263 cluster (%)** | **Crude OR (95% CI)** | ***P* value** | **Adjusted OR (95% CI)** | ***P* value** |
| **(1) Epidemic periods** | | | | | | |
| T1 period | 11 | 45.46% (5/11) | 1.74 (0.37-7.92) | **0.0097**** (T1 vs. T2 = 0.492) **(T3 vs. T2 = 0.014*)** | **0.18 (0.05-0.62)** | **0.007**** |
| T2 period | 56 | 32.14% (18/56) | Reference | | | |
| T3 period | 14 | 0% (0/14) | 0 (0-0.74) | | | |
| **(2) Epicenter** | | | | | | |
| Epicenter | 50 | 16% (8/50) | 0.21 (0.06-0.64) | **0.0024**** | **0.15 (0.04-0.47)** | **0.001**** |
| Non-epicenter | 31 | 48.38% (15/31) | Reference | | | |
| **(3) Daily city/county-specific vaccination coverage of the 1st-dose COVID-19 (%)** | | | | | | |
| < 2.87% | 41 | 39.02% (16/41) | Reference | **0.0479*** | - | - |
| ≥ 2.87% | 40 | 17.5% (7/40) | 0.34 (0.1-1.02) | | | |
| **(4) Daily district-specific daily public transport ridership (per 10K passengers)** | | | | | | |
| < 3.344 | 40 | 35% (14/40) | Reference | 0.225 | - | - |
| ≥ 3.344 | 41 | 21.95% (9/41) | 0.53 (0.17-1.55) | | | |
| **(5) Daily cases** | | | | | | |
| < 24 | 26 | 34.62% (9/26) | Reference | 0.435 | - | - |
| ≥ 24 | 55 | 25.45% (14/55) | 0.65 (0.21-2.04) | | | |
| **(6) Monthly district-specific population size (per 100K peoples by district)** | | | | | | |
| < 1.8 | 16 | 56.25% (9/16) | Reference | **0.011*** | - | - |
| ≥ 1.8 | 65 | 21.54% (14/65) | 0.22 (0.06-0.79) | | | |
| **(7) Monthly district-specific population density (10K pop. size / district area km$^2$)** | | | | | | |
| < 2 | 28 | 46.43% (13/28) | Reference | **0.018*** | - | - |
| ≥ 2 | 53 | 18.87% (10/53) | 0.27 (0.09-0.83) | | | |
| **(8) Age** | | | | | | |
| < 53 | 32 | 25% (8/32) | Reference | 1 | - | - |
| ≥ 53 | 35 | 25.71% (9/35) | 1.04 (0.3-3.65) | | | |
| **(9) Gender** | | | | | | |
| Female | 43 | 25.58% (11/43) | Reference | 0.625 | - | - |
| Male | 38 | 31.58% (12/38) | 1.34 (0.46-3.97) | | | |

Seq: sequence, OR: odds ratio; CI: confidence interval; *P* value: Fisher's exact test; *: <0.05; **: <0.01.

Multivariable regression formula: binomial linear regression (Identical to ID-3445 = Epidemic periods + Epicenter), AIC = 83.627

T1 Period (April 16 – May 14; pre-Level 3 Restrictions), T2 Period (May 15 – June 22; post-Level 3 Restrictions, but pre-Zero-COVID Policy), and T3 Period (June 23 – August 31; post-Zero-COVID Policy).

We used Fisher's exact test to assess all factors between subgroups due to the small sample size. Variance inflation factors (VIF > 5) were used to evaluate collinearity among factors, and the statistically significant factors without collinearity were included in the final multivariable regression model (Tables 8 and 9).

**Table 8. Binomial linear regression and Variance Inflation Factor (VIF)**

| | (1)<br>Epidemic periods | (2)<br>Epicenter | (3)<br>Vaccination coverage | (4)<br>Population size | (5)<br>Population density |
|---|---|---|---|---|---|
| 1+2+3+4+5 | 1.553 | 5.762 | 1.395 | 1.891 | 6.751 |
| **1+2+3+4** | **1.536** | **1.818** | **1.369** | **1.597** | **-** |
| 2+3+4+5 | - | 1.604 | 5.895 | 1.365 | 5.979 |

Variance inflation factors (VIF > 5) were used to evaluate collinearity among factors, and the statistically significant factors without collinearity were included in the final multivariable regression model.

doi:10.6342/NTU202301433

**Table 9. Multivariable logistic regression (binomial) analysis associated with the frequency of SARS-CoV-2 genome sequences identical to ID-3445/1186/1263 using stepwise method by Akaike information criterion (AIC) and backward/forward**

| | Estimate | Std.<br>Error | z value | Adjusted<br>OR | P value |
|---|---|---|---|---|---|
| Identical to ID-3445/1186/1263 = Epidemic periods + Epicenter + Vaccination coverage + Population size, AIC = 86.451 | | | | | |
| Epidemic periods | -1.739 | 0.736 | -2.36 | 0.176 | 0.018* |
| Epicenter | -1.477 | 0.763 | -1.94 | 0.228 | 0.053. |
| Vaccination coverage | -0.067 | 0.691 | -0.1 | 0.935 | 0.923 |
| Population size | -0.832 | 0.776 | -1.07 | 0.435 | 0.284 |
| Identical to ID-3445/1186/1263 = Epidemic periods + Epicenter + Population size, AIC = 84.46 | | | | | |
| Epidemic periods | -1.775 | 0.64 | -2.77 | 0.17 | 0.006** |
| Epicenter | -1.499 | 0.73 | -2.06 | 0.223 | 0.04* |
| Population size | -0.82 | 0.767 | -1.07 | 0.44 | 0.285 |
| **Identical to ID-3445/1186/1263 = Epidemic periods + Epicenter, AIC = 83.627** | | | | | |
| **Epidemic periods** | -1.738 | 0.639 | -2.72 | 0.176 | **0.007**** |
| **Epicenter** | -1.934 | 0.606 | -3.19 | 0.145 | **0.001**** |

*P* value: Fisher's exact test; *: <0.05; **: <0.01.

# Figures

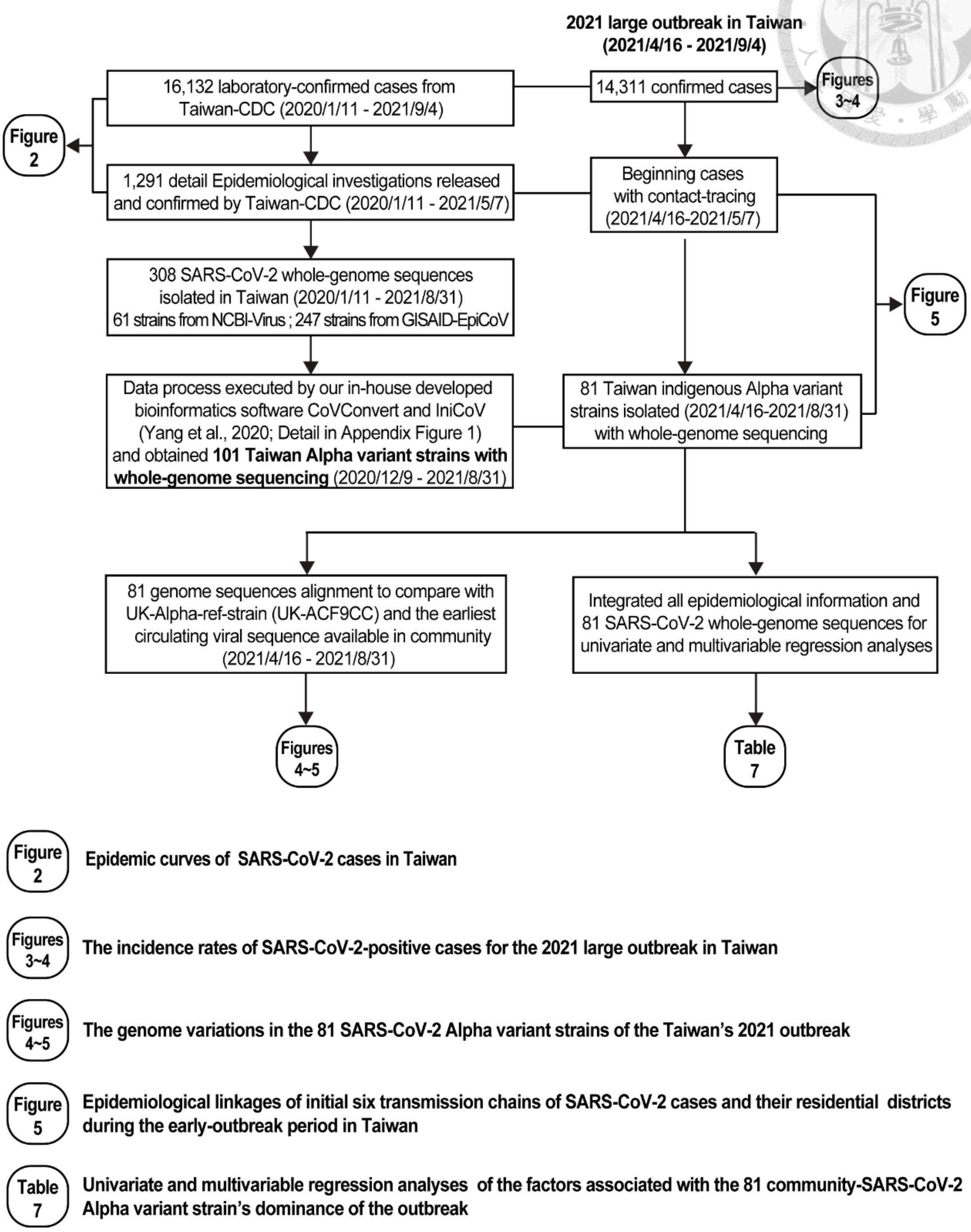

2021 large outbreak in Taiwan
(2021/4/16 - 2021/9/4)
16,132 laboratory-confirmed cases from Taiwan-CDC (2020/1/11 - 2021/9/4)
14,311 confirmed cases
Figures 3~4
Figure 2
1,291 detail Epidemiological investigations released and confirmed by Taiwan-CDC (2020/1/11 - 2021/5/7)
Beginning cases with contact-tracing (2021/4/16-2021/5/7)
308 SARS-CoV-2 whole-genome sequences isolated in Taiwan (2020/1/11 - 2021/8/31)
61 strains from NCBI-Virus ; 247 strains from GISAID-EpiCoV
Figure 5
Data process executed by our in-house developed bioinformatics software CoVConvert and IniCoV (Yang et al., 2020; Detail in Appendix Figure 1) and obtained 101 Taiwan Alpha variant strains with whole-genome sequencing (2020/12/9 - 2021/8/31)
81 Taiwan indigenous Alpha variant strains isolated (2021/4/16-2021/8/31) with whole-genome sequencing
81 genome sequences alignment to compare with UK-Alpha-ref-strain (UK-ACF9CC) and the earliest circulating viral sequence available in community (2021/4/16 - 2021/8/31)
Integrated all epidemiological information and 81 SARS-CoV-2 whole-genome sequences for univariate and multivariable regression analyses
Figures 4~5
Table 7
Figure 2
Epidemic curves of SARS-CoV-2 cases in Taiwan
Figures 3~4
The incidence rates of SARS-CoV-2-positive cases for the 2021 large outbreak in Taiwan
Figures 4~5
The genome variations in the 81 SARS-CoV-2 Alpha variant strains of the Taiwan's 2021 outbreak
Figure 5
Epidemiological linkages of initial six transmission chains of SARS-CoV-2 cases and their residential districts during the early-outbreak period in Taiwan
Table 7
Univariate and multivariable regression analyses of the factors associated with the 81 community-SARS-CoV-2 Alpha variant strain's dominance of the outbreak

**Figure 1. Flow diagram of study design to analyze SARS-CoV-2-positive cases in Taiwan from January 11, 2020 to September 4, 2021**

We used CoVConvert to check data quality and obtained different reading frames for IniCoV to identify polygenetic consensus signatures.

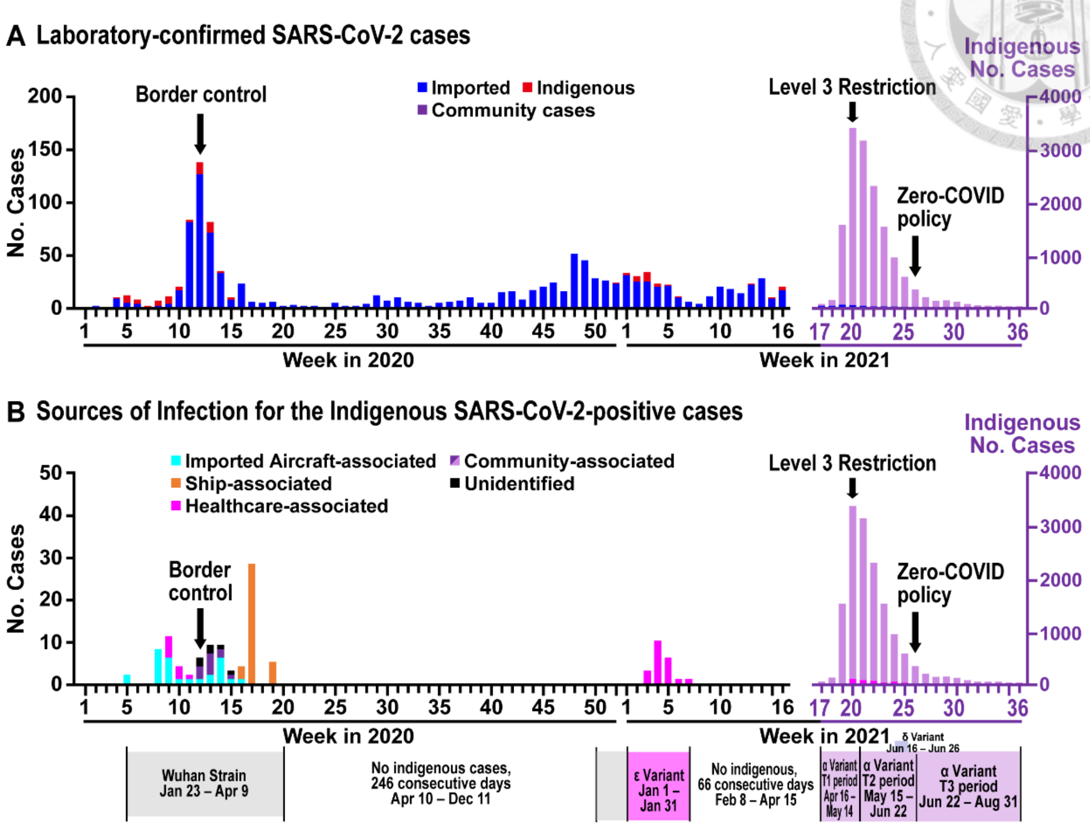


**Figure 2. Epidemic curves of laboratory-confirmed SARS-CoV-2 cases plotted with three major government countermeasures in Taiwan from January 1, 2020, to September 4, 2021**

The weekly numbers of laboratory-confirmed SARS-CoV-2 cases from the 1st week of 2020 to the 36th week of 2021 (i.e. 4 September 2021 when the daily case number dropped below 10) were obtained from Taiwan CDC Open Data Portal (https://data.cdc.gov.tw/en). The bar graphs show the distribution of cases based on the onset weeks, and the arrows indicate when countermeasures were implemented (Detail described in Supplementary). The confirmed indigenous cases caused by the three variants of SARS-CoV-2 are: (1) Alpha variants (14,311 cases, April 16-September 4, 2021), (2) Epsilon variants (19 cases, January 1-January 31, 2021), and (3) Delta variants (15 cases, June 16-June 26, 2021).

doi:10.6342/NTU202301433

(A) Weekly numbers of confirmed imported (shiny blue bars), indigenous (red bars), and the 2021 large outbreak (light purple bars, from the 16th - 36th week of 2021). The two waves of the imported cases involved western holidays:

1) the 48th week of 2020 (early December, 50 cases after Thanksgiving holidays) through the 1st week of 2021 (38 cases after New Year's holidays), and

2) 16th-19th week of 2021 after Spring breaks (mid-April, mean ± S.D.: 29.5 ± 12.95 cases/week).

(B) Sources of the infection for indigenous cases involved into five major risk groups from January 1, 2020, to September 4, 2021 before the 2021 large outbreak (1st week of 2020 to the 7th week of 2021):

1) **Imported Aircraft-associated cases** (light blue bars, contact history with imported cases) = 29/114, 25.4%,

2) **Healthcare-associated cases** (magenta bars) = 30/114, 26.3% (9 cases in the 9th - 11th weeks of 2020 and 21 cases in the 2nd -6th weeks of 2021),

3) **Community-associated cases** (purple bars, indigenous cases who had no travel history three days before the onset of illness) = 12/114, 10.5%,

4) **Ship-associated cases** (orange bars, cruise ships and naval crews) = 36/114, 31.6% (36 cases in the 16th -19th weeks of 2020), and

5) **Cases with unidentified sources** (black bars, no clear sources of infection following thorough epidemiological investigation) = 7/114, 6.2%.

**During the 2021 large outbreak (17th week of 2021 to the 36th week of 2021):**

1) **Healthcare-associated cases** (magenta bars) = 244/14,311, 1.7% (244 cases in the 20th -25th weeks of 2021),

2) **Community-associated cases and cases with unidentified sources** (light purple bars, indigenous cases who had no travel history three days before the onset of illness) = 14,067/14,311, 98.3%

doi:10.6342/NTU202301433

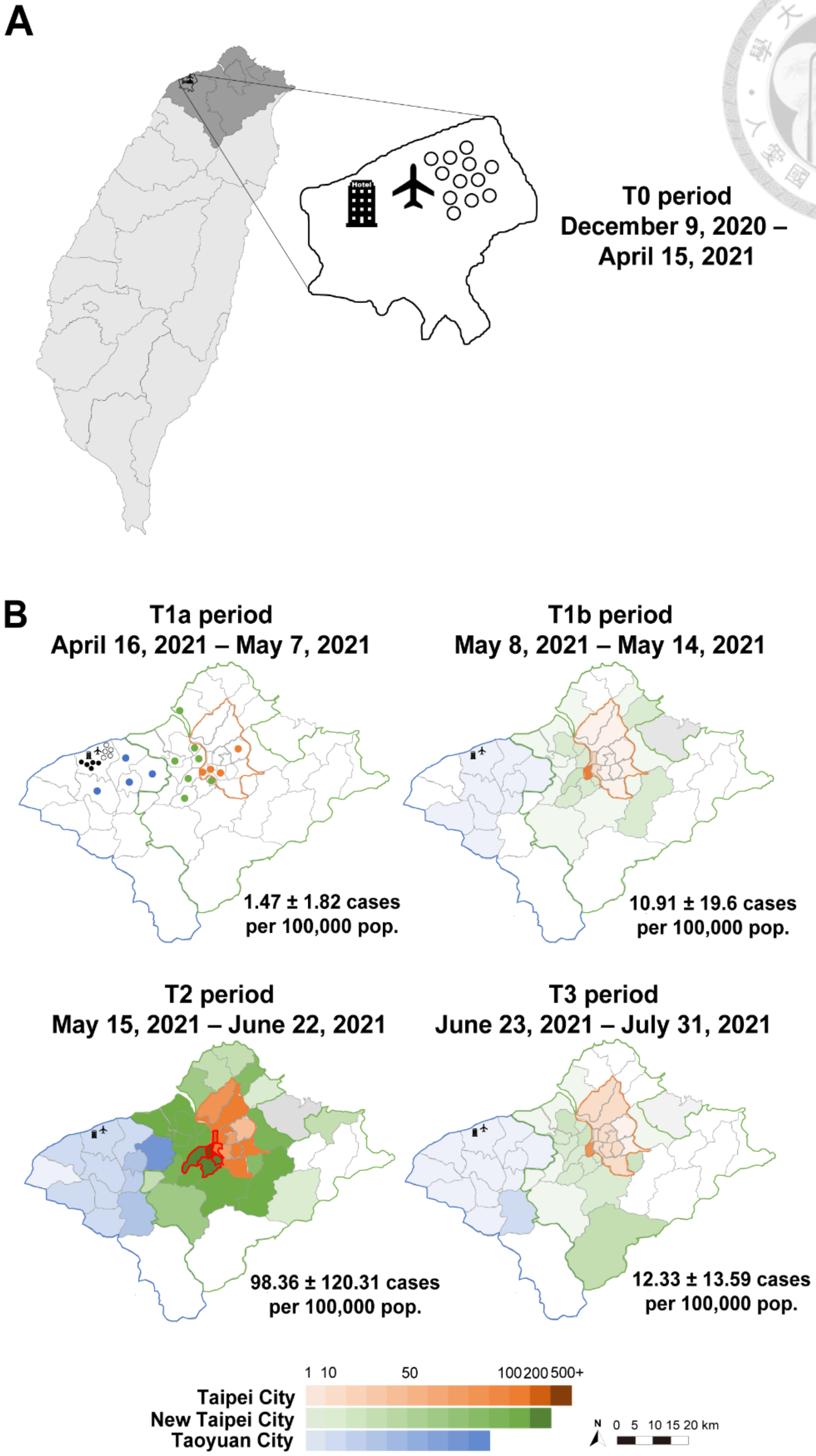

A
Hotel
T0 period
December 9, 2020 –
April 15, 2021
B
T1a period
April 16, 2021 – May 7, 2021
1.47 ± 1.82 cases
per 100,000 pop.
T1b period
May 8, 2021 – May 14, 2021
10.91 ± 19.6 cases
per 100,000 pop.
T2 period
May 15, 2021 – June 22, 2021
98.36 ± 120.31 cases
per 100,000 pop.
T3 period
June 23, 2021 – July 31, 2021
12.33 ± 13.59 cases
per 100,000 pop.
1 10 50 100 200 500+
Taipei City
New Taipei City
Taoyuan City
N
0 5 10 15 20 km

**Figure 3. The incidence rates of laboratory-confirmed SARS-CoV-2 cases in the three major affected cities and other areas of Taiwan from Pre-outbreak and during the large 2021 outbreak (from December 9, 2020 through July 31, 2021)**

**(A) Pre-outbreak (T0, December 9, 2020 through April 15, 2021)**

Symbols and lines shown the Taoyuan International Airport and quarantine hotel in Dayuan District, and imported cases (in the circle).

**(B) During the outbreak (T1-T3, April 16, 2021 through July 31, 2021)**

The colour gradients show the incidence rate (per 100,000 residents) in each district in the three major affected cities across five different time periods [T1a (April 16, 2021-May 7, 2021; early-outbreak), T1b (May 8, 2021-May 14, 2021; pre-Level 3 Restrictions), T2 (May 15, 2021-June 22, 2021; post-Level 3 Restrictions, but pre-Zero-COVID Policy), T3 (June 23, 2021-July 31, 2021; post-Zero-COVID Policy].

The early-outbreak (T1a) cutting time point on May 7 because the last pilot case ID-1183 and ID-1187 who had onset dates on May 6.

The Daily mean numbers, and red lines show the six districts neighboring the area where the epidemic began (Wanhua District).

The "epicenter" of this outbreak was defined as the district with the highest incidence and its bordering districts (Table 3). Data on district-specific population sizes was obtained using Taiwan household registry information from the Ministry of Interior population sizes in May 2021 were 2,574,704 in Taipei City, 4,026,019 in New Taipei City, and 2,270,939 in Taoyuan City (https://www.ris.gov.tw/app/en/346).

The numerator represents number of new cases occurring at that specific time period in the same studied district as the denominator. The monthly incidence rates are shown as "mean ± SD" before and after the 2021 outbreak: Taipei City: 0.152 ± 0.161 vs 46.691± 56.311 ($p$ < 0.0001), New Taipei City: 0.162 ± 0.066 vs 42.161 ± 49.067 ($p$ < 0.0001), and Taoyuan City: 0.265 ± 0.303 vs 8.288 ± 7.606 ($p$ < 0.0001).

doi:10.6342/NTU202301433

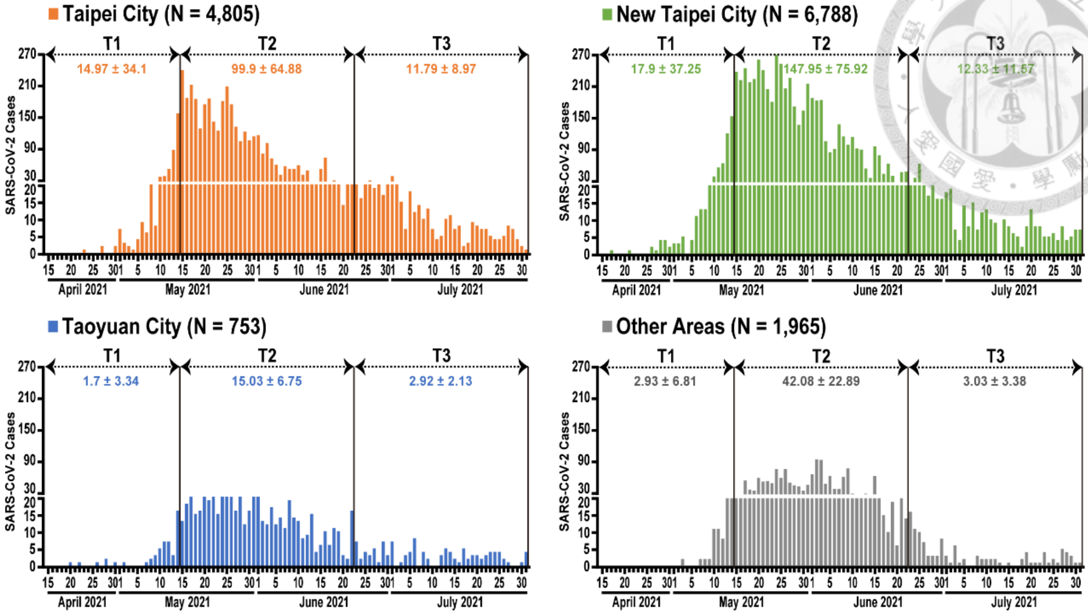


**Figure 4. Epidemic curves of laboratory-confirmed SARS-CoV-2 cases in the three major affected cities and other areas of Taiwan in the large 2021 outbreak**

The bar graphs show the distributions of cases based on onset dates from April 15 through July 31, 2021. Because the daily numbers of confirmed SARS-CoV-2 cases in Taoyuan City were much lower than those in Taipei and New Taipei cities, we used two scales (0-20 and 30-270 cases) that are separated by white lines in Taipei City, New Taipei City, and other areas.

The 81 indigenous strains involved three time periods based on population-based interventions: 1) T1 Period (April 15-May 14; pre-Level 3 restrictions), 2) T2 Period (May 15-June 21; post-Level 3 restrictions, but pre-Zero COVID policy), and 3) T3 Period (June 23-August 31; post-Zero COVID policy).

The mean weekly numbers are shown as "mean ± SD" before and after the 2021 outbreak, Taipei City: 0.516 ± 1.807 vs 1201.25 ± 1372.857 ($p$ < 0.0001), New Taipei City: 0.516 ± 2.38 vs 1939.429 ± 1975.374 ($p$ < 0.0001), and Taoyuan City: 2.323 ± 5.11 vs 215.143 ± 172.747 ($p$ < 0.0001).

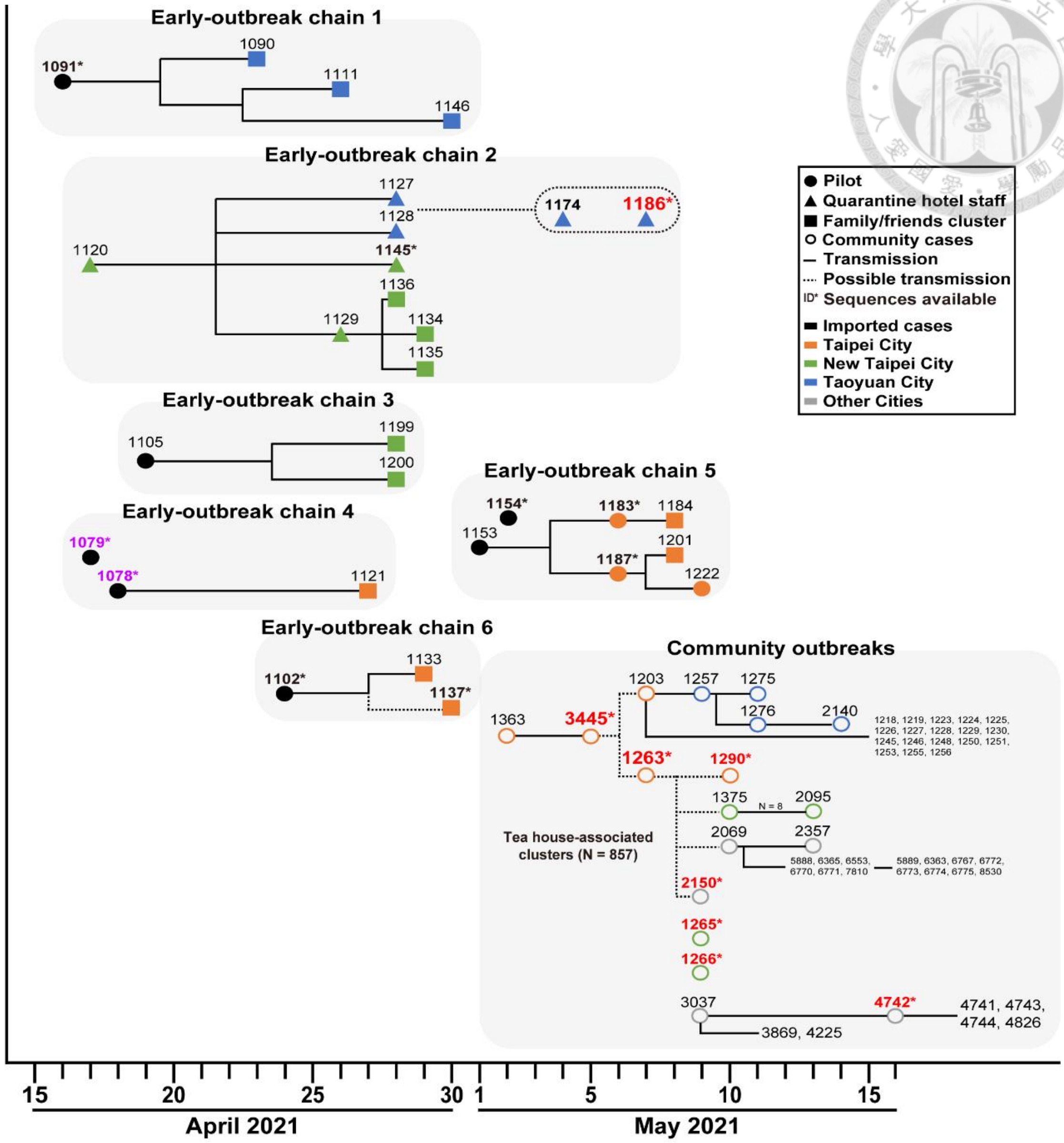


**Figure 5. Epidemiological linkages of initial six transmission chains of SARS-CoV-2 cases and their residential districts at the T1 period in the three major affected cities of Taiwan**

The initial six Early-outbreak chains were drawn according to Taiwan CDC epidemiological investigations. Symbols and lines shown in each Early-outbreak chain represent the characteristics of the subjects who transmitted the virus (pilot in the circle; hotel staff in the triangle) or new cases from family or friends contacts (in the square) through direct (solid lines) or indirect (dotted lines) transmission. The numbers shown are Case IDs. ID numbers that are red with a star sign have viral sequences available in the GISAID-EpiCoV database.

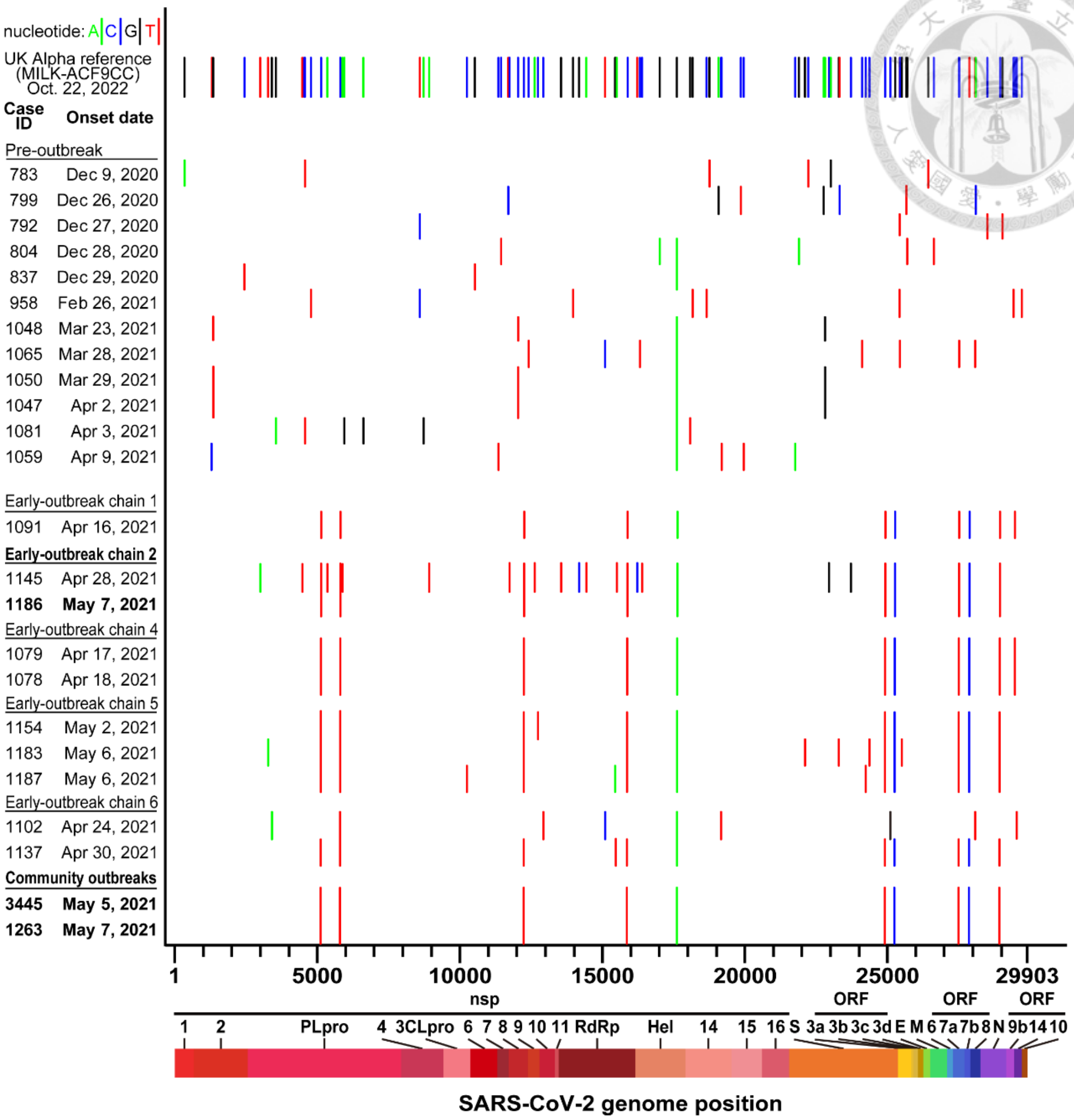


**Figure 6. Nucleotide variations of 24 SARS-CoV-2 Alpha variant strains isolated from Pre-outbreak and six different Early-outbreak chains in Taiwan (from December 9, 2020 to May 7) compared to those of the Alpha variant reference strain (UK MILK-ACF9CC)**

The whole genome sequences of 24 Taiwan SARS-CoV-2 Alpha variant strains were compared to UK-MILK-ACF9CC. The nucleotide variations between Taiwan's strains and UK-MILK-ACF9CC strain are shown in vertical lines which represent nucleotide A (green), C (blue), G (black), and T (red), respectively.

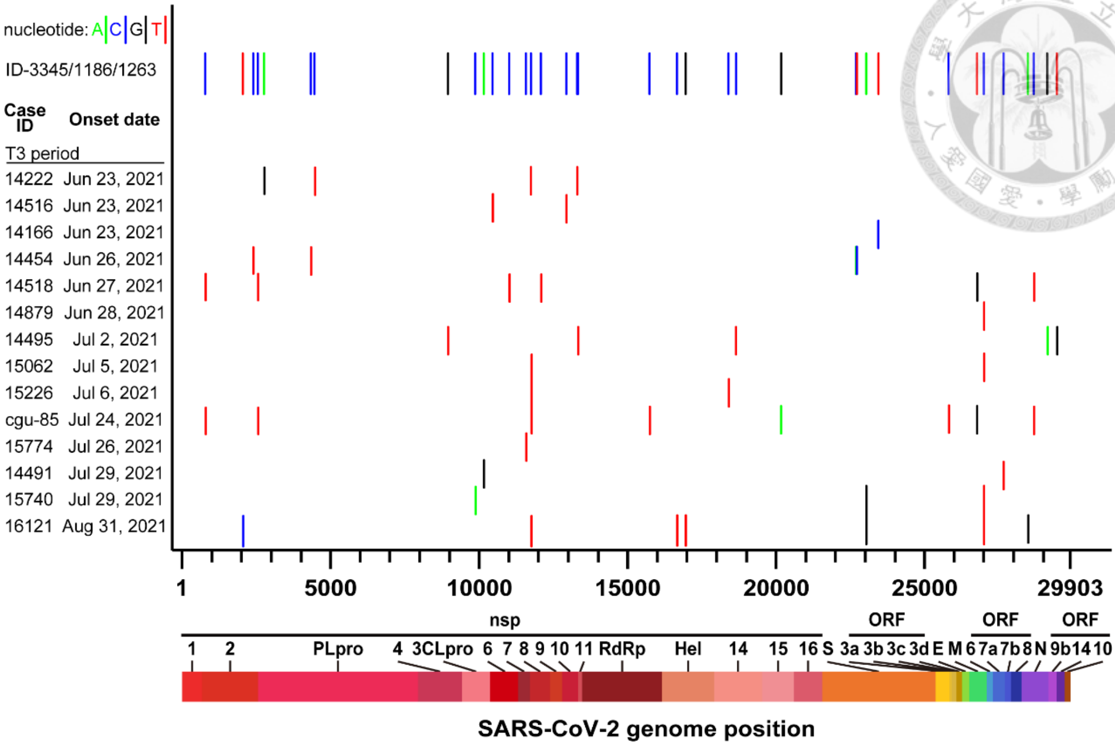


**Figure 7. Nucleotide variations of 14 SARS-CoV-2 Alpha variant strains isolated in T3 period compared to the predominant ID-3445/1186/1263 strain**

The whole genome sequences of 14 Taiwan SARS-CoV-2 Alpha variant strains isolated in T3 period were compared to ID-3445/1186/1263 strain. The nucleotide variations between T3 strains and T1/T2 predominant ID-3445/1186/1263 strain are shown in vertical lines which represent nucleotide A (green), C (blue), G (black), and T (red), respectively.

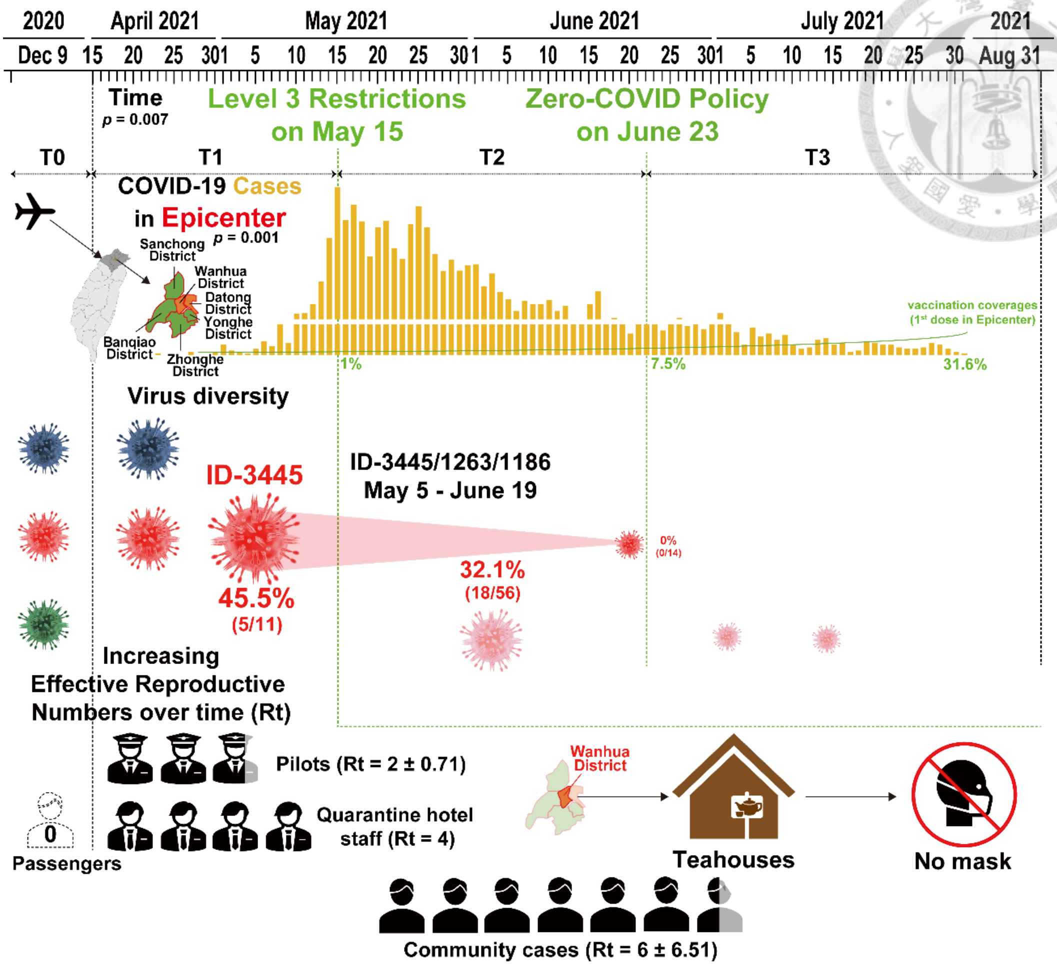


**Figure 8. The figure summarizes our major findings**

This figure summarizes our major findings in this study. Before the outbreak (T0, Pre-outbreak), the imported Alpha variant strains were heterogeneous with high viral genome divergence. However, such diversity significantly decreased during the T1 period ($p < 0.0001$), when the dominant virus strains with selective advantages appeared. We also investigated the epidemiological conditions in Taiwan that facilitated the emergence of the predominant virus strain in the T1 period. The effective reproductive numbers over time (Rt) for the viruses from the imported cases were all zero at T0 period before the outbreak. However, the mean Rt values of the viruses from the pilots to quarantine hotel staff and subsequent dominant virus strains in the community (i.e. same sequence identities as the ID-3445/1263/1186) increased rapidly. Specific epidemiological conditions, including unmasked dining in many teahouses, and customers' movement across teahouses, helped the dominant Alpha variant strains with a selective advantage. This study demonstrated that natural selection of a dominant virus strain (prior to immune selection) progressed in three stages: (1) selection started from a diverse virus pool (i.e. imported viruses at T0), (2) selection advantages increased through virus replication, in which the progeny virus had more advantages than its parent, and (3) selection of a fast-spreading strain through human-to-human transmission when a community had suitable epidemiological conditions (i.e. our T1 period and epicenter). Most importantly, COVID-19 cases dropped sharply alongside the two important population-intervention strategies (Level 3 Restrictions and Zero-COVID policy).

doi:10.6342/NTU202301433

# Chapter 6
# Perspectives

doi:10.6342/NTU202301433

In recent years, identifying and analyzing emerging infectious viruses have become increasingly crucial due to their continuous threat. The rise in episodes of diseases such as influenza virus and SARS-CoV-2 highlights the importance of developing robust sequence analysis software capable of processing and analyzing vast amounts of sequencing data generated by state-of-the-art technologies. With its segmented genome comprising eight genes, the influenza virus poses a unique challenge as these genes can reassort from other species. Therefore, careful consideration of each gene segment's identification and epidemiological information is necessary during the analysis process. Tracing the earliest evolutionary origin of the influenza virus is made even more difficult due to the absence of a reference virus strain.

On the other hand, SARS-CoV-2, consisting of a single gene with over 30,000 base pairs, benefits from a known early standard reference— Wuhan strain. This feature allows for rapid identification of amino acid residue differences through comparison. However, when comparing numerous sequences, challenges arise in calculating a consensus due to alignment performance issues. Because of these challenges, diligent improvements in excising software tools are necessary to organize, visualize, and analyze virus sequence data. Developing more advanced tools is crucial for applying these sequence data in virology, immunology, and epidemiology, gaining deeper insights and understanding.

doi:10.6342/NTU202301433

### Current web-based sequence analysis tools

Currently, several tools are available to identify viral strains. One option is to upload sequences to NCBI Virus for BLAST analysis. Another platform, GISAID-EpiCoV, enables SARS-CoV-2 sequence comparison with the reference Wuhan strain. It provides a list of amino acid differences and facilitates the determination of the virus lineage or clade. Platforms such as  BV-BRC offer the Sequence Feature Variant Type (Flu-SFVT) method for other viruses, such as the influenza virus. These platforms allow users to upload individual gene segments and identify amino acid differences. It also links this information to literature-based data on pathogenicity and drug resistance.

Furthermore, when comparing groups of viral strains within a specific population, constructing phylogenetic trees provides insights into their evolutionary trends. NextStrain is an example of a tool that performs real-time phylogenetic tree construction to track viral evolution trends. Sampling a clade of virus sequences helps understand their spatiotemporal changes and identify unique evolutionary clades (Hadfield et al., 2018). By leveraging these tools, researchers can transform virus sequences into comprehensive gene annotations with extensive descriptions of variant residues.

### Requirements of essential skills for executing the analysis

Programming skills are often necessary to effectively handle virus sequences,

doi:10.6342/NTU202301433

including sequence organization, alignment, and analysis. In recent years, the development of pipeline software and online tools has led to the emergence of packaged analysis workflows. These workflows streamline the analysis process and make it more accessible. For instance, Bioconda hosts numerous tools that can process influenza viral sequences. One such tool is the nf-flu tool, which focuses on analyzing each of the virus's eight segmented genes. It compares the viral sequences to a specified reference strain (Kruczkiewicz, 2022). Another online tool, INSaFLU, directly handles raw viral sequencing files and compares them with representative viral strains. It then utilizes the Snippy tool to generate consensus sequences and highlights divergent amino acid residues (Borges et al., 2018). While these tools provide direct residue variation annotations from raw FASTA sequence files, additional analysis may be necessary for re-aligning sequences, which can be time-consuming.

Furthermore, these tools often require the availability of reference strain for comparison. However, they may not integrate cross-segment gene analysis with other relevant information. Therefore, utilizing these tools requires a combination of programming skills and an understanding of the limitations of the available workflows.

### Unique features of our analysis software packages

Our software package offers a unique set of features that address the limitations of

doi:10.6342/NTU202301433

existing tools while incorporating additional advantages. It provides efficient virus sequence processing, whole genome sequence visualization, comparison, and consensus sequence analysis. With our software, users can automatically process FASTA sequence files, perform sequence alignment, and translate them into amino acid sequences without requiring any programming skills. Whether the viral genomes are segmented or composed of disparate segments, our software can integrate into a complete viral genome using a strain-based alignment method. This flexibility allows for easy grouping and incorporation into subsequent analyses. In addition to robust sequence processing, our software excels in sequence visualization, enabling real-time adjustments and in-depth exploration through an interactive GUI platform. This feature differentiates our software from commonly used tools like BioEdit and Integrative Genomics Viewer (IGV). While these tools can display multiple viral sequences, our software provides a more user-friendly and intuitive interface for visualizing viral sequences. Incorporating strain name information and amino acid residues into an Excel spreadsheet enhances visualization. Moreover, the software facilitates easy grouping without requiring sequence alignment, empowering users to generate group-specific consensus sequences with less efforts.

Furthermore, our software's advantages extend beyond sequence visualization. It offers automatic analytical workflows, distinguishing it from commercial software

doi:10.6342/NTU202301433

packages, such as QIAGEN CLC Genomics Workbench and bioMerieux bionumerics. These workflows enable the study of viral mutations, integration of information from Ingenuity Pathway Analysis (IPA), construction of phylogenetic trees, and other analytical methods. This comprehensive approach illuminates new directions for research, enhancing the user’s ability to delve into the complexities of viral sequence analysis.

**Limitations and future improvement**

While our software introduces new features that address limitations associated with online tools, several aspects require further improvement. Platforms like the Viral Bioinformatics Resource Center and NCBI Virus linked to PubMed offer valuable experimental corroboration and information for understanding viruses (Brister et al., 2015; Olson et al., 2022). However, our software cannot provide real-time updates and connect our findings on highly variable amino acids to scientific literature, which would reveal their epidemiological significance. Moreover, predicting the impact of unknown viral changes requires experimental demonstration to establish their relevance. Future enhancements to our software could include integrating automatic sequence generation structures, enabling a faster analysis of available therapeutics. Additionally, employing large language models (LLMs) with AI technology, such as programs integrated with ChatGPT 4 (Stokel-Walker & Van Noorden, 2023), can offer a comprehensive platform

doi:10.6342/NTU202301433

for literature review, organization, and annotation of additional points. This approach would provide great tools and insights, facilitating future research development and establishing a solid foundation for research directions.

In summary, our software combines the benefits of efficient viral sequence processing, comprehensive visualization, and automated analytical workflows. These features provide a versatile and user-friendly real-time consensus sequence analysis platform, facilitating in-depth exploration and opening new avenues for virology, epidemiology, and clinical research. While our software addresses some limitations, there is scope for improvement in real-time updates, linking findings to scientific literature, experimental demonstration of viral changes, and integrating advanced language models for enhanced literature search and annotation. These enhancements will bolster the software's capabilities, providing researchers with a robust virus analysis and exploration platform.

doi:10.6342/NTU202301433

Appendix I

BMC Bioinformatics

SOFTWARE

Open Access

# FluConvert and IniFlu: a suite of integrated software to identify novel signatures of emerging influenza viruses with increasing risk

Chin-Rur Yang[1], Chwan-Chuen King[2], Li-Yu Daisy Liu[3,4*] and Chia-Chi Ku[1*]

* Correspondence: lyliu@ntu.edu.tw; chiachiku@ntu.edu.tw
[3]Division of Biometry, Department of Agronomy, NTU, Taipei 10617, Taiwan, Republic of China
[1]Institute of Immunology, College of Medicine, National Taiwan University (NTU), 1 Jen-Ai Road Section 1, Taipei 10051, Taiwan, Republic of China
Full list of author information is available at the end of the article

## Abstract

**Background:** The pandemic threat of influenza has attracted great attention worldwide. To assist public health decision-makers, new suites of tools are needed to rapidly process and combine viral information retrieved from public-domain databases for a better risk assessment.

**Results:** Using our recently developed FluConvert and IniFlu software, we automatically processed and rearranged sequence data by standard viral nomenclature, determined the group-related consensus sequences, and identified group-specific polygenic signatures. The software possesses powerful ability to integrate viral, clinical, and epidemiological data. We demonstrated that both multiple basic amino acids at the cleavage site of the HA gene and also at least 11 more evidence-based viral amino acid substitutions present in global highly pathogenic avian influenza H5N2 viruses during the years 2009–2016 that are associated with viral virulence and human infection.

**Conclusions:** FluConvert and IniFlu are useful to monitor and assess all subtypes of influenza viruses with pandemic potential. These programs are implemented through command-line and user-friendly graphical interfaces, and identify molecular signatures with virological, epidemiological and clinical significance. FluConvert and IniFlu are available at https://apps.flutures.com or https://github.com/chinrur/FluConvert_IniFlu

**Keywords:** Highly pathogenic avian influenza viruses, H5N2, Viral and immunological informatics, Risk assessment, Pandemic potential

doi:10.6342/NTU202301433

## Background

The emergence of novel H5N1 avian influenza virus (AIV) in 1997 resulting in fatalities in humans has raised global concern [1]. As of May 8, 2020, a total of 861 human infections and 455 deaths caused by H5N1 infection had been reported [2]. Thereafter, the re-emergence of highly pathogenic avian influenza (HPAI) A H5Ny subtypes that cause widespread infections in poultry farms and in wild birds since 2003 has greatly attracted public health attention. Interestingly, the H5 AIVs in Asia have evolved faster, having higher viral diversity, greater inter-species transmission, and broader host range than those in Europe and the Americas [3]. Understanding the viral factors which determine the pathogenicity of H5 AIV by timely integration of virological, immunological and epidemiological information will be helpful to establish effective prevention and control measures to minimize future pandemic threats.

The immediate release of the genetic sequences of influenza A viruses combined with collections of tools established for analyzing all types and subtypes of influenza viral sequences have greatly advanced our understanding of the evolution of circulating viruses and their potential risk to animal and human health [4]. Given the fact that multiple mutations across gene segments of influenza viruses can exist and the genomic stability might be influenced by a particular mutation over time [5], new suites of tools are needed to integrate these databases for a better alignment of virological, epidemiological and clinical data in a real-time manner.

Several public-domain databases are available for collecting influenza genetic and epidemiological information. They include: (1) National Center for Biotechnology Information Influenza Virus Database (NCBI-IVD) [6], (2) Global Initiative on Sharing All Influenza Data (GISAID-EpiFlu) [7], and (3) Influenza Research Database (IRD) [8]. While NCBI-IVD provides the complete influenza viral sequences of gene segment across a wide range of years, GISAID-EpiFlu is recognized as a compelling mechanism for rapid sharing of partial or incomplete influenza viral sequences [9]. As for IRD, it contains human and mammalian influenza surveillance data as well as human clinical data associated with viruses, linking host surveillance data to well-characterized virus strains [8].

In this paper, we reported on development of a new suite of integrated software including FluConvert and IniFlu for data processing and analysis. FluConvert provides a series of automated packages to efficiently rearrange genetic data based on standard viral nomenclature [10] and translate the nucleotide sequences into three possible polypeptides from 0, + 1, and + 2 open reading frames (ORF) after performing simultaneous multiple sequence alignments. For IniFlu, it is programed to automatically select the correct ORF encoded from corresponding gene segment as well as the spliced isoforms (e.g. NS1, NS2 of NS gene; M1, M2 of M gene). Possible accessory proteins (e.g. PB1-N40, PB2-S1, M42) that have been reported in the literatures [11–13] can also be selected by IniFlu. The capability of IniFlu that integrates viral genetic information into clinical and epidemiological surveillance data with high efficiency provides a rapid comparison of variations in viral sequences with epidemiological significance. To this end, we provide the results from analysis of H5N2 HPAI viruses defined by the presence of the hallmark amino acid motif (XRRKRR) at the cleavage site between HA1 and HA2 domains [14]. In addition to these multiple basis amino acid residues in the HA, we demonstrate that several other amino acid substitutions across different gene segments

doi:10.6342/NTU202301433

of H5N2 avian influenza viruses could be associated with the viral virulence and mammalian infections based on IniFlu-generated polygenic HPAI consensus signature. This suggests that the data analysis platform we report here will be useful to identify novel mutations for risk assessment of AIVs with potential threat to animal and human health.

## Methods

### Installation of FluConvert and IniFlu

Both FluConvert and IniFlu are available for free download at https://apps.flutures.com or https://github.com/chinrur/FluConvert_IniFlu. The programs can be automatically installed in a desktop computer after the user perform execution files which is found in downloaded folders. The operating system of the computer requires Microsoft Windows 10 (version 1903 or later version) equipped with Microsoft Office 365 or Office Excel 2016 (or later version, 64-bit) for installation and Java 6 (or later, 64-bit) to perform each software. When open FluConvert, the user will be asked to import data which have been pre-downloaded from NCBI-IVD or GISAID-EpiFlu (Step 1 of Fig. 1)

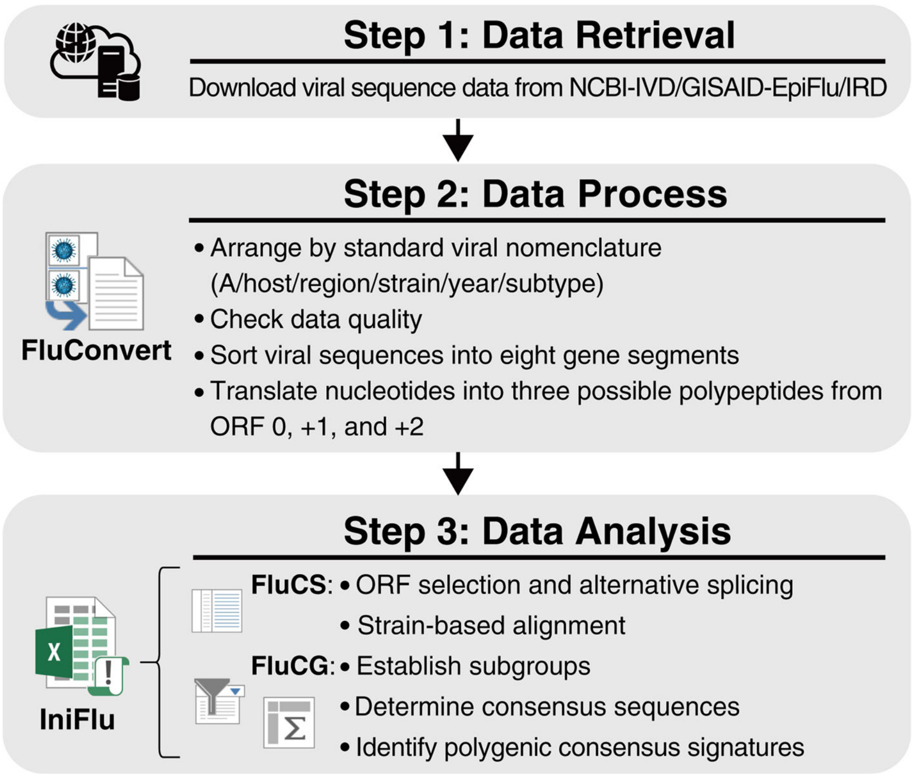

Step 1: Data Retrieval
Download viral sequence data from NCBI-IVD/GISAID-EpiFlu/IRD
Step 2: Data Process
FluConvert
• Arrange by standard viral nomenclature (A/host/region/strain/year/subtype)
• Check data quality
• Sort viral sequences into eight gene segments
• Translate nucleotides into three possible polypeptides from ORF 0, +1, and +2
Step 3: Data Analysis
IniFlu
FluCS: • ORF selection and alternative splicing
• Strain-based alignment
FluCG: • Establish subgroups
• Determine consensus sequences
• Identify polygenic consensus signatures

**Fig. 1** Workflows of data analysis executed by FluConvert and IniFlu. The stepwise processes performed by FluConvert and IniFlu to identify novel signatures of emerging influenza viruses with increasing risk are described as follows. Step 1: Viral sequences are obtained from the three databases (NCBI-IVD, GISAID-EpiFlu, and IRD). Step 2: FluConvert rearranges viral strains by viral nomenclature and ensures data quality. Viral sequences are further sorted into eight gene segments and translated into amino acid sequences. Step 3: The module FluCS of IniFlu performs strain-based alignments of FluConvert-processed viral amino sequences. The module FluCG of IniFlu regroups viral strains with epidemiological significant and computes a consensus sequence for each subgroup. Finally, the subgroup-specific unique polygenic amino acid signatures can be simultaneously identified (see details in Fig. 3)

doi:10.6342/NTU202301433

following instructions provided on the website. It will take minutes to hours to complete the process depending on the quantity of data entry. Once FluSeed Dataset has been processed by FluConvert (Step 2 of Fig. 1), IAV sequences can be analyzed by IniFlu (Step 3 of Fig. 1). We also include a detail step-by-step user's guide in Readme which can be found at https://apps.flutures.com website.

### FluConvert: a tool to process downloaded sequences

FluConvert automatically processes downloaded sequence files (*.FASTA) using the command-line interface (CLI) by batch (shell) scripts operated in a Microsoft Windows environment. It consequently performs (1) name and quality checking for downloaded sequences, (2) separation of sequences into eight gene segments, (3) multiple alignment of DNA sequences within clusters, (4) translation of DNA sequences into three possible polypeptides from ORF 0, + 1, and + 2, and (5) multiple alignment of amino acid sequences within clusters. The functions of FluConvert are to unify the arrangement of genetic information and then to convert nucleotide sequences of cDNA to amino acid sequences for multiple alignment at the protein level.

#### *Arrangement and quality checking for downloaded sequences*

All sequences downloaded from the NCBI-IVD and the GISAID-EpiFlu databases are rearranged according to the standard influenza viral nomenclature in the order of type, host, region, strain, year, and subtype within the parentheses [10]. Secondly, rearranged sequences are inspected, and the gene segments are deleted when they met any conditions in the "excluding list" generated for quality checking. Entries retrieved from NCBI-IVD and GISAID-EpiFlu databases were deleted according to "excluding list" to remove duplicates, incomplete sequences, or those with error information. Downloaded entries that are later saved to FluSeed Dataset have never been modified or corrected for any purposes. This is to ensure that the information remains original and the features of genetic sequences are kept unaltered during FluConvert processing. The three major error conditions of viral sequence information are: (1) lacking complete viral nomenclature, having mixed subtypes, belonging to lab strains or showing errata in public database records, (2) finding duplicate sequence records in any of the public databases, and (3) sequences longer than the expected lengths for different segments (e.g. PB1 > 2500 bp, PB2 > 2500 bp, PA > 2400 bp, HA > 1900 bp, NP > 1700 bp, NA > 1600 bp, M > 1150 bp, and NS > 1050 bp), or having redundant sequences or those containing more than 60 unknown nucleotides (denoted as 'n'). Finally, all the sequences that had passed the excluding list's quality check without entering the excluding list were used to create a new dataset called the "FluSeed Dataset" and subjected to IniFlu analysis.

As noted, entries retrieved from these public domain databases have never been modified or corrected after downloading. This ensures to keep information original and features of genetic sequences are not lost during FluConvert processing. Moreover, FluSeed Dataset is used for IniFlu analysis and has never been intended to make publicly accessible.

#### *Multiple sequence alignment and amino acid translation*

The genome of influenza A virus contains eight RNA segments. Therefore, FluConvert first divides the genetic sequences in FluSeed Database into eight clusters by the MAFF

doi:10.6342/NTU202301433

T multiple sequence alignment program (version 7.429) with fast Fourier transform [15]. All sequences in each of the eight gene segments are then translated into three possible polypeptides from ORF 0, + 1, and + 2 by EMBOSS Transeq (version 6.5) [16]. Nucleotide sequences and amino acid sequences in each of the eight gene segments in the FluSeed Database are subject to multiple alignment by MAFFT again with different optimizing parameters based on sequence lengths and the numbers of viral strains or files [i.e. L-INS-i (accurate) for alignment of <~200 viral strains/files; FFT-NS-2 (fast) for alignment of <~30,000 viral strains/files to obtain maximal efficiency; and PartTree (fast) for alignment of > ~ 30,000 viral strains/files] [17]. Results of sequence alignments from the same ORF were saved as comma-delimited (csv) text files.

### IniFlu: a viral information viewer and analyzer

IniFlu, a Visual Basic Application (VBA) program for Microsoft Office Excel 2019 worksheet, has a user-friendly graphical interface (GUI) to combine viral information, amino-acid sequences and epidemiological data for further analyses. IniFlu has two modules, "FluCS" (which stands for "Flu Cross-Segment alignment") and "FluCG" (which stands for "Flu Comparative Grouping"). FluCS matches the aligned sequences according to the standard viral nomenclature of the strains after encoding protein from ORF selection and alternative splicing. FluCG visualizes different epidemiologically specific (such as time-, area-, host-, age-, gender-specific) consensus signatures obtained (shown in Fig. 4), providing not only the clinical information of the viral sequences but also their epidemiological characteristics.

#### *FluCS: strain-based amino acid sequence alignment*

FluCS groups the amino-acid sequences of each gene segment according to FluSeed Dataset (Fig. 2a). FluCS is also programed to automatically select the correct ORF of each viral protein as well as the alternatively spliced isoforms. Accessory proteins, e.g. PB1-N40, PB2-S1 and M42 [11–13] can be assigned to the viral segment group of PB1, PB2, and M, respectively. PB1-F2 which is translated by a second ORF in the + 1 frame [11, 18] is selected and assigned to an independent group. As a result, a total of 11 viral segment groups (PB1, PB2, PA, HA, NA, NP, M1, M2, NS1, NS2, and PB1-F2) is established for strain-based alignment (Fig. 2b, c). The position of each residue is numbered based on the first methionine residue of that gene segment determined by FluCS (e.g. HA of H5N2 subtype is numbered by H5 numbering system) [19].

#### *FluCG: comparative sequence analysis*

FluCG chooses the most representative amino acid at each residue of a particular gene segment by computing the most frequent amino acid among all strains within the studied subgroup. If there are more than two amino acids occurring at the same frequency, one is chosen by alphabetic order. Residues that appear at stop codons or deleted codons are marked. Through this process, the consensus sequence can be created for a particular subgroup [20]. The unique residues in each subgroup can also be identified by aligning two consensus sequences and are called "consensus signatures". All the unique amino acids appearing at each residue of the consensus signature are thoroughly examined and compared to verify the unique amino acid is present only in the

doi:10.6342/NTU202301433

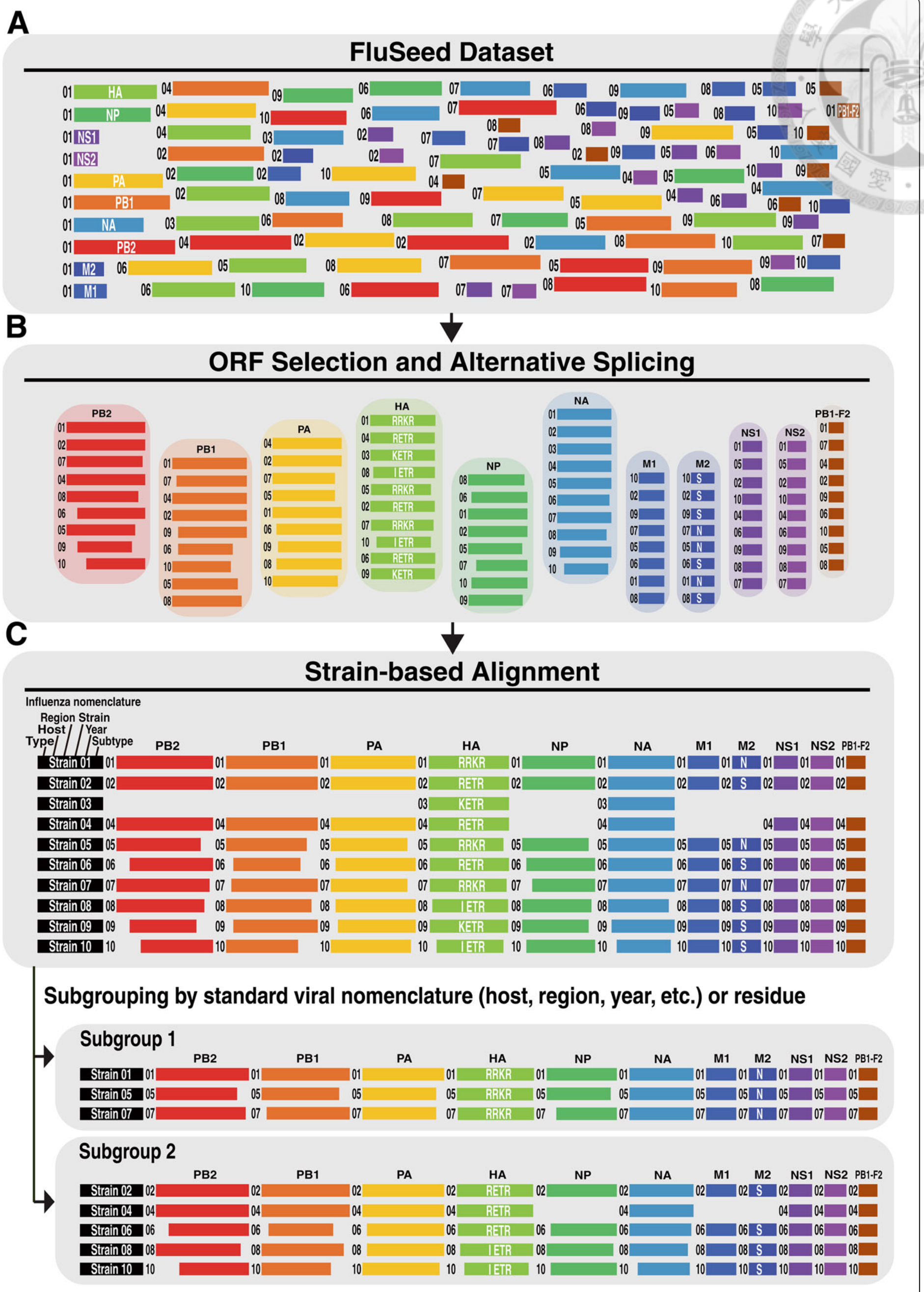


**Fig. 2** Schematic diagram of strain-based alignment approach. Influenza viral sequences are aligned by FluCS as follows: **a** The FluSeed Dataset is constructed by quality-checked and rearranged viral sequences. Blocks in different colors represent ten viral segments. The size of each block corresponds to the length of the viral sequence originally retrieved. Blocks in any color tagged with the same Arabic numbers are identified as the same strain. **b** Rearranged viral sequences are sorted into 11 protein clusters based on gene segments and well aligned within the cluster. Aligned sequences are subjected correct ORF into PB2, PB1, PA, HA, NP, NA, M1, NS1, and PB1-F2. M2 and NS2 are alternatively spliced proteins from M and NS ORF mRNAs, and respectively. **c** Delineated viral amino acid sequences are easily aligned based on standard influenza viral nomenclature. The analysis platform provides benefits for multi-layer subgrouping based on epidemiological significance

particular subgroup. Finally, all possible 20 amino acids, stop codons, and deletions are all examined and presented in a substitution table (as Fig. 5).

## Results

### Influenza viral sequences and data processing

The genetic sequences and epidemiological information of influenza viruses in one public domain database are not properly linked to the other. To maximize the information coverage for a particular AIV subtype for further analysis, we have developed the FluConvert program to combine all data available from these databases and automatically process them in one format. Viral sequences that had passed quality check after excluding incomplete or erroneous ones to ensure correct genetic information are used for constituting the FluSeed Dataset. Sequences in the FluSeed are subsequently rearranged to standard nomenclature in the order of influenza virus of type/host/region/strain/year (HxNy subtype) and segregated into eight gene segments. Amino acid sequences are translated from nucleotide sequences for alignment (Fig. 1).

### Viral strain-based sequence alignment

Continuous mutations in HPAI A (H5) viruses have been attributed to outbreaks at poultry farms and sporadic human infections [21]. Since mutations can occur across several gene segments in the genome of AIVs, multiple alignments for all viral strains based on the subgroup of interest (i.e. host, region, year, a particular residue, etc.) rather than by gene segment (i.e. HA, NA, PB2, etc.) will be useful to identify multiple amino acid types associated with viral pathogenicity in animals or the potential risk for human infections. To achieve the goal, we have developed the IniFlu platform to integrate the processed viral sequences, clinical and epidemiological information into the FluSeed Dataset. IniFlu can function to present all the information imported from FluSeed in worksheet outputs for visual cross-segment examinations simultaneously. Once all of the viral strain information is correctly arranged by FluCS, strain-based alignment can be quickly performed as illustrated in Fig. 2.

### Identification of polygenic consensus signatures

Genetic evolution of zoonotic influenza viruses is a polygenic trait. Amino acid substitutions or mutations at species-associated signature positions may increase viral pathogenicity or mammalian adaptation in a broader host range [22]. Since such mutations are not limited to one gene and can simultaneously occur in multiple gene segments, identification of the polygenic consensus signatures for a particular subgroup of viral strains offers an opportunity to monitor the changing landscape of AIVs over time with epidemiological significance. The module FluCG of IniFlu can quickly group viral strains into different subgroups and deduce the consensus sequence of each subgroup by computing and determining the most representative (i.e. most frequent) amino acid at each position of the whole genome, which can differentiate between the compared subgroups. All unique amino acid residues represented in the subgroup constitute the polygenic consensus signature (Fig. 3).

### Polygenic consensus signatures of the HPAI H5N2 viruses

Duplicate entries of downloaded influenza viral genetic sequences could possibly occur when (1) the entry was submitted to both NCBI-IVD and GISAID-EpiFlu databases, (2)

doi:10.6342/NTU202301433

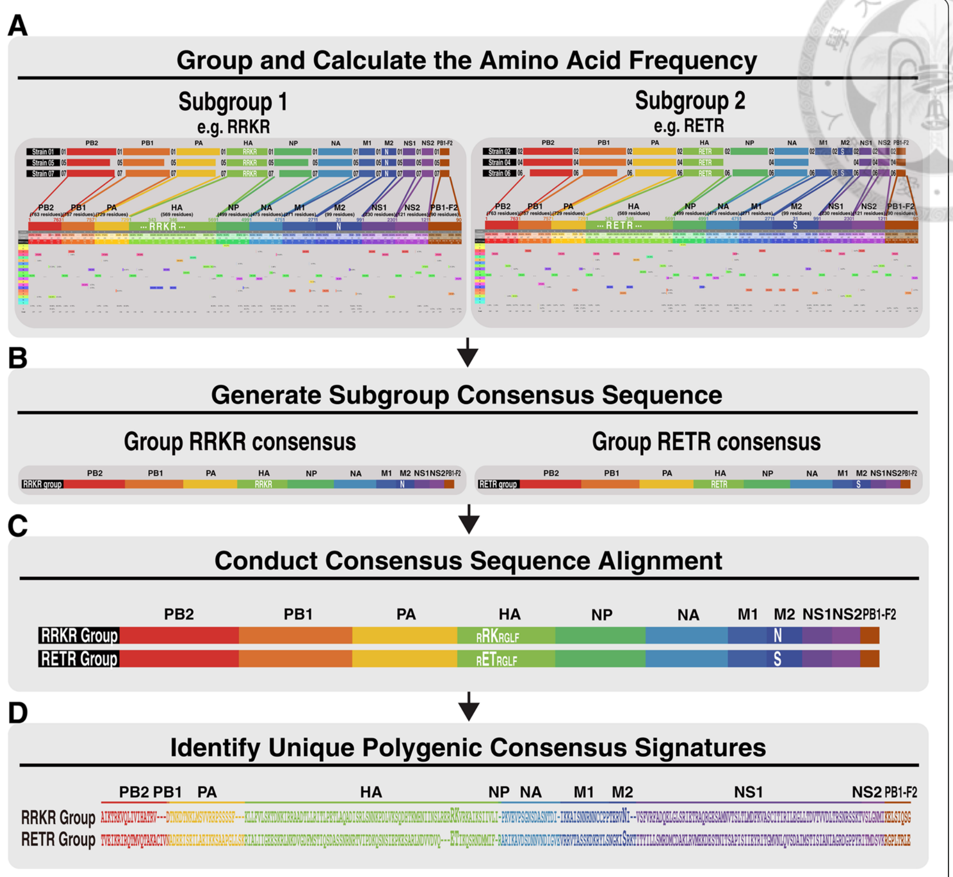


**Fig. 3** Identification of group-specific polygenic consensus signatures by FluCG. The polygenic consensus signatures are determined as follows: **a** FluCG groups viral strains into different subgroups (e.g. RETR and RRKR groups). The consensus sequence of each subgroup is determined by computing a table of 20 amino-acid substitutions to choose the most representative amino acid at each position of the whole genome. The stop codon is denoted as "X" and the deleted residue is denoted as "-". **b** One consensus sequence from each subgroup is determined. **c** FluCG aligns two group consensus sequences and identifies the signature that is unique, to distinguish between the two consensus sequences. **d** Text in different colors represents different proteins that compose the polygenic consensus signatures

the entry was submitted to one database more than once, or (3) data entries imported from NCBI-IVD co-existed in GISAID-EpiFlu database. To obtain the accurate count of H5N2 virus strains downloaded from different public domain databases, FluConvert is programmed to automatically remove duplicate entries. Only one copy of the gene segment of a single virus strain is kept in FluSeed Dataset.

As of July 1, 2017, the H5N2 FluSeed Dataset was comprised of a total of 6746 (6443 + 303 = 6746) unique gene segments that belong to 1151 (1099 + 52 = 1151) H5N2 viruses. Amongst which, 6443 gene segments of 1099 H5N2 strains were downloaded from NCBI-IVD and 303 segments of 52 H5N2 strains were downloaded from GISAID-EpiFlu, respectively. Qualified genetic sequences were rearranged by FluConvert to unify the nomenclature format. Corresponding epidemiological information and clinical data for each strain were integrated through the IniFlu platform. Since several studies have demonstrated that the presence of multiple basic amino acids at the cleavage site between HA1 and HA2 junctional sequence is a hallmark for increasing viral pathogenicity and virulence in the avian host and humans [14], we compared the molecular signature in the H5N2 viral strains with (RRKR group) or without (RETR group) polybasic residues at the cleavage site in the HA gene. The earliest record of H5N2 viruses was reported in 1972 and all of the 470 strains isolated during 1972–2008 appeared to have RETR sequence motif at the HA cleavage site. H5N2 viruses with the RRKR sequence motif in the HA only appeared after year 2009. To avoid bias towards evolutionary perspective, we excluded H5N2 viruses that were isolated before 2008 and only kept the H5N2 viruses isolated from year 2009 to 2016 in the H5N2 Flu-Seed Dataset for consensus signature analysis of both groups. As a result, there were 165 strains of H5N2 viruses with RETR marker and 138 strains with RRKR marker.

The consensus sequence analysis by FluCG identified 247 unique amino acid residues differentially presented between RRKR and RETR groups in the whole genome of H5N2 AIVs. Since these unique residues were present across several viral segments, we wanted to know which gene segment might present the most unique residues that may distinguish H5N2 viruses with the REKR marker from those RRKR. Table 1 shows the frequencies of the characteristic substitutions that occurred at a particular gene segment. We found that NS1 had the highest substitutions ($N = 69$, 30%), followed by HA ($N = 77$, 13.53%), and PB1-F2 ($N = 8$, 8.89%). There were much less substitutions in NP ($N = 1$, 0.2%) and PB1 (N = 1, 0.13%) of H5N2 viruses (Table 1).

To investigate what substitution at a particular residue or residues could be associated with the RRKR phenotype, the polygenic consensus signatures determined from the constellation of the 247 distinct residues as described in Table 1 were further analyzed (Fig. 4). In search of information on amino acid substitutions in the influenza viruses that are associated with increased viral virulence or drug resistance [23] reported in the public domain database IRD-SFVT (Sequence Feature Variant Types) by IniFlu analysis, we found that substitutions in HA, including T124I, D142E, E228K, P233S, V336S in HA, G631S in PA that are related to increasing pathogenicity [24–26] were present in the RRKR signature. Other variations in the HA of the RRKR signature involved in the increase of α-2.6 receptor binding in mammalian cells such as S139P, S145L, S149A, and I226V [27–30] were also found in our analysis. Notably, the fact that IniFlu identified the substitution of S31N in the M2 of the RRKR signature suggests that H5N2 HPAI may have a decreased sensitivity to amantadine and rimantadine [31] (Fig. 4). All of the unique 11 consensus signatures were re-examined and verified from FluCG-generated substitution table (Fig. 5). Taken together, IniFlu can identify additional substitutions across the gene segments of H5N2 that are highly associated with viral pathogenicity and/or antiviral drug resistance.

**Table 1** The 247 residues differentially occurring between RRKR and RETR consensus signatures are polygenic

| | Influenza viral proteins | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | PB2 | PB1 | PA | HA | NP | NA | M1 | M2 | NS1 | NS2 | PB1-F2 |
| **Segment size**[a] | 763 | 757 | 729 | 569 | 499 | 475 | 271 | 99 | 230 | 121 | 90 |
| **No. of consensus signatures between groups**[b] **(%)** | 20 (2.62) | 1 (0.13) | 23 (3.16) | 77 (13.53) | 1 (0.2) | 18 (3.79) | 15 (5.54) | 8 (8.08) | 69 (30) | 7 (5.79) | 8 (8.89) |

[a]: The H5N2 viral genome is composed of 4603 amino acid residues divided into 11 viral proteins. The size of each segment is indicated by the number of residues as shown

[b]: Unique amino acid residues are identified by comparing the consensus sequences between the two studied subgroups. Numbers shown are the counts of the characteristic residues in each viral protein. The variations in each viral protein are expressed by the percentage of unique residues indicated in the parentheses

doi:10.6342/NTU202301433

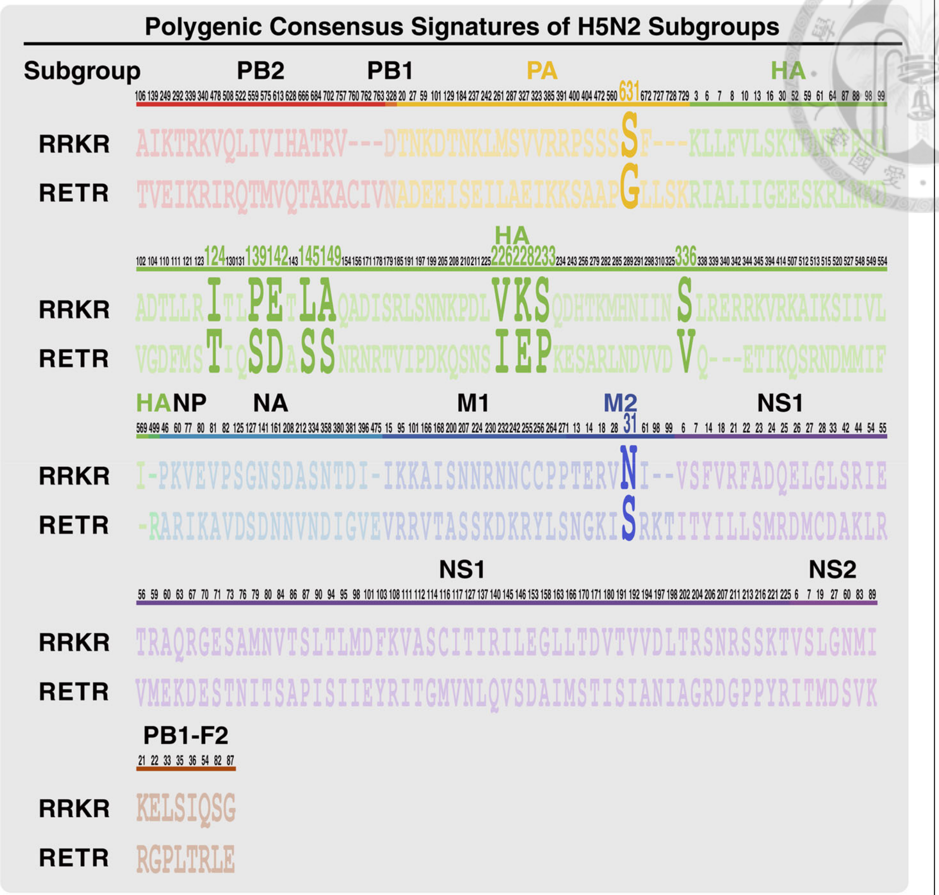


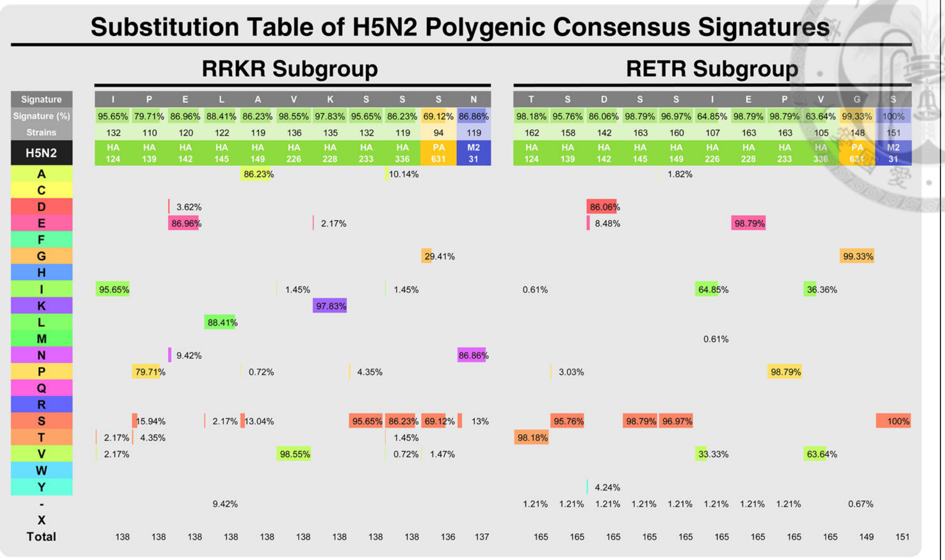
**Fig. 4** Polygenic consensus signatures of H5N2 RRKR and RETR subgroups. A total of 247 residues across 11 protein segments of H5N2 AIVs obtained from the years 2009 to 2016 are identified from the differences between the RRKR and RETR subgroups. Polygenic consensus signatures for each group are shown by the constellation of the 247 unique amino acid residues across different protein segments (distinguished by colors). The residue positions at each protein are shown as numbers. Of the 247 residues in the consensus signature, 11 evidence-based residues documented in the literature are listed in the enlarged letters with darkened colors

## Discussion

Influenza is an important disease in humans and animals. The 13,588-base-pair RNA genome segregated into eight gene segments continues to mutate randomly at 2 × $10^{-6}$ mutations per site per infectious cycle [32]. The high activity in the reassortment of segmented influenza viral genes derived from different host species has posed a great threat to public health. Numerous tools have been developed to analyze influenza genetic sequences to monitor the changes and evolution of these viruses over time in nature. In this study, we have added two integrated analysis tools, FluConvert and IniFlu, to the endeavor.

Several analysis tools for IAV genetic sequences are available online to determine antigenic characteristics of IAVs based on the genomic sequences of a particular gene segment and associated epidemiological information. Here we compare a recently published program FluPhenotype [33] with IniFlu. FluPhenotype is a web-based tool. Briefly, IAVs amino acid markers associated with human adaptation, enhanced virulence, and drug resistance, etc. that have been reported in the literatures are captured to the Data list of FluPhenotype. The input genetic sequences of IAVs are mapped with the list and the

doi:10.6342/NTU202301433



Substitution Table of H5N2 Polygenic Consensus Signatures

RRKR Subgroup

| Signature | I | P | E | L | A | V | K | S | S | S | N |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Signature (%) | 95.65% | 79.71% | 86.96% | 88.41% | 86.23% | 98.55% | 97.83% | 95.65% | 86.23% | 69.12% | 86.86% |
| Strains | 132 | 110 | 120 | 122 | 119 | 136 | 135 | 132 | 119 | 94 | 119 |
| H5N2 | HA 124 | HA 139 | HA 142 | HA 145 | HA 149 | HA 226 | HA 228 | HA 233 | HA 336 | PA 631 | M2 31 |
| A | | | | | 86.23% | | | | 10.14% | | |
| C | | | | | | | | | | | |
| D | | | 3.62% | | | | | | | | |
| E | | | 86.96% | | | | 2.17% | | | | |
| F | | | | | | | | | | | |
| G | | | | | | | | | | 29.41% | |
| H | | | | | | | | | | | |
| I | 95.65% | | | | | 1.45% | | | 1.45% | | |
| K | | | | | | | 97.83% | | | | |
| L | | | | 88.41% | | | | | | | |
| M | | | | | | | | | | | |
| N | | | 9.42% | | | | | | | | 86.86% |
| P | | 79.71% | | | 0.72% | | | 4.35% | | | |
| Q | | | | | | | | | | | |
| R | | | | | | | | | | | |
| S | | 15.94% | | 2.17% | 13.04% | | | 95.65% | 86.23% | 69.12% | 13% |
| T | 2.17% | 4.35% | | | | | | | 1.45% | | |
| V | 2.17% | | | | | 98.55% | | | 0.72% | 1.47% | |
| W | | | | | | | | | | | |
| Y | | | | | | | | | | | |
| - | | | | 9.42% | | | | | | | |
| X | | | | | | | | | | | |
| Total | 138 | 138 | 138 | 138 | 138 | 138 | 138 | 138 | 138 | 136 | 137 |

RETR Subgroup

| Signature | T | S | D | S | S | I | E | P | V | G | S |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Signature (%) | 98.18% | 95.76% | 86.06% | 98.79% | 96.97% | 64.85% | 98.79% | 98.79% | 63.64% | 99.33% | 100% |
| Strains | 162 | 158 | 142 | 163 | 160 | 107 | 163 | 163 | 105 | 148 | 151 |
| H5N2 | HA 124 | HA 139 | HA 142 | HA 145 | HA 149 | HA 226 | HA 228 | HA 233 | HA 336 | PA 631 | M2 31 |
| A | | | | | 1.82% | | | | | | |
| C | | | | | | | | | | | |
| D | | | 86.06% | | | | | | | | |
| E | | | 8.48% | | | | 98.79% | | | | |
| F | | | | | | | | | | | |
| G | | | | | | | | | | 99.33% | |
| H | | | | | | | | | | | |
| I | 0.61% | | | | | 64.85% | | | 36.36% | | |
| K | | | | | | | | | | | |
| L | | | | | | | | | | | |
| M | | | | | | 0.61% | | | | | |
| N | | | | | | | | | | | |
| P | | 3.03% | | | | | | 98.79% | | | |
| Q | | | | | | | | | | | |
| R | | | | | | | | | | | |
| S | | 95.76% | | 98.79% | 96.97% | | | | | | 100% |
| T | 98.18% | | | | | | | | | | |
| V | | | | | | 33.33% | | | 63.64% | | |
| W | | | | | | | | | | | |
| Y | | | 4.24% | | | | | | | | |
| - | 1.21% | 1.21% | 1.21% | 1.21% | 1.21% | 1.21% | 1.21% | 1.21% | | 0.67% | |
| X | | | | | | | | | | | |
| Total | 165 | 165 | 165 | 165 | 165 | 165 | 165 | 165 | 165 | 149 | 151 |

**Fig. 5** Establishment of polygenic consensus signatures of RRKR and RETR groups of H5N2 AIVs. To further confirm that each residue in the RRKR signature is unique compared with that in the RETR signature, the percentage of the most predominant amino acid (top two rows) at the corresponding position (the 3rd row) derived from the 11 evidence-based residues were re-examined by inspecting the FluCG-generated worksheet. As is shown, the percentages of the predominant amino acids in each signature range from greater than 69.12 to 95.65%

antigenic characteristics of the IAVs of interest are rapidly determined. FluPhenotype also has the capacity to predict IAV HA subtype and viral hosts based on the input genomic or protein sequences [33]. Although the Data list used in FluPhenotype is reportedly updated every half a year, any newly identified or undefined molecular markers that have not been made available in the literature would not be captured and mapped in a timely manner [33].

In comparison with FluPhenotype, FluConvert is used to sort IAV genetic entries that are downloaded from different public domain databases into eight gene segments based on the name of the gene segment (e.g. PB2, PB1, PA, … etc.). FluConvert subsequently rearranges the information tagged to each entry according to the standard IAV nomenclature in the order of type, host, region, strain, year, and subtype, thereby assigning a unique name to each virus. Therefore, gene segments that have the same name will be grouped as one strain. The capability of FluConvert that determines the correct protein sequences encoded by each viral gene segment and their spliced isoforms as well as accessory proteins results in 11 viral protein clusters in the FluSeed Dataset for strain-based alignment by FluCS.

Since FluCS can align a larger number of viral strains at one time, it saves time on cross-referring of each genetic sequence in NCBI-IVD/GISAID-EpiFlu by accession number. Additionally, the ability of FluConvert to combine information between databases can collect all available influenza genetic data as much as possible by avoiding the exclusion from incomplete information in the depository database. Data in the FluSeed Dataset can be maintained up to date by downloading newly depository of influenza viral genetic data in public domain databases by users.

There are two advantages of IniFlu-performed strain-based alignment and consensus sequence analysis. First, genetic sequences of a viral strain lacking eight complete gene segments can be compared and included for consensus sequence analysis. Second, once the information is properly aligned, sequence data can be easily re-grouped for a generating group-specific consensus sequence. As a result, polygenic consensus signatures composed of unique molecular positions across all gene segments that are associated with a particular phenotype will be determined. As demonstrated from the analysis of the sample H5N2 FluSeed Dataset by comparing the group-specific polygenic consensus signatures between the RRKR and the RETR groups, we identified that at least 247 positions of the total 303 H5N2 AIV strains from 2009 to 2016 were able to differentiate these two groups, and 11 of these substitutions have been experimentally demonstrated for the significance in crossing over between host species (e.g. S139P, S145L, S149A, and I226V in HA) [27–30], antiviral drug amantadine and rimantadine resistance (S31N in M2) [31] or increasing viral pathogenesis (e.g. T124I, D142E, E228K, P233S, and V336S in HA, and G631S in PA) [24–26]. Although there have not been reports of fatal human cases of H5N2, human infection of this AIV subtype have occurred, as documented in seroepidemiological studies [34, 35]. These substitutions together with those residues involved in enhancing receptor binding to mammalian cells [14] have suggested the potential threat to human health caused by H5N2 AIV strains with an RRKR phenotype.

Taken together, we reported the newly developed analysis tools FluConvert and IniFlu, which exhibit high capacity and efficiency in data processing, analyzing, and combining large amounts of the most comprehensive influenza viral information retrieved from different public domain databases without making any modifications on downloaded genetic information. These tools not only provide a versatile and rapid platform for real-time analysis to determine consensus sequences, but also identify molecular markers with high pathogenicity in chickens as well as with interspecies transmission to humans. FluConvert and IniFlu are particularly useful in risk assessment by monitoring and analyzing the increasing trends of important amino acids of many animal influenza viruses with pandemic potential. While IniFlu is first designed for type A influenza viruses, the software can easily adapt to investigate other emerging viruses with appropriate modifications on the worksheet template. The software reported in this study provides a useful tool for rapidly identifying molecular signatures with virological, epidemiological and clinical significance.

## Conclusions

The rapid evolution of H5 AIVs in Asia has increased the threat in agricultural safety and human health. The timely monitoring in the changes of AIV that have increasing risk are important for public health-policy makers. FluConvert and IniFlu reported in this study are demonstrated for their efficiency in combining and analyzing virological, epidemiological and clinical information from different public domain databases. Finally, identification of polygenic signature for AIVs with high risk instead of variations at one single gene segment of influenza viruses will be beneficial to assist a better risk assessment to prevent pandemic influenza.

## Availability and requirements

**Project name**: FluConvert_IniFlu

**Project home page**: https://apps.flutures.com or https://github.com/chinrur/FluConvert_IniFlu

**Operating system(s)**: Microsoft Windows 10 or later version (64-bit)

**Programming language**: Batch (shell) scripts and VBA 7.1

**Other requirements**: Microsoft Office Excel 365 or Excel 2016 or later version (64-bit); Java 6 or higher version

**License**: MIT License.

**Any restrictions to use by non-academics**: No restrictions on use by non-academics.

### Abbreviations

AIV: avian influenza virus; CLI: command-line interface; EMBOSS: The European Molecular Biology Open Software Suite; FluCG: Flu Comparative Grouping; FluCS: Flu Cross-Segment alignment; GISAID: Global Initiative on Sharing All Influenza Data; GUI: graphical interface; HPAI: highly pathogenic avian influenza; IAV: influenza A virus; IRD: Influenza Research Database; IVD: Influenza Virus Database; MAFFT: multiple sequence alignment program with fast Fourier transform; NCBI: National Center for Biotechnology Information; ORF: open-reading-frame; VBA: Visual Basic Application; WHO: World Health Organization

### Acknowledgements

We express our special thanks to Dr. Hsiang-Hsuan Sung at the Institute of Microbiology and Immunology, College of Medicine, National Defense Medical Center (NDMC), Taiwan for his insights in initiating the project.

### Authors' contributions

CRY created the software, carried out the analysis and drafted the manuscript; CCK designed epidemiological study; LYL validated the software and substantially revised the writing in Methods; CCK interpreted the data and was a major contributor in writing the manuscript. All authors read and approved the manuscript.

### Funding

CRY was the recipient of the fellowship from the Sustainability Science Research Program, Academia Sinica, Taiwan (grant no. AS-SS-106-04-4). Grants from the Ministry of Science and Technology, Taiwan (MOST 107–2320-B-002-008;108–2320-B-002-038) supported research-related logistic expenses and publication fees. None of the funding body plays any role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

### Availability of data and materials

The H5N2 Dataset generated and analyzed during the current study are available in the NCBI-IVD (https://www.ncbi.nlm.nih.gov/genomes/FLU/) and GISAID-EpiFlu databases (https://www.gisaid.org/). These analysis tools are not intended for use in public domain but only for processing data retrieved from public databases. Their use would not breach the data access agreement of GISAID. Abiding by GISAID-EpiFlu Database Access Agreement, these tools will not generate new database for public access or make any annotation, correction, or modification of data submitted to GIASID EpiFlu Database.

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

### Author details

[1]Institute of Immunology, College of Medicine, National Taiwan University (NTU), 1 Jen-Ai Road Section 1, Taipei 10051, Taiwan, Republic of China. [2]Institute of Epidemiology and Preventive Medicine, College of Public Health, NTU, Taipei 10055, Taiwan, Republic of China. [3]Division of Biometry, Department of Agronomy, NTU, Taipei 10617, Taiwan, Republic of China. [4]Department of Agronomy, National Taiwan University, No. 1, Section 4, Roosevelt Rd, Taipei 10617, Taiwan.

Received: 10 April 2020 Accepted: 9 July 2020
Published online: 18 July 2020

### References

1. Chan PKS. Outbreak of Avian Influenza A(H5N1) Virus Infection in Hong Kong in 1997. Clin Infectious Diseases. 2002; 34(Supplement_2):S58–64.
2. Cumulative number of confirmed human cases of avian influenza A(H5N1) reported to WHO [https://www.who.int/influenza/human_animal_interface/H5N1_cumulative_table_archives/en/]. Accessed 8 May 2020.
3. Neumann G, Chen H, Gao GF, Shu Y, Kawaoka Y. H5N1 influenza viruses: outbreaks and biological properties. Cell Res. 2010;20(1):51–61.
4. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, Sagulenko P, Bedford T, Neher RA. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics. 2018;34(23):4121–3.

5. Arai Y, Kawashita N, Hotta K, Hoang PVM, Nguyen HLK, Nguyen TC, Vuong CD, Le TT, Le MTQ, Soda K, et al. Multiple polymerase gene mutations for human adaptation occurring in Asian H5N1 influenza virus clinical isolates. Sci Rep. 2018;8(1):13066.
6. Bao Y, Bolotov P, Dernovoy D, Kiryutin B, Zaslavsky L, Tatusova T, Ostell J, Lipman D. The influenza virus resource at the National Center for biotechnology information. J Virol. 2008;82(2):596–601.
7. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data - from vision to reality. Euro Surveill. 2017; 22(13):30494.
8. Zhang Y, Aevermann BD, Anderson TK, Burke DF, Dauphin G, Gu Z, He S, Kumar S, Larsen CN, Lee AJ, et al. Influenza research database: an integrated bioinformatics resource for influenza virus research. Nucleic Acids Res. 2016;45(D1):D466–74.
9. Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. Global Chall. 2017;1(1):33–46.
10. WHO. A revision of the system of nomenclature for influenza viruses: a WHO Memorandum. Bull World Health Organization. 1980;58(4):585–91.
11. Wise HM, Foeglein A, Sun J, Dalton RM, Patel S, Howard W, Anderson EC, Barclay WS, Digard P. A complicated message: identification of a novel PB1-related protein translated from influenza a virus segment 2 mRNA. J Virol. 2009;83(16):8021–31.
12. Yamayoshi S, Watanabe M, Goto H, Kawaoka Y. Identification of a novel viral protein expressed from the PB2 segment of influenza a virus. J Virol. 2016;90(1):444–56.
13. Wise HM, Hutchinson EC, Jagger BW, Stuart AD, Kang ZH, Robb N, Schwartzman LM, Kash JC, Fodor E, Firth AE, et al. Identification of a novel splice variant form of the influenza a virus M2 ion channel with an antigenically distinct ectodomain. PLoS Pathog. 2012;8(11):e1002998.
14. Alexander DJ. A review of avian influenza in different bird species. Vet Microbiol. 2000;74(1–2):3–13.
15. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772–80.
16. Rice P, Longden I, Bleasby A. EMBOSS: the European molecular biology open software suite. Trends Genet. 2000;16(6):276–7.
17. Katoh K, Toh H. PartTree: an algorithm to build an approximate tree from a large number of unaligned sequences. Bioinformatics. 2007;23(3):372–4.
18. Chen W, Calvo PA, Malide D, Gibbs J, Schubert U, Bacik I, Basta S, O'Neill R, Schickli J, Palese P, et al. A novel influenza a virus mitochondrial protein that induces cell death. Nat Med. 2001;7(12):1306–12.
19. Burke DF, Smith DJ. A recommended numbering scheme for influenza a HA subtypes. PLoS One. 2014;9(11):e112302.
20. Waterman MS, Arratia R, Galas DJ. Pattern recognition in several sequences: consensus and alignment. Bull Math Biol. 1984;46(4):515–27.
21. To KK, Ng KH, Que TL, Chan JM, Tsang KY, Tsang AK, Chen H, Yuen KY. Avian influenza a H5N1 virus: a continuous threat to humans. Emerg Microbes Infect. 2012;1(9):e25.
22. Hatta M, Gao P, Halfmann P, Kawaoka Y. Molecular basis for high virulence of Hong Kong H5N1 influenza a viruses. Science. 2001;293(5536):1840–2.
23. Noronha JM, Liu M, Squires RB, Pickett BE, Hale BG, Air GM, Galloway SE, Takimoto T, Schmolke M, Hunt V, et al. Influenza virus sequence feature variant type analysis: evidence of a role for NS1 in influenza virus host range restriction. J Virol. 2012;86(10):5857–66.
24. Hulse DJ, Webster RG, Russell RJ, Perez DR. Molecular determinants within the surface proteins involved in the pathogenicity of H5N1 influenza viruses in chickens. J Virol. 2004;78(18):9954–64.
25. Gohrbandt S, Veits J, Hundt J, Bogs J, Breithaupt A, Teifke JP, Weber S, Mettenleiter TC, Stech J. Amino acids adjacent to the haemagglutinin cleavage site are relevant for virulence of avian influenza viruses of subtype H5. J Gen Virol. 2011;92(1):51–9.
26. Hiromoto Y, Saito T, Lindstrom S, Nerome K. Characterization of low virulent strains of highly pathogenic a/Hong Kong/ 156/97 (H5N1) virus in mice after passage in embryonated hens' eggs. Virology. 2000;272(2):429–37.
27. Yamada S, Suzuki Y, Suzuki T, Le MQ, Nidom CA, Sakai-Tagawa Y, Muramoto Y, Ito M, Kiso M, Horimoto T, et al. Haemagglutinin mutations responsible for the binding of H5N1 influenza a viruses to human-type receptors. Nature. 2006;444(7117):378–82.
28. Auewarakul P, Suptawiwat O, Kongchanagul A, Sangma C, Suzuki Y, Ungchusak K, Louisirirotchanakul S, Lerdsamran H, Pooruk P, Thitithanyanont A, et al. An avian influenza H5N1 virus that binds to a human-type receptor. J Virol. 2007; 81(18):9950–5.
29. Yang Z-Y, Wei C-J, Kong W-P, Wu L, Xu L, Smith DF, Nabel GJ. Immunization by avian H5 influenza Hemagglutinin mutants with altered receptor binding specificity. Science. 2007;317(5839):825–8.
30. Watanabe Y, Ibrahim MS, Ellakany HF, Kawashita N, Mizuike R, Hiramatsu H, Sriwilaijaroen N, Takagi T, Suzuki Y, Ikuta K. Acquisition of human-type receptor binding specificity by new H5N1 influenza virus sublineages during their emergence in birds in Egypt. PLoS Pathog. 2011;7(5):e1002068.
31. Bean WJ, Threlkeld SC, Webster RG. Biologic potential of amantadine-resistant influenza a virus in an avian model. J Infect Dis. 1989;159(6):1050–6.
32. Nobusawa E, Sato K. Comparison of the mutation rates of human influenza a and B viruses. J Virol. 2006;80(7):3675–8.
33. Lu C, Cai Z, Zou Y, Zhang Z, Chen W, Deng L, Du X, Wu A, Yang L, Wang D, et al. FluPhenotype-a one-stop platform for early warnings of the influenza a virus. Bioinformatics. 2020;36(10):3251–3.
34. Wu H-S, Yang J-R, Liu M-T, Yang C-H, Cheng M-C, Chang F-Y. Influenza a(H5N2) virus antibodies in humans after contact with infected poultry, Taiwan, 2012. Emerg Infect Dis. 2014;20(5):857–60.
35. Liu MD, Chan TC, Wan CH, Lin HP, Tung TH, Hu FC, King CC. Changing risk awareness and personal protection measures for low to high pathogenic avian influenza in live-poultry markets in Taiwan, 2007 to 2012. BMC Infect Dis. 2015;15:241.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.