Master Thesis

Development of Galaxy Workflows for Sequence Data Analysis of Notifiable Viral Livestock Diseases

Viktoria Isabel Schwarz

Examiner:	Prof. Dr. Rolf Backofen
Second Examiner:	Prof. Dr. Wolfgang Hess
Advisor:	Dr. Wolfgang Maier

Albert-Ludwigs-University Freiburg

Faculty of Engineering

Department of Computer Science

Bioinformatics Group

April 28th, 2023

Writing Period

October 28th, 2022 – April 28th, 2023

Examiner

Prof. Dr. Rolf Backofen

Second Examiner

Prof. Dr. Wolfgang Hess

Advisor

Dr. Wolfgang Maier

Declaration

I hereby declare that I am the sole author and composer of my thesis and that no other sources or learning aids other than those listed have been used. Furthermore, I declare that I have acknowledged the work of others by providing detailed references of said work.
I hereby also declare that my thesis has not been prepared for another examination or assignment, either wholly or excerpts thereof.

Place, Date	Signature

Acknowledgements

Abstract

auf englisch

Zusammenfassung

auf deutsch

Contents

List of Figures

List of Tables

1 Introduction

Sharing environments means sharing diseases – this simple relationship expresses how pathogens found in animal populations can spread to humans and have severe impacts. The impact can be as severe as the whole world experienced during the pandemic of Coronavirus Disease 19 (COVID-19) that originated in Wuhan, China in 2019. This highly contagious disease was caused by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), an infectious virus of presumed zoonotic origin [1]. With more than 757.26 million reported cases and more than 6.85 million confirmed deaths as of February 24, 2023 **TODO: update numbers**, this pandemic is a public health emergency that has caused estimated costs of 16 trillion U.S. dollars. Apart from this, it invoked an outstanding interest in virology research [2].

Since then, professionals from many different fields, i.e. public health specialists, researchers, biomedical staff, bioinformaticians and veterinarians have put even more effort than before into the monitoring of potentially dangerous viral diseases. International managing institutions with a globally distributed netwok work on safe and healthy environments for animal and human populations. The World Organisation for Animal Health (WOAH), founded as Office International des Epizooties (OIE), implements standards in animal health and the handling of zoonoses and other diseases. As an intergovernmental organisation following the multidisciplinary One Health principle, it supports its members in the prevention of animal diseases of concern. National veterinary authorities must notify the WOAH in case they detect cases of diseases that are listed by the WOAH. The most important definitions,

the significance, impacts and surveillance measures of animal diseases are examined below.

1.1 Viral Livestock Diseases

Infectious diseases caused by viruses that affect domesticated animals, like for example cattle, pigs, goats, sheep, and poultry are referred to as viral livestock diseases. The most frequent and known diseases include Foot-and-Mouth Disease, African Swine Fever, Avian Influenza and Newcastle Disease. They can spread quickly among animals, and in some cases are transmitted to humans, making them zoonotic diseases. There are over 200 known types of zoonoses, some of them like rabies being 100% preventable through vaccination and medication [3]. A report from the International Livestock Research Institute (ILRI) states that zoonoses account for approximately 2.5 billion illness cases in humans and 2.7 million deaths annually [4]. The Centres for Disease Control and Prevention (CDC) and its U.S. government partners listed the top eight zoonotic diseases of national concern in a report, filing zoonotic influenza and emerging coronaviruses such as SARS and Middle East Respiratory Syndrome (MERS) [5]. This joint report is used to tackle the listed diseases with a broader focus [6]. At the same time, not all livestock diseases of viral origin are zoonotic: Around 60% of all known human infectious diseases and approximately 75% of all newly emerging infections are zoonotic [7].

The term livestock is a vague term that generally refers to any breed or animal population that is kept by humans for commercial or useful purpose. According to the 20th Livestock Cencus of the Department of Animal Husbandry and Dairying, given out by the Indian government, India holds the world's largest amount of livestock with 535.78 million animals as of 2019 [8]. Globally, the ice-free surface that is dedicated to the purpose of livestock whether it is for farmlands or feed production, is up to 26% of the area [9]. Not only food production and economy, but

also global trade, the agricultural sector and employment rates highly depend on livestock resources. These numbers illustrate the impressive interconnectedness of the humans with the livestock sector. The consequences of a collapse of this important industry would therefore be significant and far-reaching. As the livestock industry is directly affected by the occurrence of zoonoses in both developed and developing countries, affected parties have a strong interest in avoiding any constraints that might be caused by disease outbreaks.

Historic Outbreaks of Zoonotic Diseases

Historically, zoonoses have shaped serious infectious events. Pathogens that cause zoonotic diseases are viruses (37.7%), and according to surveillance data also bacteria (41.4%), parasites (18.3%), fungi (2.0%) or prions (0.8%) [10]. Prior to the COVID-19 pandemic modern zoonotic diseases like Ebola virus disease and salmonellosis had high infection rates. Influenza viruses cause epidemics each year, and circulate in all parts of the world. Influenza appears in zoonotic and human-only spreads, but the different types of virus can recombine occasionally and cause events such as the 1918 Spanish flu [11, 12]. Especially for poultry, highly pathogenic avian influenza (HPAI) of the H5 subtype is an ongoing threat [13]. Since its first detection in China, 1996 it has been reported in many avian populations, both domestic and wild. Even though it has adapted to birds as the specific host, the virus can further adapt and in rare cases be transmitted between humans [14]. Avian influenza has caused more recent seasonal outbreaks, such as the 2014-15 outbreak in the United States resulting in almost 50 million birds that died as a consequence of an infection or of depopulation [15]. In 2020, there were several outbreaks reported in Europe, almost all with HPAI viruses from the H5 subtype [16]. It mainly affected farmed ducks due to the high density of animals in the facilities and the separation from wild birds due to domestication [16]. The latest outbreak of avian influenza is still ongoing, started in early 2022 and until today, February 23, 2023 **TODO: update numbers and**

source has led to more than 58 million culled or died birds. It is reported in 37 countries and so far, six human infections were reported in this outbreak [17]. This number is not nearly as high as for the animals affected, but considering that during the last 20 years, there were fewer than 900 confirmed cases of H5N1 in humans and the mortality rate of 50%, each human infection is a risk [17].

Risk Factors and Impact of Disease Outbreaks

Reasons for recurring huge outbreaks of viral diseases in animal confinements come from the advantageous circumstances for virus transmission as it is warm and humid. In general, animal husbandry practices have evolved in the sense that domestic animal species are raised in relatively small and usually confined spaces at a high density. This domestication has given plenty of opportunities to develop more pathogens of viral and bacterial origin over time. The spread of international trading of farm animals has amplified the number of infected animals and the number of infectious diseases. As transmission routes can differ depending on the disease, the other factor is how easy the infectious agent spreads (transmissibility). Vector-borne diseases are transmitted by living organisms that transfer pathogenic microorganisms to other, uninfected animals or humans. Vectors can be mosquitoes, fleas or ticks. Among others, the World Health Organization (WHO) identifies major globally present vector-borne diseases as malaria, dengue, yellow fever and Zika virus disease [18]. Another transmission mode is direct contact airborne transmission. Environmental factors such as a high temperature, humidity and precipitation can facilitate a virus to spread and keep it alive [19]. Inadequate food and water supplies, overpopulation and mass migration of animals pose additional risks for transmission of animal diseases.

Outbreaks of livestock diseases do not only affect animal and human health, but also cause high economic losses. Restrictions and containment measures, as well as the culling of animals in the case of confirmed cases of listed diseases, lead to a

loss of income for farmers – since livestock and their products, such as milk, eggs or meat, are used for further production, other businesses that rely on these products are also affected by disease outbreaks. Even if infected animals do not die or have to be culled, the medium and long-term consequences of infection can affect the health of the animals. Consequences can be poor growth or poor production and feed conversion. Another impact of depopulating infected animal populations is the loss of biodiversity [20, 21]. Wildlife populations of endangered species experiencing disease outbreak can be decimated, leading to ecological imbalances and interference with natural food chains [22, 23, 24].

As shown, the spread of viral disesaes among animal populations can have enormous impacts on dependent industries, individuals and populations.

Notifiable Animal Diseases

For reasons of biosecurity and surveillance purposes, the WOAH has agreed on a list of notifiable animal diseases that must be reported to in agricultural authorities. This list includes a total of 117 diseases, partly endemic or highly transmissable, such as Foot-and-mouth-disease, lumpy skin disease, peste des petits ruminants, classical swine fever, highly pathogenic avian influenza and Newcastle disease. The list does not cover all known zoonoses and animal diseases since not all of them pose an actual risk for outbreaks.

Reports of illness cases of animals filed by national veterinary authorities are used to detect unusual incidents, including mortality or sickness of animals and have adverse effects on socio-economic or public health. The notifiable animal diseases include more than 50 wildlife diseases which can impact livestock health [25]. As the surveillance of viral animal diseases is still of highest priority in order to avoid expensive and dangerous outbreaks, this topic is discussed in more detail in the following introductory chapter.

1.2 Prevention, Surveillance and Control

Given the potential danger of disease outbreaks to animal, human and public health, the question is how to detect, monitor, control and prevent outbreaks in farm animal populations.

To avoid the impact that a disease outbreak can have, the best method is to avoid the disease in the first place. This leads to the principle of prevention, which sees its main task as reducing the overall risk of a virus spreading. Corresponding measures are vaccinations and hygiene standards. For viral material that reassorts over time as the number of infections increases, the potential for a virus to exploit host cell genes that favour viral growth and survival may be high [26]. Other disease prevention practises include disinfection and good animal husbandry. Practitioners in the field or in veterinary clinics are obliged to follow this principle of prevention. In-depth strategies to prevent viral diseases depend heavily on the characteristics of the virus, taking into account transmission modes, environmental stability, zoonotic risk and pathogenesis. Exclusion of infected livestock and vaccination of potentially infected flocks is increasingly practised worldwide [26]. The spatial spread of disease can be contained through quarantine, separation from wildlife populations, testing and regular inspections of imported animals.

Surveillance of viral diseases involves the collection of basic information about the disease, including incidence, prevalence and transmission patterns; the systematic and regular collection and analysis of these data is crucial to obtain a detailed overview of the spread. This need for data has led the WOAH to publish the above-mentioned list of notifiable diseases. Based on the data collected, authorities can inform their decisions on the allocation of resources for disease control and other containment activities [26, 27].

Common methods for animal diseases surveillance include notifiable diseases reporting, laboratory-based surveillance and population-based surveillance. General awareness among veterinary diagnosticians and practitioners is another key to an

effective surveillance system. Most countries have their own national veterinary authorities, coordinated by the WOAH to enable a coordinated exchange of information [27]. Since efforts in tackling viral disease outbreaks or mass vaccination are very expensive, official budgets from the governments are needed. This makes it a political responsibility to prevent and control animal diseases.

One important component of modern and accurate surveillance systems of viral diseases is the access to relevant data. Technologies to produce DNA sequencing data have developed to be very cost and time efficient which makes the study of infectious diseases better and faster. At the same time, the amount of DNA sequencing data produced with next-generation sequencing (NGS; also known as high-throughput sequencing, HTS) platforms prove this change. NGS platforms include IonTorrent, Illumina and Oxford Nanopore Technologies (ONT). Advances in the biotechnological application and evaluation of these data are revolutionalising the field on the molecular level [28]. Sequencing technologies take a key role in describing viral diversity in humans and animals, in detecting pathogens and co-infections, in epidemiologic research about the evolution of viral material and in metagenomic characterisation of new microbial material. More detailed methods that are used for viral animal disease surveillance with NGS-based technologies are described in Chapter 2.

1.3 Motivation and Objectives of the Thesis

Bioinformatics and data analysis are crucial for understanding and monitoring viral diseases. However, there is a lack of knowledge and resources in many parts of the world. This is particularly true for poorer countries with small laboratories and national health organisations that are not well equipped with modern sequencers and surveillance systems. Additionally, transporting clinical samples across international borders is difficult and expensive. Nonetheless, efforts are made to establish global

networks such as the Zoonotic Disease Integrated Action (ZODIAC). It is an initiative by the International Atomic Energy Agency (IAEA), launched in 2021, with five major objectives: (1) Strenghthening member states' detection, diagnostic and monitoring capabilities, (2) Developing and making novel technologies available for the detection and monitoring of zoonotic diseases, (3) Making real-time decision-making support tools available for timely interventions, (4) Understanding the impact of zoonotic diseases on human health and (5) Providing access to an agency coordinated response for zoonotic diseases [29]. In collaboration with technical experts from different fields and from all over the world, and to support the Veterinary Diagnostic Laboratory (VETLAB) Network, the ZODIAC project has the resources to provide standardised, easy-access, public and integrated pipelines for virus surveillance on a long-term. This will enable laboratories and veterinarians to monitor and analyse their samples more effectively, leading to early detection and prevention of viral livestock diseases.

Due to the outstanding research efforts brought about by the COVID-19 pandemic, analysis pipelines for SARS-CoV-2 samples were developed on the Galaxy platform. Galaxy and the implementation of pipelines are discussed in more detail in Chapter 3. Using the knowledge and application of SARS-CoV-2 and transferring it to other viruses will lead to a more comprehensive understanding of viral diseases and better prevention strategies.

This work is part of the ZODIAC project and supports pillar (2) in the development of integrated pipelines that enable laboratories, veterinarians and other health professionals to analyse their data from samples obtained with HTS technologies. The zoonoses studied are avian influenza A for subtype identification and a poxvirus pipeline for determining poxvirus genomes sequenced as half-genomes in a tiled-amplicon approach. The poxvirus pipeline has been tested with samples of lumpy skin disease virus. These two viruses have been chosen because of available test samples that were used for validation of the pipelines.

In summary, the lack of bioinformatics knowledge and resources in poorer coun-

tries poses a major challenge to effective, globally integrated viral animal diseases surveillance systems. However, established global networks such as ZODIAC together with VETLAB can provide the necessary resources to enable effective surveillance and analysis of viral animal diseases. This in turn will lead to early detection and prevention of disease outbreaks and ultimately protect public health and reduce the impact of viral diseases on livestock.

2 State-of-the-Art

In the demand for an effective, high-quality approach to the analysis of isolates from infected animals, molecular studies help to investigate characteristics of the sample. Genome analysis has become an integral part of animal disease surveillance, especially since the advent of high-throughput sequencing technologies in the last 15 years. Next-generation techniques and applications are described below, the state of the art in poxvirus and avian influenza virus detection and analysis, and lastly the drawbacks of the methods discussed.

2.1 High-throughput Technologies in Genomics and Virology

When comparing DNA sequencing technologies, there are differences in speed, throughput and volume of sequences. The term ”next-generation” in NGS used to describe newer technologies in the field implies a next step in the evolution of sequencing technologies. As sequencing machine technologies evolve rapidly, there are gradations such as ”second-generation” and ”third-generation”. Following the original 1977 Sanger sequencing method using radioactivity and gels, second-generation sequencers are advancements of Sanger sequencing that uses sequencing by synthesis [30]. In second-generation methods, reactions run in parallel and drastically reduce overall costs compared to Sanger sequencing. They produce short sequence reads length and are able to detect reads without using electrophoresis. Reads are equal to single fragments of DNA or RNA. Third-generation sequencing technologies typically

generate longer primary reads of DNA (and RNA) molecules while maintaining the massive parallelism of the technology and taking advantage of this benefit [31]. The nowadays most commonly used next-generation technologies for DNA sequencing and their applications are described below.

2.1.1 NGS Platforms and Applications

By far the biggest player in the field of DNA sequencing is the Illumina platform, first developed by Solexa and Lync Therapeutics [32]. Illumina sequencing is based on bridge amplification, which creates clusters of copies of each DNA fragment. This technique involves repeated synthesis reactions with proprietary modified nucleotides containing a different fluorescent label for each of the four bases A, T, C and G. The reactions are performed over 300 or more rounds, and fluorescent detection allows for faster detection through direct imaging. An Illumina sequencer outputs data in the form of sequence reads, which are short DNA fragments ranging from 50 to 600 base pairs in length depending on the specific instrument and protocol used [32, 31, 30]. The output data from an Illumina sequencer typically is in the form of raw sequence files in FASTQ format, which contain the base calls and corresponding quality scores for each read. These reads can be used for downstream analyses such as viral genome assembly and variant calling.

Oxford Nanopore Technologies (ONT) is a third-generation paradigm shifting sequencing technology. It measures changes in ionic current across membranes as single-stranded DNA nucleotides pass through a nanopore [33]. Nanopore-based DNA sequencing technologies are purchasable as a portable, small MinION (ONT) device, allowing experts to use it for applications where space requirements or portability are important [34, 33]. The cyclic mode of sequencing used in second-generation approaches is replaced by sequencing in real-time with read lengths of up to 10,000 basepairs [33]. Despite its advantages, the main caveat of ONT is its relatively high error rate compared to other HTS methods [35]. This makes ONT

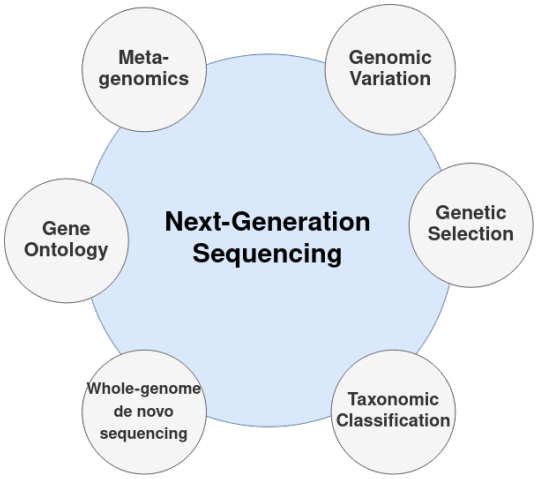

Figure 1: Overview of next-generation sequencing technology applications in virology.

less suitable for single-nucleotide variant analysis that is required in some diagnostic applications [36, 37].

Other frequently used second-generation platforms are Roche/454 sequencing, Ion Torrent (Thermo Fisher) technology and SOLiD (Sequencing by Oligonucleotide Ligation and Detection). Third-generation platforms include single molecule real-time sequencing (SMRT) by PacBio and nanopore sequencing [38].

As NGS platforms are widely used in biomedical and clinical contexts, some of the most important applications in diagnostic virology are depicted in Figure 1. In virology, metagenomics can be used to identify viruses in complex clinical samples [39]. It allows for the detection of known and novel viruses without prior knowledge of the infectious agent. Metagenomics involves the sequencing of all genetic material in a sample, including viral genomes, to identify the presence of viruses. Once a virus is identified, genomic variation refers to differences in the DNA sequence of a virus between different strains or isolates. These variants can be used for tracking the spread of an outbreak, identification of sources of an infection, or determination of the level of virus virulence [40]. Variant detection is only possible with NGS data, as they provide insight to the genome on a nearly every-base level and allow to reliably

interpret and identify the many different possible variants [41].

Genetic selection describes to the process by which certain viral strains become more prevalent in a population over time due to selective pressures. In diagnostic virology, genetic selection is used to track the evolution of a virus in the course of time and determine which strains are most likely to cause outbreaks or epidemics. This is of special interest in the backtracing of infected animals to know where the virus came from. Using gene ontology, functions and interactions of genes are described. This is crucial to identify the genes responsible for specific viral functions and to understand how these functions contribute to viral pathogenesis.

Based on their genetic and structural characteristics, viruses are classified to existing systems, called taxonomic classification. This clustering analysis can be used for the type identification of a virus causing infections and determination of its potential for transmission and pathogenicity [42].

Whole-genome de novo sequencing is the sequencing of an entire viral genome without prior knowledge of its genetic sequence. Similar to metagenomics, this technique can be used to identify novel viruses, to study mutations in viral genomes and to track the evolution of a virus over time [31].

2.1.2 Detection of Viral Pathogens

For NGS methods to be a viable tool in diagnosis and analysis of viral animal diseases, the methods must be efficient and reliable. Almost all downstream analyses depend on the data obtained by sequencing, so it is imperative to choose the most appropriate method for each application. Metagenomic-based approaches use whole-genome sequencing to characterise viral diversity in animal, human and environmental samples. The detection of rare and novel infectious pathogens and the study of mutations in the genome are crucial for developing a deeper understanding of livestock viromes and potential zoonotic agents. In addition, NGS has been shown to detect non-culturable organisms as well as co-infections that have not been detected using

traditional microbiological approaches [43]. Metagenome sequencing often relies on a low number of pathogenic reads to detect and to make diagnostic calls. As sequencing depth directly influences genome coverage that can be obtained, the optimal amount of data to cover the complete genome is necessary. It has been shown that for a full virus genome to be represented, NGS data generated from ribo-depleted total RNA with a minimum length of one million high-quality reads works best [44]. Nevertheless, validation pipelines and confirmatory tests are needed for NGS approaches to pathogen detection [45].

2.1.3 Data Analysis Issues

Since the surveillance of viral animal diseases with NGS is advancing rapidly, it is important that regions and health organizations that experience high damage of viral outbreaks but do not have their own facilities and know-how have access to the needed tools and knowledge. Costs for NGS sequencers are still high and the access to appropriate laboratories is not given everywhere. Networks like VETLAB and standardisation of techniques, for example freely available and published by the WOAH, can enable professionals worldwide independent of their equipment on site. In the scope of the ZODIAC project, this aspect is addressed by providing protocols for each step from taking samples of potentially infected animals to the detailed analysis and derived actions [29].

NGS methods themselves have downsides that need to be considered when applying these techniques. Generally, chimerical sequences are formed during sequencing, which may be interpreted as false positives for novel organisms. Chimeric products are artifacts originating from joining sequences and are represented by point mutations, insertions and deletions. Chimera formation also occurs during PCR amplification [46].
During bioinformatics analysis steps using algorithms with computationally expensive steps, the choice of the algorithm as well as its configuration settings have huge

impact on the final results obtained. This includes algorithms in steps such as filtering for quality, clustering and sequence classification [47]. The cleaning step or filtering phase eliminates low-quality reads from the dataset, whereas the error correction process distinguishes true variants from those caused by experimental noise. This is based on the concept that errors occur randomly with low frequency, while true mutations tend to be clustered, and their frequency can be measured [48]. Longer reads avoid this problem because contigs must not be assembled in the first place, avoiding clustering and filtering errors. This is why the shift in third-generation and later sequencing platforms is towards longer reads again. Due to the relatively high error rates of HTS technologies, that base on the sequencing process itself, polymerase chain reaction (PCR) amplification of the viral material, and reverse transcription of viral RNA to cDNA, it is crucial to include quality checks and filtering steps when using the HTS data [49].

Each application of software with NGS data requires expertise in resolving limitations and drawbacks of specific methods. This in turn requires skills and experience in the field and the careful interpretation of results. Still, NGS provides a large pool of methods which eases this task, although available algorithms for genome assembly and amplicon analysis have drawbacks and limitations [50].

2.2 Poxvirus Analysis

In the following, current approaches to analyse NGS data of poxviruses are described. To get into the topic, the characteristics of poxviruses are examined.

2.2.1 Poxviruses

Throughout human history, poxviruses have played a significant role with variola being the most notorious as it is the causative agent of smallpox. Smallpox has

been described in Chinese texts dating back to the 4th Century AD, and evidence of pox-like scars found on Egyptian mummies suggests the disease may have existed as far back as the 2nd millennium BC [51]. The discovery of a vaccine for smallpox made it the first disease to be eradicated by human efforts, and variola was the first human virus to be successfully eliminated [52]. Modern vaccinology owes its origins to Edward Jenner's discovery in the late 18th century that zoonotic infections with the "cowpox virus" provided immunity to smallpox [51]. Furthermore, vaccinia virus, which is now used for smallpox vaccination, was the first animal virus to be observed using electron microscopy and the first to be utilized as a vector for transporting foreign genes into animals. This is why poxviruses are among the best-known viruses. The family of poxviruses, *Poxviridae*, is a family of double-stranded DNA viruses. Its natural hosts are vertebrates and arthropods and there are currently 83 species within 22 genera in this family. The family is divided into two subfamilies, *Entemopoxvirinae* (insect-infecting viruses) and *Chordopoxvirinae* (vertebrate-infecting viruses).
Historically, poxviruses were classified based on disease symptoms and the animal species that was infected. Humans, cows, sheep, goats, horses and pigs have been studied to determine not only clinical symptoms but with the aim to classify poxviruses. This genus classisification has been confirmed by recent comparative genome analysis [53]. Symptoms of disease caused by a poxvirus infection are skin lesions that can differ in size. Depending on the type of poxvirus, the papules can vary from small and pearly papules in infections of lumpy skin disease virus (LSDV) to larger crusts and spread generalized pustules in infections with the variola virus. Other general symptoms include fever, headache and rash.

Table 1 shows ten representatives of the 18 Chordopoxvirus genera according to the newest ICTV (International Committee on Taxonomy of Viruses) Taxonomy Release from 2021, while at least five genera contain zoonotic poxviruses [54]. Orthopoxviruses have the biggest impact on human and animal health, and are remarkable for their broad host spectrum ranging from humans to wild and domestic animals [52]. The Chordopoxvirus subfamily is characterised by its large, linear double-stranded genome.

Genus	Virus Species	Natural Hosts
Avipoxvirus	Canarypox virus	Songbirds
	Fowlpox virus	Chickens, turkeys
Capripoxvirus	Sheep pox virus	Sheep
	Lumpy skin disease virus	Cattle
Centapoxvirus	Yokapox virus[1]	Humans, mosquitoes
Cervidpoxvirus	Deerpox virus	Deer
Crocodylidpoxvirus	Crocodilepox virus	Crocodiles
Leporipoxvirus	Myxoma virus	Rabbits, hares
Molluscipoxvirus	Molluscum contagiosum virus[1]	Humans, primates, birds, dogs
Orthopoxvirus	Variola virus (Smallpox)	Humans (eradicated)
	Mpox virus[1]	Humans, primates
	Cowpox virus[1]	Humans, cats, cows, elephants
	Vaccinia virus[1]	Humans, cattle, buffalos, rabbits
	Camelpox virus	Camels
Parapoxvirus	Pseudocowpox virus[1]	Humans, cattle
	Orf virus[1]	Humans, sheep, goats, etc.
Suipoxvirus	Swinepox virus	Pigs
Yatapoxvirus	Yaba monkey tumour virus[1]	Humans, rhesus monkeys

[1] Zoonotic disease

Table 1: Representative viruses from ten Chordopoxvirus genera.

Size varies between 134 to 365 kilobases [55, 56]. Chordopoxvirus genomes contain 130 to 328 open reading frames (ORF), and typically, two identical inverted terminal repeats (ITR) are located at both ends of poxvirus genomes.

Vaccination is available for smallpox, and the vaccine is even considered protective against symptoms of all orthopoxvirus infections. It is recommended for laboratory staff that works with mpox, cowpox, vaccinia and variola [57]. For animals, there is a smallpox-based vaccine that is used to protect elephants against cowpox [58]. Sheep and goats are broadly vaccinated with an orf vaccine, which is, similar to smallpox

vaccine, a live virus. The effective vaccination against existing poxvirus diseases and further microbiological studies, as well as similarities between poxviruses, motivate the expansion of existing data analysis pipelines that work for a specific poxvirus so that they can also work with other poxviruses.

Lumpy Skin Disease Virus

Lumpy Skin Disease is caused by the lumpy skin disease virus belonging to the *Capripoxvirus* (CaPV) genus within the family of poxviruses, subfamily *Chordopoxvirinae* [59]. The LSD virus genome is a double-stranded linear DNA molecule of circa 151 kilobasepairs in length. It contains between 147 and 156 open reading frames. Similar to other poxviruses, the LSDV genome consists of a central coding region which is bounded by two identical ITR regions with a length of circa 2,400 basepairs at both ends of the genome. This is a key characteristic to consider during reconstruction of the genome. With a sequence identity of over 96% with the other CaPV genus members sheep pox and goatpox, the LSDV genome is highly similar to the other CaPV genomes [60].

LSDV is not known to be transmissiable to humans and therefore not a zoonosis. Natural hosts of LSDV are cattle and Asian water buffalos. Although CaPV is considered to be host specific, sheep pox and goatpox strains can naturally cross-infect in both host species. There have been no cases of natural infection of sheep or goats with LSDV reported [61]. The three CaPV viruses are the most serious poxvirus diseases of livestock in terms of economic losses in the case of an outbreak.

Cattle infected with the LSDV typically show symptoms like fever, reduced feed and water uptake and characteristic skin nodules. The number of lesions varies from a few to many, covering the whole body [62]. From these symptoms alone, it is impossible to differentiate the diagnosis between sheep pox, goatpox and lumpy skin disease. Even with classical methods like cell culture and electron microscopy the highly similar viruses cannot be distinguished. Nowadays, PCR and sequencing are

the techniques used to provide the sensitive detection of CaPv [63].

LSDV has spread from the African continent and since 2019 reached major cattle producer countries in Asia, mainly India, Republic of China and Bangladesh. Other bigger outbreaks in south-west Europe were reported in 2014 to 2018, although these countries opted for a strict vaccination program and successfully eliminated LSDV from the region [64]. In African and Asian countries, veterinarians struggle to fight endemic LSDV outbreaks because of a lacking financial support by governments, justified by low mortality and morbidity rates.

One strain of LSDV that has been extensively studied is the Neethling strain, first isolated in Kenya in 1958. It constitutes the strain used for the live attenuated vaccine that is widely used, if accessible, for cattle against LSDV outbreaks. Some countries use sheep pox vaccines to protect cattle against LSD, even though it does not bring complete immunity. Nevertheless they are used in regions where all CaPV are prevalent [65].

2.2.2 Pipelines for Genomic Analysis with Poxvirus NGS Data

The need for rapid identification of a virus sample to distinguish between species of poxviruses requires sensitive analysis of NGS data. Challenges in alignment against a reference are the identical ITR at both ends of Capripoxviruses, which is omitted from many pipelines and not part of the analysis, as well as the high identity of 96-97% between the three Capripoxviruses. In order to reach a sufficiently high coverage in all parts of the genome, the reference and the reads can be split into two parts to map against the identical ITR regions. With a tiling approach, there is no ambiguity in where to map a read from the ITR to. However, the reads have to be sequenced in two pools, which is not a standard protocol. These challenges make it difficult to differentiate between LSDV, goatpox and sheeppox [60].

A whole genome sequencing (WGS) approach to distinguish capripoxviruses is described by Mathijs et al. [66]. They develop a sequencing protocol in two pools to separate the ITR regions. After pre-processing, the pools of reads are de novo assembled with SPAdes and the resulting contigs of each pool are merged into a single contig. To find the correct merging location, an overlap of one amplicon in the middle is assembled in both pools. The test results with various samples show that this approach reconstructs nearly complete CaPV genomes. The presented tiling amplicon approach is not usable as an automated pipeline, but can be implemented using the tool specifications in the article. Other viral genomes have been examined in a similar tiling amplicon approach with Illumina, ONT or PacBio sequenced data [67, 68, 69, 70].

A pipeline of Zhao et al. was designed to study the whole genome of monkeypox samples [71]. After the de novo assembly step, a neural network method is used for smart gap filling between the assembled contigs. The method shows that gap filling of a genome is an *all k shortest path* (KSP) problem and can be used in an automated pipeline from HTS reads to the whole genome sequence. They conclude that it is a promising method to find the "correct" sequence but it did not find the correct sequence assembly for five cases in a sample sequence of monkeypox. Therefore, this method can be used as a guiding first-shot feature, but should not be used for sensitive analyses. Also, the neural-KSP method requires knowledge in how to finetune the pipeline parameters.

Other methods to detect the species of capripoxvirus of a given sample is nucleic acid extraction and real-time PCR [72]. This approach is based on the presence of specific genes to distinguish between capripoxviruses, but since it does not work with NGS data, it does not allow for more analyses and is not comparable to the previous methods.

2.3 Avian Influenza Virus Analysis

NGS-based sequencing data from AIV samples need profound processing to gain insights into the subtype and variants within the sequence. In the following, the causative agent for avian influenza, avian influenza virus, is described in detail and state-of-the-art methods in the form of automated pipelines for the analysis of such data are presented.

2.3.1 Avian Influenza Virus

Informally known as bird flu, avian influenza is a viral infectious disease that affects wild birds and poultry. The avian influenza virus (AIV) has occasionally crossed the species barrier and infects mammals, including humans. This makes it a high-priority zoonotic viral disease that has been designated as notifiable by WHO and WOAH [25]. Avian influenza occurs in two variants that determine severity: low pathogenic avian influenza (LPAI) and high pathogenic avian influenza (HPAI), with only HPAI cases requiring notification. The virus spreads indirectly via contaminated material, e.g. feed, water supplies, feces or feathers. It is transmitted directly from bird to bird via the air, mainly through the transregional movement of wild birds and through long distance bird migration. Humans become infected through close contact with infected livestock or wild birds, and most reported human avian influenza infections are from farm workers and others who are exposed in markets, production or clinical contexts [14].
Symptoms of severe illness are characterised by influenza-like signs such as fever, nasal discharge, coughing and conjunctivitis. This applies to infections in both human and mammals, while infected birds show signs such as swollen heads, loss of appetite, breathing difficulties and a decrease in egg production.

AIV contains a negative-sense, single-stranded segmented RNA genome, and due to the segmented nature of the virus, co-infection of different influenza strains can lead

to reassortment events. Avian influenza viruses are members of the *Orthomyxoviridae* family and the four species of influenza viruses A, B, C and D are distinguished on the basis of the presence of the nucleoprotein (NP) and matrix (M1) proteins [14]. AIV subtypes are determined by the hemagglutinin (HA) and neuraminidase (NA) segments, which include all known influenza A virus subtypes H1-H16 in combination with N1-N11, resulting in subtype designations such as H5N1 or H7N9 [14, 73]. To be infectious, a virus particle must contain one of eleven proteins in each of the eight unique segments PB2 (poymerase), PB1/PB1-F2 (polymerase), PA/PA-X (polymerase), HA, NP, NA, M1/M2 and NS1/NEP (distinct non-strucutral proteins). Mutations in the HA and NA genes occur relatively frequently due to the prone-error RNA polymerase in the viral genome which lacks the proof-reading exonuclease activity. LPAI subtypes H5 and H7 usually infect poultry, although the natural hosts of avian influenza A are wild waterfowl. These subtypes can transform into HPAI during circulation in poultry stocks by recombination with other gene segments or the host genome [74]. Both LPAI and HPAI infections have been reported in domestic poultry, i.e. ducks and chickens, turkeys, caged birds, aquatic birds and wild birds. As the different influenza species can infect different animal hosts, all of them can infect pigs and humans.

Influenza A strains are the most virulent virus species, and have caused all major historic flu outbreaks through reassortment. Subtypes H5, H7 and H9 are responsible for the largest outbreaks of AIV that also spread to humans [75]. The first confirmed report of human infection with an animal avian influenza virus dates to 1958, and since then 16 subtypes have been detected in humans [76]. Zoonotic spillover events have becomeincreasingly common since the early 20th century and have led to major endemics such as a huge H5 outbreak in the U.S. in 2014-2015. It resulted in more than 25 million bird deaths [77]. Another current outbreak, resulting in more than 58 million dead birds and costs of around 661 million U.S. dollars began in 2022 and is spreading across the U.S. [78]. Vaccination against HPAI in poultry are used worldwide to ward off avian influenza. They also serve as a preventive measure in

the event of an outbreak to reduce the risk of introducing the virus into poultry populations [79, 80].

2.3.2 Pipelines for Genomic Analysis with Avian Influenza Virus NGS Data

Surveillance systems in the field of genotyping emerging viral strains include classical phylogenetic methods far classifying viral strains, assessing tree topologies, distinguishing between novel and emerging strains, and discovering novel disease-causing variants [41]. These analyses are essential given the high genetic variability of the genome, and since it consists of eight segments, specific bioinformatics workflows are required for the analysis.

The challenge in identifying subtypes and detecting variants lies in the diversity of HA and NA genes, the main targets of the host immune response. The HA and NA genes have evolved into several subfamilies and require a dynamic reference selection approach for sequencing analysis. There are a growing number of web platforms, suites and pipelines that enable the analysis of influenza-specific samples with NGS data and resources for further analysis, e.g. Influenza Research Database/Fludb [81], EpiFLU/GISAID [82], Nextflu [83], NCBI Influenza Virus Resource [84], FluNet [85] and OpenFluDB [86]. Many existing suites for automated analysis of influenza samples are based on SARS-CoV-2 research and have been adapted for the similarly large influenza genome. INSaFLU and PAIVS are two pipelines specifically designed for the analysis of NGS-generated (avian) influenza samples and are discussed in more detail below.

INSaFLU

One prominent pipeline for viral metagenomic detection and routine genomic surveillance, INSaFLU ("INSide the FLU"), provides a web-based protocol for data gener-

ated by Illumina, Ion Torrent or ONT sequencers [87]. It is the first influenza-focused suite to process NGS data to automatically generate output data and answer key questions in influenza genomic surveillance. These include type and subtype identification, reference-based mapping, consensus sequence generation, and phylogenetic tree construction. The INSaFLU pipeline consists of steps that cover some of the objectives in parallel: (1) Reads quality analysis and improvement, (2a) classification, (2b) mutation detection and consensus generation, (3a) intra-host minor variant detection, (3b) alignment/phylogeny and (3c) coverage analysis. Using the output data of step (3b), a downstream integrative phylogenetic and geotemporal analysis with Nextstrain can be started. A reference sequence for the mapping step must be provided as input data from the beginning. Currently, INSaFLU is accepting NGS data from influenza, SARS-CoV-2 and monkeypox samples [87]. The INSaFLU pipeline is installed locally via the command-line on any server instance, which requires technical knowledge to set up, but can also be used via the website. The pipeline steps cannot be customised via the web interface, instead general configurations can be set at the beginning. The pipeline is constantly being developed to integrate new features and modules.

PAIVS

PAIVS (Prediction of Avian Influenza Virus Subtype) is a pipeline specifically designed for avian influenza virus samples. It consists of five steps: (1) pre-processing, (2a) reference-based alignment or (2b) de novo assembly, (3) subtyping, (4) variant calling and identification of the closest sequences by (5) BLASTn (Basic Local Alignment Search Tool) [88]. PAIVS uses a similar approach to INSaFLU, but leaves it up to the user to decide whether to include a de novo assembly step. The results are presented in a downloadable format for the user and include a graphical summary. The pipeline is written in Python and is freely available on `http://ircgp.com/paivs`, being a web-based platform only available in Korean [88]. This is a very limiting factor for

the usability of PAIVS.

2.4 Foot-and-Mouth Disease Virus Analysis

2.4.1 Foot-and-Mouth Disease Virus

Cloven hoofed animals, small positive-sense ssRNA virus (8.3kb) Aphthovirus genus, Picomaviridae family, 7 distinct serotypes (with different subtypes each), high heterogenity of the virus in the host populations –> WGS needed for accurate variant calling

2.4.2 Pipelines for Genomic Analysis with Foot-and-Mouth Disease Virus NGS Data

* https://academic.oup.com/bib/article/21/5/1766/5565040?login=true&casa_token=1DxsiOURgvsAAAAA:m9Fyy4W5xLE-6y2jMvn8EkQHIayhjrgXZJ3sSTNnTgx61D6TLP3MhmZIIRO-k alignment with Bowtie2

* ONT serotyping: https://www.frontiersin.org/articles/10.3389/fvets.2021.656256/full and mapping, consensus sequence generation, BLASTn, But: many SNPs found (due to high error rates of 5-10% in MinION sequencer)

* https://assets.researchsquare.com/files/rs-2396402/v1/7ab80566-9e56-40a3-8117-07191dbcf2a9.pdf?c=1672243503 mapping with Bowtie2, samtools, variants calling (Mutect2), SnpEff for annotation, phylogenetic analysis. GATK4 pipeline

2.5 Tools for Genomic Analysis with NGS Data

* preprocessing (trimmomatic, fastp, fastqc) -> quality reporting
* aligner (BWA-MEM, Medaka, minimap2, Bowtie2)
* mutation detection, annotation (FreeBayes, Medaka, bcftools, SnpEff)
* consensus sequence generation (ivar, Medaka)
* coverage analysis (bedtools, deepTools)
* alignments (MAFFT)
* de novo assembler (SPAdes)
* phylogenetic trees (FastTree, IQ-Tree, PhyloCanvas)
* closest sequence (BLAST, VAPOR)
* variant caller (lofreq, medaka)

2.6 Pipelines for Genomic Analysis with Viral NGS Data

In the following, pipelines are presented that can be used with unspecified or unknown virus data. They cover some general parts of the previously mentioned pipelines but focus mainly on virus discovery, assembly and consensus sequence generation.
ViReflow is a pipeline for viral consensus sequence generation and provides a mapping-based approach to variant calling and many optional downstream analyses such as de novo assembly and lineage assignment [89]. The pipeline is based on the Reflow suite, and all computations run in an AWS container in a cloud. Reflow emphasises versioning, testing and workflow sharing and does not provide a user-friendly web interface. Instead, it is accessible via a command-line interface. As a result, it may not be as easy to use as Galaxy and its workflows, including workflow development, as this requires programming in Go language. Similar to other pipelines, ViReflow was originally created for the consensus genome construction of SARS-CoV-2 samples and has been extended for use with all viral genomes [89].

Another automated pipeline for viral genome assembly, lineage assignment, mutation and intra-host variant detection is V-Pipe, a computational pipeline assessing genetic diversity and introducing a new alignment method *ngshmmalign* specifically for small and highly diverse viral genomes. It includes local and global haplotype reconstruction and a module for detection of flow cell cross-contamination [90]. Although V-Pipe is suitable for all viral genomes, it was tested for the identification of the eight influenza segments and successfully identified them from the test sample.

Other freely available pipelines for the analysis of viral genomes from NGS data with several focuses in genomics are VirFind [91] and IRIDA (Integrated Rapid Infectious Disease Analysis) [92]. These pipelines focus on rapid identification of viral materials and do not provide steps for detailed downstream analyses. Automated pipelines for metagenomic NGS data are drVM (detect and reconstruct known Viral genomes from Metagenomes) and VirMAP [93, 94]. However, they do not consider the segmented influenza genome and do not provide output data for custom downstream analyses. To our knowledge, there is no freely available pipeline that uses a mapping-based approach that focuses on the viral segments of the AIV genome and uses the closest possible reference for each segment. For the various possible downstream analyses, depending on the specific research question, it is critical for a pipeline to provide data outputs and endpoints that enable user-specific assays. A Galaxy workflow covering the above points has been developed in this thesis and is described in the following chapter.

3 Materials and Methods

3.1 Galaxy Platform

Galaxy is a web-based scientific platform that has become a major player in many fields of life sciences and bioinformatics. Founded in 2007 it has provided an emerging amount of resources and tools to empower scientists and researchers to work with biomedical datasets. The platform is free to use and collaborative, making it one of the biggest of its kind. Resources on Galaxy cover genomics, metagenomics, transcriptomics, proteomics, drug discovery and non-biology fields like natural language processing and social sciences.

Galaxy's primary objective is to make analyses more accessible, reproducible, and easier to communicate among researchers. The platform's distinctive and success is attributed to four core elements: a very active community, a public server for analyses, an open-source software ecosystem, and the Galaxy ToolShed. The community adheres to the FAIR practices (Findable, Accessible, Interoperable and Reusable) [95].

The Galaxy community is thriving, with over 124,000 users who also contribute to subcommunities. The public server for analyses provides access to public datasets and workflows. The open-source software ecosystem ensures automated setup and deployment of all tools and services, making it simple for beginners and professionals to use. The Galaxy ToolShed is a server dedicated to hosting, sharing, and installing

tools used on the platform. A Galaxy tool is the abstraction layer that makes external software usable from within Galaxy with a frontend, i.e. lets users use the program with all its parameters and inputs from within Galaxy.
Galaxy workflows are a key feature that allow the user to stack tools in a chain and to configure them so that the workflow user only has to upload his or her data for the input fields. The automation of tools in a chain is used for modular, longer analyses that are executed repeatedly.
Workflows that are available on and accepted by the Intergalactic Workflow Commission (IWC; `https://github.com/galaxyproject/iwc`) are conform with the community's best practise standards and tested on the latest Galaxy release. Dockstore and WorkflowHub automatically publish the IWC workflows and guarantee the availability in a Docker-based environment on Dockstore [96] and on the workflow collaboratory WorkflowHub [97].

Important contributions of Galaxy, as stated by the Galaxy Community (2022), include Vertebrate Genome Project assembly workflows and collaborations on SARS-CoV-2 research. Another toolkit leveraged in Galaxy is Galaxy-ML, a set of tool that provides a suite for analyses based on machine learning. With growing publicity, more topics are covered by and moved to Galaxy. It has contributed to over 5,700 scientific publications and has many tutorials available for researchers to use. Training material and ready-to-use workflows facilitate professionals and beginners in the field to use Galaxy for their research purposes.

The platform is continuously enhanced, and it still attracts around 2,000 new users every month, indicating the quality and significance of the project. The team and infrastructure of Galaxy initially come from the Nekrutenko lab in the Center for Comparative Genomics and Bioinformatics at Penn State, the Taylor lab at Johns Hopkins University, and the Goecks Lab at Oregon Health & Science University. All of these organisations have contributed significantly to the success of Galaxy. There are 138 public servers available worldwide as of 2023, while the most prominent

general-purpose server instances are hosted by teams at University of Freiburg, Germany (for UseGalaxy.eu), Texas Advanced Computing Center (for UseGalaxy.org) and Genomics Virtual Laboratory, formerly at the University of Queensland (for UseGalaxy.org.au). These main public servers are synchronized in their tools and set of reference tools [95].

3.2 Requirements

SARS-CoV-2 Pipeline as Baseline.
annotated variants are of interest
description of basic steps

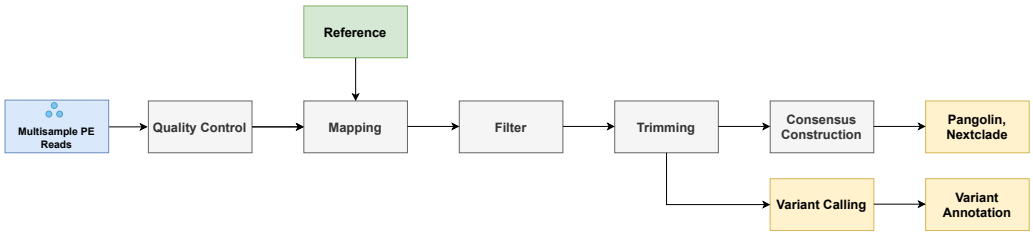

Figure 2: Simplified SARS-CoV-2 ARTIC PE reads iVar-based workflow.

tested workflow, includes 'minimal' steps:

1. Quality control
2. Mapping
3. Filtering
4. Trimming
5. Consensus Sequence Construction

Plus Variant Calling and genome annotation;
Plus phylogenetic ranking "to assign a SARS-CoV-2 genome sequence the most likely lineage based on a chosen nomenclature system" (Pangolin)

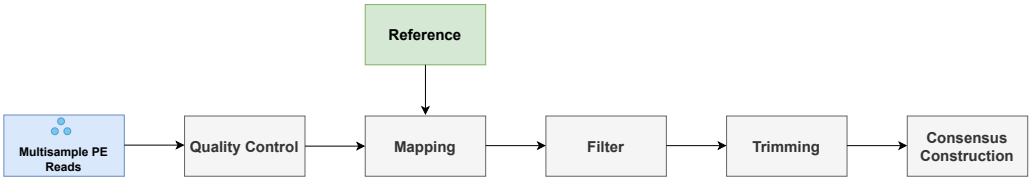

Figure 3: Simplified minimal ARTIC PE reads iVar-based workflow.

which problems should the pipeline solve?

what is "ampliconic" sequence analysis, ARTIC Illumina-sequenced data

Requirements for Poxvirus Workflow

- repetitions in the start and end regions → need to split reads into 2 pools and mask references
- after splitting, merging alignments back

Requirements for AIV Workflow

- reference for each of the 8 segments has to be chosen
- align reference of each segment with consensus sequence and add to phylogenetic analysis
- snipit for visualisation of SNPs
- trimming would dismiss too many of the already short reads
- have a reference database ready for VAPOR (multifasta) with read name pattern
- dismiss B/C/D strains and H17, H18 subtypes (occur in bats only)

a tool to get closest reference

For a detailed genetic analysis of the AIV genome, several consecutive steps are required after taking a sample from an infected host and sequencing it. NGS data need to be preprocessed in order to remove reads that are too short, have low base quality or include NGS platform-specific adapters that are ligated to the read ends and need to be trimmed.

why mapping is good and not assembly: The diversity of HA and NA segments' sequences is significant enough to make it challenging to map sequenced reads to a single Influenza A reference sequence chosen by the user using a naive approach. Although this approach may be effective for the other six segments, the mapping software would frequently be unable to locate sufficient plausible matches for sequenced reads of HA and NA origin to continue with the analysis.

Requirements for FMDV Workflow

3.3 Workflow Development

"Reference-based genomic Surveillance" (INSaFLU)

3.3.1 Poxvirus Illumina Amplicon Workflow

47 distinct steps

Tiling amplicon approach for CaPV genome. Makes up 23 primer pairs for an amplicon size of 7.5 kb each instead of smaller sizes usually used in tiling amplicon protocols. (ref2.2.2)

Workflow is composed of seven crucial steps: - preparing reference sequence for mapping (masking halves) - quality control - mapping - Filtering - merging - trimming - consensus sequence construction

CaPV genomes have the central coding region bounded by identical inverted terminal repeats, containing 156 putative genes. the repeat of the ITRs would make any mapping in these regions ambiguous. need to part the reads in two pools and do mapping in two parts: N-mask the reference (start and end by start/end position primers)

Figure 4: Simplified Poxvirus PE reads iVar-based workflow.

Efficiency: Assembly vs. Mapping; efficiency more details in discussion.

If the goal in a broad and rapid surveillance is a high number of sample throughput

and analysis, assembly is too cost and time senstive. The presented pipeline could be used in a broad context for the use in many laboraties. building index is expensive (BWT)

3.3.2 AIV Illumina Reads Workflow

We propose a fully automated pipeline for the analysis of Illumina-sequenced paired-end reads from influenza samples. The workflow is integrated in the Galaxy platform and available via **TODO: link**. It is designed to take one input sample at a time and the outputs of the analysis steps can be used for further research based on the user's interest. The workflow is outlined in Figure 5, where the nine main steps of the workflow are visualized. One upside of the workflow is the consideration of the different segments of the influenza virus genome. After uploading paired-end reads and a reference sequence database, which is available online too **TODO: link**, the workflow is designed to build a hybrid reference from the given database for each of the segments of the genome. The reference sequence database consists of eight FASTA files, one per segment, containing numerous full-length sequences for a given segment. The provided database file consists of 56 sequences for each segment. If a user decides to upload their own references, it is important to follow the sequence identifier pattern: $>segment\_name|influenza\_strain|subtype|accession\_number$. For example, one entry's identifier is $>PB1|A/duck/Manitoba/1953|A/H10N7|KF435047.1$ followed by the sequence.

After (1) preprocessing of the reads with `fastp` to dismiss reads shorter than 30 basepairs and trimming read ends, the database of reference sequences is used to (3) find the closest possible reference for each of the segments. The tool `VAPOR` outputs a table with a scoring that comes from the graph construction, and should not be confused with the identity of the sequence compared to the reference. As `VAPOR` is running once per segment but has independent inputs, this step is executed in parallel. `VAPOR` is a graph-based classifier that maps k-mers to a weighted De Bruijn

graph [98]. Its benchmarking shows that it runs significantly faster than BLAST and default configuration leads to reasonable matches, as long as the given sample is not very different from or novel to the provided sequences in the database.
Using the highest scoring sequences from the eight `VAPOR` runs, a hybrid reference sequence is built. To control the statistics of the graph and adapt the configuration, a table with the highest `VAPOR` scores of each run is generated.
The hybrid reference sequence is composed of the eight segments and is used for (3) mapping with `BWA-MEM` (Burrow-Wheeler Aligner for short-read alignment) [99].

48 distinct steps

user can decide for size of phylogenetic trees (number of sequences to include from the next closer sequences -> measure by VAPOR scores, determines size of tree)

consensus sequence per segment is included in the phylogenetic trees to visualize the taxonomy (spatial/temporal spread)

upload list of FASTA files with a reference database for each influenza segment;

Table 2

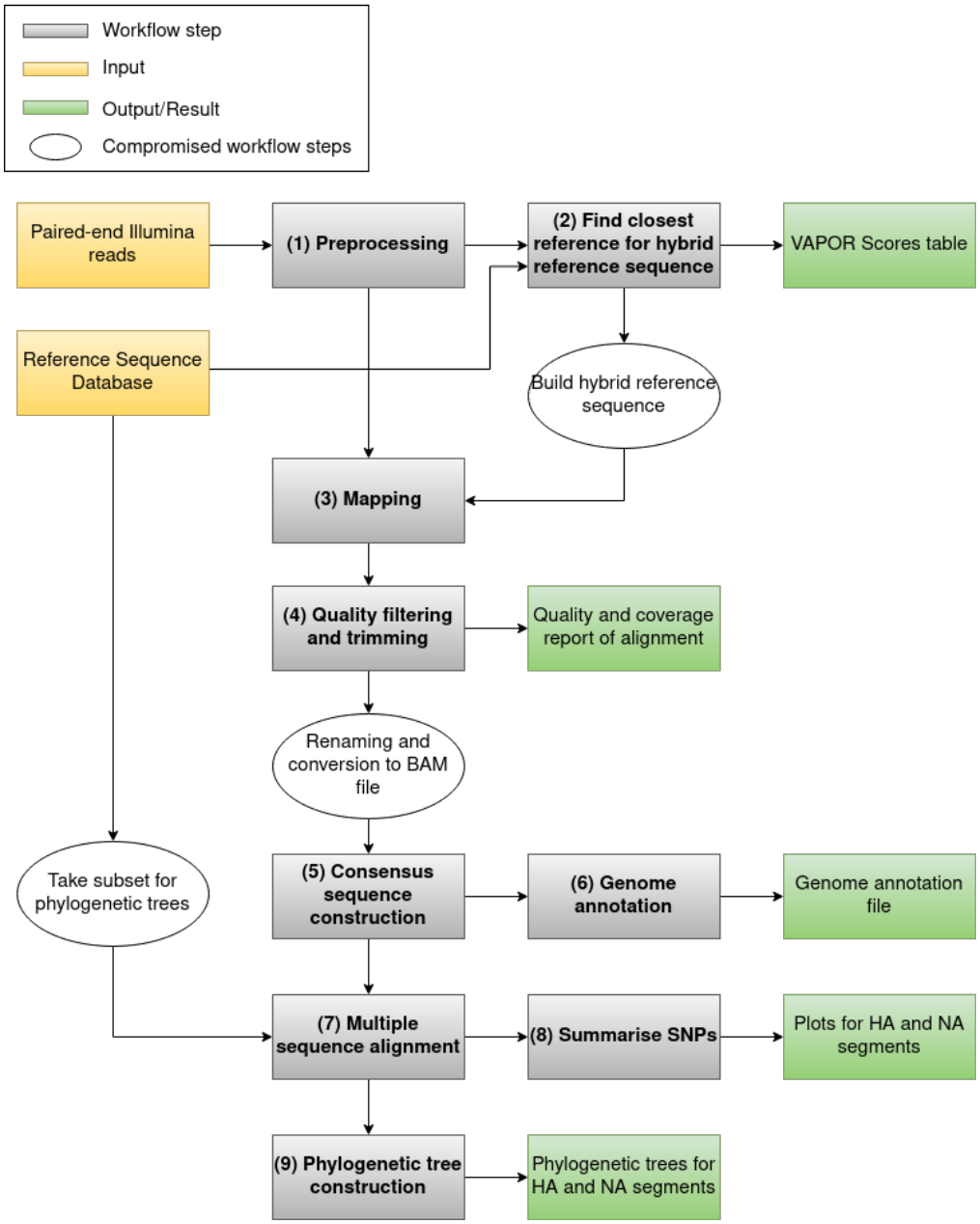

Figure 5: Simplified AIV Illumina reads iVar-based workflow.

steps: (3) Mapping with BWA-MEM: preprocessed reads against hybrid reference, BAM output (4) Filter by quality and for paired and properly mapped reads (5) Consensus sequence construction (from BAM file with useful sequence names that

include the segment name) (-) create an overview of the VAPOR scoring in form of a table, with the identifier names (6) Prokka from consensus sequences, for genome annotation (kingdom viruses) (7) Snipit to visualize and summarize SNPs relative to the hybrid reference, needs a multiple sequence alignment of the segments with MAFFT (8) IQ-Tree build phylogenetic tree from filtered reference database, size of the tree is determined by input, consensus sequences are renamed so they contain the segment name and the sample and outputs trees for HA and NA genes.

Overview of outputs in a table with datatypes (1) Quality and coverage report of alignment (2) VAPOR scores in a table (3) Genome annotation file (4) snipit SNP plots for HA and NA genes (5) IQ-Trees for HA and NA genes

Kraken2 vs. VAPOR; Efficiency: LoFreq vs. iVar consensus; both consensus identification methods using the same site-specific depth threshold

3.3.3 FMDV Illumina Amplicon Workflow

4 Results and Workflow Evaluation

real-world data provided by Belgian Sciensano laboratory to test the workflow.

4.1 Poxvirus Illumina Workflow

Samples by Sciensano by Elisabeth Mathijs

IWC link, primer scheme. tested with LSDV data, pipeline outputs on 20L70, 20L81

4.1.1 Results for LSDV Datasets 20L70 and 20L81

4.2 AIV Workflow

Samples by Sciensano s4+s8, Tunesian?

point out output for downstream analyses

4.2.1 Results for Dataset U2012100-n21_S8

Quality report, snipit plots, IQ-Tree for HA/NA, consensus reference, VAPOR scores

4.2.2 Results for Dataset U2008751-n5_S4

Quality report, snipit plots, IQ-Tree for HA/NA, consensus reference, VAPOR scores

4.3 FMDV Workflow

Samples by Pirbright Institute by Dr. Graham Freimanis

5 Discussion

5.1 Contribution to the Field

single sample vs. multi sample (reality check, what is needed?)
further pox viruses, pipelines can be more or less easily applied/adjusted
limitations
LSDV interesting for all poxviruses and adjustable due to ITR and tiled-amplicon approach
AIV downstream - everything is possible. Highlight key minor assets that indicate adaption to mammals -> databases needed to check against, detect mutation of isolates?
Generally: annotate on amino acid layer (most information)

Make phylogenetic trees publicly accessible, not one sample per strain but in high resolution and greater details, strains from different countries,

"The high sensitivity of the NGS technology ensures that major kinds of viral pathogens in mixed samples can be detected." One strength of NGS is that it can be used to detect emerging viral diseases with a high genetic variation. Like AIV. Since it can analyse a full sequence instead of targeting a specifig gene. -> makes sense to use virus-specific primers for PCR or NGS

”Comparison of the whole genome sequences of recent LSDV isolates from the 2015–2016 epidemic in southern Europe revealed only a limited number of point mutations between the isolates” WGS is essential to capture all genetic variation at once

In sequencers, falso positive variants (FPV) must be avoided (happens when too many amplifiation cycles are made)

5.2 Future Directions

further validation and improvement of the developed pipelines, expansion to other viral livestock diseases, integration with existing surveillance systems; expand the VETLAB network to entitle even more professionals to professionally analyse their samples.

AIV workflow offers many possible directions for downstream analysis:

* consensus sequence for each segment -> compare consensus sequence to others can help identify outbreaks and patterns of transmission, get more insights how the virus spreads and its evolution * Prokka annotation file. Predict the protein coding regions of the virus, to understand the function of the viral proteins and how they interact with host cells * SNPs relative to the reference sequence * MSA and phylogenetic tree for broad or detailed phylolgenetic analysis and understand evolutionary relationships between the sample and other strains. could also use clusters or subtypes within the sample. make trees available so that new isolates can be immediately arranged * more visualisation of the data

* long-term objective: build public high-resolution databases to enable researchers to detect mutation of an isolate. this is crucial for a global surveillance system to work.

6 Conclusion

Summary of objectives, achievements and discussion

By relying on raw read data rather than assembled genomes and allowing every result to be traced back to its raw data, it goes a step beyond current surveillance efforts.

Bibliography

[1] F. Wu, S. Zhao, B. Yu, Y.-M. Chen, W. Wang, Z.-G. Song, Y. Hu, Z.-W. Tao, J.-H. Tian, Y.-Y. Pei, *et al.*, "A new coronavirus associated with human respiratory disease in China," *Nature*, vol. 579, no. 7798, pp. 265–269, 2020.

[2] World Health Organization (WHO), "Weekly epidemiological update on covid-19." https://www.who.int/publications/m/item/weekly-epidemiological-update-on-covid-19—22-february-2023, 2023. Retrieved: 22nd February, 2023.

[3] World Health Organization (WHO), "Zoonoses." https://www.who.int/news-room/fact-sheets/detail/zoonoses, 2020. Retrieved: 23rd February, 2023.

[4] D. Grace, F. Mutua, P. Ochungo, R. Kruska, K. Jones, L. Brierley, M. Lapar, M. Y. Said, M. T. Herrero, P. Phuc, *et al.*, "Mapping of poverty and likely zoonoses hotspots," 2012.

[5] I. Brown, J. Banks, R. Manvell, S. Essen, W. Shell, M. Slomka, B. Londt, and D. Alexander, "Recent epidemiology and ecology of influenza A viruses in avian species in Europe and the Middle East," *Developments in biologicals*, vol. 124, pp. 45–50, 2006.

[6] Centers for Disease Control and Prevention, "Zoonotic Diseases Shared Between Animals and People of Most Concern in the US CDC Newsroom," 8.

[7] K. E. Jones, N. G. Patel, M. A. Levy, A. Storeygard, D. Balk, J. L. Gittleman, and P. Daszak, "Global trends in emerging infectious diseases," *Nature*, vol. 451, no. 7181, pp. 990–993, 2008.

[8] Ministry of Fisheries, Animal Husbandry & Dairying, "Livestock Census." https://pib.gov.in/PressReleaseIframePage.aspx?PRID=1813802, 2022.

[9] H. Steinfeld, P. Gerber, T. D. Wassenaar, V. Castel, M. Rosales, M. Rosales, and C. de Haan, *Livestock's Long Shadow: Environmental Issues and Options*. Food & Agriculture Org., 2006.

[10] S. J. Salyer, R. Silver, K. Simone, and C. B. Behravesh, "Prioritizing Zoonoses for Global Health Capacity Building—Themes from One Health Zoonotic Disease Workshops in 7 Countries, 2014—2016," *Emerging infectious diseases*, vol. 23, no. Suppl 1, p. S55, 2017.

[11] R. J. Garten, C. T. Davis, C. A. Russell, B. Shu, S. Lindstrom, A. Balish, W. M. Sessions, X. Xu, E. Skepner, V. Deyde, *et al.*, "Antigenic and Genetic Characteristics of Swine-Origin 2009 A(H1N1) Influenza Viruses Circulating in Humans," *science*, vol. 325, no. 5937, pp. 197–201, 2009.

[12] M. J. Gibbs, J. S. Armstrong, and A. J. Gibbs, "Recombination in the hemagglutinin gene of the 1918 "Spanish flu"," *Science*, vol. 293, no. 5536, pp. 1842–1845, 2001.

[13] D.-H. Lee, K. Bertran, J.-H. Kwon, and D. E. Swayne, "Evolution, global spread, and pathogenicity of highly pathogenic avian influenza H5Nx clade 2.3. 4.4," *Journal of veterinary science*, vol. 18, no. S1, pp. 269–280, 2017.

[14] R. G. Webster, W. J. Bean, O. T. Gorman, T. M. Chambers, and Y. Kawaoka, "Evolution and Ecology of Influenza A Viruses," *Microbiological reviews*, vol. 56, no. 1, pp. 152–179, 1992.

[15] D.-H. Lee, J. Bahl, M. K. Torchetti, M. L. Killian, H. S. Ip, T. J. DeLiberto, and D. E. Swayne, "Highly Pathogenic Avian Influenza Viruses and Generation of Novel Reassortants, United States, 2014–2015," *Emerging infectious diseases*, vol. 22, no. 7, p. 1283, 2016.

[16] N. S. Lewis, A. C. Banyard, E. Whittard, T. Karibayev, T. Al Kafagi, I. Chvala, A. Byrne, S. Meruyert, J. King, T. Harder, *et al.*, "Emergence and spread of novel H5N8, H5N5 and H5N1 clade 2.3. 4.4 highly pathogenic avian influenza in 2020," *Emerging Microbes & Infections*, vol. 10, no. 1, pp. 148–151, 2021.

[17] C. Adlhoch, A. Fusaro, J. L. Gonzales, T. Kuiken, S. Marangon, É. Niqueux, C. Staubach, C. Terregino, I. Aznar, *et al.*, "Avian influenza overview September-December 2022," *EFSA journal. European Food Safety Authority*, vol. 21, no. 1, p. e07786, 2023.

[18] World Health Organization, "Global vector control response 2017-2030," *Global vector control response 2017-2030*, 2017.

[19] R. Eccles, "An Explanation for the Seasonality of Acute Upper Respiratory Tract Viral Infections," *Acta oto-laryngologica*, vol. 122, no. 2, pp. 183–191, 2002.

[20] C. Lacroix, A. Jolles, E. W. Seabloom, A. G. Power, C. E. Mitchell, and E. T. Borer, "Non-random biodiversity loss underlies predictable increases in viral disease prevalence," *Journal of the Royal Society Interface*, vol. 11, no. 92, p. 20130947, 2014.

[21] S. Morand, "Emerging diseases, livestock expansion and biodiversity loss are positively related at global scale," *Biological Conservation*, vol. 248, p. 108707, 2020.

[22] R. S. Reid, C. Bedelian, M. Y. Said, R. L. Kruska, R. M. Mauricio, V. Castel, J. Olson, and P. K. Thornton, "Global Livestock Impacts on Biodiversity,"

Livestock in a Changing Landscape. Drivers, Consequences, and Responses; Steinfeld, H., Mooney, HA, Schneider, F., Neville, LE, Eds, pp. 111–138, 2010.

[23] D. J. Civitello, J. Cohen, H. Fatima, N. T. Halstead, J. Liriano, T. A. McMahon, C. N. Ortega, E. L. Sauer, T. Sehgal, S. Young, *et al.*, "Biodiversity inhibits parasites: Broad evidence for the dilution effect," *Proceedings of the National Academy of Sciences*, vol. 112, no. 28, pp. 8667–8671, 2015.

[24] R. Espinosa, D. Tago, and N. Treich, "Infectious Diseases and Meat Production," *Environmental and Resource Economics*, vol. 76, no. 4, pp. 1019–1044, 2020.

[25] World Organisation for Animal Health, "Animal Diseases." https://www.woah.org/en/what-we-do/animal-health-and-welfare/animal-diseases/, 2023. Retrieved: 25th February, 2023.

[26] "Chapter 6 - Epidemiology and Control of Viral Diseases," in *Fenner's Veterinary Virology (Fifth Edition)* (N. J. MacLachlan and E. J. Dubovi, eds.), pp. 131–153, Boston: Academic Press, fifth edition ed., 2017.

[27] WHO, OIE, "One Health," *World Health Organization*, vol. 736, 2017.

[28] G. G. D. Suminda, S. Bhandari, Y. Won, U. Goutam, K. K. Pulicherla, Y.-O. Son, and M. Ghosh, "High-throughput sequencing technologies in the detection of livestock pathogens, diagnosis, and zoonotic surveillance," *Computational and Structural Biotechnology Journal*, 2022.

[29] International Atomic Energy Agency, "Zoonotic Disease Integrated Action Initiative." https://nucleus.iaea.org/sites/zodiac/Shared2021. Retrieved: 21st February, 2023.

[30] E. R. Mardis, "Next-generation DNA sequencing methods," *Annual Review of Genomics and Human Genetics*, vol. 9, pp. 387–402, 2008.

[31] B. E. Slatko, A. F. Gardner, and F. M. Ausubel, "Overview of Next Generation Sequencing Technologies," *Current protocols in molecular biology*, vol. 122, no. 1, p. e59, 2018.

[32] Illumina, "An introduction to Next-Generation Sequencing Technology," *Illumina, Inc*, 2015.

[33] M. Jain, H. E. Olsen, B. Paten, and M. Akeson, "The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community," *Genome biology*, vol. 17, pp. 1–11, 2016.

[34] A. L. Greninger, S. N. Naccache, S. Federman, G. Yu, P. Mbala, V. Bres, D. Stryke, J. Bouquet, S. Somasekar, J. M. Linnen, *et al.*, "Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis," *Genome medicine*, vol. 7, pp. 1–13, 2015.

[35] S. Fu, A. Wang, and K. F. Au, "A comparative evaluation of hybrid error correction methods for error-prone long reads," *Genome biology*, vol. 20, no. 1, pp. 1–17, 2019.

[36] R. Bowden, R. W. Davies, A. Heger, A. T. Pagnamenta, M. de Cesare, L. E. Oikkonen, D. Parkes, C. Freeman, F. Dhalla, S. Y. Patel, *et al.*, "Sequencing of human genomes with nanopore technology," *Nature communications*, vol. 10, no. 1, p. 1869, 2019.

[37] C. P. Stefan, A. T. Hall, A. S. Graham, and T. D. Minogue, "Comparison of Illumina and Oxford Nanopore Sequencing Technologies for Pathogen Detection from Clinical Matrices Using Molecular Inversion Probes," *The Journal of Molecular Diagnostics*, vol. 24, no. 4, pp. 395–405, 2022.

[38] A. Rhoads and K. F. Au, "PacBio sequencing and its applications," *Genomics, proteomics & bioinformatics*, vol. 13, no. 5, pp. 278–289, 2015.

[39] C. Y. Chiu and S. A. Miller, "Clinical metagenomics," *Nature Reviews Genetics*, vol. 20, no. 6, pp. 341–355, 2019.

[40] M. Capobianchi, E. Giombini, and G. Rozera, "Next-generation sequencing technology in clinical virology," *Clinical Microbiology and Infection*, vol. 19, no. 1, pp. 15–22, 2013.

[41] D. C. Koboldt, K. M. Steinberg, D. E. Larson, R. K. Wilson, and E. R. Mardis, "The next-generation sequencing revolution and its impact on genomics," *Cell*, vol. 155, no. 1, pp. 27–38, 2013.

[42] B. E. Dutilh, A. Varsani, Y. Tong, P. Simmonds, S. Sabanadzovic, L. Rubino, S. Roux, A. R. Muñoz, C. Lood, E. J. Lefkowitz, *et al.*, "Perspective on taxonomic classification of uncultivated viruses," *Current opinion in virology*, vol. 51, pp. 207–215, 2021.

[43] P. G. Cantalupo and J. M. Pipas, "Detecting viral sequences in NGS data," *Current opinion in virology*, vol. 39, pp. 41–48, 2019.

[44] M. Visser, R. Bester, J. T. Burger, and H. J. Maree, "Next-generation sequencing for virus detection: covering all the bases," *Virology journal*, vol. 13, no. 1, pp. 1–6, 2016.

[45] T. D. Minogue, J. W. Koehler, C. P. Stefan, and T. A. Conrad, "Next-generation sequencing for biodefense: biothreat detection, forensics, and the clinic," *Clinical chemistry*, vol. 65, no. 3, pp. 383–392, 2019.

[46] P. Zylstra, H. S. Rothenfluh, G. F. Weiller, R. V. Blanden, and E. J. Steele, "Pcr amplification of murine immunoglobulin germline v genes: strategies for minimization of recombination artefacts," *Immunology and cell biology*, vol. 76, no. 5, pp. 395–405, 1998.

[47] E. Kopylova, J. A. Navas-Molina, C. Mercier, Z. Z. Xu, F. Mahé, Y. He, H.-W. Zhou, T. Rognes, J. G. Caporaso, and R. Knight, "Open-source sequence clustering methods improve the state of the art," *MSystems*, vol. 1, no. 1, pp. e00003–15, 2016.

[48] O. Zagordi, R. Klein, M. Däumer, and N. Beerenwinkel, "Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies," *Nucleic acids research*, vol. 38, no. 21, pp. 7400–7409, 2010.

[49] N. Beerenwinkel, H. F. Günthard, V. Roth, and K. J. Metzner, "Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data," *Frontiers in microbiology*, vol. 3, p. 329, 2012.

[50] F. Finotello, E. Lavezzo, P. Fontana, D. Peruzzo, A. Albiero, L. Barzon, M. Falda, B. Di Camillo, and S. Toppo, "Comparative analysis of algorithms for whole-genome assembly of pyrosequencing data," *Briefings in bioinformatics*, vol. 13, no. 3, pp. 269–280, 2012.

[51] F. Fenner, D. A. Henderson, I. Arita, Z. Jezek, and I. D. Ladnyi, "The history of smallpox and its spread around the world," *Smallpox and its Eradication*, pp. 209–244, 1988.

[52] F. Fenner, "Adventures with poxviruses of vertebrates," *FEMS microbiology reviews*, vol. 24, no. 2, pp. 123–133, 2000.

[53] C. Gubser, S. Hué, P. Kellam, and G. L. Smith, "Poxvirus genomes: a phylogenetic analysis," *Journal of General Virology*, vol. 85, no. 1, pp. 105–117, 2004.

[54] International Committee on Taxonomy of Viruses, "Virus Taxonomy: 2021 Release." https://ictv.global/taxonomy, 2021. Retrieved: 3rd March, 2023.

[55] C. R. Brunetti, H. Amano, Y. Ueda, J. Qin, T. Miyamura, T. Suzuki, X. Li, J. W. Barrett, and G. McFadden, "Complete genomic sequence and comparative analysis ofthe tumorigenic poxvirus yaba monkey tumorvirus," *Journal of virology*, vol. 77, no. 24, pp. 13335–13347, 2003.

[56] E. Tulman, C. Afonso, Z. Lu, L. Zsak, G. Kutish, and D. Rock, "The genome of canarypox virus," *Journal of virology*, vol. 78, no. 1, pp. 353–366, 2004.

[57] J. Cono, C. G. Casey, and D. M. Bell, "Smallpox vaccination and adverse reactions; guidance for clinicians," *MMWR. Recommendations and reports: Morbidity and mortality weekly report. Recommendations and reports*, vol. 52, no. RR-4, 2003.

[58] A. Kurth, G. Wibbelt, H.-P. Gerber, A. Petschaelis, G. Pauli, and A. Nitsche, "Rat-to-elephant-to-human transmission of cowpox virus," *Emerging infectious diseases*, vol. 14, no. 4, p. 670, 2008.

[59] P. J. Walker, S. G. Siddell, E. J. Lefkowitz, A. R. Mushegian, D. M. Dempsey, B. E. Dutilh, B. Harrach, R. L. Harrison, R. C. Hendrickson, S. Junglen, *et al.*, "Changes to virus taxonomy and the International Code of Virus Classification and Nomenclature ratified by the International Committee on Taxonomy of Viruses (2019)," *Archives of virology*, vol. 164, no. 9, pp. 2417–2429, 2019.

[60] E. Tulman, C. Afonso, Z. Lu, L. Zsak, G. Kutish, and D. Rock, "Genome of Lumpy Skin Disease Virus," *Journal of virology*, vol. 75, no. 15, pp. 7122–7130, 2001.

[61] F. Namazi and A. Khodakaram Tafti, "Lumpy skin disease, an emerging transboundary viral disease: A review," *Veterinary Medicine and Science*, vol. 7, no. 3, pp. 888–896, 2021.

[62] L. Prozesky and B. Barnard, "A study of the pathology of lumpy skin disease in cattle," *The Onderstepoort journal of veterinary research*, vol. 49, no. 3, pp. 167–175, 1982.

[63] S. Lafar, K. Zro, and M. M. Ennaji, "Capripoxvirus diseases: Current updates and developed strategies for control," in *Emerging and Reemerging Viral Pathogens*, pp. 635–655, Elsevier, 2020.

[64] FAO Sustainable Prevention, "Control and Elimination of Lumpy Skin Disease—Eastern Europe and the Balkan," *FAO Animal Production and Health Position Paper; Food and Agriculture Organization of the United Nations (FAO): Rome, Italy*, vol. 2, p. 25, 2017.

[65] J. Brenner, M. Bellaiche, E. Gross, D. Elad, Z. Oved, M. Haimovitz, A. Wasserman, O. Friedgut, Y. Stram, V. Bumbarov, *et al.*, "Appearance of skin lesions in cattle populations vaccinated against lumpy skin disease: statutory challenge," *Vaccine*, vol. 27, no. 10, pp. 1500–1503, 2009.

[66] E. Mathijs, A. Haegeman, K. De Clercq, S. Van Borm, and F. Vandenbussche, "A robust, cost-effective and widely applicable whole-genome sequencing protocol for capripoxviruses," *Journal of Virological Methods*, vol. 301, p. 114464, 2022.

[67] N. E. Freed, M. Vlková, M. B. Faisal, and O. K. Silander, "Rapid and inexpensive whole-genome sequencing of sars-cov-2 using 1200 bp tiled amplicons and oxford nanopore rapid barcoding," *Biology Methods and Protocols*, vol. 5, no. 1, p. bpaa014, 2020.

[68] S. N. Gardner, C. J. Jaing, M. M. Elsheikh, J. Peña, D. A. Hysom, and M. K. Borucki, "Multiplex degenerate primer design for targeted whole genome amplification of many viral genomes," *Advances in bioinformatics*, vol. 2014, 2014.

[69] N. D. Grubaugh, K. Gangavarapu, J. Quick, N. L. Matteson, J. G. De Jesus, B. J. Main, A. L. Tan, L. M. Paul, D. E. Brackney, S. Grewal, *et al.*, "An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using primalseq and ivar," *Genome biology*, vol. 20, no. 1, pp. 1–19, 2019.

[70] J. Quick, N. D. Grubaugh, S. T. Pullan, I. M. Claro, A. D. Smith, K. Gangavarapu, G. Oliveira, R. Robles-Sikisaka, T. F. Rogers, N. A. Beutler, *et al.*, "Multiplex pcr method for minion and illumina sequencing of zika and other virus genomes directly from clinical samples," *Nature protocols*, vol. 12, no. 6, pp. 1261–1276, 2017.

[71] K. Zhao, R. M. Wohlhueter, and Y. Li, "Finishing monkeypox genomes from short reads: assembly analysis and a neural network method," *BMC genomics*, vol. 17, pp. 527–537, 2016.

[72] B. Armson, V. Fowler, E. Tuppurainen, E. Howson, M. Madi, R. Sallu, C. Kasanga, C. Pearson, J. Wood, P. Martin, *et al.*, "Detection of capripoxvirus dna using a field-ready nucleic acid extraction and real-time pcr platform," *Transboundary and emerging diseases*, vol. 64, no. 3, pp. 994–997, 2017.

[73] F. Krammer, G. J. Smith, R. A. Fouchier, M. Peiris, K. Kedzierska, P. C. Doherty, P. Palese, M. L. Shaw, J. Treanor, R. G. Webster, *et al.*, "Influenza (primer)," *Nature Reviews: Disease Primers*, vol. 4, no. 1, p. 3, 2018.

[74] R. G. Webster and E. A. Govorkova, "H5n1 influenza – continuing evolution and spread," *New England journal of medicine*, vol. 355, no. 21, pp. 2174–2177, 2006.

[75] M.-A. Widdowson, J. S. Bresee, and D. B. Jernigan, "The global threat of animal influenza viruses of zoonotic concern: then and now," *The Journal of Infectious Diseases*, vol. 216, no. suppl_4, pp. S493–S498, 2017.

[76] V. Kluska, M. Macku, and J. Mensik, "Demonstration of antibodies against swine influenza viruses in man," *Ceskoslovenska pediatrie*, vol. 16, pp. 408–414, 1961.

[77] R. M. Seeger, A. D. Hagerman, K. K. Johnson, D. L. Pendell, and T. L. Marsh, "When poultry take a sick leave: Response costs for the 2014–2015 highly pathogenic avian influenza epidemic in the usa," *Food Policy*, vol. 102, p. 102068, 2021.

[78] Animal and Plant Health Inspection Service, U.S. Department of Agriculture, "2022-2023 confirmations of highly pathogenic avian influenza in commercial and backyard flocks." https://www.aphis.usda.gov/aphis/ourfocus/animalhealth/animal-disease-information/avian/avian-influenza/hpai-2022/2022-hpai-commercial-backyard-flocks, 2023. Retrieved: 9th March 2023.

[79] D. E. Swayne and E. Spackman, "Current status and future needs in diagnostics and vaccines for high pathogenicity avian influenza," in *Vaccines and Diagnostics for Transboundary Animal Diseases*, vol. 135, pp. 79–94, Karger Publishers, 2013.

[80] D. Swayne, G. Pavade, K. Hamilton, B. Vallat, K. Miyagishima, *et al.*, "Assessment of national strategies for control of high-pathogenicity avian influenza and low-pathogenicity notifiable avian influenza in poultry, with emphasis on vaccines and vaccination," *Revue Scientifique et Technique-OIE*, vol. 30, no. 3, p. 839, 2011.

[81] Y. Zhang, B. D. Aevermann, T. K. Anderson, D. F. Burke, G. Dauphin, Z. Gu, S. He, S. Kumar, C. N. Larsen, A. J. Lee, *et al.*, "Influenza research database: An integrated bioinformatics resource for influenza virus research," *Nucleic acids research*, vol. 45, no. D1, pp. D466–D474, 2017.

[82] Y. Shu and J. McCauley, "Gisaid: Global initiative on sharing all influenza data–from vision to reality," *Eurosurveillance*, vol. 22, no. 13, p. 30494, 2017.

[83] R. A. Neher and T. Bedford, "Nextflu: real-time tracking of seasonal influenza virus evolution in humans," *Bioinformatics*, vol. 31, no. 21, pp. 3546–3548, 2015.

[84] Y. Bao, P. Bolotov, D. Dernovoy, B. Kiryutin, L. Zaslavsky, T. Tatusova, J. Ostell, and D. Lipman, "The influenza virus resource at the national center for biotechnology information," *Journal of virology*, vol. 82, no. 2, pp. 596–601, 2008.

[85] A. Flahault, V. Dias-Ferrao, P. Chaberty, K. Esteves, A.-J. Valleron, and D. Lavanchy, "Flunet as a tool for global monitoring of influenza on the web," *Jama*, vol. 280, no. 15, pp. 1330–1332, 1998.

[86] R. Liechti, A. Gleizes, D. Kuznetsov, L. Bougueleret, P. Le Mercier, A. Bairoch, and I. Xenarios, "Openfludb, a database for human and animal influenza virus," *Database*, vol. 2010, 2010.

[87] V. Borges, M. Pinheiro, P. Pechirra, R. Guiomar, and J. P. Gomes, "Insaflu: an automated open web-based bioinformatics suite "from-reads" for influenza whole-genome-sequencing-based surveillance," *Genome Medicine*, vol. 10, pp. 1–13, 2018.

[88] H.-C. Park, J. Shin, S.-M. Cho, S. Kang, Y.-J. Chung, and S.-H. Jung, "Paivs: prediction of avian influenza virus subtype," *Genomics & Informatics*, vol. 18, no. 1, 2020.

[89] N. Moshiri, K. M. Fisch, A. Birmingham, P. DeHoff, G. W. Yeo, K. Jepsen, L. C. Laurent, and R. Knight, "The vireflow pipeline enables user friendly large scale viral consensus genome reconstruction," *Scientific reports*, vol. 12, no. 1, p. 5077, 2022.

[90] S. Posada-Céspedes, D. Seifert, I. Topolsky, K. P. Jablonski, K. J. Metzner, and N. Beerenwinkel, "V-pipe: a computational pipeline for assessing viral genetic diversity from high-throughput data," *Bioinformatics*, vol. 37, no. 12, pp. 1673–1680, 2021.

[91] T. Ho and I. E. Tzanetakis, "Development of a virus detection and discovery pipeline using next generation sequencing," *Virology*, vol. 471, pp. 54–60, 2014.

[92] T. C. Matthews, F. R. Bristow, E. J. Griffiths, A. Petkau, J. Adam, D. Dooley, P. Kruczkiewicz, J. Curatcha, J. Cabral, D. Fornika, *et al.*, "The integrated rapid infectious disease analysis (irida) platform," *BioRxiv*, p. 381830, 2018.

[93] H.-H. Lin and Y.-C. Liao, "drvm: a new tool for efficient genome assembly of known eukaryotic viruses from metagenomes," *Gigascience*, vol. 6, no. 2, p. gix003, 2017.

[94] N. J. Ajami, M. C. Wong, M. C. Ross, R. E. Lloyd, and J. F. Petrosino, "Maximal viral information recovery from sequence data using virmap," *Nature communications*, vol. 9, no. 1, p. 3205, 2018.

[95] The Galaxy Community, "The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update," *Nucleic Acids Res.*, vol. 50, no. W1, pp. W345–W351, 2022.

[96] B. D. O'Connor, D. Yuen, V. Chung, A. G. Duncan, X. K. Liu, J. Patricia, B. Paten, L. Stein, and V. Ferretti, "The Dockstore: enabling modular, community-focused sharing of Docker-based genomics tools and workflows," *F1000Research*, vol. 6, 2017.

[97] C. Goble, S. Soiland-Reyes, F. Bacall, S. Owen, A. Williams, I. Eguinoa, B. Droesbeke, S. Leo, L. Pireddu, L. Rodríguez-Navas, *et al.*, "Implementing FAIR Digital Objects in the EOSC-Life Workflow Collaboratory," *Zenodo*, 2021.

[98] J. A. Southgate, M. J. Bull, C. M. Brown, J. Watkins, S. Corden, B. Southgate, C. Moore, and T. R. Connor, "Influenza classification from short reads with vapor facilitates robust mapping pipelines and zoonotic strain detection for routine surveillance applications," *Bioinformatics*, vol. 36, no. 6, pp. 1681–1688, 2020.

[99] H. Li, "Aligning sequence reads, clone sequences and assembly contigs with bwa-mem," *arXiv preprint arXiv:1303.3997*, 2013.

Appendix

AIV Workflow Tools

No.	Step	Tool	Tool Version
1	Preprocessing	fastp	0.20.1
2	Find closest reference for hybrid reference sequence	VAPOR	1.0.2
3	Mapping	BWA-MEM	0.7.17.2
4	Quality filtering and trimming	Samtools	1.9
5	Consensus sequence construction	iVar consensus	1.3.1
6	Genome annotation	Prokka	1.14.6
7	Find and visualize SNPs	snipit	1.0.7
8	Multiple sequence alignment	MAFFT	7.508
9	Phylogenetic tree construction	IQ-Tree	2.1.2

Table 2: Galaxy tools that were used for each of the listed steps of the AIV workflow with their tool version number.

Poxvirus Workflow Tools

No.	Step	Tool	Tool Version
1	Mask reference for pool1 and pool2	maskseq	5.0.0
2	Preprocessing	fastp	0.23.2
3	Mapping	BWA-MEM	0.7.17.2
4	Quality filtering and trimming	Samtools view	1.15.1
5	Merge mappings of pool1 and pool2	Samtools merge	1.15.1
6	Primer trimming	iVar trim	1.3.1
7	Consensus sequence construction	iVar consensus	1.3.2

Table 3: Galaxy tools that were used for each of the listed steps of the Poxvirus workflow with their version number.