Master Thesis

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

Viktoria Isabel Schwarz

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. med. Marcus Panning

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

TODO

Abstract

TODO

Availability

Poxvirus Workflow: `https://workflowhub.eu/workflows/439`

Avian Influenza Virus Workflow:

Foot-and-mouth Disease Virus Workflow:

Contents

List of Figures

List of Tables

Acronyms

AIV Avian Influenza Virus

AWS Amazon Web Services, Inc.

BAM Binary Alignment Map

BED Browser Extensible Data

BLAST Basic Local Alignment Search Tool

BWA-MEM Burrow-Wheeler Aligner for short-read alignment with Maximal Exact Matches

CaPV Capripoxvirus

CDC Centres for Disease Control and Prevention

cDNA coding Deoxyribonucleic Acid

DNA Deoxyribonucleic Acid

COVID-19 Coronavirus Disease 19

drVM detect and reconstruct known Viral genomes from Metagenome

FMD Foot-and-mouth Disease

FMDV Foot-and-mouth Disease Virus

FPV False Positive Variant

GATK Genome Analysis Toolkit

HA Hemagglutinin

HPAI Highly Pathogenic Avian Influenza

HTS High-Throughput Sequencing

IAEA International Atomic Energy Agency

ICTV International Committee on Taxonomy of Viruses

ILRI International Livestock Research Institute

INSaFLU "INSide the FLU"

IRIDA Integrated Rapid Infectious Disease Analysis

ITR Inverted Terminal Repeat

iVar intrahost Variant analysis of replicates

IWC Intergalactic Workflow Commission

KSP All k shortest path

LPAI Low Pathogenic Avian Influenza

LSD Lumpy Skin Disease

LSDV Lumpy Skin Disease Virus

MAFFT Multiple Alignment using Fast Fourier Transform

MERS Middle East Respiratory Syndrome

MSA Multiple Sequence Alignment

NA Neuraminidase

NCBI National Center for Biotechnology Information

NGS Next-Generation Sequencing

NP Nucleoprotein

OIE Office International des Epizooties

ONT Oxford Nanopore Technologies

ORF Open Reading Frame

PAIVS Prediction of Avian Influenza Virus Subtype

PCR Polymerase Chain Reaction

PDF Portable Document Format

RNA Ribonucleic Acid

RSV Respiratory Syncytial Virus

SAM Sequence Alignment Map

SARS Severe Acute Respiratory Syndrome

SARS-CoV-2 Severe Acute Respiratory Syndrome Coronavirus 2

SAT Southern African Territories

SMRT Single Molecule Real-Time Sequencing

SNP Single Nucleotide Polymorphism

VCF Variant Call Format

VETLAB Veterinary Diagnostic Laboratory

WDL Workflow Description Language

WGS Whole-Genome Sequencing

WHO World Health Organization

WOAH World Organization for Animal Health

ZODIAC Zoonotic Disease Integrated Action

1 Introduction

Sharing environments means sharing diseases – this simple relationship expresses how pathogens spread among populations if they get in touch. The affected populations can be animalic or human. Impacts of disease outbreaks 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** `https://covid19.who.int/table`, 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].

Professionals from many different fields, i.e. public health specialists, researchers, biomedical staff, bioinformaticians and veterinarians are carefully monitoring potentially dangerous viral diseases. International managing institutions with a globally distributed network work on safe and healthy environments for both animal and human populations. The World Organization 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 from an animal host 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 Severe Acute Respiratory Syndrome (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 vague, and 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 between humans and livestock. The consequences of a collapse of the livestock industry would therefore be significant and far-reaching. As farming animals are 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 worldwide, both domestic and wild. Even though it has adapted to birds as the specific host, the virus can further adapt, spillover to humans and in rare cases be transmitted between humans [14]. Avian influenza has caused 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]. This is roughly a third of the national stock of laying hens. 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 HPAI is spreading worldwide. Having started in early 2022, until today, February 23, 2023 **TODO: update numbers** https://efsa.onlinelibrary.wiley.com/doi/abs/10.2903/j.efsa.2023.7786 it has led to

more than 58 million culled or died birds. Different H5 subtypes have been 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 from 2003 to 2022, there were a total of 868 confirmed cases of H5N1 in humans with a mortality rate of 52%, 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 are the advantageous circumstances for virus transmission, since it is warm and humid. In general, animal husbandry practises 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, another factor is transmissibility, determining how easy the infectious agents spread. 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 in farming surroundings.
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 lead to 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 long-term consequences of infection can affect the health of the animals. This can reflect in poor growth, 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 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 [25]. The list does not cover all known zoonoses and animal diseases since not all of them pose an actual risk.
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 and public health, the question is how to detect, monitor, control and prevent outbreaks in farm animal popula-

tions.

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 has its main task in 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 mode, 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 viral diseases 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 disease surveillance include notifiable disease 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 Deoxyribonucleic Acid (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. This is done by constructing the parts or the complete genetic information of a virus, the genome, where the nucleic acids store this information in single or double strands in a linear or circular sequence. With NGS methods, the genome sequence can be precisely determined. 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. Reusing parts of the globally used SARS-CoV-2 pipeline with NGS input data for genomic analysis can help to understand other viruses and ultimately lead to a deeper understanding of viral genomes from isolates.

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 developed pipelines concern avian influenza A for subtype identification and genome analysis, a poxvirus pipeline for determining poxvirus genomes sequenced as half-genomes in a tiled-amplicon approach and Foot-and-mouth Disease (FMD) for serotyping and genome analysis. The poxvirus pipeline has been tested with samples of lumpy skin disease virus. These viruses are chosen for the availability of test samples that were used for validation of the pipelines, for relevance within the ZODIAC project and for their importance concerning animal and public health risk. All three pipelines follow an approach that relies on raw read data and enables monitoring of intra-sample minor allelic variant frequencies. The high resolution that allows early warning of epidemiological signs of a changing viruses, specifically important for the assessment of emerging variants in pathogenicity and vaccine sensitivity.

In summary, the lack of bioinformatics knowledge and resources in poorer countries poses a major challenge to effective, globally integrated viral livestock disease 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, insights into transmission routes and changes of the virus, 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, applications and drawbacks are described below, software tools to use handle NGS data, state of the art in poxvirus and avian influenza virus analysis, and lastly pipelines for genomic analysis are 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 is used to describe newer technologies in the field and 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 applies 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 read length and are able to detect reads without using electrophoresis. Reads are equal to single fragments of DNA or Ribonucleic Acid (RNA). Third-generation sequencing technologies typically generate longer primary

reads of DNA or 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 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]. Error rates of Illumina MiSeq and HiSeq sequencers range from 0.1% to 10% depending on the experiment and platform used [32]. 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.

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 (by 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 rates of 10% to 15% compared to other HTS

methods [35, 36]. This makes ONT less suitable for single-nucleotide variant analysis that is required in some diagnostic applications [37, 38].
Other frequently used second-generation platforms are Roche/454 sequencing, IonTorrent (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 [39].

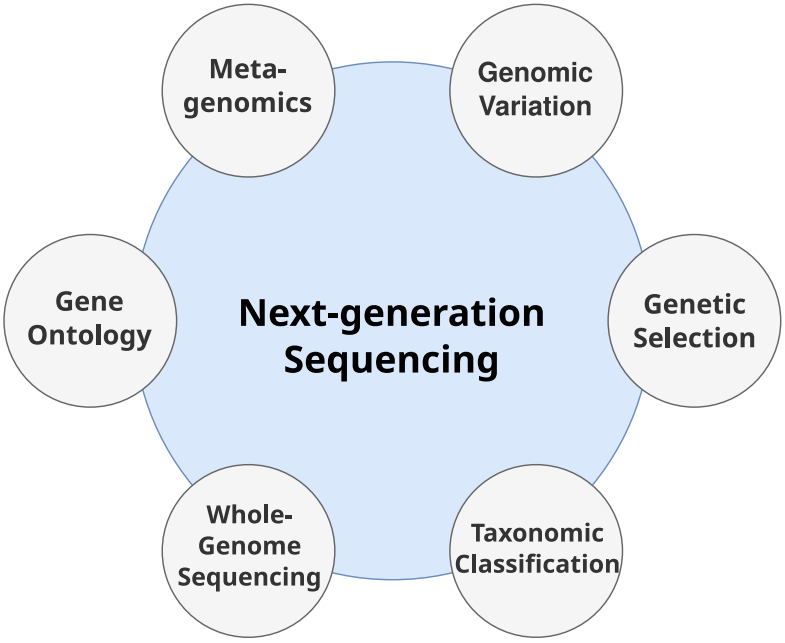

Figure 1: Next-generation sequencing technology applications in virology.

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 or environmental samples [40]. 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 information about

virus virulence [41]. Variant detection is only possible with the sequenced genome, as it provides insight to the genome on a nearly every-base level and allows to reliably interpret and identify the many different possible variants [42].
Genetic selection describes the process by which certain viral strains become more prevalent in a population over time due to selective pressures. In genomics with viral material, 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 and transmission.
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 or pathogenicity [43].
Whole-genome *de novo* assembly is the reconstruction of an entire viral genome without prior knowledge of its genetic sequence, being a costly and time-intense method with potentially high error rates. Similar to metagenomics, whole-genome sequencing and assembly can be used to identify novel viruses, to study mutations in viral genomes and to track the evolution of a virus over time [31]. In contrast, transcriptomic sequencing is used for viral identification of viruses that are actively expressed in their host, and is performed at a fraction of the cost needed for Whole-Genome Sequencing (WGS).

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, hence it is imperative to choose the most appropriate method for each application.
The reconstruction from HTS-generated reads to assemble the full-length genome can be made using different approaches, depending on the preparation of the NGS data and the

objectives. During sequencing, Polymerase Chain Reaction (PCR) amplification on the whole genome is a widely used technique to amplify specific regions of nucleic acids by producing many copies of the targetted sequence. PCR is used to sequence for example specific genes in a viral metagenomic sample. Another approach, developed by the ARTIC network is based on amplicons, i.e. fragments of the genetic sequence that cover the whole genome and are then sequenced on an Illumina platform. Amplicons are generated by tiled primers that start a PCR to generate the amplicons. For each amplicon, two primers for each end of the region of interest are needed, hence the expression tiled or tiling primers. The distance between the primers determines the size of the produced amplicon [44].

2.1.2 Data Analysis Issues

Since the surveillance of viral animal diseases with NGS is advancing rapidly, it is important that regions and health organisations 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 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 if the data are not cleaned. Chimeric products are artefacts originating from joining sequences and are represented by point mutations, insertions and deletions. Chimera formation also occurs during PCR amplification [45].

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. This includes algorithms in steps such as quality filtering, clustering and sequence classification [46]. 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 [47]. Longer reads preclude 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, PCR amplification of the viral material and reverse transcription of viral RNA to coding Deoxyribonucleic Acid (cDNA), it is crucial to include quality checks and filtering steps when using the HTS data [48].
Each application of software used 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. Nevertheless, NGS technologies provide a large pool of methods, although also available algorithms for genome assembly and amplicon analysis have drawbacks and limitations [49].

2.2 Tools for Genomic Analysis with NGS Data

A variety of suites and software packages is available to process NGS-generated data. Depending on the user's research and analysis interest, tools are used independently and/or subsequently. A tool represents a modular program to use with data input from the user. Different suites for bioinformatical analyses offer different interfaces to execute a tool, either on the commandline to work on a server, or via a web-interface. The software of a tool can be complex algorithms and expensive calculations, or simple and

fast formatting programs.
Pursuing the goal to construct the full-length genome, short NGS-sequenced raw reads in FASTQ format, which is the text-based standard format to store nucleotide sequences with numeric quality scores for each nucleotide, serve as the primary input for any analysis steps. For the central steps of the bioinformatics pipelines described in Section 2.4.2, Section 2.5.2, Section 2.3 and importantly the newly designed pipelines in Section 3.3.1, Section 3.3.2 and Section 3.3.3, tools and software suites that can work with NGS-generated data from different techniques are discussed in the following.

Tools for Preprocessing

Working with raw NGS reads requires quality control before executing any further steps. Preprocessing with quality control helps the user to understand the sequencing data and to check its overall quality so sequencing errors, PCR artefacts and contaminations can be detected. Usually, a quality filtering to keep only reads above a certain length and quality threshold is conducted in the preprocessing, as well as when working with ampliconic data, trimming typical sequencing artefacts. The remnants of adapters, artifically introduced during sequencing, need to be removed as they are not part of the transcriptome. Common tools for this purpose are `FastQC`, `Trimmomatic`, `Cutadapt` or `fastp` [50, 51, 52, 53]. Being developed specifically for adapter-trimming of Illumina and SOLiD data, `Cutadapt` is a commandline tool written in Python that at the time of publishing was the only tool to support colour-space data. It also provides some read-filtering options [52]. `Trimmomatic` was developed to solve similar tasks but with higher performance and correct handling of paired-end data. It works with Illumina sequenced data and the user can upload the adapter sequences for adapter trimming if they deviate from standard protocols. It performs quality pruning with a sliding-window cutting algorithm [51]. `FastQC` is a Java-based tool for quality control and provides per-base and per-read quality profiling options [50]. The newest tool for preprocessing `fastp` provides an all-in-one solution for quality control of FASTQ data, which includes

all of the options the formerly mentioned tools provide. Additionally, `fastp` outperforms them in terms of speed by being 2 to 5 times faster [53]. From the point of view of the user who wants to perform all the steps of quality control, filtering and trimming, none of the tools except `fastp` offers all the functions, which slows down the preprocessing because several tools have to be started. Additional features such as unique molecular identifier preprocessing and per-read polyG tail trimming are integrated into `fastp` [53]. Its multithreading implementation in C/C++ makes it much faster than the previously mentioned tools. Reporting of the quality results to compare statistics of the reads before and after the preprocessing run is possible with all tools in combination with `MultiQC` [54].

Tools for Alignment

In order to obtain the full-length sample sequence and to identify Single Nucleotide Polymorphisms (SNPs) in an isolate, short sequenced reads need to be aligned to a reference genome. Assembly of short reads can also be done as *de novo* assembly without a reference, however this approach requires greater computational capacities and greater sequencing depth to ensure a sufficient overlap of the reads for an accurate assembly [55]. The choice of alignment method depends on the amount, length and origin of read data. For reference-based alignment approaches, typically `minimap2` is used with ONT, PacBio or Illumina-sequenced data. `minimap2` is a pairwise aligner for short reads of at least 100 basepairs in length [56]. It states to be faster and more accurate than other domain-specific alignment tools [56].
Other frequently used tools are `BWA-MEM` for Illumina data and `Bowtie2` for ONT data. Like most other full-genome aligners, `BWA-MEM` follows the seed-chain-align pattern [57]. Using a Burrows-Wheeler Transform, both `BWA-MEM` and `Bowtie2` are shown to be faster aligners than others with reads of 100 basepairs in length [58].
For *de novo* assembly with Illumina, PacBio or IonTorrent data, `SPAdes` is used. It is based on a De Bruijn graph algorithm by building k-mers [59]. For ONT or PacBio

data to align long error-prone reads, `Flye` is a modern *de novo* assembler, shown to be highly performant with relatively low error rates [60, 61]. Its underlying algorithm is A-Bruijn assembly graph construction that attempts to generate arbitrary paths with overlaps, unlike other De Bruijn-based assemblers which attempt to generate long accurate contigs [60].

Consensus Sequence Construction

Representing the alignment results in the form of a full-length genome, based on the calculated order of the most frequent residues for each position the consensus sequence is constructed. With aligned Illumina reads, the `iVar` suite provides tools to generate the consensus sequence. Its features also include primer and quality trimming (`iVar trim`) and intrahost variant detection (`iVar variants`) [62]. On the `Geneious` platform, similar analyses can be executed in order to produce a consensus sequence from raw reads[63]. `bcftools` as a tool suite also offers a package for consensus sequence construction from Variant Call Format (VCF) files [64].
For ONT generated data, the `medaka` tools suite provides a module to generate the consensus sequence.

Phylogenetic Tree Construction

Having multiple virus strain samples and wanting to express their relations, evolutionary or phylogenetic trees are a common method to use with nucleotide sequences. Most common tools are `FastTree` and `IQ-Tree`, both based on infering relations using the maximum-likelihood criterion [65, 66]. Other, more inefficient algorithms rely on creating a distance matrix to compute the minimum distances. Both `FastTree` and `IQ-Tree` require a multiple sequence alignment as input data, which can be obtained by multiple sequence aligners like `MAFFT` or `ClustalW` [67, 68]. Generated phylogenetic trees can be visualised to infer topologies and study relationships of taxon-groups, while only the

`IQ-Tree` tool provides an in-built visualisation in *iqtree* format. With the phylogenetic tree in Newick format (*nhx*) and the `PhyloCanvas` tree viewer, trees can be exported, extended and visualised with other tools [69].

Classification of Sequences

Many applications of genomic analysis require the placement of the inferred sequence compared to other, similar sequences. There are many databases available for different categories of sequences, offering database searches to find the closest sequence compared to the given sequence. Basic Local Alignment Search Tool (BLAST) is a popular program to search databases, while there are different heuristics and variations depending on the specific database and search string characterisations [70]. For nucleotide sequences and databases, the National Center for Biotechnology Information (NCBI) provides a web-based `BLAST` search [71, 72]. The underlying algorithm is a similarity measure that "permits a tradeoff between speed and sensitivity" by setting a threshold parameter [72]. Having BLAST classifying full-length genomes, short sequencing reads can be used for a database search, too. Specifically developed for and tested with influenza reads, `VAPOR` is a software tool that infers a scoring based on a De Bruijn graph construction and emits the closest sequences from a database [73]. It directly maps reads to a De Bruijn graph without prior assembly and therefore accelerates the classification search as compared to a BLAST search while still achieving similar or better results [73]. `VAPOR` provides options to finetune the classification run, depending on read length, database size, k-mer size and other measures. It also has the option to output a file with the scoring, generated by a scoring function that favours sequences with a high coverage of the reference and with a high weight in the De Bruijn graph [73].

2.3 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 later 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 [74]. The pipeline is based on the Reflow suite, and all computations run in an Amazon Web Services, Inc. (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. The user chooses from a tool pool of read trimmers, mappers, variant callers and optional downstream analyses. Defaults are `iVar` for trimming, `minimap2` for mapping, `LoFreq` as a variant caller and consensus sequence calling with `bcftools`. 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 [74].

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 [75]. 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. V-Pipe is based on Snakemake as a workflow and dependency manager.

Other freely available pipelines for the analysis of viral genomes from NGS data with several focuses in genomics are VirFind [76] and Integrated Rapid Infectious Disease Analysis (IRIDA) [77]. 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 detect and reconstruct known Viral genomes from Metagenome (drVM) and VirMAP [78, 79]. 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. This holds not only for avian influenza virus samples, but also for isolates containing other viral material. Galaxy workflows covering the above points for Illumina-sequenced data have been developed in this thesis and are described in Chapter 3.

2.4 Poxvirus Analysis

Among the family of poxviruses, there are some diseases that circulate in livestock and pose a risk so that they are on the list of notifiable animal diseases. Among others, mpox, sheeppox and goat pox are the diseases of concern. In the following, characteristics of poxviruses and current approaches to analyse NGS data of poxviruses are described.

2.4.1 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 [80]. The discovery of a vaccine for smallpox made it the first disease to be eradicated by human efforts, so variola was the first human virus to be successfully

eliminated [81]. 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 [80]. 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-studied 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 *Poxviridae* 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 classification has been confirmed by recent comparative genome analysis [82]. 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 generalised pustules in infections with the variola virus. Other general symptoms include fever, headache and rash.

Genus	Virus Species	Natural Hosts
Avipoxvirus	Canarypox virus	Songbirds
	Fowlpox virus	Chickens, turkeys
Capripoxvirus	Sheeppox 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.

Table 1 shows ten representatives of the 18 Chordopoxvirus genera according to the newest International Committee on Taxonomy of Viruses (ICTV) Taxonomy Release from 2021, while at least five genera contain zoonotic poxviruses [83]. 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 [81]. The Chordopoxvirus subfamily is characterised by its large, linear double-stranded genome. Size varies between 134 to 365 kilobases [84, 85]. Chordopoxvirus genomes contain 130 to 328 Open Reading Frames (ORFs), and typically, two identical Inverted Terminal Repeats (ITRs) 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 [86]. For animals, there is a smallpox-based vaccine that is used to protect elephants against cowpox [87]. 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 (LSD) is caused by the lumpy skin disease virus belonging to the *Capripoxvirus* (CaPV) genus within the family of poxviruses, subfamily *Chordopoxvirinae* [88]. 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 sheeppox and goatpox, the LSDV genome is highly similar to the other CaPV genomes [89].
LSDV is not known to be transmissible 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, sheeppox 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 [90]. 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 [91]. From these symptoms alone, it is impossible to differentiate the diagnosis between sheeppox, 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 [92].

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 these regions [93]. 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. Some countries use sheeppox vaccines to protect cattle against LSDV, even though it does not provide complete immunity. Nevertheless they are used in regions where all CaPV are prevalent [94].

In 2017, a novel LSDV was discovered in Russia, calles the Saratov strain [95]. It seems to have arisen through recombination events between field and vaccine strains, which Gershon and his colleagues had predicted much earlier, in 1989, due to the close similarities between Capripoxviruses [96].

2.4.2 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 ITRs 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 ITRs to. However, the reads have to be sequenced in two pools, which is not the standard protocol. These challenges make it difficult to differentiate between LSDV, goatpox and sheeppox [89].

An ampliconic assembly-based approach to distinguish Capripoxviruses is described by Mathijs et al. [97]. They develop a sequencing protocol in two pools to separate the ITR regions. After preprocessing with `Trimmomatic` and `FastQC`, 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 [62, 98, 99, 100].
A pipeline of Zhao et al. was designed to study the whole genome of mpox samples [101]. After preprocessing with `FastQC` and 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, though it did not find the

correct sequence assembly for five cases in a sample sequence of mpox [101]. 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 [102]. 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.

2.5 Avian Influenza Virus Analysis

NGS-based sequencing data from AIV samples need to be thoroughly processed to gain insights into the subtype and variants within the sequence. In the following, the avian influenza pathogen, the avian influenza virus, is described in detail and modern methods in the form of automated pipelines for the analysis of such data are presented.

2.5.1 Avian Influenza Virus

Informally known as bird flu, avian influenza is a viral infectious disease that affects wild birds and poultry. The 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 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, and in the poultry industry in closed confinements. Humans become infected through close contact with infected material,

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 humans 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 and only occur in the influenza A strain, 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, 103]. 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-structural proteins). Genome size differs due to different possible combinations of proteins, though the typical size of a H5N1 genome is 13.5 kilobases. 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 [104]. Both LPAI and HPAI infections have been reported in domestic poultry, i.e. ducks and chickens, turkeys, caged birds, aquatic birds and wild birds. While some influenza species infect specific 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 [105]. 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 [106]. Zoonotic spillover events have become increasingly 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 [107]. A current AIV outbreak is resulting in more than 58 million dead birds and costs of around 661 million U.S. dollars, which began in 2022 and is spreading across the U.S. [108]. 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 [109, 110].

2.5.2 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 [42]. 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 is 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 [111], EpiFLU/GISAID [112], Nextflu [113], NCBI Influenza Virus Resource [114], FluNet [115] and OpenFluDB [116]. 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. “INSide the FLU” (INSaFLU) and Prediction of Avian Influenza Virus Sub-

type (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, provides a web-based protocol for data generated by Illumina, IonTorrent or ONT sequencers [117]. It is the a influenza-focused suite to process NGS data to automatically generate output data and answer key questions in influenza genomic surveillance. INSaFLU can be used for seasonal influenza, avian influenza, SARS-CoV-2, mpox virus, Respiratory Syncytial Virus (RSV) and for unspecified viruses a generic pipeline is provided. The INSaFLU pipeline consists of the following steps: (1) Reads quality analysis and improvement with `FastQC` and `Trimmomatic`, (2a) classification using a *de novo* assembly with `SPAdes` and searching a provided database with `ABRIcate`, (2b) mutation detection and consensus generation with `Prokka` and `Snippy` (using the `Medaka` suite for ONT data), (3a) intra-host minor variant detection using `Freebayes`, (3b) alignment/phylogeny with `FastTree` and the tree visualiser `Phylocanvas` and (3c) coverage analysis with a INSaFLU specific Python script. 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 mpox samples [117]. 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) preprocessing with `FastQC` and `Trimmomatic`, (2a) reference-based alignment with `BWA` or (2b) *de novo* assembly with `IVA`, (3) subtyping using the `samtools` suite, (4) variant calling with `bcftools` and identification of the closest sequences by (5) BLAST+ for nucleotides [118]. 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 the internet, being a web-based platform with additional material only available in Korean [118]. This is a very limiting factor for the usability of PAIVS.

2.6 Foot-and-Mouth Disease Virus Analysis

In the following, the pathogen of foot-and-mouth disease, FMDV, as a severe and highly contagious virual disease is described. It is of great importance to study in the livestock industry and estimated to circulate in 77% of the livestock population in Asia, Africa and the Middle East [119].

2.6.1 Foot-and-Mouth Disease Virus

Cloven hoofed animals such as cattle, swine, sheep and goats are the ruminants affected by FMD. It was the first viral disease the WOAH established a list for with disease-free countries and defined zones, motivated by the huge economic impacts the FMD outbreaks have. There are 40 reported cases of human infections with FMDV, but the virus is not classified as a public health risk by the WOAH [119]. FMD must not be mistaken with hand, foot and mouth disease, which occurs more often in humans.

Infected livestock show clinical symptoms of viremia, fever and lesions mainly in the mouth, tongue and feet [120]. Although infected animals can completely recover from an infection, they are oftentimes culled in order to prevent spreading and avoid production loss. Mortality is high for young calves, piglets and lambs with 20% but lower for adult animals (1%-5%) [119].

The causative virus is a small positive-sense RNA virus genome with a size of 8.3 kilobases, belonging to the Aphthovirus genus in the *Picornaviridae* family [83]. There are seven distinct strains (A, O, C, SAT1, SAT2, SAT3, and Asia1) and all of them are endemic in different regions of the world, for example the SAT strains in Southern African Territories (SAT), serotype C in the Indian sub-continent and Asia1 in southern Asia [121]. Types O and A are broadly distributed in the non-free countries mainly in Africa, southern Asia and South America. Vaccination against the strains exist, although each strain requires a specific vaccine due to the high antigenic heterogenity of the virus even within one serotype.

While there is a huge list of FMD-free countries without, which is all of North and Central America, continental Europe, Australia, New Zealand and Indonesia, there are regions that successfully put effort into the elimination of FMDV using mass vaccination. This is mainly true for Latin America, though there are sporadic outbreaks in Venezuela and Bolivia. FMD is an endemic disease in Asia, Africa and the Middle East [122]. Similar to poxviruses and AIV, FMDV is very difficult to control due to its contagiousness, wide host range, multiple transmission modes and the potential for long-term carrier status in livestock [123].

Efforts with next-generation sequencing data are made to reconstruct the viral genome in order to understand within-host diversity and downstream analyses.

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

Genomic analysis of FMDV samples gives valuable insights, and may help understand transmission routes and mutations. The analyses can vary depending on the specific objective and typically consist of several subsequent steps, starting with sequenced reads from the sample. However, there are no such ready-to-use pipelines available. Protocols that exactly describe the single steps used for genomic analysis are rare and highly depend on the input data.

One protocol by Munir et al. working with Illumina-sequenced data describes a Genome Analysis Toolkit (GATK) 4.0 pipeline to run on a local machine [124]. Its preprocessing step includes quality checks with `MultiQC` and trimming with `fastp`. Mapping is performed using `Bowtie2` and a variant calling step is performed with `Mutect2`. The variants are annotated with `SnpEff`. This protocol does not provide insight into parameters or detailed settings for the single tools.

Another FMDV-specific protocol to analyse NGS data produced using ONT-sequenced data is described by Brown et al. [125]. They compare the resulting consensus sequence with Illumina sequenced data. For this analysis, quality control is performed with `FastQC`, trimming the read ends using `Prinseq-lite` and `sickle` for quality and length filtering. Afterwards, the preprocessed reads are assembled using `IDBA_UD` and a BLASTn search. The resulting contig is used as a reference sequence for mapping with `BWA-MEM`. The final consensus sequence is obtained using `samtools` [125]. Again, this protocol is not a start-to-end pipeline but requires the manual execution of the single tools.

3 Materials and Methods

The challenges in genomic analysis of viral material using NGS raw read data are the major motivation for this thesis. Ready-to-use pipelines that can be executed without deeper biological or bioinformatical knowledge specifically designed for the viral genomes of avian influenza, pox and foot-and-mouth disease are presented below. They run on the Galaxy platform and show that for development of the pipelines, large parts of existing viral genomic analysis pipelines as such for SARS-CoV-2 can be reused and adapted.

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, as all related codebases are open-sourced on GiHub. 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 practises (Findable, Accessible, Interoperable and Reusable) [126].

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. Each program that is available as a Galaxy tool is XML-wrapped to make dependency requirements, parameter and data inputs and other settings possible via the Galaxy web-interface.
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 or enter data for the input fields. The automated subsequent order and execution of tools in a workflow is used for modular, longer analyses that are executed repeatedly. Each user gets 250 GB of disk space to run computations.
Workflows that are available on and accepted by the Intergalactic Workflow Commission (IWC) on GitHub are conform with the community's best practise standards and tested on the latest Galaxy release. Dockstore for availability in the US and WorkflowHub for EU users publish the IWC workflows and guarantee the availability in Docker-based environments [127] and on the workflow collaboratory WorkflowHub [128].

Important contributions of Galaxy, as stated by the Galaxy Community (2022), include Vertebrate Genome Project assembly workflows and research collaborations about SARS-CoV-2. Another toolkit leveraged in Galaxy is Galaxy-ML, a set of tools that form 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.

The Galaxy platform is continuously enhanced, and it still attracts around 2,000 new users every month, indicating its quality and significance. 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. 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 public servers are synchronised in a set of reference tools [126].
The platform serves as a public infrastructure that can be used in many different contexts and by professionals from all fields and backgrounds. It therefore is very suitable for offering publicly available and transparent resources for surveillance of diseases.

3.2 SARS-CoV-2 Workflow and Requirements

The COVID-19 pandemic motivated many researchers to study and develop analysis workflows of SARS-CoV-2 sequencing data. In the IWC repository, there are seven workflows available and ready to use on Galaxy for the different kind of NGS data (ONT/Illumina) and with varying objectives (variant calling/variation reporting/consensus construction). Specifically for Illumina ARTIC reads, a workflow for genomic analysis based on the `iVar` suite has been released [129]. It is conceptually similar to other existing pipelines outside of Galaxy, written in Nextflow, Snakemake and Workflow Description Language (WDL). The workflow for ampliconic Illumina paired-end reads consists of the following steps: (1) read adapters are trimmed with `fastp` and (2) mapped to a reference genome with `BWA-MEM`. The alignment is (3) quality filtered using `Samtools view`, keeping the reads with a minimum length of 20 and only if they are mapped and properly paired. After generating quality and coverage reports,

(4) `iVar trim` is run with the primer scheme to cut out the primers from the filtered alignment. The cleaned alignment file is processed (5) with `iVar consensus` to call the consensus sequence and (6) with `iVar variants` to call variants. The resulting outputs are used for variant annotation, phylogenetic assignment of the outbreak lineages and clade assignment. The structure of the workflow is depicted in Figure 2.

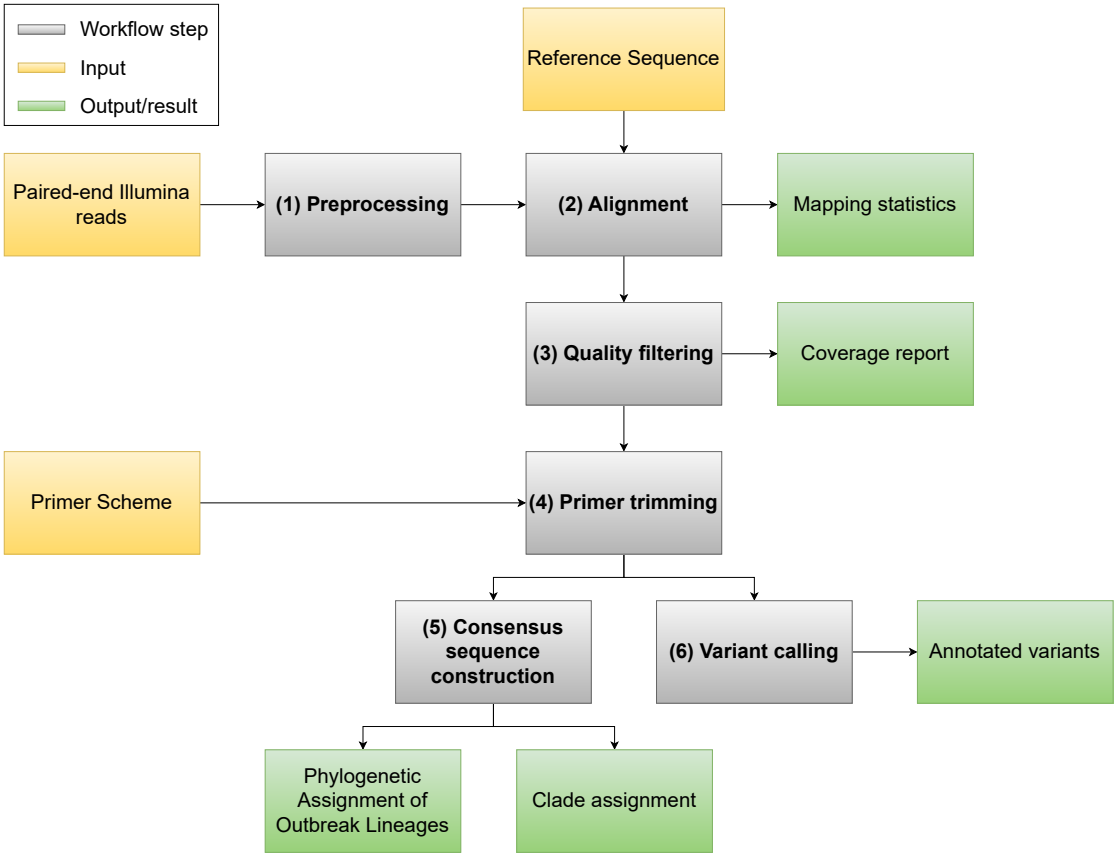

Figure 2: Simplified SARS-CoV-2 analysis workflow for ampliconic Illumina-sequenced data.

This workflow is designed for the specifics of the SARS-CoV-2 genome, however most viral genomes can be analysed in similar ways. Accounting for the genomic structure and composition of each virus, analysis workflows for poxviruses, avian influenza virus and foot-and-mouth disease virus are developed, reusing modified components of the

described SARS-CoV-2 worfklow. The requirements for the viruses and the workflows are described below, before the developed workflows are examined.

Requirements for Poxvirus Analysis Workflow

As explained in Section 2.4, the genome of most poxviruses is bound by identical sequences located at the termini of the genome. It is shown that the size of such differs for some poxviruses, such as rabbitpox and vaccinia virus, while monkeypox, cowpox and capripoxviruses have shorter ITRs [130]. For a whole-genome reconstruction from HTS-generated reads, alignment algorithms look for the unambiguous location of a read. Since this is impossible for repeated identical sequences neither for reference-based mapping approaches nor *de novo* assembly, a new approach has to be used that splits the sequencing reads into two parts, separating the identical sequences and running alignment algorithms for each of the splits. To build the full-length genome, the alignments need to be "glued" together. This approach requires the reads to be sequenced in two pools with two libraries. A similar protocol has been described by Mathijs et al. and the ARTIC network for SARS-CoV-2 data [97, 131].

As a consequence, a requirement for a reference-based surveillance of the genomics of poxviruses is the availability of the primer scheme that was used for amplicon-based sequencing with an Illumina sequencer. The Browser Extensible Data (BED) file containing the primers, their positions and the pool identifier is essential for the correct linking of the alignments when splitting the pipeline into two parts and merging it back together. Apart from the split approach with a masked reference sequence for alignment, the poxvirus reads can be processed in the same way as SARS-CoV-2 reads. In the SARS-CoV-2 workflow, clade and lineage assignment, with `Nextclade` and `Pangolin` respectively, work with SARS-CoV-2 specific databases. Although the tools are designed to work with the SARS-CoV-2 genome, the `Nextclade` tool is adapted and expanded to work with other viruses (mpox, Influenza A H1N1 and H3N2 HA gene, Influenza B Victoria and Yamagata HA) [132].

Requirements for AIV Analysis Workflow

The main objectives of surveillance of AIV on the genetic level are to get phylogenetic insights and to check for new variants that could occur in the HA and NA proteins as a consequence of reassortment.

A pipeline for an avian influenza virus sample that should build a consensus sequence in order to check for mutations needs a reference sequence that it can align the sequence to. A main caveat of many existing pipelines is the user's choice of reference sequence, since it is an arbitrary choice, and it has a direct impact on the alignment. The goal is to pick a reference that is representative of the sample being analysed. In the SARS-CoV-2 pipeline, a reference genome is recommended from a recent strain. For avian influenza virus, multiple reference sequences exist depending on the strain and subtype. Hence, a dynamic approach that is sensitive enough for the segmented structure of the AIV genome is needed to pick a representative reference. The diversity of HA and NA segments' sequences is significant enough to make it challenging to map sequenced reads to a single, full-length influenza A reference sequence. 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. By using a split approach that finds the best reference sequence from a database, the expensive assembly step is avoided and mapping to a suitable reference can be conducted.

Compared to analyses with genomes such as SARS-CoV-2 and due to the segmented structure of the AIV genome, duplicates among the mapped reads of the AIV sample should not be dismissed but kept for maintaining a reasonable high coverage for the further analyses. Downstream analyses for phylogenetic placing are useful for the HA and NA genes, as well as visual summaries of SNPs to identify genetic variation in different regions.

Requirements for FMDV Analysis Workflow

The non-segmented structure of the relatively short FMDV viral genome allows for a straight-forward analysis and reconstruction of the full-length genome by consensus sequence construction. Similar to the workflow for AIV, a dynamic reference search from a database helps to find the best reference genome for alignment. Otherwise, reference-based mapping and consensus sequence construction as well as quality and coverage reports are processed in the same way as in the SARS-CoV-2 workflow. Although the genomes of FMDV and SARS-CoV-2 are of different lengths, both can be analysed in a very similar way. For the developed FMDV workflow, genomic sequencing data are required.

3.3 Workflow Development

Galaxy workflows developed for poxviruses, AIV and FMDV that account for the genomic structure of each virus and the NGS approaches are described below.

3.3.1 Poxvirus Illumina Amplicon Workflow

The proposed Galaxy workflow for poxvirus samples that were Illumina-sequenced with a tiling amplicon approach is available on WorkflowHub (`https://workflowhub.eu/workflows/439`), Dockstore (`https://dockstore.org/workflows/github.com/iwc-workflows/pox-virus-amplicon/main:main?tab=info`) and on IWC to use on Galaxy EU (`https://github.com/galaxyproject/iwc/tree/main/workflows/virology/pox-virus-amplicon`) **TODO: links in appendix or in text?**.
This workflow is the first public pipeline for ampliconic Illumina-sequenced data that provides a ready-to-use infrastructure for genomic analysis of poxviruses with a tiled

amplicon approach. It aims at constructing the full genome from ampliconic Illumina-sequenced reads and providing alignment files, consensus sequence and intermediate results and reports that give insights into reads, mapping quality and mapping coverage. An overview of these outputs and the respective datatypes is provided in Supplementary Table 2. The pipeline is clearly and summarised shown in Figure 3.
To account for the repeated ITRs at the ends of the poxvirus genome, the workflow is based on a tiled amplicon approach. During the first steps, the reads of the two pools from each genome half are processed individually as half genomes. Input data for the workflow are two distinct collections of reads from *pool1* and *pool2*, sourced from the sequencing with two libraries; the used primer scheme in BED file format that contains an indicator for *pool1* or *pool2* in the *SCORE* column, and a reference sequence that is used for mapping.

As a first step, (1) the provided reference sequence is prepared for the mapping of the two read pools. Hence, the primer scheme is needed to determine the exact start and end position of the pools so that the remaining bases are N-masked. For mapping *pool1* against the full-length reference, the second half of the reference sequence is N-masked and therefore the interval for the remaining bases is constructed as a text parameter. The masking starts at the minimal start position of the first primer of *pool2*. If the pools and primers are of similar size, this position is in the middle part of the reference sequence. It is important that this position that separates the pools is between the ITRs so the individual mappings of each pool only contain one ITR. Accordingly for the mapping of *pool2*, the interval of the remaining bases is constructed by taking the maximal end position of the *pool1* primers and the full length of the reference sequence so that the masking of the first half can be conducted. The construction of the text parameter in the correct input format is done by multiple Galaxy-specific text-processing tools and can be looked up in the Supplementary Table 17. Using this approach, it is ensured that the ITRs are unambiguously mapped and coverage statistics are expressive, which would not be the case if mapping would be performed on the full-length reference and reads

Figure 3: Simplified poxvirus genomic analysis workflow for ampliconic Illumina-sequenced data.

from the ITR regions could be mapped to either one ITR.

The workflow is designed to process multiple samples in one run, thus for better comparison the samples of the second pool are sorted by the order of how they are listed in *pool1*. Before mapping, (2) the reads of both pools are preprocessed with `fastp` to automatically trim Illumina-specific PolyG tails of the reads. The following (3) mapping step with `BWA-MEM` takes the corresponding masked reference sequence for each genome-half. A

statistics report for each alignment is generated using `Samtools stats` and allows the user to inspect the mapping quality and coverage. The alignments are (4) filtered for quality with `Samtools view` to keep reads with a minimum length of 20 and only properly paired and mapped reads. Additionally, the pool identifiers (*pool1/pool2*) are prepended to the sample names so that using external software to check variants, the pool and sample identification is maintained and unambiguous. In the next step (5), the two alignments are merged while retaining the identifiers for each sample and pool. For the full-length mapping, a coverage report is generated with `QualiMap BamQC` so that the ITRs and the part where the mappings are merged can be inspected. The mean coverage depth is an important standard parameter when performing NGS. It indicates how often each base occurs on average in the individual reads. For smaller segments or amplicon-based data, checking the depth of coverage in each region is crucial as it provides information on how close the sequenced sample is to the reference sequence selected for mapping. Low coverage indicates incorrect mapping due to many genetic differences. Therefore, coverage plots are given for each sample.
(6) Primer-trimming with `iVar trim` removes the loose primer ends and cleans the alignment for the consensus sequence construction. The (7) consensus sequence is called with `iVar consensus` and a 10-fold minimum depth. For this step, the user can either use provided default settings (minimum quality score to count base: 20, minimum allele frequency threshold to call SNV: 0.7, minimum allele frequence to call indel: 0.8) or enter their own values before starting the workflow. These settings yield for the minimum of ten sequenced reads per base in coverage. With the final combined consensus sequence in FASTA format for each input sample, any downstream analyses can be started.
The workflow with a complete list of the 47 steps, used tools with version number, settings, outputs and connections between the tools is provided in Supplementary Table 17.

3.3.2 AIV Illumina Workflow

We propose a fully automated pipeline for the analysis with a reference-based mapping approach of Illumina-sequenced paired-end reads from avian influenza samples. The workflow is integrated in the Galaxy platform and available via **TODO: link**. Furthermore, to the best of our knowledge this pipeline is the first ready-to-use workflow that uses a hybrid reference sequence for a fast mapping and provides many outputs for various downstream analyses. It is designed to take one input sample at a time and besides a summarising results report, the outputs of the analysis steps can be used for further research based on the user's interest. The workflow is outlined in Figure 4, where the nine main steps of the workflow are visualised. The full workflow of 48 steps with the tools, tool version and settings can be found in Supplementary Table 17.
One novelty 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 provided on Galaxy **TODO: link**, the workflow builds 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 (PB2, PB1, PA, HA, NP, NA, M, and NS), containing multiple full-length sequences per 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 so that the extraction of sequence identifiers in the workflow works as expected: $>segment\_name|influenza\_strain|subtype|accession\_number$. For example, one entry's identifier is $>PB1|A/duck/Manitoba/1953|A/H10N7|KF435047.1$ followed by the sequence in the next line.

After (1) preprocessing of the reads with `fastp` to dismiss reads shorter than 30 basepairs and automatic trimming PolyG tails of the Illumina reads, the database of reference sequences is used to (2) find the closest possible reference for each of the segments. The tool `VAPOR` outputs a table with a scoring based on the weighted graph construction, and should not be confused with the identity of the sequence compared

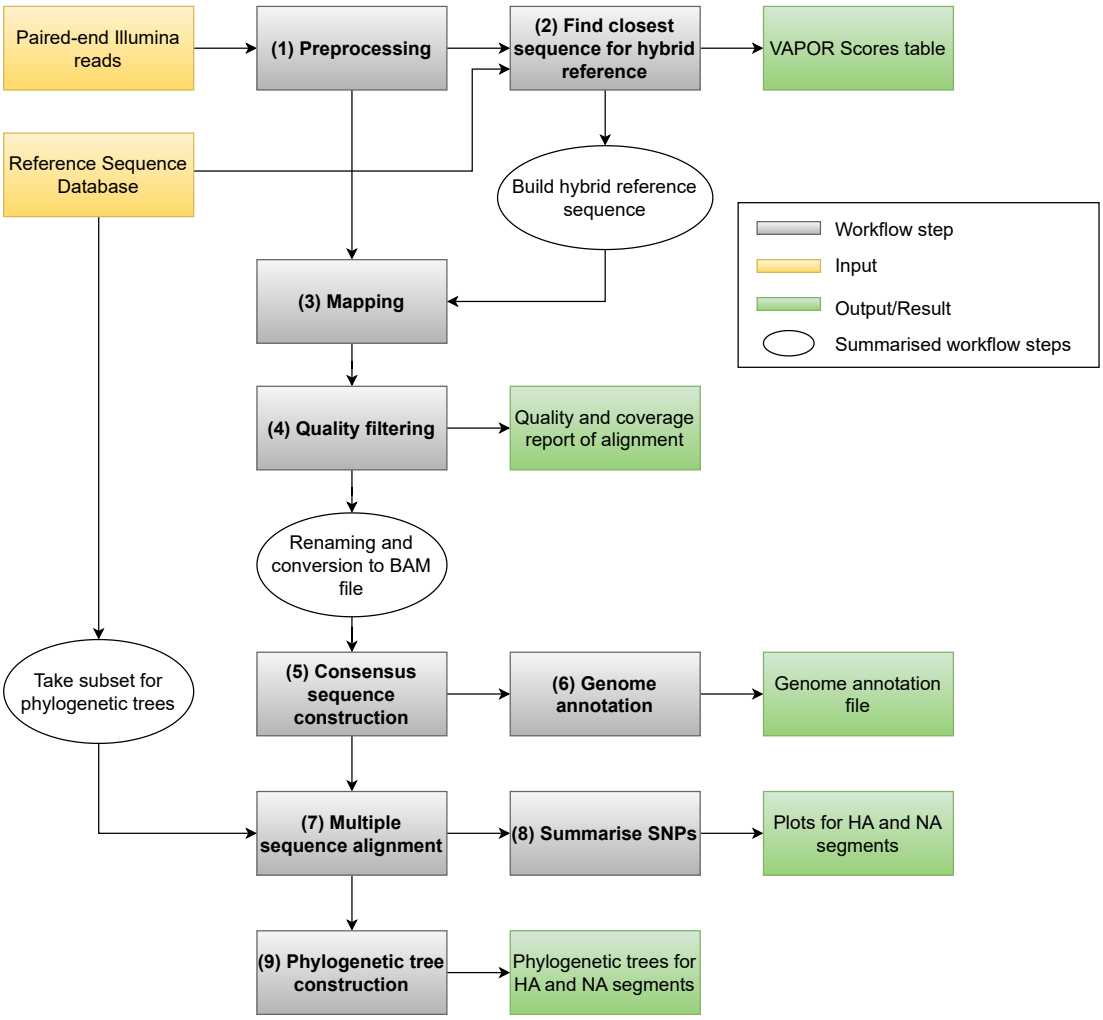

Figure 4: Simplified AIV genomic analysis workflow for Illumina-sequenced data.

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 [73]. Benchmarking shows that it runs significantly faster than BLAST and default configurations lead to reasonable matches similar to Mash, as long as the given sample is not very different from or novel to the provided sequences in the reference database.

Retrieving 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 the third step of the pipeline, (3) mapping with `BWA-MEM`. The segment names in the hybrid reference genome are truncated and shortened to just the segment identifier. Mapping of the preprocessed reads against the prepared hybrid reference is run with default parameters of `BWA-MEM`. The Burrow-Wheeler Aligner for short-read alignment with Maximal Exact Matches (BWA-MEM) algorithm aligns 70-1000 basepairs long reads by seeding alignments with maximal exact matches, and extending the seeds using the affine-gap Smith-Waterman algorithm [57]. After mapping, the resulting Binary Alignment Map (BAM) dataset is (4) quality filtered using `Samtools view`. Reads with a minimum quality of 20 and only those that are paired and mapped in a proper pair are kept. The alignment and quality results as well as coverage statistics for each segment are reported using `QualiMap BamQC`.

The subsequent steps before generating the consensus sequence of the sample prepare the BAM file and deconstruct the mapped reads into a collection of eight datasets and relabel the elements, so that (5) `iVar consensus` can perform consensus sequence construction in parallel. Per-segment consensus construction is run with a minimum quality score threshold of 20, minimum frequency threshold of 0.7, minimum depth to call consensus of 10, which does not exclude regions with smaller depth than the minimum threshold and uses N instead of "-" for regions with less than the minimum coverage. These settings accept any base as the consensus base for a genome position with a base calling quality of 20 or higher in order to avoid false bases that come from sequencing errors. If there is no consensus base to be found with the above thresholds, an N is inserted instead.

The next step using the consensus sequence is (6) generating genome annotation files with `Prokka`. As the input sample is a viral genome, the *Kingdom* parameter is set to *Viruses*. With this file, open reading frames can be predicted using other tools and further downstream analyses can be started.

To place the consensus sequence of the avian influenza segments in a set of samples from the reference sequences to generate phylogenetic data, (7) a multiple sequence

alignment for a user-specified number of sequences (i.e. determines the size of the resulting phylogenetic trees) is conducted with `MAFFT` (Multiple Alignment using Fast Fourier Transform). The consensus sequence is added using `MAFFT add`. The multiple sequence alignment is also used for (8) a visualisation of SNPs, produced with the `snipit` tool. It provides a graphical summary of the variations on base-resolution compared to the reference sequence and other close sequences from the reference database.

As a final step, (9) phylogenetic trees for the HA and NA segments are built using `IQ-Tree`. The taxonomy of the sample segments visualised in the phylogenetic trees give insight into spatial and temporal spread of the genome. The consensus sequence from the input sample is assigned to the most likely lineage [66]. Trees can be explored by downloading one of the standard tree formats (.nhx, .mldist or .iqtree) for flexible downstream analysis or using the Galaxy web-interface.

The presented workflow avoids the computationally expensive *de novo* assembly, instead uses a mapping approach with a dynamically composed reference sequence of close sequences for each of the eight influenza segmets. This accounts for a high quality mapping and is evaluated in Chapter 4.2. To control and look up intermediate outputs, quality reports are emitted during the workflow process and after finishing, which can be downloaded as a Portable Document Format (PDF) for each workflow run.

Due to a variety of possible downstream analyses that can be of the user's interest, the pipeline provides results of the individual steps so that they can be used with other tools. An overview of these outputs with their datatypes is provided in Table 4. Possible downstream analyses are discussed in Chapter 5.

3.3.3 FMDV Illumina Workflow

We propose a mapping-based workflow for the analysis of data from Foot-and-mouth disease virus using Illumina sequencing technology. The workflow takes multiple samples in a collection and is integrated into the Galaxy platform. It is available at **TODO: link** and follows the SARS-CoV-2 workflow pattern. Its full list of produced outputs

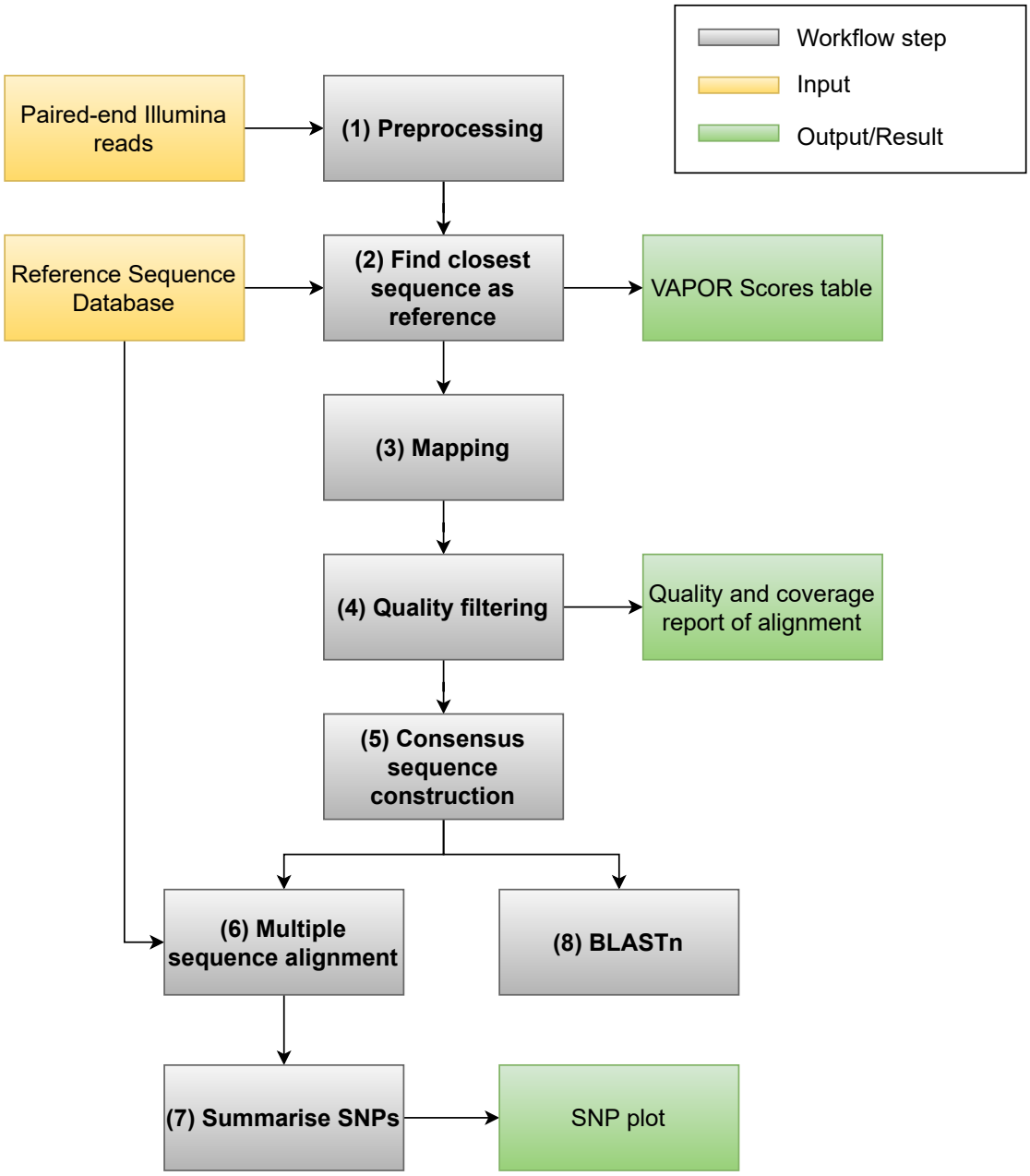

Figure 5: Simplified FMDV genomic analysis workflow for Illumina-sequenced data.

can be found in Supplementary Table 6 and the workflow steps are listed in detail in Supplementary Table 18.

The workflow starts with (1) preprocessing of the reads using `fastp`. Reads shorter

than 30 base pairs are discarded, and PolyG tails of the Illumina reads are trimmed automatically. Additionally, the reads are quality trimmed at the ends of the 5' (15 bases) and 3' (5 bases) ends **TODO: machen wir das?**.

The preprocessed reads are then used to find the closest possible reference sequence using `VAPOR`. As FMDV is not a segmented virus, the reference sequence database provided for this virus consists of one FASTA file with multiple full-length sequences from different serotypes. The provided database file consists of 201 sequences, each containing the complete and N-free genome of FMDV. The highest-scoring sequence from the `VAPOR` run is used as the reference sequence for the alignment. The highest `VAPOR` scores of each run are summarised in a table to check for the quality of the results.

The third step of the workflow is (3) mapping with `BWA-MEM`. Mapping of the preprocessed reads against the reference is run with default parameters of `BWA-MEM`. The resulting BAM dataset is (4) quality filtered using `Samtools view`. Only reads with a minimum quality of 20 and those that are paired and mapped in a proper pair are kept. Alignment and quality reports including coverage statistics are generated per sample using `QualiMap BamQC`.

Similar to the SARS-CoV-2 workflow, (5) consensus sequence construction is run with `iVar consensus` and a minimum quality score threshold of 20, minimum frequency threshold of 0.7, minimum depth to call consensus of 10, and does not exclude regions with smaller depth than the minimum threshold. These settings accept any base as the consensus base for a genome position with a base calling quality of 20 or above.

After constructing the consensus sequence, (6) multiple sequence alignment of a subset of the reference sequences with the consensus sequence is performed using `MAFFT`. The size of the subset can be adapted via user input and defaults to the 7 sequences with the highest `VAPOR` scoring. The Multiple Sequence Alignment (MSA) is then used to visualise SNPs with `snipit`. A (7) BLASTn search is performed against a database of known FMDV serotypes to confirm the serotype of the FMDV sample and preclude contamination, co-infections or recombination. The workflow produces a summary report

of the results of each step and allows the user to investigate the output of each step further for additional research. The workflow is outlined in 5.

4 Results and Workflow Evaluation

The Galaxy workflows are validated using real-world datasets from different laboratories. The analysis results for each workflow with complying test samples are described below.

4.1 Validation of Poxvirus Workflow on Lumpy Skin Disease Virus Datasets

TODO

Public samples by Sciensano by Elisabeth Mathijs

IWC link, primer scheme. tested with LSDV data from 2021

20L70 pool1 (SRR15145276) and pool2 (SRR15145275)

20L81 pool1 (SRR15145274) and pool2 (SRR15145273)

NC_003027.1 as reference sequence (South African Neethling strain. Collection date: 1959 from Bos taurus) -> also used for mostly used vaccine. Other strains are wildtype LSDV and KSGP strains (Accession: PRJNA661421, SRA: SRS7321935)

We emply our pipeline using a tiling amplicon approach with masked references for each half genome.

4.2 Validation of AIV Workflow on H4N6 and H5N8 Samples

TODO

We illustrate the utility of the workflow on the Galaxy platform, we use public real-world datasets

Reference database for each influenza gene segment, prepared the collection from public INSaFLU data. Consists of 56 sequences

within-subtype variation is not captured well by this ref. database (1 reference per subtype)

controlling influenza A/B species assignment: 50 informative references

Samples by Sciensano s4+s8

point out output for downstream analyses

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

4.3 Validation of FMDV Workflow on ? Samples

O/SAT2 samples **TODO**
Samples by Pirbright Institute by Dr. Graham Freimanis (NOT YET)

...wait for samples

4.4 Workflow Profiling

TODO

Assembly vs. mapping

5 Discussion

TODO

5.1 Contribution to the Field

TODO

Workflows that solve common problems, provide useful information, are user-friendly, customisable, extendable

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 ITRs 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 (False Positive Variant (FPV)) must be avoided (happens when too many amplifiation cycles are made)

BWA-MEM vs. BWA-MEM2 for small viral genomes: no big speed gain, error-prone as a new tool. preferably stick to known tools in Galaxy that are proved to work well and have a maintenance

-> caveat of tools with small development teams: abundance of technological advancements or error solving, but dependence on the reliablity -> choose established software

5.2 Future Directions

TODO

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.

* develop workflows for the same purpose that work with other NGS sequencing data (ONT, PacBio etc.)

6 Conclusion

TODO

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.

We show that viral genome analyses can be performed with public scientific infrastructure that is ready to use and based on a community-approved effort in an open-source software. Galaxy wfs can be exported, adapted and used in other systems such as CWL, Nextflow, Snakemake

-> transparent data analyses tool that is robust and transparent to ensure quality and efficiency, all to empower scientists and health professionals to biomedical research

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. Retrieved: 22nd February, 2023.

[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/Shared%20Documents/ZODIAC%20Project%20Document.pdf`, 2021. 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] T. Laver, J. Harrison, P. O'neill, K. Moore, A. Farbos, K. Paszkiewicz, and D. J. Studholme, "Assessing the performance of the oxford nanopore technologies minion," *Biomolecular detection and quantification*, vol. 3, pp. 1–8, 2015.

[37] 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.

[38] 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.

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

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

[41] 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.

[42] 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.

[43] 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.

[44] L. Mamanova, A. J. Coffey, C. E. Scott, I. Kozarewa, E. H. Turner, A. Kumar, E. Howard, J. Shendure, and D. J. Turner, "Target-enrichment strategies for next-generation sequencing," *Nature methods*, vol. 7, no. 2, pp. 111–118, 2010.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] S. Andrews, "Fastqc a quality control tool for high throughput sequence data." http://www.bioinformatics.babraham.ac.uk/projects/fastqc, 2010. Retrieved: 19th March, 2023.

[51] A. M. Bolger, M. Lohse, and B. Usadel, "Trimmomatic: a flexible trimmer for illumina sequence data," *Bioinformatics*, vol. 30, no. 15, pp. 2114–2120, 2014.

[52] M. Martin, "Cutadapt removes adapter sequences from high-throughput sequencing reads," *EMBnet. journal*, vol. 17, no. 1, pp. 10–12, 2011.

[53] S. Chen, Y. Zhou, Y. Chen, and J. Gu, "fastp: an ultra-fast all-in-one fastq preprocessor," *Bioinformatics*, vol. 34, no. 17, pp. i884–i890, 2018.

[54] P. Ewels, M. Magnusson, S. Lundin, and M. Käller, "Multiqc: summarize analysis results for multiple tools and samples in a single report," *Bioinformatics*, vol. 32, no. 19, pp. 3047–3048, 2016.

[55] R. Ekblom and J. B. Wolf, "A field guide to whole-genome sequencing, assembly and annotation," *Evolutionary applications*, vol. 7, no. 9, pp. 1026–1042, 2014.

[56] H. Li, "Minimap2: pairwise alignment for nucleotide sequences," *Bioinformatics*, vol. 34, no. 18, pp. 3094–3100, 2018.

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

[58] I. Borozan, S. N. Watt, and V. Ferretti, "Evaluation of alignment algorithms for discovery and identification of pathogens using rna-seq," *PloS one*, vol. 8, no. 10, p. e76935, 2013.

[59] A. Bankevich, S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham, A. D. Prjibelski, *et al.*, "Spades: a new genome assembly algorithm and its applications to single-cell sequencing," *Journal of computational biology*, vol. 19, no. 5, pp. 455–477, 2012.

[60] M. Kolmogorov, J. Yuan, Y. Lin, and P. A. Pevzner, "Assembly of long, error-prone reads using repeat graphs," *Nature biotechnology*, vol. 37, no. 5, pp. 540–546, 2019.

[61] F. Dida and G. Yi, "Empirical evaluation of methods for de novo genome assembly," *PeerJ Computer Science*, vol. 7, p. e636, 2021.

[62] 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.

[63] M. Kearse, R. Moir, A. Wilson, S. Stones-Havas, M. Cheung, S. Sturrock, S. Buxton, A. Cooper, S. Markowitz, C. Duran, *et al.*, "Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data," *Bioinformatics*, vol. 28, no. 12, pp. 1647–1649, 2012.

[64] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and . G. P. D. P. Subgroup, "The sequence alignment/map format and samtools," *bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009.

[65] M. N. Price, P. S. Dehal, and A. P. Arkin, "Fasttree: computing large minimum evolution trees with profiles instead of a distance matrix," *Molecular biology and evolution*, vol. 26, no. 7, pp. 1641–1650, 2009.

[66] B. Q. Minh, H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. Von Haeseler, and R. Lanfear, "Iq-tree 2: new models and efficient methods for phylogenetic inference in the genomic era," *Molecular biology and evolution*, vol. 37, no. 5, pp. 1530–1534, 2020.

[67] K. Katoh and D. M. Standley, "Mafft multiple sequence alignment software version 7: improvements in performance and usability," *Molecular biology and evolution*, vol. 30, no. 4, pp. 772–780, 2013.

[68] J. D. Thompson, T. J. Gibson, and D. G. Higgins, "Multiple sequence alignment using clustalw and clustalx," *Current protocols in bioinformatics*, no. 1, pp. 2–3, 2003.

[69] K. Abudahab, A. Underwood, B. Taylor, C. Yeats, and D. M. Aanensen, "Phylocanvas. gl: A webgl-powered javascript library for large tree visualisation," 2021.

[70] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of molecular biology*, vol. 215, no. 3, pp. 403–410, 1990.

[71] M. Johnson, I. Zaretskaya, Y. Raytselis, Y. Merezhuk, S. McGinnis, and T. L. Madden, "Ncbi blast: a better web interface," *Nucleic acids research*, vol. 36, no. suppl_2, pp. W5–W9, 2008.

[72] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped blast and psi-blast: a new generation of protein database search programs," *Nucleic acids research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[73] 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.

[74] 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.

[75] 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.

[76] 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.

[77] 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.

[78] 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.

[79] 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.

[80] 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.

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

[82] 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.

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

[84] 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.

[85] 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.

[86] 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.

[87] 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.

[88] 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.

[89] 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.

[90] 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.

[91] 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.

[92] 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.

[93] 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.

[94] 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.

[95] A. Sprygin, Y. Babin, Y. Pestova, S. Kononova, D. B. Wallace, A. Van Schalkwyk, O. Byadovskaya, V. Diev, D. Lozovoy, and A. Kononov, "Analysis and insights into recombination signals in lumpy skin disease virus recovered in the field," *PLoS One*, vol. 13, no. 12, p. e0207480, 2018.

[96] P. D. Gershon, R. Paul Kitching, J. M. Hammond, and D. N. Black, "Poxvirus genetic recombination during natural virus transmission," *Journal of General Virology*, vol. 70, no. 2, pp. 485–489, 1989.

[97] 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.

[98] 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.

[99] 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.

[100] 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.

[101] 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.

[102] 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.

[103] 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.

[104] 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.

[105] 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.

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

[107] 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.

[108] 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.

[109] 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.

[110] 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.

[111] 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.

[112] 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.

[113] 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.

[114] 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.

[115] 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.

[116] 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.

[117] 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.

[118] 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.

[119] World Organisation for Animal Health, "Foot and mouth disease." `https://www.woah.org/en/disease/foot-and-mouth-disease/`, 2023. Retrieved: 23rd March, 2023.

[120] E. Domingo, M. G. Mateu, M. A. Martínez, J. Dopazo, A. Moya, and F. Sobrino, "Genetic variability and antigenic diversity of foot-and-mouth disease virus," *Virus variability, epidemiology and control*, pp. 233–266, 1990.

[121] N. Knowles and A. Samuel, "Molecular epidemiology of foot-and-mouth disease virus," *Virus research*, vol. 91, no. 1, pp. 65–80, 2003.

[122] B. Brito, L. Rodriguez, J. Hammond, J. Pinto, and A. Perez, "Review of the global distribution of foot-and-mouth disease virus from 2007 to 2014," *Transboundary and emerging diseases*, vol. 64, no. 2, pp. 316–332, 2017.

[123] S. M. Firestone, Y. Hayama, R. Bradhurst, T. Yamamoto, T. Tsutsui, and M. A. Stevenson, "Reconstructing foot-and-mouth disease outbreaks: a methods comparison of transmission network models," *Scientific reports*, vol. 9, no. 1, p. 4809, 2019.

[124] A. Munir, A. A. Anjum, I. Altaf, and A. R. Awan, "Whole-genome variants discovery of fmd virus isolated from cattle population in pakistan," 2022.

[125] E. Brown, G. Freimanis, A. E. Shaw, D. L. Horton, S. Gubbins, and D. King, "Characterising foot-and-mouth disease virus in clinical samples using nanopore sequencing," *Frontiers in Veterinary Science*, vol. 8, p. 656256, 2021.

[126] 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.

[127] 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.

[128] 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.

[129] Intergalactic Workflow Commission (IWC), "COVID-19 sequence analysis on Illumina Amplicon PE data." `https://workflowhub.eu/workflows/155`, 2021. Retrieved: 24th March, 2023.

[130] R. Wittek, A. Menna, H. Müller, D. Schümperli, P. Boseley, and R. Wyler, "Inverted terminal repeats in rabbit poxvirus and vaccinia virus DNA," *Journal of Virology*, vol. 28, no. 1, pp. 171–181, 1978.

[131] J. R. Tyson, P. James, D. Stoddart, N. Sparks, A. Wickenhagen, G. Hall, J. H. Choi, H. Lapointe, K. Kamelian, A. D. Smith, *et al.*, "Improvements to the ARTIC multiplex PCR method for SARS-CoV-2 genome sequencing using nanopore," *BioRxiv*, 2020.

[132] I. Aksamentov, C. Roemer, E. B. Hodcroft, and R. A. Neher, "Nextclade: clade assignment, mutation calling and quality control for viral genomes," *Journal of open source software*, vol. 6, no. 67, p. 3773, 2021.

Appendix

Poxvirus Workflow Outputs

No.	Output	From Tool	Datatype	Remark
12	Quality report of pool1/pool2 input reads	fastp	HTML	For web-view
29	Alignment of pool1/pool2 against half-masked reference sequence	Map with BWA-MEM	BAM	Available as SAM file in hidden datasets
32	Statistics report of mapping of pool1/ pool2	Samtools stats	HTML	For web-view
31	Quality filtered alignments of pool1/ pool2	Samtools view	BAM	Available as SAM file in hidden datasets
39	Merged alignments	Samtools merge	BAM	Available as SAM file in hidden datasets
40	Quality and coverage reports of merged alignments	QualiMap BamQC	HTML	For web-view
41	Primer-trimmed aligned reads	iVar trim	BAM	Available as SAM file in hidden datasets
43	Consensus sequences of each sample	iVar consensus	FASTA	List of one FASTA file per sample
47	Combined consensus genomes	Concatenate datasets	FASTA	Consensus sequences of all samples in one file

Table 2: Outputs of the Poxvirus workflow that can be used for downstream analyses.

Output No.	Input No.	Description	Tool	Parameters	Output Datatype
2	-	Upload Primer scheme with pool1/pool2 identifiers in SCORE column	-	-	BED
9	2	Select pool1 primers	Select 1.0.4	that: Matching the pattern: [p|P]ool1	BED
10	2	Select pool2 primers	Select 1.0.4	that: Matching the pattern: [p|P]ool2	BED
14	9	Get end position of pool1	Datamash 1.1.0	Operation to perform on each group: Type: maximum On column: 3	BED
15	10	Get start position of pool2	Datamash 1.1.0	Operation to perform on each group: Type: minimum On column: 2	BED
19	14	Parse integer as text	Parse parameter value 0.1.0	-	text_param
20	15	Parse integer as text	Parse parameter value 0.1.0	-	text_param
1	-	Upload reference sequence	-	-	FASTA
8	1	Get the full length	Compute sequence length 1.0.3	-	tabular
13	8	Get the length value from table	Cut 1.0.2	Cut columns: c2	tabular
18	13	Parse integer as text	Parse parameter value 0.1.0	-	text_param

Output No.	Input No.	Description	Tool	Parameters	Output Datatype
23	19, 18	Build interval for masking the second half of the reference	Compose text parameter value 0.1.1	1: components value from 19 2: components Enter text that should be part of the computed value: - 3: components value from 18	text_param
24	20	Build interval for masking the first half of the reference	Compose text parameter value 0.1.1	1: components Enter text that should be part of the computed value: 1- 2: components value from 20	text_param
27	1, 23	Mask reference for mapping of pool1	maskseq 5.0.0	Regions to mask: value from 23	FASTA
28	1, 24	Mask reference for mapping of pool2	maskseq 5.0.0	Regions to mask: value from 24	FASTA
3	-	Upload paired-end reads of pool1 run	-	-	list of fastq/ fastqsanger pairs
4	-	Upload paired-end reads of pool2 run	-	-	list of fastq/ fastqsanger pairs
11	3	Get the sample names from pool1	Extract element identifiers 0.0.2	-	TXT
17	4, 11	Sort samples of pool 2 according to pool1	Sort collection 1.0.0	Sort type: from file	list of fastq/ fastqsanger pairs
16	11	Place one sample identifier per file	Split file 0.5.0	Select the file type to split: Text files Base name for new files in collection: split_file	TXT

Output No.	Input No.	Description	Tool	Parameters	Output Datatype
21	16	Parse identifier as text	Parse parameter value 0.1.0	-	text_param
25	21	Append _pool1 to the sample identifiers	Compose text parameter value 0.1.1	1: components value from 21 2: components Enter the text that should be part of the computed value: _pool1	text_param
26	21	Append _pool2 to the sample identifiers	Compose text parameter value 0.1.1	1: components value from 21 2: components Enter the text that should be part of the computed value: _pool2	text_param
12	3	Preprocessing of pool1 reads	fastp 0.23.2	Single-end or paired reads: Paired collection PolyG tail trimming: Automatic trimming for Illumina NextSeq/ NovaSeq data	fastq/ fastqsanger, HTML, JSON
22	17	Preprocessing of pool2 reads	fastp 0.23.2	Single-end or paired reads: Paired collection PolyG tail trimming: Automatic trimming for Illumina NextSeq/ NovaSeq data	fastq/ fastqsanger, HTML, JSON
29	27, 12, 25	Mapping of pool1 against half-masked reference sequence, retaining read group identifier	Map with BWA-MEM 0.7.17.2	Single or Paired-end reads: Paired Collection Set read groups information? Set read groups (SAM/BAM specification) Platform/technology used to produce the reads (PL): ILLUMINA	BAM

Output No.	Input No.	Description	Tool	Parameters	Output Datatype
30	28, 22, 26	Mapping of pool2 against half-masked reference sequence, retaining read group identifier	Map with BWA-MEM 0.7.17.2	Single or Paired-end reads: Paired Collection Set read groups information? Set read groups (SAM/BAM specification) Platform/technology used to produce the reads (PL): ILLUMINA	BAM
32	29	Generate statistics of mapping	Samtools stats 2.0.4	-	tabular
34	30	Generate statistics of mapping	Samtools stats 2.0.4	-	tabular
35	12, 32	Aggregate quality reports of pool1 reads and mapping statistics	MultiQC 1.11	1: Results Which tool was used to generate logs? fastp 2: Results Which tool was used to generate logs? Samtools Type of Samtools output? stats	HTML
37	22, 34	Aggregate quality reports of pool1 reads and mapping statistics	MultiQC 1.11	1: Results Which tool was used to generate logs? fastp 2: Results Which tool was used to generate logs? Samtools Type of Samtools output? stats	HTML
31	29	Quality filter the mapped reads of pool1	Samtools view 1.15.1	What would you like to look at? A filtered/subsampled selection of reads Filter by quality: 20 Require that these flags are set: Read is paired, Read is mapped in a proper pair	BAM

Output No.	Input No.	Description	Tool	Parameters	Output Datatype
33	30	Quality filter the mapped reads of pool2	Samtools view 1.15.1	What would you like to look at? A filtered/ subsampled selection of reads Filter by quality: 20 Require that these flags are set: Read is paired, Read is mapped in a proper pair	BAM
36	31, 33	Create paired collection from both pool mappings	Zip collections 1.0.0	Input 1: value from 31 Input 2: value from 33	BAM
38	36	Add rules to distinguish between pool1 and pool2	Apply rules 1.1.0	1. Add column for identifier0. 2. Add column for identifier1. 3. Set columns A and B as List Identifier(s)	BAM
39	38	Merge the alignments of both pools	Samtools merge 1.15.1	Alignments in BAM format: value from 38	BAM
40	39	Generate quality and coverage report of mapping to check for the middle part of the merged pools	QualiMap BamQC 2.2.2d	-	fastq/ fastqsanger, HTML
42	40	Remove failed reads from the dataset	Filter failed datasets 1.0.0	Input Collection: value from 40	fastq/ fastqsanger, HTML
44	42	Place the two datasets from nested collection into a list of reports	Flatten collection 1.0.0	Input Collection: value from 42 Join collection identifiers using: underscore	fastq/ fastqsanger
46	44	Aggregate results from the reports	MultiQC 1.11	1: Results Which tool was used to generate logs? Qualimap	HTML
41	39, 2	Trim the aligned reads to remove primers	ivar trim 1.3.1	BED file with primer sequences and positions: value from 39 Filter reads based on amplicon info: Yes, drop reads that extend beyond amplicon boundaries Include reads not ending in any primer binding sites? Yes	BAM

Output No.	Input No.	Description	Tool	Parameters	Output Datatype
5	-	Enter minimum quality score to call base	-	-	integer
6	-	Enter allele frequency to call SNV	-	-	float
7	-	Enter allele frequence to call indel	-	-	float
43	41	Consensus sequence construction from trimmed and merged alignments	ivar consensus 1.3.2	Minimum quality score threshold to count base: value from 5 Minimum frequency threshold: value from 6 Minimum indel frequency threshold: value from 7 Minimum depth to call consensus: 50 How to represent positions with coverage less than the minimum depth threshold: Represent as N	FASTA
45	43	Relabel consensus sequences per sample	Text transformation 1.1.1	File to process: value from 43 SED Program: /^>/s/Consensus_(.*) _threshold_.*/\1	FASTA
47	45	Get combined consensus genomes in a multifasta file	Concatenate datasets 0.1.1	Datasets to concatenate: value from 45	FASTA

Table 3: The full Poxvirus workflow with tools, parameters and input/output connections.

No.	Output	From Tool	Datatype	Remark
6	Quality report of input reads	fastp	HTML	In report
8	Scores of closest references	VAPOR	tabular	One table per segment
24	Scores of sequences chosen for hybrid reference sequence	VAPOR	tabular	Overview of scores
25	Alignment of reads against hybrid reference sequence	Map with BWA-MEM	BAM	Available as SAM file in hidden datasets
27	Quality filtered alignments	Samtools view	BAM	Available as SAM file in hidden datasets
31	Quality and coverage report on alignment	QualiMap BamQC	HTML	In report
36	Consensus sequence	iVar consensus	FASTA	One FASTA file per segment
41	Consensus sequences with segment names as sequence identifiers	Collapse Collection	FASTA	Contains eight sequences
42	Plots to visualise SNPs	snipit	PNG	HA and NA plots in report
44	Genome annotation file	Prokka	FAA	In report
43	Phylogenetic trees	IQ-Tree	iqtree	HA and NA trees in report

Table 4: Outputs of the Avian Influenza Virus workflow that can be used for downstream analyses.

Output No.	Input No.	Description	Tool	Parameters	Output Datatype
1	-	Upload reference sequence database per segment	-	-	list
2	-	Upload paired-end Illumina reads	-	-	pair of fastq/ fastqsanger
6	2	Preprocessing	fastp 0.20.1	Single-end or paired reads: Paired Collection Length required: 30 PolyG tail trimming: Automatic trimming for IlluminaNextSeq/NovaSeq data Cut by quality in front (5'): Yes Cut by quality in tail (3'): Yes Cutting mean quality: 30	pair of fastq/ fastqsanger, HTML JSON
8	1, 6	Find closest reference per segment	VAPOR 1.0.2	Type of sequencing data: Paired-end as collection Desired output: Return scores of best matches Limit number of reported matches to: 0 Kmer Length: 21 Read kmer filtering threshold: 0.1 Coverage threshold for k-mer culling: 5 Minimum k-mer proportion: 0.0 Fraction of best seeds to extend: 1.0	tabular
9	8	Get sequence identifier	Replace 1.1.4	Find pattern: ^.+\t>()$ Replace with: $1 Find-Pattern is a regular expression: Yes Replace all occurences of the pattern: Yes Find and Replace text in: entire line	tabular
12	9	Get the first identifier of each segment	Select first 1.0.2	Select first: 1	tabular
16	1, 12	Extract sequences from the database according to the identifiers	seqtk_subseq 1.3.1	Select source of sequence choices: FASTA/Q ID list	FASTA
20	16	Place all closest reference sequences into one FASTA file	Collapse Collection 5.1.0	-	FASTA
23	20	Shorten the identifiers	Replace 1.1.4	Find pattern: >([^|]+).+$ Replace with: >$1 Find-Pattern is a regular expression: Yes Replace all occurences of the pattern: Yes Find and Replace text in: entire line	FASTA
25	6, 23	Mapping against hybrid reference sequence	Map with BWA-MEM 0.7.17.2	Single or Paired-end reads: Paired Collection Select analysis mode: 1.Simple Illumina mode BAM sorting mode: Sort by chromosomal coordinates	BAM

Output No.	Input No.	Description	Tool	Parameters	Output Datatype
10	8	Pick columns with sequence identifier and score	Cut 1.0.2	Cut columns: c6,c1	tabular
14	10	Get the first identifier of each segment	Select first 1.0.2	Select first: 1	tabular
18	14	Place all closest reference sequences in one table	Collapse Collection 5.1.0	-	tabular
3	-	Generate a text file with names “segment 1”, “segment 2” etc.	Create text file 1.1.0	1: selection Characters to insert: segment 1 Specify the number of iterations by: User defined number How many times? 1 (repeat until segment 8)	TXT
22	3, 18	Build table with segment name, sequence identifier and score	Paste 1.0.0	-	tabular
24	22	Reorder the columns	Cut 1.0.2	Cut columns: c3,c1,c2	tabular
27	25	Quality filter the mapped reads	Samtools view 1.9	What would you like to look at? A filtered/ subsampled selection of reads Filter by quality: 20 Require that these flags are set: Read is paired, Read is mapped in a proper pair	BAM
29	27	Generate coverage and quality report of alignment	QualiMap BamQC 2.2.2d	Skip duplicate reads: Unselect all Number of bins to use in across-reference plots: 40	SAM
31	29	Generate HTML report	MultiQC 1.9	Which tool was used to generate logs? Qualimap (BamQC or RNASeq output)	HTML
26	25	Convert header to SAM format	Samtools view 1.9	What would you like to look at? Just the input header (-H) What would you like to have reported? The header in ... Output format: SAM	SAM
28	26	Get header lines starting with @SQ	Select 1.0.4	that: Matching the pattern: ^@SQ.+	SAM
30	28	Rewrite segment names and get subtype information from identifier	Replace 1.1.4	Find pattern: ^\@SQ\tSN:(.*)\tLN:([0-9]+) Replace with: $1@$2 Find-Pattern is a regular expression: Yes Replace all occurences of the pattern: Yes Find and Replace text in: entire line	SAM

Output No.	Input No.	Description	Tool	Parameters	Output Datatype
32	30	Place the alignment of one segment in one file	Split file 0.5.0	Select the file type to split: Text files Specify number of output files or number of records per file? Number of records per file Chunk size: 1 Base name for new files in collection: split_file Method to allocate records to new files: Alternate output files	SAM
33	30	Get columns with sequence identifiers	Cut 1.0.2	Cut columns: c1	tabular
34	27, 32	Filter by mapped intervals	Samtools view 1.9	What would you like to look at? A filtered/ subsampled selection of reads Filter by regions: Regions from BED file	BAM
35	34, 33	Rename sequence identifiers to segment names	Relabel identifiers 1.0.0	How should the new labels be specified? Using lines in a simple text file.	BAM
36	35	Consensus sequence construction	ivar consensus 1.3.1	Minimum quality score threshold to count base: 20 Minimum frequency threshold: 0.7 Minimum depth to call consensus: 10 Exclude regions with smaller depth than the minimum threshold: No Use N instead of - for regions with less than minimum coverage: Yes	FASTA
38	36	Use readable sequence identifiers	Replace 1.1.4	Find pattern: ^>Consensus_(.*)_threshold_.* Replace with: >$1 Find-Pattern is a regular expression: Yes Replace all occurences of the pattern: Yes Find and Replace text in: entire line	FASTA
41	38	Create multifasta file containing the eight consensus sequences	Collapse Collection 5.1.0	-	FASTA
44	41	Create genome annotation file from consensus sequence	Prokka 1.14.6	Kingdom: Viruses	FAA
11	9	Get identifiers of the 10 best scores per segment	Select first 1.0.2	Select first: 10	tabular
15	1, 11	Retrieve sequences of the 10 best scores per segment from reference database	seqtk_subseq 1.3.1	Select source of sequence choices: FASTA/Q ID list	tabular
19	15	Multiple sequence alignment of the 10 best scores per segment	MAFFT 7.508	Data type: Nucleic acids Matrix selection: No matrix	FASTA
4	-	Set size of phylogenetic trees	-	-	integer
13	4, 9	Select the first X lines from the VAPOR scores table	Select first 1.0.2	-	tabular

Output No.	Input No.	Description	Tool	Parameters	Output Datatype
5	1	Dismiss B strain if present in the reference sequence database	Filter FASTA 2.3	Criteria for filtering on the headers: Regular expression on the headers Regular expression pattern the header should match: \>\S*|A	FASTA
7	5	Dismiss H17 and H18 subtypes if present in the reference sequence database	Filter FASTA 2.3	Criteria for filtering on the headers: Regular expression on the headers Regular expression pattern the header should match: H1N|H2|H3|H4|H5|H6|H7|H8|H9|H10| H11|H12|H13|H14|H15|H16	FASTA
17	7, 13	Retrieve the sequences for the phylogenetic trees from the filtered reference sequence database	seqtk_subseq 1.3.1	Select source of sequence choices: FASTA/Q ID list	FASTA
21	17	Multiple sequence alignment of the sequences for phylogentic trees	MAFFT 7.508	Data type: Nucleic acids Matrix selection: No matrix	FASTA
37	36	Rename sequence identifiers from consensus sequences	Replace 1.1.4	Find pattern: ^>.+ Replace with: >sequenced_sample Find-Pattern is a regular expression: Yes Replace all occurences of the pattern: Yes Find and Replace text in: entire line	FASTA
39	37, 19	Add consensus sequences to the alignment	MAFFT add 7.508	What do you want to add to the alignment: A single sequence Preserve the original alignment: Yes Preserve the original order of sequences: Yes	FASTA
42	39	Summarise SNPs relative to the reference sequence of each segment	snipit 1.0.7	The reference sequence ...: should be picked via its ID ID of reference sequence: sequenced_sample Order of samples in the plot: Sort by number of mutations Invert sort order: Yes Flip plot orientation: Yes	Collection of PNG
45	42	Get SNP plot for HA segment	Extract dataset 1.0.1	How should a dataset be selected? Select by index Element index: 3	PNG
46	42	Get SNP plot for NA segment	Extract dataset 1.0.1	How should a dataset be selected? Select by index Element index: 5	PNG
40	38, 21	Add relabeled consensus sequences to the alignment	MAFFT add 7.508	What do you want to add to the alignment: A single sequence Preserve the original alignment: Yes	FASTA

Output No.	Input No.	Description	Tool	Parameters	Output Datatype
43	40	Build phylogenetic trees for each segment from the specific MSAs	IQ-TREE 2.1.2	Specify sequence type: DNA	nhx, mldist, iqtree
47	43	Get phylogenetic tree for HA segment	Extract dataset 1.0.1	How should a dataset be selected? Select by index Element index: 3	iqtree
48	43	Get phylogenetic tree for NA segment	Extract dataset 1.0.1	How should a dataset be selected? Select by index Element index: 5	iqtree

Table 5: The full AIV workflow with tools, parameters and input/output connections.

No.	Output	From Tool	Datatype	Remark
4	Quality report of preprocessed input reads	fastp	HTML	Trimmed and quality filtered
5	Scores of closest references	VAPOR	tabular	
13	Alignment	Map with BWA-MEM	BAM	
24	Quality filtered alignment	Samtools view/ MultiQC	HTML	Failed datasets are filtered and not included in the report
18	Consensus sequences	iVar consensus	FASTA	One FASTA file per sample
21	Search results of consensus sequence query in NCBI nucleotides database	NCBI BLAST+ blastn	tabular	
23	Plots to visualise SNPs	snipit	PNG	One plot per sample

Table 6: Outputs of the Foot-and-mouth disease virus workflow that can be used for downstream analyses.

TODO

for FMDV, not AIV

Output No.	Input No.	Description	Tool	Parameters	Output Datatype

Table 7: The full FMDV workflow with tools, parameters and input/output connections.