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

The need for comprehensive surveillance and detection of viral diseases in livestock has led to the use of Next-Generation Sequencing (NGS) data analysis pipelines for viruses such as Avian Influenza Virus (AIV), poxviruses and Foot-and-mouth Disease Virus (FMDV). To achieve accurate analysis of isolates at the molecular level, we developed three workflows on the Galaxy platform that use different approaches of reference-based mapping with an automatically compiled hybrid reference sequence to enable rapid and informative genetic monitoring of viral origins, relationships and structures. The pipelines are based on components of globally deployed SARS-CoV-2 Galaxy workflows with Illumina-sequenced input data and are characterised by avoiding computationally intensive *de novo* assembly and instead integrating reference sequence selection into the pipelines. We show that mapping-based pipelines can generate full-length consensus genomes useful for downstream tasks such as phylogenetic context analysis and mutation detection without requiring the user to have domain-specific knowledge for the selection of qualified reference. While for short viral genomes like FMDV we use a split method that integrates assembly in the reference selection process, we show that a fix reference for Capripoxviruses is sufficient for high quality mapping and consensus sequence construction. By providing ready-to-use Galaxy workflows that allow professionals in the field to perform viral genome analyses for poxviruses, AIV and FMDV we can expand our knowledge of disease outbreaks and ultimately deepen our understanding of viral genomes from animal isolates.

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

cDNA coding Deoxyribonucleic Acid

CWL Common Workflow Language

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

GISAID Global Initiative on Sharing All Influenza Data

GTPV Goatpox Virus

HA Hemagglutinin

HPAI Highly Pathogenic Avian Influenza

HTS High-Throughput Sequencing

IAEA International Atomic Energy Agency

ICTV International Committee on Taxonomy of Viruses

IGV Integrative Genomics Viewer

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

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

RNA Ribonucleic Acid

RSV Respiratory Syncytial Virus

SAM Sequence Alignment Map

SARS-CoV-2 Severe Acute Respiratory Syndrome Coronavirus 2

SAM Sequence Alignment Map

SAT Southern African Territories

SMRT Single Molecule Real-Time Sequencing

SNP Single Nucleotide Polymorphism

SNV Single Nucleotide Variant

SPPV Sheeppox Virus

UTR Untranslated Region

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 762.79 million reported cases and more than 6.89 million confirmed deaths as of 13$^{\text{th}}$ April, 2023, 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, such as public health specialists, researchers, biomedical staff, bioinformaticians and veterinarians are carefully monitoring potentially dangerous viral diseases by examining the viral genome. 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. Modern and high-resolution monitoring of viral diseases include the sequencing of samples with

NGS platforms and inspecting the viral genome on a base-by-base level. In order to generate the full-length genome sequence in a quality that reliably constitutes the viral genome from the sample, bioinformatic steps are required. Since the motivation for the genomic analysis is given by the large impact and importance in surveillance, these topics 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].

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 Census 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 [4]. 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.

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%) [5]. 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 [6, 7]. Since the first detection of Highly Pathogenic Avian Influenza (HPAI) of the H5 subtype 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 [8]. Avian influenza has caused recent seasonal outbreaks, such as the outbreak in 2014 and 2015 in the United States resulting in almost 50 million birds that died as a consequence of an infection or of depopulation [9]. 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 [10]. It mainly affected farmed ducks due to the high density of animals in the facilities and the separation from wild birds due to domestication [10]. The latest outbreak of HPAI is spreading worldwide. Having started in early 2022, 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 [11]. 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 [12]. The first fatal case of a H3 subtype has been confirmed by the World Health Organization (WHO) just on 11$^{\text{th}}$ April 2023, while the infected woman was only the third known case of H3N8 [13].

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. 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 [14]. 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.

This can reflect in poor growth, production and feed conversion. Another impact of depopulating infected animal populations is the loss of biodiversity within endangered species of wildlife populations [15, 16, 17].

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 transmissible, such as Foot-and-mouth disease, lumpy skin disease, peste des petits ruminants, classical swine fever, highly pathogenic avian influenza and Newcastle disease [18].
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 [18]. 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 populations.
To avoid the impact that a disease outbreak can have, the best method is to avoid the disease in the first place. This leads to the principle of prevention, which 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 [19]. 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 [19].

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. To gain valuable insights into the origin and characteristics of a viral genome, samples from infected animals are sequenced with modern NGS methods to derive information from the reconstructed whole-length viral genome. The 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 [19, 20]. Since efforts in tackling viral disease outbreaks for example by mass vaccination are very expensive, official budgets from the governments are required. 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 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 revolutionising the field on the molecular level [21]. 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. Algorithms and analyses of the viral genome of concern are extensively studied and developed for many different purposes. 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) Strengthening 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 [22]. In collaboration with technical experts from different fields 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. 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 allows early warning of epidemiological signs of a changing viruses, specifically important for the assessment of emerging variants in pathogenicity and vaccine sensitivity. The aim is to provide fast, sensitive pipelines that are ready-to-use for surveillance purposes and rely on concepts from SARS-CoV-2 research.

In summary, the lack of bioinformatics resources in many countries poses a major challenge to effective viral livestock disease surveillance. 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 Next-Generation Sequencing

NGS has revolutionised the field of genomics and virology. With its ability to rapidly and cost-effectively sequence large amounts of DNA or Ribonucleic Acid (RNA), NGS has enabled researchers to explore genetic and epigenetic variations, identify novel genomic features and investigate the evolution and diversity of viruses. In the following section, applications of NGS in genomics and virology are discussed, including the different NGS platforms available for sequencing. Finally, pitfalls associated with NGS, software and techniques are examined.

2.1.1 Sequencing Approaches

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 Sanger sequencing method from 1977 using radioactivity and gels based on chain termination, second-generation sequencers are advancements of Sanger sequencing that apply sequencing by synthesis [23]. However, equipment is comparably expensive and samples require a high concentration of the DNA or RNA to be sequenced. In second-generation methods, reactions run in parallel and drastically reduce overall costs compared to Sanger sequencing. They produce short sequence read lengths and are able to detect reads without using electrophoresis. Reads are equal to single fragments of DNA or RNA. Third-generation sequencing technologies typically generate longer primary reads of DNA or RNA molecules while maintaining the massive speed and parallelism of the technology [24].

Not only sequencing methods themselves differ in terms of underlying technologies, but also the overall approach how to handle samples in genomics research can be different. Ampliconic and genomic sequencing represent two alternatives in sequencing that can be applied on the varying NGS platforms. Ampliconic sequencing is a targeted sequencing approach that amplifies specific regions of a genome by using Polymerase Chain Reaction (PCR) prior to sequencing. This method is particularly useful for studying very similar sequences in order to detect genetic variations within specific genomic regions. The ARTIC network, a collaborative group of researchers, contributed to the development of ampliconic sequencing for viral genomes, particularly for the detection and analysis of SARS-CoV-2 [25, 26]. The ARTIC protocol uses a specific approach, tiled amplicon sequencing, which involves designing primers that amplify overlapping regions of the target genome. The primers are designed to have a fixed length and to be

located at a fixed distance from each other, allowing for complete coverage of the target region [27]. The amplicons can cover only specific regions of the genome but depending on the primer design also cover the full-length genome. At the same time, genomic sequencing is a Whole-Genome Sequencing (WGS) method that sequences the entire genome without prior amplification. This method allows for a comprehensive analysis of the entire genome, including detection of genomic variations, structural variations, and epigenetic modifications [25].

2.1.2 NGS Platforms and Applications

By far the biggest player in the field of high-throughput sequencing of DNA or RNA samples is the Illumina platform [28]. 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 fluorescence allows for faster detection through direct imaging. All DNA and RNA is converted to coding Deoxyribonucleic Acid (cDNA) before sequencing, because Illumina sequencing requires it to be attached to a solid surface which is done by attaching adapters to the cDNA fragments. 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, e.g. MiSeq sequencer for smaller-scale sequencing applications and HiSeq sequencer for whole-genome sequencing [23, 24, 28]. Error rates of Illumina MiSeq and HiSeq sequencers range from 0.1% to 1% depending on the experiment and platform used [28]. 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 further analyses.

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 [29]. 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 [29, 30]. The cyclic mode of sequencing used in second-generation approaches is replaced by sequencing in real-time with user-defined read lengths of up to 10,000 bp (basepairs) [29]. The real-time aspect allows for rapid samples analysis in various settings. Despite its advantages, the main pitfall of ONT is its relatively high error rates in single reads of 5% to 15% compared to other HTS methods [31, 32]. Due to the real-time nature and avoidance of PCR, the output quality highly depends on the quality and purity of the input sample. This makes ONT alone less suitable for single-nucleotide variant analysis that is required in some diagnostic applications [33, 34].
Other second-generation platforms are Roche/454 pyrosequencing, IonTorrent (Thermo Fisher) technology and SOLiD (Sequencing by Oligonucleotide Ligation and Detection). Third-generation platforms include Single Molecule Real-Time Sequencing (SMRT) by Pacific Biosciences (PacBio), nanopore sequencing by ONT and MGI sequencing [35].

As NGS platforms are widely used in biomedical and clinical contexts, some of the most important applications are depicted in Figure 1. Using different tools, hardware and software, samples are prepared for sequencing on an application-specific and available platform and the produced reads can be used for a variety of applications. In virology, metagenomics can be used to identify viruses in complex clinical or environmental samples that contain lots of different nucleic acids [36]. 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, with the aim to identify the presence of viruses and other microorganisms or microbes. This approach is also known as “shotgun sequencing” in diagnostics, infection surveillance and the discovery of mutations and pathogens.

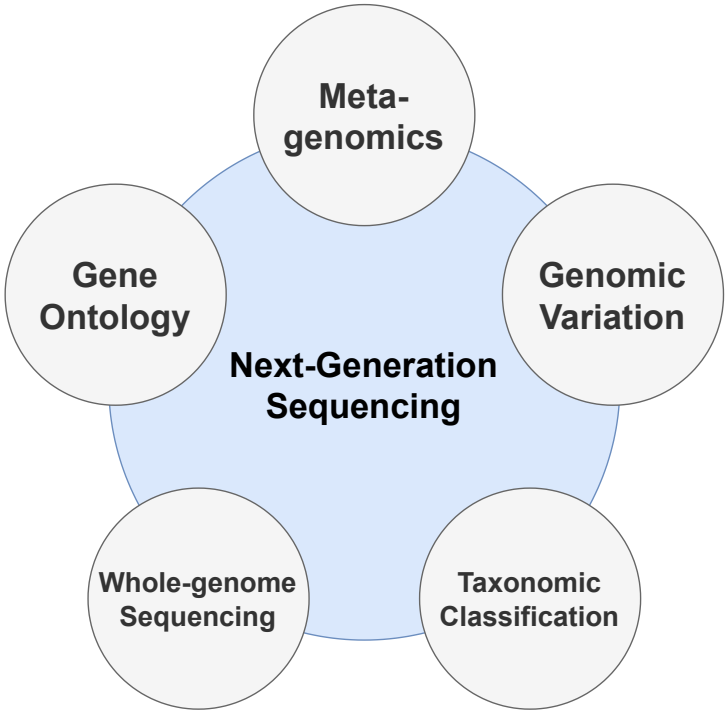


Figure 1: Applications of next-generation sequencing in different fields.

NGS produces huge amounts of data, which can be used to detect genetic variations in samples of clinical patients. Genetic variation refers to finding differences in the form of variations and associating them with particular phenotypes. Genetic variation data of patients are used for personalised treatments or complex disorders. In virology, it 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 [37]. Variant detection provides insight to the genome on an every-base level and allows to reliably interpret and identify the many different possible variants [38]. Specific regions can be sequenced from a sample which is also known as targeted sequencing. NGS data can help identify genetic variants that contribute to genomic variations, which can include differences in size, rearrangements, or epigenetic modifications in a nucleic acid sequence. By detecting these genetic variants, NGS reads provide insight into the underlying causes of complex diseases or the evolution and virulence of viruses.

Taxonomic classification describes the placement of samples to existing profiles based on their genetic and structural characteristics. It is a form of clustering analysis and is done by assigning a rank to each read so that a profile is obtained which is used to place it in taxa-specific entities. These entities label and group the sequences. Traditional general-purpose classifiers in microbiology at the genus level are Kraken, Kraken2 and Basic Local Alignment Search Tool (BLAST) [39, 40, 41]. Taxonomic classification is used for the type identification of a virus causing infections and determination of its potential for transmission or pathogenicity, but also in microbial diversity studies and diagnosis [42].

Another important application of NGS is the whole-genome sequencing, allowing a base-by-base view of the sequenced genome. With this approach, large and small variants can be detected. It helps to analyse entire genomes. Since NGS methods don't produce the full-length genome at once but generate many reads, the sequencing data are processed and assembled to the complete genome sequence. This can be done by *de novo* assembly without prior knowledge of the genetic origin of the sequenced sample. Assembly software tools generate overlaps of reads and reconstruct the genome without a reference sequence. Alternatively, a reference sequence can be used to map the reads to. WGS is used to identify novel viruses, to study mutations in viral genomes and to track the evolution of a virus over time by checking variations on base-level [24]. It also provides detailed understanding of the genetic makeup of viral pathogens and can help to identify regions that may contribute to pathogenicity or resistance to vaccines.

Lastly, gene ontology is applied to NGS data in the analysis of genes and their products. It is used to annotate genes and their functions and to compare the gene expression profiles across biological systems. With this approach, it is possible to gain insights to biological processes and molecular mechanisms. With viral samples, it helps to identify pathways that are affected by different disease states, and also to identify biomarkers for diagnosis, treatment and prevention of diseases.

2.1.3 Data Analysis Issues

Since the surveillance of viral animal diseases with NGS is advancing rapidly, it is important that 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, starting at sample extraction of potentially infected animals to the detailed analysis and derived actions [22].

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 [43].
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 [44]. 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 [45]. 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 directly sequencing longer reads. 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 cDNA, it is crucial to include quality checks and filtering steps when using the HTS data [46].
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 available algorithms for genome assembly and amplicon analysis have drawbacks and limitations [47].

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 bioinformatic analyses offer different interfaces to execute a tool, either on the command-line 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 from raw reads. For the central steps of the bioinformatics pipelines described in Section 2.3, 2.4.2, 2.5.2, 2.6.2 and importantly the proposed pipelines in Section 3.4.1, 3.4.2 and 3.4.3, tools and software suites that can work with NGS-generated data from different techniques are discussed in the following.

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 that 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 part of preprocessing, as well as when working with ampliconic data, trimming typical sequencing artefacts. The remnants of adapters, artificially 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` [48, 49, 50, 51]. Being developed specifically for adapter-trimming of Illumina and SOLiD data, `Cutadapt` is a command-line 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 [50]. `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 [49]. `FastQC` is a Java-based tool for quality control and provides per-base and per-read quality profiling options [48]. 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 previously mentioned tools provide. Additionally, `fastp` outperforms them in terms of speed by being 2 to 5 times faster [51]. 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 polyguanine (polyG) tail trimming are integrated into `fastp` [51]. Its multithreading implementation in C/C++ makes it faster than other 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` [52].

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. Assembly of short reads can be done as *de novo* assembly without a reference, however this approach requires great computational capacities and great sequencing depth to ensure a sufficient overlap of the reads for an accurate assembly [53].

The choice of alignment method depends on the amount, length and origin of read data. In genomic analyses and routine surveillance, this choice can be challenging and typically, fixed references are utilised for mapping the reads against. Often, these are arbitrarily chosen or old sequences. Especially true for RNA virus genome analysis, as with the avian influenza virus, the RNA genome is highly mutable and diverse. This leads to divergence on a regular basis and also leads to antigenic drift. Thus, an intelligent choice of reference sequence for mapping is essential for a meaningful analysis.

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 bp in length [54]. It states to be faster and more accurate than other domain-specific alignment tools [54].

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 [55]. Using a Burrows-Wheeler Transform, both `BWA-MEM` and `Bowtie2` are shown to be faster aligners than others with reads of 100 bp in length [56].

For *de novo* assembly with Illumina, PacBio or IonTorrent data, `SPAdes` can be used. It is based on a De Bruijn graph algorithm building k-mers [57]. 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 [58, 59]. Its underlying algorithm is an 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 [58]. The A-Bruijn graph is a variant of the De Bruijn graph algorithm

that claims to distinguish better between true variants and sequencing errors, and it can more effectively combine repetitive regions into a single assembly sequence [58].

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. In a consensus sequence, the most commonly observed nucleotide at each site across the full-length genome is reflected by inference from aligning the sequencing data against a specific reference genome. 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 intra-host variant detection (`iVar variants`) [60]. On the `Geneious` platform, similar analyses can be executed in order to produce a consensus sequence from raw reads [61]. `bcftools` as a tool suite also offers a package for consensus sequence construction from Variant Call Format (VCF) files [62].
For ONT generated data, the `medaka` tools suite provides a module to generate the consensus sequence from aligned reads.

Phylogenetic Analysis

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`, `RAxML` (Randomized Axelerated Maximum Likelihood) and `IQ-Tree`, all based on inferring relations using the maximum-likelihood criterion [63, 64, 65]. These tools require a multiple sequence alignment as input data, which can be obtained by multiple sequence aligners like `MAFFT` or `ClustalW` [66, 67]. The generated phylogenetic trees can be visualised to infer topologies and study deep 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 [68].
Other approaches for phylogenetic analysis include a BLAST search to identify similar sequences in a database, a widely used method to get an idea of the phylogenetic relationships of the sequence [41].

Classification of Sequences

Many applications of genomic analysis require the placement of the inferred sequence compared to other, similar sequences. Especially in molecular epidemiology, it is useful to query the assembled sequence from the raw reads against a large database of virus genomes. The results give hints for epidemiological linkage or could determine possible regions or countries of origin. This search can be achieved by global alignment or faster techniques based on string comparison. There are many databases available for different categories of sequences, offering database searches to find the closest sequence compared to the query sequence.
BLAST is a popular search program, while there are different heuristics and variations depending on the specific database and search string characterisations [69]. For nucleotide sequences and databases, the National Center for Biotechnology Information (NCBI) provides a web-based `BLAST` search form [41, 70]. The underlying algorithm is based on similarity measure that performs a trade-off between speed and sensitivity [41].
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 tool that infers a scoring based on a De Bruijn graph construction. It emits the closest sequences from a database of reference sequences [71]. `VAPOR` 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 [71]. `VAPOR` provides options to fine-tune the classification run, depending on read length, database size, k-mer size and other measures. It 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 those that cover large proportions of the reference [71]. It has been argued previously that mapping short reads directly to a De Bruijn graph is less biased than that of mapping to *de novo* assembled contigs [72].

2.3 Pipelines for Genomic Analysis with Viral NGS Data

In the following, a selection of existing pipelines is presented that can be used with unspecified or unknown virus data. They cover parts of the later mentioned pipelines in Sections 3.4.1, 3.4.2 and 3.4.3 and focus 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 [73]. 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 [73].

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 [74]. 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 [75] and Integrated Rapid Infectious Disease Analysis (IRIDA) [76]. 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 [77, 78]. 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 [79]. 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 [80]. 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 [79]. 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 [81]. 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, buffaloes, 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 [82]. 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 [80]. The Chordopoxvirus subfamily is characterised by its large, linear double-stranded genome. Size varies between 134 to 365 kilobases [83, 84]. 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 [85]. For animals, there is a smallpox-based vaccine that is used to protect elephants against cowpox [86]. 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* [87]. 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 bp 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 Virus (SPPV) and Goatpox Virus (GTPV), the LSDV genome is highly similar to the other CaPV genomes [88].
LSDV is not known to be transmissible to humans and therefore not a zoonosis. Natural

hosts of LSDV are cattle and Asian water buffaloes. Although CaPV is considered to be host specific, SPPV and GTPV strains can naturally cross-infect in both host species. There have been no cases of natural infection of sheep or goats with LSDV reported [89]. 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 [90]. 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 [91].

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 [92]. In African and Asian countries, veterinarians struggle to fight endemic LSDV outbreaks due to 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 from LSDV, even though it does not provide complete immunity. Nevertheless they are used in regions where all CaPV are prevalent [93].

In 2017, a novel LSDV was discovered in Russia, called the Saratov strain [94]. 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 [95].

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, SPPV and GTPV [88].

An ampliconic assembly-based approach to distinguish Capripoxviruses is described by Mathijs et al. [96]. 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 [25, 60, 97, 98].
A pipeline of Zhao et al. was designed to study the whole genome of mpox samples [99]. 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 [99]. 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 fine-tune the pipeline parameters.
Other methods to detect the species of Capripoxvirus of a given sample is nucleic acid extraction and real-time PCR [100]. 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 [18]. 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 [8].
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 [8]. 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 [8, 101]. To be infectious, a virus particle must contain one of eleven proteins in each of the eight unique segments PB2 (polymerase basic 2), PB1/PB1-F2 (polymerase basic 1), PA/PA-X (polymerase acidic), HA, NP, NA, M1/M2 (matrix) and NS1/NEP (non-structural gene/nuclear export protein). 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 surface proteins HA and NA 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 [102]. 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 [103]. 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 [104]. 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 and 2015. It resulted in more than 25 million bird deaths [105]. 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. [106]. 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 [107, 108].

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 [38]. 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 [109], EpiFLU on Global Initiative on Sharing All Influenza Data (GISAID) [110], Nextflu [111], NCBI Influenza Virus Resource [112], FluNet [113] and OpenFluDB [114]. 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 Subtype (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 [115]. It is an influenza-focused suite to process NGS data to automatically get outputs to answer key questions in influenza genomic surveillance. INSaFLU can be used for seasonal influenza, avian influenza, SARS-CoV-2, mpox virus or Respiratory Syncytial Virus (RSV). 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`. Alignment is performed with mapping against a user-input reference sequence and the `BWA` tool. In the following, (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 and phylogeny with `FastTree` and the tree visualiser `Phylocanvas`, and lastly (3c) coverage analysis with a INSaFLU specific Python script are performed. 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 by the user as input data from the beginning. Currently, INSaFLU is accepting NGS data from influenza, SARS-CoV-2 and mpox samples [115]. 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 [116]. 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 [116]. 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 viral 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 [117].

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 [117]. 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 [118]. 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%) [117].

The causative virus is a small positive-sense RNA virus genome with a size of approximately 8.3 kilobases, belonging to the Aphthovirus genus in the *Picornaviridae* family [82]. It contains a single ORF coding [119]. It is notable that the FMDV, similar to other Picornaviruses, has a polycytidine (polyC) tract in its 5' non-coding region that is highly conserved among the different serotypes and strains [120]. The exact length of the polyC tract in FMDV varies between 50 and 400 nucleotides and is presumably responsible for translation efficiency and pathogenicity [120]. There are seven distinct strains (A, O, C, SAT-1, SAT-2, SAT-3, and Asia-1) 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 Asia-1 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 heterogeneity of the virus even within one serotype.

While there is a huge list of FMD-free countries, 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 with samples from infected stock 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 a major motivation for this work. Ready-to-use pipelines that can be executed without deeper biological or bioinformatic 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 GitHub. 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, public servers, 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 servers for analyses provide 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 front-end, and lets users start 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 has a default of 250 GB disk space allocated on the three main public servers 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 U.S. and WorkflowHub for EU users publish the IWC workflows and guarantee the availability in Docker-based environments and on the workflow collaborative WorkflowHub [127, 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 three 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 three public servers are synchronised in a subset 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

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 generation of 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 output files are used for variant annotation, phylogenetic assignment of the outbreak lineages and clade assignment. The workflow skeleton is depicted in Figure 2.

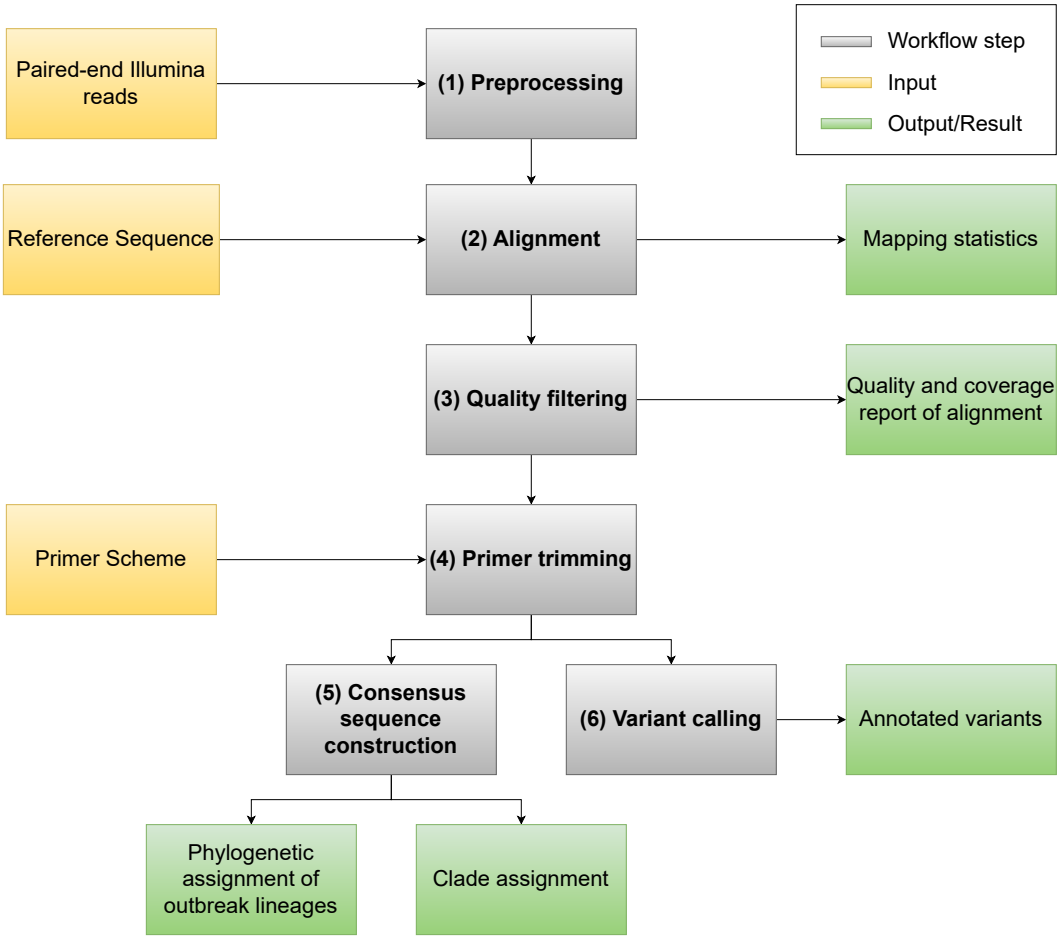


Figure 2: Simplified SARS-CoV-2 analysis workflow for ampliconic Illumina-sequenced data.

This workflow is designed for the specific attributes of the SARS-CoV-2 genome, however most viral genomes can be analysed on genomic level 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 in this work, reusing components of the described SARS-CoV-2 workflow. The requirements for the viruses and the workflows are described below, before the developed workflows are examined.

3.3 Workflow Requirements

To account for the distinct attributes of the examined viruses, automated pipelines used to achieve whole-genome insights from raw reads must be tailored to the specific attributes of the virus being examined. The requirements for each workflow are explained below.

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 like rabbitpox and vaccinia virus, while mpox, 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 to find the most agreed position for a complete alignment. Since there is no unambiguous position for repeated identical sequences neither in reference-based mapping approaches nor *de novo* assembly, a new approach has to be used so that the two ITRs are not aligned in the same run, but separately to constitute disambiguity. Therefore, we use a method 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" back together. To ensure the reads to be mapped in the appropriate and not in the wrong ITR, the reads need to be sequenced in two pools with two sequencing libraries. A similar tiling amplicon protocol has been described by Mathijs et al. and the ARTIC network for SARS-CoV-2 data [26, 96].
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 the split amplicon-based sequencing with an Illumina sequencer. Working with Illumina-generated NGS data, the workflow necessitates a quality control and trimming of the reads to remove sequencing artefacts and adapters. 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 them back together.
Mapping of each genome-half that each contains one ITR requires a reference sequence, which is a compulsory workflow input. Since poxviruses have low mutation rates in general, a fixed reference sequence accounts for a nearly unambiguous mapping and consensus sequence.
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), however not suitable for undetermined poxvirus genus members [131].

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 mutations or new variants that occur in the HA and NA proteins as a consequence of reassortment.
A pipeline for an avian influenza virus sample that builds a consensus sequence requires a reference sequence that it can align the NGS reads to. For an Illumina-based workflow, preprocessing is crucial to ensure reliable results working with the reads. Quality filtering and trimming must be included in the beginning of the workflow. A main caveat of many existing pipelines for AIV genomic analysis is the user's choice of reference sequence, since it is an arbitrary selection and has a direct impact on the alignment. Another, computationally very expensive approach would be assembly which does not require a reference sequence. Since the influenza segments can have very similar regions at the

segment's ends and mapping is computationally faster than assembly, a reference-guided mapping method is favoured for the analysis of AIV samples due to the genome size and high mutation rates of AIV. The goal is to use a reference that is representative of the sample being analysed. In the SARS-CoV-2 pipeline, it is recommended to use a reference genome from a recent SARS-CoV-2 strain. For avian influenza virus, multiple reference sequences exist depending on the strain and subtype, however this information only helps for the reference selection if the strain or subtype of the sequenced sample is known. Additionally, avian influenza viruses tend to reassort during replication and one sample may match with different possible reference sequences for the different segments. Taking a single reference for mapping, a possibly new reassortment event may not be discovered. Hence, a dynamic approach that is sensitive enough for the segmented structure of the AIV genome is needed to pick a representative reference and to relieve the user from taking the complex choice of reference. A search tool like `VAPOR` could help identifying close reference sequences based on the input reads by looking up a large user-defined database of sequences. 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 an approach that takes any (maybe imperfect) reference strain may be effective for the other six segments, the mapping software would frequently be unable to achieve sufficiently plausible matches for sequenced reads of the HA and NA segments to continue with the analysis. By introducing a method that finds the best reference sequence from a database before the actual alignment, the expensive assembly step is avoided, the user is not required to choose an arbitrary reference and mapping to a suitable reference with minimal bias can finally be performed.

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 as they are in the SARS-CoV-2 workflow, but kept for maintaining a reasonable high coverage for the further analysis. Downstream analyses for phylogenetic placing are useful for the HA and NA genes to trace viral origins and consider relations to

similar strains, as well as visual summaries of SNPs for identification of genetic variation in different regions.

Requirements for FMDV Analysis Workflow

Genomic analysis of the viral FMDV RNA genome requires a workflow that accounts for its high mutation rate. Aligning raw Illumina-sequenced reads requires quality control in a preprocessing step to remove sequencing platform specific adapters and dismiss reads with low quality. For alignment of the reads to construct a consensus sequence, mapping to a reference sequence or assembly are considered. Finding a representative reference sequence from a database with many sequences, for FMDV reads this approach would regularly fail due to the very high mutation rate and ensuing large differences between the query reads and the database sequences. Therefore, a different approach to find a suitable reference sequence for mapping is required. Since the FMDV genome is relatively short with approximately 8.3 kilobases, a *de novo* assembly takes only little amount of computational resources for a run. This is due to less contigs to assemble and fewer gaps to fill during assembly, and usually more high coverage regions that facilitate the assembler to find long contigs. The overall complexity of assembly is highly reduced with short genome lengths and therefore increases efficiency of assembly.
A *de novo* assembly of the FMDV reads to avoid an arbitrarily chosen reference sequence with a subsequent BLASTn search in the nucleotides database is one method to find similar sequences that allow for a high-quality mapping and consensus sequence construction.
Additionally, the workflow should include steps for quality control, including the removal of low-quality reads and the identification and removal of potential contaminants or other sources of error. Finally, a workflow for FMDV genomic analysis should accommodate Illumina-sequenced data and be able to scale up for working with multiple samples at a time.

3.4 Workflow Development

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

TODO: check again all wfs and add more step descriptions

3.4.1 Poxvirus Illumina Workflow

The newly designed Galaxy workflow for Illumina-sequenced poxvirus samples with a tiling amplicon approach is available on WorkflowHub, Dockstore and on IWC to use on Galaxy EU. Links can be found in Supplementary Section 1.1.
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 ampliconic data that were sequenced in two pools. It aims at constructing the full genome from ampliconic Illumina-sequenced reads and providing alignment files, sample-specific consensus sequences and intermediate results and reports that give insights into reads, mapping quality and mapping coverage. The pipeline is clearly shown in its structural elements 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 that separates the ITRs to ensure unambiguous mapping of reads. Therefore, the workflow requires the input reads in two sequencing pools that each represent one half of the genome. During the first steps, the reads of the two pools 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* (5th) column; and a reference sequence that is used for mapping which can be retrieved from the NCBI reference sequence database depending on the genus of the sequenced sample.

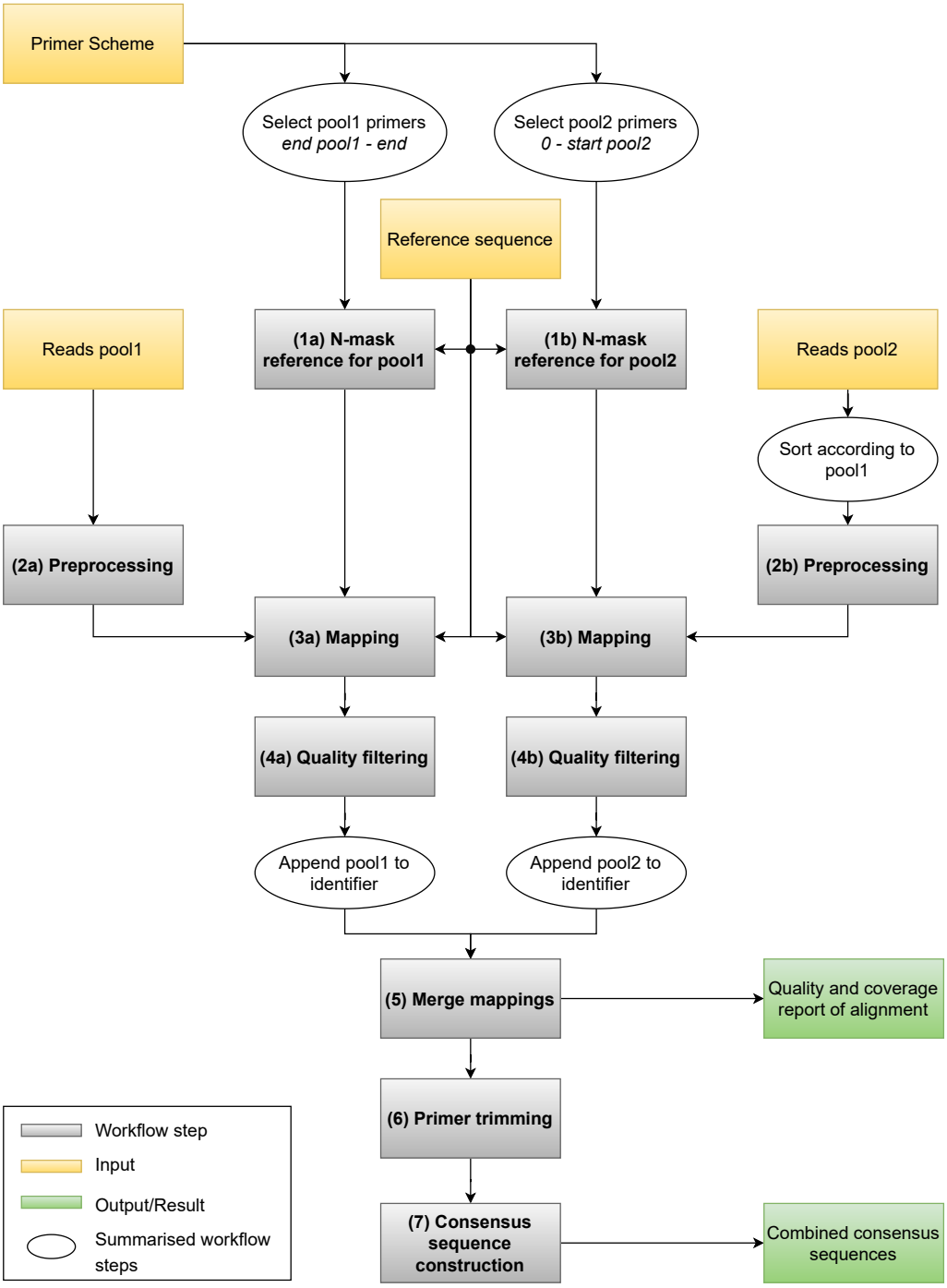


Figure 3: Simplified poxvirus genomic analysis workflow for ampliconic Illumina-sequenced data.

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 start and end position of the two pools so that the remaining bases can be N-masked. For mapping *pool1*, which accounts for the first half of primers 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 for further workflow logic. The N-masking of the reference starts at the minimal start position of the first primer of *pool2*. If the pools and amount of primers are of similar size, this position is in the middle region of the reference sequence. It is important that this position, separating the pools, is in between the two 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 as an end position 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.

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

The poxvirus workflow is designed to process multiple samples in one run. The workflow requires the raw reads to be uploaded in two distinct collections, one for *pool1* and one for *pool2*, each containing the reads for potentially multiple samples. For better comparison during the workflow, the samples in the second reads pool collection 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 and remove sequencing adapters with default settings of `fastp` to ensure further quality filtering. The following (3) mapping step with `BWA-MEM` takes the corresponding masked reference sequence for each genome-half as explained. A statistics report for each alignment is generated using `Samtools`

`stats` and allows the user to inspect the mapping quality and coverage of the alignment. Next, the alignments are (4) filtered for quality using `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 for variants, the pool and sample identification is maintained and unambiguous for the user. In the next step, (5) the two alignments are merged while still retaining the identifiers for each sample and pool. For the full-length mapping, a coverage report is generated with `QualiMap BamQC` which allows the inspection of the ITRs and to examine the part where the mappings are merged together. 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 compared to the reference sequence that was selected for mapping. Low coverage of an alignment indicates incorrect mapping due to genetic differences. Therefore, coverage plots are provided in the workflow for each sample.
To prepare the alignment for consensus sequence construction, (6) primer-trimming with `iVar trim` removes the loose primer ends. The (7) consensus sequence is called with `iVar consensus` and a 50-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 Single Nucleotide Variant (SNV): 0.7, minimum allele frequency to call indel: 0.8) or enter their own values before starting the workflow. These settings yield for the minimum of 50 sequenced reads per base in coverage. With the final combined consensus sequence in FASTA format for each input sample, further downstream analyses can be started after finishing the workflow.

3.4.2 AIV Illumina Workflow

We designed 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 is available with all related material via links provided in Supplementary Section 1.2. 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 various outputs for downstream analyses. It is designed to take one input sample at a time and quality report emitted during the workflow run, the outputs of the analysis steps can be used for any consecutive research based on the user's research objective. The AIV workflow is outlined in Figure 4, where the nine main steps of the workflow are visualised. A link to the full workflow and its supplementary material can be found in Supplementary Section 1.2.

One novelty of our AIV workflow is the consideration of the different segments of the influenza virus genome in the composition of the reference sequence. After uploading paired-end reads and a reference sequence database, the workflow builds a hybrid reference from the database for each of the genome segments. The reference sequence database is described in detail in the subsequent Section 3.4.2. If a user decides to upload their own curated references, it is important to follow the sequence identifier pattern so that the extraction of sequence identifiers in the workflow works as expected: $>segment\_name|influenza\_strain|subtype|accession\_number$. Spaces must be avoided. For instance, one entry's identifier is $>PB1|A/duck/Manitoba/1953|A/H10N7|KF435047.1$ followed by the nucleotide sequence in the next line.

The AIV workflow takes the reference sequences in datasets split by segment and the paired-end Illumina-sequenced reads, and an additional numeric parameter to determine the size of the produced phylogenetic tree. After (1) preprocessing of the reads with `fastp` with default quality trimming options, and additionally to dismiss reads shorter than 30 bp, filter out 5' and 3' ends with a mean quality of below Q30 (a Phred quality score of at least 30 is required) and automatic trimming of 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 to

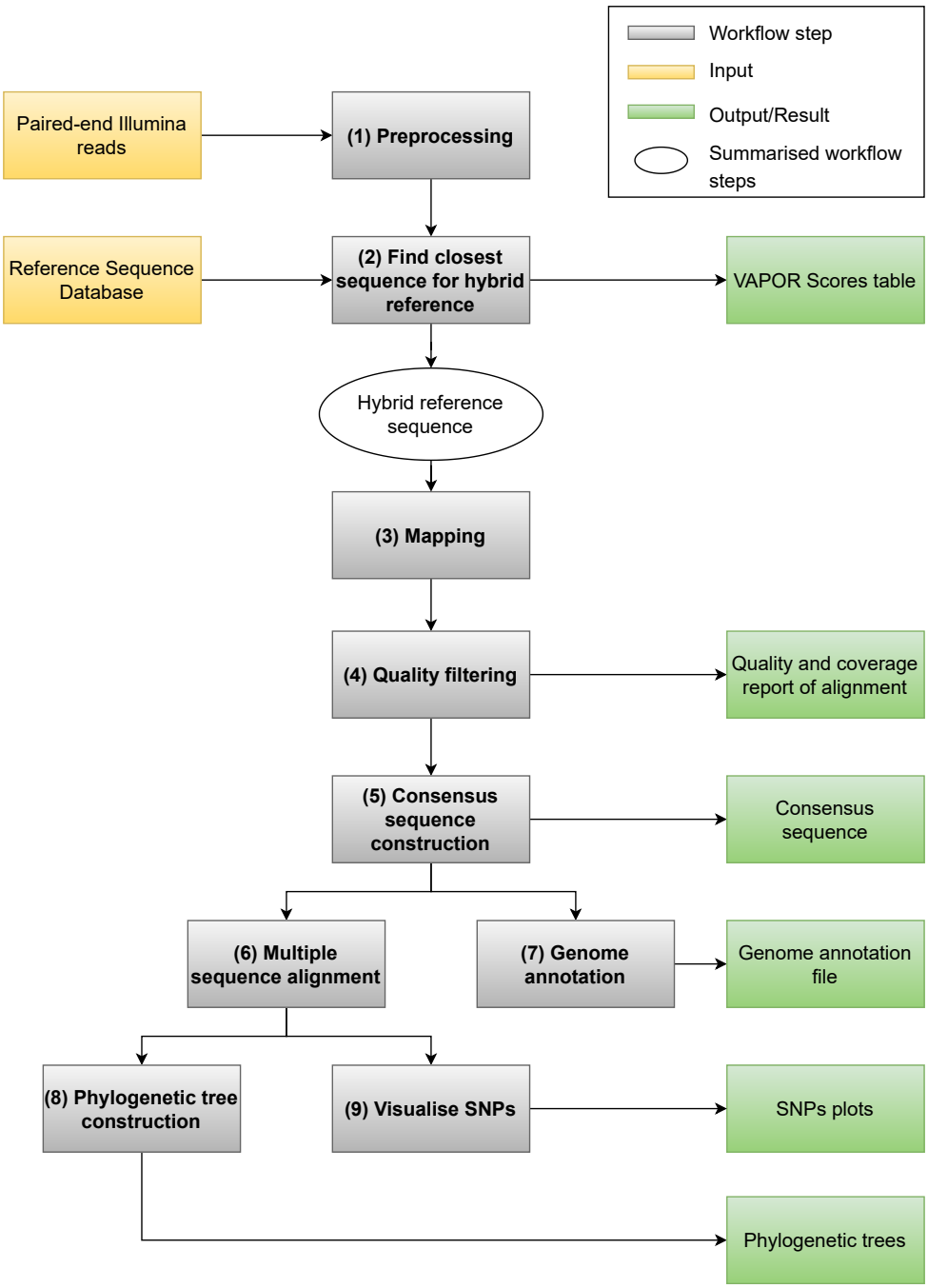


Figure 4: Simplified AIV genomic analysis workflow for Illumina-sequenced data.

the reference. As VAPOR is running once per segment yet with independent inputs, the eight jobs are executed in parallel and do not depend on each other's outputs. VAPOR is a graph-based classifier that maps k-mers to a weighted De Bruijn graph to find sequences from a database that are as similar as possible to the query sequence [71]. 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, an integral part of the workflow is to build a hybrid reference sequence used for mapping. To inspect the statistics of the graph and adapt the configuration, a table with the highest VAPOR scores of each run is generated and visible in the workflow history on Galaxy.

The hybrid reference sequence is composed of the eight segments and is used as the reference sequence in 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 to 1000 bp long reads by seeding alignments with maximal exact matches, and extending the seeds using the affine-gap Smith-Waterman algorithm [55]. 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 generation of the consensus sequence from the reads alignment prepare the BAM file and deconstruct the mapped reads into a collection of eight datasets by relabelling the elements so that (5) iVar consensus can perform consensus sequence construction. 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 criteria, an N is inserted instead.

To place the consensus sequence of the genome segments in a set of samples from the reference sequences to generate phylogenetic tree data, (6) a multiple sequence alignment for a user-specified number of sequences that determines the size of the resulting phylogenetic trees is done with `MAFFT` (Multiple Alignment using Fast Fourier Transform). The consensus sequence is added using `MAFFT add`. The multiple sequence alignment is also used for (9) a visualisation of SNPs, produced with the `snipit` tool. It provides a graphical summary of the mutations on base-resolution by comparing the consensus sequence to other close sequences from the reference database. To make the `snipit` tool available on Galaxy, a tool wrapper in XML format was written as part of this research. The link to the file is provided in Supplementary Section 1.4. These sequences are selected from the top hits that resulted in the `VAPOR` run and therefore are suitable to flag up possible mutations or mis-aligned consensus bases in the consensus sequence of each influenza segment.

The next step in the pipeline using the consensus sequence is the (7) generation of genome annotation files with `Prokka`. Because 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. (8) 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 [65]. Trees can be explored by downloading one of the standard tree formats (*.nhx, .mldist* or *.iqtree*) for further analysis, visualisation or using the Galaxy web-interface with visualisation tools.

The presented AIV workflow avoids the computationally expensive *de novo* assembly, instead uses a mapping approach with a dynamically composed reference genome of

close sequences for each of the eight influenza segments. This accounts for a high quality mapping and is evaluated in Chapter 4.2. To trace the individual steps and look up intermediate outputs, quality reports are emitted during the workflow process and after finishing, which can be downloaded as a PDF for each workflow run. Due to a variety of possible downstream analyses that can be interesting for a user, our workflow provides results of the individual steps which can be used with various other tools.

AIV Reference Database

For reference-based mapping of the AIV reads in the AIV workflow, a reference sequence is required. To choose a representative sequence for each of the eight segments of the influenza genome, a database with sequences is used, split into files by segment. The reference database consists of eight FASTA files, one per segment (PB2, PB1/PB1-F2, PA/PA-X, HA, NP, NA, M1/M2, and NS1/NEP), containing multiple full-length sequences per segment.

The sequences were downloaded from the NCBI Influenza Virus Database in nucleotide FASTA format. In addition, it is ensured that the 56 sequences from INSaFLU which are provided in their reference database are part of the reference collection. Only full-length sequences with complete coding regions that include start and stop codons are used. Search results from the NCBI Influenza Virus Database show that for some few sequences, the segment genome including start and stop codons is encoded, however includes additional sequence artefacts possibly from other segments in the front or tail. Therefore, sequences with a length of more than 105% of the mean segment genome length according to Chauhan et al. (2022) are dismissed. Similarly, short sequences that hold less than 80% of the mean segment nucleotide count are dismissed. This criterion ensures that a segment can be shorter than the mean length due to deletions, yet does not include sequences that are too short to reliably identify and compare with other sequences.

Vaccine strains and mixed subtypes are exluded from the results. Duplicate sequences

are dismissed and the sequences are prepared so that the header is in the required format, does not contain spaces and has the segment name in the sequence header before the first pipe. Additionally, all sequences containing non-ACTG bases were dismissed. The remaining sequences ensure within-subtype variation of influenza strain A sequences by providing multiple sequences of the same subtype for each segment. Gene subtypes that occur only in bats as H17, H18, N10 and N11 do, are dismissed from the reference collection [132]. Due to the strict criteria, there are not always all eight segments present for one genome. The filtering criteria mentioned above are essential for maintaining the quality and reliability of the data for the `VAPOR` tool, for which the reference sequences are used as the collection to query the sample reads on. An overview of the resulting reference database and the filter criteria is provided in Supplementary Table 9, as well as an overview in Supplementary Table 10 of the occurrences of each subtype in the HA and NA genes. Since the AIV is split into its gene segments in each dataset, the reference collection that composes subtypes from the different HA subfamilies (1-18) and NA (1-11) subfamilies is not required to contain all possible combinations and therefore not all subtypes are required to be present. In the AIV workflow, the `VAPOR` tool looks for HA sequences that are highly similar to the query reads, and provides the reference for the specific HA subfamily sequence only from the dataset containing HA segments. Similarly, the NA subfamily is determined by querying the NA dataset. This means that in the HA `VAPOR` results and therefore in the reference used for mapping, only the HA subfamily part (e.g. H5 from H5N10) determines the HA subfamily. Accordingly, the NA subfamily is derived by the most similar sequence (e.g. N8 from H3N8) within the NA dataset. In combination, the subtype of the sample is derived (e.g. H5N8). The reference collection with a total of 137 507 unique sequences is ready to import into a new history and publicly available on Galaxy EU. The link is provided in Supplementary Section 1.2.

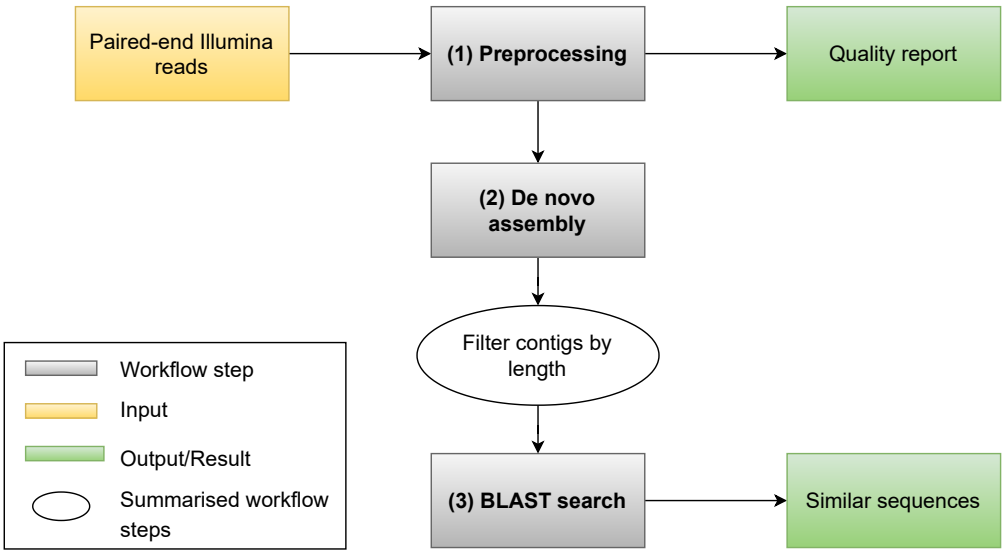


Figure 5: Simplified FMDV workflow (1/2) with *de novo* assembly and BLASTn search.

3.4.3 FMDV Illumina Workflow

We developed a workflow for the genomic analysis of reads from Foot-and-mouth disease virus samples using Illumina sequencing technology. The workflow is split into two single workflow parts and requires the user to take action in the process of reference sequence selection. The FMDV workflow takes multiple samples in a collection and is integrated into the Galaxy platform. Links to the two ready-to-use workflows in *.ga* and *.cwl* format can be found in Supplementary Section 1.3.

The first part of the workflow accounts for the choice of reference sequence and 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. Default settings of `fastp` account for automatic quality filtering to remove sequencing adapters and unqualified reads by additional quality filters. Before assembly, a step to create a datastructure that allows multisample processing by the assembler is inserted to the workflow. Using the `Apply Rules` tool, a nested collection from the reads is generated,

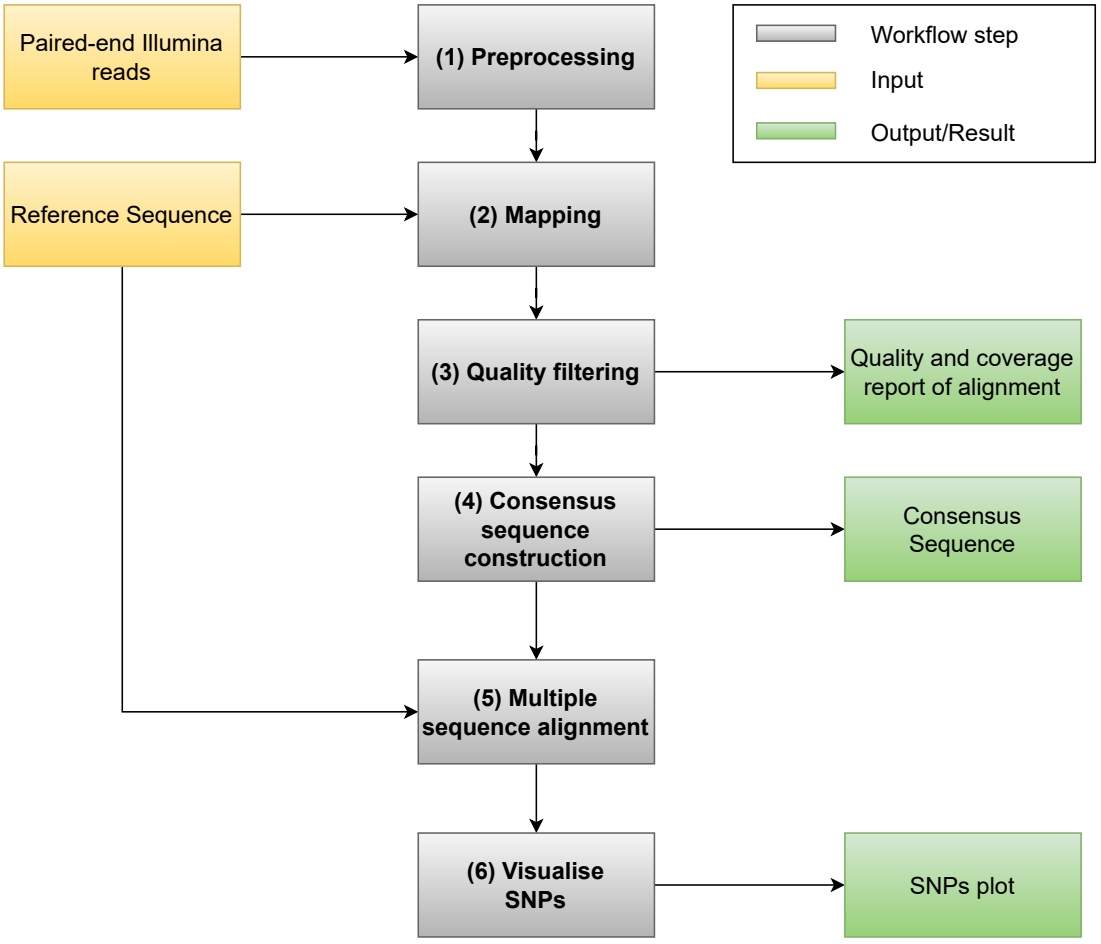


Figure 6: Simplified FMDV workflow (2/2) with reference-based mapping and consensus sequence construction.

while each sample is a list containing the forward and reverse reads. This step ensures that the samples are treated separately during assembly and invoke one assembly run per sample.

In order to find the most similar existing sequence that can be used as a reference for mapping, this workflow first includes a *de novo* assembly with `rnaviralSPAdes`. It is an assembler tailored for short RNA viral data and was initially developed in a specific version for the assembly of SARS-CoV-2 samples. It improves the existing `SPAdes`

assembler by making use of knowledge about the structure of viral RNA genomes [133]. The FMDV contigs, which may be varying in terms of amount of sequences and length of contig, are then (2) filtered by length to dismiss short contigs that represent only fragments of the viral genome. The cut-off is set to almost half the size of the FMDV (minimal length 4.0 kilobases). This allows the subsequent (3) megablast search on the NCBI NT database to find similar genomes even in the case of co-infection of the sample or recombination within the viral genome. After the BLASTn search, the user is required to inspect the hits of the megablast search, and to identify the best matching reference genome based on their own criteria, such as coverage and identity. The resulting sequence that is chosen as a reference may prove plausibility of the sample or reveal the presence of other nucleic material if one of the top BLASTn results matches an unexpected virus. To avoid the automatic choice of reference selection by taking the top BLASTn match as reference for the following mapping, the workflow is split up. This implies that if a user has a different reference sequence independent of the first FMDV workflow, the sequence analysis can start directly with the second FMDV workflow and by uploading the reference sequence to a new Galaxy history.

The second workflow for FMDV genomic analysis is designed for multiple raw reads in a collection that are mapped to the same reference sequence. This is useful in an outbreak scenario where a workflow user has multiple sequenced samples from the same outbreak and seeks to compare these samples with each other, identify similarities, relations and origin of the virus. However, this workflow can be used as a stand-alone pipeline without the first workflow in case the user aims to map the raw reads to a specific, arbitrarily chosen reference sequence. After uploading the reference sequence in FASTA format, identified from the megablast search and retrieved via the NCBI upload tool (`NCBI Accession Download`) or from other sources, the workflow runs a (1) preprocessing with `fastp` to dismiss reads shorter than 30 bp and to trim polyG tails of the reads. The second step involves (2) mapping of the preprocessed reads to the reference genome using `BWA-MEM` with default configurations. This step generates a

Sequence Alignment Map (SAM) file, which is then (3) filtered for quality using `samtools`. Paired and mapped reads are kept that have a minimum quality of 20. Alignment and quality reports including coverage statistics are generated per sample using `QualiMap BamQC`. The filtered SAM file is then used to generate a consensus sequence for each sample using the (4) `iVar consensus` tool with a minimum quality score threshold of 20, minimum frequency threshold of 0.7, minimum indel frequency threshold of 0.7 and a minimum depth of 10 to call consensus. This step allows the workflow to generate a high-quality consensus sequence for each sample, which can be used for downstream analyses, such as multiple sequence alignment and phylogenetic analysis. The resulting consensus sequences from all samples are aligned to the reference genome using (5) `MAFFT`, and the resulting alignment is used to (6) identify and visualise SNPs with `snipit`. The FMDV workflow produces a summary report of the results of each step and allows the investigation in additional research with the output of each step from within the Galaxy history.

4 Results of Workflow Validation

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 Poxvirus Workflow with Lumpy Skin Disease Virus Datasets

We employed our pipeline using a tiling amplicon approach with masked reference sequences for each half genome to ensure an unambiguous mapping to the identical two ITR regions of the poxvirus genome. Two public LSDV samples, 20L70 (MZ577075.1) and 20L81 (MZ577076.1), that were sequenced using a primer scheme with two primer pools are used and retrieved from GenBank on 10^{th} April, 2023. Collected from cattle in 2020 during a lumpy skin disease outbreak in Northern Vietnam (20L70_Dinh-To/VNM/20 and 20L81_Bang-Thanh/VNM/20), both samples were sequenced on a MiSeq System using a Nextera XT library preparation kit.
The used CaPV primer scheme in BED format contains information about the primer pairs used for the amplicons. Each primer pair has one positive and one negative strand primer, indicated by the strand in the sixth column and by the *LEFT* and *RIGHT* label in the name. Primers are labeled in an alternating way: *pool1* primer pairs are denoted as *pool1a* and *pool1b*. We use the same method for *pool2* with *pool2a* and *pool2b*. We reuse the *SCORE* column from the BED file to unambiguously identify primer and strand for each amplicon. The annotated primer scheme for Capripoxvirus (CaPV) is part of

the workflow run to which links are provided in Supplementary Section 1.2. The LSDV "Neethling" strain was used as reference genome (NC_003027.1, retrieved on GenBank 10th April, 2023) for mapping. The raw FASTQ files for each sample were quality trimmed with `fastp` and mapped to each half-masked reference, which is explained in detail below. Preprocessing with default `fastp` settings includes automatically detected adapter trimming, a quality filter to discard reads below an average quality of Q15, with more than 5 uncalled (N) bases and 40% unqualified bases, a length filter to discard reads below a threshold of 30, automatic trimming of polyG tails for Illumina NextSeq/NovaSeq data and a minimum length of 10 to detect polyG tails.

Output Metric	**20L70**	**20L81**
Paired-end raw reads	863 820	1 016 168
Paired-end reads after quality trimming	856 138	947 064
Proportion of reads mapping to reference	99.6%	77.3%
Proportion of reference covered	99.68%	99.68%
Mean coverage	2 705.2×	2 411.4×
Alignment error rate	1.25%	1.30%

Table 2: Metrics after preprocessing and mapping for datasets 20L70 and 20L81.

The used primer scheme contains a total of 23 primers, while the first set of 12 primers covers the 5' genome end labeled with *pool1*, and the remaining 11 primers cover the 3' genome end, labeled with *pool2* as depicted at the top in blue (*pool1*) and red (*pool2*) in Figure 7. This scheme was designed for the tiling amplicon approach all three members of Capripoxviruses, which includes the LSDV sample. To use the primer scheme with SPPV or GTPV samples, the primer positions need to be shifted to the correct coordinates.

The used primer scheme is provided in the Galaxy history of the test runs and is available via links in Supplementary Section 1.1. Inspection of the masking intervals for N-masking the reference confirms that the right-most position of Ns of masking the 5' half (i.e.

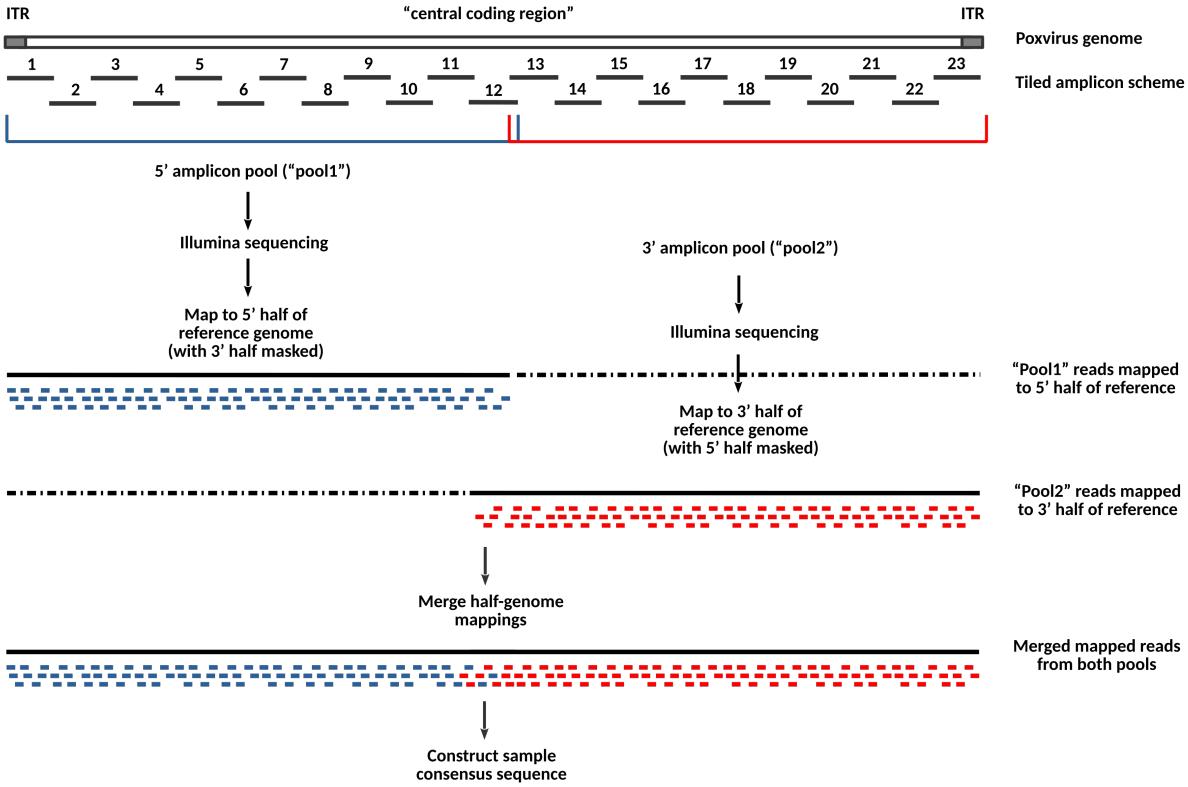


Figure 7: Tiling amplicon scheme emphasising masking of the reference, mapping in two pools and merging of mappings with almost no overlap. Source: adapted from **TODO: Poxvirus tutorial**

preparing the reference for mapping *pool2* reads) is the minimal start position of *pool2* primers ("1 – 79081", and 79081 being the start position of primer 13). Accordingly for N-masking the reference for mapping of *pool1* reads, the position of the right-most primer end of *pool1* primers (primer 12) is 80202, resulting in the interval "80202 – 150773". This is clearly visible in Figure 8, where reads from *pool1* are labeled in red and *pool2* in blue, and mapped to the same reference sequence with different masked parts. As expected, the coverage in the region where reads from both pools mapped to is higher. The final position is the maximal end position and the total length of the reference sequence. Since the reference genome and primer scheme are the same for both datasets 20L70 and 20L81, the N-masked references are used for both mappings. Mapping of each pool is done with `BWA-MEM` and default settings for Illumina-sequenced reads, using the

N-masked reference for *pool1* and *pool2* respectively. This results in a mapping with a small overlap in the central part of the genome, where primer 12 ends and primer 13 starts as indicated in the top of Figure 7. After merging the mappings with `Samtools merge`, statistics for preprocessing and mapping are reported and summarised in Table 2. Figure 8 shows a screenshot from Integrative Genomics Viewer (IGV), where the mapping of reads from *pool1* (red) and *pool2* (blue) are merged and demonstrate a higher coverage in the overlapping part of the reference sequence which was mapped to in both pools. For both samples 20L70 and 20L81, almost the complete reference genome was covered during mapping (99.68%) with a mean coverage of 2705.2× and 2411.4× respectively. Primer end positions of *pool2* primers are highlighted with blue circles and mark the start position of the overlapping region and the end of the masking of the reference sequence for *pool2*.

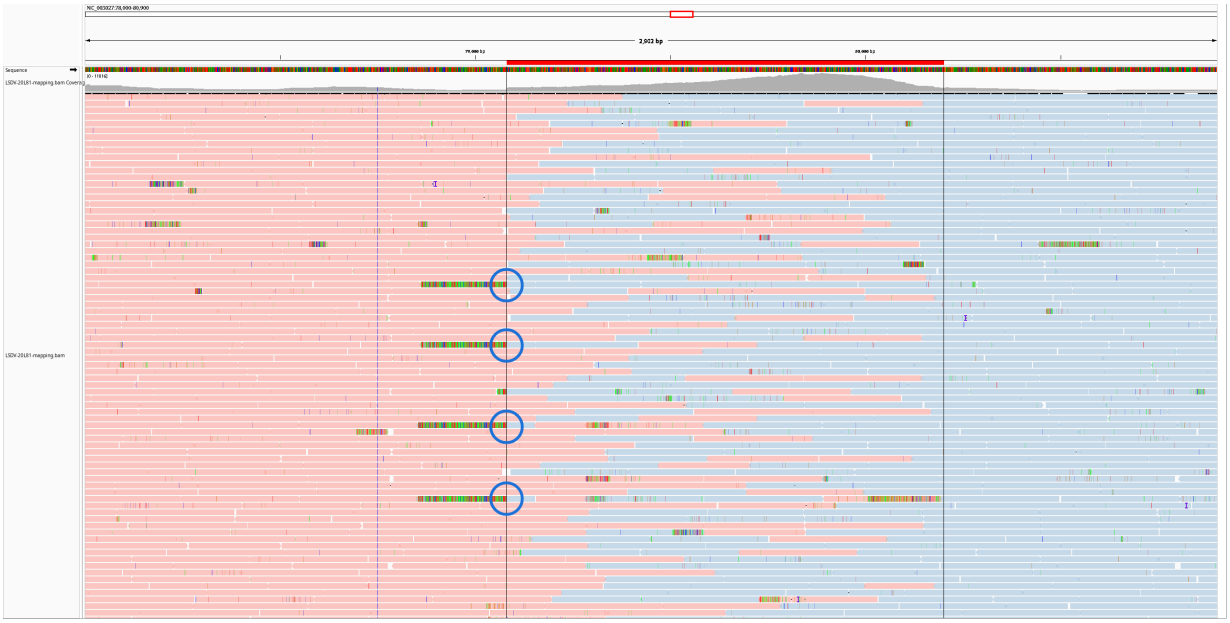


Figure 8: Overlapping, seamlessly merged mapping region of the two amplicon pools of the 20L81 sample. Coloured reads indicate read groups from *pool1* (red) and *pool2* (blue), primers are soft-clipped and end where the mapping of *pool2* reads starts, as marked with blue circles for primer 13 to which four reads from *pool2* bind. More primers may bind outside of the cropped snapshot.

The merged mapping of both read pools was quality trimmed with `iVar trim` to remove primers and mapped reads with a length of less than 30 bp. The remaining reads were used for full-length consensus sequence construction with *iVar consensus*, developed for amplicon-based sequencing data. Inspection of the consensus sequences for both samples shows that apart from low coverage regions in the 5' front and 3' tail due to ampliconic sequencing, a consensus sequence was produced and a base at each position could be found. A base could not be called in the consensus at the first 264 positions in both the 20L70 and 20L81 sample, which corresponds to the primer binding site on the forward strand, and no bases at the last 268 positions indicate the remaining part where the last primer on the reverse strand starts. This was expected due to ampliconic sequencing where the first primer ends at genomic position 264, and where the target sequence starts. The workflow ends by producing a collection of the per-sample consensus sequences.

Poxvirus Workflow Validation using other CaPV Reference Sequences

Until 2017, genomic analysis of different LSDV samples from all over the world suggested very limited genomic variation. The two phylogenetic clusters that were thought to be stable: the Neethling-based strain (used for vaccine development) and a wild-type strain [134]. In 2018, a report has been filed from a vaccine-like field isolate that revealed the presence of multiple recombination sites within the sample while one of the found parental strains was the Neethling vaccine strain (referred to as Saratov strain as in Section 2.4) [94]. Following current sequencing efforts, the existence of more recombinant clusters has been demonstrated from field isolates that took the Neethling-based vaccine. Analysis of this vaccine revealed the presence of multiple Capripoxviruses within the sample: the Neethling-vaccine strain, a KSGP-based vaccine strain, a GTPV strain from Sudan and most importantly other different recombinants between KSGP and Neethling strains. These studies suggest that this recombinant LSDV strain that currently spreads in Asia is likely caused by a vaccine spillover [135]. The Kenyan Sheep

and Goat Pox (KSGP) strain was first detected in LSDV-infected sheep and goats from Kenya, although it has been used for sheeppox and goatpox vaccines [136].
The test samples 20L70 and 20L81 are shown to be from this recombinant vaccine strain, and in order to validate the presence of recombinant sites within these samples, we compare the recombinant sites that differentiate the strains from each other. For this purpose, we run the poxvirus workflow with the two LSDV samples, but use reference sequences from other clades of the Capripoxviruses and thereby highlight the recombinant sites relative to the recombinant Neethling-vaccine strain.

	LSDV-Neethling	**LSDV-KSGP**	**LSDV-Herbivac**	**SPPV**	**GTPV**
LSDV-Neethling (NC_003027.1)	0	0.0006	1.1	2.1	1.8
LSDV-KSGP (KX683219.1)		0	1.1	2.1	1.8
LSDV-Herbivac (KX764644.1)			0	2.1	1.9
SPPV (NC_004002.1)				0	2.4
GTPV (NC_004003.1)					0

Table 3: Distance Matrix based on a maximum likelihood criterion of all clades within the CaPV genus. The LSDV clades are more closely related with each other than SPPV and GTPV, while the LSDV Neethling and KSGP strains are almost identical, and the recombinant LSDV-Herbivac vaccine strain is distinct.

To determine the relationship among the used reference sequences, a distance matrix as shown in Table 3 shows the four different known clades within the Capripoxvirus genus. While the LSDV-Neethling strain and KSGP strain are very closely related, the recombinant LSDV-vaccine strain differs from this clade. SPPV and GTPV clades are

shown to be more different to the LSDV clades. By converting a multiple sequence alignment of the reference used for mapping, the constructed consensus sequences and the LSDV-vaccine reference (Herbivac, GenBank: KX764644.1) to a VCF file and detect variants relative to the vaccine strain. The tool `faToVcf` was wrapped as part of this work to make it accessible from within Galaxy. The tool was initially published by the University of California, Santa Cruz (UCSC) and generates a VCF file.
Using this approach to compare the consensus sequences, constructed by mapping the 20L70 and 20L81 sample reads to other CaPV reference sequences, and comparing variation sites relative to the LSDV-vaccine strain, the found sites show different locations that differentiate the clade from the recombinant strain. The used primer scheme was adapted for mapping to SPPV and GTPV due to shorter genome lengths and different primer locations.Although the `faToVcf` tool is not capable of displaying indels, it clearly shows unmapped regions and the SNPs relative to the vaccine strain.

SPPV compared to recomb. LSDV vaccine strain (Ns): (20L81) - 18,479-18,498 - 110,083-110,106

SPPV compared to recomb. LSDV vaccine strain (Ns): (20L70) - 18,477-18,505 - 110,090-110,113 - 129,582-129,583 (low cov.?)

The impact of the found SNPs was not examined because a gene annotation file for poxviruses with genomic features on th amino acid layer was not available.

4.2 AIV Workflow with H4N6 and H5N8 Samples

The AIV Illumina workflow on the Galaxy platform was evaluated using two field isolates provided by Sciensano, the Belgian national health institute. The isolates were extracted in Belgium in 2020 from an H4N6 infected magpie (EPI_ISL_7593059) and an H5N8 infected duck (EPI_ISL_7596571). For reasons of readability, we refer to the samples as H4N6 and H5N8 samples. The two samples were sequenced on an Illumina

Output Metric	H4N6	H5N8
Paired-end raw reads	1 537 722	858 610
Paired-end reads after quality trimming	1 507 396	830 176

Table 4: Metrics about Illumina reads before and after preprocessing of H4N6 and H5N8 samples.

platform in paired-end mode and are utilised one sample per workflow run on Galaxy. For the AIV Illumina workflow, a reference database in FASTA format is required as a collection, i.e. a list of one dataset per AIV segment, which is uploaded in Galaxy. The used database contains multiple sequences per segment as described in Section 3.4.2. The amount of sequences and distribution of subtypes within each segments suggests that variation within each subtype is generally captured well with the given database. Due to the filtering criteria, not all eight segments of one sample were found suitable for the database.

After starting the workflow, the paired-end reads were preprocessed and serve as query reads for `VAPOR`. Metrics of before and after preprocessing are shown in Table 4 and count more than 1.50 million reads after preprocessing for the H4N6 sample and 0.83 million reads for the H5N8 dataset. Since the reference database contains eight FASTA files in a collection, `VAPOR` runs once per dataset and outputs the highest scoring sequences per segment, which represents the most similar sequences from the database compared to the query sequences.

The `VAPOR` search was able to successfully identify the avian influenza virus subtypes present in each sample: for the H5N8 sample, the most similar sequence of HA segment origins from a sample with the H5 subtype, while the most similar sequence of the NA segment origins from a sample with the N8 subtype. Both found gene sequences contain 100% of the input reads. Similarly, the H4N6 sample was correctly identified with concordance of query bases of 98.7% each. The results of the `VAPOR` run for the HA

Segment	Proportion of query bases in reads		AIV subtype of hit	
	H4N6 sample	**H5N8 sample**	**H4N6 sample**	**H5N8 sample**
HA	98.7%	100.%0	H4	H5
NA	98.7%	100.0%	N6	N8

Table 5: The best scoring sequence of the VAPOR run for each AIV test samples, indicating a perfect match (100% of the query bases are in the reads) of the HA and NA segments of the H5N8 sample sequence, and almost all query bases of the H4N6 sample in the found sample.

and NA genes are summarised in Table 5.

Consensus sequences for each genome segment were constructed with `iVar consensus` and while a consensus could be found at each position, the produced consensus sequences are 100% identical to the originally assembled reads that were uploaded to GISAID by the Sciensano laboratory from the same sample (EPI_ISL_7593059 and EPI_ISL_7596571). Due to slightly different genome sizes of the reference sequence used for mapping compared to the assembly, the resulting consensus sequences in the AIV workflow are shorter than the assembled sequences on GISAID. The consensus sequence for the HA segment of the H4N6 sample is differing in 1 bp in the 5' end, and 4 bp in the 3' end. The consensus sequence constructed for the NA gene misses 18 bp compared to the assembled sequence on GISAID in the 5' end and 33 bp in the 3' end. Similarily, the HA consensus sequence of sample H5N8 is 28 bp shorter in the front and 44 bp shorter in the end. The NA consensus sequence differs in 20 and 28 bp in the front and tail. These differences indicate the missing Untranslated Regions (UTRs) on the whole-length genome. While the hybrid reference that was compiled from the eight influenza segments contains complete segments including start and stop codons, which is a criterion for each sequence to be part of the reference database, the aligned consensus sequence does not contain the UTRs at the 5' and 3' ends as opposed to the *de novo* assembled sequences uploaded to GISAID. This was validated by running the `getorf` tool in Galaxy and reporting the ORFs as nucleotide sequences between start and stop codons. This analysis showed the presence

of full ORFs in each segment in both test samples.

Other workflow outputs for the AIV samples include plots that visually emphasise SNPs relative to the top hits of the `VAPOR` run, indicating the most similar sequences from the reference collection. The plots for the HA and NA genes for the H4N6 sample (Figure 11) show 30 SNPs compared to the first sequence which was also used as reference for mapping (LC121412.1) and 31 SNPs compared to the second best result (MK192399.1). Similarly, the NA gene consensus sequence has 29 SNPs compared to the reference sequence (MW19994.1). For the H5N8 sample, the `VAPOR` run found a sequence with 100% of the query bases in the reads, and therefore the number of SNPs is expected to be low. Supplementary Figure 10 shows there is one SNP in the HA gene compared to the reference (MZ166252.1) at position 1002 and one SNP at position 497 compared to the NA reference sequence (MZ166270.1). The SNPs indicate point mutations or mapping errors, low coverage or close calls during consensus sequence construction.

A protein FASTA file containing gene annotations is generated based on the consensus sequence with `Prokka` and can be used for detailed gene expression analysis. Changes on the amino acid level give valuable insights into the adaptation of the virus and to compare differences among strains in more detail. In virology, working on amino acid level is more common than with nucleotides alone. The protein FASTA file can be downloaded or used within Galaxy.

Phylogenetic classification relative to the 30 most similar sequences in the AIV reference database, queried by `VAPOR`, is done for each segment to reveal temporal and geographical relations of the isolate. Generated phylogenetic trees with `IQ-Tree` are depicted in Supplementary Figures 12 and 13. The trees are unrooted and for the HA gene of the H4N6 sample, shows the nearest clade being from an isolate which is a H4N6 infected mallard from the Netherlands, taken in 2017, and for the NA a single and very early split is made where the gene is clustered to a H6N6 infected duck from Hunan, China. The obtained results for the H5N8 sample show clustering of both the HA and NA genes with sequences from a H5N8 infected mule duck from France, taken in 2020 and 2021.

Users of the workflow with real-world samples could investigate in more detail by uploading

their own reference collection and hereby finding links to other sequenced samples from previous outbreaks in their region. The depth of the split in a phylogenetic tree can indicate the degree of relatedness between the sample and the other samples within that split. A deeper split suggests a closer relationship between the sample and the other samples in that particular branch.

4.3 FMDV Workflows with Asia-1, A, SAT-1 and SAT-2 Samples

The samples used for workflow validation are downloaded from the NCBI and were chosen exemplatory for four of the seven different FMDV serotypes. All four samples were sequenced on an Illumina NextSeq 550 platform. Two samples (Asia-1 serotype, SRR17960053 and A serotype, SRR18751245) were taken from infected cattle and buffalo during an outbreak in Pakistan from 2008 to 2012. One sample (SAT-1 serotype, SRR18685689) was isolated from buffaloes in Kenya in 2016 and plaque purified before sequencing, and the fourth sample (SAT-2 serotype, SRR9328470) was taken from an FMD outbreak in Nigeria in 2014.
The results of before and after preprocessing of the raw reads are described in Table 6. The SAT-2 sample contains a very low number of reads with only 11 576 reads after preprocessing, however to show the ability of the developed workflows, it was kept in the test sample collection.

After *de novo* assembly with `rnaviralSPAdes`, contigs less than half the length of the FMDV genome size were discarded. This resulted in one contig per sample for the BLASTn search, except for the A serotype reads, for which two contigs were assembled. As the longer contig with 12 133 bases is far larger than the FMDV genome size, a contamination or co-infection with another virus is indicated. The BLASTn search was

Output Metric	Asia-1	A	SAT-1	SAT-2
Paired-end raw reads	577 360	2 297 706	903 052	11 816
Paired-end reads after quality trimming	561 280	2 112 856	806 712	11 576
Length of assembled contigs	7 760	12 133	7 329	7 696
with > 4000 bp		7 558		

Table 6: Metrics about Illumina reads after preprocessing and *de novo* assembly of Asia-1, A, SAT-1 and SAT-2 serotype reads.

performed against the NCBI nucleotide database to identify the closest viral sequence matches. The results of the BLASTn search showed that all the contigs were closely related to FMDV as listed in Table 7. The highest sequence identity was observed for the Asia-1 serotype sample, with 96.74% identity, followed by the A, SAT-1 and SAT-2 serotypes, with 94.87%, 93.77% and 91.42% identity, respectively. These results were consistent with the clinical samples being positive for FMDV infection and the specific serotype. However, the second long contig of the A sample resulted in a BLASTn hit for pestivirus (formerly known as bovine viral diarrhea virus 1) with 93.162% identity [137]. This suggests the presence of a co-infection with the pestivirus in the given sample. It shows that the presented workflow is capable of assembling and identifying other viruses present. For consensus sequence construction in the second FMDV workflow, the assembled pestivirus contig is ignored and the reference sequence for mapping can be chosen from the FMDV hits for the other contig in the BLASTn search. This sample shows that during the workflow, the user is required to attentively check the results for plausibility and the reference selection process should not be automated without exact validation of the desired virus. Note that BLAST runs on the Galaxy EU servers use a locally installed database to ensure replicability of experiments. Hence results on the web form of the NCBI BLASTn search may result in different hits due to a different and newer database state. In the megablast search made in this workflow, the NCBI NT database from 22nd January 2018 was used. The multisample run with the discussed

samples of this first FMDV workflow is provided in a Galaxy history and is available via link which is provided in Supplementary Section 1.3.

Sample	Alignment length [bases]	Query coverage	Identical matches
A	7602	92.21%	94.87%
Asia-1	7690	99.10%	96.74%
SAT-1	7331	100.0%	93.77%
SAT-2	7669	99.65%	91.42%

Table 7: Results of the BLASTn run with four FMDV samples. Query coverage refers to the percentage of identical matches in the alignment compared to the BLASTn hit, hereby indicating the quality of the alignment. Alignment length describes the alignment compared to the query length, indicating how much of the query sequence is covered with the alignment.

In order to run the second workflow for reference-based mapping and consensus sequence construction for each of the four samples, the top BLASTn hit of each sample is downloaded in FASTA format to be used as reference sequence for the respective sample. Except for the A sample, the top hit is a FMDV genome sequence of the serotype of the query sample. In this case, user control is crucial for the selection of a representative reference sequence of the respective virus and not the contaminating viral sequence.
With the `NCBI Accession Download` tool, the sequence of each sample that has the highest similarity to the assembled contig is added to the Galaxy history and with `Collapse Collection` the FASTA file is extracted from the list to a single file, so that it is in the required format to start the second part of the FMDV workflow.

For each of the testing samples, the accession numbers used as reference sequence for mapping with `BWA-MEM` are listed in Table 8 as well as quality and coverage measures after mapping. As expected due to the low number of reads, the SAT-2 sample had a low mean coverage of 188.0× and a relatively high error rate of 8.18% compared to the other samples. Consensus sequences are accurately obtained for the Asia-1 and SAT-1

Output Metric	Asia-1	A	SAT-1	SAT-2
Accession no. of reference	KM268898.1	JN006722.1	KM268899.1	JX014256.1
Proportion of reads mapping to reference	100%	100%	100%	100%
Proportion of reference covered	99.67%	100%	98.16%	99.60%
Mean coverage	1 525.9×	15 895.5×	9 302.8×	188.0×
Alignment error rate	3.47%	5.08%	5.87%	8.18%

Table 8: Quality and coverage metrics of the alignment in the second FMDV workflow.

sample that each contain low-coverage regions in the 5' and 3' ends, so the consensus was lost with decreasing coverage in both genomic ends. Sample A has its only low coverage region immediately adjacent to the polyA tail at positions 7626–7634. The Asia-1 and SAT-1 samples show additional low coverage regions adjacent to the polyC tract (Asia-1 sample: genomic positions 372–383; SAT-1: 321–335). Using the same coverage threshold for consensus calling for a sample with a low amount of reads like SAT-2, the coverage criteria were not met in several regions (genomic positions 1–37, 293, 346–528, 3605–3639 and 8039–9131). The consensus sequences can be found in the Galaxy histories of the respective workflow runs for each sample which are linked in Supplementary Section 1.3. Part of the reported results after consensus construction is a summarising SNPs plot, produced by the `snipit` tool. Even though the emitted PNG is too large to display at once, it provides an overview of the distribution of SNP relative to the reference sequence used for mapping. In case of a recombination event, where genetic material from two or more virus strains that are infecting the same cell combines to form a new strain, the distribution of SNPs in the plot would deviate from statistical expectations, in such that there occur regions of high genetic diversity or a sudden shift in allele frquency. Using the visual approach without further tools for identification or confirmation of recombination events is a fast method to detect recombination in the sample, however requires expertise

in the expected pattern of variation based on the reference genome. Besides recombination events, the distribution of variation on the reference genome indicates how representative the reference sequence, selected based on the BLASTn search, captures the aligned reads. If the reference sequence only covers a large part of the genome and not the entire length, it would be reflected in the plot by indicating large numbers of SNPs. In each `snipit` plot produced for the four test samples, a high number of SNPs are displayed, relating to the high mutation rate of FMDV, yet not showing any irregularities that could indicate recombination events. The plots can be retrieved as part of the Galaxy histories of the test sample runs, which are linked in Supplementary Section 1.3.

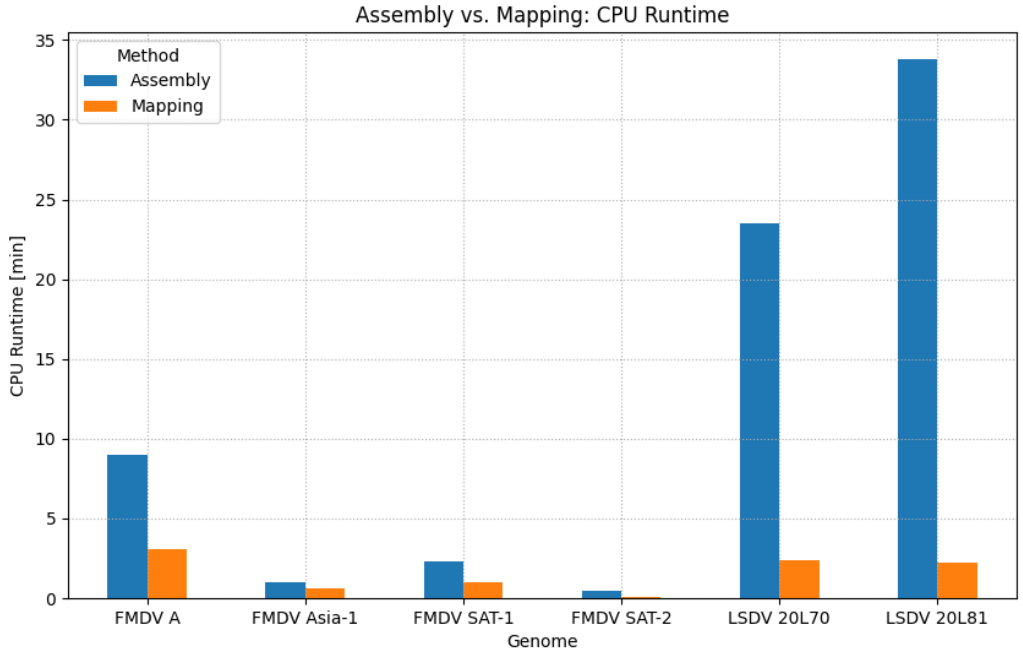
4.4 Workflow Profiling

TODO

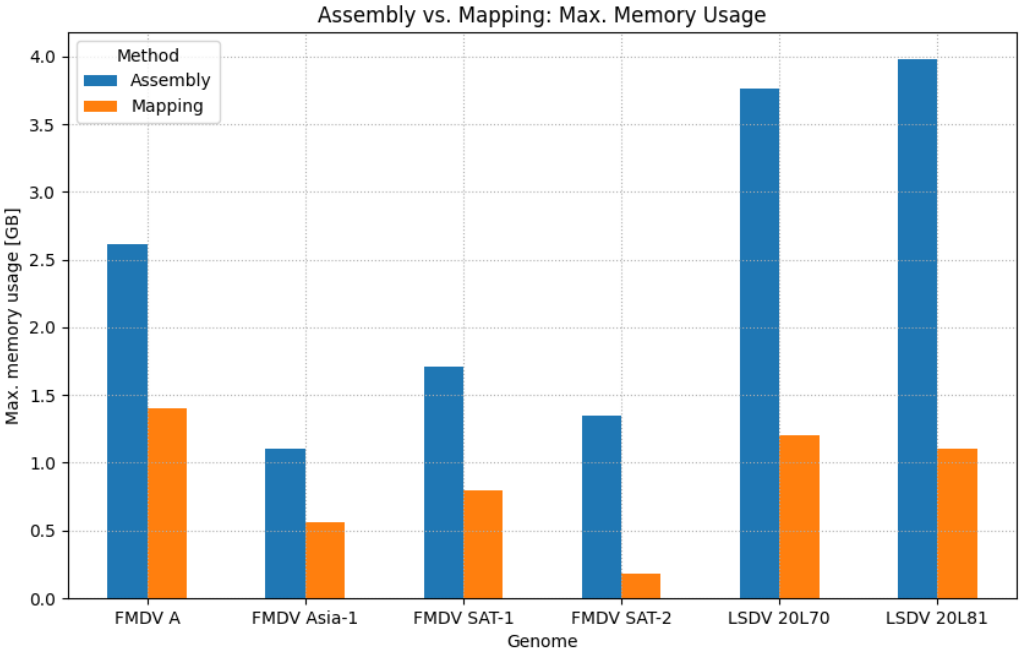
Assembly vs. mapping

for long/short genomes, for rnaviralSPAdes and BWA-MEM, compare CPU runtime and wall clock time, max.mem usage determines the scheduling on a cluster on galaxy (more requested memory means longer queuing times due to fewer available machines)

on a locally installed galaxy server: depends on the cpu alone

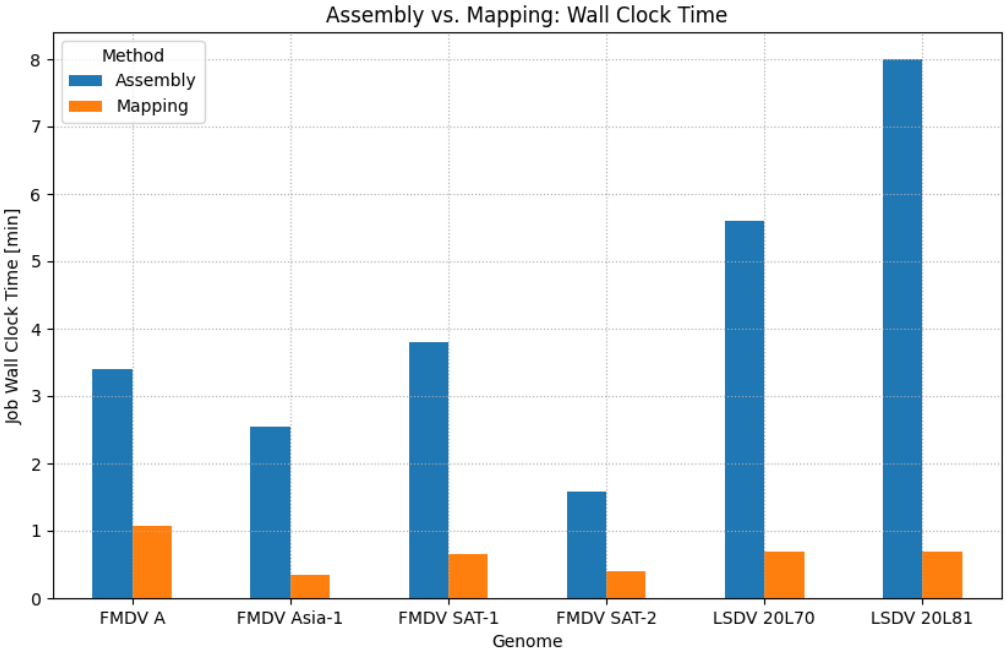


(a) Profiling of assembly versus mapping



(b) Profiling of assembly versus mapping

Figure 9: Profiling



(c) Profiling of assembly versus mapping

Figure 9: Profiling (cont.)

5 Discussion and Outlook

Despite enormous efforts in the prevention and containment of viral livestock diseases, viruses are a constant threat to the economy and the livestock industry. Extensive information on viruses is being collected and shared at the international level, and research advances in diagnostics, epidemiology and virology are increasingly providing valuable insights into the genomic characteristics of specific viruses. Sequence analysis plays an important role in identifying outbreak sources and monitoring antigenic changes within known virus strains. Continuous surveillance, vaccine development and quality control are important aspects of global disease control programs to protect animals, livestock industry and public health. Therefore, a public infrastructure for modern surveillance methods for the use of NGS is of global interest. Reconstruction of the full-length viral genome at the base level from raw sequencing reads requires practical knowledge in tool selection and parameter settings. However, WGS is valuable for virus surveillance as it provides comprehensive genetic information about the virus sample. This information is used to identify virus strains and disease-causing mutations, detect within-host variation and determine relationships between populations and individual samples. Moreover, research on SARS-CoV-2 which was used to combat the COVID-19 pandemic, has made such great strides in understanding viral structures that the introduction of a public infrastructure for experts and laypeople suggests that this newfound knowledge can be applied to other emerging viruses. Existing pipelines for full genomic sequence analysis from raw read data are specific to a small subset of viruses, require domain expertise and a server infrastructure to run computations on local computers. Different sequencing

platforms produce different raw read data types with specific needs. As part of the ZODIAC project, we have developed three ready-to-use, publicly available workflows on the Galaxy platform for sequence analysis of poxviruses, AIV and FMDV from Illumina-sequenced reads to analyse these viral diseases in livestock by generating the full-length consensus sequence. With accompanying training resources and user-friendly workflows on an open-source platform, we ultimately contribute to research-based surveillance methods and enable health professionals without in-house resources to analyse viral outbreaks in livestock.
In this chapter, the contributions to the field, limitations of the work and an outlook for future directions with the workflows are discussed.

Aiming at the challenging whole-genome construction from raw sequencing data, we developed three Galaxy workflows that work with relevant livestock viruses and apply reference-based mapping approaches to determine the sample-specific full-length genome. By carefully choosing reference sequences for mapping, it is possible to avoid *de novo* assembly and to integrate fast reference-based mapping tools. Our workflows for poxviruses, avian influenza virus and foot-and-mouth disease virus are all set up for usage with little to no knowledge required from the user. NGS data for all three workflows are required to be produced from an Illumina platform which is one commonly used sequencing technology. We maintain the workflows on the community-based Galaxy platform that aims at providing a transparent, public infrastructure research platform for biomedical applications. Since Galaxy is a globally known platform, it is a reasonable choice for sharing the workflows and making them accessible to researchers with different backgrounds all over the world. The workflow *.ga* format is convertable to other standard pipeline formats (e.g. Common Workflow Language (CWL)) and can be exported for usage in other environments than Galaxy.
Our workflows are designed to streamline the process of whole-genome construction and reduce the time and effort required for this complex task. By leveraging the power of reference-based mapping, the computational challenges of *de novo* assembly are avoided

and accurate results are achieved more quickly. Comparing average run times and memory usage of a virus-specific assembler, `rnaviralSPAdes`, mapping with a standard alignment tool like `BWA-MEM` outperforms assembly with a small viral genome in all means. However, when it comes to larger genomes, assembly can become significantly slower and more resource-intensive process. Reference-based mapping remains a fast and efficient option even in such cases.

The methods used in the developed workflows take into account the accuracy, speed and usability of the choice of alignment and consider virus-specific characteristics throughout the entire workflow.

Adaption of curated Galaxy workflows that were designed for sequence analysis of SARS-CoV-2 samples allows to use tested processes and components within all newly developed workflows. This includes preprocessing, reference-based mapping, quality filtering and consensus sequence construction. Advantages of reusing components of the SARS-CoV-2 workflows are that they have been optimised for specific analysis tasks and have been exhaustively tested with real-world samples by the community. This encourages to leverage the expertise of others and avoid potential pitfalls and errors in the analysis.

Especially with the poxvirus workflow that not only performs fast for the large genome, but also succeeds to map the repeated region of identical ITRs at each genome end, a workflow with many applications has been developed. To avoid ambiguous mapping of the reads to either one of the ITRs, our workflow employs a tiling mapping approach from ampliconic sequencing data that are required to be sequenced in two pools. This method assures the correct read mapping throughout the whole genome and works for all poxviruses. The workflow requires raw Illumina-sequenced reads, an annotated primer scheme in BED format that contains information about the sequencing pools, and a reference sequence for mapping. The user has the option to change default thresholds for the consensus calling step by configuring values for the minimum quality score to call base, the allele frequency threshold to call SNVs and the allele frequency threshold

to call a consensus indel. The two LSDV test samples which were used to assess the workflow quality yield considerable results and complete consensus sequences for each sample, however more and diverse test samples would be needed to determine the ability of the workflow to construct the consensus sequence from raw reads.

TODO: add results/discussion about lsdv/capv experiment. which ref. to choose -> which clade, pre-workflow with vapor for sub-sequence of meaningful positions to determine best ref. among capvs The poxvirus workflow is published and available for download on WorkflowHub and Dockstore, two popular and widely used databases for versioned pipelines. Additional training material and the used primer scheme for Capripoxviruses are linked and provided alongside the workflow, so that a user can be guided through the steps for a deeper understanding of the single workflow steps. Despite the specific amplicon-based sequencing in two pools is no generally applied method in the real world, it can be implemented in laboratories that work with the given workflow and facilitates the read data in the required formats. However in future version of this workflow, alternative sequencing platforms and simplified primer schemes can be considered to make more analyses possible from other raw sequencing data. Especially for Capripoxviruses, a new workflow prior to the presented workflow could help determining the reference sequence that should be used with the sample. In this pre-workflow, the genomic positions that differentiate the LSDV strains and the SPPV and GTPV from each other, could be extracted to find the best fitting reference sequence with a `VAPOR` run against a CaPV sequences database. Using this approach, similar to the FMDV workflows, the user would be asked to infer the most similar sequence from the database search, however it ensures a high-quality alignment and low-error consensus sequence.

A similar challenge in sequence analysis of the avian influenza virus lays in finding the ideal reference sequence for mapping. Assembling raw reads may result in a complete genome too, however is more computationally intensive respecting the size of the genome and may end in misassembly regions where coverage criteria are not met. In the presented AIV workflow, we use a new approach that compiles a hybrid reference from a large database that contains thousands of sequences for each of the eight AIV segments. The user of the

workflow is not required to enter an arbitrarily selected reference genome, instead the workflow automatically finds the most similar sequence from the database by using the search classification tool `VAPOR` and stacks its results together to one reference sequence. While it works fast on a large number of input reads and a query database consisting of thousands of sequences, it selects the highest scoring sequence from the database per segment based on a scoring function on a weighted De Bruijn graph. This method is less biased than mapping raw reads to an assembled contig and yields reliable results with high identities between the query reads and the found sequence for each of the segments in all test samples. Alternative methods to find similar sequences to use as reference, such as read classification by looking up large databases of full influenza genomes are often complex, slow on the large number of reads and require expertise from the user which is not necessarily available. Kraken2 as a taxonomic classification system, also based on k-mer matches similar to `VAPOR`, provides large databases of viral sequences. However, this tool requires more maintenance and computational resources than `VAPOR`, which was specifically developed for influenza reads. Using Kraken2 on one of the large Galaxy server instances comes with the caveat of older database versions that are maintained from within Galaxy and need regular updates, whereas the reference collection that `VAPOR` is based on can be easily maintained and expanded by the user. The overall good quality in the H4N6 and H5N8 test samples emphasises that this method finds the same consensus sequence as the assembly of the reads. These results encourage to test and use the workflow with more AIV samples. The criteria for sequences within the reference genome database assure that complete proteins are captures within each gene by only retaining sequences that contain the complete start and stop codons and no ambiguous nucleotides. Moreover, the amount of sequences per subtype guarantees to capture within-subtype variation which is specifically vital for the detection and identification of nucleotide sites that differ from known strains. The reference collection can be extended with custom references by the user and therefore permits outbreak-specific analysis and taxonomic classification while searching closely related sequences. For investigation on base level on an AIV genome, possible adaptations or reassortment events are captured

and provide a useful starting point for further downstream analyses. The developed workflow provides multiple datapoints for examination within or outside Galaxy, and includes a summary of SNPs on gene level. Phylogenetic classification with `IQ-Tree` based on sequences from the reference database or extended with other samples outside the collection is a part of the workflow to allow insights into relations and clusters among the samples. In future versions of the AIV workflow, this step could be exchanged with the `UShER`, a faster tree generation tool. However, the tool is not available on the Galaxy EU server instance in its high performant version. Additionally, for lineage classification, tools like Pangolin that are SARS-CoV-2 specific could be extended in the future to use it with other viruses. Outbreak tracing and analyses to determine transmission events, intra-host variation and virus origins can be started from these data. Other possible analyses include variant calling, lineage assignment, gene annotation, functional analysis and many more. These opportunities for downstream analysis are more exhaustive than other existing pipelines for genomic sequence analysis of AIV samples, and while subtype identification remains an essential part of surveillance methods in the field, modern monitoring networks for zoonotic animal diseases require detailed study of the sequenced samples on the full genome. Although the workflow design is different to the underlying concept of the SARS-CoV-2 workflow, it picks up its major components while considering the AIV specific viral attributes. Yet there is no integration of variant calling and gene annotation tools, which should be part of the workflow for meaningful analysis of the results. Also, the consensus calling tool `iVar consensus` as opposed to `LoFreq` which is used in the SARS-CoV-2 workflow, is a faster alternative that identifies the most frequent base at each position. Since this first version of the workflow ends by constructing the consensus sequence and marks some further directions for biomedical analysis, depending on research objectives and routine applications it could be expanded in the future.

For large viral genomes, the generation of the genome sequence from NGS reads by assembly is very cost and time sensitive. However, small viral genomes such as FMDV perform reasonably faster in a *de novo* assembly than larger DNA genomes and achieve

sufficiently good results. We therefore developed a Galaxy workflow that provides the necessary steps in two parts which first assembles long contigs from the raw Illumina-sequenced reads and searches the BLASTn database for similar sequences, and secondly builds the consensus sequence from a reference-based alignment of the reads. Although the BLASTn database that is used within Galaxy requires regular updates and is not always up-to-date compared to the BLAST web form provided by NCBI, it nevertheless obtains highly similar sequences from its database to proceed with. However, strains from more recent outbreaks are not guaranteed to be included in the database builds, and results may miss better matches and exclude results from recently examined samples. After the BLASTn search, the user is asked to select a plausible reference sequence from the results, with which the second workflow for mapping and consensus generation can be started. Users without expertise in the field are guided to select a reference from the result by accompanying materials. The trade-off for integrating a *de novo* assembly to determine a suitable reference sequence and achieving a high quality consensus sequence, but investing the time and computational resources in the assembly of the short viral genome is a convenient approach for genomic analysis of FMDV samples. A characteristic of the FMDV genome and of other Picornaviruses is its polyC tract in the 5' end. It is highly conserved across serotypes and both for mapping and assembly approaches poses a challenge to unambiguously map the reads in this location. In the four test samples used for workflow validation, the coverage decreased in these regions and a consensus base could not be called at each position. Despite this caveat, the workflow finds a reference sequence according to the sample serotype, reliably maps the reads to the reference and provides opportunities to adapt tool settings such as coverage thresholds to account for a low number of reads.

The high quality of consensus sequences obtained in the test samples show that the workflow is capable of providing genomic information on base resolution that allows insights into phylogenetic analysis and genomic surveillance. With intermediate steps that split the workflow in two parts, the user can early on detect possible contamination

and co-infection in the sample. This insight leverages the workflow to be a valuable tool to detect genetic variations, mutations and viral characterisation for FMDV samples.

The presented workflows for poxviruses, AIV and FMDV are embedded on the Galaxy platform and provide transparent analysis tools for genomic research on viral samples. They contribute to modern surveillance pipelines in the biomedical field by using fast, accurate reference-based mapping methods to generate consensus sequences that require no user expertise and prior knowledge about the sequenced samples. Despite more validation and improvements should be employed on the workflows, they help the

6 Conclusion

TODO

"good quality assemblies/mappings rely on proper preprocessing and filtering to reduce the number of misassembly events, ambiguous assemblies and of incorporation of sequencing errors into the final assembly"

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] Ministry of Fisheries, Animal Husbandry & Dairying, "Livestock Census." `https://pib.gov.in/PressReleaseIframePage.aspx?PRID=1813802`, 2022. Retrieved: 22nd February, 2023.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] World Health Organization (WHO), "Cumulative number of confirmed human cases for avian influenza A(H5N1) reported to WHO, 2003-2023." `https://cdn.who.int/media/docs/default-source/influenza/human-animal-interface-risk-assessments/cumulative-number-of--confirmed-human-cases-for-avian-influenza-a(h5n1)-reported-to-who--2003-2023.pdf?sfvrsn=c6600b55_1&download=true`, 2023. Retrieved: 17th April, 2023.

[13] World Health Organization (WHO), "Avian Influenza A(H3N8) – China." `https://www.who.int/emergencies/disease-outbreak-news/item/2023-DON456`, 2023. Retrieved: 15th April, 2023.

[14] 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.

[15] 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.

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

[17] 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.

[18] 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.

[19] "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.

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

[21] 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.

[22] 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.

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

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] Illumina, "An introduction to Next-Generation Sequencing Technology," *Illumina, Inc*, 2015. Retrieved: 30th March, 2023.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

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

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

[37] 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.

[38] 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.

[39] D. E. Wood and S. L. Salzberg, "Kraken: ultrafast metagenomic sequence classification using exact alignments," *Genome biology*, vol. 15, no. 3, pp. 1–12, 2014.

[40] D. E. Wood, J. Lu, and B. Langmead, "Improved metagenomic analysis with kraken 2," *Genome biology*, vol. 20, pp. 1–13, 2019.

[41] 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.

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

[43] P. 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

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

[51] 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.

[52] 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.

[53] 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.

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

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

[56] 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.

[57] 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.

[58] 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.

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

[60] 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.

[61] 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.

[62] 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.

[63] 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.

[64] A. Stamatakis, "RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies," *Bioinformatics*, vol. 30, no. 9, pp. 1312–1313, 2014.

[65] 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.

[66] 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.

[67] 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.

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

[69] 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.

[70] 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.

[71] 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.

[72] A. Limasset, B. Cazaux, E. Rivals, and P. Peterlongo, "Read mapping on de Bruijn graphs," *BMC bioinformatics*, vol. 17, no. 1, pp. 1–12, 2016.

[73] 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.

[74] 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.

[75] 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.

[76] 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.

[77] 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.

[78] 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.

[79] 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.

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

[81] 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.

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

[83] 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.

[84] 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.

[85] 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.

[86] 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.

[87] 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.

[88] 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.

[89] 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.

[90] 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.

[91] 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.

[92] 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.

[93] 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.

[94] 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.

[95] 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.

[96] 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.

[97] 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.

[98] 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.

[99] 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.

[100] 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.

[101] 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.

[102] 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.

[103] 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.

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

[105] 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.

[106] 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.

[107] 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.

[108] 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.

[109] 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.

[110] 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.

[111] 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.

[112] 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.

[113] 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.

[114] 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.

[115] 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.

[116] 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.

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

[118] 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.

[119] M. D. Ryan, G. J. Belsham, and A. M. King, "Specificity of enzyme-substrate interactions in foot-and-mouth disease virus polyprotein processing," *Virology*, vol. 173, no. 1, pp. 35–45, 1989.

[120] V. Penza, S. J. Russell, and A. J. Schulze, "The long-lasting enigma of polycytidine (polyc) tract," *PLoS pathogens*, vol. 17, no. 8, p. e1009739, 2021.

[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] 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.

[132] S. Tong, X. Zhu, Y. Li, M. Shi, J. Zhang, M. Bourgeois, H. Yang, X. Chen, S. Recuenco, J. Gomez, *et al.*, "New world bats harbor diverse influenza a viruses," *PLoS pathogens*, vol. 9, no. 10, p. e1003657, 2013.

[133] D. Meleshko, I. Hajirasouliha, and A. Korobeynikov, "coronaSPAdes: from biosynthetic gene clusters to RNA viral assemblies," *Bioinformatics*, vol. 38, no. 1, pp. 1–8, 2022.

[134] S. Biswas, R. S. Noyce, L. A. Babiuk, O. Lung, D. M. Bulach, T. R. Bowden, D. B. Boyle, S. Babiuk, and D. H. Evans, "Extended sequencing of vaccine and wild-type

capripoxvirus isolates provides insights into genes modulating virulence and host range," *Transboundary and emerging diseases*, vol. 67, no. 1, pp. 80–97, 2020.

[135] F. Vandenbussche, E. Mathijs, W. Philips, M. Saduakassova, I. De Leeuw, A. Sultanov, A. Haegeman, and K. De Clercq, "Recombinant LSDV strains in Asia: vaccine spillover or natural emergence?," *Viruses*, vol. 14, no. 7, p. 1429, 2022.

[136] E. S. Tuppurainen, C. R. Pearson, K. Bachanek-Bankowska, N. J. Knowles, S. Amareen, L. Frost, M. R. Henstock, C. E. Lamien, A. Diallo, and P. P. Mertens, "Characterization of sheep pox virus vaccine for cattle against lumpy skin disease virus," *Antiviral research*, vol. 109, pp. 1–6, 2014.

[137] D. B. Smith, G. Meyers, J. Bukh, E. A. Gould, T. Monath, A. S. Muerhoff, A. Pletnev, R. Rico-Hesse, J. T. Stapleton, P. Simmonds, *et al.*, "Proposed revision to the taxonomy of the genus Pestivirus, family Flaviviridae," *The Journal of general virology*, vol. 98, no. 8, p. 2106, 2017.

[138] R. P. Chauhan and M. L. Gordon, "An overview of influenza A virus genes, protein functions, and replication cycle highlighting important updates," *Virus Genes*, vol. 58, no. 4, pp. 255–269, 2022.

Appendix

1 Links to Workflows and Additional Materials

All workflows in Galaxy-specific GA and CWL formats are available on GitHub:
`https://github.com/kciy/msc-thesis/tree/main/workflows/`

1.1 Poxvirus Workflow

- Galaxy EU:
  `https://usegalaxy.eu/u/vyetoria/w/pox-virus-illumina-amplicon`
- GitHub Intergalactic Workflow Commission:
  `https://github.com/galaxyproject/iwc/tree/main/workflows/virology/pox-virus-amplicon`
- WorkflowHub:
  `https://workflowhub.eu/workflows/439`
- Dockstore:
  `https://dockstore.org/workflows/github.com/iwc-workflows/pox-virus-amplicon/main:main?tab=info`
- Galaxy Training Material: **TODO**

- Galaxy history with workflow test run (LSDV samples 20L70 and 20L81):
  `https://usegalaxy.eu/u/vyetoria/h/pox-virus-illumina-amplicon-sample-lsdv`

1.2 Avian Influenza Virus Workflow

- Galaxy EU:
  `https://usegalaxy.eu/u/vyetoria/w/aiv-illumina-analysis`
- Reference Database:
  `https://usegalaxy.eu/u/vyetoria/h/aiv-reference-sequences`
- Galaxy Training Material:
  `https://training.galaxyproject.org/training-material/topics/variant-analysis/tutorials/aiv-analysis/tutorial.html`
- Galaxy history with workflow test run (H4N6 sample):
  `https://usegalaxy.eu/u/vyetoria/h/aiv-seq-analysis-test-h4n6`
- Galaxy history with workflow test run (H5N8 sample):
  `https://usegalaxy.eu/u/vyetoria/h/aiv-seq-analysis-test-h5n8`

1.3 Foot-and-mouth Disease Virus Workflows

- Part 1 (*De novo* assembly and BLASTn search) Galaxy EU:
  `https://usegalaxy.eu/u/vyetoria/w/fmdv-sequence-analysis-part-1-assembly-and-blastn-search`
- Part 2 (Mapping and consensus sequence construction) Galaxy EU:
  `https://usegalaxy.eu/u/vyetoria/w/fmdv-sequence-analysis-part-2-mapping-consensus`
- Galaxy history with workflow 1 test run (SRR17960053, SRR18751245, SRR18685689 and SRR9328470):
  `https://usegalaxy.eu/u/vyetoria/h/fmdv-sequence-analysis-1-2`
- Galaxy histories of workflow 2 test runs:
  - SRR17960053 (Asia-1 sample):
    `https://usegalaxy.eu/u/vyetoria/h/fmdv-seq-analysis-2-2-asia-1`
  - SRR18751245 (A sample):
    `https://usegalaxy.eu/u/vyetoria/h/fmdv-seq-analysis-2-2-a`

  - SRR18685689 (SAT-1 sample):
    `https://usegalaxy.eu/u/vyetoria/h/fmdv-seq-analysis-2-2-sat-1`
  - SRR9328470 (SAT-2 sample):
    `https://usegalaxy.eu/u/vyetoria/h/fmdv-seq-analysis-2-2-sat-2`

1.4 Links to Tool Wrappers

These tools have been wrapped in the Galaxy Tools IUC to make them available on the Galaxy EU server instance.

- `snipit` XML Tool Wrapper:
  `https://github.com/galaxyproject/tools-iuc/tree/main/tools/snipit`
- `VAPOR` XML Tool Wrapper:
  `https://github.com/galaxyproject/tools-iuc/tree/main/tools/vapor`
- `UCSC faToVcf` XML Tool Wrapper:
  `https://github.com/galaxyproject/tools-iuc/tree/main/tools/ucsc_tools/fatovcf`

2 Reference Collection for AIV Workflow

2.1 Filter Criteria and Amount of Sequences

			Filter criteria [1]		**Reference collection**		
Segment		Length [2]	Minimum length	Maximum length	# of sequences	Length range	Mean length
1	PB2	2 316	1 945	2 432	17 714	2 244 - 2 417	2 306.7
2	PB1	2 316	1 945	2 432	17 355	2 265 - 2 414	2 304.9
3	PA	2 208	1 766	2 318	17 569	1 905 - 2 275	2 192.4
4	**HA**	**1 752**	**1 401**	**1 840**	**20 753**	**1 659 - 1 805**	**1 717.5**
5	NP	1 540	1 232	1 617	16 024	1 440 - 1 616	1 529.9
6	**NA**	**1 434**	**1 147**	**1 506**	**17 646**	**1 308 - 1 498**	**1 418.8**
7	MP	1 002	801	1 052	14 684	979 - 1 047	1 001.0
8	NS	865	692	908	15 762	752 - 907	860.9

[1] Minimum cutoff length is 80% of segment length, maximal cutoff length is 105% of segment length.

[2] For strain A/swine/Iowa/18Tosu0505/2018(H1N1) [138].

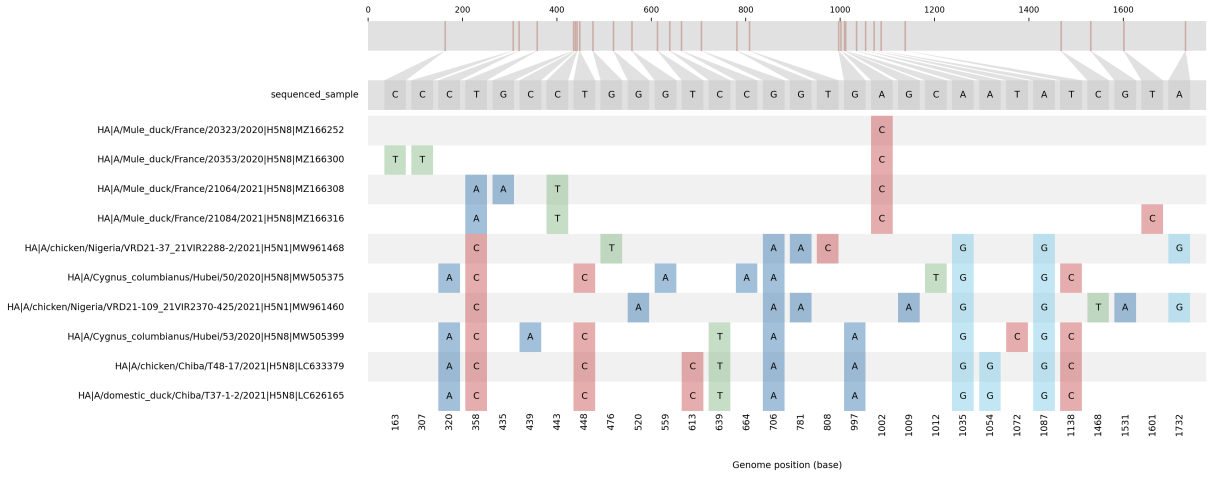
Table 9: Summary of reference collection obtained from search criteria on the NCBI Influenza Virus Database.

2.2 Reference Collection by Amount of Sequence per Subtype

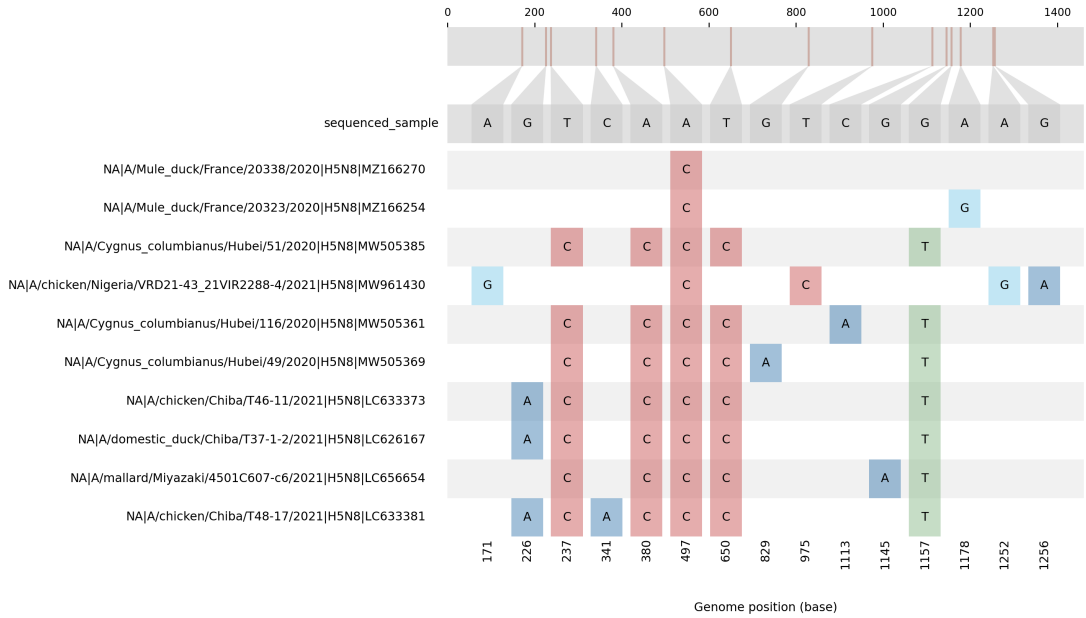
HA Subtype	# of HA sequences	# of NA sequences	NA Subtype	# of HA sequences	# of NA sequences
H1	779	787	**N1**	4 317	3 797
H2	512	450	**N2**	6 346	5 419
H3	1 857	1 724	**N3**	1 585	1 387
H4	1 561	1 520	**N4**	344	304
H5	5 199	4 345	**N5**	523	442
H6	1 817	1 808	**N6**	2 454	2 360
H7	1 898	1 655	**N7**	838	761
H8	169	155	**N8**	2 255	2 077
H9	3 530	2 835	**N9**	1 260	1 092
H10	866	851	**N10**	0	0
H11	749	672	**N11**	0	0
H13	370	310			
H14	42	38			
H15	15	12			
H16	225	177			
H12	337	266			
H17	0	0			
H18	0	0			

Table 10: Amount of sequences per HA and NA gene dataset of reference sequences, divided by 18 HA and 11 NA subtypes present in the AIV reference database, retrieved from NCBI Influenza Virus Database. Subtypes H17, H18, N10 and N11 are excluded for irrelevance in livestock (H17N10 and H18N11 only known in bats [132]).

3 Results of H4N6 and H5N8 Samples for AIV Workflow

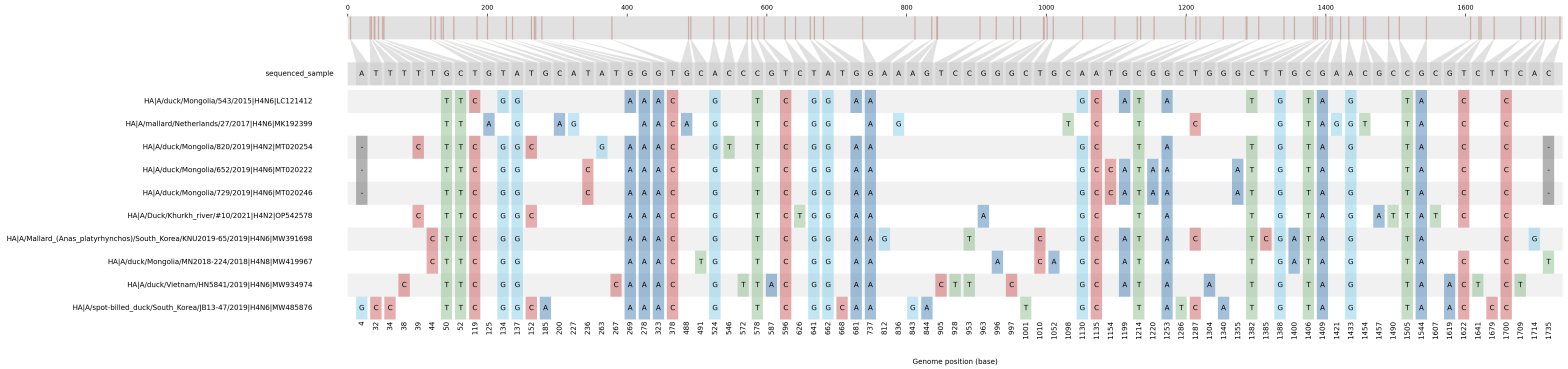


(a) SNPs of HA gene of H5N8 sample.

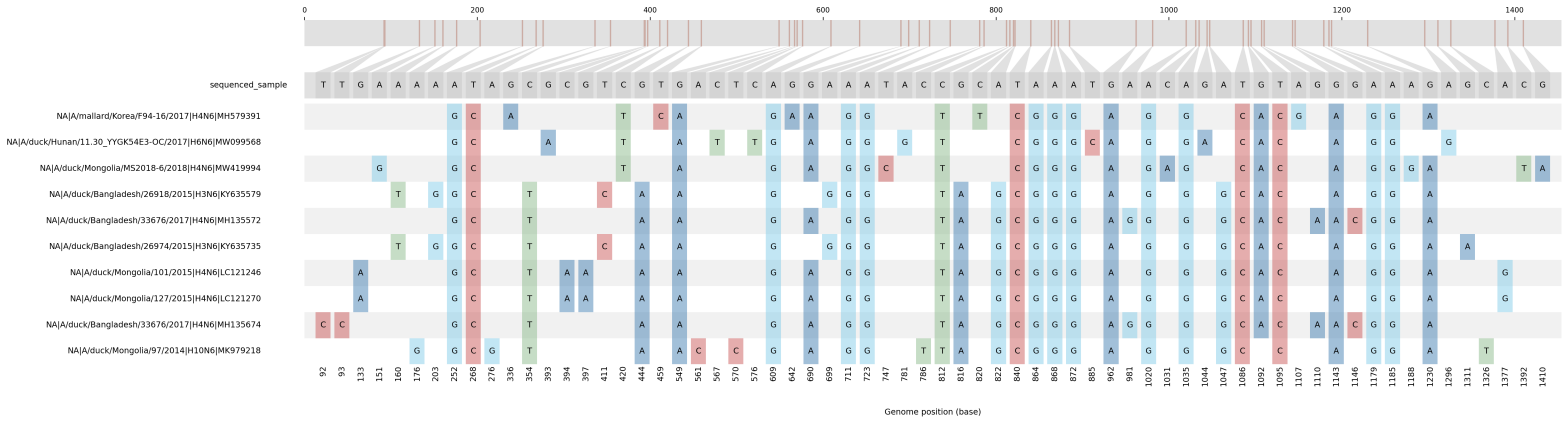


(b) SNPs of NA gene of H5N8 sample.

Figure 10: Visual summaries of SNPs in H5N8 sample. The consensus sequence of the gene is the reference at the top of each plot.

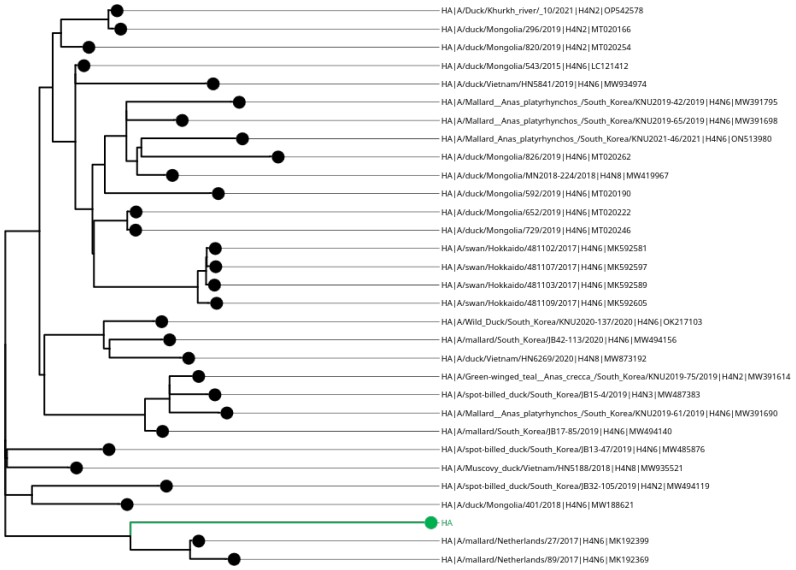


(a) SNPs of HA gene of H4N6 sample.

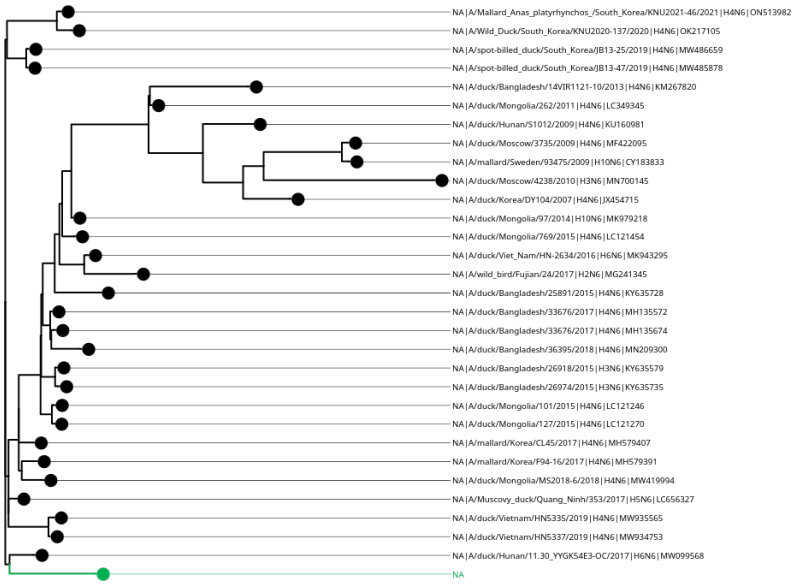


(b) SNPs of NA gene of H4N6 sample.

Figure 11: Visual summaries of SNPs in H4N6 sample. The consensus sequence of the gene is the reference at the top of each plot.

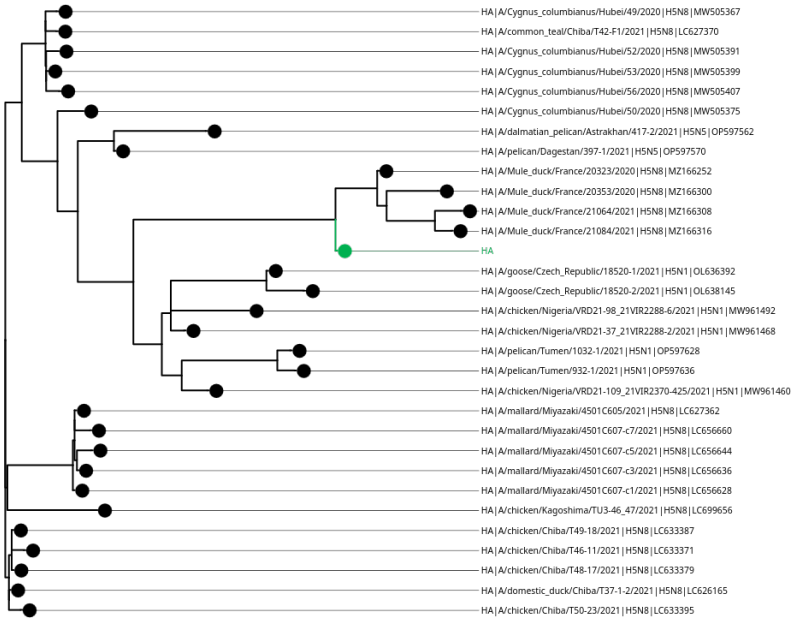


(a) Phylogenetic tree of HA gene of H4N6 sample.

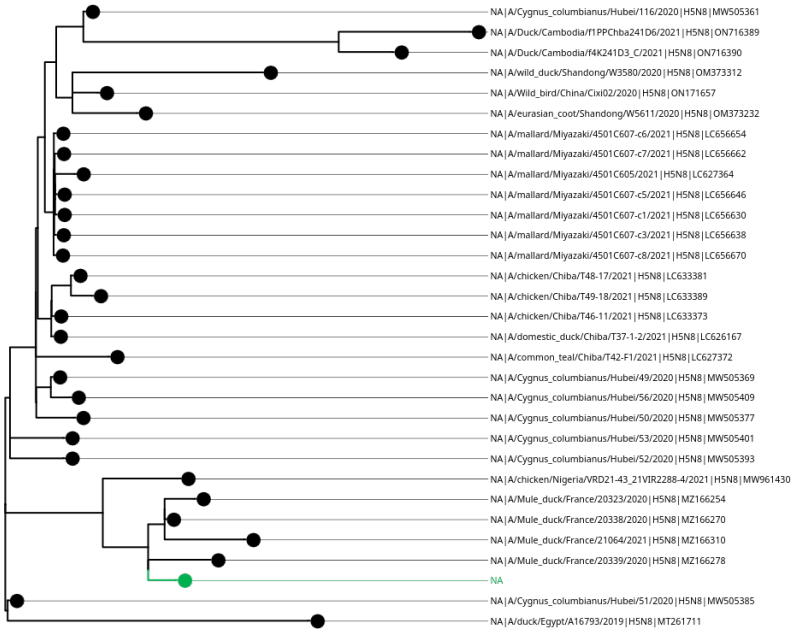


(b) Phylogenetic tree of NA gene of H4N6 sample.

Figure 12: Phylogenetic trees of HA and NA genes for H4N6 sample, indicating linkage to *A/mallard/Netherlands* (HA) and to *A/duck/Hunan* (NA). The consensus sequence is marked in green. The other sequences are the 30 highest scoring sequences from the `VAPOR` run.



(a) Phylogenetic tree of HA gene of H5N8 sample.



(b) Phylogenetic tree of NA gene of H5N8 sample.

Figure 13: Phylogenetic trees of HA and NA genes for H5N8 sample, indicating linkage to *A/Mule_duck/France* in both the HA and NA segments.