Master Thesis

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

Viktoria Isabel Schwarz

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

Albert-Ludwigs-University Freiburg

Faculty of Engineering

Department of Computer Science

Bioinformatics Group

April 28th, 2023

Writing Period

October 28th, 2022 – April 28th, 2023

Examiner

Prof. Dr. Rolf Backofen

Second Examiner

Prof. Dr. Wolfgang Hess

Advisor

Dr. Wolfgang Maier

Declaration

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

Place, Date	Signature

Acknowledgements

Abstract

auf englisch

Zusammenfassung

auf deutsch

Contents

List of Figures

List of Tables

1 Introduction

(to do: put One Health Principle: multidisciplinary perspective on how humans, animals and environmental health are interconnected)

Sharing environments means sharing diseases – this simple relationship expresses how pathogens found in animal populations can spread to humans and have severe impacts. The impact can be as severe as the whole world experienced during the pandemic of Coronavirus Disease 19 (COVID-19) that originated in Wuhan, China in 2019. This highly contagious disease was caused by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), an infectious virus of presumed zoonotic origin [1]. With more than 757.26 million reported cases and more than 6.85 million confirmed deaths as of February 24, 2023 **(to do: update numbers)**, this pandemic is a public health emergency that has caused estimated costs of 16 trillion U.S. dollars. Apart from this, it invoked an outstanding interest in virology research [2].
Since then, professionals from many different fields, i.e. public health specialists, researchers, biomedical staff, bioinformaticians and veterinarians have put even more effort than before into the monitoring of potentially dangerous viral diseases. International managing institutions with a globally distributed netwok work on safe and healthy environments for animal and human populations. The World Organisation for Animal Health (WOAH), founded as Office International des Epizooties (OIE), implements standards in animal health and the handling of zoonoses and other diseases. As an intergovernmental organisation, it supports its members in the prevention of animal diseases of concern. National veterinary authorities must notify

the WOAH in case they detect cases of diseases that are listed by the WOAH. The most important definitions, the significance of animal diseases, impacts and more surveillance measures of diseases spread among domestic animals and humans are examined below.

1.1 Viral Livestock Diseases

Zoonoses and Livestock

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. Some of the most frequent of those diseases include Foot-and-Mouth Disease, African Swine Fever, Avian Influenza and Newcastle Disease. They can spread quickly among animals, and in some cases can be transmitted to humans, making them zoonotic diseases. There are over 200 known types of zoonoses, some of them like rabies being 100% preventable through vaccination and medication [3]. When viral livestock diseases become zoonotic, they pose a significant public health risk, potentially leading to widespread illness and death. A report from the International Livestock Research Institute (ILRI) states that zoonoses account for approximately 2.5 billion illness cases in humans and 2.7 million deaths annually [4]. The Centres for Disease Control and Prevention (CDC) and its U.S. government partners listed the top eight zoonotic diseases of national concern in a report, filing zoonotic influenza and emerging coronaviruses such as SARS and Middle East Respiratory Syndrome (MERS) [5]. This collaborative report is used for focussing on the listed diseases since they are of greatest concern [6]. At the same time, not all livestock diseases of viral origin are zoonotic. Still, around 60% of all known human infectious diseases and approximately 75% of all newly emerging infections are zoonotic [7].

The term livestock is a vague term that generally refers to any breed or animal

population that is kept by humans for commercial or useful purpose. According to the 20th Livestock Cencus of the Department of Animal Husbandry and Dairying, given out by the Indian government, India holds the world's largest amount of livestock with 535.78 million animals as of 2019 [8]. Globally, the ice-free surface that is dedicated to the purpose of livestock whether it is for farmlands or feed production, is up to 26% of the area [9]. Not only food production and economy, but also global trade, the agricultural sector and employment rates highly depend on livestock resources. These numbers illustrate the impressive interconnectedness of the human population with the livestock sector and show how many households and industries worldwide are linked to the still growing and thriving livestock industry. The consequences of a collapse of this important industry would therefore be significant and far-reaching. As the livestock industry is directly affected by the occurrence of zoonoses in both developed and developing countries, it has a strong interest in avoiding any constraints that might be caused by disease outbreaks. Some diseases cause high costs for the industry every year, as many animals are affected or infected and have to be culled.

Historic Outbreaks of Zoonotic Diseases

Historically, zoonoses have shaped serious infectious events. Pathogens that cause zoonotic diseases can not only be viruses (37.7%), but according to surveillance data also bacteriums (41.4%), parasites (18.3%), fungi (2.0%) or prions (0.8%) [10]. Prior to the COVID-19 pandemic, modern zoonotic diseases like Ebola virus disease and salmonellosis have had high infection rates. Influenza viruses cause epidemics each year, and circulate in all parts of the world. There are four types of seasonal influenza viruses (A, B, C and D), while only influenza A and B cause yearly epidemics. Influenza strains appear in zoonotic and human-only spreads, but the viruses can recombine occasionally and cause events such as the 1918 Spanish flu [11, 12]. Especially for poultry, highly pathogenic avian influenza (HPAI) of the H5 subtype is an ongoing threat [13]. Since its first case in China, 1996 it has been detected

in many avian populations, both domestic and wild. It is the avian influenza type with the greatest risk. Even though it has adapted to birds as the specific host, the virus can adapt and be transmitted between humans [14]. Avian influenza has caused ongoing seasonal outbreaks, such as the 2014-15 outbreak in the United States resulting in almost 50 million birds that died as a consequence of an infection or of depopulation [15]. In 2020, there were several outbreaks reported in Europe, almost all with HPAI viruses from the H5 subtype [16]. It mainly affected farmed ducks due to the high density of animals in the facilities and the separation from wild birds due to domestication [16]. The latest outbreak of avian influenza is still ongoing, started in early 2022 and until today, February 23, 2023 **(to do: update numbers and source)** has led to more than 58 million culled or died birds. It is reported in 37 countries and so far, 6 human infections were reported in this outbreak [17]. This number is not nearly as high as for the animals affected, but considering that during the last 20 years, there were fewer than 900 confirmed cases of H5N1 in humans and the mortality rate of 50%, each human infection is at risk [17].

Risk Factors and Impact of Disease Outbreaks

Reasons for recurring huge outbreaks of viral diseases in animal confinements come from the good circumstances for virus transmission. In general, animal husbandry practices have evolved in the sense that domestic animal species are raised in relatively small and usually confined spaces at a high density. This domestication has given plenty of opportunities to develop more pathogens of viral or bacterial origin over time as seen in recent years. Then, the spread of international trading of farm animals has amplified the number of infected animals and the number of infectious diseases. As transmission routes can differ depending on the disease, the other factor is how easy the infectious agent spreads (transmissibility). Vector-borne diseases are transmitted by living organisms that transfer pathogenic microorganisms to other, uninfected animals or humans. Vectors can be mosquitoes, fleas or ticks.

Among others, the WHO identifies major globally present vector-borne diseases as malaria, dengue, yellow fever and Zika virus disease [18]. Other transmission modes are direct contact airborne transmission. Environmental factors such as a high temperature, humidity and precipitation can facilitate a virus to spread and keeping it alive [19]. Overpopulation, inadequate food and water supplies and mass migration of populations pose additional risks for transmission of animal diseases.
(to do: add paragraph about wildlife; interface) Outbreaks of livestock diseases do not only affect animal and human health, but also cause high economic losses. Restrictions and containment measures, as well as the culling of animals in the case of confirmed cases of listed diseases, lead to a loss of income for farmers – since livestock and their products, such as milk, eggs or meat, are used for further production, other businesses that rely on these products are also affected by disease outbreaks. Even if infected animals do not die or have to be culled, the medium- and long-term consequences of infection can affect the health of the animals. This can lead to poor growth or poor production and feed conversion. Another impact of depopulating infected animal populations is the loss of biodiversity [20, 21]. Wildlife populations of endangered species experiencing disease outbreak can be decimated, leading to ecological imbalances and interference with natural food chains [22, 23, 24].
As shown, the spread of viral disesaes among animal populations can have enormous impacts on dependent industries, individuals and populations.

Notifiable Animal Diseases

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

risk for costly outbreaks. This is due to severity, transmissibility or environmental stability as discussed above.

Reports of illness cases of animals are 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 may have impact on livestock health [25]. As the surveillance of viral animal diseases is still of highest priority in order to avoid costly and dangerous outbreaks, this topic is discussed in more detail in the following introductory chapter.

1.2 Prevention, Surveillance and Control

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

To avoid the impact that a disease outbreak can have, the best method is to prevent the disease in the first place. This gives rise to the principle of prevention, which sees its main task as reducing the overall risk of a virus spreading. Corresponding measures can be vaccinations and the establishment of hygiene standards. For viral material that recombines over time, as the number of infections increases, the potential for the virus to exploit host cell genes that favour viral growth and survival may be high [26]. Therefore, it seems logical to reduce the overall number of infections. Other disease prevention practices primarily include disinfection and good animal husbandry. Practicioners in the field or in veterinary clinics are obliged to follow this principle of prevention. In-depth strategies to prevent viral diseases depend heavily on the characteristics of the virus, taking into account transmission mode(s), environmental stability, zoonotic risk and pathogenesis. Exclusion of livestock and the use of vaccines from potentially infected flocks is increasingly practised [26]. The spatial spread of disease can be contained through quarantine, testing and regular

inspections of imported animals.

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

Common methods for animal diseases surveillance include notifiable diseases reporting, laboratory-based surveillance and population-based surveillance. General awareness among veterinary diagnosticians and practitioners is another key to an effective surveillance system. Most countries have their own national veterinary authorities, coordinated by the WOAH to enable a coordinated exchange of information [27].

It is vital to analyse collected data promptly in order to influence necessary follow-up actions. National databases may contain reliable and annotated data, but they often reflect information gathered several weeks or months ago. On the other hand, early warning signs of a potential disease outbreak may be found in local media reviews, unusual social media activity, and unverified individual reports on the internet. However, such sources may provide well-intentioned but inaccurate information. Timely action and communication of information, particularly to local veterinary practitioners, is a crucial component of effective surveillance systems. Nevertheless, it is important to exercise caution to prevent unnecessary public concern.

One important component of modern and accurate surveillance systems of viral diseases is the access to relevant data. Technologies to produce DNA sequencing data have developed to be very cost and time efficient, which makes the study of infectious diseases better and faster. At the same time, the amount of DNA sequencing data produced with next-generation sequencing (NGS; also known as high-throughput sequencing, HTS) platforms prove this change. NGS platforms include IonTorrent,

Illumina HiSeq/MiSeq (for different read lengths) and Oxford Nanopore Technologies (ONT). Advances in the biotechnological application and evaluation of these data are revolutionalizing the field of studying these data on its molecular level [28]. Sequencing technologies take a key role nowadays 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 characterization of new microbial material. More detailed methods that are used for viral animal disease surveillance with NGS-based technologies are described in Chapter 2.

1.3 Motivation and Objectives of the Thesis

Bioinformatics and data analysis are crucial for understanding and monitoring viral diseases. However, there is a lack of knowledge and resources in many parts of the world. This is particularly true for poorer countries with small laboratories and national health organizations that are not well established. Nonetheless, efforts are made to establish global networks such as the Zoonotic Disease Integrated Action (ZODIAC). It is an initiative by the International Atomic Energy Agency (IAEA), launched in 2021, with five major objectives: (1) Strenghthening member states' detection, diagnostic and monitoring capabilities, (2) Developing and making novel technologies available for the detection and monitoring of zoonotic diseases, (3) Making real-time decision-making support tools available for timely interventions, (4) Understanding the impact of zoonotic diseases on human health and (5) Providing access to an agency coordinated response for zoonotic diseases [29]. In collaboration with technical experts from different fields and from all over the world, and to support the Veterinary Diagnostic Laboratory (VETLAB) Network, the ZODIAC project has the resources to provide standardised, easy-access, public, and integrated pipelines for virus surveillance on a long-term. This will enable laboratories and veterinarians to monitor and analyse their samples more effectively, leading to early detection and prevention of viral livestock diseases.

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

This work is part of the ZODIAC project and supports pillar (2) in the development of integrated pipelines that enable laboratories, veterinarians and other health professionals to analyse their data from samples obtained with HTS technologies. The zoonoses studied are avian influenza A for subtype identification and a poxvirus pipeline for determining poxvirus genomes sequenced as half-genomes in a tiled-amplicon approach. This pipeline has been tested with samples of lumpy skin disease virus, but can also be applied to other poxviruses.

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

2 State-of-the-Art

Surveillance measures (currently taken); development of vaccines to protect humans is one central pillar of why the constant; ensure adequate hygiene, know and be alert to symptoms; vaccinate animals if possible and necessary –> control/surveillance of zoonoses is crucial

SARS-CoV-2 high similarity to SARS-CoV virus
animal origin is suspected (!) globally, in science, advances in bioinformatics: new surveillance techniques, e.g. clinical testing for self-monitoring
systematic global collection of data that allows health professionals and policymakers to react with appropriate measures at hand to protect the public

molecular surveillance studies: genetic analysis is an integral part of animal disease surveillance.

surveillance systems include classical phylogenetic methods to genotype novel emerging strains, classify viral lineages or assess tree topologies to distinguish between novel and emerging strains

approaches to assess correlations between similarities of nucleotide sequences and related epidemiological characterstics

investigate spatio-temporal and evolutionary dynamics of the virus isolates in a disjointed analytical framework

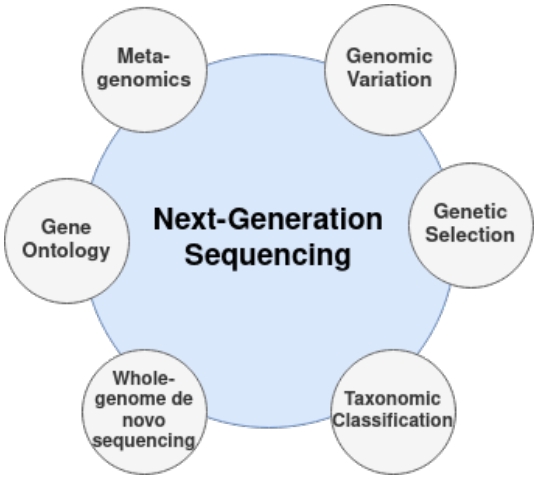

Figure 1: Overview of Applications of Next-Generation Sequencing Technologies in diagnostic virology.

2.1 High-throughput Technologies in Diagnostic Virology

2.1.1 Overview of NGS Platforms and Applications

2.1.2 Detection of Viral Pathogens

- NGS-based methods: VIDISCA protocol for SARS-CoV in 2004 - metagenomics-based strategies: higher sensitivity (compared to microarray-based assays), detect full spectrum of viruses - NGS-based: Illumina GA platform for detection of new viruses (influenza A) and de novo assembly (only works with high enough number of reads generated) - bats: coronavirus consensus PCR and unbiased HTS (high-throughput pyrosequencing) -> reveal presence of sequences of new coronavirus related to SARS-CoV

2.1.3 Data Analysis Issues

no standardization of techniques

no knowledge sharing/best practices/common source of knowledge

livestock diseases do not receive much attention from bioinformaticians and therefore not enough monitoring; although in regions where outbreaks have larger impacts on economy and other fields, the surveillance is crucial for prevention and control.

This emphasizes the importance of free and easy-access platforms like Galaxy that entitle professionals to analyse their samples. Technical know-how to develop and maintain servers for analysis of NGS data is not a global standard and also not needed everywhere when multidisciplinary approaches can be shared through platforms.

costs are relatively high, especially for countries or health organizations in development countries that are still affected by diseases they want to monitor

format of sequencing data -> data accuracy

Illumina shows low coverage of AT rich regions [30]

* chimerical sequences (artifacts originating from joining sequences), point mutations, insertions/deletions which occur during reverse transcription, PCR amplification or sequencing itself * cleaning step or filtering phase removes low-quality reads from the dataset, while the error correction separates true variants from those due to experimental noise. -> idea that errors are randomly distributed with low frequency, and real mutations are clustered and their abundance are quantified

Each application of software with NGS data requires expertise in resolving limitations and drawbacks of specific methods. This in turn requires bioinformatic skills and the careful interpretation of results. Still, NGS provides a large pool of methods which eases the tasks, although available algorithms for genome assembly and amplicon analysis have drawbacks and limitations [31].

2.2 NGS Methods for Poxviruses

difference/similarities between monkeypox, goat pox, sheep pox etc.
advances in surveillance (4 genera of poxviruses are zoonotic and may infect humans)

classification based on phenotypic characteristics.

2.2.1 Poxviruses

2 subfamilies, one infects vertebrates and one infects insects -> 10 genera infect vertebrates

characteristic ITR that is left out in other pipelines (Yale University primer scheme starts after and ends before ITR)

with monkeypox outbreak in 2022, professionals of the field had the event to examine another pox virus, seeing similarities among pox viruses motivates to extend existing pipelines for data analyses to work with samples of all pox viruses

one poxvirus to mention in more detail: Lumpy Skin Disease (LSD) is caused by the virus belonging to the Capripoxvirus (CaPV) genus within the family Poxviridae, subfamily Chordopoxvirinae [32]. The LSDV genome is a double-stranded linear DNA molecule of circa 151 kilobasepairs in length. It contains between 147 and 156 open reading frames. With a sequence identity of over 96% with the other CaPV genus members sheep pox and goatpox, the lSDV genome is very similar to the other CaPV genomes [33]. These three viruses of the capripoxvirus genus are the most serious poxvirus diseases of livestock in terms of economic losses in the case of an outbreak.

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 for cattle against LSDV outbreaks. Similar to other poxviruses, the LSDV genome consists of a central coding region which is bounded by two identical inverted terminal repeat (ITR) regions with a length of circa 2,400 basepairs at both ends of the genome. This is a key characteristic to consider during reconstruction of the genome.

capripox not transmissable to humans -> NOT a zoonosis

viral disease that affects cattle, transmission through blood-feeding insects such as ticks or some species of mosquitoes and flies.
* symptoms
vaccination programme (mass-vaccination by EU commission protecting >1.8million cattle)
spreading in African countries, since 2012 present in middle east to south-east europe

Zoonotic: A disease of humans caused by pathogen coming from non-human host and vice versa.

Transmission modes, limited host range

2.2.2 Application of NGS Technologies in Poxvirus Diagnostics

https://www.sciencedirect.com/science/article/pii/S0166093422000118 explains Primer scheme and why tiling amplicon approach makes sense even for large genome size of CaPV genome and complex structure with repetitive ITR regions

2.3 NGS Methods for Avian Influenza Virus

2.3.1 Avian Influenza Virus

informally known as bird flu, the avian influenza is a viral infectious disease whose hosts are wild waterbirds.
occurring in two variants determining the severeness, therefor low/highly pathogenic and in a variety of subtypes, that are composed by two viral segments H1-H16 in combination with N1-N11

* symptoms

* LPAIV, HPAIV

* Etiology: virus composition, taxonomy, origin, mutation rates

Human Influenza Virus: AIV has more subtypes as there are more prevalent subtypes in many different populations; more variants

2.3.2 Application of NGS Technologies in Avian Influenza Virus Diagnostics

SARS-CoV-2 tracking is of huge global interest, resulting in a highly regarded topic with ongoing scientific activity in terms of publications
includes established institutions in bioinformatics that hand out approaches, guidelines, recommendations to govern outbreaks of viral livestock diseases. includes comprehensive pipelines for bioinformaticians, veterinarians and other health professionals. major platforms that offer exhaustive approaches to analyse genomic samples from infected stock.

efforts for end-to-end tools and pipelines for sequencing data

INSaFLU, ViReflow? (SARS-CoV-2 samples)
VAPOR (ref datasets)

* INSaFLU (inside the flu) -> for influenza, NGS towards metagenomic virus detection, routine genomic surveillance,

* Nextstrain -> for RT SARS-CoV-2, Influenza, Ebola pathogen populations * Kraken2 -> taxonomic sequence classifier (using database and k-mers of FASTA sequences) * VirFind (by Arkansas High Performance Computing Center) -> for fasta/Illumina fastq files, to detect new samples (trimming, mapping to ref, de novo assembly, Blastn, Blastx) * ARTIC Network -> RAMPART for Ebola, yellow fever virus (read assignment, mapping, phylogenetic analysis on ONT data) * IRIDA -> Integrated Rapid Infectious Disease Analysis for NGS data e.g. de novo assembly (FLAsh, SPAdes, Prokka)

tracking viruses using genomic sequence data collection; effective surveillance does not require exhaustive case surveillance, instead the collection of enough data from representative populations. This enables health professionals to detect newly evolved variants and to monitor trends in the circulating variants.
wastewater

3 Materials and Methods

3.1 Galaxy Platform

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

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

The Galaxy community is thriving, with over 124,000 users who also contribute to subcommunities. This community is dedicated to supporting scientists and researchers, and it plays a critical role in the growing importance of the platform. The public server for analyses provides access to public datasets, enabling scientists to perform their analyses with ease. 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 all tools used on the platform.

Important contributions of Galaxy, as stated by the Galaxy Community (2022), include Vertebrate Genome Project assembly workflows and collaborations on SARS-CoV-2 research. Another toolkit leveraged in Galaxy is Galaxy-ML, a set of tool that provides a suite for analyses based on machine learning. With growing publicity, more topics are covered by and moved to Galaxy.
Workflows that are available on and accepted by the Intergalactic Workflow Commission (IWC; `https://github.com/galaxyproject/iwc`) are conform with the community's best practice standards and tested on the latest Galaxy release. Dockstore and WorkflowHub automatically publish the IWC workflows and guarantee the availability in a Docker-based environment on Dockstore [34] and on the workflow collaboratory WorkflowHub [35].

Galaxy is a highly professional platform that offers an impressive array of resources for researchers in the biomedical field. It has grown in scientific significance over the years, and its value is not just limited to in-person events. It has contributed to over 5,700 scientific publications and has many tutorials available for researchers to use. Training material and ready-to-use workflows facilitate professionals and beginners in the field to use Galaxy for their research purposes.

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

Germany (for UseGalaxy.eu), Texas Advanced Computing Center (for UseGalaxy.org) and Genomics Virtual Laboratory, formerly at the University of Queensland (for UseGalaxy.org.au). These main public servers are synchronized in their tools and set of reference tools. [36]

3.2 Workflow Design

3.2.1 SARS-CoV-2 Pipeline as Baseline

annotated variants are of interest
description of basic steps

well-established workflow, includes 'minimal' steps:

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

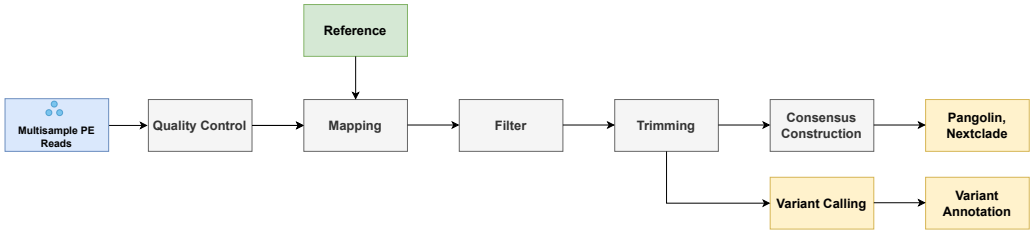

Figure 2: Simplified SARS-CoV-2 ARTIC PE reads iVar-based workflow.

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

3.2.2 Requirements

which problems should the pipeline solve?

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

Requirements for LSDV Workflow

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

Requirements for AIV Workflow

- reference for each of the 8 segments has to be chosen
- align reference of each segment with consensus sequence for phylogenetic analysis
- snipit for visualisation of SNPs
- trimming would dismiss too many of the already short reads

a tool to get closest reference

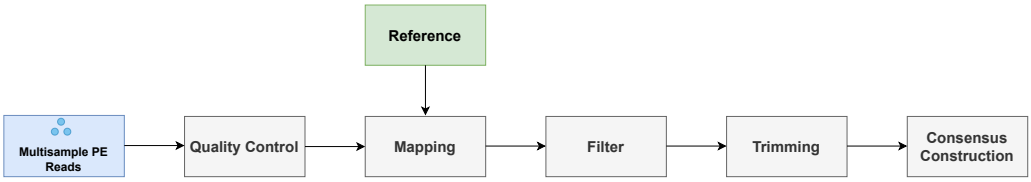

Figure 3: Simplified minimal ARTIC PE reads iVar-based workflow.

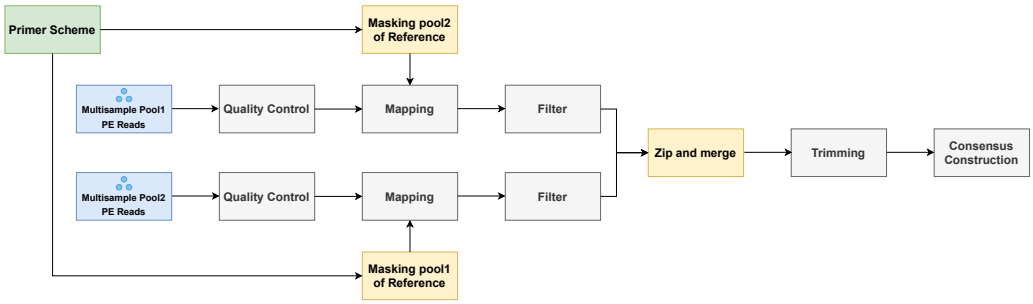

Figure 4: Simplified LSDV ARTIC PE reads iVar-based workflow.

3.3 Workflow Development

"Reference-based genomic Surveillance" (INSaFLU)

3.3.1 Pox Virus Illumina Amplicon Workflow

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

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

LSDV genome has its central coding region bounded by identical inverted terminal repeats, containing 156 putative genes. the repeat of the ITRs would make any mapping in these regions ambiguous. need to part the reads in two pools and do mapping in two parts: N-mask the reference

Efficiency: Assembly vs. Mapping!!; efficiency (hier nur kurz, ausführlicher in Diskussion). Wenn Ziel viele Samples/flächendeckende Überwachung ist, dann ist Assembly zu teuer. Im großen Stil soll das hier genutzt werden)

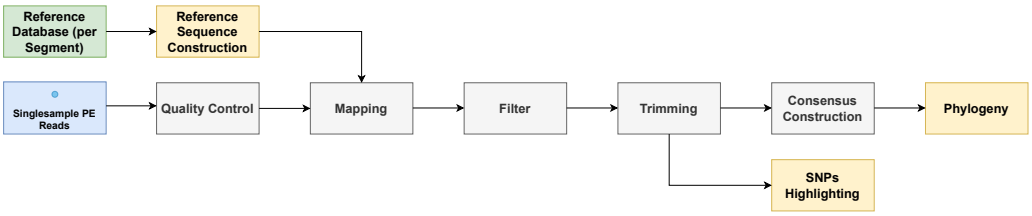

Figure 5: Simplified AIV ARTIC PE reads iVar-based workflow.

building index is expensive (BWT)

3.3.2 AIV Illumina Amplicon Workflow

explain VAPOR here

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

3.4 Workflow Evaluation

3.4.1 Evaluation of AIV Workflow Using Test Datasets

Sciensano s4+s8, Tunesian?

3.4.2 Evaluation of Pox Virus Workflow Using LSDv Test Dataset

Sciensano by Elisabeth Mathijs

4 Results

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

4.1 Pox Virus Illumina Workflow

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

4.1.1 Results for LSDv Datasets 20L70 and 20L81

4.2 AIV Workflow

IWC link if existent. point out output for downstream analysis

4.2.1 Results for Dataset U2012100-n21_S8

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

4.2.2 Results for Dataset U2008751-n5_S4

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

5 Discussion

5.1 Contribution to the Field

single sample vs. multi sample (reality check, what is needed?)
further pox viruses, pipelines can be more or less easily applied/adjusted
limitations
LSDV für alle Pox-Viren interessant
AIV downstream alles. Es wäre gut key minor assets zu highlighten, die auf Adaption bei Säugetiere hinweist -> Databases werden benötigt zum Abgleichen ob ein Isolat mutiert ist?
Generell will man auf Aminosäure-Ebene annotieren (meiste Information)

Stammbäume zugänglich öffentlich, wäre gut die öffentlich zu haben, auch detailliert also >1 Sample pro Land, sehr feingliedrig um echt einordnen zu können (1 Isolat pro Kontinent bringt nicht so viel)

5.2 Future Directions

further validation and improvement of the developed pipelines, expansion to other viral livestock diseases, integration with existing surveillance systems; expand the

VETLAB network to entitle even more professionals to professionally analyse their samples.

AIV workflow offers many possible directions for downstream analysis:

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

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

6 Conclusion

Summary of objective and discussion

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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[30] O. Harismendy, P. C. Ng, R. L. Strausberg, X. Wang, T. B. Stockwell, K. Y. Beeson, N. J. Schork, S. S. Murray, E. J. Topol, S. Levy, *et al.*, "Evaluation of next generation sequencing platforms for population targeted sequencing studies," *Genome biology*, vol. 10, no. 3, pp. 1–13, 2009.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

Appendix

* table with WF tools, versions?

* results of datasets