THE UNIVERSITY
of EDINBURGH

New Approaches to Characterise Viral Pathogens in Aquaculture

A thesis presented for the degree of Doctor of Philosophy at the University of Edinburgh

2020

THE UNIVERSITY *of* EDINBURGH

Michael D. Gallagher

MSci - University of Aberdeen

Declaration

I hereby declare that this thesis was composed by myself. The research presented within this thesis was conducted by myself between October 2016 and March 2020 with the following exceptions: i) the Sanger sequencing of SAV PCR amplicons carried out by Dr Iveta Matejusova at Marine Scotland Science in Aberdeen (chapter 2); ii) targeted sequence capture library preparations and Illumina sequencing of SAV-infected tissues, carried out by the Centre for Genome-Enabled Biology and Medicine (CGEBM) in Aberdeen (chapter 4); iii) the Snakemake pipeline was initially constructed by Dr Alicia Bertolotti at the University of Edinburgh (Chapter 5). This Thesis has not been submitted, in whole or in part, in any previous application for a higher degree. All sources of information have been acknowledged in the text.

Michael D. Gallagher

March 2020

Contents

List of Figures

List of Tables

List of Abbreviations

AMR	Anti-microbial resistance
BLAST	Basic Local Alignment Search Tool
Bp	Base pairs
cDNA	Complementary DNA
CyHV-3	Cyprinid herpesvirus 3
DNA	Deoxyribonucleic acid
dsRNA	Double stranded RNA
ESS	Effective sample size
FAO	Food and Agriculture Organisation
HIV	Human immunodeficiency virus
HPD	Highest posterior density
HPR	Highly polymorphic region
IGV	Integrative Genomics Viewer
ISAV	Infectious salmon anaemia virus
MCMC	Markov chain Monte Carlo
MRCA	Most recent common ancestor
mRNA	Messenger RNA
NCBI	National Centre for Biotechnology Information
NGS	Next-Generation sequencing
NOV	Norovirus
OIE	World Organisation for Animal Health
ONT	Oxford Nanopore Technologies
ORF	Open reading frame
OsHV-1	Ostreid Herpesvirus 1
PCR	Polymerase chain reaction
PD	Pancreas disease
PMCV	Piscine myocarditis virus
PRV	Piscine reovirus
qPCR	Quantitative PCR
RNA	Ribonucleic acid
SAV	Salmonid alphavirus
SD	Sleeping disease
SH-aLRT	Shimodaira–Hasegawa approximate likelihood ratio test
SNV	Single nucleotide variant
SPDV	Salmon pancreas disease virus
+ssRNA	Positive single-strand RNA
-ssRNA	Negative single-strand RNA
SVCV	Spring viremia of carp virus
TiLV	Tilapia lake virus
VHSV	Viral haemorrhagic septicaemia virus
WGS	Whole genome sequencing
WSSV	White spot syndrome virus
YHV	Yellow head virus

Acknowledgments

First and foremost, I would like to thank my PhD supervisor, Dr Dan Macqueen, for his support and mentorship that stretches back to my Honours project, 2 years before I even started this PhD. Without him, I would not be the scientist I am today. I am also extremely grateful for my co-supervisor Dr Iveta Matejusova from Marine Scotland for not only providing much needed funding and samples, but also grounding some of my more outlandish ideas in reality.

I was exceptionally lucky for the collaborators I gained throughout this project. Thank you to Dr Neil Ruane from Marine Institute Ireland whose generosity in donating samples time and time again was crucial to the success of this project. Thank you also to Dr Alex Douglas at the University of Aberdeen whose support and guidance when it came to getting codes and computational methods working, including giving me access to his personal computing resources – indispensable in the early days of this project. Finally a particularly huge thank you needs to be given to my PHARMAQ collaborators – Elin Petterson, Øyvind Haugland, and especially to Marius Karlsen – for taking an interest in my work from the beginning and being the reason I was able to translate my academic work into applied research relevant to the aquaculture industry. Without this collaboration, this project would have been much more difficult!

I would like to thank all the people I met at both the University of Edinburgh and the University of Aberdeen who made working in both places fantastic experiences. To the 'Macqueen Minions', Alicia for always having my best interests at heart and being ready to bribe me with food/wine to meet those interests, and Manu for his unending patience for my self-inflicted spiralling sessions and a constant flow of coffee to deal with the stress at the end of the project. To Craig, who to this day I don't know how he put up with living with me for so many years and was always ready for beers and pizza! And to Rob, Rose and Fiona who have made the move to the Roslin so enjoyable and are always supportive, whether that takes the form of coffee, junk food or drinking! Finally to my family who have supported me constantly through this process and never doubted that the end would come, despite my constant doomsday predictions!

Abstract

Aquaculture is the fastest growing food-producing sector in the world and of considerable economic, cultural and environmental relevance. This sector will be vital to achieving future food security demands, but its continued sustainable expansion is severely threatened by infectious diseases, with viral diseases amongst the most problematic to control. Unlike farmed livestock, fish are generally reared in open systems with constant circulation between farms and the natural aquatic environment. This routinely exposes the animals to naturally occurring viruses in the water, both pathogenic and non-pathogenic, which are generally uptaken through mucosal surfaces (i.e. gills and gut surfaces). However, with the increase in globalisation, aquatic species are frequently farmed in non-native habitats, thus exposing them not only to the pathogens present in wild relatives of the same species, but to pathogens of other species in their introduced habitat. Moreover, wild fish are threatened by viral disease outbreaks on fish farms due to the high density of individuals available to carry and transmit the pathogen. Characterising viral infections is therefore important to support the prevention and control of disease outbreaks, as understanding the disease agent enables both fish farmers and regulating agencies to tailor appropriate mitigation strategies.

The routine use of whole genome sequencing to screen infected animals is not yet commonplace in the aquaculture industry, where genetic screening of viruses is largely done using PCR for 1 or 2 marker genes. However, the 'genomic surveillance' approach has been used to great effect in cases of disease outbreaks relevant to human health, and could be applied in aquaculture to enhance the resolution of molecular epidemiology investigations and diagnostic tests. Moreover, with the under-researched genetic diversity of aquatic viruses, significant advances in the understanding of host-pathogen interactions could be achieved with a denser and better curated genomic database of viruses. To address these knowledge gaps, I have developed and optimised several approaches to characterise aquatic viruses up-taking various sequencing methods depending on the resolution required for the specific study, using salmonid alphavirus (SAV) as a primary study system.

To rapidly and accurately generate consensus-level genomes of specific pathogenic viruses, I developed a targeted PCR approach using overlapping long amplicons tiled across the SAV genome for full coverage. These amplicons are sequenced on the Oxford Nanopore Technologies MinION long-read platform. An analysis workflow was then optimised to generate consensus genomes while maintaining capability to discover SAV subtype-level co-infections by simultaneously mapping to multiple reference sequences. This approach can generate highly accurate consensus sequences (as judged by independent Sanger sequencing) and detect co-infections of strains with ≥ 95% pairwise identity over a 2kb region, even when

minor infecting strains are present at just 5% frequency. This approach was used to investigate the population dynamics and phylogeography of the SAV3 epidemic in Norwegian aquaculture, revealing repeated seedings of SAV3 from 'source' to 'sink' counties.

To characterise viral genetic diversity within a host, I applied a targeted sequence capture strategy to obtain SAV genomes at high coverage (using Illumina technology) from infected fish using both pooled and individual tissue samples. This approach utilises RNA baits to capture and enrich for specific DNA (or cDNA) strands in a sample, and allows for greater sequencing efficiency. These baits, while designed from specific templates, are less specific than PCR primers and can tolerate a certain amount of template mismatches, thus capturing all genetic variation of a specific viral species within a sample. This approach was used to compare the genetic diversity of SAV in farmed Atlantic salmon and rainbow trout, in addition to two wild flatfish species, sampled from multiple regions in Scottish and Irish waters. In the same study, I provided evidence of complex infections on single fish farms, and for co-infections within single wild fish.

Finally, I developed a pipeline to detect viral infections in metagenomics samples, which can be applied even when the infectious agent is unknown. This involves an optional step of mapping to the host reference genome to increase efficiency of later steps, assembly of the remaining reads with a transcriptome assembler, and identifying viral transcripts using homology-based tools. Before implementation, this pipeline was benchmarked against several datasets, including a simulated virome and a simulated co-infection of two strains of the same virus. It was also tested against datasets with known pathogens, resulting in similar efficiencies of detection as a mapping-based approach. Finally the pipeline was used on datasets with unknown viromes to demonstrate its applicability to detect novel viral species.

Overall, my research has led to the development of several cutting-edge approaches for the genomic analysis of aquatic viruses and other pathogens, and helps clarify which approach is most useful in different epidemiological settings. I also demonstrate that genome-wide analyses of viral pathogens impacting salmonid aquaculture generates valuable additional information on viral diversity compared to standard surveillance methods using particular marker genes, advocating for route use of genomic approaches in this sector.

Lay Summary

One of the biggest threats to the expansion of the aquaculture industry is the spread of infectious diseases, particularly viruses, as many viral diseases in fish lack effective therapeutics. As the fastest growing food-producing sector, the sustainability of this sector is hugely impactful both economically and as a useful tool to combat food security concerns. However as seen with several human viruses in recent years including Ebola, Zika, and the recent outbreak of novel coronavirus, being able to identify and characterise the viral pathogen rapidly and accurately is crucial to mitigation and disease control efforts. Having a diverse toolkit of methods when handling infectious diseases is key as it gives researchers and industry room to tailor their approach to handling specific outbreaks. This Thesis outlines several methods of characterising viruses infecting farmed fish. These methods have been developed for a range of sample sizes, viral species and importantly affordability. In Chapter 2, I outline a rapid and affordable approach that accurately generates whole genome sequences of specific viral species, useful when the pathogen is already known and widespread screening is the goal. Chapter 3 then implements this approach to investigate the current epidemic of salmonid alphavirus subtype 3 in Norway and how the virus is moving between salmon-producing regions of the country. Chapter 4 presents a more comprehensive method of characterising specific viral species or groups of species, able to detect co-infecting strains at low frequency within individual fish. This method was used to identify the circulation of multiple viral strains of the same species both within salmon farm sites and co-infecting individual wild fish. Finally Chapter 5 outlines an unbiased method of detecting viruses even when the investigator hasn't already identified the pathogen, or when characterising pathogens not seen before. The findings of this Thesis contribute to the establishment of a more comprehensive toolkit available for investigating and controlling infectious viral diseases in aquaculture.

Chapter 1. General Introduction

Summary

In this Thesis, I aim to develop and optimise approaches to characterise viral pathogens in aquaculture. This initial chapter provides a general background to contextualise the findings reported, including the importance of studying aquatic viruses, the common viruses affecting aquaculture, particularly of salmonid fishes as the major focus of my research, and the potential of cutting-edge genomics technologies to help control and mitigate viral disease outbreaks in fish farming. Finally, this chapter ends by outlining the specific objectives of my PhD project.

1.1 Aquatic viruses

Viruses are the most abundant biological agents in the aquatic environment with millions of viral particles being estimated to exist in each millilitre of sea water (Suttle, 2005). In fact some 94% of nucleic-acid containing particles in water are viruses, outnumbering all other lifeforms (Wen et al., 2004; Suttle, 2007). However many of these viruses (between 30 and 99%) are unknown to science and have no sequence homology to any characterised organism (Marhaver et al., 2008; McDaniel et al., 2008; Aggarwala et al., 2017). This proves problematic when attempting to identify novel viral species due to a lack of data to inform widely used tools (e.g. BLAST) built on sequence homology comparison. Even when sequences encode open reading frames (ORFs), this 'unknown fraction' often don't have protein homologs in existing databases. Such orphan genes have been found to be three-fold more common in viruses than for bacteria (Yin and Fischer, 2008).

Even within known viral lineages, extensive genetic diversity exists which can complicate their further characterization. New viral strains and genotypes are routinely discovered after a viral species has been identified and this genetic diversity has the potential to impact viral phenotypes and disease outcomes. For instance, in the case of influenza, the periodic transfer of genetic material between related viral strains often results in an increase in virulence (Tscherne and García-Sastre, 2011); such 'untapped' genetic diversity thus poses direct threats to human and wildlife health. Additionally, as research is increased in this field, a more complex network of disease dynamics is being discovered, with viral lineages traditionally considered to have restricted host ranges often found to infect a wider range of organisms (Shi et al., 2016b, 2018; Geoghegan et al., 2018). Such 'host jumping' has resulted in several pandemics in humans including HIV (Sharp and Hahn, 2010), influenza A (commonly known as the Spanish Flu) (Webby and Webster, 2001), SARS (Li et al., 2005), Hendra and Nipah

viruses (Chua, 2000), measles (Furuse et al., 2010), and recently SARS-CoV-2 (Zhou et al., 2020). Indeed with the increase in global movement, anthropogenic introduction of viruses to new environments and regions poses an unprecedented threat (McMichael, 2002; Wilson, 2005; Marano et al., 2007; Lindahl and Grace, 2015).

The enormity of the scale of unknown or poorly characterised aquatic microbes has attracted significant scientific research in recent years, particularly due to the threat that such organisms pose to humans (Cabelli et al., 1979; Yates et al., 1985; Sobsey et al., 1986; Rose et al., 1987; Lipp et al., 2002; Griffin et al., 2003; Melnick, 2015; Jennings et al., 2016), and the globally important aquaculture industry (reviewed in: Crane and Hyatt, 2011; Walker and Winton, 2010). However, even with this increased interest, the current state of knowledge of aquatic viral species, their genetic diversity, transmission routes, host reservoirs and geographic ranges is largely unknown, with notable exceptions of recent metagenomics studies on aquatic viruses (Shi et al., 2016b, 2018; Geoghegan et al., 2018). As illustrated in Figure 1.1, knowledge of many different aspects of viruses, their hosts and the environment in which infection is happening, is required to determine whether there is a risk to human or animal health.

1.2 Global aquaculture

Aquaculture is the world's fastest growing food-producing sector, with an average of 6% growth per year since 2000 (FAO, 2018). In 2014, for the first time, aquaculture produced more seafood for human consumption than capture fisheries (Figure 1.2), which is a remarkable statistic when considering that forty years ago, aquaculture contributed only ~7% of all fish consumed globally (FAO, 2018). Increasing production is a crucial part of the global effort to address the growing human population (FAO, 2018), which is demanding an ever increasing access to high quality animal protein that cannot be delivered from conventional agriculture without dramatically increasing land usage (Ramankutty et al., 2018). However aquaculture is not without its downsides, with significant environmental impacts linked to the expansion of the industry, including mangrove forest destruction for shrimp farms (Naylor et al., 2000; Harper et al., 2007), the salinization and acidification of soils in former farms (Alejandro Rodríguez-Valencia et al., 2010), and the eutrophication and nitrification of surrounding ecosystems (Burford and Williams, 2001; Jackson et al., 2003; Tacon and Forster, 2003; Focardi et al., 2005; Casillas-Hernández et al., 2007; Crab et al., 2007). Nonetheless, aquaculture is hugely economically important, contributing ~$249 billion to the global economy in 2017, with $151 billion from finfish production and $91.4 billion from shellfish production (FishStatJ). The vast majority (95% - $133b) of finfish production revolves around 46 species in 26 countries. Shellfish aquaculture is a less varied industry based largely on 21 species produced in 14 countries (95% - $87b) (Madsen and Dalsgaard, 1999).

1.2.1 Salmonid aquaculture

While salmonid species contributed just ~6.4% of the total finfish aquaculture production by weight in 2017 (3.4 billion of 53 billion tonnes), salmonid farming contributed ~14.6% of the total farmed finfish by value ($22 billion) (FishStatJ). In the same year, Atlantic salmon (*Salmo salar*) contributed $16.7 billion to the global economy and was the most economically valuable salmonid species. Its farming began in Norway in the 1960s and since has spread to several countries around the world including Chile, Scotland and Canada, which together with Norway produce >90% of the total global salmonid production by value (Figure *1.3*). Following Atlantic salmon, rainbow trout (*Oncorhynchus mykiss*) is the second most farmed salmonid species with $3.6 billion produced annually, and a 16% share of total salmonid production (FishStatJ). Other salmonids are also farmed extensively, especially in Canada and the USA, where other Pacific salmon species (chinook - *Oncorhynchus tshawytscha* and coho salmon - *Oncorhynchus kisutch*) are also major contributors to local economies. New Zealand has also recently begun to farm chinook salmon and currently produces around half of the global farmed production of this species (FishStatJ). While Chile is the second largest producer of Atlantic salmon, it also farms the vast majority of globally produced coho salmon with a 2017 value of $1.1 billion. Salmonid production also creates and sustains large numbers of jobs, especially in rural communities, for instance with over 2,000 in the UK alone being linked to Atlantic salmon aquaculture (FAO 2018).

1.3 Viral diseases and aquaculture - background

A major bottleneck to the expansion of a sustainable global aquaculture industry is infectious disease, which affects the industry by causing enormous financial loss (~$6 billion loss per annum; The World Bank, 2014), while also negatively impacting the welfare of farmed fish, along with the health of the environment (e.g. by treatments/interventions) and wild fish populations (e.g. by disease transfer) near fish farms, which together cause reputational damage to commercial fish farming organizations. Agents of disease include bacterial, viral, fungal and multicellular parasites, among which bacteria cause the greatest number of distinct characterized diseases (Kibenge et al. 2012). However, viruses - which cause 20% of all known infectious diseases in aquaculture - are more difficult to control, as few effective anti-viral therapeutics or preventative vaccines have been developed to date (McLoughlin and Graham 2007; Dhar, Manna, and Allnutt 2014).

One of the reasons that relatively few effective antiviral therapeutics have been developed can be attributed to the basic biology of viruses. Viruses accumulate mutations faster than either eukaryotes or bacteria (Holland et al., 1982), with RNA viruses having particularly rapid rates of evolution due to the lack of proof-reading activity within the virus replication machinery (Steinhauer and Holland, 1987; Holmes, 2009; Domingo-Calap and Sanjuán, 2011; Lauring

et al., 2013). This along with recombination, which is relatively common in viruses (Carr et al., 1998; Walling et al., 1999; Uzcategui et al., 2001; Sugauchi et al., 2003; Moya et al., 2004; Su et al., 2016), enables a biological 'arms race' between the virus and host, with potential for novel pathogenic strains to arise rapidly, with little warning. Therefore many targeted therapeutics, including vaccines, can quickly become ineffective when faced with these new and divergent viruses. There is also evidence that by constantly attempting to eradicate pathogens in aquaculture, we might be driving the evolution of increased pathogenic and virulent strains (Kennedy et al., 2016), which in turn can lead to future complications in disease control.

The prevalence of infectious diseases associated with intensive fish farming is likely to further increase as more farms are created and greater stocking densities are achieved. Combined with new areas of the world adopting aquaculture as a useful source of food and economic security, this creates further potential for the spread of existing, and the emergence of new, aquatic diseases (Tables 1.1 & 1.2). In the following section, I will give a brief overview of economically damaging viral diseases in some of the major aquaculture sectors, affecting cyprinid, shellfish and salmonid production, the last being the major focus of this Thesis.

1.3.1 Examples of major viral diseases in aquaculture

Carp farming is one of the largest finfish aquaculture sectors in the world, worth $42 billion to the global economy each year. However there are several viral diseases that negatively affect this industry including Spring Viremia of Carp Virus (SVCV) and Cyprinid Herpesvirus-3 (CyHV-3). SVCV is a disease endemic in Europe that causes significant losses in cyprinids, particularly carp species (Ahne et al., 2002), with up to 15% of fry being lost to SVCV each year. SVCV has been detected throughout Europe including Russia (Oreshkova et al., 1999), in America (Miller et al., 2007; Warg et al., 2007), and parts of Asia (Teng et al., 2007; Zhang et al., 2009). As a highly pathogenic virus currently listed on the OIE list of notifiable diseases, SVCV poses a particular threat to Asian carp aquaculture where it has yet to take hold. However as the centre for global carp production, this industry would be specifically threatened by the introduction of such a viral pathogen.

CyHV-3 causes significant morbidity and mortalities in common and koi carp, both highly valuable species. This virus has decimated carp populations in many parts of the world including the Middle East, North America, Europe and East Asia (Rodgers et al., 2011; Baumer et al., 2013; Ito et al., 2014) and control efforts have been hampered by the tendency to stock carp in high densities (Gotesman et al., 2013). CyHV-3 has been an OIE notifiable disease since 2007, and while there are vaccines available to control disease, immunised fish may still carry the virus and pose a threat to naïve fish stocks (Bergmann and Kempter, 2011).

Viral disease also seriously impacts shellfish farming, with ~60% of global losses attributed to viruses, at an estimated annual cost of ~$1 billion (Flegel et al., 2008). An example with devastating economic impacts was a 1995 outbreak of yellow-head virus (YHV) in Thailand (the world's largest shrimp exporter) that decreased production of shrimp by 5,000 tonnes at a cost of ~$40 million. In the immediate subsequent years, white-spot syndrome virus (WSSV) caused even greater losses, with an estimated cumulative loss in export revenue of $1 billion over 3 years (Flegel, 2006). In China, outbreaks of WSSV began in 1993, shrinking shrimp production from >250,000 to 80,000 tonnes, with subsequent recovery of the industry being slow (Flegel, 2006).

Ostreid herpesvirus 1 (OsHV-1) is another serious viral shellfish pathogen, specifically affecting the Pacific oyster (*Crassostrea gigas*), with mass mortalities common (Hine et al., 1992; Renault and Arzul, 2001; Friedman et al., 2005; Moss et al., 2007; Garcia et al., 2011). The disease caused by OsHV-1 is temperature dependent, with most mortalities occurring above 16ºC, mainly affecting juveniles (EFSA, 2010). OsHV-1 has been found worldwide with cases in North America, Europe, East Asia and Oceania (Friedman et al., 2005; Segarra et al., 2010; Shimahara et al., 2012; Jenkins et al., 2013; Mello et al., 2018). Currently there is no cure or effective vaccine, with the best mitigation strategy being to minimise the movement of stock, although this is difficult due to the coastal nature of farming sites.

Several viral pathogens are a significant threat to the salmonid aquaculture sector (summarised in Table 1.2). One of the most important viruses in global salmonid aquaculture is infectious salmon anaemia virus (ISAV). In 2007, a major ISAV outbreak cost the Chilean salmon industry more than $1 billion, 15,000 jobs and reduced Atlantic salmon production from 400,000 to 100,000 tonnes in just two years (Asche et al., 2009). This outbreak arose from a Norwegian ISAV strain being imported and infecting farmed coho salmon in the late 1990s (Kibenge et al., 2009; Vike et al., 2009; Cottet et al., 2010; Castro-Nallar et al., 2011). This virus subsequently became established, before rapidly evolving increased virulence and pathogenicity to Atlantic salmon, causing the 2007 outbreak and subsequent crash in the sector (Kibenge et al., 2009).

There are also examples of salmonid viral diseases that have become endemic in a region, causing continual losses to the sector. Salmonid alphavirus (SAV), the causative agent of pancreas disease (PD), is a major hindrance to Atlantic salmon aquaculture in Europe. It has been estimated that the direct costs of a PD outbreak to a site of 500,000 smolts is ~£1.3 million (Aunsmo et al., 2010). SAV is considered to be endemic in many regions of Europe (Aunsmo et al., 2010), including Norway where 137 salmon farm sites were confirmed to have been infected with SAV in 2020 as of the 13th March 2020 (www.BarentsWatch.no). There

are six recognised SAV subtypes, thought to have been introduced to aquaculture independently from wild reservoirs (Karlsen et al., 2014), each with varying pathogenicities and clinical impacts (Graham et al., 2011). The extent of these wild reservoirs are as yet unknown, though flatfish (e.g. plaice and dab) and Ballan wrasse (*Labrus bergylta*) all carry the virus without signs of disease (Snow et al., 2010; Bruno et al., 2014; Ruane et al., 2018).

Viral haemorrhagic septicaemia virus (VHSV) is one of the most important viral pathogens in salmonid farming, particularly for rainbow trout. VHSV has been isolated from at least 48 species of fish from the Northern Hemisphere (Skall et al., 2005) and causes significant economic consequences. Split into several genogroups, North American and European strains are genetically distinct, though serological analyses have been unable to clearly distinguish them (Skall et al., 2005). European VHSV is widely distributed throughout continental Europe with intensive rainbow trout production, while coastal areas of Norway, UK, continental Finland, Ireland, Germany, France, Denmark and Spain are recognised as VHSV-free (Skall et al., 2005). However with the ever-growing list of susceptible marine species to VHSV, the disease-free status of coastal regions is under review due to the potential for reintroductions of the virus from wild host reservoirs (OIE, 2019a).

Additionally, there are several emerging viral diseases affecting salmonid aquaculture, including piscine myocarditis virus (PMCV) and piscine reovirus (PRV) (Table 1.1). Although the diseases that they cause have been known for some time (1985 in the case of PMCV - Amin and Trasti, 1988), the causative viral agents have only recently been characterised (Løvoll et al., 2010; Palacios et al., 2010; Haugland et al., 2011). In 2002, PMCV, causing cardiomyopathy syndrome, cost the Norwegian aquaculture industry an estimated €4.5 - €8.8 million in losses (Brun et al., 2003), with this number increasing to an estimated €25 million in Norway by 2007 (Garseth et al., 2018). While no official figure on the economic losses from PRV infection has been released, it is widely accepted that this virus, and the disease it causes, lead to significant economic losses (Morton et al., 2017; Madhun et al., 2018).

1.4 Disease diagnostics and pathogen characterisation in salmonid aquaculture

As one of the most value aquaculture sectors in Europe, control of infectious diseases is key to salmonid production. The standard disease characterisation pipeline involves several steps, employing both clinical and molecular methods to confirm the presence of the pathogen and a relevant disease phenotype (OIE, 2017a). Initially clinical symptoms of illness must be demonstrated, including post-mortem indications of infected tissues (e.g. pancreatic tissue loss, skeletal muscle degeneration and cardiomyocytic necrosis/inflammation).

Immunohistochemical testing for the presence of relevant antibodies is recommended for tissues showing acute necrosis, along with infectious agent detection. Pathogen detection often involves the use of cell culture to isolate viral particles, though this is laborious and time consuming for many viruses, and challenging to perform for high throughput studies (Graham et al., 2007; Petterson et al., 2013; Arseneau et al., 2019), followed by PCR and qPCR to detect specific viral nucleic acids (e.g. Fringuelli et al., 2008; Hodneland and Endresen, 2006). Depending on the virus, different host tissues are recommended to maximise the viral load in a sample (OIE, 2017a, 2017b). Additionally, serological analysis can be used to detect neutralising antibodies up to two weeks post-infection, and may help identify animals which have been exposed and subsequently recovered from the infection during a disease outbreak.

While many such diagnostic tools have been optimised for the rapid detection of pathogens and identification of associated disease outcomes, routine pathogen strain characterisation in salmonid aquaculture classically relies on the PCR amplification of marker genes, often ~500bp in length (e.g. Fringuelli et al., 2008; Hodneland and Endresen, 2006) with subsequent Sanger sequencing. While this approach has proven very effective at confirming the presence of nucleic acids from specific viruses (Hodneland and Endresen, 2006; Snow et al., 2006; Fringuelli et al., 2008; OIE, 2017b; Lewisch et al., 2018), determining genome structure and sequences (McLoughlin and Graham, 2007; Fringuelli et al., 2008; Kulshreshtha et al., 2010; Rimstad et al., 2011), and revealing the dominant strains in samples (Cunningham et al., 2002; Mjaaland et al., 2002; Nylund et al., 2007; Kibenge et al., 2007, 2009; Markussen et al., 2008; McBeath et al., 2009; Lyngstad et al., 2012; Cárdenas et al., 2014; Christiansen et al., 2017; Gagné and LeBlanc, 2018), it is generally low-throughput and not well suited to characterise the genetic diversity in virus populations (see Section 1.5). Additionally, by not sequencing whole viral genomes, large amounts of potentially informative genomic information is lost, including potential virulence markers.

1.5 Next generation sequencing of RNA viruses: tools and approaches

As mentioned above for salmonids, the most common approach to characterize the identity and origin of viral diseases impacting aquatic organisms involves the analysis of a limited number of marker genes via PCR followed by Sanger sequencing. Such approaches have proven extremely useful to determine the broad viral strain, but provide a partial representation of the genome, restricting comprehensive investigations into the link between sequence variation and pathogenicity, or other aspects of population dynamics relevant to epidemiology (Gontcharov et al., 2004; Pearson et al., 2004; Castresana, 2007; Martens et al., 2008). Recombination and reassortment are additional concerns (Lai, 1992; Vijaykrishna et al., 2015b), as viral genomes may represent a chimera of multiple evolutionary histories, and this will often be missed when characterizing only short gene fragments. Thus, whole genomes

are preferable when performing sequence-based analyses of viral pathogens, as this provides the maximal possible phylogenetic signal to reconstruct evolutionary history and events.

Whole genome sequencing (WGS) of viruses has been used extensively for human pathogens such as Influenza or HIV with whole genome data from over 49,000 strains of Influenza A being deposited in the Influenza Research Database (www.fludb.org). There are many methods of obtaining whole genome sequences for viral strains, ranging from highly targeted and low-throughput Sanger-sequencing of PCR amplicons, to unbiased shotgun metagenomics, each with advantages and disadvantages (Figure 1.4). The applicability of any given approach depends on the research question and the viral species studied; following sections will outline the main approaches for viral WGS and consider under what scenarios different approaches are most appropriate (Figure 1.4)

1.5.1 PCR based techniques using specific primers

PCR-based enrichment is perhaps the most common method for targeted analysis of viral genes. A well-designed PCR assay allows for efficient sequencing, as the vast majority of data should be for the target viral gene or genomic region. This is important especially when using Sanger sequencing, which does not produce large quantities of data (Morozova and Marra, 2008). When using Next Generation Sequencing (NGS) methods, PCR amplification allows for ultra-deep coverage across the genome, enabling low-frequency strains to be detected in viral populations (e.g. Margeridon-Thermet et al., 2009; Markussen et al., 2013; Schönherz et al., 2016; Wang et al., 2007). This high-throughput approach also allows for massive multiplexing in a single sequencing run, which reduces the cost of per sample WGS sequencing (Morozova and Marra, 2008).

However to generate PCR amplicons, reference viral genome sequences are required to design primers to bind to DNA or cDNA templates. Knowledge of common genetic variations or SNPs is also desirable to avoid mismatches in primer binding regions, which may otherwise reduce the efficiency, or prevent the success of the PCR (Stadhouders et al., 2010). Due to the rapid evolution of RNA viruses, PCR primers must be routinely re-examined to ensure newly evolved strains are captured by existing primer sets. Additionally, the potential to discover new viral species, or highly divergent strains of known species, is limited as no universal viral gene exists akin to the bacterial 16S gene (Woese and Fox, 1977; Weisburg et al., 1991; Coenye and Vandamme, 2003) or CO1 gene in animals (Lobo et al., 2013; Pentinsaari et al., 2016). A major bottleneck to PCR-based sequencing is the integrity of the nucleic acids in the original sample. DNA-based viruses tend to survive archiving and sub-optimal storage better than RNA viruses due to the greater stability of DNA compared to RNA (Lesnik and Freier, 1995). Additionally, PCR reactions are limited by the maximum size of amplified fragments

possible, which is determined not only by the polymerase enzyme used, but also the integrity of the DNA template (Jia et al., 2014). Therefore, PCR amplification requires the DNA or cDNA template to be of a certain length, and dependent on the primer sets, while degradation of the sample can inhibit successful amplification, giving rise to false-negative tests.

Overall, PCR amplification of viral genomes has been highly successful in its ability to generate large quantities of DNA of viral origin to be sequenced on a variety of platforms, at cost-effective prices. However there are several limitations to this approach which need to be considered, especially in an under-researched area such as aquaculture viral genomics, where the genetic variability is not yet fully described.

1.5.2 Targeted Sequence Capture

Targeted sequence capture involves using RNA or oligonucleotide baits designed against a set of target loci or genomic regions, hybridising the baits to a DNA or cDNA sample, and removing unbound (i.e. off-target) DNA. These baits can target genomic regions that range from hundreds of base pairs to millions of base pair in length, by binding to fragmented DNA matching that region to the exclusion of the rest of the genome. This library preparation strategy allows for samples to be highly enriched for specific DNA strands covering a much greater genomic area than possible with PCR amplification. Consequently, this approach produces a highly enriched library of target genomic regions which enables higher sequencing efficiency. This approach has been used widely for selectively sequencing specific genomic regions of host DNA including exome capture (Choi et al., 2009; Ng et al., 2009; Fisher et al., 2011; Lonigro et al., 2011; Parla et al., 2011; Futema et al., 2012), capturing panels of cancer-related loci (Hagemann et al., 2013; Pritchard et al., 2014; Sakr et al., 2016), and ecological and evolutionary comparative genomics (Grover et al., 2012; Smith et al., 2014; Jones and Good, 2016; Lappin et al., 2016). In addition to sequencing specific regions of the host genome, capture approaches have also been used to enrich DNA samples for pathogens including viruses (Chandler et al., 1993; Miller et al., 1995; Shyamala et al., 2004; Depledge et al., 2011; Hammoumi et al., 2016), bacteria (Jacobsen, 1995; Noble and Weisberg, 2005; Koo et al., 2009), and eukaryote parasites such as *Plasmodium* spp. (Chen et al., 1991; Bright et al., 2012). One of the benefits of using capture sequencing when characterising pathogen populations is the relative tolerance of the RNA baits to template mismatches compared to PCR primers. Baits can hybridise to target sequences with as much as 15% pairwise divergence for some capture platforms e.g. Agilent SureSelect (Lappin et al., 2016), which may allow the simultaneous enrichment and sequencing of distinct pathogen strains and/or subtypes, even when the baits are not designed for all the strains present.

However, while whole virome bait designs do exist (Briese et al., 2015; Wylie et al., 2015), such approaches are still dependent on the quality and completeness of genomic databases of the organisms of interest. The identification of novel species, especially viruses, is challenging due to a low degree of sequence homology between distantly related groups (Belshaw et al., 2009), which may exceed the maximum limits of hybridization-based capture platforms. Additionally, the efficiency of the hybridisation and capture process crucially depends on the initial quantity of the target material (Depledge et al., 2011; Hammoumi et al., 2016). High viral titres result in a higher proportion of viral reads sequenced, which can limit the benefits of using capture approaches for very low-titre infections over traditional shotgun sequencing (see Section 1.5.3).

Finally, the use of baits to enrich for target genomic regions significantly reduces the cost of sequencing per sample due to the potential to multiplex several samples in a single sequencing run, without wasting data sequencing off-target loci. However, with the ever reducing cost of NGS, the expense of designing, producing and using such a panel of baits needs to be compared with shotgun sequencing entire samples and subsequently filtering genomic regions of interest, including pathogens (Shi et al., 2016b, 2018; Geoghegan et al., 2018). Overall, this approach has been shown to enable ultra-deep sequencing of specific pathogens with enough flexibility to get an accurate representation of genetic diversity within a sample.

1.5.3 Shotgun sequencing – viral metagenomics and metatranscriptomics

A standard approach for generating genomic and transcriptomic sequences is to shotgun sequence DNA or cDNA. Enrichment methods are typically applied to RNA sequencing (hereafter referred to as RNA-seq) including ribosomal depletion or mRNA capture using oligo-dT primers and oligo-dT coated magnetic beads (Hrdlickova et al., 2017). This reduces the amount of ribosomal RNA (rRNA) that is sequenced, which if left untreated is likely to overwhelm any mRNA sequences due to the relative abundance of rRNA compared to mRNA (O'Neil et al., 2013; Zhao et al., 2018). In studies of many organisms, RNA-seq approaches are also useful for sequencing commensal and pathogenic microbes that are transcriptionally active and produce RNA alongside the host, while microbes with DNA-based genomes may be captured during host genome sequencing. Such shotgun sequencing approaches are extremely valuable for identifying and characterising the genomes of new microorganisms either in healthy individuals (i.e. commensal microbiomes) or in instances of undiagnosed diseases (i.e. potentially novel pathogens) (Palacios et al., 2010; Macklaim et al., 2013; Zhang et al., 2015; Bacharach et al., 2016; Petersen et al., 2017).

Various high-throughput methods can be used for shotgun sequencing that have relative merits and limitations (see Section 1.5.4 and Figure 1.4). Many shotgun sequencing studies use short-

read platforms to take advantage of low sequencing error rates and extremely high data outputs (e.g. the latest Illumina NovaSeq can produce 3.2 Tb of paired-end sequence reads in a run). However with the continual improvement in read quality and data quantity of long read sequencing platforms (i.e. PacBio and Oxford Nanopore Technologies platforms), the use of these platforms for metagenomics has increased (Greninger et al., 2015; Frank et al., 2016; Brown et al., 2017; Driscoll et al., 2017; Schmidt et al., 2017).

Regardless of the platform used, shotgun sequencing data should be largely unbiased, with most microbiome/metagenome samples containing reads from multiple organisms, usually overwhelmingly the host, but also diverse microbes (Pereira-Marques et al., 2019). And while this makes shotgun sequencing inefficient to characterise a specific infecting pathogen, this approach allows for novel taxonomic groups to be identified without prior knowledge of their presence or even existence. Importantly, both short and long-read metagenomics studies require significant computational resources and expertise (Wooley and Ye, 2010; Howe and Chain, 2015; Quince et al., 2017; Breitwieser et al., 2018). It also generally costs more financially to sequence so much off-target DNA/RNA (up to 99% in some microbiome samples).

1.5.4 Comparison of short- and long-read technologies

In recent years, the gold-standard for high-throughput sequencing are Illumina's short-read sequencing platforms due to the low per-base error rate and high per-run outputs. However, over the past decade third-generation long-read sequencing technologies such as Oxford Nanopore Technologies and PacBio have proven to be disruptive to the NGS industry.

The primary difference between second and third generation NGS platforms is the maximum length of sequenced reads. Illumina platforms have a maximum read length of 300bp (MiSeq v3), although the insert size between paired-end reads can be much larger depending on the library preparation method. While large insert sizes can be useful for scaffolding contigs in genome assemblies, this nonetheless leaves gap regions where the base information is missing. Long read platforms on the other hand can produce reads of thousands through to hundreds of thousands of base pairs in length, with the longest contiguous strand of DNA sequenced to date being over two million base pairs long (Payne et al., 2019). This ability to produce ultra-long reads enables far more contiguous genome assembly than short-read technologies as individual reads can span repeat regions that are difficult to assemble with short reads alone (Guo et al., 2018; Bongartz, 2019; Sone et al., 2019). Long reads can also help identify and phase closely related pathogen strains that only differ by a few number of variants that may exist further apart than paired-end reads can link (Sumpter et al., 2018; Hamlin et al., 2019; Amarasinghe et al., 2020). However specialist short-read libraries can be generated that

capture ultra-long genomic information in the form of paired-end reads with massive inserts including linked reads (Ott et al., 2018) and Hi-C technologies (Kempfer and Pombo, 2019).

The second major difference between these sequencing approaches is the per-base quality produced. Illumina platforms generally have a very low per-base error rate ($0.24 \pm 0.06\%$ per base (Pfeiffer et al., 2018) which allows for highly accurate genome assemblies, and reliable within-sample microbial population analyses. Long read technologies typically have a much higher error rate with Nanopore's error rate ranging from 5-15% dependent on the type of molecule and library preparations (Rang et al., 2018), though improvements in base-calling algorithms can reduce this to 2-5% (Kono and Arakawa, 2019). PacBio sequencing does offer high accuracy per read sequencing due to their circular consensus sequence (CCS) method which involves circularising individual DNA strands and then sequencing the same strand multiple times to enable error corrections (Rhoads and Au, 2015). While the error rate for raw PacBio data is similar to Nanopore's at 13-15%, the CCS error rate can be as low as 1.7% (Buck et al., 2017), and the recent development of high-fidelity (HiFi) PacBio reads can generate long reads with an accuracy of 99.8% (Wenger et al., 2019).

A third major difference is the potential for using NGS in resource-limited or field conditions. All the short-read Illumina sequencers are bench-top appliances or larger, and require specialist training to both run and upkeep. Similarly the PacBio sequel machines and Oxford Nanopore's GridION and PromethION machines are large and cannot be taken outside of a laboratory environment. However Oxford Nanopore's MinION system is portable and powered via a USB cable to a PC or laptop. This enables in-field sequencing in rapid response scenarios (i.e. field hospitals, remote expeditions, disease epidemics requiring real-time results) (Edwards et al., 2016, 2019; Faria et al., 2016; Hoenen et al., 2016; Euskirchen et al., 2017; Johnson et al., 2017). However there is a trade-off between portability and data output potential with the MinION having a theoretical max output of 50 Gb (though outputs are usually much lower), and the PromethION having a theoretical max output per flow cell of over 200 Gb, and Illumina's NovaSeq machine having a max output of over 2 Tb per flow cell.

Each sequencing platform has its own merits and disadvantages, and these must be considered when undertaking sequencing as the technology chosen can significantly impact the likelihood of success depending on the research question.

1.6 Viral molecular epidemiology and 'phylodynamics'

As described above, due to their short generation times and lack of proof-reading enzymes, RNA viruses rapidly accumulate genetic variation. Interpreting patterns of this genetic variation in a structured analysis framework has developed over the past decade and a half

into the field of phylodynamics (Grenfell et al., 2004). The primary goal of phylodynamic studies is to make inferences on disease epidemiology from phylogenetics by combining immunology, epidemiology and evolutionary biology (Grenfell et al., 2004) and inferences on the phylodynamics of human viruses have been extensively made. The topology of these phylogenetic trees can inform on the virus population dynamics (outlined in Fig. E-G; adapted from Grenfell et al., 2004). The relative length of branches will be affected by virus population growth patterns with rapid population expansions (i.e. an epidemic) having relatively long external branches compared to internal branches (Fig. 1.5.A), and stable virus populations (i.e. endemic viruses) showing topologies with shorter external branches than internal (Fig. 1.5.b). Host population structure is also evident in phylogenetic topologies as viruses with similar hosts are expected to be more closely genetically related as transmissions are thought to be more common within the host species than between different species (Figs. 1.5.C & 1.5.D). Finally the effects of selection pressures will alter the tree topology (Fig. 5). Ladder-like phylogenies exhibit traits of strong directional selection with surviving lineages driven by immune escape (Fig. 1.5.F), commonly found in viruses like influenza A-H3N2's hemagglutinin gene. In contrast, virus populations not under strong directional selection have a more balanced topology (Fig. 1.5.E). Some viruses like HIV can show both topologies at different scales, with trees derived from sequences of different patients resembling a balanced tree like Fig. 1.5.E, while phylogenies of viruses in chronically-infected hosts resemble a ladder-like tree as in Fig. 1.5.F. These inferences can then be used to understand certain viral phenotypes including virulence, viral transmissibility, and antigenic escape from host immunity (Volz et al., 2013).

In this section, I describe methods commonly used in phylodynamic studies and their application to understand and control infectious diseases. I also summarise several examples of phylodynamic analyses in both aquaculture and public health settings.

1.6.1 Commonly employed phylodynamic methods

The first step in most phylodynamic analyses is to generate a phylogenetic tree of virus sequences originating from an outbreak or outbreaks. This can be achieved with multiple methods, but most use either a Maximum Likelihood or Bayesian approach, with the latter being particularly popular due to their ability to fit complex demographic models while handling phylogenetic uncertainty (Drummond et al., 2005; Kühnert et al., 2011). Using sequences that have been sampled at different time points during the epidemic or outbreak, phylogenetic trees can be calibrated to absolute time. This enables rates of substitution to be estimated, which in turn can inform estimates of the most recent common ancestor (MRCA) of the viral strains sampled using molecular clock models (Drummond et al., 2002).

As a consequence of the accumulation of genetic variation throughout the time-scale of an epidemic, viruses can exhibit genetic differentiation based on geographic location, providing a fundamental understanding of evolutionary dynamics (Holmes, 2004; Rambaut et al., 2008; Russell et al., 2008; Lemey et al., 2009). Such phylogeography analyses compare genetic relatedness of virus sequences to geographic proximity to inform the geographic population structure. The presence of population structure is determined by the clustering of sequences from similar geographic locations beyond what is expected under a non-structured model (Chen and Holmes, 2009). The location and movement of ancestral viral lineages between geographic locations can also be reconstructed by phylogeographic analyses (Volz et al., 2013). Viruses in some regions may mix more readily than in other regions, and this asymmetry can be visible in phylogenetic topology (Figure 1.5). Additionally, some connections between geographic regions may be unbalanced, with more movement of viruses from 'region A' to 'region B' than vice versa, in a source-sink manner (Pulliam, 1988).

1.6.2 Applications of viral phylodynamic analyses in public health

1.6.2.1 Viral origins

As mentioned above, phylodynamic models can help estimate the MRCA of viral outbreaks or specific lineages within outbreaks. The age of a MRCA is always a lower-bound estimate as the actual ancestral virus of a lineage must have existed earlier than the MRCA of the sample used (Volz et al., 2013). For example in one study by Arias et al. (2016), 554 Ebola virus genomes were generated from patients in Sierra Leone over a period of 10 months, representing 23.8% of all cases in Sierra Leone in 2015. Multiple overlapping PCR amplicons, designed to cover the entire viral genome, were generated for Ebola virus before being sequenced on the Ion Torrent PGM System. Such dense sampling and sequencing of whole viral genomes enabled the authors to identify at least nine distinct lineages circulating in Sierra Leone during this time period with estimates of transmission chains and timings of strain divergence. Timing estimates of an individual's infection time with HIV has also used similar approaches (Lemey et al., 2006). By sequentially sampling HIV from a chronically infected patient, time-calibrated analyses can be performed and using root-to-tip analysis tools can estimate the time since seroconversion (used analogously to time of infection).

The recent 2019-2020 Wuhan coronavirus outbreak is an ongoing pandemic of SARS-CoV-2, a virus related to the SARS-CoV virus, responsible for a major public health pandemic in 2003. By the 28th July, 73,374 SARS-CoV-2 genomes have been uploaded to the GISAID platform (https://www.gisaid.org/) with phylogenetic analysis showing the spread of the virus both within China and between China and the rest of the world. The rapid sequencing of whole genomes from this epidemic has been another useful trial of the impact that viral phylodynamics and molecular epidemiology can have on understanding disease dynamics in

real time. The data generated thus far has been used to estimate the substitution rate of the virus, and time-calibrated phylogenetic trees. These analyses have then been used to estimate the connections between infected patients in distant geographic regions.

1.6.2.2 Viral spread and phylogeography

Along with time-calibrated phylogenetic trees, phylodynamic models can also shed light on the spread of viral lineages, both across the world and within chronic patients. In 2015-16, a study used real-time sequencing on the MinION platform to generate 142 Ebola virus genomes from Guinea (Quick et al., 2016). The resulting data was used to characterise the viral strains circulating in Guinea and compare them to those in Sierra Leone. This identified several closely related strains between the two countries, indicating the transport and transmission of Ebola between the two countries.

During the 2015-2016 Zika virus epidemic, hundreds of thousands of cases were reported across the Americas. Phylogenetic analysis of 61 Zika virus genome sequences, sampled from Central America and Mexico by Thézé et al. (2018) indicated that almost all the viral strains fell into a single monophyletic group, closely related to Zika strains in Brazil. Phylogeographic analysis was performed and although multiple introductions of Zika virus into Central America were identified, most of the infections were a result of a single introduction from Brazil to Honduras in mid-2014 and circulated in Central America for at least a year before being first detected.

These approaches can also be used to estimate the spread of viruses not globally, but within chronically infected patients. Sampling hepatitis C virus over the course of a chronic infection has shown that the time point and sample type (e.g. serum, liver biopsies, other cell types) impacts the accuracy of genetic diversity estimates (Gray et al., 2012). This disagreement in viral strains identified indicates that different strains are found in different areas of the body at certain times, and that the shed of hepatitis C virus from the liver (where the virus persists) to the circulatory and nervous system may occur at different rates.

1.6.2.3 Viral control efforts

Phylodynamics can also be used to determine the efficacy of viral control programmes. For example, a study on HIV patients before and after antiviral treatment showed that viral substitution rates, used as a proxy for genetic diversity, dropped considerably after treatment (Drummond et al., 2001). This decrease in genetic diversity was interpreted as effectively blocking viral replication, which is a known factor in the progression of HIV-infection to AIDS (Lemey et al., 2007). A similar approach was taken to assess the genetic diversity of hepatitis B virus in the Netherlands after the implementation of a vaccination programme (Van

Ballegooijen et al., 2009). The genetic diversity declined after widespread vaccination and this was used as evidence for the reduction of hepatitis B prevalence.

Determining the selection pressures on viral populations after antiviral treatments is important to detect and prevent the evolution of drug resistance. Using phylodynamic models to assess the level of selection pressures required to shift the viral population from susceptible to resistant is thus of critical public health importance (Bloom et al., 2010). Resistance of influenza A (H1N1) to the drug oseltamivir has been modelled in this manner, and found that resistant strains had a fitness advantage even in untreated hosts (Chao et al., 2012).

1.6.3 Examples of phylodynamic analyses in aquaculture

Molecular epidemiological approaches have been applied to many viruses important to aquaculture around the world (see reviews Bayliss et al., 2017; Snow, 2011). Applications of such studies include tracking viral transmission of WSSV on shrimp farms in Vietnam using PCR-amplified repeat sequences in the WSSV genome (Hoa et al., 2011). Two different farming methods were analysed; semi-intensive and improved extensive and it was found that in the semi-intensive farms viral transmission was predominantly from neighbouring ponds, while in improved extensive farms viral transmission was mainly due to the recycling of the virus over time in the same pond. These findings suggested a nuanced control strategy that could be tailored to the type of farm involved.

Phylodynamic analyses have also been applied to larger-scale geographic movements of viruses in aquaculture (Godoy et al., 2013). As discussed in Section 1.3.1, ISAV caused a major outbreak in Chile in 2007-08 and is thought to have been imported from Europe. The viral strain involved was the highly pathogenic ISAV-HPR7b, which was eventually replaced by a low pathogenic ISAV-HPR0 strain as the dominant one in Chilean aquaculture. The origins of another, smaller, ISAV outbreak in Chile in 2013 was investigated using sequences of the segment 6 Hemagglutinin gene. Sequences from the 2013 outbreak consisted of genogroups ISAV-HPR3 and HPR14, but clustered with the low pathogenic HPR0, indicating that this new outbreak was the result of a mutation of HPR0 to a high-pathogenic HPRΔ strain.

Additionally, the evolutionary landscape and geographic distribution of SAV in Europe was characterised using phylodynamic analyses of partial genomic sequences of SAV strains (Karlsen et al., 2014). This study estimated the evolutionary rate of SAV, along with the MRCA of five of the SAV subtypes and related these estimates to known historical events in European salmonid aquaculture. This enabled the authors to conclude that there has been multiple independent introductions of SAV from wild reservoirs into aquaculture, and become established epizootics.

1.7 Project Objectives

The overall aim of my PhD project was to develop approaches to accurately characterise whole genomes of viral pathogens in farmed aquatic animals, using salmonid alphavirus (SAV) as a model pathogen, alongside a range of lab techniques, sample types, sequencing technologies and analysis tools to assess the efficiency of viral characterisation at different genomic scales.

The specific objects to achieve this aim were as follows:

1. **Develop and validate a useful long-read sequencing approach to characterise the genomes of known salmonid viral infections**. The development and benchmarking of this method against classic Sanger sequencing methods is reported in Chapter 2. Furthermore, as a validation of this method's usefulness in molecular epidemiology, I applied it to sequence 24 SAV3 genomes, which were used to characterize the PD epidemic in Norwegian aquaculture. The findings are reported in Chapter 3.

2. **Develop an ultra-deep, short-read sequencing approach using targeted sequence capture to characterise the genetic diversity of the viral population in a known infection.** To validate the usefulness of this method I applied it to sequence the complete SAV population from eighteen naturally infected fish hearts, including both wild and farmed fish from across Scottish and Irish waters. The findings are reported in Chapter 4.

3. **Develop a useful analysis pipeline to detect and characterise both known and novel viral infections in host shotgun sequencing samples.** I validated this metagenomic pipeline against mock infections (both single-virus and virome), real infections of known virus species, and previously characterised viromes. This pipeline was then used to characterise the virome of Pacific oysters challenged with norovirus, sampled throughout the infection time course. The findings are reported in Chapter 5.

Table 1.1. Economically important viral pathogens in aquaculture

Virus	Abbreviation	Genome	Taxonomic classification	Virus identified and characterised	OIE listed (2020)	Host
Cyprinid herpesvirus	CyHV-3	dsDNA	Alloherpesviridae	1998 (Hedrick et al., 2000)	Yes	Carp
Spring viraemia of carp virus	SVCV	(−)ssRNA	Rhabdoviridae	1999 (Fijan 1999)	Yes	Carp
Ostreid herpesvirus 1	OsHV-1	dsDNA	Malacoherpesviridae	1970s (Farley et al., 1972)	No	Pacific Oyster
Red sea bream iridovirus	RSIV	dsDNA	Iridoviridae	1990 (Inouye et al., 1992)	Yes	Red sea breams
Mourilyan virus	MoV	(−)ssRNA	Bunyaviridae	1996 (Cowley et al., 2005)	No	Shrimp (penaeid)
Taura syndrome virus	TSV	(+)ssRNA	Dicistroviridae	1994 (Hasson et al., 1995)	Yes	Shrimp (penaeid)
Laem-Singh virus	LSNV	(+)dsRNA	Luteoviridae	2006 (Sritunyalucksana et al., 2006)	No	Shrimp (penaeid)
White spot syndrome virus	WSSV	dsDNA	Nimaviridae	2001 (Van Hulten et al., 2001)	Yes	Shrimp (penaeid)
Macrobrachium rosenbergii nodavirus	MrNV	(+)ssRNA	Nodaviridae	1999 (Arcier et al. 1999)	Yes	Shrimp (penaeid)
Penaeus vannamei nodavirus	PvNV	(+)ssRNA	Nodaviridae	2004 (Tang et al., 2007)	No	Shrimp (penaeid)
Infectious hypodermal and haematopoietic necrosis virus	IHHNV	ssDNA	Parvoviridae	1984 (Lightner and Redman, 1985)	Yes	Shrimp (penaeid)
Hepatopancreatic parvovirus	HPV	ssDNA	Parvoviridae	1985 (Chong and Loh, 1984; Lightner and Redman, 1985)	No	Shrimp (penaeid)
Infectious myonecrosis virus	IMNV	dsRNA	Totiviridae	2006 (Poulos et al., 2006)	Yes	Shrimp (penaeid)
Yellow head virus	YHV	(+)ssRNA	Roniviridae	1999 (Tang and Lightner, 1999)	Yes	Shrimp (penaeid)
Tilapia lake virus	TiLV	(−)ssRNA	Orthomyxoviridae	2014 (Eyngor et al., 2014)	No	Tilapia
Viral haemorrhagic septicaemia virus	VHSV	(−)ssRNA	Rhabdoviridae	1962 (Jensen, 1965)	Yes	Various fish (marine and freshwater)
Viral nervous necrosis virus	VNNV	(+)ssRNA	Nodaviridae	1990 (Glazebrook et al., 1990; Mori et al., 1991)	No	Various fish (marine and freshwater)
Epizootic haematopoietic necrosis virus	EHNV	dsDNA	Iridoviridae	1985 (Langdon and Humphrey, 1987)	Yes	Various marine fish
Hirame rhabdovirus	HIRRV	(-)ssRNA	Rhabdoviridae	1984 (Kimura et al., 1985)	No	Various marine fish

Table 1.2. Economically important viral pathogens of salmonid aquaculture

Virus	Abbreviation	Genome	Taxonomic classification	Clinical Signs first reported	Virus identified and characterised	OIE listed (2020)
Infectious pancreatic necrosis virus	IPNV	dsRNA	Birnaviridae	1940 (McGonigle, 1941)	1960 (Wolf et al., 1960)	No
Infectious haematopoietic necrosis virus	IHNV	(−) ssRNA	Rhabdoviridae	1950s (Rucker et al., 1953)	1969 (Wingfield et al., 1969)	Yes
Salmonid alphavirus	SAV	(+) ssRNA	Alphaviridae	1976 (Munro et al., 1984)	1995 (Boucher and Laurencin, 1994; Castric et al., 1997)	Yes
Infectious salmon anaemia virus	ISAV	(−) ssRNA	Orthomyxoviridae	1984 (Thorud, K., 1988)	1995 (Dannevig et al., 1995; Mjaaland et al., 1997)	Yes
Piscine myocarditis virus	PMCV	dsRNA	Totiviridae	1985 (Amin and Trasti, 1988)	2010 (Løvoll et al., 2010; Haugland et al., 2011)	No
Piscine reovirus	PRV	dsRNA	Reoviridae		2011 (Palacios et al., 2010)	No

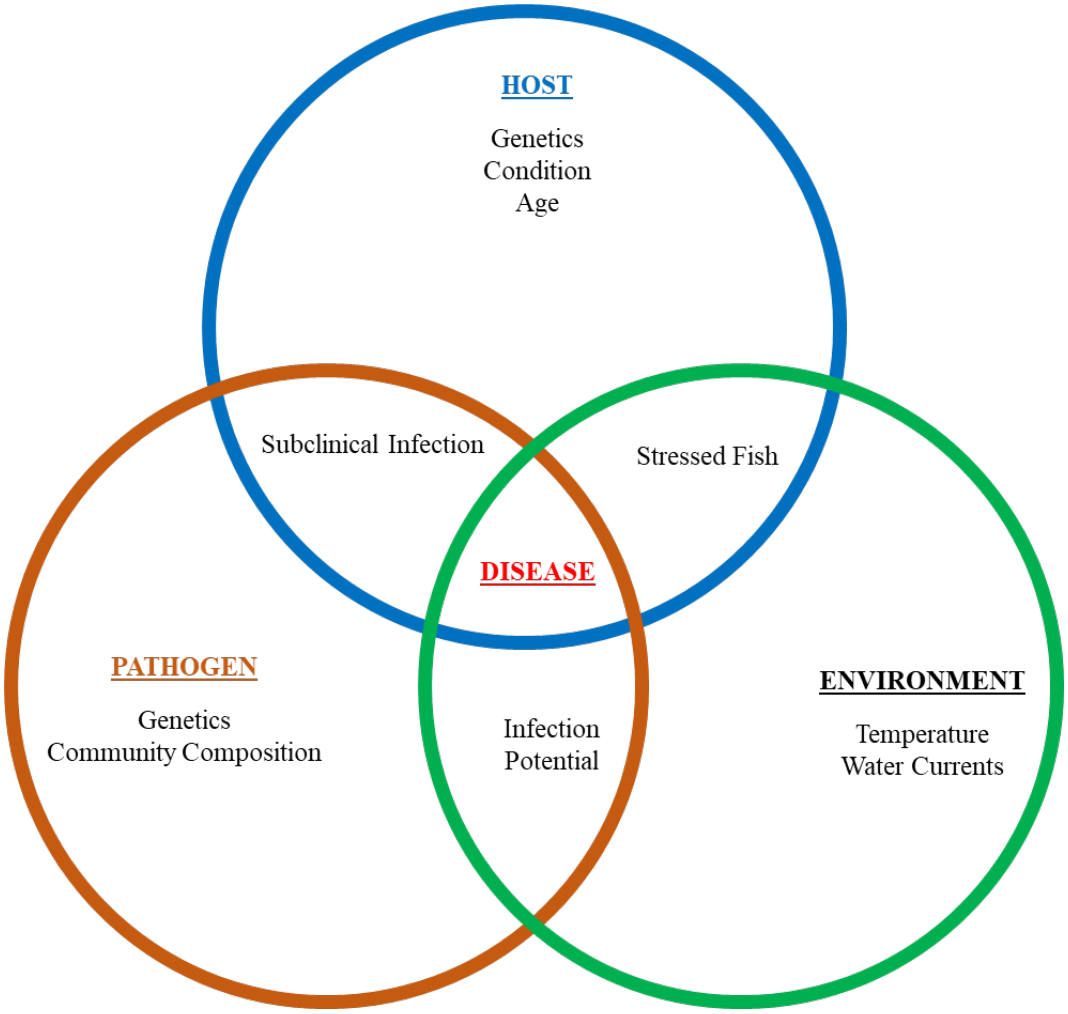

Figure 1.1. Venn diagram of the factors required for a disease phenotype. Without any of the three main components, the host may remain healthy and asymptomatic. In the absence of pathogens, non-ideal environmental conditions may cause the host to become stressed. In favourable environmental conditions, infections with pathogens can result in subclinical or asymptomatic infections, where the host carries the pathogen but does not exhibit symptoms. Finally, environments without suitable hosts cannot result in diseased hosts, but these conditions do pose a potentially infectious environment should suitable hosts become available

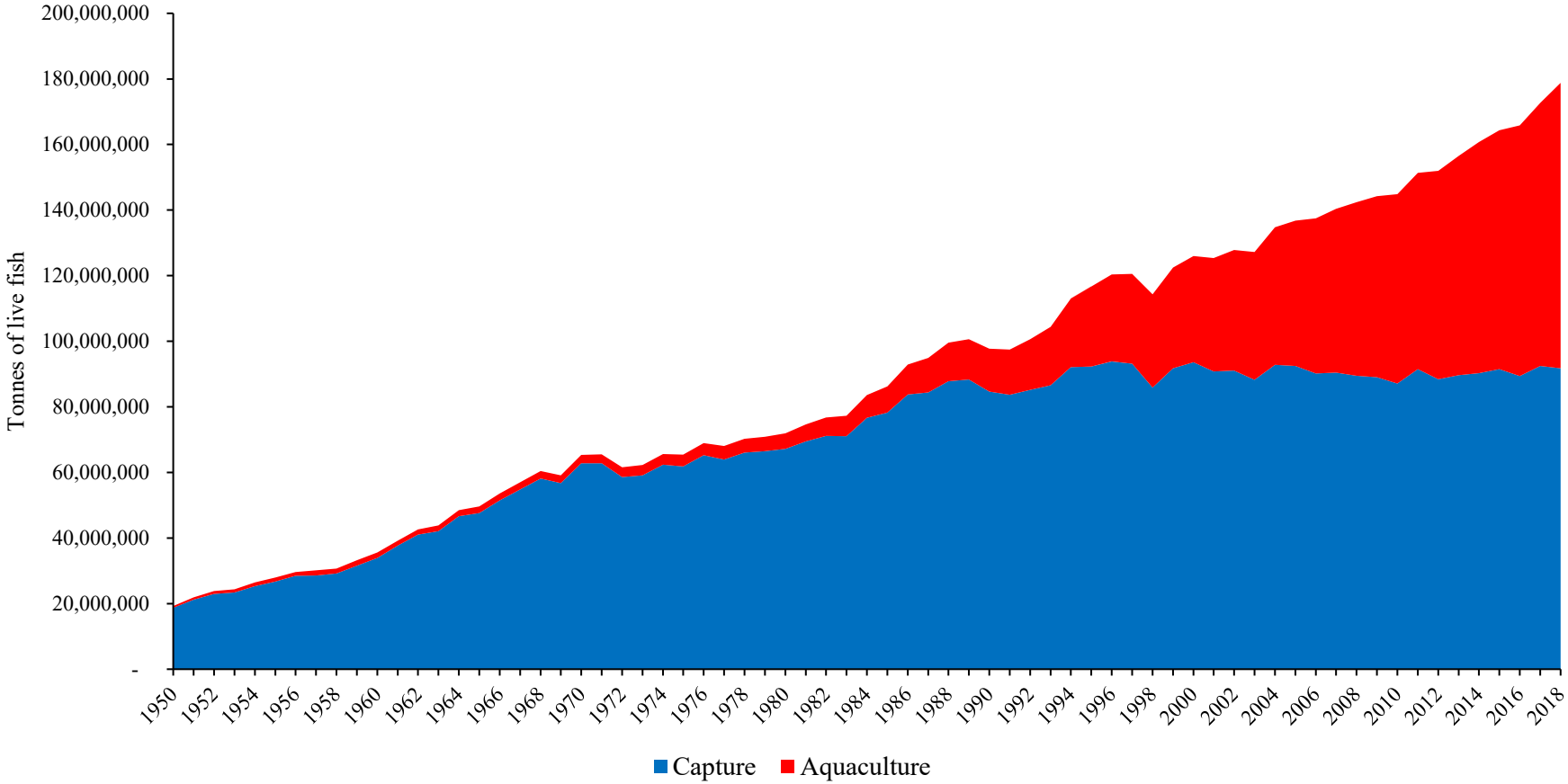

Figure 1.2. Aquaculture vs capture fisheries production quantity trends. (Source – FishStatJ FAO)

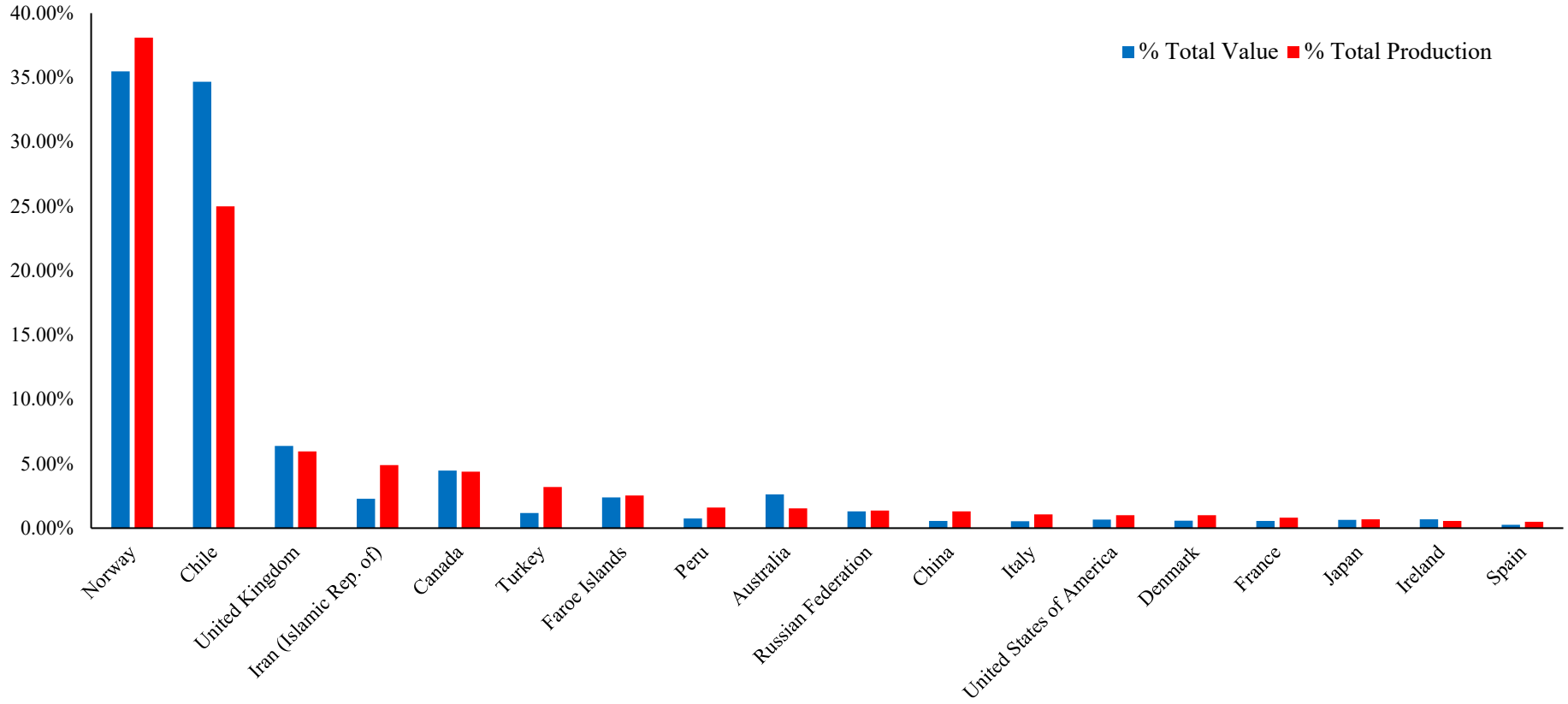

Figure 1.3. Salmonid aquaculture value vs production weight by country of origin. (Source – FishStatJ FAO)

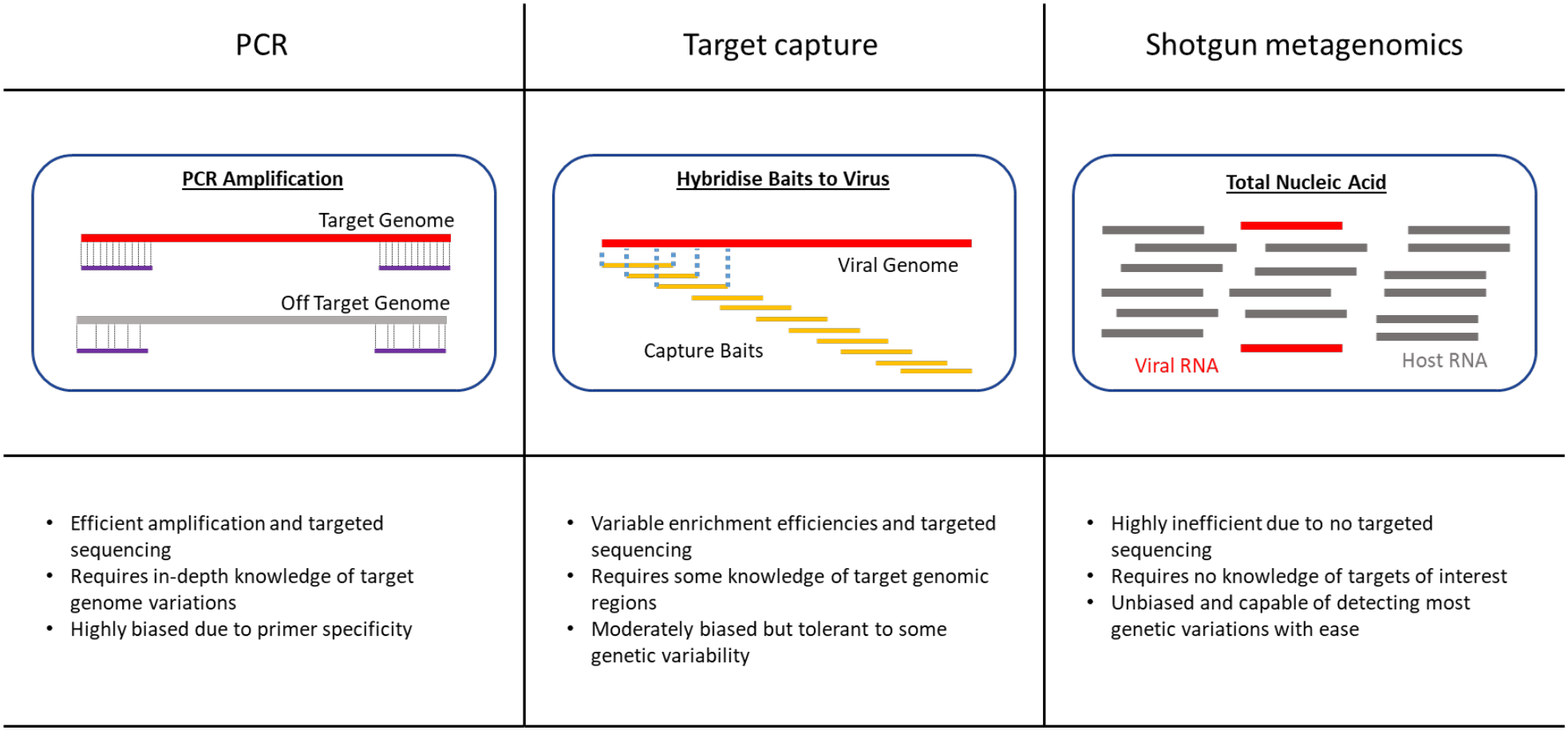

Figure 1.4. Schematic of different sequencing approaches ranging from highly efficient but template specific PCR amplification, to the unbiased but inefficient shotgun metagenomics.

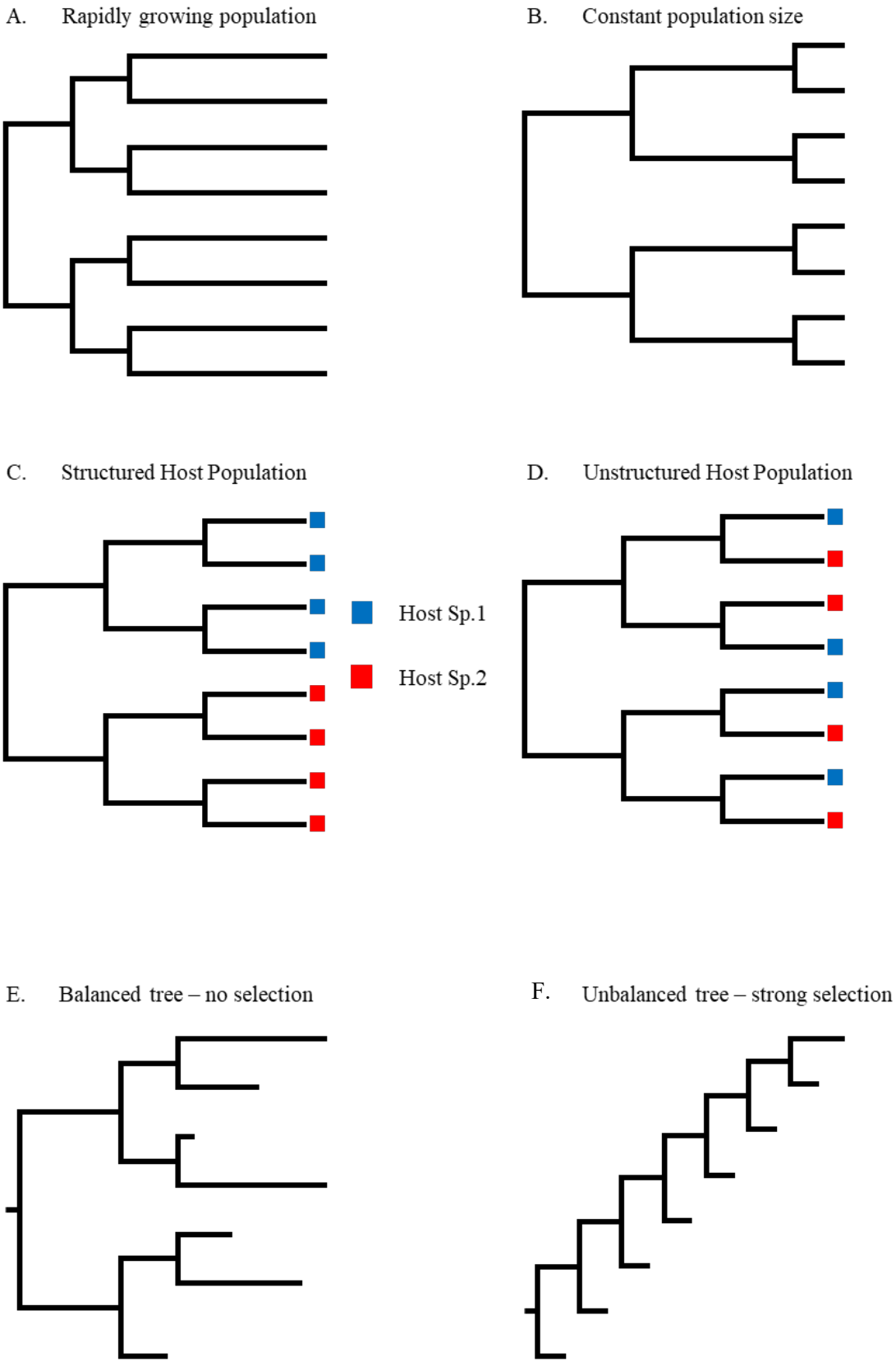

Figure 1.5. Basic viral phylodynamics using tree topology to infer viral population dynamics. (Adapted from Grenfell et al., 2004)

Chapter 2. Nanopore sequencing for rapid diagnostics of salmonid RNA viruses.

The data presented in this Chapter was published as **Gallagher, M.D**., Matejusova, I., Nguyen, L., Ruane, N.M., Falk, K., Macqueen, D.J. *Nanopore sequencing for rapid diagnostics of salmonid RNA viruses*. Sci Rep 8, 16307 (2018).

https://doi.org/10.1038/s41598-018-34464-x

Summary

This Chapter describes the development and validation of a long-read sequencing approach to characterise the genomes of known salmonid viral pathogens. Analysis of pathogen genome variation is essential for informing disease management and control measures in farmed animals. For farmed fish, the standard approach is to use PCR and Sanger sequencing to study partial regions of pathogen genomes, with second and third-generation sequencing tools yet to be widely applied. This approach uses PCR to amplify viral cDNA in long overlapping amplicons from infected samples and Oxford Nanopore's MinION platform to perform long-read sequencing on the product. I use this method to present the first SAV subtype-6 genome, which branches as the sister to all other SAV lineages in a genome-wide phylogenetic reconstruction. The Chapter then compares the required sequencing depth to achieve comparative accuracy to Sanger sequencing of the same samples.

2.1 Introduction

As described in the previous Chapter, whole genome sequencing of pathogens greatly enhances the study of viral disease evolution, phylogeography and epidemiology (Houldcroft et al., 2017), including human epidemics such as Ebola (Holmes et al., 2016), HIV (Worobey et al., 2016), and influenza (Su et al., 2015; Vijaykrishna et al., 2015a). Second-generation sequencing platforms (e.g. Illumina) are now used routinely for genome-wide monitoring and investigations of viral disease, and generate accurate short-read data at massive throughput (Datta et al., 2015; Jones et al., 2017; Qureshi et al., 2018), typically requiring computationally-intensive analysis pipelines. Third-generation platforms, including single-molecule real time (SMRT) (Rhoads and Au, 2015) and Oxford Nanopore (Laver et al., 2015) show high promise for genome-wide analysis of viruses (Li et al., 2017; Quick et al., 2017), and bring the additional benefit of longer sequencing reads offset by higher error rates. The

MinION Nanopore sequencer is a particularly promising technology for viral research and diagnostics, owing to several unique features (outlined in Section 1.5.4) that have, for example, allowed human pathogens to be rapidly characterized in the field without high-power computing or major laboratory infrastructure (Edwards et al., 2016; Faria et al., 2016; Hoenen et al., 2016; Euskirchen et al., 2017; Johnson et al., 2017).

Aquaculture is the fastest growing food production sector (FAO, 2016), yet its sustainability and expansion is threatened by infectious diseases. Among a list of concerning pathogens, several known viral disease agents cause major animal health and welfare issues, accompanied by massive financial losses through mortalities, slow growth, poor flesh quality, treatment interventions and control protocols (e.g. culling) (Aunsmo et al., 2010; Lafferty et al., 2015). Accurate diagnosis of viral diseases is an essential part of strategic planning to manage existing and limit future outbreaks, and is especially important considering the lack of fully-effective treatments and vaccines for most fish viral pathogens (e.g. Karlsen et al. 2012; Garver, LaPatra, and Kurath 2005; Munang'andu et al. 2012). Recommended diagnostic procedures of viral disease include demonstration of clinical pathology coupled to the presence of pathogen DNA/RNA, followed by culturing to establish the presence of viable pathogen (OIE, 2017b). Diagnostic sequencing of aquatic viruses is typically done by PCR and Sanger sequencing, which benefits from high accuracy and established protocols. However, such approaches are limited to relatively short sequences (i.e. up to 1500 bp when sequencing both directions) and cannot gain a genome-wide representation of viruses and their variants without non-routine effort. Second and third generation sequencing tools hold promise for the characterization of aquatic viruses (reviewed in Nkili-Meyong et al. 2016; Bayliss et al. 2017), including pathogens affecting global fish aquaculture, yet they are being up-taken relatively slowly. The utility of such approaches have been demonstrated by the characterisation of novel pathogens such as Tilapia Lake Virus (TiLV) using Ion Torrent sequencing (Bacharach et al., 2016), the discovery of Piscine Reovirus (PRV) (Palacios et al., 2010) and Piscine myocarditis virus (PMCV) (Palacios et al., 2010) with pyrosequencing, and the analysis of Cyprinid herpesvirus 3 genomes using a target enrichment and Illumina sequencing approach to identify mixed genotype infections (Hammoumi et al., 2016). However, as far as I am aware, at the time of publication, no other published studies had successfully used MinION sequencing to study viral diseases impacting farmed fish.

In this Chapter, I demonstrate rapid genome-wide sequencing of fish viral pathogens using nanopore sequencing on the MinION platform. I focused on two disease agents affecting farmed Atlantic salmon, salmonid alphavirus (SAV) and infectious salmon anaemia virus

(ISAV). SAV is a single-strand positive-strand RNA virus (Family *Togaviridae*) and the causative agent of pancreas disease, prevalent across European salmon aquaculture, with six SAV subtypes (SAV1-6) established (Fringuelli et al., 2008). All SAV sequences published to date have been generated using the Sanger method, including full genomes for SAV1-3 (Weston et al., 2002; Hodneland et al., 2005; Karlsen et al., 2006; Matejusova et al., 2013; Petterson et al., 2013), and partial genomic regions primarily encoding a glycoprotein (E2) or a non-structural protein (nsP3) (neither representing known virulence markers), for samples representing all six subtypes (e.g. Fringuelli et al. 2008; Bruno et al. 2014). ISAV is a highly pathogenic, segmented, negative-strand RNA virus (Family *Orthomyxoviridae*) often resulting in high mortality rates (Dannevig et al., 1995; OIE, 2017a), with containment and culling being the only effective mitigation strategy (Stagg, 2003). ISAV genomes have been Sanger-sequenced from several 'genogroups' (Clouthier et al., 2002; Markussen et al., 2008; Cottet et al., 2010; Merour et al., 2011; Toro-Ascuy et al., 2015; Christiansen et al., 2017), while segments 5 and 6, which contain known virulence markers and respectively encode the fusion and hemagglutinin surface proteins, are routinely used for Sanger genotyping, but have also been characterized using Illumina sequencing (Markussen et al., 2013). Overall, in common with other fish viruses, there is a lack of genome-wide data for SAV and ISAV, limiting power to define virulence markers and understand the evolution of different viral lineages. This study linked MinION sequencing to standard PCR enrichment to accurately sequence and genotype both SAV and ISAV. In addition to reporting the first full genome sequence for SAV6, I discuss the potentially transformative applications of MinION sequencing in diagnostics and molecular epidemiology of viruses impacting aquaculture.

2.2 Materials and Methods

2.1.1 Sample preparation and PCR

Total RNA was extracted from SAV and ISAV samples (Table 2.1) using a phenol-chloroform extraction method, except for the SAV6 sample, which was extracted using a Viral RNA Isolation kit (Qiagen). cDNA was synthesised using Protoscript II (New England Biolabs) reverse transcriptase and a mix of random hexamer and oligo dT ($dT_{23}VN$) primers (New England Biolabs) as per the manufacturers' instructions. First-strand cDNA was used as template for long-range PCR reactions.

To amplify the SAV1/6 genomes, degenerate PCR primers targeting three ~4 kb overlapping amplicons (Fig. 2.1) were designed in regions of the genome conserved in the five subtypes where sequence data is available (Table 2.2). PCR was conducted using LongAmp polymerase (New England Biolabs) with cycling conditions as follows: 30 s at 94 °C, followed by 35

cycles of 15 s at 94 °C, 1 min at 56 °C and 3 min 50 s at 65 °C, with a final extension for 10 min at 65 °C. ISAV segments 5 and 6 were amplified using the same approach and primers designed to conserved 5′ and 3′ regions of segment 5/6 (Table 2.2) under the same conditions, except that the PCR extension time was 2 min 30 s. PCR products were visualised on a 1% agarose gel, purified using QIAquick Gel Extraction Kit (Qiagen) and stored at −80 °C until sequencing.

2.1.2 Sanger sequencing of novel SAV genomes

Seven overlapping PCRs were performed in triplicates for five SAV isolates (Table 2.3) according to the methods published by Matejusova et al. (2013). The complete SAV genomes were generated by Sanger sequencing, assembled using Sequencher v5.4.6 and used in the phylogenetic analysis presented in Fig. 2.2.

2.1.3 Preparation of SAV Library and sequencing

1000 ng of equimolar pooled amplicon from each SAV isolate was the input to a library generated with the Ligation Sequencing Kit 1D SQK-LSK108 (Oxford Nanopore Technologies). Before ligating sequencing adaptors, DNA was end-repaired using the NEBNext Ultra II End Repair/dA Tailing kit (New England Biolabs), purified using AMPure XP beads (Beckman Coulter) in a ratio of 1:1 volume of beads per sample and eluted in 30 µl of nuclease-free water (Sigma). Sequencing adapters (AMX1D) (ONT) were ligated to the DNA using Blunt/TA Ligation Master Mix (New England Biolabs) by incubation at room temperature for 10 min. The adapter-ligated DNA library was purified with AMPure XP beads in a ratio of 1:2.5 volume of beads per sample, followed by a wash with Adapter Bead Binding buffer (ABB) (ONT) and elution in 15 µl nuclease-free water. DNA concentrations were determined between each step using a Qubit fluorimeter (Fisher Thermo). Each cleaned library was loaded onto a separate MinION Flow Cell Mk1 R9.4 (ONT) and run via MinKNOW software (without real-time basecalling) for 2 and 3 hours for SAV6 (F1045-96) and SAV1 (SCO/4640/08) respectively.

2.1.4 Preparation of ISAV Library and Sequencing

The ISAV library was prepared using the Ligation Sequencing Kit 1D SQK-LSK108 and a Native Barcoding Kit EXP-NBD103 (Oxford Nanopore Technologies). Segments 5 and 6 from the same virus isolate were pooled in equimolar amounts and 300 ng of each isolate end-repaired using the NEBNext Ultra II End Repair/dA Tailing kit. DNA was purified using AMPure XP beads in a ratio of 1:1 volume of beads per sample and eluted in 30 µl nuclease-

free water. Native barcodes were ligated to 200 ng of end-repaired DNA using Blunt/TA Ligation Master Mix. The barcoded DNA was purified using AMPure XP beads in a ratio of 1:1 volume of beads to sample to remove excess barcodes and eluted in 26 µl nuclease-free water. The barcoded samples were pooled in equimolar amounts to a total of 200 ng library DNA (~0.2 pmol as per Oxford Nanopore Technologies instructions). Barcode adapter mix (BAM) (ONT) was ligated to the library DNA using NEBNext Quick Ligation Reaction Buffer and Quick T4 DNA Ligase (New England Biolabs), and incubated at room temperature for 10 min. Library DNA was purified using AMPure XP beads in a ratio of 1:2.5 volume of beads per sample and subsequently washed with Adapter Bead Binding buffer (ABB) before elution in 15 µl nuclease-free water. DNA concentrations were determined between each step as above. Libraries were loaded according to the native barcoding kit protocol (ONT) onto a MinION Flow Cell Mk1 R9.5, using a 3-hour sequencing run via MinKNOW without real-time basecalling.

2.1.5 Basecalling and consensus assembly

MinION data basecalling and demultiplexing for barcoded ISAV samples was performed using Albacore v.2.1.7 on Windows command line. Base-called FASTQ files were loaded into Geneious v.10 (Kearse et al., 2012) for mapping and analysis. SAV1 (SCO/4640/08) sample reads were mapped to the SAV1 Sanger-reference sequence (Matejusova et al., 2013). SAV6 (F1045-96) sample reads were individually mapped to the partial gene E2 and nsP3 sequences of SAV6 (Fringuelli et al., 2008) and reference genomes for SAV1 (Matejusova et al., 2013), SAV2 (Weston et al., 2002), SAV3 (Petterson et al., 2013), SAV4 (generated in this study; isolate SAV 04-44) and SAV5 (generated in this study; isolate SCO10-684). In order to reconstruct the whole SAV6 genome, mapping was set at 5 iterations and a 65% consensus threshold. The 5 generated SAV6 consensus sequences were then manually inspected and any single base ambiguities resolved by parsimony, giving a final F1045-96 (SAV6) consensus sequence. For example, at position 2235, 4 out of 5 consensus sequences were the base G, whereas one consensus sequence was A: in this case, G was adopted for the final consensus. The ISAV samples were individually mapped to the previously sequenced segment 5 and 6 of the Scot157/08 isolate (Plarre et al., 2012) using the same parameters.

Reads for ISAV NO/Glessvær/2/90 segments 5 and 6, and SAV1 (SCO/4640/08) were subjected to random subsampling to determine the depth of coverage necessary to generate an accurate consensus (i.e. Fig. 2.3). Subsampling was performed in Geneious v.10 using the 'Randomly Sample Sequences' workflow. Subsampled reads were realigned to the reference sequences using the same mapping methods as above and consensus sequences were generated

from each alignment and compared to the reference Sanger sequence using pairwise alignment. Consensus sequences were aligned against all published genome sequences using MAFFT v.7 (Katoh and Standley, 2013) and manually inspected for errors in the mapping that disrupted the protein coding sequences in BioEdit software v.7.2.5 (Hall, 1999). Sequence pairwise similarities were calculated using Geneious statistics of the MAFFT-aligned whole genome sequences.

2.1.6 Genome-wide SAV phylogenetic analyses

Multiple sequence alignment of 23 SAV genomes (Table 2.3) was done using MAFFT v.7, generating an 11,638 bp alignment, which was uploaded to the IQ-TREE server (Trifinopoulos et al., 2016) to determine the best-fitting nucleotide substitution model (GTR) and generate a phylogenetic tree with support values gained from 1,000 Ultrafast Bootstrap iterations (Minh et al., 2013). Bayesian phylogenetic analysis was done using the same dataset in BEAST2 (Bouckaert et al., 2014) employing a relaxed clock model (Drummond et al., 2006), a Coalescent Bayesian Skyline tree model (Pybus and Rambaut, 2009), the GTR substitution model and a Markov Chain Monte Carlo (MCMC) chain of 200 million generations. Tracer (Rambaut et al., 2018) was used to assess MCMC convergence and estimate effective sample sizes for all sampled parameters (>2,000 in all cases). TreeAnnotator was used to remove the first 10% of sampled trees as burn-in and produce a Maximum Credibility Clade (MCC) tree. RootAnnotator (Calvignac-Spencer et al., 2014) was used to estimate posterior support for alternative root positions. MCC trees were visualized using FigTree (http://tree.bio.ed.ac.uk/software/figtree/).

2.3 Results and Discussion

2.3.1 SAV genome-wide sequencing

Using primers matching conserved regions of the SAV genome (Table 2.2), three overlapping PCR amplicons (approx. 4 kb each; Fig. 2.1) were obtained from two samples known to represent SAV1 (SCO/4640/08) and SAV6 (F1045-96) (Table 2.1) and sequenced on separate MinION flow cells (R9.4) for 2–3 hours. Over 98% of each SAV genome was recovered with 90 bp missing at the 5′ and 30 bp at the 3′ region of the genome due to the location of the highly conserved primer binding sites. The average read length from both sequencing runs was ~3800 bp per amplicon, indicating limited DNA shearing during the library preparation. The sequencing of sample SCO/4640/08 was stopped after 3-hours producing over 400 Mb of 'pass' reads (Q-score ≥ 7), resulting in almost 40,000x coverage throughout the genome (Table 2.4). By mapping against the Sanger sequenced SAV1 reference sequence for SCO/4640/08 (Matejusova et al., 2013), a consensus sequence was generated that showed a

99.97% similarity to the reference. The three mismatched bases were called as the degenerate base R (A/G), and were located in the nsP1, E3 and E1 genes.

As the above approach led to an accurate representation of a verified SAV genome sequence, I can be confident in its application to discovering entirely novel variation. For this reason, I decided to sequence SAV6 (sample F1045-96), which has only been identified once, as partial E2 and nsP3 sequences, from a single Irish sample (Fringuelli et al., 2008), and is highly distinct from all other subtypes. After two hours of sequencing, a genome-wide average of 21,000x coverage was achieved. The SAV6 genome consensus showed 100% similarity to Sanger-sequenced nsP3 (EF675499) and E2 (EF675547) gene sequences. Table 2.5 shows consistent genome-wide pairwise similarities contrasting the genome of SAV6 to the other SAV sub-types at both nucleotide and amino acid level (88.6–89.2% and 93.8–94.6% respectively). Variability among SAV subtypes differed based on the gene of interest and the greatest variability was seen in the nsP3 gene (82.0–83.8% and 87.7–89.8% nucleotide/amino acid similarity). In conclusion, these data gained by MinION sequencing confirm for the first time using genome-wide evidence that SAV6 represents a highly-divergent SAV subtype.

2.3.2 Genome-wide SAV phylogeny

Previous studies have failed to establish the position of SAV6 within the SAV phylogeny based on E2 and nsP3 sequences (e.g. Fringuelli et al. 2008; Bruno et al. 2014). I performed genome-wide phylogenetic reconstructions incorporating the new SAV6 genome gained by MinION sequencing, along with 17 SAV genomes available in NCBI, and 5 new (i.e. previously unpublished) Sanger-sequenced genomes for SAV2, 4 and 5 (Table 2.3). I used two probabilistic methods, the first a Bayesian approach incorporating a relaxed clock model (Drummond et al., 2006) allowing estimation of the tree root (Calvignac-Spencer et al., 2014) and the second an unrooted maximum-likelihood (ML) approach (Fig. 2.2). The root of the SAV phylogeny was estimated with high confidence (posterior probability: 0.97), and split SAV6 from all other SAV sub-types. Branching of other subtypes was maximally supported (posterior probability: 1.0; ML bootstrap values > 95%), with SAV3 and 2 forming a monophyletic group separate from a clade containing SAV1, 4 and 5 (Fig. 2.2). The basal phylogenetic position of SAV6 highlights particular importance for the new MinION genome sequence in future investigations of the evolution and phylogeography of the major SAV lineages

2.3.3 ISAV segment 5 and 6 sequencing

To test MinION sequencing on a distinct fish virus, I focused my efforts on ISAV, which exists in eight genomic segments (length: 740–2169 bp). This inherent aspect of the virus limits one of the main benefits of MinION sequencing: its capacity to generate genome-wide representation of a virus with a small number of overlapping PCR amplicons, as done successfully for SAV. I instead focused on ISAV segments 5 and 6, which are widely studied and known to contain ISAV virulence markers, this time testing a barcoding approach to sequence multiple samples on a single MinION flow cell. PCR amplicons (primers in Table 2.2) amplifying 97% of segment 5 and 93% of segment 6 including both virulence markers, were obtained from seven ISAV isolates (Table 2.1) and pooled in equimolar amounts for sequencing after barcoding. After 3 hours, approximately 9,000x mean coverage was achieved per sample. Only one of the isolates used in this study has a reference Sanger sequence (NO/Glessvær/2/90); basecalling accuracy was estimated for segments 5 and 6 of this isolate and 100% similarity was observed.

ISAV segment 6 contains a highly polymorphic region (HPR) at the 3′ end of the gene which is a known virulence marker. The putatively non-pathogenic ISAV, called HPR0, is characterized by a full length of the HPR comprising 35 amino acids and all pathogenic ISAV strains to date (called HPR-deleted) contain a deletion in the HPR region of varying length (Nylund et al., 2003). While none of the isolates used in this study were HPR0, the HPR of all the ISAV isolates used in this study were successfully classified with several different deletions being identified including three samples CA/NB04-85-1/04, CA/NB7178/08, CA/F679/99 which have a deletion previously found only once before and not yet fully characterised (Kibenge et al. 2006) (Table 2.4). In addition, the consensus sequences for each segment 5 captured another proposed virulence marker, the substitution $Q_{266}L$ (Markussen et al., 2008; Cottet et al., 2011; Cárdenas et al., 2014), with all but one isolate (CA/NB04-85-1/04) possessing the L variant. CA/NB04-85-1/04 instead encodes for a proline at this position which while unusual, is also present in a Canadian isolate from the EU/NA genogroup (EF432567) (Kibenge et al., 2007). These data thus demonstrate that MinION sequencing effectively recaptures sequence-level virulence markers.

2.3.4 Optimal sequence coverage

Future studies would benefit from establishing the necessary coverage required to determine confident consensus sequences using MinION. Thus, I randomly sampled MinION reads mapping to segments 5 and 6 of one ISAV sample (NO/Glessvær/2/90) and the SAV1 genome (sample: SCO/4640/08) at different coverages to establish the impact on consensus sequence

accuracy (Fig. 2.3.A–C). 50x and 500x coverage of either ISAV segment achieved a consensus sequence >99% and 100% identical to the Sanger reference, respectively (Fig. 2.3.A,B). For SAV1, just 20x coverage led to 99% similarity with the Sanger reference, while 1,000x coverage led to 99.97% similarity (Fig. 2.3.C). Thus, despite its high error rate (e.g. Laver et al. 2015), a highly-accurate consensus sequence can be generated with very modest MinION sequencing time.

2.3.5 Broader perspectives and comparisons with other platforms

Rapid sequencing of two structurally-distinct fish RNA viruses was achieved with high accuracy using MinION sequencing coupled with PCR. While the samples used were from cultured viruses, I have had equal success using the same protocols and infected tissues with much lower virus titres (see Chapter 3 where tissue samples were used). The methods described were achieved within 24 hours lab-time (Fig. 2.4), exploiting PCR primers matching conserved genomic regions, which allowed a highly divergent viral genome (SAV6) to be sequenced with little prior knowledge of sequence variation. Combining such turn-around and ease of application with the accuracy gained from moderate sequencing coverage opens the doorway to routine high-confidence viral genotyping at shallow phylogenetic scales, sufficient for robust diagnostics supporting disease management and regulatory decisions. Elsewhere, it has also been shown that MinION sequencing can be used to recover viral RNA genomes from infected samples without prior PCR enrichment, which has advantages in the field (Kafetzopoulou et al., 2018) and can also potentially identify viruses beyond the target pathogen. The ease of generating genome-wide sequencing data for non-segmented viruses such as SAV has revolutionary potential for diversifying the relatively restricted current repertoire of publicly-available fish virus genomes, bringing benefits for fundamental research and disease management. However, it is important to acknowledge that this approach is best-suited to generating consensus viral genome sequences, and less useful for identifying population variation within samples, which is well-established for RNA viruses (Descloux et al., 2009; Iqbal et al., 2009; Agoti et al., 2010; Hoelzer et al., 2010; McKinley et al., 2011; Vibin et al., 2018), as the PCR enrichment may introduce biases toward particular variants, and the high sequencing error rate of MinION reduces power to call low frequency variants *de novo*.

Future efforts should also aim to reduce the cost of genome-wide sequencing using multiplexing to exploit the high coverage possible on a single MinION flow cell (e.g. see Chapter 3 for 12x multiplexing). I estimate that the single SAV genomes (~12 kb) generated in this study cost approx. £850 each, including all consumables and an entire flow cell;

however, multiplexing using 96 samples and the same approach would reduce this cost to approx. £50–60 per sample. By comparison, it would not be possible to perform a direct-equivalent Sanger sequencing approach, as the amplicon length exceeds the possible length of sequenced reads. Assuming an SAV genome was tiled across 7 PCR amplicons (e.g. Matejusova et al. 2013) and sequenced directly using Sanger with no cloning step (which would add further costs), I estimate a cost of approximately £100 per SAV consensus genome, including all reagents and bi-directional sequencing. In addition to a per-genome saving, the MinION approach is more convenient and time-efficient when a large number of genomes need to be sequenced, being done in-house in a single sequencing run with fewer amplicons, avoiding the need for cloning and the use of an external Sanger provider. It is more challenging to directly compare costs of this MinION strategy with alternative high-throughput approaches, as there are many platforms and variations in library preparation strategy, and this would also be affected by the extent of sub-contracting to an external provider. However, I estimate that the costs of generating complete SAV genomes using Illumina at the same scale (i.e. 96 samples), assuming the same amplicon strategy followed by in-house library preparation/indexing (Nextera XT DNA kit) and sequencing on the MiSeq platform by an external provider to be approx. £50–65 (i.e. very comparable). While Illumina brings advantages in terms of data accuracy, e.g. giving more scope for detecting viral population variation, the MinION avoids use of an external provider, which typically leads to a lag of weeks to months for delivery. Overall, this MinION approach has some cost and/or time advantages when compared to Sanger and Illumina approaches if the aim is to recover a consensus SAV genome with high accuracy, and future work is needed to develop this approach for robust analysis of viral population variation.

In conclusion, once low cost MinION sequencing of fish viral genomes is achieved, considering the unique portability of the sequencer alongside the modest computational power needed to analyse the resultant data, it seems reasonable to anticipate in-field diagnostic applications in the near future, including the monitoring of viral genotypes and subtypes directly on fish farms and in the field.

Data Availability

MinION sequences for SAV isolates: SRA BioProject Accession SRP142226. SAV6 consensus genome: NCBI accession MH238448. MinION sequences for ISAV: SRA BioProject Accession SRP155694. ISAV segments 5 and 6: NCBI accessions: MH708654-MH708667. Sanger-sequenced SAV genomes: NCBI accessions MH341514 and MH708650-MH708653.

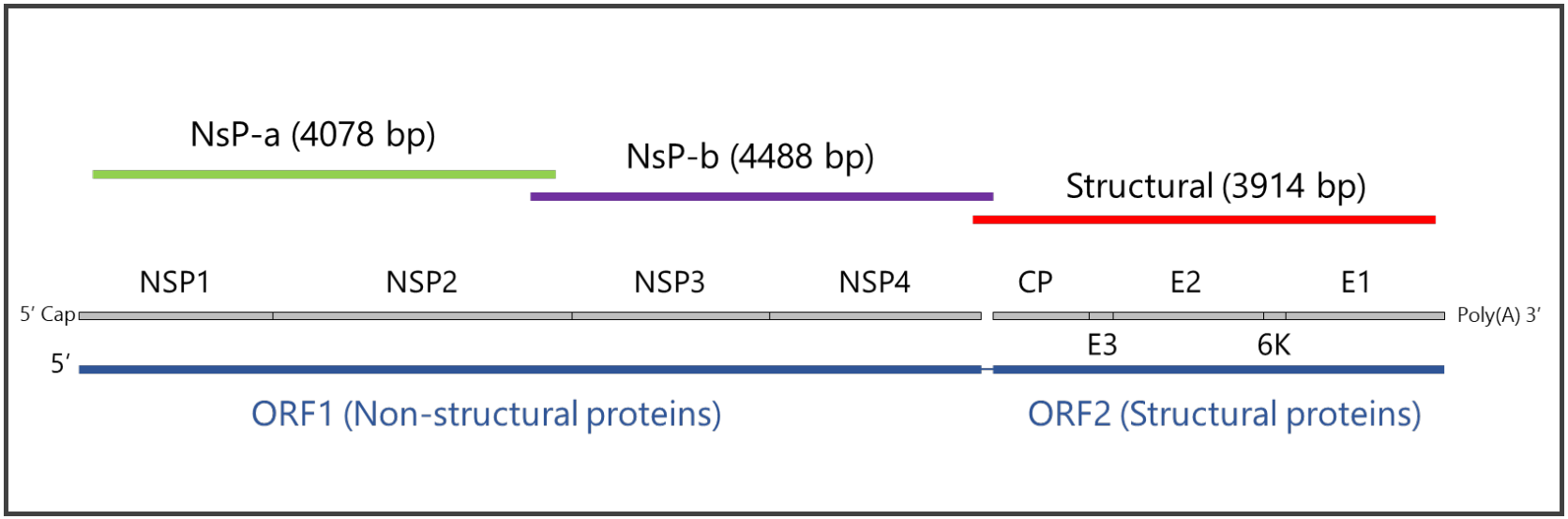

Figure 2.1. Schematic of the three overlapping PCR amplicons covering >98% of the SAV genome. Amplicons are coloured on top (nsP-a, nsP-b, Structural) with corresponding genomic locations to scale in grey and blue.

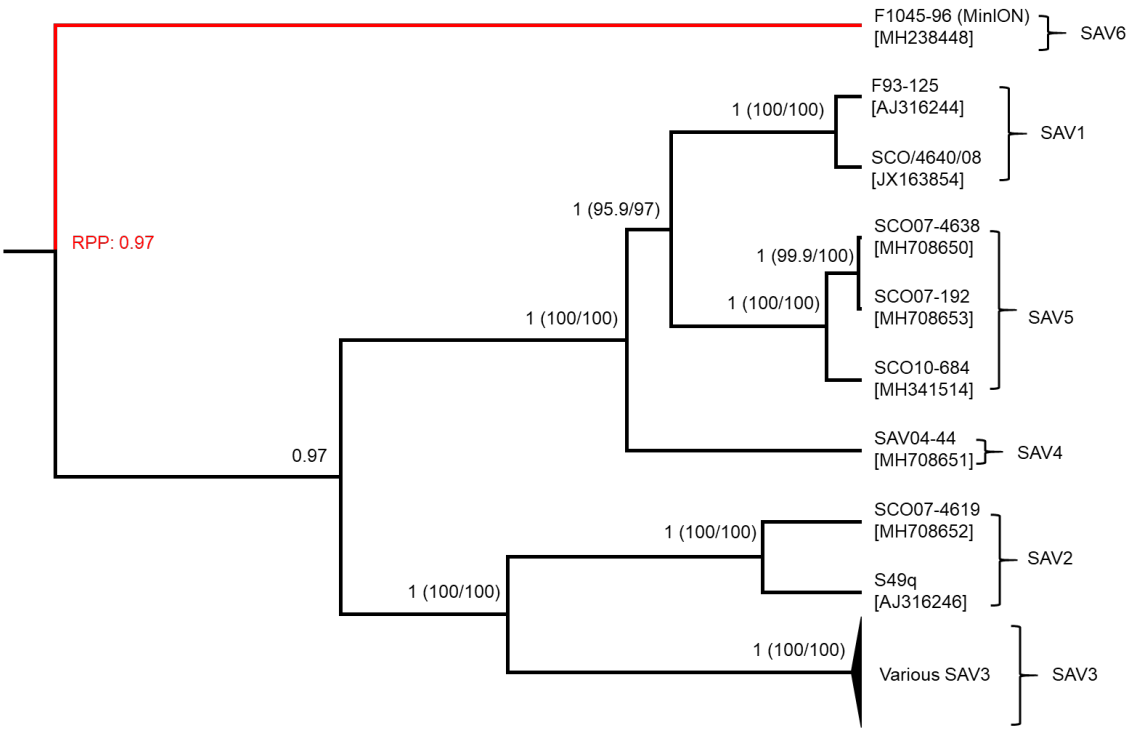

Figure 2.2. Genome-wide Bayesian phylogeny for SAV lineages including the SAV6 sequence generated by MinION sequencing (shown in red).

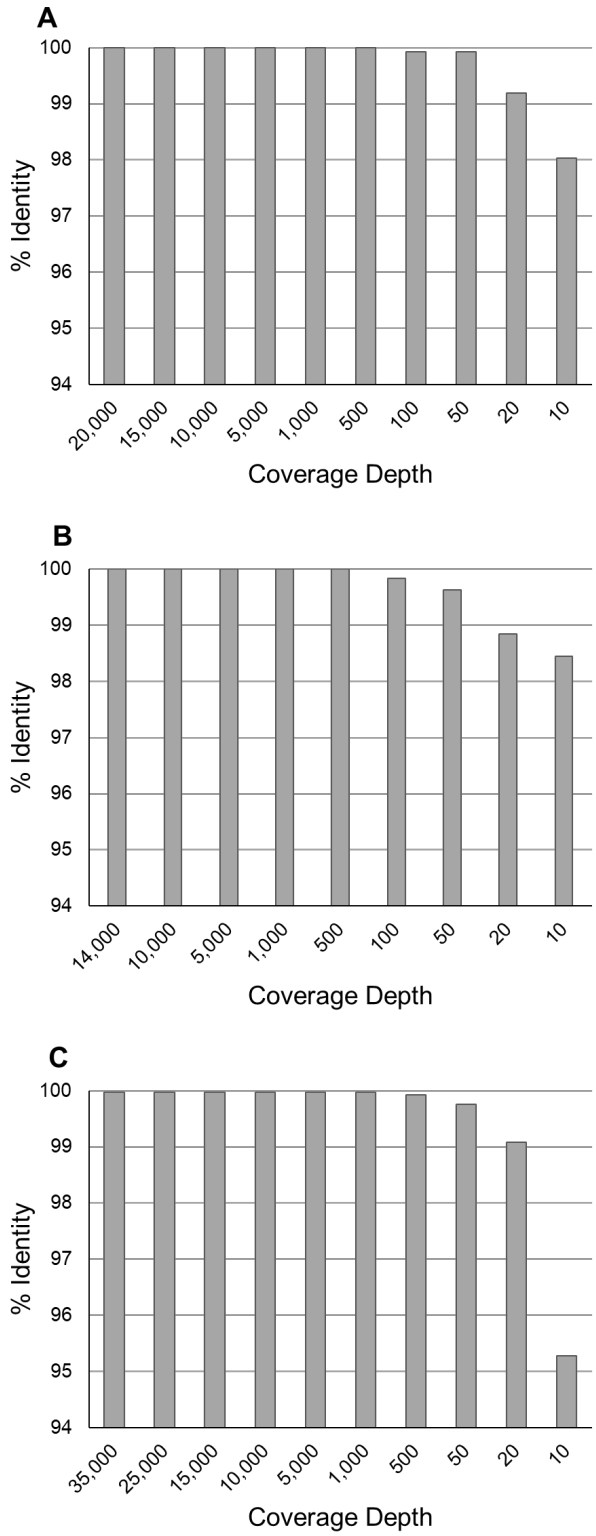

Figure 2.3. Impact of MinION read coverage on accuracy of consensus sequence generation. '% identity' is shown between reference Sanger sequences and consensus sequences generated from randomly sampling MinION reads at multiple sequence coverages for: (A) Segment 5 of ISAV NO/Glessvær/2/90; (B) Segment 6 of ISAV NO/Glessvær/2/90; (C) SAV1 genome (sample SCO/4640/08).

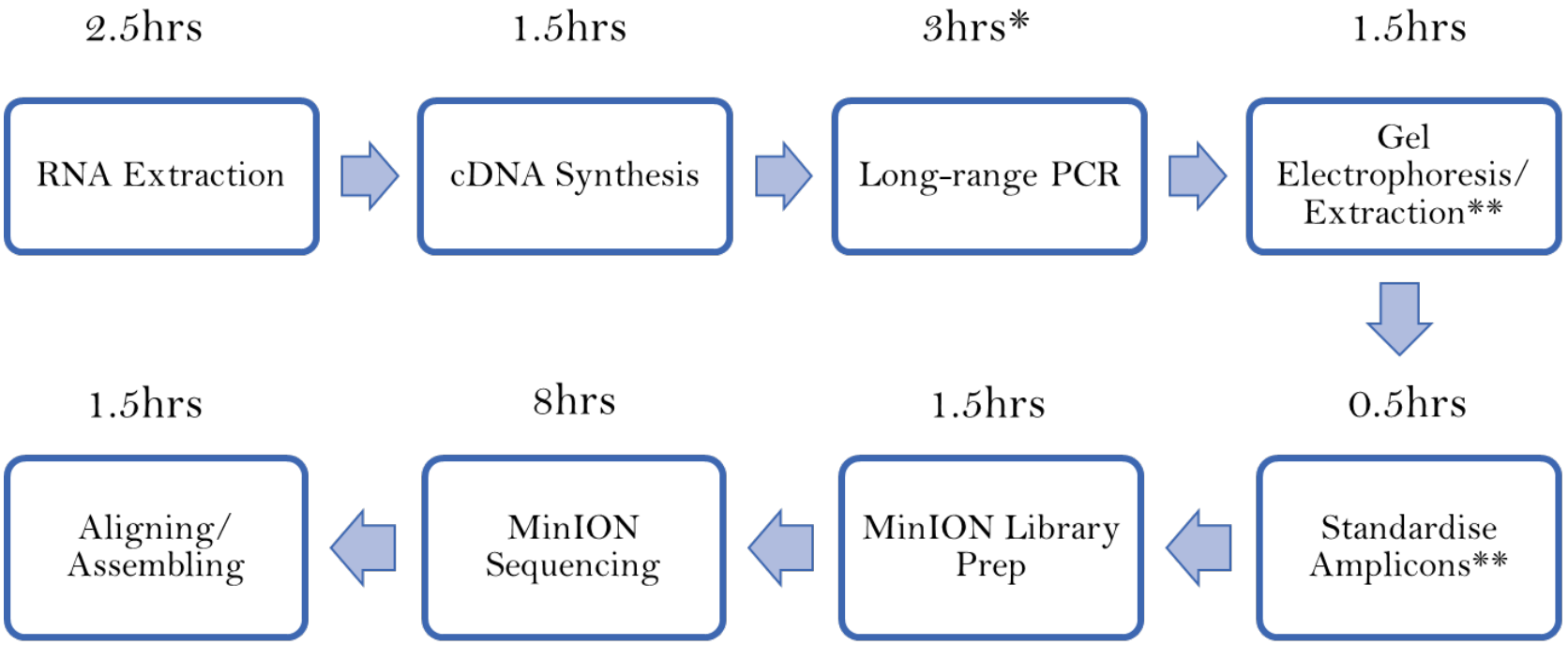

Figure 2.4. Schematic of the MinION sequencing workflow resulting in a <24hr protocol. The long-range PCR step (with an asterisk *) is customisable depending on the length of the amplicons that you are sequencing. Steps with a double asterisk (**) denote optional steps that might improve efficiency or data quality but are not required for the protocol to be successful.

Table 2.1. Details of isolates used for MinION sequencing

Virus	Isolate	Year of Isolation	Country of Isolation	Cell Line	Subtype
SAV	SCO/4640/08	2008	UK	CHSE	SAV1
	F1045-96	1996	Ireland	BF2/EPC	SAV6
ISAV	SCO/4750/09	2009	UK	NA	EU-G1
	CA/NB04-85-1/04	2004	Canada	NA	EU-NA
	CA/NB7178/08	2008	Canada	NA	EU-NA
	CA/F679/99	1999	Canada	NA	EU-NA
	NO/Sotra/B797/92	1992	Norway	NA	EU-G3
	SCO/4661/08	2008	UK	NA	EU-G1
	NO/Glessvær/2/90	1990	Norway	NA	EU-G2

Table 2.2. Primer sequences used for genomic amplification

Virus	Primer Name	Primer sequence (5' - 3')	Melting Temperature (°C)	Amplicon Length
SAV				
	Structural	CMAACTCAGCCTAYCGCCAG	60.65	3914
		GCACTTCTTCACCACGCAG	56.98	
	nsPa	AGACTGCGTTTCCAGGGTT	59.02	4078
		ATGTCGGTCAGTTGAGGGC	57.84	
	nsPb	AGTGGGAYWCTAAGCCGAGAGG	59.55	4488
		TACACGGGGAAGGTGCTCTG	60.72	
ISAV				
	Fusion	ATGGCTTTTCTAACAATTTTAGTCT	56.14	1301
		AGCACCACCAACACAACTACA	56.11	
	Hemagglutinin	GGCACGATTCATAATTTTATTCCT	59.25	1236
		GAACAGAGCAATCCCAAAACCT	60.08	

Table 2.3. Accession details of the SAV strains used in phylogenetic analyses

Virus Strain	Year	Country of Origin	Subtype	Accession Number
F93-125	1993	Ireland	SAV1	AJ316244
S49q	1995	France	SAV2	AJ316246
SavH20/03	2003	Norway	SAV3	AY604235
SAVH10/02	2002	Norway	SAV3	AY604236
PD97-N3	1997	Norway	SAV3	AY604237
SavSF21/03	2003	Norway	SAV3	AY604238
H10	2007	Norway	SAV3	JQ799139
SAV 4640	2008	United Kingdom	SAV1	JX163854
SAV3-7-R/09	2009	Norway	SAV3	KC122918
SAV3-9-R/10	2010	Norway	SAV3	KC122919
SAV3-5-H/10	2010	Norway	SAV3	KC122920
SAV3-8-R/10	2010	Norway	SAV3	KC122921
SAV3-6-H/10	2010	Norway	SAV3	KC122922
SAV3-4-SF/10	2010	Norway	SAV3	KC122923
SAV3-1-T/10	2010	Norway	SAV3	KC122924
SAV3-3-MR/10	2010	Norway	SAV3	KC122925
SAV3-2-MR/10	2010	Norway	SAV3	KC122926
4619	2007	United Kingdom	SAV2	MH708652
SAV04-44	2004	Ireland	SAV4	MH708651
SAV684	2010	United Kingdom	SAV5	MH341514
4638	2007	United Kingdom	SAV5	MH708650
F07-192	2007	Ireland	SAV5	MH708653
F1045-96	1996	Ireland	SAV6	MH238448

Table 2.4. MinION sequencing details after basecalling and quality control.

Virus	Isolate	# of Reads sequenced	# of Reads mapped	% reads mapped	Average Genome Coverage	ISAV HPR
SAV	F1045-96	73,574	66,705	91	21,306	—
	SCO/4640/08	112,805	93,998	83	39,012	—
ISAV	SCO/4750/09	25,009	24,797	99	9,609	HPR35
	CA/NB04-85-1/04	11,816	11,201	95	9,932	Uncharacterised
	CA/NB7178/08	20,136	19,672	98	4,464	Uncharacterised
	CA/F679/99	18,650	18,308	98	4,593	Uncharacterised
	NO/Sotra/B797/92	13,410	13,192	98	4,950	HPR1
	SCO/4661/08	23,793	23,584	99	9,710	HPR35
	NO/Glessvær/2/90	39,343	38,463	98	19,232	HPR2

Table 2.5. Pairwise similarities between SAV6 and reference genomes for SAV1-5

Gene	**SAV6/SAV1 (NUC/AA)**	**SAV6/SAV2 (NUC/AA)**	**SAV6/SAV3 (NUC/AA)**	**SAV6/SAV4 (NUC/AA)**	**SAV6/SAV5 (NUC/AA)**
nsP1	91.8/94.7	91.6/95.5	92.7/95.7	91.8/95.3	92.3/95.7
nsP2	89.9/95.6	89.8/95.9	89.8/96.7	89.5/95.8	89.8/96.2
nsP3	83.8/88.9	82.8/88.2	82.0/87.7	83.2/89.4	83.8/89.8
nsP4	87.9/95.6	89.3/96.1	88.5/96.4	87.5/95.2	88.3/96.2
CP	90.2/91.8	89.2/90.8	90.3/93.3	90.7/93.3	91.5/92.9
E3	88.7/93.0	85.4/93.0	86.9/94.4	83.6/88.7	85.9/93.0
E2	87.8/92.7	87.1/91.8	87.2/92.9	85.5/92.5	87.0/93.4
6K	91.2/95.6	89.7/94.1	93.1/97.1	91.7/97.1	91.2/95.6
E1	91.4/96.7	90.5/96.2	90.6/95.6	90.9/96.9	91.9/97.1
Genome	89.2/93.9	88.6/93.8	88.7/94.3	88.3/94.2	89.0/94.6

Chapter 3. Genome Sequencing of SAV3 reveals repeated seeding events of viral strains in Norwegian aquaculture

The data reported in this chapter was published as **Gallagher, M.D**., Karlsen, M., Petterson, E., Haugland, Ø., Matejusova, I. and Macqueen, D.J. *Genome Sequencing of SAV3 Reveals Repeated Seeding Events of Viral Strains in Norwegian Aquaculture*. Front Microbiol, 11, p.740 (2020)

Summary

This Chapter details an example of the utility of the Nanopore sequencing method outlined in Chapter 2. The goal of this study was to understand recent transmission dynamics of salmonid alphavirus (SAV) in Norway. To avoid the bias introduced by culturing viral isolates, genome sequences from twenty-four naturally infected SAV3 tissue samples from Norway and collected between 2016 and 2019 were generated. Phylogenetic analyses revealed that the currently active SAV3 strains sampled comprise four distinct lineages sharing an ancestor that existed ~15 years ago (95% highest posterior density interval: 12.51 to 17.7 years) and likely in Hordaland. Furthermore, the ancestor of the strains that were sampled outside of Hordaland (Sogn of Fjordane and Rogaland) existed less than eight years ago, indicating a lack of long-term viral reservoirs in these counties. This evident lack of geographically distinct subclades is compatible with a source-sink transmission dynamic explaining the long-term movements of SAV around Norway. Such anthropogenic transport of the virus indicates that at least for the sink counties, biosecurity strategies might be effective in mitigating the ongoing SAV epidemic.

3.1 Introduction

Salmon pancreas disease virus (SPDV), commonly known as salmonid alphavirus (SAV) is a major economically damaging pathogen of European salmonid aquaculture, causing pancreas disease (PD) in Atlantic salmon and sleeping disease (SD) in freshwater rainbow trout (Weston et al., 1999; McLoughlin and Graham, 2007). SAV is a (+)ssRNA virus (family *Togaviridae*) with a ~12kb genome consisting of two open reading frames (ORFs), encoding the structural polyprotein (~4kb) and the non-structural polyprotein (~8kb) (Weston et al., 2002). The viral genome exists as a polyadenylated ~12kb genomic RNA molecule from which the non-structural polyprotein is translated, and a transcribed ~4kb sub-genomic RNA, also polyadenylated, from which the structural polyprotein is translated (Weston et al., 1999, 2002; Villoing et al., 2000).

Six subtypes of SAV have been identified by phylogenetic analysis (SAV1-6) (Fringuelli et al. 2008; Graham et al. 2012), which are separated geographically, with Scotland reporting cases of SAV1, SAV2, SAV4, and SAV5, Ireland reporting cases of SAV1, SAV2, SAV5, SAV6 (Graham et al., 2012), and Norway presenting sustained epidemics of SAV2 and SAV3 (Hodneland et al., 2005; Hjortaas et al., 2013). Previous work has shown that the different subtypes diverged prior to the onset of modern aquaculture, and that it is likely that each SAV subtype had independent introductions to farmed salmonids (Karlsen et al., 2014). Additionally, multiple wild fish species testing positive for SAV have been identified as being potential viral reservoirs including common dab (*Limanda limanda*), long rough dab (*Hippoglossoides platessoides*), European plaice (*Pleuronectes platessa*) and ballan wrasse (*Labrus bergylta*) (Snow et al., 2010; Bruno et al., 2014; McCleary et al., 2014; Ruane et al., 2018), with the ancestral source of SAV likely being centred in the North Sea (Karlsen et al., 2014). Such discoveries have shown that salmonids are not the exclusive host range of SAV, but is instead present in a range of other fish species, though SAV has not yet been shown to cause mortalities in non-salmonids.

Similar to other RNA viruses, the high rate of evolution in SAV can be detected at a genetic level within timeframes of a few years (Karlsen et al., 2006, 2014). Previous work on Norwegian SAV3 showed the presence of two co-circulating strains that overlap both temporally and spatially, across wide geographic distances, indicating significant genetic diversity within SAV3 available for molecular epidemiological studies (Karlsen et al., 2014). Both SAV subtypes present in Norway are geographically structured with the counties of Rogaland and Hordaland having a SAV3 epidemic, Trøndelag having a SAV2 epidemic, and Møre og Romsdal reporting cases of both SAV2 and SAV3 (Norwegian Veterinary Institute, 2018). While the overwhelming majority of SAV cases in Sogn og Fjordane are SAV3, a small number of SAV2 outbreaks have occurred (Norwegian Veterinary Institute, 2018). These regions of overlapping epidemics provide ample opportunity for co-infections to arise between the two subtypes, observed recently in Møre og Romsdal in a case of PD (Norwegian Veterinary Institute, 2018).

Both the transport of fish stock and passive drift along water currents are suggested to be involved in the spread of SAV strains (Karlsen et al., 2006; Kristoffersen et al., 2009; Viljugrein et al., 2010). However it is likely that the two routes of viral transmission play roles at different spatial and temporal scales. While passive drift along water currents is probably important in local outbreak clusters and for several years at a time, anthropogenic transport of virus (i.e. the movement of fish stock) may play a more important role in viral transmission across large geographic distances and the seeding of previously uninfected regions. Assuming the major explanatory factor for SAV3 transmission was transport by water currents, we would

expect to recover geographically distinct groups of SAV in a phylogenetic analysis, particularly when considering the high evolutionary rate of SAV (Karlsen et al., 2014) and length of the SAV3 epidemic in Norway (Poppe et al., 1989; Hjortaas et al., 2016). However, it has not yet been possible to test alternative predictions about SAV transmission routes, as the most recent publicly available genome sequences from Norwegian SAV3 were sampled in 2010 (Petterson et al., 2013), leaving the current status of SAV genetic diversity and SAV evolutionary dynamics in Norwegian aquaculture poorly characterized.

Therefore to characterize the current phylogenetic structure of SAV3 and thus test the hypothesis that passive drift by water currents has been the dominating transmission mechanism, an established Nanopore amplicon sequencing approach (see Chapter 2) was used in the current study. After first confirming the accuracy of this approach for distinguishing subtype level co-infections of SAV2 and SAV3, which we considered a hypothetical possibility in regions with overlapping epidemics, twenty-four SAV3 genomes were sequenced from heart tissues samples between 2016 and 2019. This dataset allowed us to characterise the current genetic diversity of SAV3 in Norwegian aquaculture. We found that several distinct lineages of SAV3, representing independent transmission chains, are active in Norway. Lack of clear geographic structuring in these lineages further suggested that anthropogenic transport of the virus likely played a significant role in shaping the current genetic structure and is an important transmission route that acts not only over large geographic distances, but also locally.

3.2 Methods

3.2.1 Sample Preparation and PCR:

Twenty-three naturally SAV-infected heart tissues of Atlantic salmon and one heart tissue from rainbow trout stored in RNAlater were obtained from PHARMAQ Analytiq (Zoetis) (Table 3.1). Samples were selected from across as wide a geographic range as possible and were sampled between 2016 and 2019. Total RNA was extracted from each sample using a phenol-chloroform approach, and integrity assessed using gel electrophoresis. cDNA was synthesised using Protoscript II Reverse Transcriptase (New England Biolabs) and a mix of random hexamer and anchored-dT ($dT_{23}VN$) primers. First strand cDNA was used as template for PCR reactions.

Viral genomes were amplified in six overlapping PCR amplicons of roughly 2kb in length (primer sequences available in Table 3.2) using LongAmp polymerase (New England Biolabs) with the following PCR cycling conditions: 30 s at 94 °C, followed by 35 cycles of 15 s at 94 °C, 1 min at 56 °C and 2 min 15 s at 65 °C, with a final extension for 10 min at 65 °C. Primers were designed in regions of the SAV genome conserved between SAV2 and SAV3

(Table 3.2). All PCR products were visualised on a 1% agarose gel before being excised and purified using Monarch Gel Extraction Kit (New England Biolabs), eluted in 10µl of elution buffer and stored at -80ºC until sequencing.

3.2.2 MinION library Preparation:

500ng of PCR product from each sample, quantified with the Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific), was used as the input for MinION library preparation with the Ligation Sequencing Kit 1D SQK-LSK109. DNA was end-repaired with NEBNext Ultra II End Repair/dA Tailing kit (New England Biolabs) and purified with AMPure XP beads (Beckman Coulter) in a 1:1 ratio before being eluted in 25µl of nuclease-free water (Sigma-Aldrich). 350ng of each cleaned DNA sample was barcoded using the Native Barcoding Expansion 1-12 and 13-24 kits (ONT: EXP-NBD104 and EXP-NBD114) and Blunt/TA Ligation Master Mix (New England Biolabs) before being purified with AMPure XP beads in a 1:1 ratio and eluted in 26µl of nuclease-free water. The barcoded samples were pooled in equal quantities to a total of 250ng in 45ul of water. Sequencing adapters (AMII - ONT) were ligated to the DNA using Blunt/TA Ligation Master Mix before being purified with AMPure XP beads in a 1:2 ratio and washed with Short Fragment Buffer (SFP) (ONT) before being eluted in 15µl of elution buffer. Samples were sequenced in two libraries on separate R9.4.1 MinION flow cells (ONT) (Table 3.1) without live basecalling.

3.2.3 Data Analysis:

MinION sequence basecalling and demultiplexing was performed with Guppy v3.1.5 using the high accuracy basecaller on a Linux CPU system with default parameters. Resulting FASTQ files were aligned to reference sequences for SAV2 (MH708652) and SAV3 (JQ799139) simultaneously using MiniMap2 (Li, 2018) with default parameters. Resulting alignment files were visualised in Geneious v.2019.0.4 and consensus sequences were generated using the 'Highest Quality' threshold parameter. Consensus sequences were inspected manually for alignment errors or frameshift mutations that would disrupt the protein coding sequence of the genome.

FASTQ files were also aligned to a reference genome using the NGMLR mapper and analysed for structural variants (deletions, duplications and inversions) using Sniffles (Sedlazeck et al., 2018) with the following parameters: a minimum coverage of 50 reads per variant and a minimum variant size of 10bp. Anything smaller than 10bp was not reliably detected by this software due to the prevalence of random indels generated during Nanopore sequencing (Table 3.4). Resulting variant calling files were visualised in IGV and structural variants were manually inspected to reduce false positive calls.

Phylogenetic analysis was performed using the twenty-four consensus sequences generated in this study as well as all previously published whole genome sequences of SAV3 strains with sampling dates and locations, obtained from NCBI (Table 3.3). Sequences were aligned using MAFFT v.7 with default parameters (Katoh and Standley, 2013) and the best fit substitution model was determined using IQTREE (Nguyen et al., 2015; Kalyaanamoorthy et al., 2017) (TIM2+G4; Posada, 2003). To further imply the evolutionary rate and divergence estimations, a regression of root-to-tip genetic distances against date of sampling was performed using TempEst (Rambaut et al., 2016) with the input tree generated in IQTREE (Nguyen et al., 2015) with the same best fit substitution model as above. A 11,681bp alignment of 37 sequences was used for phylogenetic analysis in BEAST2 (Bouckaert et al., 2014) using tip-date time calibration, an uncorrelated relaxed clock model (Drummond et al., 2006) and a Coalescent Bayesian Skyline tree prior (Drummond et al., 2005). Ancestral state location reconstruction was used to estimate the likely geographic location of each node of the tree (Lemey et al., 2009). A single MCMC chain was run for 200 million generations, sampled every 20,000 generations and sampling convergence was confirmed with Tracer v1.7.1, evidenced by effective sample sizes >200 for all parameters (Rambaut et al., 2018). A maximum clade credibility tree was created using TreeAnnotator v2.5.1 (Drummond et al., 2012) after removing the first 10% of trees as burn-in. The resulting trees were visualised in FigTree v1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/).

3.2.4 Validation of Nanopore sequencing to detect subtype-level co-infections

As several of the samples sequenced in this study were from regions where both SAV2 and SAV3 have been detected, the effectiveness of this sequencing method at detecting subtype-level co-infections was determined. This rationale also follows work elsewhere in this Thesis (see Chapter 4), which provided evidence for SAV subtype-level infections in the same samples using a short-read sequencing approach (Gallagher et al., 2020). In the current study, Nanopore reads from samples of confirmed single-subtype infections were individually mapped to a structural polyprotein reference sequence of the relevant SAV subtype using MiniMap2 (Li, 2018) with default parameters. Mapped reads were extracted and only reads of >1,500bp were used for subsequent analyses. Reads from each sample were sequenced on separate flow cells and so were labelled with a different run ID. To simulate subtype-level co-infections, reads were combined in different proportions from each sample so that each 'co-infection' had 10,000 reads in total ranging from 5% SAV2 reads to 95% SAV2 reads, and the corresponding ratio of SAV3 reads. These artificial 'co-infections' were then simultaneously mapped to reference sequences of both SAV2 and SAV3 using MiniMap2 and default parameters. The alignment files were visualised in Geneious v.2019.0.4 and the

number of reads that mapped to the incorrect reference was calculated as the mapping error rate.

3.3 Results

I sequenced 24 SAV3-infected fish hearts from Norway using an overlapping PCR amplicon approach and the MinION Nanopore platform (Table 3.1). This resulted in an average coverage of 6,459x across samples (minimum average coverage of 1,089x and maximum average coverage of 12,434x); significantly more than the minimum requirement for high consensus sequence accuracy (Gallagher et al., 2018). Of the 24 samples sequenced, near full genomes were recovered from 21 samples (approximately 11,600bp in length), while partial genomes were recovered from 3 samples (~ 9,500bp). Overall, the SAV sequences generated in this study were found to be highly conserved, with an average nucleotide and amino acid similarity of the SAV3 sequences being 99.7% and 99.8% respectively.

3.3.1 Validation of Nanopore sequencing to detect subtype-level co-infections

Combining Nanopore reads from single-subtype infections sequenced using the above approach allowed us to test this method's ability to correctly detect SAV subtype level co-infections by mapping reads simultaneously to multiple reference genomes. A similar method has been used recently to provide evidence of SAV co-infections using high accuracy Illumina data (Gallagher et al., 2020), but its applicability to error-prone long-reads was not previously established. Reads from SAV2 and SAV3 were combined in a range of ratios, producing bioinformatic mimics of co-infection scenarios. All ratios of SAV2:SAV3 tested resulted in highly accurate mapping with less than 0.3% of the reads being mapped to the incorrect reference sequence across all samples (mean: 0.16% error rate). All naturally infected samples sequenced in this study were analysed with this approach to detect any subtype co-infections, however all samples proved to be single subtype in origin.

3.3.2 Evolutionary rate analysis

Heterochronous gene sequences (i.e. sequences sampled at different time points) can be used to infer time-constrained phylogenies, especially those of rapidly evolving RNA viruses. However for reliable estimation of a time-scaled phylogenetic tree, sequences should contain enough temporal signal to reconstruct the relationship between time and genetic distance (Rambaut et al., 2016). To determine whether such an analysis is appropriate for this data, and to estimate the evolutionary rate of SAV3, an analysis on the clock-like behaviour of SAV was performed using TempEst. A root-to-tip regression analysis showed that the whole dataset (37 SAV3 genome sequences; 11,861bp alignment) showed temporal signal (correlation coefficient, 0.765) and was subsequently used to estimate the evolutionary rate of SAV3. The evolutionary history was reconstructed with a relaxed molecular clock and a coalescent

skyline population demographic model (Drummond et al., 2005). The estimated evolutionary rate for SAV3 was 7.351 $x10^{-5}$ substitutions per site per year (95% highest posterior density [HPD], 5.33 $x10^{-5}$ to 9.994 $x10^{-5}$).

3.3.3 Phylogenetic inference and phylogeography of SAV in Norwegian aquaculture

To better understand patterns of SAV movement in Norwegian aquaculture, I performed a Bayesian phylogeographic analysis (Lemey et al., 2009) including samples generated in the study, along with other publicly available SAV3 genome sequences (Figure 3.1). The phylogeny suggests that SAV3 consists of two distinct clades (previously observed in Karlsen et al., 2006; Jansen et al., 2010), here defined as SAV3a and SAV3b (Figure 3.1), which diverged approximately 18.9 years ago (95% HPD, 17.15 to 21.08 years) (Figure 3.2), relatively early on in the SAV3 epidemic . These two clades differ by silent substitutions in just two locations in the genome, one in nsP2 (703_K) and the other in E2 (539_S). The two sequences that branched basal to the two SAV3 clades contain a variant from each clade, indicating that this may have been the ancestral genotype of SAV3 in Norway before splitting into two clades. Moreover, while strains sampled in 2009-2010 fall in both clades, all of the sequences generated in this study (sampled from 2016-2019) fall into the SAV3b clade and are highly conserved. Interestingly, the absence of any SAV3a clade strains in the sequences generated in this study - even in samples from the heavily populated Hordaland region - suggests that this lineage may have gone extinct. While the estimated backbone of the phylogeny is Hordaland (Figure 3.1), there appears to be several distinct lineages that have been evolving separately for around 15 years (95% HPD, 12.51 to 17.7 years), which indicates that small-scale, local epidemics are co-circulating at the same time (Figure 3.2; indicated by red-coloured node). All but one of the strains from Sogn of Fjordane were monophyletic, indicating that an individual seeding event from Hordaland (independent of previous strains in this county) resulted in this outbreak. While only a single strain from Rogaland was sequenced in this study, it branched internal to many Hordaland strains, again indicating that the ancestor of this strain originated in Hordaland and was recently introduced to Rogaland.

3.3.4 Characterisation of structural deletions in natural SAV infections

Natural SAV3 infections have previously been shown to possess numerous defective genomes characterized by deletion variants (Petterson et al., 2013). To further explore this finding, all samples sequenced on the MinION platform were screened for structural variants. The size of deletions detected ranged widely between and within samples, from 11bp (FR16934408 and FR14304869) to 378bp (FR14696209). In all but one of the samples (FR14700631) deletions were found, again with a wide range of prevalence (between 1 and 40 deletions per sample). While deletions were found in all genes of the SAV genome, they were not evenly distributed

across genes, with relatively fewer deletions detected in the capsid and E3 genes (Figure 3.3). Several strains from different fish contained the deletions in the same or similar loci (Figure 3.3). While some closely related strains contained similar deletion variants, which may have been transmitted from one fish to another in the event of a viral genome containing an in-frame deletion being packaged into an infective virion (See Section 3.4), there was little phylogenetic signal in the overall distribution of many apparently similar deletions across the tested samples. Additionally, the estimated frequency of commonly observed deletions varied widely across samples, including closely related strains (Figure 3.3).

3.4 Discussion

In this study, 24 near-complete genomes were generated from SAV3-positive samples, more than doubling the publicly available genome sequences of Norwegian SAV. I show that SAV3, similar to other alphaviruses (Tan et al., 2018; Ling et al., 2019), evolves relatively slowly in comparison to many RNA viruses with an estimated evolutionary rate of 7.351 x 10^{-5} substitutions per site per year. This substitution rate is similar to previous whole genome estimates but slower than reported for shorter fragments of the genome (e.g. E2) (Karlsen et al., 2014). However sufficient genetic variation exists within Norwegian SAV3 strains to make genome sequences informative for fine-scale reconstructions of SAV evolution and phylogeographic patterns. Additionally, the temporal signal of SAV3 determined by a TempEst analysis showed strong clock-like signal, and thus further supports the use of genomic sequences in epidemiological studies. The contemporary SAV3 sequences sampled in 2016-2019 were distributed into five distinct clades that were estimated to have had a common ancestor 15 years ago, likely in Hordaland (Figure 3.1 Figure 3.2). These five clades were all represented in a relatively limited geographical area of Western Norway during the same timeframe. The clades did not show any clear, long-term geographical pattern within Western Norway, and sequences from Sogn og Fjordane and Rogaland all shared a common ancestor with sequences from Hordaland, which was less than 8 years old, again likely in Hordaland (Figure 3.1l indicted by the green node). This is much more recent than the first reports of SAV3 in these counties and suggests repeated reintroductions to these areas; my phylogeographic analysis suggested that a Hordaland reservoir was the more likely source. Furthermore, I could not find evidence for long-term (>10 years) persistence of local SAV3 reservoirs in Sogn og Fjordane and Rogaland, suggesting that local epidemics (ie. outbreaks of related stains that have persisted in a geographic region for longer than any individual cohort of salmon is out to sea) eventually burn out in these areas. While the presence of defective viruses may affect the infection dynamics of individual fish due to strong antiviral responses (see below), the accumulation of defective viral genomes is unlike to affect epidemic dynamics. A more likely explanation to these apparent local epidemic burn outs is

that the density of hosts in 'sink' regions (i.e. Rogaland and Sogn og Fjordane) is too low to support long term SAV3 epidemics. However it is possible that older SAV3 strains still persist in these sink regions, though no evidence of this was detected in this study.

Considering the presence of sequence diversity within each host, I considered the consensus sequence (i.e. the most abundant viral strain in a sample) to be informative for epidemiological analysis. This data suggests that Hordaland sources are seemingly seeding introductions of SAV3 strains to surrounding counties where pathogen persistence is shorter than in Hordaland (i.e. Rogaland and Sogn og Fjordane). This dynamic is compatible with a source-sink model (Tan et al., 2018) where viral lineages move from an area which can support a sustained epidemic (the source) to regions that cannot support the epidemic indefinitely (the sink). This is particularly apparent when comparing strains sampled in 2010 and those sampled between 2016 and 2019. In 2010, there were two distinct co-circulating clades of SAV3 (Figure 3.1). However samples from 2016-2019 fell into only one of these clades, and while this data cannot confirm the extinction of SAV3a, the absence of any recent SAV3a strains supports a source-sink model. These two SAV3 clades have been reported before (Karlsen et al., 2006; Jansen et al., 2010), however the timing of the split between the two clades had not yet been estimated. It is somewhat unsurprising to find Hordaland as an occasional source of virus, since the county has the highest density of seawater sites and the highest number of reported SAV3 cases per year (Norwegian Veterinary Institute, 2018). The lack of long-term established SAV local reservoirs in Sogn og Fjordane and Rogaland is however more noteworthy. A possible explanation to this pattern could be that a significant number of SAV3 transmissions across large geographic distances are not a result of passive transport with water currents, but rather of anthropogenic origin. However it is still likely that passive transport plays a role in short-term local outbreaks.

An interesting finding was the presence of multiple presumably defective viral RNA sequences carrying numerous deletions across the genome (Figure 3.3). All samples apart from one contained deletions ranging from 1-40 deletions per isolate, across a range of lengths from 11bp to 378bp. Deletion mutations have previously been described in Venezuelan equine encephalitis, a closely related member of the alphavirus genus (Forrester et al., 2011), and in SAV (Petterson et al., 2013, 2016), with the latter showing the presence of the same or similar deletions in multiple samples. The non-random distribution of deletions across the SAV genome, seen clearly in this analysis (Figure 3.3), suggests that template switching during RNA synthesis is a plausible source of origin, with factors such as secondary RNA structure, sequence identity and the kinetics of transcription influencing template switching (Baird et al., 2006; Simon-Loriere and Holmes, 2011). While it is unlikely that viral genomes missing large sections of either the structural or non-structural polyprotein are viable, it is impossible to rule

that all of the deletions observed result in a defective viral particle, considering that approximately 34% of the deletions did not cause a disruption to the protein coding sequence (in-frame deletions) (Table 3.4). Additionally, in-frame deletions located in the non-structural polyprotein may only affect the replication of the virus, not the packaging of the virus particles. As SAV has been shown to recombine in vivo to rescue viral genomes with otherwise fatal mutation errors (Petterson et al., 2016), it is possible that defective RNA molecules may be packaged into viral particles in the eventuality that more than one particle infects a single cell. Additionally, very similar deletions were found in closely related strains (Figure 3.3), which is compatible with deleted genomes being transmitted between infected fish. However, equally similar deletions are found across relatively distant strains which strongly suggests an independent origin of at least the majority of the deletion mutations.

Finally, as several of the samples used in this study were from regions that have had outbreaks of both SAV2 and SAV3 in the past, this sequencing method was tested on the sensitivity and accuracy of detecting both subtypes in the same sample. Other work in this Thesis (see Chapter 4) has shown that multiple SAV subtypes are commonly co-circulating on the same farm, and occasionally are found as co-infections within individual fish (Gallagher et al., 2020), and industry have anecdotally reported that co-infections of SAV2 and SAV3 are not uncommon. At an error rate of less than 0.3%, these results show that sequencing ~2kb amplicons and sequencing on the MinION platform enables highly accurate detection of such co-infections, even when the second subtype is at a relatively low titre (as low as 5% of the total SAV reads). However these data are based on *in-silico* co-infections as none of the samples sequenced in this study showed natural co-infections of SAV2 and SAV3, and future work should include samples that have been mixed in the lab prior to sequencing, as well as natural co-infections when identified.

In conclusion, whole genome analyses has helped increase our knowledge of the genetic diversity found in SAV infections impacting farmed salmon, and can thus be used to understand transmission pathways, viral population dynamics and the potential role that wild reservoirs play in ongoing PD epidemics. The apparent repeated seeding of SAV3 from ‘source’ counties like Hordaland to surrounding ‘sink’ counties implies that effective mitigating strategies might be able to limit the PD epidemic in ‘sink’ regions with improved biosecurity approaches. However more work is required to understand the relative impact that passive transmission (i.e. water currents) has on viral spread compared to the transportation of viral material via infected fish, especially on different geographical and temporal scales.

Data Availability

Raw sequence files are available under SRA BioProject PRJNA599578. Genome sequences are available in Genbank under the accession numbers: MN906915- MN906938

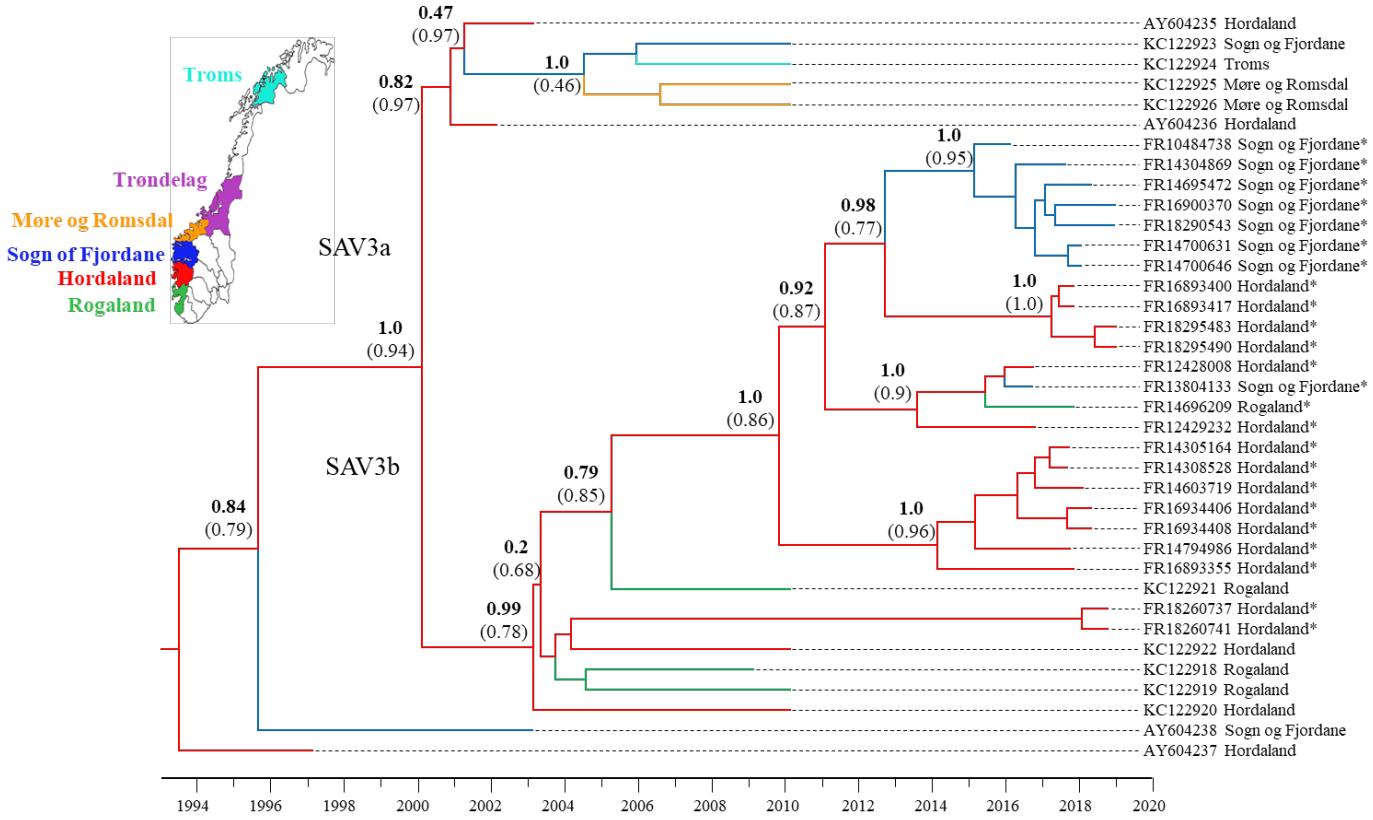

Figure 3.1. Bayesian phylogeny of the 24 SAV3 genomes generated in this study along with all publicly available SAV3 genome sequences from NCBI. The tree was built from an 11,681bp alignment and analysed in BEAST2 using the best fit nucleotide substitution model (TIM2+G4), a relaxed molecular clock model, tip-dating and a coalescent Bayesian Skyline population model. A discrete phylogeographical analysis was performed using ancestral reconstruction with branch colours indicating the estimated geographic location of each node. Statistical support for key nodes is indicated by posterior probability values in bold, and the ancestral location probability in brackets.

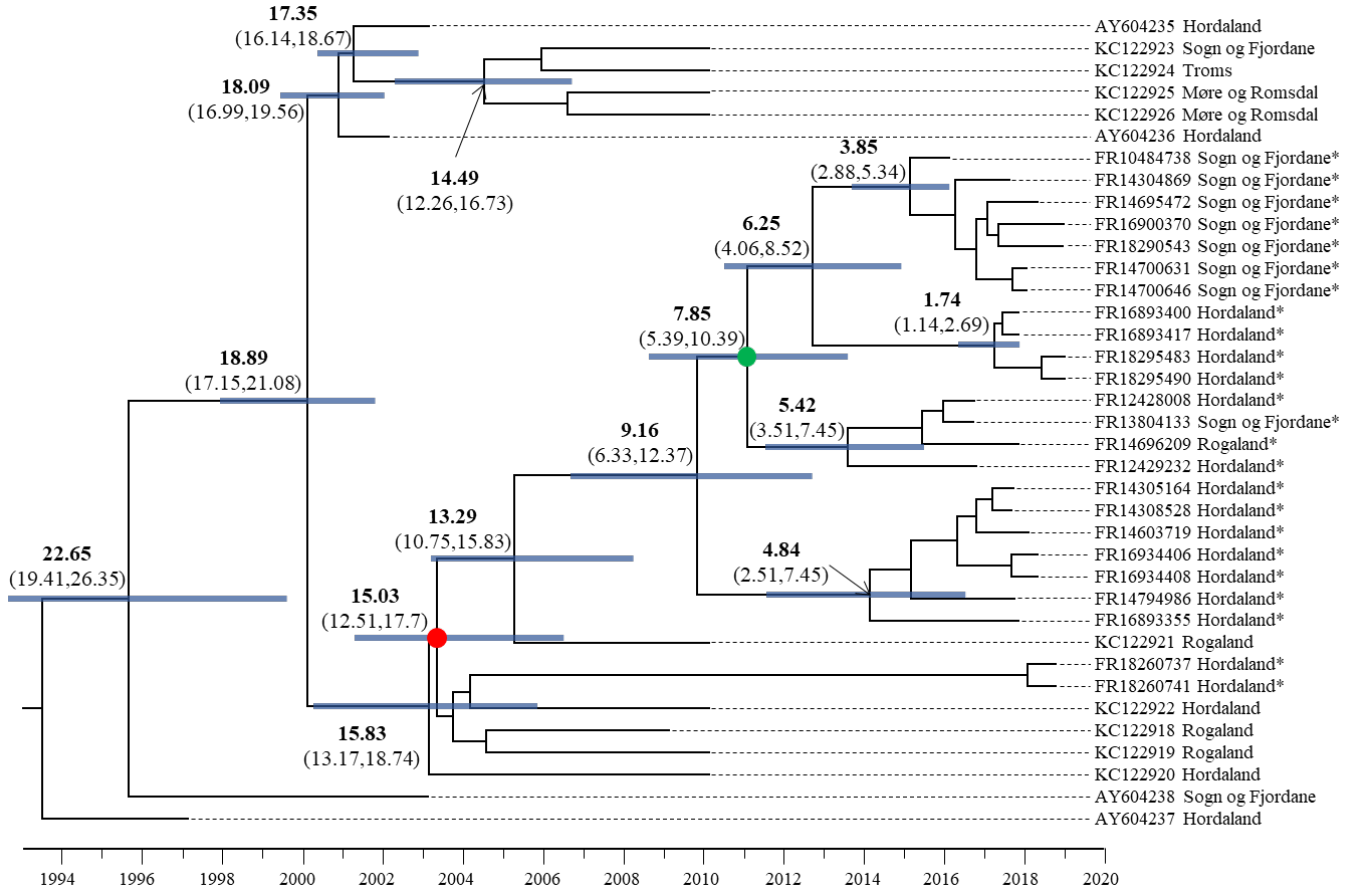

Figure 3.2. Time-calibrated Bayesian phylogenetic tree of SAV3 built from an 11,681 alignment and analysed in BEAST2 using the best fit nucleotide substitution model (TIM2+G4), a relaxed molecular clock model, tip-dating and a coalescent Bayesian Skyline population model. The values on branches indicate years before 2019 in bold, and the 95% highest posterior density (HPD) values in brackets. Node bars represent 95% HPD age range.

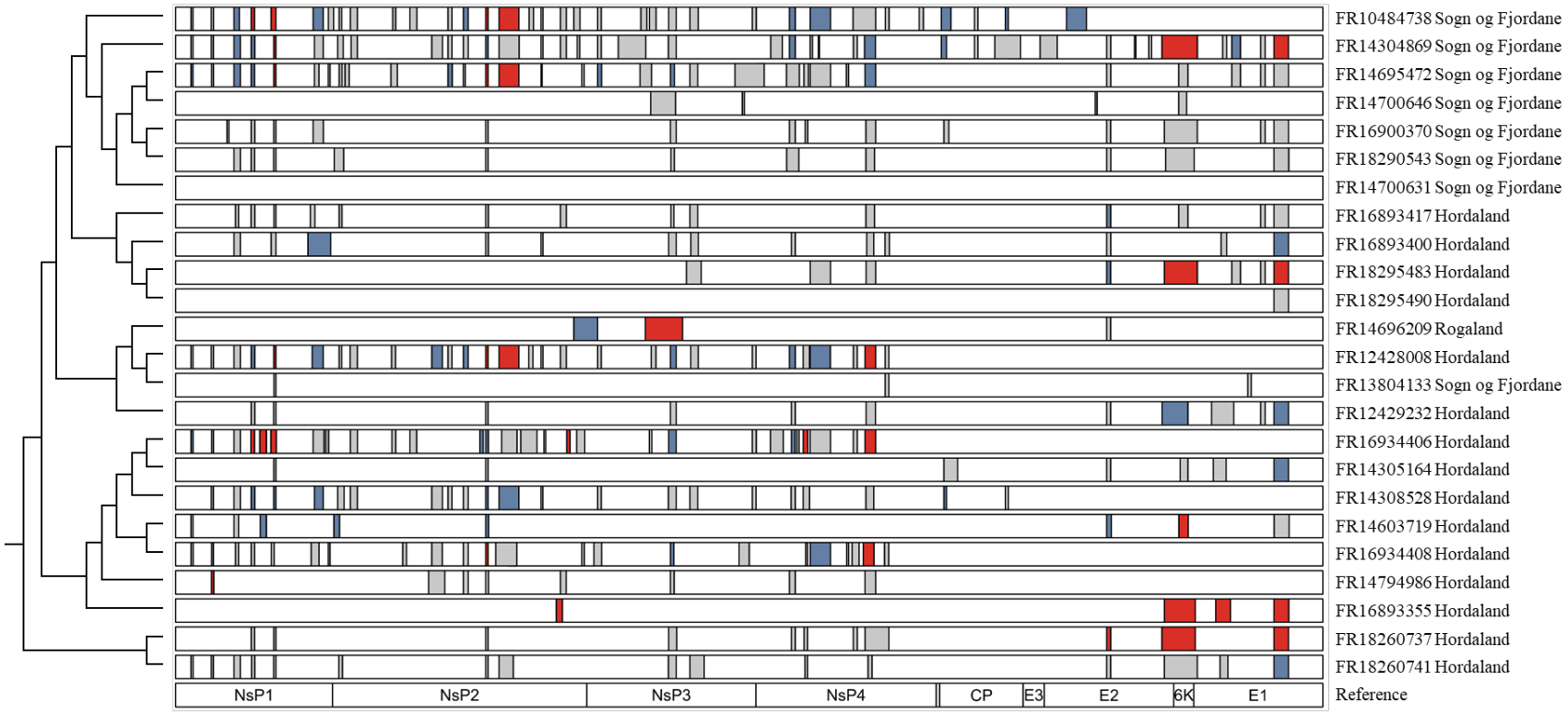

Figure 3.3. Distribution of deletions (≥ 10bp) throughout the SAV3 genome of isolates sequenced on the MinION platform. Only deletions with >50 supporting reads were considered, and all deletions were manually inspected to reduce the rate of false-positive calls. Bars indicate regions with a deletion and are coloured by estimated frequency. Isolates are plotted according to phylogenetic relationships shown elsewhere in this study (Figure 3.1) and the genomic position of each gene is used as a reference. Full details of each deletion can be found in Table 3.4

Table 3.1. Details of the SAV-infected tissue samples used in this study. Average coverage represents sequencing depth across genome

Sample ID	Accession	Location	Day	Month	Year	Species	Average Coverage
FR14696209	MN906920	Rogaland	8	3	2018	A. salmon	5,730
FR12429232	MN906917	Hordaland	20	2	2017	A. salmon	12,434
FR12428008	MN906916	Hordaland	30	1	2017	A. salmon	9,668
FR13804133	MN906918	Sogn of Fjordane	16	1	2017	A. salmon	12,331
FR14304869	MN906919	Sogn of Fjordane	13	12	2017	A. salmon	9,491
FR10484738	MN906915	Sogn of Fjordane	14	6	2016	A. salmon	8,483
FR14308528	MN906921	Hordaland	2	1	2018	A. salmon	8,073
FR14305164	MN906922	Hordaland	24	1	2018	A. salmon	4,713
FR16893400	MN906923	Hordaland	13	3	2018	A. salmon	6,510
FR16893417	MN906924	Hordaland	13	3	2018	A. salmon	10,273
FR16893355	MN906925	Hordaland	13	3	2018	R. trout	5,184
FR14700631	MN906926	Sogn og Fjordane	22	5	2018	A. salmon	4,783
FR14700646	MN906927	Sogn og Fjordane	22	5	2018	A. salmon	7,911
FR14603719	MN906928	Hordaland	7	6	2018	A. salmon	5,340
FR14695472	MN906929	Sogn og Fjordane	31	8	2018	A. salmon	7,314
FR16934408	MN906930	Hordaland	27	8	2018	A. salmon	2,829
FR16934406	MN906931	Hordaland	27	8	2018	A. salmon	4,506
FR18260741	MN906932	Hordaland	13	2	2019	A. salmon	6,053
FR18260737	MN906933	Hordaland	13	2	2019	A. salmon	3,455
FR18290543	MN906934	Sogn og Fjordane	15	4	2019	A. salmon	1,089
FR16900370	MN906935	Sogn og Fjordane	29	4	2019	A. salmon	7,555
FR18295490	MN906936	Hordaland	30	4	2019	A. salmon	5,396
FR14794986	MN906937	Hordaland	31	1	2018	A. salmon	1,561
FR18295483	MN906938	Hordaland	30	4	2019	A. salmon	4,323

Table 3.2. Details of PCR primers used to amplify SAV in overlapping amplicons

Primer Name	Primer Sequence 5' - 3'	Amplicon Length
Amplicon 1	AGACTGCGTTTCCAGGGTTC	2156
	CCCGTAGATGCCAATCGTGT	
Amplicon 2	GAATACGTTTACGAATTGTCCTCC	1966
	ACCGAGACGGACTTGAAATACC	
Amplicon 3	GACCTGGTGTTTTGTGACGC	2438
	TCCCGTGTTAGCCCTCTAGG	
Amplicon 4	GCAGCGTCCACRGCCATAGT	2014
	CATCAGGCGTTTTACAGGGTC	
Amplicon 5	TTGTGGCGGCTTCCTGTTAC	2110
	GTAAACGTCTGGGAGTCGCTG	
Amplicon 6	AGAGAACGCAGCAAGGGC	2402
	GGCACTTCTTCACCACGCA	

Table 3.3. Additional SAV3 genome sequences used in phylogenetic analysis

Isolate	Location	Sampling Date
AY604235	Hordaland	2003
AY604236	Hordaland	2002
AY604237	Hordaland	1997
AY604238	Sogn og Fjordane	2003
KC122918	Rogaland	2009
KC122919	Rogaland	2010
KC122920	Hordaland	2010
KC122921	Rogaland	2010
KC122922	Hordaland	2010
KC122923	Sogn og Fjordane	2010
KC122924	Troms	2010
KC122925	Møre og Romsdal	2010
KC122926	Møre og Romsdal	2010

Table 3.4. Summary of deletions characterised in 24 naturally infected SAV3 samples from Norway.

Sample_ID	start	end	Length	Effect	Gene
SAV3_BC01	363	383	20	Frameshift	nsP1
SAV3_BC01	592	657	65	Frameshift	nsP1
SAV3_BC01	768	803	35	Frameshift	nsP1
SAV3_BC01	996	1018	22	Frameshift	nsP1
SAV3_BC01	1409	1498	89	Frameshift	nsP1
SAV3_BC01	1644	1709	65	Frameshift	nsP2
SAV3_BC01	1773	1844	71	Frameshift	nsP2
SAV3_BC01	2599	2706	107	Frameshift	nsP2
SAV3_BC01	2764	2804	40	Frameshift	nsP2
SAV3_BC01	2919	2970	51	In-frame	nsP2
SAV3_BC01	3147	3171	24	In-frame	nsP2
SAV3_BC01	3281	3482	201	In-frame	nsP2
SAV3_BC01	3707	3726	19	Frameshift	nsP2
SAV3_BC01	4282	4318	36	In-frame	nsP3
SAV3_BC01	4999	5079	80	Frameshift	nsP3
SAV3_BC01	5217	5296	79	Frameshift	nsP3
SAV3_BC01	5850	5888	38	Frameshift	nsP3-nsP4
SAV3_BC01	6244	6286	42	In-frame	nsP4
SAV3_BC01	6366	6429	63	In-frame	nsP4
SAV3_BC01	7002	7083	81	In-frame	nsP4
SAV3_BC01	7793	7821	28	Frameshift	Cp
SAV3_BC01	8417	8447	30	In-frame	Cp
SAV3_BC02	997	1019	22	Frameshift	nsP1
SAV3_BC02	3147	3171	24	In-frame	nsP2
SAV3_BC02	7793	7931	138	In-frame	Cp
SAV3_BC02	9443	9484	41	Frameshift	E2
SAV3_BC02	10190	10269	79	Frameshift	6K
SAV3_BC02	10526	10656	130	Frameshift	E1
SAV3_BC02	11140	11288	148	Frameshift	E1
SAV3_BC03	594	657	63	In-frame	nsP1
SAV3_BC03	969	1018	49	Frameshift	nsP1
SAV3_BC03	1342	1574	232	Frameshift	nsP1
SAV3_BC03	3147	3172	25	Frameshift	nsP2
SAV3_BC03	3707	3727	20	Frameshift	nsP2
SAV3_BC03	5000	5079	79	Frameshift	nsP3
SAV3_BC03	5227	5302	75	In-frame	nsP3
SAV3_BC03	6246	6286	40	Frameshift	nsP4
SAV3_BC03	7010	7083	73	Frameshift	nsP4
SAV3_BC03	7196	7239	43	Frameshift	nsP4
SAV3_BC03	9443	9485	42	In-frame	E2
SAV3_BC03	10605	10661	56	Frameshift	E1
SAV3_BC03	11140	11289	149	Frameshift	E1
SAV3_BC04	608	639	31	Frameshift	nsP1
SAV3_BC04	767	803	36	In-frame	nsP1
SAV3_BC04	998	1018	20	Frameshift	nsP1
SAV3_BC04	1367	1414	47	Frameshift	nsP1
SAV3_BC04	1660	1688	28	Frameshift	nsP2
SAV3_BC04	3147	3172	25	Frameshift	nsP2
SAV3_BC04	3904	3964	60	In-frame	nsP2
SAV3_BC04	5024	5055	31	Frameshift	nsP3
SAV3_BC04	5217	5296	79	Frameshift	nsP3
SAV3_BC04	7002	7089	87	In-frame	nsP4
SAV3_BC04	9443	9485	42	In-frame	E2
SAV3_BC04	10175	10269	94	Frameshift	6K
SAV3_BC04	11006	11052	46	Frameshift	E1
SAV3_BC04	11140	11288	148	Frameshift	E1
SAV3_BC05	3863	3923	60	In-frame	nsP2

SAV3_BC05	10028	10343	315	In-frame	E2-6K
SAV3_BC05	10551	10700	149	Frameshift	E1
SAV3_BC05	11140	11291	151	Frameshift	E1
SAV3_BC07	4820	5070	250	Frameshift	nsP3
SAV3_BC07	5749	5771	22	Frameshift	nsP3
SAV3_BC07	9329	9347	18	In-frame	E2
SAV3_BC07	10175	10256	81	In-frame	6K
SAV3_BC09	157	176	19	Frameshift	nsP1
SAV3_BC09	593	640	47	Frameshift	nsP1
SAV3_BC09	859	921	62	Frameshift	nsP1
SAV3_BC09	1606	1664	58	Frameshift	nsP2
SAV3_BC09	3147	3176	29	Frameshift	nsP2
SAV3_BC09	9443	9490	47	Frameshift	E2
SAV3_BC09	10176	10273	97	Frameshift	6K
SAV3_BC09	11143	11293	150	In-frame	E1
SAV3_BC11	157	176	19	Frameshift	nsP1
SAV3_BC11	366	384	18	In-frame	nsP1
SAV3_BC11	592	657	65	Frameshift	nsP1
SAV3_BC11	768	802	34	Frameshift	nsP1
SAV3_BC11	997	1018	21	In-frame	nsP1
SAV3_BC11	1409	1454	45	In-frame	nsP1
SAV3_BC11	1549	1567	18	In-frame	nsP1
SAV3_BC11	1660	1689	29	Frameshift	nsP2
SAV3_BC11	1719	1761	42	In-frame	nsP2
SAV3_BC11	2185	2250	65	Frameshift	nsP2
SAV3_BC11	2764	2806	42	In-frame	nsP2
SAV3_BC11	2919	2938	19	Frameshift	nsP2
SAV3_BC11	3147	3171	24	In-frame	nsP2
SAV3_BC11	3281	3482	201	In-frame	nsP2
SAV3_BC11	3707	3720	13	Frameshift	nsP2
SAV3_BC11	4120	4144	24	In-frame	nsP2
SAV3_BC11	4282	4320	38	Frameshift	nsP3
SAV3_BC11	4710	4827	117	In-frame	nsP3
SAV3_BC11	5020	5060	40	Frameshift	nsP3
SAV3_BC11	5217	5296	79	Frameshift	nsP3
SAV3_BC11	5674	5972	298	Frameshift	nsP3
SAV3_BC11	6197	6327	130	Frameshift	nsP4
SAV3_BC11	6374	6418	44	Frameshift	nsP4
SAV3_BC11	6438	6644	206	Frameshift	nsP4
SAV3_BC11	6805	6825	20	Frameshift	nsP4
SAV3_BC11	6993	7101	108	In-frame	nsP4
SAV3_BC11	9443	9485	42	In-frame	E2
SAV3_BC11	10175	10269	94	Frameshift	6K
SAV3_BC11	10713	10801	88	Frameshift	E1
SAV3_BC11	11006	11052	46	Frameshift	E1
SAV3_BC11	11140	11289	149	Frameshift	E1
SAV3_BC12	158	178	20	Frameshift	nsP1
SAV3_BC12	366	383	17	Frameshift	nsP1
SAV3_BC12	608	639	31	Frameshift	nsP1
SAV3_BC12	768	802	34	Frameshift	nsP1
SAV3_BC12	970	999	29	Frameshift	nsP1
SAV3_BC12	1376	1455	79	Frameshift	nsP1
SAV3_BC12	1549	1566	17	Frameshift	nsP1
SAV3_BC12	2304	2343	39	In-frame	nsP2
SAV3_BC12	2599	2705	106	Frameshift	nsP2
SAV3_BC12	2919	2970	51	In-frame	nsP2
SAV3_BC12	3147	3171	24	In-frame	nsP2
SAV3_BC12	3248	3462	214	Frameshift	nsP2
SAV3_BC12	4120	4146	26	Frameshift	nsP2
SAV3_BC12	4243	4320	77	Frameshift	nsP3

SAV3_BC12	5020	5055	35	Frameshift	nsP3
SAV3_BC12	5716	5819	103	Frameshift	nsP3
SAV3_BC12	6392	6410	18	In-frame	nsP4
SAV3_BC12	6438	6642	204	In-frame	nsP4
SAV3_BC12	6807	6826	19	Frameshift	nsP4
SAV3_BC12	6861	6933	72	In-frame	nsP4
SAV3_BC12	6975	7084	109	Frameshift	nsP4
SAV3_BC12	7191	7233	42	In-frame	nsP4
SAV3_BC13	157	177	20	Frameshift	nsP1
SAV3_BC13	367	384	17	Frameshift	nsP1
SAV3_BC13	594	657	63	In-frame	nsP1
SAV3_BC13	765	802	37	Frameshift	nsP1
SAV3_BC13	855	921	66	In-frame	nsP1
SAV3_BC13	968	1022	54	In-frame	nsP1
SAV3_BC13	1395	1506	111	In-frame	nsP1
SAV3_BC13	1524	1553	29	Frameshift	nsP1
SAV3_BC13	1773	1845	72	In-frame	nsP2
SAV3_BC13	2194	2233	39	In-frame	nsP2
SAV3_BC13	2376	2445	69	In-frame	nsP2
SAV3_BC13	3086	3120	34	Frameshift	nsP2
SAV3_BC13	3147	3172	25	Frameshift	nsP2
SAV3_BC13	3308	3463	155	Frameshift	nsP2
SAV3_BC13	3502	3665	163	Frameshift	nsP2
SAV3_BC13	3737	3755	18	In-frame	nsP2
SAV3_BC13	3969	4001	32	Frameshift	nsP2
SAV3_BC13	4069	4150	81	In-frame	nsP2
SAV3_BC13	4805	4831	26	Frameshift	nsP3
SAV3_BC13	5000	5079	79	Frameshift	nsP3
SAV3_BC13	5850	5890	40	Frameshift	nsP3-nsP4
SAV3_BC13	6033	6163	130	Frameshift	nsP4
SAV3_BC13	6246	6282	36	In-frame	nsP4
SAV3_BC13	6301	6323	22	Frameshift	nsP4
SAV3_BC13	6366	6413	47	Frameshift	nsP4
SAV3_BC13	6436	6642	206	Frameshift	nsP4
SAV3_BC13	6875	6915	40	Frameshift	nsP4
SAV3_BC13	6993	7101	108	In-frame	nsP4
SAV3_BC14	157	177	20	Frameshift	nsP1
SAV3_BC14	363	383	20	Frameshift	nsP1
SAV3_BC14	594	656	62	Frameshift	nsP1
SAV3_BC14	772	803	31	Frameshift	nsP1
SAV3_BC14	996	1018	22	Frameshift	nsP1
SAV3_BC14	1656	1694	38	Frameshift	nsP2
SAV3_BC14	3147	3171	24	In-frame	nsP2
SAV3_BC14	3281	3426	145	Frameshift	nsP2
SAV3_BC14	4999	5079	80	Frameshift	nsP3
SAV3_BC14	5217	5361	144	In-frame	nsP3
SAV3_BC14	6385	6410	25	Frameshift	nsP4
SAV3_BC14	7025	7065	40	Frameshift	nsP4
SAV3_BC14	9443	9484	41	Frameshift	E2
SAV3_BC14	10028	10364	336	In-frame	E2-6K
SAV3_BC14	10594	10674	80	Frameshift	E1
SAV3_BC14	11140	11288	148	Frameshift	E1
SAV3_BC15	768	802	34	Frameshift	nsP1
SAV3_BC15	997	1019	22	Frameshift	nsP1
SAV3_BC15	3147	3171	24	In-frame	nsP2
SAV3_BC15	5002	5082	80	Frameshift	nsP3
SAV3_BC15	6246	6286	40	Frameshift	nsP4
SAV3_BC15	6375	6410	35	Frameshift	nsP4
SAV3_BC15	6875	6915	40	Frameshift	nsP4
SAV3_BC15	6994	7236	242	Frameshift	nsP4

SAV3_BC15	9443	9485	42	In-frame	E2
SAV3_BC15	10004	10343	339	In-frame	E2-6K
SAV3_BC15	11140	11289	149	Frameshift	E1
SAV3_BC16	592	657	65	Frameshift	nsP1
SAV3_BC16	768	802	34	Frameshift	nsP1
SAV3_BC16	996	1018	22	Frameshift	nsP1
SAV3_BC16	1611	1706	95	Frameshift	nsP2
SAV3_BC16	3147	3171	24	In-frame	nsP2
SAV3_BC16	5024	5058	34	Frameshift	nsP3
SAV3_BC16	6198	6322	124	Frameshift	nsP4
SAV3_BC16	7002	7089	87	In-frame	nsP4
SAV3_BC16	9444	9487	43	Frameshift	E2
SAV3_BC16	10043	10331	288	In-frame	E2-6K
SAV3_BC16	11141	11291	150	In-frame	E1
SAV3_BC17	521	541	20	Frameshift	nsP1
SAV3_BC17	768	803	35	Frameshift	nsP1
SAV3_BC17	997	1018	21	In-frame	nsP1
SAV3_BC17	1395	1499	104	Frameshift	nsP1
SAV3_BC17	3147	3171	24	In-frame	nsP2
SAV3_BC17	5020	5079	59	Frameshift	nsP3
SAV3_BC17	6227	6286	59	Frameshift	nsP4
SAV3_BC17	6387	6412	25	Frameshift	nsP4
SAV3_BC17	7002	7101	99	In-frame	nsP4
SAV3_BC17	7793	7844	51	In-frame	Cp
SAV3_BC17	9443	9484	41	Frameshift	E2
SAV3_BC17	10028	10364	336	In-frame	E2-6K
SAV3_BC17	11006	11052	46	Frameshift	E1
SAV3_BC17	11140	11289	149	Frameshift	E1
SAV3_BC18	11140	11289	149	Frameshift	E1
SAV3_BC20	365	389	24	In-frame	nsP1
SAV3_BC20	2567	2730	163	Frameshift	nsP2
SAV3_BC20	2919	2970	51	In-frame	nsP2
SAV3_BC20	3147	3173	26	Frameshift	nsP2
SAV3_BC20	3904	3963	59	Frameshift	nsP2
SAV3_BC20	5020	5057	37	Frameshift	nsP3
SAV3_BC20	6227	6286	59	Frameshift	nsP4
SAV3_BC20	6993	7101	108	In-frame	nsP4
SAV3_BC21	5182	5332	150	In-frame	nsP3
SAV3_BC21	6438	6642	204	In-frame	nsP4
SAV3_BC21	7004	7101	97	Frameshift	nsP4
SAV3_BC21	9443	9484	41	Frameshift	E2
SAV3_BC21	10028	10364	336	In-frame	E2-6K
SAV3_BC21	10713	10801	88	Frameshift	E1
SAV3_BC21	11006	11052	46	Frameshift	E1
SAV3_BC21	11140	11289	149	Frameshift	E1
SAV3_BC22	4040	4281	241	Frameshift	nsP2
SAV3_BC22	4764	5142	378	In-frame	nsP3
SAV3_BC22	9443	9484	41	Frameshift	E2
SAV3_BC23	768	803	35	Frameshift	nsP1
SAV3_BC23	996	1018	22	Frameshift	nsP1
SAV3_BC23	3147	3171	24	In-frame	nsP2
SAV3_BC23	5020	5079	59	Frameshift	nsP3
SAV3_BC23	6246	6286	40	Frameshift	nsP4
SAV3_BC23	7004	7099	95	Frameshift	nsP4
SAV3_BC23	9443	9484	41	Frameshift	E2
SAV3_BC23	10007	10269	262	Frameshift	E2-6K
SAV3_BC23	10509	10733	224	Frameshift	E1
SAV3_BC23	11006	11052	46	Frameshift	E1
SAV3_BC23	11140	11289	149	Frameshift	E1
SAV3_BC24	157	176	19	Frameshift	nsP1

SAV3_BC24	363	384	21	In-frame	nsP1
SAV3_BC24	592	657	65	Frameshift	nsP1
SAV3_BC24	768	803	35	Frameshift	nsP1
SAV3_BC24	997	1018	21	In-frame	nsP1
SAV3_BC24	1387	1498	111	In-frame	nsP1
SAV3_BC24	1660	1686	26	Frameshift	nsP2
SAV3_BC24	1771	1844	73	Frameshift	nsP2
SAV3_BC24	2191	2232	41	Frameshift	nsP2
SAV3_BC24	2599	2706	107	Frameshift	nsP2
SAV3_BC24	2764	2804	40	Frameshift	nsP2
SAV3_BC24	2919	2970	51	In-frame	nsP2
SAV3_BC24	3147	3171	24	In-frame	nsP2
SAV3_BC24	3281	3482	201	In-frame	nsP2
SAV3_BC24	3582	3625	43	Frameshift	nsP2
SAV3_BC24	3707	3727	20	Frameshift	nsP2
SAV3_BC24	3904	3964	60	In-frame	nsP2
SAV3_BC24	4282	4318	36	In-frame	nsP3
SAV3_BC24	4825	4873	48	In-frame	nsP3
SAV3_BC24	5020	5079	59	Frameshift	nsP3
SAV3_BC24	5225	5301	76	Frameshift	nsP3
SAV3_BC24	5850	5888	38	Frameshift	nsP3-nsP4
SAV3_BC24	6227	6286	59	Frameshift	nsP4
SAV3_BC24	6366	6429	63	In-frame	nsP4
SAV3_BC24	6438	6641	203	Frameshift	nsP4
SAV3_BC24	6876	6914	38	Frameshift	nsP4
SAV3_BC24	6993	7101	108	In-frame	nsP4
SAV3_BC24	7197	7235	38	Frameshift	nsP4
SAV3_BC25	998	1018	20	Frameshift	nsP1
SAV3_BC25	7197	7236	39	In-frame	nsP4
SAV3_BC25	10876	10911	35	Frameshift	E1
SAV3_BC26	157	176	19	Frameshift	nsP1
SAV3_BC26	363	383	20	Frameshift	nsP1
SAV3_BC26	592	657	65	Frameshift	nsP1
SAV3_BC26	768	803	35	Frameshift	nsP1
SAV3_BC26	997	1018	21	In-frame	nsP1
SAV3_BC26	1409	1498	89	Frameshift	nsP1
SAV3_BC26	1644	1701	57	In-frame	nsP2
SAV3_BC26	1781	1844	63	In-frame	nsP2
SAV3_BC26	2599	2707	108	In-frame	nsP2
SAV3_BC26	2764	2804	40	Frameshift	nsP2
SAV3_BC26	2919	2970	51	In-frame	nsP2
SAV3_BC26	3147	3171	24	In-frame	nsP2
SAV3_BC26	3281	3485	204	In-frame	nsP2
SAV3_BC26	3707	3727	20	Frameshift	nsP2
SAV3_BC26	3904	3964	60	In-frame	nsP2
SAV3_BC26	4073	4099	26	Frameshift	nsP2
SAV3_BC26	4282	4318	36	In-frame	nsP3
SAV3_BC26	4491	4769	278	Frameshift	nsP3
SAV3_BC26	4999	5079	80	Frameshift	nsP3
SAV3_BC26	6037	6154	117	In-frame	nsP4
SAV3_BC26	6227	6286	59	Frameshift	nsP4
SAV3_BC26	6436	6459	23	Frameshift	nsP4
SAV3_BC26	6521	6532	11	Frameshift	nsP4
SAV3_BC26	6876	6914	38	Frameshift	nsP4
SAV3_BC26	6989	7099	110	Frameshift	nsP4
SAV3_BC26	7766	7821	55	Frameshift	Cp
SAV3_BC26	8104	8138	34	Frameshift	Cp
SAV3_BC26	8309	8571	262	Frameshift	Cp
SAV3_BC26	8769	8943	174	In-frame	E3-E2
SAV3_BC26	9443	9485	42	In-frame	E2

SAV3_BC26	9726	9741	15	In-frame	E2
SAV3_BC26	9873	9900	27	In-frame	E2
SAV3_BC26	10004	10364	360	In-frame	E2-6K
SAV3_BC26	10620	10661	41	Frameshift	E1
SAV3_BC26	10713	10801	88	Frameshift	E1
SAV3_BC26	11006	11052	46	Frameshift	E1
SAV3_BC26	11140	11288	148	Frameshift	E1
SAV3_BC27	158	176	18	In-frame	nsP1
SAV3_BC27	363	384	21	In-frame	nsP1
SAV3_BC27	592	648	56	Frameshift	nsP1
SAV3_BC27	768	802	34	Frameshift	nsP1
SAV3_BC27	969	1018	49	Frameshift	nsP1
SAV3_BC27	1395	1498	103	Frameshift	nsP1
SAV3_BC27	1550	1603	53	Frameshift	nsP1-nsP2
SAV3_BC27	1660	1686	26	Frameshift	nsP2
SAV3_BC27	1773	1844	71	Frameshift	nsP2
SAV3_BC27	2203	2233	30	In-frame	nsP2
SAV3_BC27	2376	2446	70	Frameshift	nsP2
SAV3_BC27	2764	2804	40	Frameshift	nsP2
SAV3_BC27	2919	2970	51	In-frame	nsP2
SAV3_BC27	3147	3171	24	In-frame	nsP2
SAV3_BC27	3281	3482	201	In-frame	nsP2
SAV3_BC27	3586	3631	45	In-frame	nsP2
SAV3_BC27	3707	3727	20	Frameshift	nsP2
SAV3_BC27	3904	3964	60	In-frame	nsP2
SAV3_BC27	4040	4101	61	Frameshift	nsP2
SAV3_BC27	4282	4318	36	In-frame	nsP3
SAV3_BC27	4720	4780	60	In-frame	nsP3
SAV3_BC27	4807	4874	67	Frameshift	nsP3
SAV3_BC27	5001	5079	78	In-frame	nsP3
SAV3_BC27	5217	5301	84	In-frame	nsP3
SAV3_BC27	5850	5890	40	Frameshift	nsP3-nsP4
SAV3_BC27	6227	6286	59	Frameshift	nsP4
SAV3_BC27	6438	6643	205	Frameshift	nsP4
SAV3_BC27	6872	7101	229	Frameshift	nsP4
SAV3_BC27	7200	7237	37	Frameshift	nsP4
SAV3_BC27	7542	7586	44	Frameshift	nsP4
SAV3_BC27	7766	7865	99	In-frame	Cp
SAV3_BC27	8104	8138	34	Frameshift	Cp
SAV3_BC27	8417	8446	29	Frameshift	Cp
SAV3_BC27	9037	9240	203	Frameshift	E2

Chapter 4. Genome-wide target enriched viral sequencing reveals extensive 'hidden' salmonid alphavirus diversity in farmed and wild fish populations.

The data presented in this chapter was published as **Gallagher, M.D**., Matejusova, I., Ruane, N.M., Macqueen, D.J. *Genome-wide target enriched viral sequencing reveals extensive 'hidden' salmonid alphavirus diversity in farmed and wild fish populations*. Aquaculture 522, (2020).

https://doi.org/10.1016/ j.aquaculture.2020.735117

Summary

In this Chapter, I investigate the use of targeted sequence capture to characterise intra-host SAV genetic diversity in naturally infected fish. Pooling tissues from multiple infected animals is a standard method of sampling for molecular diagnostics of pathogens in aquaculture; however the impacts of pooling on detection of viral diversity remain poorly understood. Therefore I included both pooled and individual fish samples in this study from farmed Atlantic salmon and rainbow trout, in addition to two wild flatfish species, sampled from multiple regions in Scottish and Irish waters. Mixed subtype infections were present in three of the four species studied, and in both farmed and wild samples. This involved pairs of SAV subtypes known to previously exist in the sampled geographical locations. Evidence of subtype-level SAV co-infections were also shown in individual fish (i.e. not pooled), including wild fish such as dab. My findings confirm the circulation of multiple SAV subtypes on the same fish farm and abundant within-subtype genetic diversity in all studied samples.

4.1 Introduction

The rapid evolutionary rate of RNA viruses leads to high levels of genetic diversity and the potential for co-existence of multiple strains in host populations, including within single hosts (Duffy et al., 2008; Sanjuán et al., 2010). Such diversity poses challenges to both human health and agricultural systems, as the effectiveness of disease control relies on knowledge of both viral diversity and evolutionary dynamics (García-Arenal and McDonald, 2003; Grenfell et al., 2004; Acosta-Leal et al., 2011). Aquaculture, as the fastest growing food production industry (FAO, 2016), plays an increasingly important role in global food and economic security (Jennings et al., 2016). However, viral disease remains a major threat to the

sustainability and expansion of this sector, due to a lack of effective therapeutics and vaccines (Garver et al., 2005; Karlsen et al., 2012; Munang'andu et al., 2012) and limited understanding of disease transmission between farmed populations and wild reservoir fish (Snow et al., 2010; Bruno et al., 2014; Ruane et al., 2018).

Molecular characterization of pathogens plays an important role during investigations of viral disease outbreaks on fish farms, helping to understand the transmission of pathogens between farms, and contributing to improvement of disease control measures to further limit pathogen transmission. Sanger sequencing of one (up to a) few candidate/marker genes is often applied to characterize the disease agent (Nishizawa et al., 2006; Matejusova et al., 2013; Holopainen et al., 2017) and this method is accurate and well suited to low-throughput applications aiming to reveal the dominant viral strain(s). Second generation high-throughput sequencing (e.g. Illumina, Roche 454, Ion Torrent etc.) is now well established for genome-wide investigations of animal viruses (Bodewes et al., 2013; Ferretti et al., 2018; Pfaff et al., 2019) and, like Sanger, provides highly accurate data. Recent studies have demonstrated the potential of such platforms to reveal intrahost diversity of fish viral pathogens, e.g. of viral hemorrhagic septicaemia virus (VHSV) (Schönherz et al., 2016) and Cyprinid herpesvirus 3 (Hammoumi et al., 2016). Third generation Nanopore sequencing has recently been used for rapid genome-wide analysis of fish RNA viruses (Gallagher et al., 2018), but this platform still suffers from higher error rates compared to Sanger and Illumina platforms, making finer-scale investigations of viral genetic diversity more challenging.

Understanding the genetic diversity of natural viral infections is essential, as different viral strains may be associated with unique pathological outcomes, demanding alternative control strategies. For example, while infectious salmon anaemia virus (ISAV) typically causes high mortality rates in Atlantic salmon, a specific ISAV strain (HPR0) is non-pathogenic (Nylund et al., 2007). However, as co-infections with both pathogenic and non-pathogenic ISAV strains have been reported (Kibenge et al., 2009; Kulshreshtha et al., 2010), and ISAV HPR0 has been shown to mutate into a pathogenic form (Christiansen et al., 2017), the ability to accurately capture all viral forms within each host is necessary for effective decisions on disease control. Similarly, different strains of VHSV have distinct outcomes for pathogenicity in salmonids (Skall et al., 2004; Dale et al., 2009) and the introduction of the exotic VHSV genotype IV into Europe could have devastating consequences (Lumsden et al., 2007). Thus the ability to detect co-infecting viral strains is central to viral epidemiological studies and the control of viral disease outbreaks.

In this Chapter, I used a capture-based approach to enrich the whole genome of target viruses and used it to characterize genetic diversity of salmonid alphavirus (SAV, *Togaviridae*), a

single-strand positive-sense RNA virus. This virus was recently added to the World Organisation for Animal Health ('OIE') list as a notifiable disease agent (OIE, 2019b). SAV causes pancreas disease (PD) in Atlantic salmon (*S. salar)* and sleeping disease (SD) in rainbow trout (*O. mykiss)*, resulting in significant mortality, reduced growth and poor flesh quality (Aunsmo et al., 2010). Six SAV subtypes (SAV1-6) are widely recognized (Fringuelli et al., 2008) that are loosely geographically structured across Europe, with Scotland reporting cases of SAV1, 2, 4 and 5, Ireland SAV1, 4 and 6 (Graham et al., 2012), and Norway SAV2 and 3 (Hodneland et al., 2005; Hjortaas et al., 2013). SAV has also been detected in wild species including flatfish (Snow et al., 2010; Bruno et al., 2014) and Ballan wrasse (*L. bergylta)* (Ruane et al., 2018). While one past study provided evidence for complex population structure in SAV3 (Petterson et al., 2013), including the presence of non-random deletion variants in natural infections, epidemiological studies of SAV have been limited to the subtype-level; omitting intrahost variation. Considering that different SAV subtypes are known to have unique pathogenicity (e.g. infections with SAV1 and SAV3 show the most pronounced histopathological changes) (Graham et al., 2011), and the recent addition of SAV to the OIE notifiable list, gaining a deeper understanding of SAV genetic diversity and population structure is currently extremely timely.

I thus performed a genome-wide analysis of SAV diversity within infected tissues from farmed salmonid and wild flatfish samples from several locations, representing both single hosts and pools of different fish. This data revealed extensive genetic diversity on several levels, including SAV subtype co-infections in single wild hosts, the presence of multiple SAV subtypes at the farm level, and extensive within-subtype SAV diversity in all samples. These findings have implications for sampling strategies of epidemiological and transmission studies and disease management where strain-level information is relevant.

4.2 Methods

4.2.1 Sample Preparation

Tissue homogenate or RNA from eighteen heart tissue samples from either individual fish or pools (n=5 fish) with natural SAV infections were obtained from Marine Scotland Science or Marine Institute Ireland (Table 4.1). Irish flatfish from Marine Institute Ireland were previously published in (McCleary et al., 2014). Total RNA was extracted using a phenol-chloroform method and RNA integrity was assessed by agarose gel electrophoresis. Single-strand (ss)-cDNA was synthesized using Superscript III reverse transcriptase (Invitrogen) and cleaned by AMPure XP bead purification (Beckman Coulter). Ss-cDNA was converted to double-strand (ds)-cDNA using NEBNext Ultra II Non-directional RNA Second Strand Synthesis Module (New England Biolabs) following the manufacturer's instructions. Ds-

cDNA concentration was determined using a Qubit system with a ds-DNA HS Assay kit (ThermoFisher). Relative viral load was estimated using qPCR (Table 4.2), employing a primer pair designed in a region of the SAV genome conserved across all subtypes: 5' - TGC CCG ACA GAG CAC CTT - 3' (sense) and 5' - CTC GGC GAC CTG GAA CTT GAT - 3' (antisense). 15 µl qPCR reactions were performed for each isolate including 5 ng ds-cDNA, 7.5 µl Brilliant III Ultra-fast SYBR Green (Agilent Technologies) and 500 nM sense/antisense primers. Cycling conditions were as follows: 1 cycle of 3 min at 95°C, followed by 40 cycles of 20 s at 64°C, finishing with 30 s at 55°C. The ds-cDNA samples were kept at -80ºC until library preparation for sequence capture.

4.2.2 Sequence capture probe design, library preparation and sequencing

Agilent SureSelect^{XT2} 120-mer RNA oligomer baits were generated at 4-fold tiling to cover reference genomes for SAV1 to SAV5 (where possible representing two representatives of the most phylogenetically distant clades) (accession numbers: SAV1: JX163854, AJ316244; SAV2: AJ316246, MH708652; SAV3: DQ149204, SAV4: MH708651; SAV5: MH708653, MH708650), as well as two fragments of SAV6 (EF675547, EF675499). Sequence capture library preparation and Illumina sequencing were performed by the Centre of Genomic Enabled Biology and Medicine (CGEBM) at the University of Aberdeen. 100ng of ds-cDNA from each sample was sheared using sonication, end-repaired and purified with AMPure XP beads. The pre-capture SureSelect^{XT2} reagent kit was used to ligate indexing adapters to the DNA fragments and the libraries were amplified using PCR before quality assessment on an Agilent TapeStation. Sequence capture was performed with the custom baits following the manufacturer's instructions. Indexed samples were pooled together for the hybridisation step, where RNA baits bound to the virus cDNA were captured using a streptavidin bead-based separation. Captured libraries were amplified using PCR (12-14 cycles) and the amplified library confirmed using the Agilent BioAnalyser. The pooled library was run on a single NextSeq500 flowcell (2x150bp pair-end configuration) according to Illumina specifications.

4.2.3 SAV genome analysis

Demultiplexed FASTQ files were trimmed of sequencing adapters and poor-quality bases using TrimGalore v.0.4 (min q-score of 30) (Krueger, 2015). The average cDNA fragment size prior to adapter ligation was 192 bp, leaving an overlap between 2 x 150 bp paired reads. Paired reads were merged into longer contigs when possible using the BBMap (Bushnell, 2016) programme BBMerge (default settings). Both merged and unmerged reads were used for subsequent analyses (average sequence length: 187 bp). PCR duplicates were removed using DeDupe (BBMap package) with default parameters. BBSplit (BBMap package) was used to align all quality controlled passed reads of each isolate to a reference of each of the six subtypes (SAV1 - AJ316244; SAV2 - AJ316246; SAV3 - DQ149204; SAV4 -

MH708651; SAV5 - MH708653; SAV6 - MH238448). Different SAV subtypes were considered present in the same sample when >50 reads mapped to locations of the reference sequence that contained subtype-unique variants (e.g. as visualized in Figure S1). The resulting bins of reads were mapped to the corresponding reference genome with BWA-MEM (Li, 2013), using default settings and the alignments were then processed using SAMtools v1.3.1 (Li et al., 2009). Consensus sequences were generated using FreeBayes variant-calling (Garrison and Marth, 2012) and the VCF manipulation package vcflib (Garrison, 2012) to produce a FASTA file for each subtype-specific isolate. The percentage of the SAV genome captured in each consensus sequence was calculated (Table 4.2) by comparison to a reference sequence of the same subtype. Proportions of each subtype per sample were estimated by comparing sequencing depth of each assembly across ORF2 (Figure 4.1)

4.2.4 Bayesian phylogenetic analysis

Isolate sequences from which ≥85% of the structural polyprotein (ORF2) was assembled were used for phylogenetic analysis, along with all unique published SAV ORF2 sequences (isolate names given in Figure *4.2*). Sequence alignment (n=51 sequences) was performed using MAFFT v.7 with default parameters (Katoh and Standley, 2013). The final alignment was 3,917 bp in length and is available in the Supplementary Data of the published manuscript (https://www.sciencedirect.com/science/article/pii/S0044848619321970#s0075). Prior to phylogenetic analysis, the best fitting nucleotide substitution model was estimated using IQTREE v1.5.3 (Nguyen et al., 2015; Trifinopoulos et al., 2016). Bayesian phylogenetic analysis was performed using BEAST v.2.4.4 (Bouckaert et al., 2014), employing the best fitting nucleotide substitution model (general time reversible model; Tavaré, 1986, with gamma distribution of among-site rate variation estimated under 4 rate categories) an uncorrelated lognormal relaxed clock model (Drummond et al., 2006), a random starting tree and a Bayesian coalescent constant population model (Drummond et al., 2005). The Markov Chain Monte Carlo (MCMC) chain length was 200 million generations, with sampled parameters logged every 20,000 generations. Convergence and mixing were assessed using Tracer v.1.6 where all effective sample size statistics (ESS) were >200. A maximum clade credibility tree was created using TreeAnnotator (Drummond et al., 2012) after removing the first 10% of trees as burn-in, The resulting tree (Figure *4.2*) was visualised in FigTree (http://tree.bio.ed.ac.uk/software/figtree/).

4.2.5 Subtype-specific SAV genetic diversity

Bins of sequence reads representing different SAV subtypes were mapped back to their consensus sequence (generated as outlined above) using BWA-MEM, with alignment processing performed using SAMtools. SNV detection was then performed using FreeBayes, with stringent parameters to reduce type-I error: a minimum base quality of 30, a minimum

mapping score of 30, a minimum variant frequency of 0.05, a minimum coverage of 50 reads, and a p-value of < e^{-7}. All SNVs were visually inspected and their effect on coding sequence determined. Genomic location of each SNV from each isolate was determined and plotted in reference to an SAV genome alignment (n=31) using the R package ggplot2 (Wickham, 2009) and coloured by both codon effect (Figure 4.4) and novelty of the SNV in question (Figure 4.5). Samples from which a subtype-consensus sequence could be obtained were further analysed to determine the percentage of the genome which was variable (i.e. proportion of the nucleotide sites which contained an SNV) (Table 4.3).

4.2.6 Intra-subtype haplotype reconstruction and phylogenetic analysis

Due to inconsistent results obtained from several haplotype reconstruction softwares, haplotype reconstruction of small genomic regions (~350 bp) was performed manually using visualised alignments on IGV (Thorvaldsdóttir et al., 2013). Isolates were chosen for analysis based on visualisation of SNVs; only those with several SNVs each located within the maximum read length (~250bp) of the next SNV were considered (samples IRE/3/12, IRE/38/11, SCO/G415/09, SCO/G572/09, SCO/G573/09, and SCO/G582/07). A genomic region overlapping the E3 and E2 genes was selected to maximise the number of SNVs present in the samples. SNVs were considered to belong to the same haplotype if >99% of the reads with one SNV also contained a second SNV, itself present in <1% of reads not containing the first SNV. Haplotype sequences generated for samples IRE/3/12, IRE/38/11, SCO/G415/09, SCO/G572/09, SCO/G573/09, and SCO/G582/07 were aligned against the same regions of consensus SAV1 sequences recovered for other samples within the study, before a Bayesian phylogenetic analyses was performed as described above.

4.3 Results

4.3.1 Sequence capture for genome-wide SAV analysis

Eighteen tissue samples from four fish species (Atlantic salmon, rainbow trout, European plaice (P platesa) and common dab (L. limanda)) were used in this study (Table 4.1). For the farmed salmonid samples, heart tissue from a representative sample of five fish, pooled together at the farm level, were obtained during routine visits to fish farms on the west coast of Scotland and Shetland. The pooled wild flatfish were caught on the east coast of Scotland, while the individual flatfish samples were caught in Dublin bay and the Celtic sea. All samples tested positive for SAV using qPCR analysis (Table 4.2). The samples were sequenced on an Illumina NextSeq 500 following enrichment of SAV cDNA using the Agilent SureSelect^{XT2} platform. The sequence capture probes covered the full genomes of SAV1-SAV5 genotypes and also included two partial fragments of SAV6 (Fringuelli et al. 2008).

In total, 136,914,992 reads (20.5 Gb total DNA) were obtained that passed quality control (Table 4.2). The proportion of SAV reads among all sequenced reads ranged from 2.9% to 91.4% per sample (Table 4.2) and was positively correlated with viral RNA load estimates obtained by qPCR (Pearson's $R=0.92$, $p<0.001$). Sequence coverage was uneven across the genome, with more reads mapping to ORF2 (structural polyprotein) than ORF1 (non-structural polyprotein) for SAV1, and a higher 3' coverage of each polyprotein for SAV2 (Figure 4.6).

4.3.2 Evidence for co-presence of different SAV subtypes

An approach was designed to map reads from single samples to all six SAV reference genome sequences (note: an SAV6 genome became available while the work was in preparation; (Gallagher et al., 2018). Sequence reads from all eighteen samples mapped exclusively to SAV1, SAV2 and SAV5. Sequence reads generated from two out of five samples collected from individual fish (common dab) mapped to both SAV1 and SAV5 (Table 4.2), indicating the presence of viral RNA from two different SAV subtypes in the same host tissue (e.g. Figure 4.7). Additionally, eleven out of the thirteen pooled samples contained sequence reads that mapped to two SAV subtypes (Table 4.2), with both SAV1-SAV2 and SAV1-SAV5 pairings observed. These included pooled samples from both farmed Atlantic salmon and rainbow trout, as well as wild caught Scottish common dab. In all cases, the Sanger-generated sequence confirmed the presence of only the SAV subtype present at higher coverage across a greater proportion of the genome (Table 4.2).

For all samples showing a co-presence of reads from two SAV subtypes, at least half of the genome was represented; including the majority of the ORF2; sufficient to discriminate between the SAV genotypes (Table 4.2). A Bayesian phylogenetic analysis of ORF2 was performed, including all the unique publicly available SAV sequences and the consensus sequences assembled during this study; inclusive of SAV subtypes co-present in the same samples (Figure 4.2). SAV sequences of co-present subtypes belong to a diverse range of phylogenetic lineages, including distantly related clades for SAV1 and SAV2 (Figure 4.2) and were unique compared to the published list of sequenced SAV genes and genomes as well as the list of commonly handled SAV isolates in both Marine Science Scotland and Marine Institute. This abundant novel genetic diversity is incompatible with scenarios where the co-presence of SAV subtypes in the same sample resulted from contamination (see Section 4.4).

4.3.3 Within-subtype SAV diversity

To test for the existence of within-subtype SAV diversity, I mapped the recovered reads back to consensus sequences for the different SAV genomes. Any samples showing co-presence of two SAV subtypes were mapped to two unique references. Genome-wide single nucleotide

variant (SNV) calls were generated for all samples, excluding SNVs with frequencies <5% to avoid false-positives. All samples showed evidence for minor SNVs with large differences observed between samples (0.01 to 0.77% of genome affected) (Table 4.3).

Considering the evidence favouring a co-presence of phylogenetically distinct SAV subtypes in single samples, it seemed plausible that distinct subtype strains might also be commonly present in the same sample. However, the short sequencing reads obtained (average fragment length: 192 bp) limits our ability to distinguish such scenarios from viral strains that evolved within-host (Domingo et al., 2012). In an attempt to address this issue, I performed phylogenetic analysis on a short (manually-assembled, see Materials and Methods and Figure 4.8) fragment of the SAV E3 and E2 genes for a subset of samples (samples IRE/F3/12, IRE/F38/11, SCO/G415/09, SCO/G572/09, SCO/G573/09, SCO/G582/07) that contained enough variation for reliable viral haplotype phasing (311bp fragment). SAV1 subtype haplotypes identified within each sample were characterised by phylogenetic analysis (Figure 4.3). Two distinct SAV1 clades were observed (posterior probability of 0.84), sharing a pairwise nucleotide similarity of 98%, compatible with a co-infection scenario for single-individual samples (IRE/F3/12 and IRE/F38/11) and (at minimum) a co-circulation scenario for the represented pooled samples (SCO/G572/09 and SCO/G415/09). Additionally isolates IRE/F3/12 and SCO/G572/09 also contained SAV1 haplotypes that branched within monophyletic clades for each sample (e.g. isolate IRE/F3/12 has three SAV1 sequences sharing 99.6% pairwise identity on average). While it cannot be assumed to be the case for SCO/G572/09 (a pooled sample), for isolate IRE/F3/12, these results are consistent with intrahost evolution of SAV.

A broader definition of the SNV landscape of these samples is visualized in Figure 4.4, including the proportion of synonymous and non-synonymous variants across different genes in the SAV genome. The caveat to this analysis is that the data represents viral strains from both single fish samples and pools of five individuals. Hence the presence of multiple distinct SAV strains in samples may represent one or multiple of the following scenarios: SAV co-infection in individual fish, intrahost SAV evolution during the time course of an infection, or, and most likely, the co-circulation of several SAV strains infecting different fish at the farm level. Consequently, this limits the value of statistical analyses to formally contrast differences in SNV rate across different genes and host species, due to confounding effects of possible co-circulations, co-infections and intra-host evolution. Informally, it seems notable that large variation in the number of observed SNVs across samples was inclusive of the different species, with no obvious differences between farmed and wild fish (Table 4.3). However, I did observe a notably higher number of SNVs in SAV1 compared to SAV2/SAV5

(Table 4.3), which may reflect a higher natural diversity of this subtype in Scottish and Irish waters.

4.4 Discussion

Second and third generation sequencing platforms have been widely used to study pathogen genomic variation. However, the uptake of such tools to characterize genetic diversity for pathogenic viruses affecting farmed fishes has been slow, leaving knowledge gaps in our understanding of commercially important diseases including PD and SD. The few studies that have achieved a deep profiling of pathogenic viral diversity in farmed fish provide ample evidence for intrahost viral diversity. However, until recently only a few studies considered natural SAV infections, with most work done on cultured viral isolates, which likely lack the natural genetic diversity, instead accumulating novel genetic variation associated with passaging in cell culture (Karlsen et al., 2006). This should be especially true for cultured material consisting of multiple pooled fish, as any genetic variation present will be combined and presumably removed/reduced by selection or drift from the onset of cell culture.

In this Chapter, I developed a target enrichment sequencing approach to characterise the genetic diversity of SAV in natural infections and compared the levels of diversity between single-individual and pooled samples. The generated in this Chapter provides evidence for common co-circulation (and potentially co-infections) of two SAV subtypes and within-subtype strain diversity, both on Atlantic salmon and rainbow trout farms, as well as in wild flatfish populations. Furthermore, SAV1 and SAV5 were co-detected in two single-individual flatfish samples from Dublin Bay representing, to my knowledge, the first empirical evidence for subtype-level SAV co-infection. This finding cannot be explained by contamination, due to the extensive novel phylogenetic diversity of the viral sequences identified. Under a scenario of contamination, for example resulting from SAV PCR amplicons previously generated in the lab, I would expect the repeated presence of one or a few contaminating samples. The only isolate used previously in our lab was SAV4640 (accession: JX163854), to which none of the new SAV1 sequences matched. Additionally, several isolates previously identified as SAV1 were found here to contain SAV2 or SAV5; no samples of these subtypes had been subjected to PCR in the laboratory where amplicon libraries were prepared. Finally, under a scenario of contamination, I would expect the issue to impact all samples, given that they were processed together; however, this was not observed, as several contained reads mapping to a single SAV subtype (isolates IRE/F3/12, IRE/F10/12, IRE/F39/11, SCO/G576/07, and SCO/G865/15). I thus conclude that the presence of multiple SAV subtypes circulating on single salmonid farms and wild fish populations is a true reflection of natural infections. Importantly, these findings have previously been hidden to consensus sequence approaches reliant on Sanger sequencing of PCR products (Karlsen et al., 2006;

Domingo et al., 2012; Hjortaas et al., 2013; Matejusova et al., 2013), advocating a need for routine uptake of higher-resolution sequencing methods from individual rather than pooled samples for epidemiological studies and diagnostics. The high prevalence of within-sample SAV diversity indicates that the dynamics of PD are markedly more complex than widely recognised, suggesting a need for extensive reappraisal and expansion of existing genetic databases to support ongoing disease management decisions.

The sequence capture method employed in this study allows for an unbiased characterisation of viral diversity. My data corroborates previous findings (Hammoumi et al., 2016) that the efficiency of viral sequence capture depends on the initial viral load and is variable between analysed samples. Additionally, the SureSelect^{XT2} protocol used requires pooling of barcoded samples prior to capture; while reducing handling and reagent costs, this prevents the normalization of library quantities between isolates. A difference was also observed in coverage across the SAV genome, which might reflect the natural abundance of the two mRNAs in SAV. In many samples the structural polyprotein, encoded by a ~4kb 3' mRNA, showed higher coverage than the ~8kb 5' mRNA encoding the non-structural polyprotein, which has been observed for other members of the alphavirus genus (Carrasco et al., 2018).

While I discovered evidence for within-subtype intrahost variation of SAV in wild fish, my data was unsuitable to phase complete viral genomes for closely related strains. This is due to the use of short-read sequence data, which allows high confidence SNV calling, but makes it challenging to link SNVs separated by distances greater than the sequenced fragment length (192bp average in this study), even when using software dedicated to this problem (data not shown). Hence, future studies of within-subtype SAV diversity will benefit from longer sequence information, which could be generated using the same capture strategy and larger fragment sizes. However, even better results are envisaged by adopting third-generation long-read sequencing tools (e.g. Oxford Nanopore and Pacific Biosciences) (Posada-Cespedes et al., 2017) and/or linked-read sequencing (Russell et al., 2018), which represent promising tools for ongoing genomic investigations into aquaculture pathogens.

Previous work has characterised the presence of distinct SAV subtypes in proximal geographical locations around Ireland and Scotland (Graham et al., 2012). Interestingly, the industry (https://www.fishfarmingexpert.com/article/msd-survey-vaccination-significantly-reducing-pd-positive-results/) has noted that different SAV subtypes have affected individual Scottish salmon farms in subsequent years, consistent with co-circulation of distinct viral lineages on small spatial scales. These results are consistent with such non-published reports, as all SAV subtypes detected were previously detected in the same regions (Figure *4.1*) and demonstrate the presence of multiple SAV strains and subtypes on single farms.

From a practical perspective, it will be important to document the conditions under which complex SAV infections arise, are maintained, and impact pathological outcomes. High-throughput sequencing of pooled samples can accurately identify SAV strains and subtypes present on an individual fish farm; however it does not allow for the characterisation of SAV co-infections and subsequent association between genetic diversity levels (or the presence of individual pathogenic strains) and disease prognosis at the scale of individual fish. Additionally, if Sanger sequencing is used alone for identification of viral genotypes present at the farm level from pooled material, even more information is lost. This is demonstrated by the fact that all Scottish samples tested had the dominant SAV strain correctly genotyped, but missed the secondary strains present in the pooled samples. This is of importance in countries or regions that regulate the salmon industry based on SAV strain presence, such as Norway. Determining the spread and transmission of particularly virulent strains, or even identifying the presence of these virulent strains at low frequencies is challenging, if not impossible using Sanger sequencing approaches. It is now well-established that the implications of co-infections for disease progression can be highly varied, and may range from detrimental to beneficial (McArdle et al., 2018). In this respect, a priority will be to determine whether higher SAV diversity (either intrahost or co-circulating in a salmon cage) is associated with different disease progression or alteration in mortality rates. Such analyses would be most powerful if done using high coverage genome-wide sequencing of samples from individual fish.

While pooling samples may be appropriate for routine statutory disease surveillance (OIE, 2017b), care needs to be taken to ensure that the sensitivity of detection assays is sufficient for detection of viral nucleic acid even when present in lower titres and/or few individuals in the pool (Hall, 2013). Additionally, sampling at the individual level is required for epidemiological studies investigating the origin or relatedness of disease outbreaks where a lack of accurate sequence data can infer incorrect transmission patterns and population dynamics.

Finally, several PD vaccines, with varying efficiencies, are available on the market (Karlsen et al., 2012; Xu et al., 2012). While it is yet to be established whether SAV intrahost variation (or subtype co-circulation on the same farm) impacts the efficacy of PD vaccination, this is an important line of investigation considering previous work, which showed that virulence can vary in multiple-genotype infections compared to single infections (Lancaster and Pfeiffer, 2012; Bose et al., 2016).

In conclusion, I have demonstrated an unbiased approach to enrich viral RNA in infected fish tissues and used it to define previously unrecognized diversity in a viral pathogen responsible for significant commercial losses and welfare issues in salmonid aquaculture. A more

thorough definition of the genetic diversity characterising viral infections in aquaculture, especially the associated implications for pathogenicity and disease outcomes, along with a suitable sampling strategy, will be essential in the ongoing battle against viral diseases threatening the expansion of global aquaculture.

Data Availability

Raw sequence files are available under SRA BioProject PRJNA599596. Genome sequences are provided in the Supplementary Data of the published manuscript (https://www.sciencedirect.com/science/article/pii/S0044848619321970)

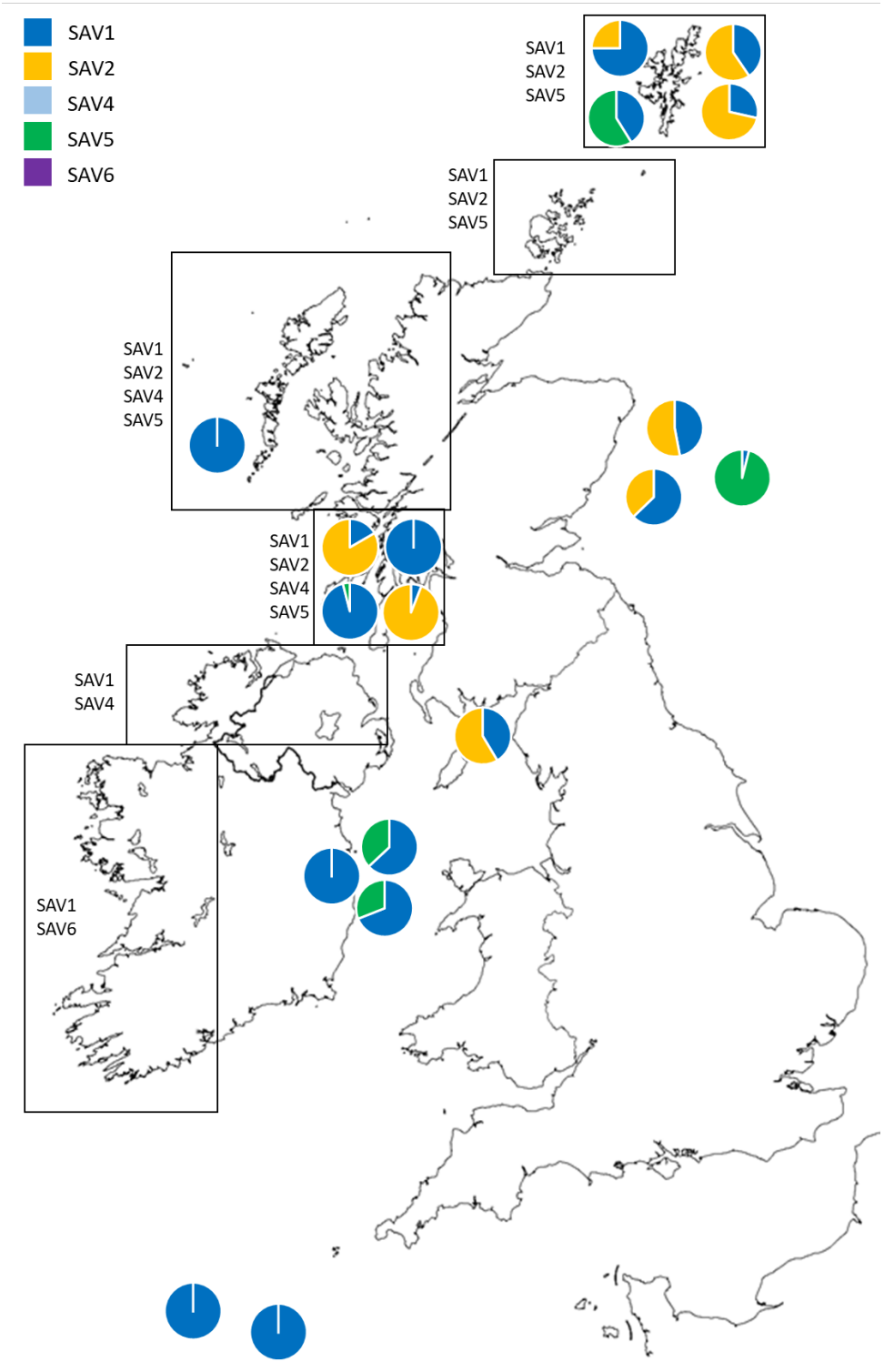

Figure 4.1. Geographic distribution of SAV in Scotland and Ireland. Regions previously characterised with the presence of multiple SAV subtypes are indicated with boxes. Pie charts represent isolates characterised in this study with the estimated proportion of co-circulating strains shown by colour

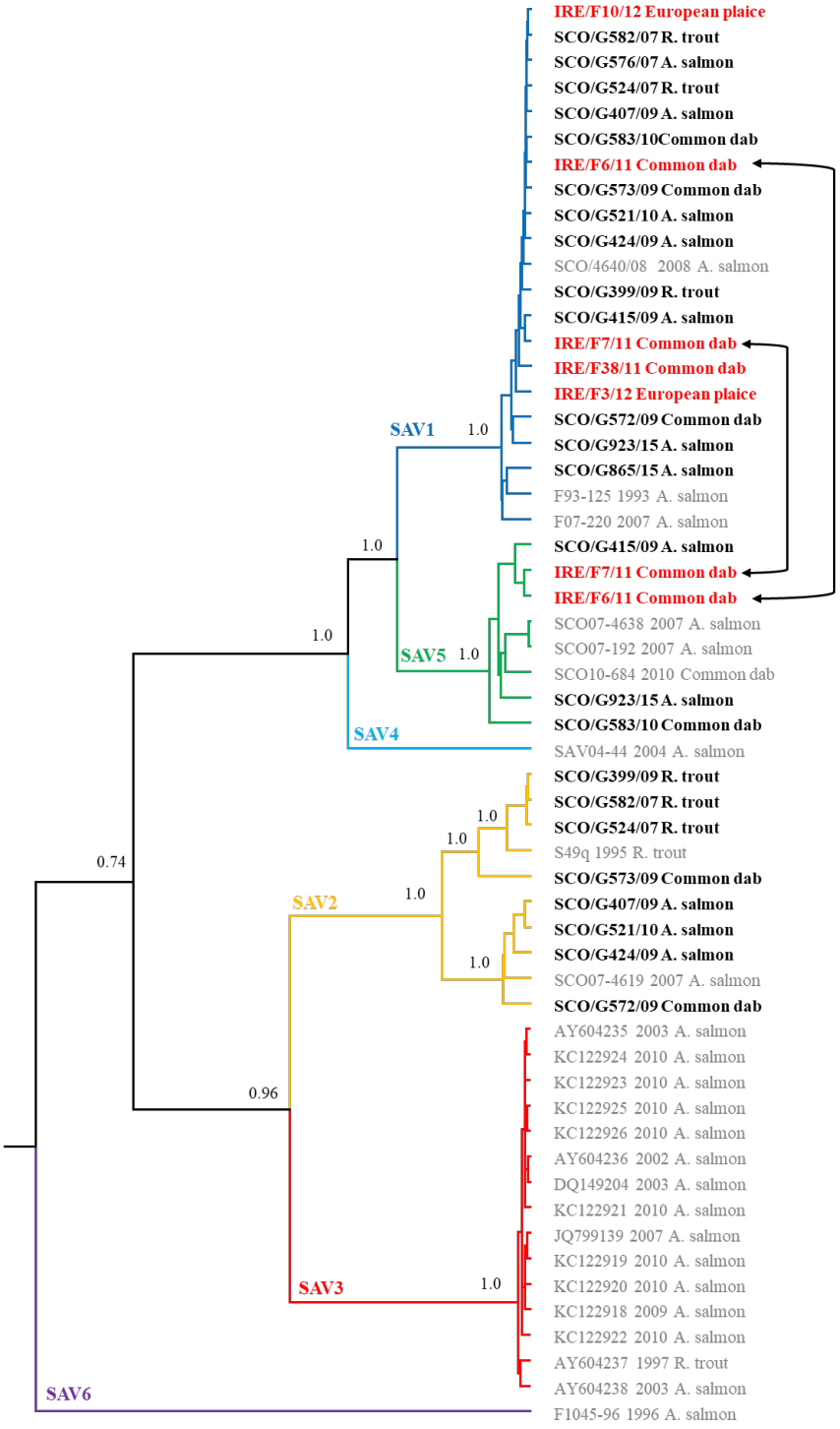

Figure 4.2. Bayesian phylogeny showing evidence for two SAV subtypes within single samples used in this study. The tree was built from a 3,935 bp nucleotide alignment of the SAV structural polyprotein (ORF2). The analysis was performed in BEAST2 using the best fit nucleotide substitution model (GTR+G), a relaxed molecular clock model, and a coalescent constant population model. Statistical support for key nodes is indicated by posterior probability values. Consensus sequences from single fish samples have red font titles, while sequences from pooled fish samples have black font titles. Arrows joining branches indicate subtype-level co-infections within a single fish. Samples with grey font titles were downloaded from NCBI GenBank

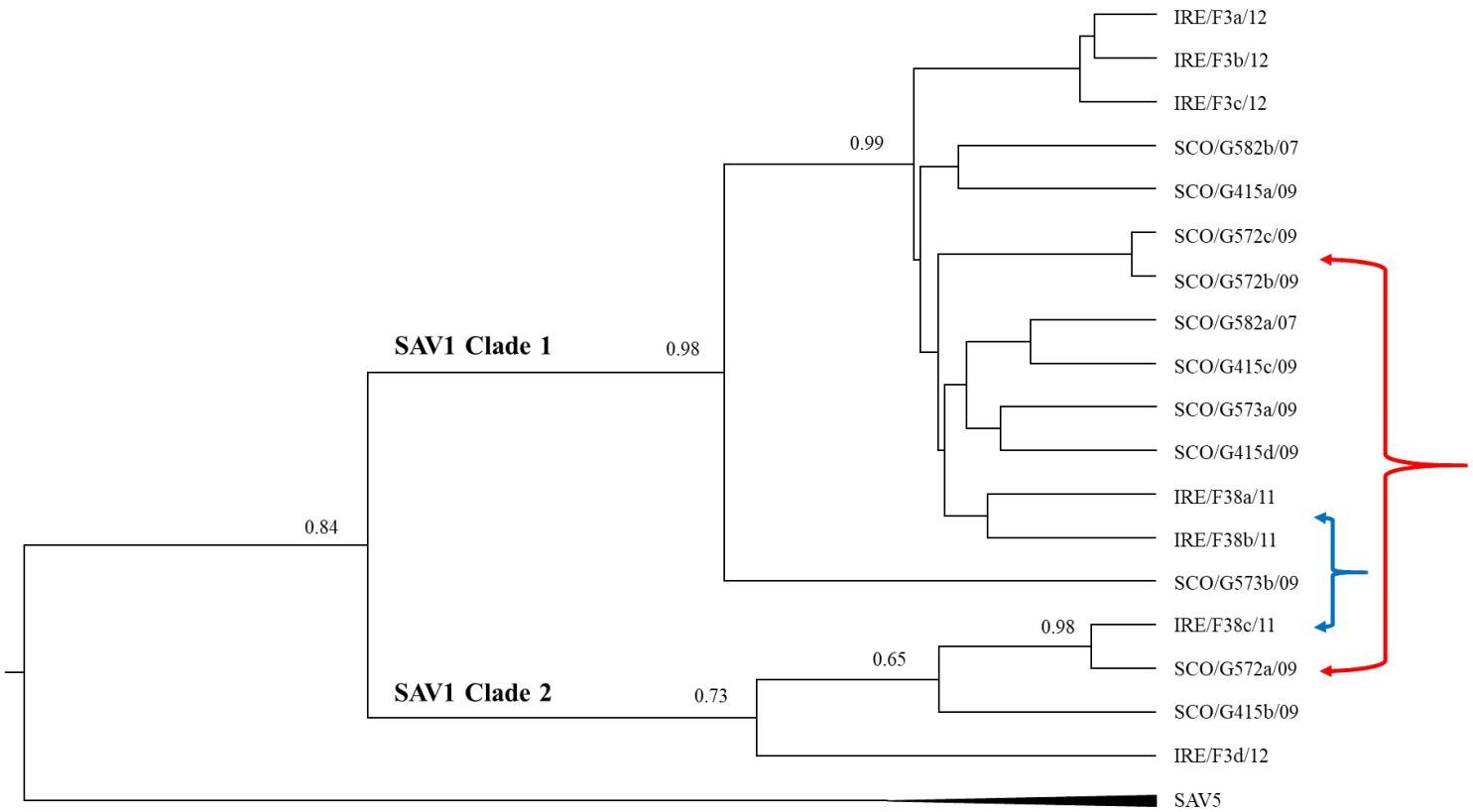

Figure 4.3. Bayesian phylogenetic analysis of a 311 bp fragment of the SAV E3/E2 genes to identify the relationship of manually-phased haplotypes in six SAV1 samples. Strains are labelled by isolate with letter-based identifiers (e.g. 415a, 415b) indicating multiple strains per isolate. Example co-infections in single-individual samples are indicated by blue arrows, while strains from pooled samples are shown by red arrows. The analysis was performed in BEAST2 using the best-fit nucleotide substitution model (TN93+G), a relaxed clock model and a coalescent constant population tree prior. Statistical support for key nodes is indicated with posterior probability values.

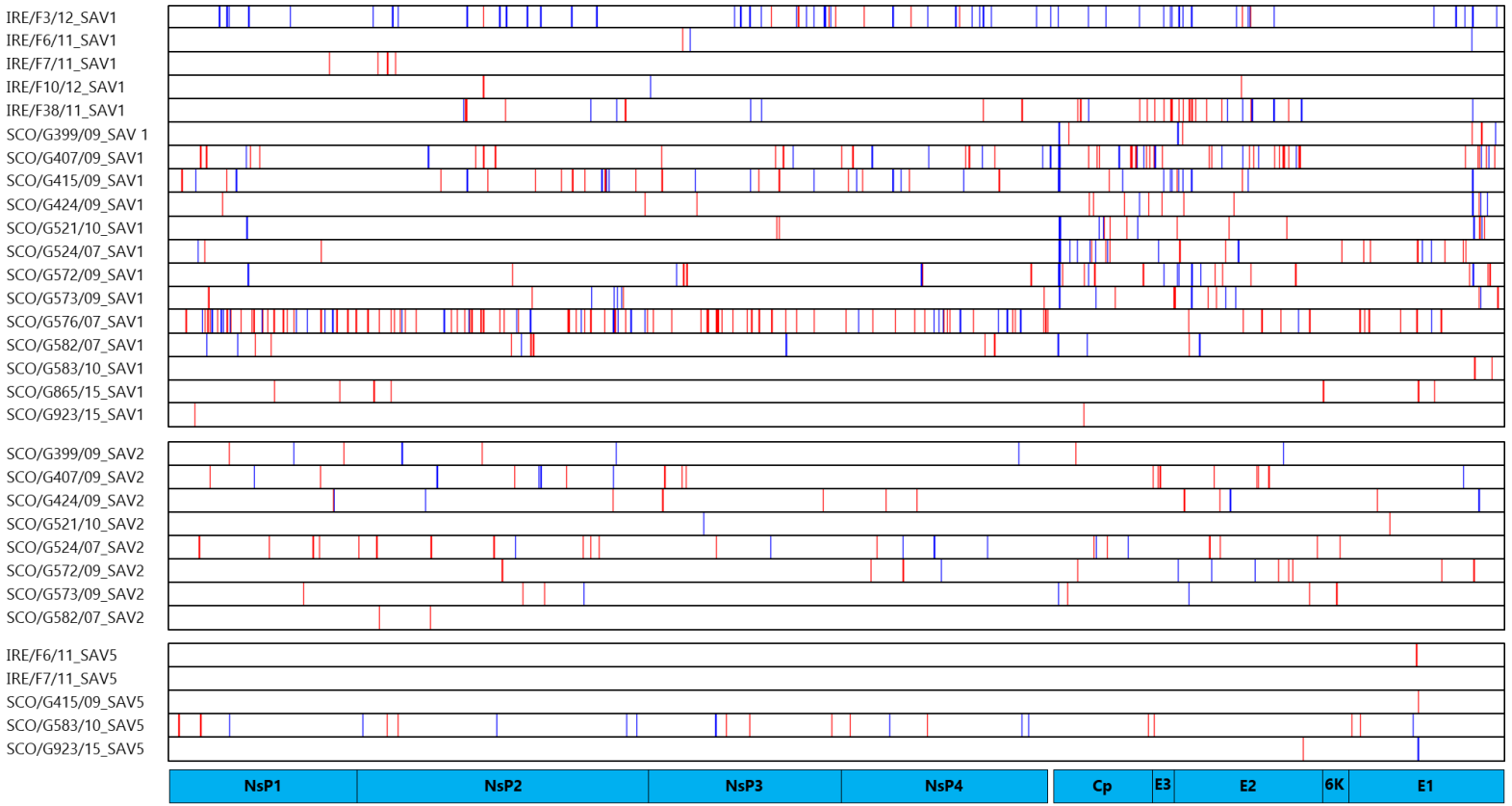

Figure 4.4. SNV landscape of all samples including secondary strains (n=31) with position of SNV representing genomic location. Synonymous and nonsynonymous SNVs are coloured blue and red respectively. Approximate gene regions are indicted for reference.

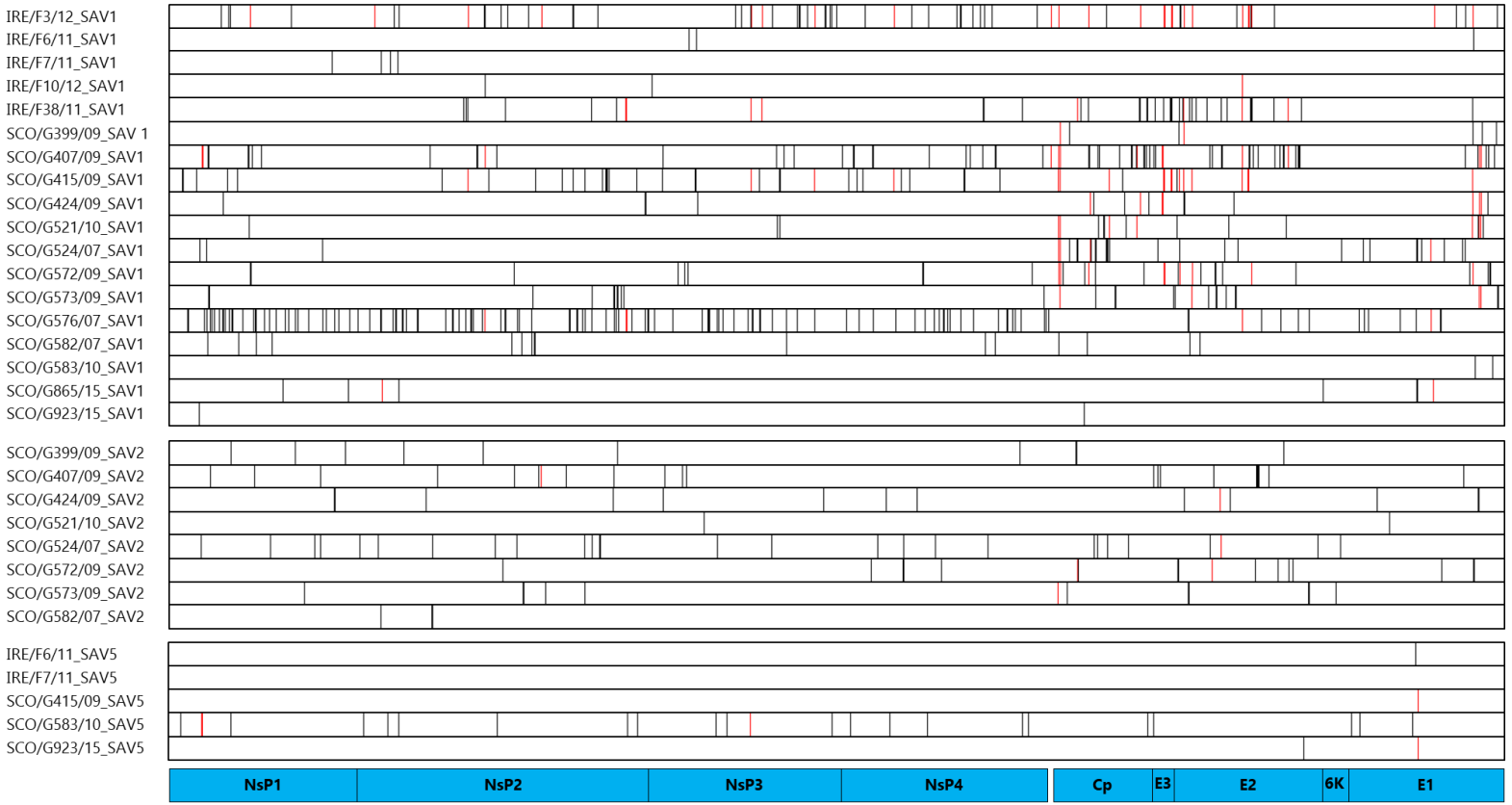

Figure 4.5. SNV landscape of all samples including secondary strains (n=31) with position of SNV representing genomic location. Unique and shared SNVs are coloured black and red respectively. Approximate gene regions are indicted for reference.

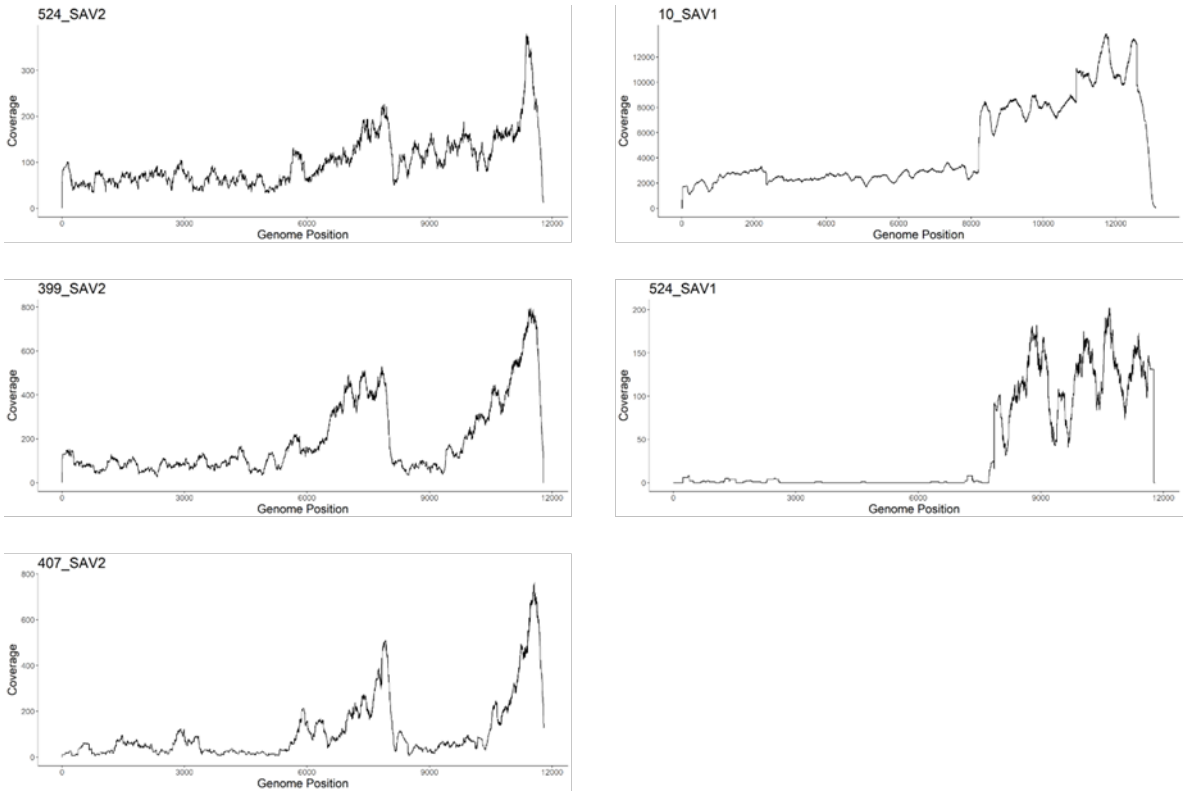

Figure 4.6. Coverage plots of representative isolates showing an increase in coverage over the structural polyprotein for SAV1 samples, but an increase at the 3' end of each polyprotein in the SAV2 samples.

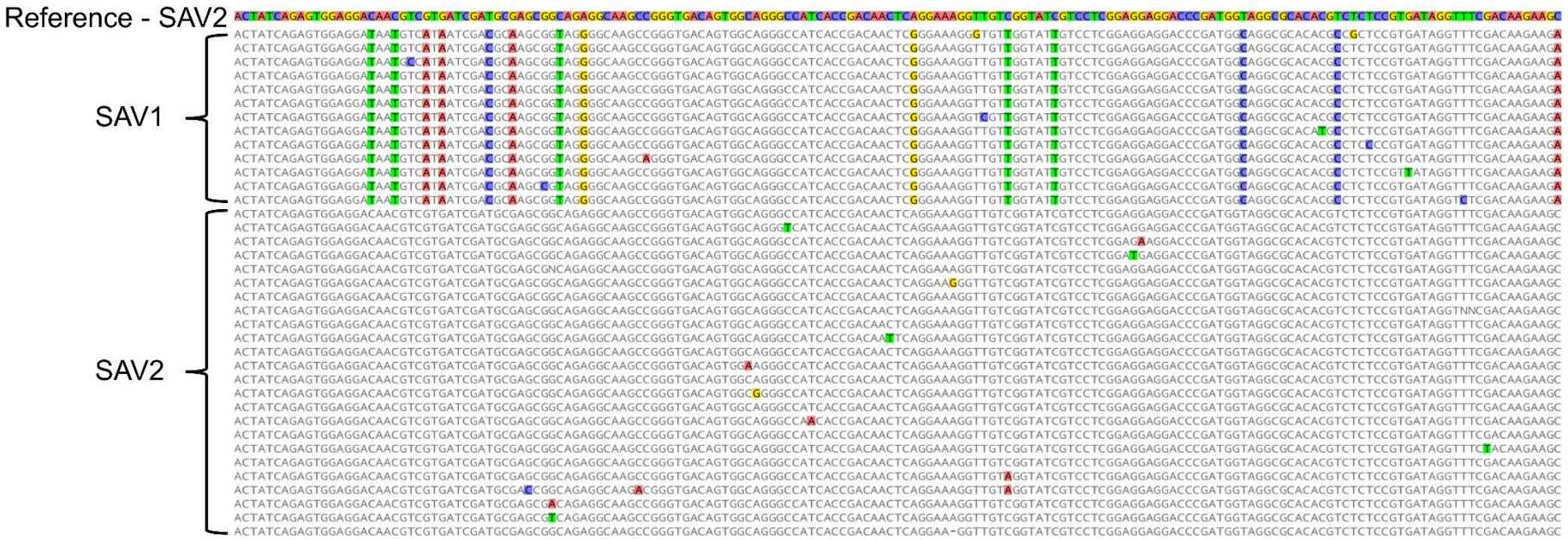

Figure 4.7. Visualisation of a subtype-level 'co-infection' in a pooled sample where reads from SAV1 clearly distinguishable from those of SAV2.

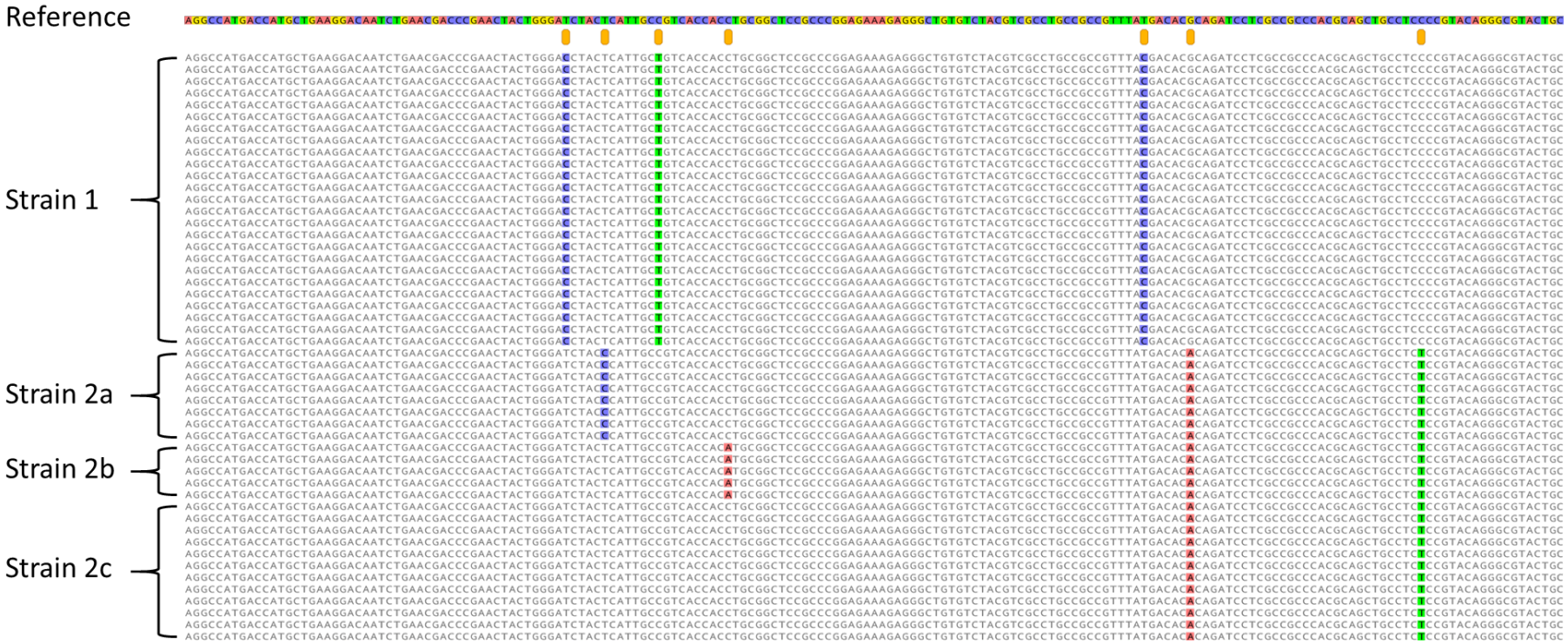

Figure 4.8. Visualisation of haplotype-level reconstruction of the viral population within an infected sample. Haplotypes 2a, 2b and 2c are all more closely related to each other than they are to haplotype 1, indicating the presence of a coinfection of two viral strains (strain 1 and strain 2) and the subsequent intra-host diversification of one of these stains (strain 2)

Table 4.1. Isolate details for the eighteen SAV-infected heart tissues analysed

Sample ID	Year of Isolation	Location of Isolation	Host Species	Source	Sampling Method
IRE/F3/12*	2012	Celtic Sea	Plaice	Wild	Individual
IRE/F6/11*	2011	Dublin Bay	Dab	Wild	Individual
IRE/F7/11*	2011	Dublin Bay	Dab	Wild	Individual
IRE/F10/12*	2012	Celtic Sea	Plaice	Wild	Individual
IRE/F38/11*	2011	Dublin Bay	Dab	Wild	Individual
SCO/G399/09	2009	Argyll	R. trout	Farmed	Pooled (5x)
SCO/G407/09	2009	Shetland	A. salmon	Farmed	Pooled (5x)
SCO/G415/09	2009	Argyll	A. salmon	Farmed	Pooled (5x)
SCO/G424/09	2009	Shetland	A. salmon	Farmed	Pooled (5x)
SCO/G521/10	2010	Shetland	A. salmon	Farmed	Pooled (5x)
SCO/G524/07	2007	Central	R. trout	Farmed	Pooled (5x)
SCO/G572/09	2009	East Coast	Dab	Wild	Pooled (5x)
SCO/G573/09	2009	East Coast	Dab	Wild	Pooled (5x)
SCO/G576/07	2007	South Uist	A. salmon	Farmed	Pooled (5x)
SCO/G582/07	2007	Argyll & Bute	R. trout	Farmed	Pooled (5x)
SCO/G583/10	2010	East Coast	Dab	Wild	Pooled (5x)
SCO/G865/15	2015	Argyll & Bute	A. salmon	Farmed	Pooled (5x)
SCO/G923/15	2015	Shetland	A. salmon	Farmed	Pooled (5x)

*Samples obtained from Marine Institute Ireland are indicated with *, all other samples were obtained from Marine Science Scotland.*

Table 4.2. Summary of genome-wide SAV data following sequence capture and Illumina sequencing

Sample ID	Species	Relative Viral Load (Cq)	Total Reads per Sample	% SAV Reads	% Host Reads	Sanger Subtyped	% SAV1 Genome Covered [mean coverage]	% SAV2 Genome Covered [mean coverage]	% SAV5 Genome Covered [mean coverage]
SCO/G865/15	Atlantic salmon	23.5	39,697,464	91.39	4.63	SAV1	100% [15,126x]		
SCO/G415/09	Atlantic salmon	31.8	9,988,198	45.13	16.22	SAV1	100% [75x]		84% [3x]
SCO/G521/10	Atlantic salmon	34.2	9,548,438	50.65	14.45	SAV1	74% [5x]	46% [3x]	
SCO/G424/09	Atlantic salmon	33.9	9,053,958	17.97	22.80	SAV2	77% [5x]	97% [13x]	
SCO/G407/09	Atlantic salmon	32.2	7,760,266	41.90	32.15	SAV2	100% [27x]	100% [107x]	
SCO/G582/07	Rainbow trout	29.6	5,640,002	9.16	55.24	SAV2	91% [52x]	100% [513x]	
SCO/G399/09	Rainbow trout	31.7	4,657,156	16.87	40.78	SAV2	73% [2x]	100% [191x]	
IRE/F7/11	Common dab	30	167,848	6.40	NA	SAV1	94% [24x]		54% [1x]
IRE/F6/11	Common dab	32.7	209,960	6.60	NA	SAV1	88% [30x]		72% [2x]
IRE/F38/11	Common dab	33.2	3,506,568	19.70	NA	SAV1	96% [422x]		
SCO/G573/09	Common dab	32.9	5,156,948	24.70	NA	SAV2	86% [4x]	100% [18x]	
SCO/G572/09	Common dab	32.4	5,198,444	48.70	NA	SAV2	87% [7x]	98% [15x]	
SCO/G583/10	Common dab	32.7	7,723,112	18.00	NA	SAV5	74% [3x]		100% [47x]
IRE/F3/12	European plaice	32	3,607,684	13.60	NA	SAV1	100% [314x]		
IRE/F10/12	European plaice	23.8	13,304,474	76.90	NA	SAV1	100% [4,864x]		
SCO/G524/07	Rainbow trout	31.1	3,939,136	4.10	28.86	SAV2	58% [40x]	100% [101x]	
SCO/G576/07	Atlantic salmon	31.9	4,404,064	2.90	31.94	SAV1	99% [89x]		
SCO/G923/15	Atlantic salmon	33.9	3,351,272	9.35	43.92	SAV5	47% [1.5x]		54% [5x]

Relative viral load estimated using highly conserved primers in the capsid gene. Only reads with a cut-off q-score of 30 were used. All samples were subtyped using Sanger sequencing for a fragment of the E2 gene. Proportion of reads from the host was not possible for the flatfish species due to a lack of available genomic resources

Table 4.3. Genome-wide SAV genetic diversity present within SAV subtypes

Sample	Species	Subtype	# SNVs	% Genome Variable	% of SNVs non-synonymous
IRE/F3/12	European plaice	SAV1	78	0.66%	24.36%
IRE/F10/12	European plaice	SAV1	3	0.03%	66.67%
IRE/F38/11	Common dab	SAV1	26	0.22%	38.46%
SCO/G407/09	Atlantic salmon	SAV1	44	0.37%	72.73%
SCO/G415/09	Atlantic salmon	SAV1	46	0.39%	41.30%
SCO/G576/07	Atlantic salmon	SAV1	90	0.76%	73.33%
SCO/G865/15	Atlantic salmon	SAV1	7	0.06%	0.00%
SCO/G399/09	Rainbow trout	SAV2	5	0.04%	60.00%
SCO/G407/09	Atlantic salmon	SAV2	19	0.16%	78.95%
SCO/G424/09	Atlantic salmon	SAV2	4	0.03%	50.00%
SCO/G524/07	Rainbow trout	SAV2	17	0.14%	82.35%
SCO/G572/09	Common dab	SAV2	11	0.09%	72.73%
SCO/G573/09	Common dab	SAV2	5	0.04%	60.00%
SCO/G582/07	Rainbow trout	SAV2	1	0.01%	100.00%
SCO/G583/10	Common dab	SAV5	8	0.07%	87.50%

SNVs called only for consensus SAV sequence for which >95% of total genome length recovered. Only SNVs with frequencies >5% included

Chapter 5. RNA Virus characterisation using metagenomics in aquaculture

The data presented in this chapter are being used in manuscript under preparation for submission to a peer-reviewed journal

Summary

Metagenomics is a powerful tool for identifying both known and novel viral species, particularly when characterising emerging viral diseases. In this Chapter, I report a new analysis pipeline for characterising RNA-based viromes from short-read RNA-seq datasets, which can be implemented on datasets with and without genome sequences for the host species. This Chapter goes on to report the benchmarking and value of the pipeline using a range of simulated viromes and existing host RNA-Seq datasets, before demonstrating its application for *de novo* virus discovery and biological characterization using sequencing data from Pacific oyster (*C. gigas*) infected with norovirus and sampled at multiple time points post-infection.

5.1 Introduction

Viruses are the most abundant biological entity on Earth and likely infect every species of cellular life across all habitable environments (Koonin et al., 2006). However, viruses are traditionally characterised when causing symptomatic diseases to humans or economically important plants and animals. Viral pathogens of commercially important aquatic species are currently predominantly characterised using cell culture to isolate the virus in question, which remains effective for viral diagnostics and characterising economically important diseases. However the majority of aquatic viruses are thought to be unculturable (Wang et al., 2002) and as RNA viruses do not contain a universally conserved gene (like 16S rRNA in bacteria, or CO1 in animals), species-specific PCR primers can only be designed after the virus has been identified.

With the continuing increase in sequence data quantity and quality, alongside decreasing sequencing costs, the application of NGS for metagenomics is emerging as a powerful approach to characterise viral infections, identify novel viral species, and understand host responses to infections involving complex viral populations (Xu et al., 2015, 2016a, 2016b; Munang'andu et al., 2017). Additionally, NGS approaches can reveal viromes in the absence of overt disease (Li et al., 2015; Shi et al., 2016b, 2018; Geoghegan et al., 2017, 2018).

Understanding the baseline compositions of viral populations, together with investigations into the factors that mediate host jumps, may help reveal key determinants of the process of disease emergence (Geoghegan et al., 2016, 2017; Geoghegan and Holmes, 2017). Metagenomic studies have greatly accelerated the pace of virus discovery, with some individual studies characterising hundreds of new viral species, particularly in non-mammalian species (Essbauer and Ahne, 2001; Batts et al., 2011; Stenglein et al., 2012; Mikalsen et al., 2014; Shi et al., 2016b, 2016a, 2018). Databases that catalogue expanded genetic diversity in particular viral groups or species, along with their host ranges, can be used to develop tools to inform control strategies in the event of a virus jumping hosts between species.

To match this pace of discovery and data production, several virome analysis tools have been developed to aid in the identification of both novel viral species/groups, and to detect known strains of viral species (e.g. Huson et al., 2011; Roux et al., 2014, 2015; Wood and Salzberg, 2014). However, the primary challenge of many of these tools is that they rely on sequence homology to known genomic databases in order to categorise sequences. This combined with the evidence that viromes are characterised by a large amount of sequences with little to no homology to anything in databases (up to 90% - Aggarwala et al 2017), limits many of these tools to the 'completeness' of these databases.

In this Chapter, I present an analysis workflow to detect both known and unknown viruses in short-read RNA-seq datasets. Having validated the workflow against both mock and real datasets, this new approach was used to characterise the viromes of Pacific oyster samples challenged with norovirus and sampled at time points post-infection, revealing 35 putative novel viral genome sequences from 31 putative species. This Chapter also highlights the importance of re-examining datasets previously screened for viruses as large amounts of viruses may exist in the 'unknown fraction' of reads with no homology to anything in the current databases.

5.2 Methods

5.2.1 Workflow overview:

A bioinformatic pipeline was constructed in Snakemake (Köster and Rahmann, 2012) with the goal of reliably identifying RNA viruses in host RNA-seq datasets, together with a series of analysis steps employed individually that annotate and polish viral genome sequences. This section will explain the steps in the Snakemake pipeline and subsequent individual analysis steps that require manual inputs. A visualisation of the workflow including the Snakemake steps and recommended manual curation of resulting viral genomic sequences is presented in Fig. 5.1.

5.2.1.1 Snakemake pipeline

Initially, raw paired-end sequence data is generated or can be downloaded from public databases (e.g. NCBI SRA), which are trimmed of poor quality and short reads using TrimGalore (Krueger, 2015) using a minimum read quality of 30 and a minimum read length of 50bp. Following quality control, reads are mapped to a reference genome of the host organism using BowTie2 (Langmead and Salzberg, 2012), with default parameters to reduce time and memory usage for future analysis steps. All reads that fail to map to the input reference genome are extracted using Samtools (Li et al., 2009) and Seqtk (Li, 2015), are used for *de novo* transcriptome assembly using Trinity RNA-Seq (Grabherr et al., 2011) under default settings. Several RNA-seq *de novo* assembly softwares were assessed on key criteria (quality of species identification and contig length) before Trinity was chosen as the preferred option (see Section 5.2.2). If no host genome is available, then quality-controlled sequence data is used as the input for transcriptome assembly directly without the prior mapping step. The assembled contigs are then annotated based on sequence similarity to the non-redundant protein (Nr) NCBI database using DIAMOND BLASTx (Buchfink et al., 2014) and an expected e-value threshold of $1x10^{-5}$ to remove false positive hits. These steps are included in a Snakemake file (Chapter 5; Supplementary File 1) and the resulting viral contigs can be used in further manual curation of viral genome sequences (outlined below).

5.2.1.2 Annotation of new viral sequences

The assembled contigs resulting from the Snakemake pipeline can be annotated using a range of possible approaches. Here, I explain the approach taken in this Chapter, which led to the characterization of novel viral genomes sequences and therefore has proven applicability (and can be viewed as a recommended practice).

The contigs were assigned taxonomically based on their closest BLASTx hit using the NCBI taxonomy and the R package 'taxonomizr', which assigns taxonomies to NCBI accession numbers. Assembled contigs over 1kb in length with a viral BLASTx match were extracted, subjected to a 6-frame translation to identify putative ORFs, and the protein sequences scanned for viral-specific protein domain signatures using NCBI Conserved Domain Search (Lu et al., 2020). Though several viral-specific protein sequences can be used to identity viral contigs, in this study only contigs containing viral RNA-dependent RNA polymerase (RdRp) proteins were retained for further analysis.

The quality controlled reads and assembled viral contigs were imported into Geneious 2019.0.4 (Kearse et al., 2012) and mapped back to each contig sequence to estimate the relative abundance of each virus, and to elongate viral contigs. Iterative mapping with the Geneious Mapper (Kearse et al., 2012) was performed for each contig (using the 'Medium-

Low' mapping sensitivity), mapping multiple best matches randomly until no further elongation was achieved with further iterations. The resulting alignments were visualised in Geneious and the presence of co-infecting viral strains was assumed when SNV's were detected throughout the genome sequence that would result in ≤95% nucleotide pairwise identity between the two (or more) infecting strains. When detected, manual phasing of viral strains was attempted as automated viral phasing softwares had already proven unreliable in Chapter 4 (see Section 4.2.6). Therefore, manual phasing 'by eye' was deemed to be the most accurate approach. Consensus sequences were generated for each alignment using a '50% - Strict' threshold for calling ambiguous bases and ORFs were then predicted for each sequence using a minimum size of 400bp and a start codon of ATG. Finally contigs were translated according to ORF predictions and the protein sequences were used for viral phylogenetics.

To infer the phylogenetic relationships of novel viruses, viral contigs containing an ORF encoding an RdRp protein were used in BLASTp searches against the non-redundant (Nr) protein NCBI database and the top 100 hits for each contig were retrieved. Duplicated BLASTp hits were removed and sequences containing the same RdRp protein domain family aligned using MAFFT v.7 (Katoh and Standley, 2013; Katoh et al., 2019) employing the E-INS-I algorithm. Low-confidence aligned regions were trimmed using TrimAL 1.2 (Capella-Gutiérrez et al., 2009), employing the 'Strict' method. Maximum likelihood trees were generated using the IQTree online server (Nguyen et al., 2015; Trifinopoulos et al., 2016), which is able to estimate the best-fit amino acid substitution model prior to tree building. 1,000 ultrafast bootstrap alignments (Minh et al., 2013) and 1,000 SH-aLRT branch test replicates (Guindon et al., 2010), were used to acquire confidence scores for each node. Phylogenetic trees were drawn and annotated in FigTree v1.4.4 (https://github.com/rambaut/figtree/releases/tag/v1.4.4).

5.2.2 Pipeline Validation – Test Datasets

The pipeline was validated on several test datasets to establish its efficiency to detect known and unknown viruses. To simulate a virome of known composition, synthetic reads from a selection of 32 different RNA viruses spanning a range of evolutionary distances and genome architectures (Table 5.1) were generated using InsilicoSeq (Gourlé et al., 2019) with the default HiSeq error model settings. InSilicoSeq uses a log-normal distribution to generate abundance estimates of each input genome sequence, and a dataset of 300,000 reads was generated from the 32 virus species genome templates (the number of reads from each species are given in Table 5.2). These synthetic reads were combined with Atlantic salmon (*S. salar*) RNA-seq reads (SRR7139950) to simulate a sample with 3% total viral reads (i.e. a total of 10 million reads), which was analysed with the new pipeline. The coverage of each virus in

the mock dataset is given in Table 5.2. Several *de novo* transcriptome assembly programmes were assessed at this stage to determine the most suitable for viral characterisation including Trinity (Grabherr et al., 2011), SOAPdenovo-Trans (Xie et al., 2014), rnaSPAdes (Bushmanova et al., 2019), MetaSPAdes (Nurk et al., 2017) and Velvet (Zerbino, 2010), the last being implemented in Geneious v2019.0.4. The relative performance of each assembler was judged against the correct identification of each viral species, and the proportion of viral templates covered by a single assembled contig (Table 5.2).

To simulate an infection with a known pathogenic virus, synthetic 2 x 125bp paired-end reads were generated from an SAV1 reference genome template (Accession number JX163854) using InSilicoSeq under the default HiSeq error model settings. This error model and read length was chosen as the host Atlantic salmon reads that these synthetic viral were added to match these specifications, in addition to being of similar length and from the same Illumina platform as the published datasets outlined below. These synthetic reads were then mixed with Atlantic salmon RNA-seq paired-end reads (SRR7139950) for a total of 10 million paired-end reads. Nine datasets were generated in this manner, ranging from SAV titres of 0.05% (5,000 reads) to 5% (500,000 reads) (Table 5.3). These samples were then analysed with the pipeline outlined in Section 5.2.1 using the Trinity RNA-seq assembler. Effectiveness of the pipeline was assessed by whether full length SAV genomes could be generated from all simulated viral titres, and whether the sequences of the assembled viral genomes were identical to the template for synthetic read generation (Table 5.3).

Finally, this pipeline was tested against several published RNA-seq datasets including Atlantic salmon with known ISAV infections (BioProject Accession: PRJNA517818) consisting of 2 x 100bp reads sequenced on an Illumina HiSeq 2000 platform, prepared using the TruSeq Stranded mRNA Sample Preparation Kit (Illumina) (LeBlanc et al., 2018) and a mixed fish sample previously analysed for virome characterisation (SRA accession: SRR6291373) consisting of 2 x 150bp reads sequenced on an Illumina HiSeq 4000 platform, prepared using the Ribo-Zero Gold rRNA Removal Kit (Illumina) followed by the TruSeq total RNA Library Preparation protocol (Illumina) (Shi et al., 2018). The ISAV-infected datasets were analysed according to the pipeline in Section 5.2.1 with reference genome mapping to reduce the proportion of host reads being assembled. The mixed fish sample consists of sequences from species without reference genomes and was therefore analysed without reference genome mapping. The pipeline was assessed by its ability to recover the relevant ISAV genome in all datasets where present (confirmed by mapping trimmed reads to an ISAV reference genome – isolate CA/NL/G0010/2012) (Table 5.4), and asking whether all viruses identified by Shi et al. (2018) were detected, or if any additional new viruses were revealed (Table 5.5).

5.2.3 *De novo* virome characterisation in Pacific oyster

Once validated, as described in Section 5.2.2, the new pipeline was used to characterise the virome of several previously published Pacific oyster datasets, which had been experimentally challenged with norovirus (BioProject PRJNA353875; SRR5043896-99) (Ma et al., 2017). Briefly, as outlined in Ma et al. (2017), wild Pacific oysters were harvested from Aoshanwei, Qingdao, China and kept in control conditions to ensure no natural contamination of norovirus was present. Oysters were then inoculated with norovirus and sampled at specific time points during the infection period (0hr, 12hr, 24hr, and 48hrs post-infection). RNA-seq was then performed on oyster digestive tissues pooled from 5 individuals using poly-A mRNA isolated with oligo-dT beads. Finally, the Illumina HiSeq 2500 platform was used to sequence 2 x 125bp reads from ds-cDNA libraries. To maximise the data available to identify new viruses, reads from all four time points were initially combined and analysed as a single dataset. The Pacific oyster has a published genome (Wang et al. 2012), which was used for read mapping prior to *de novo* assembly. To test the efficacy of host mapping in reducing host reads, the combined dataset was also analysed without host mapping and the relative abundance of taxonomic groups that contributed ≥0.05% of the total contigs in the two assemblies were compared (Fig. 5.12). Virus characterisation was performed as outlined in Section 5.2.1, and new viral species from these datasets, along with those found in the mixed fish sample, were classified using phylogenetics of the RdRp protein sequences. Each timepoint dataset was then analysed individually with this workflow to detect any changes in relative virome composition between the sampling time points (Fig. 5.13).

5.3 Results

5.3.1 Pipeline Validation

The analysis pipeline was validated against several mock and real datasets to determine its usefulness in identifying both novel and characterised viruses. Initially, a mock virome consisting of synthetic paired-end reads from 32 viruses, possessing a range of genome sizes, structures and phylogenetic positions (Table 5.1), was analysed independently and combined with 'host' reads (Atlantic salmon). Six de novo assembly programmes were tested for their ability to assemble contigs identifiable as specific virus species, and the proportion of virus template sequences covered in single contigs. Trinity, RNA-Spades, MetaSpades and Spades were able to detect all the virus species present in the mock viromes and assemble contigs that covered the majority of the genome templates (>95% of the genomic template) (Table. 5.2), with Trinity having a slightly better performance than the other programmes. SOAPdenovo-Trans, while comparable to the former four assemblers for most viral species, failed to identify one species entirely, and produced contigs that only covered half of another species. Velvet

performed comparatively poorly, missing several species and producing contigs only partially covering templates for several other species (Table 5.2). Therefore, Trinity was chosen as the most suitable programme for use in this pipeline and for all future analyses in this Chapter.

When tested against a range of mock SAV infections, the pipeline successfully identified SAV contigs in all simulated titres (0.05% to 5% of total reads), and full length genomes were recovered in single contigs in all datasets (Table 5.3). When tested against real RNA-seq datasets of ISAV-infected Atlantic salmon samples (LeBlanc et al., 2018), the pipeline identified ISAV contigs in all the high-titre datasets (Table 5.4). While no ISAV was identified in the low-titre samples, mapping the reads to the ISAV genome using BWA-Mem (Li, 2013) also failed to identify ISAV in these samples. Therefore this pipeline was deemed at least as sensitive and accurate as a mapping approach. Additionally, most ISAV segments were assembled into a single contig, thus confirming the ability of Trinity to produce full-length contigs of segmented virus genomes (Table 5.4).

The mixed-fish sample previously used for RNA virus characterisation by Shi et al. (2018) was analysed without reference genome mapping, as the species contributing to this dataset lack publicly available genomes. Fourteen viral species were previously identified in this dataset (Shi et al., 2018) and the new pipeline identified all of these species. Additionally, twenty putative novel viral species were also identified (Table 5.5) belonging to eight different viral families and each novel virus was classified using phylogenetics based on RdRp protein sequences (Figs. 5.2-5.11), including a relatively high-abundance virus responsible for 1.236% of the total reads in the dataset (SCS picorna-like virus 2 - Table 5.5). Potential explanations for the identification of these new viruses in this study and not in the original are elaborated on in Section 5.4.

5.3.2 Pacific Oyster virome characterisation

Four RNA-seq datasets representing the digestive organs of five pooled Pacific oysters experimentally infected with norovirus (Ma et al., 2017) were analysed to detect novel viral sequences. After analysis with the pipeline and manual validation, fifteen putative novel viral sequences belonging to thirteen putative novel species were identified based on amino acid homology to known viral sequences (Table 5.5) and taxonomically classified using phylogenetics based on RdRp sequences (Figs. 5.2-5.11).

The relative abundance of major taxonomic groups represented in the assembled contigs for both ‘combined’ datasets were compared to assess the efficacy of mapping to the host genome before assembling (Fig. 5.12). Overall, mapping reduced the proportion of contigs that had homology to a bivalve protein by over 10% (67.2% when unmapped; 55.5% when mapped) and increased the proportion of contigs that had no BLASTx hit in the Nr protein database by

8% (31.9% to 40.1%). Additionally, the proportion of contigs mapping to other taxonomic groups, including viruses and bacteria, increased by more than four-fold (0.95% to 3.84%), with viral contigs increasing by almost five-fold in prevalence (0.16% to 0.81%).

The trimmed reads from each dataset representing the time points of the experiment (0hrs, 12hrs, 24hrs & 48hrs) were mapped back to the new viral sequences along with a genome sequence of norovirus (KC175323) and relative abundance (the % of total reads mapping exclusively to each contig) of each virus was estimated (Fig. 5.13). As expected, at time point 0 (T0), norovirus was completely absent in this dataset while being abundant at the other three time points. While several of the novel viral species had low abundance in all time points of the experiment, two particular new viral sequences 'Qingdao picorna-like virus 14' and 'Qingdao picorna-like virus 15' were present at high levels at T0 but dropped to undetectable levels at further time points (Fig. 5.13).

5.4 Discussion

This Chapter shows that the new metagenomics pipeline I developed is useful for detecting both known and unknown viruses in RNA-seq datasets. It has been widely reported in both viral and bacterial metagenomics studies that the choice of sequencing method and analysis strategy (including assembly method) have a significant impact on the accuracy of the reconstructed metagenome (Mavromatis et al., 2007; Lindgreen et al., 2016; Greenwald et al., 2017; Vollmers et al., 2017; Sutton et al., 2019). Viral metagenomics assembly is particularly complex due to many features of viral genomes that inhibit assembly including within-host strain variation (Roux et al., 2017), hypervariable genomic regions (Warwick-Dugdale et al., 2019), and high proportions of repeat regions within viral genomes (Minot et al., 2012). While the use of highly accurate short read sequencing technologies (i.e. Illumina) are the gold standard for virome and microbiome studies, there are challenges with using such data types in these studies. The short length of the reads inhibit full length assembly of complex genome types as repeat regions often fail to assemble accurately, and assembly of closely related viral species or strains in the same sample is highly challenging as the identifying SNPs are often further apart than the maximum read length. The advent of newer, long-read sequencing technologies would greatly help the contiguous assembly of viruses in these challenging sample. However while the per-base accuracy is steadily increasing, and can be greatly enhanced with approaches such as unique molecular identifiers (UMI – see Section 6.2.1), the relative lower read quality compared to Illumina sequencing remains a challenge for accurate assembly and strain phasing of viruses.

With that in mind, the use of both artificial and real virome datasets in benchmarking my short-read virome analysis pipeline allowed for reliable inferences. The pipeline presented in

this Chapter proved to be at least as sensitive at detecting known viruses as a mapping-based approach using mock viromes (Table 5.2), mock SAV infections (Table 5.3), and real ISAV-infected Atlantic salmon RNA-seq datasets (Table 5.4), thus conclusively demonstrating its utility for applications with additional sample types.

In this study, 35 putative novel viral sequences belonging to 31 potential species were identified from publicly available RNA-seq datasets. While four of the datasets from Pacific oyster (BioProject PRJNA353875; SRR5043896-99) had not been screened for novel viruses before and contained 15 of the new viral sequences, a dataset previously used for virome analyses (SRR6291373) (Shi et al., 2018) yielded 20 new viral sequences (Table 5.5). While many of the new sequences belong to the Picornavirales order, including all the viruses identified in BioProject PRJNA353875, there were several other viral lineages represented in these new sequence (Figs. 5.2-5.11). Additionally, the novel sequences showed a wide range of divergence compared to previously annotated viruses, with several sharing ≤30% amino acid identity to their closest BLASTx match (Table 5.5). This included SCS toti-like virus 1, whose closest known relative (Beihai barnacle virus 15) shares only a 30% amino acid pairwise identity to this new virus, and was itself only identified by Shi et al. (2018). Such a high level of sequence divergence in viruses identified from previously characterised datasets highlights the potential for identifying novel viral species from publicly available datasets, even when they have already been characterised in published work. As viral genome databases become more densely populated, newer and more diverged viral lineages are likely to become 'visible' in existing datasets on the basis of homology. Consequently, I advocate that regularly re-examining publicly available datasets offers a valuable source of untapped information in ongoing efforts to understand the evolutionary landscape of RNA viruses.

The Pacific oyster datasets (Ma et al., 2017) were analysed for changes in virome composition at different time points following norovirus challenge (Fig. 5.13). However as biological replicates were not sequenced in the original study, no tests could be performed for statistically significant changes in virome composition. Predictably, norovirus was not present at T0 and was at relatively high levels for the other time points. Most of the novel viral species showed little change over the time course, with some having a slight increase in prevalence at T12, but similar titres at both earlier and later time points (Fig. 5.13). Due to the fact that the T12 dataset is over twice the size of T0 and T48, it is therefore possible that the slight increase in several of these species is due to the specific library and not due to fluctuations in virome composition. However two new putative species, Qingdao picorna-like virus 14 and Qingdao picorna-like virus 15, showed particularly high titres at T0, but were largely undetectable at all later time points. This finding is parsimonious with legitimate infections that the oyster cleared itself of during the course of the challenge experiment, or an infection of an organism

in the oyster's microbiome that did not survive the experiment itself. Previous work has shown that duration of oysters in clean water environments can help purge biological contaminants (Son and Fleet, 1980; Sobsey et al., 1987; de Abreu Corrêa et al., 2007), though the value of such practices in removing viral particles is disputed (Gill et al., 1983; Loisy et al., 2005; Ueki et al., 2007; Nappier et al., 2008). As samples were not taken from this oyster population before being purged in laboratory conditions (i.e. representing the natural microbiome of the oysters), it is impossible to say if these two high titre viruses were present at the beginning of the experiment, what prevalence they were originally present at, and the reason for their apparent removal from the oysters.

To assess the effect of mapping to a host genome on the resulting microbiome, the combined Pacific oyster datasets were analysed both with and without host mapping (Fig. 5.12). While mapping the reads to the host genome before assembly did reduce the overall proportion of bivalve (proxy for host) contigs from 67% to 55.5%, and the total number of contigs from 174,294 to 45,492 (> three-fold reduction), over half the assembled contigs still had an estimated bivalve origin. This finding indicates that while mapping does reduce the total number of host reads, it is not an infallible method, and is highly dependent on the quality and 'completeness' of the host reference genome. However this reduction in host contigs did result in an overall 4.9x increase in the relative proportion of viral contigs and a 1.6x increase in bacterial contigs. Additionally, the proportion of contigs with no known match in the Nr NCBI protein database increased from 32% to 40%, representing >20,000 contigs in the dataset that mapped to the host genome. These 'unknown' contigs represent both novel, uncharacterised species as has been seen before in other studies (Marhaver et al., 2008; McDaniel et al., 2008; Yin and Fischer, 2008; Aggarwala et al., 2017), but also RNA sequences from the host missed during annotation or absent from the genome assembly. This has been noted in other study systems where improvements in data quality, data quantity, and assembly and annotation softwares have identified genes absent from earlier assembly drafts (Florea et al., 2011; Denton et al., 2014), and it is likely that improvements in the *C. gigas* genome would reduce the quantity of contigs without a known hit in the Nr protein database. This approach of assigning taxonomies to contigs (and indeed raw reads) can also be used to identify lab-based contamination sources. In both the host mapped and unmapped datasets, contigs matching mammalian proteins, and specifically human proteins, were in relatively high abundance (1.86% and 0.35% respectively). While it is possible some might be legitimate contamination from human effluent entering estuaries and bays, it is at least as likely that contamination during sample and library preparation are the sources of the human contigs. Finally while mapping to the host genome did not in this case remove all the host reads from the sample, it

did significantly reduce the amount of reads for assembly, and accordingly reduced the number of contigs produced.

To summarise, this Chapter presents a useful analysis pipeline to characterise both known and novel viruses in RNA-seq datasets, regardless of the original purpose of the data. While there are several established analysis tools to characterise viromes in RNA-seq samples (e.g. Huson et al., 2011; Roux et al., 2014, 2015; Wood and Salzberg, 2014), many of these are highly effective at identifying known viruses, but are challenged with identified highly divergent novel species. This Chapter also points to the potential usefulness of analysing datasets that have previously been screened for viruses, as updates to viral databases can enable the identification of highly divergent sequences that may have lacked sufficient sequence homology to databases at the time of the original screening.

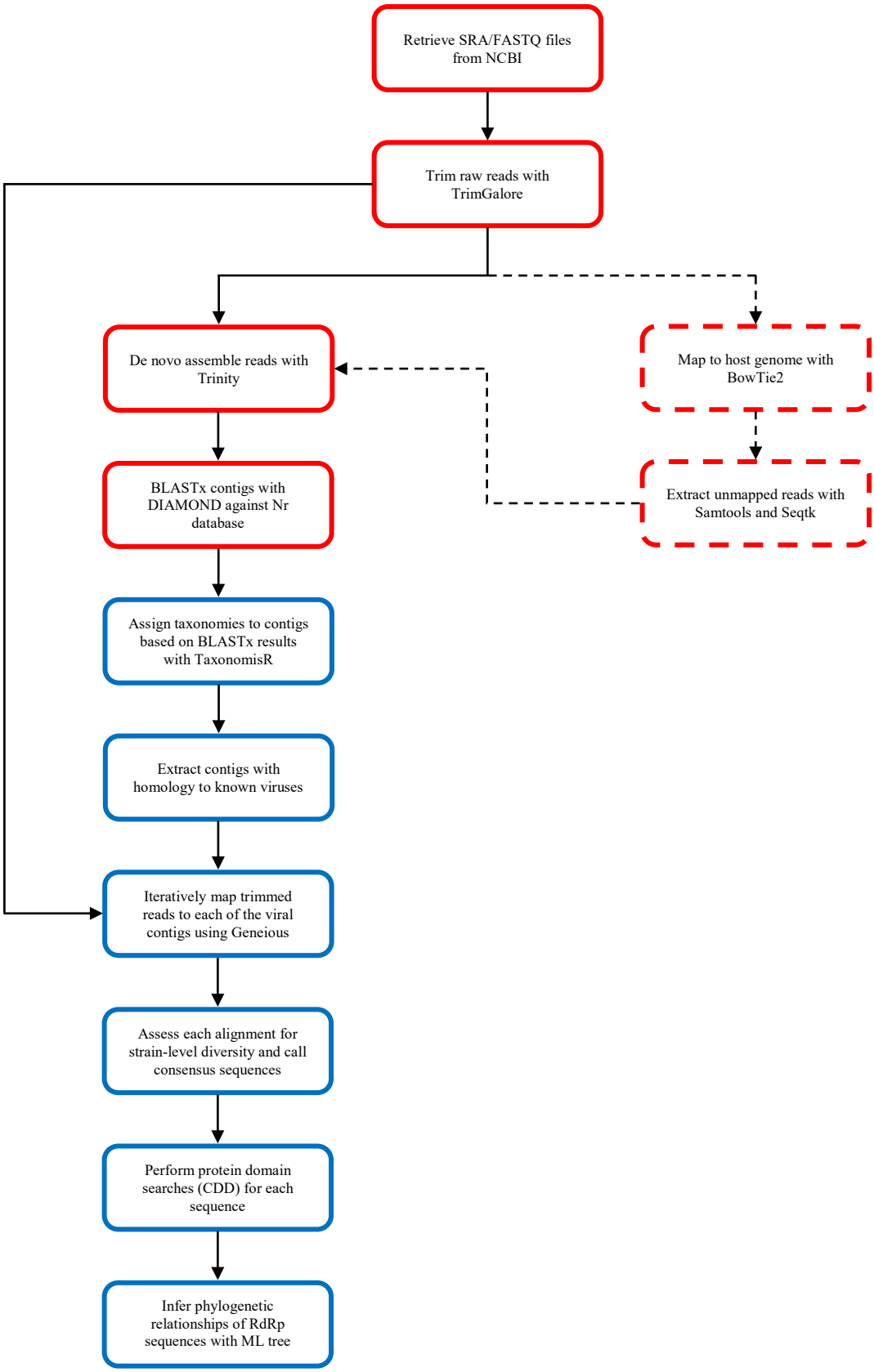

Figure 5.1. Visualisation of the new virus discovery workflow. Steps taken as part of the Snakemake pipeline are highlighted in red, and steps taken individually are highlighted in blue. Hashed arrows and boxes indicate optional steps to be included when there is a high-quality reference genome for the host species.

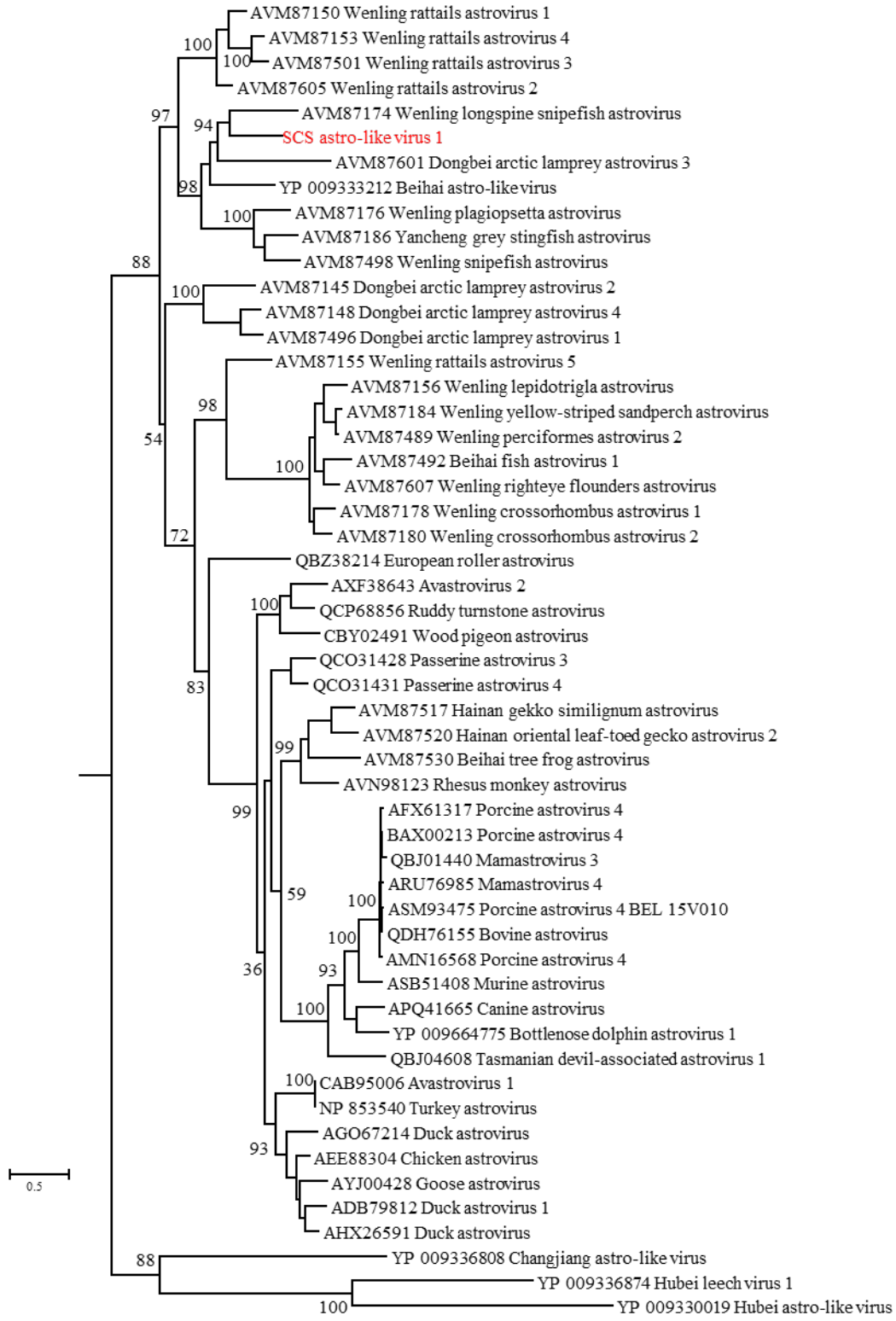

Figure 5.2. Maximum likelihood phylogeny of the "Astroviridae" clade. For each phylogeny presented in this Chapter, the names of new viral sequences identified from both the Pacific oyster samples and the Shi et al. (2018) dataset are coloured red. Viral RdRp sequences related to those identified in this study were identified by BLASTx against the Nr protein database. The top 100 non-redundant matching hits were retrieved and aligned with MAFFT v.7 and trimmed using trimAL v1.2. The ML analysis was performed on the IQTREE server including estimation of the best-fit substitution model (*LG*+I+G4; according to the Bayesian Information Criterion), with ultrafast bootstrapping done to assess confidence in the branching patterns.

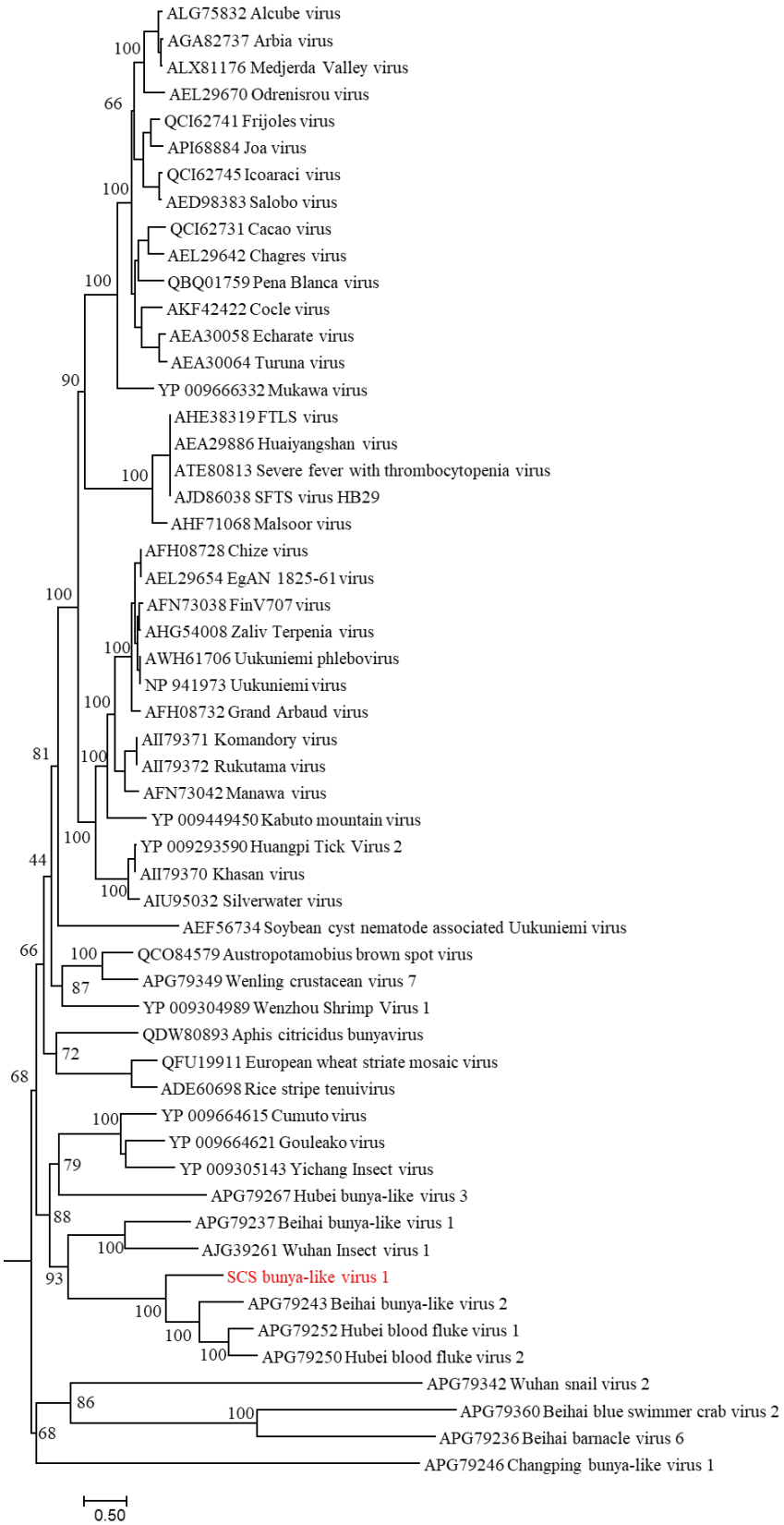

Figure 5.3. Maximum likelihood phylogeny of the "Bunyaviridae" clade. Figure legend follows Figure 5.2 with the best-fit substitution model of LG+F+I+G4.

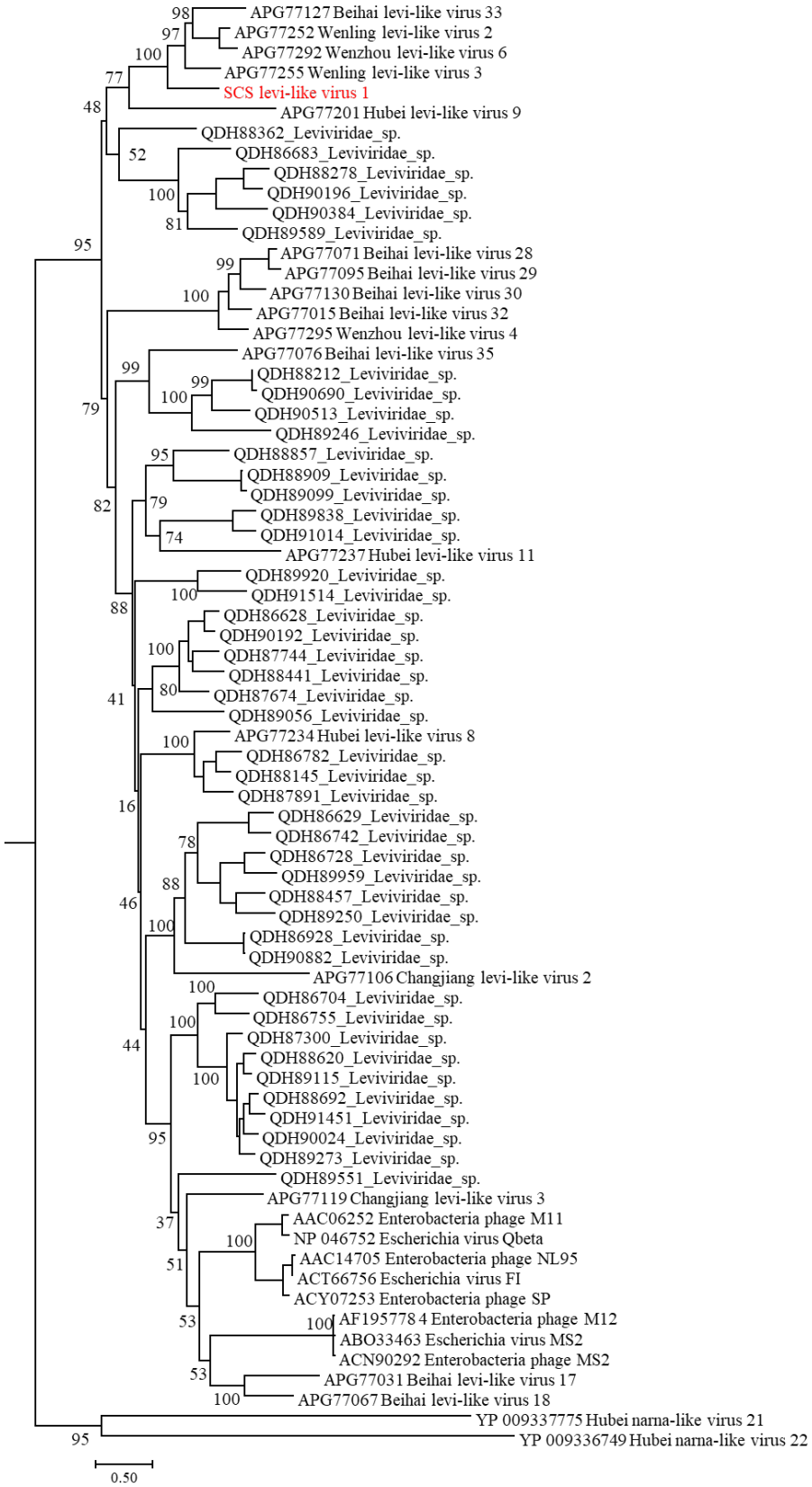

Figure 5.4. Maximum likelihood phylogeny of the "Leviviridae" clade. Figure legend follows Figure 5.2 with the best-fit substitution model of rtREV+F+I+G4

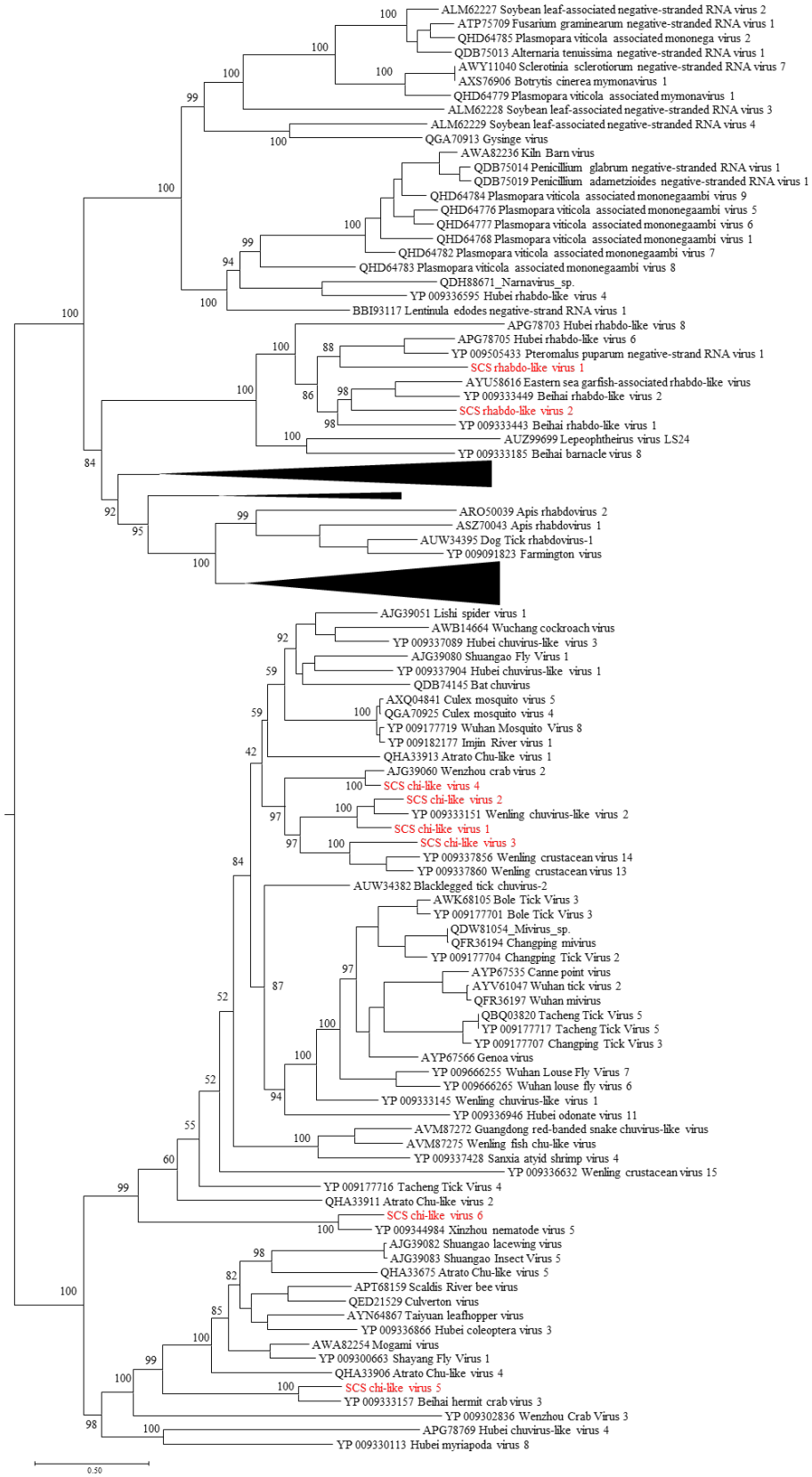

Figure 5.5. Maximum likelihood phylogeny of the "Mononegavirales-Chuviridae" clade. Figure legend follows Figure 5.2 with the best-fit substitution model of LG+F+I+G4.

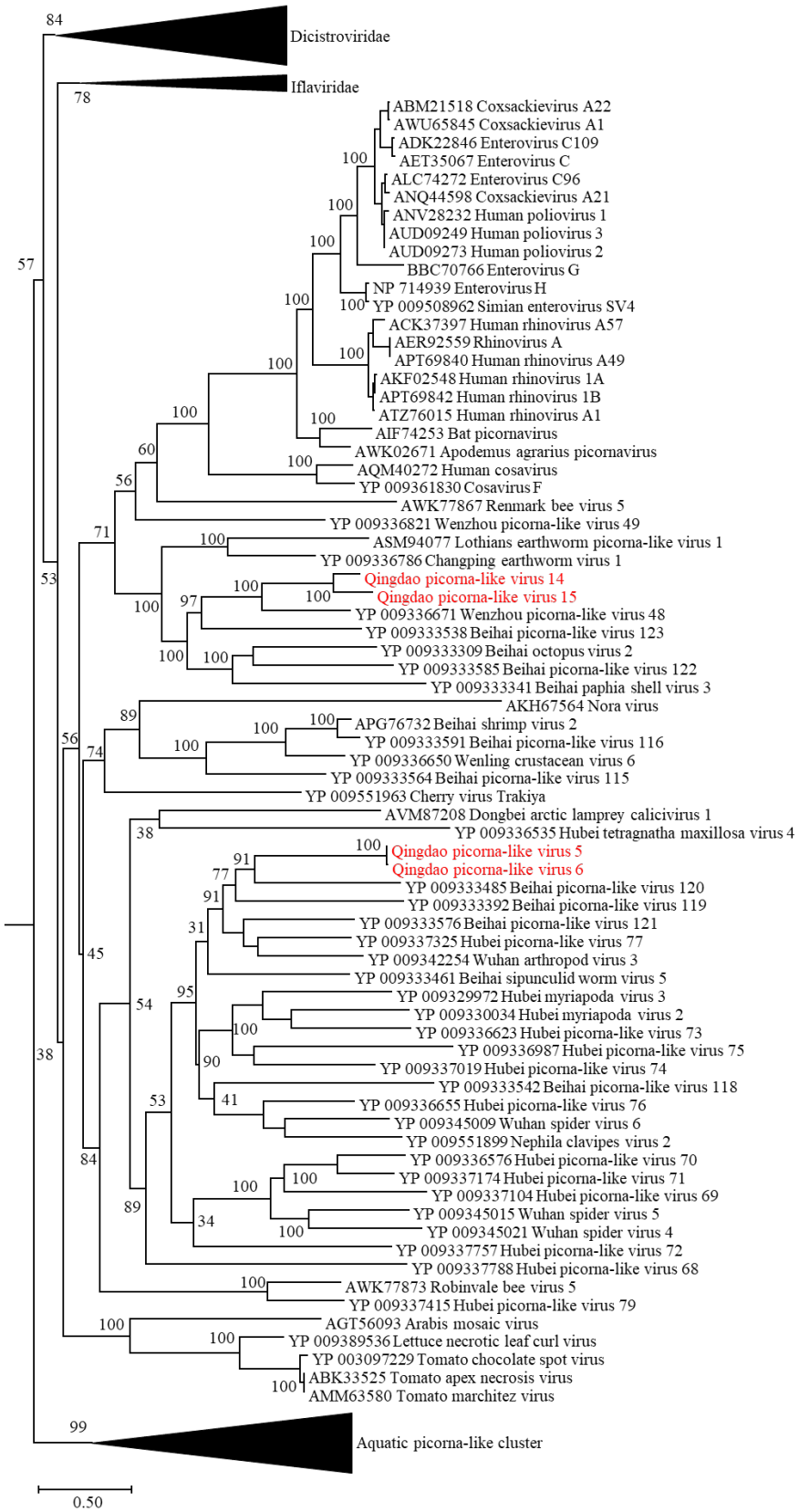

Figure 5.6. Maximum likelihood phylogeny of the 'Picornaviridae' clade. Figure legend follows Figure 5.2 with the best-fit substitution model of LG+F+I+G4.

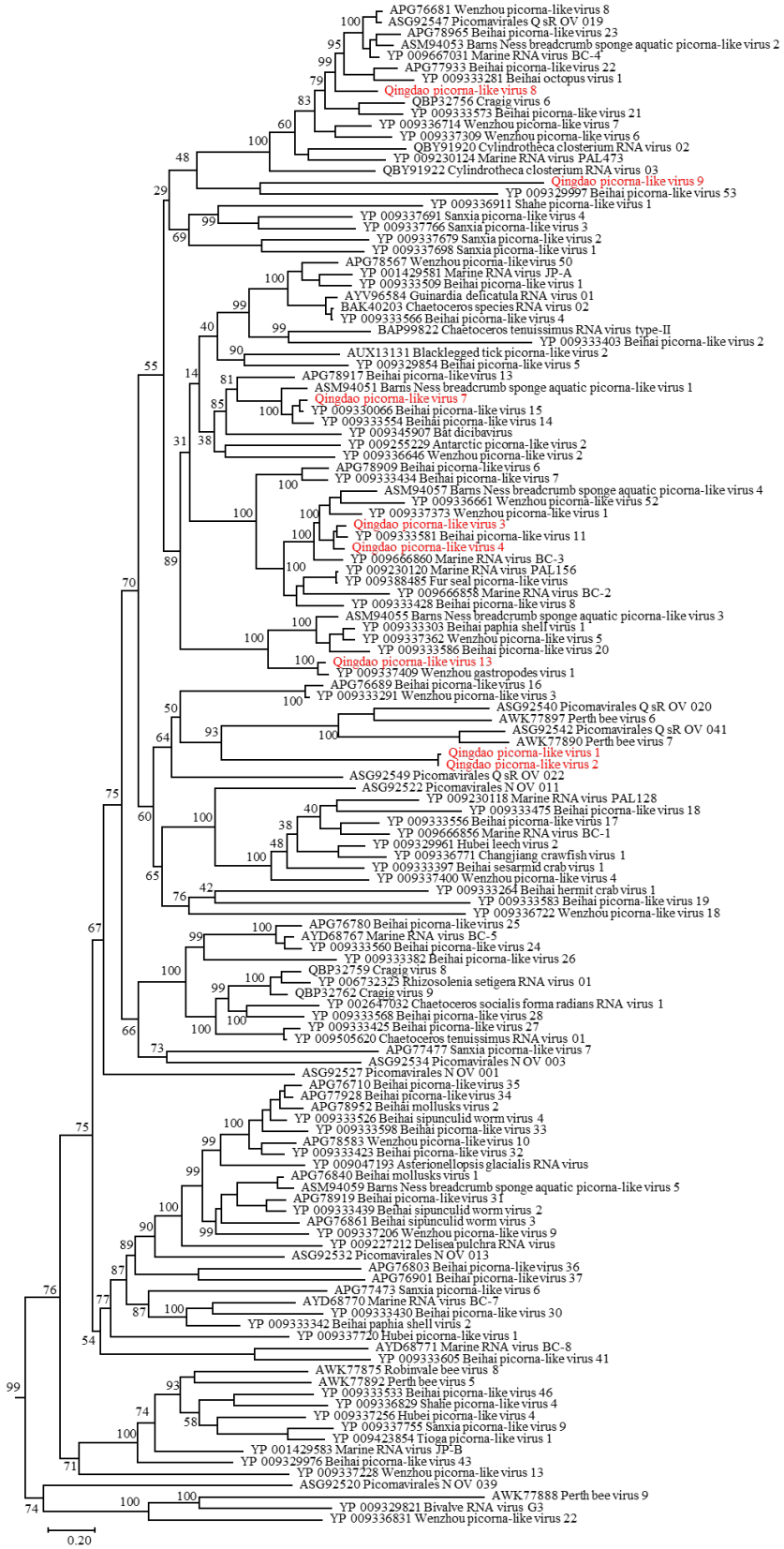

Figure 5.7. Maximum likelihood phylogeny of the "Aquatic picorna-like cluster" within the "Picornaviridae" clade. Figure legend follows Figure 5.1 with the best-fit substitution model of LG+F+I+G4.

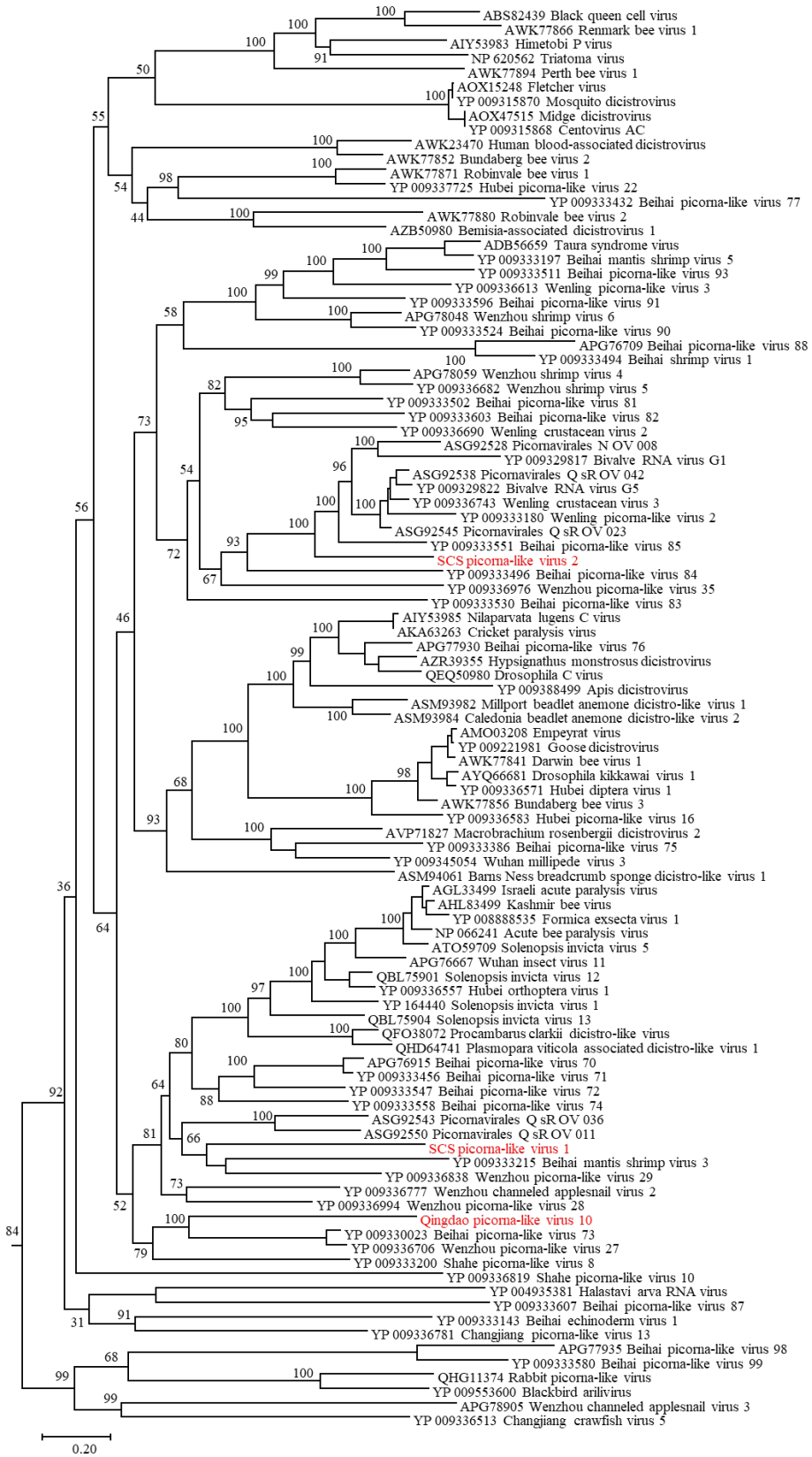

Figure 5.8. Maximum likelihood phylogeny of the "Dicistroviridae" clade of the "Picornaviridae" clade. Figure legend follows Figure 5.1 with the best-fit substitution model of LG+F+I+G4

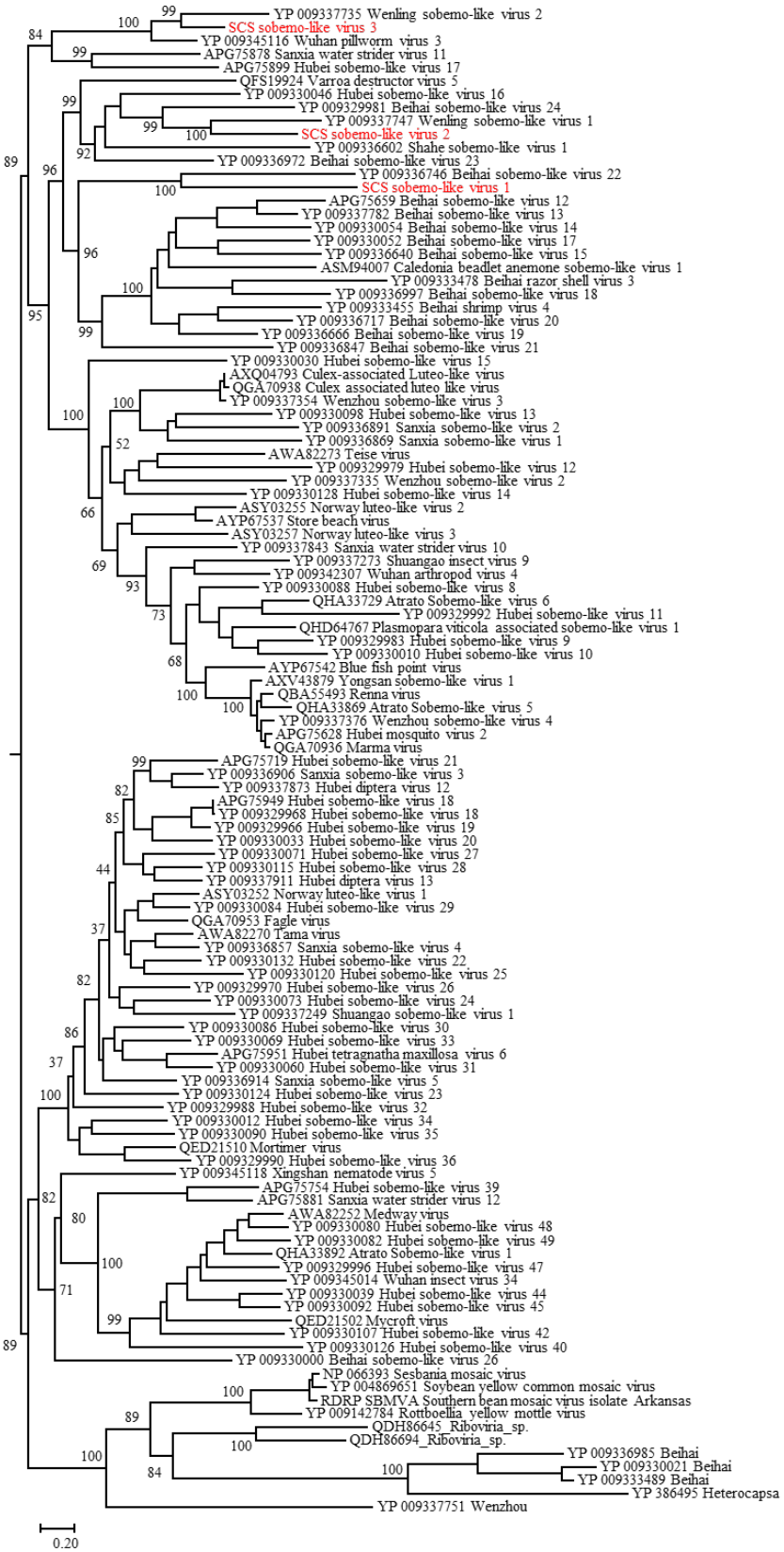

Figure 5.9. Maximum likelihood phylogeny of the "Sobemoviridae" clade. Figure legend follows Figure 5.1 with the best-fit substitution model of rtREV+F+I+G4.

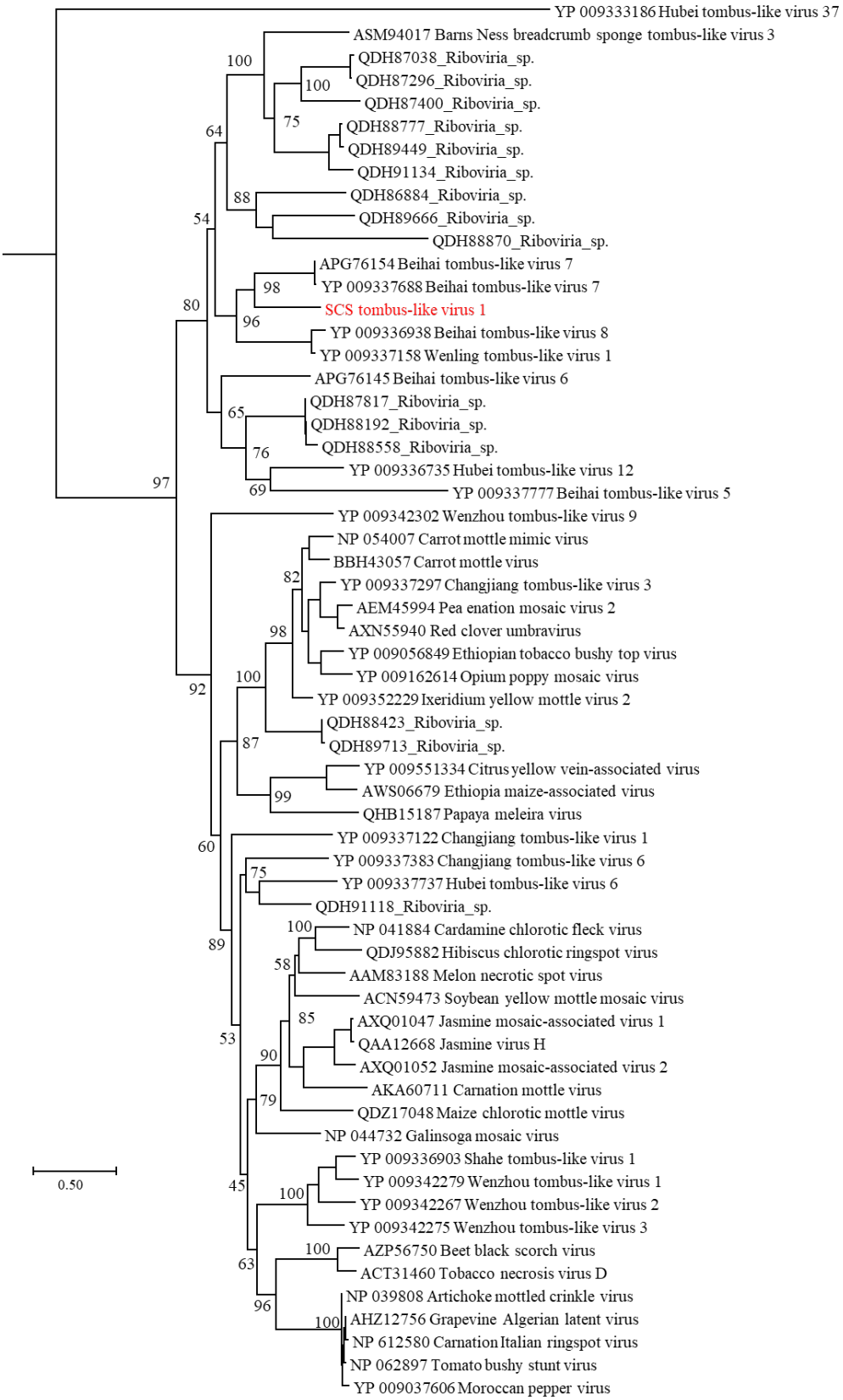

Figure 5.10. Maximum likelihood phylogeny of the "Tombusviridae" clade. Figure legend follows Figure 5.1 with the best-fit substitution model of LG+F+G4

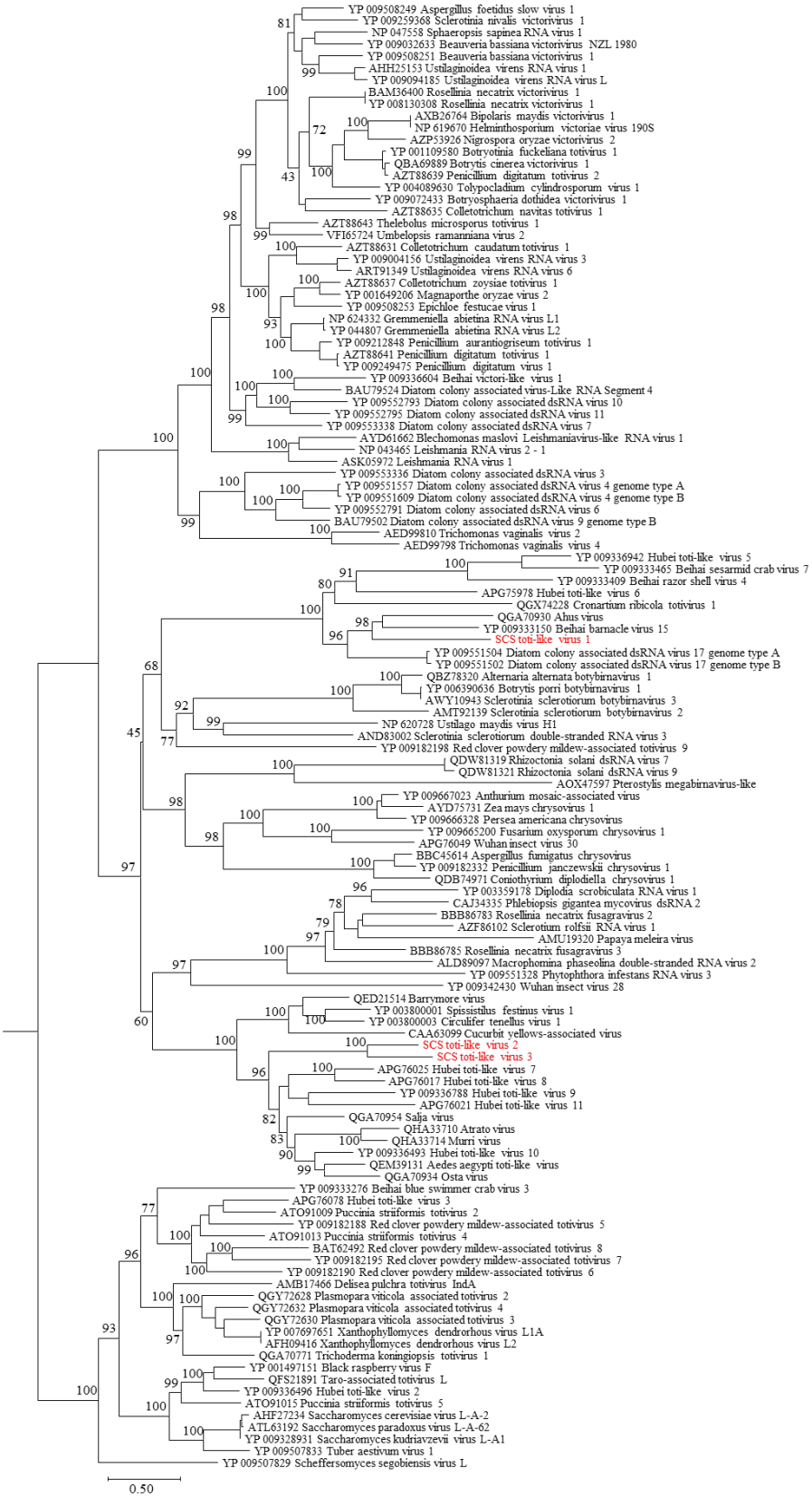

Figure 5.11. Maximum likelihood phylogeny of the "Totiviridae" clade. Figure legend follows Figure 5.1 with the best-fit substitution model of LG+F+I+G4.

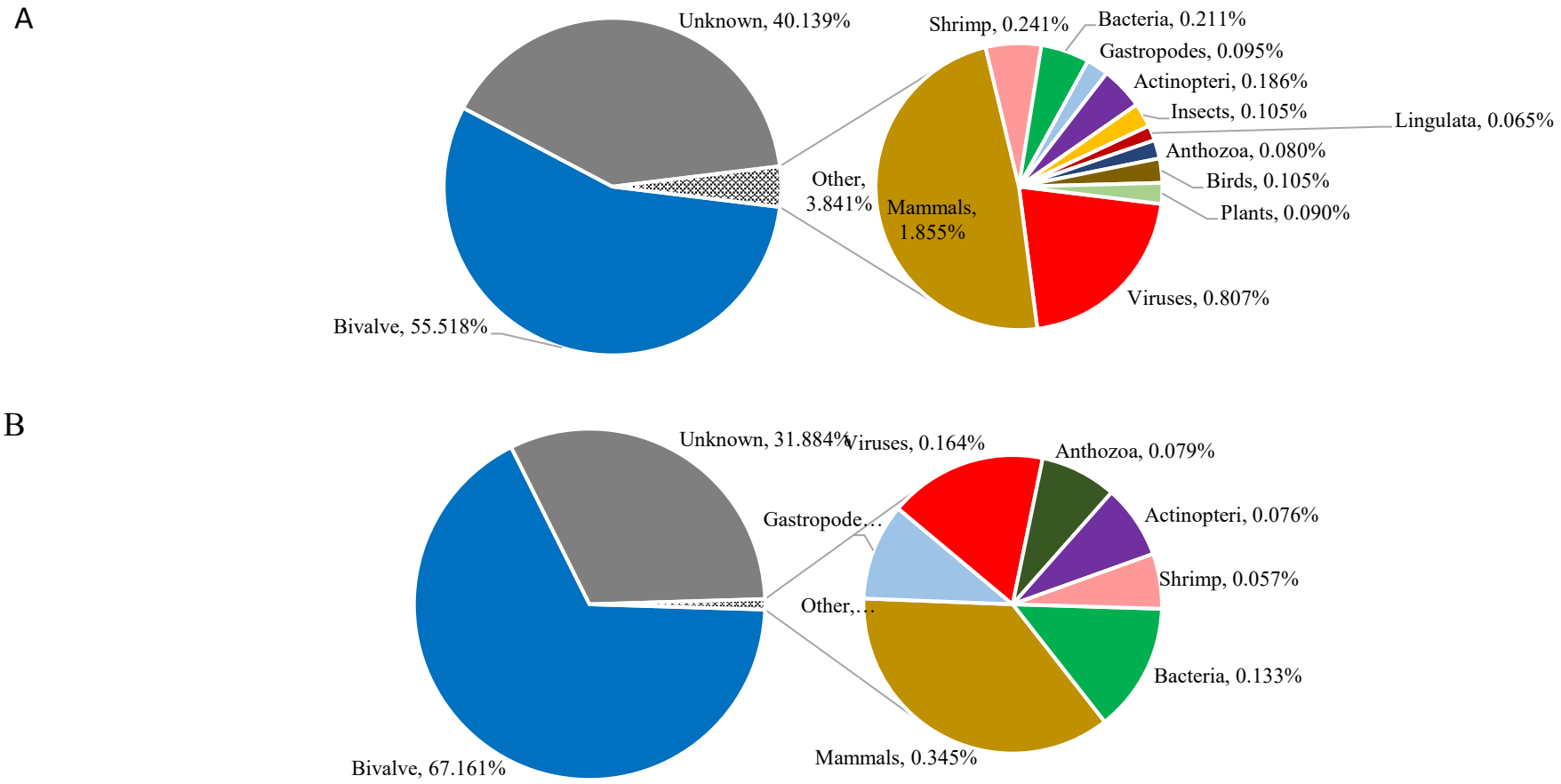

Figure 5.12. Taxonomic classification of assembled contigs from the Pacific oyster samples (BioProject PRJNA353875) against the Nr database (NCBI). Contigs classified as 'Unknown' had no clear homology with sequences in the Nr database. Only taxonomic groupings contributing ≥0.05% of total contigs were included. A) Taxonomic classification after mapping trimmed reads to the host genome (C. gigas). B) Taxonomic classification without mapping reads to the host genome.

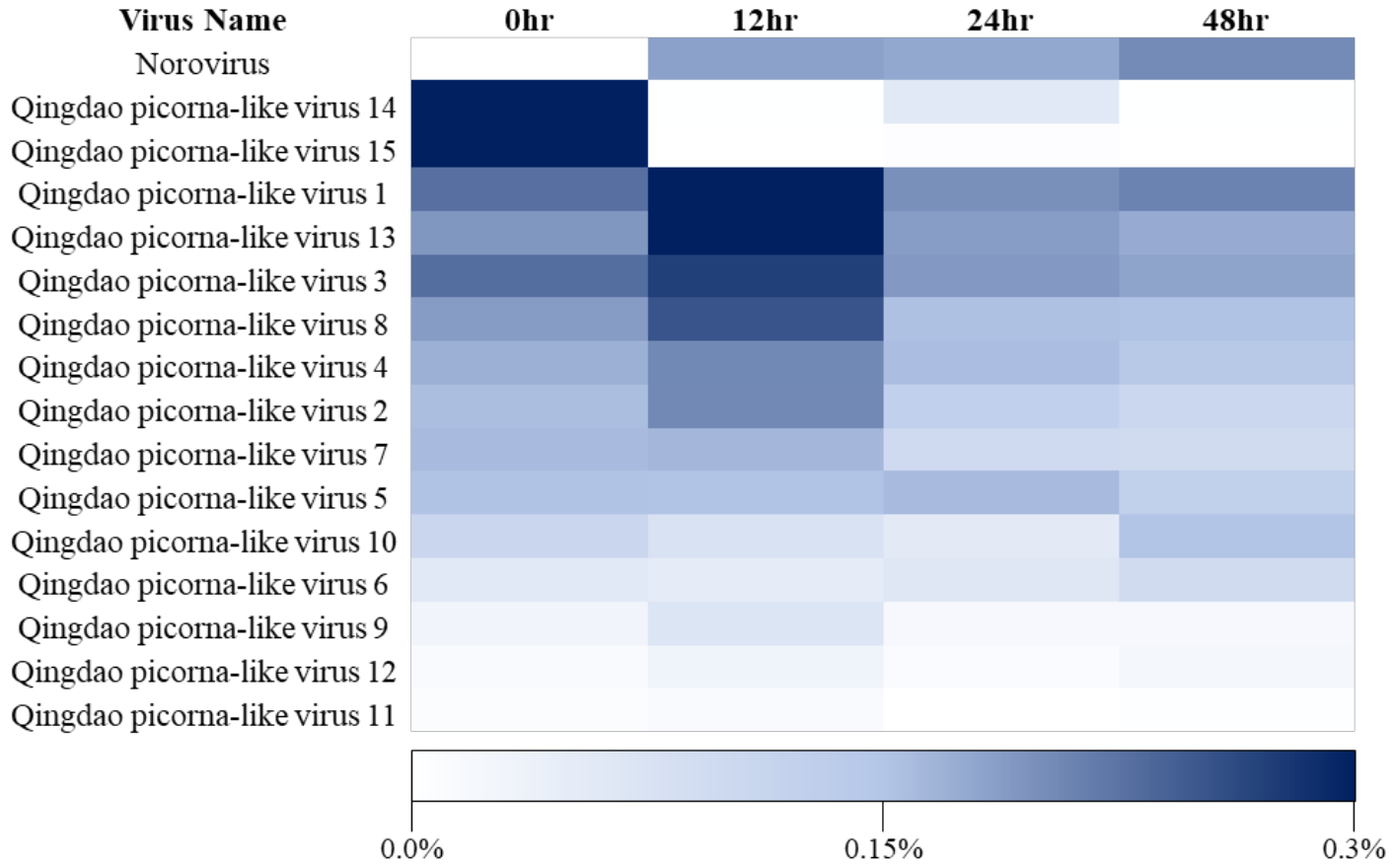

Figure 5.13. Heat map of the relative proportion of reads for different 16 viral species at each time-point of the norovirus infection experiment. Data are reported as % of total reads per dataset mapping to each viral genome exclusively as indicated by the coloured scale. Viruses are named according to Table 3.

Table 5.1 Viral species included in the mock virome.

Classification	Family	Virus	Genome Structure	Length (kb)
(-)ssRNA	*Arenaviridae*	Lassa mammarenavirus	Bipartite	10.6
(-)ssRNA	*Qinviridae*	Beihai yingvirus	Bipartite	7.4
(-)ssRNA	*Deltavirus*	Hepatitis delta virus	Circular	1.7
(-)ssRNA	*Filoviridae*	Zaire ebolavirus	Monopartite	19
(-)ssRNA	*Paramyxoviridae*	Measles morbillivirus	Monopartite	16.7
(-)ssRNA	*Rhabdoviridae*	Rabies lyssavirus	Monopartite	13.1
(-)ssRNA	*Orthomyxoviridae*	Isavirus	Octopartite	12
(-)ssRNA	*Aspiviridae*	Citrus psorosis ophiovirus	Tripartite	11.2
(-)ssRNA	*Hantaviridae*	Hantaan orthohantavirus	Tripartite	11.8
(-)ssRNA	*Peribunyaviridae*	Bunyamwera orthobunyavirus	Tripartite	12
(+)ssRNA	*Alphaflexiviridae*	Indian citrus ringspot virus	Monopartite	7.5
(+)ssRNA	*Arteriviridae*	Alphaarterivirus equid	Monopartite	12.7
(+)ssRNA	*Betaflexiviridae*	Apple chlorotic leaf spot virus	Monopartite	7.5
(+)ssRNA	*Coronaviridae*	Middle Eastern Respiratory syndrome coronavirus	Monopartite	30
(+)ssRNA	*Iflaviridae*	Infectious flacherie virus	Monopartite	9.6
(+)ssRNA	*Marnaviridae*	Heterosigma akashiwo RNA virus	Monopartite	8.6
(+)ssRNA	*Mesoniviridae*	Alphamesonivirus 1	Monopartite	20
(+)ssRNA	*Okavirus*	Gill-associated virus	Monopartite	26
(+)ssRNA	*Picornaviridae*	Salivirus A	Monopartite	8
(+)ssRNA	*Tymoviridae*	Turnip yellow mosaic virus	Monopartite	6.3
dsRNA	*Reoviridae*	Aquareovirus A	11 Segments	23
dsRNA	*Birnaviridae*	Infectious Pancreatic necrosis virus	Bipartite	5.5
dsRNA	*Megabirnaviridae*	Rosellinia necatrix megabirnavirus 1	Bipartite	16.1
dsRNA	*Partitiviridae*	Cryptosporidium parvum virus 1	Bipartite	3.3
dsRNA	*Picobirnavirus*	Human picobirnavirus	Bipartite	4.2
dsRNA	*Amalgaviridae*	Blueberry latent virus	Monopartite	3.4
dsRNA	*Endornaviridae*	Cluster bean endornavirus 1	Monopartite	12.9
dsRNA	*Hypoviridae*	Cryphonectria hypovirus 1	Monopartite	12.7
dsRNA	*Totiviridae*	Saccharomyces cerevisiae virus L-A	Monopartite	4.5
dsRNA	*Chrysoviridae*	Penicillium chrysogenum virus	Quadripartite	12.5
dsRNA	*Quadriviridae*	Rosellinia necatrix quadrivirus 1	Quadripartite	16.8
dsRNA	*Cystoviridae*	Pseudomonas phage phi6	Tripartite	14

Table 5.2. Mock virome assembly comparison results. Viral species were mixed in variable proportions (the number of reads) and assembled with each of the de novo assembly programmes listed. Statistics reported are the % of the template (genome or segment) assembled in a single contig. Results <90% are highlighted in red. Table continued on next page

Virus Name	Accession	# of Reads	Mean Coverage	SOAPdenovo-Trans	Velvet	Trinity	RNA-SPAdes	MetaSPAdes	SPAdes
Alphaarterivirus equid	GQ903794	6,025	59.62	99.90%	99.90%	100.00%	100.00%	100.00%	100.00%
Alphamesonivirus 1	KC807167	24,451	148.31	99.90%	100.00%	100.00%	100.00%	100.00%	100.00%
Apple chlorotic leaf spot virus	AB326225	35,930	598.85	100.00%	100.00%	100.00%	100.00%	99.90%	100.00%
Aquareovirus A	AF418304	3,886	610.92	99.90%	99.90%	99.90%	99.90%	99.90%	99.90%
Aquareovirus A	AF418303	6,649	841.28	99.90%	69.50%	99.90%	99.90%	99.90%	99.90%
Aquareovirus A	AF418302	9,114	1021.12	99.90%	88.50%	99.90%	99.90%	99.90%	99.90%
Aquareovirus A	AF418301	9,541	912.72	99.90%	99.90%	99.90%	93.50%	99.90%	99.90%
Aquareovirus A	AF418300	19,376	1749.27	99.90%	93.00%	99.90%	99.90%	99.90%	99.90%
Aquareovirus A	AF418297	44,805	3493.47	99.90%	53.70%	99.90%	99.90%	99.90%	99.90%
Aquareovirus A	AF418299	1,447	88.79	99.80%	99.90%	100.00%	100.00%	100.00%	100.00%
Aquareovirus A	AF418298	7,414	416.68	99.90%	100.00%	100.00%	100.00%	100.00%	100.00%
Aquareovirus A	AF418296	6,247	213.28	99.80%	99.90%	99.90%	99.90%	99.90%	99.90%
Aquareovirus A	AF418295	3,110	101.34	99.80%	99.90%	99.90%	99.90%	99.90%	99.90%
Aquareovirus A	AF418294	72,728	2324.98	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Beihai yingvirus	KX883991	3,332	238.54	99.90%	99.90%	99.90%	99.90%	99.90%	99.90%
Beihai yingvirus	KX883990	10,372	232.57	99.90%	100.00%	100.00%	100.00%	99.90%	100.00%
Blueberry latent virus	AB608991	952	34.93	99.00%	0.00%	99.50%	99.50%	99.50%	99.50%
Bunyamwera orthobunyavirus	D00353	19,348	2504.79	99.90%	87.10%	99.90%	99.90%	99.90%	99.90%
Bunyamwera orthobunyavirus	M11852	1,142	32.28	99.50%	0.00%	99.70%	99.70%	99.70%	99.70%
Bunyamwera orthobunyavirus	X14383	21,101	386.78	100.00%	100.00%	100.00%	100.00%	99.90%	100.00%
Citrus psorosis ophiovirus	AY654894	18,357	1599.07	99.90%	81.70%	99.90%	99.90%	99.90%	99.90%
Citrus psorosis ophiovirus	AY654893	9,073	694.90	99.90%	99.90%	99.90%	99.90%	99.90%	99.90%
Citrus psorosis ophiovirus	AY654892	3,686	56.74	99.70%	77.90%	99.50%	99.50%	99.50%	99.50%
Cluster bean endornavirus 1	MG764084	6,205	60.62	99.80%	99.90%	99.90%	99.90%	99.90%	99.90%
Cryphonectria hypovirus 1	KY471627	18,149	179.57	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Cryptosporidium parvum virus 1	KY884721	11,581	966.92	99.90%	99.90%	99.90%	99.90%	99.90%	99.90%
Cryptosporidium parvum virus 1	KY884720	33,534	2303.51	99.90%	85.20%	99.90%	93.80%	99.90%	99.90%
Gill-associated virus	AF227196	9,343	44.84	100.00%	0.00%	100.00%	100.00%	100.00%	100.00%
Hantaan orthohantavirus	M14626	16,560	1231.65	99.90%	99.90%	99.90%	99.90%	99.90%	99.90%
Hantaan orthohantavirus	M14627	7,354	256.26	99.90%	99.90%	99.90%	99.90%	99.90%	99.90%
Hantaan orthohantavirus	X55901	17,351	334.67	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Hepatitis delta virus	D01075	2,326	174.25	99.60%	99.80%	99.80%	99.80%	99.80%	99.80%

Virus Name	Accession	# of Reads	Mean Coverage	SOAPdenovo-Trans	Velvet	Trinity	RNA-SPAdes	MetaSPAdes	SPAdes
Heterosigma akashiwo RNA virus	AY337486	7,094	104.09	99.80%	99.90%	99.90%	99.90%	99.90%	99.90%
Human picobirnavirus	AB186898	16,988	1227.73	99.90%	77.40%	99.90%	99.90%	99.90%	99.90%
Human picobirnavirus	AB186897	7,049	351.71	99.90%	100.00%	100.00%	100.00%	100.00%	100.00%
Indian citrus ringspot virus	AF406744	11,880	198.03	100.00%	100.00%	100.00%	100.00%	100.00%	99.70%
Infectious Pancreatic necrosis virus	KY548520	14,790	679.11	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Infectious Pancreatic necrosis virus	KY548509	24,748	1006.08	100.00%	97.90%	100.00%	100.00%	100.00%	100.00%
Infectious flacherie virus	HM245295	4,722	61.50	99.60%	99.90%	99.70%	99.70%	99.70%	99.70%
Isavirus	HQ259678	45,332	6230.20	99.90%	35.00%	99.90%	89.20%	99.90%	99.90%
Isavirus	HQ259677	18,963	2110.86	99.90%	74.50%	99.90%	99.90%	99.90%	99.90%
Isavirus	HQ259676	6,831	651.48	99.90%	95.50%	99.90%	99.90%	99.90%	99.90%
Isavirus	HQ259675	22,554	1915.40	99.90%	99.90%	99.90%	92.50%	99.90%	99.90%
Isavirus	HQ259674	22,088	1524.82	99.90%	99.90%	99.90%	90.40%	99.90%	99.90%
Isavirus	HQ259673	19,012	1172.63	99.90%	100.00%	100.00%	100.00%	100.00%	99.90%
Isavirus	HQ259672	8,176	458.07	99.80%	100.00%	100.00%	100.00%	100.00%	99.90%
Isavirus	HQ259671	57,430	3196.01	100.00%	96.50%	100.00%	100.00%	100.00%	100.00%
Lassa mammarenavirus	MG8126 75	22,258	818.99	99.90%	100.00%	100.00%	100.00%	100.00%	99.90%
Lassa mammarenavirus	MG812674	5,551	95.44	99.80%	99.90%	99.90%	99.90%	99.90%	99.90%
Measles morbillivirus	LC420351	50,538	379.11	100.00%	99.80%	100.00%	100.00%	100.00%	100.00%
MERS	MK129253	4,095	17.11	98.00%	0.00%	99.40%	99.90%	99.90%	99.90%
Penicillium chrysogenum virus	AF296442	13,213	573.60	99.90%	100.00%	100.00%	100.00%	99.20%	99.20%
Penicillium chrysogenum virus	AF296441	2,855	120.83	99.80%	99.90%	99.90%	99.90%	99.90%	99.70%
Penicillium chrysogenum virus	AF296440	17,438	687.33	100.00%	94.20%	100.00%	100.00%	100.00%	100.00%
Penicillium chrysogenum virus	AF296439	7,407	261.98	99.90%	99.90%	100.00%	100.00%	100.00%	100.00%
Pseudomonas phage phi6	M12921	5,758	246.11	99.90%	99.90%	99.90%	99.90%	99.90%	99.90%
Pseudomonas phage phi6	M17462	24,192	750.45	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Pseudomonas phage phi6	M17461	11,660	230.50	99.90%	100.00%	100.00%	100.00%	99.80%	100.00%
Rabies lyssavirus	MH660455	9,658	92.53	0.00%	89.20%	100.00%	100.00%	99.80%	99.80%
Rosellinia necatrix megabirnavirus 1	AB512283	3,989	69.99	99.60%	99.80%	99.80%	99.80%	99.80%	99.80%
Rosellinia necatrix megabirnavirus 1	AB512282	7,822	110.36	99.90%	100.00%	100.00%	100.00%	100.00%	100.00%
Rosellinia necatrix quadrivirus 1	AB620064	5,170	176.78	99.70%	99.80%	99.90%	99.80%	99.80%	99.80%
Rosellinia necatrix quadrivirus 1	AB620063	20,708	636.72	100.00%	100.00%	100.00%	100.00%	100.00%	99.90%
Rosellinia necatrix quadrivirus 1	AB620062	5,172	149.74	99.80%	99.90%	99.90%	99.90%	99.90%	99.90%
Rosellinia necatrix quadrivirus 1	AB620061	4,377	111.57	99.70%	99.80%	99.90%	99.90%	99.90%	99.90%
Saccharomyces cerevisiae virus L-A	KU845301	2,617	71.91	99.70%	99.90%	99.90%	99.90%	99.90%	99.90%
Salivirus A	KT240115	3,543	55.64	99.50%	0.00%	99.70%	99.70%	99.70%	99.70%
Turnip yellow mosaic virus	KJ690173	19,674	392.33	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Zaire ebolavirus	KU174139	4,153	27.18	82.00%	0.00%	99.70%	99.70%	96.20%	96.70%

Table 5.3. Validation of the new pipeline against mock SAV infections ranging in viral titre from 0.05% to 5% of total reads

			Mapping		Assembly	
Total Dataset	**SAV Reads**	**Host Reads**	**# Reads Mapped**	**Average Coverage**	**Longest Viral Contig***	**Identity to template**
0.05%	5,000	9,995,000	4,994	53.271	11,838	100%
0.10%	10,000	9,990,000	9,983	106.508	11,857	100%
0.25%	25,000	9,975,000	24,945	264.653	11,857	100%
0.50%	50,000	9,950,000	49,883	529.274	11,857	100%
1.00%	100,000	9,900,000	99,804	1,059.18	11,857	100%
2.00%	200,000	9,800,000	199,581	2,119.01	11,857	100%
3.00%	300,000	9,700,000	299,364	3,177.26	11,857	100%
4.00%	400,000	9,600,000	399,147	4,239.31	11,857	100%
5.00%	500,000	9,500,000	498,969	5,299.93	11,857	100%

***** Length of the reference genome template was 11,858 bp

Table 5.4. Validation of the new pipeline against real ISAV infections.

Total Dataset	Titre Level	Mapping		Assembly	
		# Reads Mapped	# Segments Mapped	# Segments Assembled	Mean Contigs per Segment
SRR8506602	High	6,186	8	8	1.25
SRR8506607	High	8,801	8	8	1.25
SRR8506608	High	493,895	8	8	1.5
SRR8506609	High	13,594	8	8	1.5
SRR8506610	High	157,692	8	8	1.25
SRR8506611	High	91,623	8	8	1.25
SRR8506614	High	59,926	8	8	1.125
SRR8506617	High	58,883	8	8	1.25
SRR8506622	High	665	8	5	1.8
SRR8506630	High	6,333	8	8	2
SRR8506599	Low	0	0	0	-
SRR8506603	Low	6	0	0	-
SRR8506604	Low	44	0	0	-
SRR8506612	Low	33	0	0	-
SRR8506613	Low	21	0	0	-
SRR8506615	Low	8	0	0	-
SRR8506616	Low	13	0	0	-
SRR8506618	Low	101	0	0	-
SRR8506623	Low	0	0	0	-
SRR8506631	Low	32	0	0	-

Table 5.5. Summary of putative novel viral sequences identified in this study. Sequences were named according to the geographic location of origin, the estimated taxonomic classification based on maximum likelihood phylogenies (Figs. 5.2-5.11), and a virus number. Coverage and % reads were calculated from the library with the highest abundance in the case of datasets from BioProject PRJNA353875, and represent the average sequencing depth across the genome, and the % of total reads mapping to the viral genome respectively.

Classification	Virus Name	Length	Coverage	% reads	Closest BLASTx hit (amino acid identity)	Library
Picornaviridae	Qingdao picorna-like virus 1	9386	15.880	0.031%	Perth bee virus 6 (28.31%)	PRJNA353875
Picornaviridae	Qingdao picorna-like virus 10	6801	9.770	0.014%	Beihai picorna-like virus 70 (40.52%)	PRJNA353875
Picornaviridae	Qingdao picorna-like virus 11	1210	15.270	0.001%	Wenzhou picorna-like virus 28 (33.50%)	PRJNA353875
Picornaviridae	Qingdao picorna-like virus 12	3468	4.050	0.003%	Perth bee virus 7 (75.47%)	PRJNA353875
Picornaviridae	Qingdao picorna-like virus 13	9002	33.500	0.062%	Wenzhou gastropodes virus 1 (79.97%)	PRJNA353875
Picornaviridae	Qingdao picorna-like virus 14	8929	33.000	0.088%	Wenzhou picorna-like virus 48 (31.61%)	PRJNA353875
Picornaviridae	Qingdao picorna-like virus 15	9510	85.210	0.240%	Wenzhou picorna-like virus 48 (29.82%)	PRJNA353875
Picornaviridae	Qingdao picorna-like virus 2	9417	40.330	0.078%	Perth bee virus 7 (28.42%)	PRJNA353875
Picornaviridae	Qingdao picorna-like virus 3	9021	16.930	0.052%	Beihai picorna-like virus 11 (79.50%)	PRJNA353875
Picornaviridae	Qingdao picorna-like virus 4	8856	28.000	0.031%	Beihai picorna-like virus 11 (83.50%)	PRJNA353875
Picornaviridae	Qingdao picorna-like virus 5	9568	15.320	0.016%	Beihai sipunculid worm virus 5 (27.47%)	PRJNA353875
Picornaviridae	Qingdao picorna-like virus 6	9548	5.270	0.008%	Beihai sipunculid worm virus 5 (27.74%)	PRJNA353875
Picornaviridae	Qingdao picorna-like virus 7	7803	10.930	0.018%	Beihai picorna-like virus 15 (88.55%)	PRJNA353875
Picornaviridae	Qingdao picorna-like virus 8	8599	26.010	0.046%	Marine RNA virus BC-4 (59.24%)	PRJNA353875
Picornaviridae	Qingdao picorna-like virus 9	4752	6.240	0.006%	Wenzhou gastropodes virus 1 (35.80%)	PRJNA353875
Astroviridae	SCS astro-like virus 1	3169	12.929	0.001%	Wenling longspine snipefish astrovirus (42.79%)	SRR6291373
Bunyavirales	SCS bunya-like virus 1	5996	71.935	0.010%	Beihai bunya-like virus 2 (34.32%)	SRR6291373
Leviviridae	SCS levi-like virus 1	2665	9.232	0.001%	Wenling levi-like virus 2 (40.64%)	SRR6291373
Luteo-sombeo	SCS sobemo-like virus 1	3029	267.330	0.019%	Beihai sobemo-like virus 22 (38.87%)	SRR6291373
Luteo-sombeo	SCS sobemo-like virus 2	3171	25.675	0.002%	Wenling sobemo-like virus 1 (51.71%)	SRR6291373
Luteo-sombeo	SCS sobemo-like virus 3	2891	15.097	0.001%	Wuhan pillworm virus 3 (53.68%)	SRR6291373
Mono-chu	SCS chi-like virus 1	5395	9.781	0.001%	Wenling chuvirus-like virus 2 (50.71%)	SRR6291373
Mono-chu	SCS chi-like virus 2	6921	133.152	0.021%	Wenling chuvirus-like virus 2 (51.61%)	SRR6291373

Mono-chu	SCS chi-like virus 3	6617	16.576	0.003%	Wenling crustacean virus 13 (47.54%)	SRR6291373
Mono-chu	SCS chi-like virus 4	7139	28.975	0.005%	Wenzhou crab virus 2 (78.23%)	SRR6291373
Mono-chu	SCS chi-like virus 5	8060	31.280	0.006%	Beihai hermit crab virus 3 (46.03%)	SRR6291373
Mono-chu	SCS chi-like virus 5	4262	17.904	0.002%	Xinzhou nematode virus 5 (60%)	SRR6291373
Mono-chu	SCS rhabdo-like virus 1	9014	20.307	0.004%	Beihai rhabdo-like virus 2 (35.47%)	SRR6291373
Mono-chu	SCS rhabdo-like virus 2	7650	9.285	0.002%	Hubei rhabdo-like virus 5 (43.07%)	SRR6291373
Picornaviridae	SCS picorna-like virus 1	6144	15.823	0.002%	Beihai picorna-like virus 74 (31.34%)	SRR6291373
Picornaviridae	SCS picorna-like virus 2	9818	5318.594	1.236%	Picornavirales Q_sR_OV_042 (40.32%)	SRR6291373
Tombus-noda	SCS tombus-like virus 1	3098	91.955	0.007%	Beihai tombus-like virus 7 (40.87%)	SRR6291373
Toti-chryso	SCS toti-like virus 1	8624	30.494	0.006%	Beihai barnacle virus 15 (30.01%)	SRR6291373
Toti-chryso	SCS toti-like virus 2	3913	24.133	0.002%	dsRNA virus environmental sample (30.07%)	SRR6291373
Toti-chryso	SCS toti-like virus 3	4895	10.858	0.001%	Hubei toti-like virus 9 (35.13%)	SRR6291373

Supplementary File 1. Snakefile of the data curation, assembly and viral identification steps of Chapter 5.

```
SAMPLES = ["sample_name"]


rule all:
  input:
     "results/diamond/combined_results_mock_host.txt"

 rule trim:
  input:
      forward = "data/samples/mock_virome/mock_sav_infections/{sample}_1.fastq.gz",
      reverse = "data/samples/mock_virome/mock_sav_infections/{sample}_2.fastq.gz"
  output:
      forward = "data/samples/mock_virome/mock_sav_infections/{sample}_1_val_1.fq.gz",
      reverse = "data/samples/mock_virome/mock_sav_infections/{sample}_2_val_2.fq.gz"
  conda:
      "snakefiles/myenvs/py27.yml"
  shell:
      "trim_galore -q 30 --length 50 -o data/samples/mock_virome/mock_sav_infections/ --paired
{input.forward} {input.reverse}"

rule minimap:
  input:
      reference = "data/ref/ICSASG_v2_genomic.fna.gz",
      forward = "data/samples/mock_virome/mock_sav_infections/{sample}_1_val_1.fq.gz",
      reverse = "data/samples/mock_virome/mock_sav_infections/{sample}_2_val_2.fq.gz"
  output:
      "mapping/mapped_reads/mock_virome/mock_sav_infections/{sample}.sam"
  log:
      "logs/minimap_{sample}.log"
  shell:
      "(minimap2 -ax sr {input.reference} {input.forward} {input.reverse} "
      "> {output})"

rule ID_unmapped:
  input:
       "mapping/mapped_reads/mock_virome/mock_sav_infections/{sample}.sam"
  output:
       "mapping/unmapped_reads/mock_virome/mock_sav_infections/{sample}.unmapped.sam"
  log:
       "logs/unmapped_{sample}.log"
  shell:
       "samtools view -S -f 4 {input} > {output}"

rule unmapped_names:
  input:
       "mapping/unmapped_reads/mock_virome/mock_sav_infections/{sample}.unmapped.sam"
  output:
       "mapping/unmapped_reads/mock_virome/mock_sav_infections/{sample}.unmapped.txt"
  log:
       "logs/unmapped_names_{sample}.log"
  shell:
       "cut -f1 {input} | sort | uniq > {output}"

rule extract_unmappedF:
  input:
       idlist =
"mapping/unmapped_reads/mock_virome/mock_sav_infections/{sample}.unmapped.txt",
```

```
        forward = "data/samples/mock_virome/mock_sav_infections/{sample}_1_val_1.fq.gz"
    output:
        "mapping/unmapped_reads/mock_virome/mock_sav_infections/{sample}_1.unmapped.fastq"
    log:
        "logs/unmapped_names_{sample}.log"
    shell:
        "seqtk subseq {input.forward} {input.idlist} > {output}"

rule extract_unmappedR:
    input:
        idlist =
"mapping/unmapped_reads/mock_virome/mock_sav_infections/{sample}.unmapped.txt",
        reverse = "data/samples/mock_virome/mock_sav_infections/{sample}_2_val_2.fq.gz"
    output:
        "mapping/unmapped_reads/mock_virome/mock_sav_infections/{sample}_2.unmapped.fastq"
    shell:
        "seqtk subseq {input.reverse} {input.idlist} > {output}"

rule denovo:
    input:
        forward =
"mapping/unmapped_reads/mock_virome/mock_sav_infections/{sample}_1.unmapped.fastq",
        reverse =
"mapping/unmapped_reads/mock_virome/mock_sav_infections/{sample}_2.unmapped.fastq"
    output:
        directory("trinity_out_dir_{sample}")
    shell:
        "Trinity --seqType fq --max_memory 64G --left {input.forward} --right {input.reverse} --output
{output}"

rule organise:
    input:
        "trinity_out_dir_{sample}"
    output:
        "assembly/trinity_contigs_{sample}/Trinity.fasta"
    shell:
        "mv {input}/Trinity.fasta {output}"

rule blastx_diamond:
    input:
        "assembly/trinity_contigs_{sample}/Trinity.fasta"
    output:
        "results/diamond/diamond_blastx_hits_{sample}.txt"
    shell:
        "diamond blastx -d data/databases/ref_prot_database --threads 1 -q {input} -o {output}"

rule combine_results:
    input:
        expand("results/diamond/diamond_blastx_hits_{sample}.txt", sample=SAMPLES)
    output:
        "results/diamond/combined_results_mock_host.txt"
    shell:
        "cat {input} > {output}"
```

Chapter 6. General Discussion

6.1 Thesis main findings

This Thesis presents three different approaches to characterise viruses in aquaculture using NGS methods to generate whole viral genomes. These approaches range in utility based on the virus of interest, the number and type of samples available, the financial scope of the project and the turnaround time required.

Chapter 2 outlines a rapid, long-read sequencing approach of known viruses using tiled PCR amplicons to ensure full genome sequencing coverage. Using a <24hr lab-based workflow, users can take infected tissue or cultured samples, amplify the virus genome in overlapping 2kb or 4kb PCR amplicons, and sequence the products on Oxford Nanopore's MinION platform (Fig. 2.4). This approach is compatible with studies requiring real-time analysis of disease outbreaks and is highly customisable depending on the virus of interest and the sample type including low-titre samples. However it is heavily dependent on an effective PCR amplification step along with primers that reduce the amplification bias which limits it's utility to well characterised viral species and requires a continuous screening of variants in PCR primer regions to ensure that the primers are not introducing unacceptable levels of bias in the amplification of certain strains over others. Chapter 3 then uses this approach to investigate the current molecular epidemiological landscape of the SAV3 epidemic in Norway using whole genome sequencing and phylodynamic analyses. This approach is also potentially useful for higher resolution studies of virus transmission and evolution as has been shown extensively in studies on human viruses including several on the ongoing SARS-COV2 pandemic (e.g. COG-UK, 2020), though this remains to be validated in the case of aquaculture viruses. In this Chapter, the structural variant landscape of Norwegian SAV3 samples was also characterised and while the results corroborated previous work (Petterson et al. 2013) and demonstrated the usefulness of Nanopore sequencing of amplicons to detect structural variants, the biological implications of such results would need to be carefully analysed. By using infected tissue samples instead of virus culture supernatant, the true viral population can be captured, but the viral RNA being sequenced is a mixture of both genomic RNA and mRNA. This mixture of RNA molecules results in both deleted viral genomes and mRNA splice variants being sequenced, and therefore distinguishing between the two becomes challenging. More research and finer scale resolution into alternative splicing break-points in viruses would help distinguish splice variants presenting as deleted mRNAs from true deleted viral genomes (see Section 6.2.1)

Chapter 4 presents an approach for ultra-deep characterisation of viral populations in a sample using targeted sequence capture and short-read Illumina sequencing. This Chapter employs

the use of Agilent SureSelect^{XT2} 120-mer RNA oligomer baits designed from a selection of viral genome templates to enrich for the viral cDNA and remove the host nucleic acids. This approach was used to characterise the viral diversity present in natural SAV infections of both farmed and wild fish from Scottish and Irish waters, leading to the identification of both subtype-level co-infections in individual fish, and the circulation of multiple SAV subtypes in individual farm sites. Accurate, high-resolution data on the population genetics of viruses on farms in Ireland and Scotland is critical to large scale epidemiological studies on the transmission routes of SAV. Additionally, knowing what strains and subtypes of SAV are present in farms over a longitudinal time series provides a direct assessment of the efficacy of biosecurity controls and fallowing techniques following local epidemics, something Sanger sequencing a viral culture is unable achieve without considerably more effort. While less flexible or customisable than the sequencing approach used in Chapters 2 and 3, the improved data quality of this approach (i.e. Illumina sequencing – see Section 1.5) allows for high resolution of viral epidemiology studies and comprehensive characterisation of complex viral populations. However the lab work involved in this approach does not facilitate a rapid response time, as well as the likelihood of having to employ commercial services for the Illumina sequencing. Additionally, the efficiency of the sequence capture depends on the initial viral load in the samples, with low titre samples resulting in less efficient sequencing results.

Chapter 5 outlines a bioinformatics workflow including a Snakemake pipeline to characterise viruses, both known and unknown, from metagenomics and meta-transcriptomic datasets (i.e. RNA-seq). This workflow was validated against several mock and real datasets before being used to robustly characterise the virome of Pacific oysters before and during an experimental challenge with norovirus, resulting in the identification of 35 putative new viral sequences belonging to 31 putative viral species. Unlike Chapters 2, 3 & 4, this approach is unbiased towards the viral species of interest and can easily detect and characterise known viruses as well as unclassified viruses. As has been sharply demonstrated by this current COVID-19 pandemic, there is an ever-present risk of emerging diseases from viruses not currently in circulation among human populations. While a long-understudied area of infectious diseases, both creating a comprehensive and annotated database of viruses found in our food sources and developing tools which can accurately analyse viruses highly divergent from any other in said database is one of the most effective forms of proactive research regarding emerging diseases. This approach can also be used to investigate host-pathogen dynamics as the transcriptional response of the host to infection is captured along with the pathogen causing any changes, something that is often poorly understood in non-human and non-profitable agricultural diseases. However this is a highly inefficient method of sequencing specific

viruses if already known to be in a sample as the vast majority of reads produced are likely to be from the host species. The analysis of such datasets is also relatively laborious and real-time results are difficult if not impossible to achieve with most sequencing platforms (though with the continuous improvement in Nanopore platforms, this might change in the near future).

This final Chapter concludes the Thesis by briefly discussing the future of viral genomics in aquaculture, pointing to future areas of interest including more comprehensive genomic surveillance strategies, advances in NGS technologies, and the use of such data in multidisciplinary infectious disease control.

6.2 Future research in aquaculture viral genomics

As mentioned already, aquaculture is a hugely important industry, both economically and from a food security point of view. However a major factor in the expansion of this industry is the control of infectious diseases of which viruses are of particular concern due to the lack of effective therapeutics (McLoughlin and Graham, 2007; Dhar et al., 2014). In the context of pathogen surveillance, viral diagnostics have been developed and optimised extensively to be able to reliably detect the presence of most, if not all the commonly impactful viral species (e.g. Fringuelli et al., 2008; Hodneland and Endresen, 2006). However despite the fact that genomic surveillance has been used to great effect in scenarios of public health epidemics (e.g. Van Ballegooijen et al., 2009; Arias et al., 2016; Quick et al., 2016; Thézé et al., 2018), many of the diagnostic tests used in aquaculture rely on simply confirming the presence of the virus and occasionally sequencing a portion of a gene to reveal the subtype (see Section 1.4). Developing methods to generate reliable genomic surveillance should be an important part of future effective control programmes for viral diseases in aquaculture, particularly with the improvements in cost and quality of NGS (see Section 1.5). Having a range of options for infectious disease surveillance allows researchers and regulators to choose the most appropriate tool for specific outbreaks. Importantly these options need to cover a range of affordability (e.g. cost per sample or cost per project), turnaround time (e.g. as close to real-time as possible for outbreak scenarios), tolerance to strain-level genetic variability, and lab- and computational-based labour intensity. For example, even with the ever-decreasing cost of NGS platforms, many of the short-read technologies (i.e. Illumina) only become cost effective when sequencing highly multiplexed libraries, due to most flow cells run-time not being customisable in length of time and reusability, while Nanopore flow cells solve both of these issues by allowing the user to determine the run-time and number of runs per flow cell. The initial price of the sequencing equipment and flow cells are often the inhibitive cost of small sequencing projects, including either studies that only have a few samples for sequencing, or studies that require real-time sequencing and analysis and therefore cannot wait until enough samples have been accumulated to make it cost effective. However, when sequencing large

numbers of samples over a longer time period, the cost per project of real-time sequencing considerably increases and approaches such as using the Illumina NovaSeq for highly multiplexed libraries become more attractive. Additionally, the sensitivity of studies to per-base accuracy often drives method selection. While sequencing amplicons, or genomic DNA, to get a single consensus sequence for a sample is rapidly, accurately and affordably achieved with current long-read technologies, finer resolution studies into intra-host variation, haplotyping or microbial population diversity requires individual reads to be of high quality and the originating DNA strands to be easily distinguishable. For example, in the ongoing COVID-19 epidemic, rapid genomic sequencing of SARS-COV2 has been occurring around the world. Between the need to near real-time sequencing of infected patients, and the knowledge of a lack of genetic diversity in the viral populations to-date, many national sequencing efforts have been heavily relying on Nanopore sequencing of PCR amplicons as it gives fast results, has a highly customisable library prep, and is well-suited to producing genome-wide consensus sequences (COG-UK 2020)

Future research needs to take these factors into account when validating further approaches, with a goal of bodies such as the OIE and regional reference labs giving official accreditation to a range of diagnostic and characterisation tests beyond PCR/qPCR and Sanger sequencing of partial gene sequences. With more genome-wide characterisation of viral strains affecting aquaculture available, opportunities to understand both the biology of viruses and the dynamics of the diseases they cause will become increasingly available.

6.2.1 Advances in NGS will improve pathogen analyses

As already mentioned in Section 1.5, NGS promises significant improvements in viral genome assembly and characterisation over traditional Sanger sequencing. In particular third generation (i.e. long-read) sequencing may solve many issues with viral genome sequencing and intra-host strain phasing that short-read technologies struggle with, such as phasing closely-related strains, as the relevant variants are often further apart than the longest Illumina platform's read. To my knowledge Chapter 4 of this Thesis was the first time that NGS had been used to characterise SAV populations, and Chapter 2 the first use of third-generation sequencing in any aquatic virus. As these long-read technologies mature, data quality and quantity will improve which will open opportunities to study viral populations and genetic variation across whole viral genomes. However as it stands right now, the raw error rate of PacBio and Nanopore sequencing remains too high for reliable intra-host variant calling without bespoke library preparation techniques such as unique molecular identifiers (UMI's) (Karst et al. BioRxiv). In addition to long-read technologies, the development of affordable linked read sequencing methods (e.g. Chen et al., 2019; Redin et al., 2019) opens the possibility of performing high accuracy, short-read Illumina sequencing that contains long-

range information, equivelent to the more expensive 10x Genomics linked-read sequencing approach (https://www.10xgenomics.com/linked-reads/). While linked-read methods have been optimised to generate ultra-long phased blocks of larger genomes (i.e. eukaryotes), the same principles that enable the phasing of haplotypes from diploid or polyploidy organisms, should also enable the phasing of viral strains from a sample with a complex viral population. The affordability of such library preparation approaches is attractive (for example, $720 for 12 samples using Tell-Seq; https://www.universalsequencing.com/shop) as it becomes competitive with the cheapest metagenomics RNA-seq projects and even rivals the price of sequencing PCR amplicons on Oxford Nanopore's MinION platform (see Section 2.3.5), though the cost of the Illumina sequencing would need to be factored into any future project.

Finally, the development of direct RNA sequencing on Oxford Nanopore's platforms permits the investigation of viral transcriptomics without introducing biases involved in cDNA synthesis (Keller et al., 2018; Depledge et al., 2019). This is an exciting development in the field of viral genomics and transcriptomics as directly detecting base modifications, splice variants and transcriptional changes is crucial to understanding the transcriptional landscape of viral pathogens. This in turn is important in understanding how viruses overcome host cell defences and ultimately may help shape vaccine production efforts. Specifically, the knowledge of mRNA structure and structural variants found in viral populations would greatly be enhanced with direct RNA sequencing of samples. In Chapter 3 of this Thesis I characterised deletions in natural SAV3 infections in Norway. However, while a portion of these deletions may be viral particles with true deleted genomes, it is likely that at least some of these deletions represent isoforms or splice variants of viral mRNA, and others possible represent artifacts of cDNA synthesis and PCR. Directly sequencing the viral RNA would remove the uncertainty surrounding the latter sources of error, while sequencing from viral supernatant would address the former (as discussed in Section 6.1). However, as with all uses of novel technologies, the effects of sample selection, sample storage conditions, and library/sample handling on the accuracy of such studies is yet unknown.

6.2.2 Improvements in viral genomic surveillance in aquaculture

By increasing the amount of NGS data produced by genomic surveillance screening (of both symptomatic and asymptomatic hosts), the density of genomic databases will increase, which are vital resources in molecular epidemiological studies. Future research in this area needs to build on the work presented in this Thesis and focus on the standardisation of sequencing approaches that can reliably generate whole genome sequences from a wide range of sample types. Specifically, approaches should be tolerant of sample degradation (i.e. in the case of historically archived samples), viral titres and genetic variability, while also being affordable enough to perform on large numbers of samples with as wide a range of sampling time and

geographic origins as possible. Due to the wide variation in genome structures and transcriptomic profiles of many viruses in aquaculture (see Tables 1.2 and 1.2), sequencing methods may need to be species-specific as it is unlikely that a one-size-fits-all approach would be appropriate.

A point to consider before wide-spread genomic surveillance can begin, is that while PCR-based tested are likely to be the gold-standard approaches due to the low cost and high efficiency of sequencing PCR products, they are highly sensitive to primer mismatches. As RNA viruses have such a fast mutation rate, and viral populations tend to be genetically heterogeneous, a comprehensive knowledge of common variants across the genome in each species needs to be accumulated so as to avoid accidental biases in genomic surveillance programmes. This database of conserved variants would also be of great use for transmission studies as it would allow for the identification of specific lineages rapidly, and without the necessity of complex and time-consuming phylogenetic analyses. This is have implemented with great success in the current COVID-19 pandemic with the development of the Pangolin software (Rambaut et al., 2020) (https://pangolin.cog-uk.io/). This software identifies lineages in a large genomic dataset based on variants that are shared by multiple sequences within a monophyletic phylogeny. Thus allowing for a hierarchical nomenclature that is adaptable to future evolution of the virus (e.g. lineages B1 and B2 are related but distinct, but both having emerged from the parent lineage of B). While this type of database requires large amounts of highly accurate genome sequences, the potential utility for transmission studies from both academia and industry is enormous and would greatly assist traditional epidemiological methods in tracing common sources of infections in new farms.

With the development and standardisation of such approaches, widespread genome sequencing of virus-infected samples can be performed with the goal of generating dense genomic databases, ideally along with suitable metadata (e.g. isolation date, location, host species, etc). Improving the density of genomic databases will enable much higher resolution studies of viral transmission routes, the identification of which is important in effective disease control efforts, as discussed in Chapter 3 of this Thesis. Additionally genome-wide association studies (GWAS), which requires large numbers of high quality genomes and their associated metadata, have enormous potential in the identification of virulence markers and drug resistance (Power et al., 2016b, 2016a; Genissel et al., 2017), though several hurdles do remain in the field of microbial GWAS (Power et al., 2016b). In particular, reducing false positive causal variants from non-heterogeneous population structures remains a major issue (Ioannidis et al., 2009) with studies on diverse population resulting in the identification of variants linked to ancestry, rather than the biology of disease (e.g. antimicrobial resistance). However the

potential to identify causal variants and test their phenotypic effect in the lab may reduce the concerns of false positives typically associated with human GWAS.

Another potential avenue of research for some hypervariable viral pathogens is the creation of a comprehensive database of variants found in the species. Mimicking the concept of pan-genomes or graph genomics in bacteria and eukaryotes (Brandt et al., 2015; Dilthey et al., 2015; Limasset et al., 2016; Novak et al., 2017), the generation of a 'master-genome' for viral species, or at least a dataset of variants known to be associated with certain phenotypes of interest, is a tempting option in the pursuit of characterising the genetic diversity found in natural viral populations. Though this approach would require extensive lab validation, these data could then be instrumental in understanding the effects of variants (both SNVs and indels) on the RNA coding sequence and ultimately the phenotypic effect on the viral packaging and replication processes (Nikolaitchik et al., 2006; Tong and Revill, 2016).

Finally, with the likely increase in rate of sequencing for many aquaculture viruses due to the uptake of NGS approaches, nomenclature of viral strains and isolates should be standardised as it is for many public health viruses (e.g. influenza).

6.3 Viral diseases in aquaculture; a multidisciplinary control effort

To conclude this Thesis, these are very exciting times for the fields of infectious disease dynamics and genomic epidemiology. The emergence of a wider appreciation for the effect of the whole microbiome (including the virome) on the health and wellbeing of animals is of particular interest as aquatic microbes and viruses have traditionally been understudied compared to their terrestrial counterparts. Particularly with the increase in prevalence of anti-microbial resistance in both agriculture and aquaculture settings, the ability to link disease phenotypes to individual pathogen infections will become more complicated as complex diseases often are associated with multiple pathogen infections, in which viruses usually contribute significantly. However, while viral genomics and molecular epidemiology can play a vital supporting role, these approaches cannot control infectious diseases in aquaculture without a range of other interventions and strategies. Multidisciplinary efforts must be employed if viral diseases are to be controlled, or even eradicated. Traditional diagnostic methods should be complemented by routine whole virus genome sequencing, which in turn is reliant on accurate, detailed and large-scale epidemiological data to make inferences on viral transmissions and population dynamics. While not a focus of this Thesis, genetic selection of fish stocks and the development of specific 'lines' of species that may exhibit desired traits (e.g. faster growth, better food conversion ratios, or resistance to certain diseases) will be central to any comprehensive disease control programme. And while there has been huge progress in this field for some well-studied species, there are many more species of fish and

shellfish that have not been intensely selectively bred and therefore have considerable genetic potential still present in different populations. As shown in Figure 1.1, many factors are required to result in animal disease, but these cannot be treated as independent areas of research. Although not all these efforts are commonly employed yet in aquaculture, the advances being made in the genomic epidemiology of human viruses open many avenues of research which will undoubtedly help address many of the remaining mysteries of viral dynamics in aquaculture.

References

Acosta-Leal, R., Duffy, S., Xiong, Z., Hammond, R.W., Elena, S.F., 2011. Advances in plant virus evolution: translating evolutionary insights into better disease management. Phytopathology 101, 1136–1148. https://doi.org/10.1094/PHYTO-01-11-0017

Aggarwala, V., Liang, G., Bushman, F.D., 2017. Viral communities of the human gut: Metagenomic analysis of composition and dynamics. Mob. DNA 8, 1–10. https://doi.org/10.1186/s13100-017-0095-y

Agoti, C.N.N., Mbisa, J.L.L., Bett, A., Medley, G.F.F., Nokes, D.J.J., Cane, P.A.A., 2010. Intrapatient Variation of the Respiratory Syncytial Virus Attachment Protein Gene. J. Virol. 84, 10425–10428. https://doi.org/10.1128/jvi.01181-10

Ahne, W., Bjorklund, H., Essbauer, S., Fijan, N., Kurath, G., Winton, J., 2002. Spring viremia of carp (SVC). Dis. Aquat. Organ. 52, 261–272. https://doi.org/10.3354/dao052261

Alejandro Rodríguez-Valencia, J., Crespo, D., López-Camacho, M., 2010. LA CAMARONICULTURA Y LA SUSTENTABILIDAD DEL GOLFO DE CALIFORNIA.

Amarasinghe, S.L., Su, S., Dong, X., Zappia, L., Ritchie, M.E., Gouil, Q., 2020. Opportunities and challenges in long-read sequencing data analysis. Genome Biol. https://doi.org/10.1186/s13059-020-1935-5

Amin, A.B., Trasti, J., 1988. Endomyocarditis in Atlantic salmon in Norwegian seafarms; A case report.

Arias, A., Watson, S.J., Asogun, D., Tobin, E.A., Lu, J., Phan, M.V.T., Jah, U., Wadoum, R.E.G., Meredith, L., Thorne, L., Caddy, S., Tarawalie, A., Langat, P., Dudas, G., Faria, N.R., Dellicour, S., Kamara, A., Kargbo, B., Kamara, B.O., Gevao, S., Cooper, D., Newport, M., Horby, P., Dunning, J., Sahr, F., Brooks, T., Simpson, A.J.H., Groppelli, E., Liu, G., Mulakken, N., Rhodes, K., Akpablie, J., Yoti, Z., Lamunu, M., Vitto, E., Otim, P., Owilli, C., Boateng, I., Okoror, L., Omomoh, E., Oyakhilome, J., Omiunu, R., Yemisis, I., Adomeh, D., Ehikhiametalor, S., Akhilomen, P., Aire, C., Kurth, A., Cook, N., Baumann, J., Gabriel, M., Wölfel, R., Di Caro, A., Carroll, M.W., Günther, S., Redd, J., Naidoo, D., Pybus, O.G., Rambaut, A., Kellam, P., Goodfellow, I., Cotten, M., 2016. Rapid outbreak sequencing of Ebola virus in Sierra Leone identifies transmission chains linked to sporadic cases. Virus Evol. 2, vew016. https://doi.org/10.1093/ve/vew016

Arseneau, J.R., Gautreau, C., Boston, L., Goguen, M.L., Laflamme, M., 2019. Accelerated ISAV replication detection by cell culture methods combined with time-monitoring RT-qPCR. J. Fish Dis. 42, 257–267. https://doi.org/10.1111/jfd.12925

Asche, F., Hansen, H., Tveteras, R., Tveterås, S., 2009. The salmon disease crisis in Chile. Mar. Resour. Econ. 24, 405–411. https://doi.org/10.1086/mre.24.4.42629664

Aunsmo, A., Valle, P.S., Sandberg, M., Midtlyng, P.J., Bruheim, T., 2010. Stochastic modelling of direct costs of pancreas disease (PD) in Norwegian farmed Atlantic salmon (Salmo salar L.). Prev. Vet. Med. 93, 233–241. https://doi.org/10.1016/j.prevetmed.2009.10.001

Bacharach, E., Mishra, N., Briese, T., Zody, M.C.C., Tsofack, J.E.K.E.K., Zamostiano, R., Berkowitz, A., Ng, J., Nitido, A., Corvelo, A., Toussaint, N.C.C., Abel Nielsen, S.C.C., Hornig, M., del Pozo, J., Bloom, T., Ferguson, H., Eldar, A., Lipkin, W.I.I., 2016. Characterization of a novel orthomyxo-like virus causing mass die-offs of Tilapia. MBio 7, 1–7. https://doi.org/10.1128/mBio.00431-16

Baird, H.A., Galetto, R., Gao, Y., Simon-Loriere, E., Abreha, M., Archer, J., Fan, J., Robertson, D.L., Arts, E.J., Negroni, M., 2006. Sequence determinants of breakpoint location during HIV-1 intersubtype recombination. Nucleic Acids Res. 34, 5203–5216. https://doi.org/10.1093/nar/gkl669

Batts, W., Yun, S., Hedrick, R., Winton, J., 2011. A novel member of the family Hepeviridae from cutthroat trout (Oncorhynchus clarkii). Virus Res. 158, 116–123. https://doi.org/10.1016/j.virusres.2011.03.019

Baumer, A., Fabian, M., Wilkens, M., Steinhagen, D., Runge, M., 2013. Epidemiology of cyprinid herpesvirus-3 infection in latently infected carp from aquaculture. Dis. Aquat. Organ. 105, 101–108. https://doi.org/10.3354/dao02604

Bayliss, S.C., Verner-Jeffreys, D.W., Bartie, K.L., Aanensen, D.M., Sheppard, S.K., Adams, A., Feil, E.J., 2017. The promise of whole genome pathogen sequencing for the molecular epidemiology of emerging aquaculture pathogens. Front. Microbiol. 8, 1–18. https://doi.org/10.3389/fmicb.2017.00121

Belshaw, R., de Oliveira, T., Markowitz, S., Rambaut, A., 2009. The RNA Virus Database. Nucleic Acids Res. https://doi.org/10.1093/nar/gkn729

Bergmann, S.M., Kempter, J., 2011. Detection of koi herpesvirus (KHV) after re-activation in persistently infected common carp (Cyprinus carpio L.) using non-lethal sampling methods. Bull. Eur. Assoc. Fish Pathol.

Bloom, J.D., Gong, L.I., Baltimore, D., 2010. Permissive secondary mutations enable the evolution of influenza oseltamivir resistance. Science (80-.). https://doi.org/10.1126/science.1187816

Bodewes, R., van der Giessen, J., Haagmans, B.L.L., Osterhaus, A.D.M.E.D.M.E., Smits, S.L.L., 2013. Identification of Multiple Novel Viruses, Including a Parvovirus and a Hepevirus, in Feces of Red Foxes. J. Virol. 87, 7758–7764. https://doi.org/10.1128/jvi.00568-13

Bongartz, P., 2019. Resolving repeat families with long reads. BMC Bioinformatics. https://doi.org/10.1186/s12859-019-2807-4

Bose, J., Kloesener, M.H., Schulte, R.D., 2016. Multiple-genotype infections and their complex effect on virulence. Zoology 119, 339–349. https://doi.org/10.1016/j.zool.2016.06.003

Boucher, P., Laurencin, F., 1994. Sleeping Disease (SD) of salmonids. Bull. Eur. Ass. Fish Pathol 14, 179–180.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.H., Xie, D., Suchard, M.A., Rambaut, A., Drummond, A.J., 2014. BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. PLoS Comput. Biol. 10, 1–6. https://doi.org/10.1371/journal.pcbi.1003537

Brandt, D.Y.C., Aguiar, V.R.C., Bitarello, B.D., Nunes, K., Goudet, J., Meyer, D., 2015. Mapping bias overestimates reference allele frequencies at the HLA genes in the 1000 genomes project phase I data. G3 Genes, Genomes, Genet. https://doi.org/10.1534/g3.114.015784

Breitwieser, F.P., Lu, J., Salzberg, S.L., 2018. A review of methods and databases for metagenomic classification and assembly. Brief. Bioinform. https://doi.org/10.1093/bib/bbx120

Briese, T., Kapoor, A., Mishra, N., Jain, K., Kumar, A., Jabado, O.J., Ian Lipkina, W., 2015. Virome capture sequencing enables sensitive viral diagnosis and comprehensive virome analysis. MBio. https://doi.org/10.1128/mBio.01491-15

Bright, A.T., Tewhey, R., Abeles, S., Chuquiyauri, R., Llanos-Cuentas, A., Ferreira, M.U., Schork, N.J., Vinetz, J.M., Winzeler, E.A., 2012. Whole genome sequencing analysis of Plasmodium vivax using whole genome capture. BMC Genomics. https://doi.org/10.1186/1471-2164-13-262

Brown, B.L., Watson, M., Minot, S.S., Rivera, M.C., Franklin, R.B., 2017. MinIONTM nanopore sequencing of environmental metagenomes: A synthetic approach. Gigascience. https://doi.org/10.1093/gigascience/gix007

Brun, E., Poppe, T., Skrudland, A., Jarp, J., 2003. Cardiomyopathy syndrome in farmed Atlantic salmon Salmo salar: occurrence and direct financial losses for Norwegian aquaculture. Dis. Aquat. Organ. 56, 241–247. https://doi.org/10.3354/dao056241

Bruno, D.W., Noguera, P.A., Black, J., Murray, W., Macqueen, D.J., Matejusova, I., 2014. Identification of a wild reservoir of salmonid alphavirus in common dab Limanda limanda, with emphasis on virus culture and sequencing. Aquac. Environ. Interact. 5, 89–98. https://doi.org/10.3354/aei00097

Buchfink, B., Xie, C., Huson, D.H., 2014. Fast and sensitive protein alignment using DIAMOND.

Nat. Methods. https://doi.org/10.1038/nmeth.3176

Buck, D., Weirather, J.L., de Cesare, M., Wang, Y., Piazza, P., Sebastiano, V., Wang, X.-J.J., Au, K.F., Buck, D., Au, K.F., 2017. Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. F1000Research 6, 100. https://doi.org/10.12688/f1000research.10571.1

Burford, M.A., Williams, K.C., 2001. The fate of nitrogenous waste from shrimp feeding. Aquaculture 198, 79–93. https://doi.org/10.1016/S0044-8486(00)00589-5

Bushmanova, E., Antipov, D., Lapidus, A., Prjibelski, A.D., 2019. RnaSPAdes: A de novo transcriptome assembler and its application to RNA-Seq data. Gigascience. https://doi.org/10.1093/gigascience/giz100

Bushnell, B., 2016. BBMap short read aligner [WWW Document]. http://sourceforge. net/projects/bbmap.

Cabelli, V.J., Dufour, A.P., Levin, M., McCabe, L.J., Haberman, P.W., 1979. Relationship of microbial indicators to health effects at marine bathing beaches. Am. J. Public Health 69, 690–696. https://doi.org/10.2105/AJPH.69.7.690

Calvignac-Spencer, S., Schulze, J.M., Zickmann, F., Renard, B.Y., 2014. Clock Rooting Further Demonstrates that Guinea 2014 EBOV is a Member of the Zaïre Lineage. PLoS Curr. 5–9. https://doi.org/10.1371/currents.outbreaks.c0e035c86d721668a6ad7353f7f6fe86

Capella-Gutiérrez, S., Silla-Martínez, J.M., Gabaldón, T., 2009. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. https://doi.org/10.1093/bioinformatics/btp348

Cárdenas, C., Carmona, M., Gallardo, A., Labra, A., Marshall, S.H., 2014. Coexistence in field samples of two variants of the infectious salmon anemia Virus: A putative shift to pathogenicity. PLoS One 9, 1–9. https://doi.org/10.1371/journal.pone.0087832

Carr, J.K., Salminen, M.O., Albert, J., Sanders-Buell, E., Gotte, D., Birx, D.L., McCutchan, F.E., 1998. Full genome sequences of human immunodeficiency virus type 1 subtypes G and A/G intersubtype recombinants. Virology 247, 22–31. https://doi.org/10.1006/viro.1998.9211

Carrasco, L., Sanz, M.A., González-Almela, E., 2018. The Regulation of Translation in Alphavirus-Infected Cells. Viruses 10, 70. https://doi.org/10.3390/v10020070

Casillas-Hernández, R., Nolasco-Soria, H., García-Galano, T., Carrillo-Farnes, O., Páez-Osuna, F., 2007. Water quality, chemical fluxes and production in semi-intensive Pacific white shrimp (Litopenaeus vannamei) culture ponds utilizing two different feeding strategies. Aquac. Eng. 36, 105–114. https://doi.org/10.1016/j.aquaeng.2006.09.001

Castresana, J., 2007. Topological variation in single-gene phylogenetic trees. Genome Biol. 8. https://doi.org/10.1186/gb-2007-8-6-216

Castric, J., Baudin Laurencin, F., Brémont, M., Jeffroy, J., Le Ven, A., Bearzotti, M., 1997. Isolation of the virus responsible for sleeping disease in experimentally infected rainbow trout (Oncorhynchus mykiss). Bull. Eur. Assoc. Fish Pathol.

Castro-Nallar, E., Cortez-San Martín, M., Mascayano, C., Molina, C., Crandall, K.A., 2011. Molecular phylodynamics and protein modeling of infectious salmon anemia virus (ISAV). BMC Evol. Biol. 11, 349. https://doi.org/10.1186/1471-2148-11-349

Chandler, L.J., Blair, C.D., Beaty, B.J., 1993. Detection of dengue-2 viral RNA by reversible target capture hybridization. J. Clin. Microbiol. https://doi.org/10.1128/jcm.31.10.2641-2647.1993

Chao, D.L., Bloom, J.D., Kochin, B.F., Antia, R., Longini, I.M., 2012. The global spread of drug-resistant influenza. J. R. Soc. Interface. https://doi.org/10.1098/rsif.2011.0427

Chen, G.X., Zhu, J., Plitt, J.R., Weiler, A.K., Werner Zolg, J., 1991. A Plasmodium falciparum-specific reverse target capture assay. Mol. Biochem. Parasitol. https://doi.org/10.1016/0166-6851(91)90002-N

Chen, R., Holmes, E.C., 2009. Frequent inter-species transmission and geographic subdivision in avian influenza viruses from wild birds. Virology 383, 156–161. https://doi.org/10.1016/j.virol.2008.10.015

Chen, Z., Pham, L., Wu, T.-C., Mo, G., Xia, Y., Chang, P., Porter, D., Phan, T., Che, H., Tran, H., Bansal, V., Shaffer, J., Belda-Ferre, P., Humphrey, G., Knight, R., Pevzner, P., Pham, S., Wang, Y., Lei, M., 2019. Ultra-low input single tube linked-read library method enables short-read NGS systems to generate highly accurate and economical long-range sequencing information for de novo genome assembly and haplotype phasing. bioRxiv. https://doi.org/10.1101/852947

Choi, M., Scholl, U.I., Ji, W., Liu, T., Tikhonova, I.R., Zumbo, P., Nayir, A., Bakkaloğlu, A., Özen, S., Sanjad, S., Nelson-Williams, C., Farhi, A., Mane, S., Lifton, R.P., 2009. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc. Natl. Acad. Sci. U. S. A. https://doi.org/10.1073/pnas.0910672106

Chong, Y.., Loh, P., 1984. Hepatopancreas chlamydial and parvovirus infections of farmed marine prawns in Singapore. Singapore Vet. J.

Christiansen, D.H., McBeath, A.J.A., Aamelfot, M., Matejusova, I., Fourrier, M., White, P., Petersen, P.E., Falk, K., 2017. First field evidence of the evolution from a non-virulent HPR0 to a virulent HPR-deleted infectious salmon anaemia virus. J. Gen. Virol. 98, 595–606.

Chua, K.B., 2000. Nipah virus: A recently emergent deadly paramyxovirus. Science (80-.). 288, 1432–1435. https://doi.org/10.1126/science.288.5470.1432

Clouthier, S.C., Rector, T., Brown, N.E.C., Anderson, E.D., 2002. Genomic organization of infectious salmon anaemia virus. J. Gen. Virol. 83, 421–428. https://doi.org/10.1099/0022-1317-83-2-421

Coenye, T., Vandamme, P., 2003. Intragenomic heterogeneity between multiple 16S ribosomal RNA operons in sequenced bacterial genomes. FEMS Microbiol. Lett. 228, 45–49. https://doi.org/10.1016/S0378-1097(03)00717-1

Cottet, L., Cortez-San Martin, M., Tello, M., Olivares, E., Rivas-Aravena, A., Vallejos, E., Sandino, A.M., Spencer, E., 2010. Bioinformatic Analysis of the Genome of Infectious Salmon Anemia Virus Associated with Outbreaks with High Mortality in Chile. J. Virol. 84, 11916–11928. https://doi.org/10.1128/jvi.01202-10

Cottet, L., Rivas-Aravena, A., Cortez-San Martin, M., Sandino, A.M., Spencer, E., 2011. Infectious salmon anemia virus-Genetics and pathogenesis. Virus Res. https://doi.org/10.1016/j.virusres.2010.10.021

The COVID-19 Genomics UK (COG-UK) consortium, 2020. An integrated national scale SARS-CoV-2 genomic surveillance network. The Lancet Microbe. 1 (3), 99-100. https://doi.org/10.1016/S2666-5247(20)30054-9

Cowley, J.A., McCulloch, R.J., Spann, K.M., Cadogan, L.C., Walker, P.J., 2005. Preliminary molecular and biological characterization of Mourilyan virus (MoV): a new bunya-related virus of peaneid prawns. Dis. Asian Aquac. V 113–124.

Crab, R., Avnimelech, Y., Defoirdt, T., Bossier, P., Verstraete, W., 2007. Nitrogen removal techniques in aquaculture for a sustainable production. Aquaculture. https://doi.org/10.1016/j.aquaculture.2007.05.006

Crane, M., Hyatt, A., 2011. Viruses of fish: An overview of significant pathogens. Viruses 3, 2025–2046. https://doi.org/10.3390/v3112025

Cunningham, C.O., Gregory, A., Black, J., Simpson, I., Raynard, R.S., 2002. A novel variant of the infectious salmon anaemia virus (ISAV) haemagglutinin gene suggests mechanisms for virus diversity. Bull. Eur. Assoc. Fish Pathol.

Dale, O.B., Ørpetveit, I., Lyngstad, T.M., Kahns, S., Skall, H.F., Olesen, N.J., Dannevig, B.H., 2009. Outbreak of viral haemorrhagic septicaemia (VHS) in seawater-farmed rainbow trout in Norway caused by VHS virus Genotype III. Dis. Aquat. Organ. 85, 93–103. https://doi.org/10.3354/dao02065

Dannevig, B.H., Falk, K., Namork, E., 1995. Isolation of the causal virus of infectious salmon anaemia (ISA) in a long-term cell line from Atlantic salmon head kidney. J. Gen. Virol. 76, 1353–1359. https://doi.org/10.1099/0022-1317-76-6-1353

Datta, S., Budhauliya, R., Das, B., Chatterjee, S., Vanlalhmuaka, Veer, V., 2015. Next-generation sequencing in clinical virology: Discovery of new viruses. World J. Virol. 4, 265. https://doi.org/10.5501/wjv.v4.i3.265

de Abreu Corrêa, A., Albarnaz, J.D., Moresco, V., Poli, C.R., Teixeira, A.L., Oliveira Simões, C.M., Monte Barardi, C.R., 2007. Depuration dynamics of oysters (Crassostrea gigas) artificially contaminated by Salmonella enterica serovar Typhimurium. Mar. Environ. Res. https://doi.org/10.1016/j.marenvres.2006.12.002

Denton, J.F., Lugo-Martinez, J., Tucker, A.E., Schrider, D.R., Warren, W.C., Hahn, M.W., 2014. Extensive Error in the Number of Genes Inferred from Draft Genome Assemblies. PLoS Comput. Biol. 10, e1003998. https://doi.org/10.1371/journal.pcbi.1003998

Depledge, D.P., Palser, A.L., Watson, S.J., Lai, I.Y.C., Gray, E.R., Grant, P., Kanda, R.K., Leproust, E., Kellam, P., Breuer, J., 2011. Specific Capture and Whole-Genome Sequencing of Viruses from Clinical Samples. PLoS One 6. https://doi.org/10.1371/journal.pone.0027805

Depledge, D.P., Srinivas, K.P., Sadaoka, T., Bready, D., Mori, Y., Placantonakis, D.G., Mohr, I., Wilson, A.C., 2019. Direct RNA sequencing on nanopore arrays redefines the transcriptional complexity of a viral pathogen. Nat. Commun. https://doi.org/10.1038/s41467-019-08734-9

Descloux, E., Cao-Lormeau, V.M., Roche, C., De Lamballerie, X., 2009. Dengue 1 diversity and microevolution, French Polynesia 2001-2006: Connection with epidemiology and clinics. PLoS Negl. Trop. Dis. 3. https://doi.org/10.1371/journal.pntd.0000493

Dhar, A.K., Manna, S.K., Thomas Allnutt, F.C., 2014. Viral vaccines for farmed finfish. Indian J. Virol. 25, 1–17. https://doi.org/10.1007/s13337-013-0186-4

Dilthey, A., Cox, C., Iqbal, Z., Nelson, M.R., McVean, G., 2015. Improved genome inference in the MHC using a population reference graph. Nat. Genet. https://doi.org/10.1038/ng.3257

Domingo-Calap, P., Sanjuán, R., 2011. Experimental evolution of RNA versus DNA viruses. Evolution (N. Y). 65, 2987–2994. https://doi.org/10.1111/j.1558-5646.2011.01339.x

Domingo, E., Sheldon, J., Perales, C., 2012. Viral Quasispecies Evolution. Microbiol. Mol. Biol. Rev. 76, 159–216. https://doi.org/10.1128/mmbr.05023-11

Driscoll, C.B., Otten, T.G., Brown, N.M., Dreher, T.W., 2017. Towards long-read metagenomics: Complete assembly of three novel genomes from bacteria dependent on a diazotrophic cyanobacterium in a freshwater lake co-culture. Stand. Genomic Sci. https://doi.org/10.1186/s40793-017-0224-8

Drummond, A., Forsberg, R., Rodrigo, A.G., 2001. The Inference of Stepwise Changes in Substitution Rates Using Serial Sequence Samples. Mol. Biol. Evol. 18, 1365–1371. https://doi.org/10.1093/oxfordjournals.molbev.a003920

Drummond, A.J., Ho, S.Y.W., Phillips, M.J., Rambaut, A., 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4, 699–710. https://doi.org/10.1371/journal.pbio.0040088

Drummond, A.J., Nicholls, G.K., Rodrigo, A.G., Solomon, W., 2002. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics 161, 1307–1320. https://doi.org/10.1016/0370-2693(87)91197-X

Drummond, A.J., Rambaut, A., Shapiro, B., Pybus, O.G., 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. Mol. Biol. Evol. 22, 1185–1192. https://doi.org/10.1093/molbev/msi103

Drummond, A.J., Suchard, M.A., Xie, D., Rambaut, A., 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. 29, 1969–1973. https://doi.org/10.1093/molbev/mss075

Duffy, S., Shackelton, L.A., Holmes, E.C., 2008. Rates of evolutionary change in viruses: Patterns

and determinants. Nat. Rev. Genet. 9, 267–276. https://doi.org/10.1038/nrg2323

Edwards, A., Debbonaire, A.R., Nicholls, S.M., Rassner, S.M., Sattler, B., Cook, J.M., Davy, T., Soares, A.R., Mur, L.A., Hodson, A.J., 2019. In-field metagenome and 16S rRNA gene amplicon nanopore sequencing robustly characterize glacier microbiota. bioRxiv. https://doi.org/10.1101/073965

Edwards, A., Debbonaire, A.R., Sattler, B., Mur, L.A.J., Hodson, A.J., 2016. Extreme metagenomics using nanopore DNA sequencing: a field report from Svalbard, 78 °N. https://doi.org/10.1101/073965

EFSA, 2010. Scientific Opinion on the increased mortality events in Pacific oysters, Crassostrea gigas. EFSA J. 8, 1894–1953. https://doi.org/10.2903/j.efsa.2010.1894

Essbauer, S., Ahne, W., 2001. Viruses of lower vertebrates. J. Vet. Med. Ser. B. https://doi.org/10.1046/j.1439-0450.2001.00473.x

Euskirchen, P., Bielle, F., Labreche, K., Kloosterman, W.P., Rosenberg, S., Daniau, M., Schmitt, C., Masliah-Planchon, J., Bourdeaut, F., Dehais, C., Marie, Y., Delattre, J.Y., Idbaih, A., 2017. Same-day genomic and epigenomic diagnosis of brain tumors using real-time nanopore sequencing. Acta Neuropathol. https://doi.org/10.1007/s00401-017-1743-5

Eyngor, M., Zamostiano, R., Tsofack, J.E.K., Berkowitz, A., Bercovier, H., Tinman, S., Lev, M., Hurvitz, A., Galeotti, M., Bacharach, E., Eldar, A., 2014. Identification of a novel RNA virus lethal to tilapia. J. Clin. Microbiol. https://doi.org/10.1128/JCM.00827-14

FAO, 2018. WORLD FISHERIES AND AQUACULTURE.

FAO, 2016. The state of world fisheries and aquaculture 2016, The state of world fisheries and aquaculture. https://doi.org/92-5-105177-1

Faria, N.R., Sabino, E.C., Nunes, M.R.T., Alcantara, L.C.J., Loman, N.J., Pybus, O.G., 2016. Mobile real-time surveillance of Zika virus in Brazil. Genome Med. 8, 2–5. https://doi.org/10.1186/s13073-016-0356-2

Farley, C.A., Banfield, W.G., Kasnic, G., Foster, W.S., 1972. Oyster Herpes-Type Virus. Science (80-.). https://doi.org/10.1126/science.178.4062.759

Ferretti, L., Di Nardo, A., Singer, B., Lasecka-Dykes, L., Logan, G., Wright, C.F., Pérez-Martín, E., King, D.P., Tuthill, T.J., Ribeca, P., 2018. Within-host recombination in the foot-and-mouth disease virus genome. Viruses 10, 1–14. https://doi.org/10.3390/v10050221

Fisher, S., Barry, A., Abreu, J., Minie, B., Nolan, J., Delorey, T.M., Young, G., Fennell, T.J., Allen, A., Ambrogio, L., Berlin, A.M., Blumenstiel, B., Cibulskis, K., Friedrich, D., Johnson, R., Juhn, F., Reilly, B., Shammas, R., Stalker, J., Sykes, S.M., Thompson, J., Walsh, J., Zimmer, A., Zwirko, Z., Gabriel, S., Nicol, R., Nusbaum, C., 2011. A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. Genome Biol. https://doi.org/10.1186/gb-2011-12-1-r1

Flegel, T.W., 2006. Detection of major penaeid shrimp viruses in Asia, a historical perspective with emphasis on Thailand. Aquaculture. https://doi.org/10.1016/j.aquaculture.2006.05.013

Flegel, T.W., Lightner, D. V, Lo, C.H.U.F., Owens, L., 2008. Shrimp Disease Control : Past , Present and Future. Dis. Asian Aquac. VI 355–378.

Florea, L., Souvorov, A., Kalbfleisch, T.S., Salzberg, S.L., 2011. Genome Assembly Has a Major Impact on Gene Content: A Comparison of Annotation in Two Bos Taurus Assemblies. PLoS One 6, e21400. https://doi.org/10.1371/journal.pone.0021400

Focardi, S., Corsi, I., Franchi, E., 2005. Safety issues and sustainable development of European aquaculture: New tools for environmentally sound aquaculture. Aquac. Int. https://doi.org/10.1007/s10499-004-9036-0

Forrester, N.L., Guerbois, M., Adams, A.P., Liang, X., Weaver, S.C., 2011. Analysis of Intrahost Variation in Venezuelan Equine Encephalitis Virus Reveals Repeated Deletions in the 6-

Kilodalton Protein Gene. J. Virol. 85, 8709–8717. https://doi.org/10.1128/jvi.00165-11

Frank, J.A., Pan, Y., Tooming-Klunderud, A., Eijsink, V.G.H., McHardy, A.C., Nederbragt, A.J., Pope, P.B., 2016. Improved metagenome assemblies and taxonomic binning using long-read circular consensus sequence data. Sci. Rep. https://doi.org/10.1038/srep25373

Friedman, C., Estes, R., Stokes, N., Burge, C., Hargove, J., Barber, B., Elston, R., Burreson, E., Reece, K., 2005. Herpes virus in juvenile Pacific oysters Crassostrea gigas from Tomales Bay, California, coincides with summer mortality episodes. Dis. Aquat. Organ. 63, 33–41. https://doi.org/10.3354/dao063033

Fringuelli, E., Rowley, H.M., Wilson, J.C., Hunter, R., Rodger, H., Graham, D.A., 2008. Phylogenetic analyses and molecular epidemiology of European salmonid alphaviruses (SAV) based on partial E2 and nsP3 gene nucleotide sequences. J. Fish Dis. 31, 811–823. https://doi.org/10.1111/j.1365-2761.2008.00944.x

Furuse, Y., Suzuki, A., Oshitani, H., 2010. Origin of measles virus: Divergence from rinderpest virus between the 11th and 12th centuries. Virol. J. 7, 2–5. https://doi.org/10.1186/1743-422X-7-52

Futema, M., Plagnol, V., Whittall, R.A., Neil, H.A.W., Humphries, S.E., 2012. Use of targeted exome sequencing as a diagnostic tool for Familial Hypercholesterolaemia. J. Med. Genet. https://doi.org/10.1136/jmedgenet-2012-101189

Gagné, N., LeBlanc, F., 2018. Overview of infectious salmon anaemia virus (ISAV) in Atlantic Canada and first report of an ISAV North American-HPR0 subtype. J. Fish Dis. https://doi.org/10.1111/jfd.12670

Gallagher, M.D.M.D., Matejusova, I., Nguyen, L., Ruane, N.M.N.M., Falk, K., Macqueen, D.J.D.J., 2018. Nanopore sequencing for rapid diagnostics of salmonid RNA viruses. Sci. Rep. 8, 1–9. https://doi.org/10.1038/s41598-018-34464-x

Gallagher, M.D.M.D., Matejusova, I., Ruane, N.M.N.M., Macqueen, D.J.D.J., 2020. Genome-wide target enriched viral sequencing reveals extensive 'hidden' salmonid alphavirus diversity in farmed and wild fish populations. Aquaculture 522. https://doi.org/10.1016/j.aquaculture.2020.735117

García-Arenal, F., McDonald, B.A., 2003. An analysis of the durability of resistance to plant viruses. Phytopathology 93, 941–952. https://doi.org/10.1094/PHYTO.2003.93.8.941

Garcia, C., Thébault, A., Dégremont, L., Arzul, I., Miossec, L., Robert, M., Chollet, B., François, C., Joly, J.P., Ferrand, S., Kerdudou, N., Renault, T., 2011. Ostreid herpesvirus 1 detection and relationship with Crassostrea gigas spat mortality in France between 1998 and 2006. Vet. Res. 42. https://doi.org/10.1186/1297-9716-42-73

Garrison, E., 2012. A C++ library for parsing and manipulating VCF files. [WWW Document]. https://github.com/vcflib/vcflib.

Garrison, E., Marth, G., 2012. Haplotype-based variant detection from short-read sequencing 1–9.

Garseth, H., Fritsvold, C., Svendsen, J.C., Bang Jensen, B., Mikalsen, A.B., 2018. Cardiomyopathy syndrome in Atlantic salmon Salmo salar L.: A review of the current state of knowledge. J. Fish Dis. https://doi.org/10.1111/jfd.12735

Garver, K.A., LaPatra, S.E., Kurath, G., 2005. Efficacy of an infectious hematopoietic necrosis (IHN) virus DNA vaccine in Chinook Oncorhynchus tshawytscha and sockeye O. nerka salmon. Dis. Aquat. Organ. 64, 13–22. https://doi.org/10.3354/dao064013

Genissel, A., Confais, J., Lebrun, M.H., Gout, L., 2017. Association genetics in plant pathogens: Minding the gap between the natural variation and the molecular function. Front. Plant Sci. https://doi.org/10.3389/fpls.2017.01301

Geoghegan, J.L., Di Giallonardo, F., Cousins, K., Shi, M., Williamson, J.E., Holmes, E.C., 2018. Hidden diversity and evolution of viruses in market fish. Virus Evol. 4, 1–11. https://doi.org/10.1093/ve/vey031

Geoghegan, J.L., Duchêne, S., Holmes, E.C., 2017. Comparative analysis estimates the relative frequencies of co-divergence and cross-species transmission within viral families. PLoS Pathog. https://doi.org/10.1371/journal.ppat.1006215

Geoghegan, J.L., Holmes, E.C., 2017. Predicting virus emergence amid evolutionary noise. Open Biol. 7. https://doi.org/10.1098/rsob.170189

Geoghegan, J.L., Senior, A.M., Holmes, E.C., 2016. Pathogen population bottlenecks and adaptive landscapes: Overcoming the barriers to disease emergence. Proc. R. Soc. B Biol. Sci. https://doi.org/10.1098/rspb.2016.0727

Gill, O.N., Cubitt, W.D., McSwiggan, D.A., Watney, B.M., Bartlett, C.L., 1983. Epidemic of gastroenteritis caused by oysters contaminated with small round structured viruses. Br. Med. J. 287, 1532–1536. https://doi.org/10.1136/bmj.287.6404.1532

Glazebrook, J.S., Heasman, M.P., De beer, S.W., 1990. Picorna-like viral particles associated with mass mortalities in larval barramundi, Lates calcarifer Bloch. J. Fish Dis. https://doi.org/10.1111/j.1365-2761.1990.tb00780.x

Godoy, M.G., Kibenge, M.J., Suarez, R., Lazo, E., Heisinger, A., Aguinaga, J., Bravo, D., Mendoza, J., Llegues, K.O., Avendaño-Herrera, R., Vera, C., Mardones, F., Kibenge, F.S., 2013. Infectious salmon anaemia virus (ISAV) in Chilean Atlantic salmon (Salmo salar) aquaculture: Emergence of low pathogenic ISAV-HPR0 and re-emergence of virulent ISAV-HPR: HPR3 and HPR14. Virol. J. 10. https://doi.org/10.1186/1743-422X-10-344

Gontcharov, A.A., Marin, B., Melkonian, M., 2004. Are Combined Analyses Better Than Single Gene Phylogenies? A Case Study Using SSU rDNA and rbcL Sequence Comparisons in the Zygnematophyceae (Streptophyta). Mol. Biol. Evol. 21, 612–624. https://doi.org/10.1093/molbev/msh052

Gotesman, M., Kattlun, J., Bergmann, S.M., El-Matbouli, M., 2013. CyHV-3: The third cyprinid herpesvirus. Dis. Aquat. Organ. https://doi.org/10.3354/dao02614

Gourlé, H., Karlsson-Lindsjö, O., Hayer, J., Bongcam-Rudloff, E., 2019. Simulating Illumina metagenomic data with InSilicoSeq. Bioinformatics 35, 521–522. https://doi.org/10.1093/bioinformatics/bty630

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., Di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., Regev, A., 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29, 644–652. https://doi.org/10.1038/nbt.1883

Graham, D.A., Fringuelli, E., Rowley, H.M., Cockerill, D., Cox, D.I., Turnbull, T., Rodger, H., Morris, D., Mc Loughlin, M.F., 2012. Geographical distribution of salmonid alphavirus subtypes in marine farmed Atlantic salmon, salmo salar L., in Scotland and Ireland. J. Fish Dis. 35, 755–765. https://doi.org/10.1111/j.1365-2761.2012.01401.x

Graham, D.A., Frost, P., Mclaughlin, K., Rowley, H.M., Gabestad, I., Gordon, A., Mcloughlin, M.F., 2011. A comparative study of marine salmonid alphavirus subtypes 1-6 using an experimental cohabitation challenge model. J. Fish Dis. 34, 273–286. https://doi.org/10.1111/j.1365-2761.2010.01234.x

Graham, D.A., Staples, C., Wilson, C.J., Jewhurst, H., Cherry, K., Gordon, A., Rowley, H.M., 2007. Biophysical properties of salmonid alphaviruses: Influence of temperature and pH on virus survival. J. Fish Dis. 30, 533–543. https://doi.org/10.1111/j.1365-2761.2007.00811.x

Gray, R.R., Salemi, M., Klenerman, P., Pybus, O.G., 2012. A new evolutionary model for hepatitis C virus chronic infection. PLoS Pathog. 8. https://doi.org/10.1371/journal.ppat.1002656

Greenwald, W.W., Klitgord, N., Seguritan, V., Yooseph, S., Venter, J.C., Garner, C., Nelson, K.E., Li, W., 2017. Utilization of defined microbial communities enables effective evaluation of meta-genomic assemblies. BMC Genomics. https://doi.org/10.1186/s12864-017-3679-5

Grenfell, B.T., Pybus, O.G., Gog, J.R., Wood, J.L.N., Daly, J.M., Mumford, J.A., Holmes, E.C.,

2004. Unifying the Epidemiological and Evolutionary Dynamics of Pathogens. Science (80-.). 303, 327–332. https://doi.org/10.1126/science.1090727

Greninger, A.L., Naccache, S.N., Federman, S., Yu, G., Mbala, P., Bres, V., Stryke, D., Bouquet, J., Somasekar, S., Linnen, J.M., Dodd, R., Mulembakani, P., Schneider, B.S., Muyembe-Tamfum, J.J., Stramer, S.L., Chiu, C.Y., 2015. Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. Genome Med. 7, 1–13. https://doi.org/10.1186/s13073-015-0220-9

Griffin, D.W., Donaldson, K.A., Paul, J.H., Rose, J.B., 2003. Pathogenic human viruses in coastal waters. Clin. Microbiol. Rev. 16, 129–143. https://doi.org/10.1128/CMR.16.1.129-143.2003

Grover, C.E., Salmon, A., Wendel, J.F., 2012. Targeted sequence capture as a powerful tool for evolutionary analysis. Am. J. Bot. 99, 312–319. https://doi.org/10.3732/ajb.1100323

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. Syst. Biol. https://doi.org/10.1093/sysbio/syq010

Guo, R., Li, Y.R., He, S., Ou-Yang, L., Sun, Y., Zhu, Z., 2018. RepLong: De novo repeat identification using long read sequencing data. Bioinformatics. https://doi.org/10.1093/bioinformatics/btx717

Hagemann, I.S., Cottrell, C.E., Lockwood, C.M., 2013. Design of targeted, capture-based, next generation sequencing tests for precision cancer therapy. Cancer Genet. https://doi.org/10.1016/j.cancergen.2013.11.003

Hall, B.G., 2013. Building phylogenetic trees from molecular data with MEGA. Mol. Biol. Evol. 30, 1229–1235. https://doi.org/10.1093/molbev/mst012

Hall, T., 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp. Ser. 41, 95–98. https://doi.org/citeulike-article-id:691774

Hamlin, J.A.P., Dias, G.B., Bergman, C.M., Bensasson, D., 2019. Phased diploid genome assemblies for three strains of Candida albicans from oak trees. G3 Genes, Genomes, Genet. https://doi.org/10.1534/g3.119.400486

Hammoumi, S., Vallaeys, T., Santika, A., Leleux, P., Borzym, E., Klopp, C., Avarre, J.-C.C., 2016. Targeted genomic enrichment and sequencing of CyHV-3 from carp tissues confirms low nucleotide diversity and mixed genotype infections. PeerJ 2016, 1–17. https://doi.org/10.7717/peerj.2516

Harper, G.J., Steininger, M.K., Tucker, C.J., Juhn, D., Hawkins, F., 2007. Fifty years of deforestation and forest fragmentation in Madagascar. Environ. Conserv. 34, 325–333. https://doi.org/10.1017/S0376892907004262

Hasson, K.W., Lightner, D. V., Poulos, B.T., Redman, R.M., White, B.L., Brock, J.A., Bonami, J.R., 1995. Taura syndrome in Penaeus vannamei: Demonstration of a viral etiology. Dis. Aquat. Organ. https://doi.org/10.3354/dao023115

Haugland, O., Mikalsen, A.B., Nilsen, P., Lindmo, K., Thu, B.J., Eliassen, T.M., Roos, N., Rode, M., Evensen, O., 2011. Cardiomyopathy Syndrome of Atlantic Salmon (Salmo salar L.) Is Caused by a Double-Stranded RNA Virus of the Totiviridae Family. J. Virol. 85, 5275–5286. https://doi.org/10.1128/jvi.02154-10

Hedrick, R.P., Gilad, O., Yun, S., Spangenberg, J. V., Marty, G.D., Nordhausen, R.W., Kebus, M.J., Bercovier, H., Eldar, A., 2000. A herpesvirus associated with mass mortality of juvenile and adult koi, a strain of common carp. J. Aquat. Anim. Health. https://doi.org/10.1577/1548-8667(2000)012<0044:AHAWMM>2.0.CO;2

Hine, P., Wesney, B., Hay, B., 1992. Herpesviruses associated with mortalities among hatchery-reared larval Pacific oysters Crassostrea gigas. Dis. Aquat. Organ. 12, 135–142. https://doi.org/10.3354/dao012135

Hjortaas, M.J., Bang Jensen, B., Taksdal, T., Olsen, A.B., Lillehaug, A., Trettenes, E., Sindre, H., 2016. Genetic characterization of salmonid alphavirus in Norway. J. Fish Dis. 39, 249–257. https://doi.org/10.1111/jfd.12353

Hjortaas, M.J., Skjelstad, H.R., Taksdal, T., Olsen, A.B., Johansen, R., Bang-Jensen, B., Ørpetveit, I., Sindre, H., 2013. The first detections of subtype 2-related salmonid alphavirus (SAV2) in Atlantic salmon, Salmo salar L., in Norway. J. Fish Dis. 36, 71–74. https://doi.org/10.1111/j.1365-2761.2012.01445.x

Hoa, T.T.T., Zwart, M.P., Phuong, N.T., Vlak, J.M., De Jong, M.C.M., 2011. Transmission of white spot syndrome virus in improved-extensive and semi-intensive shrimp production systems: A molecular epidemiology study. Aquaculture 313, 7–14. https://doi.org/10.1016/j.aquaculture.2011.01.013

Hodneland, K., Bratland, A., Christie, K.E.E., Endresen, C., Nylund, A., 2005. Erratum: New subtype of salmonid alphavirus (SAV), Togaviridae, from Atlantic salmon Salmo salar and rainbow trout Oncorhynchus mykiss in Norway (Diseases of Aquatic Organisms (2005) 66 (113-120)). Dis. Aquat. Organ. 67, 181. https://doi.org/10.3354/dao066113

Hodneland, K., Endresen, C., 2006. Sensitive and specific detection of Salmonid alphavirus using real-time PCR (TaqMan®). J. Virol. Methods 131, 184–192. https://doi.org/10.1016/j.jviromet.2005.08.012

Hoelzer, K., Murcia, P.R.R., Baillie, G.J.J., Wood, J.L.N.L.N., Metzger, S.M.M., Osterrieder, N., Dubovi, E.J.J., Holmes, E.C.C., Parrish, C.R.R., 2010. Intrahost Evolutionary Dynamics of Canine Influenza Virus in Naive and Partially Immune Dogs. J. Virol. 84, 5329–5335. https://doi.org/10.1128/jvi.02469-09

Hoenen, T., Groseth, A., Rosenke, K., Fischer, R.J., Hoenen, A., Judson, S.D., Martellaro, C., Falzarano, D., Marzi, A., Squires, R.B., Wollenberg, K.R., De Wit, E., Prescott, J., Safronetz, D., Van Doremalen, N., Bushmaker, T., Feldmann, F., McNally, K., Bolay, F.K., Fields, B., Sealy, T., Rayfield, M., Nichol, S.T., Zoon, K.C., Massaquoi, M., Munster, V.J., Feldmann, H., 2016. Nanopore sequencing as a rapidly deployable Ebola outbreak tool. Emerg. Infect. Dis. 22, 331–334. https://doi.org/10.3201/eid2202.151796

Holland, J., Spindler, K., Horodyski, F., Grabau, E., Nichol, S., VandePol, S., 1982. Rapid evolution of RNA genomes. Science (80-.). 215.

Holmes, E.C., 2009. The Evolutionary Genetics of Emerging Viruses. Annu. Rev. Ecol. Evol. Syst. 40, 353–372. https://doi.org/10.1146/annurev.ecolsys.110308.120248

Holmes, E.C., 2004. The phylogeography of human viruses. Mol. Ecol. https://doi.org/10.1046/j.1365-294X.2003.02051.x

Holmes, E.C., Dudas, G., Rambaut, A., Andersen, K.G., 2016. The evolution of Ebola virus: Insights from the 2013–2016 epidemic. Nature 538, 193–200. https://doi.org/10.1038/nature19790

Holopainen, R., Eriksson-Kallio, A.M., Gadd, T., 2017. Molecular characterisation of infectious pancreatic necrosis viruses isolated from farmed fish in Finland. Arch. Virol. 162, 3459–3471. https://doi.org/10.1007/s00705-017-3525-8

Houldcroft, C.J., Beale, M.A., Breuer, J., 2017. Clinical and biological insights from viral genome sequencing. Nat. Rev. Microbiol. 15, 183–192. https://doi.org/10.1038/nrmicro.2016.182

Howe, A., Chain, P.S.G., 2015. Challenges and opportunities in understanding microbial communities with metagenome assembly (accompanied by IPython Notebook tutorial). Front. Microbiol. https://doi.org/10.3389/fmicb.2015.00678

Hrdlickova, R., Toloue, M., Tian, B., 2017. RNA-Seq methods for transcriptome analysis. Wiley Interdiscip. Rev. RNA. https://doi.org/10.1002/wrna.1364

Huson, D.H., Mitra, S., Ruscheweyh, H.J., Weber, N., Schuster, S.C., 2011. Integrative analysis of environmental sequences using MEGAN4. Genome Res. 21, 1552–1560. https://doi.org/10.1101/gr.120618.111

Inouye, K., Yamano, K., Maeno, Y., Nakajima, K., Matsuoka, M., Wada, Y., Sorimachi, M., 1992. Iridovirus Infection of Cultured Red Sea Bream, Pagrus major. Fish Pathol. https://doi.org/10.3147/jsfp.27.19

Ioannidis, J.P.A., Thomas, G., Daly, M.J., 2009. Validating, augmenting and refining genome-wide association signals. Nat. Rev. Genet. https://doi.org/10.1038/nrg2544

Iqbal, M., Xiao, H., Baillie, G., Warry, A., Essen, S.C., Londt, B., Brookes, S.M., Brown, I.H., McCauley, J.W., 2009. Within-host variation of avian influenza viruses. Philos. Trans. R. Soc. B Biol. Sci. 364, 2739–2747. https://doi.org/10.1098/rstb.2009.0088

Ito, T., Kurita, J., Yuasa, K., 2014. Differences in the susceptibility of Japanese indigenous and domesticated Eurasian common carp (Cyprinus carpio), identified by mitochondrial DNA typing, to cyprinid herpesvirus 3 (CyHV-3). Vet. Microbiol. 171, 31–40. https://doi.org/10.1016/j.vetmic.2014.03.002

Jackson, C., Preston, N., Thompson, P.J., Burford, M., 2003. Nitrogen budget and effluent nitrogen components at an intensive shrimp farm. Aquaculture 218, 397–411. https://doi.org/10.1016/S0044-8486(03)00014-0

Jacobsen, C.S., 1995. Microscale detection of specific bacterial DNA in soil with a magnetic capture-hybridization and PCR amplification assay. Appl. Environ. Microbiol. https://doi.org/10.1128/aem.61.9.3347-3352.1995

Jansen, M.D., Gjerset, B., Modahl, I., Bohlin, J., 2010. Molecular epidemiology of salmonid alphavirus (SAV) subtype 3 in Norway. Virol. J. 7, 1–8. https://doi.org/10.1186/1743-422X-7-188

Jenkins, C., Hick, P., Gabor, M., Spiers, Z., Fell, S., Gu, X., Read, A., Go, J., Dove, M., O'Connor, W., Kirkland, P., Frances, J., 2013. Identification and characterisation of an ostreid herpesvirus-1 microvariant (OsHV-1 µ-var) in Crassostrea gigas (Pacific oysters) in Australia. Dis. Aquat. Organ. 105, 109–126. https://doi.org/10.3354/dao02623

Jennings, S., Stentiford, G.D., Leocadio, A.M., Jeffery, K.R., Metcalfe, J.D., Katsiadaki, I., Auchterlonie, N.A., Mangi, S.C., Pinnegar, J.K., Ellis, T., Peeler, E.J., Luisetti, T., Baker-Austin, C., Brown, M., Catchpole, T.L., Clyne, F.J., Dye, S.R., Edmonds, N.J., Hyder, K., Lee, J., Lees, D.N., Morgan, O.C., O'Brien, C.M., Oidtmann, B., Posen, P.E., Santos, A.R., Taylor, N.G.H., Turner, A.D., Townhill, B.L., Verner-Jeffreys, D.W., 2016. Aquatic food security: insights into challenges and solutions from an analysis of interactions between fisheries, aquaculture, food safety, human health, fish and human welfare, economy and environment. Fish Fish. 17, 893–938. https://doi.org/10.1111/faf.12152

Jensen, M.H., 1965. Research on the Virus of Egtved Disease. Ann. N. Y. Acad. Sci. 126, 422–426. https://doi.org/10.1111/j.1749-6632.1965.tb14292.x

Jia, H., Guo, Y., Zhao, W., Wang, K., 2014. Long-range PCR in next-generation sequencing: Comparison of six enzymes and evaluation on the MiSeq sequencer. Sci. Rep. 4. https://doi.org/10.1038/srep05737

Johnson, S.S., Zaikova, E., Goerlitz, D.S., Bai, Y., Tighe, S.W., 2017. Real-time DNA sequencing in the antarctic dry valleys using the Oxford nanopore sequencer. J. Biomol. Tech. https://doi.org/10.7171/jbt.17-2801-009

Jones, M.R., Good, J.M., 2016. Targeted capture in evolutionary and ecological genomics, Molecular Ecology. Blackwell Publishing Ltd. https://doi.org/10.1111/mec.13304

Jones, S., Baizan-Edge, A., MacFarlane, S., Torrance, L., 2017. Viral diagnostics in plants using next generation sequencing: Computational analysis in practice. Front. Plant Sci. 8. https://doi.org/10.3389/fpls.2017.01770

Kafetzopoulou, L.E., Efthymiadis, K., Lewandowski, K., Crook, A., Carter, D., Osborne, J., Aarons, E., Hewson, R., Hiscox, J.A., Carroll, M.W., Vipond, R., Pullan, S.T., 2018. Assessment of Metagenomic MinION and Illumina sequencing as an approach for the recovery of whole genome sequences of chikungunya and dengue viruses directly from clinical samples. bioRxiv

355560. https://doi.org/10.1101/355560

Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., Von Haeseler, A., Jermiin, L.S., 2017. ModelFinder: Fast model selection for accurate phylogenetic estimates. Nat. Methods 14, 587–589. https://doi.org/10.1038/nmeth.4285

Karlsen, M., Gjerset, B., Hansen, T., Rambaut, A., 2014. Multiple introductions of salmonid alphavirus from a wild reservoir have caused independent and self-sustainable epizootics in aquaculture. J. Gen. Virol. 95, 52–59. https://doi.org/10.1099/vir.0.057455-0

Karlsen, M., Hodneland, K., Endresen, C., Nylund, A., 2006. Genetic stability within the Norwegian subtype of salmonid alphavirus (family Togaviridae). Arch. Virol. 151, 861–874. https://doi.org/10.1007/s00705-005-0687-6

Karlsen, M., Tingbø, T., Solbakk, I.T., Evensen, Ø., Furevik, A., Aas-Eng, A., 2012. Efficacy and safety of an inactivated vaccine against Salmonid alphavirus (family Togaviridae). Vaccine 30, 5688–5694. https://doi.org/10.1016/j.vaccine.2012.05.069

Karst, S.M., Ziels, R.M., Kirkegaard, R.H., Sørensen, E.A., McDonald, D., Zhu, Q., Knight, R., Albertsen, M. 2020. Enabling high-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing. BioRxiv DOI: https://doi.org/10.1101/645903

Katoh, K., Rozewicki, J., Yamada, K.D., 2019. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief. Bioinform. 20, 1160–1166. https://doi.org/10.1093/bib/bbx108

Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol. Biol. Evol. 30, 772–780. https://doi.org/10.1093/molbev/mst010

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., Drummond, A., 2012. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28, 1647–1649. https://doi.org/10.1093/bioinformatics/bts199

Kearse, M., Sturrock, S., Meintjes, P., n.d. The Geneious 6.0.3 read mapper.

Keller, M.W., Rambo-Martin, B.L., Wilson, M.M., Ridenour, C.A., Shepard, S.S., Stark, T.J., Neuhaus, E.B., Dugan, V.G., Wentworth, D.E., Barnes, J.R., 2018. Direct RNA Sequencing of the Coding Complete Influenza A Virus Genome. Sci. Rep. https://doi.org/10.1038/s41598-018-32615-8

Kempfer, R., Pombo, A., 2019. Methods for mapping 3D chromosome architecture. Nat. Rev. Genet. https://doi.org/10.1038/s41576-019-0195-2

Kennedy, D.A., Kurath, G., Brito, I.L., Purcell, M.K., Read, A.F., Winton, J.R., Wargo, A.R., 2016. Potential drivers of virulence evolution in aquaculture. Evol. Appl. 9, 344–354. https://doi.org/10.1111/eva.12342

Kibenge, F.S.B., Godoy, M.G., Fast, M., Workenhe, S., Kibenge, M.J.T., 2012. Countermeasures against viral diseases of farmed fish. Antiviral Res. 95, 257–281. https://doi.org/10.1016/j.antiviral.2012.06.003

Kibenge, F.S.B., Godoy, M.G., Wang, Y., Kibenge, M.J.T., Gherardelli, V., Mansilla, S., Lisperger, A., Jarpa, M., Larroquete, G., Avendaño, F., Lara, M., Gallardo, A., Avendão, F., Lara, M., Gallardo, A., Godoy, M.G., Wang, Y., 2009. Infectious salmon anaemia virus (ISAV) isolated from the ISA disease outbreaks in Chile diverged from ISAV isolates from Norway around 1996 and was disseminated around 2005, based on surface glycoprotein gene sequences. Virol. J. 6, 1–16. https://doi.org/10.1186/1743-422X-6-88

Kibenge, F.S.B., Kibenge, M.J.T., Groman, D., McGeachy, S., 2006. In vivo correlates of infectious salmon anemia virus pathogenesis in fish. J. Gen. Virol. 87, 2645–2652. https://doi.org/10.1099/vir.0.81719-0

Kibenge, F.S.B., Kibenge, M.J.T., Wang, Y., Qian, B., Hariharan, S., McGeachy, S., 2007. Mapping of putative virulence motifs on infectious salmon anemia virus surface glycoprotein genes. J. Gen. Virol. 88, 3100–3111. https://doi.org/10.1099/vir.0.83097-0

Kimura, T., Yoshimizu, M., Gorie, S., 1985. A new rhabdovirus isolated in Japan from cultured hirame (Japanese flounder) Paralichthys olivaceus and ayu Plecoglossus altivelis. Dis. Aquat. Organ. https://doi.org/10.3354/dao001209

Kono, N., Arakawa, K., 2019. Nanopore sequencing: Review of potential applications in functional genomics. Dev. Growth Differ. https://doi.org/10.1111/dgd.12608

Koo, O.K., Liu, Y.S., Shuaib, S., Bhattacharya, S., Ladisch, M.R., Bashir, R., Bhunia, A.K., 2009. Targeted capture of pathogenic bacteria using a mammalian cell receptor coupled with dielectrophoresis on a biochip. Anal. Chem. https://doi.org/10.1021/ac9000833

Koonin, E. V., Senkevich, T.G., Dolja, V. V., 2006. The ancient virus world and evolution of cells. Biol. Direct 1, 1–27. https://doi.org/10.1186/1745-6150-1-29

Köster, J., Rahmann, S., 2012. Snakemake-a scalable bioinformatics workflow engine. Bioinformatics. https://doi.org/10.1093/bioinformatics/bts480

Kristoffersen, A.B., Viljugrein, H., Kongtorp, R.T., Brun, E., Jansen, P.A., 2009. Risk factors for pancreas disease (PD) outbreaks in farmed Atlantic salmon and rainbow trout in Norway during 2003-2007. Prev. Vet. Med. 90, 127–136. https://doi.org/10.1016/j.prevetmed.2009.04.003

Krueger, F., 2015. Trim Galore [WWW Document]. https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.

Kühnert, D., Wu, C.H., Drummond, A.J., 2011. Phylogenetic and epidemic modeling of rapidly evolving infectious diseases. Infect. Genet. Evol. https://doi.org/10.1016/j.meegid.2011.08.005

Kulshreshtha, V., Kibenge, M., Salonius, K., Simard, N., Riveroll, A., Kibenge, F., 2010. Identification of the 3' and 5' terminal sequences of the 8 rna genome segments of european and north american genotypes of infectious salmon anemia virus (an orthomyxovirus) and evidence for quasispecies based on the non-coding sequences of transcripts. Virol. J. 7, 1–19. https://doi.org/10.1186/1743-422X-7-338

Lafferty, K.D., Harvell, C.D., Conrad, J.M., Friedman, C.S., Kent, M.L., Kuris, A.M., Powell, E.N., Rondeau, D., Saksida, S.M., 2015. Infectious Diseases Affect Marine Fisheries and Aquaculture Economics. Ann. Rev. Mar. Sci. 7, 471–496. https://doi.org/10.1146/annurev-marine-010814-015646

Lai, M.M.C., 1992. Genetic recombination in RNA viruses. Curr. Top. Microbiol. Immunol. https://doi.org/10.1007/978-3-642-77011-1_2

Lancaster, K.Z., Pfeiffer, J.K., 2012. Viral population dynamics and virulence thresholds. Curr. Opin. Microbiol. 15, 525–30. https://doi.org/10.1016/j.mib.2012.05.007

Langdon, J.S., Humphrey, J.D., 1987. Epizootic haematopoietic necrosis, a new viral disease in redfin perch, Perca fluviatilis L., in Australia. J. Fish Dis. https://doi.org/10.1111/j.1365-2761.1987.tb01073.x

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359. https://doi.org/10.1038/nmeth.1923

Lappin, F.M., Shaw, R.L., Macqueen, D.J., 2016. Targeted sequencing for high-resolution evolutionary analyses following genome duplication in salmonid fish: Proof of concept for key components of the insulin-like growth factor axis. Mar. Genomics 30, 15–26. https://doi.org/10.1016/j.margen.2016.06.003

Lauring, A.S., Frydman, J., Andino, R., 2013. The role of mutational robustness in RNA virus evolution. Nat. Rev. Microbiol. 11, 327–36. https://doi.org/10.1038/nrmicro3003

Laver, T., Harrison, J., O'Neill, P.A.A., Moore, K., Farbos, A., Paszkiewicz, K., Studholme, D.J.J.J., O'Neill, P.A., Moore, K., Farbos, A., Paszkiewicz, K., Studholme, D.J.J.J., 2015. Assessing the

performance of the Oxford Nanopore Technologies MinION. Biomol. Detect. Quantif. 3, 1–8. https://doi.org/10.1016/j.bdq.2015.02.001

LeBlanc, F., Leadbeater, S., Laflamme, M., Gagné, N., 2018. In vivo virulence and genomic comparison of infectious Salmon Anaemia Virus isolates from Atlantic Canada. J. Fish Dis. 41, 1373–1384. https://doi.org/10.1111/jfd.12832

Lemey, P., Kosakovsky Pond, S.L., Drummond, A.J., Pybus, O.G., Shapiro, B., Barroso, H., Taveira, N., Rambaut, A., 2007. Synonymous substitution rates predict HIV disease progression as a result of underlying replication dynamics. PLoS Comput. Biol. https://doi.org/10.1371/journal.pcbi.0030029

Lemey, P., Rambaut, A., Drummond, A.J., Suchard, M.A., 2009. Bayesian phylogeography finds its roots. PLoS Comput. Biol. 5. https://doi.org/10.1371/journal.pcbi.1000520

Lemey, P., Rambaut, A., Pybus, O.G., 2006. HIV evolutionary dynamics within and among hosts. AIDS Rev.

Lesnik, E.A., Freier, S.M., 1995. Relative Thermodynamic Stability of DNA, RNA, and DNA:RNA Hybrid Duplexes: Relationship with Base Composition and Structure. Biochemistry 34, 10807–10815. https://doi.org/10.1021/bi00034a013

Lewisch, E., Frank, T., Soliman, H., Schachner, O., Friedl, A., El-Matbouli, M., 2018. First confirmation of salmonid alphavirus infection in Arctic char Salvelinus alpinus and in Austria. Dis. Aquat. Organ. https://doi.org/10.3354/dao03265

Li, C.X., Shi, M., Tian, J.H., Lin, X.D., Kang, Y.J., Chen, L.J., Qin, X.C., Xu, J., Holmes, E.C., Zhang, Y.Z., 2015. Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. Elife. https://doi.org/10.7554/eLife.05378

Li, H., 2018. Minimap2: Pairwise alignment for nucleotide sequences. Bioinformatics 34, 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Li, H., 2015. Seqtk: Toolkit for processing sequences in FASTA/Q formats [WWW Document]. GitHub.

Li, H., 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM 00, 1–3.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, J., Wang, M., Yu, D., Han, Y., Yang, Z., Wang, L., Zhang, X., Liu, F., 2017. A comparative study on the characterization of hepatitis B virus quasispecies by clone-based sequencing and third-generation sequencing. Emerg. Microbes Infect. 6, e100. https://doi.org/10.1038/emi.2017.88

Li, W., Shi, Z., Yu, M., Ren, W., Smith, C., Epstein, J.H., Wang, H., Crameri, G., Hu, Z., Zhang, H., Zhang, J., McEachern, J., Field, H., Daszak, P., Eaton, B.T., Zhang, S., Wang, L.F., 2005. Bats are natural reservoirs of SARS-like coronaviruses. Science (80-.). 310, 676–679. https://doi.org/10.1126/science.1118391

Lightner, D. V., Redman, R.M., 1985. A parvo-like virus disease of penaeid shrimp. J. Invertebr. Pathol. https://doi.org/10.1016/0022-2011(85)90048-5

Limasset, A., Cazaux, B., Rivals, E., Peterlongo, P., 2016. Read mapping on de Bruijn graphs. BMC Bioinformatics. https://doi.org/10.1186/s12859-016-1103-9

Lindahl, J.F., Grace, D., 2015. The consequences of human actions on risks for infectious diseases: a review. Infect. Ecol. Epidemiol. 5, 30048. https://doi.org/10.3402/iee.v5.30048

Lindgreen, S., Adair, K.L., Gardner, P.P., 2016. An evaluation of the accuracy and speed of metagenome analysis tools. Sci. Rep. https://doi.org/10.1038/srep19233

Ling, J., Smura, T., Lundström, J.O., Pettersson, J.H.-O., Sironen, T., Vapalahti, O., Lundkvist, Å., Hesson, J.C., 2019. Introduction and Dispersal of Sindbis Virus from Central Africa to Europe.

J. Virol. 93. https://doi.org/10.1128/jvi.00620-19

Lipp, E.K., Jarrell, J.L., Griffin, D.W., Lukasik, J., Jacukiewicz, J., Rose, J.B., 2002. Preliminary evidence for human fecal contamination in corals of the Florida Keys, USA. Mar. Pollut. Bull. 44, 666–670. https://doi.org/10.1016/S0025-326X(01)00332-0

Lobo, J., Costa, P.M., Teixeira, M.A.L., Ferreira, M.S.G., Costa, M.H., Costa, F.O., 2013. Enhanced primers for amplification of DNA barcodes from a broad range of marine metazoans. BMC Ecol. 13, 34. https://doi.org/10.1186/1472-6785-13-34

Loisy, F., Atmar, R.L., Le Saux, J.C., Cohen, J., Caprais, M.P., Pommepuy, M., Le Guyader, F.S., 2005. Use of rotavirus virus-like particles as surrogates to evaluate virus persistence in shellfish. Appl. Environ. Microbiol. 71, 6049–6053. https://doi.org/10.1128/AEM.71.10.6049-6053.2005

Lonigro, R.J., Grasso, C.S., Robinson, D.R., Jing, X., Wu, Y.M., Cao, X., Quist, M.J., Tomlins, S.A., Pienta, K.J., Chinnaiyan, A.M., 2011. Detection of somatic copy number alterations in cancerusing targeted exome capture sequencing. Neoplasia. https://doi.org/10.1593/neo.111252

Løvoll, M., Wiik-Nielsen, J., Grove, S., Wiik-Nielsen, C.R., Kristoffersen, A.B., Faller, R., Poppe, T., Jung, J., Pedamallu, C.S., Nederbragt, A.J., Meyerson, M., Rimstad, E., Tengs, T., 2010. A novel totivirus and piscine reovirus (PRV) in Atlantic salmon (Salmo salar) with cardiomyopathy syndrome (CMS). Virol. J. 7, 1–7. https://doi.org/10.1186/1743-422X-7-309

Lu, S., Wang, J., Chitsaz, F., Derbyshire, M.K., Geer, R.C., Gonzales, N.R., Gwadz, M., Hurwitz, D.I., Marchler, G.H., Song, J.S., Thanki, N., Yamashita, R.A., Yang, M., Zhang, D., Zheng, C., Lanczycki, C.J., Marchler-Bauer, A., 2020. CDD/SPARCLE: the conserved domain database in 2020. Nucleic Acids Res. 48, D265–D268. https://doi.org/10.1093/nar/gkz991

Lumsden, J., Morrison, B., Yason, C., Russell, S., Young, K., Yazdanpanah, A., Huber, P., Al-Hussinee, L., Stone, D., Way, K., 2007. Mortality event in freshwater drum Aplodinotus grunniens from Lake Ontario, Canada, associated with viral haemorrhagic septicemia virus, Type IV. Dis. Aquat. Organ. 76, 99–111. https://doi.org/10.3354/dao076099

Lyngstad, T.M., Kristoffersen, A.B., Hjortaas, M.J., Devold, M., Aspehaug, V., Larssen, R.B., Jansen, P.A., 2012. Low virulent infectious salmon anaemia virus (ISAV-HPR0) is prevalent and geographically structured in Norwegian salmon farming. Dis. Aquat. Organ. 101, 197–206. https://doi.org/10.3354/dao02520

Ma, L., Su, L., Liu, H., Zhao, F., Zhou, D., Duan, D., 2017. Norovirus contamination and the glycosphingolipid biosynthesis pathway in Pacific oyster: A transcriptomics study. Fish Shellfish Immunol. 66, 26–34. https://doi.org/10.1016/j.fsi.2017.04.023

Macklaim, J.M., Fernandes, A.D., Di Bella, J.M., Hammond, J.A., Reid, G., Gloor, G.B., 2013. Comparative meta-RNA-seq of the vaginal microbiota and differential expression by Lactobacillus iners in health and dysbiosis. Microbiome. https://doi.org/10.1186/2049-2618-1-12

Madhun, A.S., Isachsen, C.H., Omdal, L.M., Einen, A.C.B., Mæhle, S., Wennevik, V., Niemelä, E., Svåsand, T., Karlsbakk, E., 2018. Prevalence of piscine orthoreovirus and salmonid alphavirus in sea-caught returning adult Atlantic salmon (Salmo salar L.) in northern Norway. J. Fish Dis. 41, 797–803. https://doi.org/10.1111/jfd.12785

Marano, N., Arguin, P.M., Pappaioanou, M., 2007. Impact of globalization and animal trade on infectious disease ecology. Emerg. Infect. Dis. 13, 1807–1809. https://doi.org/10.3201/eid1312.071276

Margeridon-Thermet, S., Shulman, N.S., Ahmed, A., Shahriar, R., Liu, T., Wang, C., Holmes, S.P., Babrzadeh, F., Gharizadeh, B., Hanczaruk, B., Simen, B.B., Egholm, M., Shafer, R.W., 2009. Ultra-Deep Pyrosequencing of Hepatitis B Virus Quasispecies from Nucleoside and Nucleotide Reverse-Transcriptase Inhibitor (NRTI)–Treated Patients and NRTI-Naive Patients. J. Infect. Dis. 199, 1275–1285. https://doi.org/10.1086/597808

Marhaver, K.L., Edwards, R.A., Rohwer, F., 2008. Viral communities associated with healthy and bleaching corals. Environ. Microbiol. 10, 2277–2286. https://doi.org/10.1111/j.1462-

2920.2008.01652.x

Markussen, T., Jonassen, C.M., Numanovic, S., Braaen, S., Hjortaas, M., Nilsen, H., Mjaaland, S., 2008. Evolutionary mechanisms involved in the virulence of infectious salmon anaemia virus (ISAV), a piscine orthomyxovirus. Virology 374, 515–527. https://doi.org/10.1016/j.virol.2008.01.019

Markussen, T., Sindre, H., Jonassen, C.M., Tengs, T., Kristoffersen, A.B., Ramsell, J., Numanovic, S., Hjortaas, M.J., Christiansen, D.H., Dale, O.B., Falk, K., 2013. Ultra-deep pyrosequencing of partial surface protein genes from infectious salmon anaemia virus (ISAV) suggest novel mechanisms involved in transition to virulence. PLoS One 8, 1–11. https://doi.org/10.1371/journal.pone.0081571

Martens, M., Dawyndt, P., Coopman, R., Gillis, M., De Vos, P., Willems, A., 2008. Advantages of multilocus sequence analysis for taxonomic studies: A case study using 10 housekeeping genes in the genus Ensifer (including former Sinorhizobium). Int. J. Syst. Evol. Microbiol. 58, 200–214. https://doi.org/10.1099/ijs.0.65392-0

Matejusova, I., Lester, K., Li, Z., Bravo, J., Bland, F., Collet, B., 2013. Comparison of complete polyprotein sequences of two isolates of salmon alphavirus (SAV) type I and their behaviour in a salmonid cell line. Arch. Virol. 158, 2143–2146. https://doi.org/10.1007/s00705-013-1689-4

Mavromatis, K., Ivanova, N., Barry, K., Shapiro, H., Goltsman, E., McHardy, A.C., Rigoutsos, I., Salamov, A., Korzeniewski, F., Land, M., Lapidus, A., Grigoriev, I., Richardson, P., Hugenholtz, P., Kyrpides, N.C., 2007. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. Nat. Methods. https://doi.org/10.1038/nmeth1043

McArdle, A.J., Turkova, A., Cunnington, A.J., 2018. When do co-infections matter? Curr. Opin. Infect. Dis. https://doi.org/10.1097/QCO.0000000000000447

McBeath, A.J.A., Bain, N., Snow, M., 2009. Surveillance for infectious salmon anaemia virus HPR0 in marine Atlantic salmon farms across Scotland. Dis. Aquat. Organ. https://doi.org/10.3354/dao02128

McCleary, S., Giltrap, M., Henshilwood, K., Ruane, N.M., 2014. Detection of salmonid alphavirus RNA in Celtic and Irish Sea flatfish. Dis. Aquat. Organ. 109, 1–7. https://doi.org/10.3354/dao02719

McDaniel, L., Breitbart, M., Mobberley, J., Long, A., Haynes, M., Rohwer, F., Paul, J.H., 2008. Metagenomic analysis of lysogeny in Tampa Bay: Implications for prophage gene expression. PLoS One 3. https://doi.org/10.1371/journal.pone.0003263

McGonigle, R.H., 1941. Acute Catarrhal Enteritis of Salmonid Fingerlings. Trans. Am. Fish. Soc. https://doi.org/10.1577/1548-8659(1940)70[297:aceosf]2.0.co;2

McKinley, T.J., Murcia, P.R., Gog, J.R., Varela, M., Wood, J.L.N., 2011. A bayesian approach to analyse genetic variation within RNA viral populations. PLoS Comput. Biol. 7. https://doi.org/10.1371/journal.pcbi.1002027

McLoughlin, M.F., Graham, D.A., 2007. Alphavirus infections in salmonids - A review. J. Fish Dis. 30, 511–531. https://doi.org/10.1111/j.1365-2761.2007.00848.x

McMichael, A.J., 2002. Population, environment, disease, and survival: Past patterns, uncertain futures. Lancet 359, 1145–1148. https://doi.org/10.1016/S0140-6736(02)08164-3

Mello, D.F., Danielli, N.M., Curbani, F., Pontinha, V.A., Suhnel, S., Castro, M.A.M., Medeiros, S.C., Wendt, N.C., Trevisan, R., Magalhães, A.R.M., Dafre, A.L., 2018. First evidence of viral and bacterial oyster pathogens in the Brazilian coast. J. Fish Dis. 41, 559–563. https://doi.org/10.1111/jfd.12755

Melnick, J.L., 2015. Etiologic Agents and Their Potential for Causing Waterborne Virus Diseases. Karger Publishers, pp. 1–16. https://doi.org/10.1159/000409113

Merour, E., LeBerre, M., Lamoureux, A., Bernard, J., Bremont, M., Biacchesi, S., 2011. Completion of the full-length genome sequence of the infectious salmon anemia virus, an aquatic

orthomyxovirus-like, and characterization of mAbs. J. Gen. Virol. 92, 528–533. https://doi.org/10.1099/vir.0.027417-0

Mikalsen, A.B., Nilsen, P., Frøystad-Saugen, M., Lindmo, K., Eliassen, T.M., Rode, M.R., Evensen, Ø., 2014. Characterization of a novel calicivirus causing systemic infection in Atlantic salmon (Salmo salarL.): Proposal for a new genus of caliciviridae. PLoS One 9. https://doi.org/10.1371/journal.pone.0107132

Miller, M.D., Bor, Y.C., Bushman, F., 1995. Target DNA capture by HIV-1 integration complexes. Curr. Biol. 5, 1047–1056. https://doi.org/10.1016/S0960-9822(95)00209-0

Miller, O., Fuller, F.J., Gebreyes, W.A., Lewbart, G.A., Shchelkunov, I.S., Shivappa, R.B., Joiner, C., Woolford, G., Stone, D.M., Dixon, P.F., Raley, M.E., Levine, J.F., 2007. Phylogenetic analysis of spring virema of carp virus reveals distinct subgroups with common origins for recent isolates in North America and the UK. Dis. Aquat. Organ. 76, 193–204. https://doi.org/10.3354/dao076193

Minh, B.Q., Nguyen, M.A.T., Von Haeseler, A., 2013. Ultrafast approximation for phylogenetic bootstrap. Mol. Biol. Evol. 30, 1188–1195. https://doi.org/10.1093/molbev/mst024

Minot, S., Grunberg, S., Wu, G.D., Lewis, J.D., Bushman, F.D., 2012. Hypervariable loci in the human gut virome. Proc. Natl. Acad. Sci. U. S. A. https://doi.org/10.1073/pnas.1119061109

Mjaaland, S., Hungnes, O., Teig, A., Dannevig, B.H., Thorud, K., Rimstad, E., 2002. Polymorphism in the infectious salmon anemia virus hemagglutinin gene: Importance and possible implications for evolution and ecology of infectious salmon anemia disease. Virology. https://doi.org/10.1006/viro.2002.1658

Mjaaland, S., Rimstad, E., Falk, K., Dannevig, B.H., 1997. Genomic characterization of the virus causing infectious salmon anemia in Atlantic salmon (Salmo salar L.): an orthomyxo-like virus in a teleost. J. Virol. https://doi.org/10.1128/jvi.71.10.7681-7686.1997

Mori, K., Nakai, T., Nagahara, M., Muroga, K., Mekuchi, T., Kanno, T., 1991. A Viral Disease in Hatchery-reared Larvae and Juveniles of Redspotted Grouper. Fish Pathol. 26, 209–210. https://doi.org/10.3147/jsfp.26.209

Morozova, O., Marra, M.A., 2008. Applications of next-generation sequencing technologies in functional genomics. Genomics 92, 255–264. https://doi.org/10.1016/j.ygeno.2008.07.001

Morton, A., Routledge, R., Hrushowy, S., Kibenge, M., Kibenge, F., 2017. The effect of exposure to farmed salmon on piscine orthoreovirus infection and fitness in wild Pacific salmon in British Columbia, Canada. PLoS One 12, e0188793. https://doi.org/10.1371/journal.pone.0188793

Moss, J.A., Burreson, E.M., Cordes, J.F., Dungan, C.F., Brown, G.D., Wang, A., Wu, X., Reece, K.S., 2007. Pathogens in Crassostrea ariakensis and other Asian oyster species: Implications for non-native oyster introduction to Chesapeake Bay. Dis. Aquat. Organ. 77, 207–223. https://doi.org/10.3354/dao01829

Moya, A., Holmes, E.C., González-Candelas, F., 2004. The population genetics and evolutionary epidemiology of RNA viruses. Nat. Rev. Microbiol. 2, 279–288. https://doi.org/10.1038/nrmicro863

Munang'andu, H.M., Fredriksen, B.N., Mutoloki, S., Brudeseth, B., Kuo, T.Y., Marjara, I.S., Dalmo, R.A., Evensen, Ø., 2012. Comparison of vaccine efficacy for different antigen delivery systems for infectious pancreatic necrosis virus vaccines in Atlantic salmon (Salmo salar L.) in a cohabitation challenge model. Vaccine 30, 4007–4016. https://doi.org/10.1016/j.vaccine.2012.04.039

Munang'andu, H.M., Mugimba, K.K., Byarugaba, D.K., Mutoloki, S., Evensen, Ø., 2017. Current advances on virus discovery and diagnostic role of viral metagenomics in aquatic organisms. Front. Microbiol. 8, 1–11. https://doi.org/10.3389/fmicb.2017.00406

Munro, A., Ellis, A., McVicar, A., McLay, H., Needham, E., 1984. An exocrine pancreas disease of farmed Atlantic salmon in Scotland. Helgolander Meeresuntersuchungen.

Nappier, S.P., Graczyk, T.K., Schwab, K.J., 2008. Bioaccumulation, retention, and depuration of enteric viruses by Crassostrea virginica and Crassostrea ariakensis oysters. Appl. Environ. Microbiol. 74, 6825–6831. https://doi.org/10.1128/AEM.01000-08

Naylor, R.L., Goldburg, R.J., Primavera, J.H., Kautsky, N., Beveridge, M.C.M., Clay, J., Folke, C., Lubchenco, J., Mooney, H., Troell, M., 2000. Effect of aquaculture on world fish supplies. Nature. https://doi.org/10.1038/35016500

Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E.E., Bamshad, M., Nickerson, D.A., Shendure, J., 2009. Targeted capture and massively parallel sequencing of 12 human exomes. Nature. https://doi.org/10.1038/nature08250

Nguyen, L.T., Schmidt, H.A., Von Haeseler, A., Minh, B.Q., 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32, 268–274. https://doi.org/10.1093/molbev/msu300

Nikolaitchik, O., Rhodes, T.D., Ott, D., Hu, W.-S., 2006. Effects of Mutations in the Human Immunodeficiency Virus Type 1 gag Gene on RNA Packaging and Recombination. J. Virol. https://doi.org/10.1128/jvi.80.10.4691-4697.2006

Nishizawa, T., Savaş, H., Işidan, H., Üstündağ, C., Iwamoto, H., Yoshimizu, M., 2006. Genotyping and pathogenicity of viral hemorrhagic septicemia virus from free-living turbot (Psetta maxima) in a Turkish coastal area of the Black Sea. Appl. Environ. Microbiol. 72, 2373–2378. https://doi.org/10.1128/AEM.72.4.2373-2378.2006

Nkili-Meyong, A.A., Bigarré, L., Labouba, I., Vallaeys, T., Avarre, J.C., Berthet, N., 2017. Contribution of Next-Generation Sequencing to Aquatic and Fish Virology. Intervirology 59, 285–300. https://doi.org/10.1159/000477808

Noble, R.T., Weisberg, S.B., 2005. A review of technologies for rapid detection of bacteria in recreational waters. J. Water Health. https://doi.org/10.2166/wh.2005.051

Norwegian Veterinary Institute, 2018. Fish Health Report 2018.

Novak, A.M., Hickey, G., Garrison, E., Blum, S., Connelly, A., Dilthey, A., Eizenga, J., Elmohamed, M.A.S., Guthrie, S., Kahles, A., Keenan, S., Kelleher, J., Kural, D., Li, H., Lin, M.F., Miga, K., Ouyang, N., Rakocevic, G., Smuga-Otto, M., Zaranek, A.W., Durbin, R., McVean, G., Haussler, D., Paten, B., 2017. Genome Graphs. bioRxiv. https://doi.org/10.1101/101378

Nurk, S., Meleshko, D., Korobeynikov, A., Pevzner, P.A., 2017. MetaSPAdes: A new versatile metagenomic assembler. Genome Res. 27, 824–834. https://doi.org/10.1101/gr.213959.116

Nylund, A., Devold, M., Plarre, H., Isdal, E., Aarseth, M., 2003. Emergence and maintenance of infectious salmon anaemia virus (ISAV) in Europe: A new hypothesis. Dis. Aquat. Organ. 56, 11–24. https://doi.org/10.3354/dao056011

Nylund, A., Plarre, H., Karlsen, M., Fridell, F., Ottem, K.F., Bratland, A., Sæther, P.A., 2007. Transmission of infectious salmon anaemia virus (ISAV) in farmed populations of Atlantic salmon (Salmo salar). Arch. Virol. 152, 151–179. https://doi.org/10.1007/s00705-006-0825-9

O'Neil, D., Glowatz, H., Schlumpberge, M., 2013. Ribosomal RNA depletion for efficient use of RNA-seq capacity. Curr. Protoc. Mol. Biol. https://doi.org/10.1002/0471142727.mb0419s103

OIE, 2019a. Manual of Diagnostic Tests for Aquatic Animals - Infection with Viral Haemorrhagic Sepeticaemia. OIE Aquat. Anim. Dis. Cards 2.3.10, 1–24.

OIE, 2019b. OIE-Listed diseases, infections and infestations in force in 2019 [WWW Document].

OIE, 2017a. Manual of Diagnostic Tests for Aquatic Animals - Infection with Salmonid Alphavirus. OIE Aquat. Anim. Dis. Cards 2.3.6, 1–14.

OIE, 2017b. Manual of diagnostic tests for aquatic animals - Infection with Infectious Salmon Anaemic Virus. OIE Aquat. Anim. Dis. Cards 2.3, 1–16. https://doi.org/www.oie.int

Oreshkova, S.F., Shchelkunov, I.S., Tikunova, N. V, Shchelkunova, T.I., Puzyrev, A.T., Ilyichev,

A.A., 1999. Detection of spring viremia of carp virus isolates by hybridization with non-radioactive probes and amplification by polymerase chain reaction. Virus Res. 63, 3–10. https://doi.org/10.1016/s0168-1702(99)00052-0

Ott, A., Schnable, J.C., Yeh, C.T., Wu, L., Liu, C., Hu, H.C., Dalgard, C.L., Sarkar, S., Schnable, P.S., 2018. Linked read technology for assembling large complex and polyploid genomes. BMC Genomics. https://doi.org/10.1186/s12864-018-5040-z

Palacios, G., Lovoll, M., Tengs, T., Hornig, M., Hutchison, S., Hui, J., Kongtorp, R.T., Savji, N., Bussetti, A. V., Solovyov, A., Kristoffersen, A.B., Celone, C., Street, C., Trifonov, V., Hirschberg, D.L., Rabadan, R., Egholm, M., Rimstad, E., Lipkin, W.I., 2010. Heart and skeletal muscle inflammation of farmed salmon is associated with infection with a novel reovirus. PLoS One 5, 3–9. https://doi.org/10.1371/journal.pone.0011487

Parla, J.S., Iossifov, I., Grabill, I., Spector, M.S., Kramer, M., McCombie, W.R., 2011. A comparative analysis of exome capture. Genome Biol. https://doi.org/10.1186/gb-2011-12-9-r97

Payne, A., Holmes, N., Rakyan, V., Loose, M., 2019. Bulkvis: A graphical viewer for Oxford nanopore bulk FAST5 files. Bioinformatics. https://doi.org/10.1093/bioinformatics/bty841

Pearson, T., Busch, J.D., Ravel, J., Read, T.D., Rhoton, S.D., U'Ren, J.M., Simonson, T.S., Kachur, S.M., Leadem, R.R., Cardon, M.L., Van Ert, M.N., Huynh, L.Y., Fraser, C.M., Keim, P., 2004. Phylogenetic discovery bias in Bacillus anthracis using single-nucleotide polymorphisms from whole-genome sequencing. Proc. Natl. Acad. Sci. U. S. A. 101, 13536–13541. https://doi.org/10.1073/pnas.0403844101

Pentinsaari, M., Salmela, H., Mutanen, M., Roslin, T., 2016. Molecular evolution of a widely-adopted taxonomic marker (COI) across the animal tree of life. Sci. Rep. 6, 1–12. https://doi.org/10.1038/srep35275

Pereira-Marques, J., Hout, A., Ferreira, R.M., Weber, M., Pinto-Ribeiro, I., Van Doorn, L.J., Knetsch, C.W., Figueiredo, C., 2019. Impact of host DNA and sequencing depth on the taxonomic resolution of whole metagenome sequencing for microbiome analysis. Front. Microbiol. https://doi.org/10.3389/fmicb.2019.01277

Petersen, L.M., Bautista, E.J., Nguyen, H., Hanson, B.M., Chen, L., Lek, S.H., Sodergren, E., Weinstock, G.M., 2017. Community characteristics of the gut microbiomes of competitive cyclists. Microbiome. https://doi.org/10.1186/s40168-017-0320-4

Petterson, E., Guo, T.C., Evensen, Ø., Mikalsen, A.B., 2016. Experimental piscine alphavirus RNA recombination in vivo yields both viable virus and defective viral RNA. Sci. Rep. 6, 36317. https://doi.org/10.1038/srep36317

Petterson, E., Stormoen, M., Evensen, Ø., Mikalsen, A.B., Haugland, Ø., 2013. Natural infection of Atlantic salmon (Salmo salar L.) with salmonid alphavirus 3 generates numerous viral deletion mutants. J. Gen. Virol. 94, 1945–1954. https://doi.org/10.1099/vir.0.052563-0

Pfaff, F., Müller, T., Freuling, C.M., Fehlner-Gardiner, C., Nadin-Davis, S., Robardet, E., Cliquet, F., Vuta, V., Hostnik, P., Mettenleiter, T.C., Beer, M., Höper, D., 2019. In-depth genome analyses of viruses from vaccine-derived rabies cases and corresponding live-attenuated oral rabies vaccines. Vaccine 37, 4758–4765. https://doi.org/10.1016/j.vaccine.2018.01.083

Pfeiffer, F., Gröber, C., Blank, M., Händler, K., Beyer, M., Schultze, J.L., Mayer, G., 2018. Systematic evaluation of error rates and causes in short samples in next-generation sequencing. Sci. Rep. 8, 1–14. https://doi.org/10.1038/s41598-018-29325-6

Plarre, H., Nylund, A., Karlsen, M., Brevik, Ø., Sæther, P.A., Vike, S., 2012. Evolution of infectious salmon anaemia virus (ISA virus). Arch. Virol. 157, 2309–2326. https://doi.org/10.1007/s00705-012-1438-0

Poppe, T., Rimstad, E., Hyllseth, B., 1989. Pancreas disease in Atlantic salmon (Salmo salar) postsmolts infected with infectious pancreatic necrosis virus (IPNV). Bull. Eur. Ass. Fish Pathol 9, 83–85.

Posada-Cespedes, S., Seifert, D., Beerenwinkel, N., 2017. Recent advances in inferring viral diversity

from high-throughput sequencing data. Virus Res. 239, 17–32. https://doi.org/10.1016/j.virusres.2016.09.016

Posada, D., 2003. Using MODELTEST and PAUP* to Select a Model of Nucleotide Substitution. Curr. Protoc. Bioinforma. 00, 6.5.1-6.5.14. https://doi.org/10.1002/0471250953.bi0605s00

Poulos, B.T., Tang, K.F.J., Pantoja, C.R., Bonami, J.R., Lightner, D. V., 2006. Purification and characterization of infectious myonecrosis virus of penaeid shrimp. J. Gen. Virol. https://doi.org/10.1099/vir.0.81127-0

Power, R.A., Davaniah, S., Derache, A., Wilkinson, E., Tanser, F., Gupta, R.K., Pillay, D., De Oliveira, T., 2016a. Genome-wide association study of HIV whole genome sequences validated using drug resistance. PLoS One 11, 1–14. https://doi.org/10.1371/journal.pone.0163746

Power, R.A., Parkhill, J., De Oliveira, T., 2016b. Microbial genome-wide association studies: lessons from human GWAS. Nat. Rev. Genet. https://doi.org/10.1038/nrg.2016.132

Pritchard, C.C., Salipante, S.J., Koehler, K., Smith, C., Scroggins, S., Wood, B., Wu, D., Lee, M.K., Dintzis, S., Adey, A., Liu, Y., Eaton, K.D., Martins, R., Stricker, K., Margolin, K.A., Hoffman, N., Churpek, J.E., Tait, J.F., King, M.C., Walsh, T., 2014. Validation and implementation of targeted capture and sequencing for the detection of actionable mutation, copy number variation, and gene rearrangement in clinical cancer specimens. J. Mol. Diagnostics. https://doi.org/10.1016/j.jmoldx.2013.08.004

Pulliam, H.R., 1988. Sources, sinks and population regulation. Am. Nat. https://doi.org/10.1086/284880

Pybus, O.G., Rambaut, A., 2009. Evolutionary analysis of the dynamics of viral infectious disease. Nat. Rev. Genet. 10, 540–550. https://doi.org/10.1038/nrg2583

Quick, J., Grubaugh, N.D.D., Pullan, S.T.T., Claro, I.M.M., Smith, A.D.D., Gangavarapu, K., Oliveira, G., Robles-Sikisaka, R., Rogers, T.F.F., Beutler, N.A.A., Burton, D.R.R., Lewis-Ximenez, L.L.L., Goes de Jesus, J., Giovanetti, M., Hill, S.C., Black, A., Bedford, T., Carroll, M.W.W., Nunes, M., Alcantara, L.C.C., Sabino, E.C.C., Baylis, S.A.A., Faria, N.R., Loose, M., Simpson, J.T.T., Pybus, O.G.G., Andersen, K.G.G., Loman, N.J.J., De Jesus, J.G., Giovanetti, M., Hill, S.C., Black, A., Bedford, T., Carroll, M.W.W., Nunes, M., Alcantara, L.C.C., Sabino, E.C.C., Baylis, S.A.A., Faria, N.R., Loose, M., Simpson, J.T.T., Pybus, O.G.G., Andersen, K.G.G., Loman, N.J.J., 2017. Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. Nat. Protoc. 12, 1261–1266. https://doi.org/10.1038/nprot.2017.066

Quick, J., Loman, N.J., Duraffour, S., Simpson, J.T., Severi, E., Cowley, L., Bore, J.A., Koundouno, R., Dudas, G., Mikhail, A., Ouédraogo, N., Afrough, B., Bah, A., Baum, J.H.J., Becker-Ziaja, B., Boettcher, J.P., Cabeza-Cabrerizo, M., Camino-Sánchez, Á., Carter, L.L., Doerrbecker, J., Enkirch, T., Garciá-Dorival, I., Hetzelt, N., Hinzmann, J., Holm, T., Kafetzopoulou, L.E., Koropogui, M., Kosgey, A., Kuisma, E., Logue, C.H., Mazzarelli, A., Meisel, S., Mertens, M., Michel, J., Ngabo, D., Nitzsche, K., Pallasch, E., Patrono, L.V., Portmann, J., Repits, J.G., Rickett, N.Y., Sachse, A., Singethan, K., Vitoriano, I., Yemanaberhan, R.L., Zekeng, E.G., Racine, T., Bello, A., Sall, A.A., Faye, Ousmane, Faye, Oumar, Magassouba, N., Williams, C. V., Amburgey, V., Winona, L., Davis, E., Gerlach, J., Washington, F., Monteil, V., Jourdain, M., Bererd, M., Camara, Alimou, Somlare, H., Camara, Abdoulaye, Gerard, M., Bado, G., Baillet, B., Delaune, D., Nebie, K.Y., Diarra, A., Savane, Y., Pallawo, R.B., Gutierrez, G.J., Milhano, N., Roger, I., Williams, C.J., Yattara, F., Lewandowski, K., Taylor, J., Rachwal, P., Turner, D.J., Pollakis, G., Hiscox, J.A., Matthews, D.A., O'Shea, M.K., Johnston, A.M.D., Wilson, D., Hutley, E., Smit, E., Di Caro, A., Wolfel, R., Stoecker, K., Fleischmann, E., Gabriel, M., Weller, S.A., Koivogui, L., Diallo, B., Keita, S., Rambaut, A., Formenty, P., Gunther, S., Carroll, M.W., 2016. Real-time, portable genome sequencing for Ebola surveillance. Nature 530, 228–232. https://doi.org/10.1038/nature16996

Quince, C., Walker, A.W., Simpson, J.T., Loman, N.J., Segata, N., 2017. Shotgun metagenomics, from sampling to analysis. Nat. Biotechnol. https://doi.org/10.1038/nbt.3935

Qureshi, A., Tantray, V.G., Kirmani, A.R., Ahangar, A.G., 2018. Next-Generation Sequencing (NGS) Based Detection of Viral Pathogens. Expert Biol. 3.

Ramankutty, N., Mehrabi, Z., Waha, K., Jarvis, L., Kremen, C., Herrero, M., Rieseberg, L.H., 2018. Trends in Global Agricultural Land Use: Implications for Environmental Health and Food Security. Annu. Rev. Plant Biol. 69, 789–815. https://doi.org/10.1146/annurev-arplant-042817-040256

Rambaut, A., Drummond, A.J., Xie, D., Baele, G., Suchard, M.A., 2018. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. Syst. Biol. 67, 901–904. https://doi.org/10.1093/sysbio/syy032

Rambaut, A., Lam, T.T., Max Carvalho, L., Pybus, O.G., 2016. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol. 2, vew007. https://doi.org/10.1093/ve/vew007

Rambaut, A., Pybus, O.G., Nelson, M.I., Viboud, C., Taubenberger, J.K., Holmes, E.C., 2008. The genomic and epidemiological dynamics of human influenza A virus. Nature 453, 615–619. https://doi.org/10.1038/nature06945

Rambaut, A., Holmes, E.C., Hill, V., O'Toole, A., McCrone, J.T., Ruis, C., du Plessis, L., Pybus, O.G., 2020. A dynamic nomenclature proposal for SARS-CoV-2 to assist genomic epidemiology. BioRxiv. https://doi.org/10.1101/2020.04.17.046086

Rang, F.J., Kloosterman, W.P., de Ridder, J., 2018. From squiggle to basepair: Computational approaches for improving nanopore sequencing read accuracy. Genome Biol. https://doi.org/10.1186/s13059-018-1462-9

Redin, D., Frick, T., Aghelpasand, H., Käller, M., Borgström, E., Olsen, R.A., Ahmadian, A., 2019. High throughput barcoding method for genome-scale phasing. Sci. Rep. 9, 1–8. https://doi.org/10.1038/s41598-019-54446-x

Renault, Arzul, 2001. Herpes-like virus infections in hatchery-reared bivalve larvae in Europe: specific viral DNA detection by PCR. J. Fish Dis. 24, 161–167. https://doi.org/10.1046/j.1365-2761.2001.00282.x

Rhoads, A., Au, K.F., 2015. PacBio Sequencing and Its Applications. Genomics, Proteomics Bioinforma. 13, 278–289. https://doi.org/10.1016/j.gpb.2015.08.002

Rimstad, E., Dale, O.B., Dannevig, B.H., Falk, K., 2011. Infectious salmon anaemia, in: Fish Diseases and Disorders. https://doi.org/10.1079/9781845935542.0143

Rodgers, C., Peeler, E., Rodgers, C.J., Mohan, C. V, Peeler, E.J., 2011. The spread of pathogens through trade in aquatic animals and their products. Rev. sci. tech. Off. int. Epiz 30, 241–256. https://doi.org/10.20506/rst.30.1.2034

Rose, J.B., Mullinax, R.L., Singh, S.N., Yates, M. V., Gerba, C.P., 1987. Occurrence of rotaviruses and enteroviruses in recreational waters of Oak Creek, Arizona. Water Res. 21, 1375–1381. https://doi.org/10.1016/0043-1354(87)90012-1

Roux, S., Emerson, J.B., Eloe-Fadrosh, E.A., Sullivan, M.B., 2017. Benchmarking viromics: An in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. PeerJ. https://doi.org/10.7717/peerj.3817

Roux, S., Enault, F., Hurwitz, B.L., Sullivan, M.B., 2015. VirSorter: Mining viral signal from microbial genomic data. PeerJ 2015, e985. https://doi.org/10.7717/peerj.985

Roux, S., Tournayre, J., Mahul, A., Debroas, D., Enault, F., 2014. Metavir 2: New tools for viral metagenome comparison and assembled virome analysis. BMC Bioinformatics 15, 76. https://doi.org/10.1186/1471-2105-15-76

Ruane, N.M., Swords, D., Morrissey, T., Geary, M., Hickey, C., Collins, E.M., Geoghegan, F., Swords, F., 2018. Isolation of salmonid alphavirus subtype 6 from wild-caught ballan wrasse, Labrus bergylta (Ascanius). J. Fish Dis. https://doi.org/10.1111/jfd.12870

Rucker, R.R., Whipple, W.J., Parvin, J.R., Evans, C.A., 1953. A contagious disease of salmon, possibly of virus origin. Fish. Bull.

Russell, A.B., Trapnell, C., Bloom, J.D., 2018. Extreme heterogeneity of influenza virus infection in single cells. Elife 7. https://doi.org/10.7554/eLife.32303

Russell, C.A., Jones, T.C., Barr, I.G., Cox, N.J., Garten, R.J., Gregory, V., Gust, I.D., Hampson, A.W., Hay, A.J., Hurt, A.C., De Jong, J.C., Kelso, A., Klimov, A.I., Kageyama, T., Komadina, N., Lapedes, A.S., Lin, Y.P., Mosterin, A., Obuchi, M., Odagiri, T., Osterhaus, A.D.M.E., Rimmelzwaan, G.F., Shaw, M.W., Skepner, E., Stohr, K., Tashiro, M., Fouchier, R.A.M., Smith, D.J., 2008. The global circulation of seasonal influenza A (H3N2) viruses. Science (80-.). 320, 340–346. https://doi.org/10.1126/science.1154137

Sakr, R.A., Schizas, M., Carniello, J.V.S., Ng, C.K.Y., Piscuoglio, S., Giri, D., Andrade, V.P., de Brot, M., Lim, R.S., Towers, R., Weigelt, B., Reis-Filho, J.S., King, T.A., 2016. Targeted capture massively parallel sequencing analysis of LCIS and invasive lobular cancer: Repertoire of somatic genetic alterations and clonal relationships. Mol. Oncol. https://doi.org/10.1016/j.molonc.2015.11.001

Sanjuán, R., Nebot, M.R.R., Chirico, N., Mansky, L.M.M., Belshaw, R., Sanjuan, R., Nebot, M.R.R., Chirico, N., Mansky, L.M.M., Belshaw, R., Sanjuán, R., Nebot, M.R.R., Chirico, N., Mansky, L.M.M., Belshaw, R., 2010. Viral Mutation Rates. J. Virol. 84, 9733–9748. https://doi.org/10.1128/jvi.00694-10

Schmidt, K., Mwaigwisya, S., Crossman, L.C., Doumith, M., Munroe, D., Pires, C., M. Khan, A., Woodford, N., Saunders, N.J., Wain, J., O'Grady, J., Livermore, D.M., 2017. Identification of bacterial pathogens and antimicrobial resistance directly from clinical urines by nanopore-based metagenomic sequencing. J. Antimicrob. Chemother. 72, 104–114. https://doi.org/10.1093/jac/dkw397

Schönherz, A.A., Lorenzen, N., Guldbrandtsen, B., Buitenhuis, B., Einer-Jensen, K., 2016. Ultra-deep sequencing of VHSV isolates contributes to understanding the role of viral quasispecies. Vet. Res. 47, 1–12. https://doi.org/10.1186/s13567-015-0298-5

Sedlazeck, F.J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., Von Haeseler, A., Schatz, M.C., 2018. Accurate detection of complex structural variations using single-molecule sequencing. Nat. Methods 15, 461–468. https://doi.org/10.1038/s41592-018-0001-7

Segarra, A., Pépin, J.F., Arzul, I., Morga, B., Faury, N., Renault, T., 2010. Detection and description of a particular Ostreid herpesvirus 1 genotype associated with massive mortality outbreaks of Pacific oysters, Crassostrea gigas, in France in 2008. Virus Res. 153, 92–99. https://doi.org/10.1016/j.virusres.2010.07.011

Sharp, P.M., Hahn, B.H., 2010. The evolution of HIV-1 and the origin of AIDS. Philos. Trans. R. Soc. B Biol. Sci. 365, 2487–2494. https://doi.org/10.1098/rstb.2010.0031

Shi, M., Lin, X.-D., Chen, X., Tian, J.-H., Chen, L.-J., Li, K., Wang, W., Eden, J.-S., Shen, J.-J., Liu, L., Holmes, E.C., Zhang, Y.-Z., 2018. The evolutionary history of vertebrate RNA viruses. Nature 556, 197–202. https://doi.org/10.1038/s41586-018-0012-7

Shi, M., Lin, X.-D., Vasilakis, N., Tian, J.-H., Li, C.-X., Chen, L.-J., Eastwood, G., Diao, X.-N., Chen, M.-H., Chen, X., Qin, X.-C., Widen, S.G., Wood, T.G., Tesh, R.B., Xu, J., Holmes, E.C., Zhang, Y.-Z., 2016a. Divergent Viruses Discovered in Arthropods and Vertebrates Revise the Evolutionary History of the Flaviviridae and Related Viruses. J. Virol. 90, 659–669. https://doi.org/10.1128/jvi.02036-15

Shi, M., Lin, X.D., Tian, J.H., Chen, L.J., Chen, X., Li, C.X., Qin, X.C., Li, J., Cao, J.P., Eden, J.S., Buchmann, J., Wang, W., Xu, J., Holmes, E.C., Zhang, Y.Z., 2016b. Redefining the invertebrate RNA virosphere. Nature 540, 539–543. https://doi.org/10.1038/nature20167

Shimahara, Y., Kurita, J., Kiryu, I., Nishioka, T., Yuasa, K., Kawana, M., Kamaishi, T., Oseko, N., 2012. Surveillance of Type 1 Ostreid Herpesvirus (OsHV-1) Variants in Japan. Fish Pathol. 47, 129–136. https://doi.org/10.3147/jsfp.47.129

Shyamala, V., Arcangel, P., Cottrell, J., Coit, D., Medina-Selby, A., McCoin, C., Madriaga, D., Chien, D., Phelps, B., 2004. Assessment of the target-capture PCR hepatitis B virus (HBV) DNA quantitative assay and comparison with commercial HBV DNA quantitative assays. J.

Clin. Microbiol. https://doi.org/10.1128/JCM.42.11.5199-5204.2004

Simon-Loriere, E., Holmes, E.C., 2011. Why do RNA viruses recombine? Nat. Rev. Microbiol. 9, 617–626. https://doi.org/10.1038/nrmicro2614

Skall, H.F., Olesen, N.J., Mellergaard, S., 2005. Viral haemorrhagic septicaemia virus in marine fish and its implications for fish farming - A review. J. Fish Dis. 28, 509–529. https://doi.org/10.1111/j.1365-2761.2005.00654.x

Skall, H.F., Slierendrecht, W.J., King, J.A., Olesen, N.J., 2004. Experimental infection of rainbow trout Oncorhynchus mykiss with viral haemorrhagic septicaemia virus isolates from European marine and farmed fishes. Dis. Aquat. Organ. 58, 99–110. https://doi.org/10.3354/dao058099

Smith, B.T., Harvey, M.G., Faircloth, B.C., Glenn, T.C., Brumfield, R.T., 2014. Target capture and massively parallel sequencing of ultraconserved elements for comparative studies at shallow evolutionary time scales. Syst. Biol. https://doi.org/10.1093/sysbio/syt061

Snow, M., 2011. The contribution of molecular epidemiology to the understanding and control of viral diseases of salmonid aquaculture. Vet. Res. 42, 1–12. https://doi.org/10.1186/1297-9716-42-56

Snow, M., Black, J., Matejusova, I., McIntosh, R., Baretto, E., Wallace, I.S., Bruno, D.W., 2010. Detection of salmonid alphavirus RNA in wild marine fish: Implications for the origins of salmon pancreas disease in aquaculture. Dis. Aquat. Organ. 91, 177–188. https://doi.org/10.3354/dao02265

Snow, M., McKay, P., McBeath, A.J.A., Black, J., Doig, F., Kerr, R., Cunningham, C.O., Nylund, A., Devold, M., 2006. Development, application and validation of a Taqman® real-time RT-PCR assay for the detection of infectious salmon anaemia virus (ISAV) in Atlantic salmon (Salmo salar), in: Developments in Biologicals.

Sobsey, M.D., Davis, A.L., Rullman, V.A., 1987. PERSISTENCE OF HEPATITIS A VIRUS AND OTHER VIRUSES IN DEPURATED EASTERN OYSTERS., in: Oceans Conference Record (IEEE). IEEE, pp. 1740–1745. https://doi.org/10.1109/oceans.1987.1160616

Sobsey, M.D., Shields, P.A., Hauchman, F.H., Hazard, R.L., Caton, L.W., 1986. SURVIVAL AND TRANSPORT OF HEPATITIS A VIRUS IN SOILS, GROUNDWATER AND WASTEWATER. Water Sci. Technol. 18, 97–106. https://doi.org/10.2166/wst.1986.0116

Son, N.T., Fleet, G.H., 1980. Behavior of pathogenic bacteria in the oyster, Crassostrea commercialis, during depuration, re-laying, and storage. Appl. Environ. Microbiol. 40, 994–1002. https://doi.org/10.1128/aem.40.6.994-1002.1980

Sone, J., Mitsuhashi, S., Fujita, A., Mizuguchi, T., Hamanaka, K., Mori, K., Koike, H., Hashiguchi, A., Takashima, H., Sugiyama, H., Kohno, Y., Takiyama, Y., Maeda, K., Doi, H., Koyano, S., Takeuchi, H., Kawamoto, M., Kohara, N., Ando, T., Ieda, T., Kita, Y., Kokubun, N., Tsuboi, Y., Katoh, K., Kino, Y., Katsuno, M., Iwasaki, Y., Yoshida, M., Tanaka, F., Suzuki, I.K., Frith, M.C., Matsumoto, N., Sobue, G., 2019. Long-read sequencing identifies GGC repeat expansions in NOTCH2NLC associated with neuronal intranuclear inclusion disease. Nat. Genet. https://doi.org/10.1038/s41588-019-0459-y

Sritunyalucksana, K., Apisawetakan, S., Boon-nat, A., Withyachumnarnkul, B., Flegel, T.W., 2006. A new RNA virus found in black tiger shrimp Penaeus monodon from Thailand. Virus Res. https://doi.org/10.1016/j.virusres.2005.11.005

Stadhouders, R., Pas, S.D., Anber, J., Voermans, J., Mes, T.H.M., Schutten, M., 2010. The effect of primer-template mismatches on the detection and quantification of nucleic acids using the 5′ nuclease assay. J. Mol. Diagnostics 12, 109–117. https://doi.org/10.2353/jmoldx.2010.090035

Stagg, 2003. The eradication of an outbreak of clinical infectious salmon anaemia from Scotland. In: International Response to Infectious Salmon Anaemia: Prevention, Control and Eradication. US Dep. Agric. 111–124.

Steinhauer, D.A., Holland, J.J., 1987. Rapid Evolution Of RNA Viruses. Annu. Rev. Microbiol. 41, 409–433. https://doi.org/10.1146/annurev.micro.41.1.409

Stenglein, M.D., Sanders, C., Kistler, A.L., Graham Ruby, J., Franco, J.Y., Reavill, D.R., Dunker, F., DeRisia, J.L., 2012. Identification, characterization, and in vitro culture of highly divergent arenaviruses from boa constrictors and annulated tree boas: Candidate etiological agents for snake inclusion body disease. MBio 3. https://doi.org/10.1128/mBio.00180-12

Su, S., Wong, G., Shi, W., Liu, J., Lai, A.C.K., Zhou, J., Liu, W., Bi, Y., Gao, G.F., 2016. Epidemiology, Genetic Recombination, and Pathogenesis of Coronaviruses. Trends Microbiol. 24, 490–502. https://doi.org/10.1016/j.tim.2016.03.003

Su, Y.C.F., Bahl, J., Joseph, U., Butt, K.M., Peck, H.A., Koay, E.S.C., Oon, L.L.E., Barr, I.G., Vijaykrishna, D., Smith, G.J.D., 2015. Phylodynamics of H1N1/2009 influenza reveals the transition from host adaptation to immune-driven selection. Nat. Commun. 6, 7952. https://doi.org/10.1038/ncomms8952

Sugauchi, F., Orito, E., Ichida, T., Kato, H., Sakugawa, H., Kakumu, S., Ishida, T., Chutaputti, A., Lai, C.L., Gish, R.G., Ueda, R., Miyakawa, Y., Mizokami, M., 2003. Epidemiologic and virologic characteristics of hepatitis B virus genotype B having the recombination with genotype C. Gastroenterology 124, 925–932. https://doi.org/10.1053/gast.2003.50140

Sumpter, N., Butler, M., Poulter, R., 2018. Single-Phase PacBio De Novo Assembly of the Genome of the Chytrid Fungus Batrachochytrium dendrobatidis, a Pathogen of Amphibia . Microbiol. Resour. Announc. https://doi.org/10.1128/mra.01348-18

Suttle, C.A., 2007. Marine viruses - Major players in the global ecosystem. Nat. Rev. Microbiol. 5, 801–812. https://doi.org/10.1038/nrmicro1750

Suttle, C.A., 2005. Viruses in the sea. Nature. https://doi.org/10.1038/nature04160

Sutton, T.D.S., Clooney, A.G., Ryan, F.J., Ross, R.P., Hill, C., 2019. Choice of assembly software has a critical impact on virome characterisation. Microbiome 7, 1–15. https://doi.org/10.1186/s40168-019-0626-5

Tacon, A.G.J., Forster, I.P., 2003. Aquafeeds and the environment: Policy implications, in: Aquaculture. Elsevier, pp. 181–189. https://doi.org/10.1016/S0044-8486(03)00476-9

Tan, Y., Lam, T.T.-Y., Heberlein-Larson, L.A., Smole, S.C., Auguste, A.J., Hennigan, S., Halpin, R.A., Fedorova, N., Puri, V., Stockwell, T.B., Shilts, M.H., Andreadis, T., Armstrong, P.M., Tesh, R.B., Weaver, S.C., Unnasch, T.R., Ciota, A.T., Kramer, L.D., Das, S.R., 2018. Large-Scale Complete-Genome Sequencing and Phylodynamic Analysis of Eastern Equine Encephalitis Virus Reveals Source-Sink Transmission Dynamics in the United States. J. Virol. https://doi.org/10.1128/jvi.00074-18

Tang, K.F.J., Lightner, D. V., 1999. A yellow head virus gene probe: Nucleotide sequence and application for in situ hybridization. Dis. Aquat. Organ. https://doi.org/10.3354/dao035165

Tang, K.F.J., Pantoja, C.R., Redman, R.M., Lightner, D. V., 2007. Development of in situ hybridization and RT-PCR assay for the detection of a nodavirus (PvNV) that causes muscle necrosis in Penaeus vannamei. Dis. Aquat. Organ. https://doi.org/10.3354/dao075183

Tavaré, S., 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. Am. Math. Soc. Lect. Math. Life Sci. https://doi.org/citeulike-article-id:4801403

Teng, Y., Liu, H., Lv, J.Q., Fan, W.H., Zhang, Q.Y., Qin, Q.W., 2007. Characterization of complete genome sequence of the spring viremia of carp virus isolated from common carp (Cyprinus carpio) in China. Arch. Virol. 152, 1457–1465. https://doi.org/10.1007/s00705-007-0971-8

The World Bank, 2014. Reducing Disease Risk In Aquaculture. World Bank. Agric. Environ. Serv. 119.

Thézé, J., Li, T., du Plessis, L., Bouquet, J., Kraemer, M.U.G., Somasekar, S., Yu, G., de Cesare, M., Balmaseda, A., Kuan, G., Harris, E., Wu, C. hsi, Ansari, M.A., Bowden, R., Faria, N.R., Yagi, S., Messenger, S., Brooks, T., Stone, M., Bloch, E.M., Busch, M., Muñoz-Medina, J.E., González-Bonilla, C.R., Wolinsky, S., López, S., Arias, C.F., Bonsall, D., Chiu, C.Y., Pybus, O.G., 2018. Genomic Epidemiology Reconstructs the Introduction and Spread of Zika Virus in Central America and Mexico. Cell Host Microbe 23, 855-864.e7.

https://doi.org/10.1016/j.chom.2018.04.017

Thorud, K., H.O.D., 1988. Infectious anaemia in Atlantic salmon (Salmo salar). Bull. Eur. Assoc. Fish Pathol.Eur. Assoc. Fish Pathol.

Thorvaldsdóttir, H., Robinson, J.T., Mesirov, J.P., 2013. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. Brief. Bioinform. 14, 178–192. https://doi.org/10.1093/bib/bbs017

Tong, S., Revill, P., 2016. Overview of hepatitis B viral replication and genetic variability. J. Hepatol. https://doi.org/10.1016/j.jhep.2016.01.027

Toro-Ascuy, D., Tambley, C., Beltran, C., Mascayano, C., Sandoval, N., Olivares, E., Medina, R.A., Spencer, E., Martína, M.C.S., 2015. Development of a reverse genetic system for infectious salmon anemia virus: Rescue of recombinant fluorescent virus by using salmon internal transcribed spacer region 1 as a novel promoter. Appl. Environ. Microbiol. 81, 1210–1224. https://doi.org/10.1128/AEM.03153-14

Trifinopoulos, J., Nguyen, L.-T.T.T., von Haeseler, A., Minh, B.Q.Q., von Haeseler, A., Minh, B.Q.Q., von Haeseler, A., Minh, B.Q.Q., 2016. W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. Nucleic Acids Res. 44, W232–W235. https://doi.org/10.1093/nar/gkw256

Tscherne, D.M., García-Sastre, A., 2011. Virulence determinants of pandemic influenza viruses. J. Clin. Invest. 121, 6–13. https://doi.org/10.1172/JCI44947

Ueki, Y., Shoji, M., Suto, A., Tanabe, T., Okimura, Y., Kikuchi, Y., Saito, N., Sano, D., Omura, T., 2007. Persistence of caliciviruses in artificially contaminated oysters during depuration. Appl. Environ. Microbiol. 73, 5698–5701. https://doi.org/10.1128/AEM.00290-07

Uzcategui, N.Y., Camacho, D., Comach, G., Cuello de Uzcategui, R., Holmes, E.C., Gould, E.A., 2001. Molecular epidemiology of dengue type 2 virus in Venezuela: Evidence for in situ virus evolution and recombination. J. Gen. Virol. 82, 2945–2953. https://doi.org/10.1099/0022-1317-82-12-2945

Van Ballegooijen, W.M., Van Houdt, R., Bruisten, S.M., Boot, H.J., Coutinho, R.A., Wallinga, J., 2009. Molecular sequence data of hepatitis B virus and genetic diversity after vaccination. Am. J. Epidemiol. https://doi.org/10.1093/aje/kwp375

Van Hulten, M.C.W., Witteveldt, J., Peters, S., Kloosterboer, N., Tarchini, R., Fiers, M., Sandbrink, H., Lankhorst, R.K., Vlak, J.M., 2001. The white spot syndrome virus DNA genome sequence. Virology. https://doi.org/10.1006/viro.2001.1002

Vibin, J., Chamings, A., Collier, F., Klaassen, M., Nelson, T.M., Alexandersen, S., 2018. Metagenomics detection and characterisation of viruses in faecal samples from Australian wild birds. Sci. Rep. 8, 1–23. https://doi.org/10.1038/s41598-018-26851-1

Vijaykrishna, D., Holmes, E.C., Joseph, U., Fourment, M., Su, Y.C.F., Halpin, R., Lee, R.T.C., Deng, Y.M., Gunalan, V., Lin, X., Stockwell, T.B., Fedorova, N.B., Zhou, B., Spirason, N., Küuhnert, D., Bošková, V., Stadler, T., Costa, A.M., Dwyer, D.E., Huang, Q.S., Jennings, L.C., Rawlinson, W., Sullivan, S.G., Hurt, A.C., Maurer-Stroh, S., Wentworth, D.E., Smith, G.J.D., Barr, I.G., 2015a. The contrasting phylodynamics of human inuenza B viruses. Elife 2015, 1–23. https://doi.org/10.7554/eLife.05055

Vijaykrishna, D., Mukerji, R., Smith, G.J.D., 2015b. RNA Virus Reassortment: An Evolutionary Mechanism for Host Jumps and Immune Evasion. PLoS Pathog. 11, e1004902. https://doi.org/10.1371/journal.ppat.1004902

Vike, S., Nylund, S., Nylund, A., 2009. ISA virus in Chile: Evidence of vertical transmission. Arch. Virol. 154, 1–8. https://doi.org/10.1007/s00705-008-0251-2

Viljugrein, H., Staalstrøm, A., Molvær, J., Urke, H.A., Jansen, P.A., 2010. Integration of hydrodynamics into a statistical model on the spread of pancreas disease (PD) in salmon farming. Dis. Aquat. Organ. 88, 35–44. https://doi.org/10.3354/dao02151

Villoing, S., Béarzotti, M., Chilmonczyk, S., Castric, J., Brémont, M., 2000. Rainbow trout sleeping disease virus is an atypical alphavirus. J. Virol. 74, 173–83. https://doi.org/10.1128/JVI.74.1.173-183.2000

Vollmers, J., Wiegand, S., Kaster, A.K., 2017. Comparing and evaluating metagenome assembly tools from a microbiologist's perspective - Not only size matters! PLoS One. https://doi.org/10.1371/journal.pone.0169662

Volz, E.M., Koelle, K., Bedford, T., 2013. Viral Phylodynamics. PLoS Comput. Biol. 9. https://doi.org/10.1371/journal.pcbi.1002947

Walker, P.J., Winton, J.R., 2010. Emerging viral diseases of fish and shrimp. Vet. Res. 41. https://doi.org/10.1051/vetres/2010022

Walling, D.M., Shebib, N., Weaver, S.C., Nichols, C.M., Flaitz, C.M., Webster-Cyriaque, J., 1999. The molecular epidemiology and evolution of Epstein-Barr virus: sequence variation and genetic recombination in the latent membrane protein-1 gene. J. Infect. Dis. 179, 763–74. https://doi.org/10.1086/314672

Wang, C., Mitsuya, Y., Gharizadeh, B., Ronaghi, M., Shafer, R.W., 2007. Characterization of mutation spectra with ultra-deep pyrosequencing: Application to HIV-1 drug resistance. Genome Res. 17, 1195–1201. https://doi.org/10.1101/gr.6468307

Wang, D., Coscoy, L., Zylberberg, M., Avila, P.C., Boushey, H.A., Ganem, D., DeRisi, J.L., 2002. Microarray-based detection and genotyping of viral pathogens. Proc. Natl. Acad. Sci. U. S. A. 99, 15687–15692. https://doi.org/10.1073/pnas.242579699

Wang, Jun, Zhang, Guofan, Fang, X., Guo, X., Li, L., Luo, R., Xu, F., Yang, P., Zhang, L., Wang, X., Qi, H., Xiong, Z., Que, H., Xie, Y., Holland, P.W.H., Paps, J., Zhu, Y., Wu, F., Chen, Y., Wang, Jiafeng, Peng, C., Meng, J., Yang, L., Liu, J., Wen, B., Zhang, N., Huang, Z., Zhu, Q., Feng, Y., Mount, A., Hedgecock, D., Xu, Z., Liu, Y., Domazet-Lošo, T., Du, Y., Sun, X., Zhang, Shoudu, Liu, B., Cheng, P., Jiang, X., Li, J., Fan, D., Wang, W., Fu, W., Wang, T., Wang, B., Zhang, J., Peng, Z., Li, Yingxiang, Li, Na, Wang, Jinpeng, Chen, M., He, Y., Tan, F., Song, X., Zheng, Q., Huang, R., Yang, Hailong, Du, X., Chen, L., Yang, M., Gaffney, P.M., Wang, S., Luo, L., She, Z., Ming, Y., Huang, W., Zhang, Shu, Huang, B., Zhang, Y., Qu, T., Ni, P., Miao, G., Wang, Junyi, Wang, Q., Steinberg, C.E.W., Wang, H., Li, Ning, Qian, L., Zhang, Guojie, Li, Yingrui, Yang, Huanming, Liu, X., Yin, Y., Wang, Jian, 2012. The oyster genome reveals stress adaptation and complexity of shell formation. Nature 490, 49–54. https://doi.org/10.1038/nature11413

Warg, J. V., Dikkeboom, A.L., Goodwin, A.E., Snekvik, K., Whitney, J., 2007. Comparison of multiple genes of spring viremia of carp viruses isolated in the United States. Virus Genes 35, 87–95. https://doi.org/10.1007/s11262-006-0042-3

Warwick-Dugdale, J., Solonenko, N., Moore, K., Chittick, L., Gregory, A.C., Allen, M.J., Sullivan, M.B., Temperton, B., 2019. Long-read viral metagenomics captures abundant and microdiverse viral populations and their niche-defining genomic islands. PeerJ 2019, e6800. https://doi.org/10.7717/peerj.6800

Webby, R.J., Webster, R.G., 2001. Emergence of influenza A viruses. Philos. Trans. R. Soc. B Biol. Sci. 356, 1817–1828. https://doi.org/10.1098/rstb.2001.0997

Weisburg, W.G., Barns, S.M., Pelletier, D.A., Lane, D.J., 1991. 16S ribosomal DNA amplification for phylogenetic study. J. Bacteriol. 173, 697–703. https://doi.org/10.1128/jb.173.2.697-703.1991

Wen, K., Ortmann, A.C., Suttle, C.A., 2004. Accurate estimation of viral abundance by epifluorescence microscopy. Appl. Environ. Microbiol. 70, 3862–3867. https://doi.org/10.1128/AEM.70.7.3862-3867.2004

Wenger, A.M., Peluso, P., Rowell, W.J., Chang, P.C., Hall, R.J., Concepcion, G.T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N.D., Töpfer, A., Alonge, M., Mahmoud, M., Qian, Y., Chin, C.S., Phillippy, A.M., Schatz, M.C., Myers, G., DePristo, M.A., Ruan, J., Marschall, T., Sedlazeck, F.J., Zook, J.M., Li, H., Koren, S., Carroll, A., Rank, D.R., Hunkapiller, M.W., 2019. Accurate circular consensus long-read sequencing improves variant detection and

assembly of a human genome. Nat. Biotechnol. https://doi.org/10.1038/s41587-019-0217-9

Weston, J., Villoing, S., Bremont, M., Castric, J., Pfeffer, M., Jewhurst, V., McLoughlin, M., Rodseth, O., Christie, K.E., Koumans, J., Todd, D., 2002. Comparison of Two Aquatic Alphaviruses, Salmon Pancreas Disease Virus and Sleeping Disease Virus, by Using Genome Sequence Analysis, Monoclonal Reactivity, and Cross-Infection. J. Virol. https://doi.org/10.1128/jvi.76.12.6155-6163.2002

Weston, J.H., Welsh, M.D., McLoughlin, M.F., Todd, D., 1999. Salmon pancreas disease virus, an alphavirus infecting farmed Atlantic salmon, Salmo salar L. Virology 256, 188–195. https://doi.org/10.1006/viro.1999.9654

Wickham, H., 2009. ggplot2: Elegant Graphics for Data Analysis. Media 35, 211. https://doi.org/10.1007/978-0-387-98141-3

Wilson, M.E., 2005. Travel and the emergence of infectious diseases. J. Agromedicine 9, 159–177. https://doi.org/10.1300/J096v09n02_10

Wingfield, W.H., Fryer, J.L., Pilcher, K.S., 1969. Properties of the Sockeye Salmon Virus (Oregon Strain). Proc. Soc. Exp. Biol. Med. https://doi.org/10.3181/00379727-130-33719

Woese, C.R., Fox, G.E., 1977. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. Proc. Natl. Acad. Sci. U. S. A. 74, 5088–5090. https://doi.org/10.1073/pnas.74.11.5088

Wolf, K., Snieszko, S.F., Dunbar, C.E., Pyle, E., 1960. Virus Nature of Infectious Pancreatic Necrosis in Trout. Proc. Soc. Exp. Biol. Med. https://doi.org/10.3181/00379727-104-25743

Wood, D.E., Salzberg, S.L., 2014. Kraken: Ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 15, R46. https://doi.org/10.1186/gb-2014-15-3-r46

Wooley, J.C., Ye, Y., 2010. Metagenomics: Facts and artifacts, and computational challenges. J. Comput. Sci. Technol. https://doi.org/10.1007/s11390-010-9306-4

Worobey, M., Watts, T.D.D., McKay, R.A.A., Suchard, M.A.A., Granade, T., Teuwen, D.E.E., Koblin, B.A.A., Heneine, W., Lemey, P., Jaffe, H.W.W., 2016. 1970s and "Patient 0" HIV-1 genomes illuminate early HIV/AIDS history in North America. Nature 539, 98–101. https://doi.org/10.1038/nature19827

Wylie, T.N., Wylie, K.M., Herter, B.N., Storch, G.A., 2015. Enhanced virome sequencing through solution-based capture enrichment. Genome Res. https://doi.org/10.1101/gr.191049.115

Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., Huang, W., He, G., Gu, S., Li, S., Zhou, X., Lam, T.W., Li, Y., Xu, X., Wong, G.K.S., Wang, J., 2014. SOAPdenovo-Trans: De novo transcriptome assembly with short RNA-Seq reads. Bioinformatics. https://doi.org/10.1093/bioinformatics/btu077

Xu, C., Evensen, Ø., Munang'andu, H.M., 2016a. De novo transcriptome analysis shows that SAV-3 infection upregulates pattern recognition receptors of the endosomal toll-like and RIG-I-like receptor signaling pathways in macrophage/dendritic like TO-cells. Viruses. https://doi.org/10.3390/v8040114

Xu, C., Evensen, Ø., Munang'andu, H.M., 2016b. A de novo transcriptome analysis shows that modulation of the JAK-STAT signaling pathway by salmonid alphavirus subtype 3 favors virus replication in macrophage/dendritic-like TO-cells. BMC Genomics. https://doi.org/10.1186/s12864-016-2739-6

Xu, C., Evensen, Ø., Munang'andu, H.M., 2015. De novo assembly and transcriptome analysis of Atlantic salmon macrophage/dendritic-like TO cells following type I IFN treatment and Salmonid alphavirus subtype-3 infection. BMC Genomics. https://doi.org/10.1186/s12864-015-1302-1

Xu, C., Mutoloki, S., Evensen, Ø., 2012. Superior protection conferred by inactivated whole virus vaccine over subunit and DNA vaccines against salmonid alphavirus infection in Atlantic salmon (Salmo salar L.). Vaccine 30, 3918–3928. https://doi.org/10.1016/j.vaccine.2012.03.081

Yates, M. V., Gerba, C.P., Kelley, L.M., 1985. Virus persistence in groundwater. Appl. Environ. Microbiol. 49, 778–781. https://doi.org/10.1128/aem.49.4.778-781.1985

Yin, Y., Fischer, D., 2008. Identification and investigation of ORFans in the viral world. BMC Genomics 9, 1–10. https://doi.org/10.1186/1471-2164-9-24

Zerbino, D.R., 2010. Using the Velvet de novo assembler for short-read sequencing technologies. Curr. Protoc. Bioinforma. https://doi.org/10.1002/0471250953.bi1105s31

Zhang, C., Cleveland, K., Schnoll-Sussman, F., McClure, B., Bigg, M., Thakkar, P., Schultz, N., Shah, M.A., Betel, D., 2015. Identification of low abundance microbiome in clinical samples using whole genome sequencing. Genome Biol. https://doi.org/10.1186/s13059-015-0821-z

Zhang, N.Z., Zhang, L.F., Jiang, Y.N., Zhang, T.Z., Xia, C., 2009. Molecular analysis of spring viraemia of carp virus in china: A fatal aquatic viral disease that might spread in East Asian. PLoS One 4, 6337. https://doi.org/10.1371/journal.pone.0006337

Zhao, S., Zhang, Y., Gamini, R., Zhang, B., Von Schack, D., 2018. Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: PolyA+ selection versus rRNA depletion. Sci. Rep. https://doi.org/10.1038/s41598-018-23226-4

Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W., Si, H.-R., Zhu, Y., Li, B., Huang, C.-L., Chen, H.-D., Chen, J., Luo, Y., Guo, H., Jiang, R.-D., Liu, M.-Q., Chen, Y., Shen, X.-R., Wang, X., Zheng, X.-S., Zhao, K., Chen, Q.-J., Deng, F., Liu, L.-L., Yan, B., Zhan, F.-X., Wang, Y.-Y., Xiao, G.-F., Shi, Z.-L., 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. https://doi.org/10.1038/s41586-020-2012-7