2018

Detection and serotyping of foot-and-mouth disease virus with laboratory and in silico methods

Ortega Polo, Rodrigo

Lethbridge, Alta. : Universtiy of Lethbridge, Department of Biological Sciences

DETECTION AND SEROTYPING OF FOOT-AND-MOUTH DISEASE VIRUS WITH LABORATORY AND *IN SILICO* METHODS

RODRIGO ORTEGA POLO
Bachelor of Science, University of Lethbridge, 2011

A Thesis
Submitted to the School of Graduate Studies
of the University of Lethbridge
in Partial Fulfillment of the
Requirements for the Degree

MASTER OF SCIENCE

Department of Biological Sciences
University of Lethbridge
LETHBRIDGE, ALBERTA, CANADA

DETECTION AND SEROTYPING OF FOOT-AND-MOUTH DISEASE VIRUS WITH LABORATORY AND *IN SILICO* METHODS

RODRIGO ORTEGA POLO

Date of Defense: July 16, 2018

Dr. O. Lung Co-Supervisor	Adjunct Professor	Ph.D.
Dr. R.M. Golsteyn Co-Supervisor	Associate Professor	Ph.D.
Dr. C.P. Goater Committee Member	Professor	Ph.D.
Dr. E. Taboada Committee Member	Adjunct Professor	Ph.D.
Dr. I. Kovalchuk Examination Chair	Professor	Ph.D./M.D.

Abstract

Foot-and-mouth disease virus (FMDV) is a highly contagious animal pathogen and it has a variable genome and high antigenic variation. There are seven known serotypes of this virus: A, O, C, Asia1, SAT1, SAT2, and SAT3. The rapid detection and serotype characterization of the virus is instrumental for the prompt response by animal health authorities. This thesis presents the design and development of the first electronic microarray assay for the simultaneous detection and subtyping of FMDV. The assay was evaluated *in silico* and it was tested with 19 synthetic DNA constructs representing all 7 serotypes, followed by the testing with 23 viral RNA samples representing all 7 serotypes. Also, various *in silico* methods were compared for the classification of FMDV sequences using complete genomes and next generation sequencing (NGS) data. Finally, highly specific and highly sensitive single nucleotide variant signatures that distinguish the seven FMDV serotypes were discovered.

Acknowledgments

I could not have arrived to this point without the help of many people. First of all, I would like to thank my thesis supervisor, Dr. Oliver Lung, for giving me the opportunity of pursuing graduate studies under his supervision. Thanks for always being present as a supervisor and for being supportive of my thesis work and my development as a graduate student, even from the distance. Thanks for giving me the unique opportunity to visit you at the genomics unit of the NCFAD in Winnipeg. I will take one of your phrases with me for the rest of my career: "*Everyone is your teacher*". I would not have the opportunities I have now before me if it was not for your guidance. Thank you for that.

Big thanks to Roy Golsteyn for being generous with his time and for being an great co-supervisor. Roy, you have been an amazing mentor since my undergraduate days, and your support since the moment I considered applying for admission to graduate school has been invaluable. Your work ethic and commitment to your students is admirable. I will never forget that it did not matter that you were in France on your study leave, but you were always available to sign documents for my program and to attend my committee meeting via teleconference. Thanks for your support.

Thanks also to Cam Goater, who ignited in me the passion to pursue research on infectious diseases. Your course on the biology of symbiotic interactions was fundamental in my education. Thanks for your advice and encouragement all these years. I also would like to thank Dr. Eduardo Taboada for introducing me bioinformatics and for providing his guidance. Ed: your lessons still resonate with me and I will not forget them.

None of this would not have been possible without the amazing team at the CFIA that I was privileged to join in 2012. Tara Furukawa-Stoffer: you are the best technician I have

ever worked with. Thank you for your friendship and support all these years. Mat Fisher: a considerable part of this thesis could not have occurred without your help. Thank you very for being such an amazing colleague. Cody Buchanan: your work ethic and your constant strive for always being the best are admirable. Thanks for being a fantastic colleague and friend. Anthony Erickson: your knowledge and your lessons from the very beginning of graduate school were very helpful in my development as a student and as a person. Thanks to my good friend Dillon Barker for his continuous encouragement and support.

Thanks to Defence Research and Development Canada, and the Centre for Security Science's Chemical, Biological, Radiological-Nuclear, and Explosives Research and Technology Initiative (CRTI) for providing funding for project 09-403TA which supported the research on the design and development of the electronic microarray for detection and serotyping of FMDV. Thanks also to the University of Lethbridge for providing scholarships and grants to support my graduate work, and for the opportunity to be a teaching assistant under the supervision of Helena Danyk and Laurie Pacarynuk.

Thanks to Dr. Gary Van Domselaar and Shane Thiessen from the Bioinformatics Core Facility of the National Microbiology Laboratory (Public Health Agency of Canada) for making their computer cluster available for part of this research.

Thanks to Dr. Erasmus Okine, Dr. Tim McAllister, and Dr. Rahat Zaheer. You are all great leaders who inspire me to give my very best every day. Thanks for the amazing opportunities that you have provided me and for showing me that the sky is the limit. I am very lucky to work with you and I look forward to keep learning from you.

Thanks forever to Malin and Roberta Burnham, my adopted grandparents. Pursuing a graduate degree simply could not have happened without you. The phrase "*Community Before Self*" is present with me every day, and I always strive to use collaboration and mentorship in my life.

Finally, thanks to my mother, who has been my example and inspiration all my life. Thanks for teaching me persistence and tenacity and to never give up in the presence of

adversity. You are my biggest role model, and you keep inspiring me to do my best. I love you.

Contents

List of Tables

List of Figures

List of Abbreviations

AUC Area Under the Curve
BLAST Basic Local Alignment Search Tool
BTV Bluetongue virus
BVDV Bovine viral diarrhea virus
CFIA Canadian Food Inspection Agency
CSFV Classical swine fever virus
EID Emergent and infectious diseases
ELISA Enzyme-linked immunosorbent assay
FI Fluorescence intensity
FMD Foot-and-mouth disease
FMDV Foot-and-mouth disease virus
GPS Global Positioning System
HIV Human immunodeficiency virus
HPV Human papillomavirus
MSA Multiple Sequence Alignment
NCAD National Centres for Animal Disease
NCBI National Center for Biotechnology Information
NCFAD National Centre for Foreign Animal Disease
NGS Next Generation Sequencing
NHGRI National Human Genome Research Institute
NML National Microbiology Laboratory
NSBP Non-specific binding probe
NTC Non-template control
OIE World Organization for Animal Health
OTU Operational Taxonomic Units
PCR Polymerase Chain Reaction
PN Positive-to-Negative ratio
RAM Random Access Memory
RNA Ribonucleic acid
ROC Receiver Operating Characteristic
RT-PCR Reverse-transcriptase polymerase chain reaction
SARS Severe Acute Respiratorial Syndrome
SHC Swine High Consequence
SNP Single Nucleotide Polymorphism
SRA Sequence Read Archive
SVDV Swine vesicular disease virus
TBED Transboundary and Emerging Diseases
VESV Vesicular exanthema of swine virus
VSV Vesicular stomatitis virus

Chapter 1

Literature Review

1.1 Abstract

Transboundary and Emerging Diseases (TBED) are a concern for human and animal health authorities around the world. A critical component to the response of those diseases is the detection methodology used for establishing the presence of their causative pathogens. A large number of serology and molecular biology methods have been developed and updated to detect and differentiate pathogens. Due to their higher throughput and increased automation, microarrays were once considered as viable alternative methods to more traditional techniques to detect and differentiate pathogens. Microarrays proved to be valuable in cases where diagnosis was not possible with other techniques or in the development of tools designed to detect a large number of genera of bacteria and viruses. The following review will present the state of the art in molecular testing microarrays for detection and subtyping of pathogens. Examples from human and veterinary diagnostics will be discussed, and a strong emphasis will be placed in the history of detection and serotyping methods for FMDV, a highly contagious and highly variable pathogen of cloven-hoofed animals. The advances in microarray technologies here reviewed are presented in the context of the recent increase in prevalence and affordability of next generation sequencing.

1.2 Transboundary and emerging infectious diseases of animals

Infectious diseases that appear for the first time or reappear in a population are known as emergent and infectious diseases (EID), and their impact on human and animal health

are an important issue for the general population and governments around the world. Notable outbreaks of EIDs have elicited the action of international human and animal health authorities in several occasions (Table 1.1). The mechanisms by which infectious diseases can emerge are diverse and complex, and thus present challenges to governments and international organizations in the management and control of infectious diseases.

1.2.1 Mechanisms of emergence of viral diseases

The emergence of infectious diseases can be influenced by the dynamics of the ecology and biology of the host, the pathogen, or both. Examples of important ecological factors that influence the emergence of diseases are the rapid growth of human populations, spillover effects and climate change (Daszak et al., 2000). Examples of biological factors that shape the emergence of infectious diseases are the evolution of viruses, and the response of the host immune system to infection. In the case of viral evolution, emergent viruses can arise by processes known as antigenic drift and antigenic shift. Antigenic drift occurs when mutations in the genes of viruses accumulate over time as the viruses replicate. This results in viruses that are phylogenetically related and show the same antigenic properties. Antigenic shift is a sudden change in the viral genomes that results in new infectious properties. A clear example of this was the H1N1 pandemic where reassortment of the influenza genome resulted in the transmission of highly virulent strains.

1.2.2 Policies and government-led actions

The mitigation of outbreaks of emergent zoonotic (diseases that are transmitted from animals to humans) and animal infectious diseases have required the collaboration between international organizations such as the World Health Organization, the World Organization for Animal Health (OIE)[1], and the Food and Agriculture Organization of the United Nations (FAO) to mitigate the impact of those events in human and animal health. The mandate

[1] OIE is the French acronym for *Office International des Epizooties*. The organization changed its name to *Organisation Mondiale de la Santé Animale* in 2014, but kept the acronym OIE.

of those institutions is to oversee the preparedness and response to emerging infectious diseases. Thus, these institutions set the regulations, policies and actions to prevent and respond to zoonotic and animal emergent infectious diseases.

The Canadian Food Inspection Agency (CFIA) enforces the Health of Animals Act, which establishes the mandate to protect the health of livestock animals in Canada against foreign and domestic animal diseases (Department of Justice Canada, 2015). The CFIA has collaborated with other federal and provincial institutions in mitigating and investigating high-profile outbreaks such as Bovine Spongiform Encephalopathy (BSE) outbreaks in Alberta and the highly pathogenic avian influenza virus (HPAIV) outbreak in British Columbia in 2014-2015. In 2017 the CFIA was investigating an outbreak of bovine tuberculosis in Southeastern Alberta and Southwestern Saskatchewan.

Table 1.1: Notable events of emergent infectious human, zoonotic and animal diseases.

Year	Disease	Geographical location	Hosts	Reference
2016	Zika virus	The Americas	Humans	Kindhauser et al. (2016)
2014-2016	Ebola haemorrhagic fever	West Africa	Humans	Gire et al. (2014)
2014	Highly pathogenic avian influenza (H5N2)	British Columbia, Canada	Poultry birds	Xu et al. (2016)
2012	Schmallenberg virus	Germany	Cattle	Hoffmann et al. (2012)
2012	Middle-East respiratory syndrome coronavirus (MERS-CoV)	The Middle East	Humans, camels, bats	Alagaili et al. (2014)
2009	Pandemic H1N1 influenza	Worldwide	Humans, pigs, birds	Smith et al. (2009)
2003-2004	Pandemic H5N1 influenza	Eastern Asia	Poultry and humans	Li et al. (2004)
2003	Severe acute respiratory syndrome coronavirus (SARS-CoV)	Canada, China	Humans	Ksiazek et al. (2003)
2001	Foot-and-mouth disease	United Kingdom	Livestock	Haydon et al. (2004)
1994	Hendra and Nipah virus	Australia	Horses, bats	Selvey et al. (1995)

Sharing information among animal health institutions is key during situations like the aforementioned outbreaks. This rationale led to the origin of the Canadian Animal Health Surveillance Network (CAHSN), which was created to monitor and respond to threats to animal, human and environmental health effectively under a One World-One Health initiative (Kloeze et al., 2010). This network approach has four fundamental components: laboratory diagnostics, surveillance, an information and technology platform and governance (Kloeze et al., 2010). Those components are the pillars of the specific goals of the CAHSN initiative, which include the creation of a national system to report animal disease threats as early as possible, the establishment of a laboratory network including federal and provincial laboratories for the rapid diagnosis of animal diseases that have high consequences, and the development of a network to share information between federal and provincial institutions including animal and human public health organizations (Kloeze et al., 2010).

Although initiatives like CAHSN are important efforts that address challenges for the surveillance and response to foreign animal diseases, the surveillance and control of highly contagious pathogens of animals are still difficult activities for animal health organizations. This could be explained by the fact that the transmission and spread of highly contagious pathogens is not stopped by artificial or natural boundaries. Those pathogens can cause profound effects in the global trade of livestock and, by extension, they can cause a severe impact to people who depend on livestock. The OIE and the FAO refer to those diseases as transboundary diseases, and the FAO considers them emergencies (Food and Agriculture Organization of the United Nations, 2016).

The OIE requires member countries to notify animal health authorities of the presence of some diseases that have a high risk of transmission and a high potential impact on the health of animals and the international trade of livestock. Thus, the OIE includes those diseases in its list of notifiable terrestrial and aquatic animal diseases (World Organisation For Animal Health, 2017a) [2]. In Canada, the Health of Animals Act also requires that

[2]Before the resolutions adopted by the International Committee of the OIE during its 72nd General Session in 2004 (World Organisation For Animal Health, 2004), there used to be two lists of notifiable diseases to

these diseases be reported (Department of Justice Canada, 2015). Some of the diseases in the OIE list are anthrax, bluetongue, BSE, bovine viral diarrhea, classical swine fever, foot-and-mouth disease (FMD), highly pathogenic avian influenza, rinderpest [3], rabies, and many others (World Organisation For Animal Health, 2017a). Many of those diseases are considered transboundary diseases, and regulations are in place across the world to minimize the likelihood that those diseases enter other countries. Many of the travel restrictions and policies, such as the prohibition against bringing meat products to other countries are in place to avoid those diseases. A disease from that list that is particularly important is foot-and-mouth disease. It is one of the most contagious animal diseases and it is the largest barrier to the international trade of animals and animal products. FMD is an extremely contagious vesicular disease of cloven-hoofed animals that has severe consequences in the international trade of livestock and animal welfare around the world. For decades, the rinderpest virus was one of the biggest menaces to the health of cattle across Asia, Africa, and Europe. Rinderpest was declared eradicated by the FAO and the OIE in 2011, and since then, FMD is arguably the greatest threat to the international trade of livestock.

1.3 Foot-and-mouth disease virus

The foot-and-mouth disease is caused by the foot-and-mouth disease virus (FMDV), which was the first pathogen characterized as an animal virus (Loeffler and Frosch, 1898; Brown, 2003). It is thought that the first written reports of the disease date back to 1514 and were authored by the Italian monk Fracastorius (Jamal and Belsham, 2013). The importance of FMD for animal health around world is reflected by its presence in the scientific

the OIE: List A and List B. List A was for those "transmissible diseases that have the potential for very serious and rapid spread, irrespective of national borders, that are of serious socio-economic or public health consequence and that are of major importance in the international trade of animals and animal products" (World Organisation For Animal Health, 2017b). Foot and mouth disease belonged in List A under such classification system. List B was for those "transmissible diseases that are considered to be of socio-economic and/or public health importance within countries and that are significant in the international trade of animals andanimal products" (World Organisation For Animal Health, 2018). Diseases such as anthrax, rabies, and bovine tuberculosis belonged in List B

[3] There are only two viral diseases in the world that have been eradicated: smallpox and rinderpest.

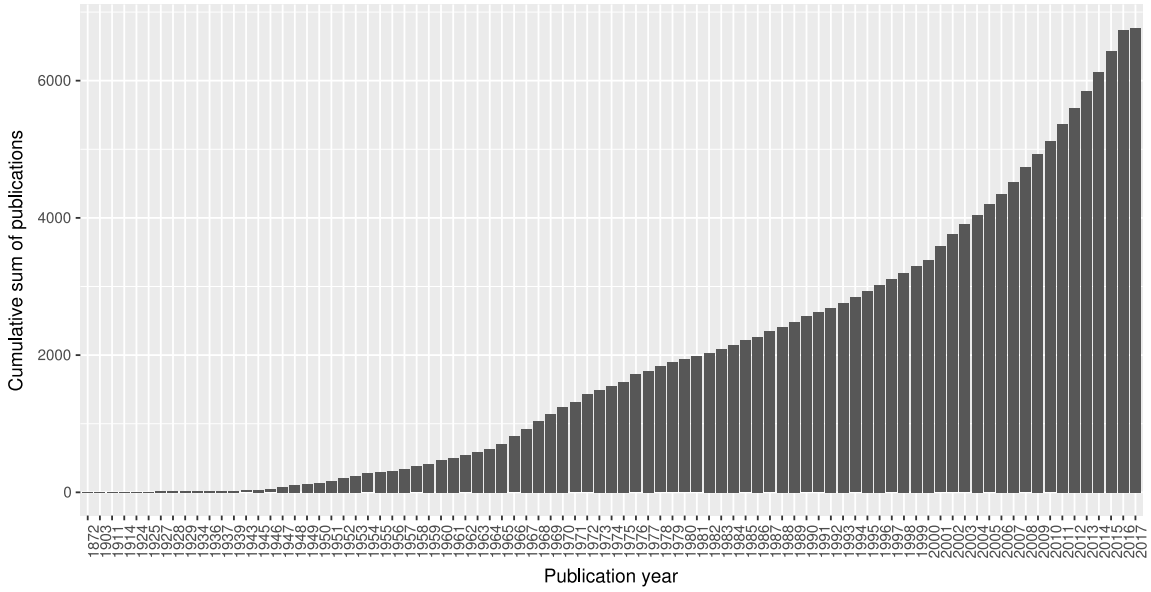


Figure 1.1: Cumulative sum of PubMed search results for the term *"foot-and-mouth disease"*. A PubMed query was submitted with the following terms: `foot-and-mouth disease[Title/Abstract] NOT hand[Title]`. That query yielded results where the term and "foot-and-mouth disease" was in the Title or Abstract of publications indexed in PubMed and that do not include the term "hand" in order to rule out results for the hand, foot and mouth disease. The PubMed search was performed on February 8, 2017. Data of number of publications per year were retrieved from PubMed's "Results by Year" option. The cumulative sum of publications over this time period was calculated using the R statistical computing language (R Core Team, 2016) and plotted with the `ggplot2` package (Wickham, 2009).

literature: a search of the term "foot-and-mouth disease" in Pubmed yielded $6,772$ results, which shows the relevance of this disease and its aetiological agent in the scientific literature since the year 1872 (Figure 1.1).

Understanding the genome of FMDV is key for studying the nature of its high virulence. FMDV is a non-enveloped, icosahedral virus with a positive sense single-stranded RNA genome of approximately 8.4 kilobases that belongs to the genus *Aphthovirus* in the family *Picornaviridae* (Alexandersen et al., 2003). The FMDV genome is composed of one open reading frame (ORF) that encodes a polyprotein that is cleaved by proteolysis into 12 structural and non-structural proteins (Alexandersen et al., 2003) (Figure 1.2, Figure 1.3). The FMDV genome has a viral protein (VPg) at its 5' end, and the long UTR at the 5'

Figure 1.2: Annotated complete genome of FMDV strain O TAW 1997 (GenBank accession: AF308157). Annotations of mature peptides are in green below the annotation for the polyprotein coding sequence (yellow).

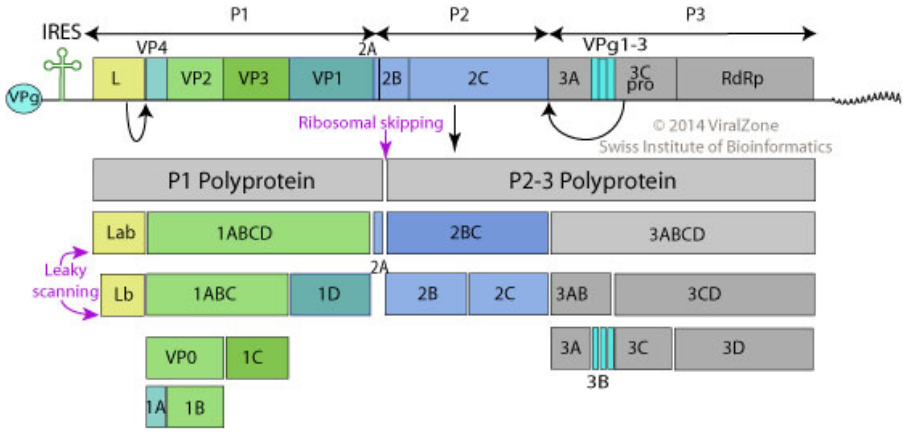


Figure 1.3: Proteolysis of the FDMV polyprotein. The schematic was reproduced from the ViralZone website of the Swiss Institute of Bionformatics (`http://viralzone.expasy.org/all_by_species/98.html`).

end contains an internal ribosome entry site (IRES). The P1 region of the genome encodes the structural polypeptides, and the P2 and P3 regions encode the nonstructural proteins associated with replication, such as the 3D polymerase. In addition, the FMDV genome encodes a N-terminal leader protease (L protease) and the 3C protease. Finally, the 3' UTR is important in negative strand synthesis (Alexandersen et al., 2003).

1.3.1 Economic impact and regulatory implications

According to the work by Knight-Jones and Rushton (2013), the economic impact of FMD can be categorized in direct losses and indirect losses. Direct losses are due to reduced production and changes in the structure of herds, while indirect losses are caused by the costs of FMD control, poor access to markets and limited use of improved production technologies (Knight-Jones and Rushton, 2013). The economic impact of FMD has been studied in simulations of the introduction of the disease to countries that have an FMD-free status (Knight-Jones and Rushton, 2013; Whiting, 2003; Carpenter et al., 2011) and in

reviews of the impact of the disease in small stakeholders (Knight-Jones et al., 2016).

Part of the economic impact of the disease comes from the notification of FMD to the OIE that triggers many restrictions for the trade of livestock and animal products. These restrictions represent important economic losses for those countries with FMD because they cannot trade live animals or livestock products with FMD free countries. If regular outbreaks occur, then only processed products can be exported to FMD free countries; if FMD is effectively controlled with vaccination and outbreaks can be detected, then meat without bones can be exported. Moreover, even if a country is free of FMD, it will have trade restrictions imposed to it if it trades with countries with FMD (Knight-Jones and Rushton, 2013).

Simulations of the economic impact of the introduction of FMD to the USA and Canada have shown that the emergence of FMD would be deleterious for the economies of these two countries that currently are free of FMD. For example, the cost of an outbreak of FMD in Canada has been estimated to be between $8.3 and $45.9 billion (CAD) (Whiting, 2003), and the cost of an outbreak in California was estimated to be between $2.3 and $69.0 billion (USD) (Carpenter et al., 2011). In both cases the range of estimated costs also take into consideration the number of days of delay in the detection of the disease.

Canadian livestock and meat export markets would be in severe risk if FMD enters the country. In 2015 Canada exported $829,842$ head of cattle, $5,768,198$ head of hogs and $5,509$ head of sheep to the United States of America, its largest export market (Agriculture and Agri-Food Canada, 2016). Canada's red meat industry exported $2.08 billion of beef and veal meat in 2016, while it exported $3.47 billion of pork in the same year (Agriculture and Agri-Food Canada, 2017). Considering the scale of those markets, steps must be taken in order to ensure the protection of the health of Canada's production animals and meat products. Furthermore, the impact of FMD in the regions where the disease is endemic is estimated at over 6.5 billion USD, which can be devastating to economically vulnerable livestock keepers and shepherds in those regions (Knight-Jones et al., 2016).

1.3.2 Clinical signs and symptoms

One of the factors that make FMD such an important threat to the international trade of livestock is its impact on the quality of meat and milk of infected animals. The clinical signs of FMD are the formation of vesicles and lesions in the hooves and inside and around the mouth, including the tongue (Alexandersen et al., 2003). The resulting pain usually leads to severe lameness which results in a decrease in meat and milk quality (Alexandersen et al., 2003). Although FMD is typically not fatal in adult animals, it is easily transmitted by aerosol and it may lead to myocarditis in young animals and to abortions in some females (Alexandersen et al., 2003). The clinical signs of FMD are very similar to multiple other viral vesicular diseases of animals such as vesicular stomatitis virus (VSV) (family *Rhabdoviridae*), vesicular exanthema of swine virus (VESV) (family *Caliciviridae*), swine vesicular disease virus (SVDV) (family *Picornaviridae*) and senecavirus A (family *Picornaviridae*). Clinical signs differ between species (Alexandersen et al., 2003), within vaccinated cattle herds, and within some cattle breeds from areas in which FMD is endemic and the disease may circulate undetected (Kitching, 2002). Therefore, FMD cannot be diagnosed on the basis of clinical signs alone and laboratory evidence of the demonstration of the antigen or nucleic acid of the virus is required (World Organisation for Animal Health, 2012).

1.3.3 Host range

Another reason why FMD is such an important disease for the international trade of livestock is that the FMDV has a very broad host range: it infects all cloven-hoofed mammals (order *Artiodactyla*), including domestic livestock species like cattle (*Bos taurus*), swine (*Sus scrofa*), sheep (*Ovis aries*), and goats (*Capra aegagrus hircus*), as well as over 70 wild species (Alexandersen and Mowat, 2005). Indeed, circulation of the virus between wild and domestic species has been observed in multiple cases and there are many literature reviews that mention documented cases of FMD in wild animals (Arzt et al., 2011; Thom-

son et al., 2003; Alexandersen and Mowat, 2005; Pinto, 2004). Notably, the African buffalo (*Syncerus caffer*) is an important reservoir host in Southern Africa, and some reports say that this species can carry the virus for up to 5 years or more (Alexandersen et al., 2003; Hedger et al., 1972). Thus, the African buffalo is known as a carrier of FMDV, and transmission of FMDV from carrier African buffalo to cattle has been identified (Alexandersen et al., 2003). Results from experimental studies have shown the susceptibility of wild animals to infection with FMDV as well as the potential for transmission to domestic species. For example, Moniwa et al. (2012) reported that Canadian white deer (*Odocoileus virginianus*) experimentally infected with FMDV were susceptible to the disease and transmitted the virus to cattle in an experimental setting.

Infection has also been described in species that belong to other taxa besides the order *Artiodactyla*, including elephants (order *Proboscidea*) (Hedger and Brooksby, 1976), bears (Order: *Carnivora*) (Officer et al., 2014), and many other species including vampire bats (Weaver et al., 2013). The literature review by Weaver et al. (2013) discusses in depth the presence of FMDV in wild animals. The taxon with the most published reports of FMD is the order *Cetartiodactyla*, although there were other 16 taxa represented (Figure 1.4).

Animals such as horses and carnivores are highly resistant to the virus, but can act as mechanical vectors and can transfer the virus if they are contaminated with it (Alexandersen and Mowat, 2005). This is why the CFIA requires that horses imported into Canada from countries with FMD are sprayed or sponged down with approved disinfectant such as Virkon, citric acid or diluted vinegar (Canadian Food Inspection Agency, 2012). CFIA regulations also stipulate that the hooves need to be cleaned and disinfected upon arrival. Disinfection steps for the import of horses to Canada also include the accompanying equipment and transport vehicles (Canadian Food Inspection Agency, 2012).

FMD is generally regarded as a disease that poses very little risk for human health, although some anecdotal accounts of infections of FMDV in humans have been reported in the literature. However, these accounts of symptoms do not meet Koch's postulates to

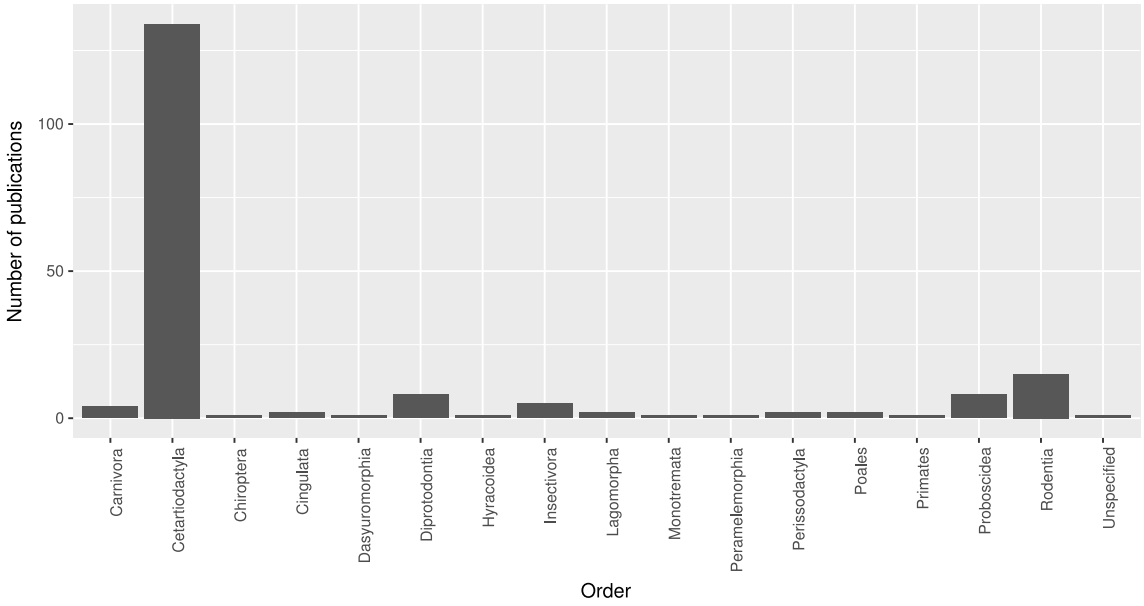


Figure 1.4: Number of publications per FMD host orders identified in the review by Weaver et al. (2013). The hosts listed in that review were identified from many papers in the literature. A Python script developed in-house (`https://github.com/ropolomx/OliVR/blob/master/retrieve_taxonomy.py`) was used to take the species names in Table 1 of Weaver et al. (2013) and automate searches of the NCBI Taxonomy database to retrieve genus, order, family and subfamily information.

confirm infection with FMDV (Alexandersen and Mowat, 2005). For example, in 1873, a British physician reported the symptoms of FMD in a female patient who had been in contact with cows infected with the disease (Briscoe, 1872). The physician also reported that he had observed milder clinical signs of the disease in three children who had drank milk from cows with the disease.

Foot-and-mouth disease is often confused by the layperson with the hand, foot and mouth disease which is caused by a coxsackievirus, a different member of the family *Picornaviridae*. Although humans are not hosts and do not play an important role in the amplification of the virus, they do play a role in the passive transport of the virus from infected animals or contaminated surfaces to susceptible animals.

1.3.4 Transmission

Besides its broad host range, another reason FMD is so difficult to control is its ease of transmission by different physical, biological and environmental means. The virus is often transmitted by direct contact via secretions and excretions from infected animals to susceptible animals (World Organisation for Animal Health, 2012), and other means of transmission like aerosol (Alexandersen and Donaldson, 2002; Mikkelsen et al., 2003) and fomites (i.e. objects that can carry pathogens) (Alexandersen et al., 2003) also make the virus hard to control. Moreover, animals that recover from FMDV infection can still excrete large numbers of virus particles (Alexandersen et al., 2003).

1.3.5 Distribution and global status

FMDV is enzootic to many countries in Africa, Asia and South America (Knowles and Samuel, 2003). Countries and regions that are currently free of the disease include Canada, the United States, Mexico, Central America, Australia, New Zealand, and parts of South America. The distribution of serotypes is roughly as follows: serotype A and O are found in many locations around the world, including South America, Africa, the Middle East, and Southern and Eastern Asia. Serotype Asia1 is found in Asia while serotypes SAT1, SAT2

and SAT3 are located in Southern Africa. Serotype C had been historically found in Europe, South America, East Africa, North Africa, and southern Asia (Knowles and Samuel, 2003). However, the OIE/FAO Foot-and- Mouth Disease Reference Laboratory Network reported in 2016 that serotype C has not been detected since 2004, and discussions were going to be held to consider the evidence in support of its extinction (OIE/FAO Foot-and-Mouth Disease Reference Laboratory Network, 2016).

The emergence of FMDV has been damaging to the economies of countries that had previously held an FMD-free status for decades. Notable examples of the emergence of FMD include the outbreaks in the United Kingdom in the years 2001 and 2007, the outbreak in Japan in the year 2000 (Sugiura et al., 2001), South Korea 2001 (Joo et al., 2002), and Taiwan in 1997 (Yang et al., 1999). Besides being one of the best documented, investigated, and studied outbreaks of FMD in history, the 2001 outbreak in the United Kingdom is an example of how devastating an FMD outbreak can be to a country's economy and agricultural sector. The cost of that outbreak was estimated to be 8 billion British pounds (Haydon et al., 2004), and there were also psychosocial consequences such as the documented stress experienced by British farmers due to the slaughter of over 6.5 million animals (Mort et al., 2005). Moreover, the trade restrictions that were imposed on the UK represented a large cost on that nation's agricultural industry for years following the 2001 outbreak.

FMD has not been reported in Canada since an outbreak occurred in Saskatchewan in 1952 (Sellers and Daggupaty, 1990), and the last time the disease was reported in the United States of America was in 1929 (Carpenter et al., 2011). The threat of this disease has lead many countries previously free of the disease to develop initiatives such as CAHSN in which laboratory diagnostics play a fundamental role.

1.3.6 FMDV serotypes and topotypes

The early detection of known and emerging animal diseases is one of the primary goals of the federal and provincial regulatory animal health agencies that are involved in the

CAHSN initiative (Kloeze et al., 2010). In the case of many bacterial and viral pathogens, the determination of the subtype or serotype of the pathogen in question is required to characterize better the circulating strains. Serotypes are determined by antigenic variants that result from variations in proteins in the viral surface, and the first report in the literature of classification of bacterial serotypes was done by Rebecca Lancefield in 1933 (Lancefield, 1933). In the case of FMDV, the serotype of an FMDV strain is determined by the VP1 surface protein which is encoded by the 1D gene. There are currently seven serotypes of FMDV: A, O, C, Asia 1 and SAT (Southern African Territories) 1, 2, and 3 (Alexandersen et al., 2003; Domingo et al., 2003).

Being a single-stranded RNA (ssRNA) virus, FMDV has high genetic and antigenic variation and it undergoes a rapid evolution, something that results in diversity that is observed even within viruses of the same serotype (Domingo et al., 2003; Sobrino et al., 1983). A key implication of the high antigenic diversity of FMDV is that protecting animals by vaccinating against one serotype does not provide protection against any of the other serotypes, and even between viruses of the same serotype (Alexandersen et al., 2003). A molecular epidemiology analysis reported that the seven FMDV serotypes can be classified into distinct genetic lineages based on differences in the sequence of the 1D gene. The variation between serotypes in this region ranges from 30% to 50% (Knowles and Samuel, 2003). Furthermore, the same study introduced the term topotype, which refers to the phylogenetic groups of strains of FMDV of the same serotype based on geography (Knowles and Samuel, 2003). The importance of knowing the serotype and topotype is reflected by the inclusion of these terms in the reports by the World Reference Laboratory for FMD on new strains that are sequenced by that laboratory.

The patterns of sequence variation between FMDV genomes are complex, and FMDV is one of most studied viruses in research to characterize and understand viral quasispecies (Domingo et al., 1985, 1992, 2005). Sequence analysis techniques for the study of viral quasispecies have been developed from work that was originally done on FMDV. Interest-

ingly, one of the first applications of split decomposition phylogenetic analysis, which is used to identify recombination and study reticular evolution, was in a study of foot-and-mouth disease virus (Dopazo et al., 1993). Because of the considerable genetic variation that is observed between FMDV genomes, there is no cross-protection between FMDV serotypes and between some topotypes, the rapid and accurate identification of the serotype of circulating strain is critical for the vaccine matching process with strains of the same serotype as the field strain (Paton et al., 2005).

Due to the importance of FMD to international animal health authorities, each year the FAO World Reference Laboratory and OIE Reference Laboratory for FMD at the Pirbright Institute (Pirbright, Surrey, United Kingdom) publishes a report on the global status of the disease. The World Reference Laboratory also publishes quarterly reports on the status of new strains for which the 1D gene has been sequenced. Phylogenetic trees and vaccine matching reports are included in such quarterly and annual reports which are available to the public. Because of the importance of characterizing the serotype of FMDV field strains and the rapid evolution and high genetic and antigenic variation of this virus the development of new and improved methods for detection and serotyping of FMDV is paramount to ensure successful monitoring of the disease should it ever enter countries that are currently free of it.

1.3.7 Methods for detection and serotyping of FMDV

Detection of FMDV is a topic of great interest in the scientific literature: according to a literature search in PubMed using the terms "foot-and-mouth disease virus detection" showed that 593 articles containing those terms have been published from 1952 to 2017 (Figure 1.5).

According to the World Organisation for Animal Health (2012), FMDV should be identified by confirmation of the viral antigen or nucleic acid from epithelial tissue. In the same document the following methods for the identification of FMDV are listed: virus isolation,

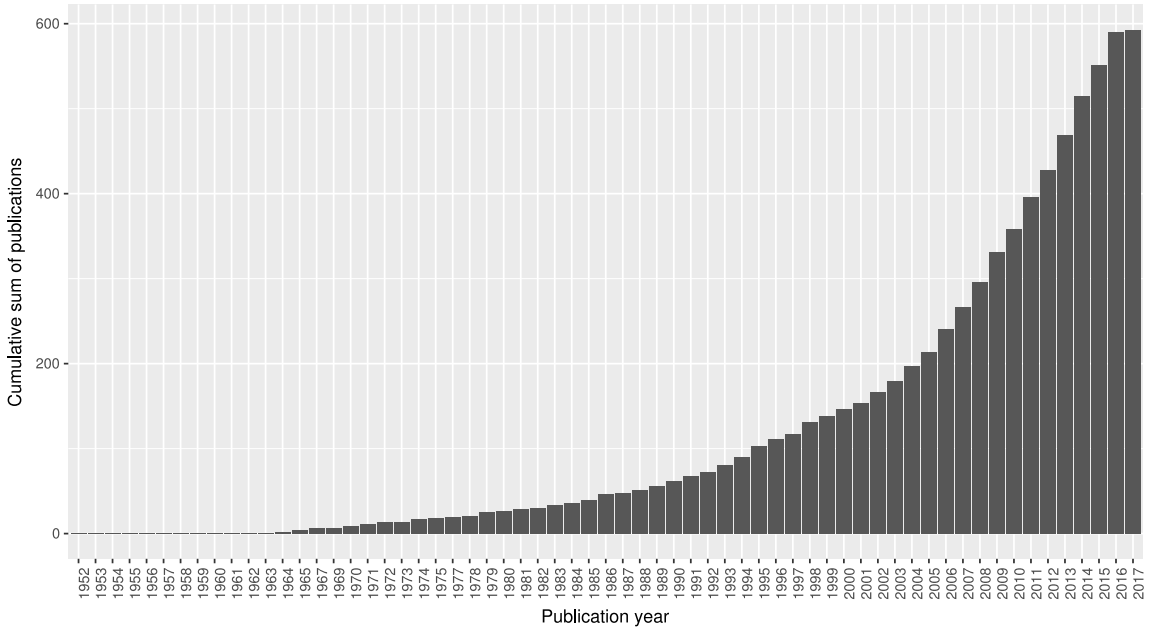


Figure 1.5: Cumulative sum of PubMed search results of the term *"foot-and-mouth disease virus detection"*. The PubMed search was performed on January 17, 2017. Data of number of publications per year were retrieved from PubMed's "Results by Year" option. The cumulative sum of publications over this time period was calculated using the R statistical computing language (R Core Team, 2016) and plotted with the `ggplot2` package (Wickham, 2009).

indirect sandwich ELISA, solid-phase competition ELISA (SPCE), liquid-phase blocking ELISA (LPBE), virus neutralisation test, complement fixation, lateral flow device tests, and nucleic acid detection methods such as end-point reverse-transcriptase PCR (RT-PCR) and real-time RT-PCR. To understand the state of the art of methods for the detection and serotyping of FMDV, a survey of those methods is presented.

1.3.7.1 Viral isolation

The gold standard of detection of FMDV is viral isolation in cell culture, for which the use of epithelium samples is preferable (World Organisation for Animal Health, 2012). The OIE recommends the use of primary bovine (calf) thyroid cells and primary pig, calf or lamb kidney cells or other cell lines such as BHK-21 (baby hamster kidney) and IB-RS-2 (porcine kidney cell line) (World Organisation for Animal Health, 2012). The cell cultures should be examined for cytopathic effect (CPE) for 48 hours. If no CPE is detected, the cells should be frozen and thawed, used to inoculate fresh cultures and examined for CPE for another 48 hours (World Organisation for Animal Health, 2012).

1.3.7.2 ELISA

The OIE's preferred method for the detection of FMDV and the identification of the serotype is the enzyme-linked immunosorbent assay Enzyme-linked immunosorbent assay (ELISA) (World Organisation for Animal Health, 2012), and it is one of the most used methods for the detection of foot-and-mouth disease virus and many other viral and bacterial pathogens. Equivalent or higher sensitivity of detection of FMDV than viral isolation and complement fixation has been achieved with ELISA assays (Ma et al., 2011). The ELISA method was first published in 1971 by Eva Engvall and Peter Perlmann (Engvall and Perlmann, 1971). The assay has been described by Dr. Engvall as a simple technique that made it easy for researchers to detect antibodies avoiding the use of radioactive isotopes (Engvall, 2010). The advantages of ELISA are that it is a fast method, the existence of highly automated ELISA protocols and that the results can be interpreted by a spectropho-

tometric scanner. In a direct ELISA test, in which antigens are detected, known monoclonal antibodies are placed in the walls of a microtitre plate and a diagnostic sample is added to each well. If the antigen, and by extension the virus, is present, the antigen will bind to the fixed antibody. To detect this reaction, secondary antibodies linked to an enzyme are added and a complex of fixed antibody, virus and enzyme-linked antibody is formed. Then, the substrate for the enzyme that is linked to the antibody is added and the reaction produces a product that causes a visible colour change. The colour change can then be measured by a spectrophotometric scanner.

In an indirect ELISA test, antibodies in a sample are detected. Instead of antibodies, in an indirect ELISA it is antigens that are adsorbed onto the walls of the well of the microtitre plate. Then serum sample is added and if antibodies against the fixed antigen are present, they will bind to the antigen. The way in which this reaction is detected is by adding an enzyme-linked anti-HISG immunoglobin that will bind to antibodies that can be used to detect expression of fusion proteins from bacterial, insect, and mammalian cells, in particular the sequence His-His-His-His-His-His-Gly (6xHis-Gly epitope). In a similar fashion to the direct ELISA, a sandwich of fixed antigen, antibody from the sample, and enzyme-linked anti-HISG is formed and substrate is added which results in a visible colour change if antibodies against the antigen are indeed present.

In the Manual of Diagnostic and Vaccines for Terrestrial Animals of the OIE it is clarified that this refers to an indirect sandwich procedure like the one described above. In the case of detection of FMDV, the wells of the microtitre plate are coated with rabbit antisera against each of the seven serotypes of FMDV. Sample suspensions are added to each of the wells of the plate as well as controls. Guinea pig antisera to each of the serotypes of FMDV are added next, followed by rabbit anti-guinea-pig serum immunoglobin conjugated to horseradish peroxidase. In this case, an absorbance reading greater than 0.1 above background indicates a positive reaction; the serotype of FMDV can also be identified. Values close to 0.1 should be confirmed by retesting or by amplification of the antigen by tissue

culture passage and testing the supernatant once a cytopathic effect has developed (World Organisation for Animal Health, 2012).

Furthermore, the OIE recognizes ELISA tests to detect antibodies against the structural proteins of the virus. These tests are serotype-specific, highly sensitive and able to detect antibodies resulting from infection as well as vaccination. These tests are appropriate for confirming previous or ongoing infection in non vaccinated animals as well as for monitoring the immunity conferred by vaccination in the field. These tests are considered by the OIE as prescribed tests for international trade of livestock (World Organisation for Animal Health, 2012). Ma et al. (2011) reviewed over 100 studies of ELISA for detection of FMDV and concluded that antibody-trapping and competitive ELISAs have high specificity and RT-PCR (oligoprobing) ELISA has extra sensitivity. The limitations of ELISA are that it does not have sufficient sensitivity and sensitivity and specificity, and that it takes hours to obtain results (Yang et al., 2013).

1.3.7.3 Virus neutralization

Like the SPCE and the LPBE, the virus neutralization test is a prescribed test for the international trade of livestock (World Organisation for Animal Health, 2012). Neutralization is the reaction between antigen and antibodies in which the antibodies block the cytopathic effect of the virus. This principle is used to detect the presence of antibodies against that virus in a sample. If the sample contains antibodies against the virus, the virus will be neutralized and it will prevent the virus from infecting cells. The virus can be titred using this method and the $TCID_{50}$ can be determined, which is the 50% tissue culture infective dose (World Organisation for Animal Health, 2012). Besides being a method used for the detection of FMDV antibodies, the virus neutralization assay is used for the process of vaccine matching.

1.3.7.4 Lateral flow assays

Lateral flow assays are emergent technologies for the detection of antibodies of FMDV. In a study on the development of lateral flow assays for the detection of FMD specific monoclonal antibodies against serotypes A, O, and Asia1 were produced and used as the capture monoclonal antibodies (Yang et al., 2013). Another multiplex lateral flow assay targeting serotypes A, O, and Asia1 that featured three test lines for the simultaneous testing of those three serotypes was also developed (Yang et al., 2015). That multiplex lateral flow strip test was able to detect and differentiate 46/46 serotype O, 45/45 serotype A, and 17/17 serotype Asia1 samples. However, there were 5/8 serotype C samples that cross-reacted with the serotype A test. Another lateral flow assay was used for detection of the viruses of the seven serotypes and for the differentiation of serotypes A, O, C and Asia1 (Morioka et al., 2015). That assay was able to detect samples representing all seven serotypes.

The lateral flow based immunological assays are relatively quick (less than 10 minutes), cheap and user-friendly; however, these tests are less sensitive than molecular assays (approximately 80%) (Yang et al., 2013, 2015; Morioka et al., 2015; Ambagala et al., 2016).

1.3.7.5 Molecular biology methods

The creation of the polymerase chain reaction (PCR) (Saiki et al., 1988) has resulted in the development of rapid, sensitive and specific laboratory methods that have been widely adopted by regulatory animal health laboratories for the detection of animal pathogens like FMDV. PCR requires the presence of a target, primers, polymerases such as *Taq* and dNTPs and ions to improve the efficiency of the enzyme. Because FMDV is a positive ssRNA virus and the thermostable *Taq* polymerase is a DNA polymerase, FMDV RNA is converted to DNA with the help of a reverse-transcriptase enzyme. This kind of assays is therefore known as reverse-transcriptase PCR (RT-PCR).

Traditionally, the amplified nucleic acids generated with PCR are visualized by agarose gel electrophoresis. In this method, the nucleic acid with a negative charge is moved

through an agarose gel matrix and ethidium bromide intercalates into the strands of DNA and it fluoresces under ultraviolet light. Size markers are used to measure the size of the amplified product.

This methodology is also known as end-point PCR, and its main challenge to the detection of foot-and-mouth disease virus is the variation of the sequences of new viruses. The rapid evolution of the FMDV makes it challenging to update the primers so that a high coverage of PCR primers remains.

Early PCR methods for the differentiation of FMDV serotypes include the work performed by Rodríguez et al. (1992) in which serotypes A, O and C were differentiated. Another multiplex PCR assay was developed to differentiate the seven serotypes (Callens and De Clercq, 1997). Constant refining and improvement of PCR methods for the detection of FMDV has been observed in the literature. The PCR procedure described by the Manual of Diagnostic Tests and Vaccines for Terrestrial Animals is performed at the OIE and FAO World Reference Laboratory at The Pirbright Institute and was developed by Reid et al. (2000).

The development of real time PCR has been a step forward in the molecular detection of FMDV and many other viral pathogens of animals. The real time reverse transcription PCR method referenced by the OIE involves the presence of a probe that generates a fluorescence reaction when it hybridizes to the target. A Ct threshold value is then measured and reported in real-time (hence the name). However, real-time PCR faces the same challenge of primer sensitivity as with end-point PCR. Although it has been suggested that the sensitivity of real time PCR can be higher than viral isolation (Pierce et al., 2010), many of the assays in the literature are designed to yield an universal product or cannot be used to discriminate between the seven serotypes. Also, some samples have not reacted with serotype-specific primers in RT-PCR experiments (Reid et al., 2001). The methods listed above normally require individual tests to be performed for each or a few serotypes.

Examples of other approaches for the detection of the nucleic acid of FMD are the

isothermal methods which do not require thermocycling such as the loop-mediated isothermal amplification (LAMP) and nucleic acid sequence-based amplification (Longjam et al., 2011). An insulated isothermal PCR (iiPCR) assay that targeted the FMDV 3D RNA polymerase gene, a highly conserved region of the FMDV genome, detected FMDV RNA in vesicular fluid samples without nucleic acid extraction (Ambagala et al., 2016). The advantages of iiPCR, and in particular of the POCKIT instrument, are that it is compact and potentially field deployable and the assay can be completed within one hour (Ambagala et al., 2016).

1.3.7.6 Next generation sequencing

Next generation sequencing (NGS) technologies (Shendure and Ji, 2008) have enabled the high throughput sequencing of many viruses, including FMDV. Sequencing FMDV is important for monitoring global patterns of virus distributions, tracing sources of outbreaks, and recognize early the emergence of new lineages, as well as antigenic prediction and vaccine selection. Advances in NGS technologies have enabled researchers with the possibility of sequencing the complete genome of this virus, which brings the advantage that the full genome sequence data has increased resolution for analyzing and tracing the spread of FMDV during outbreak situations. Another advantage of using NGS is that it allows the study of population diversity beyond consensus sequences within samples (i.e. viral quasispecies) (Freimanis et al., 2016). A limitation of NGS that has been noted in the literature is it becomes difficult to identify recombination between different strains if the samples contain parental and recombinant strains (Jamal and Belsham, 2018).

1.3.7.7 Microarrays

Microarrays occupy an interesting place in the history of pathogen detection methods because they were some of the first technologies with the ability of performing multiple molecular tests in a single reaction. Microarrays were developed in the 1990s as a method for quantifying the expression levels of many genes in parallel, and the first microarray

measured the expression of 45 *Drosophila* genes (Schena et al., 1995). Microarrays have high throughput since many samples can be tested at the same time, and these technologies contain collections of oligonucleotides which can be probed with target molecules (e.g. PCR products) to produce signals that can be quantified by a specialized instrument (Miller and Tang, 2009). One of the most important advantages of microarrays over other molecular methods such as PCR is that hundreds, thousands, or even tens of thousands of oligonucleotide probes can be used to simultaneously test multiple samples. This makes microarrays suitable for testing the presence of many pathogens simultaneously or multiple subtypes of the same pathogen in parallel, thus saving the need for conducting a large number of separate tests. The adoption of microarrays for the detection and subtyping of pathogens is the result of the integration of innovations and advances in molecular biology, biochemistry, robotics and bioinformatics (Miller and Tang, 2009).

Those innovations are presented in the literature review by Miller and Tang (2009), a comprehensive survey of the application of microarrays to clinical microbiology. That review presented the state of the art up to 2009 in detail, and discussed the technical advances and diversity of microarray technologies and the various applications of these technologies to laboratory diagnostics of microorganisms. Those applications include the detection and subtyping of pathogens, the determination of antimicrobial resistance in bacteria, the profiling of the expression of microbial genes, the profiling of the expression of host genes during microbial infections and the determination of host genetic polymorphisms related to the host immune response (Miller and Tang, 2009).

Two notable examples of the application of microarrays to clinical microbiology are the first panviral microarray (Wang et al., 2002; Chen et al., 2011) and the first panmicrobial microarray (Palacios et al., 2007). The first panviral microarray was known as the ViroChip, and it contained 70-mer oligonucleotides probes designed from highly conserved regions within viral families (Wang et al., 2002). Only the sequences of 148 complete viral genomes were used to design this chip, and although individual viruses were not explicitly

represented on the microarray, they were still detected (Wang et al., 2002), and one of the most important results obtained with the ViroChip was the discovery of the Severe Acute Respiratorial Syndrome (SARS) coronavirus (Ksiazek et al., 2003).

The panmicrobial microarray was developed by Ian Lipkin's laboratory and it was known as the GreeneChipPm (Palacios et al., 2007). The GreeneChipPm contained 29,455 60-mer oligonucleotide probes for viruses, bacteria, fungi, and parasites and it was successfully used to characterize a malarial infection of a patient who died from symptoms that appeared to be haemorrhagic fever and that it was initially thought to be viral in origin (Palacios et al., 2007). Many microarray assays have also been developed for the simultaneous detection and subtyping of animal viruses that share similar clinical signs. For example, the simultaneous detection and subtyping avian influenza virus and the simultaneous detection and pathotyping of Newcastle disease virus was achieved with electronic microarray assays (Lung et al., 2012). Another example of this kind of assay is that the detection and differentiation of vesicular diseases of livestock including FMDV, VSV, SVDV and VESV was achieved using a multiplex PCR and a glass slide microarray assay (Lung et al., 2011). Another glass slide microarray achieved the identification to genus level of many viruses that cause vesicles or vesicular-like lesions such as foot-and-mouth disease virus, vesicular stomatitis virus, swine vesicular disease virus, vesicular exanthema of swine virus, bovine herpesvirus 1, orf virus, pseudocowpox virus, bluetongue virus (bluetongue virus (BTV)) and bovine viral diarrhea virus (Bovine viral diarrhea virus (BVDV)) (Jack et al., 2009). Another slide microarray that differentiated viruses that cause vesicular diseases was the assay that discriminated between cDNA of FMDV, VSV and SVDV using padlock probes (Banér et al., 2007). Other examples of microarrays for the differentiation of pathogens include a glass slide microarray and an electronic microarray assay for the detection and differentiation of the multiple pathogens in the Porcine Respiratory Disease Complex (Lung et al., 2017), a microarray for the differentiation of lyssavirus species (Xi et al., 2012).

Microarray assays to differentiate viruses that infect the same species of animals can

be of interest to veterinarians and farmers. One example is a multiplex PCR and electronic microarray to simultaneously detect eight viruses that infect cattle: vesicular stomatitis virus, bovine viral diarrhoea virus type 1 and type 2, bovine herpesvirus 1 (BHV-1), BTV, malignant catarrhal fever virus (MCFV), rinderpest virus (RV) and parapox viruses (Lung et al., 2016). Another multiplex PCR and an electronic microarray for the differentiation of seven high consequence swine viruses which included: FMDV, swine vesicular disease virus (SVDV), vesicular exanthema of swine virus (VESV), African swine fever virus (ASFV), classical swine fever virus (CSFV), porcine respiratory and reproductive syndrome virus (PRRSV), and porcine circovirus type 2 (PCV2) (Erickson et al., 2017).

Glass slide microarrays were once the most common types of microarrays for pathogen detection assays. Some of their disadvantages are that multiple pieces of laboratory equipment are typically required when working with glass slide microarray protocols. For example, in Lung et al. (2011) slides were printed with probes that had been dried in using a speed-vacuum and were resuspended in an epoxide spotting solution. The probes were printed using a VersArray ChipWriter Pro printer with controlled humidity that was gradually raised from 60% to ambient humidity conditions and the slides were kept in a desiccating chamber until they were used (Lung et al., 2011). Then slides were pre-hybridized, then washed and dried in a slide centrifuge.

The automation of steps of the experimental workflow of microarrays allows for the rapid and accurate identification and characterization of human and animal pathogens with less human intervention. Highly automated microarray platforms such as the Nano-Chip 400 electronic microarray instrument (NC400; Nanogen Inc., San Diego, California, USA), in which probe printing, sample addressing, reporting, washing and image acquisition are programmed by the user and automatically performed by the instrument. RT-PCR and data analysis are not integrated into the Nano-Chip400 instrument and must be performed separately. In addition, the electrophoretic addressing of capture probes and hybridization driven by electrophoresis (as opposed to passive hybridization in glass slide microarray)

speed the process. The electrophoretically driven probe printing and sample addressing is done via the biotin-streptavidin bond between the 5'end of the probes and the streptavidin-coated hydrogel pads. If the probe hybridizes with the target amplicon, since the amplicon has a tag sequence which binds to a fluorescent reporter probe, the fluorescence will remain after several washing steps and will be quantified by a specialized instrument (Figure 1.6).

An important trend in the history of microarray technologies that was not captured in the review by Miller and Tang (2009) is the introduction of fully automated arrays. The development of lab-on-a chip technologies and innovations in engineering and microfluidics has allowed the creation of portable tests that can be applied bed-side and pen-side. Important advances in microfluidics include the development of glass slide lab-on-a-chip tests and the development of paper microfluidic devices (Martinez et al., 2008).

An example of full automation of array-based detection is the Rheonix platform in which nucleic acid extraction, nucleic acid amplification and detection by reverse dot blot microarray take place in a contained card with the aid of microfluidics (Spizz et al., 2012). This technology has been applied to a range of diagnostic scenarios such as the detection and differentiation of four sexually transmitted pathogens (*Neisseria gonorrhoeae*, *Chlamydia trachomatis*, *Treponema pallidum*, and *Trichomonas vaginalis*), the detection and differentiation of 20 clinically relevant human papillomavirus (Human papillomavirus (HPV)) types with cloned DNA, and the single nucleotide polymorphism genotyping of warfarin sensitivity markers (Spizz et al., 2012). Another example of fully automated Rheonix arrays is the simultaneous detection of Human immunodeficiency virus (HIV) antibodies and nucleic acids in oral fluids (Chen et al., 2013), the detection of *C. trachomatis*, *N. gonorrhoeae*, *T. vaginalis*, and *Mycobacterium genitalium* (Yasmin et al., 2016). More recently, a reverse-transcription loop-mediated isothermal amplification was coupled with automated reverse dot blot analysis for the detection of the Zika virus (Sabalza et al., 2018).

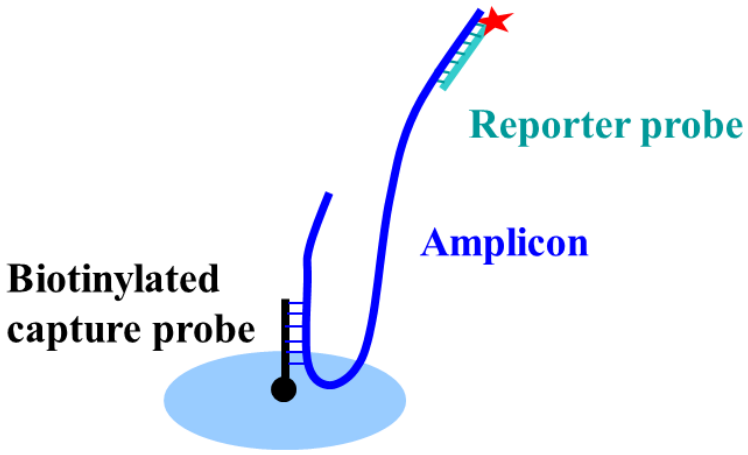


Figure 1.6: Schematic of the NC400 electronic microarray probe-target hybridization. The cartoon was reproduced from Ortega Polo et al. (2013)

.

1.4 Objectives

This thesis presents work that was developed during the transition of molecular methods and next generation sequencing being more widely adopted in the diagnostic laboratories.

The objectives of this thesis are:

1. To design, evaluate *in silico* and develop an RT-PCR and electronic microarray assay for the simultaneous detection and serotyping of FMDV (chapter 2).

2. To study the coverage and specificity of the microarray probes *in silico* detecting new strains (chapter 3).

3. To compare different *in silico* methods for the characterization of FMDV genomes and NGS data (chapter 3).

Chapter 2

Design and development of an electronic microarray assay for the simultaneous detection and serotyping of FMDV

Note to the reader: Some methods and results reported in this chapter were already presented in a poster at the 32nd Annual Meeting of the American Society for Virology (Ortega Polo et al., 2013) and published recently in the peer-reviewed article by Erickson et al. (2017). In that article, a multiplex RT-PCR and electronic microarray assay was designed to differentiate seven viruses affecting swine, while in the work presented in this chapter the focus is on the simultaneous detection and serotyping of foot-and-mouth disease virus.

2.1 Abstract

The foot-and-mouth disease virus (FMDV) is a highly contagious pathogen of cloven-hoofed animals, and one of the biggest hurdles for the international trade of livestock. The rapid and accurate detection of this virus is instrumental in preserving the livestock markets of countries that are currently free of the virus. In this study, an electronic microarray assay was developed for the simultaneous detection and serotyping of FMDV. The coverage and specificity of the assay for the detection and characterization of FMDV was evaluated *in silico*. The assay was able to detect and correctly type 19 synthetic constructs and 23 viral RNA samples in an initial laboratory validation. Further validation with additional samples, determination of the limit of detection, and transferring the assay to a portable platform are suggestions for improving this assay towards its adoption into veterinary diagnostics

programs in Canada.

2.2 Introduction

The detection of foot-and-mouth disease virus (FMDV) is a very important issue for animal health authorities and regulatory agencies around the world. Considered a reportable disease by the World Organization for Animal Health (OIE), the foot-and-mouth disease FMD is extremely contagious and has important consequences for the international trade of livestock. Because of the importance of FMDV, characterization of FMDV serotypes routinely applied as part of laboratory diagnostics of FMD. Several molecular and serological assays have been developed for the detection and serotyping. Although molecular assays such as PCR or real-time PCR have higher sensitivity than serological methods for virus detection, the genetic diversity of FMDV is very high, and molecular assays have to be monitored and updated regularly to ensure adequate coverage. Microarrays can include multiple probes which makes them suitable for subtyping highly variable and fast-evolving viruses like FMDV due to the ability of include probes that in combination provide higher coverage than individual oligonucleotides because probes that detect certain strains may not detect other strains. This property of microarrays is useful when probes that worked in the past with certain strains may not be as effective when tested with new field strains.

An example of the application of microarrays to the differentiation of viral subtypes was the development of a glass slide microarray that discriminated between the seven serotypes of FMDV (Baxi et al., 2006). In contrast to the labour-intensive and time-consuming glass slide microarray technologies, electronic microarrays are highly automated and allow the implementation of multiplexing and high-throughput testing of multiple pathogens and subtypes of pathogens. Moreover, electronic microarray assays have been successfully used in previous studies for the differentiation and subtyping of animal viruses (Gall et al., 2009; Lung et al., 2011, 2012, 2016; Erickson et al., 2017). In this study, the design, development, and initial laboratory validation of an electronic microarray for the simultaneous detection

and serotyping of FMDV are presented. This is the first report of an electronic microarray assay for the simultaneous detection and serotyping of FMDV.

2.3 Materials and methods

2.3.1 *In silico* assay design and evaluation

2.3.1.1 Reverse-Transcriptase PCR primers

The FMDV serotyping electronic microarray assay was designed to complement the multiplex Swine High Consequence (Swine High Consequence (SHC)) reverse-transcriptase PCR (RT-PCR) and microarray assay described in Erickson et al. (2017) which targeted the following viruses: swine vesicular disease virus (SVDV), classical swine fever virus (CSFV), vesicular exanthema of swine virus (VESV), African swine fever virus (ASFV), porcine circovirus (PCoV), and porcine respiratory and reproductive syndrome virus (PRRSV) (Table 2.1). In order to maximize the coverage of the genetically variable FMDV, the multiplex RT-PCR primers were designed to cover the high number of single nucleotide polymorphisms (SNPs) found across the FMDV strains. A pool of forward primers consisting of a set of 9 primers targeting the same binding region of the highly conserved 2A gene was used in this study, instead of a forward primer with multiple degeneracies to avoid the unnecessary dilution of the primers and to minimize the number of primers used in the RT-PCR. The sequence of the reverse FMDV primer contained two degenerate bases (Table 2.1).

Table 2.1: Primers of the Swine High Consequence multiplex RT-PCR

Primer name	Virus	Genomic region	Amplicon size (bp)	Sequence (5'-3')	Reference
VP3com980(1)	FMDV	1C-1D-2A-2B	971	GCTGATTACGCGTACAC	Erickson et al. (2017)
VP3com980(2)				GCTGACTACGCGTACAC	Erickson et al. (2017)
VP3com980(3)				GCGGATTACGCCTACAC	Erickson et al. (2017)
VP3com980(4)				GCGGATTACGCGTACAC	Erickson et al. (2017)
VP3com980(5)				GCTGACTATGCTTACAC	Erickson et al. (2017)
VP3com980(6)				AGTGACTTCTCCTACAC	Erickson et al. (2017)
VP3com980(7)				GCAGATTACGCGTATAC	Erickson et al. (2017)
VP3com980(8)				GCAGACTTTGCATACAC	Erickson et al. (2017)
VP3com980(9)				GCAGACTTTGCCTAYAC	Erickson et al. (2017)
FMDV 2B Rev 4026-S-Deg2*				GCGGACACCARCCGGTTRAAGTC	Erickson et al. (2017)
SVDVCV-3C-17a-F-(5875bp)	SVDV	3C/3D	791	CAGCGGCACTCCTCAGACACTAC	Lung et al. (2011)
SVDVCV-3D-3a-R-(6642bp)*				GAGTTTCAGGCACGTAAACCACAC	Lung et al. (2011)
KBH12-5 E1 Ext FWD	CSFV	E1/E2	671	AGRCCAGACTGGTGGCCNTAYGA	Paton et al. (2000)
KBH12-6 E2 Ext REV*				TTYACCACTTCTGTTCTCA	Paton et al. (2000)
King Long - Fwd Primer	ASFV	VP72	537	ATAGGATTAAAACCTACCTGGAACATCTCCG	King et al. (2003)
King Long - Rev Primer*				GGTACTGTAACGCAGCACAGCTGAACCGTTCTG	King et al. (2003)
VESVSM-2-F	VESV	Polymerase	649	CGACTCGATGGACCTGTTCACATACG	Lung et al. (2011)
VESVSM-5-R*				CGTAGAGGTCGGTTAGGTCCTTTCTG	Lung et al. (2011)
CircoV-1222F	PCV2	Capsid	534	GTAATCAATAGTGGAATCTAGGAC	Lung et al. (2017)
CircoV-1760R				TTCGTTTTCAGATATGACGTATC	Lung et al. (2017)
PRRSV-mtrx-F	PRRSV	Matrix	379	AAGGTAAGTCGCGGCCGAC	Lung et al. (2017)
PRRSV-mtrx-R				TGCCRCCCAACACGAGGC	Lung et al. (2017)

*Reverse primers contain a complimentary tag sequence (mutseq; 5'-GCAGTATATCGCTTGACA-3') for the LNA reporter probe at the 5' end (Lung et al., 2012).

Table 2.2: Concentration and volumes of the primers of the Swine High Consequence seven-plex RT-PCR

Virus	Primer	Initial concentration (μM)	Volume (μL)	Final concentration (μM)
FMDV	FMDV Forward primer pool	200	0.250	1.250
FMDV	mutseq FMDV 2B Rev 4026-S deg	200	0.250	1.250
SVDV	SVDV CV-3C-17a-F	200	0.125	0.625
SVDV	SVDV CV-3D-3a-R	200	0.125	0.625
CSFV	KBH 12-5 E1 Ext FWD	200	0.250	1.250
CSFV	mutseq KBH 12-6 E2 Ext REV	200	0.250	1.250
VESV	VESV SM-2-F	200	0.125	0.625
VESV	mutseq VESV SM-5-R	200	0.125	0.625
ASFV	ASFV-K F1	200	0.125	0.625
ASFV	mutseq ASFV-K R	200	0.125	0.625
PCV	CircoV 1222F	200	0.250	1.250
PCV	mutseq CircoV 1760 R	200	0.250	1.250
PRRSV	PRRSV-mtrx-F	200	0.125	0.625
PRRSV	mutseq PRRSV-mtrx-R	200	0.125	0.625
Total			2.500	12.500

2.3.1.2 *In silico* evaluation of FMDV primer coverage

The coverage of the FMDV primers was assessed *in silico* by mapping the sequences of the primers to a set of 288 FMDV whole genome sequences that were downloaded from GenBank on December 2013 representing all seven FMDV serotypes. The primer mapping process was started by performing a multiple sequence alignment of the 288 complete genome sequences with MAFFT v7.215 (Katoh and Standley, 2013) using the `--adjustdirection` parameter and applying the G-INS-i strategy. The sequences of the forward and reverse primers were added to the alignment using the `--addfragments` option and the `--adjustdirection` parameter of MAFFT. Once the primers were mapped to the multiple sequence alignment, a Python script developed in-house (extract_oligo_blocks.py; `https://github.com/ropolomx/OliVR/blob/master/extract_oligo_blocks.py`) was used to find the blocks of the multiple sequence alignment where the primer sequences were located.

Then, the script compared each position in the sequences of the primers versus each position in the sequences of the strains and a binary score was computed by assigning a value of 1 to matching bases and a value of 0 to mismatches. The script then calculated the number of mismatches as the Hamming distance (Hamming, 1950) by calculating the

sum of those positions where the bases of the primer and target are identical (i.e. had a score of 1), and then subtracting that result from the length of the primer. After running the script, the percentage of coverage of each primer was calculated as the number of hits against the sequence database with 0, 1, 2 or 3 mismatches divided by 288 and multiplied by 100. The binarized SNPs of primers versus targets were plotted using the `ggplot2` package (Wickham, 2009) in the R statistical computing language.

2.3.1.3 *In silico* evaluation of FMDV primer specificity

For the purposes of this *in silico* evaluation, the term "virus-specific" refers to the set of primer pairs that were designed for detecting each of the seven viruses of the SHC assay. The term "non-specific" refers to the other combinations of primer pairs of the SHC assay. For example, a non-specific pair would be the forward FMDV primer and the reverse classical swine fever virus (CSFV) primer.

The primers of the SHC assay were tested using different *in silico* methods against genetically related viruses such as other members of the genus *Aphthovirus* and the family *Picornaviridae*. The Basic Local Alignment Search Tool (BLAST) (Altschul et al., 1997), the electronic PCR (e-PCR) and reverse electronic PCR (re-PCR) tools developed by the National Center for Biotechnology Information (NCBI) (Schuler, 1997; Rotmistrovsky et al., 2004) were used to to identify potential cross-reactivities of the primers. A multi-FASTA file with the SHC multiplex RT-PCR primer sequences was manipulated with a Python script developed in-house (`multiplex_interactions.py`; `https://github.com/ropolomx/OliVR/blob/master/multiplex_interactions.py`) to generate a table with all the unique primer pairs that target viruses other than the viruses for which the primers were originally designed. That table was the input for the e-PCR program to evaluate the primers *in silico* versus all the sequences in the family *Picornaviridae* that are not FMDV. A total of 81,546 non-FMDV *Picornaviridae* sequences were downloaded from NCBI with a Perl script originally developed by NCBI in which the following

Entrez query was used: `txid12058[Organism:exp] NOT txid12110[Organism:exp] AND 100:11500[Sequence Length]`. Identical sequences were collapsed using `cdhit` v4.6 (Li and Godzik, 2006) using a sequence identity threshold of 100% which resulted in 61,995 unique sequences.

The unique non-FMDV *Picornaviridae* sequences were evaluated *in silico* with the SHC RT-PCR primers using e-PCR. The tests were performed with 1, 2, 3, 4, 5 and 6 mismatches. The specificity of the FMDV primers against host genomes was evaluated with *in silico* simulations of PCR under reaction conditions. Those PCR simulations were performed with the ThermoBLAST tool to predict possible hybridizations based on thermodynamic models and not sequence complementarity alone (SantaLucia, 2007).

2.3.1.4 *In silico* evaluation of probe coverage

Pan-FMDV and serotype-specific probes were adopted from (Baxi et al., 2006) and (Lung et al., 2011). A non-specific binding probe (NSBP) was included in all the experiments as a negative control. This oligonucleotide has been used in previous electronic microarray studies (Lung et al., 2012, 2015, 2016; Erickson et al., 2017) and it was designed from *Thermotoga maritima* sequences (Hindson et al., 2008). The evaluation methods of coverage for both pan-FMDV and serotype-specific probes is presented below.

Pan-FMDV probes The pan-FMDV probes were designed as positive controls to detect FMDV strains independently of their serotype and were designed to target a highly conserved region spanning the 2A and the 2B regions of the FMDV genome. A set of 335 sequences representing all 7 serotypes which contain that highly conserved region used to evaluate the coverage of the pan-FMDV probes (Table 2.3) were downloaded from NCBI in 2012. This database is a subset of the database of complete genomes used for the evaluation of primers in subsubsection 2.3.1.2. Probes were searches with BLAST and an identity threshold of 80% and a query coverage of 75% were used to determine matches of the probes vs. sequenced in the database.

Table 2.3: Number of sequences in the 335 FDMV sequence database with pan-FMDV probe region.

Serotype	Number of sequences
A	85
O	150
C	31
Asia1	40
SAT1	11
SAT2	9
SAT3	9

Serotype-specific probes A set of 2,888 FMDV sequences was used for the evaluation of the coverage and the specificity of the serotyping probes. Briefly, a set of accession numbers were retrieved in 2012 (n= 1,199) from the FMDV accession number databases maintained by the World Reference Laboratory for FMDV at The Pirbright Institute (Pirbright, Surrey, United Kingdom) (`http://www.wrlfmd.org/fmdv_seqs/fmdv_seq.htm`). The accession numbers were submitted to a GenBank search by means of a Batch Entrez query (`http://www.ncbi.nlm.nih.gov/sites/batchentrez`) and the matching records were downloaded from GenBank (Benson et al., 2009) in FASTA format. The sequences were added to a database previously curated by the CFIA Lethbridge Laboratory (n=1,689). Sequences of other FMDV genomic regions such as 3D (polymerase) or L (protease) and sequences that were less than 200 nucleotides were removed from the database before calculating the 2,888 sequences. Each one of the seven serotype-specific sequences files were aligned independently against a prototype sequence using the GlobalRef function of the Clone Manager 9.0 suite (Sci-Ed, Cary, North Carolina, USA). Those prototype sequences were identified from the public list of prototypes strains of the FMD World Reference

Laboratory (http://www.wrlfmd.org/fmd_genotyping/prototypes.htm) (Table 2.4). Then, to evaluate the *in silico* coverage of probes, the multiple sequence alignments were manually trimmed to 1.2 Kb using Bio-Edit v.7.1.3.0 (Hall, 1999). The 1.2 Kb region contains the 1C, 1D, 2A and 2B genes (Lung et al., 2011) and includes the 971 bp PCR product used in the present study.

Queries of the probe sequences against each of the trimmed serotype databases were submitted in local blastn searches (BLAST version 2.0) (Altschul et al., 1997) through the graphic user interface of Bio-Edit v.7.1.3.0 (Hall, 1999). The sequences of degenerate probes were expanded using a Perl script developed at the CFIA Lethbridge Laboratory (Beeston, A., https://gist.github.com/ropolomx/19220e9bd6e20088115c39ba0d3283b0). After the degenerate sequences were expanded, all probe sequences were evaluated against the sequence databases in local BLAST searches. Default Bio-Edit BLAST settings were applied. The BLAST tabular outputs were analyzed, and the probes that met the following stringent criteria were considered as matches and were selected for further testing. In this case, percent query coverage was not considered as a criterion due to the relatively short size of some of the oligonucleotides, and thus criteria based on the number of mismatches was used to evaluate the coverage of those probes.

- the probe-strain alignment was complete (100% of the probe length) with 0, 1 or 2 mismatches;
- the length of the probe-strain alignment was the probe length minus 1 base with 0 or 1 mismatches;
- the length of the probe-strain alignment was the probe length minus two bases with 0 mismatches.

Table 2.4: Prototype reference FMDV strains used in this study for guiding multiple sequence alignments

Serotype	Strain	GenBank Accession
A	A24 Cruzeiro	AY593768
O	O1 BFS/1860	AY593815
C	C Wald 32	AY593810
Asia1	Asia1 Pak 1/54	AY593795
SAT1	SAT1 BEC 30	AY593838
SAT2	SAT2 RHO	AY593847
SAT3	SAT3 BEC 1/65	AY593853

Table 2.5: Probes used in the NanoChip400 electronic microarray experiments.Size is represented as the number of nucleotides (nt). The oligonucleotide capture probes were biotinylated in the 5' end and addressed to the streptavidin-coated pads in NanoChip 400 cartridges and tested with synthetic constructs or viral RNA. The following characters represent IUPAC codes of degenerate bases: Y (C, T), B (C, T, G), M (A, C), D (A, G, T), R(A,G), S (C,G), W (A, T), and N (A, C, T, G)

Serotype	Probe	Sequence (5'-3')	Size (nt)
Pan-FMDV	FMD Common A (2A-1)	AAGTTGGCNGGAGACGTBGAGTCCAACCC	29
	FMD Common B (FV-SAT2-235)	AACTTYGACCTGTTAAAGTTGGCBGGAGACGTTGAGTC	38
	FMD Common C (FV-SAT1-227)	AACTTCGACCTGTTAAAGTTGGCYGGAGACGTTGAGTCCAACCCT	45
	FMD Common D (FMDCom2A-3)	GAGTCCAACCCTGGGCCYTTCTTCTTC	27
	FMD Common E (2AREV3915Probe(deg))	GAGAYGTBGAGTCCAACCCTGGGCCYTT	28
	FV-COM-081	ATTTTGACCTGCTCAAGTTGGCCGGAGACGTTGAGTCCAA	45
	Com81mod	AYTTYGACCTGYTMAAGTTGGCNGGAGACGTTGA	40
	FV-SAT1-226	AACTTCGACCTGTTAAAGTTGGCCGGAGACGTTGAGTCCAACCCT	45
	FV-SAT2-237	GACCTGTTAAAGTTGGCBGGAGACGTTGAGTCYAACCCTG	40
	FV-SAT2-238(237deg)	GACCTGTTAAAGTTGGCCGGAGACGTTGAGTCCAACCCTG	40
	FMDCom237deg2	GACCTGYTVAAGTTGGCNGGAGACGTBGAGTCCAACCCTG	40
	FMDCom2A–2	GACCTGYTVAAGTTGGCNGGAGACGTBGAGTC	32
A	A-T-23	GTGCGCATGAAACGGGCCGA	20
	A-T-50 deg	GARAACTACGGYGGTGAGACACAA	24
	A-t-009	CTTCTCGTGCGCATGAAACG	20
	A-T-27	AACGACACTCTGGTCGTGTCGGTTAGCGCCGGCAAGGA	38
	FMD-A-137deg	GTCGTCTCAAGACAGACACAAACAGAAGATCATTGC	36
	FMD-A-137	GTCGTCTCAAGACAGACACAAACAGAAGATCATTGC	36
	FV-A-119(008deg)	TCCTGCTTCWTTYAACTACGGTGCAATCAAGGCCV	35
	FMD A6	CACGARCTYCTCGTGCGYATGAA	23
	FV-A-008	TCCTGCTTCATTTAACTACGGTGCAATCAAGGCCG	35
	FV-A-135	GCACACCGACGTTACTTTCATAATGGACAGATTTGT	36
	A-T-1deg	GTCATTGACCTCATGCAAACCCACCARCAC	30
	FV-A-127	CTTCTCAGATCTGGAAATTGTTGTGCGGCATGACG	35
O	FV-O-148mod1	CAARTTAAYRTDTTGGACCTSATGCA	26

Table 2.5: Probes used in the NanoChip400 electronic microarray experiments (continued).

Serotype	Probe	Sequence (5'-3')	Size (nt)
O	O-T-16	ATGGGTCTGCTTGTTTCAAAT	21
	O-T-8deg	CCGGTYGACGCCCGMACCCA	20
	FV-O-014	CAATGTGAGAGGTGACCTGCAAGTATTGGCCCAAA	35
	FV-O-015	TACTACTTCGCAGATCTAGAAGTGGCAGTGAAACA	35
	FV-O-165	AACTGCTTTACCGCATGAAGAGGGCCGAAACATAC	35
	O-304-323	CAAGTGTTGGCCCAGAAGGCGGC	23
	O(Cody)	GGTGCCATCAAAGCCAC	17
	FV-O-154	ACACGGACATTGCGTTCATATTGGACAGGTTCGTG	35
	FV-O-155	AGGCAACTCGTGTTACTGAACTACTCTACAGAATGAAG	38
	FV-O-166	ACCAATGTGAGAGGTGACCTGCAAGTATTGGCTCA	35
	FV-O-167	TACTTCGCAGACCTAGAAGTGGCAGTGAAACACGA	35
	FV-O-169	GACACAAACAAAAGATTGTGGCGCCTGTGAAACAG	35
	FV-O-178	CGGACGTTGCGTTCATATTGGACAGGTTCGTGAAAGT	37
	O-T-14deg	TTRGCGACCGTCTACAACGG	20
	O-T-20deg	ACACAGRTCCAGAGGCGCCA	20
	O-T-31deg	TGGTCGTRCTGGCTAGYGCTGG	22
	FV-O-016	AATTAATGTGTTGGACCTGATGCAAACCCCTGCAC	35
	O-T-24deg	GAGATAGCAGTAAARCACGAGGGAGA	26
C	C24	TGAACTCTATTGTCCTAGGCCGATTCTTCCGATTC	35
	FMD C18	TGGAAATAGCRGTGACCCACAC	22
	C-L-002	CGATTCTTCCGATTCAGCCAC	21
	FV-C-183	WGAACTCTATTGYCCTAGGCCGATTCTTCCGATTC	35
	FV-C-192	TTGGTGCAGTTAAAGCAGAAACAGTCACTGAGCTG	35
	FV-C-195	AACACACACTGGACGTGATGCAGGTACACAAAGAC	35
	FMD VP1 643-662	CTCACATGGGTGCCCAACGG	20
	FMD VP1 481-497	GCCTTCGTTCTTGACCG	17
Asia1	ASIA208mod2	CTTGACACCACYCAVGACCGCCG	23
	ASIA-L011	ACGGYGCYGTRAAGGCYGA	19
	FV-Asia-206	CGCTGTAAAGGCTGACAACATAACTGAGCTTTTGATCCGCATGAA	45
	FV-Asia-040	AGGTCAAACTTCACAAAGTTGGTGGACACCATCAA	35
	FV-Asia-203(034deg)	TTYTCAGACCTGGARRTTGCGCTTGTYCACACAGG	35
	FV-Asia-204(036deg)	AARGAYTTTGAGTTTCGCYTRCWAGTKGAYGCACG	35
SAT1	SAT1-9	GTGGGCGACAAYCCVGTCGTCTTCTC	26
	S1-L-002	STYCTSACRCACTACGACCA	20
	SAT1-210mod2	TACAAYGGTGACTGYAAGTACAARCCCRCTGGC	33

Table 2.5: Probes used in the NanoChip400 electronic microarray experiments (continued).

Serotype	Probe	Sequence (5'-3')	Size (nt)
SAT1	SAT1 VP1-2	GARGTBGACGTRTACGTSAGGATGAA	26
	FV-SAT1-212	GGTGACTGCAAGTACAARCCCACTGGCA	28
	SAT1215mod3	GATWTACACASAGGCAGARGTGGACGTGTACSTGAGGATGAA	42
	FV-SAT1-267	TACAAAGTGGCCCTGACCAAACCTGCCAAACAATT	35
	FV-SAT1-268	GATCACCAAGCCTGTCAAACAACTGTGTAATTTCG	35
SAT2	SAT2-T-31	CCAGTCGACGTTTACTACCGGAT	23
	SAT2-T-23deg	GGCGTCGAGAAACAAYTGTG	20
	FMD S2-L-006	CTBCCNTCHACCTTCAAATTYGG	23
	FV-SAT2-261	TTTGCTGTGGACCTCTTGGACACTAAGGACAAAAC	35
	SAT2-T-13deg	GGCGACAACCCMATGGTKTTC	21
	SAT2-547-569	GACCTCATGGACACMAAGGAGAA	23
	SAT2-604-623	TACTACTTCTGTGACCTGGA	20
	SAT2-T-9deg	GACAGATTCGACGCGCCYAT	20
	FV-SAT2-255	GACAACCCAATGGTGTTTTCACACAACAACGTCAC	35
	FV-SAT2-265	GGTGAGTGTGTTTACAAGAAAACTCCCACCGCCAT	35
	Sat2-230	ACMTACTATTTYTGTGACCTGGAAATYRCVTGCCT	35
	FV-SAT2-068	ACCTACTATTTCTGTGACCTGGAAATTGCCTGCCT	35
	FV-SAT2-064	GATTCACCCATGTTCTGACAAATAGAACCGCGTTC	35
SAT3	SAT3-9	CCCACRACDTTCAACTTCGG	20
	SAT3-8	TGGGTDCCCAACGGDTGCCC	20
	S3-L-001	ACACGAAGGAACACACRYTG	20
	FV-Sat3-242	CACACCAACGTGGAGTTTCTGCTGGACAGATTCACA	36
	FV-Sat3-071	TACAAGACACCACTGGTCAAACCTGACAAGCAGAT	35
	FVSat3-244	CCCACGACATTCAACTTCGGAAGATTGTTGTGTGAA	36
Negative control	NSBP	CAAAGTGGGAGACGTCGTTG	20

2.3.1.5 *In silico* evaluation of probe specificity

To assess the specificity of the serotyping probes *in silico* for the detection of each serotype, probes of each serotype were queried in local BLAST and re-PCR searches against all the sequences of the other serotypes using the criteria mentioned above. For example, the serotype A probes were queried against all the FMDV sequences that were

not serotype A.

2.3.2 Optimization of the assay with synthetic constructs

2.3.2.1 Synthetic constructs

Synthetic plasmid DNA constructs containing partial sequences of the FMDV amplicon were used for screening FMDV detection and serotyping probes at the Lethbridge Laboratory of the National Centres for Animal Disease (CFIA). The synthetic constructs were designed by other personnel at that laboratory and more details of the design can be found in Appendix A. Plasmid vectors were cloned and propagated in *Escherichia coli* DH10B electrocompetent cells. A volume of 1 μL of a $1/100$ dilution of of each synthetic construct were added to 20 μL of *E. coli* DH10B cells. The cells were transferred to 1 mm electroporation cuvettes and electroporation was applied using a Gene Pulse X-cell system (Bio-Rad, Mississauga, Ontario) with the following conditions: voltage of 1800 V, capacitance of 25 μF, and resistance of 200 Ω. After electroporation, a volume of 480 μL of Super Optimal Catabolite (SOC) medium was added to the cuvettes. The mixtures of cells and medium were transferred to the original *E. coli* cultures and incubated for 1 hour at 37 °C. A volume of 250 μL of each transformed cell aliquot were plated on solid LB agar (Miller) medium with ampicillin and plates were incubated overnight. Plates were removed from the 37 °C oven, sealed with parafilm and stored at 4 °C overnight. Test tubes with LB broth medium with ampicillin were inoculated by picking colonies with a sterilized pipette tip grabbed with a sterilized pair of tweezers. The inoculated culture tubes were incubated overnight at 37 °C. Tubes were visually inspected for turbidity and glycerol stocks were prepared by adding 500 μL of 60%(w/w) glycerol and 500 μL of the culture grown on LB broth with ampicillin. Glycerol stocks were stored at -80 °C. A volume of 1.5 μL of overnight *E. coli* culture was transferred into a QIAcube microcentrifuge tube (QIAgen, Toronto, Ontario) by transferring 2 aliquots of 750 μL each. Those tubes were centrifuged for 3 minutes at $13,000$ rpm and the supernatant was discarded. The rest of the LB broth overnight culture

was added to the QIAcube microcentrifuge tubes (approximately 700 to 740μL). The tubes were centrifuged again for 3 minutes at 13,000 rpm and the supernatant was discarded. The pellets were temporarily stored at 4 °C. Plasmid DNA was extracted from the pellets by using the QIAprep Miniprep 1.5 mL LB culture without Buffer PB protocol using a QIAcube instrument (QIAgen, Toronto, Ontario). After the plasmids were purified and eluted, their concentrations were measured in triplicate using the Nanodrop 8000 instrument in which the EB buffer (Qiagen) that had been used for elution in the miniprep was used as blank reagent. Plasmid DNA samples were sent for Sanger sequencing to Eurofins MWG Operon (Louisville, Kentucky, USA).

Table 2.6: Synthetic constructs of FMDV used in this study

Serotype	Sample Name	FMDV Isolate Name	Topotype
A	sA-1	A_{22} Iraq 24/64	Asia
	sA-2	A_{24} Cruzeiro	Euro-SA
	sA-3	A Arg 2/2001	
	sA-4	A Arg /87	
	sA-5	A Iran 1/96	Asia
	sA-6	A_{22} Iran/99	Asia
O	sO-1	O_1 BFS 860	Euro-SA
	sO-2	O_1 Manisa	ME-SA
	sO-3	O TAW 10/97	Cathay
C	sC1 Noville	C_1 Noville	
	sC3 Resende	C_3 Resende	Euro-SA
Asia1	sAsia-1 PAK	Asia1 Pak 1/54	
	sAsia-1Shamir	Asia1 Shamir	
SAT1	sSAT1 KEN 4/98	SAT1 KEN 4/98	
	sSAT1 BOT 1/68	SAT1 BOT 1/68	III
SAT2	sSAT2 ZIM 7/83	SAT2 ZIM 7/83	G1
	sSAT2 SAU1/2000	SAT2 SAU 1/2000	G2
SAT3	sSAT3 BEC 1/65	SAT3 BEC 1/65	I
	sSAT3 ZIM 4/81	SAT3 ZIM 4/81	II

2.3.2.2 NanoChip400 electronic microarray

All the microarray detection and serotyping experiments were performed in a NanoChip 400 electronic microarray platform (Nexogen, San Diego, California, USA). Electronic microarray technologies have been described in much more detail in peer-reviewed publica-

Table 2.7: Foot-and-mouth disease virus strains used in this study.

Serotype	Sample Name	Topotype	Top BLAST Hit Accession	% Coverage	% BLAST Identity	Top BLAST hit isolate
A	A_{22} Iran1/96	Asia	AY593791.1	100	98	A/IRN/1/96
	A ARG 2/2001		AY593786.1	100	99	A ARG 39111_Cor_Gpaz
	A ARG/87	Euro-SA	KX002196.1	100	93	A ARG 39111_Cor_Gpaz
	A COL/85	Euro-SA	AY593794.1	100	100	A Sabana/COL/85
	A Iran 22/99	Asia	EF208772.1	79	96	A/Afyon/TUR/245/06/03
	A_{22} Iraq 24/64	Asia	AY593763.1	100	100	A_{22} Iraq 64 iso86
	A_{24} Cruzeiro/Bra/55	Euro-SA	AY593768.1	100	99	A_{24} Cruzeiro iso71
O	O_1 Manisa/TUR/87/67	ME-SA	AY593823.1	100	99	O_1 Manisa iso87
	O TAW 10/1997	Cathay	AY593835.1	100	99	O Taiwan 97 iso106/112
	O UKG 1/2001		KM257062.1	100	100	O UKG/1558/2001
	O_1 BFS/1860	Euro-SA	JX869188.1	100	100	O BFS 89/68
C	C_1 Noville	Euro-SA	AY593804.1	100	100	C_1 Noville/Switzerland/56/65
	C_3 Resende	Euro-SA	AY593807.1	100	99	C_3 Resende/Brazil/55
Asia1	Asia1 PAK 1/54		AY593795.1	100	100	Asia1 PAK iso3
	Asia1 Shamir		JF739177.1	100	100	Asia1/Shamir/89
SAT1	SAT1 KEN 4/98		DQ009721.1	89	100	SAT1 KEN 5/98
	SAT1 BOT 1/68	III	AY593845.1	100	99	SAT1 BOT 1/68 iso 47
SAT2	SAT2 ZIM 5/81		EF134951.1	88	100	SAT2 ZIM 5/81
	SAT2 SWA 1/69		KU821592.1	100	87	SAT2/ZAM18/2009
	SAT2 SAU 1/00	G2	AY297948.1	93	99	SAT2 SAU/6/00
	SAT2 ZIM 10/91		AF540910.1	99	93	SAT2 ZIM/7/83
SAT3	SAT3 BEC 1/65	I	AY593853.1	100	100	$SAT3_4$ Bech iso23
	SAT3 ZIM 4/81	II	KX375417.1	100	99	SAT3/ZIM/4/1981

tions (Miller and Tang, 2009; Lung et al., 2012) and in the introductory chapter of this thesis.

Samples of 3.5 μL of biotinylated oligonucleotide probes (5 μM in ultrapure water; 250 nM working concentration) were prepared in 66.5 μL of a 50mM histidine buffer containing 0.05% ProClin as described before (Lung et al., 2012). The probe and sample addressing layout of every cartridge was programmed to test multiple PCR products simultaneously with multiple capture probes by duplicate. In total, 117 capture probes were tested with FMDV synthetic constructs, including 12 pan-FMDV detection probes (positive controls for the presence of FMDV), 104 serotyping probes, and 1 non-specific binding probe (negative control). The biotinylated capture probes were electronically addressed at 350 nA/pad for 30 s to hydrogel pads containing streptavidin. Each FMDV synthetic construct sample was prepared by mixing 8.75 μL of RT-PCR product with 61.25 μL of high salt buffer (Nexogen). FMDV synthetic samples were electronically addressed to the array with a current of 88 nA/pad for 60 s.

The fluorescent reporting of the hybridization of the amplicons with the capture probes was performed with a locked nucleic acid (LNA) reporter probe (Integrated DNA Technologies, Coralville, Iowa, USA). The locked nucleic acid probe is useful for discriminating closely related targets such as the amplicons developed for this FMDV serotyping assay (You, 2006). The sequence of the locked nucleic acid was: 5'-/5Alex647N/TGT+**C**A+**A**GCG+**A**T+**A**T+**A**CT+**G**C-3', where the bases in bold letters contain the locked nucleic acid modification. Locked nucleic acids have methylene bridges that keep hybridization stable at high temperatures. The reporter probe used in this assay is complementary to the mutseq tag of the FMDV reverse primer sequence. The hybridization of the reporter probe to the mutseq tag of the synthetic construct products started at a temperature of 24 °C for 20 seconds with 2 °C increments up to 60 °C. After each temperature increment, a wash with Low Salt Buffer (Nexogen) was performed. Images were taken and fluorescence intensity (FI) was measured automatically by the NanoChip400 instrument. Electronic microarray experiments had a run time of 3 hours and 31 minutes.

2.3.3 Testing the assay with viral RNA

Details on the viral propagation and viral RNA extraction work can be found in Appendix A. Due to the highly contagious nature of FMDV that work was performed in the biocontainment level 3 facility of the National Centre for Foreign Animal Disease (NCFAD) by personnel of the Vesicular Disease Unit. The strains selected for the initial laboratory evaluation were twenty-three FMD viruses representing all seven serotypes and from different geographical origins (Table 2.7).

2.3.4 Reverse-transcriptase PCR amplification

The reaction conditions listed here are the same conditions used for the assay published by Erickson et al. (2017) in which a volume of 50 μL was prepared for each reaction. A master mix containing 19.5 μL of UltraPure water, 25 μL of 2X Superscript III Reaction Mix containing 0.4 mM of each dNTP and 3.2 mM MgSO4 (Invitrogen Life Technologies,

Burlington, Ontario), 10 μL of the SHC primer pool mix (Table 7), 2 μL of SuperScriptIII Platinum Taq DNA polymerase (Invitrogen Life Technologies) and 1 μL of RNA template. The following temperature profile was used for RT-PCR amplification: 15 min at 55 °C (reverse transcription), 2 min at 94 °C (inactivation of reverse transcriptase / activation of Taq polymerase), followed by 35 cycles of 30 s at 94 °C (denaturation) and 1 min at 50 °C (annealing) and 1 min at 68 °C (elongation) followed by a final extension of 5 min at 68 °C. RT-PCR results were analyzed by capillary electrophoresis in a QIAxcel system (Qiagen, Toronto, Ontario) in the case of the synthetic DNA constructs, or by gel electrophoresis on a 1% agarose gel and visualized with SYBR® safe DNA gel stain (Invitrogen Life Technologies) by Mr. Mathew Fisher at the NCFAD in the case of the viral RNA samples.

2.3.5 Electronic microarrayfor testing viral RNA

The microarray detection and serotyping experiments were performed in the NanoChip 400 electronic microarray platform (Nexogen). Viral RNA samples and probes were prepared in a similar fashion than in subsubsection 2.3.2.2 (NanoChip400 electronic microarray). The number of probes for each serotype that were used for validation with viral RNA can be found in Table 2.8.

Table 2.8: Number of probes per serotype and probe category used in NanoChip400 electronic microarray experiments with viral RNA.

Serotype	Number of probes
Pan-FMDV	3
A	12
O	19
C	8
Asia1	6
SAT1	9
SAT2	13
SAT3	6
Negative control	1

2.3.6 Data analysis and visualization

Raw fluorescence intensity measurements were exported as Microsoft Excel files (*.xls) from the Data Viewer proprietary software application of the NanoChip 400 instrument (Nexogen). The experimental data were processed using the R statistical computing language (R Core Team, 2016). NanoChip400 data was reshaped so that all observations were in rows and all variables were in columns, a format that is also called tidy data (Wickham, 2014, 2016). The positive-to-negative ratio (PN) was calculated for each probe by dividing the mean FI value of the viral RNA samples by the FI of the Non-template control (NTC) (Lung et al., 2011, 2012, 2013).

2.3.7 Empirical determination of optimal temperature and PN ratio

The trade-off between sensitivity and specificity at different temperatures and PN ratio cutoffs was analyzed for the electronic microarray results. First, the distributions of PN ratios at all temperatures were analyzed to identify those temperatures where negative PN ratios were observed. Negative PN ratios were considered artifacts of the NanoChip400 and a sign of lack of hybridization between probes and amplicons. Therefore, data recorded at 56 °C, 58 °C, and 60 °C were removed from the analysis because they contained negative PN ratios. As a result, further sensitivity and specificity analyses were performed only on data from temperatures from 24 °C to 54 °C. Receiver operating characteristic (ROC) curves (Lusted, 1971; Zweig and Campbell, 1993; Greiner et al., 2000) were generated after computing binary scores that assigned a value of 1 if the sample was both detected and subtyped at a given PN ratio cutoff, and a value of 0 if the sample was either not detected or not subtyped. ROC curves were plotted with the `plotROC` package of the R statistical computing language (Sachs, 2016). The area under the curve (AUC) was calculated for the ROC curves of all temperatures from 24 °C to 54 °C.

Heatmaps PN ratio data were generated using a continuous gradient of black and red colour using the TreeView software (Eisen et al., 1998), and the impact of changing the

positivity cutoff in the heatmap was explored with an interactive application that was developed using the `shiny` package of R (Chang et al., 2016) that uses the `rbokeh` visualization package of R (Hafen and Continuum Analytics, Inc., 2016) (Figure 2.10). Probe-sample hybridizations were considered positive if the positive-to-negative ratio was equal or greater than an empirically determined threshold analyzed with the ROC analysis and visualized with the aforementioned applications. The FI results at all the washing temperatures were analyzed, and the heatmap visualizations of PN ratio data at 54 °C are reported here.

2.4 Results

2.4.1 Assay design and *in silico* evaluation

2.4.1.1 FMDV primer coverage

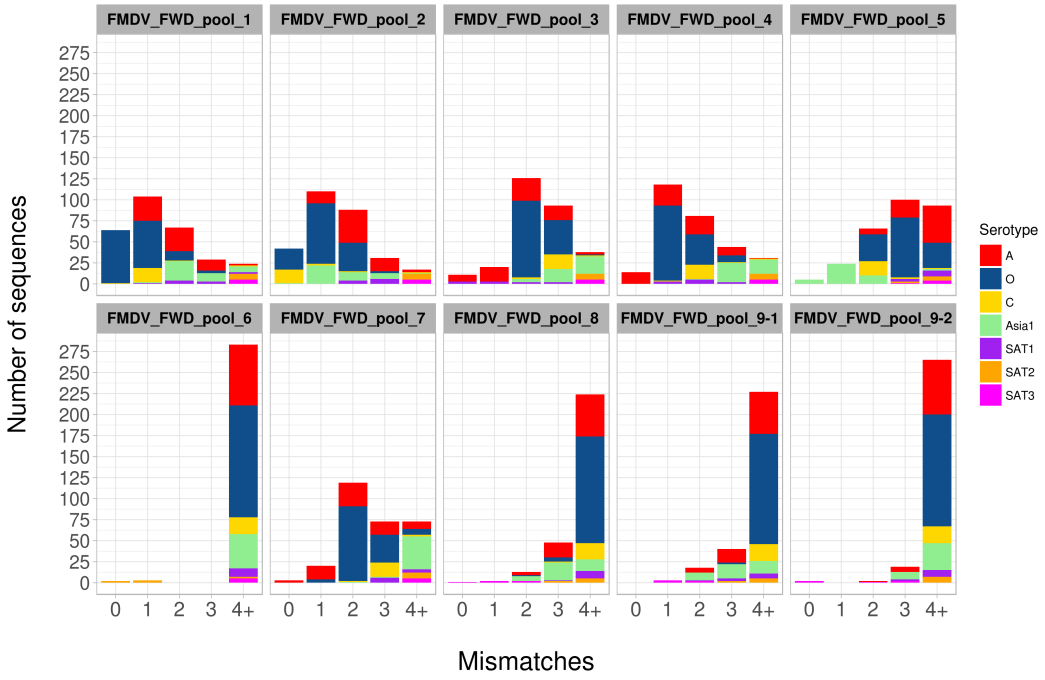
Forward primer pool Although the RT-PCR reaction developed by Erickson et al. (2017) included primers for the detection of seven viruses of swine, since the focus of this thesis is on the characterization of FMDV serotypes, only the coverage of FMDV sequences by FMDV primers was analyzed in this work. In combination, the nine primers of the forward FMD primer pool matched 286 out of 288 FMDV complete genome sequences (99.3%) with 0, 1, 2 or 3 mismatches (Figure 2.2a). There were only two sequences that were not covered by the forward primers. The first of those strains was serotype Asia1 isolate IND 21-89 (GenBank accession number DQ989316.1) that was originally obtained from a buffalo host in India in 1989 (Mohapatra et al., 2008). The reason the forward primers do not match *in silico* that region of that particular FMDV genome is that the sequence contains a segment of 582 Ns (undefined nucleotides) which includes the forward primer region. The second strain that was missed by the forward primer pool was the serotype C isolate C-S8p260d999 (GenBank Accession number DQ409184.1). This particular isolate was used in a study where FMDV strains with deletions in their genomes were produced to better understand viral evolution (García-Arriaza et al., 2006). This genome is much shorter than the typical FMDV complete genome as it only has a length of $7,116$ base pairs, and its

sequence has gaps in the multiple sequence alignment in the binding region of the forward primers. The design of this primer pool is reflected by the patterns of database coverage of the 9 primers of the pool were very different. Primers complement each other since some primers cover the sequences of FMDV strains that other primers do not cover, and there is some serotype-specificity for some of those oligonucleotides (Figure 2.1b). This design for strategically capturing the maximum number of strains possible with relatively few molecules is further supported by the patterns of mismatch distributions within and between the primers (Figure 2.2a).

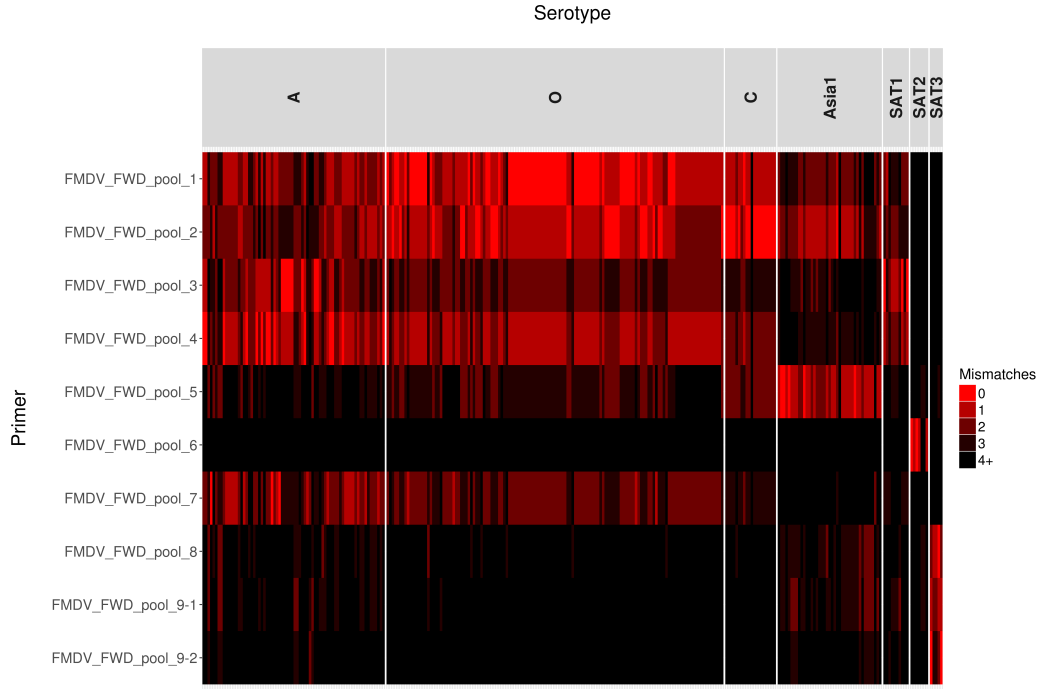
Reverse primer The reverse primer contained 2 degenerate bases which result in 4 sequence variants. Considering all four primer variants, the reverse primer covered 287 out of 288 sequences with 0, 1, 2 or 3 mismatches (Figure 2.3). The only whole genome sequence that the reverse primer did not cover with the above criteria was isolate Asia1-2 Israel 3-63 iso6 (GenBank Accession AY593796.1). This Israeli virus was isolated in 1963 and the sequence was submitted to NCBI by the African Swine Fever Research Unit of the Plum Island Animal Disease Center (Carrillo et al., 2005). As it was the case with the forward primer pool, the design of the reverse primer was intended to cover the maximum number of sequence variants as possible with few modifications. The patterns of database coverage reflected this design since those patterns were considerably different between the 4 variants of the degenerate reverse primer. Another way in which the design of the reverse primer reflects the goal to capture the diversity of FMDV sequences is that the reverse primer variants complement the coverage of each other to some extent (Figure 2.3b).

2.4.2 Primer specificity

In the study by Erickson et al. (2017), the FMDV primers did not amplify 10 clinical oral negatives nor did they amplify 12 non-target viruses and bacteria. In the present study, the specificity of the swine high consequence assay multiplex PCR primers was tested *in silico* for their potential cross-reactivity and amplification of picornaviruses other than FMDV,

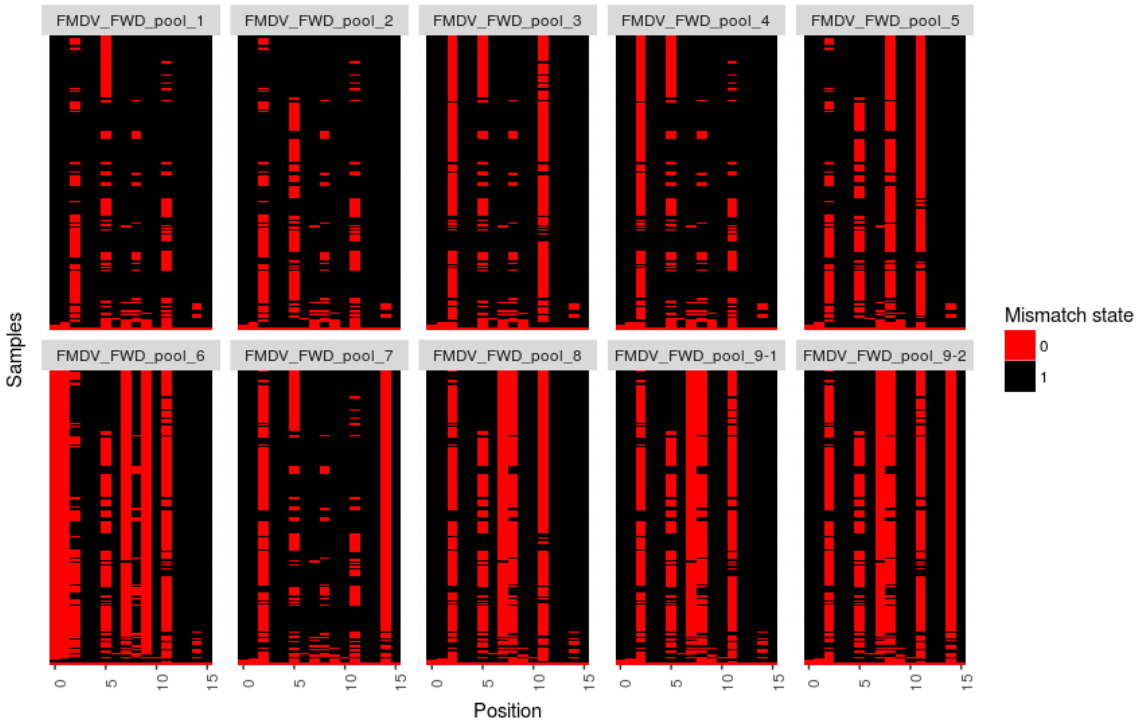


(a) Differences in the number of strains with 0, 1, 2, 3 and 4 or more mismatches between the 9 FMD primers of the forward primer pool.

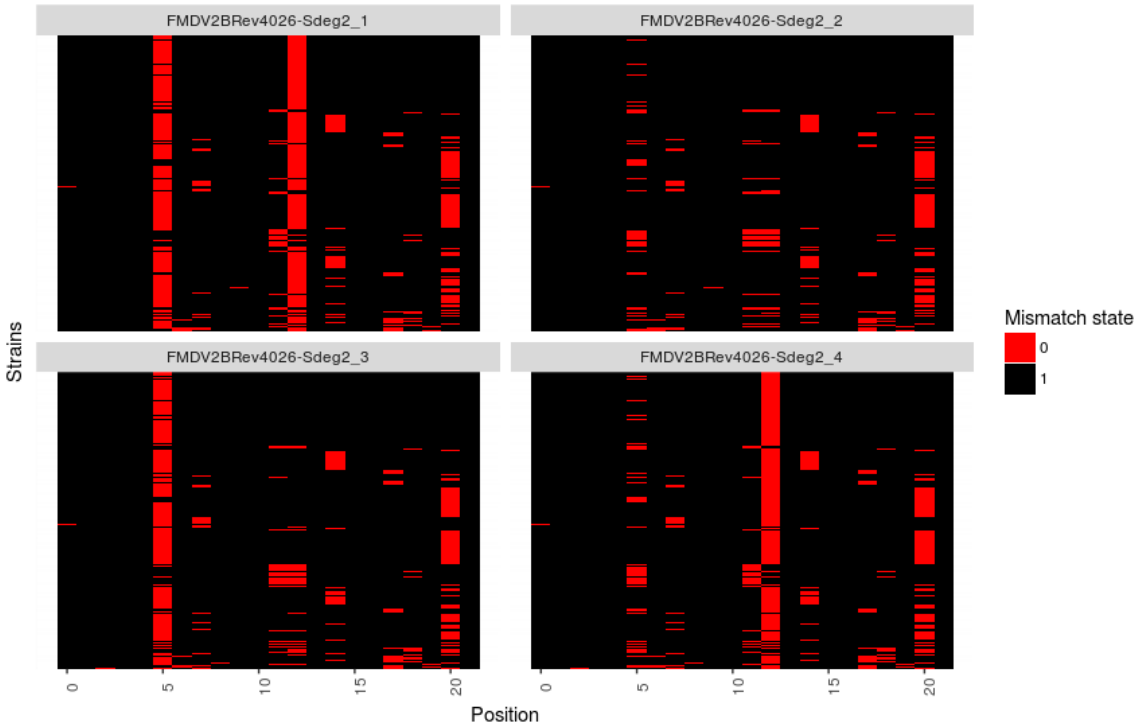


(b) Heatmap depicting coverage of the FMDV whole genome sequence database by each primer of the forward primer pool. Results are depicted using a red-black scale. The more intense the red colour, the fewer the mismatches. Strains are divided by FMDV serotype.

Figure 2.1: Coverage of the forward primer pool represented as (a) the distribution of number of mismatches, and (b)the number of mismatches per serotype. Primers 9-1 and 9-2 are two variants of the same degenerate primer.

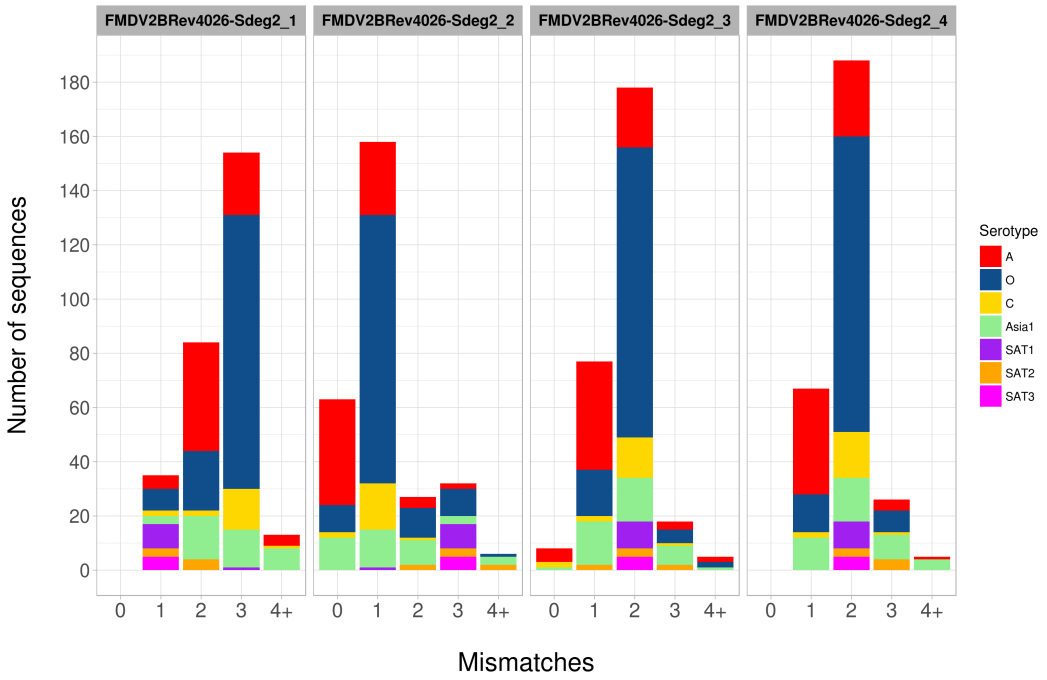


(a) Distribution of mismatches along the nucleotide positions of the forward primer pool.

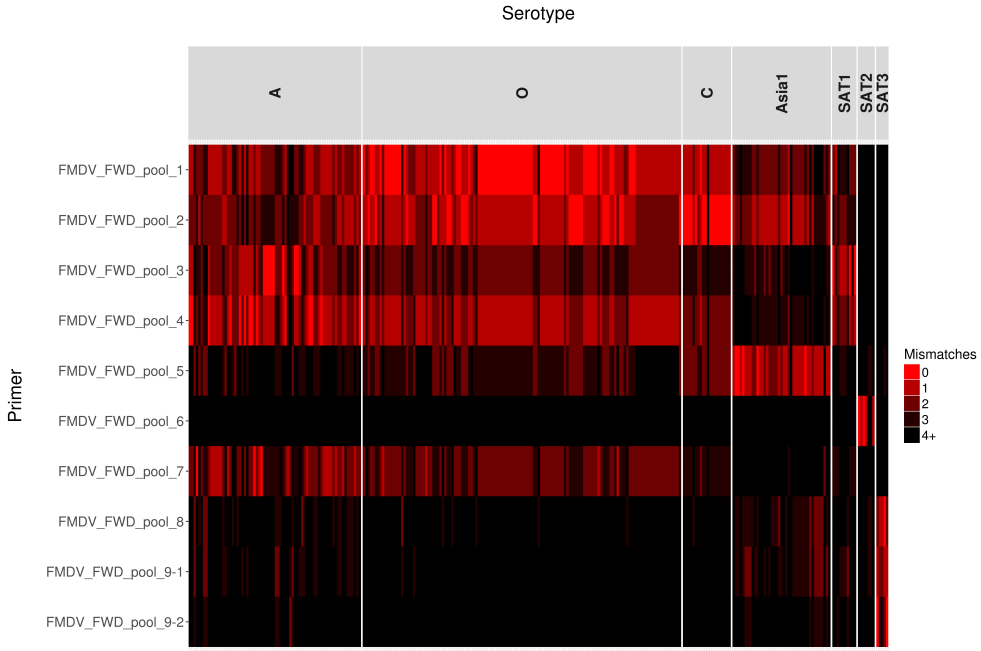


(b) Distribution of mismatches along the nucleotide positions of the reverse primer.

Figure 2.2: Distribution of mismatches along the nucleotide positions of the forward primer pool and the reverse primer.



(a) Differences in the number of strains covered with 0, 1, 2, 3, 4, 5 or 6 mismatches between the 4 variants of the reverse primer FMDV2BRev4026-Sdeg2.



(b) Heatmap depicting the FMDV strains covered by each variant of the reverse primer FMDV2BRev4026-Sdeg2. Strains are divided by FMDV serotype. The Y axis scales were set independently for each serotype.

Figure 2.3: Coverage of the reverse primer pool represented as the distribution of number of mismatches and the number of mismatches per serotype.

and for potential cross-reactivity with hosts of FMDV. The results of those evaluations are presented below:

2.4.2.1 FMDV primer pairs vs. non-FMDV picornaviruses

There was cross-reactivity predicted *in silico* of FMDV primer pairs vs. non-FMDV picornaviruses. This is expected since these primers target a highly conserved region of the FMDV genome and of picornaviruses.

2.4.2.2 All specific primer pairs vs. non-FMDV picornaviruses

As expected, the SVDV-specific primers matched SVDV sequences, although this happened at thresholds of 4, 5 and 6 mismatches. There were zero matches of Besides those matches against SVDV, there were zero non-FMDV picornaviruses matched by the other specific primer pairs of the SHC assay reported with e-PCR.

2.4.2.3 Non-specific primer pairs vs. FMDV and non-FMDV picornaviruses

Specificity tests were performed with re-PCR for a range from 1 to 10 mismatches to search if non-specific primer pairs could match FMDV sequences. No primer pair hits were found between all the non-specific primer pairs and FMDV sequences at any of the primer mismatch thresholds with re-PCR analyses. In the case of the non-specific combinations of primers vs. the non-FMDV picornaviruses, no primer pairs reported matches at thresholds of 0, 1, 2 and 3 mismatches. There was only one primer pair match at a threshold of 4 mismatches. The predicted match was a 239 bp amplicon from an Enterovirus B that is found in non-human primates (GenBank accession number: KF648613.1). The matching primer pair was composed by variant #6 of the FMDV forward primer (FMD-VP3com980(4)_6) and the SVDV forward primer (SVDVCV-3C-17a-F-(5875bp)_1). However, there is a mismatch in the terminal 3' base of that match, and therefore, amplification would be unlikely.

In addition, the simulation of polymerase chain reaction conditions with ThermoBLAST resulted in the prediction of six possible amplicons from four unique non-FMDV

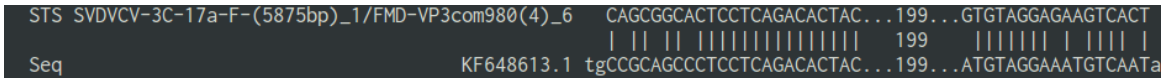


Figure 2.4: Non-FMDV picornavirus hit found by querying non-specific primer pairs with e-PCR. This hit resulted from a query of all the non-specific primer pairs with tolerance of 4 mismatches per primer.

picornaviruses. Those predicted amplicons resulted from interactions between primers that were not designed for the same target virus (Table 2.9). Further analysis of those results shows that three of those viruses (Human coxsackievirus B2 isolate KOR 04-279, Salivirus FHB and Hepatitis C virus full-length replicon pFGR-JFH1 RNA, complete sequence) infect humans or non-human primates. The only virus that infects a host of FMDV was the unclassified kobuvirus strain Kagoshima-2-24-KoV/2015/JPN (GenBank accession number NC_027918.1). This Japanese virus was found in diarrhoeic calves and it was originally identified as a novel kobu-like virus by phylogenetic analysis (Otomaru et al., 2016).

The amplification of this predicted 2,781 bp amplicon is unlikely because the hybridization of the forward PRRS_Mtrx_F2_1 primer is weak, although there is complete matching between the last twelve 3' bases of the primer and the kobuvirus (Figure 2.5). In addition, the visualization of the reverse primer hybridization generated with ThermoBLAST showed that the 3' base of the reverse primer does not match the target (Figure 2.6). Moreover, the primer pair detected in this *in silico* reaction has two forward primers (PRRS_Mtrx_F2_1 and FMDV_FWD_pool_2). Since none of these primers have the mutseq tag, even if the non-specific target was amplified, it would not be detected by the electronic microarray.

2.4.2.4 All non-specific primer pairs vs. host genomes

Pig genome No hits were reported between the non-specific primer pairs and the swine genome (Sscrofa10.2 assembly) at 0, 1, 2 and 3 mismatches with e-PCR. Results with a tolerance of 4 mismatches per primer resulted in the predicted amplification of a 313 bp amplicon derived from primer # 8 of the FMD forward primer pool (FMD-VP3com980(4)_8)

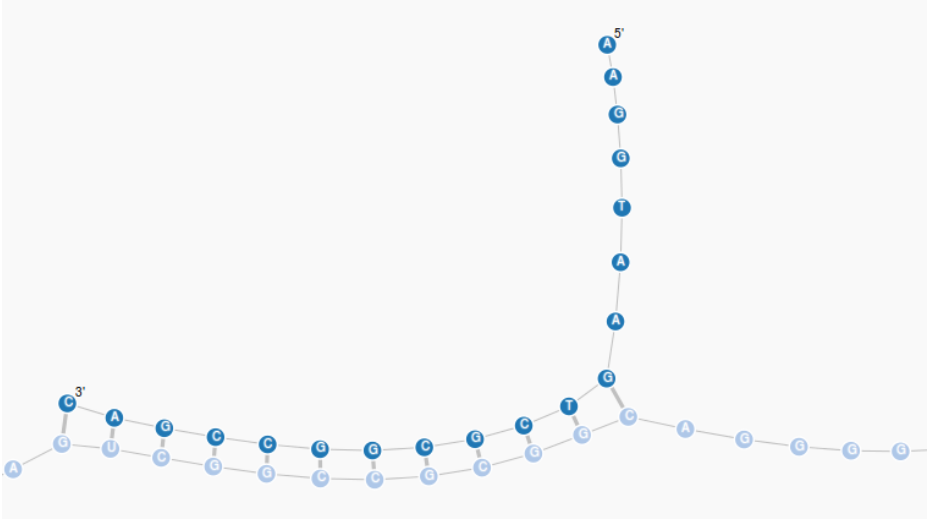


Figure 2.5: ThermoBLAST visualization of the hybridization of primer PRRS_Mtrx_F2_1 vs. cattle kobuvirus Kagoshima-2-24-KoV/2015/JPN (NC_027918). The primer sequence is represented in dark blue (top), while the kobuvirus genome sequence is represented in light blue (bottom).

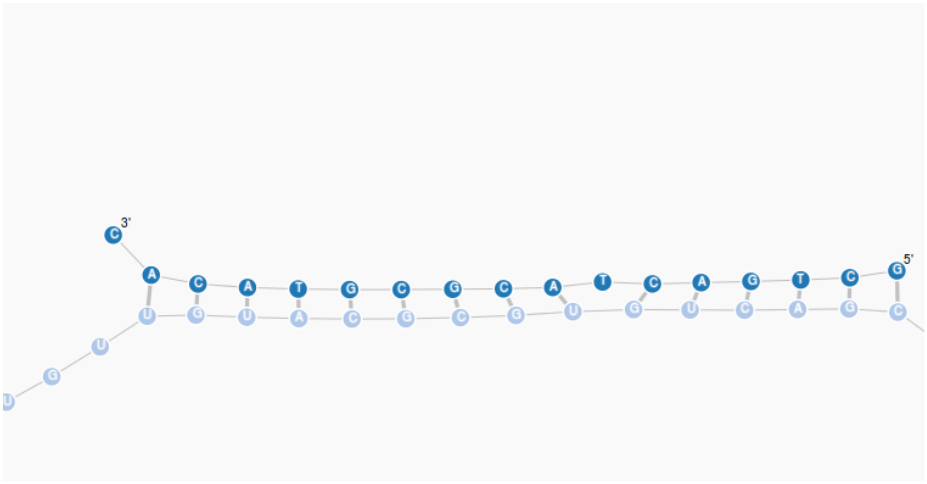


Figure 2.6: ThermoBLAST visualization of the the predicted hybridization of primer FMDV_FWD_pool_2 vs. cattle kobuvirus Kagoshima-2-24-KoV/2015/JPN (NC_027918). The primer was identified as a potential reverse primer of the predicted amplification of the kobuvirus.The primer sequence is represented in dark blue (top), while the kobuvirus genome sequence is represented in light blue (bottom).

Table 2.9: ThermoBLAST results of predicted non-specific primers vs. non-FMDV picornaviruses. Non-specific pairs are those between oligonucleotides that were designed to target different viruses.

Forward Primer	Forward Tm (°C)	Reverse Primer	Reverse Tm (°C)	Predicted Amplicon Size	Predicted Target (Accession)
PRRS_Mtrx_F2_1	59.80	FMDV_FWD_pool_2	50.40	2781	Cattle kobuvirus (NC_027918)
KBH12_6E2ExtREV_1	54.70	PRRS_Mtrx_F2_1	51.40	168	Salivirus (NC_025114)
FMDV_2BRev4026_Sdeg2_3	50.70	VESVSM_2_F_(5101bp)	55.00	1471	Human coxsackievirus B2 (EF174469)
FMDV_2BRev4026_Sdeg2_2	50.70	VESVSM_2_F_(5101bp)	55.00	1471	Human coxsackievirus B2 (EF174469)
KBH12_6E2ExtREV_2	50.40	PRRS_Mtrx_F2_1	51.40	168	Salivirus (NC_025114)
PRRS_Mtrx_R2_2	51.00	PRRS_Mtrx_F2_1	50.20	1259	Hepatitis C virus (AB237837)

and variant #16 of the CSFV primer (KBH12-5E1ExtFWD_16) (Figure 2.7).

Cow genome No hits were reported between the non-specific primer pairs and the bovine genome (Bos_tarus_UMD_3.1.1 assembly) at 0, 1 ,2 and 3 mismatches. Results with a tolerance of 4 mismatches per primer resulted in the predicted amplification of a 148 bp amplicon of chromosome 6 of the bovine genome. This *in silico* non-specific product was generated with primer # 6 of the FMD forward primer pool (FMD-VP3com980(4)_6) and the second variant of the CSFV reverse primer (KBH12-6E2ExtREV_2) (Figure 2.7).

Sheep genome No hits were reported between the non-specific primer pairs and the sheep genome (Ovis_aries_1.0 assembly) at 0, 1, 2, and 3 mismatches. Results with a tolerance of 4 mismatches per primer resulted in the prediction of the amplification of a 262 bp amplicon derived from primer # 5 of the FMD forward primer pool (FMD-VP3com980(4)_8) and the second variant of the CSFV reverse primer (KBH12-6E2ExtREV_2) (Figure 2.7).

2.4.3 Probe coverage

The pan-FMDV capture probes designed for this study provided high *in silico* coverage of the FMDV target sequence database. Twelve highly conserved pan-FMDV capture probes were evaluated against 335 sequences that contained the highly conserved region for which these probes were designed. After expansion of the degenerate bases, the total number of variants of the pan-FMDV probes was 251, and probes covered the majority of

```
STS FMD-VP3com980(4)_8/KBH12-5E1ExtFWD_16   AGGCCAGACTGGTGGCCTTATGA...313...GTGTATGCAAAGTCTGC
                                            ||| | ||| |||| ||||||||   313   |||||||| ||| | |
Seq                          NC_010450.3 aaAGGACTGACAGGTGACCTTATGA...313...GTGTATGCTAAGACAGGtt
```

(a) Predicted 313 bp non-specific product of swine target amplification. *In silico* evaluation was performed with e-PCR at a threshold of 4 mismatches.

```
STS FMD-VP3com980(4)_6/KBH12-6E2ExtREV_2   AGTGACTTCTCCTACAC...112...TGAGAACAGAAG-TGGTAAA
                                           |||  ||||||||||||   112   ||||||||| || | |||||
Seq                GK000006.2|GK000006.2 ccAGT--CTTCTCCTACAC...112...TGAGAACAG-AGCT-GTAAAac
```

(b) Predicted 112 bp non-specific product of bovine genome (*Bos taurus*) target amplification.

```
STS FMD-VP3com980(4)_5/KBH12-6E2ExtREV_2   TTTACCA-CTTCTGTTCTCA...226...GTGTAAGCATAGTCAGC
                                           ||||  | || |||||||||   226   ||||||||| ||  | |
Seq                          CM000885.1 gaTTTAAGAACT-CTGTTCTCA...226...GTGTAAGCACAGAAATCtg
```

(c) Predicted 226 bp non-specific product of sheep genome (*Ovis aries*) target amplification.

Figure 2.7: Non-specific products generated with non-specific primer pairs vs. host genomes: (a) *Sus scrofa*, (b) *Bos taurus*, and (c) *Ovis aries*. *In silico* evaluations for the three species was performed with e-PCR at a threshold of 4 mismatches. The number in the middle of the graphic is the number of base pairs between primer sequences in the host genome.

the sequences in the FMDV common region database Figure 2.8.

Table 2.10: *In silico* coverage of pan-FMDV and serotype-specific probes against the database of partial and complete FMDV sequences (n=2,888).

Serotype	Number of probes	% coverage of database (proportion)
Pan-FMDV	12	98.2 (329/335)
A	12	99.51 (615/618)
O	28	97.45 (1186/1217)
C	9	100.00 (131/131)
Asia1	6	99.16 (355/358)
SAT1	8	99.02 (202/204)
SAT2	34	97.48 (271/278)
SAT3	6	100.00 (67/67)

2.4.4 Probe specificity

2.4.4.1 Pan-FMDV probes

The pan-FMDV probes were evaluated for specificity to FMDV *in silico* by searching their sequences against the sequences of non-FMDV picornaviruses. Searches were performed with reverse-electronic PCR (re-PCR) and matches with Equine rhinovirus 3 strain P313/75 (GenBank accession: AF361253.1) were reported when performing the searches

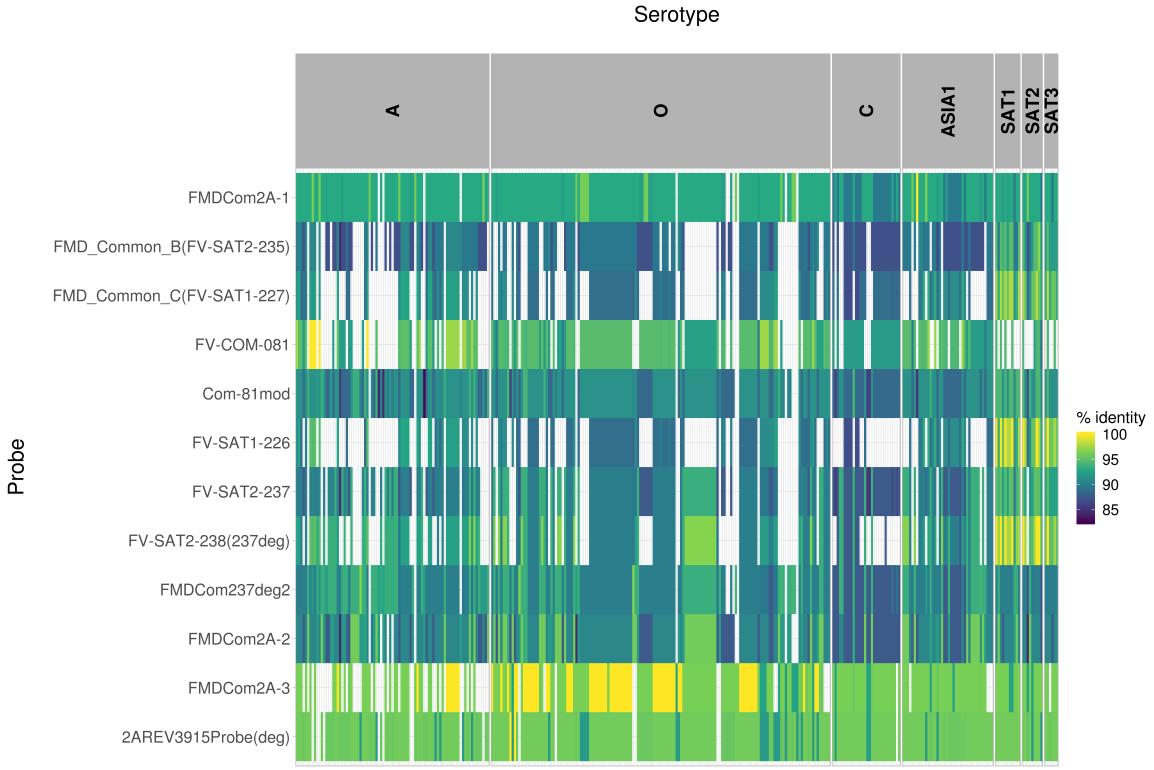


Figure 2.8: Heatmap of the *in silico* coverage of each pan-FMDV probe vs. the database of sequences with the common probe region (n=335). Probe sequences were expanded (i.e. degenerate bases were resolved) and searched with BLAST against the common probe region database. Matches were filtered using criteria such as 80% identity and 75 % query coverage. The color scale represents the % identity of the BLAST searches of the combined results of each oligonucleotide vs. the sequences in the database. Results are ordered by FMDV serotypes.

with 2 gaps. However, it is predicted that the risk of false positive reactions of the pan-FMDV probes with those equine viruses *in vitro* is very low since no amplification was predicted for that equine picornavirus according to the *in silico* analysis of the FMDV primers and the SHC primers.

2.4.4.2 Serotype-specific probes

The serotype-specific probes were tested *in silico* for specificity by searching their sequences against databases of the six heterologous serotypes combined. The serotype-specific probes showed high specificity *in silico* to each of the serotypes, although some cross-reactivities were detected.

2.4.5 Amplification of synthetic constructs and viral RNA

The singleplex FMD RT-PCR assay amplified all of the 19 synthetic constructs of FMDV strains from different geographical locations and years. There were no non-specific products generated by the singleplex RT-PCR assay that was used to amplify the synthetic constructs that were screened with the capture probes. The twenty-three viral RNA samples were successfully amplified with RT-PCR using the SHC multiplex primers with optimized reaction conditions work performed by Mr. Mathew Fisher, NCFAD) (Figure 2.14). The RT-PCR did not generate non-specific products.

2.4.6 Detection and serotyping on electronic microarray

A total of 12 pan-FMDV probes were tested with amplicons derived from the 19 synthetic FMDV constructs. All the synthetic constructs reacted with the pan-FMDV probes. As it can be seen in Figure 2.9, pan-FMDV probes A, B, and C detected all the synthetic constructs. However, after *in silico* analysis, pan-FMDV probes A, D, and E were chosen for the electronic microarray with viral RNA, due to the higher coverage of these probes in combination. This decision was made in the event strains that were not originally represented in the panel of synthetic constructs.

In addition to the pan-FMDV probes, a total of 93 serotype-specific capture probes and 1 non-specific capture probe were tested on the electronic microarray platform. Serotype-specific probes reacted with and correctly typed the 19 synthetic construct amplicons (Figure 2.9), although not all the serotype-specific probes yielded hybridizations with a PN ratio above the threshold of 2.0.

In this preliminary testing, six serotype-specific probes showed cross-reactivity with other serotypes. These probes were: FMD A6 (cross-reacted with sample C3 Resende), probe O-149 (cross-reacted with sample sA-4), probe FV-O-166 (cross-reacted with sample sA-4), probe FV-O-169 (sA-3, sA-5, C3 Resende). Probes FV-Asia-206 (cross-reacted with samples sA-3 and sA-4) and probe FV-Asia-040 (cross-reacted with samples sA-4, sO3 and Sat3 Zim). These cross-reactivities were not predicted by applying the stringent criteria to the BLAST results. However, the results were predicted with e-PCR.

The 23 RT-PCR products generated with viral RNA were successfully detected with the pan-FMDV probes (Figure 2.14), and the twenty-three strains were correctly typed with the electronic microarray assay at a temperature of 54 °C using a PN ratio of 2 as a cutoff for positive results (Figure 2.14).

2.4.6.1 Determination of the optimal temperature and PN ratio

In this study, a touch-up protocol in which temperature was raised by increments of 2 °C was used. This touch-up protocol was different than the touch-down protocol used in the electronic microarray assay by Erickson et al. (2017) in which temperature was decreased and PN ratios were determined from the final temperature wash. In this study, the results from all the experiments at all the temperatures for which results were obtained were evaluated. An initial visual inspection of the distribution of PN ratios showed that at 56 °C, 58 °C and 60 °C there were many results that had negative PN ratios. Therefore, those results were discarded from the analyses and the distributions of PN were compared for all temperatures from 24 °C to 54 °C. Although all the temperatures had PN ratio distributions

that were skewed to the left due to the majority of results having PN ratios that spanned a narrow range, PN distributions varied in the number of results with PN ratios that were higher than 5, with the most extreme case being the data at 54 °C (Figure 2.11), in which signals (amplicon-probe fluorescence intensity) were maximized in relation to the noise (non-template control fluorescence intensity). Furthermore, the sensitivity and specificity trade-offs for all PN ratios at all the temperatures were evaluated by generating ROC curves for the PN ratio results of all the computed serotyping PN results at different temperatures. The PN ratios at 54 °C had the largest area under the curve. (Figure 2.12. Visual inspection of the ROC at 54 °C showed that a PN ratio between 2 and 2.25 is likely to maximize detection and subtyping at that temperature (Figure 2.13). In this work, a PN ratio of 2 was chosen for the visualization of microarrays results because more weight was assigned to sensitivity than specificity.

2.5 Discussion

2.5.1 Comparison of testing synthetic constructs and viral RNA

The simultaneous detection and serotyping assay was developed in the NanoChip400 electronic microarray in two stages: testing with 19 FMDV synthetic constructs for preliminary capture probe screening, and testing with 23 viral RNA samples for validation. Both testing panels represented the seven serotypes of FMDV and, to some degree, the genetic diversity of the virus. One negative control sample was tested in each assay development stage. The assay was developed and screened using synthetic constructs in the biosecurity level 2 facility at the Lethbridge Laboratory of the National Centres for Animal Disease (NCAD) of the CFIA. Synthetic constructs were designed and tested as an alternative to using FMDV viral RNA for prototyping the assay when using the real virus in the biosafety containment level 2 facilities of the Lethbridge Laboratory was not a possibility due to biosecurity regulations and policies. Testing virus detection assays with synthetic constructs is a strategy that has been used before in other studies (Banér et al., 2007).

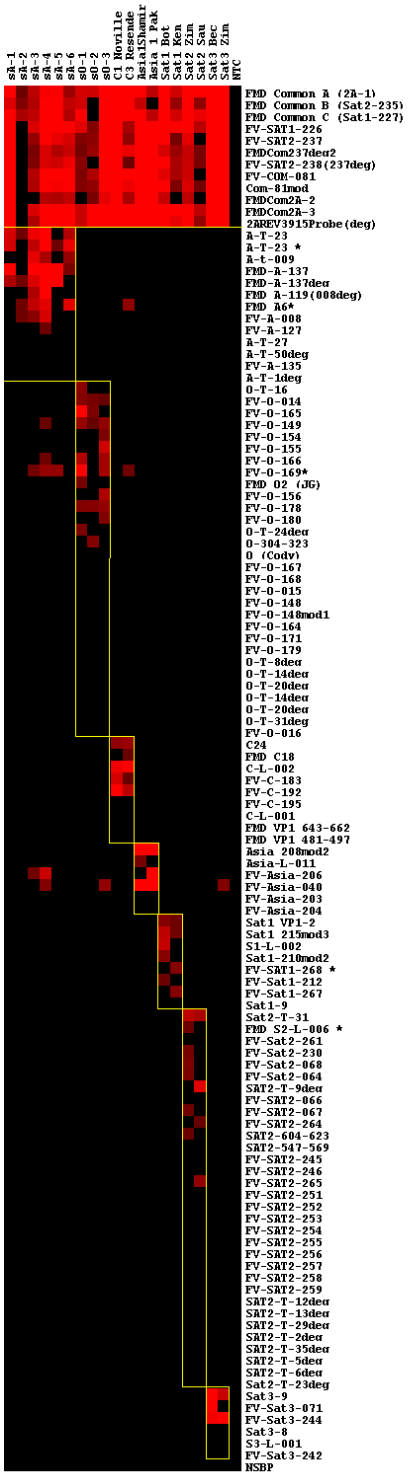


Figure 2.9: Heatmap of NanoChip 400 electronic microarray results of pan-FMDV and serotype-specific capture probes against nineteen synthetic constructs at 54 °C. Rectangles in red are results with a positive-to-negative ratio equal or greater than 2.0. The positive-to-negative ratio results of pan-FMDV and serotype-specific probes are highlighted in yellow rectangles for each of the serotypes, and in the case of pan-FMDV probes for all the serotypes.

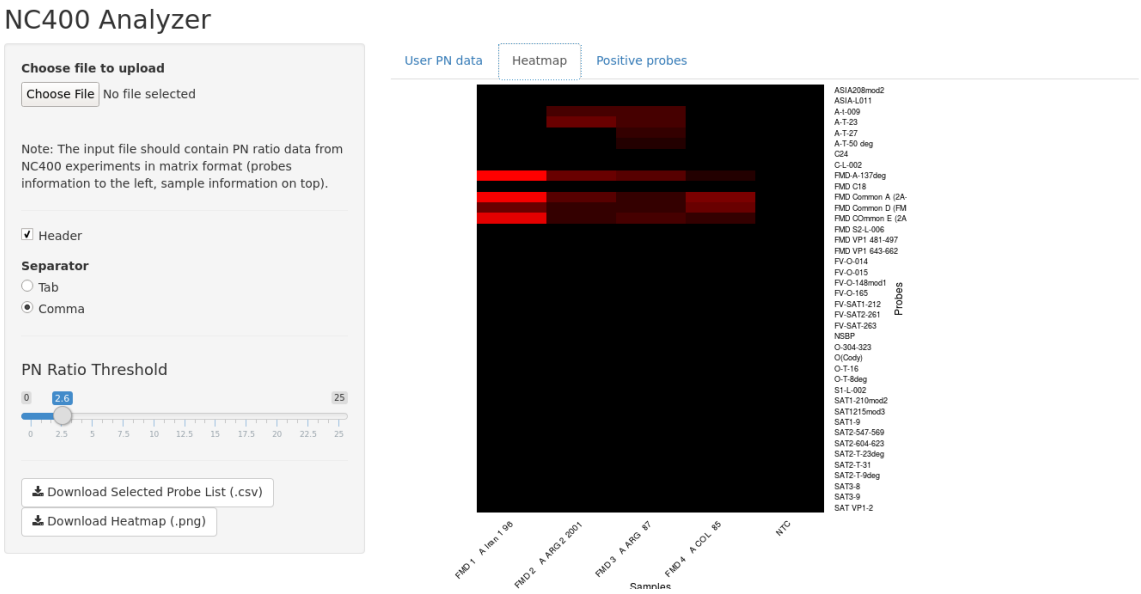


Figure 2.10: Example of the interactive application for the visualization of the results of NanoChip400 experiments.

2.5.2 *In silico* coverage of FMDV databases

A challenge for the development of detection assays for FMDV and other single stranded RNA viruses is the rapid evolution and the high amount of genetic variation present in the population of the virus. Molecular detection assays such as polymerase chain reaction (PCR), real time PCR, and microarrays are typically designed to either target a highly conserved region of the FMDV genome (which is useful for detection) (Ambagala et al., 2016), or to capture the high genetic diversity of the FMDV genome by designing degenerate primers or primer pools, whic was the strategy used in Erickson et al. (2017) and the present study. The evaluation of the forward primer pool and the reverse primers showed that the design of those molecules resulted in coverage of the different strains and serotypes of the sequences of the FMDV sequence database that was used to assess *in silico* sensitivity.

A common challenge with molecular and serological detection assays is that they need to be updated as new strains emerge. In the case of molecular detection assays, two factors are important for maintaining the high sensitivity and specificity of microarray assays for

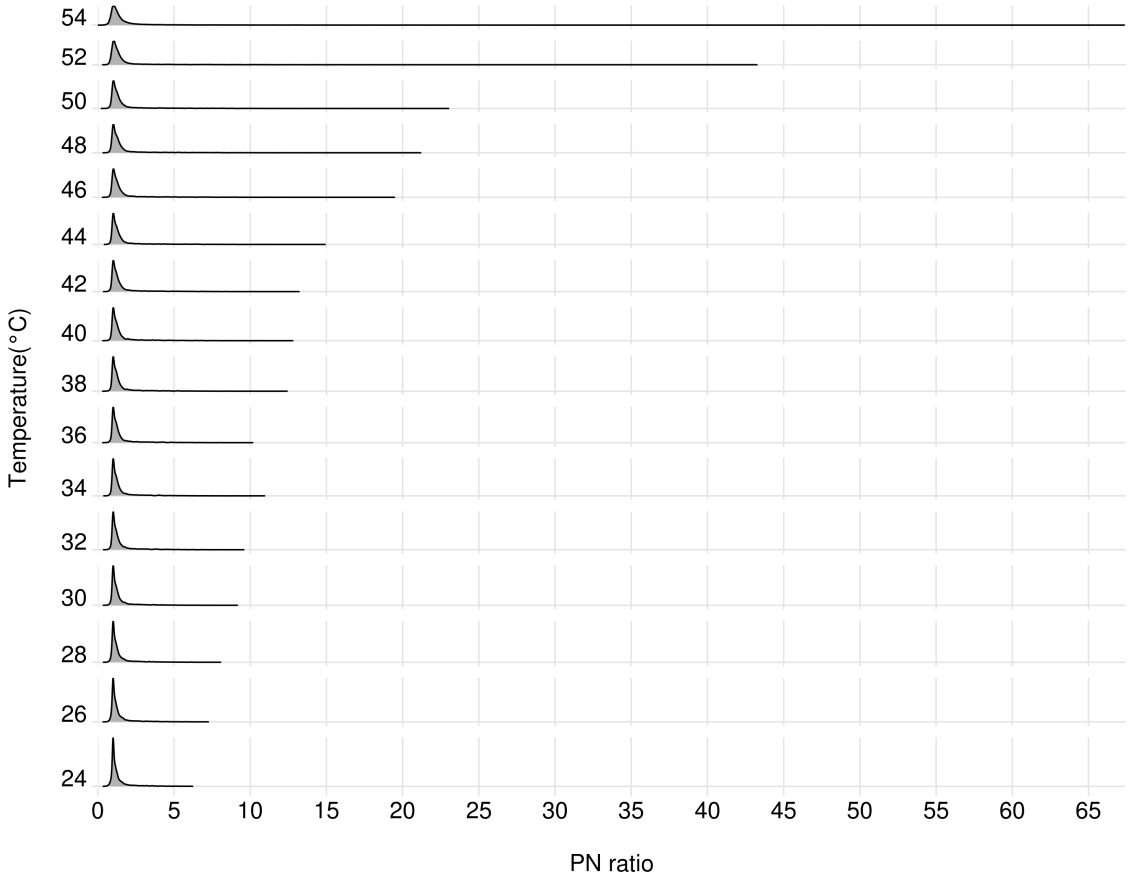


Figure 2.11: Distributions of Positive-to-Negative ratio (PN) ratios for each of the different reporting temperatures of the electronic microarray FMDV detection and serotyping assay. PN ratios represent data from all the NanoChip400 experiments performed with viral RNA at the NCFAD. Distributions with longer tails indicate a wider range of PN ratios, and thus higher signal-to-noise. The distributions were plotted as ridgeline density plots with the `ggridges` package (Wilke, 2017) of the R statistical computing language (R Core Team, 2016).

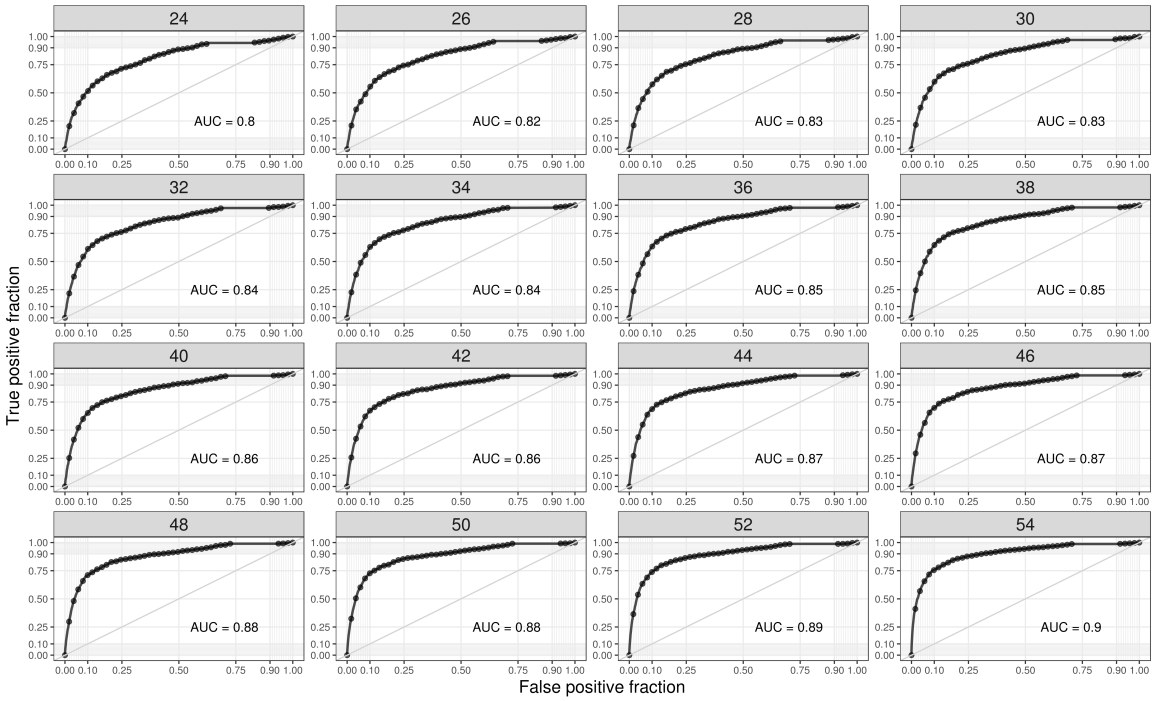


Figure 2.12: ROC curves of FMDV serotyping results with viral RNA. Each point in the ROC curves represents the trade-off between specificity (False positive fraction), and sensitivity (True positive fraction) for positive-to-negative ratio results at each of the temperatures of the NanoChip400 touchdown protocol (excluding 56 °C, 58 °C, and 60 °C due to the presence of negative PN ratios). Labels above the curves represent the temperatures in degrees Celsius. The area under the curve (AUC) was calculated for all the ROC curves.

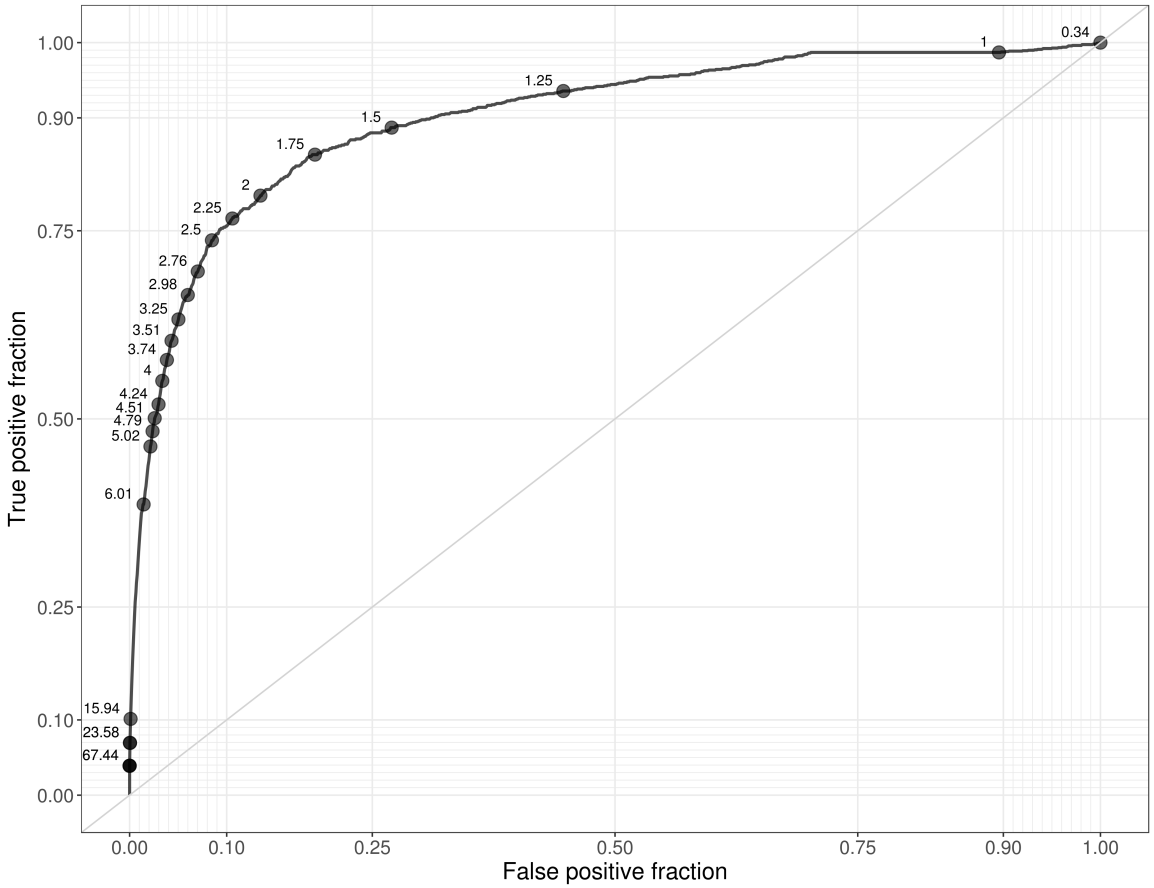


Figure 2.13: ROC curve of serotyping data at 54 °C. Points represent the trade-off between specificity (False positive fraction), and sensitivity (True positive fraction) at different PN ratio cutoffs which were included to illustrate the effect of choosing high or low cutoff values for determining which results are positive and which results are negative. A perfect test has an area under the curve value of 1, while the ROC curve for this figure shows

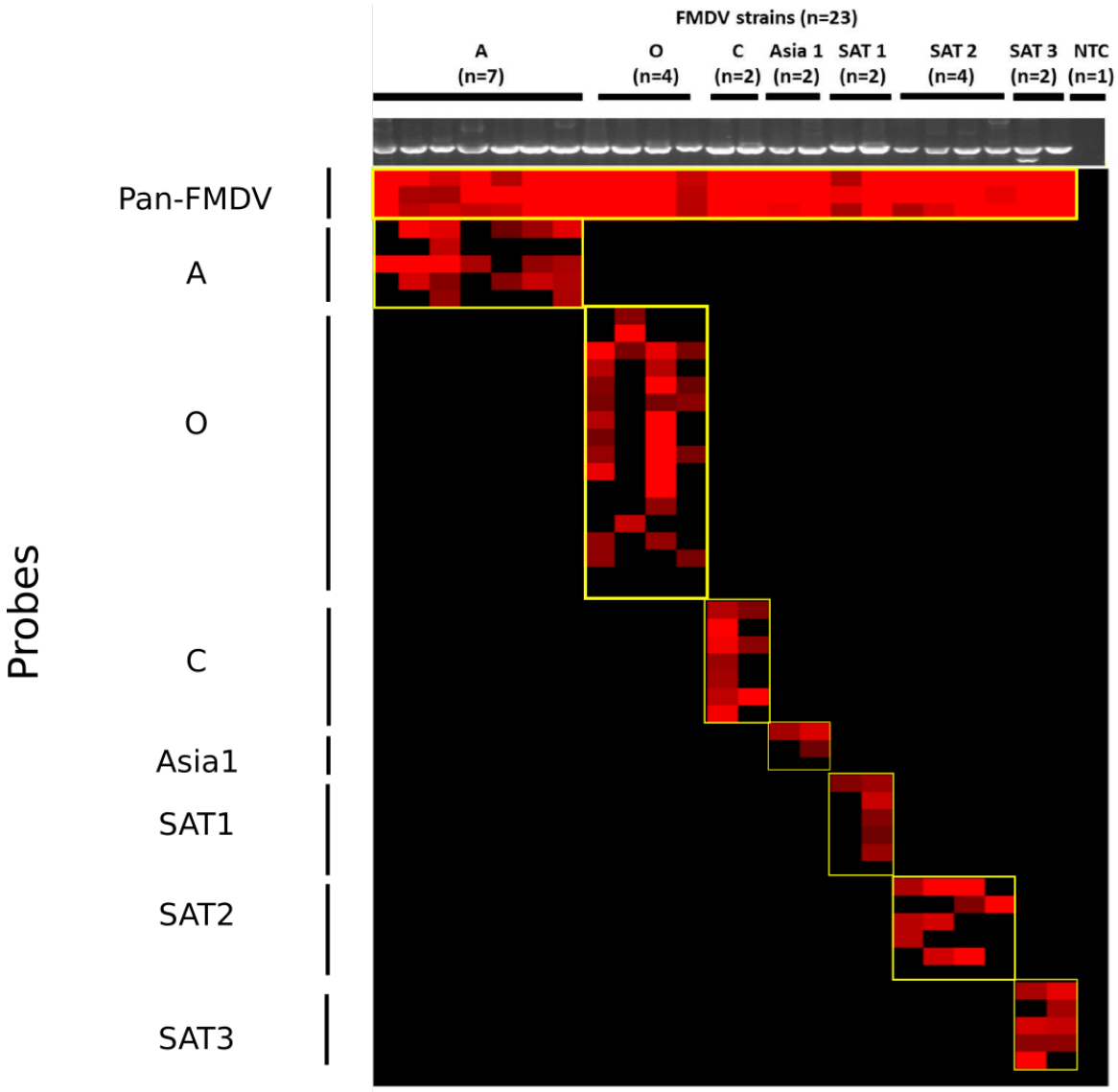


Figure 2.14: Heatmap of PN ratios of electronic microarray results. Results in the red scale represent positive results while results in black represents negative results

the detection and serotyping of viruses:

1. The continuous addition of sequences to publicly available repositories such as GenBank (National Center for Biotechnology Information), the European Nucleotide Archive (European Molecular Biology Laboratory) and the DNA Databank of Japan.

2. The emergence of new variants of the virus in the field that may or may not have their sequences deposited in said public repositories.

In the case of the first factor, there are solutions being developed by research groups around the world to automate the addition of sequences and curation of metadata. Websites such as the Virus Pathogen Resource (ViPR; `viprbrc.org`) can help researchers in human and veterinary virology to have curated, up-to-date databases with unique sequences that can be used for the evaluation and updating of databases. In the case of the second factor, the design of a molecular assay that includes the detection of broader groups can be helpful when addressing the emergence of new variants or subtypes of viruses. Similar approaches have been achieved using a panviral microarray that helped to discover new viruses such as the SARS coronavirus (Wang et al., 2002). However, due to the high genetic variation within the serotypes designing capture probes for testing with the electronic microarray may be more challenging.

2.5.3 Serotype specificity

Although some probes cross-reacted with samples from different serotypes *in vitro*, they can be removed from the assay or used as signatures for more than one serotype. Observed cross-reactivities were not predicted with the stringent criteria for filtering BLAST results for predicting hybridizations. Relaxing the match criteria can help predict the specific hybridizations. Using other alignment algorithms can help to improve the accuracy of the predictions. In this study, the e-PCR tools were better at predicting cross-reactivities than the BLAST-based method.

In this work different programs were evaluated for the *in silico* testing of primers and probes because the evaluation of oligonucleotides is one of the most commmon and important tasks for designing detection and subtyping assays. Examples of tools that are useful for this purpose include NCBI's PrimerBLAST and electronic PCR (e-PCR) and reverse PCR (re-PCR) command line tools (Schuler, 1997; Rotmistrovsky et al., 2004). In this study, *in silico* predictions performed with e-PCR and Thermo-BLAST suggested that it is unlikely that the FMDV PCR assay will cross-react vs. host genomes or any other *Aphthovirus* or *Picornavirus*. A caveat to those predictions is that that work was based on computational simulations, and the results may differ when tested *in vitro*. Also, there may be novel viruses of those genera and families or new variants as a result of evolution that have not been sequenced or deposited in publicly available databases and which might cross-react with the oligonucleotides *in vitro*. Although the *in silico* predictions suggest that it is unlikely that cross-reactivity of any of the pan-FMDV and/or serotyping probes will be observed with other aphtoviruses or picornaviruses, the pan-FMDV probes and serotyping probes were still verified against the other genetically related viruses. Although the e-PCR suite of tools was recently discontinued by NCBI, other commercial tools such as Geneious can perform similar tasks. Recently developed tools such as Oli2Go can help scientists to perform evaluation and design of oligonucleotides for multiplex PCR assays and microarray while performing specificity checks against bacteria, viruses, fungi, invertebrates, plants, protozoa, archaea and environmental samples (Hendling et al., 2018). An example of tools that were conceived for the design of oligonucleotides to detect and genotype highly variable targets is PriMux (Hysom et al., 2012), in which a multiple sequence alignments are not necessary since the search is based on k-mers. like ThermoBLAST (SantaLucia, 2007) can also help to predict results taking in consideration thermodynamic parameters, and not only sequence complementarity.

2.5.4 Evaluation of assay cross-reactivity against other viruses

In silico evidence presented in this work predicts that the FMDV primers in the RT-PCR and the microarray probes of the FMDV serotyping assay are specific to FMDV and that they should not detect other genetically related viruses with sequence information available, nor should they detect host genomes. As mentioned in the introduction chapter of this thesis, the FMDV detection and serotyping assay here reported was conceived as a companion test to a multiplex assay to detect and differentiate seven indigenous and exotic viruses that infect swine. Furthermore, according to the results presented by Erickson et al. (2017) there was no *in vitro* cross-reactivity of the FMDV primers and the FMDV detection probes with any of the other 6 viruses of the SHC assay.

2.5.5 Data analysis tools and methods for electronic microarray results

Previous studies that have used electronic microarray data for the detection and typing of viruses used the positive-to-negative ratio as a the metric for determining which probe-target hybridizations were positive (Lung et al., 2012, 2015, 2016; Erickson et al., 2017). The advantage of the PN ratio is that it is a metric that is easy to interpret which makes it user-friendly. Electronic microarrays are not well represented among the data analysis software that are specifically designed for microarray data, in contrast to platforms such as Affymetrix for which many data analysis tools exist, including many BioConductor R packages. Although the R script written by Gall et al. (2009) was designed to extract data from NanoChip400 experiments, said script was hard to debug and to adapt to data generated in this study, something that was not made easier by the fact that the comments in the code were written in German language only. Unfortunately, no publicly available repository in Github or Bitbucket for the maintenance of said code seems to exist. Another tool for the analysis of microarrays for the detection of multiple pathogen was the DetectIV R package (Watson et al., 2007), which was found useful for other microarray platforms for the detection of pathogens, but the options of visualizations that it included were limited to

bar graphs and custom work would needed to be done for creating heatmap visualizations.

2.5.6 Comparison versus other assay methods

There is strong interest from the livestock industry to implement rapid, sensitive and specific tests for the detection of animal viruses. In particular, the research and development of diagnostic tests for FMDV is growing at a fast pace. Although next generation sequencing (NGS) is a much more efficient method to detect viruses, particularly for sequencing the small whole genome of FMDV, microarrays have a niche in the rapid detection of animal viruses due to automated technologies that can produce rapid tests. In addition, the oligonucleotide probes can be used for *in silico* serotyping of new FMDV strains. This is a technique that has been used in other tools such as the Salmonella *In Silico* Typing Resource (SISTR) (Yoshida et al., 2016) and GeneSippr (Lambert et al., 2015) which characterize serotypes of *Salmonella* and *E. coli*, respectively.

2.5.7 Analytical tools

As much as microarrays provide advantages over other methods due to being able to test multiple samples against many probes simultaneously, the data generated by microarrays is complex and can be challenging to analyze and interpret. Data from microarray experiments such as the ones presented in this work have multiple dimensions. Many data analysis tools for pathogen detection microarrays have been developed for the Affymetrix platform. For example, using solutions based on the R statistical computing language (R Core Team, 2016) or other computing languages like Python can greatly help to accelerate the analysis of microarray data. The DetectIV package for the (Watson et al., 2007) was developed with the specific purpose of processing and visualizing results from pathogen detection microarrays from different platforms, although electronic microarray datasets were not considered. DetectIV is suitable for showing results of individual microarray experiments, but it is difficult to visualize aggregated results from multiple experiments with it, unless the user has a better knowledge of processing multiple datasets by programming in

R using data structures such as lists and uses functions that operate on those lists.

Although in the paper by Gall et al. (2009) the authors wrote a script in the R statistical computing language to automatically process data generated with the NanoChip400 platform, the script is not adapted to other NanoChip400 protocols. Heatmap visualization tools such as TreeView (Eisen et al., 1998) have existed for many years and have been widely used for the analysis and visualization of microarray data. However, custom tools for visualization of electronic microarray datasets and other datasets can allow the users to learn more relevant information about their data if they do not work on some of the more established microarray platforms such as Affymetrix. In the context of microarrays for virus detection and differentiation, the analysis of data using cutoffs to determine positive or suspect results is often necessary, particularly for virus subtyping assays. Unfortunately, investigating multiple positivity thresholds and seeing their impact on the heatmap visualization is a tedious task in TreeView. This underlines the need for developing and using customized analysis tools that are tailored to the specific needs of scientists who use different protocols in the NanoChip400 platform and other pathogen detection technologies. The current explosion in bioinformatic tools and in computer programming skills being acquired by bioinformaticians, but also by biologists and technicians will contribute to the development of tailored solutions for different research and development teams.

2.5.8 Future directions

The detection and serotyping assay here presented was only tested *in vitro* with twenty-three viral RNA samples, and validation with more samples would be required. Another improvement that can be performed on this assay is the determination of the limit of detection for serotyping of FMDV. The recent publication by Erickson et al. (2017) reported that the limit of detection for FMDV was 1,000 copies/μL, and an interesting follow up to that and this study would be to test how the limit of detection impacts the performance of the serotyping assay.

The characterization of the serotypes of FMDV is a very important step in the emergency response to an outbreak. Although the NanoChip400 is a highly automated platform, fully automated technologies have the potential to be even more useful for veterinary diagnostics by streamlining the process from sample to results even more. Besides the examples of fully automated arrays discussed in (subsubsection 1.3.7.7) of this thesis, an Israeli company called Savyon Diagnostics has developed the NanoCHIP Analyzer, a fully automated electronic microarray platform that is the successor of the NanoChip400, and that only requires the user to load the PCR plate, reagents and cartridge (`http://savyondiagnostics.com/product/nanochip-xl-analyzer/`; accessed: 2018-02-06). Examples of assays that have been developed by that company for that platform include an assay for detecting the Zika, dengue virus, and Chikungunya viruses in human serum and plasma (`http://savyondiagnostics.com/product/nanochip-zikvdenvchikv/`: accessed: 2018-02-06).

In countries where diagnostic laboratory services are not readily available, portable devices for detecting and serotyping could be an effective solution for characterization of animal viruses such as FMDV in the field. Preliminary work was performed for the design and development of a prototype instrument to simultaneously detect and characterize the serotype of field strains. Transferring the assay to a portable technology, or to a point-of-care device, would address the increasing need in veterinary diagnostics for the development and implementation of those technologies. For example, a recently published international effort led by the Pirbright Institute resulted in the development of a portable thermocycler for the detection and characterization of four FMDV serotypes commonly found in Eastern Africa (O, A, SAT 1, and 2), although its sensitivity was only of 33/36 samples (Howson et al., 2017). Examples of lateral flow devices that have been used for the simultaneous detection and differentiation of FMDV were discussed in subsubsection 1.3.7.4 of this thesis, and although those tests are based on antibody detection, it is important to highlight the need of portable and user-friendly technologies for rapid answer to critical needs

to protect the nation's livestock.

Besides moving the assay towards complete automation, miniaturization, or portability, other actions can be taken to improve the assay. For example, the oligonucleotide sequences can be reused for new applications. For instance, the oligonucleotides that were tested *in silico* and used for the initial laboratory validation of this assay can still be useful as virus-specific and serotype-specific markers in the analysis of next generation sequencing data. A more in-depth study of the specificity and sensitivity of the serotype-specific probes evaluated in this assay when tested with new sequences will be presented in the next chapter with new FMDV sequences that were not evaluated for this work.

Chapter 3

A comparison of methods for *in silico* characterization of FMDV serotypes

3.1 Abstract

Although molecular nucleic acid detection assays can detect and subtype animal viruses like the foot-and-mouth disease virus (FMDV), the rapid evolution of this and other RNA viruses renders those molecular assays and microarrays ineffective if enough mismatches prevent detectable hybridization between oligonucleotides and the genomes of new field strains. In this study, the performance of different *in silico* approaches for the detection and serotyping of FMDV were evaluated and compared. Publicly available complete FMDV genomes and next generation sequencing (NGS) data were tested *in silico* using various approaches including the evaluation of existing oligonucleotide signatures, the discovery of serotype-specific signatures *de novo*, the classification of sequences using k-mer based taxonomic classifiers, the usage of MinHash-based approaches, and single nucleotide polymorphism (SNP) genotyping. Results of the testing of those approaches with genomes that were not part of the reference datasets are presented. Also, notably, robust single nucleotide polymorphisms were discovered in this study for each of the seven FMDV serotypes. The results from this comparison have the potential to aid regulatory agencies, veterinary diagnostic laboratories, and animal health organizations to implement *in silico* strategies for rapidly characterizing the serotypes of new FMDV isolates and to get high-resolution information from next generation sequencing reads without the need for *de novo* assembly.

3.2 Introduction

3.2.1 Literature Review

Microarrays have higher throughput than many molecular methods for the detection and sub-typing of viruses due to their capacity to include tens, hundreds, thousands, or more oligonucleotide features for the detection of viruses and characterization of viral subtypes. However, microarrays and most molecular detection methods such as PCR may have reduced sensitivity when challenged with new viral strains that have substantial genetic differences from viral sequences used to design those methods. This is because the new strains could have more nucleotide mismatches with the primers and probes, and those mismatches may prevent hybridizations between the oligonucleotides and the targets. Multi-pathogen microarrays and microarrays for the subtyping of viruses can classify samples into different categories because oligonucleotide probes are ideally unique signatures for the detection and classification of the viruses and subtypes of viruses for which they were designed. This makes these sequences potentially useful as *in silico* markers for the differentiation of multiple pathogens or multiple subtypes/serotypes. With the increased use of next generation sequencing (NGS) for the sequencing of whole genomes, the consideration of other targets in the genome beyond amplicon targets used in current assays can improve the classification of samples by their serotype or other metadata categories.

NGS approaches do not typically have those limitations since amplicons of regions of interest and complete genomes can be sequenced for multiple strains simultaneously and products can be readily identified. Although genetic sequencing once had the high cost of $ 95,263,072 per human genome and $ 5,292.39 per megabase in September 2001, the National Human Genome Research Institute (NHGRI) reported that the cost per human genome decreased to $ 1,121 and the cost per raw megabase of DNA sequence has decreased to $ 0.012 on July 2017 (Wetterstrand, 2017). The costs of genome sequencing have outpaced Moore's law (i.e. the doubling of computer power every year) since the introduction of NGS technologies to sequencing centres (Wetterstrand, 2017). One implica-

tion of those advances is that the amount of information that can be analyzed and the scope of the analyses that can be performed has spurred the development of bioinformatic tools which enable researchers to rapidly analyze viral populations and intra-host viral genetic diversity (Orton et al., 2016).

Regulatory laboratories can benefit from the identification of unique genomic signatures or markers for the characterization of pathogen species, subtypes, serotypes, and other metadata categories. Those sequence signatures can facilitate the rapid identification and the understanding of genetic variation and population structure in outbreaks. A notable example of how the concept of universal signatures is applied in eukaryotic genomics is the usage of the cytochrome oxidase I gene as an universal barcode for differentiating species of eukaryotes (also known as the barcode of life) (Hebert et al., 2003). In the case of bacteria, the 16S ribosomal RNA gene and have become targets for the characterization of microbial communities, as this target allows for the characterization of Operational Taxonomic Units (OTU) thanks to its hypervariable regions.

In veterinary diagnostics, it is important that pathogen detection signatures have high *in silico* sensitivity and specificity for the characterization of groups of interest. However, as the number of sequences from field isolates increases, and higher variation is discovered, the sensitivity of the assay for detection and subtyping might decrease. Since many of those oligonucleotides were designed for specific subtypes or subgroups of viruses, it is possible that those sequence signatures can be used for the *in silico* characterization of new sequence data of FMDV and other animal viruses. For example, this idea has been applied to the characterization of avian influenza virus strains in the FluDB website (Zhang et al., 2016) where a curated list of eleven PCR and real-time PCR assays (as of December 2017) presents the evaluation of each oligonucleotide against the continuously updated collection of curated influenza virus sequences of the Influenza Research Database (`www.fludb.org`). Not only does that resource report the number of mismatches of oligonucleotides and new variants of the virus, but it also reports subtle nucleotide variations between strains since

it groups genomes in variants according to their matching patterns against the sequences of the primers and probes of each of the available assays. The idea that oligonucleotide sequences that have been successfully used in assays *in vitro* can be used as markers *in silico* for subtyping of NGS data is an strategy that has been used before, for example, in the characterization of *Salmonella* serotypes in the Salmonella *In Silico* Typing Resource (Yoshida et al., 2016).

Metadata-driven comparisons can help to associate sequence features such as single nucleotide polymorphisms (SNPs) to different metadata categories of pathogens (e.g. epidemiological groups or phenotypes). Tools like GenomeFisher or `feht` (Laing, 2017) can be helpful in finding those associations since those tools can aid in performing statistical analysis of the proportion of sequences in a group that have a specific SNP. Those tools are available for using locally in a Windows system through a graphical user interface (GenomeFisher) or in the command line on Linux/Unix-based systems (`feht`). Another tool for performing this kind of metadata-drive comparisons is the implementation only available online of metaCATS (Pickett et al., 2013) in the ViPR (Pickett et al., 2012) and the FluDB (Zhang et al., 2016) databases.

Another commonly used approach to find unique signatures for the characterization of epidemiological groups is the design of oligonucleotides, since many assays using those molecules have been developed to discriminate between different viral subtypes or other epidemiological groups. Whereas in molecular biology scientists usually aim to find primers and probes that meet certain biochemical parameters such as melting temperature and avoidance of folding constraints, in computational biology those signatures are text strings that are not necessarily constrained by the biochemical and thermodynamic properties that primers and probes must have for *in vitro* hybridization. For example, in the field of FMDV genomics, a computational genotyping approach that relied on a machine learning technique known as support-vector machine was used to find signature strings for each of the seven FMDV serotypes with 98.45% accuracy (Lin et al., 2008).

The *in silico* validation of known markers or signatures with genomes that were not included in the initial design of an assay can facilitate the assessment of the sensitivity and specificity of virus or subtype-specific features in predicting the presence and subtype of new strains. This idea becomes more relevant as NGS methods become more widely adopted by veterinary and public health diagnostic laboratories. There is the increasing need in those laboratories to find solutions in which subtypes and other epidemiologically relevant groups can be characterized as quickly as possible. Recent innovations in bioinformatics have opened the door, for example, to the detection and characterization of samples from NGS reads without the need of *de novo* or reference-based assembly. One such innovation that has been gaining wider adoption in the literature and is the utilization of k-mers for the *in silico* characterization of microbial genomes and metagenomes. Programs such as `kraken` (Wood and Salzberg, 2014) and `CLARK` (Ounit et al., 2015) classify reads based on their k-mer composition and what proportion of a read's k-mer composition is associated with a given taxon in a reference database. Thus, one potentially can classify FMDV subtypes by using k-mers that are unique to each of those groups. Classification of NGS reads with `kraken` is performed by associating k-mers to NCBI taxonomy ID numbers in which the lowest common ancestor (LCA) is assigned to the reads. In the case of `CLARK`, classification is achieved by using exclusive k-mers that differentiate the taxonomic ranks or epidemiological groups in question. A limitation of `CLARK` in comparison to `kraken` is that metagenomic classification can only be performed at a single taxonomic/hierarchical level at a time (i.e. no LCA approach).

Another tool which is a k-mer based signature discovery program is the Neptune bioinformatics tool Marinier et al. (2017) that searches signature strings using a k-mer based approach, and its goal is to find sequence signatures that differentiate an inclusion group from an exclusion group with high sensitivity and specificity. For example, in Marinier et al. (2017), signatures were searched for differentiating *Enterococcus hirae* versus *Enterococcus faecium* isolates. Other examples included the search for signature regions to

distinguish two *Listeria monocytogenes* serotypes (serotypes 1/2a and 4b), and the location of signatures of Shiga-toxin producing *E. coli* (STEC) (Marinier et al., 2017).

Recently, other approaches that use k-mers for characterizing genomes and metagenomes have emerged. One of those approaches is the adoption of the MinHash algorithm in comparative genomics and metagenomics studies. MinHash is a computational technique that was originally developed for estimating the similarity (i.e. the resemblance) and the containment (i.e. how much of one document is shared with another) of text documents (Broder, 1997). This computational method compares sets of strings between documents and uses the Jaccard index as a distance metric of the resemblance of those sets of strings. MinHash has been applied to comparative genomics most notably with the `mash` (Ondov et al., 2016) and `sourmash` (Brown and Irber, 2016) software tools. The `mash` implementation of MinHash has been tested for the real-time detection of pathogens, and the generation of phylogenetic trees based on Mash distances (Ondov et al., 2016). In addition, both `mash` and `sourmash` allow for the unsupervised classification of reads, which means that reads can be characterized by comparing sets of k-mers using a similarity distance measure, and only the content of the sequences is used without a reference genome.

Other techniques for *in silico* characterization of viral genotypes include SNP genotyping. Single nucleotide variants that define genetic lineages are known as canonical SNPs, which is a concept that is more commonly found in the differentiation of genetic lineages of *Bacillus anthracis* (Keim et al., 2004), but that has also been used for genotyping other bacteria such as *E. coli* (Griffing et al., 2015) and for designing assays for detecting group-specific SNPs in *Francisella tularensis* (Birdsell et al., 2014). Besides their application in bacterial genotyping, canonical SNPs have been discovered and used for the design of forensic real-time PCR assays to differentiate ebolaviruses (Zaire ebolavirus, Sudan ebolavirus, Tai Forest ebolavirus, Bundibugyo ebolavirus, and Reston ebolavirus) (Song et al., 2015).

3.2.2 Objectives

The objectives of this study were:

1. To evaluate *in silico* the sensitivity and specificity of oligonucleotides for the differentiation of FMDV serotypes by BLAST and e-PCR.

2. To discover sequence signatures that are specific to FMDV serotypes using the Neptune bioinformatics tool.

3. To evaluate the performance of different taxonomic classifiers in the characterization of FMDV serotypes.

4. To assess the performance of MinHash-based approaches for the discrimination of FMDV serotypes based on comparisons of k-mer sets.

5. To discover serotype-specific SNPs with an approach based on multiple sequence alignments (multiple sequence alignment (MSA)s) and use said SNPs for the characterization of samples that were not used in the original alignment.

3.3 Materials and Methods

3.3.1 Databases

Various sequence databases were used for the analyses of publicly available sequences and NGS data in the present study (Table 3.1):

3.3.2 Evaluation of existing oligonucleotides for the characterization of FMDV serotypes

In this *in silico* analysis the existing serotype-specific probes used in chapter 2 and the probes from two literature articles were evaluated *in silico* against the OE database (Table 3.1). The purpose of analyzing probes from older papers was to assess how sensitive and specific those oligonucleotide signatures were when evaluated against an updated

Table 3.1: Datasets used for *in silico* analysis of FMDV complete genomes and NGS databases in this study.

Database name	Abbreviation	Analysis	Number of sequences	Description
Oligonucleotide Evaluation	OE	Oligonucleotide evaluation (subsection 3.3.2)	500	Sequences longer than 2000bp retrieved from ViPR from the period of 2014 until 2017.
Sequence Marker Discovery	SMD	Identification of signatures with Neptune (subsubsection 3.4.1.2)	3,039	Partial and complete sequences database that was curated in 2012.
K-mer Classification (training)	KC-training	Kraken, CLARK and Mash (public genomes and NGS)	5,638	FMDV partial and complete genomes released by NCBI until the end of 2013.
K-mer Classification (testing)	KC-testing	Kraken, CLARK and Mash (public genomes only)	2,086	Sequences that were added to the GenBank database in the period from 2014 to the end of 2017.
SNP Discovery (training)	SNP-training	SNP discovery	298	All complete genomes downloaded from GenBank on April 10, 2013.
SNP Discovery (testing)	SNP-testing	SNP discovery	391	Sequences retrieved from ViPR: all complete genomes available until 2017.

database many years later. This was designed as a potential scenario faced by regulatory and research laboratories that may want to develop assays to detect a highly variable virus when molecular detection and subtyping methods already exist.

The first article selected for this *in silico* experiment was Wang et al. (2002), which presented the development of the landmark ViroChip microarray by the research group of Dr. Joe DeRisi at the University of California, San Francisco. The second article selected was Watson et al. (2007) which presented an R package called `DetectIV` for the analysis of high-density multi-pathogen detection microarray data in which FMDV probes were included. That software was developed by Dr. Mick Watson during his time at The Pirbright Institute in the UK (then known as the Institute for Animal Health) site of the World Reference Laboratory for FMD. The probes were part of the Virus Detection Array VA1 developed at that institute and the sequences of those oligonucleotides were downloaded from NCBI's Gene Expression Omnibus (Barrett et al., 2013) (GEO; accession GPL5725). The sequences of the electronic microarray probes and the probes from the two published studies were searched with both BLAST and an NCBI tool for reverse electronic PCR (re-PCR) (Schuler, 1997; Rotmistrovsky et al., 2004) against the OE database.

The sensitivity and specificity of the probes were evaluated using a reproducible workflow designed with the `snakemake` workflow management system (version 4.4) (Köster and Rahmann, 2012), which was implemented to run this analysis (Figure 3.1) by using `blastn` version 2.6.0+ and `re-PCR` version 2.3.12 on a Linux Mint 17.3 (Rosa) laptop with 4 cores and 6 GB RAM. A query coverage of 80% was used to determine BLAST hits and e-PCR hits were determined at 0, 1, 2, 3, 4, and 5 mismatches.

Table 3.2: Number of probes per serotype for each of the studies evaluated against the updated FMDV database.

Serotype	Present study	Watson et al. (2007)	Wang et al. (2002)
A	12	92	0
O	19	145	10
C	8	26	5
Asia1	6	36	0
SAT1	9	42	0
SAT2	13	58	5
SAT3	6	31	0

3.3.3 Discovery of sequence signatures with Neptune

The *de novo* identification of sequence signatures for detection and differentiation of FMDV serotypes was performed using Neptune version 1.2.5 (Marinier et al., 2017). In the Neptune analysis seven comparisons were performed: the FASTA files containing the sequences for each serotypes were the inclusion groups, and the FASTA files containing all the sequences that did not belong to the inclusion serotype were tested as the exclusion groups (Table 3.3). Neptune calculates a score that represents a measure of signature confidence and that is used to rank signatures by sensitivity and specificity. The score is the sum of a inclusion group component (i.e. a positive value that represents the sensitivity with which the signature matches the target group) and an exclusion group component (i.e. a negative value that represents the sensitivity with which the signature matches the non-target group). The score when a 100% sensitive and 100% specific region matches the targets in the inclusion group is 1.0

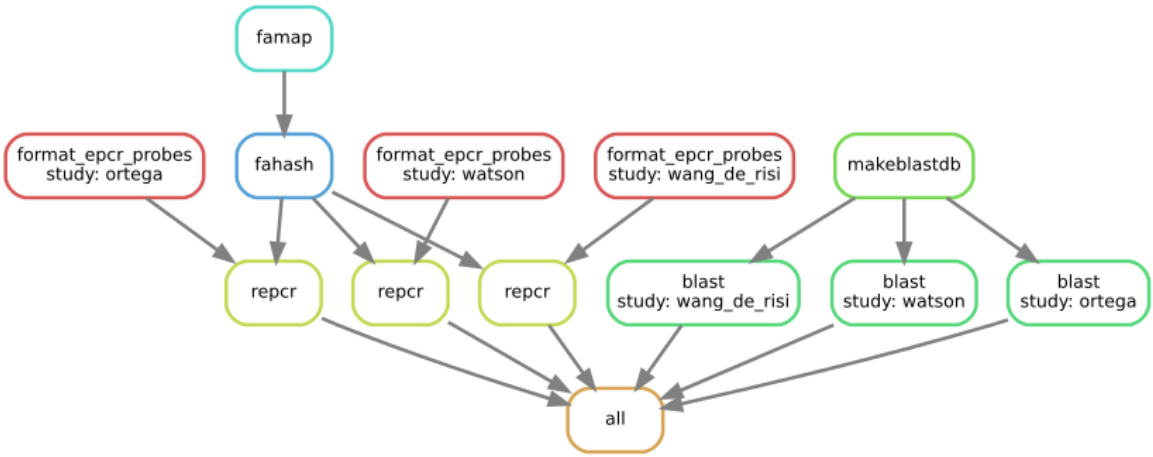


Figure 3.1: Diagram of the oligonucleotide evaluation `snakemake` workflow used for the assessment of the sensitivity of existing signatures for the discrimination of FMDV serotypes with an updated FMDV sequence database. This diagram is a representation of the directed acyclic graph (DAG) of targets and dependencies used in this workflow for the analysis of probes from three studies: chapter 2 of this thesis (`study: ortega`), the study by Wang et al. (2002) (`study: wang_de_risi`), and the study by Watson et al. (2007) (`study: watson`). Preliminary steps for `rePCR` analysis are shown: building the `famap` mapping file (rule `famap`) and the `fahash` hash table (rule `fahash`) from the reference database. Those steps were followed by changing the format of oligonucleotide sequences from FASTA to a table compatible with the `rePCR` program (rule `format_epcr_probes`). Execution of the rules for making the BLAST database (rule `makeblastdb`) and for running BLAST searches (rule `blast`) are also shown. The box labeled as "all" represents the generation of the final target of the workflow: the generation of rePCR and BLAST tabular outputs from all the other dependencies. The `snakemake` system builds the first target in the `Snakefile` which contains all the rules to build all the targets in the workflow. In this case, the "all" target is built first and therefore all the targets in the workflow will be generated.

Table 3.3: Number of sequences used in inclusion and exclusion groups for each comparison in the Neptune analysis.

Comparison	Size of inclusion group	Size of exclusion group
A vs. not A	769	2,270
O vs. not O	1,227	1,661
C vs. not C	133	2,754
Asia1 vs. not Asia1	359	2,529
SAT1 vs. not SAT1	204	2,684
SAT2 vs. not SAT2	278	2,610
SAT3 vs. not SAT3	69	2,820

3.3.4 K-mer based classification

Classifications of FMDV sequences were performed with `kraken` (Wood and Salzberg, 2014) and `CLARK` (Ounit et al., 2015). Both of those programs are fast k-mer based taxonomic classifier programs with different algorithms, and different *in silico* analyses were performed to assess the sensitivity and specificity of classified FMDV sequences by serotype. The analyses performed with those two programs are described in detail below.

3.3.5 Kraken analyses

The first `kraken` analysis involved the classification of complete, publicly available FMDV genomes (KC-testing database) with the reference Kraken database (KC-training database), instead of the RefSeq-based MiniKraken database which contained only one FMDV reference genome. The rationale behind using only FMDV sequences was to find k-mers that are exclusive to each serotype using a database with a larger number and higher diversity of FMDV k-mers. The KC-training database was built with `kraken-build`.

The `kraken` analysis was the classification of seven NGS samples representing the

seven serotypes of FMDV. Those samples were sequenced by the staff of the genomics unit of the National Centre for Foreign Animal Disease (NCFAD) of the CFIA (Winnipeg, Manitoba). The 300 bp reads showed excellent quality metrics measured with `FASTQC` (Andrews, 2016), and were pre-processed with `PRINSEQ-lite` version 0.20.4 by removing 5 bases from both the $5'$ end and the $3'$ end of each read and setting a minimum length of 80 bp. The trimmed reads were classified with `kraken` with two subsequent filtering steps:

Filter 1 Because preliminary analyses revealed the presence of bovine viral diarrhea virus in the FMDV samples, the trimmed reads were filtered with a BVDV database comprised by all complete and partial BVDV sequences obtained from GenBank on December 2016. All the reads that did not map to BVDV were processed with filter 2.

Filter 2 The reads that were unclassified in filter 1 were initially screened with FMDV.

The serotypes of the viruses were characterized by the Vesicular Disease Unit of the NCFAD. The pre-processed reads were subsampled, assembled *de novo* with SPAdes (v3.9.0) (Bankevich et al., 2012), and searched with BLAST after which the closest strains to each sample had been identified (Table 3.4). In order to run the `kraken` NGS analyses, a `snakemake` workflow was designed and implemented for classifying the FMDV reads against the KC-training Kraken database, filtering low confidence results, and generating `kraken` reports with `kraken` version 1.1 in the computer cluster of the National Microbiology Laboratory (NML) (Figure 3.2).

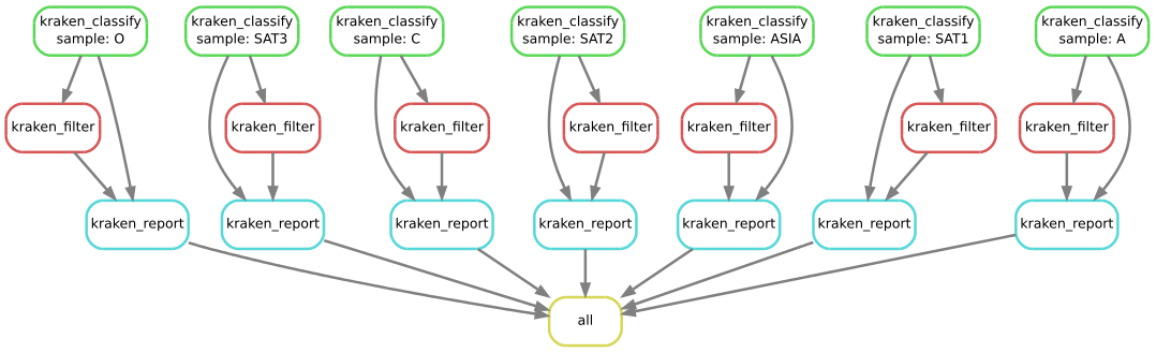


Figure 3.2: Diagram of the Kraken classification `snakemake` workflow used for the analysis of NGS data. This diagram is a representation of the directed acyclic graph (DAG) of targets and dependencies used in this workflow. Classification was performed for each of the seven NGS samples from the NCFAD (rule `kraken_classify`). Filtering of low quality assignments was also applied after classification (rule `kraken_filter`), and taxonomic classification reports were generated for both the unfiltered and filtered classifications (rule `kraken_report`). The box labeled as "all" represents the final target of the workflow, which is the generation of the filtered and unfiltered tabular Kraken reports for all the samples.

Table 3.4: Information on the seven FMDV samples sequenced at the NCFAD

Sample	Top BLAST Hit (GenBank Accession)	% Query Coverage	% Identity
A	A IRN 2005 (EF494486)	97	99
O	O UKG/11/2001 (DQ404180)	95	99
C	C KEN/1/2004 (KM268897)	97	99
Asia1	Asia1 Shamir/89 (JF739177)	100	99
SAT1	SAT1 KEN (JF749860)	99	93
SAT2	SAT2 ZIM/7/83 (AF540910)	99	92
SAT3	SAT3 ZIM/4/1981 (KX375417)	99	99

3.3.6 CLARK analyses

The objective of the first analysis performed with `CLARK` was to classify publicly available FMDV genomes at the serotype level. This was performed by searching the sequences in the KC-testing reference target database with `CLARK`. The first `CLARK` analysis involved

the classification of complete, publicly available FMDV genomes, and it was performed by using KC-testing database. Briefly, hash sketches of the KC-testing database and the KC-training database were generated with `mash sketch` and the Mash distances between the KC-testing and the KC-training database were generated with `mash dist`. Results were filtered to remove all the data with Mash distances equal to 1 (maximum dissimilarity). After that initial filter, the top 5 results for each accession number that had more than 8 hashes shared with the KC-training database were kept.

The analyses of NGS data with `CLARK` (version 1.2.4) were performed using a `snakemake` workflow in the computer cluster of the NML starting with the concatenated forward and reverse FASTQ files of each sample as inputs (Figure 3.3). The same custom database that was used to create the Kraken database (KC-training) was used for building the `CLARK` target database. Since species is the lowest taxonomic level at which `CLARK` performs classifications, a custom target list was built using the GenBank accession numbers and the NCBI Taxonomy ID of the seven FMDV serotypes for each sequence in the database.

The second `CLARK` analysis involved the classification of the seven FMDV samples sequenced at the NCFAD vs. the KC-training database. Two `CLARK` modes were compared in all the `CLARK` analysis: full and default. `CLARK`′s full mode offers higher sensitivity by loading all discriminative k-mers in memory and providing a confidence score for all the classifications, while the default mode is less sensitive, although it has a faster execution, it is precise, and uses less Random Access Memory (RAM) than the full mode.

3.3.7 MinHash-based analyses

A MinHash-based approach (`mash`) was tested in the following series of analyses: the classification of complete and partial publicly available FMDV genomes (KC-testing database), and the classification of the NGS samples sequenced at the NCFAD (n=7) by pairwise comparison of MinHash sketches ofall the samples. A `snakemake` workflow was

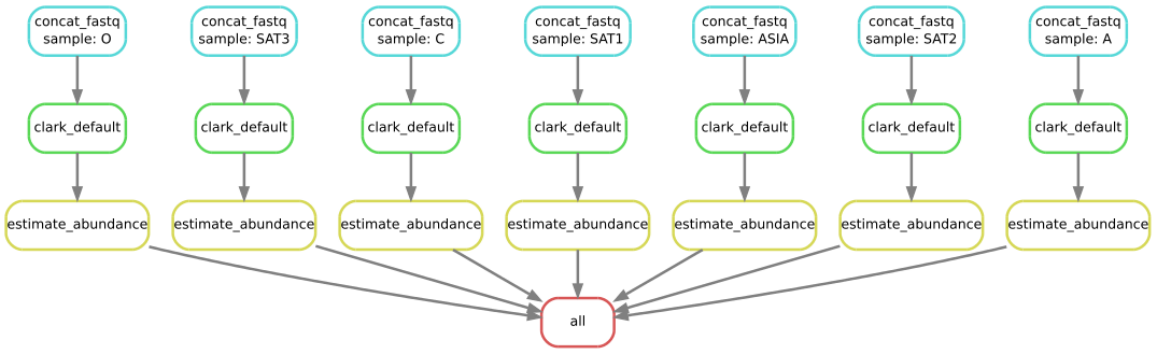


Figure 3.3: Diagram of the CLARK classification `snakemake` workflow which included the classification of concatenated forward and reverse FASTQ files for each sample (rule `concat_fastq`) with `CLARK` version 1.2.4 using the default mode (parameter `-m 1`; rule `clark_default`). Classification was followed by the estimation of serotype abundances using the `getAbundance` script of `CLARK`. A second workflow with the same structure was used for classifying reads with `CLARK`'s full mode (parameter `-m 0`). This diagram is a representation of the directed acyclic graph (DAG) of targets and dependencies used in this workflow, and the box labeled as "all" represents the final target of workflow, which is the generation of the serotype abundance estimation CSV files for all the samples.

designed and implemented for processing the seven FMDV samples sequenced at the NC-FAD with `mash` version 2.0 in the computer cluster of the NML (Figure 3.4).

3.3.8 Identification of serotype-specific SNPs

For this SNP analysis, a multiple sequence alignment (MSA) of complete genomes was generated as the reference for discovering unique SNPs that distinguish the different FMDV serotypes. The training set was generated by aligning 298 FMDV complete genome sequences (SNP-training database; Table 3.1) with ClustalX 2 (Larkin et al., 2007). The workflow contained rules for signature discovery, analysis, and prediction for the differentiation between different groups of foot-and-mouth disease virus sequences.

Single nucleotide polymorphisms were identified using a Python script (developed by P. Krusckewicz; material unpublished) in which the presence or absence of a SNP at a given position of the MSA was converted to a binary code (1 = present, 0 = absent). The statistical analysis of the association of SNPs with metadata categories was performed with `feht` (Laing, 2017). This program written in Haskell was used for analyzing the differences in the

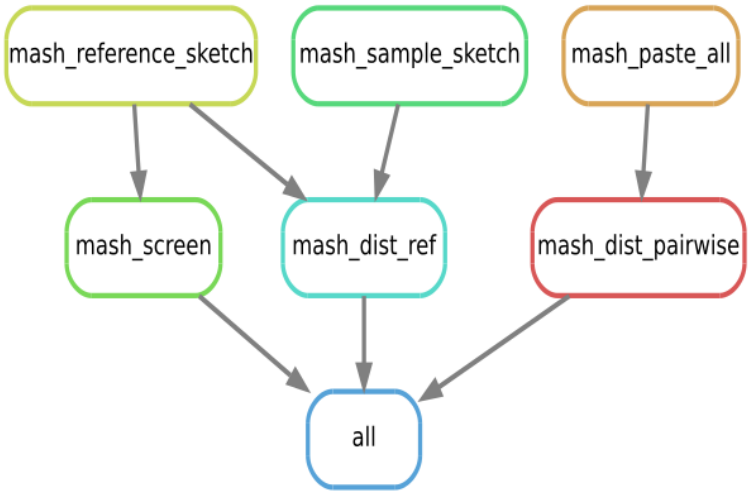


Figure 3.4: Diagram of the Mash `snakemake` workflow. This diagram is a representation of the dependency graph of rules used in this workflow, and targets are not depicted for the purpose of saving space. This workflow included the generation of reference sketches from the same dataset used for the `kraken` and `CLARK` analyses (rule `mash_reference_sketch`). Rules were also designed and implemented for the following steps: generating sample Mash sketches (rule `mash_sample_sketch`), calculating the Mash distances between samples and reference, calculation of containment, and pairwise Mash distances from the comparisons of all the samples.

proportion of SNPs between different metadata groups with Fisher's exact test. The results were corrected for multiple comparisons using the Bonferroni multiple-testing correction. In this analysis, `feht` was used to find highly sensitive and highly specific markers for the prediction of metadata groups. The `feht` program reported all the pairwise comparisons between groups, as well as the comparisons between each group and the combination of all the other groups. For example, `feht` reported the comparison between all serotype A sequences against all sequences that are not serotype A (i.e. the combination of serotypes O, C, Asia1, SAT1, SAT2, and SAT3). The null hypothesis in `feht` is that markers are present in equal proportions in both of the groups that are being compared. The null hypothesis was rejected when a marker is present or absent at a significantly higher proportion (alpha=0.05) in a group than in the combination of the other groups. For the purposes of this study, only SNPs that were present in all the sequences of a serotype and absent from all the sequences of all the other serotypes were used for the generation of SNP profiles.

To evaluate the robustness of the detected SNPs new sequences were aligned with PyNAST (Caporaso et al., 2010), which aligns new sequences to an existing MSA without introducing new gaps. This allowed the comparison of the serotype-specific bases to the new sequences introduced with the PyNAST alignment of new or unknown sequences. Comparisons between the bases identified in the reference alignment and the new PyNAST-aligned sequences were performed with the R statistical computing language (R Core Team, 2016) using the `seqinr` package (Charif and Lobry, 2007) to read the multiple sequence alignments and to convert them into matrices. The nucleotide positions identified with `feht` were used for comparing the SNPs identified in the reference alignment vs. the alignment of new sequences. If the bases at each position were identical, a value of 1 was computed and a value of 0 if they were not identical. Those profiles were used for identifying serotypes based on specific SNPs.

3.4 Results

3.4.1 Classification of publicly available genomes

3.4.1.1 Performance of microarray probes for the *in silico* classification of new strains

Previously designed oligonucleotides were analyzed to test how predictive those signatures were for the characterization of the serotypes of new sequences that were not evaluated for the development of the electronic microarray assay presented in chapter 2. In this *in silico* analysis, the probes from three microarray assays were analyzed with an updated database (OE). The evaluation of the probes from three different studies revealed that the probes in the microarray study presented in this thesis detected all the sequences in the updated database, while the probes from the study by Watson et al. (2007) were less sensitive (Table 3.5).

Table 3.5: BLAST and e-PCR results of pan-FMDV probes versus only those sequences in the Oligonucleotide Evaluation database that contain the common region (n=368/500). Numbers in parenthesis represent the number of unique sequences correctly detected and the total number of sequences of that particular serotype. The threshold for BLAST hits was 80% coverage, and the e-PCR searches were performed by evaluating hits as 0, 1, 2, 3, 4, and 5 mismatches. The study by Wang et al. (2002) did not include pan-FMDV probes.

	This study		Watson et al. (2007)	
Target Serotype	BLAST	e-PCR	BLAST	e-PCR
A	100.0 (91/91)	78.0 (71/91)	70.3 (64/91)	70.3 (64/91)
O	100.0 (175/175)	89.7 (157/175)	81.1 (142/175)	81.1 (142/175)
C	100.0 (9/9)	100.0 (9/9)	66.7 (6/9)	66.7 (6/9)
Asia1	100.0 (19/19)	47.4 (9/19)	47.4 (9/19)	47.4 (9/19)
SAT1	100.0 (29/29)	96.6 (28/29)	72.4 (21/29)	72.4 (21/29)
SAT2	100.0 (35/35)	100.0 (35/35)	77.1 (27/35)	71.4 (25/35)
SAT3	100.0 (10/10)	90.0 (9/10)	30.0 (3/10)	20.0 (2/10)

Table 3.6: BLAST and e-PCR results of querying serotyping probes versus the entire Oligonucleotide Evaluation database (n=500). Probe sequences were obtained from 3 different studies. the numbers inside the parentheses represent the number of unique sequences correctly subtyped and the total number of sequences of that particular serotype.

	This study		Watson et al. (2007)		Wang et al. (2002)	
Target Serotype	BLAST	e-PCR	BLAST	e-PCR	BLAST	e-PCR
A	93.8 (105/112)	93.8 (105/112)	93.8 (105/112)	69.6 (78/112)		
O	96.3 (180/187)	98.4 (184/187)	93.6 (175/187)	89.3 (167/187)	94.1 (176/187)	82.4 (154/187)
C	88.9 (8/9)	77.8 (7/9)	77.8 (7/9)	77.8 (7/9)	100.0 (9/9)	100.0 (9/9)
Asia1	100.0 (20/20)	100.0 (20/20)	100.0 (20/20)	100.0 (20/20)		
SAT1	75.0 (60/80)	76.2 (61/80)	73.8 (59/80)	37.5 (30/80)		
SAT2	100.0 (77/77)	96.1 (74/77)	72.7 (56/77)	31.2 (24/77)	45.5 (35/77)	36.4 (28/77)
SAT3	100.0 (15/15)	100.0 (15/15)	80.0 (12/15)	60.0 (9/15)		

3.4.1.2 Identification of signatures with Neptune

Neptune identified sequence signatures with high specificity and sensitivity for distinguishing each FMDV serotype. Although, as expected, many of these markers were located in the VP1-encoding region of the FMDV genome, there were other regions identified in non-traditional sites for serotype determination such as the 5' UTR, the 3C protease, or the

3A nonstructural protein (Table 3.7). The top sequence signatures found for serotypes A, C, and SAT1 were 100% sensitive and 100% specific to those serotypes (score = 1.0), while sequence signatures discovered for serotypes O (score = 0.8750), Asia1 (score = 0.7609), SAT2 (score = 0.9341), and SAT3 (score = 0.9670) had lower scores, representing a trade-off between sensitivity and specificity. Signature length ranged from 91 base pairs to 4104 base pairs.

Table 3.7: Sequence signatures obtained with Neptune for differentiation of FMDV strains vs. all the other serotypes. The score is a combined measure of sensitivity and specificity, and the size is the number of base pairs of the sequence signatures.

Serotype	Score	Size (bp)	Product
A	1.0	595	3C protease
	1.0	676	5' UTR
	1.0	559	L protease
	1.0	415	3A nonstructural protein
	1.0	133	5' UTR
	1.0	4104	Polyprotein
	0.9930	2839	Polyprotein
O	0.8750	104	VP1
	0.8744	223	VP1
	0.8724	243	VP1
	0.8587	92	VP1
	0.8582	268	VP1
C	1.0	91	VP1
	1.0	93	VP1
	1.0	76	VP1

Table 3.7: Sequence signatures obtained with Neptune for differentiation of FMDV strains vs. all the other serotypes. The score is a combined measure of sensitivity and specificity, and the size is the number of base pairs of the sequence signatures.

Serotype	Score	Size (bp)	Product
	1.0	88	VP1
Asia1	0.7609	138	VP1
	0.7434	152	VP1
	0.7093	375	VP1
	0.6118	85	VP1
SAT1	1.0	84	VP1
	0.8968	155	VP1
	0.8954	239	VP1
SAT2	0.9341	744	P1 polyprotein
	0.8793	174	VP1
	0.8790	124	VP1
	0.8526	251	VP1
SAT3	0.9670	424	3A
	0.9000	150	VP1
	0.8723	188	VP1
	0.8636	110	VP1

3.4.1.3 Classification of publicly available genomes with Kraken

In silico analyses were performed using `kraken` to characterize publicly available complete genomes and NGS reads. The first experiment involved using the MiniKraken database to classify 298 complete FMDV genome. In that analysis, 283 sequences were classified (94.97%), and 15 sequences were left unclassified (5.03%). The 283 sequences

that were classified by Kraken were correctly categorized as FMDV, but all of those 283 sequences were incorrectly classified as serotype O. Those results are explained by the fact that there was only one FMDV sequence in RefSeq at the moment this analysis was performed, and it was the complete genome of a serotype O strain from the 1997 Taiwan outbreak (GenBank accession number NC_004004).

In the second Kraken analysis with the database that included partial sequences and complete genomes, 2,076 sequences were classified out of 2,086 sequences (99.52% sensitivity), and 10 sequences (0.48%) were unclassified. Further investigation of the 10 unclassified sequences revealed that those 10 sequences are from vaccine patents, or from other chimeric sequences that were classified as FMDV in the NCBI taxonomy database. Out of the classified sequences, the classified serotype was correct for 2,003 sequences (96.48% specificity). In addition, there were 15 sequences that were characterized as FMDV but which were not further classified as any of the serotypes.

3.4.1.4 Classification of publicly available genomes with CLARK

Working on default mode, `CLARK` classification resulted in the correct serotype assignment of 2041 of 2086 (97.84%) of the publicly available sequences that were classified, only 45 sequences were not classified. However, that number was reduced to 38 sequences when working with the full mode of `CLARK`. Notably, three serotype sequences were detected with full mode which were not detected with the default mode (Table 3.8).

3.4.1.5 Classification of publicly available genomes with Mash

Classification of publicly available sequences was performed by using `mash dist` to calculate the Mash distance between the KC-testing database and the KC-training database.

3.4.1.6 Search and validation of serotype-specific SNPs

SNP markers that had 100% sensitivity and 100% specificity (ratio of 1.0) were found for the seven FMDV serotypes analyzed with the SNP-training database. The SNP profiles

Table 3.8: Classification of KC-testing database with CLARK. Two CLARK methods were compared: the default mode and the more computationally expensive full mode.

Target Serotype	% Subtyped - Default	% Subtyped - Full
A	20.81 (434)	20.81 (434)
O	51.97 (1084)	51.97 (1084)
C	0	0.14 (3)
Asia1	7.53 (157)	7.53 (157)
SAT1	5.37 (112)	5.56 (116)
SAT2	11.65 (243)	11.65 (243)
SAT3	0.53 (11)	0.53 (11)
Unknown	2.16 (45)	1.82 (38)

Table 3.9: Summary of Mash classification results

Query Serotype	Percent Covered
A	98.2 (431/439)
O	99.6 (1077/1081)
C	71.4 (5/7)
Asia1	100.0 (161/161)
SAT1	93.8 (105/112)
SAT2	90.6 (221/244)
SAT3	76.9 (10/13)

found in this study can be used as signatures to identify new strains that were not in the original alignment. In the approach used in this study, the SNP profiles were generated by comparing the SNPs of the reference FMDV whole genome sequence alignment vs. an alignment of the new sequences built with PyNAST using the reference alignment as a template. Comparisons were performed at the positions that were identified with the `feht` statistical tool as specific to the group in question. This method resulted in binary profiles that were highly discriminatory of the seven serotypes in the *in silico* experiments performed for this study.

The SNP profiles had different numbers of discriminatory SNPs for each serotype, and these discriminatory positions were present in the range from position 2121 to position 4156 of the reference alignment Table 3.10. These positions are various sites in the polyprotein region of the FMDV genome.

The SNP genotyping method here described was able to classify correctly 100% (391/391) of the sequenced downloaded from ViPR that were not part of the original alignment with which the SNP profiles were determined. There was only one case in which the binary profile matches less than half of positions of the target sequence (Figure 3.5g). In that case, the serotype SAT3 profile matched only 15/34 of the UGA/1/13 strain (GenBank Accession KJ820999). The other three serotype SAT3 strains ZIM/4/1981 (KX375417), KNP/10/90 (KR108950), and ZIM/6/91(KM268901) were matched by 32, 32, and 34 SNPs of the SAT3 profile. In addition, the two serotype SAT3 strains (FV536932 and FV536933) present in the "unidentified" pool of 53 genomes were matched by 26/34 SNPs of the SAT3 profile.

When tested with sequences that had not been in the previous alignment, the method identified the serotypes of 391 FMDV complete genomes Figure 3.5. There were 53 accessions in the testing data set of which ViPR or NCBI did not identify a serotype or in which the curators likely did not include the serotype information. These sequences were used as "blind" test cases to assess the accuracy of the SNP markers in characterizing the seven

serotypes of FMDV. The SNP genotyping method here described correctly identified the serotype of 100% of the uncharacterized sequences. Further investigation of the accession numbers and the metadata present in Genbank (i.e. record titles or notes in the sequence features) further confirmed the 100% sensitivity and specificity of the characterization of the serotypes of these genomes.

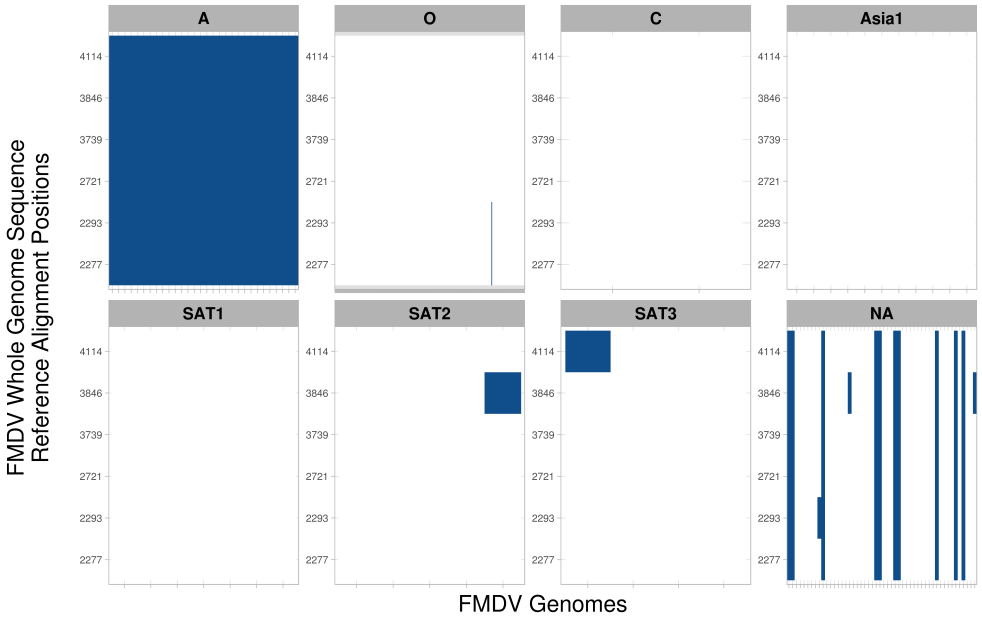
Table 3.10: Number, p-values, and positions in the reference multiple sequence alignment MSA of serotype-specific SNPs. The SNPs were located in the regions of the FMDV genome encoding for the VP4, VP2, VP3, and VP1 peptides. All the SNPs in this table were found to be present in 100% of the genomes of eachserotype, and absent from 100% from the other serotypes. P-values were calculated using the Fisher's Exact Test with `feht` (Laing, 2017). The reference MSA file and a list of the GenBank accession numbers for each genome sequence used in the alignment can be accessed in `https://github.com/ropolomx/fmdv_snp_analysis`.

Serotype	Number of SNPs	P-value	Positions in reference MSA
A	6	1.22×10^{-68}	2277, 2293, 2721, 3739, 3846, 4114
O	9	1.29×10^{-84}	2298, 2359, 2385, 2398, 2685, 3251, 3444, 3960, 4129
C	5	1.82×10^{-29}	2292, 2293, 3898, 3940, 3941
Asia1	13	2.69×10^{-46}	2121, 2292, 2293, 2386, 2574, 2746, 2779, 3013, 3024, 3030, 3031, 3807, 4059
SAT1	25	4.38×10^{-13}	2151, 2289, 2314, 2401, 2506, 2664, 2728, 2743, 2997, 3039, 3040, 3175, 3178, 3256, 3422, 3798, 3841, 3916, 3918, 3919, 3922, 3937, 4026, 4035, 4104
SAT2	34	2.14×10^{-8}	2289, 2290, 2296, 2691, 2700, 2721, 2742, 2749, 2758, 2922, 2935, 2938, 3072, 3178, 3310, 3341, 3353, 3366, 3369, 3377, 3379, 3380, 3419, 3492, 3542, 3680, 3753, 3765, 3772, 3790, 4033, 4135, 4137, 4156
SAT3	36	6.14×10^{-5}	2226, 2227, 2289, 2314, 2371, 2395, 2448, 2458, 2469, 2526, 2581, 2713, 2751, 2752, 3027, 3039, 3040, 3057, 3081, 3354, 3392, 3428, 3437, 3591, 3598, 3742, 3744, 3799, 3825, 3843, 3978, 3982, 4034, 4039, 4105, 4148

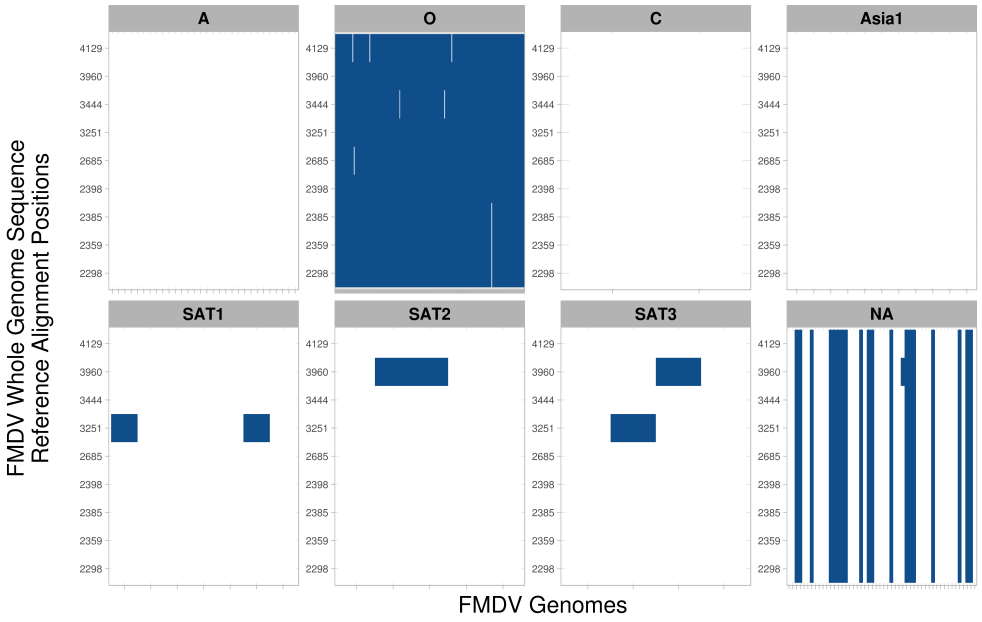
3.4.2 Classification of NGS data

3.4.2.1 Kraken analysis

In the Kraken analysis the NCFAD NGS samples were classified against the reference database. This analysis resulted in more complex results than the previous two experiments. The majority of reads of the serotype A, Asia1, O, SAT1, and SAT2 samples corresponded to the correct serotype. However, in the case of the SAT3 sample, the classification of samples was very close between the SAT3 (n=1,197,219 reads; 38.54% of input reads), and

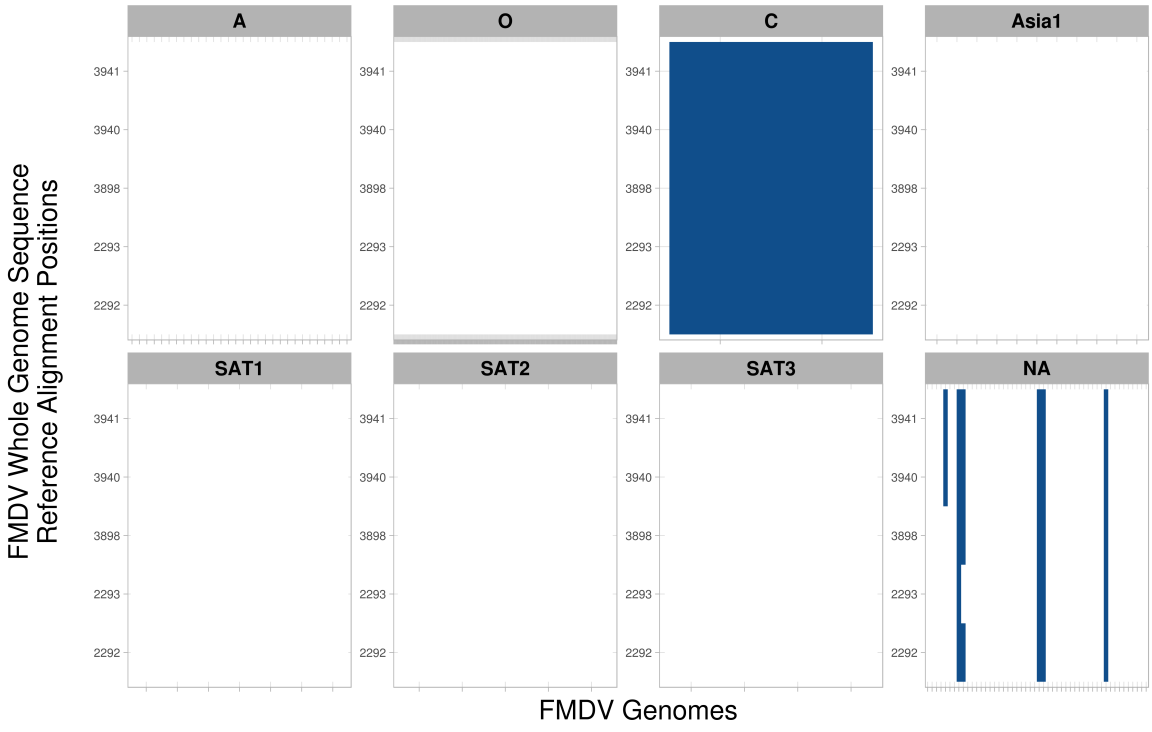


(a) Serotype A-specific SNPs

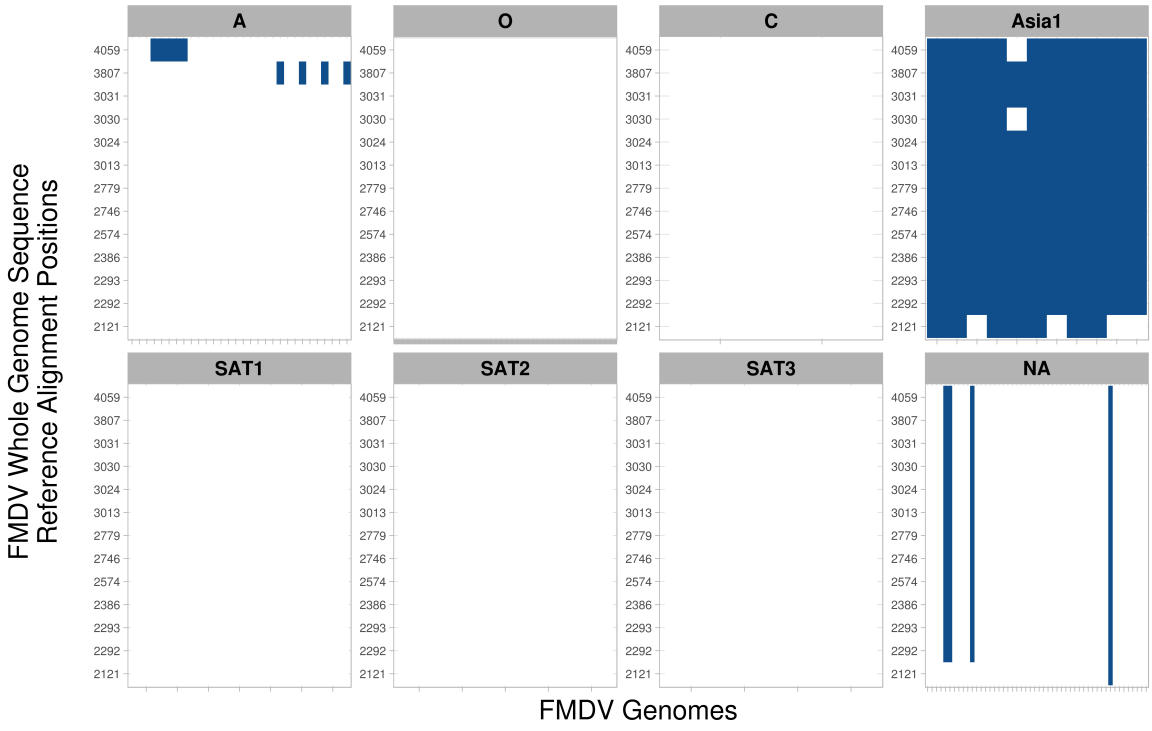


(b) Serotype O-specific SNPs

Figure 3.5: Presence and absence of serotype-specific SNPs in (a) serotype A and (b) serotype O FMDV genomes. The heatmaps represent the presence and absence of serotype-specific SNPs in genomes of the seven FMDV serotypes. SNP positions from the reference alignment are in the y-axis,and isolates are in the x-axis. Blue colour represents that the base at the position in the alignment of the new sequences is identical to the base in the reference alignment, and white colour represents that the base at the position in the alignment of the new sequences is not identical to the same position in the reference alignment. The "NA" category represents 53 genomes that did not have an assigned serotype in the ViPR database. This figure continues in the next three pages with results for the other 5 serotypes.

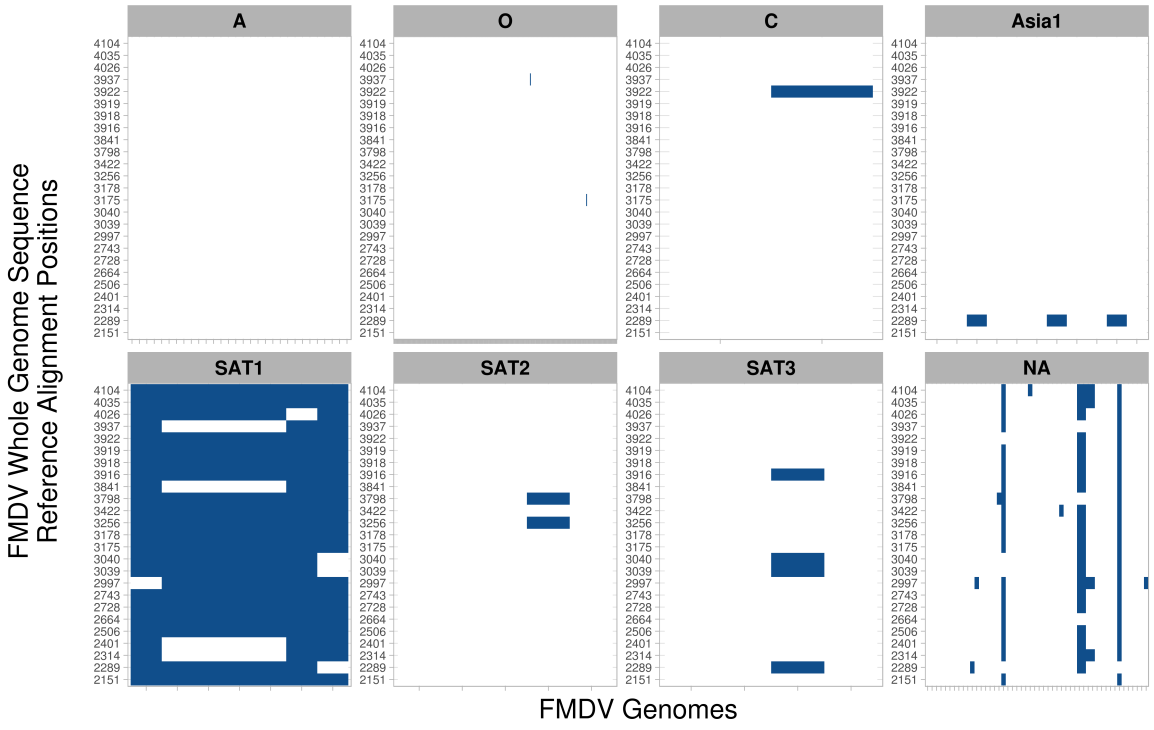


(c) Serotype C-specific SNPs

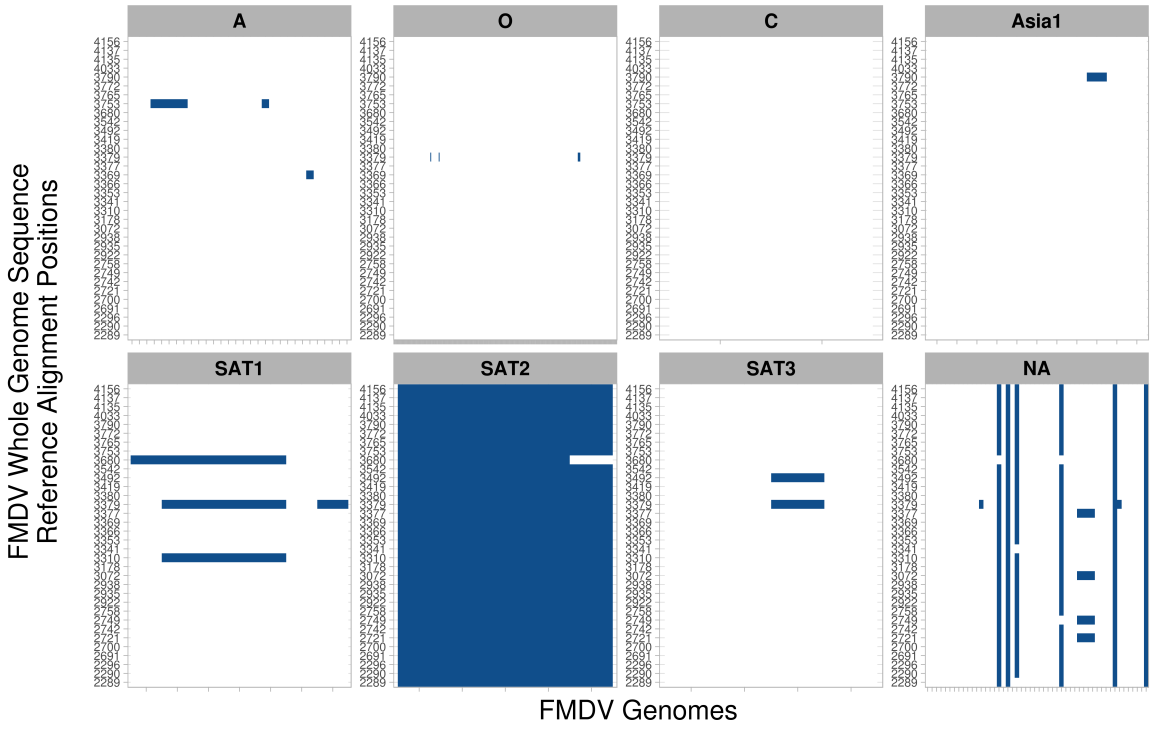


(d) Serotype Asia1-specific SNPs

Figure 3.5: Presence and absence of serotype-specific SNPs in (c) serotype C and (d) serotype Asia1 FMDV genomes.

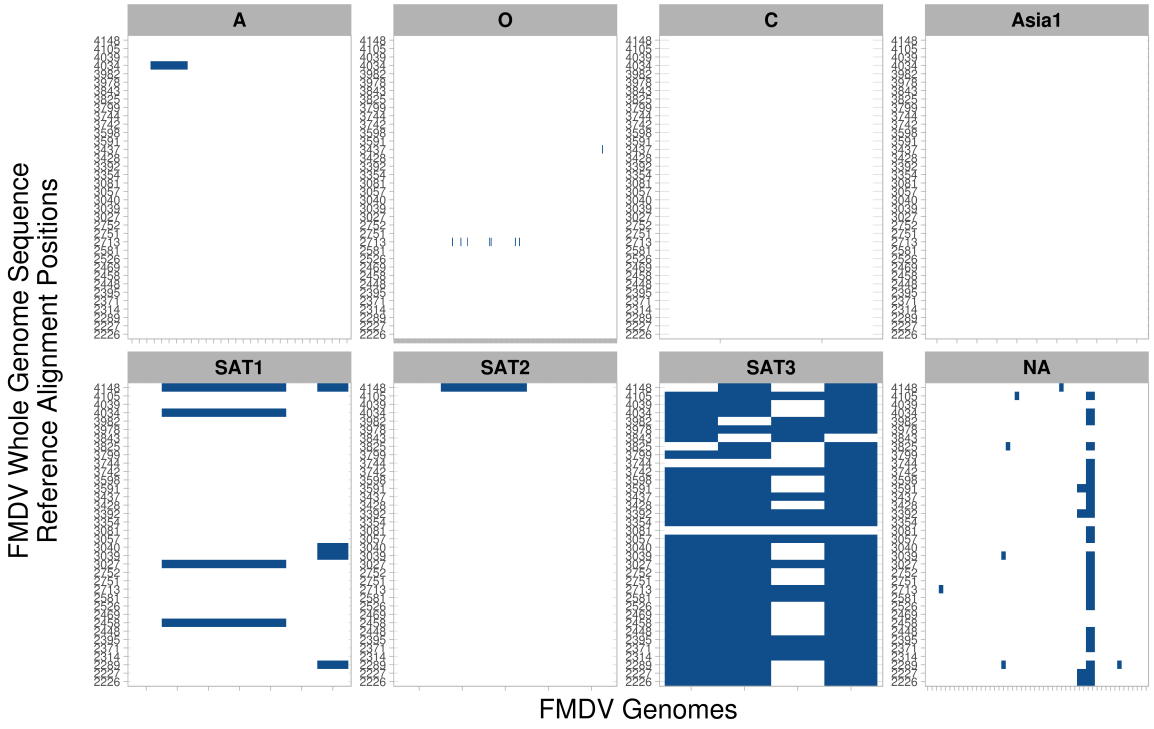


(e) Serotype SAT1-specific SNPs



(f) Serotype SAT2-specific SNPs

Figure 3.5: Presence and absence of serotype-specific SNPs in (e) serotype SAT1 and (f) serotype SAT2 FMDV genomes.



(g) Serotype SAT3-specific SNPs

Figure 3.5: Presence and absence of serotype-specific SNPs in (g) serotype SAT3 genomes.

SAT1 serotypes (n=1,151,282 reads; 37.06% of input reads). Two potential explanations for the apparently low specificity is the presence of low-quality classifications in Kraken's results, or classifications where the reads of multiple individual k-mers were split between SAT1 and SAT3 serotypes. This hypothesis was evaluated by using the `kraken-filter` program to filter the classifications based on Kraken's scoring method with two filtering scores assessed 0.05 and 2 (Figure 3.6). The filtering of the classification at the threshold of 0.05 did not show a considerable difference in comparison to the unfiltered classification, while the classification at the confidence score threshold of 2 lead to a reduction of sensitivity and a decrease in specificity.

3.4.2.2 CLARK analysis

This analysis involved the classification of the 7 NGS NCFAD samples using filtered fastq reads, and many reads were not identified as unique to each serotype, which is consis-

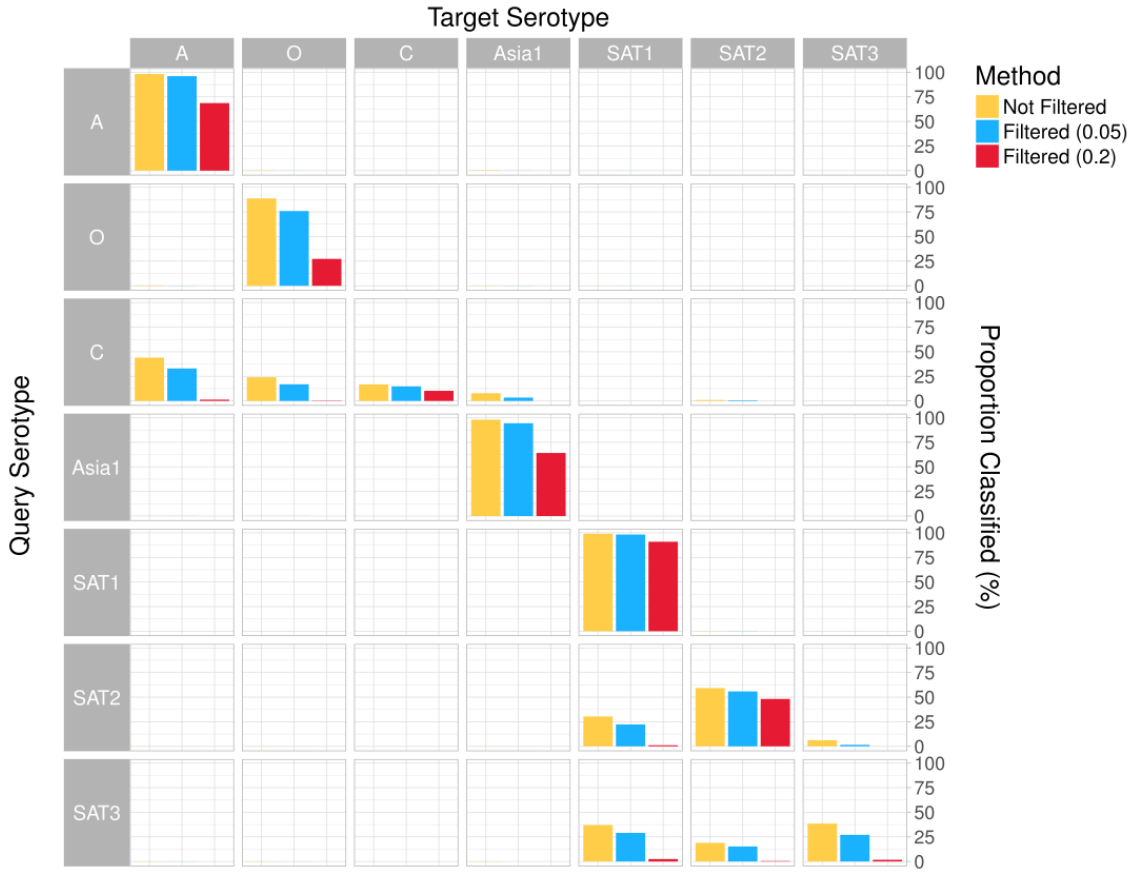


Figure 3.6: FMDV serotype classification of NCFAD samples with Kraken. Kraken assignments were performed after running `kraken` followed by the filtration of low quality assignments with `kraken-filter` at two different filtration thresholds: 0.05 and 0.2. The proportions of classified reads without using `kraken-filter` after assignment with `kraken` were also reported (Not Filtered).

tent with the observations seen in Kraken that many reads were stranded at the species level (i.e. only classified as FMDV and not classified further). A potential explanation for this is that there are many k-mers that are common between all the FMDV serotypes and only a fraction of those are specific to each serotype. In the case of the classification of the serotype C NCFAD sample, using the default mode resulted in the identification of assignments to other serotypes. However, using the default mode resulted in a mis-characterization of the sample as a mostly SAT2 sample. In the case of the SAT2 NCFAD sample, the classification at the default mode resulted in the assignment to the SAT1 and SAT3, while there was an increase of the proportion of reads that were classified as SAT2 in comparison to the proportion of reads classified as SAT1. In the case of the SAT3 sample, there was an increase in the proportion of reads classified as SAT3 in comparison to the reads classified as SAT1 Figure 3.7.

3.4.2.3 Mash analysis

The objective of this analyses was to classify seven FMDV NGS samples into the seven FMDV serotypes by using`mash`, an implementation of the MinHash algorithm for bioinformatics. In this analysis the k-mer sketches of the NCFAD FMDV samples were classified against a reference k-mer sketch built with publicly available databases and the reads were screened for containment (`mash screen`) versus the same reference database, and pairwise comparisons of all the sample NGS k-mer sketches were performed.

Plotting the screening or the analysis of containment of the NGS samples versus the reference database resulted in two types of signals in the data: the presence of multiple query hits that match more than one serotype, and the high sensitivity of Mash to detect each of the serotypes Figure 3.8. Using this approach resulted in the accurate classification of the seven NGS samples representing the seven serotypes. The pairwise comparison of the Mash sketches also showed that the NGS datasets representing the seven serotypes were differentiated by using `mash`, and that the three southern African serotypes are closer

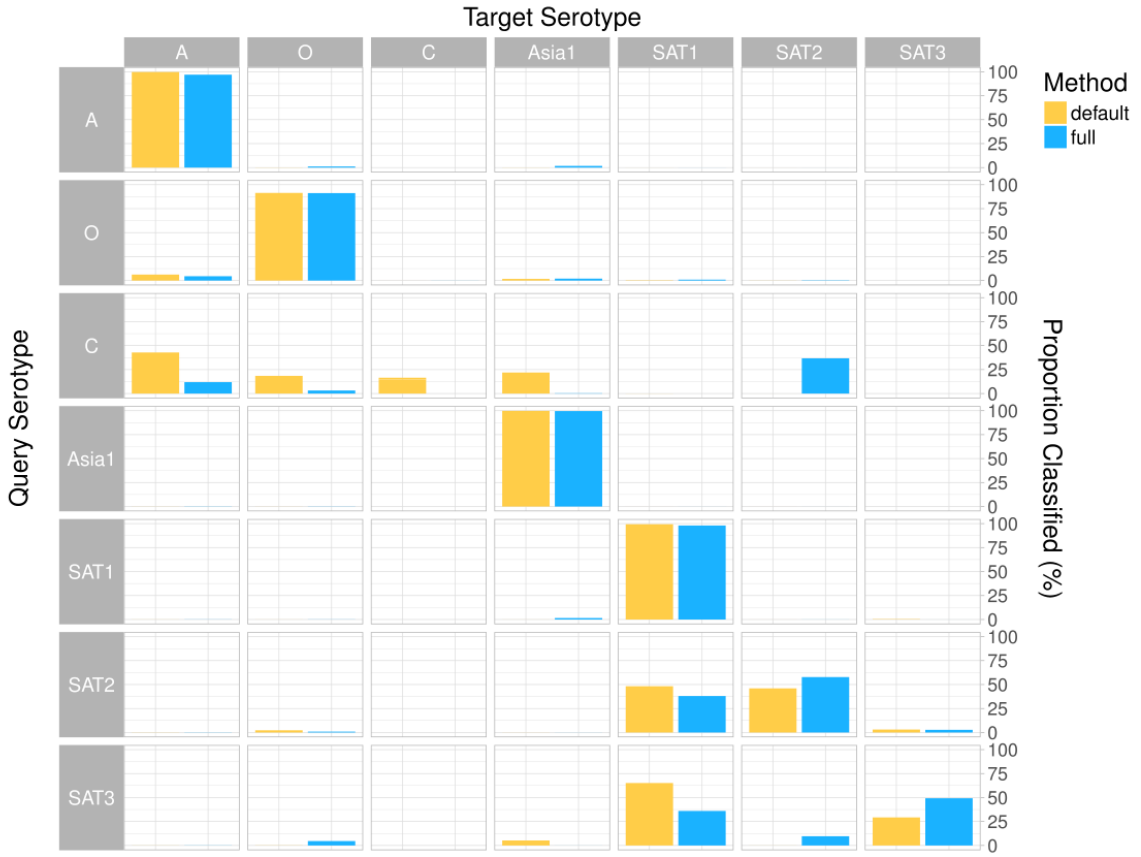


Figure 3.7: FMDV serotype classification of NCFAD samples with CLARK. Two modes of the CLARK software were tested: default and full. The proportions of classified reads for the results of both modes were reported.

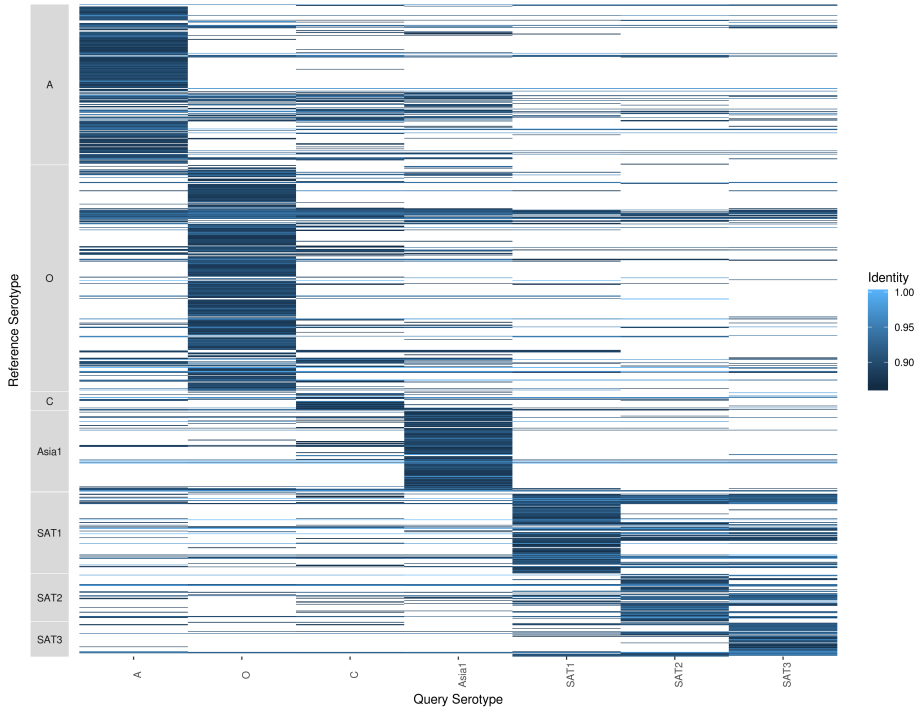


Figure 3.8: Screening of FMDV NGS samples with Mash. Seven NGS samples sequenced at the NCFAD were screened versus a reference database Mash sketch. The blue colour scale represents the identity proportion between query and reference sketches. White colour represents absence of results (i.e. there were no hits reported between the NGS sample query and a given strain in the reference sketch).

to each other than to the other serotypes, which resembles what has been observed from analyses of nucleotide and protein sequences (Figure 3.9).

3.5 Discussion

3.5.1 Evaluation of existing oligonucleotides

Updating microarray assays is particularly challenging for those assays that target ss-RNA viruses because of the rapid evolution of those viruses. However, oligonucleotides that were designed for the detection of viruses or subtypes of viruses have the potential to be used in new microarray assays in new platforms or as *in silico* signatures with NGS data provided they are sensitive and specific with updated database. The research question addressed in this analysis was: what is the sensitivity and specificity of the serotype-specific probes against new sequences in the FASTA format database? The pan-FMDV (or detec-

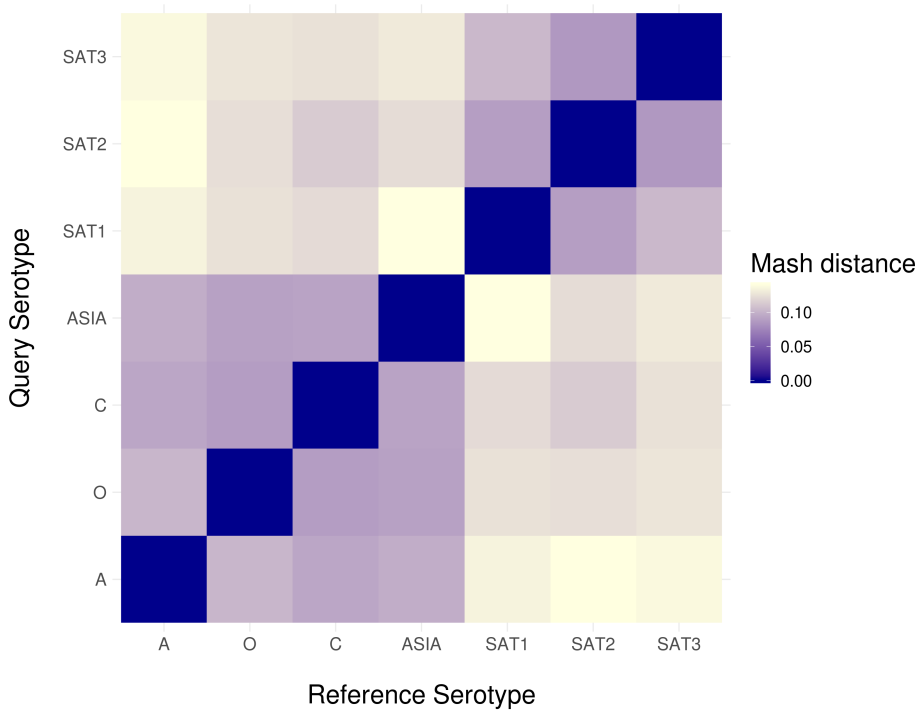


Figure 3.9: Heatmap of the pairwise Mash distances between the seven samples sequenced at the NCFAD. The `mash dist` tool was used to calculate the Mash distances between these datasets in a pairwise fashion.

tion) probes used in this study detected all the sequences in the updated database, while the probes in the older study by Watson et al. (2007) were less sensitive. In the case of the serotype-specific probes, the performance of the probes from the electronic microarray assay presented in this thesis was superior to the sensitivity of the other two studies. Notably, the performance of the serotype C probes from the ViroChip study by Wang et al. (2002) outperformed the probes from this thesis and from the study by Watson et al. (2007). However, the serotype SAT2 probes from the ViroChip study were outperformed by the probes in this study and the study by Watson et al. (2007). Those results suggest that the sensitivity of probes designed in older studies may decrease with updated databases, and that the *in silico* coverage of new strains of FMDV may not be as effective with oligonucleotides designed for older studies. As more NGS data becomes available through the Sequence Read Archive (SRA) NCBI and in GenBank, more in-depth analysis is required to evaluate the sensitivity of detection and subtyping probes.

3.5.2 Signature discovery with Neptune

The analysis with Neptune resulted in the discovery of serotype-specific sequence signatures of various sizes. These signatures can potentially become amplicon targets for discriminatory assays (i.e. a multiplex RT-PCR assay for discriminating between the serotypes), because the sequence signatures are sensitive and specific to the serotypes, and those regions can be used to design molecular assays to differentiate between the serotypes, which includes the design of oligonucleotides. More comprehensive validation is required in order to assess how predictive these signatures are for the detection and characterization of FMDV strains not present in the training database used in this study. Refining this analysis with the improvement of the training database can results in the generation of signatures that are more appropriate for *in silico* and *in vitro* analysis. Another potential application of Neptune to FMDV genomics in particular and viral genomics in general

3.5.3 K-mer based taxonomic classification methods

Kraken is one of the most widely used taxonomic classifiers of metagenomic datasets. Kraken is a fast tool due to the fast k-mer counter `jellyfish`. Interestingly, the `kraken-filter` option is a tool of this software that is not widely used and there are few entries in the literature that use it. One of the authors mentioned that they were looking for a different method to filter low-quality assignments (Salzberg, S., 2017, personal communication). Tools that were designed to improve on Kraken such as Bracken (Lu et al., 2017) and the recently developed KrakenHLL can potentially promise to further refine the assignments done with `kraken`. These results suggest two possibilities: (1) that there is more than one serotypes present in the same sample, or (2) that `kraken` does not have the sufficient precision or specificity for the classification of FMDV subtypes. In regards to the first possibility, the implementation of *de novo* metagenomic assembly could be useful for generating evidence of whether there are strains of more than one serotype present in the sample in the samples. Then, the comparison of those assemblies to the Kraken classifica-

tion could confirm if the assignments are correct. In the second case, an expansion of the reference database may improve to accurately predict the serotype of newer sequences.

In this series of k-mer based analyses, the taxonomic classification tool called `CLARK` was also used to classify FMDV sequences by serotypes. Similarly to `kraken`, `CLARK` (Ounit et al., 2015) classifies draft assemblies (in FASTA format) or sequencing reads (in FASTQ format) using k-mer profiles. In contrast to `kraken`, `CLARK` classifies reads only at one taxonomic level at a time. Two modes of this software were tested: default and full. The authors of `CLARK` reported that it was the fastest taxonomy classifier in the benchmarking tests they ran (Ounit et al., 2015). Considering that `CLARK` accepts FASTQ files as input, this approach may be suitable for characterizing the serotypes of unassembled, pre-processed NGS, and the implementation of this tool in the workflows of regulatory laboratories has the potential to make the determination of FMDV serotypes much more efficient.

Recently, alignment-based taxonomic classifiers such as `centrifuge` (Kim et al., 2016), or the protein sequence based `kaiju` (Menzel et al., 2016) can be potentially implemented for the identification of *in silico* signatures of FMDV serotypes. For instance, a recent study by Walsh et al. (2018) found that `kaiju` is one of more suitable tools for the analysis of the composition of low-complexity communities. Moreover, `kaiju` can be potentially applied to the detection of unknown viruses thanks to its translation of nucleotide sequences to amino acid sequences. Since amino acid sequences are more conserved than DNA sequences, this can improve sensitivity of detection. This functionality is also present in the newest version of Kraken (Kraken 2).

3.5.4 Mash-based classification of serotypes

The classification of publicly available genomes using `mash dist` to calculate the Mash distance showed great promise in the classification of the KC-testing database. Mash is a computational technique that is faster than typical approaches such as BLAST, and its implementation in bioinformatics is gaining more adoption for the rapid characterization of

the serotypes of pathogens and the characterization of metagenomes. The pairwise comparisons of the NGS samples performed with Mash revealed that sequences are much closer to each other. Serotypes A, O, C, and Asia1 have shorter distances to each other than to the three SAT serotypes. Although Mash was not designed explicitly for reconstructing phylogenies, it can be used to approximate phylogenies rapidly by applying hierarchical clustering to Mash distances (Ondov et al., 2016).

3.5.5 Serotype-specific SNPs of FMDV

In this study, SNPs that were statistically associated with FMDV serotypes were discovered using an approach that relies on multiple sequence alignments. To the author's knowledge, this study is the first time that barcodes of serotype-specific SNPs have been proposed for the foot-and-mouth disease virus. The SNP barcoding method here utilized shares some similarities to the strategy of oligotyping proposed by Eren et al. (2013) for amplicon-based microbial community profiling. In oligotyping, a profile or barcode is generated with SNPs that minimize the entropy between ecological groups of interest (i.e. SNPs that are present in one groups but absent in another).

From a bioinformatics point of view, a key aspect of the method here presented is the utilization of the pyNAST alignment software (Caporaso et al., 2010). `pyNAST` is often used in 16S microbial community profiling, such as in the QIIME package, when new 16S sequences are added to existing 16S alignments such as Greengenes, or SILVA. `pyNAST` returns an alignment with the same positions as the template. The alignment of new sequences to the reference alignment can be improved by generating a reference alignment with more complete genomes. This can increase the confidence in the validation of SNPs in new strains. Further characterization of the SNPs identified in this thesis is required, such as the precise annotation of the SNPs in a reference genome and the determination of whether the SNPs represent synonymous or non-synonymous mutations.

FMDV is notable for its highly variable genomes, and the concept of quasispecies was

developed by researchers who studied this virus (Domingo et al., 1985, 1992). The identification of these unique SNPs could become the basis for designing detection and subtyping assays. For example, oligonucleotides can be designed for genotyping assays based on a single SNP. An example of this is the primer extension assay developed by Spizz et al. (2012), in which SNPs were targeted for warfarin genotyping. In the case of FMDV, this could potentially allow for assays that require less oligonucleotides for serotyping due to a relative small number of stable SNPs being targeted. This could potentially mean that the number of probes in an microarray chip or card can be less than what was used for the electronic microarray assay reported in chapter 2. An interesting follow up to the search of serotype-specific SNPs would be the search of topotype-specific SNPs. The rapid identification of topotypes in field isolates has the potential to provide more in-depth knowledge of the molecular epidemiology of the virus during outbreak conditions. Validation of the SNPs found in this work would be required to consider them suitable for assay design or for the study of FMDV evolution.

3.5.6 Future directions

The identification of group-specific signatures is important for the characterization of pathogens and the understanding of their evolution. Although the robust SNPs here presented can be adapted as markers and targets of real-time PCR assays, the subtle nucleotide variation of FMDV quasispecies could be further characterized using other approaches. Another aspect to reflect upon the SNP analysis is that the robust SNPs here discovered are based on consensus sequences. If the goal is to characterize of FMDV quasispecies, this could be potentially performed with specialized tools such as ViQuaS (Jayasundara et al., 2015) which uses a reference-guided *de novo* assembly approach for building local haplotypes. ViQuas was tested with two FMDV datasets from the same sample and it outperformed similar tools in strain reconstruction. Other software tools for quasispecies characterization such as `quasitools`(`https://github.com/phac-nml/quasitools/`), which

is currently applied to the analysis of HIV quasispecies, could be potentially adapted and applied to the analysis of FMDV quasispecies and viral drug resistance. Another approach that can potentially be useful for studying the genetic diversity of quasispecies of the same FMDV serotype is the application of oligotyping (Eren et al., 2013) to the genomes or VP1 sequences of isolates from different geographical locations, different outbreak events or points in time. The oligotyping approach applies a technique known as minimum entropy decomposition (Eren et al., 2015) to multiple sequence alignments of NGS reads in order to find SNPs that differentiate ecologically important groups.

Chapter 4

Conclusions and future prospects

The consequences of the contagious nature of FMDV in the agricultural sector of the many countries where this virus is or has been present have shaped international livestock trade and animal welfare policies. The entry of transboundary and emerging diseases of animals such as FMDV into Canada is a risk for which the design and implementation of science-based solutions are required. The collection of useful information by Canada's animal health authorities during an outbreak situation is central to Canada's emergency response to mitigate the consequences for the agricultural sector. The characterization of the subtypes of FMDV field strains informs the selection of FMDV field vaccine. Given that Canada is, along with the United States of America, and Mexico, one of the owners of the North American FMDV antigen bank, obtaining this information quickly will be critical during an outbreak event. The contribution of this thesis to the detection and classification of FMDV serotypes is the development and initial validation of the first electronic microarray assay for the simultaneous detection and serotyping of FMDV. Furthermore, this work has explored the classification of FMDV serotypes of NGS and publicly available data and the generation of a novel SNP genotyping approach for the classification of FMDV serotypes. These approaches can be potentially applied to the classification of the subtypes of other viruses such as avian influenza virus.

As more regulatory and research laboratories adopt next generation sequencing for diagnostics and surveillance, innovations in sequencing technologies have resulted in the development of portable and field-deployable instruments that have the potential to be used

for the detection of viral pathogens of humans and animals in the field as outbreaks occur, including foot-and-mouth disease virus (Freimanis et al., 2016). One of those innovative sequencing technologies is the Oxford Nanopore MinION sequencer. This technology involves the reading of changes in ionic current due to base-specific fluctuations while DNA goes through a protein channel (Wang et al., 2015b).

Examples of viruses sequenced with the Oxford Nanopore include an avian influenza genome (Wang et al., 2015a), the real time genome sequencing during the 2015 Ebola outbreak in West Africa (Quick et al., 2016), and the discovery of porcine kobuvirus for the first time in Belgium (Theuns et al., 2018). Nanopore technologies are promising for the global surveillance of viruses in the field because the portability of the MinION in combination with global positioning system global positioning system (GPS) technologies can enable high-resolution tracking of viruses such as FMDV which can provide better support for tracing the spread of the virus (Freimanis et al., 2016).

Beyond the applications of nanopore sequencing for whole genome sequencing, the potential application of metagenomic sequencing for the identification of viral pathogens in clinical samples with a MinION sequencer was reported in the study by Greninger et al. (2015). In that study, sequencing followed by cloud-based basecalling of nanopore reads and real-time bionformatics analysis with the metaPORE pipeline was achieved in less than 6 hours. Using that method, Ebola virus, chikungunya virus, and hepatitis C virus were detected from human blood samples. The metaPORE pipeline is an adaptation of another metagenomic pathogen detection and discovery pipeline: SURPI (Sequence-based Ultra-Rapid Pathogen Identification) (Naccache et al., 2014). The SURPI pipeline was notably applied to the diagnosis of neuroleptospirosis with *Letospira santarosai* of a 14-year patient with meningoencephalitis where other diagnostic methods including magnetic resonance imaging, brain biopsy, and multiple PCR tests for different bacterial and viral pathogens did not reveal the causative pathogen (Wilson et al., 2014). Its innovations included the use of a fast nucleotide aligner called SNAP (Zaharia et al., 2011) and a fast aminoacid aligner

called RAPSearch (Zhao et al., 2012) that accelerated the processing time of the analysis of metagenomic samples. In that clinical case, 475 sequence reads derived from the patient's cerebrospinal fluid were mapped to the *Leptospira borgpetersenii* genome which was the closest match, and appropriate treatment was then prescribed by the medical care personnel in charge of the patient. Also, taxonomic classifiers such as the k-mer based approaches explored in chapter 3 (Kraken and CLARK), as well as alignment-based classifiers (e.g. Centrifuge) and other classification approaches that translate NGS reads and search them against an aminoacid sequence database (i.e. `kaiju` and `kraken` 2).

Generating whole genome sequences not only facilitates the work of regulatory agencies and research laboratories in understanding the molecular epidemiology of viruses such as FMDV, but it also opens the door to other applications related to the control of the virus. For instance, a critical step in the control of FMD is matching vaccines to field isolates (Paton et al., 2005), which is usually performed with serological methods such as ELISA. In the study by Reeve et al. (2010), predictions of vaccine matching and identification of antigenic variability were made based on genomic sequencing in computational work which also showed the presence of three epitopes for FMDV SAT1 for the first time in this case. One of the main results from that study was the performance of vaccine matching using sequence data was superior to the performance of serology. This work involved statistical modeling using a linear mixed-effects model to estimate r1 values from serological data (i.e. the ratio of antibody titre of reference serum against field isolates to the titre of reference serum against vaccine strains). The model also used sequence-based predictors, accounted for phylogeny-based effects, and individual areas and residues were added to this model to identify epitopes, as well as within-serotype diversity. This work allowed the prediction of the best vaccine match for any foot-and-mouth disease virus and the coverage of new vaccine candidates from their capsid sequences (Reeve et al., 2010).

Another important development in FMDV genomics research that has the potential to expand the current understanding of the epidemiology and the pathology of the disease is

the application of NGS for the characterization of quasispecies during outbreak events For example, the characterization of the genetic variation of FMDV at the host and the herd levels allowed to identify sequences not only at the consensus level, but at the sub-consensus level when studying the transmission of the virus between farms in samples from the UK 2007 outbreak (King et al., 2016) The work by Wright et al. (2011) showed that the application of NGS to the analysis of viral populations within two bovine epithelial samples showed intermediate evolution stages in the replication of the virus and evolutionary dynamics at a higher resolution. Using NGS for studying viral quasispecies can help to understand sequences that are related and understand the genetic variation of populations at each site (Wright et al., 2011).

The work in this thesis explored laboratory and *in silico* methods for the characterization of FMDV serotypes. The techniques and methods here presented could also be applied to other viruses, or to other epidemiological groups of interest for FMDV researchers (e.g. topotypes). Although additional validation work is required for the electronic microarray assay and the *in silico* analyses, the work here presented lays a foundation for the utilization of primers and probes for detection and characterization of FMDV. The refinement and expansion of training sets used in the *in silico* analyses will help to improve the performance of the classification of NGS reads of FMDV. Some of the approaches discussed in this chapter could also help to detect animal pathogens from metagenomic sequencing in cases where other diagnostic tests have failed to reveal what is the possible causative pathogen.

Bibliography

Agriculture and Agri-Food Canada (2016). Red meat and livestock export reports. http://www.agr.gc.ca/eng/industry-markets-and-trade/statistics-and-market-information/by-product-sector/red-meat-and-livestock/red-meat-and-livestock-market-information/exports/red-meat-exports-by-country/?id=1419965032803. Accessed: 2017-01-25.

Agriculture and Agri-Food Canada (2017). Red meat trade balance reports. http://www.agr.gc.ca/eng/industry-markets-and-trade/statistics-and-market-information/by-product-sector/red-meat-and-livestock/red-meat-and-livestock-market-information/trade-balance/?id=1415860000005. Accessed: 2017-01-25.

Alagaili, A. N., Briese, T., Mishra, N., Kapoor, V., Sameroff, S. C., Wit, E. d., Munster, V. J., Hensley, L. E., Zalmout, I. S., Kapoor, A., Epstein, J. H., Karesh, W. B., Daszak, P., Mohammed, O. B., and Lipkin, W. I. (2014). Middle East Respiratory Syndrome Coronavirus Infection in Dromedary Camels in Saudi Arabia. *mBio*, 5:e00884–14.

Alexandersen, S. and Donaldson, A. I. (2002). Further studies to quantify the dose of natural aerosols of foot-and-mouth disease virus for pigs. *Epidemiology and Infection*, 128:313–323.

Alexandersen, S. and Mowat, N. (2005). Foot-and-Mouth Disease: Host Range and Pathogenesis. In Mahy, P. D. B. W. J., editor, *Foot-and-Mouth Disease Virus*, Current Topics in Microbiology and Immunology, pages 9–42. Springer Berlin Heidelberg.

Alexandersen, S., Zhang, Z., Donaldson, A., and Garland, A. (2003). The Pathogenesis and Diagnosis of Foot-and-Mouth Disease. *Journal of Comparative Pathology*, 129:1–36.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25:3389–3402.

Ambagala, A., Fisher, M., Goolia, M., Nfon, C., Furukawa-Stoffer, T., Ortega Polo, R., and Lung, O. (2016). Field-Deployable Reverse Transcription-Insulated Isothermal PCR (RT-iiPCR) Assay for Rapid and Sensitive Detection of Foot-and-Mouth Disease Virus. *Transboundary and Emerging Diseases*, pages 1610–1623.

Andrews, S. (2016). FastQC A Quality Control tool for High Throughput Sequence Data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed: 2016-12-29.

Arzt, J., Baxt, B., Grubman, M. J., Jackson, T., Juleff, N., Rhyan, J., Rieder, E., Waters, R., and Rodriguez, L. L. (2011). The Pathogenesis of Foot-and-Mouth Disease II: Viral

Pathways in Swine, Small Ruminants, and Wildlife; Myotropism, Chronic Syndromes, and Molecular Virus-Host Interactions: Foot-and-Mouth Disease Pathogenesis: Atypical Syndromes. *Transboundary and Emerging Diseases*, 58:305–326.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., and Pevzner, P. A. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology*, 19:455–477.

Banér, J., Gyarmati, P., Yacoub, A., Hakhverdyan, M., Stenberg, J., Ericsson, O., Nilsson, M., Landegren, U., and Belák, S. (2007). Microarray-based molecular detection of foot-and-mouth disease, vesicular stomatitis and swine vesicular disease viruses, using padlock probes. *Journal of Virological Methods*, 143:200–206.

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S., and Soboleva, A. (2013). NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research*, 41:D991–D995.

Baxi, M. K., Baxi, S., Clavijo, A., Burton, K. M., and Deregt, D. (2006). Microarray-based detection and typing of foot-and-mouth disease virus. *The Veterinary Journal*, 172:473–481.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2009). GenBank. *Nucleic Acids Research*, 37:D26–D31.

Birdsell, D. N., Vogler, A. J., Buchhagen, J., Clare, A., Kaufman, E., Naumann, A., Driebe, E., Wagner, D. M., and Keim, P. S. (2014). TaqMan Real-Time PCR Assays for Single-Nucleotide Polymorphisms Which Identify Francisella tularensis and Its Subspecies and Subpopulations. *PLOS ONE*, 9:e107964.

Briscoe, W. T. (1872). A Case of Foot-and-Mouth Disease in the Human Subject. *British Medical Journal*, 2:464.

Broder, A. Z. (1997). On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No.97TB100171)*, pages 21–29.

Brown, C. T. and Irber, L. (2016). sourmash: a library for MinHash sketching of DNA. *The Journal of Open Source Software*. DOI: 10.21105/joss.00027.

Brown, F. (2003). The history of research in foot-and-mouth disease. *Virus research*, 91:3–7.

Callens, M. and De Clercq, K. (1997). Differentiation of the seven serotypes of foot-and-mouth disease virus by reverse transcriptase polymerase chain reaction. *Journal of Virological methods*, 67:35–44.

Canadian Food Inspection Agency (2012). Procedures and requirements for horses imported into canada from countries with foot and mouth disease (fmd). http://www.inspection.gc.ca/animals/terrestrial-animals/diseases/reportable/foot-and-mouth-disease/horses/eng/1332033335585/1332033552830. Accessed:2017-03-22.

Caporaso, J. G., Bittinger, K., Bushman, F. D., DeSantis, T. Z., Andersen, G. L., and Knight, R. (2010). PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics*, 26:266–267.

Carpenter, T. E., O'Brien, J. M., Hagerman, A. D., and McCarl, B. A. (2011). Epidemic and Economic Impacts of Delayed Detection of Foot-And-Mouth Disease: A Case Study of a Simulated Outbreak in California. *Journal of Veterinary Diagnostic Investigation*, 23:26–33.

Carrillo, C., Tulman, E. R., Delhon, G., Lu, Z., Carreno, A., Vagnozzi, A., Kutish, G. F., and Rock, D. L. (2005). Comparative Genomics of Foot-and-Mouth Disease Virus. *Journal of Virology*, 79:6487–6504.

Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. (2016). *shiny: Web Application Framework for R*. R package version 0.13.2.

Charif, D. and Lobry, J. (2007). SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In Bastolla, U., Porto, M., Roman, H., and Vendruscolo, M., editors, *Structural approaches to sequence evolution: Molecules, networks, populations*, Biological and Medical Physics, Biomedical Engineering, pages 207–232. Springer Verlag, New York. ISBN : 978-3-540-35305-8.

Chen, E. C., Miller, S. A., DeRisi, J. L., and Chiu, C. Y. (2011). Using a Pan-Viral Microarray Assay (Virochip) to Screen Clinical Samples for Viral Pathogens. *JoVE (Journal of Visualized Experiments)*, pages e2536–e2536.

Chen, Z., Abrams, W. R., Geva, E., de Dood, C. J., Gonz&#xe1, Lez, J., M, S., Tanke, H. J., Niedbala, R. S., Zhou, P., Malamud, D., and Corstjens, P. L. A. M. (2013). Development of a Generic Microfluidic Device for Simultaneous Detection of Antibodies and Nucleic Acids in Oral Fluids. *BioMed Research International*, 2013.

Daszak, P., Cunningham, A. A., and Hyatt, A. D. (2000). Emerging Infectious Diseases of Wildlife– Threats to Biodiversity and Human Health. *Science*, 287:443–449.

Department of Justice Canada (2015). Health of Animals Act.

Domingo, E., Escarmis, C., Martinez, M. A., Martinez-Salas, E., and Mateu, M. G. (1992). Foot-and-Mouth Disease Virus Populations Are Quasispecies. In Holland, J. J., editor, *Genetic Diversity of RNA Viruses*, Current Topics in Microbiology and Immunology, pages 33–47. Springer Berlin Heidelberg.

Domingo, E., Escarmıs, C., Baranowski, E., Ruiz-Jarabo, C. M., Carrillo, E., Núñez, J. I., and Sobrino, F. (2003). Evolution of foot-and-mouth disease virus. *Virus Research*, 91:47–63.

Domingo, E., Martínez-Salas, E., Sobrino, F., de la Torre, J. C., Portela, A., Ortín, J., López-Galindez, C., Pérez-Breña, P., Villanueva, N., Nájera, R., VandePol, S., Steinhauer, D., DePolo, N., and Holland, J. (1985). The quasispecies (extremely heterogeneous) nature of viral RNA genome populations: biological relevance — a review. *Gene*, 40:1–8.

Domingo, E., Pariente, N., Airaksinen, A., González-Lopez, C., Sierra, S., Herrera, M., Grande-Pérez, A., Lowenstein, P. R., Manrubia, S. C., Lázaro, E., and Escarmís, C. (2005). Foot-and-Mouth Disease Virus Evolution: Exploring Pathways Towards Virus Extinction. In Mahy, P. D. B. W. J., editor, *Foot-and-Mouth Disease Virus*, Current Topics in Microbiology and Immunology, pages 149–173. Springer Berlin Heidelberg.

Dopazo, J., Dress, A., and Haeseler, A. v. (1993). Split decomposition: a technique to analyze viral evolution. *Proceedings of the National Academy of Sciences*, 90:10320–10324.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95:14863–14868.

Engvall, E. (2010). The ELISA, Enzyme-Linked Immunosorbent Assay. *Clinical Chemistry*, 56:319–320.

Engvall, E. and Perlmann, P. (1971). Enzyme-linked immunosorbent assay (ELISA) quantitative assay of immunoglobulin G. *Immunochemistry*, 8:871–874.

Eren, A. M., Maignien, L., Sul, W. J., Murphy, L. G., Grim, S. L., Morrison, H. G., and Sogin, M. L. (2013). Oligotyping: differentiating between closely related microbial taxa using 16s rRNA gene data. *Methods in Ecology and Evolution*, 4:1111–1119.

Eren, A. M., Morrison, H. G., Lescault, P. J., Reveillaud, J., Vineis, J. H., and Sogin, M. L. (2015). Minimum entropy decomposition: Unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *The ISME Journal*, 9:968.

Erickson, A., Fisher, M., Furukawa-Stoffer, T., Ambagala, A., Hodko, D., Pasick, J., King, D. P., Nfon, C., Ortega Polo, R., and Lung, O. (2017). A multiplex reverse transcription PCR and automated electronic microarray assay for detection and differentiation of seven viruses affecting swine. *Transboundary and Emerging Diseases*, 65:e272–e283.

Food and Agriculture Organization of the United Nations (2016). Transboundary animal diseases. `http://www.fao.org/emergencies/emergency-types/transboundary-animal-diseases/en/`. Accessed: 2015-12-16.

Freimanis, G. L., Di Nardo, A., Bankowska, K., King, D. J., Wadsworth, J., Knowles, N. J., and King, D. P. (2016). Genomics and outbreaks: foot and mouth disease. *Revue Scientifique et Technique - Office International des Epizooties*, 35:175–89.

Gall, A., Hoffmann, B., Harder, T., Grund, C., Höper, D., and Beer, M. (2009). Design and Validation of a Microarray for Detection, Hemagglutinin Subtyping, and Pathotyping of Avian Influenza Viruses. *Journal of Clinical Microbiology*, 47:327–334.

García-Arriaza, J., Ojosnegros, S., Dávila, M., Domingo, E., and Escarmís, C. (2006). Dynamics of Mutation and Recombination in a Replicating Population of Complementing, Defective Viral Genomes. *Journal of Molecular Biology*, 360:558–572.

Gire, S. K., Goba, A., Andersen, K. G., Sealfon, R. S. G., Park, D. J., Kanneh, L., Jalloh, S., Momoh, M., Fullah, M., Dudas, G., Wohl, S., Moses, L. M., Yozwiak, N. L., Winnicki, S., Matranga, C. B., Malboeuf, C. M., Qu, J., Gladden, A. D., Schaffner, S. F., Yang, X., Jiang, P.-P., Nekoui, M., Colubri, A., Coomber, M. R., Fonnie, M., Moigboi, A., Gbakie, M., Kamara, F. K., Tucker, V., Konuwa, E., Saffa, S., Sellu, J., Jalloh, A. A., Kovoma, A., Koninga, J., Mustapha, I., Kargbo, K., Foday, M., Yillah, M., Kanneh, F., Robert, W., Massally, J. L. B., Chapman, S. B., Bochicchio, J., Murphy, C., Nusbaum, C., Young, S., Birren, B. W., Grant, D. S., Scheiffelin, J. S., Lander, E. S., Happi, C., Gevao, S. M., Gnirke, A., Rambaut, A., Garry, R. F., Khan, S. H., and Sabeti, P. C. (2014). Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science*, 345:1369–1372.

Greiner, M., Pfeiffer, D., and Smith, R. D. (2000). Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. *Preventive Veterinary Medicine*, 45:23–41.

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., and Chiu, C. Y. (2015). Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. *Genome Medicine*, 7:99.

Griffing, S. M., MacCannell, D. R., Schmidtke, A. J., Freeman, M. M., Hyytiä-Trees, E., Gerner-Smidt, P., Ribot, E. M., and Bono, J. L. (2015). Canonical Single Nucleotide Polymorphisms (SNPs) for High-Resolution Subtyping of Shiga-Toxin Producing Escherichia coli (STEC) O157:H7. *PLOS ONE*, 10:e0131967.

Hafen, R. and Continuum Analytics, Inc. (2016). *rbokeh: R Interface for Bokeh*. R package version 0.4.2.

Hall, T. A. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series*, 41:95–98.

Hamming, R. W. (1950). Error detecting and error correcting codes. *The Bell System Technical Journal*, 29:147–160.

Haydon, D. T., Kao, R. R., and Kitching, R. P. (2004). The UK foot-and-mouth disease outbreak — the aftermath. *Nature Reviews Microbiology*, 2:675–681.

Hebert, P. D. N., Cywinska, A., Ball, S. L., and deWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270:313–321.

Hedger, R. S. and Brooksby, J. B. (1976). FMD in an Indian elephant. *Veterinary Record*, 99:93–93.

Hedger, R. S., Condy, J. B., and Golding, S. M. (1972). Infection of some species of African wild life with foot-and-mouth disease virus. *Journal of Comparative Pathology*, 82:455–461.

Hendling, M., Pabinger, S., Peters, K., Wolff, N., Conzemius, R., and Barišić, I. (2018). Oli2go: an automated multiplex oligonucleotide design tool. *Nucleic Acids Research*, 46(W1):W252–W256.

Hindson, B. J., Reid, S. M., Baker, B. R., Ebert, K., Ferris, N. P., Tammero, L. F. B., Lenhoff, R. J., Naraghi-Arani, P., Vitalis, E. A., Slezak, T. R., Hullinger, P. J., and King, D. P. (2008). Diagnostic Evaluation of Multiplexed Reverse Transcription-PCR Microsphere Array Assay for Detection of Foot-and-Mouth and Look-Alike Disease Viruses. *Journal of Clinical Microbiology*, 46:1081–1089.

Hoffmann, B., Scheuch, M., Höper, D., Jungblut, R., Holsteg, M., Schirrmeier, H., Eschbaumer, M., Goller, K. V., Wernike, K., Fischer, M., Breithaupt, A., Mettenleiter, T. C., and Beer, M. (2012). Novel Orthobunyavirus in Cattle, Europe, 2011. *Emerging Infectious Diseases*, 18:469–472.

Howson, E. L. A., Armson, B., Lyons, N. A., Chepkwony, E., Kasanga, C. J., Kandusi, S., Ndusilo, N., Yamazaki, W., Gizaw, D., Cleaveland, S., Lembo, T., Rauh, R., Nelson, W. M., Wood, B. A., Mioulet, V., King, D. P., and Fowler, V. L. (2017). Direct detection and characterization of foot-and-mouth disease virus in East Africa using a field-ready real-time PCR platform. *Transboundary and Emerging Diseases*, pages 221–231.

Hysom, D. A., Naraghi-Arani, P., Elsheikh, M., Carrillo, A. C., Williams, P. L., and Gardner, S. N. (2012). Skip the Alignment: Degenerate, Multiplex Primer and Probe Design Using K-mer Matching Instead of Alignments. *PLoS ONE*, 7:e34560.

Jack, P. J., Amos-Ritchie, R. N., Reverter, A., Palacios, G., Quan, P.-L., Jabado, O., Briese, T., Ian Lipkin, W., and Boyle, D. B. (2009). Microarray-based detection of viruses causing vesicular or vesicular-like lesions in livestock animals. *Veterinary Microbiology*, 133:145–153.

Jamal, S. M. and Belsham, G. J. (2013). Foot-and-mouth disease: past, present and future. *Veterinary Research*, 44:116.

Jamal, S. M. and Belsham, G. J. (2018). Molecular epidemiology, evolution and phylogeny of foot-and-mouth disease virus. *Infection, Genetics and Evolution*, 59:84–98.

Jayasundara, D., Saeed, I., Maheswararajah, S., Chang, B. C., Tang, S.-L., and Halgamuge, S. K. (2015). ViQuaS: an improved reconstruction pipeline for viral quasispecies spectra generated by next-generation sequencing. *Bioinformatics*, 31(6):886–896.

Joo, Y.-S., An, S.-h., Kim, O.-K., Lubroth, J., and Sur, J.-H. (2002). Foot-and-mouth disease eradication efforts in the Republic of Korea. *Canadian journal of veterinary research*, 66:122.

Katoh, K. and Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, 30:772–780.

Keim, P., Van Ert, M. N., Pearson, T., Vogler, A. J., Huynh, L. Y., and Wagner, D. M. (2004). Anthrax molecular epidemiology and forensics: using the appropriate marker for different evolutionary scales. *Infection, Genetics and Evolution*, 4:205–213.

Kim, D., Song, L., Breitwieser, F. P., and Salzberg, S. L. (2016). Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Research*.

Kindhauser, M. K., Allen, T., Frank, V., Santhana, R. S., and Dye, C. (2016). Zika: the origin and spread of a mosquito-borne virus. *Bulletin of the World Health Organization*, 94:675–686C.

King, D. J., Freimanis, G. L., Orton, R. J., Waters, R. A., Haydon, D. T., and King, D. P. (2016). Investigating intra-host and intra-herd sequence diversity of foot-and-mouth disease virus. *Infection, Genetics and Evolution*, 44:286–292.

King, D. P., Reid, S. M., Hutchings, G. H., Grierson, S. S., Wilkinson, P. J., Dixon, L. K., Bastos, A. D. S., and Drew, T. W. (2003). Development of a TaqMan® PCR assay with internal amplification control for the detection of African swine fever virus. *Journal of Virological Methods*, 107:53–61.

Kitching, R. P. (2002). Clinical variation in foot and mouth disease: cattle. *Revue Scientifique et Technique - Office International des Epizooties*, 21:499–504.

Kloeze, H., Mukhi, S., Kitching, P., Lees, V. W., and Alexandersen, S. (2010). Effective Animal Health Disease Surveillance Using a Network-Enabled Approach. *Transboundary and Emerging Diseases*, 57:414–419.

Knight-Jones, T. J. D., McLaws, M., and Rushton, J. (2016). Foot-and-Mouth Disease Impact on Smallholders - What Do We Know, What Don't We Know and How Can We Find Out More? *Transboundary and Emerging Diseases*, pages 1079–1094.

Knight-Jones, T. J. D. and Rushton, J. (2013). The economic impacts of foot and mouth disease – What are they, how big are they and where do they occur? *Preventive Veterinary Medicine*, 112:161–173.

Knowles, N. and Samuel, A. (2003). Molecular epidemiology of foot-and-mouth disease virus. *Virus Research*, 91:65–80.

Ksiazek, T. G., Erdman, D., Goldsmith, C. S., Zaki, S. R., Peret, T., Emery, S., Tong, S., Urbani, C., Comer, J. A., Lim, W., Rollin, P. E., Dowell, S. F., Ling, A.-E., Humphrey, C. D., Shieh, W.-J., Guarner, J., Paddock, C. D., Rota, P., Fields, B., DeRisi, J., Yang, J.-Y., Cox, N., Hughes, J. M., LeDuc, J. W., Bellini, W. J., Anderson, L. J., and Group, t. S. W. (2003). A Novel Coronavirus Associated with Severe Acute Respiratory Syndrome. *New England Journal of Medicine*, 348:1953–1966.

Köster, J. and Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28:2520–2522.

Laing, C. (2017). feht: Predictive marker discovery for groups; binary data, genomic data (single nucleotide variants), arbitrary character data. https://github.com/chadlaing/feht. original-date: 2016-08-17T21:33:31Z.

Lambert, D., Carrillo, C. D., Koziol, A. G., Manninger, P., and Blais, B. W. (2015). GeneSippr: A Rapid Whole-Genome Approach for the Identification and Characterization of Foodborne Pathogens such as Priority Shiga Toxigenic Escherichia coli. *PLOS ONE*, 10:e0122928.

Lancefield, R. C. (1933). A Serological Differentiation of Human and Other Groups of Hemolytic Streptococci. *Journal of Experimental Medicine*, 57:571–595.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J., and Higgins, D. G. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*, 23:2947–2948.

Li, K. S., Guan, Y., Wang, J., Smith, G. J. D., Xu, K. M., Duan, L., Rahardjo, A. P., Puthavathana, P., Buranathai, C., and Nguyen, T. D. (2004). Genesis of a highly pathogenic and potentially pandemic H5N1 influenza virus in eastern Asia. *Nature*, 430:209.

Li, W. and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22:1658–1659.

Lin, G., Cai, Z., Wu, J., Wan, X.-F., Xu, L., and Goebel, R. (2008). Identifying a few foot-and-mouth disease virus signature nucleotide strings for computational genotyping. *BMC Bioinformatics*, 9:279.

Loeffler, F. and Frosch, P. (1898). Report of the commission for research on foot-and-mouth disease. *Zentrabl. Bacteriol. Parastenkunde Infektionkrankh*, 23:371–391.

Longjam, N., Deb, R., Sarmah, A. K., Tayo, T., Awachat, V. B., and Saxena, V. K. (2011). A Brief Review on Diagnosis of Foot-and-Mouth Disease of Livestock: Conventional to Molecular Tools. *Veterinary Medicine International*, 2011:e905768.

Lu, J., Breitwieser, F. P., Thielen, P., and Salzberg, S. L. (2017). Bracken: estimating species abundance in metagenomics data. *PeerJ Computer Science*, 3:e104.

Lung, O., Beeston, A., Ohene-Adjei, S., Pasick, J., Hodko, D., Hughes, K. B., Furukawa-Stoffer, T., Fisher, M., and Deregt, D. (2012). Electronic microarray assays for avian influenza and Newcastle disease virus. *Journal of Virological Methods*, 185:244–253.

Lung, O., Fisher, M., Beeston, A., Hughes, K. B., Clavijo, A., Goolia, M., Pasick, J., Mauro, W., and Deregt, D. (2011). Multiplex RT-PCR detection and microarray typing of vesicular disease viruses. *Journal of Virological Methods*, 175:236–245.

Lung, O., Furukawa-Stoffer, T., Burton Hughes, K., Pasick, J., King, D. P., and Hodko, D. (2016). Multiplex RT-PCR and Automated Microarray for Detection of Eight Bovine Viruses. *Transboundary and Emerging Diseases*, 64:1929–1934.

Lung, O., Nadin-Davis, S., Fisher, M., Erickson, A., Knowles, M., Furukawa-Stoffer, T., and Ambagala, A. (2013). Microarray for Identification of the Chiropteran Host Species of Rabies Virus in Canada. *Microarrays*, 2:153–169.

Lung, O., Ohene-Adjei, S., Buchanan, C., Joseph, T., King, R., Erickson, A., Detmer, S., and Ambagala, A. (2015). Multiplex PCR and Microarray for Detection of Swine Respiratory Pathogens. *Transboundary and Emerging Diseases*, 64:834–848.

Lung, O., Ohene-Adjei, S., Buchanan, C., Joseph, T., King, R., Erickson, A., Detmer, S., and Ambagala, A. (2017). Multiplex PCR and Microarray for Detection of Swine Respiratory Pathogens. *Transboundary and Emerging Diseases*, 64:834–848.

Lusted, L. B. (1971). Signal Detectability and Medical Decision-Making. *Science*, 171:1217–1219.

Ma, L.-n., Zhang, J., Chen, H.-t., Zhou, J.-h., Ding, Y.-z., and Liu, Y.-s. (2011). An overview on ELISA techniques for FMD. *Virology Journal*, 8:419.

Marinier, E., Zaheer, R., Berry, C., Weedmark, K. A., Domaratzki, M., Mabon, P., Knox, N. C., Reimer, A. R., Graham, M. R., Chui, L., Patterson-Fortin, L., Zhang, J., Pagotto, F., Farber, J., Mahony, J., Seyer, K., Bekal, S., Tremblay, C., Isaac-Renton, J., Prystajecky, N., Chen, J., Slade, P., and Van Domselaar, G. (2017). Neptune: a bioinformatics tool for rapid discovery of genomic variation in bacterial populations. *Nucleic Acids Research*, 45:e159–e159.

Martinez, A. W., Phillips, S. T., and Whitesides, G. M. (2008). Three-dimensional microfluidic devices fabricated in layered paper and tape. *Proceedings of the National Academy of Sciences*, pages 19606–19611.

Menzel, P., Ng, K. L., and Krogh, A. (2016). Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature Communications*, 7:11257.

Mikkelsen, T., Alexandersen, S., Astrup, P., Champion, H. J., Donaldson, A. I., Dunkerley, F. N., Gloster, J., Sørensen, J. H., and Thykier-Nielsen, S. (2003). Investigation of airborne foot-and-mouth disease virus transmission during low-wind conditions in the early phase of the UK 2001 epidemic. *Atmos. Chem. Phys.*, 3:2101–2110.

Miller, M. B. and Tang, Y.-W. (2009). Basic Concepts of Microarrays and Potential Applications in Clinical Microbiology. *Clinical Microbiology Reviews*, 22:611–633.

Mohapatra, J. K., Sanyal, A., Hemadri, D., Tosh, C., Biswas, S., Knowles, N. J., Rasool, T. J., Bandyopadhyay, S. K., and Pattnaik, B. (2008). Comparative genomics of serotype Asia 1 foot-and-mouth disease virus isolates from India sampled over the last two decades. *Virus Research*, 136:16–29.

Moniwa, M., Embury-Hyatt, C., Zhang, Z., Hole, K., Clavijo, A., Copps, J., and Alexandersen, S. (2012). Experimental Foot-and-Mouth Disease Virus Infection in White Tailed Deer. *Journal of Comparative Pathology*, 147:330–342.

Morioka, K., Fukai, K., Yoshida, K., Kitano, R., Yamazoe, R., Yamada, M., Nishi, T., and Kanno, T. (2015). Development and Evaluation of a Rapid Antigen Detection and Serotyping Lateral Flow Antigen Detection System for Foot-and-Mouth Disease Virus. *PLOS ONE*, 10:e0134931.

Mort, M., Convery, I., Baxter, J., and Bailey, C. (2005). Psychosocial effects of the 2001 UK foot and mouth disease epidemic in a rural population: qualitative diary based study. *BMJ*, 331:1234–0.

Naccache, S. N., Federman, S., Veeeraraghavan, N., Zaharia, M., Lee, D., Samayoa, E., Bouquet, J., Greninger, A. L., Luk, K.-C., Enge, B., Wadford, D. A., Messenger, S. L., Genrich, G. L., Pellegrino, K., Grard, G., Leroy, E., Schneider, B. S., Fair, J. N., Martínez, M. A., Isa, P., Crump, J. A., DeRisi, J. L., Sittler, T., Hackett, J., Miller, S., and Chiu, C. Y. (2014). A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. *Genome Research*.

Officer, K., Lan, N. T., Wicker, L., Hoa, N. T., Weegenaar, A., Robinson, J., Ryoji, Y., and Loukopoulos, P. (2014). Foot-and-mouth disease in Asiatic black bears (Ursus thibetanus). *Journal of Veterinary Diagnostic Investigation*, 26:705–713.

OIE/FAO Foot-and-Mouth Disease Reference Laboratory Network (2016). Annual Report 2016. `http://www.wrlfmd.org/ref_labs/ref_lab_reports/OIE-FAO%20FMD%20Ref%20Lab%20Network%20Report%202016.pdf`. Accessed: 2018-02-03.

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., and Phillippy, A. M. (2016). Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology*, 17:132.

Ortega Polo, R., Hodko, D., Zhang, Z., King, D., Furukawa-Stoffer, T., Ohene-Adjei, S., Burton-Hughes, K., Fisher, M., Buchanan, C., Ambagala, A., and Lung, O. (2013).

Electronic Microarray Assay for Detection and Serotyping of Foot-and-Mouth Disease Virus. American Society for Virology 32nd Annual Meeting, Pennsylvania State University.

Orton, R., Gu, Q., Hughes, J., Maabar, M., Modha, S., Vattipally, S., and Wilkie, G. (2016). Bioinformatics tools for analysing viral genomic data. *Revue Scientifique et Technique - Office International des Epizooties*, 35:271–285.

Otomaru, K., Naoi, Y., Haga, K., Omatsu, T., Uto, T., Koizumi, M., Masuda, T., Yamasato, H., Takai, H., Aoki, H., Tsuchiaka, S., Sano, K., Okazaki, S., Katayama, Y., Oba, M., Furuya, T., Shirai, J., Katayama, K., Mizutani, T., and Nagai, M. (2016). Detection of novel kobu-like viruses in Japanese black cattle in Japan. *The Journal of Veterinary Medical Science*, 78:321–324.

Ounit, R., Wanamaker, S., Close, T. J., and Lonardi, S. (2015). CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics*, 16:236.

Palacios, G., Quan, P.-L., Jabado, O. J., Conlan, S., Hirschberg, D. L., Liu, Y., Zhai, J., Renwick, N., Hui, J., and Hegyi, H. (2007). Panmicrobial oligonucleotide array for diagnosis of infectious diseases. *Emerging Infectious Diseases*, 13:73.

Paton, D. J., McGoldrick, A., Greiser-Wilke, I., Parchariyanon, S., Song, J. Y., Liou, P. P., Stadejek, T., Lowings, J. P., Björklund, H., and Belák, S. (2000). Genetic typing of classical swine fever virus. *Veterinary Microbiology*, 73:137–157.

Paton, D. J., Valarcher, J.-F., Bergmann, I., Matlho, O. G., Zakharov, V. M., Palma, E. L., and Thomson, G. (2005). Selection of foot and mouth disease vaccine strains : a review. *Revue Scientifique et Technique - Office International des Epizooties*, 24:981–993.

Pickett, B. E., Liu, M., Sadat, E. L., Squires, R. B., Noronha, J. M., He, S., Jen, W., Zaremba, S., Gu, Z., Zhou, L., Larsen, C. N., Bosch, I., Gehrke, L., McGee, M., Klem, E. B., and Scheuermann, R. H. (2013). Metadata-driven comparative analysis tool for sequences (meta-CATS): An automated process for identifying significant sequence variations that correlate with virus attributes. *Virology*, 447:45–51.

Pickett, B. E., Sadat, E. L., Zhang, Y., Noronha, J. M., Squires, R. B., Hunt, V., Liu, M., Kumar, S., Zaremba, S., Gu, Z., Zhou, L., Larson, C. N., Dietrich, J., Klem, E. B., and Scheuermann, R. H. (2012). ViPR: an open bioinformatics database and analysis resource for virology research. *Nucleic Acids Research*, 40:D593–D598.

Pierce, K., Mistry, R., Reid, S., Bharya, S., Dukes, J., Hartshorn, C., King, D., and Wangh, L. (2010). Design and optimization of a novel reverse transcription linear-after-the-exponential PCR for the detection of foot-and-mouth disease virus. *Journal of Applied Microbiology*, 109:180–189.

Pinto, A. A. (2004). Foot-and-Mouth Disease in Tropical Wildlife. *Annals of the New York Academy of Sciences*, 1026:65–72.

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, , Carter, L. L., Doerrbecker, J., Enkirch, T., Dorival, I. G., 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, O., Faye, O., Magassouba, N., Williams, C. V., Amburgey, V., Winona, L., Davis, E., Gerlach, J., Washington, F., Monteil, V., Jourdain, M., Bererd, M., Camara, A., Somlare, H., Camara, A., 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., J. Turner, D., Pollakis, G., Hiscox, J. A., Matthews, D. A., Shea, M. K. O., Johnston, A. M., Wilson, D., Hutley, E., Smit, E., Di Caro, A., Wölfel, R., Stoecker, K., Fleischmann, E., Gabriel, M., Weller, S. A., Koivogui, L., Diallo, B., Keïta, S., Rambaut, A., Formenty, P., Günther, S., and Carroll, M. W. (2016). Real-time, portable genome sequencing for Ebola surveillance. *Nature*, 530(7589):228–232.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reeve, R., Blignaut, B., Esterhuysen, J. J., Opperman, P., Matthews, L., Fry, E. E., de Beer, T. A. P., Theron, J., Rieder, E., Vosloo, W., O'Neill, H. G., Haydon, D. T., and Maree, F. F. (2010). Sequence-Based Prediction for Vaccine Strain Selection and Identification of Antigenic Variability in Foot-and-Mouth Disease Virus. *PLoS Computational Biology*, 6:e1001027.

Reid, S. M., Ferris, N. P., Hutchings, G. H., De Clercq, K., Newman, B. J., Knowle, N. J., and Samuel, A. R. (2001). Diagnosis of foot-and-mouth disease by RT-PCR: use of phylogenetic data to evaluate primers for the typing of viral RNA in clinical samples. *Archives of Virology*, 146:2421–2434.

Reid, S. M., Ferris, N. P., Hutchings, G. H., Samuel, A. R., and Knowles, N. J. (2000). Primary diagnosis of foot-and-mouth disease by reverse transcription polymerase chain reaction. *Journal of Virological Methods*, 89:167–176.

Rodríguez, A., Martínez-Salas, E., Dopazo, J., Dávila, M., Sáiz, J. C., and Sobrino, F. (1992). Primer design for specific diagnosis by PCR of highly variable RNA viruses: Typing of foot-and-mouth disease virus. *Virology*, 189:363–367.

Rotmistrovsky, K., Jang, W., and Schuler, G. D. (2004). A web server for performing electronic PCR. *Nucleic Acids Research*, 32:W108–W112.

Sabalza, M., Yasmin, R., Barber, C. A., Castro, T., Malamud, D., Kim, B. J., Zhu, H., Montagna, R. A., and Abrams, W. R. (2018). Detection of Zika virus using reverse-transcription LAMP coupled with reverse dot blot analysis in saliva. *PLOS ONE*, 13:e0192398.

Sachs, M. C. (2016). *plotROC: Generate Useful ROC Curve Charts for Print and Interactive Use*. R package version 2.0.1.

Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B., and Erlich, H. A. (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, 239:487–491.

SantaLucia, Jr, J. (2007). Physical Principles and Visual-OMP Software for Optimal PCR Design. In Yuryev, A., editor, *PCR Primer Design*, Methods in Molecular Biology™, pages 3–33. Humana Press. DOI: 10.1007/978-1-59745-528-2_1.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science*, 270:467–470.

Schuler, G. D. (1997). Sequence Mapping by Electronic PCR. *Genome Research*, 7:541–550.

Sellers, R. F. and Daggupaty, S. M. (1990). The epidemic of foot-and-mouth disease in Saskatchewan, Canada, 1951-1952. *Canadian Journal of Veterinary Research*, 54:457–464.

Selvey, L. A., Wells, R. M., McCormack, J. G., Ansford, A. J., Murray, K., Rogers, R. J., Lavercombe, P. S., Selleck, P., and Sheridan, J. W. (1995). Infection of humans and horses by a newly described morbillivirus. *The Medical Journal of Australia*, 162:642–645.

Shendure, J. and Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, 26:1135–1145.

Smith, G. J. D., Vijaykrishna, D., Bahl, J., Lycett, S. J., Worobey, M., Pybus, O. G., Ma, S. K., Cheung, C. L., Raghwani, J., Bhatt, S., Peiris, J. S. M., Guan, Y., and Rambaut, A. (2009). Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature*, 459:1122–1125.

Sobrino, F., Dávila, M., Ortín, J., and Domingo, E. (1983). Multiple genetic variants arise in the course of replication of foot-and-mouth disease virus in cell culture. *Virology*, 128:310–318.

Song, J., Doggett, N., Wren, M., Burr, T., Fenimore, P. W., Hatcher, E. L., Bruno, W. J., Li, P.-E., Stubben, C., and Wolinsky, M. (2015). Development of Forensic Assay Signatures for Ebolaviruses. *Journal of Forensic Sciences*, 60:315–325.

Spizz, G., Young, L., Yasmin, R., Chen, Z., Lee, T., Mahoney, D., Zhang, X., Mouchka, G., Thomas, B., Honey, W., Roswech, T., McGuire, C., Montagna, R., and Zhou, P. (2012). Rheonix CARD® Technology: An Innovative and Fully Automated Molecular Diagnostic Device. *Point of Care*, 11:42–51.

Sugiura, K., Ogura, H., Ito, K., Ishikawa, K., Hoshino, K., and Sakamoto, K. (2001). Eradication of foot and mouth disease in Japan. *Revue Scientifique et Technique - Office International des Epizooties*, 20:701–711.

Theuns, S., Vanmechelen, B., Bernaert, Q., Deboutte, W., Vandenhole, M., Beller, L., Matthijnssens, J., Maes, P., and Nauwynck, H. J. (2018). Nanopore sequencing as a revolutionary diagnostic tool for porcine viral enteric disease complexes identifies porcine kobuvirus as an important enteric virus. *Scientific Reports*, 8(1):9830.

Thomson, G. R., Vosloo, W., and Bastos, A. D. S. (2003). Foot and mouth disease in wildlife. *Virus research*, 91:145–161.

Walsh, A. M., Crispie, F., O'Sullivan, O., Finnegan, L., Claesson, M. J., and Cotter, P. D. (2018). Species classifier choice is a key consideration when analysing low-complexity food microbiome data. *Microbiome*, 6:50.

Wang, D., Coscoy, L., Zylberberg, M., Avila, P. C., Boushey, H. A., Ganem, D., and DeRisi, J. L. (2002). Microarray-based detection and genotyping of viral pathogens. *Proceedings of the National Academy of Sciences*, 99:15687–15692.

Wang, J., Moore, N. E., Deng, Y.-M., Eccles, D. A., and Hall, R. J. (2015a). MinION nanopore sequencing of an influenza genome. *Virology*, page 766.

Wang, Y., Yang, Q., and Wang, Z. (2015b). The evolution of nanopore sequencing. *Frontiers in Genetics*, 5.

Watson, M., Dukes, J., Abu-Median, A.-B., King, D. P., and Britton, P. (2007). DetectiV: visualization, normalization and significance testing for pathogen-detection microarray data. *Genome Biology*, 8:R190.

Weaver, G. V., Domenech, J., Thiermann, A. R., and Karesh, W. B. (2013). Foot and mouth disease: a look from the wild side. *Journal of Wildlife Diseases*, 49:759–785.

Wetterstrand, K. (2017). DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). https://www.genome.gov/27541954/DNA-Sequencing-Costs-Data. Accessed: 2018-01-26.

Whiting, T. L. (2003). Foreign animal disease outbreaks, the animal welfare implications for Canada: Risks apparent from international experience. *The Canadian Veterinary Journal*, 44:805–815.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wickham, H. (2014). Tidy data. *The Journal of Statistical Software*, 59.

Wickham, H. (2016). *tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions*. R package version 0.4.1.

Wilke, C. O. (2017). *ggridges: Ridgeline Plots in 'ggplot2'*. R package version 0.4.1.

Wilson, M. R., Naccache, S. N., Samayoa, E., Biagtan, M., Bashir, H., Yu, G., Salamat, S. M., Somasekar, S., Federman, S., Miller, S., Sokolic, R., Garabedian, E., Candotti, F., Buckley, R. H., Reed, K. D., Meyer, T. L., Seroogy, C. M., Galloway, R., Henderson, S. L., Gern, J. E., DeRisi, J. L., and Chiu, C. Y. (2014). Actionable Diagnosis of Neuroleptospirosis by Next-Generation Sequencing. *New England Journal of Medicine*, 370:2408–2417.

Wood, D. E. and Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, 15:R46.

World Organisation For Animal Health (2004). Resolutions adopted by the International Committee of the OIE during its 72nd General Session. In *72nd General Session*, Paris. http://www.oie.int/fileadmin/Home/eng/Animal_Health_in_the_World/docs/pdf/A_RESO_2004_WP.pdf.

World Organisation for Animal Health (2012). Chapter 2.1.5: Foot and mouth disease. In *Manual of Diagnostic Tests and Vaccines for Terrestrial Animals*.

World Organisation For Animal Health (2017a). OIE-Listed diseases, infections and infestations in force in 2017. http://www.oie.int/animal-health-in-the-world/oie-listed-diseases-2017/. Accessed: 2017-01-17.

World Organisation For Animal Health (2017b). Old Classification of Diseases Notifiable to the OIE - List A. http://www.oie.int/en/animal-health-in-the-world/the-world-animal-health-information-system/old-classification-of-diseases-notifiable-to-the-oie-list-a/. Accessed: 2016-12-20.

World Organisation For Animal Health (2018). Old Classification of Diseases Notifiable to the OIE - List B. http://www.oie.int/animal-health-in-the-world/the-world-animal-health-information-system/old-classification-of-diseases-notifiable-to-the-oie-list-b/. Accessed: 2018-02-03.

Wright, C. F., Morelli, M. J., Thébaud, G., Knowles, N. J., Herzyk, P., Paton, D. J., Haydon, D. T., and King, D. P. (2011). Beyond the Consensus: Dissecting Within-Host Viral Population Diversity of Foot-and-Mouth Disease Virus by Using Next-Generation Genome Sequencing. *Journal of Virology*, 85:2266–2275.

Xi, J., Guo, H., Feng, Y., Xu, Y., Shao, M., Su, N., Wan, J., Li, J., and Tu, C. (2012). Differentiation of the Seven Major Lyssavirus Species by Oligonucleotide Microarray. *Journal of Clinical Microbiology*, 50:619–625.

Xu, W., Berhane, Y., Dubé, C., Liang, B., Pasick, J., VanDomselaar, G., and Alexandersen, S. (2016). Epidemiological and Evolutionary Inference of the Transmission Network of the 2014 Highly Pathogenic Avian Influenza H5n2 Outbreak in British Columbia, Canada. *Scientific Reports*, 6:30858.

Yang, M., Caterer, N. R., Xu, W., and Goolia, M. (2015). Development of a multiplex lateral flow strip test for foot-and-mouth disease virus detection using monoclonal antibodies. *Journal of Virological Methods*, 221:119–126.

Yang, M., Goolia, M., Xu, W., Bittner, H., and Clavijo, A. (2013). Development of a quick and simple detection methodology for foot-and-mouth disease virus serotypes O, A and Asia 1 using a generic RapidAssay Device. *Virology Journal*, 10:125.

Yang, P. C., Chu, R. M., Chung, W. B., and Sung, H. T. (1999). Epidemiological characteristics and financial costs of the 1997 foot-and-mouth disease epidemic in Taiwan. *Veterinary Record*, 145:731–734.

Yasmin, R., Zhu, H., Chen, Z., and Montagna, R. A. (2016). A modifiable microarray-based universal sensor: providing sample-to-results automation. *Heliyon*, 2:e00179.

Yoshida, C. E., Kruczkiewicz, P., Laing, C. R., Lingohr, E. J., Gannon, V. P. J., Nash, J. H. E., and Taboada, E. N. (2016). The Salmonella In Silico Typing Resource (SISTR): An Open Web-Accessible Tool for Rapidly Typing and Subtyping Draft Salmonella Genome Assemblies. *PLOS ONE*, 11:e0147101.

You, Y. (2006). Design of LNA probes that improve mismatch discrimination. *Nucleic Acids Research*, 34:e60–e60.

Zaharia, M., Bolosky, W. J., Curtis, K., Fox, A., Patterson, D., Shenker, S., Stoica, I., Karp, R. M., and Sittler, T. (2011). Faster and More Accurate Sequence Alignment with SNAP. *arXiv:1111.5572 [cs, q-bio]*. arXiv: 1111.5572.

Zhang, Y., Aevermann, B. D., Anderson, T. K., Burke, D. F., Dauphin, G., Gu, Z., He, S., Kumar, S., Larsen, C. N., Lee, A. J., Li, X., Macken, C., Mahaffey, C., Pickett, B. E., Reardon, B., Smith, T., Stewart, L., Suloway, C., Sun, G., Tong, L., Vincent, A. L., Walters, B., Zaremba, S., Zhao, H., Zhou, L., Zmasek, C., Klem, E. B., and Scheuermann, R. H. (2016). Influenza Research Database: An integrated bioinformatics resource for influenza virus research. *Nucleic Acids Research*, page gkw857.

Zhao, Y., Tang, H., and Ye, Y. (2012). RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics*, 28(1):125–126.

Zweig, M. H. and Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, 39:561–577.

Appendix A

Supplementary methods from Chapter 2

A.1 Design of synthetic constructs

The work described in this section was performed by staff of the Lethbridge Laboratory of the National Centres for Animal Disease (CFIA), in particular by Kimberley Burton Hughes (now with Agriculture and Agri-Food Canada). Nineteen synthetic DNA constructs were synthesized by Integrated DNA Technologies (IDT; Coralville, Iowa, USA) by using pGEM3Zf(+) plasmids (Promega P2271) as vectors for the cloning of the 1.2 Kb inserts (Lung et al., 2011). The constructs were designed using sequences from vaccine strains and representative strains for each serotype after reviewing the quarterly reports of the World Reference Laboratory for FMD where recommendations of FMDV strains to be added to antigen banks are often included (The Pirbright Institute, Pirbright, Surrey, UK). In some cases, the sequences of certain strains that are available in Genbank are partial and belong to a smaller portion of the target region (600 bp). In other cases, no sequences were available in GenBank for the desired strain. In both types of those cases, the most closely related complete genome sequences available were used to generate backbones to complete to 1.2Kb sequences.

A.2 Viral propagation and viral RNA extraction

The work described in this section was performed by staff of the vesicular disease unit of the National Centre for Foreign Animal Disease (CFIA; Winnipeg, Manitoba) and Lethbridge Laboratory staff in Winnipeg. The twenty-three viruses used for initial

laboratory validation of the electronic microarray assay were propagated in primary lamb kidney (PLK) cells. Viral RNA was extracted using TriPure reagent (Roche Diagnostics Canada), and the RNA pellets were suspended in 25μL of DEPC-treated water and stored at $-80\,^{\circ}$C. Total RNA was extracted from the FMDV-infected cells as described by Lung et al. (2011).