UNIVERSITEIT GENT

FACULTEIT DIERGENEESKUNDE
approved by EAEVE

Genome sequencing by random priming methods for viral identification

Rosseel Toon

Dissertation submitted in fulfillment of the requirements for the degree of Doctor of Philosophy (PhD) in Veterinary Sciences, Faculty of Veterinary Medicine, Ghent University, 2015

Promotors:

Dr. Steven Van Borm

Prof. Dr. Hans Nauwynck

CODA - CERVA

“The real voyage of discovery consist not in seeking new landscapes, but in having new eyes”

Marcel Proust, French writer, 1923

Table of contents

Table of contents 1

List of abbreviations 3

Chapter 1 General introduction 5

1. Viral diagnostics and genomics 7
2. The DNA sequencing revolution 12
    - 2.1. Classical Sanger sequencing 12
    - 2.2. Next-generation sequencing 16
3. The viral metagenomic workflow 24
    - 3.1. Sample preparation 25
    - 3.2. Sequence independent amplification 27
    - 3.3. DNA sequencing 51
    - 3.4. Bioinformatic data analysis 52
    - 3.5. Follow-up 55
4. The challenges 56

Chapter 2 Aims of the thesis 69

Chapter 3 Implementation and Case studies 73

Chapter 3.1 Implementation of a viral metagenomic workflow 75

Chapter 3.2 Identification and complete genome sequencing of paramyxoviruses in mallard ducks (*Anas platyrhynchos*) using random access amplification and next generation sequencing technologies 81

Chapter 3.3 Phylogeographic analysis of avian influenza viruses isolated from *Charadriiformes* in Belgium confirms intercontinental reassortment in gulls 105

Chapter 3.4 What's in a strain? Viral metagenomics identifies genetic variation and contaminating circoviruses in laboratory isolates of pigeon paramyxovirus type 1 127

Chapter 3.5 DNase SISPA-next generation sequencing confirms Schmallenberg virus in Belgian field samples and identifies genetic variation in Europe 147

Chapter 3.6 False-positive results in metagenomic virus discovery, a strong case for follow-up diagnosis 161

Chapter 4 Fine-tuning of methodology 175

Chapter 4.1 The origin of biased sequence depth in sequence-independent nucleic acid amplification and optimization for efficient massive parallel sequencing 177

Chapter 4.2 Evaluation of convenient pretreatment protocols for RNA virus metagenomics in serum and tissue samples 203

Chapter 5 General discussion 229

Summary 261

Samenvatting 267

Bibliography 275

Curriculum Vitae 281

Dankwoord 285

List of abbreviations

A	Adenine
aa	Amino acids
AFLP	Amplified fragment length polymorphism
AI(V)	Avian influenza A (virus)
APMV	Avian paramyxovirus
BLAST	Basic Local Alignment Search Tool
bp	Bases or base pairs
BVDV	Bovine viral diarrhea virus
C	Cytosine
CCD	Charge-coupled device
cDNA	Complementary DNA
CDS	Coding sequence
Contigs	Contiguous overlapping sets of sequences
Cp	Crossing point value
CPE	Cytopathic effect
Ct	Threshold cycle value
Da	Dalton (molecular weight unit)
ddNTP	2’,3’-dideoxynucleotides (ddATP, ddGTP, ddCTP, ddTTP)
DNA	Deoxyribonucleic acid
DNase	Deoxyribonuclease
dNTP	Deoxynucleoside triphosphates (dATP, dGTP, dCTP, dTTP)
DOP-PCR	Degenerate-oligonucleotide primer PCR
dsDNA	Double stranded DNA
dsRNA	Double stranded RNA
EDTA	Ethylenediaminetetraacetic acid
EID50	50 percent embryo infectious dose
ELISA	Enzyme-linked immunosorbent assay
emPCR	Emulsion PCR
G	Guanine
HA	Hemagglutination or hemagglutinin gene
HI	Hemagglutination inhibition
HIV	Human immunodeficiency virus
HPAI	High-pathogenic avian influenza virus
ICPI	Intracerebral pathogenicity index
IQR	Interquartile range
LASL	Linker Amplified Shotgun Library
LPAI	Low-pathogenic avian influenza virus
MDA	Multiple displacement amplification

MID	Multiplex identifier
N	Random nucleotide (A, C, G or T)
NA	Neuraminidase
ND(V)	Newcastle disease (virus)
NGS	Next-generation DNA sequencing
nt	Nucleotides
ORF	Open reading frame
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
PEG	Polyethylene glycol
PGM	Personal Genome Machine
PHV	Parvovirus-like hybrid virus
PiCV	Pigeon circovirus
PPMV1	Pigeon type 1 paramyxoviruses
(q)[RT]-PCR	(quantitative real-time) [reverse transcription] PCR reaction
pWGA	Primase-based whole genome amplification
QV	Quality score
RACE	Rapid Amplification of cDNA Ends
RCA	Rolling circle amplification
RNA	Ribonucleic acid
RNase	Ribonuclease
rPCR	Random PCR
rRNA	Ribosomal RNA
RT	Reverse transcription
SARS	Severe Acute Respiratory Syndrome
SBV	Schmallenberg virus
SE	southeast
SISPA	Sequence independent single primer amplification
SMRT	Single molecule real time sequencing
SPF	Specific-pathogen-free
SPIA	Single primer isothermal amplification
ssDNA	Single stranded DNA
ssRNA	Single stranded RNA
T	Thymine
TCID	Tissue culture infective dose
U	Units of polymerase
v/v	Volume concentration (% volume/volume)
VIDISCA	Virus-Discovery-cDNA-Amplified fragment length polymorphism
VIDISCA-454	VIDISCA combined with 454 pyrosequencing
WGA	Whole genome amplification
WTA	Whole transcriptome amplification

CHAPTER 1

General introduction

1. Viral diagnostics and genomics

We live in a globalized society where people move increasingly between countries and continents, and livestock animals are farmed and traded extensively. Emerging and reemerging infectious diseases are a constant threat to livestock and the human population. A recent example is the current Ebola virus disease epidemic in West Africa which started in March 2014. This virus has already affected more than 22,000 people, and killed over 9,000 (according to data from World Health Organization on 8 February 2015). Furthermore, recent reviews indicate that the majority of emerging human infections are zoonotic. Therefore, contact with livestock or wildlife increases the probability of infections threatening human populations and individuals lacking immunity. About 15 million (>25 %) of 57 million annual human deaths worldwide are estimated to be linked to infectious diseases [1]. Moreover, livestock diseases may result in significant economic losses and socio-economic consequences. Viral diseases are often spread and transmitted by vectors such as bloodsucking mosquitoes, ticks and wildlife animals. Sometimes, an intermediate host like a domestic animal is the link between viral circulations in wildlife and humans. For instance, some human infections originating from bats, such as Nipah, Hendra, SARS and Ebola viral infections, may involve intermediate amplification in pigs, horses, civets and primates respectively [2]. Complex pathogen lifecycles complicate the control of diseases. Rapid diagnosis of infectious diseases is essential in order to take appropriate action to control livestock and human diseases. Characterization of new pathogens will help to understand diseases and is an important step in the development of vaccines and better diagnostic tests.

Viruses are the most abundant infectious agents on the planet. In an apparently sterile environment like sea water, the number of virus particles is estimated at 10^6 to 10^9 particles per milliliter [3]. A virus needs living cells of other organisms to reproduce. They infect all life forms, including bacteria, plants, protista, fungi and animals. The species they infect and use to replicate is called a host. Morphologically speaking, viral particles, known as virions, consist of only 2 major parts: (1) genetic material and (2) a protein coat that protects the genetic material called capsid. Some viruses may also contain an envelope of lipids that surrounds the protein coat when they are outside host cells. The genetic material can be either deoxyribonucleic acid (DNA) or ribonucleic acid (RNA), linear or circular, single stranded and double stranded, segmented or non-segmented. Viruses are classified according to the type and organization of their genomes, viral replication strategy, and the morphological structure of the virion. In Figure 1 an overview is given of the morphologies and genome structures of the 26 virus families known to cause disease in animals and humans.

The detection and identification of viruses in clinical samples relies on a range of traditional techniques. Samples are examined for the presence of virus particles (e.g. by electron microscopy), virus antigens (e.g. by *ELISA*, the enzyme-linked immunosorbent assay*)* and/or viral nucleic acids (e.g. DNA microarrays, PCR, DNA sequencing). Virus isolation attempts the multiplication of viruses in cell culture, eggs or laboratory animals. Cell cultures are checked for lysis of cells caused by infecting viruses known as cytopathic effect (CPE). In most cases, viral infection is diagnosed by the detection of specific antiviral antibodies in the serum (i.e. the phase that remains after centrifugation of clotted blood) using reference antigens (e.g. inactivated viruses).

Rapid development of novel molecular techniques like (1) PCR and (2) DNA sequencing, has resulted in the strong advancement of disease diagnostics during the last decades. PCR allows the detection of virus-specific nucleic acids directly from the clinical samples in a very sensitive way. It has become the most widely used molecular diagnostic technique in clinical virology [4]. The technique amplifies a single or few copies of a stretch of DNA in an exponential manner, thereby generating thousands to millions of copies of a particular DNA sequence in a short period of time. A pair of primers (i.e. short DNA fragments which are complementary to target regions you want to amplify) along with a thermostable DNA polymerase, are the key components to enable selective and repeated amplification [5]. By inclusion of a step converting RNA into complementary DNA (cDNA), PCR can be adapted to detect viral RNA in a reverse transcription PCR reaction (RT-PCR). Using a mix of different primer pairs, several different target nucleic acid sequences can be detected simultaneously (multiplex PCR). Using fluorescently labeled short DNA fragments (probes) or double stranded DNA (dsDNA) intercalating dyes, PCR amplification can be monitored and quantified in real-time in a closed tube format. This allows highly sensitive, accurate and rapid detection of viral nucleic acids, whilst alleviating contamination issues (real-time PCR). Most PCR-based virus detection approaches use primers of which the sequence is complementary to conserved regions in the targeted virus genome, whereby at least partial sequence knowledge of the target viral genome is needed. Unfortunately, there is no gene which may serve as universal molecular marker for viruses as no single gene is common to all viral genomes. This is in contrast to, for instance the 16S ribosomal RNA (rRNA) gene which is used as universal amplification target for identification of nearly all bacteria species [6].

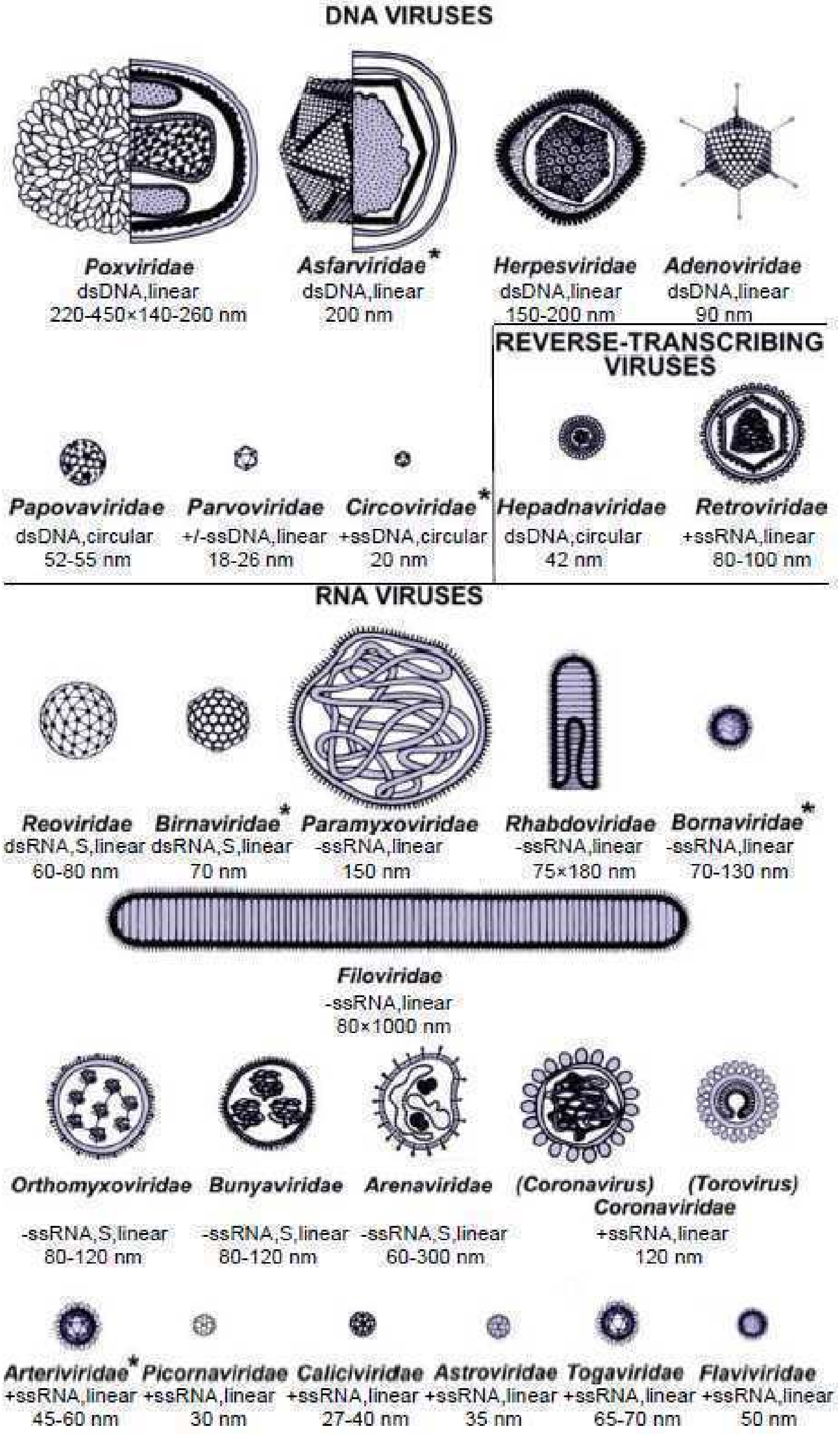


Figure 1: Morphology and genomic organization of virus families that include animal, zoonotic and human pathogens. *: viruses only known to infect animals, ds: double stranded, ss: single stranded, +: sense strand, -: antisense strand, S: segmented, nm: approximate dimensions of capsid/envelope expressed in nanometers (Reprinted from Murphy et al., 1999 [7] with kind permission from Elsevier; virion sizes are taken from http://viralzone.expasy.org).

DNA consists of a sequence of four nucleotides. Nucleotides are composed of a nitrogenous base, a five-carbon sugar and at least one phosphate group. The four possible nitrogenous bases are adenine (A), guanine (G), thymine (T), and cytosine (C). The process of determining the precise order of nucleotides within a DNA molecule is called DNA sequencing. The amount of sequenced viral nucleic acids has increased exponentially during the last 20 years (Figure 2). Currently there are over 1.8 million viral sequences available in public databases. An overview of the rapid evolution of DNA sequencing technologies and their applications is presented in section 2 of this introduction. The most obvious strategy for sequencing full viral genomes involves the design of overlapping PCR fragments (also known as amplicons) that span the entire genome followed by the targeted amplification. PCR primer design could be based on conserved regions of viral genomes, requiring prior knowledge of the genome sequence of the target pathogen. Conserved regions can be identified by aligning available sequences for a virus of interest from public databases. In the absence of perfect sequence conservation in a region, a primer containing some degenerated nucleotides may be designed (e.g. [8]). Specific viral PCR amplification is the most sensitive method for detecting and sequencing small amounts of viral nucleic acids. However, this method is not useful for obtaining genomic information from viruses with little or no available sequence data.

Correct diagnosis of viral infections is not possible in some instances. Some viruses are difficult to culture in existing cell culture or biological model systems. In other cases, the available specific diagnostic tests fail because an unexpected virus, new variant or totally new virus is involved. In addition, the viral replication biology, in particular that of RNA viruses, poses its own unique problems. The lack of proofreading mechanisms (i.e. when an incorrect base pair [bp] is incorporated during replication, proofreading polymerases have the ability to correct this error) provided by the viral RNA polymerase and a short generation time result in a very high mutation rate. Consequently, RNA viruses exist as a complex mix of differing genomes (a “swarm” of closely related viruses) within a single host, often termed as “quasispecies” (reviewed in [9]). Moreover, when a host cell is infected by more than one virus strain, recombination between genomic regions and reassortment of genomic segments (of segmented viral genomes) frequently occurs, enhancing genetic variation and viral evolution [10]. Thus, rapid virus evolution makes detection and control of virus infections even more challenging.

One promising solution for failing viral diagnostics is sequencing all genetic material directly in a clinical sample in a random manner, and subsequently analyzing the obtained sequencing data for viral sequences. This approach is often referred to as viral metagenomics (reviewed in [11, 12]) and has the benefit of not requiring prior knowledge about any virus that may be present in a sample. Besides virus identification, the viral metagenomic workflow has the ability to provide a detailed characterization of viruses, including the full genome sequence [13]. In the next section, a summary will be given of available DNA sequencing technologies on the market, and their applications in virology (section 2). In section 3, an overview will be given of frequently used steps to generate viral metagenomes (also known as viromes) in order to characterize new and known viruses from animal and human clinical samples. In the last section (4) of this general introduction, the remaining challenges of the viral metagenomic workflow will be listed for its application in viral diagnostics and genomics.

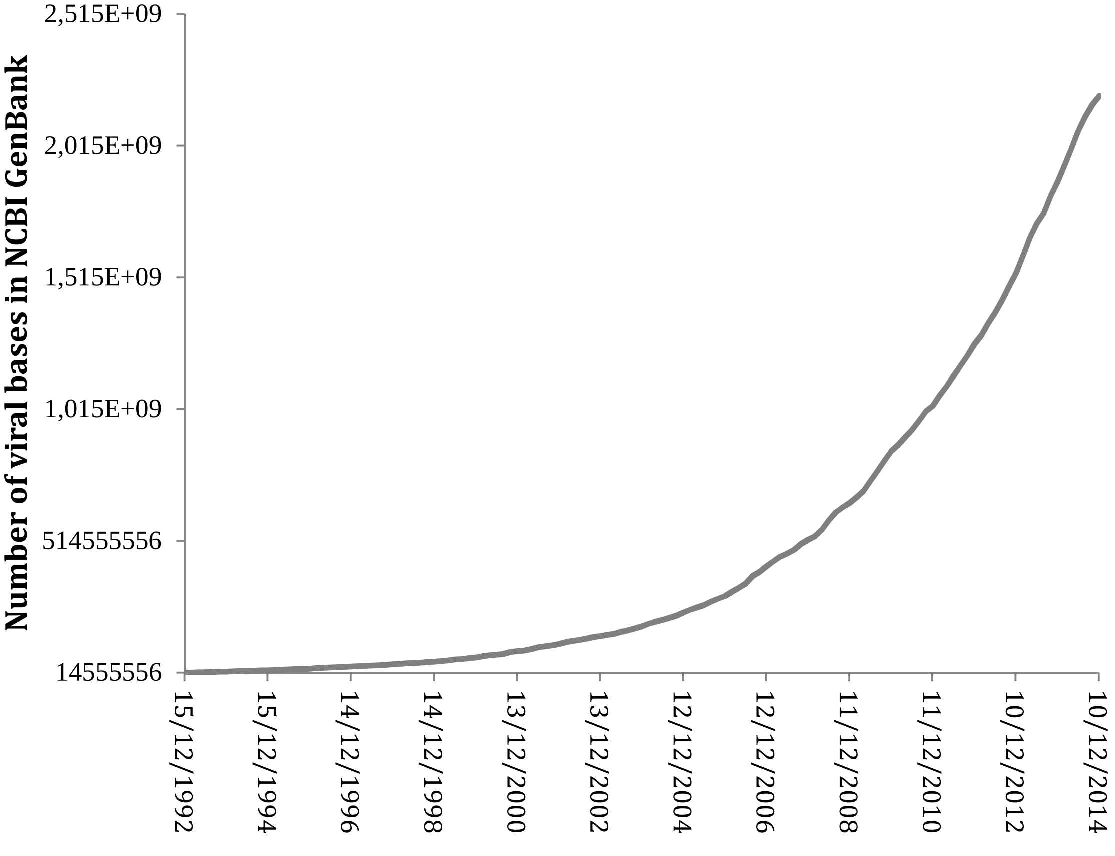


Figure 2: Number of bases from all viral DNA/RNA sequences in the NCBI genetic sequence database GenBank (extracted from the bimonthly release notes ftp://ftp.ncbi.nih.gov/genbank/release.notes/).

2. The DNA sequencing revolution

The first modern DNA sequencing technologies were developed in 1977, of which the Frederick Sanger chain-termination method was the most important one. For decades this technology has been the gold standard for DNA sequencing. Since 2005, various “next-generation” DNA sequencing (NGS) technologies were developed at a fast rate, increasing the sequencing throughput enormously and reducing the costs dramatically.

2.1 Classical Sanger sequencing

2.1.1 Method

In 1977, Sanger and coworkers developed a sequencing method that was initially known as the chain-termination or dideoxynucleotide method [14]. Briefly, a radioactive-labelled (^{32}P) primer was annealed to a specific known region on the template DNA, which provided a starting point for DNA synthesis. Subsequently complementary strand synthesis was performed in the presence of DNA polymerase, deoxynucleoside triphosphates (dNTP) and a modified 2’,3’-dideoxy nucleotides (ddNTP). Since the ddNTP lack a 3’-OH group necessary for incorporation of a next base, the reaction is terminated with the incorporation of a ddNTP into the growing chain. This method was performed in four separate chain-termination reactions, each containing one of the four terminators (ddATP, ddGTP, ddCTP, ddTTP). All the generated fragments had the same 5’ end, whereas the residue at the 3’ end was determined by the ddNTP used in the reaction. After all four reactions were completed; the mixture of different-sized DNA fragments was separated by electrophoresis on a denaturing polyacrylamide gel in four lanes, allowing a size difference resolution of one nucleotide. The pattern of bands showed the distribution of the termination in the synthesized strand of DNA and the unknown sequence could be read by autoradiography. At that time, DNA sequencing was slow, laborious and radioactive reagents were required. Since this first version of the Sanger method, several adaptations have been implemented to improve the method. The major technological breakthroughs were:

- The use of fluorescent detection integrated into a partially automated platform. A different colored fluorophore label is used for each of four terminators (ddATP, ddGTP, ddCTP, ddTTP). This makes it possible to carry out the four termination reactions - for A, C, G and T - in a single tube, and separate the termination products in a single polyacrylamide gel lane. Sequence information is acquired directly by a computer when the bands on the gel pass in front of a fluorescent detector [15].

- Introduction of capillary gel electrophoreses (a miniaturized polyacrylamide gel) for separation and detection of DNA sequencing samples [16].
- Introduction of thermal cycle sequencing which is carried out in a similar way to PCR but uses just one primer [17]. Because there is only one primer, only one of the strands of the starting molecule is copied, and the product accumulates in a linear fashion, not exponentially as is the case in a traditional PCR.

Finally, the DNA fragments could be detected in real-time by a laser and raw sequencing traces were converted by base calling software into sequence reads with associated quality. This automatic capillary sequencing was later termed as “first-generation sequencing”. Nowadays, a 96-capillary machine can produce about half a million nucleotides (0.5 megabases [Mb]) per day.

2.1.2 Genome sequencing strategies

As only relatively short DNA strands (100 to 1,000 bp) can be sequenced with the Sanger method, longer DNA fragments must be subdivided into smaller fragments and subsequently re-assembled back into the original overall sequence. Two different strategies have been developed for the sequencing of whole (large) genomes: the “whole-genome shotgun” approach and the “clone contig” approach [18].

Whole genome shotgun sequencing

The genome is randomly broken (e.g. by sonication or nebulization) into short fragments of appropriate size for sequencing (Figure 3). Subsequently these DNA fragments are inserted into a cloning vector (e.g. plasmid or lambda phage) and molecularly cloned. End sequences of the clonal DNA inserts are obtained and software is used to identify overlaps between sequences. The resulting sequences are examined for overlaps and are used to build up the contiguous genome sequence (referred as *de novo* assembly). When the original genome is not fully covered yet, gaps can be filled by (1) sequencing more clones, (2) molecular cloning using a different vector, or (3) primer walking. The primer walking approach is quite time consuming and extensive primer design is needed. First, a sequencing reaction is performed using primers which bind to part of a known sequence. A second priming site is then chosen inside the newly generated sequence, following the same sequencing direction as the first one. In this manner, you “walk” along the strand in order to complete the sequence. The whole genome shotgun approach has been used for sequencing small genomes (e.g. viruses and bacteria) and in metagenomics. The shotgun approach is straightforward and does not require any prior knowledge of the genome sequence. However, the whole genome shotgun sequencing of viruses, possesses its own unique problems (see section 3, viral metagenomics).

Clone contig sequencing

The *de novo* assembly of shotgun sequences of more complex (e.g. eukaryotic) genomes can be very computationally demanding, and is complicated by the presence of repeated sequences (i.e. patterns of nucleic acids that occur in multiple copies throughout the genome). One solution is the clone contig sequencing approach (also known as hierarchical shotgun sequencing), where a low-resolution physical map of the genome is made prior to actual sequencing (Figure 3). The genome is first broken down into large fragments, referred to as “clone contigs”, which are cloned in a high-capacity cloning vector (i.e. large inserts possible, typically 100-200 kilobases [kb] each). Subsequently, overlaps between clone contigs and their relative orientation to the genome are identified and a low-resolution physical map is made. Each piece of cloned DNA is then sequenced by the random shotgun strategy and this sequence is placed at its appropriate position on the contig map in order to gradually build up the overall genome sequence. Although this method has clear advantages, the clone contig approach is much slower and more labor-intensive than the whole genome shotgun approach.

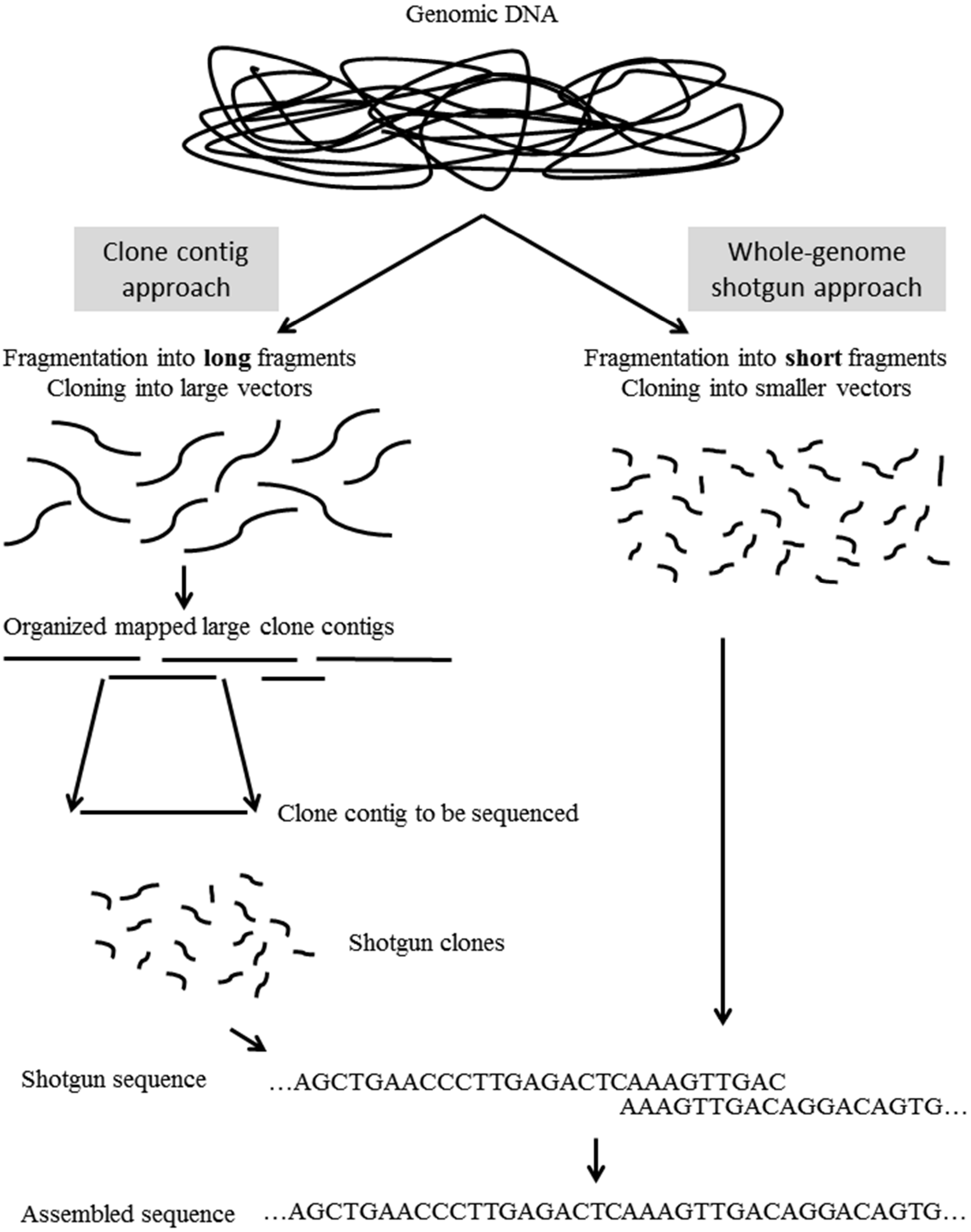

Genomic DNA
Clone contig approach
Whole-genome shotgun approach
Fragmentation into long fragments
Cloning into large vectors
Fragmentation into short fragments
Cloning into smaller vectors
Organized mapped large clone contigs
Clone contig to be sequenced
Shotgun clones
Shotgun sequence
…AGCTGAACCCTTGAGACTCAAAGTTGAC
AAAGTTGACAGGACAGTG…
Assembled sequence
…AGCTGAACCCTTGAGACTCAAAGTTGACAGGACAGTG…

Figure 3: Schematic representation of the "clone contig" and "whole genome shotgun" sequencing strategies. In the clone contig approach, larger fragments are first cloned into large-fragment cloning vectors. The genomic DNA fragments represented in the cloning library are then organized into a physical map and individual clones are selected and sequenced by the random shotgun strategy. In the whole genome shotgun strategy, the genomic DNA is directly broken into fragments of appropriate sequencing lengths, cloned into small-fragment cloning vectors and sequenced.

2.1.3 Applications

Sanger sequencing is used to sequence individual genes, larger genetic regions, full chromosomes or whole genomes. The first genome sequence ever determined was that of bacteriophage MS2 [19]. The Sanger method was also the necessary foundation for the development of PCR, which is up to now the most successful molecular diagnostic method [5] (reviewed in [20]). Using the shotgun approach, the first bacterial genome (*Haemophilus influenza*) was sequenced in 1995 [21], and more recently the first working draft sequence of the human genome [22, 23]. The whole Human Genome Project took more than a decade and involved multiple sequencing centers and cost over $3 billion. The final sequencing of this first draft sequence of the human genome is estimated to have cost $300 million and several years were needed to generate and process all of the sequencing data.

Sanger sequencing is used extensively by clinical laboratories and has been useful for a wide range of applications such as molecular cloning, breeding, finding pathogenic genes, and comparative and evolutionary studies. Also in virology it has become an important tool for virus identification and genome characterization, including pathotyping and tracing evolutionary relationships. Full viral genome sequencing often relies on extensive primer design. For instance, the molecular characterization of an uncommon paramyxovirus involved the use of degenerated primers (matching conserved regions) and a primer walking approach [24]. Whole genome shotgun sequencing of viruses is a real challenge due their short genome lengths and low amounts of viral nucleic acids compared to massive amounts of background (mainly host) nucleic acids. These virus-specific challenges are discussed in section 3.

2.2 Next-generation sequencing

Since the beginning of this PhD project, much has changed in the field of DNA sequencing. Automated capillary sequencing based on the Sanger method is widely accessible and provides high quality data. However, the application to projects such as whole genome sequencing of species with large genomes is expensive and time-consuming. Limitations in throughput, scalability and speed have motivated the development of alternative, post-Sanger sequencing technologies which are referred as “Next-Generation Sequencing” or NGS. NGS platforms provide unprecedented throughput, generating hundreds of gigabases (Gb) of data in a single experiment. Sequencing with NGS platforms is also known as “deep sequencing” because the total number of reads is often many times larger than the number of reads required to cover the complete length of the studied DNA sequence; thereby overlapping a certain position numerous times (referred to as coverage depth). Although the initial capital investment and cost per experiment remains high, the price per information unit (nucleotide) has been dramatically reduced in comparison with first-generation sequencing (Figure 4). Moreover, these technologies allow sequencing without prior knowledge of the complete DNA content in a sample (shotgun sequencing approach, but without the need for molecular cloning into vectors) whilst retaining the flexibility to allow targeted sequencing (e.g. PCR amplicons). A number of different NGS platforms are currently available, with each utilizing unique protocols and solutions for template preparation and sequencing. This has led to individual systems having their own strengths and limitations (reviewed in: [25-27]).

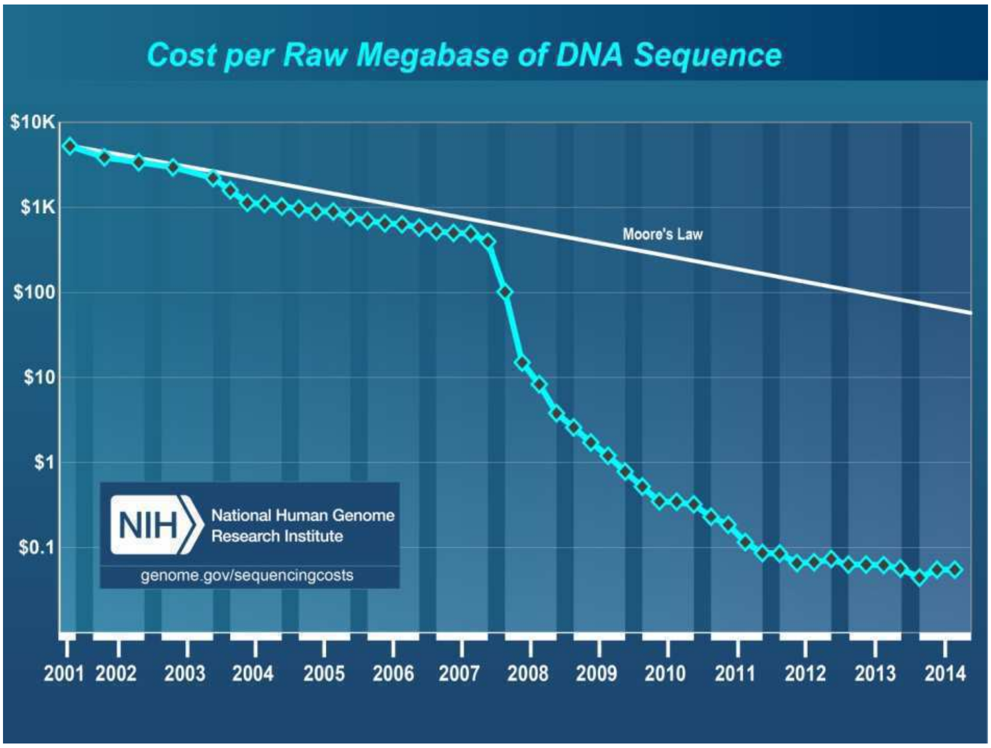


Figure 4: The average cost (in U.S. dollar, logarithmic scale) per 1000 bases sequenced since 2001 (http://www.genome.gov/sequencingcosts/).

2.2.1 Second-generation technologies

NGS platforms that require amplification of the template molecules prior to sequencing are referred to as “second-generation sequencing” platforms. Second-generation sequencing platforms vary in technology and chemistry used, but have the following properties in common:

- A DNA library is made from the sample. This library represents either all DNA in the sample without prior knowledge of the sequence or a targeted library using PCR amplification or alternative enrichment methods. As the DNA template needs to be quite short (ranging from 200-1,000 bp), a mechanical or enzymatic DNA fragmentation step is performed. Subsequently, adapter sequences, containing primer binding sites for amplification and sequencing, are joined to DNA molecules by ligation or amplification. Adapter sequences may include a barcode sequence (also known as index, MID or tag) that allows multiplexing of several samples in an experiment.

- Individual DNA molecules in each library are clonally amplified.
- Clonal DNAs are sequenced in parallel, resulting in hundreds of thousands of DNA sequence reads.

The second-generation sequencing platforms first emerged on the market with an emphasis on extreme high-throughput sequencing applications and were initially restricted to large genome sequencing centers or core facilities. These technologies use different detection principles including 454 pyrosequencing (454 Life Sciences, acquired by Roche, available since 2005, but planned to be discontinued by mid-2016 [28]), Illumina’s sequencing by synthesis (previously Solexa, available since 2007 [29]), SOLiD ligation based sequencing (Life Technologies, available since 2006), and since 2010, the Ion Torrent semiconductor sequencing technology [30].

- In **454 pyrosequencing** (http://www.454.com/), a single template molecule is annealed to a bead-bound oligonucleotide and is clonally amplified via emulsion PCR (emPCR). emPCR is a PCR reaction that occurs within aqueous microdroplets separated by oil enabling up to thousands of independent reations per microliter (µl) of volume. emPCR results in each bead having a homogeneous set of template molecules. Millions of beads are loaded together with sequencing reagents onto a flat “picotiter” plate which contains multiple wells and is designed to allow only a single bead in each well. All beads are then sequenced in parallel by pyrosequencing. When a polymerase-mediated incorporation event occurs, a chemiluminescent enzyme generates an observable light signal that is recorded by a CCD camera. Consecutive flows of A, C, G, and T reagents over the picotiter plate allow the determination of the sequence. Currently there are two platforms on the market using this technology, the GS FLX+ and the GS Junior+ system. The GS FLX+ system (developed from its ancestor GS20) can now generate around 700 Mb of sequence data in a day with read lengths of up to 1,000 bp. The GS Junior is essentially a smaller benchtop version generating around 70 Mb per run with read lengths of 700-800 bp and longer. 454 pyrosequencing has been the most commonly used technology for

NGS to date. Reagents are quite expensive and the technology shows a high error rate in homopolymer regions (i.e. three or more consecutive identical DNA bases), caused by accumulated light intensity variance [28]. Although long read lengths offer considerable advantages for some applications, this technology has the highest cost per nucleotide of information, forcing Roche to discontinue the technology from 2016.

- In **Illumina sequencing** (http://www.illumina.com/), individual library DNA fragments are captured on a solid glass surface and clonally amplified by bridge amplification. The library DNA fragments binds at both ends to oligo primers on the glass surface and successive round of PCR result in the generation of tiny islands or clusters of amplified molecules which serve as clones for subsequent sequencing using chain terminators similar to traditional Sanger sequencing. However, unlike the Sanger method, Illumina uses only fluorescently labelled reversible terminators, such that each base incorporation temporarily terminates the sequencing reaction. After imaging to determine which nucleotide is incorporated in each DNA clonal cluster, the dye is cleaved and more dye labelled terminators are added. The principle is called sequencing-by-synthesis. Currently three platforms are marketed: the HiSeq, MiSeq and NextSeq. Different read lengths may be chosen, from 36 bp to 2×300 bp on the MiSeq. HiSeq generates extreme throughput (ranging from 15 Gb to 1.8 Terabases [Tb] per run) and is most suitable for large sequencing centers. The MiSeq is a smaller benchtop version generating up to 15 Gb output per run. The NextSeq is a medium throughput platform (up to 120 Gb). At the moment Illumina is the dominant NGS technology on the market as it is a proven technology, with good error rates and acceptable operational costs.
- The **SOLiD** Sequencing by Oligonucleotide Ligation and Detection technology (http://www.lifetechnologies.com/) uses emPCR to generate clonal DNA fragments on beads. The beads are attached to a glass slide to allow sequencing by ligation and detection using fluorescently labelled di- or tri-base probes. Whilst this approach provides very high accuracy, the maximum read length is relatively short (75 bp). The SOLiD platforms have therefore been used mainly for applications that do not require *de novo* assembly of reads (i.e. assembling overlapping short reads back into larger fragments, also known as contigs), such as transcriptomics, epigenomics and resequencing of large mammalian genomes. There are currently two SOLiD platforms available, the 5500 and the 5500xl, with an output up to 320 Gb.
- **Ion Torrent** (http://www.lifetechnologies.com/) uses a sequencing strategy similar to 454 pyrosequencing, except that hydrogen ions (H+) are detected instead of

chemiluminescence signals. The hydrogen ion is detected by a semiconductor as a small change in pH. Lasers, cameras or fluorescent dyes are no longer needed which highly reduces the costs. However, the technology remains still sensitive to misreading the length of homopolymers. There are currently three different sequencing chips available for the Personal Genome Machine (PGM), the first platform for Ion Torrent sequencing, ranges in capacity from 30 Mb to 2 Gb (read length up to 400 bp). With a short run time and flexible capacity, the PGM represents an affordable and rapid benchtop system designed for small projects. Recently the Proton platform was released to enable higher throughput.

Over the last 5-7 years, all of the major platforms have made significant improvements, with notable advancements made in terms of protocol complexity, overall performance (including read length, fidelity, lower input DNA) and cost-efficiency. The development of smaller benchtop sequencers [31] have made the technology more accessible for use in routine microbiology laboratories. Whilst academic core facilities and commercial service providers have focused increasingly upon providing users with access to a range of the different sequencing technologies.

2.2.2 Third-generation technologies

In addition to the continuous improvement of existing second-generation sequencing platforms, newer methodologies are being developed. NGS platforms that do not require amplification of the template molecules prior to sequencing and directly sequence individual DNA molecules are referred to as single molecule sequencers or “third-generation” platforms (reviewed in: [27, 32]). These approaches promise additional advantages such as scalability, simplicity, long read length and low operational costs. As they do not require clonal amplification of template DNA molecules, potential errors associated with clonal amplification are avoided.

- **HeliScope** (developed by Helicos BioSciences [33]) was the first commercial available single molecule sequencer. The high cost of the instrument and the very short read lengths (average of 32 bp) resulted in limited use. Consequently, the instrument is no longer sold.
- Currently, **PacBio RS** (http://www.pacificbiosciences.com/) is the only third-generation platform available on the market (since 2011). Their SMRT sequencing technology determines sequences of long single DNA molecules in real-time. Unlike other NGS

technologies, it is the DNA polymerase that is immobilized on the bottom of a microcell. Fragmented dsDNA is ligated to hairpin adapters to create circular DNA. These are amplified linearly using primers complementary to the hairpin sequence and then captured by a single molecule of DNA polymerase and sequenced in the bottom of a well (referred to as “zero-mode waveguide”). The sequence of individual DNA strands can be determined because each dNTP has a unique fluorescent label (no chain terminators unlike Sanger and Illumina sequencing) that is detected immediately prior to being cleaved off during synthesis [34]. Following incorporation, the fluorescent label is cleaved and diffuses away, allowing the DNA polymerase to continue to incorporate multiple bases per second. Even though its throughput (around 90 Mb) is lower than most NGS platforms on the market, the PacBIO RS II still has several advantages that make it attractive for clinical laboratories, in particular for microbiology research. Sample preparation is fast, there is no introduction of amplification artefacts, run times are relatively short (finished within 1 or 2 hours), and read length is the greatest currently available (average circa 12,500 bp, using the latest chemistry sequences of up to 40,000 bp can be achieved). However, compared with other sequencing platforms, the PacBio RS has been reported to have the highest raw error rate. In addition, due to lack of clonal amplification, large amounts of input DNA material are required.

Other 3rd generation technologies are still under development [25] such as DNA sequencing in nanopores that offer the potential of simple, inexpensive, single-molecule sequencing in miniaturized or highly scalable devices [35]. Although substantial validation data is still required, these technologies have the potential to make NGS even more widely available for clinical purposes.

2.2.3 Applications in clinical virology

The rapid evolution in DNA sequencing has resulted in the exponential increase of genetic data in public databases, including viral sequence data (Figure 2). The NGS platforms were originally designed for high-throughput sequencing of large genome sequencing projects. In 2008, NGS made it possible to sequence a human genome for less than $1 million in only 2 months [36]. Nowadays, it is even possible to re-sequence a human genome for $1000 in only a few days with the Illumina sequencing HiSeqX Ten platform. By contrast, viral genomes have very small genomes. Therefore, sequencing of, for instance, a single viral genome using a full NGS run would make the project very expensive. Multiplexing possibilities which allow sequencing of different samples on the same run, and the development of smaller benchtop

NGS sequencers, are making this new technology more accessible for the average clinical virology laboratory.

NGS provides many new opportunities and is used increasingly to study etiology, genomes, evolution and outbreak management of human and animal infectious diseases as well as host-pathogen interactions (Figure 5; comprehensive reviews: [26, 37, 38]). NGS offers high sensititive diagnostic potential to detect the full spectrum of viruses, including unknown and unexpected viruses. For instance, metagenomics using 454 technology allowed the identification of a novel orthobunyavirus, subsequently named Schmallenberg virus (SBV), in an epidemiological cluster of diseased cattle in Germany [39]. The cattle showed symptoms of fever, decreased milk production, and diarrhea. The identified viral sequences were used to rapidly design targeted molecular tests that were used to confirm a clear association between the presence of the virus and affected animals. International adoption of these molecular tests identified a widespread occurrence of SBV in European countries (http://www.efsa.europa.eu/en/supporting/pub/429e.htm) and its detection in stillborn and malformed lambs [40, 41], as well as in insect vectors [42, 43]. The molecular tests were also helpful in targeting samples for isolation of the virus, which ultimately led to the development of vaccines [44-46]. This example shows the power of NGS to boost responsiveness to emerging diseases. Beside detection of candidate pathogens, NGS-based viral metagenomics often find a wealth of previously unrecognized viruses [47]. NGS is being used increasingly to explore the viral diversity in a wide range of environmental samples (determining what might be considered as “normal”) and in surveillance studies on vector-borne and zoonotic viruses (e.g. surveillance in mosquitos [48] or in bats [49]).

NGS provides the ability to sequence and compare multiple full genomes of distinct types and to identify important genetic differences between them [50]. In contrast to Sanger sequencing, one NGS read represents one DNA fragment which does not need to be the major variant. Consequently, NGS makes it possible to study variations (single nucleotide variations, deletions, insertions…) beyond consensus level with great accuracy. This is extremely useful for the study of RNA virus quasispecies. For example, Wright and colleagues investigated the genetic diversity and resulting quasispecies population after inoculation of foot-and-mouth disease virus into a single animal [51]. The researchers identified genetically distinct populations originating from different lesions. The study of the viral swarm within individual hosts has implications for understanding the evolutionary dynamics of viral populations under selection pressures, e.g. antiviral drugs or host immune response. This has been a particularly active field in human medicine, e.g. with regard to human immunodeficiency virus (HIV) antiviral drugs response, drug resistance, and viral tropism (reviewed in: [52-54]), as well as studies on human influenza A (e.g. [55]). Deep sequencing was also used for quality control of live-attenuated viral vaccines [56, 57]. In this context, the method allowed identification of mutations of vaccine strains, minority variants of vaccine strains and sequences of adventitious viruses in vaccine seeds or stocks [57].

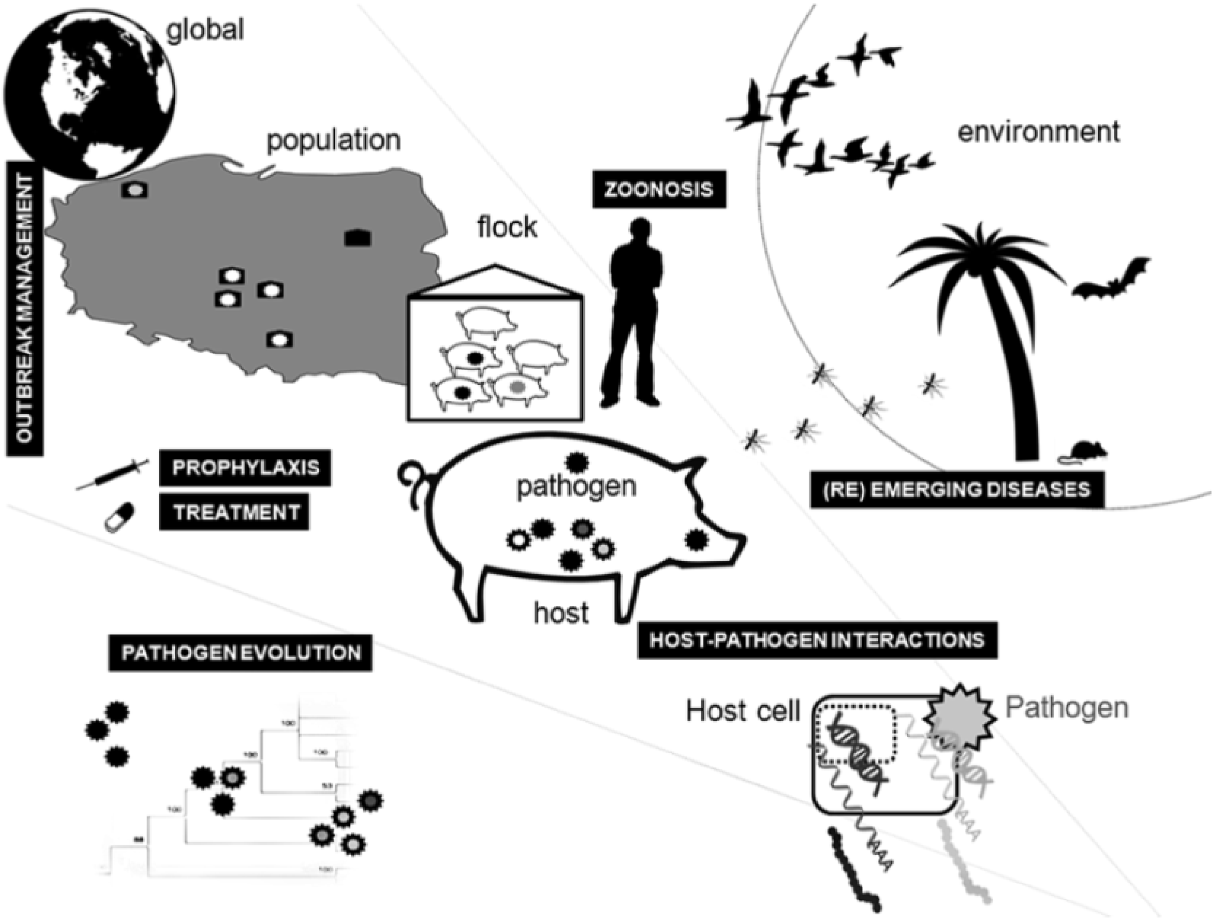


Figure 5: Different dimensions of animal infectious diseases and associated infection biology. High-throughput technologies can be applied to study etiology, genomes, evolution, and outbreak management of infectious diseases as well as host-pathogen interations. This both on the level of the individual sample as on a higher epidemiological scale (Reprinted from Van Borm et al., 2015 [37] with kind permission from Springer Science and Business Media).

3. The viral metagenomic workflow

Viral metagenomics aims to provide the genetic composition of the viral populations present in a sample by directly sequencing genetic material in the sample. However, when sequencing all nucleic acids (through shotgun sequencing approach) present in an un-manipulated biological sample, the amount of viral nucleic acids will be very low or even undetectable. This is due to of the nature of viruses in viral infections. Firstly, the length of viral genomes ranges typically from 5 to 350 kb, which are smaller than genomes of bacteria (typically ranging from 500 kb to 10,000 kb) and eukaryotes (e.g. the human genome is approximately 3,200,000 kb in size). Furthermore, in clinical samples virion concentrations (known as virus titer) are often very low, whereby after nucleic acid isolation (i.e. the process of purification of nucleic acids from a sample), virus genomes are present in extremely low quantities compared to background bacterial and host genomes in clinical samples. Therefore, one will logically sequence much more background nucleic acids. Due to the small genome sizes and low virus titer, the key to a metagenomic based viral discovery workflow is enriching the levels of viral nucleic acids whilst reducing background prokaryotic and eukaryotic nucleic acids [11].

In this section, the frequently used steps to generate viral metagenomes in order to characterize new and known viruses from animal and human samples will be outlined (partially based on reviews: [11, 12]). First, samples are selected and some virus enrichment steps may be performed (section 3.1). Subsequently, the workflow is often divided into an RNA virus and/or DNA virus discovery workflow. After isolation of nucleic acids, an amplification step is often needed to generate sufficient amounts of DNA (section 3.2). Thereafter, DNA is sequenced (section 3.3) and the resulting data are analyzed for similarities with viral sequence information in publicly available databanks (section 3.4). Depending on the purpose of the study, the outcome of a viral metagenomic study may be followed up to exclude possible false positive results or to confirm the result with different diagnostic tests (section 3.5). Besides virus discovery, this viral metagenomic workflow offers the possibility to determine the full genomic information of viruses with little or no available sequence information [13]. The more virus nucleic acids sequenced, the bigger the chance to overlap the full genome. When gaps are still present, gap filling using extensive primer design is needed (primer walking). Therefore, a sensitive workflow is critical for both virus identification and genome characterization.

3.1 Sample preparation

The first and probably most important step is sample selection and collection. Depending on the observed clinical symptoms and recommendations of medical experts (or epidemiologists), samples which are expected to contain the highest virus titer should be collected. Furthermore, rapid collection and proper preservation will reduce bacterial growth. Clinical samples with a high virus titer and low background contamination (bacteria, host cells) are more appropriate for viral metagenomic studies. For these samples, pretreatment steps to enrich for viral nucleic acids are slightly less important compared to clinical samples with a low virus titer or a high amount of contaminating cells. Moreover, samples from which cells can be easily filtered and residual host nucleic acids removed by enzymatic digestion, while viral genetic material remains protected within viral capsids, are very useful for viral metagenomics studies. Examples of such samples are plasma, serum, respiratory secretions, cerebrospinal fluid, urine and feces [11]. Viral metagenomics is much more challenging for samples in which host nucleic acids and viral nucleic acids cannot be easily separated, such as in tissues. If the unknown virus can be grown in a cell culture system (often not the case), the cell supernatant is the preferred sample as it is likely to contain a high virus titer. For clinical biopsy tissue samples, **homogenization** of the sample is the next step in order to disrupt cellular membranes and release all virions. Homogenates are often suspended in buffered solutions in a 10 % or 20 % weight to volume ratio (e.g. [58]). Solid feces samples are also suspended in buffered solutions like phosphate-buffered saline (e.g. [59]).

Regardless of the initial procedure, **low speed centrifugation** of homogenates and liquid samples is a simple and widely used method to separate free virions from larger particles such as cellular debris. Subsequently, **nucleic acids** may be directly **isolated** using either commercial viral nucleic acid extraction kits or chaotropic agents like TRIzol LS reagent (e.g. [60]). However, most random-based viral discovery and genome sequencing studies perform some virion enrichment steps because otherwise little (or even no) viral nucleic acids would be sequenced compared to high amounts of background host and bacterial genomes.

The most frequently used **virion enrichment steps** are:

- Commercially available **sterile filters** used to remove bacteria, eukaryotic cells and other large aggregates from the sample. Filter pore sizes of 0.22 µm and 0.45 µm are frequently used in viral metagenomic experiments [13, 61]. Using 0.22 µm pore size filters has the disadvantage of retaining certain large viruses (Figure 1), whilst 0.45 µm pore size filters may allow flow through of certain bacteria as the size of the largest viruses overlap the size of the smallest bacteria. Either way a compromise should be made. Treatment with antibiotics in order to destroy the cell walls of bacteria before filtration is a possible solution (e.g. [58]).

- When large volume (milliliter [ml] amounts) of starting sample is available, different options exists to **concentrate virions**:
  - **High-speed centrifugation** with an ultracentrifuge has been proven to be a reliable method, especially in combination with a density gradient layered system. The density gradient is often made of cesium chloride or sucrose. After ultracentrifugation, purified virions are harvested at the height of expected density (for further details: [62, 63]). Disadvantages are: (1) the requirement for large starting volumes; (2) virions of certain families are lysed by cesium chloride; (3) the requirement for large amounts of virions due to significant loss of virions after recovery; (4) performing a density gradient layered ultracentrifugation requires some experience. For example, concentration with a cesium chloride gradient is used directly on blood plasma for the purification of human DNA viruses [64]. Stang and colleagues used a 30 % sucrose cushion to purify viral particles from cell culture [65], but high-speed centrifugation may also be used without a density gradient (e.g. [61]). However, ultracentrifuges are not readily available in all laboratories.
  - **Filtration** with filters having pore sizes that retain virions and thereby **concentrate** them. These filter membrane types are expressed in molecular weight limit, meaning the ability to retain molecules above a specified molecular weight (expressed in Dalton). For instance 100 kDa filters have been used to concentrate viruses in feces samples ([66, 67]), 30 kDa filters have been used to study the viral metagenome in mosquitos [68].
  - **Polyethylene glycol (PEG) precipitation** is sometime used as an alternative to ultracentrifugation for concentrating virions (e.g. [69]).
- **Enzymatic removal of non-particle protected nucleic acids.** Deoxyribonuclease (DNase) and ribonuclease (RNase) are enzymes capable of cleaving DNA and RNA respectively. They are frequently used to digest free bacterial and host nucleic acids. The viral nucleic acids are protected from enzymatic activity by viral capsids, and a lipid bilayer in the case of enveloped viruses [67, 70]. The disadvantage of this is that free viral nucleic acids will also be degraded. If the focus is restricted to DNA or RNA

viruses, nucleic acids obtained after nucleic acid isolation can be further digested with the appropriate nuclease (e.g. DNase I treatment on RNA extract [67]).

- A treatment with **chloroform** is sometimes used to disrupt phospholipid bilayer membranes of eukaryotic cells, bacteria and/or mitochondria, and expose their DNA to enzymatic degradation (e.g. [64, 71]). However, this may disrupt also the stability of some lipid enveloped viruses.

Regardless whether virion enrichment steps were performed or not, viral RNA and DNA has to be isolated. From this point, the RNA and DNA virus discovery workflow is often split [13]. In order to remove some extra background RNA after RNA isolation, rRNA (e.g. [48]) or messenger RNA (e.g. [72]) may be depleted by hybridization methods.

3.2 Sequence independent amplification

Virus discovery and whole genome sequencing of viruses is often hindered because of the need for large quantities of genomic material for subsequent cloning and/or sequencing reactions. Especially when pretreatment steps, such as nuclease treatment are performed, nucleic acid quantities will be limited. Therefore, an amplification step is often required to provide sufficient DNA input for downstream sequencing. At the start of this PhD project, few virus discovery and whole genome sequencing studies were available using sequencing without performing an amplification step. Most sequencing workflows at that time required microgram-range (µg) DNA input which limited the possibility of direct sequencing to very specific cases:

- Random-based full viral genome sequencing was applied on well concentrated virus cultured samples. In the pre-NGS era, the whole genome shotgun approach was applied (Figure 3). First the genomic DNA was fragmented into random fragments (by sonication or enzymatic digestion) and subsequently subcloned into plasmid vectors and bacterial cells. This very labor intensive and time-consuming way of working was performed on various large dsDNA genome viruses like poxviruses [73-84] and herpesviruses [85-92]. With the introduction of NGS, direct 454 pyrosequencing using µg's of input DNA allowed characterization of, for instance, Gallid herpesvirus type 2 [93], pseudocowpoxvirus [84], Human herpesvirus 5 [94] and Rodent herpesvirus isolates [95].

- A new polyomavirus was identified in human Merkel cell carcinoma tissue (aggressive human skin cancer) by directly sequencing the cDNA using the 454 technology [72]. Five µg input cDNA was used to prepare the pyrosequencing library.
- The feasibility of detecting arthropod-borne viruses (known as arboviruses) was explored in mosquitoes experimentally infected with dengue virus and pooled with non-infected mosquitoes to simulate samples derived from ongoing arbovirus surveillance programs [48]. Total RNA (ng-range) was purified from mosquito pools and directly incorporated in a library preparation workflow of the GS FLX 454 pyrosequencing platform (using a slightly modified protocol from Simons et al., 2007 [96]).
- The Illumina GA platform allowed identification and characterization of the 2009 pandemic H1N1 influenza A virus from swab specimens [97].
- The Illumina GA platform allowed detection of viral pathogens in nasopharyngeal aspirates from patients with acute lower respiratory tract infections [98]. Approximately 1 µg input DNA was needed to prepare the sequencing library.
- Direct 454 pyrosequencing (cDNA Rapid library preparation protocol), using 200 nanograms (ng) of input RNA, allowed identification of contaminating viruses in vaccine cell substrate cell lines [99].

In contrast, most applications aimed to sequence from clinical samples, required amplification of the target DNA. The most frequently used random whole genome amplification (PCR-based) methods in viral metagenomics are reviewed below. The ideal whole genome amplification method should:

- Introduce as little as possible amplification errors or bias.
- Yield useful amounts of DNA needed for cloning and/or sequencing.
- Be universally applicable to all types of viral genomes (RNA & DNA, single & double stranded, linear & circular, non-segmented and segmented) and samples (blood, tissues, etc).

3.2.1 Phi29 DNA polymerase based random amplification

Method

Multiple displacement amplification (MDA) and multiple-primed rolling circle amplification are two approaches for whole genome amplification (WGA) based on the same principle. These methods use Phi29 DNA polymerase and random hexamer primers (5’-NNNNNN-3’) to amplify ng's of template DNA up to µg's amplified product. After an initial denaturation step to make the template DNA single stranded and binding of the hexamer primers, isothermal genome amplification is performed at a constant temperature of 30°C. Phi29 DNA polymerase originates from the bacteriophage Phi29 which uses this enzyme for the replication of its DNA. Besides (5'→3') DNA synthesis activity, this enzyme possesses several remarkable features which makes it suitable for efficient *in vitro* amplification of DNA [100-102]:

- 3'→5' exonuclease activity: this is an error-correcting process which is also known as proofreading activity. While moving along the template DNA (5'→3'), the polymerase incorporates nucleotides complementary to the template strand. If an incorrect nucleotide is incorporated this nucleotide will be excised by the polymerase and DNA replication can continue. This proofreading activity enhances the Phi29 DNA polymerase fidelity (fewer errors are introduced) compared to *Taq* DNA polymerase which is commonly used in standard PCR reactions.
- Strand displacement activity: When the enzyme encounters a double stranded region while moving along the template DNA, it will displace the complementary strand in the region and continue DNA synthesis.
- Long products: After DNA-primed initiation, DNA is continuously synthesized by the Phi29 polymerase resulting in products with a length higher than 70,000 bp within 20 min.
- Phi29 DNA polymerase is very stable and has a long half-life durability.

The principle for *in vitro* whole genome amplification by MDA is displayed in Figure 6. After a denaturation step, multiple random hexamer primers bind along the template DNA and initiate DNA synthesis with Phi29 DNA polymerase using its strand displacement activity. Other random hexamer primers present in the reaction mixture will bind to the displaced strands and are used as additional initiation points for DNA synthesis. Various branches are obtained and lead to an exponential amplification of the original template. The utilized random hexamers are usually modified (containing thiophosphate linkages) to prevent their degradation by the exonuclease activity of the Phi29 polymerase. The length of the amplification products can be verified by running on an agarose gel. The average product lengths exceeds 10 kb due the high processivity of the polymerase, and constant yields are obtained [103]. The reaction products are subsequently digested with restriction endonucleases to obtain smaller fragments. By eliminating the denaturing step in the MDA reaction, only single stranded DNA (ssDNA) could be selectively amplified [104].

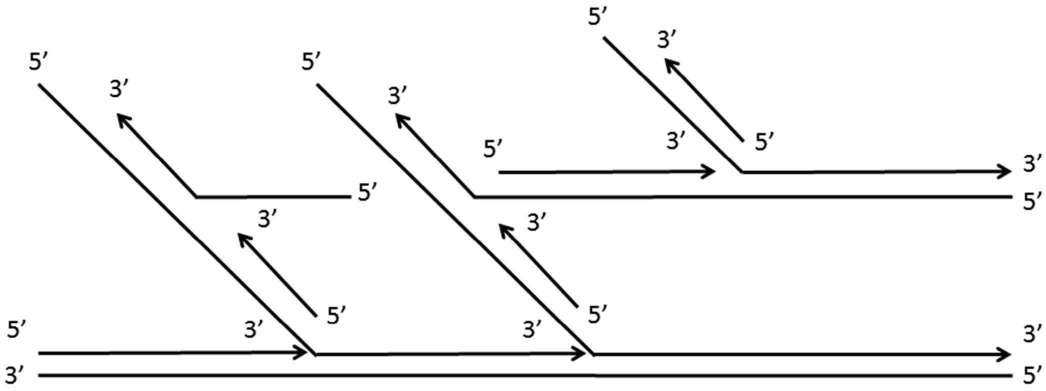


Figure 6: Principle of multiple displacement amplification (MDA) reaction using Phi29 DNA polymerase. At the bottom genomic DNA template is displayed. Arrows represent annealed random primers and following DNA synthesis of complementary strand in the direction of the arrow.

When a circular DNA template is used, the strand displacement activity of Phi29 DNA polymerase causes the formation of large concatemeric DNA molecules (Figure 7). This process is often referred to as multiple-primed rolling-circle amplification (RCA) and results in an exponential amplification of the circular template. The multiple-primed RCA mimics the rolling-circle mechanism that occurs in nature for replication of circular DNA molecules, and amplifies short circular DNA more efficiently when compared to linear DNA. Ten thousand-fold amplification can be obtained in a few hours [100]. The obtained long dsDNA products may subsequently be cut with a restriction enzyme, expected to cut once within the circle sequence, to release linear full genome fragments (Figure 7). As cloned DNA is typically obtained in circular vectors, these can easily be amplified by multiple-primed RCA [105].

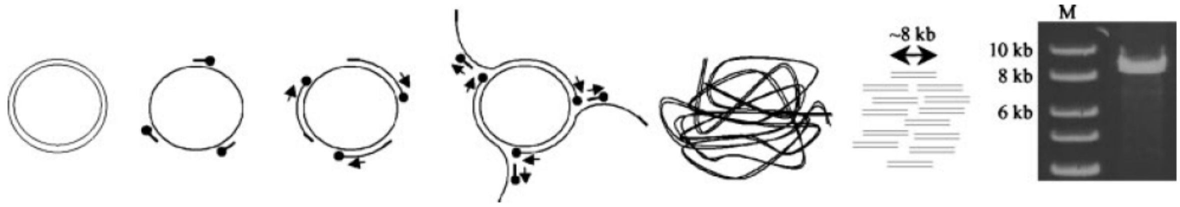


Figure 7: Principle of multiple primed rolling circle amplification (RCA) using Phi29 DNA polymerase (Reprinted from Delwart et al., 2007 [11] with kind permission from John Wiley and Sons).

Various commercial WGA kits are available that use Phi29 DNA polymerase based amplification (Table 1). Kit selection is based on the amount of input genomic DNA, desired yield, degradation state of DNA, source of DNA and whether the WGA process must be automated [101].

Phi29 DNA polymerase cannot be used to amplify RNA directly or to efficiently amplify small cDNA obtained from viral RNA genomes. However, Berthet and colleagues [106] described a Phi29 polymerase based random amplification approach for viral RNA by modifying the MDA based protocol of the QuantiTect Whole Transcriptome kit (Qiagen; Table 1). First, cDNA is made from the RNA by reverse transcription using only random hexamer primers. Then, all cDNAs are ligated together into longer linear chains which serve as the template for an MDA reaction.

Use in viral metagenomics and evaluation

Thurber and colleagues recommended use of the GenomiPhi DNA amplification kit (Sigma-Aldrich, Table 1) to amplify viral DNA in a viral metagenomic sequencing study [62]. Table 2 summarizes the viral metagenomic studies in veterinary and human medicine which made use of Phi29 DNA polymerase based random amplification. Only studies of 2011 and earlier are listed to outline the situation at the start of this PhD project. Although a version of the protocol is described to amplify RNA genomes [106], no RNA virus metagenomic studies were available in the literature. Nevertheless, the amplification method seems well suited for discovery and genome sequencing of circular DNA genomes (almost all studies in Table 2). It has been demonstrated that MDA preferentially amplifies short circular DNA over linear DNA, which results in preferential amplification of ssDNA viruses over dsDNA viruses in the viral metagenome [104, 107]. Moreover, multiple primed RCA is less suitable for amplifying linear genomes and larger circular viral genomes [102]. This method, combined with Sanger sequencing is frequently applied on tissue samples, especially on tumor tissues for the characterization of papillomaviruses. Notwithstanding lack of data demonstrating sensitivity for virus identification, the method seems fit-for-purpose for circular DNA viruses.

Yields by Phi29 DNA polymerase based random amplification methods are very reproducible and are often in µg-range (Table 1). MDA has been associated with statistically significant amplification bias relative to an unamplified control [108] and with production of artifacts such as the formation of chimeric DNA rearrangements in the amplified DNA [109]. Chimeric reads are artificial reads which map ambiguously to different genomic regions and will complicate *de novo* assembly of the genome. Phi29 DNA polymerase based random amplification methods may result in overrepresentation of certain genomic regions and/or a biased representation of the ratio between different species present in a sample [110-112]. This bias seems to be pronounced when low starting DNA amounts are used.

Table 1: Commercial whole genome amplification kits based on sequence-independent amplification (information is based on manufactures' data on their websites and/or on the manual of the kit in question). MDA: multiple displacement amplification, RCA: rolling circle amplification, RT: reverse transcription, rPCR SISPA: random PCR sequence-independent single primer amplification.

Name of kit	Company	Template	Amplification method	Input	Yield	Product size	Time
illustra GenomiPhi V2 DNA amplification kits	GE Health Care Life Sciences	genomic DNA	MDA (at 30°C)	≥10 ng	4-7 µg	>10 kb	1,5-2 h
illustra GenomiPhi HY DNA amplification kits	GE Health Care Life Sciences	genomic DNA	MDA (at 30°C)	≥ 10 ng	40-50 µg	>10 kb	4 h
illustra TempliPhi DNA amplification kits	GE Health Care Life Sciences	< 30 kb circular DNA	RCA (at 30°C)	1 pg – 10 ng	1-1,5 µg	*not described*	4-18 h
illustra TempliPhi large construct kits	GE Health Care Life Sciences	> 30 kb circular DNA	RCA (at 30°C)	1-10 ng	5 µg	*not described*	18 h
REPLI-g Mini kit	Qiagen	genomic DNA	MDA (at 30°C)	> 10 ng	10 µg	>10 kb (range 2-100 kb)	10-16 h
REPLI-g UltraFast Mini kit	Qiagen	genomic DNA	MDA (at 30°C)	> 10 ng	7-10 µg	>10 kb (range 2-100 kb)	1,5 h
REPLI-g Midi kit	Qiagen	genomic DNA	MDA (at 30°C)	> 10 ng	40 µg	>10 kb (range 2-100 kb)	8-16 h
QuantiTect Whole Transcriptome	Qiagen	total RNA	RT followed by ligation and MDA	> 10 ng	up to 40 µg	*not described*	5 -11 h
GenomePlex WGA kits	Sigma-Aldrich	genomic DNA	Chemical fragmentation + rPCR SISPA (14 cycles)	≥10 ng (1 ng for less complex genomes)	>10 µg	0.1-1 kb	< 3 h
Transplex WTA kit & Complete WTA kit	Sigma-Aldrich	total RNA	rPCR SISPA (RT + 17 cycles PCR)	5-300 ng	µg-range	100-1000 bp	< 4 h

Table 2: Review of viral metagenomic studies in veterinary and human medicine which make use of Phi29 DNA polymerase based random amplification. Only studies of 2011 and earlier are listed to outline the situation at the start this PhD project. MDA: multiple displacement amplification, RCA: multiple-primed rolling-circle amplification, rPCR SISPA: random PCR sequence-independent single primer amplification, ss: single stranded, ds: double stranded, PV: papillomavirus, *: the identified virus is a new virus species.

Sample	Reference	Application	Virus (family, genome type)	Amplification	Sequencing
Cell culture supernatant	[113]	Viruses associated with unexplained illness in human patients (as part of an enterovirus surveillance programme)	2BK polyomavirus (*Polyomaviridae*, circular dsDNA); herpes simplex virus (*Herpesviridae*, dsDNA)	MDA	cloning + Sanger
″	[114]	Determination of complete genomic coding sequence of a virus isolated from pigs showing clinical signs of African swine fever	African swine fever virus (*Asfarviridae*, dsDNA)	MDA	454 GS FLX
Lymph nodes	[115]	DNA virus discovery in pigs suffering with postweaning multisystemic wasting syndrome	Porcine boca-like virus* (*Parvoviridae*, ssDNA); Porcine circovirus type 2 (*Circoviridae*, circular ssDNA); Torque Teno virus (*Anelloviridae*, circular ssDNA)	MDA	454 GS FLX
Sputum	[116]	Exploration of DNA viruses in respiratory tract of diseased (cystic fibrosis) and non-diseased humans	Phages, human herpesviruses, human retroviruses and various other DNA viruses	MDA	454 GS FLX
Feces	[117]	Examination of DNA viral flora in feces of monozygotic twins and their mothers	Mainly phages	MDA	454 GS FLX
Blood	[64]	Screening for novel DNA viruses in blood of healthy humans (blood donors)	Human anellovirus* (*Anelloviridae*, circular ssDNA)	MDA + rPCR SISPA	cloning + Sanger
″	[118]	Screening for novel DNA viruses in serum of 8 humans (4 blood donors and 4 patients with AIDS) and 4 pigs	Human and Porcine Torque teno virus (*Anelloviridae*, circular ssDNA)	RCA	cloning + Sanger
″	[119]	Screening for novel DNA viruses in 4 samples of human plasma and 1 sample of saliva from a cat	nine anelloviruses (*Anelloviridae*, circular ssDNA)	RCA + ligation-based SISPA	cloning + Sanger

Table 2 (continued)

Sample	Reference	Application	Virus (family, genome type)	Amplification	Sequencing
Tissue		Identification and genome sequencing of viruses associated with:	Viruses belonging to the *Papillomaviridae* (circular dsDNA):		
″	[120]	• bovine fibropapillomatous wart	Bovine PV type 1*	RCA	cloning + Sanger
″	[121]	• papillomatous skin lesion of a Florida manatee	Trichechus manatus latirostris PV 1*	RCA	cloning + Sanger
″	[122]	• cutaneous papillomatous lesion of a North American porcupine	Erethizon dorsatum PV type 1*	RCA	cloning + Sanger
″	[123]	• malignant lesion of a dog with epidermodysplasia verruciformis	Canine PV 3*	RCA	cloning + Sanger
″	[124, 125]	• condylomatous lesions of bottlenose dolphins	Tursiops truncatus PV type 1, 2, 3*	RCA	cloning + Sanger
″	[126]	• healthy skin of a female goat	Capra hircus PV 1*	RCA	cloning + Sanger
″	[127]	• papillomatous lesions of the bobcat, Florida panther, and Asian lion	Lynx rufus PV 1*, Puma concolor PV 1*, Panthera leo persica PV 1*, and Uncia uncia PV 1*	RCA	cloning + Sanger
″	[128]	• cutaneous papilloma from cattle	Bovine PV type 7	RCA	cloning + Sanger
″	[129]	• papillomatous lesion on the oral mucosa of a polar bear	Ursus maritimus PV type 1*	RCA	cloning + Sanger
			Other (mostly circular) viruses:		
″	[130, 131]	• lesional tissue from bandicoots affected with papillomatosis and carcinomatosis syndrome	Bandicoot papillomatosis carcinomatosis virus type 1, 2* (*Papillomaviridae*-like and *Polyomaviridae*-like, circular dsDNA)	RCA	cloning + Sanger
″	[132]	• fibropapillomatosis in sea turtles	Sea turtle tornovirus 1* (unclassified, circular ssDNA)	MDA + rPCR SISPA	cloning + Sanger
″	[133]	• morality event in lung tissue of captive sea lions	California sea lion anellovirus* (*Anelloviridae*, circular ssDNA)	MDA + rPCR SISPA	cloning + Sanger
″	[134, 135]	• liver and spleen samples of fatally diseased birds	Finch polyomavirus*, Crow polyomavirus*, Canary polyomavirus* (*Polyomaviridae*, circular ssDNA)	RCA	cloning + Sanger

Table 2 (continued)

Sample	Ref.	Application	Virus (family, genome type)	Amplification	Sequencing
Tissue	[136]	• spleen samples of wild starlings found dead during an epidemic outbreak of septicemic salmonellosis	Starling circovirus* (*Circoviridae*, circular ssDNA)	RCA	cloning + Sanger
″	[137]	• mixture of liver and spleen from mute swans found dead	Swan circovirus (*Circoviridae*, circular ssDNA)	RCA	cloning + Sanger
″	[138]	• cardiomyopathy syndrome in heart samples of affected salmon	Piscine reovirus (*Reoviridae*, dsRNA) , Piscine totivirus*(*Totiviridae*, dsRNA)	MDA	454 GS FLX
Arthropod vectors	[68]	DNA viral flora in three mosquito samples	i.a. human papillomavirus-like virus (*Papillomaviridae*, circular dsDNA)	MDA	454 GS20, GS FLX

3.2.2 Sequence-Independent Single Primer Amplification

Commonly used sequence-independent single primer amplification (SISPA) approaches can be categorized as either (1) adapter ligation-based and (2) partially degenerated primer-based SISPA approaches. The latter is better known as “random PCR” (rPCR).

3.2.2.1 Adapter ligation-based SISPA approach

Method

The adapter ligation-mediated SISPA approach was first introduced as a technique to identify viral nucleic acids of unknown sequence present at low concentration [139]. Extracted DNA and RNA are processed separately (Figure 8). RNA is first converted into double stranded cDNA in a random primed reverse transcriptase reaction. The double stranded cDNA could either be digested first by restriction enzymes into smaller fragments [70, 140] or left intact [139, 141]. To detect viral DNA, first a complementary DNA strand is made from the DNA extract as ssDNA could be present. The dsDNA can be either fragmented by restriction enzymes at restriction enzyme recognitions sites (e.g. Csp 6.1 in [70, 140]), or physically at random sites by a shearing instrument (e.g. HydroShear, DIGILAB). The physical shearing approach is known as the Linker Amplified Shotgun Library (LASL) method [142, 143]. Subsequently a short DNA oligonucleotide linker is ligated to both ends (overhangs or blunt) of the dsDNA. A common end sequence of the adapter allows the unknown DNA to be amplified in a subsequent PCR using a single primer where the sequence is complementary to the adapter sequence. After amplification, the random fragments result in a smear (at complex samples) or more discrete bands (at low complexity samples) on an agarose gel [70]. The used adapter may contain restriction endonuclease sites to facilitate subsequent molecular cloning and sequencing.

Different variants of this approach specific for double stranded RNA (dsRNA) viruses exist, whereby oligonucleotides are directly ligated to the dsRNA [144-149]. After the ligation, the dsRNA is denaturated in presence of DMSO (dimethyl sulfoxide; a chemical that assists with strand separation) and converted to cDNA in a reverse transcription reaction. Subsequently, the remaining RNA is removed by hydrolysis with an alkali and full-length cDNA can be amplified (using a single primer), cloned and sequenced.

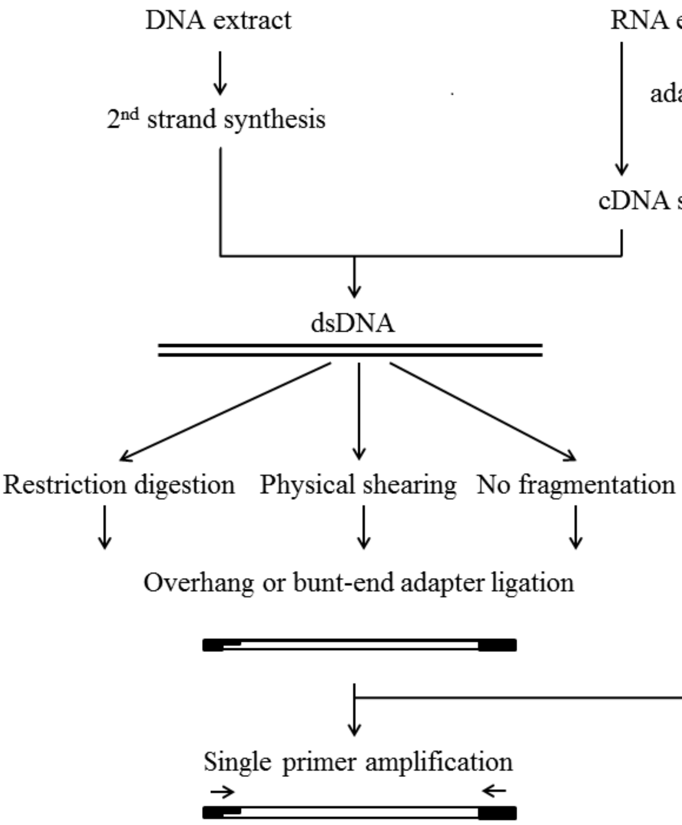


Figure 8: Schematic overview of the different adapter ligation-based SISPA workflows.

Use in viral metagenomics and evaluation

Table 3 reviews the viral metagenomic studies in veterinary and human medicine which made use of adapter ligation-based SISPA (again, limited to studies of 2011 and earlier). The method has been applied to identify and sequence nearly all types of viral genome structures, except for negative sense single stranded RNA (ssRNA) viruses for which there was no published evidence prior to 2011. The dsRNA version of the protocol seems useful for the full genome cloning and sequencing of viruses having a segmented dsRNA genomic organization like viruses of the *Reoviridae* family. Overall, the adapter ligation-based SISPA is not uniform as different versions exist. In addition, the workflow includes laborious steps such as; adapter ligations, restriction enzyme digestions or physical DNA shearing (which requires relatively high initial DNA concentration), use of denaturation enhancing chemicals and so on. It has been demonstrated that the LASL version of the method shows a clear amplification bias towards dsDNA viruses over ssDNA viruses [107]. Notably, most ligation-based SISPA were applied to liquid samples (e.g. cell culture supernatants, blood and feces). A single study documents the use of adapter ligation-based SISPA in combination with NGS performed on a tissue (intestine) sample [47].

Table 3: Review of viral metagenomic studies (2011 and earlier) in veterinary and human medicine which make use of adapter ligation-based SISPA. SISPA: sequence-independent single primer amplification, ss: single stranded, ds: double stranded, +: sense strand, -: antisense strand, dsDNA-RT: replicate through an RNA intermediate, *: the identified virus is a new virus species, °: variant method for dsRNA viruses.

Sample	Reference	Application	Virus (family, genome type)	Amplification	Sequencing
Cell culture supernatant	[150]	Complete sequence determination and genetic analysis of Banna (can cause meningoencephalitis in humans) and Kadipiro viruses	Banna virus, Kadipiro virus (*Reoviridae*, dsRNA)	ligation-based SISPA°	cloning + Sanger
″	[149, 151, 152]	Virus characterization from diseased sheep and cows during an outbreak of Bluetongue disease	Bluetongue virus serotype 6, 8 (*Reoviridae*, dsRNA)	ligation-based SISPA°	cloning + Sanger
″	[153]	Characterization of the virus that causes epizootic hemorrhagic disease in ruminants	Epizootic hemorrhagic disease virus (*Reoviridae*, dsRNA)	ligation-based SISPA°	cloning + Sanger
″	[154]	Characterization of viruses isolated from outbreaks of disease that occurred in horses, donkeys, cattle and sheep in Peru	Peruvian horse sickness virus*, Rioja virus* (*Reoviridae,* dsRNA)	ligation-based SISPA°	cloning + Sanger
″	[155]	Characterization of (zoonotic) tick-borne viruses	Great Island virus, Kemerovo virus, Lipovnik virus and Tribec virus (*Reoviridae,* dsRNA)	ligation-based SISPA°	cloning + Sanger
	[149]	Improving strategy for full genome sequencing viral dsRNA genomes	African horse sickness virus 1, Equine encephalosis virus (*Reoviridae*, dsRNA)	ligation-based SISPA°	454 GS20 & GS FLX
Feces	[156]	Characterization of Norwalk virus, an important cause of acute gastroenteritis in humans	Norwalk virus (*Caliciviridae*, +ssRNA)	ligation-based SISPA	cloning + Sanger
″	[157]	Characterization of cause of non-A and non-B hepatitis epidemic outbreaks in humans	Hepatitis E virus* (*Hepeviridae*, +ssRNA)	ligation-based SISPA	cloning + Sanger
″	[144, 158-161]	Clone and sequence all 11 segments of a Human group C rotavirus from a family outbreak of diarrhea that resulted in the death of an infant	Human group C rotavirus (*Reoviridae,* dsRNA)	ligation-based SISPA°	cloning + Sanger
″	[66, 67, 162]	Examination of viral flora in feces of healthy humans	Various unknown viruses, many phages (*Siphoviridae*, dsDNA), plant RNA viruse …	ligation-based SISPA	cloning + Sanger

Table 3 (continued)

Sample	Ref.	Application	Virus (family, genome type)	Amplification	Sequencing
Feces	[143]	Examination of viral flora in equine feces	Mosty phages, 1 Orthopoxvirus (*Poxviridae*, dsDNA)	ligation-based SISPA	cloning + Sanger
″	[149]	Improving strategy for full genome sequencing viral dsRNA genomes	Human rotavirus G9P (*Reoviridae*, dsRNA)	ligation-based SISPA°	454 GS20
Blood	[70]	Identification of viral contaminants in commercial bovine sera	Bovine parvoviruses 2 and 3* (*Parvoviridae*, ssDNA)	ligation-based SISPA	cloning + Sanger
″	[140]	Screening 25 plasma samples from humans with multiple symptoms of acute viral infection	Hepatitis G virus (*Flaviviridae*, +ssRNA), Hepatitis B virus (*Hepadnaviridae*, dsDNA-RT), Human parvovirus 4* (*Parvoviridae*, ssDNA) and small anellovirus 1, 2* (*Anelloviridae*, circular ssDNA)	ligation-based SISPA	cloning + Sanger
″	[64]	Discovery of novel DNA viruses in blood of healthy humans (blood donors)	Several anellovirus-like sequences (*Anelloviridae*, circular ssDNA)	ligation-based SISPA	cloning + Sanger
″	[163]	Screening of sera for causal virus associated with outbreak of porcine myocarditis (sudden death in piglets; birth of stillborn fetuses)	Bungowannah virus* (*Flaviviridae*, +ssRNA)	ligation-based SISPA	cloning + Sanger
″	[149]	Improving strategy for full genome sequencing viral dsRNA genomes	African horse sickness virus 2 (*Reoviridae*, dsRNA)	ligation-based SISPA°	454 GS FLX
Tissue	[47]	RNA virus community of the gut (intestine tissue) from a turkey presenting with enteric disease	Members of *Reoviridae* (dsRNA), *Picobirnaviridae* (dsRNA) , *Caliciviridae* (+ssRNA), *Picornaviridae* (+ssRNA) and *Astroviridae* (+ssRNA)	ligation-based SISPA	454 GS FLX

3.2.2.2 Random PCR SISPA approach

Method

The rPCR SISPA approach was originally described to amplify cDNA from low amounts of RNA [164, 165]. This approach makes use of a partially degenerated primer consisting of a random sequence at its 3’ end (e.g. hexamer of N’s) and a defined universal sequence tag at its 5’ end (e.g. 20 nucleotides). This partially degenerated primer is actually a mixture of similar sequences. Each N position may be an A, C, G or T nucleotide. Therefore, a partially degenerated primer containing, for instance, 6 N’s exists as a mixture of 4096 (6^4) different primer sequences.

In the first step, the random 3’ part of the primer is annealed to random places of the target RNA or DNA, and a complementary strand is made (Figure 9). When the target is RNA, reverse transcriptase is used for generating first strand cDNA; when the target is DNA, Klenow fragment DNA polymerase is used for generating complementary DNA. In a second cycle (using the Klenow fragment) dsDNA is generated which now contains the universal tag sequence at its two ends. In a subsequent PCR reaction these random dsDNA fragments are amplified using a single primer complementary to the universal 5’ end sequence of the tagged primer. After SISPA amplification, random amplified fragments are usually visualized and size selected on an agarose gel. Amplification of template DNA of high complexity (i.e. variety of DNA sequences are present) and high concentration will lead to a homogeneous size distribution of the SISPA PCR products, which resolve as a smear after size separation by agarose gel electrophoresis. A low concentration and low complexity of template DNA will produce distinct bands after agarose gel electrophoresis [65]. After size selection on agarose gel, SISPA fragments of desired length range are purified, cloned and sequenced.

Different universal 5’-tag sequences have been used in viral metagenomic experiments, as well as different numbers of random N’s at the 3’ end of the tagged primer (Table 4). The universal 5’-tag sequences may contain restriction enzyme sites which could facilitate subsequent molecular cloning and sequencing. Sometimes, in addition to random primers with a random N part, tagged oligos with a poly-T sequence at their 3’ end could be used to help amplify and sequence viruses with poly-A tails (e.g. +ssRNA viruses).

Two types of commercial rPCR SISPA-based kits are available on the market (Table 1). The GenomePlex Whole Genome Amplification kits and the Whole Transcriptome Amplification kits (Sigma-Aldrich) amplify genomic DNA and total RNA respectively, starting from ng-range input to yield µg-range amplified product.

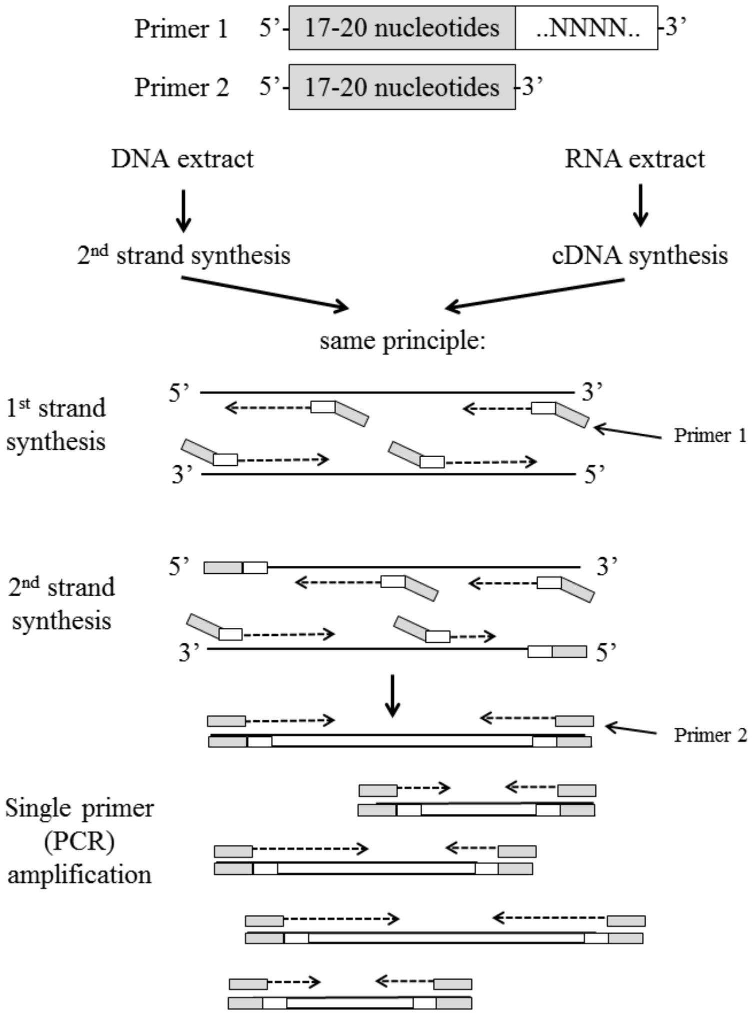


Figure 9: Schematic overview of the random PCR single primer amplification method. First complementary double stranded DNA is made using a partially degenerated primer (“primer 1”) which consists of a random sequence at its 3’ end (e.g. hexamer of N’s) and a defined universal sequence tag at its 5’ end (e.g. 20 nucleotides). In a subsequent PCR reaction these random dsDNA fragments are amplified using a single primer complementary to the universal 5’ end sequence of the tagged primer (“primer 2”).

Use in viral metagenomics and evaluation

An overview of the applications of this method for virus discovery and genome sequencing in veterinary and human medicine is given in Table 5 (studies 2011 and earlier). The method has been widely used both in combination with Sanger and next-generation sequencing, and permits amplification and subsequent detection of any type of viral genome structure. The method is relatively easy to perform, as there is no need for adapter ligation steps and enzymatic digestion. Djikeng and colleagues proved the utility of the method for full genome sequencing of various types and sources of viruses including viral isolates and feces [13]. Limited by using a cloning and Sanger sequencing approach, they observed a need for samples containing a minimum of 10^6/ml particles for generation of (nearly) full viral genomes. Careful examination of the sequence data obtained in rPCR-based studies shows a lack of homogeneous distribution of randomly generated sequence reads over the target genome [13, 48, 57, 59]. In particular, difficulties to obtain viral 3’ and 5’ end sequences have been observed. For +ssRNA viruses which have 3’ end poly-A sequence, tagged poly-T primers have been used to solve this problem [13]. If the virus genome end sequences are known, primers complementary to these sequences may be used to generate full genome sequence (e.g. [166]).

Most of the studies in Table 5 were performed on liquid sample types, although the method is also repeatedly used on tissues. rPCR SISPA is increasingly used in combinations with 454 pyrosequencing, which results in a higher sequencing output and thereby higher virus discovery sensitivity [59]. Nakamura et al. used rPCR SISPA combined with 454 pyrosequencing to sequence directly (without virion enrichment steps) nasopharyngeal aspirates from an influenza virus infection [167]. The majority (>90 %) of the reads were host sequences, which emphasizes the importance of performing virion enrichment step.

Table 4: Overview of the different random PCR SISPA “single primer” sequences used in viral metagenomic studies in veterinary and human medicine (2011 and earlier).

Primer name	Universal 5’ sequence tag	Random 3’ oligo	Publications
universal primer-dN6	GCCGGAGCTCTGCAGAATTC	NNNNNN	[164, 165, 168]
17-mer primer	GTTTCCCAGTAGGTCTC	NNNNNN	[169]
primer D/primer E	GTTTCCCAGTAGGTCTC	NNNNNNNN	[49, 170-176]
pirmer A/primer B	GTTTCCCAGTCACGATC	NNNNNNNNN	[177-182]
pirmer A/primer B	GTTTCCCAGTCACGATA	NNNNNNNNN	[183]
pirmer K-random-s/primer K-s	GACCATCTAGCGACCTCCAC	MNNMNM	[65]
pirmer K-8N/primer-K	GACCATCTAGCGACCTCCAC	NNNNNNNN	[58, 59, 184, 185]
FR26RV-N/FR20RV	GCCGGAGCTCTGCAGATATC	NNNNNN	[13, 55, 60, 61, 166, 186-193]
primer RA01 /primer RA02	"	NNNNNNNNNN	[59, 71, 194]
454-A,B,C,D,E,F,G,H,I,J	different tag / sample (10 in total)	NNNNNNNN	[57, 59, 195-199]
primer 18-Hex/primer 18-Univ	CCATGGATCCACTTCATC	NNNNNN	[200]

Table 5: Review of viral metagenomic studies (2011 and earlier) in veterinary and human medicine which make use of random PCR SISPA. rPCR: random PCR, SISPA: sequence-independent single primer amplification, ss: single stranded, ds: double stranded, +: sense strand, -: antisense strand, ssRNA-RT: replication through a DNA intermediate, *: the identified virus is a new virus species.

Sample	**Reference**	**Application**	**Virus (family, genome type)**	**Amplification**	**Sequencing**
Cell culture supernatant	[177]	Characterization of an unknown virus from a human patient suffering from severe acute respiratory syndrome (SARS)	SARS virus (*Coronaviridae*, +ssRNA)	rPCR SISPA	cloning + Sanger
″	[65]	Identification of unknown virus from mouth-washing material of a patient with chronic fatigue syndrome	Herpes simplex virus type 1 (*Herpesviridae*, dsDNA)	rPCR SISPA	cloning + Sanger
″	[178]	Identification of unknown virus isolated from stool specimens of a patient presenting with gastroenteritis	Human adenovirus-52* (*Adenoviridae*, dsDNA)	rPCR + ligation-based SISPA	cloning + Sanger
″	[179]	Identification of unknown virus isolated from stool specimens of a patient presenting fever of unknown origin	SAF-V* (*Picornaviridae*, +ssRNA)	rPCR + ligation-based SISPA	cloning + Sanger
″	[201]	Testing random based molecular amplification and sequencing for Avian influenza virus surveillance	Avian influenza virus (*Orthomyxoviridae*, -ssRNA)	rPCR SISPA	cloning + Sanger
″	[194]	Examination of a nasal swab from a seal that gave CPE in cell culture	Seal picornavirus 1* (*Picornaviridae*, +ssRNA)	rPCR SISPA	cloning + Sanger
″	[200]	Examination of an viral isolated in a patient who developed a lethal case of pneumonia following a peripheral blood stem cell transplant	Avian paramyxovirus 1 (*Paramyxoviridae*, -ssRNA)	rPCR SISPA	cloning + Sanger
″	[166]	Full genome sequencing 70 Human rhinovirus reference serotypes and 10 field sample isolates	Human rhinovirus (*Picornaviridae*, +ssRNA)	rPCR SISPA	cloning + Sanger
″	[168]	Examination of virus isolated from bat fecal samples	Bat adenovirus strain TJM* (*Adenoviridae*, dsDNA)	rPCR SISPA	cloning + Sanger

Table 5 (continued)

Sample	**Reference**	**Application**	**Virus (family, genome type)**	**Amplification**	**Sequencing**
Cell culture supernatant	[113]	Examination for viruses associated with unexplained illness in human patients (as part of an enterovirus surveillance programme)	Newcastle disease virus (*Paramyxovirinae*, -ssRNA), Saffold viruses (*Picornaviridae*, +ssRNA)	rPCR SISPA	cloning + Sanger
″	[192]	Identification of virus associated with a clinical outbreak of Cardiomyopathy syndrome at Atlantic salmon	Piscine myocarditis virus* (*Totiviridae*, dsRNA)	rPCR SISPA	cloning + Sanger
″	[193]	Investigated the causal agent of a viral infection in 3 hospitalized infants of which one has died	Astrovirus type 4* (*Astroviridae*, +ssRNA)	rPCR SISPA	cloning + Sanger
″	[169, 176]	Examination of an viral isolated from mosquitos	Leanyer virus* (*Bunyaviridae*, -ssRNA), Moussa virus* (*Rhabdoviridae*, -ssRNA)	rPCR SISPA	454 GS FLX
″	[175]	Examination of a virus isolated from an aborted fetus of a sea lion	Steller sea lion reovirus* (*Reoviridae*, dsRNA)	rPCR SISPA	454 GS FLX
″	[202]	Test rPCR cloning technique for detection and genotyping of a PRRSV strain isolated from an aborted porcine fetus	Porcine reproductive and respiratory syndrome virus (*Arteriviridae*, +ssRNA)	rPCR SISPA	cloning + Sanger
Respiratory tract samples	[61, 183, 186]	Virus detection in nasopharyngeal aspirates from patients with symptoms of acute respiratory tract infection	Coronavirus HKU1* (*Coronaviridae*, +ssRNA), Human bocavirus* (*Parvoviridae*, ssDNA), Influenza A (*Orthomyxoviridae*, -ssRNA), Adenovirus (*Adenoviridae,* dsDNA), Respiratory syncytial virus & Metapneumovirus (*Paramyxoviridae*, -ssRNA), TT virus-like (*Anelloviridae*, circular ssDNA), KI polyomavirus* & WU virus* (*Polyomaviridae*, circular dsDNA)	rPCR SISPA	cloning + Sanger
″	[167]	Analyzing nasopharyngeal samples collected during seasonal influenza	Influenza A virus (*Orthomyxoviridae*, -ssRNA), WU polyomavirus (*polyomaviridae*, circular dsDNA)	rPCR SISPA	454 GS FLX
″	[60]	Characterization of a new Avian paramyxovirus from penguin oropharyngeal and cloacal swabs	Avian paramyxovirus serotype 10* (*Paramyxoviridae*, -ssRNA)	rPCR SISPA	cloning + Sanger

Table 5 (continued)

Sample	Reference	Application	Virus (family, genome type)	Amplification	Sequencing
Feces	[203]	To recover additional sequences of a divergent enterovirus identified in nose and throat swab specimens from a patient presenting with influenza-like illness	Enterovirus 109* (*Picornaviridae*, +ssRNA)	rPCR SISPA (modified Solexa protocol)	Illumina GAII
″	[180, 181, 189]	Examination of viral flora in unexplained cases of human diarrhea	i.a. Astrovirus MLB1* (*Astroviridae*, +ssRNA), Human rhinovirus* (*Picornaviridae*, +ssRNA), Picobirnavirus (*Picobirnaviridae*, dsRNA),	rPCR SISPA	cloning + Sanger
″	[71]	Examination of stool samples of children with non-polio acute flaccid paralysis	Human cosavirus A1* (*Picornaviridae*, +ssRNA)	rPCR SISPA	cloning + Sanger
″	[59]	Examination of 35 stool samples of children with non-polio acute flaccid paralysis	i.a. Torque teno virus (*Anelloviridae*, circular ssDNA), Rotavirus (*Reoviridae*, dsRNA), Adenovirus (*Adenoviridae*, dsDNA), Picobirnavirus (*Picobirnaviridae*, dsRNA), Human enterovirus species A-C*, cosavirus, parechovirus, Aichi virus, rhinovirus, and Human cardiovirus (*Picornaviridae*, +ssRNA); nodavirus-like* (*Nodaviridae*, +ssRNA), Dicistrovirus-like* (*Dicistroviridae*, +ssRNA), Circovirus-like* (*Circoviridae*, circular ssDNA), Human bocavirus* (*Parvoviridae*, ssDNA)	rPCR SISPA	cloning + Sanger & 454 GS FLX
″	[167]	Analyzing human fecal samples collected during norovirus outbreaks	Norovirus (*Caliciviridae*, +ssRNA), Human coronavirus HKU1 (*Coronaviridae*, +ssRNA)	rPCR SISPA	cloning + Sanger
″	[184]	Examination for viruses in 3 stool samples of wild-living chimpanzees	Chimpanzee stool-associated circular viruses* (unknown family, circular DNA)	rPCR SISPA	cloning + Sanger
″	[185, 188]	Examination for viruses in porcine stool of healthy piglets (<15 days of age)	Porcine bocaviruses 1*, 2* , V1-H18*, V2-A6* (*Parvoviridae*, ssDNA)	rPCR SISPA	cloning + Sanger

Table 5 (continued)

Sample	**Reference**	**Application**	**Virus (family, genome type)**	**Amplification**	**Sequencing**
Feces	[195]	Explorations of viral flora bat guano	i.a. Bat adeno-associated virus GF-4a* (*Parvoviridae,* ssDNA), cyclovirus GF-4*, circovirus-like virus TM6* (*Circoviridae*, circular ssDNA), kobuvirus TM246k* (*Picornaviridae*, +ssRNA), adenovirus GF-4* (*Adenoviridae*, dsDNA), astrovirus GF-7a* (*Astroviridae*, +ssRNA) and coronavirus TM5* (*Coronaviridae*, +ssRNA)	rPCR SISPA	454 GS FLX
″	[204]	Characterization of virus associated with mink epizootic catarrhal gastroenteritis	Mink Coronavirus WD1127*, WD1133* (*Coronaviridae*, +ssRNA)	rPCR SISPA	cloning + Sanger
″	[197]	Characterization of fecal viral flora of California sea lions	California sea lion: Astroviruses* (*Astroviridae*, +ssRNA), Sapeloviruses* (*Picornaviridae*, +ssRNA), Sapoviruses*, Norovirus* (*Caliciviridae*, +ssRNA), Rotaviruses* (*Reoviridae*, dsRNA); Bocaviruses* (*Parvoviridae*, ssDNA)	rPCR SISPA	454 GS FLX
″	[199]	Characterization of fecal viral flora of 105 wild rodents (mouse, vole, and rat)	i.a. Rodent circo-like viruses* (circular ssDNA), Rodent picobirnaviruses* (*Picobirnaviridae*, dsRNA), Mouse kobuvirus*, Mouse mosavirus*, Rosavirus* (*Picornaviridae*, +ssRNA), Murine astrovirus* (*Astroviridae*, +ssRNA), Adenovirus-associated virus* (*Parvoviridae*, ssDNA), Deer mouse papillomavirus 1* (*Papillomaviridae*, circular dsDNA), Adenovirus* (*Adenoviridae*, dsDNA)	rPCR SISPA	454 GS FLX
″	[196]	Characterization of fecal viral flora of 18 diarrhoeic dogs	Canine kobuvirus* (*Picornaviridae*, +ssRNA), Canine sapovirus* (*Caliciviridae*, +ssRNA), Canine parvoviruses (*Parvoviridae*, ssDNA), Coronaviruses (*Coronaviridae*, +ssRNA) , Rotaviruses (*Reoviridae*, dsRNA)	rPCR SISPA	454 GS FLX

Table 5 (continued)

Sample	**Reference**	**Application**	**Virus (family, genome type)**	**Amplification**	**Sequencing**
Blood	[171]	Full genome sequencing of Ebola virus associated with hemorrhagic fever outbreaks in Uganda (serum sample)	Bundibugyo ebolavirus (*Filoviridae*, -ssRNA)	rPCR SISPA	454 GS FLX
〃	[49]	Virus discovery in serum as part of surveillance study of Nipah virus in bats	GB virus D* (*Flaviviridae*, +ssRNA)	rPCR SISPA	454 GS FLX
〃	[182]	Quality control for detection of adventitious agents in bovine serum	Bovine parvovirus* (*Parvoviridae*, ssDNA)	rPCR SISPA	454
Tissue	[187]	Examination for viruses in intestinal samples of poultry exhibiting characteristic signs of enteric disease	Chicken parvovirus ABU/84*, Turkey parvovirus TuPV/87* (*Parvoviridae*, ssDNA)	rPCR SISPA	cloning + Sanger
〃	[58]	Examination for viruses in brain and muscle of mouse showing pathology after inoculation for investigating samples originating from insect pools, skunk brain, human feces and sewer effluent	Coxsackie A virus*, Simian sapelovirus-49* (*Picornaviridae*, +ssRNA), Eyach virus*, Skunk Orthoreovirus*, California Mosquito Pool Virus* (*Reoviridae*, dsRNA)	rPCR SISPA	cloning + Sanger
〃	[172]	Examination of tissue samples from patients who received organ transplants from a single donor and died of a febrile illness 4 to 6 weeks after transplantation	Dandenong virus* (*Arenaviridae*, -ssRNA)	rPCR SISPA	454 GS FLX
〃	[205]	Characterization of candidate pathogen in a crop biopsy tissue from a live psittacine bird (parrots) suffering from proventricular dilatation disease	Avian bornavirus* (*Bornaviridae*, -ssRNA)	rPCR SISPA (modified Solexa protocol)	Illumina GAII
〃	[173]	Virus discovery in serum and tissue samples of an outbreak of hemorrhagic fever in humans with a high case fatality rate of 80%	Lujo Virus* (*Arenaviridae*, -ssRNA)	rPCR SISPA	454 GS FLX
〃	[190]	Examination of viral flora in fecal, oral, urine, and tissue samples from bats	i.a. Appalachian Ridge Coronaviruses* (*Coronaviridae*, +ssRNA)	rPCR SISPA	454 GS FLX

Table 5 (continued)

Sample	Reference	Application	Virus (family, genome type)	Amplification	Sequencing
Tissue	[174]	Deep sequencing of heart muscle biopsies of farmed salmon with heart and skeletal muscle inflammation disease	Piscine reovirus* (*Reoviridae*, dsRNA)	rPCR SISPA	454 GS FLX
″	[191]	Deep sequencing of brain tissue of mink suffering from shaking mink syndrome	Mink astrovirus* (*Astroviridae*, +ssRNA)	rPCR SISPA	454 GS FLX
″	[198]	Investigate for viruses in spleen and lung samples from a sick wild fox	Gray fox amdovirus* (*Parvoviridae*, ssDNA)	rPCR SISPA	454 GS FLX
Arthropod vectors	[170]	Screening of candidate pathogens for significance of association with honey bee colony collapse disorder	Israeli acute paralysis virus* (*Dicistroviridae*, +ssRNA)	rPCR SISPA	454 GS FLX
Vaccines	[57]	Quality control for detection of adventitious agents in 8 live-attenuated viral vaccines	Besides viral of vaccins: Retroviruses (*Retroviridae*, ssRNA-RT), Porcine circovirus-1 (*Circoviridae*, circular ssDNA)	rPCR SISPA	454 GS FLX

3.2.3 VIDISCA

Method

Virus-Discovery-cDNA-Amplified fragment length polymorphism (VIDISCA) was originally described by van der Hoek and colleagues [206] and is similar to ligation-mediated SISPA (section 3.2.2.1). This process uses two different adapters and primers in the ligation and amplification step respectively. The workflow (Figure 10) is based on the cDNA-AFLP (Amplified fragment length polymorphism) technique [207] and begins with some virion enrichments steps, nucleic acid isolation and dsDNA synthesis using random hexamer primers (2nd strand for viral DNA, cDNA synthesis for viral RNA). Then, two frequently cutting restriction enzymes are used to digest the dsDNA (e.g. *MseI* and *HinPII*) and oligonucleotide adapter sequences are ligated. Subsequently, a PCR reaction is performed using two primers with sequence complementarity to an adapter sequence. Thereafter, a second “selective” PCR amplification is performed using the same primers extended by one extra 3’ random N. The second PCR will simplify the resultant PCR products from a DNA smear to specific bands, when visualized on agarose gel. In 2011, the method was optimized by adjusting the reverse transcription enzymes and including a step to discourage ribosomal RNA amplification [208]. The optimized protocol was called VIDISCA-454, because the amplification products were sequenced by a 454 pyrosequencing platform.

Use in viral metagenomics and evaluation

VIDISCA helped to identify and sequence a new human coronavirus (+ssRNA, *Coronaviridae*) from cell culture showing CPE [206]. The virus was isolated from a nasopharyngeal aspirate specimen of a 7-month-old child suffering from bronchiolitis and conjunctivitis. The authors were also able to amplify hepatitis B (dsDNA-RT, *Hepadnaviridae*) and parvovirus B19 (ssDNA, *Parvoviridae*) from plasma, and HIV-1 (ssRNA-RT, *Retroviridae*) from cell culture. In another study, VIDISCA was used to type uncommon picornaviruses in cell cultures which were previously untypeable using serology [209, 210]. The optimized VIDISCA-454 was successfully used to identify human respiratory syncytial virus (-ssRNA, *Paramyxoviridae*), human coronavirus OC43 (+ssRNA, *Coronaviridae*), influenza virus A & B (-ssRNA, *Orthomyxoviridae*), and adenovirus (dsDNA, *Adenoviridae*) in nasopharyngeal specimens [208]. Two new parvoviruses (ssDNA, *Parvoviridae*) were also identified by VIDISCA-454 from serum and plasma samples of bats [211].

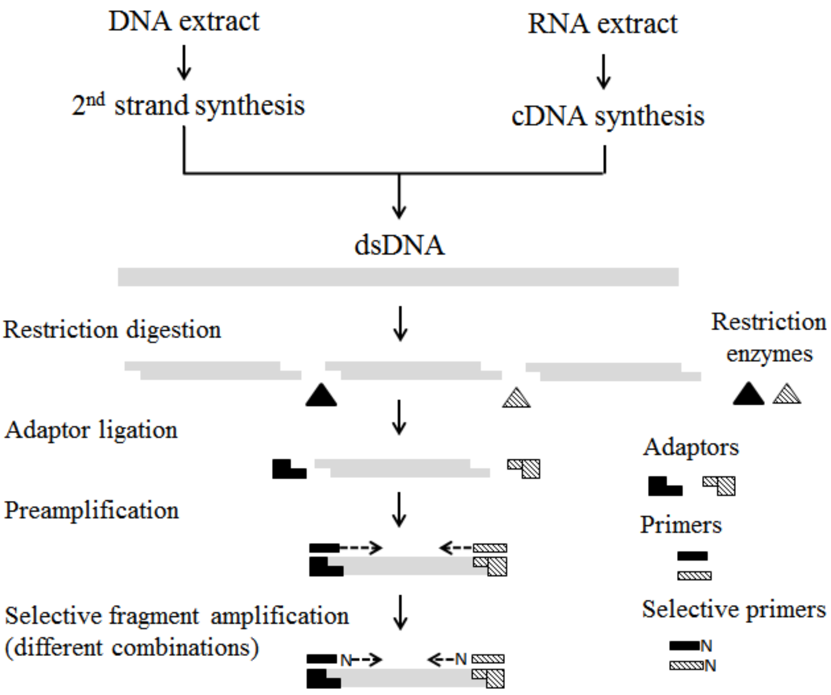


Figure 10: Schematic workflow of the Virus-Discovery-cDNA-Amplified fragment length polymorphism amplification method. dsDNA is fragmented by restriction enzyme digestion using two different enzymes. Subsequently two different adapters are ligated and their sequence is used to selectively amplify the dsDNA.

Theoretically, the method can be applied to all possible viral genome structures. By choosing frequently cutting restriction enzymes, the method could be fine-tuned such that most viruses would be amplified. In practice the method was of limited use in virus discovery and genome sequencing studies prior to 2011. In addition, it is time-consuming and technically demanding (restriction digestion, adapter ligation and two stages of PCR). Whether a virus can be identified depends on the presence of endonuclease restriction site in its genome. Furthermore, the protocol includes two amplification steps which will give a biased representation of the viral content of the sample. Like the other random amplification methods, the approach works best for the identification of viruses from cell cultures or clinical samples containing high viral concentrations [212].

3.3 DNA sequencing

As listed in Tables 2, 3 and 5, Sanger sequencing of amplified DNA has been successfully applied to all kind of sample types and using various pre-amplification methods in viral metagenomic studies. It is not surprising that most of the studies were performed on liquid samples (e.g. cell culture supernatants, blood) containing relatively low amounts of background contaminating nucleic acids. Studies on tissue samples are rare.

NGS is now changing the way we understand viruses. At the start of this PhD project, an increasing number of viral metagenomic studies were integrating high-throughput sequencing, mainly 454 pyrosequencing, in their workflow (Tables 2, 3, 5). The higher associated sequencing output enhanced sensitivity of virus discovery [59]. As such, it is not surprising that recent studies which did not use a sequence-independent amplification step to enrich viral nucleic acids make use of NGS [48, 72, 93, 97, 98]. 454 pyrosequencing has the advantage of producing relatively long sequence reads compared to other NGS platforms; increasing the chance of unique identification of novel viral sequences and facilitating *de novo* assembly (see further in section 3.4). A comparative study of the analytical sensitivity of two technologies, 454 pyrosequencing (GS FLX) and Illumina (GAII), for the detection of viruses in biological samples was done on a set of samples which were artificially spiked with 11 different single or double stranded RNA and DNA viruses [213]. MDA was used to generate appropriate input DNA amounts. The researchers experienced that the higher output of Illumina sequencing is associated with a much greater sensitivity for virus discovery compared to 454 pyrosequencing, and that nearly full-length genomes could be obtained when the virus load was sufficiently high. However, at low viral concentration, the number of reads generated by the Illumina platform (shorter reads) was too small for generating *de novo* contigs of viral sequences [213]. The use of NGS without pre-amplification can be sufficient if the viral nucleic acids are present in sufficient quantity relative to background nucleic acids. However, strategies for viral enrichment and reduction of host background sequences remain essential for detection sensitivity.

3.4 Bioinformatic data analysis

Unlike re-sequencing of viral genomes, the analysis of datasets from metagenomic studies is complicated by the fact that the datasets contain a mixture of different species. Therefore, identification of the sequences is needed. Identification of reads is based on sequence similarity to previously sequenced data and is usually done by the Basic Local Alignment Search Tool, better known as BLAST [214]. Nucleotide sequences may be aligned to (1) nucleotide databases using megablast (search for highly similar sequences; works fast) or blastn (search for more divergent similarities; is slower), (2) protein databases using blastx (very slow), or to (3) translated nucleotide databases using a translated nucleotide query with tblastx (extremely slow). BLAST tools produce alignment scores which indicate similarity of the query sequence to a certain blast hit and an E-value which describes the statistical significance of the blast hit (the closer to zero, the more significant the match is). However, these parameters depend on the length of the used query sequence and the content and size of the used database. There is also a lack of a validated scoring system that permits confident identification of a certain read. Moreover, public databases are biased to sequences of organisms which are more abundant (e.g. Human genomic data). For instance, 30% of all viral nucleotide sequences data in GenBank are HIV sequences (see Figure 11). In addition, existing reference sequence databases are fraught with annotation errors. The criterion for classifying sequences into virus-like sequences is arbitrary. An E-value threshold of <0.001 is often used to classify a viral sequence [59, 66, 67], although others have used a more stringent cutoff (e.g. E-value $<10^{-5}$ [61]; $<10^{-25}$ [13]). BLAST tools can detect known and highly divergent viruses, but it remains extremely challenging to detect the presence of completely new viral sequences. A large fraction of sequences derived from metagenomic studies shows no significant nucleotide and amino acid sequence similarities to any known sequence [11].

As reads are relatively short and the original DNA/RNA fragment is often much larger, the analysis can attempt to re-assemble sequences back into the original overall sequence. Doing this without the use of a reference sequence to map reads is called *de novo* assembly. Assembly software (or assemblers) uses various algorithms to find overlaps between single reads and combine them into contiguous overlapping sets of sequences (contigs). Subsequently, contigs can be identified using BLAST and different contigs may be combined together to obtain, for instance, a full viral genome sequence (known as scaffolding). When BLAST analysis fails to classify contigs, other strategies can be explored. For instance, Dutilh and colleagues used the cross-assembly analysis tool ‘crAss’ for comparing unknown contigs originating from different metagenomics datasets in order to identify co-occurrence of similar contigs [215]. By this means, they identified and sequenced a new bacteriophage crAssphage which has a high prevalence in published human fecal metagenomes [216].

After scaffolding contigs, genomic gaps may still occur and could be completed by direct sequencing of the gap region using e.g. PCR priming sites lying before and after the gap (e.g. using the inverse PCR technique [217]). In viral metagenomics, often the genomic 5’ and 3’ ends are challenging to sequence by the shotgun approach [13]. For RNA genomes, the Rapid Amplification of cDNA Ends (RACE) technique is often used to sequence these ends [176, 218]. In this method, an unknown sequence (5’ or 3’) end sequence is copied by first reverse transcription into cDNA using a known nearby sequence which is more centrally located in the genome. Following reverse transcription, PCR is used to amplify the unknown region from a known starting point.

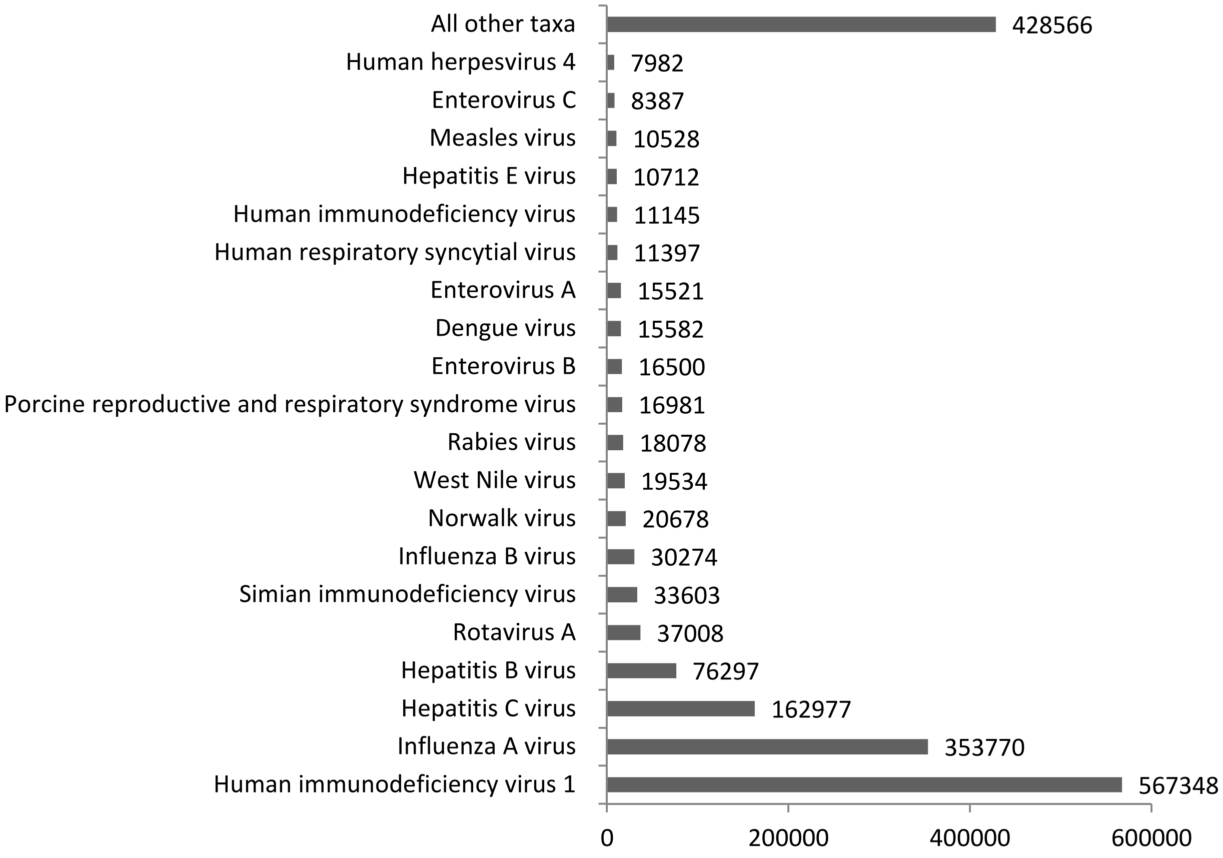


Figure 11: Classification of the 1.8 million viral sequences in NCBI Genbank on 25/12/2014.

The use of high-throughput NGS platforms in viral metagenomic workflows results in new challenges. NGS is associated with a higher error rate compared to Sanger sequencing [25]. Therefore, it is recommended to filter NGS raw sequences for quality. NGS platform specific quality scores typically indicate the probability of base miscalling and can be used to identify and remove low quality reads or parts of reads (called trimming). Moreover, the analysis of NGS data can be computationally challenging and requires minimum level of computational capacity and bioinformatics expertise. No standard data analysis strategy exists. In viral metagenomics, one strategy is to first filter for contaminating background organisms (host, bacterial and other non-viral) by computation subtract them from the dataset. This can be done by aligning all the reads to a reference sequence representing the host species (also known as reference assembly or mapping). The remaining reads are then compared with viral databases using BLAST-like software. To facilitate metagenomic analysis of the data as well as visualization of the results, several programs and platforms have been developed such as; MEGAN [219], PathSeq [220], CAMERA [221, 222], RIEMS [223] and Galaxy [224, 225].

3.5 Follow-up

Depending on the purpose of a study and on the identified viral nucleic acids, the results of a metagenomic study need to be interpreted with care. For instance in the discovery of candidate pathogens, identification of virus-like nucleic acids is only the first step in determining whether or not an association with the observed clinical symptoms is evident. The use of random instead of targeted primers to amplify all of the nucleic acids in samples and the high depth of NGS platforms (easy to generate millions to billions of sequences per run) result in significant potential for laboratory and reagent contamination, in addition to sample carryover between subsequent NGS runs. The availability of sequence information facilitates the development of, for instance, specific PCR-based assays for detection. For example, PCR assays identified contaminating mouse retroviral DNA originating from Qiagen nucleic acid extraction columns [226]. In another study of nasopharyngeal swabs taken from individuals in the 2009 H1N1 pandemic, one sample contained a pair of reads that mapped with 97% nucleotide identity to Ebola virus, but after further investigation, this finding was concluded to be contamination [97]. The assembled data of different disciplines (serology, pathology, epidemiology, metagenomic data, PCR prevalence studies, isolation, characterization, etc.) should be used to identify the most likely candidate etiologic agent and exclude possible contaminations. These strategies were previously applied to conclusively determine that the retrovirus Xenotropic murine leukemia virus-related virus (XMRV) is not associated with chronic fatigue syndrome or prostate cancer in humans and, in fact, originated as a mouse cell line-derived laboratory contaminant [227-233]. The synergistic and parallel use of molecular and classical methods not only results in detection of infectious agents and development of targeted diagnostic tests, but also has the potential to make isolates or strains available shortly after the occurrence of outbreaks. Virus isolation of candidate pathogens is required to assign causality by addressing Koch’s postulates for pathogen-disease association [234]. In addition, the availability of isolates or strains is of special importance to allow the design of effective vaccines or antimicrobial drugs. A good recent example in veterinary is the identification and follow-up of the Schmallenberg virus (see above in section 2.2.3 [39-44]).

4. The challenges

Viral metagenomic workflows attempt to detect and identify infectious agents based on the presence of their nucleic acid molecules in the sample. This application has been successfully used in the characterization of new and unexpected viruses in both veterinary and human health studies. In addition, it obviously has a value in the complete genome characterization, in a random manner, of known and unknown pathogens. However, different challenges remain:

- The application of metagenomics to viruses has not been a straightforward process and no standard workflow exists. Pretreatment steps for separation and concentration of virions in samples are key concerns for enhancing the chance of detecting viral nucleic acids. Strategies often include filtration, centrifugation, and enzymatic digestion of non-viral DNA, but little is known about their true beneficial effect or the bias they may introduce.
- Viral metagenomic investigations typically require amplification of viral nucleic acids in order to generate sufficient input DNA for subsequent DNA sequencing. Various sequence-independent amplification strategies are used to generate appropriate DNA amounts. However, they may introduce amplification bias such as (1) biased relative frequencies of different species in the sample or (2) biased genomic overlap. The origin of amplification bias should be identified and kept to a minimum.
- Although there is a lack of knowledge about the sensitivity (i.e. amount of virus needed for proper identification) of the various workflows used for viral identification, it is assumed that the sensitivity depends on both the characteristics of the clinical sample and the properties of the virus. Sensitivity issues may also limit its applicability for whole genome sequencing. The real challenge is to improve sensitivity of the metagenomic workflow in order to identify viral nucleic acids in samples with a low virus titer and in samples (e.g. tissues) containing a high level of contaminating background nucleic acids. During the course of this thesis, NGS became increasingly available for use in viral metagenomic and genomic studies, providing opportunities for improving sensitivity significantly. NGS can easily generate many million bases of sequence, which entails new challenges regarding data analysis and data storage. Decent computer skills and expertise in bioinformatics are needed, because no standard data analysis workflow exists.

- Sequenced reads are usually compared to known viral nucleotide and protein sequences in databases through BLAST programs. Although viral sequence data is increasingly becoming available in databases, truly novel sequences that do not show any homology to known viruses remain a major challenge. However, obtaining more sequences from the same new virus will enhance the chance of obtaining sequences showing at least a low degree of conservation within a virus family. Sensitivity of the workflow is of crucial importance.

References

1. Morens, D.M., G.K. Folkers, and A.S. Fauci, *The challenge of emerging and re-emerging infectious diseases.* Nature, 2004. **430**(6996): p. 242-9.
2. Wong, S., et al., *Bats as a continuing source of emerging infections in humans.* Rev Med Virol, 2007. **17**(2): p. 67-91.
3. Suttle, C.A., *Viruses in the sea.* Nature, 2005. **437**(7057): p. 356-61.
4. Becker, Y., *PCR: protocols for diagnosis of human and animal virus diseases*. 1995: Springer.
5. Mullis, K., et al., *Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction.* Cold Spring Harb Symp Quant Biol, 1986. **51 Pt 1**: p. 263-73.
6. Woo, P.C., et al., *Then and now: use of 16S rDNA gene sequencing for bacterial identification and discovery of novel bacteria in clinical microbiology laboratories.* Clin Microbiol Infect, 2008. **14**(10): p. 908-34.
7. Murphy, F.A., Gibbs, E.P.J., Horzineck, M.C. and Studdert, M.J., *Veterinary Virology*. Third Edition ed. 1999: Elsevier.
8. Höper, D., B. Hoffmann, and M. Beer, *A comprehensive deep sequencing strategy for full-length genomes of influenza A.* PLoS One, 2011. **6**(4): p. e19075.
9. Lauring, A.S. and R. Andino, *Quasispecies theory and the behavior of RNA viruses.* PLoS Pathog, 2010. **6**(7): p. e1001005.
10. Yoon, S.W., R.J. Webby, and R.G. Webster, *Evolution and ecology of influenza A viruses.* Curr Top Microbiol Immunol, 2014. **385**: p. 359-75.
11. Delwart, E.L., *Viral metagenomics.* Rev Med Virol, 2007. **17**(2): p. 115-31.
12. Blomstrom, A.L., *Viral metagenomics as an emerging and powerful tool in veterinary medicine.* Vet Q, 2011. **31**(3): p. 107-14.
13. Djikeng, A., et al., *Viral genome sequencing by random priming methods.* BMC Genomics, 2008. **9**: p. 5.
14. Sanger, F., S. Nicklen, and A.R. Coulson, *DNA sequencing with chain-terminating inhibitors.* Proc Natl Acad Sci U S A, 1977. **74**(12): p. 5463-7.
15. Smith, L.M., et al., *Fluorescence detection in automated DNA sequence analysis.* Nature, 1986. **321**(6071): p. 674-9.
16. Swerdlow, H. and R. Gesteland, *Capillary gel electrophoresis for rapid, high resolution DNA sequencing.* Nucleic Acids Res, 1990. **18**(6): p. 1415-9.
17. Sears, L.E., et al., *CircumVent thermal cycle sequencing and alternative manual and automated DNA sequencing protocols using the highly thermostable VentR (exo-) DNA polymerase.* BioTechniques, 1992. **13**(4): p. 626-33.
18. Brown, T., *Genomes*. 2nd edition ed. 2002: Oxford: Wiley-Liss.
19. Fiers, W., et al., *A-protein gene of bacteriophage MS2.* Nature, 1975. **256**(5515): p. 273-8.
20. Bartlett, J.M. and D. Stirling, *A short history of the polymerase chain reaction.* Methods Mol Biol, 2003. **226**: p. 3-6.
21. Fleischmann, R.D., et al., *Whole-genome random sequencing and assembly of Haemophilus influenzae Rd.* Science, 1995. **269**(5223): p. 496-512.
22. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
23. Venter, J.C., et al., *The sequence of the human genome.* Science, 2001. **291**(5507): p. 1304-51.
24. Subbiah, M., et al., *Complete sequence of the genome of avian paramyxovirus type 2 (strain Yucaipa) and comparison with other paramyxoviruses.* Virus Res, 2008. **137**(1): p. 40-8.

25. Glenn, T.C., *Field guide to next-generation DNA sequencers.* Mol Ecol Resour, 2011. **11**(5): p. 759-69.
26. Radford, A.D., et al., *Application of next-generation sequencing technologies in virology.* J Gen Virol, 2012. **93**(Pt 9): p. 1853-68.
27. Pareek, C.S., R. Smoczynski, and A. Tretyn, *Sequencing technologies and genome sequencing.* J Appl Genet, 2011. **52**(4): p. 413-35.
28. Margulies, M., et al., *Genome sequencing in microfabricated high-density picolitre reactors.* Nature, 2005. **437**(7057): p. 376-80.
29. Bennett, S., *Solexa Ltd.* Pharmacogenomics, 2004. **5**(4): p. 433-8.
30. Rothberg, J.M., et al., *An integrated semiconductor device enabling non-optical genome sequencing.* Nature, 2011. **475**(7356): p. 348-52.
31. Loman, N.J., et al., *Performance comparison of benchtop high-throughput sequencing platforms.* Nat Biotechnol, 2012. **30**(5): p. 434-9.
32. Schadt, E.E., S. Turner, and A. Kasarskis, *A window into third-generation sequencing.* Hum Mol Genet, 2010. **19**(R2): p. R227-40.
33. Bowers, J., et al., *Virtual terminator nucleotides for next-generation DNA sequencing.* Nat Methods, 2009. **6**(8): p. 593-5.
34. Eid, J., et al., *Real-time DNA sequencing from single polymerase molecules.* Science, 2009. **323**(5910): p. 133-8.
35. Eisenstein, M., *Oxford Nanopore announcement sets sequencing sector abuzz.* Nat Biotechnol, 2012. **30**(4): p. 295-6.
36. Wheeler, D.A., et al., *The complete genome of an individual by massively parallel DNA sequencing.* Nature, 2008. **452**(7189): p. 872-6.
37. Van Borm, S., et al., *Next-generation sequencing in veterinary medicine: how can the massive amount of information arising from high-throughput technologies improve diagnosis, control, and management of infectious diseases?* Methods Mol Biol, 2015. **1247**: p. 415-36.
38. Barzon, L., et al., *Applications of next-generation sequencing technologies to diagnostic virology.* Int J Mol Sci, 2011. **12**(11): p. 7861-84.
39. Hoffmann, B., et al., *Novel orthobunyavirus in cattle, europe, 2011.* Emerg Infect Dis, 2012. **18**(3): p. 469-72.
40. De Regge, N., et al., *Diagnosis of Schmallenberg virus infection in malformed lambs and calves and first indications for virus clearance in the fetus.* Vet Microbiol, 2013. **162**(2-4): p. 595-600.
41. van den Brom, R., et al., *Epizootic of ovine congenital malformations associated with Schmallenberg virus infection.* Tijdschr Diergeneeskd, 2012. **137**(2): p. 106-11.
42. De Regge, N., et al., *Detection of Schmallenberg virus in different Culicoides spp. by real-time RT-PCR.* Transbound Emerg Dis, 2012. **59**(6): p. 471-5.
43. Elbers, A.R., et al., *Schmallenberg virus in Culicoides spp. biting midges, the Netherlands, 2011.* Emerg Infect Dis, 2013. **19**(1): p. 106-9.
44. Wernike, K., et al., *Inactivated Schmallenberg virus prototype vaccines.* Vaccine, 2013. **31**(35): p. 3558-63.
45. *Veterinary Medicines Directorate Grants Provisional Marketing Authorisation to Merck Animal Health for First Vaccine Targeting Schmallenberg Virus.* [cited 2013 21 May]; Available from: http://www.merck-animal-health.com/news/2013-5-21.aspx.
46. *Merial Receives Approval For New Vaccine To Prevent Schmallenberg Disease In Livestock.* [cited 2013 9 August]; Available from: http://www.merial.com/EN/PressRoom/PressRelease/Pages/MerialApprovalSchmallenbergVaccine.aspx.

47. Day, J.M., et al., *Metagenomic analysis of the turkey gut RNA virus community.* Virol J, 2010. **7**: p. 313.
48. Bishop-Lilly, K.A., et al., *Arbovirus detection in insect vectors by rapid, high-throughput pyrosequencing.* PLoS Negl Trop Dis, 2010. **4**(11): p. e878.
49. Epstein, J.H., et al., *Identification of GBV-D, a novel GB-like flavivirus from old world frugivorous bats (Pteropus giganteus) in Bangladesh.* PLoS Pathog, 2010. **6**: p. e1000972.
50. Szpara, M.L., L. Parsons, and L.W. Enquist, *Sequence variability in clinical and laboratory isolates of herpes simplex virus 1 reveals new mutations.* J Virol, 2010. **84**(10): p. 5303-13.
51. Wright, C.F., et al., *Beyond the consensus: dissecting within-host viral population diversity of foot-and-mouth disease virus by using next-generation genome sequencing.* J Virol, 2011. **85**(5): p. 2266-75.
52. Vrancken, B., et al., *Covering all bases in HIV research: unveiling a hidden world of viral evolution.* AIDS Rev, 2010. **12**(2): p. 89-102.
53. Swenson, L.C., M. Daumer, and R. Paredes, *Next-generation sequencing to assess HIV tropism.* Curr Opin HIV AIDS, 2012. **7**(5): p. 478-85.
54. Swenson, L.C., et al., *Deep sequencing to infer HIV-1 co-receptor usage: application to three clinical trials of maraviroc in treatment-experienced patients.* J Infect Dis, 2011. **203**(2): p. 237-45.
55. Ghedin, E., et al., *Deep sequencing reveals mixed infection with 2009 pandemic influenza A (H1N1) virus strains and the emergence of oseltamivir resistance.* J Infect Dis, 2011. **203**(2): p. 168-74.
56. Neverov, A. and K. Chumakov, *Massively parallel sequencing for monitoring genetic consistency and quality control of live viral vaccines.* Proc Natl Acad Sci U S A, 2010. **107**(46): p. 20063-8.
57. Victoria, J.G., et al., *Viral nucleic acids in live-attenuated vaccines: detection of minority variants and an adventitious virus.* J Virol, 2010. **84**(12): p. 6033-40.
58. Victoria, J.G., et al., *Rapid identification of known and new RNA viruses from animal tissues.* PLoS Pathog, 2008. **4**(9): p. e1000163.
59. Victoria, J.G., et al., *Metagenomic analyses of viruses in stool samples from children with acute flaccid paralysis.* J Virol, 2009. **83**(9): p. 4642-51.
60. Miller, P.J., et al., *Evidence for a New Avian Paramyxovirus Serotype-10 Detected in Rockhopper Penguins from the Falkland Islands.* J Virol, 2010. **84**(21): p. 11496-11504.
61. Allander, T., et al., *Cloning of a human parvovirus by molecular screening of respiratory tract samples.* Proc Natl Acad Sci U S A, 2005. **102**(36): p. 12891-6.
62. Thurber, R.V., et al., *Laboratory procedures to generate viral metagenomes.* Nat Protoc, 2009. **4**(4): p. 470-83.
63. Lawrence, J.E., Steward, G.F., *Purification of viruses by centrifugation*, in *MANUAL of AQUATIC VIRAL ECOLOGY*. 2010, American Society of Limnology and Oceanography. p. 166-181.
64. Breitbart, M. and F. Rohwer, *Method for discovering novel DNA viruses in blood using viral particle selection and shotgun sequencing.* Biotechniques, 2005. **39**(5): p. 729-36.
65. Stang, A., et al., *Characterization of virus isolates by particle-associated nucleic acid PCR.* J Clin Microbiol, 2005. **43**(2): p. 716-20.
66. Breitbart, M., et al., *Metagenomic analyses of an uncultured viral community from human feces.* J Bacteriol, 2003. **185**(20): p. 6220-3.

67. Zhang, T., et al., *RNA viral community in human feces: prevalence of plant pathogenic viruses.* PLoS Biol, 2006. **4**(1): p. e3.
68. Ng, T.F., et al., *Broad surveys of DNA viral diversity obtained through viral metagenomics of mosquitoes.* PLoS ONE, 2011. **6**(6): p. e20579.
69. Shah, J.D., et al., *Comparison of tissue sample processing methods for harvesting the viral metagenome and a snapshot of the RNA viral community in a turkey gut.* J Virol Methods, 2014. **209**: p. 15-24.
70. Allander, T., et al., *A virus discovery method incorporating DNase treatment and its application to the identification of two bovine parvovirus species.* Proc Natl Acad Sci U S A, 2001. **98**(20): p. 11609-14.
71. Kapoor, A., et al., *A highly prevalent and genetically diversified Picornaviridae genus in South Asian children.* Proc Natl Acad Sci U S A, 2008. **105**(51): p. 20482-7.
72. Feng, H., et al., *Clonal integration of a polyomavirus in human Merkel cell carcinoma.* Science, 2008. **319**(5866): p. 1096-100.
73. Antoine, G., et al., *The complete genomic sequence of the modified vaccinia Ankara strain: comparison with other orthopoxviruses.* Virology, 1998. **244**(2): p. 365-96.
74. Massung, R.F., et al., *Analysis of the complete genome of smallpox variola major virus strain Bangladesh-1975.* Virology, 1994. **201**(2): p. 215-40.
75. Afonso, C.L., et al., *The genome of fowlpox virus.* J Virol, 2000. **74**(8): p. 3815-31.
76. Afonso, C.L., et al., *The genome of swinepox virus.* J Virol, 2002. **76**(2): p. 783-90.
77. Tulman, E.R., et al., *Genome of lumpy skin disease virus.* J Virol, 2001. **75**(15): p. 7122-30.
78. Tulman, E.R., et al., *The genomes of sheeppox and goatpox viruses.* J Virol, 2002. **76**(12): p. 6054-61.
79. Tulman, E.R., et al., *The genome of canarypox virus.* J Virol, 2004. **78**(1): p. 353-66.
80. Delhon, G., et al., *Genomes of the parapoxviruses ORF virus and bovine papular stomatitis virus.* J Virol, 2004. **78**(1): p. 168-77.
81. Afonso, C.L., et al., *Genome of deerpox virus.* J Virol, 2005. **79**(2): p. 966-77.
82. Tulman, E.R., et al., *Genome of horsepox virus.* J Virol, 2006. **80**(18): p. 9244-58.
83. Afonso, C.L., et al., *Genome of crocodilepox virus.* J Virol, 2006. **80**(10): p. 4978-91.
84. Hautaniemi, M., et al., *The genome of pseudocowpoxvirus: comparison of a reindeer isolate and a reference strain.* J Gen Virol, 2010. **91**(Pt 6): p. 1560-76.
85. Cullinane, A.A., F.J. Rixon, and A.J. Davison, *Characterization of the genome of equine herpesvirus 1 subtype 2.* J Gen Virol, 1988. **69 (Pt 7)**: p. 1575-90.
86. Telford, E.A., et al., *The DNA sequence of equine herpesvirus-1.* Virology, 1992. **189**(1): p. 304-16.
87. Gompels, U.A., et al., *The DNA sequence of human herpesvirus-6: structure, coding content, and genome evolution.* Virology, 1995. **209**(1): p. 29-51.
88. Telford, E.A., et al., *The DNA sequence of equine herpesvirus-4.* J Gen Virol, 1998. **79 (Pt 5)**: p. 1197-203.
89. Delhon, G., et al., *Genome of bovine herpesvirus 5.* J Virol, 2003. **77**(19): p. 10339-47.
90. Li, Y., et al., *Molecular characterization of the genome of duck enteritis virus.* Virology, 2009. **391**(2): p. 151-61.
91. Tulman, E.R., et al., *The genome of a very virulent Marek's disease virus.* J Virol, 2000. **74**(17): p. 7980-8.
92. Afonso, C.L., et al., *The genome of turkey herpesvirus.* J Virol, 2001. **75**(2): p. 971-8.
93. Spatz, S.J. and C.A. Rue, *Sequence determination of a mildly virulent strain (CU-2) of Gallid herpesvirus type 2 using 454 pyrosequencing.* Virus Genes, 2008. **36**(3): p. 479-89.

94. Jung, G.S., et al., *Full genome sequencing and analysis of human cytomegalovirus strain JHC isolated from a Korean patient.* Virus Res, 2011. **156**(1-2): p. 113-20.
95. Loh, J., et al., *Identification and sequencing of a novel rodent gammaherpesvirus that establishes acute and latent infection in laboratory mice.* J Virol, 2011. **85**(6): p. 2642-56.
96. Simons, J.F., Hutchison, S.K., *H5N1: Whole RNA Virus Sequencing Using the Genome Sequencer FLX System.* Biochemica 2007. **4**(Sequencing).
97. Greninger, A.L., et al., *A metagenomic analysis of pandemic influenza A (2009 H1N1) infection in patients from North America.* PLoS ONE, 2010. **5**(10): p. e13381.
98. Yang, J., et al., *Unbiased parallel detection of viral pathogens in clinical samples by use of a metagenomic approach.* J Clin Microbiol, 2011. **49**(10): p. 3463-9.
99. Onions, D., et al., *Ensuring the safety of vaccine cell substrates by massively parallel sequencing of the transcriptome.* Vaccine, 2011. **29**(41): p. 7117-21.
100. Dean, F.B., et al., *Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification.* Genome Res, 2001. **11**(6): p. 1095-9.
101. Silander, K. and J. Saarela, *Whole genome amplification with Phi29 DNA polymerase to enable genetic or genomic analysis of samples of low DNA yield.* Methods Mol Biol, 2008. **439**: p. 1-18.
102. Johne, R., et al., *Rolling-circle amplification of viral DNA genomes using phi29 polymerase.* Trends Microbiol, 2009. **17**(5): p. 205-11.
103. Dean, F.B., et al., *Comprehensive human genome amplification using multiple displacement amplification.* Proc Natl Acad Sci U S A, 2002. **99**(8): p. 5261-6.
104. Kim, K.H., et al., *Amplification of uncultured single-stranded DNA viruses from rice paddy soil.* Appl Environ Microbiol, 2008. **74**(19): p. 5975-85.
105. Nelson, J.R., et al., *TempliPhi, phi29 DNA polymerase based rolling circle amplification of templates for DNA sequencing.* BioTechniques, 2002. **Suppl**: p. 44-7.
106. Berthet, N., et al., *Phi29 polymerase based random amplification of viral RNA as an alternative to random RT-PCR.* BMC Mol Biol, 2008. **9**: p. 77.
107. Kim, K.H. and J.W. Bae, *Amplification methods bias metagenomic libraries of uncultured single-stranded and double-stranded DNA viruses.* Appl Environ Microbiol, 2011. **77**(21): p. 7663-8.
108. Pinard, R., et al., *Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing.* BMC Genomics, 2006. **7**: p. 216.
109. Lasken, R.S. and T.B. Stockwell, *Mechanism of chimera formation during the Multiple Displacement Amplification reaction.* BMC Biotechnol, 2007. **7**: p. 19.
110. Abulencia, C.B., et al., *Environmental whole-genome amplification to access microbial populations in contaminated sediments.* Appl Environ Microbiol, 2006. **72**(5): p. 3291-301.
111. Zhang, K., et al., *Sequencing genomes from single cells by polymerase cloning.* Nat Biotechnol, 2006. **24**(6): p. 680-6.
112. Yilmaz, S., M. Allgaier, and P. Hugenholtz, *Multiple displacement amplification compromises quantitative analysis of metagenomes.* Nat Methods, 2010. **7**(12): p. 943-4.
113. Svraka, S., et al., *Metagenomic sequencing for virus identification in a public-health setting.* J Gen Virol, 2010. **91**(Pt 11): p. 2846-56.
114. Chapman, D.A., et al., *Genomic analysis of highly virulent Georgia 2007/1 isolate of African swine fever virus.* Emerg Infect Dis, 2011. **17**(4): p. 599-605.

115. Blomström, A.L., et al., *Detection of a novel porcine boca-like virus in the background of porcine circovirus type 2 induced postweaning multisystemic wasting syndrome.* Virus Res, 2009. **146**(1-2): p. 125-9.
116. Willner, D., et al., *Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals.* PLoS ONE, 2009. **4**(10): p. e7370.
117. Reyes, A., et al., *Viruses in the faecal microbiota of monozygotic twins and their mothers.* Nature, 2010. **466**(7304): p. 334-8.
118. Niel, C., L. Diniz-Mendes, and S. Devalle, *Rolling-circle amplification of Torque teno virus (TTV) complete genomes from human and swine sera and identification of a novel swine TTV genogroup.* J Gen Virol, 2005. **86**(Pt 5): p. 1343-7.
119. Biagini, P., et al., *Circular genomes related to anelloviruses identified in human and animal samples by using a combined rolling-circle amplification/sequence-independent single primer amplification approach.* J Gen Virol, 2007. **88**(Pt 10): p. 2696-701.
120. Rector, A., R. Tachezy, and M. Van Ranst, *A sequence-independent strategy for detection and cloning of circular DNA virus genomes by using multiply primed rolling-circle amplification.* J Virol, 2004. **78**(10): p. 4993-8.
121. Rector, A., et al., *Characterization of a novel close-to-root papillomavirus from a Florida manatee by using multiply primed rolling-circle amplification: Trichechus manatus latirostris papillomavirus type 1.* J Virol, 2004. **78**(22): p. 12698-702.
122. Rector, A., et al., *Isolation and cloning of a papillomavirus from a North American porcupine by using multiply primed rolling-circle amplification: the Erethizon dorsatum papillomavirus type 1.* Virology, 2005. **331**(2): p. 449-56.
123. Tobler, K., et al., *Detection of the prototype of a potential novel genus in the family Papillomaviridae in association with canine epidermodysplasia verruciformis.* J Gen Virol, 2006. **87**(Pt 12): p. 3551-7.
124. Rehtanz, M., et al., *Isolation and characterization of the first American bottlenose dolphin papillomavirus: Tursiops truncatus papillomavirus type 2.* J Gen Virol, 2006. **87**(Pt 12): p. 3559-65.
125. Rector, A., et al., *Genomic characterization of novel dolphin papillomaviruses provides indications for recombination within the Papillomaviridae.* Virology, 2008. **378**(1): p. 151-61.
126. Van Doorslaer, K., et al., *Genetic characterization of the Capra hircus papillomavirus: a novel close-to-root artiodactyl papillomavirus.* Virus Res, 2006. **118**(1-2): p. 164-9.
127. Rector, A., et al., *Ancient papillomavirus-host co-speciation in Felidae.* Genome Biol, 2007. **8**(4): p. R57.
128. Ogawa, T., et al., *Complete genome and phylogenetic position of bovine papillomavirus type 7.* J Gen Virol, 2007. **88**(Pt 7): p. 1934-8.
129. Stevens, H., et al., *Novel papillomavirus isolated from the oral mucosa of a polar bear does not cluster with other papillomaviruses of carnivores.* Vet Microbiol, 2008. **129**(1-2): p. 108-16.
130. Bennett, M.D., et al., *Genomic characterization of a novel virus found in papillomatous lesions from a southern brown bandicoot (Isoodon obesulus) in Western Australia.* Virology, 2008. **376**(1): p. 173-82.
131. Woolford, L., et al., *A novel virus detected in papillomas and carcinomas of the endangered western barred bandicoot (Perameles bougainville) exhibits genomic features of both the Papillomaviridae and Polyomaviridae.* J Virol, 2007. **81**(24): p. 13280-90.

132. Ng, T.F., et al., *Discovery of a novel single-stranded DNA virus from a sea turtle fibropapilloma by using viral metagenomics.* J Virol, 2009. **83**(6): p. 2500-9.
133. Ng, T.F., et al., *Novel anellovirus discovered from a mortality event of captive California sea lions.* J Gen Virol, 2009. **90**(Pt 5): p. 1256-61.
134. Johne, R., et al., *Characterization of two novel polyomaviruses of birds by using multiply primed rolling-circle amplification of their genomes.* J Virol, 2006. **80**(7): p. 3523-31.
135. Halami, M.Y., et al., *Whole-genome characterization of a novel polyomavirus detected in fatally diseased canary birds.* J Gen Virol, 2010. **91**(Pt 12): p. 3016-22.
136. Johne, R., et al., *Genome of a novel circovirus of starlings, amplified by multiply primed rolling-circle amplification.* J Gen Virol, 2006. **87**(Pt 5): p. 1189-95.
137. Halami, M.Y., et al., *Detection of a novel circovirus in mute swans (Cygnus olor) by using nested broad-spectrum PCR.* Virus Res, 2008. **132**(1-2): p. 208-12.
138. Lovoll, M., et al., *A novel totivirus and piscine reovirus (PRV) in Atlantic salmon (Salmo salar) with cardiomyopathy syndrome (CMS).* Virol J, 2010. **7**: p. 309.
139. Reyes, G.R. and J.P. Kim, *Sequence-independent, single-primer amplification (SISPA) of complex DNA populations.* Mol Cell Probes, 1991. **5**(6): p. 473-81.
140. Jones, M.S., et al., *New DNA viruses identified in patients with acute viral infection syndrome.* J Virol, 2005. **79**(13): p. 8230-6.
141. Culley, A.I., A.S. Lang, and C.A. Suttle, *Metagenomic analysis of coastal RNA virus communities.* Science, 2006. **312**(5781): p. 1795-8.
142. Breitbart, M., et al., *Genomic analysis of uncultured marine viral communities.* Proc Natl Acad Sci U S A, 2002. **99**(22): p. 14250-5.
143. Cann, A.J., S.E. Fandrich, and S. Heaphy, *Analysis of the virus population present in equine faeces indicates the presence of hundreds of uncharacterized virus genomes.* Virus Genes, 2005. **30**(2): p. 151-6.
144. Lambden, P.R., et al., *Cloning of noncultivatable human rotavirus by single primer amplification.* J Virol, 1992. **66**(3): p. 1817-22.
145. Vreede, F.T., et al., *Sequence-independent amplification and cloning of large dsRNA virus genome segments by poly(dA)-oligonucleotide ligation.* J Virol Methods, 1998. **72**(2): p. 243-7.
146. Attoui, H., et al., *Strategies for the sequence determination of viral dsRNA genomes.* J Virol Methods, 2000. **89**(1-2): p. 147-58.
147. Maan, S., et al., *Rapid cDNA synthesis and sequencing techniques for the genetic study of bluetongue and other dsRNA viruses.* J Virol Methods, 2007. **143**(2): p. 132-9.
148. Potgieter, A.C., A.D. Steele, and A.A. van Dijk, *Cloning of complete genome sets of six dsRNA viruses using an improved cloning method for large dsRNA genes.* J Gen Virol, 2002. **83**(Pt 9): p. 2215-23.
149. Potgieter, A.C., et al., *Improved strategies for sequence-independent amplification and sequencing of viral double-stranded RNA genomes.* J Gen Virol, 2009. **90**(Pt 6): p. 1423-32.
150. Attoui, H., et al., *Complete sequence determination and genetic analysis of Banna virus and Kadipiro virus: proposal for assignment to a new genus (Seadornavirus) within the family Reoviridae.* J Gen Virol, 2000. **81**(Pt 6): p. 1507-15.
151. Maan, S., et al., *Sequence analysis of bluetongue virus serotype 8 from the Netherlands 2006 and comparison to other European strains.* Virology, 2008. **377**(2): p. 308-18.

152. Maan, S., et al., *Full genome characterisation of bluetongue virus serotype 6 from the Netherlands 2008 and comparison to other field and vaccine strains.* PLoS ONE, 2010. **5**(4): p. e10323.
153. Anthony, S.J., et al., *Genetic and phylogenetic analysis of the core proteins VP1, VP3, VP4, VP6 and VP7 of epizootic haemorrhagic disease virus (EHDV).* Virus Res, 2009. **145**(2): p. 187-99.
154. Attoui, H., et al., *Peruvian horse sickness virus and Yunnan orbivirus, isolated from vertebrates and mosquitoes in Peru and Australia.* Virology, 2009. **394**(2): p. 298-310.
155. Belhouchet, M., et al., *Complete sequence of Great Island virus and comparison with the T2 and outer-capsid proteins of Kemerovo, Lipovnik and Tribec viruses (genus Orbivirus, family Reoviridae).* J Gen Virol, 2010. **91**(Pt 12): p. 2985-93.
156. Matsui, S.M., et al., *The isolation and characterization of a Norwalk virus-specific cDNA.* J Clin Invest, 1991. **87**(4): p. 1456-61.
157. Reyes, G.R., et al., *Isolation of a cDNA from the virus responsible for enterically transmitted non-A, non-B hepatitis.* Science, 1990. **247**(4948): p. 1335-9.
158. Fielding, P.A., et al., *Molecular characterization of the outer capsid spike protein (VP4) gene from human group C rotavirus.* Virology, 1994. **204**(1): p. 442-6.
159. Deng, Y., et al., *Molecular characterization of the 11th RNA segment from human group C rotavirus.* Virus Genes, 1995. **10**(3): p. 239-43.
160. James, V.L., et al., *Molecular characterization of human group C rotavirus genes 6, 7 and 9.* J Gen Virol, 1999. **80 (Pt 12)**: p. 3181-7.
161. Chen, Z., et al., *Human group C rotavirus: completion of the genome sequence and gene coding assignments of a non-cultivatable rotavirus.* Virus Res, 2002. **83**(1-2): p. 179-87.
162. Breitbart, M., et al., *Viral diversity and dynamics in an infant gut.* Res Microbiol, 2008. **159**(5): p. 367-73.
163. Kirkland, P.D., et al., *Identification of a novel virus in pigs--Bungowannah virus: a possible new species of pestivirus.* Virus Res, 2007. **129**(1): p. 26-34.
164. Froussard, P., *A random-PCR method (rPCR) to construct whole cDNA library from low amounts of RNA.* Nucleic Acids Res, 1992. **20**(11): p. 2900.
165. Froussard, P., *rPCR: a powerful tool for random amplification of whole RNA sequences.* PCR Methods Appl, 1993. **2**(3): p. 185-90.
166. Palmenberg, A.C., et al., *Sequencing and analyses of all known human rhinovirus genomes reveal structure and evolution.* Science, 2009. **324**(5923): p. 55-9.
167. Nakamura, S., et al., *Direct metagenomic detection of viral pathogens in nasal and fecal specimens using an unbiased high-throughput sequencing approach.* PLoS ONE, 2009. **4**(1): p. e4219.
168. Li, Y., et al., *Host range, prevalence, and genetic diversity of adenoviruses in bats.* J Virol, 2010. **84**(8): p. 3889-97.
169. Savji, N., et al., *Genomic and phylogenetic characterization of Leanyer virus, a novel orthobunyavirus isolated in northern Australia.* J Gen Virol, 2011. **92**(Pt 7): p. 1676-87.
170. Cox-Foster, D.L., et al., *A metagenomic survey of microbes in honey bee colony collapse disorder.* Science, 2007. **318**(5848): p. 283-7.
171. Towner, J.S., et al., *Newly discovered ebola virus associated with hemorrhagic fever outbreak in Uganda.* PLoS Pathog, 2008. **4**(11): p. e1000212.
172. Palacios, G., et al., *A new arenavirus in a cluster of fatal transplant-associated diseases.* N Engl J Med, 2008. **358**(10): p. 991-8.

173. Briese, T., et al., *Genetic detection and characterization of Lujo virus, a new hemorrhagic fever-associated arenavirus from southern Africa.* PLoS Pathog, 2009. **5**(5): p. e1000455.
174. Palacios, G., et al., *Heart and skeletal muscle inflammation of farmed salmon is associated with infection with a novel reovirus.* PLoS ONE, 2010. **5**(7): p. e11487.
175. Palacios, G., et al., *Discovery of an orthoreovirus in the aborted fetus of a Steller sea lion (Eumetopias jubatus).* J Gen Virol, 2011. **92**(Pt 11): p. 2558-65.
176. Quan, P.L., et al., *Moussa virus: a new member of the Rhabdoviridae family isolated from Culex decens mosquitoes in Cote d'Ivoire.* Virus Res, 2010. **147**(1): p. 17-24.
177. Wang, D., et al., *Viral discovery and sequence recovery using DNA microarrays.* PLoS Biol, 2003. **1**(2): p. E2.
178. Jones, M.S., 2nd, et al., *New adenovirus species found in a patient presenting with gastroenteritis.* J Virol, 2007. **81**(11): p. 5978-84.
179. Jones, M.S., et al., *Discovery of a novel human picornavirus in a stool sample from a pediatric patient presenting with fever of unknown origin.* J Clin Microbiol, 2007. **45**(7): p. 2144-50.
180. Finkbeiner, S.R., et al., *Metagenomic analysis of human diarrhea: viral detection and discovery.* PLoS Pathog, 2008. **4**(2): p. e1000011.
181. Finkbeiner, S.R., C.D. Kirkwood, and D. Wang, *Complete genome sequence of a highly divergent astrovirus isolated from a child with acute diarrhea.* Virol J, 2008. **5**: p. 117.
182. Onions, D. and J. Kolman, *Massively parallel sequencing, a new method for detecting adventitious agents.* Biologicals, 2010. **38**(3): p. 377-80.
183. Gaynor, A.M., et al., *Identification of a novel polyomavirus from patients with acute respiratory tract infections.* PLoS Pathog, 2007. **3**(5): p. e64.
184. Blinkova, O., et al., *Novel circular DNA viruses in stool samples of wild-living chimpanzees.* J Gen Virol, 2010. **91**(Pt 1): p. 74-86.
185. Shan, T., et al., *Genomic characterization and high prevalence of bocaviruses in swine.* PLoS ONE, 2011. **6**(4): p. e17292.
186. Allander, T., et al., *Identification of a third human polyomavirus.* J Virol, 2007. **81**(8): p. 4130-6.
187. Zsak, L., K.O. Strother, and J. Kisary, *Partial genome sequence analysis of parvoviruses associated with enteric disease in poultry.* Avian Pathol, 2008. **37**(4): p. 435-41.
188. Cheng, W.X., et al., *Identification and nearly full-length genome characterization of novel porcine bocaviruses.* PLoS ONE, 2010. **5**(10): p. e13583.
189. van Leeuwen, M., et al., *Human picobirnaviruses identified by molecular screening of diarrhea samples.* J Clin Microbiol, 2010. **48**(5): p. 1787-94.
190. Donaldson, E.F., et al., *Metagenomic analysis of the viromes of three North American bat species: viral diversity among different bat species that share a common habitat.* J Virol, 2010. **84**(24): p. 13004-18.
191. Blomström, A.L., et al., *Detection of a novel astrovirus in brain tissue of mink suffering from shaking mink syndrome by use of viral metagenomics.* J Clin Microbiol, 2010. **48**(12): p. 4392-6.
192. Haugland, O., et al., *Cardiomyopathy syndrome of atlantic salmon (Salmo salar L.) is caused by a double-stranded RNA virus of the Totiviridae family.* J Virol, 2011. **85**(11): p. 5275-86.
193. Wunderli, W., et al., *Astrovirus infection in hospitalized infants with severe combined immunodeficiency after allogeneic hematopoietic stem cell transplantation.* PLoS ONE, 2011. **6**(11): p. e27483.

194. Kapoor, A., et al., *A highly divergent picornavirus in a marine mammal.* J Virol, 2008. **82**(1): p. 311-20.
195. Li, L., et al., *Bat guano virome: predominance of dietary viruses from insects and plants plus novel mammalian viruses.* J Virol, 2010. **84**(14): p. 6955-65.
196. Li, L., et al., *Viruses in diarrhoeic dogs include novel kobuviruses and sapoviruses.* J Gen Virol, 2011. **92**(Pt 11): p. 2534-41.
197. Li, L., et al., *The fecal viral flora of California sea lions.* J Virol, 2011. **85**(19): p. 9909-17.
198. Li, L., et al., *Novel amdovirus in gray foxes.* Emerg Infect Dis, 2011. **17**(10): p. 1876-8.
199. Phan, T.G., et al., *The fecal viral flora of wild rodents.* PLoS Pathog, 2011. **7**(9): p. e1002218.
200. Goebel, S.J., et al., *Isolation of avian paramyxovirus 1 from a patient with a lethal case of pneumonia.* J Virol, 2007. **81**(22): p. 12709-14.
201. Afonso, C.L., *Sequencing of avian influenza virus genomes following random amplification.* Biotechniques, 2007. **43**(2): p. 188, 190, 192.
202. Van Doorsselaere, J., et al., *Characterization of a circulating PRRSV strain by means of random PCR cloning and full genome sequencing.* Virol J, 2011. **8**: p. 160.
203. Yozwiak, N.L., et al., *Human enterovirus 109: a novel interspecies recombinant enterovirus isolated from a case of acute pediatric respiratory illness in Nicaragua.* J Virol, 2010. **84**(18): p. 9047-58.
204. Vlasova, A.N., et al., *Molecular characterization of a new species in the genus Alphacoronavirus associated with mink epizootic catarrhal gastroenteritis.* J Gen Virol, 2011. **92**(Pt 6): p. 1369-79.
205. Kistler, A.L., et al., *Recovery of divergent avian bornaviruses from cases of proventricular dilatation disease: identification of a candidate etiologic agent.* Virol J, 2008. **5**: p. 88.
206. van der Hoek, L., et al., *Identification of a new human coronavirus.* Nat Med, 2004. **10**(4): p. 368-73.
207. Bachem, C.W., et al., *Visualization of differential gene expression using a novel method of RNA fingerprinting based on AFLP: analysis of gene expression during potato tuber development.* Plant J, 1996. **9**(5): p. 745-53.
208. de Vries, M., et al., *A sensitive assay for virus discovery in respiratory clinical samples.* PLoS ONE, 2011. **6**(1): p. e16118.
209. de Vries, M., et al., *Human parechovirus type 1, 3, 4, 5, and 6 detection in picornavirus cultures.* J Clin Microbiol, 2008. **46**(2): p. 759-62.
210. de Souza Luna, L.K., et al., *Identification of a contemporary human parechovirus type 1 by VIDISCA and characterisation of its full genome.* Virol J, 2008. **5**: p. 26.
211. Canuti, M., et al., *Two novel parvoviruses in frugivorous New and Old World bats.* PLoS ONE, 2011. **6**(12): p. e29140.
212. Tan le, V., et al., *Random PCR and ultracentrifugation increases sensitivity and throughput of VIDISCA for screening of pathogens in clinical specimens.* J Infect Dev Ctries, 2011. **5**(2): p. 142-8.
213. Cheval, J., et al., *Evaluation of high-throughput sequencing for identifying known and unknown viruses in biological samples.* J Clin Microbiol, 2011. **49**(9): p. 3268-75.
214. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.
215. Dutilh, B.E., et al., *Reference-independent comparative metagenomics using cross-assembly: crAss.* Bioinformatics, 2012. **28**(24): p. 3225-31.

216. Dutilh, B.E., et al., *A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes.* Nat Commun, 2014. **5**: p. 4498.
217. Yu, Q., et al., *Rapid acquisition of entire DNA polymerase gene of a novel herpesvirus from green turtle fibropapilloma by a genomic walking technique.* J Virol Methods, 2001. **91**(2): p. 183-95.
218. Huang, J.C. and F. Chen, *Simultaneous amplification of 5' and 3' cDNA ends based on template-switching effect and inverse PCR.* BioTechniques, 2006. **40**(2): p. 187-9.
219. Huson, D.H., et al., *Integrative analysis of environmental sequences using MEGAN4.* Genome Res, 2011. **21**(9): p. 1552-60.
220. Kostic, A.D., et al., *PathSeq: software to identify or discover microbes by deep sequencing of human tissue.* Nat Biotechnol, 2011. **29**(5): p. 393-6.
221. Seshadri, R., et al., *CAMERA: a community resource for metagenomics.* PLoS Biol, 2007. **5**(3): p. e75.
222. Sun, S., et al., *Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource.* Nucleic Acids Res, 2011. **39**(Database issue): p. D546-51.
223. Scheuch, M., D. Hoper, and M. Beer, *RIEMS: a software pipeline for sensitive and comprehensive taxonomic classification of reads from metagenomics datasets.* BMC Bioinformatics, 2015. **16**(1): p. 69.
224. Blankenberg, D., et al., *Galaxy: a web-based genome analysis tool for experimentalists.* Curr Protoc Mol Biol, 2010. **Chapter 19**: p. Unit 19 10 1-21.
225. Goecks, J., A. Nekrutenko, and J. Taylor, *Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences.* Genome Biol, 2010. **11**(8): p. R86.
226. Erlwein, O., et al., *DNA extraction columns contaminated with murine sequences.* PLoS ONE, 2011. **6**(8): p. e23484.
227. Simmons, G., et al., *Failure to confirm XMRV/MLVs in the blood of patients with chronic fatigue syndrome: a multi-laboratory study.* Science, 2011. **334**(6057): p. 814-7.
228. Lee, D., et al., *In-depth investigation of archival and prospectively collected samples reveals no evidence for XMRV infection in prostate cancer.* PLoS ONE, 2012. **7**(9): p. e44954.
229. Knox, K., et al., *No evidence of murine-like gammaretroviruses in CFS patients previously identified as XMRV-infected.* Science, 2011. **333**(6038): p. 94-7.
230. Alter, H.J., et al., *A multicenter blinded analysis indicates no association between chronic fatigue syndrome/myalgic encephalomyelitis and either xenotropic murine leukemia virus-related virus or polytropic murine leukemia virus.* MBio, 2012. **3**(5).
231. Paprotka, T., et al., *Recombinant origin of the retrovirus XMRV.* Science, 2011. **333**(6038): p. 97-101.
232. Oakes, B., et al., *Failure to Detect XMRV-Specific Antibodies in the Plasma of CFS Patients Using Highly Sensitive Chemiluminescence Immunoassays.* Adv Virol, 2011. **2011**: p. 854540.
233. Sato, E., R.A. Furuta, and T. Miyazawa, *An endogenous murine leukemia viral genome contaminant in a commercial RT-PCR kit is amplified using standard primers for XMRV.* Retrovirology, 2010. **7**: p. 110.
234. Fredricks, D.N. and D.A. Relman, *Sequence-based identification of microbial pathogens: a reconsideration of Koch's postulates.* Clin Microbiol Rev, 1996. **9**(1): p. 18-33.

CHAPTER 2

Aims of the thesis

Rapid diagnosis of an infectious disease is essential in order to take appropriate action to control it. During the last decades, rapid and highly sensitive molecular diagnostics (e.g. PCR) became extremely important in disease control. Being highly specific and dependent upon knowledge of the genome sequence of the target pathogen, these tests are vulnerable to emerging variants that have altered sequences. Moreover, sequence information for the design of specific tests is missing for emerging novel pathogens. Therefore, sequence independent methods are increasingly being used for the identification of unknown viruses.

One promising approach is sequencing all of the genetic material within a clinical sample in a random way, and subsequently analyzing the obtained sequencing data for presence of viral sequences. Veterinary diagnostic samples include liquid samples like serum and more complex samples like feces and tissue. Due to the small viral genome sizes and low virus titers, when compared to the enormous background genetic material from host and bacterial cells, enrichment and amplification of viral nucleic acids is of crucial importance. This virus-focused random sequencing workflow is referred to as viral metagenomics, and generally includes the following steps: sample selection, sample preparation, random amplification (if included), sequencing, bioinformatics data analysis and follow up.

Various workflows have been described, using different methodologies to target the viral nucleic acids within a sample as much as possible. They provide an interesting new approach for the identification of unknown pathogens and the characterization of their genomic sequence, complementing classical virology and specific molecular diagnostic tools, both in terms of virus identification & virus genome characterization. However, different challenges remain. This thesis aims to prove the diagnostic potential of viral metagenomics, and to address some of the remaining challenges associated with viral metagenomics.

The main aims of this thesis were:

1. To adopt and fine-tune a universal viral identification method based on random amplification and DNA sequencing, and to evaluate its applicability and value for a diagnostic laboratory. At the start of this PhD project, NGS technologies became more and more accessible. After initial workflow optimizations using cloning and Sanger sequencing, most final studies were carried out using NGS platforms (Chapter 3).
2. To investigate the feasibility and relative value of NGS of random amplified (c)DNA libraries for viral identification in clinical samples, and characterization of high quality complete genome sequences in virus isolates and clinical samples (Chapter 3).

3. Furthermore, to implement appropriate metagenomic bioinformatics analysis tools (Chapter 3, 4).
4. To investigate the nature and origin of amplification bias resulting from the most promising pre-amplification method in viral metagenomic workflows (Chapter 4).
5. To investigate the value of different sample preparation methods and their relative sensitivity for virus discovery (Chapter 4).

Due to the continued improvement of data outputs generated by NGS platforms during the course of the project, a sixth aim was added:

6. To evaluate whether current NGS technologies allow direct characterization and identification of viral genomes in a sample without being preceded by a random amplification step, and compare both approaches in terms of sensitivity and specificity (bias) (Chapter 4).

The general objectives of this study were thus to implement, optimize, and apply novel NGS-based viral metagenomics workflows for the generation, without prior knowledge, of viral sequence information for the identification and molecular characterization of viral pathogens.

CHAPTER 3

Implementation and Case studies

The first goal of this thesis was to adopt one promising and widely applicable viral metagenomics workflow which allows both sensitive virus detection and high quality full genome sequencing. The initial implementation of the workflow was combined with molecular cloning and Sanger sequencing, and is briefly outlined in Chapter 3.1. After successful implementation in the lab, the workflow was combined with NGS (454 pyrosequencing) and applied to ongoing diagnostic and research projects (Chapter 3.2-3.5).

CHAPTER 3.1

Implementation of a viral metagenomic workflow

In this Chapter the selection of a viral metagenomic workflow is discussed and an overview is given of the initial tests of the workflow using different model viruses representing different genomic structures.

An initial important consideration which had to be made was the selection of an appropriate sequence independent amplification method, as various methods are described in the literature. An amplification method is often required to provide sufficient DNA input for the sequencing reaction (e.g. the rapid library preparation workflow of 454 pyrosequencing needs 500 ng of input DNA). The four most frequently used random amplification methods in viral metagenomics are: Phi29 DNA polymerase based random amplification, adapter ligation-mediated SISPA, random PCR SISPA (rPCR; uses a partially degenerated primer) and VIDISCA. The ideal whole genome amplification method should (1) introduce no amplification errors or biases, (2) yield useful amounts of DNA needed for cloning and/or sequencing, and (3) be universal applicable to all types of viral genomes (RNA & DNA, single & double stranded, linear & circular, non-segmented & segmented) and samples (blood, tissues...).

Careful literature review of these four amplification methods was performed in Chapter 1. All methods seem to produce appropriate amounts of DNA needed for the subsequent sequencing. Applications of VIDISCA have been very limited. The method also seems time-consuming and technically demanding. The Phi29 DNA polymerase based random amplification methods are powerful in amplifying genomes of DNA viruses, but are not efficiently applicable to viruses possessing RNA genomes. Furthermore, the method is biased to preferentially amplify circular DNA over linear DNA and ssDNA over dsDNA [1-3]. For the adapter ligation-based SISPA, different versions of the workflow exist, all of which include some laborious steps (adaptor ligations, restriction enzyme digestions, physical DNA shearing, use of denaturation enhancing chemicals, etc.). In addition, only one study prior to 2011 reported using the method on a clinical tissue sample, which could indicate limited sensitivity of the assay. On the other hand, the rPCR SISPA method has been widely used on different sample types. The workflow is relatively easy to perform and permits detection and sequencing of any type of viral genome structure. Based on the above arguments, the rPCR SISPA method was selected as the preferred sequence independent amplification method for the workflow. For the rPCR SISPA primer, the FR26RV-N primer was selected (5’-GCCGGAGCTCTGCAGATATCNNNNNN-3’), as this was the most widely used partially degenerated primer (Chapter 1: Table 4).

Initial adoption of the protocol was based on the workflow in the paper of Djikeng et al. and Allander et al. [4, 5]. As virion enrichment steps, low-speed centrifugation, 0.22 µm filtration, DNase I treatment (overkill DNase amount) and nucleic acid isolation were included.

Subsequently, the workflow was subdivided into an RNA virus discovery workflow (requiring cDNA synthesis) and/or a DNA virus discovery workflow (requiring dsDNA synthesis), and sequence independent amplification was performed by rPCR SISPA. After some initial tests using different SISPA primer concentrations, the method was implemented in the lab in combination with molecular cloning (TOPO TA cloning with pCR2.1 plasmid vector and competent E. coli cells, Invitrogen, Life technologies) and Sanger sequencing. By sequencing a limited amount of bacterial colonies (i.e. their plasmid inserts), correct identification for a range of model viruses with diverse genomic characteristics was possible (Table 1).

Table 1: Summary of used model viruses to test the viral metagenomic workflow.

Virus (family)	**Genome structure (size)**	**Lipid envelope**	**Sample type**	**Virus quantity**	**Identification**
Newcastle disease virus, La Sota strain *(Paramyxoviridae*)	-ssRNA (15 kb)	none	isolate - allantoic fluid	10^8 EID50/ml	14 of 20 colonies
Chicken infectious anemia virus (*Circoviridae*)	circular, ssDNA (2.3 kb)	none	live attenuated vaccine	$\geq 10^6$ TCID50/ml	28 of 50 colonies
Avian influenza virus, H5N1 strain (*Orthomyxoviridae*)	segmented, -ssRNA (14 kb)	yes	isolate - allantoic fluid	$\geq 10^9$ EID50/ml	12 of 16 colonies
Bluetongue virus, BTV-8 strain (*Reoviridae*)	segmented, dsRNA (19 kb)	yes	isolate – cell culture supernatant	$\geq 10^5$ TCID50/50µl	2 of 6 colonies
Goat pox virus, Bangladesh strain (*Poxviridae*)	dsDNA (150 kb)	yes	isolate – cell culture supernatant	unknown	50 of 57 colonies (DNA workflow)
Bovine viral diarrhea virus (Flaviviridae)	+ssRNA (12.5 kb)	yes	contaminant in above Goat pox virus isolate	unknown	2 of 10 colonies (RNA workflow)

For the Goat pox virus isolate, a filter with pore size of 0.45 µm was needed as members of the *Poxviridae* family (size range: 220-450 nm long and 140-260 nm wide) are often too big to pass efficiently through a filter with 0.22 µm pore size (Chapter 1: Figure 1). In addition to the correct identification of the pox virus in the DNA virus identification workflow, a contaminating Bovine viral diarrhea virus (a common cell culture contaminant) could be detected in the RNA virus identification workflow. In addition to virus identification, the percentage of genome coverage for Newcastle disease virus at an infectivity titer of 10^9 EID50/ml was determined by sequencing different numbers of colonies. When sequencing 24, 48, 96, 192 and 288 colonies, 37 %, 60 %, 76 %, 91 % and 97 % of the genome was covered respectively. Remarkably, genome coverage was unevenly spread, with high overlap in some regions and low overlap in other regions. This coverage bias will be further investigated in depth using NGS (Chapter 4.1).

Furthermore, a preliminary evaluation was performed to compare different virus concentration techniques. Newcastle disease virus (an RNA virus belonging to the *paramyxoviridae* family) was used as test virus and diluted from an isolate virus stock (10^{10} EID50/ml) to a concentration of 10^{7} EID50/ml with PBS (phosphate buffered saline) buffer as well as allantoic fluid. Each time 4 ml sample was concentrated using each of the different concentration methods under evaluation to a final volume of 40 µl, and the virus RNA quantity was estimated by specific real-time quantitative RT-PCR assay [6]. An unconcentrated control was each time processed along. The results are displayed in Table 2. The tested PEG precipitation protocol seemed inappropriate to concentrate the tested virus from four ml sample. Ultracentrifugation and ultrafiltration performed similarly for the allontoic fluid samples with a concentration factor of 20× to 25× (calculate using a standard curve). Ultrafiltration with an Amicon Ultra filter seems a good and easy to perform virion concentration method. Moreover, there is no need to have an ultracentrifuge. Ultrafiltration filters exist for several starting volumes, but it is possible use higher sample volumes on the filter as long as the filter is not saturated.

Table 2: Preliminary evaluation of different virion concentration techniques. Starting sample volume was each time four ml PBS or allontoic fluid containing 10^{7} EID50/ml Newcastle disease viruses. Elution volume at RNA extraction was 40 µl (using the QIAamp Viral RNA Mini kit from Qiagen).

Concentration technique	Sample diluent	Ct value concentrated	Ct value control
PEG Virus Precipitation Kit (BioVision) - protocol according to manufacturer's intrustions	PBS	26.47	23.93
	Allontoic fluid	21.1	23.64
Ultracentrifugation - 30,000 RPM for 2h at 4°C	PBS	23.1	23.72
	Allontoic fluid	18.93	23.59
Ultrafiltration with Amicon Ultra-4 centrifugal filter; 50 kDA membrane (Merck Millipore) – 1,500×g for 40min	PBS	17.51	21.66
	Allontoic fluid	17.22	21.75
Ultrafiltration with Amicon Ultra-4 centrifugal filter; 100 kDA membrane (Merck Millipore) – 1,500×g for 40min	PBS	18.41	21.66
	Allontoic fluid	17.49	21.75

In conclusion, a viral genome sequencing method based on random genome amplification and Sanger sequencing was implemented in the lab. Tests with model viruses, representing different genome structures, were successful and demonstrate the potential of the workflow as a tool for virus genome sequencing.

References

1. Johne, R., et al., *Rolling-circle amplification of viral DNA genomes using phi29 polymerase.* Trends Microbiol, 2009. **17**(5): p. 205-11.
2. Kim, K.H., et al., *Amplification of uncultured single-stranded DNA viruses from rice paddy soil.* Appl Environ Microbiol, 2008. **74**(19): p. 5975-85.
3. Kim, K.H. and J.W. Bae, *Amplification methods bias metagenomic libraries of uncultured single-stranded and double-stranded DNA viruses.* Appl Environ Microbiol, 2011. **77**(21): p. 7663-8.
4. Djikeng, A., et al., *Viral genome sequencing by random priming methods.* BMC Genomics, 2008. **9**: p. 5.
5. Allander, T., et al., *Cloning of a human parvovirus by molecular screening of respiratory tract samples.* Proc Natl Acad Sci U S A, 2005. **102**(36): p. 12891-6.
6. Wise, M.G., et al., *Development of a real-time reverse-transcription PCR for detection of Newcastle Disease virus RNA in clinical samples.* J. Clin. Microbiol., 2004. **42**(1): p. 329-338.

CHAPTER 3.2

Identification and complete genome sequencing of paramyxoviruses in mallard ducks (*Anas platyrhynchos*) using random access amplification and next generation sequencing technologies

Toon Rosseel, Bénédicte Lambrecht, Frank Vandenbussche, Thierry van den Berg and Steven Van Borm

Published in Virology Journal (2011) 8: p. 463.

After successful implementation of a viral metagenomic workflow combined with traditional Sanger sequencing, the workflow was combined with NGS (454 pyrosequencing) and the feasibility was tested on different case studies. As a first case study, the workflow was tested on two non-characterized avian paramyxoviruses (-ssRNA genome organization) isolated during a wildlife screening program for avian influenza A and paramyxoviruses. Avian paramyxoviruses (APMV) belong to the *paramyxoviridae* family with a negative-stranded RNA genome. APMV consist of 11 distinct known serotypes at the time of the writing of this thesis (APMV1, APMV2… APMV11). APMV1 includes Newcastle disease virus which causes Newcastle disease, a devastating disease in poultry, and is included in List A of the Office International des Epizooties. Other serotypes appear to be present in natural reservoirs of specific feral avian species, although other host species are usually susceptible. APMV infection may vary in virulence depending on strain, serotype or host species.

The term *sequence independent single primer amplification* or *SISPA* used in this Chapter refers to rPCR SISPA as defined in Chapter 1. The expression *random access amplification* refers to sequence independent amplification as defined in Chapter 1.

Abstract

During a wildlife screening program for avian influenza A viruses (AIV) and avian paramyxoviruses (APMV) in Belgium, we isolated two hemagglutinating agents from pools of cloacal swabs of wild mallards (*Anas platyrhynchos*) caught in a single sampling site at two different times. AIV and APMV1 were excluded using hemagglutination inhibition (HI) testing and specific real-time RT-PCR tests.

To refine the virological identification of APMV2-10 realized by HI subtyping tests and in lack of validated molecular tests for APMV2-10, random access amplification was used in combination with next generation sequencing for the sequence independent identification of the viruses and the determination of their genomes.

Three different APMVs were identified. From one pooled sample, the complete genome sequence (15,054 nucleotides) of an APMV4 was assembled from the random sequences. From the second pooled sample, the nearly complete genome sequence of an APMV6 (genome size of 16,236 nucleotides) was determined, as well as a partial sequence for an APMV4. This APMV4 was closely related but not identical to the APMV4 isolated from the first sample. Although a cross-reactivity with other APMV subtypes did not allow formal identification, the HI subtyping revealed APMV4 and APMV6 in the respective pooled samples but failed to identify the co-infecting APMV4 in the APMV6 infected pool.

These data further contribute to the knowledge about the genetic diversity within the serotypes APMV4 and 6, and confirm the limited sensitivity of the HI subtyping test. Moreover, this study demonstrates the value of a random access nucleic acid amplification method in combination with massive parallel sequencing. Using only a moderate and economical sequencing effort, the characterization and full genome sequencing of APMVs can be obtained, including the identification of viruses in mixed infections.

Keywords: APMV4, APMV6, avian paramyxovirus, mallard, next generation sequencing, random amplification, SISPA

Introduction

A large number of viruses of humans and animals are classified in the family *Paramyxoviridae* [1]. Their single stranded, unsegmented, RNA genomes of negative orientation vary in length from 13-19 kb and contain 6-10 genes encoding up to 12 different proteins [1]. Avian paramyxoviruses (APMV) are frequently isolated from domestic and wild birds throughout the world. Recently they are classified in the genus *Avulavirus* of the subfamily *Paramyxovirinae*, family *Paramyxoviridae* [2]. Ten serological types (APMV1–10) of APMVs are described so far based on hemagglutination inhibition (HI) and neuraminidase inhibition tests [3-5]. APMV1, including Newcastle disease virus (NDV, defined in [6]) is the most characterized among all APMV types because it can cause severe disease outbreaks in poultry. In contrast to the well-studied APMV1 or NDV, very little is known about the biological characteristics, pathogenicity, and diversity (both genetic and antigenic) of other APMV serotypes 2–10. APMV types 2, 3, 6 and 7 have been associated with disease in domestic poultry [7-13]. APMV6 viruses have been associated with mild respiratory disease and decreased egg production in turkeys [14]. APMV3 and APMV5 (Kunitachi virus) caused severe pulmonary disease in wild birds [15, 16]. Other serotypes, including APMV4, 8, 9 and 10 have been isolated from ducks, waterfowls, and other wild birds with no clinical signs of disease [3, 5, 17-20]. APMV4 viruses have been isolated predominantly from feral birds of the order *Anseriformes* [21, 22] and from commercial ducks and geese, presumably as a result of their direct contact with feral waterfowl [21, 23, 24]. Experimental infection of chickens with APMV4 and APMV6 showed mild respiratory pathology, suggestive of possible viral disease in poultry [8, 25].

Molecular characterization through whole genome sequencing of APMV2-10 remains technically challenging because these viruses are poorly represented in public sequence databases, complicating the design of sequencing primers. Recent efforts to sequence whole genomes of representative strains for all serotypes (APMV2 [26], APMV3 [27, 28], APMV4 [29, 30], APMV5 [31], APMV6 [25, 32], APMV7 [33], APMV8 [34], APMV9 [35], APMV10 [5]) have significantly contributed to our understanding of the *Avulavirus* genus genome organisation. However, further studies are needed to explore the diversity within the serotypes.

Random access sequencing using sequence independent single primer amplification was previously described for NDV genome sequencing [36], based on resource demanding sequencing of high number of cloned random amplicons to achieve completion of a genome. This protocol contains efficient steps to enrich viral nucleic acids and deplete contaminating and host sequences, including size selective filtration and extensive nuclease treatments [36, 37]. It was also used for the molecular identification of an APMV in penguins [5] where existing protocols did not allow a starting point for primer walking. This resulted in the identification of a new serotype, APMV10.

Massive parallel sequencing technologies were developed to accommodate the need of higher sequencing capacity and lower costs per nucleotide for large genome sequencing projects [38]. One main advantage of these second generation sequencing technologies is the possibility to sequence DNA samples without any prior knowledge of the sequence, which is required for priming [38].

During a wildlife screening program for avian influenza A viruses (AIV) and APMVs, we isolated two hemagglutinating agents from two pools consisting of each four cloacal swabs of wild mallards. The birds were caught in a same location at two different times. AIV and APMV1 were excluded using HI testing and specific real-time RT-PCR tests. To refine HI test based identification of these viruses and in lack of validated molecular tests for APMV2-10, this study applied random access amplification in combination with next generation sequencing for the sequence independent identification of the viruses and the determination of their complete genome sequences.

Materials and Methods

Viruses

Two non-characterized APMVs (mallard/Belgium/12245/07 and mallard/Belgium/15129/07) were isolated from two pools consisting of each four cloacal swabs from healthy wild mallard ducks according to standard diagnostic procedures (OIE, diagnostic manual 2005/94/CE). The wild birds were caught in a funnel trap located along a pond at 20 km SE of Brussels in Belgium. The trap was visited every two to three days during the entire survey period. All new birds were ringed, weighted, the wings measured, and a cloacal swab was collected. A maximum of four cloacal swabs from the same bird species, sex and sampling time were pooled for laboratory analysis.

HI-tests

Briefly, the hemagglutination (HA) titer of the different viruses was standardized to a concentration of four units of HA activity/25 µl to perform the test (methodology according to Council Directive 92/66/EC (1992)). All HI tests referenced in this study were conducted with the AIV and APMV1-9 reference sera provided by the European reference laboratory VLA (Weybridge, U.K.). The titer of a serum is defined by the last dilution giving a complete inhibition of HA. A titer below 16 is considered as negative and a titer above or equal to 16 is considered as positive. Absence of APMV1 was confirmed using specific real-time RT-PCR assays (data not shown).

Random access to viral nucleic acids using DNAse I SISPA

Virus particles from samples mallard/Belgium/12245/07 and mallard/Belgium/15129/07 were purified starting from one ml of allantoic fluid. This was first centrifuged at 3,200 × g for 15 minutes at four °C to remove cell debris. The supernatants were then filtered at 3,000 × g for eight minutes or longer at four °C (200 µl/filter) using 0.22 µm filters (Ultrafree-MC GV sterile, Millipore) to remove remaining cell fragments and bacteria. The resulting eluates were subsequently subjected to nuclease treatment with 100 U of DNase I (New England Biolabs) at 37 °C for one hour to remove all nucleic acids that are not protected within virions. The resulting virion-enriched samples were used for viral RNA extraction using the QIAamp Viral RNA Mini Kit (Qiagen) according to the manufacturer’s instructions.

Sequence independent single primer amplification (SISPA) was performed essentially as previously described [36] with some modifications. Briefly, the extracted RNA was converted into single-stranded cDNA using the Transcriptor First Strand cDNA Synthesis Kit (Roche) and one µM (final concentration) random primer FR26RV-N (5’-GCC GGA GCT CTG CAG ATA TCN NNN NN-3’, [36, 37]). Ten µl extracted RNA was denatured at 95 °C for five minutes in the presence of primer FR26RV-N, immediately followed by cooling on ice. The remaining reagents were added. The 20 µl reaction mix contained 1× Transcriptor Reverse Transcriptase Reaction Buffer, dNTP mix (1mM final concentration each), 20 U Protector RNase Inhibitor, ten units Transcriptor Reverse Transcriptase and one µl PCR-grade H_2O. The reaction was incubated at 25 °C for ten minutes followed by 50 °C for 60 minutes. After a reverse transcriptase inactivation step at 85 °C for five minutes and chilling on ice, 2.5 U of 3’-5’ exo^- Klenow Fragment of DNA polymerase (New England Biolabs) were added for second strand synthesis using random primer FR26RV-N for one hour at 37 °C. An enzyme inactivation step was performed at 75°C for ten minutes.

Five microliters of the reaction mix was used as template for a subsequent PCR amplification. The 50µl reaction mix consisted of 1× AmpliTaq Gold® 360 DNA buffer, 2.5 mM MgCl2, dNTP mix (0.2 mM final concentration each), 2.5 U AmpliTaq Gold® 360 DNA polymerase (Applied Biosystems), 32.7 µl RNase free water and 1.6 µM FR20RV primer (5'-GCC GGA GCT CTG CAG ATA TC-3', [36, 37]). This PCR primer is complementary to the amplification tag of FR26RV. The reaction was incubated at 95 °C for ten minutes, 40 cycles at 95 °C for one minute, 48 °C for one minute and 72 °C for two minutes followed by a final elongation for seven minutes at 72 °C.

The random amplified DNA fragments were visualised on a one % agarose gel. Fragments of 400-1,000 base pairs (bp) were excised and purified from the gel with the High Pure PCR Product Purification Kit (Roche). The purified PCR fragments were quantified by spectrophotometry (Nanodrop-1000).

Sequencing

Five micrograms of size selected (400-1,000 bp) purified random amplified DNA was sequenced on a GS FLX (Roche, Mannheim, Germany) by the Genomics Core of the University Hospital, University of Leuven, Belgium. They used multiplex identifier (MID) identification during library preparation (standard Roche MID tag sequences) and GS FLX Titanium series reagents (Roche, Mannheim, Germany) according to their standard procedures, aiming for 5,000-10,000 reads per library. Briefly, adaptors including standard MID tag sequences (for our samples RL3 and RL10) were ligated to the size selected double stranded DNA library (Rapid Library Preparation Method Manual, GS FLX Titanium Series reagents, Roche, Mannheim, Germany), followed by single stranded DNA library isolation and library quality assessment and quantitation. The resulting libraries were then pooled with other MID identified libraries and emulsion PCR clonal amplification was performed as described by the provider. The amplified libraries were then loaded on a Pico Titer Plate for sequencing by the Genome Sequencer FLX. Data were provided to the authors by secured ftp-server.

DATA Analysis

The obtained raw sequence data were assembled using SeqMan NGen® version 3.0 (DNASTAR, Madison, WI, USA). The reads were trimmed to remove primer sequences as well as low quality ends. Standard assembling and filtering parameters were used. First we performed a de novo assembly and entered the resulting contigs (i.e. sets of overlapping sequence reads) into a Blastn similarity search against public sequence databases (http://blast.ncbi.nlm.nih.gov/Blast.cgi; [39]) for identification. When we identified a certain APMV serotype, we used the blast hit with the highest identity of the biggest APMV contig as reference genome for a subsequent reference assembly with the same raw data set. The resulting reference assembly was used to obtain a complete genome consensus sequence. The sequence reads contributing to the consensus were also checked for variability. When at a certain position along the consensus two different nucleotides were present, the variability was indicated as an ambiguous nucleotide when the minor nucleotide exceeded the threshold of one third of the reads.

Analysis of the virus specificity of the protocol

Sequences failing to align with the used reference genome were subjected to a metagenomics assembly in SeqMan NGen. The obtained contigs containing more than two sequence reads were identified with megablast (http://blast.ncbi.nlm.nih.gov/Blast.cgi). Sequences were classified as previously described [36]. Briefly, viral blast results were considered reliable if the best hit had an E-value less than 10^{-25}. Non-viral sequences were identified as *Gallus gallus* (chicken embryos were used for virus isolation), other birds, bacteria, … if their best hit was below an E-value of 10^{-10}. If no blast results were found or the E-value was below the 10^{-10} cut off value, the sequences were not given a specific designation.

Phylogenetic analysis

Consensus sequences were edited, aligned and translated, and sequence identities were calculated using Bioedit v7.0.5.3 (http://www.mbio.ncsu.edu/bioedit/bioedit.html) [40]. Nucleotide (nt) sequence identities with selected complete genome sequences were determined (GenBank accession codes: EU877976, FJ177514, EF569970, NC003043, GQ406232, EU622637). Amino acid (aa) alignments (ClustalW algorithm) using all available complete coding sequences for the F and HN genes of APMV4 (GenBank accession codes FJ177514, EU877976) and APMV6 (GenBank accession codes EF569970, NC003043, EU622637, GQ406234, GQ406233) and selected sequences representative of other APMV serotypes (GenBank accession codes AF077761, DQ097393, HQ896024, HQ896023, HM159993-HM159995, AY129676, EU338414, EU782025, EU403085, GU206351, FJ231524, FJ619036, FJ215864, GU068584-GU068587, EU910942, HM147142) were used for phylogenetic analysis. Mega v5.01 [41] was used to construct phylogenetic trees by bootstrap analysis (1,000 replicates) using the neighbour-joining of the Poisson-corrected values for aa differences. All positions containing alignment gaps and missing data were eliminated only in pairwise sequence comparisons, allowing the inclusion of the incomplete aa F and HN sequences of APMV4/mallard/Belgium/12245/07 (APMV4-BE12245) in the analysis.

Results

Identification and genome sequence of avian paramyxoviruses

Two pooled samples, consisting of each four swab samples from wild mallards, were positive for hemagglutinating agents without inducing mortality of embryonated chicken eggs. AIV and APMV1 could be excluded using specific real-time RT-PCR tests (data not shown) and HI tests using reference sera for AIV and APMV1. The HI assays with reference sera specific for APMV2-9 identified sample mallard/Belgium/15129/07 as APMV4 positive and sample mallard/Belgium/12245/07 as APMV6 positive. A cross reactivity with the APMV2 reference serum P/Robin/Hiddensee/57 was observed for both samples, but not with another APMV2 reference serum P/chicken/Yucaipa/Cal/56. The HI titers for the APMV3 and APMV7 reference sera showed for sample mallard/Belgium/15129/07 the borderline value of 16, still we considered this as nonspecific reactivity (Table 1).

Table 1: Hemagglutination inhibition tests (HI-titers) of pooled samples mallard/Belgium/15129/07 (07/15129) and mallard/Belgium/12245/07 (07/12245).

Antisera (used inactivated virus strain)	**07/15129**	**07/12245**
Polyclonal serum APMV1	<4	<4
Monoclonal serum 12B7 anti APMV1		<4
Monoclonal serum 8C11 + 4D6 anti APMV1		<4
Polyclonal serum APMV2 (P/robin/Hiddensee/15/75)	32	64
Polyclonal serum APMV2 (P/chicken/Yucaipa/Cal/56)	<4	8
Polyclonal serum APMV3 (P/tk/1087/82)	16	4
Polyclonal serum APMV4 (P/duck/Hong Kong/D3/75)	128	<4
Polyclonal serum APMV6 (P/duck/Hong Kong/199/77)	<4	64
Polyclonal serum APMV7 (P/dove/TN/4/75)	16	8
Polyclonal serum APMV8 (P/goose/Del/1053/76)	<4	4
Polyclonal serum APMV9 (P/duck/NY/22/78)	<4	<4
28 influenza A reference sera tested *	<4	<4
Genetic identification	APMV4	APMV6 + APMV4

* sera specifications available on request

Combining the advantages of random amplification and massive parallel sequencing, 5,225 and 12,310 sequence reads were produced from the library resulting respectively from sample mallard/Belgium/12245/07 and mallard/Belgium/15129/07. More than 95 % of these reads were specific for APMVs, and host-derived or contaminating sequences were negligible.

Assembly of random generated sequences for sample mallard/Belgium/15129/07 produced a 15,054 nucleotides (nt) contig representing the complete genome sequence of an APMV4. APMV4/mallard/Belgium/15129/07 (APMV4-BE15129) was assembled from 9,767 sequence reads of raw data (APMV4/KR/YJ/06 [GenBank: EU877976] used as reference in the reference assembly). Assembly of 4,715 sequences generated for sample mallard/Belgium/12245/07 produced a nearly complete (98.89%) APMV6 genome of length 16,236 nt (APMV6/mallard/Belgium/12245/07; APMV6-BE12245). APMV6/Goose/FarEast/4440/2003 [GenBank: EF569970] was used as a reference sequence in this reference assembly. Surprisingly, APMV4 sequences were also identified in sample mallard/Belgium/12245/07. APMV4/KR/YJ/06 [GenBank: EU877976] was used as a reference and 21 sequences mapped to various regions (total of 2,977 nt representing 19.75 % of the APMV4 genome sequence). The APMV4 virus was named APMV4/mallard/Belgium/12245/07 (APMV4-BE12245). Unfortunately the original individual cloacal swabs were no longer available at the time of the genetic analysis, so we could not find out which of the four animals in the pool were infected and whether we were dealing with a mixed infection of one bird. The missing 1.11 % of the APMV6 genome represents two small internal gaps and some nucleotides (24 nt at 5’ and 37 nt at 3’) at the genome termini. A low coverage at the genome termini was also observed for the fully sequenced APMV4 genome (41 terminal 5’ nt and 42 terminal 3’ nt, with depth ≤ 3x).

Database accession numbers

The consensus sequences were submitted to GenBank under the following accession numbers: JN571485 (APMV4/mallard/Belgium/15129/07, complete genome), JN571486 (APMV6/mallard/Belgium/12245/07, nearly complete genome) and JN571487, JN571488, JN571489, JN571490 (APMV4/mallard/Belgium/12245/07, partial sequences of phosphoprotein, fusion protein, hemagglutinin-neuraminidase and large polymerase genes).

Genomic features of APMV4/mallard/Belgium/15129/07

The virus has a genome length of 15,054 nt as previously described for APMV4 viruses, consisting of six transcriptional units (Table 2) encoding from 3’ to 5’ the NP (nucleoprotein), P/V/W (phosphoprotein and additional proteins through RNA editing), M (matrix), F (fusion), HN (hemagglutinin-neuraminidase) and L (large polymerase) proteins. The 3’ leader and 5’ trailer sequences of the genome were respectively 55 nt and 17 nt in length. Gene start and gene end sequences were as previously described for APMV4 [30]. The NP protein encoded a 457 amino acids (aa) protein, as previously described for other APMV4. The P gene encodes a 393 aa phosphoprotein. A putative RNA editing site at genome position 2,057-2,065 (5’AAAGGGGGG-3’) was identified, where insertion of one non-templated G residue would encode a 224 aa V protein. Alternatively, the insertion of two non-templated G residues would result in a putative W protein of 137 aa. The matrix gene open reading frame (ORF) encodes a 370 aa long matrix protein, unlike the 367 aa or 369 aa previously described for APMV4 genomes [29, 30]. The lengths of the other proteins encoded by their ORF’s are the same as previously described for APMV4 (F 566 aa; HN 569 aa; L 2,211 aa). The fusion protein has a monobasic cleavage site (DIQPR↓F).

Table 2: APMV4/mallard/Belgium/15129/07 genome organization and characterization (genome size 15,054 nt).

Gene	Genome position	5'UTR (nt) (incl. gene-start)	ORF(nt)	3'UTR (nt) (incl. gene-end)	Intergenic region (nt)	Deduced protein size (aa)
NP	56-1,606	60	1,374	117	9	457
P (V; W)*	1,616-2,979	46	1,182 (675; 413)	136	34	393 (224; 137)
M	3,014-4,306	77	1,113	103	16	370
F	4,323-6,210	71	1,701	116	40	566
HN	6,251-8,161	78	1,698	135	45	565
L	8,207-15,037	92	6,636	103		2,211

*putative RNA editing site at position 2,057-2,065 (5'-AAAGGGGGG-3')

Genomic features of APMV6/mallard/Belgium/12245/07

The genome length of 16,236 nt is consistent with that of "class I" of APMV6 [32], containing seven transcriptional units (Table 3) encoding from 3' to 5' the NP, P/V/W , M, F, SH (small hydrophobic protein), HN and L proteins. The F protein has a monobasic cleavage site, PEP<u>R</u>↓L. The 3' leader and 5' trailer sequences of the genome were respectively 55 and 54 nt in length. Gene start and gene end sequences were as previously described for APMV6 [32]. The lengths of the proteins encoded by the ORF's are the same as previously described for APMV6 (NP 465 aa; P 430 aa, V 268 aa, W 177 aa, M 366 aa, F 555 aa, 142 aa, HN 613 aa, L 2,241 aa).

Table 3: APMV6/mallard/Belgium/12245/07 genome organization and characterization (genome size 16,236 nt).

Gene	Genome position**	5'UTR (nt) (incl. gene-start)	ORF	3'UTR(nt) (incl. gene-end)	Intergenic region (nt)	Deduced protein size (aa)
NP	56-1,626	72	1,398	101	7	465
P (V; W)*	1,634-,3119	53	1,293 (807; 534)	140	2	430 (268; 177)
M	3,122-4,526	113	1,101	191	59	366
F	4,586-6,420	12	1,668	155	49	555
SH	6,470-7,043	72	429	73	28	142
HN	7,072-9,102	50	1,842	139	63	613
L	9,166-16,182	112	6,726	179		2,241

* putative RNA editing site at position 2,148-2,156 (5'-AAAAAAGGG-3')

** The indicated position is relative to the position of the used reference APMV6/Goose/FarEast/4440/2003 [GenBank: EF569970]

Phylogenetic analysis based on F and HN proteins

Phylogenetic trees based on amino acid sequence alignments of the F and HN proteins clearly classify APMV4-BE15129 and APMV6-BE12245 within respectively serotype APMV4 and APMV6 (Figure 1 and Figure 2). APMV6-BE12245 is most closely related to the “class I” of APMV6 viruses described by Xiao and colleagues [32]. This is confirmed by its high whole genome nucleotide sequence identity with APMV6/Goose/FarEast/4440/2003 (GenBank: EF569970; Table 4). The F and HN amino acid sequences of APMV4-BE15129 are most closely related to APMV4/KR/YJ/06 (GenBank: EU877976), which is confirmed by a high whole genome nucleotide homology to this virus (Table 5). APMV4-BE15129 is more closely related to both previously sequenced APMV4 whole genomes than these are to each other (Table 5).

Table 4: APMV6 complete genomes nucleotide identity matrix. Sites with missing data were deleted from the alignment.

	Hong Kong/18/199/77	**Italy/4524-2/07**	**Taiwan/Y1/98**	**FarEast/4440/03**	**APMV-6 BE/12245/07**
APMV-6/duck/HongKong/18/199/77	ID				
APMV-6/duck/Italy/4524-2/07	70%	ID			
APMV-6/duck/Taiwan/Y1/98	94.2%	70.2%	ID		
APMV-6/Goose/FarEast/4440/2003	93.8%	70.3%	98.3%	ID	
APMV-6/Anas platyrhynchos/BE/12245/2007	91.9%	69.2%	96%	96.5%	ID

Table 5: APMV4 complete genomes nucleotide identity matrix.

	Hong Kong/D3/75	**KR/YJ/06**	**BE/15129/2007**
APMV4/duck/Hong Kong/D3/75	ID		
APMV4/KR/YJ/06	91.6%	ID	
APMV4/Anas platyrhynchos/BE/15129/2007	91.9%	97.6%	ID

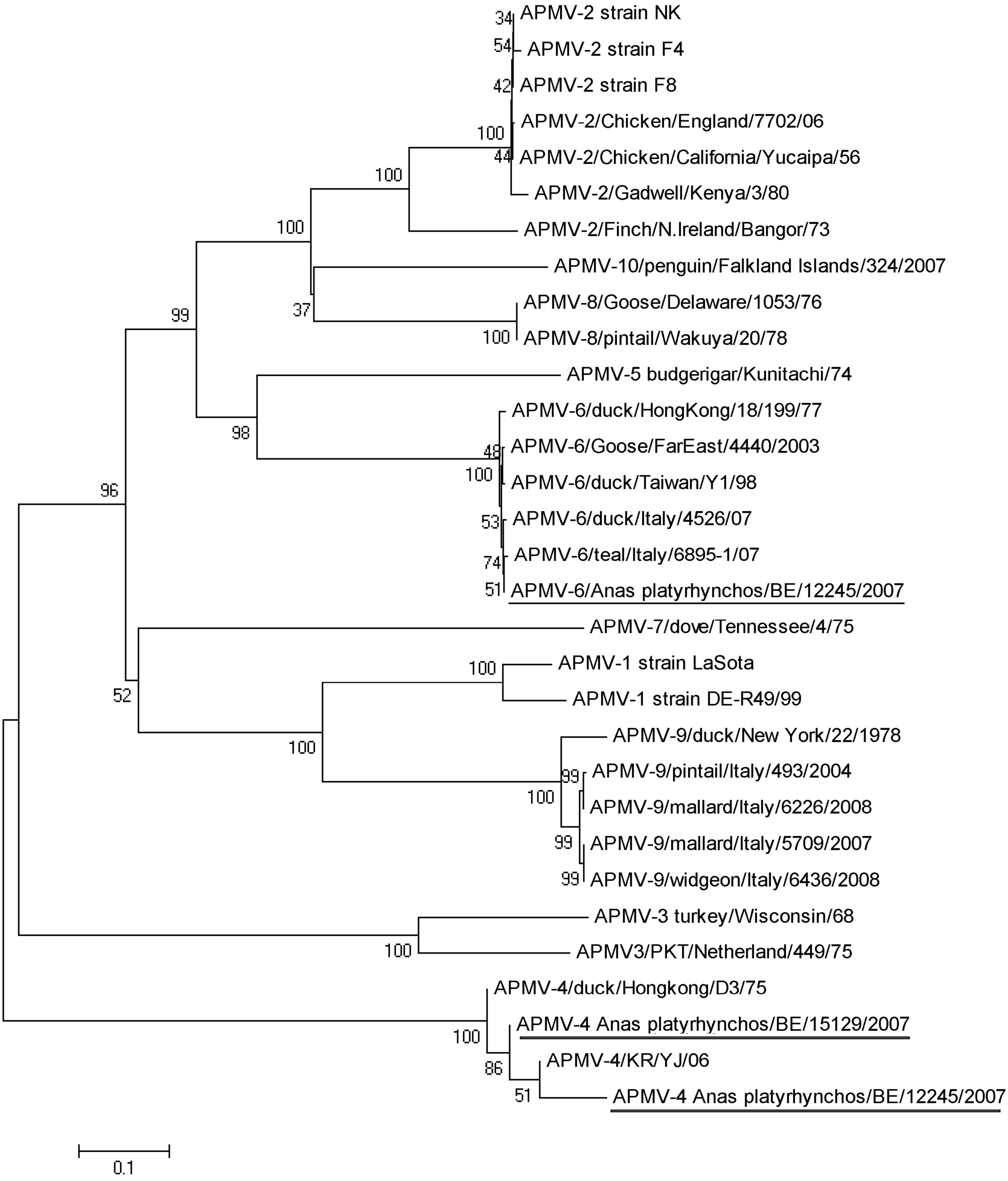

34
APMV-2 strain NK
54
APMV-2 strain F4
42
APMV-2 strain F8
APMV-2/Chicken/England/7702/06
100
44
APMV-2/Chicken/California/Yucaipa/56
100
APMV-2/Gadwell/Kenya/3/80
100
APMV-2/Finch/N.Ireland/Bangor/73
APMV-10/penguin/Falkland Islands/324/2007
99
37
APMV-8/Goose/Delaware/1053/76
100
APMV-8/pintail/Wakuya/20/78
APMV-5 budgerigar/Kunitachi/74
APMV-6/duck/HongKong/18/199/77
98
48
APMV-6/Goose/FarEast/4440/2003
100
APMV-6/duck/Taiwan/Y1/98
96
53
APMV-6/duck/Italy/4526/07
74
APMV-6/teal/Italy/6895-1/07
51
APMV-6/Anas platyrhynchos/BE/12245/2007
APMV-7/dove/Tennessee/4/75
100
APMV-1 strain LaSota
52
APMV-1 strain DE-R49/99
APMV-9/duck/New York/22/1978
100
99
APMV-9/pintail/Italy/493/2004
100
APMV-9/mallard/Italy/6226/2008
99
APMV-9/mallard/Italy/5709/2007
99
APMV-9/widgeon/Italy/6436/2008
APMV-3 turkey/Wisconsin/68
100
APMV3/PKT/Netherland/449/75
APMV-4/duck/Hongkong/D3/75
100
APMV-4 Anas platyrhynchos/BE/15129/2007
86
APMV-4/KR/YJ/06
51
APMV-4 Anas platyrhynchos/BE/12245/2007
0.1

Figure 1: Phylogenetic tree based on the aa sequences of the F protein.

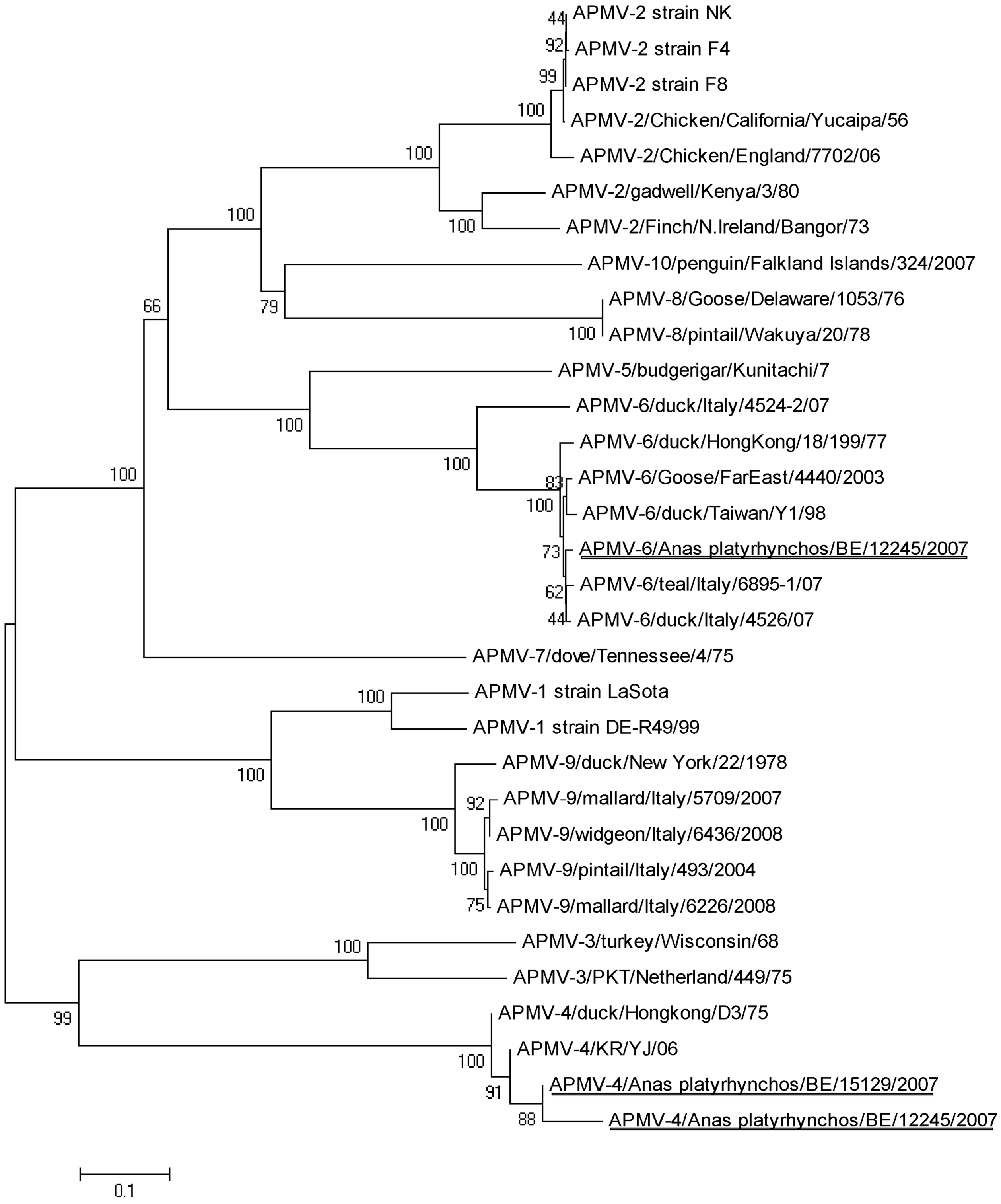

44
APMV-2 strain NK
92
APMV-2 strain F4
99
APMV-2 strain F8
100
APMV-2/Chicken/California/Yucaipa/56
100
APMV-2/Chicken/England/7702/06
100
APMV-2/gadwell/Kenya/3/80
100
APMV-2/Finch/N.Ireland/Bangor/73
APMV-10/penguin/Falkland Islands/324/2007
66
79
APMV-8/Goose/Delaware/1053/76
100
APMV-8/pintail/Wakuya/20/78
APMV-5/budgerigar/Kunitachi/7
APMV-6/duck/Italy/4524-2/07
100
APMV-6/duck/HongKong/18/199/77
100
100
83
APMV-6/Goose/FarEast/4440/2003
100
APMV-6/duck/Taiwan/Y1/98
73
APMV-6/Anas platyrhynchos/BE/12245/2007
62
APMV-6/teal/Italy/6895-1/07
44
APMV-6/duck/Italy/4526/07
APMV-7/dove/Tennessee/4/75
100
APMV-1 strain LaSota
APMV-1 strain DE-R49/99
APMV-9/duck/New York/22/1978
100
92
APMV-9/mallard/Italy/5709/2007
100
APMV-9/widgeon/Italy/6436/2008
100
APMV-9/pintail/Italy/493/2004
75
APMV-9/mallard/Italy/6226/2008
100
APMV-3/turkey/Wisconsin/68
APMV-3/PKT/Netherland/449/75
99
APMV-4/duck/Hongkong/D3/75
APMV-4/KR/YJ/06
100
APMV-4/Anas platyrhynchos/BE/15129/2007
91
88
APMV-4/Anas platyrhynchos/BE/12245/2007
0.1

Figure 2: Phylogenetic tree based on the aa sequences of the HN protein.

Although no complete F and HN sequences were available for APMV4-BE12245 (in sample with APMV6-BE12245), we included the partial sequence information in the phylogenetic analysis using pairwise deletion of positions with gaps and missing data. This may have resulted in biased distance estimations. However, there are clear indications that although it is most closely related to APMV4- BE15129, it is not identical. This is also evident from nucleotide sequence identity calculated over all available sequence information (2,977 nt) for the partial genome APMV4-BE12245 (Table 6). The partial sequence APMV4-BE12245 is 98.4% identical to APMV4- BE15129 considering all positions allowed by the partial sequence of APMV-BE12245. In contrast, its identity with previously sequenced APMV4 genomes is only 97.5 % (APMV4/KR/YJ/06; GenBank: EU877976) and 90.9 % (APMV4/duck/Hong Kong/D3/75; GenBank: FJ177514).

Table 6: Nucleotide identity of partial genome sequence (2,977nt) of APMV4/mallard/Belgium/12245/07 with APMV4/mallard/Belgium/15129/07 and other APMV4 genomes. Sites with missing data were deleted from the alignment.

	Hong Kong/D3/75	**KR/YJ/06**	**BE/15129/2007**	**APMV-4 BE/12245/2007**
APMV4/duck/Hong Kong/D3/75	ID			
APMV4/KR/YJ/06	91.4%	ID		
APMV4/Anas platyrhynchos/BE/15129/2007	91.3%	97.5%	ID	
APMV4/Anas platyrhynchos/BE/12245/2007	90.9%	97.1%	98.4%	ID

Discussion

Wild birds are increasingly recognized as a reservoir for important livestock diseases. This has been extensively shown for avian influenza A viruses (AIV) and to a lesser degree for avian paramyxoviruses of serotype 1 (APMV1). Moreover, other viruses, including APMV2-10 have been shown to circulate in wild birds. Some of these viruses have been shown to infect poultry species and induced major outbreaks in flocks.

Apart from the well-characterized serotype APMV1 associated with the economically important Newcastle disease in poultry, knowledge of the antigenic and genetic diversity in the APMV serotypes of the genus *Avulavirus* is limited. The determination of complete genome sequences of an additional APMV4 and APMV6 widens our understanding of the genetic diversity in these serotypes. Interestingly, we could identify two different viruses from single pooled samples. In one tested pool of four cloacal swabs, taken in beginning of September, at least one of the four animals was infected with an APMV4. In the other tested pool, taken at the end of this month in the same capture location, two different APMV serotypes APMV6 and APMV4 were identified. The latter APMV4, although closely related to the APMV4 in the first pool, was not identical to it. Contamination artifacts during virus isolation are very unlikely to have occurred as the two APMV4 viruses characterized in this study are not identical based on the sequence information (2,977 nt partial sequence of APMV4-BE12245) obtained, and no other APMV4 viruses were manipulated in the laboratory.

It is difficult to assess whether both APMV4 viruses characterized in this study fall within the normal range of quasispecies genetic variation. This is because of the limited availability of sequence information for this serotype and the lack of studies investigating the genetic variability within circulating populations of paramyxoviruses. To prove the economic feasibility of the method of random amplification combined with deep sequencing, the number of sequence reads per sample was intentionally kept below 10 000 in this study. This turned out to be sufficient for the completion of the APMV4 genome in one pool. In the mixed APMV infected pool, this number of reads did not allow the determination of the last 1.11 % of the APMV6 genome because part of the sequencing effort resulted in 19.75% of the genome of a co-infecting APMV4. Most probably, the APMV4 virus was present in a lower amount in the original samples, and a higher number of sequence reads would have resulted in completion of the APMV6 genome. However, we cannot fully exclude preferential growth of either virus during virus isolation or a slight bias in our random amplification protocol. This means that quantitative statements about the relative presence of either virus in the original pooled sample based on the distribution of sequence reads are not possible. As the original swabs were no longer available, we could not determine (1) in which proportion the two viruses were present in the original sample/pool before the propagation in eggs, (2) which of the four animals in the pool were infected and (3) whether we were dealing with a mixed infection of one bird. Moreover, the analytical sensitivity of the method remains to be determined and may limit the applicability to field samples containing relatively high virus titers. The presented methodology has the potential to identify viruses present in minor proportions in a pooled sample, and mixed infections in single samples. Clearly our methodology, using a sequence independent methodology for genome determination, has allowed the detection of sequence information from both viruses without bias. In contrast, the use of serotype specific tests such as HI or serotype specific PCR methods may fail to characterize the full complexity of an isolate. Further passage of “double isolates” may give a selective advantage to either virus, changing the biological properties of the isolate, as was suggested by Shihmanter and colleagues [42]. They described that an APMV1 had a selective advantage over co-infecting APMV viruses during passaging in embryonated chicken eggs.

Our genetic identification of the APMVs revealed some difficulties in the HI based identification of APMVs other than APMV1. The APMV6 reference serum did detect the APMV6 virus in sample 07/12245 (titer 1/64) and the APMV4 reference serum detected the APMV4 virus in sample 07/15129 (titer 1/128). However, the HI test failed to detect the APMV4 virus co-present at low titer with the APMV6 virus in pooled sample 07/12245. This most likely indicates that our molecular method is much more sensitive to the identification of viruses present at very low concentrations. Additionally, a cross reactivity with the APMV2 reference serum P/Robin/Hiddensee/57 was observed for both samples (titer 1/32 or 1/64 – Table 1). However another APMV2 reference serum P/chicken/Yucaipa/Cal/56 did not show cross reactivity with these samples, which makes the HI subtyping interpretation difficult. In the context of mixed infections, where it’s likely that one virus has a higher concentration than the other, genetic information seems more informative for the identification. Further studies are obviously needed to gain insight in the genetic and antigenic diversity of APMV2-10.

Recently Xiao and colleagues [32] increased the amount of whole genome sequences available for APMV6 to six, identifying two classes with APMV6. APMV6 class I isolates differed less than five % from each other but differed 29-31 % to the single class II isolate IT4524-2. The additional APMV6 genome identified in this study clustered within class I, maintaining the separation with class II (31 % distance) while slightly increasing the genetic diversity within class I to a maximum of 8 % distance.

On the other hand, whole genome sequences of only two representative strains of APMV4 have been reported so far [29, 30]. The complete genome of APMV4-BE15129 determined in this study further extends our knowledge of this serotype. This additional APMV4 complete genome does not increase the maximum genetic distance previously documented within the APMV4 serotype. The genetic distance now ranges from two to eight % nucleotide sequence distance (based on only three complete genome sequences). The amount of sequence data compared to APMV1 remains low and further studies are needed to get a better estimate of genetic diversity within serotypes APMV2-10. The sequencing methodology used in this study may facilitate this.

The genome length of 15,054 nt for APMV4 and 16,236 nt for APMV6 complies with the “rule of six*” for efficient genome replication of *Paramyxovirinae* [43]. The genomic characteristics and genome organizations, including putative mRNA editing of the P gene, are as previously described for APMV4 and APMV6 genomes [29, 30, 32, 33]. Further variability in protein length of the APMV4 M protein was shown. Variability in the intergenic sequence length, as is known for the genus *Avulavirus*, was also confirmed here. A monobasic fusion protein cleavage site was present in both viruses. However, fusion protein cleavage site sequences in APMV2-9 are not necessarily predictive of protease activation phenotype [33], as it is in Newcastle disease virus [44]. Interestingly, the terminal amino acid of the fusion protein cleavage site of APMV4/mallard/Belgium/15129/07 is a phenylalanine. As previously shown for other APMV4 [29, 30], this did not require an exogenous exonuclease for in vitro replication on chicken embryonic fibroblasts [29]. A phenylalanine at this position is known to contribute to the in vitro growth characteristics and in vivo pathogenicity of velogenic Newcastle disease [4]. Further in vivo and in vitro phenotypic characterization of this virus would be interesting.

This study clearly demonstrates the value of a sequencing strategy combining next generation sequencing and random access amplification for the identification and whole genome determination of APMVs. Although the method allows sequencing of complete APMV genomes, an unequal distribution of sequencing depth results in low coverage at the genome termini when only a modest sequencing effort is applied. Efforts to optimize the homogenous distribution of sequencing reads along the genome and to determine the optimal sequencing effort for reproducible whole genome sequencing, could further improve the applicability of the method. Previous studies determining complete genomes of APMV2-9 often relied on a round of amplification using degenerated or custom designed oligonucleotides, followed by primer walking [29, 31-35]. The use of random access amplification alleviates the problem of oligonucleotide design in a context of poor representation in sequence databases. Moreover, it allows for the identification of potential co-infection with other APMVs or other viruses

** Many paramyxovirus genomes follow the "rule of six". The total length of the genome is almost always a multiple of six. This is probably due to the advantage of having all RNA bound by NP protein (since NP binds hexamers of RNA). If RNA is left exposed, the virus does not replicate efficiently.*

without methodological bias. Sequence independent single primer amplification (SISPA) was originally described by Reyes and Kim [45]. It was later modified to include enrichment steps for viral nucleic acids using filtration and nuclease treatment (DNase-SISPA, [36, 37]). Miller and colleagues [5] used a similar approach for the identification and sequencing of a new serotype of APMV10 in penguins. Unlike their method, that relied on the molecular cloning and sequencing of hundreds of random amplicons, this study used the power of next generation to provide the necessary sequence information. The preparation of a next generation sequencing library includes the process of emulsion PCR, which isolates single DNA molecules on beads and clonally amplifies them ([38], reviewed in [46]). There is no longer a need for molecular cloning and the generated random amplicons can directly be processed in the sequencing library workflow. An additional advantage is that this methodology avoids biological biases induced by the virological analysis of mixed infections.

Conclusion

Within a single sampling location, three different APMVs were identified in wild mallards using random access amplification in combination with next generation sequencing. From one pooled sample, the complete genome sequence of an APMV4 was assembled from the random sequences. From a second pooled sample, the nearly complete genome sequence of an APMV6 (genome size of 16,236 nt) was determined, as well as a partial sequence for an APMV4 closely related but not identical to the APMV4 virus isolated from the first sample.

These data further contribute to the knowledge about the genetic diversity within serotypes APMV4 and APMV6. Moreover, this study demonstrates the value of a random access nucleic acid amplification method in combination with massive parallel sequencing for the characterization and full genome sequencing of APMVs. Moreover, the sequence independent nature of this method allows the detection of potential co-infections with other viruses and is applicable to other viruses.

Acknowledgements

We thank N. Dumont and M. Decaesstecker for excellent technical assistance and Dr. Dirk Höper (Friedrich Loeffler Institute, Germany) for helpful discussions. The professional services of the Genomics Core of the University Hospital of Leuven are greatly appreciated. This project was supported by internal research and development grant *RandSeq09* of the Veterinary and Agrochemical Research Center.

References

1. Lamb, R.A. and G. Parks, *Paramyxoviridae: the viruses and their replication 5th edition.* 2007, Philadelphia: Lippincott Williams and Wilkins.
2. Lamb, R.A., et al., *Family paramyxoviridae*, in *Virus Taxonomy: The classification and nomenclature of viruses. The eighth report of the international committee in taxonomy of viruses*, C.M. Fauquet, Editor. 2005.
3. Alexander, D.J., et al., *Characterization of viruses which represent further distinct serotypes (PMV-8 and PMV-9) of avian paramyxoviruses.* Arch Virol, 1983. **78**(1-2): p. 29-36.
4. Alexander, D.J., *Newcastle disease*, in *Disease of poultry*, Y.M. Saif, et al., Editors. 2003, Iowa State Press: Ames IA. p. 64-87.
5. Miller, P.J., et al., *Evidence for a New Avian Paramyxovirus Serotype-10 Detected in Rockhopper Penguins from the Falkland Islands.* J Virol, 2010. **84**(21): p. 11496-11504.
6. OIE, *Terrestrial animal health code.* World organisation for animal health, 2010.
7. Alexander, D.J. and M.S. Collins, *Pathogenicity of PMV-3/parakeet/Netherlands/449/75 for chickens.* Avian Pathol, 1982. **11**(1): p. 179-85.
8. Warke, A., et al., *Comparative study on the pathogenicity and immunogenicity of wild bird isolates of avian paramyxovirus 2, 4, and 6 in chickens.* Avian Pathol, 2008. **37**(4): p. 429-34.
9. Saif, Y.M., et al., *Natural and experimental infection of turkeys with avian paramyxovirus-7.* Avian Dis, 1997. **41**(2): p. 326-9.
10. Tumova, B., et al., *A further member of the Yucaipa group isolated from the common wren (Troglodytes troglodytes).* Acta Virol, 1979. **23**(6): p. 504-7.
11. Bankowski, R.A., J. Almquist, and J. Dombrucki, *Effect of paramyxovirus yucaipa on fertility, hatchability, and poult yield of turkeys.* Avian Dis, 1981. **25**(2): p. 517-20.
12. Redmann, T., et al., *[Isolation of a paramyxovirus-3 from turkeys with respiratory tract disease in Germany]*. Dtsch Tierarztl Wochenschr, 1991. **98**(4): p. 138-41.
13. Zhang, G.Z., et al., *Isolation, identification, and comparison of four isolates of avian paramyxovirus serotype 2 in China.* Avian Dis, 2006. **50**(3): p. 386-90.
14. Alexander, D.J., *Newcastle disease and other avian paramyxoviridae infections.*, in *Diseases of poultry*, B.W. Calnek, Editor. 1997, Iowa State University Press: Ames. p. 541-569.
15. Jung, A., et al., *Avian paramyxovirus serotype 3 infection in Neopsephotus, Cyanoramphus, and Neophema species.* J Avian Med Surg, 2009. **23**(3): p. 205-8.
16. Nerome, K., et al., *Isolation of a new avian paramyxovirus from budgerigar (Melopsittacus undulatus).* J Gen Virol, 1978. **38**(2): p. 293-301.
17. Gough, R.E. and D.J. Alexander, *Avian paramyxovirus type 4 isolated from a ringed teal (Calonetta leucophrys).* Vet Rec, 1984. **115**(25-26): p. 653.
18. Stallknecht, D.E., et al., *Avian paramyxoviruses from migrating and resident ducks in coastal Louisiana.* J Wildl Dis, 1991. **27**(1): p. 123-8.
19. Maldonado, A., et al., *Serological survey for avian paramyxoviruses from wildfowl in aquatic habitats in Andalusia.* J Wildl Dis, 1995. **31**(1): p. 66-9.
20. Capua, I., et al., *Isolation of an avian paramyxovirus type 9 from migratory waterfowl in Italy.* Vet Rec, 2004. **155**(5): p. 156.
21. Alexander, D.J., *The classification, host range and distribution of avian paramyxoviruses*, in *Acute virus infections of poultry*, J.B. McFerran and M.S. MCNulty, Editors. 1986, Martinus Nijhoff Publishers: The Netherlands. p. 52-66.

22. Stanislawek, W.L., et al., *Avian paramyxoviruses and influenza viruses isolated from mallard ducks (Anas platyrhynchos) in New Zealand.* Arch Virol, 2002. **147**(7): p. 1287-302.
23. Shortridge, K.F. and D.J. Alexander, *Newcastle disease virus surveillance in Hong Kong on local and imported poultry.* Res Vet Sci, 1978. **25**(2): p. 204-6.
24. Turek, R., M. Gresikova, and B. Tumova, *Isolation of influenza A virus and paramyxoviruses from sentinel domestic ducks.* Acta Virol, 1984. **28**(2): p. 156-8.
25. Chang, P.C., et al., *Complete nucleotide sequence of avian paramyxovirus type 6 isolated from ducks.* J Gen Virol, 2001. **82**(Pt 9): p. 2157-68.
26. Subbiah, M., et al., *Complete sequence of the genome of avian paramyxovirus type 2 (strain Yucaipa) and comparison with other paramyxoviruses.* Virus Res, 2008. **137**(1): p. 40-8.
27. Kumar, S., et al., *Complete genome sequence of avian paramyxovirus-3 strain Wisconsin: evidence for the existence of subgroups within the serotype.* Virus Res, 2010. **149**(1): p. 78-85.
28. Kumar, S., et al., *Complete genome sequence of avian paramyxovirus type 3 reveals an unusually long trailer region.* Virus Res, 2008. **137**(2): p. 189-97.
29. Nayak, B., et al., *Molecular characterization and complete genome sequence of avian paramyxovirus type 4 prototype strain duck/Hong Kong/D3/75.* Virol J, 2008. **5**: p. 124.
30. Jeon, W.J., et al., *Full-length genome sequence of avain paramyxovirus type 4 isolated from a mallard duck.* Virus Genes, 2008. **37**(3): p. 342-50.
31. Samuel, A.S., et al., *Complete genome sequence of avian paramyxovirus (APMV) serotype 5 completes the analysis of nine APMV serotypes and reveals the longest APMV genome.* PLoS One, 2010. **5**(2): p. e9269.
32. Xiao, S., et al., *Complete genome sequences of avian paramyxovirus serotype 6 prototype strain Hong Kong and a recent novel strain from Italy: evidence for the existence of subgroups within the serotype.* Virus Res, 2010. **150**(1-2): p. 61-72.
33. Xiao, S., et al., *Complete genome sequence of avian paramyxovirus type 7 (strain Tennessee) and comparison with other paramyxoviruses.* Virus Res, 2009. **145**(1): p. 80-91.
34. Paldurai, A., et al., *Complete genome sequences of avian paramyxovirus type 8 strains goose/Delaware/1053/76 and pintail/Wakuya/20/78.* Virus Res, 2009. **142**(1-2): p. 144-53.
35. Samuel, A.S., et al., *Complete sequence of the genome of avian paramyxovirus type 9 and comparison with other paramyxoviruses.* Virus Res, 2009. **142**(1-2): p. 10-8.
36. Djikeng, A., et al., *Viral genome sequencing by random priming methods.* BMC Genomics, 2008. **9**: p. 5.
37. Allander, T., et al., *A virus discovery method incorporating DNase treatment and its application to the identification of two bovine parvovirus species.* Proc Natl Acad Sci U S A, 2001. **98**(20): p. 11609-14.
38. Margulies, M., et al., *Genome sequencing in microfabricated high-density picolitre reactors.* Nature, 2005. **437**(7057): p. 376-80.
39. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
40. Hall, T.A., *BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT.* http://www.mbio.ncsu.edu/bioedit/bioedit.html. Nucl Acids Symp Ser, 1999. **41**: p. 95-98.

41. Tamura, K., et al., *MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods.* Mol Biol Evol, 2011. **Epub ahead of print**.
42. Shihmanter, E., et al., *Mixed paramyxovirus infection of wild and domestic birds in Israel.* Vet Microbiol, 1997. **58**(1): p. 73-8.
43. Kolakofsky, D., et al., *Paramyxovirus RNA synthesis and the requirement for hexamer genome length: the rule of six revisited.* J Virol, 1998. **72**(2): p. 891-9.
44. Morrison, T., et al., *The role of the amino terminus of F1 of the Newcastle disease virus fusion protein in cleavage and fusion.* Virology, 1993. **193**(2): p. 997-1000.
45. Reyes, G.R. and J.P. Kim, *Sequence-independent, single-primer amplification (SISPA) of complex DNA populations.* Mol Cell Probes, 1991. **5**(6): p. 473-81.
46. Metzker, M.L., *Sequencing technologies - the next generation.* Nat Rev Genet, 2010. **11**(1): p. 31-46.

CHAPTER 3.3

Phylogeographic analysis of avian influenza viruses isolated from *Charadriiformes* in Belgium confirms intercontinental reassortment in gulls

Steven Van Borm, Toon Rosseel, Didier Vangeluwe, Frank Vandenbussche, Thierry van den Berg and Bénédicte Lambrecht

Published in Archives of Virology (2012) 157(8): p. 1509-22.

As a second case study, the workflow was used for the molecular characterization of 9 different avian influenza virus isolates (segmented -ssRNA genome organization) from a surveillance program in wild gulls and shorebirds.

Avian influenza viruses (AIV) are influenza type A viruses belonging to the *orthomyxoviridae* family containing a segmented negative-stranded RNA genome (8 segments). Influenza A viruses are divided into subtypes on the basis of two proteins on the surface of the virus, hemagglutinin (HA) and neuraminidase (NA). There are 18 known HA subtypes and 11 known NA subtypes. Many different combinations of HA and NA proteins are possible. Three prominent subtypes of AIV are known to infect both birds and humans, influenza A H5, H7 and H9. Avian influenza A viruses are designated as low pathogenicity avian influenza (LPAI) or highly pathogenic avian influenza (HPAI).

Abstract

Nine influenza viruses isolated from gulls and shorebirds in Belgium (2008-2010), including H3N8, H5N2, H6N1, H11N9, H12N5, H13N6, H13N8, and H16N3 subtypes, were targeted using random amplification and next-generation sequencing. The gene segments of these viruses segregated into three phylogeographic lineage types: (1) segments circulating in waterfowl in Eurasia with sporadic introduction in other species and in the Americas ("Eurasian avian"), (2) segments circulating in American waterfowl with sporadic introduction to other species and regions ("American avian"), and (3) segments circulating exclusively in gulls and shorebirds and having increased connectivity between the two hemispheres ("*Charadriiformes* specific"). Notably, an H6N1 and an H5N2 isolated from *L. argentatus* had mainly Eurasian avian genes but shared a matrix segment of American avian origin (first documentation in European gulls of transhemispheric reassortment). These data support the growing evidence of an important role of *Charadriiformes* birds in the dynamic nature of avian influenza ecology.

Introduction

Avian influenza A viruses (AIV, *Orthomyxoviridae*) possess a segmented single stranded RNA genome with negative polarity. Eight unlinked RNA segments encode 11 proteins [1, 2], including the hemagglutinin (HA) and neuraminidase (NA) surface glycoproteins. 16 HA and 9 NA subtypes have been described [3, 4], of which the HA16 and HA13 seem to be specific to birds of the order *Charadriiformes* (gulls, terns and shorebirds) [4]. Genetic diversity is mainly generated by mutation and reassortment of complete RNA segments during coinfection with multiple viruses [1, 5]. The genetic diversity of AIV is maintained by circulation in aquatic wild birds. The highest prevalence of avian influenza infection is observed in birds belonging to the orders of *Anseriformes* (geese, ducks and swans) and *Charadriiformes* (gulls, terns and shorebirds) [6]. These viruses are occasionally transmitted to other species like domesticated poultry and mammals [1] where they can establish themselves, including strains with pandemic potential [7-9]. Regarding avian influenza, wild birds typically carry low-pathogenic strains (LPAI), some of which - confined to subtypes H5 and H7 - can mutate and produce highly pathogenic forms (HPAI) after introduction into poultry species [1].

Geographical constraints on migratory bird movements result in a broad division of all genes of avian influenza A viruses into two large phylogenetic lineages evolving independently, the American and the Eurasian lineages (including Oceania and Africa) [1, 6]. Within hemispherically separated gene pools, wild migratory bird ecology contributes to frequent interspecies reassortment of gene segments resulting in transient ‘‘genome constellations’’ in wild birds. These gene segment combinations are continually reshuffled by reassortment, in contrast to the spread of a limited number of stable genome constellations that characterizes the evolution of mammalian-adapted influenza A viruses [10]. Exchange of AIV gene segments between these geographically separated gene pools is a relatively rare event [10-17]. Although AIV prevalence in *Charadriiformes* is typically low [18], genomic data of AIV isolated in these birds in North America provide evidence of intercontinental reassortment across the North Atlantic ocean [17, 19] and across the Bering Sea between Alaska and Russia [13, 15, 20, 21], representing mostly areas along continental margins where migratory flyways overlap. Interspecies transmission of AIV may occur in wild birds at migration stopover sites or breeding habitats where flocks of mixed species congregate [20, 22]. These introduced Eurasian AIV genome segments in the Americas are sporadically established in other species of the wild avifauna [17, 19]. On the contrary, North American origin genes in Europe are less frequently detected [11, 17]. The current evidence of American origin gene segments in Eurasia is limited to individual cases in guillemots in Sweden [12], a tern in Australia [23], ducks in Japan[14, 24] and India [25], and domestic ducks in Italy [26]. Entire genomes of Eurasian origin in America or vice versa have not been reported. Although transhemispheric movement of gene segments is relatively infrequent [11], implications for AIV population structure may be large. For example, all currently circulating PB1, PA and HA of the H6 lineages in North America have a Eurasian origin and have replaced previously circulating American lineages [27].

In addition to sporadically demonstrated transhemispheric reassortment events in AIV isolated from *Charadriiformes*, genetic lineages exist that are specific to *Charadriiformes* [6, 28, 29]. For example, the H13 and H16 hemagglutinin genes seem to occur almost exclusively in *Charadriiformes* birds [4, 15, 30]. *Charadriiformes* specific clades are also documented in PB2, PB1, PA, NP, M, NS (allele B) genes [17]. These *Charadriiformes* specific H13 and H16 viruses have all been suggested to have genomes with a mosaic of geographic origins [10, 11], and increased intercontinental mobility of gene segments compared to other AIV occurring in wild birds. *Charadriiformes*, however are also infected with viruses whose sequences are shared with those from other wild bird hosts.

During the annual active surveillance for avian influenza in wild birds in Belgium from 2008 to 2010, several avian influenza viruses were isolated from shorebirds (*Arenaria interpres*, Ruddy Turnstone) and gulls (*Larus* sp.). In this study, we targeted the complete genome of these viruses using next-generation sequencing technology and compare the genetic data with sequences available in public databases in a phylogeographic analysis framework to provide additional insights in the evolutionary ecology of avian influenza viruses in *Charadriiformes* in Western Europe.

Materials and Methods

Viruses

We selected nine viruses isolated from the target species *Arenaria interpres* (Ruddy Turnstone) and gull species, including *Larus argentatus* (Herring Gull) and *Larus michahellis* (Yellow-legged Gull), for whole genome sequencing (Table 1). The HA and NA subtype diversity in these samples included: H3N8, H5N2, H6N1, H11N9, N13N6, H13N8, and H16N3 (Table 1). These virus isolates originated in ongoing active AIV monitoring during the period 2008-2010. These samples were taken in the framework of active surveillance of wildlife for LPAI. The description of the results of the monitoring program is not within the scope of this study. Briefly, the screening methodology was as follows. Captured birds were ringed or identified, their age and sex were determined and cloacal and oropharyngeal swabs were taken from individual birds. Swabs were placed in brain heart infusion broth (BHI) medium containing appropriate antibiotics concentrations (gentamicin 0.5 % [v/v] and glutamine (1% [v/v]). Swab samples were kept at 4°C until laboratory analysis or frozen below -70 °C if they could not be processed immediately. The swabs were screened for AIV by realtime RT-PCR as previously described [31]. Samples positive for influenza A were processed for virus isolation by inoculation of 8-10 day old specific-pathogen-free embryonated chicken eggs using standard procedures (OIE, diagnostic manual 2005/94/CE). The subtypes of hemagglutinating allantoic fluids were determined using reference sera (OIE, diagnostic manual 2005/94/CE) and partial sequencing of the HA and NA genes (amplification primers and sequencing protocol available on request).

Table 1: Origin of the viruses isolated from Charadriiformes in Belgium, 2008-2010

	***A.interpres*/BE/ 17044-27/08 (H3N8)**	***A.interpres*/BE/ 06765cls2/09 (H11N9)**	***A.interpres*/BE/ 02936pcs1/10 (H12N5)**	***L.michahellis*/BE/ 17429-3/08 (H13N6)**	***L.argentatus*/BE/ 17429-1'/08 (H13N6)**	***L.argentatus*/BE/ 14469-20/08 (H13N8)**	***L.argentatus*/BE/ 14469-16/08 (H16N3)**	***L.argentatus*/BE/ 02936pcs3/10 (H6N1)**	***L.argentatus*/BE/ 02936cls9/10 (H5N2)**
Sampling location	Heist	Zeebrugge	Oostende	Mont-Saint-Guibert	Mont-Saint-Guibert	Mont-Saint-Guibert	Mont-Saint-Guibert	Oostende	Oostende
Sampling habitat	coast	coast	coast	landfill	landfill	landfill	landfill	coast	coast
Sampling date	15/09/2008	1/05/2009	27/01/2010	25/09/2008	25/09/2008	18/08/2008	18/08/2008	27/01/2010	27/01/2010

Random access to viral nucleic acids using DNAse I SISPA

Virions in samples were purified starting from 1 ml of allantoic fluid (non-passaged original from virus isolation described above), which was first centrifuged at 3,200 × g for 15 minutes at 4 °C to remove cell debris. The supernatant was then filtered at 3,000 × g for 8 min or longer at 4 °C (200 µl/filter) using 0.22 µm filters (Ultrafree-MC GV sterile, Millipore) to remove remaining cell fragments and bacteria. The resulting eluate was subsequently subjected to nuclease treatment with 100 U of DNase I (New England Biolabs) at 37 °C for 1 hour to remove all nucleic acids not protected within virions. The resulting virion-enriched samples were used for viral RNA extraction using the QIAamp Viral RNA Mini Kit (Qiagen) according to the manufacturer's instructions. Sequence independent single primer amplification (SISPA) was performed essentially as previously described [32]. Briefly, the extracted RNA was converted into single-stranded cDNA using the Transcriptor First Strand cDNA Synthesis Kit (Roche) and 1 µM (final concentration) random primer FR26RV-N (5'GCC GGA GCT CTG CAG ATA TCN NNN NN 3', [33]). Ten µl extracted RNA was denatured at 95 °C for 5 min in the presence of primer FR26RV-N, immediately followed by cooling on ice. The remaining reagents were added. The 20 µl reaction mix contained 1× Transcriptor Reverse Transcriptase Reaction Buffer, 2 µl dNTP mix (10 mM each), 20 U Protector RNase Inhibitor, 10 U Transcriptor Reverse Transcriptase and 1 µl PCR-grade H_2O. The reaction was incubated at 25 °C for 10 min followed by 50 °C for 60 min. After a transcriptase inactivation step at 85 °C for 5 min and chilling on ice, 2.5 U of 3'-5' exo$^-$ Klenow Fragment of DNA polymerase (New England Biolabs) were added for second strand synthesis using random primer FR26RV-N for 1 h at 37 °C. An enzyme inactivation step was performed at 75 °C for 10 min. Five microliters of the reaction mix was used as template for a subsequent PCR amplification. The 50 µl reaction mix consisted of 1× AmpliTaq Gold® 360 DNA buffer, 2.5 mM MgCl2, dNTP mix (0.2 mM each), 2.5 U AmpliTaq Gold® 360 DNA polymerase (Applied Biosystems), 32.7 µl RNase free water and 1.6 µM FR20RV primer (5'-GCC GGA GCT CTG CAG ATA TC-3', [33]) which is complementary to the amplification tag of FR26RV. The reaction was incubated at 95 °C for 10 min, 40 cycles of 95 °C 1 min, 48 °C 1 min and 72 °C 2 min followed by a final elongation for 7 minutes at 72°C. The random amplified DNA fragments were visualised on a 1 % agarose gel. Fragments of 400-1,000 bp were excised and purified from the gel with the High Pure PCR Product Purification Kit (Roche). The purified PCR fragments were quantified by spectrophotometry (Nanodrop-1000).

Sequencing and sequence assembly

Five micrograms of size selected (400-1,000 bp) purified random amplified DNA was sequenced on a GS-FLX (Roche, Mannheim, Germany) by the Genomics Core of the University Hospital, University of Leuven, Belgium, using multiplex identifier (MID) identification during library preparation and their standard procedures using GS FLX Titanium series reagents (Roche, Mannheim, Germany), aiming for 5,000-1,0000 reads per library. The obtained raw sequence data were assembled using SeqMan NGen® version 2 (DNASTAR, Madison, WI, USA). The reads were trimmed to remove primer sequences (including the primer-encoded random N positions) as well as low quality ends. Standard assembling and filtering parameters were used. Contigs (i.e. sets of overlapping sequence reads) produced in an initial *de novo* assembly were compared to public sequence databases for identification (http://blast.ncbi.nlm.nih.gov/Blast.cgi; [34]). The complete coding sequence (CDS) with the highest identity score to the largest contig representing each identified segment was then used as reference genome for a subsequent reference assembly with the same raw data set (closest genbank entries summarized in Figure 3) . The resulting reference assembly was used to deduce a genome consensus sequence. At polymorphic positions (contradictory sequence reads), we included a degenerate nucleotide in the consensus sequence if the minor nucleotide alternative was present in at least one third of the sequence reads. Polymorphic sites were considered as an indication of genetic heterogeneity in the sampled virus population if the positional sequencing depth was at least 20 x and the number of minor variant nucleotide reads was larger than 10 and represented at least 10 % of the positional sequencing depth (variability analysis summarized in Table 3).

Phylogenetic and phylogeographic analysis

The resulting 72 genome segment consensus sequences (GISAID epiFlu accession codes EPI345365-EPI345436; http://platform.gisaid.org) were compared to other publicly available sequences using a BLASTn search (http://blast.ncbi.nlm.nih.gov/Blast.cgi; [34]). For each segment, a selection of 10 complete coding sequences form the first 20 Blastn results, supplemented with selected reference sequences representing Eurasian and American lineages and gull H13/H16 viruses and a selection of viruses isolated from *Charadriiformes* were included in the phylogenetic analysis (genbank accession codes of reference sequences available on request). Selected sequences were aligned using the ClustalW algorithm within BioEdit v7.0.5.3 software [35]. Neighbor-joining phylogenetic trees (pair wise deletion of gaps/missing data and maximum composed likelihood substitution model) were deduced using MEGA v5.01 [36] using 1,000 bootstrap replicates to assess the statistical significance of nodes.

To determine the phylogeographic origin of each segment, the highest order monophyletic bootstrap supported clades (bootstrap support value > 80%) were determined containing each segment. The isolation location of the majority of viruses from these clades determined the phylogeographic origin of these clades as either American (North + South American continents) or Eurasian (Europe + Asia + Oceania + Africa). However, several segments of H13 and H16 that circulate exclusively in *Charadriiformes* birds have a dynamic phylogeographic pattern with near equal likelihood of being isolated in Eurasia or Americas. These clades were labeled *Charadriiformes* specific. It should be noted that phylogeographic grouping of isolates does not necessary signify a close phylogenetic relationship, as (1) considerable genetic diversity may be present within these highest order monophyletic clades, and (2) in principle multiple monophyletic highest order highly bootstrap supported clades of the same phylogeographic origin may arise from a phylogenetic analysis, indicating separate well established introductions in a geographic region.

Results

Next-generation sequencing of random amplified viral RNA

We sequenced partial or complete open reading frames of all 72 individual segments from the 9 virus isolates. Although a variable number of AIV sequence reads was available per sample (ranging from 2,043 to 10,330 reads), we obtained for most viruses > 90 % of the coding sequence of the genome in a single experiment not needing prior sequence information for primer design (Table 2). The largest RNA segments (PB2, PB1, PA, HA, NP) have an excellent coverage well above 90 % of the coding sequence (CDS). Only the PB1 and NP genes from the sample with the smallest number of available sequence reads (*L. argentatus*/BE/17429-1’/2008, H13N6), and the HA gene from the sample *L. argentatus*/Belgium/14469-20/2008 and NP from sample *L.argentatus*/BE/14469-16/08 had <90 % of its CDS covered (Table 2). The average sequencing depth (number of reads available at a given position) for these large segments is as expected from the limited sequencing effort and depended on the number of AIV reads available per sample. The sequence coverage of the smaller segments, especially the Matrix and NS gene segments is considerably lower for some samples. The lowest coverage of 27 % was again documented in the sample with the lowest number of available reads (M1/M2 segment). For regions with sufficient sequencing depth (≥ 20), genetic polymorphisms in the virus isolates were analysed (Table 3). Genetic polymorphisms were detected both in shorebird and gull viruses. The *Charadriiformes* specific viruses from gulls (H13 and H16 strains) contained less polymorphic sites than the other isolated viruses. The H12N5 and H6N1 viruses had the highest number of polymorphic sites, respectively eight and six. The number of polymorphic sites in regions with sufficient sequence depth was independent of the total sequencing effort since the amount of used AIV specific reads for H12N5 (5,851) and H6N1 (3,737) was not the highest among sequenced viruses. The number of polymorphic sites was also independent of the average depth of a specific RNA segment. For example the PB1, NP and NA segments of H11N9 had no polymorphic sites while their average depths were respectively 290, 760 and 302. However, the segment with the highest number of polymorphic sites (HA of the H12N5 virus) had only an average sequence depth of 152. Polymorphic sites were most frequently found in the HA segments. The sequence depth of the small M and NS segments was too low to allow a reliable variability analysis.

Phylogenetic analysis of AIV from *Charadriiformes*

Our phylogenetic analysis (Figure 1, Figure 2, Supplementary Figures S1-S17) indicated the presence of three phylogeographic lineage types for all internal genes: (1) gene segments circulating mostly in wild waterfowl in Eurasia whith sporadic introduction in other species and in the Americas (“Eurasian avian”); (2) gene segments circulating almost exclusively in American wild waterfowl with sporadic introduction to other species and regions (“American avian”) and (3) gene segments circulating exclusively in gulls and shorebirds and having increased connectivity between the hemispheres (“*Charadriiformes* specific”). The H3, H5, H6, H11, and H12 hemagglutining segments of our studied viruses were of Eurasian avian origin (Supplementary Figures S6-S10), while the H13 and H16 were *Charadriiformes* specific genes which are known to make frequent contact between different hemispheres (Figures 2). The neuraminidase segments N1, N2, N5, N8, and N9 showed a clear Eurasian avian origin (Supplementary Figures S11, S12, S14, S16), while the N3 and N6 belonged to *Charadriiformes* specific clades (Supplementary Figures S13, S15).

Table 2: Summary of the next-generation sequencing results of 9 avian influenza genomes.

	A.interpres/BE/17044-27/08 (H3N8)	*A.interpres*/BE/06765cls2/09 (H11N9)	*A.interpres*/BE/02936pcs1/10 (H12N5)	*L.michahellis* /BE/17429-3/08(H13N6)	*L.argentatus*/BE/17429-1’/08(H13N6)	*L.argentatus*/BE/14469-20 /08(H13N8)	*L.argentatus*/BE/14469-16 /08(H16N3)	*L.argentatus*/BE/02936pcs3/10 (H6N1)	*L.argentatus*/BE/02936cls9/10 (H5N2)
AIV reads (Total reads)	2,395 (10,420)	10,330 (11,009)	5,851 (6,388)	6,535 (11,893)	2,043 (5,742)	4,544 (4,816)	4,167 (7,694)	3,737 (4,055)	3,102 (3,310)
Coverage of genome/of CDS (%)	90.4/92.6	95.5/96.9	93.9/95.0	90.2/91.8	77.5/79.3	87.2/89.0	92.0/93.6	96.3/97.6	94.2/96.3
PB2 CDS coverage (%)	98.4	96.0	100	100	90.3	99.5	100	100	100
PB2 seq. depth avg. (max.)	96.7 (417)	110.9 (274)	217.9 (872)	184.3 (1,155)	57.1 (368)	86.95 (394)	162.23 (865)	71.9 (224)	61.7 (165)
PB1 CDS coverage (%)	93.5	99.4	100	98.9	89.3	91.8	97.2	100	99.7
PB1 seq. depth avg. (max.)	52.5 (178)	290 (1,200)	133.1 (574)	316.4 (1,528)	119.2 (550)	44.6 (154)	96.5 (256)	82.2 (145)	66.6 (236)
PA CDS coverage (%)	97.7	99.7	97.5	97.1	91.7	99.9	100	100	99
PA seq. depth avg. (max.)	19.1 (53)	22.5 (82)	59.6 (574)	33.1 (130)	8.3 (27)	48.4 (127)	59.1 (167)	41.4 (138)	38.2 (85)
HA CDS coverage (%)	92.4	99.3	100	97.5	90.1	**67**	91.7	100	98.9
HA seq. depth avg. (max.)	9.0 (21)	113.4 (326)	152.7 (382)	152.2 (433)	42.1 (138)	237.5 (521)	127.8 (422)	44.2 (93)	97.58 (190)
NP CDS coverage (%)	93.4	100	98.1	95.7	81.8	92.6	87.7	100	99.5
NP seq. depth avg. (max.).	43.9 (111)	756.9 (1,700)	107.7 (249)	94.9 (287)	29.5 (80)	44.6 (93)	18.7 (50)	119.2 (223)	86 (228)
NA CDS coverage (%)	96	94.7	100	75.3	**64.9**	95.3	83	90.5	94.5
NA seq. depth avg. (max.)	100.6 (250)	301.6 (713)	90.2 (303)	17 (68)	3 (8)	286.6 (630)	7.7 (34)	110.2 (259)	26.1 (60)
M1/M2 CDS coverage (%)	91.9	97.4	74.9	81.7	**26.9**	**50.7**	81.9	92.2	98.3
M1/M2 seq. depth avg. (max.)	3 (6)	24.3 (56)	6.45 (12)	3.5 (6)	1.5 (2)	3.9 (8)	8.2 (20)	14.8 (28)	5.7 (12)
NS1/NS2 CDS coverage (%)	**55.5**	78.5	**60.4**	**58**	**49.4**	93.2	95.6	86.2	**60.1**
NS1/NS2 seq. depth avg. (max.)	0.9 (1)	6.7 (12)	1.9 (3)	6.3 (28)	1.5 (6)	1.8 (3)	3.35 (6)	8 (16)	3.2 (5)

Genome segments are ranked according to decreasing size. Seqment CDS coverages < 70% are highlighted in bold type.

Figure 1: Evolutionary relationship between the M1/M2 complete coding sequences of nine AIV isolated from *Charadriiformes* and selected reference isolates (inferred using the Neigbor-Joining method). Confidence levels in bootstrap analysis (1,000 replications) above 70% are indicated at nodes. Bootstrap support values used for defining the phylogeographic groupings are circled. Colored boxes indicate the phylogeographic clades (Orange: Eurasian avian, Light blue: American avian, Green: Charadriiformes specific). The green star indicates the node defining an America-to-Eurasia reassortment event. Viruses characterized in this study are indicated in bold type. Viruses isolated from poultry are indicated in red. Asian HPAI H5N1 viruses are indicated in blue.

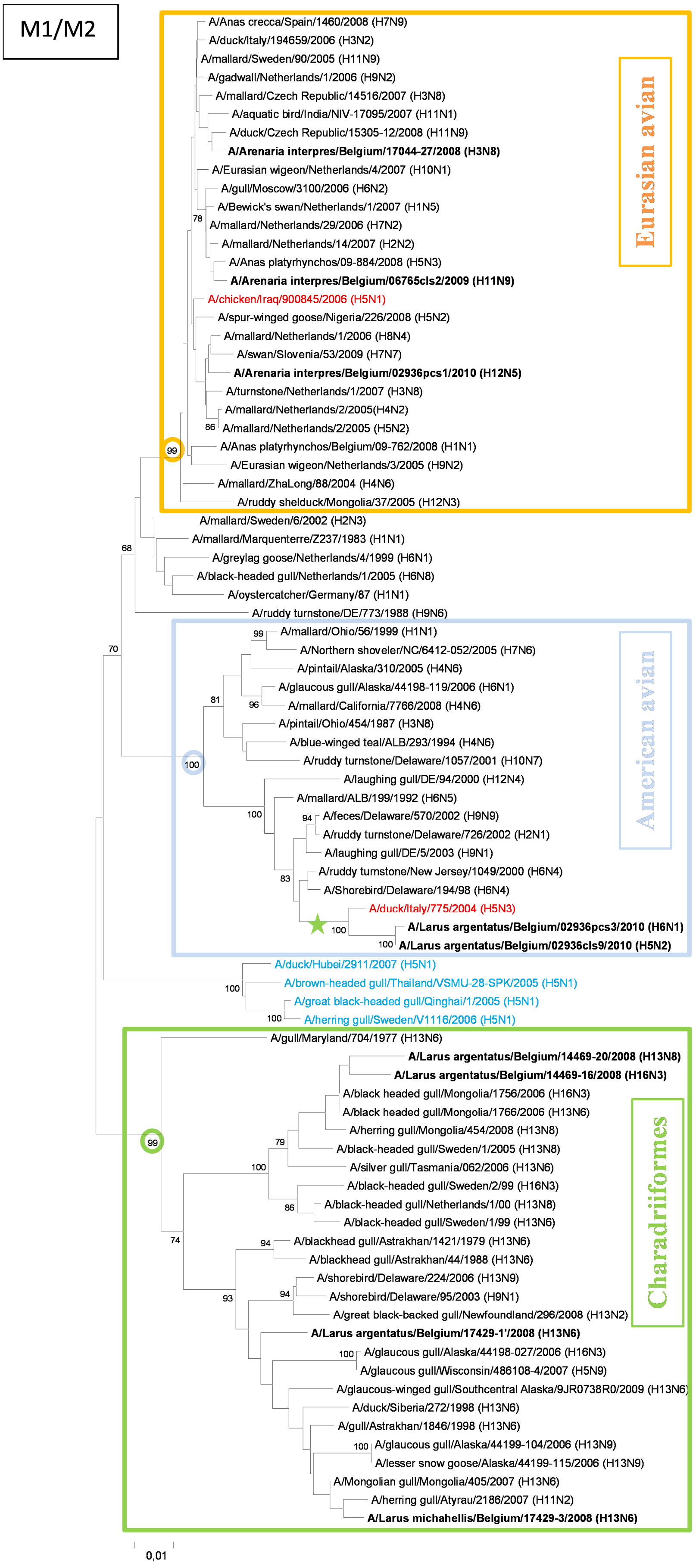

M1/M2
Eurasian avian
A/Anas crecca/Spain/1460/2008 (H7N9)
A/duck/Italy/194659/2006 (H3N2)
A/mallard/Sweden/90/2005 (H11N9)
A/gadwall/Netherlands/1/2006 (H9N2)
A/mallard/Czech Republic/14516/2007 (H3N8)
A/aquatic bird/India/NIV-17095/2007 (H11N1)
A/duck/Czech Republic/15305-12/2008 (H11N9)
A/Arenaria interpres/Belgium/17044-27/2008 (H3N8)
A/Eurasian wigeon/Netherlands/4/2007 (H10N1)
A/gull/Moscow/3100/2006 (H6N2)
A/Bewick's swan/Netherlands/1/2007 (H1N5)
78
A/mallard/Netherlands/29/2006 (H7N2)
A/mallard/Netherlands/14/2007 (H2N2)
A/Anas platyrhynchos/09-884/2008 (H5N3)
A/Arenaria interpres/Belgium/06765cls2/2009 (H11N9)
A/chicken/Iraq/900845/2006 (H5N1)
A/spur-winged goose/Nigeria/226/2008 (H5N2)
A/mallard/Netherlands/1/2006 (H8N4)
A/swan/Slovenia/53/2009 (H7N7)
A/Arenaria interpres/Belgium/02936pcs1/2010 (H12N5)
A/turnstone/Netherlands/1/2007 (H3N8)
A/mallard/Netherlands/2/2005(H4N2)
86
A/mallard/Netherlands/2/2005 (H5N2)
99
A/Anas platyrhynchos/Belgium/09-762/2008 (H1N1)
A/Eurasian wigeon/Netherlands/3/2005 (H9N2)
A/mallard/ZhaLong/88/2004 (H4N6)
A/ruddy shelduck/Mongolia/37/2005 (H12N3)
A/mallard/Sweden/6/2002 (H2N3)
A/mallard/Marquenterre/Z237/1983 (H1N1)
68
A/greylag goose/Netherlands/4/1999 (H6N1)
A/black-headed gull/Netherlands/1/2005 (H6N8)
A/oystercatcher/Germany/87 (H1N1)
A/ruddy turnstone/DE/773/1988 (H9N6)
70
American avian
99
A/mallard/Ohio/56/1999 (H1N1)
A/Northern shoveler/NC/6412-052/2005 (H7N6)
A/pintail/Alaska/310/2005 (H4N6)
A/glaucous gull/Alaska/44198-119/2006 (H6N1)
81
96
A/mallard/California/7766/2008 (H4N6)
A/pintail/Ohio/454/1987 (H3N8)
A/blue-winged teal/ALB/293/1994 (H4N6)
100
A/ruddy turnstone/Delaware/1057/2001 (H10N7)
A/laughing gull/DE/94/2000 (H12N4)
A/mallard/ALB/199/1992 (H6N5)
100
94
A/feces/Delaware/570/2002 (H9N9)
A/ruddy turnstone/Delaware/726/2002 (H2N1)
A/laughing gull/DE/5/2003 (H9N1)
83
A/ruddy turnstone/New Jersey/1049/2000 (H6N4)
A/Shorebird/Delaware/194/98 (H6N4)
A/duck/Italy/775/2004 (H5N3)
100
A/Larus argentatus/Belgium/02936pcs3/2010 (H6N1)
100
A/Larus argentatus/Belgium/02936cls9/2010 (H5N2)
A/duck/Hubei/2911/2007 (H5N1)
100
A/brown-headed gull/Thailand/VSMU-28-SPK/2005 (H5N1)
A/great black-headed gull/Qinghai/1/2005 (H5N1)
100
A/herring gull/Sweden/V1116/2006 (H5N1)
Charadriiformes
A/gull/Maryland/704/1977 (H13N6)
A/Larus argentatus/Belgium/14469-20/2008 (H13N8)
A/Larus argentatus/Belgium/14469-16/2008 (H16N3)
A/black headed gull/Mongolia/1756/2006 (H16N3)
A/black headed gull/Mongolia/1766/2006 (H13N6)
A/herring gull/Mongolia/454/2008 (H13N8)
99
79
A/black-headed gull/Sweden/1/2005 (H13N8)
100
A/silver gull/Tasmania/062/2006 (H13N6)
A/black-headed gull/Sweden/2/99 (H16N3)
86
A/black-headed gull/Netherlands/1/00 (H13N8)
A/black-headed gull/Sweden/1/99 (H13N6)
74
94
A/blackhead gull/Astrakhan/1421/1979 (H13N6)
A/blackhead gull/Astrakhan/44/1988 (H13N6)
A/shorebird/Delaware/224/2006 (H13N9)
93
94
A/shorebird/Delaware/95/2003 (H9N1)
A/great black-backed gull/Newfoundland/296/2008 (H13N2)
A/Larus argentatus/Belgium/17429-1'/2008 (H13N6)
100
A/glaucous gull/Alaska/44198-027/2006 (H16N3)
A/glaucous gull/Wisconsin/486108-4/2007 (H5N9)
A/glaucous-winged gull/Southcentral Alaska/9JR0738R0/2009 (H13N6)
A/duck/Siberia/272/1998 (H13N6)
A/gull/Astrakhan/1846/1998 (H13N6)
100
A/glaucous gull/Alaska/44199-104/2006 (H13N9)
A/lesser snow goose/Alaska/44199-115/2006 (H13N9)
A/Mongolian gull/Mongolia/405/2007 (H13N6)
A/herring gull/Atyrau/2186/2007 (H11N2)
A/Larus michahellis/Belgium/17429-3/2008 (H13N6)
0,01

Figure 2: Evolutionary relationship between the hemagglutinin H13 and H16 complete coding sequences of nine AIV isolated from *Charadriiformes* and selected reference isolates (inferred using the Neigbor-Joining method). Confidence levels in bootstrap analysis (1,000 replications) above 70 % are indicated at nodes. Viruses of American geographic isolation location are indicated in green to demonstrate the frequent transhemispheric movements of these genes.

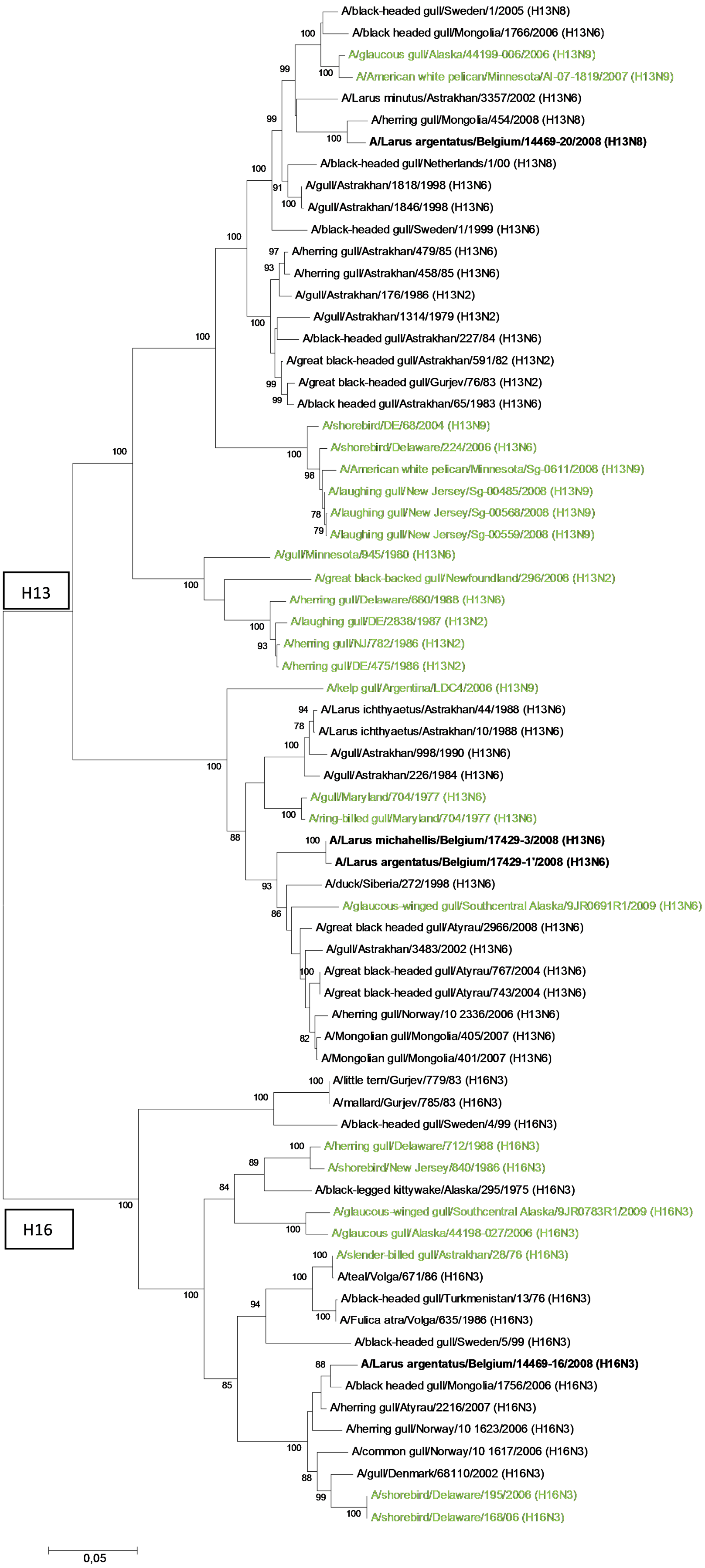

A/black-headed gull/Sweden/1/2005 (H13N8)
A/black headed gull/Mongolia/1766/2006 (H13N6)
A/glaucous gull/Alaska/44199-006/2006 (H13N9)
A/American white pelican/Minnesota/AI-07-1819/2007 (H13N9)
A/Larus minutus/Astrakhan/3357/2002 (H13N6)
A/herring gull/Mongolia/454/2008 (H13N8)
A/Larus argentatus/Belgium/14469-20/2008 (H13N8)
A/black-headed gull/Netherlands/1/00 (H13N8)
A/gull/Astrakhan/1818/1998 (H13N6)
A/gull/Astrakhan/1846/1998 (H13N6)
A/black-headed gull/Sweden/1/1999 (H13N6)
A/herring gull/Astrakhan/479/85 (H13N6)
A/herring gull/Astrakhan/458/85 (H13N6)
A/gull/Astrakhan/176/1986 (H13N2)
A/gull/Astrakhan/1314/1979 (H13N2)
A/black-headed gull/Astrakhan/227/84 (H13N6)
A/great black-headed gull/Astrakhan/591/82 (H13N2)
A/great black-headed gull/Gurjev/76/83 (H13N2)
A/black headed gull/Astrakhan/65/1983 (H13N6)
A/shorebird/DE/68/2004 (H13N9)
A/shorebird/Delaware/224/2006 (H13N6)
A/American white pelican/Minnesota/Sg-0611/2008 (H13N9)
A/laughing gull/New Jersey/Sg-00485/2008 (H13N9)
A/laughing gull/New Jersey/Sg-00568/2008 (H13N9)
A/laughing gull/New Jersey/Sg-00559/2008 (H13N9)
A/gull/Minnesota/945/1980 (H13N6)
A/great black-backed gull/Newfoundland/296/2008 (H13N2)
A/herring gull/Delaware/660/1988 (H13N6)
A/laughing gull/DE/2838/1987 (H13N2)
A/herring gull/NJ/782/1986 (H13N2)
A/herring gull/DE/475/1986 (H13N2)
H13
A/kelp gull/Argentina/LDC4/2006 (H13N9)
A/Larus ichthyaetus/Astrakhan/44/1988 (H13N6)
A/Larus ichthyaetus/Astrakhan/10/1988 (H13N6)
A/gull/Astrakhan/998/1990 (H13N6)
A/gull/Astrakhan/226/1984 (H13N6)
A/gull/Maryland/704/1977 (H13N6)
A/ring-billed gull/Maryland/704/1977 (H13N6)
A/Larus michahellis/Belgium/17429-3/2008 (H13N6)
A/Larus argentatus/Belgium/17429-1'/2008 (H13N6)
A/duck/Siberia/272/1998 (H13N6)
A/glaucous-winged gull/Southcentral Alaska/9JR0691R1/2009 (H13N6)
A/great black headed gull/Atyrau/2966/2008 (H13N6)
A/gull/Astrakhan/3483/2002 (H13N6)
A/great black-headed gull/Atyrau/767/2004 (H13N6)
A/great black-headed gull/Atyrau/743/2004 (H13N6)
A/herring gull/Norway/10 2336/2006 (H13N6)
A/Mongolian gull/Mongolia/405/2007 (H13N6)
A/Mongolian gull/Mongolia/401/2007 (H13N6)
A/little tern/Gurjev/779/83 (H16N3)
A/mallard/Gurjev/785/83 (H16N3)
A/black-headed gull/Sweden/4/99 (H16N3)
A/herring gull/Delaware/712/1988 (H16N3)
A/shorebird/New Jersey/840/1986 (H16N3)
A/black-legged kittywake/Alaska/295/1975 (H16N3)
H16
A/glaucous-winged gull/Southcentral Alaska/9JR0783R1/2009 (H16N3)
A/glaucous gull/Alaska/44198-027/2006 (H16N3)
A/slender-billed gull/Astrakhan/28/76 (H16N3)
A/teal/Volga/671/86 (H16N3)
A/black-headed gull/Turkmenistan/13/76 (H16N3)
A/Fulica atra/Volga/635/1986 (H16N3)
A/black-headed gull/Sweden/5/99 (H16N3)
A/Larus argentatus/Belgium/14469-16/2008 (H16N3)
A/black headed gull/Mongolia/1756/2006 (H16N3)
A/herring gull/Atyrau/2216/2007 (H16N3)
A/herring gull/Norway/10 1623/2006 (H16N3)
A/common gull/Norway/10 1617/2006 (H16N3)
A/gull/Denmark/68110/2002 (H16N3)
A/shorebird/Delaware/195/2006 (H16N3)
A/shorebird/Delaware/168/06 (H16N3)
0,05

Table 3: Polymorphisms detected in the genomes of the sequenced virus isolates.

Virus strain	Segment	Position	Read distribution (major/minor)	% of minor variant	Consensus at position
A.interpres/BE/06765cls2/2009	PB2	113	14A/10G	41.6 %	R
(H11N9)	HA	1,224	278G/37T	11.7 %	G
	HA	1,239	236T/73G	23.6 %	T
A.interpres/BE/02936pcs1/2010	PB2	657	106C/21T	16.5 %	C
(H12N5)	PB1	1,852	36C/18A	33.3 %	M
	HA	127	294A/80T	21.4 %	A
	HA	195	331T/40C	10.8 %	T
	HA	1,270	157A/30G	16 %	A
	HA	1,313	165A/27G	14.1 %	A
	HA	1,443	112T/13A	10.4 %	T
	NP	1,108	156C/34A	17.9 %	C
L. michahellis/BE/17429-3/2008 (H13N6)	PA	239	59G/20A	25.3 %	G
L. argentatus/BE/14469-16/2008 (H16N3)	HA	545	137A/73G	34.8 %	R
L. argentatus/BE/02936pcs3/2010	PB1	729	63T/61A	49.2 %	W
(H6N1)	PB1	782	60C/53T	46.9 %	Y
	PA	183	58C/23A	28.4 %	C
	HA	834	26A/14C	35 %	M
	HA	846	22G/11A	33.3 %	R
	NA	733	167G/57A	25.4 %	G
L. argentatus/BE/02936cls9/2010	HA	540	74A/46C	38.3 %	M
(H5N2)	HA	1,023	66C/44T	40 %	Y

The polymorphism analysis includes only genomic sites with positional sequencing depth ≥ 20-fold, percentage of minor variant ≥ 10 %, and the number of minor variant nucleotide reads ≥ 10. Degenerate nucleotide codes were included at the position when presence of minor variant was ≥ one third of the positional depth.

Phylogeographic analysis of AIV from *Charadriiformes*

Figure 3 summarizes the phylogeographic origin of all RNA segments of all viruses. All gene segments of the three Ruddy Turnstone viruses (H3N8, H11N8, H12N5) sampled over 3 consecutive years on the North Sea coast had a typical Eurasian avian origin (Figure 3) and were most closely related to virus segment sequences found in *Anseriformes* from western and central Europe as indicated by the closest BLASTn results for each segment (Figure 3). Although having a similar geographic origin, these viruses had phylogenetically distinguishable internal genes (Figure 1; Supplementary Figures 1-5). The H3N8 virus was the only virus in this study having an allele B NS segment (Supplementary Figure S5). The viruses isolated from gulls fell into two groups: typical *Charadriiformes* H13 and H16 viruses on the one hand and Eurasian avian viruses (H5N2 and H6N1) on the other hand. The two

H13N6 viruses isolated from *L. michahellis* and *L. argentatus* on a single sampling location and time (Mont-Saint-Guibert landfill, September 25, 2008) in 2008 were highly similar for all segments (Figures 1-3, Supplementary Figures S1-S5, S15). Apart from the well-known H13 gull specific hemagglutinin subtype, their PB2, PB1, NP, NA6, M1/M2 and NS1/NS2 (allele A) segments clustered in *Charadriiformes*-specific lineages (Figure 3). These *Charadriiformes* specific clades included viruses isolated in both Eurasian and American hemispheres. Both viruses were reassorted: they shared a PA segment of Eurasian avian origin most closely related to A/Mongolian gull/Mongolia/ 405/2007 (H13N6) (Figure 3; Supplementary Figure S3). Two viruses isolated on the same landfill location from *Larus argentatus*, one H16N3 virus and one H13N8 virus also shared similar segments (PB2, PB1, PA, NP, M, NS; Figure 1; Figure 2, Supplementary Figures S1-S5). The NP gene of the H13N8 virus was closely related to the NP genes of the two previous mentioned H13N6 viruses which were also isolated on the same location, but one month later (Supplementary Figure S4). All segments of these viruses clustered in supported *Charadriiformes* specific lineages, including the neuraminidase N3 and the neuraminidase N8, although the N8 belonged to a lower order *Charadriiformes* specific clade within a large Eurasian avian origin clade (Figure 3; Supplementary Figures S13, S16). The latter belonged to a bootstrap supported clade containing only NA sequences from *Charadriiformes* sampled in Eurasia. The two viruses (H5N2 and H6N1) isolated from *Larus argentatus* on the North Sea shore in Oostende on the same day in 2010, were of reassorted origin (Figure 3). They had a majority of gene segments of Eurasian avian origin (Figure 3). They shared a near identical PB2 and M1/M2 gene (Figure 1; Supplementary Figure S1), but have distinguishable other genes (Supplementary Figures S2-S5, S7-8, S11-12). The NP gene of the H6N1 isolated from gulls was similar to the NP gene of the H11N9 virus isolated a year earlier in the same coastal region from a Ruddy Turnstone (Supplementary Figure S4). The PA and NS genes of H5N2 isolated from a gull were similar to the respective RNA segments from the H12N5 virus isolated from a Ruddy Turnstone on same day and place (Supplementary Figures S3, S5). Most interestingly, the shared Matrix segment of the H5N2 and H6N1 gull viruses was of American avian origin (Figure 1; Figure 3) and was closely related to a previously documented transhemispheric reassortant Matrix segment from domestic ducks in Italy.

	A. interpres/BE/17044-27/ 2008 (H3N8)	*A. interpres*/BE/06765cls2/ 2009 (H11N9)	*A. interpres*/BE/02936pcs1/ 2010 (H12N5)
PB2	mallard/Italy/3401/2005 (H5N1) CY095584	teal/Chany/444/2009 (H8N8) CY098521	mallard/Czech Republic/ 14516/2007 (H3N8) JF682611
PB1	mallard/Sweden/48/2002 (H11N9) CY060292	mallard/Netherlands/14/2007 (H2N2) CY041240	mute swan/Hungary/5973/ 2007(H7N7) GQ240811
PA	black-headed gull/ NL/1/2005(H6N8) CY041383	tufted duck/PT/13771/ 2006(H7N3) HM849008	duck/Beijing/40/04(H3N8) EU492500
HA	mallard/Netherlands/5/2001 (H3N6) CY060343	mallard/Netherlands/17/2007 (H11N8) CY043880	teal/Norway/10_1836/2006 (H12N2) FM179754
NP	mallard/Czech Republic/ 14924-1/2007(H6N5)JF789627	greylag goose/Netherlands/4/ 1999 (H6N1) CY060199	mallard/Netherlands/28/2006 (H3N1) CY076908
NA	mallard/Netherlands/17/2007 (H11N8) CY043882	goose/Czech Republic/1848-T14/2009 (H7N9) HQ244417	mallard/Switzerland/WV406016 7/2006 (H3N5) GQ415323
M1/M2	mallard/Sweden/90/2005 (H11N9) CY076961	mallard/Netherlands/29/2006 (H7N2) CY043833	mallard/Czech Republic/15307-17/2008 GQ404573
NS1/NS2	mallard/Czech Republic/15902-17K/2009 (H6N2) HQ244434 **Allele B**	mallard/Netherlands/14/2007 (H2N2) CY041238 **Allele A**	mallard/Netherlands/7/2007 (H4N2) CY076925 **Allele A**

	L. michahellis/BE/17429-3/ 2008 (H13N6)	*L. argentatus*/BE/17429-1'/ 2008 (H13N6)	*L. argentatus*/BE/14469-20/ 2008 (H13N8)
PB2	Mongolian gull/Mongolia/ 401/2007 (H13N6) GQ907317	black-headed gull/Sweden/ 1/2005 (H13N8) CY077007	herring gull/Mongolia/454/2008 (H13N8) JF775477
PB1	black-headed gull/Sweden/ 1/2005 (H13N8) CY077006	black-headed gull/Sweden/ 1/2005 (H13N8) CY077006	black headed gull/Mongolia/ 1756/2006 (H16N3) GQ907300
PA	Mongolian gull/Mongolia/ 405/2007(H13N6) GQ907323	Mongolian gull/Mongolia/ 405/2007(H13N6) GQ907323	black headed gull/Mongolia/ 1756/2006 (H16N3) GQ907299
HA	gull/Astrakhan/3483/2002 (H13N6) EU835897	gull/Astrakhan/3483/2002 (H13N6) EU835897	herring gull/Mongolia/454/2008 (H13N8) JF775470
NP	great black-headed gull/Atyrau/ 743/2004 (H13N6) GU982289	great black-headed gull/Atyrau/ 743/2004 (H13N6) GU982289	great black-headed gull/Atyrau/ 743/2004 (H13N6) GU982289
NA	great black-headed gull/Atyrau/ 773/2004 (H13N6) GU982288	great black-headed gull/Atyrau/ 773/2004 (H13N6) GU982288	black-headed gull/Sweden/1/ 2005 (H13N8) CY077002
M1/M2	Mongolian gull/Mongolia/405/ 2007 (H13N6) GQ907319	Mongolian gull/Mongolia/405/ 2007 (H13N6) GQ907319	black-headed gull/Sweden/1/ 2005 (H13N8) CY077001
NS1/NS2	Mongolian gull/Mongolia/405/ 2007 (H13N6) GQ907322 **Allele A**	Mongolian gull/Mongolia/405/ 2007 (H13N6) GQ907322 **Allele A**	black headed gull/Mongolia/ 1756/2006 (H16N3) GQ907298 **Allele A**

	L. argentatus/BE/14469-16/ 2008 (H16N3)	*L. argentatus*/BE/02936pcs3/ 2010 (H6N1)	*L. argentatus*/BE/02936cls9/ 2010 (H5N2)
PB2	herring gull/Mongolia/454/2008 (H13N8) JF775477	mallard/Netherlands/20/2005 (H12N8) CY076975	mallard/Netherlands/20/2005 (H12N8) CY076975
PB1	black headed gull/Mongolia/ 1756/2006 (H16N3) GQ907300	duck/Primorie/2633/2001 (H5N3) GQ227611	common eider/Netherlands/ 1/200 (H3N8) CY041344
PA	black-headed gull/Sweden/ 1/99 (H13N6) AY684883	duck/Shiga/8/2004 (H4N6) AB304146	mallard/Netherlands/17/2007 (H11N8) CY043885
HA	black headed gull/Mongolia/ 1756/2006 (H16N3) GQ907294	mallard/Czech Republic/15902-17K/2009 (H6N2) HQ244430	mallard/Sweden/74/2003 (H5N2) CY076929
NP	great black-headed gull/Atyrau/ 773/2004 (H13N6) GU982292	duck/Italy/775/2004 (H5N3) CY024749	mallard/Sweden/48/2002 (H11N9) CY060295
NA	black headed gull/Mongolia/ 1756/2006 (H16N3) GQ907296	aquatic bird/India/NIV-17095/ 2007 (H11N1) CY055177	gull/Moscow/3100/2006 (H6N2) EU152239
M1/M2	black headed gull/Mongolia/ 1766/2006 (H13N6) GQ907303	duck/Italy/775/2004 (H5N3) CY024747 ruddy turnstone/New Jersey/ 1049/2000 (H6N4) GU051408	duck/Italy/775/2004 (H5N3) CY024747 ruddy turnstone/New Jersey/ 1049/2000 (H6N4) GU051408
NS1/NS2	black headed gull/Mongolia/ 1756/2006 (H16N3) GQ907298 **Allele A**	common gull/Ust-Ilimsk/121/ 2008 CY080586 **Allele A**	mallard/Czech Republic/13438-29K/2010 (H11N9) JF789606 **Allele A**

Figure 3 (previous page): Geographical mosaicism in AIV isolated from *Charadriiformes* (gulls and shorebirds) isolated in Belgium, 2008-2010. The phylogeographic origin of each segment was deduced in a detailed phylogenetic analysis. A color code is used for the geographical origin of the corresponding gene segments (orange: Eurasian avian, light blue: American avian, green: *Charadriiformes* specific, grey: Eurasian *Charadriiformes* specific subclade). For each segment, the accession code and name of the most identical complete coding sequence in a BLASTn search is indicated (Genbank accessed 14 October 2011). Grey table gridlines: isolates from Ruddy Turnstone (*Arenaria interpres*); black table gridlines: gull isolates (*Larus* sp.).

Discussion

Although less frequently infected than waterfowl [18], gulls and shorebirds have an important role in AIV ecology in wild birds. Their migratory biology has been shown to facilitate sporadic contacts between the mainly separated American and Eurasian gene pools of avian influenza [16, 17]. In addition to infections with typical avian hemisphere-specific AIV, *Charadriiformes*, notably gulls, perpetuate lineages of AIV that seem to circulate exclusively in *Charadriiformes*. We selected nine AIV isolated from gulls and shorebirds for whole genome sequencing. Due to the diversity of strains and subtypes, a generic sequencing protocol was needed. As part of a feasibility study testing the applicability scope of random amplification, we opted for random access amplification in combination with next-generation sequencing (NGS).

The protocol resulted in an efficient enrichment of the RNA library in viral RNA. Although we intentionally only used a moderate sequencing effort (< 10,000 sequencing reads) to test the feasibility of sequencing a high number of viruses, the methodology proved to target all segments with a typical whole genome coverage of 90-95 %. This allowed a phylogeographic analysis covering all gene segments of all nine viruses. The methodology seems to be less efficient for smaller segments M and NS where the % of the coding sequence (CDS) covered and the average sequencing depth were considerably lower. One sample only yielded 2,043 AIV specific sequence reads, resulting in <65 % coverage of the CDS of the three smallest RNA segments. Regions where sequence information was missing were mostly situated at segment extremities and sequence depth at extremities was generally speaking lower (data not shown). It is obvious that to achieve complete genomes including the 5’ and 3’ noncoding regions, more AIV specific reads would be needed. Alternatively, additional Sanger sequencing may be used to complete the genome sequences. Since completing this study, other methods based on AIV-specific PCR amplification [37, 38] have become available that may be more suitable for targeted influenza A virus sequencing. AIV specific library preparation tools may be especially useful for studies aiming at genome wide polymorphism analysis and quasispecies characterization. Our data show the value of NGS technology to produce large volumes of sequence data of segmented RNA genomes using a universal amplification and sequencing approach in a single experiment.

An important added value of NGS methods is the documentation of genetic variants and multiple infections [32, 39]. Multiple infections are a prerequisite for AIV dynamic evolution through reassortment [1], and have been previously demonstrated [31, 39]. Although the sequencing effort in this study was not sufficient to allow a genome wide quantitative analysis of genetic polymorphisms, our data show the power of NGS technology for the characterization of genetic variability within samples. This contributes to an increased reliability of the consensus sequence as uploaded to public databases compared to Sanger sequencing-based consensus sequences which are mostly based on a sequence depth below 5. Degenerate nucleotide codes in consensus sequences now become an indication of an actual polymorphism in the sample, instead of a doubtful base calling as a result of a limited number of contradictory sequence reads. Although a prerequisite for using NGS for large numbers of viruses is the diminished cost of NGS (can be facilitated through multiplexed sequencing, evolution of sequencing platform and service market, enhanced sample preparation methods), these data clearly show the power of complete genome sequencing using next-generation sequencing to provide improved insights in the phylogenomic/geographic ecological dynamics of avian influenza in wild birds.

Our phylogeographic analysis confirms the circulation of diverse viruses in gulls and shorebirds, with strict segregation of gene segments in three phylogeographic lineage types: (1) gene segments circulating mostly in waterfowl in Eurasia with sporadic introduction in other species and in the Americas (“Eurasian avian”); (2) gene segments circulating almost exclusively in American waterfowl with sporadic introduction to other species and regions (“American avian”) and (3) gene segments circulating exclusively in gulls and shorebirds and having increased connectivity between the two hemispheres (“*Charadriiformes* specific”). The single gene segment for which our phylogenetic analysis did not result in a strict geographical separation, the hemagglutinin H6, can be attributed to the historical replacement of American H6 sequences by a sustainably introduced Eurasian-origin H6 lineage. In fact, all currently circulating H6 genes in the Americas are of Eurasian origin [27]. In addition, interaction between these AIV lineages was demonstrated as shown by chimeric *Charadriiformes* specific viruses having an Avian Eurasian PA segment and also by chimeric Eurasian avian H5N2 and H6N1 viruses having an American avian Matrix gene. The M1/M2 segments of the intercontinental reassortant viruses we documented in gulls are closely related to American origin Matrix gene sequences from Italian domestic ducks in 2004, where also an American origin H11 gene was reported [26]. This introduction of American segments in Eurasia seems to have resulted in sustained circulation in Eurasian avifauna and sporadic occurrence in domesticated ducks. Gulls may have had important role in the introduction event. Importantly, the other segments of these chimeric gull viruses were of diverse avian Eurasian origin, indicating a dynamic evolutionary ecology of these segments in their natural reservoir. These Eurasian avian chimeric H5 and H6 viruses were isolated during the winter at a coastal location. On the contrary, the typical *Charadriiformes* specific H13 and H16 viruses we isolated from gulls were isolated at an inland landfill site during late summer. In Belgium, a year-round population of herring gulls is joined during the winter by the population of *L.argentatus* breeding in the arctic regions and moving south for the winter. Based on only a small sample size, it remains unclear whether this difference in gull population or the sampling site may have contributed to the detection of completely different viruses. Intercontinental reassortment events involving typically avian American and Eurasian clades have been mainly documented in North America. The migration ecology of the main reservoir species (*Anseriformes*) of these clades and the general phylogeographic patterns indicate a stringent separation between American and Eurasian avian clades. This has led to the suggestion that other species (*Charadriiformes*) may act as “bridge species” to facilitate contact between geographical regions. Indeed, both Eurasian avian and American avian lineages seem to be able to sporadically infect *Charadriiformes* species. The migratory ecology of these species allows overlap of Eurasian and American population in arctic breeding grounds. Moreover these birds come together in breeding grounds or migration stopover sites with numerous other migratory birds including migratory *Anseriformes*, allowing exchange of viruses between these bird species. In North America, this has led to increasing evidence of hemispheric chimeric viruses in *Charadriiformes* species. On the contrary, documentations of American genes in Eurasia remain limited in spite of a representative sampling of these species in Eurasia [18]. Notably, this is the first documentation of such a transatlantic reassortment in a Eurasian gull. The three AIV genomes from Ruddy Turnstones characterized in this study showed a strictly Eurasian gene constellation. In contrast to the well-studied shorebird population in North America, very limited AIV sequence data exists from shorebirds in Eurasia. The fact that no geographic reassortments were documented in these species in Eurasia may be due to a difference in sampling size [18]. It remains unexplained why the patterns of intercontinental reassortment differ between the continents. It has been suggested that AIV ecology is different in both continents [17]. Our data indicate that these chimeric viruses do occur in Western Europe, although their frequency must be low considering that sampling of *Charadriiformes* in Europe is representative [17, 18].

In conclusion, we show that next-generation sequencing technology allows efficient phylogenetic and phylogeographic analyses targeting all influenza viral segments. This may facilitate new insights in avian influenza in wild birds in the near future. Our data confirm the dynamic nature of LPAI in wild birds. Moreover, we provide the first evidence of a transhemispheric reassortant of AIV in gulls in Europe, reinforcing available evidence of a key role of *Charadriiformes* birds in the dynamic nature of avian influenza ecology. These birds should be considered as important target species for surveillance programs.

Supplementary material

Supplementary figures are available in the online version of this publication:
http://link.springer.com/article/10.1007%2Fs00705-012-1323-x

Acknowledgements

The authors would like to thank all technicians of the Veterinary and Agrochemical Research Center that contributed to this work. This work was supported by funding from the EU Network of Excellence, EPIZONE (FOODCT-2006-016236) and EU Project INN-FLU (SSPE-CT-2006-44372).

References

1. Webster, R.G., et al., *Evolution and ecology of influenza A viruses.* Microbiol Rev, 1992. **56**(1): p. 152-79.
2. Kawaoka, Y., et al., *Orthomyxoviridae*, in *Virus Taxonomy: Eighth Report of the International Committee for the Taxonomy of Viruses.*, C.M. Fauquet, et al., Editors. 2005, Elsevier Academic Press: San Diego, U.S.A. p. 681-693.
3. Alexander, D.J., *Newcastle disease and other avian Paramyxoviruses.* Rev Sci Tech Off int Epiz, 2000. **19**(2): p. 443-462.

4. Fouchier, R.A., et al., *Characterization of a novel influenza A virus hemagglutinin subtype (H16) obtained from black-headed gulls.* J Virol, 2005. **79**(5): p. 2814-22.
5. Rambaut, A., et al., *The genomic and epidemiological dynamics of human influenza A virus.* Nature, 2008. **453**(7195): p. 615-9.
6. Olsen, B., et al., *Global patterns of influenza a virus in wild birds.* Science, 2006. **312**(5772): p. 384-8.
7. Scholtissek, C., et al., *On the origin of the human influenza virus subtypes H2N2 and H3N2.* Virology, 1978. **87**(1): p. 13-20.
8. Kawaoka, Y., S. Krauss, and R.G. Webster, *Avian-to-human transmission of the PB1 gene of influenza A viruses in the 1957 and 1968 pandemics.* J Virol, 1989. **63**(11): p. 4603-8.
9. Li, K.S., et al., *Genesis of a highly pathogenic and potentially pandemic H5N1 influenza virus in eastern Asia.* Nature, 2004. **430**(6996): p. 209-13.
10. Dugan, V.G., et al., *The evolutionary genetics and emergence of avian influenza viruses in wild birds.* PLoS Pathog, 2008. **4**(5): p. e1000076.
11. Krauss, S., et al., *Influenza in Migratory Birds and Evidence of Limited Intercontinental Virus Exchange.* PLoS Pathog, 2007. **3**(11): p. e167.
12. Wallensten, A., et al., *Multiple gene segment reassortment between Eurasian and American lineages of influenza A virus (H6N2) in Guillemot (Uria aalge).* Arch Virol, 2005. **150**(8): p. 1685-92.
13. Koehler, A.V., et al., *Genetic evidence of intercontinental movement of avian influenza in a migratory bird: the northern pintail (Anas acuta).* Mol Ecol, 2008. **17**(21): p. 4754-62.
14. Manzoor, R., et al., *Phylogenic analysis of the M genes of influenza viruses isolated from free-flying water birds from their Northern Territory to Hokkaido, Japan.* Virus Genes, 2008. **37**(2): p. 144-52.
15. Ramey, A.M., et al., *Intercontinental reassortment and genomic variation of low pathogenic avian influenza viruses isolated from northern pintails (Anas acuta) in Alaska: examining the evidence through space and time.* Virology, 2010. **401**(2): p. 179-89.
16. Pearce, J.M., et al., *Limited evidence of trans-hemispheric movement of avian influenza viruses among contemporary North American shorebird isolates.* Virus Res, 2010. **148**(1-2): p. 44-50.
17. Wille, M., et al., *Extensive geographic mosaicism in avian influenza viruses from gulls in the northern hemisphere.* PLoS ONE, 2011. **6**(6): p. e20664.
18. Munster, V.J., et al., *Spatial, temporal, and species variation in prevalence of influenza A viruses in wild migratory birds.* PLoS Pathog, 2007. **3**(5): p. e61.
19. Wille, M., et al., *Reassortment of American and Eurasian genes in an influenza A virus isolated from a great black-backed gull (Larus marinus), a species demonstrated to move between these regions.* Arch Virol, 2011. **156**(1): p. 107-15.
20. Ramey, A.M., et al., *Transmission and reassortment of avian influenza viruses at the Asian-North American interface.* Virology, 2010. **406**(2): p. 352-9.
21. Wahlgren, J., et al., *Gene segment reassortment between American and Asian lineages of avian influenza virus from waterfowl in the Beringia area.* Vector Borne Zoonotic Dis, 2008. **8**(6): p. 783-90.
22. Garamszegi, L.Z. and A.P. Moller, *Prevalence of avian influenza and host ecology.* Proc Biol Sci, 2007. **274**(1621): p. 2003-12.
23. Kishida, N., et al., *H2N5 influenza virus isolates from terns in Australia: genetic reassortants between those of the Eurasian and American lineages.* Virus Genes, 2008. **37**(1): p. 16-21.

24. Liu, J.H., et al., *Interregional transmission of the internal protein genes of H2 influenza virus in migratory ducks from North America to Eurasia.* Virus Genes, 2004. **29**(1): p. 81-6.
25. Pawar, S., et al., *An avian influenza A(H11N1) virus from a wild aquatic bird revealing a unique Eurasian-American genetic reassortment.* Virus Genes, 2010. **41**(1): p. 14-22.
26. Fusaro, A., et al., *Gene segment reassortment between Eurasian and American clades of avian influenza virus in Italy.* Arch Virol, 2010. **155**(1): p. 77-81.
27. zu Dohna, H., et al., *Invasions by Eurasian avian influenza virus H6 genes and replacement of the virus' North American clade.* Emerg Infect Dis, 2009. **15**(7): p. 1040-5.
28. Obenauer, J.C., et al., *Large-scale sequence analysis of avian influenza isolates.* Science, 2006. **311**(5767): p. 1576-80.
29. Widjaja, L., et al., *Matrix gene of influenza a viruses isolated from wild aquatic birds: ecology and emergence of influenza a viruses.* J Virol, 2004. **78**(16): p. 8771-9.
30. Kawaoka, Y., et al., *Is the gene pool of influenza viruses in shorebirds and gulls different from that in wild ducks?* Virology, 1988. **163**(1): p. 247-50.
31. Van Borm, S., et al., *Genetic characterization of low pathogenic H5N1 and cocirculating avian influenza viruses in wild mallards (Anas platyrhynchos) in Belgium, 2008.* Avian Pathology, 2011. **available online: 22 Sep 2011**.
32. Rosseel, T., et al., *Identification and complete genome sequencing of paramyxoviruses in mallard ducks (Anas platyrhynchos) using random access amplification and next generation sequencing technologies.* Virol J, 2011. **8**: p. 463.
33. Allander, T., et al., *Cloning of a human parvovirus by molecular screening of respiratory tract samples.* Proc Natl Acad Sci U S A, 2005. **102**(36): p. 12891-6.
34. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
35. Hall, T.A., *BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT.* http://www.mbio.ncsu.edu/bioedit/bioedit.html. Nucl Acids Symp Ser, 1999. **41**: p. 95-98.
36. Tamura, K., et al., *MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0.* Mol Biol Evol, 2007. **24**(8): p. 1596-9.
37. Höper, D., B. Hoffmann, and M. Beer, *A comprehensive deep sequencing strategy for full-length genomes of influenza A.* PLoS One, 2011. **6**(4): p. e19075.
38. Kampmann, M.L., et al., *A simple method for the parallel deep sequencing of full influenza A genomes.* J Virol Methods, 2011. **178**(1-2): p. 243-8.
39. Dugan, V.G., et al., *Phylogenetic analysis of low pathogenicity H5N1 and H7N3 influenza A virus isolates recovered from sentinel, free flying, wild mallards at one study site during 2006.* Virology, 2011. **417**(1): p. 98-105.

CHAPTER 3.4

What's in a strain? Viral metagenomics identifies genetic variation and contaminating circoviruses in laboratory isolates of pigeon paramyxovirus type 1

Steven Van Borm, Toon Rosseel, Mieke Steensels, Thierry van den Berg and Bénédicte Lambrecht

Published in Virus Research (2013) 171(1): p. 186-93.

In a third case study the workflow was used to characterize 11 pigeon type 1 paramyxoviruses (PPMV1) which were isolated during routine diagnostics for avian influenza and avian paramyxoviruses in pigeons. The random access to sequence information allowed the identification of contaminating pigeon circoviruses.

PPMV1 are serotype 1 avian paramyxoviruses which infects pigeons (see introduction of Chapter 3.2 for a brief introduction about APMV). In the past, a number of Newcastle disease outbreaks in poultry have been attributed to PPMV-1, which makes these pigeon-derived viruses a real and continuous threat to the poultry industry.

Pigeon ciroviruses are classified in the *circoviridae* family, having a circular single-stranded DNA genome. Pigeon circovirus infections have been described in several regions of the world. Infected pigeons may develop symptoms such as anorexia, weight loss, reduced race performance, lethargy, respiratory distress and diarrhea.

Abstract

We used next-generation sequencing on random amplified viral nucleic acids to determine the genome sequence of 11 pigeon paramyxovirus type 1 (PPMV-1) isolates from Belgium (period 1998–2011). The PPMV-1 deep sequence data allowed identification of sequence variability in multiple PPMV-1 isolates, including one STOP codon in the Matrix gene which was present in 15 % of the viral population of one isolate. Notably, mutations that were previously associated with pathogenicity in chickens were identified as minor sequence variants in one parent laboratory strain. A phylogenetic analysis of the consensus PPMV-1 genome sequences was performed. In addition to providing nearly complete paramyxovirus genome sequences, our sequence-independent approach identified the presence of pigeon circovirus (PiCV) sequences in four of these viral stocks. Real-time quantitative RT-PCR analysis specific for PMV-1 and PiCV showed that these contaminations were present in seven viral stocks consisting of allantoic fluids and was occasionally also detected in stocks passaged in embryonated chicken eggs. Phylogenetic analysis of the PiCV consensus genome sequences showed a circulation of PiCV covering the full genetic diversity of known PiCV. This study shows the value of novel sequence independent technologies for access to sequence information for the control of reference virus stocks and other biological materials, as co-infecting viruses or sequence variants from the original sample may persist in the stocks without being identified by the routine virus-specific diagnostic tools. The exact role of PiCV in pigeon disease - in particular Newcastle disease - and its potential interference with PPMV-1 diagnostics remains to be investigated.

Keywords: next-generation sequencing, avian paramyxovirus type 1, pigeon circovirus, viral metagenomics, random access amplification

Introduction

Newcastle disease (ND) is caused by avian paramyxovirus type 1 (APMV-1) strains (family *Paramyxoviridae*, genus *Avulavirus*) [1]. It affects many species of birds, and has the potential to cause severe economic losses in the poultry sector worldwide. According to their virulence in poultry, ND virus (NDV) isolates can be classified as lentogenic*, mesogenic* or velogenic [2]. Velogenic strains cause severe disease and high mortality in poultry. The pathogenicity in poultry can be determined using *in vivo* tests, such as the intracerebral pathogenicity index** (ICPI). Alternatively, the determination of the amino acid sequence at the cleavage site of the fusion gene precursor glycoprotein F0 is recognized as a molecular marker of pathogenicity [3]. During the 1980's, the disease spread worldwide among racing and show pigeons. APMV-1 viruses isolated from affected birds were shown to be antigenically and genetically similar to each other but distinct from classical NDV strains. These APMV-1 variants could be discriminated by monoclonal antibodies [4] and are referred to as pigeon paramyxovirus type 1 (PPMV-1) [5, 6], forming a distinguished phylogenetic group, lineage VIb [7, 8], and retaining the potential to cause disease in poultry [5, 9]. PPMV-1 continues to circulate worldwide [7, 10] and are regularly isolated from unvaccinated racing and show pigeons. However, vaccinated adults can also asymptomatically carry the virus and infect young pigeons after the waning of maternally derived antibody and before efficient vaccination*** could be established. Interestingly, PPMV-1 provides some examples of unusual viruses that have a cleavage site motif that is generally associated with virulent viruses and a contrasting ICPI result indicating an avirulent virus [11]. Fuller and collaborators [12] isolated several PPMV-1 clones by limiting dilution in embryonated specific-pathogen-free (SPF) egg starting from a Belgian PPMV-1 isolate with an ICPI of 0.32 (avirulent classification) and a velogenic F protein cleavage site. They identified two genetically similar PPMV-1 clones that differed significantly in their

** lentogenic strains are NDV strains which are nonvirulent and cause only mild clinical signs; mesogenic strains have an intermediate virulence and cause coughing, affect egg quality and production and result in up to 10% mortality*

*** The Intracerebral pathogenicity index (ICPI) is a bioassay for the characterization of a Newcastle disease virus isolate recommended in international diagnostic guidelines, and involves the intracerebral injection of chicks, followed by a standardized scoring of the pathology of the birds.*

**** Both live and killed vaccines for Newcastle disease exist. An introduction about available vaccines worldwide can be found at http://www.fao.org/docrep/005/ac802e/ac802e04.htm* pathogenicity for 1-day-old chickens, despite the fact that they were comparable in *in vitro* infection characteristics such as syncytium formation, plaque size and morphology, and cell-to-cell spread. One virus had an ICPI of 0.025 and the other of 1.3 [12]. The consensus genome of these variants having an ICPI of 0.025 and 1.3 was completely characterized [13] and sequence differences were identified. We previously demonstrated the value of combining random amplification methods with next-generation sequencing technology for the generation of deep sequence data from avian paramyxoviruses [14]. In this study, we apply this sequence independent methodology to historical and recent PPMV-1 isolates, and investigate the viral nucleic acid sequence diversity detected in these isolates.

In addition, the analysis of viral sequence data showed the presence of PiCV in addition to PPMV-1 in some virus stocks. We further investigated this issue using PPMV-1 and PiCV specific quantitative real time (RT-)PCR.

Materials and Methods

Viruses

Eleven viruses isolated from pigeons were selected for in-depth molecular characterization. These viruses were isolated during the routine diagnostic activities of the Belgian reference laboratory for avian influenza and Newcastle disease and were selected to represent a wide timescale (1998-current) and to include isolates that were previously biologically characterized and have different pathogenicity indices (Table 1). The viruses were isolated from diverse avian sample types according to routine virus isolation procedures (OIE, diagnostic manual 2005/94/CE). Briefly, samples were inoculated into the allantoic cavity of 8-10 day-old embryonated specific pathogen free chicken eggs. After incubation for 3-5 days at 37 °C allantoic liquids were harvested and their hemagglutinating activity for chicken red blood cells was investigated. Allantoic liquids with hemagglutinating activity were further characterized in hemagglutination inhibition tests using reference sera [4, 15]. Where possible, the original allantoic fluid was selected for molecular characterization (i.e. without further passages on chicken embryos). Passages on chicken embryos were carried out following routine virological techniques (methodology according to Council Directive 92/66/EC (1992)).

DNAse SISPA and next-generation sequencing

DNase SISPA and 454 sequencing

Sample pretreatment and SISPA were performed independently on the 11 samples (to avoid contamination) as previously described [14, 16]. Briefly, after a centrifugation and filtration step using 0.22 µm filters, the eluate was subjected to DNase I treatment (100 U/200µl sample). The resulting virion-enriched samples were subjected to a viral RNA extraction using the QIAamp Viral RNA Mini Kit (Qiagen). The random first- and second strand cDNA synthesis (using Transcriptor reverse transcriptase, Roche, and Klenow Fragment (exo-), New England Biolabs) was performed as previously described [14] with the primer FR20RV-12N (5'-GCC GGA GCT CTG CAG ATA TCN NNN NNN NNN NN-3'). Subsequently, primer FR20RV (5'-GCC GGA GCT CTG CAG ATA TC-3') was used in PCR (using AmpliTaq Gold 360, Applied Biosystems) to amplify the resulting cDNA as previously described [14]. After visualisation of the random amplified DNA fragments on a 1 % agarose gel, the fragments larger than 400 bp were excised, purified and quantified using a Nanodrop spectrophotometer (Nanodrop Technologies). Five micrograms of each size selected and purified random amplified sample were sequenced on a GS FLX (Roche, Mannheim, Germany) by the Genomics Core of the University Hospital (University of Leuven, Belgium) using multiplex identifier (MID) identification during library preparation and their standard procedures using GS FLX Titanium series reagents (Roche, Mannheim, Germany).

Table 1: Selected virus isolates used for viral (meta)genomic characterization.

Isolate	Year isolated	F cl site sequence	Additional information (reference)
98/238	1998	^{112}RRQKRF^{117}	ICPI 1.25 [11]
98/248	1998	^{112}RRQKRF^{117}	ICPI 0.32 [11]. Cloning and sequencing [12, 13]
98/321	1998	^{112}RRQKRF^{117}	ICPI 0.33 [11]
98/324	1998	^{112}RRKKRF^{117}	ICPI 1.27 [11]
03/05843	2003	^{112}RRQKRF^{117}	From brain sample
05/01824	2005	^{112}RRQKRF^{117}	From lung sample
05/03936/8	2005	^{112}RRQKRF^{117}	From intestine sample
07/04943	2007	^{112}RRQKRF^{117}	From cloacal swab sample
11/07574	2011	^{112}RRQKRF^{117}	From intestine sample
11/08304	2011	^{112}RRQKRF^{117}	From lung sample
11/09620	2011	^{112}RRQKRF^{117}	From brain sample

Data analysis

The raw sequence data were assembled using SeqMan NGen® version 3.1 (DNASTAR, Madison, WI, USA). The reads were trimmed to remove primer sequences (including the primer-encoded random N positions) as well as low quality ends. Standard assembling and filtering parameters were used. Contigs (i.e. sets of overlapping sequence reads) produced in an initial *de novo* assembly were compared to public sequence databases for identification (http://blast.ncbi.nlm.nih.gov/Blast.cgi; [17]). When we identified a PPMV-1, we used the complete genome with the highest identity score to the largest PPMV-1 contig as reference genome for a subsequent reference assembly with the same raw data set. Any other identified viruses were analyzed in the same way. The resulting reference assembly was used to deduce a genome consensus sequence. At polymorphic positions (contradictory sequence reads), we included a degenerate nucleotide in the consensus sequence if the minor nucleotide alternative was present in at least one third of the sequence reads. Polymorphic sites were considered as an indication of genetic heterogeneity in the sampled virus population if the positional sequencing depth was at least 20-fold. Reference assemblies were manually checked for presence of significantly deviating sequences. If this was the case, reads were manually sorted and separate assemblies and consensus sequences were made. All consensus sequences were submitted to Genbank (accession codes in Table 2). Polymorphisms in the reference assemblies were identified using the single nucleotide polymorphism calling tool in SeqMan NGen® version 3.1 (DNASTAR, Madison, WI, USA). The combined error rate of the enzymes used in the SISPA protocol was estimated to be below 1.5×10^{-3} errors/base. However, given the unequal distribution of sequence coverage and the low average coverage in several samples, only clear polymorphisms present in at least 10 % of the reads and at genome positions with a minimum sequence depth of 20-fold were taken into account. Moreover, sequence variants present only in the last 20 nucleotides of a read were excluded.

Phylogenetic analysis and molecular characterization

Consensus sequences were edited, aligned and translated, and sequence identities were calculated using Bioedit v 7.0.5.3 [18]. Selected reference sequences were included to reflect the available sequence diversity of the analyzed viruses in public databases (accession codes specified in figures 1 and 2). For PPMV-1, complete genome sequences were analyzed, while for PiCV an 822 bp region of the ORF2 encoding the capsid protein was analyzed to allow inclusion of more reference data. Multiple sequence alignments were processed using

ClustalW in Bioedit v 7.0.5.3 [18]. Neighbor-Joining phylogenetic trees were constructed in Mega v5.01 [19] using the Tamura-Nei model, 1,000 bootstrap pseudoreplicates, and pairwise deletion of missing data and alignment gaps, allowing the inclusion of incomplete sequences in the analysis. The observed phylogenetic trees were confirmed using Maximum Likelihood analysis (Mega v5.01, data not shown).

Virus specific q(RT-)PCR

To allow a more sensitive and quantitative measurement of PPMV-1 and PiCV in allantoic fluid stocks, specific realtime quantitative (RT-) PCR assays were performed.

PPMV-1 specific quantitative real time RT-PCR

Viral RNA was extracted from allantoic liquids using the QIAamp Viral RNA Mini Kit (Qiagen). Viral RNA was quantified as previously described using a specific real time RT-qPCR [20] using the QuantiTect Probe RT-PCR Kit (Qiagen) on a LightCycler® 480 real time PCR system (Roche, Mannheim, Germany). A standard curve consisting of tenfold dilutions of viral RNA ranging from 7 to 1 $\log_{10}$ RNA copies per reaction was included in each reaction.

Table 2: Sequence coverage statistics for 11 virus stocks targeted by random amplification and next generation sequencing.

Isolate	**Raw reads**	**Reads PPMV1**	**% genome PPMV1 (# nt not covered)**	**Avg coverage PPMV1**	**Reads PiCV**	**% genome PiCV**	**Avg coverage PiCV**	**GB accession PPMV-1**	**GB accession PiCV**
98/238	18,229	2,716	100	48.18				JX901109	
98/248	181,711	46,984	99.99 (1)	871.8				JX901110	
98/321	14,746	7,273	99.24 (116)	125.76				JX901111	
98/324	37,971	418	47.3 (8,008)	8.31	6,011	98.77	812.78	JX901112- JX901114	JX901125
98/324; 98/321-like		69	30.47 (10,568)	2.01				JX901115- JX901117	
03/05843	42,226	871	91.75 (1,253)	18.75				JX901118	
05/01824	36,809	30,759	99.91 (13)	574.42	26	60.96	3.66	JX901119	JX901126
05/03936/8	40,333	5,086	98.55 (220)	90.71				JX901120	
07/04943	51,421	20,724	99.7 (46)	383.57				JX901121	
11/07574	33,988	22,823	98.77 (186)	396.88	236	96.65	32.05	JX901122	JX901127
11/08304	14,733	11,839	99.34 (101)	208.45	149	88.06	20.9	JX901123	JX901128
11/09620	41,788	10,795	98.54 (22)	193.13				JX901124	

PiCV specific quantitative real time PCR

PiCV viral DNA was extracted using QIAamp DNA Mini Kit (Qiagen) and quantified using a specific realtime qPCR [21] as previously described using the LightCycler® 480 SYBR Green I Master kit on a LightCycler® 480 real time PCR system (Roche, Mannheim, Germany). A standard curve consisting of tenfold dilutions of plasmid DNA (a 367 bp amplicon, spanning the realtime PCR priming sites in a pcDNA3.1/V5-his-TOPO vector kindly provided by J.P. Duchatel) ranging from 7 to 1 $\log_{10}$ plasmid copies per reaction was included in each reaction.

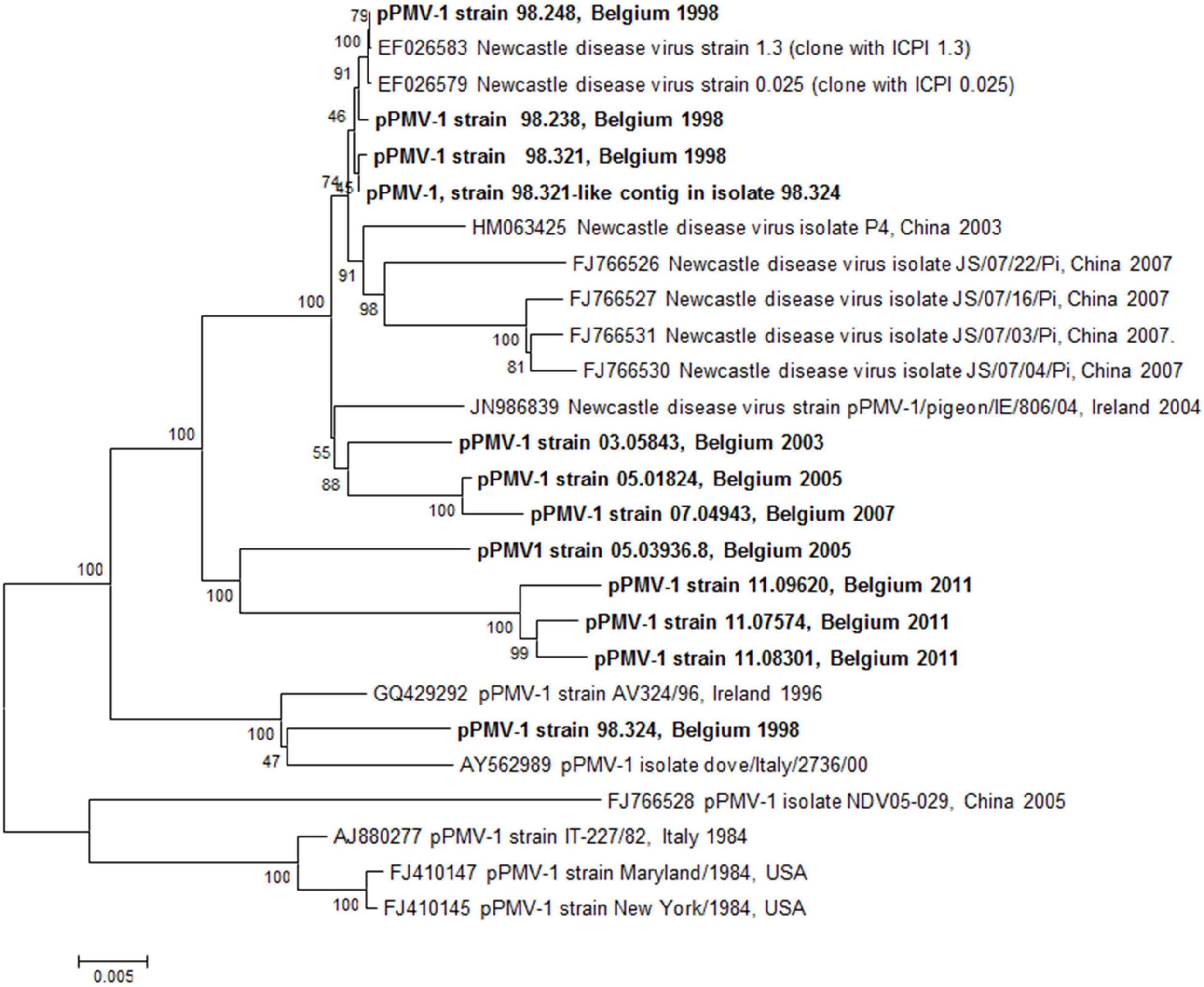

79
pPMV-1 strain 98.248, Belgium 1998
100
EF026583 Newcastle disease virus strain 1.3 (clone with ICPI 1.3)
91
EF026579 Newcastle disease virus strain 0.025 (clone with ICPI 0.025)
46
pPMV-1 strain 98.238, Belgium 1998
pPMV-1 strain 98.321, Belgium 1998
74
45
pPMV-1, strain 98.321-like contig in isolate 98.324
HM063425 Newcastle disease virus isolate P4, China 2003
91
FJ766526 Newcastle disease virus isolate JS/07/22/Pi, China 2007
100
98
FJ766527 Newcastle disease virus isolate JS/07/16/Pi, China 2007
100
FJ766531 Newcastle disease virus isolate JS/07/03/Pi, China 2007.
81
FJ766530 Newcastle disease virus isolate JS/07/04/Pi, China 2007
JN986839 Newcastle disease virus strain pPMV-1/pigeon/IE/806/04, Ireland 2004
100
55
pPMV-1 strain 03.05843, Belgium 2003
88
pPMV-1 strain 05.01824, Belgium 2005
100
pPMV-1 strain 07.04943, Belgium 2007
pPMV1 strain 05.03936.8, Belgium 2005
100
100
pPMV-1 strain 11.09620, Belgium 2011
100
pPMV-1 strain 11.07574, Belgium 2011
99
pPMV-1 strain 11.08301, Belgium 2011
GQ429292 pPMV-1 strain AV324/96, Ireland 1996
100
pPMV-1 strain 98.324, Belgium 1998
47
AY562989 pPMV-1 isolate dove/Italy/2736/00
FJ766528 pPMV-1 isolate NDV05-029, China 2005
AJ880277 pPMV-1 strain IT-227/82, Italy 1984
100
FJ410147 pPMV-1 strain Maryland/1984, USA
100
FJ410145 pPMV-1 strain New York/1984, USA
0.005

Figure 1: Neighbor-Joining phylogenetic tree of pigeon paramyxovirus-1 complete genomes determined in this study and available complete genome sequences. Bootstrap support values > 70 are indicated next to nodes. Bold type: characterized in this study.

Results

DNAse SISPA and next-generation sequencing

The MID-sorted sequence output varied for the different sequencing libraries (14,733 to 181,711 raw sequence reads, average read length 211.55 nucleotides, Table 2). This was probably due to a poor quantification of the sequencing libraries prior to pooling them for the 454 titanium emulsion PCR and sequencing workflow (out of our hands due to outsourcing of the sequencing workflow including library preparation, emulsion PCR and sequencing). Using a reference assembly (PPMV-1 sequence EF025683 as reference, except for isolate 98/324 where GQ429292 was used), 99-100 % of the genome sequence of six isolates could be determined. For three additional isolates, about 98 % of the genome was assembled. The two remaining isolates where only < 1,000 PPMV-1 reads were available had respectively 47 and 92 % of their genomes sequenced, indicating the need of a minimum sequencing effort for completion of paramyxovirus genomes by this method. The sequence depth was not homogenously distributed over the genome, but the distribution of sequence depth in function of the PPMV-1 genome position was reproducible across all PPMV-1 assemblies (data not shown).

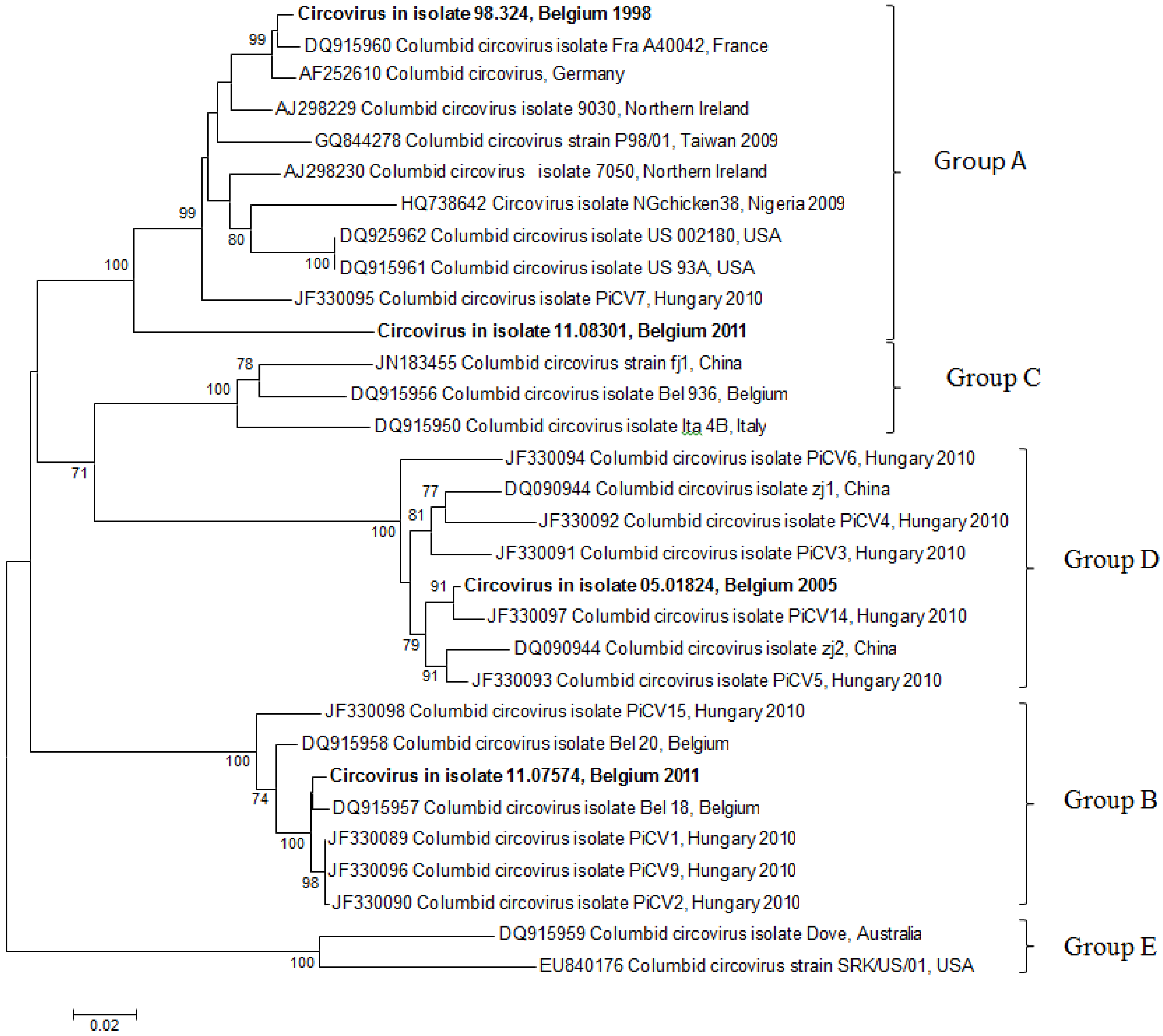

Circovirus in isolate 98.324, Belgium 1998
DQ915960 Columbid circovirus isolate Fra A40042, France
AF252610 Columbid circovirus, Germany
AJ298229 Columbid circovirus isolate 9030, Northern Ireland
GQ844278 Columbid circovirus strain P98/01, Taiwan 2009
AJ298230 Columbid circovirus isolate 7050, Northern Ireland
HQ738642 Circovirus isolate NGchicken38, Nigeria 2009
DQ925962 Columbid circovirus isolate US 002180, USA
DQ915961 Columbid circovirus isolate US 93A, USA
JF330095 Columbid circovirus isolate PiCV7, Hungary 2010
Circovirus in isolate 11.08301, Belgium 2011
JN183455 Columbid circovirus strain fj1, China
DQ915956 Columbid circovirus isolate Bel 936, Belgium
DQ915950 Columbid circovirus isolate Ita 4B, Italy
JF330094 Columbid circovirus isolate PiCV6, Hungary 2010
DQ090944 Columbid circovirus isolate zj1, China
JF330092 Columbid circovirus isolate PiCV4, Hungary 2010
JF330091 Columbid circovirus isolate PiCV3, Hungary 2010
Circovirus in isolate 05.01824, Belgium 2005
JF330097 Columbid circovirus isolate PiCV14, Hungary 2010
DQ090944 Columbid circovirus isolate zj2, China
JF330093 Columbid circovirus isolate PiCV5, Hungary 2010
JF330098 Columbid circovirus isolate PiCV15, Hungary 2010
DQ915958 Columbid circovirus isolate Bel 20, Belgium
Circovirus in isolate 11.07574, Belgium 2011
DQ915957 Columbid circovirus isolate Bel 18, Belgium
JF330089 Columbid circovirus isolate PiCV1, Hungary 2010
JF330096 Columbid circovirus isolate PiCV9, Hungary 2010
JF330090 Columbid circovirus isolate PiCV2, Hungary 2010
DQ915959 Columbid circovirus isolate Dove, Australia
EU840176 Columbid circovirus strain SRK/US/01, USA
Group A
Group C
Group D
Group B
Group E
99
99
80
100
100
78
100
71
77
81
100
91
79
91
100
74
100
98
100
0.02

Figure 2: Neighbor-Joining phylogenetic tree of pigeon circovirus ORF2 sequences detected in laboratory isolates of PPMV-1 and available reference sequences. Previously described clade terminology [22] is depicted next to the clusters. Bootstrap support values > 70 are indicated next to nodes. Bold type: characterized in this study.

Molecular characterization

Fusion gene cleavage site

All isolates sequenced in this study have amino acid sequence 112-RRQKRF-117 at their fusion gene cleavage site (Table 1), except 98/324 which has amino acid sequence 112-RRKKRF-117 like the other viruses in its phylogenetic cluster (Figure 1: AV324/6 and dove/Italy/2736/00). No sequence variability could be detected at the cleavage site.

Sequence variability

Strain 98/248 was previously determined to have an ICPI of 0.32. Two clones with contrasting ICPI values (0.025 vs. 1.3) were previously isolated [12] from this virus and sequence differences between these clones were determined at the consensus level [13]. We sequenced the parental isolate 98/248 that was used by Fuller et al for cloning. 98/248 had average genome coverage of > 850-fold (more than 45,000 reads mapped to PPMV-1 reference sequence), which allowed a detailed analysis of variability. Interestingly, all previously documented mutations differentiating the consensus genome of clones 0.025 and 1.3 [13] were present at variable prevalence in the quasispecies population of the parental virus isolate (Table 3).

Table 3: Sequence variability detected in PPMV-1 isolate 98/248 in relation to previously detected markers of pathogenicity in a cloning experiment by [12, 13]. The last column indicates the consensus nucleotide when comparing all available PPMV-1 genome sequences at this position. Degenerate nucleotide R = A+G. Nt: nucleotide.

Genome position (gene)	Minor/major nt (%minor)	Consensus sequence	Strain "0.025" EF026579	Strain "1.3" EF026583	Other PPMV-1
1,771 (NP, synonymous)	43T/92C (31.9 %)	C	C	C	C
3,593 (M)	46A/90G (33.8 %)	R	G (Asp)	A (Asn)	G
5,906 (F)	3C/2555T (0.12 %)	T	C (Pro)	T (Ser)	T
12,518 (L)	108G/373A (22.5 %)	A	G (Val)	A (Met)	A
14,220 (L)	5G/1559A (0.32 %)	A	G (Asn)	A (Ser)	A

Sequence variability in other samples: Reliable sequence variants were detected in 98/248 (see above) and 5 additional isolates (Table 4). We identified a single nucleotide polymorphism in isolate 98/324 in the coding region of the F gene, although we only had limited sequence information for this sample because more PiCV was present than PPMV-1. Isolate 05/03936/8 showed one non-synonymous and one synonymous minor sequence variant in the polymerase gene (L). One sequence variant in the 3' synonymous region and two additional synonymous sequence variants (one in the matrix (M) gene and one in the polymerase (L) gene) were observed in 07/04943. The nucleoprotein gene (NP) of 11/08304 showed one synonymous sequence variant, while one non-synonymous sequence variant was detected at 18% prevalence in its phosphoprotein (P) gene. Isolate 11/09620 showed three synonymous sequence variants in its matrix, fusion and polymerase genes. Interestingly, a sequence variant resulting in a stop codon at amino acid position 118 of its matrix gene was present at 14.8 % in isolate 11/09620.

Table 4: Sequence variability detected in other isolates. Bold type: non-synonymous mutations. Degenerate nucleotides: R = A+G, Y = C+T. AA: amino acid. Nt: nucleotide

Isolate	Genome position cf. EF026583	Major/minor nt (%minor)	Consensus sequence	Gene AA position
98/324	6,305	308A/94T (23.4 %)	A	F synonymous
05/03936/8	8,802	8T/4C (33,3 %)	Y	**L Leu139Ser**
	11,011	35T/5C (12.5 %)	T	L Cys875Cys
07/04943	95	31C/17T (35.4 %)	Y	3’ NCR
	4,303	357T/76C (17.6 %)	T	M Ala336Ala
	14,572	204T/120C (37.0 %)	Y	L Ser2062Ser
11/08304	418	197G/91A (31.6 %)	G	NP Leu99Leu
	2,158	62C/14T (18.04 %)	C	**P Pro89Leu**
11/09620	3,647	98A/17T (14.8 %)	A	**M Lys118STOP**
	4,441	155C/41T (20.9 %)	C	M synonymous
	6,276	36A/25G (41.0 %)	R	F synonymous
	11,675	15T/9C (37.5 %)	Y	L Val1096Val

Phylogenetic analysis

Pigeon paramyxovirus type 1. The Belgian viruses isolated in 1998 clustered in two distinctive groups (Figure 1). Three isolates cluster with derived strains 0.025 en 1.30 [12, 13]. These include the parental strain of the latter clones, 98/248 (deep sequenced in the present study), and two additional isolates 98/238 and 98/321 isolated in Belgium during the same period. A single isolate (98/324) clusters with dove/Italy/2736/2000 and AV324/96 (strain from Ireland), and also contained 69 sequence reads of a second virus identical to 98/321. This isolate represents a co-infection of 2 viruses.

The Belgian isolates from 2003 to 2007 cluster in a separate clade, related to a PPMV-1 isolated in Ireland in 2004, with the exception of a single Belgian strain (05/3936/8) from 2005 that clusters with the most recent Belgian isolates (Figure 1).

Pigeon circovirus. PiCV sequences were present in four isolates from 1998, 2005, and 2011. An 822 bp region of the PiCV ORF2 encoding the capsid protein was chosen for phylogenetic analysis to allow inclusion of a large number of PiCV data from public databases. These Belgian PiCV sequences encompass the complete diversity of PiCV sequence data previously demonstrated to circulate endemically in Europe [22, 23], including Genotypes A (two viruses), B (one virus) and D (one virus) (Figure 2). As previously documented [23], there was no evidence of clustering according to geographical origin. There seemed to be no co-segregation of PiCV and PPMV-1 sequence data.

Quantification of PiCV and PPMV1

Specific realtime (RT-)PCR assays allowed for the quantification of PiCV and PPMV1 in allantoic fluid stocks (Table 5). Apart from confirming the presence of high quantities of PiCV (> 4 $\log_{10}$ genome copies) in isolates where the virus was detected by metagenomics, lower quantities of PiCV DNA were detected in three additional virus stocks. In some instances PiCV could still be detected when these viruses were passaged on embryonated chicken eggs (Table 5). In one incidence (11/07574) PiCV was maintained after passage, while the quantity of PPMV1 dramatically decreased. However, in most cases the PiCV quantities detected after passage on embryonated chicken eggs decreased.

Interestingly, in one isolate (98/324), more PiCV was present than PPMV1, although this isolate was readily identified using classical virology tools as PPMV1. This virus stock proved unviable upon our attempt to passage it in eggs.

Table 5: Quantification of PiCV and PPMV1 in viral stocks by real time (RT-) PCR. na: material was not available for analysis. Isolates indicated in bold type: PiCV sequences detected in original allantoic fluid by SISPA-NGS. “<det” : below detection limit.

Isolate	**Original allantoic fluid**		**Allantoic fluid first passage (1P)**	
	PPMV-1 $\log_{10}$ genome copies/ml	PiCV $\log_{10}$ genome copies/ml	PPMV-1 $\log_{10}$ genome copies/ml	PiCV $\log_{10}$ genome copies/ml
98/238	na	na	7.65	2.83
98/248	na	na	7.23	<det
98/321	na	na	7.05	1.16
98/324	**2.33**	**5.34**	**na**	**na**
03/05843	5.98	3.07	7.93	<det
05/01824	**7.65**	**4.09**	**7.20**	**2.60**
05/03936/8	5.93	<det	7.50	<det
07/04943	7.76	<det	7.79	1.48
11/07574	**5.53**	**4.53**	**2.60**	**5.00**
11/08304	**7.31**	**4.75**	**7.79**	**3.36**
11/09620	8.06	<det	na	na

Discussion

In an effort to characterize historical and recent PPMV-1 isolates by random access next-generation sequencing as previously described [14], we identified pigeon circovirus contamination in four laboratory isolates diagnosed with virus-specific diagnostic tools as PPMV-1. This is somewhat surprising as there have been no reports describing the isolation or propagation of PiCV in culture cells or allantoic cavities [21]. Typically these viruses were isolated from suspected clinical samples (identified as PMV-1 positive by real time RT-PCR for the recent samples) by inoculation in the allantoic cavity of embryonated specific pathogen free chicken eggs. Allantoic fluids are then harvested and checked for presence of hemagglutinating activity, and an identification of hemagglutinating agent was done the using hemagglutination inhibition tests using anti-NDV or anti-AIV reference sera. Although it is not sure that the biological characteristics (notably virulence as measured by ICPI) of PPMV1 stocks are affected by the mixed infection, their presence in PPMV1 positive pigeon material needs further consideration. This must be related, at least in part, to the limitation of current diagnostics to detect PiCV and to the pooling of suspected organs. Although the exact role of PiCV in pigeon disease and its potential interference with diagnostic tests is the subject of an ongoing follow-up study (in preparation), our data show the importance of in-depth characterization of viral isolates for studies on virus pathogenicity and/or evolution. In particular, the potential immunosuppressive role of PiCV that has previously been suggested [24] may interfere with in vivo paramyxovirus infections.

Nearly complete genome sequences could be obtained for all but one PPMV1 and for the four pigeon circoviruses. Although our SISPA-NGS workflow was focused on the enrichment of RNA virus sequence data (use of an RNA extraction kit), we did not treat the enriched virions with an RNase. The resulting DNA virus reads are thus the result of either (1) DNA virus transcribed RNA co-purified with the virions (no RNase treatment of virions) and/or (2) viral DNA co-purified during the viral RNA extraction procedure (as no additional DNase treatment was performed during the RNA extraction procedure). The fact that we obtained near complete PiCV genomes, including non transcribed sequences (e.g. the origin of replication and the *Rep* promotor), suggests that viral DNA co-purification is, at least in part, contributing to the detection of viral DNA sequences.

Applying PPMV-1 and PiCV specific quantitative real time RT-PCR to DNA and RNA extracted from these virus isolates, we confirmed the presence of PiCV in PPMV-1 stocks and showed that PiCV seemed to be present in high quantities in several PPMV-1 stocks.

PiCV has a worldwide distribution with high prevalence in young pigeons [21, 23]. A high prevalence in young feral and domestic pigeons has been described [25, 26], although it can also be present in older pigeons [21]. Its worldwide distribution and lack of geographical clustering is most likely related to the extensive worldwide trade in ornamental and racing pigeons. PiCV is involved in the "young pigeon disease syndrome", where morbidity and mortality are reported in pigeons aged 4-12 weeks [24]. However, experimental infection with PiCV could not reproduce the disease, indicating a multifactorial cause [27]. A potential immunosuppressive role has been suggested [24], facilitating other pathogenic infections.

The selected Belgian isolates for deep sequencing included strain 98/248, the parent strain with an intermediate pathogenicity index of 0.32 that was used by Fuller et al [12] for the isolation by cloning of variants with contrasting pathogenicity indices (1.3 and 0.025). These clones with contrasting pathogenicity phenotype were later completely sequenced [13]. Our deep sequencing data (average coverage > 850 x) allowed a quasispecies analysis of the viral population in the parental strain used for these cloning experiments. Interestingly, all sequence differences previously identified between these variants were confirmed as minority mutations in the parental strain 98/248. The intermediate low pathogenic phenotype (ICPI 0.32) of the parent strain was thus most likely related to the presence of multiple genetic variants, some of which contained markers for increased pathogenicity in poultry that were subsequently selected from the quasispecies virus stock population in the cloning experiments by Fuller et al. In addition, a mutation in the synonymous region of the NP gene was identified in the parent strain that did not appear in the selected variant viruses. The deep sequencing approach allows for a quantification of these quasispecies in the original isolate. While some genetic variants seem to be present in about one third of the viral population, others seem to be present at a very low frequency in the seed viral stock.

Another study documented molecular changes of PPMV-1 isolate Av324 after passaging in chickens [28]. None of the 3 mutations that were associated with enhanced replication efficiency in that study (at positions 2,001, 13,077 and 13,467) were found in any of the viruses characterized in our study, although one virus was closely related to Av324.

The sequence depth in several of our strains allowed the documentation of additional polymorphic sites.

These data show the power of deep sequencing technologies in characterizing the quasispecies composition of RNA virus isolates. We could confirm that the intermediate pathogenicity phenotype that Fuller and colleagues [12] attributed to mixed viral populations is indeed at the genetic level due to a complex mixture of genetic variants.

For some isolates, although sufficient sequence information (average coverage > 300x) was available (e.g. 05/01824 and 11/07874), no sequence variation present in at least 10 % of the reads was detected. As we have no formal or experimental data describing the error introduction of the random amplification and 454 sequencing processes used in this study, we did not want to hypothesize on the importance of these minor sequence variants (< 10%).

The sequence independent approach used here allowed the identification of contaminating viruses in biological material, including unexpected circoviruses and co-infection (or contamination) with multiple avian paramyxoviruses in a single laboratory stock. These findings show the power of next-generation on random amplified viral nucleic acids for the quality control of virus reference isolates and other biological materials. As the methodology provides a sequence independent probing of the viral nucleic acid content of a sample, minority variants and adventitious viruses present in the stock can be identified while identifying high quality genomic data of the reference isolates. As such, NGS technology may provide a powerful quality control tool to reference laboratories and producers of biological materials such as vaccines and cell lines.

Caution should be taken not to overestimate the sensitivity of this methodology for the identification of contaminating viruses. Our preliminary data (not shown) indicated that at least 5 $\log_{10}$ virions need to be present in the sample to allow identification in a background of host and contaminating nucleic acids. This was confirmed in this study, where about 4 $\log_{10}$ PiCV copies were sufficient to allow detection by SISPA-NGS. Lower copy numbers did not yield PiCV sequence data. Samples containing high virus titers (as determined by realtime quantitative (RT-) PCR) yielded high quality genome sequences. In one instance (98/324) the isolate actually contained more PiCV than PPMV-1, but a relatively low quantity (2.3 $\log_{10}$) of PPMV-1 could still reliably be identified by SISPA-NGS. Follow-up (multiplex) PCRs should be designed to confirm and accommodate the limited sensitivity of metagenomic findings.

As future technology developments promise to increase the accessibility of NGS to virology labs and to decrease its costs, its application as a quality control and diagnostic technique with high information content for biological reagents such as viral stocks will be worth considering as a complementary methodology to virus-specific test design.

Acknowledgements

We are grateful to Orkun Ozhelvaci and Mireille Decaesstecker for excellent technical assistance and to Frank Vandenbussche for critically proofreading the manuscript. Jean-Pierre Duchatel (FVM Liège) kindly provided control reagents for the PiCV realtime PCR. The professional services of the Genomics Core of the University Hospital of Leuven are appreciated. This work was supported by Epi-SEQ: a research project supported under the 2nd joint call for transnational research projects by EMIDA ERA-NET (FP7 project nr 219235), the EU Network of Excellence EPIZONE (FOODCT-2006-016236) and an internal research and development fund at CODA-CERVA (RANDSEQ).

References

1. Lamb, R.A. and G. Parks, *Paramyxoviridae: the viruses and their replication 5th edition.* 2007, Philadelphia: Lippincott Williams and Wilkins.
2. Alexander, D.J., *Newcastle disease*, in *Disease of poultry*, Y.M. Saif, et al., Editors. 2003, Iowa State Press: Ames IA. p. 64-87.
3. Alexander, D.J., *Newcastle disease and other avian Paramyxoviruses.* Rev Sci Tech Off int Epiz, 2000. **19**(2): p. 443-462.
4. Collins, M.S., et al., *Evaluation of mouse monoclonal antibodies raised against an isolate of the variant avian paramyxovirus type 1 responsible for the current panzootic in pigeons.* Arch Virol, 1989. **104**(1-2): p. 53-61.
5. Alexander, D.J., et al., *Antigenic and biological characterisation of avian paramyxovirus type I isolates from pigeons--an international collaborative study.* Avian Pathol, 1985. **14**(3): p. 365-76.
6. Alexander, D.J., et al., *Avian paramyxovirus type 1 infection of racing pigeons: 3 epizootiological considerations.* Vet Rec, 1984. **115**(9): p. 213-6.
7. Aldous, E.W., et al., *A molecular epidemiological investigation of isolates of the variant avian paramyxovirus type 1 virus (PPMV-1) responsible for the 1978 to present panzootic in pigeons.* Avian Pathol, 2004. **33**(2): p. 258-269.
8. Aldous, E.W., et al., *A molecular epidemiological study of avian paramyxovirus type 1 (Newcastle disease virus) isolates by phylogenetic analysis of a partial nucleotide sequence of the fusion protein gene.* Avian Pathol, 2003. **32**(3): p. 239-56.
9. Alexander, D.J., G. Parsons, and R. Marshall, *Infection of fowls with Newcastle disease virus by food contaminated with pigeon faeces.* Vet Rec, 1984. **115**(23): p. 601-2.
10. Alexander, D.J., *Gordon Memorial Lecture. Newcastle disease.* Br Poult Sci, 2001. **42**(1): p. 5-22.

11. Meulemans, G., et al., *Evolution of pigeon Newcastle disease virus strains.* Avian Pathol, 2002. **31**(5): p. 515-9.
12. Fuller, C.M., et al., *Partial characterisation of five cloned viruses differing in pathogenicity, obtained from a single isolate of pigeon paramyxovirus type 1 (PPMV-1) following passage in fowls' eggs.* Arch Virol, 2007. **152**(8): p. 1575-82.
13. Dortmans, J.C., et al., *Two genetically closely related pigeon paramyxovirus type 1 (PPMV-1) variants with identical velogenic fusion protein cleavage sites but with strongly contrasting virulence.* Vet Microbiol, 2010. **143**(2-4): p. 139-44.
14. Rosseel, T., et al., *Identification and complete genome sequencing of paramyxoviruses in mallard ducks (Anas platyrhynchos) using random access amplification and next generation sequencing technologies.* Virol J, 2011. **8**: p. 463.
15. Alexander, D.J., P.H. Russell, and M.S. Collins, *Paramyxovirus type 1 infections of racing pigeons: 1 characterisation of isolated viruses.* Vet Rec, 1984. **114**(18): p. 444-6.
16. Van Borm, S., et al., *Phylogeographic analysis of avian influenza viruses isolated from Charadriiformes in Belgium confirms intercontinental reassortment in gulls.* Arch Virol, 2012. **157**(8): p. 1509-22.
17. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
18. Hall, T.A., *BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. http://www.mbio.ncsu.edu/bioedit/bioedit.html.* Nucl Acids Symp Ser, 1999. **41**: p. 95-98.
19. Tamura, K., et al., *MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods.* Mol Biol Evol, 2011. **Epub ahead of print**.
20. Wise, M.G., et al., *Development of a real-time reverse-transcription PCR for detection of Newcastle Disease virus RNA in clinical samples.* J. Clin. Microbiol., 2004. **42**(1): p. 329-338.
21. Duchatel, J.P., et al., *Quantification of pigeon circovirus in serum, blood, semen and different tissues of naturally infected pigeons using a real-time polymerase chain reaction.* Avian Pathol, 2009. **38**(2): p. 143-8.
22. Csagola, A., et al., *Genetic diversity of pigeon circovirus in Hungary.* Virus Genes, 2012. **44**(1): p. 75-9.
23. Todd, D., et al., *Sequence comparison of pigeon circoviruses.* Res Vet Sci, 2008. **84**(2): p. 311-9.
24. Raue, R., et al., *A disease complex associated with pigeon circovirus infection, young pigeon disease syndrome.* Avian Pathol, 2005. **34**(5): p. 418-25.
25. Franciosini, M.P., et al., *Development of a polymerase chain reaction-based in vivo method in the diagnosis of subclinical pigeon circovirus infection.* Avian Dis, 2005. **49**(3): p. 340-3.
26. Krapez, U., et al., *Prevalence of pigeon circovirus infections in feral pigeons in Ljubljana, Slovenia.* Avian Dis, 2012. **56**(2): p. 432-5.
27. Schmidt, V., et al., *Experimental infection of domestic pigeons with pigeon circovirus.* Avian Dis, 2008. **52**(3): p. 380-6.
28. Dortmans, J.C., et al., *Passaging of a Newcastle disease virus pigeon variant in chickens results in selection of viruses with mutations in the polymerase complex enhancing virus replication and virulence.* J Gen Virol, 2011. **92**(Pt 2): p. 336-45.

CHAPTER 3.5

DNase SISPA-next generation sequencing confirms Schmallenberg virus in Belgian field samples and identifies genetic variation in Europe

Toon Rosseel, Matthias Scheuch, Dirk Höper, Nick De Regge and Ann Brigitte Caij,

Frank Vandenbussche and Steven Van Borm

Published in PLoS ONE (2012) 7(7): p. e41967.

In the 3 previous Chapters virus isolates were used to evaluate the protocol, containing high virion concentrations and low background nucleic acid contaminants. The case study in this Chapter aimed to test the sensitivity of the workflow on real clinical samples. Brain tissue samples of sheep naturally infected with a recently emerging orthobunyavirus were selected. The expression “SISPA” refers to rPCR SISPA amplification method as defined in Chapter 1, and “DNase” emphasizes that pretreatment steps were performed, including a DNase treatment for the removal of unprotected nucleic acids.

Orthobunyavirus is a genus of the *bunyaviridae* family with a segmented negative-stranded RNA genome (3 segments). The genus is most diverse in Africa, Australia and Oceania, but occurs almost worldwide. Most orthobunyavirus species are transmitted by mosquitoes or ticks and cause diseases of cattle; on the other hand, the California encephalitis virus and the La Crosse virus, are North American species that cause encephalitis in humans.

Schmallenberg virus (SBV), a novel orthobunyavirus, was first isolated in 2011 in Germany. SBV preferentially infects the central nervous system of cattle and sheep and causes fever, diarrhea, a drop in milk yields, congenital malformations and stillbirths.

Abstract

In 2011, a novel Orthobunyavirus was identified in cattle and sheep in Germany and the Netherlands. This virus was named Schmallenberg virus (SBV). Later, presence of the virus was confirmed using real time RT-PCR in cases of congenital malformations of bovines and ovines in several European countries, including Belgium. In the absence of specific sequencing protocols for this novel virus we confirmed its presence in RT-qPCR positive field samples using DNase SISPA next generation sequencing (NGS), a virus discovery method based on random amplification and next generation sequencing. An in vitro transcribed RNA was used to construct a standard curve allowing the quantification of viral RNA in the field samples. Two field samples of aborted lambs containing 7.66 and 7.64 log10 RNA copies per mL total RNA extract allowed unambiguous identification of SBV. One sample yielded 192 SBV reads covering about 81 % of the L segment, 56 % of the M segment and 13 % of the S segment. The other sample resulted in 8 reads distributed over the L and M segments. Three weak positive field samples (one from an aborted calf, two from aborted lambs) containing virus quantities equivalent to 4.27-4.89 log10 RNA copies per mL did not allow identification using DNase SISPA-NGS. This partial sequence information was compared to the whole genome sequence of SBV isolated from bovines in Germany, identifying several sequence differences. The applied viral discovery method allowed the confirmation of SBV in RT-qPCR positive brain samples. However, the failure to confirm SBV in weak PCR-positive samples illustrates the importance of the selection of properly targeted and fresh field samples in any virus discovery method. The partial sequences derived from the field samples showed several differences compared to the sequences from bovines in Germany, indicating sequence divergence within the epidemic.

Introduction

During the summer and autumn of 2011, a novel disease with symptoms including fever, decreased milk production and diarrhea, was identified in dairy cattle in Germany and The Netherlands [1, 2]. Using a metagenomic analysis on next generation sequence data produced from the blood of symptomatic animals, a novel Orthobunyavirus was shown to be associated with the disease [1]. The virus was preliminary named Schmallenberg virus (SBV) according to the geographical location of the index case. A specific real time RT-PCR test was developed, confirming the presence of the virus in diseased bovines. Animal experiments with the isolated virus further supported a causal relationship between the virus and the disease [1]. In addition, the virus proved to be associated with an outbreak of congenital malformations and abortions in both ovine and bovine [1, 3, 4]. The dissemination of real time RT-PCR protocols to laboratories throughout Europe allowed the detection of Schmallenberg virus in six additional countries, including Belgium, France, Luxembourg, United Kingdom, Italy, and Spain [5] and provided evidence for the involvement of *Culicoides* sp. midges as possible vectors [6].

In the absence of targeted sequencing protocols for this novel virus, we applied a virus discovery strategy based on random amplification of purified nucleic acids in combination with next generation sequencing on SBV real time RT-PCR positive tissue samples to confirm the presence of SBV and obtain preliminary sequence data on SBV from Belgium. We previously validated this DNase SISPA (Sequence Independent Single Primer Amplification, [7]) approach on other RNA viruses [8, 9]. It consists of a viral nucleic acid enrichment step (size selective filtration in combination with a nuclease treatment to remove nucleic acids that are not encapsidated in virions) followed by a random cDNA synthesis and amplification step. The random amplicons are subsequently exploited by next generation sequencing [10]. The combination with quantitative real time RT-PCR results from the field tissue samples allowed a first estimate of the sensitivity of this approach using field samples infected with an emerging disease.

Materials and Methods

Samples

Diagnostic field samples from suspected cases of SBV related congenital malformations in lambs and calves were selected based on their geographical location (Table 1) and on the Cp values of the RT-qPCR detecting the L-segment of the virus that was used for diagnosis [1]. This study was conducted under the authorization and supervision of the Bioethics Committee at the Veterinary and Agrochemical Research Center (VAR), following national and European regulations.

Viral RNA quantification

Briefly, approximately 0.5 cm^3 of brain tissue was added to 1ml PBS and homogenized (2min, 25Hz) in a TissueLyser (Qiagen, Venlo, The Netherlands). The RNA was extracted using the RNeasy minikit (Qiagen) following manufacturer instructions and eluted in 50µl. 2µl of this RNA mixture was further analyzed by a one-step PCR using the LightCycler 480 RNA Master Hydrolysis Probes kit (Roche Diagnostics, Vilvoorde, Belgium) on a LightCycler 480 Real-time PCR system following manufacturer instructions. Primers and probe sequences for the SBV L segment detection were kindly provided by Dr. B. Hoffmann (FLI, Germany, available on request) and used at a final concentration of 1 and 0.1875 µM respectively. Two highly positive brain tissue samples ($Cp < 21$) from aborted lambs were selected for confirmation by DNase SISPA-NGS. In addition, two weak positive brain tissue samples from lambs and one weak positive sample form a calf were selected ($Cp > 28$). The viral RNA load in RNA directly extracted from these samples was quantified in the above described real time RT-PCR using a standard curve consisting of in vitro transcribed L segment RNA spanning the diagnostic RT-qPCR, which was run in five replicates. The in vitro transcribed RNA was independently quantified using three different approaches: NanoDrop Spectrophotometer (Nanodrop Technologies, Wilmington DE, USA), Qubit® RNA assay kit on Qubit® fluorometer (Invitrogen-Life Technologies, Gent, Belgium), and RNA 6000 Pico chip on the Agilent 2100 Bioanalyser (Agilent Technologies, Diegem, Belgium).

DNase SISPA and 454 sequencing

The maximum available quantity of the limited tissue samples (< 1.5 g; Table 1) was homogenized in about 1,500 µL PBS per g of tissue using gentle homogenization in a TissueLyser (Qiagen). Sample pretreatment and SISPA was largely performed as previously described [8, 9]. Briefly, after a centrifugation and filtration step using 0.22 µm filters, the eluate was subjected to DNase I treatment (100 U/200µl sample). The resulting virion-enriched samples were subjected to a viral RNA extraction using the QIAamp Viral RNA Mini Kit (Qiagen). Forty units of Protector RNase inhibitor (Roche) were added to the eluted RNA, and RNA quality was checked using a Bioanalyzer 2100 (Agilent Technologies). The random first- and second strand cDNA synthesis was performed with the primer FR26RV-N (5'GCC GGA GCT CTG CAG ATA TCN NNN NN 3') and primer FR20RV (5'-GCC GGA GCT CTG CAG ATA TC-3') was used in subsequent PCR to amplify the resulting cDNA. After visualisation of the random amplified DNA fragments on a 1 % agarose gel, the fragments between 200 and 1,000 bp were excised, purified and quantified using a Nanodrop spectrophotometer (Nanodrop Technologies). Five micrograms of each size selected (200-1,000 bp) and purified random amplified sample were sequenced on a GS FLX+ (Roche, Mannheim, Germany) by the Genomics Core of the University Hospital (University of Leuven, Belgium) using multiplex identifier (MID) identification during library preparation and their standard procedures using GS FLX Titanium series reagents (Roche, Mannheim, Germany). The DNA fragmentation step by nebulization was skipped and the intention was to obtain 30,000-40,000 reads per library.

Metagenomic analysis

As an additional control to exclude reads originating from potential DNA contamination during the library preparation steps, only reads containing the SISPA primer sequence were included in the assembly. Subsequently, the SISPA primer sequences plus additional six bases were trimmed off the reads. By using a combination of BLAST [11] and sequence mapping with the 454 reference mapper application (version 2.6; Roche), contigs (i.e. sets of overlapping sequence reads) and reads were classified into different taxa.

Reference assembly

To map the obtained raw sequence data to the genome of Schmallenberg virus, the complete coding sequence of SBV isolate BH80/11-4 (accession codes HE649912-HE649914) was used as reference genome in the reference assemblies of our different field samples using

SeqMan NGen® version 3 (DNASTAR, Madison, WI, USA). The reads were first trimmed to remove primer sequences (including the primer-encoded random N positions) as well as low quality ends. Standard assembling and filtering parameters were used, except for a reduced minimum match percentage. The partial sequence information was made accessible through GenBank accession numbers JQ861686-JQ861692, except for fragments that were less than 200 bp in length due to the minimum fragment length requirements dictated by GenBank.

Variability analysis

The obtained partial sequence information of SBV was compared to the sequence of isolate BH80/11-4 (accession codes HE649912-HE649914). Single nucleotide polymorphisms (SNP's) were identified using SeqMan Pro version 9 (DNASTAR, Madison, WI, USA) and are listed in Table 3. Only sequence differences where we had at least 2 sequence reads were included. Observed sequence variants at a position with more than 2 reads and/or with reads in both cDNA and complementary sense and/or with reads having different start positions were assigned a high support (Table 3).

Results and Discussion

To test the feasibility of virus identification using DNase SISPA-next generation sequencing (NGS) and to get a first estimate of its sensitivity on field samples, we selected both strong and weak positive SBV infected field samples. For viral RNA quantification, in vitro transcribed RNA was used as a standard curve in the previously described L gene real time RT-PCR. The curve showed a linear range at least from 2.75 to 7.75 $\log_{10}$ RNA copies per µL, and a sensitivity of less than 2.75 $\log_{10}$ RNA copies per µL (Figure 1).

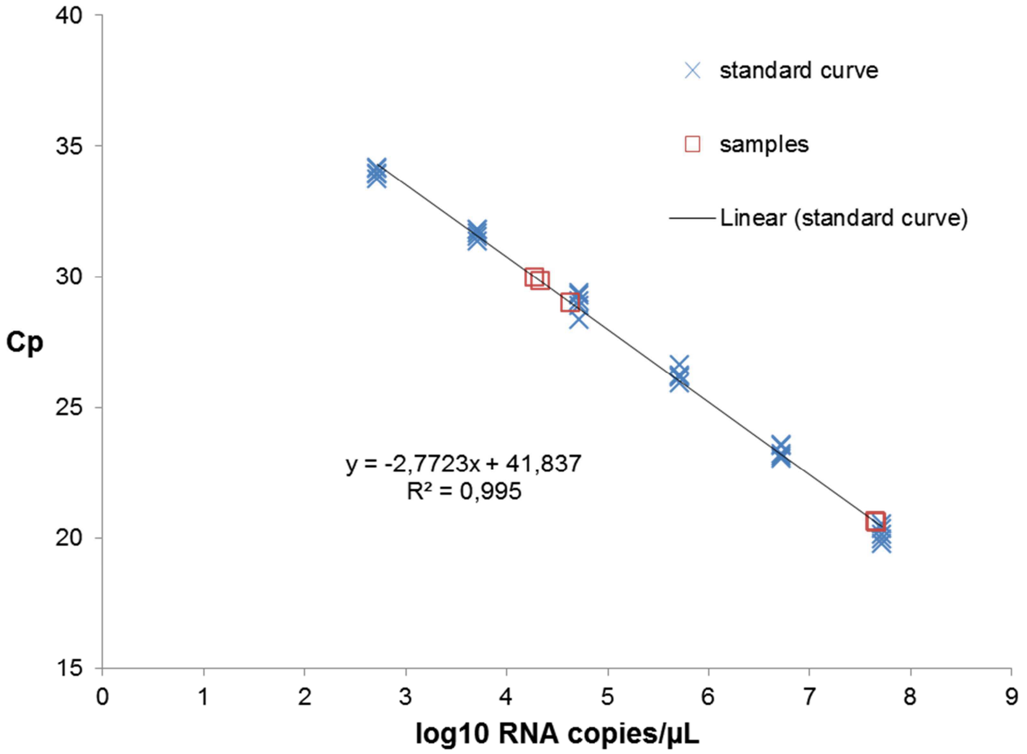


Figure 1: Quantification of Schmallenberg virus L segment RNA using quantitative real time RT-PCR. The viral RNA load in field samples was quantified with real time RT-PCR using a standard curve consisting of in vitro transcribed L segment RNA spanning the diagnostic RT-qPCR, which was run in five replicates (blue crosses). The linear trendline and the associated standard curve equation are displayed. The samples are indicated by red squares.

Two field samples of aborted lambs with Cp values of 20.59 and 20.65 corresponding to 7.66 and 7.64 $\log_{10}$ RNA copies per µL (Table 1) allowed unambiguous identification of Schmallenberg virus. One sample (BE/12-2478) yielded 2 S segment sequences (covering about 13 % of the S segment), 81 M segment sequences (covering about 56 % of the M segment) and 109 L segment sequences (covering about 81 % of the L segment) (Table 1, Figure 2). The other strong positive sample (BE/12-2068) resulted in a total of 8 SBV specific sequence reads distributed over the L and M genomic segments. This difference in sequence coverage for two samples with a comparable SBV RNA load is most likely due to the difference in the amount of raw sequence data (Table 1). While the sequencing of BE/12-2478 yielded about 95,000 reads, BE/12-2068 only resulted in circa 25,000 reads probably due to DNA library quantification issues at the sequencing facility. Moreover, more tissue sample was available for DNase SISPA protocol from sample BE/12-2478 (Table 1), although the ratio of viral reads to total raw reads was superior (0.01) in sample BE/12-2068 compared to sample BE/12-2478 (0.002 ; Table 2).

Table 1: SBV virus quantification and confirmation by DNase SISPA-NGS in selected field samples from Belgium.

Sample (location, date)	**Description**	**SBV log10 RNA copies/µl (Cp value)**	**Total no. reads**	**No. SBV reads per segment (% of RNA segment covered)**		
				S	**M**	**L**
BE/12-2068 (Ghoy, 13.01.2012)	Brain tissue, 180 mg, aborted lamb	7.664 (20.59)	25,701	-	1 (8.59)	6 (8.69)
BE/12-2478 (Deinze, 18.01.2012)	Brain tissue, 1000 mg, aborted lamb	7.642 (20.65)	94,722	2(13.68)	81(59.1)	109(81.9)
BE/12-2649 (Reningelst, 22.01.2012)	Brain tissue, 1130 mg, aborted lamb	4.331 (29.83)	50,308	-	-	-
BE/12-1235 (Sivry, 06.01.2012)	Brain tissue homogenate, 1.8 ml, aborted calf	4.630 (29.00)	27,979	-	-	-
BE/12-3610 (Izenberge, 28.01.2012)	Brain tissue, 620 mg, aborted lamb	4.270 (30.00)	98,648	-	-	-
Isolate BE/12-2068 (Ghoy, 13.01.2012)	Tissue culture supernatant. 5 $\log_{10}$ TCID50/ml	8.418 (18.50)	nd	nd	nd	nd

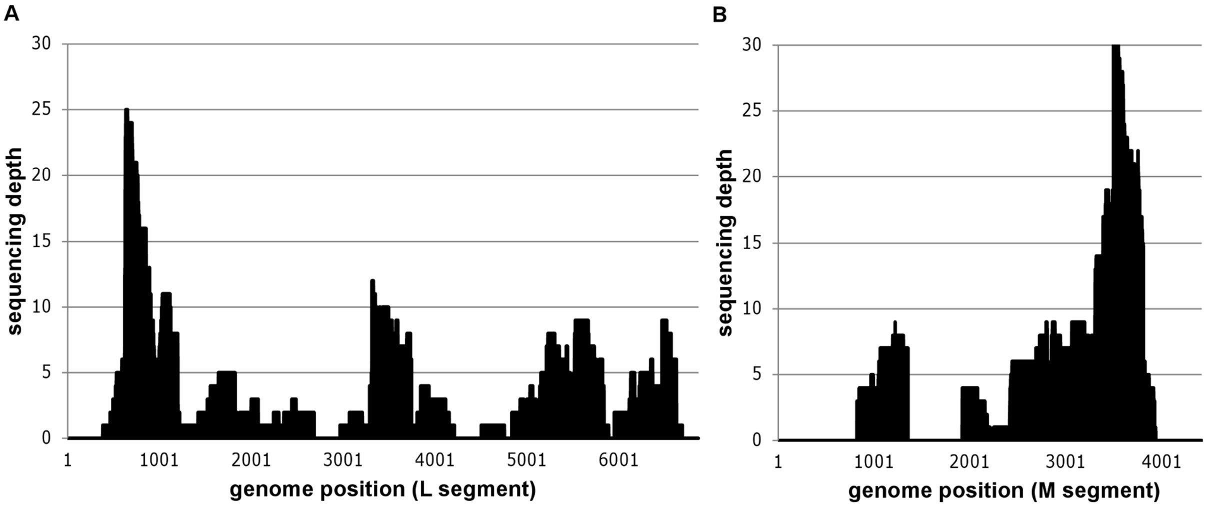


Figure 2: Mapping of Schmallenberg virus specific reads of sample BE/12-2479 against the German isolate BH80/11-4. Positional sequence coverage (number of sequence reads for given nucleotide position) of the M and L segments of sample BE/12-2478, based on reference assembly to HE649912 and HE649913.

Table 2: Output of the metagenomic analysis on raw sequence data from the sequencing libraries from SBV-positive samples.

Sample	**Total no. reads**	**Reads with primer tag identification**	**No. reads classified into superkingdom**				**No. unclassified reads**
			Eukaryota	**Archaea**	**Bacteria**	**Viruses**	
BE/12-2068	25,701	23,370	2,270	5	15,543	284 (*Phycodnaviridae, Myoviridae, Siphoviridae, Podoviridae, Bunyaviridae, Mimiviridae*)	3,977
BE/12-2478	94,722	86,178	36,690	11	36,106	167 (*Myoviridae, Siphoviridae, Podoviridae, Bunyaviridae*)	5,543
BE/12-2649	50,308	46,181	6,730	-	36,044	8 (*Siphoviridae, Podoviridae, Mimiviridae*)	2,468
BE/12-1235	27,979	26,214	13,216	5	17,89	1 (*Siphoviridae*)	9,812
BE/12-3610	98,648	91,458	12,403	13	69,057	15 (*Myoviridae, Podoviridae*)	5,053

Three weak positive field samples (one from an aborted calf, two from aborted lambs) containing virus quantities equivalent to 4.27-4.63 $\log_{10}$ RNA copies per µL did not allow identification using DNase SISPA-NGS (Table 1). This is consistent with an approximate sensitivity of 10^4-10^6 virions per ml estimated in previous studies using in vitro virus dilutions [12] or tissue biopsy samples [13]. However, it should be noted that the exact ratio between viral RNA quantities and intact virion quantities in field samples (the intact virions being detected in this method) remains to be determined. Our preliminary data indicate that RNA extracted from an SBV isolate containing 10^5 TCID50/ml may contain up to 8.46 $\log_{10}$ RNA copies per µL (Table 1). This indicates that precaution should be taken in interpreting RNA quantities in terms of DNase SISPA-NGS sensitivity, which is determined by the amount of viral nucleic acids that remain protected in intact virions during nuclease treatment. Moreover, a comparison of approximate sensitivity with other viral discovery methods is almost impossible, as the utilized sequencing effort varies from a few 100 Sanger sequencing reads [12, 14] to about 30 million Illumina GAII reads [13]; and different sample types (targeted tissue selection, freshness of the sample) may result in different levels of host and contaminating nucleic acids. Given the limited approximate sensitivity of DNase SISPA-

NGS, as any virus discovery method, careful selection of properly targeted and fresh field samples is necessary.

As expected in any metagenomic approach, a considerable part of the sequence reads represented diverse bacterial species and host nucleic acids (Table 2). It should be noted that the field samples were stored for a considerable time before the pretreatment and RNA extraction, during which opportunistic bacteria probably grew in the samples. Although we use a 0.22 µm filter to remove remaining cell fragments and bacteria, nucleic acids from disrupted cells can pass through the filter.

Other viral reads could be mainly identified as bacteriophages belonging to the families *Myoviridae*, *Siphoviridae* and *Podoviridae* (Table 2). Three reads showed partial similarity to a virus belonging to the *Phycodnaviridae* and two reads showed some similarity to viruses of the *Mimiviridae* family. None of these viruses have relatives known to infect animals. These sequences most likely represent contamination of the tissue samples during storage until analysis.

Compared to the metagenomic approach used by Hoffmann and colleagues [1] that initially identified this novel Orthobunyavirus by shotgun sequencing of total RNA extracted from clinical samples, our virus discovery protocol attempts an enrichment in viral nucleic acids by selective filtration and nuclease treatment. A direct comparison between both data sets is impossible as we treated limited amounts of tissue samples representing a different host species. Moreover, the sequencing effort per library was not identical. Both studies indicate the need of high sequence throughput and proper sample selection as critical factors for successful virus discovery using metagenomics.

The partial SBV sequence info we obtained was compared to the whole genome sequence that was determined from the virus originally isolated from diseased bovines in Germany. Several coding and noncoding mutations could be observed (Table 3). The partial data from the two Belgian ovine field samples showed together 16 nucleotide differences (of which 9 well-supported by the sequence data, Table 3) corresponding to 9 amino acid differences (of which 5 well-supported). Although this can be expected for an RNA virus that has now shown a distribution throughout a large part of Western Europe, this is to our knowledge the first documentation of genetic diversity within the Schmallenberg virus outbreak. Based upon the 8134 nucleotides in common between our partial sequence (BE/12/2478) and the genome of the virus isolated from diseased bovine in Germany (accession codes HE649912-HE649914), a mutation frequency of 1.7 10^{-3} mutations per site was observed. The time frame between the two samples was about three months and the geographical distance between the sampling sites roughly 325 km. Although only based on a three month period, the observed mutation frequency is within the documented range of Bunyavirus variability (10^{-2} to 10^{-4} mutations/site/year documented for Hantavirus ; [15]) and within the documented range of RNA virus variability (10^{-3} to 10^{-4} mutations/site/year ; [16, 17]). It should be noted that, while the samples in Germany were taken from acutely infected bovines, the ovine samples from Belgium present aborted lambs, making an estimation of the infection time of the maternal animal impossible. Future targeted molecular epidemiological studies including samples from the complete geographic range of the virus may shed light on the origin and time of introduction of this novel virus in Europe.

Our data show that DNase SISPA-NGS viral discovery technology can be used on limited amounts of field tissue samples to identify emerging diseases. However, the sensitivity of the method seems to limit its applicability to samples containing about 10^4 to 10^6 virions per ml. Consequently, when applying this methodology to a cluster of cases of an undiagnosed disease, it is important to select properly targeted and fresh samples as well as to test multiple diseased animals to allow correct identification of an associated virus.

Table 3: Differences observed in Belgian SBV sequences in comparison with the German genome sequence of isolate BH80/11-4.

Strain, genome segment	Covered regions*	Number of reads	Differences observed in Belgian sequences compared to the genome of the German isolate BH80/11-4			
			Nucleic acid	Amino acid	Depth	Support°
BE/12-2068, M segment	1,095-1,456	1				
BE/12-2068, L segment	1,271-1,426	2				
	3,313-3,744	4	G 3,490 C	E 1,159 Q	4	Low
			G 3,637 A	A 1,208 T	4	Low
BE/12-2478, S segment	428-467	1				
	669-733	1				
BE/12-2478, M segment	823-1,354	12	G 836 A	S 275 S	3	Low
			A 983 G	T 324 T	5	High
			C 998 T	F 329 F	4	High
			A 1,041 G	K 344 E	4	High
			T 1,201 C	F 397 S	8	High
	1,920-2,211	4	G 1,930 A	R 640 Q	4	Low
			A 1,969 T	Q 653 L	4	Low
	2,248-2,390	1				
	2,412-3,935	64	A 3,558 G	N 1,183 D	29	High
BE/12-2478, L segment	383-1,844	44	G 1,017 A	E 334 E	9	High
	1,873-2,682	6				
	2,966-3,744	19	C 3,097 T	H 1,028 Y	2	High
	3,770-4,209	4	C 3,937 T	H 1,308 Y	3	High
	4,502-4,754	1				
	4,835-5,888	21	T 5,736 A	P 1,907 P	6	High
	5,950-6,690	13	C 6,045 T	P 2,010 P	2	Low
			T 6,156 C	F 2,047 F	5	Low

* the indicated position is relative to the position of the used reference sequence: Schmallenberg virus, isolate BH80/11-4 (Genbank: HE649912, HE649913, HE649914).

° assessment based on low/high depth, single/double orientation of the reads, equal/different starting and end position of the reads, and quality of reads.

Acknowledgments

We thank Dr. Martin Beer and Dr. Bernd Hoffmann (FLI) for providing the SBV real time RT-PCR protocol and control reagents. We are grateful to Orkun Ozhelvaci for excellent technical assistance. The professional services of the Genomics Core of the University Hospital of Leuven are appreciated.

References

1. Hoffmann, B., et al., *Novel orthobunyavirus in cattle, europe, 2011.* Emerg Infect Dis, 2012. **18**(3): p. 469-72.
2. Muskens, J., et al., *[Diarrhea and loss of production on Dutch dairy farms caused by the Schmallenberg virus]*. Tijdschr Diergeneeskd, 2012. **137**(2): p. 112-5.
3. van den Brom, R., et al., *Epizootic of ovine congenital malformations associated with Schmallenberg virus infection.* Tijdschr Diergeneeskd, 2012. **137**(2): p. 106-11.
4. Bilk, S., et al., *Organ distribution of Schmallenberg virus RNA in malformed newborns.* Vet Microbiol, 2012.
5. ProMED-Mail (2012) *Schmallenberg virus - Europe: update, international impact.* . http://www.promedmail.com (accessed 2012 March 29) **archive no. 20120324.1079633**.
6. ProMED-mail (2012) *Schmallenberg virus - Europe: vector, morphology* http://www.promedmail.com, archive no.20120311.1066949 **2012**.
7. Allander, T., et al., *A virus discovery method incorporating DNase treatment and its application to the identification of two bovine parvovirus species.* Proc Natl Acad Sci U S A, 2001. **98**(20): p. 11609-14.
8. Rosseel, T., et al., *Identification and complete genome sequencing of paramyxoviruses in mallard ducks (Anas platyrhynchos) using random access amplification and next generation sequencing technologies.* Virol J, 2011. **8**: p. 463.
9. Van Borm, S., et al., *Phylogeographic analysis of avian influenza viruses isolated from Charadriiformes in Belgium confirms intercontinental reassortment in gulls.* Arch Virol, 2012. **In Press**.
10. Margulies, M., et al., *Genome sequencing in microfabricated high-density picolitre reactors.* Nature, 2005. **437**(7057): p. 376-80.
11. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
12. Djikeng, A., et al., *Viral genome sequencing by random priming methods.* BMC Genomics, 2008. **9**: p. 5.
13. Daly, G.M., et al., *A viral discovery methodology for clinical biopsy samples utilising massively parallel next generation sequencing.* PLoS One, 2011. **6**(12): p. e28879.
14. Allander, T., et al., *Cloning of a human parvovirus by molecular screening of respiratory tract samples.* Proc Natl Acad Sci U S A, 2005. **102**(36): p. 12891-6.
15. Ramsden, C., et al., *High rates of molecular evolution in hantaviruses.* Mol Biol Evol, 2008. **25**(7): p. 1488-92.
16. Jenkins, G.M., et al., *Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis.* J Mol Evol, 2002. **54**(2): p. 156-65.
17. Hanada, K., Y. Suzuki, and T. Gojobori, *A large variation in the rates of synonymous substitution for RNA viruses and its relationship to a diversity of viral infection and transmission modes.* Mol Biol Evol, 2004. **21**(6): p. 1074-80.

CHAPTER 3.6

False-positive results in metagenomic virus discovery, a strong case for follow-up diagnosis

Toon Rosseel, Bart Pardon, Kris De Clercq, Orkun Ozhelvaci and Steven Van Borm

Published in Transboundary and Emerging Diseases (2014) 61(4): p. 293-9.

After gaining experience with the viral discovery protocol, the time had come to test the workflow on a case study where presence and type of virus were unknown. In an attempt to increase the virus discovery sensitivity, the protocol was slightly changed when compared to that used in previous Chapters. To be sure not to exclude identification of large viruses, 0.45 µm pore size filters were used instead of 0.22 µm. In the nuclease treatment step, RNases were added in addition to DNase in order to remove non-particle protected RNA from the sample. The composition of the nuclease mixture was optimized by comparing different combinations of commercial DNases and RNases on a model DNA/RNA extract, thereby measuring the nucleic acid concentrations before and after treatment (data not shown). The workflow was divided into a RNA and DNA virus discovery workflow. The RNA and DNA extracts were treated with DNase and RNase enzymes respectively to remove remaining DNA or RNA.

During the preceding studies, we often noticed that a large part of contaminating background nucleic acids consisted of host and bacterial ribosomal RNA (rRNA). Preliminary testing with a rRNA depletion kit indicated a good removal of rRNA, thereby enhancing virus discovery sensitivity (data not shown). This led to addition of an rRNA removal step in the RNA virus discovery workflow.

A different reverse transcriptase enzyme and a different random SISPA primer concentration were used. Complementary DNA was also purified in order to remove remaining random SISPA primers to be sure they could not interfere with the subsequent PCR reaction.

Finally, bioinformatics data analysis was optimized to allow quick and reliable classification of contigs and singletons. BLAST databases were installed locally on a powerful computer server, and a dedicated software tool was used to process and visualize BLAST output files.

As a case study, a cluster of diseased dairy cattle with undiagnosed etiology was selected. We could not associate any viral sequences with the disease. However, our careful follow-up approach provided clear warnings for the potential of false positive results in metagenomics.

Abstract

A viral metagenomic approach using virion enrichment, random amplification and next-generation sequencing was used to investigate an undiagnosed cluster of dairy cattle presenting with high persistent fever, unresponsive to anti-microbial and anti-inflammatory treatment, diarrhoea and redness of nose and teat. Serum and whole blood samples were taken in the predicted hyperviraemic state of an animal that a few days later presented with these clinical signs. Bioinformatics analysis of the resulting data from the DNA virus identification workflow (a total of 32,757 sequences with average read length 335 bases) initially demonstrated the presence of parvovirus-like sequences in the tested blood sample. Thorough follow-up using specific real-time RT-PCR assays targeting the detected sequence fragments confirmed the presence of these sequences in the original sample as well as in a sample of an additional animal, but a contamination with an identical genetic signature in negative extraction controls was demonstrated. Further investigation using an alternative extraction method identified a contamination of the originally used Qiagen extraction columns with parvovirus-like nucleic acids or virus particles. Although we did not find any relevant virus that could be associated with the disease, these observations clearly illustrate the importance of using a proper control strategy and follow-up diagnostic tests in any viral metagenomic study.

Keywords: Contamination, next generation sequencing, parvovirus, viral metagenomics

Introduction

Next Generation Sequencing technologies' (NGS) ability to produce a spectacular number of sequence reads from a nucleic acid sample resulted in an exponential increase of characterized pathogens over recent years (reviewed in [1-3]). In veterinary virology, virion enrichment followed by random amplification [4, 5] and NGS has resulted in the characterization of an increasing number of viral nucleic acid sequences from outbreaks of unknown etiology [6, 7], well-known disorders presumed to be of multifactorial etiology [8-10] and reservoir species and vectors [11-15].

In Belgium, since 2009, over 50 dairy herds had been confronted with a sudden decrease in milk drop, with or without other clinical signs. Despite, that in several herds, pathogens such as bovine viral diarrhea virus (BVDV), *Anaplasma phagocytophilum*, *Leptospira hardjo* or Schmallenbergvirus had been identified (besides to nutritional issues), several outbreaks remained without etiology. In July 2013, again 4 outbreaks of milk drop were reported in Flanders, and this time typically accompanied by high fever (40-41°C), unresponsive to antimicrobial or anti-inflammatory treatment, characteristic nose lesions ("red nose"), and vasculitis-suggesting teat injuries (redness, necrosis, edema). Given the negative test results for known pathogens with a mucosal distribution, a metagenomic virus discovery workflow was used in an attempt to identify whether a viral pathogen was associated with these symptoms.

Materials and Methods

Herd anamnesis, clinical signs and diagnostics

In July 2013, in an 81 head mixed herd (Holstein Friesian and Belgian Blue beef cattle), in one week five dairy cows and one dairy heifer developed anorexia and milk drop. The animals were between a few days up to one month in lactation. Next to milk drop, the animals showed a marked reddening of the nasal ("red nose") and vaginal mucosae, teat and in some cases the interdigital space. All animals demonstrated high fever (40-41°C), which was unresponsive to treatment with non-steroidal anti-inflammatory drugs or corticosteroids. Fever could last up to 14 days. Four of the five initially ill animals also demonstrated moderate to severe diarrhea, which was bloody in one animal. Two of them showed marked tachypnea. In contrast, Belgian Blue animals and young stock were unaffected.

Two animals became recumbent and died after two weeks. The other animals needed more than six weeks for full recovery. At necropsy (one animal) the redness of nose, teat and vagina was no longer visible, suggesting a vascular (hyperemia) origin. Histologically, there was only a mild perivascular infiltration of lymphocytes and plasma cells. Other lesions were a pyometra (*Escherichia coli* and *Trueperella pyogenes*), focal ulcerative glossitis, reactive splenitis, udder abscess (*Escherichia coli*) and a multifocal ulcerative abomasitis, with histologically intralesional mycosis. It was concluded that the animal died from a secondary septicemia. No signs of vasculitis were seen. Case animals tested negative for BVDV, malignant catarrhal fever, bluetongue, epizootic haemorrhagic disease virus, Schmallenbergvirus by PCR and *A. phagocytophilum*. A gE-specific antibody ELISA for bovine herpesvirus 1 was negative as well.

Sampling and pretreatment

After these negative results, the herd was daily monitored for the appearance of new cases. Rectal temperature was taken daily. Whole blood (with EDTA as anticoagulant) and serum were taken from a cow on the 2nd day of high fever (>40 °C), aiming to sample a hyperviremic state should a viremic agent be involved, and immediately frozen at -20°C. A severe neutropenia (0.27 x 10^9 cells/ml (reference: 0.68-6.94)), with left shift was present at that time. Three days later a red discoloration of nose and teat appeared which was maximal at day 8. In addition, blood and serum samples were stored from a diseased animal from the first clinical series, which were sampled on the 10th day of high fever. Sample pretreatment and sequence independent single primer amplification (SISPA) was performed on the blood sample of the first animal similar as described before [16], with multiple modifications to increase viral identification sensitivity. Briefly, after a centrifugation step (45 min at 4,000×g at 4°C) the supernatants was filtered through 0.45 µm filters (Ultrafree-MC HV sterile, Millipore) at 4,000 ×g. Then the workflow was split for RNA virus and DNA virus identification. Before nucleic acid extraction, non-particle protected RNA and DNA were digested for 1 hour at 37°C with a mixture of filtrate (100 µl in RNA virus workflow; 134µl in DNA virus workflow), TURBO DNase (2 U/µl, Ambion; 50 U in RNA virus workflow; 68 U in DNA virus workflow), 1× TURBO DNase Buffer and RiboShredder RNase Blend (1 U/µl, Epicentre; 10 U in RNA virus workflow; 13 U in DNA virus workflow).

RNA virus workflow

RNA extraction was performed with the QIAamp Viral RNA Mini Kit (Qiagen, Venlo, the Netherlands) according the manufacturer's instructions. Low binding tubes were used for storing the extract or performing further treatments. Forty units of Protector RNase inhibitor (Roche) were added to the eluted RNA, and residual DNA was subsequently digested with 2.8 U of TURBO DNase at 37 °C (15 min). Before heat inactivation at 75 °C (for 15min), EDTA was added to a final concentration of 15mM to protect the RNA from chemical scission. RNA concentration was measured with the Qubit RNA Assay kit (Molecular Probes). Subsequently a ribosomal RNA (rRNA) removal step was performed with the Ribo-Zero Magnetic Gold Epidemiology kit (Epicentre) using the low input protocol described in the manual of the ScriptSeq Complete Gold Epidemiology Kit (Epicentre). First strand cDNA synthesis was subsequently done with the SuperScript III First-Strand Synthesis System (Invitrogen) using 10 µl RNA and 3 µM (final concentration) of FR26RV-N primer (5'-GCC GGA GCT CTG CAG ATA TCN NNN NN-3'). An initial denaturation was done at 95 °C for 5 min and placed on ice for at least 1 minute. After adding the cDNA synthesis mix, the mixture was incubated at 25°C for 10 min, 50 °C for 50 min and 5 min at 85 °C. After 2 min incubation on ice, an RNaseH treatment was performed according to the manual instructions. Second strand cDNA synthesis was accomplished by adding 2.5 U of 3'-5' exo$^-$ Klenow Fragment of DNA polymerase (New England Biolabs) and incubation for one hour at 37 °C. After having inactivated the enzymes at 75 °C for 20 min, the cDNA was purified with Agencourt AMPure XP beads (1.8:1 bead:DNA ratio; Beckman Coulter) to remove residual primers and eluted in 20 µl nuclease free water.

DNA virus workflow

DNA extraction was performed with the QIAamp DNA Blood Mini Kit (Qiagen) according to manufacturer's instructions. The extracted DNA was treated for 10 min at 37 °C with 4 U RiboShredder RNase Blend (Epicentre) to remove residual RNA. The extracted DNA was amplified as described in Allander et al., 2005 [4] with modified FR26RV-N primer concentration (1 µM final concentration). After the second denaturation, binding and elongation cycle the polymerase was heat inactivated at 75 °C for 20 min.

Random amplification and NGS

A subsequent PCR was performed using 5 µl amplified DNA (respectively cDNA) as descripted before [16] with minor modifications (1.2 µM final concentration FR20RV primer 5’-GCC GGA GCT CTG CAG ATA TC-3’ and 51.8 °C annealing temperature). The random amplified fragments were size selected on 1% agarose gel and 300-800 bp fragments were excised and purified from the gel with the High Pure PCR Product Purification Kit (Roche). The purified PCR fragments were quantified by spectrophotometry (Nanodrop-1000).

Sequencing libraries for the Genome Sequencer Junior (GS Junior, Roche) were prepared from the purified and size selected random amplified DNA according to the manufacturer’s instructions for Titanium Series reagents making use of multiplex identifiers (MID) to identify the different libraries. The resulting libraries were sequenced with a GS Junior with Titanium Series reagents and run protocol (200 cycles).

Metagenomic data analysis

The sequence output files were sorted per sequencing library according their MID sequences. The PCR primer sequence was trimmed off the reads and *de novo* assembly was performed using the GS *De Novo* Assembler software v2.7 with default parameters. The resulting contigs (contiguous sequences assembled from overlapping sequence reads) and singletons (single sequences) were subjected to a BLAST (Basic Local Alignment Search Tool) sequence alignment analysis using locally installed blast-2.2.27+ software [17]. First, a megablast search was carried out. If no result (hits) was found, we proceeded to a blastn search, followed by a blastx and ultimately a tblastx search if no hits were found in the preceding BLAST analyses. The E-value cutoff was set at 0.001. BLAST output files were subsequently processed with the MEtaGenome ANalyzer software (MEGAN v4.70.4 , [18]) to classify the reads according to the NCBI taxonomy. As lowest common ancestor (LCA) parameters default values were used, except for the Min Support (1) and Min Complexity (0.3) parameters.

Follow-up SYBR green real-time PCR

Relevant singletons* were subjected to a *de novo assembly with* Lasergene SeqMan pro

* Singletons with similarity to viral sequences

(DNASTAR) in order to find overlapping regions. Specific SYBR green real-time PCR assays were designed targeting relevant contig or singleton sequences using Primer-BLAST [19] and the reactions were performed with the LightCycler 480 SYBR Green I Master kit (Roche) using 1 µM final concentration of each primer and an annealing temperature of 52 °C. A melting curve analysis was included to verify the specificity of the PCR reaction. PCR's not yielding a specific melting curve profile were abandoned, resulting in 3 reliable and specific real-time SYBR green PCR tests targeting Parvovirus-like contigs and singletons, and 1 test targeting 2 overlapping circovirus-like singletons [supplementary Table S1: primer sequences). DNA was re-extracted (QIAamp DNA Blood Mini Kit (Qiagen)] from the original samples and tested to confirm the presence of Parvovirus-like nucleic acids.

Follow-up nucleic acid extraction

Three negative control extractions were performed with the QIAamp Viral RNA Mini Kit (Qiagen) using only elution buffer. As an alternative extraction method the NucleoSpin RNA Virus kit (Macherey-Nagel) was used according to the manufacturers' instructions.

Results and Discussion

As we have emphasized previously [20], the selection of properly targeted and fresh (field) samples is of great importance in any virus discovery method. The sample that was subjected to our optimized metagenomic virus discovery workflow was therefore selected with great care in order to increase the probability to detect viruses. The DNA virus identification workflow yielded in total 32,757 sequences (average length of 335 bases). A *de novo* assembly resulted in 1,463 contigs (comprising 16,993 reads) and 12,421 singletons. Table 1 summarizes the BLAST analysis results. A very low number of host genome sequences indicated a successful removal of host nucleic acids by the nuclease treatments. The BLAST analysis could identify 56 contigs and 450 singletons similar to viruses. Most viral contigs/singletons were similar to bacteriophage sequences. Five singletons showed a nearly exact match to the goatpox and lumpy skin disease virus genome. These sequences matched exactly the sequence of a goatpox virus strain of which we made a sequencing library in the lab 1 week earlier. As a specific pox qPCR tested negative on a new DNA extraction of the original sample (data not shown), we could exclude involvement of a poxvirus. Three contigs (comprising 25 reads) and 17 singletons were similar to marine virus sequences, mostly to regions associated with capsid or replication proteins. One contig (comprising 17 reads) and 9 singletons were similar (ranging from 68 % to 99 % identity when identified by megablast or blastn; ranging from 29 % to 50 % identity when identified by blastx) to the Parvovirus NIH-CQV (GenBank: KC617868.1). Another singleton was similar (82 % identity) to a densovirus which also belongs to the *parvoviridae*. Parvovirus NIH-CQV was discovered and sequenced by Xu et al. [21] using a similar metagenomic virus discovery approach. This study claimed this virus was highly prevalent in human patients with seronegative hepatitis and phylogenetic analysis indicated that the NIH-CQV was located at the interface of *Parvoviridae* and *Circoviridae*. The origin of this new virus was later questioned by Naccache and colleagues as a likely reagent contaminant [22-25]. Our blast analysis (Table 1) identified also 8 circovirus-like singletons (≥ 71% identity when identified by blastn; ≥ 43% identity when identified by blastx). To confirm the presence of these parvo/circovirus-like sequences in our sample we developed 4 qPCR tests based on relevant contigs and singletons. We did not develop qPCR tests for every singleton as multiple singletons represented similarity to the same viral (replication) protein. The tests confirmed the presence of these sequences in new DNA extractions of the original blood and serum sample, as well as in DNA extractions of a blood and serum sample from a second animal with the same characteristic symptoms (Table 2). Surprisingly, a negative extraction control (elution buffer extracted with QIAamp DNA Blood mini kit) also contained qPCR detectable Parvovirus-like sequences, while negative extraction controls using a different extraction kit (NucleoSpin, Macherey-Nagel) tested negative (Table 2). After our initial metagenomic data analysis, new Parvovirus-like hybrid virus (PHV) sequences were submitted in GenBank (accession numbers KF170373 and KF214637 to KF214647) by Naccache et al. [23]. The authors used a similar NGS strategy and discovered a novel highly divergent DNA virus that was at the interface between the *Parvoviridae* and *Circoviridae*. The genome sequence was nearly identical to the NIH-CQV sequence. They demonstrated the origin of the PHV DNA to be contaminated silica-binding spin columns used for the nucleic acid extraction (RNeasy MinElute, RNeasy Mini, QIAamp MiniElute and QIAampMini spin columns). The authors detected also PHV sequences in environmental metagenomic datasets of coastal marine waters of North America. They suggested that a potential association between the hybrid virus and algae (diatoms), that generated the silica matrix used in the spin columns, may have resulted in unintended viral contamination during manufacture. Our findings confirm the risk of reagent contamination stressed by Naccache et al. [23] and are also in accordance with Lysholm et al. [26] where the authors used a similar metagenomic sequencing strategy to characterize viruses in human patients with severe lower respiratory tract infection. Lysholm et al. could link detected densovirus-like and circovirus-like contigs to the use of the QIAamp DNA Blood Mini Kit (Qiagen). The remaining identified viral sequences in our study (Herpesvirus-like, Mimivirus-like; Table 1) were low complexity reads or showed only limited similarity to the blast hits. Low complexity sequences frequently occur in metagenomic datasets and are most likely artifactual amplification products. The RNA virus identification workflow yielded in total 31,365 sequences with an average length of 323 bases. *De novo* assembly resulted in 648 contigs (comprising 26,019 reads) and 1,088 singletons. BLAST analysis (Table 1) indicated again a successful removal of host nucleic acids by the nuclease treatments as almost no sequences were similar to the host genome. Eighteen contigs and 26 singletons were similar to virus sequences, again mostly bacteriophages and marine virus-like sequences. As in the DNA virus workflow, we identified 6 goatpox and lumpy skin disease virus-like sequences. The fact that we found these DNA virus sequences also in the RNA virus discovery workflow (where a DNase treatment was performed on the RNA extract), confirms the likelihood of contamination after amplification. Three singletons were similar to primate retroviruses and genomic sequences, unlikely to be involved in bovine pathologies. These sequences were found in the root of the MEGAN analysis, indicating that BLAST results with both eukaryotic and viral genome taxonomy were found (most likely due to un-annotated retroviral sequences in mammalian genomic sequences in public sequence databases).

In conclusion, using a metagenomic virus discovery workflow, we did not find any indication for the involvement of a viral agent in the disease of these dairy cattle. However, this case study clearly demonstrates the importance of potential contamination at different levels in metagenomic workflows. Contamination can occur for instance at the time of sampling, during the laboratory work (cf. the goatpox virus-like sequences in our dataset), reagent contamination (cf. parvo/circo hybrid virus sequences), etc. The false positive results detected here, stress the importance of careful interpretation of metagenomic sequencing data and emphasize the importance of thorough follow-up diagnosis. We illustrated that PCR-based prevalence studies in matching disease cases and healthy controls can provide a tool to detect false positive metagenomic results. While such studies can provide further evidence for disease association [27], isolation of candidate pathogens is required to assign causality by addressing the Koch's postulates [28]. The assembled data from such a multidisciplinary approach (pathology, epidemiology, metagenomic data, PCR prevalence studies, isolation, characterization, …) should be used to identify the most likely candidate etiological agent and to make informed intervention decisions.

Table 1: Results of BLAST analysis (E-value threshold = 0.001) of contigs (≥ 100 bp) and singletons from a *de novo* assembly with GS *De Novo* Assembler.

		DNA virus identification workflow		**RNA virus identification workflow**	
		32,757 raw reads (10,984,762 bases) 31,308 trimmed reads (9,649,226 bases)		31,365 raw reads (10,116,997 bases) 28,742 trimmed reads (8,884,507 bases)	
Classification	**Virus family**	**Number of contigs (reads)**	**Number of singletons**	**Number of contigs (reads)**	**Number of singletons**
Archaea		4 (62)	27	2(56)	4
Bacteria		514 (6,036)	3,575	242 (9,402)	271
Eukaryota		285 (3,461)	1,642	222 (9,569)	312
Unclassified		604 (6,803)	6,727	164 (6,375)	475
Viruses		56 (631)	450	18 (617)	26
Bacteriophage-like	*Myoviridae, Podoviriae, Siphoviridae, Microviridae, …*	52 (608)	400	16 (575)	17
(uncultured) Marine virus -like	*Phycoddnaviridae, environmental samples*	3 (25)	17	2 (73)	
Parvovirus NIH-CQV-like, densovirus-like	*Parvoviridae*	1 (17)	10		
Circovirus-like (circovirus-like genome RW-E, beak and feather disease virus, ...)	*Circoviridae*		8		
Herpesvirus-like (ROOT)	*Herpesviridae*		6		
Goat pox/ lumpy skin disease/ sheep pox virus-like	*Poxviridae*		5		6
Mimivirus-like (ROOT, low complexity)	*Mimiviridae*		4		
Retrovirus-like (ROOT)	*Retroviridae*				3

Table 2: Cp (crossing point = the threshold cycle on a Roche LightCycler480 real-time PCR instrument) values of qPCR tests on new DNA extractions of blood and serum samples from two diseased animals and negative extraction controls. N = negative result.

	PCR contig 1 (parvo-like)	PCR contig 2 (parvo-like)	PCR contig 3 (parvo-like)	PCR contig 4 (circo-like)
Animal 1 – blood	33,96	30,19	40	N
Animal 1 – serum	32,62	31,43	39,96	35,11
Animal 2 – blood	34,76	31,6	40	34,87
Animal 2 – serum	33,88	31,79	40	36,86
Qiagen extraction control 1	33,44	31,18	40	32,22
Qiagen extraction control 2	33,19	30,9	N	36,57
Qiagen extraction control 3	34,19	28,85	36,87	32,71
NucleoSpin extraction control	N	N	N	N
Negative qPCR control	N	N	N	N

Supplementairy material

Table S1 – Primers used for the specific SYBR green PCR detection of contigs and singletons detected in the metagenomic dataset.

Primer	Sequence (5’-3’)	Contig classification	Length (n° of reads)
Contig1F	CGACATGGACGTGGAATCGT	parvovirus NIH-CQV-like contig	533 bp (17)
Contig1R	GGGAAGATCGAAGACGCTCG		
Contig2F	TACGCGGACTTTCCAGTTCC	2 overlapping parvovirus NIH-CQV-like singletons	328 bp (2)
Contig2R	TTCAAGAATACCCGGCGCCTT		
Contig3F	CCACATCTGGCTATCTGGCG	2 overlapping parvovirus NIH-CQV-like singletons	271 bp (2)
Contig3R	TGCCGTGATCCACGTCTTG		
Contig4F	GCCGGATCGAGTGTTGGATA	2 overlapping circovirus-like genome RW-E singletons	404 bp (2)
Contig4R	GGAGATATCGACGTGGCACA		

Acknowledgements

We are grateful Dr. M. Speeckaert for access to animals and permission to use samples. We would also like to thank Dr. Eric Delwart (Blood systems Research Institute, San Francisco, CA, USA) for helpful comments. This study was supported by internal research and development grant RandSeq11 of the Veterinary and Agrochemical Research Center, European Union FP7 project RAPIDIA-FIELD (FP7-289364), and Epi-SEQ: a transnational research project supported under the 2nd joint call for transnational research projects by EMIDA ERA-NET (FP7 project nr 219235).

References

1. Barzon, L., et al., *Applications of next-generation sequencing technologies to diagnostic virology.* Int J Mol Sci, 2011. **12**(11): p. 7861-84.
2. Belak, S., et al., *New viruses in veterinary medicine, detected by metagenomic approaches.* Vet Microbiol, 2013. **165**(1-2): p. 95-101.
3. Blomstrom, A.L., *Viral metagenomics as an emerging and powerful tool in veterinary medicine.* Vet Q, 2011. **31**(3): p. 107-14.
4. Allander, T., et al., *Cloning of a human parvovirus by molecular screening of respiratory tract samples.* Proc Natl Acad Sci U S A, 2005. **102**(36): p. 12891-6.
5. Djikeng, A., et al., *Viral genome sequencing by random priming methods.* BMC Genomics, 2008. **9**: p. 5.
6. Honkavuori, K.S., et al., *Novel picornavirus in Turkey poults with hepatitis, California, USA.* Emerg Infect Dis, 2011. **17**(3): p. 480-7.
7. Blomström, A.L., et al., *Detection of a novel astrovirus in brain tissue of mink suffering from shaking mink syndrome by use of viral metagenomics.* J Clin Microbiol, 2010. **48**(12): p. 4392-6.
8. Cox-Foster, D.L., et al., *A metagenomic survey of microbes in honey bee colony collapse disorder.* Science, 2007. **318**(5848): p. 283-7.
9. Blomström, A.L., et al., *Detection of a novel porcine boca-like virus in the background of porcine circovirus type 2 induced postweaning multisystemic wasting syndrome.* Virus Res, 2009. **146**(1-2): p. 125-9.
10. Granberg, F., et al., *Metagenomic detection of viral pathogens in Spanish honeybees: co-infection by Aphid Lethal Paralysis, Israel Acute Paralysis and Lake Sinai Viruses.* PLoS ONE, 2013. **8**(2): p. e57459.
11. Bishop-Lilly, K.A., et al., *Arbovirus detection in insect vectors by rapid, high-throughput pyrosequencing.* PLoS Negl Trop Dis, 2010. **4**(11): p. e878.
12. Li, L., et al., *Bat guano virome: predominance of dietary viruses from insects and plants plus novel mammalian viruses.* J Virol, 2010. **84**(14): p. 6955-65.
13. Blomstrom, A.L., et al., *Viral metagenomic analysis of bushpigs (Potamochoerus larvatus) in Uganda identifies novel variants of Porcine parvovirus 4 and Torque teno sus virus 1 and 2.* Virol J, 2012. **9**: p. 192.
14. Ge, X., et al., *Metagenomic analysis of viruses from bat fecal samples reveals many novel viruses in insectivorous bats in China.* J Virol, 2012. **86**(8): p. 4620-30.
15. Bodewes, R., et al., *Identification of multiple novel viruses, including a parvovirus and a hepevirus, in feces of red foxes.* J Virol, 2013. **87**(13): p. 7758-64.
16. Rosseel, T., et al., *Identification and complete genome sequencing of paramyxoviruses in mallard ducks (Anas platyrhynchos) using random access amplification and next generation sequencing technologies.* Virol J, 2011. **8**: p. 463.
17. Tao, T. *Standalone BLAST Setup for Windows PC. In: BLAST® Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2008-.* 2010 [Updated 2012]; Available from: http://www.ncbi.nlm.nih.gov/books/NBK52637/.
18. Huson, D.H., et al., *Integrative analysis of environmental sequences using MEGAN4.* Genome Res, 2011. **21**(9): p. 1552-60.
19. Ye, J., et al., *Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction.* BMC Bioinformatics, 2012. **13**: p. 134.
20. Rosseel, T., et al., *DNase SISPA-next generation sequencing confirms Schmallenberg virus in Belgian field samples and identifies genetic variation in Europe.* PLoS One, 2012. **7**(7): p. e41967.

21. Xu, B., et al., *Hybrid DNA virus in Chinese patients with seronegative hepatitis discovered by deep sequencing.* Proc Natl Acad Sci U S A, 2013. **110**(25): p. 10264-9.
22. Editorial, *Correction for Xu et al., Hybrid DNA virus in Chinese patients with seronegative hepatitis discovered by deep sequencing.* Proc Natl Acad Sci U S A, 2014.
23. Naccache, S.N., et al., *The perils of pathogen discovery: origin of a novel parvovirus-like hybrid genome traced to nucleic acid extraction spin columns.* J Virol, 2013. **87**(22): p. 11966-77.
24. Naccache, S.N., et al., *Concerns over the origin of NIH-CQV, a novel virus discovered in Chinese patients with seronegative hepatitis.* Proc Natl Acad Sci U S A, 2014.
25. Smuts, H., et al., *Novel hybrid parvovirus-like virus, NIH-CQV/PHV, contaminants in silica column-based nucleic acid extraction kits.* J Virol, 2014. **88**(2): p. 1398.
26. Lysholm, F., et al., *Characterization of the viral microbiome in patients with severe lower respiratory tract infections, using metagenomic sequencing.* PLoS One, 2012. **7**(2): p. e30875.
27. Mokili, J.L., F. Rohwer, and B.E. Dutilh, *Metagenomics and future perspectives in virus discovery.* Curr Opin Virol, 2012. **2**(1): p. 63-77.
28. Fredericks, D.N. and D.A. Relman, *Sequence-based identification of microbial pathogens: a reconsideration of Koch's postulates.* Clin Microbiol Rev, 1996. **9**(1): p. 18-33.

CHAPTER 4

Fine-tuning of methodology

This chapter tackles another objective of this thesis, namely the investigation of potential bias and added value of sample preparation and amplification methods. Moreover, the potential for direct sequencing without an extra pre-amplification step is evaluated.

CHAPTER 4.1

The origin of biased sequence depth in sequence-independent nucleic acid amplification and optimization for efficient massive parallel sequencing

Toon Rosseel, Steven Van Borm, Frank Vandenbussche, Bernd Hoffmann, Thierry van den Berg, Martin Beer and Dirk Höper

Published in PLoS ONE (2013) 8(9): p. e76144.

In this chapter, evidence is provided that the rPCR SISPA pre-amplification method results in an amplification bias. In this chapter rPCR SISPA is referred as just “SISPA”.

Abstract

Sequence independent single primer amplification (SISPA) is one of the most widely used random amplification approaches in virology for sequencing template preparation. This technique relies on oligonucleotides consisting of a 3’ random part used to prime complementary DNA synthesis and a 5’ defined tag sequence for subsequent amplification. Recently, this amplification method was combined with next generation sequencing to obtain viral sequences. However, these studies showed a biased distribution of the resulting sequence reads over the analyzed genomes. The aim of this study was to elucidate the mechanisms that lead to biased sequence depth when using random amplification. Avian paramyxovirus type 8 was used as a model RNA virus to investigate these mechanisms. We showed, based on *in silico* analysis of the sequence depth in relation to GC-content, predicted RNA secondary structure and sequence complementarity to the 3’ part of the tag sequence, that the tag sequence has the main contribution to the observed bias in sequence depth. We confirmed this finding experimentally using both fragmented and non-fragmented viral RNAs as well as primers differing in random oligomer length (6 or 12 nucleotides) and in the sequence of the amplification tag. The observed oligonucleotide annealing bias can be reduced by extending the random oligomer sequence and by *in silico* combining sequence data from SISPA experiments using different 5’ defined tag sequences. These findings contribute to the optimization of random nucleic acid amplification protocols that are currently required for downstream applications such as viral metagenomics and microarray analysis.

Introduction

The determination of complete viral genome sequences is a growing field in human, animal, and plant virology. Complete genome sequences and their exponential growth in public databases (roughly 1.5 million sequences representing more than 100,000 viral taxa in GenBank at the moment of writing of this manuscript) not only allow for a better understanding of virus evolution, molecular phylogeny (phylogenomics) and epidemiology, but also facilitate functional analysis of virus genes in comparison with other sequences in databases. Traditionally, viral genome sequencing approaches are based on amplification of overlapping genome regions followed by Sanger sequencing [1]. As a result, efficient sequencing approaches rely very much on prior sequence knowledge and are often focused on specific groups of viruses to allow for robust design of amplification primers (e.g.[2]). Viral isolates from highly divergent families or less frequently studied viruses often require a cumbersome approach for genome completion, partly because of the lack of sufficient available sequence information for robust primer design, and partly because of frequent need for primer walking and redesigning primers.

Next generation sequencing (NGS) technologies were developed to accommodate the need of higher sequencing capacity and lower cost per nucleotide for large genome sequencing projects (e.g. [3], reviewed in [4]). One main advantage of NGS platforms is the possibility to sequence DNA samples without any prior knowledge of the sequence for priming [3]. However, virus samples are typically loaded with host and contaminating nucleic acids. Enrichment for nucleic acids of interest is thus needed before these technologies become useful. This enrichment is often established by a targeted amplification of viral nucleic acids using virus or taxon specific primers. Examples include streamlined sequencing protocols for influenza A viruses [5, 6], classical swine fever virus [7] and foot-and-mouth disease virus [8]. These protocols allow completion of the viral genome(s) in a single experiment and provide sufficient sequencing depth to analyze the variability of RNA virus populations in a single sample (e.g. [9, 10]).

Truly sequence independent access methods to viral genomes have been developed in the field of viral discovery (reviewed in [11-13]). One of the most prominent technologies for random access to viral nucleic acids is Sequence Independent Single Primer Amplification (SISPA), and was originally described by Reyes and Kim [14]. Several modifications have been published, some including enrichment steps for viral nucleic acids using filtration and nuclease treatment (DNase SISPA, [15, 16]). After a filtration step and nuclease treatment, nucleic acids protected within virion particles are purified. The random primers used in subsequent complementary DNA production have a fixed amplification tag which is used in downstream PCR amplification. The resulting random amplicons are cloned and selected clones from this library are sequenced. Although the method was developed as a tool for identification of unknown viruses, Djikeng and colleagues [16] demonstrated its potential use for full genome sequencing of different model genomes, albeit at a high sequencing effort (100's of colonies picked and sequenced for genome completion) and requiring a reasonable amount of virus (minimum 10^6 virus particles). This method was also applied to the partial sequencing of a novel paramyxovirus in penguins [17], influenza viruses [18] and the identification of unknown viruses from experimentally infected mice [19].

Recent studies have combined random priming approaches with NGS to obtain sequence information from viruses. These include the identification of a novel mink astrovirus [20], the metagenomic analysis of Dengue virus infected mosquitoes [21], metagenomic analyses of viruses in human stool samples [22], and the control of live-attenuated vaccines [23] and other biological products [24].

Careful examination of the sequence data obtained in these studies shows a lack of homogeneous distribution of randomly generated sequence reads over the target genome [16, 21-23, 25] exposing one limitation of these random access methods for the determination of complete viral genomes. This does not only lead to gaps and areas of low coverage, but also to areas of exaggerated sequence depth that may result in bioinformatic artifacts during sequence assembly, even with high sequencing efforts. Although these shortcomings were noted before, no systematic experimental analysis of this phenomenon was undertaken.

The aim of this study was to elucidate the mechanisms that lead to biased sequence depth when using random amplification. Moreover, we sought to use gained knowledge to improve random amplification methods, aiming for high quality viral genomes at a limited cost without prior sequence knowledge. Avian paramyxovirus type 8 was used as a model RNA virus to investigate these mechanisms.

Materials and Methods

Viral sample

Avian paramyxovirus type 8 virus (APMV-8) was kindly provided by the German reference laboratory for Newcastle disease of the Friedrich-Loeffler-Institut, Greifswald - Insel Riems, Germany. APMV-8 has a linear single stranded RNA genome of negative orientation with a length of 15,342 bases. Virus propagation was performed in 8-10 day old specific pathogen free embryonated chicken eggs. Allantoic fluid was collected and the virus titer was determined by hemagglutination assays according to the Council Directive 92/66/EC (1992) using 1% chicken erythrocytes.

Random access to viral nucleic acids using DNase SISPA

APMV-8 virions were purified starting from one milliliter (ml) of allantoic fluid. Centrifugation, filtration with 0.22 µm filters, nuclease treatment with 100 units DNase I, viral RNA extraction and sequence independent single primer amplification (SISPA) were performed as previously described [26]. Briefly, the RNA was denatured at 95 °C for five minutes in the presence of random SISPA primer FR20RV-6N (5’-GCCGGAGCTCTGCAGATATCNNNNNN-3’, [15]) which was used in the double stranded complementary DNA (cDNA) synthesis reaction. This primer was composed of a random 6N oligomer tagged with a known sequence which was subsequently used as PCR primer binding-extension sequence with complementary primer FR20RV (5’-GCCGGAGCTCTGCAGATATC-3’). Purified, size selected (400-1,200 nucleotides [nt]) random PCR fragments were quantified with the Nanodrop-1000 spectrophotometer and used for the preparation of 454 sequencing libraries as described below.

Optimization of DNase SISPA

The following modifications of the DNase SISPA protocol (on APMV-8) were performed to test whether annealing effects during cDNA synthesis and/or RNA secondary structures contributed to biased sequence depth.

RNA fragmentation. To test whether viral RNA secondary structures assisted in causing the unequal sequencing depth, we fragmented the extracted RNA according to the GS FLX Titanium cDNA Rapid Library Preparation Method Manual protocol (Roche, Mannheim, Germany, October 2009 Rev. Jan2010) starting from approximately 40 ng of viral RNA.

RNA size distribution before and after fragmentation was measured using the RNA 6000 Pico chip (Agilent, Böblingen, Germany) on the Agilent 2100 Bioanalyzer. The fragmented RNA was used in SISPA under identical reaction conditions as described above.

Effect of the Primer-tag sequence. To check whether the primer tag sequence (designed for downstream PCR amplification) had an influence on the binding of the random primer along the genome during cDNA synthesis, alternative primer sequences were tested during first and second strand cDNA synthesis (summarized in Table 1). A primer with an alternative PCR amplification tag (K-6N, 5’-GACCATCTAGCGACCTCCACNNNNNN-3’, modified from [27]) was tested to investigate whether other regions in the genome would be preferentially targeted compared to the original FR20RV-6N (5’-GCCGGAGCTCTGCAGATATCNNNNNN-3’) primer. Primer K (5’-GACCATCTAGCGACCTCCAC-3’, modified from [27]) was used for downstream PCR amplification under identical reaction conditions as for FR20RV. Additionally, a 12N random sequence version was tested for both tag sequences in comparison to the 6N random sequence primers (FR20RV-12N, 5’-GCCGGAGCTCTGCAGATATCNNNNNNNNNNNN-3’ and K-12N, 5’-GACCATCTAGCGACCTCCACNNNNNNNNNNNN-3’).

Sequencing

Purified, size selected (400-1,200 nt), random amplified DNA originating from the different random amplifications of APMV-8 was used to prepare sequencing libraries for the Genome Sequencer FLX (GS FLX; Roche, Mannheim, Germany). This was performed according to the manufacturer’s instructions for Titanium Series reagents, using multiplex identifiers (MID) to identify the different libraries. The resulting libraries were sequenced with a GS FLX with Titanium Series reagents and run protocol (200 cycles).

Data analysis

The sequence output file was sorted per sequencing library according their MID sequences. All raw sequence files were submitted to the Sequence Read Archive (SRA) under accession number SRP028373. Sequence reads were trimmed to remove the primer sequence including the random (6N or 12N) part as well as low quality ends. Non-APMV-8 specific reads were filtered out. Of each dataset approximately 7.7 Mb of raw data (≈ 500× theoretical genome wide sequencing depth) were randomly picked to allow direct comparison between all conditions. Reference guided assemblies were performed relative to APMV-8/pintail/Wakuya/20/78 (GenBank: FJ215864) using the GS Reference Mapper software (version 2.6; Roche, Mannheim, Germany). Data output files were further processed with R ([28]; http://www.r-project.org/). To investigate if the 5’ specific amplification tag of SISPA primer influenced the sequence depth distribution, we mapped short sections of the amplification tag of the SISPA primers of increasing length adjacent to the random part of the primer along the genome sequence using R. Consensus sequences from each of the amplification strategies were compared in a clustalW alignment. Variant analysis was performed by mapping of all raw sequencing reads (complete datasets) of a certain condition relative to reference sequence APMV-8/pintail/Wakuya/20/78 (GenBank: FJ215864) using SeqMan NGen® version 3.0 (DNASTAR, Madison, WI, USA). After mapping, the reference was deleted and single-nucleotide polymorphisms were called. At polymorphic positions, we included a degenerate nucleotide in the consensus sequence if the minor nucleotide alternative was present in at least 30 % of the sequence reads.

Modeling secondary structure

RNAfold from the Vienna RNA Package version 2.0 ([29], http://www.tbi.univie.ac.at/RNA/) was used to predict partition function and base pairing probability matrix (dot plot). The pair probabilities were extracted and plotted in a mountain plot (Perl script mountain.pl, http://www.tbi.univie.ac.at/RNA/utils.html). This plot represents the secondary structure in a plot of height (number of base pairs enclosing the base at a certain position, i.e. a measure of local secondary structure complexity) versus position. As our SISPA protocol used 95 °C to denature the RNA before cDNA synthesis and 50 °C at first strand cDNA synthesis, we modeled the minimum free energy secondary structure of the RNA genome sequence at these temperatures.

Positional genomic GC-content

Positional GC percentage was calculated with a sliding window of fixed size of 401 bp. The window was centered at a particular position and expanded 200 bp to either side of the center.

Analysis of the virus specificity of the protocol

In order to establish the virus specificity of the protocol, all sequence data generated for each library were classified according to the species they belong to. To this end, a combination of BLAST and the GS FLX software suite (v2.6; Roche) was used to sort the reads. Subsequently, the percentages of reads identified as viral sequences were calculated as a measure of the virus specificity.

Results

Standard DNase SISPA results in highly variable sequence depth

Using the 6N SISPA primer FR20RV-6N [16], the complete coding sequence of APMV-8 could be determined using a reference assembly with approximately 7.7 Mb of raw data. This 7.7 Mb of raw reads were randomly picked from the complete dataset and corresponds to about 500 x sequence depth under the assumption of even sequence depth along the genome (Table 1). Despite the median sequencing depth of 326.5 x, extreme variation in sequence depth (1 to 3,286 x) was observed (Table 1; Figure 1 A, repeat 1). 23 % of the genome nucleotides were covered less than 100 times, which we set as a minimum sequence depth to allow quantitative variant analysis (Table 1). An independent repetition starting from the same virus stock was made, producing a similar distribution pattern of sequencing depth (Figure 1 A, repeat 2). Apart from a lower maximum sequence depth (2,109 x), the coverage statistics were reproducible (Table 1). The resulting consensus sequence was submitted to GenBank under accession number JX901129.

Table 1: Summary of experimental conditions tested on APMV-8 and their assembly and coverage statistics.

Primer condition	RNA (F/N)*	Assembly statistics			Coverage statistics						
		n raw bases	n mapped bases	n raw reads (% mapped reads)	Min	Q1	Med	Q3	Max	IQR	% bases depth <100×
FR20RV-6N°	N rep1	7 695 675	7 695 553	24 522 (100)	1	138	326.5	653	3 286	515	23.09%
	N rep2	7 695 161	7 695 029	31 372 (100)	1	139	333	734	2 109	595	19.56%
	F	7 695 068	7 694 932	26 683 (100)	1	149	307	650	2 458	501	17.31%
FR20RV-12N	N	7 693 226	7 692 897	26 770 (100)	1	253	394	647	1 609	394	2.11%
	F	7 693 689	7 693 254	27 565 (100)	1	268	435	672	1 418	404	3.25%
K-6N°	N rep1	7 695 878	7 695 698	25 834 (100)	1	140	271	579	3 096	439	15.72%
	N rep2	7 695 826	7 695 623	32 746 (100)	1	97	236	607	3 709	510	25.65%
K-12N	N	7 694 987	7 694 677	31 097 (100)	1	171	320	556	3 611	385	8.98%
FR20RV-6N + K-6N	N rep1	7 696 111	7 695 907	25 181 (100)	1	206	362	697	2 026	491	5.91%
	N rep2	7 695 231	7 695 032	32 115 (100)	1	178	371	724	2 208	546	7.68%
FR20RV-12N + K-12N	N	7 694 566	7 694 158	28 885 (100)	1	261	385	553	2 387	292	2.31%

° To rule out any random effects, the experimental conditions FR20RV-6N and K-6N were repeated independently starting from fresh virus culture aliquots of the same virus lot

* RNA fragmented (F) or not fragmented (N)

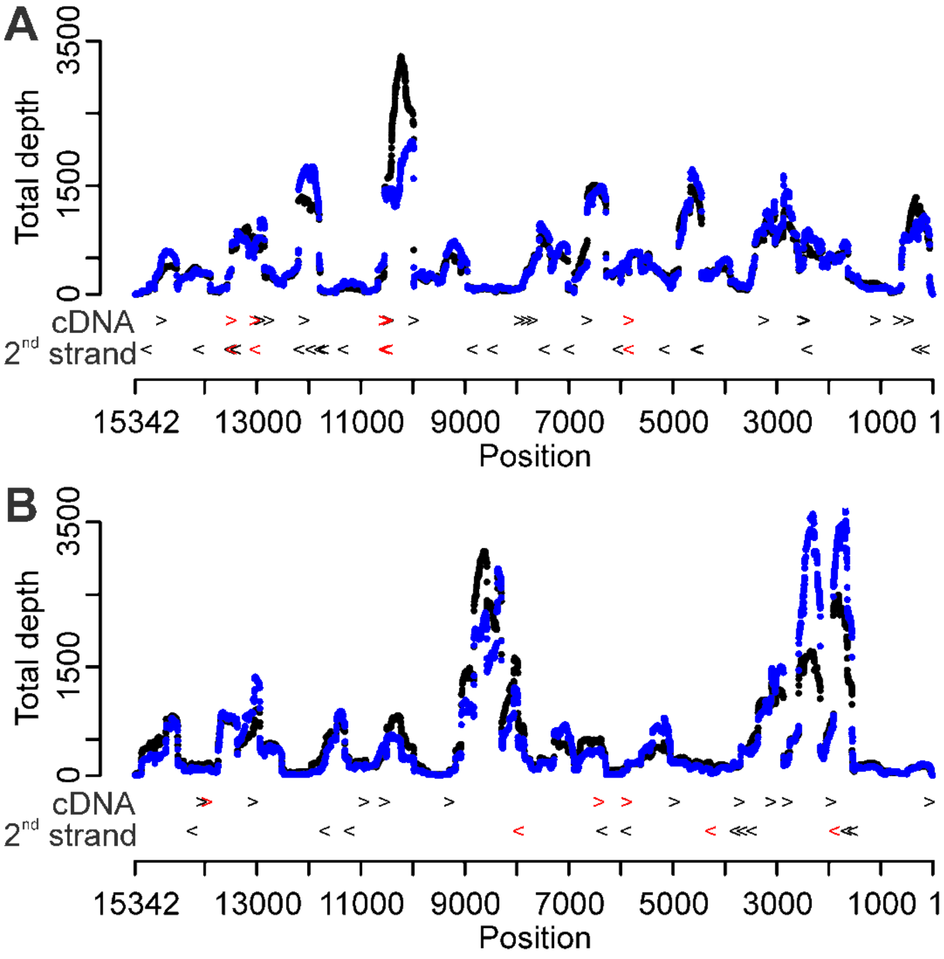


Figure 1: Distribution of sequence depth and enhanced annealing sites for two alternative 6N SISPA primers. (A) FR20RV-6N primer condition, repeat 1 (black) & 2 (blue). (B) K-6N primer condition, repeat 1 (black) & 2 (blue). The X-axis represents the position on the APMV-8 genome. Arrows to the right (>) represent potential enhanced annealing sites in the first strand cDNA synthesis direction, arrows to the left (<) in the second strand cDNA synthesis reaction. Black arrows symbolize enhanced annealing sites on the genome with five consecutive nucleotides in common with the 3' end of the tag of the given SISPA primer; red arrows represent six or more common consecutive nucleotides.

RNA secondary structure and GC-content do not significantly influence sequence depth

We could find no evidence of an overall correlation between areas below average sequence depth and global secondary structure of the viral RNA at denaturation and annealing temperatures (Figure 2 compared to Figure 1). However, the most complex region in the viral RNA secondary structure model (peaking approximately at position 6,000) coincides with a region of low coverage under all conditions tested. To investigate the effect of RNA secondary structure experimentally, we compared libraries obtained from fragmented and non-fragmented RNA. The fragmentation of the RNA was successful with a shift of median RNA fragment size from 1,400 nt for non-fragmented RNA to 400 nt for fragmented RNA (RNA 6000 Pico chip results on the Agilent 2100 Bioanalyzer). To allow comparison with other conditions, we used approximately 7.7 Mb of raw data in the reference assembly. The coverage statistics were similar to the data obtained from non-fragmented RNA versus fragmented RNA (Table 1, FR20RV-6N and FR20RV-12N), and the distribution of coverage along the genome did not change (Figure 3 A).

The unique depth, roughly described as depth where reads with identical starts are omitted, did not show extreme variability over the genome (Figure 3 B).

We found no clear correlation between the GC-content (calculated with a shifting frame of size 401) and sequence depth (Figure 3 B; Spearman rank correlation coefficient of repetition 1: 0.09, P-value < 0.005; repetition 2: 0.06, P-value < 0.005).

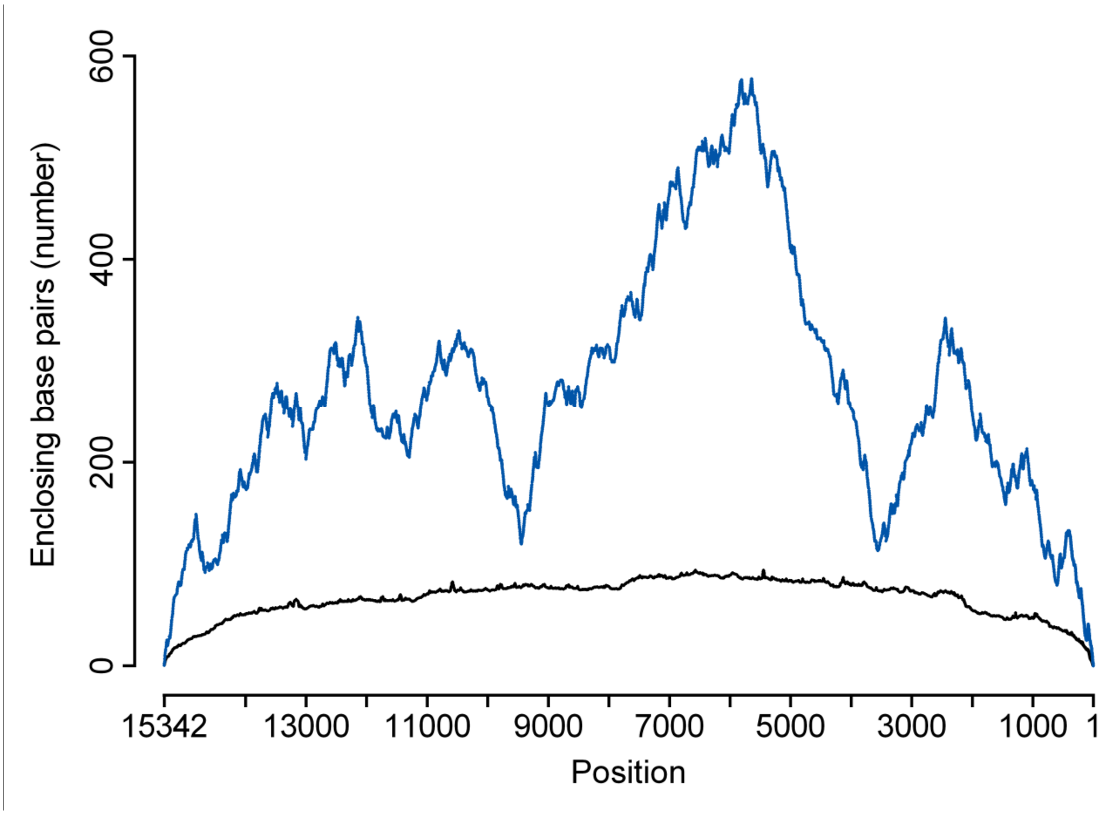


Figure 2: Mountain plot representation of viral RNA secondary structure. Based on base pair probabilities of genomic position. The Y-axis represents the number of base pairs enclosing the base at a certain position in the predicted RNA secondary structure (direct measure of secondary structure complexity). The X-axis represents the position on the APMV-8 genome. The equilibrium pair probabilities were predicted at 95°C (black) and 50°C (blue).

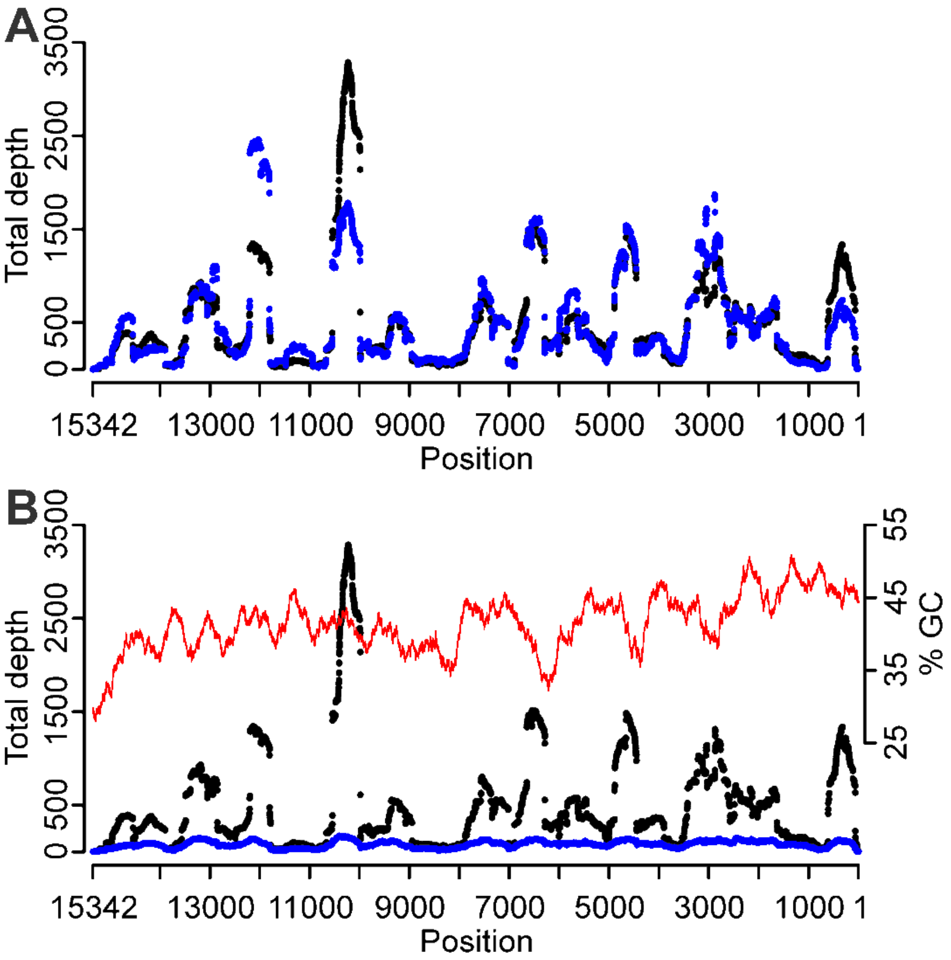


Figure 3: Effect of RNA fragmentation and GC-content on sequence depth distribution. (A) Fragmented (blue) and unfragmented (black) RNA, random amplified with the FR20RV-6N SISPA primer. The X-axis represents the position on the APMV-8 genome. (B) Total (black) and unique (blue) depth distribution of the FR20RV-6N SISPA primer condition. At unique depth, reads with identical starts are omitted. Also positions GC-content (window size 401) of the APMV-8 genome is displayed (red).

Fixed SISPA primer amplification tag sequence induces biased annealing

To analyze if the 5' specific amplification tag of SISPA primer FR20RV-6N influenced the sequence depth, we mapped stretches of nucleotides adjacent to the random part of the primer along the genome sequence. The first one, two, three and four consecutive 3' nucleotides of the amplification tag sequence occurred frequently along the genome (data not shown). Mapping of the possible annealing events on the APMV-8 genome during first strand or second cDNA strand synthesis aided by five, six, or even seven consecutive 3' nucleotides revealed a clear correlation with the variability of the sequence depth (arrows on Figure 1 A).

The six-nucleotide-enhanced annealing sites (red arrows) were found at the positions 5,843; 10,487; 10,549; 13,031; 13,490 along the APMV-8 genome. They were found both in first and second strand cDNA synthesis direction because of the palindromic nature of this sequence. Position 10,487 in the first strand direction and position 13,031 in the second strand direction even had 7 nucleotides in common with the 3’ end of the tag sequence. This resulted in an extreme increase of sequence depth in the adjacent regions (Figure 1 A). The combination of first strand annealing positions with some adjacent second strand positions located in antisense direction were associated with pronounced peaks in sequencing depth in both independent SISPA-NGS repetitions. In addition, a higher density of possible annealing sites resulted in more pronounced peaks in sequence depth (Figure 1 A; data not shown for shorter stretches). In contrast, genomic areas with a low density of enhanced annealing sites had lower sequence depth. It is worth noting that although region 12,851-13,483 had the highest density of potential annealing sites (stretches ranging from 3 to 7 nucleotides) of the whole genome, this region did not have the highest depth of the genome.

Optimization of DNase SISPA annealing

In order to experimentally confirm the primer annealing bias and to improve the randomness of the SISPA protocol, we repeated the SISPA library preparation with modified conditions. These modifications included (a) a random annealing sequence extended from 6 to 12 nucleotides, (b) an alternative primer amplification tag sequence, and (c) the combination of the two primers with alternative tag sequences. The amount of raw data used in all assemblies was again kept at about 7.7 Mb to allow direct comparison between all conditions.

Extended random annealing part of the SISPA primer reduces bias. In order to minimize the impact of the tag sequences on primer annealing we extended the random part of the primer while keeping the tag sequence constant. The 12N oligomer condition was performed on both non-fragmented and fragmented RNA. The use of a 12N SISPA primer resulted in a lower maximum depth, to the advantage of a higher first quartile and median sequence depth (Table 1). The number of genomic bases with depth less than 100 × was almost negligible (2.11 %) compared to the 6N oligomer conditions (23.1 %). The number of bases with depth above 500× stayed the same, but extreme heights disappeared (Figure 4 A). A decreased interquartile range (IQR, Table 1) indicated more homogenous distribution of sequence depth across the genome. Increases and decreases in sequence depths could still be correlated with annealing influence of the amplification tag, but the effect was reduced compared to the 6N oligomer condition resulting in a more equally distributed sequence depth over the genome (Figure 4 A). Again, fragmentation of the RNA prior to cDNA production did not seem to exert any significant effect (Table 1) and no clear correlation between GC-content and sequence depth was found (Spearman rank correlation coefficient: 0.02, P-value < 0.005).

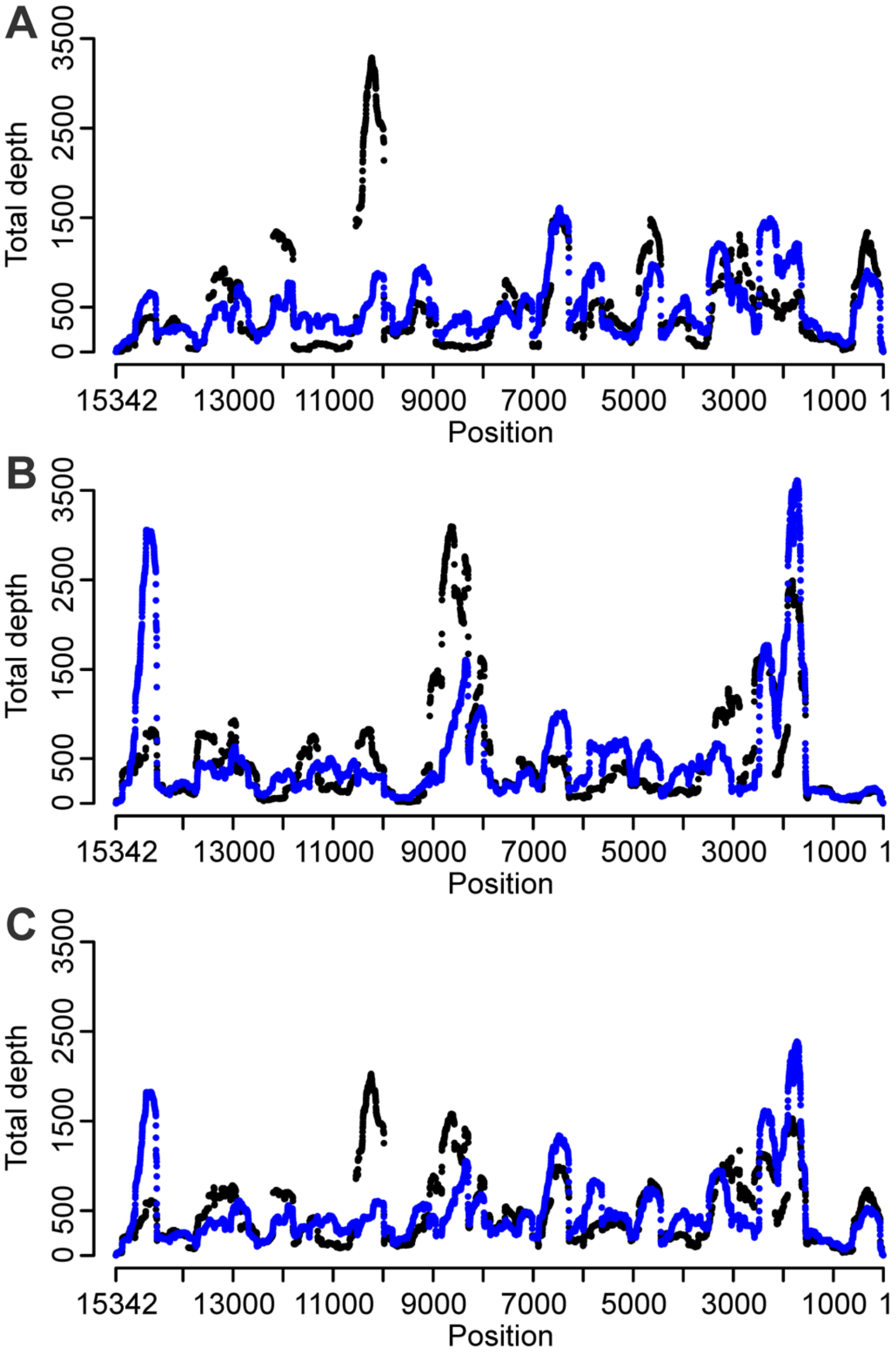


Figure 4: Effect of the SISPA primer sequence (random part and tag) on sequence depth distribution**.** (A) FR20RV-6N (black) and FR20RV-12N (blue). (B) K-6N (black) and K-12N (blue). (C) *In silico* combined conditions "FR20RV-6N + K-6N" (black) and "FR20RV-12N + K-12N" (blue).

An alternative amplification tag sequence relocates bias. Using a random 6N oligomer with an alternative amplification tag sequence, K-6N, the median and Q3 sequence depth were lower compared to sequences resulting from SISPA primer FR20RV-6N (Table 1). Again, highly variable sequence depth was obvious with peaks up to 3,709 × coverage and a high proportion (15.72 %) of genomic positions with depth less than 100 × (Table 1). The use of an alternative amplification tag relocated the areas of extreme coverage to other genomic areas (Figure 1 B, repeat 1). When mapping the possible annealing bias introduced by the alternative amplification tag sequence to the APMV-8 consensus sequence, the correlation between strong annealing sites (of five, six , seven and now even eight 3' nucleotides of the tag sequence) and coverage extremes was confirmed for this alternative tag sequence (arrows on Figure 1 B). The six nucleotide tag sequences (red arrows) were found at the positions 5,888; 6,416 and 13,957 along the APMV-8 genome in first strand cDNA synthesis direction and at positions 1,886, 4,271 and 7,966 in the second strand cDNA synthesis direction. Position 5,888 had seven nucleotides in common with the alternative tag sequence, position 7,966 even eight nucleotides. The high coverage peak in region after position 7,966 could be explained by annealing bias. On the other hand, no extreme coverage depth was visible near position 5,888. An independent SISPA repetition was performed from the same virus stock and confirmed our general findings (Table 1; Figure 1 B, repeat 2). 25.62 % of genomic positions had a depth less than 100 ×. The maximum sequence depth of the peak observed in region 2,166-2,580 differed substantially between the two repetitions (repeat 1: 1,716×; repeat 2: 3,610×).

When comparing the number of enhanced annealing sites on the APMV-8 genome using primer K-6N and FR20RV-6N, we observed fewer enhanced annealing sites at K-6N (arrows on Figure 1 B compared to 1 A). This could contribute to the lower median depth of this K-6N condition (repeat 1: 271 ×; repeat 2: 236 ×) compared to the FR20RV-6N condition (repeat 1: 327 ×; repeat 2: 333 ×).

Using a 12N SISPA primer with identical amplification tag to K-6N (K-12N), the Q1 and median depth was increased, and both the IQR and number of positions with a depth of less than 100 × was decreased compared to the K-6N condition (Table 1). The extremes in the coverage plot were respectively flattened or raised, but the effect was less pronounced compared to the FR20RV-12N SISPA primer (Figure 4 B). The sequence depth extreme of 3 709 × in region 1,560-2,210 of the genome (Figure 4 B) might be explained by a remaining annealing bias effect as this region had the highest concentration of enhanced annealing places of 3 and 4 common nucleotides with the 3’ end of the primer tag, especially in the second strand direction (data not shown). A second unexpected observation was the peak at the 3’ end (14,500-15,000) which had increased enormously compared to the 6N oligomer condition. Moreover, this region did not show a concentration of enhanced annealing places. For logistical reasons, we were unable to perform an independent repetition of the K-12N condition for confirmation.

Combined primer sets reduce bias in silico. Because of the use of an alternative amplification tag sequence resulted in relocation of areas where enhanced annealing induced sequence depth peaks, we tested whether combined libraries from two different SISPA primers might result in a more homogenous sequence depth distribution along the entire genome. We modeled this possibility by performing a combined APMV-8 reference assembly with equal amounts of raw sequence data (each about 3.8 Mb; reads were randomly picked) resulting from the libraries produced using 6N SISPA primers FR20RV-6N and K-6N. As each of these libraries was repeated independently, we modeled two combined libraries resulting from the two repeats of the respective libraries (Table 1, Figure 4 C). Combining the two first repeats resulted in improved coverage statistics (increased Q1 and median depth, and decreased number of bases with depth < 100×) compared to the single SISPA primer conditions alone (Table 1). Combining the two second repeats confirmed this more equal distribution. The decreased IQR indicated a better distribution of depth which was also visible on the sequence depth plot. A similar combined assembly using the 12N SISPA primer conditions (FR20RV-12N and K-12N) resulted in an improvement of the coverage statistics of the stand-alone K-12N data (Table 1, Figure 4 C). Interestingly, this condition had the lowest IQR, with 50 % of the genome positions having a coverage depth in between 261 x and 553 x (Table 1). However, combination of the two primer sets in a single reaction during reverse transcription/second strand cDNA synthesis/PCR repeatedly failed, probably due to interaction between the 2 different SISPA primers. Primer purification after 1st and 2nd cDNA strand synthesis would most likely solve this problem.

DNase SISPA-NGS has high sequence fidelity and allows reliable variant calling. When aligning the consensus sequences of the tested reaction conditions we did not observe any sequence differences. We looked for variation in the largest available dataset (FR20RV-12N condition) making use of all raw data. Only variant positions with depth >100 × and a percentage of minor nucleotide > 10 % were examined to allow reliable variant analysis (Table 2). In comparison we also evaluated variability in K-12N condition. Despite a variation in average depth between the two conditions (FR20RV-12N: 2,300×; K-12N: 1,158×), we observed a very similar variation ratio at the variant positions. For example at genome position 3,279 there was a variant calling of 20 % ‘G’ and 80% ‘A’ in both FR20RV-12N (depth: 5,463×) and K-12N (depth: 1,378×). We found only big differences between conditions when the position was part of a homopolymer stretch (typical base-call errors caused by the 454 pyrosequencing technology). For example, position 2,102 had as major nt ‘A’ and as minor nt ‘G’. At the FR20RV-12N condition the minor percentage was 11 % against 35 % at K-12N. This position is preceded by 4 A’s and followed by 3 G’s. Overall, these data show that sequence variants present in at least 10 % of the reads can be reproducibly identified using DNase SISPA-NGS.

DNase SISPA-NGS is highly specific for viral sequences

A metagenomics analysis of all sequenced raw data showed that all studied conditions were highly specific for APMV-8 RNA (Table 3). The proportion of APMV-8 specific sequences ranges from 87.7 % using SISPA primer FR20RV-12N condition to 95.8 % using SISPA primer K-6N (repeat 1). The remaining fractions contained almost no contaminating host or bacterial sequences, and most of the unassigned sequences in all datasets could be attributed to unknown sequences (most likely technical or sequence database artifacts). The protocols based on 12N SISPA primers were slightly less specific to the advantage of the eukaryotic/host sequences compared to the 6N primers.

The efficient targeting of the protocol enables sequencing complete viral genomes with very little sequencing effort. To investigate the relation between sequencing effort and sequence depth, we randomly picked increasing amounts of reads from the largest available dataset (FR20RV-12N) and looked at its coverage statistics. This random sampling was repeated 10 times and averages are displayed in Figure 5. As expected, the median depth increased linearly with an increasing sequencing effort. We observed also a higher IQR for higher sequencing efforts, which indicated a higher variation of sequence depth (Figure 5). In summary, it was possible to obtain nearly complete (99.9 %) high quality genome sequences missing only a few nucleotides at the genome extremities using as few as 2,500 reads.

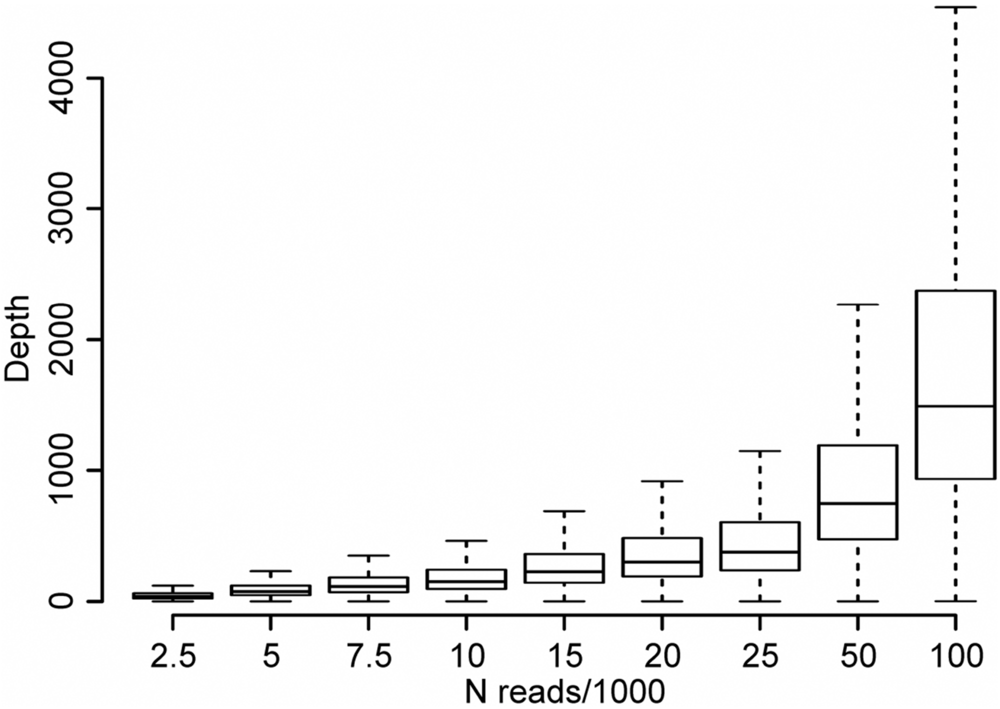
Table 2: Variant calling in the 12-mer datasets at genome positions with depth larger than 100 × (except for position 31 in the K-12N condition) and minor nucleotide percentage of at least 10 %.

position	condition	depth	major nt	minor nt	% minor
31	FR20RV-12N	366	C	T	28 %
	K-12N	63	C	T	28 %
1,908	FR20RV-12N	4,500	G	A	20 %
	K-12N	5,279	G	A	22 %
2,102	FR20RV-12N	3,946	A	G	11 %
	K-12N	1,110	A	G	35 %
3,188	FR20RV-12N	5,216	A	G	19 %
	K-12N	1,073	A	G	19 %
3,279	FR20RV-12N	5,463	A	G	20 %
	K-12N	1,378	A	G	20 %
4,972	FR20RV-12N	1,123	A	G	17 %
	K-12N	515	A	G	13 %
6,526	FR20RV-12N	6,535	G	A	19 %
	K-12N	2,136	G	A	17 %
6,705	FR20RV-12N	4,039	G	A	18 %
	K-12N	1,357	G	A	15 %
7,760	FR20RV-12N	1,481	G	A	20 %
	K-12N	321	G	A	17 %
7,890	FR20RV-12N	1,396	C	G	41 %
	K-12N	724	C	G	34 %
7,895	FR20RV-12N	1,457	G	A	11 %
	K-12N	756	G	A	11 %
8,055	FR20RV-12N	990	A	G	26 %
	K-12N	214	A	G	14 %
12,803	FR20RV-12N	2,652	A	C	33 %
	K-12N	1,026	A	C	32 %
13,410	FR20RV-12N	1,797	T	G	39 %
	K-12N	872	T	G	38 %
14,913	FR20RV-12N	1,611	A	G	18 %
	K-12N	2,786	A	G	17 %
15,155	FR20RV-12N	484	A	T	20 %
	K-12N	430	A	T	28 %

Table 3: Metagenomic analysis of all sequence raw data from the different studied conditions.

Condition	Used reads	APMV-8	Bacteria	Eukaryotes	Unassigned
FR20RV-6N, repeat 1	90,137	94.99 %	0.09 %	0.31 %	3.72 %
FR20RV-6N, repeat 2	69,860	94.40 %	0.58 %	0.81 %	3.62 %
FR20RV-6N, fragmented RNA	80,473	93.61 %	0.34 %	1.13 %	4.34 %
FR20RV-12N	147,004	87.73 %	1.00 %	2.45 %	6.35 %
FR20RV-12N, fragmented RNA	133,895	88.87 %	0.79 %	2.35 %	6.40 %
K-6N, repeat 1	141,400	95.75 %	0.15 %	1.10 %	2.69 %
K-6N, repeat 2	116,494	93.84 %	0.44 %	0.86 %	4.34 %
K-12N	80,258	91.20 %	0.42 %	2.52 %	3.87 %

Figure 5: Effect of increasing sequencing effort on coverage statistics. Reads were randomly picked from the largest dataset (FR20RV-12N). This random sampling was repeated 10 times and averages of coverage statistics are displayed in boxplots.



Discussion

Sequence Independent Single Primer Amplification [14], also known as “random PCR”, is one of the most prominent random access techniques used in viral discovery research. The method provides an efficient enrichment in viral nucleic acids, avoiding excessive host and other contaminant sequence reads [15, 16]. The metagenomics analysis of our data shows that the enrichment procedure for viral nucleic acids is highly efficient, while being applicable to any virus genome and requiring no prior sequence knowledge. A current limitation of the random access protocol described here is the need for sufficient virus material. Although combination of SISPA with NGS makes the protocol more sensitive, full genome sequencing of low titer samples directly from clinical material may be difficult as we recently discussed [30]. Virus genome specific amplification strategies may be better suited to increase the applicability to field samples for known viruses (e.g. [6]). Additionally, the method requires virion protected viral nucleic acids and as such is not applicable to cellular forms of viral nucleic acids (latent infections).

When combining random amplification with NGS, its tendency to create a biased distribution of sequence depth becomes evident (e.g. [22, 23]). Extreme variability in sequence depth does not only imply problematic regions of low coverage, also areas of exaggerated sequence depth compared to the median coverage may result in bioinformatic artifacts during data analysis such as *de novo* contig assembly and variant analysis. The tendency of the 454 technology to produce duplicate reads as a result of multiple beads present in the same micro-reactor during emulsion PCR [31, 32] alone cannot explain the dramatic differences between the coverage plots based on the total depth and those based on unique depth. Our study is the first to formally analyze the possible causes of this sequence depth variability.

Victoria and colleagues [22] suggested that regions with overrepresented sequencing depth are associated with annealing bias of the used primer, while secondary structure may result in areas of low sequencing depth. Formally modeling secondary structure and annealing factors on the one hand and using experimental data on the other hand, our study identified annealing bias as the main cause of the overrepresented regions after performing random priming in the SISPA method. Annealing of the random 3’ end of SISPA primers seems to be locally enhanced when 3, 4, 5, 6, 7 or even 8 nucleotides from their specific 5’ amplification tag (20 nucleotides in length) designed for PCR amplification assist the random oligomer part of the primer in annealing during the first strand and/or second strand cDNA synthesis. This results in regions of exaggerated sequence. The importance of this highly reproducible annealing effect seems to depend on the length of the complementary region between the primer and the genome and the proximity of first strand and second strand enhanced annealing positions. Going beyond a theoretical mapping of these annealing effects, our data show that the majority of sequence depth bias can be explained by this effect.

Using a different specific amplification tag sequence, we observed a clear shift of the coverage extremes to other genomic regions caused by the shift of possible enhanced annealing sites. Again, the coverage depth was very variable, but the median depth was lower compared to original used primer. This is probably due to the lower number of potential strong annealing places on the APMV-8 genome for the alternative random primer K-6N. This confirms that the distribution of possible enhanced annealing sites depends on the SISPA primer amplification tag sequence and on the target genome sequence, which implies possible differential annealing dynamics when targeting other virus genomes.

Using any of these SISPA primers, regions in the genome existed where predicted enhanced annealing sites did not result in extreme sequence depth, for example the region around positions 6,416 and 5,888 at the K-6N primer condition have respectively 6 and 7 nucleotides in common with the tag sequence. Similarly, the region around position 5,843 at the FR20RV-6N primer condition has high similarity with the tag sequence. However, these predicted enhanced annealing sites do not seem to coincide with increased sequence coverage in this region. When looking at the RNA secondary structure mountain plot at first strand cDNA annealing temperature (Figure 2, 50 °C) we see that these positions are located in the most complex region of the genome. It should be noted that we started our first strand cDNA synthesis step with a denaturation at 95 °C for a better removal of RNA secondary (Figure 2, 95 °C) and tertiary structures. Nevertheless, it could be that the RNA locally refolds at lower temperature during cDNA synthesis. In addition, extreme coverage was also observed in regions which did not show a high concentration of enhanced annealing sites, for example genome region 14,500-15,000 at the K-12N primer condition. This region had a very low GC-content and no complex RNA structures, but did not yield high coverage when targeted with primer K-6N. In none of the conditions, a clear overall correlation was found between GC% and sequencing depth (Spearman rank correlation coefficient was always between -0.15 and 0.09, P-value < 0.005). These observations show that, although being a major correlate of sequence depth, annealing bias is not the only factor contributing to variations in sequence depth. These sequence coverage variations that remain unexplained by SISPA primer annealing/extension bias highlight the complexity of annealing and extension dynamics which can be locally influence by multiple interacting factors such as RNA secondary structure, GC content and oligomer length. Victoria and colleagues [22] suggested already that complex RNA secondary structures contributed to regions with problematic sequence depth. We did not observe a change in distribution of sequence depth after fragmentation of the viral RNA for the studied avian paramyxovirus genome. Although the global secondary structure of the viral RNA is thus successfully fragmented, it cannot be excluded that local RNA folding may still have influenced primer annealing. However, we could not further fragment the RNA in order to assure compatibility with the RNA sequencing workflow.

These findings confirm earlier observations by Wong and colleagues [33]. That study used SISPA amplification in the context of a pathogen detection DNA microarray. In initial experiments using random priming amplification to identify pathogens they observed frequently incomplete hybridization of the pathogen genomes marked by interspersed genomic regions not detected by the probes that could not be explained by sequence polymorphisms, probes GC content and genome secondary structure. The composition of the SISPA primer tag had a significant impact on the efficiency of viral genome amplification – as suggested in our study. Using an algorithm to optimize primer sequences for uniform amplification efficiency across the viral genomes included in the DNA array, they managed to increase the sensitivity of pathogen detection of their microarray.

Previous studies have used different lengths of the 3’ random annealing part of the SISPA primer. Examples include hexamer [16, 17, 20, 26, 30, 34-38], octamer [19, 22, 23, 39-42], nonamer [24, 43-45] and decamer random 3’ annealing parts [46]. Stangegaard and colleagues [47] studied the impact of different random primers (without 5’ specific amplification tag sequence) on the yield and quality of synthesized cDNA, concluding that reverse transcription using random pentadecamer primers increases yield and quality of resulting cDNA compared with hexa- and nonamers. In our study, we compared SISPA-primers with hexamer and dodecamer random annealing stretches. Theoretically, the number of possible annealing sequences for a 12N primer is 4^{12} (=16,777,216), compared to 4^{6} (=4,096) for a hexamer. In our study, increasing the random part of the SISPA primer from 6N to 12N reduced the amplification tag induced sequence depth bias to the advantage of the regions with problematic coverage.

We combined two factors affecting the distribution of sequence reads over the genome in a single assembly: (a) alternative SISPA primers are biased to alternative genome regions, and (b) longer oligomer-based SISPA primers tend to have a more equal distribution. The combination of data from libraries resulting from two different 12N SISPA primers (FR20RV-12N and K-12N) resulted in the best distribution of sequence reads over the APMV-8 genome. We suggest that in future efforts (using DNase SISPA for the determination of complete viral genomes), a combination of a longer random annealing part (e.g.12N) and the combined assembly of data resulting from libraries amplified with alternative amplification tags, may result in an improved homogeneity of sequence depth distribution over the genome. Ultimately, a more homogenous sequence depth distribution reduces the sequencing effort (and thus cost) needed for genome completion. In this study, 2 500 GS-FLX titanium reads were enough to sequence 99.9 % of the APMV-8 viral genome with a median coverage of 38.3 ×. We have shown that median coverage increased with the increasing sequencing effort. 7,500 reads covered the full genome with a median depth of more than a 100 ×. This confirms our previous experience that only about 5,000 GS-FLX titanium reads from a library of SISPA amplified viral RNA was enough in determining the complete genome of uncharacterized avian paramyxoviruses [26]. It should be noted that amplified viral stock was used in this study.

Next generation sequencing is becoming increasingly accessible to laboratories, both through the evolution of sequencing platforms and chemistries and through the increasing availability of sequencing service providers. Combined with opportunities to multiplex samples during sequencing, this technology is now evolving towards a cost-effective methodology for genome sequencing. Generic, sequence independent access methods such as optimized DNase SISPA may facilitate access to viral genome sequences, without the need for prior sequence knowledge. In addition, our findings may be of value to other technologies requiring random nucleic acid amplification such as DNA microarrays.

Acknowledgements

We thank the Friedrich-Loeffler-Institut for supplying the avian paramyxovirus and Matthias Scheuch for performing the metagenomic analysis. We are also grateful to Orkun Ozhelvaci, Moctezuma Reimann and Mireille Decaesstecker for excellent technical assistance.

References

1. Sanger, F., S. Nicklen, and A.R. Coulson, *DNA sequencing with chain-terminating inhibitors.* Proc Natl Acad Sci U S A, 1977. **74**(12): p. 5463-7.
2. Obenauer, J.C., et al., *Large-scale sequence analysis of avian influenza isolates.* Science, 2006. **311**(5767): p. 1576-80.
3. Margulies, M., et al., *Genome sequencing in microfabricated high-density picolitre reactors.* Nature, 2005. **437**(7057): p. 376-80.
4. Metzker, M.L., *Sequencing technologies - the next generation.* Nat Rev Genet, 2010. **11**(1): p. 31-46.
5. Höper, D., B. Hoffmann, and M. Beer, *Simple, Sensitive, and Swift Sequencing of Complete H5N1 Avian Influenza Virus Genomes.* J Clin Microbiol, 2009. **47**(3): p. 674-9.
6. Höper, D., B. Hoffmann, and M. Beer, *A comprehensive deep sequencing strategy for full-length genomes of influenza A.* PLoS One, 2011. **6**(4): p. e19075.
7. Leifer, I., et al., *Molecular epidemiology of current classical swine fever virus isolates of wild boar in Germany.* J Gen Virol, 2010. **91**(Pt 11): p. 2687-97.
8. Wright, C.F., et al., *Beyond the consensus: dissecting within-host viral population diversity of foot-and-mouth disease virus by using next-generation genome sequencing.* J Virol, 2011. **85**(5): p. 2266-75.
9. Ghedin, E., et al., *Deep sequencing reveals mixed infection with 2009 pandemic influenza A (H1N1) virus strains and the emergence of oseltamivir resistance.* J Infect Dis, 2011. **203**(2): p. 168-74.

10. Neverov, A. and K. Chumakov, *Massively parallel sequencing for monitoring genetic consistency and quality control of live viral vaccines.* Proc Natl Acad Sci U S A, 2010. **107**(46): p. 20063-8.
11. Bexfield, N. and P. Kellam, *Metagenomics and the molecular identification of novel viruses.* Vet J, 2010.
12. Jarrett, R.F., et al., *Molecular methods for virus discovery.* Dev Biol (Basel), 2006. **123**: p. 77-88; discussion 119-32.
13. Ambrose, H.E. and J.P. Clewley, *Virus discovery by sequence-independent genome amplification.* Rev Med Virol, 2006. **16**(6): p. 365-83.
14. Reyes, G.R. and J.P. Kim, *Sequence-independent, single-primer amplification (SISPA) of complex DNA populations.* Mol Cell Probes, 1991. **5**(6): p. 473-81.
15. Allander, T., et al., *A virus discovery method incorporating DNase treatment and its application to the identification of two bovine parvovirus species.* Proc Natl Acad Sci U S A, 2001. **98**(20): p. 11609-14.
16. Djikeng, A., et al., *Viral genome sequencing by random priming methods.* BMC Genomics, 2008. **9**: p. 5.
17. Miller, P.J., et al., *Evidence for a New Avian Paramyxovirus Serotype-10 Detected in Rockhopper Penguins from the Falkland Islands.* J Virol, 2010. **84**(21): p. 11496-11504.
18. Afonso, C.L., *Sequencing of avian influenza virus genomes following random amplification.* Biotechniques, 2007. **43**(2): p. 188, 190, 192.
19. Victoria, J.G., et al., *Rapid identification of known and new RNA viruses from animal tissues.* PLoS Pathog, 2008. **4**(9): p. e1000163.
20. Blomström, A.L., et al., *Detection of a novel astrovirus in brain tissue of mink suffering from shaking mink syndrome by use of viral metagenomics.* J Clin Microbiol, 2010. **48**(12): p. 4392-6.
21. Bishop-Lilly, K.A., et al., *Arbovirus detection in insect vectors by rapid, high-throughput pyrosequencing.* PLoS Negl Trop Dis, 2010. **4**(11): p. e878.
22. Victoria, J.G., et al., *Metagenomic analyses of viruses in stool samples from children with acute flaccid paralysis.* J Virol, 2009. **83**(9): p. 4642-51.
23. Victoria, J.G., et al., *Viral nucleic acids in live-attenuated vaccines: detection of minority variants and an adventitious virus.* J Virol, 2010. **84**(12): p. 6033-40.
24. Onions, D. and J. Kolman, *Massively parallel sequencing, a new method for detecting adventitious agents.* Biologicals, 2010. **38**(3): p. 377-80.
25. Greninger, A.L., et al., *A metagenomic analysis of pandemic influenza A (2009 H1N1) infection in patients from North America.* PLoS ONE, 2010. **5**(10): p. e13381.
26. Rosseel, T., et al., *Identification and complete genome sequencing of paramyxoviruses in mallard ducks (Anas platyrhynchos) using random access amplification and next generation sequencing technologies.* Virol J, 2011. **8**: p. 463.
27. Stang, A., et al., *Characterization of virus isolates by particle-associated nucleic acid PCR.* J Clin Microbiol, 2005. **43**(2): p. 716-20.
28. R Development Core Team, *R: A Language and Environment for Statistical Computing*. 2011, R Foundation for Statistical Computing: Vienna, Austria.
29. Lorenz, R., et al., *ViennaRNA Package 2.0.* Algorithms Mol Biol, 2011. **6**: p. 26.
30. Rosseel, T., et al., *DNase SISPA-next generation sequencing confirms Schmallenberg virus in Belgian field samples and identifies genetic variation in Europe.* PLoS One, 2012. **7**(7): p. e41967.
31. Dong, H., et al., *Artificial duplicate reads in sequencing data of 454 Genome Sequencer FLX System.* Acta Biochim Biophys Sin (Shanghai), 2011. **43**(6): p. 496-500.

32. Niu, B., et al., *Artificial and natural duplicates in pyrosequencing reads of metagenomic data.* BMC Bioinformatics, 2010. **11**: p. 187.
33. Wong, C.W., et al., *Optimization and clinical validation of a pathogen detection microarray.* Genome Biol, 2007. **8**(5): p. R93.
34. Allander, T., et al., *Cloning of a human parvovirus by molecular screening of respiratory tract samples.* Proc Natl Acad Sci U S A, 2005. **102**(36): p. 12891-6.
35. Yu, X., et al., *A random PCR screening system for the identification of type 1 human herpes simplex virus.* J Virol Methods, 2009. **161**(1): p. 91-7.
36. Goebel, S.J., et al., *Isolation of avian paramyxovirus 1 from a patient with a lethal case of pneumonia.* J Virol, 2007. **81**(22): p. 12709-14.
37. Zsak, L., K.O. Strother, and J. Kisary, *Partial genome sequence analysis of parvoviruses associated with enteric disease in poultry.* Avian Pathol, 2008. **37**(4): p. 435-41.
38. Van Borm, S., et al., *Phylogeographic analysis of avian influenza viruses isolated from Charadriiformes in Belgium confirms intercontinental reassortment in gulls.* Arch Virol, 2012. **157**(8): p. 1509-22.
39. Adams, I.P., et al., *Next-generation sequencing and metagenomic analysis: a universal diagnostic tool in plant virology.* Mol Plant Pathol, 2009. **10**(4): p. 537-45.
40. Li, J.S., et al., *[Sequence the complete sequence of bocavirus I with SISPA-PCR].* Zhonghua Shi Yan He Lin Chuang Bing Du Xue Za Zhi, 2010. **24**(1): p. 14-6.
41. Briese, T., et al., *Genetic detection and characterization of Lujo virus, a new hemorrhagic fever-associated arenavirus from southern Africa.* PLoS Pathog, 2009. **5**(5): p. e1000455.
42. Wang, D., et al., *Microarray-based detection and genotyping of viral pathogens.* Proc Natl Acad Sci U S A, 2002. **99**(24): p. 15687-92.
43. Wang, D., et al., *Viral discovery and sequence recovery using DNA microarrays.* PLoS Biol, 2003. **1**(2): p. E2.
44. Finkbeiner, S.R., et al., *Metagenomic analysis of human diarrhea: viral detection and discovery.* PLoS Pathog, 2008. **4**(2): p. e1000011.
45. Gaynor, A.M., et al., *Identification of a novel polyomavirus from patients with acute respiratory tract infections.* PLoS Pathog, 2007. **3**(5): p. e64.
46. Kapoor, A., et al., *A highly divergent picornavirus in a marine mammal.* J Virol, 2008. **82**(1): p. 311-20.
47. Stangegaard, M., I.H. Dufva, and M. Dufva, *Reverse transcription using random pentadecamer primers increases yield and quality of resulting cDNA.* Biotechniques, 2006. **40**(5): p. 649-57.

CHAPTER 4.2

Evaluation of convenient pretreatment protocols for RNA virus metagenomics in serum and tissue samples

Toon Rosseel, Orkun Ozhelvaci, Graham Freimanis, Steven Van Borm

Published in Journal of Virological Methods (2015), pii: S0166-0934(15)00209-8. doi:
10.1016/j.jviromet.2015.05.010. [Epub ahead of print]

Due to recent advances in library preparation methods, requiring significant lower input, direct sequencing without an extra pre-amplification step is evaluated in this chapter. Illumina sequencing on the MiSeq platform using the Nextera XT library preparation kit was selected. In addition, different convenient and widely used pretreatment methods are compared and evaluated for their relative virus discovery sensitivity in serum and tissue samples. TriPure isolation Reagent was selected as preferred extraction method to isolate high quality RNA.

Abstract

Viral metagenomic approaches are increasingly being used for viral discovery. Various strategies are applied to enrich viral sequences, but there is often a lack of knowledge about their effective influence on the viral discovery sensitivity. We evaluate some convenient and widely used approaches for RNA virus discovery in clinical samples in order to reveal their sensitivity and potential bias introduced by the enrichment or amplifications steps. An RNA virus was artificially spiked at a fixed titer in serum and lung tissue, respectively low and high nucleic acid content matrices. For serum, a simple DNase treatment on the RNA extract gave the maximum gain in proportion of viral sequences (83×), and a subsequent ribosomal RNA removal nearly doubled once more the proportion of viral sequences. For lung tissue, a ribosomal RNA depletion step on the RNA extract had the biggest gain in proportion of viral sequences (32×). We show also that direct sequencing of cDNA is recommended above an extra random PCR amplification step, and a that the virion enrichment strategy (filtration and nuclease treatment) has a beneficial effect for sequencing-based virus discovery. Our findings provide sample-dependent guidelines for targeted virus discovery strategies.

Keywords: RNA virus discovery, clinical samples, viral metagenomics, pretreatment, next-generation sequencing, rRNA depletion

1. Introduction

Viral metagenomics approaches can provide insights into the composition and structure of environmental viral communities [1] and are being increasingly used for viral discovery in diseased humans or animals [2-6]. The applied protocols usually start with a virion enrichment step such as (ultra)centrifugation, polyethylene glycol (PEG) precipitation, (ultra)filtration, chloroform treatment, or nuclease treatment [7]. Following pretreatment steps, a sequence independent genome amplification method may be used to generate sufficient DNA (reviewed in Ambrose et al., 2006). Finally, DNA sequencing is performed, followed by bioinformatics analysis of the sequences (comparing results to publicly available sequence information). In the past, viral metagenomics was accomplished by molecular cloning and Sanger sequencing [8, 9], whereas more recent techniques such as next generation sequencing (NGS) have made both the cloning and amplification (in the case of sufficient quantities of initial template DNA) steps unnecessary. Concordantly, the revolutions in sequencing output, coupled with the downward trends in running costs have greatly improved the potential of viral discovery [10-12]. However, particularly in complex biological samples (e.g. tissues), the relatively low abundance of viral nucleic acids in comparison to a background of host, bacterial and other contaminating genetic material [13] remains a challenge. In short, increasing levels of viral genetic material, whilst reducing background signal (host) is the key to a successful metagenomic-based viral discovery pipeline. Numerous viral enrichment and sequence independent amplification approaches have been applied to a wide variety of sample types [7, 14, 15], but there has often been a paucity of knowledge regarding their impact on the sensitivity of viral discovery. Furthermore, few studies have specifically evaluated the efficiencies of either the viral enrichment steps, the sequence independent amplification methods or the sequencing platforms [11, 15-21]. Moreover, as a consequence of the lack of standardization between these studies, it is often difficult to directly compare conclusions. Therefore, additional evaluations of proof of concept focusing on the analytical sensitivity of the tested protocols in different sample types are required to estimate the diagnostic performance of the tested metagenomics approaches.

In this study we focus on RNA virus discovery, as the majority of relevant (re)emerging viral threats are caused by RNA viruses [22]. We compare the sensitivity and bias of rapid and easy to use viral metagenomics approaches in different sample types. Newcastle disease virus (NDV) was taken to be representative of how pretreatment protocols would impact upon viral

RNA. NDV has a single-stranded, non-segmented, negative-sense RNA genome of approximately 15 kb in size [23]. A widely used 2-step virion enrichment method (filtration and nuclease-treatment) was compared to a control without enrichment. Moreover, the impact of DNase treatment and ribosomal RNA (rRNA) depletion on the extracted RNA was also evaluated. Finally, the influence of sequence independent genome amplification was compared to direct sequencing of the complementary DNA (cDNA).

2. Materials and Methods

2.1. Sample selection and virus spiking

The LaSota vaccine strain of Newcastle disease virus (NDV; family *Paramyxoviridae*, genus *Avulavirus*) was selected for the purposes of optimization and comparison of different pretreatment protocols. Samples were collected from specific pathogen free white leghorn chickens (lung, brain, intestine, whole blood, serum). Extracted total nucleic acids (ssDNA, dsDNA, RNA) were quantified using the respective Qubit assays (Life Technologies) and quality assessed using the Agilent 2100 Bioanalyzer (Agilent Technologies) (data not shown). Phosphate buffered saline, containing a mixture of antibiotics (10^7 U/l penicillin, 10 g/l streptomycin, 0.25 g/l gentamycin), was added to lung tissue (10 % w/v) and incubated for 1 h at 4 °C. The suspension was homogenized for 3 min at 30 Hz using a TissueLyser II (Qiagen). The lung homogenate and serum were centrifuged* and spiked with NDV to a final concentration of 10^4 EID50/ml, prior to storage at –80 °C.

2.2. Sample pretreatments and RNA extraction

Different pretreatment protocols were tested on both serum (*'S'*) and tissue (*'T'*) samples (Table 1 and Figure 1). The direct RNA extraction without enrichment (Section 2.2.1) was compared with the 2-step virion enrichment method (Section 2.2.2). For each group, untreated RNA was compared with DNase (Section 2.2.3) and rRNA removal (Section 2.2.4) treatments on the RNA extracts. Nucleic acid content of all RNA extracts were determined with Qubit RNA HS and dsDNA HS assays (Life Technologies).

Table 1: Overview of the used pretreatment protocols, virus specific CP values, nucleic acid concentrations of RNA extracts and amplification strategy. Both serum (S) and (lung) tissue (T) were spiked with 10^4 EID50/ml Newcastle disease virus (NDV). Viral RNA was quantified with an NDV specific qRT-PCR, The displayed CP values are the average of 2 replicates. Total RNA was quantified with Qubit RNA and dsDNA assays (Life Technologies).

Protocol short name*	Sample type	Pretreatment protocols			NDV CP value	Nucleic acid concentration of RNA extract		Amplification strategy
		2-step enrichment	Dnase treatment on RNA extract	rRNA remvoval on total RNA		RNA (ng/µl)	dsDNA (ng/µl)	
S-Control1	Serum	No	No	No	28.58	11.65	12.55	cDNA synthesis
S-Control2	Serum	No	No	No	29.07	6.64	5	cDNA synthesis
S-Control1 + Dnase	Serum	No	Yes	No	29.18	<0.02	<0.0005	cDNA synthesis
S-Control2 + Dnase	Serum	No	Yes	No	29.35	<0.02	<0.0005	cDNA synthesis
S-Control2 + Dnase + Removal	Serum	No	Yes	Yes	30.69	<0.02	<0.0005	cDNA synthesis
S-Control1 + Removal1	Serum	No	No	Yes	29.02	Not tested	Not tested	cDNA synthesis
S-Control1 + Removal2	Serum	No	No	Yes	28.69	Not tested	Not tested	cDNA synthesis
S-2-step	Serum	Yes	No	No	29.68	<0.02	0.413	cDNA synthesis
S-2-step + Removal1	Serum	Yes	No	Yes	30.2	Not tested	Not tested	cDNA synthesis
S-2-step + Removal2	Serum	Yes	No	Yes	28.75	Not tested	Not tested	cDNA synthesis
S-2-step + rPCR	Serum	Yes	No	No	29.68	<0.02	0.413	cDNA + random PCR
S-2-step + Removal2 + rPCR	Serum	Yes	No	Yes	28.75	Not tested	Not tested	cDNA + random PCR
T-Control1	Lung tissue	No	No	No	29.56	105	14.6	cDNA synthesis
T-Control2	Lung tissue	No	No	No	30.15	74.8	11.6	cDNA synthesis
T-Control1 + Dnase	Lung tissue	No	Yes	No	30.14	Not tested	Not tested	cDNA synthesis
T-Control2 + Dnase	Lung tissue	No	Yes	No	29.39	72.5	10.3	cDNA synthesis
T-Control2 + Dnase + Removal	Lung tissue	No	Yes	Yes	31.02	42.7	4.18	cDNA synthesis
T-Control1 + Removal1	Lung tissue	No	No	Yes	Not tested	47.6	Not tested	cDNA synthesis
T-Control1 + Removal2	Lung tissue	No	No	Yes	30.57	Not tested	Not tested	cDNA synthesis

Table 1 (continued)

T-2-step	Lung tissue	Yes	No	No	31.28	2.4	0.152	cDNA synthesis
T-2-step + Removal1	Lung tissue	Yes	No	Yes	33.01	Not tested	Not tested	cDNA synthesis
T-2-step + Removal2	Lung tissue	Yes	No	Yes	33.23	Not tested	Not tested	cDNA synthesis
T-2-step + rPCR	Lung tissue	Yes	No	No	31.28	2.4	0.152	cDNA + random PCR
T-2-step + Removal2 + rPCR	Lung tissue	Yes	No	Yes	33.23	Not tested	Not tested	cDNA + random PCR

* 1 and 2 are independent repetitions of either RNA extraction or rRNA removal reaction.

Table 2: Average sequencing depth and NDV genome coverage per evaluated pretreatment approach. Both serum (S) & lung tissue (T) sample were spiked with 10^4 EID50/ml NDV. Control = no virion enrichment, 2-step = 2-step virion enrichment, DNase = DNase treatment on RNA extract, Removal = rRNA removal on RNA extract, rPCR = random PCR amplification, 1 and 2 are independent repetitions of either RNA extraction or rRNA removal.

Serum condition	**Average depth**	**% Overlap**	**Lung tissue condition**	**Average depth**	**% Overlap**
S-Control1	0.193	10.338	**T-Control1**	0.029	1.442
S-Control2	0.365	17.839	**T-Control2**	0.099	7.151
S-Control1 + Dnase	17.546	97.847	**T-Control1 + Dnase**	0.062	3.339
S-Control2 + Dnase	37.852	99.513	**T-Control2 + Dnase**	0.022	2.180
S-Control2 + Dnase + Removal	50.647	99.987	**T-Control2 + Dnase + Removal**	1.364	45.423
S-Control1 + Removal1	0.083	5.676	**T-Control1 + Removal1**	0.989	39.813
S-Control1 + Removal2	0.055	5.143	**T-Control1 + Removal2**	0.360	19.913
S-2step	0.739	35.263	**T-2-step**	0.109	7.211
S-2-step + Removal1	0.476	24.891	**T-2-step + Removal1**	0.865	33.926
S-2-step + Removal2	5.181	95.173	**T-2-step + Removal2**	0.655	33.537
S-2-step + rPCR	2.050	35.506	**T-2-step + rPCR**	0	0
S-2-step + Removal2 + rPCR	1.940	23.074	**T-2-step + Removal2 + rPCR**	0.013	1.277

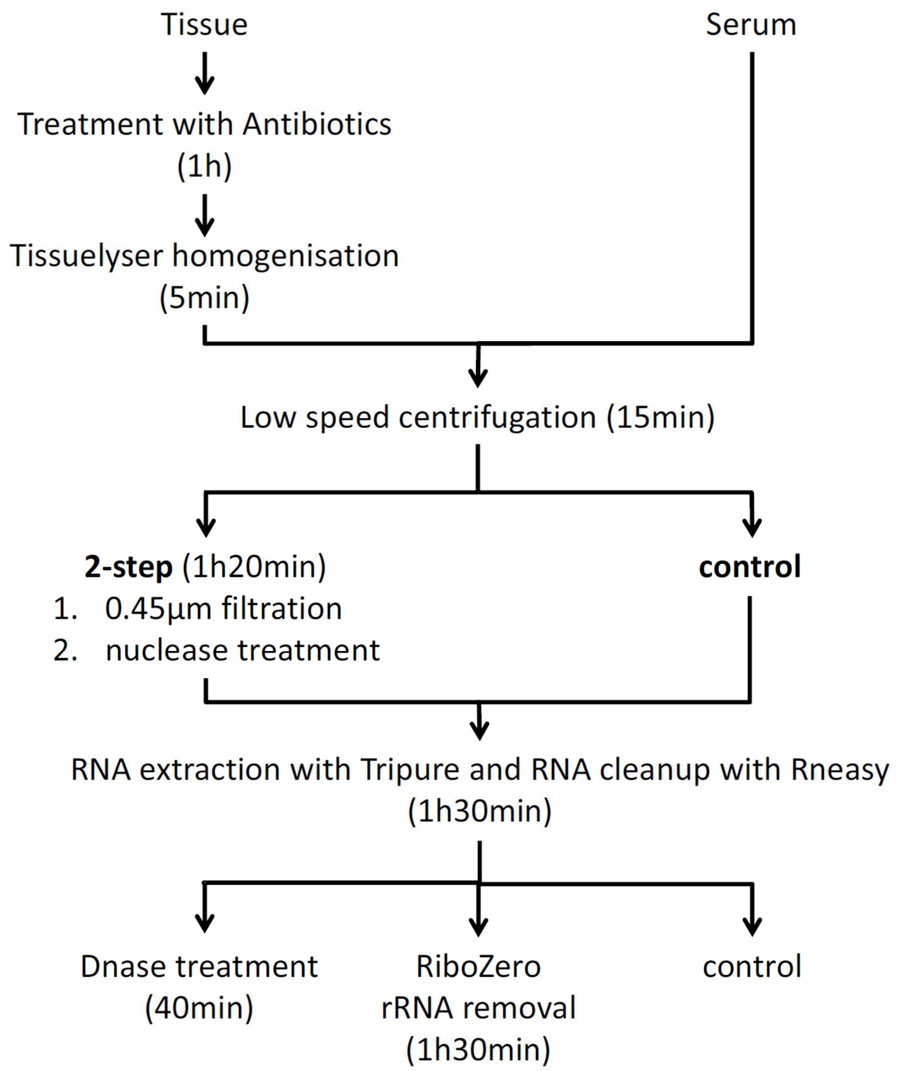

Tissue
Serum
Treatment with Antibiotics (1h)
Tissuelyser homogenisation (5min)
Low speed centrifugation (15min)
2-step (1h20min)
1. 0.45µm filtration
2. nuclease treatment
control
RNA extraction with Tripure and RNA cleanup with Rneasy (1h30min)
Dnase treatment (40min)
RiboZero rRNA removal (1h30min)
control

Figure 1: Schematic overview of the tested sample pretreatment steps in the viral metagenomic workflow.

2.2.1. No virion enrichment ('control')

RNA was extracted (in duplicate - control1 and control2), from 1 ml of virus-spiked sample (processed in the absence of viral enrichment) using the TriPure isolation Reagent (Roche) according to manufacturer's instructions. The aqueous phase was subsequently purified with the RNeasy Mini Kit (Qiagen) prior to elution in 60 µl nuclease-free water.

2.2.2. Two-step viral enrichment ('2-step')

Two pretreatment steps (0.45 µm filtration followed by nuclease treatment) were performed on 1 ml of spiked sample as described previously [24]. The nuclease treatment step was performed using 50 U TURBO DNase (2 U/µl, Life Technologies), 1× TURBO DNase Buffer and 10 U RiboShredder RNase Blend (1 U/µl, Epicentre Technologies) as reported previously [24].

2.2.3. DNase treatment on RNA

Extracted RNA duplicates were treated with TURBO DNase (20 U/30 µl RNA; Life Technologies) for 30 min at 37 °C, prior to purification using the RNeasy mini kit (Qiagen).

2.2.4. Ribosomal RNA removal on RNA

Cytoplasmic and mitochondrial ribosomal RNA depletion of both the ‘no virion enrichment’ control and ‘2-step enrichment’ were performed using the Ribo-Zero Magnetic Gold Epidemiology kit (Epicentre Technologies). rRNA removal was performed using the Ribo-Zero kit (Epicentre Technologies) as per manufacturer’s instructions. If RNA quantity was less than 500 ng per 28 µl RNA, then a low-input protocol was used as described in the manual of the ScriptSeq Complete Gold Epidemiology Kit (Epicentre Technologies). Ribosomal RNA depletions were performed in duplicate (‘Removal1’ and ‘Removal2’). Additionally, an rRNA removal treatment upon previously DNase treated RNA (Section 2.2.3) was also performed (‘DNase + Removal’).

2.3. Virus quantification by q(RT-)PCR

To measure the impact of pretreatment steps on the viral RNA, a specific real-time quantitative RT-PCR assay was performed (in duplicate) on the RNA extracts and on DNase or rRNA-removal treated RNA. Viral RNA was quantified using QuantiTect Probe RT-PCR (Qiagen) on a LightCycler 480 real-time PCR system (Roche) as previously described [25].

2.4. cDNA synthesis and amplification

cDNA synthesis using random hexamer primers was performed according to the guidelines in the 454 sequencing cDNA Rapid library preparation method manual (Roche, version March 2012, Section 3.2), but excluding RNA fragmentation. Briefly, double-stranded cDNA was synthesized with the cDNA Synthesis System kit (Roche) from 17 µl of RNA using 400 µM Primer “random” (Roche) and subsequently purified using Agencourt AMPure XP beads (1.6:1 bead/cDNA ratio; Beckman Coulter).

Sequence independent PCR amplification was performed as previously described [24] on a single replicate of the 2-step virion enriched samples (control & rRNA removal). Random

PCR fragments were size selected on a 1 % agarose gel, and fragments between 200 and 800 bp were excised and purified using the High Pure PCR Product Purification Kit (Roche).

The purified random fragments and PCR-free cDNA were both quantified with the appropriate Qubit assay kit (Life Technologies) and quality checked with the Agilent 2100 Bioanalyzer (Agilent Technologies).

2.5. Sequencing

Sequencing libraries were prepared using the Nextera XT sample preparation kit (Illumina) according the manufacturer's instructions. Total dsDNA input of 1 ng was used unless cDNA was present at lower concentrations. In these cases, the maximum available cDNA was used as an input (i.e. 5 µl). Libraries were quantified with the Kapa library quantification kit for Illumina sequencing (Kapa Biosystems) and their fragment length distribution was verified using the Agilent Bioanalyzer with the High Sensitivity DNA Kit (Agilent Technologies). Sequencing was performed on the MiSeq benchtop sequencer at The Pirbright Institute using a MiSeq Reagent kit v3 (Illumina) with 2×300 bp paired end sequencing. Eighteen libraries were multiplexed using standard Illumina indexing primers on an initial sequencing run. The remaining 6 libraries were sequenced on a second sequencing run. On the first run, samples were calculated to generate approximately 2 million paired end reads, with samples on the second run generating 4 million paired end reads.

2.6. Data analysis

2.6.1. Quality trimming

Quality of sequences was checked with the FastQC tool v0.10.1 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). In the metagenomic data analysis we treated both paired end reads as independent reads in order not to lose information when one of the paired end reads would be discarded during trimming. Raw reads containing unidentified nucleotides ("N") and those nucleotides adjacent to the 3' end were trimmed using Cutadapt v1.3 [26] prior to quality trimming using Sickle v1.210 (Q-score <20 and reads with length <30 bp were removed) [27]. Cutadapt was also used to trim any PCR primer sequences in samples generated using random PCR.

2.6.2. Data filtering and classification

Ribosomal RNA reads were filtered from quality trimmed data using the SortMeRNA tool v1.7 [28]. Thereafter, non-rRNA reads were mapped against the host (chicken) reference genome (Genbank accession numbers: NC_006088.3, NC_006115.3, NW_001471666.1, NC_006119.2, NC_006126.3, NC_006127.3, NC_008465.2, NC_008466.2, NC_001323.1) using BWA-MEM v.0.7.5a-r405 [29]. The unmapped reads were subjected to a megablast sequence alignment analysis against the NCBI nt database using blast-2.2.27+ software [30] employing an E value threshold of 0.001. Blast output files were processed with the MetaGenome Analyzer software MEGAN v4.70.4 [31] to classify the reads according to NCBI taxonomy. Reads were further categorized as ribosomal RNA (rRNA filtered reads), non-rRNA host (host mapped reads + unmapped reads which were classified by megablast/MEGAN as species *Gallus*) and non-rRNA bacteria, other eukaryotes, viruses (NDV & other), unclassified, and no megablast result.

2.6.2 Genome coverage

For each pretreatment protocol, raw sequences were mapped against a reference genome (GenBank accession number: AY845400.2) using GS Mapper v2.9 (Roche). The GATK v3.2 (Genome Analysis Toolkit) ‘DepthOfCoverage’ tool was applied to the BAM file to determine coverage depth at each genomic position [32].

3. Results

3.1. Quantification of nucleic acid content

The effectiveness of pretreatment protocols for RNA virus discovery in simulated clinical samples was compared (Table 1). The results from the NDV-specific quantitative RT-PCR assay indicated that different treatments had a limited impact upon yields of extracted viral RNA (in each case maximum 1 CP difference compared to controls). The only observable differences were between the Tissue 2-step and 2-step + Removal which differed by 2 CP’s). The RNA concentration in the serum control (no virion enrichment) extraction was comparable to its dsDNA concentration (S-Control) whereas the tissue sample contained approximately 7-fold higher amount of RNA present compared to dsDNA (T-Control).

The 2-step virion enrichment reduced the nucleic acid content in both serum and tissue to almost undetectable levels (S/T-Control vs. S/T-2-step). A single DNase treatment reduced the nucleic acid content to below the detection limit in serum, whereas it only marginally impacted upon the DNA/RNA yield of the tissue sample (S/T-Control vs. S/T-Control + DNase). Increased quantities of DNase enzyme/different DNase enzymes were tested unsuccessfully in an attempt to eliminate the remaining 10 ng/µl dsDNA. By contrast, the 2-step enrichment successfully removed dsDNA completely. We were unable to measure nucleic acid quantity in all rRNA depleted RNA extracts due to limited volumes, however in tissue-derived samples rRNA removal halved the RNA content (T-Control1 vs. T-Control1 + Removal, T-Control2 + DNase vs. T-Control2 + DNase + Removal).

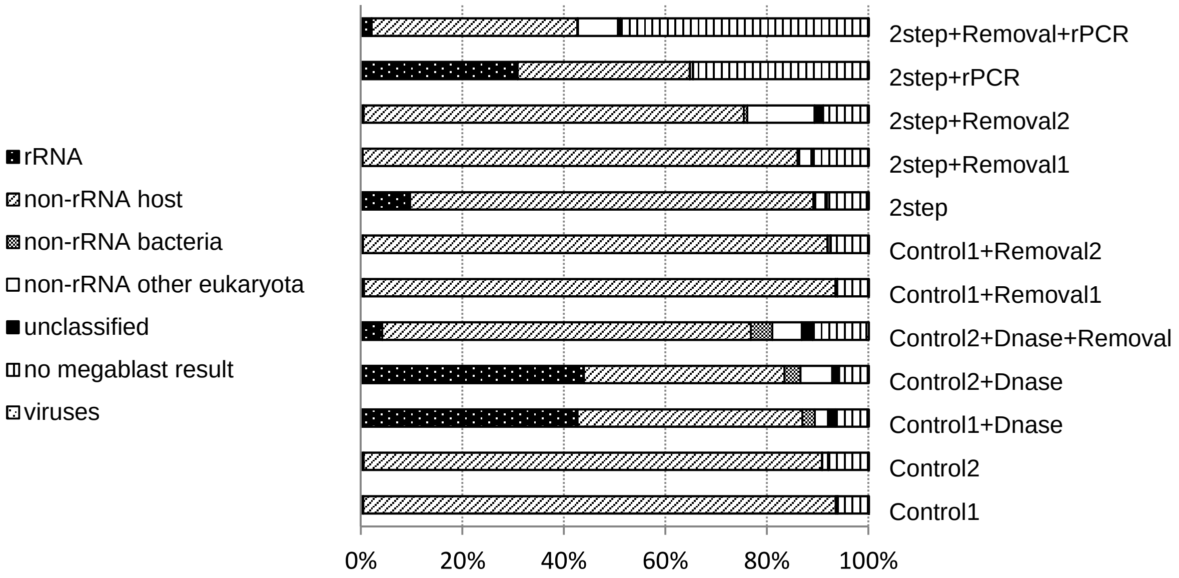
3.2. Metagenomic analysis of the different protocols

We investigated the impact of (1) 2-step virion enrichment (filtration and nuclease treatment), (2) DNase treatment on the RNA extract, (3) rRNA removal on the RNA extract, and (4) random PCR amplification, compared to direct sequencing of cDNA. In the tissue control sample, over 90 % of the reads were identified as rRNA, with the majority derived from the (chicken – *Gallus gallus domesticus*) host (Figure 2). The serum control sample contained only 0.45 % rRNA, whereas the largest class was classified as non-rRNA host (>90 % compared to 6.9 % in tissue; Figure 2). In these control samples more NDV specific reads could be identified in serum compared to tissue (Figure 3 & Table S1).

3.2.1. Effect of 2-step virion enrichment

Performing both the filtration and nuclease pretreatment improved the identification of NDV reads in the context of both serum (Figure 3 & Table S1: 17× compared to Control1; 19× compared to Control2) and tissue (Figure 3 & Table S1: 3.5× compared to Control1; 1.5× compared to Control2). This increased incidence of NDV specific reads resulted in larger overlaps of the genome (Table S1: S/T-Control vs. S/T-2-step). The distribution of taxonomic classes was similar to that observed in the control sample in tissue (Figure 2), whereas in serum the amount of non-rRNA host decreased in favor of more rRNA (9.61 %) (Figure 2). Furthermore, the amount of other (non-NDV) viral reads increased in both serum and tissue (Table S1).

(A) Serum



(B) Tissue

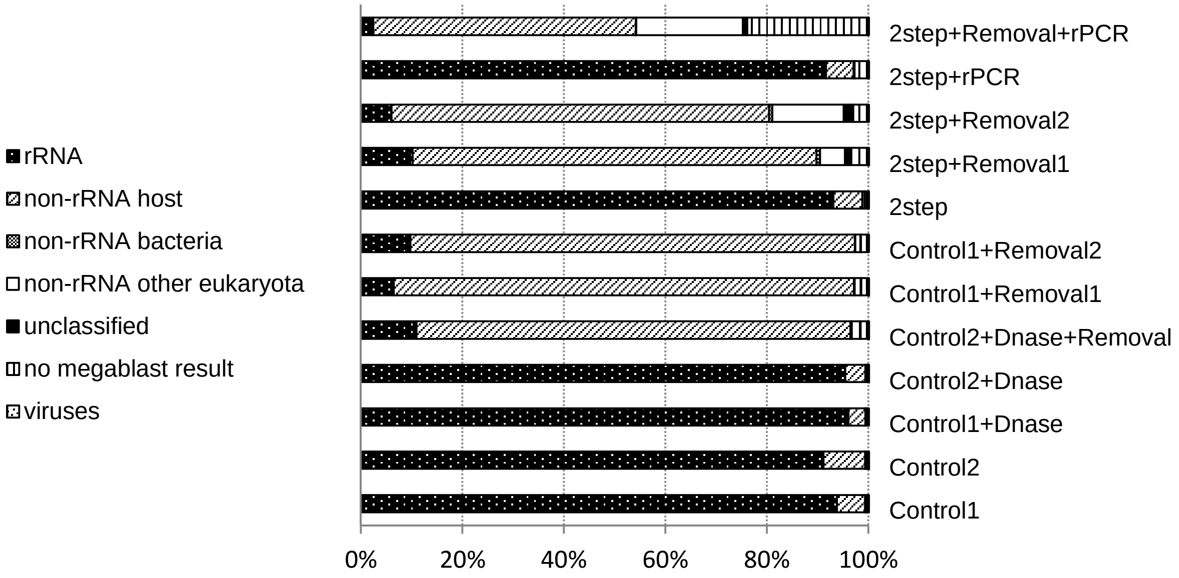


Figure 2: Taxonomic classification of sequencing output per investigated pretreatment approach in serum sample (A) and lung tissue sample (B). Control = no virion enrichment, 2-step = 2-step virion enrichment, DNase = DNase treatment on RNA extract, Removal = rRNA removal on RNA extract, rPCR = random PCR amplification, 1 and 2 are independent repetitions of either RNA extraction or rRNA removal.

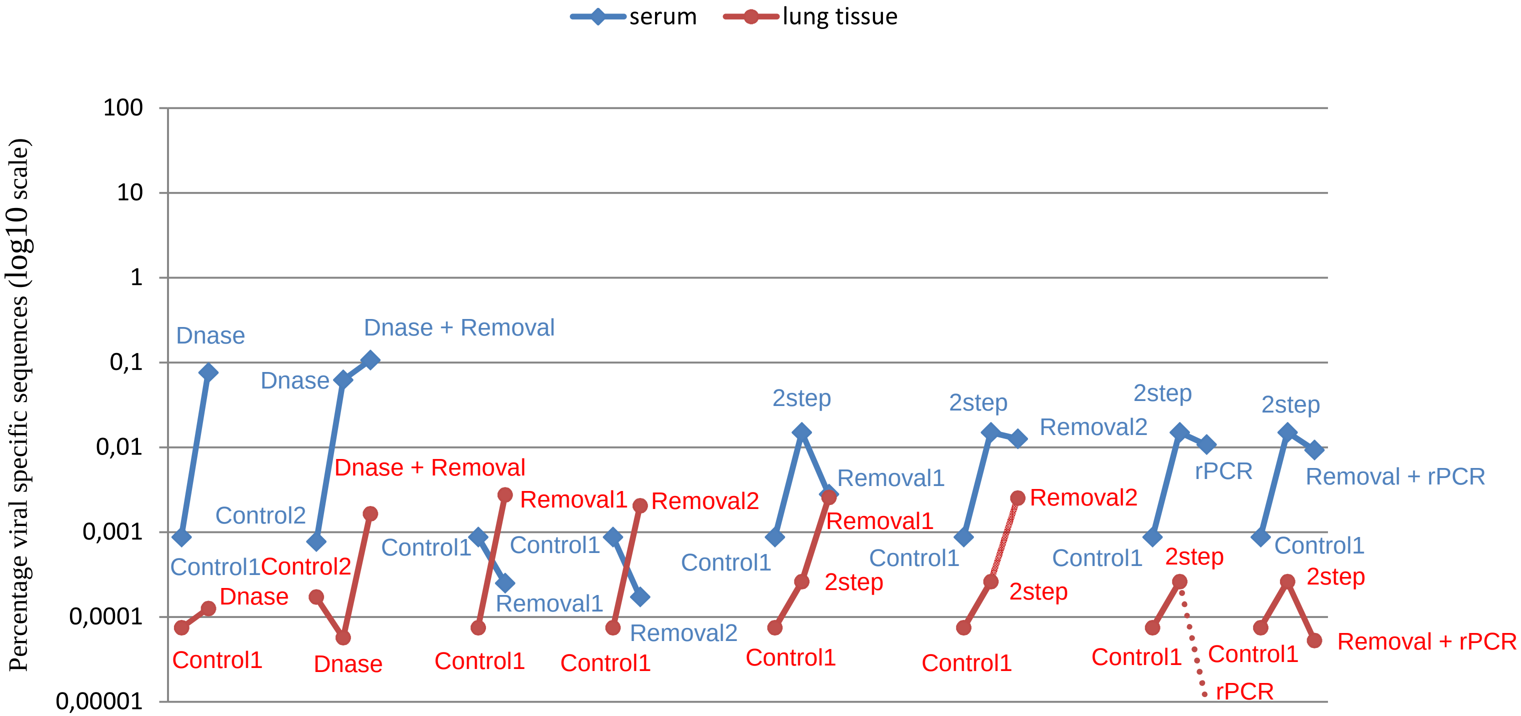


Figure 3: Proportion of viral specific reads per tested pretreatment condition. Serum (blue lines) & lung tissue (red lines) sample were spiked with 104 EID50/ml NDV. Control = no virion enrichment, 2-step = 2-step virion enrichment, DNase = DNase treatment on RNA extract, Removal = rRNA removal on RNA extract, rPCR = random PCR amplification, 1 and 2 are independent repetitions of either RNA extraction or rRNA removal.

3.2.2. Effect of DNase treatment on extracted RNA

A single DNase treatment on extracted RNA resulted in the identification of increased numbers of NDV reads in the serum-derived sample (Figure 3 & Table S1: 87× compared to Control1, 80× compared to Control2) and an increased genome coverage from 10 % to 17 % in the control samples to, respectively, 97.8 % and 99.5 % in the DNase treated samples (Table 2). In tissue samples, DNase did not improve the proportion of NDV reads (Figure 3 & Table S1: 1.69× compared to Control1, 0.33× compared to Control2), with similar trends observed for other viral reads (Table S1). In serum, non-rRNA host reads decreased significantly in favor of more rRNA reads (Figure 2: circa 42 % rRNA reads at Control-DNase compared to 0.45 % in Control). This was also true for tissue (Figure 2: circa 96 % rRNA reads at Control+DNase compared to circa 92 % in Control).

3.2.3. Effect of rRNA removal

The depletion of rRNA from serum-derived RNA extracts did not improve the detection of NDV reads (Figure 3 & Table S1: decrease of 3.5× in Control1+Removal1, 5× decrease in Control1+Removal2). Conversely, rRNA depletion of tissue RNA resulted in the identification of increased numbers of NDV reads (Figure 3 & Table S1: 37× in Control1+Removal1, 27× in Control1+Removal2) resulting in increased genome coverage (Table 2). The same trends were observed for non-NDV viral reads (Table S1). The overall distribution of taxonomic classes was not impacted in serum (Figure 2), whereas in tissue almost all rRNA has disappeared in favor of mainly non-rRNA host sequences (rRNA: 6.41 %/ 9.69 % compared to 93.77 %/ 91.14 % in control samples).

Similar trends were also observed for the rRNA depleted 2-step virion enrichment samples (Figure 2; Table S1). Ribosomal RNA removal from tissue samples decreased the amount of rRNA by 90 % (93 % in repetition), thereby increasing the number of NDV reads, ten-fold, in both duplicates (Figure 2). Similar proportions of NDV reads could be identified in both ‘Control1+Removal’ and ‘2-step+Removal’ groups (Table S1). Although in the serum samples the rRNA removal on the 2-step virion enrichment sample successfully removed all rRNA completely in both repetitions, this did not lead to increases in NDV reads. The independent repetition of ‘2-step+rRNA removal’ yielded 383 NDV reads spanning approximately 95 % of the genome (Table 2).

When performing a rRNA depletion step on previously DNase treated RNA (DNAse+removal), the amount of NDV reads doubled in the serum sample compared to the ‘DNase-only’ treated group (Figure 3 & Table S1). Using this pretreatment regimen, the entire NDV genome was obtained to an average depth of 50× (Table 3). Furthermore, we identified no evidence of biases in coverage distribution, as was introduced with rRNA depletion protocols (Figure S1, A). The tissue ‘DNAse+removal’ sample exhibited increased numbers of NDV reads compared to its no virion enrichment control group (9.6×) and the ‘DNase-only’ treated group (29×; Figure 3 & Table S1). In both serum and tissue, the depletion step reduced the numbers of rRNA reads that were present in the ‘DNase-only’ treated group (Figure 2). The efficiency of rRNA removal (Figure 2: decrease of 85 % rRNA compared to its control Tissue Control2+DNase) was similar to the efficiency of rRNA removal on No-DNase treated RNA (Figure 2: decrease of 87 % at Tissue Control1+Removal1 and 84 % at Tissue Control1+Removal2 compared to Tissue Control1).

3.2.4. Effect random PCR amplification

In both tissue and serum samples, random PCR (rPCR) amplification had a negative impact on the number of obtained NDV (and other viruses) reads compared to that of directly sequenced cDNA (Figure 3; Table S1). In the tissue rRNA depletion (2-step+Removal2+rPCR) and control (2-step+rPCR) pretreatment one and no NDV sequences could be identified respectively. In the serum rPCR groups, non-rRNA host sequences were reduced in favor of more reads, of which megablast was unable to identify any result (Figure 2: no megablast result: 34 % in 2-step+rPCR compared to 8 % in control; 49 % in 2-step+Removal2+rPCR compared to 9 % average in 2-step+Removal2). In this instance, coverage appeared unevenly distributed along the genome (Figure S1, B).

4. Discussion

Numerous protocols for viral enrichment and genome amplification have been described in literature; however, their effect on the sensitivity of viral discovery methodologies has not always been documented. Moreover, the use of different reagents, sequencing platforms, sample matrices and virus types has made the conclusions of these studies difficult to compare directly with one another. In this investigation we directly compared viral enrichment approaches to determine their effect on RNA virus discovery in spiked tissue and serum, thereby simulating clinical samples.

Although increased sequencing capacity improves the chances of virus detection [17], not every lab has the access and budget to use the massive throughput NGS platforms. Currently, three smaller benchtop-level NGS sequencers are commonly used: 454 GS Junior (Roche), Illumina MiSeq and Ion Torrent Personal Genome Machine (PGM) ([33] update http://www.molecularecologist.com/next-gen-fieldguide-2014/). The GS Junior reagents are rather expensive and the library preparation requires a minimum input of 500 ng of DNA (GS Junior Titanium series, Rapid Library Preparation Method) or 200 ng of RNA (GS Junior Titanium series, cDNA Rapid Library Preparation Method). Moreover, its throughput is limited to 100k reads per run and the platform is only supported until mid-2016. Both the MiSeq and PGM have much higher throughput and reduced cost per sequenced Mb. The MiSeq has been reported previously to be better suited to viral identification within a metagenomic context [11]. For these reasons, we chose to utilize the MiSeq in this study, although the PGM platform was also selected in a recent study evaluating a protocol for metagenomics virus detection in clinical specimens [21].

The sample matrix, in addition to sequencing throughput, has a significant impact upon method sensitivity, notably, the ability to detect small amounts of virus in high background of host nucleic acids [3]. For this reason, the detection of viral nucleic acids down to the "picogram range", within tissue samples, poses significant challenges for sequencing-based virus discovery over liquid biological samples (e.g. plasma, serum, cerebrospinal fluid, respiratory secretion). We selected serum as model sample type for liquid biological samples (with a low level of contaminating and host nucleic acids) and lung tissue as model for high background nucleic acid content. NDV Virus was artificially spiked in these samples at a constant titer of 10^4 EID50 /ml to allow comparison. Although this may represent a bias compared to infected clinical samples, this enabled an exact quantification of the impact of pretreatment protocol modifications on the amount of NDV reads obtained. We believe that our approach is a correct approximation for the clinical reality, and will allow for the quantification of differences in viral discovery sensitivity.

In this study, the DNase treatment of RNA extracted from serum exhibited the greatest increase in numbers of NDV reads (average coverage depth of 83.5× for the 2 independent repetitions). After DNase treatment we could no longer measure nucleic acids in the RNA extracts (DNA < 0.0005 ng/µl; RNA < 0.02 ng/µl). Nevertheless, quantifiable Nextera XT (Illumina) sequencing libraries could be made although the use of ultra-low nucleic acid input concentrations may be incompatible with alternative library preparation methods. DNase appeared to preferentially remove host DNA, which resulted in more rRNA, non-rRNA bacterial and non-host eukaryotic reads being sequenced. By performing an additional rRNA depletion step on the ‘DNase treated’ RNA, we were able to double the numbers of NDV specific reads permitting full genome sequence to be obtained. The removal of rRNA from non-DNAse-treated RNA from serum had no positive effect as the initial rRNA content of the control serum was already negligible. To a lesser extent, simple 2-step virion enrichment (0.45µm filtration and nuclease treatment) was also advantageous for the identification of NDV-specific reads in serum (average of 18× improvement for the 2 independent repetitions) due a significant reduction of host nucleic acids. This confirmed earlier findings of Hall and colleagues, who tested enrichment protocols in both bacterial and human cells, in recovery of DNA and RNA viruses [15]. However, we demonstrated that serum samples subjected to a DNase treatment resulted in increases in viral sequences for RNA virus identification. Furthermore, in-spite of the rRNA content of serum in the 2-step virion enrichment sample being minimal and completely removed after both rRNA depletion repetitions, we did not observe increases in NDV reads. It is probable that an additional DNase treatment on the RNA (whether or not followed by a rRNA depletion step) would improve the NDV identification power although this may adversely impact upon the cDNA synthesis. Although only one RNA virus was evaluated, the same trends were observed for the other identified viruses (Table S1). The other identified viruses consisted mainly of bacteriophages and retroviruses.

Less viral reads (NDV) were identified in the tissue samples in comparison to serum. This was expected as the background nucleic acid content was much higher, whilst the spiked virus titer was kept constant. We demonstrated that the nucleic acid background of lung tissue consisted primarily of rRNA (92 %). This corresponded with earlier reports that a significant proportion of total cellular RNA consisted of rRNA [18] and was therefore unsurprising that the rRNA removal step exhibited the greatest impact in tissue samples. In this study we demonstrated a 93 % reduction in the rRNA that was present in the tissue sample. This was highly reproducible (independent repetition removed 90 % of the rRNA) and resulted in a 37× increase of viral reads (27× in the repetition). When performing a rRNA depletion on the 2-step virion enrichment sample, the same trends were observed. Again, these findings were also observed for the other (non-NDV) viruses (Table S1). rRNA depletion pretreatments have been successfully used in transcriptomics research [34-36] and virus discovery in clinical samples [18, 21, 37, 38]. Bishop-Lilly and colleagues used a similar viral metagenomics approach combined with direct cDNA sequencing using a 454 GS FLX platform to identify Dengue virus type 1 (DENV-1, enveloped, +ssRNA 10.7 kb genome) in mosquito samples. They demonstrated rRNA depletion (using the RiboMinus Kit, Life Technologies) doubled the sensitivity of DENV-1 detection, but skewed genome coverage towards the 5’end of the genome. Kohl and colleagues used a eukaryote specific rRNA depletion kit and observed a reduction of virus nucleic acids after using the kit [21]. Our study failed to identify such biases using the Ribo-Zero Magnetic Gold Epidemiology Kit (Epicentre Technologies). It should be noted that the cost for including a rather expensive rRNA removal kit (approximately 70 € per sample at the time of writing this manuscript) should be balanced with its added value to viral discovery. An alternative approach to decrease rRNA sequences for detection of RNA viruses in clinical samples could be the use of a mixture of non-ribosomal hexanucleotides and/or rRNA blocking oligo’s as primers in the reverse transcription reaction [18, 39]. Although these non-rRNA hexamers were thoroughly tested to prime all the known mammalian viruses listed in public databases, biases may be introduced through the preferential amplification of certain sequence context, e.g. the inefficient amplification of low complexity regions. The performance of the 2-step virion enrichment or DNase treatment on extracted RNA derived from tissue also did not confer any advantage in terms of improving the virus identification power in tissue samples. After the filtration and nuclease treatment step, 93% of sequence reads were identified as rRNA (mainly host cytoplasmic 18S and 28S rRNA), indicating that the ribosomes were able to pass through the 0.45 µm filter and the rRNA stays well protected from the RNase treatment until they are released during nucleic acid extraction. Consequently, any pre-extraction viral enrichment step (e.g. ultracentrifugation, nucleases treatments, etc.) may also enrich ribosomes. Daly and colleagues tested the efficiency of a viral enrichment procedure based on repeated cycles of tissue homogenization and freeze/thaw using dry ice, followed by nuclease digestion of non-protected nucleic acids, nucleic acid extraction, rPCR amplification and NGS sequencing on an Illumina GAII sequencer [16]. Using Hepatitis C virus they tested enrichment protocols, demonstrating the resulting pretreatment regimens increased the proportion of HCV reads circa 15 times. The authors used the same sequence independent single primer random PCR amplification approach as we evaluated in this our study (rPCR). We observed that this random PCR amplification negatively impacted upon viral identification power. The observed biases in sequence distributing confirms our previous findings [40]. Moreover, rPCR amplification increased the proportion of reads for which megablast could not find similar sequences in the NCBI nt database, suggesting the creation of amplification artefacts.

Sequence independent genome amplification has been reported previously to introduce significant bias, including coverage bias and/or biased species representation [41, 42]. Several attempts have been performed to reduce amplification introduced bias [40, 43]. Although the sequencing library preparation kit (Illumina Nextera XT) used in this study is not completely PCR-free, it supports ultra-low DNA input of only a single nanogram which eliminates the need for extra sequence independent genome amplification step. Even with our lowest concentrated cDNA's (some were <0.5 pg/µl), we were able to successfully produce quantifiable libraries. Currently, PCR-free sequencing library preparation methods (e.g. Illumina TruSeq) require "microgram range" DNA input, making them incompatible with direct cDNA sequencing of certain clinical samples. Although the Nextera technology and use of low copy number template input is not completely bias-free [44-46], we demonstrated that this library preparation method can be useful for viral identification studies. Furthermore, the Nextera XT workflow is rapid and non-laborious.

In this experiment, we multiplexed different samples on a MiSeq sequencing run using default Illumina (Nextera XT) index barcodes. This allows a more economical use of the available sequencing capacity. In spite of a correct application of good laboratory practice and appropriate precautions, we identified a low level of cross contamination of a non-NDV virus that was sequenced along on the same run with a different dual index adaptor combination. The cross contamination was especially seen in libraries that share 1 index of the dual index combination of this library (data not shown). We also identified at least 0.1% carryover sequences of another non-NDV virus from previous sequencing runs. Barcode cross contamination and sample carryover seems to be an issue for the MiSeq platform (http://core-genomics.blogspot.ca/2013/04/miseq-and-2500-owners-better-read-this.html, http://seqanswers.com/forums/showthread.php?t=29110&page=2). NGS workflows are highly sensitive and can detect low amounts of contaminant nucleic acids in biological samples or reagents [47]. For virus discovery experiments this is an important issue as false conclusions could be made. Of course molecular identification of viral sequences is only the first step in determining whether or not it is associated with disease. Follow-up studies are needed to prove a link between a candidate infectious agent and disease [24, 48, 49]. Some recommendations to minimize contamination issues would be (1) alternating use of a large set of barcodes, (2) do not simultaneously make libraries in the same PCR cabinet, (3) perform the recommended maintenance wash steps of your sequencing instrument in between 2 subsequent sequencing runs, (4) avoiding multiplexing samples of different metagenomics experiments on the same sequencing run, and (5) using a different NGS platform with less carryover issues.

5. Conclusions

Metagenomics-based detection of viruses in clinical biopsy samples remains a challenge as viral nucleic acids are predominantly present at a very low ratio compared to host and contaminating nucleic acids. Our study evaluated the impact of convenient pretreatment protocols on serum and lung tissue for RNA virus discovery. We demonstrated that the efficiency, and consequently the recommended strategy, depends strongly on sample nucleic acid amount and composition. Serum samples predominantly contained host DNA, and a single DNase treatment on the RNA extract resulted in the biggest gain in proportion of identified viral reads (circa 83× improvement). A subsequent rRNA removal almost doubled this effect. Lung tissues background consisted mainly of rRNA sequences, and once again a rRNA removal resulted in the greatest gain (circa 32× improvement) in the proportion of viral sequences. We also showed that random PCR amplification was not beneficial for the viral identification, and direct sequencing of the cDNA s recommended. Some of the tested pretreatment approaches reduced nucleic acid concentration in the RNA extract to below detection limits (<20 pg/µl RNA, <0.5 pg/µl DNA). Nevertheless, our workflow allowed successfully cDNA synthesis and NGS library preparation. The present technical evaluation contributes to a more targeted virus discovery strategy. It would be interesting to test pretreatment efficiencies on a wider range of tissue sample and virus types. Moreover, in an attempt to further improve the virus detection sensitivity on clinical tissue samples, it would also be interesting to evaluate the analytical sensitivity of widely used virion concentration approaches like ultracentrifugation, ultrafiltration and PEG precipitation.

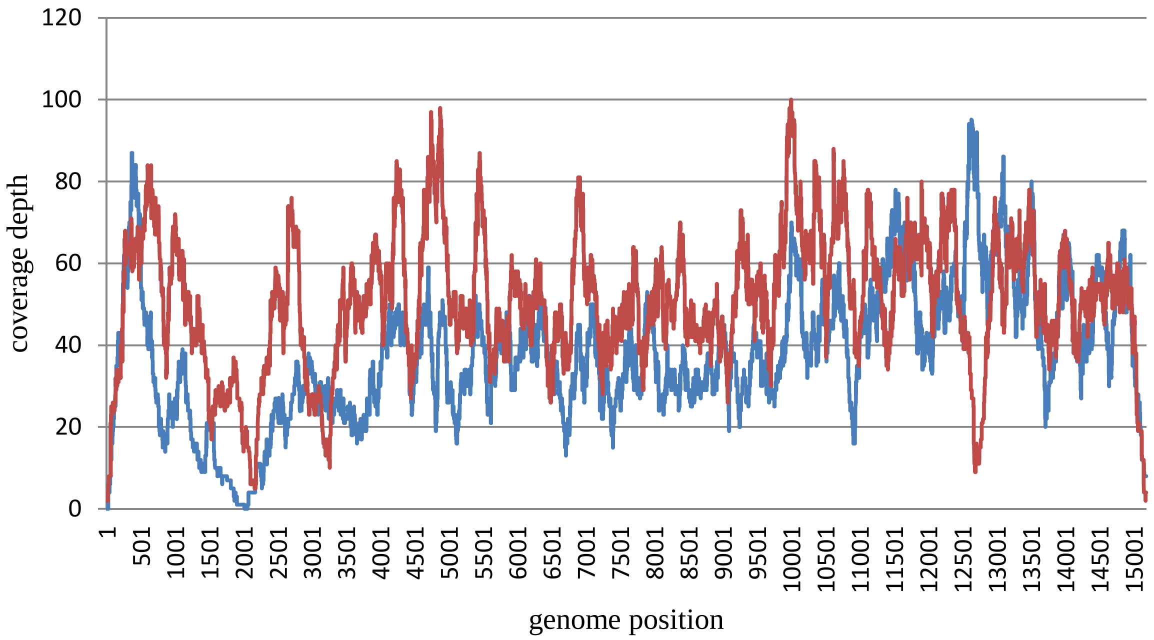
Supplementairy material

Table S1: Sequencing output and identified NDV reads per investigated condition. Serum (S) & lung tissue (T) sample were spiked with 104 EID50/ml NDV. Control = no virion enrichment, 2-step = 2-step virion enrichment, DNase = DNase treatment on RNA extract, Removal = rRNA removal on RNA extract, rPCR = random PCR amplification, 1 and 2 are independent repetitions of either RNA extraction or rRNA removal.

condition name	reads	NDV reads	% of NDV reads	% of other viral reads	condition name	reads	NDV reads	% of NDV reads	% of other viral reads
S-Control1	1,144,348	10	0.000874%	**0.000961%**	**T-Control1**	2,691,116	2	0.000074%	0.000186%
S-Control2*	3,229,212	25	0.000774%	**0.000434%**	**T-Control2***	4,676,814	8	0.000171%	0.000086%
S-Control1 + Dnase	2,072,774	1,577	0.076082%	**0.112217%**	**T-Control1 + Dnase**	3,177,102	4	0.000126%	0.000630%
S-Control2 + Dnase*	5,444,839	3,383	0.062132%	**0.111592%**	**T-Control2 + Dnase***	3,511,400	2	0.000057%	0.037051%
S-Control2 + Dnase + Removal*	4,339,431	4,603	0.106074%	**0.283539%**	**T-Control2 + Dnase + Removal***	7,995,759	131	0.001638%	0.009655%
S-Control1 + Removal1	2,410,250	6	0.000249%	**0.000622%**	**T-Control1 + Removal1**	2,985,232	82	0.002747%	0.005159%
S-Control1 + Removal2	2,325,942	4	0.000172%	**0.000516%**	**T-Control1 + Removal2**	1,966,193	40	0.002034%	0.006307%
S-2step	413,464	62	0.014995%	**0.014028%**	**T-2step**	3,456,036	9	0.000260%	0.001128%
S-2step + Removal1	1,438,088	40	0.002781%	**0.013907%**	**T-2step + Removal1**	2,737,682	70	0.002557%	0.045769%
S-2step + Removal2	3,043,698	383	0.012583%	**0.024083%**	**T-2step + Removal2**	2,147,566	54	0.002514%	0.037624%
S-2step + rPCR	1,502,810	161	0.010713%	**0.000532%**	**T-2step + rPCR**	2,251,576	0	0.000000%	0.000178%
S-2step + Removal2 + rPCR	1,525,072	141	0.009245%	**0.000656%**	**T-2step + Removal2 + rPCR**	1,903,854	1	0.000053%	0.012028%

* Conditions that were sequenced on the second MiSeq sequencing run (more details: see Section 2.5).

(A)



(B)

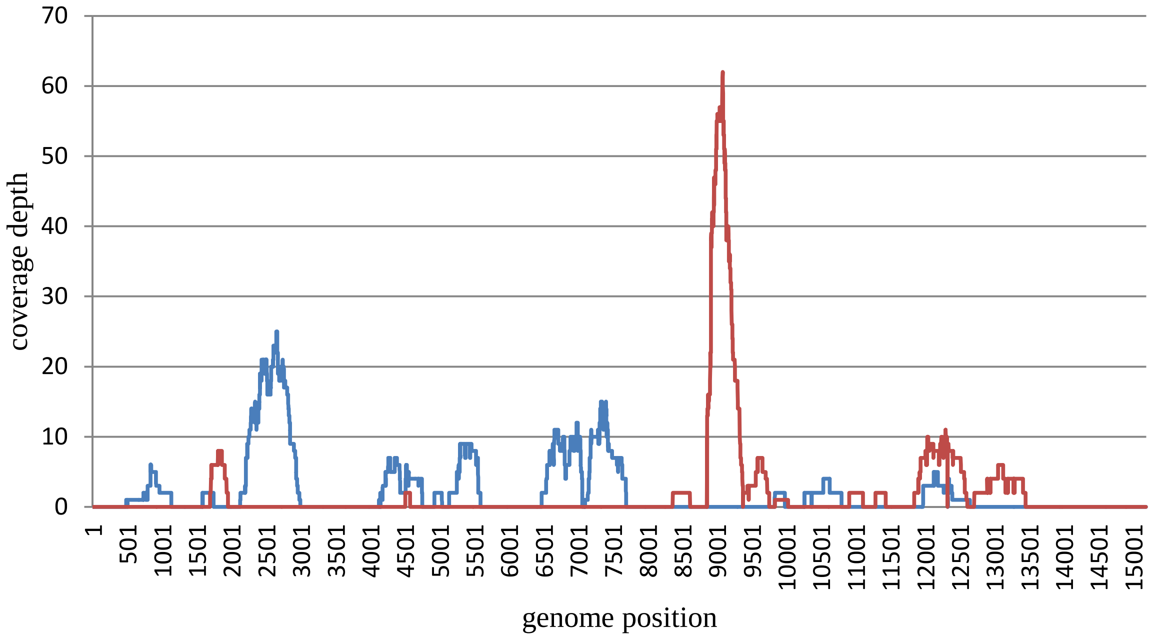


Figure S1: Depth of coverage plots for serum conditions: (A) Control2+Dnase (blue line) & Control2+DNase+Removal (red line), and (B) 2-step+rPCR (blue line) & 2-step & 2-step+Removal2+rPCR (red line). Serum was each time spiked with 104 EID50/ml NDV. Control = no virion enrichment, 2-step = 2-step virion enrichment, DNase = DNase treatment on RNA extract, Removal = rRNA removal on RNA extract, rPCR = random PCR amplification.

Acknowledgements

This study was supported by internal research and development grant RandSeq11 of the Veterinary and Agrochemical Research Center, European Union FP7 project RAPIDIA-FIELD (grant number FP7-289364), and Epi-SEQ: a transnational research project supported under the 2nd joint call for transnational research projects by EMIDA ERA-NET (FP7 project nr 219235). Cristine Kinross of Epicentre Technologies, Madison WI, USA (Illumina) kindly provided rRNA removal reagents.

References

1. Edwards, R.A. and F. Rohwer, *Viral metagenomics.* Nat Rev Microbiol, 2005. **3**(6): p. 504-10.
2. Tang, P. and C. Chiu, *Metagenomics for the discovery of novel human viruses.* Future Microbiol, 2010. **5**(2): p. 177-89.
3. Delwart, E.L., *Viral metagenomics.* Rev Med Virol, 2007. **17**(2): p. 115-31.
4. Blomstrom, A.L., *Viral metagenomics as an emerging and powerful tool in veterinary medicine.* Vet Q, 2011. **31**(3): p. 107-14.
5. Bibby, K., *Metagenomic identification of viral pathogens.* Trends Biotechnol, 2013. **31**(5): p. 275-9.
6. Belak, S., et al., *New viruses in veterinary medicine, detected by metagenomic approaches.* Vet Microbiol, 2013. **165**(1-2): p. 95-101.
7. Thurber, R.V., et al., *Laboratory procedures to generate viral metagenomes.* Nat Protoc, 2009. **4**(4): p. 470-83.
8. Allander, T., et al., *Cloning of a human parvovirus by molecular screening of respiratory tract samples.* Proc Natl Acad Sci U S A, 2005. **102**(36): p. 12891-6.
9. Djikeng, A., et al., *Viral genome sequencing by random priming methods.* BMC Genomics, 2008. **9**: p. 5.
10. Radford, A.D., et al., *Application of next-generation sequencing technologies in virology.* J Gen Virol, 2012. **93**(Pt 9): p. 1853-68.
11. Frey, K.G., et al., *Comparison of three next-generation sequencing platforms for metagenomic sequencing and identification of pathogens in blood.* BMC Genomics, 2014. **15**: p. 96.
12. Barzon, L., et al., *Applications of next-generation sequencing technologies to diagnostic virology.* Int J Mol Sci, 2011. **12**(11): p. 7861-84.
13. Rosseel, T., et al., *DNase SISPA-next generation sequencing confirms Schmallenberg virus in Belgian field samples and identifies genetic variation in Europe.* PLoS One, 2012. **7**(7): p. e41967.
14. Ambrose, H.E. and J.P. Clewley, *Virus discovery by sequence-independent genome amplification.* Rev Med Virol, 2006. **16**(6): p. 365-83.
15. Hall, R.J., et al., *Evaluation of rapid and simple techniques for the enrichment of viruses prior to metagenomic virus discovery.* J Virol Methods, 2014. **195**: p. 194-204.
16. Daly, G.M., et al., *A viral discovery methodology for clinical biopsy samples utilising massively parallel next generation sequencing.* PLoS One, 2011. **6**(12): p. e28879.
17. Cheval, J., et al., *Evaluation of high-throughput sequencing for identifying known and unknown viruses in biological samples.* J Clin Microbiol, 2011. **49**(9): p. 3268-75.

18. de Vries, M., et al., *A sensitive assay for virus discovery in respiratory clinical samples.* PLoS One, 2011. **6**(1): p. e16118.
19. Sachsenroder, J., et al., *Simultaneous identification of DNA and RNA viruses present in pig faeces using process-controlled deep sequencing.* PLoS ONE, 2012. **7**(4): p. e34631.
20. Li, L., et al., *Comparing viral metagenomics methods using a highly multiplexed human viral pathogens reagent.* J Virol Methods, 2015. **213**: p. 139-46.
21. Kohl, C., et al., *Protocol for metagenomic virus detection in clinical specimens.* Emerg Infect Dis, 2015. **21**(1): p. 48-57.
22. Jones, K.E., et al., *Global trends in emerging infectious diseases.* Nature, 2008. **451**(7181): p. 990-3.
23. Cattoli, G., et al., *Newcastle disease: a review of field recognition and current methods of laboratory detection.* J Vet Diagn Invest, 2011. **23**(4): p. 637-56.
24. Rosseel, T., et al., *False-positive results in metagenomic virus discovery: a strong case for follow-up diagnosis.* Transbound Emerg Dis, 2014. **61**(4): p. 293-9.
25. Wise, M.G., et al., *Development of a real-time reverse-transcription PCR for detection of Newcastle Disease virus RNA in clinical samples.* J. Clin. Microbiol., 2004. **42**(1): p. 329-338.
26. Martin, M., *Cutadapt removes adapter sequences from high-throughput sequencing reads.* EMBnet.journal, 2011. **17, n. 1, 10-12, may. ISSN 2226-6089.** (1).
27. Joshi NA, F.J., *Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]. Available at https://github.com/najoshi/sickle.* 2011.
28. Kopylova, E., L. Noe, and H. Touzet, *SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data.* Bioinformatics, 2012. **28**(24): p. 3211-7.
29. Li, H., *Toward better understanding of artifacts in variant calling from high-coverage samples.* Bioinformatics, 2014. **30**(20): p. 2843-51.
30. Tao, T. *Standalone BLAST Setup for Windows PC. In: BLAST® Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2008-.* 2010 [Updated 2012]; Available from: http://www.ncbi.nlm.nih.gov/books/NBK52637/.
31. Huson, D.H., et al., *Integrative analysis of environmental sequences using MEGAN4.* Genome Res, 2011. **21**(9): p. 1552-60.
32. McKenna, A., et al., *The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.* Genome Res, 2010. **20**(9): p. 1297-303.
33. Glenn, T.C., *Field guide to next-generation DNA sequencers.* Mol Ecol Resour, 2011. **11**(5): p. 759-69.
34. Lim, Y.W., et al., *Metagenomics and metatranscriptomics: windows on CF-associated viral and microbial communities.* J Cyst Fibros, 2013. **12**(2): p. 154-64.
35. He, S., et al., *Validation of two ribosomal RNA removal methods for microbial metatranscriptomics.* Nat Methods, 2010. **7**(10): p. 807-12.
36. Huang, R., et al., *An RNA-Seq strategy to detect the complete coding and non-coding transcriptome including full-length imprinted macro ncRNAs.* PLoS One, 2011. **6**(11): p. e27288.
37. Bishop-Lilly, K.A., et al., *Arbovirus detection in insect vectors by rapid, high-throughput pyrosequencing.* PLoS Negl Trop Dis, 2010. **4**(11): p. e878.
38. Moore, R.A., et al., *The sensitivity of massively parallel sequencing for detecting candidate infectious agents associated with human tissue.* PLoS One, 2011. **6**(5): p. e19838.

39. Endoh, D., et al., *Species-independent detection of RNA virus by representational difference analysis using non-ribosomal hexanucleotides for reverse transcription.* Nucleic Acids Res, 2005. **33**(6): p. e65.
40. Rosseel, T., et al., *The origin of biased sequence depth in sequence-independent nucleic acid amplification and optimization for efficient massive parallel sequencing.* PLoS One, 2013. **8**(9): p. e76144.
41. Pinard, R., et al., *Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing.* BMC Genomics, 2006. **7**: p. 216.
42. Karlsson, O.E., S. Belak, and F. Granberg, *The effect of preprocessing by sequence-independent, single-primer amplification (SISPA) on metagenomic detection of viruses.* Biosecur Bioterror, 2013. **11 Suppl 1**: p. S227-34.
43. Motley, S.T., et al., *Improved multiple displacement amplification (iMDA) and ultraclean reagents.* BMC Genomics, 2014. **15**: p. 443.
44. Marine, R., et al., *Evaluation of a transposase protocol for rapid generation of shotgun high-throughput sequencing libraries from nanogram quantities of DNA.* Appl Environ Microbiol, 2011. **77**(22): p. 8071-9.
45. Perkins, T.T., et al., *Choosing a benchtop sequencing machine to characterise Helicobacter pylori genomes.* PLoS One, 2013. **8**(6): p. e67539.
46. Parkinson, N.J., et al., *Preparation of high-quality next-generation sequencing libraries from picogram quantities of target DNA.* Genome Res, 2012. **22**(1): p. 125-33.
47. Tosar, J.P., et al., *Mining of public sequencing databases supports a non-dietary origin for putative foreign miRNAs: underestimated effects of contamination in NGS.* RNA, 2014. **20**(6): p. 754-7.
48. Lipkin, W.I., *Microbe hunting in the 21st century.* Proc Natl Acad Sci U S A, 2009. **106**(1): p. 6-7.
49. Chiu, C.Y., *Viral pathogen discovery.* Curr Opin Microbiol, 2013. **16**(4): p. 468-78.

CHAPTER 5

General discussion

Viral metagenomics, i.e. characterization of the complete viral diversity in a clinical sample or environment, is very promising for identifying viral pathogens in clinical and public health settings. Various workflows have been described in the literature and were discussed in Chapter 1. Briefly, viral metagenomic workflows can be divided into 5 main parts: sample preparation, sequence independent amplification, sequencing, data analysis and follow-up of the results. At each step challenges remain, of which some have been addressed in the presented thesis.

The first aim was to adopt a promising and widely applicable viral metagenomic workflow, which allows both sensitive virus detection and high quality full genome sequencing. Based on a thorough literature review (Chapter 1), rPCR SISPA was selected as sequence independent amplification method.

The initial viral metagenomic workflow was similar to the protocol presented by Djikeng et al. [1], and combines virion enrichment steps, rPCR SISPA amplification, molecular cloning and limited Sanger sequencing. After successful testing of different model viruses representing different genomic structures, the workflow was combined with NGS and applied to ongoing diagnostic and research projects. For the initial NGS technology, the 454 pyrosequencing technology was selected. This was because at the start of this thesis 454 pyrosequencing was the most accessible and appropriate NGS technology (at that time most alternative platforms produced insufficient read lengths) for metagenomic purposes [2]. That is why most NGS-based viral metagenomic studies use this technology (reviewed in Tables 2, 3, 5 of Chapter 1).

In Table 1 (of this Chapter) an overview is given of the different case studies that have been performed. First, the method was successfully applied on 3 case studies, using virus cell culture samples (Chapters 3.2-3.4). Besides correct viral identification, it provided (nearly) complete genome sequences with only limited sequencing efforts (i.e. amount of sequencing output per sample; Table 1). Limited sequencing efforts ensure the economic feasibility of similar sequencing projects. When no or limited sequence information is available for a studied virus, performing a metagenomic workflow on virus isolates in order to obtain a full genome sequence is less cumbersome compared to a primer walking strategy, which entails extensive rounds of primer design and PCR amplification.

Determination of full viral genomes has many advantages for outbreak management. It allows:

- Proper phylogenetic analysis (which widens the understanding of genetic diversity of these viruses)
- Pathotyping (e.g. documentation of known molecular markers of avian influenza viruses)
- Detailed molecular epidemiology analysis
- Outbreak tracing during (or after) disease outbreaks
- Development of new molecular diagnostic tests

The benefit of full viral genome sequening was clearly demonstrated by the identification of transhemispheric reassortment of genomic segments of avian influenza viruses in European gulls (Chapter 3.3).

Traditional diagnostic methods may fail to properly identify certain variants and/or co-infecting variants (as in Chapter 3.2) or unexpected viruses (as in Chapter 3.4) in a single sample.

In chapter 3.4 the aim was to characterize RNA viruses (paramyxoviruses). Beside these viruses, co-existing DNA viruses possessing circular single stranded genomes could be identified using the same workflow (Table 1). These findings demonstrate that the NGS-based metagenomic workflow is of great use for quality control of biological samples like viral stocks. This approach has also been used for quality control of vaccines [3, 4] and cell culture substrates [5-7].

In addition, NGS-based metagenomics allowed identification of genetic mutations and minority variants. For instance, in Chapter 3.4 a STOP codon in the matrix gene of the pigeon paramyxovirus genome was identified in 15 % of the viral population in one isolate. Although the exact virion concentrations were not known in all 3 virus isolate case studies, virions were well concentrated in cultured samples (often $\geq 10^7$ virions/ml). Furthermore, the allantoic fluid containing these virus isolates (Table 1), was not significantly contaminated with genetic material from cells.

The application of a NGS-based viral metagenomic workflow to samples rich in genetic material samples is much more challenging. Such samples contain limited amounts of viral nucleic acids compared to high amounts of background nucleic acids. This was clearly experienced in the SBV case study in Chapter 3.5. Using a relatively low sequencing effort

(454 technology), it was only possible to identify SBV in brain tissue samples containing high viral loads. The two samples in which SBV could be identified, contained comparable viral RNA loads (circa 7.65 $\log_{10}$ RNA copies/µl), but more SBV reads were identified in one sample compared to the other, which resulted in a larger genome coverage. This was most likely due to differences in sequencing efforts between the two samples (95,000 reads versus 25,000 reads), confirming previous findings of increased virus discovery sensitivity at larger sequencing efforts [8]. It is obvious that using NGS platforms generating higher amount of reads will increase the chance of detecting virus in clinical biopsy samples (e.g. Illumina platforms are generating millions of reads with acceptable read lengths up to 2×300 bp). Nevertheless, it was shown that, using only limited sequencing efforts (circa 60,000 reads), SBV identification was possible in brain tissue samples containing about 10^4 to 10^6 virions per ml. Although this observed sensitivity is consistent with sensitivities estimated in other studies (e.g. [1, 9]), it is extremely difficult to compare virus detection sensitivities between different case studies as this depends on many parameters. Besides the sequencing effort, sequencing platform, and sample type and condition (liquid versus tissue; targeted sample selection; stage in the viral life cycle; storage conditions), the applied virion enrichment steps, amplification method and studied virus type (genome length; RNA vs DNA genome; virion encapsidated vs integrated in host genome) will all influence the virus discovery sensitivity.

After gaining initial experience with the viral metagenomic workflow (Chapters 3.2-3.5), the workflow was further optimized and applied to a case study where presence and type of virus were unknown. A cluster of diseased dairy cattle were selected for which the cause of the disease was unknown (Chapter 3.6). Diseased animals presented high fever, anorexia, milk drop, diarrhea, redness of nose and teat, and two animals had died. All animals tested negative following different diagnostic tests for possible causative pathogens. Serum and whole blood samples were taken from an animal that started to show high fever and a few days later presented with the clinical signs described above. Because it is known that a viremia (i.e. viral particles are present in the blood) is mostly accompanied by fever, we considered that these samples were ideally optimized to target potential disease related viruses. The contigs and singletons that were classified as viruses in the viral metagenomic workflow were further analyzed. Different potential pathogens were excluded by newly developed qPCR tests on original samples and controls. This follow-up allowed us to identify a laboratory contaminant (a previously sequenced library) and a reagent contaminant (from the extraction kit used). Although no possible causal virus was discovered, this study clearly demonstrated the importance of potential contamination at different levels in metagenomic workflows, and of thorough follow-up diagnosis.

Identification of potential viral sequences is only the first step and should be compared to complementary data from serology, pathology, epidemiology, PCR prevalence studies, isolation, electron microscopy, etc. Comparing metagenomic outputs between samples of diseased and healthy animals will also help in identification of causal virus(es) of a certain disease or syndrome (e.g. [10]). Ultimately, the Koch's postulates should be fulfilled to establish a causative relationship between the viral pathogen(s) and a disease [11].

Most recently, our optimized workflow, which was based on recommendations from Chapter 4, was applied on a selection of unexplained abortion cases in cattle. Although a BVDV specific ELISA test gave negative results (based on antigen detections on the skin of the fetus), the applied metagenomic workflow allowed the identification of a type 1 BVDV in one sample. Follow-up investigations are still ongoing, but it can already be confirmed that besides the nearly full genome sequence of BVDV, the complete genome sequence of a type 1 Bovine polyomavirus was also characterized. These data confirm the sensitivity and diagnostic potential of the workflow. However, a correlation with disease was not possible and difficult given the small sample size and a context of much dispersed epidemioligy of the symptoms.

Table 1 (next page): Overview of the different case studies performed in the present thesis. Per case study sequencing effort, identified viruses, percentage genome covered and the added value of the viral metagenomic workflow are shown.

Case study	Chapter 3.2	Chapter 3.3	Chapter 3.4	Chapter 3.5	Chapter 3.6
Application	Two avian parayxoviruses (APMV) isolated during a wildlife surveillance program in mallards	9 different avian influenza (AIV) virus isolates from a surveillance program in wild gulls and shorebirds	11 pigeon paramyxoviruses (PPMV1) isolated during routine diagnostics in pigeons	5 field samples of sheep infected with a novel emerging orthobunyavirus	Undiagnosed cluster of diseased cattle showing symptoms of viral infection. Besides clear symptoms of illness, two animals died
Sample type (workflow)	Isolates -Allantoic fluid (RNA viruses)	Isolates -Allantoic fluid (RNA viruses)	Isolates -Allantoic fluid (RNA viruses)	Brain tissues (RNA viruses)	Blood and serum (DNA & RNA viruses)
Sequencing effort	5,000 - 12,000 reads per virus library; multiplexed on a GS FLX	3,300 - 12,000 reads per virus library; multiplexed on a GS FLX	Average of 45,000 reads per virus library; multiplexed on a GS FLX	Average of 59,400 reads per virus library; multiplexed on a GS FLX	32,000 reads each for the DNA & RNA virus discovery workflows; multiplexed on a GS Junior
Identified viruses (family, genome type)	APMV serotype 4 & 6 (*Paramyxoviridae*, -ssRNA)	AIV subtypes H3N8, H5N2, H6N1, H11N9, H12N5, H13N6, H13N8 and H16N3 (*Orthomyxoviridae*, -ssRNA, segmented)	All 11 samples: PPMV1 (*Paramyxoviridae*, -ssRNA) In 4 samples also: pigeon circoviruses, PiCV (*circoviridae*, circular ssDNA)	Schmallenberg virus, SBV (*Bunyaviridae*, -ssRNA, segmented)	Background: phages (dsDNA), herpesvirus-like (dsDNA), … Extraction kit contaminant: parvovirus-like (ssDNA) , circovirus-like (ssDNA) Lab contaminant: Goat pox (*Poxviridae*, dsDNA)
% genome covered	Sample1: AMPV4: 100% Sample2: AMPV4: 20%, AMPV6: 99%	Ranging from 78% to 96%	PPMV1's: mostly 92 - 100%, one sample 47.3% PiCV's: 61 - 99%	0% - 50%	n/a
Added value of approach	Identification and full genome sequencing of (unknown) APMVs; Identification of co-existing viruses	Subtyping influenza variants; Full genome sequencing of segmented viruses; Identification of transhemispheric reassortment	Full genome sequencing; Identification of genetic variations and co-infecting viruses	Viral identification in tissue samples; Illustrates importance of sample selection	Identification of contaminating viral sequences; Illustrates importance of control strategy and follow-up

Another aim of this thesis was to investigate the nature and origin of amplification biases resulting from the most promising pre-amplification method, which was the rPCR SISPA method. The rPCR SISPA amplification technique relies on a primer consisting of a 3' random part (to serve as primer in complementary DNA synthesis) and a 5' defined tag sequence (used as primer annealing site in subsequent amplification). Own experience with this method and literature review revealed an uneven sequence depth along sequenced viral genomes.

In Chapter 4.1, it was shown that the tag sequence was the main contributor to the observed uneven sequence depth. Furthermore, it was proven that this oligonucleotide annealing bias can be reduced by extending the random oligomer sequence and by combining sequence data from multiple rPCR SISPA reactions using different 5' tag sequences. However, as long as there is an amplification step used, some kind of amplification bias will remain. PCR amplification is known to produce amplification artefacts such chimeras, mutations, and heteroduplexes [12]. Karlsson and colleagues proved that direct sequencing of an unamplified viral metagenomic sample gives a more even distribution of sequencing reads along the complete genome [13].

In Table 2, the viral metagenomic studies in the veterinary field are listed which were published during the time frame of this PhD. It is notable that various studies are no longer using a pre-amplification step. This is made possible by recent advances of NGS platforms and library preparation methods which enable less input DNA for sequencing and increased feasibility of direct sequencing while still allowing sufficient read length for data analysis (reviewed in [14]; [2] update http://www.molecularecologist.com/next-gen-fieldguide-2014/).

Although truly amplification-free NGS library preparations still require considerable amounts of input material (e.g. 1 µg for Illumina TruSeq PCR-free kit), new library preparation strategies have been developed allowing ultra-low DNA input. One example is the Nextera XT kit from Illumina. Although this method is not completely PCR and bias free [15-17], it proved its usefulness in viral metagenomic studies in Chapter 4.2**.** Not only was the genome coverage much more evenly distributed compared to rPCR SISPA amplified template DNA, direct sequencing of low input DNA or cDNA was possible (< 0.5 pg/µl). Furthermore, an increased proportion of unassigned reads were found in the rPCR SISPA samples when compared to the direct cDNA sequenced samples, suggesting the creation of amplification artifacts. This is in accordance with other reports of increased unassigned reads following sequence independent amplification of DNA for NGS sequencing [13, 18]. The usefulness of rPCR SISPA for virus identification and qualitative genome sequencing was proven by means of different case studies in this thesis (Chapter 3.2-3.6). However, these new insights suggest that even better and higher quality results would be obtained in these case studies if the rPCR SISPA pre-amplification step would be omitted and direct sequencing would be performed using an illumina sequencing based NGS platform with the Nextera library preparation method.

It should be noted that some commercial available NGS library preparation kits are based on sequence independent amplification methods. For instance the Illumina ScriptSeq v2 RNA-Seq library preparation kit (Epicenter) is based on the rPCR SISPA approach and uses a random hexamer with sequence tag for production of cDNA. This sequence tag is then used in subsequent rounds of PCR amplification to directly incorporate the sequencing adaptor.

Cell culture samples containing high viral amounts and low levels of contaminating nucleic acids are the most accessible sample for sequence independent generation of full viral genomes (Table 2). However due to increased sequencing output, decreased sequencing costs and development of novel bioinformatics tools, it has also become possible to obtain full viral genomes in samples containing lower viral amounts or sample types containing more background nucleic acids like tissues (Table 2).

A typical strategy for determining full viral genomes in a random manner is the following: (1) performing a NGS-based viral metagenomics workflow to identify as much as possible viral sequences; (2) designing PCR primers to close remaining gaps in the genome sequence, and then sequencing (Sanger technology) of these PCR amplicons; (3) optionally, perfoming a RACE or similar strategy to get the genome end sequences which are are typically difficult to cover (e.g. [19]).

Even in a strong background of eukaryotic and prokaryotic DNA, viral sequences have been identified without the use of specific viral enrichment protocols (Table 2, e.g. [20]). However, in order to keep the sequencing effort limited and increase the chance of detecting viruses present at a very low level of abundance, pretreatment steps are still of great importance and can make the workflow more sensitive for virus discovery (Table 2). Nevertheless, caution should be taken because sample preparation methods can have a large impact on the composition of the characterized virome (e.g. [21]).

At the start of this PhD, not much was known about the real value of different sample preparation methods and their relative sensitivity for virus discovery. Therefore, another aim of this PhD was to investigate the value of different sample preparation methods and their relative sensitivity for virus discovery.

In Chapter 4.2, different convenient and widely used pretreatment steps for RNA virus identification in both a serum sample and lung tissue samples were compared. The Illumina MiSeq platform was selected as the NGS platform. In the past, 454 pyrosequencing platforms were the NGS platforms of choice for viral metagenomics due their long read lengths [2, 22]. However, other NGS technologies have evolved in producing longer reads and are more cost effective then 454, making them very suitable for viral metagenomics (Table 2). The Illumina MiSeq reagents with the longest possible read length of 2×300bp paired end was selected (i.e. one library DNA molecule is sequenced in both directions, keeping track of the link between the 300bp read for each strand). In addition, the bioinformatics data analysis workflow was optimized to cope with the relatively high sequencing output of the Illumina MiSeq platform (circa 44-50 million paired end reads per run using a V3 sequencing reagent kit). A paramyxovirus (-ssRNA, 15kb genome) was spiked in 2 clinical samples, serum and lung tissue, at a fixed virus concentration of 10^4 EID50/ml. In preliminary experiments with a 454 GS Junior platform (output ranges from 100,000 to 200,000 reads per run), it was found that 10^5 EID50/ml of this virus, spiked on lung tissue, was the limit of detection when using half of the sequencing capacity of this NGS platform (data not shown). The illumina MiSeq allowed us to multiplex 18 conditions on one run, and still generating circa 2 million reads per sample. The main conclusion of this comparative study was that the recommended pretreatment strategy depended strongly on the amount and composition of the sample background nucleic acids. A filtration and subsequent nuclease treatment step increased virus discovery sensitivity in both serum and tissue. Simple DNase treatment on the RNA extract resulted in the highest improvement in RNA virus discovery sensitivity in the serum sample, while an rRNA removal on the RNA extract had the biggest improvement in the tissue sample.

The tissue samples of Chapter 4.2 were before homogenization incubated in buffer containing lytic antibiotics (i.e. kills bacteria directly). We included this step because we wanted to use 0.45 µm pore size filters in the 2-step pretreatment condition in order to allow flow through of most virus types (“universal protocol”). As this filter pore size allows also flow through of small bacteria (e.g. of the genus *Mycoplasma*), we wanted to destroy these first and subsequently degrade their nucleic acids in the nuclease treatment step. It should be noted that RNases are not very effective in the degradation of rRNA. Nevertheless, our lung tissue control sample contained mainly eukaryotic rRNA (bacterial rRNA <0.04% of all rRNA reads).

During the completion of our optimization study (Chapter 4.2), two studies were published that compared the impact of alternative methods used to generate viral metagenomes [23, 24]. Li and colleagues used a low complexity biological sample (mixture of mainly cell culture supernatants and allantoic fluids) containing 25 different RNA and DNA viral pathogens [23]. Filtration and nuclease treatment were compared to a control sample without treatment, rPCR SISPA amplification was compared to no pre-amplification, and different RNA extraction kits were used. Filtration and nuclease treatment slightly decreased the number of viral reads and number of viruses identified. These results are probably a consequence of the use of a low complexity sample type as such samples contained limited background host and bacterial nucleic acids when compared to more complex clinical samples such as tissue, plasma, feces and respiratory secretions. As in Chapter 4.2, the authors also discourage the use of rPCR SISPA amplification if sufficient input DNA is available for NGS library preparation. Kohl and colleagues used a clinical sample (mixture of different organ tissues) containing 4 different viral pathogens at different concentrations to compare methods of virus purification and enrichment for metagenomic virus detection [24]. Different homogenization, filtration, nucleic acid digestion, concentration, nucleic acid extraction and random amplification methods were compared based on qPCR results to an unprocessed control extract. Subsequently, they validated their optimized workflow on the tested clinical sample and a diagnostic lung tissue sample by sequencing on the Ion Torrent PGM platform. Although they were not able to identify all 4 viruses in the test sample, a good improvement was observed compared to some conventional pretreatment and direct sequencing approaches for the Ion Torrent PGM.

The effect of different pretreatment/amplification methods seems to depend on various parameters like sample type, virus type/concentration, amount/quality of nucleic acids, etc. Evidence from this and other studies is increasingly indicating that no universal optimal viral metagenomic workflow exists. Results of protocol comparison studies (Chapter 4.2; [21, 23-25]) are therefore difficult to compare as many variable parameters exist, but they all provide useful data. Their results emphasize that metagenomic workflows have to be flexible. Depending on the case study concerned, a variety of steps may be included or excluded from the workflow. Here, I outline some recommendations which could be useful when planning future metagenomic virus discovery studies:

- The clinical sample should be selected with great care. It is advisable to have good prior knowledge about: clinical symptoms, the epidemiological context of the disease or syndrome, sampling time and type of target in order to increase the chance of finding viruses in large amounts. Sample storage conditions and/or previously performed diagnostic tests should also be considered.
- Inclusion of control samples may help to exclude possible contaminations during the workflow, or to identify the “normal virome” of the target host species. For instance, a similar sample type from a healthy animal could be processed and sequenced along with the clinical samples.
- If large sample volumes are readily available, methods for concentration can be considered. For instance, our own experience of concentration using Amicon Ultra centrifugal filters (Merck Millipore) indicated similar concentration efficiencies to ultracentrifugation. Such filters provide a convenient and efficient way to improve virus discovery sensitivity.
- Generally, virus enrichment steps like nuclease treatment work well, but are dependent on the virus target and sample type. For instance, nucleic acids not protected by a protein coat or lipid envelope may be of great importance (e.g. retrovirus integration). In such case it is recommended not to include a nuclease treatment step.
- After the nucleic acid isolation, nucleic acid concentration should be measured and quality checked with the Agilent bioanalyzer (Agilent technologies) or equivalent instruments. This will help to make decisions on whether to perform additional treatments like rRNA removal (if clear ribosomal peaks are visible on the bioanalyzer profile) or DNase treatment (if high DNA concentration is measured in the total RNA extract of the RNA virus discovery workflow) etc.
- The choice of NGS platform and library preparation method is of great importance. Recent library preparation methods allow sequencing with low inputs. However, as amplification is difficult to omit completely, amplification bias should be taken into account (coverage bias and biased representation of the different species present in sample). In the future, it will be important to critically evaluate the evolution of NGS technologies, because they are still evolving towards lower cost, greater flexibility, and increased accuracy.

- For the data analysis part, it's critical to realize that no ideal nucleic acid database exists. Databases will always be biased towards the known species and towards species which are most frequently sequenced. This will inevitably affect metagenomic analysis workflows that are based on similarity searching to know sequences.

In conclusion, this PhD thesis realized the following achievements:

- A universal and widely used diagnostic workflow based on random amplification and next generation DNA sequencing was successfully adapted, evaluated for virus diagnostics and characterization of high quality complete genome sequences through different case studies, and further optimized. Benefits and challenges were discussed.
- The nature and origin of amplification bias of the rPCR SISPA pre-amplification method was identified and solutions were proposed to minimize the bias.
- Different sample preparation methods and their relative sensitivity for virus discovery were compared. In addition, direct random-based viral genome sequencing without a preliminary pre-amplification step was evaluated. Guidelines have been provided for future metagenomic experimental design.

The list of diseases caused by viral pathogens is ever changing and growing [26]. Although metagenomic approaches pre-dates next-generation sequencing, current evolutions in NGS technologies are about to revolutionize diagnostics of infectious diseases in both human health and veterinary medicine [27-29]. Metagenomic workflows enable comprehensive characterization of all viruses present in a clinical sample. Thorough follow-up analysis is required to exclude possible false-positive results. In the future, this strategy may complement a front-line diagnostic test for infectious diseases. Currently, it is recommended to use it in parallel with conventional diagnostic tests.

The work presented in this thesis offers a clear documentation of the potential of NGS based methods for the identification of viruses and for the characterization of their complete genome sequences. Solid workflows, both for sample treatment and data analysis are now available to allow virus identification and further genome characterization.

Table 2: Review of viral metagenome studies in the veterinary field after 2011. Also studies on zoonosis vectors are included. Studies are listed according to sample type. S: sample type, C: cell culture (supernatants), R: respiratory/eye swabs, F: feces, B: blood, T: tissue, A: arthropod vectors, PT: pretreatment/viral enrichment methods were used, AM: amplification method, rPCR: random PCR, pWGA: primase-based whole genome amplification [30]**,** SPIA: single primer isothermal amplification, DOP-PCR: degenerate-oligonucleotide primer PCR [31]**,** rPCR TruSeq: variation of Illumina Truseq protocol where rPCR SISPA method is used to generate cDNA, *: novel virus or novel variant.

S	Ref	Application	Viruses (family)	PT	AM	Sequencing
C	[32]	Full genome characterization of a cytomegalovirus isolated from cynomolgus macaques	Cynomolgus macaque cytomegalovirus* (*Herpesviridae*)	yes	none	Illumina GAII
C	[33]	Full genome characterization of reference isolate of bluetongue virus serotype 16	Bluetongue virus type 16 (*Reoviridae*)	no	ligation-based SISPA	Cloning + Sanger
C	[34]	Full genome characterization of a reovirus isolated from liver of sick geese	Goose orthoreovirus (*Reoviridae*)	no	ligation-based SISPA	Cloning + Sanger
C	[35]	Full genome characterization of bluetongue virus serotype 4	Bluetongue virus subtype 4 (*Reoviridae*)	yes	ligation-based SISPA	Illumina MiSeq
C	[36]	Characterization of virus isolated from a liver tissue of a dog with idiopathic acute hepatitis	Canine adenovirus type 1 (*Adenoviridae*)	yes	VIDISCA-454	454 GS FLX
C	[37]	Characterization of a parapoxvirus ovis isolate for identification of useful ORFs	Orf virus (*Poxviridae*)	?	none	Cloning + Sanger
C	[38]	Characterization of Cotia virus isolate from mice in order to allow correct classification	Cotia virus (*Poxviridae*)	yes	none	454 GS FLX & Illumina GAIIx
C	[39]	Full genome characterization of rescued bluetongue viruses by reverse genetics	Bluetongue virus type 6, 8 (*Reoviridae*)	no	ligation-based SISPA	454 GS 20
C	[40]	Full genome sequencing of Eubenangee viruses isolated from mosquitoes and *Culicoides*	Eubenangee virus, Pata virus, Tilligerry virus (*Reoviridae*)	no	ligation-based SISPA	Cloning + Sanger
C	[41]	Full genome characterization of a viral isolate from an outbreak of disease in a collection of pythons	Sunshine virus* (*Paramyxoviridae*)	yes	none	Illumina
C	[42]	Full genome characterization of a bat virus isolated	Bat betaherpesvirus (*Herpesviridae*)	yes	none	454 GS FLX

Table 2 (continued)

S	Ref	Application	Viruses (family)	PT	AM	Sequencing
C	[43]	Genome sequencing of c owpox virus mutant lacking host range factor CP77	Cowpox virus (*Poxviridae*)	yes	none	454 GS FLX
C	[44]	Monitor virus evolution during serial passaging of the causative agent of Marek's disease in chickens	Gallid herpesvirus 2 (*Herpesviridae*)	yes	none	454 GS 20
C	[45]	Full genome characterization of a duck virus isolate	Avian Paramyxovirus type 4 (*Paramyxoviridae*)	no	rPCR SISPA	Illumina
C	[46]	Full genome characterization of a virus isolated from the brain of a civet displaying clinical signs of rabies	Ikoma lyssavirus (*Rhabdoviridae*)	yes	none	454 GS FLX
C	[47]	Full genome characterization virus isolated from pool of mosquitos	Corriparta virus (*Reoviridae*)	no	ligation-based SISPA	Cloning + Sanger
C	[48]	Full genome characterization virus isolated from tissues of bats	Fikirini bat rhabdovirus* (*Rhabdoviridae*)	yes	SPIA	Ion torrent PGM
C	[49]	Full genome characterization of two viruses isolated from turkey and pheasant presenting diarrhea	Pheasant rotavirus A, Turkey rotavirus A (*Reoviridae*)	yes	rPCR SISPA	454 GS FLX
C	[50]	Characterization of virus isolated from a dead bat	Bokeloh bat lyssavirus (*Rhabdoviridae*)	yes	none	Illumina HiSeq
C	[51]	Genome characterization of a Fowl adenovirus strain	Fowl adenovirus 5 (*Adenoviridae*)	yes	none	Illumina GAIIx
C	[52]	As a method of typing bluetongue virus isolates	Bluetongue virus (*Reoviridae*)	no	none	Illumina GAII
C	[53]	Genome haracterization of a virus in laryngeal tissue of a dead stranded dolphin	Dolphin polyomavirus 1* (*Polyomaviridae*)	yes	none	Ion torrent PGM
C	[54]	Compare sensitivity for rotavirus A detection between antigen detection kits, rt-PCR and NGS	Bovine rotavirus A (*Reoviridae*)	yes	none	Illumina MiSeq
C	[55]	Full genome characterization of a Schmallenberg virus isolated from the brain of a malformed lamb	Schmallenberg virus (*Bunyaviridae*)	yes	none	Illumina MiSeq
C	[56]	Full genome characterization to investigate a contamination event in a bluetongue virus vaccination trial	Bluetongue virus serotype 11 viruses (*Reoviridae*)	no	ligation-based SISPA	454 GS FLX

Table 2 (continued)

S	Ref	Application	Viruses (family)	PT	AM	Sequencing
C	[57-59]	Complete genome sequencing of a Bovine herpesvirus vaccine and field strains	Bovine herpesviruses 1 (*Herpesviridae*)	no	none	SOLiD & Illumina MiSeq
C	[60]	Characterization of Lone Star Virus which was isolated from ticks	Lone Star Virus* (*Bunyaviridae*)	yes	none	Illumina MiSeq
C	[61]	Genome characterization of Lyssaviruses from viral isolates (brain & cell culture)	Different lyssavirus strains (*Rhabdoviridae*)	yes	none	454 GS FLX+
C	[62]	Genome characterization of nasal isolates from an acute respiratory outbreak in a baboon colony	Simian adenovirus C* (*Adenoviridae*)	yes	none	Illumina HiSeq
C	[63]	Genome characterization of virus isolated from a green bush viper presenting with multiple disorders	Bush viper reovirus (*Reoviridae*)	yes	rPCR SISPA	Ion torrent PGM
C	[64]	Genome characterization of 2 turkey adenovirus isolates	Turkey adenovirus 4, 5 (*Adenoviridae*)	yes	none	Illumina HiSeq & GAIIx
C	[65]	Characterization of virus isolated from a deer with clinical signs of epizootic haemorrhagic disease	Mobuck virus*, Epizootic haemorrhagic disease virus (*Reoviridae*)	yes	ligation-based SISPA	Ion torrent PGM
C	[66]	Genome characterization of virus isolated in a crow presenting neurological disorders	Tvärminne avian virus (*Reoviridae*)	yes	rPCR SISPA	Ion torrent PGM
C	[67, 68]	Characterization of guinea pig cytomegalovirus isolate or directly from salivary gland of infected animals	Caviid herpesvirus 2 (*Herpesviridae*)	yes	none	Illumina & PacBio RS
C	[69]	Characterization of virus isolate, isolated from mosquitoes	Tibet orbivirus* (*Reoviridae*)	no	rPCR SISPA	454 GS FLX
C	[70]	Genome characterization of virus responsible for severe clinical distemper in dogs and wolves	Canine distemper virus (*Paramyxoviridae*)	yes	none	Ion Torrent PGM
C	[71]	Characterization of virus isolated from ticks	Hunter Island virus* (*Bunyaviridae*)	yes	yes	454 GS FLX
C	[72]	In depth investigation of isolates from a recent outbreak of a virulent bovine viral diarrhea virus	Bovine viral diarrhea virus type 2 (*Flaviviridae*)	no	none	454 GS FLX & Illumina MiSeq
C	[73]	Characterization of virus isolated from a diseased Mississippi sandhill crane	Highlands J virus (*Togaviridae*)	yes	rPCR SISPA	Illumina MiSeq

Table 2 (continued)

S	Ref	Application	Viruses (family)	PT	AM	Sequencing
C	[74]	Genome characterization of a virus isolated from a wild sooty mangabey monkey found dead	Monkeypox virus (*Poxviridae*)	no	none	Ion Torrent PGM & Illumina HiSeq
C	[75]	Genome characterization of a dolphin virus isolate	Dolphin rhabdovirus (*Rhabdoviridae*)	yes	rPCR SISPA	454 GS Junior
C	[76]	Characterization of a virus isolated from a bat	Fruit bat alphaherpesvirus 1* (*Herpesviridae*)	no	none	454 GS Junior
C	[77]	Identification of viruses isolated from bats and which cause severe disease in mice	Nairoviruses (*Bunyaviridae*)	no	rPCR SISPA	454 GS Junior
C	[78]	Genome characterization of a West Nile virus strain isolated from ticks sampled from livestock	West Nile virus (*Flaviviridae*)	yes	rPCR SISPA	454 GS FLX
C	[79]	Australian mosquitos causing cytopathic effects in mammalian cell cultures	Beaumont virus*, North Creek virus* (*Rhabdoviridae*); Murrumbidgee virus*, Salt Ash virus* (*Bunyaviridae*), Liao Ning virus, Warrego virus, Wallal virus (*Reoviridae*) …	yes	rPCR SISPA none	454 GS FLX Illumina MiSeq
C	[80]	Characterization of a virus isolated from mosquitoes	Nhumirim virus* (*Flaviviridae*)	no	rPCR SISPA-like	Illumina MiSeq
C	[81]	Full genome sequencing of a novel simian adenovirus isolated from the urine of rhesus macaques	Simian mastadenovirus D* (*Adenoviridae*)	yes	none	Illumina MiSeq
C	[82]	Studying quasispecies PRRSV variantions	Porcine reproductive and respiratory syndrome virus (*Arteriviridae*)	yes	none	Illumina MiSeq
C	[83]	Characterization of virus which causes inclusion body rhinitis in newborn piglets	Porcine cytomegalovirus (*Herpesviridae*)	yes	none	Illumina GAIIX
C	[7, 84]	Test biological reagents and model cell culture substrates for adventitious agents	Endogenous retroviruses (*Retroviridae*), Bovine viral diarrhoea virus (*Flaviviridae*), Hokovirus (*Parvoviridae*), nodavirus (*Nodaviridae*), bracovirus (*Polydnaviridae*)	yes & no	DOP-PCR	454 GS FLX & Illumina MiSeq & HiSeq
R	[85]	Eye swab samples that were obtained from semi-domesticated reindeer during an outbreak of infectious eye disease	Rangifer tarandus papillomavirus 2,3* (*Papillomaviridae*), Cervid herpesvirus 3* (*Herpesviridae*)	yes	rPCR SISPA	454 GS Junior

Table 2 (continued)

S	Ref	Application	Viruses (family)	PT	AM	Sequencing
R	[86]	Analysis for viruses in oral swab samples and 1 serum sample collected during outbreak of disease in seals	Phocine herpesvirus 7* (*Herpesviridae*), Seal anellovirus 4, 5, 6, 7* (*Anelloviridae*)	yes	rPCR SISPA	454 GS Junior
R	[87]	Monitoring wild caribou for viruses in eye and nose swabs	Rangifer tarandus granti: papillomavirus (*Papillomaviridae*), parvovirus* (*Parvoviridae*), polyomavirus (*Polymaviridae*), nidovirus* (unclassified Nidovirales)	yes	rPCR SISPA	454 GS Junior
R	[88]	Metagenomics approach in the nasal swab of young dairy cattle with symptoms of bovine respiratory disease	Bovine adenovirus-3 (*Adenoviridae*), Bovine adeno-associated virus, Bovine parvovirus 2 (*Parvoviridae*), Bovine influenza D (*Orthomyxoviridae*) , Bovine herpesvirus 6 (*Herpesviridae*), Bovine rhinitis A and B (*Picornaviridae*), Bovine astrovirus* (*Astroviridae*) and picobirnaviruses* (*Picobirnaviridae*)	yes	rPCR	Illumina HiSeq
F	[89]	Analysis of the viral flora of pine marten and European badger feces	i.a. Pine marten bocavirus* (*Parvoviridae*), Pine marten torque teno virus 1* (*Anelloviridae*), Meles meles fecal virus* (unclassified), Meles meles circovirus-like virus* (*Circoviridae*), paramyxoviridae-like	yes	rPCR SISPA	454 GS Junior
F	[90, 91]	Virome analysis for identification of novel mammalian viruses in pharyngeal and anal swab samples of bats	i.a. Bat viruses* belonging to *Herpesvirirdae, Papillomaviridae, Circoviridae, Parvoviridae, Picornaviridae, Flaviviridae, Retroviridae, Coronaviridae*	yes	rPCR SISPA	Illumina GA II
F	[92]	Analysis of the viral flora in pig feces	i.a. Pig stool-associated circular ssDNA virus*(unclassified), Pig kobuvirus, Pig enterovirus B (*Picornaviridae*), Pig rotavirus C (*Reoviridae*), Pig astrovirus (*Astroviridae*), Pig sapovirus (*Caliciviridae*), Pig picobirnavirus (*Picobirnaviridae*)	yes	rPCR SISPA	454 GS FLX
F	[93]	Metagenomic analysis of viruses from bat fecal samples	Bat: parvoviruses, AAV, feces associated picorna-like virus (*Parvoviridae*), circovirus (*Circoviridae*) …	yes	rPCR SISPA	Illumina GA

Table 2 (continued)

S	Ref	Application	Viruses (family)	PT	AM	Sequencing
F	[94]	Discovery of novel viruses in rectal swabs from bat, pigs, cattle, stray dogs, stray cats and monkeys	i.a. Miniopterus schreibersii papillomavirus type 1* (*Papillomaviridae*)	yes	pWGA	454 GS FLX
F	[95]	Analysis for causal virus of a high-mortality outbreak of enterocolitis in a colony of domestic rabbits	Rabbit astrovirus (*Astroviridae*)	yes	rPCR SISPA	Ion torrent PGM
F	[96-98]	Analysis of the viral flora in porcine fecal samples	Parechovirus-like virus PLV-CHN* (*Picornaviridae*), astrovirus MLB1 (*Astroviridae*), Porcine bocavirus* (*Parvoviridae*) …	yes	rPCR SISPA	454 GS FLX
F	[99, 100]	Analysis of the viral flora in fecal samples of wild boars	Porcine kobuvirus (*Picornaviridae*), Astrovirus wild boar (*Astroviridae*)	yes	rPCR SISPA	454 GS FLX
F	[101]	Analysis of the viral flora in feces of 13 red foxes	i.a. Fox: parvovirus*, bocavirus*, adeno-associated virus* (*Parvoviridae*), hepatitis E virus* (*Hepeviridae*), astroviruses* (*Astroviridae*), picobirnaviruses* (*Picobirnaviridae*)	yes	rPCR SISPA	454 GS FLX
F	[102]	Analysis of intestinal content of freshwater carp	Banna-like virus* (*Reoviridae*)	yes	rPCR SISPA	454 GS FLX
F	[103]	Analysis of the viral flora in diarrheal fecal samples from sick pigs	Porcine stool-associated circular virus 2, 3* (unclassified ssDNA virus)	yes	rPCR SISPA	454 GS FLX
F	[104]	Metagenomic analysis of the ferret fecal viral flora	Ferret: kobuviruses*, parechovirus* (*Picornaviridae*), papillomavirus* (*Papillomaviridae*), coronavirus (*Coronaviridae*), hepatitis E virus (*Hepeviridae*); Aleutian mink disease virus (*Parvoviridae*), Murine astrovirus STL1 (*Astroviridae*) …	yes	rPCR SISPA	454 GS Junior
F	[105]	Analysis diversity of viral flora in pooled fecal samples of diarrhoeic piglets	i.a. Porcine: stool-associated single-stranded DNA virus*, posavirus-1* (unclassified), circovirus-like virus* (*Circoviridae*), picobirnavirus*, kobuvirus (*Picobirnaviridae*), epidemic diarrhea virus, torovirus (*Coronaviridae*), sapovirus (*Caliciviridae*), bocavirus-4 (*Parvoviridae*), Torque teno sus virus 2 (*Anelloviridae*)	yes	none	Illumina HiSeq

Table 2 (continued)

S	Ref	Application	Viruses (family)	PT	AM	Sequencing
F	[106]	Metagenomic analysis of the fecal viral flora of a healthy cat	Feline: sakobuvirus A* (*Picornaviridae*), bocavirus 2*, astrovirus Viseu (*Astroviridae*), rotavirus Viseu* (*Reoviridae*) and picobirnavirus* (*Picobirnaviridae*) …	yes	rPCR SISPA	Illumina MiSeq
F	[107, 108]	Identify novel DNA viruses associated with Adélie penguins	Pygoscelis adeliae papillomavirus (*Papillomaviridae*), Adelie penguin polyomavirus (*Polyomaviridae*)	yes	RCA	Illumina HiSeq
F	[109]	Viral metagenomic analysis of 268 bat rectal swabs	Bat coronaviruses* (*Coronaviridae*)	yes	rPCR SISPA	Illumina
F	[110]	Analysis of feces samples from an epizootic outbreak of diarrhea in cows	Bovine rotavirus A* (*Reoviridae*)	yes	none	Illumina MiSeq
F	[111]	Analysis of viral flora in fecal samples collected from turkeys with light turkey syndrome	Turkey picobirnavirus (*Picobirnaviridae*)	no	none	Illumina HiSeq
F	[112]	Genetic characterization of porcine group A rotaviruses in stool samples of asymptomatic pigs	Rotaviruses A (*Reoviridae*)	yes	none	Illumina MiSeq
F	[113]	Analysis of viral flora in feces of 25 cats	i.a. Feline astroviruses* (*Astroviridae*), feline bocaviruses* (*Parvoviridae*), Feline cyclovirus* (*Circoviridae*), feline coronavirus type 1 (*Coronaviridae*), felid herpes 1 (*Gerpesviridae*)	yes	none	Illumina MiSeq
F	[114]	Analysis of viral diversity in feces of wild small carnivores	Genet fecal theilovirus* (*Picornaviridae*), Otter fecal bunyavirus*, Fox fecal bunyaviru* (*Bunyaviriae*), Red fox fecal amdovirus*, Red fox fecal kobuvirus* (*Parvovirinae*), and Red fox fecal picobirnavirus* (*Picobirnaviridae*).	yes	rPCR SISPA	454 GS Junior
F	[115]	Analysis of viral flora in feces of healthy dromedaries	Viruses of the *Picobirnaviridae, Circoviridae, Picornaviridae, Parvoviridae, Astroviridae, Hepeviridae* …	yes	rPCR SISPA	Illumina HiSeq
B	[116]	Screening for viruses in sera of domestic pigs during surveillance for African swine fever	African swine fever virus (*Asfarviridae*), Torque teno viruses (*Anelloviridae*), Ndumu virus (*Togaviridae*), P orcine endogenous retroviruses (*Retroviridae*)	yes	rPCR SISPA	454 GS FLX

Table 2 (continued)

S	Ref	Application	Viruses (family)	PT	AM	Sequencing
B	[117]	Screening for viruses in sera of wild bush pigs as part of a research project on African swine fever	Torque teno sus virus 1 & 2 * (*Anelloviridae*), Porcine parvovirus 4* (*Parvoviridae*) …	yes	rPCR SISPA	454 GS FLX
B	[118]	Screening for causal agent of Theiler's disease in serum from horses in a recent outbreak	Theiler's disease-associated virus (*Flaviviridae*)	yes	rPCR SISPA	Illumina HiSeq
B	[119, 120]	Screen ing for viruses in serum of wild living monkeys	Simian immunodeficiency viruses* (*Retroviridae*); Guereza hepacivirus* (*Flaviviridae*)	yes	none	Illumina MiSeq & 454 GS Junior
B	[121]	Pilot detection of RNA viruses in feces of dead ducks	Duck coronavirus* (*Coronaviridae*)	yes	none	Ion Torrent PGM
B	[122]	Virus discovery in a fibromatosis tumor from a female pig-tailed macaque	Retroperitoneal fibromatosis-associated herpesvirus (*Herpesviridae*)	no	none	Illumina GAII & HiSeq
B	[123]	Screening for causal viruses in serum of aborted pig fetuses	Porcine parvovirus 6 (*Parvoviridae*)	no	ligation-based SISPA	Cloning + Sanger
B	[124]	Screening for viruses in serum from a slow loris with diffuse histiocytic sarcoma	Slow loris parvovirus 1* (*Parvoviridae*)	yes	VIDISCA-454	454 GS FLX
B	[125]	Genome characterization of causal virus of viral erythrocytic necrosis in herring	Erythrocytic necrosis virus (*Iridoviridae*)	yes	none	454 GS FLX
B	[126]	Examination of simian arterivirus infections in baboon populations	Mikumi yellow baboon virus 1*, Southwest baboon virus 1* (*Arteriviridae*)	yes	none	Illumina MiSeq
T	[127]	Identify causative agent of a mass mortality event in wild and captive birds (liver and spleen samples)	Usutu virus (*Flaviviridae*)	no	none	454 GS FLX
T	[128]	Examination of tissues for candidate etiological agents for snake inclusion body disease	Golden Gate virus*, CAS virus* (*Arenaviridae*)	yes	rPCR SISPA	Illumina HiSeq
T	[129]	Examination of kidney samples of snakes that were diagnosed with inclusion body disease	Boa arenaviruses* (*Arenaviridae*)	yes	rPCR SISPA	454 GS Junior
T	[130]	Analysis of central nervous system tissue of horses that died during an outbreak of neurologic disease	Murray Valley encephalitis virus, West Nile virus (*Flaviviridae*)	yes	none	Illumina HiSeq
T	[131]	Identifying the causative agent in brain tissue samples from cattle presenting neurologic disease	Bovine astrovirus NeuroS1* (*Astroviridae*)	yes	none	Illumina MiSeq

Table 2 (continued)

S	Ref	Application	Viruses (family)	PT	AM	Sequencing
T	[132]	Examination of viral flora in the hepatopancreas of shrimps	Farfantepenaeus duorarum circovirus*(*Circoviridae*), Farfantepenaeus duorarum nodavirus* (*Nodaviridae*)	yes	MDA & rPCR SISPA	Cloning + Sanger
T	[133]	Examination for viruses in liver of a dog with disease of unknown cause	Dog circovirus* (*Circoviridae*)	yes	rPCR SISPA	454 GS FLX
T	[134]	Examination for viruses in liver of a dog with disease of unknown cause	Canine bocavirus 3* (*Parvoviridae*)	yes	rPCR SISPA none	454 GS FLX Illumina MiSeq
T	[135]	Characterization of new virus found in the cloacal bursa in dead gulls	Gull adenovirus* (*Adenoviridae*)	yes	rPCR SISPA	454 GS Junior
T	[136]	Identification of viruses in tissue of a harbor seal with chronic non-suppurative meningo-encephalitis	Seal parvovirus* (*Parvoviridae*), Seal anelloviruses 2, 3 * (*Anelloviridae*)	yes	rPCR SISPA	454 GS Junior
T	[137]	Metagenomic study of bat viruses in urine, throat swabs & lung tissue	i.a. Eidolon helvum: kobuvirus* (*Picornaviridae*), parvoviruses* (*Parvoviridae*), poxvirus* (*Poxviridae*), herpesviruses*(*Herpesviridae*), adenoviriruses* (*Adenoviridae*), papillomaviruses* (*Papillomaviridae*), retroviruses* (*Retroviridae*)	yes	rPCR SISPA	Illumina GAII
T	[138, 139]	Study virome in bat tissues	i.a. Bat: astroviruses* (*Astroviridae*), bocaviruses* (*Parvoviridae*), circoviruses* (*Circoviridae*), Iflaviruses* (*Iflaviridae*), orthohepadnaviruses* (*Hepadnaviridae*) , Rotavirus A* (*Reoviridae*)	yes	rPCR SISPA	Illumina
T	[140]	Virus screening in postmortem tissue (lung & lever) samples from rabies-negative big brown bats	American bat vesiculovirus* (*Rhabdoviridae*)	yes	(MDA +) rPCR SISPA	454 & Illumina
T	[141]	Characterization of viruses in fibroma tissue on the trunk of an African elephant	African elephant polyomavirus 1* (*Polyomaviridae*)	no	RCA	454 GS FLX
T	[142]	Virus screening in tissue samples for rabies-negative dead bats	Bat bornaviruses* (*Bornaviridae*), Bat nairovirus* (*Bunyaviridae*), Bat rotavirus* (*Reoviridae*), Bat gammaretrovirus* (*Retroviridae*) …	no	rPCR SISPA	Illumina HiSeq

Table 2 (continued)

S	Ref	Application	Viruses (family)	PT	AM	Sequencing
T	[143]	Characterization of virus associated with disease outbreak in pigs (intestine homogenate + 2 isolates)	Porcine epidemic diarrhea virus (*Coronaviridae*)	no	none	Illumina MiSeq
T	[144]	Characterization of causal agent (in intestinal content) of an outbreak of fulminating enteritis in guinea fowl	Guinea fowl coronavirus (*Coronaviridae*)	yes	rPCR SISPA	Illumina MiSeq
T	[145]	Characterization of virus associated with sick raccoon dogs in lesion tissues of spleens and kidneys	Raccoon dog and fox amdoparvovirus* (*Parvoviridae*)	?	rPCR SISPA	Illumina MiSeq
T	[21]	RNA viral community in a gut tissue of a healthy turkey	Viruses belonging to *Reoviridae*, *Astroviridae*, *Bunyaviridae*, *Paramyxoviridae*, *Rhabdoviridae* …	yes	none	Illumina MiSeq
T	[20, 146]	Virus screening for etiology agent associated with a fatal respiratory disease in pythons	Ball python nidovirus* (unclassified Nidovirales)	no yes	none	Illumina HiSeq & Ion torrent PGM
T	[147, 148]	Complete viral genome sequencing from heart/tongue tissue of a diseased elephants (fatal case)	Elephant endotheliotropic herpesvirus 1A, 1B, 5 (*Herpesviridae*)	no	none	Illumina HiSeq & MiSeq
T	[149]	Virus screening in papilloma lesions of stranded sea otters	Enhydra lutris papillomavirus 1* (*Papillomaviridae*)	yes	rPCR SISPA	Illumina MiSeq
T	[10]	Virus screening in tissue from chickens with transmissible viral proventriculitis disease	Chicken proventriculitis virus *, Avian encephalomyelitis virus (*Picornaviridae*), Chicken anemia virus (*Circoviridae*), parvovirus (*Parvoviridae*), avastrovirus (*Astroviridae*), calicivirus (*Caliciviridae*), adenovirus (*Adenoviridae*)	yes	rPCR SISPA	Illumina
A	[150]	Identification of causal viruses in a diseased Spanish honeybee colony	i.a. Aphid lethal paralysis virus, Israel acute paralysis virus* (*Dicistroviridae*), Lake sinai viruses (unclassified)	yes	rPCR SISPA	454 GS FLX
A	[151]	Proof of principle – detecting viruses in infected mosquitoes	Dengue virus, Yellow fever virus (*Flaviviridae*), Chikungunya virus (*Togaviridae*)	no	rPCR SISPA MDA	Ion Torrent PGM
A	[152]	Examine viromes of ticks	Long Island tick rhabdovirus (*Rhabdoviridae*)	yes	none	Ion Torrent PGM
A	[153]	Metagenomics sequencing in wild-caught *Aedes aegypti*	Phasi Charoen-like virus (*Bunyaviridae*)	yes	none	Illumina GAII

References

1. Djikeng, A., et al., *Viral genome sequencing by random priming methods.* BMC Genomics, 2008. **9**: p. 5.
2. Glenn, T.C., *Field guide to next-generation DNA sequencers.* Mol Ecol Resour, 2011. **11**(5): p. 759-69.
3. Victoria, J.G., et al., *Viral nucleic acids in live-attenuated vaccines: detection of minority variants and an adventitious virus.* J Virol, 2010. **84**(12): p. 6033-40.
4. Switzer, W.M., et al., *No evidence of murine leukemia virus-related viruses in live attenuated human vaccines.* PLoS ONE, 2011. **6**(12): p. e29223.
5. Onions, D. and J. Kolman, *Massively parallel sequencing, a new method for detecting adventitious agents.* Biologicals, 2010. **38**(3): p. 377-80.
6. Onions, D., et al., *Ensuring the safety of vaccine cell substrates by massively parallel sequencing of the transcriptome.* Vaccine, 2011. **29**(41): p. 7117-21.
7. McClenahan, S.D., C. Uhlenhaut, and P.R. Krause, *Evaluation of cells and biological reagents for adventitious agents using degenerate primer PCR and massively parallel sequencing.* Vaccine, 2014. **32**(52): p. 7115-21.
8. Cheval, J., et al., *Evaluation of high-throughput sequencing for identifying known and unknown viruses in biological samples.* J Clin Microbiol, 2011. **49**(9): p. 3268-75.
9. Moore, R.A., et al., *The sensitivity of massively parallel sequencing for detecting candidate infectious agents associated with human tissue.* PLoS One, 2011. **6**(5): p. e19838.
10. Kim, H.R., et al., *Identification of a picornavirus from chickens with transmissible viral proventriculitis using metagenomic analysis.* Arch Virol, 2015. **160**(3): p. 701-9.
11. Fredricks, D.N. and D.A. Relman, *Sequence-based identification of microbial pathogens: a reconsideration of Koch's postulates.* Clin Microbiol Rev, 1996. **9**(1): p. 18-33.
12. Qiu, X., et al., *Evaluation of PCR-generated chimeras, mutations, and heteroduplexes with 16S rRNA gene-based cloning.* Appl Environ Microbiol, 2001. **67**(2): p. 880-7.
13. Karlsson, O.E., S. Belak, and F. Granberg, *The effect of preprocessing by sequence-independent, single-primer amplification (SISPA) on metagenomic detection of viruses.* Biosecur Bioterror, 2013. **11 Suppl 1**: p. S227-34.
14. Head, S.R., et al., *Library construction for next-generation sequencing: overviews and challenges.* BioTechniques, 2014. **56**(2): p. 61-4, 66, 68, passim.
15. Marine, R., et al., *Evaluation of a transposase protocol for rapid generation of shotgun high-throughput sequencing libraries from nanogram quantities of DNA.* Appl Environ Microbiol, 2011. **77**(22): p. 8071-9.
16. Parkinson, N.J., et al., *Preparation of high-quality next-generation sequencing libraries from picogram quantities of target DNA.* Genome Res, 2012. **22**(1): p. 125-33.
17. Perkins, T.T., et al., *Choosing a benchtop sequencing machine to characterise Helicobacter pylori genomes.* PLoS One, 2013. **8**(6): p. e67539.
18. Malboeuf, C.M., et al., *Complete viral RNA genome sequencing of ultra-low copy samples by sequence-independent amplification.* Nucleic Acids Res, 2013. **41**(1): p. e13.
19. Alfson, K.J., M.W. Beadles, and A. Griffiths, *A new approach to determining whole viral genomic sequences including termini using a single deep sequencing run.* J Virol Methods, 2014. **208**: p. 1-5.
20. Stenglein, M.D., et al., *Ball python nidovirus: a candidate etiologic agent for severe respiratory disease in Python regius.* MBio, 2014. **5**(5): p. e01484-14.

21. Shah, J.D., et al., *Comparison of tissue sample processing methods for harvesting the viral metagenome and a snapshot of the RNA viral community in a turkey gut.* J Virol Methods, 2014. **209**: p. 15-24.
22. Wommack, K.E., J. Bhavsar, and J. Ravel, *Metagenomics: read length matters.* Appl Environ Microbiol, 2008. **74**(5): p. 1453-63.
23. Li, L., et al., *Comparing viral metagenomics methods using a highly multiplexed human viral pathogens reagent.* J Virol Methods, 2015. **213**: p. 139-46.
24. Kohl, C., et al., *Protocol for metagenomic virus detection in clinical specimens.* Emerg Infect Dis, 2015. **21**(1): p. 48-57.
25. Hall, R.J., et al., *Evaluation of rapid and simple techniques for the enrichment of viruses prior to metagenomic virus discovery.* J Virol Methods, 2014. **195**: p. 194-204.
26. Fauci, A.S. and D.M. Morens, *The perpetual challenge of infectious diseases.* N Engl J Med, 2012. **366**(5): p. 454-61.
27. Miller, R.R., et al., *Metagenomics for pathogen detection in public health.* Genome Med, 2013. **5**(9): p. 81.
28. Temmam, S., et al., *Viral metagenomics on animals as a tool for the detection of zoonoses prior to human infection?* Int J Mol Sci, 2014. **15**(6): p. 10377-97.
29. Van Borm, S., et al., *Next-generation sequencing in veterinary medicine: how can the massive amount of information arising from high-throughput technologies improve diagnosis, control, and management of infectious diseases?* Methods Mol Biol, 2015. **1247**: p. 415-36.
30. Li, Y., et al., *Primase-based whole genome amplification.* Nucleic Acids Res, 2008. **36**(13): p. e79.
31. Nanda, S., et al., *Universal virus detection by degenerate-oligonucleotide primed polymerase chain reaction of purified viral nucleic acids.* J Virol Methods, 2008. **152**(1-2): p. 18-24.
32. Marsh, A.K., et al., *Genomic sequencing and characterization of cynomolgus macaque cytomegalovirus.* J Virol, 2011. **85**(24): p. 12995-3009.
33. Maan, S., et al., *Complete genome sequence analysis of a reference strain of bluetongue virus serotype 16.* J Virol, 2012. **86**(18): p. 10255-6.
34. Yun, T., et al., *Complete genomic sequence of goose-origin reovirus from China.* J Virol, 2012. **86**(18): p. 10257.
35. Yang, H., et al., *Full genome sequence of bluetongue virus serotype 4 from China.* J Virol, 2012. **86**(23): p. 13122-3.
36. van der Heijden, M., et al., *Sequence-independent VIDISCA-454 technique to discover new viruses in canine livers.* J Virol Methods, 2012. **185**(1): p. 152-5.
37. McGuire, M.J., S.A. Johnston, and K.F. Sykes, *Novel immune-modulator identified by a rapid, functional screen of the parapoxvirus ovis (Orf virus) genome.* Proteome Sci, 2012. **10**(1): p. 4.
38. Afonso, P.P., et al., *Biological characterization and next-generation genome sequencing of the unclassified Cotia virus SPAn232 (Poxviridae).* J Virol, 2012. **86**(9): p. 5039-54.
39. van Gennip, R.G., et al., *Rescue of recent virulent and avirulent field strains of bluetongue virus by reverse genetics.* PLoS ONE, 2012. **7**(2): p. e30540.
40. Belaganahalli, M.N., et al., *Full genome sequencing and genetic characterization of Eubenangee viruses identify Pata virus as a distinct species within the genus Orbivirus.* PLoS ONE, 2012. **7**(3): p. e31911.
41. Hyndman, T.H., et al., *Isolation and molecular identification of Sunshine virus, a novel paramyxovirus found in Australian snakes.* Infect Genet Evol, 2012. **12**(7): p. 1436-46.

42. Zhang, H., et al., *A novel bat herpesvirus encodes homologues of major histocompatibility complex classes I and II, C-type lectin, and a unique family of immune-related genes.* J Virol, 2012. **86**(15): p. 8014-30.
43. Schuenadel, L., B.K. Tischer, and A. Nitsche, *Generation and characterization of a Cowpox virus mutant lacking host range factor CP77.* Virus Res, 2012. **168**(1-2): p. 23-32.
44. Spatz, S.J., et al., *Dynamic equilibrium of Marek's disease genomes during in vitro serial passage.* Virus Genes, 2012. **45**(3): p. 526-36.
45. Abolnik, C., M. de Castro, and J. Rees, *Full genomic sequence of an African avian paramyxovirus type 4 strain isolated from a wild duck.* Virus Genes, 2012. **45**(3): p. 537-41.
46. Marston, D.A., et al., *Complete genome sequence of Ikoma lyssavirus.* J Virol, 2012. **86**(18): p. 10242-3.
47. Belaganahalli, M.N., et al., *Full genome sequencing of Corriparta virus, identifies California mosquito pool virus as a member of the Corriparta virus species.* PLoS ONE, 2013. **8**(8): p. e70779.
48. Kading, R.C., et al., *Isolation and molecular characterization of Fikirini rhabdovirus, a novel virus from a Kenyan bat.* J Gen Virol, 2013. **94**(Pt 11): p. 2393-8.
49. Trojnar, E., et al., *Identification of an avian group A rotavirus containing a novel VP4 gene with a close relationship to those of mammalian rotaviruses.* J Gen Virol, 2013. **94**(Pt 1): p. 136-42.
50. Picard-Meyer, E., et al., *Isolation of Bokeloh bat lyssavirus in Myotis nattereri in France.* Arch Virol, 2013. **158**(11): p. 2333-40.
51. Marek, A., et al., *The first whole genome sequence of a Fowl adenovirus B strain enables interspecies comparisons within the genus Aviadenovirus.* Vet Microbiol, 2013. **166**(1-2): p. 250-6.
52. Rao, P.P., et al., *Deep sequencing as a method of typing bluetongue virus isolates.* J Virol Methods, 2013. **193**(2): p. 314-9.
53. Anthony, S.J., et al., *Identification of a novel cetacean polyomavirus from a common dolphin (Delphinus delphis) with Tracheobronchitis.* PLoS ONE, 2013. **8**(7): p. e68239.
54. Minami-Fukuda, F., et al., *Detection of bovine group a rotavirus using rapid antigen detection kits, rt-PCR and next-generation DNA sequencing.* J Vet Med Sci, 2013. **75**(12): p. 1651-5.
55. Hulst, M., et al., *Genetic characterization of an atypical Schmallenberg virus isolated from the brain of a malformed lamb.* Virus Genes, 2013. **47**(3): p. 505-14.
56. Vandenbussche, F., et al., *Full-Genome Sequencing of Four Bluetongue Virus Serotype 11 Viruses.* Transbound Emerg Dis, 2013.
57. d'Offay, J.M., R.W. Fulton, and R. Eberle, *Complete genome sequence of the NVSL BoHV-1.1 Cooper reference strain.* Arch Virol, 2013. **158**(5): p. 1109-13.
58. Fulton, R.W., J.M. d'Offay, and R. Eberle, *Bovine herpesvirus-1: comparison and differentiation of vaccine and field strains based on genomic sequence variation.* Vaccine, 2013. **31**(11): p. 1471-9.
59. Fulton, R.W., et al., *Bovine herpesvirus-1: evaluation of genetic diversity of subtypes derived from field strains of varied clinical syndromes and their relationship to vaccine strains.* Vaccine, 2015. **33**(4): p. 549-58.
60. Swei, A., et al., *The genome sequence of Lone Star virus, a highly divergent bunyavirus found in the Amblyomma americanum tick.* PLoS ONE, 2013. **8**(4): p. e62083.

61. Marston, D.A., et al., *Next generation sequencing of viral RNA genomes.* BMC Genomics, 2013. **14**: p. 444.
62. Chiu, C.Y., et al., *A novel adenovirus species associated with an acute respiratory outbreak in a baboon colony and evidence of coincident human infection.* MBio, 2013. **4**(2): p. e00084.
63. Banyai, K., et al., *Whole-genome sequencing of a green bush viper reovirus reveals a shared evolutionary history between reptilian and unusual mammalian orthoreoviruses.* Arch Virol, 2014. **159**(1): p. 153-8.
64. Marek, A., et al., *Whole-genome sequences of two turkey adenovirus types reveal the existence of two unknown lineages that merit the establishment of novel species within the genus Aviadenovirus.* J Gen Virol, 2014. **95**(Pt 1): p. 156-70.
65. Cooper, E., et al., *Mobuck virus genome sequence and phylogenetic analysis: identification of a novel Orbivirus isolated from a white-tailed deer in Missouri, USA.* J Gen Virol, 2014. **95**(Pt 1): p. 110-6.
66. Dandar, E., et al., *Complete genome analysis identifies Tvarminne avian virus as a candidate new species within the genus Orthoreovirus.* J Gen Virol, 2014. **95**(Pt 4): p. 898-904.
67. Yang, D., et al., *Complete genome sequence of pathogenic Guinea pig cytomegalovirus from salivary gland homogenates of infected animals.* Genome Announc, 2013. **1**(2): p. e0005413.
68. Schleiss, M.R., et al., *Molecular and biological characterization of a new isolate of guinea pig cytomegalovirus.* Viruses, 2014. **6**(2): p. 448-75.
69. Li, M., et al., *Tibet Orbivirus, a novel Orbivirus species isolated from Anopheles maculatus mosquitoes in Tibet, China.* PLoS ONE, 2014. **9**(2): p. e88738.
70. Marcacci, M., et al., *Whole genome sequence analysis of the arctic-lineage strain responsible for distemper in Italian wolves and dogs through a fast and robust next generation sequencing protocol.* J Virol Methods, 2014. **202**: p. 64-8.
71. Wang, J., et al., *Novel phlebovirus with zoonotic potential isolated from ticks, Australia.* Emerg Infect Dis, 2014. **20**(6): p. 1040-3.
72. Jenckel, M., et al., *Mixed triple: allied viruses in unique recent isolates of highly virulent type 2 bovine viral diarrhea virus detected by deep sequencing.* J Virol, 2014. **88**(12): p. 6983-92.
73. Ip, H.S., et al., *Identification and characterization of Highlands J virus from a Mississippi sandhill crane using unbiased next-generation sequencing.* J Virol Methods, 2014. **206**: p. 42-5.
74. Radonic, A., et al., *Fatal monkeypox in wild-living sooty mangabey, Cote d'Ivoire, 2012.* Emerg Infect Dis, 2014. **20**(6): p. 1009-11.
75. Siegers, J.Y., et al., *Genetic relatedness of dolphin rhabdovirus with fish rhabdoviruses.* Emerg Infect Dis, 2014. **20**(6): p. 1081-2.
76. Sasaki, M., et al., *Isolation and characterization of a novel alphaherpesvirus in fruit bats.* J Virol, 2014. **88**(17): p. 9819-29.
77. Ishii, A., et al., *A nairovirus isolated from African bats causes haemorrhagic gastroenteritis and severe hepatic disease in mice.* Nat Commun, 2014. **5**: p. 5651.
78. Lwande, O., et al., *Whole genome phylogenetic investigation of a West Nile virus strain isolated from a tick sampled from livestock in north eastern Kenya.* Parasit Vectors, 2014. **7**(1): p. 542.
79. Coffey, L.L., et al., *Enhanced arbovirus surveillance with deep sequencing: Identification of novel rhabdoviruses and bunyaviruses in Australian mosquitoes.* Virology, 2014. **448**: p. 146-58.

80. Pauvolid-Correa, A., et al., *Nhumirim virus, a novel flavivirus isolated from mosquitoes from the Pantanal, Brazil.* Arch Virol, 2015. **160**(1): p. 21-7.
81. Malouli, D., et al., *Full genome sequence analysis of a novel adenovirus of rhesus macaque origin indicates a new simian adenovirus type and species.* Virol Rep, 2014. **3-4**: p. 18-29.
82. Lu, Z.H., et al., *Genomic variation in macrophage-cultured European porcine reproductive and respiratory syndrome virus Olot/91 revealed using ultra-deep next generation sequencing.* Virol J, 2014. **11**: p. 42.
83. Gu, W., et al., *Genomic organization and molecular characterization of porcine cytomegalovirus.* Virology, 2014. **460-461**: p. 165-72.
84. McClenahan, S.D., C. Uhlenhaut, and P.R. Krause, *Optimization of virus detection in cells using massively parallel sequencing.* Biologicals, 2014. **42**(1): p. 34-41.
85. Smits, S.L., et al., *Identification and characterization of two novel viruses in ocular infections in reindeer.* PLoS ONE, 2013. **8**(7): p. e69711.
86. Bodewes, R., et al., *Identification of DNA sequences that imply a novel gammaherpesvirus in seals.* J Gen Virol, 2014.
87. Schurch, A.C., et al., *Metagenomic survey for viruses in Western Arctic caribou, Alaska, through iterative assembly of taxonomic units.* PLoS ONE, 2014. **9**(8): p. e105227.
88. Ng, T.F., et al., *A metagenomics and case-control study to identify viruses associated with bovine respiratory disease.* J Virol, 2015.
89. van den Brand, J.M., et al., *Metagenomic analysis of the viral flora of pine marten and European badger feces.* J Virol, 2012. **86**(4): p. 2360-5.
90. Wu, Z., et al., *Virome analysis for identification of novel mammalian viruses in bat species from Chinese provinces.* J Virol, 2012. **86**(20): p. 10999-1012.
91. Yang, L., et al., *Novel SARS-like betacoronaviruses in bats, China, 2011.* Emerg Infect Dis, 2013. **19**(6): p. 989-91.
92. Sachsenroder, J., et al., *Simultaneous identification of DNA and RNA viruses present in pig faeces using process-controlled deep sequencing.* PLoS ONE, 2012. **7**(4): p. e34631.
93. Ge, X., et al., *Metagenomic analysis of viruses from bat fecal samples reveals many novel viruses in insectivorous bats in China.* J Virol, 2012. **86**(8): p. 4620-30.
94. Tse, H., et al., *Identification of a novel bat papillomavirus by metagenomics.* PLoS ONE, 2012. **7**(8): p. e43986.
95. Stenglein, M.D., et al., *Complete genome sequence of an astrovirus identified in a domestic rabbit (Oryctolagus cuniculus) with gastroenteritis.* Virol J, 2012. **9**: p. 216.
96. Yu, J.M., et al., *Identification of a novel picornavirus in healthy piglets and seroepidemiological evidence of its presence in humans.* PLoS ONE, 2013. **8**(8): p. e70137.
97. Yu, J.M., et al., *Detection of novel viruses in porcine fecal samples from China.* Virol J, 2013. **10**: p. 39.
98. Yang, W.Z., et al., *Genome characterization of a novel porcine bocavirus.* Arch Virol, 2012. **157**(11): p. 2125-32.
99. Reuter, G., et al., *Astrovirus in wild boars (Sus scrofa) in Hungary.* Arch Virol, 2012. **157**(6): p. 1143-7.
100. Reuter, G., et al., *Porcine kobuvirus in wild boars (Sus scrofa).* Arch Virol, 2013. **158**(1): p. 281-2.
101. Bodewes, R., et al., *Identification of multiple novel viruses, including a parvovirus and a hepevirus, in feces of red foxes.* J Virol, 2013. **87**(13): p. 7758-64.

102. Reuter, G., et al., *Novel seadornavirus (family Reoviridae) related to Banna virus in Europe.* Arch Virol, 2013. **158**(10): p. 2163-7.
103. Cheung, A.K., et al., *A divergent clade of circular single-stranded DNA viruses from pig feces.* Arch Virol, 2013. **158**(10): p. 2157-62.
104. Smits, S.L., et al., *Metagenomic analysis of the ferret fecal viral flora.* PLoS ONE, 2013. **8**(8): p. e71595.
105. Zhang, B., et al., *Viral metagenomics analysis demonstrates the diversity of viral flora in piglet diarrhoeic faeces in China.* J Gen Virol, 2014. **95**(Pt 7): p. 1603-11.
106. Ng, T.F., et al., *Feline fecal virome reveals novel and prevalent enteric viruses.* Vet Microbiol, 2014. **171**(1-2): p. 102-11.
107. Varsani, A., et al., *A novel papillomavirus in Adelie penguin (Pygoscelis adeliae) faeces sampled at the Cape Crozier colony, Antarctica.* J Gen Virol, 2014. **95**(Pt 6): p. 1352-65.
108. Varsani, A., et al., *Identification of an avian polyomavirus associated with Adelie penguins (Pygoscelis adeliae).* J Gen Virol, 2015. **96**(Pt 4): p. 851-7.
109. He, B., et al., *Identification of diverse alphacoronaviruses and genomic characterization of a novel severe acute respiratory syndrome-like coronavirus from bats in China.* J Virol, 2014. **88**(12): p. 7070-82.
110. Masuda, T., et al., *Identification of novel bovine group A rotavirus G15P[14] strain from epizootic diarrhea of adult cows by de novo sequencing using a next-generation sequencer.* Vet Microbiol, 2014. **171**(1-2): p. 66-73.
111. Verma, H., et al., *Prevalence and complete genome characterization of turkey picobirnaviruses.* Infect Genet Evol, 2015. **30**: p. 134-9.
112. Amimo, J.O., et al., *Detection and genetic characterization of porcine group A rotaviruses in asymptomatic pigs in smallholder farms in East Africa: predominance of P[8] genotype resembling human strains.* Vet Microbiol, 2015. **175**(2-4): p. 195-210.
113. Zhang, W., et al., *Faecal virome of cats in an animal shelter.* J Gen Virol, 2014. **95**(Pt 11): p. 2553-64.
114. Bodewes, R., et al., *Viral metagenomic analysis of feces of wild small carnivores.* Virol J, 2014. **11**: p. 89.
115. Woo, P.C., et al., *Metagenomic analysis of viromes of dromedary camel fecal samples reveals large number and high diversity of circoviruses and picobirnaviruses.* Virology, 2014. **471-473**: p. 117-25.
116. Masembe, C., et al., *Viral metagenomics demonstrates that domestic pigs are a potential reservoir for Ndumu virus.* Virol J, 2012. **9**: p. 218.
117. Blomstrom, A.L., et al., *Viral metagenomic analysis of bushpigs (Potamochoerus larvatus) in Uganda identifies novel variants of Porcine parvovirus 4 and Torque teno sus virus 1 and 2.* Virol J, 2012. **9**: p. 192.
118. Chandriani, S., et al., *Identification of a previously undescribed divergent virus from the Flaviviridae family in an outbreak of equine serum hepatitis.* Proc Natl Acad Sci U S A, 2013. **110**(15): p. E1407-15.
119. Lauck, M., et al., *A novel hepacivirus with an unusually long and intrinsically disordered NS5A protein in a wild Old World primate.* J Virol, 2013. **87**(16): p. 8971-81.
120. Lauck, M., et al., *Discovery and full genome characterization of two highly divergent simian immunodeficiency viruses infecting black-and-white colobus monkeys (Colobus guereza) in Kibale National Park, Uganda.* Retrovirology, 2013. **10**: p. 107.
121. Chen, G.Q., et al., *Identification and survey of a novel avian coronavirus in ducks.* PLoS ONE, 2013. **8**(8): p. e72918.

122. Bruce, A.G., et al., *Next-generation sequence analysis of the genome of RFHVMn, the macaque homolog of Kaposi's sarcoma (KS)-associated herpesvirus, from a KS-like tumor of a pig-tailed macaque.* J Virol, 2013. **87**(24): p. 13676-93.
123. Ni, J., et al., *Identification and genomic characterization of a novel porcine parvovirus (PPV6) in china.* Virol J, 2014. **11**(1): p. 203.
124. Canuti, M., et al., *Persistent viremia by a novel parvovirus in a slow loris (Nycticebus coucang) with diffuse histiocytic sarcoma.* Front Microbiol, 2014. **5**: p. 655.
125. Emmenegger, E.J., et al., *Molecular identification of erythrocytic necrosis virus (ENV) from the blood of Pacific herring (Clupea pallasii).* Vet Microbiol, 2014. **174**(1-2): p. 16-26.
126. Bailey, A.L., et al., *Two novel simian arteriviruses in captive and wild baboons (Papio spp.).* J Virol, 2014. **88**(22): p. 13231-9.
127. Becker, N., et al., *Epizootic emergence of Usutu virus in wild and captive birds in Germany.* PLoS ONE, 2012. **7**(2): p. e32604.
128. Stenglein, M.D., et al., *Identification, characterization, and in vitro culture of highly divergent arenaviruses from boa constrictors and annulated tree boas: candidate etiological agents for snake inclusion body disease.* MBio, 2012. **3**(4): p. e00180-12.
129. Bodewes, R., et al., *Detection of novel divergent arenaviruses in boid snakes with inclusion body disease in The Netherlands.* J Gen Virol, 2013. **94**(Pt 6): p. 1206-10.
130. Mann, R.A., et al., *Molecular characterization and phylogenetic analysis of Murray Valley encephalitis virus and West Nile virus (Kunjin subtype) from an arbovirus disease outbreak in horses in Victoria, Australia, in 2011.* J Vet Diagn Invest, 2013. **25**(1): p. 35-44.
131. Li, L., et al., *Divergent astrovirus associated with neurologic disease in cattle.* Emerg Infect Dis, 2013. **19**(9): p. 1385-92.
132. Ng, T.F., et al., *Metagenomic identification of a nodavirus and a circular ssDNA virus in semi-purified viral nucleic acids from the hepatopancreas of healthy Farfantepenaeus duorarum shrimp.* Dis Aquat Organ, 2013. **105**(3): p. 237-42.
133. Li, L., et al., *Circovirus in tissues of dogs with vasculitis and hemorrhage.* Emerg Infect Dis, 2013. **19**(4): p. 534-41.
134. Li, L., et al., *A novel bocavirus in canine liver.* Virol J, 2013. **10**: p. 54.
135. Bodewes, R., et al., *Identification and characterization of a novel adenovirus in the cloacal bursa of gulls.* Virology, 2013. **440**(1): p. 84-8.
136. Bodewes, R., et al., *Novel B19-like parvovirus in the brain of a harbor seal.* PLoS ONE, 2013. **8**(11): p. e79259.
137. Baker, K.S., et al., *Metagenomic study of the viruses of African straw-coloured fruit bats: detection of a chiropteran poxvirus and isolation of a novel adenovirus.* Virology, 2013. **441**(2): p. 95-106.
138. He, B., et al., *Virome profiling of bats from Myanmar by metagenomic analysis of tissue samples reveals more novel Mammalian viruses.* PLoS ONE, 2013. **8**(4): p. e61950.
139. He, B., et al., *Characterization of a novel G3P[3] rotavirus isolated from a lesser horseshoe bat: a distant relative of feline/canine rotaviruses.* J Virol, 2013. **87**(22): p. 12357-66.
140. Ng, T.F., et al., *Distinct lineage of vesiculovirus from big brown bats, United States.* Emerg Infect Dis, 2013. **19**(12): p. 1978-80.
141. Stevens, H., et al., *Characterization of a novel polyomavirus isolated from a fibroma on the trunk of an African elephant (Loxodonta africana).* PLoS ONE, 2013. **8**(10): p. e77884.

142. Dacheux, L., et al., *A preliminary study of viral metagenomics of French bat species in contact with humans: identification of new mammalian viruses.* PLoS ONE, 2014. **9**(1): p. e87194.
143. Chen, Q., et al., *Isolation and characterization of porcine epidemic diarrhea viruses associated with the 2013 disease outbreak among swine in the United States.* J Clin Microbiol, 2014. **52**(1): p. 234-43.
144. Liais, E., et al., *Novel avian coronavirus and fulminating disease in guinea fowl, France.* Emerg Infect Dis, 2014. **20**(1): p. 105-8.
145. Shao, X.Q., et al., *Novel amdoparvovirus infecting farmed raccoon dogs and arctic foxes.* Emerg Infect Dis, 2014. **20**(12): p. 2085-8.
146. Uccellini, L., et al., *Identification of a novel nidovirus in an outbreak of fatal respiratory disease in ball pythons (Python regius).* Virol J, 2014. **11**: p. 144.
147. Wilkie, G.S., et al., *Complete genome sequences of elephant endotheliotropic herpesviruses 1A and 1B determined directly from fatal cases.* J Virol, 2013. **87**(12): p. 6700-12.
148. Wilkie, G.S., et al., *First fatality associated with elephant endotheliotropic herpesvirus 5 in an Asian elephant: pathological findings and complete viral genome sequence.* Sci Rep, 2014. **4**: p. 6299.
149. Ng, T.F., et al., *Oral Papillomatosis Caused by Enhydra Lutris Papillomavirus 1 (Elpv-1) in Southern Sea Otters (Enhydra Lutris Nereis) in California, USA.* J Wildl Dis, 2015.
150. Granberg, F., et al., *Metagenomic detection of viral pathogens in Spanish honeybees: co-infection by Aphid Lethal Paralysis, Israel Acute Paralysis and Lake Sinai Viruses.* PLoS ONE, 2013. **8**(2): p. e57459.
151. Hall-Mendelin, S., et al., *Detection of arboviruses and other micro-organisms in experimentally infected mosquitoes using massively parallel sequencing.* PLoS ONE, 2013. **8**(2): p. e58026.
152. Tokarz, R., et al., *Genome characterization of Long Island tick rhabdovirus, a new virus identified in Amblyomma americanum ticks.* Virol J, 2014. **11**: p. 26.
153. Chandler, J.A., et al., *Metagenomic shotgun sequencing of a Bunyavirus in wild-caught Aedes aegypti from Thailand informs the evolutionary and genomic history of the Phleboviruses.* Virology, 2014. **464-465**: p. 312-9.

Summary

Rapid diagnosis of infectious diseases is essential in order to take appropriate action to control livestock and human diseases. The development of novel molecular techniques like PCR and DNA sequencing has contributed to the advancement of disease diagnostics during the last decades. Most of these molecular virus detection approaches rely on minimal sequence knowledge of the target viral genome. DNA sequencing has become an important tool for virus identification and genome characterization.

Since the beginning of this PhD project, much has changed in the field of DNA sequencing. “Next-generation” DNA sequencing technologies (NGS) became increasingly available; improving the sequencing throughput enormously and reducing costs dramatically when compared to the classical Sanger sequencing technology.

Specific molecular tests target known regions of the viral genome. Such tests will fail to detect (and sequence) novel viruses with little or no available sequence information. Moreover, viruses (in particular RNA viruses) evolve quickly, making detection and control of viral infections more challenging.

One promising solution to the vulnerability of specific diagnostics for the identification of unknown viruses is to directly sequence all the genetic material within a clinical sample, and subsequently analyse the data for the presence of viral sequences. Due to the small size of virus genomes and low virus concentrations, when compared to the enormous background genetic material from host and bacterial cells in typical clinical samples, enrichment and amplification of viral nucleic acids is of crucial importance. This virus-focused random sequencing is referred to as “viral metagenomics”. Various viral metagenomic workflows have been described, using different methodologies to optimise targeting of the viral nucleic acids within a sample. Strategies often include filtration, centrifugation and enzymatic digestion of non-viral DNA/RNA, but little is known about the true benefits or biases that these manipulations may introduce.

Furthermore, various sequence independent amplification strategies are used to generate appropriate amounts of input DNA for sequencing. The nature of the amplification bias resulting from these strategies is often not fully understood. Moreover, viral metagenomic identification workflows are associated with low sensitivity compared to specific diagnostic tests due to their random sequencing process. The major challenge is to improve sensitivity of the metagenomic workflow in order to sequence viral nucleic acids in samples with a low

virus titer, and in clinical samples containing many contaminating background nucleic acids (e.g. tissues).

Complementary to classical virology and specific molecular diagnostic tools, viral metagenomics provides an interesting new approach for the identification of unknown pathogens and the characterization of their genomic sequence. This thesis aimed to demonstrate the diagnostic potential of metagenomics for virus identification and the characterization of high quality genomes, whilst addressing some of the challenges associated with viral metagenomics.

A promising and widely applicable viral metagenomics workflow was adopted from literature (**Chapter 3.1**). The method combined virion enrichment steps, a sequence independent pre-amplification step, molecular cloning and Sanger sequencing. The random PCR sequence independent single primer amplification (rPCR SISPA) method was selected as a widely used and viral genome structure independent amplification approach. After successful implementation in the laboratory and testing with different model viruses, the workflow was combined with NGS (454 pyrosequencing) and applied to ongoing diagnostic and research projects.

In the first case study described, the workflow was tested on two un-characterized avian paramyxoviruses (APMV) isolated during a wildlife screening program in mallards for avian influenza A viruses and APMV (**Chapter 3.2**). Traditional diagnostics failed to subtype these virus isolates, whereas the applied metagenomics workflow allowed characterization and full genome sequencing using only a moderate and economical sequencing effort. Furthermore, in one virus isolate, a multiple infection with two different APMV serotypes was identified.

In the second case study, the workflow was used for the molecular characterization of nine different avian influenza virus (AIV) isolates from a surveillance program in wild gulls and shorebirds (**Chapter 3.3**). The obtained nearly full genome sequences of these segmented viruses allowed correct subtyping and a detailed phylogenetic analysis. A transhemispheric reassortment of the genomic segments of AIV’s in European gulls was documented.

In the third case study the workflow was used to characterize 11 pigeon type 1 paramyxoviruses (PPMV1) which were isolated during routine diagnostics for AIV and APMV in pigeons (**Chapter 3.4**). In addition to the full genome sequences of the PPMV1 isolates (RNA virus), the obtained sequence information also documented considerable

genetic variability within multiple isolates and the presence of contaminating pigeon circovirus genome sequences (DNA virus) in four of these viral stocks. These findings demonstrate the value of NGS-based metagenomic workflows for quality control of biological products such as virus stocks.

The three initial case studies targeted virus isolates, containing high virion concentrations and low background nucleic acid contaminants. A forth case study aimed to test the sensitivity of the workflow on clinical samples (**Chapter 3.5**). Brain tissue samples of sheep naturally infected with the recently emerging novel Schmallenberg virus (SBV) were selected. The metagenomic workflow only allowed identification of SBV in strong positive samples, illustrating the importance of properly targeted samples and demonstrating a limited virus discovery sensitivity of the workflow applied in theis chapter.

After gaining experience with the protocol, the workflow was tested on a case study where the presence and type of virus were unknown (**Chapter 3.6**). In an attempt to increase the virus discovery sensitivity, the workflow was slightly adapted. A cluster of dairy cattle affected by a disease of unknown causality were selected as a case study. Using both a DNA and RNA virus specific discovery workflow, no viral sequences were found to be associated with the disease. However, careful follow-up analysis provided clear warnings for the potential of false positive results in metagenomics, as a laboratory contaminant and an extraction kit contaminant were identified.

Another objective of this thesis was to fine-tune the workflow to address some of the remaining challenges. Firstly, potential bias and the added value of the pre-amplification method used were investigated (**Chapter 4.1**). Evidence is provided that the rPCR SISPA amplification method results in a biased distribution of the sequence reads along the full genome sequence. The rPCR SISPA primer consists of a random oligomer sequence at its 3’ end and a defined universal sequence tag at its 5’ end. It was shown that the tag sequence of the rPCR SISPA primer was the main contributor to the uneven sequence depth observed. Furthermore, it was proven that this primer annealing bias can be reduced by extending the random 3’ oligomer sequence and by combining sequence data from multiple different rPCR SISPA reactions using different 5’ tag sequences.

Due to recent advances of NGS platforms and library preparation methods, direct sequencing of low amounts of input DNA without an extra pre-amplification step could be evaluated (**Chapter 4.2**). Illumina sequencing on the MiSeq platform, using the Nextera XT library preparation kit, was selected as the preferred NGS technology. In addition, the bioinformatics data analysis pipeline was adjusted in anticipation of the higher sequencing output. It was shown that rPCR SISPA amplification was not beneficial for viral identification, and that direct sequencing of the complementary DNA is recommended. In addition, different convenient and widely used pretreatment methods were compared and evaluated for their relative (RNA) virus discovery sensitivity in clinical samples by spiking a fixed (low) virus titer in serum and lung tissue samples. The main conclusion of this comparative study was that the recommended pretreatment strategy depended strongly on the amount and composition of the sample background nucleic acids. A filtration and subsequent nuclease treatment step increased virus discovery sensitivity in both serum and tissue. Simple DNase treatment on the RNA extract resulted in the highest improvement in the serum sample, whereas a ribosomal RNA removal on the RNA extract resulted in the biggest improvement in the tissue sample. Evidence from this and other studies in the literature is increasingly indicating that no universal optimal viral metagenomic workflow exists. Results of protocol comparison studies (including Chapter 4.2 and other published investigations) are therefore difficult to compare as many variables exist, but they all provide useful guidelines. Their results emphasize that metagenomic workflows have to be flexible. Depending on the case study concerned, a variety of steps may be included or excluded from the workflow. Some recommendations are provided which could be useful when planning future metagenomic virus discovery studies (**Chapter 5**). Furthermore, it was also shown that our optimized workflows can now easily detect viruses present at low virion concentrations (**Chapter 4.2**).

The work presented in this thesis offers a clear documentation of the potential of NGS-based methods for the identification of viruses and for the characterization of their complete genome sequences. Solid workflows, both for sample treatment and data analysis are now available to allow virus identification and further genome characterization. Sequencing technologies are expected to continuously evolve and become more accurate, economical, and accessible. Viral metagenomics will be established as a complementary tool to traditional virological tools and specific molecular diagnostic methods for both the identification and characterization of viral pathogens. However, as with any diagnostic method, its weaknesses have to be properly monitored. An integrative approach should be used, combining the added values of classical virology, specific diagnostic tests, and novel sequencing approaches, in order to exploit the full complementarity of these approaches.

Samenvatting

Voor het nemen van gepaste maatregelen om infectieziekten bij landbouwhuisdieren en mensen te bestrijden is een snelle diagnose essentieel. De ontwikkeling van nieuwe moleculaire technieken zoals PCR en DNA sequentiebepaling hebben de afgelopen decennia bijgedragen tot de vooruitgang van de diagnostiek van virale ziekten. De meeste van deze moleculaire virusdetectie benaderingen vertrouwen op ten minste een gedeeltelijke kennis van de virale genoomsequentie. De DNA sequentiebepalingstechniek is uitgegroeid tot een belangrijk instrument voor virus identificatie en genoomkarakterisatie.

Sinds het begin van dit doctoraatsonderzoek is er veel veranderd op het gebied van DNA sequentiebepaling. "Next-generation" DNA sequentiebepalingstechnologieën (NGS) werden in toenemende mate beschikbaar. Vergeleken met de klassieke Sanger sequentiebepalingstechnologie genereren ze een hogere output en verminderen ze de kosten drastisch.

Specifieke moleculair diagnostische testen richten zich op gekende regio’s in het virale genoom. Dergelijke testen zullen nieuwe virussen met weinig of geen beschikbare sequentie-informatie moeilijk kunnen detecteren. Virussen (in het bijzonder RNA-virussen) evolueren ook snel, wat detectie en bestrijding van virale infecties nog uitdagender maakt.

Een veelbelovende oplossing voor de kwetsbaarheid van deze specifieke diagnostische testen voor de identificatie van onbekende virussen is om de sequentie te bepalen van al het genetische materiaal in een staal, en vervolgens de sequentiegegevens te analyseren op de aanwezigheid van virale sequenties. Door de kleine afmetingen van virus genomen en de lage virusconcentraties vergeleken met de enorme hoeveelheid genetisch materiaal van gastheer en bacteriële cellen in typische klinische stalen, is de aanrijking en amplificatie van virale nucleïnezuren van cruciaal belang. Deze op virus gefocuste willekeurige sequentiebepaling wordt ook wel “virus metagenomics” genoemd. Diverse virus metagenomische workflows zijn beschreven in de literatuur, elk gebruik makend van verschillende methoden om de focus op virale nucleïnezuren in een staal te optimaliseren. Strategieën omvatten vaak stappen zoals filtratie, centrifugatie en enzymatische afbraak van niet-viraal DNA/RNA, maar er is weinig bekend over de echte meerwaarde of mogelijke bias die deze manipulaties kunnen introduceren.

Verder bestaan er verschillende sequentie onafhankelijke amplificatiestrategieën om geschikte hoeveelheden input DNA voor de sequentiebepaling te genereren. De aard van de amplificatie bias als gevolg van deze strategieën is vaak niet volledig begrepen. Bovendien zijn virale metagenoom workflows geassocieerd met een lagere virus detectie gevoeligheid vergeleken met specifieke diagnostische testen vanwege het willekeurig karakter van hun sequentiebepalingsproces. De grootste uitdaging is om de gevoeligheid van de metagenomische workflow voor sequentiebepaling van virale nucleïnezuren te verbeteren in stalen met een lage virus titer en met veel contaminerende nucleïnezuren (bv weefsels).

Aanvullend aan de klassieke virologie en specifieke moleculaire diagnostische instrumenten is virus metagenomics een interessante nieuwe benadering voor de identificatie van onbekende pathogenen en karakterisatie van hun genomische sequentie. Deze thesis had als doel om de diagnostische mogelijkheden van metagenomics voor virus identificatie en de karakterisatie te demonstreren en daarbij ook enkele uitdagingen geassocieerd met virus metagenomics aan te pakken.

Een veelbelovend en breed toepasbaar virus metagenomisch protocol werd geïmplementeerd uit de literatuur (**hoofdstuk 3.1**). De workflow combineerde viruspartikel aanrijkingsstappen, een sequentie onafhankelijke pre-amplificatiestap, moleculaire klonering en Sanger sequentiebepaling. De *random PCR sequence independent single primer amplification* (rPCR SISPA) werd geselecteerd als een veelgebruikte en virale genoomstructuur onafhankelijke amplificatie methode. Na een succesvolle implementatie in het laboratorium en validatie met behulp van verschillende model virussen, werd de workflow in combinatie met NGS (454 pyrosequencing) toegepast op enkele lopende diagnostische- en onderzoeksprojecten.

In de eerste casus werd de workflow getest op twee niet gekarakteriseerde aviaire paramyxovirussen (APMV) geïsoleerd tijdens een screeningsprogramma in wilde eenden voor aviaire influenza A virussen en APMV (**hoofdstuk 3.2**). Traditionele diagnostische testen slaagde er niet in om deze virusisolaten te subtyperen, terwijl de toegepaste metagenomische workflow subtypering en volledige sequentiebepalingen van het genoom toeliet, gebruik makende van slechts een gemiddelde en economische sequenbepalingsinspanning. Daarnaast werd er in één virusisolaat een co-infectie met twee verschillende APMV serotypes geïdentificeerd.

In de tweede casus werd de workflow gebruikt voor de moleculaire karakterisering van negen verschillende aviaire influenza A virussen (AIV) isolaten uit een bewakingsprogramma in wilde meeuwen en steltlopers (**hoofdstuk 3.3**). De verkregen, bijna volledige, genoom sequenties van deze gesegmenteerde virussen lieten een juiste subtypering en een gedetailleerde fylogenetische analyse toe. Er werd ook een transhemisferische re-assortering van de AIV genoomsegmenten in Europese meeuwen gedocumenteerd.

In de derde casus werd de workflow gebruikt om elf type 1 duif paramyxovirussen (PPMV1) te karakteriseren die geïsoleerd werden tijdens routine diagnostiek voor AIV en APMV in duiven (**hoofdstuk 3.4**). Naast de volledige genoom sequenties van de PPMV1 isolaten (een RNA virus), werd in de verkregen sequentie-informatie ook een aanzienlijke genetische variabiliteit waargenomen bij meerdere isolaten, alsook de aanwezigheid van duif circovirus genoomsequenties (DNA virus) in vier van deze virusstocks. Deze bevindingen tonen aan dat NGS gebaseerde metagenomische workflows geschikt zijn voor kwaliteitscontrole van biologische producten zoals de controle van virusstocks.

De drie eerste casestudies onderzochten virus isolaten die hoge viruspartikelconcentraties en lage hoeveelheden contaminerende nucleïnezuren bevatten. Een vierde casus had als doel de gevoeligheid van de workflow op klinische stalen te testen (**hoofdstuk 3.5**). Hersenweefsels van schapen werden geselecteerd welke op natuurlijke wijze besmet waren met het recente nieuwe Schmallenbergvirus (SBV). De metagenomische workflow liet enkel identificatie van SBV toe in sterk positieve stalen, hetgeen het belang van gerichte staalname demonstreert en een beperkte gevoeligheid van het gebruikte protocol aantoont.

Na ervaring te hebben opgedaan met het protocol, werd de workflow getest op een casus waarin de aanwezigheid en het virustype niet bekend waren (**hoofdstuk 3.6**). In een poging om het virus identificatiegevoeligheid te verhogen, werd de workflow enigszins aangepast. We onderzochten een cluster van melkkoeien, welke getroffen waren door een ziekte met een nog onbekende oorzaak. Gebruik makende van zowel een DNA als RNA virus specifieke identificatie workflow, konden geen virale sequenties gevonden worden die in verband konden gebracht worden met het ziektebeeld. Nauwkeurige opvolganalyse gaf duidelijke waarschuwingen voor mogelijke vals-positieve resultaten in de metagenomics analyse; er werden namelijk een laboratorium contaminant en een extractiekit contaminant geïdentificeerd.

Een ander doel van deze thesis was de workflow te optimaliseren om zo enkele resterende uitdagingen uit de virus metagenomics aan te pakken. Ten eerste werd mogelijke bias en de toegevoegde waarde van de gebruikte pre-amplificatiemethode onderzocht (**hoofdstuk 4.1**). We konden aantonen dat de rPCR SISPA amplificatie methode resulteert in een ongelijke verdeling van de sequenties over het genoom. De rPCR SISPA primersequentie bestaat uit een random oligomeer sequentie aan zijn 3’ uiteinde en een vaste universele “tag”-sequentie aan zijn 5’ uiteinde. We toonden aan dat deze sequentietag van de rPCR SISPA primer de belangrijkste bijdrage leverde aan de waargenomen ongelijke sequentiediepte. Bovendien werd er aangetoond dat deze primer annealing bias kan worden verminderd door de 3' random oligomeersequentie te verlengen en door sequentiebepalingsgegevens van meerdere rPCR SISPA reacties met verschillende 5' sequentietags te combineren.

Door recente ontwikkelingen van NGS platformen en kits, was het mogelijk geworden om rechtstreekse sequentiebepaling zonder een extra pre-amplificatie stap te evalueren van lage hoeveelheden input DNA (**hoofdstuk 4.2**). Illumina sequentiebepaling op het MiSeq platform, met behulp van de Nextera XT library preparatiekit, werd geselecteerd als aangewezen NGS technologie en reagentia. Daarenboven werd de bioinformatica analyse aangepast in anticipatie van een hogere sequentie output. Er werd aangetoond dat rPCR SISPA amplificatie niet gunstig is voor virus identificatie, en dat rechstreekse sequentiebepaling van het cDNA aanbevolen is. Verder werden verschillende gemakkelijke en veel gebruikte voorbehandelingsmethodes met elkaar vergeleken en geëvalueerd op hun relatieve gevoeligheid voor (RNA) virus identificatie in klinische stalen, door artificiëel serum en longweefsel te besmetten met een vaste (lage) virus titer. De belangrijkste conclusie van dit vergelijkend onderzoek was dat de aanbevolen voorbehandelingsstrategie sterk afhankelijk is van de hoeveelheid en de samenstelling van de achtergrondnucleïnezuren in het staal. Een filtratie en daaropvolgende nuclease behandelingsstap verhoogde de gevoeligheid voor virus identificatie in zowel serum en weefsel. Een DNase behandeling op het RNA extract leidde tot de hoogste verbetering van de gevoeligheid voor RNA virus identificatie in het serum, terwijl een ribosomaal RNA verwijderingsbehandeling op het RNA extract tot de grootste verbetering leidde in het weefsel staal. De resultaten uit deze en andere studies uit de literatuur geven in toenemende mate aan dat er geen universele optimale virus metagenomische workflow bestaat. Resultaten uit verschillende studies (waaronder hoofdstuk 4.2 en andere gepubliceerde onderzoeken) zijn moeilijk met elkaar te vergelijken omdat er zoveel variabelen zijn; maar ze bieden allemaal wel nuttige richtlijnen. Hun resultaten benadrukken dat metagenomische workflows flexibel moeten zijn. Afhankelijk van de betrokken casus kunnen verschillende stappen worden opgenomen of uitgesloten uit de metagenomische workflow. Enkele aanbevelingen worden geformuleerd in deze thesis en kunnen nuttig zijn bij de planning van toekomstige virus metagenomische studies (**hoofdstuk**

5). Verder werd er ook aangetoond dat onze geoptimaliseerde protocols nu gemakkelijk virussen aanwezig in lage concentraties kunnen detecteren (**hoofdstuk 4.2**).

Het in dit proefschrift gepresenteerde werk biedt een duidelijke documentatie van het potentieel van NGS-gebaseerde methoden voor de identificatie van virussen en voor de karakterisering van hun complete genoomsequenties. Degelijke protocols, zowel voor staalbehandeling en data analyse zijn nu beschikbaar, en kunnen gebruikt worden voorvirus identificatie en genoomkarakterisatie. Het ligt in de verwachtingen dat sequentiebepalingstechnologieën zullen blijven evolueren naar meer accuraatheid, een lagere kostprijs en een betere toegankelijkheid. Virus metagenomics zal zich profileren als een complementair instrument aan traditionele virologische tools en specifieke moleculaire diagnostische methoden, zowel voor de identificatie als voor karakterisering van virale ziekteverwekkers. Zoals met elke diagnostische methode moeten de zwakke punten ervan nauwlettend in het oog gehouden worden. Een geïntegreerde benadering zou beoogd moeten worden, waarin de toegevoede waarde van klassieke virologie, specifieke diagnostische testen en nieuwe sequentiebepalingsstrategieën gecombineerd worden, om zo de complementariteit van deze benaderingen ten volle te benutten.

Bibliography

List of peer-reviewed publications

Rosseel T., Lambrecht B., Vandenbussche F., van den Berg T., and Van Borm S., Identification and complete genome sequencing of paramyxoviruses in mallard duck (*Anas platyrhynchos*) using random access amplification and next generation sequencing technologies (2011). *Virology Journal,* 8: 463.

Van Borm S., **Rosseel T.**, Vangeluwe D., Vandenbussche F., van den Berg T. and Lambrecht B., Phylogeographic analysis of avian influenza viruses isolated from *Charadriiformes* in Belgium confirms intercontinental reassortment in gulls (2012). *Archives of Virology,* 157(8): 1509-1522.

Rosseel T., Scheuch M., Höper D., De Regge N., Caij A.B., Vandenbussche F., and Van Borm S., DNase SISPA-next generation sequencing confirms Schmallenberg virus in Belgian field samples and identifies genetic variation in Europe (2012). *PLoS One*, 7(7): e41967.

Van Borm S., **Rosseel T.**, Steensels M., van den Berg T., and Lambrecht B., What's in a strain? Viral metagenomics identifies genetic variation and contaminating circoviruses in laboratory isolates of pigeon paramyxovirus type 1 (2013). *Virus Research,* 171(1): 186-193.

Vandenbussche F., Sailleau C., **Rosseel T.**, Desprat A., Viarouge C., Richardson J., Eschbaumer M., Hoffman B., De Clercq K., Bréard E., and Zientara S., Full-genome sequencing of four bluetongue virus serotype 11 viruses (2013). *Transboundary and Emerging Diseases*, doi: 10.1111/tbed.12178.

Rosseel T., Van Borm S., Vandenbussche F., Hoffmann B., van den Berg T., Beer M. and Höper D., The Origin of Biased Sequence Depth in Sequence-Independent Nucleic Acid Amplification and Optimization for Efficient Massive Parallel Sequencing (2013). *PLoS ONE,* 8(9): e76144.

Rosseel T., Pardon B., De Clercq K., Ozhelvaci O., and Van Borm S., False-positive results in metagenomic virus discovery: a strong case for follow-up diagnosis (2014). *Transboundary and Emerging Diseases*, 61(4):293-90

Boland C., Bertrand S., Mattheus W., Dierick K., Jasson V., **Rosseel T.**, Van Borm S., Mahillon J. and Wattiau P. Extensive Genetic Variability Linked to IS26 Insertions in the fljB

Promoter Region of Atypical Monophasic Variants of Salmonella Typhimurium (2015). *Applied and environmental microbiology*, doi:10.1128/AEM.00270-15.

Rosseel T., Ozhelvaci O., Freimanis G. and Van Borm S. Evaluation of convenient pretreatment protocols for RNA virus metagenomics in serum and tissue samples. *Journal of Virological Methods, pii: S0166-0934(15)00209-8. doi: 10.1016/j.jviromet.2015.05.010.*

Dridi M., **Rosseel T.**, Orton R., Johnson P., Lecollinet S., Muylkens B., Lambrecht B., Van Borm S. Next generation sequencing shows different quasispecies dynamic of West Nile virus in SPF chicken and carrion crow (*Corvus corone*) *in vivo* infections models. *Submitted in Journal of General Virology.*

Book chapters

Van Borm S., Bélak S., Freimanis G., Fusaro A., Granberg F., Höper D., King D., Monne I., Orton R. and **Rosseel T.,** Next-generation sequencing in veterinary medicine: how can the massive amount of information arising from high-throughput technologies improve diagnosis, control, and management of infectious diseases? (2015) *Methods in molecular biology,* 1247:415-36.

Conference contributions

Rosseel T., Lambrecht B., van den Berg T., Vandenbussche F. and Van Borm S. Combining random access amplification and massive parallel sequencing on avian influenza virus genomes: towards cost effective sequence independent whole genome determination? Poster presentation at 5th Annual Meeting of EPIZONE “Science on alert”, Arnhem, NL, April 2011.

Rosseel T., Müller I., Höper D., Van Borm S., Grund C., Hoffmann B., van den Berg T., Beer B. and Harder T.C. Optimize strategies for whole genome sequencing of avian paramyxoviruses: Towards the 100 euro viral genome? Oral presentation (Co-author) at 1st International Avian Respiratory Disease Conference, University of Georgia, Athens, GA, USA, May 2011.

Van Borm S., **Rosseel T.**, Lambrecht B., van Geluwe D. and van den Berg T.. Phylogeographic analysis of nine avian influenza viruses isolated from Charadriiformes in Belgium confirms intercontinental reassortment in gulls. Oral presentation at 8th International Symposium on Avian Influenza, University of London, UK, April 2012.

Van Borm S., Lambrecht B., van den Berg T., Vandenbussche F. and **Rosseel T**. “What’s in a strain?” Random access next generation sequencing of virus isolates as a diagnostic and quality control tool. Oral presentation (co-author) at 6th Annual Meeting of EPIZONE “Viruses on the move”, Brighton, UK, June 2012.

Van Borm S., Monne I., King D. and **Rosseel T.** Next generation sequencing opens new views on virus evolution and epidemiology. Oral presentation (co-author) at 16th International Symposium of World Association of Veterinary Laboratory Diagnosticians, Berlin, Germany, June 2013.

Rosseel T., Steensels M., Nguyen T.G., Lambrecht B. and Van Borm S. Next Generation sequencing to investigate in vitro evolution of low pathogenic avian influenza H5N1 virus under strong antiviral (oseltamivir) selection pressure. Poster presentation (Poster **award** for the best poster) at the 7th Annual Meeting of EPIZONE “Nothing permanent, except change”, Brussels, BE, October 2013.

Rosseel T., Ozhelvaci O., Freimanis G., Van Borm S. Technical evaluation of pretreatment protocols for viral metagenomics of RNA viruses. Poster presentation at 8th Annual Meeting of EPIZONE “Primed for tomorrow”, Copenhagen, DK, September 2014.

Curriculum Vitae

Toon Rosseel werd geboren op 19 maart 1985 te Brugge. Na het behalen van het diploma hoger secundair onderwijs aan het Sint-Andreasinstituut Garenmarkt te Brugge (richting Wetenschappen-Wiskunde) starte hij in 2003 met de studie Bio-ingenieur aan de KU Leuven Kulak te Kortrijk om deze dan een jaar later verder te zetten in Leuven. Hij behaalde in 2008 met onderscheiding het master diploma van Bio-Ingenieur in de Landbouwkunde (specialisaties: Dierproductie, Cel- en genbiotechnologie). In september 2014 startte hij een bachelor opleiding Bioinformatica aan de Howest te Brugge.

Als onderzoeker begon hij op 1 mei 2009 te werken aan het project RandSeq met als titel *Identificatie van onbekende virussen door random genoomamplificatie en sekweneringstechnieken*. Dit project werd gefinancieerd door het Centrum voor Onderzoek in Diergeneeskunde en Agrochemie (CODA-CERVA). Na een succesvolle afronding van dit twee jaar durende project, werd een verlenging goedgekeurd en een doctoraatsonderzoek gestart in samenwerking met de vakgroep Virologie, parasitologie en immunologie van de Faculteit Dierengeneeskunde aan de Universiteit Gent. Na het afronden van het RandSeq project, werkte hij nog op het Europees project Epi-SEQ met als titel *Molecular epidemiology of epizootic diseases using next generation sequencing technology* (gefinancieerd door het CODA-CERVA), en het BELSPO project *Next Generation sequencing en bioinformatica workflows voor identificatie en karakterisatie van virale ziekten*. Sinds 1 mei 2015 werkt hij als bioinformaticus in het Centrum voor Medische Genetica Gent.

Toon Rosseel is auteur of mede-auteur van meerdere wetenschappelijke pubicaties in internationale tijdschriften en lichtte zijn onderzoeksresultaten toe op verschillende internationale congressen.

Dankwoord

Mijn doctoraat is af! Fini! Finished! De kers op de taart van een zeer boeiende periode in mijn leven. Zes jaar lang heb ik me verdiept in de wereld van de virussen, moleculaire biologie, DNA sequentiebepalingstechnologieën en bioinformatica. Dit lijvig boekje is dan ook het bewijs dat ik me zeker niet heb verveeld. Onderzoek voeren doe je echter niet alleen. Een groot aantal mensen hebben bijgedragen tot het tot stand komen van dit doctoraat. Ik wil hen hiervoor dan ook oprecht bedanken.

Mijn grootste dank gaat uit naar mijn promoter Steven Van Borm. Ik bewonder jouw passie, je gedrevenheid, en je klare kijk. Wat startte met een ‘DNA smeer’ van op een agarosegel, is geëindigd in 9 accepted peer-reviewed publicaties, 1 publicatie nog onder review, 1 boekhoofdstuk en deelname aan 2 Europese projecten, waaronder een met jou als projectleider. Je wist steeds het overzicht te bewaren en juiste inschattingen te maken. Ik kon me echt geen betere promotor voorstellen! Bedankt voor al je bijdragen tot dit doctoraat, voor het nalezen van vele draft versies, voor het vertrouwen dat je steeds in mij had, voor je steun om me bij te scholen in de bioinformatica, en natuurlijk ook voor de vele leuke momenten in onze ‘Office’ en tijdens de verschillende buiten- en binnenlandse meetings.

Ik dank ook mijn co-promotor Prof. Dr. Hans Nauwynck van de faculteit Diergeneeskunde van de Universiteit Gent om het promotorschap op zich te nemen, en om mijn doctoraat grondig na te lezen.

Thanks to all the members of the Examination Commission (Prof. Dr. Herman Favoreel, Prof. Dr. Niek Sanders, Prof. Dr. Xavier Saelens, Prof. Dr. Linos Vandekerckhove and Dr. Sebastiaan Theuns of UGent; Dr. Dirk Höper of Friedrich-Loeffler-institut; Dr. Steven Van Gucht of WIV-ISP; and Dr. Thierry van den Berg of CODA-CERVA) for accepting this task, for the critical proofreading of my thesis and for the useful comments and suggestions.

Graag wil ik directeur Pierre Kerkhofs en de voltallige directieraad van het CODA/CERVA bedanken voor de financiële steun gedurende deze 6 jaar. Dankzij hen heb ik de kans gekregen om me op professioneel vlak ten volle te kunnen ontplooien.

Dank ook aan de collega’s van het Moleculair Platform voor aangename werksfeer. Frank Vandenbussche & Elisabeth Mathijs, nog eens bedankt voor het nalezen van mijn doctoraat. En Orkun Ozhelvaci... jou bijdrage tot deze thesis is enorm groot. Merci voor het vele labowerk bij zowat alle experimenten, merci voor je flexibiliteit (niet enkel flexibele metagenomics protocollen zijn van onschatbare waarde), merci om taxi te spelen vanuit Gent tijdens mijn ‘beenbreuk-periode’, merci dat jij nooit gestopt bent met koffie te drinken, merci om me te voeren naar Aalter in het midden van de nacht omdat ik mijn laatste trein ‘gemist’ had na het afdrukken van de eerste versie van dit doctoraat, en zo kan ik nog wel een tijdje doorgaan met opsommen eigenlijk. Thanks mate!

Graag had ik nog alle andere collega’s van het CODA/CERVA bedankt met wie ik heb samengewerkt. En particulier, tous les collègues du laboratoire Virologie et Immunologie Aviaires pour de nombreux moments agréables dans “la zone”, mais aussi sur les activités après le travail. Merci Bénédict Lambrecht dat ik uw labo mocht gebruiken om veel van mijn experimenten uit te voeren. Bedankt ‘Ridder’ Marc Boschmans voor het aanleren van verschillende labotechnieken en om je ervaringen te delen, voor de vele voetbal-gerelateerde gesprekken, en voor de sfeervolle Halle Carnaval edities. Merci Mireille Decaesstecker voor je professionele hulp bij verschillende experimenten. En bedankt Hugo Van der Vaet voor de aangename babbels tussen pot en pint. Ik hoop mee met jou dat het goed komt met “den eendracht”.

Also, I would like to express my gratitude to all my international colleagues for the several great collaborations. Special thanks to Dr. Dirk Höper and his team for the nice time I had during my short term mission on “the island”, and for the excellent collaborations.

Bedankt ook aan al mijn vrienden voor de nodige afleiding tijdens het schrijven van dit boekje. Ik beloof hierbij plechtig opnieuw wat meer tijd voor jullie vrij te maken.

En last but not least, dank aan mijn familie voor hun onvoorwaardelijke steun. Mama, Papa, bedankt voor alles! Buiten het feit dat ik zonder jullie niet zou rondlopen op deze aardbol (als dat al geen grote bijdrage is tot dit doctoraat), ben ik jullie erg dankbaar voor de kansen die jullie mij hebben gegeven. Het betekent zo veel dat jullie altijd voor me klaar staan.