Contribution of Next-Generation Sequencing to Aquatic and Fish Virology
Andriniaina Andy Nkili-Meyong, Laurent Bigarré, Ingrid Labouba, Tatiana Vallaeys, Jean-Christophe Avarre, Nicolas Berthet

To cite this version:
Andriniaina Andy Nkili-Meyong, Laurent Bigarré, Ingrid Labouba, Tatiana Vallaeys, Jean-Christophe Avarre, et al.. Contribution of Next-Generation Sequencing to Aquatic and Fish Virology. Intervirology, Karger, 2016, 59 (5-6), pp.285-300. 10.1159/000477808. hal-02922791

HAL Id: hal-02922791
https://hal.archives-ouvertes.fr/hal-02922791
Submitted on 27 Aug 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NoDerivatives| 4.0 International License
Contribution of next-generation sequencing to aquatic and fish virology

Andriniaina Andy Nkili-Meyong¹, Laurent Bigarré², Ingrid Labouba¹, Tatiana Vallaey³, Jean-Christophe Avarre⁴*, Y, Nicolas Berthet¹,⁵*, Y

¹ Centre International de Recherches Médicales de Franceville (CIRMF), Franceville, Gabon
² ANSES, Laboratoire de Ploufragan-Plouzané, Université Bretagne-Loire, Plouzané, France
³ Université de Montpellier, Place Eugene Bataillon, Montpellier, France
⁴ Institut des Sciences de l'Evolution de Montpellier, UMR 226 IRD-CNRS-UM-EPHE, Montpellier, France
⁵ Centre National de Recherche Scientifique (CNRS) UMR3569, 25 rue du docteur Roux, Paris, France
⁶ Institut Pasteur, Unité Environnement et risques infectieux, Cellule d'Intervention Biologique d’Urgence, 25 rue du Docteur Roux, Paris, France

Corresponding authors:

Dr Jean-Christophe Avarre
Institut des Sciences de l'Evolution de Montpellier, UMR 226 IRD-CNRS-UM-EPHE, Place Eugène Bataillon, 34095 Montpellier, France
Email: jean-christophe.avarrer@ird.fr

Dr Nicolas Berthet
Centre International de Recherches Médicales de Franceville (CIRMF), Department of Zoonosis and Emerging Diseases, Franceville, Gabon
E-mail: nicolas.berthet@ird.fr

*Y equally contributed.
Abstract

The recent technological advances in nucleic acid sequencing, called next-generation sequencing (NGS), have revolutionized the field of genomics and have also influenced viral research. Aquatic viruses, and especially those infecting fish, have also greatly benefited from NGS technologies, which provide a huge amount of molecular information at a low cost in a relatively short period of time. Here, we review the use of the current high-throughput sequencing platforms with a special focus on the associated challenges (regarding sample preparation and bioinformatics) in their applications to the field of aquatic virology, especially for (i) discovering novel viruses that may be associated with fish mortalities; (ii) elucidating the mechanisms of pathogenesis; and finally, (iii) studying the molecular epidemiology of these pathogens.

Keywords: viruses, fish, diagnostics, next-generation sequencing, aquatic virology
1. Introduction

The first clue that aquatic viruses play a major role in their environment came from investigations in the early 1990s that revealed their high abundance (Bergh et al., 1989; Hara et al., 1991). For example, viruses are the most abundant organisms in the marine environment (Suttle, 2005). It was estimated that each milliliter of seawater contained millions of viral particles (Proctor and Fuhrman, 1990) and that viral particles accounted for about 94% of nucleic-acid-containing particles (Wen et al., 2004; Suttle, 2007). Some environmental studies on aquatic viral communities have shown that the “unknown fraction” (i.e., without homology to any sequence available in the databases) is substantial, and may range from 30% to 99% (Marhaver et al., 2008; McDaniel et al., 2008). A recent study focusing on the analysis of virus open reading frames (ORFs) showed that a large fraction had no homologs in databases. These “ORFans” are more frequently present in viruses than in bacteria (respectively 30% and 9%), suggesting that the diversity among viruses may be even greater than suspected (Yin and Fischer, 2008).

The overwhelming presence of viruses in aquatic environments has attracted the attention of the scientific community (Fuhrman, 1999; Wilhelm and Suttle, 1999; Wommack and Colwell, 2000). Moreover, many studies have also targeted aquatic viruses essentially because of their devastating economic effects on aquaculture (Bloch et al., 1991; van Hulten et al., 2001; Zhang et al., 2004; Yoganandhan et al., 2006; Bacharach et al., 2016). Indeed, the impact on the worldwide trade of some infected species, such as fishes or shrimps, has resulted in estimated losses of billions of U.S. dollars (Mardones et al., 2010; Bacharach et al., 2016). Nevertheless, viruses also play a central role in ecological niches, in terms of the variety of infected hosts and nutrient/carbon cycling, further fostering research in this group. In fact, aquatic viruses infect a wide range of aquatic species, from phyto- and zooplankton (Suttle, 2005; Dunlap et al., 2013), shrimp (Lightner, 2011), fish (Walker and Winton, 2010), and seals (Bodewes et al., 2013) to even the largest animals in the oceans, whales (Birkun, 1996). The ability of viruses to infect such a wide variety of species indicates a remarkable capacity for adaptation and a potential public health hazard. For example, in 2011, an avian H3N8 virus caused the deaths of over 160 juvenile seals. This virus can be transmitted via respiratory droplets and it has already been involved in human pandemics (Valleron et al., 2010; Karlsson et al., 2014). All of these aspects bring viruses into the spotlight (Suttle et al., 1990). However, knowledge on aquatic viruses, their diversity, their transmission modes or their reservoir(s), is relatively scarce, exemplifying an urgent need to dive deeper into this vast unexplored world.

In this context, this review describes the latest developments of next-generation sequencing (NGS) tools in virology, from their emergence in the field of human health to their recent applications for the investigation of aquatic and more specifically fish viruses.

2. Evolution of technologies for pathogen discovery across the history of virology

The history of virology is tightly linked with the evolution of technologies that allowed major breakthroughs in the field. Technological developments during the last century have improved the detection and characterization of viral agents and have simultaneously led to a better understanding of their evolutionary history.

Viruses were first isolated on Chamberlain columns but, unlike most other microorganisms, viruses long remained invisible under conventional light microscopes. Furthermore, given that virus diameter varies from a few tens of nm to several hundreds of nm, the observation of viruses only became possible with the advent of electronic microscopy (EM) and its resolution on nanometer scales. With the visualization of virus, virology then underwent important development because virus morphology constitutes a fundamental feature in the
classification and recognition of viruses. One of the first EM applications in clinical virology was the distinction between poxviruses and the varicella-zoster virus (Nagler and Rake, 1948; Van Rooyen and Scott, 1948). Later developments, such as negative staining technology or immuno-electron microscopy, paved the way to virological diagnosis and detection of viral agents (Gust et al., 1971; Welch, 1971; Kapikian et al., 1972). However, for better identification of viruses at the genus or the species level, EM generally must be combined with other methods such as cellular culture or molecular approaches. Because viruses are necessarily intracellular parasites, and symptomatic effects very often lag behind infection peaks — thereby hindering detection —, viral replication in living cells was initially a prerequisite for viral identification. Historically, specific cells were used to isolate viruses either from cellular culture, embryonated eggs, or laboratory animals. These cells, traditionally derived from newborn mice (in the case of mammal viruses), constitute the classic tools for the isolation of a large spectrum of viruses from a wide diversity of biological samples. Therefore, viral replication in cellular culture, still considered the gold standard for diagnosis, remains a key feature for the discovery and in vitro investigation of novel viruses.

The main advantages of cell replication-based methods for virological diagnosis include the potential for isolation and amplification of viruses, improved sensitivity compared with antigen-based detection methods, and the possibility to isolate unexpected infectious viruses. Although mammal virology has benefited from these advantages, cellular models are still needed in aquatic virology (Villena, 2003). Despite the large number of fish cell lines have been developed over the last few years, efforts are still needed for shellfish and other aquatic organisms for which cell culture is often inaccessible (Rinkevich, 2011) and for which novel and integrative approaches remain to be designed (Burge et al., 2016; Pernet et al., 2016).

Since the advent of molecular biology, nucleic-acid-based detection and identification methods have revolutionized virology and continue to play an important role in the detection and characterization of viral infections in many fields, including aquatic virology, at the fundamental and clinical levels. For instance, the polymerase chain reaction (PCR) is the most widely used method for viral detection (Van Belkum and Niesters, 1995; Hayden and Persing, 2001; Cobo, 2012). One of its main strengths is its high sensitivity, although false negatives can occur when (i) the viral load is below the threshold level, which is often the case in healthy carriers; (ii) the biological samples contain chemical or biological substances that inhibit the enzymatic amplification process; (iii) there is genetic divergence between the sequence of the screened agent and the primers used for the amplification of a targeted area, which is often the case for poorly characterized viruses such as aquatic ones (Engelsma et al., 2013); or (iv) the virus evolves rapidly, as is the case for RNA viruses, which undergo reassortment mechanisms. PCR generally targets one or at most a few species and variants. Nevertheless, identification of the pathogen can be performed quickly and directly on clinical or environmental samples, which can influence public health policy. More recently, high-throughput approaches, such as microarrays, have facilitated the parallel detection of a large number of pathogens, thus providing access to the high genetic diversity of viral agents (Berthet et al., 2008). Though very powerful, this technology still requires the a priori knowledge of viral sequences. This barrier has been overcome by the latest advances in sequencing technologies that have unlocked access to a wider world of viral agents, including those that cannot be readily grown or are not culturable, or those that persist and even develop in their environment before (re)emerging to give rise to sporadic or epidemic cases.

3. From Sanger to next-generation sequencing: overcoming technological challenges

3.1 Sequencing technologies
Ever since genomic DNA was identified as the main tool for characterizing species and individuals, progress in nucleotide sequencing has been a basic challenge. Many improvements have indeed been made since the first descriptions of DNA sequencing in 1977 by Frederick Sanger and Walter Gilbert (Maxam and Gilbert, 1977; Sanger et al., 1977). Both methods are today referred to as first-generation sequencing technologies, Sanger’s method being adopted as the standard one. Subsequent improvements have aimed at reducing, on the one hand, the cost of the overall sequencing process, and, on the other hand, increasing its efficiency and practicality. Major advances include, among others, the introduction of DNA amplification before sequencing or the development of fluorescence-based capillary electrophoresis (Figure 1) (Mardis, 2013). Thus, since the first Sanger sequencer was put on the market in 1987, numerous technological challenges have been overcome, eventually leading to automated sequencing (Ansorge et al., 1986; Ansorge et al., 1987). However, despite these major advances, Sanger sequencing remained limited by (i) the quantity of DNA required and the number of samples that could be treated simultaneously; (ii) the size of the sequenced DNA fragment (ranging from 0.5 to 1 kb); and (iii) the low throughput of the process (only 115 kb/day for 1 kb length) (Figure 1). In the late 1990s, the human genome sequencing project gave rise to major improvements in sequencing technologies and approaches (Lander et al., 2001; Venter et al., 2001) that quickly found numerous applications, including in virology. This project greatly contributed to the emergence of the so-called next-generation sequencing (NGS), largely used for screening and characterizing pathogen communities (Naidoo et al., 2011), first in the human environment and then in environmental sciences, including aquatic virology. The first NGS apparatuses were developed and released in 2005 as the second generation of sequencers, and they introduced significant improvements in DNA amplification, sequencing chemistry and fluorescence detection; they were also accompanied by powerful bioinformatics tools for treating the unprecedented amounts of generated data. NGS thus made it possible to carry out parallel analyses of complex microbial communities with high sensitivity. Second-generation (2G) technologies are mainly based on sequencing a huge number of short DNA fragments via distinct but simultaneous massive reactions, resulting in millions to billions of reads. They rely on clonal amplification that generates sequence-able clusters, but exploit a different type of sequencing chemistry. The first 2G sequencer to be commercialized (the 454 from Roche) was based on pyrosequencing and produced nearly 200,000 reads of 100-150 nt in length, corresponding to approximately 20 Mb of output data (Liu et al., 2012). The last version of the 454 (GS FLX titanium XL+) could generate up to 1,000,000 reads of 700 nt, corresponding to 700 Mb of data (Source: Roche website, 454.com) (Figure 1). The Roche 454 pyrosequencer is no longer marketed due to the development of other 2G platforms and the recent emergence of the third-generation (3G) sequencing technology (Bleidorn, 2016).

In parallel, several other 2G platforms arose, further improving sequencing output by increasing the length, quantity, and quality of the generated sequences (Rothberg and Leamon, 2008; Liu et al., 2012). One leading 2G sequencing approach relies on enzymatic synthesis with reversible dye terminators. Initially developed by Solexa in 2006, this technology was acquired and improved by Illumina in 2007 (Chi, 2008; Liu et al., 2012). Illumina’s platforms use a specific flow cell as a solid support carrying adapter sequences that bind the DNA fragments to be sequenced. Illumina’s technology is also characterized by a clustering step based on clonal amplification carried out with a bridge PCR approach, thus making paired-end sequencing possible (Figure 1) (Mardis, 2008b; Liu et al., 2012). The length of generated reads, as well as the quantity of output data, have significantly increased since, from 25 nt/1 Gb to up to 150 nt/1800 Gb. Today, Illumina offers two main series of benchtop sequencers, namely MiSeq (Figure 1) and NextSeq that provide 15 Gb/25 x 10^6 reads and 120 Gb/400 x 10^6 reads as maximum output data, respectively. The HiSeq series,
incompatible with benchtop use, can generate up to 1800 Gb/6 x 10^9 reads. This scale of output ability allows for a large panel of biological applications as well as optimizing time and costs according to research needs (Mardis, 2008b;© Illumina, 2016).

More recently, another type of technology has been put on the market and is based on monitoring changes in pH (Ion Torrent, from Life Technologies). Developed in 2010, this brand offers several platforms that vary according to output data, namely Ion S5, Proton and a benchtop version, the Ion PGM. The sequencing principle is completely different from the 2G sequencers Roche 454 and Illumina. The clonal amplification step involves emulsion PCR using small beads linking DNA fragments to oil micelles (as in the 454 system), and the sequencing process consists in measuring the variation in pH due to the release of H^+ protons during DNA synthesis (Figure 1). Only uncoupled native nucleotides are used for DNA synthesis, and no specific marker is needed, wherein lies its major advantage. This semiconductor technology does not require any laser or scanner to read fluorescence, greatly simplifying the sequence reading process and saving time. Moreover, this approach prevents the accumulation of background noise due to fluorophores. In comparison with Illumina’s benchtop sequencer, for equivalent accuracy and sequence quality, the Ion PGM is more expensive — USD 1000/1 Gb/run versus USD 502 with Illumina — but much more rapid — 2 h versus ≈27 h (Figure 1) (Quail et al., 2012).

Pacific Biosciences (PacBio) was the first company to commercialize (in 2009) 3G technology known as single-molecule real-time (SMRT) DNA sequencing. Its technology is based on the direct monitoring of polymerase activity during the DNA synthesis with fluorescently labeled nucleotides allowing for real-time imaging of the newly synthesized DNA strand. Initially able to generate ≈100 base reads, today PacBio platforms can produce reads up to 10 kb within 2 h (Figure 1) — the longest generated read is ≈54 kb — thus helping to solve the genome assembly issue in part (Eid et al., 2009;Lee et al., 2014). Despite this high efficiency, a 5–20% error rate has been reported (Quail et al., 2012;Lee et al., 2014;Rhoads and Au, 2015). The ongoing development of fourth generation (4G) technologies could fill this gap by both increasing accuracy and reducing time and cost of high-throughput sequencing. Possibilities include combining 3G single-molecule sequencing with Nanopore technology, based on the ability of DNA molecules to pass through nanoscale diameter biopores. Oxford Nanopore Technology launched its MiniION sequencer, which detects the specific electrical signature of each nucleotide after cleavage by an exonuclease of the single DNA molecule that passes through the pore (Feng et al., 2015). The rapidity of the reaction (20 ms/base), combined with the accuracy of the exonuclease reaction, holds great promise. Despite a reported high error sequencing rate, the MiniION allowed to rapidly sequence a full genome of influenza virus (Wang et al., 2015), or to achieve the identification of a virus at the strain level (Kilianski et al., 2015). For a more detailed description of these technologies, readers can refer to Goodwin et al. (Goodwin et al., 2016).

3.2 Preparation of libraries for viral genome detection: Issues and challenges

As in all other fields of biology using NGS technologies, sample treatment and library preparation raise the issue of the pathogen detection threshold. The advantage, and also the main difficulty with NGS, is the ability to sequence all DNA present in a sample, which produces a large proportion of noisy data, potentially masking the targeted sequence. Bioinformatics tools may eventually be able to prune unwanted sequences; however, data that is clean from the start will facilitate the subsequent bioinformatics analyses. In aquatic metagenomics, the problem is particularly complex because samples may include high quantities of contaminants. Various filtration procedures, such as size fractioning, have been optimized to efficiently concentrate viruses and eliminate the main source of noisy sequences
(e.g. eukaryotes, bacteria) (Tringe and Rubin, 2005; Datta et al., 2015). Ultimately, the real challenge lies in detecting the targeted viruses among all those present in the samples.

The problem is slightly different when working directly on fish specimens, in which the quantity of contaminants is limited compared with water samples. Here, although viral diversity is restricted to the targeted host and host bacteria, the notion of viral load has a significant impact. It is extremely laborious to detect the genome of a virus present in low quantities. Viruses in low proportions in a sample cannot be detected without any prior enrichment strategy (Daly et al., 2011; Hall et al., 2014). Several procedures have already been developed to enrich samples or derivative DNA/RNA. For example, DNase pre-treatment (Allander et al., 2001) and rRNA removal (He et al., 2010; Li et al., 2013) are efficient and robust options for RNA virus exploration by respectively eliminating host DNA contaminants and host ribosomal RNA that represent more than 90% of total RNA of a sample (Datta et al., 2015). Another strategy consists in enriching studied samples with fractions that are likely to contain the targeted viral material. Random whole genome or transcriptome amplification based on non-specific techniques can also be used (Dupinay et al., 2012; Datta et al., 2015).

However, the efficiency of this type of amplification for the most diluted fractions — namely viruses with very low loads — is still uncertain. Finally, it is important to first enrich the sample with the targeted viral fraction and then amplify it to ensure a higher efficacy (Dupinay et al., 2012; Lefterova et al., 2015). For these reasons, the option of targeted sequence capture systems is very attractive, because it can ensure specific selection, and thus specific enrichment, of targeted sequences (Mertes et al., 2011). These approaches, despite their promising efficiency, remain expensive and imply that the sequence of the targeted pathogen is known; they are thus not suitable for pathogen discovery (Jones and Good, 2016).

Global enrichment processes by capture have already been tested and proven mainly in the field of the human virology (Briese et al., 2015; Lefterova et al., 2015; Wylie et al., 2015). Although they appear promising, their use is still limited in fish virology (Hammoumi et al., 2016).

Considering the current state of the art, it is clearly necessary to optimize both library preparation, to increase the probability of sequencing the targeted viruses, and bioinformatics tools, to efficiently exploit the available data.

### 3.3 Bioinformatics challenges

The complexity of analyzing fish viruses is intrinsically linked to their genomic diversity. Fish species represent about half of all known vertebrates and have colonized very different aquatic ecosystems, from marine to freshwater. This tremendous diversity of hosts and environments has certainly contributed to the extraordinary wealth in viruses that has been reported, particularly over the last decade, and which is undoubtedly only a small part of the existing virome.

#### 3.3.1 Identification of viral organisms

The polyphyletic nature of viruses makes it even more complicated to find the proverbial needle in the haystack. Contrary to the 16S ribosomal RNA gene that is commonly used for bacterial taxonomic classification, there is no known sequence shared among all viruses (Mizrahi-Man et al., 2013; Chaudhary et al., 2015). A recent discovery revealed that ribonucleotide reductases (RNRs) can be used as potential viral markers because they are widely distributed among dsDNA viral families (Sakowski et al., 2014). For instance, the predicted nucleic acid distribution among viroplankton indicates that dsDNA viruses can comprise 70% of the total viruses and that 93% of dsDNA viroplankton may contain RNRs (Bidle, 2014; Sakowski et al., 2014). Moreover, RNRs are present in other viral lineages such as Cyanophages, Iridoviruses, Cyanopodoviruses, and many others.
However, ensuring accurate identification of short sequences (singleton and/or contig(s)) is difficult at best when no homologs are available in databases. Classification tools such as CLARCK or GOTTCHA can overcome these hurdles with their target-specific or unique signature approach (Freitas et al., 2015; Ounit et al., 2015). However, unique signature models can be altered depending on how big the genome is and how sequence divergence is distributed across the genome. Consistently unique or specific k-mers will fail to be recognized in targeted reads if divergence is dispersed instead of being grouped throughout the genome. In parallel, standard mapping tools (e.g. Bowtie, (Langmead et al., 2009)) are not suitable for this situation, because they fail to correctly assign some reads related to similar virus types. Most commonly used mapping tools assume low divergence between datasets and references, and only some are more flexible regarding error rates, insertion/deletions (indels) or substitution (Caboche et al., 2014). New tools have been developed to overcome this issue and can correctly classify reads displaying high divergence with reference sequences; (MaxSSmap and NexGenMap (both graphics processing unit programs) or Stampy (hash-based program) tolerating divergence up to 30%, 15% and 10%, respectively (Lunter and Goodson, 2011; Sedlazeck et al., 2013; Turki and Roshan, 2014).

3.3.2 Lack of dedicated databases and tools to limit the amount of unknown data
Metagenomic sequencing gives a clear and sharp view of the microbial community present in a given sample. With its ability to detect uncultured microorganisms, this sequencing method can characterize the abundance and diversity of microorganisms that compose a sample and lead to a better understanding of the biological interactions in a community. Identifying new sequences relies heavily on the availability and quality of databases. Therefore, database curation is of prime importance. Bioinformatics tools and databases dedicated to viral metagenomics have been reviewed elsewhere (Alavandi and Poornima, 2012; Dolled-Filhart et al., 2013). In terms of virus discovery, the purpose of these tools is to help identify viral species and characterize their diversity and their function. In this regard, there are two basic approaches to analyze metagenomics samples after sequence filtering. The first is a taxonomic classification based on reads to determine the diversity of species present followed by the assembly of organisms of the selected taxa (Alavandi and Poornima, 2012). The second involves de novo assembly based on the entire pre-processed set of reads where resulting contigs are classified (Meiring et al., 2012; Thomas et al., 2012). In both cases, once a full genome has been covered, its characterization usually involves the prediction of possible ORFs and the comparison with those found in protein databases. Although efforts have been made to develop tools for viral discovery pipelines, namely ViromeScan and MetaVir2 (Roux et al., 2014; Rampelli et al., 2016), database integration still requires improvements. The size of viral aquatic databases is growing and has the potential for tremendous increases because a vast number of viral organisms have yet to be discovered. However, these databases should not expand as a single structure isolated from other biological databases. Accordingly, Vinithkumar et al. propose a multi-linked approach, where databases could have their own internal compartmentalized structure while being integrated with one another (Vinithkumar, 2006). No matter how well curated, sequences without any similarities to any species recorded in the available databases remain unresolved. Due to the limits of similarity-based methods, new strategies have been developed to decipher this “unknown fraction”. Because unrelated sequences are still a serious impediment to the understanding of marine viral communities, protein clustering (PC) has been used to organize the unassigned sequences. PC consists of predicting ORFs and then grouping sequences regarding their ORF similarities. Hurwitz et al. (2013) applied this method first as a way of
mapping the ocean virome (Hurwitz and Sullivan, 2013). Beyond sorting the unknown sequences, they highlighted the variation in functional viral diversity according to sampling parameters such as season, distance to shore and depth. In the same year, Ignaio et al. asserted that the marine viral diversity may be lower than the commonly accepted paradigm (Ignacio-Espinoza et al., 2013). The number of PCs as a function of the number of metagenomes appears to reach saturation near 30 metagenomes, showing that diversity stagnates despite the increasing number of genomes sequenced. In 2014, Hurwitz et al. defined “core” PCs as the ones found in every sample of the Pacific Ocean virome (POV) and demonstrated that those “core” PCs were more abundant in photic zones than in aphotic zones, assuming gene flow between surface waters and the depths (Hurwitz and Sullivan, 2013; Hurwitz et al., 2015). These examples demonstrate the type of information that clustering methods can provide. Another approach relies on artificial neural networks (ANNs) (Seguritan et al., 2012) for the identification of viral structural proteins, which are present in all viruses. These ANNs correctly classified major capsid and tail proteins of phages. The classification has been validated experimentally and turns out to be more sensitive than a commonly-used pattern analysis tool, MEME (Bailey et al., 2009). However, this procedure is not straightforward because results may be distorted by network architecture (number of neurons and hidden layers), data complexity, or the training and validation datasets. Given that sequences populating those sets come from available databases, the whole process and its accuracy still depend, to a certain extent, on what is accessible. Considerable efforts have been made to label the unknown fraction and those new tools and methods are paving the way to a clearer picture of the aquatic viral world.

### Including the host in the analysis

Likewise, another major challenge arises when viruses share homologies with their hosts. Here, recoverable viral data is hampered, because, insofar as possible, viral sequences must be distinguished from cellular signals. For example, the increasing number of complete genomes from marine viruses has led to the identification of auxiliary metabolic genes (AMGs) in cyanophages (Roux et al., 2014). In addition to bona fide viral genes, viral genomes may also contain genes coding for proteins involved in photosynthesis (Mann et al., 2003), overturning the idea that viruses only carry genes responsible for viral reproduction. These genes, shared by viruses and bacteria, need to be differentiated to retrieve correctly only the virome component of the studied samples. Consequently, new computational approaches that discriminate viral and host versions of photosynthesis genes have emerged (Tzahor et al., 2009). The use of support vector machines (SVMs) has helped to solve such taxonomic problems. SVMs are learning algorithms designed for binary qualitative variable prediction. Determining a classifier or a discrimination function with the greatest forecast quality guides the prediction. Here, discriminating viral from bacterial genes is similar to discriminating a dichotomous variable; in this case, SVMs are able to correctly separate fragments without adhering to previous observations. They have been used to classify core photosystem-ii genes of cyanobacteria and cyanophages, and were even tested on partial sequences with good performance (Tzahor et al., 2009).

### Challenges in assembling fish viral genomes in metagenomics

In aquatic viral metagenomics, whole genome sequencing is required for better virus characterization. Unfortunately, high genetic diversity implies more complexity in the assembly process (Miller et al., 2010). “High diversity” can involve (i) high inter-taxon diversity, whereby a taxon has many members that are not necessarily genetically remote one from another or (ii) high intra-taxon diversity, whereby a taxon shows large dissimilarities between members. Each situation can lead to different approaches regarding assembly (figure
2) or taxonomic classification. First, for genome assembly, high intra-taxon diversity is bothersome in cases of multi-infection, because assembling short reads from close viruses from the same taxon can lead to chimeric contigs. Closely related sequences lead to similar k-mers, resulting in false paths in the graph assembly (Yang et al., 2012;Archer et al., 2014;Toseland et al., 2014). A sample of water containing such genetically close viral particles would thus be assimilated to a multi-infection case. In contrast, high inter-taxon diversity can ease the assembly step since k-mers in the data set would be quite distinct from one another. Additionally, uneven coverage is also problematical for the assembly step because coverage variations may result in loss of accuracy. Many assemblers assume uniform coverage, with some of them removing low coverage regions, yielding fragmented assemblies (Miller et al., 2010;Chen et al., 2013;Gawad et al., 2016). Assemblers successfully used in single cell sequencing (SCS) should be preferred since they have been designed for handling non-uniform coverage (Chitsaz et al., 2011;Bankevich et al., 2012;Peng et al., 2012).

3.3.5 Bioinformatics regarding third and fourth generation sequencing

Novel sequencing technologies (3G and 4G) rely on single molecules, and as such, do not require any amplification step prior to sequencing. In addition to removing the potential biases induced by amplification (Lopez-Barragan et al., 2011;Van Nieuwerburgh et al., 2011), these technologies will very soon change challenging issues regarding bioinformatics analysis. For instance, the read assembly issue will no longer deal with “putting small parts together in the right order” since longer reads will be produced with 3G and 4G sequencing methods. Assembly of 100 bp sequences is not the same problem as assembly of 1000 bp sequences, especially concerning k-mer and overlap sizes. Longer reads should facilitate assembly because longer k-mers involve less computational operations; for overlap layout consensus-based methods, it will require fewer operations to compute overlaps and less ambiguity related to repeats; for the de Bruijn graph-based approaches, it means less edges and possibly less ambiguity (bulges and alternative paths). Also, most medium-size repeats (couples of kbp) will be solved with longer k-mers. Nevertheless, current mapping tools will also benefit from 3G/4G sequencing. K-mer size depends on read size. Hence, assembly with larger k-mer sizes and fewer reads should be faster. More specifically, 3G/4G provide a single copy of a genomic region, whereas 2G returns a larger depth of coverage with multiple copies of given regions. Therefore, mapping with 3G/4G reads will require fewer operations by avoiding redundant computation. Even assembly validation using read mapping will be irrelevant knowing that coverage will no longer be an important issue. When it comes to the taxonomic classification of reads, the result can be questionable with low-load entities, which may be represented by a few reads and in extreme cases by only one read. Classification based on only one short read is not very reliable, particularly in cases of multi-infections. Obviously, longer reads allow more robust classification calls. Organisms identified with a very low proportion of reads are commonly thought to result from contamination. For this reason, contamination is a thorny issue because the difference between low loads and contamination cannot be solved. Unfortunately, no sequencing technologies are 100% error-free and 3G/4G is no exception (Lee et al., 2014;Feng et al., 2015;Rhoads and Au, 2015). Therefore, assembled genomes from longer reads that may contain errors, can be misinterpreted as possible variants. Single error-prone copies of genomic regions can lead to calling false single nucleotide polymorphisms (SNPs), contrary to 2G with sufficient coverage depth for reliable SNP calls. New bioinformatics challenges will therefore arise with the emergence of 3G/4G.

3.4 Portability and field sequencing
Another feature of 4G sequencing that may be particularly relevant to aquatic/fish virology is its compatibility with field experiments. In this regard, the ultra-portable MinION device from Oxford Nanopore Technologies showed good performances in terms of rapidity and robustness in difficult environments (Quick et al., 2016; Johnson et al., 2017). Combined with the newly launched VolTRAX device for automated library preparation, Oxford Nanopore Technologies is about to provide a complete and fully portable sequencing workflow for non-laboratory environments (https://nanoporetech.com/publications/voltrax-rapid-programmable-portable-disposable-device-sample-and-library-preparation). Although this technology still requires a deep evaluation, especially with regards to the nature of input material and the limit of detection, it seems very promising for the quick detection and identification of viruses from water or fish samples. Finally, single molecule-based technologies will also dramatically reduce the required amount of input DNA (or RNA), and in turn increase the ability to detect viruses at low concentrations.

4. Applications of NGS to fish virology: some successful examples

4.1 Discovery of new viruses
An interesting application of NGS is the search for viruses as etiological agents of new syndromes. This approach is extremely interesting when all the traditional methods to identify viruses have failed and when the viral pathogen is not culturable in vitro. One example is the heart and skeletal muscle inflammation (HSMI) of salmon, a fatal disease that appeared in 1999 in hundreds of farms in Norway and the United Kingdom and remained unexplained for a decade. The presence of an infectious agent was demonstrated by horizontal transmission experiments, but it was only in 2010 that a virus was identified using NGS. Total RNAs were extracted from the heart (a putative target of this agent) and serum, and were pyrosequenced. This approach led to the identification of a new virus related to the Reoviridae and named piscine reovirus (PRV) (Palacios et al., 2010). Although Koch’s postulate was not fulfilled, several pieces of evidence point to the implication of PRV in the syndrome. During the same period, a second novel virus was identified by NGS from salmon affected by a syndrome called cardiomypathy syndrome (CMS), similar in some aspects to HSMI. Using GS FLX titanium chemistry, sequences identical to those of PRV were found, but a new virus, related to the totivirus genus, was also identified in the same samples (Lovoll et al., 2010; Haugland et al., 2011). Further research suggested that this piscine totivirus was the causative agent of CMS whereas PRV may only have been an opportunistic agent. Indirect evidence involving the totivirus was the absence of any other viral sequences in the massive set of reads produced. Another example is the recent description of the so-called tilapia lake virus (TiLV). Ion Torrent sequencing of cDNA libraries built from RNA extracted from the brain of diseased tilapia led to the identification of an orthomyxo-like virus as the etiological agent of a disease responsible for mass mortalities of tilapia in Israel (Bacharach et al., 2016). The segmented RNA genome of this new virus, tentatively named TiLV, shows almost no homology with any known sequence, except for one segment. Surprisingly, sequencing the genome of a TiLV PCR-positive sample from a diseased fish in Ecuador showed >97% nucleotide identity with the Israeli virus. Given that the clinical signs observed in Ecuador were very different from those observed in Israel, these findings indicate that a similar (or nearly similar) viral agent may be responsible for two diseases in two distant geographical locations. These data also highlight how an emerging virus can rapidly become a global threat for an important aquaculture species.

Although from these examples it appears that NGS may provide clues to explain new pathologies, complementary evidence is still needed to link formally identified viral
sequences to specific symptoms and to verify Koch’s postulates. This issue is well illustrated by the recent findings of three viruses from different families in a single sea bream (Sparus aurata) affected by lymphocystis disease, which is frequent in the Mediterranean region and generally attributed to an iridovirus (LCDV). In this study, the combination of 454-Roche and Illumina sequencing identified not only a LCDV genome very distinct and much longer than the ones already known, but also new viruses from the papilloma- and a polyomavirus families. This finding raised the issue of the respective roles of these new viruses in triggering the disease (Lopez-Bueno et al., 2016).

### 4.2 Contribution of full-genome sequencing to taxonomy and classification

Another application of NGS is the full-length sequencing of already described viruses, for which complete genomes are likely to provide more complete phylogenetic and/or epidemiological information. In a recent study, full-genome sequencing of the European sea bream virus, a ranavirus from the Iridoviridae family, was useful to genetically distinguish it from EHNV, another ranavirus (Mavian et al., 2012). Before this sequencing work, ESV and EHNV were differentiated in testing laboratories by sequencing a short conserved region of the major capsid protein, both viruses differing by only few nucleotides in this portion. The alignment of full-length EHNV and ESV genomes show many more differences at the genome scale, such as indels and inversions, which will be useful for designing a simple PCR-based diagnostic test not relying on sequencing. Full-genome sequencing has also successfully been applied to cyprinid herpesvirus 2 (CyHV-2), which is responsible for a severe pathology in goldfish (Davison et al., 2013) and more recently for cyprinid herpesvirus 3 (CyHV-3) (Li et al., 2015). The complete sequence of the 290 kb long CyHV-2 genome has been compared with those of two other cyprinid herpesviruses, namely CyHV-1 and CyHV-3, infecting the common and ornamental carps (Cyprinus carpio). Partial sequences previously obtained by PCR as well as distinct biological properties can distinguish the three herpesviruses. However, alignment of their complete genomes revealed significant genetic differences, with genes or blocks of genes rearranged between the viruses. For instance, ORF4 present in the terminal repeat of CyHV-3 is inverted and located in the unique region of CyHV-2.

Although whole-genome sequencing has helped to distinguish CyHVs in a phylogenetic perspective, the poor understanding of the function of the genes prevents the elucidation of the biological differences between species. However, the complete sequencing of a series of variants of CyHV-3, possibly less pathogenic than the typical virulent strains and for which only short sequences are available, would certainly help formulate hypotheses on the function of some genes involved in virulence (Engelsma et al., 2013).

### 4.3 Evolution and virulence

NGS greatly improves the understanding of the evolution of piscine viruses, and more specifically their pathogenicity/virulence. The infectious salmon anemia virus (ISAV), a devastating virus in the salmon industry, is the best example of the use of NGS in the study of the evolution of virulence in a fish virus. ISAV is an orthomyxovirus whose major surface glycoproteins, fusion protein (F) and hemagglutinin-esterase (HE), have a crucial function in virus uptake and the release of viral particles, respectively (Falk et al., 2004). The HE protein contains a highly polymorphic region (HPR) whose size has a strong impact on the virulence of the virus. A full-length HPR (HPR0) is associated with avirulent strains, not culturable and primarily found in the gills, while shortened HPR versions are associated with virulent viruses (vISAV), systemic in the fish and responsible for severe losses. In addition to the HPR of HE, a single amino acid in the F gene is correlated with virulence. A common view is that the avirulent strains circulate in reservoirs and that some deleted variants appear. Although the
mechanisms of action of the proteins are rather well understood, the genomic transition of full-length viral populations to strains harboring deletions was poorly documented until recently (Fourrier et al., 2015). By performing ultra-deep sequencing targeting the F and HE genes of viruses in the same location before and after an outbreak, the composition of different quasi-species could be elucidated, which in turn led to several hypotheses regarding their link with the different patterns of virulence (Markussen et al., 2013). Intra-sample genetic variation, not detected by Sanger sequencing, was found for each type of sample. Surprisingly, there were more mutants in samples before the outbreak than after. Mutations were more concentrated in the HPR region compared with the other parts of F and HE. A total of 14 deletion types were observed. Importantly, both deleted and non-deleted variants were found in fish before and after the outbreak. Moreover, low-frequency reads identified deleted variants in fish before the outbreak, a result with important consequences in terms of diagnosis in cases where the PCR-Sanger method did not detect these variants. Among all the variants found in fish before the outbreak, one carried a single mutation which was of particular interest because it dramatically changed the predicted secondary structure of HE RNA. The authors suggested that this new structure was more prone to deletion events. This hypothesis suggests that the intra-sample genetic diversity includes variants whose potential for change to higher virulence may differ.

There is less documentation on VHSV, a novirhabdovirus responsible for severe outbreaks in farmed rainbow trout. There are four main genotypes (I-IV) and a considerable number of variants, some with very different virulence levels. Historically, the virulent strains were thought to originate from avirulent strains circulating among wild marine fish before spreading through the trade of infected fish. It was surmised that a few mutations were responsible for changes from low to high virulence (Campbell et al., 2009; Kim et al., 2014). However, it has not been established whether these virulent variants pre-existed in wild fish or if they arose and were selected for in farmed trout. In a recent study, ultra-deep sequencing on four “isolates” of VHSV passaged in vitro (Schonherr et al., 2016) showed that the number of variants within a specimen is very variable, from 100 to only three. No fewer than 89 polymorphic sites were observed for single isolate specimen. The authors deliberately excluded poorly represented (<5%) variants which may have arisen from artefacts. Although the links between virulence and sequence diversity need to be clarified, these findings illustrate the great diversity of genotypes that can co-occur within a single specimen. They also highlight the importance of sequencing viral genomes directly from their hosts, especially for gaining insights into their evolution and pathogenesis. To this end, targeted genomic enrichment (TGE) may be very useful to overcome the high discrepancy in the proportions of cellular and viral genomic DNA in raw samples obtained from infected tissues. TGE has been described for a large range of applications (Mertes et al., 2011), but surprisingly, the latter in virology has remained very limited. As far as we know, the only examples of successful use involve human herpesviruses (Depledge et al., 2011; Depledge et al., 2014). In the latter study, authors accurately quantified the allele frequencies of viral populations contained in herpes zoster skin lesions and studied the pathogenesis of varicella-zoster virus and its attenuated vaccine strain. In a very recent study, 10 complete genomes of CyHV-3 were recovered directly from infected carp tissues in one single run of sequencing, with mean read depth varying from 219x to 9179x (Hammoumi et al., 2016). Using this strategy, it now seems possible to study the dynamics of in vivo infections and to trace outbreaks of this deadly virus. For comparison, 24 genomic DNA libraries of Epstein-Barr virus (EBV) were needed to recover one full genome from a tumor tissue with an average read depth of 17x (Lui et al., 2011).

NGS was also recently applied to identify microRNAs (miRNAs) in cells infected by CyHV-3, which displays the largest genome among herpesviruses known so far (295 kbp).
Preliminary bioinformatics analysis of the viral genome suggested the presence of pre-miRNA, and deep-sequencing of the microtranscriptome confirmed the existence of some of the predicted miRNAs (Donohoe et al., 2015). Two pre-miRNAs were found in cells infected in vitro and the corresponding miRNAs were detected in infected animals. There is experimental evidence that the most highly expressed miRNA are able to regulate the expression of ORF123, whose function is unknown. Furthermore, some of these miRNAs are conserved in the two other cyprinid herpesviruses, CyHV-1 and CyHV-2, suggesting an important function of these sequences. Although much work is still needed to understand the role of miRNAs in the virulence of CyHV-3, and of other aquatic viruses, NGS definitely opens new avenues of investigation.

4.4 Characterization of viromes through metagenomics

The presence of fish viruses in the water has been rarely documented due to their low concentrations in the environment. Metagenomics provides new opportunities to detect already-known viruses as well as to discover new viruses, both in marine and freshwater environments. As such, this type of approach has great potential for epidemiology because it offers the possibility to study the persistence and/or the spread of pathogenic viruses. One application is the search of pathogens in ballast waters which are transported worldwide and contaminate new ecosystems. In this regard, the presence of nucleic acids related to several pathogenic viruses was investigated in the ballast waters of boats in the North American Great Lakes region (Kim et al., 2015). It is estimated that 6.6 million tons of water is discharged annually from boats in the Great Lakes, which poses a serious threat for the introduction of new organisms and pathogens in the whole region. The emerging genotype IV of VHSV may have been introduced this way. In an attempt to characterize the viromes present in the ballast waters, viral nucleic acids extracted from 60 L sampled from several boats were sequenced. Among all the contigs identified, several exhibited genetic links either with a betanodavirus, a megalocytivirus or CyHV3, which are all among the most economically important fish viruses. Unfortunately, the complete sequences of these viruses could not be obtained, probably due to an insufficient number of reads and/or an insufficient volume of water analyzed; combined with the difficulty of assembling rare sequences in contigs, as explained above. The first obstacle will certainly be overcome in the near future with the dramatic increase of the number of reads forecasted with the emergence of 3G and 4G technologies. Therefore, the number of reads appears to have crucial importance, not only to ensure an adequate coverage of a targeted viral genome (as highlighted in the previous section), but also to reduce the probability of missing rare sequences.

5. Conclusions and avenues for future research

To conclude, the different technical approaches for virological diagnosis have improved according to scientific discoveries and technological innovations. Traditional techniques, such as cellular culture and immunological methods, have significantly contributed to the identification of numerous pathogens during the last century, whereas modern molecular methods such as PCR or microarrays have played an ever increasing role in clinical virology in the last decade. Methods based on metagenomics have allowed significant advances in the identification of pathogens directly from a biological sample, bypassing the need for cell cultures or for any a priori knowledge on the potential agents present. NGS and bioinformatics analysis have now become essential tools for viral pathogen discovery, for screening viromes and for monitoring potential emerging pathogens. In this regard, the recently developed portable sequencers (4G) offer new perspectives by their ability to rapidly produce sequencing data outside laboratory facilities. Although the error rate is high today,
one can reasonably expect that technological improvements will soon be made to decrease this rate to an acceptable range. If this is the case, this new generation of sequencers will dramatically increase our capacity to monitor potential emerging pathogens. However, associating a candidate pathogen with a disease will still remain a key issue, particularly considering Koch’s postulate or Rivers’ modification (Rivers, 1936; Byrd and Segre, 2016). In practice, these postulates are not always entirely met because some viruses still resist isolation and culture. The initial criteria have since been revised to include the use of nucleic acid-based detection methods (Fredricks and Relman, 1996; Lipkin, 2013). Sequencing without prior knowledge of what is in the sample widens the scope of viral discovery. With its great diversity and abundance, the marine/aquatic virome may harbor viral pathogens that can trigger an epidemic in fish, and potentially in other organisms. Consequently, increasing our knowledge on fish and other aquatic viruses requires the development of sharper methods and tools for both molecular and bioinformatics analyses, which fit with the ever-changing sequencing technologies.

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
We acknowledge Heïdi Lançon and Dr. Carolyn Engel-Gautier for her revision of English usage. The CIRMF is supported by the government of Gabon, Total-Fina-Elf Gabon, and the Ministère de la Coopération Française. This review article was also made possible by the ERA-NET EMIDA funded project MOLTRAQ (molecular tracing of viral pathogens in aquaculture)
Legends to figures

**Figure 1.** Marketed sequencing platforms: from the first to the fourth generation. The chemical principle of the main sequencing technologies is illustrated. Information on the manufacturer, benchtop model and sequencer performance is also given. A. First-generation sequencing is represented by the Sanger technology implemented and released by Applied Biosystems. B. Second-generation sequencing is illustrated by three distinct technologies from different manufacturers: pyrosequencing by Roche, semiconductor (pH based-) sequencing by Life Technologies and synthesis with reversible chain terminators by Illumina. C. Third-generation sequencing uses single-molecule, real-time sequencing commercialized by Pacific Biosciences. D. Fourth generation offers single molecule, real-time and even portable sequencing, commercialized by Oxford Nanopore Technologies.

**Figure 2.** Viral genome assembly and viral diversity. Longer "true" contigs are produced when the genomes included in the sequencing sample share fewer similarities (high inter-taxon diversity) than in a multi-infection-like case (high intra-taxon diversity). In the latter case, contigs are generally shorter and some may be chimeric.


Dunlap, D.S., Ng, T.F., Rosario, K., Barbosa, J.G., Greco, A.M., Breithart, M., and Hewson, I.


Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K.,
Dewar, K., Doyle, M., Fitzhugh, W., Funke, R., Gage, D., Harris, K., Heaford, A.,
Howland, J., Kann, L., Lehoczky, J., Levine, R., Mcewan, P., McKernan, K., Meldrim,
J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rossetti, M., Santos,
R., Sheridan, A., Sougnez, C., Stange-Thomann, Y., Stojanovic, N., Subramanian, A.,
Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Bentley, D., Burton, J.,
Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I.,
Durbin, R., French, L., Graham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt,

correction and assembly complexity of single molecule sequencing reads. bioRxiv.

Sequencing for Infectious Disease Diagnosis and Management: A Report of the
Association for Molecular Pathology. J Mol Diagn 17, 623-634.

Comparison of next-generation sequencing systems. J Biomed Biotechnol 2012,
251364.


Effect of PCR extension temperature on high-throughput sequencing. Mol
Biochem Parasitol 176, 64-67.

(2016). Concordance of iridovirus, polyomavirus and a unique member of a new
group of fish papillomaviruses in lymphocystis disease affected gilthead

Lovell, M., Wiik-Nielsen, J., Grove, S., Wiik-Nielsen, C.R., Kristoffersen, A.B., Faller, R,
Poppe, T., Jung, J., Pedamallu, C.S., Nederbragt, A.J., Meyerson, M., Rimstad, E., and


Van Nieuwerburgh, F., Soetaert, S., Podshivalova, K., Ay-Lin Wang, E., Schaffer, L.,
bias in Illumina TruSeq and a novel post amplification barcoding strategy for


Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O.,
Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D.,
Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T.J., Higgins, M.E., Ji, R.R., Ke, Z., Ketchum, K.A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G.V., Milshina, N.,
*Science* 291, 1304-1351.

Villena, A.J. (2003). Applications and needs of fish and shellfish cell culture for disease
control in aquaculture. *Reviews in fish biology and fisheries* 13, 111-140.


sequencing of an influenza genome. *Front Microbiol* 6, 766.

Welch, A.B. (1971). Purification, morphology and partial characterization of a reovirus-


Yang, X., Charlebois, P., Gnerre, S., Coole, M.G., Lennon, N.J., Levin, J.Z., Qu, J., Ryan, E.M.,
Zody, M.C., and Henn, M.R. (2012). De novo assembly of highly diverse viral
populations. *BMC Genomics* 13, 475.


disease of the giant freshwater prawn Macrobrachium rosenbergii in Thailand.
*Dis Aquat Organ* 69, 255-258.