Development and application of -omics and bioinformatics approaches for a deeper understanding of infectious diseases systems

INAUGURALDISSERTATION

Zur

Erlangung der Würde eines Doktors der Philosophie

Vorgelegt der

Philosophisch-Naturwissenschaftlichen Fakultät

Der Universität Basel

Von

Pierre H. H. Schneeberger

Ochlenberg (BE) und Frankreich

Basel, 2017

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät auf Antrag von

Prof. Dr. Jürg Utzinger, Dr. Jürg E. Frey und PD Dr. Mauro Tonolla

Basel, den 13. Oktober 2015

Prof. Dr. Jörg Schibler

Dekan der Philosophisch-Naturwissenschaftlichen Fakultät

Acknowledgements

- My deepest thanks go to my four mentors, Jürg Frey, Jürg Utzinger, Christian Beuret and Joël Pothier for giving me the freedom of action as well as the required guidance to go through these amazing years. You gave me the opportunity to pursue this PhD and allowed me to bring in my own ideas – this was really a golden opportunity given to a PhD student! You were always open for discussion and there to provide support throughout the various steps of this PhD – sincere thanks for everything.
- I would especially like to thank my External Referee, Dr. Mauro Tonolla, for accepting to review my work in such a short notice.
- “Mes meilleurs remerciements” to all who contributed significantly within the different projects: Dr. Jean-Sebastien Reynard, Prof. Eliézer N’Goran, Dr. Brion Duffy.
- Special thanks to Sören Becker, Samuel Führimann and Andreas Bühlmann for the excellent discussions and collaborations within my different projects. The three of you have, all in different ways, positively influenced me and provided me with exceptional opportunities to further expand the scope of my thesis.
- An enormous thank to the people who helped me so much during my laboratory work, and this includes mainly, but is not limited to, Markus Oggenfuss, Beatrice Frey, Dr. Oliver Engler and Jasmine Portmann.
- Special thanks to Geoffrey Williams who kindly corrected this document.
- To all my old long-time friends from the “french” side.
- A final thank goes to my parents, Christine and Georg and to my sister and brother, Séverine and Olivier who have always supported me through sometimes difficult times during my research!

Table of Contents

Pour mes grands-parents, Roswitha et Henri, qui ne sont plus là pour partager ce moment, mais qui m'ont donné l'enthousiasme et l'envie d'arriver jusque-là…

Abbreviations

NGS = Next-generation sequencing

RDT = Rapid-diagnostic test

DNA = Deoxyribonucleic acid

RNA = Ribonucleic acid

Mb = Megabase

Mbp = Megabase pairs

Kbp = Kilobase pairs

PCR = Polymerase chain reaction

rt-PCR = Real-time polymerase chain reaction

LAMP = Loop-mediated isothermal amplification

HIV = Human immunodeficiency virus

AIDS = Acquired immune deficiency syndrome

HBV = Hepatitis B virus

HCV = Hepatitis C virus

HDV = Hepatitis D virus

mm = millimetres

µm = micrometres

MERS-CoV = Middle East respiratory syndrome

BLAST = Basic local alignment search tool

List of tables

List of figures

Abstract

Background: Research in infectious diseases underwent a revolution with the uprising of Omics approaches, including, but not limited to, genomics, metagenomics and metatranscriptomics. In fact, there are several examples where Omics approaches showed their potential to tackle different challenges related to the versatile nature of infectious diseases by promoting “studies of one” to “system-wide studies”. In the frame of this PhD programme, we focused on the development and validation of Omics approaches and bioinformatics workflow aiming at tackling mainly diagnostics but also to some extents the treatment of infectious diseases. The four applications presented in this thesis had following specific objectives; (i) to develop and validate a bioinformatics approach aiming at selecting high quality markers among a large amount of complete genomic sequences; (ii) to characterise the viral metagenome of a plant to determine aetiology of a disease that could not be identified and/or fully characterised with other tools; (iii) to assess the potential of metagenomics in the field of personalised medicine and compare its diagnostics accuracy with validated diagnostics tools; and (iv) to make a system-wide survey of microbial populations and estimate its potential to cause harm to humans.

Methods: Methodology was specific for each application but as a general rule, we only used published bioinformatics tools that have been used and validated in other studies. This includes, but is not limited to, the BLAST algorithm for the comparison of sequences to various databases and the MIRA assembler to assemble the metagenomics datasets obtained within the different projects.

Results: For clarity, the results are summarised by project, corresponding to the different applications investigated during this PhD.

Project (i): The developed bioinformatics workflow allowed the selection of highly conserved and specific molecular markers among various viral species with inputs of up to several hundred complete genomic sequences. The quality of the selected markers was successfully validated using several types of molecular assays including real-time PCR, LAMP and Sanger sequencing.

Project (ii): We were able to find the aetiology of a grapevine plant presenting leafroll symptoms. A new virus, named Grapevine Leafroll-associated virus 4 Ob, with a thirteen kilobases genome was found in the viral metagenome. Other viruses that were co-identified in the virome were known to be asymptomatic viruses for grapevine, and with the help of additional serological experiences, we were able to confirm that this GLRaV-4 Ob was the causative agent of the Leafroll symptoms.

Project (iii): The gut pathobiomes from four patients presenting persistent digestive disorders were fully characterised using a metagenomics approach. Comparison of validated diagnostics tools with this approach showed that the diagnostics rate was in favour of the latter for the detection of bacterial and helminths pathogens and in favour of the validated tools for the detection of viruses and protozoa. Using the same datasets, but compared to a different database, we were also able to screen the stool samples for antimicrobial resistance genes and retrieve potential resistance genes that might interfere with the treatment of these patients.

Project (iv): In this project, a system-wide assessment of the microbial communities of the wastewater treatment system was done using a metagenomics approach. We were able to demonstrate how closely the genetic diversity of *Escherichia coli* and the overall genetic diversity were linked in this environment. We were also able to map the repartition of different pathogenic classes, including bacteria, helminths, intestinal protozoa and viruses as well as to show if and how human waterborne pathogens spread throughout this ecosystem.

Conclusion: Omics offer new strategies of how challenges, mainly related to the vast diversity within the research area of infectious diseases, can be tackled. Meta-analyses, like metagenomics or metatranscriptomics are the applications that benefited most from the use of Next-Generation Sequencing technologies, and they now allow system-wide studies where previous studies were only focusing on one parameter (one microbe or one specific gene for instance). However, these Omics approaches have their limitations, mainly due to the bioinformatics challenges they give rise to. As a general conclusion, it is foreseeable that, because of the increased amount of results they generate, Omics approaches, once matured, will be more widely used and will replace standard approaches in the field of infectious diseases.

Résumé

Contexte : La recherche en maladies infectieuses a subi une révolution avec l'avènement des approches Omiques, incluant mais n'étant pas limitées à, la génomique, la métagénomique et la métatranscriptomique. Les approches Omiques ont été utilisées pour aborder la diversité intrinsèque des maladies infectieuses et ont permis de passer des études limitées à un paramètre aux études de systèmes complets. Dans le cadre de ce doctorat, nous nous sommes concentrés sur le développement et la validation de ces approches Omiques ainsi que des pipelines d'analyse bio-informatique dans le diagnostic ainsi que certains aspects du traitement des maladies infectieuses. Le but des quatre applications testées durant cette thèse étaient ; (i) de développer et valider une approche de bio-informatique capable d'analyser un grand nombre de séquences dans le but de sélectionner des marqueurs moléculaires et de les valider à l'aide de différents tests moléculaires; (ii) de caractériser le métagenome viral d'une plante pour déterminer l'origine d'une maladie; (iii) d'analyser le potentiel de la métagénomique dans le domaine de la médecine personnalisée ainsi que de valider son potentiel de diagnostic; et (iv) de réaliser l'analyser microbienne complète d'un environnement complexe et d'estimer le risque qu'il présente pour la santé humaine.

Méthodologie : Les méthodes utilisées sont spécifiques pour chaque application mais en règle générale, seuls des outils de bio-informatique reconnus et publiés ont été utilisés. Ces logiciels incluent, mais ne sont pas limités, à l'algorithme de BLAST pour la comparaison de séquences à différentes bases de données ou l'assembleur MIRA qui a été utilisé pour assembler les données de métagénomique.

Résultats : Pour des raisons de clarté, les résultats ont été regroupés par projet.

Projet (i) : Le pipeline de bio-informatique a permis de sélectionner des marqueurs moléculaires hautement conservés et spécifiques pour différents pathogènes viraux parmi un grand nombre de séquences génomiques. La qualité de ces marqueurs a été validée en utilisant différents types de tests moléculaires.

Projet (ii) : Il a été possible de déterminer l'organisme responsable des symptômes observables sur un plant de vigne. Un nouveau virus, nommé « Virus de l'enroulement de la vigne 4 Ob » ou « GLRaV-4 Ob », possédant un génome d'environ 13 kilobases a été détecté dans le métagenome viral. Du fait que les autres virus détectés dans le virome sont connus pour ne pas causer de symptômes dans la vigne et à l'aide d'expériences supplémentaires, il a été possible de confirmer que le virus GLRaV-4 Ob est l'agent pathogène responsable des symptômes observés.

Projet (iii) : En utilisant une approche de métagénomique, il a été possible de caractériser le pathobiome intestinal chez des patients présentant des troubles gastro-intestinaux persistants. La comparaison du diagnostic est en faveur de l'approche métagénomique pour les pathogènes bactériens ainsi que les helminthes mais les outils de diagnostic standard permettent une meilleure identification des pathogènes viraux et des protozoaires.

Projet (iv) : Ce projet a permis, avec l'utilisation d'une approche de métagénomique, de caractériser les communautés microbiennes du réseau de traitement des eaux usées de la ville de Kampala, Ouganda. Il a été possible de démontrer que la diversité génétique d'*Escherichia coli* est intimement liée à la diversité génétique bactérienne générale dans cet environnement. Il a également été possible de répertorier géographiquement les

différentes classes de pathogènes ainsi que les principaux pathogènes transmis aux humains par contact direct ou ingestion de l'eau.

Conclusion : Les approches Omiques ont permis le développement de nouvelles stratégies permettant l'analyse de la diversité intrinsèque aux maladies infectieuses. Les méta-analyses, telle que la métagénomique ou la métatranscriptomique sont les applications qui ont le plus bénéficié de l'utilisation du séquençage de nouvelle génération et elles permettent maintenant la caractérisation complète de différents systèmes. Pourtant, ces approches Omiques ont leurs limitations qui sont principalement liées aux analyses bio-informatiques. En conclusion, il est plausible que ces approches Omiques, une fois optimisées, seront de plus en plus utilisées jusqu'à remplacer les approches actuellement utilisées dans le domaine des maladies infectieuses.

Chapter I. Introduction

1. Infectious diseases

a. Definition

Infectious diseases, also known as transmissible diseases or communicable diseases, are illnesses resulting from the infection of a host by a pathogenic microorganism. The spectrum of pathogenic microorganisms is extremely wide, resulting in the fact that any living organism, including plants, animals, as well as microorganisms, can become infected and hence, a symptomatic host. An overview of the principal areas in the field of infectious diseases research (Anderson et al 1992) is shown in **Figure 1**.

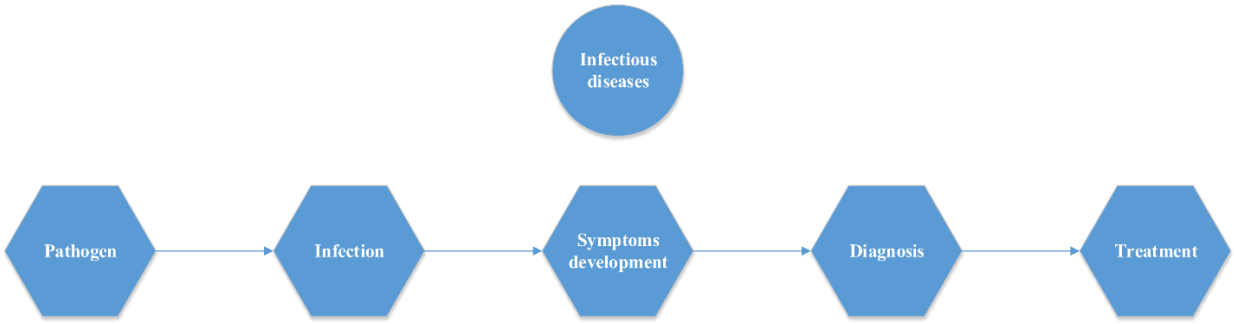


Figure 1. Areas of infectious diseases research. This figure represents the processes involved in infection, from the initial infection step to the final treatment step. Research focuses are similar for infectious diseases occurring in human, veterinarian and plant health.

b. Pathogens

The etymology of the word “pathogen” has a negative connotation, literally translating from Greek to “suffering producer” (*pathos* and *–genes*). A pathogen is a microorganism

which has the potential to infect a host organism and cause the symptomatic expression of a disease. Pathogens are, however, like any other living organism, only trying to survive and replicate (Alberts et al 2002). The strategy adopted by pathogens, as shown in **Figure 2**, is quite effective since it consists in using the hosts' energy or molecular machinery to achieve its own survival (Hilleman 2004, Hingley-Wilson et al 2003).

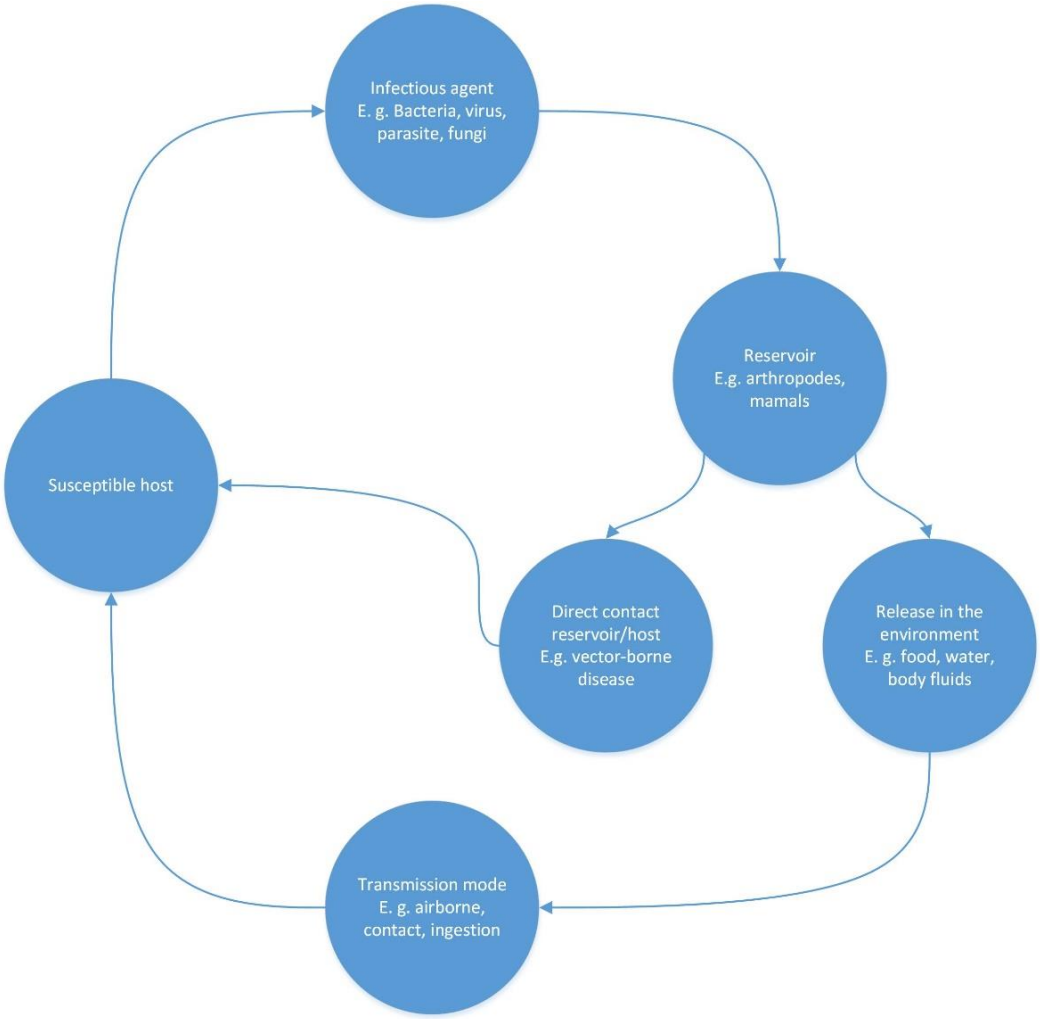
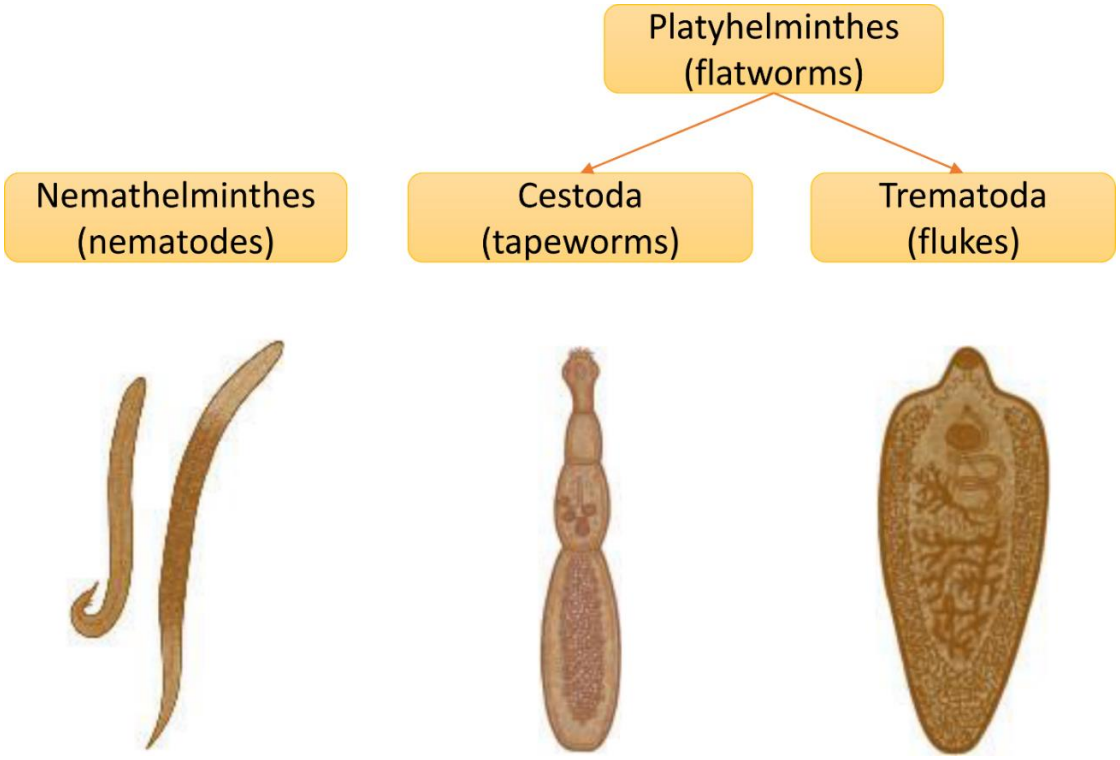


Figure 2. Generic lifecycle of a pathogen. Because of the broad diversity of pathogens, there is a wide range of variations and specificities in the lifecycles of each pathogen. These variations occur because pathogens need to adapt, among other things, to their respective reservoirs, environments, intermediate hosts, and final hosts.

c. Diversity of pathogens

From phylogenetic and phenotypic point-of-views, microbial pathogens are extremely diverse. In this brief overview, microbial pathogens have been sorted into five main classes, namely; (i) helminth parasites, (ii) protozoan parasites; (iii) bacterial pathogens, (iv) viral pathogens, and (v) fungal pathogens.

Helminths, also commonly known as parasitic worms, are large multicellular organisms which can be classified into three main groups, namely Nemathelminths, Cestodes and Trematodes as shown in **Figure 3**.

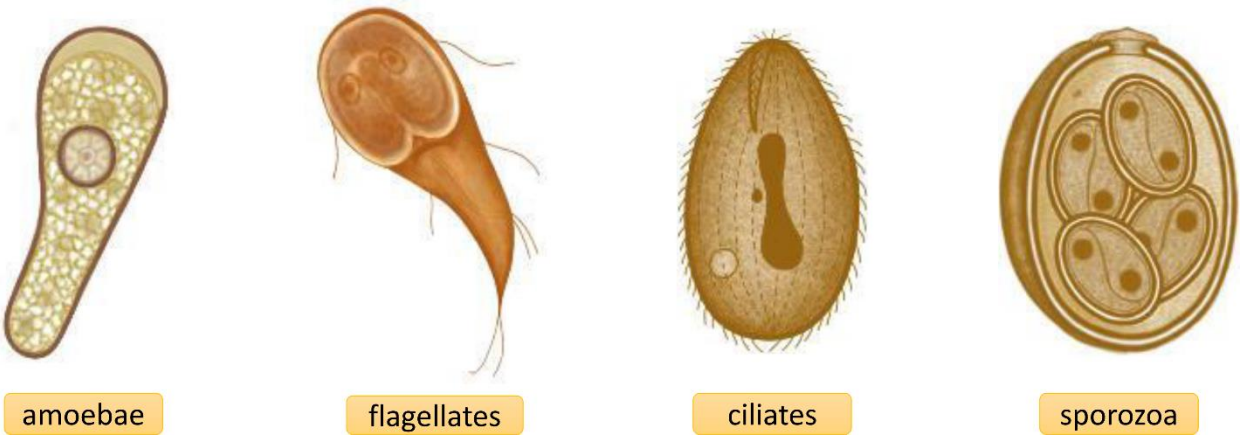


Source : http://parasite.org.au/para-site/contents/helminth-intoduction.html

Figure 3. Main groups of helminth parasites. These includes nematodes and flatworms, the second being divided into two subgroups, tapeworms and flukes.

Due to their higher complexity, genomes of helminth parasites have not yet been extensively sequenced, but estimates indicate that their genome sizes span between 50 and 500 Mb (Hotez et al 2008). However, their health impact is so important, with estimates of over 1 billion infected people, that genomics projects have become more and more common (Brindley et al 2009, Hotez et al 2008, Lustigman et al 2012) and high quality assembled genomes are expected to become available in the near future for a wider range of helminth species. In September 2015, 2'752'593 nucleotide sequences were available for flatworms as well as 1'955'922 nematodes sequences in the National Centre for Biotechnology Information sequence database, Genbank, which is the main sequence repository publicly available (Benson et al 2013).

Protozoa are unicellular eukaryotes which can be divided into four subgroups, based on their locomotion strategies, namely, (i) amoebae, (ii) flagellates, (iii) ciliates, and (iv) sporozoa as shown in **Figure 4.**

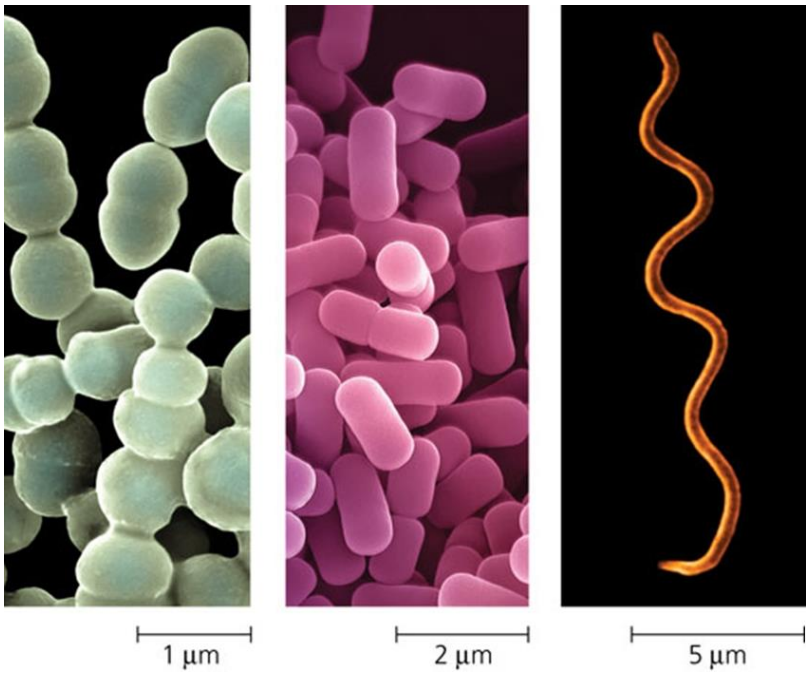


Source : http://parasite.org.au/para-site/contents/protozoa-intoduction.html

Figure 4. Subgroups of the protozoa embranchment. Protozoa are sub-divided in groups based on their locomotion strategies.

The World Health Organisation (http://www.who.int/) has identified ten major, yet neglected, infectious diseases (African trypanosomiasis, Chagas disease, dengue fever, lymphatic filariasis, leishmaniosis, leprosy, malaria, onchocerciasis, schistosomiasis, and tuberculosis) that are currently being intensively studied to provide control measures or even eradication measures for the causative agents. Four of them, namely, African trypanosomiasis, Chagas disease, leishmaniosis and malaria are caused by protozoan parasites and account for over 1.3 million deaths annually, possibly even more (Ersfeld 2003). So far, approximately 40'000 protozoa species have been described (Antonello 2007). The Welcome Trust Sanger Institute provides information on current and past protozoan sequencing projects and the genome sizes of completed projects span from approximately 8,3 Mb for *Theileria annulata* to over 62 Mb for *Neospora caninum*. To date, 84'958 protozoa nucleotide sequences are available in the Genbank database.

Bacteria are present in most of Earth's habitats and are found in various shapes including spheres, spirals and rods. Their size is typically between 0.5 and 5 µm as shown in **Figure 5** with some species, like *Thiomargarita namibiensis* reaching up to 0.75 mm (Schulz and Jørgensen 2001), making them visible to the naked eye.

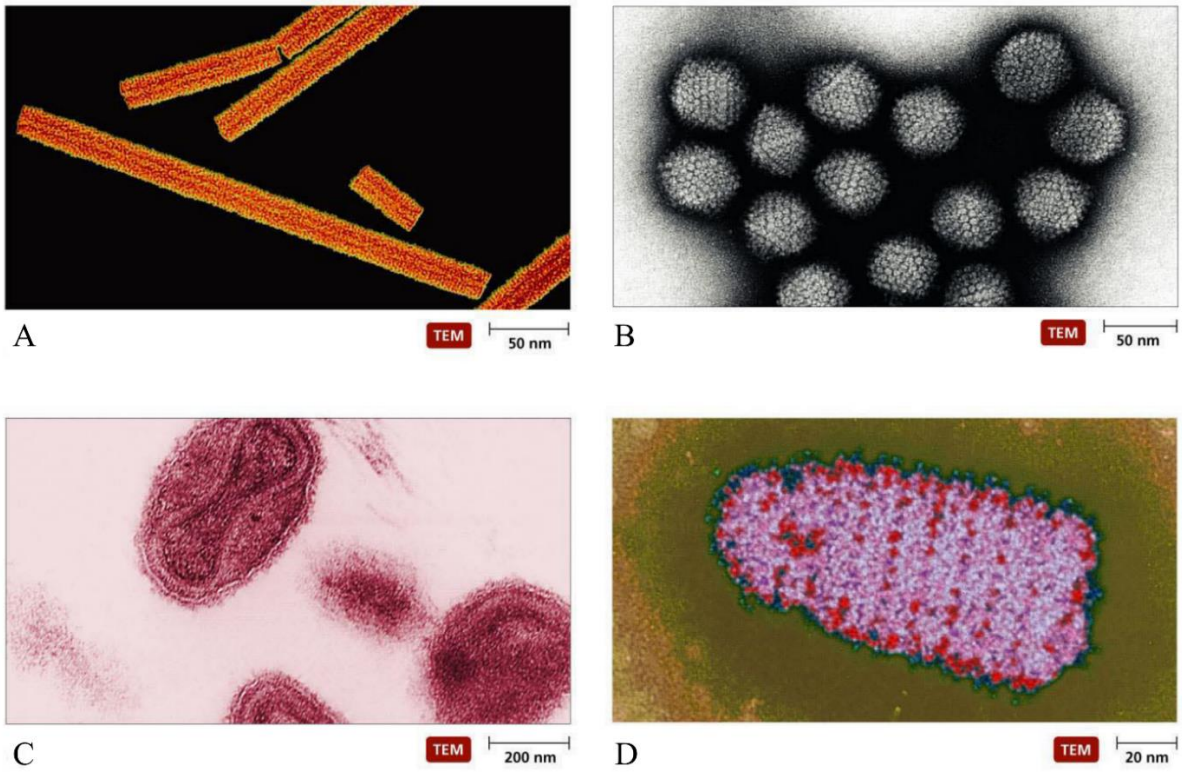


Source : http://www.ppdictionary.com/

Figure 5. Bacterial shapes and order of size. Left panel shows spherical bacteria, middle panel shows rod-shaped bacteria and right panel a spiral-shaped bacterium.

There are currently 15’974 bacterial taxa (Parte 2014) and bacterial genomes range from approximately 130 Kbp (McCutcheon and Moran 2012) to over 14 Mbp for *Sorangium cellulosum* (Han et al 2013). With 20’401’838 nucleotide sequences in the Genbank database, this is the most extensively sequenced of the five pathogenic classes presented here.

Viruses are the most important biological entities with an estimated 10^{31} viruses on Earth (Breitbart and Rohwer 2005, Edwards and Rohwer 2005). They are found in every type of ecosystem on this planet and they are present in a variety of shapes and sizes, as shown in **Figure 6**.



Source : http://academic.pgcc.edu/~kroberts/Lecture/Chapter%2013/shape.html

Figure 6. Various morphologies of viral particles. Panel A. Helical-shaped virus. Panel B. Aggregates of icosahedral shaped viruses. Panels C and D. Viral particles with random, more complex shapes.

Viruses can infect any other living organism (Koonin et al 2006) and require the hosts' cellular machinery to replicate. Viral genomes can be composed of DNA or RNA, be double-stranded or single stranded, and finally, segmented or not segmented. The International Committee on Taxonomy of Viruses, which is the reference organisation for the taxonomy of viruses, identified a list of 3'186 viral species in its annual report of 2014 (http://www.ictvonline.org/virusTaxInfo.asp). These species were classified in 505 genera distributed into 104 Families. 2'016'112 viral nucleotide sequences were available in the Genbank database as of September 2015.

Fungi include both unicellular and multicellular eukaryotic microorganisms. They are sorted in the fungi group mainly due to the fact that, unlike plants, bacteria and protozoa, their cell walls produce chitin**.** There is an estimated 1'500'000 fungal species on Earth but only 300 have been described as pathogenic for humans (Garcia-Solache and Casadevall 2010, Hawksworth 2001). With 5'452'827 available on Genbank, Fungal microorganisms are the second most represented pathogenic class in the Genbank database.

d. Pathogenic types

Pathogenic microorganisms can be either primary pathogens or opportunistic pathogens. Primary pathogens are microorganisms that cause symptoms when they cross the hosts' defensive barriers. A good example of primary pathogens are the three main parasitic species causing schistosomiasis, *Schistosoma mansoni*, *S. japonicum* and *S. haematobium*. The natural reservoirs of schistosomes are various freshwater snail species, namely *Biomphalaria* spp. for *S. mansoni*, *Oncomelania* spp. for *S. japonicum* and *Bulinus* spp. for *S. haematobium*. Human infections only occur through direct contact with water which has been contaminated with cercariae (= infectious life stage of the parasite) released by the host snails (Jordan and Webbe 1969, Sturrock et al 1993).

Opportunistic pathogens are microorganisms which are normally found in the environment or in association with various parts of the body. While they usually don't cause disease in healthy individuals, they are able to cause illness in patients with certain specific conditions such as immunocompromised individuals. Many examples of opportunistic pathogens can be found directly in the human gut microbiome, with the most

important being probably *Escherichia coli* (non-pathogenic strains). In healthy patients, *E. coli* are associated with the degradation of organic matter in the gut and is also closely related to other normal functions of the gastrointestinal tract (Chang et al 2004, Isolauri et al 2001, Kruis et al 2004). In immunocompromised patients, however, certain *E. coli* strains can cross the gastrointestinal barrier and migrate to the bladder or urinary tract and therefore cause various severe symptoms (Kaper et al 2004, Manges et al 2001). Another example of opportunistic pathogen is the bacteria *Acinetobacter baumanii*, often associated with nosocomial infections. While it is an almost ubiquitous bacteria in hospital settings, it usually only colonizes the human body without causing any symptoms, but, might give rise to pulmonary infection, septicaemia and wound infection in weakened patients (Camp and Tatum 2010, Fournier et al 2006).

e. Natural reservoirs of pathogens

A variety of environments can serve as reservoirs for pathogens. This includes both living organisms as well as environmental niches. Recent examples of diseases transmitted to humans from their natural reservoirs include bats, acting as the natural reservoir for various Ebola outbreaks (Baize et al 2014, Leroy et al 2005) or the Middle East respiratory syndrome coronavirus found in dromedary camels (Azhar et al 2014, Raj et al 2014) and infecting humans by direct contact. On the plant side, examples of reservoirs include xylem feeding leafhoppers for the bacteria *Xylella fastidiosa*, an important pathogen with a major economic impact (Blua et al 1999, Hopkins 1989, Mizell et al 2003). Moreover, these same reservoirs can also often harbour multiple pathogens at the same time, hence vectoring multiple human, veterinary or plant pathogens. Bats, for instance, are believed to be the natural reservoir of approximately 20 % of all mammalian-infecting viruses and

is considered as one of the most important reservoir for emerging and re-emerging human diseases (Calisher et al 2006, Daszak et al 2000). Similarly, there is intra-reservoir pathogen diversity in camels, which, in addition to MERS-CoV were also shown to transmit the Camelpox virus to humans by direct contact (Bera et al 2011).

f. Cumulative burden of coinfections

In parasitology, coinfection is the simultaneous infection of a host by several parasites. Data about coinfections in humans is lacking but it is thought to be extremely common (Cox 2001, Pullan and Brooker 2008), sometimes being more prevalent than single infections in specific settings (Petney and Andrews 1998). In virology and bacteriology, the term coinfection applies for cells infected with two or more viral or bacterial species. Several examples involve bacteria and viruses in coinfection events causing serious outcomes on human health. These include infections of patients with both the human immunodeficiency virus and *Mycobacterium tuberculosis*, responsible for acquired immune deficiency syndrome and tuberculosis, respectively (Pawlowski et al 2012). This particular case poses serious public health challenges, due mainly to multidrug-resistant strains of *Mycobacterium tuberculosis* which strive in immunocompromised patients and are now widely spread (Streicher et al 2015, Zignol et al 2012). Another bacterial/viral coinfection synergy example are patients infected with both Influenza virus and pneumonia-causing bacteria. A recent review showed that more than 65’000 deaths per year are attributable to influenza and pneumonia occurring together in the United States (Chertow and Memoli 2013). A final example of coinfection causing aggravated health outcomes is infection with the HIV-Viral hepatitis complex (HBV, HCV, HDV), that is

reported to cause sever liver disease and jeopardise the effectiveness of HIV treatment (Alter 2006, Casey et al 1996, Kiesslich et al 2009).

g. Pathogen genomics and associated challenges

Genomics is, with the advent of next-generation sequencing technologies and future sequencing technologies, one of the scientific areas that produces the highest amounts of data, with an expected exabase of sequence produced in the next decade (Stephens et al 2015). This consequent amount of sequencing data will pertain all types of living organisms, but will be mainly focused on human genomes with several hundred thousands sequenced genomes along with a few millions of sequenced microbes, for which the genome size is, however, smaller than the human genome (Stephens et al 2015). This large amount of information is two-sided, as, on one hand, it will allow researchers to gain new and deeper insights into multiple areas of infectious diseases, including, but not limited to, epidemiology, diagnostics, and pathogenesis of infectious diseases as well as species-species and species-host interactions (Bessen et al 2014, Depledge et al 2014, Feero et al 2011, Rappuoli 2004). On the other hand, however, this amount of data also raises questions surrounding data analysis, data safety and bioinformatics approaches, which are not developing at the same pace as sequencing technologies (Fernald et al 2011, Pop and Salzberg 2008, Stephens et al 2015).

h. Pathogen identification and genetic traits

In addition to increase accuracy for pathogen discovery and diagnostics, the amount of information accompanying the genomics era also enabled the creation of extensive gene databases pertaining different phenotypic characters pathogens. These aspects include

mobile genetic elements such as bacterial phages, plasmids, virulence factors and antimicrobial resistance genes (Chen et al 2005, Leplae et al 2004, McArthur et al 2013, Zhou et al 2007, Zuo et al 2007). The latter is a good example of bioinformatics challenges that need to be addressed before these databases make their way to the clinical setting. In this case, specific challenges exist, mainly due to the diversity of resistance mechanisms adopted by bacterial pathogens. These mechanisms can be either due to acquired plasmids carrying resistance genes, point mutations in the antibiotics targets or modified expression of genes coding for efflux pumps (Mah and O'Toole 2001, Martínez 2008, Stewart and Costerton 2001). Therefore, the related bioinformatics challenges in this specific context are due to; (i) the fact that plasmid-driven resistance is difficult to attribute to one organism as plasmids might be exchanged between bacterial species; (ii) the fact that point mutations need a deep sequencing coverage to rule out sequencing errors and confirm quality of assembled sequences; and (iii) that bioinformatics analyses involving metagenomics or metatranscriptomics approaches need to take quantitative information into account when screening for efflux-based resistance (Schneeberger et al 2015).

2. Challenges in infectious diseases research

The aspects of pathogens mentioned in the previous subchapter are all recurrent challenges where much remains to be researched. Their respective impact on the field of infectious diseases is shown in **Figure 7**. The focus of this thesis is located mainly between diagnostics and treatment as the two main objectives were i) to assess the potential of omics in the area of pathogen diagnostics and ii) to use omics techniques in

the area of patient treatment to providing advanced molecular characterization of the pathogen.

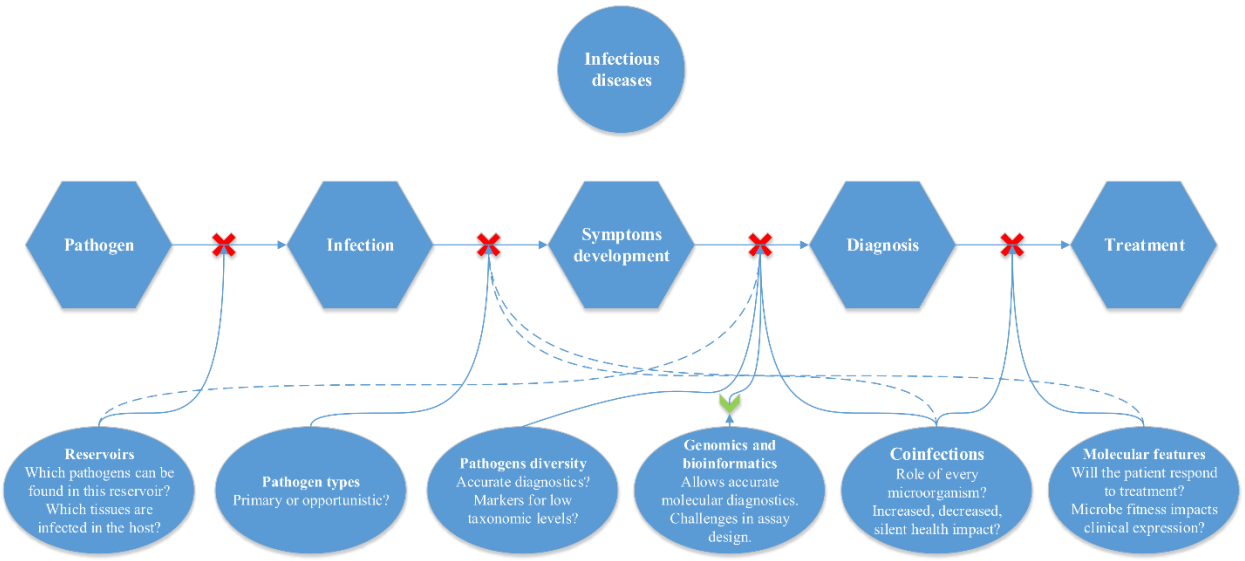


Figure 7. Pathogens traits and associated bottlenecks in infectious diseases research. Different dash types of the connectors indicate different impacts, the solid line indicates a stronger impact than the dashed line.

a. Current diagnostic approaches

Current diagnostic approaches in the field of infectious diseases rely mainly on four strategies, namely; (i) culture-based diagnostic approaches; (ii) microscopy diagnostics; (iii) immunological diagnostics; and (iv) molecular diagnostic approaches. This subchapter summarises the different tools available and their specificities.

b. Culture-based diagnostics

Culture-based diagnostics is mainly used in bacteriology (Fischbach and Dunning 2009, Washington 1996) and virology (Leland and Ginocchio 2007) and to a lesser extent in

parasitology (Visvesvara and Garcia 2002). For bacteria, the diagnostic is based on the use of selective mediums that allow the growth of bacterial species with specific biochemical properties. Bacterial pathogens are identified based on these phenotypes. Diagnosis of viral diseases relies on the isolation of viruses in adequate cell cultures. Parasite diagnosis on culture is more complex, since it may require different environments for each life stage.

Pros	Cons
Standardised protocols	Not available for all pathogens
Accurate identification	Low throughput (one culture = one identification)
	No information about the intra-species genetic diversity
	May require long incubation time for some microbes
	Infectious material, requires specific facilities

Table 1. Pros and cons of culture-based diagnostics.

c. Microscopy

Microscopy is the most common method used both for the detection of microorganisms directly in clinical specimens and for the characterisation of organisms grown on culture media. Microscopy is defined as the use of a microscope to visually enlarge objects too small to be visualised with the naked eye so that their phenotypes can become observable. There are four main classes of microscopes used in diagnostic microbiology, namely; (i) bright-field microscopes used to identify bacteria, fungi, and parasites; (ii) fluorescence microscopes which can be used for any of the five pathogen classes; (iii) dark-field microscopes used for the identification of bacteria; and (iv) electron microscopes mainly used to diagnose parasites and viruses.

Pros	Cons
Optical microscopy is fast and inexpensive	Other microscope types require expensive equipment
	Accurate diagnostics requires an experienced operator
	Complex samples are difficult to analyse
	Accurate identification at low taxonomic level is difficult

Table 2. Pros and cons of microscopy-based diagnostics (Mabey et al 2004).

d. Immunoassays

Immunoassays are protein based assays that allow the detection and/or quantification of an antibody/antigen reaction during an infection event. Antibodies are used as probes to detect a specific antigen and are linked to a reactive molecule, be it a radiolabel, a fluorescent label or a colour-forming enzyme. Immunoassays are available for a wide range of microorganisms for each pathogenic class. They are also often available in the format of rapid-diagnostics tests, making them an excellent tool for point-of-care diagnostics.

Pros	Cons
Fast and relatively inexpensive	Specific to one or a group of closely related microorganisms
Highly specific	Relies on the immune response of the host
Ease-of-use (e. G. RDTs)	Identification at low taxonomic level can be difficult

Table 3. Pros and cons of immunodiagnostics (Jacobson 1998).

e. Molecular-based assays

Molecular diagnostics is based on the amplification of a specific genomic region of a pathogen, also known as diagnostic sequence. Since genetic information is highly specific

to each microbial species, these tests are usually very accurate and have a high discriminative power. They can be used for the diagnostics of all pathogen classes, provided nucleotide sequences are available to select an amplification target. These assays include Polymerase Chain Reaction, real-time PCR, Loop-mediated isothermal amplification, DNA microarrays, Sanger sequencing and a number of other variations of PCR.

Pros	Cons
Highly discriminative and specific	Assay design requires the organism to be sequenced
Identification at any taxonomic level	➔ Not possible if intra-taxon genetic diversity is too high
Low per reaction price	➔ but expensive equipment
Standardised protocols	Quality of the assay depends on input sequences used for the selection of the amplification target
Allows phylogenetic studies	

Table 4. Pros and cons of molecular-based diagnostics. (Mancini et al 2010, Yang and Rothman 2004)

3. Next-generation sequencing and implication in pathogen diagnostics

Current diagnostics approaches, except for the specific case of diagnostics microarrays, present a shared limitation since they all follow the “one assay = one organism” rule. While this is not a problem for studies focusing specifically on e.g. the epidemiology of one microorganism, it becomes problematic to understand system-wide dynamics, e.g. to study all species-host or species-species interactions, since hosts are rarely colonized by a single microorganism. In fact, there are several examples where NGS showed its potential to tackle the different challenges related to the versatile nature of infectious diseases by providing a tool allowing this research area to upgrade from “studies of one”

to “system-wide studies”, or molecular meta-analyses. These new studies include, but are not limited to, complete characterisation of microbial populations, or microbiomes as well as system-wide characterisation of additional molecular features relevant to gain further insights into infectious diseases.

a. Evolution and impact of NGS technologies

Next-generation sequencing started a revolution in early 2000 in the field of genomics and genome-wide studies with the introduction of the 454-pyrosequencing technology (Mardis 2008, Shendure and Ji 2008, Williams et al 2006). The introduction of this technological advance, with the 454 FLX instrument from Roche (Dressman et al 2003, Margulies et al 2005), allowed the multiplication of the output of Sanger sequencing by a factor of 10000, from a thousand base pair to over 100 Mb produced in a single sequencing run (Droege and Hill 2008). As a consequence, sequence repositories, such as Genbank, have increased dramatically in size and management and storage of this massive data amount is currently one of the major challenge (Mohammed et al 2012, Stephens et al 2015), along with the flourishing nebula of non-standardised bioinformatics tools and pipelines that makes it difficult for biologists without informatics knowledge to keep an overview (Fernald et al 2011, Moore et al 2010). The most notable example of benefits NGS has brought to the field of genomics is the significant decrease in the price of sequencing human genomes, which was roughly around 70’000’000 USD in the pre-genomics era, 1’000’000 USD at the beginning of the genomics era and is now roughly around 1’000 USD (Metzker 2010, Shendure and Ji 2008, van Dijk et al 2014), making

the concept of personalised medicine come even closer to reality (Feero and Guttmacher 2014, Ingelman-Sundberg 2015, Shukla et al 2015).

b. NGS technologies in 2015

There are currently four main NGS technologies used on the market, namely, (i) pyrosequencing (454 sequencing); (ii) semiconductor-based sequencing (Ion Torrent); (iii) sequencing-by-synthesis (Illumina); and (iv) first generation single molecule sequencing (Pacific Biosciences). Technical characteristics of the different sequencing platforms are briefly summarised in **Figure 8**.

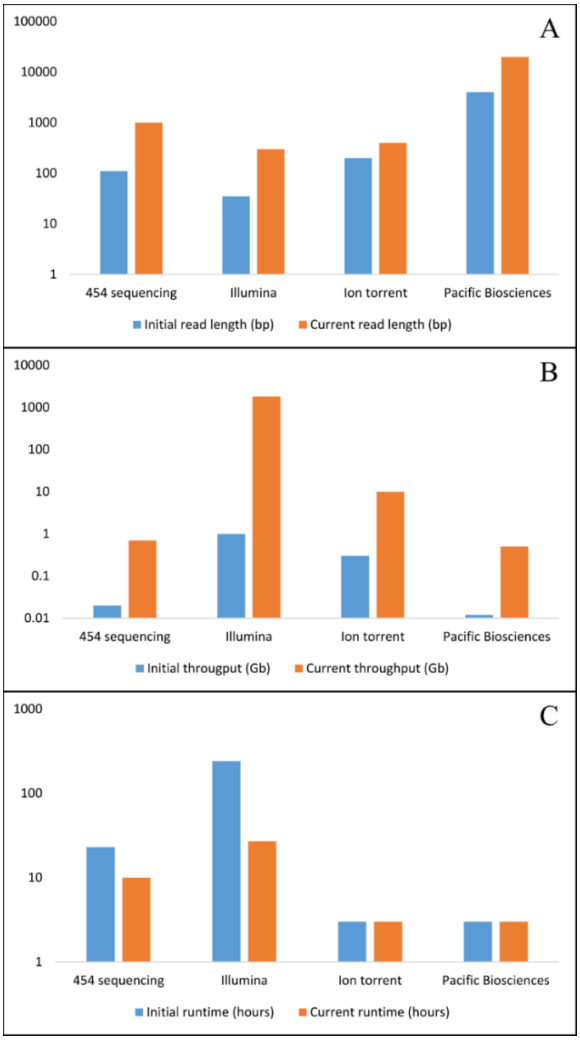


Figure 8. Technical characteristics of next-generation sequencing platforms in 2015. Panel A. Maximum read-length at the early commercial stages of the technologies and current read-length. Panel B. Sequencing output at the commercialisation of the technology and current sequencing output. Panel C. Comparison of runtime with early protocols and current protocols. (Shendure and Ji 2008, van Dijk et al 2014).

Next-generation sequencing has now been applied in a variety of studies (Ekblom and Galindo 2011, Lee et al 2013, McCormack et al 2013). This includes *de novo* sequencing of new microbes (Yin et al 2014), variant calling (Henn et al 2012), advances in transcriptomics (Wang et al 2009) as well as various types of meta-analyses (Handelsman 2004, Shi et al 2009, Tringe et al 2005).

c. NGS meta-analyses: targeted, whole-genome and -transcriptome sequencing

It is important to notice that the term metagenomics, one type of meta-analysis, is often incorrectly used in studies involving targeted sequencing, like 16S rRNA gene sequencing (Cénit et al 2014, Sankar et al 2015) as well as for studies involving whole-genome sequencing (Qin et al 2012). These different types of studies also generate different types of results, one being restricted to a specific class of microorganism, i.e. bacteria when 16S rRNA genes are analysed, and the latter being an unbiased approach in that complete microbiomes are identified, including all viral, bacterial, fungal and parasitical microorganisms (Human Microbiome Project 2012). The various applications involving metagenomics in the frame of this PhD thesis are of the latter type, as it is believed that, once fully developed and matured, this type of approach will allow “true” complete sample characterisation.

4. Overarching goals of the PhD

The overarching goals of this PhD were (i) to assess and review current molecular diagnostics tools; (ii) to develop and optimize approaches that could help improving the

molecular diagnosis of infectious diseases; and (iii) to validate these approaches with different applications.

The specific objectives were:

i) To compare “naïve” molecular diagnostics approaches, including multiplexed assays, microarrays and meta-analyses based on next-generation sequencing (metagenomics and metatranscriptomics).

ii) To develop a workflow allowing the selection of highly conserved and specific molecular markers among highly diverse taxa, which were further used as targets for molecular assays. This included the development of the bioinformatics pipeline as well as the validation on a set of selected viruses.

iii) To develop and conduct a proof-of-concept study showing the potential of meta-analyses in the field of molecular diagnostics. This proof-of-concept study focused on patients with persistent digestive disorders and the potential of metagenomics in the context of the rapidly developing area of personalised medicine.

iv) To apply a metagenomics approach in a larger study including both environmental and human samples with the aim to assess the impact of exposure to wastewater on the gut microbiome of different population groups.

Chapter II. Development and evaluation of a bioinformatics approach for designing molecular assays for viral detection

Published in "*PLOS one*", 2017. (DOI: 10.1371/journal.pone.0178195)

Pierre H. H. Schneeberger[1,2,3,4*], Joël F. Pothier[5], Andreas Bühlmann[6], Brion Duffy[5], Christian Beuret[2], Jürg Utzinger[3,4], Jürg E. Frey[1]

[1]Agroscope, Department of Methods Development and Analytics, Wädenswil, Switzerland. [2]Department of Virology, Spiez Laboratory, Federal Office for Civil Protection, Spiez, Switzerland. [3]Swiss Tropical and Public Health Institute, Basel, Switzerland, [4]University of Basel, Basel, Switzerland. [5]Zurich University of Applied Sciences (ZHAW), Institute of Natural Resource Sciences, Environmental Genomics and Systems Biology Research Group, Wädenswil, Switzerland. [6]Department of Foods of Plant Origin, Agroscope, Institute for Food Sciences IFS, Wädenswil, Switzerland.

* pierre.schneeberger@unibas.ch

Short title: Bioinformatics approach for viral detection

1. Abstract

Background: Viruses belonging to the *Flaviviridae* and *Bunyaviridae* families show considerable genetic diversity. However, this diversity is not necessarily taken into account when developing diagnostic assays, which are often based on the pairwise alignment of a limited number of sequences. Our objective was to develop and evaluate a bioinformatics workflow addressing two recurrent issues of molecular assay design: (i) the high intraspecies genetic diversity in viruses and (ii) the potential for cross-reactivity with close relatives.

Methodology: The workflow developed herein was based on two consecutive BLASTn steps; the first was utilized to select highly conserved regions among the viral taxon of interest, and the second was employed to assess the degree of similarity of these highly-conserved regions to close relatives. Subsequently, the workflow was tested on a set of eight viral species, including various strains from the *Flaviviridae* and *Bunyaviridae* families.

Principal findings: The genetic diversity ranges from as low as 0.45% variable sites over the complete genome of the Japanese encephalitis virus to more than 16% of variable sites on segment L of the Crimean-Congo haemorrhagic fever virus. Our proposed bioinformatics workflow allowed the selection – based on computing scores – of the best target for a diagnostic molecular assay for the eight viral species investigated.

Conclusions/significance: Our bioinformatics workflow allowed rapid selection of highly conserved and specific genomic fragments among the investigated viruses, while considering up to several hundred complete genomic sequences. The pertinence of this workflow will increase in parallel to the number of sequences made publicly available. We

hypothesize that our workflow might be utilized to select diagnostic molecular markers for higher organisms with more complex genomes, provided the sequences are made available.

2. Introduction

The genus *Flavivirus* (RNA virus) includes several species that cause serious human diseases. In *Flavivirus* infections, the first clinical features observed include, but are not limited to, fever, myalgia, headaches, and other nonspecific symptoms (Burke and Monath 2001, Gould and Solomon 2008, Leyssen et al 2000, Solomon 2004). These nonspecific symptoms complicate the identification of the specific causative agent. Importantly, Japanese encephalitis virus (JPEV), West Nile virus (WNV), and St. Louis encephalitis virus (SLEV) are responsible for larger outbreaks affecting both humans and animals (Erlanger et al 2009, Kopp et al 2013, Petersen and Fischer 2012). Other emerging zoonotic *Flaviviruses*, such as the Usutu virus (USUV), might become important threats to human health due to their similarities with other human pathogenic viruses, such as WNV (Nikolay et al 2011, Vazquez et al 2011). While potential vectors are expanding in the northern hemisphere, resulting in sporadic cases of WNV (Mulatti et al 2014, Nash et al 2001) and USUV infections in birds (Steinmetz et al 2011, Weissenböck et al 2002), these infections remain endemic in low- and middle-income countries. New research is needed to develop methods for rapid and accurate identification, and to validate these diagnostic tests before wider application. Additionally, while other zoonotic arboviruses, such as the Rift Valley fever virus (RVFV) and the Crimean-Congo haemorrhagic fever virus (CCHFV) within the *Bunyaviridae* family, cause serious diseases in humans, only a limited number of assays are currently available for their

identification and there is a lack of standardization in the assays used in routine diagnostics laboratories (Anon. , Hujakka et al 2003).

Virus neutralization tests (VNTs) are usually considered the ‘gold’ standard for the diagnosis of infections by these pathogens (Li 2013). VNTs, however, require a cultivation step that must be performed in laboratories with high biosafety measures, which are not widely available in low- or middle-income countries. Immunoassays are broadly used in clinical-diagnostic settings. However, while immunoassays rely on biochemistry to identify the presence or concentration of antibodies or antigens, genomic and phylogenetic information to understand the route of transmission and biology of these viruses is lacking. Various polymerase chain reaction (PCR)-based assays, including real-time PCR, have been used successfully in epidemiologic studies (Burt and Swanepoel 2005, Grobbelaar et al 2011, Pepin et al 2010). Yet, this variety of assays introduces a lack of standardization in the different routine diagnostic laboratories. It is conceivable that taxon-specific molecular assays, even though system-wide diagnostics studies become more and more common (Schneeberger et al 2016), that are relying on genomic information might help clinicians and researchers to obtain more accurate epidemiologic baseline data for neglected viral infections (Espy et al 2006, Mackay et al 2002, Sloan et al 2008). Within the *Bunyaviridae* family, viruses from the *Hantavirus* genus are responsible for several recent outbreaks (Hartline et al 2013, Montgomery et al 2012, Roehr 2012), but reliable molecular assays to trace transmission pathways and to deepen our understanding of viral epidemiology have yet to be developed and more widely implemented.

Genetic diversity among RNA viruses from the *Bunyaviridae* and *Flaviviridae* families is high compared with that of DNA viruses, as has been shown by new data produced by next-generation sequencing technologies (Beerenwinkel et al 2012, Radford et al 2012). While the development of molecular assays is quite straightforward, such approaches are mainly based on the pairwise alignments of sequences, followed by selection of the most conserved region within the aligned sequences. Although alignment algorithms are constantly being improved, computational challenges are still encountered when dealing with large numbers of sequences. Such molecular assays are of low priority for organisms with slow mutation rates because the overall genetic diversity of these organisms remains low and few sequences are sufficient to create an accurate representation. In contrast, in rapidly mutating viruses, the method may become restrictive because of the small number of sequences, which may not necessarily represent the complete genetic diversity within the species. Thus, overall, this alignment approach may give rise to two challenges: (i) the selected region is only conserved among a few genetic variants and not among the complete taxon and (ii) lack of information about the degree of sharing between the selected regions and the sequences of other closely related organisms, potentially causing cross-reactions.

We developed a workflow based on the well-established BLASTn algorithm (Altschul et al 1990) to address the aforementioned challenges. Subsequently, the workflow was tested on a set of viruses from the *Flaviviridae* and *Bunyaviridae* families. Our data may be applicable for rapid selection of highly conserved and taxon-specific regions for any viral family and, perhaps, for other higher organism for which sufficient genomic data are

available. This may further improve various nucleic acid-based molecular tools, such as real-time PCR or loop-mediated isothermal amplification (LAMP).

3. Methods

a. Hardware and software requirements

Version 2.2.28+ (64 bits) of the standalone BLAST algorithm was employed in the workflow. A backbone script written in PERL was utilized to automate the process and to parse and retrieve the intermediate and final result files. The workflow was tested on two versions of PERL (versions 5.16 x64 and 5.10 x32). Of note, the script will work with any other PERL version compatible with the BioPerl package v.1.6.901 (Stajich et al 2002). Version 2.3.4 of the Primer3 package (Untergasser et al 2012) was utilized to select primers for the real-time PCR assays. For each species, a subset of highly conserved fragments (HCFs; n = 2) selected by the workflow was used to design a primer pair for real-time PCR analysis. In order to test different assay configurations, we used the “pick primers tool” from Primer3 with a primer size range set to 18–24-mer primers, and a target amplification product size set between 300 and 400 bp for members of the *Flaviviridae* family. The same “pick primers tool” was used for members of the *Bunyaviridae* family; however, because of the higher genetic variability, the primer size range was adjusted to generate 25–30-mer primers, and the amplification product target size was set between 100 and 400 bp.

The same sets of HCFs selected for real-time PCR assays were used as the amplification target to test LAMP assays. The HCFs for SLEV and USUV were submitted to the online LAMP primer design tool Primer Explorer V4 (Fujitsu, Japan; see:

https://primerexplorer.jp/). A set of six LAMP primers (F3, B3, FIP, BIP, LoopF, and LoopB) was automatically selected for each of the two species.

To demonstrate the flexibility of this workflow, two different computer configurations were used. Configuration “1” was a conventional notebook, running Windows 7 (x64) with 8 Gigabyte (Gb) of RAM and an i7 quad core CPU to run up to eight BLASTn instances in parallel. Configuration “2” was a more powerful workstation running Windows 7 (x64), with 32 Gb of RAM and an i7 hexacore CPU able to run up to 12 BLASTn instances in parallel.

b. Input Data Used for the Workflow

A file containing all publicly available complete genome sequences was downloaded on January 17, 2013 for each tested virus species from GenBank (Benson et al 2013). The number of sequences available on this date ranged from only six sequences for USUV up to 608 sequences for WNV (Table S1).

c. Phylogenetic Analyses

Phylogenetic analysis was performed using MEGA v.6.0 software (Tamura et al 2013). The ClustalW pairwise alignment algorithm (Larkin et al 2007) was used with default parameters, and the trees were generated from the sequence alignments using the neighbour-joining approach (Saitou and Nei 1987) with 700 bootstrap replications.

d. Viral Samples

Eight viral species from the *Flaviviridae* and the *Bunyaviridae* families were used to test the results of the workflow. Two WNV strains (i.e., NY99 and Dakar) were included in this

study. For the remaining seven viral species, we included a single species sample and did not test various strains. The viral samples were obtained from various European collections and cultivated using various methods, as reported in **Table 1**. Upon receipt, each virus was propagated in appropriate cell cultures within a biosafety level 3 (BSL-3) facility at Spiez Laboratory (Spiez, Switzerland) and virus titres were measured using the respective validated rt-qPCR protocols. An aliquot of each sample was stored at -80°C.

Taxonomy (family, genus, species)			Abbreviation	Subtype	Cell type	Origin[a]
Flaviviridae						
	Flavivirus					
		St. Louis encephalitis virus	SLEV	Type 1	Vero E6	NCPV
		Usutu virus	USUV	Bologna	Vero E6	UNIBO
		Tick-borne encephalitis virus	TBEV	Hanzalova	Porcine kidney	IP ASCR
		Japanese encephalitis virus	JPEV	Nakayama	Vero E6	NCPV
		West Nile virus	WNV	NY99	Vero E6	NCPV
		West Nile virus	WNV	Dakar	Vero E6	NCPV
Bunyaviridae						
	Nairovirus					
		Crimean-Congo hemorrhagic fever virus	CCHFV	N.A.[b]	BNI	BNI
	Phlebovirus					
		Rift Valley fever virus	RVFV	H13/96	Vero E6	NCPV
	Hantavirus					
		Seoul virus	SEOV	R22	Vero E6	NCPV

Table 1. Virus species used for the validation of the diagnostic assays developed with the workflow designed in this study. [a]NCPV, National Collection of Pathogenic Viruses (Porton Down, United Kingdom). BNI, Bernhard-Nocht-Institute for Tropical Medicine (Hamburg, Germany). IP ASCR, Institute of Parasitology - Academy of Sciences of the Czech Republic (Prague, Czech Republic). UNIBO, University of Bologna (Bologna, Italy). [b]N.A., not available.

The viral titres were measured as follow: SLEV = 8.1×10^9 PFU/ml, USUV = 1.35×10^9 PFU/ml, TBEV = 1.66×10^9 PFU/ml, JPEV = 5.34×10^7 PFU/ml, WNV NY99 = 1.5×10^{10} PFU/ml, WNV Dakar = 1.61×10^{10} PFU/ml, CCHFV = 9.6×10^8 PFU/ml, RVFV = 9.92×10^7 PFU/ml, and SEOV = 4.66×10^7 PFU/ml.

e. Nucleic acid isolation

Prior to extraction, each cell culture supernatant was concentrated from 1 ml to 100 µl using 10-kDa AMICON Ultra centrifugal units (Merck Millipore; Billerica, MA, United States of America) at 4,000 × *g* for 4 min. After concentration, RNA was isolated and extracted on an EZ1 Advanced XL platform (Qiagen; Hilden, Germany). The EZ1 Virus Mini Kit v2.0 (Qiagen) was used, adhering to the manufacturer's protocol.

f. Real-time PCR and LAMP assays

Real-time PCR assays were performed on a ViiA 7 real-time system (Applied Biosystems; Carlsbad, CA, United States of America) using the Power SYBR RNA-to-Ct One-Step kit (Thermofisher Scientific; Bremen, Germany). Reverse transcription was performed at 48°C for 30 min, and samples were subjected to 40 cycles of PCR amplification (95°C for 15 s and 55°C for 1 min) for flaviviruses. The same conditions were used for the members of the *Bunyaviridae* family, except that 52°C was used for the second step of the cycles, instead of 55°C. Amplification was performed in a reaction volume of 50 µl, and amplification products were detected using SYBR Green staining. Due to higher concentrations for the *Flaviviridae*, 3 µl from the initial solution was used as a template instead of 5 µl for CCHFV, RVFV, and SEOV. A final concentration of 0.2 µM was used for both the forward and reverse primers for each reaction. The melting curves were done with temperatures ranging from 55°C to 95°C with a ramp rate of 0.05°C/s. LAMP assays were performed on a 7500 Fast Real-Time PCR System (Applied Biosystems; Carlsbad, CA, United States of America). Isothermal MMX (OptiGene; Horsham, United Kingdom) was used at a 1× concentration in a 12-µl reaction volume. Primers were used at the

following concentrations: F3 and B3, 0.2 µM; FIP and BIP, 2 µM; and loopF and loopB, 1 µM.

4. Results

a. Workflow Concept

The presented approach consisted of two consecutive BLASTn steps to assess the degree of conservation of a sequence among a taxon of interest and to test for its specificity towards closely related organisms, as detailed in **Figure 1**.

Fig 1. Bioinformatics analysis workflow.

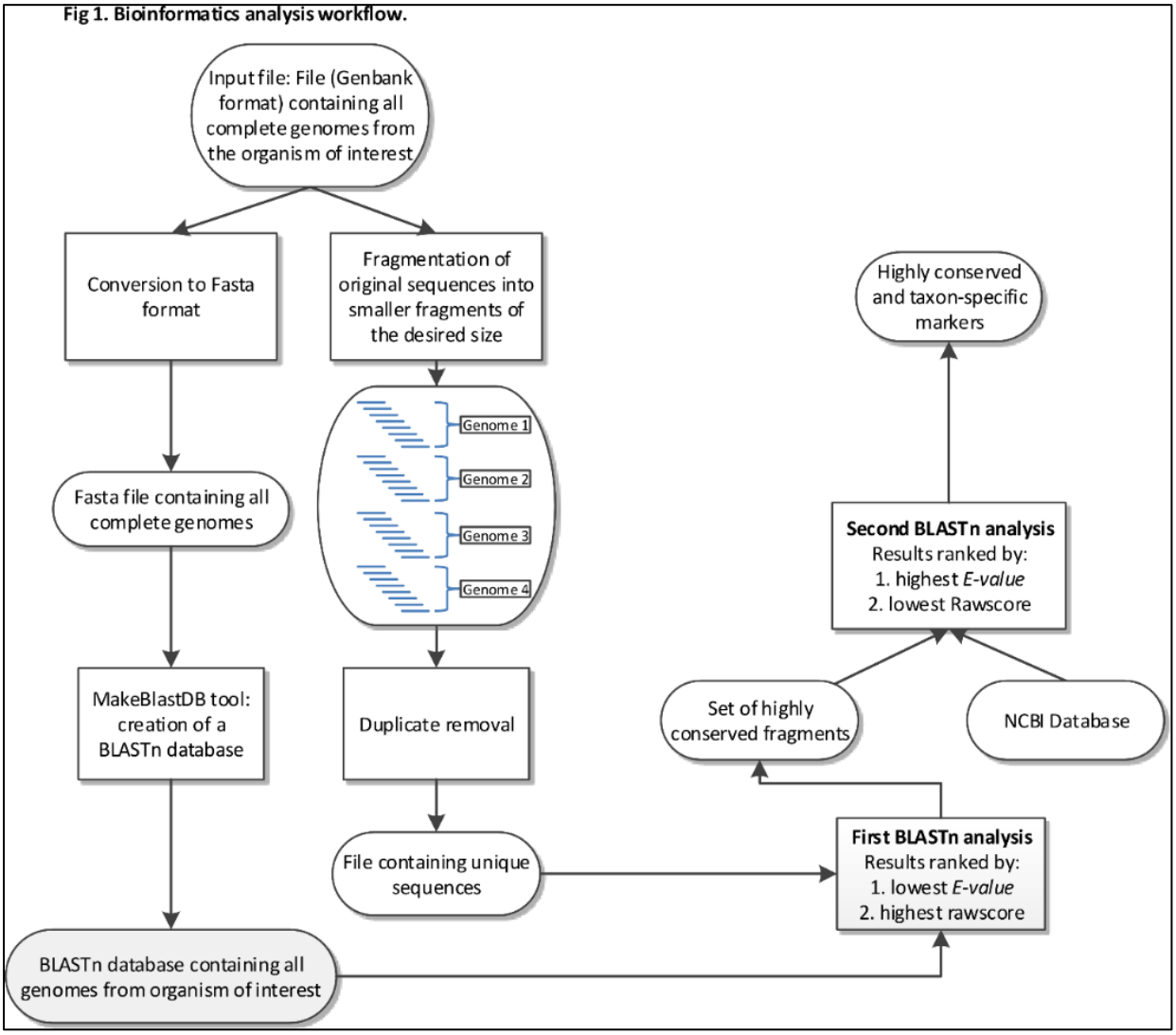


Figure 1. Bioinformatics analysis workflow. Input sequences were processed through a "dual-BLASTn" pipeline in order to select for the most conserved and at the same time specific molecular markers.

Pre-processing of the whole genomic sequences used as input was carried out in two steps. Genomic samples were first fragmented to 400 bp. Because consecutive fragments shared an overlap of 390 bp, they allowed accurate representation of the various genomic regions for the next processing steps. Two additional filtering steps were used to remove sequences showing suboptimal thermodynamic parameters from this pool of organism-specific fragments (OSFs). The first filter selected only fragments with a GC content of 30–70%, and the second filter checked the remaining fragments for homopolymers or repeated regions, which are generally considered inappropriate targets for molecular assays. In parallel, genomic sequences in GenBank format were converted to Fasta format and further converted into an organism-specific database (OSD) using the appropriate tool provided within the NCBI software suite. Subsequently, the first BLASTn step was carried out to select the HCFs among the taxon of interest. In order to perform this action, OSFs were compared to the OSD. The scores resulting from this analysis, including the total amount of hits in the OSD, *E*-values and bitscores, were retrieved in order to assess the degree of conservation of each OSF in the taxon of interest. Moreover, OSFs were ranked by decreasing number of hits, decreasing sum of bitscores, and increasing sum of *E*-values. A subset (n = 100) of the fragments with the best scores was selected for further analysis.

The second part of this workflow aimed to assess the specificity of the subset of HCFs toward the organism of interest, thus providing information on potential cross-reactions

with close relatives. This step consisted of an additional BLASTn step against the NCBI's nt database. In contrast to the ranking system from the previous step, HCFs were ranked by increasing number of hits, increasing sums of bitscores, and decreasing *E*-values, thus enabling ranking to be carried out in accordance with the complete database. Hence, this step allowed us to assess the specificity of each of HCF and served as an assessment of the potential for cross-reactions when using the selected HCFs as targets for molecular assays.

b. Genetic Diversity among the Tested Viruses

The consensus sequences from 10 and 60 segment L complete sequences from the CCHFV were generated in order to assess whether using different numbers of sequences could influence the selection of a target for identification assays. For the same reason, two consensus sequences from 10 and 153 complete JPEV genomes were also generated. The results of these alignments are reported in **Table 2**. The consensus generated from 60 CCHFV sequences had 871 additional ambiguities when compared with the consensus generated from 10 CCHFV sequences. This represents approximately 16% of the overall length of the consensus (5,372 bp). On the other hand, the consensus generated from 153 JPEV sequences had only 48 additional ambiguities when compared with the consensus generated with 10 JPEV sequences, suggesting that only 0.44% of the genome (10,980 bp) represented variable sites.

	Consensus CCHFV	**Consensus JPEV**	**Consensus CCHFV**	**Consensus JPEV**	**Variation CCHFV**	**Variation JPEV**
Sequences	10	10	60	153	N.A.[a]	N.A.
Length (bp)	5,370	10,979	5,372	10,980	2	1
GC (%):	38.18	46.33	30.49	46.45	-7.68	-0.11

A (%):	28.29	24.98	23.59	25.17	-4.70	-0.19
C (%):	18.90	20.13	14.99	20.15	-3.92	-0.02
G (%):	19.27	26.20	15.51	26.30	-3.77	-0.10
T (%):	22.09	17.68	17.09	17.81	-5.00	-0.14
Y (%):	5.51	5.16	11.58	5.11	6.07	0.05
W (%):	0.73	0.44	1.73	0.40	1.00	0.04
V (%):	0.02	0.00	0.24	0.00	0.22	0.00
S (%):	0.04	0.30	0.50	0.26	0.47	0.05
R (%):	4.41	3.96	10.67	3.73	6.25	0.23
N (%):	0.02	0.15	0.58	0.16	0.56	-0.02
M (%):	0.34	0.57	1.62	0.53	1.28	0.05
K (%):	0.24	0.43	0.60	0.36	0.35	0.06
H (%):	0.04	0.00	0.73	0.01	0.69	-0.01
D (%):	0.06	0.00	0.35	0.00	0.30	0.00
B (%):	0.06	0.00	0.24	0.00	0.19	0.00

Table 2. Ambiguity-based comparison of consensus sequences generated using different amounts of Crimean-Congo hemorrhagic fever virus (CCHFV) or Japanese encephalitis virus (JPEV) genomes. [a]N.A., not applicable.

c. Workflow Output

While using configuration 1, it was not possible to align all 608 complete WNV genome sequences with the ClustalW algorithm or the MUSCLE algorithm (Edgar 2004). Using our workflow allowed us to select candidate molecular markers from different numbers of complete genome sequences, from as few as six sequences for USUV to as many as 608 sequences for WNV. Selected molecular markers were used to generate real-time PCR primer sets for the detection of viruses from both the *Bunyaviridae* and *Flaviviridae* families (**Table 3**). Because of the lack of published LAMP assays and to demonstrate that the molecular markers selected using this workflow were multipurpose, we used the HCFs for USUV and SLEV to design LAMP primer sets (**Table 4)**.

Species	Target	Forward primer (5'–3')	Reverse primer (5'–3')	Sequence number	Size (bp)
JPEV	NSP 5	GGTACTACTGGGGCGAATGG	CCAAAAGGGGTGGTGTCAGT	153	342
SLEV	PreMP	ACAAGACTGACGCTCAAAGC	GGATTGCGCAAAACCCAGTT	8	352
TBEV	NSP 5	ACAGCTAAACTTGCCTGGCT	ACGGTTTTTCCACTGCTCCA	42	348
USUV	NSP 5	TCATGGAGCGCTTGGAAGTT	CAGGTCCGATATGGGTGGTC	6	343
WNV	NSP 1	ACCAGAACTCGCCAACAACA	TCTCAAGGATTCCATCGCCC	608	341
CCHFV	Seg.[a] L	GCATCTCTGAAGTAACTGAAACAACA	GTTGAGATAGCACCGAGTTTCTTTAG	41	154
	Seg. M	AGAAACAAGCTTATCAATTGAGGCAC	TGTCCTTTCTTCCAGCTTCATAATTG	60	175
	Seg. S	GATGAGATGAACAAGTGGTTTGAAGA	GTAGATGGAATCCTTTTGTGCATCAT	65	159
SEOV	Seg. L	GTCTCACTTAGTACGAGTAAGGTTGA	AATTTTTGTCAGACATGCCTATACCG	7	178
	Seg. M	CCTTGCAACAATTGATTCTTTTCAAT	ACAAGGATTCTCAGCCAAATTTTCAA	18	160
	Seg. S	GAAGAAATCCAGAGAGAAATCAGTGC	ATTTTTGATTGTATTGAAGCTGCGAC	19	161
RVFV	Seg. L	ATGATGAATGACGGGTTTGATCATTT	AACCTCATACTTAGCGAGTTTAGTCA	86	150
	Seg. M	GGCCCTTAGAGTTTTTAACTGTATCG	GGGCTCTCAATGAAAGAAAAGCTATT	91	192
	Seg. S	AACAATCATTTTCTTGGCATCCTTCT	ATAATGGACAACTATCAAGAGCTTGC	141	180

Table 3. List of selected targets and real-time PCR primer pairs designed for different viral species employed in this study. [a]Seg., segment; WNV, West Nile virus; SLEV, St. Louis encephalitis virus; JPEV, Japanese encephalitis virus; USUV, Usutu virus; TBEV, Tick-borne encephalitis virus; SEOV, Seoul virus; CCHFV, Crimean-Congo hemorrhagic fever virus; RVFV, Rift Valley fever virus.

Species	Primer[a]	Primer sequence (5'–3')	Input sequences
SLEV	F3	GAGCACTTGATGTGGGAG	8
	B3	CAATGATTGCCGAATCGC	
	FIP	CTTCCATCCGTAATCCAACTCATCCTGACTTGTCAGTTGTAGTGC	
	BIP	AACACATTTGTTGTTGATGGACCCGAGTGAACACCATGCCAA	
	LoopF	CCAGCTTCTTCAGGCGTC	
	LoopB	CAAGGAGTGTCCAACAGCA	
USUV	F3	GCTGCCAATGAATACGGA	6
	B3	TAGTGGAGGGTAGCCAGA	
	FIP	GTGAGAACCACTGTGCTCCCTACCCTCCATGAACGCTT	
	BIP	TCAGAATACATCACAACATCTCTGGCGTAGGTTGAACAAAGACCCA	
	LoopF	GGTCGCAAATCCAATGCC	
	LoopB	TTCAATAAGCGCTCAGGC	

Table 4. List of LAMP primer sets designed for Usutu virus (USUV) and St. Louis encephalitis virus (SLEV). [a]F3 and B3, forward outer and reverse outer primers for LAMP,

respectively; FIP and BIP, inner LAMP primers; LoopF and LoopR, forward and reverse loop primers.

The selected primer pairs were tested against a panel of virus species, including two WNV (NY99 and Dakar) strains, as shown in **Figure 2**. CCHFV was amplified with an average between the different genomics segments of 21.9 cycles, RVFV with an average of 23 cycles, and SEOV a C_t value average of 27.8. SLEV, WNV NY99, USUV, and WNV Dakar reached the threshold between 23 and 26 cycles (23.8, 24.1, 25.3, and 25.4, respectively). TBEV and JPEV were amplified within 27.8 and 28.1 cycles, respectively. The efficiency of the reactions was measured between 82% (RVFV Segment M) at the lowest and 141% (JPEV) at the highest. The efficiency of 11 of the other 13 reactions was comprised between 90% and 110% except for TBEV (115%) and CCHFV Segment S (86%).

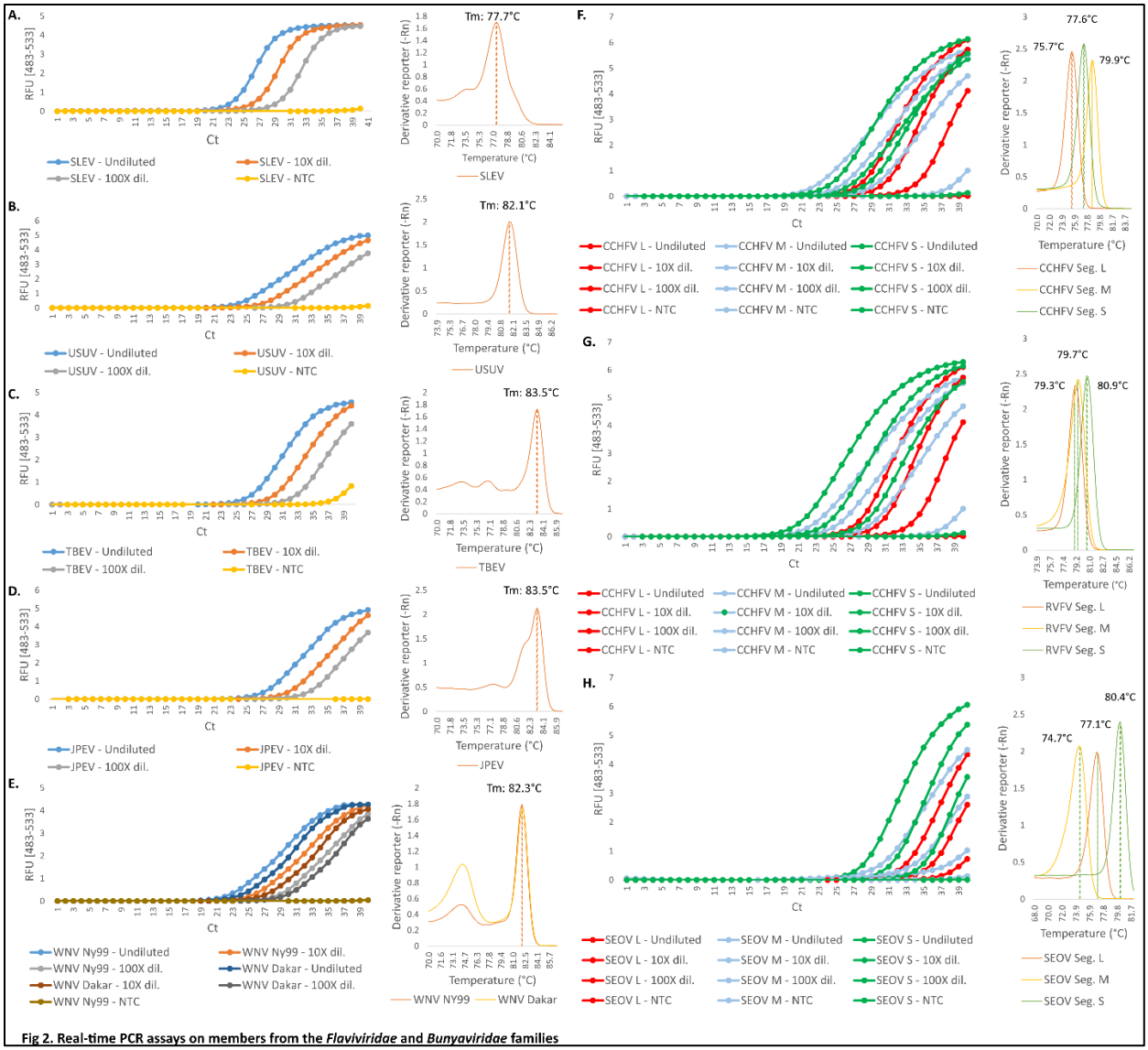


Figure 2. Real-time PCR assays of members from the *Flaviviridae* and *Bunyaviridae* families. Amplification and melting curves for five different flaviviruses species are shown. Each sample was tested undiluted, with a 10-fold dilution and with a 100-fold dilution. (A) St. Louis encephalitis virus (SLEV). (B) Usutu virus (USUV). (C) Tick-borne encephalitis virus (TBEV). (D) Japanese encephalitis virus (JPEV). (E) West Nile virus (WNV; 2 strains, NY99 and Dakar). The right half of the panel shows the amplification and melting curves of the different genomic segments of the members from the *Bunyaviridae* family

tested in this study. (F) Crimean-Congo hemorrhagic fever virus (CCHFV). (G) Rift Valley fever virus (RVFV). (H) Seoul virus (SEOV). NTC, no template control; RFU, relative fluorescence units; C_t, cycle threshold; Dil., dilution; Seg., Segment.

A phylogenetic tree of the *Flaviviridae* family was generated, as shown in **Figure 3**, in order to test for cross-reactivity between the closest relatives, namely JPEV, USUV, and WNV. As previously shown, fragments of all three species were amplified using the corresponding primer sets at 28.1 cycles for JPEV, 25.3 cycles for USUV, and 24.1 cycles for WNV (NY99).

There was no cross-amplification when mixing the JPEV template with the primer pairs selected for USUV and WNV. Similarly, for WNV, amplification occurred only with the corresponding WNV primers and not with the USUV or JPEV primer pairs. However, while there was no amplification with the USUV template and WNV primers, there was amplification when using JPEV primers around cycle 29.

In order to make sure that this cross-reaction does not involve the selection of the target regions but rather the selection of the primer pairs designed to amplify this region, we sequenced the amplicons from the three relevant reactions, namely (i) the JPEV template amplified with JPEV primers; (ii) the USUV template amplified with JPEV primers and (iii) the USUV template amplified with USUV primers (S1 Supporting Information). The obtained sequences were compared to the NCBI database and the USUV template amplified with the USUV primers showed 61.8% identity with JPEV genomic sequences and 70.3% identity when using the primer pair selected for JPEV.

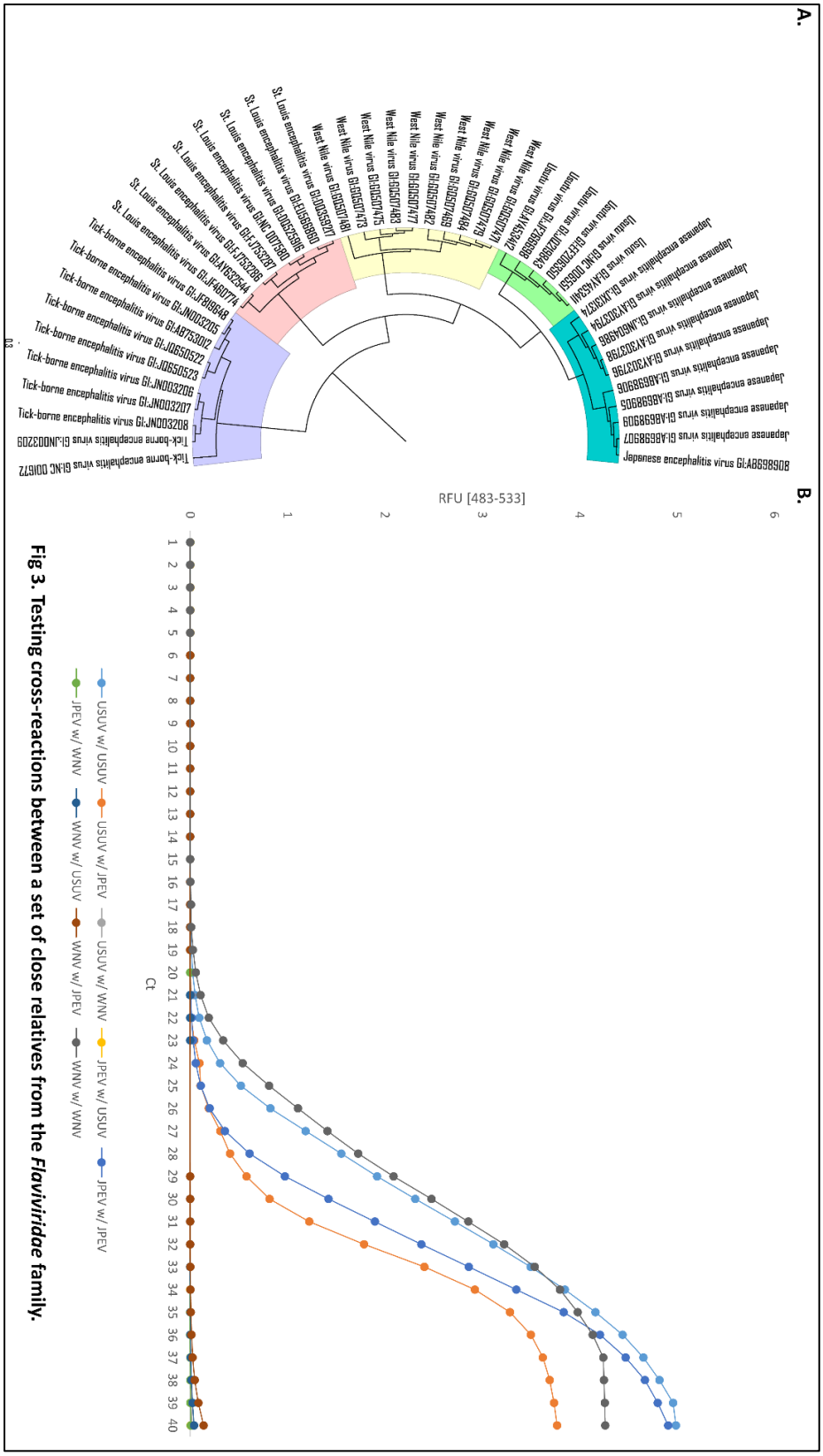


Fig 3. Testing cross-reactions between a set of close relatives from the *Flaviviridae* family.

Figure 3. Testing cross-reactions between a set of close relatives from the *Flaviviridae* family. West Nile virus (WNV), Japanese encephalitis virus (JPEV), and Usutu virus (USUV) were tested. (A) Phylogenetic analysis of a subset of 6–10 sequences from members of the *Flaviviridae* family. (B) Real-time amplification of viruses with master mixes containing different primer pairs. RFU, relative fluorescence units; C_t, cycle threshold.

Both USUV and SLEV were successfully amplified with corresponding LAMP assay primers. Amplification occurred after 46 min for SLEV RNA, including the reverse transcription step. The LAMP primer set selected for USUV successfully amplified the template within 40 min, also including the reverse transcription step (**Figure 4**). The included controls excluded the formation of primer dimers, which is likely to happen due to the nested nature of LAMP assays.

Fig 4. Loop-mediated isothermal amplification of Usutu virus (USUV) and St. Louis encephalitis virus (SLEV).

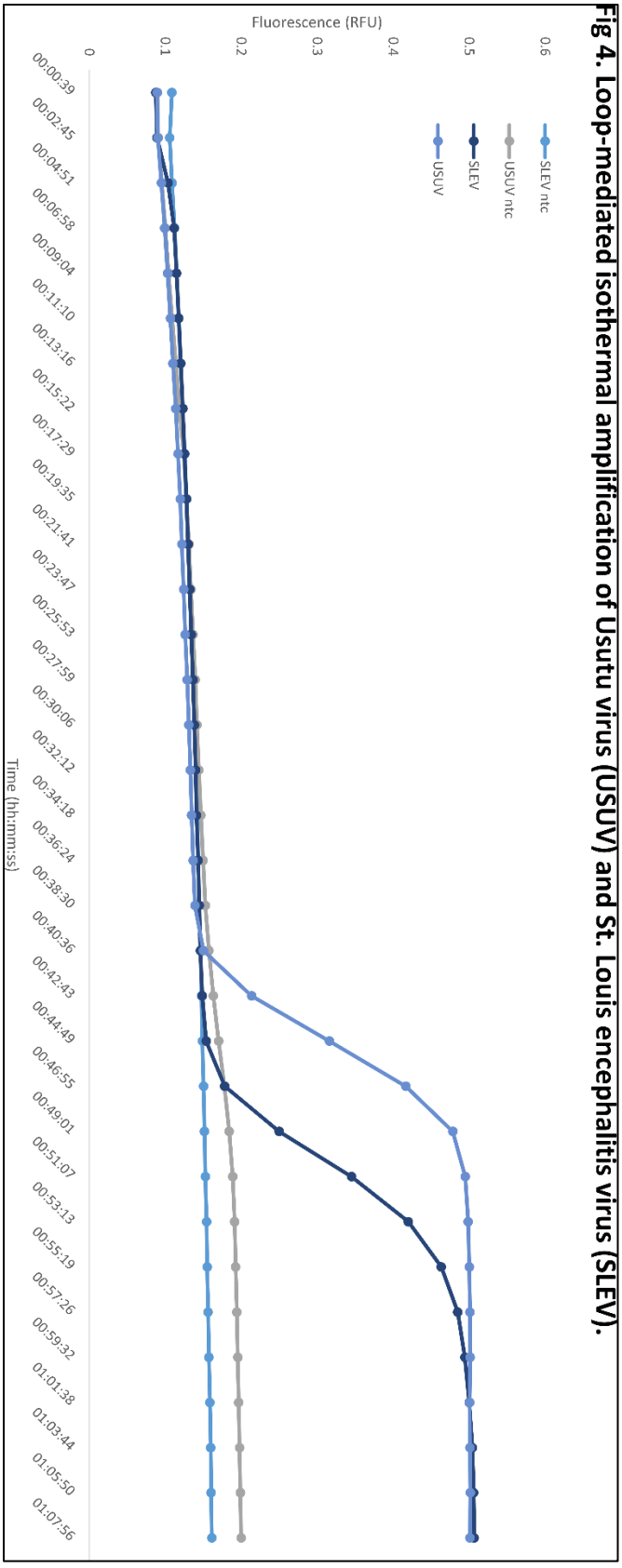


Figure 4. Loop-mediated isothermal amplification of Usutu virus (USUV) and St. Louis encephalitis virus (SLEV). NTC, no template control; RFU, relative fluorescence units; C_t, cycle threshold.

5. Discussion

We developed and evaluated a bioinformatics workflow that readily addresses the high intra-species genetic diversity of viruses and takes into consideration the potential for cross-reactivity between close relatives. These are two key issues that complicate the design of molecular assays (Avise 2012, Espy et al 2006). Our workflow allowed for rapid selection of highly conserved and specific genomic fragments among the investigated viruses, while considering up to several hundred complete genomic sequences.

With the advent of next-generation sequencing, an increasing number of sequences have been, and continue to be, made publicly available (Montgomery et al 2012, Stephens et al 2015). Although this has greatly improved our knowledge of the dynamics of viral populations, the massive amount of data available also renders bioinformatics analysis more complex. In the case of CCHFV, for example, the difference in the consensus sequences between analyses utilizing 10 and 60 genomic sequences was 17.39%, which is a challenge for selecting an appropriate target for a molecular assay. For JPEV, the amount of variable positions was much lower, only representing 0.45% of the complete genome; nonetheless, 50 additional ambiguities were observed throughout the whole consensus. Yet, even such a small difference might still negatively influence the performance of a molecular assay by affecting the thermodynamic parameters of the reaction, particularly the primer annealing step.

Aligning a few genomic sequences is usually straightforward with widely available bioinformatics tools (Ferreira et al 2014, Wen et al 2014). In the case of organisms that have not been as thoroughly sequenced, alignment may not be an issue at all because all available variants may simply be included in the alignment; thus, the overall genetic diversity is considered. In the case of extensively sequenced organisms, however, the issue of “masked” diversity might rise, since only a subset of all the available sequences will be selected for the alignment and finally only a subset of the genetic diversity is taken into account for the design of the molecular assay. By using reproducible computing scores, including bitscores, *E*-values, and the number of “hits” in a database, the workflow also removed the potential bias that could be introduced by manual selection of an adequately amplified region by the user. This workflow allowed us to select highly specific molecular markers in less than an hour for all tested viruses using the more powerful configuration 2. In order to assess the impact of the hardware, we ran the workflow with a single species on both configurations. While the task could be successfully completed on both computer platforms, we noted a drop in the time requirement of approximately 30% from configuration 1 to 2. This drop-in performance was thought to be due to the well-optimized parallelization capacity of the BLASTn algorithm. Therefore, we expected that the overall runtime could be reduced by increasing the number of CPU cores and providing sufficient RAM. In future studies, we will examine the importance of this feature in terms of increased sequencing capacity and the increased resulting genomic data generated every year (Stephens et al 2015). The performance of this workflow will also allow rerunning the analyses when new sequences for a given species of interest become available. This would facilitate identification of shifts in the viral population and could

reveal whether previously selected molecular markers are still valid (*i.e.*, to keep the molecular assay up-to-date and to have it further refined as new data become available). In specific cases, if enough sequences are available, this workflow could also be utilized to generate strain-specific molecular markers. Having strain-specific assays, particularly in the case of neglected tropical diseases, could be a great asset when tracking/investigating transmission events and risk factors, in resource-constrained settings (Fankhauser et al 2002, Van Belkum et al 2001). This workflow also has the advantage of manual design, and hence, it can be entirely customized to the needs of the user. In fact, the output from the workflow only depended on the input sequences, and the user should be able to select, for example, only geographically related strains to design a "geographically specific" assay in order to quickly demonstrate whether outbreaks are caused by a new or re-emerging pathogen (Knowles and Samuel 2003).

All molecular markers that were selected with the workflow could be used as inputs for primer design. Real-time PCR assays were all performed successfully, from the single amplification target selected for the flaviviruses to the three regions selected for each genomic fragment of the members from the *Bunyaviridae* family. Similarly, the same markers selected for USUV and SLEV were successfully used to design LAMP primer sets, and the corresponding LAMP assays performed well. These assays confirmed that the first BLASTn step of this workflow functioned well for selecting highly conserved regions among a pool of species-specific fragments.

The results generated within this study offer a preliminary overview of the assays sensitivity and specificity. However, additional experiments would be required to optimize these assays, especially concerning the efficiency of reaction. In general, the melting

curves show a high specificity, except for WNV for which some primer-dimers seem to be forming. Regarding the suboptimal efficiencies, one lead to optimize could be to remove either inhibitors (especially in the case of JPEV and TBEV, which show an increased reaction efficiency), test various primer concentrations as well as a range of more adapted, reaction-specific, PCR conditions.

In order to further improve this workflow, we added a second BLASTn step to assess the degree of sharing of highly conserved species-specific fragments in a general database also containing genomic data from close relatives. The tested cross-reactions showed that the primers selected for WNV and USUV were specific for those species, whereas the JPEV primers cross-reacted with the USUV template, but not with the WNV template. In order to determine whether this cross-reaction occurred because of the primers or poor selection of the molecular markers, we used Sanger sequencing to sequence the amplicons from the two USUV reactions (both with USUV and JPEV primers) and the JPEV reaction (with the JPEV primers). Sequencing revealed that the amplified regions (*i.e.*, the selected molecular markers) were highly specific to their corresponding species. An online BLASTn of the JPEV primers against USUV sequences showed that the forward primer had nine nucleotides matching the USUV virus at the 3′ end and 19 common nucleotides on the reverse primer (only one mismatch, data not shown). This issue highlights two additional controls that should be performed using this workflow after selecting the target regions, namely (i) an additional online BLASTn control of the primer selected by the various software programs, be it for real-time PCR or LAMP assays, and (ii) since cross-reactions are difficult to predict, the designed assay should be tested with a gradient PCR first to ensure that the thermodynamic parameters of the

reaction are optimal. However, sequencing of the amplification product is still considered the ‘gold’ standard for validating the molecular assay and ensuring high specificity of the assay.

In conclusion, the workflow presented here for viral detection provides a promising approach as it addresses the recurrent issue of bioinformatics analysis of large amounts of sequencing data, which is expected to be an even greater challenge as publicly available data are rapidly increasing. This workflow removes user-introduced bias by being solely based on well-established computing scores (bitscore, *E*-value, and number of hits). Hence, our workflow addresses two issues encountered in the manual design of a molecular assay, as it takes into account the complete genetic diversity of an organism, and provides timely information on potential cross-reactions. We speculate that our workflow is applicable to a variety of DNA-based assays, and hence, it should theoretically work for higher organisms, such as bacteria or parasites, facilitating the selection of future diagnostic markers.

6. Supporting Information

S1 Table. Accession numbers.

S1 Supporting Information. Sequenced amplicons.

7. Acknowledgements

We are thankful to Dr. Oliver Engler and Ms. Jasmine Portman from Laboratory Spiez for preparing and providing us with the virus samples. We would also like to thank Mr. Markus Oggenfuss and Ms. Beatrice Frey at Agroscope for support during the laboratory work.

8. Author Contributions

Conceived and designed the experiments: PHHS. Performed the experiments: PHHS, JFP, and AB. Analysed the data: PHHS, JFP, AB, CB, and JEF. Contributed reagents/materials/analysis tools: JFP, CB, and JEF. Contributed significantly to the manuscript: PHHS, JFP, AB, BD, CB, JU, and JEF.

9. References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). Basic local alignment search tool. *J Mol Biol* **215:** 403-410.

Anon. DISCONTOOLS project. *Available:* http://wwwdiscontoolseu/home/index.

Avise JC (2012). *Molecular markers, natural history and evolution*. Springer Science & Business Media.

Beerenwinkel N, Gunthard HF, Roth V, Metzner KJ (2012). Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. *Frontiers in microbiology* **3:** 329.

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J *et al* (2013). GenBank. *Nucleic acids research* **41:** D36-42.

Burke D, Monath T (2001). Flaviviruses. *Fields virology* **1:** 1043-1125.

Burt F, Swanepoel R (2005). Molecular epidemiology of African and Asian Crimean-Congo haemorrhagic fever isolates. *Epidemiology and infection* **133:** 659-666.

Edgar RC (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* **32:** 1792-1797.

Erlanger TE, Weiss S, Keiser J, Utzinger J, Wiedenmayer K (2009). Past, present, and future of Japanese encephalitis. *Emerging infectious diseases* **15:** 1.

Espy M, Uhl J, Sloan L, Buckwalter S, Jones M, Vetter E *et al* (2006). Real-time PCR in clinical microbiology: applications for routine laboratory testing. *Clinical microbiology reviews* **19:** 165-256.

Fankhauser RL, Monroe SS, Noel JS, Humphrey CD, Bresee JS, Parashar UD *et al* (2002). Epidemiologic and molecular trends of “Norwalk-like viruses” associated with outbreaks of gastroenteritis in the United States. *Journal of Infectious Diseases* **186:** 1-7.

Ferreira AS, Costa P, Rocha T, Amaro A, Vieira ML, Ahmed A *et al* (2014). Direct Detection and Differentiation of Pathogenic Leptospira Species Using a Multi-Gene Targeted Real Time PCR Approach.

Gould EA, Solomon T (2008). Pathogenic flaviviruses. *The Lancet* **371:** 500-509.

Grobbelaar AA, Weyer J, Leman PA, Kemp A, Paweska JT, Swanepoel R (2011). Molecular epidemiology of Rift Valley fever virus. *Emerging infectious diseases* **17:** 2270-2276.

Hartline J, Mierek C, Knutson T, Kang C (2013). Hantavirus infection in North America: a clinical review. *The American journal of emergency medicine* **31:** 978-982.

Hujakka H, Koistinen V, Kuronen I, Eerikäinen P, Parviainen M, Lundkvist Å *et al* (2003). Diagnostic rapid tests for acute hantavirus infections: specific tests for Hantaan, Dobrava

and Puumala viruses versus a hantavirus combination test. *Journal of virological methods* **108:** 117-122.

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

Kopp A, Gillespie TR, Hobelsberger D, Estrada A, Harper JM, Miller RA *et al* (2013). Provenance and geographic spread of St. Louis encephalitis virus. *MBio* **4:** e00322-00313.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H *et al* (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* **23:** 2947-2948.

Leyssen P, De Clercq E, Neyts J (2000). Perspectives for the Treatment of Infections with Flaviviridae. *Clinical Microbiology Reviews* **13:** 67-82.

Li D (2013). A highly pathogenic new bunyavirus emerged in China. *Emerging Microbes & Infections* **2:** e1.

Mackay IM, Arden KE, Nitsche A (2002). Real-time PCR in virology. *Nucleic acids research* **30:** 1292-1305.

Montgomery JM, Blair PJ, Carroll DS, Mills JN, Gianella A, Iihoshi N *et al* (2012). Hantavirus pulmonary syndrome in Santa Cruz, Bolivia: outbreak investigation and antibody prevalence study. *PLoS neglected tropical diseases* **6:** e1840.

Mulatti P, Ferguson HM, Bonfanti L, Montarsi F, Capelli G, Marangon S (2014). Determinants of the population growth of the West Nile virus mosquito vector Culex pipiens in a repeatedly affected area in Italy. *Parasit Vectors* **7:** 26.

Nash D, Mostashari F, Fine A, Miller J, O'Leary D, Murray K *et al* (2001). The outbreak of West Nile virus infection in the New York City area in 1999. *New England Journal of Medicine* **344:** 1807-1814.

Nikolay B, Diallo M, Boye CSB, Sall AA (2011). Usutu virus in Africa. *Vector-borne and zoonotic diseases* **11:** 1417-1423.

Pepin M, Bouloy M, Bird BH, Kemp A, Paweska J (2010). Rift Valley fever virus (Bunyaviridae: Phlebovirus): an update on pathogenesis, molecular epidemiology, vectors, diagnostics and prevention. *Veterinary research* **41:** 61.

Petersen LR, Fischer M (2012). Unpredictable and difficult to control--the adolescence of West Nile virus. *The New England journal of medicine* **367:** 1281-1284.

Radford AD, Chapman D, Dixon L, Chantrey J, Darby AC, Hall N (2012). Application of next-generation sequencing technologies in virology. *The Journal of general virology* **93:** 1853-1868.

Roehr B (2012). US officials warn 39 countries about risk of hantavirus among travellers to Yosemite. *Bmj* **345:** e6054.

Saitou N, Nei M (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution* **4:** 406-425.

Schneeberger PH, Becker SL, Pothier JF, Duffy B, N'Goran EK, Beuret C *et al* (2016). Metagenomic diagnostics for the simultaneous detection of multiple pathogens in human stool specimens from Cote d'Ivoire: a proof-of-concept study. *Infect Genet Evol* **40:** 389-397.

Sloan LM, Duresko BJ, Gustafson DR, Rosenblatt JE (2008). Comparison of real-time PCR for detection of the tcdC gene with four toxin immunoassays and culture in diagnosis of Clostridium difficile infection. *Journal of clinical microbiology* **46:** 1996-2001.

Solomon T (2004). Flavivirus Encephalitis. *New England Journal of Medicine* **351:** 370-378.

Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C *et al* (2002). The Bioperl toolkit: Perl modules for the life sciences. *Genome research* **12:** 1611-1618.

Steinmetz HW, Bakonyi T, Weissenböck H, Hatt J-M, Eulenberger U, Robert N *et al* (2011). Emergence and establishment of Usutu virus infection in wild and captive avian species in and around Zurich, Switzerland—Genomic and pathologic comparison to other central European outbreaks. *Veterinary Microbiology* **148:** 207-212.

Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ *et al* (2015). Big Data: Astronomical or Genomical? *PLoS biology* **13:** e1002195.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013). MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Molecular biology and evolution* **30:** 2725-2729.

Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M *et al* (2012). Primer3--new capabilities and interfaces. *Nucleic acids research* **40:** e115.

Van Belkum A, Struelens M, de Visser A, Verbrugh H, Tibayrenc M (2001). Role of genomic typing in taxonomy, evolutionary genetics, and microbial epidemiology. *Clinical microbiology reviews* **14:** 547-560.

Vazquez A, Jimenez-Clavero M, Franco L, Donoso-Mantke O, Sambri V, Niedrig M *et al* (2011). Usutu virus: potential risk of human disease in Europe. *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin* **16**.

Weissenböck H, Kolodziejek J, Url A, Lussy H, Rebel-Bauder B, Nowotny N (2002). Emergence of Usutu virus, an African mosquito-borne flavivirus of the Japanese encephalitis virus group, central Europe. *Emergence*.

Wen H, Wang K, Liu Y, Tay M, Lauro FM, Huang H *et al* (2014). Population dynamics of an Acinetobacter baumannii clonal complex during colonization of patients. *Journal of clinical microbiology* **52:** 3200-3208.

Supplementary S1

>Usutu virus template with JPEV primers

gaGgagggTTaACCTTGgAAGTGGAACaAGAGCCGTTGGGAAACCCCaGcCACATACC
AACcaGGAGAAGaTTAAAGCcaGGATTCAAAGATTGAAAGAGGAGTATGCAGCCACA
TGGCACCACGATAAGGACCACCCATATCGGACCTGgaCCTACCACGGAAGTTATGA
AGTGAAACCGACCGGTTCAGCAAGCTCCTTGGTCAACGGAGTTGTCCGCCTAATG
AGCAAGCCCTGGgATGCAATTCTCAACGTgaCCACCATGGCGATGACTGACAccaCC
CTTTTGGa

>Usutu virus template with USUV primers

tTtcaAccaTGAGatgTACTGGGTCAGTGGAGCTGCTGGCAACatTGTCCACGCAGTGA
ACATGACGAGTCAAGTGCTCATAGGGCGAATGGAGAAGAGAACATGGCATGGACC
AAAATACGAGGAGGATGTTAACCTTGGAAGTGGAACAAGAGCCGTTGGGAAACCC
CAGCCACATACCAACCAGGAGAAGATTAAAGCCAGGATTCAAAGATTGAAAGAGG
AGTATGCAGCCACATGGCACCACGATAAGGACCACcctacgGACCtggaa

>Japanese encephalitis virus template with JPEV primer

gtatgagGagaTTCACCTAGGgagCGGAGAGCCGTGGGAAAGGGAGAAGTCCATAGCA
ATCAGGAGAAAATCAAGAAGAGAATCCAGAAGCTTAAAGAAGAATTCGCCACAACg
TGGCACAAAGACCCCGAGCATCCATACCGTACTTGgACATACCACGgaAGCTATGA
AGTGAAGGCTACTGGCTCAGCCAGCTCTCTCGTCAATGGAGTGGTGAAGCTCATG
AGCAAACCTTGGGACGCCATCGCCAACGTCACCACCATGGCCATGACTGACACCa
cCCCTTTTGGa

Chapter III. Biological, serological and molecular characterisation of a highly divergent strain of GLRaV-4 causing grapevine leafroll disease

Published in “Phytopathology”, 2015. (DOI: 10.1094/PHYTO-12-14-0386-R)

Jean-Sébastien Reynard[1], Pierre H.H. Schneeberger[2,3], Jürg Ernst Frey[4], Santiago Schaerer[1]

[1]Agroscope-Virology and Phytoplasmology, Nyon, Switzerland. [2]Deparment of Epidemiology and Public Health, Swiss Tropical and Public Health Institute, Basel, Switzerland. [3]University of Basel, Basel, Switzerland. [4]Department of Methods Development and Analytics, Agroscope, Wädenswil, Switzerland.

Keywords: Grapevine; Leafroll disease; Closteroviridae; Ampelovirus; Deep sequencing; Grapevine leafroll-associated virus

1. Abstract

The complete genome sequence of a highly divergent strain of *Grapevine leafroll-associated virus 4* (GLRaV-4) was determined using 454 pyrosequencing technology. This virus, designated GLRaV-4 Ob, was detected in *Vitis vinifera* cv. Otcha bala from our grapevine virus collection at Agroscope. The GLRaV-4 Ob genome length and organization share similarities with members of subgroup II in the genus *Ampelovirus* (fam. *Closteroviridae*). Otcha bala was graft-inoculated onto indicator plants of cv. Gamay to evaluate the biological properties of this new strain, and typical leafroll symptoms were induced. A monoclonal antibody for the rapid detection of GLRaV-4 Ob by enzyme-linked immunosorbent assay (ELISA) is available, thus facilitating large-scale diagnostics of this virus. Based on the relatively small size of the coat protein, the reduced amino acid identity and the distinct serological properties, our study clearly shows that GLRaV-4 Ob is a divergent strain of GLRaV-4. Furthermore, molecular and serological data revealed that the AA42 accession from which GLRaV-7 was originally reported is in fact co-infected with GLRaV-4 Ob and GLRaV-7. This finding challenges the idea that GLRaV-7 is a leafroll-causing agent.

2. Introduction

Similar to other woody perennial crops, grapevines (*Vitis* spp.) are prone to infection by diverse viruses. Currently, more than 60 viruses have been reported to infect grapevines (Martelli 2014). Grapevine leafroll disease (GLRD) is one of the most economically important viral diseases of grapevines, and its effects on yield and harvest quality have been documented for several grapevine cultivars (Komar et al 2010, Lee and Martin 2009,

Mannini et al 2012, Spring et al 2012). Cultivars infected with GLRD generally exhibit yield reduction and poor fruit quality. For red grape cultivars, one of the primary effect of GLRD is lower anthocyanin accumulation, thus resulting in poor berry colour development. For white cultivars, GLRD symptoms are visually less evident; however, infected grapevines may show chlorotic mottling of leaves toward the end of the growing season.

GLRD has a complex aetiology associated with different filamentous viruses referred to as *grapevine leafroll-associated viruses* (GLRaVs). All GLRaVs identified to date belong to the family *Closteroviridae*. In total, 5 different GLRaV species have been identified: one in the genus *Closterovirus* (GLRaV-2), three in the genus *Ampelovirus (*GLRaV-1, GLRaV-3 and GLRaV-4*)* and one in the recently defined genus *Velarivirus* (GLRaV-7) (Al Rwahnih et al 2012). The genus *Ampelovirus* is further divided into subgroup I, consisting of GLRaV-1 and GLRaV-3 and subgroup II, consisting of all the genetically divergent GLRaV-4 strains (Abou Ghanem-Sabanadzovic et al 2012). According to the most recent taxonomic revision of the genus *Ampelovirus*, GLRaV-5, GLRaV-6, GLRaV-9, GLRaV-Pr and GLRaV-Car were classified as strains of GLRaV-4 and not, as had been previously assumed, as distinct species in the genus *Ampelovirus* (Abou Ghanem-Sabanadzovic et al 2010, Abou Ghanem-Sabanadzovic et al 2012, Maliogka et al 2009, Martelli et al 2012).

Herein, we report the description of a filamentous virus infecting a grapevine accession and showing leafroll symptoms when grafted onto cv. Gamay indicators. We present its complete genome sequence, describe the genome organization and serological features, and show that this virus is a highly divergent strain of GLRaV-4. Finally, using a combination of serological and molecular diagnostic techniques, we show

that accession AA42 is co-infected with GLRaV-7 and GLRaV-4 Ob. The implication of these findings for leafroll aetiology is discussed.

3. Materials and methods

a. Virus isolates and biological indexing

The primary grapevine materials used for this study were collected from the grapevine virus collection at Agroscope in Nyon (Switzerland), which contains more than 600 clones of distinct plant accessions (Gugerli et al 2009a). Three cuttings from the Otcha bala grapevine accession (Nos. 10,496, 10,497, and 10,498) were used for biological, serological and molecular characterization. Additional grapevine accessions used for this study included AA42, Y276 and Chiliaki Chjornyj, which were kindly provided by W. Jelkmann, O. Lemaire and the National Institute of Agrobiological Sciences (Japan), respectively. The accession Chiliaki Chjornyj was shown to be coinfected by GLRaV-7 and GLRaV-4 strain Ru (Ito et al 2013). Three additional GLRaV-7-infected accessions were provided by A. Rowhani from UC Davis: Siar, Takhani and Sultanina rose. Using microsatellite analysis, the cultivar identity of the Otcha bala plant accession was verified, and grapevine accession AA42 was identified as the grapevine cultivar Sultanine (E. Droz, personal communication).

Otcha bala canes were graft-inoculated onto the leafroll-specific indicator *Vitis vinifera* cv. Gamay Rouge de la Loire. Eight replicates were planted in the field, and symptoms were evaluated over a 3-year period. Graft-inoculated GLRaV-1-infected vines were grown as positive controls.

b. Virus particle purification and serology

Virus particles were purified from mature leaves as described previously (Gugerli et al 1984). Purified virus particles were observed using a Philips CM10 transmission electron microscope, as described by Gugerli and Ramel (Gugerli and Ramel 2004).

A cell line producing the monoclonal antibody MAb37a was generated against viruses purified from accession Y276 (Rigotti et al 2006). The serological tests used in this study consisted of double antibody sandwich enzyme-linked immunosorbent assays (DAS-ELISAs), immunoprecipitation electron microscopy (IPEM) and Western blot. These tests were performed essentially as described previously (Gugerli and Ramel 2004).

Commercially available ELISA kits (GLRaV-1 DAS, GLRaV-2 DAS, GLRaV-3 DAS and GLRaV-6 DAS from Bioreba AG, Switzerland) were used to screen for the indicated grapevine leafroll-associated viruses according to the manufacturer's instructions. DAS-ELISAs using reference antisera and monoclonal antibodies developed at Agroscope were used to test for GLRaV-4 infection. Briefly, ELISA plates were first coated with rabbit antiserum (1 µg/ml) in carbonate buffer and then incubated with grapevine crude leaf extracts for 16 hours at 6°C. Then, the wells were washed, and alkaline phosphatase-conjugated monoclonal antibodies were added. To detect GLRaV-4, GLRaV-4 strain 5 and GLRaV-4 strain 9, the following monoclonal antibodies were used: MAb 3-1, MAb 3-3 and MAb 27-1, respectively (Besse et al 2009, Gugerli et al 2009b). The reaction with the chromogenic substrate p-nitrophenyl phosphate was performed at room temperature, and the absorbance at 405 nm was read using a spectrophotometer after 3 hours.

c. Nucleic acid extraction, RT-PCR amplification and Sanger sequencing

Total RNA was extracted from mature leaf petioles using RNeasy Plant Mini Kits (Qiagen, Germany). One-step reverse transcription-polymerase chain reaction (RT-PCR) was performed using the AMV reverse transcriptase (Promega, Germany) and GoTaq polymerase (Promega, Germany) with total RNA as the template. RT-PCR was performed with primer pairs specific for each virus using the conditions described in the original publications (Supplementary Table 1).

For sequencing purposes, purified PCR products were cloned into the vector pGEM-T (Promega, Germany) and were sequenced at Fasteris SA (Switzerland). To sequence the 3'-end of the GLRaV-4 Ob genome, viral RNAs were polyadenylated using an A-Plus Poly (A) Tailing Kit (Epicentre Biotechnologies, Madison, USA), and the tailed viral RNA was used as the template in a reverse-transcription reaction. Sequences of the 5' and 3' viral termini were obtained using a 5'/3' RACE kit (Roche). Two independent clones were sequenced from each 5' and 3' terminus.

d. Viral particle enrichment, pyrosequencing, assembly and sequence analyses

Purified viral particles were treated with nucleases (DNase and RNaseA) to remove *Vitis* DNA and RNA contaminants. Then, viral RNA was extracted from purified viruses using RNeasy Plant Mini kits (Qiagen, Germany) and randomly amplified using a Whole Transcriptome kit (Sigma-Aldrich) for sequencing on a Roche 454 GS Junior platform (Roche Diagnostics Corp., Branford, CT). Sequencing libraries were prepared with a Rapid Library Preparation kit according to the manufacturer's protocol and sequenced on one PicoTiter plate using Titanium chemistry. Quality control analysis and assembly of

the produced reads were performed using DNASTAR's NGen assembler (Madison, USA) with 454-specific parameters. Filtered reads were converted to fasta files and subjected to BLASTN analysis (Altschul et al 1997) with the GenBank non-redundant nucleotide database using decreasing wordsize options of 400, 200, 100, 50 and 28.

Gene annotation was performed following a comparison with sequences from other leafroll-associated viruses and using GeneMarkS software (Besemer and Borodovsky 2005). Amino acid and nucleotide alignments were created using ClustalW (Goujon et al 2010). The sequences and accession numbers of the viral species/strains used for the amino acid sequence comparisons with GLRaV-4 Ob are provided in Supplementary Table 2. The phylogenetic relationships were determined using Molecular Evolutionary Genetic Analysis software MEGA version 6 with the best amino acid substitution model (Tamura et al 2013). Phylogenetic trees were generated using the maximum likelihood algorithm with 500 bootstrap replicates.

4. Results

a. Electron microscopy and biological indexing

Viral particles were isolated from leaf samples of the Otcha bala accession. Electron micrographs showed filamentous particles consistent with the family *Closteroviridae*, with the most frequent length being 1600 nm (data not shown). The presence of leafroll disease was assessed by biological indexing onto the leafroll-specific indicator cv. Gamay. Mild leafroll symptoms, including reddening and down curling of the leaves, were observed during the 3 consecutive years following the graft inoculation (**Figure 1**).

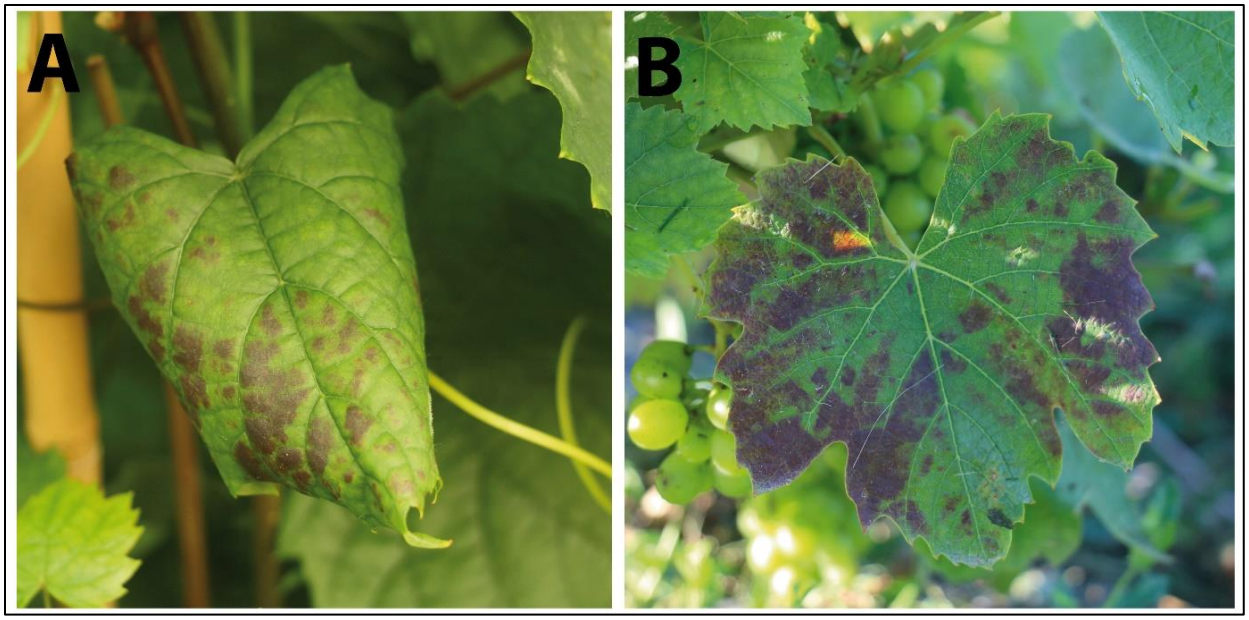


Figure 1. Leafroll symptoms on Gamay graft-inoculated with Otcha bala accession: A. downward curling o leaf margins and B. interveinal red coloration.

Original Otcha bala accessions and graft-inoculated Gamay plants repeatedly tested negative for GLRaV-1, GLRaV-2 GLRaV-3, GLRaV-4 (and its strains 5, 6, and 9) viruses by ELISA, thus justifying further investigation to characterize the cause of the disease.

b. Molecular characterization by pyrosequencing

RNA isolated from virus particles purified from the Otcha bala grapevine was submitted to 454 high-throughput sequencing. The analysis yielded 59,087 high-quality reads with an average read length of 430 bp. In total, 13,173 reads were *de novo* assembled into a 12,882-nt contig with homology to members of the *Closteroviridae* family. The coverage over this contig ranged from 1- to 1918-fold, as shown in **Figure 2**.

To verify the results provided by deep sequencing, specific primers were designed using the pyrosequencing data (Supplementary Table 3). Sanger sequencing of PCR products validated the pyrosequencing results. Completion and polishing of the sequence's termini was performed by RACE PCR using Otcha bala cDNA as the template. RACE sequencing of viral termini led to the modification of the 12,882-nt initial contig's extremities, resulting in a complete genome length of 12,849 nt. The genome sequence was deposited in the GenBank database under accession number KP313764.

Virus-derived fragments were identified in the total fragment pool based on their similarities to the nucleotide sequences archived in GenBank using BLASTN. The closterovirus-like virus, which we propose to name GLRaV-4 Ob, was the most prevalent species among the 454 dataset (**Table 1**).

Virus species	Virus family	Total hits
Grapevine leafroll-associated virus 4 Ob	*Closteroviridae*	19,572 reads
Grapevine fleck virus	*Tymoviridae*	9,002 reads
Grapevine red globe virus	*Tymoviridae*	2,687 reads
Grapevine virus A	*Betaflexiviridae*	111 reads

Table 1. High-throughput sequencing reads for viral species identified from the Otcha bala grapevine using BLASTN analysis.

Three other viruses were also identified in the 454 dataset: two viruses of the family *Tymoviridae* (*Grapevine fleck virus* [GFkV] and Grapevine red globe virus [GRGV]) and one member of the *Betaflexiviridae* (*Grapevine virus A* [GVA]). No other closterovirus-related reads were identified from the 454 run. The presence of these viral species was confirmed by specific RT-PCR analysis or ELISA.

The GLRaV-4 Ob genome is 12,849 nt in length and contains six putative open reading frames (ORFs) (**Figure 2**).

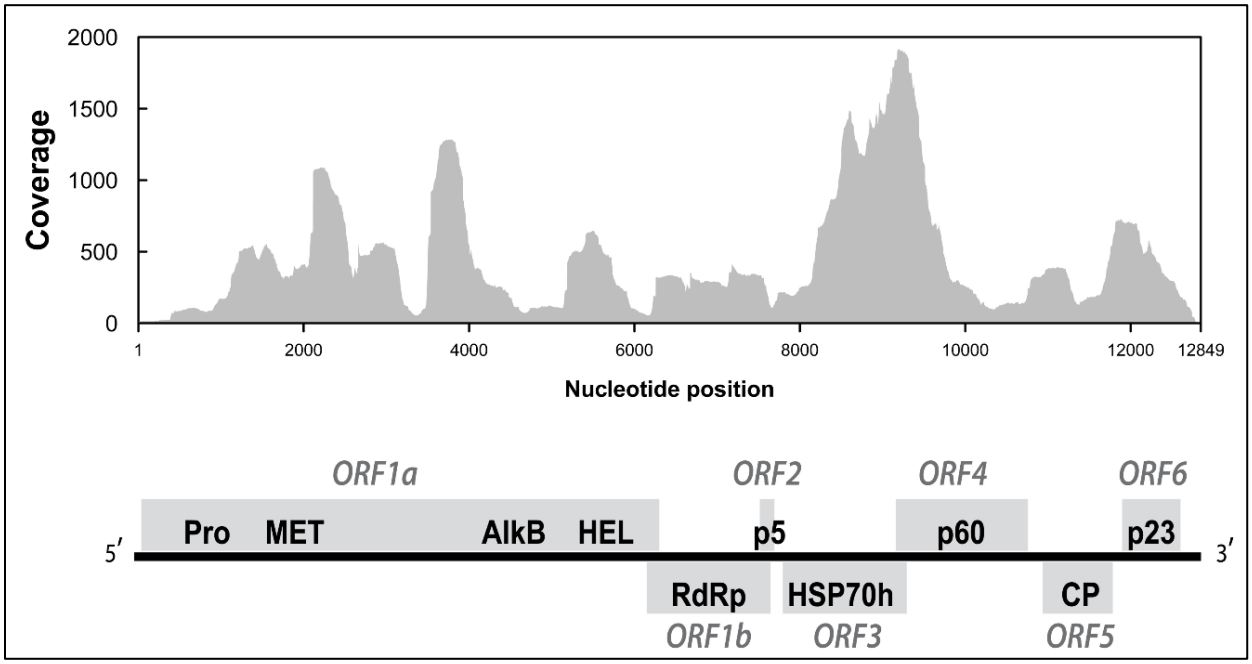


Figure 2. Sequence coverage and nucleotide positions along the *Grapevine leafroll-associated virus 4* strain Ob (GLRaV-4 Ob) genome. The schematic representation of the GLRaV-4 Ob genome organization is presented to scale. Putative open reading frames (ORFs) are shown in boxes: ORF1a with corresponding domains: Pro = protease, MET = methyltransferase, AlkB = 2OG-Fe(II) oxygenase domain, HEL = helicase; ORF1b = RNA-dependent RNA polymerase; ORF2 = small 5 K protein; ORF3 = heat shock 70 protein homolog; ORF4 = 60 K; ORF5 = coat protein; and ORF6 = 23 K protein.

The GLRaV-4 Ob genome starts with a short 37-nt-long non-coding region. ORF1a encodes a polyprotein (2076 aa). Different domains were identified in ORF1a, including a methyltransferase (MET, pfam 01660, Pfam database 27.0 (Finn et al 2014)), AlkB (pfam 03171) and helicase (HEL, pfam 01443). Additionally, ORF1a contains a papain-

like protease domain with the catalytic residues Cys^{225} and His^{268} and a predicted cleavage site after Gly^{285} (Peng et al 2001).

ORF1a terminates with the sequence augguUAG (the stop codon of ORF1a is shown in capital letters, while the start codon of ORF1b is underlined); this sequence is presumably involved in a +1 ribosomal frameshift as described for other closteroviruses (Dolja et al 2006). ORF1b overlaps the last 8 nt of ORF1a and potentially encodes a 526-aa-long protein. ORF1b shows high homology to the RNA-dependent RNA polymerase (RdRp) domain (pfam 00978). The small ORF2 partially overlaps ORF1b by 26 nucleotides and potentially encodes a 46-aa-long hydrophobic protein (p5). ORF3 is situated downstream of p5 after a 144-bp intergenic region and encodes a 533-aa HSP70-homolog (HSP70h) protein similar to other sequenced GLRaV-4s. ORF4 partially overlaps the previous ORF and encodes a 546-aa-long protein homologous to the p60 proteins of other closteroviruses. After a 69-nt-long intergenic region, ORF5 encodes a 261-aa-long protein corresponding to a viral coat protein (CP). The 3'-end proximal ORF (ORF6) encodes a putative p23 protein. ORF6 is in accordance with similarly positioned small peptides encoded by other closteroviruses at the genome's 3' end (Dolja et al 2006). The genome ends with a 131-nt-long 3' non-coding region.

c. Serological characterization

The monoclonal antibody MAb37a reacted with Otcha bala grapevine extracts in a DAS-ELISA (**Figure 3**).

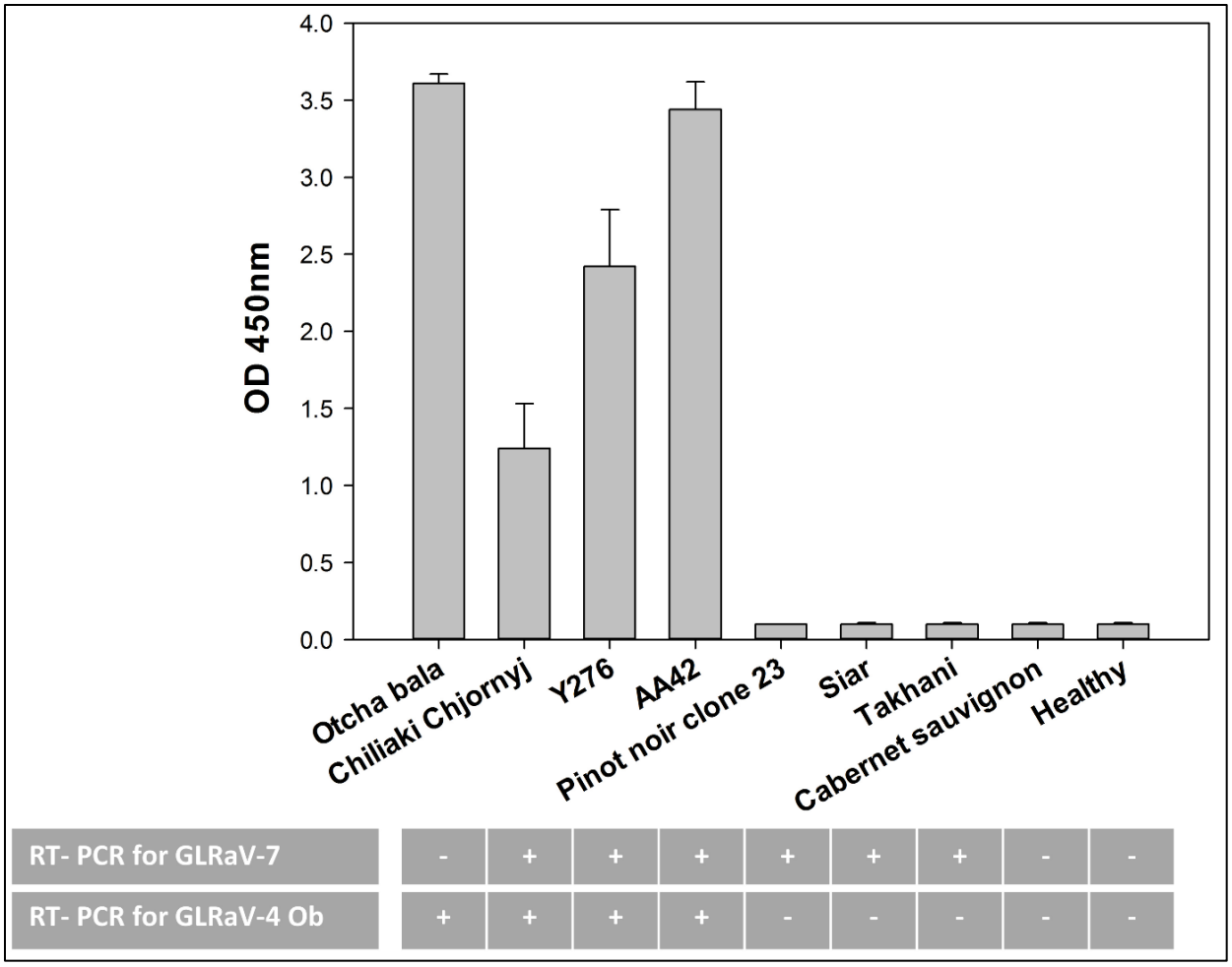


Figure 3. Specific detection of *Grapevine leafroll-associated virus 4* strain Ob (GLRaV-4 Ob) in crude leaf extracts of different grapevine accessions by homologous double antibody sandwich enzyme-linked immunosorbent assays using Mab37a (four samples were analysed for each accession; error bars represent standard deviations). Reverse transcription-polymerase chain reaction results using GLRaV-4 Ob- and GLRaV-7-specific primer pairs are shown underneath for each accession (+, specific positive amplification: -, no amplification). Leaf extracts from GLRaV-4 strain 9-infected Cabernet sauvignon and a healthy grapevine were also tested. The absorbance was read after a 3-h incubation with the substrate.

Otcha bala leaf extracts produced OD values 30 times higher than healthy controls after 3 hours of incubation. MAb37a was highly specific because it did not react with other GLRaV-4-like viruses (i.e., GLRaV-4, GLRaV-4 strain 5, GLRaV-4 strain 6 and GLRaV-4 strain 9) from infected grapevines in our collection.

MAb37a activity was further assayed by immunoprecipitation electron microscopy. The filamentous virions of the Otcha bala grapevine were heavily decorated with MAb37a (**Figure 4**).

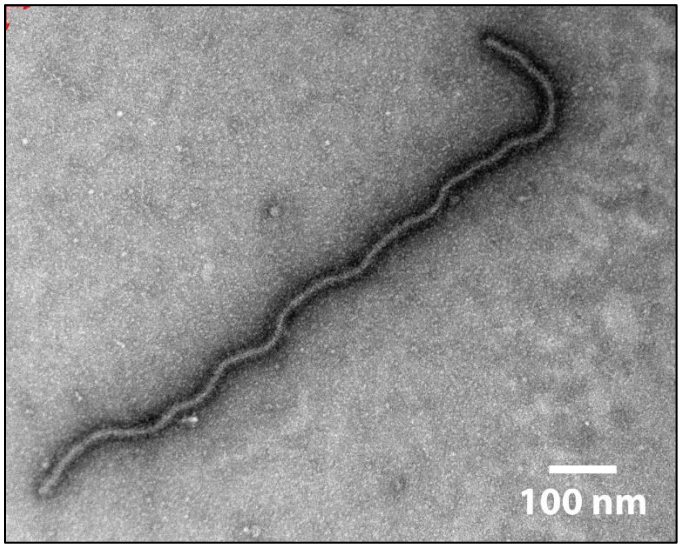


Figure 4. Immuno-precipitation electron microscopy of *Grapevine leafroll-associated virus 4* strain Ob virions decorated with Mab37a.

In Western blot analysis, MAb37a reacted to a dominant protein with an estimated molecular mass of approximately 33,000 Da (**Figure 5**).

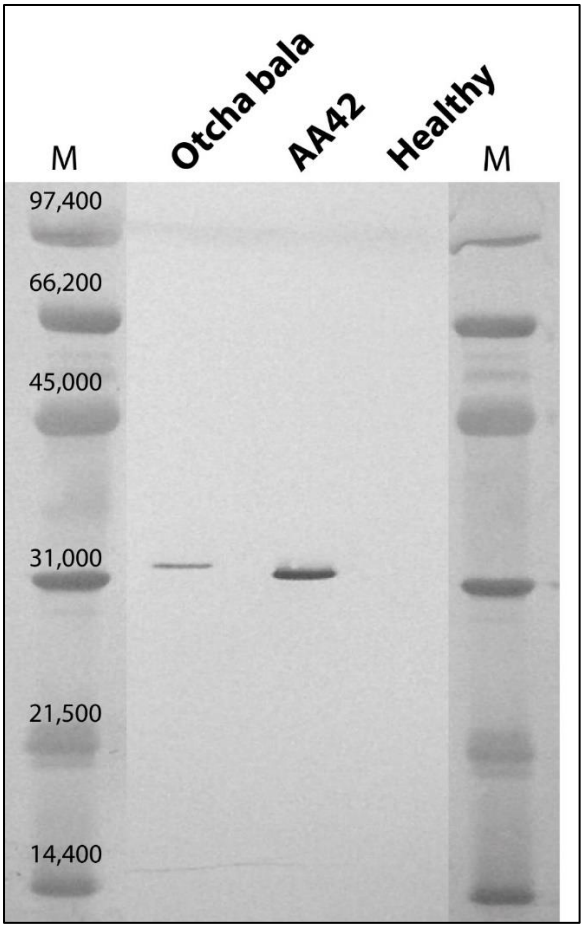


Figure 5. Detection of the Grapevine leafroll-associated virus 4 strain Ob by western blot analysis with Mab37a. M, molecular mass marker in Daltons.

d. RT-PCR assays and GLRaV-4 Ob survey of Agroscope virus collection

A DAS-ELISA assay using MAb37a was used to monitor the prevalence of GLRaV-4 Ob in our grapevine virus collection. An RT-PCR test targeting the HEL domain of ORF1a of GLRaV-4 Ob was developed using GLRaV-4 Ob-F/R primers (Supplementary Table 1) to confirm the infection status. Three other accessions from our collection tested positive for GLRaV-4 Ob by ELISA and by RT-PCR: Chiliaki Chjornyj, Y276 and AA42 (Fig. 5). To ascertain viral infection, amplicons from these accessions were sequenced, yielding nucleotide sequences with 88 to 96% identity to GLRaV-4 Ob. The three accessions

infected with GLRaV-4 Ob were also tested with the primer set LRamp-F/R reported by Abou Ghanem-Sabanadzovic (2012). With this primer set, a fragment with the expected size was amplified from all three accessions (data not shown). Amplicons were sequenced to verify the specificity of PCR products: amplicon sequence identity varied from 88% to 98%.

Because Ito *et al.* (Ito et al 2013) reported a mixed infection of GLRaV-4 and GLRaV-7 in a grapevine, we decided to evaluate the GLRaV-7 infection status of the different materials used in this study. GLRaV-7 infection was assessed by RT-PCR using different pairs of specific primers. Each RT-PCR amplification product was sequenced to verify its identity. Six accessions tested positive for GLRaV-7 (Fig. 5). The accessions Chiliaki Chjornyj, Y276 and AA42 were co-infected with GLRaV-4 Ob and GLRaV-7. The Otcha bala accession repeatedly tested negative for GLRaV-7 by RT-PCR using 5 different primer pairs.

5. Discussion

Grapevine leafroll disease has a complex aetiology; different viral species belonging to different genera in the family *Closteroviridae* are associated with the disease (Martelli 2014). In this study, we described the infection of an Otcha bala grapevine accession from our viral collection by clostero-like virus particles. Graft-inoculation of this grapevine accession to the leafroll-indicator Gamay resulted in typical leafroll symptoms. To identify the virus responsible for the leafroll symptoms, we characterized the virome of the Otcha bala accession using a pyrosequencing approach. *De novo* assembly generated a consensus sequence and revealed the presence of a divergent strain of GLRaV-4, which we propose to name GLRaV-4 Ob. Four viruses were identified in the diseased Otcha

bala grapevine, including GLRaV-4 Ob in the family *Closteroviridae*. However, GLRaV-4 Ob was the only closterovirus detected in this vine; therefore, this virus was considered to be the agent responsible for the leafroll symptoms observed on the Gamay grapevine. Similar to other viruses in the family *Closteroviridae*, the GLRaV-4 Ob genome possesses two large gene modules. One module is responsible for genome replication (MET, HEL and RdRp), whereas the other module includes five genes (p5, HSP70h, p60, CP and p23) responsible for intercellular transport and virion assembly (Dolja et al 2006). GLRaV-4 Ob's genomic organization and size are similar to viruses of subgroup II of the genus *Ampelovirus* (Martelli et al 2012, Thompson et al 2012). For example, the p23 ORF of GLRaV-4 Ob does not show any significant homology with CP ORFs and does not contain a closterovirus coat protein domain (pfam 01785). Thus, minor CP (CPm) is absent in GLRaV-4 Ob, as in all other GLRaV-4 strains (Naidu et al 2014). In contrast, members of subgroup I of the genus *Ampelovirus,* such as GLRaV-1 and GLRaV-3, all possess at least one CPm ORF in their genomes (Maree et al 2013). Furthermore, GLRaV-4 Ob consistently grouped with viruses of the GLRaV-4 cluster in subgroup II of the genus *Ampelovirus* in phylogenetic analyses performed on the HSP70h gene (**Figure 6**).

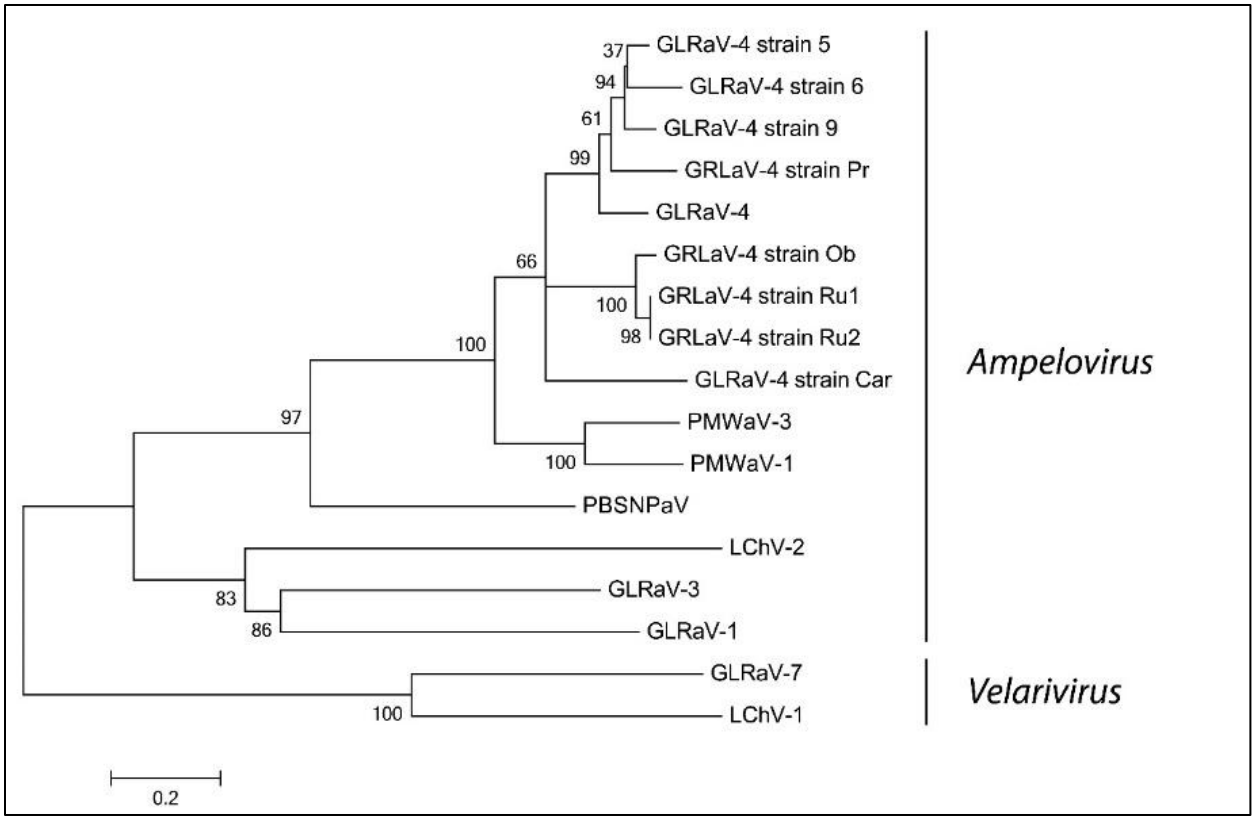


Figure 6. Unrooted phylogram constructed using a multiple alignment of heat shock 70 protein homolog amino acid sequences from some members of the genera *Ampelovirus* and *Velarivirus*. The scale represents 0.2 amino acid substitutions per site. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test is shown next to the branches.

Despite sequence similarity with other GLRaV-4 strains, GLRaV-4 Ob has several genomic features that differentiate this strain from others. i) This virus contains the smallest genome among viruses associated with grapevine leafroll disease. ii) The length of the RdRp ORF in GLRaV-4 Ob is larger than that in other GLRaV-4 strains (**Table 2**).

	RdRp	**HSP70h**	**CP**

	Identity (%)	Size (aa)	Identity (%)	Size (aa)	Identity (%)	Size (aa)
GLRaV-4 Ob		526		533		261
GLRaV-4 AA42	98	*189	98	*173	93	261
GLRaV-4 strain Ru1	n.a.	n.a.	94	*520	93	261
GLRaV-4 strain Ru2	n.a.	n.a.	94	*520	98	261
GLRaV-4	74	517	71	533	77	272
GLRaV-4 strain 5	76	517	69	533	78	269
GLRaV-4 strain 6	78	517	69	533	76	269
GLRaV-4 strain 9	76	517	70	533	75	272
GLRaV-4 strain Pr	77	517	69	533	78	273
GLRaV-4 strain Car	76	516	68	534	77	267
PMWaV-1	58	525	59	509	57	257
PMWaV-3	56	525	60	533	67	262
PBNSPaV	38	525	50	529	30	325
GLRaV-3	33	533	33	549	15	313

Table 2. Amino acid sequence identities and the sizes of different genome products from viruses of the genus *Ampelovirus*. n.a.: not available; *: partial sequence.

iii) The p5 ORF of GLRaV-4 Ob overlaps the RdRp ORF, whereas other members of the GLRaV-4 cluster have an intergenic region between those two ORFs (Abou Ghanem-Sabanadzovic et al 2012, Thompson et al 2012). Other ampeloviruses that share these features with GLRaV-4 Ob include *Plum bark necrosis stem pitting-associated virus* (PBNSPaV) and *Pineapple mealybug wilt-associated viruses 1 and 3* (PMWaV-1 and PMWaV-3) (Melzer et al 2008, Sether et al 2009).

MAb37a was raised against accession Y276, which was initially thought to be infected only by GLRaV-7, and was therefore reported to be specific to GLRaV-7 (Rigotti

et al 2006). In this work, the double infection (GLRaV-7 and GLRaV-4 Ob) of the Y276 source was demonstrated. Western blot analysis indicated that MAb37a reacted to a structural protein with an apparent molecular weight of approximately 33 kDa (Fig. 4). Coat proteins of other GLRaV-4 strains have been reported to have similar molecular weights ranging from 31 to 35 kDa as estimated by SDS-PAGE (Besse et al 2009, Gugerli et al 2009b, Rigotti et al 2006). This molecular weight is larger than the expected molecular weight calculated from the CP amino acid sequences of GLRaV-4 strains (circa 30 kDa). However, such differences between theoretical molecular weight and SDS-PAGE estimates are common (Rubinson et al 1997). MAb37a also reacted with the source AA42 (coinfected with GLRaV-7 and GLRaV-4 Ob), but not with the Pinot noir 23 source (infected only by GLRaV-7) as demonstrated by Western blot (Fig. 4), ELISA (Fig. 5), and IPEM (data not shown). These two GLRaV-7 isolates have been sequenced, and their coat proteins share high amino acid sequence homology (identity: 96.3%, similarity: 99%) (Al Rwahnih et al 2012, Jelkmann et al 2012). Moreover, three additional accessions infected by GLRaV-7, but not by GLRaV-4 Ob, were also tested and they did not react with MAb37a. Therefore, common epitopes between GLRaV-4 Ob and GLRaV-7 do not seem to exist and Mab37a should be considered to be specific to GLRaV-4 Ob and not to GLRaV-7, as stated previously.

The GLRaV-4 Ob sequences determined in this study showed 93-98% identity with the previously reported GLRaV-4 Ru sequences at the amino acid level (Ito et al 2013). Furthermore, the serological relatedness between GLRaV-4 Ob and GLRaV-4 Ru was demonstrated in this study using MAb37a in DAS-ELISA (Fig. 5). GLRaV-4 Ob and the published partial sequences of GLRaV-4 Ru1 and 2 are 87 and 88% identical at the

nucleotide level, respectively. These two variants share a common epitope recognized by MAb37a; however, this epitope is not present in other GLRaV-4 strains because no cross-reactivity was observed in DAS-ELISA against GLRaV-4, GLRaV-4 strain 5, GLRaV-4 strain 6 and GLRaV-4 strain 9 (data not shown).

The serological data were in agreement with the molecular data and strongly supported the conclusion of Ito et al. (2013) that GLRaV-4 Ob, is similar to variant GLRaV-4 Ru, belongs to a distinct strain of the GLRaV-4 species. Furthermore, the amino acid identities between taxonomically relevant genes of GLRaV-4 Ob and other members of the GLRaV-4 species were between 68-78% (Table 2). The International Committee on Taxonomy of Viruses (ICTV) adopted an amino acid divergence threshold of 25% for RdRp, HSP70h and CP for the genus *Ampelovirus* (Thompson et al 2012), making GLRaV-4 Ob a highly divergent strain. Therefore, GLRaV-4 Ob should be considered a more diverse strain of the GLRaV-4 species. GLRaV-4 strain Car is another example of a more diverged member of GLRaV-4 cluster (Abou Ghanem-Sabanadzovic et al 2010).

A number of studies have utilized different starting materials for deep sequencing, including purified viral particles (Melcher et al 2008), total RNA (Al Rwahnih et al 2009, Wylie and Jones 2011), small interfering RNAs (Kreuze et al 2009, Seguin et al 2014) and double-stranded RNAs (Al Rwahnih et al 2011, Al Rwahnih et al 2012, Coetzee et al 2010). In this study, virus particles were first purified using ultracentrifugation before applying the deep sequencing techniques. This approach allowed us to obtain the complete genomic sequence of a closterovirus. The characterization of a new virus or strain is particularly tedious and laborious for woody crops due to low concentrations of the virus or due to the presence of inhibitors such as polyphenols that may interfere with

virus purification and/or nucleic acid amplification techniques (Candresse et al 2013). Furthermore, mixed infections and the extreme diversity of viruses infecting grapevines represent challenges for studying grapevine viruses. As previously reported (Al Rwahnih et al 2009, Giampetruzzi et al 2012, Studholme et al 2011, Wu et al 2010), the results presented here demonstrate the utility and value of applying deep sequencing technology to characterize new viral pathogens and to study viral disease aetiology.

Serological and molecular data revealed that three grapevine accessions in our collection (Y276, Chiliaki Chjornyj, and AA42) are co-infected with GLRaV-4 Ob and GLRaV-7. Grapevines are prone to infection with several viruses and viral variants; thus, simultaneous infection by two or more viruses in the same grapevine plant is common (Al Rwahnih et al 2009, Goszczynski 2013, Hu et al 1990, Prosser et al 2007, Sharma et al 2011). For example, Chasselas 8/22 is co-infected with GLRaV-2, GLRaV-4 strain 5, GLRaV-4 strain 6 and an unidentified virus with isometric morphology (Gugerli et al 1997, Poudel et al 2012). Previously, Chiliaki Chjornyj was reported to be co-infected with GLRaV-7 and GLRaV-4 Ru (Ito et al 2013). Importantly, for the first time, our molecular and serological examinations of grapevine accession AA42 revealed a mixed infection of two members of the family *Closteroviridae*, GLRaV-7 and GLRaV-4 Ob.

GLRaV-7 was originally reported in a symptomless white-berried accession from Albania (AA42) that induced leafroll symptoms when grafted onto Cabernet Sauvignon indicators (Choueiri et al 1996). Because no other closteroviruses were identified in AA42, GLRaV-7 was considered the causal agent responsible for leafroll symptoms (Martelli et al 2012). However, different authors have reported that GLRaV-7 infections cause no or uncertain leafroll symptoms (Al Rwahnih et al 2012, Avgelis and Boscia 2001, Morales

and Monis 2007, Rowhani et al 2012). Our findings suggest that the leafroll symptoms from the AA42 isolate may not be related to GLRaV-7 infection as reported previously but is due to GLRaV-4 Ob. To the best of our knowledge, no case of GLRaV-7 infection associated with leafroll symptoms has been reported where co-infection with other closteroviruses can be completely excluded. Pinot noir 23 is the only grapevine accession in which GLRaV-7 Swi is present as a unique closterovirus (Al Rwahnih et al 2012), and this isolate does not induce any leafroll symptoms in Pinot noir and Cabernet sauvignon (Al Rwahnih et al 2012). Because GLRaV-7 cannot be conclusively associated with symptomatic infection, this virus may not be a leafroll-causing agent. Our findings support the proposition made by Al Rwahnhi et al. (2012) to replace the name “GLRaV-7”.

Interestingly, a situation similar to leafroll disease and GLRaV-7 may exist in cherries, another long-lived vegetatively propagated plant species. Little cherry viruses 1 and 2 are two species of the family *Closteroviridae* reported to be associated with little cherry disease (Jelkmann and Eastwell 2011). LChV-2 from the genus *Ampelovirus* induces typical little cherry disease symptoms in sweet and sour cherries (Jelkmann et al 2008). In contrast, LChV-1, similar to GLRaV-7 belongs to the newly proposed genus *Velarivirus*; symptoms of LChV-1 infection are milder or absent because some isolates may not produce typical symptoms of little cherry disease (Katsiani et al 2014, Matic et al 2009, Schröder and Petruschke 2010).

In conclusion, this study describes a new virus that induces leafroll symptoms on cv. Gamay indicators. The serological and sequencing data reported here indicate that this virus belongs to subgroup II of the genus *Ampelovirus*. Therefore, we suggest the name G*rapevine leafroll-associated virus 4 strain* Ob (GLRaV-4 Ob) for this virus. This

work clearly demonstrates that two closteroviruses are co-infecting the AA42 grapevine, the accession from which GLRaV-7 was originally reported. The results presented here, together with previous reports of symptomless infection, suggest that GLRaV-7 is not associated with leafroll disease of grapevines. Future studies will be necessary to evaluate the phenotype of GLRaV-7 infections in grapevines definitively.

6. Acknowledgments

The technical assistance of J. Brodard, N. Dubuis, M. Borgeaud and E. Droz is gratefully acknowledged. We are indebted to Dr. Olivier Lemaire, INRA, France; Prof. Wilhelm Jelkmann, JKI, Germany; Dr. Adib Rowhani, UC Davis, USA; and Dr. Takao Ito, NIFTS, Japan for providing some of the grapevine accessions used in this study. We kindly acknowledge Dr. P. Gugerli for critical reading of the manuscript and for fruitful discussions.

7. References

Abou Ghanem-Sabanadzovic N, Sabanadzovic S, Uyemoto JK, Golino D, Rowhani A (2010). A putative new ampelovirus associated with grapevine leafroll disease. *Archives of Virology* **155:** 1871-1876.

Abou Ghanem-Sabanadzovic N, Sabanadzovic S, Gugerli P, Rowhani A (2012). Genome organization, serology and phylogeny of Grapevine leafroll-associated viruses 4 and 6: Taxonomic implications. *Virus Research* **163:** 120-128.

Al Rwahnih M, Daubert S, Golino D, Rowhani A (2009). Deep sequencing analysis of RNAs from a grapevine showing Syrah decline symptoms reveals a multiple virus infection that includes a novel virus. *Virology* **387:** 395-401.

Al Rwahnih M, Daubert S, Urbez-Torres JR, Cordero F, Rowhani A (2011). Deep sequencing evidence from single grapevine plants reveals a virome dominated by mycoviruses. *Archives of Virology* **156:** 397-403.

Al Rwahnih M, Dolja VV, Daubert S, Koonin EV, Rowhani A (2012). Genomic and biological analysis of Grapevine leafroll-associated virus 7 reveals a possible new genus within the family *Closteroviridae*. *Virus Research* **163:** 302-309.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W *et al* (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25:** 3389-3402.

Avgelis A, Boscia D (2001). Grapevine leafroll-associated closterovirus 7 in Greece. *Phytopathologia mediterranea* **40:** 289-292

Besemer J, Borodovsky M (2005). GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Research* **33:** 451-454.

Besse S, Bitterlin W, Gugerli P (2009). Development of an ELISA for the simultaneous detection of *Grapevine leafroll-associated virus* 4, 5, 6, 7 and 9. *Proc 16th Congr Int Counc Study Virus Virus-like Dis Grapevine (ICVG), Dijon, France***:** 296-298.

Candresse T, Marais A, Faure C, Gentit P (2013). Association of Little cherry virus 1 (LChV1) with the Shirofugen Stunt Disease and Characterization of the Genome of a Divergent LChV1 Isolate. *Phytopathology* **103:** 293-298.

Choueiri E, Boscia D, Digiaro M, Castellano MA, Martelli GP (1996). Some properties of a hitherto undescribed filamentous virus of the grapevine. *Vitis* **35:** 91-93.

Coetzee B, Freeborough MJ, Maree HJ, Celton JM, Rees DJG, Burger JT (2010). Deep sequencing analysis of viruses infecting grapevines: Virome of a vineyard. *Virology* **400:** 157-163.

Dolja VV, Kreuze JF, Valkonen JPT (2006). Comparative and functional genomics of closteroviruses. *Virus Research* **117:** 38-51.

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR *et al* (2014). Pfam: the protein families database. *Nucleic Acids Res* **42:** D222-230.

Giampetruzzi A, Roumi V, Roberto R, Malossini U, Yoshikawa N, La Notte P *et al* (2012). A new grapevine virus discovered by deep sequencing of virus- and viroid-derived small RNAs in Cv Pinot gris. *Virus Research* **163:** 262-268.

Goszczynski DE (2013). Brief Report of a New Highly Divergent Variant of Grapevine leafroll-associated virus 3 (GLRaV-3). *Journal of Phytopathology* **161:** 874-879.

Goujon M, McWilliam H, Li WZ, Valentin F, Squizzato S, Paern J *et al* (2010). A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Research* **38:** W695-W699.

Gugerli P, Brugger JJ, Bovey R (1984). L'enroulement de la vigne: mise en évidence de particules virales et développement d'une méthode immuno-enzymatique pour le diagnostic rapide (Grapevine leafroll: presence of virus particles and development of an immuno-enzyme method for diagnosis and detection). *Revue Suisse de Viticulture, Arboriculture, Horticulture* **16:** 299-304.

Gugerli P, Brugger JJ, Ramel ME (1997). Identification immuno-chimique du 6ème virus associé à la maladie de l'enroulement de la vigne et amélioration des techniques de diagnostic pour la sélection sanitaire en viticulture. *Revue Suisse de Viticulture, Arboriculture, Horticulture* **29:** 137-141.

Gugerli P, Ramel ME (2004). Production of monoclonal antibodies for the serological identification and reliable detection of Apple stem pitting and pear yellow vein viruses in apple and pear. *Proceedings of the XIX International Symposium on Virus and Virus-Like Diseases of Temperate Fruit Crops-Fruit Tree Diseases***:** 59-69.

Gugerli P, Brugger JJ, Ramel MA, Besse S (2009a). Grapevine virus collection at Nyon: a contribution to a putative network of a worldwide grapevine virus reference collection. *Proc 16th Congr Int Counc Study Virus Virus-like Dis Grapevine (ICVG), Dijon, France***:** 40-41.

Gugerli P, Rigotti S, Ramel MA, Habili N, Rowhani A, Bitterlin W *et al* (2009b). Production of monoclonal antibodies to *Grapevine leafroll-associated virus 9* (GLRaV-9). *Proc 16th Congr Int Counc Study Virus Virus-like Dis Grapevine (ICVG), Dijon, France***:** 44-45.

Hu JS, Gonsalves D, Boscia D, Namba S (1990). Use of Monoclonal-Antibodies to Characterize Grapevine Leafroll Associated Closteroviruses. *Phytopathology* **80:** 920-925.

Ito T, Nakaune R, Nakano M, Suzaki K (2013). Novel variants of grapevine leafroll-associated virus 4 and 7 detected from a grapevine showing leafroll symptoms. *Archives of Virology* **158:** 273-275.

Jelkmann W, Leible S, Rott ME (2008). Little cherry closteroviruses-1 and -2, their genetic variability and detection by real-time-PCR. *Acta horticulturae* **781:** 321-329.

Jelkmann W, Eastwell KC (2011). Little cherry virus-1 and -2. *Virus and Virus-Like Diseases of Pome and Stone Fruits.* American Phytopathological Society Press, St. Paul, MN. pp 153-159.

Jelkmann W, Mikona C, Turturo C, Navarro B, Rott ME, Menzel W *et al* (2012). Molecular characterization and taxonomy of grapevine leafroll-associated virus 7. *Archives of Virology* **157:** 359-362.

Katsiani AT, Maliogka VI, Amoutzias GD, Efthimiou KE, Katis NI (2014). Insights into the genetic diversity and evolution of Little cherry virus 1. *Plant Pathology***:** n/a-n/a.

Komar V, Vigne E, Demangeat G, Lemaire O, Fuchs M (2010). Comparative Performance of Virus-Infected *Vitis vinifera* cv. Savagnin rose Grafted onto Three Rootstocks. *American Journal of Enology and Viticulture* **61:** 68-73.

Kreuze JF, Perez A, Untiveros M, Quispe D, Fuentes S, Barker I *et al* (2009). Complete viral genome sequence and discovery of novel viruses by deep sequencing of small

RNAs: A generic method for diagnosis, discovery and sequencing of viruses. *Virology* **388:** 1-7.

Lee JM, Martin RR (2009). Influence of grapevine leafroll associated viruses (GLRaV-2 and-3) on the fruit composition of Oregon *Vitis vinifera* L. cv. Pinot noir: Phenolics. *Food Chem* **112:** 889-896.

Maliogka VI, Dovas CI, Lotos L, Efthimiou K, Katis NI (2009). Complete genome analysis and immunodetection of a member of a novel virus species belonging to the genus Ampelovirus. *Archives of Virology* **154:** 209-218.

Mannini F, Mollo A, Credi R (2012). Field Performance and Wine Quality Modification in a Clone of *Vitis vinifera* cv. Dolcetto after GLRaV-3 Elimination. *American Journal of Enology and Viticulture* **63:** 144-147.

Maree HJ, Almeida RPP, Bester R, Chooi KM, Cohen D, Dolja VV *et al* (2013). Grapevine leafroll-associated virus 3. *Frontiers in Microbiology* **4**.

Martelli GP, Abou Ghanem-Sabanadzovic N, Agranovsky AA, Al Rwahnih M, Dolja VV, Dovas CI *et al* (2012). Taxonomic Revision of the Family Closteroviridae with Special Reference to the Grapevine Leafroll-Associated Members of the Genus Ampelovirus and the Putative Species Unassigned to the Family. *Journal of Plant Pathology* **94:** 7-19.

Martelli GP (2014). Directory of Virus and Virus-Like Diseases of the Grapevine and Their Agents. *Journal of Plant Pathology* **96:** 1-136.

Matic S, Minafra A, Sanchez-Navarro JA, Pallas V, Myrta A, Martelli GP (2009). 'Kwanzan Stunting' syndrome: Detection and molecular characterization of an Italian isolate of Little cherry virus 1. *Virus Research* **143:** 61-67.

Melcher U, Muthukumar V, Wiley GB, Min BE, Palmer MW, Verchot-Lubicz J *et al* (2008). Evidence for novel viruses by analysis of nucleic acids in virus-like particle fractions from *Ambrosia psilostachya*. *Journal of Virological Methods* **152:** 49-55.

Melzer MJ, Sether DM, Karasev AV, Borth W, Hu JS (2008). Complete nucleotide sequence and genome organization of pineapple mealybug wilt-associated virus-1. *Archives of Virology* **153:** 707-714.

Morales RZ, Monis J (2007). First detection of Grapevine leafroll associated virus-7 in California vineyards. *Plant Disease* **91:** 465-465.

Naidu R, Rowhani A, Fuchs M, Golino D, Martelli GP (2014). Grapevine Leafroll: A Complex Viral Disease Affecting a High-Value Fruit Crop. *Plant Disease* **98:** 1172-1185.

Peng CW, Peremyslov VV, Mushegian AR, Dawson WO, Dolja VV (2001). Functional specialization and evolution of leader proteinases in the family *Closteroviridae*. *Journal of Virology* **75:** 12153-12160.

Poudel B, Sabanadzovic S, Bujarski J, Tzanetakis IE (2012). Population structure of Blackberry yellow vein associated virus, an emerging crinivirus. *Virus Research* **169:** 272-275.

Prosser SW, Goszczynski DE, Meng BZ (2007). Molecular analysis of double-stranded RNAs reveals complex infection of grapevines with multiple viruses. *Virus Research* **124:** 151-159.

Rigotti S, Bittterlin W, Gugerli P (2006). Production of monoclonal antibodies to grapevine leafroll associated virus 7 (GLRaV-7). *Proc 15th Congr Int Counc Study Virus Virus-like Dis Grapevine (ICVG), Stellenbosch, South Africa***:** 200-202.

Rowhani A, Golino D, Sim S, Alrwahnih M (2012). Grapevine leafroll associated viruses effects on yields, vine performance and grape quality. *Proc 17th Congr Int Counc Study Virus Virus-like Dis Grapevine (ICVG), Davis, US***:** 52-53.

Rubinson E, Galiakparov N, Radian S, Sela I, Tanne E, Gafny R (1997). Serological detection of grapevine virus a using antiserum to a nonstructural protein, the putative movement protein. *Phytopathology* **87:** 1041-1045.

Schröder M, Petruschke M (2010). Occurrence of Little cherry virus-1 on Prunus species in the State of Baden- Württemberg, Germany. *Proceedings of the 21st International Conference on Virus and other Graft Transmissible Diseases of Fruit Crops***:** 268-271.

Seguin J, Rajeswaran R, Malpica-Lopez N, Martin RR, Kasschau K, Dolja VV *et al* (2014). De Novo Reconstruction of Consensus Master Genomes of Plant RNA and DNA Viruses from siRNAs. *Plos One* **9**.

Sether DM, Melzer MJ, Borth WB, Hu JS (2009). Genome organization and phylogenetic relationship of Pineapple mealybug wilt associated virus-3 with family Closteroviridae members. *Virus Genes* **38:** 414-420.

Sharma AM, Wang JB, Duffy S, Zhang SM, Wong MK, Rashed A *et al* (2011). Occurrence of Grapevine Leafroll-Associated Virus Complex in Napa Valley. *Plos One* **6**.

Spring JL, Reynard JS, Viret O, Maigre D, Gugerli P (2012). Effets du virus 1 associé à l'enroulement (GLRaV-1) et du virus de la marbrure (GFkV) sur le comportement agronomique et la qualité des vins de Gamay. *Revue Suisse de Viticulture, Arboriculture, Horticulture* **44:** 180–188.

Studholme DJ, Glover RH, Boonham N (2011). Application of High-Throughput DNA Sequencing in Phytopathology. *Annu Rev Phytopathol* **49:** 87-105.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013). MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. *Mol Biol Evol* **30:** 2725-2729.

Thompson JR, Fuchs M, Perry KL (2012). Genomic analysis of Grapevine leafroll associated virus-5 and related viruses. *Virus Research* **163:** 19-27.

Wu QF, Luo YJ, Lu R, Lau N, Lai EC, Li WX *et al* (2010). Virus discovery by deep sequencing and assembly of virus-derived small silencing RNAs. *P Natl Acad Sci USA* **107:** 1606-1611.

Wylie SJ, Jones MGK (2011). The complete genome sequence of a Passion fruit woodiness virus isolate from Australia determined using deep sequencing, and its relationship to other potyviruses. *Archives of Virology* **156:** 479-482.

Chapter IV. Metagenomic diagnostics for the simultaneous detection of multiple pathogens in human stool specimens from Côte d'Ivoire: a proof-of-concept study

Published in "*Infection, Genetics and Evolution*", 2015.

Pierre H. H. Schneeberger[1,2,3,4,§], Sören L. Becker[3,4,5], Joël F. Pothier[6], Brion Duffy[6], Eliézer K. N'Goran[7,8], Christian Beuret[2], Jürg E. Frey[1], Jürg Utzinger[3,4]

1 Department of Diagnostics and Risk Assessment Plant Protection, Agroscope, Institute for Plant Production Sciences IPS, Wädenswil, Switzerland, **2** Department of Virology, Spiez Laboratory, Federal Office for Civil Protection, Spiez, Switzerland, **3** Department of Epidemiology and Public Health, Swiss Tropical and Public Health Institute, Basel, Switzerland, **4** University of Basel, Basel, Switzerland, **5** Institute of Medical Microbiology and Hygiene, Saarland University Medical Centre, Homburg/Saar, Germany, **6** Institute of Natural Resource Sciences, Zurich University of Applied Sciences, Wädenswil, Switzerland, **7** Unité de Formation et de Recherche Biosciences, Université Félix Houphouët-Boigny, Abidjan, Côte d'Ivoire, **8** Centre Suisse de Recherches Scientifiques en Côte d'Ivoire, Abidjan, Côte d'Ivoire

§ **Corresponding author:** Pierre H. H. Schneeberger, Department of Epidemiology and Public Health, Swiss Tropical and Public Health Institute, P.O. Box, CH-4002 Basel, Switzerland. Tel.: +41 78 619-7067, E-mail: pierre.schneeberger@unibas.ch

E-mail addresses:

PHHS	pierre.schneeberger@unibas.ch
SLB	soeren.becker@becker-malente.de
JFP	joel.pothier@zhaw.ch
BD	brion.duffy@zhaw.ch
EKN	eliezerngoran@yahoo.fr
CB	christian.beuret@babs.admin.ch
JEF	juerg.frey@agroscope.admin.ch
JU	juerg.utzinger@unibas.ch

Keywords: Bacterial strain typing, Côte d'Ivoire, diarrhoea, metagenomics, pathobiome, resistome

1. Abstract

Background: The intestinal microbiome is a complex community and its role in influencing human health is poorly understood. While conventional microbiology commonly attributes digestive disorders to a single microorganism, a metagenomics approach can detect multiple pathogens simultaneously and might elucidate the role of microbial communities in the pathogenesis of intestinal diseases. We present a proof-of-concept that a shotgun metagenomics approach provides useful information on the diverse composition of intestinal pathogens and antimicrobial resistance profiles in human stool samples.

Methods: In October 2012, we obtained stool specimens from patients with persistent diarrhoea in south Côte d'Ivoire. Four stool samples were purposefully selected and subjected to microscopy, multiplex polymerase chain reaction (PCR), and a metagenomics approach. For the latter, we employed the National Centre for Biotechnology Information nucleotide database and screened for 36 pathogenic organisms (bacteria, helminths, intestinal protozoa, and viruses) that may cause digestive disorders. We further characterized the bacterial population and the prevailing resistance patterns by comparing our metagenomics datasets with a genome-specific markers database and with a comprehensive antibiotic resistance database.

Results: In the four patients, the metagenomics approach identified between eight and 11 pathogen classes that potentially cause digestive disorders. For bacterial pathogens, the diagnostic agreement between multiplex PCR and metagenomics was high; yet, metagenomics diagnosed several bacteria not detected by multiplex PCR. In contrast, some of the helminth and intestinal protozoa infections detected by microscopy were

missed by metagenomics. The antimicrobial resistance analysis revealed the presence of genes conferring resistance to several commonly used antibiotics.

Conclusions: A metagenomics approach provides detailed information on the presence and diversity of pathogenic organisms in human stool samples. Metagenomics studies allow for in-depth molecular characterization such as the antimicrobial resistance status, which may be useful to develop setting-specific treatment algorithms. While metagenomics approaches remain challenging, the benefits of gaining new insights into intestinal microbial communities call for a broader application in epidemiologic studies.

2. Background

An accurate diagnosis of diseases is crucial to identify the causative pathogen(s) giving rise to specific clinical syndromes, and is necessary for targeted treatment and personalized patient management (Khoury and Evans 2015, Pawlowski et al 2009). The interpretation of diagnostic test results can be straightforward when a specific pathogen is detected in a sample obtained from a normally sterile body compartment (e.g., synovial fluid in joint infections or cerebrospinal fluid in meningitis). However, diagnosis in specimens from the upper respiratory tract or the gastrointestinal tract remains challenging, as colonization with various microorganisms commonly occurs and their pathogenic potential and virulence may vary considerably (Frickmann et al 2015, Wessels et al 2014).

Diarrheal diseases and related digestive disorders may be caused by more than 40 different bacterial, parasitic, and viral pathogens (Becker et al 2013). A combination of several laboratory procedures has to be performed to cover the most common infectious agents with sufficient diagnostic accuracy (Knopp et al 2008). The epidemiology of the

causative pathogens is highly setting-specific, e.g., parasitic pathogens (helminths and intestinal protozoa) are much more common in tropical and subtropical areas, whereas bacteria prevail in industrialized countries (Fagundes-Neto 2013). However, quality data from tropical countries are scarce. In recent years, molecular diagnostic techniques, such as polymerase chain reaction (PCR) assays, have considerably improved the diagnostic yield of microbiologic stool examinations (Halligan et al 2014, McAuliffe et al 2013). In parallel, the application of these highly sensitive tools in clinical practice has brought to light that co-infection with several pathogens is the rule rather than the exception (Frickmann et al 2015, Steinmann et al 2010). Additionally, infectious agents that were previously thought to occur exclusively in symptomatic patients have been detected in healthy controls (Becker et al 2015b), which calls for the inclusion of asymptomatic controls in epidemiologic studies (Becker et al 2015a, Dubourg and Fenollar 2015). Taken together, these observations suggest that complex interactions between several intestinal pathogens and the 'normal' gut microflora rather than the presence of a single infectious agent may determine whether or not an individual develops clinical symptoms (e.g., diarrhoea) (Kinross et al 2011). However, little is known regarding the exact composition of the gut microbiome in patients and asymptomatic controls and its implications for human health.

Conventional diagnostic methods for pathogen detection in human stool samples are relatively cheap and easy to use (e.g., stool microscopy for parasites and bacterial stool culture), but they are less sensitive than molecular assays (Zboromyrska et al 2014). As an alternative, commercially available molecular assays such as the Luminex Gastrointestinal Pathogen Panel (GPP) (Wessels et al 2014) for the detection of a broad

range of pathogens have been developed. However, flexibility and adaptation of these assays with regard to genetic variation and new pathogens remain difficult. The emergence of new technologies, such as next-generation sequencing, as well as parallel optimization of bioinformatics software bring about opportunities in the field of infectious disease diagnostics. The rapidly growing field of metagenomics already provided new insights into the gut microflora (Human Microbiome Project 2012, Karlsson et al 2013, Sun and Relman 2013) and has deepened our understanding of the importance and links of intestinal micro-organisms (De Filippo et al 2010, Proal et al 2011, Scher et al 2013) with various health conditions. Metagenomics is a combination of research methodologies aimed at characterizing complex microbial communities without isolating or culturing organisms (Handelsman 2004). It is a powerful tool to study the complete range of pathogenic organisms, the so-called pathobiome, and thus generates highly valuable baseline information on rare pathogens, unculturable bacteria and multiple infections, which are common in low- and middle-income countries (Phan et al 2014, Steinmann et al 2010). Another important application of a metagenomics approach is its potential to analyse sequence datasets with several databases, thereby allowing for a distinct characterization of pathogens and antimicrobial resistance genes. Hence, this approach can provide specific health-relevant information on the patient and may guide the clinician toward personalized health care.

Here, we present a proof-of-concept study using a metagenomics approach to investigate the composition of the intestinal bacterial, parasitic, and viral flora in stool samples obtained from patients with persistent diarrhoea in Côte d'Ivoire. Additionally, the metagenomics data were compared to conventional diagnostic techniques and

multiplex PCR. Practical aspects of metagenomics in the field of medical diagnostics and public health are discussed.

3. Methods

a. Ethics statement

The stool samples analysed here were obtained during a study on persistent diarrhoea in southern Côte d'Ivoire (Becker et al 2015b). The study protocol was approved by the institutional research commissions of the Swiss Tropical and Public Health Institute (Swiss TPH; Basel, Switzerland) and the Centre Suisse de Recherches Scientifiques en Côte d'Ivoire (CSRS; Abidjan, Côte d'Ivoire). The study was cleared by the Directorate of the Hôpital Méthodiste in Dabou. The study is registered on Current Controlled Trials (http://www.controlled-trials.com; identifier ISRCTN86951400). All participants were informed in detail about the aims and procedures of the study and written informed consent was obtained before stool collection and any laboratory investigation.

b. Study area and population

The study was conducted in October 2012 in Dabou and surrounding villages, located some 30 km west of Abidjan, the economic capital of Côte d'Ivoire. The study was embedded in a preliminary investigation to identify a suitable setting for a subsequent multi-country study on the aetiology of persistent diarrhoea (≥2 weeks) and persistent abdominal pain (≥2 weeks) in resource-constrained settings of Africa and Asia (Becker et al 2015b, Polman et al 2015).

c. Field and laboratory procedures

Details of the study area, inclusion criteria, patients and asymptomatic controls, and clinical and laboratory procedures have been described elsewhere (Becker et al 2015b). In brief, fresh stool samples from individuals with persistent diarrhoea as defined by the World Health Organization (WHO; ≥3 loose stools per day for ≥2 weeks) were obtained and a short clinical questionnaire was administered. The stool samples were processed at the laboratory of the Hôpital Méthodiste in Dabou, using the following suite of laboratory examinations for parasite diagnosis: (i) Kato-Katz technique for *Schistosoma mansoni* and soil-transmitted helminths (*Ascaris lumbricoides*, hookworm, and *Trichuris trichiura*); (ii) Baermann funnel concentration for *Strongyloides stercoralis* and hookworm; (iii) Koga agar plate for *S. stercoralis* and hookworm; and (iv) formalin-ether concentration technique applied to fixed stool samples for helminths and intestinal protozoa. Additionally, rapid diagnostic tests (RDTs) were used for the detection of *Clostridium difficile* (synonymous: *Peptoclostridium difficile*), *Cryptosporidium* spp., and *Giardia intestinalis*. Small aliquots of stool samples were transferred at ambient temperature to a reference laboratory in Europe (Institute of Medical Microbiology and Hygiene; Homburg, Germany) for *post-hoc* molecular diagnosis using the Luminex GPP multiplex PCR (Becker et al 2015b). Upon arrival in the reference laboratory, samples were stored at -20°C pending further examination. For the current study, four purposefully selected stool aliquots were sent to Spiez Laboratory (Spiez, Switzerland) for in-depth analysis using shotgun metagenomics.

d. Preparation of nucleic acids

From each of the four stool samples, 150 mg were taken to extract nucleic acids in 50 µl nuclease-free water using an Isolate Faecal DNA Kit (Bioline; London, UK), following the manufacturer's instructions. Concentrations were measured on a Qubit 2.0 fluorometer (Life Technologies; Darmstadt, Germany) using the dsDNA high-sensitivity assay.

e. Sequencing and data availability

Data libraries were prepared with 10 µl of each sample using Nextera XT library kits (Illumina; San Diego, USA) and the MiSeq platform (Illumina; San Diego, USA) was used to sequence the libraries in 2×250 base pairs (bp) paired-end mode. An in-house developed Perl pipeline was used to process the sequence datasets. The pipeline consists of three main steps: (i) pre-processing and curation of the datasets; (ii) assembly of the curated sequences; and (iii) comparison of the assembled sequences to various databases.

Pre-processing was further sub-divided in three steps; namely (i) quality control; (ii) filtering of human sequences; and (iii) assembly of the datasets. The tool FastQC (Andrews 2010) version 10.1 was used to generate quality reports. The suite ea-utils (Aronesty 2011) version 1.1.2 was used to remove reads not passing the proprietary CASAVA filter of Illumina. The same software was employed to remove bases with a quality score below Q20 both at the 5' and 3' ends. Using Bowtie 2 version 2.2.3 (Langmead and Salzberg 2012) against *Homo sapiens* reference genomes allowed to filter and remove human-related sequences. Of note, no sequences pertaining to the human genome were analysed during this study.

The remaining reads were assembled using MIRA (Chevreux et al 1999) version 4.0.2 on an Ubuntu-based 24 cores and 256 GB RAM workstation. MIRA was used in "*de novo*" and "accurate" mode with four assembly passes (nop = 4).

f. Databases employed for metagenomics

Three databases were used to characterize the four stool samples: (i) National Centre for Biotechnology Information nucleotide (NCBI nt); (ii) genome-specific markers (GSMer); and (iii) comprehensive antibiotic resistance database (CARD; McArthur et al., 2013). Key features of the three databases are summarized in **Table 1**. Sequences were compared to these databases using the basic local alignment search tool (Camacho et al 2009) version 2.2.28 configured with database-specific parameters.

The NCBI nt database (Benson et al 2013) is the largest sequence repository and is widely used for genomic sequences analyses. We employed NCBI nt database to screen for pathogenic parasites, viruses, and bacteria focusing on 36 sequenced organisms that may give rise to persistent digestive disorders (Becker et al 2013). Due to high redundancy of sequences between closely related organisms in the NCBI nt database, taxonomic results, including identified strains, serovars, and pathovars, were discarded in order to ensure that only highly significant results were kept. Moreover, BLASTn parameters were kept rather stringent with four BLAST steps with decreasing wordsizes (300, 150, 100, and 50) and an *E*-value cut-off of 10^{-5}.

The recently published GSMer database (Tu et al 2014) was utilized to screen for bacterial strains with a high accuracy. Briefly, GSMer database contains strain-specific markers that were selected using a novel *k-mer*-based approach. It contains over 2 million

50-mers markers for 5,418 bacterial strains. For the screening against this database, the filtered reads were used and the BLAST analysis was performed using a wordsize of 50.

The CARD (McArthur et al 2013) is a curated and well-maintained database containing sequences of antimicrobial resistance genes. The CARD was added to investigate whether a metagenomics approach might also be suitable for identification of additional health-relevant molecular characteristics, such as genes that may confer resistance to antibiotics.

The BLAST algorithm was used with the same parameters as for the screening against NCBI's GenBank database. Results from the three BLAST searches were analysed in a sample-specific report file using the BioPerl toolkit (Stajich et al 2002). The complete taxonomic information for each BLAST hit was retrieved using the NCBI taxonomy identifier (taxid).

4. Results

a. Data analysis and patient characteristics

The conceptual framework of the analysis pipeline employed in the current study is shown in **Figure 1**. All four patients (A-D) whose faecal samples were subjected to a metagenomics approach had persistent diarrhoea. Additionally, three of these individuals concurrently complained of persistent abdominal pain. Of note, one participant was infected with the human immune deficiency virus (HIV). **Table 2** summarizes patient characteristics.

b. Identified organisms according to different diagnostic approaches

Each metagenomics dataset was screened against 36 pathogens, including diarrheagenic bacteria, helminths, intestinal protozoa, and DNA viruses (**Table 3**). The key findings from this metagenomics approach, in comparison to conventional stool microscopy, RDTs, and multiplex PCR, are summarized in **Table 4**. In brief, using shotgun metagenomics, we identified between eight and 11 potential pathogen classes in the four patients. Some pathogens, including *Aeromonas caviae*, *Escherichia coli* (represented by 7, 1, 6 and 32 different strains, respectively), *Campylobacter* spp., and microsporidia were found in all four samples. *Vibrio parahaemolyticus* was only found in sample A. One *C. difficile* strain was observed in samples A and B. Traces of *Salmonella enterica* and *Shigella* spp. were detected in samples A, C, and D. *Entamoeba histolytica* was found in samples B and D. *Mycobacterium abscessus* was detected in sample D. One *Vibrio cholerae* strain and *Yersinia* spp. were found in samples B and D, respectively. Sequences belonging to the nematode *A. lumbricoides* and the trematode *S. mansoni* were found in sample B.

None of the target organisms were identified in sample A, using standard diagnostic tools (microscopy and RDTs) and multiplex PCR. Using microscopy, *Entamoeba coli* was found in samples B and C. In addition, *A. lumbricoides*, *Chilomastix mesnili*, and *G. intestinalis* were found in sample B, whilst *B. hominis* was detected in sample C. Use of the dual-strip RDT to concurrently test for *G. intestinalis* and *Cryptosporidium* spp. revealed a positive reaction in sample B for *G. intestinalis* and for *Cryptosporidium* spp. in sample D. The presence of *G. intestinalis* and *Cryptosporidium* spp. in samples B and D, respectively, was confirmed using multiplex PCR.

c. Performance of metagenomics approach

A total of 431, 691, 3,967, and 1,056 million bases (MB) were generated for samples A, B, C, and D, respectively. The percentage of reads with more than 70% of bases with a Phred quality score over Q20 was 82% for sample A, 86% for sample B, 72% for sample C, and 78% for sample D. While it took between 171 and 464 min for three of the samples to assemble, sample D was running for 6,406 min for a complete assembly on a 48 cores machine using the MIRA assembler. **Figure 2** shows the quality of assembly for the four samples.

Subsamples of randomly selected reads among the dataset for sample C were selected in order to assess the detection rate of the different parameters with various amounts of reads. Five subsamples with 1.1, 3.3, 5.5, 7.7, and 9.9 million reads randomly selected from the initial 11 million reads of sample C were generated. **Figure 3** summarizes the number of unique taxonomic IDs, unique antimicrobial resistance genes, and several pathogenic classes found in the different subsamples.

Number of taxonomic IDs identified in the different subsamples started at 309 in subsample 1 and reached 921 of identified taxonomic IDs for the complete sample. For the pathogenic classes, the range spanned from five identified pathogenic classes for subsample 1 to 10 identified pathogenic classes for the full sample. Finally, regarding the number of unique antimicrobial resistance genes, it ranged from 33 identified genes for subsample 1 and 68 identified genes for the complete sample. The number of assembled contigs ranged from 22,233 up to 66,341, while the percentage of hits against the NCBI nt and CARD databases ranged from 56.31% to 90.98% and from 0.23% to 3.52%, respectively.

d. Antimicrobial resistance analysis

Aiming to identify other important health-related aspects, the samples were screened for a host of antimicrobial resistance genes, using CARD. **Figure 4** summarizes the potential antibiotic resistances based on the detection of the corresponding antimicrobial resistance genes.

Resistance genes for four antibiotics from two different antibiotic classes were observed in sample A. For sample B, genes were found for 14 antibiotics from six classes, including the recently introduced glycylcycline class (i.e., tigecycline). Patient C provided a sample where eight different antibiotics resistances from five classes were found. In sample D, genes that could potentially provide resistance to 25 different healthcare-relevant antibiotics from eight classes were detected.

5. Discussion

We present a proof-of-concept using a novel metagenomics approach for the diagnosis of a wide range of pathogens that may give rise to persistent digestive disorders. We purposefully selected four stool samples from well-characterized patients with persistent digestive disorders who presented to the hospital of Dabou, south Côte d'Ivoire. Sample A was provided by a 1-year-old female with persistent diarrhoea (≥2 weeks). Using all the aforementioned diagnostics methods, including multiplex PCR techniques, did not help to diagnose the cause of the clinical symptoms. The metagenomics approach enabled the identification of various pathogenic organisms that could potentially have caused the symptoms. Samples B and C were obtained from children (aged 5 and 12 years), one living in a rural and the other in an urban setting. These two samples were selected

because persistent diarrhoea in this age group is more likely to be caused by an infection, whereas other non-communicable aetiologies occur more often in older age groups. The fourth sample stemmed from a 34-year old female, HIV-infected patient. This sample was included to determine whether the spectrum of detected pathogens and resistance genes would differ significantly between an immunocompromised patient and other individuals from the same setting. Due to the immunocompromised state of this patient, it is conceivable that the individual may have experienced multiple previous infections (e.g., pneumonia and infective gastroenteritis). Because of the scope of this study, placing particular emphasis on diagnostic agreement rate between standard diagnostic tools (including microscopy and RDTs), a validated molecular tool (Luminex xTAG), and an experimental molecular tool (metagenomics), we did not include stool samples from asymptomatic controls. We believe that the results of the metagenomics analysis of a non-related stool sample would have limited outcomes since the microbiome itself is highly diverse and specific to an individual. The stool samples were examined with a suite of diagnostic techniques (i.e., microscopy, RDTs, and multiplex PCR), and subsequently subjected to a novel metagenomics approach. With regard to helminth diagnosis, our metagenomics approach holds promise. While *A. lumbricoides* was detected in sample B both by classical microscopy and metagenomics, *S. mansoni* was detected in sample B, but only with metagenomics. Of note, the full genome sequence of *S. mansoni* has been published in 2009 (Berriman et al 2009). Our findings suggest that metagenomics, provided that complete sequence data are available, has a higher detection capacity than currently more widely used methods, most importantly stool microscopy (Utzinger et al., 2015).

For intestinal protozoa, particularly *Cryptosporidium* spp., *G. intestinalis*, and *Entamoeba coli*, conventional diagnostic techniques were superior compared to our metagenomics approach. Indeed, metagenomics failed to detect these pathogens, while they were diagnosed with standard microscopy. On the other hand, our metagenomics approach proved useful for bacterial diagnosis and allowed retrieval of detailed taxonomic information. In comparison to other diagnostic techniques, which usually require a series of specific tests for the detection of various pathogens that may give rise to a clinical syndrome (e.g., various selective agar plate media for enteropathogenic bacteria), metagenomics can – in a single sequencing run – identify an extensive range of human health-relevant bacterial pathogens down to the strain level. Indeed, a very large number of bacterial genomes are well assembled and annotated. Possibly explained by the higher complexity, thus far, only a limited number of eukaryotic genomes have been completely assembled and annotated. While this represents a shortcoming for metagenomics, numerous projects aim at providing improved genomic data for higher organisms. In view of these developments, we speculate that diagnosis of intestinal protozoa using a metagenomics approach will become feasible in the not too distant future.

In order to standardize and further improve the diagnostic yield of a metagenomics approach, the establishment of complete syndrome-specific lists (e.g., persistent digestive disorders, persistent fevers, and persistent neurological disorders) (Becker et al 2013, Yansouni et al 2012, Yansouni et al 2013) as well as the establishment of the corresponding sequence databases would facilitate such analyses and would further reduce the required time to perform these in-depth diagnostics. Additionally, such syndrome-specific databases might allow the generation of pathogen-symptoms profiles

that could be compared between patients affected with the same syndrome, but with different profiles of signs and symptoms. Thereby, variations in the symptomatology could potentially be linked with the presence of multiple pathogens and would provide new insights into the complex interactions between multiple enteric pathogens, the intestinal microbiome, and arising clinical signs and symptoms (Kinross et al 2011). In particular, the impact and combined effects of multiple infections could be studied, which is of critical importance, as co-infections are the norm rather than the exception in many tropical settings (Raso et al 2004, Steinmann et al 2008, Steinmann et al 2010). Further potential benefits at the population-level could be the establishment of comprehensive databases that provide setting-specific information on prevailing antibiotic resistance mechanisms and could thus guide the adaptation and development of context-sensitive guidelines for empiric anti-infective treatment of common clinical syndromes. For individual patient management, metagenomics data might be used to provide personalized treatment, e.g., following the rapid identification of causative pathogens and their antimicrobial resistance profile. For example, we could find an elevated number of antibiotic resistance genes in sample D, which might be explained by previous anti-infective treatments, which in turn would guide personalized intervention. Such highly targeted treatments may even help to monitor and prevent the spread of antibiotic resistance development.

Nevertheless, our study has several limitations. First, the application of metagenomics on only four samples – all selected purposefully – does not allow drawing inference that could be more widely extrapolated. However, we conducted this study as a proof-of-concept and found that metagenomics indeed provides highly relevant data on

the composition of the intestinal flora and other health-related factors that may improve our understanding of the aetiology and pathogenesis of diarrheal diseases.

Second, while metagenomics provides highly accurate data, it requires the use of next-generation sequencing techniques, which are currently too expensive to be applied in resource-constrained settings and many diagnostic centres also in industrialized countries. The combined costs of microscopy, including Kato-Katz thick smear (US$ 2 per sample) and FLOTAC (US$ 2.5 per sample), RDTs (approximately US$ 4 per sample), and multiplex Luminex GPP (~US$ 80 per test) approached US$ 100 per sample in this study. Compared to a single sequencing experiment, which now costs approximately US$ 250, it is still a factor 2-3 more expensive and therefore not yet applicable in most contexts. It should be kept in mind, however that metagenomics provides significantly more health-relevant information and if optimized and further standardized, it might become the method of choice as it allows multiple pathogen identification at once.

Third, we cannot exclude that the quality of extracted nucleic acids has been negatively affected by the interruption of the cold chain from the collection of stool samples in the hospital in Dabou until final analysis in a European laboratory (Becker et al., 2015c). Hence, some information may have been lost.

Fourth, the application of further diagnostic techniques such as RNA-based meta-transcriptomics analyses would have further increased the diagnostic yield, i.e., by obtaining even more information on the current status of infections, co-infections, ‘pathogenic synergies’, infections with RNA viruses as well as phenotypically expressed antibiotic resistance patterns.

Fifth, our metagenomics approach is directly linked to resistance mechanisms of various pathogen species. The fact is that resistances might be acquired by different mechanisms, including horizontal DNA transfer (e.g., plasmids), and hence, they might result from a single point mutation in a gene or the resistance itself might be directly linked to the expression level of the corresponding gene. While a metagenomics approach allows to draw a general picture of the resistome, it might require some additional improvements (e.g., RNA-sequencing), to be able to link antimicrobial resistances to a specific bacterial strain.

In conclusion, we provide a proof-of-concept that a metagenomic approach is a powerful diagnostic tool that holds promise to deepen our understanding of infectious diseases and their pathogenesis. A large variety of pathogens could be diagnosed in clinical samples that remained undetected despite the use of a suite of sensitive diagnostic assays, including commercially available multiplex PCR assays. The diagnostic accuracy of metagenomics was high for a wide range of bacteria, but less so for the detection of parasitic pathogens, which can be explained by the current unavailability of sequence data for many human parasites. Hence, before wider application, metagenomics need further improvements pertaining to the duration of sample analysis, the high costs associated with sequencing, database content and quality, and additional tools for sequence comparison need to become available. However, it is conceivable that the insights gained from in-depth diagnostic studies employing metagenomics will considerably enhance the etiologic understanding, diagnosis, and management of diarrheal diseases and potentially other important clinical syndromes on local, regional, and global scales.

6. Competing interests

The authors declare that they have no competing interests.

7. Funding

This work was supported by armasuisse project ARAMIS no. 2011/22-16 / 353003285. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. This study was conducted as a part of the NIDIAG European research network (Collaborative Project), supported by the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 260260.

8. References

Andrews S (2010). FastQC: A quality control tool for high throughput sequence data. *http://wwwbioinformaticsbabrahamacuk/projects/fastqc/*.

Aronesty E (2011). ea-utils: Command-line tools for processing biological sequencing data. *https://codegooglecom/p/ea-utils/*.

Becker SL, Vogt J, Knopp S, Panning M, Warhurst DC, Polman K *et al* (2013). Persistent digestive disorders in the tropics: causative infectious pathogens and reference diagnostic tests. *BMC infectious diseases* **13:** 37.

Becker SL, Chappuis F, Polman K, N'Goran EK, von Müller L, Utzinger J (2015a). Epidemiological studies need asymptomatic controls. *Clinical microbiology and infection:*

the official publication of the European Society of Clinical Microbiology and Infectious Diseases.

Becker SL, Chatigre JK, Gohou JP, Coulibaly JT, Leuppi R, Polman K *et al* (2015b). Combined stool-based multiplex PCR and microscopy for enhanced pathogen detection in patients with persistent diarrhoea and asymptomatic controls from Cote d'Ivoire. *Clin Microbiol Infect*.

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J *et al* (2013). GenBank. *Nucleic acids research* **41:** D36-42.

Berriman M, Haas BJ, LoVerde PT, Wilson RA, Dillon GP, Cerqueira GC *et al* (2009). The genome of the blood fluke *Schistosoma mansoni*. *Nature* **460:** 352-358.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K *et al* (2009). BLAST+: architecture and applications. *BMC bioinformatics* **10:** 421.

Chevreux B, Wetter T, Suhai S (1999). Genome sequence assembly using trace signals and additional sequence information. *German conference on bioinformatics*.

De Filippo C, Cavalieri D, Di Paola M, Ramazzotti M, Poullet JB, Massart S *et al* (2010). Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proceedings of the National Academy of Sciences* **107:** 14691-14696.

Dubourg G, Fenollar F (2015). Epidemiologic studies need asymptomatic controls. *Clinical microbiology and infection: the official publication of the European Society of Clinical Microbiology and Infectious Diseases*.

Fagundes-Neto U (2013). Persistent diarrhea: still a serious public health problem in developing countries. *Current gastroenterology reports* **15:** 345.

Frickmann H, Schwarz NG, Rakotozandrindrainy R, May J, Hagen RM (2015). PCR for enteric pathogens in high-prevalence settings. What does a positive signal tell us? *Infectious diseases (London, England)***:** 1-8.

Halligan E, Edgeworth J, Bisnauthsing K, Bible J, Cliff P, Aarons E *et al* (2014). Multiplex molecular testing for management of infectious gastroenteritis in a hospital setting: a comparative diagnostic and clinical utility study. *Clin Microbiol Infect* **20:** O460-467.

Handelsman J (2004). Metagenomics: application of genomics to uncultured microorganisms. *Microbiology and molecular biology reviews* **68:** 669-685.

Human Microbiome Project C (2012). Structure, function and diversity of the healthy human microbiome. *Nature* **486:** 207-214.

Karlsson FH, Tremaroli V, Nookaew I, Bergstrom G, Behre CJ, Fagerberg B *et al* (2013). Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* **498:** 99-103.

Khoury MJ, Evans JP (2015). A Public Health Perspective on a National Precision Medicine Cohort: Balancing Long-term Knowledge Generation With Early Health Benefit. *JAMA* **313:** 2117-2118.

Kinross JM, Darzi AW, Nicholson JK (2011). Gut microbiome-host interactions in health and disease. *Genome medicine* **3:** 14.

Knopp S, Mgeni AF, Khamis IS, Steinmann P, Stothard JR, Rollinson D *et al* (2008). Diagnosis of soil-transmitted helminths in the era of preventive chemotherapy: effect of multiple stool sampling and use of different diagnostic techniques. *PLoS neglected tropical diseases* **2:** e331.

Langmead B, Salzberg SL (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods* **9:** 357-359.

McArthur AG, Waglechner N, Nizam F, Yan A, Azad MA, Baylay AJ *et al* (2013). The comprehensive antibiotic resistance database. *Antimicrobial agents and chemotherapy* **57:** 3348-3357.

McAuliffe GN, Anderson TP, Stevens M, Adams J, Coleman R, Mahagamasekera P *et al* (2013). Systematic application of multiplex PCR enhances the detection of bacteria, parasites, and viruses in stool samples. *Journal of Infection* **67:** 122-129.

Pawlowski SW, Warren CA, Guerrant R (2009). Diagnosis and treatment of acute or persistent diarrhea. *Gastroenterology* **136:** 1874-1886.

Phan TG, Nordgren J, Ouermi D, Simpore J, Nitiema LW, Deng X *et al* (2014). New astrovirus in human feces from Burkina Faso. *Journal of Clinical Virology* **60:** 161-164.

Polman K, Becker SL, Alirol E, Burza S, Bottieau E, Bratschi MW *et al* (2015). Diagnosis of neglected tropical diseases among patients with persistent digestive disorders: A multi-country, prospective, non-experimental case-control study. *BMC infectious diseases* **15**: 338

Proal A, Albert P, Marshall T (2011). Autoimmune Disease and the Human Metagenome. In: Nelson KE (ed). *Metagenomics of the Human Body*. Springer New York. pp 231-275.

Raso G, Luginbühl A, Adjoua CA, Tian-Bi NT, Silué KD, Matthys B *et al* (2004). Multiple parasite infections and their relationship to self-reported morbidity in a community of rural Côte d'Ivoire. *International journal of epidemiology* **33:** 1092-1102.

Scher JU, Sczesnak A, Longman RS, Segata N, Ubeda C, Bielski C *et al* (2013). Expansion of intestinal *Prevotella copri* correlates with enhanced susceptibility to arthritis. *eLife* **2:** e01202.

Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C *et al* (2002). The Bioperl toolkit: Perl modules for the life sciences. *Genome research* **12:** 1611-1618.

Steinmann P, Du Z-W, Wang L-B, Wang X-Z, Jiang J-Y, Li L-H *et al* (2008). Extensive multiparasitism in a village of Yunnan province, People's Republic of China, revealed by a suite of diagnostic methods. *The American journal of tropical medicine and hygiene* **78:** 760-769.

Steinmann P, Utzinger J, Du Z-W, Zhou X-N (2010). Multiparasitism: a neglected reality on global, regional and local scale. *Advances in parasitology* **73:** 21-50.

Sun CL, Relman DA (2013). Microbiota's 'little helpers': bacteriophages and antibiotic-associated responses in the gut microbiome. *Genome biology* **14:** 127.

Tu Q, He Z, Zhou J (2014). Strain/species identification in metagenomes using genome-specific markers. *Nucleic acids research* **42:** e67.

Wessels E, Rusman LG, van Bussel MJ, Claas EC (2014). Added value of multiplex Luminex Gastrointestinal Pathogen Panel (xTAG(R) GPP) testing in the diagnosis of infectious gastroenteritis. *Clin Microbiol Infect* **20:** O182-187.

Yansouni CP, Bottieau E, Chappuis F, Phoba M-F, Lunguya O, Ifeka BB *et al* (2012). Rapid diagnostic tests for a coordinated approach to fever syndromes in low-resource settings. *Clinical infectious diseases* **55:** 610-611.

Yansouni CP, Bottieau E, Lutumba P, Winkler AS, Lynen L, Büscher P *et al* (2013). Rapid diagnostic tests for neurological infections in central Africa. *The Lancet infectious diseases* **13:** 546-558.

Zboromyrska Y, Hurtado JC, Salvador P, Alvarez-Martinez MJ, Valls ME, Mas J *et al* (2014). Aetiology of traveller's diarrhoea: evaluation of a multiplex PCR tool to detect different enteropathogens. *Clin Microbiol Infect* **20:** O753-759.

Table 1. Key features of the three databases employed during a proof-of-concept metagenomics approach for the diagnosis of multiple pathogens in human stool samples in Dabou, south Côte d'Ivoire in October 2012. bp, base-pair; CARD, Comprehensive Antimicrobial Resistance Database; GSMer, Genome-Specific Markers database; N.A., Not available; NCBI nt; National Centre for Biotechnology Information nucleotide database.

Database	Sequence type	Number of sequences	Organism spectrum	Sequence size (bp)
NCBI nt	Publicly available sequences	182,188,746	Any sequenced organisms	N.A.
GSMer	Bacterial strain-specific markers	>2,000,000	5,418 bacterial strains	50 bp

CARD	Antibiotic resistance genes	2,993	All bacteria	N.A.

Table 2. Epidemiological and clinical characteristics of four patients with persistent diarrhoea in Dabou, south Côte d'Ivoire, in October 2012.

Characteristics	Sample A	Sample B	Sample C	Sample D
Residency	Dabou (town)	Rural village	Dabou (town)	Dabou (town)
Sex	Female	Male	Male	Female
Age (years)	1	5	12	34
Signs and symptoms	Persistent diarrhoea, nausea, vomiting	Persistent diarrhoea, abdominal pain	Persistent diarrhoea, abdominal pain	Persistent diarrhoea, abdominal pain, weight loss
Previous anti-infective treatment	No	No	No	Yes (unknown)
Comorbidity	None	None	None	HIV infection

Table 3. Summary of 36 pathogens, including bacteria, intestinal protozoa, helminths and viruses screened using the metagenomics approach.

Bacteria	**Helminths**
Aeromonas spp.; *Campylobacter* spp.; *Clostridium difficile*; *Escherichia coli*; *Mycobacterium* spp.; *Plesiomonas shigelloides*; *Salmonella* spp.; *Shigella* spp.; Tropheryma *whipplei*; *Vibrio* spp.; *Yersinia* spp.	*Ancylostoma duodenale; Ascaris lumbricoides; Capillaria* spp.*; Digenea (intestinal flukes); Diphyllobotrhium latum*; *Hymenolepsis* spp.; *Necator americanus*; *Schistosoma* spp.; Strongyloides spp.; *Taenia* spp.; *Trichuris trichuria*
Intestinal protozoa	**Viruses**
Chilomastix mesneli; *Cryptosporidium* spp.; *Cyclospora cayetanensis; Cystoisospora belli; Dientamoeba fragilis; Entamoeba* spp.; *Giardia intestinalis;* Microsporidia; *Naegleria fowleri; Neobalantidium coli*; *Toxoplasma gondii*	*Adenoviridae*; *Bocavirus*; *Cytomegalovirus*

Table 4. Comparison of conventional parasitology, rapid diagnostic tests (RDTs), a commercial Luminex multiplex PCR and a metagenomics approach for detection of intestinal pathogens in four human stool specimens obtained in Dabou, Côte d'Ivoire, in October 2012.

	Organism list	Conventional parasitology	Rapid diagnostic tests	Luminex GPP	Metagenomics	% DNA in total microbiome
Sample A	*Aeromonas* spp.				*A. caviae* (1 strain)	0,006
	Peptoclostridium difficile				1 strain	0,034
	Campylobacter spp.				Positive	0,004
	Escherichia coli				7 strains	0,091
	Microsporidia spp.				Positive	0,010
	Salmonella spp.				1 strain	0,013
	Shigella spp.				Positive	0,038
	Vibrio parahaemolyticus				Positive	0,053
Sample B	*Ascaris lumbricoides*	Positive			Positive	0,013
	Aeromonas spp.				*A. caviae* (1 strain)	0,010
	Peptoclostridium difficile				1 strain	3,485
	Campylobacter spp.			Positive	*C. jejuni* (1 strain)	0,044
	Chilomastix mesnili	Positive				
	Entamoeba spp.	*E. coli*			*E. histolytica*	0,032
	Escherichia coli				1 strain	0,012
	Giardia intestinalis	Positive	Positive	Positive		
	Mycobacterium abscessus				Positive	0,002
	Microsporidia spp.				Positive	0,004
	Schistosoma mansoni				Positive	0,118
	Shigella spp.			Positive	Positive	0,001
	Vibrio spp.				*V. cholerae* (1 strain)	0,026
	Yersinia spp.				Positive	0,006
Sample C	*Aeromonas* spp.				*A. caviae* (1 strain)	0,157
	Blastocystis spp.	*B. hominis*				
	Campylobacter spp.				Positive	0,002
	Escherichia coli			ETEC	6 strains	0,054
	Entamoeba spp.	*E. coli*			*E. histolytica*	0,002
	Giardia intestinalis			Positive	Positive	0,001
	Microsporidia spp.				Positive	0,002
	Mycobacterium abscessus				Positive	0,041
	Norovirus GI/GII			Positive		
	Salmonella spp.				Positive	0,248
	Shigella spp.				Positive	0,017

Sample D	*Aeromonas* spp.		Positive	Positive	*A. caviae* (1 strain)	0,008
	Campylobacter spp.				Positive	0,001
	Cryptosporidium spp.					
	Escherichia coli			ETEC	32 strains	2,997
	Mycobacterium abscessus				Positive	0,003
	Microsporidia spp.				Positive	0,001
	Salmonella spp.			Positive	3 strains	0,098
	Shigella spp.				1 strain	0,127
	Vibrio cholerae				1 strain	0,005
	Yersinia spp.				Positive	0,001

Figure captions

Figure 1. Bioinformatics pipeline used to retrieve information relevant to the patients' health from the metagenomics datasets. This graph summarizes the steps required for processing raw sequencing reads until the comparison of the prepared reads against three different databases.

Figure 2. Comparison of shotgun assembly metrics between four human stool samples that were provided by patients with persistent diarrhoea in Dabou, south Côte d'Ivoire, in October 2012. The values are summarized in stacked histograms showing the proportion of each parameter from each sample compared to the rest of the samples

Figure 3. Assembly comparison of sub-samples of one patient with persistent diarrhoea (sample C) in Dabou, south Côte d'Ivoire in October 2012. (A) Observed abundance of taxonomic IDs, antimicrobial resistance genes and pathogenic classes from randomly selected subsamples of sample C. (B) Number of assembled contigs from the same subsamples of sample C and the percentage having a BLASTn hit against the NCBI nucleotide database and the comprehensive antibiotic resistance database.

Figure 4. Resistome of four diarrheic human stool samples in Dabou, south Côte d'Ivoire, in October 2012. The detected antibiotic resistance genes in the stool specimens are indicated by black bars.

Figure 1.

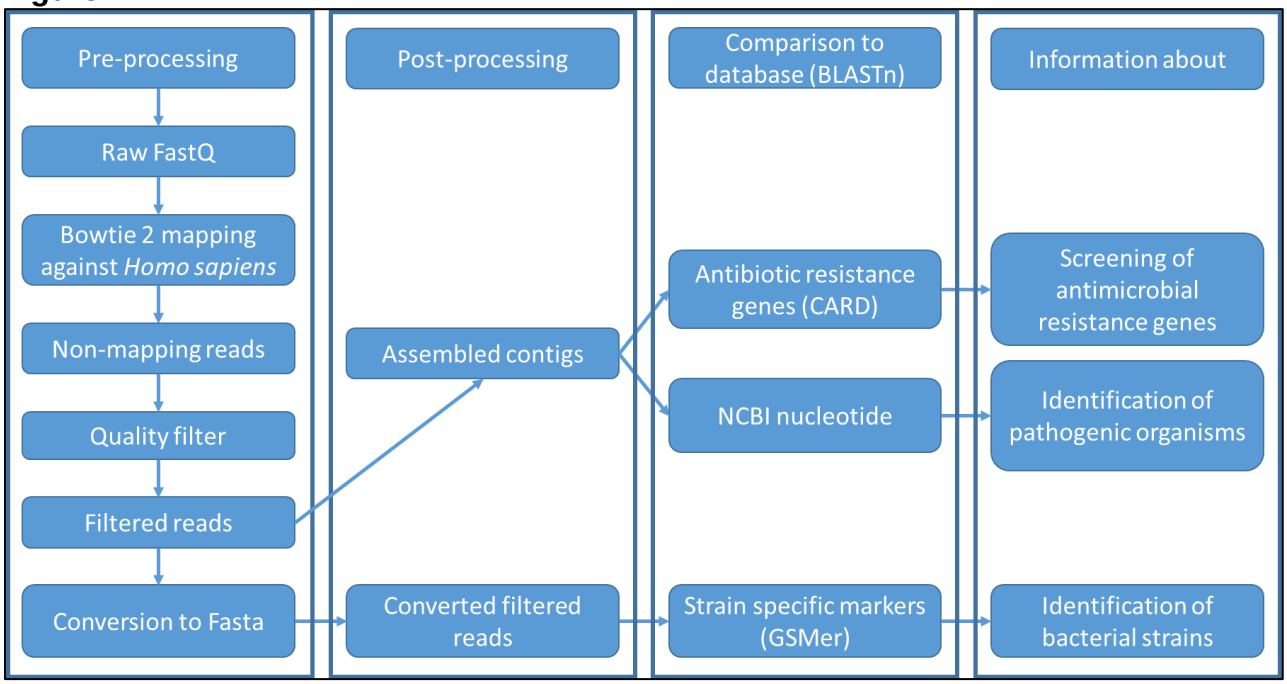


Figure 2.

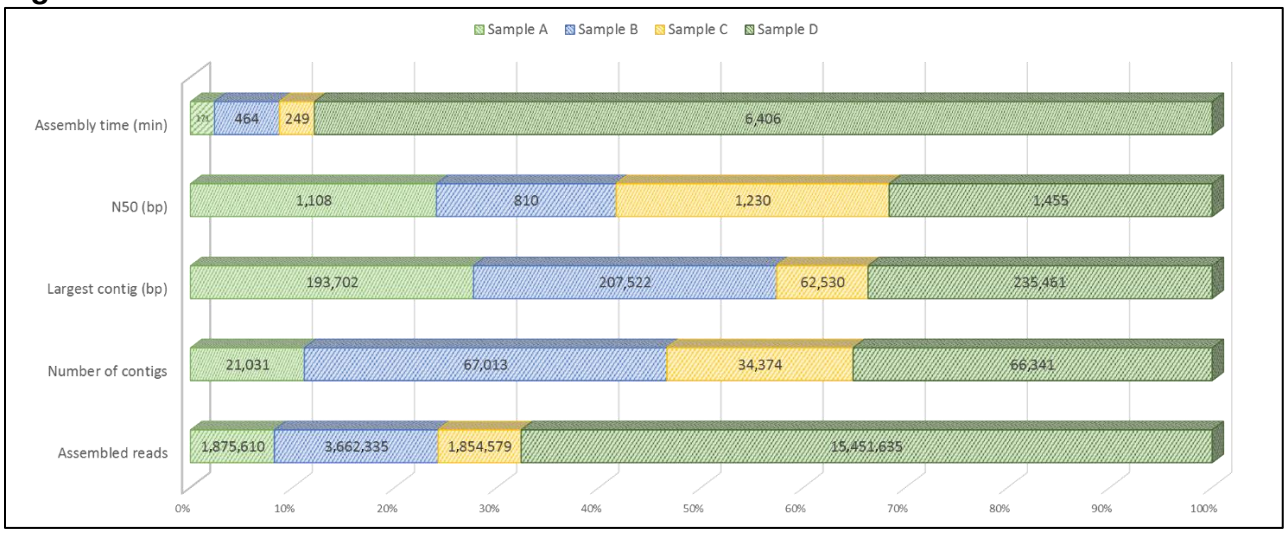
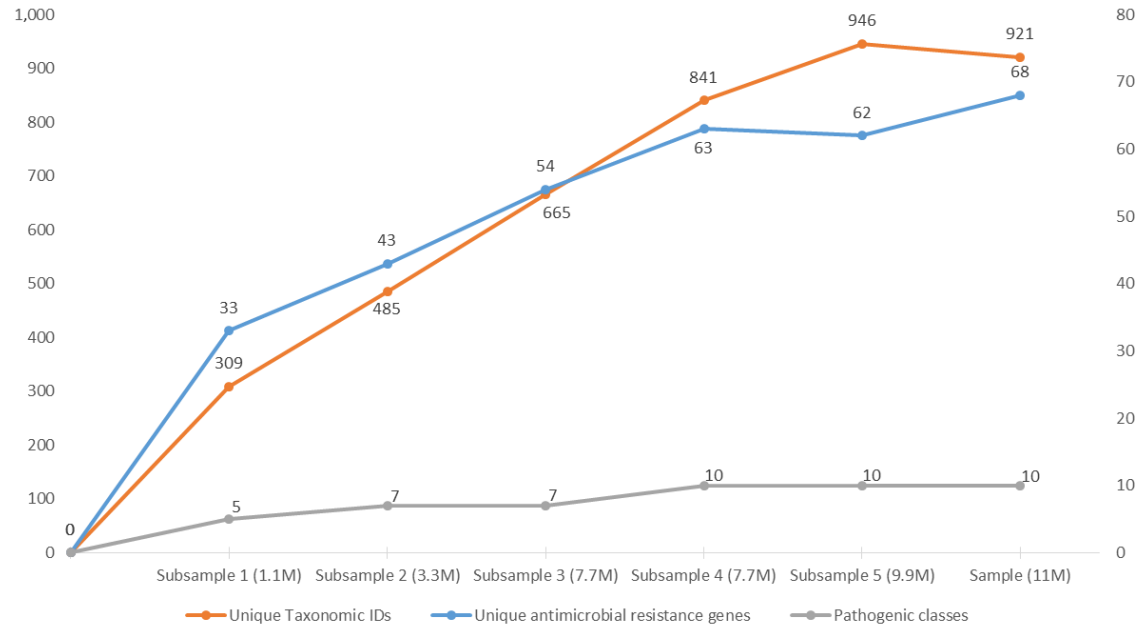


Figure 3.

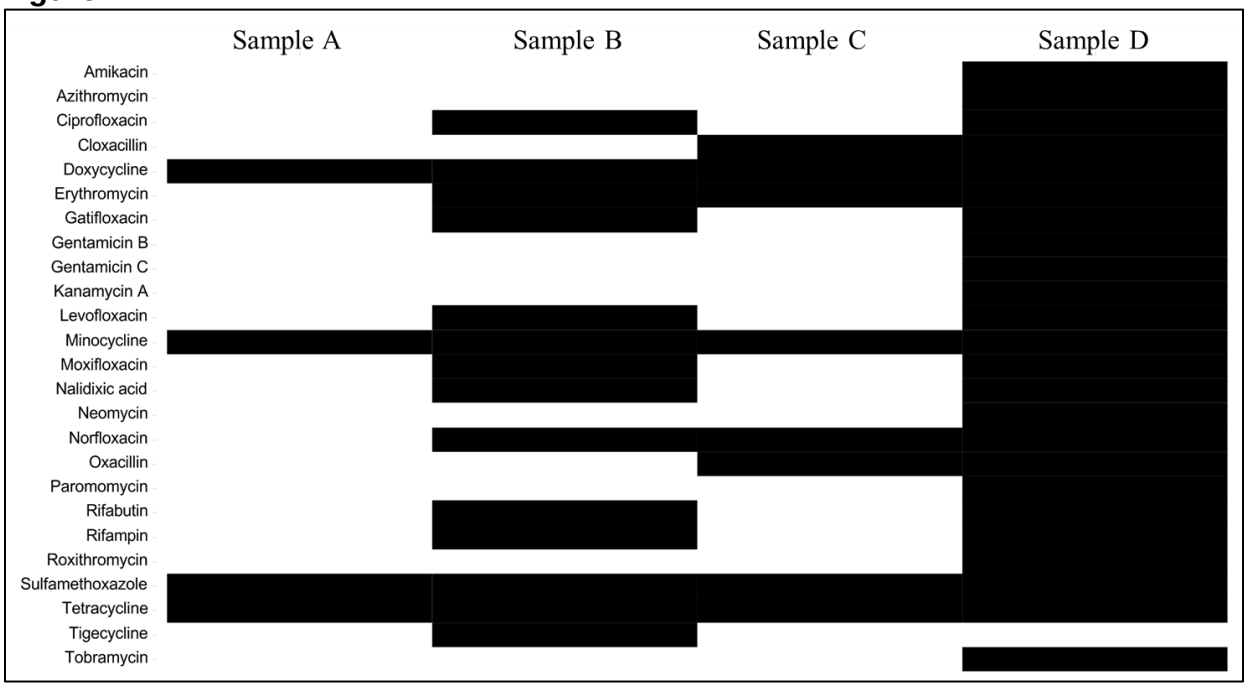
A.



B.

Sample name	Assembled contigs	Percentage hits against NCBI nt	Percentage hits against CARD
Subsample 1 (10%)	22,233	61.55	0.23
Subsample 2 (30%)	26,299	90.98	1.16
Subsample 3 (50%)	36,446	86.45	2.24
Subsample 4 (70%)	49,082	68.70	3.52
Subsample 5 (90%)	64,916	56.31	0.41
Sample C (100%)	66,341	56.88	0.43

Figure 4.



Chapter V. Microbiome profiling for an accurate assessment of microbiological health threats along a major wastewater system in Kampala, Uganda

Pierre H.H. Schneeberger[1,2,3,4§], Samuel Fuhrimann[3,4], Sören L. Becker[3,4,5], Joël F. Pothier[1,6], Brion Duffy[1,6], Christian Beuret[2], Jürg E. Frey[1], Jürg Utzinger[3,4]

1 Department of Diagnostics and Risk Assessment Plant Protection, Agroscope, Institute for Plant Production Sciences IPS, Wädenswil, Switzerland, **2** Department of Virology, Spiez Laboratory, Federal Office for Civil Protection, Spiez, Switzerland, **3** Swiss Tropical and Public Health Institute, Basel, Switzerland, **4** University of Basel, Basel, Switzerland, **5** Institute of Medical Microbiology and Hygiene, Saarland University, Homburg/Saar, Germany, **6** Environmental Genomics and Systems Biology Research Group, Institute of Natural Resource Sciences, Zurich University of Applied Sciences, Wädenswil, Switzerland

§ **Corresponding author:** Pierre H. H. Schneeberger, Department of Epidemiology and Public Health, Swiss Tropical and Public Health Institute, P.O. Box, CH-4002 Basel, Switzerland. Tel.: +41 78 619-7067, E-mail: pierre.schneeberger@unibas.ch

E-mail addresses:

PHHS	pierre.schneeberger@swisstph.ch
SF	samuel.fuhrimann@swisstph.ch
SLB	soeren.becker@swisstph.ch
JFP	poth@zhaw.ch
BD	dufy@zhaw.ch
CB	christian.beuret@babs.admin.ch
JEF	juerg.frey@agroscope.admin.ch
JU	juerg.utzinger@swisstph.ch

1. Abstract

Background: Kampala, the capital city of Uganda, is rapidly growing with annual population growth rate of up to 5.6%. Combined with unplanned urbanization, this is leading to a lack in wastewater treatment infrastructures for most of the 1.8 million inhabitants. Consequently, it is plausible that water streams and natural ecosystems around the city are heavily polluted, both with organic and inorganic contamination. However, specific data on pathogenic microorganisms, which are potentially arising health threats for the population, remain scarce. Hence, we performed an in-depth analysis using a metagenomics approach to characterize the Nakivubo system including its wastewater channel, the surroundings wetlands, and, to some extent, the inner Murchison bay.

Methods: In October 2013, we obtained water samples from 23 locations distributed homogeneously within three ecosystems: the Nakivubo channel itself, the wetland areas located around it, as well as four samples collected at the Murchison bay, on Lake Victoria. The samples were concentrated on-site using tangential flow filtration and transferred to Switzerland, where they were sequenced using Illumina's technology. A total of 1.2 billion sequencing reads were generated and compared either after quality filtering to a bacterial strain-specific database (GSMer), or after de novo assembly to the NCBI nt database.

Results: Based on the composition of the bacterial communities in the water samples, three clearly differentiated clusters could be identified with regards to their microbial diversity. A high correlation between *Escherichia coli* intra-species diversity and the total bacterial strains was identified. Using linear regression analyses, it was possible to find

strong correlation of the taxonomic composition with distance of several non-channel samples with their closest channel samples, allowing us to highlight potential leakage points of the Nakivubo channel. The three environmental clusters harbour several significant differences in their respective microbial structure. Several pathogenic microorganisms could be strongly correlated with wastewater contamination across the system.

Conclusion: This study is an example of the power of metagenomics approaches to perform system wide-characterization of the environment. To our knowledge, it is the first of its kind, attempting to characterize the main wastewater system in a booming African megapolis and trying to find the potentially related health hazards for the human populations. We were able to make important statements concerning the system including i) that it has a sub-optimal wastewater treatment capacity, ii) that containment potential of the wastewater isn't sufficient, iii) highlight the potential leakage points, and iv) to pinpoint potential risks for human health arising from contamination of the environment by wastewater.

2. Introduction

Kampala is the capital city of Uganda and is located on the northern shores of Lake Victoria, at an altitude of 1140 m above sea level. The climate in Kampala is tropical with precipitations throughout the year, mainly concentrated during two rainy seasons, the first one occurring between March and May and the second one from October to November. With an annual population growth of 5,6% and a population of more than 1.5 million inhabitants in 2014, it is among the fastest growing cities in sub-Saharan Africa (Statistics 2001, Vermeiren et al 2012). Despite its fast demographic and overall economic

development, arising social and health-related challenges have not yet been fully addressed, as can be seen e.g. in the relatively moderate increase of funding in the field of water supply (Okuonzi 2004). Considerable population growth in combination with rapid urbanization are putting pressure on existing wastewater infrastructures. Indeed, only approximately 10% of Kampala's total population is connected to a sewer, while the large majority is relying on on-site sanitation systems such as pit latrines and septic tanks (Fuhrimann et al 2014, Fuhrimann et al 2015, Kansiime and Maimuna 1999). Moreover, industrial development and urban farming have led to a reduction of wetland systems around the city that have previously served as natural wastewater treatment resources (Fuhrimann et al 2015, Kansiime and Maimuna 1999, Mbabazi et al 2010). With a surface of 5.29 km^2 and a total catchment area of over 40 km^2 (Emerton et al 1999), the Nakivubo wetland is the largest of a series of 12 wetland areas surrounding the city of Kampala. It is divided by an old railway line, with the area located north of the railway being composed mainly of drained wetland and the area located south of it composed mainly of floating wetlands. The Nakivubo wetland also serves as an agricultural ground, with yams and sugar cane being the main cultivated crops. As farmers directly re-use wastewater for irrigation purposes, any health threats present in the water will impact on the farmers' health and the safety of agricultural products grown in this area. The Nakivubo wetland area is also subjected to flooding events, especially during the rainy season, and this puts an estimated 12'000 individuals living in the surrounding slums at risk of direct contact with wastewater (Kayima et al 2008, Mbabazi et al 2010). Several studies elucidated the potential risk of exposure to waste water on human health (Al-Jassim et al 2015, Becerra-Castro et al 2015, Lu et al 2015, Youenou et al 2016). The World Health Organization

(WHO) recommends to assess and quantify standard indicators of water faecal contamination, e.g. by microbiological analysis of the number of faecal bacteria in water samples (WHO 2011). A recent study by Fuhrimann *et al.* (2015) reported that the counts of colony-forming units (CFUs) of both *Escherichia coli* and *Salmonella* spp. along the main wastewater treatment system in Kampala, including the Nakivubo channel and wetlands, were above the thresholds set by WHO for unrestricted use for irrigation in agriculture (WHO 2011). However, detailed phylogenetic and microbiological information on the exact composition of pathogenic organisms is scarce in Kampala and other rapidly growing urban areas in sub-Saharan Africa (Bateganya et al 2015, Youenou et al 2016). Hence, we employed a shotgun metagenomics approach on a set of water samples collected along the Nakivubo channel to characterise the microbiological composition of this dynamic environment and to further assess potentially arising health consequences for the exposed population.

The overarching aim of this study was to use an advanced molecular approach in a resource-constraint setting, namely, metagenomics, to provide an in-depth representation of the effective microbial contamination, and potentially human-health specific risks, along the main wastewater treatment network in the city of Kampala. We aimed at providing a system-wide analysis of the Nakivubo system by, i) grouping samples with regards to their microbial profiles, ii) characterizing each group's specificities and iii) focusing our analysis on the distribution of bacterial, eukaryotic, viral and fungal pathogens and their relations with wastewater contamination, in this specific context.

3. Methods

a. Sampling strategy

In the frame of this study, a total of 23 water samples were collected at different locations distributed all along the wastewater collection network of Kampala city, as shown in **Figure 1.** The samples were collected within one day, to provide a comparable snapshot of the bacterial communities. Sample collection was performed in a relatively dry period in October 2013.

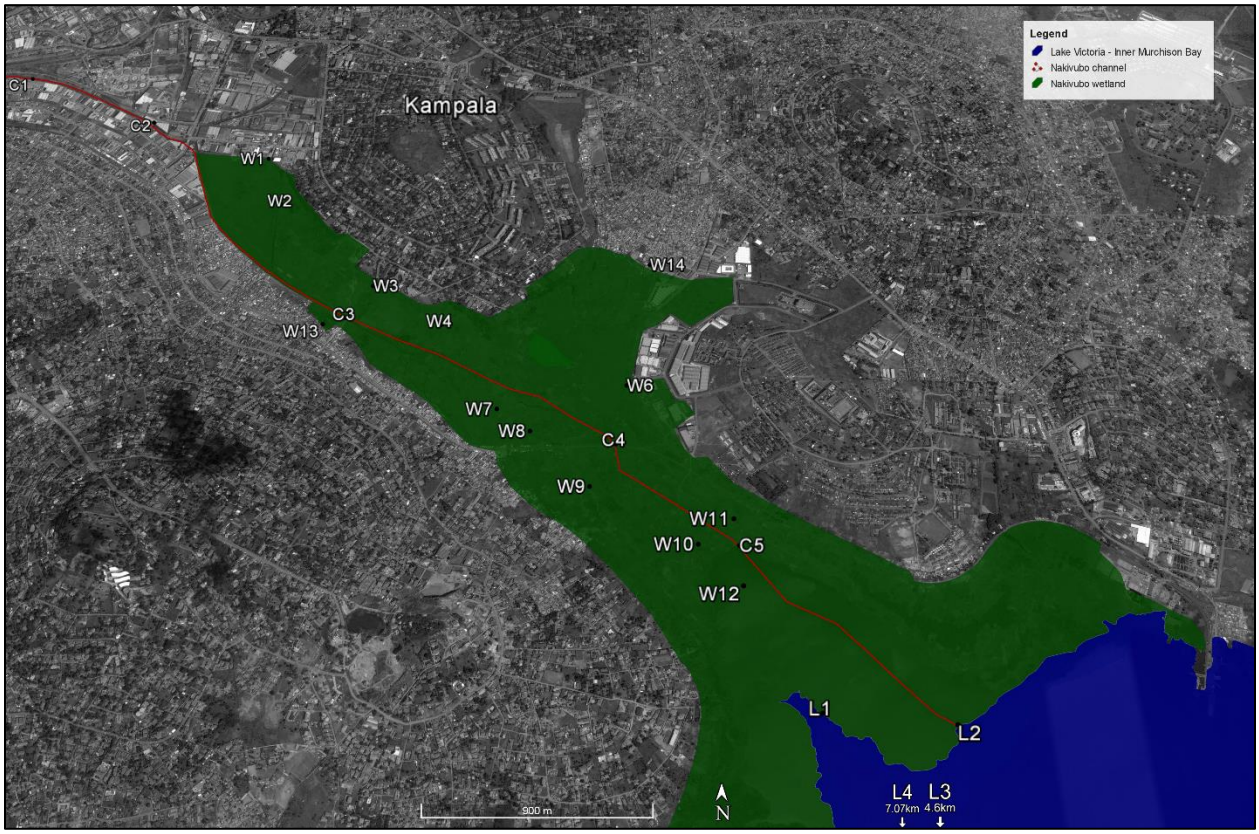


Figure 1. Map of the study area, including Kampala city, the Nakivubo wetland area and the inner Murchison Bay (Lake Victoria). This map shows the sampling locations of the 22 water samples. Samples collected on the channel are coded with C (1-5), samples

collected in the wetlands are coded with W (1-14) and samples collected on Lake Victoria are coded with L (1-4).

The 23 samples were labelled belonging to one of the three ecosystems we planned to study, namely, (i) the Nakivubo wastewater channel (Samples C1 to C5); (ii) the Nakivubo wetlands (Samples W1-W14); and (iii) the inner Murchison bay of Lake Victoria (Samples L1-L4). Two samples from the W series (Samples W13 and W14) were collected directly at informal communities' outlets flowing directly into the wetlands. One sample from the L series (L3) was collected directly at the city's freshwater intake, located in Gaba, approximately 5.8 kilometres south from the Nakivubo channel outlet. Sample L4, located in the middle of the Murchison Bay, approximately 8.2 kilometres south from the Nakivubo outlet, serves as a low human-related contamination control. Of note, one tube, which contained sample W5 broke during transportation, therefore this sample could not be analysed and no results on this specimen can be reported. Hence, 22 samples were used for the final analysis.

b. Sample collection procedure, storage and nucleic acid extraction.

At each location, a minimum of one litre of surface water was collected. Upon arrival in the local laboratory, the samples were directly stored in a fridge at a temperature of 4°C. Each sample was then concentrated using a tangential flow filtration unit into a smaller volume of approximately 50 ml. The concentrated samples were frozen at -20°C and transferred to Switzerland in a cooling box to avoid microbial growth. Upon arrival in Switzerland, the 50 ml samples were further concentrated with Amicon® Ultra Centrifugal

Filter Units with a molecular cut off of 10K Daltons (Millipore; Billerica, MA, United States of America) into a smaller volume of approximately 150 µl.

Nucleic acids were isolated from 150 µl of the concentrated samples using a PowerSoil DNA isolation kit (MO-BIO; Carlsbad, CA, United States of America) following the manufacturer's instruction except for the elution step that was done in 60 µl purified water. Extracted samples were quantified on a Qubit 2.0 fluorimeter (Life Technologies; Darmstadt, Germany) using the dsDNA high-sensitivity assay.

c. Sequencing and data analysis.

DNA libraries were prepared from 30 µl of the different samples using NEBNext Ultra (New England Biolabs; Ipswich, MA, USA) library preparation kits. Samples were pooled on an Illumina HiSeq 2500 (Illumina; San Diego, CA, USA) in 2^x125 base pairs (bp) paired-end mode for sequencing. An in-house developed Perl pipeline was used to automatize the dataset analysis in three steps, namely (i) a pre-processing step of the raw sequences datasets; (ii) an assembly of the curated datasets; and (iii) the comparison of the obtained sequences to various databases.

Pre-processing of the raw datasets was further divided into two sub-steps, including (i) a quality control; and (ii) a quality filtering of raw sequences. The tool FastQC (Andrews) in version 10.1 was used to assess the overall sequencing quality and the software suite EA-utils (Aronesty 2011) in its version 1.1.2 was used to remove the reads not passing the proprietary CASAVA filter from Illumina. The same tool suite was used to remove bases with a quality score below Q20 at both 5' and 3' ends. The assembly was performed using the MIRA assembler (Chevreux 2007) in its version 4.0.2 in *de novo* and accurate modes with four assembly passes for each sample (nop= 4). Computing-wise,

a high-performance cluster running under CentOS 6.5 was used and 24 computing cores and >512 gigabyte RAM were allocated to each assembly job.

The third step of the analysis was the comparison to two different databases, namely (i) the National Centre for Biotechnology Information nucleotide (NCBI nt) database (Benson et al 2013); and (ii) the genome-specific markers database (GSMer; (Tu et al 2014)). The comparison was performed using the BLAST software in its version 2.2.28+ (Camacho et al 2009). The features of the comparative analyses against the two different databases are summarized in **Table 1**. The complete taxonomic information for each BLAST hit was retrieved using the NCBI taxon identifier (taxid) and the corresponding BioPerl (version 1.2.9; (Stajich et al 2002)) features.

Database	**NCBI nt**	**GSMer**
Sequence type	Publicly available sequences	Bacterial strain-specific markers
Number of sequences	182’188’746	> 2’000’000
Organism spectrum	Any sequenced organisms	5418 bacterial strains
Sequence size (bp)	N.A.	50 bp
BLASTn parameters	WS: 300 -> 150 -> 100 -> 50; EVC: 10^-5	WS: 50; EVC: 10^-5
Sequenced compared	Assembled contigs	Curated reads
Information obtained	Prevalence of non-bacterial pathogens with quantitative information	Prevalence of bacterial strains with quantitative information

Table 1. Databases used in the metagenomics approach. This table shows the databases that were used in the bioinformatics workflow and highlights their characteristics. bp = base pairs; WS = Word size; EVC = *E-value* cut-off.

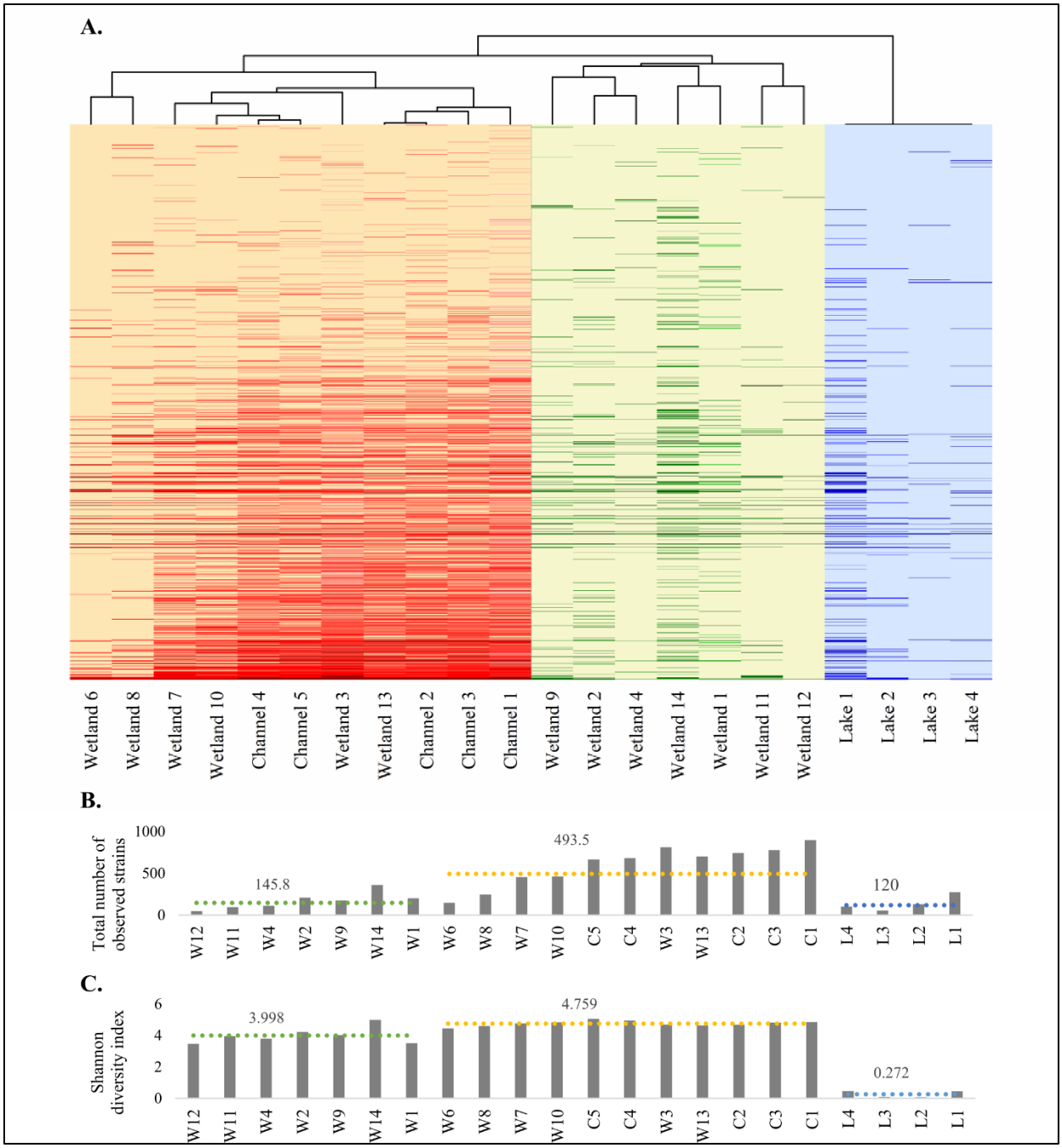
4. Results

a. Sequencing profiles.

Measured DNA concentrations ranged from 1.77 ng/µl to 19.2 ng/µl in a volume of 60 µl for a total DNA minimum of 106.2 ng and up to a maximum of 1’152 ng per sample. A total of 1’240’255’828 reads were sequenced for the twenty-two samples with an average of 56’275’265 reads per sample. On average, 22% of the reads were assembled into approximately 292’000 contigs with a N50 size of 645 base pairs (bp). Out of these ~300’000 contigs, approximately one quarter were contigs larger than 500 bp with a N50 of 1’421 bp. When utilising BLAST analysis, an average 44 % of the assembled contigs in each sample had a hit in the NCBI database. With regard to the GSMer database, an average of 11’588 matching reads or markers were found for each sample. Detailed results are provided in Supplementary Table 1.

b. Spatial relationships.

Using the taxonomic profiles derived from the comparison of the datasets with the GSMer database, we performed a hierarchical analysis of all samples, as shown in **Figure 2**.


A.
Wetland 6
Wetland 8
Wetland 7
Wetland 10
Channel 4
Channel 5
Wetland 3
Wetland 13
Channel 2
Channel 3
Channel 1
Wetland 9
Wetland 2
Wetland 4
Wetland 14
Wetland 1
Wetland 11
Wetland 12
Lake 1
Lake 2
Lake 3
Lake 4
B.
Total number of observed strains
1000
500
0
145.8
493.5
120
W12
W11
W4
W2
W9
W14
W1
W6
W8
W7
W10
C5
C4
W3
W13
C2
C3
C1
L4
L3
L2
L1
C.
Shannon diversity index
6
4
2
0
3.998
4.759
0.272
W12
W11
W4
W2
W9
W14
W1
W6
W8
W7
W10
C5
C4
W3
W13
C2
C3
C1
L4
L3
L2
L1

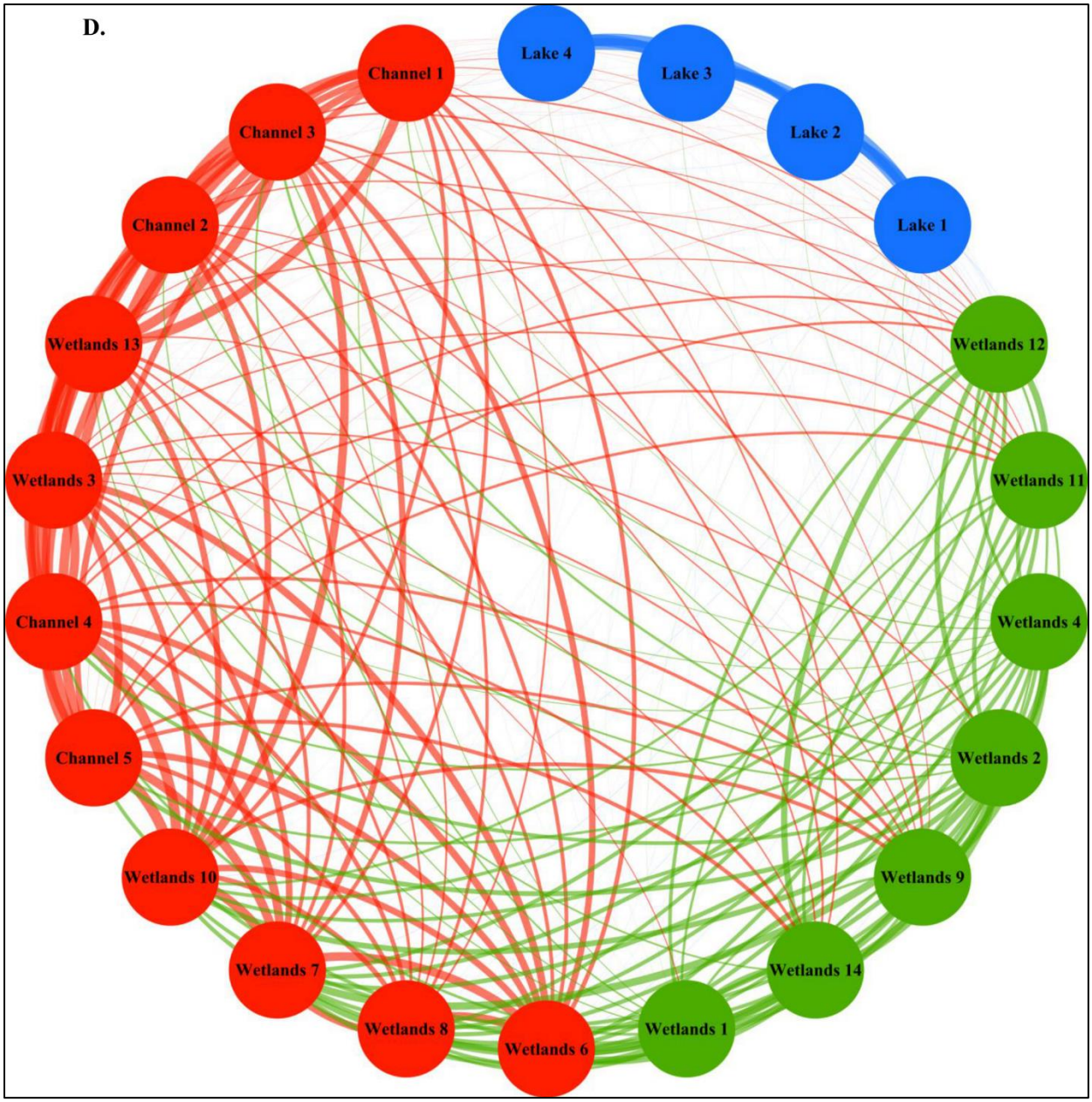


Figure 2. Sample-to-sample relationships. Panel A. Correlation-based hierarchical cluster analysis of water samples based on the relative abundance of bacterial strains. The tree is constructed using the group average method and a Pearson correlation matrix. The dendrogram shows the degree of similarity between the different samples (scale = 1). The three clusters that were selected for the rest of the study are highlighted in orange colour (cluster 1), green colour (cluster 2) and blue colour (cluster 3). Panel B shows the

total number of observed strains, per sample. The average cluster diversity is shown in dashed lines. Panel C shows the Shannon diversity index, per sample. The average value of the Shannon diversity per cluster is shown in dashed lines. Panel D. Radial representation of the sample-to-sample correlations coloured by cluster. The thicker the connecting line, the stronger the correlation between the samples.

All samples collected on Lake Victoria (L1-L4) cluster together in the most distant ramification, that we named cluster 3. The rest of the samples is separated in two distinct branches, namely, (i) the samples collected at the channel locations (C1-C5) together with samples collected at six wetland locations (W3, W6, W7, W8, W10 and W13); and (ii) the remaining six wetland samples (W1, W2, W4, W11 and W12) that we subsequently refer to as cluster 1 and 2, respectively. The average number of bacterial strains found in cluster 1 (n = 493.5) was significantly higher than the average number of strains found in cluster 2 (p-value = $4.3*10E^{-4}$) and cluster 3 (p-value = 0.003). In contrast, the difference between cluster 2 and 3 was not significant. Similarly, the mean Shannon diversity index (SDI), an indicator taking into account both abundance and evenness within a sample, was significantly lower in cluster 2 (p-value = 0.003) and cluster 3 (p-value = $2.2*10E^{-15}$) than in cluster 1. SDI of cluster 2 and 3 is also significantly different (p-value = $2.4*10E^{-7}$).

We further used the GSMer derived profiles compare the number of identified *E. coli* strains against the total amount of bacterial strains per sample, using a linear regression analysis as shown in **Figure 3**.

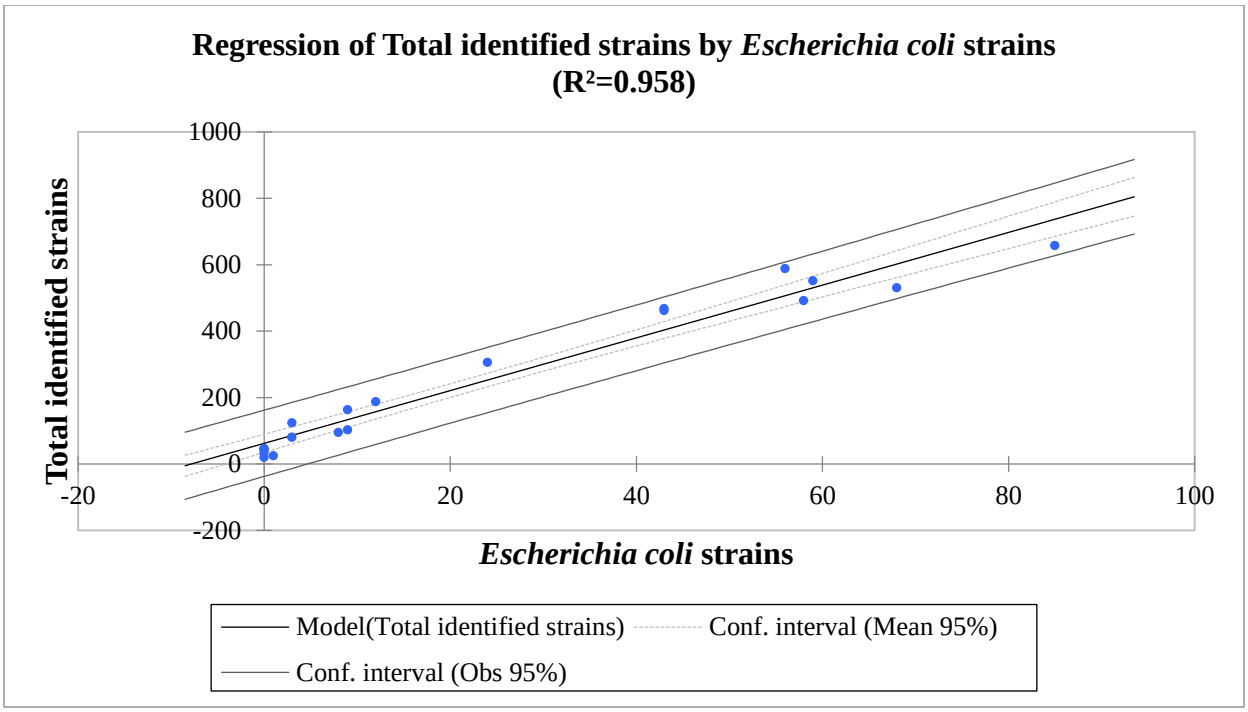


Figure 3. Linear regression analysis of *E. coli* strains (ECS) and the total number of observed strains (NOS). This regression shows the relation between the intra species diversity of Escherichia coli and the total diversity. (R^2 = 0.958, *p*-value < $1E10^{-4}$). The total number of observed species can be estimated along the Nakivubo channel using the following equation: NOS = 62.2 + 7.94*ECS.

To assess the effect of distance to channel on the bacterial composition, we performed a linear regression analysis to assess the relation between distance and taxonomic correlation throughout the Nakivubo channel (Supplementary 1A) as well as the relations between the distance to the channel and the composition of the wetland samples from cluster 1 (Supplementary 1B-1G).

c. Specificities of the environmental clusters.

Using the Lefse pipeline, we screened the GSMer profiles for bacterial strains that were significantly different in relative abundance between the three enviromental clusters. The results of this analysis are shown in **Figure 4.**

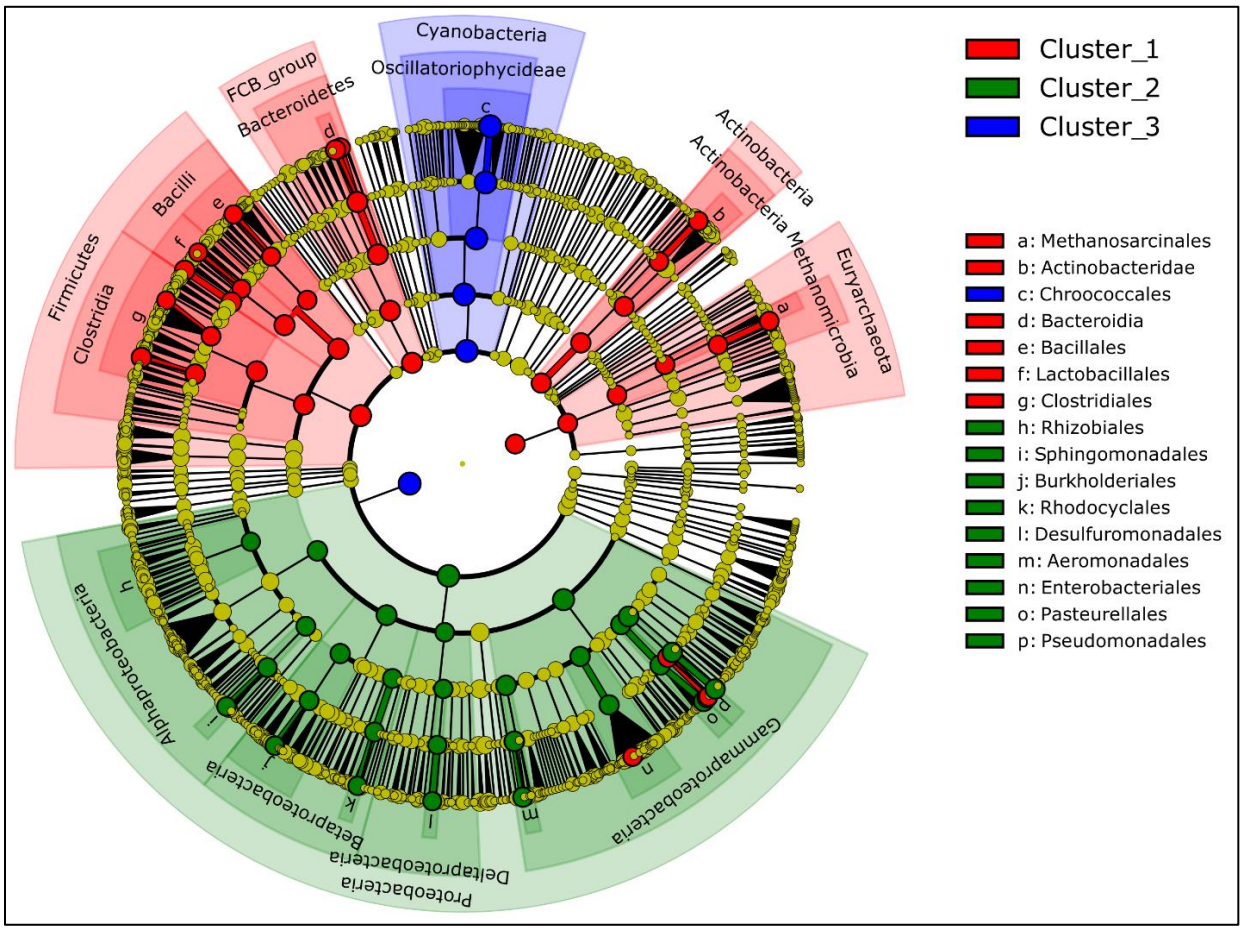


Figure 4. Cluster related-biomarkers. The cladogram shows the statistically significant differences (*p*-value < 0.01, LDA > 4.0) in abundance of bacterial taxa between the three environmental clusters. For clarity purpose, we show here only the significant differences, down to the taxonomical level of order. For complete list of identified organisms, see Supplementary Table 2. Members from the Archaea and Bacteria domains are shown here. FCB_group = *Bacteroidetes/Chlorobi* group.

The bacterial composition of the clusters harbours various specificities. All results discussed in this section which involve taxonomic levels below “order” can be consulted in Supplemental Table 2. In brief, abundance differences include gram-positive bacteria from the *Firmicutes* and *Actinobacteria* phyla, which are over-abundant in samples from cluster 1. The differences in *Clostridia* abundance include species from the *Blautia* and *Ruminococcus* genera whereas differences in the *Bacilli* include species from the *Enterococcus*, *Streptococcus* and *Exiguobacterium* genera. *Escherichia coli* and *Enhydrobacter aerosaccus* are also over-represented in cluster 1. Over-abundance of *Prevotella* and *Bacteroides* genera, within the gram-negative *Bacteroidetes* phylum and of *Methanosaeta concilii* (Phylum *Euryarcaeota*) are additional characteristics of cluster 1 samples. For cluster 2, we highlighted the overabundance of one phylum, namely the *Proteobacteria*. This overrepresentation includes members from the *Alpha-* (*Novosphingobium* spp.), *Beta-* (*Comamonas testosteroni* and *Dechloromonas aromatica*), *Delta-* (*Geobacter* spp.) and *Gammaproteobacteria* classes (*Pseudomonas*, *Acinetobacter*, *Pasteurella* and *Aeromonas* genera). Cluster 3 is characterized by the over-abundance of *Cyanobacteria*, with *Microcystis aeruginosa* being the main driver of this difference.

d. Risks associated with wastewater contamination**.**

To assess the potential risk caused by wastewater contamination on human health, we assessed the relative abundances of a set of known waterborne bacterial pathogens throughout the Nakivubo system, which we refer to as the “pathobiome” (Figure 5A). We expanded this analysis to eukaryotic parasites and viral pathogens (Figure 5B) that were identified from the comparison of *de novo* assembled contigs with the NCBI nt database.

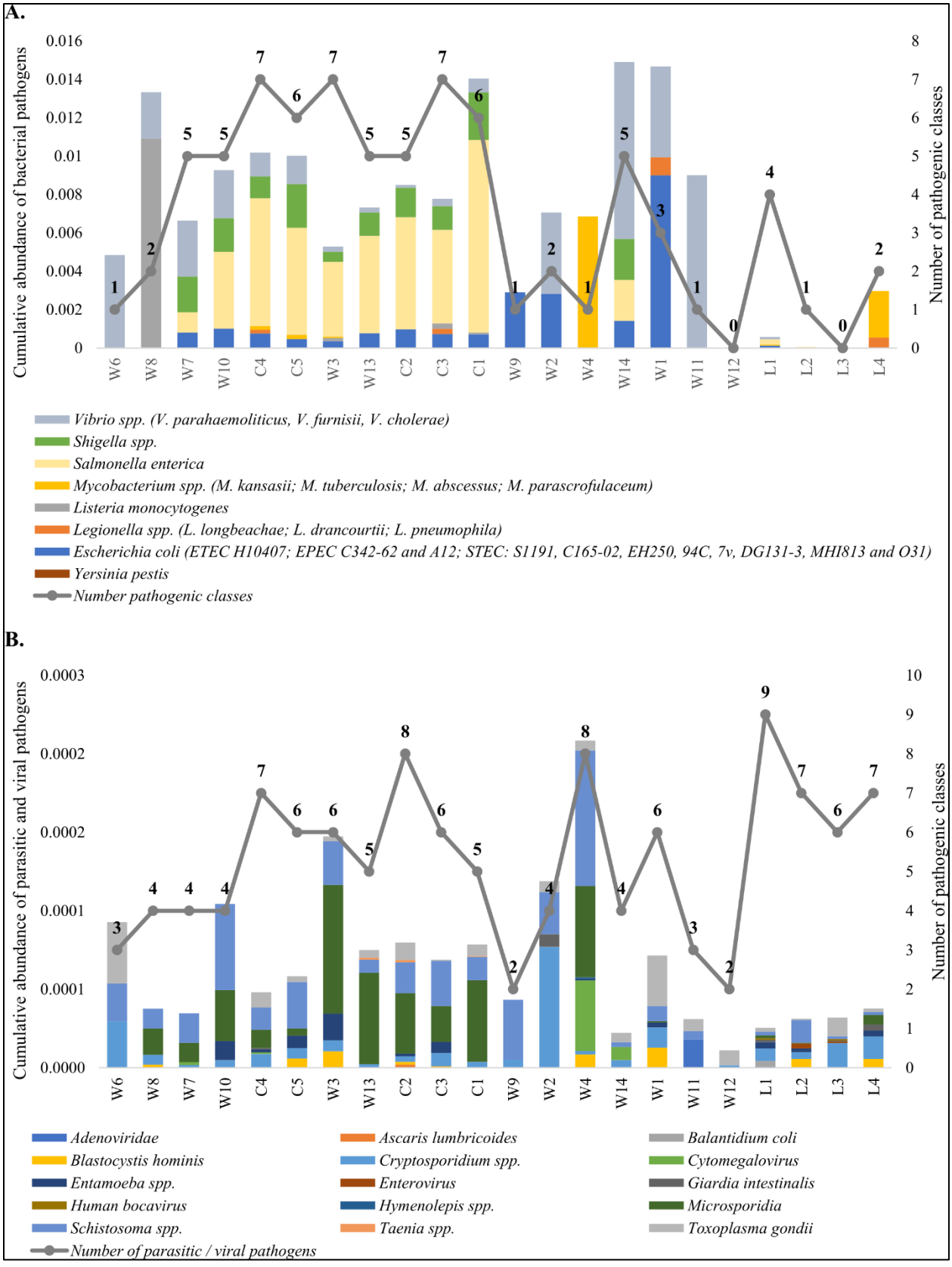

A.
Cumulative abundance of bacterial pathogens
Number of pathogenic classes
0.016
0.014
0.012
0.01
0.008
0.006
0.004
0.002
0
8
7
6
5
4
3
2
1
0
W6 W8 W7 W10 C4 C5 W3 W13 C2 C3 C1 W9 W2 W4 W14 W1 W11 W12 L1 L2 L3 L4
1 2 5 5 7 6 7 5 5 7 6 1 2 1 5 3 1 0 4 1 0 2
Vibrio spp. (V. parahaemoliticus, V. furnisii, V. cholerae)
Shigella spp.
Salmonella enterica
Mycobacterium spp. (M. kansasii; M. tuberculosis; M. abscessus; M. parascrofulaceum)
Listeria monocytogenes
Legionella spp. (L. longbeachae; L. drancourtii; L. pneumophila)
Escherichia coli (ETEC H10407; EPEC C342-62 and A12; STEC: S1191, C165-02, EH250, 94C, 7v, DG131-3, MHI813 and O31)
Yersinia pestis
Number pathogenic classes
B.
Cumulative abundance of parasitic and viral pathogens
Number of pathogenic classes
0.0003
0.0002
0.0002
0.0001
0.0001
0.0000
10
9
8
7
6
5
4
3
2
1
0
W6 W8 W7 W10 C4 C5 W3 W13 C2 C3 C1 W9 W2 W4 W14 W1 W11 W12 L1 L2 L3 L4
3 4 4 4 7 6 6 5 8 6 5 2 4 8 4 6 3 2 9 7 6 7
Adenoviridae
Ascaris lumbricoides
Balantidium coli
Blastocystis hominis
Cryptosporidium spp.
Cytomegalovirus
Entamoeba spp.
Enterovirus
Giardia intestinalis
Human bocavirus
Hymenolepis spp.
Microsporidia
Schistosoma spp.
Taenia spp.
Toxoplasma gondii
Number of parasitic / viral pathogens

Figure 5. Prevalence of important waterborne pathogens across the Nakivubo system. Panel A. Bar chart representing the cumulative abundance of bacterial pathogens at the different sampling locations. The line indicates the number of pathogenic classes for each sample. Panel B. Bar chart representing the cumulative abundances of parasitic, fungal and viral pathogens in the Nakivubo system. The line shows the number of pathogenic classes found at each sampling location. For clarity, samples are sorted per environmental cluster.

In terms of diversity, the most diverse bacterial pathobiomes are found in samples C3, C4 and W3 with seven pathogenic classes (PC) followed by C1 and C5 (PC = 6) and samples collected at locations C2, W13, W10, W7 and W14 (PC = 5). The bacterial pathobiome represents more than 1% of the total bacteria in seven samples (C1, C4, C5, W8, W10, W14, W1) and close to 1% in samples C2 (0.97%) and C3 (0.90%). Using a Kruskal-Wallis test, we tested whether the abundance of these bacterial pathogens is significantly different between the environmental clusters and found that high abundance of agents causing salmonellosis, yersinosis and shigellosis is significantly associated with cluster 1 (*p*-value = 0.018, 0.036 and 0.008, respectively).

Regarding the remaining organisms composing the pathobiome, including parasitic, viral and fungal pathogens, the most diverse pathobiome was found in L1 (PC = 9), followed by locations W4 and C2 with 8 pathogenic classes and C2, W4, L2 and L4 with 7 pathogenic classes. The highest cumulative abundances of these pathogens were found in samples obtained from the wetlands (W4>W3>W2>W10>W6). Agents responsible for microsporidiosis were significantly more abundant in cluster 1 (*p*-value = 0.003), while the mean relative abundance of *Giardia intestinalis* was higher in samples

collected in Lake Victoria (cluster 3). DNA traces of *Cryptosporidium* spp., *Schistosoma* spp. and *Toxoplasma gondii* were found homogeneously across all sampling locations with a prevalence of 95%, 95% and 81%, respectively.

5. Discussion

In this study, we used a metagenomics approach to perform an in-depth analysis of the microbial composition of water samples taken along a wastewater channel in Kampala, Uganda, in an attempt to characterize and elucidate associated health threats. To our knowledge, this is the first study to apply such an approach based on ultra-deep shotgun sequencing with subsequent strain-level characterization in a sub-Saharan African setting.

Several findings obtained during this work are worth discussing. First, the correlation of the bacterial strain profiles showed a clustering of all samples into three different groups, with one group containing all samples from Lake Victoria, while another one was not only composed of all samples collected on the Nakivubo channel, but also of some of the samples collected directly on the surrounding wetlands. The third group represented the majority of wetland samples. The heterogeneous repartition of wetland samples in different clusters clearly indicates that contamination of wastewater from the Nakivubo channel occurs throughout the wetlands, which might have important implications for human health. As the sampling period was performed during a rain-free period, we strongly believe that this contamination cannot be explained by, e.g., temporary flooding, but rather by permanent leakage around the wastewater channel.

Second, our results yield interesting patterns regarding the intra-species distribution of *E. coli* strains in the different clusters. Indeed, *E. coli* plating and counting

is used as the standard method to assess the faecal-related contamination of water, and it has been shown previously that the counts of total faecal coliforms and *E. coli* in the Nakivubo wastewater system are above the threshold recommended by WHO (means of 2.9*105 and 9.9*104 colony forming units per 100mL, respectively (Fuhrimann et al 2015)). Here, we showed that, in addition to such a quantification of *E. coli* colonies, the intra-species diversity of *E. coli* is also strongly correlated to the total bacterial diversity. The establishment of diversity thresholds based on e.g. the average cluster means of *E. coli* strains might be helpful to provide a better description of the microbiological diversity of a certain ecosystem and the arising health consequences.

Third, sampling locations C1 and C2 are located upstream and downstream respectively of the Bugolobi sewage treatment facility which is located at the latitude of 0.3182079° and longitude of 32.6070297°. The number of observed strains decreased from 899 to 745 between both samples while the Shannon diversity index decreased only slightly from 4.85 to 4.68 indicating a slight effect of the decontamination process on the microbial composition. Among this effect, we noticed the apparition of several bacterial genera related to the sludge treatment process, including but not limited to the *Aminobacterium* or *Aminomonas* genera (Baena et al 1998, Baena et al 1999) as well as several methanogens. It is worthwhile noticing that some bacterial genera including *Erysipelothrix* or *Parasutterella* that have been isolated from faecal material (Morotomi et al 2011, Wood 1974) are also introduced in the process hinting towards previous and potentially permanent contamination of the infrastructure. Shigellosis, salmonellosis, and yersinosis causing agents, among others, see their relative abundances decrease.

Finally, abundances of several bacteria including some genera commonly known to contain human pathogenic species are increasing between both locations. This specific aspect, although based on two samples, shows that the microbial treatment capacity or process of the Bugolobi sewage treatment facility is sub-optimal.

To pinpoint the potential leakage points of the Nakivubo channel, we tested whether taxonomic correlation is a function of the distance between each sampling points, along the Nakivubo channel. We showed that distance explains more than 90% of the taxonomic correlation between all samples from the channel. We further compared these two metrics between individual wetland samples that grouped with the channel samples in environmental cluster 1. For four of the points, namely W6-8 and W10, we found a strong correlation between the bacterial composition and their closest sampling points on the wastewater channel. W6-8 are spatially closest to C4 while W10 is closest to C5. This enables us to hypothesize that containment of wastewater is insufficient around locations C4 and C5 on the Nakivubo channel and that leakage happen in the wetlands around these points. This approach, combined with additional sampling points to increase resolution could help us establish an exact map of the system's weakness.

After demonstrating that the Nakivubo channel clearly impacts its immediate surroundings, and where this effect is the strongest, we aimed at characterizing the exact compositional differences between the three environmental clusters. Samples from cluster 1 are characterized by an over-abundance of bacterial strains from the *Firmicutes*, *FCB*-group, *Actinobacteria* and *Euryarchaeota* phylum which are phylum commonly found in the human gastrointestinal tract (Eckburg et al 2005, Human Microbiome Project 2012). *Escherichia coli* overabundance in samples from cluster 1 also hints towards the

same faecal origin. It is worthwhile to notice the overabundance of members from the *Proteobacteria* in cluster 2. This overabundance however doesn't include any bacteria strictly related to the gut microbiota and mainly include known environmental bacteria (Shin et al 2015). Samples collected in the inner Murchison bay are characterized by the over-abundance of *Microcystis aeruginosa*, which is commonly found in freshwater and under warm temperatures, which correlate with the context of Lake Victoria.

Fourth, when looking specifically at the bacterial pathobiome in the Nakivubo system, there is evidence that a high absolute diversity of pathogens, or number of pathogenic classes, is associated with the samples from cluster 1. There is also strong evidence that *Salmonella enterica*, *Shigella* spp. and *Yersinia pestis* contamination are directly related to wastewater, while other potential bacterial pathogens that we focused on aren't. In the case of *Mycobacterium* spp. and *Vibrio* spp., the lack of cluster specificity is probably due to the ubiquitous aspect of these bacterial species (Primm et al 2004, Raszl et al 2016, Thompson et al 2004) and a more targeted screening of, e.g. virulence factors might result in a different picture. It is worthwhile noticing that *Listeria monocytogenes* is present in 4 samples from cluster 1 but that the very low relative abundance in three of these samples isn't sufficient to statistically link the species with wastewater contamination. Legionellosis agent is present in a minority of samples from each cluster, indicating that the Nakivubo channel isn't the source of it. The last observation we made is about the heterogeneous repartition of pathogenic *E. coli* strains which speaks in favour of past contamination events, as it is present in a vast majority of samples from cluster 1 as well as in some samples from cluster 2, in a relatively high abundance. This could also be interpreted as the natural treatment function of the

wetland, with pathogenic *E. coli* being the only bacteria in our study that isn't naturally removed over time.

Regarding parasitic, fungal and viral pathogens, the situation is different, as all but two species are not specifically related to wastewater contamination by the Nakivubo channel. Microsporidia, a fungal agent that is known as an emerging opportunistic pathogen (Curry and Smith 1998, Stentiford et al 2016) causing, among others, gastrointestinal-related symptoms, is found to be strongly linked to samples from cluster 1, indicating a strong effect from the Nakivubo channel on this potentially important pathogen. While found in one sample of both clusters 1 and 2, a high relative abundance of *Giardia intestinalis* seems to be correlated with samples from cluster 3, ruling out the effect of wastewater contamination in this specific case. It is interesting to notice that three of the parasitic pathogens, namely *Schistosoma* spp., *Toxoplasma gondii* and *Cryptosporidium* spp. are found in over 80% of the samples, indicating that the Nakivubo system is potentially a strong source of contamination with these pathogens.

To conclude this study, we showed that system-wide characterization is possible, using an ultra-deep metagenomics approach and state-of-the-art bioinformatics and that it yielded in-depth insights of a complex system such as the Nakivubo system. While the resolution of the study is limited by the number of sampling locations and by the lack of temporally distinct sampling, the sequencing depth of each samples already allowed us to highlight several specificities of the Nakivubo channel, wetlands and of the inner Murchison bay. In this specific setting, we are able to draw four main conclusions, namely, i) that the system harbours, based on the microbial composition, three distinct environmental groups, ii) that the Nakivubo channel has a clear impact on the wetland

microbiota in specific locations and that it's containment potential isn't sufficient, iii) that leakage of wastewater occurs around two sampling locations, in dry periods and iv) that several potentially harmful microorganisms for human health are found to be spread by wastewater contamination.

6. Conclusions

- The bacterial composition in the wetlands is very heterogeneous with some hotspots of contamination, which indicates that some wetland areas may pose a significant health threat to humans.
- Intra-species diversity of *Escherichia coli* is proportional to the total number of observed strains. An *E. coli* diversity assay could be used to estimate the contamination status of the wetland with a lower cost and denser resolution.
- Contamination with several human pathogens around the system is associated with wastewater. Leakage points indicate that the containment potential of the Nakivubo channel is sub-optimal and poses a threat to human health.

7. Competing interests

The authors declare that they have no competing interests.

8. References

Al-Jassim N, Ansari MI, Harb M, Hong P-Y (2015). Removal of bacterial contaminants and antibiotic resistance genes by conventional wastewater treatment processes in Saudi Arabia: Is the treated wastewater safe to reuse for agricultural irrigation? *Water research* **73:** 277-290.

Andrews S FastQC A Quality Control tool for High Throughput Sequence Data. http://wwwbioinformaticsbabrahamacuk/projects/fastqc/.

Aronesty E (2011). ea-utils: Command-line tools for processing biological sequencing data. https://codegooglecom/p/ea-utils/.

Baena S, Fardeau ML, Labat M, Ollivier B, Thomas P, Garcia JL *et al* (1998). Aminobacterium colombiensegen. nov. sp. nov., an amino acid-degrading anaerobe isolated from anaerobic sludge. *Anaerobe* **4:** 241-250.

Baena S, Fardeau ML, Ollivier B, Labat M, Thomas P, Garcia JL *et al* (1999). Aminomonas paucivorans gen. nov., sp. nov., a mesophilic, anaerobic, amino-acid-utilizing bacterium. *International journal of systematic bacteriology* **49 Pt 3:** 975-982.

Bateganya NL, Nakalanzi D, Babu M, Hein T (2015). Buffering municipal wastewater pollution using urban wetlands in sub-Saharan Africa: a case of Masaka municipality, Uganda. *Environmental Technology* **36:** 2149-2160.

Becerra-Castro C, Lopes AR, Vaz-Moreira I, Silva EF, Manaia CM, Nunes OC (2015). Wastewater reuse in irrigation: A microbiological perspective on implications in soil fertility and human and environmental health. *Environment international* **75:** 117-135.

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J *et al* (2013). GenBank. *Nucleic acids research* **41:** D36-42.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K *et al* (2009). BLAST+: architecture and applications. *BMC bioinformatics* **10:** 421.

Chevreux B (2007). MIRA: an automated genome and EST assembler.

Curry A, Smith HV (1998). Emerging pathogens: Isospora, Cyclospora and microsporidia. *Parasitology* **117 Suppl:** S143-159.

Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M *et al* (2005). Diversity of the human intestinal microbial flora. *Science* **308:** 1635-1638.

Emerton L, Iyango L, Luwum P, Malinga A (1999). The present economic value of Nakivubo urban wetland, Uganda. *IUCN—The World Conservation Union, Eastern Africa Regional Office, Nairobi and National Wetlands Programme, Wetlands Inspectorate Division, Ministry of Water, Land and Enviornment, Kampala*.

Fuhrimann S, Winkler MS, Schneeberger PH, Niwagaba CB, Buwule J, Babu M *et al* (2014). Health risk assessment along the wastewater and faecal sludge management and reuse chain of Kampala, Uganda: a visualization. *Geospatial health* **9:** 241-245.

Fuhrimann S, Stalder M, Winkler MS, Niwagaba CB, Babu M, Masaba G *et al* (2015). Microbial and chemical contamination of water, sediment and soil in the Nakivubo wetland area in Kampala, Uganda. *Environmental Monitoring and Assessment* **187:** 1-15.

Human Microbiome Project C (2012). Structure, function and diversity of the healthy human microbiome. *Nature* **486:** 207-214.

Kansiime F, Maimuna N (1999). *Wastewater treatment by a natural wetland: the Nakivubo swamp, Uganda*, vol. 21. CRC Press.

Kayima J, Kyakula M, Komakech W, Echimu SP (2008). A study of the degree of Pollution in Nakivubo Channel, Kampala, Uganda. *Journal of Applied Sciences and Environmental Management* **12**.

Lu Y, Song S, Wang R, Liu Z, Meng J, Sweetman AJ *et al* (2015). Impacts of soil and water pollution on food safety and health risks in China. *Environment international* **77:** 5-15.

Mbabazi J, Kwetegyeka J, Ntale M, Wasswa J (2010). Ineffectiveness of Nakivubo wetland in filtering out heavy metals from untreated Kampala urban effluent prior to discharge into Lake Victoria, Uganda. *African Journal of Agricultural Research* **5:** 3431-3439.

Morotomi M, Nagai F, Watanabe Y (2011). Parasutterella secunda sp. nov., isolated from human faeces and proposal of Sutterellaceae fam. nov. in the order Burkholderiales. *International journal of systematic and evolutionary microbiology* **61:** 637-643.

Okuonzi SA (2004). Dying for economic growth? Evidence of a flawed economic policy in Uganda. *the Lancet* **364:** 1632-1637.

Primm TP, Lucero CA, Falkinham JO (2004). Health Impacts of Environmental Mycobacteria. *Clinical Microbiology Reviews* **17:** 98-106.

Raszl SM, Froelich BA, Vieira CR, Blackwood AD, Noble RT (2016). Vibrio parahaemolyticus and Vibrio vulnificus in South America: water, seafood and human infections. *Journal of applied microbiology* **121:** 1201-1222.

Shin N-R, Whon TW, Bae J-W (2015). Proteobacteria: microbial signature of dysbiosis in gut microbiota. *Trends in biotechnology* **33:** 496-503.

Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C *et al* (2002). The Bioperl toolkit: Perl modules for the life sciences. *Genome research* **12:** 1611-1618.

Statistics UBO (2001). Statistical Abstract. *Uganda Bureau of Statistics*.

Stentiford GD, Becnel JJ, Weiss LM, Keeling PJ, Didier ES, Williams BAP *et al* (2016). Microsporidia – Emergent Pathogens in the Global Food Chain. *Trends in parasitology* **32:** 336-348.

Thompson FL, Iida T, Swings J (2004). Biodiversity of Vibrios. *Microbiology and Molecular Biology Reviews* **68:** 403-431.

Tu Q, He Z, Zhou J (2014). Strain/species identification in metagenomes using genome-specific markers. *Nucleic acids research* **42:** e67.

Vermeiren K, Van Rompaey A, Loopmans M, Serwajja E, Mukwaya P (2012). Urban growth of Kampala, Uganda: Pattern analysis and scenario development. *Landscape and Urban Planning* **106:** 199-206.

WHO (2011). Guidelines for Drinking-water Quality, fourth edition. *WHO chronicle* **38:** 104-108.

Wood RL (1974). Isolation of pathogenic Erysipelothrix rhusiopathiae from feces of apparently healthy swine. *American journal of veterinary research* **35:** 41-43.

Youenou B, Hien E, Deredjian A, Brothier E, Favre-Bonté S, Nazaret S (2016). Impact of untreated urban waste on the prevalence and antibiotic resistance profiles of human opportunistic pathogens in agricultural soils from Burkina Faso. *Environmental Science and Pollution Research* **23:** 25299-25311.

Chapter VI. Discussion and perspectives

1. Impact of NGS on the field of infectious diseases research

Next-generation sequencing, due to its naïve and relatively unbiased nature, revolutionized how studies in the field of infectious diseases are conducted. Major examples of how these new approaches influenced the field of infectious diseases include, but are not limited to; (i) the unravelling of new insights into the microbial genetic diversity (Cirulli and Goldstein 2010, Prosperi et al 2011); (ii) the discovery of new pathogenic microorganisms previously unknown (Mardis 2009, Palacios et al 2008); and (iii) highlighting worsened symptomatic expression of diseases caused by coinfection events as well as gaining insights into dynamics of microbiomes or other complex microbial communities (Cho and Blaser 2012, David et al 2014, Koenig et al 2011, Salipante et al 2014). This PhD thesis sought to apply these techniques within several complex situations, extending their application beyond the established frontier, and further developing the optimized usage of these types of studies.

a. A bioinformatics tool to improve accuracy and specificity of molecular assays

The first part of this thesis was aimed at developing a bioinformatics workflow that would adapt to the challenge caused by the tremendous amount of nucleotide sequences made available since NGS was brought on the market. In fact, while increased amounts of sequences have offered decisive insights in research on infectious diseases (Capobianchi et al 2013, Radford et al 2012), they are also accompanied with challenges, one of them being the lack of bioinformatics approaches able to process such datasets. We developed a bioinformatics workflow that allowed the selection of highly conserved

and specific molecular markers, and we validated it by testing different types of molecular assays, including real-time PCR, LAMP and Sanger sequencing assays, using the selected markers. The study focused on thirteen neglected viral pathogens from the *Flaviviridae* and *Bunyaviridae* families that pose a serious threat to human health and were responsible for several outbreaks (Aradaib et al 2011, Balogh et al 2010, Lanciotti et al 1999, Lvov et al 2000, Spinsanti et al 2008, Woods et al 2002). This approach, based on the widely available BLAST algorithm enabled the selection of these markers among several hundreds of complete genomic sequences for the most extensively sequenced viral species used in this study. This study confirms the usefulness of this tremendous amount of sequencing data, provided suitable analysis workflows are available, to improve accuracy and specificity of molecular diagnostics.

b. Identification of a new virus from a complex plant microbiome

The second part of this thesis was focusing on the discovery and molecular characterisation of the microbial organism causing leafroll symptoms in a *Vitis vinifera* cv. Otcha bala plant. This example showed how a metagenomics approach could successfully complement standard diagnostic tools. In this case, a preliminary analysis using electron microscopy allowed the determination of the origin of the causative agent, which was in fact an unidentifiable virus. Leafroll disease in grapevine has a complex aetiology since various viral species were usually associated with this disease. The viral metagenome present in this sample could be completely characterised using 454 sequencing and an in-house developed bioinformatics pipeline. Complete or partial genomes from four viral species could be reconstructed, namely (i) the Grapevine fleck virus and the Grapevine red globe virus that are both members of the *Tymoviridae* family,

(ii) the *Grapevine virus A* from the *Betaflexviridae* family; and (iii) a thirteen Kb long contig closely related to *Closteroviridae*, that we proposed to name *Grapevine leafroll-associated virus 4* strain Ob (GLRaV-4 Ob). Since none of the three first viruses were known to cause symptoms in Grapevine, and with the confirmation obtained by additional serological tests, we were able to demonstrate that this new viral variant was in fact the causative agent of the Grapevine disease. This study confirmed, that omics approaches are useful to determine aetiology of a disease that could not be identified and/or fully characterised with other tools.

c. Metagenomics and its application in personalized medicine

Microbiome characterisation is a powerful tool and one pillar of the rapidly growing field of personalized medicine (Collins 2010, Isaacs and Ferraccioli 2011, Nicholson et al 2005, Nicholson 2006, Tsai and Coyle 2009). The third part of this thesis focused on assessing the potential of a metagenomics approach for the characterisation of the gut pathobiome (microbiome restricted to pathogenic microorganisms) and for the generation of additional individual health-related information. In this proof-of-concept study, stool samples from four patients presenting persistent digestive disorders were screened for a wide range of pathogenic microorganisms. We were able to demonstrate the consistency of this approach for the diagnosis of pathogens associated with this digestive syndrome by comparing it with a set of validated diagnostic approaches, including microscopy, rapid diagnostics tests and multiplex PCR assays. The detection rate of bacterial pathogens and helminths was in favour of the metagenomics approach, which permitted the identification of a wider range of species belonging to these pathogenic classes, including *Mycobacterium* spp. and *Schistosoma mansoni*. However, for viruses and intestinal

protozoa, the detection rate was more in favour of the standard tools, mainly for two reasons, namely, (i) because the sequencing depth was not sufficient to detect viral sequencing reads in this complex sample; and (ii) because there is a lack of sequence data for protozoan species. These two issues are only temporary, since sequencing technologies are permanently improved and are expected to provide a higher sequencing depth making even the rarest organisms clearly identifiable. Also, current protozoa sequencing projects will provide additional sequence data for these species, closing up the sequencing gap between these and other microorganisms for which more complete genomic sequences are available, like helminths (Berriman et al 2009, Park et al 2011, Young et al 2012, Zhou et al 2009). In addition to the confirmation of the diagnostic potential of this metagenomics approach, we also generated additional health-related information by screening the sequence datasets for antimicrobial resistance genes. While this approach has several limitations, mainly bioinformatics-wise due to the diversity of mechanisms involved in antibiotics resistance in bacteria, it provides important information about the resistome and potential resistances that are present or could quickly spread in an environment as complex as the human gut microbiome.

d. Wastewater microbiota and its impact on human health

The final part of this PhD thesis was to apply a metagenomics approach to enable a system-wide microbial survey and to assess the potential risk on human health. In this study, we were able to conduct a complete survey of microbial communities present in the wastewater network from the city of Kampala, Uganda. We demonstrated how closely the diversity of *E. coli*, which is a standard indicator of faecal contamination recommended by the WHO, was linked to the overall diversity of bacteria in different aquatic ecosystems.

This is an important result for future studies that could then specifically concentrate on analysing the diversity of *E. coli* in order to assess the impact of human activity on the bacterial communities in the surrounding environment. Analysing the overall bacterial diversity confirmed the pertinence of our three environmental groups, by clearly generating three distinct sample clusters. Using assembled sequences from the 22 metagenomics datasets also allowed us to characterise the geographical repartition of the different pathogenic classes among the Nakivubo wetland. Similarly, we were also able to summarise the main waterborne pathogens causing symptoms in humans and how they spread through this ecosystem using this metagenomics approach.

In addition to this environmental assessment, this project included a second part with the aim of investigating the impact of wastewater exposure on the human gut microbiome. For this purpose, 114 stool samples were collected from different population groups around Kampala. We selected three population groups that we expected to be at different level of exposure to wastewater from the Nakivubo channel and wetlands. This included 38 samples (S001-S038) from what we described as the high exposure group. The individuals in this group were directly exposed to water contact on a daily basis due to their farming activities in the Nakivubo wetlands. Sample S039 to sample S078 were collected from individuals living in the slum areas surrounding the Nakivubo swamps, at occasional risk of exposure to wastewater, mainly because of occasional flooding events. The last group, including samples S079 to S114 originated from a control population, rarely or never coming into contact with the water from the Nakivubo channel/wetlands. A first result of the ongoing analyses is shown in **Figure 1** where we were able to show how the bacterial composition of the environmental samples impacted the gut bacterial

microbiomes of the surrounding populations. With this analysis, we were able to demonstrate that substantial differences in the bacterial communities exist between the three population groups from this study. In fact, 15 out of 18 stool samples clustering together with the environmental samples were from Group 1 suggesting that the gut microbiomes of this group, in which individuals are exposed to water on a daily basis, are deeply impacted by their exposure to wastewater. This is only a snapshot of this study and we hope that the aggregation of these data with the answers collected using health questionnaires will bring further insights in the impact of wastewater management in Kampala on human health.

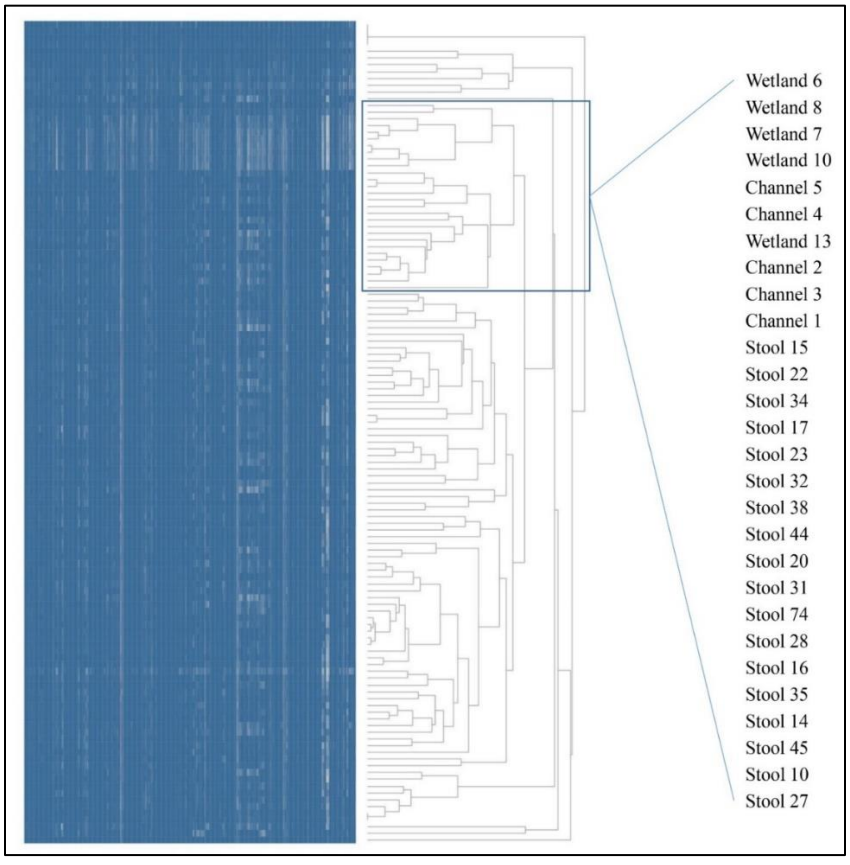


Figure 1. Hierarchical clustering of the bacterial communities from both environmental and human samples. A blue gradient indicates the number of markers found for each bacterial strain. The lighter the shade, the more markers were found.

2. Future of omics approaches and associated challenges

a. The future of NGS

Next-generation sequencing is rapidly pervading all areas of infectious diseases: improving speed, precision, and, last but not least, the breadth of diagnostics. This statement is even truer since sequencing technologies are constantly evolving and becoming more straightforward and easy to apply to different projects. The field of single molecule sequencing through a nanopore is one of the area of NGS which holds most promises and interesting applications (Bayley 2015, Feng et al 2015, Garaj 2014, Jones 2015). Recently, a number of studies involving nanopore sequencing have shown that the technology, while it has yet to be optimized, already shows great potential (Kilianski et al 2015, Loman and Watson 2015). It is predicted that the omics applications of NGS presented in this thesis will be greatly improved by the use of these new nanopore sequencing technologies. For instance, studies show early promising results in the screening of antimicrobial resistance genes using this technology (Ashton et al 2014, Judge et al 2015)

The other great improvement lies in the miniaturisation of the sequencing devices. Until now, NGS instruments could only be used in an equipped laboratory and some, like the Pacific Biosciences RSII instrument, even required a specific room due to its massive size. The nanopore sequencing instrument from Oxford Nanopore, the MinION, has the size of an usb stick and protocols are currently being optimized to allow the direct deposit of a sample without pre-purification steps. This improvement will allow to bring this handheld device directly to the patient as a point-of-care diagnostics device, allowing

cheaper, faster yet more accurate results on-site. In a closer future, we can expect that the technical characteristics of current NGS technologies will be continuously improved. Current applications, like metagenomics or metatranscriptomics will greatly benefit from increased sequencing read-length and deeper coverage for a lower price.

b. Associated bioinformatics challenges

Besides many insights, the expected data deluge also brings questions, especially in the area of bioinformatics. This includes challenges with handling, processing and moving information, challenges that were historically reserved for astronomers and high-energy physicists (Marx 2013). Biologists now have to store, analyse, compare and share massive amount of sequencing data – which is not a simple task when a single sequenced human genome is already 140 gigabytes in size (Marx 2013, Stephens et al 2015). Increased computing resources, including additional and faster storage, as well as additional computing cores are nowadays a requirement for any laboratory wanting to embark in the omics field. There is also a need to find consensuses on the software side, with the need of standardised bioinformatics workflows for applications involving NGS. All in all, bioinformatics should, with specific educational programs for future scientists, now benefit from the same attention and development pace as NGS technologies.

3. General conclusion

Omics approaches, facilitated by the advancements of NGS technologies, have revolutionised the way research is conducted in the field of infectious diseases. Many

challenges that were mainly due to the vast diversity of pathogenic microorganisms can now be approached differently. There are many examples of applications improving research in infectious diseases. This includes the field of genomics, which by multiplying the number of sequenced genomes by a factor of over 1000 between the years 2000 and 2015 (Stephens et al 2015), permitted further insights in the genetic diversity of many pathogens. Meta-analyses, like metagenomics or metatranscriptomics are the applications that benefited most from the use of NGS technologies, and they now allow system-wide studies, where previous studies were only focusing on one parameter (one microbe or one specific gene for instance). However, these omics approaches have their limitations, mainly due to the bioinformatics challenges they give rise to. In conclusion, it is foreseeable that these approaches, once matured, due to the increased amount of results they allow to generate, will be widely used and will replace standard approaches in the field of infectious diseases.

4. References (chapters 1 and 6)

Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P (2002). *Molecular Biology of the Cell, Fourth Edition*. Garland Science.

Alter MJ (2006). Epidemiology of viral hepatitis and HIV co-infection. *Journal of hepatology* **44:** S6-S9.

Anderson RM, May RM, Anderson B (1992). *Infectious diseases of humans: dynamics and control*, vol. 28. Wiley Online Library.

Antonello SD (2007). *Frontiers in ecology research*. Nova Publishers.

Aradaib IE, Erickson BR, Karsany MS, Khristova ML, Elageb RM, Mohamed ME *et al* (2011). Multiple Crimean-Congo hemorrhagic fever virus strains are associated with disease outbreaks in Sudan, 2008–2009.

Ashton PM, Nair S, Dallman T, Rubino S, Rabsch W, Mwaigwisya S *et al* (2014). MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nature biotechnology*.

Azhar EI, El-Kafrawy SA, Farraj SA, Hassan AM, Al-Saeed MS, Hashem AM *et al* (2014). Evidence for camel-to-human transmission of MERS coronavirus. *New England Journal of Medicine* **370:** 2499-2505.

Baize S, Pannetier D, Oestereich L, Rieger T, Koivogui L, Magassouba NF *et al* (2014). Emergence of Zaire Ebola virus disease in Guinea. *New England Journal of Medicine* **371:** 1418-1425.

Balogh Z, Ferenczi E, Szeles K, Stefanoff P, Gut W, Szomor KN *et al* (2010). Tick-borne encephalitis outbreak in Hungary due to consumption of raw goat milk. *Journal of virological methods* **163:** 481-485.

Bayley H (2015). Nanopore Sequencing: From imagination to reality. *Clinical chemistry* **61:** 25-31.

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J *et al* (2013). GenBank. *Nucleic acids research* **41:** D36-42.

Bera B, Shanmugasundaram K, Barua S, Venkatesan G, Virmani N, Riyesh T *et al* (2011). Zoonotic cases of camelpox infection in India. *Veterinary microbiology* **152:** 29-38.

Berriman M, Haas BJ, LoVerde PT, Wilson RA, Dillon GP, Cerqueira GC *et al* (2009). The genome of the blood fluke *Schistosoma mansoni*. *Nature* **460:** 352-358.

Bessen DE, McShan WM, Nguyen SV, Shetty A, Agrawal S, Tettelin H (2014). Molecular epidemiology and genomics of group A *Streptococcus*. *Infection, Genetics and Evolution*.

Blua M, Phillips P, Redak R (1999). A new sharpshooter threatens both crops and ornamentals. *California Agriculture* **53:** 22-25.

Breitbart M, Rohwer F (2005). Here a virus, there a virus, everywhere the same virus? *Trends in microbiology* **13:** 278-284.

Brindley PJ, Mitreva M, Ghedin E, Lustigman S (2009). Helminth genomics: The implications for human health. *PLoS neglected tropical diseases* **3:** e538.

Calisher CH, Childs JE, Field HE, Holmes KV, Schountz T (2006). Bats: Important Reservoir Hosts of Emerging Viruses. *Clinical Microbiology Reviews* **19:** 531-545.

Camp C, Tatum OL (2010). A review of *Acinetobacter baumannii* as a highly successful pathogen in times of war. *Lab Medicine* **41:** 649-657.

Capobianchi M, Giombini E, Rozera G (2013). Next-generation sequencing technology in clinical virology. *Clinical Microbiology and Infection* **19:** 15-22.

Casey JL, Niro GA, Engle RE, Vega A, Gomez H, McCarthy M *et al* (1996). Hepatitis B virus (HBV)/hepatitis D virus (HDV) coinfection in outbreaks of acute hepatitis in the Peruvian Amazon basin: the roles of HDV genotype III and HBV genotype F. *Journal of Infectious Diseases* **174:** 920-926.

Cénit M, Matzaraki V, Tigchelaar E, Zhernakova A (2014). Rapidly expanding knowledge on the role of the gut microbiome in health and disease. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease* **1842:** 1981-1992.

Chang D-E, Smalley DJ, Tucker DL, Leatham MP, Norris WE, Stevenson SJ *et al* (2004). Carbon nutrition of *Escherichia coli* in the mouse intestine. *Proceedings of the National Academy of Sciences of the United States of America* **101:** 7427-7432.

Chen L, Yang J, Yu J, Yao Z, Sun L, Shen Y *et al* (2005). VFDB: a reference database for bacterial virulence factors. *Nucleic acids research* **33:** D325-D328.

Chertow DS, Memoli MJ (2013). Bacterial coinfection in influenza: a grand rounds review. *Jama* **309:** 275-282.

Cho I, Blaser MJ (2012). The human microbiome: at the interface of health and disease. *Nature Reviews Genetics* **13:** 260-270.

Cirulli ET, Goldstein DB (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews Genetics* **11:** 415-425.

Collins F (2010). *The language of life: DNA and the revolution in personalised medicine*. Profile Books.

Cox F (2001). Concomitant infections, parasites and immune responses. *Parasitology* **122:** S23-S38.

Daszak P, Cunningham AA, Hyatt AD (2000). Emerging infectious diseases of wildlife--threats to biodiversity and human health. *Science* **287:** 443-449.

David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE *et al* (2014). Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505:** 559-563.

Depledge DP, Kundu S, Jensen NJ, Gray ER, Jones M, Steinberg S *et al* (2014). Deep sequencing of viral genomes provides insight into the evolution and pathogenesis of varicella zoster virus and its vaccine in humans. *Molecular biology and evolution* **31:** 397-409.

Dressman D, Yan H, Traverso G, Kinzler KW, Vogelstein B (2003). Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proceedings of the National Academy of Sciences* **100:** 8817-8822.

Droege M, Hill B (2008). The Genome Sequencer FLX™ System—Longer reads, more applications, straight forward bioinformatics and more complete data sets. *Journal of biotechnology* **136:** 3-10.

Edwards RA, Rohwer F (2005). Viral metagenomics. *Nature Reviews Microbiology* **3:** 504-510.

Ekblom R, Galindo J (2011). Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity* **107:** 1-15.

Ersfeld K (2003). Genomes and genome projects of protozoan parasites. *Current issues in molecular biology* **5:** 61-74.

Feero WG, Guttmacher AE, Relman DA (2011). Microbial genomics and infectious diseases. *New England Journal of Medicine* **365:** 347-357.

Feero WG, Guttmacher AE (2014). Genomics, personalized medicine, and pediatrics. *Academic pediatrics* **14:** 14-22.

Feng Y, Zhang Y, Ying C, Wang D, Du C (2015). Nanopore-based Fourth-generation DNA Sequencing Technology. *Genomics, proteomics & bioinformatics* **13:** 4-16.

Fernald GH, Capriotti E, Daneshjou R, Karczewski KJ, Altman RB (2011). Bioinformatics challenges for personalized medicine. *Bioinformatics* **27:** 1741-1748.

Fischbach FT, Dunning MB (2009). *A manual of laboratory and diagnostic tests*. Lippincott Williams & Wilkins.

Fournier PE, Richet H, Weinstein RA (2006). The epidemiology and control of *Acinetobacter baumannii* in health care facilities. *Clinical infectious diseases* **42:** 692-699.

Garaj S (2014). Nucleic Acid Sequencing and Analysis with Nanopores. *Nucleic Acid Nanotechnology*. Springer. pp 287-303.

Garcia-Solache MA, Casadevall A (2010). Global warming will bring new fungal diseases for mammals. *MBio* **1:** e00061-00010.

Han K, Li ZF, Peng R, Zhu LP, Zhou T, Wang LG *et al* (2013). Extraordinary expansion of a *Sorangium cellulosum* genome from an alkaline milieu. *Scientific reports* **3:** 2101.

Handelsman J (2004). Metagenomics: application of genomics to uncultured microorganisms. *Microbiology and molecular biology reviews* **68:** 669-685.

Hawksworth D (2001). The magnitude of fungal diversity: the 1· 5 million species estimate revisited. *Mycological research* **105:** 1422-1432.

Henn MR, Boutwell CL, Charlebois P, Lennon NJ, Power KA, Macalalad AR *et al* (2012). Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. *PLoS Pathog* **8:** e1002529.

Hilleman MR (2004). Strategies and mechanisms for host and pathogen survival in acute and persistent viral infections. *Proceedings of the National Academy of Sciences* **101:** 14560-14566.

Hingley-Wilson SM, Sambandamurthy VK, Jacobs WR (2003). Survival perspectives from the world's most successful pathogen, *Mycobacterium tuberculosis*. *Nature immunology* **4:** 949-955.

Hopkins D (1989). *Xylella fastidiosa*: xylem-limited bacterial pathogen of plants. *Annual review of phytopathology* **27:** 271-290.

Hotez PJ, Brindley PJ, Bethony JM, King CH, Pearce EJ, Jacobson J (2008). Helminth infections: the great neglected tropical diseases. *Journal of Clinical Investigation* **118:** 1311-1321.

Human Microbiome Project C (2012). Structure, function and diversity of the healthy human microbiome. *Nature* **486:** 207-214.

Ingelman-Sundberg M (2015). Personalized medicine into the next generation. *Journal of internal medicine* **277:** 152-154.

Isaacs JD, Ferraccioli G (2011). The need for personalised medicine for rheumatoid arthritis. *Annals of the rheumatic diseases* **70:** 4-7.

Isolauri E, Sütas Y, Kankaanpää P, Arvilommi H, Salminen S (2001). Probiotics: effects on immunity. *The American journal of clinical nutrition* **73:** 444s-450s.

Jacobson R (1998). Validation of serological assays for diagnosis of infectious diseases. *Revue scientifique et technique (International Office of Epizootics)* **17:** 469-526.

Jones B (2015). Technology: Nanopore sequencing for clinical diagnostics. *Nature Reviews Genetics* **16:** 68-68.

Jordan P, Webbe G (1969). Human schistosomiasis. *Human schistosomiasis*.

Judge K, Harris SR, Reuter S, Parkhill J, Peacock SJ (2015). Early insights into the potential of the Oxford Nanopore MinION for antimicrobial resistance gene detection.

Kaper JB, Nataro JP, Mobley HL (2004). Pathogenic *Escherichia coli*. *Nature Reviews Microbiology* **2:** 123-140.

Kiesslich D, Crispim MA, Santos C, Ferreira FL, Fraiji NA, Komninakis SV *et al* (2009). Influence of hepatitis B virus (HBV) genotype on the clinical course of disease in patients coinfected with HBV and hepatitis delta virus. *The Journal of infectious diseases* **199:** 1608-1611.

Kilianski A, Haas JL, Corriveau EJ, Liem AT, Willis KL, Kadavy DR *et al* (2015). Bacterial and viral identification and differentiation by amplicon sequencing on the MinION nanopore sequencer. *GigaScience* **4:** 12.

Koenig JE, Spor A, Scalfone N, Fricker AD, Stombaugh J, Knight R *et al* (2011). Succession of microbial consortia in the developing infant gut microbiome. *Proceedings of the National Academy of Sciences* **108:** 4578-4585.

Koonin EV, Senkevich TG, Dolja VV (2006). The ancient Virus World and evolution of cells. *Biol Direct* **1:** 29.

Kruis W, Frič P, Pokrotnieks J, Lukáš M, Fixa B, Kaščák M *et al* (2004). Maintaining remission of ulcerative colitis with the probiotic *Escherichia coli* Nissle 1917 is as effective as with standard mesalazine. *Gut* **53:** 1617-1623.

Lanciotti R, Roehrig J, Deubel V, Smith J, Parker M, Steele K *et al* (1999). Origin of the West Nile virus responsible for an outbreak of encephalitis in the northeastern United States. *Science* **286:** 2333-2337.

Lee C-Y, Chiu Y-C, Wang L-B, Kuo Y-L, Chuang EY, Lai L-C *et al* (2013). Common applications of next-generation sequencing technologies in genomic research. *Translational Cancer Research* **2:** 33-45.

Leland DS, Ginocchio CC (2007). Role of cell culture for virus detection in the age of technology. *Clinical Microbiology Reviews* **20:** 49-78.

Leplae R, Hebrant A, Wodak SJ, Toussaint A (2004). ACLAME: A CLAssification of Mobile genetic Elements. *Nucleic acids research* **32:** D45-D49.

Leroy EM, Kumulungui B, Pourrut X, Rouquet P, Hassanin A, Yaba P *et al* (2005). Fruit bats as reservoirs of Ebola virus. *Nature* **438:** 575-576.

Loman NJ, Watson M (2015). Successful test launch for nanopore sequencing. *Nature methods* **12:** 303-304.

Lustigman S, Prichard RK, Gazzinelli A, Grant WN, Boatin BA, McCarthy JS *et al* (2012). A research agenda for helminth diseases of humans: the problem of helminthiases. *PLoS neglected tropical diseases* **6:** e1582.

Lvov D, Butenko A, Gromashevsky V, Larichev VP, Gaidamovich SY, Vyshemirsky O *et al* (2000). Isolation of two strains of West Nile virus during an outbreak in southern Russia, 1999. *Emerging infectious diseases* **6:** 373.

Mabey D, Peeling RW, Ustianowski A, Perkins MD (2004). Tropical infectious diseases: diagnostics for the developing world. *Nature Reviews Microbiology* **2:** 231-240.

Mah T-FC, O'Toole GA (2001). Mechanisms of biofilm resistance to antimicrobial agents. *Trends in microbiology* **9:** 34-39.

Mancini N, Carletti S, Ghidoli N, Cichero P, Burioni R, Clementi M (2010). The Era of Molecular and Other Non-Culture-Based Methods in Diagnosis of Sepsis. *Clinical Microbiology Reviews* **23:** 235-251.

Manges AR, Johnson JR, Foxman B, O'Bryan TT, Fullerton KE, Riley LW (2001). Widespread distribution of urinary tract infections caused by a multidrug-resistant *Escherichia coli* clonal group. *New England Journal of Medicine* **345:** 1007-1013.

Mardis ER (2008). The impact of next-generation sequencing technology on genetics. *Trends in genetics* **24:** 133-141.

Mardis ER (2009). New strategies and emerging technologies for massively parallel sequencing: applications in medical research. *Genome medicine* **1:** 40.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA *et al* (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437:** 376-380.

Martínez JL (2008). Antibiotics and antibiotic resistance genes in natural environments. *Science* **321:** 365-367.

Marx V (2013). Biology: The big challenges of big data. *Nature* **498:** 255-260.

McArthur AG, Waglechner N, Nizam F, Yan A, Azad MA, Baylay AJ *et al* (2013). The comprehensive antibiotic resistance database. *Antimicrobial agents and chemotherapy* **57:** 3348-3357.

McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT (2013). Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution* **66:** 526-538.

McCutcheon JP, Moran NA (2012). Extreme genome reduction in symbiotic bacteria. *Nature reviews microbiology* **10:** 13-26.

Metzker ML (2010). Sequencing technologies—the next generation. *Nature reviews genetics* **11:** 31-46.

Mizell RF, Andersen PC, Tipping C, Brodbeck B (2003). *Xylella fastidiosa diseases and their leafhopper vectors*. University of Florida Cooperative Extension Service, Institute of Food and Agricultural Sciences, EDIS.

Mohammed MH, Dutta A, Bose T, Chadaram S, Mande SS (2012). DELIMINATE—a fast and efficient method for loss-less compression of genomic sequences Sequence analysis. *Bioinformatics* **28:** 2527-2529.

Moore JH, Asselbergs FW, Williams SM (2010). Bioinformatics challenges for genome-wide association studies. *Bioinformatics* **26:** 445-455.

Nicholson JK, Holmes E, Wilson ID (2005). Gut microorganisms, mammalian metabolism and personalized health care. *Nature Reviews Microbiology* **3:** 431-438.

Nicholson JK (2006). Global systems biology, personalized medicine and molecular epidemiology. *Molecular systems biology* **2:** 52.

Palacios G, Druce J, Du L, Tran T, Birch C, Briese T *et al* (2008). A new arenavirus in a cluster of fatal transplant-associated diseases. *New England journal of medicine* **358:** 991-998.

Park YC, Kim W, Park J-K (2011). The complete mitochondrial genome of human parasitic roundworm, Ascaris lumbricoides. *Mitochondrial DNA* **22:** 91-93.

Parte AC (2014). LPSN—list of prokaryotic names with standing in nomenclature. *Nucleic acids research* **42:** D613-D616.

Pawlowski A, Jansson M, Skold M, Rottenberg ME, Kallenius G (2012). Tuberculosis and HIV co-infection. *PLoS Pathog* **8:** e1002464.

Petney TN, Andrews RH (1998). Multiparasite communities in animals and humans: frequency, structure and pathogenic significance. *International journal for parasitology* **28:** 377-393.

Pop M, Salzberg SL (2008). Bioinformatics challenges of new sequencing technology. *Trends in Genetics* **24:** 142-149.

Prosperi MC, Prosperi L, Bruselles A, Abbate I, Rozera G, Vincenti D *et al* (2011). Combinatorial analysis and algorithms for quasispecies reconstruction using next-generation sequencing. *BMC bioinformatics* **12:** 5.

Pullan R, Brooker S (2008). The health impact of polyparasitism in humans: are we under-estimating the burden of parasitic diseases? *Parasitology* **135:** 783-794.

Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F *et al* (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490:** 55-60.

Radford AD, Chapman D, Dixon L, Chantrey J, Darby AC, Hall N (2012). Application of next-generation sequencing technologies in virology. *Journal of General Virology* **93:** 1853-1868.

Raj VS, Farag EA, Reusken CB, Lamers MM, Pas SD, Voermans J *et al* (2014). Isolation of MERS coronavirus from a dromedary camel, Qatar, 2014. *Emerging infectious diseases* **20:** 1339.

Rappuoli R (2004). From Pasteur to genomics: progress and challenges in infectious diseases. *Nature medicine* **10:** 1177-1185.

Salipante SJ, Hoogestraat DR, Abbott AN, SenGupta DJ, Cummings LA, Butler-Wu SM *et al* (2014). Coinfection of *Fusobacterium nucleatum* and *Actinomyces israelii* in mastoiditis diagnosed by next-generation DNA sequencing. *Journal of clinical microbiology* **52:** 1789-1792.

Sankar SA, Lagier J-C, Pontarotti P, Raoult D, Fournier P-E (2015). The human gut microbiome, a taxonomic conundrum. *Systematic and applied microbiology* **38:** 276-286.

Schneeberger PH, Becker SL, Pothier JF, Duffy B, N'Goran EK, Beuret C *et al* (2015). Metagenomic diagnostics for the simultaneous detection of multiple pathogens in human stool specimens from Côte d'Ivoire: a proof-of-concept study. *Infection, Genetics and Evolution*.

Schulz HN, Jørgensen BB (2001). Big bacteria. *Annual Reviews in Microbiology* **55:** 105-137.

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

Shi Y, Tyson GW, DeLong EF (2009). Metatranscriptomics reveals unique microbial small RNAs in the ocean's water column. *Nature* **459:** 266-269.

Shukla SK, Murali N, Brilliant M (2015). Personalized medicine going precise: from genomics to microbiomics. *Trends Mol Med* **21:** 461-462.

Spinsanti LI, Díaz LA, Glatstein N, Arselán S, Morales MA, Farías AA *et al* (2008). Human outbreak of St. Louis encephalitis detected in Argentina, 2005. *Journal of Clinical Virology* **42:** 27-33.

Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ *et al* (2015). Big Data: Astronomical or Genomical? *PLoS biology* **13:** e1002195.

Stewart PS, Costerton JW (2001). Antibiotic resistance of bacteria in biofilms. *The Lancet* **358:** 135-138.

Streicher E, Sampson S, Dheda K, Dolby T, Simpson J, Victor T *et al* (2015). Molecular epidemiological interpretation of the extensively drug resistant tuberculosis epidemic in South Africa. *Journal of clinical microbiology***:** JCM. 01414-01415.

Sturrock R, Jordan P, Webbe G (1993). The parasites and their life cycles. *Human schistosomiasis***:** 1-32.

Tringe SG, Von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW *et al* (2005). Comparative metagenomics of microbial communities. *Science* **308:** 554-557.

Tsai F, Coyle WJ (2009). The microbiome and obesity: is obesity linked to our gut flora? *Current gastroenterology reports* **11:** 307-313.

van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C (2014). Ten years of next-generation sequencing technology. *Trends in genetics* **30:** 418-426.

Visvesvara GS, Garcia LS (2002). Culture of Protozoan Parasites. *Clinical Microbiology Reviews* **15:** 327-328.

Wang Z, Gerstein M, Snyder M (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10:** 57-63.

Washington J (1996). *Medical Microbiology. 4th edition. Principles of Diagnosis.*

Williams R, Peisajovich SG, Miller OJ, Magdassi S, Tawfik DS, Griffiths AD (2006). Amplification of complex gene libraries by emulsion PCR. *Nature methods* **3:** 545-550.

Woods CW, Karpati AM, Grein T, McCarthy N, Gaturuku P, Muchiri E *et al* (2002). An outbreak of Rift Valley fever in northeastern Kenya, 1997-98. *Emerging infectious diseases* **8:** 138-144.

Yang S, Rothman RE (2004). PCR-based diagnostics for infectious diseases: uses, limitations, and future applications in acute-care settings. *The Lancet infectious diseases* **4:** 337-348.

Yin H, Zhang X, Liang Y, Xiao Y, Niu J, Liu X (2014). Draft genome sequence of the extremophile *Acidithiobacillus thiooxidans* A01, isolated from the wastewater of a coal dump. *Genome announcements* **2:** e00222-00214.

Young ND, Jex AR, Li B, Liu S, Yang L, Xiong Z *et al* (2012). Whole-genome sequence of *Schistosoma haematobium*. *Nature genetics* **44:** 221-225.

Zhou C, Smith J, Lam M, Zemla A, Dyer MD, Slezak T (2007). MvirDB—a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications. *Nucleic acids research* **35:** D391-D394.

Zhou Y, Zheng H, Chen Y, Zhang L, Wang K, Guo J *et al* (2009). The *Schistosoma japonicum* genome reveals features of host–parasite interplay. *Nature* **460:** 345-351.

Zignol M, Gemert Wv, Falzon D, Sismanidis C, Glaziou P, Floyd K *et al* (2012). Surveillance of anti-tuberculosis drug resistance in the world: an updated analysis, 2007-2010. *Bulletin of the World Health Organization* **90:** 111-119.

Zuo D, Mohr SE, Hu Y, Taycher E, Rolfs A, Kramer J *et al* (2007). PlasmID: a centralized repository for plasmid clone information and distribution. *Nucleic acids research* **35:** D680-D684.

Curriculum Vitae

Personal data

Name Pierre SCHNEEBERGER

Date of birth 27 April 1987 (Basel, Switzerland)

Citizenship Swiss and French

Work address University of Basel
Swiss Tropical and Public Health Institute
Department of Epidemiology and Public Health
CH-4051 Basel, Socinstrasse 57
E-mail : pierre.schneeberger@unibas.ch

Private address Unit 2701, 118 Balliol St., M4S 0A9, Toronto, Ontario, Canada
Phone : +1 996 8044
E-mail : pierre.schneeberger@gmail.com

Education

Since 10.2011 Universität Basel
PhD student at the Swiss Tropical and Public Health Institute
Epidemiology and Molecular Microbiology

2009-2011 Université De Strasbourg, France
Master's degree
Molecular and Cellular Biology, Genetics, Biochemistry

2006-2009 Université De Strasbourg, France

Bachelor's degree
Cellular Biology and Physiology

2005 Université de Strasbourg, France
School of Pharmacy

2002-2005 Lycée Jean Mermoz, Saint-Louis, France
(High School Diploma with option Biology)

Experience

Since 10.2011 PhD Candidate
Swiss Tropical and Public Health Institute, Basel
"Development and application of omics and bioinformatics approaches for a deeper understanding of infectious diseases systems."

1.2011-6.2011 Master thesis
University of Strasbourg / Syngenta, Stein-Säckigen
"Using molecular markers to identify three oilseed rape pests and characterization of resistance to pyrethroids"

7.2010-9.2010 Internship with Syngenta, Stein
Monitoring susceptibility of fungal pathogens to different fungicides

7.2009-8.2009 Internship with Syngenta, Stein
Developing projects, set up of agronomical tests

2009-2011 Part-time student job with Actelion, Allschwil
Organisation of clinical trial archives

6.2009 Internship with Ecole Nationale Superieure de Chimie Mulhouse

	Laboratory assistant
1.2002.2.2002	Internship with Syngenta, Basel Laboratory assistant
2002-2011	Part time student job with COOP, Basel Loading, Responsible for drinks display

Skills

Languages	French (native), German, English
Computer skills	Bioinformatics tools, Linux and Windows environments, high-performance computing, Perl and Python programming

Publications (October 2015)

Fuhrimann, S., M. Stalder, M. S. Winkler, C. B. Niwagaba, M. Babu, G. Masaba, N. B. Kabatereine, A. A. Halage, **P. H. H. Schneeberger** and J. Utzinger (2015). "Microbial and chemical contamination of water, sediment and soil in the Nakivubo wetland area in Kampala, Uganda." Environmental Monitoring and Assessment **187**(7): 1-15.

Fuhrimann, S., M. S. Winkler, **P. H. H. Schneeberger**, C. B. Niwagaba, J. Buwule, M. Babu, K. Medlicott, J. Utzinger and G. Cissé (2014). "Health risk assessment along the wastewater and faecal sludge management and reuse chain of Kampala, Uganda: a visualization." Geospatial health **9**(1): 241-245.

Polman, K., S. Becker, E. Alirol, N. Bhatta, N. Bhattarai, E. Bottieau, M. Bratschi, S. Burza, J. Coulibaly, M. Doumbia, N. Horie, J. Jacobs, B. Khanal, A. Landoure, Y. Mahendradhata, F. Meheus, P. Mertens, F. Meyanti, E. Murhandarwati, E. N'Goran, R. Peeling, R. Ravinetto, S. Rijal, M. Sacko, R. Saye, **P. H. H. Schneeberger**, C. Schurmans, K. Silue, J. Thobari and M. Traore (2015). "Diagnosis of neglected tropical

diseases among patients with persistent digestive disorders (diarrhoea and/or abdominal pain [greater than or equal to] 14days): A multi-country, prospective, non-experimental case-control study." BMC Infect Dis **15**(1): 338.

Reynard, J.-S. S., **P. H. H. Schneeberger**, J. Frey and S. Schaerer (2015). "Biological, serological and molecular characterization of a highly divergent strain of GLRaV-4 causing grapevine leafroll disease." Phytopathology.

Schneeberger, P. H. H., S. L. Becker, J. F. Pothier, B. Duffy, E. K. N'Goran, C. Beuret, J. E. Frey and J. Utzinger (2015). "Metagenomic diagnostics for the simultaneous detection of multiple pathogens in human stool specimens from Côte d'Ivoire: A proof-of-concept study." Infection, Genetics and Evolution.