ANA CAROLINA DE OLIVEIRA CONDEZ
Licenciada em Biologia Celular e Molecular

HUMAN POLYOMAVIRUSES IN WASTE AND ENVIRONMENTAL WATERS IN THE LISBON METROPOLITAN AREA

MESTRADO EM MICROBIOLOGIA MÉDICA
Universidade NOVA de Lisboa
Novembro, 2021

ANA CAROLINA DE OLIVEIRA CONDEZ
Licenciada em Biologia Celular e Molecular

HUMAN POLYOMAVIRUSES IN WASTE AND ENVIRONMENTAL WATERS IN THE LISBON METROPOLITAN AREA

MESTRADO EM MICROBIOLOGIA MÉDICA
Universidade NOVA de Lisboa
Novembro, 2021

HUMAN POLYOMAVIRUSES IN WASTE AND ENVIRONMENTAL WATERS IN THE LISBON METROPOLITAN AREA

ANA CAROLINA DE OLIVEIRA CONDEZ

Licenciada em Biologia Celular e Molecular

Orientador:	Prof. Doutor Ricardo Parreira, Professor Associado, Instituto de Higiene e Medicina Tropical, Universidade NOVA de Lisboa
Coorientador:	Doutora Mónica Nunes, Investigadora Júnior, Instituto de Biologia Experimental e Tecnológica

Júri:

Presidente:	Prof.ª Doutora Rita Sobral, Professora Auxiliar, FCT-NOVA
Arguente:	Doutora Sílvia Monteiro Investigadora, Instituto Superior Técnico, Universidade de Lisboa
Orientador:	Prof. Doutor Ricardo Parreira Professor Associado, Instituto de Higiene e Medicina Tropical, Universidade NOVA de Lisboa

MESTRADO EM MICROBIOLOGIA MÉDICA

Universidade NOVA de Lisboa
Novembro, 2021

Human polyomaviruses in waste and environmental waters in the Lisbon Metropolitan Area

Agradecimentos

Ao meu orientador, Prof. Ricardo Parreira, agradeço pela confiança depositada em mim nesta jornada, por todos os ensinamentos que me transmitiu e pela enorme paciência para todas as minhas questões. Obrigada por todos os conselhos, apoio, incrível disponibilidade, e por saber sempre quando precisava de uma palavra extra de encorajamento.

À Dra. Mónica Nunes, minha coorientadora, agradeço imenso por todo o apoio, confiança, paciência, ajuda incansável, e por demonstrar estar sempre disponível para o que quer que fosse preciso durante este percurso.

Aos restantes colegas do "Food Safety & Microbiology Lab" do iBET, em particular à Andreia e à Inês, o meu mais sincero obrigada por me terem deixado ocupar o vosso espaço, responderem a todas as minhas dúvidas e me terem ajudado sempre que precisei.

Aos meus amigos, por estarem sempre presentes e por todos os momentos de descontração que vinham sempre na altura certa, com uma clara menção honrosa para a minha "crew informática". Um especial obrigada à "Su Marie", por ter estado sempre presente quando precisava de dar voz a todas as minhas preocupações durante estes últimos dois anos.

Ao Diogo, por todo o incentivo, compreensão gigantesca, e por acreditar sempre em mim mesmo nos momentos em que eu não era capaz de o fazer.

À minha família, pela motivação, apoio e por me terem proporcionado esta oportunidade de aprender ainda mais um pouco. Sem vocês, isto nunca teria sido possível. Um especial destaque para os meus avós, D. Eva e Sr. Mário, por serem o maior porto de abrigo que alguma vez poderia pedir. Obrigada por todo o apoio e entusiasmo, mesmo naqueles momentos em que não entendiam nada do que eu estava a dizer. Palavras nunca serão suficientes para descrever o quão grata estou por vocês.

Bibliographic Elements

From the work described in this dissertation resulted the following published article:

Condez AC, Nunes M, Filipa-Silva A, Leonardo I, Parreira R. 2021. Human Polyomaviruses (HPyV) in Wastewater and Environmental Samples from the Lisbon Metropolitan Area: Detection and Genetic Characterization of Viral Structural Protein-Coding Sequences. Pathogens 10:1309 doi:10.3390/pathogens10101309.

This dissertation also contains data presented at the following scientific meeting:

Condez AC, Nunes M, Filipa-Silva A, Leonardo I, Parreira R. 2021. Genetic characterization of viral structural-protein coding-sequences of Human polyomaviruses present in wastewater and environmental samples in Lisbon. MicroBiotec'21 - Congress of Microbiology and Biotechnology 2021. 25th of November 2021. Online. Oral presentation.

Abstract

Human polyomaviruses (HPyVs) are commonly associated with asymptomatic infections, albeit some have been linked to diseases in immunocompromised hosts. Moreover, most of these viruses are present in human excreta, being routinely found in wastewater and environmental waters worldwide. Although the human health risk associated with the presence of fecal pollution in environmental waters has traditionally been assessed by the presence of fecal indicator bacteria, recently, HPyVs have been proposed as indicators of human waste contamination. This is mostly due to their host-specificity, non-seasonal distribution, and relatively high titers in untreated sewage, and other aquatic environments.

To date, there appears to be no substantial data surrounding the geographic distribution and genetic diversity of HPyVs in Portugal. Therefore, this study aims to address these issues by targeting the HPyVs' structural protein-coding region, which was partially amplified using two touch-down PCR multiplex protocols. To do so, wastewater influents and environmental water samples were collected between 2018 and 2020 in the Lisbon Metropolitan area, one of the most populated hubs in Portugal. The results disclosed the circulation of HPyV1, HPyV2, HPyV5 and HPyV6 in 35.3 % (n = 6), 29.4 % (n = 5), 47.1 % (n = 8) and 29.4 % (n = 5) of the water samples analyzed, respectively. Genetic characterization revealed the coexistence of various genotypes, with this being particularly true for HPyV2, for which five genotypes were identified (including a putative new genotype 9). The phylogenetic trees obtained for HPyV5 and HPyV6 appeared to be less robust than those of HPyV1/HPyV2, but their genetic diversity was still apparent.

These results support the clear usefulness of the molecular analysis of HPyVs in disclosing their epidemiological distribution patterns in the general population. Furthermore, this work demonstrates that studies involving their detection/genetic characterization are key to assess HPyVs' potential as human fecal markers.

Keywords: Human polyomavirus, Lisbon metropolitan area, wastewater influents, environmental waters, multiplex PCR.

Resumo

Os poliomavírus humanos (HPyVs) são normalmente associados a infeções assintomáticas, embora alguns sejam agentes etiológicos de patologias que ocorrem em hospedeiros imunocomprometidos. A maioria destes vírus encontra-se em excreções humanas, sendo frequentemente identificados em águas residuais e ambientais. Apesar da poluição fecal (e respetivo risco de saúde pública) em águas ambientais ser tradicionalmente avaliada pela presença de bactérias indicadoras de contaminação fecal, os HPyVs foram recentemente propostos como marcadores de contaminação fecal humana. Tal deve-se à sua especificidade para o hospedeiro humano, distribuição não sazonal e concentrações consideravelmente elevadas em águas residuais e outros ambientes aquáticos.

Atualmente, não existem dados claros sobre a distribuição geográfica e diversidade genética dos HPyVs em Portugal. Este estudo teve como objetivo abordar essas questões, através da amplificação de parte da região codificante das proteínas estruturais destes vírus, utilizando dois protocolos de PCR touch-down em multiplex. Amostras de influentes de águas residuais e ambientais foram recolhidas entre 2018 e 2020 na área metropolitana de Lisboa, o centro urbano mais populoso do país. Os resultados revelaram a deteção de HPyV1, HPyV2, HPyV5 e HPyV6 em 35,3% (n = 6), 29,4% (n = 5), 47,1% (n = 8) e 29,4% (n = 5) nas amostras de água analisadas, respetivamente. A sua caracterização genética revelou a coexistência de vários genótipos, especialmente para HPyV2, no qual foram identificados cinco dos genótipos (incluindo um possível novo genótipo 9). As árvores filogenéticas de HPyV5 e HPyV6 apresentaram um menor poder resolutivo do que o registado para HPyV1/HPyV2, mas foi possível constatar a sua diversidade genética.

Estes resultados demonstram a utilidade da análise molecular destes vírus no esclarecimento dos seus padrões de distribuição epidemiológica na população. Este trabalho demonstra ainda que estudos que envolvem a deteção/caracterização genética dos HPyVs são fundamentais para avaliar o seu potencial como marcadores de contaminação fecal humana.

Palavras-chave: Poliomavírus humano, área metropolitana de Lisboa, influentes de águas residuais, águas ambientais, multiplex PCR.

Contents

List of Figures

List of Tables

Abbreviations

ADB – Agarose dissolving buffer

aLRT – Approximate likelihood-ratio test

ALTO – Alternative Tumor Antigen

AML – Lisbon Metropolitan Area, from the Portuguese *Área Metropolitana de Lisboa*

bp – Base pairs

COVID-19 – Coronavirus disease 2019

dAMP – Deoxyadenosine 5′-monophosphate

DMSO – Dimethyl sulfoxide

DNA – Desoxyribonucleic acid

EDTA – Ethylenediaminetetraacetic acid

EPA – US Environmental Protection Agency

FIB – Fecal Indicator Bacteria

HPyV – Human polyomavirus

ICTV – International Committee on Taxonomy of Viruses

JCPyV – John Cunningham virus

KIPyV – Karolinska Institute polyomavirus

L-A – Lineage A

L-B – Lineage B

LB – Lysogeny Broth

LIPyV – Lyon-IARC polyomavirus

LTAg – Large Tumor Antigen

LTIIa – Heat-labile enterotoxin IIa

MCC – Merkel Cell Carcinoma

MCMC – Markov-Chain Monte-Carlo

MCPyV – Merkel Cell polyomavirus

ML – Maximum Likelihood

MST – Microbial Source Tracking

MTAg – Middle Tumor Antigen

NCBI – National Center for Biotechnology Information

NCCR – Non-coding control region

NGS – Next Generation Sequencing

NJPyV – New Jersey polyomavirus

NNn – NeighborNet Networks

nt – Nucleotides

ORF – Open reading frame

PCOORD – Principal Coordinate Analysis

PCR – Polymerase chain reaction

PEG – Polyethylene Glycol

PML – Progressive multifocal leukoencephalopathy

qPCR – quantitative (real-time) PCR

QPyV – Quebec polyomavirus

RNA – Ribonucleic acid

rRNA – Ribosomal RNA

SDS – Sodium dodecyl sulfate

SOC – Super Optimal Broth with Catabolite Repression

STAg – Small Tumor Antigen

STIb – Heat-stable enterotoxin Ib

STLPyV – Saint Louis polyomavirus

TAE – Tris-acetate EDTA

TSPyV – Trichodysplasia spinulosa polyomavirus

TSS – Transformation and Storage Solution

VLP – Viral-like particle

WHO – World Health Organization

WUPyV – Washington University polyomavirus

1 | Introduction

1.1 Water Quality Control - Fecal pollution

In recent years, the concern surrounding water quality has risen partly due to the numerous contamination episodes of water with pathogenic bacteria, viruses, and protozoan, with many cases of disease resulting from the transmission of serious waterborne infectious agents to humans and animals (1). Although microorganisms are naturally part of these aquatic ecosystems and are essential for their maintenance, due to the health risk that some of them may pose, it is important to guarantee microbial safety is secured when assessing water quality. Furthermore, by evaluating the microbial profile of the water, it is possible to determine if the water is appropriate for recreational use as well as other activities, by following either global or regional established quality standards (2, 3). One of the main contributors to the dissemination of pathogens in water is fecal pollution, which can turn any aquatic environments inappropriate for drinking, food production (including fish farming and shellfish harvesting), and recreational activities, thus posing a serious public health risk to the population (4).

Fecal pollution is frequently anthropogenic and zoogenic in origin and may stem from a point source discharge when the fecal material is derived from a specific identifiable source, namely raw sewage, wastewater treatment plant effluents, and industrial wastewater. Additionally, the contamination can also derive from a non-point pollution source, when the pollutants affect a wider area and cannot be attributed to a unique source, such as a draining pipe. This includes agricultural, urban, and forestry run-off (refer to Figure 1.1 for examples) (5).

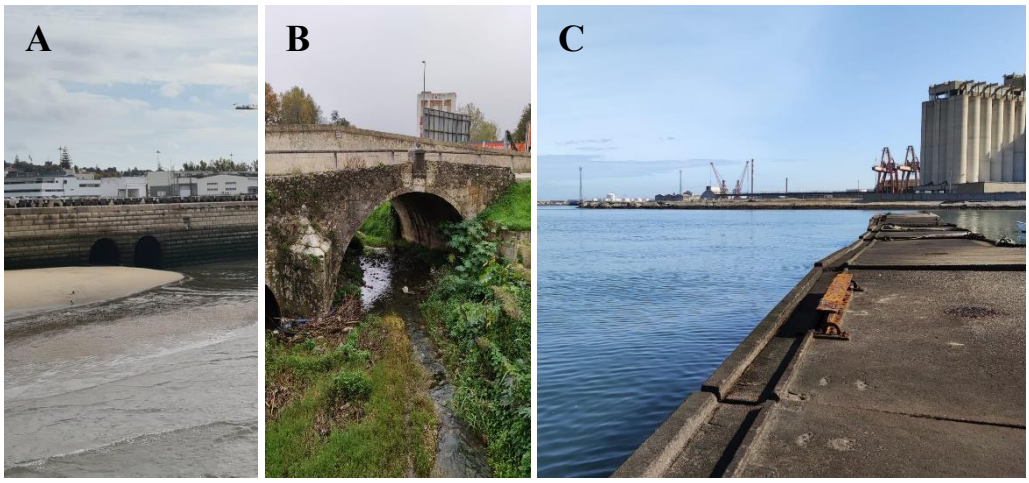


Figure 1.1. Possible examples of point and non-point source pollution in the Lisbon Metropolitan Area. (**A**) A sewage discharge point that leads to the Tagus River in Algés, Oeiras. (**B**) An open sewer pipe below the bridge arch, in Ribeira do Carenque, Amadora. (**C**) An industrial complex next to the Tagus River in Barreiro, Setúbal.

In order to reduce the potential human health risk and interrupt enteric pathogen transmission, it is very important that the fecal pollution source is quickly identified and conveniently eliminated. However, considering the numerous possible pollution sources and the diversity of microorganisms that inhabit the gastrointestinal tract and excreta (i.e., feces and urine) of humans and animals, it would be unrealistic to track and monitor all the possible bacterial, viral, and protozoan pathogens. Additionally, the detection and quantification processes for some microorganisms may be labor-intensive and expensive, with their concentration normally being low in natural waters and their presence or absence depending on the characteristics of the human and animal population in that area (6).

In an effort to determine the presence of fecal water contamination, several microorganisms were suggested as fecal indicators, whose levels should be monitored during routine microbiological water analyses. According to the US Environmental Protection Agency (EPA) and the World Health Organization (WHO), the ideal indicators should fulfill as many as possible of the established criteria, such as (i) be present in all types of water, (ii) share the same habitat as enteric pathogens, (iii) be unable to increase their numbers in the environment, (iv) display resistance to environmental conditions and disinfection, (v) be consistently present in feces and in the gastrointestinal tract of warm-blooded animals and (vi) allow for easy and simple laboratory processing (2, 7). However, so far, no microbiological indicator has fulfilled every single criterion.

Traditionally, fecal indicator bacteria (FIB) have been used to predict the presence of fecal pollution and to assess the associated public health risk (6). This group of indicator organisms includes *Escherichia coli*, several *Enterococcus* species, such as *E. faecalis* and *E. faecium*, and *Clostridium perfringens* (6, 8). Although some of these organisms have been used as fecal pollution indicators for centuries, numerous limitations have been acknowledged. For example, the presence of FIB might not actually be correlated with the presence of pathogens, considering different microorganisms possess different growth, ecological and physiologic characteristics, thus resulting in an inaccurate representation of the existence of waterborne pathogens and an improper evaluation of health risks (6). Other previously reported limitations include having a source other than fecal matter and greater weakness against disinfection methods such as UV radiation and chlorination, hence increasing the difficulty of FIB detection and leading to a lack of correlation with some of the pathogens present in the waters (1, 8). FIB have also been found to persist and maybe even multiply in some environments, while some enteric pathogens do not, leading to an even bigger discrepancy between the concentrations of the indicator organisms and pathogenic microorganisms (6, 9). Perhaps one of the biggest shortcomings with the application of FIB is that it does not allow to determine the source of fecal contamination, whether it is point or non-point, as well as if it has a human, domestic or wild animal source (4, 10). This greatly impairs the implementation of measures to quickly wipe out the pollution source and reduce

the public health risk. Moreover, since this group of indicator organisms might not be accurate at predicting the presence of all enteric pathogens, more research efforts were made to determine alternative indicators that allow the mitigation of at least some of the aforementioned limitations.

1.2 Microbial Source Tracking

In the last decade of the 20th century, numerous studies reported possible markers or indicators that would either allow for the distinction between anthropogenic and zoogenic fecal pollution or aid in the identification of the source of contamination. Hence, Microbial Source Tracking (MST) emerged as a collection of methods that would identify and differentiate the source of fecal contamination in water bodies and, subsequently, determine the influence that the fecal source might have on human health in case individuals come in contact with contaminated water (6). Currently, MTS methodologies can be divided into two major groups: library-dependent, and library-independent methods. Library-dependent methods require the creation of a database (i.e., library) of characteristics (sometimes referred to as fingerprints, patterns, or profiles) of bacteria isolated from various fecal sources and water samples. This approach involves the phenotypic or genotypic typing of several isolates, creating patterns that will later be used as a comparison with environmental strains to identify the source of contamination (11). This includes antibiotic resistance analysis and carbon source utilization profiling, as an example of biochemical typing methods, and pulse-field gel electrophoresis and ribotyping, as examples of the genotypic tests that can be utilized for bacterial characterization (12). The rationale behind this type of method is based on the principle that certain characteristics of FIB can be associated with a specific animal host and that bacteria isolated from the environment possess similar characteristics. However, it is hard to determine what defines a representative library and the number of isolates it should include. Additionally, the typing methods used for each microorganism might be time-consuming and expensive, which can significantly restrict not only library size, but also the specific isolates it includes, leading to the creation of reduced-size libraries that can, in turn, increase the difficulty of the contamination source identification (13).

Due to the shortcomings mentioned above, other methodologies have been explored, giving rise to library-independent methods. In this case, the identification of the pollution source does not rely on the comparison of environmental-obtained data with a defined library created *a priori*. Instead, it usually depends on the detection of host-specific markers in different microorganisms using methods that can either be qualitative or quantitative in nature (11). Although the most frequently used library independent methods involve the detection of genetic markers and are culture-independent, male-specific (F^+) RNA coliphage typing has been proposed as a quite effective method for the determination

of the contamination source. Coliphages are viruses that infect *E. coli* and, in the specific case of F^+ RNA phages, can be divided into four subgroups (I, II, III, IV) that are present in wastewaters, with II and III being usually associated with human feces and I and IV with animal feces. Hence, by serotyping or nucleic acid hybridization, it is possible to classify the F^+ RNA coliphages found in water bodies, allowing for the differentiation of fecal pollution sources. However, serological typing of coliphages has shown ambiguous results (8). Even though genotypic characterization produces more accurate data, it is a laborious method that requires the multiplication of coliphages in bacterial strains (8). Furthermore, it appears some subgroups may persist for longer periods of time in the environment, as well as being more resistant to disinfection methods, which might interfere with the results and consequently with their respective interpretation (14).

Other MST methods involving bacteriophages have already been tested. This investment is partly due to the hypothesis that bacteriophages might be better indicators of the presence of human enteric viruses in the environment than FIB, as both types of viruses present similar resistance to stress (15). Among the bacteriophages already proposed as possible indicators for human fecal pollution, are those infecting *Bacteroides fragilis* and *Bacteroides thetaiotaomicron*, anaerobic bacteria, members of the *Bacteroidales* order. Due to being abundant in the gastrointestinal tract of warm-blooded animals and apparently not being able to multiply in the environment, *Bacteroides* spp. have been proposed as indicators of fecal contamination, and are often used as a target for library-independent MST methods (16). Hence, bacteriophages that infect specific *B. fragilis* and *B. thetaiotaomicron* strains, such as HSP40 and GB-124 for the first, and GA17 for the latter, have been proposed as human-associated fecal indicators (17). Although *Bacteroides* spp. itself does not exclusively infect humans, bacteriophages as the ones mentioned above appear to be only found in anthropogenic fecal material, therefore allowing to distinguish between sources of fecal contamination (18). However, the usefulness of this method is limited by the fact that these strains are incapable of efficiently detecting phages in some geographical areas, thus jeopardizing its universal application (17).

These library-independent methods based on the detection and/or quantification of a host-specific target, are normally dependent on either a standard polymerase chain reaction (PCR) or take the form of a real-time PCR assay (also known as quantitative PCR or qPCR), allowing for the detection of bacteria, as well as viruses in water (6, 11). One of the first library-independent methods to be developed was the detection of a variable region of the 16S rRNA gene of *Bacteroides* spp. that allowed to differentiate between human and ruminant fecal matter (6). Since then, other systems utilizing non-16S rRNA markers have been designed, such as a qPCR protocol targeting *gyrB* of *B. fragilis*, a single copy gene that encodes for one of the subunits of DNA gyrase, as well as another qPCR assay that is based on the quantification of the putative α-1,6-mannanase of *B. thetaiotaomicron*. While these last two

methods are supposed to account for human fecal pollution, further evaluation of their performance should be undertaken, since the studies available for the evaluation of their sensitivity and specificity are scarce. Moreover, using the *gyrB* marker, a positive result was obtained against a sample of pig fecal matter, questioning its specificity (6).

Among other genetic markers, the enterococcal surface protein (*esp*) gene from *E. faecium* was also proposed as human-specific, with Scott et. al (19) offering a PCR detection method for this putative virulence factor that still relied on an enrichment culture step. Another gene marker approach proposed involved the detection of toxin genes from enterotoxigenic *E. coli*. In this case, through the PCR-based identification of the heat-stable enterotoxin Ib (STIb) coding gene (which is associated with human fecal waste), it is possible to detect anthropogenic fecal pollution (20). In parallel, by detecting the heat-labile enterotoxin IIa (LTIIa) coding sequence, it is possible to identify cattle fecal pollution in water (21). However, similarly to the non-16S rRNA markers for *Bacteroides* spp., *esp* and STIb have also been found in fecal samples from other animals, such as dogs and gulls, indicating that these genes might not be consistent markers of anthropogenic fecal pollution (6, 22).

1.3 Library-independent methods – Enteric viruses

Despite these MST methodologies solving some of the limitations that the traditional water quality control indicator microorganisms impose, whether the method used is library-dependent or independent, the fact remains that one marker might not correctly represent the presence of all the other microorganisms in the water body. Besides bacteriophages, none of the indicators mentioned above account for the presence of human enteric viruses. However, the search for an indicator that would allow for the direct monitoring of this group of microorganisms is of utmost importance, in order to try and eliminate any uncertainty surrounding their presence in the water that may result from the use of an inaccurate marker (23).

Enteric viruses are one of the most common waterborne viruses, many of which are known human waterborne pathogens, with more than 100 different viruses being associated with the human gastrointestinal tract (23). This group of microorganisms is commonly associated with diarrhea and self-limiting acute gastroenteritis, although they may also lead to other types of infections, such as conjunctivitis, hepatitis, encephalitis, and respiratory infections, particularly in immunocompromised individuals (1). As of 2016, diarrhea remained one of the leading causes of mortality in all age groups, with rotavirus (family *Reoviridae*) being the etiologic agent behind the biggest number of deaths (24). Besides rotavirus, other enteric viruses that are frequently studied as possible waterborne pathogens are

polyomaviruses (family *Polyomaviridae*), adenoviruses (family *Adenoviridae*), astroviruses (family *Astroviridae*), enteroviruses and hepatitis A viruses (family *Picornaviridae*), hepatitis E viruses (family *Hepeviridae*), and noroviruses (family *Caliciviridae*), with the latter being responsible for most cases of non-bacterial gastroenteritis (1, 25).

These viruses are excreted at high concentrations in the feces and urine of infected hosts, thus leading to their presence in wastewater and other surface waters when contaminated with untreated or improperly treated sewage, agricultural or urban runoff, subsequently creating possible routes of infection (1, 26). Furthermore, these viruses are known to be stable in the environment, displaying stronger resistance than bacteria to disinfection processes, with chlorination and UV radiation proving to be insufficient for the complete inactivation of some of these viruses. Additionally, most of them possess a relatively small infectious dose, especially when compared to other waterborne bacteria (1). This translates into the dissolution of the idea that FIB could potentially account for the presence of enteric viruses in water bodies, with previous studies already reporting the detection of viruses in waters where FIB counts were considerably low (27, 28).

Since these viruses pose a major health concern and bacterial indicators appear to be insufficient for the assessment of water quality, human enteric viruses have been proposed as MST tools that would allow for the discrimination between zoogenic and anthropogenic fecal contamination (23). Considering most of these viruses cannot be detected by cultivation methods, the majority of the designed assays involve the molecular detection of a specific genetic marker using various PCR techniques (1, 23). Among the viruses already suggested as possible water quality indicators are human polyomaviruses (HPyVs), which have drawn attention in this field of studies in recent years, due to their host-specificity and high prevalence in the human population. Moreover, HPyVs appear to be present at relatively high concentrations in a variety of water matrices, are found in various geographical locations and, unlike noroviruses, do not display seasonality, revealing a consistent presence year-round. The lack of envelope in the structure of the virion also contributes to the choice of HPyVs as a conceivable indicator for the presence of fecal pollution, as the absence of an envelope grants the viral particle higher resistance to environmental challenges, including the regular disinfection processes applied in wastewater treatment plants. Lastly, the detection of DNA instead of RNA, which constitutes the genome of most enteric viruses, makes the assay more cost-effective, as it obviates the otherwise needed reverse transcription step (29). Hence, several nested PCR and qPCR methods have been developed for the detection of the HPyVs in wastewater and environmental waters, with one of the most widely quoted methods being a qPCR that targets the conserved Large Tumor Antigen sequence (30) of the so-called BK and JC polyomaviruses (BKPyV and JCPyV, respectively; Table 1.1).

1.4 Human Polyomaviruses – General Characteristics

HPyVs are a group of viruses that belong to the *Polyomaviridae* family which, according to the International Committee on Taxonomy of Viruses (ICTV) can be divided into six different genera: *Alphapolyomavirus*, *Betapolyomavirus*, *Deltapolyomavirus*, *Epsilonpolyomavirus*, *Gammapolyomavirus,* and *Zetapolyomavirus*, defined by the phylogenetic relationships of the Large Tumor Antigen protein sequence (31, 32). The several species of HPyVs are distributed throughout the first three genera (31) (Table 1.1).

Table 1.1. Human polyomaviruses and respective associated human diseases.

Virus	Genus	Species	Abbreviation[a]	Associated clinical diseases
BKPyV	*Betapolyomavirus*	*Human polyomavirus 1*	HPyV1	Hemorrhagic cystitis, nephropathy
JCPyV	*Betapolyomavirus*	*Human polyomavirus 2*	HPyV2	Progressive multifocal leukoencephalopathy
KIPyV	*Betapolyomavirus*	*Human polyomavirus 3*	HPyV3	None
WUPyV	*Betapolyomavirus*	*Human polyomavirus 4*	HPyV4	None
MCPyV	*Alphapolyomavirus*	*Human polyomavirus 5*	HPyV5	Merkel cell carcinoma
HPyV6	*Deltapolyomavirus*	*Human polyomavirus 6*	HPyV6	Pruritic dermatitis
HPyV7	*Deltapolyomavirus*	*Human polyomavirus 7*	HPyV7	Pruritic dermatitis
TSPyV	*Alphapolyomavirus*	*Human polyomavirus 8*	HPyV8	Trichodysplasia spinulosa
HPyV9	*Alphapolyomavirus*	*Human polyomavirus 9*	HPyV9	None
HPyV10	*Deltapolyomavirus*	*Human polyomavirus 10*	HPyV10	None
STLPyV	*Deltapolyomavirus*	*Human polyomavirus 11*	HPyV11	None
HPyV12[b]	n.d.	*Human polyomavirus 12*	HPyV12	None
NJPyV	*Alphapolyomavirus*	*Human polyomavirus 13*	HPyV13	None
LIPyV	*Alphapolyomavirus*	*Human polyomavirus 14*	HPyV14	None
QPyV[c]	n.d.	n.d.	-	None

"n.d." stands for "non-determined".
[a] How each virus will be referred to during the course of this work.
[b] Excluded from the family on the ICTV 2019 report.
[c] Recently described, not yet featured on the current family *Polyomaviridae* taxonomy report (ICTV 2020).

The viruses of this family possess small, non-enveloped virions, with a circular dsDNA genome with approximately 5,000 base pairs (bp) linked to cellular histones (Figure 1.2A). The structural

organization of the genome is quite similar throughout the different species of polyomaviruses, having two different transcriptional regions, the early and the late region, named after the stage of infection in which each is transcribed (Figure 1.2B). These regions are separated by the non-coding control region (NCCR), the part of the genome that encompasses the origin of replication and the promoters that allow for the transcription of both the early and late regions (33).

The early region is normally comprised of the genes that encode the Large Tumor Antigen (LTAg) and the Small Tumor Antigen (STAg), two regulatory proteins obtained by alternative splicing. The LTAg is a multifunctional protein that is essential for viral replication and is also involved in cellular transformation, while the STAg is usually associated with an increase in the efficiency of viral proliferation and optimal cellular transformation, even though these processes are not completely dependent on its expression (34). On the other hand, the late region encodes for the structural proteins that compose the viral capsid. For most polyomaviruses, those proteins are VP1, VP2, and VP3, with the first two being synthesized in an infected cell as a consequence of mRNA alternative splicing, while VP3 is translated as a result of the use of an alternative starting codon present in the transcript that originates VP2. The capsomers that compose the viral capsid are formed by five VP1s that interact with either one VP2 or one VP3 molecule (33).

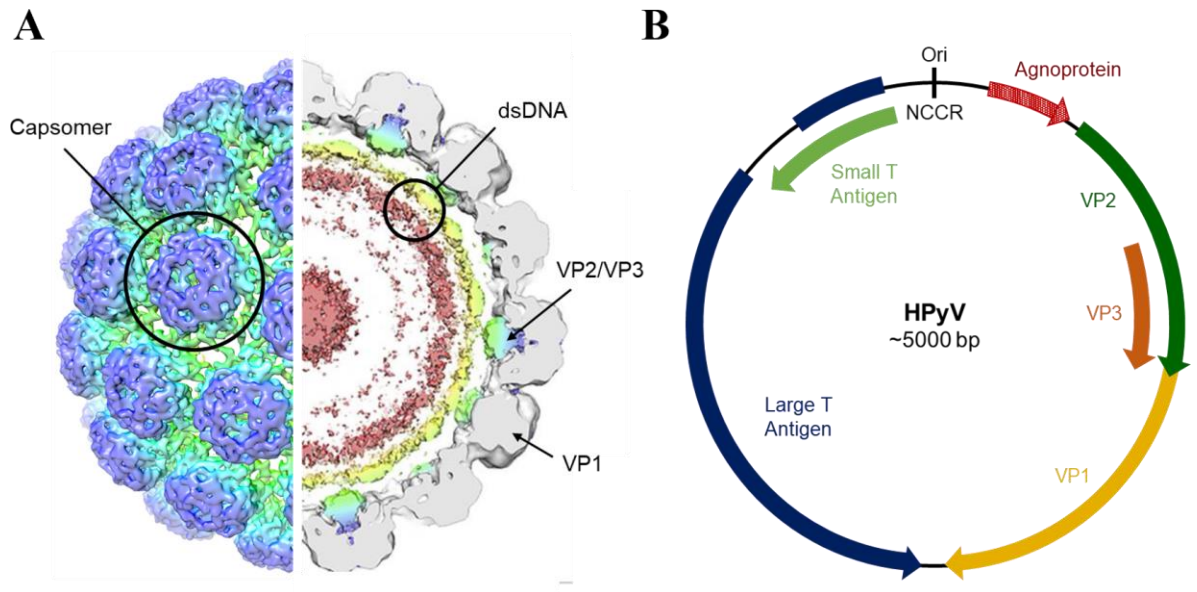


Figure 1.2. Virion and genome of a Human polyomavirus (HPyV). (**A**) The CryoEm structure of the external side of the BK virion (left) and the internal side (right). Adapted from Hurdiss et. al. (35) (**B**) Schematic representation of the genome of a HPyV. The genomic organization of a typical HPyV is illustrated, with the open reading frames being depicted by curved colored arrows.

In the genome of some polyomaviruses, an additional open reading frame (ORF) located upstream of the late region may be present. This region of the viral genome corresponds to the *agnogene*, a gene that encodes for the agnoprotein, a multifunctional hydrophobic peptide that plays a role in viral replication and transcription, seemingly also being essential for viral assembly (36). Another extra ORF might be present in the early region of the genome, being expressed either as the Alternative Tumor Antigen (ALTO) or as one of the exons of the Middle Tumor Antigen (MTAg). The function of both of these alternative early proteins is still not well understood, but it appears they may be involved in cellular transformation (37). In the HPyV group, only the JC polyomavirus (JCPyV) and BK polyomavirus (BKPyV) possess an *agnogene*, while the ALTO protein is present in Merkel Cell Polyomavirus (MCPyV) and Trichodysplasia spinulosa polyomavirus (TSPyV), with the latter also expressing MTAg (36, 38).

The viral life cycle begins when the virus enters the host cell via endocytosis through the interaction of viral proteins and cell surface receptors (Figure 1.3). In the endoplasmic reticulum, the capsid structure is rearranged, resulting in the release of a partly coated virion in the cytoplasm. Polyomaviruses are known to infect quiescent cells, even though their replication is fully dependent on the host. Hence, LTAg is essential during this process, ensuring that the cell enters the S phase, and the host's machinery is readily available to be used for viral replication. The viral DNA is subsequently brought into the host nucleus, where the transcription of the early genes by the host transcriptional machinery ensues, with the expression of LTAg eventually leading to genome replication. After replication has been initiated, transcription of the late genes begins in the opposite strand as well as direction from the early genes. The viral capsid proteins are then synthesized in the cytosol before being transported to the nucleus, where viral chromatin has been assembled into nucleosomes with host histones. The newly formed virions are released by lysis of the host cell or by other nonlytic mechanisms (Figure 1.3) (33).

So far, 13 different viral species are identified by the ICTV as being part of the HPyV group, with most of them having already been described in various geographical locations (32). All these viruses are thought to only infect humans, being ubiquitous in the population, with seroprevalence values reaching values as high as 90%. Accompanied by the increase in seropositivity in the first years of life, this information suggests that HPyVs infect humans from an early age (39, 40). HPyVs are usually associated with persistent yet asymptomatic infections. However, in instances where, for some reason, the host's immune system is compromised, clinical disease may ensue due to viral reactivation. The clinical manifestation of a HPyV infection will depend on the virus in question, with only some of these viruses being associated with disease in immunocompromised hosts (41).

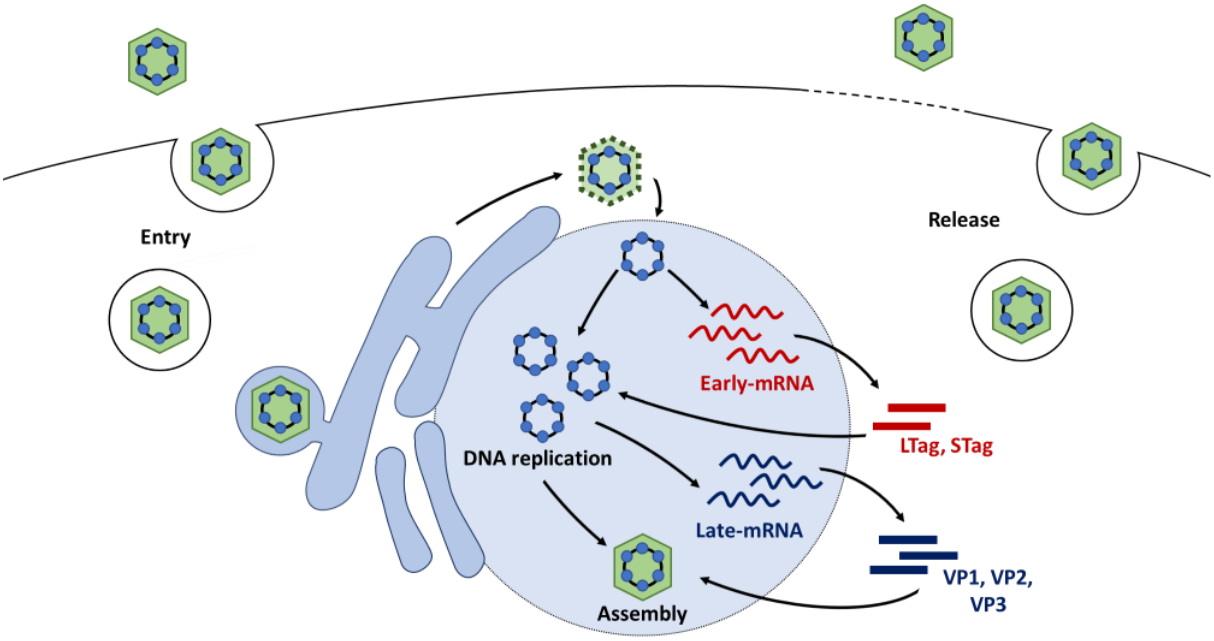


Figure 1.3. Model of the HPyV life cycle. The viral particle enters the host cell by endocytosis and transits through the endoplasmic reticulum, where the capsid suffers conformational changes. The partly coated virion is then transported to the cytoplasm and the DNA is imported to the nucleus, where the early genes (red) will be transcribed for DNA replication to begin. Afterwards, the production of the late proteins (blue) begins, and the viral life cycle enters its assembly stage. The assembled virions either exit the cell by lysis or through a nonlytic egress.

1.5 BK and JC polyomavirus

The first human polyomaviruses to be described were BK polyomavirus (species *Human polyomavirus 1*, henceforth designated HPyV1) and JC polyomavirus (species *Human polyomavirus 2*, from now on referred to as HPyV2), having both been identified in 1971 (42, 43). Evidence from previous studies shows fecal/urine-oral transmission as a possible path of infection, due to the fact that HPyV1 is excreted in urine and feces and HPyV2 in urine (44, 45). Furthermore, since these viruses have been detected in tonsillar tissues, respiratory transmission has also been suggested, creating the hypothesis that the respiratory tract might be the initial infection site (46, 47). Both of these viruses may show seroprevalence values above 80% and 60%, for HPyV1 and HPyV2, respectively in the healthy population (39, 40).

After the initial infection, HPyV1 is known to persist in the kidneys and urinary tract (48). Although viral reactivation has been reported in immunocompetent individuals, when this process occurs in immunocompromised hosts, HPyV1 has been associated with two major illnesses in transplant patients (41, 44). In renal transplant patients, HPyV1 infection is known to cause polyomavirus-associated nephropathy, while in bone marrow transplant recipients it is linked to hemorrhagic cystitis (41).

This virus has been classified into four different genotypes (I – IV) according to the analysis of the genetic variability of the VP1 coding region (49). Genotype I is distributed worldwide, being the most frequently detected. It is further divided into four subtypes: Ia, Ib1, Ib2, and Ic, with Ia being the most prevalent in Africa, Ib1 in Southeast Asia, Ib2 in Europe, and Ic in East Asia (50, 51). Another genotype that is widely distributed is IV, which is classified into six subtypes: IVa1, IVa2, IVb1, IVb2, IVc1, and IVc2, all of them prevalent in East Asia with the exception of IVc2, which is also present in Europe and Northeast Asia (52). In contrast with genotypes I and IV, genotypes II and III are more scarcely detected worldwide (53).

HPyV2 also causes persistent infections in the urinary tract and can infect cells of the central nervous as well as the lymphatic system (48, 54). Similarly to HPyV1, when reactivation of HPyV2 replication occurs in the brain of immunocompromised individuals, it can lead to progressive multifocal leukoencephalopathy (PML), a neurological demyelinating disorder characterized by lesions in the white matter of the brain, often resulting in cognitive impairment and motor dysfunction. PML usually occurs in individuals infected with HIV-1 but it has also been associated with patients receiving immunomodulating therapy (41).

Several algorithms have been used for genotyping HPyV2, classifying this virus into several different genotypes that, as with HPyV1, are associated with specific human populations and geographical areas (55–57). Indeed, studying the molecular epidemiology of both these viruses may come as an extremely useful tool in disclosing the diversity of a particular population, which resulted from various migration processes that have occurred throughout the centuries. Considering this fact, it has been hypothesized that HPyV1 and HPyV2 may have co-diverged with humans, with the latter even being proposed as a human migration marker, and thus being used to clarify the history of the human population (58). Although the same proposal has been done regarding HPyV1, the geographical distribution displayed by its genotypes implies that a more complex evolution process took place when compared to HPyV2. This is mostly due to the differences between genotype I and IV since the first might have been originated in Africa and afterwards diverged with the "Out-of-Africa" movement (i.e., a theory that proposes that the majority of current non-African populations descend from humans that left Africa approximately 100-200 millennia ago), while genotype IV is not prevalent in Africa and displays an uneven geographic distribution when compared to genotype I (58).

Two nomenclatures have been proposed for HPyV2 genotypes, with the one proposed and reviewed by Stoner et. al (57) being the one used in this work. Genotype 1 (further divided into 1A and 1B) is mostly associated with Europeans and North Americans, with genotype 4 registering a similar pattern (56). Genotype 2 is more complex, having been classified in five different subtypes: 2A, 2B, 2C, 2D,

and 2E. Subtypes 2A and 2C form a paraphyletic group, representing the HPyV2 mostly linked to Eastern Asians and Native Americans. Subtype 2B is associated with Europeans and Eastern Asians whereas subtype 2D, which is subdivided into 2D1, 2D2, and 2D3, is commonly distributed in Asia (59–61). Lastly, 2E is the subtype that predominates in Oceania (62). Genotype 3 has been associated with some African and Asian populations, with subtype 3A being most prevalent in South, West, and Central Asia and almost the entirety of the African continent, except for the South, where subtype 3B predominates (59, 63). Likewise, genotype 6 is linked to African populations, being more commonly found in Central and Western Africa (63). Genotype 7 is classified into subtypes 7A, 7B (encompassing 7B1 and 7B2), and 7C (further divided into 7C1 and 7C2) and is predominantly found in Asia, with subgroup 7C2 also being found in Mauritius (59). Genotype 8 is the most prevalent in the Western Pacific region, where both 8A and 8B dominate in Polynesia and Malesia, with the last one also being present in Papua New Guinea (62, 64). Lastly, genotype Eu-c appears to be only present in Northeast Siberia and Japan (56).

1.6 KI and WU polyomavirus

In 2007, Karolinska Institute polyomavirus, KIPyV (species *Human polyomavirus 3*, from now on referred to as HPyV3) and Washington University polyomavirus, WUPyV (species *Human polyomavirus 4*, hereafter HPyV4) were identified as two new HPyVs (65, 66). Both of these viruses have been mostly reported in respiratory samples taken from children, but have also been detected in specimens from immunocompromised adults (67–69). Additionally, multiple studies have reported the presence of HPyV3 and HPyV4 in stool samples (67, 70, 71). So far, neither virus has been identified as an etiologic agent of any clinical disorder, even though both have mostly been detected in samples from patients with respiratory illnesses (67). Nonetheless, HPyV3 and HPyV4 normally register seroprevalences of over 80% in the healthy adult population (39, 40).

Unlike HPyV3, HPyV4 has already been classified into three major clades (I - III) and seven subtypes: Ia, Ib, Ic and Id, IIIa, IIIb, and IIIc, according to the phylogenetic analysis of the whole-genome sequence (72, 73). Correspondingly, genotyping methods involving either only a small region of VP2 (denominated VP2 typing region) or a region that encompasses the interface between VP2 and VP1 have already been proposed (72, 74).

1.7 Merkel cell polyomavirus

MCPyV (species *Human polyomavirus 5*, henceforth HPyV5) is frequently found as part of the skin microbiota (75). First identified in 2008, it is the only HPyV with a known association to an oncogenic illness, the Merkel Cell Carcinoma (MCC), a rare, aggressive form of skin cancer that usually occurs in elderly and/or immunocompromised individuals, with HPyV5 having been linked to approximately 80% of MCC cases (76). Similar to the other HPyVs, this virus displays a high seroprevalence, varying between 70% and 80% (39, 40). Besides the skin, HPyV5 has also been found in respiratory secretions, urine, and stool samples, illustrating the fact that, like its predecessors, this virus appears to also be shed by the fecal route (77, 78).

Based on the analysis of a fragment composed of the LTAg and VP1 coding-regions, HPyV5 has been divided into five genotypes, each with a suggested strong geographical association: Europe/North America, Africa, Asia, Oceania, and South America (79). Other genotyping methods have been proposed since, with a study showing that it's possible to achieve similar results with only VP1's coding region and another proposing the NCCR as a new typing region (80, 81).

1.8 Human polyomavirus 6 and Human polyomavirus 7

First described in 2010, HPyV6 (species *Human polyomavirus 6*) and HPyV7 (species *Human polyomavirus 7*) are skin-tropic viral agents that, like HPyV5, are chronically shed from human skin (75). Viral DNA is frequently found in skin swabs, with prevalence in healthy individuals reaching values as high as 18% and 12% for HPyV6 and HPyV7, respectively (82). On the other hand, seroprevalences of above 55% have been reported for both viruses, with HPyV7 registering slightly lower values than HPyV6 (39, 40). Besides the skin, these viruses have also been detected in feces and respiratory samples (83). HPyV6 and HPyV7 are routinely associated with asymptomatic infections, but recent reports suggest a possible association with inflammatory skin disorders such as pruritic and dyskeratotic rashes (84). Furthermore, HPyV6 and HPyV7's viral DNA was found in several types of skin tumors, with HPyV7 also being detected in thymic epithelial tumors, which might suggest that these viruses may have oncogenic potential (84).

Due to the low number of available HPyV6 and HPyV7 sequences in databases, studies regarding the molecular epidemiology of these viruses remain scarce. However, through the analysis of whole-genome sequences and/or partial sequences of HPyV6, two main clades have been identified. The first group is comprised of HPyV6 sequences from various geographical locations, while the second group mainly includes sequences of Asian origin. Furthermore, another clustering phenomenon appears to

occur inside the first group, where HPyV6 sequences form subgroups according to their geographical origin, resulting in four distinct groups from Europe, North and South America, Oceania, and Asia (58).

1.9 Trichodysplasia spinulosa polyomavirus and Human polyomavirus 9

Trichodysplasia spinulosa polyomavirus, also known as TSPyV (species *Human polyomavirus 8*, hereafter HPyV8) was first identified in 2010 as the possible etiologic agent of trichodysplasia spinulosa, a rare skin disease that occurs exclusively in immunocompromised patients, usually due to a transplant or immunosuppressive treatment for leukemia (85). Although the presence of viral DNA on the skin is strongly correlated with disease, HPyV8 registers a seroprevalence of approximately 80% in the adult healthy population (39, 40, 86). This virus has been detected in other regions of the body besides the skin, including the upper respiratory tract, blood, cerebral spinal fluid, urine, and feces (83, 87, 88). Nonetheless, not enough evidence has been collected to allow for the identification of HPyV8 as the possible causative agent of any other illness.

In 2011, HPyV9 (species *Human polyomavirus 9*) was detected for the first time in the serum and urine of a kidney transplant patient and, unlike HPyV8, an association between this virus and a clinical disorder is yet to be established (89). In regards to seroprevalence, this virus displays values of approximately 20% in immunocompetent populations (39, 40).

1.10 Human polyomavirus 10 and Saint Louis polyomavirus

In 2012, HPyV10 (species *Human polyomavirus 10*) was identified in feces from healthy and diarrheic individuals, and in anal warts of an immunosuppressed individual (90–92). In addition to the detection in stool samples, HPyV10 viral DNA has also been found in respiratory and serum samples (83). Saint Louis polyomavirus, STLPyV (species *Human polyomavirus 11*, henceforth HPyV11) was found in 2013 in a stool sample taken from a child (93). In healthy populations, the seroprevalence is usually higher than 80% for HPyV10 and 60% for HPyV11 (39, 40). Despite having both been found in diarrheic feces, up to this date, HPyV10 and HPyV11 appear to have no clinical significance (94).

1.11 Human polyomavirus 12, New Jersey polyomavirus, Lyon-IARC polyomavirus, and Quebec polyomavirus

From 2013 to 2019, four other HPyV were described: HPyV12 (species *Human polyomavirus 12*), New Jersey polyomavirus, NJPyV (species *Human polyomavirus 13*, from now on referred to as HPyV13), Lyon-IARC polyomavirus, LIPyV (species *Human polyomavirus 14*, hereafter HPyV14), and Quebec polyomavirus, QPyV (unassigned species) (95–98). The seroprevalence registered for the first three viruses is much lower than the ones reported for the aforementioned HPyVs, with seropositivity only reaching 5% among immunocompetent individuals (40). For QPyV, seroprevalence values are yet to be assessed. Considering the low seroprevalences and seroreactivity reported for these viruses, their status as HPyV has been questioned and, in some cases, even revoked.

HPyV12 was first identified in 2013 in the human gastrointestinal tract. However, due to the discovery that this virus infects common shrews, it has since been excluded as a human virus in the ICTV taxonomy 2019 report (32, 99). Although HPyV14 was already detected in human blood and skin samples, it has also been found in diarrheic feces from felines, which suggests that similarly to HPyV12, this virus might not be a *bona fide* HPyV (96, 100, 101).

In regard to QPyV, it was first identified in the feces of an elderly patient in Canada and has escaped proper classification as a HPyV by the ICTV (95). A recent study detected QPyV DNA in urine samples from pregnant women, multiple sclerosis, and systemic lupus erythematosus patients, being the first-ever report on the prevalence of this virus (102). However, further studies will be needed to determine whether QPyV is a genuine HPyV.

1.12 Justification for this Dissertation and objectives of this work

In order to establish new MST tools and determine new microbiological fecal pollution indicators, it is necessary to consider that the performance of any MST method and the presence of any indicator can vary with time and geographical location (11). In the particular case of HPyVs, the epidemiology of the different polyomaviruses within the group may differ according to the analyzed location, leading to the presence of various epidemiological patterns. Furthermore, although HPyVs constitute a group of highly distributed viruses, the distribution of the different genotypes/subtypes might also be affected by the geographic region (58).

To date, epidemiological data related to HPyVs in Portugal appears to be inexistent, with no consistent data surrounding its geographical distribution having been reported. Therefore, the objectives

of this work were to describe the molecular epidemiology of the several HPyVs in the Lisbon metropolitan area (AML, from the Portuguese *Área Metropolitana de Lisboa*) and by analyzing raw sewage (i.e., influent) and environmental water samples, assess whether HPyVs would be a viable indicator of human fecal contamination in Portuguese waters.

Considering all the aforementioned facts, two nested touch-down multiplex PCR protocols were developed in order to assess the presence of HPyVs in each sample. Using (essentially) a molecular cloning-based approach, the recovery of viral DNA sequences from the wastewater and environmental samples allowed us to determine which strains of HPyVs are circulating in the population of the AML. The genetic diversity of such strains was analyzed using phylogenetic, network-reconstruction, and Principal Coordinate Analysis, with the identification of the viral species being additionally confirmed using Next Generation Sequencing (NGS) data.

2 | Materials and Methods

2.1 Sample Collection

For this work, 17 water samples were collected in 14 different locations spread across the AML. The AML is composed of 18 municipalities that are divided between the Lisbon and Setubal districts, located on the north and south side of the Tagus River, respectively. It is the second most populated area in Portugal, registering 2 821 876 inhabitants in 2011, close to a third of Portugal's population (103).

All samples were transported to the laboratory at room temperature, and they were either immediately processed or stored for a maximum of 48 h at 4 ºC after collection. The collection sites are illustrated in Figure 2.1 and the list of samples is summarized in Table 2.1.

Table 2.1. List of collected samples and respective collection sites and dates.

Sample type	Collection site	GPS Coordinates	Date of collection
Environmental	Algés (Tagus River)	38º 41’ 41’’ N 9º 13’ 48’’ W	October 2020
	Barreiro (Tagus River)	38º 39’ 55’’ N 9º 4’ 39’’ W	October 2020
	Industrial Area of Setúbal (Sado River)	38º 30’ 22’’ N 8º 50’ 52’’ W	November 2020
	Lizandro River	38º 56’ 24’’ N 9º 24’ 47’’W	November 2020
	Port of Setúbal (Sado River)	38º 31’ 14’’ N 8º 53’ 14’’ W	November 2020
	Ribeira das Lajes	38º 41’ 9’’ N 9º 18’ 51’’ W	October 2020
	Ribeira de Carenque	38º 45’ 24’’ N 9º 14’ 55’’ W	November 2020
	Trancão River	38º 47’ 47’’ N 9º 5’ 57’’ W	October 2020[a]
Wastewater (influent)	Wastewater treatment plant A	n.a.	October 2018 April 2019 July 2020[a]
	Wastewater treatment plant B	n.a.	October 2018[a] April 2019[a]
	Wastewater treatment plant C	n.a.	November 2020[a]
	Wastewater treatment plant D	n.a.	November 2020
	Wastewater treatment plant E	n.a.	November 2020[a]
	Wastewater treatment plant F	n.a.	November 2020[a]

º indicates degrees,’ minutes and ’’ seconds.
“n.a.” stands for “not applied” (the wastewater samples were collected under a confidentiality agreement).
[a] DNA extracts implicated in the NGS analysis.

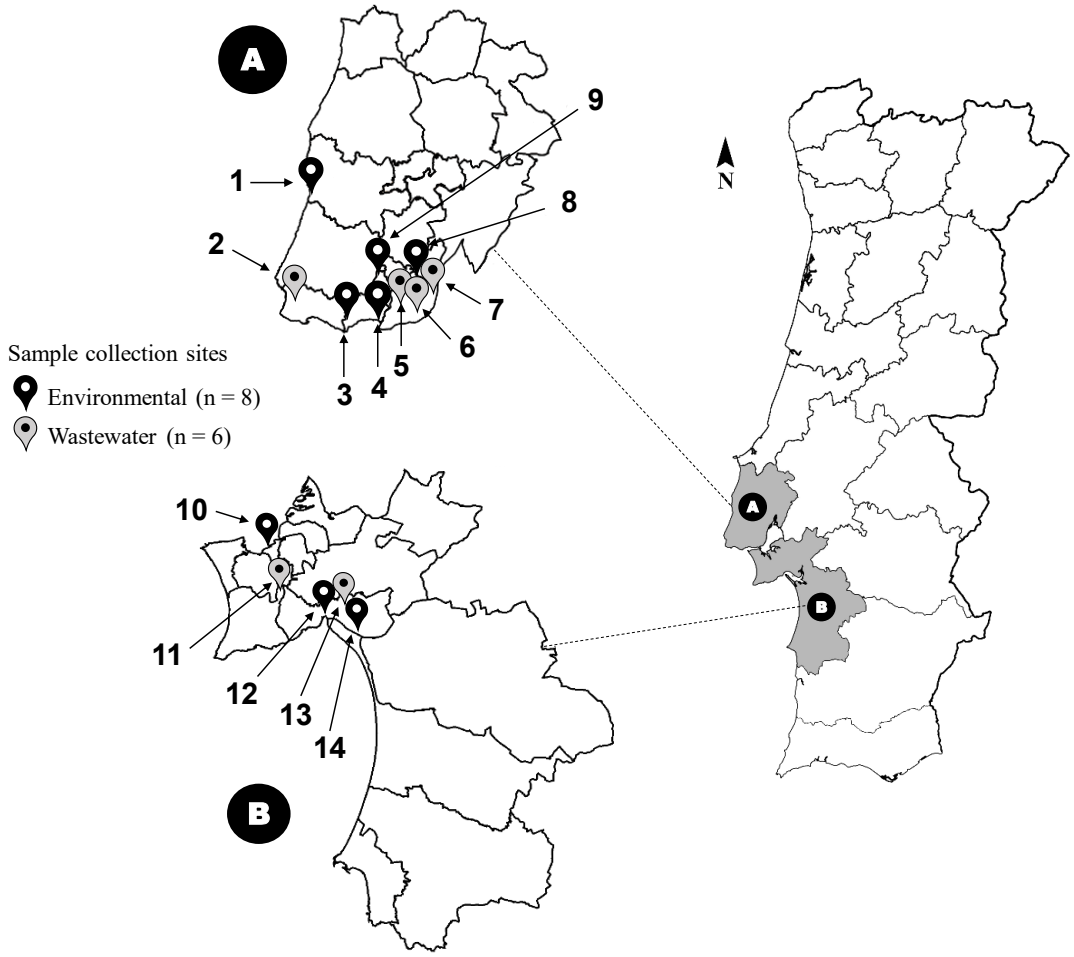


Figure 2.1. Geographic distribution of the wastewater and environmental sampling sites in the AML. All the collection sites (numbered 1 to 14) are distributed between both the districts that are part of the AML, with sites 1 to 9 being marked in Lisbon (indicated by A) and sites 10 to 14 in Setúbal (indicated by B). The total number of collection sites is indicated between brackets.

2.1.1 Environmental water samples

A total of eight environmental water samples (10 L) were collected throughout the months of October and November of 2020. Two of the samples were taken from the margins of the Tagus River, two from the margins of the Sado River, one from the mouth of the Lizandro River, and lastly, one from the Trancão River, a tributary of the Tagus River. The last two samples were taken from two different creeks, Ribeira do Carenque and Ribeira das Lajes, both located in the Lisbon district. All the environmental collection sites were selected based on their proximity with possible point and non-point fecal pollution sources.

2.1.2 Wastewater samples

For the wastewater samples, the solid particle-free influent of six different wastewater treatment plants scattered across the AML was taken in several different periods. Depending on the different collection sites, 1 L of wastewater was collected either in October 2018, April 2019, July, October, or November of 2020, with some of the sites being sampled on more than one occasion. In total, nine wastewater samples were collected.

2.2 Virus Concentration by skimmed milk flocculation

The viral-like particles (VLP) present in both environmental and wastewater samples were concentrated based on the procedures previously described by Calgua *et al.* (104). Firstly, a skimmed milk solution of 1% (w/v) was prepared using skimmed milk powder (Difco, USA) and 3.2% (w/v) synthetic seawater (Paragon Scientific, UK). This solution was added directly to the collected water, with 100 mL and 10 mL being used for environmental and wastewater samples, respectively. For milk protein flocculation to occur, the pH of the resulting solution was adjusted to 3.5 using 1 M HCl. The samples were gently stirred for 8 h at room temperature, and the flocculants were then left to sediment by gravity for another 8 h. Afterwards, so as to not disturb the sediment, the supernatant was carefully removed with a vacuum pump, and the remaining volume (of about 500 mL) transferred to a centrifuge container and centrifuged at 5,500 *g*, for 45 min, at 4 ºC. The supernatant was removed, and the pellet resuspended in 8 mL of phosphate buffer (1:2, v/v of 0.2 M Na_2HPO_4, 0.2 M NaH_2PO_4, pH 7.5). The VLP concentrate was stored at -20 ºC until DNA extraction was performed.

2.3 DNA extraction from VLP concentrates

The DNA from environmental and wastewater samples was extracted with the QIAmp® DNA Mini Kit (QIAGEN, USA). Firstly, 200 µL of each sample was mixed with 200 µL of Buffer AL followed by an incubation period at 70 ºC (10 min). Afterwards, 200 µL of ethanol (100%) was added and the entire mix was loaded onto a column (QIAmp Mini Spin) and centrifuged at 6,000 *g* for 1 min. This was followed by two washing steps, being added 500 µL of the first washing buffer (AW1) to the column and, after a second centrifugation, 500 µL of the second washing buffer (AW2). After centrifugation at 20,000 *g* for 3 min, the DNA was eluted with 100 µL of elution buffer (AE). DNA concentrations and purification were determined with NanoDrop™ 1000 (ThermoFisher Scientific, USA) and the extracts were stored at -20 ºC until further analysis.

2.4 Primer Design

The primers designed in this study for the detection of HPyVs were created based on the resulting alignment of several structural protein-coding sequences from the genome of several HPyVs, identified via their accession numbers in GenBank®. This alignment was rendered using the multiple alignment tool MAFFT version 7 (105), and exploiting the iterative refinement method G-INS-i. By giving priority to more conserved areas of the alignment, the primers were designed using a combination of the software PrimerDesign-M (106) (available at https://www.hiv.lanl.gov/content/sequence/PRIMER_DESIGN/primer_design.html) and manual visual inspection of the alignments. Furthermore, in order to ensure specific and sensitive detection, the construction of the primers was made considering several standardized characteristics, such as the G+C content being over 30%, the presence of one or more G/C residues at the end of the primer (commonly referred to as "GC clamp"), the melting temperature not being lower than 45 ºC, the inclusion of no more than three degenerate positions in the primer sequence, the length of each primer being at least 19 nucleotides, and that no primer dimers or secondary structures would form. This latter feature was verified using the tool Multiple Primer Analyzer – ThermoFisher Scientific (available at https://www.thermofisher.com/pt/en/home/brands/thermo-scientific/molecular-biology/molecular-biology-learning-center/molecular-biology-resource-library/thermo-scientific-web-tools/multiple-primer-analyzer.html). The melting temperature of each primer was determined using the tool Tm Calculator – New England Biolabs (available online at https://tmcalculator.neb.com/#!/main). A total of 24 primers were designed, allowing for the amplification of DNA fragments of similar size (from 881 bp for HPyV3, to 1175 bp for HPyV10) in the 2nd-round of amplification. The primers designed during this study are summarized in Table 2.2.

Table 2.2. List of primers used for the HPyV screening in the collected water samples.

Reaction		Primers (5'-3')	Number of mismatches[a]										
			HPyV1	HPyV2	HPyV3	HPyV4	HPyV5	HPyV6	HPyV7	HPyV8	HPyV9	HPyV10	HPyV11
PCR A	1st round	HPyV1/2Fo: CTGCTCCTCAATGGATGTTGC (from 1489 to 1509)[b]	0	0	-	-	10	-	-	7	9	7	17
		HPyV1/2Ro: ATCATRTCTGGGTCCCCTGGAAG (from 2606 to 2584)[b]	0	0	-	-	8	-	-	4	7	7	8
		HPyV5Fo: GAAAATAGCTTGCTGCATTCTG (from 879 to 900)[c]	35	15	-	-	0	-	-	7	10	10	14
		HPyV5Ro: GGGCCCACTCCATTCTCATC (from 1979 to 1960)[c]	5	6	-	-	0	-	-	2	2	5	8
		HPyV8/9Fo: AGGAGGRGCAMATCAAAGAG (from 1198 to 1217)[d]	3	2	-	-	9	-	-	0	0	3	10
		HPyV8/9Ro: ATAAAYTCTGACTTCTTCMAC (from 2336 to 2316)[d]	6	5	-	-	4	-	-	0	0	10	8
		HPyV10/11Fo: CCTGGATAYAGACAMTTTSA (from 806 to 825)[e]	9	17	-	-	10	-	-	11	13	0	0
		HPyV10/11Ro: TTAMAGGATAAGGATTTCTVA (from 2334 to 2313)[e]	7	8	-	-	6	-	-	5	1	0	1
		Expected amplicon size (bps)[f]	1113	1094	-	-	1101	-	-	1139	1136	1529	1474
	2nd round	HPyV1/2Fi: GTACGGGACTGTAACACCTGC (from 1526 to 1546)[b]	0	0	-	-	12	-	-	8	8	11	14
		HPyV1/2Ri: CCATACATAGGCTGCCCATC (from 2534 to 2515)[b]	0	0	-	-	7	-	-	7	7	3	7
		HPyV5Fi: CAATCAAACCTAGTGAATCTG (from 951 to 971)[c]	31	18	-	-	0	-	-	12	10	13	10
		HPyV5Ri: GGATCAGGACACCATACTTC (from 1859 to 1840)[c]	7	5	-	-	0	-	-	7	5	6	4
		HPyV8/9Fi: GGWTTGTATGGTGATATAAC (from 1250 to1269)[d]	7	6	-	-	9	-	-	0	0	5	7
		HPyV8/9Ri: ATTAAARTAYCTAGGTAGGCCTCT (from 2192 to 2170)[d]	4	5	-	-	6	-	-	0	0	5	5
		HPyV10/11Fi: AGAGCTTTTTGGGARGCTKT (from 881 to 900)[e]	8	7	-	-	14	-	-	6	10	0	1
		HPyV10/11Ri: CCCAGGCCTCYACWGGATAR (from 2055 to 2036)[e]	3	6	-	-	5	-	-	4	5	0	1
		Expected amplicon size (bps)[f]	989	985	-	-	909	-	-	943	940	1175	1136

(Continues on the next page).

Reaction		Primers (5'-3')	Number of mismatches[a]										
			HPyV1	HPyV2	HPyV3	HPyV4	HPyV5	HPyV6	HPyV7	HPyV8	HPyV9	HPyV10	HPyV11
PCR B	1st round	HPyV3/4Fo: GGACGTGTTCAATAGAATTGC (from 980 to 1000)[g]	-	-	0	0	-	10	8	-	-	-	-
		HPyV3/4Ro: CCAATGCCATTTTCATCCAA (from 2282 to 2263)[g]	-	-	0	0	-	5	4	-	-	-	-
		HPyV6/7Fo: GACTCGGCCCAAGARTTGG (from 708 to 726)[h]	-	-	9	13	-	0	0	-	-	-	-
		HPyV6/7Ro: GCACCTGTGGCTTCTGRGG (from 2220 to 2202)[h]	-	-	7	6	-	0	0	-	-	-	-
		Expected amplicon size (bps)[f]	-	-	1303	1318	-	1513	1513	-	-	-	-
	2nd round	HPyV3/4Fi: CATCATATTACAATRCGGGG (from 1015 to 1034)[g]	-	-	0	0	-	8	7	-	-	-	-
		HPyV3/4Ri: GTTTCCATTCTRTACAGCTC (from 1895 to 1876)[g]	-	-	0	0	-	7	5	-	-	-	-
		HPyV6/7Fi: TGGCACTTCAAYTGTGGTTG (from 738 to 757)[h]	-	-	27	18	-	0	0	-	-	-	-
		HPyV6/7Ri: WCCAATKACATCCAAGGGGC (from 1730 to 1711)[h]	-	-	16	5	-	0	0	-	-	-	-
		Expected amplicon size (bps)[f]	-	-	881	893	-	993	1002	-	-	-	-

K (G or T), M (A or C), R (A or G), S (C or G), W (A or T), Y (C or T), V (A, C or G)

This work is the reference for all the primers listed.

[a] According to the GenBank® reference sequences for each HPyV. The number of nucleotides in indels was counted as a mismatch.

[b] Positions numbered according to the HPyV1 GenBank® reference sequence NC_001538.

[c] Positions numbered according to the HPyV5 GenBank® reference sequence NC_010277.

[d] Positions numbered according to the HPyV8 GenBank® reference sequence NC_014361.

[e] Positions numbered according to the HPyV10 GenBank® reference sequence NC_018102.

[f] Estimated according to the GenBank® reference sequence for each HPyV.

[g] Positions numbered according to the HPyV3 GenBank® reference sequence NC_009238.

[h] Positions numbered according to the HPyV6 GenBank® reference sequence NC_014406.

2.5 Human polyomavirus screening

For the detection of several different HPyVs in the DNA extracts obtained from the collected water samples, two different nested touch-down multiplex PCR protocols targeting part of VP1's and VP2's coding sequence (VP1-2) were designed. These were designated PCR-A, which aimed for the amplification of HPyV1, 2, 5, 8, 9, 10, 11, and PCR-B, which targeted the amplification of HPyV3, 4, 6, and 7. The primers used in each reaction are described in Table 2.2 and the cycling conditions are summarized in Table 2.3. For all the amplification reactions, the commercial mix NZYTaq 2x Green MasterMix (NZYTech, Portugal) was used.

After each round of amplification, the PCR products obtained were analyzed by gel electrophoresis on a 1% agarose gel with Tris-Acetate-EDTA (TAE) buffer 0.5x, using ethidium bromide (5 µg/mL) as a fluorescent dye. To determine the size of the amplified products, NZYLadder VI (NZYTech, Portugal) was used as a molecular marker.

Table 2.3. Cycling conditions and reaction mixes used in the amplification reactions for the HPyV screening in the collected water samples.

Reaction		PCR-A	PCR-B
1st-round	Master mix	1x	1x
	Forward primers (10 pmol/µL)	HPyV1/2Fo – 0.54 pmol/µL HPyV5Fo – 0.54 pmol/µL HPyV8/9Fo – 0.54 pmol/µL HPyV10/11Fo – 0.54 pmol/µL	HPyV3/4Fo – 0.56 pmol/µL HPyV6/7Fo – 0.56 pmol/µL
	Reverse primers (10 pmol/µL)	HPyV1/2Ro – 0.54 pmol/µL HPyV5Ro – 0.54 pmol/µL HPyV8/9Ro – 0.54 pmol/µL HPyV10/11Ro – 0.54 pmol/µL	HPyV3/4Ro – 0.56 pmol/µL HPyV6/7Ro – 0.56 pmol/µL
	DNA	3 µL	3 µL
	H_2O	-	-
	Total volume	37 µL	21 µL
	Cycling conditions	95 ºC – 3’ (1x); [95 ºC – 30’’; 55 ºC – 30’’ (-1 ºC per cycle); 72 ºC – 1’15’’] (10x); (95 ºC – 30’’; 45ºC – 30’’; 72 ºC – 1’15’’) (30x); 72 ºC – 7’ (1x); 4 ºC - ∞	95 ºC – 3’ (1x); [95 ºC – 30’’; 63 ºC – 30’’ (-1 ºC per cycle); 72 ºC – 1’15’’] (10x); (95 ºC – 30’’; 53 ºC – 30’’; 72 ºC – 1’15’’) (30x); 72 ºC – 7’ (1x); 4 ºC - ∞
2nd-round	Master mix	1x	1x
	Forward primers (10 pmol/µL)	HPyV1/2Fi – 0.95 pmol/µL HPyV5Fi – 0.95 pmol/µL HPyV8/9Fi – 0.95 pmol/µL HPyV10/11Fi – 0.95 pmol/µL	HPyV3/4Fi – 1 pmol/µL HPyV6/7Fi – 1 pmol/µL
	Reverse primers (10 pmol/µL)	HPyV1/2Ri – 0.95 pmol/µL HPyV5Ri – 0.95 pmol/µL HPyV8/9Ri – 0.95 pmol/µL HPyV10/11Ri – 0.95 pmol/µL	HPyV3/4Ri – 1 pmol/µL HPyV6/7Ri – 1 pmol/µL
	DNA	2 µL	2 µL
	H_2O	-	-
	Total volume	36 µL	20 µL
	Cycling conditions	95 ºC – 3’ (1x); [95 ºC – 30’’; 56 ºC – 30’’ (-1 ºC per cycle); 72 ºC – 1’15’’] (10x); (95 ºC – 30’’; 46 ºC – 30’’; 72 ºC – 1’15’’) (30x); 72 ºC – 7’ (1x); 4 ºC - ∞	95 ºC – 3’ (1x); [95 ºC – 30’’; 59 ºC – 30’’ (-1 ºC per cycle); 72 ºC – 1’15’’] (10x); (95 ºC – 30’’; 49 ºC – 30’’; 72 ºC – 1’15’’) (30x); 72 ºC – 7’ (1x); 4 ºC - ∞

’ indicates minutes, ’’ indicates seconds, and ∞ symbolizes undetermined time.

2.6 Purification of PCR products from agarose gels

After each amplification reaction, every time the gel electrophoresis revealed a band of the expected size, regardless of the presence of any spurious products, the band corresponding to the desired product was purified with the Zymoclean™ Gel DNA Recovery Kit (Zymo Research, USA). Firstly, the section of the gel containing the DNA of interest was excised and three volumes of agarose dissolving buffer (ADB) were added before incubation at 55 ºC (10 min). After the agarose was completely dissolved, the mix was transferred into a column (Zymo-Spin™ Column) where, after centrifugation, the DNA was

adsorbed. After washing the column twice with 200 µL of wash buffer, the DNA was eluted in 15 µL of elution buffer. The yield of the purification reaction was then assessed by electrophoretic analysis of a small fraction of the product (as described in section 2.5).

2.7 Molecular Cloning of DNA molecules in a plasmid vector

During this work, two different cloning systems were used. One of them – NZY-A PCR Cloning Kit (NZYTech, Portugal) – allows for the selection of bacterial clones carrying recombinant plasmid DNA molecules through a "blue-white screening" strategy, based on the α-complementation phenomenon. In this case, by exploiting the activity of the β-galactosidase, an enzyme that is coded by the *lacZ* gene in *E. coli*, together with a chromogenic metabolite, it is possible to determine if an insert is present or absent by the white or blue color of the bacterial colonies, respectively. The other cloning system used was the CloneJET™ Cloning Kit system (ThermoFisher Scientific, USA), which utilizes pJet2.1/blunt as a cloning vector, allowing for a selection of recombinant clones based on the presence of a gene that encodes for a restriction enzyme (*eco45IR*) in the plasmid. In the absence of an insert, the vector will re-circularize and the expression of the enzyme will ensue, leading to the destruction of the bacterial DNA and the consequential impediment of the replication of clones that internalized the plasmid. Conversely, in the presence of an insert, the *eco45IR* gene will be interrupted, thus allowing the propagation of bacteria transformed with the recombinant plasmid. In each case, the manufacturer's instructions were followed during the cloning protocol with the ligation mixtures being attained either after overnight incubation at 4 ºC, or for 30 min at room temperature, depending on which one of the kits was used.

In order to obtain recombinant clones, the aforementioned ligation mixtures were used for the transformation of *E. coli* NovaBlue (Novagen, USA) competent cells. The preparation and transformation of these cells were made as previously described by Chung *et al.* (107) with some modifications. In summary, the cells were first grown in liquid LB (Lysogeny Broth), supplemented with tetracycline (12 µg/mL), until the optical density of the bacterial culture reached 0.4 (measured at 600 nm). The culture was then transferred into a 50 mL Falcon® tube and centrifuged at 3,000 *g* for 10 min (4 ºC). After removing the supernatant fraction, the pellet was resuspended in 500 µL of cold TSS (Transform and Storage Solution, containing 10% PEG [Polyethylene glycol], 5% DMSO [Dimethyl sulfoxide] and 20 mM $MgCl_2$ in liquid LB). Subsequently, half of the ligation mixture was added to 150 µL of the suspension, followed by an incubation period of 30 min on ice, during which the mixture of cells and DNA was occasionally gently agitated. After this period, the mixture was put in a water bath at 42 ºC for 45 sec, before being quickly put back on ice for an additional 5 min. For the recovery of the

cells after the heat shock, 850 µL of SOC (Super Optimal Broth with Catabolite repression (108)) medium was added before incubation at 37 ºC for 1 h. Lastly, the mixture was plated on solid LB (1.5% agar) supplemented with ampicillin (100 µg/mL), X-Gal (40 µg/mL), and IPTG (0.2 mM).

2.8 Isolation of plasmid DNA by alkaline lysis method

The extraction of plasmid DNA was carried out following the alkaline lysis method, initially described by Birnboim and Doly (109), with some minor adaptions. Briefly, 2 mL of a saturated bacterial culture was centrifuged at 17,900 *g* for 2 min to pellet the cells in suspension. After discarding the supernatant, the sedimented cells were resuspended in 250 µL of TEG (25 mM Tris-HCl pH 8, 1 mM EDTA, 1% Glucose). To this mixture, an equal volume of lysis solution (0.2 M NaOH, 1.5% SDS) was added to promote cell lysis, followed by the addition of an equal volume of potassium acetate (3 M, pH 5.2) to neutralize the lysate, being this last step the one that allowed the separation of chromosomal DNA (which precipitates with the cell debris) from plasmid DNA (which remains in the soluble fraction). The tubes containing the resulting mixture were then centrifuged at 17,900 *g*, for 10 min, at room temperature. To precipitate the plasmid DNA, 750 µL of isopropanol was added before a centrifugation step at 17,600 *g*, for 30 min, at room temperature. The resulting sediments were washed with 250 µL of ethanol at 70%, dried under vacuum, and afterwards resuspended in 40 µL of TE (10 mM Tris-HCl pH 8, 1 mM EDTA) supplemented with RNase A (50 µg/mL).

After extraction of the plasmid DNA, the recombinant clones were selected by electrophoretic analysis (as previously described in section 2.5). By comparing the migration profile with the one displayed by the vector, the recombinant DNA molecules that showed retarded migration were chosen for sequencing. Alternatively, when the yield of the plasmid DNA extraction was low, the desired screening was carried out by PCR (refer to section 2.10).

2.9 Plasmid DNA purification

In order to obtain the nucleotide sequence of the obtained recombinant plasmids, they were first extracted using the alkaline lysis method (as previously described in section 2.8) and then purified using chloroform ($CHCl_3$). To do so, to the existing volume of purified plasmid DNA in TE/RNase A, Tris-HCl (10 mM) was added to a total volume of 100 µL. Afterwards, 100 µL of $CHCl_3$ was added and the mixture thoroughly emulsified. After a centrifugation step at 17,600 *g*, for 3 min, at room temperature, the aqueous phase was collected and 10 µL of sodium acetate (NaOAc) was added prior to the addition

of an equal volume of isopropanol. The mixture was left to incubate for 5 min at room temperature before being centrifuged at 17,600 *g* for 30 min at room temperature. The sedimented nucleic acids were washed with 350 µL of ethanol, dried under vacuum, and lastly, resuspended in 20 µL of Tris-HCl (10 mM). The yield of the reaction was then assessed by electrophoretic analysis (refer to section 2.5), with 1 µL of the purified DNA extract.

2.10 PCR based protocol for screening of recombinant clones

To carry out the recombinant plasmid screening using PCR, the amplification reactions were performed using 2 µL of a dilution of plasmid DNA (obtained as described in section 2.8) as input. The reaction mix was made using 1x NZTaq Green MasterMix (NZYTech, Portugal), forward and reverse primers at a final concentration of 0.8 pmol/µL, and H_2O in a final volume of 25 µL. Depending on the vector utilized for the molecular cloning step (see section 2.7), the pairs of primers used were either the pJET1.2 forward sequencing primer (5' – CGACTCACTATAGGGAGAGCGGC – 3' [ThermoFisher Scientific, USA]) and the pJET1.2 reverse sequencing primer (5' – AAGAACATCGATTTTCCATGGCAG – 3' [ThermoFisher Scientific, USA]) or the T7promoterGGG (5' – TAATACGACTCACTATAGGG – 3') and M13pUC reverse primer (5' – CCAGGGTTTTCCCAGTCACGAC – 3').

The thermocycling conditions for this protocol consisted of an initial denaturation step at 95 ºC for 3 min, followed by 35 amplification cycles comprised of a denaturation step at 95 ºC for 1 min, an annealing step at either 55 ºC (when the pJET1.2 sequencing primers were used) or 50 ºC (when the T7promoterGGG and M13pUC reverse primers were used) for 30 sec, and an extension step at 72 ºC for 30 sec. After a final extension step at 72 ºC for 7 min, the samples were kept at 4 ºC and afterwards stored at -20 ºC until being screened by electrophoretic analysis (see section 2.5).

2.11 Editing and analysis of nucleotide sequences

Depending on the efficiency of the transformation process, a minimum of five and a maximum of 14 recombinant plasmids were sequenced for each sample. All the nucleotide sequences obtained during this work were produced at STABVIDA Lda. (Portugal), using the Sanger method (110). Whether the samples sent for sequencing were DNA fragments inserted in a cloning vector or a PCR product obtained from previously extracted plasmid DNA (refer to section 2.10), the primers specifically designed for the vectors were used for the attainment of most of the nucleotide sequences. In the event where more than

one primer had to be used to obtain the desired sequence, the software tool CAP3 (available at http://doua.prabi.fr/software/cap3) was utilized for the assembly of contigs. The analysis of the chromatograms resulting from each sequencing reaction and their editing was carried out using the software program BioEdit Sequence Alignment Editor version 7.2.5 (111). Preliminary taxonomic identification of the sequences was made using the tool BLASTn/MegaBlast from NCBI (National Center of Biotechnology Information: https://blast.ncbi.nlm.nih.gov/Blast.cgi), which allowed to search for homologous sequences in public access GenBank®/DDBJ/EMBL sequence databases.

For the analysis of the obtained sequences, five nucleotide datasets were constructed using homologous whole-genome and protein coding (VP1-2) reference sequences available at the GenBank® database. These sequences were identified via their accession number or added to the dataset as a result of a similarity search using either BLASTn or NCBI-Virus (available at https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/). The five datasets included four HPyV-specific datasets (HPyV1, HPyV2, HPyV5, and HPyV6), and one dataset composed of reference whole-genome sequences for all the existing HPyVs. In every case, only the sequences (obtained in this work) with a minimum length of 650 nt were included in these datasets, with 65 out of a total of 73 DNA sequences being analyzed by phylogenetic reconstruction. The multiple alignment nucleotide sequences were performed with MAFFT version 7 (105), using the iterative alignment option G-INS-i. The obtained alignments were then edited using Gblocks (112), resorting to the options that allowed for a less restrictive selection, and, to ensure no major errors could be found in the alignment, the multiple sequence alignments were further verified using BioEdit 7.2.5 (111).

Before beginning the phylogenetic analysis, the phylogenetic signal for each dataset was determined by likelihood-mapping, using TREE-PUZZLE version 5.3 (113). To assess if different phylogeny reconstruction approaches could lead to phylogenetic trees varying in topology, phylogenetic analyses were carried out by two distinct approaches. Hence, the analysis was done using the maximum likelihood (ML) optimization criterion as well as Bayesian inference, with the best fitting evolution model being GTR+ Γ+I, as determined by IQ-TREE (114). For the maximum likelihood analysis, the topological stability of the obtained trees was assessed with 1,000 bootstraps and 1,000 iterations of the approximate likelihood-ratio test (aLRT), with the analysis being carried out on an IQ-TREE run on an Ubuntu server. On the other hand, for the Bayesian approach, the BEAST version 1.10 software package (115) was used. In each case, the default priors were assumed, with the exception of the relaxed uncorrelated lognormal molecular clock model, which was suggested by the ML Clock Test provided in MEGA X, as well as the smooth skyline demographic prior (116, 117). For every phylogenetic reconstruction, two independent runs of the Markov-Chain Monte-Carlo (MCMC) method were performed, with tree sampling occurring for $1x10^8$ generations. The first 10% of sampled trees were

discarded as burn-in, and the results attained from the independent runs were combined as a maximum clade credibility tree using logCombiner and TreeAnnotator. To determine whether the sampling remained random after the removal of burn-in, as well as to verify the convergence of the sampling made by both MCMC runs, Tracer software version 1.7.1 (available at http://beast/bio/ed.ac.uk/Tracer) was used. All the obtained phylogenetic trees were visualized and graphically manipulated with FigTree version 1.4.2 software (available at http://tree.bio.ed.ac.uk/software/figtree/).

To analyze the presence of divergent sequences, and to further inspect the phylogenetic relationships presented by the obtained trees, split graphs (also known as networks) were constructed using the NeighborNet algorithm implemented in SplitsTree 4 software (118). The distance matrixes were corrected with the HKY model and the resulting graphs (hereafter NNn, from NeighborNet networks) edited using the same software. Complementary genetic assessments explored the use of Principal Coordinate Analysis, which was carried out using PCOORD (available at https://www.hiv.lanl.gov/content/sequence/PCOORD/PCOORD.html). To ensure none of the sequences obtained during this study were possible recombinants, RDP4 (119) was used to determine the occurrence of putative genetic recombination events.

All the nucleotide sequences obtained during this study were deposited to DDBJ dataset under the accession numbers LC636333-LC636405.

2.12 Next Generation Sequencing (NGS) analysis

Sequence analysis using an NGS approach was performed on a pooled sample containing ten purified PCR products obtained from DNA extracts from six wastewater samples and one environmental sample. This included fragments obtained using both PCR-A and PCR-B protocols, with three of the ten PCR products corresponding to the ones obtained with the latter. The DNA extracts implicated in this analysis are marked in Table 2.1.

Before beginning the library preparation step, the pooled sample was spiked with 5% of bacteriophage PhiX174 DNA, which worked as an internal control for the reaction. The DNA was then fragmented by sonication, using a Bioruptor® (Diagenode, Belgium), and the DNA library prepared using the NEBNext Ultra II Library Prep Kit (New England Biolabs, USA). In brief, the protocol began with the end repair of the previously obtained DNA fragments, using the NEBNext Ultra II End Prep enzyme mix and reaction buffer. This step ensured that each fragment is free of overhangs, with the 5' extremity possessing a phosphate group, and the 3' extremity containing a dAMP molecule that is essential for the adaptor (NEBNext Adaptor) ligation. Afterwards, the adaptor was incorporated in each

DNA molecule, with the resulting adaptor-ligated DNA pairs subsequently purified using magnetic beads (NEBNext Sample Purification Beads). The fragments were then amplified by PCR and the resulting enriched DNA library sequenced using a MiSeq Illumina system. Lastly, the analysis of the obtained paired-end reads (150 bp x 2) was carried out by the Genome Detective Virus Tool software (120), thus filtering the low-quality reads, trimming the part of the sequence that corresponds to the adaptor, as well as assembling the contigs and identifying the organism to each they belong, readily separating viral from non-viral sequences.

3 | Results

3.1 Primer and PCR protocol design

This study aimed for the detection and genetic analyses of HPyV genomic sequences obtained from a total of 17 water samples. These included nine influent wastewater and eight environmental samples collected at 14 different sites spread across the AML at five time-points: October of 2018, April 2019, July, October, and November of 2020. To do so, it was first necessary to design primers that targeted a section of the late region of the viral genome, as we searched for a conserved section of the genome other than the one used for taxonomical identification, as determined by ICTV. Additionally, it was also necessary to design a PCR protocol that allowed for the amplification of the viral DNA. A dataset composed of viral reference sequences of the structural-protein coding region of 15 lineages of HPyV was constructed, with a minimum of one and a maximum of six sequences being used to represent each viral lineage. The construction of the datasets of viral sequences took into consideration the uncertainty surrounding the human-specificity and/or low prevalence in the human population of polyomaviruses HPyV12, HPyV13, HPyV14, and QPyV (40, 95, 99, 100), which were not considered during primer design, thus leading to the exclusion of these four viruses from our analysis.

A preliminary phylogenetic analysis using this dataset demonstrated that the viral sequences segregated into two major monophyletic groups. The first group included HPyV1, HPyV2, HPyV5, HPyV8, HPyV9, HPyV10, HPyV11, HPyV12, HPyV13, and HPyV14, while a smaller second group clustered together HPyV3, HPyV4, HPyV6, HPyV7, and the putative QPyV (Figure 3.1). These two phylogenetic lineages were named Lineage-A (L-A) and Lineage-B (L-B), respectively, and they were used to guide the design of the amplification primers. Considering the sub-clusters occurring inside each lineage, by using a software tool that allows for primer design in combination with the visual inspection of multiple sequence alignments of the structural-protein coding regions of HPyVs, a total of 24 primers were proposed (Table 2.2).

Sixteen of the PCR amplification primers designed were used for the detection of the genomes clustering within L-A, while eight were used to amplify sequences of the HPyVs grouping in L-B. These primers allowed us to define two parallel nested touch-down multiplex PCR protocols, named PCR-A and PCR-B, separating the detection of HPyVs that were part of L-A from those that were included in L-B, respectively. For PCR-A, four pairs of primers were used in each amplification round and these targeted the amplification of HPyV1/HPyV2 (the slash indicates that the detection of these viruses was tentatively done simultaneously, utilizing the same pair of primers), HPyV5, HPyV8/HPyV9, and HPyV10/HPyV11, while only two pairs of primers were used per amplification round to target

HPyV3/HPyV4 and HPyV6/HPyV7 in PCR-B (Table 2.2). In each case, after two amplification rounds, it was possible to obtain DNA amplicons of similar size. Moreover, by aligning each primer to a reference HPyV sequence, it was possible to determine that the number of mismatched positions was usually higher when the primer sequence was aligned with a heterologous HPyV (from a minimum of 2 to a maximum of 35 positions), contrasting with the minimal (n = 1) to nonexistent mismatching with the respective homologous HPyV (Table 2.2). This observation appeared to further support an *a priori* (as suggested by bioinformatic analysis) primer specificity, indicating the unlikelihood of cross-amplification due to primer hybridization to non-homologous targets.

Even so, since the starting samples contained DNA from multiple sources (including expected large amounts of non-viral/heterologous viral DNA), there are numerous possible targets with which the HPyV-specific primers might hybridize. Furthermore, the high number of primers used in each PCR reaction and the occasional presence of degenerate positions in these primers may result in spurious amplification. Thus, to minimize nonspecific hybridization, we used multiplexed touch-down PCR protocols. Additionally, due to the expected low viral titer in the collected water samples, a nested-PCR approach was also chosen. In combination, the implemented temperature gradient in the initial amplification cycles and the two consecutive amplification rounds would ensure an increase in specificity and sensitivity of the amplification reactions (121, 122).

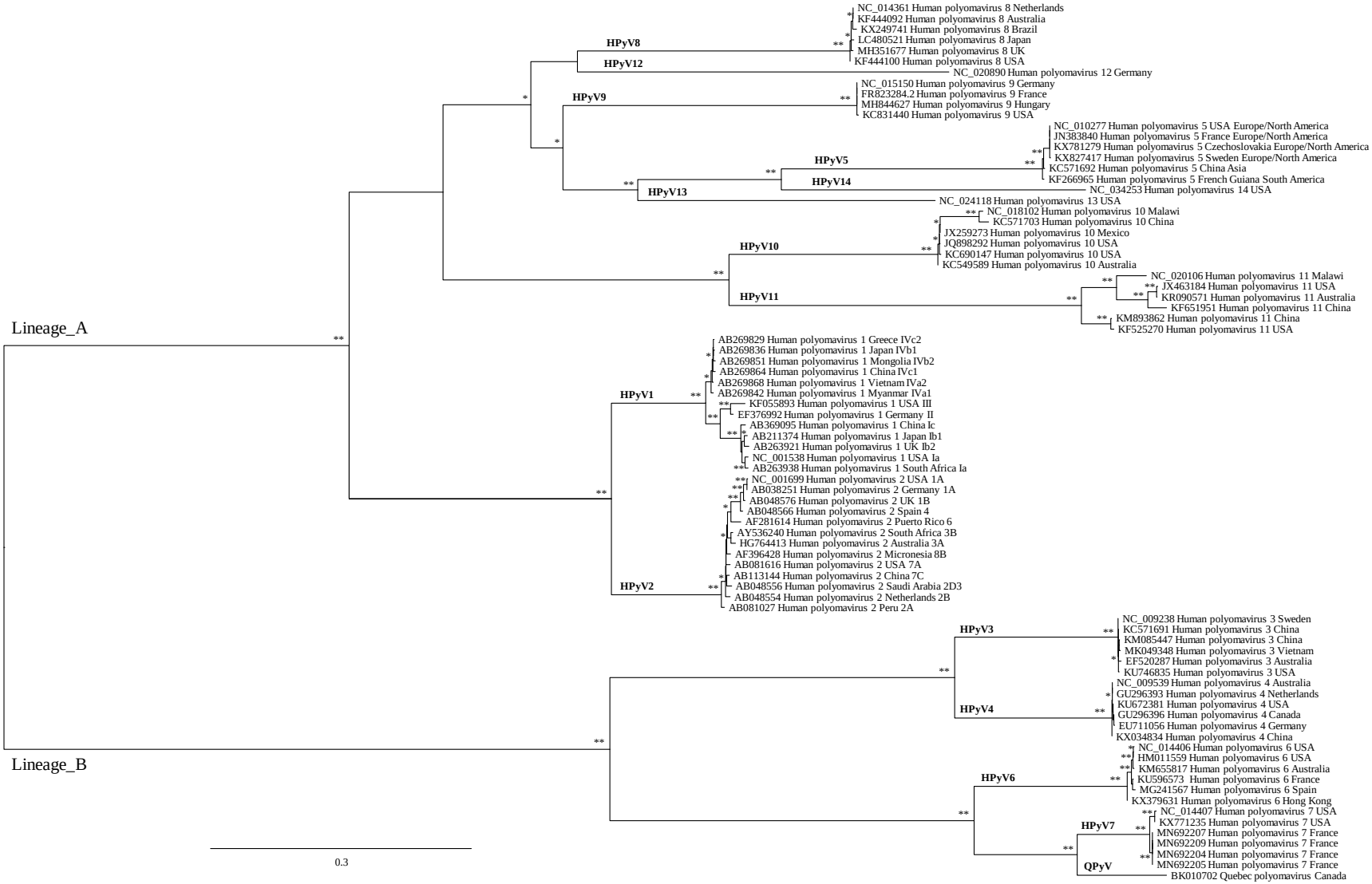


Figure 3.1. Phylogenetic analysis of nucleotide sequences of the structural-protein coding region of 15 different lineages of HPyV. Each sequence is identified by its accession number, HPyV type, and country of origin. When applicable, the viral genotype and/or subtype is also indicated. The number of "*" indicates the number of methods that support the demonstrated topology, considering 75% and above as relevant aLRT and bootstrap values.

3.2 PCR performance assessment – Multiplex versus singleplex approach

To determine whether our designed primers would support the detection of HPyV DNA, as well as to assess if the use of a multiplex approach impaired the expected performance of the PCR reaction, a preliminary analysis was performed using a HPyV2 DNA extract (previously obtained from a clinical isolate) as a positive control. To accomplish this, alongside PCR-A and PCR-B, an extra singleplex reaction was set up, where only primers HPyV1/2Fo and HPyV1/2Ro or HPyV1/2Fi and HPyV1/2Fi, for the 1st-round and 2nd-round respectively, were added.

An electrophoretic analysis of the 1st-round of amplification products showed that it was only possible to detect the DNA fragment of expected size (approximately 1000 bps) in an aliquot of an amplification reaction where only one pair of primers was used (lane 3, Figure 3.2). On the other hand, the multiplex approach allowed for the detection of HPyV2 DNA (lane 7, Figure 3.2) after only the 2nd-round of amplification. This appeared to suggest that, even in the presence of a known homologous matrix, the combination of multiple primers in the multiplexed reaction led to an apparent loss in its sensitivity.

Additionally, as products of the 2nd-round of the amplification reaction, it was possible to observe the presence of spurious amplicons in the lanes corresponding to the multiplex and singleplex reactions (lanes 7 and 8, respectively, Figure 3.2), with the first showing a slight smear. Considering that in the singleplex reaction only one unspecific product is visible, it seemed that the presence of multiple primers might also affect the specificity of the intended amplification. However, but as expected, PCR-B was unable to detect the HPyV2 DNA in both amplification rounds (lanes 5 and 10, Figure 3.2), demonstrating that no cross-amplification reactions had occurred.

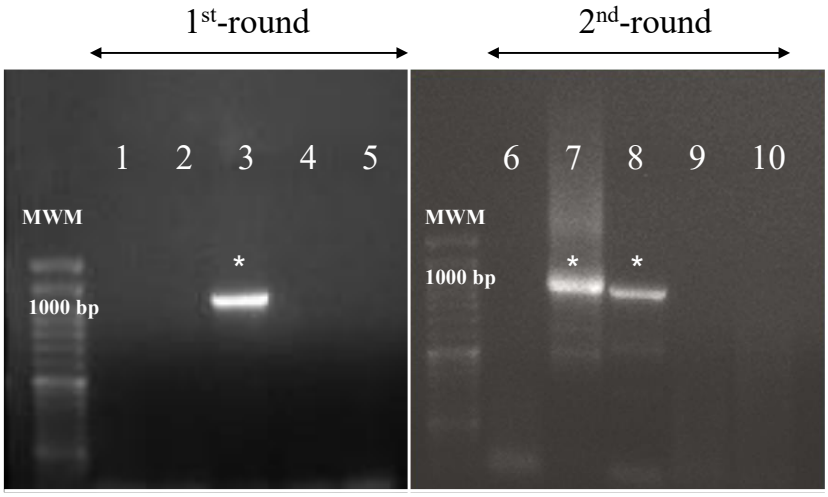


Figure 3.2. Electrophoretic analysis of the amplification results obtained using a multiplex and singleplex approach with a HPyV2 DNA extract. Lanes 1 and 6 correspond to the 1st and 2nd-round of the negative controls (H_2O) of PCR-A, respectively; lanes 2 (1st-round) and 7 (2nd-round) correspond to the multiplex reactions of PCR-A; lanes 3 (1st-round) and 8 (2nd-round) to the singleplex reaction (using the thermocycling conditions of PCR-A); lanes 4 and 9 correspond to the 1st and 2nd-round of the negative controls (H_2O) of PCR-B, respectively, and lanes 5 (1st-round) and 10 (2nd-round) of PCR-B. The input for every reaction (except for the negative controls) was a HPyV2 DNA extract (diluted 1:10,000) in the 1st-round and a 1:50 dilution of the 1st-round product in the 2nd-round. The lane marked with "MWM" corresponds to the NZYLadder VI (NZYTech, Portugal). The "*" indicates the fragments of expected size: 1100 bp and 990 bp (approximate size) for the 1st and 2nd-round, respectively.

3.3 Detection and genetic analysis of HPyV nucleotide sequences

3.3.1 HPyV screening in the collected water samples

After the design of primers and the setup of the PCR protocols, the collected water samples were screened for the presence of HPyV DNA. The VLPs present in both types of samples were first concentrated using a skimmed milk flocculation protocol and the DNA was afterwards extracted using a commercial DNA extraction kit (described in sections 2.2 and 2.3 in "Materials and Methods"). The obtained DNA extracts were then used as input for both PCR-A and PCR-B, and the amplification results assessed by electrophoretic analysis (refer to section 2.5). Regardless of the PCR reaction involved, we considered an amplification as successful whenever a fragment of the expected size was observed on the agarose gel (refer to Figure 3.3 for an example).

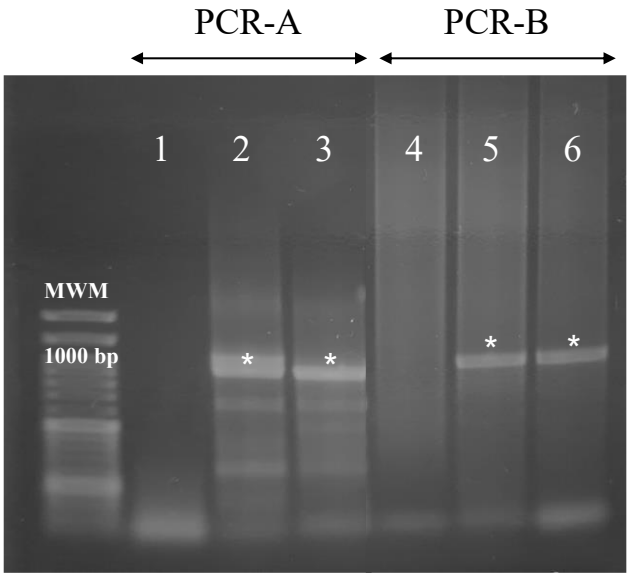


Figure 3.3. Electrophoretic analysis of the amplification products from the 2nd-round of PCR-A and PCR-B, obtained using two wastewater samples. The lane marked with "MWM" corresponds to the NZYLadder VI (NZYTech, Portugal). Lanes 1 and 4 correspond to the negative control (H_2O) for PCR-A and PCR-B, respectively; lanes 2 (2nd-round of PCR-A) and 5 (2nd-round of PCR-B) correspond to PCR reactions made with a DNA extract obtained from the water sample collected in wastewater treatment plant B in 2018, and lanes 3 (2nd-round of PCR-A) and 6 (2nd-round of PCR-B) correspond to the amplification reactions made with a DNA extract obtained from the sample collected in the same wastewater treatment as the latter in 2019. The "*" indicates the fragments of expected size (approximately 1000 bp PCR-A and ranging between 900 and 1000 bp for PCR-B).

For PCR-A, from a total of 17 water samples, ten were deemed positive (58.8%) for the presence of HPyV DNA, whilst with PCR-B, amplification was only considered successful in association with three wastewater samples (17.6%) (Table 3.1). Although the presence of spurious amplicons sometimes occurred after the 2nd-round of PCR-A, PCR-B never yielded non-specific amplification products, while specific amplicons were only ever obtained using the HPyV6/HPyV7-specific primers.

Taking into account all the water samples used, those corresponding to wastewater had an associated higher amplification success than the environmental samples, with only one of the latter (Trancão River, site 8 in Figure 2.1) considered positive for the detection of HPyV DNA (12.5%). In contrast, 100% of the wastewater samples were deemed positive for the presence of HPyV DNA using a combination of multiplex and singleplex approaches (mentioned later in this section). Furthermore, one other environmental sample (Ribeira das Lajes, site 3 in Figure 2.1) appeared to be positive for the presence of HPyV DNA, but we could not confirm this result as the individual recombinant plasmids (containing the obtained amplicons; see below) that were obtained and analyzed did not seem to contain inserts with a match in the GenBank® database. This result was as such deemed as a "false-positive" (Table 3.1). A similar case occurred with two sewage samples that were collected in wastewater treatment plant C and

wastewater treatment plant D, where all the sequenced recombinant molecules contained either bacterial DNA (from species *Aliarcobacter cryaerophilus* and *Cloacibacterium normanense*) or crustacean DNA (from a freshwater shrimp, species *Macrobrachium nipponense*), respectively. Hence, another singleplex assay was done, where every single primer pair was used individually against the DNA extracts obtained from both the aforementioned wastewater samples, as well as from other two environmental water samples that had been deemed "negative" by the previously established multiplex approach (Barreiro (site 10) and Industrial area of Setúbal (site 14)). These last two samples were chosen due to their proximity to industrial discharge pipes and a wastewater treatment plant, respectively, which are features that do not extend to Ribeira das Lajes (site 3). This would possibly increase the chances of finding HPyV DNA in environmental waters.

Table 3.1. Water samples HPyV screening with the touch-down multiplex PCR-based protocols.

Sample type	Collection site	PCR-A	PCR-B
Environmental	Algés (Tagus River)	-	-
	Barreiro (Tagus River)	-	-
	Industrial Area of Setúbal (Sado River)	-	-
	Lizandro River	-	-
	Port of Setúbal (Sado River)	-	-
	Ribeira das Lajes	+[c]	-
	Ribeira de Carenque	-	-
	Trancão River	+	-
Wastewater	Wastewater treatment plant A[a]	+	-
	Wastewater treatment plant B[b]	+	+
	Wastewater treatment plant C	+[c]	-
	Wastewater treatment plant D	+[c]	-
	Wastewater treatment plant E	+	-
	Wastewater treatment plant F	+	+

[a] Three samples were collected from this site, with all displaying the same result after the screening.
[b] Two samples were collected from this site, with all displaying the same result after the screening.
[c] Later deemed as a "false-positive".

From the total of four water samples that were analyzed using the singleplex approach, only the one collected in the Industrial area of Setúbal (site 14) remained negative for the presence of HPyV DNA, regardless of the PCR protocol used. On the contrary, the environmental water sample collected in Barreiro (site 10) was deemed positive for the presence of HPyV1 and/or HPyV2. However, due to the low amplification yield, we were unable to confirm it as HPyV DNA by Sanger sequencing, and as such, this result will no longer be accounted for henceforth in this work. For the wastewater samples, the singleplex approach allowed to detect HPyV1/HPyV2 and HPyV6/HPyV7 DNA in both, while HPyV5

DNA was only amplified in the wastewater treatment plant C sample (Table 3.2). These positive results were all confirmed by either direct sequencing of the obtained HPyV5 and HPyV6 amplicons, or by the analysis of individual recombinant plasmids (carrying HPyV1/HPyV2 DNA inserts). All the sequences obtained during this assay were included in the genetic analysis that followed, with the exception of the two HPyV6 sequences, which due to their low quality, were only used to confirm the presence of HPyV DNA using BLASTn.

Regarding the sensitivity of the singleplex reaction, the fact that all these samples were first regarded as negative by the multiplex PCR approach (PCR-A and/or PCR-B), or later by the confirmation of the non-viral nature of the cloned amplicons, suggests that the presence of multiple primers, possibly competing for the same matrix, leads to a decrease in the sensitivity of the amplification protocol. Moreover, the presence of amplification products after only the 1st-round of amplification with the singleplex approach (Table 3.2) further demonstrated the increase of sensitivity of this format. Indeed, for other water samples where the amplification of HPyV DNA was considered successful, we were only able to observe the presence of amplicons after the 2nd-round of amplification with the multiplex format. These results are corroborated by the one obtained using HPyV2 DNA from a clinical isolate (refer to section 3.2).

Table 3.2. Analysis of four water samples using a singleplex approach.

1st-round

	PCR-A				PCR-B	
	HPyV1/2	HPyV5	HPyV8/9	HPyV10/11	HPyV3/4	HPyV6/7
WWTP C	+	+	-	-	-	+
WWTP D	+	-	-	-	-	+
Barreiro	-	-	-	-	-	-
Setúbal (IA)	-	-	-	-	-	-

2nd-round

	PCR-A				PCR-B	
	HPyV1/2	HPyV5	HPyV8/9	HPyV10/11	HPyV3/4	HPyV6/7
WWTP C	+	+	-	-	-	+
WWTP D	+	-	-	-	-	+
Barreiro	+	-	-	-	-	-
Setúbal (IA)	-	-	-	-	-	-

"WWTP" stands for "Wastewater Treatment Plant", while "IA" stands for "Industrial Area".

Lastly, an increase in specificity also seemed to occur when only a single primer-pair was used, considering that with the singleplex approach, we were able to detect HPyV viral DNA in both wastewater samples (instead of non-viral DNA). Indeed, by reducing the number of primers in the amplification reaction, we are simultaneously reducing their competition for the non-homologous templates, which might significantly impact the performance of the intended amplifications.

Altogether, a total of 73 nucleotide sequences were obtained from the HPyV genome screening step. From these, 68 (93.15%) were obtained from the analysis of individual recombinant molecules, by use of a molecular cloning approach; four (5.48%) sequences were attained by a NGS analysis of a pooled sample composed of different previously obtained amplicons (discussed further down in this section), and one (1.37%) sequence was obtained by direct sequencing of an amplicon attained during the singleplex assay. Whenever possible, and especially in the case of the sequences obtained from the analysis of recombinant plasmids, their ends were inspected for the presence of the amplification primer sequences used. Due to the sequencing approach used, in those instances where the start/end of the chromatographs were unusable, it was impossible to properly inspect the extremities of the corresponding sequences and, as such, determine the primers used to amplify them. Even so, in the cases where it was indeed possible to verify which primers had been used in the amplification reaction, in nine of them (13.0%), one of the primers used was not the one expected *a priori*, taking into account the type of sequence. This translated into unexpected primer combinations being found at the ends of four HPyV1 sequences, three HPyV2 sequences, and two HPyV5 sequences. Surprisingly, in most cases, the heterologous primer found belonged to either the HPyV10/11 or HPyV8/9 primer combination.

After being edited (i.e., after the removal of the amplification primers), the obtained sequences were analyzed aiming at a first taxonomical identification of the viruses in question using BLASTn, with this preliminary identification being later confirmed for most sequences by phylogenetic reconstruction (as discussed below). However, before performing any sort of phylogenetic analysis, it was first necessary to verify if any of the Portuguese viral sequences corresponded to recombinant ones, as these should be discarded. This preliminary analysis is of utmost importance, considering that prior to suggesting that any sequence might be a representative of an undescribed genotype, it is necessary to ensure that it is not an intragenic mosaic sequence that resulted from a recombination event between other sequences from the previously established genotypes. Furthermore, the accuracy of the phylogenetic tree may be impaired if these events go undetected (119). Thus, in order to accomplish this goal, RDP4, a recombination detection tool, was used. Nevertheless, no evidence that any of the Portuguese sequences corresponded to recombinant sequences was found.

Hence, after confirming that the sequences obtained in this work were in fact not recombinant sequences, a dataset comprised of 132 HPyV sequences (including 65 Portuguese sequences, thus excluding the sequences that were smaller than 650 nt from this analysis) was designed and subsequently used for the construction of a phylogenetic tree (Figure 3.4) that would confirm the assignment of most of the nucleotide sequences to a viral species. Considering the high phylogenetic signal of this dataset disclosed by a likelihood mapping analysis (95.60% of the random sequence quartets were resolved), the performed phylogenetic analysis undoubtedly supports the viral species assignment process.

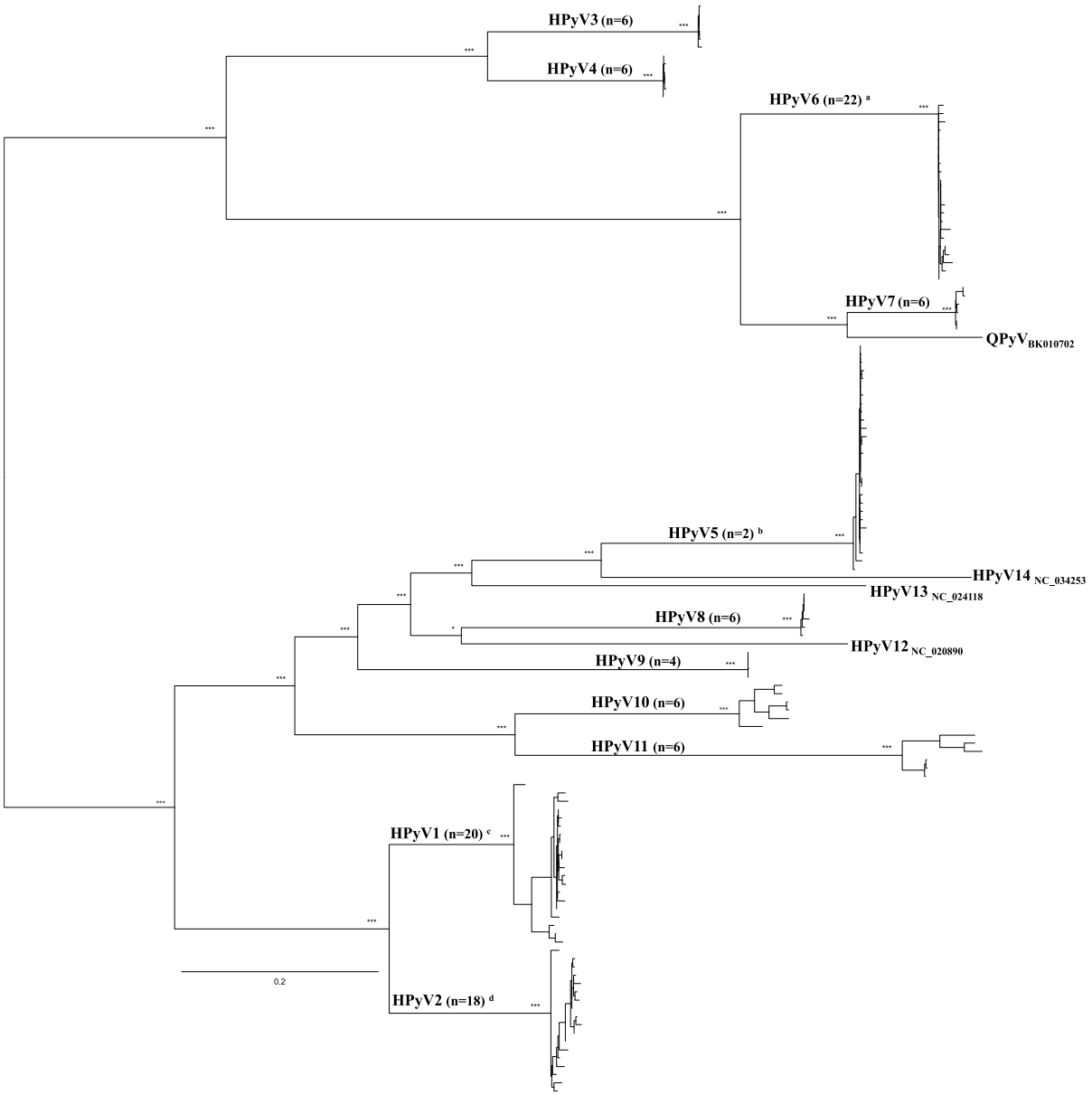


Figure 3.4. Phylogenetic analysis by maximum likelihood of nucleotide sequences of the structural-protein coding region of all species of HPyV. Above the main branches, the identity of each HPyV monophyletic cluster is indicated, with the respective number of sequences indicated between brackets. The letters in superscript refer to the number of sequences (> 650 nt) obtained during this work, with "a" equaling 16 HPyV6 sequences, "b" 23 HPyV5 sequences, "c" 14 HPyV1 sequences, and "d" 12 HPyV2 sequences. The reference sequences for HPyV12, HPyV13, HPyV14, and QPyV are individually identified by their accession numbers in their respective branch. The number of "*" indicates the number of methods that support the demonstrated topology, considering 75% and above as relevant aLRT and bootstrap values, and 0.80 and above for posterior probability, considering a Bayesian approach was used to further confirm the proposed topology. The bar indicates the average number of substitutions per site.

Overall, considering the heterogeneity of the population that inhabits the geographic area screened during this work, finding various HPyV viral types would not be unexpected. Indeed, in combination with the identification of the eight remaining sequences done exclusively by BLASTn, this analysis allowed to disclose the presence of sequences assigned to HPyV1 (n = 15, 20.55%), HPyV2 (n = 16, 21.91%), HPyV5 (n = 26, 35.62%), and HPyV6 (n = 16, 21.91%), having been found in six (35.3%), five (29.4%), eight (47.1%), and five (29.4%) water samples, respectively.

3.3.2 Viral DNA analysis by NGS

Although the experimental approach used allowed for the identification of the presence of HPyV1, HPyV2, HPyV5, and HPyV6 viral DNA in ten samples (9 wastewater/1 environmental) from the total of 17 that were collected, the hypothesis that some of the choices made in our viral DNA detection algorithm may have contributed to a biased result should be considered. Thus, we opted to perform an additional NGS analysis to ensure that the obtained results were not skewed by our experimental approach, and particularly by the molecular cloning step.

For this, a pooled sample of ten PCR-A and PCR-B purified amplicons obtained from six wastewater samples and one environmental sample (n_{total} = 7) collected in 2018, 2019, and 2020 was used to construct a DNA library. These samples were, for the most part, randomly selected, only making sure that both types of water samples were represented. After sequencing, the NGS results were analyzed using the software tool Genome Detective, revealing a total of 576,822 short-sequence reads that appeared to correspond to viral DNA (accounting for 87.7% of the obtained reads). From this number, 496,197 corresponded to HPyV-specific reads that were assembled into contigs, with 148,851 being assigned to HPyV1 (30.0%, coverage 21.3%), 111,781 to HPyV2 (22.5%, coverage 20.2%), 201,255 to HPyV5 (40.6%, coverage 17.5%) and 34,310 to HPyV6 (6.9%, coverage 20.3%) (Figure 3.7). This distribution of the HPyV sequences was supported by the fact that the sequencing of the PhiX174 DNA, used as an internal control, displayed an extremely high coverage (100%), resulting in a full-length sequence. The identification of HPyV1, HPyV2, HPyV5, and HPyV6 by NGS was congruent with the results obtained with individual plasmid analysis since these four viral types were also the only ones to be identified with the latter. Hence, the molecular cloning step did not appear to introduce any qualitative biases in our analysis. However, considering that this resulted from the analysis of purified amplicons, an accurate representation of the HPyVs that are present in the environment is yet to be achieved. Moreover, the comparison between these two approaches does not discard the possibility that the DNA cloning step could skew the results regarding any sequence quantitative representation amongst the recombinant plasmids analyzed.

Although the qualitative distribution of the HPyV sequences was equal to the one previously reported in this study, there is a quantitative difference when the distribution of HPyV reads is compared to the number of obtained unique sequences for each viral type (Figure 3.5). In fact, besides HPyV5 (which accounts for the biggest number of HPyV assigned reads as well as the highest number of obtained individual sequences), and HPyV2 (where the percentage of assigned reads is quite similar to the one of obtained sequences), HPyV1 and HPyV6 both display a numerical difference when the two different approaches are compared (Figure 3.5). For HPyV1, the number of reads was slightly higher than the number of obtained individual nucleotide sequences, while for HPyV6 the opposite occurred, as the percentage of reads attributed to this viral species is much lower than the one corresponding to the nucleotide sequences.

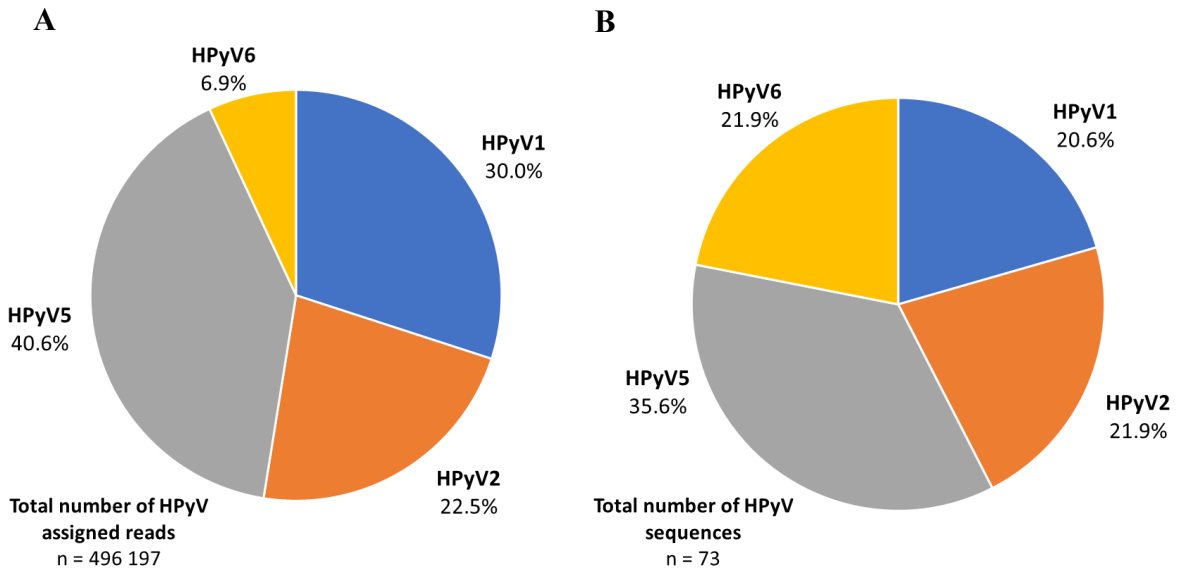


Figure 3.5. Graphical distribution of the short-sequencing reads (A) and unique nucleotide sequences obtained with the analysis of recombinant plasmids (B) assigned to four HPyVs. In both pie charts, each color represents one viral species: HPyV1 (blue), HPyV2 (orange), HPyV5 (grey), and HPyV6 (yellow).

3.3.3 Genetic characterization of HPyV sequences

After taxonomically identifying (assignment of a viral species to a given viral sequence) each one of the 73 nucleotide sequences obtained during this work, a molecular characterization of the HPyVs circulating the AML was performed by the construction of four HPyV-specific datasets. These datasets were used to create phylogenetic trees using a combination of maximum likelihood and Bayesian

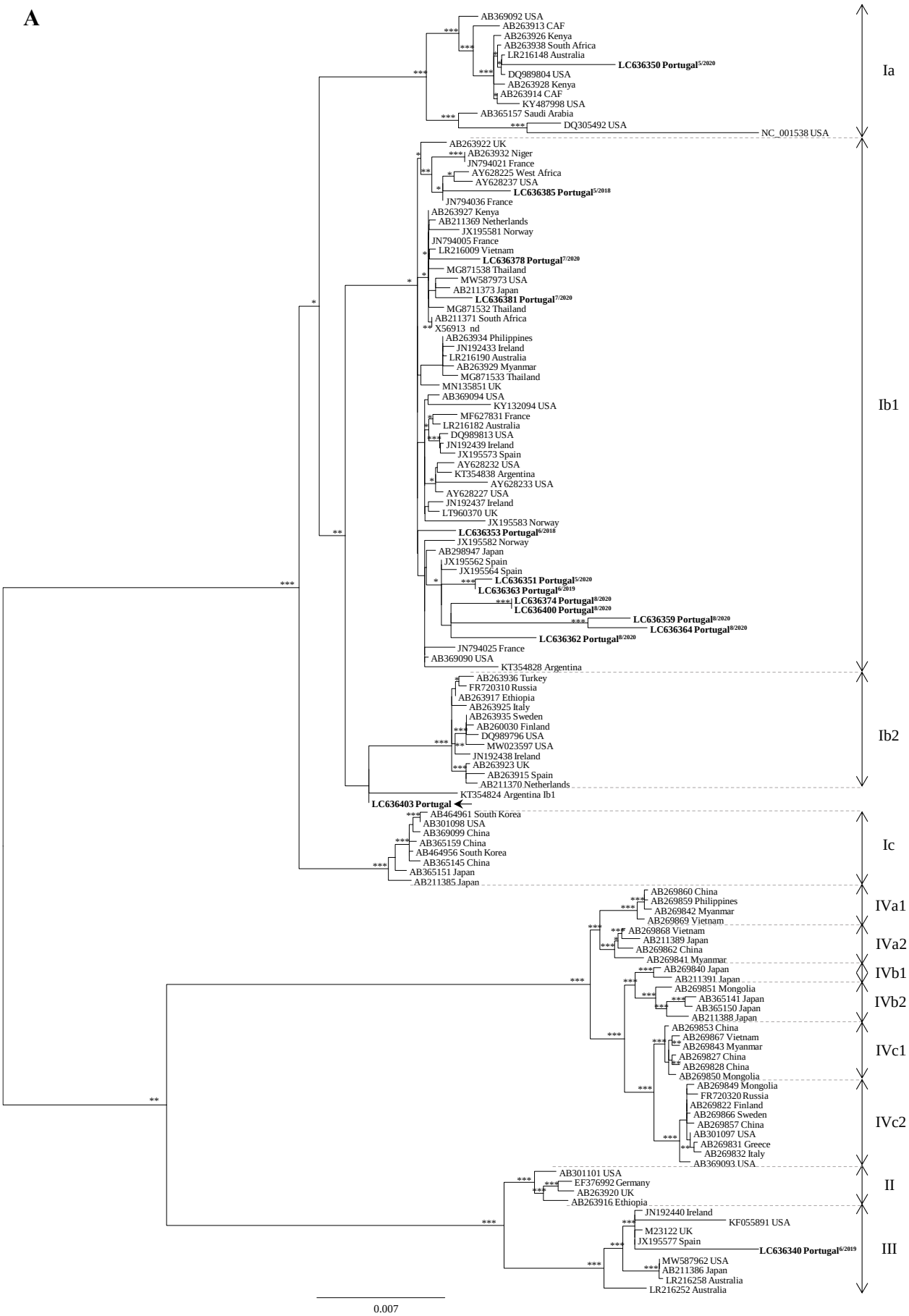
approaches, that allowed the assignment of a viral sequence to a genotype/subtype (Table 3.3). This analysis more accurately disclosed the genetic variability of the Portuguese sequences since, in the majority of the cases, these sequences were associated with different reference sequences assigned to different genotypes, with only a couple of the Portuguese sequences clustering together in a monophyletic group that did not include viral references (as the one observed in Figure 3.6A). Particularly for HPyV1 and HPyV2, the species-specific phylogenetic trees demonstrated that the Portuguese sequences did not conform to a single viral genotype, with most of these associations being highly supported by the phylogenetic reconstruction methods (Figure 3.6A and Figure 3.6B). Although the resolution of the HPyV5 tree was lower than the one observed for the abovementioned HPyVs (as there were more single segregating branches), its topological stability appeared to be sufficient to unambiguously define different genotypes (Figure 3.6C). However, the same did not occur for HPyV6, where the confirmation of the common shared ancestry of the two geographically linked clades previously identified by Torres et al. (58) (refer to section 1.8 of "Introduction") could not be confirmed (Figure 3.6D). Even so, this result could simply be due to the analyzed regions being different in both works. In fact, when looking at the phylogenetic signals of each species-specific dataset, a significant discrepancy becomes evident, with the percentage of resolved randomly generated sequence quartets being 90.22% and 89.62% for the HPyV1 and HPyV2 datasets, respectively, and only 47.56% and 43.22% for the HPyV5 and HPyV6 datasets, correspondingly. Hence, the aforementioned observation that not all trees allowed to properly attribute a sequence to a certain genetic type, came as no surprise.

Table 3.3. Distribution of the nucleotide sequences described in this work by HPyV species.

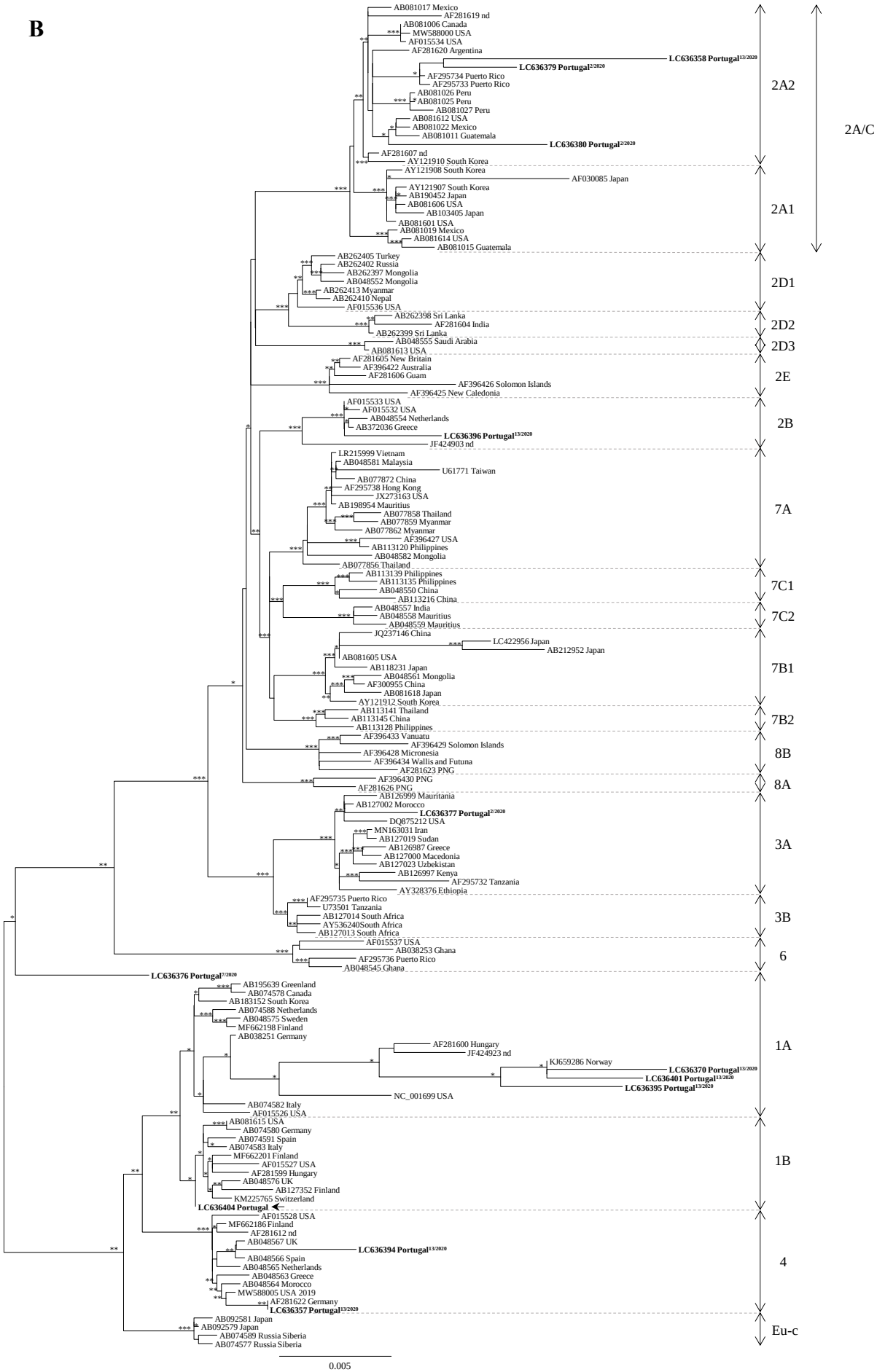
HPyV	Genotype/Subtype/Subgroup ($\%_{total}$; $\%_{per\ virus\ type}$)	Accession Number
HPyV1	Ia (n = 2/2.74%$_{total}$; 13.33%$_{HPyV1}$)	LC636348[5/2020,b], LC636350
	Ib1 (n = 11/15.07%$_{total}$;73.33%$_{HPyV1}$)	LC636351, LC636353, LC636359, LC636362, LC636363, LC636364, LC636374, LC636378, LC636381, LC636385, LC636400
	III (n = 1/1.37%$_{total}$;6.67%$_{HPyV1}$)	LC636340
	I-like (n = 1/1.37%$_{total}$;6.67%$_{HPyV1}$)	LC636403
HPyV2	1A (n = 3/4.11%$_{total}$; 18.75%$_{HPyV2}$)	LC636370, LC636395, LC636401
	1B (n = 3/4.11%$_{total}$; 18.75%$_{HPyV2}$)	LC636349[5/2020,b], LC636360[13/2020,b], LC636404
	2A2 (n = 3/4.11%$_{total}$; 18.75%$_{HPyV2}$)	LC636358, LC636379, LC636380
	2B (n = 1/1.37%$_{total}$; 6.25%$_{HPyV2}$)	LC636396
	3A (n = 2/2.74%$_{total}$; 12.50%$_{HPyV2}$)	LC636377, LC636399[11/2020,b]
	4 (n = 3/4.11%$_{total}$; 18.75%$_{HPyV2}$)	LC636357, LC636365[13/2020,b], LC636394
	9 (n = 1/1.37%$_{total}$; 6.25%$_{HPyV2}$)	LC636376
HPyV5	Africa (n = 8/10.96%$_{total}$; 30.77%$_{HPyV5}$)	LC636333, LC636335, LC636352, LC636355, LC636356, LC636384, LC636387[5/2018,b], LC636398
	Europe/North America (n = 17/23.29%$_{total}$; 65.38%$_{HPyV5}$)	LC636341, LC636342, LC636343, LC636347, LC636354, LC636369, LC636373, LC636375, LC636386, LC636388[5/2018,b], LC636389, LC636390, LC636391[5/2019,b], LC636392, LC636393, LC636397, LC636402
	Eu/NAm/Af (n = 1/1.37%$_{total}$; 3.85%$_{HPyV5}$)	LC636361
HPyV6	n.a.[a] (n = 16/21.91%$_{total}$; 100.00%$_{HPyV6}$)	LC636334, LC636336, LC636337, LC636338, LC636339, LC636344, LC636345, LC636346, LC636366, LC636367, LC636368, LC636371, LC636372, LC636382, LC636383, LC636405

[a] Not applicable.

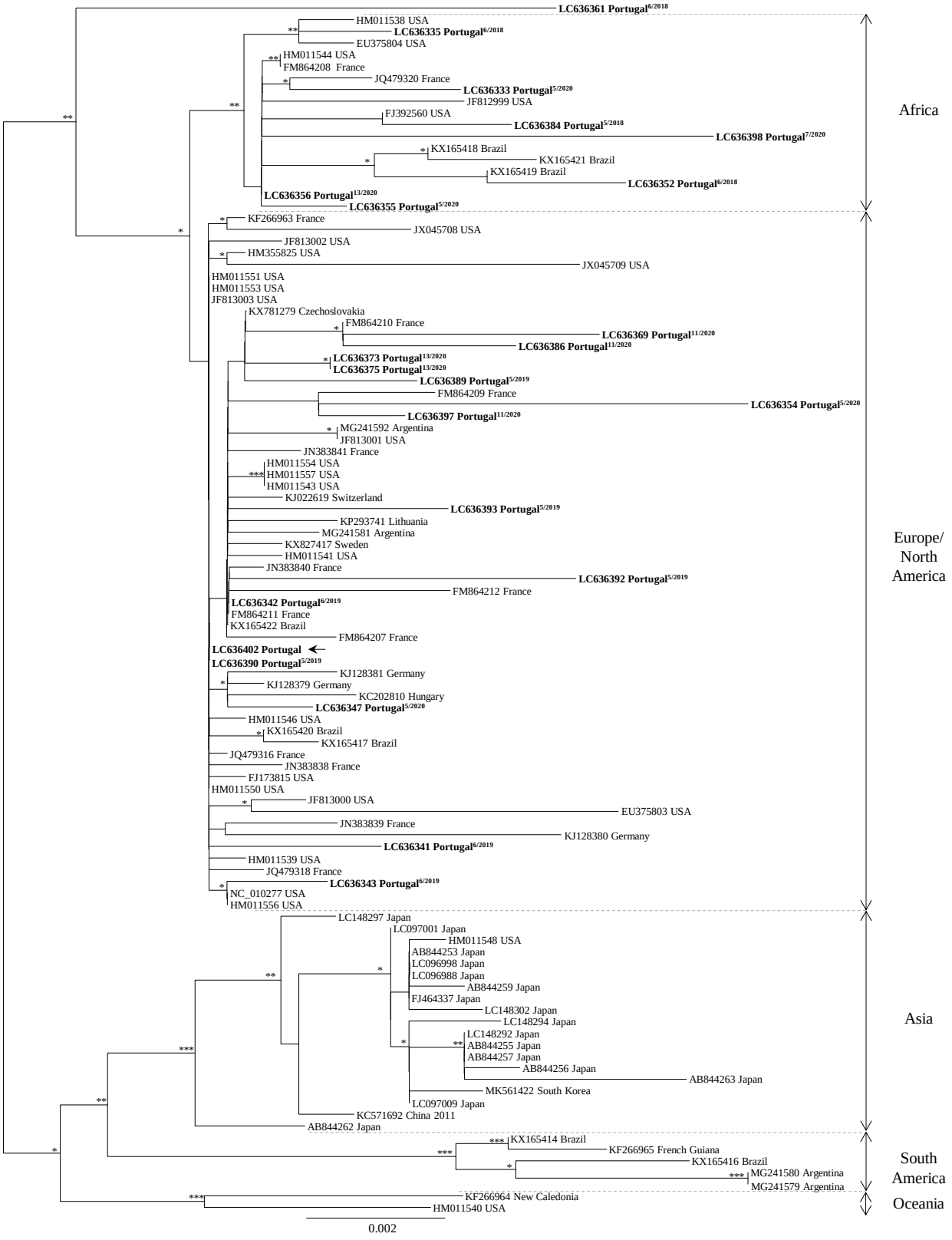
[b] Sequences typed using only BLASTn results. The collection site of the samples of these sequences is marked using the same annotation as the one used in the trees (refer to Figure 3.6).


A
AB369092 USA
AB263913 CAF
AB263926 Kenya
AB263938 South Africa
LR216148 Australia
LC636350 Portugal5/2020
DQ989804 USA
AB263928 Kenya
AB263914 CAF
KY487998 USA
AB365157 Saudi Arabia
DQ305492 USA
NC_001538 USA
AB263922 UK
AB263932 Niger
JN794021 France
AY628225 West Africa
AY628237 USA
LC636385 Portugal5/2018
JN794036 France
AB263927 Kenya
AB211369 Netherlands
JX195581 Norway
JN794005 France
LR216009 Vietnam
LC636378 Portugal7/2020
MG871538 Thailand
MW587973 USA
AB211373 Japan
LC636381 Portugal7/2020
MG871532 Thailand
AB211371 South Africa
X56913 nd
AB263934 Philippines
JN192433 Ireland
LR216190 Australia
AB263929 Myanmar
MG871533 Thailand
MN135851 UK
AB369094 USA
KY132094 USA
MF627831 France
LR216182 Australia
DQ989813 USA
JN192439 Ireland
JX195573 Spain
AY628232 USA
KT354838 Argentina
AY628233 USA
AY628227 USA
JN192437 Ireland
LT960370 UK
JX195583 Norway
LC636353 Portugal6/2018
JX195582 Norway
AB298947 Japan
JX195562 Spain
JX195564 Spain
LC636351 Portugal5/2020
LC636363 Portugal6/2019
LC636374 Portugal8/2020
LC636400 Portugal8/2020
LC636359 Portugal8/2020
LC636364 Portugal8/2020
LC636362 Portugal8/2020
JN794025 France
AB369090 USA
KT354828 Argentina
AB263936 Turkey
FR720310 Russia
AB263917 Ethiopia
AB263925 Italy
AB263935 Sweden
AB260030 Finland
DQ989796 USA
MW023597 USA
JN192438 Ireland
AB263923 UK
AB263915 Spain
AB211370 Netherlands
KT354824 Argentina Ib1
LC636403 Portugal
AB464961 South Korea
AB301098 USA
AB369099 China
AB365159 China
AB464956 South Korea
AB365145 China
AB365151 Japan
AB211385 Japan
AB269860 China
AB269859 Philippines
AB269842 Myanmar
AB269869 Vietnam
AB269868 Vietnam
AB211389 Japan
AB269862 China
AB269841 Myanmar
AB269840 Japan
AB211391 Japan
AB269851 Mongolia
AB365141 Japan
AB365150 Japan
AB211388 Japan
AB269853 China
AB269867 Vietnam
AB269843 Myanmar
AB269827 China
AB269828 China
AB269850 Mongolia
AB269849 Mongolia
FR720320 Russia
AB269822 Finland
AB269866 Sweden
AB269857 China
AB301097 USA
AB269831 Greece
AB269832 Italy
AB369093 USA
AB301101 USA
EF376992 Germany
AB263920 UK
AB263916 Ethiopia
JN192440 Ireland
KF055891 USA
M23122 UK
JX195577 Spain
LC636340 Portugal6/2019
MW587962 USA
AB211386 Japan
LR216258 Australia
LR216252 Australia
0.007
Ia
Ib1
Ib2
Ic
IVa1
IVa2
IVb1
IVb2
IVc1
IVc2
II
III

B


AB081017 Mexico
AF281619 nd
AB081006 Canada
MW588000 USA
AF015534 USA
AF281620 Argentina
LC636358 Portugal13/2020
LC636379 Portugal2/2020
AF295734 Puerto Rico
AF295733 Puerto Rico
AB081026 Peru
AB081025 Peru
AB081027 Peru
AB081612 USA
AB081022 Mexico
AB081011 Guatemala
LC636380 Portugal2/2020
AF281607 nd
AY121910 South Korea
AY121908 South Korea
AF030085 Japan
AY121907 South Korea
AB190452 Japan
AB081606 USA
AB103405 Japan
AB081601 USA
AB081019 Mexico
AB081614 USA
AB081015 Guatemala
AB262405 Turkey
AB262402 Russia
AB262397 Mongolia
AB048552 Mongolia
AB262413 Myanmar
AB262410 Nepal
AF015536 USA
AB262398 Sri Lanka
AF281604 India
AB262399 Sri Lanka
AB048555 Saudi Arabia
AB081613 USA
AF281605 New Britain
AF396422 Australia
AF281606 Guam
AF396426 Solomon Islands
AF396425 New Caledonia
AF015533 USA
AF015532 USA
AB048554 Netherlands
AB372036 Greece
LC636396 Portugal13/2020
JF424903 nd
LR215999 Vietnam
AB048581 Malaysia
U61771 Taiwan
AB077872 China
AF295738 Hong Kong
JX273163 USA
AB198954 Mauritius
AB077858 Thailand
AB077859 Myanmar
AB077862 Myanmar
AF396427 USA
AB113120 Philippines
AB048582 Mongolia
AB077856 Thailand
AB113139 Philippines
AB113135 Philippines
AB048550 China
AB113216 China
AB048557 India
AB048558 Mauritius
AB048559 Mauritius
JQ237146 China
LC422956 Japan
AB212952 Japan
AB081605 USA
AB118231 Japan
AB048561 Mongolia
AF300955 China
AB081618 Japan
AY121912 South Korea
AB113141 Thailand
AB113145 China
AB113128 Philippines
AF396433 Vanuatu
AF396429 Solomon Islands
AF396428 Micronesia
AF396434 Wallis and Futuna
AF281623 PNG
AF396430 PNG
AF281626 PNG
AB126999 Mauritania
AB127002 Morocco
LC636377 Portugal2/2020
DQ875212 USA
MN163031 Iran
AB127019 Sudan
AB126987 Greece
AB127000 Macedonia
AB127023 Uzbekistan
AB126997 Kenya
AF295732 Tanzania
AY328376 Ethiopia
AF295735 Puerto Rico
U73501 Tanzania
AB127014 South Africa
AY536240South Africa
AB127013 South Africa
AF015537 USA
AB038253 Ghana
AF295736 Puerto Rico
AB048545 Ghana
LC636376 Portugal7/2020
AB195639 Greenland
AB074578 Canada
AB183152 South Korea
AB074588 Netherlands
AB048575 Sweden
MF662198 Finland
AB038251 Germany
AF281600 Hungary
JF424923 nd
KJ659286 Norway
LC636370 Portugal13/2020
LC636401 Portugal13/2020
LC636395 Portugal13/2020
NC_001699 USA
AB074582 Italy
AF015526 USA
AB081615 USA
AB074580 Germany
AB074591 Spain
AB074583 Italy
MF662201 Finland
AF015527 USA
AF281599 Hungary
AB048576 UK
AB127352 Finland
KM225765 Switzerland
LC636404 Portugal
AF015528 USA
MF662186 Finland
AF281612 nd
AB048567 UK
LC636394 Portugal13/2020
AB048566 Spain
AB048565 Netherlands
AB048563 Greece
AB048564 Morocco
MW588005 USA 2019
AF281622 Germany
LC636357 Portugal13/2020
AB092581 Japan
AB092579 Japan
AB074589 Russia Siberia
AB074577 Russia Siberia
0.005
2A2
2A1
2A/C
2D1
2D2
2D3
2E
2B
7A
7C1
7C2
7B1
7B2
8B
8A
3A
3B
6
1A
1B
4
Eu-c

C


LC636361 Portugal6/2018
HM011538 USA
LC636335 Portugal6/2018
EU375804 USA
HM011544 USA
FM864208 France
JQ479320 France
LC636333 Portugal5/2020
JF812999 USA
FJ392560 USA
LC636384 Portugal5/2018
LC636398 Portugal7/2020
KX165418 Brazil
KX165421 Brazil
KX165419 Brazil
LC636352 Portugal6/2018
LC636356 Portugal13/2020
LC636355 Portugal5/2020
KF266963 France
JX045708 USA
JF813002 USA
HM355825 USA
JX045709 USA
HM011551 USA
HM011553 USA
JF813003 USA
KX781279 Czechoslovakia
FM864210 France
LC636369 Portugal11/2020
LC636386 Portugal11/2020
LC636373 Portugal13/2020
LC636375 Portugal13/2020
LC636389 Portugal5/2019
FM864209 France
LC636354 Portugal5/2020
LC636397 Portugal11/2020
MG241592 Argentina
JF813001 USA
JN383841 France
HM011554 USA
HM011557 USA
HM011543 USA
KJ022619 Switzerland
LC636393 Portugal5/2019
KP293741 Lithuania
MG241581 Argentina
KX827417 Sweden
HM011541 USA
JN383840 France
LC636392 Portugal5/2019
FM864212 France
LC636342 Portugal6/2019
FM864211 France
KX165422 Brazil
FM864207 France
LC636402 Portugal
LC636390 Portugal5/2019
KJ128381 Germany
KJ128379 Germany
KC202810 Hungary
LC636347 Portugal5/2020
HM011546 USA
KX165420 Brazil
KX165417 Brazil
JQ479316 France
JN383838 France
FJ173815 USA
HM011550 USA
JF813000 USA
EU375803 USA
JN383839 France
KJ128380 Germany
LC636341 Portugal6/2019
HM011539 USA
JQ479318 France
LC636343 Portugal6/2019
NC_010277 USA
HM011556 USA
LC148297 Japan
LC097001 Japan
HM011548 USA
AB844253 Japan
LC096998 Japan
LC096988 Japan
AB844259 Japan
FJ464337 Japan
LC148302 Japan
LC148294 Japan
LC148292 Japan
AB844255 Japan
AB844257 Japan
AB844256 Japan
AB844263 Japan
MK561422 South Korea
LC097009 Japan
KC571692 China 2011
AB844262 Japan
KX165414 Brazil
KF266965 French Guiana
KX165416 Brazil
MG241580 Argentina
MG241579 Argentina
KF266964 New Caledonia
HM011540 USA
0.002
Africa
Europe/ North America
Asia
South America
Oceania

D

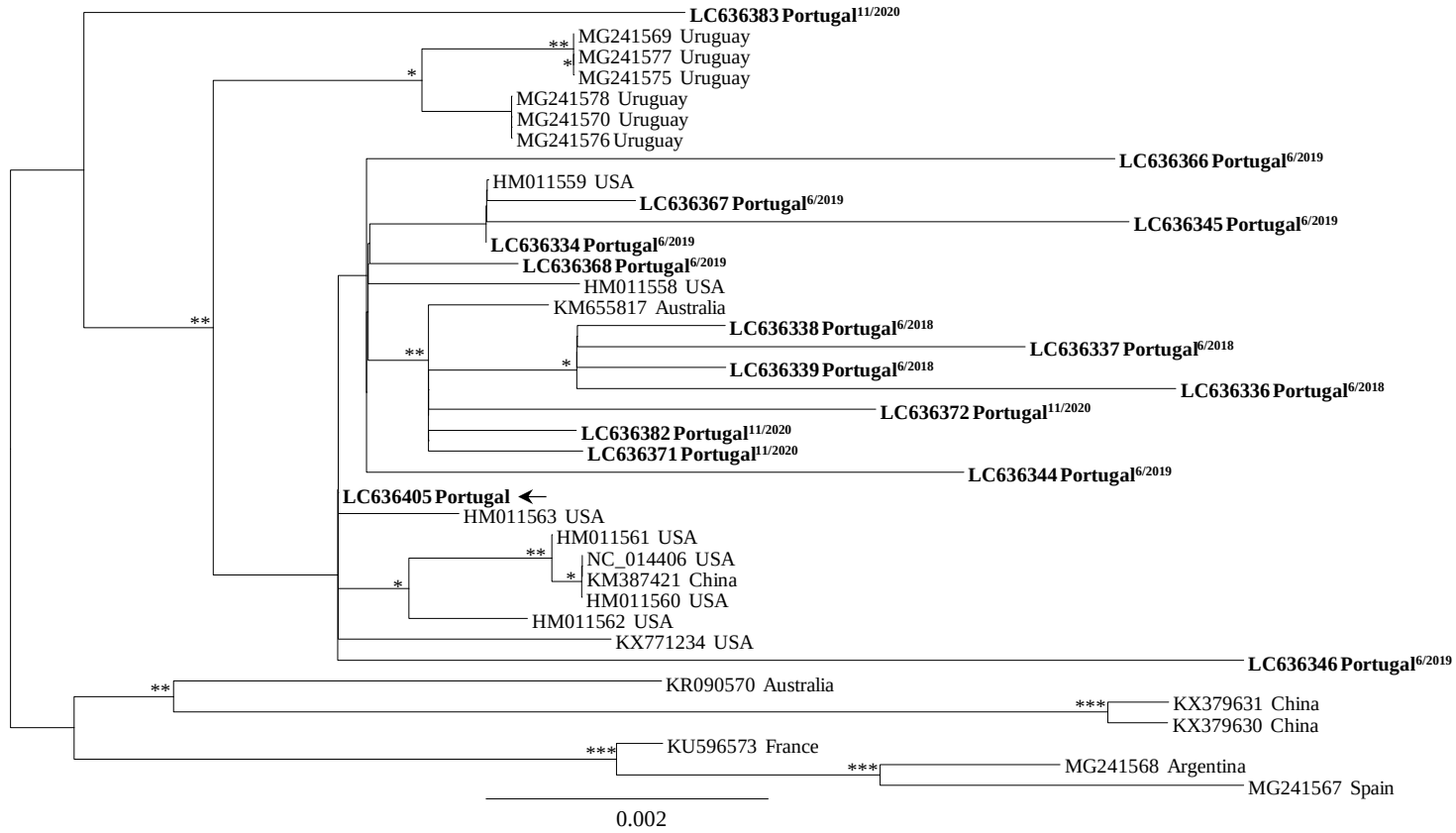


Figure 3.6. Phylogenetic analysis by maximum likelihood of species-specific HPyV sequences considering the structural protein-coding region. Phylogenetic reconstruction using a HPyV1 (A), HPyV2 (B), HPyV5 (C), and HPyV6 (D) dataset. For the first three analyses, the previously described viral genotypes/subtypes are indicated. All sequences are identified by their accession number and, whenever possible, country of origin. The sequences obtained during this work are indicated in bold, with the numbers in superscript referring to the sample collection site, as indicated in Figure 2.1 (in "Materials and Methods"). The arrows indicate the consensus sequences generated by NGS. The number of "*" indicates the number of phylogenetic reconstruction methods that confirm the observed topology, assuming 75% and above as relevant aLRT and bootstrap values, and 0.80 and above for posterior probability. The "nd" after some of the access numbers in the tress stands for "non-determined". The bars indicate the average number of substitutions per site.

To confirm the genetic lineages established by the phylogenetic tree analysis, additional tools were explored. These tools extended our genetic characterization of HPyV sequences, including the establishment of relationships between sequences that could not be detected while simply assuming the usual bifurcated pattern of evolution and the formation of clades, as seen in "classical/cladistic" phylogenetic reconstruction. One of them allowed for the construction of NeighborNet networks (NNn), while the other was PCOORD, which tries to identify patterns in an alignment of sequences and creates a graph where the position of each sequence is influenced by the genetic distance between them (in this particular case, the distance is measured by the total number of mismatches that exist between them, divided by the total number of compared nucleotides i.e., Hamming (or proportion) distances).

For HPyV1, the ML/Bayesian phylogenetic analysis revealed that the analyzed viral sequences segregated into four major lineages (corresponding to genotypes I to IV), with two of them further dividing into several subtypes (Figure 3.6A). From a total of 15 sequences analyzed (including the one sequence that was only identified by BLASTn), all but one were assigned to either genotype I (subtype Ia and Ib1) or genotype III (Table 3.3). Altogether, the majority of the HPyV1 sequences were identified as part of the Ib1 subtype (approximately 73%), while three of the remaining sequences fell into subtype Ia or genotype III (Table 3.3). These associations were almost entirely confirmed by both the NNn and PCOORD. However, the sequence identified by LC636403 failed to associate with any of the previously identified genetic subtypes in the phylogenetic tree. Whilst the PCOORD graph only suggested that this sequence might be associated with genotype I (Figure 3.7B1), in the NNn this sequence appears to most likely be an Ib1 sequence (Figure 3.7A1). As such, LC636403 has been designated as a genotype I-like sequence.

Regarding HPyV2, the phylogenetic reconstruction disclosed the presence of eight different genotypes that included twelve Portuguese sequences (Figure 3.6B). In this case, the heterogeneity of the sequences described in this work became apparent, considering the variety of viral types that were identified. Overall, the same number of sequences (n = 3) were attributed to subtype 1A, 1B, 2A2, and genotype 4, accounting for 75% of the total of HPyV2 sequences. The remaining HPyV2 sequences conformed to either subtype 2B (n = 1) or subtype 3A (n = 2) (Table 3.3). As in the case of HPyV1, one sequence (LC636376) failed to be assigned to any of the previously identified genotypes, appearing as an isolated branch between the division that separates genotypes 1, 4, and Eu-c from the remainder (Figure 3.6B). This singularity was further confirmed by the NNn, where it was possible to observe an isolated sequence (LC636376) that yet again appeared amidst the same genotype division as in the phylogenetic tree (Figure 3.7A2). Moreover, the PCOORD graph also revealed that this particular sequence is equally isolated from the rest (Figure 3.7B2). Thus, we suggest that this might be the first representative of a still undescribed genotype 9. The PCOORD graph also showed another sequence

(LC636358) that appeared isolated from all the others (Figure 3.7B2). However, this same sequence was assigned to subtype 2A/C by the phylogenetic analysis and the constructed NNn, suggesting this result may only be a consequence of the used bioinformatic tool (Figure 3.6B and Figure 3.7A2).

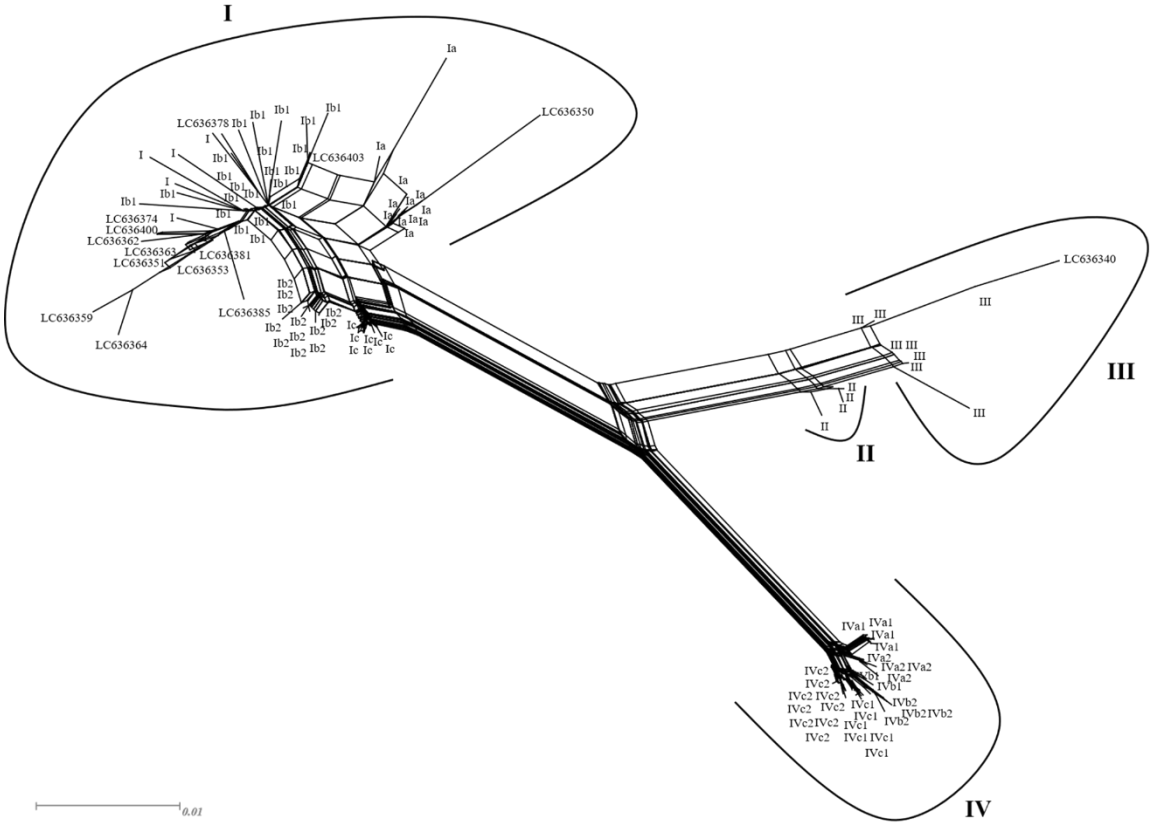
As for HPyV5, the majority of the sequences (approximately 65%) were assigned to the Europe/North American cluster, with only eight of the 26 sequences being ascribed to the African cluster (Table 3.3). Once again, one of the sequences (LC636361) failed to conform to any of the clusters proposed by Martel-Jantin et. al (79), emerging as a single, long branch, away from de African and Europe/North America clusters (Figure 3.6C). This genetic singularity was confirmed by the NNn, but not entirely by the PCOORD analysis, where the sequence appeared to be a divergent member of the Europe/North American cluster (Figure 3.7A3 and Figure 3.7B3), being thus classified as either (i) a more ancient member of a possible "supercluster", that encompasses both European/North American and African sequences, or (ii) a representative of a new cluster altogether.

As previously mentioned, from all the phylogenetic reconstructions that were performed in this study, the ones including only HPyV6 sequences presented a lower topological stability. This may be a direct consequence of the limited number of HPyV6 sequences available in the public databases. This lead to the construction of a dataset containing a smaller number of reference sequences and, subsequently, to a tree where it was impossible to unambiguously identify the expected clusters. Even so, the phylogenetic reconstruction showed that the 16 sequences obtained in this work clustered with references with distinct geographical origins, suggesting substantial geographic heterogeneity between them (Figure 3.6D). However, by analyzing the corresponding NNn and PCOORD graphs, this distinctiveness did not appear so obvious, as most Portuguese sequences seemed to be more closely related as they appeared to cluster together (Figure 3.7A4 and Figure 3.7B4). The only exception to this grouping was sequence LC636383, which clearly segregated from the remaining sequences in both NNn and PCOORD analyses while having been the only sequence to be completely isolated from the others in the tree. Furthermore, in the PCOORD analysis, one sequence (LC636346) was very close, but not part of, the major cluster containing the great majority of the obtained sequences (Figure 3.7B4). Although this was not replicated in the NNn (Figure 3.7A4), in the phylogenetic analysis LC636346 was, indeed, in a different monophyletic grouping than most of the other sequences (Figure 3.6D).

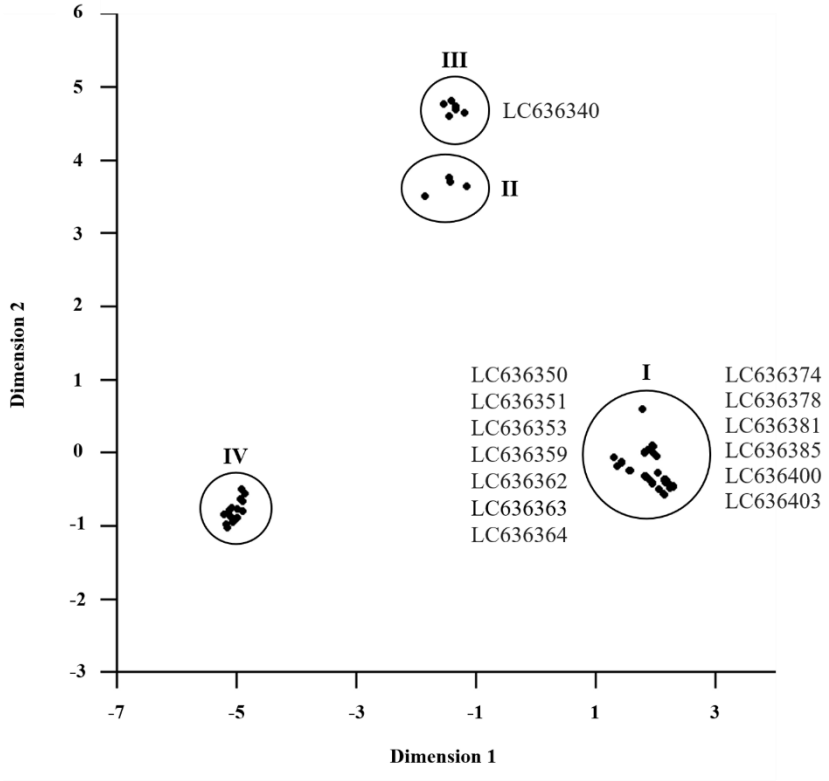
In the obtained phylogenetic trees (Figure 3.6), the hypothesis that the Portuguese sequences might be segregated by sampling point (geographic clustering) and/or year was taken into consideration. Our analysis revealed that whenever a monophyletic cluster containing Portuguese sequences was identified, in most cases its members were indeed obtained from samples collected either in the same or a nearby sampling location, in the exact same year. However, when multiple sequences were obtained from the

same sampling location, in each species-specific phylogenetic tree, they commonly associated with different reference sequences instead of creating a monophyletic group or clustering with the same references.

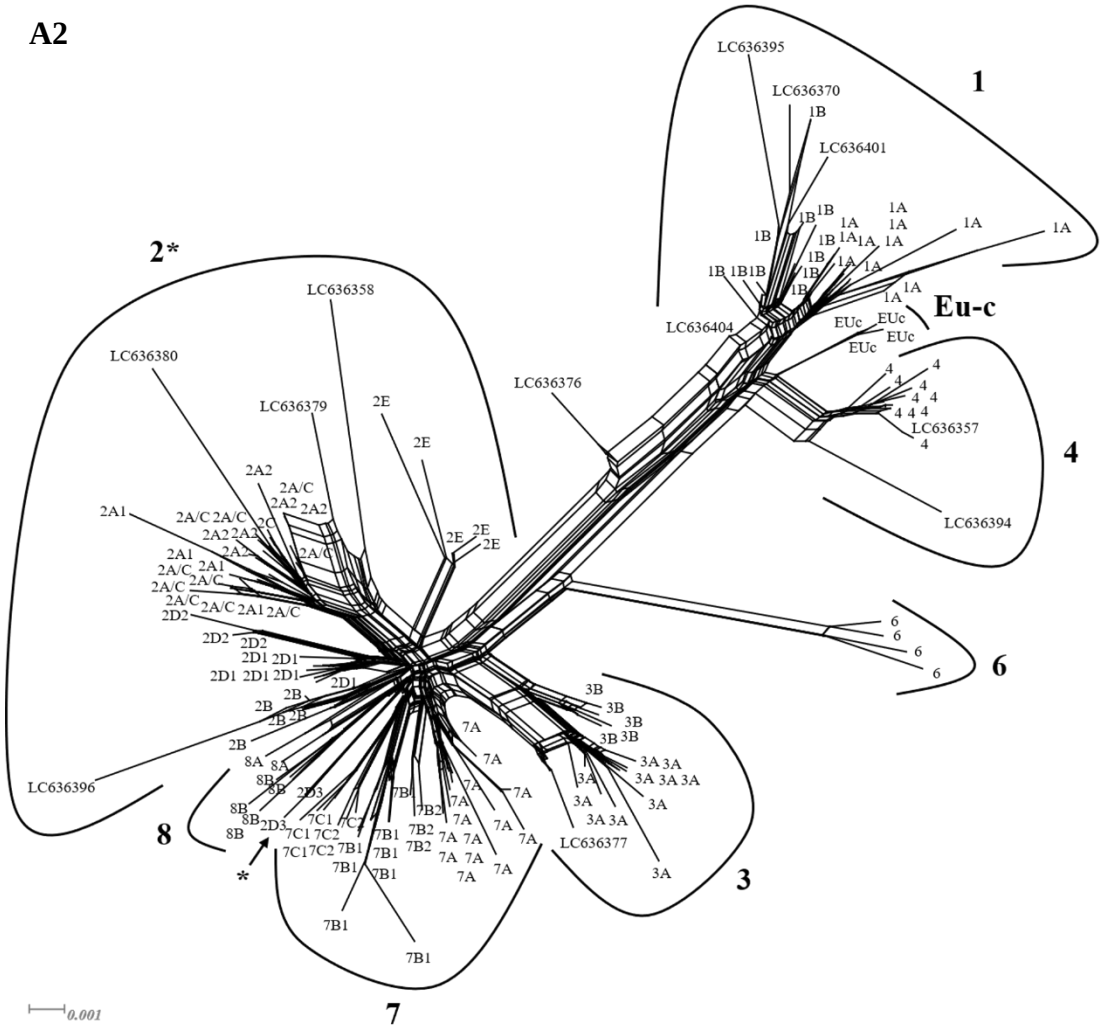
A1


I
II
III
IV
LC636340
LC636350
LC636359
LC636364
LC636385
0.01

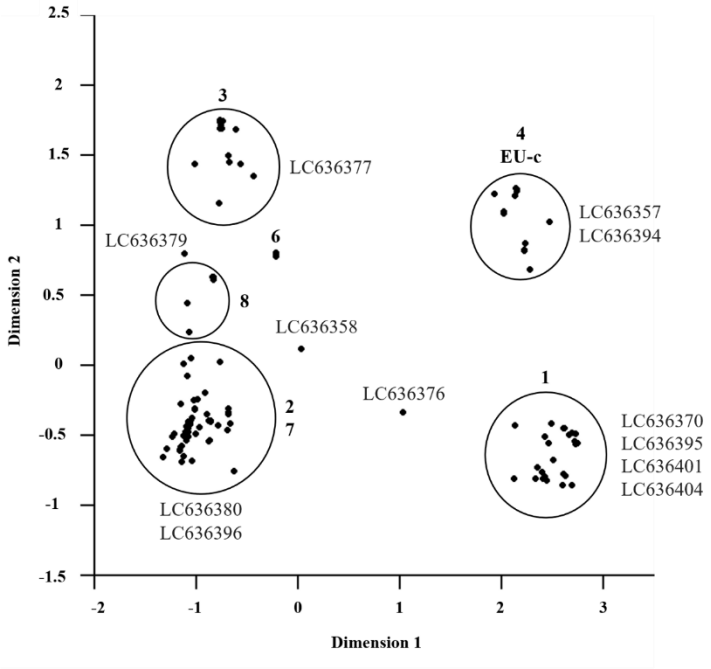
B1


III
LC636340
II
I
LC636350
LC636351
LC636353
LC636359
LC636362
LC636363
LC636364
LC636374
LC636378
LC636381
LC636385
LC636400
LC636403
IV
Dimension 2
Dimension 1
6
5
4
3
2
1
0
-1
-2
-3
-7
-5
-3
-1
1
3

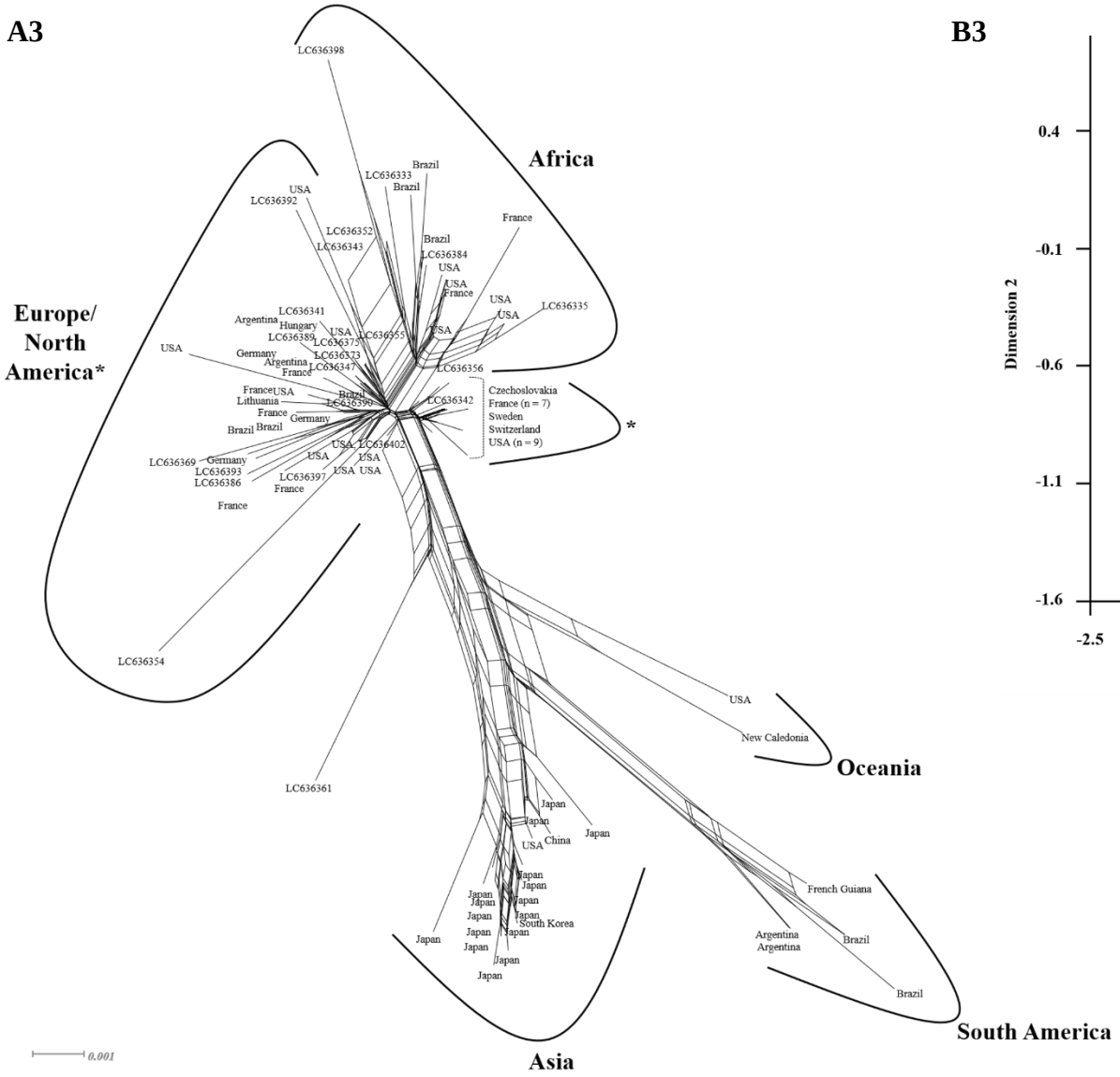
A2


LC636395
LC636370
LC636401
LC636404
LC636376
LC636358
LC636380
LC636379
LC636357
LC636394
LC636396
LC636377
1
2*
Eu-c
4
6
8
3
7
*
0.001

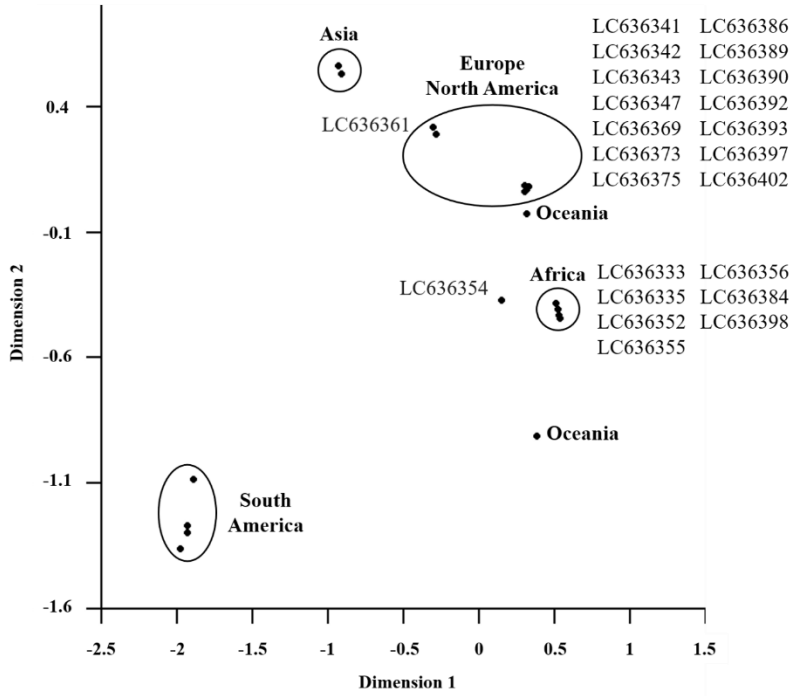
B2


Dimension 2
Dimension 1
3
LC636377
4
EU-c
LC636357
LC636394
LC636379
6
8
LC636358
LC636376
1
2
7
LC636370
LC636395
LC636401
LC636404
LC636380
LC636396

A3


Africa
Europe/
North
America*
Oceania
South America
Asia
*
Czechoslovakia
France (n = 7)
Sweden
Switzerland
USA (n = 9)
LC636398
LC636354
LC636361
0.001

B3


Asia
Europe
North America
Oceania
Africa
South
America
LC636361
LC636354
LC636341 LC636386
LC636342 LC636389
LC636343 LC636390
LC636347 LC636392
LC636369 LC636393
LC636373 LC636397
LC636375 LC636402
LC636333 LC636356
LC636335 LC636384
LC636352 LC636398
LC636355
Dimension 1
Dimension 2
0.4
-0.1
-0.6
-1.1
-1.6
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5

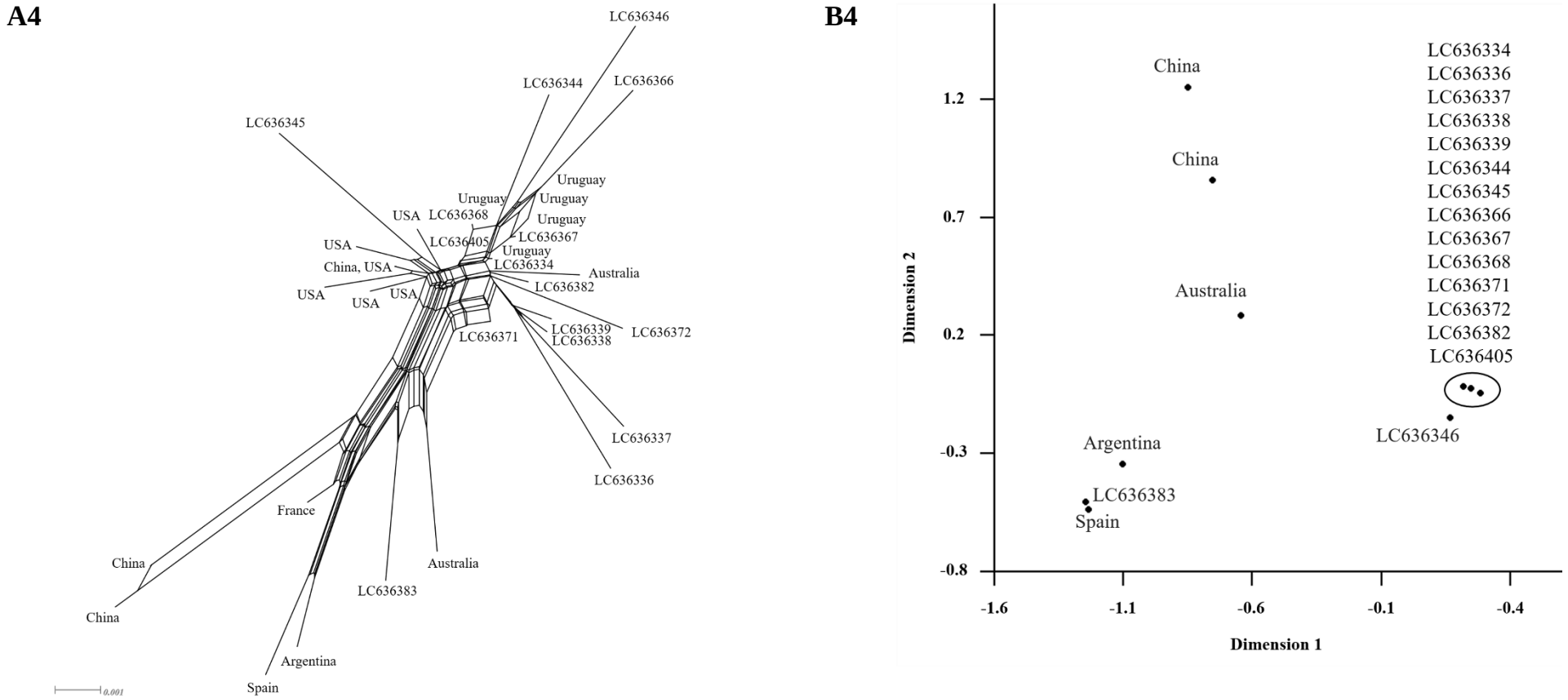


Figure 3.7. NeighborNet networks (A) and PCOORD analysis (B) of HPyV sequences. These genetic analyses were done using HPyV1 (A1/B1), HPyV2 (A2/B2), HPyV5 (A3/B3), and HPyV6 (A4/B4) datasets. The sequences obtained during this work are indicated by their accession number. For the NNn, the reference sequences are either indicated by their respective genotype/subtype (A1 to A3) or by their country of origin (A4) In the PCOORD graphs, the two first dimensions account for 50.30%, 47.02%, 73.95%, and 40.20% of the total nucleotide differences from B1 to B4, respectively. Additionally, the first ten axis cover 90.60% (B1), 78.78% (B2), 90.88% (B3), and 67.75% (B4) of the sequence variation in each dataset. .

4 | Discussion

The search for an indicator that would allow more precise monitoring for the presence of enteric viruses in water bodies should be regarded as a pressing matter. As it is known, most microbial detection tools used for the assessment of human and/or animal waste contamination involve either the (i) direct detection of a certain species of bacteria (assessing bacterial growth) or (ii) call for the partial amplification of a specific marker in its DNA, thus neglecting the potential presence of other types of cellular organisms (e.g. protozoa) or even viruses, that may pose a serious public health concern (6, 123). Furthermore, the use of a microbiological marker that would allow unambiguous discrimination between human and animal excreta would be a powerful tool to promptly identify the source of contamination and, subsequently, correct it.

Human polyomaviruses (HPyVs) are one of the biological indicators recently proposed as a water quality marker that shows potential whenever one addresses both shortcomings mentioned above. Besides their host-specificity, these viruses appear to be present in the environment all year-round, are highly prevalent in the human population and are present in several types of water-associated samples, including treated/untreated sewage, and environmental bodies of waters (29). If implemented, the detection of HPyVs could be a powerful Microbial Source Tracking tool, allowing to discriminate between types of fecal contamination (human or animal), and thus rapidly assess the human health risk that this type of pollution may pose (6). Hence, various PCR protocols have been developed, and several studies have been carried out, surrounding the detection of HPyVs in multiple types of water matrices, with some reports simultaneously studying both their role as possible indicators of human fecal contamination and the distribution patterns displayed by these viruses in a particular geographical region (29, 30, 124–129).

Considering (i) the importance of molecular epidemiological studies in disclosing the diversity of HPyVs circulating in the human population, and (ii) the consolidation of HPyVs as a microbiological fecal pollution indicator and since, to our knowledge, (iii) no study had ever previously assessed the distribution of this group of viruses in any particular region of Portugal, this work focused on the molecular analysis of HPyV genomic DNA found in influent (i.e., untreated sewage) and environmental water samples collected in the so-called Lisbon Metropolitan Area (AML). Thus, to gain more insights into the molecular epidemiology of these viruses and, in addition, aid in reinforcing the proposal of HPyVs as human fecal markers, this study focused on the detection, and further genetic characterization, of HPyV sequences identified in water samples collected in several municipalities of the AML. This region includes some of the most populated cities of Portugal, one of them being Lisbon (i.e., the capital

city of Portugal), which was well known as a preferred tourist destination before the Coronavirus disease 2019 (COVID-19) pandemic. Besides the thousands of tourists that visited Lisbon annually, the AML population is by itself very diverse, integrating various ethnic backgrounds and cultures. Moreover, some of the samples analyzed in this work were collected before the beginning of the COVID-19 pandemic. Hence, the fact that various genotypes/subtypes belonging to four different species of HPyV (HPyV1, 2, 5, and 6) were identified in this study, as will later be discussed, came as no surprise.

To accomplish the objectives set for this work, two independent touch-down multiplex nested PCR protocols that targeted, for amplification, part of the late region of the viral genome (i.e., the structural protein-coding region) of most HPyV species, were designed. Although the undertaken experimental approach involved the analysis of alignments of sequences that encompassed part of VP1's and part of VP2's coding region (henceforth referred to as VP1-2 sequences), the taxonomic differentiation of HPyVs is traditionally done using amino acid sequences of the LTAg, according to ICTV (31). Since LTAg sequences are considerably less polymorphic than the ones regarding the late region of the genome, it is easier to obtain a good quality alignment containing enough information to conduct a reliable phylogenetic analysis (130). Although the phylogenetic relationships revealed by the analysis of the structural protein-coding region do not completely correspond to the ones disclosed by the analysis of the LTAg, the analysis of the late region of the viral genome still allows for the identification of not only the different viral species but also the various genotypes and subtypes that the latter may include. Admittedly, several studies have reported that the phylogenetic resolution capability of the late region of the viral genome is suitable to accurately attribute a viral genotype and/or subtype to a HPyV sequence, with the analysis of this section of the HPyV genome having been already used to define multiple genotypes and subtypes, particularly for HPyV1, HPyV2, and HPyV5 (49, 80, 131–133). Indeed, a likelihood mapping analysis carried out in this study demonstrated that the phylogenetic signal of a dataset composed of several reference sequences from all HPyV species, as well as by the sequences obtained in the course of this work, was high, resolving 95.60% of the sequence quartets randomly sampled from an aligned sequence dataset. Therefore, this analysis is consistent with the observation regarding the use of the late region of the viral genome as a target in phylogenetic studies.

To implement both PCR amplification protocols, it was first necessary to design pairs of primers that targeted the structural protein-coding region of HPyVs. The definition of the two independent PCR protocols grew from a preliminary phylogenetic analysis of the viral genome, using a dataset of various representative sequences of all HPyV species. Hence, the group that included HPyV1, HPyV2, HPyV5, HPyV8, HPyV9, HPyV10, and HPyV11 was targeted by PCR-A, while the group that contained HPyV3, HPyV4, HPyV6, and HPyV7 was identified by PCR-B. While a similar strategy had been already established by Torres at. al in Argentina (125), for the screening of wastewater, river, and urine samples

for the presence of HPyVs, the latter was designed in such a way that it became possible to, at least in some instances, perform a direct preliminary identification of the obtained amplification products by their size estimation using a simple electrophoretic analysis. In contrast, the approach used here was planned in a way that the obtained 2nd-round amplicons would have relatively similar sizes, most of which were considerably larger than those obtained by Torres and collaborators. Indeed, the smallest possible amplifiable DNA fragment would correspond to an 881 bp amplicon (corresponding to HPyV3), and the biggest would comprehend 1175 bp (referring to HPyV10). In this regard, one should not forget that the shear sequence length of the intended amplicons may impact, for example, the performance of phylogenetic reconstruction. By creating multiple sequence alignments with longer sequences, presumably more information regarding them (i.e., a bigger number of nucleotides) will be considered when constructing the phylogenetic tree, usually resulting in an improved genetic resolution capability, as long as the dataset of sequences used displays enough phylogenetic signal (134). Furthermore, when trying to detect HPyV DNA in a complex matrix such as untreated sewage water, spurious amplification is a frequent occurrence. This might corroborate the hypothesis that the preliminary visual identification of the viral origin of the fragment might not be that helpful, especially when considering that in the work of Torres at. al., the viral identity of the amplicons still had to be confirmed by DNA sequencing in most cases.

To indisputably identify, and genetically characterize, different HPyV genomes in the collected water samples, this study also involved sequencing of VP1-2 amplicons as DNA inserts ligated to a vector. To do so, several independent recombinant plasmids were selected and analyzed, producing sequence data that would later be analyzed by (i) phylogenetic inference, (ii) network reconstruction, and (iii) PCOORD analysis. However, the hypothesis that some of the choices made in our viral DNA detection algorithm may have contributed to a biased result could not be ignored. Particularly, the use of a molecular cloning step may have inaccurately portrayed the circulation of the different HPyVs in the environment, mostly due to the fact that only a small number of recombinant molecules was analyzed. Moreover, some of the amplified viral sequences might have been lost during replication of the bacterial strain even under the selective pressure of the antibiotic-supplemented media, subsequently leading to a loss of the bacterial culture. Thus, to try and confirm our results and ensure no bias appeared to be introduced by the molecular cloning step, a NGS analysis was also used.

The detection of HPyVs was a clear success when screening wastewater, with HPyV-specific amplicons detected in all samples (100%) of this type of water matrix. On the other hand, for environmental samples, the detection efficiency was considerably lower (12.5%). This difference in amplification efficiency when considering both types of water samples is not completely unexpected and most likely translates a substantial difference in viral titer. Considering all the environmental

samples were collected during autumn, and some already close to winter, the increased occurrence of rainfall may have water-downed the existing viral particles, especially in the environmental samples, resulting in a lower HPyV titer. Additionally, some of these environmental samples were collected from river estuaries (Tejo and Sado basins), where the possible diluting effect of a large body of water under the influence of the tides cannot be ignored.

Over time, several studies have reported the detection of HPyVs in different types of environments, such as rivers, coastal water, and stormwater, as well as wastewater samples collected in various geographical locations, with HPyV1 and HPyV2 being the HPyVs most commonly used as human fecal pollution indicators (4, 29, 125, 127, 129, 135–137). Since the presence of these types of viruses in untreated sewage is frequently disclosed (125, 126, 129, 138, 139), the fact that in this work HPyV detection was performed with high efficiency on the wastewater samples was not unexpected. Furthermore, even though numerous studies have attested to the presence of HPyV-specific DNA in various water matrices, the efficiency of detection also appears to fluctuate considerably when comparing samples collected in different geographical regions. In particular, the detection of viral DNA in samples collected in countries from central/western Europe (such as Germany, Italy, and Greece) appears to display lower efficiency values (between 50% and 75%) (127, 140, 141) than the ones disclosed by studies performed in other geographical regions, such as Japan, Chile and several states of the United States of America (29, 129). These discrepancies might not only be due to geographically related variables that may impact viral concentration in the waters (such as the influence of distinct ocean currents, or differences in the sewage systems, for example), but also be conditioned by the experimental approach undertaken. Admittedly, the disparity between these studies may simply translate a difference in PCR protocols and/or the viral targets used.

While sequencing the recombinant plasmids, we detected bacterial and crustacean DNA in association with wastewater treatment plants C and D, respectively, whereas simultaneously not finding HPyV DNA in any of the sequenced individual plasmids obtained from both samples. At first, this appeared to be quite an odd occurrence, considering that the probability of finding HPyV DNA in this type of water sample is, in fact, very high. However, we must also consider the fact that these detections of nonviral DNA might be due to the experimental approach followed, where the simultaneous usage of multiple primers in a single reaction, combined with the high "DNA-complexity" of the samples used may have led to a loss in detection sensitivity and specificity. Indeed, even though both samples appeared to be positive for the detection of HPyV DNA by electrophoretic analysis (i.e., the amplicon obtained was of the expected size), no viral DNA was found. This may purely reflect the limitations that come with including a molecular cloning approach in the workflow when only a finite, small number of molecules is analyzed. One way to obviate this issue would be to analyze the amplicon by direct

sequencing once it had been purified from the agarose gel. Unfortunately, due to the experimental design that we decided to follow, involving primers that are not always HPyV species-specific, this option came as an impossibility, since multiple sequences belonging to different HPyVs may be present in the same amplification product, and their genomes detected all at once.

Altogether, four distinct species of HPyV were detected. From those, the identification of HPyV1, HPyV2, and HPyV5 sequences was the least unexpected, considering these are the most commonly found HPyVs in wastewater (125). Particularly for HPyV1 and HPyV2, this might be a direct consequence of the fact that most studies regarding HPyV detection and analysis only use the genome of both these viruses as targets (i.e. only aim at the detection of HPyV1 and HPyV2 as a proxy for the presence of HPyVs). Furthermore, it is speculated that HPyV1 and HPyV2 may explore the fecal/oral route of transmission, having been detected in untreated sewage water in numerous studies (29, 30, 128, 129, 142). Although HPyV5 is less frequently used as an indicator of human fecal contamination, it has been identified in samples of human excreta and its presence in water samples has been thoroughly documented (78, 125, 143, 144). On the other hand, the identification of several sequences of HPyV6 was not anticipated, considering that even though this virus has been detected in fecal and urine samples (145–147), less evidence exists surrounding its possible transmission route.

When observing the distribution of each viral type, HPyV5 accounted for 35.62% of the obtained sequences, being followed by HPyV2 and HPyV6 (each corresponding to 21.92/1%), which were, in turn, closely followed by HPyV1 (20.55%). Besides being the most frequently identified HPyV in this work, HPyV5 was also the virus that was detected in the biggest number of samples (n = 8). However, the question of whether these values are a true depiction of the HPyV distribution in the analyzed samples remains to be clarified. Firstly, by introducing an amplification step in the experimental workflow, our analysis may be impacted by the use of primers with different amplification efficiencies. Curiously, it may not be a coincidence that the only virus that was detected using type-specific primers (i.e., exclusively targeting HPyV5) in a multiplex format was also the one most often detected. Moreover, the PCR strategy used might also have an effect on the qualitative aspect of the analysis, especially considering that we were unable to detect other HPyVs that have been previously found in wastewater and/or in urine and fecal samples or that are shed by the skin, such as HPyV3, HPyV4, HPyV7, HPyV10 and HPyV11 (83, 94, 144, 146, 148). Considering that the number of mismatches between the primers designed to target these viruses and their viral genomes was either zero or one (refer to Table 2.2 of "Materials and Methods"), this should not be a consequence of poor primer design. Even so, in a few select cases, we were able to find HPyV8/9 or HPyV10/11 primers at the extremities of HPyV1 and HPyV2 amplicons, which are primers that possessed less than 10 mismatches with the HPyV1/2 genome. Although we did use a touch-down PCR protocol to ensure high

amplification specificity, we must not forget the impact that the presence of multiple pairs of primers in the same amplification reaction may have, particularly when subjected to lower annealing temperatures. Another aspect worth mentioning is the fact that, besides HPyV2, HPyV5, and HPyV6, HPyV3 and HPyV4 have also been detected in one and two sewage samples, respectively, collected in Spain (29, 144, 146, 149), Portugal's "next-door neighbor" and the "second half" of the Iberian Peninsula. This may suggest that these viruses can also be found in Portuguese waters since the lack of border restrictions creates the tendency to freely travel between Portugal and Spain. As discussed above, the inclusion of a molecular cloning protocol might also have skewed the obtained results due to the fact that only a small, limited number of recombinant molecules of each sample is analyzed. However, the analysis of a pooled sample of amplicons obtained from multiple water samples using NGS revealed the exact same four viral species that we had already identified with the analysis of individual plasmid clones via Sanger sequencing. This appears to indicate that the DNA cloning step included in the experimental approach did not affect the ability to detect HPyV sequences. Even so, considering the NGS step was performed using a pooled sample of previously amplified and purified PCR products, this analysis does not exclude the possibility that the PCR amplification step might have biased (at least partially) our results. Hence, the fact that we were unable to detect HPyV DNA in most of our environmental samples may simply translate the absence of viral particles or might purely reflect that the sensitivity of our amplification assay is not sufficient to detect HPyV sequences when present in a low titer.

Even considering possible biases that might have impacted some of the obtained results, we were able to identify 15 HPyV1 sequences that, either by phylogenetic reconstruction or by a sequence similarity search using BLASTn, were attributed to genotypes I (subtypes Ia and Ib1) and III. As previously mentioned, this virus has been classified into four distinct genotypes (I-IV), with genotypes I and IV being widely distributed throughout almost the entire world (49, 52). Therefore, the fact that the majority of the HPyV1 sequences found belonged to genotype I (86.66%) was not unexpected. Moreover, most of the viral sequences were attributed to Ib1, a subtype that has been associated with South-East Asians (53). This is congruent not only with the history of the population of the AML, considering the former colonization of East Timor by Portugal, but as well as with its current state, considering the circulation of individuals from China, and some of its neighboring countries (such as Nepal and Bangladesh), is quite common in the region. Additionally, the detection of subtype Ia is also compatible with the diverse nature of the population that currently inhabits the AML, considering this viral subtype has been frequently associated with Africans (53). Unlike the aforementioned genotypes, II and III are more scarcely detected, displaying somewhat of a similar distribution, having been detected in mostly the same geographical areas (58). Even so, genotype II appears to be more common in Northern and East Africa, while genotype III is frequently associated with Western Africa (58). As such,

considering that the AML possesses a considerably high number of African or African-descendent individuals, the fact that one of the detected sequences in this work was assigned to genotype III, came as no surprise. Lastly, the phylogenetic reconstruction revealed a singled-out sequence (LC636403) that did not appear to correspond to any of the previously established genotypes. However, neither the NNn nor the PCOORD analysis appeared to suggest the same, with this genotype I-like sequence seemingly being attributed to the Ib1 subtype in the first, and clustering together with the remainder genotype I sequences in the latter.

Unlike HPyV1, where most of the sequences found were identified as being a part of subtype Ib1, the data collected regarding HPyV2 revealed a much more heterogeneous distribution. Indeed, six different genotypes/subtypes/subgroups were detected within the total of the sixteen HPyV2 sequences. Indeed, HPyV2 appears to be naturally more genetically diverse than HPyV1, and it has been classified by Stoner et. al (57) into eight different genotypes. Some of the latter have been further divided into subtypes and, in some instances, even subgroups. Like HPyV1, HPyV2 genotypes are commonly associated with specific human populations and geographical regions (55–57). For example, genotypes 1 and 4 are commonly associated with Europeans and North Americans, while genotype Eu-c is usually only found in Northeast Siberia and Japan (56). In this work, six HPyV2 sequences were associated with genotype 1 (three with subtype 1A, and three with subtype 1B), while three others were attributed to genotype 4. Considering the abovementioned distribution pattern displayed by both these genotypes, the fact that more than half of the identified HPyV2 sequences belongs to either genotype 1 or 4 was, again, not surprising. Furthermore, two sequences were identified as part of subtype 3A, which has previously been associated with the majority of the African territory, as well as with South, West, and Central Asia, and only rarely found in Southern Europe (63, 150, 151). Four genotype 2 sequences were also found, with three belonging to subgroup 2A2, and one to subtype 2B, with the latter being linked to Europeans and Eastern Asians (60, 61). All these findings appear to be congruent with either the current population of the AML or its origin, with perhaps the most curious finding being the presence of three sequences of 2A2, a subgroup of the 2A/C subtype that is usually associated with Eastern Asians, and North and South Americans (61). Even so, this data seems to purely reflect the multicultural background of the AML. Moreover, one sequence (LC636376) was not ascribed to any previously determined HPyV2 genotypes by any of the three of the genetic analyses performed. Therefore, it has been tentatively classified as a maiden representative of a putative HPyV2 genotype 9.

As aforementioned, the most frequently represented HPyV type among the obtained HPyV sequences was HPyV5, (n = 26). This virus has been classified into five geographically related genotypes: Europe/North America, Asia, Oceania, and South America (79). Almost all the sequences obtained were either part of the Europe/North America cluster (65.38%) or belonged to the African

genotype (30.77%), just as expected when considering AML's demographics, as well as the close relationships that Portugal has established with its past African colonies (Angola, Mozambique, Cape Verde, São Tomé and Príncipe and Guinea-Bissau). Additionally, one sequence (LC636361) did not cluster (by phylogenetic analysis) inside any of the previously established genotype-defining lineages but instead segregated away from the Europe/North America and African genotypes. Although this finding was not entirely confirmed by the other two genetic analyses (considering that in the PCOORD graph, LC636361 appeared as a distant member of the Europe/North America genotype), this sequence might be an ancient member of a possible larger cluster that encompasses both the viral genotypes mentioned above. Nonetheless, this result may purely be due to a low phylogenetic signal of the used dataset, and, as such, further studies involving a larger number of representative sequences would be required. In fact, to better comprehend the genetic diversity of this virus, additional studies based on sampling HPyVs from several geographical regions will be necessary, since we may be downplaying the real diversity of HPyV5.

Finally, and somewhat unexpectedly, we were able to identify a relatively high number (n = 16) of HPyV6 sequences. While this virus appears to be less frequent in water samples than HPyV1, HPyV2, or HPyV5, our results raise the hypothesis that, besides being shed from the skin (75), this virus may also be excreted in the feces and/or urine, with a couple of studies already reporting on its presence in fecal, and urine samples (83, 147). As mentioned in section 3.3.3 of "Results", we were unable to replicate the distribution of the sequences into the previously reported "Worldwide" and Asian clades (58). This may be due to the lack of information available in the databases, notably regarding the structural-protein coding region of HPyV6, which when combined with the low topological stability of the phylogenetic tree regarded us unable to assign the HPyV6 Portuguese sequences to individual genetic groups. In this case, sequence LC636383 was in a single segregating branch in the phylogenetic tree and associated with different references in the other two genetic analyses. Yet again, this may just be an artifact of the lack of representation of HPyV6 sequences in public databases and, consequently, in our dataset. Considering this fact, although HPyV6 appears to be frequent in the AML, more studies will be needed to further confirm this result, as well as to better understand the genetic characterization of this virus.

4.1 Final Remarks

In the course of this work, we were able to detect DNA sequences from HPyV1, HPyV2, HPyV5, and HPyV6 in a total of ten water samples collected in the AML. This appears to be the first-ever study to explore the genetic diversity of the HPyVs circulating in Portuguese territory (particularly

in the AML), describing newly designed primers featured in two independent multiplex touch-down PCR protocols. Furthermore, alongside other studies, this work uncovers, yet again, the presence of some type of HPyV in wastewaters, and reinforces the idea that these viruses might be a useful Microbial Source Tracking tool, serving as indicators of human fecal pollution in waters.

By observing the AML's current demographic, the genetic heterogeneity of the viral sequences found suddenly becomes an expected occurrence. Indeed, this region possesses a diverse population, hosting people from various ethnic backgrounds, beyond the already multiculturally diverse native population. This group of non-natives includes individuals from Eastern Europe, India, Brazil, China, and Africa, with most of the individuals with African ancestry originating from the former Portuguese colonies, such as Angola, Cape Verde, Guinea-Bissau, and Mozambique. Hence, even though most of the water samples were collected during the first year of the COVID-19 pandemic, the obtained results still accurately convey the diversity that the population of the AML carries at its core. In fact, approximately 7% of the entire population of the Lisbon district by itself was deemed as foreigner in the 2011 census, with at least 3.9% of the population in the Setúbal district representing non-Portuguese individuals (103). Studies like the one here described, where the sampled areas are characterized by a relatively heterogeneous population, are essential to further investigate the genetic diversity of HPyVs. This becomes particularly relevant when assessing sewage samples from wastewater treatment plants that receive the discharge sewage from the mentioned diverse population, and environmental samples from locations in close proximity with discharge points. Additionally, in order to disclose HPyVs possible role as human fecal indicators in a certain geographical region, it is first necessary to study the molecular epidemiology of these viruses in that same place. However, the latter is dependent on the efficiency of the experimental approach used, as well as on the limitations that come with working with complex matrices such as wastewater influent. As mentioned above, not only is there a risk for false-negative detection results, especially when the viruses being detected are present in low numbers, as well as the use of multiple primers in a single amplification reaction may impact the sensitivity of viral detection, even when a nested-PCR protocol is used.

In sum, the detection and genetic characterization of HPyV sequences are essential to obtain information regarding the distribution patterns this group of viruses presents in the general population. Besides aiding in disclosing the true genetic diversity of HPyVs, studies as the one here described emphasize the potential that these viruses might have as markers of human fecal contamination. Even so, we must not forget that further developments still need to be made in regard to methodology in order to fully uncover the potential that an environmental viral analysis can bring. In the future, similar studies focusing on the analysis of HPyVs over a broader geographic range will be of great use to truly comprehend the diversity of this group of viruses in Portugal. Selecting a larger number of sampling

locations, and perhaps even including other, less explored types of water matrices, such as lakes, water fountains, ponds, and maybe, even swimming pools, can aid in obtaining a more complete picture regarding the different HPyVs circulating in the studied area. However, we must not forget that due to the ongoing globalization and disruptive population diasporas (caused by religious or ethnic prosecutions, or simply by the dream of a better life), the HPyV genotypes circulation in a specific location may change over time. Thus, to accurately portray the genetic diversity of these viruses in a specific location, studies like this one would have to be repeated periodically. Lastly, it would also be interesting to perform infectivity assays in untreated sewage vs. nearby environmental waters, to verify the effectiveness that the treatments carried out by the wastewater treatment plant have in neutralizing HPyVs.

5 | Bibliography

1. Fong T, Lipp EK. 2005. Enteric Viruses of Humans and Animals in Aquatic Environments: Health Risks, Detection, and Potential Water Quality Assessment Tools. Microbiol Mol Biol Rev 69:357–371 doi:10.1128/MMBR.69.2.357.
2. WHO. 2017. Guidelines for Drinking-water Quality, Fourth Ed. Geneva ISBN:9789241549950.
3. EU. 2006. Directive 2006/7/EC of the European Parliament and of the Council of 15 February 2006 concerning the management of bathing water quality and repealing Directive 76/160/EEC. Off J Eur Union L64/37-L64/51.
4. McQuaig SM, Scott TM, Harwood VJ, Farrah SR, Lukasik JO. 2006. Detection of Human-Derived Fecal Pollution in Environmental Waters by Use of a PCR-Based Human Polyomavirus Assay. Appl Environ Microbiol 72:7567–7574 doi:10.1128/AEM.01317-06.
5. Seurinck S, Verstraete W, Siciliano SD. 2005. Microbial source tracking for identification of fecal pollution. Rev Environ Sci Bio/Technology 4:19–37 doi:10.1007/s11157-005-4997-7.
6. Harwood VJ, Staley C, Badgley BD, Borges K, Korajkic A. 2014. Microbial source tracking markers for detection of fecal contamination in environmental waters: relationships between pathogens and human health outcomes. FEMS Microbiol Rev 38:1–40 doi:10.1111/1574-6976.12031.
7. USEPA. 2006. Bacteria: Indicators of Potential Pathogens, p. 25. *In* Agency, USEP (ed.), Volunteer Estuary Monitoring: A Methods Manual. Washigton, DC ISBN:EPA-842-B-06-003.
8. Scott TM, Rose JB, Jenkins TM, Farrah SR, Lukasik J. 2002. Microbial source tracking: Current methodology and future directions. Appl Environ Microbiol 68:5796–5803 doi:10.1128/AEM.68.12.5796-5803.2002.
9. Rochelle-Newall E, Nguyen TMH, Le TPQ, Sengtaheuanghoung O, Ribolzi O. 2015. A short review of fecal indicator bacteria in tropical aquatic ecosystems: Knowledge gaps and future directions. Front Microbiol 6:308 doi:10.3389/fmicb.2015.00308.
10. Field KG, Bernhard AE, Brodeur TJ. 2003. Molecular approaches to microbiological monitoring: fecal source detection. Environ Monit Assess 81:313–326.
11. Stoeckel DM, Harwood VJ. 2007. Performance, Design, and Analysis in Microbial Source Tracking Studies. Am Soc Microbiol 73:2405–2415 doi:10.1128/AEM.02473-06.
12. Simpson JM, Domingo JWS, Reasoner DJ. 2002. Microbial Source Tracking: State of the Science. Environ Sci Technol 36:5279–5288 doi:10.1021/es026000b.
13. Ahmed W. 2007. Limitations of library-dependent microbial source tracking methods. J Aust Water Assoc 34:96–101.
14. Muniesa M, Payan A, Moce-Llivina L, Blanch AR, Jofre J. 2009. Differential persistence of F-specific RNA phage subgroups hinders their use as single tracers for faecal source tracking in surface water. Water Res 43:1559–1564 doi:10.1016/j.watres.2008.12.038.
15. Sinton LW, Hall CH, Lynch PA, Davies-Colley RJ. 2002. Sunlight inactivation of fecal indicator bacteria and bacteriophages from waste stabilization pond effluent in fresh and saline waters. Appl Environ Microbiol 68:1122–1131 doi:10.1128/AEM.68.3.1122-1131.2002.
16. Ahmed W, Hughes B, Harwood VJ. 2016. Current status of marker genes of bacteroides and related taxa for identifying sewage pollution in environmental waters. Water 8:1–27 doi:10.3390/w8060231.

17. Jofre J, Blanch AR, Lucena F, Muniesa M. 2014. Bacteriophages infecting Bacteroides as a marker for microbial source tracking. Water Res 55:1–11 doi:10.1016/j.watres.2014.02.006.

18. Tartera C, Jofre J. 1987. Bacteriophages active against Bacteroides fragilis in sewage-polluted waters. Appl Environ Microbiol 53:1632–1637 doi:10.1128/aem.53.7.1632-1637.1987.

19. Scott TM, Jenkins TM, Lukasik J, Rose JB. 2005. Potential use of a host associated molecular marker in Enterococcus faecium as an index of human fecal pollution. Environ Sci Technol 39:283–287 doi:10.1021/es035267n.

20. Oshiro RK, Olson BH. 1997. Occurrence of STh toxin gene in wastewater, p. 255–259. *In* Kay, D, Fricher, C (eds.), Coliforms and E. coli: Problem or Solution?, First Ed. The Royal Society of Chemsitry, Cambridge ISBN:0854047719.

21. Khatib LA, Tsai YL, Olson BH. 2002. A biomarker for the identification of cattle fecal pollution in water using the LTIIa toxin gene from enterotoxigenic Escherichia coli. Appl Microbiol Biotechnol 59:97–104 doi:10.1007/s00253-002-0959-y.

22. Field KG, Chern EC, Dick LK, Fuhrman J, Griffith J, Holden PA, LaMontagne MG, Le J, Olson B, Simonich MT. 2003. A comparative study of culture-independent, library-independent genotypic methods of fecal source tracking. J Water Health 1:181–194 doi:10.2166/wh.2003.0020.

23. McQuaig SM, Noble RT. 2011. Viruses as Tracers of Fecal Contamination, p. 113–135. *In* Hagedorn, C, Blanch, RA, Harwood, VJ (eds.), Microbial Source Tracking: Methods, Applications, and Case Studies ISBN:9781441993854.

24. Troeger C, Blacker BF, Khalil IA, Rao PC, Cao S, Zimsen SR, Albertson SB, Stanaway JD, Deshpande A, Abebe Z, Alvis-Guzman N, Amare AT, Asgedom SW, Anteneh ZA, Antonio CAT, Aremu O, Asfaw ET, Atey TM, Atique S, Avokpaho EFGA, Awasthi A, Ayele HT, Barac A, Barreto ML, Bassat Q, Belay SA, Bensenor IM, Bhutta ZA, Bijani A, Bizuneh H, Castañeda-Orjuela CA, Dadi AF, Dandona L, Dandona R, Do HP, Dubey M, Dubljanin E, Edessa D, Endries AY, Eshrati B, Farag T, Feyissa GT, Foreman KJ, Forouzanfar MH, Fullman N, Gething PW, Gishu MD, Godwin WW, Gugnani HC, Gupta R, Hailu GB, Hassen HY, Hibstu DT, Ilesanmi OS, Jonas JB, Kahsay A, Kang G, Kasaeian A, Khader YS, Khan EA, Khan MA, Khang YH, Kissoon N, Kochhar S, Kotloff KL, Koyanagi A, Kumar GA, Magdy Abd El Razek H, Malekzadeh R, Malta DC, Mehata S, Mendoza W, Mengistu DT, Menota BG, Mezgebe HB, Mlashu FW, Murthy S, Naik GA, Nguyen CT, Nguyen TH, Ningrum DNA, Ogbo FA, Olagunju AT, Paudel D, Platts-Mills JA, Qorbani M, Rafay A, Rai RK, Rana SM, Ranabhat CL, Rasella D, Ray SE, Reis C, Renzaho AM, Rezai MS, Ruhago GM, Safiri S, Salomon JA, Sanabria JR, Sartorius B, Sawhney M, Sepanlou SG, Shigematsu M, Sisay M, Somayaji R, Sreeramareddy CT, Sykes BL, Taffere GR, Topor-Madry R, Tran BX, Tuem KB, Ukwaja KN, Vollset SE, Walson JL, Weaver MR, Weldegwergs KG, Werdecker A, Workicho A, Yenesew M, Yirsaw BD, Yonemoto N, El Sayed Zaki M, Vos T, Lim SS, Naghavi M, Murray CJ, Mokdad AH, Hay SI, Reiner RC. 2018. Estimates of the global, regional, and national morbidity, mortality, and aetiologies of diarrhoea in 195 countries: a systematic analysis for the Global Burden of Disease Study 2016. Lancet Infect Dis 18:1211–1228 doi:10.1016/S1473-3099(18)30362-1.

25. Mawatari M, Kato Y. 2009. Norovirus Gastroenteritis. N Engl J Med 361:1776–1785 doi:10.1016/B978-0-12-416975-3.00016-9.

26. Upfold NS, Luke GA, Knox C. 2021. Occurrence of Human Enteric Viruses in Water Sources and Shellfish: A Focus on Africa.Food and Environmental Virology. Springer US doi:10.1007/s12560-020-09456-8.

27. Noble RT, Fuhrman JA. 2001. Enteroviruses detected by reverse transcriptase polymerase chain reaction from the coastal waters of Santa Monica Bay, California: Low correlation to bacterial indicator levels. Hydrobiologia 460:175–184 doi:10.1023/A:1013121416891.

28. Jiang S, Noble R, Chu W. 2001. Human adenoviruses and coliphages in urban runoff-impacted coastal waters of Southern California. Appl Environ Microbiol 67:179–184 doi:10.1128/AEM.67.1.179-184.2001.

29. Rachmadi AT, Torrey JR, Kitajima M. 2016. Human polyomavirus: Advantages and limitations as a human-specific viral marker in aquatic environments. Water Res 105:456–469 doi:10.1016/j.watres.2016.09.010.

30. McQuaig SM, Scott TM, Lukasik JO, Paul JH, Harwood VJ. 2009. Quantification of human polyomaviruses JC virus and BK Virus by TaqMan quantitative PCR and comparison to other water quality indicators in water and fecal samples. Appl Environ Microbiol 75:3379–3388 doi:10.1128/AEM.02302-08.

31. Calvignac-Spencer S, Feltkamp MCW, Daugherty MD, Moens U, Ramqvist T, Johne R, Ehlers B. 2016. A taxonomy update for the family Polyomaviridae. Arch Virol 161:1739–1750 doi:10.1007/s00705-016-2794-y.

32. Moens U, Calvignac-Spencer S, Lauber C, Ramqvist T, Feltkamp MC., Daugherty MD, Verschoor EJ, Ehlers B. 2020. Report on the taxonomy of the Polyomaviridae. Virus Taxonomy: 2020 Release.

33. Imperiale MJ, Major EO, DeCaprio JA. 2013. Polyomaviruses, p. 1633–1661. *In* Knipe, DM, Howley, PM (eds.), Fields Virology, Sixth Ed. Wolter Kluwer/Lippincott Williams & Williams ISBN:9781451105636.

34. Baez CF, Brandão Varella R, Villani S, Delbue S. 2017. Human Polyomaviruses: The Battle of Large and Small Tumor Antigens. Virol Res Treat 8:1–12 doi:10.1177/1178122X17744785.

35. Hurdiss DL, Frank M, Snowden JS, Macdonald A, Ranson NA. 2018. The Structure of an Infectious Human Polyomavirus and Its Interactions with Cellular Receptors. Structure 26:839-847.e3 doi:10.1016/j.str.2018.03.019.

36. Saribas AS, Coric P, Hamazaspyan A, Davis W, Axman R, White MK, Abou-Gharbia M, Childers W, Condra JH, Bouaziz S, Safak M. 2016. Emerging From the Unknown: Structural and Functional Features of Agnoprotein of Polyomaviruses. J Cell Physiol 231:2115–2127 doi:10.1002/jcp.25329.

37. van der Meijden E, Feltkamp M. 2018. The human polyomavirus middle and alternative T-antigens; thoughts on roles and relevance to cancer. Front Microbiol 9:1–8 doi:10.3389/fmicb.2018.00398.

38. van der Meijden E, Kazem S, Dargel CA, van Vuren N, Hensbergen PJ, Feltkamp MCW. 2015. Characterization of T Antigens, Including Middle T and Alternative T, Expressed by the Human Polyomavirus Associated with Trichodysplasia Spinulosa. J Virol 89:9427–9439 doi:10.1128/jvi.00911-15.

39. Gossai A, Waterboer T, Nelson HH, Michel A, Willhauck-Fleckenstein M, Farzan SF, Hoen AG, Christensen BC, Kelsey KT, Marsit CJ, Pawlita M, Karagas MR. 2016. Seroepidemiology of Human Polyomaviruses in a US Population. Am J Epidemiol 183:61–69 doi:10.1093/aje/kwv155.

40. Kamminga S, Meijden E Van Der, Feltkamp MCW, Zaaijer HL. 2018. Seroprevalence of fourteen human polyomaviruses determined in blood donors. PLoS One 13:1–11 doi:10.1371/journal.pone.0206273.

41. Ciotti M, Prezioso C, Pietropaolo V. 2019. An overview on human polyomaviruses biology and related diseases. Future Virol 14:487–501 doi:10.2217/fvl-2019-0050.

42. Gardner SD, Field AM, Coleman D V., Hulme B. 1971. New Human Papovavirus (B.K.) Isolated From Urine After Renal Transplantation. Lancet 297:1253–1257 doi:10.1016/S0140-

6736(71)91776-4.

43. Spencer ES, Andersen HK, Padgett BL, Zurhein GM, Walker DL, Eckroade RJ, Dessel BH. 1971. Cultivation of papova-like virus from human brain with progressive multifocal leucoencephalopathy. Lancet 297:1257–1260 doi:https://doi.org/10.1016/S0140-6736(71)91777-6.

44. Polo C, Pérez JL, Mielnichuck A, Fedele CG, Niubó J, Tenorio A. 2004. Prevalence and patterns of polyomavirus urinary excretion in immunocompetent adults and children. Clin Microbiol Infect 10:640–644 doi:10.1111/j.1469-0691.2004.00882.x.

45. Vanchiere JA, Nicome RK, Greer JM, Demmler GJ, Butel JS. 2005. Frequent detection of polyomaviruses in stool samples from hospitalized children. J Infect Dis 192:658–664 doi:10.1086/432076.

46. Goudsmit J, Dillen PW, van Strien A, van der Noordaa J. 1982. The role of BK virus in acute respiratory tract disease and the presence of BKV DNA in tonsils. J Med Virol 10:91–99 doi:10.1002/jmv.1890100203.

47. Monaco MCG, Jensen PN, Hou J, Durham LC, Major EO. 1998. Detection of JC Virus DNA in Human Tonsil Tissue: Evidence for Site of Initial Viral Infection. J Virol 72:9918–9923 doi:10.1128/jvi.72.12.9918-9923.1998.

48. Chesters PM, Heritage J, Mccance DJ. 1983. Persistence of DNA Sequences of BK Virus and JC Virus in Normal Human Tissues and in Diseased Tissues. J Infect Dis 147:676–684 doi:10.1093/infdis/147.4.676.

49. Jin L, Gibson PE, Booth JC, Clewley JP. 1993. Genomic typing of BK virus in clinical specimens by direct sequencing of Polymerase Chain Reaction products. J Med Virol 41:11–17 doi:10.1002/jmv.1890410104.

50. Ikegaya H, Saukko PJ, Tertti R, Metsa KP, Carr MJ, Crowley B, Sakurada K, Zheng H, Kitamura T, Yogo Y. 2006. Identification of a genomic subgroup of BK polyomavirus spread in European populations. J Gen Virol 87:3201–3208 doi:10.1099/vir.0.82266-0.

51. Takasaka T, Goya N, Tokumoto T, Tanabe K, Toma H, Ogawa Y, Hokama S, Momose A, Funyu T, Fujioka T, Omori S, Akiyama H, Chen Q, Zheng H, Ohta N, Kitamura T, Yogo Y. 2004. Subtypes of BK virus prevalent in Japan and variation in their transcriptional control region. J Gen Virol 85:2821–2827 doi:10.1099/vir.0.80363-0.

52. Nishimoto Y, Zheng H, Zhong S, Ikegaya H, Chen Q, Sugimoto C, Kitamura T, Yogo Y. 2007. An Asian Origin for Subtype IV BK Virus Based on Phylogenetic Analysis. J Mol Evol 65:103–111 doi:10.1007/s00239-006-0269-6.

53. Zheng H, Nishimoto Y, Chen Q, Hasegawa M. 2007. Relationships between BK virus lineages and human populations. Microbes Infect 9:204–213 doi:10.1016/j.micinf.2006.11.008.

54. Houff SA, Major EO, Katz DA, Kufta C V., Sever JL, Pittaluga S, Roberts JR, Gitt J, Saini N, Lux W. 1988. Involvement of JC virus-infected mononuclear cells from the bone marrow and spleen in the pathogenesis of progressive multifocal leukoencephalopathy. N Engl J Med 318:301–305 doi:10.1056/NEJM198802043180507.

55. Sugimoto C, Kitamura T, Guo J, Al-ahdal MN, Shchelkunov SN, Otova B, Ondrejka P, Chollet JY, El-Safi S, Ettayebi M, Grésenguet G, Kocagoz T, Chaiyarasamee S, Thant KZ, Thein S, Moe K, Kobayashi N, Taguchi F, Yogo Y. 1997. Typing of urinary JC virus DNA offers a novel means of tracing human migrations. Proc Natl Acad Sci U S A 94:9191–9196 doi:10.1073/pnas.94.17.9191.

56. Yogo Y, Sugimoto C, Zheng H, Ikegaya H, Takasaka T, Kitamura T. 2004. JC virus genotyping offers a new paradigm in the study of human populations. Rev Med Virol 14:179–191

doi:10.1002/rmv.428.

57. Stoner GL, Jobes D V, Fernandez M, Agostini HT, Chima SC, Ryschkewitsch CF. 2000. JC virus as a marker of human migration to the Americas. Microbes Infect 2:1905–1911 doi:10.1016/s1286-4579(00)01339-3.

58. Torres C. 2020. Evolution and molecular epidemiology of polyomaviruses. Infect Genet Evol 79:1–17 doi:10.1016/j.meegid.2019.104150.

59. Saruwatari L, Sugimoto C, Kitamura T, Ohno N, Sakai E, Shresta P, Hoa BK, Phi PTP, An HPH, Tuyet NTA, Honjo T, Kobayashi N, Takasaka T, Yogo Y. 2002. Asian domains of four major genotypes of JC virus, Af2, B1-b, CY and SC. Arch Virol 147:1–10 doi:10.1007/s705-002-8299-4.

60. Agostini T, Shishido-hara Y, Baumhefner RW, Singer EJ, Ryschkewitsch CF, Stoner GL. 1998. JC virus Type 2 : definition of subtypes based on DNA sequence analysis of ten complete genomes. J Gen Virol 79:1143–1151 doi:10.1099/0022-1317-79-5-1143.

61. Cui X, Wang JC, Deckhut A, Joseph BC, Eberwein P, Cubitt CL, Ryschkewitsch CF, Agostini HT, Stoner GL. 2004. Chinese Strains (Type 7) of JC Virus Are Afro-Asiatic in Origin But Are Phylogenetically Distinct from the Mongolian and Indian Strains (Type 2D) and the Korean and Japanese Strains (Type 2A). J Mol Evol 58:568–583 doi:10.1007/s00239-003-2579-2.

62. Jobes D V, Friedlaender JS, Mgone CS, Agostini HT, Koki G. 2001. New JC virus (JCV) genotypes from Papua New Guinea and Micronesia (Type 8 and Type 2E) and evolutionary analysis of 32 complete JCV genomes. Arch Virol 146:2097–2113 doi:10.1007/s007050170023.

63. Takasaka T, Kitamura T, Sugimoto C, Guo J, Zheng H, Yogo Y. 2006. Phylogenetic Analysis of Major African Genotype (Af2) of JC Virus : Implications for Origin and Dispersals of Modern Africans. Am J Phys Anthropol 129:465–472 doi:10.1002/ajpa.20208.

64. Yanagihara R, Nerurkar VR, Scheirich I, Agostini HT, Mgone CS, Cui X, Jobes D V, Cubitt CL, Ryschkewitsch F, Hrdy DB, Friedlaender JS, Stoner L. 2002. JC Virus Genotypes in the Western Pacific Suggest Asian Mainland Relationships and Virus Association with Early Population Movements. Hum Biol 74:473–488 doi:10.1353/hub.2002.0037.

65. Allander T, Andreasson K, Gupta S, Bjerkner A, Bogdanovic G, Persson MAA, Dalianis T. 2007. Identification of a Third Human Polyomavirus. J Virol 81:4130–4136 doi:10.1128/JVI.00028-07.

66. Gaynor AM, Nissen MD, Whiley DM, Mackay IM, Lambert SB, Wu G, Brennan DC, Storch GA, Sloots TP, Wang D. 2007. Identification of a Novel Polyomavirus from Patients with Acute Respiratory Tract Infections. PloS Pathog 3:595–604 doi:10.1371/journal.ppat.0030064.

67. Mourez T, Bergeron A, Ribaud P, Scieux C, Latour RP De, Tazi A, Socié G, Simon F. 2009. Polyomaviruses KI and WU in Immunocompromised Patients with Respiratory Disease. Emerg Infect Dis 15:107–109 doi:10.3201/1501.080758.

68. Abed Y, Wang D, Boivin G. 2007. WU Polyomavirus in Children, Canada. Emerg Infect Dis 13:1939–1941 doi:10.3201/eid1312.070909.

69. Bialasiewicz S, Whiley DM, Lambert SB, Wang D, Nissen MD, Sloots TP. 2007. A newly reported human polyomavirus, KI virus, is present in the respiratory tract of Australian children. J Clin Virol 40:15–8 doi:10.1016/j.jcv.2007.07.001.

70. Motamedi N, Mairhofer H, Nitschko H, Jäger G, Koszinowski UH. 2012. The polyomaviruses WUPyV and KIPyV: a retrospective quantitative analysis in patients undergoing hematopoietic stem cell transplantation. Virol J 9:209 doi:10.1186/1743-422X-9-209.

71. Prezioso C, Ciotti M, Obregon F, Ambroselli D, Maria D, Cudillo L, Gaziev J, Mele A, Nardi

A, Favalli C, Arcese W, Palamara AT, Pietropaolo V. 2019. Polyomaviruses shedding in stool of patients with hematological disorders : detection analysis and study of the non-coding control region's genetic variability. Med Microbiol Immunol 208:845–854 doi:10.1007/s00430-019-00630-9.

72. Bialasiewicz S, Rockett R, Whiley DW, Abed Y, Allander T, Binks M, Boivin G, Cheng AC, Chung J, Ferguson PE, Gilroy NM, Leach AJ, Lindau C, Rossen JW, Sorrell TC, Nissen MD, Sloots TP. 2010. Whole-Genome Characterization and Genotyping of Global WU Polyomavirus Strains. J Virol 84:6229–6234 doi:10.1128/JVI.02658-09.

73. Zhu T, Lu Q, Zhang S, Wo Y, Zhuang L, Zhang P. 2017. Molecular epidemiology of WU polyomavirus in hospitalized children with acute respiratory tract infection in China. Future Microbiol 12:481–489 doi:10.2217/fmb-2016-0144.

74. Venter M, Visser A, Lassauniere R. 2009. Human polyomaviruses, WU and KI in HIV exposed children with acute lower respiratory tract infections in hospitals in South Africa. J Clin Virol 44:230–234 doi:10.1016/j.jcv.2008.12.007.

75. Schowalter RM, Pastrana D V, Pumphrey KA, Moyer AL, Buck CB. 2010. Merkel Cell Polyomavirus and Two Previously Unknown Polyomaviruses Are Chronically Shed from Human Skin. Cell Host Microbe 7:509–515 doi:10.1016/j.chom.2010.05.006.

76. Feng H, Shuda M, Chang Y, Moore PS. 2008. Clonal Integration of a Polyomavirus in Human Merkel Cell Carcinoma. Science 319:1095–1100 doi:10.1126/science.1152586.

77. Kantola K, Sadeghi M, Lahtinen A, Koskenvuo M, Aaltonen L, Möttönen M, Rahiala J, Saarinen-pihkala U, Riikonen P. 2009. Merkel cell polyomavirus DNA in tumor-free tonsillar tissues and upper respiratory tract samples: implications for respiratory transmission and latency. J Clin Virol 45:292–295 doi:10.1016/j.jcv.2009.04.008.

78. Signorini L, Belingheri M, Ambrogi F, Pagani E, Binda S, Ticozzi R, Ferraresso M, Ghio L, Giacon B, Ferrante P, Delbue S. 2014. High frequency of Merkel cell polyomavirus DNA in the urine of kidney transplant recipients and healthy controls. J Clin Virol 61:565–570 doi:10.1016/j.jcv.2014.10.012.

79. Martel-Jantin C, Filippone C, Tortevoye P, Afonso P V, Betsem E. 2014. Molecular Epidemiology of Merkel Cell Polyomavirus: Evidence for Geographically Related Variant Genotypes. J Clin Microbiol 52:1687–1690 doi:10.1128/JCM.02348-13.

80. Baez CF, Diaz NC, Venceslau MT, Luz FB, Guimarães MAAM, Zalis MG, Varella RB. 2016. Phylogenetic and structural analysis of merkel cell polyomavirus VP1 in Brazilian samples. Virus Res 221:1–7 doi:10.1016/j.virusres.2016.05.004.

81. Hashida Y, Higuchi T, Matsui K, Shibata Y, Nakajima K, Sano S, Daibata M. 2018. Genetic Variability of the Noncoding Control Region of Cutaneous Merkel Cell Polyomavirus: Identification of Geographically Related Genotypes. J Infect Dis 217:1601–1611 doi:10.1093/infdis/jiy070.

82. Hasihda Y, Higuchi T, Matsui S, Nakajima K, Sano S, Daibata M. 2018. Prevalence and Genetic Variability of Human Polyomaviruses 6 and 7 in Healthy Skin Among Asymptomatic Individuals. J Infect Dis 217:483–493 doi:10.1093/infdis/jix516.

83. Rockett RJ, Sloots TP, Bowes S, Neill NO, Ye S, Robson J, Whiley DM, Lambert SB, Wang D, Nissen MD, Bialasiewicz S. 2013. Detection of Novel Polyomaviruses, TSPyV, HPyV6, HPyV7, HPyV9 and MWPyV in Feces, Urine, Blood, Respiratory Swabs and Cerebrospinal Fluid. PLoS One 8:e62764 doi:10.1371/journal.pone.0062764.

84. Klufah F, Mobaraki G, Liu D, Alharbi RA, Kurz AK, Speel EJM, Winnepenninckx V, zur Hausen A. 2021. Emerging role of human polyomaviruses 6 and 7 in human cancers. Infect

Agent Cancer 16:1–12 doi:10.1186/s13027-021-00374-3.

85. Janssens WA, Lauber C, Nico J, Bavinck B, Alexander E, Meijden E Van Der. 2010. Discovery of a New Human Polyomavirus Associated with Trichodysplasia Spinulosa in an Immunocompromized Patient. PloS Pathog 6:1–10 doi:10.1371/journal.ppat.1001024.

86. Kazem S, Meijden E Van Der, Kooijman S, Rosenberg AS, Hughey LC, Browning JC, Sadler G, Busam K, Pope E, Benoit T, Fleckman P, Vries E De, Eekhof JA, Feltkamp MCW. 2012. Trichodysplasia spinulosa is characterized by active polyomavirus infection. J Clin Virol 53:225–230 doi:10.1016/j.jcv.2011.11.007.

87. van der Meijden E, Horváth B, Nijland M, Vries K De. 2017. Primary Polyomavirus Infection, Not Reactivation, as the Cause of Trichodysplasia Spinulosa in Immunocompromised Patients. J Infect Dis 215:1–5 doi:10.1093/infdis/jiw403.

88. Chamseddin BH, Anh B, Tran PD, Lee EE, Diana V, Buck CB, Wang RC, Yasmine A. 2019. Trichodysplasia spinulosa in a child: Identification of trichodysplasia spinulosa-associated polyomavirus in skin, serum, and urine. Pediatr Dermatol 36:723–724 doi:10.1111/pde.13857.Trichodysplasia.

89. Scuda N, Ruprecht K, Liman P, Ku J, Hengel H, Ehlers B. 2011. A Novel Human Polyomavirus Closely Related to the African Green Monkey-Derived Lymphotropic Polyomavirus. J Virol 85:4586–4590 doi:10.1128/JVI.02602-10.

90. Siebrasse EA, Reyes A, Lim ES, Zhao G, Mkakosya RS, Manary MJ, Gordon JI, Wang D. 2012. Identification of MW Polyomavirus, a Novel Polyomavirus in Human Stool. J Virol 86:10321–10326 doi:10.1128/JVI.01210-12.

91. Buck CB, Phan GQ, Raiji MT, Murphy PM, Mcdermott DH, Mcbride AA. 2012. Complete Genome Sequence of a Tenth Human Polyomavirus. J Virol 86:10887 doi:10.1128/JVI.01690-12.

92. Yu G, Greninger AL, Isa P, Phan TG, Martı MA, Santos-preciado I, Parsonnet J, Miller S, Sanchez L, Contreras JF, Derisi JL, Delwart E, Arias CF, Chiu CY. 2012. Discovery of a Novel Polyomavirus in Acute Diarrheal Samples from Children. PLoS One 7:1–10 doi:10.1371/journal.pone.0049449.

93. Lim ES, Reyes A, Antonio M, Saha D, Ikumapayi UN, Adeyemi M, Stine OC, Skelton R, Brennan DC, Mkakosya RS, Manary MJ, Gordon JI, Wang D. 2013. Discovery of STL polyomavirus, a polyomavirus of ancestral recombinant origin that encodes a unique T antigen by alternative splicing. Virology 436:295–303 doi:10.1016/j.virol.2012.12.005.

94. Pinheiro MS, Mendes GS, Santos N. 2020. Human polyomaviruses 10 and 11 in faecal samples from Brazilian children. Brazilian J Microbiol 51:585–591 doi:10.1007/s42770-019-00166-3.

95. Ondov BD, Starrett GJ, Sappington A, Kostic A, Koren S, Buck CB, Phillippy AM. 2019. Mash Screen: high-throughput sequence containment estimation for genome discovery. Genome Biol 20:232 doi:10.1186/s13059-019-1841-x.

96. Gheit T, Dutta S, Oliver J, Robitaille A, Hampras S, Combes D, Mckay-chopin S, Calvez-kelm F Le, Fenske N, Cherpelis B, Giuliano AR, Franceschi S, Mckay J, Rollison DE, Tommasino M. 2017. Isolation and characterization of a novel putative human polyomavirus. Virology 506:45–54 doi:10.1016/j.virol.2017.03.007.

97. Mishra N, Pereira M, Rhodes RH, An P. 2014. Identification of a Novel Polyomavirus in a Pancreatic Transplant Recipient With Retinal Blindness and Vasculitic Myopathy. J Infect Dis 210:1595–1599 doi:10.1093/infdis/jiu250.

98. Trusch F, Moens U, Sauer I, Voigt S, Schmuck R, Ehlers B. 2013. Identification of a Novel Human Polyomavirus in Organs of the Gastrointestinal Tract. PLoS One 8:e58021

doi:10.1371/journal.pone.0058021.

99. Gedvilaite A, Tryland M, Ulrich RG, Schneider J, Kurmauskaite V, Moens U, Preugschas H. 2017. Novel polyomaviruses in shrews (Soricidae) with close similarity to human polyomavirus 12. J Gen Virol 98:3060–3067 doi:10.1099/jgv.0.000948.

100. Fahsbender E, Altan E, Estrada M, Seguin MA, Young P, Leutenegger CM. 2019. Lyon-IARC Polyomavirus DNA in Feces of Diarrheic Cats. Microbiol Resour Announc 8:e00550-19 doi:10.1128/MRA.00550-19.

101. Kamminga S, van der Meijden E, Brouwer C De, Feltkamp M, Zaaijer H. 2019. Prevalence of DNA of fourteen human polyomaviruses determined in blood donors. Transfusion 59:3689–3697 doi:10.1111/trf.15557.

102. Prezioso C, Ghelue M Van, Pietropaolo V. 2021. Detection of Quebec Polyomavirus DNA in Samples from Different Patient Groups. Microorganisms 9:1082 doi:10.3390/microorganisms9051082.

103. Instituto Nacional de Estatística. 2011. Censos 2011 Resultados Definitivos - Portugal. Lisboa ISBN:978-989-25-0181-9.

104. Calgua B, Mengewein A, Grunert A, Bofill-Mas S, Clemente-Casares P, Hundesa A, Wyn-Jones AP, López-Pila JM, Girones R. 2008. Development and application of a one-step low cost procedure to concentrate viruses from seawater samples. J Virol Methods 153:79–83 doi:10.1016/j.jviromet.2008.08.003.

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

106. Yoon H, Leitner T. 2015. PrimerDesign-M: A multiple-alignment based multiple-primer design tool for walking across variable genomes. Bioinformatics 31:1472–1474 doi:10.1093/bioinformatics/btu832.

107. Chung CT, Niemela SL, Miller RH. 1989. One-step preparation of competent Escherichia coli: transformation and storage of bacterial cells in the same solution. Proc Natl Acad Sci U S A 86:2172–2175 doi:10.1073/pnas.86.7.2172.

108. Hanahan D. 1983. Studies on transformation of Escherichia coli with plasmids. J Mol Biol 166:557–580 doi:10.1016/S0022-2836(83)80284-8.

109. Birnboim HC, Doly J. 1979. A rapid alkaline extraction procedure for screening recombinant plasmid DNA. Nucleic Acids Res 7:1513–1523 doi:10.1093/nar/7.6.1513.

110. Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain terminating inhibitors. Proc Natl Acad Sci U S A 74:5463–5467 doi:10.1073/pnas.74.12.5463.

111. Hall TA. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser.

112. Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17:540–552 doi:10.1093/oxfordjournals.molbev.a026334.

113. Schmidt HA, Strimmer K, Vingron M, von Haeseler A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18:502–504 doi:10.1093/bioinformatics/18.3.502.

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

115. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. 2018. Bayesian

phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol 4:1–5 doi:10.1093/ve/vey016.

116. Ho SYW, Phillips MJ, Drummond AJ, Cooper A. 2005. Accuracy of rate estimation using relaxed-clock models with a critical focus on the early metazoan radiation. Mol Biol Evol 22:1355–1363 doi:10.1093/molbev/msi125.

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

118. Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol 23:254–267 doi:10.1093/molbev/msj030.

119. Martin DP, Murrell B, Golden M, Khoosal A, Muhire B. 2015. RDP4: Detection and analysis of recombination patterns in virus genomes. Virus Evol 1:1–5 doi:10.1093/ve/vev003.

120. Vilsker M, Moosa Y, Nooij S, Fonseca V, Ghysens Y, Dumon K, Pauwels R, Alcantara LC, Vanden Eynden E, Vandamme AM, Deforche K, De Oliveira T. 2019. Genome Detective: An automated system for virus identification from high-throughput sequencing data. Bioinformatics 35:871–873 doi:10.1093/bioinformatics/bty695.

121. Korbie DJ, Mattick JS. 2008. Touchdown PCR for increased specificity and sensitivity in PCR amplification. Nat Protoc 3:1452–1456 doi:10.1038/nprot.2008.133.

122. Green MR, Sambrook J. 2019. Nested polymerase chain reaction (PCR). Cold Spring Harb Protoc 2019:175–179 doi:10.1101/pdb.prot095182.

123. Griffith JF, Cao Y, McGee CD, Weisberg SB. 2009. Evaluation of rapid methods and novel indicators for assessing microbiological beach water quality. Water Res 43:4900–4907 doi:10.1016/j.watres.2009.09.017.

124. Calgua B, Fumian T, Rusiñol M, Rodriguez-Manzano J, Mbayed VA, Bofill-Mas S, Miagostovich M, Girones R. 2013. Detection and quantification of classic and emerging viruses by skimmed-milk flocculation and PCR in river water from two geographical areas. Water Res 47:2797–2810 doi:10.1016/j.watres.2013.02.043.

125. Torres C, Barrios ME, Cammarata RV, Cisterna DM, Estrada T, Martini Novas S, Cahn P, Blanco Fernández MD, Mbayed VA. 2016. High diversity of human polyomaviruses in environmental and clinical samples in Argentina: Detection of JC, BK, Merkel-cell, Malawi, and human 6 and 7 polyomaviruses. Sci Total Environ 542:192–202 doi:10.1016/j.scitotenv.2015.10.047.

126. Fumian TM, Guimarães FR, Vaz BJP, Da Silva MTT, Muylaert FF, Bofill-Mas S, Gironés R, Leite JPG, Miagostovich MP. 2010. Molecular detection, quantification and characterization of human polyomavirus JC from waste water in Rio de Janeiro, Brazil. J Water Health 8:438–445 doi:10.2166/wh.2010.090.

127. Iaconelli M, Petricca S, Libera DS, Di Bonito P, La Rosa G. 2015. First Detection of Human Papillomaviruses and Human Polyomaviruses in River Waters in Italy. Food Environ Virol 7:309–315 doi:10.1007/s12560-015-9203-7.

128. Bofill-Mas S, Pina S, Girones R. 2000. Documenting the epidemiologic patterns of polyomaviruses in human populations by studying their presence in urban sewage. Appl Environ Microbiol 66:238–245 doi:10.1128/AEM.66.1.238-245.2000.

129. Levican J, Levican A, Ampuero M, Gaggero A. 2019. JC polyomavirus circulation in one-year surveillance in wastewater in Santiago, Chile. Infect Genet Evol 71:151–158 doi:10.1016/j.meegid.2019.03.017.

130. Yang Z. 1998. On the Best Evolutionary Rate for Phylogenetic Analysis. Syst Biol 47:12–17 doi:10.1080/106351598261067.

131. Morel V, Martin E, François C, Helle F, Faucher J, Mourez T, Choukroun G, Duverlie G, Castelain S, Brochot E. 2017. A Simple and Reliable Strategy for BK Virus Subtyping and Subgrouping. J Clin Microbiol 55:1177–1185 doi:10.1128/JCM.01180-16.

132. Dubois V, Moret H, Lafon ME, Brodard V, Icart J, Ruffault A, Guist'hau O, Buffet-Janvresse C, Abbed K, Dussaix E, Ingrand D. 2001. JC virus genotypes in France: Molecular epidemiology and potential significance for progressive multifocal leukoencephalopathy. J Infect Dis 183:213–217 doi:10.1086/317927.

133. Agostini HT, Yanagihara R, Davis V, Ryschkewitsch CF, Stoner GL. 1997. Asian genotypes of JC virus in Native Americans and in a Pacific Island population: Markers of viral evolution and human migration. Proc Natl Acad Sci U S A 94:14542–14546 doi:10.1073/pnas.94.26.14542.

134. Strimmer K, von Haeseler A. 1997. Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. Proc Natl Acad Sci U S A 94:6815–9.

135. Ahmed W, Wan C, Goonetilleke A, Gardner T. 2010. Evaluating Sewage-Associated JCV and BKV Polyomaviruses for Sourcing Human Fecal Pollution in a Coastal River in Southeast Queensland, Australia. J Environ Qual 39:1743–1750 doi:10.2134/jeq2010.0062.

136. Sidhu JPS, Hodgers L, Ahmed W, Chong MN, Toze S. 2012. Prevalence of human pathogens and indicators in stormwater runoff in Brisbane, Australia. Water Res 46:6652–6660 doi:10.1016/j.watres.2012.03.012.

137. Dias J, Pinto RN, Vieira CB, de Abreu Corrêa A. 2018. Detection and quantification of human adenovirus (HAdV), JC polyomavirus (JCPyV) and hepatitis A virus (HAV) in recreational waters of Niterói, Rio de Janeiro, Brazil. Mar Pollut Bull 133:240–245 doi:10.1016/j.marpolbul.2018.05.031.

138. Rusiñol M, Fernandez-Cassi X, Hundesa A, Vieira C, Kern A, Eriksson I, Ziros P, Kay D, Miagostovich M, Vargha M, Allard A, Vantarakis A, Wyn-Jones P, Bofill-Mas S, Girones R. 2014. Application of human and animal viral microbial source tracking tools in fresh and marine waters from five different geographical areas. Water Res 59:119–129 doi:10.1016/j.watres.2014.04.013.

139. Bofill-Mas S, Albinana-Gimenez N, Clemente-Casares P, Hundesa A, Rodriguez-Manzano J, Allard A, Calvo M, Girones R. 2006. Quantification and stability of human adenoviruses and polyomavirus JCPyV in wastewater matrices. Appl Environ Microbiol 72:7894–7896 doi:10.1128/AEM.00965-06.

140. Jurzik L, Hamza IA, Wilhelm M. 2015. Investigating the Reduction of Human Adenovirus (HAdV) and Human Polyomavirus (HPyV) in a Sewage Treatment Plant with a Polishing Pond as a Tertiary Treatment. Water Air Soil Pollut 226:1–8 doi:10.1007/s11270-015-2545-9.

141. Kokkinos PA, Ziros PG, Mpalasopoulou G, Galanis A, Vantarakis A. 2011. Molecular detection of multiple viral targets in untreated urban sewage from Greece. Virol J 8:11–16 doi:10.1186/1743-422X-8-195.

142. Tandukar S, Ghaju Shrestha R, Malla B, Sthapit N, Sherchand JB, Sherchan SP, Haramoto E. 2021. Virus reduction at wastewater treatment plants in Nepal. Environ Challenges 5:100281 doi:10.1016/j.envc.2021.100281.

143. Di Bonito P, Libera S Della, Petricca S, Iaconelli M, Accardi L, Muscillo M, La Rosa G. 2015. Frequent and Abundant Merkel Cell Polyomavirus Detection in Urban Wastewaters in Italy. Food Environ Virol 7:1–6 doi:10.1007/s12560-014-9168-y.

144. Bofill-Mas S, Rodriguez-Manzano J, Calgua B, Carratala A, Girones R. 2010. Newly described

human polyomaviruses Merkel Cell, KI and WU are present in urban sewage and may represent potential environmental contaminants. Virol J 7:141 doi:10.1186/1743-422X-7-141.

145. Cantalupo PG, Calgua B, Zhao G, Hundesa A, Wier AD, Katz JP, Grabe M, Hendrix RW, Girones R, Wang D, Pipas JM. 2011. Raw sewage harbors diverse viral populations. MBio 2:e00180-11 doi:10.1128/mBio.00180-11.

146. Torres C, Barrios ME, Cammarata RV, Victoria M, Fernandez-Cassi X, Bofill-Mas S, Colina R, Blanco Fernández MD, Mbayed VA. 2018. Phylodynamics of Merkel-cell polyomavirus and human polyomavirus 6: A long-term history with humans. Mol Phylogenet Evol 126:210–220 doi:10.1016/j.ympev.2018.04.025.

147. Prezioso C, Van Ghelue M, Moens U, Pietropaolo V. 2021. HPyV6 and HPyV7 in urine from immunocompromised patients. Virol J 18:1–7 doi:10.1186/s12985-021-01496-1.

148. Babakir-Mina M, Ciccozzi M, Alteri C, Polchi P, Picardi A, Greco F, Lucarelli G, Arcese W, Perno CF, Perno M. 2009. Excretion of the novel polyomaviruses KI and WU in the stool of patients with hematological disorders. Antivir Ther 81:1668–1673 doi:10.1002/jmv.21559.

149. Rusiñol M, Fernandez-Cassi X, Timoneda N, Carratalà A, Abril JF, Silvera C, Figueras MJ, Gelati E, Rodó X, Kay D, Wyn-Jones P, Bofill-Mas S, Girones R. 2015. Evidence of viral dissemination and seasonality in a Mediterranean river catchment: Implications for water pollution management. J Environ Manage 159:58–67 doi:10.1016/j.jenvman.2015.05.019.

150. Pagani E, Delbue S, Mancuso R, Borghi E, Tarantini L, Ferrante P. 2003. Molecular analysis of JC virus genotypes circulating among the Italian healthy population. J Neurovirol 9:559–566 doi:10.1080/13550280390241269.

151. Agostini HT, Deckhut A, Jobes VD, Girones R, Schlunck G, Prost MG, Frias C, Pérez-Trallero E, Ryschkewitsch CF, Stoner GL. 2001. Genotypes of JC virus in East, Central and Southwest Europe. J Gen Virol 82:1221–1331 doi:10.1099/0022-1317-82-5-1221.