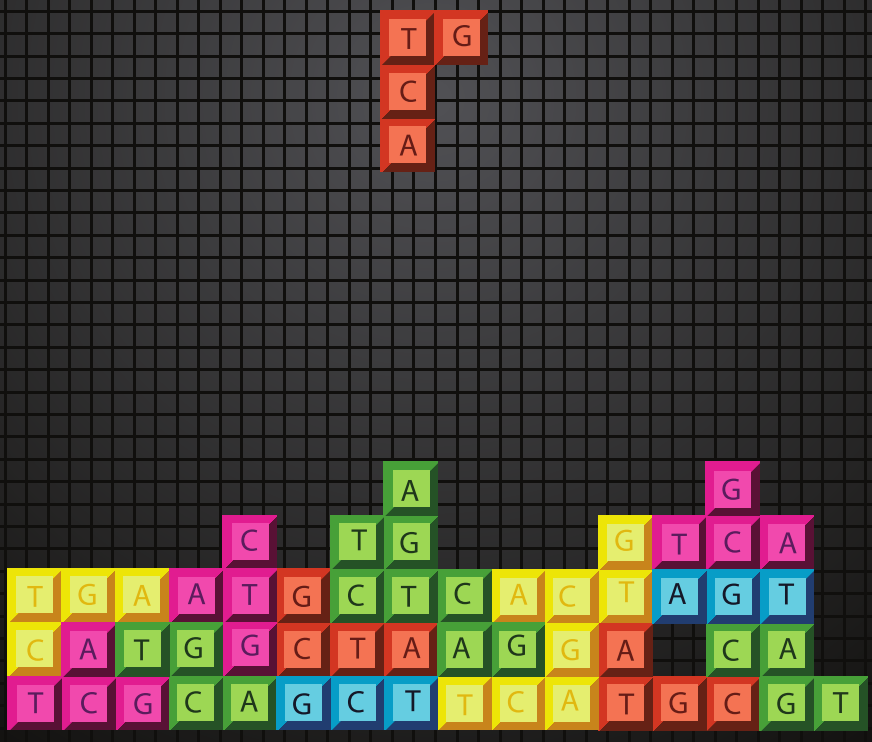
Mètodes computacionals per a la identificació de virus emergents analitzats per seqüenciació en massa en aigua i aliments

Natàlia Timoneda Solé

Tesi Doctoral

Mètodes computacionals per a la identificació de virus emergents analitzats per seqüenciació en massa en aigua i aliments



Natàlia Timoneda Solé / Maig 2017

— Imatge de Portada —

El "joc" d'ensamblar metagenomes...

Ideada per:
Josep F. Abril

Creada per:
Ricardo Montes de Oca

— Imatge de Contraportada —

Script de Perl...

Creada per:
Natàlia Timoneda

Facultat de biologia

Departament de genètica, microbiologia i estadística

Programa de doctorat en microbiologia ambiental i biotecnologia

—Any 2012—

TESI DOCTORAL

Mètodes computacionals per a la identificació de virus emergents analitzats per seqüenciació en massa en aigua i aliments.

Memòria presentada per

Natàlia Timoneda Solé

per optar al grau de

Doctora per la Universitat de Barcelona

Tesi doctoral realizada sota la direcció de
la Dra. Rosina Girones Llop i el Dr. Josep F. Abril Ferrando
al Departament de Genètica, Microbiologia i Estadística
de la Facultat de Biologia, de la Universitat de Barcelona.

Co-directora i Tutora

Dra. Rosina Gironés

Co-director

Dr. Josep F. Abril

L'autora

Natàlia Timoneda Solé

Barcelona, 28 de Maig de 2017

Agraïments

Ja he arribat aquí! I no ha estat fàcil, però sola no hagués estat capaç de fer-ho, i per això vull agrair a tota la gent que ha estat al meu costat d'alguna manera o una altra.

Primer vull donar les gràcies al Pep, al principi vaig pensar que em vas dir que sí per muntar estanteries d'Ikea, però després ja vaig veure que no, que també incloïa armaris de servidors!! Ara en serio, moltes gràcies per tot, per la teva paciència, per agunatar-me, per acollir-me al teu grup, per tot lo que m'has ensenyat, pels consells i per tots els projectes: començant per les planaries, el EXOBLAMER, *subobscura*, i finalment el gran món dels petits virus!

També agrair a la Rosina, per diversos motius, però en especial per veure la importància dels bioinformàtics en els grups de recerca; encara que de vegades sembla que parlem idiomes diferents, amb paciència ens hem acabat entenent. Gràcies per l'oportunitat de treballar al teu grup i per tot lo que he après del món dels virus, que abans d'arribar a Vircont era ben poc!!

Al Lab 8; gràcies per acollir-me amb els braços oberts des del primer dia; per la paciència, per tot lo que m'heu ensenyat i ajudat sempre en tot. A la Martus, per l'experiència i tots els consells; al David per les nostres xerrades d'ordinadors; a l'Eloy por toda la estadística, a la Sílvia per tots els consells, a la Sandra per tota l'energia, i gràcies a tota la resta de companys de Vircont: Eva, Ayalke, Laura, Natàlia, Texe, i tots els que han passat per allà. I a tu Xavi, Persi, què dir-te, gràcies per ser el meu company durant tot aquest llarg (i dur) viatge, que sense tu no hagués estat el mateix, més ben dit, sense tu no hagués estat possible!!! Gràcies per la paciència en la meva petita incursió al món de la *poyata*, i tot l'aprenentatge!!

Gràcies també a la resta de companys del departament de Micro, especialment a la Eli per les xerrades i compartir històries de les tietes, a la Montse per patir i resoldre els enigmes de les beques de doctorat, als del lab 5 per aguantar a la pesada sempre preguntant pel Xavi i moltes gràcies també a la gent que fa possible que la burocràcia sigui almenys una mica més fàcil i que no sé què faríem sense tots ells: Bea, Susana, Rosario, Manolo i a l'Aiora per la seva eficiència i ser la còmplice dels bioenigmes!

També agrair a CompGen; a tu Gus, gràcies per tot, tant a dins com a fora, pels consells, coneixements, totes les nostres xerrades i la plantilla de LaTeX! Frias, gràcies per tots els riures i per ser la meva companya de crims just arribar. Al Sergio per tota la ajuda amb el JavaScript i ser el meu estilista personal de pàgines web! A tu Sílvia, pel català correcte i tots els cotis!! I a tots els que han passat per aquí que no són pocs!! Les Martes, Joel, Josué, Tano, Marc, Bernat, Pepe i Núria.

Al grup del Xavier Barril, que durant molt temps van compartir despatx amb nosaltres, Peter, Mousomi, Laura, Sergi, Kevin, Yvonne, Dani i Montse. Els divendres eren una altra cosa amb una partideta a OpenArena. I a la nouvinguda, a la Gemma, per no espantar-se i marxar al veure les nostres pantalles!! Moltes gràcies per totes les nostres xerrades, consells i ser la proveïdora oficial de llet!

Gràcies a Gertjan Medema i a tot al seu grup de KWR Water de Utrecht, Holanda, per l'acollida i per tot l'aprenentatge de QMRA, sobretot a l'Helena pel curs accelerat del paquet *mcd2* en R. Al Frank Borgart, per deixar-me una bici, acollir-me a casa seva i ser el meu guia personal per Amsterdam i Utrecht.

A la *pandilla* dels dinars, a l'Emili per les històries de la facultat i departament; al Francesc per ser un bon lider de Clan i al Moi per la teva vitalitat i idees.

A les meves nenes: Sara, por aguantarnos mútuamente, las charlas y las broncas cuando te lo pido! Núria, m'has ajudat i ensenyat més del que et penses, gràcies sobretot per les sortides a la muntanya i rutes impressionants! Cris, per la xocolata i gelats per compensar el *running* per les morisques, i els cotis de YouTube!! Pau, gràcies per compartir el meu petit món friqui al Kaburi i Glups, per les escapades al MMVV, i al mercat medieval! (sempre amb el permís del Gerard). Ludo, por nuestras charlas y encuentros inesperados que me gustaria que fueran más a menudo, (difícil por estar en países distintos). Alex, què dir-te que no sàpigues, vas viure l'inici de tot aquest viatge quan encara no havia ni començat, moltes gràcies per estar i confiar amb mi!

Al bàsquet en general per ser la meva via d'escapament a tot això i més concretament a les floretes del St Ramon Nonat (Júlia, Go, Morera, Pursa, Adri, Alba, Elia, Nayra, Estela, i el més floreta de tots, el Dani), al DrinkTeam (Albes, Laura, Estefania, Núria, Gina, Elena, Mercè, Marina) pels partits, els bons moments, i sobretot les braves!! A tots els equips on he estat per aquí a Barna: St Ramon Nonat, St. Joan de Mata, Drink Team i Ramon Llull!!

Finalment, a la meva família, gràcies pel suport encara que de vegades no heu entès molt bé lo que estava fent o que representava fer un tesi. Mama, qui t'hauria dit que la teva filla que no volia estudiar batxillerat acabés fent una tesi? (sí, ni jo m'ho crec). Papa, a continuar fent hort i cuidar les gallines, que ningú té unes hortalisses i uns ous tan bons com els teus! Algun dia els albercocs sortiran. Aileen, tata, sé que pot sonar una mica egoista però m'agrada tenir-te per aquestes terres; moltes gràcies per tot i per les trucades nocturnes de tornada a casa. Richi gracias por la portada! (aún me debes una clase de *skate*).

"*Computers are useless. They can only give you answers*"
— Pablo Picasso

Índex

Índex III

Llista de Figures VI

Llista de Taules VII

Abreviatures IX

Introducció 1

1 Introducció 3
- 1.1 Els virus com a contaminants d'aigua i aliments . . . 3
- 1.2 Mètodes analítics per virus . . . 4
- 1.3 Taxonomia dels virus . . . 5
- 1.4 Principals virus patògens transmesos a través d'aigua i aliments . 9
- 1.5 Metagenòmica de virus . . . 10
- 1.6 Caracterització de la composició dels ecosistemes virals . . . 11
  - 1.6.1 Metagenòmica dirigida envers seqüenciació de genomes complets . . . 13
- 1.7 Seqüenciació d'àcids nucleics . . . 14
  - 1.7.1 Cronologia de la seqüenciació de l'ADN . . . 14
  - 1.7.2 Tècniques de seqüenciació . . . 15
- 1.8 Ensambladors: mètodes i programes . . . 27
  - 1.8.1 Història . . . 27
  - 1.8.2 Mètodes . . . 28

Objectius 35

2 Objectius 37

Material i Mètodes 39

3 Origen de les dades 41
- 3.1 Projectes on s'han obtingut les mostres . . . 41
- 3.2 Protocols de tractament de mostres ambientals i d'aliments . . . 42
  - 3.2.1 Protocols per la recuperació i concentració de virus . . . 42

3.2.2 Preparació de les llibreries i seqüenciació 44

4 Desenvolupament del protocol Bioinformàtic per genomes virals 45
4.1 Anàlisi pre-ensamblat 47
4.1.1 Qualitat de les seqüències 47
4.1.2 Complexitat de les seqüències 50
4.1.3 Redundància de les seqüències 52
4.2 Ensamblat de Metagenomes 54
4.3 Anàlisi post-ensamblat 57
4.3.1 Riquesa del metaviroma (*Richness*) 60
4.3.2 Anàlisis filogenètics 61
4.4 Metagenòmica dirigida 64

Resultats 65

5 Resultats 67
5.1 Anàlisi pre-ensamblat 67
5.2 Ensamblat de Metagenomes 75
5.3 Anàlisi post-ensamblat 82
5.4 Anàlisis filogenètics per famílies específiques 91

6 Visualització i accessibilitat de les dades 97
6.1 Base de dades 97
6.2 Visualització de dades 98
6.2.1 Estadístiques de l'ensamblat 99
6.2.2 Gràfics "Krona" 106
6.2.3 Taules dinàmiques 107

Discussió 111

7 Discussió 113
7.1 Metagenòmica sobre genomes complets (WGS) 114
7.2 Dissenys experimentals 115
7.3 Protocols de neteja de seqüències pre-ensamblat 116
7.4 Programes d'ensamblat 117
7.5 Anotació taxonòmica basada en homologia 118
7.6 Anàlisis filogenètics i caracterització de nous virus 118
7.7 Millores futures en els protocols 119

Conclusions 121

8 Conclusions 123

Bibliografia 125

Annexos 145

Article 1 147
A metagenomic assessment of viral contamination on fresh parsley plants irrigated with fecally tainted river water 147

Article 2 165
Identification of sapovirus GV.2, astrovirus VA3 and novel anelloviruses in serum from patients with acute hepatitis of unknown etiology 165

Article 3 197
Metagenomics for the study of viruses in urban sewage as a tool in public health surveillance 197

Altres publicacions 1 241
Evidence of viral dissemination and seasonality in a Mediterranean river catchment: Implications for water pollution management 241

Altres publicacions 2 253
Health Risks Derived from Consumption of Lettuces Irrigated with Tertiary Effluent Containing Norovirus 253

Altres publicacions 3 263
Evaluation of methods for the concentration and extraction of viruses from sewage in the context of metagenomic sequencing 263

Notes 283

Índex de figures

1.1 Esquema de les cadenes i mètode de replicació emprats en la classificació de Baltimore . . . 6
1.2 Grans trets de la classificació taxonòmica de Baltimore . . . 8
1.3 Mètode de terminació en cadena (Sanger) . . . 17
1.4 Mètode de piroseqüenciació (454) . . . 17
1.5 Mètode SOLiD . . . 19
1.6 Mètode Illumina GA/HiSeq/MiSeq/NextSeq . . . 19
1.7 Mètode Ion PGM . . . 22
1.8 Mètode Pac-Bio . . . 22
1.9 Mètode Nanopore . . . 24
1.10 Comparació del rendiment de les diferents tecnologies de seqüenciació . . . 26
1.11 Mètode de grafs de solapament . . . 29
1.12 Mètode de grafs *de Bruijn* . . . 30
1.13 Mètode de grafs *Greedy* . . . 32

4.1 Diagrama general de flux del protocol desenvolupat . . . 46
4.2 Estudi de la qualitat de les seqüències: composició nucleotídica i qualitat *Phred* . . . 48
4.3 Calculant la Complexitat Lingüística . . . 51
4.4 *Scatterplot* de la complexitat dels *reads* d'un experiment de seqüenciació . . . 53
4.5 Esquema del funcionament del ensamblador `Velvet` . . . 56
4.6 Esquema de les optimitzacions de `MetaVelvet` . . . 57
4.7 *Boxplots* comparatius dels diferents mètodes paramètrics per calcular la riquesa amb el programa `CatchAll` . . . 62
4.8 *Boxplots* comparatius dels diferents mètodes no paramètrics per calcular la riquesa amb el programa `CatchAll` . . . 63

5.1 Corbes de complexitat per les mostres d'aigua residual. . . . 72
5.2 Corbes de complexitat per les mostres de julivert i aigua de riu. . 73
5.3 Corbes de complexitat per les mostres d'aigua residual. . . . 74
5.4 Distribució dels *contigs* de dos ensamblats en les mostres d'aigua residual. . . . 79
5.5 Distribució dels *contigs* de dos ensamblats en les mostres de julivert i aigua de riu. . . . 80
5.6 Distribució dels *contigs* de dos ensamblats en les mostres de sèrum humà. . . . 81

5.7 *Heatmap* de les mostres d'aigua residual. 88
5.8 *Heatmap* de les mostres de julivert i agua de riu 89
5.9 *Heatmap* de les mostres de sèrum humà 90
5.10 Arbre filogenètic de la família *Adenoviridae* a partir de dades de metagenòmica dirigida 92
5.11 Arbre filogenètic de la família *Anelloviridae* 93
5.12 Arbre filogenètic de les seqüències relacionades amb Hepelivirus . 94
5.13 Alineament de la regió conservada per RdRp de les seqüències relacionades amb hepelivirus 95

6.1 Esquema UML de les taules MySQL del nucli de la nostra base de dades 98
6.2 Taula principal d'entrada per accedir a les taules dinàmiques . . 100
6.3 Taula sumari de les mostres d'un *run* 100
6.4 Taula informativa sobre els filtrats de les seqüències *raw* 101
6.5 Taula resum amb informació sobre complexitat de les seqüències 102
6.6 Taula descriptiva dels paràmetres de seqüència i de complexitat avaluats 102
6.7 Gràfic de les distribucions de densitats basades en les freqüències dels diferents paràmetres analitzats 103
6.8 Gràfiques de la distribució dels diferents paràmetres analitzats respecte la longitud de seqüència 104
6.9 Comparació dos a dos dels tres paràmetres de complexitat analitzats. 105
6.10 Estadístiques de les seqüències filtrades 106
6.11 Estadístiques de les seqüències que s'han pogut assignar a un grup taxonòmic (a partir de les cerques amb el BLAST) 106
6.12 Plots *Krona* per navegar pels grups taxonòmics 107
6.13 Taula dinàmica colapsada d'una mostra en concret 109
6.14 Taula dinàmica desplegada d'una mostra en concret 109
6.15 Taula dinàmica de seqüències per una família en concret 110

Índex de taules

1.1 Principals famílies de virus contaminants que poden causar malalties en humans . . . 10
1.2 Resum de les característiques dels diferents seqüenciadors 25
1.3 Resum de les característiques dels diferents ensambladors 33

4.1 Taula amb els valors de qualitat *Phred* . . . 49
4.2 Bases de dades emprades per l'anotació taxonòmica . . . 59

5.1 Resum de les anàlisis pre-ensamblat de les mostres d'aigua residual 69
5.2 Resum de les anàlisis pre-ensamblat de les mostres de julivert i riu 70
5.3 Resum de les anàlisis pre-ensamblat de les mostres de sèrum humà 71
5.4 Resum de les anàlisis sobre ensamblat de les mostres d'aigua residual . . . 76
5.5 Resum de les anàlisis sobre ensamblat de les mostres de julivert i riu . . . 77
5.6 Resum de les anàlisis sobre ensamblat de les mostres de sèrum humà . . . 78
5.7 Resum dels resultats de diversitat calculats sobre l'ensamblat per les mostres d'aigua residual . . . 85
5.8 Resum dels resultats de diversitat calculats sobre l'ensamblat per les mostres de julivert i riu . . . 86
5.9 Resum dels resultats de diversitat calculats sobre l'ensamblat per les mostres de sèrum humà . . . 87
5.10 Codificació de les seqüències anotades en l'arbre filogenètic dels hepelivirus . . . 96

Abreviatures

AdV	*Adenovirus*
ADN	Àcid DesoxiriboNucleic
ARN	Àcid RiboNucleic
ARNm	Àcid RiboNucleic missatger
CL	Complexitat Lingüística
DBG	Grafs de *Bruijn*
EDAR	Estació Depuradora d'Agües Residuals
GO	Ontologia funcional desl gens (*Gene Ontology*)
GPL	Llicència Pública General GNU (*GNU General Public License*)
HSP	Segment alineat de millor puntuació (*High-scoring Segment Pair*)
ICTV	Comitè Internacional de Taxonomia de Virus (*International Committee on Taxonomy of Viruses*)
NGS	Noves tecnologies de seqüenciació (*Next Generation Sequencing*)
OLC	Grafs de solapament (*Overlap-Layout-Consensus*)
OTU	Unitat Taxonòmica Operativa (*Operational Taxonomic Unit*)
PCR	Reacció en cadena de la polimerasa (*Polymerase Chain Reaction*)
PE	Traces de seqüenciació aparellades (*Pair-end reads*)
SE	Traces de seqüenciació desaparellades (*Single-end reads*)
SMF	Floculació amb llet descremada (*skimmed milk floculation*)
WG	Genoma complet (*Whole Genome*)
WGS	Seqüenciació de Genomes Complets (*Whole Genome Sequencing*/*"Shotgun" sequencing*)
WHO	Organització mundial de la salut (*World Health Organisation*)
ZMW	Múltiples guies d'ona en mode 0 (*Zero-Mode Waveguides*)
ddNTP	Didesoxinucleòtid
dNTP	Desoxirribonucleòtid

dsDNA	ADN bicatenari (*Double-strand DNA*)
dsDNA-RT	ADN bicatenari retrotranscrit
dsRNA	ARN bicatenari (*Double-strand RNA*)
ssDNA	ADN monocatenari (*Single-strand DNA*)
ssRNA+	ARN monocatenari positiu (*Single-strand RNA*)
ssRNA-RT	ARN monocatenari retrotranscrit
ssRNA-	ARN monocatenari negatiu

Unitats

bp	Parells de bases (*Base Pairs*)
CG	Còpies genòmiques
GB	10^9 octets (*Giga bytes*)
kbp	Kilobases (10^6 *Base Pairs*)
MB	10^6 octets (*Mega bytes*)
Mbp	Megabases (10^9 *Base Pairs*)
pH	Potencial hidrogeniònic
TB	10^{12} octets (*Tera bytes*)

Introducció

1 Introducció

1.1 Els virus com a contaminants d'aigua i aliments

Segons l'Organització Mundial de la Salut (OMS), les malalties d'origen alimentari són un problema de salut pública en expansió. Actualment l'aigua es considera un recurs limitat, afectat per les variacions extremes en les condicions climàtiques, fent necessari en algunes regions realitzar polítiques de canvi en el seu ús que poden haver generat noves vies de transmissió de patògens vírics fins ara no presents (World Health Organization [WHO], 2016). Un d'aquests canvis és la utilització d'aigües regenerades com a font d'aigua i nutrients que poden reaprofitar-se en les indústries, recàrrega de aqüífers, rec de jardins i rec de productes frescos, com per exemple les hortalisses.

La població humana i els animals excreten una gran diversitat de virus patògens a les seves orines i femtes; com a resultat, l'aigua residual que es genera representa un dels principals vehicles per a la disseminació de patògens a les aigües superficials, subterrànies o costaneres, i com a conseqüència en aliments (Carter, 2005). La contaminació del medi ambient suposa un risc greu de salut pública. S'ha estimat que en el conjunt del planeta, aproximadament 3 000 milions de persones no disposen d'aigua potable i que el 95% de l'aigua residual domèstica és abocada al medi ambient sense tractar (Langford, 2005). Tot i les mesures destinades a millorar la qualitat de l'aigua i la seguretat alimentària, a nivell mundial s'identifiquen de manera recurrent brots causats per virus transmesos per aigua o aliments (Carvalho *et al.*, 2012; Fournet *et al.*, 2012; Koroglu *et al.*, 2011; Nenonen *et al.*, 2012). També a escala global, les malalties transmeses per l'aigua o aliments constitueixen un problema que en els darrers 20 anys, lluny de disminuir per la millora de les condicions higièniques de molts països, s'ha complicat considerablement (Newell *et al.*, 2010). Entre els principals factors responsables d'aquesta tendència podem destacar els següents: creixement i envelliment de la població humana; comerç internacional de vegetals, carns, aliments exòtics i animals de granja entre països amb estàndards microbiològics diferents; canvis en certes pràctiques agrícoles per abaratir costos; i com a rerefons el canvi climàtic.

Les fonts de contaminació de virus per transmissió ambiental són molt diverses i poden classificar-se segons si el seu origen és puntual o difús. Les fonts de contaminació puntuals són habitualment més fàcils d'identificar i s'originen per l'excreció directa dels patògens al medi o per les descàrregues d'aigua residual

tractada o sense tractar en localitzacions molt concretes. D'altra banda, les fonts de contaminació difuses s'originen principalment durant episodis de precipitacions que poden rentar el sòl urbà i terres de cultiu. S'ha demostrat que la població pot contraure infeccions per vies molt diverses, pel consum o contacte amb aigua contaminada, consum de mol·luscs bivalves crus o poc cuinats, que hagin estat cultivats en aigües contaminades (Woods *et al.*, 2016), així com pel consum de vegetals irrigats amb aigües contaminades o abonats amb biosòlids (Müller *et al.*, 2016). El biosòlid, o fang de depuració, és el residu produït en la depuració d'aigua residual, on es concentra gran part de la càrrega contaminant separada al llarg d'aquest procés.

1.2 Mètodes analítics per virus

Existeixen diverses aproximacions que tradicionalment s'han emprat per la detecció de virus en mostres ambientals concentrades; des de l'observació de partícules víriques mitjançant microscòpia electrònica fins a la detecció de l'efecte citopàtic en línies cel·lulars o els mètodes moleculars.

La microscòpia electrònica presenta una sensibilitat molt limitada, ja que requereix matrius molt netes i concentracions molt elevades per poder visualitzarles (Atmar i Estes, 2001). Un altre inconvenient és que té un cost elevat i requereix molt esforç. Com a conseqüència, és una tècnica que està limitada a estudis de caracterització d'estructura de virus cultivables.

Un altre mètode és la utilització de les tècniques de cultiu cel·lular, les quals tenen l'avantatge de poder mesurar quantitativament la presència d'espècies víriques infeccioses. Aquesta tècnica s'aplica principalment en estudis d'estabilitat i desinfecció (Wyn-Jones *et al.*, 2011). Dins d'aquesta, es poden diferenciar els assajos d'unitats formadores de clapa (PFU, *Plaque Forming Units*) i el de dosi infecciosa en cultiu de teixits 50% (TCID50, *Tissue Culture Infectious Dose* 50%). Aquests dos mètodes requereixen però de l'existència d'una línia cel·lular susceptible de ser infectada pel virus que s'estigui analitzant (Hamza *et al.*, 2011).

Una de la tècniques més emprades en l'actualitat és la reacció en cadena de la polimerasa (PCR, *Polymerase Chain Reaction*). En aquesta metodologia es pot diferenciar entre la PCR simple (PCR/RT-PCR), la PCR niada o *Nested-PCR* (nPCR/nRT-PCR) i la PCR quantitativa (qPCR/ qRT-PCR; Girones *et al.*, 2010). Aquesta tècnica requereix una extracció d'àcids nucleics prèvia a l'amplificació de les seqüències genòmiques. La PCR és una de les tècniques més utilitzades de manera rutinària als laboratoris de virologia i permet quantificar

el nombre de còpies genòmiques (CG) presents a una mostra de manera ràpida i poc costosa a nivell econòmic, però no dóna dades sobre la infectivitat (Heid *et al.*, 1996). Per obtenir informació sobre la infectivitat es poden fer assajos combinats amb cultius cel·lulars, els quals ens permeten estimar la capacitat infectiva del virus estudiat. Són les anomenades tècniques integrades de PCR-cultiu cel·lular (ICC-PCR, *Integrated cell culture-PCR*; Dong *et al.*, 2010), que analitzen la presència d'àcids nucleics vírics en els cultius infectats amb la mostra problema, incrementant la sensibilitat en la detecció.

1.3 Taxonomia dels virus

Els orígens de la taxonomia es remunten als inicis del llenguatge, quan la gent anomenava organismes similars amb el que hom coneix com "noms vulgars". El principal objectiu de la taxonomia és el de classificar en sentit literal, i en el cas de la biologia poder classificar els organismes. En el segle XVIII, en Carl Linneo va establir les normes bàsiques per definir la nomenclatura de les espècies en base a la semblança morfològica dels organismes (Stearn, 1959), el que avui en dia formaria part del que denominem fenotip i que inclou caràcters no tan sols morfològics sinó també moleculars. Aquest sistema va permetre tipificar i classificar més de 8 000 espècies d'animals i 6 000 de plantes. Els criteris més importants en el moment de classificar noves espècies de virus encara estan basats en els caràcters fenotípics, però també es té en compte, a diferència de la resta d'organismes, el tipus morfològic, el tipus d'àcid nucleic del seu genoma i el mode de replicació, les espècies d'hoste a les que parasiten i el tipus de malaltia que poden arribar a provocar. Actualment hi ha dos grans aproximacions per classificar els virus: el sistema de Baltimore i la classificació del Comitè Internacional de Taxonomia de Virus (ICTV)[1]. En la classificació de Baltimore, que es va definir a l'any 1971, els virus s'agrupen en 7 grups, en funció del tipus d'àcid nucleic, el nombre de cadenes, el sentit i el mètode de replicació (Baltimore, 1971). A la Figura 1.1 de la pàgina 6 es descriuen de manera general cadascun dels grups de Baltimore, amb un petit esquema de l'estructura de les seves cadenes i el seu mètode de replicació, on es pot veure la relació entre el genoma viral i els ARN missatgers que seran traduïts a proteïnes víriques. La Figura 1.2, de la pàgina 8, resumeix aquests 7 grups de Baltimore, amb les principals famílies que contenen i quin és el seu hoste.

- **Grup I: virus d'ADN bicatenari (dsDNA)**
  L'ARNm es transcriu directament a partir del genoma del virus, el qual és de doble cadena d'ADN; posteriorment les proteïnes reguladores i es-

[1] `https://talk.ictvonline.org/`

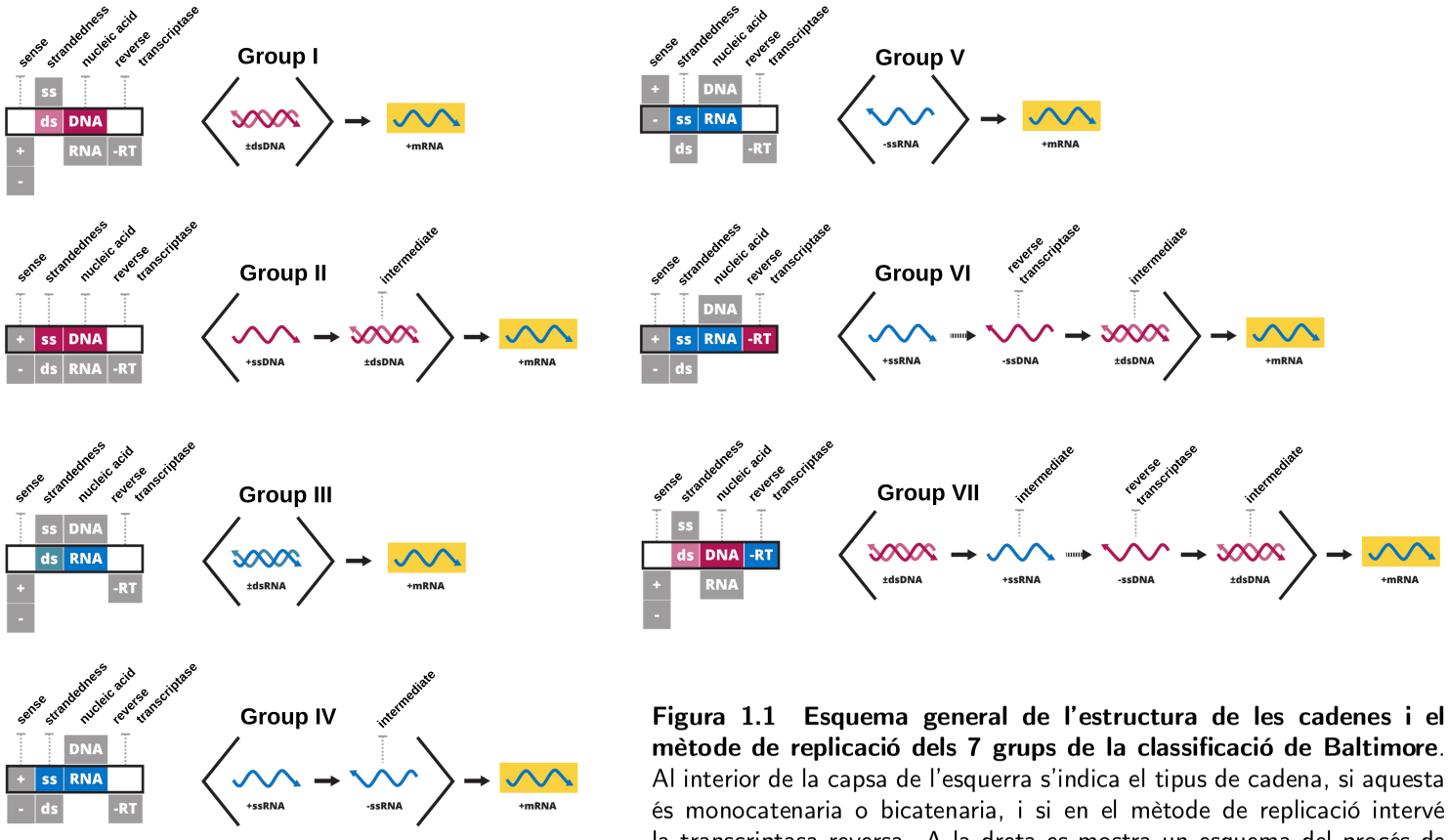


Figura 1.1 Esquema general de l'estructura de les cadenes i el mètode de replicació dels 7 grups de la classificació de Baltimore. Al interior de la capsa de l'esquerra s'indica el tipus de cadena, si aquesta és monocatenaria o bicatenaria, i si en el mètode de replicació intervé la transcriptasa reversa. A la dreta es mostra un esquema del procés de replicació. Adaptat a partir d'una Figura de Confalonieri, (2014), de la `Wikimedia Commons`, amb llicència Creative Commons (CC BY-SA 3.0).

tructurals (per exemple, de la càpside) es tradueixen a partir d'aquest ARNm.

- **Grup II: virus d'ADN monocatenari (ssDNA)**
  L'ADN monocatenari es fa servir de motlle per recuperar l'ADN bicatenari utilitzant enzims del seu hoste; a partir d'aquest, l'ARNm es transcriu igual que en el grup I. La majoria presenten genomes circulars.

- **Grup III: virus d'ARN bicatenari (dsRNA)**
  A partir de l'ARN bicatenari s'obté l'ARN monocatenari que actua directament com a ARNm, el qual de nou és traduït per obtenir les proteïnes reguladores i estructurals.

- **Grup IV: virus ARN monocatenari positiu (ssRNA+)**
  A partir de la traducció de l'ARN monocatenari s'obtenen les proteïnes reguladores; en el cas del grup **IVa** també les proteïnes estructurals, amb un ARNm policistrònic que es tradueix a una poliproteïna. En el cas del grup **IVb**, els virus tenen processos de transcripció complexos i es generen ARNm subgenòmics.

- **Grup V: virus d'ARN monocatenari negatiu (ssRNA-)**
  A partir de l'ARN monocatenari negatiu, mitjançant una transcriptasa viral es produeixen els diferents ARNm, els quals son traduïts en les proteïnes virals.

- **Grup VI: virus d'ARN monocatenari retrotranscrit (ssRNA-RT)**
  Mitjançant una transcriptasa inversa, a partir de l'ARN viral de sentit positiu, es genera primer una cadena d'ADN monocatenari i després bicatenari. Aquest ADN bicatenari s'integra en el genoma de l'hoste i un cop allí es transcriu a ARN, que a la vegada es traduirà en proteïnes virals.

- **Grup VII: virus d'ADN bicatenari retrotranscrit (dsDNA-RT)**
  L'ADN viral de doble cadena és completat a dins de la cèl·lula. L'ARN transcrit és l'intermediari replicatiu per la síntesi de nous genomes d'ADN mitjançant una retrotranscriptasa viral.

El Comitè Internacional de Taxonomia de Virus (ICTV) ha desenvolupat un esquema universal per la classificació taxonòmica viral amb l'objectiu de descriure tots els virus però que funcioni de manera estàndard, regulant la descripció formal de noves espècies i definint la seva ubicació dins de l'esquema classificatori. Aquest comitè ha intentat promoure que les regles de nomenclatura s'assemblin el més possible als estàndards tradicionals de la classificació de la resta d'organismes. Una de les principals diferències en aquest sistema de nomenclatura és que els noms dels ordres i famílies estan en format itàlic (van Regenmortel

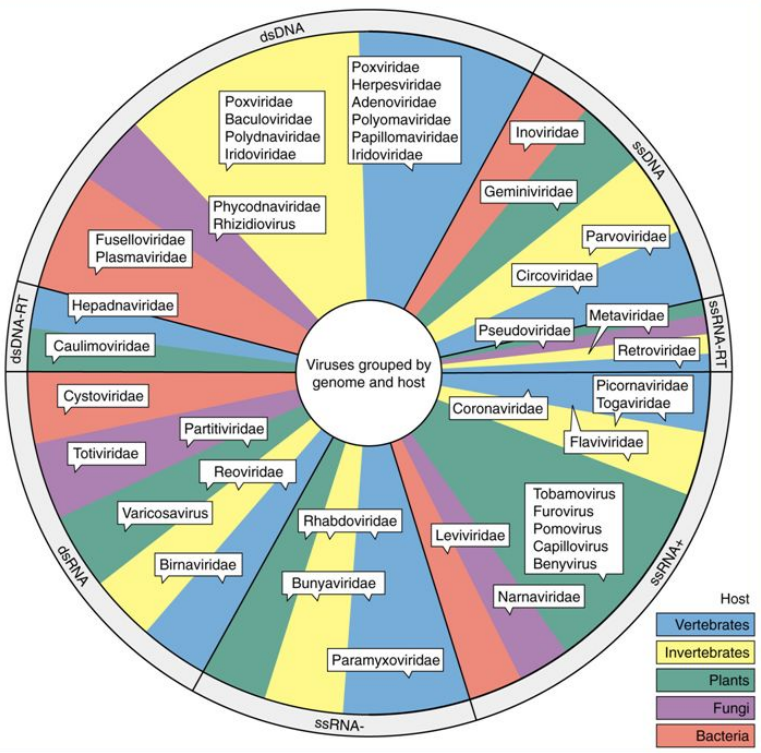


Figura 1.2 Grans trets de la classificació taxonòmica en base al sistema Baltimore. L'anell exterior defineix els set grans grups en funció del seu genoma. Els requadres contenen diferents famílies de virus, agrupats a partir de l'hoste que infecten. Els colors assignats als diferents subsectors corresponen als 5 grans grups que comprenen la resta de tots els organismes no virals. Adaptat de la presentació *Viral Taxonomy* de Kumar Das, (2014).

i Mahy, 2004). Per assolir aquest objectiu s'han definit sufixes que indiquen el rang taxonòmic, o sigui el nivell taxonòmic equivalent a ordre, família, gènere i espècie. A continuació s'enumeren els sufixes pels diferents nivells definits en la taxonomia de virus: "*-virales*" per definir "ordre", "*-viridae*" per "família", "*-virinae*" per "subfamília" i "*-virus*" per anotar "gènere". Finalment, el nom de les espècies sovint s'assigna en funció de la malaltia que causa el virus. S'han publicat diversos articles (Morgan, 2016; van Regenmortel i Mahy, 2004) on es suggereixen canvis en aquesta nomenclatura, ja que l'actual presenta problemes, com per exemple que en més de 50 grups s'engloben més de 1 500 espècies diferents, o que hi ha virus que es podrien classificar en més d'una espècie i altres en cap.

1.4 Principals virus patògens transmesos a través d'aigua i aliments

Els virus patògens que es transmeten per l'ús o consum d'aigua i aliments contaminats són molt diversos. Els més freqüents solen causar gastroenteritis, però n'hi ha d'altres que originen la malaltia un cop han migrat a altres òrgans, com per exemple, el fetge en el cas de l'hepatitis o el sistema nerviós central en el cas dels enterovirus. Tot i que generalment es caracteritzen per provocar infeccions asimptomàtiques o de pronòstic lleu, en alguns casos i en general en grups de risc, com nens, ancians, dones embarassades o pacients immuno-suprimits, en els quals es poden produir complicacions més greus i, ocasionalment, fins i tot la mort (Koopmans i Duizer, 2004).

La majoria de virus humans amb transmissió possible per aigua i aliments pertanyen a les famílies *Adenoviridae*, *Astroviridae*, *Caliciviridae*, *Hepeviridae*, *Picornaviridae*, *Reoviridae* i *Polyomaviridae* (veure resum a la Taula 1.1). Molts dels virus considerats com a contaminants ambientals són responsables d'infeccions subclíniques i només ocasionen símptomes clínics en una petita part de la població afectada. Alguns virus com l'Adenovirus humà i el Poliomavirus produeixen infeccions durant la infantesa que poden esdevenir infeccions persistents (Pavia, 2011). L'exposició, fins i tot en dosis molt baixes, com succeeix amb els Norovirus, pot causar la infecció i la malaltia (Teunis *et al.*, 2008). En tots els casos, l'evolució de la infecció depèn de factors com la via de transmissió, la dosi infecciosa del patogen, l'edat, l'estat immunològic del pacient i la possibilitat d'accedir a atenció sanitària de qualitat.

Família	Gènere	Propietats	Malalties Associades
Adenoviridae	Mastadenovirus (Adenovirus humà)	90-100nm dsDNA	Conjuntivitis, gastroenteritis, malalties respiratòries.
Astroviridae	Mamstrovirus	28-30nm ssRNA	Gastroenteritis.
Caliciviridae	Norovirus Sapovirus	27-38nm ssRNA	Gastroenteritis. Gastroenteritis.
Hepeviridae	Orthohepevirus	25-30nm ssRNA	Hepatitis aguda.
Picornaviridae	Enterovirus Hepatovirus (VHA) Kobuvirus (Aichivirus) Salivirus	24-30nm ssRNA	Gastroenteritis, encefalitis, meningitis, conjuntivitis. Hepatitis aguda. Gastroenteritis. Gastroenteritis.
Reoviridae	Rotavirus	70-75nm dsRNA	Gastroenteritis.
Polyomaviridae	Alphapolyomavirus Betapolyomavirus Deltapolyomavirus	50-60nm dsDNA	Nefropaties (BKPyV), leucoencepalopatia multifocal progressiva(JCPyV), càncer de pell (MCPyV).

Taula 1.1 Principals famílies de virus contaminants que poden causar malalties en humans. Les dues primeres columnes ens mostren la família i els gèneres dins de la mateixa per als que s'han descrit espècies patogèniques en humans. En les propietats es resumeixen el rang de diàmetres de les partícules víriques dins de cada grup, així com el tipus de genoma (classificació Baltimore). La darrera columna enumera algunes de les malalties associades al virus. Adaptat de Guerrero-Latorre, (2016).

1.5 Metagenòmica de virus

S'estima que més del 99% dels organismes presents a la natura no són cultivables utilitzant tècniques estàndards com els cultius cel·lulars (Hugenholtz *et al.*, 1998; Streit i Schmitz, 2004). Tradicionalment, la seqüenciació genòmica s'ha basat en la clonació específica d'un gen (en bacteris normalment la regió del gen 16S de l'ARNr) mitjançant mètodes com la PCR, s'obtenen mostres no modificades d'un gen marcador en tots els membres de la comunitat de la mostra en estudi, i així determinar un perfil específic de la biodiversitat d'una mostra ambiental o clínica (Methé *et al.*, 2012; Oude Munnink *et al.*, 2013; Wang *et al.*, 2016). El problema en aquesta aproximació és que s'estudia un gen en concret o un gen

de les espècies d'una família o gènere, però si es vol tenir una visió més general i àmplia del microbioma present en una mostra ambiental, aquesta no arriba a generar suficient informació.

Podem definir la metagenòmica com l'estudi del material genètic que ha estat recuperat directament de les mostres que es volen analitzar. La primera vegada que es va parlar de metagenòmica va ser al 1998, quan Jo Handlesman, Jon Clardly i Robert M. Goodman van utilitzar el terme "metagenòmica" com a referència a un abordatge que pretenia analitzar una col·lecció de gens seqüenciats d'una mostra ambiental com si es tractés d'un únic genoma (Handelsman *et al.*, 1998). Més recentment, Kevin Chen i Lior Pachter (investigadors de la Universitat Californiana de Berkeley) van definir la metagenòmica com l'aplicació de tècniques genòmiques modernes per l'estudi directe de comunitats de microorganismes al seu entorn natural, evitant la necessitat d'aïllar i cultivar cadascuna de les espècies que composen la comunitat (Chen i Pachter, 2005).

1.6 Caracterització de la composició dels ecosistemes virals

En estudis de metagenòmica de genomes complets amb mostres d'origen ambiental, com per exemple aigua, aliments i ocasionalment també en mostres clíniques, la concentració del material genètic víric és molt baixa i s'ha d'arribar a la concentració mínima necessària per poder aplicar tècniques de seqüenciació. Per tant s'ha d'amplificar el material genètic de les mostres. Donat que els virus no tenen un marcador com el 16S bacterià ni el 18S eucariòtic o similar, aquesta amplificació es realitza mitjançant una PCR amb encebadors aleatoris o *random primers* (Wang *et al.*, 2002; Wang *et al.*, 2003). També s'utilitzen mètodes basats en l'amplificació de desplaçament múltiple (*multiple displacement amplification*, MDA) per arribar a aquesta concentració mínima (Angly *et al.*, 2006; Dinsdale *et al.*, 2008).

Com a resultat s'augmentarà qualsevol tipus de material genètic present a la mostra i per tant, és necessari que com a pas previ a l'amplificació, es realitzin tractaments per tal de reduir la presència d'ADN/ARN no víric; com per exemple, mètodes de concentració i extracció que beneficiïn les partícules víriques o tractaments amb nucleases per eliminar ADN/ARN lliure (majoritàriament d'origen bacterià i eucariòtic) i filtres per poder aïllar partícules víriques.

Els estudis de metagenòmica dirigida en virus es centren en la seqüenciació d'alguna família, grup de virus o espècie en concret, en el pas d'amplificació amb PCR utilitzen regions conservades d'aquests com a encebador (Hall *et al.*, 2014).

Aquests estudis permeten tenir una caracterització molt exhaustiva de l'espècie o família en la mostra i fer anàlisis filogenètics. Amb aquestes aproximacions s'han pogut detectar noves espècies dins de famílies ja conegudes (Ng *et al.*, 2012) i incrementar el nombre de seqüències virals a les bases de dades generals (Batovska *et al.*, 2017; Ogorzaly *et al.*, 2015).

A mesura que s'han anat realitzant més estudis de metagenòmica, ha sorgit la necessitat de poder consultar i comparar els resultats de tots aquest estudis; per aquest motiu s'han desenvolupat diverses eines. La major part d'estudis es realitzen en bacteris i caracteritzant la regió 16S, per tant la majoria de les bases de dades que s'han creat estan centrades a metagenòmica dirigida de bacteris. Un exemple d'aquestes eines és EBI Metagenomics[2] (Mitchell *et al.*, 2016), la qual proporciona un servei que a part d'aplicar un *pipeline* automatitzat per analitzar dades de metagenòmica de 16S rRNAs i emmagatzemar-los, permet a l'usuari seleccionar diversos projectes i fer comparacions en base a ontologies gèniques (GO). L'aplicació MG-RAST[3] (Meyer *et al.*, 2008), realitza anàlisis i emmagatzema dades de metagenòmica dirigida de 16S i 18S. A més a més, facilita les comparacions entre mostres en base diversos criteris, com per exemple composició, funcionalitat i tipus de mostra, utilitzant *barcharts*, taules, Principal Coordinates Analysis (PCoA) i mapes de KEGG. Dins del projecte Bioconductor (Huber *et al.*, 2015), on l'objectiu és analitzar i entendre dades de *high-throughput* en genòmica i biologia molecular, s'han desenvolupat paquets en R (R Core Team, 2016) com *PathoStat* (Manimaran *et al.*, 2016) per a la visualització d'abundàncies relatives, estimacions de diversitat i anàlisis d'OTU's per genòmica dirigida o també *metavizr* (Corrada Bravo *et al.*, 2017), una pàgina web interactiva per a la visualització d'anàlisis de metagenòmica.

De bases de dades específiques de virus n'hi ha moltes menys degut a que el nombre d'estudis en metagenòmica en virus i de genomes complets és molt menor a la metagenòmica dirigida en bacteris. Entre altres podem citar MetaVir[4], la qual fa anotacions de seqüències virals i arbres filogenètics. A més a més, hi ha un apartat de projectes públics amb els quals es poden realitzar comparacions entre mostres. També hi ha l'eina MGmapper[5] (Petersen *et al.*, 2017), que permet realitzar anotacions taxonòmiques de *reads* provinents de metagenòmica, indistintament de si són mostres de bacteris o virus.

[2] `https://www.ebi.ac.uk/metagenomics/`
[3] `http://metagenomics.anl.gov/`
[4] `http://metavir-meb.univ-bpclermont.fr/`
[5] `https://cge.cbs.dtu.dk/servic/`

1.6.1 Metagenòmica dirigida envers seqüenciació de genomes complets

Podem diferenciar dos grans tipus de metagenòmica: la metagenòmica que permet seqüenciar zones concretes del genoma i/o genomes específics (metagenòmica dirigida) i la metagenòmica basada en seqüenciació de genomes complets (*whole genome sequencing* o WGS). El primer tipus d'aproximació permet centrar-se en l'anàlisi d'àrees específiques d'interès i tenir una cobertura més gran durant la seqüenciació. Com a conseqüència, això permet fer estudis exhaustius en alguna família en concret o grup de virus, identificar variants rares, mutacions i noves espècies dins d'aquella família, per posteriorment realitzar un estudi filogenètic i taxonòmic (Brussaard *et al.*, 2004). En el cas dels bacteris normalment el gen que es vol seqüenciar és el 16S, ja que és un fragment curt, generalment molt conservat dins d'una espècie, fet que ajuda a la caracterització de la mostra. Aquest tipus de seqüenciació pot estar esbiaixada per l'amplificació diferencial del fragment "diana" de cadascun dels organismes; tot i que s'utilitzin *primers* universals, aquests poden unir-se amb més afinitat a soques concretes en funció de la complementarietat de bases. També poden existir seqüències flanquejants que interfereixin amb el lloc d'unió de l'encebador (*primer*), disminuint així l'eficàcia de l'amplificació (Hansen *et al.*, 1998). Això pot tenir com a conseqüència que la llibreria sigui poc representativa, especialment a nivell quantitatiu (Farrelly *et al.*, 1995).

La metagenòmica basada en la seqüenciació de genomes complets consisteix en identificar tots els organismes presents en una mostra; sense tenir cap *target* en concret. El principal avantatge d'aquesta metodologia és la gran quantitat d'informació que es pot extreure a partir de les seqüències obtingudes, tenint una representació més global del microbioma present a la mostra. No obstant, aquesta aproximació també presenta diverses limitacions com per exemple la profunditat de seqüenciació, ja que és complicat poder detectar aquelles espècies que estan menys representades en la mostra perquè seran "amagades" o desplaçades durant l'amplificació per les que estan més representades (Kalyuzhnaya *et al.*, 2008). A més, les abundàncies relatives que s'obtenen després de la seqüenciació poden resultar afectades en gran mesura pels protocols d'extracció i seqüenciació (Morgan *et al.*, 2010). Les seqüències obtingudes amb aquesta tècnica tenen una qualitat i una llargada inferiors que les obtingudes pels mètodes de metagenòmica dirigida, ja que no se sap *"a priori"* quines espècies estan presents a la mostra. També cal tenir en compte que durant l'anàlisi postseqüenciació té lloc el procés d'ensamblat, el qual si es fa de manera incorrecta pot complicar la identificació dels organismes d'una mostra. Finalment, si la longitud de les seqüencies obtingudes en la seqüenciació és massa curta, pot dificultar el pas de l'ensamblat i l'anotació funcional i/o taxonòmica (Oulas *et al.*, 2015).

1.7 Seqüenciació d'àcids nucleics

1.7.1 Cronologia de la seqüenciació de l'ADN

Abans de parlar sobre els ensambladors, cal fer una revisió sobre la història i els inicis de la seqüenciació de l'ADN. El 1977, Frederick Sanger va desenvolupar una tecnologia per seqüenciar l'ADN que es basava en el mètode de terminació de cadena, també conegut com seqüenciació *de Sanger* (Sanger i Coulson, 1975; Sanger *et al.*, 1977a). El mateix any, Walter Gilbert i Allan Maxam van desenvolupar una altra tècnica de seqüenciació basada en modificacions químiques de l'ADN i el posterior ancoratge a bases específiques, també coneguda com mètode de seqüenciació química (Maxam i Gilbert, 1977).

Amb el mètode de Sanger es va seqüenciar el primer genoma d'ADN, el del bacteriòfag X174 (PhiX; Sanger *et al.*, 1977b). Aquesta tècnica va ser la més utilitzada entre els anys 80 i mitjans dels 2000, i juntament amb el mètode de modificacions químiques van ser anomenades les tecnologies de la "primera generació", utilitzant-se tant en laboratoris com en aplicacions comercials (Anderson, 1981). Comparant els dos mètodes, el de Sanger era més eficient i utilitzava menys productes químics tòxics i radioactius que el de Maxam i Gilbert. Tot i així, la seqüenciació d'ADN era laboriosa i es necessitaven materials radioactius en ambdues tècniques. En anys posteriors es va seguir investigant, fins que Lloyd M. Smith amb Applied Biosystems, va substituir els materials radioactius pels nucleòtids marcats amb fluorocroms i com a resultat el 1986, va introduir la primera màquina automàtica de seqüenciació anomenada AB370 (Smith *et al.*, 1986); la qual utilitzava un mètode d'electroforesi capil·lar, el qual era més ràpid i més precís (Prober *et al.*, 1987). El model més recent és el AB3730xl, el qual pot arribar a generar 2.88Mbp per dia. La llargada pot arribar a fins a 900bp des de l'any 1995 (Dawson *et al.*, 2013).

El 1998, Shankar Balasubramanian i David Klenerman van fundar Solexa, i van crear el mètode de seqüenciació basat nucleòtids fluorescents i de terminació reversible (Bentley *et al.*, 2008; Cronn *et al.*, 2008). A l'any 2005, el mètode de piroseqüenciació va ser llençat per 454, el 2006 Solexa va fer public el Genome Analyzer i Agentcourt SOLiD (Sequencing by Oligo Ligation Detection; Valouev *et al.*, 2008). Totes aquestes tecnologies van ser adoptades com els sistemes de seqüenciació més típicament usades per seqüenciació en "next-generation sequencing" (NGS). Posteriorment aquestes companyies van anar sent absorbides; al 2006 Agencourt va ser comprada per Applied Biosystems; al 2007, 454 va ser comprada per Roche; i el mateix any Solexa per Illumina.

El 2009, Clifford Reid, Radoje Drmanac i John Curson van crear una companyia, Complete Genomics, en la que van desenvolupar una nova tècnica amb l'objectiu de seqüenciar el genoma complet d'un organisme. La principal diferència amb les altres companyies va ser que no oferien les eines per seqüenciar sinó que només oferien el servei. Al febrer de l'any 2010, la empresa Torrent Systems Inc. va llençar al mercat Ion Torrent, presentat com un seqüenciador ràpid, eficaç, compacte i econòmic que és podia utilitzar en un gran nombre de laboratoris com un aparell més de la poiata (Katsnelson, 2010); aquesta tècnica està basada en la detecció de protons alliberats durant el procés de polimerització de l'ADN; es diferencia dels altres sequenciadors perquè no utilitza nucleòtids modificats químicament, ja que per la detecció no s'utilitzen mètodes optics sinò detecció del pH. Juntament amb les màquines millorades desenvolupades per Illumina, aquestes tècniques formen part de la seqüènciació de tercera generació (Niedringhaus *et al.*, 2011). A partir de mitjans de la dècada del 2010, han irromput en el panorama genòmic les tècniques de seqüenciació de molècules individuals (*single-molecule sequencing*) com les que proporcionen les empreses PacBio als Estats Units o Oxford Nanopore a UK.

1.7.2 Tècniques de seqüenciació

A continuació es pretén incloure una descripció una mica més detallada de les diferents metodologies de seqüènciació disponibles.

Mètode terminació en cadena

El principi fonamental del mètode Sanger és l'ús de didesoxinucleids trifosfats (ddNTPs) com a terminadors de la síntesi de la cadena complementària de l'ADN. Es dissenya un oligonucleotid sintètic d'unes 17-20 bases, complementari al fragment d'ADN que es vol seqüenciar i situat aproximadament a 20 parells de bases de distància del fragment de seqüència que es vol llegir. El dúplex format entre l'oligo i l'ADN complementari de cadena senzilla es converteix en el substrat de l'ADN polimerasa I (DNApolI), que anirà estenent la cadena des del grup OH lliure de l'extrem 3' de l'oligo, incorporant dNTPs i copiant del motlle d'ADN per sintetitzar la cadena complementaria. Durant la seqüenciació es realitzen 4 reaccions de síntesi separades, incloent a cada una d'elles petites quantitats de dideoxidnucleids (ddNTP: ddGTP, ddATP, ddCTP, ddTTP) que no tenen l'extrem 3' OH lliure, els quals, al incorporar-se a la cadena d'ADN que s'està sintetitzant acaben amb l'allargament de la mateixa (Chidgeavadze *et al.*, 1984). La incorporació a l'atzar d'un ddNTP en competència amb el dNTP corresponent, implica la síntesi d'una barreja de cadenes de diferents longituds,

totes elles comencen al extrem 5' i acaben en totes les diferents posicions possibles on un ddNTP pot incorporar-se en lloc d'un dNTP. La longitud de les seqüències obtingudes dependrà de la relació entre dNTP/ddNTP que es posin en la reacció. Els nucleòtids marcats amb fluorocroms permeten la visualització de bandes de diferent longitud en un gel de poliacrilamida on cadascuna de les reaccions es carrega en un carril. Cada banda del gel representa les diferents llargades obtingudes, i la seqüencia es podrà determinar llegint cadascun dels quatre carrils (Sanger *et al.*, 1977a). En aquest tipus de seqüenciació es poden arribar a llegir fins 400/900 bases. Aquesta tècnica ha estat usada en diversos tipus d'anàlisi com per identificar mutacions en gens (Bu *et al.*, 2016), estudiar la diversitat microbiològica en biofilms (da Silva *et al.*, 2014) o en estudis filogenètics (Combosch *et al.*, 2017).

Mètode de piroseqüenciació (454)

El mètode de 454 (`http://www.454.com`) s'aprofita d'una reacció química coneguda com a piroseqüenciació, on es detecta el pirofosfat alliberat durant la incorporació nucleotídica mitjançant una reacció luminescent. Durant el procés de preparació de les llibreries per seqüenciació, fragments únics d'ADN s'enganxen a una única perla per a l'amplificació clonal dels fragments mitjançant una PCR en emulsió. Aquest procés va ser definit com "Un fragment, una perla, una lectura"; de l'anglès, *"One fragment, one bead, one read"*. D'aquesta manera, una vegada incorporat el nucleòtid corresponent per complementarietat de bases, s'allibera un pirofosfat que donarà lloc a una molècula d'ATP, que serà usada per un enzim que produeix llum de forma proporcional a la quantitat de pirofosfats alliberats. Inicialment aquesta plataforma podia obtenir seqüències entre 100-150bp i fins a 20MB per *run* (AllSeq, 2015). Les diferents evolucions del sistema han permès arribar a una longitud de lectura de 700 bases amb un rendiment total de 0,7 Gbases per procés de seqüenciació (AllSeq, 2015). Donada l'elevada longitud de les seves lectures, aquest mètode ha estat des dels seus inicis, un molt bon sistema per a la seqüenciació "*de novo*" (Pearson *et al.*, 2007). L'elevat preu per base seqüenciada i l'aposta de la companyia pels mètodes de seqüenciació de lectura única, han fet que aquest tipus de plataforma no tingui suport més enllà de 2016 (Karow, 2013). Aquesta tècnica s'ha utilitzat per exemple en la detecció de regions hipervariables (Kim *et al.*, 2015), en el mapatge de paràsits com *Plasmodium falciparum* (Samarakoon *et al.*, 2011) o en la detecció de mutacions en gens concrets (Moskalev *et al.*, 2013).

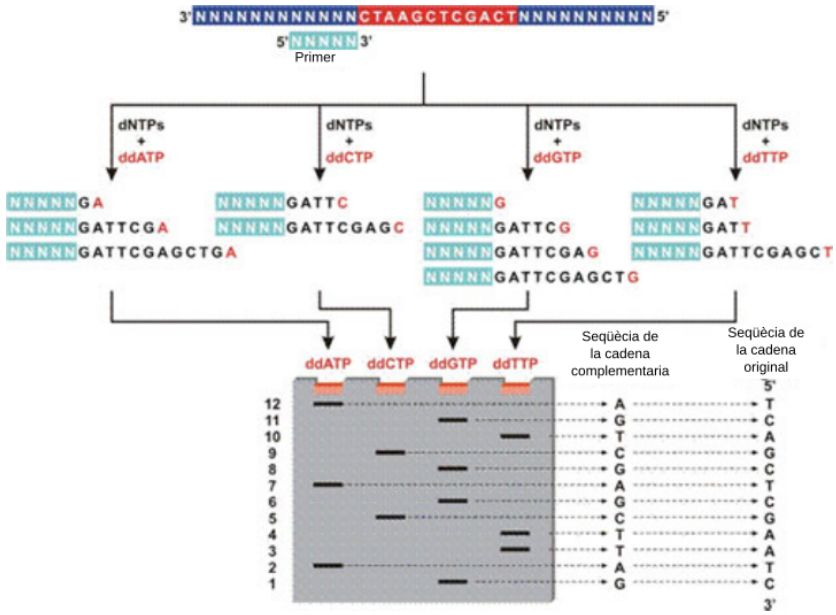


Figura 1.3 Esquema del mètode de terminació en cadena. S'utilitzen nucleòtids marcats amb fluorocroms i les mostres corren per un gel de poliacrilamida per poder seqüència la cadena d'ADN. Adaptat del portal acadèmic de la UNAM, Portal académico, (2015).

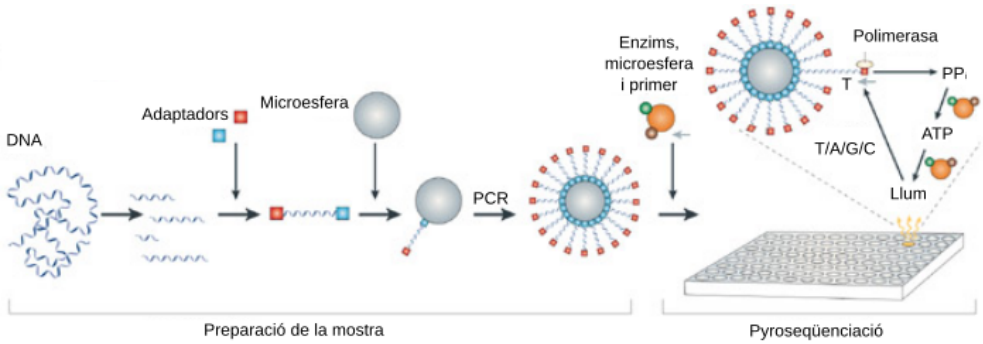


Figura 1.4 Esquema del mètode de piroseqüenciació 454. La cadena d'ADN és amplificada utilitzant la tècnica "Un fragment, una perla, una lectura" i seqüenciada mitjançant reaccions de luminescència. Adaptat de Medini *et al.*, (2008).

Mètode SOLiD

El mètode de SOLiD, (*Sequencing by Oligo Ligation Detection*) és un mètode de seqüenciació desenvolupat per Applied Biosystems (ara Life Technologies), comercialment disponible des de l'any 2006 (`http://solid.appliedbiosystems.com/`). El seu mètode de seqüenciació adopta l'anomenada tecnologia SBL, un mètode que usa la sensibilitat d'una lligasa d'ADN per identificar el nucleòtid present en una determinada posició d'una seqüència d'ADN. Durant la construcció de la llibreria, aquesta seqüència d'ADN és flanquejada per, com a mínim, un extrem de seqüència coneguda, i serà amplificada en una emPCR (Shendure *et al.*, 2005). Així, la reacció de seqüenciació consta de successius cicles de lligació, detecció i trencament entre la molècula d'ADN i diferents sondes marcades amb fluorocroms, on cada color representa un parell de bases, aquesta tècnica es va anomenar codificació de 2 bases o "color space". Les contínues evolucions del sistema (5 entre 2007 a 2010) han donat lloc a la versió 5500XL W, amb un rendiment de 320 Gbases per experiment de seqüenciació. Un dels principals problemes d'aquesta tècnica apareix amb les seqüències palindròmiques (Huang *et al.*, 2012). Life Technologies és també la propietària d'una altra plataforma de seqüenciació coneguda com Ion Torrent, sobre la que van concentrat tots els seus esforços. Això, juntament amb la manca de desenvolupament de programari específic per processar les dades en el format del "color space" per part de la comunitat científica, ha fet que la plataforma SOLiD estigui actualment en retrocés respecte a les seves competidores. Aquesta plataforma s'ha utilitzat per a realitzar varis estudis sobre metagenomes en mostres clíniques (Tyakht *et al.*, 2017) o sediments (Costa *et al.*, 2015).

Mètode Illumina GA/HiSeq/MiSeq/NextSeq

Illumina és una de les empreses que actualment domina el mercat de la seqüenciació massiva. Inicialment, el que avui coneixem com mètode Illumina va ser desenvolupat per Solexa. El seqüenciador Genome Analyzer (GA) adoptà la tecnologia de Terminació Cíclica Reversible (CRT) per seqüenciar una llibreria d'ADN obtinguda seguint el protocol d'amplificació clonal en fase sòlida. Aquest concepte va ser inventat l'any 1994 per Bruno Canard i Simon Sarfati a l'institut Pasteur de Paris (Canard i Sarfati, 1994), i fou presentat l'any 1998 per Pascal Mayer i col·laboradors al 5è Congrés d'Automatització en el Mapeig i Seqüenciació d'ADN (Mayer *et al.*, 1998). Aquesta metodologia consisteix en primer crear una llibreria amb els fragments d'ADN, que són lligats als adaptadors de seqüenciació. Després els fragments de la llibreria s'enganxen a l'atzar a la cel·la de flux, de l'anglés *flow cell*, gràcies a les seqüències fixades sobre la placa, que

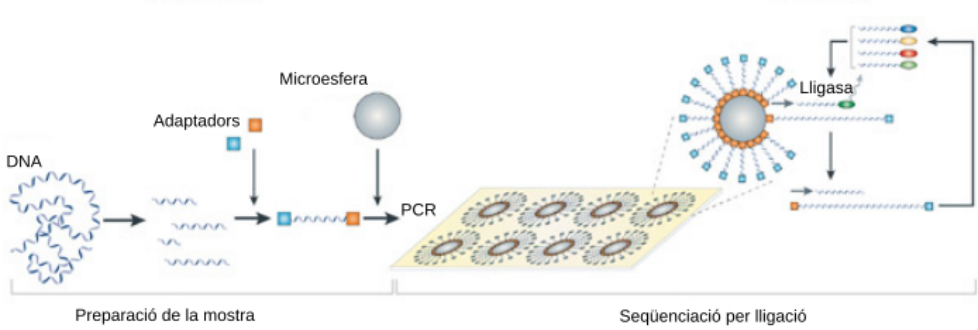


Figura 1.5 Esquema del mètode SOLiD. Utilitza cicles de lligació-detecció-trencament juntament amb sondes marcades amb fluorocroms per tal de seqüenciar la cadena d'ADN. Adaptat de Medini *et al.*, (2008).

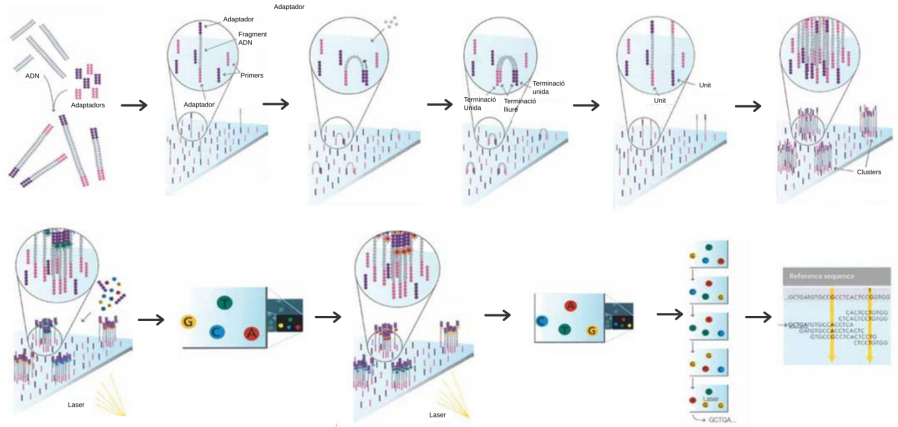


Figura 1.6 Esquema del mètode Illumina GA/HiSeq/MiSeq/NextSeq. Aplica el mètode de terminació cíclica reversible i un protocol d'amplificació clonal en fase sòlida. Adaptat de Illumina, (2017).

són complementàries a un dels adaptadors de seqüenciació. A continuació els fragments s'amplifiquen mitjançant una ADN polimerasa per donar lloc a colònies d'ADN separades, conegudes com a *clusters* o polonies (colònies clonals de la polimerasa), cadascuna d'elles amb aproximadament 1 000 *amplicons* idèntics. Milions de clusters poden ser amplificats a cadascun dels 8 carrils que existeixen en una cel·la de flux. La seqüenciació tindrà lloc mitjançant la repetició dels següents passos: incorporació d'un únic nucleòtid marcat amb un fluorocrom; adquisició de la imatge; i trencament i eliminació del grup bloquejant a 3' del nucleòtid i de l'etiqueta fluorescent per permetre un nou cicle d'amplificació.

El seqüenciador GA de Solexa tenia un rendiment inicial d'una Gigabase per experiment de seqüenciació. Gràcies a millores en l'eficiència de la polimerasa, en els tampons de reacció, en la cel·la de flux i el programari; a l'agost del 2009 van aconseguir augmentar el rendiment a 20Gbp per experiment (*run*), a 30Gbp a l'octubre, i fins a 50 Gbp al desembre d'aquell mateix any. A principis del 2010, Illumina comercialitza el seqüenciador `HiSeq` 2000, que té la mateixa estratègia de seqüenciació que el GA, però un rendiment de fins 200Gbp per *run* de seqüenciació, actualment esta millorat fins a 1 000Gbp. Utilitzant aquesta tècnica s'han realitzat estudis de tot tipus com en mostres de sang (Winn *et al.*, 2011), seqüenciació de genomes petits d'insectes (Kanda *et al.*, 2015), o gens bacterians 16S (Burke i Darling, 2016).

Mètode Ion PGM

La tecnologia desenvolupada per Ion Torrent en els seus seqüenciadors Ion Torrent Personal Machine (PGM) i Ion Proton va permetre la fabricació dels primers seqüenciadors que no necessiten fluorescència, ni càmera CCD, fet que s'ha traduït en una velocitat de seqüenciació més elevada, un cost més baix, però una mida de *reads* més reduïda (Rothberg *et al.*, 2011). Desenvolupada per Jonathan M. Rothberg, un dels fundadors de l'empresa 454 Life Sciences, aquesta tecnologia de seqüenciació fa ús de la seqüenciació per síntesi sobre una llibreria d'ADN construïda mitjançant emPCR. A diferència de la resta de mètodes, l'Ion Torrent detecta el protò o àtom d'hidrogen (H^+) alliberat durant el procés d'incorporació d'un nucleòtid, mitjançant un sensor de pH, cosa que dóna lloc a un pols elèctric que serà transcrit a seqüència d'ADN, evitant així l'ús de qualsevol sistema òptic. L'existència d'un xip semiconductor on té lloc la construcció de la llibreria i la reacció de seqüenciació han fet que sovint la tecnologia es conegui com a seqüenciació Ion semiconductora. D'entre les fites més importants d'aquest tipus de seqüenciador podem destacar la detecció del gen codificant d'una toxina a la soca O104:H4 d'*Escherichia coli*, que va causar 50 morts a principis de maig de 2011 (Mellmann *et al.*, 2011). Amb l'Ion Proton, el 2010, Life Tech-

nologies, va arribar a seqüenciar aproximadament 100 Gb de *reads* per *run* amb una longitud mitjana de lectura de 200 nucleòtids per *read*. Actualment, amb l'Ion PGM, poden arribar a seqüenciar 5.5 milions de *reads*, amb una llargada mitjana entre 200 i 400pb (Kim *et al.*, 2017). Aquesta tècnica s'ha utilitzat en detecció de mutacions en casos cardiomiopaties (Zhao *et al.*, 2016), en mostres d'aigua de mar (Lim *et al.*, 2014) o anàlisis de diversitat genètica en Influenza A (van den Hoecke *et al.*, 2015).

Mètode PacBio

Pacific BioSciences va treure al mercat el seqüenciador PacBio RS l'any 2009 (GenomeWeb, 2009a; GenomeWeb, 2009b). L'avantatge d'aquesta tècnica és la possibilitat de seqüenciar una molècula "única" en temps real. El mètode consisteix en trencar l'ADN i afegir una "A" als fragments obtinguts, que posteriorment permetrà lligar-la amb adaptadors que tenen una "T". Els adaptadors són unes molècules de cadena senzilla d'ADN que prenen la conformació d'una forquilla intramolecular; això genera una molècula amb una forma característica coneguda com *SMRTbell DNA*. A continuació es col·loca damunt un xip de vidre que conté múltiples pous que s'han denominat com a guies d'ona en mode 0 (*o zero-mode waveguides, ZMW*). Les ZMW són estructures circulars amb forma de pouet d'uns 70nm de diàmetre i 100nm de profunditat que permeten confinar ens fotons en un volum aproximat de $20x10^{-21}$L (Levene *et al.*, 2003). En cada ZMW únicament hi cap una polimerasa i una molècula d'ADN motlle. Un cop finalitzada la fase de fixació, la polimerasa va afegint els nucleòtids, específicament marcats amb fluoròfors diferents, a la cadena replicada d'ADN. En el procés d'incorporació dels nucleòtids, aquests s'acosten a la base del pou ZMW on hi ha una càmera d'alta resolució que captura el canvi de fluorescència que emet el nucleòtid en alliberar un fòsfor; aquest procés es repeteix fins a desxifrar la seqüència completa (Brakmann, 2010). Donada la sensibilitat de la càmera, és possible capturar aquest senyal emès per un fotó, el que implica menor distorsió de la seqüència per no haver d'amplificar el senyal, com passa amb altres tècniques ja sigui per haver de replicar la mostra via PCR en polònies (colònies de polimerasa) o per amplificar la reacció química per observar un senyal més intens. Abans que es comercialitzessin els *primers* aparells PacBio, es van publicar dades obtingudes amb aquesta metodologia de seqüenciació d'una única molècula en temps real (Eid *et al.*, 2009), es va aplicar també per detectar isoformes (Sharon *et al.*, 2013) o els metilomes de 6 sis bacteris diferents (Murray *et al.*, 2012).

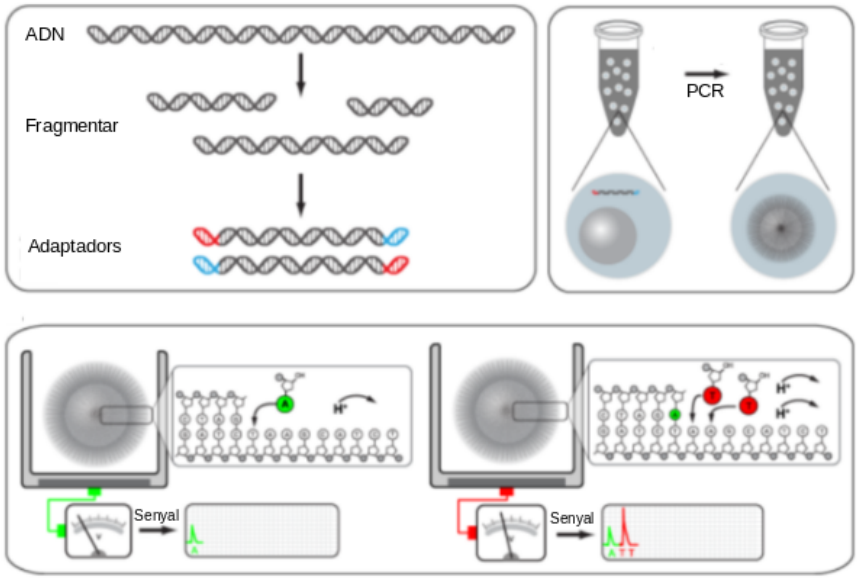


Figura 1.7 Esquema del mètode Ion PGM. Aquest mètode està basat en la detecció de protons alliberats durant el procés de polimerització de la rèplica sobre el motllo d'ADN. Adaptat de Genomics, (2015).

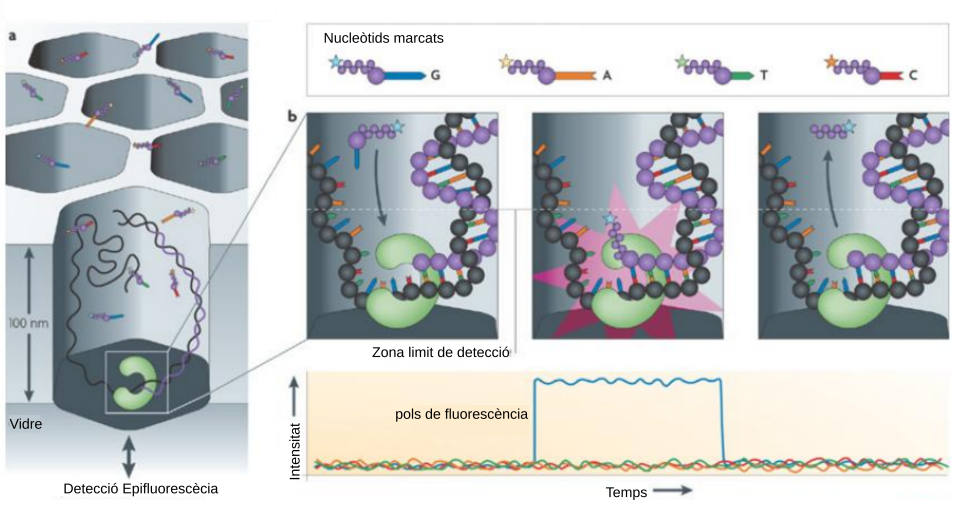


Figura 1.8 Esquema del mètode Pac Bio. Els fluoròfors marcats dels nucleòtids són detectats en ser alliberats, després de ser estimulats per un làser, són detectats per una càmera d'alta resolució. Adaptat de RNA-Seq blog, (2017).

Mètode Oxford Nanopore

El mètode de seqüenciació basat en nanopors va ser desenvolupat per Hagan Bayley el 2012 per la companyia Oxford Nanopore. Es classifica com a mètode de seqüenciació de tercera generació, ja que no requereix modificacions de nucleòtids i els protocols s'executen a temps real. El mètode consisteix en fer passar les seqüències d'ADN per uns nanopors amb un diàmetre de 10^{-9}m (dela Torre *et al.*, 2012), que es troba submergit en un fluid conductor i s'aplica un voltatge. L'efecte resultant és que s'observa un petit corrent a través del nanopor. A continuació es registren els canvis de corrent que es produeixen al passar les molècules dels nucleòtids de la cadena d'ADN a través del nanopor, en funció dels pics de voltatge es pot determinar la composició de les bases de la seqüència (Chang *et al.*, 2010; Pathak *et al.*, 2012). Actualment la taxa d'error és del 5% en la determinació dels nucleòtids. Un dels principals avantatges d'aquesta tècnica és que no cal fer cap pas previ d'amplificació de l'ADN; s'obtenen seqüències molt llargues, de fins a 150kbp; i el preu es molt més baix comparat amb els mètodes de segona generació ja existents. Aquesta metodologia té capacitat per produir entre 50-250 bases per segon i porus, que com a conseqüència seqüencia 500 bases per segon (McNally *et al.*, 2010). Aquesta tècnica s'ha utilitzat ja en diversos estudis com la detecció de haplotips en HLA (Ammar *et al.*, 2015), monitoritzar l'Ebola (Loman *et al.*, 2015), o per identificar de manera ràpida patògens vírics (Greninger *et al.*, 2015).

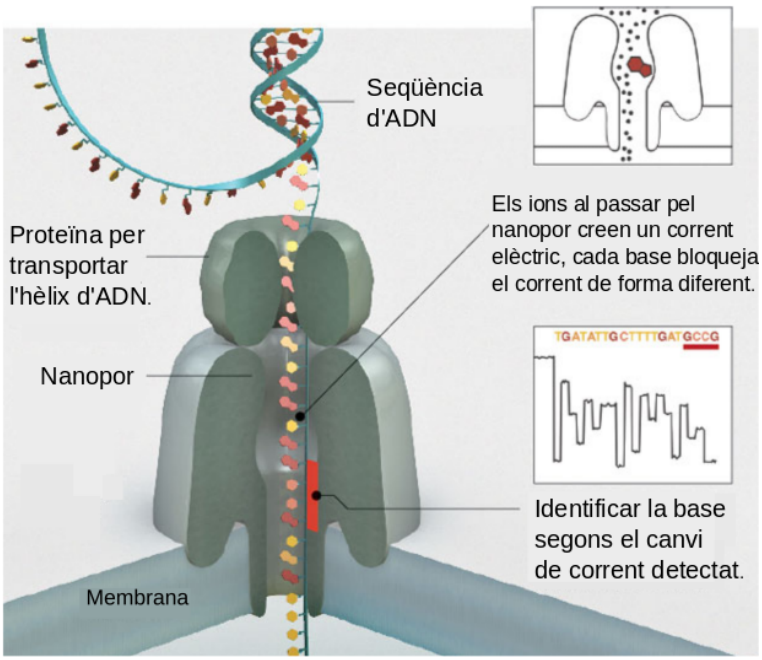


Figura 1.9 Esquema del mètode Nanopore. Uns nanopors de proteïna s'encaixen entre dos compartiments. Al passar les molècules a través del porus, el dispositiu permet detectar els canvis de corrent elèctric entre els dos compartiments. Adaptat de Nanoporetech, (2017).

Mètode	Llargada *Reads*	# *Reads*	Cost x Mpb	Duració del *run*	Principals Avantatges	Desavantatges
Sanger clàssic	400-900 pb	1̃96	2400$	20 min-3 hores	*Reads* llargs	Costòs
Piroseqüenciació (454)	700 pb	10^6 M	10$	24 hores	Ràpid Llargada gran	Costòs Errors homopolímers
Per lligació (SOLiD)	50+35 pb 50+50 pb	1.2-1.4x10^9	0.13$	1-2 setmanes	Costòs	Temps (lent) Errors seq. palindròmiques
Seqüenciació per síntesi (Illumina)			$0.05-$0.15	1-11 dies	Alt nombre de *reads*	Costòs Concentració alta d'ADN a la mostra
NextSeq	75-300 bp	1-25x10^6				
MiSeq	50-600 bp	1-25x10^6				
HiSeq	300 bp	3x10^9				
Ion semiconductor (Ion Torrent)	>400 pb	>80x10^6	1$	2 hores	Ràpid Equipament barat	Errors homopolimers
PacBio	10-15 kb	3.65x10^5	800$ per SMRT	2 hores	No amplificació sensibilitat alta	Baix nombre de *reads*
Oxford Nanopore	500 kb	tria usuari	500-999$ per flow cell	2 dies	Llargada gran Portable	Baix nombre de *reads*

Taula 1.2 Resum de les característiques dels diferents seqüenciadors. La primera columna ens mostra el nom del mètode de seqüenciació; la segona la llargada mitjana dels *reads* que s'obtenen. La tercera columna correspon al número de *reads* que es poden arribar a obtenir amb la tècnica. La quarta columna és el cost aproximat per base. La cinquena columna correspon a la duració del procés de seqüenciació. La sisena i setena columnes presenten respectivament un petit resum dels avantatges i inconvenients de cada tècnica.

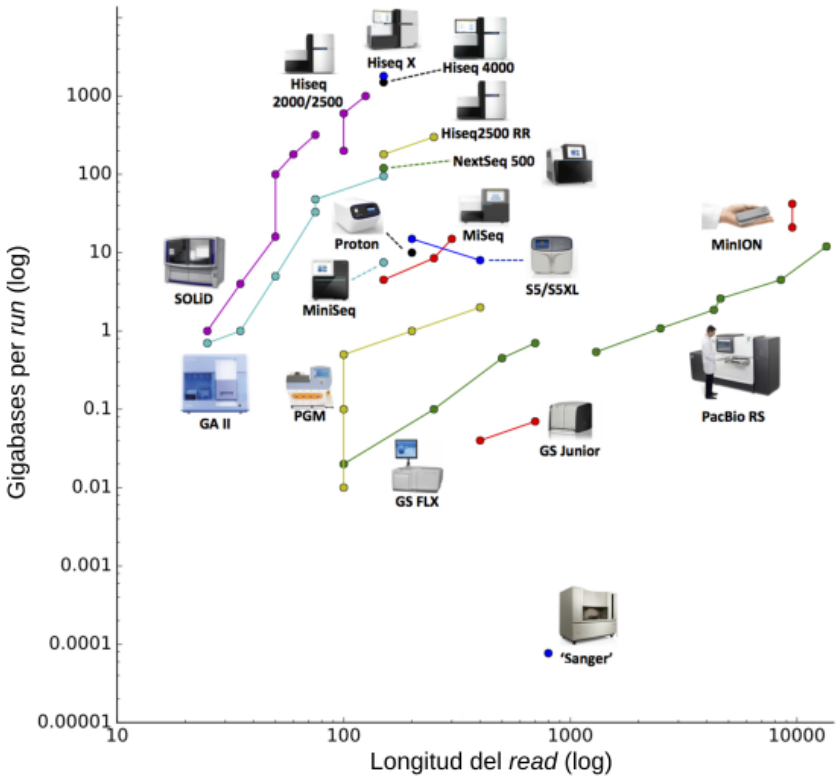


Figura 1.10 Comparació del rendiment dels diferents aparells de seqüenciació. En aquest gràfic els punts representen el rendiment obtingut per cada metodologia de seqüenciació NGS. S'ha tingut en compte la llargada mitjana dels *reads* que es generen en un experiment, respecte les Gigabases totals que s'obtenen en aquest experiment. Les línies connecten punts obtinguts amb les diferents versions de cada tecnologia; a mesura que es milloren els protocols i els *kits*, s'obtenen millors rendiments. Adaptat d'una presentació de Lex Nederbragt (2012-2016).

1.8 Ensambladors: mètodes i programes

1.8.1 Història

Un cop finalitzada la seqüenciació, el següent pas és intentar agrupar i fusionar els fragments de seqüències obtinguts. Actualment, aquest procés és necessari perquè encara que les tècniques puguin arribar a seqüenciar cadenes llargues, aquestes no tenen la llargada suficient per poder reconstruir directament les seqüències originals dels cromosomes o fragments sencers de les llibreries, ni tan sols un gen o proteïna completa.

El problema d'ensamblar seqüències va començar a principis dels anys 80 amb l'arribada dels primers projectes de seqüenciació de l'ADN. El primer programa per ensamblar va ser introduït per Sanger el 1982 (Sanger *et al.*, 1982). Aquests ensambladors utilitzaven diverses estratègies per poder manipular les seqüències repetitives i els errors de seqüenciació que podien confondre a l'ensamblador. Tanmateix no podien analitzar genomes més grans que els de bacteris de l'ordre de pocs milions de bases.

La majoria dels algorismes formats dels ensambladors utilitzen els *reads* com a *input* d'estructura per crear un graf que connecta els *reads* entre ells. La diferència entre els ensambladors és la manera de construir aquest graf inicial, la seva configuració, com el recorren i el procés de simplificació (Miller *et al.*, 2010). El graf és una estructura de dades abstracta que descriu les relacions de similaritat en un conjunt de *reads*. Matemàticament, un graf representa un grup de vèrtexs (nodes) i arestes. En el graf d'ensamblat, els nodes representen cadenes o subcadenes de nucleòtids dels *reads*, i les arestes representen el sufix/prefix dels solapaments entre els *reads* (o sigui que les arestes serveixen per representar la informació dels alineaments que agrupen els *reads* entre si). Hi ha moltes aproximacions per construir grafs, les quals es poden agrupar en tres: *overlap/layout/consensus*, basat en grafs de solapament; els grafs *de Bruijn*, que utilitzen k-mers per agrupar els *reads*; i el mètode *Greedy*.

1.8.2 Mètodes

Mètode de grafs de solapament (OLC)

Aquest mètode és utilitzat normalment en ensambladors que processen seqüències de *reads* llargs, com els que s'obtenen pel mètode Sanger (DiGuistini *et al.*, 2009). Està optimitzat per genomes grans, com els d'eucariotes; era un dels més utilitzats per ensamblar *de novo* i es divideix en tres fases (Peltola *et al.*, 1984). El primer pas consisteix en calcular i descobrir tots els possibles solapaments per parelles entre totes les seqüències fent servir una variant de l'algorisme d'alineament local basat en la comparació de sufixes/prefixes. Posteriorment, aquesta informació és organitzada en un graf on els *reads* són representats pels nodes i les arestes són els solapaments entre aquestes (Myers, 1995). L'objectiu és trobar el camí Hamiltonià més curt, que visiti cadascun dels nodes del graf exactament un cop, i aquest camí representarà la solució al problema d'ensamblat. Per finalitzar, es combinen els solapaments entre els nodes (Figura 1.11). Aquest mètode és utilitzat per diversos ensambladors com `Newbler` (Margulies *et al.*, 2005), `CABOG` (Miller *et al.*, 2008), `Shorty` (Hossain *et al.*, 2009), `Forge` (DiGuistini *et al.*, 2009), `Edena` (Hernandez *et al.*, 2008), `SGA` (Simpson i Durbin, 2012), `Fermi` (Li, 2012), `Phrap` (Gordon *et al.*, 1998; Gordon *et al.*, 2001) i `Readjoiner` (Gonnella i Kurtz, 2012).

Mètode de grafs *de Bruijn* (DBG)

Els ensambladors basats en grafs *de Bruijn* modelen la relació entre subcadenes exactes de longitud *k* dins dels *reads*. De manera semblant al mètode OLC, els nodes representen els k-mers i les arestes indiquen quins k-mers adjacents es solapen per $k-1$ lletres, pel que la longitud del k-mer es correlaciona amb la longitud del solapament que l'ensamblador és capaç de detectar (Pevzner *et al.*, 2001). En aquest mètode no es modelen els nodes directament a partir dels *reads*, sinó que aquests estan implícitament representats pels connectors del graf *de Bruijn* (Idury i Waterman, 1995; Figura 1.12). Alguns ensambladors que apliquen aquest mètode són `Euler-SR` (Chaisson i Pevzner, 2008; Chaisson *et al.*, 2004), `ALLPATHS-LG` (Butler *et al.*, 2008; Gnerre *et al.*, 2011; Maccallum *et al.*, 2009), `Velvet` (Zerbino i Birney, 2008), `ABySS` (Simpson *et al.*, 2009), `SOAPdenovo` (Luo *et al.*, 2012), `CLCBio` (QIAGEN, 2008) i `SparseAssembler` (Ye *et al.*, 2012).

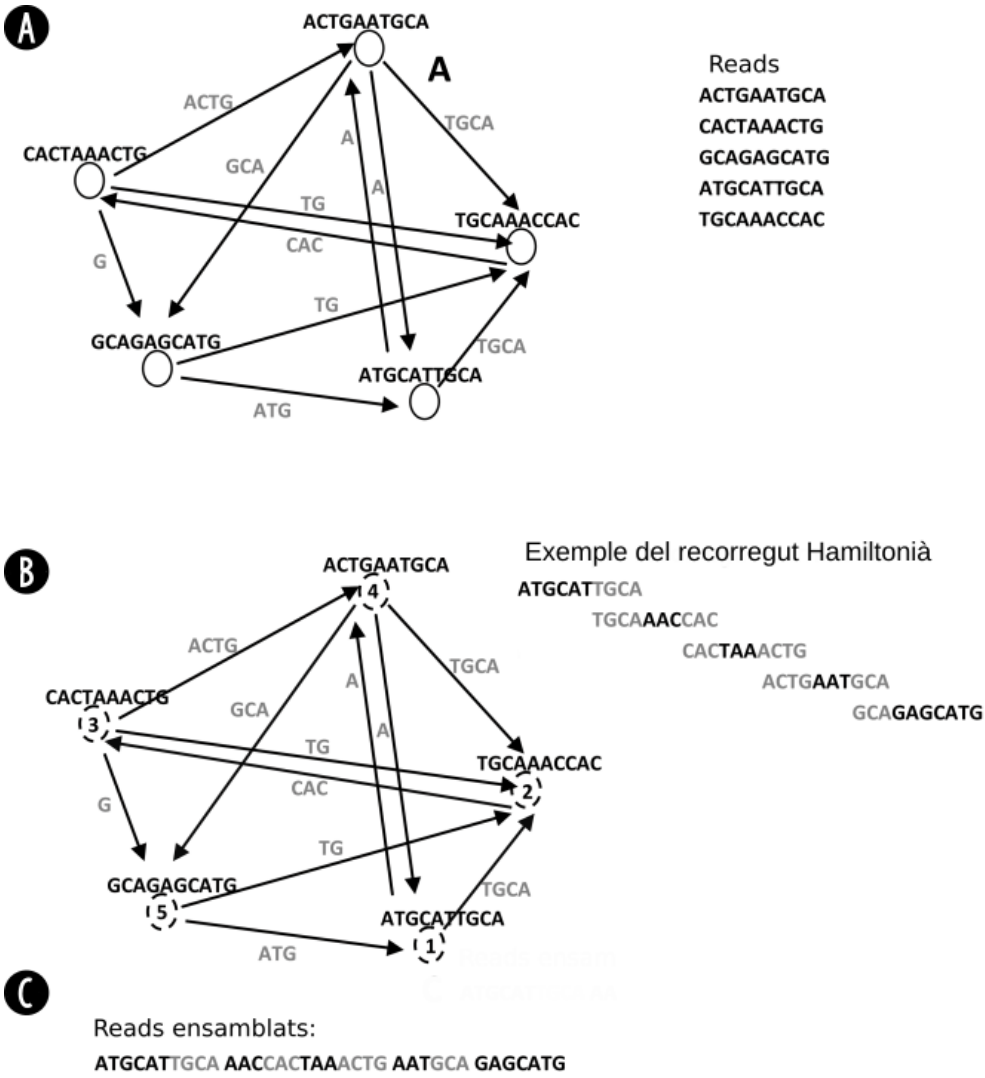


Figura 1.11 Esquema del mètode de grafs de solapament. A: Graf de solapament on els nodes són els *reads* i les arestes els solapaments entre ells. B: Exemple del recorregut Hamiltonià que visita cada node un cop. C: *Reads* ensamblats a partir dels nodes que visita el recorregut Hamiltonià. Adaptat de El-Metwally *et al.*, (2013).

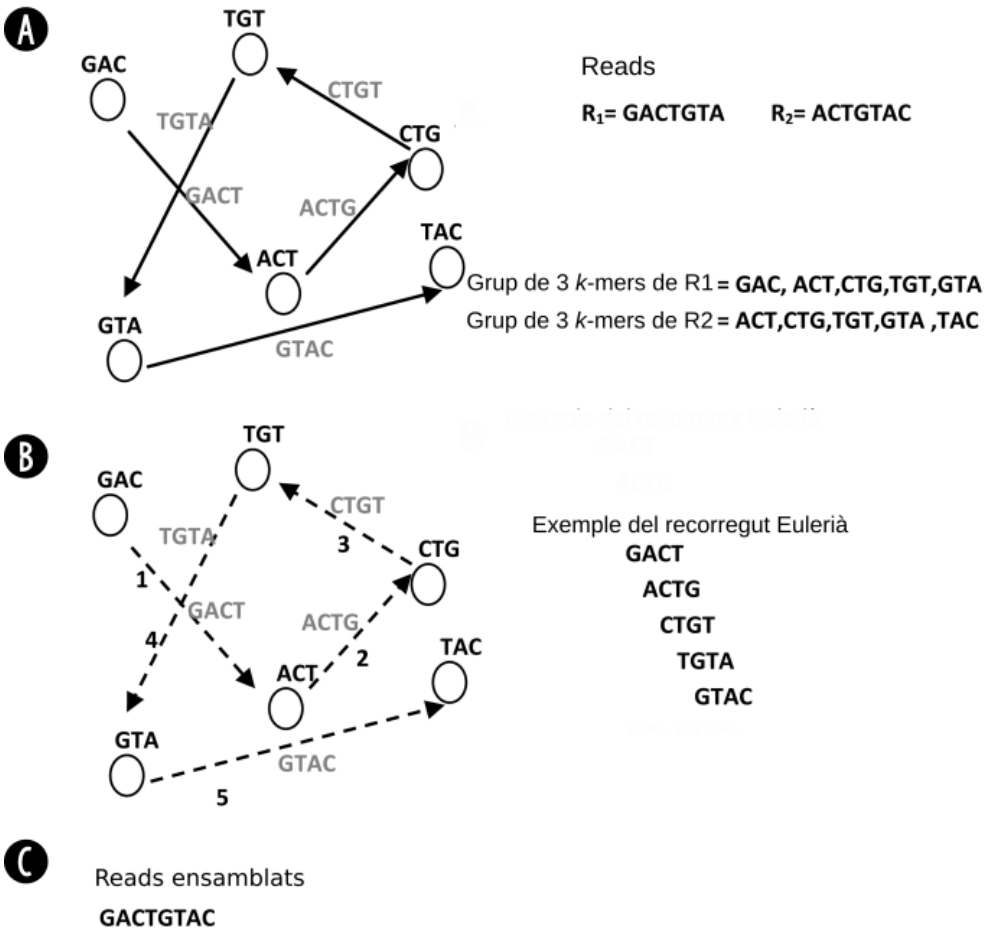


Figura 1.12 Esquema del mètode *de Bruijn*. **A**: Graf *de Bruijn* on els nodes són els k-mers i les arestes els solapaments $k-1$. **B**: Exemple del recorregut Eulerià que visita sols un cop cada node. **C**: *Reads* ensamblats a partir dels nodes visitats seguint el recorregut Eulerià. Adaptat de El-Metwally *et al.*, (2013).

Mètode de grafs *Greedy*

Aquest mètode, que era utilitzat pels primers ensambladors desenvolupats per processar *reads* obtinguts per NGS, està optimitzat per una operació bàsica. Tenint un graf dibuixat, on novament cada node és un *read* i les arestes són el solapament entre elles, es tria el següent node que té el solapament que puntua més alt. Aquest pas es realitza tantes vegades com sigui possible. Per calcular la puntuació del solapament es tenen en compte el número de bases coincidents en el solapament (Tarhio i Ukkonen, 1988). Un dels problemes d'aquest tipus d'ensamblador és la incorporació de solapaments erronis. Aquests falsos solapaments són producte de la presència de seqüències repetitives, ja que la seva puntuació serà superior a la produïda pels solapaments entre dos nodes que no tenen repeticions. El fet que un solapament sigui un fals negatiu comporta unir seqüències no relacionades entre elles produint una quimera (Figura 1.13). Aquest mètode és utilitzat per ensambladors com `SSAKE` (Warren *et al.*, 2007), `SHARCGS` (Dohm *et al.*, 2007), `QSRA` (Bryant *et al.*, 2009) i `VCAKE` (Jeck *et al.*, 2007).

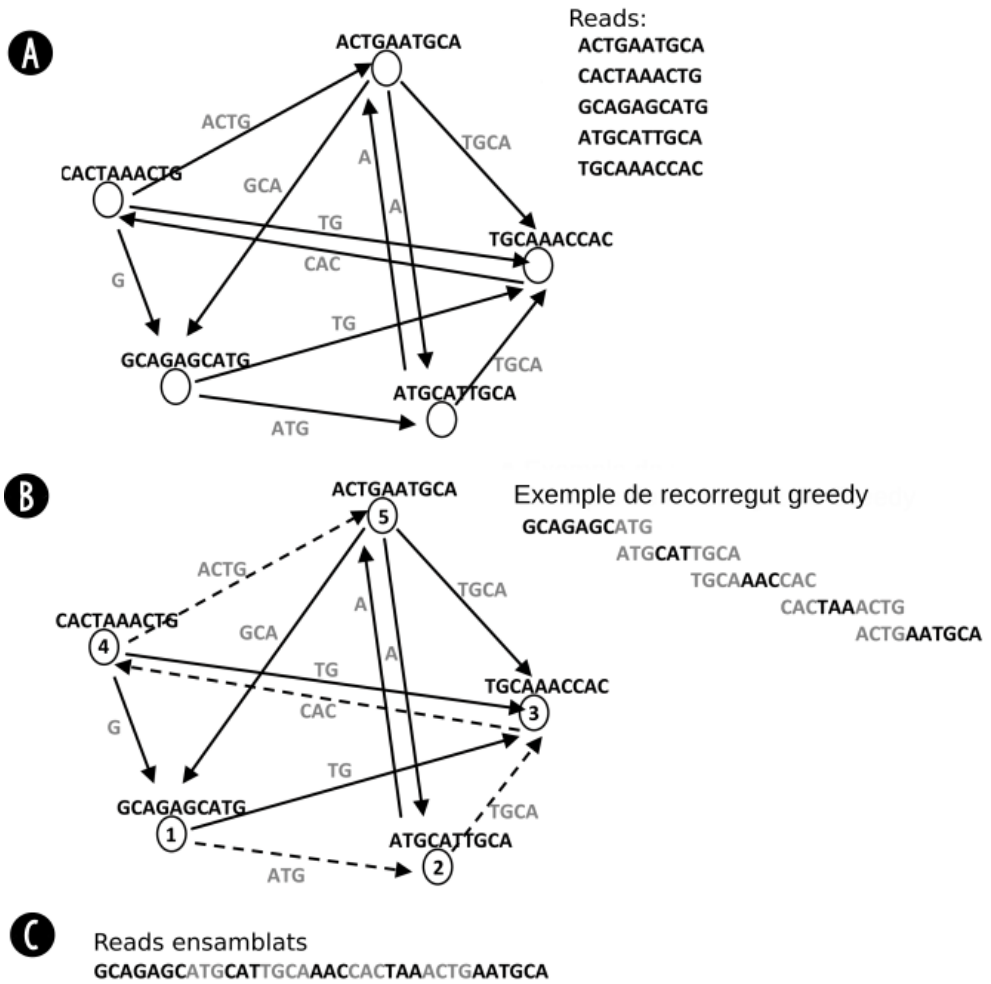


Figura 1.13 Esquema del mètode *Greedy*. **A**: Graf de *greedy* on els nodes representen els *reads* i les arestes els solapements entre ells. **B**: Exemple de recorregut de *greedy* en el qual es visiten els nodes en funció del màxim solapament (el primer node és triat a l'atzar). **C**: *Reads* ensamblats a partir dels nodes que visita el recorregut de *greedy*. Adaptat de El-Metwally *et al.*, (2013).

Nom del programa	Tipus Input	Tecnologies	Construcció del graf	Disponible des de	Llicència	Referència
`Velvet`	Genomes petits	Sanger, 454, Illumina, SOLiD	DBG	2007	Obert (GPLv3)	Zerbino i Birney, 2008
`SSAKE`	Genomes petits	Illumina	Greedy	2007	Obert (GPLv2)	Warren *et al.*, 2007
`SHARCGS`	Genomes petits	Illumina	Greedy	2007	Obert (GPLv3)	Dohm *et al.*, 2007
`VCAKE`	Genomes petits	Illumina	Greedy	2007	Obert (GPLv2)	Jeck *et al.*, 2007
`CLCbio`	Genomes	Sanger, 454, Illumina, SOLiD	DBG	2008	Comercial	QIAGEN, 2008
`ABySS`	Genomes llargs	Illumina, SOLiD	DBG	2008	Obert (GPLv3)	Simpson *et al.*, 2009
`SOAPdenovo`	Genomes	Illumina	DBG	2009	Obert (GPLv3)	Luo *et al.*, 2012
`Newbler`	Genomes, ESTs	454, Sanger	OLC	2004	Comercial	Margulies *et al.*, 2005
`Forge`	Genomes llargs EST	454, Illumina, SOLID, Sanger	OLC	2010	Obert(GPLv3)	DiGuistini *et al.*, 2009

Taula 1.3 Resum de les característiques dels diferents ensambladors. La primera columna correspon al nom del programa d'ensamblat, la segona fa referència al tipus de seqüències d'entrada (*input*) en el qual està especialitzat cadascun dels ensambladors. La tercera columna ens indica la tecnologia de seqüenciació específica per la qual es va desenvolupar el programa. La quarta columna ens diu el tipus de graf en el qual està basat l'algorisme de resolució de l'ensamblat, la cinquena és l'any de creació del programa, la sisena ens indica si el programari és lliure (codi obert, `GPL`) o si es necessita pagar una llicència per utilitzar-lo (comercial). La darrera columna és la referència on es va publicar l'algorisme i la metodologia de l'ensamblador.

Objectius

2 Objectius

L'objectiu general plantejat en aquesta tesi és desenvolupar la metodologia necessària per fer una descripció àmplia i detallada del viroma present en mostres de diferents matrius i aportar nova informació a la comunitat científica en la investigació dels virus presents en aigua i aliments. Aquest objectiu es concreta en l'anàlisi de dades de metagenòmica de virus basada en `MiSeq` que inclou:

- Desenvolupar un protocol computacional per l'anàlisi de dades de seqüenciació en massa de `MiSeq` per l'estudi de virus en diferents tipus matrius.

- Analitzar estadísticament els resultats obtinguts amb els protocols desenvolupats i caracteritzar filogenèticament els conjunts de seqüències virals amb que hem treballat.

- Construir un entorn web per gestionar i accedir a les dades de seqüència, especialment a nivell d'ensamblats, i als resultats obtinguts pels diferents anàlisis realitzats.

- Dissenyar eines per generar taules dinàmiques i *krones* que ens permetin visualitzar els resultats de la seqüenciació de manera intuitiva i integrada.

- Automatitzar el protocol computacional una vegada avaluats, ajustats i optimitzats els paràmetres dels programes que s'han incorporat al mateix.

- Anàlisi i anotació funcional i taxonòmica de seqüències per la descripció del metaviroma de les aigües residuals.

- Anàlisi i anotació funcional i taxonòmica de seqüències virals presents a l'aigua de riu utilitzada com aigua de reg i aliments contaminats.

- Anàlisi i anotació funcional i taxonòmica de seqüències virals de mostres clíniques de patologies associades a virus transmesos per aigua i aliments.

Material i Mètodes

3 Origen de les dades

Tal i com s'ha mencionat als objectius, el protocol bioinformàtic desenvolupat durant la tesi doctoral s'ha aplicat sobre diverses mostres de diferents característiques i procedències. A continuació, passem a donar detalls dels projectes en que s'emmarca aquesta tesi, en concret sobre els estudis i els dissenys experimentals plantejats per a l'obtenció de dades.

3.1 Projectes on s'han obtingut les mostres

Els treballs de la tesi s'emmarquen en el desenvolupament de tres projectes finançats per diferents institucions: RecerCaixa, Asssesora i JPI. En aquests projectes era necessari definir una metodologia per caracteritzar el metaviroma de l'aigua residual, de l'aigua utilitzada per la irrigació d'aliments, i de mostres clíniques. Es van processar mostres d'aigua residual de diferents punts del territori de Catalunya, recollides en diferents estacions de l'any per tal de tenir una perspectiva de les poblacions víriques presents.

Una matriu de dades correspon a aliments que es poden consumir crus, com els vegetals; en aquest cas l'estudi de la contaminació viral en el julivert. Per dur a terme aquest projecte es va realitzar la metagenòmica sobre plantes de julivert cultivats en camps experimentals i regats amb una solució nutritiva per les plantes emprades com a control negatiu, i unes altres plantes de julivert regades amb aigua potencialment contaminada per restes fecals, obtinguda del riu Besós. Paral·lelament també s'han estudiat les dues mostres d'aigua de riu amb les que es van regar les plantes de julivert. Una de les principals patologies associades a virus transmesos per aigua i aliments són les hepatitis agudes; actualment es detecten casos clínics d'hepatitis sense agent etiològic identificat. Per aquest motiu es van analitzar mostres de sèrum humà provinents de pacients que presentaven un quadre agut de símptomes d'infecció per hepatitis sense agent etiològic identificat, per poder així caracteritzar virus que puguin tenir un rol causal en casos clínics d'hepatitis. Es van analitzar cinc *pools* compostos a partir d'un total de 32 pacients positius i quatre *pools* amb mostres de 20 voluntaris sans com a controls negatius.

A continuació es detalla pas a pas com es van obtenir les diferents mostres i com es va procedir amb la concentració i extracció dels àcids nucleics de cadascuna. Cal especificar, que aquests protocols van ser duts a terme principalment per altres membres del grup d'investigació (Fernàndez Cassi, 2017).

3.2 Protocols de tractament de mostres ambientals i d'aliments

3.2.1 Protocols per la recuperació i concentració de virus

Com a pas previ a la seqüenciació en massa de les mostres, hi ha un pre-processat d'aquestes per tal de tenir un concentrat de virus representatiu i el més purificat possible.

Mostres d'aliments (julivert) i aigua de riu utilitzat pel reg

Les mostres de riu provenen del Besòs, un riu de 17,7 km de longitud, en el qual hi van a parar els afluents de 27 plantes depuradores d'aigua (EDAR) i desemboca al mar Mediterrani. Per a l'estudi, es van agafar 10L d'aigua de riu per regar les plantes de julivert i 10L per realitzar la concentració de partícules víriques i per la posterior caracterització dels virus presents a l'aigua mitjançant qPCR i NGS. Amb la mostra d'aigua de riu el pas de concentració de partícules víriques consisteix en una floculació orgànica de la mostra de 10L amb llet descremada, la qual té una eficiència de recuperació d'un 50% (20%-95%, Calgua *et al.* 2013). En resum, aquest protocol consisteix en fer una preparació de 10 grams de llet descremada en pols amb 1L d'aigua de mar artificial, ajustant a un pH de 3,5 utilitzant 1N HCl; així s'obté una pre-solució de llet descremada del 1%(w/v PSM). Les mostres d'aigua de riu s'ajusten a una conductivitat de 1,5mS/cm i s'acidifiquen a un pH de 3,5 utilitzant 1N HCl. S'afegeixen 10mL de PSM a les mostres. Les mostres estan en agitació durant 8 hores a temperatura ambient i posteriorment els flocs són sedimentats per gravetat durant 8 hores. S'extreu el sobrenedant i es conserven el pel·lets, fins a un volum aproximat de 500mL, el qual es centrifuga a 8 000xg durant 30 minuts a 4°C. Els pel·lets són resuspesos en 8mL de tampó fosfat (pH 7,5) i són guardats a una temperatura de -80°C fins que s'extreguin els àcids nucleics.

El julivert ha estat cultivat en els serveis de camps experimentals de la Facultat de Biologia en condicions controlades a una temperatura ambiental de 22°C, una humitat relativa del 60% i una condició de llum de 110nm de radiació fotosintètica activa (*photosynthetical active radiation*, PAR). Les llavors han estat regades amb una solució nutritiva de ferro, nitrogen i fòsfor. Després de 6 setmanes, 12 plantes van ser traslladades a l'hivernacle de les mateixes instal·lacions on van ser regades dues vegades a la setmana durant tot el temps de creixement amb la mateixa solució nutritiva. El mètode de reg fou per inundació de la safata que contenia els tests. Per la meitat de les plantes, a part de ser regades amb la solució nutritiva, les seves fulles van ser polvoritzades amb 15mL d'aigua de riu cada dia. Les fulles de les plantes utilitzades com a control

negatiu es polvoritzaven amb 15mL de la solució nutritiva. Al final de l'estudi, els dos grups de plantes van ser irrigats amb 450mL d'aigua de riu i la solució nutritiva. Després d'un mes de regar-ho diàriament, es van agafar 25 grams de fulles de julivert i es van guardar en una bossa estèril a 4°C durant menys de 48 hores fins que es va procedir a la concentració. El mètode de concentració es basa en rentar els 25 grams dins d'una bossa amb un filtre i afegir 50mL de *buffer* de glicina, a un pH de 9,5 durant 40 minuts sobre un *stomacher*; posteriorment s'ajusta el pH a 7 amb 0,1N HCl. Es centrifuga a 8 000xg durant 10 minuts a 4°C i es recupera el líquid, per eliminar els bacteris i altres materials orgànics. El sobrenedant s'ultracentrifuga a 90 000xg durant 1 hora per concentrar els virus, posteriorment el pel·let és resuspès a un volum final de 500 µL amb tampó de fosfat (pH 7,5) i és guardat a -80°C fins a l'extracció dels àcids nucleics.

Mostres de sèrum humà

Les mostres de sèrum humà es van obtenir a partir d'extraccions sanguínies de pacients de l'Hospital de la Vall d'Hebron de Barcelona, dins dels protocols del diagnòstic rutinari que es duen a terme a l'Hospital. L'estudi ha estat aprovat prèviament pel comitè ètic de l'Hospital i les mostres són analitzades en *pools* anonimitzats sense la identificació individual dels pacients. Es comptava amb 32 pacients: 19 homes i 13 dones d'entre 1 i 92 anys. Vuit d'aquests pacients tenien malalties autoimmunes o estaven immunosuprimits(Ai+ImSP). Deu dels pacients eren positius per hepatitis E per PCR. A més a més, com a control negatiu, es van incloure 20 individus sans. Totes les mostres es van agrupar en diferents grups (o *pools*); cinc grups de pacients malalts (2 grups d'homes, 1 de dones, un grup de Ai+ImSP, i un grup de HEV positiu) i quatre grups d'individus sans. Els diferents *pools* es van purificar amb filtres de baixa absorció proteica de 0,45µm (Millipore Corp.). Posteriorment, el filtrat va ser ultracentrifugat a 100 000xg durant 90 minuts a 4°C i una resuspensió del pèl·let en 500µL de PBS. S'agafaren 300µL i van ser tractats amb DNAsa per eliminar el *background* d'ADN eucariòtic i bacterià que hi pogués haver, emprant 200U TURBO DNase (Ambion, Thermo Fisher Scientific). Les mostres de sèrum es van guardar a -80°C fins la realització del següent anàlisis metagenòmic.

Mostres d'aigua residual

El protocol per la recollida de les mostres d'aigua residual consisteix en agafar entre 500mL i 10L d'aigua residual de l'entrada de la planta de tractament d'aigües residuals de Sant Adrià del Besòs. Les mostres són processades com a màxim dues hores després. En les mostres d'aigua residual es van aplicar dos tipus de concentració: el mètode d'ultracentrifugació (Pina *et al.*, 1998) i el mètode de concentració per floculació orgànica amb llet descremada (Calgua *et al.*, 2008), en el qual s'agafen 500mL d'aigua residual que s'ajusta a un pH

de 3,5. Posteriorment s'afegeix 500µL d'una suspensió de sals marines amb llet descremada (PSM) a pH de 3,5. Les mostres estan en agitació durant 8 hores a temperatura ambient i posteriorment els flocs són sedimentats per gravetat durant 8 hores. El mateix protocol s'utilitzat en les mostres de 10L, ajustant proporcionalment les quantitats de PSM. Els sediments són centrifugats a 8 000xg durant 40 minuts, el pèl·let és resuspès en 4mL de tampó fosfat. Les mostres es van guardar a -80°C fins a realitzar l'extracció d'àcids nucleics.

3.2.2 Preparació de les llibreries i seqüenciació

L'últim pas abans de la seqüenciació, un cop ja realitzada la concentració de les partícules víriques és la preparació de les llibreries de seqüenciació. EL protocol va ser aplicat per igual a totes les mostres.

Primer s'extreuen els àcids nucleics amb el *kit* de QIAmp Viral RNAa MiniKit (Qiagen Inc). Per tal de poder detectar tant els virus ADN com els d'ARN, es generà el cDNA amb una transcriptasa reversa, el Superscript I o II (Life Technologies) i la segona cadena de cDNA i ADN mitjançant la polimerasa Klenow o la Sequenasa 2.0 (USB/Affymetrix). Això ens permet analitzar mostres amb genomes virals de diferents tipus de cadena (dsDNA, ssDNA, dsRNA, ssRNA+, ssRNA-, ssRNA-RT, dsDNA-RT). Per poder obtenir una concentració de la mostra d'1ng/µL per a la preparació de les llibreries, es realitzà una amplificació per PCR amb AmpliTaqGold (Life Technologies). El producte resultant és purificat i eludit en 15µL d'aigua amb el *kit* de Zymo DNA clean (Zymo Research). La construcció de les llibreries es realitzà amb el *kit* de Nextera XT DNA (Illumina Inc.) i la seqüenciació per *pair-ends* amb `MiSeq` 2x300. En aquest mètode es seqüencien els dos extrems dels fragments d'ADN de les llibreries, donant com a resultat dos fitxers `FASTQ`, R1 i R2, que corresponen a les seqüències *forward* i *reverse* respectivament. Finalment tenim un informe inicial sobre el rendiment i la qualitat de l'experiment proporcionat pel servei de seqüenciació.

4 Desenvolupament del protocol Bioinformàtic per genomes virals

L'objectiu principal d'aquesta tesis és el desenvolupament d'un protocol bioinformàtic per analitzar mostres de metagenòmica provinents d'experiments de seqüenciació massiva, en especial, derivats de la tècnica de seqüenciació `MiSeq`. El protocol que es descriurà a continuació es pot aplicar sobre els diferents tipus de matrius explicats en el capítol on es descriu l'origen de les dades. Les diferents anàlisis estan integrades en un sol protocol; però per descriure amb més claredat tot el procés, es pot dividir en tres parts ben diferenciades: l'anàlisi pre-ensamblat, l'ensamblat pròpiament dit i l'anàlisi post-ensamblat. Amb els resultats d'aquest protocol es realitzen diverses anàlisis, incloent estadístiques descriptives de la diversitat taxonòmica de la mostra, així com continuar el processat de resultats amb altres protocols computacionals per anotar genomes i calcular filogènies, en funció del tipus de pregunta biològica que es vulgui abordar. En el diagrama es mostra un resum del protocol desenvolupat que s'anirà detallant en les seccions següents de la Figura 4.1.

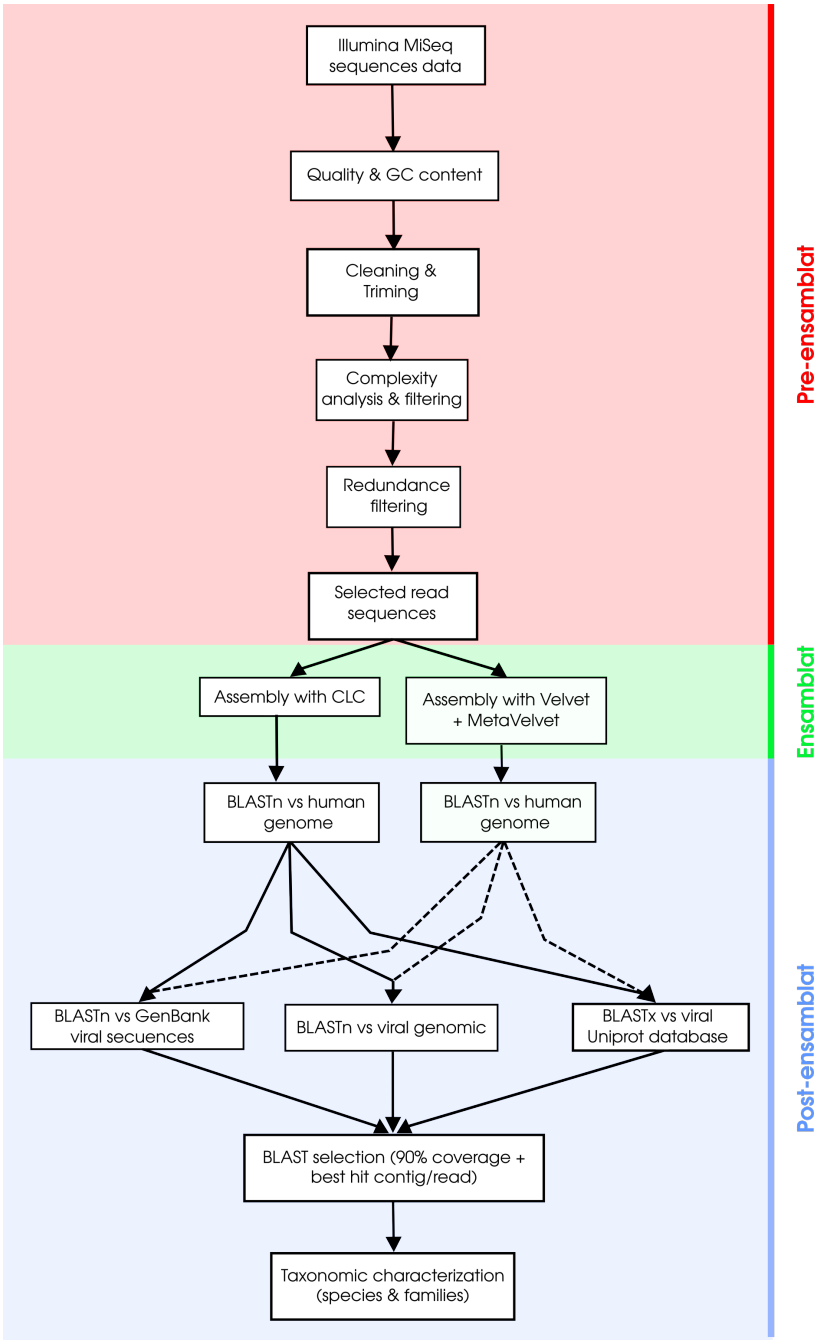


Figura 4.1 Diagrama de flux del protocol desenvolupat. Les capses detallen els diferents passos en que es divideix el protocol. Cada bloc de color diferencia les tres parts més importants del protocol, l'anàlisi pre-ensamblat, l'ensamblat i l'anàlisi post-ensamblat.

4.1 Anàlisi pre-ensamblat

La finalitat d'aquest pre-processat és poder eliminar aquelles seqüències de baixa qualitat, les que puguin ser redundants, i totes aquelles que no ens aporten informació per a realitzar un bon ensamblat de manera eficient i correcta.

4.1.1 Qualitat de les seqüències

El format estàndard de sortida de la major part d'aparells de seqüenciació massiva es el `FASTQ`, en el qual, a part del vector de caràcters de la seqüència de nucleòtids, hi ha un vector de puntuacions amb el valor de qualitat de cadascuna de les posicions nucleotídiques (Cock *et al.*, 2010). Això permet fer una anàlisi de la qualitat de cadascuna de les seqüències, i en funció del vector de qualitats per nucleòtid descartar, ja sigui parcialment eliminant prefixes o sufixes (caps 5' o cues 3' respectivament), o completament descartant tota la seqüència del *read*.

El primer pas per analitzar la qualitat de les sequències, és obtenir una visualització global de la qualitat de cadascuna de les posicions en totes les seqüències; per realitzar això s'utilitza el programa `FASTX-Toolkit` versió 0.0.14 de Hannon Lab (Hannon Lab, 2010), en el qual a partir dels fitxers FASTQ que ens proporciona el seqüenciador, obtenim el gràfic on se'ns mostra la composició nucleotídica com a percentatge de les diferents bases possibles (A, C, T, G o N) de cadascuna de les posicions, tal i com es pot veure a la Figura 4.2A de la pàgina 48. Es pot projectar la distribució de qualitats per posició emprant *boxplots*, com els que es mostren a la Figura 4.2 B de la pàgina 48.

En el gràfic de la composició de la seqüència, el qual podem veure a la figura 4.2A; de cadascuna de les 4 bases (A, C, T i G), s'espera veure una tendència del 25% aproximadament, respecte el total de cada posició. Això es correspon a les freqüències que esperaríem en alinear moltes seqüències escollides a l'atzar. En el cas ideal, el percentatge de Ns seria nul, ja que el seqüenciador assigna a una posició el caràcter "N" quan no ha estat capaç d'adjudicar una base a cap dels quatre nucleòtids. El fet que una mostra presenti regions on no aparegui aquesta distribució d'equiprobabilitat, implica que hi ha algun problema específic amb la seqüenciació que s'ha de corregir abans de continuar el protocol.

Tots els sistemes de seqüenciació tenen una estimació de la probabilitat de que cadascun dels nucleòtids seqüenciats sigui erroni, aquest paràmetre se l'anomena qualitat. Cada seqüenciador calcula una estimació de qualitat a través del programari específic de l'equip. Per poder facilitar l'anàlisi i interpretació dels

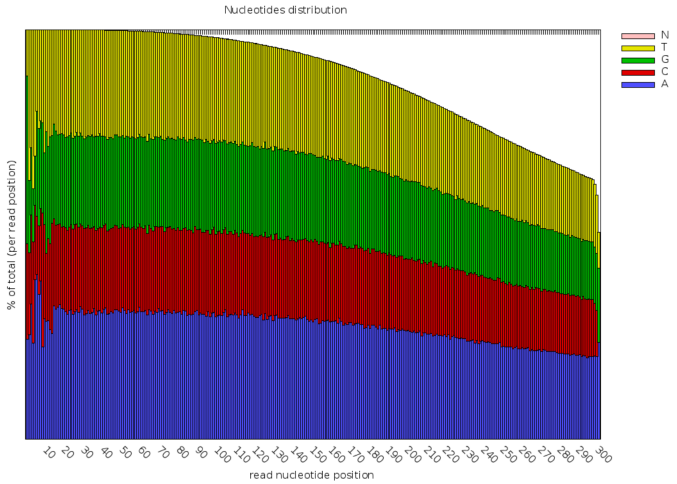
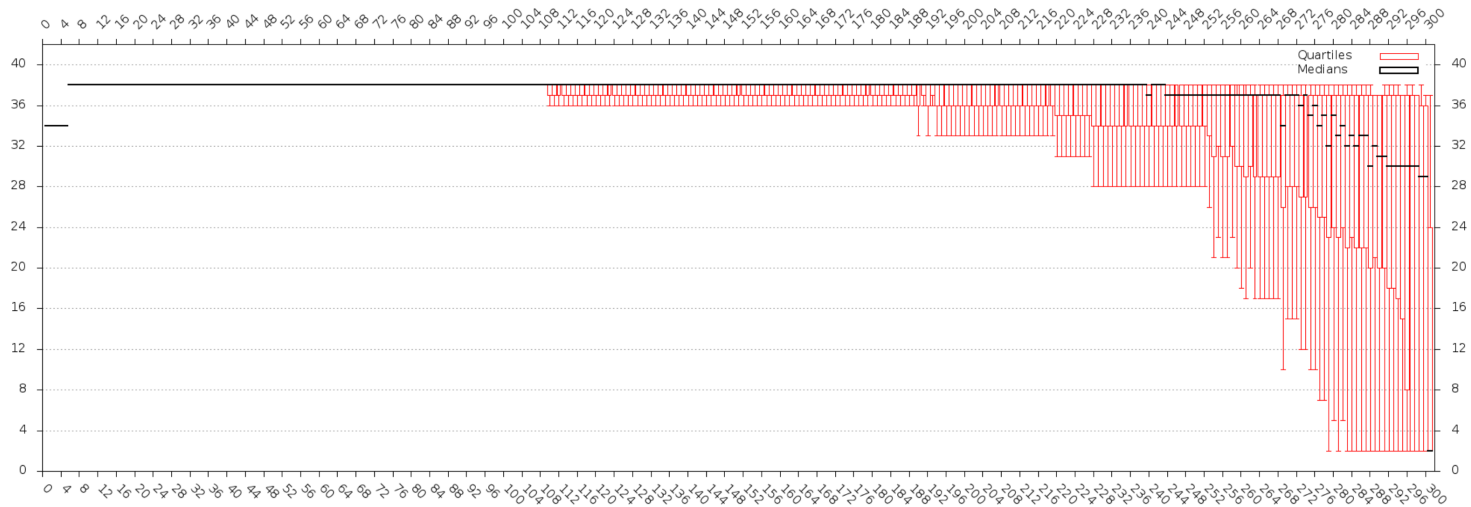


Figura 4.2 Estudi de la qualitat de les seqüències; per composició nucleotídica i valor de qualitat *Phred*.
A: Percentatge de distribució de la composició de nucleòtids per cada una de les posicions de les seqüències. A l'eix de les X es representa la posició dintre de tots els *reads* i a l'eix de les Y el percentatge de composició nucleotídica.
B: *Boxplots* amb la distribució del valor de la qualitat dels *reads* tenint com a valor més baix el 0 i com a valor més alt el 41, per a cadascuna de les posicions de les totes les seqüències. A l'eix de les X apareix la posició de seqüència i a l'eix de les Y el valor de qualitat. En ambdues figures s'han visualitzat dades provinents de *pair-ends* de `MiSeq` de 300bp.



Valor qualitat Phred	Freqüència de base incorrecta	Fiabilitat de seqüenciació
10	1 de 10	90%
20	1 de 100	99%
30	1 de 1.000	99.9%
40	1 de 10.000	99.99%
50	1 de 100.000	99.999%

Taula 4.1 Taula amb els valors de puntuació *Phred* i les correspondències de probabilitat i precisió. Els valors superiors a 60 són generats per programes de validació.

resultats, aquests valors es solen ajustar a una escala normalitzada que utilitzen totes les tecnologies de seqüenciació: l'escala *Phred* (Ewing i Green, 1998), el qual és una mesura logarítmica de la probabilitat d'error de la identificació del nucleòtid generat pel seqüenciador, els valors poden anar de 0 a 60: si s'obté un valor superior a 60 és que s'ha generat pels programes de validació, un exemple del gràfic produit amb aquests valors és la Figura 4.2 B. En la Taula 4.1 es poden veure les relacions entre aquest valor i la freqüència d'error a la qual correspon cada valor. En el gràfic de la distribució de qualitats per posició, tal i com es pot veure a la Figura 4.2 B, podem observar la variació del valor *Phred* al llarg de les seqüències. En aquest gràfic esperem que els valors de *Phred* no variïn molt i que aquests s'aproximin a 40, encara que sovint s'escull un valor mínim suficient com *Phred* 20 o Q20.

En el gràfic de la Figura 4.2A a la pàgina 48 es pot observar que els primer 15bp no segueixen una proporció del 25%; aquestes bases podrien correspondre als adaptadors de seqüenciació utilitzats, per tant no són part de les seqüències víriques. En el gràfic de la Figura 4.2B es pot observar que la qualitat de les posicions finals de les seqüències és baixa; però en aquest cas no es pot establir a partir de quina posició tallar, per tant s'haurà de realitzar a partir del valor de *Phred* i no per posició. Per eliminar aquestes primeres posicions (15bp aproximadament) i totes aquelles posicions que tinguin un valor de qualitat *Phred* inferior a 20, s'utilitza l'eina `FASTX`.

Una altre factor a tenir en compte en la qualitat de les seqüències és la possible presència de *primers* d'amplificació, *barcodes* provinents de la creació de la llibreria o de la pròpia seqüenciació, que no s'han pogut eliminar en el procés anterior en funció de la qualitat. Aquestes seqüències es poden eliminar utilitzant un programa específic com `Trimmomatic` (Bolger *et al.*, 2014) i una base de dades de totes les possibles traces i *primers* que s'empren en la creació,

en el nostre cas, de la llibreria Nextera i del seqüenciador Illumina.

4.1.2 Complexitat de les seqüències

Com ja s'ha comentat, un punt clau en el pre-processat és eliminar aquelles seqüències que no ens aporten informació en el moment de fer l'ensamblat. Un exemple d'aquestes seqüències són aquelles que tenen una complexitat baixa, sovint associades a regions que contenen una composició esbiaixada de repeticions (Hancock, 2002).

Per poder identificar aquestes seqüències es van analitzar diversos paràmetres: nivell de compressió, entropia i valor de Trifonov de cadascuna d'elles. A partir d'aquests valors es va elaborar un criteri de selecció per retenir les seqüències amb major complexitat pels passos posteriors del protocol.

Per nivell de compressió entenem el procés de codificació de dades que utilitzi el mínim nombre possible de bits. Es basa fonamentalment en buscar repeticions en sèries de dades per després emmagatzemar només la dada amb el nombre de vegades que es repeteix (Ziv i Lempel, 1978). Per exemple, si una seqüència o una part d'ella està formada per un homopolímer de 6 A's; sense comprimir ocuparia 6 bytes; però comprimida es podria guardar com a "6A", que correspondria només a 2 bytes. En aquesta situació el nivell de compressió seria del 66,66%. El cas contrari correspondria a una seqüència sense repeticions, amb un nivell de compressió molt baix. Per tant, analitzant aquest paràmetre les seqüències amb una complexitat elevada serien les que tenen el valor de compressió més baix.

La Complexitat Lingüística (CL) de Trifonov és la mesura de "la riquesa del vocabulari". Quan una seqüència de nucleòtids s'analitza com un text escrit amb un alfabet de quatre lletres, es pot calcular la repetició de certs k-nucleòtids (paraules), i utilitzar-la com a mesura de la complexitat de la seqüència. Com més complexa és una seqüència, més ric serà el vocabulari dels oligonucleòtids d'aquesta, mentre que les seqüències repetitives tindran complexitat baixa i poca diversitat de k-nucleòtids (o paraules; Trifonov, 1990). La CL es pot entendre com la representació d'un arbre de subcadenes sobre una cadena de símbols, en el nostre cas un arbre de subseqüències envers la seqüència original. Les seqüències complexes seran representades per un arbre més complet i simètric; la mesura del nivell d'asimetria ens servirà per tenir una mesura de la complexitat. El número de nodes a un nivell j de l'arbre és igual a la mida del vocabulari de les paraules de llargada j de la seqüència; el número de nodes en l'arbre més simètric, correspondrà a la seqüència més complexa d'una llargada N, al nivell j del arbre (Gabrielian i Bolshoy, 1999). La complexitat es pot calcular com el

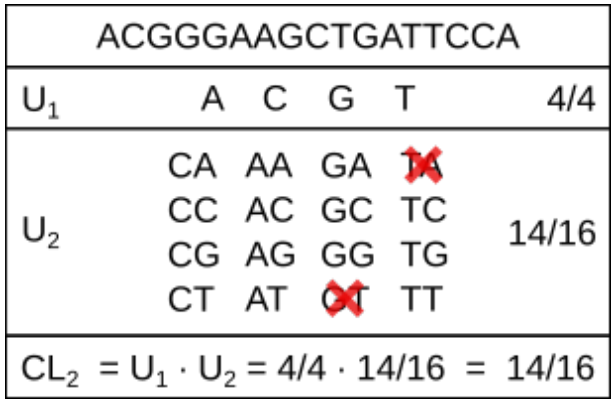


Figura 4.3 Esquema del càlcul del valor de la complexitat lingüística de Trifonov. La primera línia correspon a la seqüència problema. A la segona i tercera caixa es mostra com s'ha calculat el valor de la mesura del vocabulari (U) a nivells 1 i 2; les diferents combinacions de nucleòtids i el valor de U en la seqüència problema. El darrer requadre correspon al càlcul del valor final de la CL.

producte de la mesura del vocabulari a tots els nivells:

$$CL = U_1 U_2 \cdots U_j$$

on la mesura del vocabulari U és el ràtio del vocabulari present respecte al valor màxim de vocabulari de la mateixa mida. Per exemple, U_2 de la seqüència ACGGGAAGCTGATTCCA seria 14/16 ja que conté 14 dinucleòtids diferents dels 16 possibles (més detalls a la Figura 4.3).

D'altra banda, el valor de l'entropia (CE) mesura la quantitat d'informació mitjana que contenen cadascun dels símbols utilitzats; els símbols amb menor probabilitat són els que aporten més informació.

$$CE = -\sum_{i=1}^{K} (n_1/N) log_K (n_i/N)$$

En l'anterior fórmula, N és la mida de la subseqüència, n_i és el número de nucleòtids diferents en la subseqüència i k és la mida de l'alfabet, en el nostre cas 4 (Orlov i Potapov, 2004).

Analitzant la distribució d'aquest valors i totes les seves possibles combinacions, al dibuixar els valors de la complexitat lingüística de Trifonov i el ràtio de compressió de les seqüències, en un gràfic de punts (*scatterplot*) s'observa una tendència a dibuixar una corba de creixement asimptòtic, com es pot veure a la Figura 4.4 de la pàgina 53. Les dades es centraven majoritàriament prop

del punt de la corba on es passa d'un creixement exponencial a estabilitzar-se de manera asimptòtica. Per tant, es podria considerar aquest punt com un bon punt de tall. Per no perdre falsos negatius així com informació de les seqüències, es calcula una línia que passa per aquest punt d'inflexió amb un pendent de 45° i posteriorment es baixa un 5% en paral·lel de la línia original.

Per calcular els diferents valors de complexitat es va implementar un *script* en `Perl`, on per cada seqüència es calculaven els tres valors de complexitat: la complexitat lingüística de Trifonov, el ràtio de compressió i l'entropia. Els gràfics i els càlculs de les línies ajustades sobre els punts d'inflexió es van realitzar a través d'un *script* desenvolupat dins la suit R (R Core Team, 2016) amb el paquet *ggplot2* (Hadley, 2009). Per acabar, mitjançant un altre *script* escrit en `Perl`, es van seleccionar les seqüències que passen aquests filtres.

4.1.3 Redundància de les seqüències

A causa dels clústers que es formen durant la seqüenciació i fragmentació aleatòria de les seqüències originals en crear les llibreries, ens trobem que hi ha seqüències totalment idèntiques o iguals en una part d'una altra seqüència, el que hom anomenaria cadenes duplicades i subcadenes respectivament. El fet de tenir aquestes seqüències ens dificulta l'ensamblat, i el fet que hi hagi menys seqüències en el moment de l'ensamblat ens pot ajudar a reduir el temps d'execució dels programes. Per eliminar aquestes seqüències s'ha utilitzat un programa en `Python` que detecta les seqüències duplicades o que formen part d'altres seqüències i les descarta.

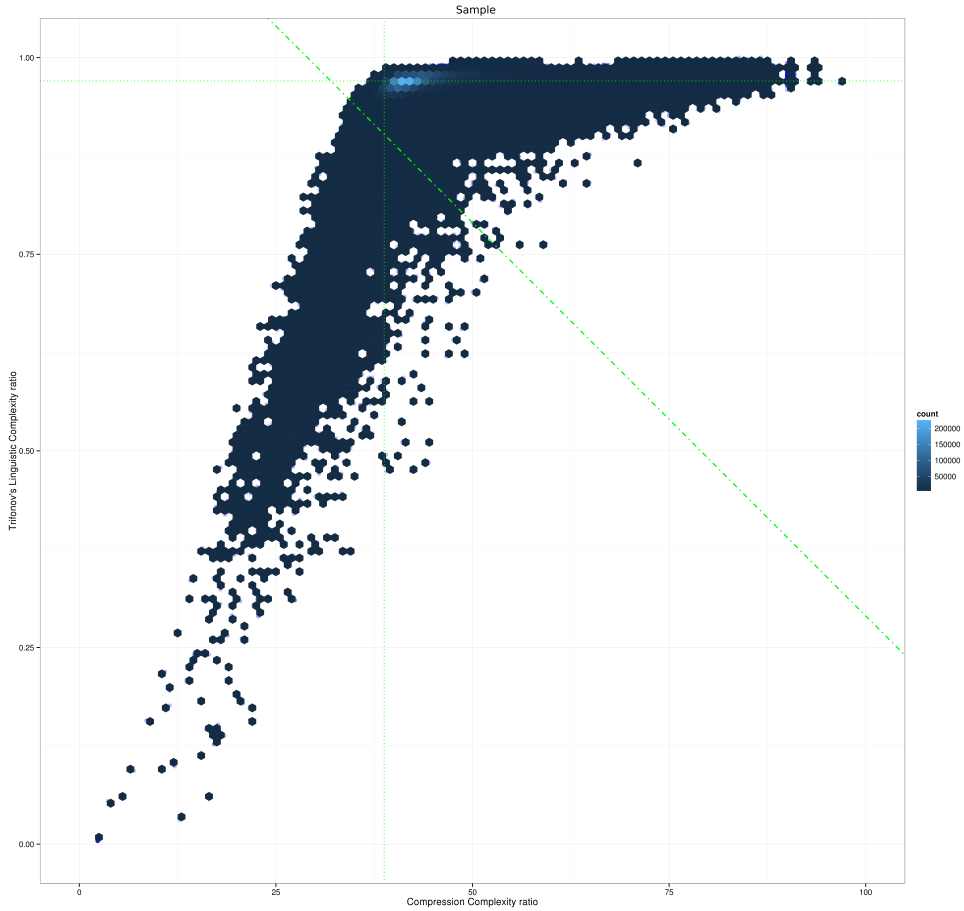


Figura 4.4 ***Scatterplot*** **de la complexitat dels *reads* d'un experiment de seqüenciació**. En l'eix de les X tenim el ràtio de compressió i en l'eix de les Y el valor de complexitat lingüística de Trifonov. Cada punt correspon a un *read* i la diferència en intensitat ens indica el nombre de *reads* amb valors similars (escala de color de la dreta on es veuen els colors associats als comptatges). Les línies verdes de punts es creuen sobre el punt d'inflexió del gràfic, mentre que la línia verda discontínua diagonal és la línia final de tall, que ens defineix el criteri de selecció, per evitar perdre *reads* assignats com a falsos negatius.

4.2 Ensamblat de Metagenomes

En aquest punt del protocol l'objectiu més important és intentar aconseguir seqüències més llargues a partir de les obtingudes en la seqüenciació en massa (*reads*); el que hom defineix com a construir *contigs* a partir dels *reads*. Tal i com s'explica a la introducció, el fet de treballar amb seqüències més llargues, un cop acabat l'ensamblat, permet obtenir una millor caracterització de les mateixes, fent que les cerques d'homologia, així com la resta de passos "*downstream*", siguin més ràpids i eficients. Es poden utilitzar diversos algorismes per realitzar aquest procediment d'ensamblat, en el nostre cas ens vam decidir per programes basats en grafs *de Bruijn*.

Després d'avaluar diferents eines, utilitzarem dos programes: `Velvet` (Versió 1.2.10; Zerbino i Birney, 2008) i `CLC Assembly Cell` (versió 4.4, CLC Inc, Aarhus, Denmark); el primer està dissenyat especialment per *reads* curts, va ser desenvolupat per Daniel Zerbino i Ewan Birney a l'European Bioinformatics Institute (EMBL-EBI) al Regne Unit; el segon programa va ser desenvolupat per QIAGEN (QIAGEN, 2008).

Tal i com s'ha esmentat anteriorment, tant `Velvet` com `CLCBio`, utilitzen el graf *de Bruijn* com a algorisme per ensamblar *reads* curts, representant cadascun dels k-mers obtinguts dels diferents *reads* com a un únic node al graf. Dos nodes estan connectats si els seus k-mers tenen un solapament de $k-1$; és a dir, dos nodes estan units si les últimes $k-1$ posicions del node A són exactament iguals que les primeres $k-1$ posicions del node B. Simultàniament, el mateix procediment es realitza amb les seqüències complementàries reverses de cada un dels nodes, per tal de tenir en compte els solapaments entre *reads* de les cadenes complementàries. Tenint en compte aquest graf s'incorporen optimitzacions com la simplificació de branques i bombolles, així com l'eliminació d'errors.

El pas de simplificació ens serveix per no gastar tanta memòria i consisteix en unificar nodes que no alteren el camí generat pel graf. Per exemple, si un node només té una aresta en comú amb un segon node, s'uneixen els dos en un de gran, unificant la seva informació. Durant el processat del graf poden haver-hi errors causats per la seqüenciació o simplement que la mostra biològica presenti polimorfismes. El programa reconeix tres tipus d'errors, branques, bombolles i connexions errònies, que estan representades en la Figura 4.5 de la pàgina 56.

Un node és considerat com error extrem i és eliminat si és el darrer node del camí i la seva longitud té una llargada inferior a $2k$. Donat que durant la construcció del graf aquest node té un número molt baix de connexions amb altres nodes, és a dir, té molt pocs nodes amb els que té solapaments. Una bombolla es genera quan dos camins comencen i acaben pel mateix node inicial

i final. Normalment són causats per variants biològiques (al·lels). Aquest tipus d'errors són eliminats utilitzant l'Algorisme de *Tour Bus*, el qual detecta quin és el millor camí, decidint quin node s'elimina i quin queda representat en el graf. Finalment, les connexions errònies són aquelles que no generen camins correctes o no creen cap estructura reconeixible al graf. Tots aquests errors s'eliminen després del algorisme de *Tour Bus*, aplicant un valor mínim de cobertura (*coverage cut-off*), que ve definit per l'usuari.

Quan s'utilitza `Velvet` hi ha un pas addicional, l'extensió `MetaVelvet`, creada per l'anàlisis d'ensamblats de múltiples genomes provinents d'una barreja de *reads* de múltiples espècies d'una mateixa comunitat. La majoria dels ensambladors tracten els *reads* com si fossin part d'un únic genoma, i com a resultat, seqüències conservades entre diferents espècies sovint causen quimeres en els *contigs*, però també es poden considerar com a repeticions d'un únic genoma. Per resoldre aquest problema, `MetaVelvet` modifica el graf *de Bruijn* generat per `Velvet`, on la idea fonamental és descomposar-lo en sub-grafs individuals i llavors construir els *contigs* com si cada sub-graf fos una espècie diferent, tal i com es mostra en la Figura 4.6 de la pàgina 57. Amb `MetaVelvet` s'augmenta el número de *contigs* ensamblats i el valor de N50 augmenta (N50 és la longitud dels *contigs* que utilitzant *contigs* d'igual o major longitud produeixen la meitat de les bases de la suma de totes les bases).

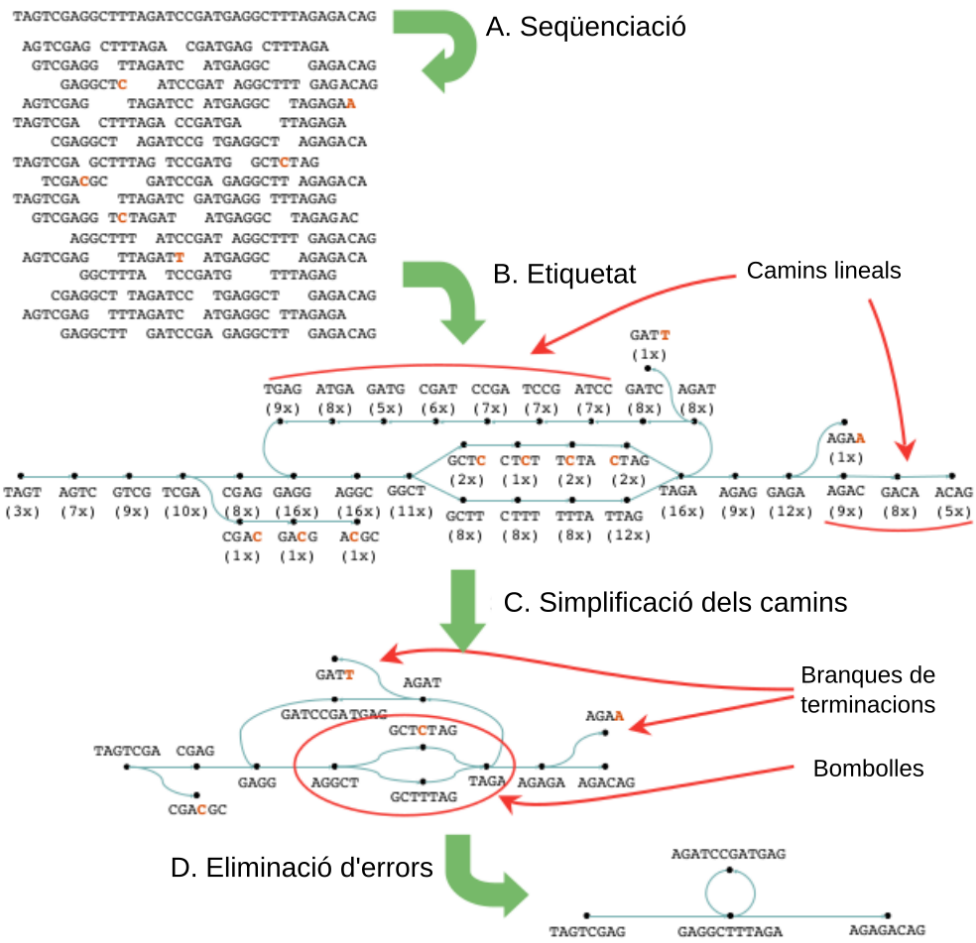


Figura 4.5 Esquema del procés d'extensió i creació del graf *de Bruijn*. S'inclouen a la figura també les optimitzacions aplicades pel Velvet.

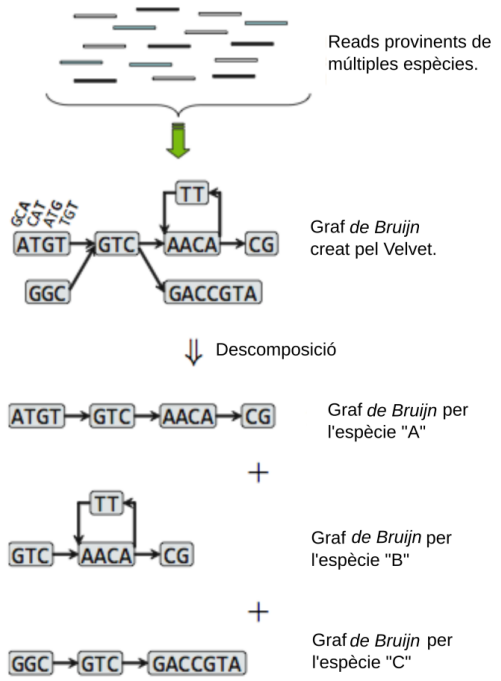


Figura 4.6 Esquema de les optimitzacions de `MetaVelvet`. Es descriuen els canvis en l'algorisme de processat del graf *de Bruijn* que aplica aquesta extensió de `Velvet`, per generar *contigs* espècie-específics.

4.3 Anàlisi post-ensamblat

Un cop tenim les seqüències netes i ensamblades és el moment de començar a treballar amb elles per identificar-les, caracteritzar-les i anotar-les, però també per tenir una idea de com ha anat la seqüenciació i el procés d'ensamblat, a més a més de saber què hem pogut detectar en cadascuna de les mostres.

Una de les contaminacions més freqüents en les nostres mostres de metagenòmica de virus correspon a seqüències humanes; per aquest motiu, un dels primers passos en aquesta fase d'anàlisi és realitzar un alineament amb el programa `Bowtie2` (Langmead i Salzberg, 2012) contra una base de dades del genoma humà (versió hg18) i eliminar totes aquelles seqüències que tinguin almenys un *hit*. El següent pas és identificar cadascun dels *contigs* i/o *singletons* obtinguts en l'ensamblat, ja sigui amb `CLCBio` o `Velvet-MetaVelvet`. Per realitzar aquest pas, hem de fer cerques d'homologia amb el `BLAST` (Altschul *et al.*, 1990; Altschul *et al.*, 1997) contra una base de dades amb seqüències conegudes. En el nostre cas utilitzem tres bases de dades diferents: el primer `BLAST` consisteix en un `BLASTN` contra una base de dades que conté tots els genomes virals sencers

obtinguts del repositori del NCBI-GENOMES; el segon es tracta de realitzar un altre BLASTN; en aquest cas contra una base de dades que conté totes les seqüències virals presents a la base de dades de nucleòtids del NCBI-GENBANK. Finalment el tercer BLAST consisteix en realitzar un BLASTX contra una base de dades generada a partir de totes les proteïnes virals filtrades d'UNIPROT. La Taula 4.2 de la pàgina 59 ens resumeix el nombre de seqüències i el total de nucleòtids o aminoàcids de cadascuna d'aquestes bases de dades mencionades.

Les anotacions taxonòmiques s'obtenen en aplicar una sèrie de filtres sobre els conjunts d'alineaments (HSPs) del BLAST, que permet agrupar les seqüències (ensamblades o no) en funció dels diferents grups o nivells taxonòmics de genomes virals coneguts o bé en funció dels hostes. Per cadascuna de les seqüències que ha fet algun *hit* amb la base de dades, s'obtenen diversos alineaments que s'anomenen segments aparellats d'alta puntuació (HSPs, *High-scoring Segment Pairs*). Amb l'ajut d'un `script` de `Perl`, analitzem diversos paràmetres com l'*E-value*, la llargada de l'alineament i el percentatge d'identitat, per tal de seleccionar el millor HSP per a cada seqüència de cada mostra. Posteriorment, amb una sèrie d'*scripts* en `Perl`, descartem aquelles seqüències que tenen una llargada d'alineament inferior a 100bp, excepte en el cas de la base de dades de proteïna. D'aquests resultats separem la llista de seqüències en tres blocs, les que tenen una identitat superior a 80% i un *coverage* igual o superior a 90%, les seqüències que tenen una identitat inferior a 80% i un *coverage* igual o superior a 90% i les que tenen un *coverage* inferior al 90%. A més a més, es crea un fitxer FASTA amb totes aquelles seqüències que no han obtingut cap *hit* amb el BLAST.

Amb els resultats del BLAST es poden fer alineaments relatius a espècies concretes, però l'objectiu és poder fer una classificació taxonòmica fins al nivell de família i grup Baltimore. Per això s'ha extret de la base de dades de taxonomia de NCBI (Benson *et al.*, 2017; NCBI Resource Coordinators, 2013) la relació entre espècie-família-grup Baltimore; a partir d'aquest fitxer es pot projectar cadascun dels *hits* obtinguts pel BLAST amb la seva taxonomia completa.

Un cop arribat a aquest punt, tenim la informació de cadascuna de les seqüències obtingudes per l'ensamblador, si té o no *hit* amb alguna base dades, i en cas afirmatiu, la informació sobre quina o quines bases de dades a fet *hit*, amb quines condicions (*coverage*, identitat, llargada, *e-value*, *gaps*), i finalment a quin grup taxonòmic pot pertànyer.

Nom de la Base de Dades	Número Total de Genomes	Número de Seqüències	Bases Totals	Seqüència més llarga	Data darrera Actualització
GenBank Virus i fags, seqüències nucleotídiques	—	2 011 686	3 054 343 645	2 473 870	23-05-2016
GenBank Genomes virals, seqüències nucleotídiques	5 516 (5 493)	7 156	218 441 626	2 473 870	23-05-2016
Uniprot, proteïnes víriques	—	2 659 517	822 926 418	8 573	17-05-2016

Taula 4.2 Bases de dades per l'anotació. Resum de les característiques de les diferents bases de dades utilitzades per anotar taxonòmicament i funcionalment les noves seqüències obtingudes a partir de les mostres de metagenòmica. A la columna del nombre de genomes s'indica entre parèntesis el nombre total de genomes complets.

4.3.1 Riquesa del metaviroma (*Richness*)

Un dels paràmetres més interessants a estudiar un cop tenim totes les seqüències de cada mostra anotades a nivell taxonòmic, és la diversitat viral que presenta cadascuna d'aquestes mostres. Per calcular aquest valor de *richness* s'utilitza el programa `CatchAll` (Allen *et al.*, 2013; Bunge *et al.*, 2012). El qual requereix que indiquem quantes seqüències estan relacionades amb cadascuna de les espècies identificades (Bunge, 2011).

`CatchAll`, de fet, és un grup de programes que permeten analitzar la freqüència de comptatges de les dades, basats en abundància i incidència. `CatchAll` fa una estima del valor total del número d'espècies, observades i no observades, utilitzant diferents models que es poden classificar en dos grans grups: els paramètrics i els no paramètrics.

En els models paramètrics sorgeix una complicació, ja que en les dades normalment hi ha un elevat nombre d'espècies que apareixen pocs cops i molt poques espècies que apareixen molts cops. Els models paramètrics no accepten aquest tipus de dades, per solucionar-ho s'ha de procedir a eliminar els valors extrems (*outliers*). Per exemple, es pot ajustar el model paramètric posant un límit en la freqüència τ, eliminant tots aquells comptatges on la freqüència sigui superior o igual a τ, per tant, els models calculats dependran d'aquest valor. El programa permet treballar amb 5 models paramètrics: el model Poisson, el model de Poisson exponencial, una barreja de dos models de Poisson exponencials, una barreja de tres models de Poisson exponencials i un barreja de quatre models de Poisson exponencials (Irene *et al.*, 2011). El programa calcula, per cada conjunt de dades, tots els models per a cada valor de τ, fet que genera molts valors. *A posteriori* es prenen decisions heurístiques basades en experiències empíriques per triar el "millor" model.

En els models no paramètrics l'estimació del nombre total d'espècies es basa directament o indirectament en el valor estimat del *coverage* de la mostra i la proporció de la població que està representada per les espècies (Chao i Lee, 1992). En aquest cas, també s'apliquen els diferents models per diversos valors de τ, però només s'avaluen quan la τ és 10 (o el valor més pròxim), ja que en aquest valor es tendeix a ser molt sensible als valors extrems (*outliers*). Els models calculats són: el *Good-turing*, també anomenat "model homogeni"; el Chao 1; ACE, que estima el *coverage* basant-se en l'abundància; el ACE1, que estima el *coverage* basant-se en l'abundància en mostres heterogènies; i el Chao-Bungue gamma-Poisson, utilitzat quan la mostra segueix una distribució binomial negativa.

Es van fer diferents provatures amb tots els paràmetres per explorar el funci-

onament d'aquesta eina i per poder escollir el millor model. Aquest experiment va consistir en agafar una mostra a l'atzar d'aigua residual, seguidament es van definir 11 grups amb diferents números de seqüències agafades a l'atzar (5 000, 10 000, 15 000, 20 000, 25 000, 30 000, 35 000, 40 000, 45 000, 50 000, 55 000) i per cadascun d'aquests grups es va calcular la riquesa en tots els models. Aquest procés es va repetir 10 cops. Per representar aquests resultats es va utilitzar un *Boxplot* per poder comparar les diferents rèpliques i els diferents models possibles, que es mostra a la Figura 4.8 de la pàgina 63. Primerament es van descartar els models paramètrics, ja que les nostres dades no segueixen una distribució normal; a més a més en els models ACE_Non-P3 i Chao1 els seus respectius *boxplots* pels valors de riquesa presenten menys dispersió. Amb tot això, es va decidir que el model a seguir seria el Chao1, ja que s'adaptava millor a una corba de saturació i presentava una dispersió menor en funció de les dades mostrals, per tant, resulta menys sensible a les variacions que esperem trobar quan treballem amb mostres de metagenòmica de virus.

4.3.2 Anàlisis filogenètics

Una vegada recopilada tota la informació d'homologia de les seqüències, ens podem trobar amb alguns *contigs* que han estat anotats en funció d'un alineament amb una identitat baixa i que a més a més compren tot o una gran part del fragment genòmic que volem utilitzar per caracteritzar la família. En aquests casos pot ser interessant realitzar un estudi filogenètic per inferir les relacions entre les diferents seqüències.

Quan hem tingut un cas d'aquestes característiques, el disseny experimental ha estat el següent: s'han agafat seqüències representatives de les espècies representatives de la variabilitat de la família; s'ha calculat un alineament múltiple amb el programa `Clustal` juntament amb la seqüència que hem trobat a la mostra, aplicant el mètode *Neighbor-joinning* i un *bootstrap* de 1 000. Degut a que la seqüència de la mostra és sovint considerablement molt més curta que la resta de seqüències homòlogues capturades pel `BLAST`, hem aplicat el programa `Gblocks` (Castresana, 2000; Talavera i Castresana, 2007), per tal de només quedar-nos amb les columnes dels alineaments sense *gaps* i amb més conservació. Aquestes posicions de l'alineament són realment on hi ha la informació necessària per poder reconstruir l'arbre filogenètic. En alguns casos, per poder realitzar un arbre amb arrel, s'afegeix una seqüència d'alguna família pròxima de la proteïna o el gen que volem fer servir per caracteritzar les espècies, o fins i tot alguna seqüència de la mateixa família però d'un altre gènere. A l'hora de representar l'arbre i poder modificar-lo per tenir una millor visualització hem fet servir el programa `Itol` (*Interactive tree of live* de `EMBL`; Letunic i Bork, 2016).

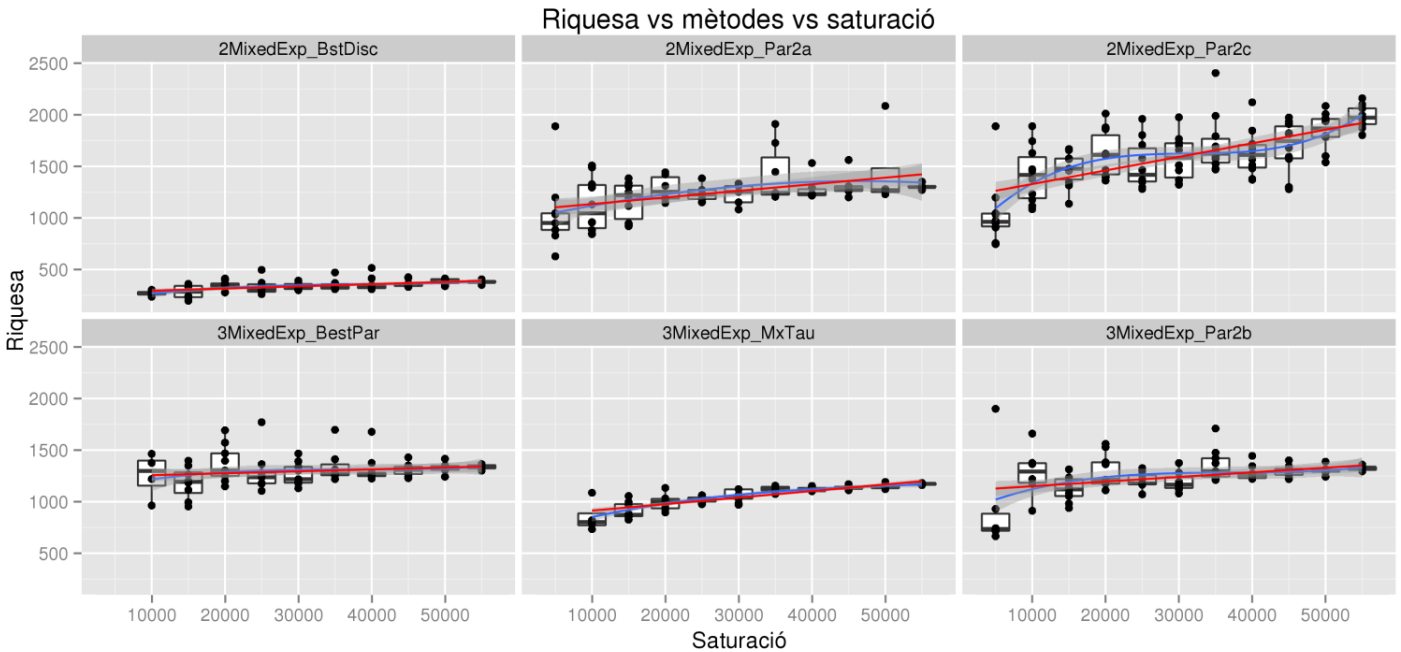


Figura 4.7 ***Boxplots*** **comparatius dels diferents mètodes paramètrics per calcular la riquesa amb el programa** `CatchAll`**.** L'eix de les Y representa el valor de la riquesa, l'eix de les X representa les 10 rèpliques dels 11 grups de diferents nombres de seqüències. La línia vermella representa la mitjana dels valors per cada conjunt de mostres.

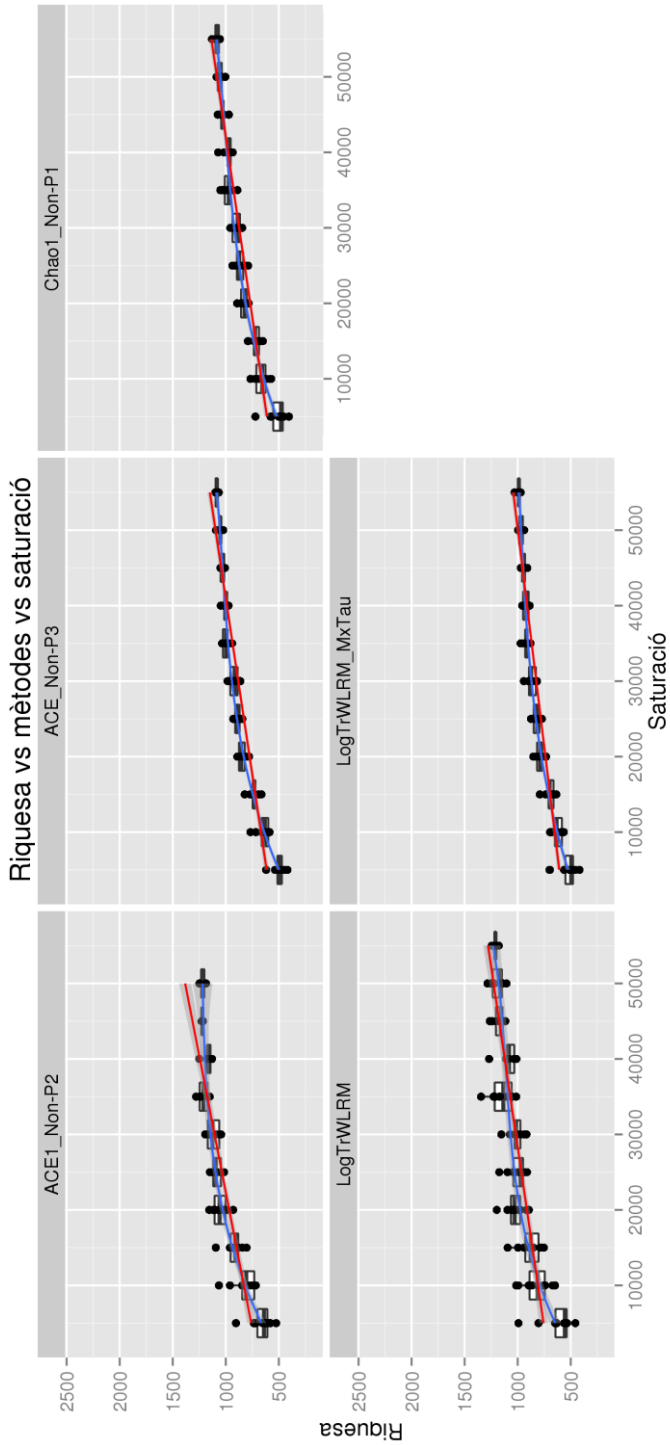


Figura 4.8 ***Boxplots*** **comparatius dels diferents mètodes no paramètrics per calcular la riquesa amb el programa** `CatchAll`. L'eix de les Y representa el valor de la riquesa, l'eix de les X representa les 10 rèpliques dels 11 grups de diferents nombres de seqüències. La línia vermella representa la mitjana dels valors per cada conjunt de mostres.

4.4 Metagenòmica dirigida

Per tal de detectar i tipificar el virus del gènere *Mastadenovirus*, de la família *Adenoviridae*, excretats per la població en les aigües residuals de l'entrada de la planta de tractament d'aigua residual de Sant Adrià del Besòs, es va implementar un segon protocol per generar i processar dades de metagenòmica dirigida derivat del protocol principal WGS. Es van dissenyar *primers* específics sobre l'exó d'AdV. La regió va ser seleccionada per la seva versatilitat, perquè és un segment genòmic amb regions conservades (pel reconeixement d'encebadors i amplificació dirigida) i variables (per acumular canvis que ens permetin fer filogènies; Hernroth *et al.*, 2002). Aquesta mostra es va seqüenciar amb la tècnica 454 de Life Science-Roche. A partir del fitxer de sortida en format SFF, s'extreu la informació dels *reads* i la seva qualitat i són transformats amb el programa `FASTQ` a `fastaq`; a continuació els adaptadors de seqüenciació es van eliminar amb el programa `Cutadapt` (Martin, 2011).

El següent pas va consistir en definir les unitats taxonòmiques operatives (OTUs, *Operational Taxonomic Unit*), es va realitzar amb el programa CD-HIT[1] (Fu *et al.*, 2012). Es van determinar els OTU's a diferents nivells de distància: 0,00; 0,01; 0,02; 0,05 i 0,10. Es va escollir un valor de distància de 0,02. Tot seguit es va crear una base de dades específica que contenia la regió seleccionada de l'exò d'AdV per 153 *Adenovirus* disponibles a `NCBI-GenBank`, i que s'agrupaven en cinc gèneres diferents: *Aviadenovirus* (9), *Atadenovirus* (12), *Mastadenovirus* (122), *Siadenovirus* (4) i *Ichtadenovirus*(1). Finalment, es va inferir un arbre filogenètic amb el programa `RAxML` (Stamatakis, 2014) considerant un valor de *bootstrap* de 1 000.

[1] `http://weizhong-lab.ucsd.edu/cd-hit/`

Resultats

5 Resultats

Al llarg de la tesis s'ha pogut aplicar el protocol desenvolupat que s'ha descrit en el capítol de metodologia en diverses mostres virals de metagenòmica obtingudes a partir de mostres de diferents matrius. De cara a facilitar la lectura dels resultats obtinguts, cada apartat s'ha dividit en 3 grups en funció del tipus de matriu de les mostres: aigua residual, aliments (julivert) i sèrum humà.

5.1 Anàlisi pre-ensamblat

Totes les dades que es descriuen a continuació fan referència a mostres provinents de les diferents matrius que s'han anat multiplexant en diversos experiments de seqüenciació; en alguns casos el *run* era exclusiu i en altres compartit, però en cadascun d'ells s'han inclòs entre 4 i 20 mostres. Això implica que el número de seqüències, obtingudes amb la metodologia `MiSeq` de Illumina, sigui força variable. El número mínim de seqüències obtingudes ha estat 600 000 i el màxim 8 145 076. En les Taules 5.1, 5.2 i 5.3 (de les pàgines 69, 70 i 71 respectivament) es llisten aquests valors i també les estadístiques, un cop aplicats els filtres de qualitat, complexitat i redundància, corresponents de cadascuna de les mostres considerades en el moment de redactar aquesta tesi.

En realitzar l'anàlisi de la complexitat, tal i com es comenta en la Secció 4, es va generar un gràfic de distribució dels valors de CL de Trifonov i el ràtio de compressió de les seqüències. A les Figures 5.1, 5.2 i 5.3 (de les pàgines 72, 73 i 74), es poden veure els resultats desglossats per cadascuna de les mostres de les diferents matrius (aigua residual, julivert i aigua de riu, i sèrum humà respectivament).

En l'estudi d'aigua residual es pot observar que totes les mostres segueixen una corba de creixement asimptòtic, però hi ha algunes diferències entre elles que cal remarcar (Figures 5.1). La mostra R4, que correspon a la mostra d'hivern, no té un punt definit on es concentrin la majoria de les seqüències, sinó que es troben repartides al voltant de la recta de saturació del valor de CL de Trifonov. En les altres mostres es pot distingir clarament un punt just abans d'aquesta saturació. Un altra diferència apreciable es troba a la "cua" del gràfic, on s'acumulen les seqüències de baixa complexitat; per exemple R6 i R8 presenten una major densitat de punts (un punt equival a un *read*) mentre que R4 és molt més difús en aquesta regió del gràfic. Cal destacar la dispersió de les dades que en alguns casos crea una mena de "panxa", fet que ha dificultat estimar-hi les línies de tall.

Per les mostres de julivert i agua de riu aquestes diferències són més exagerades, i ens permeten diferenciar clarament els dos tipus de mostres. En les mostres de julivert, les "cues" de baixa complexitat del gràfic són molt més curtes i la dispersió de les dades en la regió asimptòtica és molt petita; en canvi, en les mostres d'aigua de riu s'observen "cues" més llargues i una dispersió més elevada.

En les mostres de sèrum humà, primerament cal comentar que en el tractament d'una mostra (SH4, veure taula 5.3 i següents) hi va haver algun problema en el processat del material biològic, que es va detectar al fer l'anàlisi computacional. Això ens va fer decidir per repetir l'anàlisi complet de la mostra, els resultats del qual correspon a la mostra SH6. De manera global podem observar que la dispersió de les dades és molt més elevada que en la resta de tipus de matriu.

De les seqüències eliminades no sempre es treia la parella (*forward* i *reverse*), per tant quedaven dos grups de seqüències filtrades, els *pair-end* i els *single-end*. El nombre de seqüències més elevat recau al grup de *paired-ends*, mentre que els *single-ends* normalment representen aproximadament un 0,25% del total de seqüències que han passat el filtre. Cal comentar que en el moment de realitzar el filtrat de complexitat hi va haver dues mostres que presentaven una dispersió de les dades més gran en la part baixa del gràfic provocant que el *script* emprat en la resta de mostres no funcionés adequadament i per això es va haver de calcular manualment la línia de tall.

Indirectament, quan s'aplica aquest filtre les seqüències curtes també es descarten, i com a conseqüència la mitjana de la longitud de les seqüències s'incrementa. La majoria de les seqüències de totes les mostres conserven la seva parella, les seqüències que es queden com a *single-end* van de 4 a 197 320, un número molt inferior a les aparellades (*pair-end*), que presenten valors entre 329 520 i 5 740 372. En les mostres d'aigua residual el percentatge de mostres filtrades varia entre el 81% i el 95%, les mostres de julivert entre un 27% i un 43%, mentre que en el control negatiu el percentatge és el més baix de totes les mostres analitzades amb un 6,5%; finalment per l'aigua de riu entre un 72% i un 98%. El percentatge en les mostres de sèrum humà presenta un rang d'entre el 50% i el 75%.

Identificador de la Mostra	Descripció de la Mostra	Núm. Inicial PE Seqs	Nucleòtids Totals	PE Seqs Filtrades	SE Seqs Filtrades
R4	hivern	2 862 464	653 150 240	2 520 868	7 351
R6	primavera	1 225 920	290 054 808	1 172 444	9 158
R8	estiu	951 664	223 053 943	779 690	14 678
R10	SMF	3 846 130	962 076 715	3 654 800	176
R11	SMF	4 079 546	1 039 169 432	3 841 750	998
R12	ultra	4 078 152	1 015 816 890	3 871 162	170
R13	ultra	3 441 666	924 822 940	3 237 524	136

Taula 5.1 Resum de les anàlisis de pre-ensamblat de les mostres d'aigua residual. A la primera columna tenim el codi de la mostra i en la segona és una breu descripció d'aquesta. La tercera columna conté el número de seqüències *pair-ends* obtingudes (aquest valor dividit per dos fa referència al nombre total de *reads* aparellats obtinguts). La quarta columna indica el número total de nucleòtids seqüenciats. Les dues columnes següents mostren el número de seqüències *pair-ends* (PE Seqs) que han passat tots els filtres de pre-ensamblat i el número de seqüències *single-ends* (SE Seqs) que també han passat els filtres de pre-ensamblat però que han perdut a la seva parella, independentment de si són R1 o R2 (*forward* o *reverse*).

Identificador de la Mostra	Descripció de la Mostra	Núm. Inicial PE Seqs	Nucleòtids Totals	PE Seqs Filtrades	SE Seqs Filtrades
P1	Julivert 30C	6 677 622	1 847 021 184	2 882 080	733
P2	Julivert 20C	1 192 716	327 348 654	329 520	1 808
PN1	Julivert CN	8 067 728	2 199 101 314	5 301 560	866
RI1	Riu	5 426 788	1 450 732 550	3 955 116	663
RI2	Riu	2 476 054	577 865 642	2 445 144	158

Taula 5.2 Resum de les anàlisis de pre-ensamblat de les mostres de julivert i riu. A la primera columna tenim el codi de la mostra i en la segona és una breu descripció d'aquesta. La tercera columna conté el número de seqüències *pair-ends* obtingudes (aquest valor dividit per dos fa referència al nombre total de *reads* aparellats obtinguts). La quarta columna indica el número total de nucleòtids seqüenciats. Les dues columnes següents mostren el número de seqüències *pair-ends* (PE Seqs) que han passat tots els filtres de pre-ensamblat i el número de seqüències *single-ends* (SE Seqs) que també han passat els filtres de pre-ensamblat però que han perdut a la seva parella, independentment de si són R1 o R2 (*forward* o *reverse*).

Identificador de la Mostra	Descripció de la Mostra	Núm. Inicial PE Seqs	Nucleòtids Totals	PE Seqs Filtrades	SE Seqs Filtrades
SH1	Homes	5 255 854	1 293 579 184	3 614 220	6 928
SH2	Homes	2 669 124	585 112 642	1 769 992	4 738
SH3	Dones	12 029 238	2 782 049 793	7 470 502	18 074
SH4	HEV	625 634	173 695 698	473 882	5 232
SH5	AI+IMSP	6 000 606	1 529 991 937	3 887 728	197 320
SH6	HEV	8 145 076	2 168 472 837	5 740 372	2 012
SH7	Sans A.1	3 413 928	849 803 466	1 873 370	286
SH8	Sans A.2	3 588 692	902 903 493	1 796 830	298
SH9	Sans B.1	3 457 150	849 330 701	2 119 224	250
SH10	Sans B.2	3 494 586	894 462 829	1 934 704	4

Taula 5.3 Resum de les anàlisis de pre-ensamblat de les mostres de sèrum humà. A la primera columna tenim el codi de la mostra i en la segona és una breu descripció d'aquesta. La tercera columna conté el número de seqüències *pair-ends* obtingudes (aquest valor dividit per dos fa referència al nombre total de *reads* aparellats obtinguts). La quarta columna indica el número total de nucleòtids seqüenciats. Les dues columnes següents mostren el número de seqüències *pair-ends* (PE Seqs) que han passat tots els filtres de pre-ensamblat i el número de seqüències *single-ends* (SE Seqs) que també han passat els filtres de pre-ensamblat però que han perdut a la seva parella, independentment de si són R1 o R2 (*forward* o *reverse*).

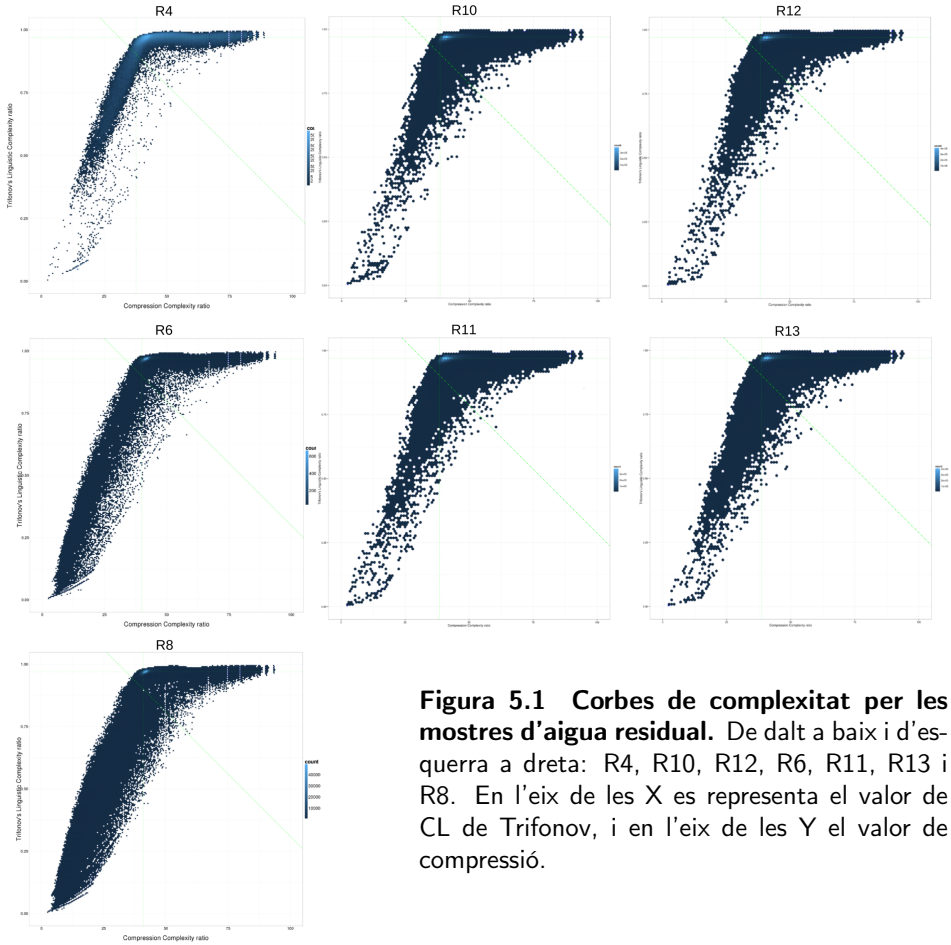


Figura 5.1 Corbes de complexitat per les mostres d'aigua residual. De dalt a baix i d'esquerra a dreta: R4, R10, R12, R6, R11, R13 i R8. En l'eix de les X es representa el valor de CL de Trifonov, i en l'eix de les Y el valor de compressió.

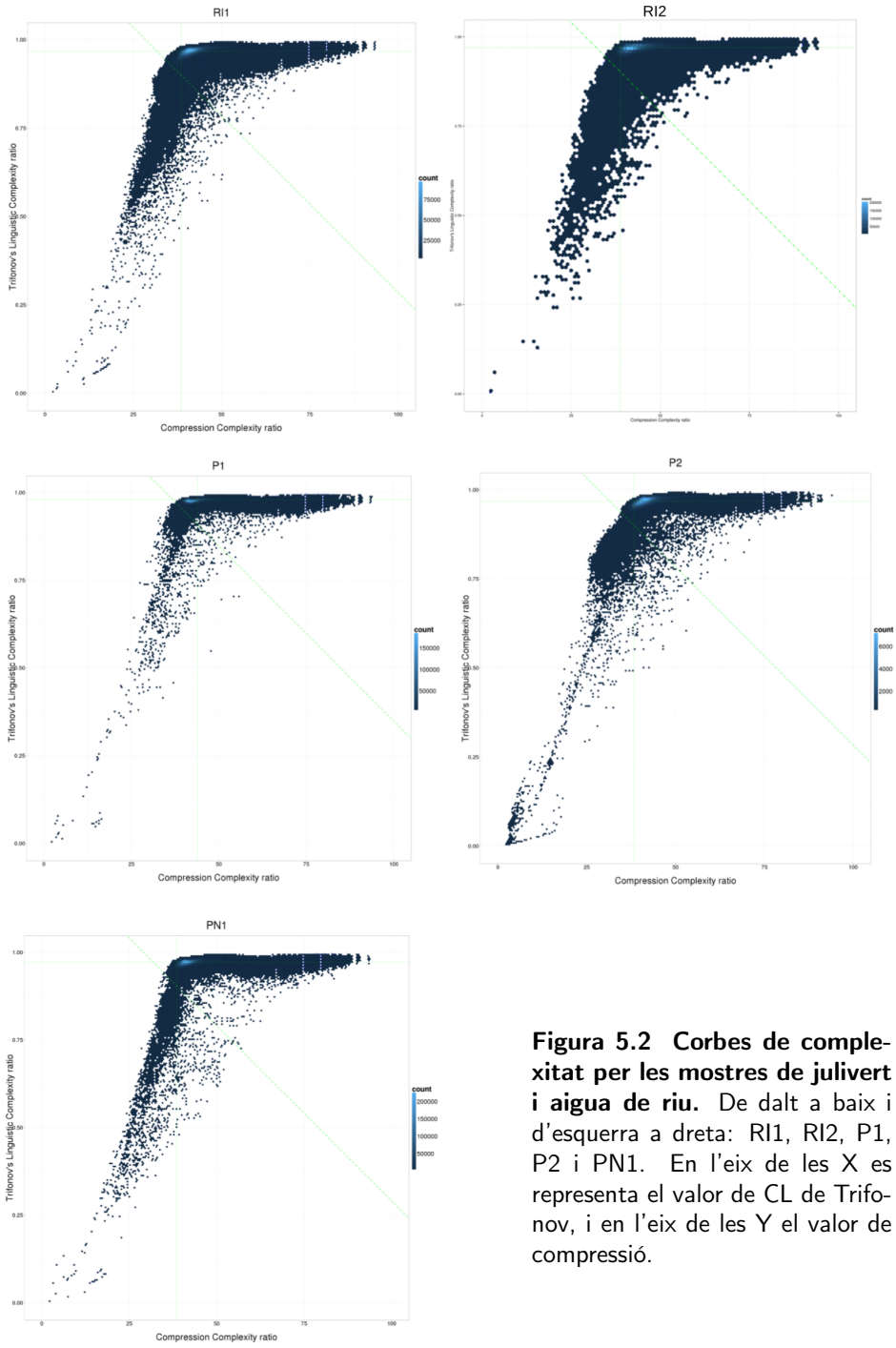


Figura 5.2 Corbes de complexitat per les mostres de julivert i aigua de riu. De dalt a baix i d'esquerra a dreta: RI1, RI2, P1, P2 i PN1. En l'eix de les X es representa el valor de CL de Trifonov, i en l'eix de les Y el valor de compressió.

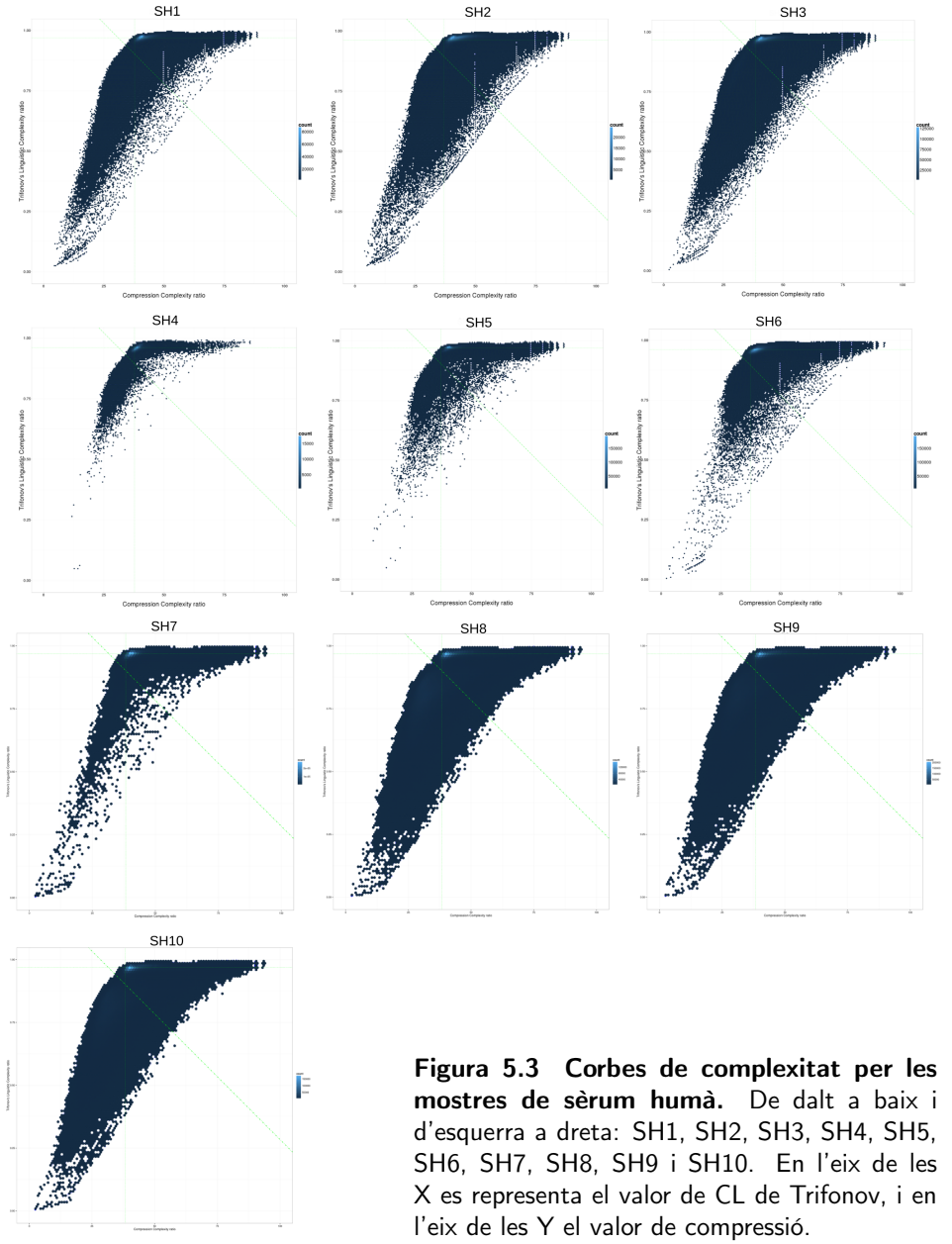


Figura 5.3 Corbes de complexitat per les mostres de sèrum humà. De dalt a baix i d'esquerra a dreta: SH1, SH2, SH3, SH4, SH5, SH6, SH7, SH8, SH9 i SH10. En l'eix de les X es representa el valor de CL de Trifonov, i en l'eix de les Y el valor de compressió.

5.2 Ensamblat de Metagenomes

En el moment de construir els ensamblats es van utilitzar dos programes diferents: `CLCBio` i `Velvet-MetaVelvet`. En el cas de les mostres de julivert i riu es presenten i discuteixen els resultats obtinguts amb el programa `Velvet-MetaVelvet`. En les mostres de sèrum humà i aigua residual es mostren i descriuen els resultats obtinguts amb l'ensamblador `CLCBio`. En les Taules 5.4, 5.5, 5.6 (de les pàgines 76, 77, 78) on es poden apreciar els següents paràmetres per cada mostra: el número de *contigs* i *singletons* obtinguts; la llargada del *contig* més llarg creat; la mitjana de la llargada dels *contigs*; i el valor del N50 de la llargada dels *contigs*. El valor N50 és una estimació de la qualitat de l'ensamblat que ens diu quina és la longitud del *contig* que ens trobem en el 50% de la llargada estimada del genoma, al sumar les llargades de tots els *contigs* ensamblats (ordenats abans de més llarg a més curt).

Una conclusió que hom pot extreure en mirar aquestes dades, és que amb el programa `CLCBio` s'obtenen *contigs* més llargs. En el cas de les mostres d'aigua residual, la longitud del *contig* més llarg és de 31 725bp i la llargada mitjana dels *contigs* és de 350bp. En les mostres de sèrum, on també s'ha utilitzat el programa `CLCBio`, les longituds dels *contigs* més llargs són més curts que en el cas anterior de les aigües residuals, i s'observa un rang de valors d'entre 2 499–5 560bp; però la mitjana de la longitud és de 370bp, lleugerament superior a les analitzades en les mostres d'aigua residual. Emprant el programa `Velvet-MetaVelvet` per les mostres de julivert i aigua de riu, el rang de longitud dels *contigs* més llargs és de 2 636–8 923bp; mentre que la mitjana de la longitud dels *contigs* és de 260bp (aquestes dades es resumeixen en les Figures 5.5, 5.6 i 5.4, de les pàgines 80, 81 i 79).

Mostra	Descripció	Núm. de *Contigs*	Núm. de *Singletons*	*Contig* més llarg (bp)	Llargada mitjana dels *Contigs* (bp)	N50 (bp)
R4	hivern	83 758	714 161	31 725	370,10	391
R6	primavera	138 906	501 137	6 432	263,92	251
R8	estiu	92 208	113 903	5 839	268,68	251
R10	SMF	157 402	3 372 435	9 077	382,89	391
R11	SMF	142 298	3 509 087	26 525	379,00	386
R12	ultra	212 737	3 466 437	14 069	382,16	387
R13	ultra	282 605	2 663 042	24 961	395,07	396

Taula 5.4 Resum de les anàlisis d'ensamblat de les mostres d'aigua residual. A la primera columna tenim el codi de la mostra i la segona és una breu descripció d'aquesta. La tercera conté el número de *contigs* generats per l'ensamblador, la quarta ens mostra el número de *singletons* (seqüències que no s'han pogut concatenar per construir cap *contig*). La cinquena llista la llargada del *contig* més llarg que s'ha obtingut en parells de bases (bp). La sisena presenta la mitjana de la llargada dels *contigs*. La darrera columna ens informa del valor N50 com a estimador de la qualitat de l'ensamblat.

Mostra	Descripció	Núm. de *Contigs*	Núm. de *Singletons*	*Contig* més llarg (bp)	Llargada mitjana dels *Contigs* (bp)	N50 (bp)
P1	Julivert 30C	15 833	2 807 447	4 086	276,02	328
P2	Julivert 20C	20 961	191 276	8 923	280,09	324
PN1	Julivert CN	10 296	5 169 800	3 215	306,94	372
RI1	Riu	243 043	25 556 829	3 665	242,34	277
RI2	Riu	192 139	1 664 874	2 636	230,51	258

Taula 5.5 Resum de les anàlisis d'ensamblat de les mostres de julivert i riu. A la primera columna tenim el codi de la mostra i la segona és una breu descripció d'aquesta. La tercera conté el número de *contigs* generats per l'ensamblador, la quarta ens mostra el número de *singletons* (seqüències que no s'han pogut concatenar per construir cap *contig*). La cinquena llista la llargada del *contig* més llarg que s'ha obtingut en parells de bases (bp). La sisena presenta la mitjana de la llargada dels *contigs*. La darrera columna ens informa del valor N50 com a estimador de la qualitat de l'ensamblat.

Mostra	Descripció	Núm. de *Contigs*	*Contig* més llarg (bp)	Llargada mitjana dels *Contigs* (bp)	N50 (bp)
SH1	Homes	43 188	2 971	384,89	408
SH2	Homes	19 000	2 660	352,13	356
SH3	Dones	83 518	3 102	374,19	389
SH4	HEV	4 273	2 931	294,09	290
SH5	AI+IMSP	5 889	2 902	396,64	447
SH6	HEV	13 060	2 996	487,81	543
SH7	Sans A.1	189 820	3 269	328,20	331
SH8	Sans A.2	166 167	5 560	328,67	331
SH9	Sans B.1	227 469	3 376	331,75	334
SH10	Sans B.2	185 359	2 499	342,30	344

Taula 5.6 Resum de les anàlisis d'ensamblat de les mostres de sèrum humà. A la primera columna tenim el codi de la mostra i la segona és una breu descripció d'aquesta. La tercera conté el número de *contigs* generats per l'ensamblador, la quarta ens mostra el número de *singletons* (seqüències que no s'han pogut concatenar per construir cap *contig*). La cinquena llista la llargada del *contig* més llarg que s'ha obtingut en parells de bases (bp). La sisena presenta la mitjana de la llargada dels *contigs*. La darrera columna ens informa del valor N50 com a estimador de la qualitat de l'ensamblat.

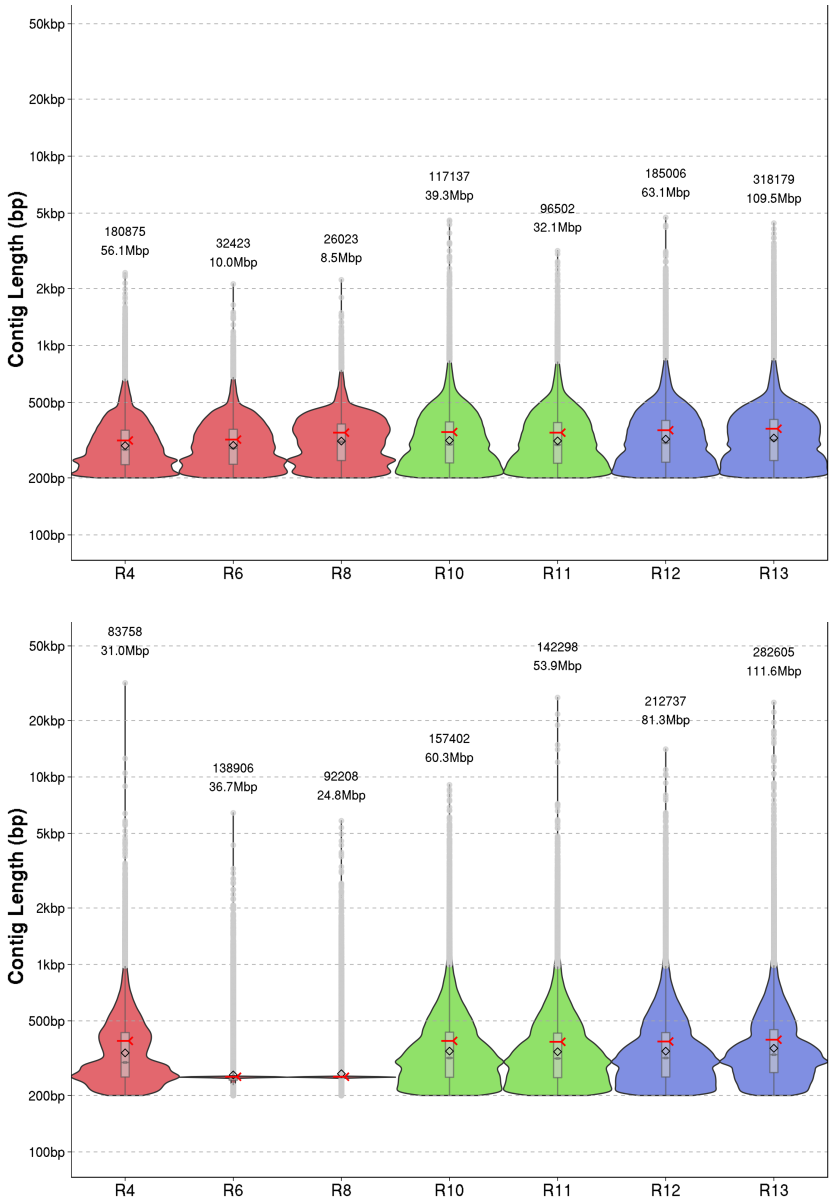


Figura 5.4 Distribució dels *contigs* obtinguts amb els dos ensambladors (`Velvet-MetaVelvet` i `CLCBio`) en les mostres d'aigua residual. A l'eix de les X estan representades cadascuna de les mostres i a l'eix de Y la longitud de les seqüències en escala $\log_{10}$. El símbol ≺ representa el valor de N50, el rombe la mitjana. Els punts grisos del *boxplot* són *outliers*. Els números superiors fan referència al nombre total de *contigs* obtinguts i a la longitud total del metagenoma ensamblat (en bp). El gràfic superior correspon als resultats obtinguts amb l'ensamblador `Velvet-MetaVelvet` i l'inferior als resultats de `CLCBio`.

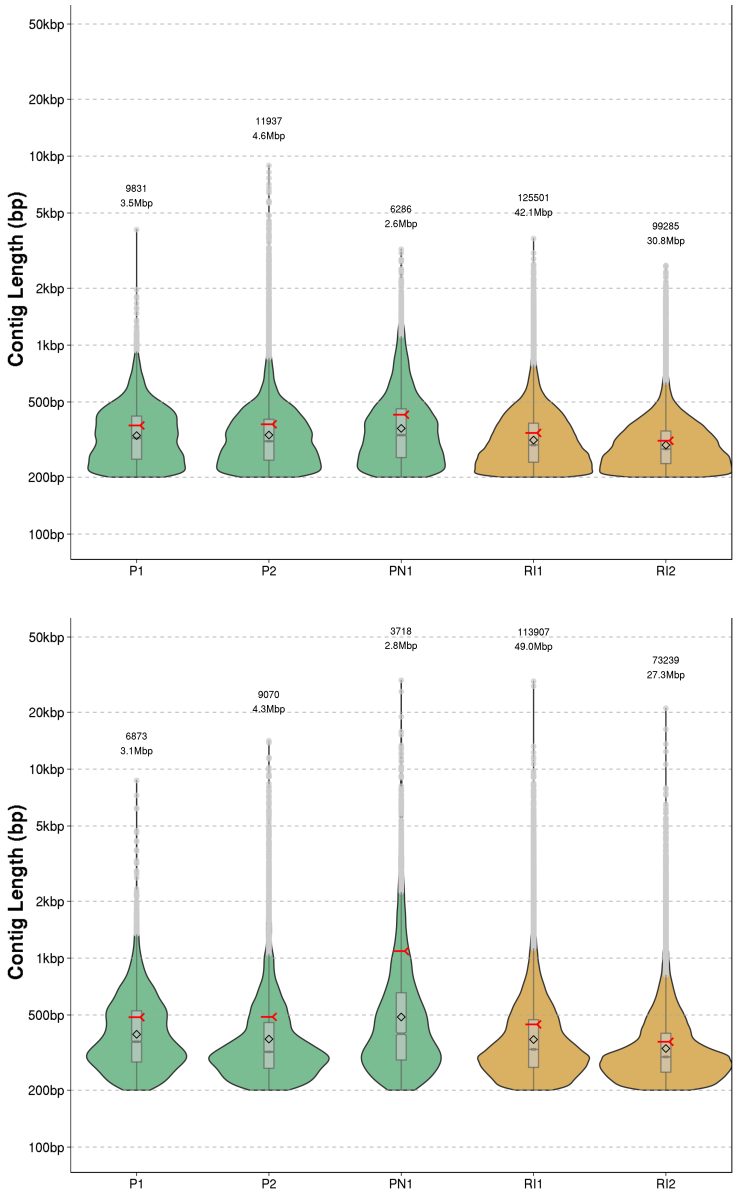


Figura 5.5 Distribució dels *contigs* obtinguts amb els dos ensambladors (`Velvet-MetaVelvet` i `CLCBio`) en les mostres de julivert i agua de riu. A l'eix de les X estan representades cadascuna de les mostres i a l'eix de Y la longitud de les seqüències en escala $\log_{10}$. El símbol ≺ representa el valor de N50, el rombe la mitjana. Els punts grisos del *boxplot* són *outliers*. Els números superiors fan referència al nombre total de *contigs* obtinguts i a la longitud total del metagenoma ensamblat (en bp). El gràfic superior correspon als resultats obtinguts amb l'ensamblador `Velvet-MetaVelvet` i l'inferior als resultats de `CLCBio`.

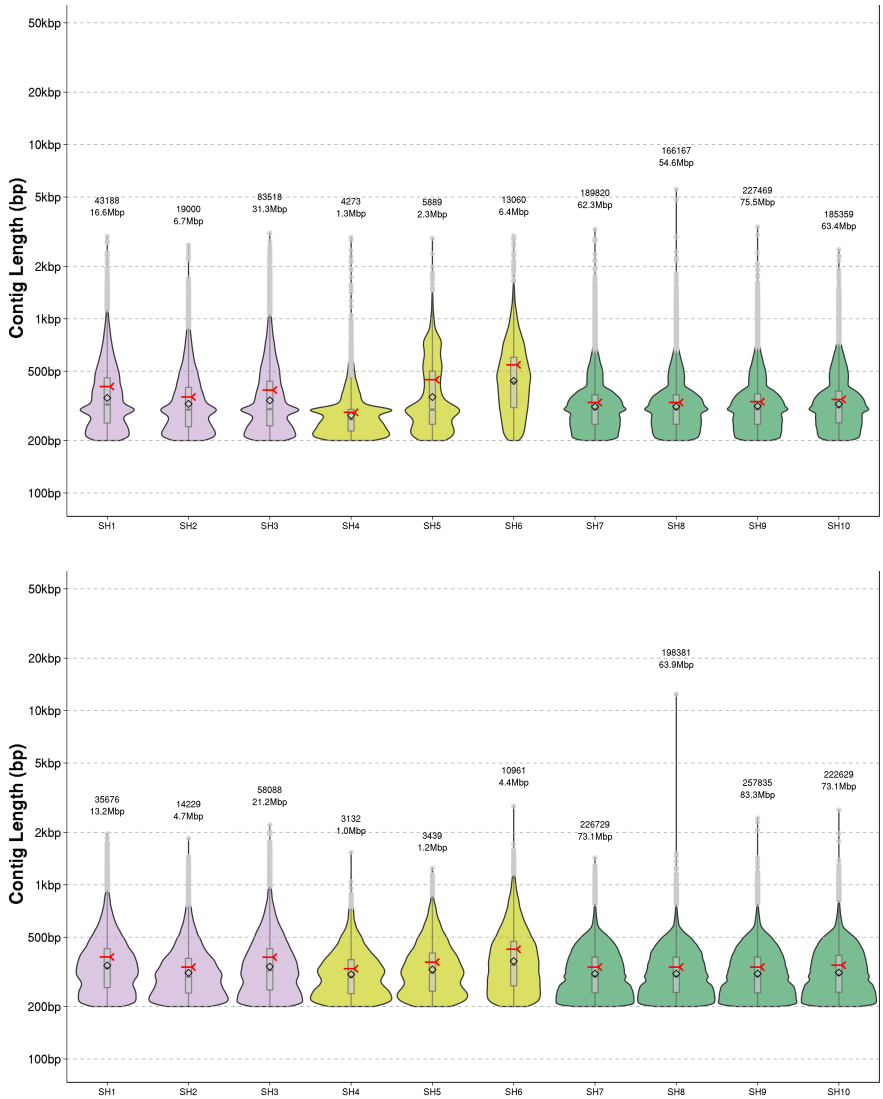


Figura 5.6 Distribució dels *contigs* obtinguts amb els dos ensambladors (`Velvet-MetaVelvet` i `CLCBio`) en les mostres de sèrum humà. A l'eix de les X estan representades cadascuna de les mostres i a l'eix de Y la longitud de les seqüències en escala $\log_{10}$. El símbol ≺ representa el valor de N50, el rombe la mitjana. Els punts grisos del *boxplot* són *outliers*. Els números superiors fan referència al nombre total de *contigs* obtinguts i a la longitud total del metagenoma ensamblat (en bp). El gràfic superior correspon als resultats obtinguts amb l'ensamblador `Velvet-MetaVelvet` i l'inferior als resultats de `CLCBio`.

5.3 Anàlisi post-ensamblat

Després de l'ensamblat s'han de dur a terme diferents cerques d'homologia mitjançant el programa `BLAST` tal i com s'explica en el capítol de "Material i mètodes". Els resultats contenen les llistes de seqüències homòlogues, sobre els que s'apliquen una sèrie de filtres. Tot seguit es procedeix a classificar taxonòmicament totes les seqüències de cada mostra per determinar l'espècie, família i el grup de Baltimore. En les Taules 5.7, 5.8 i 5.9, de les pàgines 85, 86 i 87 respectivament, es pot veure, per cadascuna de les mostres, el valor calculat i la desviació estàndard de la riquesa, així com el número de seqüències assignades putativament a un virus i el nombre total de famílies i espècies diferents trobades.

Per facilitar la visualització de quantes espècies diferents han estat assignades a cadascuna de les famílies, aquesta informació s'ha integrat en *heatmaps* on cada línia representa una mostra i cada columna una família en concret. El número de cada casella de les matrius dels *heatmaps* correspon al total d'espècies diferents que hi ha en aquella família. Els números anotats als marges del gràfic corresponen als subtotals; al sumatori de les espècies diferents trobades en una mostra en la part superior, i el sumatori de les espècies trobades en cada família a la dreta. Tota aquesta informació està disponible per als tres grups de mostres en les Figures 5.7, 5.8, 5.9; de les pàgines 88, 89, 90 respectivament.

En el moment d'analitzar els resultats de diversitat obtinguts pels conjunts de les diferents mostres, podem veure, almenys qualitativament, que hi ha diferències entre els conjunts que podrien ser degudes al programa d'ensamblat utilitzat en passos anteriors. Si ens fixem en els valors de famílies diferents detectades en cadascuna de les mostres podem observar que en les mostres d'aigua residual tenim un rang entre 30 i 36 famílies diferents, en les mostres de julivert entre 20 i 25 (16 en el control negatiu); a l'aigua de riu entre 22 i 26, i finalment en les mostres de sèrum entre 20 i 31. Es podria dir que els valors són bastant similars entre tots els conjunts de mostres, destacant que els valors de l'aigua residual són més alts. En canvi, en el nombre de espècies sí que s'observa una diferència més gran entre els diferents ensambladors. En l'aigua residual es detecten entre 718 i 1 289 espècies diferents, en sèrum humà entre 492 i 13 330 famílies. En les mostres de julivert i aigua de riu tan sols caracteritzem entre 66 i 464 espècies diferents; en aquest cas es va utilitzar `Velvet-MetaVelvet` i en les altres mostres `CLCBio`.

Si ens centrem en les dades de les mostres de julivert i agua de riu, podem comentar que la riquesa de la mostra P2, en la qual es va fer un amplificació de 20 cicles per PCR, presenta un valor del doble que per la mostra P1, on

l'amplificació per PCR va ser de 30 cicles. Aquest fet es veu directament reflectit en el nombre d'espècies i famílies detectades. En les mostres de riu el valor de la riquesa i nombre de famílies i especies és molt més elevat, ja que és una matriu molt més complexa. En aquest cas s'han pogut detectar famílies que infecten a animals i humans, com ara *Adenoviridae*, *Astroviridae*, *Caliciviridae*, *Picobirnaviridae*, *Reoviridae*, *Picornaviridae*, *Parvoviridae* i *Circoviridae*. En la mostra de julivert regada amb aigua control no s'han detectat famílies que puguin ser patògens humans; les famílies més abundants han estat *Podoviridae*, *Siphoviridae* i *Dicistroviridae*. En la mostra de julivert regada amb aigua de riu s'han pogut detectar diferents famílies amb virus patogènics com *Astroviridae*, *Caliciviridae*, *Flaviviridae*, *Hepeviridae*, *Parvoviridae*, i *Picornaviridae*.

En les mostres de sèrum humà les famílies majoritàries detectades pertanyen a la família *Anelloviridae*, a totes les mostres, encara que s'observa major abundància en les mostres de malalts que en les dels individus sans. Entre els grups taxonòmics caracteritzats trobem els següents: *Astroviridae* en una mostra de homes (SH1), les dones (SH3) i els AI+IMSP (SH5); *Caliciviridae* en totes les mostres menys a les dels individus; *Hepeviridae* únicament a les mostres de HEV (SH6) i Ai+ImSP (SH5); de la família *Flaviviridae* el virus GBvirusC, en les mostres de dones, dos mostres d'individus sans, i Ai+ImSP (SH3, SH5, SH9 i SH10); finalment, seqüències de retrovirus endògens en totes les mostres menys en una mostra d'homes (SH1) i els AI+IMSP (SH5).

Focalitzant en el grup de mostres d'aigua residual, les famílies més abundants que infecten a humans i animals són *Parvoviridae* i *Picornaviridae*. També es detecten altres famílies amb importants patògens com *Astroviridae*, *Caliciviridae*, *Hepeviridae*, *Circoviridae* i *Polyomaviridae*. En el cas de les famílies d'*Astroviridae* i *Caliciviridae*, s'ha detectat una abundància més gran en la mostra d'hivern, com a conseqüència de la seva estacionalitat.

Una de les diferències importants a nivell experimental entre les mostres d'aigua residual mostrejades en diferents estacions (hivern, primavera i estiu) respecte les altres quatre mostres d'aigua residual, és el volum de la mostra analitzada i el protocol de concentració de virus. En dues d'elles es van analitzar 500ml d'aigua residual i floculació amb llet descremada (R10 i R11); en les altres dues mostres es va utilitzar un protocol d'ultracentrifugació (R12 i R13). Altres diferències inclouen els cicles de PCR utilitzats per l'amplificació; en les mostres R10, R11, R12 i R13 es va utilitzar un volum molt inferior i 25 cicles de PCR enlloc de 30. Als *heatmaps* es pot observar que hi ha famílies com *Adenoviridae*, *Polyomaviridae* i *Papillomaviridaeque*, que només són detectades en aquestes últimes mostres.

Comparant les mostres obtingudes amb els dos mètodes de concentració, la

ultracentrifugació i SMF, s'ha vist que hi ha espècies virals que són patogèniques per als humans que només són detectades per ultracentrifugació, com *Anelloviridae*, *Alloherpesviridae*, *Geminiviridae*, *Hepeviridae*, *Papillomaviridae*, *Totiviridae*, *Geminiviridae* i *Polyomaviridae*. Per acabar, voldria destacar la diferència a nivell de diversitat d'espècies en famílies de fags com *Luteoviridae*, *Nanoviridae* i *Baculoviridae*, que ha influït molt en el valor de la riquesa d'algunes de les mostres.

Mostra	Descripció	Riquesa (*Richness*)	Desviació estàndard	Núm. de seqüències	Núm. de famílies	Núm. d'espècies
R4	hivern	831,2	27,2	20 231	31	775
R6	primavera	843,5	37,9	11 371	30	735
R8	estiu	865,6	43,8	7 469	30	718
R10	SMF	755,8	23,6	21 717	32	778
R11	SMF	541,0	18,9	12 761	30	572
R12	ultra	1 066,4	23,2	29 393	34	1 094
R13	ultra	1 318,5	27,6	33 432	36	1 289

Taula 5.7 Resum dels resultats de diversitat calculats sobre l'ensamblat per les mostres d'aigua residual. A la primera columna tenim el codi de la mostra i a la segona columna una breu descripció de la mateixa. La tercera columna conté el valor de riquesa; la quarta es correspon amb la desviació estàndard de la riquesa de la mostra. La cinquena columna presenta el número de seqüències que han passat tots els filtres post-ensamblat i han estat anotades a nivell taxonòmic com a virus. La sisena i setena columna ens llisten el número de famílies i d'espècies diferents presents a la mostra, respectivament.

Mostra	Descripció	Riquesa (*Richness*)	Desviació estàndard	Núm. de seqüències	Núm. de famílies	Núm. d'espècies
P1	Julivert 30C	171,8	40,9	5 558	20	66
P2	Julivert 20C	255,6	34,7	59 380	25	139
PN1	Julivert CN	130,3	34,7	3 878 957	16	60
RI1	Riu	632,4	44,4	77 626	22	361
RI2	Riu	923,5	70,9	44 607	26	464

Taula 5.8 Resum dels resultats de diversitat calculats sobre l'ensamblat per les mostres de julivert i riu. Valoració comparativa d'amplificació de 20 i 30 cicles de PCR en les mostres de julivert (P1 i P2). A la primera columna tenim el codi de la mostra i a la segona columna una breu descripció de la mateixa. La tercera columna conté el valor de riquesa; la quarta es correspon amb la desviació estàndard de la riquesa de la mostra. La cinquena columna presenta el número de seqüències que han passat tots els filtres post-ensamblat i han estat anotades a nivell taxonòmic com a virus. La sisena i setena columna ens llisten el número de famílies i d'espècies diferents presents a la mostra, respectivament.

Mostra	Descripció	Riquesa (*Richness*)	Desviació estàndard	Núm. de seqüències	Núm. de famílies	Núm. d'espècies
SH1	Homes	133,6	30,2	2 708	31	2 115
SH2	Homes	61,9	19,8	1 057	25	848
SH3	Dones	109,7	24,3	4 079	29	3 518
SH4	HEV	50,1	16,4	565	21	492
SH5	AI+IMSP	41,0	3,7	2 109	21	2 000
SH6	HEV	243,0	91,7	1 344	22	1 175
SH7	Sans A.1	26,3	13,2	14 865	20	13 330
SH8	Sans A.2	48,0	24,0	8 313	20	7 293
SH9	Sans B.1	49,0	24,1	10 430	24	9 338
SH10	Sans B.2	15,6	2,2	9 386	22	8 329

Taula 5.9 Resum dels resultats de diversitat calculats sobre l'ensamblat per les mostres de sèrum humà. A la primera columna tenim el codi de la mostra i a la segona columna una breu descripció de la mateixa. La tercera columna conté el valor de riquesa; la quarta es correspon amb la desviació estàndard de la riquesa de la mostra. La cinquena columna presenta el número de seqüències que han passat tots els filtres post-ensamblat i han estat anotades a nivell taxonòmic com a virus. La sisena i setena columna ens llisten el número de famílies i d'espècies diferents presents a la mostra, respectivament.

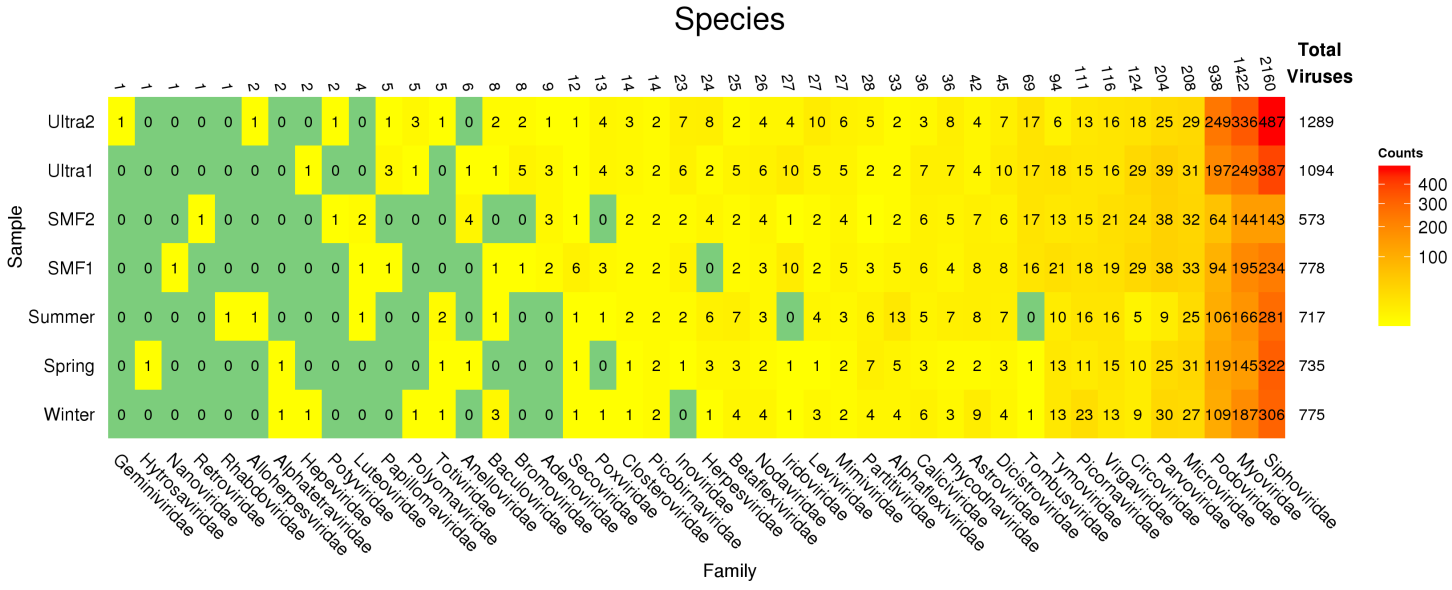


Figura 5.7 ***Heatmap*** **de les mostres d'aigua residual.** Es representa l'abundància relativa de les diferents famílies (columnes) detectades en cadascuna de les mostres (files). La xifra dins de cada cel·la representa el número de espècies diferents que s'han pogut assignar a cada família. El rang de colors escollit va des de verd, que representa l'absència d'espècies a la mostra, fins el vermell, que representa l'abundància relativa més alta. La llegenda de la dreta ens indica la relació entre els comptatges i la intensitat dels colors del gradient, la qual s'ajusta al màxim de cada comparació entre mostres (487 en aquest cas).

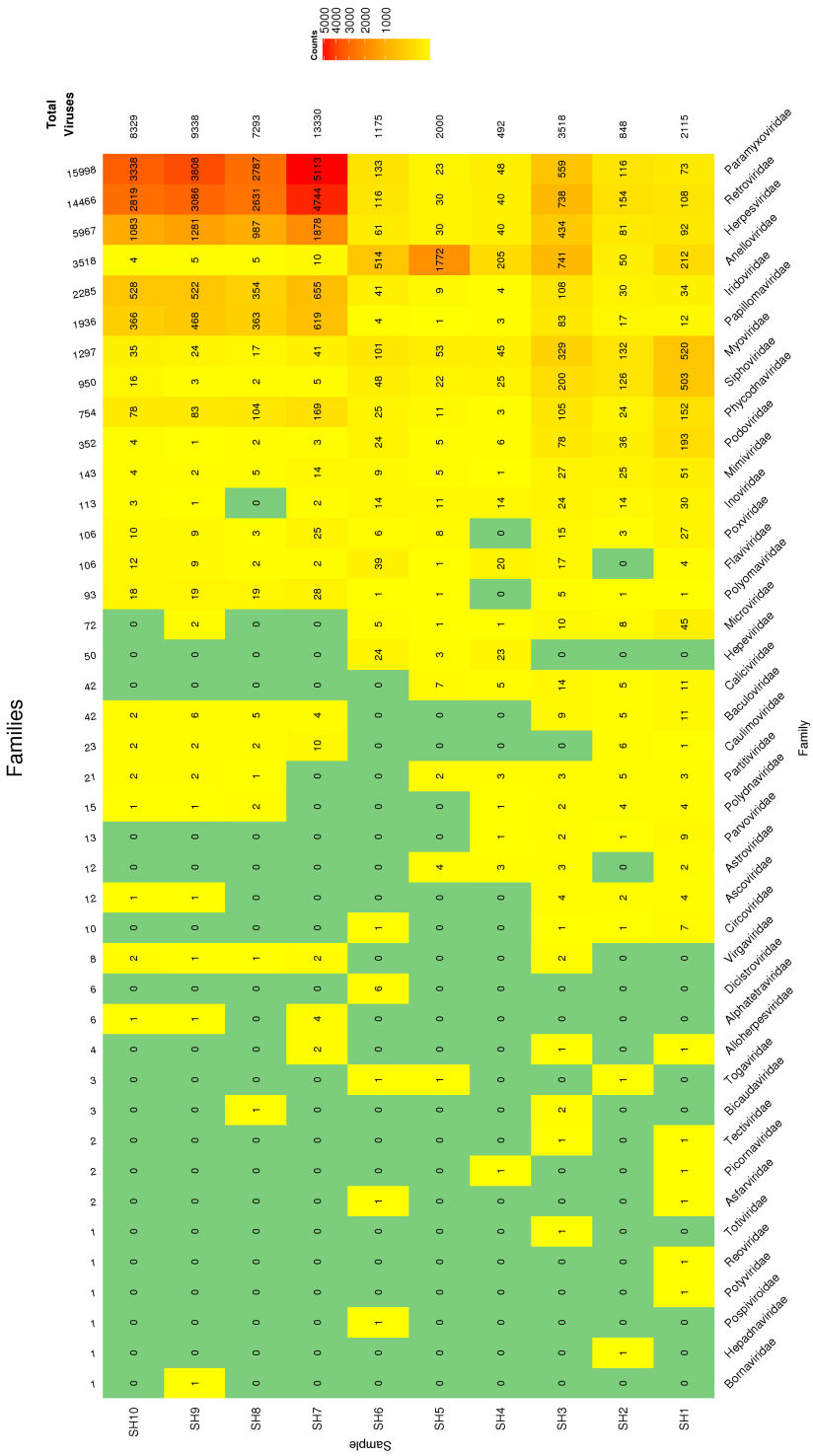


Figura 5.8 ***Heatmap*** **de les mostres de julivert i agua de riu.** Es representa l'abundància relativa de les diferents famílies (columnes) detectades en cadascuna de les mostres (files). La xifra dins de cada cel·la representa el número de espècies diferents que s'han pogut assignar a cada família. El rang de colors escollit va des de verd, que representa l'absència d'espècies a la mostra, fins el vermell, que representa l'abundància relativa més alta. La llegenda de la dreta ens indica la relació entre els comptatges i la intensitat dels colors del gradient, la qual s'ajusta al màxim de cada comparació entre mostres (169 en aquest cas).

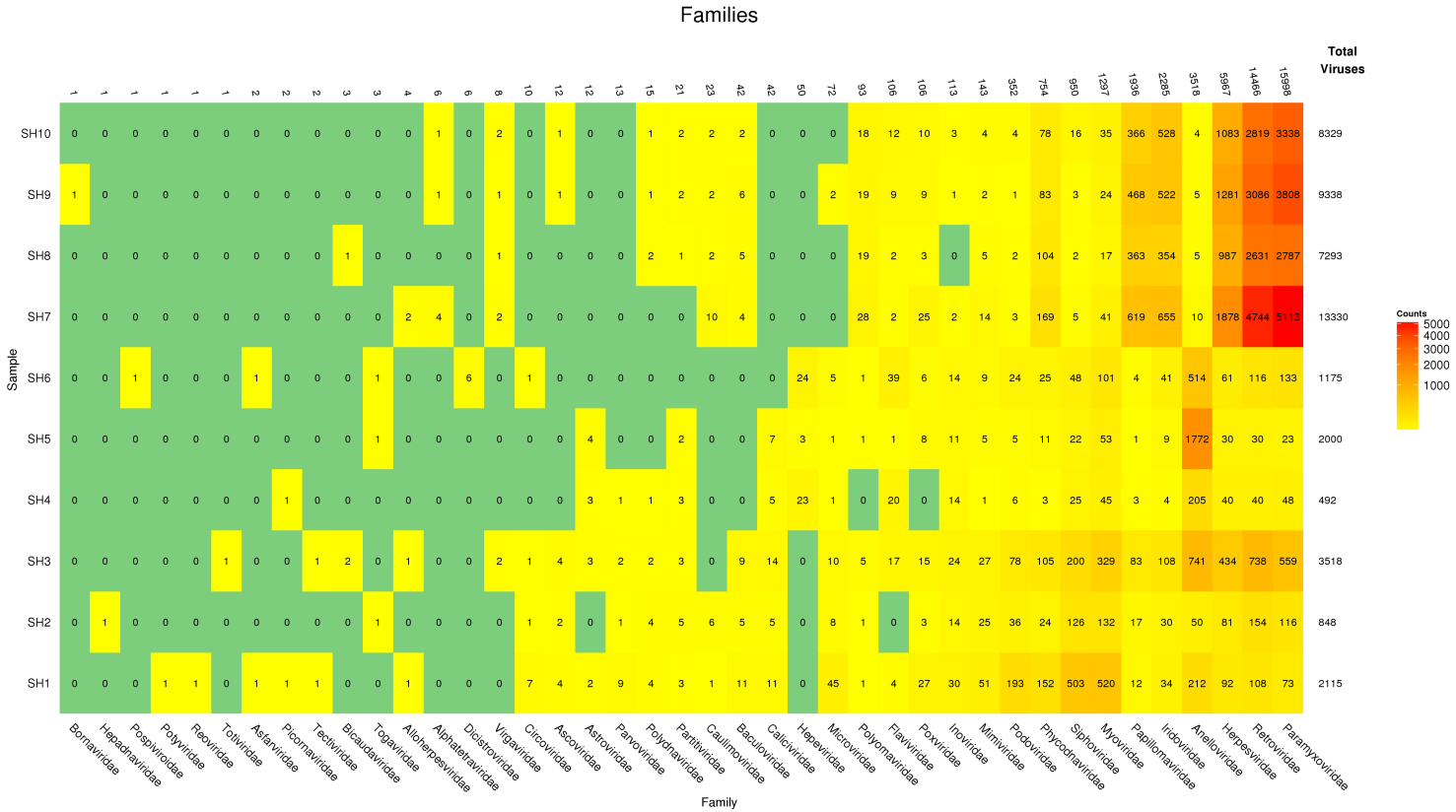


Figura 5.9 ***Heatmap*** **de les mostres de sèrum humà.** Es representa l'abundància relativa de les diferents famílies (columnes) detectades en cadascuna de les mostres (files). La xifra dins de cada cel·la representa el número de espècies diferents que s'han pogut assignar a cada família. El rang de colors escollit va des de verd, que representa l'absència d'espècies a la mostra, fins el vermell, que representa l'abundància relativa més alta. La llegenda de la dreta ens indica la relació entre els comptatges i la intensitat dels colors del gradient, la qual s'ajusta al màxim de cada comparació entre mostres (5 113 en aquest cas).

5.4 Anàlisis filogenètics per famílies específiques

En l'estudi de les mostres d'aigua residual i les mostres de sèrum humà es va poder aprofundir en l'anàlisi filogenètic de dues famílies específiques. En el cas de les mostres de sèrum humà es van estudiar els *Torque teno virus* de la família *Anelloviridae*, mentre que en el cas de les mostres d'aigua residual va ser la família *Adenoviridae*.

En paral·lel als estudis de metagenòmica *whole genome*, per les mostres d'aigua residual es va realitzar metagenòmica dirigida amb *primers* específics dissenyats sobre seqüències característiques de la família *Adenoviridae*. Després de completar les anàlisis filogenètiques explicades a l'apartat 4.4 de Material i Mètodes, es va inferir un arbre filogenètic de totes les espècies identificades (veure Figura 5.10 de la pàgina 92), sobre el que es van marcar aquelles famílies que són patogèniques pels humans.

En les mostres de sèrum humà una de les famílies més abundants ha estat *Anelloviridae*. En estudis previs s'havia detectat aquest grup de virus en pacients que presentaven un quadre d'hepatitis però no s'havia pogut relacionar cap agent causal. Es va reconstruir un arbre filogenètic amb totes les seqüències que van poder ser anotades per aquesta família juntament amb les seqüències de referència ja descrites; l'arbre es mostra en la Figura 5.11 de la pàgina 93. Malgrat que la regió genòmica que es va recuperar de l'ensamblat no tenia més de 2 000 nucleòtids, l'arbre filogenètic ens permet distingir tres grans grups de seqüències: les que estaven alineades amb un percentatge d'identitat molt elevat respecte seqüències ja descrites anteriorment dins d'aquest grup taxonòmic; les seqüències que havien estat assignades a seqüències classificades fora de la família; i finalment, seqüències amb molta poca identitat, les quals podrien correspondre a nous virus encara no descrits i que podrien jugar algun paper en l'etiologia relacionada amb aquests casos d'hepatitis.

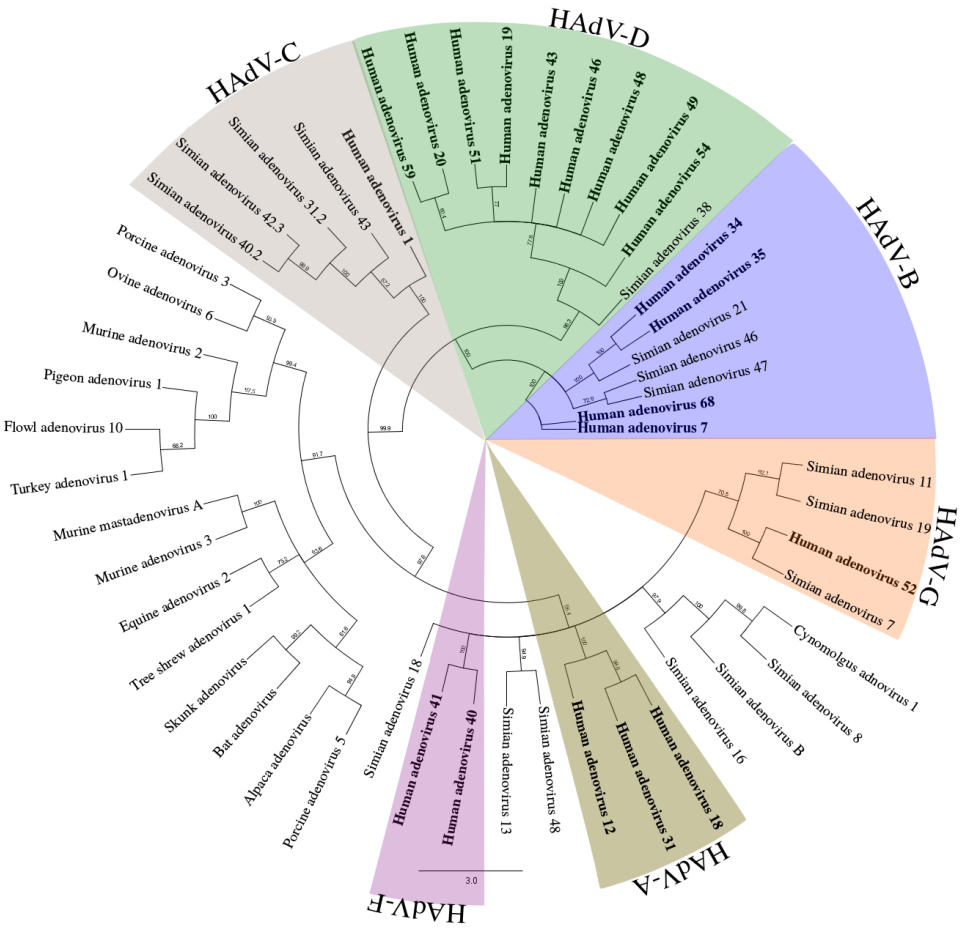


Figura 5.10 Arbre filogenètic de la família *Adenoviridae* a partir de dades de metagenòmica dirigida. Representació filogenètica de les diferents espècies identificades en les mostres d'aigua residual de la família d'*Adenoviridae*. En negreta es representen les espècies que s'han descrit com a patogèniques en humans. Els sectors de colors ressalten les subfamílies a les quals pertanyen aquestes espècies. Els nombres que apareixen sobre les branques corresponen als valors de *bootstrap*.

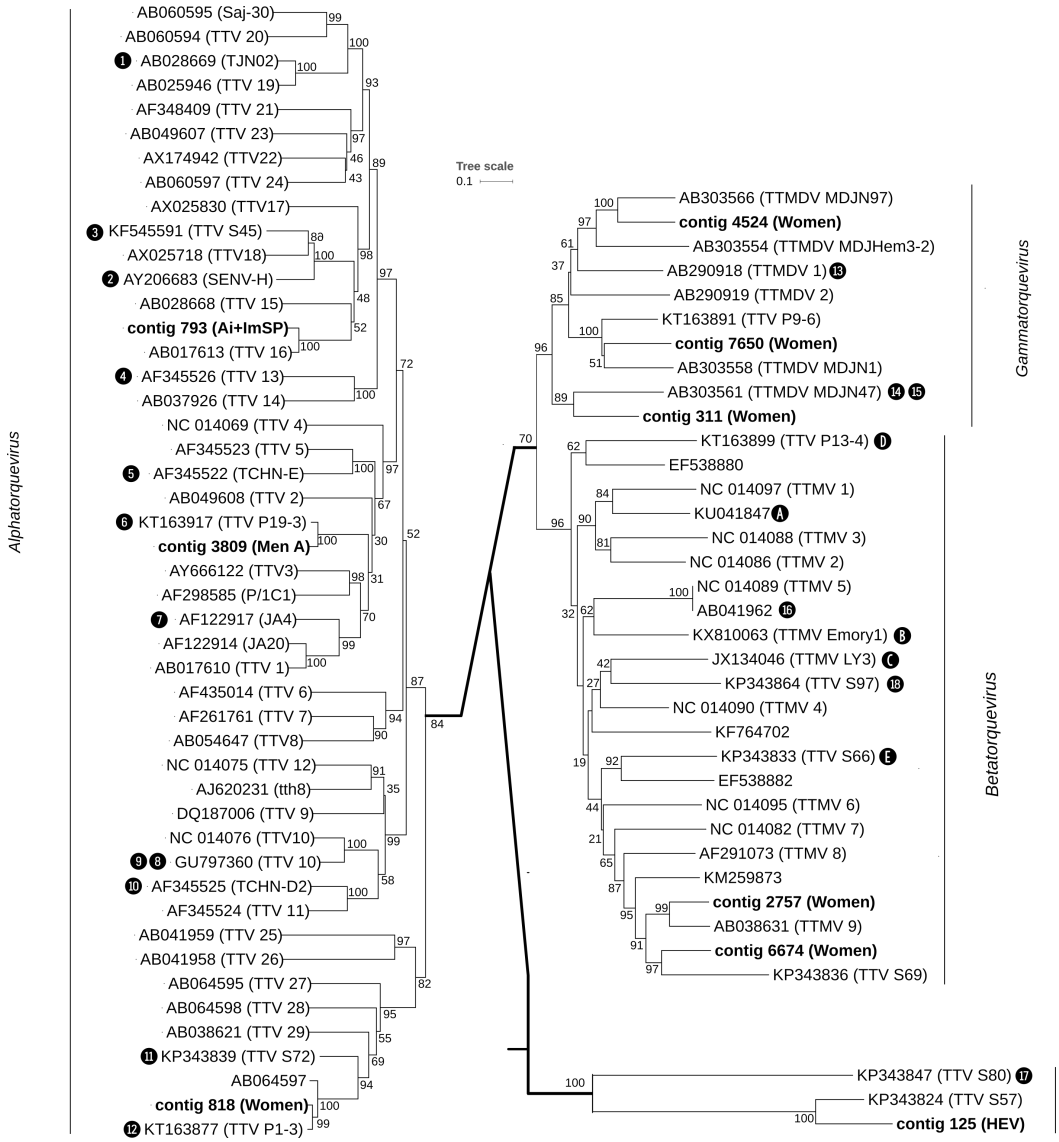


Figura 5.11 Arbre filogenètic de la família *Anelloviridae*. S'ha mantingut la topologia de l'arbre, però per poder-ho visualitzar en la pàgina les línies centrals més gruixudes no respecten l'escala de la resta de l'arbre. S'han destacat en negreta els *contigs* detectats a les mostres de sèrum humà que no tenen la identitat necessària per formar part d'una espècie ja descrita. Els números i lletres encerclats al costat de cadascun dels identificadors de seqüència es corresponen amb els *contigs* de les mostres que es podrien classificar dins la mateixa espècie, ja que presenten un percentatge d'identitat igual o superior al requerit pels criteris del ICTV. La Taula 5.10 de la pàgina 96 conté informació més detallada per les seqüències etiquetades amb els números i lletres encerclats.

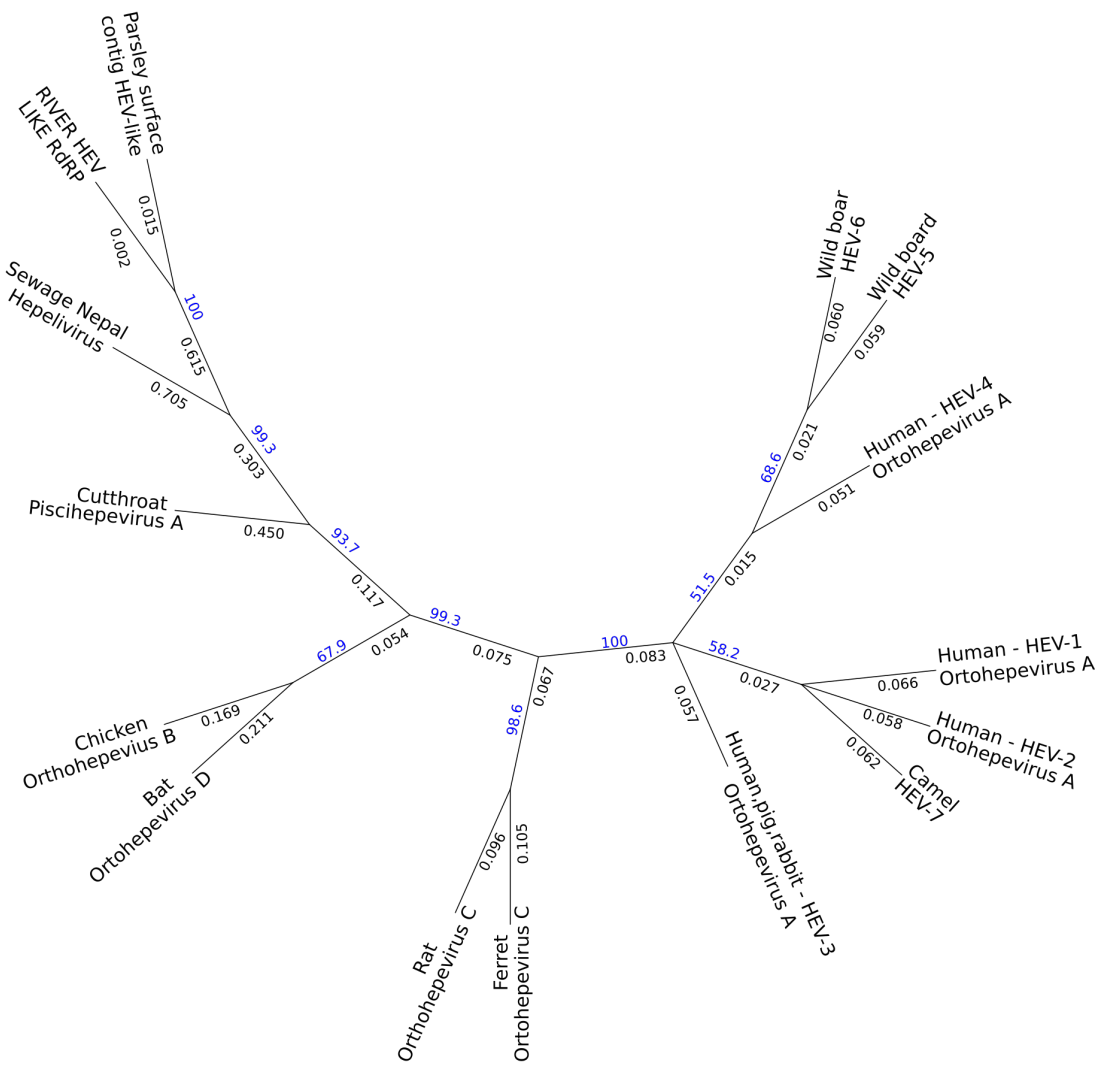


Figura 5.12 Arbre filogenètic de les seqüències relacionades amb Hepelivirus El resultat de l'anàlisis filogenètic de la regió *RdRp* de l'Hepilivirus i la família *Hepeveridae* inclou seqüències de les mostres d'aigua de riu i julivert. En l'arbre s'observa que les seqüències del julivert i aigua de riu s'agrupen més a prop de Hepilivirus que de la família *Hepeveridae*. L'arbre ha estat construït usant el programa `Geneious`, clusteritzant segons el mètode *Neighbor-joining*, amb els paràmetres del model Jukes-Cantor, amb un *bootstrap* de 1 000. Els números sobre les branques en blau ens indiquen els valors de *bootstrap* i en negre les distàncies filogenètiques.

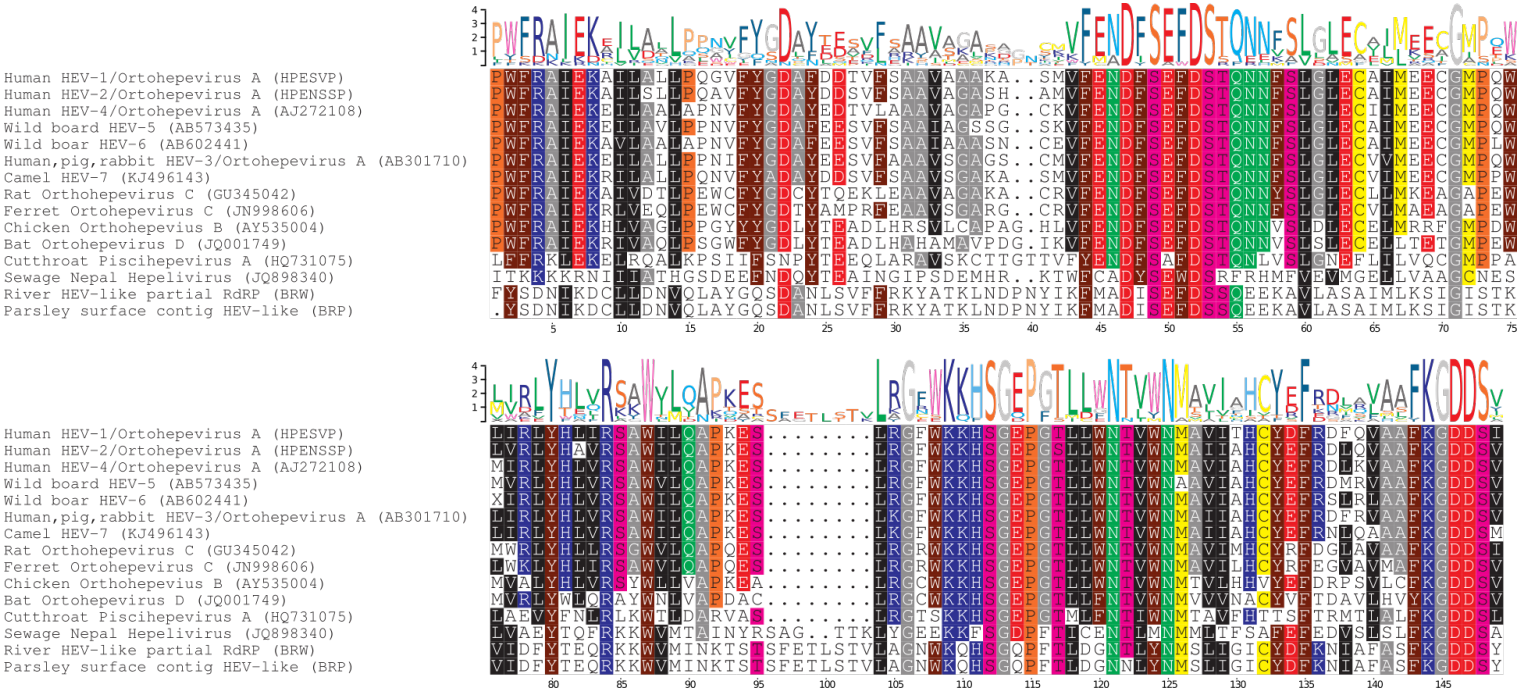


Figura 5.13 Alineament de la regió conservada per RdRp de les seqüències relacionades amb hepelivirus L'arbre filogenètic ha estat generat a partir de les posicions conservades d'aquest alineament obtingut a partir d'un fragment conservat de 204 aminoàcids de la proteïna RdRp. Cada color correspon a un grup aminoacídic conservat en l'alineament. El logo de seqüència de la part superior reflexa les freqüències dels aminoàcids a cada posició, ajustades a la quantitat d'informació (escala numèrica de l'esquerra del logo en bits). Figura generada amb el paquet TEXshade de LATEX.

Codi	Identificador	Mostra	Longitud (bp)	% Identitat
1	contig_5911	Female	1 500	88,9
2	contig_268	Female	1 951	90,1
3	contig_2837	HEV	1 845	86,6
4	contig_1475	Female	1 899	87,5
5	contig_236	HEV	1 643	72,0
6	contig_2366	HEV	1 514	93,8
7	contig_9035	Female	2 243	94,0
8	contig_6533	Female	1 530	83,3
9	contig_1709	AI+IMSP	1 832	85,7
10	contig_7929	Male A	1 875	79,2
11	contig_129	Female	1 506	79,8
12	contig_1	Female	1 548	92,4
13	contig_1946	Female	1 644	71,6
14	contig_16376	Female	1 513	68,3
15	contig_1013	AI+IMSP	1 687	66,0
16	contig_506	HEV	2 046	73,7
17	contig_1199	AI+IMSP	1 586	66,8
18	contig_2151	AI+IMSP	1 985	65,6
A	contig_626	Female	1 503	89,7
B	contig_118	HEV	1 781	68,6
C	contig_2	HEV	1 923	66,6
D	contig_1199	Male A	1 512	69,3
E	contig_66	HEV	1 904	71,9

Taula 5.10 Llista de seqüències de les mostres de sèrum humà utilitzades en l'arbre filogenètic de les seqüències relacionades amb hepelivirus. La primera columna correspon al codi emprat en l'arbre de la figura 5.11, la segona a l'identificador de la seqüència, la tercera és el nom de la mostra d'on prové la seqüència. La quarta és la llargada de la seqüència, la última columna és el valor d'identitat respecte l'alineament entre el *contig* i la seqüència de referència amb la que ha estat anotada.

6 Visualització i accessibilitat de les dades

Un cop tenim processada tota la informació de cada mostra, un pas vital perquè les dades siguin útils i de fàcil accés pel seu anàlisi i interpretació, consisteix en integrar-les en una interfície que permeti als usuaris accedir-hi de manera eficient i intuïtiva. Per fer això possible ens cal reestructurar les dades per emmagatzemar-les fent servir un gestor de bases de dades, el qual facilitarà i accelerarà les cerques sobre grans conjunts de dades. A sobre d'aquesta infraestructura de dades cal incorporar una capa de programes que ens permetin fer cerques i recuperar informació de manera interactiva, ja sigui des de la línia de comandes per usuaris avançats o des de formularis web per la resta.

6.1 Base de dades

Per cada*run* d'Illumina de mitjana s'han seqüenciat 10 mostres; cada mostra té dos fitxers de dades crues o en brut (*raw data*), un *forward* i un *reverse*, que es corresponen amb els *reads* R1 i R2. Cadascun d'aquests fitxers comprimit ocupa de mitjana uns 250MB, per tant cada mostra 500MB, i un *run* 5GB totals (suposant una taxa de compressió entre el 60$\tilde{7}$0%, les dades en brut ocuparien unes tres vegades més). Els fitxers resultants del protocol bioinformàtic complet ocupen aproximadament 3MB per cada mostra sense comprimir. Fins ara hem analitzat 11 *runs*, amb un total de 73 mostres diferents, que equival a 11TB, entre fitxers *raw*, fitxers intermedis i temporals i els fitxers finals d'ensamblat i de cerques d'homologia amb el `BLAST`. Com que el volum de mostres i dades és bastant elevat, es va decidir implementar una base de dades en MySQL (DuBois, 2005; Welling i Thomson, 2005) per tal d'emmagatzemar les dades finals de cadascuna de les mostres i posteriorment poder treballar d'una manera més ràpida i eficient en accedir-hi i visualitzar-les.

En la base de dades de MySQL, per a cada mostra, es va crear un nucli de tres taules: una amb les seqüències de nucleòtids, una altra amb els resultats de l'ensamblador `CLCBio` i una tercera amb els resultats del `Velvet-MetaVelvet`; l'esquema d'aquest nucli de la base de dades es mostra en la Figura 6.1 de la pàgina 98. En aquesta figura es poden veure les diferents taules i les relacions entre elles. Una taula auxiliar integra els codis de cada mostra i *run*, i serveix de punt d'entrada unificat a les dades estructurades.

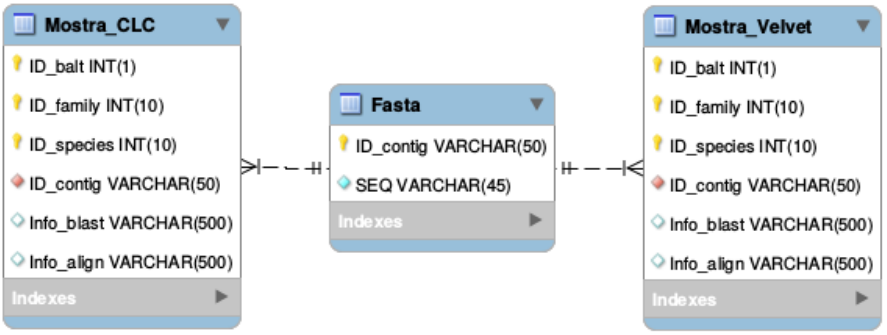


Figura 6.1 Esquema UML de les taules MySQL i les relacions indexades per cadascuna de les mostres analitzades. En la taula central `Fasta` "Id_contig" és el codi identificatiu únic per a cadascuna de les seqüències, aquest camp és el que permetrà relacionar amb les altres taules; "SEQ" fa referència a la seqüència de nucleòtids dels *contigs*. Els camps definits en les taules que emmagatzemen les dades dels ensamblats són els següents: "ID_balt" correspon al número d'identificació pel grup de Baltimore al que hem assignat la seqüència; "ID_family" és el número d'identificació per la família a la que hem assignat la seqüència; "ID_species" ens indica el número de identificació de l'espècie al que hem assignat la seqüència; "ID_contig" és el codi identificatiu únic per a cadascuna de les seqüències; "Info_blast" correspon al fitxer de text on hi ha la informació sobre els *hits* en les diferents bases de dades; "Info_align" fa referència al fitxer de text on hi ha l'alineament de la seqüència en les diferents bases de dades.

6.2 Visualització de dades

Un cop tenim totes les dades indexades a la base de dades, el següent pas és la visualització. S'han desenvolupat diversos mètodes per a que l'usuari pugui accedir específicament a la informació desitjada. Les aplicacions que s'han implementat estan basades en taules dinàmiques, gràfics *krona* i taules estàtiques. Per poder accedir a tots aquests resultats l'usuari ha de tenir accés a la pàgina web del grup de recerca en Genòmica Computacional (CompGen)[1]. Un cop allà es trobarà amb una taula estàtica on hi ha les llistes de totes les mostres analitzades, classificades segons el tipus de matriu (Figura 6.2 de la pàgina 100). Per cadascuna de les mostres es pot accedir a les diferents taules dinàmiques explicades amb més detall en els apartats següents. La taula estàtica es crea directament sobre un article de la Tiki[2], el motor amb el que treballa la web de CompGen, fent servir marques per tabular dades segons la sintaxi de la mateixa

[1] `https://compgen.bio.ub.edu`
[2] `https://tiki.org`

Tiki. Totes les captures de pantalla de les Figures 6.2, 6.3, 6.4, 6.5, 6.6, 6.7 i 6.8 han estat realitzades amb l'aplicació `shutter`[3]. La `TikiWiki` és un motor de gestió de pàgines web similar a Wikipedia, amb un conjunt definit de símbols per fer marques i macros per automatitzar la creació de pàgines web i la gestió dels seus continguts.

6.2.1 Estadístiques de l'ensamblat

Per l'usuari és molt important tenir informació sobre com ha funcionat l'experiment de seqüenciació i de com ha anat el protocol de pre-ensamblat i ensamblat, a més de poder accedir als resultats finals obtinguts. Per aquest motiu, a la mateixa web s'han creat pàgines específiques per cada *run*, on hi ha tota la informació prèvia a la anotació de les seqüències, enllaçades a partir de la pàgina principal que conté les taules resum. Cadascuna d'aquestes pàgines està dividida en 3 apartats: informació sobre la seqüenciació, informació sobre l'ensamblat i informació de les assignacions als grups taxonòmics i filtrats després de la cerca d'homologia amb `BLAST`. Totes aquestes pàgines i les taules que contenen es defineixen també amb el sistema de gestió documental basat en `Tiki`.

En la primera secció a la que l'usuari podrà accedir hi ha les taules amb la informació sobre el procés de seqüenciació. En aquesta taula estan disponibles els informes que faciliten els serveis tècnics de seqüenciació un cop finalitzat tot el procés. Per a cada mostra és pot veure el nom de l'experiment en el qual està englobat, el codi identificatiu i una descripció de la mostra (tipus de matriu, si és control negatiu, o especificacions importants a tenir en compte). Es pot veure un exemple a la Figura 6.3 a la pàgina 100.

En la segona secció s'inclouen els anàlisis per avaluar les dades de seqüències obtingudes en els experiments de seqüenciació. Es tracten separadament les seqüències R1 i R2 de les mostres multiplexades per experiment. En els gràfics d'aquest apartat de la pàgina web, l'usuari podrà veure el número de seqüències, el percentatge de GC, la llargada mitjana i la longitud total de les seqüències tres gràfics, dos dels quals explicats anteriorment a les seccions 4.1 i 4.2 (veure Figura 4.2 de la pàgina 48). El tercer és un gràfic de distribució del contingut del GC en les seqüències, on l'eix de les X correspon al percentatge de GC i l'eix de les Y a la llargada de la seqüència (veure Figura 6.4 de la pàgina 101).

En la secció següent es resumeix la informació relativa al processament dels *reads* en funció de la seva complexitat. En la Figura 6.5 a la pàgina 102 apareix una taula que conté una columna amb un resum de les diferents estadístiques descriptives calculades sobre els principals paràmetres analitzats de cada mostra

[3] `https://shutter-project.org/`

Experiment	Run Code	Run Name	CLC assembler	Velvet assembler
Residual Water	R4_DNA_RNA	Winter DNA+RNA	Link	Link
Residual Water	R5_DNA	Winter DNA	Link	Link
Residual Water	R6	Primavera	Link	Link
Serum clinical	SH1	Human serum autoinmune+inmunodepressed	Link	Link
Serum clinical	SH2	Human serum positive in HEV	Link	Link
Serum clinical	SH3	Human serum without cause, 8 men A	Link	Link
Parsley & River	P1	Parsley 30 cicles	Link	Link
Parsley & River	P2	Parsley 20 cicles	Link	Link

Figura 6.2 Exemple de la taula principal d'entrada on es llisten els enllaços per accedir a les taules dinàmiques corresponents. Aquesta taula està definida directament amb el sistema de marcat de la `Tiki`, com s'explica en el text.

Run Code	Run name	Name input file
R5_R1	Winter DNA	DNA-SEWAGE_S2_L001_R1_001.fastq
R5_R2	Winter DNA	DNA-SEWAGE_S2_L001_R2_001.fastq
H1_R11	Urine	HUMAN_S5_L001_R1_001.fastq
H1_R2	Urine	HUMAN_S5_L001_R2_001.fastq
R4_R1	Winter DNA+RNA	RAW-SEWAGE_S4_L001_R1_001.fastq
R4_R2	Winter DNA+RNA	RAW-SEWAGE_S4_L001_R2_001.fastq
S1_R1	Rhesus serum 1, week 1	SEREUM-W1_S3_L001_R1_001.fastq
S1_R2	Rhesus serum 1, week 1	SEREUM-W1_S3_L001_R2_001.fastq

Figura 6.3 Taula sumari per les mostres presents en un *run*, on s'indica el seu codi, la descripció breu i els fitxers dels *raw reads*. Com amb la Taula anterior, aquesta també és una taula estàtica definida amb la sintaxi de la Tiki.

Run Code	Run name	# sequences	Avg %GC	graph GC	Avg Len	Total Len	seq quality	nucleotide
R5_R1	Winter DNA	1431232	47.76		227.99	326299732		
R5_R2	Winter DNA	1431232	48.10		228.37	326850508		
H1_R11	Urine	1823781	40.71		217.30	396309613		
H1_R2	Urine	1823781	40.84		217.72	397072756		
R4_R1	Winter DNA+RNA	1969416	43.20		213.52	420500684		
R4_R2	Winter DNA+RNA	1969416	43.83		215.15	423717197		
S1_R1	Rhesus serum 1, week 1	1518494	45.96		214.22	325291378		
S1_R2	Rhesus serum 1, week 1	1518494	47.10		217.91	330893125		

Figura 6.4 Informació sobre els filtrats *raw* d'un dels experiments. Exemple de la taula on podem accedir als gràfics i a la informació sobre la qualitat de l'experiment de seqüenciació per cadascuna de les mostres del *run*.

(llargada, percentatge en GC, Entropia, Trifonof i nivell de compressió). La Figura 6.6 a la pàgina 102 ens permet veure en detall aquesta Taula descriptiva per una de les mostres en concret. La columna *histograms* permet accedir als gràfics de les distribucions de densitat per cada conjunt de seqüències respecte els valors de les variables representades en la Taula d'estadístiques, (Figura 6.7 de la pàgina 103). La columna relativa a la *complexity vs lenght* mostra la distribució dels valors respecte la llargada per cadascuna de les seqüències (Figura 6.8 de la pàgina 104). Finalment, a la darrera columna es pot accedir als *scatterplots* on es comparen dos a dos els diferents paràmetres; això ens ha de facilitar la valoració de quins són els factors que ens ajudaran a analitzar la complexitat de les seqüències i definir posteriorment els paràmetres del filtrat per complexitat més escaients (Figura 6.9 a la pàgina 105).

La última taula amb informació sobre els ensamblats consisteix en un resum de les seqüències que han superat tots els filtres i que han estat utilitzades per generar l'ensamblat corresponent. La taula inclou el número de seqüències de tipus *pair-ends* i *single-ends* que han passat tots els filtres, la llargada mitjana, número de *contigs* construïts pel programa ensamblador i la seva llargada mitjana (pels dos ensambladors amb els que hem treballat, `CLCBio` i `Velvet-MetaVelvet`). Per completar la descripció de les pàgines on es resumeixen els resultats de tot el protocol, tenim la secció on es llisten els resultats sobre les cerques d'homologia del `BLAST` i les assignacions a nivell taxonòmic. La última taula és un resum dels tres tipus de cerques realitzades: `BLASTN` contra la base de dades de genomes complets virals de `NCBI-GenBank`, `BLASTN` contra la base de dades de seqüències nucleotídiques de `NCBI-GenBank`; i un `BLASTX` contra la base de

Run Code	Run name	summary table	histograms	complexity vs length	data distribution
R4DR_R1	Winter DNA				
R4DR_R2	Winter DNA				
R5D_R1	Urine				
R5D_R2	Urine				
H1_R1	Winter DNA+RNA				
H1_R2	Winter DNA+RNA				
S1_R1	Rhesus serum 1, week 1				
S1_R2	Rhesus serum 1, week 1				

Figura 6.5 Taula resum amb la informació sobre la complexitat de les seqüències per cadascuna de les mostres del *run*. A més de mostrar dades descriptives de les variables que juguen un paper important en la complexitat, també permet accedir a les diferents gràfiques relacionades amb aquest càlcul de complexitat.

SEQlen	NUClen	GCpct	LZWpct	Entropy	CLingTrif
Min. : 35	Min. : 0.0	Min. : 5.578	Min. : 2.80	Min. :0.0000	Min. :0.0090
1st Qu.:247	1st Qu.:247.0	1st Qu.:42.629	1st Qu.:41.04	1st Qu.:0.9680	1st Qu.:0.9660
Median :251	Median :251.0	Median :49.200	Median :41.83	Median :0.9830	Median :0.9710
Mean :228	Mean :227.9	Mean :47.679	Mean :43.03	Mean :0.9733	Mean :0.9651
3rd Qu.:251	3rd Qu.:251.0	3rd Qu.:52.988	3rd Qu.:42.80	3rd Qu.:0.9900	3rd Qu.:0.9750
Max. :251	Max. :251.0	Max. :84.800	Max. :88.57	Max. :1.1570	Max. :0.9980
NA	NA	NA's :2306	NA	NA	NA

Figura 6.6 Taula descriptiva dels paràmetres de seqüència i complexitat avaluats. Aquests paràmetres inclouen longitud (NUCLen), %GC, compressió (LZWpct), entropia i Complexitat Lingüística de Trifonov (CLingTrif). Les files corresponen respectivament al valor mínim, el valor del primer quartil, la mediana, la mitjana, el valor del tercer quartil, el valor màxim i la quantitat de valors nuls definits com (*Not Available*, "NA").

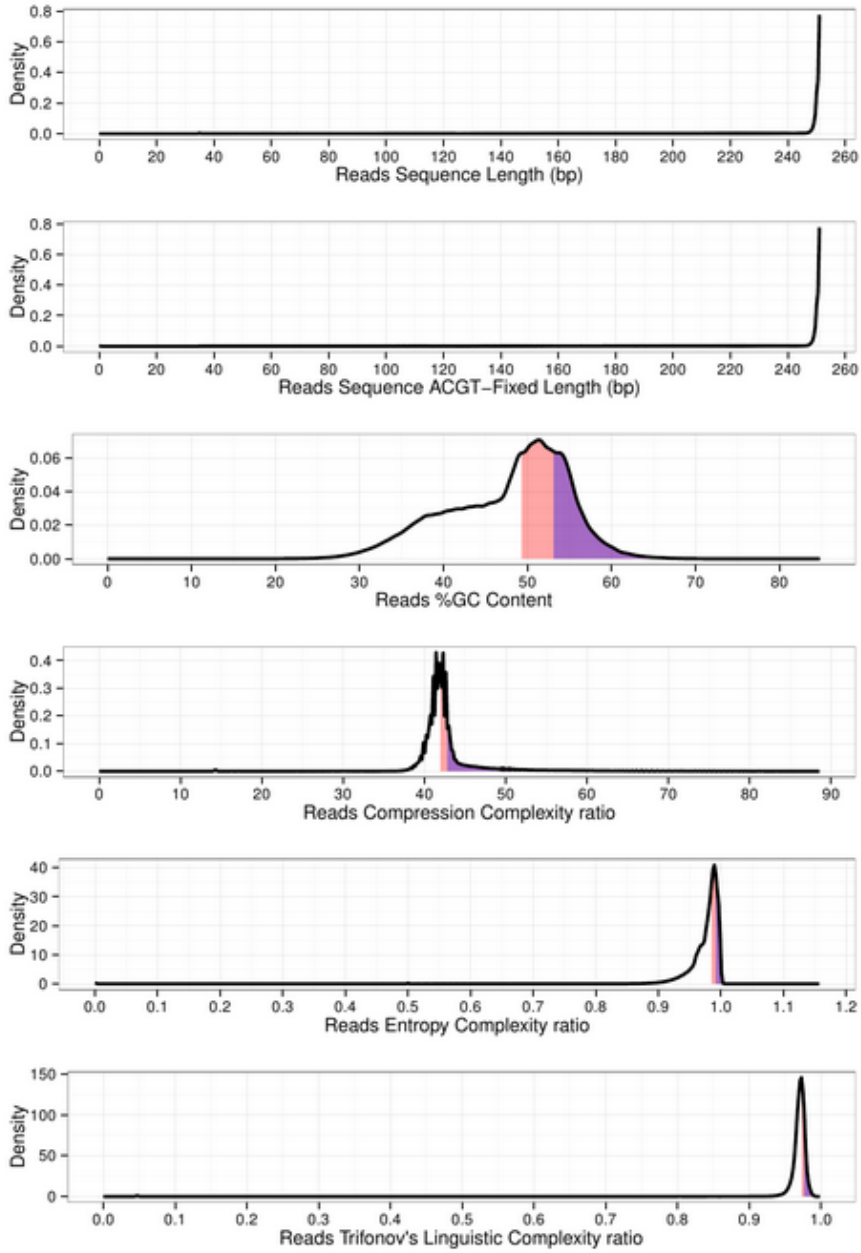


Figura 6.7 Gràfic de les distribucions de densitats basades en les freqüències dels diferents paràmetres analitzats. De dalt a baix, els panells mostren la longitud de la seqüencia (en bp), el percentatge en GC, el percentatge de compressió, l'entropia i la complexitat lingüística. En tots els gràfics, l'àrea sota la corba a partir del segon i tercer quartils s'ha remarcat en rosa i lila respectivament.

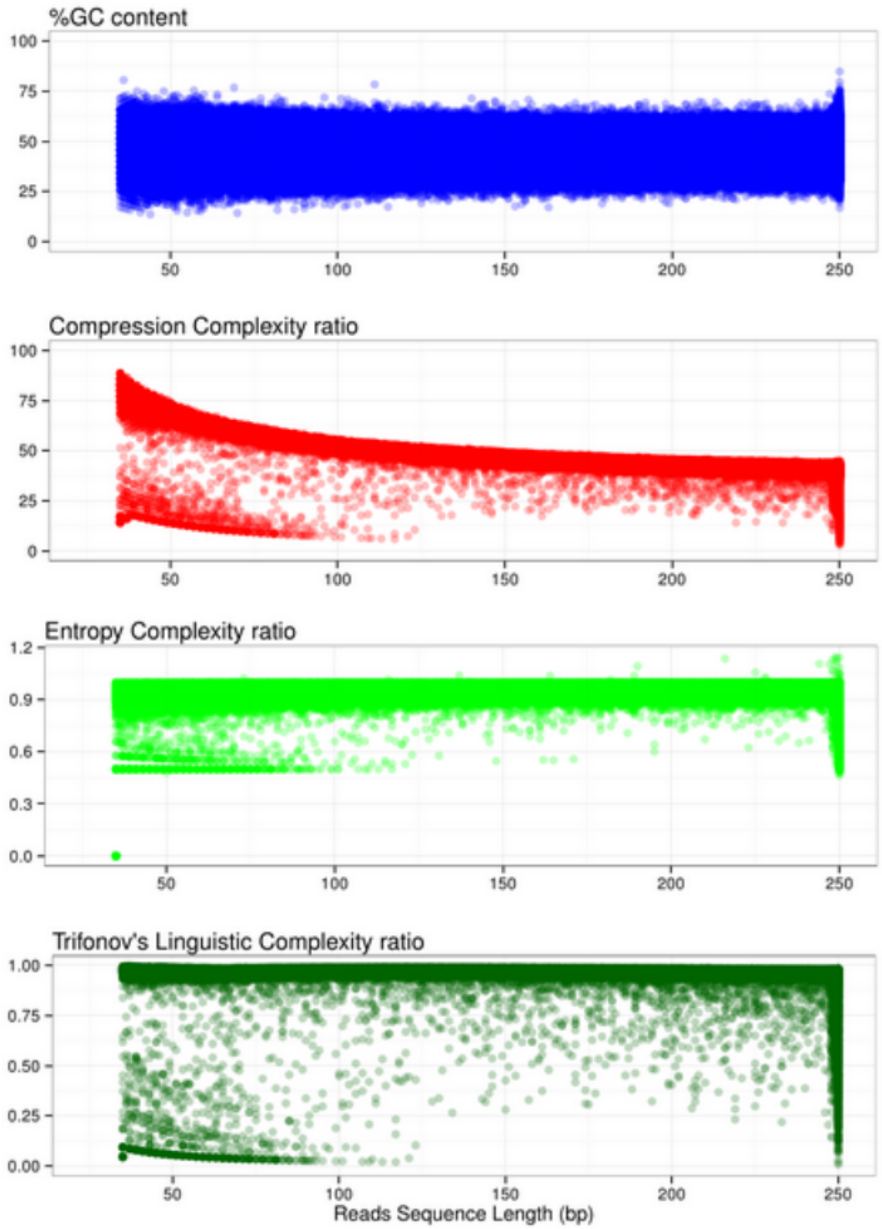


Figura 6.8 Gràfiques de la distribució dels diferents paràmetres analitzats respecte la longitud de seqüència. S'ha afegit un canal alfa per dibuixar els punts, el que permet ressaltar algunes variacions en la densitat del nombre de seqüències que presentin valors similars. En el cas del %GC s'observa una distribució més o menys similar per les diferents longituds de seqüència; en canvi, en les mesures de complexitat es fa evident que hi ha un *pool* de seqüències que s'agrupa al voltant de valors relativament baixos (com passa en les cantonades inferiors esquerres dels tres panells)

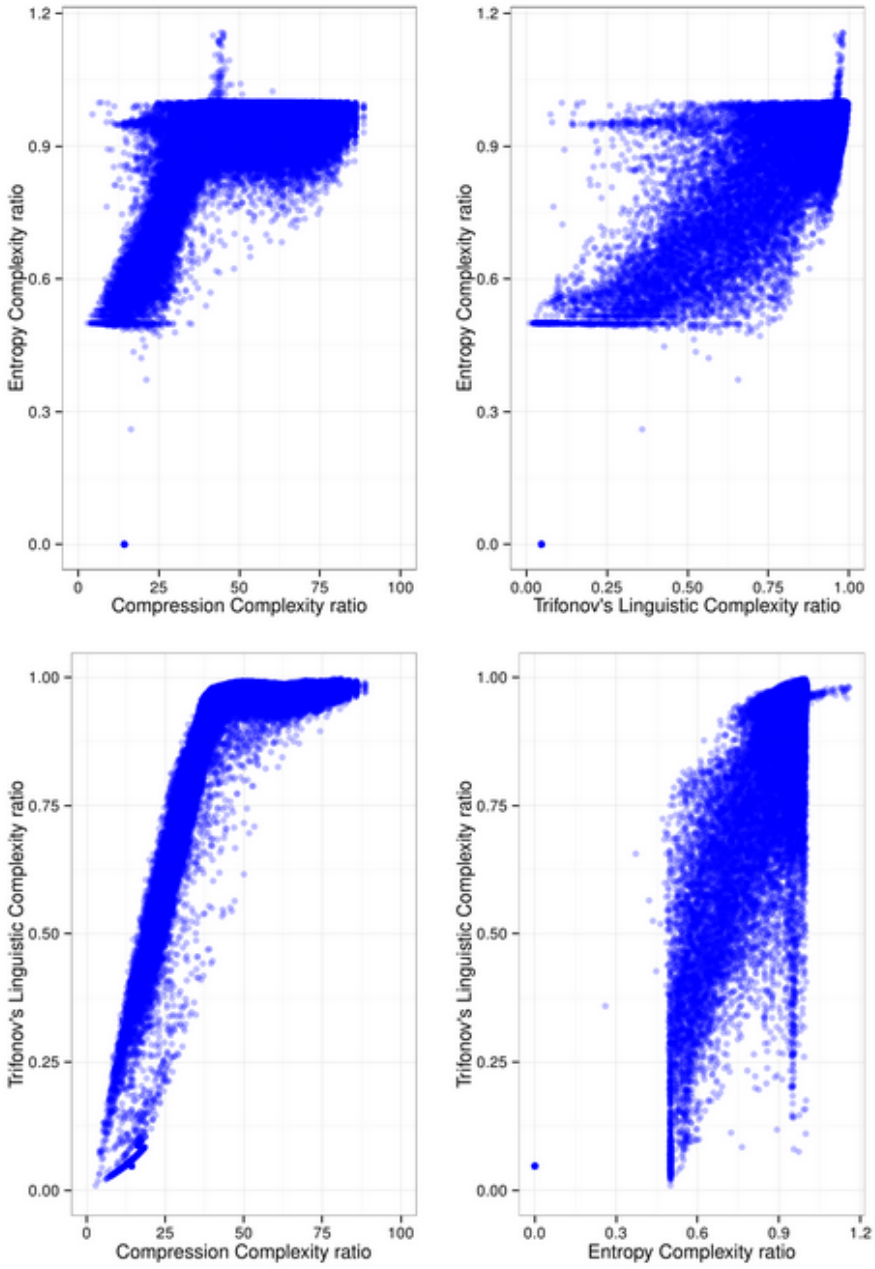


Figura 6.9 Comparació dos a dos dels tres paràmetres de complexitat analitzats. D'aquests gràfics, el que presenta una distribució més interessant des del punt de vista de la nostra recerca, és el que ens compara els valors de compressió i la puntuació de CL de Trifonov (panell inferior esquerre).

Run code	Run name	Paired-end Mi-Seq	Singletons after filters	Paired-end after filter	Contigs Velvet	Singletons Velvet	Mean lenght Velvet	Contigs CLC	Mean lenght CLC
R4	Winter DNA	1410434	7351	1410434	494487	2024082	231,74	83773	370,01
R5	Urine	1786538	8243	1786538	829715	1990575	218,29	171366	353,92
H1	Winter DNA+RNA	1327152	47726	1327152	158894	964394	229,92	88498	417,11
S1	Rhesus serum 1, week 1	892859	38159	892859	103444	1460947	231,61	24158	328,47

Figura 6.10 Estadístiques de les seqüències filtrades. Informació sobre les seqüències que han passat tots els filtres pre-ensamblat i els *contigs* generats pels dos programes d'ensambladors utilitzats en el nostre protocol (`CLCBio` i `Velvet-MetaVelvet`).

Run code	Run name	HSPs vrl	HSPs gnms	HSPs Uniprot	Putative viral seqs vrl	Putative viral seqs gnms	Putative viral seqs Uniprot	Seqs without blast hit vrl	Seqs without blast hit gnms	Seqs without blast hit prot
R4	Winter DNA	107648	44283	1437608	28429	12111	24927	72911	73992	48948
R5	Winter DNA+RNA	159218	25140	1199568	25545	10087	13961	164883	165743	152148
H1	Urine	298664	8246	163235	12952	2335	2120	73685	87392	85015
S1	Rhesus serum 1, week 1	105604	300	42082	3137	570	303	19258	24042	23770

Figura 6.11 Estadístiques de les seqüències que s'han pogut assignar a un grup taxonòmic a partir de les cerques amb `BLAST`. Es pot apreciar les diferències entre les diferents bases de dades: la de genomes complerts virals de `NCBI-GenBank`, la de seqüències nucleotídiques virals de `NCBI-GenBank` i la de proteïnes virals d'`Uniprot`. En aquest exemple, només es mostren els resultats obtinguts a partir dels *contigs* generats per `CLCBio`.

dades de proteïnes virals filtrades d'`Uniprot`. En aquesta taula es resumeix el número total d'alineaments (HSPs), el número de seqüències que putativament s'han pogut assignar a un virus i el nombre de seqüències que no han fet cap *hit* amb el `BLAST`. La diferència entre el número de virus putatius i el número de seqüències que no han fet cap *hit* amb el `BLAST` no es correspon amb la suma del total de seqüències, ja que no totes les seqüències que potencialment tenen un *hit* d'homologia amb alguna seqüència de les bases de dades passen els filtres descrits en el capítol anterior en el moment de fer l'assignació taxonòmica (Figura 6.11 a la pàgina 106).

6.2.2 Gràfics "Krona"

Krona és un tipus de gràfic interactiu que serveix per explorar les dades agrupades en forma de diagrama de sectors. És una eina molt compacta que facilita l'exploració de les dades taxonòmiques i l'obtenció d'una representació global dels resultats obtinguts. En el nostre cas, es van definir tres nivells diferents en la taxonomia dels virus, concretament grup Baltimore, família i espècie. En tot

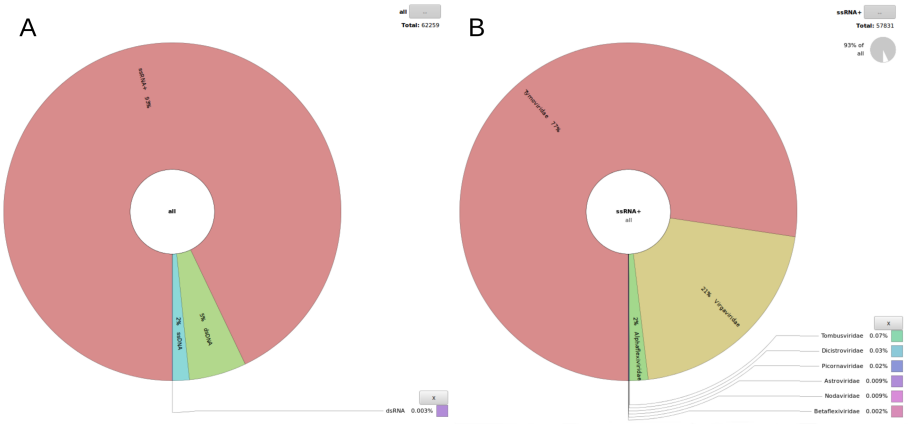


Figura 6.12 Gràfica *krona* d'una mostra en concret. Les gràfiques `Krona` aprofiten les capacitats implementades en llibreries JavaScript per facilitar la navegació a través dels grups taxonòmics. **A**: `Krona` a nivell de grup Baltimore. **B**: Resultat de fer clic sobre la porció on estan agrupats tots els *hits* relatius a seqüències de virus ssRNA+.

moment es pot veure quin és el percentatge sobre el total de les seqüències al que pertany a cadascun dels grups, representats com a sectors en aquests diagrames interactius. L'usuari pot anar cap a un nivell superior o inferior tot fent clic sobre els diferents sectors per ampliar els detalls corresponents taxonòmics inferiors (veure les figures 6.12 A i B de la pàgina 107). Un dels principals avantatges d'aquest gràfic és que l'usuari pot explorar els resultats de manera ràpidament i senzilla, sense entrar en detalls dels resultats obtinguts. L'usuari pot identificar si hi ha o no alguna seqüència que pugui ser representativa d'una família o espècie en concret, i simultàniament saber la proporció de seqüències que s'han trobat a la mostra per grup Baltimore, família o espècie. També pot saber al mateix temps quin ha estat el grup més o menys representat a la mostra analitzada.

6.2.3 Taules dinàmiques

Les taules dinàmiques han estat creades per a que l'usuari pugui navegar pels resultats que s'han obtingut en cadascuna de les mostres d'una manera ordenada i fàcil, alhora que li permet recuperar subconjunts de seqüències que poden estar relacionades taxonòmicament. Les dades s'estructuren a partir d'un arbre basat en la classificació taxonòmica dels virus, que permet desplegar grups a diferents nivells perquè l'usuari pugui aprofundir en aquelles famílies i espècies en les que estigui més interessat.

Els enllaços per accedir a aquestes taules estan disponibles a partir de la taula

estàtica del llistat que conté totes les mostres seqüenciades (Figura 6.2 a la pàgina 100). El primer nivell de la taula són els diferents grups de Baltimore per als que almenys hi ha una seqüència anotada a la mostra. Es poden realitzar cerques directes tant per nom de família i espècie com per l'identificador numèric del registre taxonòmic del NCBI (*NCBI-Taxonomy Browser*).

Al fer clic sobre els diferents grups, llistats a la columna "Taxon" es van desplegant els diferents sub-nivells i sub-sub-nivells, que corresponen a famílies i espècies respectivament; el nivell inicial, com ja hem dit, està associat al tipus de genoma viral seguint la classificació de Baltimore que ha estat descrita a la Introducció (1.3). Si es torna a clicar es pot anar minimitzant el grup taxonòmic escollit. Per facilitar cerques a nivell taxonòmic, també s'inclou una columna on apareix el codi `NCBI` pel taxò corresponent, al clicar sobre el qual s'obre una pestanya nova amb la pàgina web específica del *Taxonomy Browser* del `NCBI` (Benson *et al.*, 2017; NCBI Resource Coordinators, 2013). Aquest identificador només surt en els nivells de família i espècie, ja que la classificació Baltimore no té un codi específic que sigui equivalent per exemple a nivell de *Phylum*. La tercera columna ens mostra el número de seqüències que han estat classificades dins d'aquell grup. El color de fons d'aquesta casella és una escala de verds, on la intensitat es proporcional al nombre total de seqüències assignades a aquell grup. La quarta i cinquena pcolumna permeten visualitzar i recuperar les seqüències en format FASTA i els alineaments del BLAST que han determinat la classificació de les seqüències dintre d'aquest grup respectivament. En clicar sobre les icones apareix un panell emergent amb la informació corresponent. A nivell d'espècie apareixen nou columnes més: les tres primeres són el valor mínim, la mitjana, i el valor màxim de la llargada dels *hits* obtinguts amb `BLAST`; de la 4 a la 6 tenim el valor mínim, la mitjana, i el valor màxim de les longituds en bp de les seqüències; i de la columna 7 a la 9 es mostren el valor mínim, la mitjana, i el valor màxim per la identitat dels alineaments de `BLAST` (Figura 6.14 a la pàgina 109).

Hi ha l'opció de recuperar a quina base de dades ha fet *hit* cadascuna de les seqüències, així com la identitat o la llargada de cadascuna d'elles. Per accedir a aquesta informació s'ha de clicar, estant a qualsevol nivell de la taula, sobre una cel·la de la columna 3, corresponent al número de *hits*. Als panells emergents, a més de botons per retornar a la pàgina anterior o tancar el panell, podem tenir un botó per moure el contingut del panell a una nova pestanya o finestra del navegador web. El nombre de taules que apareixen en el nou panell correspon al nombre d'espècies diferents presents dins del nivell de taxonomia que s'hagi clicat. Cada fila de les taules que es mostren en els panells detallats que acabem de comentar és un *contig*; la primera columna correspon al identificador d'aquest *contig*, la segona i tercera columna son els enllaços per descarregar

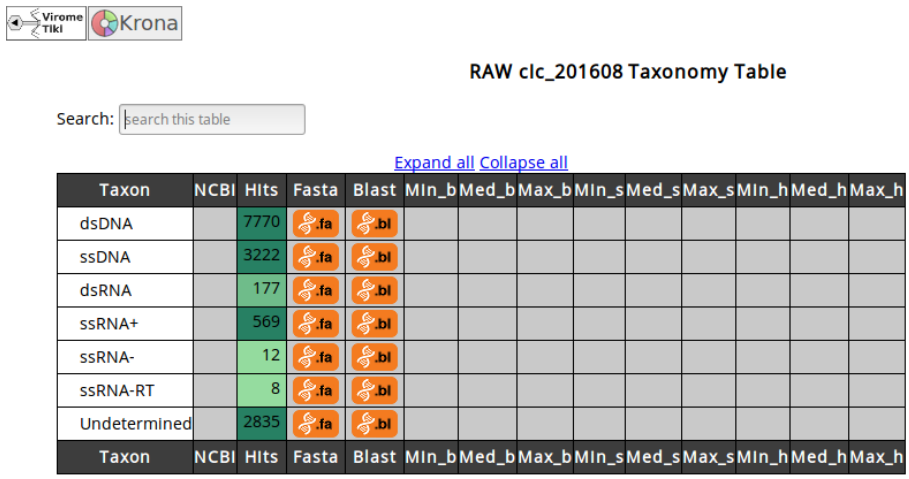


Figura 6.13 Taula dinàmica d'una mostra en concret *Raw*. En aquesta taula tots els subnivells estan colapsats i tan sols es mostren 7 taxons a nivell de classificació Baltimore, incloent-hi un calaix de sastre per aquelles seqüències víriques que encara no estan ben classificades (*undetermined*).

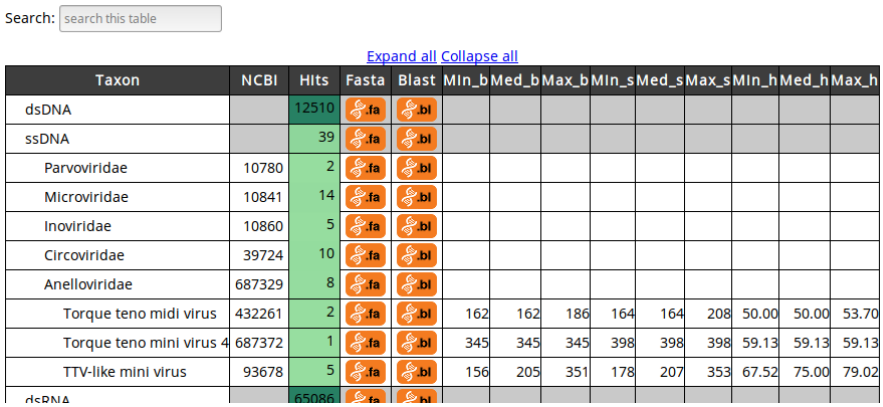


Figura 6.14 Taula dinàmica d'una mostra en concret desplegada. Partint de la taula de la Figura 6.13, hem desplegat el grup de Baltimore dels virus ssDNA i, dintre d'aquest, la família *Anelloviridae*. A nivell d'espècies, a banda de poder recuperar el subconjunt de *contigs* i els resultats del BLAST que donen suport a l'assignació taxonòmica, també veiem algunes dades descriptives sobre aquests alineaments. Per detalls sobre les columnes veure text.

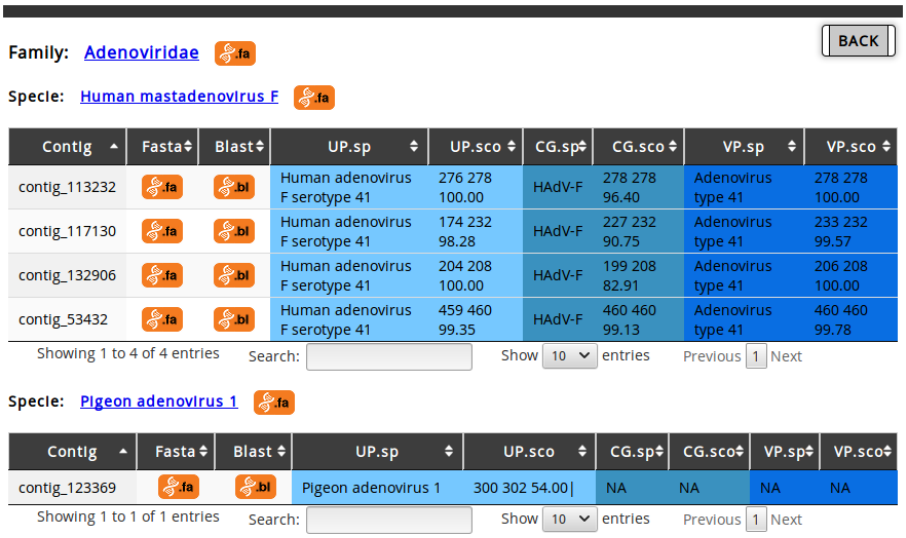


Figura 6.15 Taula dinàmica per una família en concret. Quan l'usuari fa clic sobre una cel·la de la columna *hits* a la taula dinàmica, per exemple la mostrada a la Figura 6.14, s'obre un panell on es pot recuperar el llistat per seqüència. En l'exemple, s'ha seleccionat una cel·la a nivell de família, "*Adenoviriade*", s'ha generat una sub taula per a cada espècie de la família. Les diferents columnes mostren una descripció de l'alineament de les diferents bases de dades.

el FASTA i l'alineament corresponent. Les següents columnes, agrupades de dos en dos i fent servir diferents tonalitats de blau, ens donen informació sobre l'espècie i la puntuació del millor candidat trobat al fer les cerques amb `BLAST`; els prefixes "UP", "CG" i "VP" ens remeten a les tres bases de dades utilitzades: `UniProt`, genomes virals sencers de `NCBI-Genomes` (*Complete Genomes*) i seqüències virals de `NCBI-GenBank` respectivament. La columna de la informació de l'alineament consta de 3 paràmetres: la llargada del *contig*, la llargada del *hit*, i la identitat de l'alineament. Finalment, en aquestes taules també es poden fer cerques i ordenar per la variable espècie o identitat de les tres bases de dades que es vulgui, tot clicant al nom de la columna corresponent a la línia d'encapçalament.

Discussió

7 Discussió

Avui en dia els estudis de metagenòmica es centren majoritàriament en bacteris, amplificant regions de d'ARN 16S ribosomal, seguits dels eucariotes on es sol utilitzar el 18S ribosomal. Les tècniques de metagenòmica s'estan introduint en el camp de la virologia principalment per realitzar anàlisis de metagenòmica dirigida envers alguna família o espècie en concret, el que permet caracteritzar fàcil i ràpidament moltes variants, així com inferir reconstruccions filogenètiques més acurades.

Els estudis de metagenòmica sobre genomes complets de virus (NGS) s'han aplicat ja amb èxit en diversos treballs on es tractaven mostres fecals (Reyes *et al.*, 2010), mostres d'oceans (Hurwitz *et al.*, 2013), mostres de pacients amb malalties respiratòries (Lysholm *et al.*, 2012), en sediments pelàgics (Yoshida *et al.*, 2013), en aire (Whon *et al.*, 2012), o fins i tot per l'anàlisi de mostres sanguínies (Moustafa *et al.*, 2017). Com a resultat, tots aquests estudis han incrementat el nombre de seqüències víriques disponibles a les bases de dades genèriques com `NCBI-Genbank` o `UniProt`.

Paral·lelament, s'han creat bases específiques per metagenòmica on dipositar les seqüències en brut generades pels experiments de seqüenciació, però també en alguns casos per fer accessibles els resultats corresponents (Eisen, 2007), com passa per exemple amb `MetaVir`[1], on s'emmagatzemen més d'una cinquantena d'anàlisis publicats. Una altra aplicació és `MG-Rast`[2] (Meyer *et al.*, 2008)), que permet fer estudis d'anotació funcionals i filogenètics de mostres microbials sobre dades de metagenòmica dirigida amb 16S i 18S. La plataforma `IMG/VR` (*Integrated Microbial Genomes/Virus*, Paez-Espino *et al.* 2017), creada a l'estiu del 2016, a banda d'emmagatzemar els resultats de les assignacions, permet també executar protocols d'anotació i comparacions de genomes sobre mostres de diferents localitzacions.

Un dels avantatges de la creació d'aquestes pàgines web és que al estar totes les dades de metagenòmica centralitzades, ens permet accedir a través a una interfície unificada i uniforme a les dades en cru, *raw data*, o als resultats de les anotacions i que en algun cas ens pot facilitar la comparació entre mostres. Els problemes amb el que ens podem trobar amb les dades en cru en les bases de dades convencionals com `NCBI-GenBank` és que aquestes seqüències no hagin estat verificades, una fracció pot tenir baixa qualitat, i a més poden presentar

[1] `http://metavir-meb.univ-bpclermont.fr/`
[2] `http://metagenomics.anl.gov/`

una gran redundància. La creació de bases de dades especialitzades fa que l'usuari pugui saber en tot moment quin tipus de mostra i experiment s'han obtingut les seqüències.

7.1 Metagenòmica sobre genomes complets (WGS)

Entre els principals obstacles que ens trobem en el moment d'abordar un projecte de metagenòmica viral és que els virus no tenen un marcador específic. Per aquesta motiu, si no es pretén estudiar una família en concret, s'utilitzen *random primers* per amplificar les mostres i generar les llibreries de seqüenciació; el que pot comportar l'obtenció d'un gran percentatge de seqüències que no són virals. Tal i com s'ha explicat en el capítol corresponent, per solucionar aquest problema es realitzen tractaments previs a l'etapa de seqüenciació, per tal que el material genètic que anem a seqüenciar presenti el major nombre de seqüències víriques.

En els nostres estudis hem utilitzat mètodes de concentració com SMF (Calgua *et al.*, 2013) o la ultracentrifugació (Annex, article 3), tractaments de DNAsa per digerir i eliminar ADN lliure eucariota o procariota de les mostres, i centrifugats per concentrar encara més les partícules víriques. Tot i els esforços que s'han dut a terme, tal com es pot observar en les Taules 5.7, 5.8 i 5.9 de les pàgines 85, 86 i 87, el percentatge de seqüències que estan anotades en alguna de les tres bases de dades de seqüències virals conegudes és molt baix. Per exemple, en les mostres d'aigües residuals el percentatge de seqüències anotades va de 0,58% a 4,81% mentre que en les mostres clíniques representa una mitjana de 7,8% i en les mostres de julivert es troba entre un 1,81% i un 26,11%. Aquests percentatges es podrien explicar per dues causes principals. La primera, que el tractament que es realitza a les mostres, previ a la seqüenciació, no sigui prou eficient; això implica que s'haurien de millorar o canviar els protocols experimentals per aconseguir concentrar encara més el material genètic víric de les mostres. L'altra causa dels baixos percentatges podria ser la completitud de les bases de dades utilitzades per l'anotació de les seqüències generades durant el procés de seqüenciació; és a dir, el percentatge de seqüències víriques podria ser més elevat però que no es detectessin correctament, el que hom descriuria com falsos negatius.

Per les anàlisis descrites en aquesta tesi s'han utilitzat tres bases de dades diferents per intentar detectar i anotar el màxim de seqüències virals possibles. La base de dades amb la que s'obtenen més resultats és genera a partir de totes les proteïnes víriques enregistrades a `UniProt`; deixant de banda que és la base de dades amb més seqüències, el programa escollit per fer les cerques d'homologia és `BLASTX`, el qual tradueix les seqüències de nucleòtids de les nostres mostres

a les sis possibles pautes de lectura per traduïr-les a aminoàcids i comparar-les amb les seqüències de la base de dades. Els errors puntuals de la seqüenciació poden ser parcialment evitats gràcies a el codi genètic degenerat, però el factor més important és el fet que les puntuacions d'alineaments entre proteïnes són millors que les dels nucleòtids. La segona base de dades utilitzada s'ha construït a partir d'una selecció de la base de dades de `NCBI-Genbank`, filtrant únicament sols que conté les seqüencies que estan classificades com a genomes complets de virus. L'avantatge de fer cerques sobre aquesta base de dades és que les seqüències són més llargues; el seu principal inconvenient és el baix nombre de seqüències ben anotades de genomes complets disponibles (5 400).

El fet de combinar cerques sobre les tres bases de dades mencionades ens dóna una major fiabilitat en l'assignació taxonòmica de les seqüències; és a dir, si una seqüència ha estat classificada dins d'un grup taxonòmic a partir de l'alineament amb les tres bases de dades, la fiabilitat que aquesta assignació sigui correcte és molt alta. En canvi, si una seqüència tan sols presenta evidències d'homologia per la base de dades de nucleòtids extreta de `NCBI-Genbank`, és més probable que sigui un fals positiu especialment quan les puntuacions de l'alineament són "pobres" (baixa cobertura, *E-values* grans, baixa identitat i alineaments curts)

7.2 Dissenys experimentals

En els estudis metagenòmics, ja siguin de mostres ambientals o clíniques, es poden analitzar mostres individuals o agrupar-les en *pools*; però en molts pocs casos s'en fan rèpliques o controls negatius. Això és degut per dos motius, el primer es l'obtenció de la mostra i el segon econòmic. Tal i com s'ha esmentat anteriorment, obtenir la concentració requerida per poder construir les llibreries de seqüènciació a partir d'una mostra pot resultar complicat. Pel que fa a la segona raó esmentada, els motius econòmics (Knight *et al.*, 2012; Prosser, 2010), degut al cost encara elevat de processar les moltes llibreries i experiments de seqüenciació (o *runs*) que caldrien per augmentar la potència estadística en estudis quantitatius.

En els experiments sobre mostres d'aigua residual es van realitzar controls negatius amb aigua destil·lada de qualitat i aigua potable de distribució. En les mostres on es comparaven diferents mètodes de concentració, per ultracentrifugació i SMF hi havia dues rèpliques per mostra. Però en aquestes mostres no s'obtenen resultats exactament iguals, fet que ens fa pensar que en els estudis de metagonòmica, sobretot en els que s'utilitzen *random primers*, s'hauria de dissenyar estudis amb rèpliques per tal de tenir uns resultats més fiables, però per causes econòmiques o d'obtenció de la mostra sovint és impossible de realitzar.

Encara que els costos de seqüenciació s'han reduït dràsticament entre la primera i la segona generació de seqüenciació, continuen sent tecnologies molt cares, a les que no tothom hi pot accedir. Illumina ja ha anunciat que a part d'incrementar el rendiment i la llargada de les seqüències, un dels seus objectius és reduir encara més el cost per base. Tecnologies com Ion torrent o PacBio ja han aconseguit el repte de reduir els costos i augmentar les longituds de seqüència, pel simple fet de que no necessiten tants reactius (per exemple nucleòtids marcats). En els pròxims anys s'espera que aquests costos disminueixin encara més; així hi podrà accedir un espectre més gran de la comunitat científica i es podran realitzar més rèpliques. Un dels paràmetres que s'utilitza per quantificar el cost de seqüenciar ADN és el total de diners que costaria seqüenciar un genoma humà. En la primera versió d'aquest genoma, utilitzant la tècnologia Sanger i el mètode de seqüenciació per *shotgun*, el cost va ser aproximadament de tres mil milions de dòlars; amb les tècniques de la segona generació el cost seria d'uns quatre mil dòlars i amb les més modernes tecnologies de la tercera generació el cost pot baxar fins a 100 dòlars.

7.3 Protocols de neteja de seqüències pre-ensamblat

Tant en el capítol de metodologia com en el dels resultats s'ha comentat que el pas de neteja de seqüències abans de l'ensamblat ens serveix per aconseguir un millor ensamblat i reduir el nombre inicial de seqüències. En el nostre cas s'apliquen diverses eines desenvolupades per altres grups, com `Trimmomatic` o `FASTX`; però també s'ha desenvolupat una eina nova per mesurar la complexitat dels *raw reads*, a través de diferents valors com la Complexitat Lingüística de Trifonov o el nivell de compresssió. Això ens permet filtrar aquelles seqüències de baixa complexitat, com per exemple les seqüències repetitives, les quals poden ocasionar solapaments erronis en el moment d'ensamblar els *reads*. El fet d'incloure aquestes seqüències de baixa complexitat podria reduir la riquesa de la mostra. Una altra contraprestació és que en eliminar seqüències de poca qualitat o de baixa complexitat, dues propietats que no tenen perquè anar correlacionades, el temps d'execució dels programes utilitzats per resoldre els ensamblats disminueix de manera dràstica.

7.4 Programes d'ensamblat

En la introducció s'ha exposat que avui en dia hi ha molts tipus d'eines desenvolupades per resoldre el problema de l'ensamblat, tenint com a base un dels tres algorismes (OCL, Greedy o DBG). Aquest pas d'ensamblat és un punt clau en el protocol, ja que si el programa és massa estricte en el moment de connectar els *reads* entre sí, no es reduiria el número de seqüències i la llargada dels *contigs* no seria molt gran. Al contrari, si l'ensamblador és molt laxe, s'obtindran *contigs* d'una llargada elevada, però hi ha el risc de que s'ensamblin seqüències que no corresponguin a la mateixa seqüència genòmica original, o fins i tot a la mateixa espècie en el cas de la metagenòmica; el que hom anomena quimeres d'ensamblat. Les quimeres entre genomes de diferents espècies es deuen majoritàriament al fet que espècies víriques poden compartir dominis funcionals dintre d'una mateixa família, això faria que en algun punt de la creació del graf d'ensamblat sorgeixin nodes comuns a les diferents espècies i al recòrrer aquest graf mal conectat es generin les quimeres.

Després de provar-ne diversos, es va escollir per aquest projecte dos ensambladors diferents: `CLCBio` i `Velvet-MetaVelvet`. Tots dos estan basats en el mètode de grafs de solapament, però malgrat això els resultats obtinguts són lleugerament diferents ente ells. Quan s'utilitza el programa `Velvet-MetaVelvet` es genera una quantitat més elevada de *contigs* i *singletons*; és a dir, s'obtenen més *contigs* però utilitzant menys *reads* per construir-los i la mitjana de la llargada dels *contigs* és més baixa. En el cas del programa de `CLCBio`, es produeixen menys *contigs* i menys *singletons*; és a dir, en la construcció dels *contigs* s'utilitzen més *reads* i la llargada mitjana dels *contigs* és més elevada, comparada amb els del `Velvet-MetaVelvet`. Els *contigs* més llargs obtinguts en el conjunt de les mostres han estat produïts mitjançant el programa `CLCbio`. Això és fa evident per exemple en la comparació de les distribucions de longituds dels *contigs* de les Figures 5.5, 5.6 i 5.4 que es troben en la secció de resultats Si comparem els valors de de riquesa (*richness*) calculats amb les seqüències provinents de l'ensamblat amb `Velvet-MetVelvet` respecte als de `CLCBio`, els primers són més grans, i això implica que el nombre de famílies i espècies diferents identificades en les mostres és superior. Per altra banda, utilitzant les seqüències construides amb l'ensamblador `CLCBio`, encara que s'obtenia una diversitat inferior, s'aconseguia una llargada dels *contigs* superiors. Aquest resultat era més útil per anotar les seqüències a nivell d'espècie i per fer estudis filogenètics. En conclusió, si es vol tenir una visualització taxonòmica més global, sense voler entrar en detalls a nivell filogenètic, s'obtenen millors resultats en general utilitzant `Velvet-MetaVelvet`.

7.5 Anotació taxonòmica basada en homologia

Fer cerques simultàniament en tres bases de dades diferents ens ajuda a identificar moltes més seqüències que si només féssim ús d'una. En el cas de les mostres d'aigua residual, utilitzant únicament els genomes complets de la base de dades de GenBank podríem identificar 22 437 seqüències, amb tota la divisió vírica de la base de dades de GenBank podem identificar 38 239 seqüències; emprant la base de dades de proteïnes víriques d'UniProt, 23 527 seqüències; i considerant el conjunt de les tres bases de dades obtenim 24 429 seqüències. A banda de recuperar més o menys anotacions taxonòmiques, podem estimar la sensibilitat de les anotacions predites i calcular la fiabilitat de cadascuna d'elles, donant més robustesa als resultats finals.

En el cas de les mostres d'aigua residual s'han detectat fins a 35 famílies i un total de 1 289 espècies diferents. Amb l'anàlisi posterior basat en metagenòmica dirigida hem pogut detectar més de 50 soques o tipus diferents de la família d'*Adenovirus*, al contrari que amb el mètode de metagenòmica basat en genomes complets, on només es detectava un tipus en concret d'aquesta família. Amb aquest estudi ens hem adonat que la metagenòmica de genomes complets té limitacions de profunditat; és a dir, per una banda té l'avantatge que podem detectar espècies i famílies no descrites anteriorment en aquest tipus de mostres, però si es vol tenir una informació més detallada sobre alguna d'aquestes espècies s'han de dissenyar anàlisis dirigits a una família o grup de virus.

7.6 Anàlisis filogenètics i caracterització de nous virus

Un dels objectius en els estudis de metagenòmica és intentar fer anàlisis filogenètics per caracteritzar la diversitat d'algun grup taxonòmic en particular i, si és possible, intentar descriure noves variants o espècies dins d'aquest grup (Cornelissen-Keijsers *et al.*, 2012; Ng *et al.*, 2012). Per tal de poder realitzar aquestes anàlisis és necessari modelar dominis de seqüència que permetin tipificar els diferents grups.

D'altra banda, en el cas de la metagenòmica de genomes complets no es pot seleccionar quina és la regió o regions que es seqüenciaran. Per tant, és possible que en la seqüenciació, no obtinguem les regions més idònies per realitzar estudis filogenètics, que el solapament dels *reads* provinents de diferents espècies no sigui uniforme o que els *reads* de cada espècie es projectin sobre regions diferents i no arribin a solapar-se per poder alinear-los de cara a construir els

arbres filogenètics. En aquests casos es presenten diverses opcions: que la seqüència obtinguda sigui la regió que segons l'ICTV (*International Committee on Taxonomy of Viruses*) és la que s'utilitza per fer les anàlisis filogenètiques; que la seqüència contingui únicament una part de la regió específica, en aquest cas, en el moment de fer l'alineament amb les seqüències de referència s'extreu la mateixa regió que s'ha obtingut en la seqüència de la mostra. Una tercera opció és que la seqüència no contingui cap tros de la regió que ens interessa, però sigui relativament llarga, en aquest cas agafaríem la regió homòloga de les seqüències de referència de la família que estem estudiant.

En les mostres d'aigua residual és va detectar un *contig* que es relacionava amb seqüències de la família dels *Hepeviridae*, però amb una identitat molt baixa. En agafar la regió d'una proteïna RdRp de les espècies de referència per aquesta família i les seqüències que s'havien descrit en un article publicat anteriorment (Ng *et al.*, 2012), es va poder observar que filogenèticament era més propera a una nova espècie descrita, però encara no classificada, que als *Hepeviridae*. Malgrat els esforços, el fet de que no es pogués recuperar la regió completa de la RdRp fa que no es pugui determinar o establir una classificació definitiva. Aquest és un exemple més que a partir d'estudis de metagenòmica de genomes complets es poden arribar a identificar noves espècies, les quals poden ser posteriorment caracteritzades per altres mitjans. Una alternativa al fet de no disposar d'una regió conservada és la que es va aplicar en l'estudi de les mostres d'aigua residual (Annex, article 3). En aquest cas, a banda de caracteritzar el viroma complet d'aquestes mostres, hi havia un interès específic per la família *Adenoviridae*. En el moment del disseny experimental es va decidir realitzar, paral·lelament a la metagenòmica de *whole genome*, un *run* basat en la tècnica 454 de Roche amb metagenòmica dirigida, aplicant encebadors específics de la família *Adenoviridae* per la regió que es fa servir per tipificar aquesta família. Com a conseqüència, els *reads* obtinguts eren majoritàriament de la família *Adenoviridae*, i en base a aquestes seqüències es va poder construir un arbre filogenètic molt detallat per aquesta família.

7.7 Millores futures en els protocols

Un dels punts febles del nostre protocol és l'anotació taxonòmica de les seqüències, ja que en darrer terme depenem de la completitud de les bases de dades. Els falsos negatius, és a dir, les seqüències virals que no són anotades com a tal, potser no són tan greus si volem centrar-nos en l'especificitat del protocol. La fiabilitat de les anotacions que estan disponibles avui en dia i quina seria la base de dades més òptima a utilitzar poden ser factors molt rellevants, però que no

podem controlar al ser externs. Amb l'objectiu de poder caracteritzar millor les mostres s'han proposat aproximacions basades en etiquetes curtes d'ADN, els anomenats *"DNA barcodes"*, emprats com a marcadors genètics propis de cada espècie, el que ens permetria, en teoria, determinar la composició dels organismes que es troben en una mostra (Hajibabaei *et al.*, 2007). Aplicar el *"DNA barcoding"* a la metagenòmica de virus podria obrir perspectives interessants, tal i com s'ha suggerit en algun estudi (Chakraborty *et al.*, 2014). El principal problema que hauria de superar en aquesta aproximació és el de la cobertura variable o parcial que s'obté al generar *reads* curts, els quals també són més dificils d'ensamblar (Freitas *et al.*, 2015).

No es agosarat predir que en el moment en què la nova generació de mètodes de seqüenciació *single molecule* es vagin popularitzan, sigui més fàcil aconseguir fragments genòmics més llargs a partir de les mostres. Això resultarà en millors ensamblats i una millor caracterització dels grups taxonòmics sobre les seqüències genòmiques. A mesura que els costos de seqüenciació s'abarateixin més, també s'obriran les portes a la caracterització de mostres en temps real, el que segur revertirà en una millor i major capacitat per fer front a la detecció d'organismes patògens i la prevenció de brots pandèmics.

Conclusions

8 Conclusions

El desenvolupament dels objectius plantejats en aquesta tessi han donat lloc a una sèrie de resultats que han estat publicats o estan en fase de publicació, tal i com es pot comprovar en els annexos corresponents. Les principals conclusions de la tesi es detallen a continuació:

1. S'ha desenvolupat un protocol bioinformàtic per tractar les seqüències generades per experiments de seqüenciació en massa abans de l'ensamblat. El punt clau ha estat eliminar seqüències que poden alentir, dificultar i/o introduir errors en el moment d'aplicar els programes específics d'ensamblat.

2. S'ha estudiat en detall un tipus particular de seqüències que fan més difícil la tasca dels ensambladors. Són les seqüències de baixa complexitat, sovint associades a seqüències repetitives. S'han desenvolupat eines i filtres per poder eliminar aquest tipus de seqüències en la fase de pre-ensamblat.

3. S'ha desenvolupat un programa per eliminar seqüències redundants. Aquest pas facilita el procés d'ensamblar, ja que disminuint el número de seqüències s'agilitza l'ensamblat sense pèrdua d'informació a nivell de seqüència.

4. S'ha observat que el `CLCBio` i el `Velvet-MetaVelvet` ens donaven els millors resultats per ensamblar seqüències quan es tracta de mostres obtingudes per metodologies *Whole Genome Sequencing*.

5. S'ha desenvolupat un protocol automatitzat, en el que s'han ajustat també els paràmetres dels programes que el composen. Com a conseqüència, agilitza tot el procés i minimitza l'error humà.

6. S'ha treballat en el desenvolupament d'eines bioinformàtiques, així com la definició dels criteris necessaris per l'anotació taxonòmica a nivell de família, espècies i tipus vírics, de les seqüències ensamblades.

7. S'han desenvolupat i implementat diferents eines, com les que es generen gràfics tipus *Krona*, i les taules basades en pàgines web interactives, perquè l'usuari pugui accedir i visualitzar tots els resultats de cada una de les mostres de manera fàcil i eficient.

8. S'ha caracteritzat el metaviroma de l'aigua residual urbana, identificant més de 36 famílies, entre les que cal destacar algunes d'elles que poden afectar a l'home.

9. S'ha identificat en mostres de sèrum de pacients amb hepatitis aguda sense agent causal identificat, virus pertanyents a les famílies de *Calicivirus*, *Astrovirus* i *Anellovirus*

10. S'ha definit un protocol d'elevada eficiència per aplicar les tècniques de la metagenòmica a l'estudi de virus contaminants en aigua i aliments.

11. S'han identificat patògens humans en aigua de riu i en vegetals cultivats experimentalment, mostrant el risc elevat que representa la utilització d'aigua amb contaminació viral en irrigació de productes frescos que es consumeixen crus.

Bibliografia

Bibliografia

A

Allen, Heather K., John Bunge, James A. Foster, Darrell O. Bayles i Thaddeus B. Stanton (2013). "Estimation of viral richness from shotgun metagenomes using a frequency count approach". *Microbiome* **1**[1]: 5.p1 - 7.

AllSeq (2015). "Sequencing-platforms, 454". `http://allseq.com/knowledge-bank/sequencing-platforms/454-roche/`.

Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers i David J. Lipman (1990). "Basic local alignment search tool". *Journal of Molecular Biology* **215**[3]: 403 - 410.

Altschul, Stephen F., Thomas L. Madden, Aejandro A Schäffer, Jinghui Zhang, Zheng Zhang *et al.* (1997). "Gapped `BLAST` and `PSI-BLAST`: a new generation of protein database search programs". *Nucleic Acids Research* **25**[17]: 3389 - 3402.

Ammar, Ron, Tara A. Paton, Dax Torti, Adam Shlien i Gary D. Bader (2015). "Long read nanopore sequencing for detection of *HLA* and *CYP2D6* variants and haplotypes". *F1000–Research* **4**, 17.p1 - 19.

Anderson, Stephen (1981). "Shotgun DNA sequencing using cloned DNase I-generated fragments". *Nucleic Acids Research* **9**[13]: 3015 - 3027.

Angly, Florent E., Ben Felts, Mya Breitbart, Peter Salamon, Robert A. Edwards *et al.* (2006). "The Marine Viromes of Four Oceanic Regions". *PLoS Biology* **4**[11]: e368.p2121 - 2131.

Atmar, Robert L. i Mary K. Estes (2001). "Diagnosis of Noncultivatable Gastroenteritis Viruses, the Human Caliciviruses". *Clinical Microbiology Reviews* **14**[1]: 15 - 37.

B

Baltimore, David (1971). "Expression of animal virus genomes". *Bacteriological Reviews* **35**[3]: 235 - 241.

Batovska, Jana, Stacey E. Lynch, Noel O. I. Cogan, Karen Brown, Jonathan M. Darbro *et al.* (2017). "Effective mosquito and arbovirus surveillance using metabarcoding". *Molecular Ecology Resources* **00**, 1 - 9.

Benson, Dennis A., Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J. Lipman *et al.* (2017). "`GenBank`". *Nucleic Acids Research* **45**[DataBases Special Issue Suppl.1]: D37 - D42.

Bentley, David R., Shankar Balasubramanian, Harold P. Swerdlow, Geoffrey P. Smith, John Milton *et al.* (2008). "Accurate whole human genome sequencing using reversible terminator chemistry". *Nature* **456**[7218]: 53 - 59.

Bolger, Anthony M., Marc Lohse i Bjoern Usadel (2014). "`Trimmomatic`: a flexible trimmer for Illumina sequence data". *Bioinformatics* **30**[15]: 2114 - 2120.

Brakmann, Susanne (2010). "Single-molecule analysis: A ribosome in action". *Nature* **464**[7291]: 987 - 988.

Brussaard, Corina P.D., Steven M. Short, Cindy M. Frederickson i Curtis A. Suttle (2004). "Isolation and phylogenetic analysis of novel viruses infecting the phytoplankton *Phaeocystis globosa* (Prymnesiophyceae)". *Applied and Environmental Microbiology* **70**[6]: 3700 - 3705.

Bryant, Douglas W., Weng-Keen Wong i Todd C. Mockler (2009). "`QSRA`: a quality-value guided *de novo* short read assembler". *BMC Bioinformatics* **10**[1]: 69.p1 - 6.

Bu, Rong, Abdul K. Siraj, Khadija A.S. Al-Obaisi, Shaham Beg, Mohsen Al Hazmi *et al.* (2016). "Identification of novel *BRCA* founder mutations in Middle Eastern breast cancer patients using capture and Sanger sequencing analysis". *International Journal of Cancer* **139**[5]: 1091 - 1097.

Bunge, John (2011). "Estimating the number of species with `CatchAll`". *Pacific Symposium on Biocomputing*, pp.121 - 130.

Bunge, John, Linda Woodard, Dankmar Böhning, James A. Foster, Sean Connolly *et al.* (2012). "Estimating population diversity with `CatchAll`". *Bioinformatics* **28**[7]: 1045 - 1047.

Burke, Catherine M. i Aaron E. Darling (2016). "A method for high precision sequencing of near full-length 16S rRNA genes on an Illumina MiSeq". *PeerJ* **4**, e2492.p1 - 20.

Butler, Jonathan, Iain MacCallum, Michael Kleber, Ilya A. Shlyakhter, Matthew K. Belmonte *et al.* (2008). "`ALLPATHS`: *De novo* assembly of whole-genome shotgun microreads". *Genome Research* **18**[5]: 810 - 820.

C

Calgua, Byron, Anett Mengewein, A. Grunert, Silvia Bofill-Mas, Pilar Clemente-Casares *et al.* (2008). "Development and application of a one-step low cost procedure to concentrate viruses from seawater samples". *Journal of Virological Methods* **153**[2]: 79 - 83.

Calgua, Byron, Tulio Fumian, Marta Rusiñol, Jesus Rodriguez-Manzano, Viviana Mbayed *et al.* (2013). "Detection and quantification of classic and emerging viruses by skimmed-milk flocculation and PCR in river water from two geographical areas". *Water Research* **47**, 2797 - 2810.

Canard, Brunno i Robert S. Sarfati (1994). "DNA polymerase fluorescent substrates with reversible 3'-tags". *Gene* **148**[1]: 1 - 6.

Carter, Michael J. (2005). "Enterically infecting viruses: pathogenicity, transmission and significance for food and waterborne infection". *Journal of Applied Microbiology* **98**[6]: 1354 - 1380.

Carvalho, Carlos F.A., Helen L. Thomas, Koye Balogun, Richard S. Tedder, Richard G. Pebody *et al.* (2012). "A possible outbreak of hepatitis A associated with semi-dried tomatoes, England, July-November 2011". *Euro surveillance: Bulletin Europeen sur les maladies transmissibles. European communicable disease bulletin* **17**[6]: pii=20083.p1 - 4.

Castresana, Jose (2000). "Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis". *Molecular Biology and Evolution* **17**[4]: 540 - 552.

Chaisson, Mark J. i Pavel A. Pevzner (2008). "Short read fragment assembly of bacterial genomes". *Genome Research* **18**[2]: 324 - 330.

Chaisson, Mark J., Pavel A. Pevzner i Haixu Tang (2004). "Fragment assembly with short reads". *Bioinformatics* **20**[13]: 2067 - 2074.

Chakraborty, Chiranjib, C. George Priya Doss, Bidhan C. Patra i Sanghamitra Bandyopadhyay (2014). "DNA barcoding to map the microbial communities: current advances and future directions". *Applied Microbiology and Biotechnology* **98**[8]: 3425 - 3436.

Chang, Shuai, Shuo Huang, Jin He, Feng Liang, Peiming Zhang *et al.* (2010). "Electronic Signatures of all Four DNA Nucleosides in a Tunneling Gap". *Nano Letters* **10**[3]: 1070 - 1075.

Chao, Anne i Shen-Ming Lee (1992). "Estimating the number of classes via sample coverage". *Journal of the American Statistical Association* **87**[417]: 210 - 217.

Chen, Kevin i Lior Pachter (2005). "Bioinformatics for whole-genome shotgun sequencing of microbial communities". *PLoS Computational Biology* **1**[2]: 106 - 112.

Chidgeavadze, Zurab G., Robert S. Beabealashvilli, Alexey M. Atrazhev, Marina K. Kukhanova, Alexey V. Azhayev *et al.* (1984). "2',3'–Dideoxy–3' aminonucleoside 5'–triphosphates are the terminators of DNA synthesis catalyzed by DNA polymerases". *Nucleic Acids Research* **12**[3]: 1671 - 1686.

Cock, Peter J.A., Christopher J. Fields, Naohisa Goto, Michael L. Heuer i Peter M. Rice (2010). "The Sanger `FASTQ` file format for sequences with quality scores, and the Solexa/Illumina `FASTQ` variants". *Nucleic Acids Research* **38**[6]: 1767 - 1771.

Combosch, David J., Timothy M. Collins, Emily A. Glover, Daniel L. Graf, Elizabeth M. Harper *et al.* (2017). "A family-level Tree of Life for bivalves based on a Sanger-sequencing approach". *Molecular Phylogenetics and Evolution* **107**, 191 - 208.

Confalonieri, Sara (2014). "Wikimedia Commons: Wikipedia Virus classification". http:///en.wikipedia.org/wiki/Virus_classification.

Cornelissen-Keijsers, Vivian, Alexandra Jiménez-Melsió, Denny Sonnemans, Martí Cortey, Joaquim Segalés *et al.* (2012). "Discovery of a novel Torque teno sus virus species: genetic characterization, epidemiological assessment and disease association". *Journal of General Virology* **93**[Pt12]: 2682 - 2691.

Corrada Bravo, Hector, Florin Chelaru, Justin Wagner, Jayaram Kancherla i Joseph N. Paulson (2017). "metavizr: R Interface to the metaviz web app for interactive metagenomics data analysis and visualization". *GitHub repository*. https://github.com/epiviz/metavizr.

Costa, Patrícia S., Mariana P. Reis, Marcelo P. Ávila, Laura R. Leite, Flávio M.G. de Araújo *et al.* (2015). "Metagenome of a Microbial Community Inhabiting a Metal-Rich Tropical Stream Sediment". *PLoS ONE* **10**[3]: e0119465.

Cronn, Richard, Aaron Liston, Matthew Parks, David S. Gernandt, Rongkun Shen *et al.* (2008). "Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology". *Nucleic Acids Research* **36**[19]: e122 - e122.

D

da Silva, Ennyo S.C., Magda Feres, Luciene C. Figueiredo, Jamil A. Shibli, Fernanda S. Ramiro *et al.* (2014). "Microbiological diversity of peri-implantitis biofilm by Sanger sequencing". *Clinical Oral Implants Research* **25**[10]: 1192 - 1199.

Dawson, Paul A., Pearl Sim, David W. Mudge i David Cowley (2013). "Human *SLC26A1* Gene Variants: A Pilot Study". *The Scientific World Journal* **2013**, 1 - 7.

dela Torre, Ruby, Joseph Larkin, Alon Singer i Amit Meller (2012). "Fabrication and characterization of solid-state nanopore arrays for high-throughput DNA sequencing". *Nanotechnology* **23**[38]: 385308.p1 - 6.

DiGuistini, Scott, Nancy Y. Liao, Darren Platt, Gordon Robertson, Michael Seidel *et al.* (2009). "*De novo* genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data". *Genome Biology* **10**[9]: R94.1 - 12.

Dinsdale, Elizabeth A., Robert A. Edwards, Dana Hall, Florent Angly, Mya Breitbart *et al.* (2008). "Functional metagenomic profiling of nine biomes". *Nature* **452**[7187]: 629 - 632.

Dohm, Juliane C., Claudio Lottaz, Tatiana Borodina i Heinz Himmelbauer (2007). "SHARCGS, a fast and highly accurate short-read assembly algorithm for *de novo* genomic sequencing". *Genome Research* **17**[11]: 1697 - 1706.

Dong, Yanhong, Jinheung Kim i Gillian D. Lewis (2010). "Evaluation of methodology for detection of human adenoviruses in wastewater, drinking water, stream water and recreational waters". *Journal of Applied Microbiology* **108**[3]: 800 - 809.

DuBois, Paul (2005). *MySQL: The Definitive Guide to Using, Programming, and Administering MySQL 4.1 and 5.0*. 3^{rd} ed. SAMS Publishing.

E

Eid, John, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle *et al.* (2009). "Real-Time DNA Sequencing from Single Polymerase Molecules". *Science* **323**[5910]: 133 - 138.

Eisen, Jonathan A. (2007). "Environmental shotgun sequencing: its potential and challenges for studying the hidden world of microbes". *PLoS Biology* **5**[3]: e82.0384 - 0388.

El-Metwally, Sara, Taher Hamza, Magdi Zakaria i Mohamed Helmy (2013). "Next-Generation Sequence Assembly: Four Stages of Data Processing and Computational Challenges". *PLoS Computational Biology* **9**[12]: e1003345.1 - 19.

Ewing, Brent i Phil Green (1998). "Base-calling of automated sequencer traces using `phred`. II. Error probabilities." *Genome Research* **8**[3]: 186 - 194.

F

Farrelly, Vincent, Frederick A. Rainey i Erko Stackebrandt (1995). "Effect of genome size and *rrn* gene copy number on PCR amplification of 16S rRNA genes from a mixture of bacterial species". *Applied and Environmental Microbiology* **61**[7]: 2798 - 2801.

Fernàndez Cassi, Xavi (2017). "Aplicació de tècniques se seqüenciació massiva a l'estudi de virus potencialment contaminants d'aigües i aliments". Tesi doct. Universitat de Barcelona.

Fournet, Nelly, Dominique C. Baas, Wilfrid van Pelt, Corien M. Swaan, H.J. Ober *et al.* (2012). "Another possible food-borne outbreak of hepatitis A in the Netherlands indicated by two closely related molecular sequences, July to October 2011". *Euro surveillance: Bulletin Europeen sur les maladies transmissibles. European communicable disease bulletin* **17**[6]: pii=20079.p1 - 3.

Freitas, Tracey Allen K., Po-E Li, Matthew B. Scholz i Patrick S.G. Chain (2015). "Accurate read-based metagenome characterization using a hierarchical suite of unique signatures". *Nucleic Acids Research* **43**[10]: e69 - e69.

Fu, Limin, Beifang Niu, Zhengwei Zhu, Sitao Wu i Weizhong Li (2012). "CD-HIT: accelerated for clustering the next-generation sequencing data". *Bioinformatics* **28**[23]: 3150 - 3152.

G

Gabrielian, Andrei i Alexander Bolshoy (1999). "Sequence complexity and DNA curvature". *Computers & Chemistry* **23**[3-4]: 263 - 274.

GenomeWeb (2009a). "PacBio Reveals Beta System Specs for RS". https://www.genomeweb.com/sequencing/pacbios-q3-losses-widen-it-readies-launch-rs-system.

– (2009b). "PacBio Ships First Two Commercial Systems". https://www.genomeweb.com/sequencing/pacbio-ships-first-two-commercial-systems-order-backlog-grows-44.

Genomics (2015). "Lloc web de Genomics". http://www.genomics.cn/en.

Girones, Rosina, Maria Antonia Ferrús, José Luis Alonso, Jesus Rodriguez-Manzano, Byron Calgua *et al.* (2010). "Molecular detection of pathogens in water—The pros and cons of molecular techniques". *Water Research* **44**[15]: 4325 - 4339.

Gnerre, Sante, Iain Maccallum, Dariusz Przybylski, Filipe J. Ribeiro, Joshua N. Burton *et al.* (2011). "High-quality draft assemblies of mammalian genomes from massively parallel sequence data". *Proceedings of the National Academy of Sciences of the United States of America (PNAS)* **108**[4]: 1513 - 1518.

Gonnella, Giorgio i Stefan Kurtz (2012). "Readjoiner: a fast and memory efficient string graph-based sequence assembler". *BMC Bioinformatics* **13**[1]: 82.1 - 19.

Gordon, David, Chris Abajian i Phil Green (1998). "Consed: a graphical tool for sequence finishing". *Genome Research* **8**[3]: 195 - 202.

Gordon, David, Cindy Desmarais i Phil Green (2001). "Automated finishing with autofinish". *Genome Research* **11**[4]: 614 - 625.

Greninger, Alexander L., Samia N. Naccache, Scot Federman, Guixia Yu, Placide Mbala *et al.* (2015). "Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis". *Genome Medicine* **7**[1]: 99.1 - 13.

Guerrero-Latorre, Laura (2016). "Estudis sobre la contaminació i desinfecció de virus entèrics en contexts d'ajuda humanitària". Tesi doct. Universitat de Barcelona.

H

Hadley, Wickham (2009). *ggplot2: elegant graphics for data analysis*. 1^{st} ed. Springer New York.

Hajibabaei, Mehrdad, Gregory A.C. Singer, Paul D.N. Hebert i Donal A. Hickey (2007). "DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics". *Trends in Genetics* **23**[4]: 167 - 172.

Hall, Richard J., Jing Wang, Angela K. Todd, Ange B. Bissielo, Seiha Yen *et al.* (2014). "Evaluation of rapid and simple techniques for the enrichment of viruses prior to metagenomic virus discovery". *Journal of Virological Methods* **195**, 194 - 204.

Hamza, Ibrahim Ahmed, Lars Jurzik, Klaus Überla i Michael Wilhelm (2011). "Methods to detect infectious human enteric viruses in environmental water samples". *International Journal of Hygiene and Environmental Health* **214**[6]: 424 - 436.

Hancock, John M. (2002). "Genome size and the accumulation of simple sequence repeats: implications of new data from genome sequencing projects". *Genetica* **115**[1]: 93 - 103.

Handelsman, Jo, Michelle R. Rondon, Sean F. Brady, Jon Clardy i Robert M. Goodman (1998). "Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products". *Chemistry & Biology* **5**[10]: R245 - R249.

Hannon Lab (2010). "`FASTX-toolkit`: `FASTQ/A` short-reads pre-processing tools". `http://www.hannonlab.org`.

Hansen, Martin Christian, Tim Tolker-Nielsen, Michael Givskov i Søren Molin (1998). "Biased 16S rDNA PCR amplification caused by interference from DNA flanking the template region". *FEMS Microbiology Ecology* **26**[2]: 141 - 149.

Heid, Christian A., Junko Stevens, Kenneth J. Livak i P. Mikey Williams (1996). "Real time quantitative PCR". *Genome Research* **6**[10]: 986 - 994.

Hernandez, David, Patrice François, Laurent Farinelli, Magne Østerås i Jacques Schrenzel (2008). "*De novo* bacterial genome sequencing: Millions of very short reads assembled on a desktop computer". *Genome Research* **18**[5]: 802 - 809.

Hernroth, Bodil E., Ann-Christine Conden-Hansson, Ann-Sofi Rehnstam-Holm, Rosina Girones i Annika K. Allard (2002). "Environmental factors influencing human viral pathogens and their potential indicator organisms in the blue mussel, *Mytilus edulis*: the first Scandinavian report". *Applied and Environmental Microbiology* **68**[9]: 4523 - 4533.

Hossain, Mohammad, Navid Azimi i Steven Skiena (2009). "Crystallizing short-read assemblies around seeds". *BMC Bioinformatics* **10**[Suppl 1]: S16.1 - 12.

Huang, Yu-Feng, Sheng-Chung Chen, Yih-Shien Chiang, Tzu-Han Chen i Kuo-Ping Chiu (2012). "Palindromic sequence impedes sequencing-by-ligation mechanism". *BMC Systems Biology* **6**[Suppl 2]: S10.1 - 7.

Huber, Wolfgang, Vincent J. Carey, Robert Gentleman, Simon Anders, Marc Carlson *et al.* (2015). "Orchestrating high-throughput genomic analysis with `Bioconductor`". *Nature Methods* **12**[2]: 115 - 121.

Hugenholtz, Philip, Brett M. Goebel i Norman R. Pace (1998). "Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity". *Journal of Bacteriology* **180**[18]: 4765 - 4774.

Hurwitz, Bonnie L., Li Deng, Bonnie T. Poulos i Matthew B. Sullivan (2013). "Evaluation of methods to concentrate and purify ocean virus communities through comparative, replicated metagenomics". *Environmental Microbiology* **15**[5]: 1428 - 1440.

I

Idury, Ramana M. i Michael S. Waterman (1995). "A New Algorithm for DNA Sequence Assembly". *Journal of Computational Biology* **2**[2]: 291 - 306.

Illumina (2017). "Lloc web d'`Illumina`". `https://www.illumina.com`.

Irene, Rocchetti, Bunge John i Böhning Dankmar (2011). "Population size estimation based upon ratios of recapture probabilities". *The Annals of Applied Statistics* **5**[1]: 1512 - 1533.

J

Jeck, William R., Josephine A. Reinhardt, David A. Baltrus, Matthew T. Hickenbotham, Vincent Magrini *et al.* (2007). "Extending assembly of short DNA sequences to handle error". *Bioinformatics* **23**[21]: 2942 - 2944.

K

Kalyuzhnaya, Marina G., Alla Lapidus, Natalia Ivanova, Alex C. Copeland, Alice C. McHardy *et al.* (2008). "High-resolution metagenomics targets specific functional types in complex microbial communities". *Nature Biotechnology* **26**[9]: 1029 - 1034.

Kanda, Kojun, James M. Pflug, John S. Sproul, Mark A. Dasenko i David R. Maddison (2015). "Successful Recovery of Nuclear Protein-Coding Genes from Small Insects in Museums Using Illumina Sequencing". *PloS ONE* **10**[12]: e0143929.1 - 53.

Karow, Julia (2013). "Following Roche's decision to shut down 454, customers make plans to move to other platforms". `https://www.genomeweb.com/sequencing/following-roches-decision-shut-down-454-customers-make-plans-move-other-platform`.

Katsnelson, Alla (2010). "Nature News: DNA sequencing for the masses". *Nature*. `http://www.nature.com/doifinder/10.1038/news.2010.674`.

Kim, Hanna, Henry A. Erlich i Cassandra D. Calloway (2015). "Analysis of mixtures using next generation sequencing of mitochondrial DNA hypervariable regions". *Croatian Medical Journal* **56**[3]: 208 - 217.

Kim, Nan-Ok, Hae-young Na, Su-Mi Jung, Gyung Tae Chung, Hyo Sun Kawk *et al.* (2017). "Genome Sequencing Analysis of Atypical *Shigella flexneri* Isolated in Korea". *Osong Public Health and Research Perspectives* **8**[1]: 78 - 85.

Knight, Rob, Janet Jansson, Dawn Field, Noah Fierer, Narayan Desai *et al.* (2012). "Unlocking the potential of metagenomics through replicated experimental design". *Nature Biotechnology* **30**[6]: 513 - 520.

Koopmans, Marion i Erwin Duizer (2004). "Foodborne viruses: an emerging problem". *International Journal of Food Microbiology* **90**[1]: 23 - 41.

Koroglu, Mehmet, Yusuf Yakupogullari, Baris Otlu, Serhat Ozturk, Mehmet Ozden *et al.* (2011). "A waterborne outbreak of epidemic diarrhea due to group A rotavirus in Malatya, Turkey". *The New Microbiologica* **34**[1]: 17 - 24.

Kumar Das, Bikram (2014). "`SlideShare`: Viral Taxonomy". `https://www.slideshare.net/bikramkdas/viral-taxonomy`.

L

Langford, Malcolm (2005). "The United Nations Concept of Water as a Human Right: A New Paradigm for Old Problems?" *International Journal of Water Resources Development* **21**[2]: 273 - 282.

Langmead, Ben i Steven L. Salzberg (2012). "Fast gapped-read alignment with `Bowtie 2`". *Nature Methods* **9**[4]: 357 - 359.

Letunic, Ivica i Peer Bork (2016). "Interactive tree of life (`iTOL`) v3: an online tool for the display and annotation of phylogenetic and other trees". *Nucleic Acids Research* **44**[Web Servers Special Issue Suppl.2]: W242 - W245.

Levene, Michael J., Jonas Korlach, Stephen W. Turner, Mathieu Foquet, Harold G. Craighead *et al.* (2003). "Zero-Mode Waveguides for Single-Molecule Analysis at High Concentrations". *Science* **299**[5607]: 682 - 686.

Li, Heng (2012). "Exploring single-sample SNP and INDEL calling with whole-genome *de novo* assembly". *Bioinformatics* **28**[14]: 1838 - 1844.

Lim, Yan Wei, Daniel A. Cuevas, Genivaldo Gueiros Z. Silva, Kristen Aguinaldo, Elizabeth A. Dinsdale *et al.* (2014). "Sequencing at sea: challenges and expe-

riences in Ion Torrent PGM sequencing during the 2013 Southern Line Islands Research Expedition". *PeerJ* **2**, e520.1 - 23.

Loman, Nicholas J., Mick Watson, Scot Federman, Guixia Yu, Placide Mbala *et al.* (2015). "Successful test launch for nanopore sequencing". *Nature Methods* **12**[4]: 303 - 304.

Luo, Ruibang, Binghang Liu, Yinlong Xie, Zhenyu Li, Weihua Huang *et al.* (2012). "SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler". *GigaScience* **1**[1]: 18.1 - 6.

Lysholm, Fredrik, Anna Wetterbom, Cecilia Lindau, Hamid Darban, Annelie Bjerkner *et al.* (2012). "Characterization of the Viral Microbiome in Patients with Severe Lower Respiratory Tract Infections, Using Metagenomic Sequencing". *PLoS ONE* **7**[2]: e30875.1 - 12.

M

Maccallum, Iain, Dariusz Przybylski, Sante Gnerre, Joshua Burton, Ilya Shlyakhter *et al.* (2009). "ALLPATHS 2: small genomes assembled accurately and with high continuity from short paired reads". *Genome Biology* **10**[10]: R103.1 - 10.

Manimaran, Solaiappan, Matthew Bendall, Sandro Valenzuela Diaz, Eduardo Castro, Tyler Faits *et al.* (2016). "PathoStat: PathoScope Statistical Microbiome Analysis Package". *GitHub repository*. https://github.com/mani2012/PathoStat.

Margulies, Marcel, Michael Egholm, William E. Altman, Said Attiya, Joel S. Bader *et al.* (2005). "Genome sequencing in microfabricated high-density picolitre reactors". *Nature* **437**[7057]: 376 - 380.

Martin, Marcel (2011). "Cutadapt removes adapter sequences from high-throughput sequencing reads". *EMBnet Journal* **17**[1]: 10 - 12.

Maxam, Allan M. i Walter Gilbert (1977). "A new method for sequencing DNA". *Proceedings of the National Academy of Sciences of the United States of America (PNAS)* **74**[2]: 560 - 564.

Mayer, P., L. Farinelli, G. Matton, C. Adessi, G. Turcatti *et al.* (1998). "DNA colony massively parallel sequencing **ams98** presentation: A very large scale, high throughput and low cost DNA sequencing method based on a new 2-dimensional DNA auto-patterning process". *Fifth International Automation in Mapping and DNA Sequencing Conference, St. Louis, MO, USA*.

McNally, Ben, Alon Singer, Zhiliang Yu, Yingjie Sun, Zhiping Weng *et al.* (2010). "Optical recognition of converted DNA nucleotides for single-molecule DNA sequencing using nanopore arrays". *Nano Letters* **10**[6]: 2237 - 2244.

Medini, Duccio, Davide Serruto, Julian Parkhill, David A. Relman, Claudio Donati *et al.* (2008). "Microbiology in the post-genomic era". *Nature Reviews Microbiology* **6**[6]: 419 - 430.

Mellmann, Alexander, Dag Harmsen, Craig A. Cummings, Emily B. Zentz, Shana R. Leopold *et al.* (2011). "Prospective genomic characterization of the German enterohemorrhagic *Escherichia coli* O104:H4 outbreak by rapid next generation sequencing technology". *PloS ONE* **6**[7]: e22751.1 - 9.

Methé, Barbara A., Karen E. Nelson, Mihai Pop, Heather H. Creasy, Michelle G. Giglio *et al.* (2012). "A framework for human microbiome research". *Nature* **486**[7402]: 215 - 221.

Meyer, Folker, Daniel Paarmann, Mark D'Souza, Robert Olson, Elizabeth M. Glass *et al.* (2008). "The metagenomics RAST server—A public resource for the automatic phylogenetic and functional analysis of metagenomes". *BMC Bioinformatics* **9**, 386.1 - 8.

Miller, Jason R., Arthur L. Delcher, Sergey Koren, Eli Venter, Brian P. Walenz *et al.* (2008). "Aggressive assembly of pyrosequencing reads with mates". *Bioinformatics* **24**[24]: 2818 - 2824.

Miller, Jason R., Sergey Koren i Granger Sutton (2010). "Assembly algorithms for next-generation sequencing data". *Genomics* **95**[6]: 315 - 327.

Mitchell, Alex, Francois Bucchini, Guy Cochrane, Hubert Denise, Petra ten Hoopen *et al.* (2016). "EBI metagenomics in 2016—An expanding and evolving resource for the analysis and archiving of metagenomic data". *Nucleic Acids Research* **44**[DataBases Special Issue Suppl.1]: D595 - D603.

Morgan, Gregory J. (2016). "What is a virus species? Radical pluralism in viral taxonomy". *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* **59**, 64 - 70.

Morgan, Jenna L., Aaron E. Darling i Jonathan A. Eisen (2010). "Metagenomic Sequencing of an *In Vitro*-Simulated Microbial Community". *PLoS ONE* **5**[4]: e10209.1 - 10.

Moskalev, Evgeny A., Robert Stöhr, Ralf Rieker, Simone Hebele, Florian Fuchs *et al.* (2013). "Increased detection rates of *EGFR* and *KRAS* mutations in NSCLC specimens with low tumour cell content by 454 deep sequencing". *Virchows Archiv* **462**[4]: 409 - 419.

Moustafa, Ahmed, Chao Xie, Ewen Kirkness, William Biggs, Emily Wong *et al.* (2017). "The blood DNA virome in 8,000 humans". *PLoS Pathogens* **13**[3]: e1006292.1 - 20.

Müller, Luise, Lasse Dam Rasmussen, Tenna Jensen, Anna Charlotte Schultz, Charlotte Kjelsø *et al.* (2016). "Series of Norovirus Outbreaks Caused by Consumption of Green Coral Lettuce, Denmark, April 2016". *PLOS Currents Outbreaks*. `http://currents.plos.org/outbreaks/article/series-of-norovirus-outbreaks-caused-by-consumption-of-green-coral-lettuce-denmark-april-2016/`.

Murray, Iain A., Tyson A. Clark, Richard D. Morgan, Matthew Boitano, Brian P. Anton *et al.* (2012). "The methylomes of six bacteria". *Nucleic Acids Research* **40**[22]: 11450 - 11462.

Myers, Eugene W. (1995). "Toward Simplifying and Accurately Formulating Fragment Assembly". *Journal of Computational Biology* **2**[2]: 275 - 290.

N

Nanoporetech (2017). "Lloc web de `Nanoporetech`". `http://www.nanoporetech.com`.

NCBI Resource Coordinators (2013). "Database resources of the National Center for Biotechnology Information". *Nucleic Acids Research* **41**[DataBases Special Issue Suppl.1]: D8 - D20.

Nenonen, Nancy P., Charles Hannoun, Chralotte U. Larsson i Tomas Bergstrom (2012). "Marked Genomic Diversity of Norovirus Genogroup I Strains in a Waterborne Outbreak". *Applied and Environmental Microbiology* **78**[6]: 1846 - 1852.

Newell, Diane G., Marion Koopmans, Linda Verhoef, Erwin Duizer, Awa Aidara-Kane *et al.* (2010). "Food-borne diseases—The challenges of 20 years ago still persist while new ones continue to emerge". *International Journal of Food Microbiology* **139**, S3 - S15.

Ng, Terry Fei Fan, Rachel Marine, Chunlin Wang, Peter Simmonds, Beatrix Kapusinszky *et al.* (2012). "High variety of known and new RNA and DNA viruses of diverse origins in untreated sewage". *Journal of Virology* **86**[22]: 12161 - 12175.

Niedringhaus, Thomas P., Denitsa Milanova, Matthew B. Kerby, Michael P. Snyder i Annelise E. Barron (2011). "Landscape of next-generation sequencing technologies". *Analytical Chemistry* **83**[12]: 4327 - 4341.

O

Ogorzaly, Leslie, Cécile Walczak, Mélissa Galloux, Stéphanie Etienne, Benoît Gassilloud *et al.* (2015). "Human Adenovirus Diversity in Water Samples Using a Next-Generation Amplicon Sequencing Approach". *Food and Environmental Virology* **7**[2]: 112 - 121.

Orlov, Yury L. i Vladimir N. Potapov (2004). "`Complexity`: an internet resource for analysis of DNA sequence complexity". *Nucleic Acids Research* **32**[Web Servers Special Issue Suppl.2]: W628 - W633.

Oude Munnink, Bas B., Seyed Mohammad Jazaeri Farsani, Martin Deijs, Jiri Jonkers, Joost T.P. Verhoeven *et al.* (2013). "Autologous Antibody Capture to Enrich Immunogenic Viruses for Viral Discovery". *PLoS ONE* **8**[11]: e78454.1 - 6.

Oulas, Anastasis, Christina Pavloudi, Paraskevi Polymenakou, Georgios A. Pavlopoulos, Nikolas Papanikolaou *et al.* (2015). "Metagenomics: tools and insights for analyzing next-generation sequencing data derived from biodiversity studies". *Bioinformatics and Biology Insights* **9**, 75 - 88.

P

Paez-Espino, David, I-Min A. Chen, Krishna Palaniappan, Anna Ratner, Ken Chu *et al.* (2017). "IMG/VR: a database of cultured and uncultured DNA Viruses and retroviruses". *Nucleic Acids Research* **45**[DataBases Special Issue Suppl.1]: D457 - D465.

Pathak, Biswarup, Henrik Löfås, Jariyanee Prasongkit, Anton Grigoriev, Rajeev Ahuja *et al.* (2012). "Double-functionalized nanopore-embedded gold electrodes for rapid DNA sequencing". *Applied Physics Letters* **100**[2]: 023701.1 - 3.

Pavia, Andrew T. (2011). "Viral Infections of the Lower Respiratory Tract: Old Viruses, New Viruses, and the Role of Diagnosis". *Clinical Infectious Diseases* **52**[Suppl.4]: S284 - S289.

Pearson, Bruce M., Duncan J.H. Gaskin, Ruud P.A.M. Segers, Jerry M. Wells, Piet J.M. Nuijten *et al.* (2007). "The complete genome sequence of *Campylobacter jejuni* strain 81116 (NCTC11828)". *Journal of Bacteriology* **189**[22]: 8402 - 8403.

Peltola, Hannu, Hans Söderlund i Esko Ukkonen (1984). "SEQAID: a DNA sequence assembling program based on a mathematical model". *Nucleic Acids Research* **12**[1]: 307 - 321.

Petersen, Thomas Nordahl, Oksana Lukjancenko, Martin Christen Frølund Thomsen, Maria Maddalena Sperotto, Ole Lund *et al.* (2017). "MGmapper: Reference based mapping and taxonomy annotation of metagenomics sequence reads". *PLoS ONE* **12**[5]: e0176469.1 - 13.

Pevzner, Pavel A., Haixu Tang i Michae S. Waterman (2001). "An Eulerian path approach to DNA fragment assembly". *Proceedings of the National Academy of Sciences of the United States of America (PNAS)* **98**[17]: 9748 - 9753.

Pina, Sonia, Joan Jofre, Suzanne U. Emerson, Robert H. Purcell i Rosina Girones (1998). "Characterization of a strain of infectious hepatitis E virus isolated from sewage in an area where hepatitis E is not endemic". *Applied and Environmental Microbiology* **64**[11]: 4485 - 4488.

Prober, J.M., George L. Trainor, Rudy J. Dam, Frank W. Hobbs, C.W. Robertson *et al.* (1987). "A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides". *Science* **238**[4825]: 336 - 341.

Prosser, James I. (2010). "Replicate or lie". *Environmental Microbiology* **12**[7]: 1806 - 1810.

Q

QIAGEN (2008). "CLC Assembly Cell, mapping reads of next-generation sequencing data". https://www.qiagenbioinformatics.com/products/clc-assembly-cell/.

R

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

Reyes, Alejandro, Matthew Haynes, Nicole Hanson, Florent E. Angly, Andrew C. Heath *et al.* (2010). "Viruses in the fecal microbiota of monozygotic twins and their mothers". *Nature* **466**[7304]: 334 - 338.

RNA-Seq blog (2017). "Lloc web de RNA-Seq-Blog". http://www.rna-seqblog.com.

Rothberg, Jonathan M., Wolfgang Hinz, Todd M. Rearick, Jonathan Schultz, William Mileski *et al.* (2011). "An integrated semiconductor device enabling non-optical genome sequencing". *Nature* **475**[7356]: 348 - 352.

S

Samarakoon, Upeka, Allison Regier, Asako Tan, Brian A. Desany, Brendan Collins *et al.* (2011). "High-throughput 454 resequencing for allele discovery and recombination mapping in *Plasmodium falciparum*". *BMC Genomics* **12**, 116.1 - 14.

Sanger, F, A R Coulson, G F Hong, D F Hill i G B Petersen (1982). "Nucleotide sequence of bacteriophage lambda DNA". *Journal of Molecular Biology* **162**[4]: 729 - 773.

Sanger, Frederick i Alan R. Coulson (1975). "A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase". *Journal of Molecular Biology* **94**[3]: 441 - 446.

Sanger, Frederick, Steve Nicklen i Alan R. Coulson (1977a). "DNA sequencing with chain-terminating inhibitors". *Proceedings of the National Academy of Sciences of the United States of America (PNAS)* **74**[12]: 5463 - 5467.

Sanger, Frederick, Gillian M. Air, Bart G. Barrell, Nigel L. Brown, Alan R. Coulson *et al.* (1977b). "Nucleotide sequence of bacteriophage ϕX174 DNA". *Nature* **265**[5596]: 687 - 95.

Sharon, Donald, Hagen Tilgner, Fabian Grubert i Michael Snyder (2013). "A single-molecule long-read survey of the human transcriptome". *Nature Biotechnology* **31**[11]: 1009 - 1014.

Shendure, Jay, Gregory J. Porreca, Nikos B. Reppas, Xiaoxia Lin, John P. McCutcheon *et al.* (2005). "Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome". *Science* **309**[5741]: 1728 - 1732.

Simpson, Jared T. i Richard Durbin (2012). "Efficient *de novo* assembly of large genomes using compressed data structures". *Genome Research* **22**[3]: 549 - 556.

Simpson, Jared T., Kim Wong, Shaun D. Jackman, Jacqueline E. Schein, Steven J.M. Jones *et al.* (2009). "ABySS: A parallel assembler for short read sequence data". *Genome Research* **19**[6]: 1117 - 1123.

Smith, Lloyd M., Jane Z. Sanders, Robert J. Kaiser, Peter Hughes, Chris Dodd *et al.* (1986). "Fluorescence detection in automated DNA sequence analysis". *Nature* **321**[6071]: 674 - 679.

Stamatakis, Alexandros (2014). "RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies". *Bioinformatics* **30**[9]: 1312 - 1313.

Stearn, William T. (1959). "The Background of Linnaeus's Contributions to the Nomenclature and Methods of Systematic Biology". *Systematic Zoology* **8**[1]: 4 - 22.

Streit, Wolfgang R. i Ruth A. Schmitz (2004). "Metagenomics—The key to the uncultured microbes". *Current Opinion in Microbiology* **7**[5]: 492 - 498.

T

Talavera, Gerard i Jose Castresana (2007). "Improvement of Phylogenies after Removing Divergent and Ambiguously Aligned Blocks from Protein Sequence Alignments". *Systematic Biology* **56**[4]: 564 - 577.

Tarhio, Jorma i Esko Ukkonen (1988). "A greedy approximation algorithm for constructing shortest common superstrings". *Theoretical Computer Science* **57**[1]: 131 - 145.

Teunis, Peter F.M., Christine L. Moe, Pengbo Liu, Sara E. Miller, Lisa Lindesmith *et al.* (2008). "Norwalk virus: How infectious is it?" *Journal of Medical Virology* **80**[8]: 1468 - 1476.

Trifonov, Edward N. (1990). "Human Genome Initiative & DNA Recombination: Proceedings of the 6^{th} Conversation in the Discipline Biomolecular Stereodynamics". Ed. de R.H. Sarma i M.H. Sarma. 1^{st} ed. Vol. 1. Structure & Methods. Albany, New York: Adenine Press. Cap. Making sense of the human genome, pp.69 - 77.

Tyakht, Alexander V., Veronika B. Dubinkina, Vera Y. Odintsova, Konstantin S. Yarygin, Boris A. Kovarsky *et al.* (2017). "Data on gut metagenomes of the patients with alcoholic dependence syndrome and alcoholic liver cirrhosis". *Data in Brief* **11**, 98 - 102.

U

UNAM, Portal académico (2015). "`Universidad Nacional Autónoma de México`: Secuenciación de ADN". `https://tinyurl.com/unam-sanger`.

V

Valouev, Anton, Jeffrey Ichikawa, Thaisan Tonthat, Jeremy Stuart, Swati Ranade *et al.* (2008). "A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning". *Genome Research* **18**[7]: 1051 - 1063.

van den Hoecke, Silvie, Judith Verhelst, Marnik Vuylsteke i Xavier Saelens (2015). "Analysis of the genetic diversity of influenza A viruses using next-generation DNA sequencing". *BMC Genomics* **16**[1]: 79.1 - 23.

van Regenmortel, Marc H.V. i Brian W.J. Mahy (2004). "Emerging issues in virus taxonomy". *Emerging Infectious Diseases* **10**[1]: 8 - 13.

W

Wang, Cheng, Da Dong, Haoshu Wang, Karin Müller, Yong Qin *et al.* (2016). "Metagenomic analysis of microbial consortia enriched from compost: new insights into the role of Actinobacteria in lignocellulose decomposition". *Biotechnology for Biofuels* **9**, 22.1 - 17.

Wang, David, Laurent Coscoy, Maxine Zylberberg, Pedro C. Avila, Homer A. Boushey *et al.* (2002). "Nonlinear partial differential equations and applications: Microarray-based detection and genotyping of viral pathogens". *Proceedings of the National Academy of Sciences of the United States of Americ (PNAS)* **99**[24]: 15687 - 15692.

Wang, David, Anatoly Urisman, Yu-Tsueng Liu, Michael Springer, Thomas G. Ksiazek *et al.* (2003). "Viral discovery and sequence recovery using DNA microarrays". *PLoS Biology* **1**[2]: 257 - 260.

Warren, René L., Granger G. Sutton, Steven J.M. Jones i Robert A. Holt (2007). "Assembling millions of short DNA sequences using `SSAKE`". *Bioinformatics* **23**[4]: 500 - 501.

Welling, Luke i Laura Thomson (2005). *`PHP and MySQL Web Development`*. 3^{rd} ed. SAMS Publishing.

Whon, Tae Woong, Min-Soo Kim, Seong Woon Roh, Na-Ri Shin, Hae-Won Lee *et al.* (2012). "Metagenomic Characterization of Airborne Viral DNA Diversity in the Near-Surface Atmosphere". *Journal of Virology* **86**[15]: 8221 - 8231.

Winn, Mary E., Marian Shaw, Craig April, Brandy Klotzle, Jian-Bing Fan *et al.* (2011). "Gene expression profiling of human whole blood samples with the Illumina WG-DASL assay". *BMC Genomics* **12**, 412.1 - 8.

Woods, Jacquelina W., Kevin R. Calci, Joey G. Marchant-Tambone i William Burkhardt (2016). "Detection and molecular characterization of norovirus from oysters implicated in outbreaks in the US". *Food Microbiology* **59**, 76 - 84.

World Health Organization [WHO] (2016). "Estimación de la carga mundial de las enfermedades de transmisión alimentaria". *Centro De Prensa WHO*. pp.1–7; `http://www.who.int/mediacentre/news/releases/2015/foodborne-disease-estimates/es/`.

Wyn-Jones, A. Peter, Annalaura Carducci, Nigel Cook, Martin D'Agostino, Maurizio Divizia *et al.* (2011). "Surveillance of adenoviruses and noroviruses in European recreational waters". *Water Research* **45**[3]: 1025 - 1038.

Y

Ye, Chengxi, Zhanshan Sam Ma, Charles H. Cannon, Mihai Pop i Douglas W. Yu (2012). "Exploiting sparseness in *de novo* genome assembly". *BMC Bioinformatics* **13**[Suppl.6]: S1.1 - 8.

Yoshida, Mitsuhiro, Yoshihiro Takaki, Masamitsu Eitoku, Takuro Nunoura i Ken Takai (2013). "Metagenomic Analysis of Viral Communities in (Hado)Pelagic Sediments". *PLoS ONE* **8**[2]: e57271.1 - 14.

Z

Zerbino, Daniel R. i Ewan Birney (2008). "`Velvet`: algorithms for *de novo* short read assembly using *de Bruijn* graphs". *Genome Research* **18**[5]: 821 - 829.

Zhao, Yue, Hong Cao, Yindi Song, Yue Feng, Xiaoxue Ding *et al.* (2016). "Identification of novel mutations including a double mutation in patients with inherited cardiomyopathy by a targeted sequencing approach using the Ion Torrent PGM system". *International Journal of Molecular Medicine* **37**[6]: 1511 - 1520.

Ziv, Jacob i Abraham Lempel (1978). "Compression of individual sequences via variable-rate coding". *IEEE Transactions on Information Theory* **24**[5]: 530 - 536.

Annexos

Article 1

A metagenomic assessment of viral contamination on fresh parsley plants irrigated with fecally tainted river water

X.Fernandez-Cassi, **N. Timoneda**, E. Gonzales-Gustavson, J.F. Abril, S. Bofill-Mas, R. Girones.

Manuscrit en procés de segona revisió a International Journal of Food Microbiology.

International Journal of Food Microbiology xxx (2017) xxx-xxx

Contents lists available at ScienceDirect

International Journal of Food Microbiology

journal homepage: www.elsevier.com

A metagenomic assessment of viral contamination on fresh parsley plants irrigated with fecally tainted river water

X. Fernandez-Cassi[a, *, 1], N. Timoneda[a, b, c, 1], E. Gonzales-Gustavson[a], J.F. Abril[b, c], S. Bofill-Mas[a], R. Girones[a]

[a] Laboratory of Virus Contaminants of Water and Food, Department of Genetics, Microbiology and Statistics, University of Barcelona, Barcelona, Catalonia, Spain
[b] Computational Genomics Lab, Department of Genetics, Microbiology and Statistics, University of Barcelona, Barcelona, Catalonia, Spain
[c] Institute of Biomedicine (IBUB), University of Barcelona, Barcelona, Catalonia, Spain

ARTICLE INFO

Keywords:
Food metagenomics
Viral pathogens
Food safety
Caliciviridae
Picornaviridae
Next generation sequencing

ABSTRACT

Microbial food-borne diseases are still frequently reported despite the implementation of microbial quality legislation to improve food safety. Among all the microbial agents, viruses are the most important causative agents of food-borne outbreaks. The development and application of a new generation of sequencing techniques to test for viral contaminants in fresh produce is an unexplored field that allows for the study of the viral populations that might be transmitted by the fecal-oral route through the consumption of contaminated food. To advance this promising field, parsley was planted and grown under controlled conditions and irrigated using contaminated river water. Viruses polluting the irrigation water and the parsley leaves were studied by using metagenomics. To address possible contamination due to sample manipulation, library preparation, and other sources, parsley plants irrigated with nutritive solution were used as a negative control. In parallel, viruses present in the river water used for plant irrigation were analyzed using the same methodology. It was possible to assign viral taxons from 2.4 to 74.88% of the total reads sequenced depending on the sample. Most of the viral reads detected in the river water were related to the plant viral families *Tymoviridae* (66.13%) and *Virgaviridae* (14.45%) and the phage viral families *Myoviridae* (5.70%), *Siphoviridae* (5.06%), and *Microviridae* (2.89%). Less than 1% of the viral reads were related to viral families that infect humans, including members of the *Adenoviridae*, *Reoviridae*, *Picornaviridae* and *Astroviridae* families. On the surface of the parsley plants, most of the viral reads that were detected were assigned to the *Dicistroviridae* family (41.52%). Sequences related to important viral pathogens, such as the hepatitis E virus, several picornaviruses from species A and B as well as human sapoviruses and GIV noroviruses were detected. The high diversity of viral sequences found in the parsley plants suggests that irrigation on fecally-tainted food may have a role in the transmission of a wide diversity of viral families. This finding reinforces the idea that the best way to avoid food-borne viral diseases is to introduce good field irrigation and production practices. New strains have been identified that are related to the *Picornaviridae* and distantly related to the *Hepeviridae* family. However, the detection of a viral genome alone does not necessarily indicate there is a risk of infection or disease development. Thus, further investigation is crucial for correlating the detection of viral metagenomes in samples with the risk of infection. There is also an urgent need to develop new methods to improve the sensitivity of current Next Generation Sequencing (NGS) techniques in the food safety area.

1. Introduction

Food-borne diseases remain a significant cause of illness worldwide, and consumers are exposed to microbiological and chemical contaminants. From a microbiological point of view, food can be a vehicle for protozoan, bacterial, viral, and prion infections. Although most fecally excreted microorganisms cause gastroenteritis or acute hepatitis, other pathologies such as meningitis, myocarditis, and neurological disorders are also possible.

Food contamination can occur at several stages of food chain production, from the irrigation and collection stages on farms to contamination during food processing in industrial settings, food preparation at a restaurant, or at home. In high income countries, measures have been

* Corresponding author.
Email address: x.fernandez-cassi@ub.edu (X. Fernandez-Cassi)
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.ijfoodmicro.2017.06.001
Received 24 December 2016; Received in revised form 30 May 2017; Accepted 3 June 2017
Available online xxx
0168-1605/

implemented to reduce the risk of fecal contamination of water and food such as proper sewer pipeline systems and hygienic measures during food handling, manufacturing, and preparation. Countries use legislation to measure the microbiological quality of water and food, yet food-borne outbreaks are still reported (Bernard et al., 2014; Ethelberg et al., 2010).

Among all the food-borne etiological agents, viruses are the most important causative agents of food-borne outbreaks with noroviruses accounting for 125 million cases a year (Kirk et al., 2015; Painter et al., 2013). The increase in fresh food consumption, probably as a way for consumers to develop a healthier diet, has been linked to an increase in viral food-borne outbreaks (Callejón et al., 2015; Kozak et al., 2013). Coleman et al. (2013) noted that of the 127 outbreaks associated with leafy greens in the United States from 2004 to 2008, 64% of cases were attributed to a viral infectious agent such as noroviruses, sapoviruses, and hepatitis A virus. As mentioned, viruses can be accidentally introduced at different steps in food chain production, and crop irrigation with fecally contaminated water is one of the most critical points. Recently, Maunula et al. (2013) reported that 9.5% of irrigation water samples used to water berries were positive for human adenoviruses (HAdV), underlining the presence of fecal contamination. Although there are laws to control the microbiological water quality, none of them currently include specific viral parameters, and, therefore, water safety monitoring relies only on the use of fecal indicator bacteria (FIB). The usefulness of these laws for minimizing the viral presence in food matrices is unclear because the FIB do not always correlate with the presence of viral pathogens, and viruses are more resistant to water treatments than bacteria (Gerba et al., 1979; Pusch et al., 2005; Savichtcheva and Okabe, 2006).

While bacterial contamination in food has been widely reported, notifications of viral outbreaks have been hampered in many cases by a lack of specific, sensible, and standardized concentration/detection methods. These important factors are mandatory for the inclusion of viral parameters in food legislation. Recently, a standard ISO 15216-1:2017 was published for the concentration and quantification of the two food-borne viruses hepatitis A virus (HAV) and human noroviruses (HNoV) (https://www.iso.org/standard/65681.html), and qualitative detection is under revision (ISO/TS 15216-2:2013). The concentration methods are limited by their low recoveries while detection and quantification methods, which are usually based on RT-PCR or RT-qPCR, are restricted to specific targeted viruses. Many different pathogens may contaminate water and food simultaneously, especially if fecally contaminated irrigation water is used on fresh vegetables. The introduction of next generation sequencing (NGS) techniques in the food safety field allows for the simultaneous analysis of diverse viral pathogens in a single assay. To date, NGS has been applied to study the viral species that are present in all types of environmental and clinical samples as follows: oceans (Hurwitz et al., 2013), lakes (Djikeng et al., 2009), raw sewage (Cantalupo et al., 2011), reclaimed water (Rosario et al., 2009), and infectious clinical samples with unknown etiological agents (Greninger et al., 2015). However, the use of NGS in the food safety field has not been exhaustively explored (Aw et al., 2016; Park et al., 2011; Zhang et al., 2014).

The focus of the present study is to provide more information on the applicability of NGS techniques in the food safety field by studying viral contamination in fresh vegetables, with parsley plants (*Petroselinum crispum* L.) that were irrigated with river water containing fecal contamination as a model.

2. Materials and methods

2.1. River water samples

The Besòs River (Sant Adrià de Besòs, Barcelona, Catalonia, Spain) is a 17.7 km long, and it displays irregular discharge due to the Mediterranean climate. Along its path, the river collects the secondary effluent of 27 wastewater treatment plants (WWTP) and ends in the Mediterranean Sea next to Barcelona. Since the mid-1990, it has been subjected to a recovery process to improve its water quality. As a consequence of these efforts, the Besòs River is being used to irrigate crops by some local farmers. Twenty liter river water samples from the Besòs River were collected in May 2014. Ten liters of each river sample was used to irrigate the parsley plants, and the remaining 10 L was used to concentrate the viral particles to characterize the viruses that were present through qPCR and NGS. After two weeks, the procedure was repeated. The river water samples collected for parsley irrigation were kept in the dark at 4 °C for 15 days and used to irrigate the plants.

2.2. River water viral particle concentration

The viruses present in the 10 L river water samples were concentrated using skimmed milk organic flocculation. This method has a recovery efficiency of 50% (20–95%) and it was applied as previously described by Calgua et al. (2013). Viral concentrates were suspended in 8 mL of phosphate buffer (pH 7.5) and stored at − 80 °C until further use. The two river water viral concentrates were labeled as the Besòs River water (BRW1 and BRW2) samples.

2.3. Parsley plant growth and irrigation

Parsley (*Petroselinum crispum* L.) seeds were planted and cultivated in a climate room at 22 °C, 60% relative humidity, and light conditions equivalent to 110 μmol of photosynthetically active radiation (PAR) at the "Serveis de Camps Experimentals" at the University of Barcelona. The seeds were irrigated using a nutritive solution containing iron, nitrogen and phosphorus (Hoagland solution 50%). After 6 weeks, a total of 12 different parsley pots were moved to a greenhouse from the same facility.

All 12 parsley pots were irrigated twice a week during the whole growth process (from May to June) by using 4 L of the same nutritive Hoagland solution. This irrigation procedure was performed by inundating the tray containing the parsley pots, and thus the parsley leaves were not washed. Half of the 12 pots were irrigated daily, in addition to receiving nutritive solution, by spraying the leaves with 15 mL of Besòs River water. These samples were labeled as Besòs River Parsley (BRP).

In the same way, the remaining pots were used as a Negative Control Parsley group (NCP), and they were irrigated daily by spraying 15 mL of the same nutritive solution that was used for irrigation, but by inundation, into both the control and test pots. The negative control was used as a blank sample to identify viral sequences that could be naturally present in the parsley plants or due to other external factors (greenhouse, irrigation nutritive solution, facility users, manipulation, reagent contaminants or equipment).

Two weeks later, a fresh BRW sample was collected to continue the irrigation process for another 15 days. At the end of the study, both plant groups were irrigated with 450 mL of river water or control nutritive solution water. After one month of daily irrigation, 25 g of parsley leaves from NCP and BRP were hand-cut by investigators who were wearing sterile gloves, and the leaves were placed in sterile bags (Bag-

Page® filter-bag, Interscience, France). The samples were kept at 4 °C for < 48 h until the concentration method was applied.

2.4. Viral concentration from plants

Twenty-five grams of NCP and BRP were washed in a sterile filter-bag with 50 mL of glycine buffer (pH of 9.5, 0.25 N) for 40 min using a stomacher. Afterwards, the sample pH was adjusted to 7.0 (± 0.2) by using HCl 0.1 N. To remove bacteria and other suspended organic material, the samples were centrifuged at 8000 × *g* for 10 min at 4 °C. The supernatant was carefully collected without disturbing the pellet and ultracentrifuged at 90,000 × *g* for 1 h. The pellet containing the viruses was suspended in a final volume of 500 μL of phosphate buffer (pH 7.5) and stored at − 80 °C until further analysis.

2.5. Nucleic acid extraction, library preparation, and sequencing

Volumes of NCP, BRP, BRW1 and BRW2 viral concentrates were treated with 156 units of Turbo DNAse (cat no AM1907, Ambion, Lithuania) for 1 h at 37 °C to remove free DNA prior to nucleic acid extraction, and 280 μL of the treated DNAse viral concentrate was extracted using the QIAamp®Viral RNA Mini Kit from QIAGEN (Qiagen, Valencia, CA, USA). The nucleic acids (NA) were eluted in 60 μL and stored at − 80 °C for further analysis. To enable the detection of both DNA and RNA viruses, the total NAs were reverse-transcribed as previously described in Wang et al. (2002, 2003). In short, SuperScript II (Life Technologies, California, USA) was used to retrotranscribe RNA to cDNA with primerA (5′-GTTTCCCAGTCACGATCNNNNNNNNN-3′). Second strand cDNA and DNA were constructed with the primer sequences using Sequenase 2.0 (USB/Affymetrix, Cleveland, OH, USA). A PCR amplification with AmpliTaqGold (Life Technologies, Austin, Texas, USA) was performed using primerB (5′-GTTTCCCAGTCACGATC-3′) with 20–30 cycles; this step was run in duplicate. The PCR products were purified and eluted in 15 μL using a Zymo DNA clean and concentrator (cat no D4013, Zymo Research, USA) to yield enough DNA for the library preparation. Amplified DNA samples were quantified by Qubit 2.0 (Life Technologies, Oregon, USA) and libraries were constructed using a Nextera XT DNA sample preparation kit (Illumina Inc.). The samples were sequenced on an Illumina MiSeq 2 × 300 in base-pair paired end format. Raw read sequences for the four analyzed samples were made publicly available at NCBI Sequence Read Archive database (NCBI-SRA), under the PRJNA381682 bioproject.

2.6. Determination of the level of human fecal contamination in river water by HAdV qPCR

To evaluate the level of human fecal contamination in the Besòs River water, specific real-time qPCR assays for human adenovirus were performed with TaqMan® Environmental Master Mix 2.0 (Life Technologies, Foster City, CA, USA). Real-time primers and probes for HAdV were previously described in (Bofill-Mas et al., 2006).

2.7. Bioinformatic analyses

The quality of the raw and clean read sequences was assessed using FASTX-Toolkit software, version 0.0.14 (Hannon Lab, http://www.hannonlab.org). The read sequences were cleaned by Trimmomatic version 0.32 (Bolger et al., 2014), while attending to the sequencing adaptors and linker contamination. Low quality ends were trimmed by using an average threshold Phred score above Q15 over a running-window of 4 nucleotides. Low complexity sequences, which were mostly biased to repetitive sequences that would affect the performance of downstream procedures in the computational protocol, were then discarded after estimating a linear model based on Trifonov's linguistic complexity (Sarma et al., 1990). Finally, duplicated reads were removed in a subsequent step to speed up the downstream assembly.

Assemblies were obtained using Velvet (Zerbino and Birney, 2008) and Meta-Velvet tools (Namiki et al., 2012) versions 1.2.10 and 1.2.02, respectively. Afterwards, the contigs and singletons longer than 100 bp were queried for sequence similarity using NCBI-BLASTN and NCBI-BLASTX (Altschul et al., 1997, 1990) against the NCBI viral complete genome database (Brister et al., 2015), the viral division from the GenBank nucleotide database (Benson et al., 2015), and viral proteins from UniProt (UniProt Consortium 2015, ftp://ftp.uniprot.org/pub/databases/uniprot/current_release). The species nomenclature and classification were assigned according to the NCBI Taxonomy database standards and the basic Baltimore classification. The HSPs considered for taxonomic assessment must have an E-value that is lower than 10^{-5} and a minimum length of 100 bp. On the basis of the best BLAST results and a 90% coverage cutoff, the sequences were classified into their likely taxonomic groups of origin. Contigs and singletons with viral taxonomic assignments were scaffolded by using the Geneious software assembler (Geneious 9, Kearse et al., 2012). The scaffolded sequences were subsequently mapped using the Geneious mapper tool. Phylogenetic trees were constructed, also using Geneious software and a neighbor-joining method with 1000 bootstrap replicates was chosen.

2.8. Richness

Tables summarizing the number of sequences from the assembly matching of each taxonomic unit were built. From those tables, the richness ratios were calculated with CatchAll software, version 4.0 (Allen et al., 2013). Among all the models included in the package, the Chao1 non-parametric model was chosen.

3. Results

3.1. Mi-Seq sequence output

Parsley that was irrigated with river water was assayed using 20 (BRP20) and 30 amplification cycles (BRP30). A summary of the bioinformatics filtering and assembly of the reads can be found on Table 1 (data for BRP30 not shown, as explained on next section). Total number of sequences having at least an homology hit to a viral sequence from the distinct databases after BLAST searches, as well as the total viral species and families from which each sample richness score was estimated are also listed on that table. Only a small amount of sequences were taxonomically assigned to known viruses, except for the NCP sample where most of the hits accounted for viruses infecting aphids parasiting the leaves. In this case, despite having much more reads from the sequencing experiment, the taxonomic complexity was lower anyway.

3.2. Estimating the viral diversity by using CatchAll

The estimated total richness ratios for the BRP20 and BRP30 samples were 255.6 (± 34.7 s.d.) and 171.8 (± 40.9 s.d.), respectively. The higher estimated viral richness obtained in BRP20 (20 cycle amplification) over BRP30 (30 cycles amplification) is consistent with the observed higher diversity of the detected viral species. For this reason, all further analyses were conducted only taking into account BRP20. The resulting richness ratios for NCP, BRW1 and BRW2 were 130.3 (± 34.7 s.d.) and 632.4 (± 44.4 s.d.) and 923.5 (± 70.9 s.d.), respectively. The richness value obtained from the river water is about 3 to 4 times greater than that obtained for the other two samples (Table 1). This finding reflects a higher diversity in the viral population that was present in a more complex ecosystem, such as river water, which is

Table 1
Metagenomics sequencing summary statistics. Sequences and nucleotide counts are total values; number of pairs is half of the shown values. Percent of reads sequence refer to the total number of raw reads, while the percent of sequences having, or not, a BLAST hit corresponds to the total number of assembled sequences (contigs plus singletons). Difference between the sequences assigned to known viruses and the sequences without a BLAST hit relates to those sequences having a BLAST hit not passing all the filtering criteria for a valid species assignment.

Samples	BRW1			BRW2			BRP			NCP		
Description	Besòs River water			Besòs River water			Parsley irrigated with river water (20C)			Parsley irrigated with control water		
	Sequences		Nucleotides	Sequences		Nucleotides	Sequences		Nucleotides	Sequences		Nucleotides
Raw reads (MiSeq)	5,426,788		1,450,732,550	2,476,054		577,865,642	1,192,716		327,348,654	8,067,728		2,199,101,314
Clean reads												
Pair-ends	3,955,116		930,629,132	2,445,144		508,229,216	329,520		73,820,015	5,301,560		1,264,595,255
Single-ends	663		123,178	158		26,175	1808		148,588	866		243,907
Total	3,955,779	72.89%		2,445,302	98.76%		331,328	27.78%		5,302,426	65.72%	
Assembly (MetaVelvet)												
Contigs	243,043		58,899,523	192,193		44,300,670	20,961		5,871,015	10,296		3,160,250
Singletons	2,556,829		599,432,209	1,664,874		339,082,359	191,276		43,401,095	5,169,800		1,233,416,680
$N50_{contigs\ +\ singletons}$	1,149,201		270	715,216		232	83,416		324	2,125,927		372
Homology (BLAST)												
Putative viral seqs	77,626	2.77%		44,607	2.40%		59,380	27.98%		3,878,957	74.88%	
Seqs without BLAST hit	2,702,102	96.51%		1,772,659	95.45%		135,759	63.97%		914,807	17.66%	
# distinct viral families	22			26			25			16		
# distinct viral species	361			464			139			60		
	Estimated value		SE	Estimated value		SE	Estimated value		SE	Estimated value		SE
Richness	632.4		44.4	923.5		70.9	255.6		34.7	130.3		34.7

heavily impacted by human activity. As expected, the NCP richness was lower than that of BRP, which was irrigated using BRW.

3.3. Identification of human and animal viruses in river water samples used for irrigation (BRW)

River water samples from the Besòs River contained human viral fecal contamination, as shown by the finding that both samples were positive by qPCR for Human Adenovirus (HAdV) with low concentration $1.2 \times 10^{+3}$ GC/L and $5.62 \times 10^{+2}$ GC/L, respectively. Parsley irrigated with Bèsos River water was qPCR-positive for HAdV, also with low concentration $7.2 \times 10^{+1}$ CG/25 g. Interestingly, only the BRW2 presented reads assigned to *Adenoviridae* family, specifically to HAdV41 (HAdV-F). Plant-infecting ssRNA + viruses from the *Tymoviridae*, *Virgaviridae*, and *Alphaflexiviridae* families were found. These plant viral families accounted for a 66.13%, 14.45%, and 1.11% of the viral reads, respectively. The primary phage sequences found in the river water were annotated as *Myoviridae*, *Siphoviridae*, *Podoviridae*, and dsDNA *Microviridae* families, accounting for 5.70%, 5.06%, 3.69% and 2.88% of the viral reads, respectively. The human and potentially zoonotic viruses found here are summarized in Table 2. By applying the described metagenomics approach, up to 26 different viral families were identified in river water used for irrigation. Among the pathogenic and potentially pathogenic viruses, members of the *Astroviridae*, *Adenoviridae*, *Reoviridae*, *Picobirnaviridae*, *Picornaviridae* and *Parvoviridae* families were detected. Although important pathogenic viruses were detected using the metagenomics procedure, the sum of all the reads from all the putative pathogenic families did not reach the 1% of viral reads of the sample.

3.3.1. ssRNA + viruses detected from Picornaviridae, Astroviridae and Hepeviridae families in BRW

Human pathogenic viruses within the *Picornaviridae* family, such as Aichi virus species, were found to have high identities (94.4%). Compared to reference strains several sequences detected in BRW1 presented low identity scores and high coverage (> 99%); being distantly related to the recently described *Ampivirus* genus (NC_027214.1). The BLASTx sequences matched up to 80% of the identity of putative RHV-like sequences in the ampivirus species of the *Picornaviridae* family, suggesting that these detected viral sequences could represent a new viral species within the same genus. Furthermore, sequences of 159 and 206 base pairs resembling rodent hepatovirus and bat picornavirus were detected, revealing shared identities of 77 and 79%, respectively.

Several sequences that were distantly related to the *Hepeviridae* family were detected in Besòs River water. A total of 4 sequences with lengths ranging from 242 to 636 nucleotides showed a distant relation to a sequence from a newly suggested virus named tentatively hepelivirus (accession AFR11847) according to a BLASTx against the entire GenBank database. All the translated reads were distantly aligned to the RdRp region of this new proposed virus with identities ranging

Table 2
BLASTN statistics for human/animal viruses found using metagenomics in Besòs River water samples.

Sample	Viral family	Viral specie	# related sequences	Maximum contig length	Blast output statistics		
					Larger contig nucleotide identity (%)	Average query coverage (%)	Match GenBank accession number
River water samples used for irrigation (BRW)	*Adenoviridae*	HAdV-41	2	220	100%	100%	KY316164
	Astroviridae	MAstV-1	5	374	94%	99%	KF039911
	Caliciviridae	Goose Calicivirus	2	221	74%	93%	KJ473715
	Picobirnaviridae	Human picobirnaviridae	2	304	98%	100%	KJ663813
		Porcine picobirnavirus	2	360	83%	92%	HM070240
	Reoviridae	Human Rotavirus A	6	296	99%	100%	KU048625
	Picornaviridae	Aichi virus	4	385	95%	100%	GQ927712
		Ampivirus	7	716	74%	99%	KP770140
		Bat picornavirus	1	206	77%	87%	HQ595341
		Rat hunnivirus	1	394	91%	100%	KT944214
		Rodent hepatovirus	2	159	80%	99%	KT452641
	Parvoviridae	Kilham rat virus	1	204	79%	80%	AF321230
		Avian adeno-associated virus	5	258	99%	100%	NC_006263
		Caprine adeno-associated virus	1	376	99%	98%	DQ335246
		Simian adeno-associated virus	1	202	97%	99%	EU285562
		Porcine bocavirus	3	344	87%	100%	KJ622366
		Mouse parvovirus	1	448	84%	100%	KY489986
		Bovine adeno-associated virus	4	274	79%	95%	AY388617
		Rat bocavirus	5	438	94%	100%	KT454517
		Human adeno associated	12	441	98%	100%	AY530578
	Circoviridae	Gull circovirus	3	174	76%	85%	NC_026625
		Avon-Heathcote Estuary associated circular virus 3	2	369	76%	94%	KT454927
		Cyclovirus	13	487	74%	99%	GQ404854
		Human feces pecovirus	2	232	98%	100%	KT600066
		Porcine circovirus	10	148	76%	82%	NC_027796

from 39.3 to 42.2% and query coverage from 54.43 to 100%. A translated sequence from the read with query coverage of 100% was used to construct a phylogenetic tree, together with reference to *Hepeviridae* GenBank entries covering the RdRp region (Fig. 1). Within the *Hepeviridae* family, the sequence detected in river water shares its closest common ancestor with Hepelivirus, which was recently described in a sewage sample from Nepal. Both sequences seem to have diverged from species within the *Orthohepevirus* genus and are closer to the recently characterized *Piscihepevirus*. Sequences belonging to the *Astroviridae* species Mamastrovirus 1 were detected in river water that was used for irrigation, with nucleotide identities ranging from 92.6% to 97.4%. Most of the resulting sequence reads had a hit with the serine

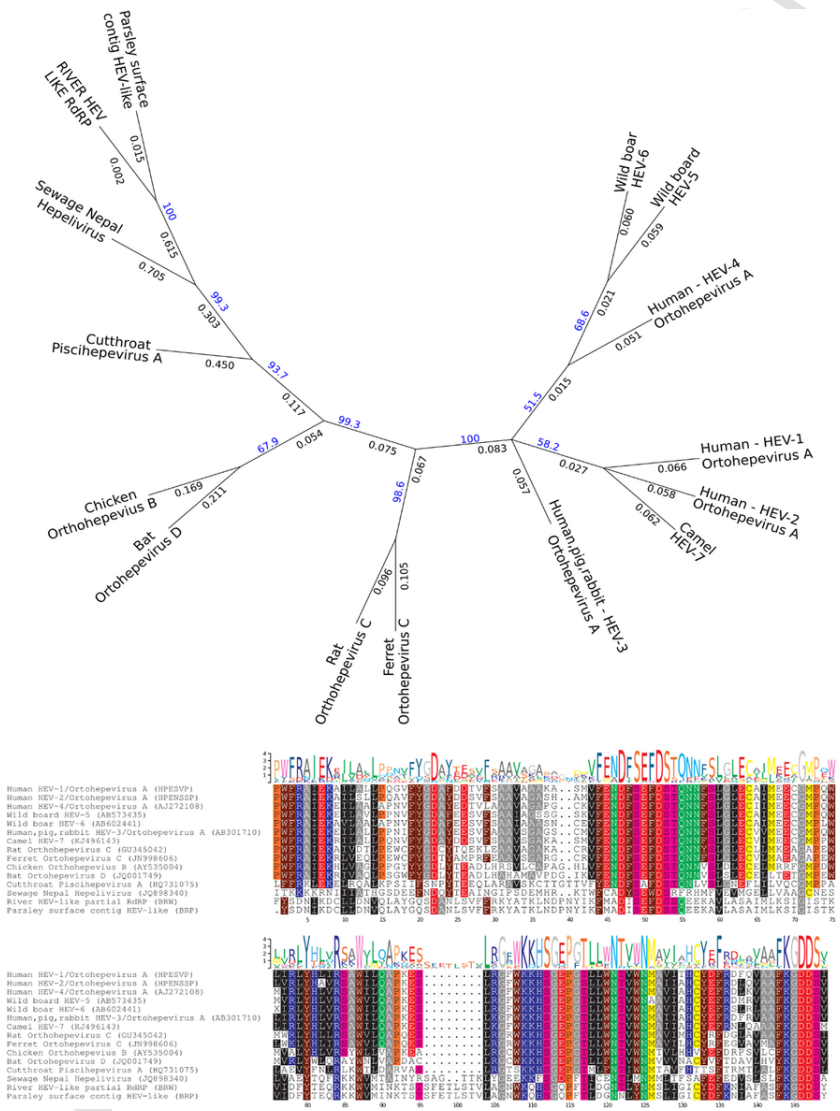


Fig. 1. Phylogenetic analysis of the partial RdRp region for known Hepelivirus and *Hepeviridae* families, including the River and Parsley surface sequences. A) It is evident that both novel sequence candidates cluster together, and closer to Hepelivirus rather that to *Hepeviridae* family. Phylogenetic tree was built with Geneious software using Jukes-Cantor model, and clustering method was Neighbor-joining with 1000 bootstrap replicates. Two different scores were included over the branches: bootstrap scores in blue, phylogenetic distances in black. B) The phylogenetic tree was generated from the conserved positions of the RdRp region (204aa), which are shown on this alignment summary produced by Geneious software with default parameters. Sequence accession numbers from GenBank were also annotated with the labels used in the above tree figure. Colors in the alignment correspond to distinct amino acids. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

protease encoded in this viral family, showing high identities at the amino acid level (≥ 90%).

3.3.2. ssDNA viruses detected in from families Parvoviridae and Circoviridae in BRW

The ssDNA viral reads assigned to the *Parvoviridae* family represented 0.067% of the total viral sequences. Most of the reads associated with this family share a high identity with several densoviruses. Those species infect a wide range of hosts including cockroaches, mosquitoes, and ants, among other. However, some of the sequences presented a heterogeneous identity with viruses that infect mammals (swine, rats, and mice) and with several adeno-associated parvoviruses that infect humans (84–100%). ssDNA viral reads matching the *Circoviridae* family made up 0.07% of the viral sequences presenting a high identity with several circoviral sequences that were previously described in, birds, insects, porcine stool and sewage.

3.3.3. dsRNA viruses of Reoviridae, Picobirnaviridae families detected in BRW

Human Rotaviruses from species A were detected only in BRW2 sharing higher identities (99–100%) whereas BRW1 had no reads taxonomically assigned to the family *Reoviridae*.

Sequences remotely homologous to the double-stranded RNA human picobirnaviruses were detected presenting similarity, at a protein level, to the RNA-dependent RNA polymerase (RdRp) region of the human picobirnaviruses, with identities ranging from 60 to 90%.

3.4. Identification of human and animal viruses in parsley plants irrigated with control water (NCP)

NCP was used as a negative control to discard the sequences already present in the plants or introduced due to external factors. Despite the higher number of viral sequences obtained (7.50%) in comparison with the sequences from BRW (2.41%), none of them matched the pathogenic enteric viruses. Most of the taxonomically assigned sequences from this sample were classified as the dsDNA phage families *Podoviridae* (32.09%) and *Siphoviridae* (0.09%), whereas the majority of ssRNA + viral sequences (57%) were related to the insect virus *Dicistroviridae* family.

3.5. Identification of human and animal viruses in parsley plants irrigated with river water samples (BRP)

Viral concentrates from the BRP contained viruses from 18 different families. Despite the low MiSeq output (596,358 raw reads), members of the families *Astroviridae*, *Caliciviridae*, *Flaviviridae*, *Hepeviridae*, *Parvoviridae*, and *Picornaviridae*, which included several viral pathogenic species, were detected. The *Dicistroviridae* family of insect viruses accounts for the majority of reads (41.52%). Among bacteriophages, the *Podoviridae* family are the most abundant, accounting for 14.06% of the total viral reads. A more detailed list of human viral families is summarized in Table 3.

3.5.1. ssRNA + viruses from the Picornaviridae, Caliciviridae, Hepeviridae and Flaviviridae families detected in BRP

The majority of human pathogenic viral reads from BRP were assigned to the *Picornaviridae* family (0.035% total reads from sample) and several enteroviruses from species A and B. For enterovirus species A, most of the reads were similar to the reads for coxsackievirus A6 (CV-A6), coxsackievirus A10 (CV-A10), and coxsackievirus A16 (CV-A16). After scaffolding almost the entire genome, CV-A16 was recovered. Enterovirus species B reads displayed matches up with echovirus E6 and enteroviruses B4 and B5.

Viral sequences related to the ssRNA + *Caliciviridae* family were identified and taxonomically assigned to human sapovirus GI and norovirus genogroup IV. The sapovirus reads were scaffolded with Geneious software. A contig of 1592 base pairs covering the RdRp/VP1 junction was found to be closely related to GI.2 sapoviruses, sharing 96% of the identity of GenBank accession AB614356.

Four sequences presented identities above 98% with norovirus GIV. One out of 4 NoVGIV sequences recovered here shared 100% of its nucleotide identity with the VP1 region (GenBank accession LC150859).

Within the *Hepeviridae* family, several sequences related to hepatitis E genotype 3 ranging from 154 to 548 base pairs long were detected in the parsley leaves. Those sequences matched up with variable identities at nucleotide levels ranging from 86% to 98% and clustered with HEV genotype III sub genotype f (data not shown). A sequence of 448 nucleotides taxonomically related to this family was detected by BLASTx presenting a low amino acid identity with the RdRP domain from the Hepatitis E virus (34% amino acid identity, 80% coverage, Accession ANJ02846). The retrieved sequence contains an FKGDDS domain as well, which is commonly present in *Hepeviridae* family members. The phylogenetic tree suggests that the RdRp region from this sequence detected on the parsley surface is distantly related to the Hepelivirus sequence detected in sewage from Nepal (Fig. 1).

The ssRNA + viral species GBV-C virus from the *Flaviviridae* family were also reported. These detected sequences had an identity between 84.7 and 94% at the nucleotide level, and they covered several regions of the GBV-C virus including the E1 and E2 proteins, helicase, NS3, and NS5A domains.

3.5.2. Members of ssDNA viral families Parvoviridae, Anelloviridae and Circoviridae detected in BRP

The ssDNA viral reads belonging to the *Parvoviridae* family represented 0.016% of the total viral sequences. Most of the reads were associated with human bocavirus.

Sequences related to the ssDNA viral *Anelloviridae* family are highly prevalent, accounting for 35.10% of the total reads from the samples. A variety of sequences related to Torque teno mini virus (from TTMV 1 to 9), Torque teno midi virus (from TTMDV 1 and 2) and Torque teno virus (TTV2, 3, 18 and 19) were found. The most prevalent species were TTMVs 4, 5 and 8.

4. Discussion

In this manuscript, NGS techniques were applied within the food safety field. River water contaminated with human feces was used to irrigate parsley as a plant model. To our knowledge, only two studies using NGS techniques in the field of food safety have been published, with one on meat and another one on lettuce (Aw et al., 2016; Zhang et al., 2014), and many questions and technical improvement are still pending. In this study, a high diversity of viruses including pathogens have been identified in water and more importantly on the surface of vegetables. Contaminated parsley could present a risk for the population, and it could represent a source of different viral diseases such as hepatitis and gastroenteritis due to the presence of viral that are transmissible by the fecal-oral route.

BRW contains human viral fecal contamination as shown by the positive HAdV qPCR results and the results obtained by NGS. HAdV has been widely used as a fecal marker of human viruses as reviewed in Bofill-Mas et al. (2013). Therefore, the Besòs River water was appropriate for this simulation assay as a representative irrigation water with human viral fecal contamination used for crop irrigation. Despite the low concentration of HAdV for the two river water types collected ($1.2 \times 10^{+3}$ GC/L and $5.62 \times 10^{+2}$ GC/L) and at the parsley surface (7.2×10^{1} GC/25 g), only HAdV41 sequences were detected in BRW2 by using untargeted metagenomics. Viral metagenomics is usually lim-

Table 3
BLASTN statistics for human/vertebrates viruses found using metagenomics in Parsley Plants irrigated with Besòs River water.

Sample	Viral family	Viral specie	Number of related sequences	Maximum contig length	Blast output statistics		
					Larger contig nucleotide identity (%)	Average query coverage (%)	Match GenBank accession number
Parsley plants irrigated with river water samples (BRP)	*Caliciviridae*	Human norovirus GIV.1	4	360	98%	100%	JQ613567
		Human sapovirus GI.2	16	1780	95%	100%	AB614356
	Hepeviridae	Hepatitis E genotype 3	8	544	86%	99%	JQ953666
	Flaviviridae	Hepatitis G virus	13	1050	93%	100%	KU685422
	Parvoviridae	Human bocavirus 2a	10	1287	99%	100%	FJ170280
	Anelloviridae	Torque teno midi virus 1	263	306	95%	88%	AB290918
		Torque teno midi virus 2	388	269	77%	87%	AB290919
		Torque teno mini virus 1	44	237	85%	80%	AB026931
		Torque teno mini virus 2	30	241	80%	82%	AB038629
		Torque teno mini virus 3	68	303	88%	89%	AB038630
		Torque teno mini virus 4	2529	315	77%	82%	AB041963
		Torque teno mini virus 5	918	564	91%	83%	AB041962
		Torque teno mini virus 6	280	314	84%	81%	AB026929
		Torque teno mini virus 7	198	297	93%	88%	AB038627
		Torque teno mini virus 8	2503	305	78%	85%	AF291073
		Torque teno mini virus 9	5	189	94%	89%	AB038631
		Torque teno virus 18	1	150	75%	99%	AX025718
		Torque teno virus 2	1	225	75%	95%	AB049608
		Torque teno virus 29	1	386	87%	100%	AB038621
		Torque teno virus 3	2	191	79%	85%	AY666122
	Circoviridae	Avon-Heathcote Estuary associated circular virus 3	1	278	73%	91%	NC_026625
	Picornaviridae	Enterovirus species A	11	3706	97%	100%	LT617117
		Enterovirus species B	12	394	97%	99%	KU574626

ited by the quantities of DNA/RNA of viral origin that are present in the tested samples. Therefore, after the RT and sequenase reactions, a PCR-based random amplification method (SISPA) is used to obtain enough DNA for library preparation. This PCR step might introduce bias by amplifying the most abundant genomes and producing %GC bias (Duhaime et al., 2012). Less abundant genomes might not be sequenced or may be underrepresented (Karlsson et al., 2013). A balance between the number of cycle amplifications and DNA concentrations for library preparation must be achieved. The application of a higher number of PCR cycles produced higher DNA concentrations for library preparation but decreased the estimated viral richness of the river water-irrigated parsley, as in BRP30 (data not shown). For these reasons, metagenomes cannot be interpreted quantitatively but only as a relative abundance of species for comparison. The detection of low numbers of DNA viruses, such as Human Adenovirus, whose presence has been quantified in River water and Parsley leaves by qPCR and have only been detected by metagenomics in one river water sample, arises a sensitivity question regarding metagenomics applied to food safety. This point is especially important due to the fact that viruses are present in food and environmental samples in low concentration, frequently close to the LODs and LOQs of PCR and (RT) qPCR systems to detect them. To reduce this possible lack of sensitivity of viral metagenomics targeted/capture sequence NGS could be a more suitable approach to be applied as a tool to detect all diverse viral strains in one specific group of viruses in food and drinking water. As sequencing costs are dramatically diminishing with the novel high-throughput technologies, this kind of approaches will be affordable to run periodic controls and to check potential risks in real time in the near future.

Other factors that might affect the detected viral sequence diversity can be the different efficiencies of the given concentration methods and the representative sample volume tested for each matrix as well as the inactivation and degradation of viral genomes during crop production.

Despite this, the results presented in this work support the use of a viral Metagenomics approach for a general assessment of viral contaminants present in food matrices by using a single assay. Taking into account all the samples, viruses were assigned to > 34 different families. Important viral pathogens belonging to the families *Astroviridae*, *Adenoviridae*, *Reoviridae*, *Caliciviridae*, *Circoviridae*, *Hepeviridae*, *Picornaviridae*, and *Parvoviridae* have been detected.

The parsley that was irrigated with river water shows a very high diversity of viral strains, and more pathogenic viruses were detected in parsley than in river water. This finding could be related to a lower background of DNA/RNA input in the sample and potentially high absorption or stability capabilities for some viral species at the vegetable

surface. Two different concentration protocols have been used to concentrate viral particles due to the different nature of matrices tested. The viral concentration method used for parsley leaves is based on elution using glycine at basic pH and ultracentrifugation whereas the concentration from river samples was done by using SMF. Although the recovery efficiency of both methods have showed to be equivalent for viral indicators, the impact of different concentration and extraction methods might have played a role as proven to occur in sewage on the context of metagenomics sequencing (Hjelmsø et al., 2017). Parsley that was irrigated with a nutritive solution and used as a negative control showed an expected absence of viruses associated with fecal contamination.

The *Caliciviridae* viral family has been detected in the parsley plants. Noroviruses (NoV) are the leading cause of food-borne disease outbreaks worldwide (Koo et al., 2010) whereas human sapoviruses (HSaV) are increasingly recognized as a food-borne outbreak etiological agent (Kobayashi et al., 2012). Human gastroenteritis is primarily caused by the specific genotype NoV GII.4 (Verhoef et al., 2015). Noroviruses from GI and GII are abundantly found in fecally contaminated water but were neither detected in vegetal produce nor in the river water tested. The absence of these pathogens can be explained by their seasonality (Haramoto et al., 2006; Nordgren et al., 2009) and, probably, by the low titer of these viruses during the period of the year when the study was conducted (May). Other minority HNoV species, such as GIV NoV (which are not commonly included in environmental or food safety studies) were detected in the parsley leaves. NoV GIV food-borne outbreaks have rarely been reported, but the association of this virus with human pathology has been shown (Ao et al., 2014). The availability of environmental NoV GIV information is scarce; therefore, information about the seasonality of the virus is not clear. The presence of norovirus in fresh vegetables is a matter of concern with respect to the norovirus genogroup, and more data regarding NoV GIV seasonality and epidemiology is needed. Human sapoviruses (HSaV) belonging to GI, GII, and GV were detected. Most of the sequences were assigned to HSaV GI.2, which is considered a minor genogroup with very few outbreaks reported (Iwakiri et al., 2009). However, an increase in the number of outbreaks ligated to HSaV GI.2 has been observed in recent years (Lee et al., 2012; Svraka et al., 2010).

In the present study, all the HAstV-related sequences were only detected in BRW and were assigned to the genus *Mamastrovirus*, species MAstV-1. The involvement of HAstV in foodborne outbreaks has also been documented (Oishi et al., 1994), but its occurrence seems to be of lower importance compared to the number of outbreaks caused by NoV. The importance of HAstV might be underestimated due that in sporadic gastroenteritis cases the etiological agent is not investigated.

To our knowledge, HEV Food-borne outbreaks have been linked to the consumption of contaminated meat (Li et al., 2005; Yazaki et al., 2003) but no single outbreak has been linked to the consumption of fresh vegetables yet. However, Kokkinos et al. (2012), who studied the prevalence of HEV in leafy green irrigation water and in point-of-sale lettuce, have found the virus in 1 out of 20 (5,0%) and 4 out of 125 (3,2%) of the tested samples, respectively. HEV-3 has been detected in the parsley; this is a potentially zoonotic HEV strain, and its finding is consistent with the results of Kokkinos et al., which suggests that vegetables could represent a potential vehicle of transmission for HEV. In Spain, HEV-3 in the clade of subtype 3f strains have been detected in most autochthonous human cases (Fogeda et al., 2009) and in pigs, including HEV cases associated to pork meat consumption (Riveiro-Barciela et al., 2015). In this study, several sequences that are distantly related to the RdRp from genus *orthohepevirus* and closely related to hepeliviruses found in sewage from Nepal have been detected (Ng et al., 2012). These results show how little knowledge there is about the *Hepeviridae* family, but it will surely be expanded with the application of NGS techniques to different samples. However, the role and importance of these viral sequences as a food-borne agent have yet to be fully understood.

Different viruses in the *Picornaviridae* family have been detected in river water and parsley. *Picornaviridae* is a family grouping of > 30 different genera of ssRNA + viruses that infect vertebrates and includes historically important human pathogens, such as hepatitis A virus (HAV), enteroviruses (EV), and poliovirus. Aichi virus (AiV), which has been detected in river water, is a viral specie within genus *Kobuvirus* related to gastroenteritis (Yamashita et al., 1991). Recent studies have shown that AiV may co-infect with other enteric viruses (Räsänen et al., 2010).

Different *Enteroviruses* (EV) from species A and B were detected in parsley surface. An increase in enterovirus outbreaks has been reported recently by emerging recombinant EV strains (Holm-Hansen et al., 2016; Zhang et al., 2010). Nearly the entire genome of CV-A16, an emerging enterovirus genotype from species A which is related to hand, foot, and mouth disease (HFMD), was sequenced. Although EV can be transmitted person-to-person, the EV from these species have been linked to water-borne (Beller et al., 1997; Häfliger et al., 2000) and food-borne outbreaks (Le Guyader et al., 2008).

Closely related animal *Picornaviruses* sequences with low nucleotide identity to *Ampivirus* (Reuter et al., 2015), a new genus infecting amphibians, and with rat hepatovirus (Drexler et al., 2015), a proposed ancestral virus with a putative common origin with HAV, have been detected.

Currently, several new picornaviruses have been discovered with the introduction of NGS technologies to virology (Holtz et al., 2009).

The low shared identities for the sequences found in this study suggest that more members of the *Picornaviridae* family will be discovered in the years to come.

Rotaviruses were only detected in BRP2. Other dsRNA viruses from the family *Picobirnaviridae* were detected in tested samples. Therefore, according to our results, there is no data that clearly indicate a bias for rotaviruses or other viruses not detected, and the absence of these viruses might be related to a lower relative abundance compared to other viral species present.

Other viral sequences detected in the parsley or irrigation water from the families *Flaviviridae* and ssDNA viruses from the families *Parvoviridae*, *Circoviridae and Anelloviridae*, whose transmission route through fecally contaminated water or food is not fully understood and are known to infect human without causing any known disease were detected.

Several sequences that were taxonomically assigned to *Dicistroviridae* family aligned with the Aphid lethal paralysis virus (ALPV). The presence of greenflies in the greenhouse area could explain their high abundance. Plant viruses from *Virgaviridae* and *Tymoviridae* families as well as bacteriophages from the *Siphoviridae*, *Podoviridae*, *Microviridae* and *Inoviridae* families were detected. The quantities of plant and bacteriophage viruses matching our read sequences were expected because previous environmental studies already highlighted their high concentrations (Cantalupo et al., 2011; Ng et al., 2012).

Recently, some authors have described the ability of lettuce to internalize viral particles belonging to the *Caliciviridae* family (DiCaprio et al., 2015a, b; Esseili et al., 2012b), which suggests the possibility that infectious viruses may not only be found on food surfaces. Also the risk of infection trough the consumption of internalized NoV has been quantified using QMRA approaches (Sales-Ortells et al., 2015). Present study has only evaluated viruses attached at leaves surfaces. Future studies should gaze at this internalization scenario to evaluate viral risk of infection through internalized viruses.

Finally, it is remarkable that the Metagenomics approach facilitates the capture and assignment of a wide diversity of DNA/RNA viruses from a particular sample in a single assay. In the present work, several viral species, the transmission routes of which are not yet fully under-

stood, have been detected in food or in river water used for irrigation. Viral persistence has been associated with specific and non-specific attachment to carbohydrate moieties (Esseili et al., 2012a). This association has been described for human NoV VLPs, which are bound with different strengths to the extracts of different plants including coriander, iceberg lettuce, spinach, or romaine lettuce (DiCaprio et al., 2015b). Data involving parsley extracts or extending to other viral models that were different from caliciviruses have not been published to date.

Once viral genomic sequences have been detected, the question whether those viruses represent a risk arises. The sole presence of viral genomes in food does not necessarily represent a biological hazard, as natural inactivation processes may occur during food harvesting. More specific studies on differences in the attachment, stability and potential internalization of viruses should be conducted. These data might be useful when evaluating the risk associated to the consumption of food containing viral genomes.

5. Conclusions

1. A protocol with a high sensitivity for pathogenic viruses present on the surface of fresh vegetables has been described.
2. The application of viral Metagenomics to water and food safety surveillance is a useful tool for investigating, within a single assay, the potential risk associated to presence of viral pathogens in irrigation river water and vegetable food matrices.
3. New viral sequences related to the *Hepeviridae* and *Picornaviridae* families, which may represent new variants/genera within those families, have been detected. These sequences indicate the merit of further studies.
4. Viral pathogens can represent a threat at low concentrations; thus, high sensitivity to detect low number of viral pathogens is critical for water and food safety. Biases related to the relative viral abundance, the difficulty of obtaining viral concentrates from food and water samples, and the limited availability of viral sequences in public databases are among the problems that remain to be solved. The NGS sequencing output has to be carefully analyzed downstream later on, to improve the taxonomic classification of more robust assemblies.
5. New standardized protocols that will be adjusted to the uniqueness of specific food matrices with the aim of addressing all the aforementioned points are needed to introduce Metagenomics effectively to the water and food safety fields.

Uncited references

Acknowledgments

The study reported here was partially funded by the Programa RecerCaixa 2012 (ACUP-00300), AGL2011-30461-C02-01/ALI from the Spanish Ministriy of Science and Innovation and AGL2014-55081-R from the Ministry of Economy. This study was partially funded by a grant from the Catalan Government to Consolidated Research Group VirBaP (2014SRG914), the JPI Water project METAWATER (4193-00001B) and with the collaboration of the Institut de Recerca de l'Aigua (IdRA). During the development of this study, Xavier Fernandez-Cassi was a fellow of the Catalan Government "AGAUR" (FI-DGR); Natalia Timoneda is a fellow of the Spanish Ministry of Science. The authors would like to thank the "Servei de Camps Experimentals" at the Universitat de Barcelona, and especially Josep Matas for their assistance with the parsley growth and irrigation. Also we would like to thank Hagar Azab for her English language assistance.

References

Allen, H.K., Bunge, J., Foster, J.A., Bayles, D.O., Stanton, T.B., 2013. Estimation of viral richness from shotgun metagenomes using a frequency count approach. Microbiome 1, 5. http://dx.doi.org/10.1186/2049-2618-1-5.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410. http://dx.doi.org/10.1016/S0022-2836(05)80360-2.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. http://dx.doi.org/10.1093/nar/25.17.3389.

Ambert-Balay, K., Lorrot, M., Bon, F., Giraudon, H., Kaplon, J., Wolfer, M., Lebon, P., Gendrel, D., Pothier, P., 2008. Prevalence and genetic diversity of Aichi virus strains in stool samples from community and hospitalized patients. J. Clin. Microbiol. 46, 1252–1258. http://dx.doi.org/10.1128/JCM.02140-07.

Ao, Y., Yu, J., Li, L., Jin, M., Duan, Z., 2014. Detection of human norovirus GIV.1 in China: a case report. J. Clin. Virol. 61, 298–301. http://dx.doi.org/10.1016/j.jcv.2014.08.002.

Aw, T.G., Wengert, S., Rose, J.B., 2016. Metagenomic analysis of viruses associated with field-grown and retail lettuce identifies human and animal viruses. Int. J. Food Microbiol. 223, 50–56. http://dx.doi.org/10.1016/j.ijfoodmicro.2016.02.008.

Beller, M., Ellis, A., Lee, S.H., Drebot, M.A., Jenkerson, S.A., Funk, E., Sobsey, M.D., Simmons, O.D., Monroe, S.S., Ando, T., Noel, J., Petric, M., Middaugh, J.P., Spika, J.S., 1997. Outbreak of viral gastroenteritis due to a contaminated well. International consequences. JAMA 278, 563–568. http://dx.doi.org/10.1001/jama.278.7.563.

Benson, D.A., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Sayers, E.W., 2015. GenBank. Nucleic Acids Res. 43, D30–D35. http://dx.doi.org/10.1093/nar/gku1216.

Bernard, H., Faber, M., Wilking, H., Haller, S., Höhle, M., Schielke, A., Ducomble, T., Siffczyk, C., Merbecks, S.S., Fricke, G., Hamouda, O., Stark, K., Werber, D., 2014. Large multistate outbreak of norovirus gastroenteritis associated with frozen strawberries, Germany, 2012. Euro. Surveill. 19, 20719.

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. http://dx.doi.org/10.1128/AEM.00965-06.

Bofill-Mas, S., Rusiñol, M., Fernandez-Cassi, X., Carratalà, A., Hundesa, A., Girones, R., 2013. Quantification of human and animal viruses to differentiate the origin of the fecal contamination present in environmental samples. Biomed. Res. Int. 2013, 192089. http://dx.doi.org/10.1155/2013/192089.

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120. http://dx.doi.org/10.1093/bioinformatics/btu170.

Brister, J.R., Ako-Adjei, D., Bao, Y., Blinkova, O., 2015. NCBI viral genomes resource. Nucleic Acids Res. 43, D571–D577. http://dx.doi.org/10.1093/nar/gku1207.

Calgua, B., Fumian, T., Rusiñol, M., Rodriguez-Manzano, J., Mbayed, V.A., 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. http://dx.doi.org/10.1016/j.watres.2013.02.043.

Callejón, R.M., Rodríguez-Naranjo, M.I., Ubeda, C., Hornedo-Ortega, R., Garcia-Parrilla, M.C., Troncoso, A.M., 2015. Reported foodborne outbreaks due to fresh produce in the United States and European Union: trends and causes. Foodborne Pathog. Dis. 12, 32–38. http://dx.doi.org/10.1089/fpd.2014.1821.

Cantalupo, P.G., Calgua, B., Zhao, G., 2011. Raw Sewage Harbors Diverse Viral Populations. 2, 1–11. http://dx.doi.org/10.1128/mBio.00180-11.Editor.

Coleman, E., Delea, K., Everstine, K., Reimann, D., Ripley, D., 2013. Handling practices of fresh leafy greens in restaurants: receiving and training. J. Food Prot. 76, 2126–2131. http://dx.doi.org/10.4315/0362-028X.JFP-13-127.

DiCaprio, E., Culbertson, D., Li, J., 2015. Evidence of the internalization of animal caliciviruses via the roots of growing strawberry plants and dissemination to the fruit. Appl. Environ. Microbiol. 81, 2727–2734. http://dx.doi.org/10.1128/AEM.03867-14.

DiCaprio, E., Purgianto, A., Ma, Y., Hughes, J., Dai, X., Li, J., 2015. Attachment and localization of human norovirus and animal caliciviruses in fresh produce. Int. J. Food Microbiol. 211, 101–108. http://dx.doi.org/10.1016/j.ijfoodmicro.2015.07.013.

Djikeng, A., Kuzmickas, R., Anderson, N.G., Spiro, D.J., 2009. Metagenomic analysis of RNA viruses in a fresh water lake. PLoS One 4, http://dx.doi.org/10.1371/journal.pone.0007264.

Drexler, J.F., Corman, V.M., Lukashev, A.N., van den Brand, J.M.A., Gmyl, A.P., Brünink, S., Rasche, A., Seggewiß, N., Feng, H., Leijten, L.M., Vallo, P., Kuiken, T., Dotzauer, A., Ulrich, R.G., Lemon, S.M., Drosten, C., 2015. Evolutionary origins of hepatitis A virus in small mammals. Proc. Natl. Acad. Sci. 112, 201516992. http://dx.doi.org/10.1073/pnas.1516992112.

Duhaime, M.B., Deng, L., Poulos, B.T., Sullivan, M.B., 2012. Towards quantitative metagenomics of wild viruses and other ultra-low concentration DNA samples: a rigorous assessment and optimization of the linker amplification method. Environ. Microbiol. 14, 2526–2537. http://dx.doi.org/10.1111/j.1462-2920.2012.02791.x.

Esseili, M.A., Wang, Q., Saif, L.J., 2012. Binding of human GII.4 norovirus virus-like particles to carbohydrates of romaine lettuce leaf cell wall materials. Appl. Environ. Microbiol. 78, 786–794. http://dx.doi.org/10.1128/AEM.07081-11.

Esseili, M.A., Wang, Q., Zhang, Z., Saif, L.J., 2012. Internalization of sapovirus, a surrogate for norovirus, in romaine lettuce and the effect of lettuce latex on virus infectivity. Appl. Environ. Microbiol. 78, 6271–6279. http://dx.doi.org/10.1128/AEM.01295-12.

Ethelberg, S., Lisby, M., Böttiger, B., Schultz, A.C., Villif, A., Jensen, T., Olsen, K.E., Scheutz, F., Kjelso, C., Muller, L., 2010. Outbreaks of gastroenteritis linked to lettuce, Denmark, January 2010. Eur. Secur. 15, 1 (doi:19484 [pii]).

Fogeda, M., Avellón, A., Cilla, C.G., Echevarría, J.M., 2009. Imported and autochthonous hepatitis E virus strains in Spain. J. Med. Virol. 81, 1743–1749. http://dx.doi.org/10.1002/jmv.21564.

Gerba, C.P., Goyal, S.M., LaBelle, R.L., Bodgan, G.F., 1979. Failure of indicator bacteria to reflect the occurrence of enteroviruses in marine waters. Am. J. Public Health 69, 1116–1119. http://dx.doi.org/10.2105/AJPH.69.11.1116.

Greninger, A.L., Naccache, S.N., Federman, S., Yu, G., Mbala, P., Bres, V., Stryke, D., Bouquet, J., Somasekar, S., Linnen, J.M., Dodd, R., Mulembakani, P., Schneider, B.S., Muyembe-Tamfum, J.-J., Stramer, S.L., Chiu, C.Y., 2015. Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. Genome Med. 7, 99. http://dx.doi.org/10.1186/s13073-015-0220-9.

Häfliger, D., Hübner, P., Lüthy, J., 2000. Outbreak of viral gastroenteritis due to sewage-contaminated drinking water. Int. J. Food Microbiol. 54, 123–126. http://dx.doi.org/10.1016/S0168-1605(99)00176-2.

Haramoto, E., Katayama, H., Oguma, K., Yamashita, H., Tajima, A., Nakajima, H., Ohgaki, S., 2006. Seasonal profiles of human noroviruses and indicator bacteria in a wastewater treatment plant in Tokyo, Japan. Water Sci. Technol. 54, 301–308.

Hjelmsø, M.H., Hellmér, M., Fernandez-Cassi, X., Timoneda, N., Lukjancenko, O., Seidel, M., Elsässer, D., Aarestrup, F.M., Löfström, C., Bofill-Mas, S., Abril, J.F., Girones, R., Schultz, A.C., 2017. Evaluation of methods for the concentration and extraction of viruses from sewage in the context of metagenomic sequencing. PLoS One 12, e0170199http://dx.doi.org/10.1371/journal.pone.0170199.

Holm-Hansen, C.C., Midgley, S.E., Fischer, T.K., 2016. Global emergence of enterovirus D68: a systematic review. Lancet Infect. Dis. http://dx.doi.org/10.1016/S1473-3099(15)00543-5.

Holtz, L.R., Finkbeiner, S.R., Zhao, G., Kirkwood, C.D., Girones, R., Pipas, J.M., Wang, D., 2009. Klassevirus 1, a previously undescribed member of the family Picornaviridae, is globally widespread. Virol. J. 6, 86. http://dx.doi.org/10.1186/1743-422X-6-86.

Hurwitz, B.L., Deng, L., Poulos, B.T., Sullivan, M.B., 2013. Evaluation of methods to concentrate and purify ocean virus communities through comparative, replicated metagenomics. Environ. Microbiol. 15, 1428–1440. http://dx.doi.org/10.1111/j.1462-2920.2012.02836.x.

Iwakiri, A., Ganmyo, H., Yamamoto, S., Otao, K., Mikasa, M., Kizoe, S., Katayama, K., Wakita, T., Takeda, N., Oka, T., 2009. Quantitative analysis of fecal sapovirus shedding: Identification of nucleotide substitutions in the capsid protein during prolonged excretion. Arch. Virol. 154, 689–693. http://dx.doi.org/10.1007/s00705-009-0358-0.

Karlsson, O.E., Belák, S., Granberg, F., 2013. The effect of preprocessing by sequence-independent, single-primer amplification (SISPA) on metagenomic detection of viruses. Biosecur. Bioterror. 11 (Suppl. 1), S227–S234. http://dx.doi.org/10.1089/bsp.2013.0008.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., Drummond, A., 2012. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28, 1647–1649. http://dx.doi.org/10.1093/bioinformatics/bts199.

Kirk, M.D., Pires, S.M., Black, R.E., Caipo, M., Crump, J.A., Devleesschauwer, B., Döpfer, D., Fazil, A., Fischer-Walker, C.L., Hald, T., Hall, A.J., Keddy, K.H., Lake, R.J., Lanata, C.F., Torgerson, P.R., Havelaar, A.H., Angulo, F.J., 2015. World Health Organization estimates of the global and regional disease burden of 22 foodborne bacterial, protozoal, and viral diseases, 2010: a data synthesis. PLoS Med. 12, e1001921http://dx.doi.org/10.1371/journal.pmed.1001921.

Kobayashi, S., Fujiwara, N., Yasui, Y., Yamashita, T., Hiramatsu, R., Minagawa, H., 2012. A foodborne outbreak of sapovirus linked to catered box lunches in Japan. Arch. Virol. 157, 1995–1997. http://dx.doi.org/10.1007/s00705-012-1394-8.

Kokkinos, P., Kozyra, I., Lazic, S., Bouwknegt, M., Rutjes, S., Willems, K., Moloney, R., de Roda Husman, A.M., Kaupke, A., Legaki, E., D'Agostino, M., Cook, N., Rzeżutka, A., Petrovic, T., Vantarakis, A., 2012. Harmonised investigation of the occurrence of human enteric viruses in the leafy green vegetable supply chain in three European countries. Food Environ. Virol. 4, 179–191. http://dx.doi.org/10.1007/s12560-012-9087-8.

Koo, H.L., Ajami, N., Atmar, R.L., DuPont, H.L., 2010. Noroviruses: the leading cause of gastroenteritis worldwide. Discov. Med. 10, 61–70. http://dx.doi.org/10.1007/s12560-010-9038-1.Noroviruses.

Kozak, G.K., MacDonald, D., Landry, L., Farber, J.M., 2013. Foodborne outbreaks in Canada linked to produce: 2001 through 2009. J. Food Prot. 76, 173–183. http://dx.doi.org/10.4315/0362-028X.JFP-12-126.

Le Guyader, F.S., Le Saux, J.C., Ambert-Balay, K., Krol, J., Serais, O., Parnaudeau, S., Giraudon, H., Delmas, G., Pommepuy, M., Pothier, P., Atmar, R.L., 2008. Aichi virus, norovirus, astrovirus, enterovirus, and rotavirus involved in clinical cases from a French oyster-related gastroenteritis outbreak. J. Clin. Microbiol. 46, 4011–4017. http://dx.doi.org/10.1128/JCM.01044-08.

Lee, L.E., Cebelinski, E.A., Fuller, C., Keene, W.E., Smith, K., Vinjé, J., Besser, J.M., 2012. Sapovirus outbreaks in long-term care facilities, Oregon and Minnesota, USA, 2002–2009. Emerg. Infect. Dis. 18, 873–876. http://dx.doi.org/10.3201/eid1805.111843.

Li, T.C., Chijiwa, K., Sera, N., Ishibashi, T., Etoh, Y., Shinohara, Y., Kurata, Y., Ishida, M., Sakamoto, S., Takeda, N., Miyamura, T., 2005. Hepatitis E virus transmission from wild boar meat. Emerg. Infect. Dis. 11, 1958–1960.

Maunula, L., Kaupke, A., Vasickova, P., Söderberg, K., Kozyra, I., Lazic, S., van der Poel, W.H.M., Bouwknegt, M., Rutjes, S., Willems, K.A., Moloney, R., D'Agostino, M., de Roda Husman, A.M., von Bonsdorff, C.-H., Rzeżutka, A., Pavlik, I., Petrovic, T., Cook, N., 2013. Tracing enteric viruses in the European berry fruit supply chain. Int. J. Food Microbiol. 167, 177–185. http://dx.doi.org/10.1016/j.ijfoodmicro.2013.09.003.

Namiki, T., Hachiya, T., Tanaka, H., Sakakibara, Y., 2012. MetaVelvet: an extension of velvet assembler to de novo metagenome assembly from short sequence reads. Nucleic Acids Res. 40, http://dx.doi.org/10.1093/nar/gks678.

Ng, T.F.F., Marine, R., Wang, C., Simmonds, P., Kapusinszky, B., Bodhidatta, L., Oderinde, B.S., Wommack, K.E., Delwart, E., 2012. High variety of known and new RNA and DNA viruses of diverse origins in untreated sewage. J. Virol. 86, 12161–12175. http://dx.doi.org/10.1128/JVI.00869-12.

Nordgren, J., Matussek, A., Mattsson, A., Svensson, L., Lindgren, P.-E., 2009. Prevalence of norovirus and factors influencing virus concentrations during one year in a full-scale wastewater treatment plant. Water Res. 43, 1117–1125. http://dx.doi.org/10.1016/j.watres.2008.11.053.

Oishi, I., Yamazaki, K., Kimoto, T., Minekawa, Y., Utagawa, E., Yamazaki, S., Inouye, S., Grohmann, G.S., Monroe, S.S., Stine, S.E., Carcamo, C., Ando, T., Glass, R.I., 1994. A large outbreak of acute gastroenteritis associated with astrovirus among students and teachers in Osaka, Japan. J. Infect. Dis. 170, 439–443. http://dx.doi.org/10.1093/infdis/170.2.439.

Painter, J.A., Hoekstra, R.M., Ayers, T., Tauxe, R.V., Braden, C.R., Angulo, F.J., Griffin, P.M., 2013. Attribution of foodborne illnesses, hospitalizations, and deaths to food commodities by using outbreak data, United States, 1998–2008. Emerg. Infect. Dis. 19, 407–415. http://dx.doi.org/10.3201/eid1903.111866.

Park, E.-J., Kim, K.-H., Abell, G.C.J., Kim, M.-S., Roh, S.W., Bae, J.-W., 2011. Metagenomic analysis of the viral communities in fermented foods. Appl. Environ. Microbiol. 77, 1284–1291. http://dx.doi.org/10.1128/AEM.01859-10.

Pusch, D., Oh, D.Y., Wolf, S., Dumke, R., Schröter-Bobsin, U., Höhne, M., Röske, I., Schreier, E., 2005. Detection of enteric viruses and bacterial indicators in German environmental waters. Arch. Virol. 150, 929–947. http://dx.doi.org/10.1007/s00705-004-0467-8.

Räsänen, S., Lappalainen, S., Kaikkonen, S., Hämäläinen, M., Salminen, M., Vesikari, T., 2010. Mixed viral infections causing acute gastroenteritis in children in a waterborne outbreak. Epidemiol. Infect. 138, 1227–1234. http://dx.doi.org/10.1017/S0950268809991671.

Reuter, G., Boros, , Tóth, Z., Gia Phan, T., Delwart, E., Pankovics, P., 2015. A highly divergent picornavirus in an amphibian, the smooth newt (Lissotriton vulgaris). J. Gen. Virol. 96, 2607–2613. http://dx.doi.org/10.1099/vir.0.000198.

Riveiro-Barciela, M., Mínguez, B., Gironés, R., Rodriguez-Frías, F., Quer, J., Buti, M., 2015. Phylogenetic demonstration of hepatitis E infection transmitted by pork meat ingestion. J. Clin. Gastroenterol. 49, 165–168. http://dx.doi.org/10.1097/MCG.0000000000000113.

Rosario, K., Nilsson, C., Lim, Y.W., Ruan, Y., Breitbart, M., 2009. Metagenomic analysis of viruses in reclaimed water. Environ. Microbiol. 11, 2806–2820. http://dx.doi.org/10.1111/j.1462-2920.2009.01964.x.

Sales-Ortells, H., Fernandez-Cassi, X., Timoneda, N., Dürig, W., Girones, R., Medema, G., 2015. Health risks derived from consumption of lettuces irrigated with tertiary effluent containing norovirus. Food Res. Int. 68, 70–77. http://dx.doi.org/10.1016/j.foodres.2014.08.018.

Sarma, R.H., Saram, M.H., Mukti, H., State University of New York at Albany, 1990. Structure & Methods: Proceedings of the Sixth Conversation in the Discipline Biomolecular Stereodynamics Held at the State University of New York at Albany, June 6–10, 1989. Adenine Press.

Savichtcheva, O., Okabe, S., 2006. Alternative indicators of fecal pollution: Relations with pathogens and conventional indicators, current methodologies for direct pathogen monitoring and future application perspectives. Water Res. 40, 2463–2476. http://dx.doi.org/10.1016/j.watres.2006.04.040.

Svraka, S., Vennema, H., Van Der Veer, B., Hedlund, K.O., Thorhagen, M., Siebenga, J., Duizer, E., Koopmans, M., 2010. Epidemiology and genotype analysis of emerging sapovirus-associated infections across Europe. J. Clin. Microbiol. 48, 2191–2198. http://dx.doi.org/10.1128/JCM.02427-09.

Vega, E., Barclay, L., Gregoricus, N., Williams, K., Lee, D., Vinjé, J., 2011. Novel surveillance network for norovirus gastroenteritis outbreaks, United States. Emerg. Infect. Dis. 17, 1389–1395. http://dx.doi.org/10.3201/eid1708.101837.

Verhoef, L., Hewitt, J., Barclay, L., Ahmed, S.M., Lake, R., Hall, A.J., Lopman, B., Kroneman, A., Vennema, H., Vinjé, J., Koopmans, M., 2015. Norovirus genotype profiles associated with foodborne transmission, 1999–2012. Emerg. Infect. Dis. 21, 592–599. http://dx.doi.org/10.3201/eid2104.141073.

Wang, D., Coscoy, L., Zylberberg, M., Avila, P.C., Boushey, H.A., Ganem, D., DeRisi, J.L., 2002. Microarray-based detection and genotyping of viral pathogens. Proc. Natl. Acad. Sci. U. S. A. 99, 15687–15692. http://dx.doi.org/10.1073/pnas.242579699.

Wang, D., Urisman, A., Liu, Y.T., Springer, M., Ksiazek, T.G., Erdman, D.D., Mardis, E.R., Hickenbotham, M., Magrini, V., Eldred, J., Latreille, J.P., Wilson, R.K., Ganem, D., DeRisi, J.L., 2003. Viral discovery and sequence recovery using DNA microarrays. PLoS Biol. 1, http://dx.doi.org/10.1371/journal.pbio.0000002.

Wu, F.-T., Oka, T., Takeda, N., Katayama, K., Hansman, G.S., Muo, C.-H., Liang, S.-Y., Hung, C.-H., Dah-Shyong Jiang, D., Hsin Chang, J., Yang, J.-Y., Wu, H.-S., Yang, C.-F., 2008. Acute gastroenteritis caused by GI/2 sapovirus, Taiwan, 2007. Emerg. Infect. Dis. 14, 1169–1171. http://dx.doi.org/10.3201/eid1407.071531.

Yamashita, T., Kobayashi, S., Sakac, K., Nakata, S., Chiba, S., Ishihara, Y., 1991. Isolation of cytopathic small round viruses with BS-C-1 cells from patients with gastroenteritis. J. Infect. Dis. http://dx.doi.org/10.1093/infdis/164.5.954.

Yazaki, Y., Mizuo, H., Takahashi, M., Nishizawa, T., Sasaki, N., Gotanda, Y., Okamoto, H., 2003. Sporadic acute or fulminant hepatitis E in Hokkaido, Japan, may be food-borne, as suggested by the presence of hepatitis E virus in pig liver as food. J. Gen. Virol. 84, 2351–2357. http://dx.doi.org/10.1099/vir.0.19242-0.

Zerbino, D.R., Birney, E., 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18, 821–829. http://dx.doi.org/10.1101/gr.074492.107.

Zhang, Y., Zhu, Z., Yang, W., Ren, J., Tan, X., Wang, Y., Mao, N., Xu, S., Zhu, S., Cui, A., Zhang, Y., Yan, D., Li, Q., Dong, X., Zhang, J., Zhao, Y., Wan, J., Feng, Z., Sun, J., Wang, S., Li, D., Xu, W., 2010. An emerging recombinant human enterovirus 71 responsible for the 2008 outbreak of hand foot and mouth disease in Fuyang city of China. Virol. J. 7, 94. http://dx.doi.org/10.1186/1743-422X-7-94.

Zhang, W., Li, L., Deng, X., Kapusinszky, B., Delwart, E., 2014. What is for dinner? Viral metagenomics of US store bought beef, pork, and chicken. Virology 468, 303–310. http://dx.doi.org/10.1016/j.virol.2014.08.025.

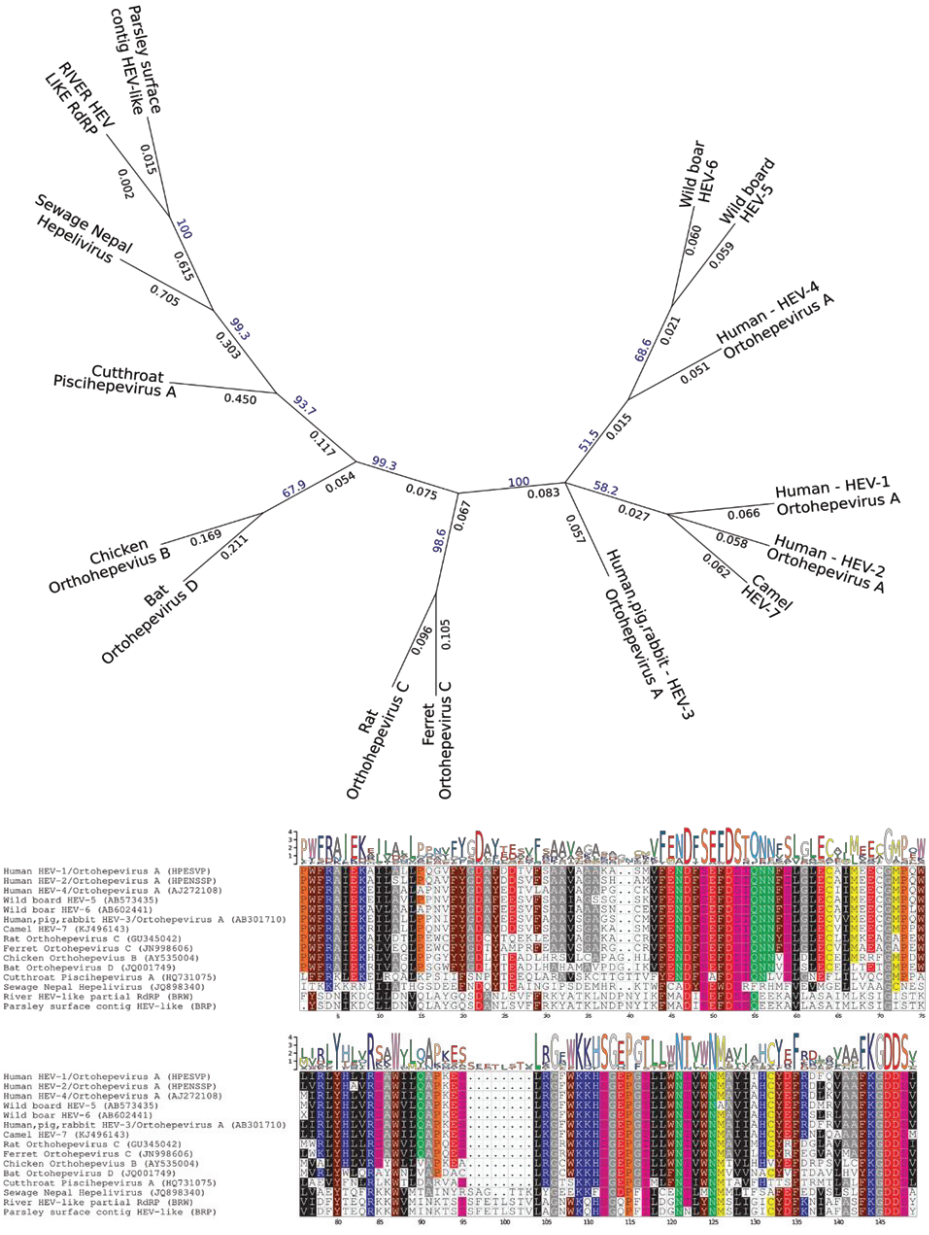


Fig. 1. Phylogenetic analysis of the partial RdRp region for known Hepelivirus and Hepeviridae families, including the River and Parsley surface sequences. A) It is evident that both novel sequence candidates cluster together, and closer to Hepelivirus rather than to Hepeviridae family. Phylogenetic tree was built with Geneious software using Jukes-Cantor model, and clustering method was Neighbor-joining with 1000 bootstrap replicates. Two different scores were included over the branches: bootstrap scores in blue, phylogenetic distances in black. B) The phylogenetic tree was generated from the conserved positions of the RdRp region (204aa), which are shown on this alignment summary produced by Geneious software with default parameters. Sequence accession numbers from GenBank were also annotated with the labels used in the above tree figure. Colors in the alignment correspond to distinct amino acids.

SAMPLES	BRW1			BRW2			BRP			NCP		
Description	Besòs River water			Besòs River water			Parsley Irrigated with river water (20C)			Parsley irrigated with control water		
	Sequences		Nucleotides	Sequences		Nucleotides	Sequences		Nucleotides	Sequences		Nucleotides
Raw Reads (MiSeq)	5,426,788		1,450,732,550	2,476,054		577,865,642	1,192,716		327,348,654	8,067,728		2,199,101,314
Clean Reads												
Pair-Ends	3,955,116		930,629,132	2,445,144		508,229,216	329,520		73,820,015	5,301,560		1,264,595,255
Single-Ends	663		123,178	158		26,175	1,808		148,588	866		243,907
Total	3,955,779	72.89%		2,445,302	98.76%		331,328	27.78%		5,302,426	65.72%	
Assembly (MetaVelvet)												
Contigs	243,043		58,899,523	192,193		44,300,670	20,961		5,871,015	10,296		3,160,250
Singletons	2,556,829		599,432,209	1,664,874		339,082,359	191,276		43,401,095	5,169,800		1,233,416,680
$N50_{contigs+singletons}$	1,149,201		270	715,216		232	83,416		324	2,125,927		372
Homology (BLAST)												
Putative viral seqs	77,626	2.77%		44,607	2.40%		59,380	27.98%		3,878,957	74.88%	
Seqs without BLAST hit	2,702,102	96.51%		1,772,659	95.45%		135,759	63.97%		914,807	17.66%	
# Distinct viral families	22			26			25			16		
# Distinct viral species	361			464			139			60		
	Estimated Value		**SE**	**Estimated Value**		**SE**	**Estimated Value**		**SE**	**Estimated Value**		**SE**
Richness	632.4		44.4	923.5		70.9	255.6		34.7	130.3		34.7

Table 1. Metagenomics sequencing summary statistics. Sequences and nucleotide counts are total values; number of pairs is half of the shown values. Percent of reads sequence refer to the total number of raw reads, while the percent of sequences having, or not, a BLAST hit corresponds to the total number of assembled sequences (contigs plus singletons). Difference between the sequences assigned to known viruses and the sequences without a BLAST hit relates to those sequences having a BLAST hit not passing all the filtering criteria for a valid species assignment.

Table 2. BLASTN statistics for human/animal viruses found using metagenomics in Besòs River Water samples.

Sample	Viral family	Viral specie	# related sequences	Maximum contig length	Blast output statistics		
					Larger contig nucleotide identity (%)	Average query coverage (%)	Match GenBank accession number
River water samples used for irrigation (BRW)	*Adenoviridae*	HAdV-41	2	220	100%	100%	KY316164
	Astroviridae	MAstV-1	5	374	94%	99%	KF039911
	Caliciviridae	Goose Calicivirus	2	221	74%	93%	KJ473715
	Picobirnaviridae	Human Picobirnaviridae	2	304	98%	100%	KJ663813
		Porcine Picobirnavirus	2	360	83%	92%	HM070240
	Reoviridae	Human Rotavirus A	6	296	99%	100%	KU048625
	Picornaviridae	Aichi virus	4	385	95%	100%	GQ927712
		Ampivirus	7	716	74%	99%	KP770140
		Bat picornavirus	1	206	77%	87%	HQ595341
		Rat hunnivirus	1	394	91%	100%	KT944214
		Rodent hepatovirus	2	159	80%	99%	KT452641
	Parvoviridae	Kilham rat virus	1	204	79%	80%	AF321230
		Avian adeno-associated virus	5	258	99%	100%	NC_006263
		Caprine Adeno-associated virus	1	376	99%	98%	DQ335246
		Simian adeno-associated virus	1	202	97%	99%	EU285562
		Porcine bocavirus	3	344	87%	100%	KJ622366
		Mouse parvovirus	1	448	84%	100%	KY489986
		Bovine adeno-associated virus	4	274	79%	95%	AY388617
		Rat bocavirus	5	438	94%	100%	KT454517
		Human adeno associated	12	441	98%	100%	AY530578
	Circoviridae	Gull circovirus	3	174	76%	85%	NC_026625
		Avon-Heathcote Estuary associated circular virus 3	2	369	76%	94%	KT454927
		Cyclovirus	13	487	74%	99%	GQ404854
		Human feces pecovirus	2	232	98%	100%	KT600066
		Porcine circovirus	10	148	76%	82%	NC_027796

Table 3. BLASTN statistics for human/vertebrates viruses found using metagenomics in Parsley Plants irrigated with Besòs River water

Sample	Viral family	Viral specie	Number of related sequences	Maximum contig length	Blast output statistics		
					Larger contig nucleotide identity (%)	Average query coverage (%)	Match GenBank accession number
Parsley plants irrigated with river water samples (BRP)	*Caliciviridae*	Human Norovirus GIV.1	4	360	98%	100%	JQ613567
		Human Sapovirus GI.2	16	1780	95%	100%	AB614356
	Hepeviridae	Hepatitis E genotype 3	8	544	86%	99%	JQ953666
	Flaviviridae	Hepatitis G virus	13	1050	93%	100%	KU685422
	Parvoviridae	Human Bocavirus 2a	10	1287	99%	100%	FJ170280
	Anelloviridae	Torque teno midi virus 1	263	306	95%	88%	AB290918
		Torque teno midi virus 2	388	269	77%	87%	AB290919
		Torque teno mini virus 1	44	237	85%	80%	AB026931
		Torque teno mini virus 2	30	241	80%	82%	AB038629
		Torque teno mini virus 3	68	303	88%	89%	AB038630
		Torque teno mini virus 4	2529	315	77%	82%	AB041963
		Torque teno mini virus 5	918	564	91%	83%	AB041962
		Torque teno mini virus 6	280	314	84%	81%	AB026929
		Torque teno mini virus 7	198	297	93%	88%	AB038627
		Torque teno mini virus 8	2503	305	78%	85%	AF291073
		Torque teno mini virus 9	5	189	94%	89%	AB038631
		Torque teno virus 18	1	150	75%	99%	AX025718
		Torque teno virus 2	1	225	75%	95%	AB049608
		Torque teno virus 29	1	386	87%	100%	AB038621
		Torque teno virus 3	2	191	79%	85%	AY666122
	Circoviridae	Avon-Heathcote Estuary associated circular virus 3	1	278	73%	91%	NC_026625
	Picornaviridae	Enterovirus Species A	11	3706	97%	100%	LT617117
		Enterovirus species B	12	394	97%	99%	KU574626

Article 2

Identification of sapovirus GV.2, astrovirus VA3 and novel anelloviruses in serum from patients with acute hepatitis of unknown etiology

Eloy A Gonzales Gustavson, **N.Timoneda**, X. Fernandez-Cassi, A. Caballero, J.F. Abril, M. Buti ,F. Rodriguez-Frias ,R.Girones.

Manuscrit en procés de revisió a *PLOS ONE*.

Identification of sapovirus GV.2, astrovirus VA3 and novel anelloviruses in serum from patients with acute hepatitis of unknown aetiology

Eloy A. Gonzales Gustavson[a,†], N. Timoneda[a,b,†], X. Fernandez-Cassi[a], A. Caballero[c], J. F. Abril[b,d], M. Buti[c], F. Rodriguez-Frias[c], R. Girones[a]

Eloy A. Gonzales Gustavson: egonzagu15@alumnes.ub.edu

N. Timoneda: natalia.timoneda@gmail.com

X. Fernandez-Cassi: xaviako@gmail.com

A. Caballero: ancaballero@vhebron.net

J. F. Abril: Jabril@ub.edu

M. Buti: mbuti@vhebron.net

F. Rodriguez-Frias: frarodri@gmail.com

R. Girones: rgirones@ub.edu

† These authors contributed equally to this work

[a]Laboratory of Virus Contaminants of Water and Food, Department of Genetics, Microbiology and Statistics, Faculty of Biology, University of Barcelona, Av. Diagonal 643, 08028 Barcelona, Catalonia, Spain

[b]Computational Genomics Lab, Department of Genetics, Microbiology and Statistics, Faculty of Biology, University of Barcelona, Av. Diagonal 643, 08028 Barcelona, Catalonia, Spain

[c]Hospital Universitari Vall d'Hebron, Barcelona, Catalonia, Spain

[d]Institut de Biomedicina de la Universitat de Barcelona (IBUB), Av. Diagonal 643, 08028 Barcelona, Catalonia, Spain

Corresponding author

Prof. Rosina Girones. Email: rgirones@ub.edu. Laboratory of virus contaminants of water and food. Section of Microbiology, Virology and Biotechnology. Department of Genetics, Microbiology and Statistics. Faculty of Biology. University of Barcelona. Diagonal, 643, 08028-Barcelona, Spain. Tel: +34 934021483. Fax: +34 934110592.

http://www.ub.edu/microbiologia_virology/

Abstract

Hepatitis is a general term meaning inflammation of the liver, which can be caused by a variety of viruses. However, a substantial number of cases remain with unknown aetiology. We analysed the serum of patients with clinical signs of hepatitis using a metagenomics approach to characterize their viral species composition. Four pools of patients with hepatitis without identified aetiological agents were evaluated. Additionally, one pool of patients with hepatitis E (HEV) and pools of healthy volunteers were included as controls. The most abundant viruses in pools from patients with hepatitis of unknown aetiology belonged to the *Anelloviridae* family, followed by sapovirus GV.2 and astrovirus VA3. Most of the HEV genome was recovered from the HEV pool. Additionally, GB virus C and human endogenous retrovirus were found in the HEV and healthy pools. Our study provides an overview of the virome in serum from hepatitis patients and describes potentially novel viruses.

Keywords: hepatitis of unknown aetiology, metagenomics, sapovirus GV.2, *Anelloviridae*, astrovirus VA3, hepatitis E

Introduction

Hepatitis is a general term meaning inflammation of the liver and can be caused by a variety of viruses, such as hepatitis A, B, C, D and E [1]. Infectious agents such as bacteria, fungi or parasites, as well as non-infectious agents such as alcohol, drugs or autoimmune diseases, may cause hepatitis. According to the estimates of the Global Burden of Disease study, viral hepatitis is responsible for approximately 1.5 million deaths each year, which is comparable to the number of annual deaths from HIV/AIDS (1.3 million), malaria and tuberculosis (TB) (0.9 million and 1.3 million, respectively) [2] (Lozano et al., 2012).

Viral hepatitis is still one of the key causes of acute liver failure (ALF) in the world. ALF is a devastating clinical syndrome associated with high mortality in the absence of immediate care, specific treatment or liver transplantation [3] (Manka et al., 2016). Globally, hepatitis A, B and E infections are probably responsible for the majority of ALF cases. However, despite significant progress in the diagnosis and treatment of hepatitis, in a considerable number of patients, the aetiological agents remain unknown. Previous studies have found that between 3.8% and 33.9% of hospital inpatients with acute hepatitis had non-A-E-hepatitis [4–8] (Alter et al., 1997; Cacopardo et al., 2000; Chu et al., 1999; Delic et al., 2010; Tassopoulos et al., 2008). Additionally, 10% of patients with ALF had non-A-E hepatitis [9] (Yeh et al., 2006).

Therapeutic trials using interferon-α to treat hepatitis of unknown aetiology have consistently resulted in response rates of approximately 50%, indicating a virological aetiology [10] (Van Thiel et al., 1994). This evidence suggests that other viruses may be responsible for hepatitis. As a result, new viruses, including a *Flaviviridae* GB virus type C (GBV-C) [11] (Simons et al., 1995) and *Anelloviridae* TTV and SEN virus [12] (Nishizawa et al., 1997), have been reported in recent years to be associated with hepatitis. However, epidemiological data

failed to confirm a causative role for those viruses in the development of hepatitis, and a high percentage of individuals infected by them were found to be healthy carriers. Recent investigation has shown that other viral infections such as cytomegalovirus and Epstein Barr virus may mimic viral hepatitis [13] (Conrad and Knodell, 2014). Less frequently, hepatitis may be present in people with herpes simplex virus [14] (Kaufman et al., 1997), parvovirus B19 [15] (Bihari et al., 2013), and adenoviruses 1, 2, 5, 12 and 32 [16,17] (Kawashima et al., 2015; Mateos et al., 2012).

Epidemiologic information related to non-A-E hepatitis is scarce. History of blood transfusion, drug use or other parenteral exposure were not associated with the onset of illness [7] (Delic et al., 2010). If the viral nature of non-A-E hepatitis is proven, it should spread primarily by non-parenteral means. Moreover, some patients diagnosed with acute non-A-E hepatitis show biochemical features at admission similar to those associated with other viral hepatitis. Apparently, acute non-A-E hepatitis is distributed worldwide, and progression to chronicity was observed in approximately 9% of patients [7,18] (Chu et al., 2001; Delic et al., 2010).

The cause of acute non-A-E hepatitis remains unknown. It seems likely that another as-yet-unidentified infectious agent(s) exists [18] (Chu et al., 2001). Recent rapid progress in sequencing technologies and associated bioinformatics methodologies has enabled a more in-depth view of the structure and functioning of viral communities, supporting the characterization of emerging viruses [19] (Ogilvie and Jones, 2015). With the advent of metagenomics studies, our knowledge of the different components and the complexity of the microbiome greatly expanded. The eukaryotic virome comprises viruses infecting the host, endogenous viral

elements, and viruses associated with other eukaryotic components of the ingesta [20] (Hugenholtz and Tyson, 2008).

In this study, next-generation sequencing (NGS) was used to identify viruses in serum samples from patients suffering from acute hepatitis signs. For that purpose, the viromes in the serum of patients with Non-A-E hepatitis were analysed and the results were compared with the viromes from patients with acute hepatitis E (positive controls) and healthy patients (negative controls).

Materials and Methods

Serum Samples

Serum samples were collected from patients with acute viral hepatitis from the Vall d'Hebron Hospital, Barcelona, Spain. The clinical diagnosis of acute viral hepatitis was based on the lack of previous history of chronic liver disease, a rise in serum aminotransferase (AST, ALT) activity of at least 200 IU/L, high values of total (TB) and direct bilirubin (DB) and exclusion of other causes of liver disease such as hepatitis A (Ig-M negative), hepatitis B (surface-antigen-HBsAg- and anti-core antibodies-anti-HBc-negative-), hepatitis C (anti-VHC-negative) and hepatitis E (HEV) (IgG, IgM and RT-PCR, all negatives). Of the 32 patients selected, 19 were male, and 13 were female, with ages ranging from 1 to 92 years old. Eight of those patients were diagnosed with an autoimmune or immunosuppressed (Ai+ImSP) condition. Additionally, serum from 10 patients positive for HEV by nested RT-PCR were included as positive controls. Finally, serum samples from 20 healthy volunteers were also evaluated.

The serum samples were pooled according to the following criteria. Patients with acute hepatitis were grouped into five pools: male pool A (age range from 1 to 43), male pool B (age

range from 48 to 78), a female pool (age range from 6 to 92), and an Ai+ImSP pool (age range from 2 to 84) that included patients with the Ai+ImSP condition. Each pool included eight individual serum samples. Finally, there was an HEV pool (age range from 6 to 84) aggregating 10 HEV RNA-positive patients. Volunteers' serum samples were grouped in two pools and evaluated in duplicate: Healthy A1 and A2 pools, with 10 females (age range between 27 and 63), and Healthy B1 and B2 pools, with two males and eight females (age range between 26 and 58).

Sample Preparation

Serum samples were kept at -80 °C prior to the metagenomics analysis protocol. Pools were prepared with the corresponding serum samples to achieve an initial volume of 500 µL. Briefly, the pools were first filtered through a pore size of 0.45 µm (Millipore Corp., Billerica, MA, USA) to remove cellular debris, ultracentrifuged at 100,000 × g for 90 min at 4 °C and re-suspended in 500 µL of PBS 1X. Next, 300 µL of the re-suspended pool was subjected to DNAse treatment to eliminate background DNA with 20 U TURBO™ DNase (Ambion, Thermo Fisher Scientific, Waltham, MA, USA). Then, viral nucleic acids (NAs) were extracted with QIAmp Viral RNA Mini Kit (Qiagen, Inc., Valencia, CA), without carrier RNA, according to the manufacturer`s instructions. To enable the detection of both DNA and RNA viruses, total NAs were reverse-transcribed as previously described [21,22] (Wang et al., 2002, 2003). In short, SuperScript II (Life Technologies, California, USA) was used to retro-transcribe RNA to cDNA with primerA (5'-GTTTCCCAGTCACGATCNNNNNNNNN-3'). Second-strand cDNA and DNA were constructed with the primer sequences using Sequenase 2.0 (USB/Affymetrix, Cleveland, OH, USA). PCR amplification with AmpliTaqGold (Life Technologies, Austin,

Texas, USA) was performed using primerB (5'-GTTTCCCAGTCACGATC-3') with 30 cycles; this step was run in duplicate. The PCR products were purified and eluted in 15 µL using a Zymo DNA Clean and Concentrator kit (cat nº D4013, Zymo Research, USA) to yield enough DNA for the library preparation.

Sequencing Protocol

NGS sequencing was performed at SGB-UAB, Barcelona. dsDNA samples were quantified by Qubit 2.0 (Life technologies), and libraries were constructed using a Nextera XT DNA sample preparation kit (Illumina Inc). Samples were sequenced on Illumina MiSeq 2x300; all samples were multiplexed and distributed within three independent sequencing runs.

NGS Data Processing

The quality of raw and clean read sequences was assessed using FASTX-Toolkit software, version 0.0.14 (Hannon Lab) [23] (Hannon, 2015). The sequenced reads were cleaned by Trimmomatic version 0.32 [24] (Bolger et al., 2014) while the sequencing adaptors and linker contamination were removed. Low-quality ends were trimmed using a Phred score average threshold above Q15 over a running window of four nucleotides. Low-complexity sequences, mostly repetitive sequences that would affect the performance of downstream procedures in the computational protocol, were then discarded after estimating a linear model based on Trifonov's linguistic complexity and the sequence string-compression ratio. The discrimination criteria for that linear model assumes low complexity scores below the line having a -45º slope and crossing data distribution at 5% below the complexity inflexion point found by the model, which is

specific to each sequence set. Finally, duplicated reads were removed in a subsequent step to speed up the downstream assembly.

Sequence Assembly and Taxonomic Assignment

Clean and filtered MiSeq reads were assembled using as parameters 90% identity over a minimum of 50% of the read total length in CLC Genomics Workbench 4.4 (CLC bio USA, Cambridge, MA) [25] (CLC Bio, 2008). Afterwards, contigs longer than 100 bp were queried for sequence similarity using BLASTN and BLASTX (NCBI-BLAST ([26] Altschul et al., 1990)) against the NCBI complete viral genomes database [27,28] (Altschul et al., 1997; Brister et al., 2015), the viral division of the GenBank nucleotide database [29,30] (Clark et al., 2016; NCBI, 2016), and viral proteins from UniProt [31] (Bateman et al., 2015). The species nomenclature and classification followed NCBI Taxonomy database standards and the basic Baltimore classification. The alignments reported by BLAST (High-scoring Segment Pairs, HSPs) were required to have an E-value lower than 10^{-5} and a minimum length of 100 bp to be considered for taxonomical assessment. On the basis of the best BLAST results and a 90% coverage cut-off, the sequences were classified into their most likely taxonomic groups of origin.

Phylogenetic Analysis

For *Anelloviridae*, we included in the phylogenetic tree all contigs covering the complete ORF1 with all the representative members of this family previously reported in humans that have been used as reference strains. Additionally, we also included some contigs longer than 1,500 bp that overlapped a large segment of ORF1 or a region upstream for individual trees. We compared each tree with the main tree generated from the reference strains to confirm equivalent

distribution of species. In this manuscript, the following notation criteria were applied to name sequences on the phylogenetic trees: sequences covering ORF1, partially or not, were assigned to a number; contigs having some part outside ORF1 were identified with letters. For *Hepeviridae*, we also generated individual trees for each filtered contig; those were also named using numbers when the generated tree was similar to the reference tree. All the phylogenetic trees were computed in Geneious 10® using the neighbour-joining method under the Jukes Cantor model. The robustness of the trees was assessed by bootstrap analysis of 1000 replicates each; finally, the branches are proportional to the corresponding phylogenetic distance.

Ethical statement

The study has been approved by the corresponding ethical committee: ethical committee on clinical investigation and research projects of the Hospital Universitari Vall D'Hebron. Serum samples were pooled at the hospital and in this study we do not have information on the identity of the patients

Results

Nine libraries, consisting of 62 serum samples, were obtained and sequenced using paired-end 300-base runs on the Illumina MiSeq platform, generating a total of 48 million reads (see Table 1 for a summary of the sequencing statistics of individual pools). Raw reads were binned by pool-based library barcodes and quality-filtered, leaving 30.5 million high-quality reads, which were assembled *de novo* within each pool subset. The resulting sequence contigs and singlets were compared to NCBI complete viral genomes, the viral division of the GenBank nucleotide database, and viral proteins from UniProt. Most of the viral sequences detected were

related to the *Anelloviridae*, *Astroviridae*, *Caliciviridae*, *Hepeviridae*, *Flaviviridae* and *Retroviridae* families (see Fig. 1); those near-to-complete or partial genomes were characterized and are described in the following sections.

Volunteer samples that were analysed in the Healthy pools, in duplicate, show similar number of reads, and contigs. Additionally, the same families were found in those replicates, demonstrating that those results are highly consistent (Table and Fig. 1).

Hepeviridae

A total of 27 contigs were matched to the *Hepeviridae* family. The HEV and Ai+ImSP pools produced sequences related to this family. A total of 76.1% (5,508 of 7,238 bp) of the HEV genome was sequenced from the HEV pool, with an average pairwise identity of 85.5% against the genotype 3 HEV (AF082843, Reference sequence genotype 3 ICTV). To identify the genotypes present in the pools and because metagenomics amplified different regions of the genome at random, individual phylogenetic trees were computed from contigs mapping over the same reference genome locations. The individual trees were compared to a reference species tree based on the reference-genomic sequences. Contigs that produced trees similar to the reference are marked in Fig. 2 using numeric indexes, and information about each of those contigs is displayed on Table 2. We were able to generate phylogenetic trees similar to the reference for eighteen contigs (the individual trees are available in Supplementary File A). Fifteen contigs from the HEV pool aligned to genotype 3f or closely related genotypes. The three contigs from the Ai+ImSP pool aligned with genotype 3a.

Anelloviridae

A total of 3,286 contigs matched sequences from the *Anelloviridae* family. All the pools produced sequences related to this family; however, the number of contigs was significantly higher in the pools with signs of hepatitis compared to the healthy pools (Wilcoxon rank-sum test, p= 0.009) and much more abundant in the Ai+ImSP pool (Fig. 1). Contigs completely covering the ORF1 region of *Anelloviridae* family – or longer than 1,500 bp and overlapping this region – were found in the male A (less than 48 years old), female, HEV, and Ai+ImSP pools. Those particularly long sequences were used to build a phylogenetic tree to obtain a more accurate characterization of the species (Fig. 3 and Table 3). The main members detected were Torque Teno Viruses (TTV - genus *Alphatorquevirus*) 1, 5, 10, 11, 13, 16, 18, 19, SEN virus H, Torque Teno Mini Viruses (TTMV - genus *Betatorquevirus*) 5, 9 and 18, Torque Teno Midi Viruses (TTMDV - genus *Gammatorquevirus*) 1, MDJN47, MDJN97, and other unclassified anelloviruses: TTV P19-3 (KT163917), TTV S72 (KP343839), TTV P1-3 (KT163877), TTV P13-4 (KT163899), TTMV Emory1 (KX810063), TTV S97 (KP343864), TTMV LY3 (JX134046), TTV S66 (KP343833), TTV S69 (KP343836), TTV S45 (KF545591), TTV P9-6 (KT163891), TTV S80 (KP343847), and TTV S57 (KP343824). Furthermore, contigs matching to the last two reference sequences do not belong to the three known genera of *Anelloviridae* previously identified in humans; thus, it seems they define a new cluster/genus for this family. Moreover, 60% (19/32) of the longest contigs have less than 80% identity to the already described sequences from the NCBI database. Table 3 shows the contigs that were considered for the phylogenetic analysis. Each contig is identified by its name, length, identity to the blast HSPs, and bootstrap on the corresponding phylogenetic tree (individual trees are provided in the Supplementary File B). Fewer and shorter contigs were found in the pools from healthy

individuals in comparison with the other pools (median of 300 bp); they correspond to TTV 1, 19 and TTMV 6.

Caliciviridae

A total of 35 contigs between 200 and 654 bp aligned to the *Caliciviridae* family. They were found in the male A and B, female and Ai+ImSP pools. No sequences of this family were detected in healthy volunteer pools. All contigs were assigned to sapovirus Hu/Nagoya/NGY (AB775659), genogroup 5 strain 2 (GV.2), with identities varying between 97% and 100%. Those contigs map over several regions of the non-structural protein and major structural protein, including eleven that aligned to a partial capsid fragment.

Astroviridae

As few as eight contigs between 214 and 493 bp long matched the *Astroviridae* family. They were found in the male A (less than 48 years old), female, and Ai+ImSP pools. No sequences of this family were detected in the healthy-volunteers pools. These contigs correspond to a recently discovered astrovirus, clade VA strain 3 (VA3, also known as HMO-C) (7 matching JX857868, 1 matching JX083288), with identities ranging from 97% to 100%.

Flaviviridae

A total of 65 contigs between 219 and 2778 bp matched the *Flaviviridae* family. They were found in the female, Ai+ImSP, and healthy B1 and B2 pools. All the sequences aligned to several entries of GB virus C from GenBank, with identities between 97% and 100%.

Retroviridae

In this case, 285 contigs between 300 and 1,032 bp were assigned to the *Retroviridae* family. They were found in the male B (more than 48 years old), female, and HEV pools and in all healthy pools. All the sequences matched several entries of human endogenous retrovirus type K and HCML-ARV with identities greater than 70%.

The raw sequencing data used to perform this analysis along with the FASTQ file are located in the NCBI Sequence Read Archive; BioProject (PRJNA379441).

Discussion

The aim of this study was to investigate viruses infecting patients diagnosed with acute hepatitis. Different groups of patients presenting with acute hepatitis but without serological infection markers of the most common viral hepatitis were studied to determine possible causal agents of non-A-E hepatitis. Our findings demonstrate the presence of a high variety of viral sequences in all evaluated samples.

HEV viruses were detected in two pools (HEV and Ai+ImSP). We found a variety of contigs related to genotype 3f in HEV pools. Genotype 3f has been described in hepatitis outbreaks in Catalonia [32] (Riveiro-Barciela et al., 2015), Spain [33] (Rivero-Juarez et al., 2017) and the south of France [34] (Legrand-Abravanel et al., 2009). This strain has also been related to swine and wild boar consumption, which can be considered a food-borne and an emerging zoonotic infection [32,35,36] (Banks et al., 2004; Meng et al., 1997; Riveiro-Barciela et al., 2015). Individual samples from the Ai+ImSP pool were re-analysed afterwards by nPCR, and one patient was identified as HEV-positive in this second round, which would explain the presence of HEV contigs in this pool. Metagenomics approaches have the advantage of

identifying more than one genotype in the pools; this facilitates description of traces of possible multiple infections in a single sample.

An increased number of contigs aligning to anelloviruses was observed in this study, supporting the hypothesis that these viruses are not causative agents. Previous studies have suggested titres of TTV in plasma as an indicator of immune status [37] (Touinssi et al., 2001). Another study showed that anellovirus load in plasma increases substantially during immunosuppressive therapy and in immunocompromised patients [38] (Li et al., 2013). Shotgun sequencing from plasma samples that were collected over several months post-transplantation also revealed that viral loads increased, whereas the bacterial composition remained unchanged [39] (Hofer, 2014).

TTV-1, the first member identified in the *Anelloviridae* family, was reported in hepatitis patients in whom no causative agents were detected [12] (Nishizawa et al., 1997). The *Anelloviridae* are unenveloped viruses with icosahedral capsids and a diameter of approximately 30 nm. This family includes three genera that have been identified in humans: *Alphatorquevirus* (TTV), *Betatorquevirus* (TTMV), and *Gammatorquevirus* (TTMDV) [40] (Biagini et al., 2012). However, the role of those viruses in hepatitis or in other diseases remains uncertain [41–43] (Hsiao et al., 2016; Okamoto, 2009; Spandole et al., 2015). Numerous recent studies have demonstrated a prevalence between 5 and 90% in the blood of the general population, depending on the geographic region [43] (Spandole et al., 2015). Moreover, the genetic diversity among anelloviruses is far greater than it is within any other group of ssDNA viruses. The considerable genetic heterogeneity is exemplified by the large number of highly divergent sequences being identified in this family. There are at least 41 species infecting humans that are recognized by the ICTV based on the ORF1 region [40] (Biagini et al., 2012). Some viruses, such as TTV 1, 12,

13, 16, SEN virus D and H, have been considered potential causal agents of hepatitis. A mixed infection of genotypes or a combination with other microorganisms has been proposed as the cause of the high diversity of pathologies associated with these viruses [42,44–46] (Bostan, 2013; Kakkola et al., 2008; Mi et al., 2014; Okamoto, 2009).

We have found at least three different kinds of *Anelloviridae* contigs: a) contigs that match previously characterized sequences; b) contigs that are closely related to unclassified sequences; and, c) contigs poorly related to classified and unclassified sequences (potential new viruses). The demarcation criteria of the genus establish a cut-off value of 35% nucleic-acid identity in the ORF1 region. Due to the number of quasispecies discovered in this family [43] (Spandole et al., 2015), it is difficult to establish a clear cut-off at the species level.

Metagenomics analyses are driving the discovery of new potential sequences in this family; Bzhalava et al. (2016) described for first time a group of sequences detected from human samples, spawning a new branch of the *Anelloviridae* family. We found two contigs (125 and 1199) falling into this new potential genus of *Anelloviridae*, yet they have less than 70% of identity to those sequences, which were described in serum samples from pregnant women. Such results suggest that there will be more viruses within this family that have not yet been identified.

We also describe in this paper viruses that have been previously associated with hepatitis such as TTV-1, 11, 16 and SEN virus H [42,47] (Luo et al., 2002; Okamoto, 2009); other viruses have been recently described in serum samples from HIV patients (P19-3, P13-4, P9-6, P1-3); yet other sets were described in patients with various conditions, including lymphocytic leukaemia (TTV 10) [48](Chu et al., 2011), gingival periodontitis (TTMV 18) [49] (Zhang et al., 2016), and haemophilia (TTMDV MDJN47 and MDJN97) [50] (Ninomiya et al., 2007) and in

pregnant women whose offspring developed leukaemia and lymphomas (TTV S45, S57, S66, S69, S72, S80 and S97) [51] (Bzhalava et al., 2016).

Unfortunately, anelloviruses cannot be propagated *in vivo* due to the lack of compatible cell systems. However, they have a high *in vivo* replication capacity. Infection with TTV is characterized by persistent lifelong viremia in humans, with circulation levels of up to 10^6 genomic copies/ml in the general population [42,43] (Okamoto, 2009; Spandole et al., 2015). TTV replicates in the liver and is excreted at high levels in bile and faeces [52] (Ohbayashi et al., 2001). Additionally, viral particles are frequently found in wastewater and are used as an indicator of viral contamination [53] (Griffin et al., 2008). Additionally, studies have shown that TTV does not have a particular tropism; replication can occur in the bone marrow, lymph nodes, spleen, pancreas, thyroids, muscles, lungs, kidney and peripheral blood mononuclear cells [43,54] (Okamoto et al., 2004; Spandole et al., 2015). Metagenomic analyses have also shown that TTV is a common finding in several sample types [55] (Delwart, 2007). For that reason, determining the causative factors of illness can be difficult.

The results described in this study also show the presence of sapovirus strain GV.2 in all the pools of patients with clinical hepatitis of unknown aetiology. This strain has been recently characterized from faecal samples from a suspected foodborne gastroenteritis outbreak in Japan using a metagenomics sequencing approach [56] (Shibata et al., 2015). Partial fragments of that virus were described early from another gastroenteritis outbreak in Italy [57] (Medici et al., 2012), in river water from Barcelona (the same region where this study was conducted) [58] (Sano et al., 2011), and in wastewater from Japan [59] (Hansman et al., 2007), suggesting prevalent circulation of this virus around the world. Sapovirus are positive-sense single-stranded RNA viruses from the family *Caliciviridae*. Members of this family are known to cause

gastroenteritis with self-limited infections and low mortality rates; severe infections or serious clinical complications are usually reported in immunocompromised patients [60] (Oka et al., 2015). Further studies would be required to analyse the possible pathogenic role of sapovirus GV.2 in our study.

Few contigs of the *Astroviridae* family were detected in this work. Astrovirus VA3 was identified in most of the pools of hepatitis of unknown aetiology. However, those contigs were less abundant and shorter than the sapovirus contigs. The first description of astrovirus VA3 was from the stool of paediatric patients with diarrhoea from India [61] (Finkbeiner et al., 2009), and it was later completely sequenced [62] (Jiang et al., 2013). This virus has also been described in stools from southern China [63] (Xiao et al., 2011), Kenya, and the Gambia [64] (Meyer et al., 2015). However, the role of this virus in health and disease remain largely unknown.

The presence of gastrointestinal viruses in serum samples has been described before. Noroviruses (NoV), important members of the *Caliciviridae* family, have been identified in blood from children with acute gastroenteritis, immunocompromised adults, and gnotobiotic pigs and calves [65–70] (Cheetham et al., 2006; Frange et al., 2012; Fumian et al., 2013; Lemes et al., 2014; Souza et al., 2008; Takanashi et al., 2009), suggesting that NoV infection may not be limited to the gut. Nevertheless, NoV was not present at detectable levels in serum samples from immunocompetent human adults [71] (Newman et al., 2015). The potential pathogenic role of sapovirus V.2 and astrovirus VA3 in blood remains still uncertain, although our results suggest that the presence of these viruses in pools from patients with non A-to-E hepatitis, including the AI+ImSp pool, merits further research, since there is no previous evidence relating those viruses to hepatitis.

GB virus C, also known as pegivirus or hepatitis G virus, is a human virus of the *Flaviviridae* family that is structurally and epidemiologically closest to hepatitis C virus [72] (Chivero and Stapleton, 2015). Most GBV-C infections appear to be asymptomatic, transient, and self-limiting, with slight or no elevation of ALT levels. Those infections are rarely identified and very difficult to evaluate. The role of GBV-C in the aetiology of hepatitis has not been fully established [73] (Leary and Mushahwar, 2004). Moreover, it is commonly reported in metagenomics studies [55] (Delwart, 2007), suggesting its limited role in the development of illness, including hepatitis. We have detected this virus in one healthy pool and in a hepatitis pool, and our results support the hypothesis that this species may be widely distributed within the population.

Human endogenous retroviruses (HERVs) are remnants of germ-line retrovirus integration and are considered functionally defective [74] (Van der Kuyl, 2012). They has been described in metagenomics studies at high levels [38,74] (Li et al., 2013; Van der Kuyl, 2012) without association with any particular pathology [75] (Canuti et al., 2015). Our findings support previous results that this virus is present in healthy people.

Conclusions

In summary, metagenomics was applied in this study to detect a broad spectrum of viral species based on sequences found in the samples, including HEV in pools of patients with confirmed HEV; these samples allowed the characterization of the most prevalent genotypes. Additionally, we were able to identify novel undescribed sequences of anellovirus, sapovirus GV.2, and astrovirus VA3 in patients with acute hepatitis of unknown aetiology. We did not attempt to determine causality or to describe epidemiologic results; our purpose was to

409 characterize the virome of patients diagnosed with hepatitis to describe new potential causal
410 agents. Metagenomics analyses offer unprecedented possibilities for diagnostics, characterization
411 and identification of possible co-infections of rare and novel viruses that will be relevant to
412 understanding the aetiology of current pathologies without known causative agents.
413
414 **Acknowledgements**
415 The study reported here was partially funded by the Programa RecerCaixa 2012 (ACUP-00300)
416 and AGL2011-30461-C02-01/ALI from the Spanish Ministry of Science and Innovation. This
417 study was partially funded by a grant from the Catalan Government to Consolidated Research
418 Group VirBaP (2014SRG914), the JPI Water project METAWATER (4193-00001B) and with
419 the collaboration of the Institut de Recerca de l'Aigua (IdRA). During the development of this
420 study, Eloy Gonzales-Gustavson is a fellow of the Peruvian Government; Natalia Timoneda is a
421 fellow of the Spanish Ministry of Science and Xavier Fernandez-Cassi was a fellow of the
422 Catalan Government "AGAUR" (FI-DGR).
423
424 **Declaration of Interest**
425 The authors have no conflicts of interest to declare
426

References

1. Previsani N, Lavanchy D, Siegl G. Hepatitis A. In: Mushahwar Isa K, editor. Viral Hepatitis: Molecular Biology, Diagnosis, Epidemiology and Control. 1st ed. Amsterdam: ElSevier; 2004. pp. 1–30.

2. Lozano R, Naghavi M, Foreman K, Lim S, Shibuya K, Aboyans V, et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: A systematic analysis for the Global Burden of Disease Study 2010. Lancet. 2012;380: 2095–2128. doi:10.1016/S0140-6736(12)61728-0

3. Manka P, Verheyen J, Gerken G, Canbay A. Liver Failure due to Acute Viral Hepatitis (A-E). Visc Med. 2016; 80–85. doi:10.1159/000444915

4. Alter MJ, Gallagher M, Morris TT, Moyer LA, Meeks EL, Krawczynski K, et al. Acute Non-A–E Hepatitis in the United States and the Role of Hepatitis G Virus Infection. N Engl J Med. Massachusetts Medical Society; 1997;336: 741–746. doi:10.1056/NEJM199703133361101

5. Cacopardo B, Nunnari G, Berger A, Doer HW, Russo R. Acute non A-E hepatitis in eastern Sicily: the natural history and the role of hepatitis G virus. Eur Rev Med Pharmacol Sci. 2000;4: 117–121. Available: http://www.ncbi.nlm.nih.gov/pubmed/11710508

6. Chu C-M, Lin S-M, Hsieh S-Y, Yeh C-T, Lin D-Y, Sheen I-S, et al. Etiology of sporadic acute viral hepatitis in Taiwan: The role of hepatitis C virus, hepatitis E virus and GB virus-C/hepatitis G virus in an endemic area of hepatitis A and B. J Med Virol. John Wiley & Sons, Inc.; 1999;58: 154–159. doi:10.1002/(SICI)1096-9071(199906)58:2<154::AID-JMV9>3.0.CO;2-E

7. Delic D, Mitrovi N, Radovanovi A. Epidemiological characteristics and clinical manifestations of acute non-A-E hepatitis. Vojnosanit Pregl. 2010;67: 903–909.

8. Tassopoulos NC, Papatheodoridis G V., Delladetsima I, Hatzakis A. Clinicopathological features and natural history of acute sporadic non-(A-E) hepatitis. J Gastroenterol Hepatol. 2008;23: 1208–1215. doi:10.1111/j.1440-1746.2008.05454.x

9. Yeh C-T, Tsao M-L, Lin Y-C, Tseng I-C. Identification of a novel single-stranded DNA fragment associated with human hepatitis. J Infect Dis. 2006;193: 1089–97. doi:10.1086/501474

10. Van Thiel DH, Gavaler JS, Baddour N, Friedlander L, Wright HI. Treatment of putative non-A, non-B, non-C hepatitis with alpha interferon: a preliminary trial. J Okla State Med Assoc. UNITED STATES; 1994;87: 364–368.

11. Simons JN, Leary TP, Dawson GJ, Pilot-Matias TJ, Muerhoff a S, Schlauder GG, et al. Isolation of novel virus-like sequences associated with human hepatitis. Nat Med. 1995;1: 564–569. doi:10.1038/nm0695-564

12. Nishizawa T, Okamoto H, Konishi K, Yoshizawa H, Miyakawa Y, Mayumi M. A novel DNA virus (TTV) associated with elevated transaminase levels in posttransfusion hepatitis of unknown etiology. Biochem Biophys Res Commun. 1997;241: 92–7. doi:10.1006/bbrc.1997.7765

13. Conrad M, Knodell R. Viral hepatitis—1975. JAMA. 2014;312: 654. Available: http://dx.doi.org/10.1001/jama.2013.279664

14. Kaufman B, Gandhi SA, Louie E, Rizzi R, Illei P. Herpes simplex virus hepatitis: case report and review. Clin Infect Dis. 1997;24: 334–338.

15. Bihari C, Rastogi a, Saxena P, Rangegowda D, Chowdhury a, Gupta N, et al. Parvovirus B19 Associated Hepatitis. Hepat Res Treat. 2013;2013: 472027. doi:10.1155/2013/472027

16. Kawashima N, Muramatsu H, Okuno Y, Torii Y, Kawada J, Narita A, et al. Fulminant adenovirus hepatitis after hematopoietic stem cell transplant: Retrospective real-time PCR analysis for adenovirus DNA in two cases. J Infect Chemother. Elsevier Ltd; 2015;21: 857–863. doi:10.1016/j.jiac.2015.08.018

17. Mateos ME, López-Laso E, Pérez-Navero JL, Peña MJ, Velasco MJ. Successful response to cidofovir of adenovirus hepatitis during chemotherapy in a child with hepatoblastoma. J Pediatr Hematol Oncol. 2012;34: e298-300. doi:10.1097/MPH.0b013e318266ba72

18. Chu C-M, Lin D-Y, Yeh C-T, Sheen I-S, Liaw Y-F. Epidemiological Characteristics, Risk Factors, and Clinical Manifestations of Acute Non-A±E Hepatitis. J Med Virol J Med Virol. 2001;65.

19. Ogilvie L, Jones B. The human gut virome: a multifaceted majority. Front Microbiol. 2015;6: 918. doi:10.3389/fmicb.2015.00918

20. Hugenholtz P, Tyson GW. Metagenomics. Nature. 2008;455: 481–483. doi:10.1038/455481a

21. Wang D, Coscoy L, Zylberberg M, Avila PC, Boushey H a, Ganem D, et al. Microarray-based detection and genotyping of viral pathogens. Proc Natl Acad Sci U S A. 2002;99: 15687–92. doi:10.1073/pnas.242579699

22. Wang D, Urisman A, Liu YT, Springer M, Ksiazek TG, Erdman DD, et al. Viral discovery and sequence recovery using DNA microarrays. PLoS Biol. 2003;1. doi:10.1371/journal.pbio.0000002

23. Hannon G. FASTX-Toolkit [Internet]. 2015. Available: http://hannonlab.cshl.edu/fastx_toolkit/index.html

24. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30: 2114–2120. doi:10.1093/bioinformatics/btu170

25. CLC Bio. CLC Genomics Workbench [Internet]. 2008. Available: https://www.qiagenbioinformatics.com/products/clc-genomics-workbench/

26. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215: 403–410. doi:http://dx.doi.org/10.1016/S0022-2836(05)80360-2

27. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res. 1997;25: 3389–3402. doi:10.1093/nar/25.17.3389

28. Brister JR, Ako-Adjei D, Bao Y, Blinkova O. NCBI viral Genomes resource. Nucleic Acids Res. 2015;43: D571–D577. doi:10.1093/nar/gku1207

29. Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res. 2016;44: D67–D72. doi:10.1093/nar/gkv1276

30. NCBI. National Center for Biotechnology Information (NCBI) assembled genomes [Internet]. 2016. Available: ftp://ftp.ncbi.nlm.nih.gov/genomes/

31. Bateman A, Martin MJ, O'Donovan C, Magrane M, Apweiler R, Alpi E, et al. UniProt: A hub for protein information. Nucleic Acids Res. 2015;43: D204–D212. doi:10.1093/nar/gku989

32. Riveiro-Barciela M, Mínguez B, Gironés R, Rodriguez-Frías F, Quer J, Buti M. Phylogenetic demonstration of hepatitis E infection transmitted by pork meat ingestion. J Clin Gastroenterol. 2015;49: 165–8. doi:10.1097/MCG.0000000000000113

33. Rivero-Juarez A, Frias M, Martinez-Peinado A, Risalde MA, Rodriguez-Cano D, Camacho A, et al. Familial Hepatitis E Outbreak Linked to Wild Boar Meat Consumption. Zoonoses Public Health. 2017; 1–5. doi:10.1111/zph.12343

34. Legrand-Abravanel F, Mansuy JM, Dubois M, Kamar N, Peron JM, Rostaing L, et al. Hepatitis E virus genotype 3 diversity, France. Emerg Infect Dis. 2009;15: 110–114. doi:10.3201/eid1501.080296

35. Banks M, Bendall R, Grierson S, Heath G, Mitchell J, Dalton H. Human and porcine hepatitis E virus strains, United Kingdom. Emerg Infect Dis. 2004;10: 953–955. doi:10.3201/eid1005.030908

36. Meng XJ, Purcell RH, Halbur PG, Lehman JR, Webb DM, Tsareva TS, et al. A novel virus in swine is closely related to the human hepatitis E virus. Proc Natl Acad Sci U S A. 1997;94: 9860–9865.

37. Touinssi M, Gallian P, Biagini P, Attoui H, Vialettes B, Berland Y, et al. TT virus infection: prevalence of elevated viraemia and arguments for the immune control of viral load. J Clin Virol. 2001;21: 135–41. Available: http://www.ncbi.nlm.nih.gov/pubmed/11378494

38. Li L, Deng X, Linsuwanon P, Bangsberg D, Bwana MB, Hunt P, et al. AIDS alters the commensal plasma virome. J Virol. 2013;87: 10912–5. doi:10.1128/JVI.01839-13

39. Hofer U. Microbiome: anelloviridae go viral. Nat Rev Microbiol. Nature Publishing

Group; 2014;12: 4–5. doi:10.1038/nrmicro3192

40. Biagini P, Bendinelli M, Hino S, Kakkola L, Mankertz A, Niel C, et al. Anelloviridae. In: King A, Adams M, Carstens E, Lefkowitz E, editors. Virus taxonomy, clasification and nomenclature of viruses. Ninth Repo. Davis: ElSevier Academic Press; 2012. pp. 331–341.

41. Hsiao KL, Wang LY, Lin CL, Liu HF. New phylogenetic groups of torque teno virus identified in eastern Taiwan indigenes. PLoS One. 2016;11: 1–10. doi:10.1371/journal.pone.0149901

42. Okamoto H. History of discoveries and pathogenicity of TT viruses. In: de Villiers E-M, zur Hausen H, editors. TT viruses - The still elusive human pathogens. First. Berlin: Springer; 2009. pp. 1–20. doi:10.1007/978-3-540-70972-5_1

43. Spandole S, Cimponeriu D, Berca LM, Mihaescu G. Human anelloviruses: an update of molecular, epidemiological and clinical aspects. Arch Virol. Austria; 2015;160: 893–908. doi:10.1007/s00705-015-2363-9

44. Bostan N. Current and Future Prospects of Torque Teno Virus. J Vaccines Vaccin. 2013; 1–9. doi:10.4172/2157-7560.S1-004

45. Kakkola L, Bondén H, Hedman L, Kivi N, Moisala S, Julin J, et al. Expression of all six human Torque teno virus (TTV) proteins in bacteria and in insect cells, and analysis of their IgG responses. Virology. Elsevier Inc.; 2008;382: 182–189. doi:10.1016/j.virol.2008.09.012

46. Mi Z, Yuan X, Pei G, Wang W, An X, Zhang Z, et al. High-throughput sequencing exclusively identified a novel Torque teno virus genotype in serum of a patient with fatal fever. Virol Sin. 2014;29: 112–118. doi:10.1007/s12250-014-3424-z

47. Luo K, He H, Liu Z, Liu D, Xiao H, Jiang X, et al. Novel variants related to TT virus distributed widely in China. J Med Virol. 2002;67: 118–126. doi:10.1002/jmv.2200

48. Chu C, Zhang L, Dhayalan A, Agagnina BM, Magli AR, Fraher G, et al. Torque Teno Virus 10 Isolated by Genome Amplification Techniques from a Patient with Concomitant Chronic Lymphocytic Leukemia and Polycythemia Vera. Mol Med. 2011;17: 1. doi:10.2119/molmed.2010.00110

49. Zhang Y, Li F, Shan T-L, Deng X, Delwart E, Feng X-P. A novel species of torque teno mini virus (TTMV) in gingival tissue from chronic periodontitis patients. Sci Rep. Nature Publishing Group; 2016;6: 26739. doi:10.1038/srep26739

50. Ninomiya M, Takahashi M, Shimosegawa T, Okamoto H. Analysis of the entire genomes of fifteen torque teno midi virus variants classifiable into a third group of genus Anellovirus. Arch Virol. 2007;152: 1961–1975. doi:10.1007/s00705-007-1046-6

51. Bzhalava D, Hultin E, Muhr LSA, Ekstrom J, Lehtinen M, Villiers EM de, et al. Viremia during pregnancy and risk of childhood leukemia and lymphomas in the offspring: nested case-control study. Int J Cancer. 2016;138: 2212–2220. doi:10.1002/ijc.29666

52. Ohbayashi H, Tanaka Y, Ohoka S, Chinzei R, Kakinuma S, Goto M, et al. TT virus is shown in the liver by in situ hybridization with a PCR-generated probe from the serum TTV-DNA. J Gastroenterol Hepatol. 2001;16: 424–428. doi:10.1046/j.1440-1746.2001.02460.x

53. Griffin JS, Plummer JD, Long SC. Torque teno virus: an improved indicator for viral pathogens in drinking waters. Virol J. 2008;5: 112. doi:10.1186/1743-422X-5-112

54. Okamoto H, Nishizawa T, Takahashi M. Torque teno virus (TTV): molecular virology and clinical implications. In: Mushawar IK, editor. Viral Hepatitis: Molecular Biology, Diagnosis, Epidemiology and Control. Amsterdan: Elsevier; 2004. pp. 241–251.

55. Delwart EL. Viral metagenomics. Rev Med Virol. John Wiley & Sons, Ltd.; 2007;17: 115–131. doi:10.1002/rmv.532

56. Shibata S, Sekizuka T, Kodaira A, Kuroda M, Haga K, Doan YH, et al. Complete Genome Sequence of a Novel GV.2 Sapovirus Strain, NGY-1, Detected from a Suspected Foodborne Gastroenteritis. Genome Announc. 2015;3: 2–3. doi:10.1128/genomeA.01553-14.Copyright

57. Medici MC, Tummolo F, Albonetti V, Abelli LA, Chezzi C, Calderaro A. Molecular detection and epidemiology of astrovirus, bocavirus, and sapovirus in Italian children admitted to hospital with acute gastroenteritis, 2008-2009. J Med Virol. United States; 2012;84: 643–650. doi:10.1002/jmv.23231

58. Sano D, Pérez-Sautu U, Guix S, Pintó RM, Miura T, Okabe S, et al. Quantification and genotyping of human sapoviruses in the Llobregat river catchment, Spain. Appl Environ Microbiol. 2011;77: 1111–1114. doi:10.1128/AEM.01721-10

59. Hansman GS, Sano D, Ueki Y, Imai T, Oka T, Katayama K, et al. Sapovirus in water, Japan. Emerg Infect Dis. 2007;13: 133–135. doi:10.3201/eid1301.061047

60. Oka T, Wang Q, Katayama K, Saif LJ. Comprehensive review of human sapoviruses. ClinMicrobiolRev. 2015;28: 32–53. doi:10.1128/CMR.00011-14

61. Finkbeiner S, Holtz L, Jiang Y, Rajendran P, Franz C, Zhao G, et al. Human stool contains a previously unrecognized diversity of novel astroviruses. Virol J. 2009;6: 161. doi:10.1186/1743-422X-6-161

62. Jiang H, Holtz LR, Bauer I, Franz CJ, Zhao G, Bodhidatta L, et al. Comparison of novel MLB-clade, VA-clade and classic human astroviruses highlights constrained evolution of the classic human astrovirus nonstructural genes. Virology. 2013;436: 8–14. doi:10.1016/j.virol.2012.09.040

63. Xiao J, Li J, Hu G, Chen Z, Wu Y, Chen Y, et al. Isolation and phylogenetic characterization of bat astroviruses in southern China. Arch Virol. 2011;156: 1415–1423. doi:10.1007/s00705-011-1011-2

64. Meyer CT, Bauer IK, Antonio M, Adeyemi M, Saha D, Oundo JO, et al. Prevalence of classic, MLB-clade and VA-clade Astroviruses in Kenya and The Gambia. Virol J.

Virology Journal; 2015;12: 78. doi:10.1186/s12985-015-0299-z

65. Cheetham S, Souza M, Meulia T, Grimes S, Han MG, Saif LJ. Pathogenesis of a genogroup II human norovirus in gnotobiotic pigs. J Virol. 2006;80: 10372–81. doi:10.1128/JVI.00809-06

66. Frange P, Touzot F, Debré M, Héritier S, Leruez-Ville M, Cros G, et al. Prevalence and clinical impact of norovirus fecal shedding in children with inherited immune deficiencies. J Infect Dis. 2012;206: 1269–1274. doi:10.1093/infdis/jis498

67. Fumian TM, Justino MCA, Mascarenhas JDAP, Reymão TKA, Abreu E, Soares L, et al. Quantitative and molecular analysis of noroviruses RNA in blood from children hospitalized for acute gastroenteritis in Belém, Brazil. J Clin Virol. Elsevier B.V.; 2013;58: 31–35. doi:10.1016/j.jcv.2013.06.043

68. Lemes LGN, Corrêa TS, Fiaccadori FS, Cardoso D das Dô de P, Arantes A de M, Souza KMC, et al. Prospective study on Norovirus infection among allogeneic stem cell transplant recipients: Prolonged viral excretion and viral RNA in the blood. J Clin Virol. 2014;61: 329–333. doi:10.1016/j.jcv.2014.08.004

69. Souza M, Azevedo MSP, Jung K, Cheetham S, Saif LJ. Pathogenesis and immune responses in gnotobiotic calves after infection with the genogroup II.4-HS66 strain of human norovirus. J Virol. 2008;82: 1777–1786. doi:10.1128/JVI.01347-07

70. Takanashi S, Hashira S, Matsunaga T, Yoshida A, Shiota T, Tung PG, et al. Detection, genetic characterization, and quantification of norovirus RNA from sera of children with gastroenteritis. J Clin Virol. 2009;44: 161–163. doi:10.1016/j.jcv.2008.11.011

71. Newman KL, Marsh Z, Kirby AE, Moe CL, Leon JS. Immunocompetent Adults from Human Norovirus Challenge Studies do not Exhibit Norovirus Viremia. J Virol. 2015;89: JVI.00392-15. doi:10.1128/JVI.00392-15

72. Chivero ET, Stapleton JT. Tropism of human pegivirus (formerly known as GB virus C/hepatitis G virus) and host immunomodulation: insights into a highly successful viral infection. J Gen Virol. 2015;96: 1521–1532. doi:10.1099/vir.0.000086

73. Leary TP, Mushahwar IK. The GB Viruses. In: Mushawar IK, editor. Viral Hepatitis: Diagnosis, Therapy, and Prevention. Amsterdam: Elsevier; 2004. pp. 223–240.

74. Van der Kuyl AC. HIV infection and HERV expression: a review. Retrovirology. 2012;9: 6. doi:10.1186/1742-4690-9-6

75. Canuti M, van Beveren NJM, Jazaeri Farsani SM, de Vries M, Deijs M, Jebbink MF, et al. Viral metagenomics in drug-naïve, first-onset schizophrenia patients with prominent negative symptoms. Psychiatry Res. Elsevier; 2015;229: 678–684. doi:10.1016/j.psychres.2015.08.025

Figure Captions

Fig. 1. Heatmap describing number of contigs identified in each pool after their characterization and classification into taxonomic families. The colour gradient is proportional to the counts, which are also available in each cell.

Fig. 2. Phylogenetic tree of *Hepeviridae* based on complete genomes, including the main members of genotype 3. Numbers in blank bullets correspond to contigs identified in the HEV and Ai+ImSP pools (see Table 2); they are located beside the reference sequence where specific individual alignments of sequenced fragments over the same region in the reference sequences generated an equivalent tree topology (further results available from Supplementary Materials File A). Labels within the square brackets define the species subtype. Small numbers on the tree branches show the bootstrap score of those branches.

Fig. 3. Phylogenetic tree for the *Anelloviridae* family based on ORF1 region and including only contigs that fully overlap with that region. Numbers and letters within black bullets refer to contigs longer than 1,500 bp (see Table 3) that partially aligned with ORF1 or with the ORF1 upstream region, respectively. See Figure 2 for further details about notation used in this tree.

Tables

Table 1. Summary of the sequences produced for each pool of serum samples in the sequencing experiment. All read counts correspond to total values, and the paired-reads real counts are half the values shown in the table. PE: paired-end reads; SE: single-end reads.

Pool ID	Raw Reads (MiSeq)	Clean Reads (PE + SE)	Contigs (after assembly)
Male A	5,255,854	3,614,220 + 6,928	43,188
Male B	2,669,124	1,769,992 + 4,738	19,000
Female	12,029,238	7,470,502 +18,074	83,518
Ai+ImSP	6,000,606	3,887,728 + 197,320	5,889
HEV	8,145,076	5,769,136 + 2,025	13,060
Healthy A.1	3,413,928	1,873,370 + 286	189,820
Healthy A.2	3,588,692	1,796,830 +298	166,167
Healthy B.1	3,457,150	2,119,224 +250	227,469
Healthy B.2	3,494,586	1,934,704 + 4	185,359

Table 2. Summary of similarity searches for those detected from the HEV and Ai+ImSP pools. The first column corresponds to the numbers in the black bullets shown on some of the branches of the *Hepeviridae* phylogenetic tree from Figure 2.

Code	Pool	Contig ID	Length	Bootstrap	%Identity
1	AI+IMSP	contig_953	992	100.0	91.63
2	AI+IMSP	contig_1893	686	97.4	91.40
3	AI+IMSP	contig_3606	412	84.1	91.02
4	HEV	contig 3810	573	60.0	86.83
5	HEV	contig_2453	1,416	68.4	91.05
6	HEV	contig_533	590	70.0	90.51
7	HEV	contig_1542	575	76.7	89.04
8	HEV	contig 749	541	51.3	88.54
9	HEV	contig_6571	1,572	99.9	91.35
10	HEV	contig_747	1,415	74.6	88.30
11	HEV	contig _3424	1,032	83.3	90.31
12	HEV	contig_6979	944	100.0	91.58
13	HEV	contig_7146	929	95.0	89.26
14	HEV	contig_8370	557	96.0	93.33
15	HEV	contig 10460	314	95.4	86.29
16	HEV	contig 3914	297	65.2	87.21
17	HEV	contig_1444	1,534	98.3	87.11
18	HEV	contig 3007	333	86.7	89.33

673 **Table 3.** Summary information for contigs longer than 1,500 bp that were found in the pooled
674 samples and assigned to the *Anelloviridae* family. The number and letter codes from the first
675 column correspond to those in the blank bullets shown on some of the branches of the
676 phylogenetic tree from Figure 3. Those without codes were placed directly on the tree, as they
677 defined new branches.

Code	Sample	Contig ID	Length (bp)	Sequence Name	Bootstrap	% Identity
	Male A	contig_3809	2,798	TTV P19-3	100.0	92.8
D	Male A	contig_1199	1,512	TTV P13-4	68.0	69.3
10	Male A	contig_7929	1,875	TCHN-D2/TTV 11	100.0	79.2
12	Female	contig_1	1,548	TTV P1-3	100.0	92.4
11	Female	contig_129	1,506	TTV S72	100.0	79.8
14	Female	contig_16376	1,513	TTMDV MDJN47	91.9	68.3
	Female	contig_2757	2,142	TTMV 9	99.0	76.6
	Female	contig_4524	1,977	TTMDV MDJN97	100.0	72.1
1	Female	contig_5911	1,500	TJN2/TTV19	100.0	88.9
A	Female	contig_626	1,503	TTMV 18	100.0	89.7
	Female	contig_6674	2,381	TTV S69	97.0	71.2
	Female	contig_818	2,264	TTV P1-3	99.0	95.1
7	Female	contig_9035	2,243	JA4	100.0	94.0
4	Female	contig_1475	1,899	TCHN-A	100.0	87.5
13	Female	contig_1946	1,644	TTMDV1	100.0	71.6
2	Female	contig_268	1,951	SENV-H	100.0	90.1
	Female	contig_311	2,086	TTMDV MDJN47	89.0	66.6
8	Female	contig_6533	1,530	TTV10	100.0	83.3
	Female	contig_7650	1,730	TTV P9-6	51.0	69.6
17	AI+IMSP	contig_1199	1,586	TTV S80	99.8	66.8
	AI+IMSP	contig_793	2,367	TTV 16	100.0	90.8
15	AI+IMSP	contig_1013	1,687	TTMDV MDJN47	95.8	66.0
9	AI+IMSP	contig_1709	1,832	TTV 10	100.0	85.7
18	AI+IMSP	contig_2151	1,985	TTV S97	79.0	65.6
B	HEV	contig_118	1,781	TTMV Emory1	68.0	68.6
3	HEV	contig_2837	1,845	TTV S45	100.0	86.6
	HEV	contig_125	2,303	TTV S57	100.0	70.1
C	HEV	contig_2	1,923	TTMV LY3	79.0	66.6
5	HEV	contig_236	1,643	TTV TCHN-E	99.9	72.0
6	HEV	contig_2366	1,514	TTV P19-3	100.0	93.8
16	HEV	contig_506	2,046	TTMV 5	96.0	73.7
E	HEV	contig_66	1,904	TTV S66	49.0	71.9

678

Figures

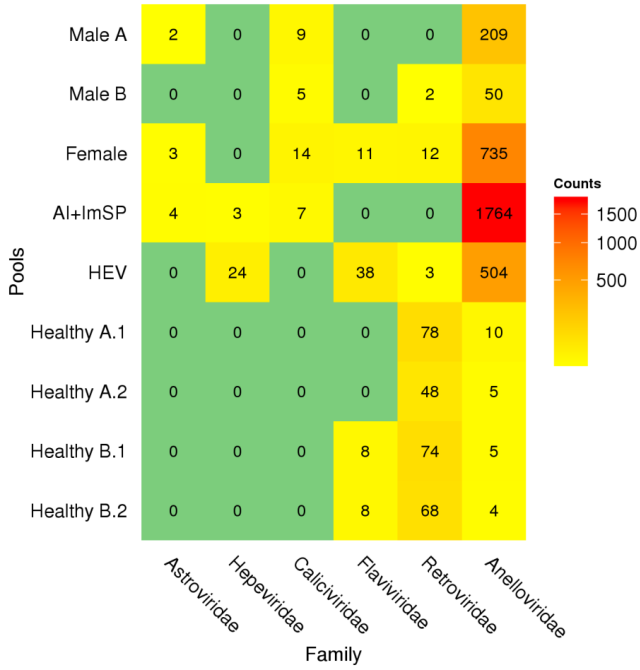


Fig 1

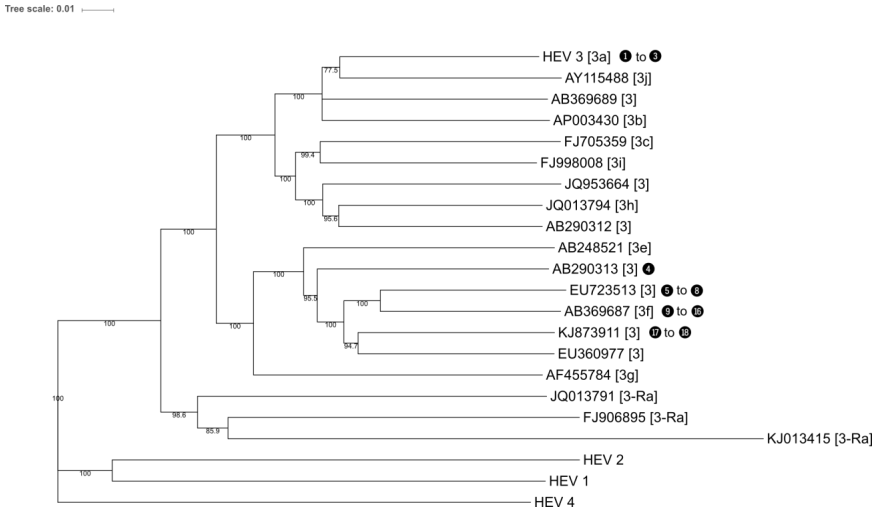
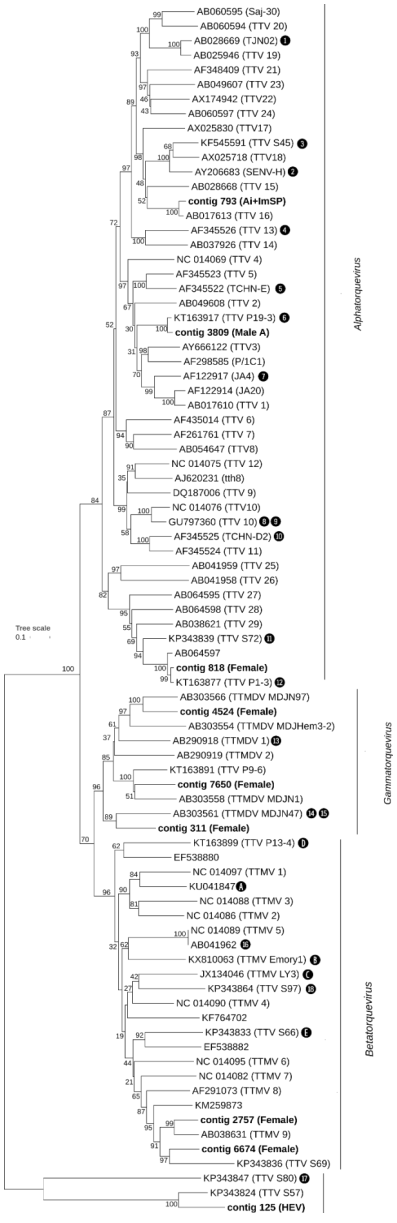


Fig. 2


AB060595 (Saj-30)
AB060594 (TTV 20)
AB028669 (TJN02) 1
AB025946 (TTV 19)
AF348409 (TTV 21)
AB049607 (TTV 23)
AX174942 (TTV22)
AB060597 (TTV 24)
AX025830 (TTV17)
KF545591 (TTV S45) 3
AX025718 (TTV18)
AY206683 (SENV-H) 2
AB028668 (TTV 15)
contig 793 (Ai+ImSP)
AB017613 (TTV 16)
AF345526 (TTV 13) 4
AB037926 (TTV 14)
NC 014069 (TTV 4)
AF345523 (TTV 5)
AF345522 (TCHN-E) 5
AB049608 (TTV 2)
KT163917 (TTV P19-3) 6
contig 3809 (Male A)
AY666122 (TTV3)
AF298585 (P/1C1)
AF122917 (JA4) 7
AF122914 (JA20)
AB017610 (TTV 1)
AF435014 (TTV 6)
AF261761 (TTV 7)
AB054647 (TTV8)
NC 014075 (TTV 12)
AJ620231 (tthB)
DQ187006 (TTV 9)
NC 014076 (TTV10)
GU797360 (TTV 10) 8 g
AF345525 (TCHN-D2) 10
AF345524 (TTV 11)
AB041959 (TTV 25)
AB041958 (TTV 26)
AB064595 (TTV 27)
AB064598 (TTV 28)
AB038621 (TTV 29)
KP343839 (TTV S72) 9
AB064597
contig 818 (Female)
KT163877 (TTV P1-3) 12
AB303566 (TTMDV MDJN97)
contig 4524 (Female)
AB303554 (TTMDV MDJHem3-2)
AB290918 (TTMDV 1) 11
AB290919 (TTMDV 2)
KT163891 (TTV P9-6)
contig 7650 (Female)
AB303558 (TTMDV MDJN1)
AB303561 (TTMDV MDJN47) d k
contig 311 (Female)
KT163899 (TTV P13-4) 0
EF538880
NC 014097 (TTMV 1)
KU041847 A
NC 014088 (TTMV 3)
NC 014086 (TTMV 2)
NC 014089 (TTMV 5)
AB041962 m
KX810063 (TTMV Emory1) s
JX134046 (TTMV LY3) c
KP343864 (TTV S97) 13
NC 014090 (TTMV 4)
KF764702
KP343833 (TTV S66) f
EF538882
NC 014095 (TTMV 6)
NC 014082 (TTMV 7)
AF291073 (TTMV 8)
KM259873
contig 2757 (Female)
AB038631 (TTMV 9)
contig 6674 (Female)
KP343836 (TTV S69)
KP343847 (TTV S80) n
KP343824 (TTV S57)
contig 125 (HEV)
Tree scale
0.1
Alphatorquevirus
Gammatorquevirus
Betatorquevirus

686

687 **Fig. 3**

Article 3

Metagenomics for the study of viruses in urban sewage as a tool in public health surveillance

X.fernandez-Cassi1, **N. Timoneda**, S. Martínez-Puchol, M. Rusiñol, J. Rodríguez-Manzano, N. Figuerola1, R.H. Purcell, S. Bofill-Mas, J.F. Abril, R. Girones.

Paper en procés de revisió.

Metagenomics for the study of viruses in urban sewage as a tool in public health surveillance

Authorship

X.Fernandez-Cassi[1,*,¶], N. Timoneda[1,2,¶], S. Martínez-Puchol[1], M. Rusiñol[1], J. Rodríguez-Manzano, N. Figuerola[1], R.H. Purcell[3], S. Bofill-Mas[1], J.F. Abril[2], R. Girones[1].

[1] Department of Genetics, Microbiology and Statistics, Faculty of Biology, University of Barcelona, Barcelona, Spain.

[2]Computational Genomics Lab, University of Barcelona; and Institute of Biomedicine (IBUB), University of Barcelona; Barcelona, Catalonia, Spain

[3]*National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD 20892, USA.*

[*] Corresponding author

[¶] These authors contributed equally to this work.

Abstract/summary

The development of a public health viral metagenomics surveillance of the viruses circulating in a population based on the study of urban sewage and contaminating viruses in the environment is of high interest although still in its initial steps. The application of next-generation sequencing (NGS) techniques to study viruses present in urban sewage has been limited to very few studies,, which is in part due to the lack of reliable and sensitive protocols to study viral diversity as well the difficulties with processing NGS data. One important step in the methodology needed is an efficient virus concentration protocol for sewage samples. In this study, different protocols for virus concentration in urban sewage were evaluated. The application of a concentration method based on organic flocculation of viruses using skimmed milk (SMF) in 10 L of sewage has allowed the detection of many viruses, producing very valuable information on the virome of urban sewage in different seasons, however some viruses as the human Adenovirus (HAdV) were not always detected using the metagenomics approach even when a qPCR assay was positive. In order to evaluate the diversity of adenoviruses present in sewage, a targeted metagenomics assay using one of the previously tested sewage samples by untargeted metagenomics, was studied using general degenerated primers. . The results of the targeted

metagenomics assay showed the presence of a high diversity of adenoviral strains, most of them taxonomically assigned to murine Adenoviruses (60%), HAdV-41 (29%) and HAdV-9 (3,6%). In order to increase the sensitivity of the metagenomics assay in urban sewage improvements in the concentration protocols were evaluated. Two different protocols for the virus concentration were comparatively analysed: an ultracentrifugation protocol and a lower-volume SMF protocol (500 ml) producing robust results in the virome with both protocols proving that the bioinformatics pipeline was efficient. The sewage virome presented 41 viral families, including pathogenic viral species that were taxonomically assigned to *Caliciviridae, Adenoviridae, Astroviridae*, *Picornaviridae, Polyomaviridae, Papillomaviridae* and *Hepeviridae.* The contribution of urine to the viral composition of sewage was also evaluated by analysing a pools of urine samples by viral metagenomics and it seems to be restricted to few specific DNA viral families, including Polyomavirus and Papillomavirus species. Amplification of viral strains present in sewage by experimental infections using the Rhesus macaque model allowed for the identification of infective human hepatitis E (HEV) and JC Polyomavirus (JCPyV) but no novel viruses were identified in the rhesus serum samples. The protocol for the analysis of the virome in urban sewage developed showed to be reliable and the list of species in the sewage virome has been defined including members of the human virome, classical pathogens and emerging strains. A sensitive protocol for the analysis of viruses in sewage and other environmental samples by metagenomics has been proposed. Urban raw sewage consists of the excreta of thousands of inhabitants; therefore, it is a representative sample for epidemiological surveillance purposes. Therefore, the study of the metavirome present in sewage can provide important information of public health significance, highlighting the presence of viral strains circulating within a population while acting as a complex matrix for viral discovery.

Keywords: viral metagenomics, Human adenovirus, viral pathogens, sewage, next-generation sequencing

1. Introduction

In recent years, water scarcity and the application of more sustainable water reuse practices has favoured the use of treated sewage for several purposes, such as crop and green area irrigation, river catchment replenishment and toilet flushing. Conventional treatments applied in wastewater treatment plants are known to be less efficient for viral removal compared to faecal indicator bacteria (FIB) (Gerba et al., 1979; Pusch et al., 2005). This higher viral survival in waste water treatment plants (WWTP) treatments can represent a threat for consumers because WWTP effluents with viruses can contaminate water or food. Raw urban sewage is a complex matrix consisting of urine, faeces and skin desquamation from people. Therefore, raw sewage contains a large variety of viruses, bacteria and protozoa excreted from thousands of inhabitants. Sewage contains pathogenic and commensal viruses, the latter of which might play a beneficial role in the human gut microbiome; also, a high number of plant viruses pass though the human intestines. Sewage additionally contains other non-human inputs, which increases the diversity of this complex ecosystem. Viruses do not have a conserved gene marker, such as 16s, that is shared across all species, hampering the study of viral metagenomes. However, the application of random-primer based sequencing approaches in combination with next-generation sequencing techniques (NGS) has opened a new path for viral discovery, increasing the viral species described each year. The application of viral metagenomics in sewage constitute an excellent tool to monitor and identify potentially known and unknown viral pathogens circulating among the population, contributing to public health surveillance.

Although some viral metagenomics protocols are available for clinical samples (Kohl et al., 2015), only a few manuscripts describe the application of metagenomics approaches to analyse the viruses present in sewage (Cantalupo et al., 2011; Ng et al., 2012). Previous studies have shown that viruses prevalent in sewage are not always detected by metagenomics, suggesting that protocols should be improved to increase sensitivity. For example, HAdV were hardly detected by Cantalupo and collaborators by NGS, although they had high genome numbers by qPCR.

In the present manuscript, we have studied the diversity of viruses present in raw sewage by testing samples from three different seasons using metagenomics. The application of this methodology allowed for description of the human virome and evaluation of the sensitivity of the technique using HAdVs as a reference virus. Human Adenoviruses (HAdVs) were selected because their double role as pathogens and as specific human viral faecal indicator (Bofill-Mas et al., 2013).

With this purpose, we compared the performance of untargeted metagenomics to an adenovirus-targeted NGS assay and HAdV qPCR values. To increase the number of different viral species identified in sewage, different protocols for virus concentration in urban sewage were evaluated. and an efficient protocol for the analysis of viruses in sewage and other environmental samples by metagenomics has been proposed.

The application of metagenomics in different human body parts has facilitated the study of viral communities in the oral cavity (Ly et al., 2014), gut (Minot et al., 2011), respiratory tract (Willner et al., 2009), skin (Foulongne et al., 2012), blood (Sauvage et al., 2016) and cerebrospinal fluid (Perlejewski et al., 2015). Viral faecal viromes have been studied in healthy (Minot et al., 2011) and unhealthy patients (Linsuwanon et al., 2015) as well as in domestic animals (Mihalov-Kovács et al., 2014); hence, the viral contribution of faeces to raw sewage seems clear. Of note, the viral communities excreted through urine remain poorly studied (Tasha M Santiago-Rodriguez et al., 2015), which is probably because urine was previously considered a sterile environment. To assess the contribution of urine to the virome of raw sewage and to study its viral composition, viruses in pooled urine samples have been also analysed by metagenomics in this study.

The infectivity of known and unknown viral species present in raw sewage has been explored by the intravenous inoculation of a sewage sample to Rhesus monkeys as a potential enrichment step prior the application of the metagenomics approach in the rhesus serum samples.

Finally, a tailored protocol to analyse sewage and other environmental samples using metagenomics has been proposed. Bioinformatics-specific parameters were adjusted at different levels and new tools were tested to filter out the best set of raw reads, such as those containing the most

informative sequences. Those reads were combined into assembled contigs that were later used to detect the known species genomes present in the samples and the relative abundances of the taxonomic groups found in the species mixture. Similarity searches also provided a basic characterization of the pathogenic species present in the samples.

2. Materials and methods

2.1 Concentration of viral particles from tested samples

2.1.1. Concentration of viral particles from raw sewage using skimmed milk organic flocculation (SMF).

Three 10-L samples of raw sewage from a UWWTP in Sant Adrià del Besós were collected in Winter, Spring and Summer 2013. Samples were processed after 2 hours of collection. Viral particles were concentrated by applying the skimmed milk organic flocculation (SMF) method described by Cantalupo et al. (2011). Free DNA from viral concentrates was removed, nucleic acids (NAs) were extracted, and libraries were prepared as explained in section 2.2.

In a second protocol, the reduction of the volume of the sewage sample was also evaluated in order to reduce inhibitory compounds and interfering materials in the viral concentrate.

Briefly, the SMF-adapted protocol used 500 mL of raw sewage that was preconditioned to a pH of 3.5. A volume of 500 µL of a pre-flocculated skim milk solution at pH 3.5 was added to the samples. After 8 h of stirring, flocks were centrifuged at 8000xg for 40 minutes, and the pellet was suspended in 4 mL of phosphate buffer [vol/vol] (0.2 M Na_2HPO_4 and 0.2 M NaH_2PO_4). The viral concentrate was kept at -80°C until further use.

A third protocol based on ultracentrifugation was evaluated in comparison the 500ml SMF protocol. Two samples of 600 mL of raw sewage were collected. Samples were divided into two aliquots: 500 mL for processing according to the SMF protocol adapted from Calgua et al., 2008, and 42 mL for the ultracentrifugation protocol adapted from Pina et al., 1998a. The ultracentrifugation protocol used 42 mL of sewage that was processed as described by Pina et al., 1998a. The obtained SMF and Ultracentrifugation viral concentrates were filtered in 0.45-µm Sterivex filters. Free DNA was removed, NAs were extracted, and libraries were prepared as explained in

section 2.2. For both methodologies, the equivalent of 7 mL of a raw sewage sample was analysed in the final constructed libraries. HAdV qPCR was performed on NA extractions as described in section 2.5.

2.1.2 Concentration of viral particles from urine

To explore the viruses excreted by urine and the contribution from urine to the raw sewage virome, 100 mL of urine from 14 healthy volunteers of different ages and origins (7 males and 7 females from 25-63 years old) although most of them living in Barcelona was collected. Individual urine samples were ultracentrifuged for 1 h at 90,000xg at 4°C. The obtained viral pellets were suspended in 300 µL of PBS1X and kept at -80°C until further use. A pooled sample with 1000 µL of each individual urine viral concentrate was obtained. From the pooled sample, 500 µL was DNAse treated, NAs were extracted, and a library was prepared as explained in section 2.2.

2.2 Free DNA removal, nucleic acid extraction, library preparation and sequencing

In all samples, DNAse treatment was performed with the same conditions. Then, 300 µL of raw sewage viral concentrate was treated with 160 U of Turbo DNase (Ambion Cat nº AM1907, Ambion) to remove non-viral free DNA during 1 h at 37°C. DNAse was inactivated using the provided inactivation reagent and centrifugation at 10,000xg for 1.5 minutes. Treated supernatant was collected and kept at 4°C until nucleic viral acid extraction. Then, 280 µL of viral concentrate was extracted using the Qiagen RNA Viral Mini Kit (cat no. 22906, Qiagen, Valencia, CA, USA) without RNA carrier. NAs were eluted using 60 µL of AVE buffer.

For all samples, libraries were prepared following the same protocol. To detect both RNA and DNA viruses, NAs were retrotranscribed using random nonamer Primer A (5'-GTTTCCCAGTCACGATANNNNNNNNN'-3) as previously described in Wang et al., 2003. Briefly, RNA templates were reverse transcribed using SuperScript III (cat nº 18080093, Life Technologies) and Primer A, which contains a 17-nucleotide specific sequence followed by 9 random nucleotides for random priming. A second cDNA strand was constructed using Sequenase 2.0 (cat nº USBM70775Y200UN,

USB/Affymetrix, Cleveland, OH, USA). To address PCR inhibition, 2 library preparations were constructed using 1:2 dilutions of viral NAs. To obtain sufficient DNA for library preparation, a PCR amplification step using Prime rB (5'-GTTTCCCAGTCACGATANNNNNNNNN'-3) and AmpliTaqGold (cat nº 4311806, Life Technologies, Austin, Texax, USA) was performed. After 10 min at 95°C to activate DNA polymerase, the following PCR program was applied: 30 s at 94°C, 30 s at 40°C, 30 s at 50°C for 25 cycles for the ultracentrifugation and low-volume adapted SMF and 40 cycles for the 10L SMF protocol, and finally 60 s at 72°C. PCR products were cleaned and concentrated in a small volume (15 µL) using Zymo DNA clean and concentrator (D4013, Zymo research, USA). Amplified DNA samples were quantified by Qubit 2.0 (cat nº Q32854, Life Technologies, Oregon, USA), and libraries were constructed using a Nextera XT DNA sample preparation kit (Illumina Inc) according to the manufacturer's instructions. Samples were sequenced on Illumina MiSeq 2x250 bp and 2x300 bp, producing paired end reads.

2.3. Bioinformatic pipeline and quality filtering

The quality of raw and clean read sequences was assessed using FASTX-Toolkit software, version 0.0.14 (Hannon Lab). Read sequences were cleaned using Trimmomatic, version 0.32 (Bolger et al., 2014), taking care of sequencing adaptors and linker contamination. Low quality ends were trimmed considering an average threshold Phred score above Q15 over a running-window of 4 nucleotides. Low complexity sequences, which were mostly biased to repetitive sequences that affect the performance of downstream computational procedures, were then discarded after estimating a linear model based on Trifonov's linguistic complexity (Sarma et al., 1990) and the sequence string compression ratio. Discrimination criteria for the linear model assume low complexity scores below a line with a 45° slope and crossing at 5% below the complexity inflexion point found by the model, which is specific to each sequence set. Finally, duplicated reads were removed in a subsequent step to speed up the downstream assembly. Virome reads were assembled using 90% identify over a minimum of 50% of the read length using CLC Genomics Workbench 4.4 (CLC bio USA, Cambridge, MA), and the resulting contig spectra were used as the primary

input for the index. After that, contigs longer than 100 bp were queried for sequence similarity using BLASTN and BLASTX (Altschul et al., 1997, 1990) against the NCBI viral complete genomes database (Brister et al. 2015), the viral division from GenBank nucleotide database (Benson et al., 2015), and viral protein sequences from Uniprot (UniProt Consortium 2015). Species nomenclature and classification was performed according to the NCBI Taxonomy database (Baltimore, 1971) standards. HSPs considered for taxonomical assessment must have an E-value of 10^{-5} and minimum length of 100 bp. Based on the best BLAST result and 90% coverage cut-off, sequences were classified into their likely taxonomic groups of origin. Contigs were merged by Geneious software assembler (Geneious 9, Kearse et al. 2012), and scaffold sequences were subsequently mapped using Geneious mapper tool. Phylogenetic trees were constructed for selected alignments using Geneious software, and the neighbour-joining method was chosen with 1000 bootstrap replicates. Tables summarizing the number of sequences from the assembly matching each taxonomic unit were built. From those tables, richness ratios were calculated by Catchall software, version 4.0 (Allen et al., 2013); among all the models included in the package, the non-parametric model Chao1 was chosen, which was the model providing the best results on the data-sets. Heatmaps were created using heatmaps from the ggplot2 R graphics library (Kolde, 2015).

2.4 HAdV qPCR as a faecal indicator marker

HAdV were quantified in the urban sewage samples by qPCR as described in previous studies ,(Bofill-Mas et al., 2006).

2.5 Targeted metagenomics for the characterization of adenovirus

To detect and typify all *Mastadenovirus* and other potential adenoviruses present in raw sewage, general primers for AdV hexon were designed. To do so, the hexon region from 149 AdV genomes, recognized by the Adenovirus taxonomy group and retrieved from GenBank, was analysed. The hexon region was selected considering its versatility as a very conserved/variable region (Hernroth et al., 2002), conserved for the design of common primers and variable in the internal sequences useful for typification. Due to the specific requirements of the Roche 454 Junior GS protocol, designed primers were flanked by an adaptor and key sequences to identify samples. Primers and conditions for *Adenovirus* PCR are presented in Supplementary material 3. PCR products were purified using Zymo clean and concentrator (cat nº

D4013, Zymo Research). Purified amplicons were then pyrosequenced in a 454 GS Junior System (Life Science-Roche). Obtained raw reads in SFF were transformed to FASTQ using sff_extract from Roche. Adaptors were removed by cutadapt (Martin, 2011a); the complexity and quality of reads were assessed by PrintSeq and FastQC (Schmieder and Edwards, 2011), which were then trimmed using FASTX-Toolkit software, version 0.0.14 (Hannon Lab). To define non-redundant Operational Taxonomic Units (OTUs), CD-Hit was used and tested at different distance levels from which 0.02 was chosen. A local database was built that contained the hexon region of 153 Adenovirus genomes available from GenBank (2016) and representing different species within the 5 Adenoviridae genus: *Aviadenovirus* (9), *Atadenovirus* (12), *Mastadenovirus* (122), *Siadenovirus* (4) and *Ichtadenovirus*(1). OTUs that matched the 0.02 criteria were blasted against the Adenovirus local database using BLASTN (Altschul et al., 1997, 1990). A phylogenetic tree using Raxml with 1000 bootstrap replicates was computed using Geneious (Geneious 9, Kearse et al. 2012).

2.4. Virus amplification by experimental infection

In collaboration with Dr. Robert H. Purcell (Hepatitis Viruses Section, Laboratory of Infectious Diseases, NIAID, NIH, USA), experimental infections of two Rhesus macaques (*Macaca mulatta*) that were previously immunized for HAV were carried out as a part of a wider study in Bioqual, Rockville, MD in compliance with the guidelines of Bioqual's and NIAID'S Institutional Animal Care and Use Committees. The rhesus macaques were inoculated intravenously with 27 mL of 0.45-µm filtered raw sewage from Barcelona mixed with 3 mL of 10X PBS. Blood from both Rhesus macaques was extracted on a weekly basis over two months to study the potential replication of human viruses present in raw sewage. A blood sample from both animals was extracted a week before inoculation of raw sewage that was used as a negative control.

Nucleic acids and libraries were processed according to section 2.1.2. In total, the following 4 different library preparations were sequenced: a pooled library prior to raw sewage inoculation from the two rhesus monkeys (PW1), two different libraries from each of the animals one week after inoculation (RW1 and RW2), and a pooled library from both rhesus monkeys 4 weeks

after inoculation (RW4). Free DNA from viral concentrates was removed, NAs were extracted, and libraries were prepared as described in section 2.2.

3. Results and discussion

3.1. MI-Seq run outputs in 10L sewage samples from 3 different seasons

Mi-Seq results obtained for sequenced samples are summarized in Supplementary material 1. The virome of urban sewage collected in three different seasons, February, May, September was analysed using 10 L samples of raw sewage, and 37 different viral families were identified. The numbers of different viral species assigned to a given viral family are graphically presented in Figure 1.

Bacteriophage families *Siphoviridae, Myoviridae, Podoviridae* and *Microviridae* show a higher diversity degree in urban sewage. This finding agrees with other studies that bacteriophages are the most abundant organisms on earth (Clokie et al., 2011). The ssDNA Parvoviruses, closely followed by Picornaviruses, are viral families infecting animals/humans with higher diversity. Viral plant *Virgaviridae* species were also abundantly represented in samples. Important human viral pathogens that are taxonomically assigned to *Astroviridae, Caliciviridae*, *Hepeviridae* and *Polyomaviridae* were detected too. Also reads related to viruses belonging to the *Circoviridae* and *Picobirnaviridae* families were sequenced. A summary of the number of reads and contigs associated with those viral families can be found in Table 1. A complete list of detected viral sequences is provided as Supplementary material 2.

A wide diversity and abundance of human and animal astroviruses were detected in the winter sample. The majority of the reads from that sample belonged to the MAstV-1 genogroup, whereas MAstV-6, 8 and 9 were less frequent. Similarly, more sequences taxonomically assigned to the *Caliciviridae* viral family, and specifically to different norovirus GI and GII species and human sapoviruses, were detected in winter. The seasonality of Astroviruses and Caliciviruses during low-temperature seasons has been well-documented (Bosch et al., 2014; Haramoto et al., 2006). Within the *Picornaviridae* family, several human and animal Picornaviruses were sequenced, including the recently described Human Salivirus/Klassevirus, several Aichi viruses, and the recently described genus *Cosavirus*. Aichi virus read counts were higher during summer compared to the other tested

seasons. Human Enteroviruses from species A, B, C and D had similar numbers, irrespective of the analysed season. Important viral pathogens causing hepatitis transmitted through the consumption of water/food contaminated with faecal material, such as HAV and HEV, were only detected in low numbers in the winter sample although this will be related to the low prevalence of these infections in the studied area..

Viral faecal markers present in urban raw sewage, such as human Adenoviruses, were not detected by the metagenomics approach when the 10-L SMF protocol was applied. This contrasts with the detection of human Adenoviruses by conventional qPCR in the three seasons tested, Winter, Spring and Summer, with $3{,}18x10^4$ GC/L, $5{,}32x10^5$ GC/L and $1{,}23x10^5$ GC/L, respectively.

3.1.2. Targeted metagenomics for Adenovirus characterization

To address the lack of detection of HAdV and to study the diversity of the genus *Mastadenoviridae* in raw sewage, a target enrichment assay using broadly degenerated primers for the *hexon* region was conducted. Previously concentrated SMF from summer was used because it contained higher numbers of genome copies of HAdV. A total of 55,903 raw reads were generated by pyrosequencing. All raw reads passed the cleaning cut-offs and were used for subsequent analyses. A sequence similarity of 98% was chosen as a cut-off for the homology searches, which resulted in a total of 3,677 different OTUs, accounting for 52,370 sequences from the sample (93.7%). The obtained OTUs were blasted against the custom-built Adenovirus database, falling into 52 phylogenetically different AdV taxons, HAdV A, B, C, D, F and G-. Detected AdVs from raw sewage are shown in Figure 2, and a complete list detailing the abundance of detected AdV is available in Supplementary material 3. Most of the sequences were assigned to Murine Adenovirus-2 (60%) as well as to HAdV from species F, such as to HAdV-41 (29%) and HAdV-40 (0.7%). HAdV-9, from species type D, was the second most abundant HAdV, accounting for a total 3.6% of the reads. The degenerated primers facilitated the detection of a wide range of AdVs with a high variability of the hosts. However, given that some of the detected sequences were from AdV exotic animals and that they clustered with other well-known HAdV species, the used AdV database might not reflect the true diversity within the *Adenoviridae* family, and other excreted human/non-

human adenoviruses are yet to be discovered. This is exemplified by some of the detected Simian Adenovirus (SAdV), which is closely related to HAdV40 and 41 (see Figure 2). Therefore, detected SAdV could be variants of the closely related HAdV40 and 41. It should also be considered that in the short region analyzed, few changes are important and errors may be introduced during PCR amplification and sequencing process, it has been described that 454 GS Junior has an overall error rate of 0.18% in a study by Niklas et al(2013),and it is known also that the distribution of errors in the sequences is not homogenous.

3.2. Comparative evaluation of ultracentrifugation and SMF of small volumes for virus concentration in sewage

In order to increase sensitivity, two protocols were comparatively evaluated for the concentration of viruses in sewage and the metegenomics analysis. A modified protocol of virus concentration based on SMF with lower volume of sample, 500ml, and a protocol based on ultracentrifugation. Despite the small volume tested when compared with the results obtained concentrating 10-L of urban sewage, the new modified flocculation protocol allowed the detection of members of viral families that were previously not detected, such as *Adenoviridae, Polyomaviridae* and *Papillomaviridae*, identified when using this methodology.

Previous studies in the laboratory did evaluate the effect of different cycle amplifications (25 vs 35 PCR cycles) on the observed viral diversity by estimating the viral richness. Libraries that amplified 35 cycles had a lower average estimated viral richness, affecting the different species of detected dsDNA viruses (data not shown). Both protocols ultracentrifugation and the SMF of 500 ml were used with 25 cycles of amplification before the library construction.

Ultracentrifugation is an efficient technique to concentrate viruses, yielding good recoveries. However, the difficulty to simultaneously concentrate viral particles from several samples and the requirement of an ultracentrifuge device hampers its applicability. A recent comparative study published in collaboration with Hjelmsø et al., (2017) showed that the analysis of 10-L

SMF, as described in section 2.1.1, in combination with QIAgen extraction columns (Iker et al., 2013) had inhibition problems, as evidenced by HAdV qPCR quantifications. This observed inhibition might have affected the subsequent detection of some viral species by NGS. A simplified version of SMF using an initial smaller volume that avoided the ultracentrifugation step was compared against the reference ultracentrifugation protocol developed by Pina et al., (1998) to improve and minimize the observed limitations of the reference protocol. Both protocols assayed the same two collected sewage samples, testing the same volume of 7 raw sewage millilitre equivalents per library.

In this comparative study of two samples tested each one in parallel with both methods, producing very good results with the detection of a wide variety of RNA and DNA pathogens with a light increase in the number of sequences (bacteriophages principally) when using the ultracentrifugation method. The four viral concentrates were analysed for HAdV by qPCR showing $8.14x10^5$ GC/L, $1.23x10^5$ GC/ L, $2.19x10^5$GC/L, and $1.48x10^5$ GC/L for HAdV in SMF1, SMF2, Ultra1, and Ultra2, respectively. MI-seq results are summarized in Supplementary material 4. The estimated viral richness values were 755.8 (779.4-732.2), 541.0 (559.9-522.1), 1,066.4 (1,089.6-1,043.2), and 1,318.5 (1,345.6-1,290.4) for SMF1, SMF2, Ultra1, and Ultra2, respectively. These results demonstrate a higher estimated viral richness when using ultracentrifugation compared to SMF. In total, 41 different viral families were detected considering all samples. A complete list of the detected viral families is highlighted in Figure 3. The modified SMF protocol with reduction of the sample volume, allowed the detection of 36 different viral families compared to the 38 different viral families detected by ultracentrifugation. A higher diversity in viral phage families was observed by this methodology, which impacts the calculated richness by significantly increasing it (see Figure 3). Higher estimated viral richness was also reflected in the detection of few viral human species showing a low number of contigs only detected by the ultracentrifugation protocol, such as the *Anelloviridae*, *Alloherpesviridae, Geminiviridae, Hepeviridae*, *Totiviridae, Geminiviridae,* and *Polyomaviridae* families. Other viral families, such as the *Luteoviridae, Nanoviridae,* and *Baculoviridae* families, were only detected using SMF. For most important viral families, including human pathogenic viruses, such as *Adenoviridae, Caliciviridae, Parvoviridae, Circoviridae, Astroviridae, and*

Picornaviridae, a high diversity of viral species was detected with similar results by both SMF-500ml and ultrafiltration protocols, demonstrating the suitability of these concentration methods for the detection of pathogens such as Caliciviruses, the main virus responsible for gastroenteritis outbreaks (Ahmed et al., 2014). Hence, the availability of an effective concentration method to detect pathogenic viruses is crucial if NGS metagenomic data will be used for surveillance purposes. The efficacy of SMF to concentrate ssRNA+ viral particles agrees with previously published results by Hjelmsø et al.(2017). The low-volume SMF protocol allowed for detection of a previously undetected family, *Adenoviridae.* Human Adenoviruses detected by untargeted metagenomics were taxonomically assigned to Human Adenovirus F species (HAdV40 and HAdV41). Higher sensitivity is observed for a specific group of viruses when using targeted metagenomics. The results obtained in urban sewage using the specific targeted metagenomics assay for adenovirus showed a wide diversity of adenoviruses. Murine Adenovirus 2 was found to be the most abundant *Adenoviridae* representative in this specific sample analysed and HAdV 40 and 41 and low numbers of other adenoviral species were also detected. The specific characteristics of the sample and a possible biased preference for Murine Adenoviruses of the highly degenerated Adenovirus *hexon* primers used in the targeted assay may contribute to explain the high number of sequences assigned to this viral species

Larger volumes of analysed sample (10 L vs 500 mL) could represent a higher chance to detect rare viral families on sewage. However, larger volumes also have a higher proportion of inhibitors (Schrader et al., 2012). Inhibitors might have affected the PCR amplification step, considering that 40 cycles were needed for 10-L SMF to prepare libraries, while only 25 amplification cycles were needed when 500 mL of SMF was used. Viral metagenomics is limited by the low levels of viral DNA/RNA present in the samples, requiring, in most cases, a PCR-based random amplification step after the RT and sequenase reactions to obtain sufficient DNA for library preparation. Interestingly, viral richness was quite similar despite the different PCR amplification cycles that were applied. The PCR amplification step might introduce bias by amplifying the most abundant genomes such that less abundant genomes might not be sequenced or may be underrepresented (Karlsson et al., 2013). This might be the case for HAdV, as

data obtained in previous assays showed that PCR random amplification methods more significantly decreased the estimated viral richness of dsDNA genomes compared to other viral genomes (data not shown). Overall, the data indicate that a concentrated of 500 mL of urban raw sewage is a representative sample volume to study the virome of raw sewage.

One of the main objectives of this research was to shed light on the viral families that are present in raw sewage, which we define as the sewage virome. This list should be periodically reviewed using the developed protocols for environmental surveillance and to identify the introduction of pathogens, novel or emerging viral strains in the population and the environment. A complete list of all different viral species detected in raw sewage using the metagenomics approach in this manuscript is detailed in Supplementary material 2.

In total, more than 11 different viral families considered, or putatively considered, as pathogenic have been detected in raw sewage from Barcelona. *Astroviridae* is a single-stranded, positive-sense RNA viral family of 6.2-7.8 kilobases (kbp) that infects mammals. Human astroviruses (HAstV) are suspected to be involved in 0.5 to 15% of all acute diarrhoea outbreaks in children (Bosch et al., 2014). In the present study, a high diversity of sequences, mainly assigned to the MastV-1 genotype, was detected in all tested samples, but several recombinant genotypes, such as MAstV-6, -8 and -9, were observed in lower abundance. More precisely, the application of NGS techniques has facilitated the detection of these later mentioned animal recombinant astroviruses (Finkbeiner et al., 2009; Kapoor et al., 2009), which are related to neurological disorders in immunocompromised patients (Brown et al., 2015).

Noroviruses (NoV), within the *Caliciviridae* family, are the leading aetiological agent of food-borne disease outbreaks worldwide (Koo et al., 2010). NoV from both genogroups GI and GII were detected in all sewage samples, reflecting a wide diversity within this variable viral family. Those included sequences were taxonomically assigned to NoV GII.4 and NoV GII.17, which are the more frequently reported gastroenteritis genotypes (Chan et al., 2015; Vega et al., 2011). Within the same family, human sapoviruses (HSaV) that belong to GI, GII, GIV, and GV were also found; they have been previously reported as gastroenteritis agents (Oka et al., 2015).

Picornaviridae is a family grouping of more than 30 different genera of ssRNA+ viruses infecting vertebrates, and it includes important human pathogens, such as hepatitis A virus and poliovirus. Several species of the genera *Kobuvirus*, *Enterovirus (EV)*, *Cosavirus, Salivirus,* and *Cardiovirus* were detected in sewage. Aichi virus (AiV) has been recovered in all seasons and in all tested sewage samples, which agrees with available data (Lodder et al., 2013). Recent studies have suggested that AiV may co-infect with other enteric viruses, causing gastroenteritis (Ambert-Balay et al., 2008; Räsänen et al., 2010). EV is one of the most important genera within the *Picornaviridae* family; it contains 12 different species that infect humans, including *EV* species A to D and *Rhinovirus* species A to C (Plyusnin et al., 2011). Different EV from species A, B, and C and animal enteroviruses from species G and J were also noted. Most of the identified human enteroviruses belong to species A and B, but important enteroviruses from species C, such as Enterovirus-A71, were caught. An increase in enterovirus outbreaks has recently been reported to be caused by emerging recombinant EV strains (Holm-Hansen et al., 2016; Zhang et al., 2010). Other sequences related to the *Salivirus* and *Cosavirus* genera, whose causal role in gastroenteritis is suspected, have been detected (Li et al., 2009; Tseng et al., 2007).

Parvoviridae is a large viral family with a wide range of hosts, from mammals to insects, and constitutes an important component of urban sewage. Several sequences resembling animal parvoviruses that infect dogs, rats, cattle and swine as well as several densoviruses with invertebrate hosts have been identified. Human bocavirus (HBoV) species HBoV1, 2, 3 and, 4 and human bufaviruses have been observed, yet the implications of those parvoviruses in human disease is controversial (Nawaz et al., 2012; Phan et al., 2012), and further studies should be conducted to better characterize their pathogenic role or consider them as part of the human gut viral community.

Sequences that are taxonomically assigned to the *Circoviridae* family have been detected in all sewage samples. To date, the Circoviridae family contains two genera, namely, *Circovirus* and *Gyrovirus,* with a third of the proposed genus *Cyclovirus* under revision (Dayaram et al., 2013). Because circoviruses are prevalent in several human fluids, their detection in raw sewage seems reasonable. Their relationship with disease remains unclear, but cycloviruses have been involved in acute nervous system infections (Tan

et al., 2013).

Orthohepevirus, within the *Hepeviridae* family*,* is a genus with the specie Orthoherpesvirus A that includes the viruses causing hepatitis in humans. Genotypes 1 and 2 have been described to infect only humans, while genotypes 3 and 4 are zoonotic (Legrand-Abravanel et al., 2009). The finding in one sample of the HEV genotype 3, frequently detected in swine, illustrates its low prevalence compared to other faecal transmitted viruses causing gastroenteritis (RUTJES et al., 2014).

Surprisingly, no members of the Reoviridae family were detected. Important pathogenic viruses within this family include the human Rotaviruses, which are already known as an important gastroenteritis agent in children and cause approximately 453,000 deaths in 2008 (Tate et al., 2012). Although Rotaviruses are detected with similar concentrations as other enteric viruses in sewage (Prado et al., 2011), their prevalence is lower and influenced by seasonality patterns compared to HAdV (El-Senousy et al., 2015; Zhou et al., 2016). Other metagenomic studies failed to detect rotaviruses although they included sewage samples from endemic rotavirus areas (Cantalupo et al., 2011; Ng et al., 2012). *Picobirnaviridae* viruses from the family*,* which also have dsRNA segmented genomes, have been detected in all tested raw sewage samples. Human Picobirnaviruses are prevalent by conventional PCR in 100% of sewage samples and have been detected at high concentrations (Symonds et al., 2009). Again, a higher relative abundance of this viral family compared to Rotaviruses should be expected.

In the present study, dsDNA viral families, such as *Polyomaviridae*, *Adenoviridae* and *Papillomaviridae*, have been detected. Polyomaviruses and Adenoviruses are excreted by symptomatic and asymptomatic carriers, independent of the seasonality or geographical area. Therefore, they are present in nearly 100% of untreated sewage, which makes them suitable as human viral faecal indicators (Bofill-Mas et al., 2013). Human Papillomaviruses (HPV) have recently been reported in raw sewage (La Rosa et al., 2013). The transmission of papillomaviruses through the consumption of faecal contaminated water or food remains unproven, and further studies on the significance of their molecular detection are needed. Families with insect viruses, such as *Dicistroviridae*, *Iridoviridae,* and *Nodaviridae*, have also been detected, insects could be expected through the sewage system

of a city..

A high abundance and diversity of plant viruses was found in our samples. Viruses from the *Virgaviridae, Closteroviridae, Partitiviridae, Alphaflexiviridae, Betaflexiviridae, Tombusviridae, Bromoviridae, Secoviridae, Potyviridae, and Tymoviridae* families seem to be abundant and important components of the sewage virome. Especially diverse are the members of *Virgaviridae* family, which were the second most diverse detected family, irrespective of the concentration method or volume. Plant viruses are highly abundant in human faeces (Zhang et al., 2006); for example, PMMV has been recently related to specific immune responses, fever, and abdominal pains in humans by Colson et al. (Colson et al., 2010). The infectivity of human excreted plant viruses has already been demonstrated (Tomlinson et al., 1982; Zhang et al., 2006). As a result, their presence in WWTP effluents could represent an economic threat for farmers if reclaimed water without a suitable quality control is used for crop irrigation.

Bacteriophages were the major fraction from the sewage virome with sequences spotted from *Microviridae*, *Podoviridae*, *Myoviridae*, *Leviviridae*, *Siphoviridae* and *Myoviridae* families. *Microviridae* is the family with a higher level of diversity. Detected phage viral families in the present study agree with other untargeted metagenomic analyses (Tamaki et al., 2012). It is likely that the number of bacteriophages sequences has been underestimated due to the taxonomical assignment of prophages as bacterial DNA.

The application of NGS techniques to environmental and clinical samples facilitates the simultaneous analysis of millions of sequences. Of note, a significant fraction of sequences remains unassigned to known taxonomic units after bioinformatics analyses. In the present study, samples were virion-enriched by the applied concentration methods, and the viral concentrate was filtered to remove bacteria, while DNAse was used to remove free DNA. Nevertheless, the percentage of sequences assigned to a known virus taxon was extremely low, but it agreed with previous publications.

The evaluated sewage virome is only an initial attempt to address complex water matrices. The lack of a universal viral marker-compared to bacterial

16S and the need to sequence all available RNA/DNA present in samples requires concentration methods for viral particles while removing other DNA sources to increase the sensitivity of viral metagenomics. It is expected that the development and availability of improved sequencing technologies, such as single-molecule nanopore sequencers, in the forthcoming years will provide a more accurate and detailed composition description of the viral mixtures from different types of samples, including those of the sewage virome.

The annotation of the urban sewage virome by applying NGS methods describes the catalogue of the viral species circulating across a given population, which increasingly achieves an important role in public health surveillance. Viruses are more resistant than bacteria to specific treatments applied in WWTP; therefore, they can be present in reclaimed water produced for crop irrigation, surpassing FIB microbiological quality parameters. A previous study by Rosario et al., (2009) demonstrated that reclaimed water contains 1000-fold more virus-like particles than potable water. Although no pathogenic viruses were detected in that study, pathogenic infectious viruses have been detected in reclaimed water in other studies (Rodriguez-Manzano et al., 2012).

3.3. **The contribution of urine to the viral composition of sewage**

Detected viral sequences from the human urine samples analysed are summarized in Supplementary material 5. The urine viral concentrate contained different DNA viral families infecting humans: *Papillomaviridae,* *Polyomaviridae,* and sequences distantly related to circular ssDNA families *Circoviridae,* and *Anelloviridae*. Those results highlight that urine contributes to the highly diverse viral composition of urban sewage by introducing principally DNA viruses. Human polyomaviruses were the most abundant, specifcally JC polyomavirus (JCPyV) known to be excreted through urine principally, and with a lower number of sequences BK Polyomavirus (BKPyV),, the 0.76% of the total reads were associated with this family. This excretion route for Polyomaviruses has already been documented in the literature (Egli et al., 2009; Shinohara et al., 1993); for this reason, the group has been widely used as a specific indicator of human excreta in water (Harwood et al., 2009). In recent years, new Polyomaviruses have

been described, including up to 13 human Polyomaviruses (Mishra et al., 2014). MCPyV is not excreted through urine (Loyo et al., 2010); instead, it is frequently detected in skin samples in conjunction with human Polyomaviruses 6, 7, and 9 (Foulongne et al., 2012). The lack of detection of the new polyomavirus from urine samples suggests that the excretion patterns of these polyomaviruses might occur through faeces and skin desquamation. Reads of HPV (0.03% of total reads), matching HPV129 and HPV170, which probably come from epithelium desquamation during urination, were identified. The detection of HPVs has been reported in faeces (Di Bonito et al., 2015), raw sewage (La Rosa et al., 2013), and urine (Tasha M. Santiago-Rodriguez et al., 2015). In a prior study, several β-HPV (HPV49, HPV92, and HPV96) and γHPV (HPV121 and HPV178) samples were detected. HPV species detected in this study have not been reported in any of the urine metagenomic studies available to date (Tasha M. Santiago-Rodriguez et al., 2015; Smelov et al., 2016, 2014). Although skin desquamation and excretion through faeces might be the main modes through which human Papillomaviruses land in sewage, the excretion of specific papillomaviruses like the skin-specific γHPV, which might have tropism for the urinary tract, is an interesting finding. Because none of the volunteers participating in this study had been diagnosed with HPV infections or genital warts, HPV could be part of the virome of the urinary tract without causing any known disease. More urine-focused studies, such as those applying specific PCR target enrichment to sequencing, would improve our knowledge of the diversity of HPVs in urine. Sequences distantly related at the protein level to *Circoviridae* and *Anelloviridae* were also observed. ssDNA viruses seem to be ubiquitously present in blood (Vasilyev et al., 2009); therefore, the detection of these specific viral families in urine seems very plausible. With the advent of NGS techniques, there has been a significant increase in viruses classified under these two ssDNA viral families (Kim et al., 2011) and other ssDNA circular viral particles still unclassified (Kim et al., 2012).

3.4. Identified infective hHuman viruses present in raw sewage amplified by experimental infection

One week after inoculation, the first Rhesus monkey presented reads matching JCPyV and Hepatitis E virus, supporting the active replication of these two human viruses identified in urban raw sewage when using animal models. The HEV strain found in the rhesus blood sample was annotated as genotype 3. The inoculation of the environmental HEV strain into Rhesus monkeys is an effective method to replicate the virus (Pina et al., 1998a). Sequences classified within this genotype are frequently reported in the geographical area of the study, in Europe (Clemente-Casares et al., 2009), and this genotype is one of the most commonly detected HEV genotypes in Europe and North America (Clemente-Casares et al., 2003). The second Rhesus monkey did not presented JCPyV or HEV sequences in serum at the studied dates (one week and one month after inoculation). The pooled sample from both Rhesus monkeys after 4 weeks post-inoculation did not have any sequences related to Hepatitis E virus or JCPyV, supporting the model of the acute asymptomatic infection.. Pooled serum samples collected one week before the inoculation showed the presence of several viral plants from the *Virgaviridae* family as well as some phages from the *Microviridae* and *Inoviridae* families. Large fractions of genomic plant DNA have been detected in blood (Spisák et al., 2013), suggesting the possibility that viral DNA/RNA could also be circulating through blood and thus be detected by metagenomics. A total of 1,462 reads (0.08%) in sera sample after inoculation were taxonomically assigned to the *Anelloviridae* family, more specifically human Torque teno virus viruses (TTV) 26 and 27. These two viral species were captured in all Rhesus sera samples, supporting the wide distribution and prevalence of those viruses among mammals (de Villiers and Hausen, 2009). The presence of these viruses in blood has also been reported in humans without any associated disease (Biagini et al., 2013).

Conclusions

Raw sewage harbours a vast number of different viral families that may contaminate the environment since commonly viruses are not completely removed in WWTPs. The methodology developed based in ultracentrifugation and if a ultracentrifuge is not available, using the SMF protocol for 500 ml samples is useful and produce robust results for the description of the virome of urban sewage detecting both DNA and RNA viruses. The information of the virome of urban sewage may constitute an

important data base for known and novel and emerging viral strains excreted in the population in a specific time. Among human viral families, important human pathogens have been detected by NGS, including members of the *Parvoviridae, Caliciviridae, Hepeviridae, Adenoviridae, Polyomaviridae, Papillomaviridae,* and *Astroviridae* families. The implementation and application of a low-volume SMF protocol minimized the inhibition problems detected when sampling larger volumes, while offering a representative volume that yielded comparable results to the tested ultracentrifugation method. However, the sensitivity for analysing specific viral groups and reduce representation biases on relatively less abundant viral species is increased by using targeted metagenomics assays designed to amplify specific viral species.

The amplification of viruses excreted in sewage through experimental infection in Rhesus macaques allowed for detection of infective HEV and JCPyV from urban sewage showing interesting information on the presence of plant virus in the serum of the macaques and small cDNA viruses still unclassified that will merit further studies.

The contribution of urine to sewage seems limited to DNA viral families, mainly to Polyomaviruses JCPyV which appear to be highly excreted and with lower quantities BKPyV,.

The use of NGS techniques for sewage analysis can pinpoint major pathogens that circulate in the population and environment, constituting an interesting tool for epidemiologic studies and public health surveillance.

Acknowledgements

The study reported was partially funded by the Programa RecerCaixa 2012 (ACUP-00300) and AGL2011-30461-C02-01/ALI from the Spanish Ministry of Science and Innovation. This study was partially funded by grants from the Catalan Government to Consolidated Research Group VirBaP (2014SRG914) and JPI Water project METAWATER (4193-00001B), with the collaboration of the Institut de Recerca de l'Aigua (IdRA) and the Spanish Adenovirus Network (AdenoNet, BIO2015-68990-REDT) from the Ministerio de Economía y Competitividad of Spain (www.mineco.gob.es). During the development of this study, Xavier Fernandez-Cassi was a fellow of the Catalan Government "AGAUR" (FI-DGR); Natalia Timoneda and Sandra Martínez-Puchol were fellows of the Spanish Ministry of Science.

Bibliography

Ahmed, S.M., Hall, A.J., Robinson, A.E., Verhoef, L., Premkumar, P., Parashar, U.D., Koopmans, M., Lopman, B.A., 2014. Global prevalence of norovirus in cases of gastroenteritis: a systematic review and meta-analysis. Lancet Infect. Dis. 14, 725–730. doi:10.1016/S1473-3099(14)70767-4

Allen, H.K., Bunge, J., Foster, J.A., Bayles, D.O., Stanton, T.B., 2013. Estimation of viral richness from shotgun metagenomes using a frequency count approach. Microbiome 1, 5. doi:10.1186/2049-2618-1-5

Ambert-Balay, K., Lorrot, M., Bon, F., Giraudon, H., Kaplon, J., Wolfer, M., Lebon, P., Gendrel, D., Pothier, P., 2008. Prevalence and genetic diversity of Aichi virus strains in stool samples from community and hospitalized patients. J. Clin. Microbiol. 46, 1252–1258. doi:10.1128/JCM.02140-07

Biagini, P., Bédarida, S., Touinssi, M., Galicher, V., de Micco, P., 2013. Human gyrovirus in healthy blood donors, France. Emerg. Infect. Dis. 19, 1014–5. doi:10.3201/eid1906.130228

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, 7894–6. doi:10.1128/AEM.00965-06

Bofill-Mas, S., Rusiñol, M., Fernandez-Cassi, X., Carratalà, A., Hundesa, A., Girones, R., 2013. Quantification of human and animal viruses to differentiate the origin of the fecal contamination present in environmental samples. Biomed Res. Int. 2013, 192089. doi:10.1155/2013/192089

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120. doi:10.1093/bioinformatics/btu170

Bosch, A., Pintó, R.M., Guix, S., 2014. Human astroviruses. Clin. Microbiol. Rev. 27, 1048–1074. doi:10.1128/CMR.00013-14

Brown, J.R., Morfopoulou, S., Hubb, J., Emmett, W.A., Ip, W., Shah, D., Brooks, T., Paine, S.M.L., Anderson, G., Virasami, A., Tong, C.Y.W., Clark, D.A., Plagnol, V., Jacques, T.S., Qasim, W., Hubank, M., Breuer, J., 2015.

Astrovirus VA1/HMO-C: an increasingly recognized neurotropic pathogen in immunocompromised patients. Clin. Infect. Dis. 60, 881–8. doi:10.1093/cid/ciu940

Calgua, B., Mengewein, A., Grunert, A., Bofill-Mas, S., Clemente-Casares, P., Hundesa, A., Wyn-Jones, A.P., López-Pila, J.M., 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

Cantalupo, P.G., Calgua, B., Zhao, G., 2011. Raw Sewage Harbors Diverse Viral Populations 2, 1–11. doi:10.1128/mBio.00180-11.Editor

Chan, M.C.W., Lee, N., Hung, T.-N., Kwok, K., Cheung, K., Tin, E.K.Y., Lai, R.W.M., Nelson, E.A.S., Leung, T.F., Chan, P.K.S., 2015. Rapid emergence and predominance of a broadly recognizing and fast-evolving norovirus GII.17 variant in late 2014. Nat. Commun. 6, 10061. doi:10.1038/ncomms10061

Clemente-Casares, P., Pina, S., Buti, M., Jardi, R., MartIn, M., Bofill-Mas, S., Girones, R., 2003. Hepatitis E virus epidemiology in industrialized countries. Emerg. Infect. Dis. 9, 448–54. doi:10.3201/eid0904.020351

Clemente-Casares, P., Rodriguez-Manzano, J., Girones, R., 2009. Hepatitis E virus genotype 3 and sporadically also genotype 1 circulate in the population of Catalonia, Spain. J. Water Health 7, 664. doi:10.2166/wh.2009.120

Clokie, M.R., Millard, A.D., Letarov, A. V, Heaphy, S., 2011. Phages in nature. Bacteriophage 1, 31–45. doi:10.4161/bact.1.1.14942

Colson, P., Richet, H., Desnues, C., Balique, F., Moal, V., Grob, J.-J., Berbis, P., Lecoq, H., Harlé, J.-R., Berland, Y., Raoult, D., 2010. Pepper Mild Mottle Virus, a Plant Virus Associated with Specific Immune Responses, Fever, Abdominal Pains, and Pruritus in Humans. PLoS One 5, e10041. doi:10.1371/journal.pone.0010041

Dayaram, A., Potter, K.A., Moline, A.B., Rosenstein, D.D., Marinov, M., Thomas, J.E., Breitbart, M., Rosario, K., Argüello-Astorga, G.R., Varsani, A., 2013. High global diversity of cycloviruses amongst dragonflies. J. Gen. Virol. 94, 1827–1840. doi:10.1099/vir.0.052654-0

de Villiers, E.-M., Hausen, H. zur (Eds.), 2009. TT Viruses, Current Topics in Microbiology and Immunology. Springer Berlin Heidelberg, Berlin, Heidelberg. doi:10.1007/978-3-540-70972-5

Di Bonito, P., Della Libera, S., Petricca, S., Iaconelli, M., Sanguinetti, M., Graffeo, R., Accardi, L., La Rosa, G., 2015. A large spectrum of alpha and beta papillomaviruses are detected in human stool samples. J. Gen. Virol. 96, 607–613. doi:10.1099/vir.0.071787-0

Egli, A., Infanti, L., Dumoulin, A., Buser, A., Samaridis, J., Stebler, C., Gosert, R., Hirsch, H.H., 2009. Prevalence of polyomavirus BK and JC infection and replication in 400 healthy blood donors. J. Infect. Dis. 199, 837–46.

El-Senousy, W.M., Ragab, A.M.E.-S., Handak, E.M.A.E.H., 2015. Prevalence of Rotaviruses Groups A and C in Egyptian Children and Aquatic Environment. Food Environ. Virol. 7, 132–141. doi:10.1007/s12560-015-9184-6

Finkbeiner, S.R., Li, Y., Ruone, S., Conrardy, C., Gregoricus, N., Toney, D., Virgin, H.W., Anderson, L.J., Vinjé, J., Wang, D., Tong, S., 2009. Identification of a novel astrovirus (astrovirus VA1) associated with an outbreak of acute gastroenteritis. J. Virol. 83, 10836–9. doi:10.1128/JVI.00998-09

Foulongne, V., Sauvage, V., Hebert, C., Dereure, O., Cheval, J., Gouilh, M.A., Pariente, K., Segondy, M., Burguière, A., Manuguerra, J.-C., Caro, V., Eloit, M., 2012. Human Skin Microbiota: High Diversity of DNA Viruses Identified on the Human Skin by High Throughput Sequencing. PLoS One 7, e38499. doi:10.1371/journal.pone.0038499

Gerba, C.P., Goyal, S.M., LaBelle, R.L., Cech, I., Bodgan, G.F., 1979. Failure of indicator bacteria to reflect the occurrence of enteroviruses in marine waters. Am. J. Public Health 69, 1116–9.

Haramoto, E., Katayama, H., Oguma, K., Yamashita, H., Tajima, A., Nakajima, H., Ohgaki, S., 2006. Seasonal profiles of human noroviruses and indicator bacteria in a wastewater treatment plant in Tokyo, Japan. Water Sci. Technol. 54, 301–8.

Harwood, V.J., Brownell, M., Wang, S., Lepo, J., Ellender, R.D., Ajidahun, A., Hellein, K.N., Kennedy, E., Ye, X., Flood, C., 2009. Validation and field

testing of library-independent microbial source tracking methods in the Gulf of Mexico. Water Res. 43, 4812–4819. doi:10.1016/j.watres.2009.06.029

Hernroth, B.E., Conden-Hansson, A.C., Rehnstam-Holm, A.S., Girones, R., Allard, A.K., 2002. Environmental factors influencing human viral pathogens and their potential indicator organisms in the blue mussel, Mytilus edulis: The first Scandinavian report. Appl. Environ. Microbiol. 68, 4523–4533. doi:10.1128/AEM.68.9.4523-4533.2002

Hjelmsø, M.H., Hellmér, M., Fernandez-Cassi, X., Timoneda, N., Lukjancenko, O., Seidel, M., Elsässer, D., Aarestrup, F.M., Löfström, C., Bofill-Mas, S., Abril, J.F., Girones, R., Schultz, A.C., 2017. Evaluation of Methods for the Concentration and Extraction of Viruses from Sewage in the Context of Metagenomic Sequencing. PLoS One 12, e0170199. doi:10.1371/journal.pone.0170199

Holm-Hansen, C.C., Midgley, S.E., Fischer, T.K., 2016. Global emergence of enterovirus D68: A systematic review. Lancet Infect. Dis. doi:10.1016/S1473-3099(15)00543-5

Iker, B.C., Bright, K.R., Pepper, I.L., Gerba, C.P., Kitajima, M., 2013. Evaluation of commercial kits for the extraction and purification of viral nucleic acids from environmental and fecal samples. J. Virol. Methods 191, 24–30. doi:10.1016/j.jviromet.2013.03.011

Kapoor, A., Li, L., Victoria, J., Oderinde, B., Mason, C., Pandey, P., Zaidi, S.Z., Delwart, E., 2009. Multiple novel astrovirus species in human stool. J. Gen. Virol. 90, 2965–2972. doi:10.1099/vir.0.014449-0

Karlsson, O.E., Belák, S., Granberg, F., 2013. The effect of preprocessing by sequence-independent, single-primer amplification (SISPA) on metagenomic detection of viruses. Biosecur. Bioterror. 11 Suppl 1, S227-34. doi:10.1089/bsp.2013.0008

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., Drummond, A., 2012. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28, 1647–1649.

doi:10.1093/bioinformatics/bts199

Kim, H.K., Park, S.J., Nguyen, V.G., Song, D.S., Moon, H.J., Kang, B.K., Park, B.K., 2012. Identification of a novel single-stranded, circular DNA virus from bovine stool. J. Gen. Virol. 93, 635–639. doi:10.1099/vir.0.037838-0

Kim, M.-S., Park, E.-J., Roh, S.W., Bae, J.-W., 2011. Diversity and abundance of single-stranded DNA viruses in human feces. Appl. Environ. Microbiol. 77, 8062–70. doi:10.1128/AEM.06331-11

Kohl, C., Brinkmann, A., Dabrowski, P.W., Radonić, A., Nitsche, A., Kurth, A., 2015. Protocol for metagenomic virus detection in clinical specimens. Emerg. Infect. Dis. 21, 48–57. doi:10.3201/eid2101.140766

Kolde, R., 2015. pheatmap: Pretty Heatmaps.

Koo, H.L., Ajami, N., Atmar, R.L., DuPont, H.L., 2010. Noroviruses: The leading cause of gastroenteritis worldwide. Discov. Med. 10, 61–70. doi:10.1007/s12560-010-9038-1.Noroviruses

La Rosa, G., Fratini, M., Accardi, L., D'Oro, G., Della Libera, S., Muscillo, M., Di Bonito, P., 2013. Mucosal and Cutaneous Human Papillomaviruses Detected in Raw Sewages. PLoS One 8, e52391. doi:10.1371/journal.pone.0052391

Legrand-Abravanel, F., Mansuy, J.-M., Dubois, M., Kamar, N., Peron, J.-M., Rostaing, L., Izopet, J., 2009. Hepatitis E virus genotype 3 diversity, France. Emerg. Infect. Dis. 15, 110–4. doi:10.3201/eid1501.080296

Li, L., Victoria, J., Kapoor, A., Blinkova, O., Wang, C., Babrzadeh, F., Mason, C.J., Pandey, P., Triki, H., Bahri, O., Oderinde, B.S., Baba, M.M., Bukbuk, D.N., Besser, J.M., Bartkus, J.M., Delwart, E.L., 2009. A novel picornavirus associated with gastroenteritis. J. Virol. 83, 12002–6. doi:10.1128/JVI.01241-09

Linsuwanon, P., Poovorawan, Y., Li, L., Deng, X., Vongpunsawad, S., Delwart, E., 2015. The fecal virome of children with hand, foot, and mouth disease that tested PCR negative for pathogenic enteroviruses. PLoS One 10. doi:10.1371/journal.pone.0135573

Lodder, W.J., Rutjes, S.A., Takumi, K., de Roda Husman, A.M., 2013. Aichi virus in sewage and surface water, the Netherlands. Emerg. Infect. Dis.

19, 1222–1230. doi:10.3201/eid1908.130312

Loyo, M., Guerrero-Preston, R., Brait, M., Hoque, M.O., Chuang, A., Kim, M.S., Sharma, R., Liégeois, N.J., Koch, W.M., Califano, J.A., Westra, W.H., Sidransky, D., 2010. Quantitative detection of Merkel cell virus in human tissues and possible mode of transmission. Int. J. Cancer 126, NA-NA. doi:10.1002/ijc.24737

Ly, M., Abeles, S.R., Boehm, T.K., Robles-Sikisaka, R., Naidu, M., Santiago-Rodriguez, T., Pride, D.T., 2014. Altered oral viral ecology in association with periodontal disease. MBio 5, e01133-14. doi:10.1128/mBio.01133-14

Mihalov-Kovács, E., Fehér, E., Martella, V., Bányai, K., Farkas, S.L., 2014. The fecal virome of domesticated animals. Virusdisease 25, 150–7. doi:10.1007/s13337-014-0192-1

Minot, S., Sinha, R., Chen, J., Li, H., Keilbaugh, S.A., Wu, G.D., Lewis, J.D., Bushman, F.D., 2011. The human gut virome: inter-individual variation and dynamic response to diet. Genome Res. 21, 1616–25. doi:10.1101/gr.122705.111

Mishra, N., Pereira, M., Rhodes, R.H., An, P., Pipas, J.M., Jain, K., Kapoor, A., Briese, T., Faust, P.L., Lipkin, W.I., 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

Nawaz, S., Allen, D.J., Aladin, F., Gallimore, C., Iturriza-Gómara, M., 2012. Human bocaviruses are not significantly associated with gastroenteritis: results of retesting archive DNA from a case control study in the UK. PLoS One 7, e41346. doi:10.1371/journal.pone.0041346

Ng, T.F.F., Marine, R., Wang, C., Simmonds, P., Kapusinszky, B., Bodhidatta, L., Oderinde, B.S., Wommack, K.E., Delwart, E., 2012. High Variety of Known and New RNA and DNA Viruses of Diverse Origins in Untreated Sewage. J. Virol. 86, 12161–12175. doi:10.1128/JVI.00869-12

Niklas, N., Pröll, J., Danzer, M., Stabentheiner, S., Hofer, K., Gabriel, C., 2013. Routine performance and errors of 454 HLA exon sequencing in diagnostics. BMC Bioinformatics 14, 176. doi:10.1186/1471-2105-14-

176

Oka, T., Wang, Q., Katayama, K., Saif, L.J., 2015. Comprehensive review of human sapoviruses. Clin. Microbiol. Rev. 28, 32–53. doi:10.1128/CMR.00011-14

Perlejewski, K., Popiel, M., Laskus, T., Nakamura, S., Motooka, D., Stokowy, T., Lipowski, D., Pollak, A., Lechowicz, U., Caraballo Cortés, K., Stępień, A., Radkowski, M., Bukowska-Ośko, I., 2015. Next-generation sequencing (NGS) in the identification of encephalitis-causing viruses: Unexpected detection of human herpesvirus 1 while searching for RNA pathogens. J. Virol. Methods 226, 1–6. doi:10.1016/j.jviromet.2015.09.010

Phan, T.G., Vo, N.P., Bonkoungou, I.J.O., Kapoor, A., Barro, N., O'Ryan, M., Kapusinszky, B., Wang, C., Delwart, E., 2012. Acute Diarrhea in West African Children: Diverse Enteric Viruses and a Novel Parvovirus Genus. J. Virol. 86, 11024–11030. doi:10.1128/JVI.01427-12

Pina, S., Jofre, J., Emerson, S.U., Purcell, R.H., Girones, R., 1998a. Characterization of a strain of infectious hepatitis E virus isolated from sewage in an area where hepatitis E is not endemic. Appl. Environ. Microbiol. 64, 4485–8.

Pina, S., Puig, M., Lucena, F., Jofre, J., Girones, R., 1998b. Viral pollution in the environment and in shellfish: human adenovirus detection by PCR as an index of human viruses. Appl. Environ. Microbiol. 64, 3376–82.

Plyusnin, A., Beaty, B.J., Elliott, R.M., Goldbach, R., Kormelink, R., Lundkvist, A., Schmaljohn, C.S., Tesh, R.B., 2011. Virus Taxonomy: Ninth Report of the International Committee on Taxonomy of Viruses, vol. 9, in: Edited by King, AMQ, Adams, J. & Lefkowitz, E.. Amsterdam: Elsevier. pp. 735–741.

Prado, T., Silva, D.M., Guilayn, W.C., Rose, T.L., Gaspar, A.M.C., Miagostovich, M.P., 2011. Quantification and molecular characterization of enteric viruses detected in effluents from two hospital wastewater treatment plants. Water Res. 45, 1287–1297. doi:10.1016/j.watres.2010.10.012

Pusch, D., Oh, D.-Y., Wolf, S., Dumke, R., Schröter-Bobsin, U., Höhne, M., Röske, I., Schreier, E., 2005. Detection of enteric viruses and bacterial indicators in German environmental waters. Arch. Virol. 150, 929–47.

doi:10.1007/s00705-004-0467-8

Räsänen, S., Lappalainen, S., Kaikkonen, S., Hämäläinen, M., Salminen, M., Vesikari, T., 2010. Mixed viral infections causing acute gastroenteritis in children in a waterborne outbreak. Epidemiol. Infect. 138, 1227–1234. doi:10.1017/S0950268809991671

Rodriguez-Manzano, J., Alonso, J.L., Ferrús, M.A., Moreno, Y., Amorós, I., Calgua, B., Hundesa, A., Guerrero-Latorre, L., Carratala, A., Rusiñol, M., Girones, R., 2012. Standard and new faecal indicators and pathogens in sewage treatment plants, microbiological parameters for improving the control of reclaimed water. Water Sci. Technol. 66, 2517–23. doi:10.2166/wst.2012.233

Rosario, K., Nilsson, C., Lim, Y.W., Ruan, Y., Breitbart, M., 2009. Metagenomic analysis of viruses in reclaimed water. Environ. Microbiol. 11, 2806–2820. doi:10.1111/j.1462-2920.2009.01964.x

RUTJES, S.A., BOUWKNEGT, M., van der GIESSEN, J.W., de RODA HUSMAN, A.M., REUSKEN, C.B.E.M., 2014. Seroprevalence of Hepatitis E Virus in Pigs from Different Farming Systems in The Netherlands. J. Food Prot. 77, 640–642. doi:10.4315/0362-028X.JFP-13-302

Santiago-Rodriguez, T.M., Ly, M., Bonilla, N., Pride, D.T., 2015. The human urine virome in association with urinary tract infections. Front. Microbiol. 6, 14. doi:10.3389/fmicb.2015.00014

Santiago-Rodriguez, T.M., Ly, M., Bonilla, N., Pride, D.T., 2015. The human urine virome in association with urinary tract infections. Front. Microbiol. 6, 14. doi:10.3389/fmicb.2015.00014

Sarma, R.H., Saram, M.H. (Mukti H.., State University of New York at Albany., 1990. Structure & methods : proceedings of the Sixth Conversation in the Discipline Biomolecular Stereodynamics held at the State University of New York at Albany, June 6-10, 1989. Adenine Press.

Sato, M., Kuroda, M., Kasai, M., Matsui, H., Fukuyama, T., Katano, H., Tanaka-Taya, K., 2016. Acute encephalopathy in an immunocompromised boy with astrovirus-MLB1 infection detected by next generation sequencing. J. Clin. Virol. 78, 66–70. doi:10.1016/j.jcv.2016.03.010

Sauvage, V., Laperche, S., Cheval, J., Muth, E., Dubois, M., Boizeau, L.,

Hébert, C., Lionnet, F., Lefrère, J.-J., Eloit, M., 2016. Viral metagenomics applied to blood donors and recipients at high risk for blood-borne infections. Blood Transfus. 14, 400–7. doi:10.2450/2016.0160-15

Schrader, C., Schielke, A., Ellerbroek, L., Johne, R., 2012. PCR inhibitors - occurrence, properties and removal. J. Appl. Microbiol. 113, 1014–1026. doi:10.1111/j.1365-2672.2012.05384.x

Shinohara, T., Matsuda, M., Cheng, S.H., Marshall, J., Fujita, M., Nagashima, K., 1993. BK virus infection of the human urinary tract. J. Med. Virol. 41, 301–5.

Smelov, V., Arroyo Mühr, L.S., Bzhalava, D., Brown, L.J., Komyakov, B., Dillner, J., 2014. Metagenomic sequencing of expressed prostate secretions. J. Med. Virol. 86, 2042–2048. doi:10.1002/jmv.23900

Smelov, V., Bzhalava, D., Arroyo Muhr, L.S., Eklund, C., Komyakov, B., Gorelov, a, Dillner, J., Hultin, E., 2016. Detection of DNA viruses in prostate cancer. Sci Rep 6, 25235. doi:10.1038/srep25235

Spisák, S., Solymosi, N., Ittzés, P., Bodor, A., Kondor, D., Vattay, G., Barták, B.K., Sipos, F., Galamb, O., Tulassay, Z., Szállási, Z., Rasmussen, S., Sicheritz-Ponten, T., Brunak, S., Molnár, B., Csabai, I., 2013. Complete Genes May Pass from Food to Human Blood. PLoS One 8, e69805. doi:10.1371/journal.pone.0069805

Symonds, E.M., Griffin, D.W., Breitbart, M., 2009. Eukaryotic viruses in wastewater samples from the United States. Appl. Environ. Microbiol. 75, 1402–1409. doi:10.1128/AEM.01899-08

Tamaki, H., Zhang, R., Angly, F.E., Nakamura, S., Hong, P.-Y., Yasunaga, T., Kamagata, Y., Liu, W.-T., 2012. Metagenomic analysis of DNA viruses in a wastewater treatment plant in tropical climate. Environ. Microbiol. 14, 441–452. doi:10.1111/j.1462-2920.2011.02630.x

Tan, L. Van, van Doorn, H.R., Nghia, H.D.T., Chau, T.T.H., Tu, L.T.P., de Vries, M., Canuti, M., Deijs, M., Jebbink, M.F., Baker, S., Bryant, J.E., Tham, N.T., BKrong, N.T.T.C., Boni, M.F., Loi, T.Q., Phuong, L.T., Verhoeven, J.T.P., Crusat, M., Jeeninga, R.E., Schultsz, C., Chau, N.V.V., Hien, T.T., van der Hoek, L., Farrar, J., de Jong, M.D., 2013. Identification of a new cyclovirus in cerebrospinal fluid of patients with acute central nervous

system infections. MBio 4. doi:10.1128/mBio.00231-13

Tate, J.E., Burton, A.H., Boschi-Pinto, C., Steele, A.D., Duque, J., Parashar, U.D., WHO-coordinated Global Rotavirus Surveillance Network, 2012. 2008 estimate of worldwide rotavirus-associated mortality in children younger than 5 years before the introduction of universal rotavirus vaccination programmes: a systematic review and meta-analysis. Lancet. Infect. Dis. 12, 136–41. doi:10.1016/S1473-3099(11)70253-5

Tomlinson, J.A., Faithfull, E., Flewett, T.H., Beards, G., 1982. Isolation of infective tomato bushy stunt virus after passage through the human alimentary tract. Nature 300, 637–8.

Tseng, C.-H., Tsai, H.-J., Kirkwood, C.D., Wang, D., Hay, C., Hallenbeck, P., Knowles, N., Lemon, S., Minor, P., Pallansch, M., Palmenberg, A., Skern, T., 2007. Sequence analysis of a duck picornavirus isolate indicates that it together with porcine enterovirus type 8 and simian picornavirus type 2 should be assigned to a new picornavirus genus. Virus Res. 129, 104–114. doi:10.1016/j.virusres.2007.06.023

Vasilyev, E. V, Trofimov, D.Y., Tonevitsky, A.G., Ilinsky, V. V, Korostin, D.O., Rebrikov, D. V, 2009. Torque Teno Virus (TTV) distribution in healthy Russian population. Virol. J. 6, 134. doi:10.1186/1743-422X-6-134

Vega, E., Barclay, L., Gregoricus, N., Williams, K., Lee, D., Vinjé, J., 2011. Novel surveillance network for norovirus gastroenteritis outbreaks, United States. Emerg. Infect. Dis. 17, 1389–1395. doi:10.3201/eid1708.101837

Wang, D., Urisman, A., Liu, Y.T., Springer, M., Ksiazek, T.G., Erdman, D.D., Mardis, E.R., Hickenbotham, M., Magrini, V., Eldred, J., Latreille, J.P., Wilson, R.K., Ganem, D., DeRisi, J.L., 2003. Viral discovery and sequence recovery using DNA microarrays. PLoS Biol. 1. doi:10.1371/journal.pbio.0000002

Willner, D., Furlan, M., Haynes, M., Schmieder, R., Angly, F.E., Silva, J., Tammadoni, S., Nosrat, B., Conrad, D., Rohwer, F., 2009. Metagenomic Analysis of Respiratory Tract DNA Viral Communities in Cystic Fibrosis and Non-Cystic Fibrosis Individuals. PLoS One 4, e7370. doi:10.1371/journal.pone.0007370

Zhang, T., Breitbart, M., Lee, W.H., Run, J.-Q., Wei, C.L., Soh, S.W.L., Hibberd, M.L., Liu, E.T., Rohwer, F., Ruan, Y., 2006. RNA viral community in human feces: prevalence of plant pathogenic viruses. PLoS Biol. 4, e3. doi:10.1371/journal.pbio.0040003

Zhang, Y., Zhu, Z., Yang, W., Ren, J., Tan, X., Wang, Y., Mao, N., Xu, S., Zhu, S., Cui, A., Zhang, Y., Yan, D., Li, Q., Dong, X., Zhang, J., Zhao, Y., Wan, J., Feng, Z., Sun, J., Wang, S., Li, D., Xu, W., 2010. An emerging recombinant human enterovirus 71 responsible for the 2008 outbreak of hand foot and mouth disease in Fuyang city of China. Virol. J. 7, 94. doi:10.1186/1743-422X-7-94

Zhou, N., Lv, D., Wang, S., Lin, X., Bi, Z., Wang, H., Wang, P., Zhang, H., Tao, Z., Hou, P., Song, Y., Xu, A., 2016. Continuous detection and genetic diversity of human rotavirus A in sewage in eastern China, 2013–2014. Virol. J. 13, 153. doi:10.1186/s12985-016-0609-0

Nicholas, K. B. and H. B. Nicholas (2006). GeneDoc: Multiple sequence alignment editor & shading utility.

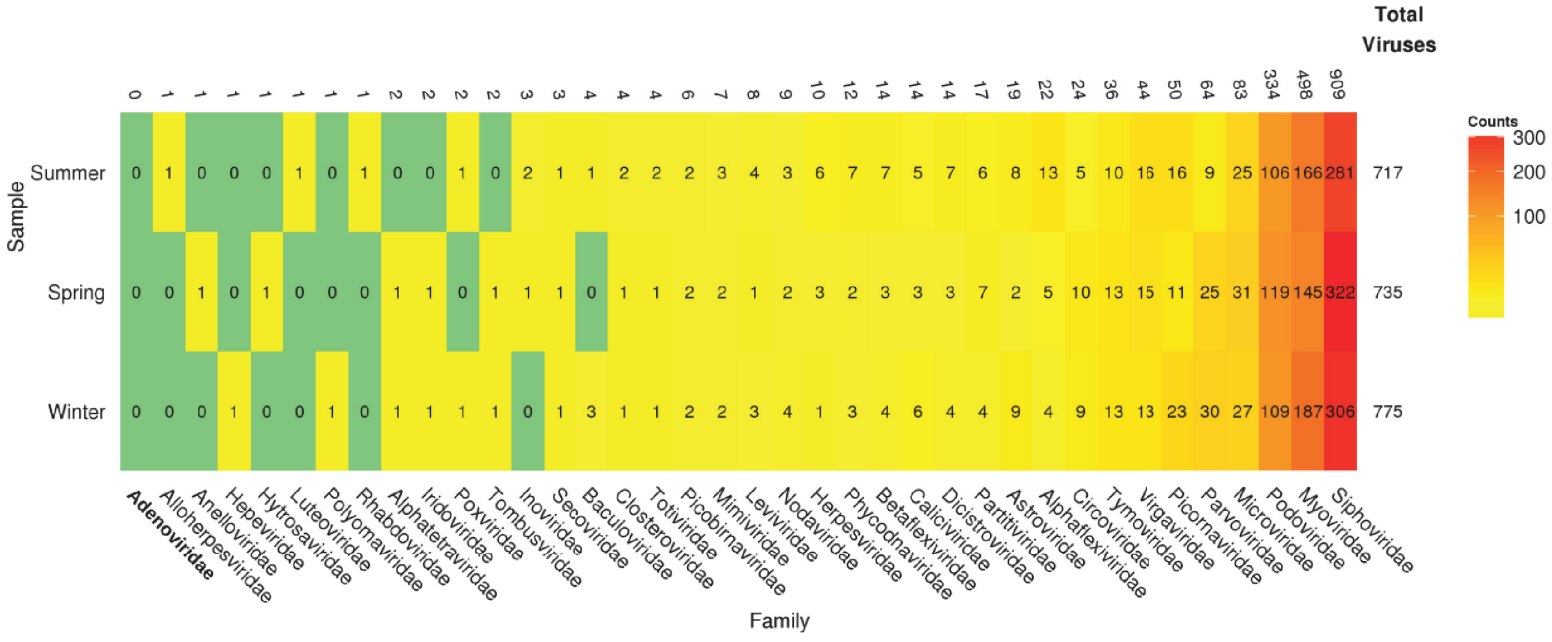
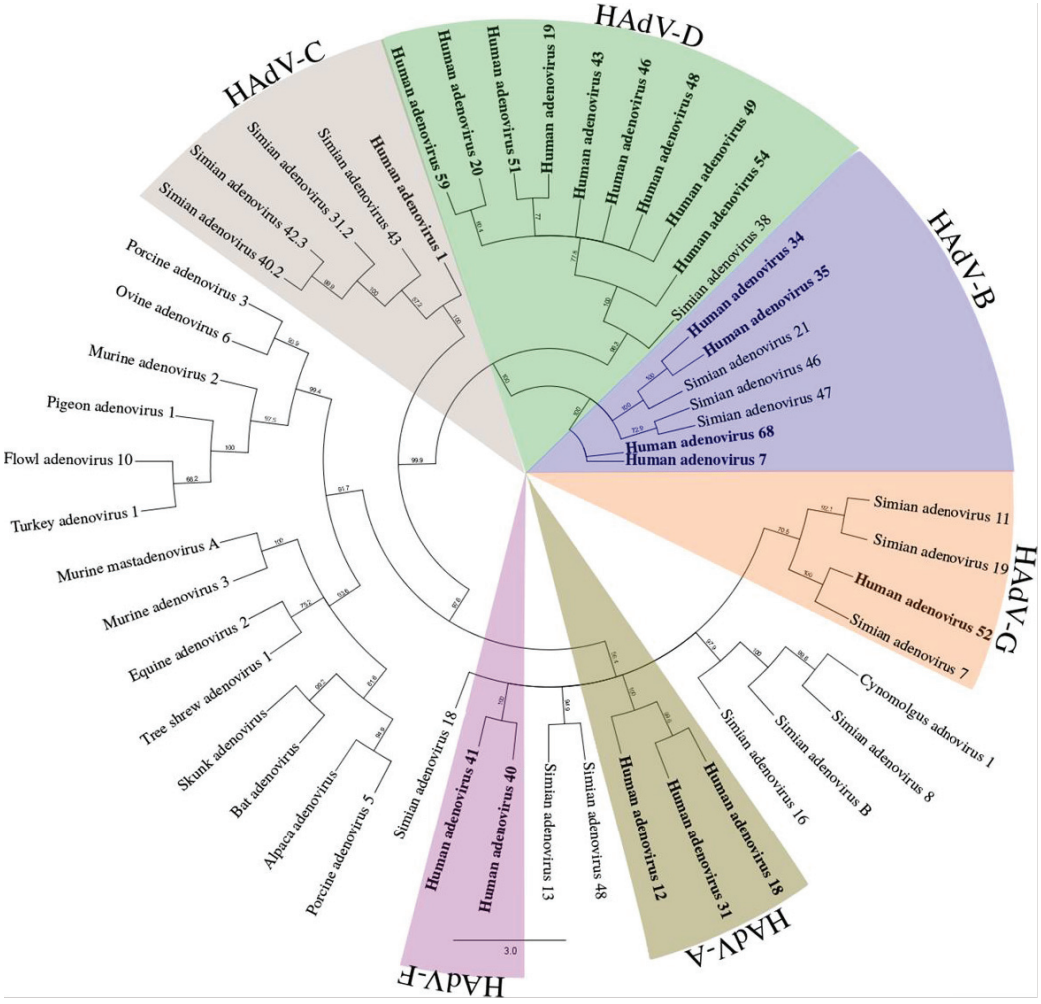


Figure 1. Relative abundance of viral species classified by family. The heatmap shows the relative abundance of 28 viral families detected over 3 different seasonal samples. Numbers within each cell represent the number of sequences that had at least a positive BLAST hit to known species and passed all the selection criteria. The colours range from green (not detected), to red (high relative abundance). Top row and right column correspond to the count sums by category and sample respectively.

Figure 2. Phylogenetic tree based on the complete nucleotide Adenovirus hexon region. Detected Adenovirus sequences from 454 sequencing experiment that presented a match in sewage were aligned. Human adenovirus species are shown in boldface, where colored arcs highlight the distinct taxon groups ranging from HAdV-A to HAdV-F. The tree was built using the neighbor-joining method and 1000 bootstrap replicates (bootstrap values are shown on the tree branches).



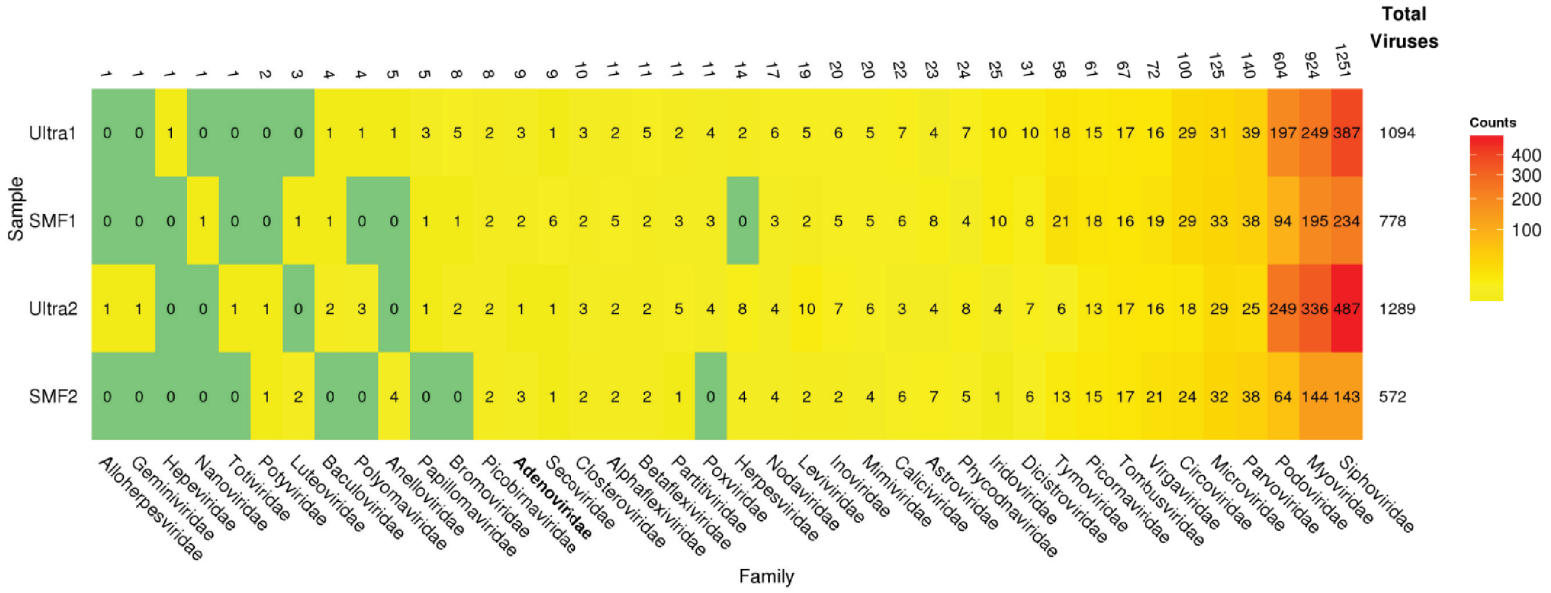


Figure 3. Relative abundance of viral species classified by family. The heatmap shows the relative abundance of 35 viral families detected over 4 different concentration method samples. Numbers within each cell represent the number of sequences that had at least a positive BLAST hit to known species and passed all the selection criteria. The colours range from green (not detected), to red (high relative abundance). Top row and right column correspond to the count sums by category and sample respectively.

Table 1.- Important and potentially pathogenic Human viral families detected in raw sewage by using 10L SMF

Family	Genus/Species	Winter	Spring	Summer	Total hits
Astroviridae	MAstV-1	48	7	5	60
	MAstV-6	8	1	10	19
	MAstV-8/9	18	0	2	20
Parvoviridae	HBoV-2	6	1	0	8
	HBoV-3	3	0	0	3
	HBoV-4	1	2	0	3
	AAV2	12	4	1	17
	AAV3	2	0	0	2
	AAV5	2	0	0	2
Caliciviridae	HSaVGI	7	2	4	13
	HSaVGII	11	0	3	14
	NoVGI	17	6	3	26
	NoVGII	20	1	1	22
Polyomaviridae	JCPyV	4	0	0	4
Circoviridae	Human circovirus	5	12	2	19
Picornaviridae	HAV	2	0	0	2
	Salivirus	20	20	22	62
	Cosavirus	0	2	1	3
	Rosavirus	1	0	0	1
	Enterovirus A	4	1	1	6
	Enterovirus B	10	4	5	19
	Enterovirus C	5	2	1	8
	Enterovirus D	1	0	0	1
	Enterovirus J	0	0	1	1
	Rabovirus	10	0	1	11
	Cardiovirus	8	7	5	20
	Aichi virus	16	12	42	70
Hepeviridae	HEV	4	0	0	4
Picobirnaviridae	Human picobirnavirus	37	11	15	63

Samples	WINTER			SPRING			SUMMER		
	Sequences		Nucleotides	Sequences		Nucleotides	Sequences		Nucleotides
MiSeq Raw READS	2,862,464		653,150,240	1,225,920		290,054,808	951,664		223,053,943
Clean READS									
Pair-Ends	2,820,868		644,330,961	1,172,444		266,982,894	779,690		172,598,345
Single-Ends	7,351		1,802,212	9,158		1,676,711	14,678		3,050,553
Total	2,828,219	98.80%		1,181,602	96.38%		794,368	83.47%	
Assembly (Contigs)									
Sequences	83,758		30,999,497	138,906		36,661,374	92,208		24,775,032
N50			391			251			251
Putative viral seqs	2,127	2.54%		3,342	2.41%		1,982	2.15%	
Seqs without BLAST hit	78,125			131,932			88,346		
Assembly (Singlets)									
Sequences	714,161			501,137			113,903		
Putative viral seqs	18,107	2.53%		8,029	1.61%		5,487	4.81%	
Seqs without BLAST hit	517,133			413,810			38,781		
	Estimated Value		**SE**	**Estimated Value**		**SE**	**Estimated Value**		**SE**
Richness	831.2		27.2	843.5		37.9	865.6		43.8

Supplementary material 1. Metagenomic sequencing summary statistics (waste water seasonal samples). Sequences and nucleotide counts are total values, number of pairs is half of the shown values. Percent of reads sequence refer to the total amount of raw reads, while the percent of sequences having a BLAST hit, or not, corresponds to the total number of assembled sequences (contigs and singletons). Difference between the sequences assigned to known viruses and the sequences without a BLAST hit relates to those sequences having a BLAST hit that has not passed all the filtering criteria for a valid species assignment.

Supplementary material 3.

Table 1.- Distribution of adenovirus types identified in wastewater sample. The number of assigned OTU's, the number of 454 reads associated to those taxonomic units and the % of representative sequences are provided.

Adenoviridae member	Adenovirus type	GenBank Accesion number	OTU's	454 reads	% of sequences
Murine adenovirus 2		NC_014899	965	31440	59.41043084
Human adenovirus 41	HAdV-F	DQ315364	655	15473	29.23847317
Simian adenovirus 48		HQ241818	384	1238	2.339380197
Human adenovirus type 51	HAdV-D	DQ149642	248	864	1.632653061
Ovine adenovirus 6		DQ630759	268	677	1.279289494
Simian adenovirus 19		KP329565	158	500	0.944822373
Human adenovirus type 48	HAdV-D	EF153473	48	382	0.721844293
Human adenovirus 40	HAdV-F	L19443	119	342	0.646258503
Equine adenovirus 2		AEEHEXEND	103	318	0.600907029
Human adenovirus type 46	HAdV-D	AY875648	34	260	0.491307634
Simian adenovirus 47		FJ025929	112	250	0.472411187
Human adenovirus type 35	HAdV-B	AY271307S	73	175	0.330687831
Simian adenovirus 38		FJ025922	74	146	0.275888133
Bat adenovirus TJM		GU226970	53	143	0.270219199
Simian adenovirus B		KC693021	73	133	0.251322751
Human adenovirus 59	HAdV-D	JF799911	70	118	0.22297808
Simian adenovirus 18		FJ025931	43	82	0.154950869
Human adenovirus type 19	HAdV-D	DQ149618	23	63	0.119047619
Human adenovirus type 49	HAdV-D	DQ393829	14	57	0.107709751
Human adenovirus 68	HAdV-B	JN860678	8	36	0.068027211
Skunk adenovirus PB1		KP238322	21	28	0.052910053
Alpaca adenovirus		GQ499375	21	28	0.052910053
Simian adenovirus 16		NC_028105	8	21	0.03968254
Human adenovirus type 43	HAdV-D	DQ149636	2	16	0.030234316
Simian adenovirus 7		DQ792570	13	13	0.024565382
Simian adenovirus 42.3		FJ025925	10	13	0.024565382
Simian adenovirus 21		AC_000010	5	11	0.020786092
Human adenovirus type 1	HAdV-C	AF534906	4	9	0.017006803
Human adenovirus 54	HAdV-D	AB333801	2	9	0.017006803

Murine adenovirus 3		NC_012584	8	8	0.015117158
Simian adenovirus 11		KP329562	8	8	0.015117158
Simian adenovirus 43		FJ025900	4	6	0.011337868
Human adenovirus 12	HAdV-A	NC_001460	2	6	0.011337868
Simian adenovirus 13		NC_028103	5	5	0.009448224
Pigeon adenovirus 1		FN824512	4	5	0.009448224
Tree shrew adenovirus 1		AF258784	3	5	0.009448224
Human adenovirus type 31	HAdV-A	AM749299	4	4	0.007558579
Turkey adenovirus 1		NC_014564	1	4	0.007558579
Cynomolgus adenovirus		KT013209	3	3	0.005668934
Porcine adenovirus 5		AF289262	3	3	0.005668934
Simian adenovirus 40.2		FJ025926	2	3	0.005668934
Simian adenovirus 8		NC_028113	2	2	0.003779289
Human adenovirus 18	HAdV-A	GU191019	2	2	0.003779289
Human adenovirus 52	HAdV-G	DQ923122	2	2	0.003779289
Human adenovirus type 34	HAdV-B	AY737797	2	2	0.003779289
Murine adenovirus A		NC_000942	1	1	0.001889645
Simian adenovirus 46		FJ025930	1	1	0.001889645
Simian adenovirus 31.2		FJ025904	1	1	0.001889645
Fowl adenovirus 10		FAU26221	1	1	0.001889645
Human adenovirus type 20	HAdV-D	DQ149619	1	1	0.001889645
Human adenovirus type 7	HAdV-B	AC_000018	1	1	0.001889645
Porcine adenovirus 3		AB026117	1	1	0.001889645

Table 2.- Primers and conditions used for target enrichment assay for HAdV hexon are expressed in the following table:

PCR round	Gene (size)	Primer ID	Sequence (5'-3')	PCR Mix	Amplification conditions
1st	300	Forward	GCCSCARTGGKCNTACATGCACAT	In a final volume of 50 µl containing 1× Gold buffer (Applied Biosystems, Inc) at 50 mM, MgCl2 25 mM, 0,1 mM of each deoxynucleotide, 0,5 µM of each primer, 1U Taq polymerase and 10ul of the concentrate extraction	35 cycles of denaturation at 94ºC 30s, annealing at 50ºC 30s and extension at 72ºC 1min
		Reverse	CARNACVCCNCKRATGTCAAA		
2nd (introduction of barcodes & key index)	360	Forward	CCATCTCATCCCTGCGTGTCTCCGACTCAGGCCSCARTGGKCNTACATGCACAT	In a final volume of 50 µl containing 1× Gold buffer (Applied Biosystems, Inc) at 50 mM, MgCl2 25 mM, 0,1 mM of each deoxynucleotide, 0,5 µM of each primer, 1U Taq polymerase and 3ul of product from the first PCR	25 cycles of denaturation at 94ºC 30s, annealing at 50ºC 30s and extension at 72ºC 1min
		Reverse	CCTATCCCCTGTGTGCCTTGGCAGTCTCAGCARNACVCCNCKRATGTCAAA		

Samples	SMF1			SMF2			Ultra1			Ultra2		
	Sequences		Nucleotides	Sequences		Nucleotides	Sequences		Nucleotides	Sequences		Nucleotides
MiSeq Raw READS	3,846,130		962,076,715	4,079,546		1,039,169,432	4,078,152		1,015,816,890	3,441,666		924,822,940
Clean READS												
Pair-Ends	3,654,800		830,379,508	3,841,750		884,294,646	3,871,162		879,354,886	3,237,524		416,596,088
Single-Ends	176		40,654	998		238,833	170		40,190	136		32,184
Total	3,654,976	95.03%		3,842,748	94.20%		3,871,332	94.93%		3,237,660	94.07%	
Assembly (Contigs)												
Sequences	157,402		60,268,319	142,298		53,932,002	212,737		81,299,715	282,605		111,648,857
N50			391			386			387			396
Putative viral seqs	3,070	1.95%		2,017	1.42%		4,568	2.15%		5,629	1.99%	
Seqs without BLAST hit	147,950			135,870			199,878			265,817		
Assembly (Singlets)												
Sequences	3,372,435			3,509,087			3,466,437			2,663,042		
Putative viral seqs	18,647	0.55%		10,744	0.31%		24,825	0.71%		27,803	1.04%	
Seqs without BLAST hit	3,069,901			3,114,207			3,152,828			2,428,443		
	Estimated Value		SE	Estimated Value		SE	Estimated Value		SE	Estimated Value		SE
Richness		755.8	23.6		541	18.9		1066.4	23.2		1318.5	27.6

Supplementary material 4. Metagenomic sequencing summary statistics (concentration methods comparison). Sequences and nucleotide counts are total values, number of pairs is half of the shown values. Percent of reads sequence refer to the total number of raw reads, while the percent of sequences having a BLAST hit, or not, corresponds to the total number of assembled sequences (contigs and singletons). Difference between the sequences assigned to known viruses and the sequences without a BLAST hit relates to those sequences having a BLAST hit that has not passed all the filtering criteria for a valid species assignment.

Altres publicacions 1

Evidence of viral dissemination and seasonality in a Mediterranean river catchment: Implications for water pollution management

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..

Journal of Environmental Management (2015) Aug 15;159:58-67

Journal of Environmental Management 159 (2015) 58–67

Contents lists available at ScienceDirect

Journal of Environmental Management

journal homepage: www.elsevier.com/locate/jenvman

Research paper

Evidence of viral dissemination and seasonality in a Mediterranean river catchment: Implications for water pollution management

Marta Rusiñol [a], Xavier Fernandez-Cassi [a], Natàlia Timoneda [a, b], Anna Carratalà [a], Josep Francesc Abril [b, c], Carolina Silvera [d], Maria José Figueras [d], Emiliano Gelati [e], Xavier Rodó [e], David Kay [f], Peter Wyn-Jones [f], Sílvia Bofill-Mas [a], Rosina Girones [a, *]

[a] Laboratory of Virus Contaminants of Water and Food, Department of Microbiology, University of Barcelona, Barcelona, Catalonia, Spain
[b] Institute of Biomedicine of the University of Barcelona (IBUB), University of Barcelona, Barcelona, Catalonia, Spain
[c] Computational Genomics Laboratory, Department of Genetics, University of Barcelona, Barcelona, Catalonia, Spain
[d] Microbiology Unit, Faculty of Medicine and Health Sciences, IISPV, University Rovira and Virgili, Reus, Catalonia, Spain
[e] Catalan Institute of Climate Sciences (IC3), Barcelona, Catalonia, Spain
[f] Institute of Geography and Earth Sciences (IGES), Aberystwyth University, Aberystwyth, United Kingdom

ARTICLE INFO

Article history:
Received 25 November 2014
Received in revised form
13 May 2015
Accepted 16 May 2015
Available online 2 June 2015

Keywords:
Human adenovirus
Merkel cell polyomavirus
Norovirus
Hepatitis E virus
River water

ABSTRACT

Conventional wastewater treatment does not completely remove and/or inactive viruses; consequently, viruses excreted by the population can be detected in the environment. This study was undertaken to investigate the distribution and seasonality of human viruses and faecal indicator bacteria (FIB) in a river catchment located in a typical Mediterranean climate region and to discuss future trends in relation to climate change. Sample matrices included river water, untreated and treated wastewater from a wastewater treatment plant within the catchment area, and seawater from potentially impacted bathing water. Five viruses were analysed in the study. Human adenovirus (HAdV) and JC polyomavirus (JCPyV) were analysed as indicators of human faecal contamination of human pathogens; both were reported in urban wastewater (mean values of 10^6 and 10^5 GC/L, respectively), river water (10^3 and 10^2 GC/L) and seawater (10^2 and 10^1 GC/L). Human Merkel Cell polyomavirus (MCPyV), which is associated with Merkel Cell carcinoma, was detected in 75% of the raw wastewater samples (31/37) and quantified by a newly developed quantitative polymerase chain reaction (qPCR) assay with mean concentrations of 10^4 GC/L. This virus is related to skin cancer in susceptible individuals and was found in 29% and 18% of river water and seawater samples, respectively. Seasonality was only observed for norovirus genogroup II (NoV GGII), which was more abundant in cold months with levels up to 10^4 GC/L in river water. Human hepatitis E virus (HEV) was detected in 13.5% of the wastewater samples when analysed by nested PCR (nPCR). Secondary biological treatment (i.e., activated sludge) and tertiary sewage disinfection including chlorination, flocculation and UV radiation removed between 2.22 and 4.52 $\log_{10}$ of the viral concentrations. Climate projections for the Mediterranean climate areas and the selected river catchment estimate general warming and changes in precipitation distribution. Persistent decreases in precipitation during summer can lead to a higher presence of human viruses because river and sea water present the highest viral concentrations during warmer months. In a global context, wastewater management will be the key to preventing environmental dispersion of human faecal pathogens in future climate change scenarios.

1. Introduction

Understanding the environmental fate of pathogens is useful for minimising the risk to humans. Human viruses are excreted at high concentrations in faeces and urine and can be transmitted through improperly treated wastewater. As part of the EU-FP7-funded VIROCLIME project, the present study developed a surveillance program centred on a typically Mediterranean climate region: the Llobregat River basin (Catalonia, northeast of Spain).

Water microbiological quality is traditionally defined and

* Corresponding author. Department of Microbiology, University of Barcelona, Avinguda Diagonal 643, 08028 Barcelona, Spain.
E-mail address: rgirones@ub.edu (R. Girones).

http://dx.doi.org/10.1016/j.jenvman.2015.05.019
0301-4797/

regulated by faecal indicators that are more sensitive to water treatment and environmental conditions than viral pathogens (de RodaHusman et al., 2009; Figueras and Borrego, 2010). Human adenoviruses (HAdV) and JC polyomavirus (JCPyV) have been proposed as specific human faecal indicators based on their high prevalence in all geographical areas surveyed to date (Bofill-Mas et al., 2000a,b; Pina et al., 1998). These viruses have been widely used to trace faecal pollution in the environment (Bofill-Mas et al., 2013; Rusiñol et al., 2013). Both viruses are also human pathogens related with enteric and respiratory illness, eye infections and severe disease in immunocompromised patients (Crabtree et al., 1997; Imperiale, 2000). Merkel cell polyomavirus (MCPyV), which has been found integrated in a very high percentage of Merkel cell carcinomas, has also been isolated from urban sewage river water (Bofill-Mas et al., 2010; Calgua et al., 2013a). Although a subcutaneous route seems to be the most likely transmission pathway, the identification of this cancer-related polyomavirus in sewage has been recognised as a significant research question (Spurgeon and Lambert, 2013).

Norovirus genogroup II (NoV GGII) is a single stranded RNA virus that is recognised as the major cause of self-limiting viral gastroenteritis (Craun et al., 2010; Kroneman et al., 2008). Furthermore, NoV GGII is believed to be the most significant etiological agent in documented recreational water-borne outbreaks, followed by adenoviruses (Sinclair et al., 2009). In the wider community, NoV GGII has been associated with the majority of recorded gastroenteritis cases (Lopman et al., 2004). Person-to-person transmission is the most common pathway, but NoV is spread by several routes that include contaminated shellfish, fresh food, processed food and water (Mathijs et al., 2012). NoVs are highly infectious, with a single virus particle having a probability of infection approaching 49% (Teunis et al., 2008).

Hepatitis E virus (HEV) also has a water-borne route of transmission (Orrù et al., 2004). Although HEV is endemic in low-income countries where it produces acute and self-limited hepatitis, it also circulates in industrialised countries (Clemente-Casares et al., 2003; Legrand-Abravanel et al., 2009). In Spain, HEV is found in 30% of urban sewage, which is considered an important source of HEV dissemination (Rodriguez-Manzano et al., 2010).

During the 18-month study period, HAdV, JCPyV, MCPyV, NoV GGII and HEV together with two faecal indicator bacteria (FIB) (*Escherichia coli* (EC) and intestinal enterococci (IE)) were surveyed in a Mediterranean river catchment. Given that the main viral inputs were likely derived from raw or treated effluents, raw and treated water samples were also tested. In this study, we also assessed the repeatability of the skimmed milk flocculation protocol used to concentrate viruses from different water matrices, designed a new qPCR method for the specific detection of MCPyV in water and finally discussed further trends of virus pollution considering climate projections for the Llobregat river basin.

Climate change models continue to predict higher stress on water resources that may contribute to pathogen dispersion, including bacteria, viruses and protozoa (IPCC, 2007). The overall goal of this investigation was to provide empirical data on the spatial and temporal patterns of viral pathogens and indicators in a changing river basin to allow public health managers to assess risks in future scenarios.

2. Materials and methods

2.1. Sample and data collection

The Llobregat River flows approximately 170 km from the Pyrenees Mountains to the Mediterranean Sea, discharging near the city of Barcelona. The 4950-km2 river basin accommodates 5 million people, including more than half of the Catalan population. Treated urban sewage, industrial effluents and agricultural runoff affect the quality of raw river water, which is the main source of treated drinking water for Barcelona and its metropolitan area. In fact, the urban water supply constitutes 65% of the total Llobregat water demand (ACA, 2012). The annual average river discharge volume is 690 cubic hectometers/second (hm^3/s), of which over 40% consists of effluents from the approximately 50 wastewater treatment plants (WWTP) located within the Llobregat river catchment area (annual mean discharges: 300 hm^3/s) (ACA, 2012).

A total of 196 water samples were collected from January 2011 to June 2012 from 4 different sampling sites (Fig. 1). Site A is located 80 km upstream of the river mouth ($n = 36$ samples). The water flow rate or mean river discharge volume at site A was 0.5 m^3/s, corresponding to 14 hm^3/s of mean annual discharge. Ten WWTP are located upstream of the sampling site and present a mean annual effluent discharge of 6 hm^3/s. Site B ($n = 37$) is located approximately 70 km downstream from site A and 10 km from the river mouth close to the city of Barcelona. The Llobregat River at site B has 255 hm^3/s of annual discharge, of which 154 hm^3/s come from fifty upstream WWTP effluents.

Most of the sewage generated in the metropolitan area of Barcelona is treated at site C, a WWTP designed to handle approximately 153 hm^3/s of water using secondary and tertiary treatment. Ninety-one samples were collected from the WWTP, first at the inflow (C_1, $n = 37$), at the outlet of secondary treatment consisting of a biological reactor and sedimentation (C_2, $n = 32$) and at the outflow from the tertiary treatment (C_3, $n = 22$) that comprised chlorination, Actiflo® filtration and flocculation and UV disinfection. These samples were collected when the complete treatment system was operating. Site D consists of a marine bathing water site affected by the river plume diluted by Mediterranean seawater ($n = 32$). Water samples of 10 L were collected from the river, the sea and the tertiary effluent, whereas only 50 mL were collected from the secondary effluent and raw wastewater.

All viral (HAdV, JCPyV, MCPyV and NoV GGII) and bacterial (EC and IE) parameters were quantified bimonthly in all river water samples (sites A and B) and in all seawater samples from the beach affected by the Llobregat river plume (site D). Raw wastewater and two differently treated effluents from a WWTP (site C) were simultaneously collected and also tested for the presence of HEV by nPCR.

Stream flow measurements were recorded in two gauges located at sites A and B by the Catalan Water Agency (ACA, 2012). We assumed wastewater treatment plant (WWTP) effluents were equivalent to the recorded WWTP inflows. Conductivity, pH and water temperature data were determined in the field for each sample. Precipitation data were acquired from the Catalan Meteorological Institute (Meteocat, 2014). Data from seasonal river flow, precipitation and temperature is shown in the Supplementary Material (S1).

2.2. Virus analyses

2.2.1. Concentration and nucleic acid extraction

The detection of viruses in the environment requires the concentration of the viral particles into small volumes. This study was conducted using optimised Standardised Operational Procedures (SOPs) for virus concentration, nucleic acid extraction and quantitative PCR (qPCR and RT-qPCR) detection. The SOPs incorporated process controls and standard plasmid preparation. First, samples with high levels of organic matter or sand were permitted to sediment for 1 h to avoid any interference with the concentration method. Then, clarified water was transferred to a new container without disturbing the sediment. As established in ISO 5667-

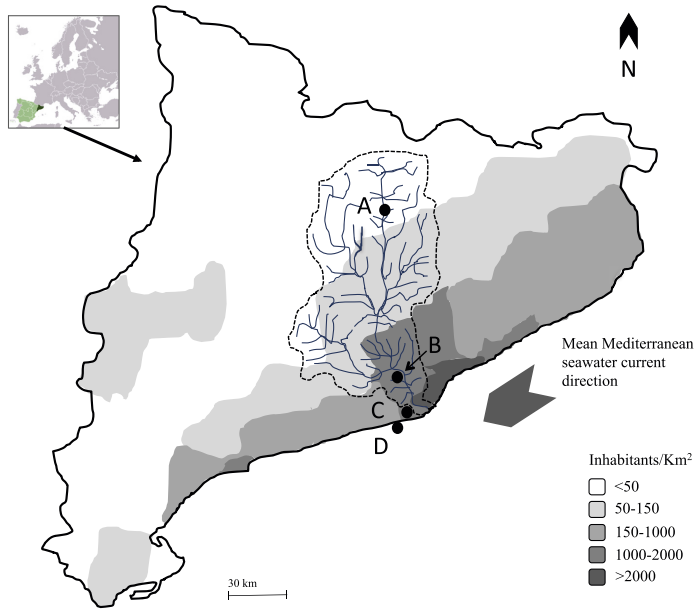


Fig. 1. Sampling sites location in the Llobregat river catchment (Catalonia, Spain). Site A: upstream river water; Site B: downstream river water; Site C: raw sewage (C_1), secondary (C_2) and terciary (C_3) effluents from a WWTP; Site D: seawater.

9:1992, seawater samples were collected from one meter below sea level to avoid contamination with sand and suspended macro algae.

Water samples were concentrated as previously described by Calgua and co-workers (Calgua et al., 2013a, 2013b, 2008). Briefly, a pre-flocculated 1% (w/v) skimmed milk solution (PSM) was prepared by dissolving 10 g skimmed milk powder (Difco, Detroit, MI, USA) in 1 L of artificial seawater at pH 3.5 (Sigma–Aldrich Chemie GMBH, Steinheim, Germany). All samples were carefully acidified to pH 3.5 using 1N HCl and the conductivity was adjusted to 1.5 mS/cm^2. Then, the PSM solution was added to each of the previously conditioned samples to obtain a final concentration of skimmed milk in the sample of 0.01%. Samples were stirred for 8 h at room temperature and flocks were allowed to sediment by gravity for 8 h. Supernatants were carefully removed using a vacuum pump without disturbing the sediment. The sediment was resuspended using 10 mL of phosphate buffer at pH 7.5, with the exception of the raw sewage samples and secondary treated effluents that were suspended in 1 mL. On the same day, viral DNA was extracted from all samples using the QIAamp Viral RNA kit (Qiagen, Inc., Valencia, CA). Adenovirus type 35 and UltraPure™ DNase/RNase-Free distilled water were used as positive and negative controls of the nucleic acid (NA) extraction experiment, respectively. Finally, NA elutants were stored at −20 °C until use.

2.2.2. Quantitative and nested PCR assays

Specific real-time quantitation of DNA viruses (HAdV and JCPyV) by qPCR or an RNA virus (NoV GGII) by quantitative reverse transcription PCR (qRT-PCR) was performed as previously described (Table 1) using TaqMan® Universal PCR Master Mix and the RNA UltraSense™ One-Step qRT-PCR System, respectively (Invitrogen, Carlsbad, CA, USA) (Hernroth et al., 2002; Kageyama et al., 2003; Loisy et al., 2005; Pal et al., 2006). Undiluted and log10 dilutions of the nucleic acid extracts were analysed in duplicate. The equivalence of 35 mL for DNA viruses and 17.5 mL for the RNA virus were tested from river and seawater, whereas 1.7 mL and 0.9 mL, respectively, were tested from wastewater. All qPCRs included more than one non-template control (NTC) to demonstrate that the mix did not produce fluorescence due to contamination. Quantitation was performed with an MX3000P sequence detector system (Stratagene, La Jolla, CA, USA).

The amplification conditions of the HEV nested RT-PCR (nRT-PCR) methods used for qualitative detection were described elsewhere (Erker et al., 1999). Primers are described in Table 1. The reverse transcription of the extracted RNA was performed with a one-step RT-PCR Kit (Qiagen, Valencia, CA, USA) and the nRT-PCR was performed with AmpliTaq™ Gold DNA polymerase (Applied Biosystems Foster City, CA, USA).

2.2.3. Development of the MCPyV qPCR assay

A new qPCR assay was designed for MCPyV detection in water matrices. A fragment of the VP1 gene (132 bp) was obtained by applying a specific PCR (Bofill-Mas et al., 2010) and cloned into a pGEM-T Easy vector (Promega, Madison, WI, USA). Standard curves were generated by transferring the plasmid construct into *E. coli* DH5α cells (Invitrogen, Carlsbad, CA, USA). After verifying the presence of transformed colonies containing the target sequence by conventional PCR and purifying them with the Qiagen Plasmid Midi kit (Qiagen GmbH Inc., Hilden, Germany), the constructions were linearised with the Sal I restriction enzyme (Promega, Madison, WI). Then, two primers and a hydrolysis probe were designed based on the TaqMan® assay to amplify the specific VP1 fragment of the viral genome (Table 1). Annealing temperatures and primer and

Table 1
Oligonucleotide primers and probes used for the detection and quantification of viral pathogens.

Virus	Primers and probes	Position[a]	Sequence (5′-3′)	References
Human adenovirus (HAdV)	ADF	Hexon gene	CWTACATGCACATCKCSGG	Hernroth et al., 2002
	ADR	18869–18937	CRCGGGCRAAYTGCACCAG	
	ADP1		FAM-CCGGGCTCAGGTACTCCGAGGCGTCCT-BHQ1	
JC Polyomavirus (JCPyV)	JE3F	T-antigen gene	ATGTTTGCCAGTGATGATGAAAA	Pal et al., 2006
	JE3R	4251–4482	GGAAAGTCTTTAGGGTCTTCTACCTTT	
	JE3P		FAM-AGGATCCCAACACTCTACCCCACCTAAAAAGA-BHQ1	
Merkel Cell Polyomavirus (MCPyV)	MCF	VP1 gene 1232–1293	ATTTGGGTAATGCTATCTTCTC	This study
	MCR		CTAATGTTGCCTCAGTTCCAA	
	MCP		FAM- AACCACAGATAATACTTCCACTCCT-BHQ1	
Norovirus genogroup II (NoV GGII)	QNIF2d	Capsid protein gene	ATGTTCAGRTGGATGAGRTTCTCWGA	Primers: Loisy et al., 2005
	COG2R	5012–5100	TCGACGCCATCTTCATTCACA	Probe: Kageyama et al., 2003
	QNIFS		FAM-AGCACGTGGGAGGCGATCG-TAMRA	
Hepatitis E[b] virus (HEV)	HEVORF2con-a1	ORF2 region	GACAGAATTRATTTCGTCGGCTGG	Erker et al., 1999
	HEVORF2con-s1	6283–6479	CTTGTTCRTGYTGGTTRTCATAATC	
	HEVORF2con-s2		GTYGTCTCRGCCAATGGCGAGC	

[a] The sequence positions are referred to strains J01917.1 (HAdV), NC001699.1 (JCPyV), HM011557.1 (MCPyV), AF145896 (NoVGII) and AF060668 (HEV).
[b] Seminested PCR.

probe concentrations for the novel MCPyV qPCR were optimised by assaying primer concentrations ranging from 0.4 to 0.9 μM and probe concentrations ranging from 0.225 to 0.4 μM for each reaction. Amplifications were performed in a mixture containing 10 μl DNA, 15 μl TaqMan® Universal PCR Master Mix (Applied Biosystems), 0.9 μM of each primer (MCF and MCR) and 0.225 μM of the probe (MCP). Following activation with AmpliTaq™ Gold DNA polymerase for 10 min at 95 °C, 40 cycles (15 s at 95 °C and 1 min at 60 °C) were performed with an MX3000P detector system (Stratagene, La Jolla, CA, USA). Serial dilutions of the confirmed standard ranging from 10^0 to 10^5 molecules per 10 μL were performed with TE buffer and stored at −80 °C until use. Known amounts of standard DNA containing 10^0, 2×10^0, 5×10^0, 10^1, 2×10^1, 5×10^1 and 10^2 GC/reaction were analysed according to MIQE guidelines using 6 replicates to determine the sensitivity of the qPCR assay (Bustin et al., 2009). The specificity was verified with available standard plasmids from other human polyomaviruses (JCPyV, BKPyV, KIPyV and WUPyV) and animal polyomaviruses (ovine PyV and bovine PyV).

2.2.4. Control viruses and plasmid DNA for the viral qPCR assays

Human adenovirus type 35 (HAdV35) and norovirus genogroup II type 13 (NoV GGII.13) stocks were kindly donated by Dr. A. Allard of the University of Umeå (Sweden) and were used as positive process controls. On each sampling day, an extra sample was collected and spiked with HAdV35 (10^5 viral particles/mL) as a process control for flocculation, NA extraction and DNA quantification. The NoV GGII.13 genome was also extracted from each sample as a positive control for nucleic acid extraction and RNA quantification (10^4 genome copies in each reaction of 5 μL).

Plasmid DNA was used as a positive control and as a quantitative standard. The hexon region (8961 bp) and the whole genome (5130 bp) of HAdV41 and JCPyV Mad1, respectively, were cloned into pBR322. To reduce the possibility of DNA contamination in the laboratory, 10 μg of plasmid DNA was linearised with BamHI for HAdV and NruI for JCPyV (Promega, Madison, WI) and the reaction products were purified and quantified. The capsid protein region of NoV GGII.13 cloned into the pTrueBlue®-Pvu II vector (kindly donated by Dr. J. Vinjé, CDC, Atlanta) was used as a qRT-PCR standard. A total of 10 μg of this construct was linearised with XhoI (Promega) to prevent contamination. Serial dilutions ranging from 10^0 to 10^5 molecules per 5 or 10 μL for RNA or DNA virus qPCR, respectively, were performed with TE buffer. Standard dilutions were distributed into tubes and stored at −80 °C until use. Specific primers and hydrolysis probes are described in Table 1. UltraPure™ DNase/RNase-Free distilled water was used as a negative control for the NA extraction and qPCR assays.

2.2.5. Recovery efficiency and repeatability of the concentration procedure

An intra-laboratory assay was performed to test the repeatability of the concentration method. Briefly, 200 L of seawater, river water or mineral water spiked with HAdV35 and NoV GGII.13 was mixed in a large plastic container. Water was mixed by manual stirring, acidified and then distributed into twenty 10-L sample containers. This procedure was repeated for each matrix and included mineral water as a control. Viral particles were concentrated as described in 2.2.1. HAdV and NoV GGII.13 were quantified by qPCR and qRT-PCR after nucleic acid extraction, respectively. Additionally, indigenous JCPyV was quantified in the 20 seawater replicates.

2.3. Faecal indicator bacteria (FIB) detection

For FIB detection, 100 mL of each sample was collected in parallel from all sites. All samples were kept on ice and processed within 24 h. The enumeration of EC was carried out in a 96-well microplate (MUG/EC 355-3782, BioRad, Barcelona, Spain®) according to ISO 9308-2:2012 using the most probable number (MPN) procedure (Donovan et al., 1998) to detect EC based on the expression of the enzyme β-D-glucuronidase present in most EC strains. Intestinal enterococci were also quantified by MPN in a 96-well microplate (MUG/EC 355-3783, BioRad®) following the ISO 7899-1:1998 procedure based on the detection of the expression of the β-glycosidase enzyme characteristic of enterococci.

2.4. Statistical analyses

Statistical analysis was performed using R software version 2.15.1 (Verzani, 2004). The data distribution was tested with the Shapiro–Wilk and Kolmogorov–Smirnov tests. After proving that the raw data were not normally distributed, the viral and bacterial raw data were $\log_{10}$ transformed to improve normality and facilitate parametric analyses. Correlation analysis of the transformed ($\log_{10}$) data was completed using both parametric (Pearson correlation) and non-parametric (Spearman correlation) approaches. The results were considered to be significant when p was <0.01. Values less than the detection limit were given the value of 1. Correlations between the temperature, flow, precipitation and microbial concentration were calculated. Coefficients of variability

of the concentration procedure calculated for each water matrix with raw (GC/L), and $\log_{10}$ transformed data were tested for normality and variability (Coefficient of Variation (CV), where CV = standard deviation/mean value x 100) among the replicate concentrations.

2.5. Climate projections for a Mediterranean river catchment: methods and datasets

Temperature and precipitation records were extracted from the ERA-interim reanalysis database (http://data-portal.ecmwf.int). Because data can contain gaps, we established the criteria that for a grid point to be eligible it must contain a minimal fraction of 30 valid points for each grid point comprised in the box 0 to 5°E and 40°N-45°N. Precipitation is reported in mm/day and temperature in degrees Kelvin (°K). Both precipitation and average, minimum and maximum temperature simulations were downloaded from the Climate Model Intercomparison Project version 5 (CMIP5) at http://cmip-pcmdi.llnl.gov/cmip5/data_portal.html. The data for each point is the monthly multi-model mean of historical + rcp26 experiments for the interval 1860–2100 of the following group of climate models: BCC-CSM1-1, BCC-CSM1-1-m, BNU-ESM, CanESM2, CCSM4, CESM1-CAM5, CNRM-CM5, CSIRO-Mk3-6-0, EC-EARTH, FGOALS-g2, FIO-ESM, GFDL-CM3, GFDL-ESM2G, GFDL-ESM2M, GISS-E2-H, GISS-E2-R, HadGEM2-AO, HadGEM2-ES, IPSL-CM5A-LR, IPSL-CM5A-MR, MIROC5, MIROC-ESM, MIROC-ESM-CHEM, MPI-ESM-LR, MPI-ESM-MR, MRI-CGCM3, NorESM1-M and NorESM1-ME (see Supplementary Information S2).

3. Results and discussion

3.1. Performance of the assays

3.1.1. Recovery efficiency and repeatability of the skimmed milk flocculation protocol

Based on the outcome of the normality test, CV values were calculated with the $\log_{10}$-transformed data generated from the replicate enumerations for each determinant. Indigenous JCPyV presented the most repeatable results with a CV of 12.4%. Smaller variations in HAdV results were found among the 20 river water replicates (CV: 14.8%), whereas NoV GGII.13 measurements were more repeatable in mineral water (CV: 20.3%). The highest CV was calculated for NoV GGII.13 in seawater (CV: 36.3%). The concentration protocol by means of skimmed milk flocculation proved to be repeatable for use in the study.

3.1.2. Sensitivity and specificity of the new MCPyV quantitative assay

The qPCR assay developed for the quantification of MCPyV was shown to be specific both by “in silico” sequence analysis of the primers and the probe considering nucleotide sequence databases (NCBI BLAST) and by experimental assays. No false positive results were detected due to cross-reactivity with non-target DNA from the viruses contained in the different plasmid constructs assayed (human polyomaviruses JCPyV, BKPyV, KIPyV and WUPyV and animal polyomaviruses OPyV and BPyV). A total of 20 and 10 DNA genome copies were detected in 100% and 80% of the performed qPCR reactions, respectively. Sensitivity did not vary even when high levels of exogenous but related viral DNA (samples with high levels of JCPyV) were added to the test tubes.

3.2. Occurrence of viral pathogens and indicator microorganisms in river and seawater

The geometric mean genomic copies ($\log_{10}$ GC/L or MPN/100 mL) and concentration ranges of HAdV, JCPyV, MCPyV, NoV GGII, EC and IE found in each sampling location and season are shown in Fig. 2. HAdV and JCPyV were commonly detected in river water samples with prevalences of 80% and 59%, respectively, and at similar concentrations to those previously described (Albinana-Gimenez et al., 2009). These concentrations were stable throughout the year, with mean values of 8.1×10^2 GC/L in site A and 1.4×10^3 GC/L in site B. Overall, 100% of river and seawater samples were positive for HAdV in the summer and mean values were at times as high as those quantified in treated sewage effluent from tertiary treatment plants. Despite the lack of detection of JCPyV in the spring from site A, the prevalence was still high and the mean concentration was similar to that reported for HAdV. No JCPyV was detected in marine samples collected in winter. The lower frequency of HAdV and JCPyV in seawater was consistent with other studies that suggested marine dispersion, dilution and disinfection due to the high salinity may be the cause of the low concentrations reported (Hawley and Garver, 2008; Wyn-jones et al., 2010).

MCPyV was principally detected among the river sampling sites, with a higher prevalence in seawater during the summer (6.9×10^2 GC/L) (Fig. 2). Prevalence and concentrations of MCPyV in river water were similar to those previously reported (Calgua et al., 2013a). The most significant new finding was the detection of MCPyV in seawater during the summer bathing season, which may indicate a risk of transmission for recreators. The presence of this polyomavirus that has been described to have a skin tropism may also be linked to the recreational activities themselves rather than being exclusively derived from sewage effluent inputs into the bathing zone. This is the first description of the presence of MCPyV in seawater samples and bathing water, suggesting a putative transmission pathway for this pathogen.

NoV GGII prevalence showed a marked seasonality: the highest concentrations were measured during winter months with colder temperatures. A high prevalence of NoV GGII was detected in the studied river in winter (83% and 100% of positive samples in sampling points A and B, respectively) and autumn (77% and 88% in sampling points A and B, respectively) (Fig. 2). Outbreaks of winter vomiting syndrome caused by noroviruses in the environment are well reported (Kitajima et al., 2012, 2010; Nordgren et al., 2009; Lopman et al., 2009). Recently, other authors performed a norovirus survey in the same river catchment and reported similar results to those confirmed in the present investigation (Collado et al., 2010; Pérez-Sautu et al., 2012).

EC and IE were detected during all seasons at both river sites (A and B) with mean values of 10^3 MPN/100 mL, whereas in seawater FIB were under the limit of detection of the technique (15 MPN/100 mL) and were only detected in spring (EC) and autumn (EC and IE). Occasionally after spring and early summer rain events FIB levels exceeded those limits, but no significant FIB and rain correlations were observed (p-value>0.01). Moreover, most FIB levels in seawater samples complied with the ‘good’ EU bathing water criteria, with less than 100 cfu/100 mL of EI and less than 250 cfu/100 mL of EC (SpanishGovernment, 2007b).

3.3. Viruses and FIB in wastewater samples and treatment removal efficiencies

HAdV were detected in all 37 influent raw sewage samples with similar geometric mean values during the year (data not shown). High viral titers were observed in untreated wastewater (8.4×10^5 GC/L), secondary treated effluents (1.2×10^5 GC/L) and tertiary treated effluents (1.9×10^3 GC/L). Moreover, JCPyV was consistently prevalent in raw wastewater and secondary treated effluents with mean concentrations of 7.5×10^5 GC/L and

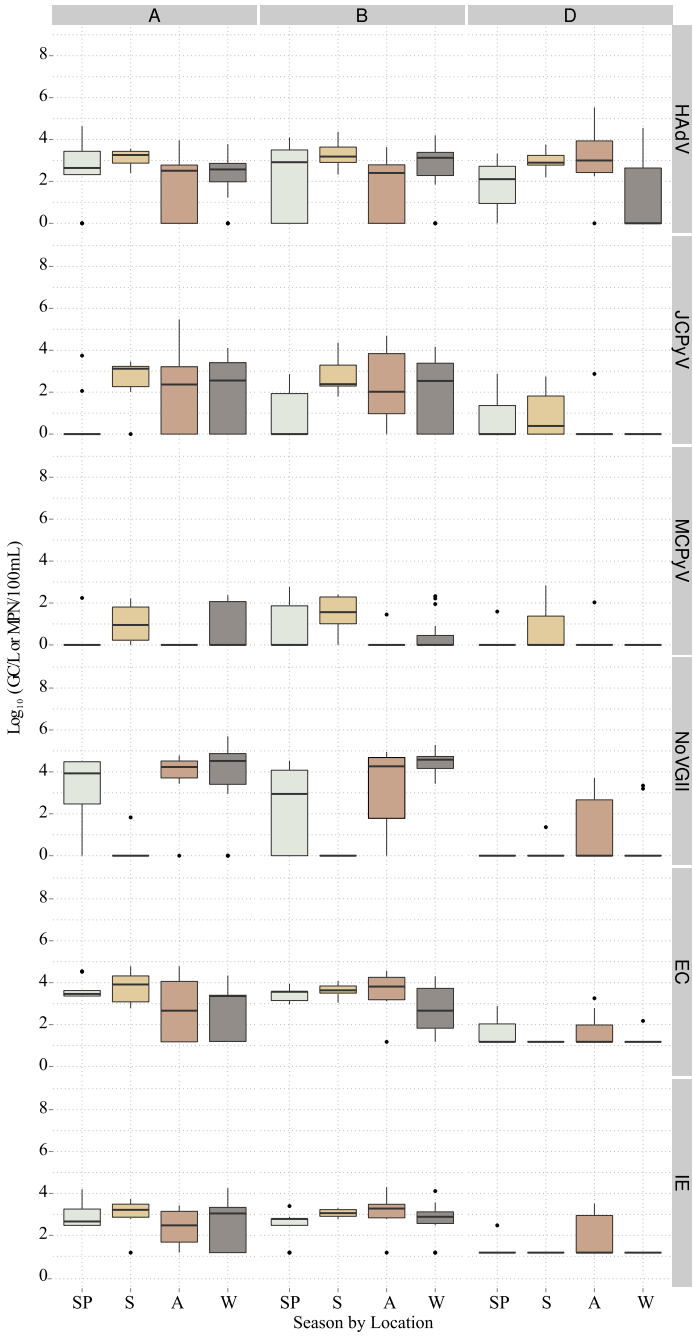


Fig. 2. Boxplots representing interquartile pathogen concentrations (Log_{10} GC/L or MPN/100 mL) by season (SP: spring, S: summer, A: autumn and W: winter) and matrices. Lines extending vertically from the boxes indicate variability outside the upper and lower quartiles, and the outlier values are plotted as individual points. A: upstream river site, B: dowstram river site and D: seawater site.

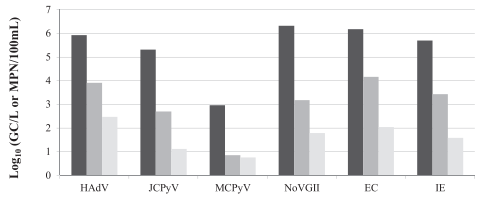
1.6×10^5 GC/L, respectively. Tertiary treated effluents had only 7 positive samples for JCPyV out of 22 tested samples. The high prevalence and concentrations of HAdV and JCPyV observed in raw sewage inputs to the treatment plant and treated effluents from the secondary and tertiary treatment stages were similar to published data (Bofill-Mas et al., 2000a,b; Fong et al., 2010; Hewitt et al., 2013; McQuaig et al., 2009; Rodriguez-Manzano et al., 2012).

MCPyV was detected at median concentrations of 1.6×10^4 GC/L in 75% of the raw wastewater. Although its prevalence in secondary and tertiary effluents was extremely variable, MCPyV was mostly present in summer and winter. The frequent detection of MCPyV in urban sewage samples (31/37) suggested a persistent excretion of this virus by the contributing human population. Prevalence and concentrations of MCPyV in sewage were similar to those previously reported (Bofill-Mas et al., 2010).

The highest NoV GGII mean values from the WWTP were observed in winter, with 2.5×10^8 GC/L in raw sewage, 2.1×10^7 GC/L in secondary treated effluent and 2.2×10^5 GC/L after tertiary treatment. Positive HEV samples were detected in 9 of the 91 samples collected from the influent raw sewage (5/37) and the secondary treatment effluent (4/32).

HAdV, JCPyV, NoV GGII and MCPyV were widely detected in raw sewage samples, whereas HEV was found in low percentages (13.5%). Previous studies reported HEV to be present in 43.5% of the raw sewage from Barcelona (Clemente-Casares et al., 2003); the low occurrence detected in this study could be related to a short excretion period or the small volume of sample analysed using the concentration method. Nevertheless, this is the first study that has analysed HEV in wastewater after concentrating the viral particles with the skimmed milk flocculation method. In the majority of the river water samples, the EC levels complied with the Spanish standards for irrigation of non-processed food (i.e., less than 100 colony-forming units (cfu) in 100 mL) (SpanishGovernment, 2007a).

The reduction of viral and bacterial concentrations was calculated for all treatment steps (Fig. 3). All tested pathogens and faecal indicators were detected in the samples collected to characterise different levels of treatment with exception of HEV, which was not detected in any tertiary treatment samples. Pathogen removal was quantified by $\log_{10}$ reduction throughout the treatment process, after the secondary treatment (C_2) and after the tertiary treatment (C_3). The results showed that the activated sludge process was the most important step for the removal of pathogens, with maximum reductions of 3.14 $\log_{10}$ for NoVGGII. Similar findings have been described in the literature (Rodriguez-Manzano et al., 2012). The combination of the two processes (C_2+C_3) resulted in 4.13 $\log_{10}$, 4.11 $\log_{10}$ and 4.52 $\log_{10}$ reduction values for EC, IE and NoV GGII, respectively. Human DNA viruses exhibited variable total removal efficiencies (2.22 $\log_{10}$ for MCPyV, 3.44 $\log_{10}$ for HAdV and 4.19 $\log_{10}$ for JCPyV). Despite achieving a maximum decay of 4-$\log_{10}$ when compared to the initial concentrations in raw sewage, viruses were also detected in the majority of treated final effluent, including those in compliance with bacterial indicator standards (SpanishGovernment, 2007b). Although infectivity should be highly reduced after tertiary treatment, previous studies detected infectious HAdV using immunofluorescence assays after UV disinfection at the same WWTP (Rodriguez-Manzano et al., 2012).



	HAdV	JCPyV	MCPyV	NoVGII	EC	IE
Secondary (C2)	2.00	2.61	2.11	3.14	2.01	2.27
Tertiary (C3)	1.44	1.58	0.11	1.38	2.12	1.84
TOTAL(C2+C3)	3.44	4.19	2.22	4.52	4.13	4.11

Fig. 3. Summary of WWTP log10 concentrations and log10 reduction values by treatment for each viral and bacterial pathogen.

Considering that the average viral load of secondary treatment effluent is 10^5 GC/L, we can assume that every year a total amount of 3×10^{16} potentially infectious viruses are discharged into the river from WWTPs. These have a direct impact on the presence of human viruses in the river water and in the seawater bathing area studied. Treating sewage with only a secondary or tertiary process will not prevent impairment of riverine and marine receiving waters or guarantee satisfactory quality of raw surface waters used for the drinking water supply. The extent to which the concentrations (as GC number) identified after disinfection treatment with UV reflect viable infective organisms is of course debatable. However, our results are consistent with studies reporting viral loading from sewage treatment plants (Simmons and Xagoraraki, 2011). These findings provide evidence of the inefficacy of current interventions in removing qPCR signals indicating viral pathogens. New legislation including viral quality standards may be required as viability and infectivity information become increasingly available, especially after disinfection treatments.

3.4. Correlations among pathogens, river flow, precipitation, and water and air temperatures

The calculated Pearson correlations between microbial parameters showed that $\log_{10}$ EC and $\log_{10}$ IE correlated ($r = 0.92$, $p \approx 0$) along the river basin locations and seasons. In wastewater, the $\log_{10}$ transformed viral data exhibited significant correlations between viruses ($p < 0.01$), but with lower correlation coefficients compared to the bacterial indicators (Table S3 in the supplementary information). The concentrations of NoV GGII correlated with both river and sea water and air temperatures, highlighting the seasonality of the viral infection ($r = -0.64$ for both river water temperature and sea water temperature; $p \approx 0$). HAdV, NoV GGII and FIB presented significant correlations with accumulated precipitation only in seawater samples (r = 0.75, 0.95, 0.75, and 0.78 for EC, IE, HAdV and NoV GGII, respectively). When taking all samples into account, HAdV significantly correlated with JCPyV ($p < 0.01$), indicating that HAdV may potentially be used to predict the concentration of JCPyV. Statistically significant but lower correlations were also found between HAdV and the other tested pathogens. HAdV was the best predictor of the majority of pathogen concentrations (correlation values ranged from 0.53 to 0.70). As recently noted by other authors (Payment and Locas, 2011), direct correlation between pathogens and indicators of faecal contamination becomes improbable in water bodies receiving sewage discharges because of different dilution, transportation and inactivation rates. Nevertheless, the high prevalence of HAdV in all type of water matrices, seasons and river locations strongly supports their applicability as a viral indicator with potential for indexing the fate of viral pathogens and faecal contamination.

3.5. Future trends in virus concentrations considering climate projections for the Mediterranean river catchment

The risk associated with waterborne viral infections would be

influenced by the exposure (concentration of virus in water) and stability of the viruses in water. Thus, the temporal pattern and population burden of several infectious diseases may shift and human exposure may differ under future scenarios (Morse et al., 2012). To study future trends in viral concentrations in the selected river catchment, we generated climate projections to 2100 with the suite of climate models contained in the CMIP5+ project database (see Methods). Due to the large uncertainty associated with detailed spatial resolution simulations, a grid of 5 × 5 degrees was preferred and applied. The results for the trend in the seasonality of temperature and precipitation are presented in Fig. 4 together with the observed values for the same region derived from the European Centre for Medium-Range Weather Forecasts (ECMWF) interim re-analysis (ERA) (http://data-portal.ecmwf.int/data/d/interim_daily/). A bias between current climate conditions compared to those simulated with the ensemble of climate models is clearly evident when compared to the observations (derived from the ERA interim reanalysis). While it is true that there may be a noticeable difference between station data and reanalysis fields for a small region, we thought it convenient to highlight these disparities to better define the scope of our study. Projections for regional precipitation yield an estimate of a slight increase in total amounts between the simulated winter months of 2070–2100 and 1980–2010 (on average approximately 1–3 mm/month) and persistent decreases between May and October (on average 2.5 mm/month) (Fig. 4A). While these differences might be deemed low, values rise considerably in relative terms when compared to current observations (a general increase throughout the year with localised maximum increases of approximately 30% in the main winter months). Temperature projections for the region are presented in Fig. 4B. In this case, systematic differences between observations and simulations for the common 1980–2010 interval are much lower than for the rainfall described above. The overall change in temperature between the two intervals (2070–2100 vs 1980–2010) indicates a general warming trend throughout the year, with an increase in the minimum winter temperatures that is positive when compared to observations. Increases are above 13% in summer months and approximately 11% in winter.

Globally, climate change is expected to shrink potable water availability. In the Mediterranean climate regions it is predicted to increase the intensity and/or frequency of floods and droughts. Faecal contamination by means of viral concentrations may increase, but at the same time environmental factors producing viral inactivation could also increase in intensity. For example, UVB radiation and the biotic activity indirectly associated with higher temperatures (Carratalà et al., 2013) could reduce viral viability, compensating for the increasing viral loads. Where climate change scenarios predict more frequent floods, the pollution events caused by sewage overflows will increase, posing a challenge for water managers. It is likely that the predicted reduction in the number of summer rainfall events in the Mediterranean climate regions will produce more frequent low river flows with greater proportions of treated effluents entering surface water bodies, possibly increasing viral pathogen concentrations in river water and the impacted beaches (Figueras and Borrego, 2010; Cann et al., 2013).

4. Conclusions

1. The concentration of HAdV is stable in raw sewage throughout the year, with mean values of 8.38×10^5 GC/L.
2. Secondary biological treatment reduces pathogen concentrations between 2.0 and 3.1 $\log_{10}$ (as GC/L). Because conventional WWTPs discharge secondary effluents into rivers, the Llobregat River is persistently impacted by human faecal pollution as evidenced by the presence of HAdV, JCPyV and FIB in river water samples.
3. Seasonal NoV GGII patterns are observed at all sampling sites including wastewater and environmental samples. From the most upstream sampling site to the seawater impacted by the river discharge, NoV GGII was detected during spring, autumn and especially in winter when more outbreaks are identified in the population.
4. A new quantitative PCR method has been developed for the detection of the emergent MCPyV related to Merkel cell

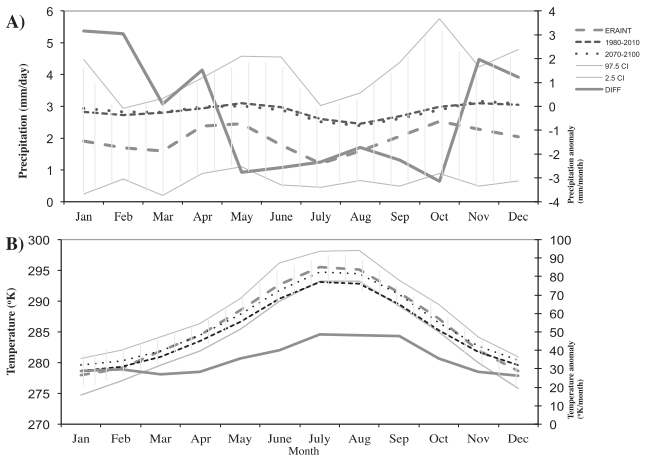


Fig. 4. A) Precipitation and B) mean temperature changes between the intervals 2070–2100 and 1980–2010 according to the suite of climate models contained in the rcp26 experiment of CMIP5+ (see Methods for details). 95% confidence interval is denoted by hatched area plots. DIFF denotes the difference in values for each variable obtained from 2100 to 2070 minus 1980–2010 and are referred to the secondary Y axis. ERA interim line yields values for the observational reanalysis of the region's temperature and precipitation values in the historical period (1980–2010). Temperature is in degrees Kelvin (°K) and precipitation in mm/day. Secondary Y axes have units referred to monthly totals.

carcinoma in water matrices. This is the first study reporting detection and quantification of MCPyV in seawater samples (with 18% prevalence and mean concentration of 1.18×10^2 GC/L). The high prevalence of MCPyV in sewage (75%) suggests that the human population commonly sheds this virus.

5. Temperature and precipitation predictions to 2100 for the selected Mediterranean river catchment suggest an increase in temperatures throughout the year and increased precipitation during winter with a decrease in precipitation between May and October. The highest virus concentrations are detected in summer when minimum precipitation and river flow occurs. The reduced dilution of treated effluents by rivers will result in elevated summer pathogen concentrations, with the potential for increased concentrations in winter resulting from intermittent combined sewage overflows discharging during the more frequent storm events.

Acknowledgements

The VIROCLIME Project is funded under the EU Seventh Framework Program, Contract No. 243923. The described study was supported by a collaborative European project coordinated by David Kay and Peter Wyn-Jones as vice-coordinator from the University of Aberystwyth, United Kingdom (VIROCLIME, contract no. 243923). We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modelling groups for producing and making available their model output. For CMIP the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals. During the development of the study Anna Carratalà and Natalia Timoneda were fellows of the Spanish Ministry of Science, Marta Rusiñol was a fellow of the Catalan Government "AGAUR" (FI-DGR) and Xavier Fernández was a fellow of the University of Barcelona (APIF). Finally, we would like to thank the EMSSA wastewater treatment plant for kindly providing the wastewater samples.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jenvman.2015.05.019.

References

ACA, 2012. Catalan Water Agency, Government of Catalonia [WWW Document]. URL. http://www.gencat.cat/aca/.

Albinana-Gimenez, N., Clemente-Casares, P., Calgua, B., Huguet, J.M., Courtois, S., Girones, R., 2009. Comparison of methods for concentrating human adenoviruses, polyomavirus JC and noroviruses in source waters and drinking water using quantitative PCR. J. Virol. Methods 158, 104–109.

Bofill-Mas, S., Pina, S., Girones, R., 2000a. Documenting the epidemiologic patterns of polyomaviruses in human populations by studying their presence in urban sewage. Appl. Environ. Microbiol. 66, 238–245.

Bofill-Mas, S., Pina, S., Girones, R., 2000b. Documenting the epidemiologic patterns of polyomaviruses in human populations by studying their presence in urban sewage. Appl. Environ. Microbiol. 66, 238–245.

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.

Bofill-Mas, S., Rusiñol, M., Fernandez-Cassi, X., Carratala, A., Hundesa, A., Girones, R., Carratalà, A., 2013. Quantification of human and animal viruses to differentiate the origin of the fecal contamination present in environmental samples. Biomed. Res. Int. 2013, 192089.

Bustin, S.A., Benes, V., Garson, J.A., Hellemans, J., Huggett, J., Kubista, M., Mueller, R., Nolan, T., Pfaffl, M.W., Shipley, G.L., Vandesompele, J., Wittwer, C.T., 2009. The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. Clin. Chem. 55, 611–622.

Calgua, B., Fumian, T., Rusiñol, M., Rodriguez-Manzano, J., Mbayed, V., Bofill-Mas, S., Miagostovich, M., Girones, R., 2013a. 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.

Calgua, B., Mengewein, A., Grunert, A., Bofill-Mas, S., Clemente-Casares, P., Hundesa, A., Wyn-Jones, a P., López-Pila, J.M., 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.

Calgua, B., Rodriguez-Manzano, J., Hundesa, A., Suñen, E., Calvo, M., Bofill-Mas, S., Girones, R., 2013b. New methods for the concentration of viruses from urban sewage using quantitative PCR. J. Virol. Methods 187, 215–221.

Cann, K.F., Thomas, D.R., Salmon, R.L., Wyn-Jones, A.P., Kay, D., 2013. Extreme water-related weather events and waterborne disease. Epidemiol. Infect. 141, 671–686.

Carratalà, A., Rusiñol, M., Rodriguez-Manzano, J., Guerrero-Latorre, L., Sommer, R., Girones, R., 2013. Environmental effectors on the inactivation of human adenoviruses in water. Food. Environ. Virol. http://dx.doi.org/10.1007/s12560-013-9123-3.

Clemente-Casares, P., Pina, S., Buti, M., Jardi, R., MartIn, M., Bofill-Mas, S., Girones, R., 2003. Hepatitis E virus epidemiology in industrialized countries. Emerg. Infect. Dis. 9, 448–454.

Collado, L., Kasimir, G., Perez, U., Bosch, A., Pinto, R., Saucedo, G., Huguet, J.M., Figueras, M.J., 2010. Occurrence and diversity of *Arcobacter spp.* along the Llobregat River catchment, at sewage effluents and in a drinking water treatment plant. Water Res. 44, 3696–3702.

Crabtree, K.D., Gerba, C.P., Rose, J.B., Haas, C.N., 1997. Waterborne adenovirus: a risk assessment. Water Sci. Technol. 35, 1–6.

Craun, G.F., Brunkard, J.M., Yoder, J.S., Roberts, V. a, Carpenter, J., Wade, T., Calderon, R.L., Roberts, J.M., Beach, M.J., Roy, S.L., 2010. Causes of outbreaks associated with drinking water in the United States from 1971 to 2006. Clin. Microbiol. Rev. 23, 507–528.

De RodaHusman, A.M., Lodder, W.J., Rutjes, S.A., Schijven, J.F., Teunis, P.F.M., 2009. Long-term inactivation study of three enteroviruses in artificial surface and groundwaters, using PCR and cell culture. Appl. Environ. Microbiol. 75, 1050–1057.

Donovan, T.J., Gallacher, S., Andrews, N.J., Greenwood, M.H., Graham, J., Russell, J.E., Roberts, D., Lee, R., 1998. Commun. Dis. Public Health 1, 188–196.

Erker, J.C., Desai, S.M., Mushahwar, I.K., 1999. Rapid detection of Hepatitis E virus RNA by reverse transcription-polymerase chain reaction using universal oligonucleotide primers. J. Virol. Methods 81, 109–113.

Figueras, M.J., Borrego, J.J., 2010. New perspectives in monitoring drinking water microbial quality. Int. J. Environ. Res. Public Health 7, 4179–4202.

Fong, T.-T., Phanikumar, M.S., Xagoraraki, I., Rose, J.B., 2010. Quantitative detection of human adenoviruses in wastewater and combined sewer overflows influencing a Michigan river. Appl. Environ. Microbiol. 76, 715–723.

Hawley, L.M., Garver, K.A., 2008. Stability of viral hemorrhagicsepticemia virus (VHSV) in freshwater and seawater at various temperatures. Dis. Aquat. Organ 82, 171–178.

Hernroth, B., Conden-Hansson, A., Rehnstam-Holm, A., Girones, R., Allard, A., 2002. Environmental factors influencing human viral pathogens and their potential indicator organisms in the blue mussel, Mytilusedulis: the first Scandinavian report. Appl. Environ. Microbiol. 68, 4523–4533.

Hewitt, J., Greening, G.E., Leonard, M., Lewis, G.D., 2013. Evaluation of human adenovirus and human polyomavirus as indicators of human sewage contamination in the aquatic environment. Water Res. 47, 6750–6761.

Imperiale, M.J., 2000. The human polyomaviruses, BKV and JCV: molecular pathogenesis of acute disease and potential role in cancer. Virology 267, 1–7.

IPCC, 2007. Climate Change 2007: an Assessment of the Intergovernmental Panel on Climate Change, pp. 12–17.

ISO 5667–9, 1992. Water Quality, Sampling, Part 9: Guidance on Sampling from Marine Waters.

Kageyama, T., Kojima, S., Shinohara, M., Uchida, K., Fukushi, S., Hoshino, F.B., Takeda, N., Katayama, K., 2003. Broadly reactive and highly sensitive assay for Norwalk-like viruses based on real-time quantitative reverse transcription-PCR. J. Clin. Microbiol. 41, 1548–1557.

Kitajima, M., Haramoto, E., Phanuwan, C., Katayama, H., Furumai, H., 2012. Molecular detection and genotyping of human noroviruses in influent and effluent water at a wastewater treatment plant in Japan. J. Appl. Microbiol. 112, 605–613.

Kitajima, M., Oka, T., Haramoto, E., Takeda, N., Katayama, K., Katayama, H., 2010. Seasonal distribution and genetic diversity of genogroups I, II, and IV noroviruses in the Tamagawa River. Jpn. Environ. Sci. Technol. 44, 7116–7122.

Kroneman, A., Verhoef, L., Harris, J., Vennema, H., Duizer, E., van Duynhoven, Y., Gray, J., Iturriza, M., Böttiger, B., Falkenhorst, G., Johnsen, C., von Bonsdorff, C.-H., Maunula, L., Kuusi, M., Pothier, P., Gallay, A., Schreier, E., Höhne, M., Koch, J., Szücs, G., Reuter, G., Krisztalovics, K., Lynch, M., McKeown, P., Foley, B., Coughlan, S., Ruggeri, F.M., Di Bartolo, I., Vainio, K., Isakbaeva, E., Poljsak-Prijatelj, M., Grom, A.H., Mijovski, J.Z., Bosch, A., Buesa, J., Fauquier, A.S., Hernandéz-Pezzi, G., Hedlund, K.-O., Koopmans, M., 2008. Analysis of integrated virological and epidemiological reports of norovirus outbreaks collected within the Foodborne Viruses in Europe network from 1 July 2001 to 30 June 2006. J. Clin. Microbiol. 46, 2959–2965.

Legrand-Abravanel, F., Mansuy, J.M., Dubois, M., Kamar, N., Peron, J.M., Rostaing, L., Izopet, J., 2009. Hepatitis E virus genotype 3 diversity, France. Emerg. Infect. Dis. 15, 110–114.

Loisy, F., Atmar, R.L., Guillon, P., Le Cann, P., Pommepuy, M., Le Guyader, F.S., 2005.

Real-time RT-PCR for norovirus screening in shellfish. J. Virol. Methods 123, 1–7.
Lopman, B., Armstrong, B., Atchison, C., Gray, J.J., 2009. Host, weather and virological factors drive norovirus epidemiology: time-series analysis of laboratory surveillance data in England and Wales. PLoS One 4, e6671.
Lopman, B., Vennema, H., Kohli, E., Pothier, P., Sanchez, A., Negredo, A., Buesa, J., Schreier, E., Reacher, M., Brown, D., Gray, J., Iturriza, M., Gallimore, C., Bottiger, B., Hedlund, K.-O., Torvén, M., von Bonsdorff, C.-H., Maunula, L., Poljsak-Prijatelj, M., Zimsek, J., Reuter, G., Szücs, G., Melegh, B., Svennson, L., van Duijnhoven, Y., Koopmans, M., 2004. Increase in viral gastroenteritis outbreaks in Europe and epidemic spread of new norovirus variant. Lancet 363, 682–688.
Mathijs, E., Stals, A., Baert, L., Botteldoorn, N., Denayer, S., Mauroy, A., Scipioni, A., Daube, G., Dierick, K., Herman, L., Van Coillie, E., Uyttendaele, M., Thiry, E., 2012. A review of known and hypothetical transmission routes for noroviruses. Food Environ. Virol. 4, 131–152.
McQuaig, S.M., Scott, T.M., Lukasik, J.O., Paul, J.H., Harwood, V.J., 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.
Meteocat, 2014. Meteocat, Generalitat de Catalunya [WWW Document]. http://www.meteo.cat/servmet/index.html (accessed 02.27.14.).
Morse, S.S., Mazet, J.A.K., Woolhouse, M., Parrish, C.R., Carroll, D., Karesh, W.B., Zambrana-Torrelio, C., Lipkin, W.I., Daszak, P., 2012. Prediction and prevention of the next pandemic zoonosis. Lancet 380, 1956–1965.
Nordgren, J., Matussek, A., Mattsson, A., Svensson, L., Lindgren, P.-E., 2009. Prevalence of norovirus and factors influencing virus concentrations during one year in a full-scale wastewater treatment plant. Water Res. 43, 1117–1125.
Orrù, G., Masia, G., Orrù, G., Romanò, L., Piras, V., Coppola, R.C., 2004. Detection and quantitation of hepatitis E virus in human faeces by real-time quantitative PCR. J. Virol. Methods 118 (2), 77–82.
Pal, A., Sirota, L., Maudru, T., Peden, K., Lewis, A.M., 2006. Real-time, quantitative PCR assays for the detection of virus-specific DNA in samples with mixed populations of polyomaviruses. J. Virol. Methods 135, 32–42.
Payment, P., Locas, A., 2011. Pathogens in water: value and limits of correlation with microbial indicators. Ground Water 49, 4–11.
Pérez-Sautu, U., Sano, D., Guix, S., Kasimir, G., Pintó, R.M., Bosch, A., 2012. Human norovirus occurrence and diversity in the Llobregat river catchment, Spain. Environ. Microbiol. 14, 494–502.
Pina, S., Puig, M., Lucena, F., Jofre, J., Girones, R., 1998. Viral pollution in the environment and in shellfish: human adenovirus detection by PCR as an index of human viruses. Appl. Environ. Microbiol. 64, 3376–3382.
Rodriguez-Manzano, J., Miagostovich, M., Hundesa, A., Clemente-Casares, P., Carratala, A., Buti, M., , et al.Girones, R., 2010. Analysis of the evolution in the circulation of HAV and HEV in eastern Spain by testing urban sewage samples. J. Water Health 8 (2), 346–354.
Rodriguez-Manzano, J., Alonso, J.L., Ferrús, M.A., Moreno, Y., Amorós, I., Calgua, B., Hundesa, A., Guerrero-Latorre, L., Carratala, A., Rusiñol, M., Girones, R., 2012. Standard and new faecal indicators and pathogens in sewage treatment plants, microbiological parameters for improving the control of reclaimed water. Water Sci. Technol. 66, 2517–2523.
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., 2013. 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.
Simmons, F.J., Xagoraraki, I., 2011. Release of infectious human enteric viruses by full-scale wastewater utilities. Water Res. 45, 3590–3598.
Sinclair, R.G., Jones, E.L., Gerba, C.P., 2009. Viruses in recreational water-borne disease outbreaks: a review. J. Appl. Microbiol. 107, 1769–1780.
SpanishGovernment, 2007a. Real Decreto 1620/2007. Régimen jurídico de la reutilización de las aguas regeneradas, 2007. Boletín Oficial del Estado, 294–21092.
SpanishGovernment, 2007b. Real Decreto 1341/2007. Gestión de la calidad de las aguas de baño. Boletín Oficial del Estado, vol. 257, p. 18581.
Spurgeon, M.E., Lambert, P.F., 2013. Merkel cell polyomavirus: a newly discovered human virus with oncogenic potential. Virology 435, 118–130.
Teunis, P.F.M., Moe, C.L., Liu, P., Miller, S.E., Lindesmith, L., Baric, R.S., Le Pendu, J., Calderon, R.L., 2008. Norwalk virus: how infectious is it? J. Med. Virol. 80, 1468–1476.
Verzani, J., 2004. simpleR – Using R for Introductory Statistics. the CSI Math department.
Wyn-jones, A.P., Carducci, A., Cook, N., D'Agostino, M., Divizia, M., Fleischer, J., Gantzer, C., Gawler, A., Girones, R., Höller, C., de RodaHusman, A.M., Kay, D., Kozyra, I., López-Pila, J., Muscillo, M., Nascimento, M.S.J., Papageorgiou, G., Rutjes, S., Sellwood, J., Szewzyk, R., Wyer, M., Agostino, M.D., Ho, C., Maria, A., Husman, D.R., Sa, M., Lo, J., 2010. Surveillance of adenoviruses and noroviruses in European recreational waters. Water Res. 5, 1025–1038.

Altres publicacions 2

Health Risks Derived from Consumption of Lettuces Irrigated with Tertiary Effluent Containing Norovirus

Helena Sales-Ortells Xavier Fernandez-Cassi, **Natàlia Timoneda**, Wiebke Dürig, Rosina Girones, Gertjan Medema.

Food Research International (2015) Feb 68; 70-77

Food Research International 68 (2015) 70–77

Contents lists available at ScienceDirect

Food Research International

journal homepage: www.elsevier.com/locate/foodres

Health risks derived from consumption of lettuces irrigated with tertiary effluent containing norovirus

Helena Sales-Ortells [a,b,*], Xavier Fernandez-Cassi [c], Natàlia Timoneda [c,d], Wiebke Dürig [a], Rosina Girones [c], Gertjan Medema [a,b]

[a] KWR Watercycle Research Institute, P.O. Box 1072, 3430 BB Nieuwegein, The Netherlands
[b] Section Sanitary Engineering, Department of Water Management, Faculty of Civil Engineering and Geosciences, Delft University of Technology, P.O. Box 5048, 2600 GA Delft, The Netherlands
[c] Department of Microbiology, Faculty of Biology, University of Barcelona, Av. Diagonal 643, Barcelona 08028, Spain
[d] Computational Genomics Laboratory, Department of Genetics, Institute of Biomedicine, University of Barcelona, Av. Diagonal 643, Barcelona 08028, Spain

ARTICLE INFO

Article history:
Received 22 May 2014
Received in revised form 6 August 2014
Accepted 14 August 2014
Available online 23 August 2014

Keywords:
Norovirus
Water reclamation
Crop irrigation
Health risks
Quantitative Microbial Risk Assessment

ABSTRACT

Wastewater is a valuable resource for water-scarce regions, and is becoming increasingly important due to the rising frequency of droughts as a result of climate change. The health risks derived from ingestion of lettuce that has been irrigated with effluent from a wastewater treatment plant (WWTP) in Catalonia (Spain) were estimated following a quantitative microbial risk assessment (QMRA) approach using site-specific data. Norovirus (NoV) was selected for this analysis, since it is the most common cause of acute gastroenteritis outbreaks in Catalonia. Two scenarios, irrigation with secondary and with tertiary effluent, were analysed. An uncertainty analysis was conducted to determine the impact of possible internalization of NoV into edible parts of the lettuce. The mean disease burden for ingestion of lettuce irrigated with secondary and tertiary effluent was 7.8×10^{-4} Disability Adjusted Life Years (DALYs) per person per year (pppy) and 3.9×10^{-4} DALYs pppy, respectively. A sensitivity analysis revealed that the model parameters with higher influence on the probability of disease are the concentration of NoV in the effluent and the consumption of lettuce. In order to decrease the disease burden to the guidance level of 10^{-6} DALYs pppy, the tertiary treatment should be able to achieve a 4.3 log reduction of the concentration of NoV. If internalization of NoV into lettuces occurs, this would require a reduction of 7.6 log. This is the first time that site specific data and virus internalization in crops are incorporated in a QMRA of irrigation of lettuce and its impact is quantified.

1. Introduction

Wastewater has been widely used in the past for irrigation purposes. It is still in use in developing countries due to water scarcity, the associated nutrient value of these waters for crop growth, and economic limitations. In developed countries, the use of treated wastewater is increasingly seen as a way to deal with water scarcity (exacerbated by climate change), as a more economical alternative to inter-basin transfers, and as an environmentally sustainable practice (Drechsel, Scott, Raschid-Sally, Redwood, & Bahri, 2010).

Uses of reclaimed water include irrigation of landscapes, recreational fields, plants' nurseries, or agricultural lands for food crops, amongst others. In Spain, 362.2 Hm^3 of reclaimed water (42.39 Hm^3 in Catalonia) are used annually, corresponding to 10.6% of the total volume of treated wastewater. 71% of it is used for agricultural irrigation (Iglesias, Ortega, Batanero, & Quintas, 2010).

Although domestic wastewater is treated by secondary or tertiary wastewater treatment, reclaimed water can contain infectious pathogens, posing a risk for public health. Wastewater treatment methodologies are used to reduce concentrations of faecal indicators, e.g. faecal coliforms (FC) or *Escherichia coli* (EC), to below certain standards (BOE, 2007). However, wastewater treatment can be considerably less effective in the elimination of enteric pathogens, such as enteric viruses (EV) and protozoa (Montemayor et al., 2008). Whilst concentrations of FC and EC are usually monitored at the wastewater treatment plants (WWTP), EV, which are relatively resistant to treatment technologies, are not (BOE, 2007), and concentrations of faecal indicators below the standards do not imply absence of EV hazards.

The health risks derived from irrigation of fresh produce with reclaimed water have been previously studied for EV (Hamilton, Stagnitti, Premier, Boland, & Hale, 2006; Petterson, Ashbolt, & Sharma, 2001, 2002; Seidu et al., 2008; Shuval, Lampert, & Fattal, 1997; Stine, Song, Choi, & Gerba, 2005). Few studies focused on the norovirus

* Corresponding author at: KWR Watercycle Research Institute, P.O. Box 1072, 3430 BB Nieuwegein, The Netherlands. Tel.: +31 30 6069 649.
E-mail addresses: hsalesortells@gmail.com (H. Sales-Ortells), xaviako@gmail.com (X. Fernandez-Cassi), Natalia.timoneda@gmail.com (N. Timoneda), wiebkedeurig@gmail.com (W. Dürig), rgirones@ub.edu (R. Girones), gertjan.medema@kwrwater.nl (G. Medema).

http://dx.doi.org/10.1016/j.foodres.2014.08.018
0963-9969/

(NoV) risks (Mara & Sleigh, 2010a,b; Mok, Barker, & Hamilton, 2014). Mara and Sleigh (2010a,b) found infection risks of NoV to range between 10^{-5} and 1 per person per year (pppy), depending on the initial concentration, and concluded that additional reduction of the NoV concentration in wastewater is needed, but easily achievable by water treatment. Mok et al. (2014) found, for an estimated concentration of 6.0×10^7 virus/L in raw sewage, a 90% Confidence Interval (CI) of 4.7×10^{-4} to 4.4×10^{-3} Disability Adjusted Life Years (DALY) pppy in lettuce irrigated with wastewater treated by stabilization ponds. Other wastewater treatment methods (Actiflo, chlorination, ozone or UV) did not reduce the disease burden below the WHO recommendation of 10^{-6} DALY pppy (WHO, 2006), but this reduction could be achieved by a combination of the stabilization pond with any of the other treatment technologies. Those studies, however, did not use site-specific data on NoV concentrations in reclaimed water, and only considered the viruses deposited on lettuce surface (and not internalization of viruses through the roots). Furthermore, all QMRA studies have used a model derived from *Bacteroides fragilis* bacteriophage B40-8 (Petterson et al., 2001, 2002) to estimate the NoV field-decay, whilst recent studies (Carratalà et al., 2013; Hirneisen & Kniel, 2013) have provided more specific data to estimate the inactivation of NoV, not only in-field, but also during crops transport and storage.

The objective of this study was to quantify the health risks of lettuce irrigation with treated domestic wastewater in Catalonia (Spain) and the effect of secondary versus tertiary wastewater treatment on these health risks. This study followed a Quantitative Microbial Risk Assessment (QMRA) approach and NoV was selected as reference pathogen, since it is the most common cause of acute gastroenteritis outbreaks in Catalonia (Martínez et al., 2013). Recent literature indicates the ability and extent of lettuces to internalize virus particles (Dicaprio, Ma, Purgianto, Hughes, & Li, 2012; Esseili, Wang, Zhang, & Saif, 2012; Wei, Jin, Sims, & Kniel, 2011). This is an important element that influences the outcome of the risk assessment and has not been considered in previous QMRA studies. We introduced this as an alternative scenario in the QMRA model and quantified the effect on the health risks.

2. Methods

2.1. Study-site description

The WWTP is located on the North-East coast of Spain and is designed to treat wastewater from 175,000 inhabitants with a flow capacity of 35,000 m^3/day. The conventional secondary treatment consists of sedimentation and activated sludge. The tertiary treatment, with a design capacity of 600 m^3/h, consists of flocculation by addition of iron chloride, followed by filtration (pulsed-bed sand filters), UV treatment (2 banks with 4 medium pressure lamps each, with a UV dose of 25–30 mJ/cm^2, according to the UV supplier) and chlorination (dosing of 3 to 6 mg/L of sodium hypochlorite with a contact time of 30 to 90 min). This tertiary effluent is used for irrigation and its production depends on the demand of the users, being higher from April to October, with a peak in July–August. Characteristics of the secondary and tertiary effluent measured by the WWTP system can be found in the supplementary material (S1).

The tertiary effluent is intended to irrigate several vegetable farms located in the vicinity of the WWTP. At the farms, different vegetables are irrigated through sprinkler, furrow, or drip irrigation. Most of them, however, with lettuce in particular, are irrigated with a sprinkler system every other day, in the evening. Lettuce is harvested, transported to the local market, and sold to the customers twice per week.

2.2. Hazard identification

NoV is a single stranded RNA virus that belongs to the Caliciviridae family (Lodder & de Roda Husman, 2005; Patel, Hall, Vinjé, & Parashar, 2009). They are found in water and food worldwide and are a leading cause of acute gastroenteritis (Mok et al., 2014), specially genogroups NoVGI and NoVGII (Rajko-Nenow et al., 2013). A study on epidemiological data from a 10-year period revealed that NoV was the most common cause of acute gastroenteritis outbreaks in Catalonia from 2004 to 2010 (Martínez et al., 2013).

Concentrations of NoV in wastewater range from 10^0 to 10^5 virus/L (Katayama et al., 2008; Lodder & de Roda Husman, 2005), with higher concentrations usually found in winter. In secondary effluents, NoV concentrations of the range of 10^1 to 10^3 have been found (Katayama et al., 2008). NoV has been linked with food outbreaks, including salad crops (Ethelberg et al., 2010; Gallimore et al., 2005; Makary et al., 2009; Rutjes, van den Berg, Lodder, & de Roda Husman, 2006; Wadl et al., 2010).

2.3. Exposure assessment

A conceptual exposure model was designed to describe the virus fate and transport from the secondary effluent to the consumers' fork (S2).

2.3.1. Norovirus concentration in effluent

Data on the concentration of NoV was obtained from the secondary ($n = 8$) and tertiary effluents ($n = 8$), and from a reservoir to store tertiary effluent ($n = 8$). Monthly samples were gathered for a period of 8 months. Detailed methodology is described in the supplementary material (S3). Briefly, viruses present in 10 L samples were concentrated using the skimmed milk organic flocculation method as described by Calgua, Fumian, et al. (2013). A negative control was included in each sampling event using tap water as matrix, and neutralizing the free chlorine by adding 100 mL of 10% sodium thiosulfate solution. Viral extraction of RNA from 140 μl of concentrates was done with the QIAamp® Viral RNA Mini Kit (Qiagen, Valencia, CA, USA) employing the automated system QIACube (Qiagen, Valencia, CA, USA) following the manufacturer's instructions. Extracts were stored at −80 °C until analysed. A negative control of extraction was included in each extraction batch using free DNAse/RNAse molecular water. Samples were tested using specific real-time RT-qPCR for the viral pathogens NoVGI (Loisy et al., 2005) and NoVGII (Kageyama et al., 2003). Duplicate alliquots of undiluted and log10 diluted extracts were analysed. More than one non-template control (NTC) were included in the RT-qPCRs. MX3000Pro sequence detector system (Stratagene, La Jolla, CA, USA) was used to quantify the samples. Detection limits are 10 genome copies (gc) per reaction tube (Kageyama et al., 2003), equivalent to 570 gc/L. Plasmid DNA was used as a positive control and as a quantitative standard. RT-qPCR standards were generated as described by Calgua, Rodriguez-Manzano, et al. (2013). Recovery of the method can be found in Calgua, Fumian, et al. (2013).

Although most of the NoV outbreaks and clinical cases in Catalonia are related to NoVGII, both genogroups were added up since NoVGI is also a human pathogen. An ANOVA test was run to check for significant differences between the NoV combined concentrations in the three sampling points. Gamma and lognormal distributions were fitted to the data using the maximum likelihood estimation and the method of matching moments. These distributions were used because they have shown before to give a good fit to pathogens concentrations in water (Tanaka, Asano, Schroeder, & Tchobanoglous, 1998; Westrell et al., 2006). Goodness of fit was analysed graphically and by the Kormogorov–Smirnov test.

2.3.2. Norovirus removal by tertiary treatment

The absence of a statistical difference between the concentration of NoV in secondary and tertiary effluent (see Results) indicated little removal by the coagulation and sand filtration in the tertiary treatment. Also UV and chlorination did not reduce the concentration, but this could at least partly be due to the fact that RT-qPCR detects both active and inactive (i.e. non-infective) viruses (Sobsey, Battigelli, Shin, & Newland, 1998). Concentrations of *EC* in secondary and tertiary effluent

indicate that the tertiary treatment is reducing the *EC* concentrations very significantly throughout the year (Table S1). To model the reduction of (infectious) NoV concentration by the tertiary treatment adequately, we used data on surrogate viruses to determine the virus elimination during the tertiary treatment processes. Studies have shown the disinfection efficiency of surrogates (MS2 virus, Feline Calicivirus, Sapovirus, etc.) through chlorine and UV treatment (Hijnen, Beerendonk, & Medema, 2006; Tree, Adams, & Lees, 1997, 2003). The efficiency of this particular WWTP was determined in a previous study, in which the reduction of the concentration of faecal indicator bacteria (EC, enterococci, etc.) and surrogate viruses (somatic, F-specific and *Bacteroides* phages) was determined for UV treatment, for chlorine disinfection, and for the combination of the two (Montemayor et al., 2008). The UV dose was 25 mJ/cm^2 (according to UV supplier) when used alone or combined with chlorine, and the UV_{254nm} transmittance of the secondary effluent was 46 +/− 5%. The chlorine dose was 10 ppm (alone), and 5 ppm (when combined with UV), and in combination with chlorine decay and contact times this obtained average Ct-values (concentration of disinfectant times the contact time) of 216 and 100 mgCl_2 min/L, respectively. Results of this study showed that, whilst chlorine (with or without UV) was effective for disinfection of EC and enterococci, it had very little effect on the reduction of F-RNA and other bacteriophages (Table 1). For the latter, UV treatment (with or without chlorine) managed 2 to 6 times higher inactivation than chlorination alone (Table 1) (Montemayor et al., 2008). The data on NoV concentration in the secondary effluent was combined with the data on inactivation of NoV by UV to estimate the concentration of infectious NoV in tertiary effluent. A PERT distribution was introduced, using the mean and 95% CI of the log reduction, to include the variability on the inactivation data.

2.3.3. Transfer of Norovirus to lettuce by irrigation

The crop fields are located in the immediate surroundings of the WWTP and do not store the tertiary effluent. Therefore, the time between tertiary effluent production and use is very short and it is assumed that no inactivation of viruses occurs in this period. Irrigation of lettuce is done with an overhead sprinkler system through which lettuce surfaces receive a considerable amount of water. Mok and Hamilton (2014) studied the volume of water that clings to lettuces after irrigation by such a system and found that it was best described by a lognormal3 (−4.75, 0.50, 0.006) distribution. We assumed that all viruses in the water that retained the lettuce after irrigation are and remain attached to the lettuce.

2.3.4. Virus internalization

Several studies have indicated that enteric viruses can be internalized into crops. This can occur through the roots, where the viruses are transported via the lettuce vascular system to the leaves, or though the stomata and wounds present on the leaves (Hirneisen, Sharma, & Kniel, 2012; Warriner, Ibrahim, Dickinson, Wright, & Waites, 2003). Few manuscripts have been published on the ability of NoV and surrogates to be internalized by lettuce (Dicaprio et al., 2012; Esseili, Wang, Zhang, et al., 2012; Wei et al., 2011). Only one study used human NoV and obtained a high proportion of internalization in the leaves (0.24 to 0.72 virus/g) of lettuce grown in hydroponic solution (Dicaprio et al., 2012). The internalization rate of Murine norovirus (MNV) into edible parts of lettuce ranged from 4×10^{-6} to 2×10^{-3} virus/g (Wei et al., 2011) and of Sapovirus was 1×10^{-7} to 5×10^{-7} virus/g (Esseili, Wang, Zhang, et al., 2012). Different laboratory conditions (growth substrate, relative humidity (RH), initial virus titer, etc.) were associated with the differences in internalization (Esseili, Wang, Zhang, et al., 2012). No quantitative information has been found in the literature on internalization through the surface of lettuce leaves.

Table 1
Mean (95% CI) of log10 reduction of *E. coli* and F-RNA bacteriophages by tertiary treatment processes (extracted from Montemayor et al., 2008).

Disinfection method	EC	F-RNA bacteriophages
UV	1.80 (1.52–2.10)	0.94 (0.57–1.30)
Cl	5.00 (4.82–5.22)	0.30 (0–0.66)
UV + Cl	5.05 (4.82–5.40)	0.85 (0.63–0.93)

2.3.5. Virus inactivation in the field and during storage and transport

Sunlight and high temperatures influence virus inactivation in the field. During storage and transport, inactivation might also happen, but at a slower rate because of lower temperatures and absence of sunlight. Recently, studies have investigated the persistence of surrogates for human NoV on crop surfaces. MS2 virus on lettuce surface was inactivated by almost 3 logs after 25 h at 30 °C and exposed to artificial sunlight. In the dark at 30 °C, the decrease after 25 h was maximum 1 log, whilst at 4 °C in the dark no inactivation was found (Carratalà et al., 2013). Hirneisen and Kniel (2013) found no difference in survival between MNV, Tulane virus (TV) (with tissue culture and qPCR) and NoVGII (with RT-qPCR) on spinach surface. The decimal reduction times (D), i.e. the time needed for 1 log reduction in virus titer, for MNV and TV, ranged from 1.40 ± 0.14 to 5.73 ± 2.41 days at 18 °C, depending on the spinach leaf type, the inoculation location (abaxial, adaxial or whole plant) and the presence of UV-A and B light.

All these data suggest that viruses located on the surface of lettuces, which are exposed to sunlight and high temperatures, will be inactivated by 1 to 2 logs in the period between the last irrigation and harvesting at our study site: 12 to 36 h total, with 6 to 18 h of sunlight. During dark hours, inactivation ranges from 0 to 1 log. Internalized viruses are only affected by high temperatures; therefore, their inactivation during the field period is assumed to be the same as that of the virus on the lettuce surface during dark hours, hence 0 to 1 log for the 12 to 36 h. During transport and storage (before selling, in the market, and in the households), lettuce is not exposed to sunlight, but is to different temperatures. Times between harvest and consumption are estimated to range from 13 to 55 h. Therefore, the reduction during this period is also between 0 and 1 log. Since no period appears to be most probable, a uniform distribution was used to define the degree of inactivation between last irrigation and consumption.

As an alternative scenario, the model from Petterson et al. (2001, 2002) from *B. fragilis* bacteriophage B40-8 has been used for inactivation in the field (only) as in previous studies, to quantify this uncertainty.

2.3.6. Virus removal by washing

Viruses on the plant surface will be partially removed by washing practices. Since usually no disinfection products are used for washing salad crops in Spain, a log reduction of the viruses on the surface is defined by PERT (0.1, 1, 2), based on Mok et al. (2014).

2.3.7. Consumption of lettuce

A survey of food consumption was conducted in Spain during 2009 and 2010 (AESAN, 2011). 3000 people covering both sexes, different age ranges, geographic regions, and urban settlements were interviewed. The retrospective intake questionnaire consisted on a dietary history (three days), a 24 h recall, and a food-frequency survey. The Spanish Agency (AESAN) provided the daily lettuce ingestion data in percentiles, average and standard deviation (personal communication). Total population consumed an average of 20.7 (±26.4) g per person per day (pppd), with 95% UCL of 74.2 g pppd. Average and standard deviation were used to construct several distributions, and the lognormal distribution was chosen because it resulted in percentile values closest to the survey data (mean = 19.4, 95% UCL = 87.2 g pppd).

2.3.8. Dose

The daily dose of virus on lettuce surface (d_S) ingested by the consumers of the market where the reclaimed water irrigated lettuce is sold was calculated as shown by Eq. (1):

$$d_s = 10^{\left(\log10\left(C_{eff}\times V_{surf}\right)-R_S-R_T-R_{wash}\right)} \times I \quad (1)$$

where *Ceff* is the concentration of NoV in secondary or tertiary effluent, *Rs* is the reduction of virus on the surface due to exposure to UV and high temperatures in the field, R_T is the reduction of viruses achieved during the lapsed time between harvest and consumption, *Rwash* is the reduction of surface viruses due to washing with water, and *I* the lettuce ingestion. The exposure model inputs are summarized in Table 2.

Uniform distributions were combined with the internalization ratios found in the literature (Dicaprio et al., 2012; Esseili, Wang, Zhang, et al., 2012; Wei et al., 2011) to define three different internalization scenarios with high (0.24–0.72), medium (4×10^{-6} to 2×10^{-3}) and low (1×10^{-7} to 5×10^{-7}) internalization ratios. In all three scenarios, the in-field reduction of internalized viruses (R_I) was considered uniform (0, 1) log 10 units (Carratalà et al., 2013; Hirneisen & Kniel, 2013). The dose of internalized virus (d_I) is calculated with Eq. (2).

$$d_I = 10^{\left(\log10\left(C_{eff}\times F_{int}\right)-R_I-R_T\right)} \times I \quad (2)$$

where *Fint* is the internalized fraction of viruses in the irrigation water that is found in the leaves (in viruses/g of lettuce).

2.4. Dose–response

In order to estimate the individual risk of NoV infection per event, the Beta-Poisson model defined by (Teunis et al., 2008) was used (Eq. (3))

$$P_d = 1-{}_1F_1(\alpha,\alpha+\beta,-d) \quad (3)$$

where ${}_1F_1$ is the Kummer confluent hypergeometric function, α and β are shape parameters with values 0.04 and 0.055, respectively, and *d* is the dose (either d_S or $d_S + d_I$). The illness given infection risk (Pdill) is calculated by multiplying the infection risk by an illness given infection ratio, which is, for NoV, 0.67 (Atmar et al., 2014).

2.5. Risk characterization

Farmers irrigate the crops with tertiary effluent during warm months, and not in winter. Therefore, frequency of ingestion of reclaimed water irrigated lettuce (*f*) happens from April to October, i.e. 214 days per year, assuming all the lettuce consumed comes from the street market. The annual probability is estimated using Eq. (4) (Haas, Rose, & Gerba, 1999).

$$Py = 1-(1-Pd)^f \quad (4)$$

Annual disease burden was calculated using DALYs. DALYs account for the years lived with disability (YLD) (Eq. (5)) plus the years of life lost (YLL) (Eq. (6)) due to the hazard, as compared to the average expected age of death in a community.

$$YLD = Pyill \times Dw \times Dt \quad (5)$$

$$YLL = Pyill \times Nd \times L \quad (6)$$

where *Dw* is the disability weight, *Dt* is the duration of illness, *Nd* is the number of deaths per illness and *L* is the average years lost per fatality. The *Dw* for acute gastroenteritis is 0.0007 for cases who do not visit the general practitioner (GP), which constitute 83.1% of NoV disease cases in Catalonia and 0.0062 for patients who do, which are 16.9% in Catalonia (Domínguez et al., 2008; Kemmeren, Mangen, Duynhoven, & Havelaar, 2006). The *Dt* was estimated, in The Netherlands, from 3 to 6 days, with average 3.8 days, in people who did not visit the GP, and from 5.73 to 7.23 in those who did (Kemmeren et al., 2006). We introduced this variability by defining a lognormal distribution (mean = log10(3.8), sd = 0.1) and uniform (5.73–7.23) days, for non-visiting and visiting GP respectively. No information on NoV mortality in Spain has been found. In The Netherlands, the annual mortality is 0.009% of NoV cases and the years lost due to premature death was estimated to be 20.7 (Havelaar et al., 2012).

Risks were calculated using Monte Carlo simulations with random sampling of 10,000 values from each distribution input. The Anderson–Darling test was run to see if the differences between the base scenario and the internalization scenarios were significant. Sensitivity analysis was conducted to know how sensitive the model outputs are to the inputs. The effect of the value of a model parameter on the probability of illness was calculated by varying a model parameter to the 95% CI limits of its variability, whilst keeping the variability of the other parameters. The effect of tertiary treatment on the DALYs was checked by running the model every time with a different log reduction of the NoV concentration. Monte Carlo simulations, and uncertainty and sensitivity analysis, were performed using R version 3.0.1 (The R Foundation for Statistical Computing, 2013).

3. Results

3.1. Concentration of norovirus in secondary and tertiary effluent

NoVGI was found in 71% and NoVGII in 100% of the samples. Concentrations of NoVGI and NoVGII were summed up because both

Table 2
Exposure assessment inputs, units, distribution and parameter values, and references.

Model inputs	Units	Distribution (parameter values)[a]	Literature references
Cseff: concentration of NV in secondary effluent	Genome copies/L	Gamma(0.3, 1.2×10^{-6})	This manuscript
Cteff: concentration of NV in tertiary effluent	Genome copies/L	Gamma(0.3, 1.2×10^{-6})	This manuscript
Rw: log reduction due to tertiary treatment (alternative scenario)	Log10 units	PERT(0.57, 0.94, 1.30)	Montemayor et al. (2008)
Vsurf: water that clings to lettuce surface through sprinkler irrigation	mL/g	Lognormal3 (−4.57, 0.50, 0.006)	Mok & Hamilton (2014)
Rs: in-field reduction of surface viruses	Log10 units	Uniform (1, 2)	Carratalà et al. (2013)
Rt: reduction of viruses during transport and storage	Log10 units	Uniform (0, 1)	Carratalà et al. (2013)
Rwash: reduction of surface viruses due to washing	Log10 units	PERT (0.1, 1, 2)	Mok et al. (2014)
I: daily consumption of lettuce	g pppd	Lognormal (20.72, 26.35, inf = 0, sup = 120)	AESAN (2011)

[a] Second parameter of the gamma distribution is the rate.

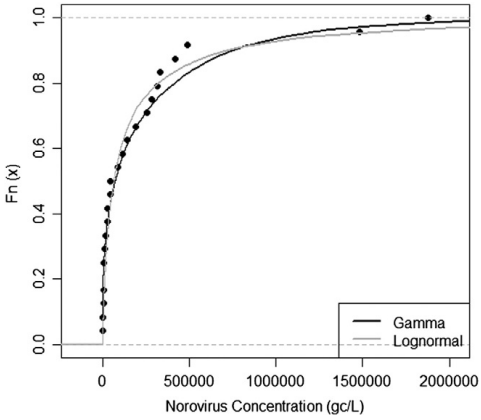


Fig. 1. Empirical cumulative distribution function (ecdf) graphs of the NoV data and the gamma and lognormal distribution fitted to the data. Parameters of the gamma distribution are shape = 3.2×10^{-1}, rate = 1.2×10^{-6} (in gc/L). Parameters of the lognormal distribution are meanlog = 11.1, sdlog = 1.9 (in gc/L).

genogroups are able to infect humans. Concentrations in secondary effluent ranged from 2.0×10^4 to 1.9×10^6, in tertiary effluent from 4.4×10^3 to 1.5×10^6, and in the reservoir from 1.8×10^3 to 3.1×10^5 gc/L. The NoV combined concentrations from the three sampling points was pooled after an ANOVA analysis showed no significant differences between the log transformed data of the three locations. Both distributions (lognormal and gamma) gave a good fit to the pooled data and the gamma distribution was selected because it has previously described concentrations of NoV in water (Westrell et al., 2006). Fig. 1 shows the concentration of NoV (NoVGI and NoVGII combined) and the fitted curves.

3.2. Impact of tertiary treatment on burden of disease

Virus concentration at the main steps of the exposure assessment, dose, and risk estimates is shown in Table 3. The mean individual probability of developing gastroenteritis after eating lettuce irrigated with secondary and tertiary water containing NoV was 2.3×10^{-2} and 5.2×10^{-3} pppd, respectively. The mean annual disease burden was 7.8×10^{-4} and 3.9×10^{-4} DALYs pppy. Due to the limited efficiency of virus removal by the tertiary treatment, the disease burden reduction by the use of tertiary effluent as compared to the use of secondary effluent is very limited.

3.3. Impact of internalization

Internalization scenarios when using internalization rates derived from Dicaprio et al. (2012) (high internalization rate) and Wei et al. (2011) (medium internalization rate) resulted in higher disease burden, but not when using data from Esseili, Wang, Zhang, et al. (2012) (low internalization rate) (Fig. 2). In the first two cases, the concentration of internalized viruses (49.3 and 0.03 gc/g, respectively) was higher than that on the lettuce surface (0.02 gc/g), and was, therefore, driving the probability of disease. In the latter case, the concentration inside lettuce leaves was much lower (3.1×10^{-5} gc/g) and, therefore, the concentration on the lettuce surface was responsible for the probability of disease. The Anderson–Darling test showed that the disease burden of the base scenario was statistically different from all the internalization scenarios (p-value < 0.0001), except from the low internalization scenario (p-value = 0.8).

3.4. Sensitivity analysis

The sensitivity analysis showed that the concentration of virus in the tertiary effluent and the consumption of lettuce were major factors influencing the variability of the risks (Fig. 3). Washing the lettuce, in-field inactivation, inactivation during transport and storage, and the volume of water clinging on the lettuce surface had little effect on the variability of the probability of disease.

An alternative scenario was built using the virus decay of Petterson et al. (2001), as done in other NoV QMRA in crops (Barker et al., 2013; Mok et al., 2014). This resulted in lower inactivation of NoV and, hence, higher disease burden (mean of 1.4×10^{-3} DALYs pppy), than when using NoV and surrogate inactivation data under several temperature and sunlight conditions (S4).

The disease burden was plotted against the efficiency of the tertiary treatment to show the impact of additional virus removal by the tertiary treatment on the burden of disease, both for the scenario without internalization and the scenario with the highest internalization (Fig. 4). The shoulder in this figure is the result of the high NoV doses and the dose–response function. The graph shows that 4.3 log reduction of the NV concentration (approx.) is required to achieve a burden of disease of 10^{-6} DALYs pppy, if no internalization of viruses into lettuce is considered, and 7.6 log reduction (approx.) if internalization as described by Dicaprio et al. (2012) is incorporated. The graph also shows that the reduction currently achieved by the tertiary treatment plant is not enough to reduce the DALYs under the WHO guideline value, even when no internalization is happening.

4. Discussion

The health risks associated with consumption of lettuce irrigated with secondary and tertiary treated effluent containing NoV in the north-east of Spain were estimated, based on NoV data collected at this site. Although the tertiary treatment was efficient enough to reduce the concentration of EC below the regulatory threshold for reclaimed water uses for crops irrigation (Table S1 and BOE, 2007), additional removal is needed in the system in order to meet the 10^{-6} DALYs pppy recommended by the WHO. If this is to be achieved by the tertiary treatment alone, then a 4.3 decimal logarithmic reduction of the NoV concentration must be ensured.

The concentrations of NoV in secondary and tertiary effluent were not significantly different. Concentrations were measured with RT-qPCR, detecting, specifically, NoVGI and NoVGII genes. No infectivity

Table 3
Mean (95% UCL) of theQMRA results for irrigation of lettuces with secondary and tertiary effluent.

Results	Units	Irrigation with secondary effluent	Irrigation with tertiary effluent
Concentration in water	Virus/mL	263.1 (1169)	31.7 (139.3)
Concentration on lettuce after irrigation	Virus/g	4.66 (20.3)	0.56 (2.47)
Concentration on lettuce at consumption	Virus/g	0.01 (0.04)	0.001 (0.005)
Dose	Pppd	0.19 (0.73)	0.02 (0.09)
Pd illness	Pppd	0.02 (0.15)	0.005 (0.02)
Py illness	Pppy	0.45 (1)	0.24 (0.99)
Disease burden	DALYs/year	7.8×10^{-4} (1.9×10^{-3})	3.9×10^{-4} (1.9×10^{-3})

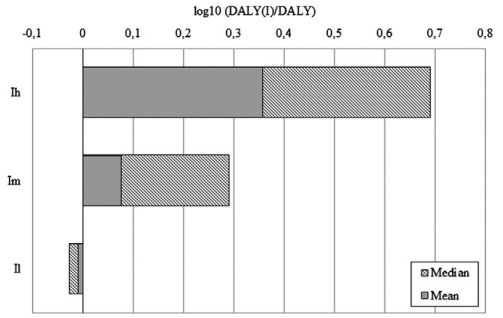


Fig. 2. Impact of the internalization scenarios on the burden of disease. Ih, Im and Il stand for high, medium and low internalization rates, respectively.

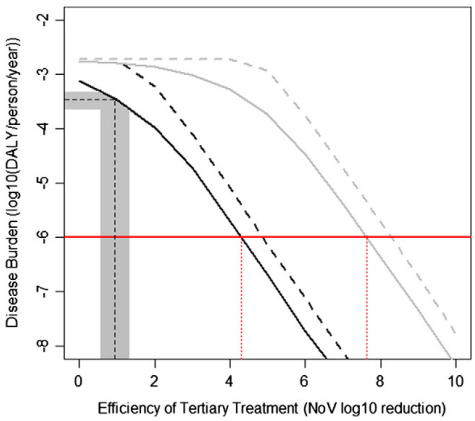


Fig. 4. Annual disease burden plotted against the efficiency of the tertiary treatment. The black solid line is the mean disease burden considering only surface deposition of virus on lettuce, with the dashed black line representing the 95% UCL. The grey solid line is the mean disease burden when internalization of viruses into lettuce (using DiCaprio data) is included, with the dashed grey line representing the 95% UCL. Red horizontal line is the recommended risk threshold of 10^{-6} DALY/person/year (WHO), with dotted vertical lines showing the needed treatment efficiency to achieve the threshold disease burden (4.3 and 7.6 decimal log reduction for surface only and internalized virus in lettuces, respectively). Dashed black vertical line corresponds to the known efficiency of the treatment plant and the corresponding disease burden (when virus is not internalized in lettuces), the grey area being the 95% CI.

assays are currently available for NoV due to the lack of specific cell culture lines. The infectivity of NoV in secondary effluent is, therefore, not known. However, the NoV dose–response model was derived from a study where a solution of Norovirus with unknown infective particles was administered to human volunteers. Since the technologies used for secondary treatment might remove virus particles but do not result in further inactivation, we consider the ratio of genomic copies to infectious NoV particles in secondary effluent comparable to the ratio in the (aged) samples used for the dose response studies.

However, during tertiary treatment it is likely that NoV is affected by the UV and chlorination. We used site-specific data on bacteriophage inactivation to estimate NoV inactivation. Montemayor et al. (2008) used three different disinfection methods (UV, chlorine, and a combination of the two) to study the reduction efficiency of the WWTP for different indicators. When using chlorine alone, the dose applied was 10 ppm, which is higher than the one used in the studied tertiary treatment. This, however, did not result in a high inactivation of bacteriophages, whilst it reduced the concentration of EC by 5 logs. UV yielded higher inactivation of bacteriophages, either when used alone or in combination with chlorine doses of 5 ppm. Unfortunately, no validated UV dose was given. The WWTP uses a combination of UV and chlorine, at doses between 3 and 6 ppm, although sometimes the UV treatment is bypassed, because the chlorine treatment is effective to reduce the concentration of EC to below the legal requirements. Therefore, tertiary treatment without UV may occasionally occur. This may result in periods of higher risk, given that the chlorination was not very effective against bacteriophages.

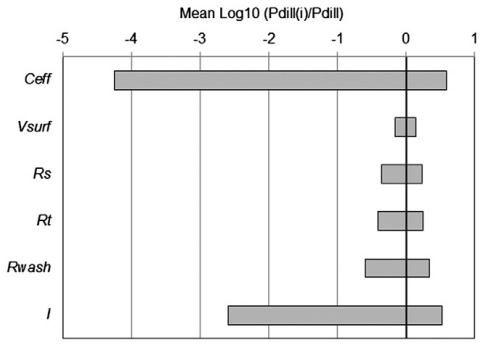


Fig. 3. Sensitivity of Pdill by varying each parameter to its extreme values.

Recently, the study of virus internalization into vegetables has experienced increasing attention. However, different studies show very different outputs for internalization of NoV and its surrogates into edible parts of lettuces (Dicaprio et al., 2012; Esseili, Wang, Zhang, et al., 2012; Wei et al., 2011). Sources of this variability can be the growth media (soil vs hydroponic solution), the RH of the environment, the applied virus dose, and species of lettuce and virus used in the experiments (Hirneisen et al., 2012; Wei et al., 2011). Overall, results show that lettuces are able to internalize NoV and surrogates, and these can reach edible parts of the crops, under laboratory conditions. It is not so clear to what extent this happens under field-conditions, with different kind of soils, lower soil saturation, lower concentration of NoV in irrigation water, and different climatic conditions influencing internalization. Specifically, the use of lower concentrations of viruses in irrigation water compared to the ones used in laboratory studies could result in internalized concentrations below the LOD of the methods used. This risk assessment showed that virus internalization into lettuces can have a large impact on the risk estimates (if high and medium internalization rates are considered) or no influence at all (if low internalization rates are considered), demonstrating the need for further research on the ability of lettuce on internalizing NoV under field conditions. This is the first time that virus internalization into crops has been incorporated in a QMRA study.

The amount of virus attached to the lettuce surface through overhead sprinkler irrigation was estimated by Mok (Mok & Hamilton, 2014). This is a conservative approach because it was assumed that all pathogens in the wastewater captured on the lettuce attach to its surface, and might lead to an overestimation of the risks. However, NoV have been shown to bind specifically to the carbohydrates of the cell wall of lettuce leaves surface (Esseili, Wang, & Saif, 2012).Virus attachment and survival differ on virus type, plants properties, and weather conditions (Hirneisen & Kniel, 2013; Vega, Smith, Garland, Matos, & Pillaii, 2005). Furthermore, differences might exist between the experimental conditions of the water retention study, and our field conditions.

For instance, the volume of water used for irrigation, the position of the overhead sprinklers (and distance from the irrigated vegetables and height), or environmental conditions (wind direction and speed, temperature, etc.).

Other QMRA studies have estimated the in-field decay of NoV on lettuce surface with the *B. fragilis* bacteriophage B40-8 model and assumed post-harvest decay as negligible (Barker et al., 2013; Mok et al., 2014). This is a conservative assumption because the bacteriophage is very resistant to environmental conditions (Petterson et al., 2001) and because viruses can undergo post-harvest degradation, depending on temperature conditions and time (Carratalà et al., 2013). In an attempt to use a more specific approach, we have used data derived from studies on the decay of NoV surrogates (Carratalà et al., 2013; Hirneisen & Kniel, 2013). Although human NoV data is not available, we believe that this is more appropriate, since MS2, TV, and MNV have been shown to be good surrogates (Bae & Schwab, 2008; Hirneisen & Kniel, 2013) and because virus inactivation is dependent on different temperature and solar radiation conditions (Carratalà et al., 2013). However, specific NoV data would provide the best assessment, but would require human volunteer studies (Richards, 2012).

Other approaches to understand NoV inactivation are being studied, such as the combination of enzymatic treatment with real-time nucleic acid sequence-based amplification (Lamhoujeb, Fliss, Ngazoa, & Jean, 2008), the combination of RT-qPCR with RNAse treatment (Mormann, Dabisch, & Becker, 2010; Topping et al., 2009), or the quantitative evaluation of oxidative damages on viral capsid protein (Sano, Pintó, Omura, & Bosch, 2009). However, more research is needed in this field, for instance to know if these methods are able to identify loss of infectivity due to different causes (heat, UV, chlorine, pH, etc.) to the same extent.

Our results show that the annual disease burden of consuming lettuce irrigated with reclaimed water exceeds the recommended threshold of 10^{-6} DALYs pppy. The risks are even higher when internalization is considered. Further measures should be applied in the system to reduce the virus load in the irrigation water, prevent lettuce contamination, or inactivate or remove the virus after contamination. To reduce the virus load in reclaimed water, improving of the UV system should be considered, e.g. adding a pre-treatment step that increases the transmittance of the water, as done in other WWTP (Montemayor et al., 2008). Measures to prevent contamination include using an irrigation method that results in lower surface deposition (such as subsurface drip irrigation), although this would not reduce the virus internalization. To inactivate the viruses after contamination, farmers could be advised to irrigate with a different source water on the last irrigation event, or increase the time between the last irrigation and the harvest, increasing the inactivation of viruses already deposited on the surface.

5. Conclusions

We conducted a stochastic QMRA to quantify the disease burden of NoV through ingestion of lettuce that has been irrigated with reclaimed water. This is the first time that site-specific data on human NoVGI and NoVGII in sewage effluent were used in a QMRA of NoV on lettuce. Decay data of NoV and surrogates has been used to describe the virus inactivation in the field and during transport and storage of lettuce, in contrast with the more conservative commonly used decay model derived from Bacteriophage B40-8.

- The recently discovered internalization of viruses in crops can have a significant impact on the disease burden if internalization occurs in the field. More research is needed to better understand and quantify virus internalization into lettuce under field conditions.
- Although the tertiary effluent of the target WWTP meets the EC requirements of national guidelines, additional barriers (either in the treatment or in irrigation practices) would be needed to meet the WHO recommendation for gastrointestinal disease burden.

Acknowledgements

This work was supported by KWR Innovation Fund (grant number 400414/001/002) and by the project AGL2011-30461-C02-01/ALI, funded by the Spanish Ministry of Science and Innovation. Xavier Fernandez-Cassi is a fellow of the Catalan Government "AGAUR" (FI-DGR) (grant number 2014FI_B1 00086) and Natalia Timoneda is a fellow of the Spanish Ministry of Science and Innovation (grant number BES-2012-053084). The authors would like to thank the WWTP manager and one of the farmers for all the information and help provided during the field visits.

Appendix A. Supplementary data

More information on the secondary and tertiary effluent, the conceptual exposure model, the methodology for norovirus concentration determination, and the virus decay uncertainty, can be found in the supplementary material to this manuscript. Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.foodres.2014.08.018.

References

AESAN (2011). *Encuesta Nacional de Ingesta Dietética Española 2009–2010 (ENIDE).* Agencia Española de Seguridad Alimentaria y Nutrición (AESAN).

Atmar, R. L., Opekun, A. R., Gilger, M. A., Estes, M. K., Crawford, S. E., Neill, F. H., et al. (2014). Determination of the 50% human infectious dose for Norwalk virus. *Journal of Infectious Diseases, 209*(7), 1016–1022.

Bae, J., & Schwab, K. J. (2008). Evaluation of murine norovirus, feline calicivirus, poliovirus, and MS2 as surrogates for human norovirus in a model of viral persistence in surface water and groundwater. *Applied and Environmental Microbiology, 74*(2), 477–484.

Barker, S. F., O'Toole, J., Sinclair, M. I., Leder, K., Malawaraarachchi, M., & Hamilton, A. J. (2013). A probabilistic model of norovirus disease burden associated with greywater irrigation of home-produced lettuce in Melbourne, Australia. *Water Research, 47*(3), 1421–1432.

BOE (2007). *REAL DECRETO 1620/2007, de 7 de diciembre, por el que se establece el régimen jurídico de la reutilización de las aguas depuradas.* (1620/2007).

Calgua, B., Fumian, T., Rusiñol, M., Rodriguez-Manzano, J., Bofill, S., Miagostovich, M., et al. (2013). Detection and quantification of classical and emerging viruses by skimmed-milk flocculation and PCR in river water from two geografical areas. *Water Research, 47*(8), 2797–2810.

Calgua, B., Rodriguez-Manzano, J., Hundesa, A., Suñen, E., Calvo, M., Bofill-Mas, S., et al. (2013). New methods for the concentration of viruses from urban sewage using quantitative PCR. *Journal of Virological Methods, 187*(2), 215–221.

Carratalà, A., Rodriguez-Manzano, J., Hundesa, A., Rusiñol, M., Fresno, S., Cook, N., et al. (2013). Effect of temperature and sunlight on the stability of human adenoviruses and MS2 as fecal contaminants on fresh produce surfaces. *International Journal of Food Microbiology, 164*(2–3), 128–134.

Dicaprio, E., Ma, Y., Purgianto, A., Hughes, J., & Li, J. (2012). Internalization and dissemination of human norovirus and animal caliciviruses in hydroponically grown romaine lettuce. *Applied and Environmental Microbiology, 78*(17), 6143–6152.

Domínguez, A., Torner, N., Ruíz, L., Martínez, A., Barrabeig, I., Camps, N., et al. (2008). Aetiology and epidemiology of viral gastroenteritis outbreaks in Catalonia (Spain) in 2004–2005. *Journal of Clinical Virology, 43*(1), 126–131.

Drechsel, P., Scott, C. A., Raschid-Sally, L., Redwood, M., & Bahri, A. (2010). *Wastewater Irrigation and Health. Assessing and mitigating risk in low-income countries.* London: Earthscan.

Esseili, M. A., Wang, Q., & Saif, L. J. (2012). Binding of human GII.4 norovirus virus-like particles to carbohydrates of romaine lettuce leaf cell wall materials. *Applied and Environmental Microbiology, 78*(3), 786–794.

Esseili, M. A., Wang, Q., Zhang, Z., & Saif, L. J. (2012). Internalization of sapovirus, a surrogate for norovirus, in romaine lettuce and the effect of lettuce latex on virus infectivity. *Applied and Environmental Microbiology, 78*(17), 6271–6279.

Ethelberg, S., Lisby, M., Bottiger, B., Schultz, A. C., Villif, A., Jensen, T., et al. (2010). Outbreaks of gastroenteritis linked to lettuce, Denmark, January 2010. *Eurosurveillance, 15*(6).

Gallimore, C. I., Pipkin, C., Shrimpton, H., Green, A. D., Pickford, Y., McCartney, C., et al. (2005). Detection of multiple enteric virus strains within a foodborne outbreak of gastroenteritis: An indication of the source of contamination. *Epidemiology and Infection, 133*(1), 41–47.

Haas, C. N., Rose, J. B., & Gerba, C. P. (1999). *Quantitative microbial risk assessment.* New York: John Willey & Sons, Inc (Pub).

Hamilton, A. J., Stagnitti, F., Premier, R., Boland, A. -M., & Hale, G. (2006). Quantitative microbial risk assessment models for consumption of raw vegetables irrigated with reclaimed water. *Applied and Environmental Microbiology, 72*(5), 3284–3290.

Havelaar, A. H., Haagsma, J. A., Mangen, M. -J. J., Kemmeren, J. M., Verhoef, L. P. B., Vijgen, S. M. C., et al. (2012). Disease burden of foodborne pathogens in the Netherlands, 2009. *International Journal of Food Microbiology, 156*(3), 231–238.

Hijnen, W. A. M., Beerendonk, E. F., & Medema, G. J. (2006). Inactivation credit of UV radiation for viruses, bacteria and protozoan (oo)cysts in water: A review. *Water Research*, *40*(1), 3–22.
Hirneisen, K. A., & Kniel, K. E. (2013). Norovirus surrogate survival on spinach during preharvest growth. *Phytopathology*, *103*(4), 389–394.
Hirneisen, K. A., Sharma, M., & Kniel, K. E. (2012). Human enteric pathogen internalization by root uptake into food crops. *Foodborne Pathogens and Disease*, *9*(5), 396–405.
Iglesias, R., Ortega, E., Batanero, G., & Quintas, L. (2010). Water reuse in Spain: Data overview and costs estimation of suitable treatment trains. *Desalination*, *263*(1–3), 1–10.
Kageyama, T., Kojima, S., Shinohara, M., Uchida, K., Hoshino, F. B., Takeda, N., et al. (2003). Broadly reactive and highly sensitive assay for Norwalk-like viruses based on real-time quantitative reverse transcription-PCR. *Journal of Clinical Microbiology*, *41*(4), 1548–1557.
Katayama, H., Haramoto, E., Oguma, K., Yamashita, H., Tajima, A., Nakajima, H., et al. (2008). One-year monthly quantitative survey of noroviruses, enteroviruses, and adenoviruses in wastewater collected from six plants in Japan. *Water Research*, *42*(6–7), 1441–1448.
Kemmeren, J. M., Mangen, M. -J. J., Duynhoven, Y. v, & Havelaar, A. H. (2006). *Priority setting of foodborne pathogens: Disease burden and costs of selected enteric pathogens.* Bilthoven (The Netherlands): RIVM.
Lamhoujeb, S., Fliss, I., Ngazoa, S. E., & Jean, J. (2008). Evaluation of the persistence of infectious human noroviruses on food surfaces by using real-time nucleic acid sequence-based amplification. *Applied and Environmental Microbiology*, *74*(11), 3349–3355.
Lodder, W. J., & de Roda Husman, A. M. (2005). Presence of noroviruses and other enteric viruses in sewage and surface waters in the Netherlands. *Applied and Environmental Microbiology*, *71*(3), 1453–1461.
Loisy, F., Atmar, R. L., Guillon, P., Le Cann, P., Pommepuy, M., & Le Guyader, F. S. (2005). Real-time RT-PCR for norovirus screening in shellfish. *Journal of Virological Methods*, *123*, 1–7.
Makary, P., Maunula, L., Niskanen, T., Kuusi, M., Virtanen, M., Pajunen, S., et al. (2009). Multiple norovirus outbreaks among workplace canteen users in Finland, July 2006. *Epidemiology and Infection*, *137*(3), 402–407.
Mara, D., & Sleigh, A. (2010a). Estimation of norovirus and Ascaris infection risks to urban farmers in developing countries using wastewater for crop irrigation. *Journal of Water and Health*, *8*(3), 572–576.
Mara, D., & Sleigh, A. (2010b). Estimation of norovirus infection risks to consumers of wastewater-irrigated food crops eaten raw. *Journal of Water and Health*, *8*(1), 39–43.
Martínez, A., Torner, N., Broner, S., Bartolomé, R., Guix, S., de Simón, M., et al. (2013). Norovirus: A growing cause of gastroenteritis in Catalonia (Spain)? *Journal of Food Protection*, *76*(10), 1810–1816.
Mok, H. -F., Barker, S. F., & Hamilton, A. J. (2014). A probabilistic quantitative microbial risk assessment model of norovirus disease burden from wastewater irrigation of vegetables in Shepparton, Australia. *Water Research*, *54*, 347–362.
Mok, H. -F., & Hamilton, A. J. (2014). Exposure factors for wastewater-irrigated Asian vegetables and a probabilistic rotavirus disease burden model for their consumption. *Risk Analysis*, *34*(4), 602–613.
Montemayor, M., Costan, A., Lucena, F., Jofre, J., Munoz, J., Dalmau, E., et al. (2008). The combined performance of UV light and chlorine during reclaimed water disinfection. *Water Science and Technology*, *57*(6), 935–940.
Mormann, S., Dabisch, M., & Becker, B. (2010). Effects of technological processes on the tenacity and inactivation of norovirus genogroup II in experimentally contaminated foods. *Applied and Environmental Microbiology*, *76*(2), 536–545.
Patel, M. M., Hall, A. J., Vinjé, J., & Parashar, U. D. (2009). Noroviruses: A comprehensive review. *Journal of Clinical Virology*, *44*(1), 1–8.
Petterson, S. R., Ashbolt, N. J., & Sharma, A. (2001). Microbial risks from wastewater irrigation of salad crops: A screening-level risk assessment. *Water Environment Research*, *73*(6), 667–672.
Petterson, S. R., Ashbolt, N. J., & Sharma, A. (2002). Of: Microbial risks from wastewater irrigation of salad crops: A screening-level risk assessment. *Water Environment Research*, *74*(4), 411.
Rajko-Nenow, P., Waters, A., Keaveney, S., Flannery, J., Tuite, G., Coughlan, S., et al. (2013). Norovirus genotypes present in oysters and in effluent from a wastewater treatment plant during the seasonal peak of infections in Ireland in 2010. *Applied and Environmental Microbiology*, *79*(8), 2578–2587.
Richards, G. P. (2012). Critical review of norovirus surrogates in food safety research: Rationale for considering volunteer studies. *Food and Environmental Virology*, *4*, 6–13.
Rutjes, S. A., van den Berg, H. H. J. L., Lodder, W. J., & de Roda Husman, A. M. (2006). Real-time detection of noroviruses in surface water by use of a broadly reactive nucleic acid sequence-based amplification assay. *Applied and Environmental Microbiology*, *72*(8), 5349–5358.
Sano, D., Pintó, R. M., Omura, T., & Bosch, A. (2009). Detection of oxidative damages on viral capsid protein for evaluating structural integrity and infectivity of human norovirus. *Environmental Science & Technology*, *44*(2), 808–812.
Seidu, R., Heistad, A., Amoah, P., Drechsel, P., Jenssen, P. D., & Stenstrom, T. A. (2008). Quantification of the health risk associated with wastewater reuse in Accra, Ghana: A contribution toward local guidelines. *Journal of Water and Health*, *6*(4), 461–471.
Shuval, H., Lampert, Y., & Fattal, B. (1997). Development of a risk assessment approach for evaluating wastewater reuse standards for agriculture. *Water Science and Technology*, *35*(11–12), 15–20.
Sobsey, M. D., Battigelli, D. A., Shin, G. A., & Newland, S. (1998). RT-PCR amplification detects inactivated viruses in water and wastewater. *Water Science and Technology*, *38*(12), 91–94.
Stine, S. W., Song, I., Choi, C. Y., & Gerba, C. P. (2005). Application of microbial risk assessment to the development of standards for enteric pathogens in water used to irrigate fresh produce. *Journal of Food Protection*, *68*(5), 913–918.
Tanaka, H., Asano, T., Schroeder, E. D., & Tchobanoglous, G. (1998). Estimating the safety of wastewater reclamation and reuse using enteric virus monitoring data. *Water Environment Research*, *70*(1), 39–51.
Teunis, P. F., Moe, C. L., Liu, P., Miller, S. E., Lindesmith, L., Baric, R. S., et al. (2008). Norwalk virus: How infectious is it? *Journal of Medical Virology*, *80*(8), 1468–1476.
The R Foundation for Statistical Computing (2013). *The R project for statistical computing.*
Topping, J. R., Schnerr, H., Haines, J., Scott, M., Carter, M. J., Willcocks, M. M., et al. (2009). Temperature inactivation of Feline calicivirus vaccine strain FCV F-9 in comparison with human noroviruses using an RNA exposure assay and reverse transcribed quantitative real-time polymerase chain reaction—A novel method for predicting virus infectivity. *Journal of Virological Methods*, *156*(1–2), 89–95.
Tree, J. A., Adams, M. R., & Lees, D. N. (1997). Virus inactivation during disinfection of wastewater by chlorination and UV irradiation and the efficacy of F + bacteriophage as a 'viral indicator'. *Water Science and Technology*, *35*(11–12), 227–232.
Tree, J. A., Adams, M. R., & Lees, D. N. (2003). Chlorination of indicator bacteria and viruses in primary sewage effluent. *Applied and Environmental Microbiology*, *69*(4), 2038–2043.
Vega, E., Smith, J., Garland, J., Matos, A., & Pillaii, S. D. (2005). Variability of virus attachment patterns to butterhead lettuce. *Journal of Food Protection*, *68*(10), 2112–2117.
Wadl, M., Scherer, K., Nielsen, S., Diedrich, S., Ellerbroek, L., Frank, C., et al. (2010). Foodborne norovirus-outbreak at a military base, Germany, 2009. *BMC Infectious Diseases*, *10*, 30.
Warriner, K., Ibrahim, F., Dickinson, M., Wright, C., & Waites, W. M. (2003). Internalization of human pathogens within growing salad vegetables. *Biotechnology & Genetic Engineering Reviews*, *20*, 117–134.
Wei, J., Jin, Y., Sims, T., & Kniel, K. E. (2011). Internalization of murine norovirus 1 by Lactuca sativa during irrigation. *Applied and Environmental Microbiology*, *77*(7), 2508–2512.
Westrell, T., Teunis, P., van den Berg, H., Lodder, W., Ketelaars, H., Stenström, T. A., et al. (2006). Short- and long-term variations of norovirus concentrations in the Meuse river during a 2-year study period. *Water Research*, *40*(14), 2613–2620.
WHO (2006). Guidelines for the safe use of wastewater, excreta and greywater. In W.H. Organization (Ed.), *Wastewater Use in Agriculture*, *Vol. II*, Geneva: WHO.

Altres publicacions 3

Evaluation of methods for the concentration and extraction of viruses from sewage in the context of metagenomic sequencing

Mathis Hjort Hjelms, Maria Hellmér, Xavier Fernandez-Cassi, **Natàlia Timoneda**, Oksana Lukjancenko, Michael Seidel, Dennis Elsässer, Frank M. Aarestrup, Charlotta Löfström, Sílvia Bofill-Mas, Josep F. Abril, Rosina Girones, Anna Charlotte Schultz

PLoS ONE (2017) 12(1): e0170199

PLOS ONE

RESEARCH ARTICLE

Evaluation of Methods for the Concentration and Extraction of Viruses from Sewage in the Context of Metagenomic Sequencing

Mathis Hjort Hjelmsø[1☯]*, Maria Hellmér[2☯], Xavier Fernandez-Cassi[3], Natàlia Timoneda[3,4], Oksana Lukjancenko[1], Michael Seidel[5], Dennis Elsässer[5], Frank M. Aarestrup[1], Charlotta Löfström[2¤], Sílvia Bofill-Mas[3], Josep F. Abril[3,4], Rosina Girones[3], Anna Charlotte Schultz[2]

1 Research Group for Genomic Epidemiology, The National Food Institute, Technical University of Denmark, Kongens Lyngby, Denmark, **2** Division of Microbiology and Production, The National Food Institute, Technical University of Denmark, Søborg, Denmark, **3** Laboratory of Virus Contaminants of Water and Food, Department of Genetics, Microbiology, and Statistics, University of Barcelona, Barcelona, Catalonia, Spain, **4** Institute of Biomedicine of the University of Barcelona, University of Barcelona, Barcelona, Catalonia, Spain, **5** Institute of Hydrochemistry, Chair of Analytical Chemistry, Technical University of Munich, Munich, Germany

☯ These authors contributed equally to this work.
¤ Current address: Food and Bioscience, SP Technical Research Institute of Sweden, Lund, Sweden
* mhjhj@food.dtu.dk

Citation: Hjelmsø MH, Hellmér M, Fernandez-Cassi X, Timoneda N, Lukjancenko O, Seidel M, et al. (2017) Evaluation of Methods for the Concentration and Extraction of Viruses from Sewage in the Context of Metagenomic Sequencing. PLoS ONE 12(1): e0170199. doi:10.1371/journal.pone.0170199

Editor: Patrick Tang, Sidra Medical and Research Center, QATAR

Received: October 13, 2016

Accepted: January 2, 2017

Published: January 18, 2017

Data Availability Statement: The metagenomic sequences are available from the European Nucleotide Archive (ENA) at the European Bioinformatics Institute (EBI) under accession number PRJEB15242.

Funding: Funding and support has been provided by European Union's Horizon 2020 research and innovation programme, COMPARE, grant agreement No. 643476 (http://www.compare-europe.eu/), the JPI Water project, METAWATER

Abstract

Viral sewage metagenomics is a novel field of study used for surveillance, epidemiological studies, and evaluation of waste water treatment efficiency. In raw sewage human waste is mixed with household, industrial and drainage water, and virus particles are, therefore, only found in low concentrations. This necessitates a step of sample concentration to allow for sensitive virus detection. Additionally, viruses harbor a large diversity of both surface and genome structures, which makes universal viral genomic extraction difficult. Current studies have tackled these challenges in many different ways employing a wide range of viral concentration and extraction procedures. However, there is limited knowledge of the efficacy and inherent biases associated with these methods in respect to viral sewage metagenomics, hampering the development of this field. By the use of next generation sequencing this study aimed to evaluate the efficiency of four commonly applied viral concentrations techniques (precipitation with polyethylene glycol, organic flocculation with skim milk, monolithic adsorption filtration and glass wool filtration) and extraction methods (Nucleospin RNA XS, QIAamp Viral RNA Mini Kit, NucliSENS® miniMAG®, or PowerViral® Environmental RNA/DNA Isolation Kit) to determine the viriome in a sewage sample. We found a significant influence of concentration and extraction protocols on the detected viriome. The viral richness was largest in samples extracted with QIAamp Viral RNA Mini Kit or PowerViral® Environmental RNA/DNA Isolation Kit. Highest viral specificity were found in samples concentrated by precipitation with polyethylene glycol or extracted with Nucleospin RNA XS. Detection of viral pathogens depended on the method used. These results contribute to the understanding of method associated biases, within the field of

(http://www.waterjpi.eu/images/Kick-Off/METAWATER.pdf) and Aquavalens (EU FP7-KBBE-2012-6) (http://aquavalens.org/). This study was partially funded by a grant of the Catalan Government as the Consolidated Research Group VirBaP (2014SRG914) (http://www.ub.edu/microbiologia/grupbacterisen/index.html). During the development of the study XFC was a fellow of the Catalan Government "AGAUR" (FI-DGR) (http://www.uab.cat/web/research/uab-research-training-grants/postdoctoral-grants/catalan-government-1184220108167.html) and NT was a fellow of the Spanish Ministry of Science (http://www.uab.cat/web/research/uab-research-training-grants/postdoctoral-grants/ministry-of-education-and-science-1184220108300.html). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

viral sewage metagenomics, making evaluation of the current literature easier and helping with the design of future studies.

Introduction

Within raw sewage, feces, urine and other biological fluids from thousands of humans are mixed together with food and household waste, industrial waste, and runoff water. Every individual, who is connected to the drainage system, contributes with his or hers own microbiota [1], including infecting pathogens [2]. This makes sewage an attractive matrix for epidemiological studies [3], microbial source tracking [4], and for controlling the efficacy of pathogen removal in wastewater treatment plants [5,6]. Sewage has been shown to harbor a diverse viral population including enteric, respiratory and oncogenic viruses [7]. The high viral diversity and the continuous mutation of viral species makes identification with traditional methods difficult and time consuming, therefore many studies have turned to Next Generation Sequencing (NGS) approaches instead [7–9]. Metagenomic sequencing of the virus associated nucleic acids is considered to be an unbiased approach enabling the detection of all known viral species, as well as the discovery of novel and emergent species [10]. Three main challenges exist for viral sewage metagenomics. First, only a small fraction of the total nucleic acids are of known viral origin, hence mechanical and enzymatic viral purification is often needed [9]. Second, the low abundance of viral particles in the samples requires the use of viral concentration methods prior to nucleic acid extraction [11] and is often combined with subsequent random DNA amplification [12]. Third, the nucleic acid extraction procedure has to cover the large variety in viral structures and genome types. To overcome these biases, different methods to concentrate viruses from water samples have been developed, including: polyethylene glycol precipitation (PEG) [8], $FeCl_3$ precipitation [13], skimmed milk flocculation (SMF) [14], glass wool filtration (GW) [15] or monolithic adsorption filtration (MAF) [16]. The influence of concentration method on viral recovery has been evaluated on sea water [17], spiked tap water [15,18] and raw sewage [19], cautioning of method associated biases. To our knowledge, no major comparison studies using metagenomics have been performed with sewage water.

Biases caused by nucleic acid extraction kits have been well documented for both bacteria [20,21] and viruses [22,23]. In addition, contaminants have been found to be ubiquitous in some extraction kits [24] and laboratory reagents [25], potentially giving rise to false positive results [26,27]. A better understanding of specific method associated biases, in respect to viral wastewater metagenomics, would make evaluation of the current literature easier, and help guide future studies.

In this study we evaluated four previously published concentration methods, PEG, MAF, SMF, and GW, as well as four extraction kits, Nucleospin RNA XS (NUC), QIAamp Viral RNA Mini Kit (QIA), NucliSENS® miniMAG® (MIN), or PowerViral® Environmental RNA/DNA Isolation Kit (POW), for wastewater viral metagenomics, in a full factorial design resulting in 16 combinations of procedures. Aspects studied included viral community composition, viral selectiveness, viral richness, viral pathogen detection, and viral contaminants. Extracted nucleotides were amplified with PCR and sequenced using the Illumina MiSeq platform.

Materials and Methods

Sample collection, spiking and pooling

In July 2015 raw sewage (130 L) was collected at the waste water treatment plant BIOFOS Lynetten in Copenhagen, Denmark, receiving waste water from about 550,000 inhabitants. Approval was granted from BIOFOS Lynettefællesskabet A/S before sampling. The sewage was mixed thoroughly in a single container and spiked to a concentration of 1.74×10^8 RT-PCR units/L of murine norovirus (MNV) (kindly provided by Dr Virgin, Washington University School of Medicine, USA), and 2.13×10^9 genome copies/L of human adenovirus 35 (HAdV). The sample was mixed for 5 min before aliquoted and stored at—20˚C until further processing.

Concentration methods

Four different methods were used to concentrate virions from the sewage samples: protein precipitation with PEG, organic flocculation with SMF and filtration with positively charged filters, MAF, or GW. All concentration methods were done in triplicate together with a negative control using sterile molecular grade water (VWR—Bie & Berntsen, Søborg, Denmark).

PEG

The PEG protocol was based on the procedure as previously described [8]. Initially, 25 mL of glycin buffer (0.05 M glycine, 3% beef extract, pH 9.6) was added to 200 mL of sewage and mixed, to detach virions bound to organic material. The sample was then centrifuged at 8,000×g for 30 min, and the collected supernatant was filtered through a 0.45 μm polyethersulfone (PES) membrane (Jet Biofil, Guangzhou, China) to remove bacterial and eukaryotic cells. Viruses were precipitated from the supernatant by incubation with PEG 8000 (80 g/L) and NaCl (17.5 g/L) during agitation (100 rpm) overnight at 4˚C, followed by centrifugation for 90 min at 13,000×g. The resulting viral-containing pellet was eluted in 1 mL phosphate buffer saline (PBS) and stored at -80˚C until further processing.

MAF

The principle of the MAF adsorption/elution method was based on the procedure as previously described [18]. Monolithic discs, diameter 3.86 cm and height 1.0 cm, were synthesized by polymerization of polyglycerol-3-glycidyl ether (Ipox chemicals, Laupheim, Germany). An 80:20 mixture of toluene and tert-buthyl methyl ether was used as porogen to create monoliths with a pore size of ca. 20 μm. After synthesis, functionalization was performed by recirculating 10% diethylamine in 50% ethanol at 60˚C through the monolithic disks for 3 h to create positively charged diethylaminoethyl groups on the pore surface. Afterwards the monoliths were rinsed with ultrapure water and stored at 4˚C until further use. One liter of raw sewage was filtrated through a MAF disc (Microarray and Bioseparation Group of the Institute of Hydrochemistry, Technical University of Munich, Germany) assembled as previously described [28]. Viruses were eluted from the filter by soaking 2×2 min in a total of 20 mL high salt buffer (1.5 M NaCl, 0.05 M HEPES (4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid) buffer, pH 7). The eluate was further concentrated to 3 mL by 100 kDa Amicon ultra centrifugation filters (Merck Millipore, Cork, Ireland) according to the manufacturer's instructions. The viral concentrate was stored at -80˚C until further processing.

SMF

Organic flocculation with skimmed milk was based on the procedure as previously described [7]. In brief, 100 mL pre flocculated skimmed milk solution (1% (w/v) skimmed milk powder

(Difco, Detroit, MI, USA), 3.2% (w/w) sea salts (Sigma Aldrich Chemie GMBH, Steinheim, Germany)) at pH 3.5 was added to 10 L of acidified (using HCl to pH 3.5) raw sewage and mixed for 8 h. Flocculants were allowed to sediment for 8 h, and centrifuged at 8,000×g for 40 min. The pelleted viral concentrate was suspended in 15 mL phosphate buffer (1:2 (v/v) mixture of 0.2 M Na_2HPO_4 and 0.2 M NaH_2PO_4). The phosphate suspension was eluted in 30 mL 0.25 M glycine buffer (pH 9.5) with slow agitation for 45 min at 4˚C. Suspended solids were separated by centrifugation at 8,000×g for 40 min at 4˚C. The sample was neutralized to pH 7 by adding 1 M HCl. Virions present in the supernatant were concentrated by ultracentrifugation at 90,000×g for 90 min (Sorvall Discovery 90SE) at 4˚C and suspended in 2 mL PBS. The viral concentrate was stored at -80˚C until further processing.

GW

The glass wool filters were prepared as previously described [29]. Sodocalcic glass wool (15 g) (Ouest Isol, Alizay, France) was packed into a PVC tube with the density of 0.11 g/cm^3 and pretreated with the following solutions, 100 mL NaOH (1 M) for 15 min, 1 L sterile distilled water, 100 mL HCl (1 M) for 15 min, and 1 L sterile distilled water. Samples of raw sewage (4 L) were filtered through the glass wool column. Viruses were eluted by incubating 100 mL elution buffer (3% beef extract, 0.5 M glycine, pH 9.5) for 15 min. Secondary concentration was done by PEG precipitation (as above) to a final volume of 1 mL.

DNase/RNase treatment + chloroform-butanol treatment

All viral concentrates were treated with OmniCleave endonuclease (Epicentre, Wisconsin, USA) to remove extracellular DNA/RNA as previously described [30]. Samples were further purified by extraction using a 1:1 mixture of chloroform-butanol [31] to remove nucleases and inhibitors.

Extraction methods

Nucleic acids were extracted from 200 μL-portions of the respective viral concentrate using four different extraction kits; NUC (Macherey-Nagel, Düren, Germany), QIA (Qiagen, Valencia CA, USA), MIN (BioMerieux, Herlev, Denmark) or POW (MO BIO, Carlsbad, CA, USA). In all cases, extractions were carried out according to manufacturer's instructions.

qPCR analysis of spiked viruses

Detection of HAdV and MNV was performed on extracted nucleic acids (undiluted and 10-fold diluted) in a 96-well plate format of ABI Step One (Applied Biosystems, Naerum, Denmark). MNV RNA was detected by quantitative reverse transcriptase polymerase reaction (qRT-PCR) using the RNA UltraSense one-step qRT-PCR system (Invitrogen, Taastrup, Denmark) and previously described primers and probes [32]. Amplification was performed in a 25 μL reaction mixture containing 5 μL extracted nucleic acids and 20 μL qRT-PCR reaction mixture with 500 nM forward primer, 900 nM reverse primer, 250 nM probe, 1 × UltraSense reaction mix, 1 × ROX reference dye and 1 × UltraSense enzyme mix under the following reaction conditions, 55˚C for 1 min and 95˚C for 5 min followed by 45 cycles of 95˚C for 15 s, 60˚C for 1 min, and 65˚C for 1 min. HAdV DNA was detected by qPCR using TaqMan Universal Master Mix (Applied Biosystems, Naerum, Denmark), and previously described primers and probe [33]. Amplification was performed in a total of 25 μL reaction mixture containing 5 μL extracted nucleic acids and 20 μL qPCR reaction mixture containing 1 × TaqMan Universal Master Mix, primer and probe concentrations and qPCR running

conditions are described in [33]. Quantification was performed using standard curves generated from 10-fold dilution series, of extracted RNA of cell propagated MNV or of ds HAdV DNA segments, artificially constructed by gBlocks® Gene Fragments (Integrated DNA Technologies, Leuven, Belgium).

Reverse transcriptase, library preparation and, sequencing

To prepare extracted RNA and DNA for sequencing, each viral extract was subjected to reverse transcriptase and PCR amplified, as previously described [34]. Briefly, first strand cDNA synthesis were performed using the SuperScript® III First-Strand Synthesis SuperMix (Invitrogen, Carlsbad, California) and 1 μL Primer A (50 μM) (5′-GTTTCCCAGTCACGATCNNNNNNNNN-3′) according to the manufacturer's instructions. Second strand DNA synthesis were performed using Klenow Fragment exo-polymerase (Thermo Fisher Scientific, Waltham, MA, USA) as previously described [30]. Double stranded DNA products were PCR amplified using AmpliTaq Gold (Qiagen, Valencia CA, USA) as per manufacturer's instruction using 0.8 μM Primer B (5′- GTTTCCCAGTCACGATC-3′) and the following conditions, 10 min at 95˚C, 25 cycles of amplification (94˚C for 30 s, 40˚C for 30 s, 50˚C for 30 s and 72˚C for 1 min), and 1 cycle of elongation (72˚C for 10 min). PCR products were purified using the DNA Clean & Concentrator™-5 (Zymo Research, Irvine CA, USA). NGS library preparation was performed using the Nextera XT DNA Library Preparation kit (Illumina, Eidenhoven, The Netherlands) according to the manufacturer's instructions. The 64 samples were sequenced on three Illumina MiSeq runs with an average output of 1.4×10^6 250 bp paired-end reads per sample (S1 Table).

Bioinformatic analyses

The distribution of viral species was determined using MGmapper software version 2.2 (https://cge.cbs.dtu.dk/services/MGmapper/) [31]. The MGmapper tool follows three main steps: quality assessment of the raw reads, mapping of reads to the reference databases, and post-processing of mapping results. Quality assessment was done using cutadapt [35] which performs common adapter removal, trimming of the low-quality ends from reads with a minimum Phred quality score of 20, and later discards reads that are shorter than 40 bp. Later, already trimmed pair-end reads were aligned to a pre-defined set of reference sequence databases using bwa mem [36] ver. 0.7.7-r441 with default settings. In this study, reads were mapped against three viral reference databases (S2 Table): whole genomes virus sequences (Virus) and viral sequences extracted from nt database (Virus_nt), obtained from Genbank (http://www.ncbi.nlm.nih.gov/genbank/), as well as Vipr database (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3245011). Samtools [37] were used to remove singletons and filter reads where neither a read nor its mate is mapped. Reads were mapped in best-mode, meaning that mapping was performed against all databases, simultaneously, and later for each read pair the best hit among all alignments is chosen. A pair of reads is considered as a hit only if the sum of the alignment scores (SAS) is higher than any SAS values from other database hits. If a pair of reads has identical SAS values when mapping to several databases, the only one pair, associated with the database that was specified first in the list of reference databases, is kept. In the last, post-processing step, alignments are filtered based on matches/mis-matches threshold. In this analysis, 70% matches/mis-matches threshold needed to be satisfied in order the hit to be considered significant. The metagenomic sequences are available from the European Nucleotide Archive (ENA) at the European Bioinformatics Institute (EBI) under accession number PRJEB15242.

Statistics and plots

Viral richness was estimated using the program CatchAll [38] and the non-parametric Chao1 richness index, as a measure of number of viral species in a sample. All statistics were done in R [39], using two-way analysis of variance (ANOVA) test for determining the overall significance of concentration or extraction method on the studied factors (viral richness, etc.). Subsequently, pairwise t-tests with "Holm-Bonferroni" [40] p-value adjustments were applied to determine significant pairwise effects between individual concentration or extraction methods. Principal component analysis (PCA) were performed using prcomp and plotted with ggbiplot [41]. Heatmaps were created using pheatmap [42]. Linear regression between reads per million (RPM) and genome copies per liter were done in Excel on log transformed data.

Results

In this study different virus concentration and nucleic acid extraction methods were evaluated for metagenomic analysis of sewage samples. Sequencing results showed that the majority of the mapped reads (>80%) were of viral origin (S1 Fig). However, between 60 and 90% of the total reads were unmapped. The three main viral families detected were Adenoviridae (human viruses) including the spiked HAdV 35, Virgaviridae (plant viruses), and Siphoviridae (bacteriophages). The sequencing data were analyzed further to determine the viral community composition, viral specificity, viral richness, and detection of pathogenic species.

Viral community composition

To compare the viral community composition resulting from the individual concentration and extraction methods, a series of PCAs were made using the relative abundances from the nine most abundant viral families, accounting for more than 99% of the mapped viral sequences. The effect of extraction (Fig 1) and concentration (Fig 2) were plotted independently for easier visualization. Samples plotted close together have similar viral community compositions, whereas samples far away from each other are less alike. The negative controls clustered together far away from the samples in initial PCA plots (data not shown). To allow for better visualization of the effect of concentration and extraction on the sewage samples, they were not included in Figs 1 and 2.

Sample replicates extracted with NUC clustered away from the samples extracted with the other methods, when concentrated with PEG (Fig 1D). This was also true for the concentrates from MAF (Fig 1A), SMF (Fig 1B), and GW (Fig 1C), suggesting that the viral community composition of the NUC extractions differed from the other tested extraction methods. The samples extracted with POW and MIN clustered together, suggesting similar viral community compositions (Fig 1A–1D). The samples extracted with QIA sometimes clustered separately (Fig 1A and 1C) and sometimes together with the samples extracted with POW and MIN (Fig 1B and 1D). The four concentration methods formed separate non-overlapping clusters regardless of extraction kit used (Fig 2A–2D), although some variation between replicates were observed.

Viral specificity

The proportion of reads mapping to viruses ranged between 3.4% and 49.4%. Both the concentration and the extraction methods had a statistical significant effect on the viral specificity (two-way ANOVA, $p < 0.001$). However, a significant interacting effect (two-way ANOVA, $p < 0.001$) indicated that the effect on viral specificity by the extraction method was affected by the type of concentration method, and vice versa.

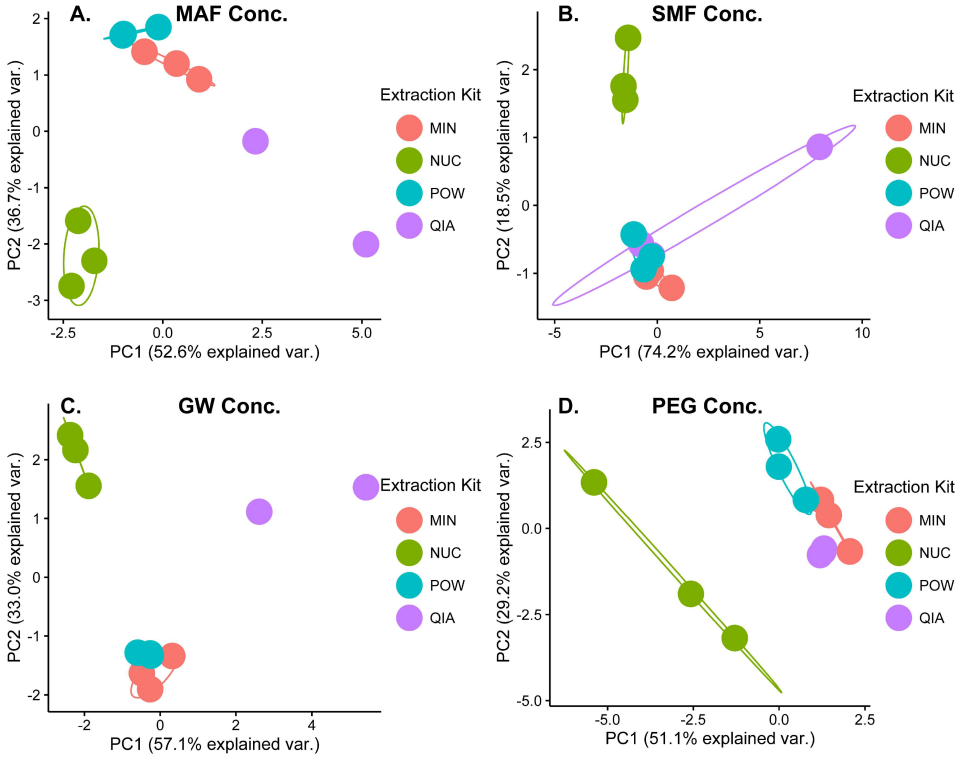


Fig 1. The influence of extraction method on the viral community composition. PCA plots made by using the relative abundances of the nine most abundant viral families. Separate PCAs were done for (A) samples concentrated with MAF, (B) SMF, (C) GW, and (D) PEG. Sample replicates were individually plotted and grouped according to the extraction method. In cases where only two samples were present, no ellipse representing the cluster was drawn.

doi:10.1371/journal.pone.0170199.g001

The PEG concentration method had a significant larger mean proportion of viral reads compared to the SMF and GW methods (pairwise t-test, $p < 0.01$) (Fig 3A). For the extraction methods, NUC had a significant larger mean proportion of viral reads compared to POW and QIA (pairwise t-test, $p < 0.01$) (Fig 3B). However, there were some interacting effects, with MIN scoring higher than NUC when used in combination with PEG and GW, implying that the MIN method depends heavily on the performance of the concentration method.

Viral richness

Both concentration and extraction methods had an effect on the viral richness. However, none of the concentration methods were statistically different from each other (pair-wise t-test) (Fig 4A). For the extraction methods, NUC had a significantly lower Chao1 richness than the other methods (Fig 4B). POW and QIA had the highest mean richness estimates of 516 and 495, respectively.

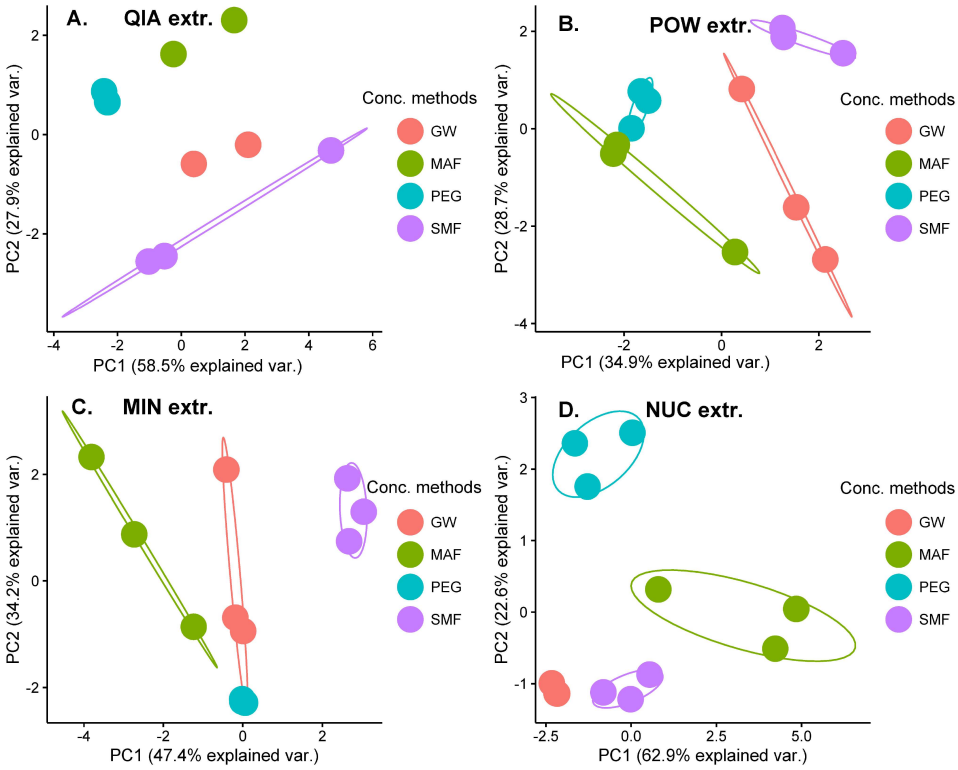


Fig 2. The influence of concentration method on the viral community composition. PCA plots made by using the relative abundances of the nine most abundant viral families. Separate PCAs were done for (A) samples extracted with QIA, (B) POW, (C) MIN, and (D) NUC. Sample replicates were individually plotted and grouped according to the concentration method. In cases where only two samples were present, no ellipse representing the cluster was drawn.

doi:10.1371/journal.pone.0170199.g002

Detection of pathogenic species

Fourteen viral families with suspected human pathogens were detected (Fig 5). The most prevalent was Adenoviridae including the spiked HAdV. The highest read count for the viral RNA families, Reoviridae, Picornaviridae, Astroviridae, Caliciviridae and Picorbinaviridae, was obtained in samples extracted with NUC. The spiked HAdV was detected at the highest abundance when extracted with MIN. The effect of the concentration methods was not as pronounced as for the extraction kits. The highest read count of the DNA virus family, Adenoviridae, was found in samples concentrated with MAF and PEG. In general, SMF had a lower performance compared with the other methods when testing for Adenoviruses such as the spiked HAdV. However, the combination of SMF and NUC had the highest read count for most of the RNA viruses.

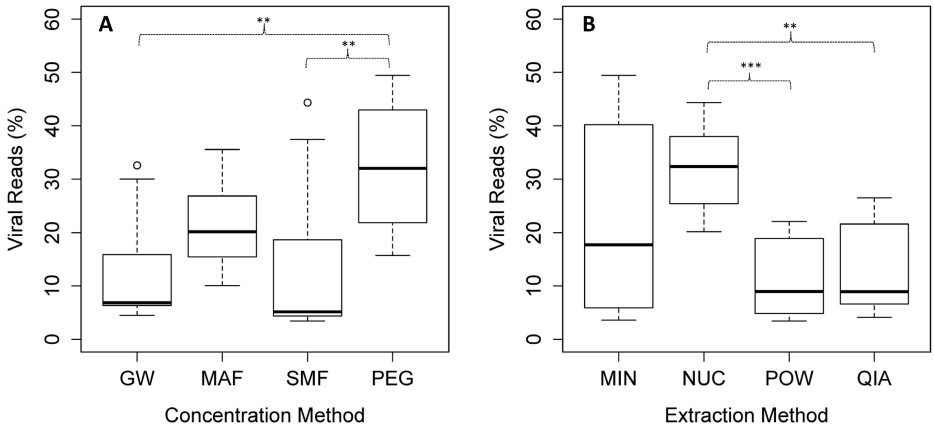


Fig 3. Viral selectivity measured in percentage of reads. (A) Viral selectivity for the tested concentration methods (B) and extraction methods. Each boxplot was made from 12 individual samples (including the four extraction/concentration methods with three replicates each). The bar, box, whiskers and circles represents median, inter-quartile range, inter-quartile range times 1.5, and outliers, respectively. Asterisks represent significance level of a pairwise t-test with "Holm-Bonferroni" adjusted p-values. ** = $p < 0.01$, *** = $p < 0.001$.

doi:10.1371/journal.pone.0170199.g003

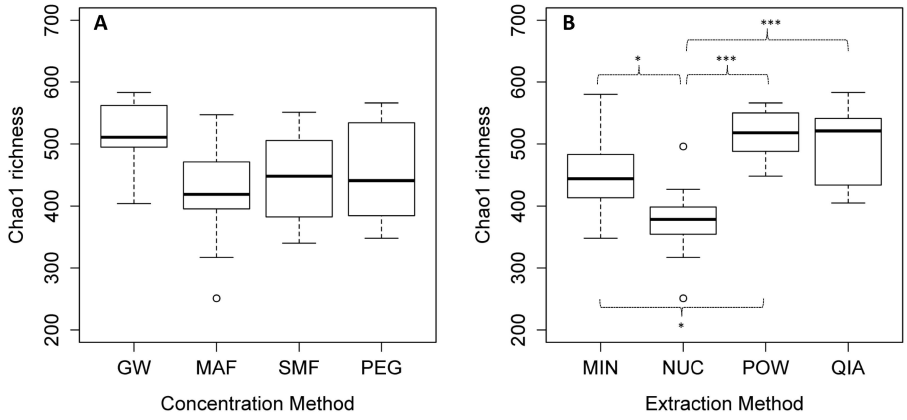


Fig 4. Viral species richness. (A) Viral Chao 1 species richness of the tested concentration methods, and (B) extraction methods. Each boxplot was made from 12 individual samples (including the four extraction/concentration methods with three replicates each). The bar, box, whiskers and circles represents median, inter-quartile range, inter-quartile range times 1.5, and outliers, respectively. Asterisks represent significance level of a pairwise t-test with "Holm-Bonferroni" adjusted p-values. * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$.

doi:10.1371/journal.pone.0170199.g004

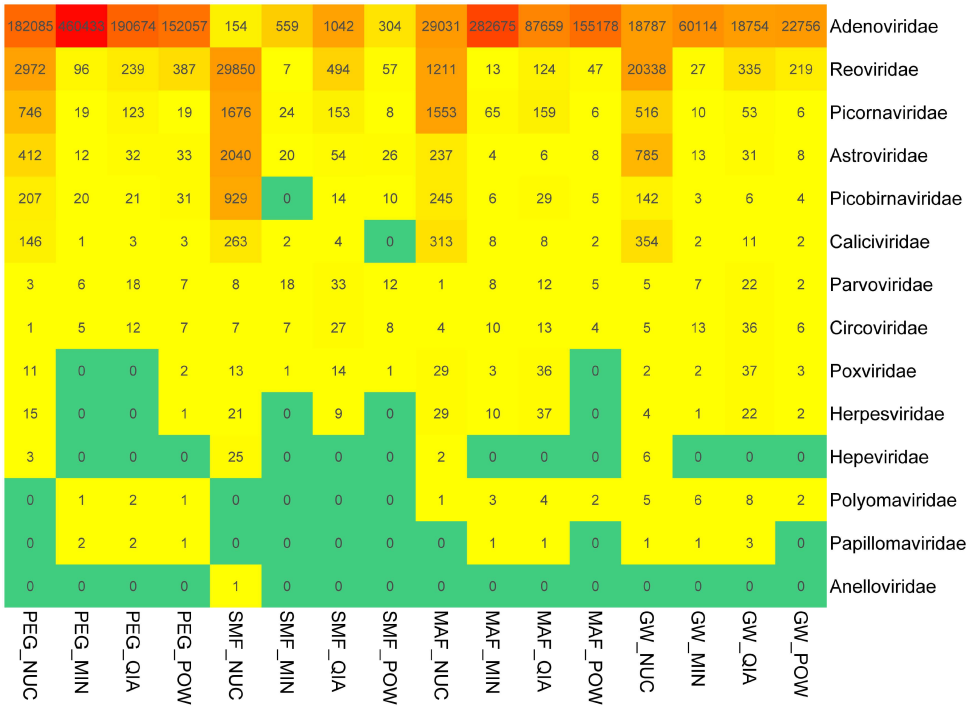


Fig 5. Detection of pathogenic viral families. Heatmap of the relative abundance of 14 human pathogenic viral families, detected by the 16 different concentration/extraction combinations. The numbers within each cell represents reads per million. The colours range from green = no detection, to red = high relative abundance.

doi:10.1371/journal.pone.0170199.g005

The spiked MNV were only detected by metagenomics in 57% of the samples, and at low read counts, from 2 to 194 reads. The combinations that could detect most MNV were MAF and GW extracted with NUC as well as MAF extracted with QIA and MIN.

qPCR analysis of spiked viruses

The detected concentrations of HAdV and MNV varied widely between the different method combinations (S2 and S3 Figs) with mean values ranging between 650 and 8.2×10^7 genome copies/L for HAdV, and 1.8×10^2 and 3.9×10^5 RT-PCR units/L for MNV. Choice of extraction method did not influence HAdV or MNV recovery. However, concentration methods had a significant impact (pairwise t-test, $p < 0.05$). The highest recovery of HAdV and MNV was obtained with PEG followed by MAF, GW, and SMF.

Inhibition

To investigate the possibility of PCR inhibition, extracts of nucleic acids (undiluted and 10-fold diluted) from all samples were analyzed for the two spiked viruses, MNV and HAdV,

with qPCR (S3 Table). The lowest inhibition of MNV and HAdV were observed in samples concentrated with PEG or extracted with MIN. Strongest inhibition was observed in samples concentrated with SMF for both MNV and HAdV. In addition samples extracted with QIA showed strong inhibition of HAdV detection.

Correlation between qPCR quantification and reads per million (RPM)

To investigate the correlation between viral concentrations and RPM, qPCR data was compared with read counts from the two spiked viruses, HAdV and MNV. There was a strong correlation between RPM and qPCR enumeration for HAdV ($R^2 = 0.82$). However, no relationship was observed for MNV ($R^2 = 0.07$).

Contamination

To detect method dependent contamination, a negative control was included, using sterile molecular grade water, for each of the 16 method combinations. Negative controls generally had a low total read count, a low percentage of viral reads (0.05–3.4%), and a high abundance of reads with human, bacterial, fungal and parasitic origin (S1 Fig). Several viral species were found in the negative controls with much higher RPM values than in the corresponding samples, suggesting that they originated from the corresponding kits or reagents. Reads mapping to pandora viruses, tupaiid herpes viruses, and *Citrobacter* phages were contaminants in all procedures except the ones using QIA extractions. However several mardi viruses were found exclusively in the QIA negative controls.

Discussion

In the presented study we evaluated the influence of four commonly applied concentration and extraction methods on viral metagenome analysis.

The viral community composition was heavily biased by the type of concentration procedure, which dramatically skewed the relative abundances (Fig 2). Choice of extraction kit did not influence the viral community composition to the same degree (Fig 1). However, the results from the NUC extraction kit were remarkably different from samples extracted with the three other kits. The NUC kit includes an "on column DNase step" after viral capsid disruption, which selects for RNA viruses and could explain the separate clustering in the PCA plots (Fig 1). Based on the results from this study it seems inadvisable to compare results, in relation to viral community composition, between studies using different concentration methods and to some degree also extraction methods.

A high species richness have been linked to several ecosystem functions [43], and is often included as a factor in ecological studies. In this study we included the measure to discern if some methods were better at catching the entire spectrum of viral species. Our results show that the choice of extraction method is of more importance than the choice of concentration method with regard to viral richness. However, samples concentrated with GW had a slightly higher richness compared to the other concentration methods (Fig 4A). The low mean richness of the samples extracted with NUC can probably be explained by the DNase step, degrading the genomes of DNA viruses and the species rich bacteriophages [44].

Viral specificity, or how large a fraction of the sequencing reads is of viral origin, is important for sensitivity reasons, increasing the chance to detect rare or less abundant species. A high viral specificity also has financial implications, causing large savings on both sequencing, and for subsequent CPU hours used in the bioinformatics analyses. In this study, the PEG protocol was the best concentration method, in respect to viral specificity (Fig 3A). This might be explained by the initial filtration step, not part of the other evaluated protocols. Pre-filtration

might have improved the viral specificity in the other concentration methods, although clotting might become a problem due to the increased volumes processed with these methods. The NUC had a consistent high viral specificity (three times that of POW and QIA), probably due to the effective removal of DNA from other organisms, and contaminants, during the DNase step. Overall, there was a 10-fold difference in viral specificity between the lowest and the highest method combination, highlighting the potential savings associated with choice of method. We observed a generally high viral specificity in this study compared to previous studies [45]. This might be due to the addition of the spiked HAdV, inflating the amount of virus in the sewage matrix, but should not have any influence on the method comparisons.

Sewage metagenomics is often used to detect human viral pathogens [8] including the important enteric RNA viruses as norovirus [46], rotavirus [47] and Hepatitis A and E virus [48] that has a big impact on public health [49]. These RNA viral families were best detected when using the NUC extraction kit compared to the other tested extraction kits, especially in combination with the concentration method SMF. However, if looking at DNA viruses exclusively, the MIN extraction combined with PEG, MAF, or GW may be preferable, since it produced the highest read counts for the spiked Adenoviridae. Low detection of Adenoviruses using SMF concentration has previously been described [7], and were also observed in this study. In addition, SMF failed to detect the low numbers of reads of polyomaviruses and papillomaviruses observed by the other methods.

The larger initial sample volume, and associated organic material and inhibitors, for SMF (10 L) compared with the other methods (4, 1 and 0.2 L for GW, MAF, and PEG, respectively), could be an explanation for the low recovery of the spiked viruses. Inhibitors can affect PCR amplification, quality of the prepared library, and subsequent virus detection. This theory was further supported by the qPCR results were extracts obtained from SMF had a high level of inhibition. Extraction with QIA has previously been shown to impair detection of HAdV in samples with high levels of organic matter [23]. This was also the case in our study, where extraction with QIA inhibited HAdV detection in all cases except when combined with PEG concentration which both had the lowest starting volume (0.2 L) and an additional filtration step.

Sampling volume is an important factor in viral metagenomics, enhancing the sensitivity and increasing the chances of detecting rare viruses. However, in this study, we did not find a positive relation between methods with high sampling volumes and increased sensitivity. This could be due to an increase in inhibitors or other aspects of the employed concentration methods, although this question was not within the scope of this study. Further studies are needed to investigate the influence of sample volumes and viral metagenomics.

In this study, the bioinformatic analyses were done using alignment of single reads to three virus databases, using the program MGmapper. The choice of bioinformatics pipeline can affect results [50] but any biases of our particular approach should be the same on all samples and should therefore not affect the conclusions of this study.

Low levels of MNV were detected in the metagenomics analysis compared to the amounts used for spiking. However, the reasonable high values that could be detected using qPCR, indicated that the initial extraction was successful. Noroviruses have previously been documented to be difficult to detect using metagenomics [51,52] possibly because of the small genome, robust nucleocapsid, or inhibitory RNA secondary structures [53]. Virus species specific extraction efficiency biases are well documented in viral metagenomics [54] and should always be considered when interpreting the results. Quantitative conclusions from viral metagenomics are not possible for all viral species, illustrated by the good correlation between RPM and qPCR data found for HAdV where no correlation was found for MNV.

Several viruses were detected in higher amounts in the negative controls than in the corresponding samples, strongly suggesting them to be procedure contaminants. Contaminating

DNA is a huge challenge for low input metagenomics [24], and contaminating viral nucleotides have previously been detected in polymerases [25], spin columns [27] and DNases [54]. The specific origin of the contaminating viruses in our study was not clear although some avian herpesviruses were only linked to the QIA extracts. The ubiquitous presence of contaminating viruses stress the importance of including negative controls in future viral metagenomics studies, as well as adding measures to reduce the problem [55,56].

When evaluating the efficiency of the tested methods, clear differences were observed. No single method was superior to the others in all of the tested parameters. However, some trends were observed for the concentration methods as PEG scored higher in viral specificity and SMF inhibited detection of both spiked viruses. In the evaluation of the tested extraction methods the NUC kit stood out in regard to viral specificity and RNA virus detection. Nevertheless, if the focus is only on DNA viruses, for example phage studies, NUC might not be the best option since it scored low in viral richness which could result in loss of rare species. Practical aspects of the concentration and extraction methods were not within the scope of this paper, but may also influence the choice of method (S4 and S5 Tables).

In conclusion, we found a significant influence of concentration and extraction protocols on viral richness, viral specificity, viral pathogen detection, and viral community composition for metagenomic analyses of sewage. This is of major importance when interpreting results from the literature and conducting meta-studies. The use of data base resources, such as the European nucleotide archive (ENA) and short read archive (SRA) are also severely hampered by this fact since extraction kit, volume sample, and concentration procedure are not usually included in the metadata of published viromes. We suggest that such metadata will be included in the future, to allow researchers to select and compare studies conducted with similar methodologies.

Supporting Information

S1 Fig. Distribution of reads on kingdom level of the 16 method combinations and their associated negative controls. Samples were processed in triplicate, and the data shown is the average. _S = sample, _C = Negative extraction control. Databases used are listed in S1 Table.
(PDF)

S2 Fig. HAdV concentration measured by qPCR. (A) HAdV concentration in extracts obtained by using four different concentration methods and (B) extraction methods. The bar, box, whiskers and circles represents median, inter-quartile range, inter-quartile range times 1.5, and outliers, respectively. Asterisks represent significance level of a pairwise t-test with "Holm-Bonferroni" adjusted p-values. ** = $p < 0.01$, *** = $p < 0.001$.
(PDF)

S3 Fig. MNV concentrations measured by qPCR. (A) MNV concentration in extracts obtained by using four different concentration methods and (B) extraction methods. The bar, box and whiskers represents the median, the inter-quartile range, and the inter-quartile range times 1.5, respectively. Asterisks represent significance level of a pairwise t-test with "Holm-Bonferroni" adjusted p-values. * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$.
(PDF)

S4 Fig. Abundance of all detected viral families. Heatmap showing the abundance of all detected viral families, measured in reads per million, in each biological replica for the different method combinations as well as the controls. _S = sample, _C = Negative control.
(PDF)

S1 Table. Sequence information. Number of raw reads, reads after quality assessment, and reads not mapping to PhiX, and thus usable for subsequent analysis. _S = sample, _C = Negative control.
(PDF)

S2 Table. Overview of reference sequence databases and associated download information. Reference sequence information can be obtained from the URL's shown in 'Download information'.
(PDF)

S3 Table. qPCR inhibition of MNV and HAdV. Inhibition of the 16 combinations of concentration and extraction methods. Inhibition was measured using qPCR of undiluted (1:1) and tenfold diluted (1:10) DNA/RNA extracts. The values in the tables represents Δct between the undiluted and 10 fold diluted samples. A Δct = -3.3 represent a perfect 10 fold dilution. Samples marked in red represents undiluted extracts that could not be quantifiable, these samples are regarded as the most inhibited.
(PDF)

S4 Table. Specifications of the four concentration methods applied in this study.
(PDF)

S5 Table. Properties of the four nucleic acid extraction kits applied in this study.
(PDF)

Acknowledgments

We are grateful to Resadije Idrizi for technical assistance and to Dines Thornbjerg and BIOFOS for providing the wastewater.

Author Contributions

Conceptualization: MHH MH XFC.

Formal analysis: MHH OL NT XFC MH.

Funding acquisition: FMA RG CL ACS.

Investigation: MHH MH.

Methodology: MHH MH XFC.

Project administration: MHH MH.

Resources: ACS RG DE MS JFA FMA SBM.

Software: OL NT.

Supervision: MHH MH.

Visualization: MHH MH XFC NT OL.

Writing – original draft: MHH MH XFC NT OL.

Writing – review & editing: MHH MH XFC NT OL MS DE FMA CL SBM JFA RG ACS.

References

1. Newton RJ, Mclellan SL, Dila DK, Vineis JH, Morrison HG, Eren a M, et al. Sewage Reflects the Microbiomes of Human Populations. MBio. 2015; 6: 1–9.
2. Pina S, Buti M, Jardí R, Clemente-Casares P, Jofre J, Girones R. Genetic analysis of hepatitis A virus strains recovered from the environment and from patients with acute hepatitis. J Gen Virol. Microbiology Society; 2001; 82: 2955–63.
3. Hellmer M, Paxeus N, Magnius L, Enache L, Arnholm B, Johansson a., et al. Detection of Pathogenic Viruses in Sewage Provided Early Warnings of Hepatitis A Virus and Norovirus Outbreaks. Appl Environ Microbiol. 2014; 80: 6771–6781. doi: 10.1128/AEM.01981-14 PMID: 25172863
4. Wong K, Fong TT, Bibby K, Molina M. Application of enteric viruses for fecal pollution source tracking in environmental waters. Environ Int. Elsevier B.V.; 2012; 45: 151–164.
5. van den Berg H, Lodder W, van der Poel W, Vennema H, de Roda Husman AM. Genetic diversity of noroviruses in raw and treated sewage water. Res Microbiol. 2005; 156: 532–40. doi: 10.1016/j.resmic.2005.01.008 PMID: 15862452
6. Rosario K, Nilsson C, Lim YW, Ruan Y, Breitbart M. Metagenomic analysis of viruses in reclaimed water. Environ Microbiol. 2009; 11: 2806–2820. doi: 10.1111/j.1462-2920.2009.01964.x PMID: 19555373
7. Cantalupo PG, Calgua B, Zhao G, Hundesa A, Wier AD, Katz JP, et al. Raw sewage harbors diverse viral populations. MBio. 2011; 2.
8. Bibby K, Peccia J. Identification of viral pathogen diversity in sewage sludge by metagenome analysis. Environ Sci Technol. 2013; 47: 1945–51. doi: 10.1021/es305181x PMID: 23346855
9. Ng TFF, Marine R, Wang C, Simmonds P, Kapusinszky B, Bodhidatta L, et al. High variety of known and new RNA and DNA viruses of diverse origins in untreated sewage. J Virol. 2012; 86: 12161–75. doi: 10.1128/JVI.00869-12 PMID: 22933275
10. Bodewes R, van der Giessen J, Haagmans BL, Osterhaus ADME, Smits SL. Identification of multiple novel viruses, including a parvovirus and a hepevirus, in feces of red foxes. J Virol. 2013; 87: 7758–64. doi: 10.1128/JVI.00568-13 PMID: 23616657
11. Daly GM, Bexfield N, Heaney J, Stubbs S, Mayer AP, Palser A, et al. A viral discovery methodology for clinical biopsy samples utilising massively parallel next generation sequencing. PLoS One. 2011; 6.
12. Conceição-neto N, Zeller M, Lefrère H, Bruyn De P, Beller L. Modular approach to customise sample preparation procedures for viral metagenomics : a reproducible protocol for virome analysis. Nat Publ Gr. Nature Publishing Group; 2015; 1–14.
13. John SG, Mendez CB, Deng L, Poulos B, Kauffman AKM, Kern S, et al. A simple and efficient method for concentration of ocean viruses by chemical flocculation. Environ Microbiol Rep. 2011; 3: 195–202. doi: 10.1111/j.1758-2229.2010.00208.x PMID: 21572525
14. Calgua B, Mengewein a., Grunert a., Bofill-Mas S, Clemente-Casares P, Hundesa a., et al. Development and application of a one-step low cost procedure to concentrate viruses from seawater samples. J Virol Methods. 2008; 153: 79–83. doi: 10.1016/j.jviromet.2008.08.003 PMID: 18765255
15. Albinana-Gimenez N, Clemente-Casares P, Calgua B, Huguet JM, Courtois S, Girones R. Comparison of methods for concentrating human adenoviruses, polyomavirus JC and noroviruses in source waters and drinking water using quantitative PCR. J Virol Methods. 2009; 158: 104–109. doi: 10.1016/j.jviromet.2009.02.004 PMID: 19428577
16. Pei L, Rieger M, Lengger S, Ott S, Zawadsky C, Hartmann NM, et al. Combination of Cross fl ow Ultra fi ltration, Monolithic A ffi nity Filtration, and Quantitative Reverse Transcriptase PCR for Rapid Concentration and Quanti fi cation of Model Viruses in Water. 2012;
17. Hurwitz BL, Deng L, Poulos BT, Sullivan MB. Evaluation of methods to concentrate and purify ocean virus communities through comparative, replicated metagenomics. Environ Microbiol. 2013; 15: 1428–40. Available: http://doi.wiley.com/10.1111/j.1462-2920.2012.02836.x doi: 10.1111/j.1462-2920.2012.02836.x PMID: 22845467
18. Kunze A, Pei L, Elsässer D, Niessner R, Seidel M. High performance concentration method for viruses in drinking water. J Virol Methods. Elsevier B.V.; 2015; 222: 132–137.
19. Calgua B, Rodriguez-Manzano J, Hundesa A, Suñen E, Calvo M, Bofill-Mas S, et al. New methods for the concentration of viruses from urban sewage using quantitative PCR. J Virol Methods. 2013; 187: 215–221. doi: 10.1016/j.jviromet.2012.10.012 PMID: 23164995
20. Wesolowska-Andersen A, Bahl MI, Carvalho V, Kristiansen K, Sicheritz-Pontén T, Gupta R, et al. Choice of bacterial DNA extraction method from fecal material influences community structure as evaluated by metagenomic analysis. Microbiome. 2014; 2: 19. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4063427&tool=pmcentrez&rendertype=abstract doi: 10.1186/2049-2618-2-19 PMID: 24949196

21. Kennedy NA, Walker AW, Berry SH, Duncan SH, Farquarson FM, Louis P, et al. The impact of different DNA extraction kits and laboratories upon the assessment of human gut microbiota composition by 16S rRNA gene sequencing. PLoS One. 2014; 9: 1–9.
22. Petrich a, Mahony J, Chong S, Broukhanski G, Gharabaghi F, Johnson G, et al. Multicenter comparison of nucleic acid extraction methods for detection of severe acute respiratory syndrome coronavirus RNA in stool specimens. J Clin Microbiol. 2006; 44: 2681–8. doi: 10.1128/JCM.02460-05 PMID: 16891478
23. Iker BC, Bright KR, Pepper IL, Gerba CP, Kitajima M. Evaluation of commercial kits for the extraction and purification of viral nucleic acids from environmental and fecal samples. J Virol Methods. Elsevier B.V.; 2013; 191: 24–30.
24. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol. 2014; 12: 87. Available: http://www.biomedcentral.com/1741-7007/12/87 doi: 10.1186/s12915-014-0087-z PMID: 25387460
25. Newsome T, Li B-J, Zou N, Lo S-C. Presence of Bacterial Phage-Like DNA Sequences in Commercial Taq DNA Polymerase Reagents. J Clin Microbiol. 2004; 42: 2264–2267. doi: 10.1128/JCM.42.5.2264-2267.2004 PMID: 15131208
26. Naccache SN, Greninger AL, Lee D, Coffey LL, Phan T, Rein-Weston A, et al. The perils of pathogen discovery: origin of a novel parvovirus-like hybrid genome traced to nucleic acid extraction spin columns. J Virol. 2013; 87: 11966–77. doi: 10.1128/JVI.02323-13 PMID: 24027301
27. Rosseel T, Pardon B, De Clercq K, Ozhelvaci O, Van Borm S. False-positive results in metagenomic virus discovery: A strong case for follow-up diagnosis. Transbound Emerg Dis. 2014; 61: 293–299. doi: 10.1111/tbed.12251 PMID: 24912559
28. Wunderlich A, Torggler C, Elsässer D, Lück C, Niessner R, Seidel M. Rapid quantification method for Legionella pneumophila in surface water. Anal Bioanal Chem. Springer Berlin Heidelberg; 2016; 408: 2203–2213.
29. Millen HT, Gonnering JC, Berg RK, Spencer SK, Jokela WE, Pearce JM, et al. Glass Wool Filters for Concentrating Waterborne Viruses and Agricultural Zoonotic Pathogens. J Vis Exp. 2012; 6–11.
30. Leeuwen Van M, Williams MMW, Simon JH, Smits SL, Albert DM, Osterhaus E, et al. Human picobirnaviruses identified by molecular screening of diarrhea samples. J Clin Microbiol. 2010; 48: 1787–94. doi: 10.1128/JCM.02452-09 PMID: 20335418
31. Nordahl Petersen T, Rasmussen S, Hasman H, Carøe C, Bælum J, Charlotte Schultz A, et al. Metagenomic analysis of toilet waste from long distance flights; a step towards global surveillance of infectious diseases and antimicrobial resistance. Sci Rep. Nature Publishing Group; 2015; 5: 11444.
32. Rawsthorne H, Phister TG, Jaykus LA. Development of a fluorescent in situ method for visualization of enteric viruses. Appl Environ Microbiol. 2009; 75: 7822–7827. doi: 10.1128/AEM.01986-09 PMID: 19854924
33. Hernroth BE, Girones R, Allard AK, Hernroth BE, Girones R, Allard AK. Environmental Factors Influencing Human Viral Pathogens and Their Potential Indicator Organisms in the Blue Mussel, Mytilus edulis : the First Scandinavian Report Environmental Factors Influencing Human Viral Pathogens and Their Potential Indicator Organ. Appl Environ Microbiol. 2002; 68: 4523.
34. Wang D, Coscoy L, Zylberberg M, Avila PC, Boushey H a, Ganem D, et al. Microarray-based detection and genotyping of viral pathogens. Proc Natl Acad Sci U S A. 2002; 99: 15687–92. doi: 10.1073/pnas.242579699 PMID: 12429852
35. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011; 17: 10. Available: http://journal.embnet.org/index.php/embnetjournal/article/view/200/479
36. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. Oxford University Press; 2010; 26: 589–95. Available: http://bioinformatics.oxfordjournals.org/content/26/5/589.full
37. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25: 2078–9. doi: 10.1093/bioinformatics/btp352 PMID: 19505943
38. Bunge J, Woodard L, Böhning D, Foster JA, Connolly S, Allen HK. Estimating population diversity with CatchAll. Bioinformatics. Oxford University Press; 2012; 28: 1045–7.
39. R Core Team. R: A language and environment for statistical computing. R Found Stat Comput Vienna, Austria. Vienna, Austria; 2016; https://www.R-project.org/
40. Holm S. A Simple Sequentially Rejective Multiple Test Procedure. Scand J Stat Scand J Stat. 1979; 6: 65–70. Available: http://www.jstor.org/stable/4615733
41. Vu VQ. ggbiplot: A ggplot2 based biplot [Internet]. 2011. http://github.com/vqv/ggbiplot
42. Kolde R. pheatmap: Pretty Heatmaps [Internet]. 2015. https://cran.r-project.org/package=pheatmap

43. Wagg C, Bender SF, Widmer F, van der Heijden MGA. Soil biodiversity and soil community composition determine ecosystem multifunctionality. Proc Natl Acad Sci U S A. 2014; 111: 5266–70. doi: 10.1073/pnas.1320054111 PMID: 24639507

44. Bibby K, Viau E, Peccia J. Viral metagenome analysis to guide human pathogen monitoring in environmental samples. Lett Appl Microbiol. 2011; 52: 386–92. doi: 10.1111/j.1472-765X.2011.03014.x PMID: 21272046

45. Yang J, Yang F, Ren L, Xiong Z, Wu Z, Dong J, et al. Unbiased parallel detection of viral pathogens in clinical samples by use of a metagenomic approach. J Clin Microbiol. 2011; 49: 3463–9. doi: 10.1128/JCM.00273-11 PMID: 21813714

46. Glass RI, Parashar UD, Estes MK. Norovirus gastroenteritis. N Engl J Med. 2009; 361: 1776–85. doi: 10.1056/NEJMra0804575 PMID: 19864676

47. Tate JE, Burton AH, Boschi-Pinto C, Steele AD, Duque J, Parashar UD, et al. 2008 estimate of worldwide rotavirus-associated mortality in children younger than 5 years before the introduction of universal rotavirus vaccination programmes: a systematic review and meta-analysis. Lancet Infect Dis. Elsevier; 2012; 12: 136–41.

48. Stanaway JD, Flaxman AD, Naghavi M, Fitzmaurice C, Vos T, Abubakar I, et al. The global burden of viral hepatitis from 1990 to 2013: findings from the Global Burden of Disease Study 2013. Lancet. 2016;

49. Scallan E, Hoekstra RM, Angulo FJ, Tauxe R V., Widdowson M-A, Roy SL, et al. Foodborne Illness Acquired in the United States—Major Pathogens. Emerg Infect Dis. 2011; 17: 7–15. doi: 10.3201/eid1701.P11101 PMID: 21192848

50. Borozan I, Watt SN, Ferretti V. Evaluation of Alignment Algorithms for Discovery and Identification of Pathogens Using RNA-Seq. 2013;

51. Bibby K, Peccia J. Identification of Viral Pathogen Diversity in Sewage Sludge by Metagenome Analysis. 2013;

52. Mee ET, Preston MD, Participants S, Minor PD, Schepelmann S, Huang X, et al. Development of a candidate reference material for adventitious virus detection in vaccine and biologicals manufacturing by deep sequencing. Vaccine. 2016; 34: 2035–2043. doi: 10.1016/j.vaccine.2015.12.020 PMID: 26709640

53. Ding Y, Tang Y, Kwok CK, Zhang Y, Bevilacqua PC, Assmann SM. In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. Nature. Nature Publishing Group; 2013; 505: 696–700.

54. Li L, Deng X, Mee ET, Collot-teixeira S, Anderson R, Schepelmann S, et al. Comparing viral metagenomics methods using a highly multiplexed human viral pathogens reagent. J Virol Methods. Elsevier B. V.; 2015; 213: 139–146.

55. Kjartansdóttir KR, Friis-Nielsen J, Asplund M, Mollerup S, Mourier T, Jensen RH, et al. Traces of ATCV-1 associated with laboratory component contamination. Proc Natl Acad Sci. National Academy of Sciences; 2015; 112: E925–E926.

56. Woyke T, Sczyrba A, Lee J, Rinke C, Tighe D, Clingenpeel S, et al. Decontamination of MDA reagents for single cell whole genome amplification. PLoS One. 2011; 6.

Notes

```
#!/usr/bin/perl
#
# ####################################################################
#
#        Copyright (C) 2017 - Natàlia TIMONEDA SOLÉ
#
# This program is free software; you can redistribute it and/or modify
# it under the terms of the GNU General Public License as published by
# the Free Software Foundation; either version 2 of the License, or
# (at your option) any later version.
#
# This program is distributed in the hope that it will be useful,
# but WITHOUT ANY WARRANTY; without even the implied warranty of
# MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.  See the
# GNU General Public License for more details.
#
# You should have received a copy of the GNU General Public License
# along with this program; if not, write to the Free Software
# Foundation, Inc., 675 Mass Ave, Cambridge, MA 02139, USA.
#
# ####################################################################

use warnings;
use strict;
use Data::Dumper;

my  $stth = $dbh->prepare("SELECT taxon_id FROM taxons WHERE taxon_name=?;");
my ($keycontg,%countspec,@compts);
foreach $keybalt (keys %abalt){
    foreach $keyfam (keys %{$abalt{$keybalt}}){
        $row = $stth->execute($keyfam);
        while (my $reff = $stth->fetchrow_arrayref){
            $relation{$keyfam}= @{$reff}[0];
        }
        foreach $keyspec (keys %{$abalt{$keybalt}{$keyfam}}){
                my ($specname,$lnk) = $keyspec =~ /^(.+)\s\((\d+)\)$/;
                push @compt, $abalt{$keybalt}{$keyfam}{$keyspec};
        foreach my $num (@compt){
            $total1 = $total1 + $num;
        };
        $countspe{$keyspec} = $total1;
        $total1 = 0;
      @compt = ();
      push @compts ,$countspe{$keyspec};
        }; #species
        foreach my $num (@compts){
            $total = $total + $num;
        };
        $countfam{$keyfam} = $total;
        @compts = ();
        $total = 0;
        push @comptb, $countfam{$keyfam};
    }; #family
    foreach my $num (@comptb){
       $total2 = $total2 + $num;
    };
    $countbalt{$keybalt} = $total2;
    @comptb = ();
    $total2 = 0;
 };#baltimore

 foreach $keybalt (keys %abalt){
     foreach $keyfam (keys %{$abalt{$keybalt}}){
        foreach $keyspec (keys %{$abalt{$keybalt}{$keyfam}}){
            push @comptc, $abalt{$keybalt}{$keyfam}{$keyspec};
        };#especie
        foreach my $num (@comptc){
            $total3 = $total3 + $num;
        };#comptc
        @comptc = ();
        $total3 = 0;
    };#family
 };#baltimore
```