

Exploring the remarkable diversity of *Escherichia coli* Phages in the Danish Wastewater Environment, Including 91 Novel Phage Species

Nikoline S. Olsen¹, Witold Kot^{1,2*} Laura M. F. Junco² and Lars H. Hansen^{1,2,*}

¹ Department of Environmental Science, Aarhus University, Frederiksborgvej 399, Roskilde, Denmark;
niso@envs.au.dk

² Department of Plant and Environmental Sciences, University of Copenhagen, Thorvaldsensvej 40, 1871
Frederiksberg C, Denmark; lhha@plen.ku.dk, wk@plen.ku.dk

* Correspondence: lhha@plen.ku.dk Phone: +45 28 75 20 53, wk@plen.ku.dk Phone: +45 35 33 38 77

Funding: This research was funded by Villum Experiment Grant 17595, Aarhus University Research Foundation
AUFF Grant E-2015-FLS-7-28 to Witold Kot and Human Frontier Science Program RGP0024/2018.
Competing interests: The authors declare no competing interests.

Abstract: Phages drive bacterial diversity - profoundly influencing diverse microbial communities, from microbiomes to the drivers of global biogeochemical cycling. The vast genomic diversity of phages is gradually being uncovered as >8000 phage genomes have now been sequenced. Aiming to broaden our understanding of *Escherichia coli* (MG1655, K-12) phages, we screened 188 Danish wastewater samples (0.5 ml) and identified 136 phages of which 104 are unique phage species and 91 represent novel species, including several novel lineages. These phages are estimated to represent roughly a third of the true diversity of *Escherichia* phages in Danish wastewater. The novel phages are remarkably diverse and represent four different families *Myoviridae*, *Siphoviridae*, *Podoviridae* and *Microviridae*. They group into 14 distinct clusters and nine singletons without any substantial similarity to other phages in the dataset. Their genomes vary drastically in length from merely 5 342 bp to 170 817 kb, with an impressive span of GC contents ranging from 35.3% to 60.0%. Hence, even for a model host bacterium, in the go-to source for phages, substantial diversity remains to be uncovered. These results expand and underlines the range of *Escherichia* phage diversity and demonstrate how far we are from fully disclosing phage diversity and ecology.

Keywords: bacteriophage; wastewater; *Escherichia coli*; diversity; genomics; taxonomy; coliphage

1. Introduction

Phages are important ecological contributors, they renew organic matter supplies in nutrient cycles and drive bacterial diversity by enabling co-existence of competing bacteria by “Killing the winner” and by serving as genomic reservoirs and transport units [1,2]. Phage genomes are known to entail auxiliary metabolism genes (AMGs), toxins, virulence factors and even antibiotic resistance genes [3–7] and through lysogeny and transduction they can transfer metabolic traits to their hosts and even confer immunity against homologous phages [1]. Still, in spite of their ecological role, potential as antimicrobials and the fact that they carry a multitude of unknown genes with great potential for biotechnological applications, phages are vastly understudied. Less than 9000 phage genomes have now been published, and though the number increases

rapidly, we may have merely scratched the surface of the expected diversity [8]. It has been estimated that at least a billion bacterial species exist [9], hence only phages targeting a tiny fraction of potential hosts have been reported. Efforts to disclose the range and diversity of phages targeting a single host, have revealed a stunning display of diversity. The most scrutinized phage host is the *Mycobacterium smegmatis*, for which the Science Education Alliance Phage Hunters program has isolated more than 4700 phages and fully sequenced 680 phages which represent 30 distinct phage clusters [10,11]. This endeavour has provided a unique insight into viral and host diversity, evolution and genetics [12–15]. No other phage host has been equally targeted, but *Escherichia coli* phages have been isolated in fairly high numbers. The International Committee on Taxonomy of Viruses (ICTV) has currently recognised 158 phage species originally isolated on *E. coli* [16], while 2732 *Escherichia* phage genomes have been deposited in GenBank [17]. As phages are expected to have an evolutionary potential to migrate across microbial populations, host species may not be an ideal indicator of relatedness, but it serves as an excellent starting point to explore phage diversity. Hierarchical classification of phages is complicated by the high degree of horizontal gene transfer, consequently several classification systems have been proposed [18–20], and we may not yet have reached a point where it is reasonable to establish the criteria for a universal system [20]. Nonetheless, a system enabling a mutual understanding and exchange of knowledge is needed. Accordingly, we have in this study chosen to classify our novel phages according to the ICTV guidelines [21].

Here we aim to expand our understanding of *Escherichia* phage diversity. Earlier studies are few, have relied on more laborious and time-consuming methods, or are based on *in silico* analyses of already published phage genomes [8]. Accordingly, Korf *et al.*, (2019) isolated 50 phages on 29 individual *E. coli* strains, while Jurczak-Kurek *et al.*, (2016) isolated 60 phages on a single *E. coli* strain, both finding a broad diversity of *Caudovirales* and also representatives of novel phage lineages [22,23]. We hypothesized, that a high-throughput screening of nearly 200 samples in time-series, using a single host, would expand the number of known phages by facilitating the interception of the prevailing phage(s) of the given day.

62 2. Materials and Methods

63 The screening for *Escherichia* phages was performed with the microplate based *High-throughput screening*
 64 *method* as described in (not published). With the exception that lysates of wells giving rise to plaques were
 65 sequenced without further purification. In short, an overnight enrichment was performed in microplates with
 66 host culture, media and wastewater (0.5 ml per well), followed by filtration (0.45 µm), a purification step by
 67 re-inoculation, a second overnight incubation and filtration (0.45µm), and then a spot-test (soft-agar overlay)
 68 to indicate positive wells. All procedures were performed under sterile conditions.

69 2.1.1 Samples Bacteria and media

70 Inlet Wastewater samples (188) were collected (40-50 ml) in time-series (2-4 days spanning 1-3 weeks),
 71 from 48 Danish wastewater treatment facilities (rural and urban) geographically distributed on Zealand,
 72 Funen and in Jutland, during July and August 2017. The samples were centrifuged (9000 x g, 4 °C, 10 min) and
 73 the supernatant filtered (0.45 µm) before storage in aliquots (-20°C) until screening. The host bacterium is *E.*
 74 *coli* (MG1655, K-12), and the media Lysogeny Broth (LB), amended with CaCl₂ and MgCl₂ (final concentration
 75 50 mM). Collected phages were stored in SM-buffer [24] at 4°C.

76 2.1.2 Sequencing and genomic characterisation

77 DNA extractions, clean-up (ZR-96 Clean and Concentrator kit, Zymo Research, Irvine, CA USA) and
 78 sequencing libraries (Nextera® XT DNA kit, Illumina, San Diego, CA USA) were performed according to
 79 manufacturer's protocol with minor modifications as described in Kot et al., (2014) [25]. The libraries were
 80 sequenced as paired-end reads on an Illumina NextSeq platform with the Mid Output Kit v2 (300 cycles). The
 81 obtained reads were trimmed and assembled in CLC Genomics Workbench version 10.1.1. (CLC BIO, Aarhus,
 82 Denmark). Overlapping reads were merged with the following settings: mismatch cost: 2, minimum score: 15,
 83 gap cost: 3 and maximum unaligned end mismatches: 0, and then assembled *de novo*. Additional control
 84 assemblies were constructed using SPAdes version 3.12.0 [26]. Phage genomes were defined as contigs with

an average coverage $> \times 20$ and a sequence length $\geq 90\%$ of closest relative. Gene prediction and annotation were performed using a customized RASTtk version 2.0 [27] workflow with GeneMark [28], with manual curation and verification using BLASTP [29], HHpred [30] and Pfam version 32.0 [31], or de novo annotated using VIGA version 0.11.0 [32] based on DIAMOND searches (RefSeq Viral protein database) and HMMer searches (pVOG HMM database). All genomes were assessed for antibiotic resistance genes, bacterial virulence genes, type I, II, III and IV restriction modification (RM) genes and auxiliary metabolism genes (AMGs) using ResFinder 3.1 [33,34], VirulenceFinder 2.0 [35], Restriction-ModificationFinder 1.1 (REBASE) [36] and VIBRANT version 1.0.1 [37], respectively. The 104 unique phage genomes were aligned to viromes of BioProject PRJNA545408 in CLC Genomics Workbench and deposited in Genbank [17].

2.1.3 Genetic analyses

Nucleotide (NT) and amino acid (AA) similarities were calculated using tools recommended by the ICTV [38], i.e. BLAST [29] for identification of closest relative (BLASTn when possible, discontinuous megaBLAST (word size 16) for larger genomes) and Gegenees version 2.2.1 [39] for assessing phylogenetic distances of multiple genomes, for both NTs (BLASTn algorithm) and AAs (tBLASTx algorithm) a fragment size of 200 bp and step size 100 bp was applied. NT similarity was determined as percentage query cover multiplied by percentage NT identity. Novel phages are categorised according to ICTV taxonomy. The criterion of 95% DNA sequence similarity for demarcation of species was applied to identify novel species representatives and to determine uniqueness within the dataset. Evolutionary analyses for phylogenomic trees were conducted in MEGA7 version 2.1 (default settings) [40]. These were based on the large terminase subunit (*Caudovirales*), a gene commonly applied for phylogenetic analysis [41,42] and on the DNA replication gene (*gpA*) (*Microviridae*). The NT sequences were aligned by MUSCLE [43] and the evolutionary history inferred by the Maximum Likelihood method based on the Tamura-Nei model [44]. The trees with the highest log likelihood are shown. Pairwise whole genome comparisons were performed with Easyfig 2.2.2 [45] (BLASTn algorithm), these were curated by adding color-codes and identifiers in Inkscape version 0.92.2. The R package iNEXT [46,47] in

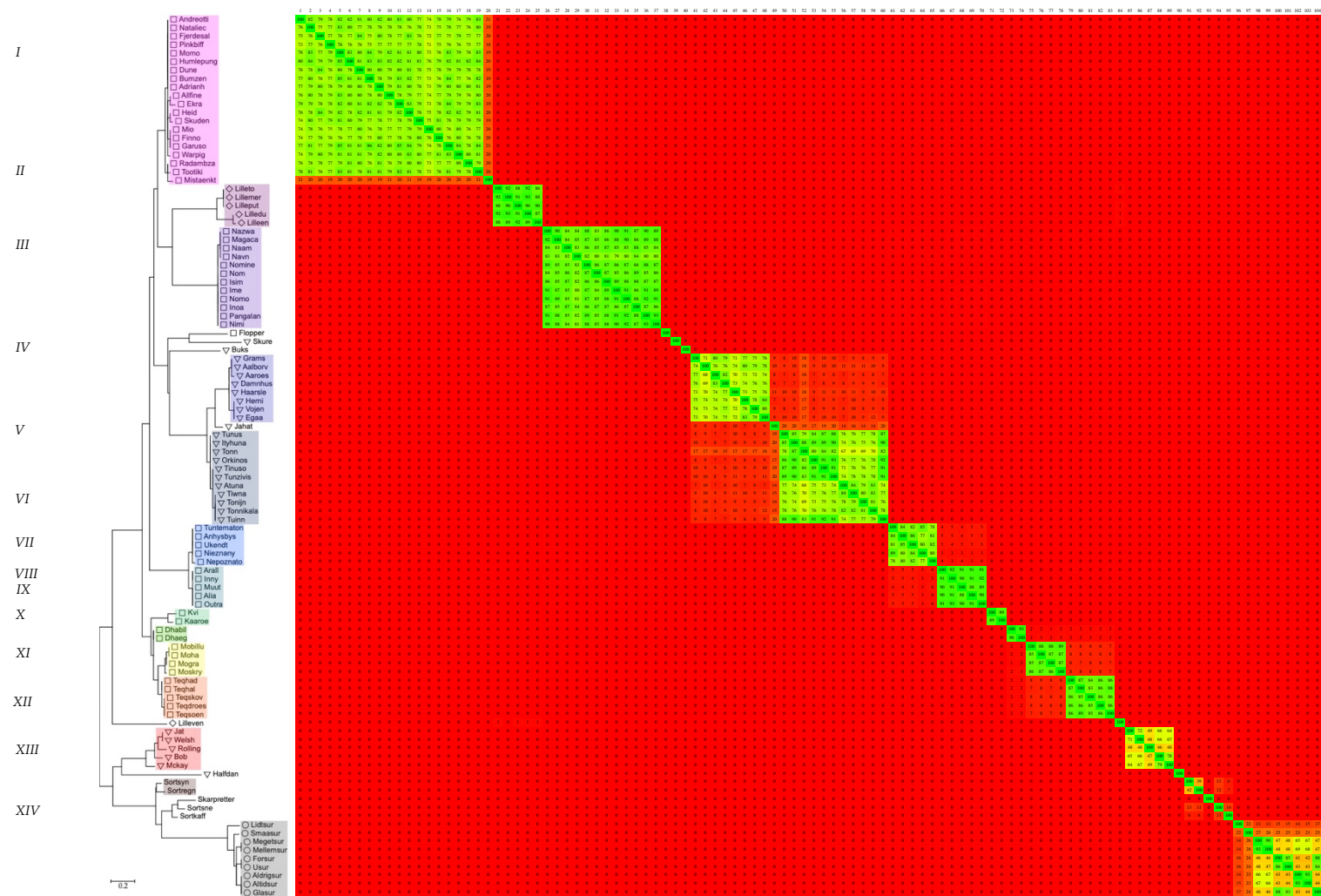
109 R studio version 1.1.456 [48] was used for rarefaction, species diversity ($q = 0$, datatype: incidence_raw),
 110 extrapolation hereof (estimatedD) and estimation of sample coverage. The visualisation of genome sizes and
 111 GC contents was prepared in Excel version 16.31.

112 3. Results and Discussion

113 3.2. Phylogenetics, taxonomy and species richness

114 The sequenced phages were analysed strictly *in silico*, focusing on their relatedness to known phages,
 115 their taxonomy and any distinctive characteristics. Based on the confirmed morphology of closely related
 116 phages, at least four different families are represented, *Myoviridae* (58%), *Siphoviridae* (25%), *Podoviridae* (7.4%)
 117 and even the single stranded DNA (ssDNA) *Microviridae* (5.9%) (Table 1). Five (3.7%) of the phages are so
 118 distinct that they could not with certainty be assigned to any family. A similar distribution was found by Korf
 119 *et al.*, (2019), i.e. 70% *Myoviridae*, 22% *Siphoviridae* and 8% *Podoviridae*, just as an analysis of the genomes of
 120 *Caudovirales* infecting *Enterobacteriaceae*, by Grose & Casjens (2014), also identified more clusters belonging to
 121 the *Myoviridae*, than the *Siphoviridae* and fewest of the *Podoviridae* [8,22]. Jurczak-Kurek *et al.*, (2016) found more
 122 siphoviruses than myovirus, but also found the *Podoviridae* to be the least abundant [23]. Nonetheless, these
 123 distributions are just as likely to be caused by culture or isolation methods, as they are to reflect true
 124 abundances. Based on DNA homology and phylogenetic analysis, the 104 unique phages identified in this
 125 study group into 14 distinct clusters and nine singletons, having a <1% Gegenees score between clusters and
 126 singletons, with the exceptions of *cluster IV*, *V* and *Jahat* which have inter-Gegenees scores of up to 20%, *cluster*
 127 *IX*, *X* and *XI* which have inter-Gegenees scores of 1-8% and *cluster XIII*, *Sortsne* and *Sortkaff* which have inter-
 128 Gegenees scores of 6-14% (Figure 1, Table 2). Excluding one phage in *cluster I* and two phages in *cluster XIV*,
 129 the intra-Gegenees score of all clusters is above 39% (Figure 1).

130



132 **Figure 1** Phylogenetic tree (Maximum log Likelihood: -1325.01, large terminase subunit (*Caudovirales*) or DNA replication protein *gpA* (*Microviridae*)), scalebar: substitutions
133 per site, morphology is indicated by symbols (*Myoviridae*: □, *Siphoviridae*: ▽, *Podoviridae*: ○, *Microviridae*: ◇) and phylogenomic nucleotide distances (Gegenes, BLASTn:
134 fragment size: 200, step size: 100, threshold: 0%)

Table 1 List of 104 unique *Escherichia* phages identified in 94 Danish wastewater samples. Novelty (*) is based on the 95% demarcation of species. Similarity is sequence identity (%) times sequence coverage (%) to closest related (Blastn). Taxonomy is based on similarity (BLASTn) to closest related and the ICTV Master Species list.

Escherichia Phage	Genome (bp)	ORFs (n)	tRNAs (n)	GC (%)	Novel	Sample	Location	Cluster	Taxonomy	Similarity (%)	Accession number
Tootiki	88257	128	22	39	*	3D	Avedøre	I	<i>Felixounavirus</i>	90,2	MN850647
Mio	83431	121	18	39,1	*	13C	Hadsten	I	<i>Felixounavirus</i>	89,7	MN850631
Allfine	86963	125	20	39	*	14A	Hammel	I	<i>Felixounavirus</i>	91,2	MN850633
Bumzen	87360	126	20	39,1	*	14B, 14D, 15A	Hammel, Hinnerup	I	<i>Felixounavirus</i>	92,5	MN850635
Dune	88511	129	20	39	*	19C	Tørring	I	<i>Felixounavirus</i>	91,5	MN850636
Warpig	86106	127	17	39	*	20A	Helsingør	I	<i>Felixounavirus</i>	93	MN850637
Radambza	86702	127	19	38,9	*	20D	Helsingør	I	<i>Felixounavirus</i>	91,6	MN850639
Ekra	87282	128	20	38,9	*	21C	Herning	I	<i>Felixounavirus</i>	92,9	MN850644
Humlepung	85311	119	19	39,1	*	12B, 25B, 37A	Drøbro, Gram, Bogense	I	<i>Felixounavirus</i>	92,1	MN850564
Finno	87554	129	20	38,9	*	31C	Skovby	I	<i>Felixounavirus</i>	89,7	MN850619
Garuso	85798	130	20	38,9	*	29A, 33A	Nustrup, Sommersted	I	<i>Felixounavirus</i>	90,9	MN850566
Momo	88168	130	20	39	*	38D	Hofmangave	I	<i>Felixounavirus</i>	90,7	MN850580
Heid	87590	126	20	39	*	8C, 24B, 24C, 34D, 37D, 38B	Esbjerg Ø, Bevtøft, Vojens, Bogense, Hofmangave	I	<i>Felixounavirus</i>	91,2	MN850577
Skuden	87263	131	20	38,9	*	41D	Odense NV	I	<i>Felixounavirus</i>	91,1	MN850585
Pinkbiff	88814	129	20	39	*	46A	Marselisborg	I	<i>Felixounavirus</i>	93,9	MN850603
Fjerdesal	87715	128	21	39	*	3A, 39A, 46C	Avedøre, Hårslev, Marselisborg	I	<i>Felixounavirus</i>	90,6	MN850605
Andreotti	83391	117	20	39,2	*	47B	Egå	I	<i>Felixounavirus</i>	91,9	MN850610
Nataliec	89137	134	20	39	*	47D, 48C	Egå, Viby	I	<i>Felixounavirus</i>	90,3	MN850611
Adrianh	88226	128	19	38,9	*	42C, 48B	Otterup, Viby	I	<i>Felixounavirus</i>	91,1	MN850614

Escherichia Phage	Genome (bp)	ORFs (n)	tRNAs (n)	GC (%)	Novel	Sample	Location	Cluster	Taxonomy	Similarity (%)	Accession number
Mistaenkt	86664	128	22	47,2	*	6A	Kolding	I	<i>Suspvirus</i>	91,1	MN850587
Nimi	137039	213	5	43,7	*	11C	Varde	III	<i>Vequintavirus</i>	93,3	MN850626
Navn	141707	224	4	43,6	*	21B	Herning	III	<i>Vequintavirus</i>	91,1	MN850642
Nomine	137991	220	5	43,6	*	23A	Lemvig	III	<i>Vequintavirus</i>	91,5	MN850649
Naswa	138583	222	5	43,6	*	45A	Ålborg Ø	III	<i>Vequintavirus</i>	93,1	MN850595
naam	137129	215	5	43,7		31D	Skovby	III	<i>Vequintavirus</i>	94,5	MN850630
Ime	137114	217	5	43,6	*	12C, 36B	Drøsbo	III	<i>Vequintavirus</i>	93,1	MN850576
Magaca	135826	217	5	43,6		48A	Viby	III	<i>Vequintavirus</i>	96	MN850612
Nom	136114	213	5	43,6	*	28D	Jegerup	III	<i>Vequintavirus</i>	92,6	MN850646
Isim	138289	219	5	43,6	*	31B	Skovby	III	<i>Vequintavirus</i>	93,8	MN850597
Nomo	137702	218	5	43,7	*	38D	Hofmansgave	III	<i>Vequintavirus</i>	93,3	MN850578
Inoa	138710	220	5	43,6	*	44C	Ålborg V	III	<i>Vequintavirus</i>	92	MN850593
Pangalan	136944	215	5	43,7		2A, 45D, 48D	Grindsted, Egå, Viby	III	<i>Vequintavirus</i>	94,8	MN850627
Tuntematon	150473	279	11	39,1	*	7B, 7C	Esbjerg V	VI	<i>Myoviridae</i>	89,6	MN850618
Anhysbys	149335	271	11	39,1	*	22C	Hillerød	VI	<i>Myoviridae</i>	91,5	MN850648
Ukendt	150947	266	11	39	*	32D	Skrydstrup	VI	<i>Myoviridae</i>	88,7	MN850565
Nepoznato	151514	265	10	38,9	*	19D, 35A, 48A, 48D	Tørring, Årøsund, Viby	VI	<i>Myoviridae</i>	85,6	MN850571
Nieznany	144998	254	11	39,1	*	45C	Ålborg Ø	VI	<i>Myoviridae</i>	88,9	MN850598
Muut	146307	243	13	37,4	*	35B	Årøsund	VII	<i>Myoviridae</i>	92	MN850573
Alia	147009	246	13	37,5	*	13D	Hadsten	VII	<i>Myoviridae</i>	93,1	MN850632
Outra	145482	246	13	37,4	*	22A	Hillerød	VII	<i>Myoviridae</i>	93,8	MN850645
Inny	147483	247	13	37,4	*	45D	Ålborg Ø	VII	<i>Myoviridae</i>	92,4	MN850601
Arall	145715	242	13	37,4		41B	Odense NØ	VII	<i>Myoviridae</i>	94,6	MN850584
Kvi	163673	266	-	40,5	*	48C	Viby	VIII	<i>Krischvirus</i>	94,2	MN850615
Kaaro	163719	267	-	40,5		35D	Årøsund	VIII	<i>Krischvirus</i>	94,7	MN850574

Escherichia	Genome	ORFs	tRNAs	GC	Novel	Sample	Location	Cluster	Taxonomy	Similarity	Accession
Phage	(bp)	(n)	(n)	(%)						(%)	number
Dhabil	165644	266	3	39,5	*	1B	Billund	IX	<i>Dhakavirus</i>	87,5	MN850621
Dhaeg	170817	278	3	39,4	*	47A, 47C	Egå	IX	<i>Dhakavirus</i>	87,4	MN850609
Mogra	168724	263	2	37,7	*	25C	Gram	X	<i>Mosigvirus</i>	91,1	MN850579
Mobillu	163063	255	2	37,7		1B	Billund	X	<i>Mosigvirus</i>	94,5	MN850622
Moha	168676	267	2	37,6		26A	Haderslev	X	<i>Mosigvirus</i>	94,8	MN850590
Moskry	169410	269	2	37,6	*	32C	Skrydstrup	X	<i>Mosigvirus</i>	93,2	MN850651
Teqskov	165017	257	6	35,4	*	10D	Skovlund	XI	<i>Tequatrovirus</i>	91,7	MN895437
Teqdroes	166833	269	10	35,4	*	12D	Drøbro	XI	<i>Tequatrovirus</i>	88,6	MN895438
Teqhad	167892	270	10	35,3	*	26B	Haderslev	XI	<i>Tequatrovirus</i>	90,1	MN895434
Teqhal	168070	266	11	35,4	*	27A, 27D	Halk	XI	<i>Tequatrovirus</i>	93,9	MN895435
Teqsoen	166468	268	10	35,5	*	43B	Søndersø	XI	<i>Tequatrovirus</i>	91,7	MN895436
Flopper	52092	78	1	44,2	*	44D	Ålborg V	-	<i>Myoviridae</i>	87	MN850594
Damhaus	51154	89	-	44,1	*	4B	Damhusåen	IV	<i>Hanriverivirus</i>	85,8	MN850602
Herni	50971	89	-	44,1	*	21B, 30D	Herning, Over Jerstal	IV	<i>Hanriverivirus</i>	87,6	MN850640
Grams	49530	83	-	44,1	*	25B	Gram	IV	<i>Hanriverivirus</i>	87,1	MN850567
Aaroes	51662	92	-	44,1	*	35B	Årøsund	IV	<i>Hanriverivirus</i>	83	MN850572
Aalborv	46660	79	-	43,9	*	44B	Ålborg V	IV	<i>Hanriverivirus</i>	86,9	MN850591
Haarsle	48613	85	-	44	*	39B, 45C	Hårslev, Ålborg Ø	IV	<i>Hanriverivirus</i>	87,1	MN850600
Egaa	51643	87	-	44,1	*	47A	Egå	IV	<i>Hanriverivirus</i>	89,7	MN850607
Vojen	50709	86	-	44,1	*	34B	Vojens	IV	<i>Hanriverivirus</i>	89,7	MN850569
Tiwna	51014	85	-	44,6	*	21B	Herning	V	<i>Tunavirinae</i>	87,2	MN850643
Tonijn	51627	86	-	44,6	*	32B, 32C	Skrydstrup	V	<i>Tunavirinae</i>	88,4	MN850641
Tonnikala	51277	86	-	44,8	*	48A	Viby	V	<i>Tunavirinae</i>	86,4	MN850613
Atuna	50732	88	-	44,6	*	7B	Esbjerg V	V	<i>Tunavirinae</i>	84,9	MN850620
Tunus	51111	87	-	44,8	*	20A	Helsingør	V	<i>Tunavirinae</i>	93,7	MN850638

Escherichia Phage	Genome (bp)	ORFs (n)	tRNAs (n)	GC (%)	Novel	Sample	Location	Cluster	Taxonomy	Similarity (%)	Accession number
Orkinos	49798	81	-	44,6	*	30B, 30D	Over Jerstal	V	<i>Tunavirinae</i>	91,3	MN850586
Ityhuna	50768	86	-	44,7	*	39D	Hårslev	V	<i>Tunavirinae</i>	93,3	MN850582
Tonn	51012	87	-	44,5	*	8C, 45C	Esbjerg Ø, Ålborg Ø	V	<i>Tunavirinae</i>	94	MN850596
Tinuso	50856	86	-	44,8		14A	Hammel	V	<i>Tunavirinae</i>	97,3	MN850634
Tunzivis	50596	84	-	44,6		46B	Marselisborg	V	<i>Tunavirinae</i>	94,5	MN850604
Tuinn	50505	86	-	44,7		46D	Marselisborg	V	<i>Tunavirinae</i>	94,8	MN850606
Jahat	51101	87	-	45,7	*	35A	Vojens	-	<i>Tunavirinae</i>	68,5	MK552105
Bob	45252	63	-	54,5	*	12C	Drøsbro	XII	<i>Dhillonvirus</i>	88,6	MN850628
Mckay	44443	63	-	54,5	*	13B	Hadsten	XII	<i>Dhillonvirus</i>	83,8	MN850629
Jat	44417	63	-	54,5	*	23C	Lemvig	XII	<i>Dhillonvirus</i>	89,4	MN850650
Rolling	46017	64	-	54,2	*	29D	Nustrup	XII	<i>Dhillonvirus</i>	80,2	MN850575
Welsh	45207	62	-	54,6	*	23A, 43D	Lemvig, Sønder sø	XII	<i>Dhillonvirus</i>	83,8	MN850589
Buks	40308	62	-	49,7	*	1A	Billund	-	<i>Jerseyvirus</i>	91,3	MN850616
Skure	59474	92	-	44,6	*	23C	Lemvig	-	<i>Seuratvirus</i>	90,4	MK672798
Halfdan	42858	57	-	53,7	*	34D	Vojens	-	<i>Siphoviridae</i>	28,8	MH362766
Lilleen	5342	6	-	46,9	*	33A	Sommersted	II	<i>Gequatrovirus</i>	93,8	MK629526
Lilleput	5490	6	-	47	*	12D	Drøsbro	II	<i>Gequatrovirus</i>	93,4	MK629525
Lilleto	5492	6	-	46,8	*	43C, 43D, 44B	Sønder sø, Ålborg V	II	<i>Gequatrovirus</i>	92,7	MK629529
Lilledu	5483	6	-	47,2	*	25C	Gram	II	<i>Gequatrovirus</i>	92,6	MK791318
Lillemer	5492	6	-	47,1		45C	Ålborg Ø	II	<i>Gequatrovirus</i>	94,5	
Lilleven	6090	9	-	44,4	*	11B	Varde	-	<i>Alphatrevirus</i>	93,9	MK629527
Sortsyn	42116	61	-	59	*	11B	Varde	XIII	<i>Unclassified</i>	92,3	MN850623
Sortregn	38200	53	-	59,3		43D	Sønder sø	XIII	<i>Unclassified</i>	97,3	MN850588
Skarpretter	42042	63	-	55,8	*	15A	Hinnerup	-	<i>Unclassified</i>	37,9	MK105855

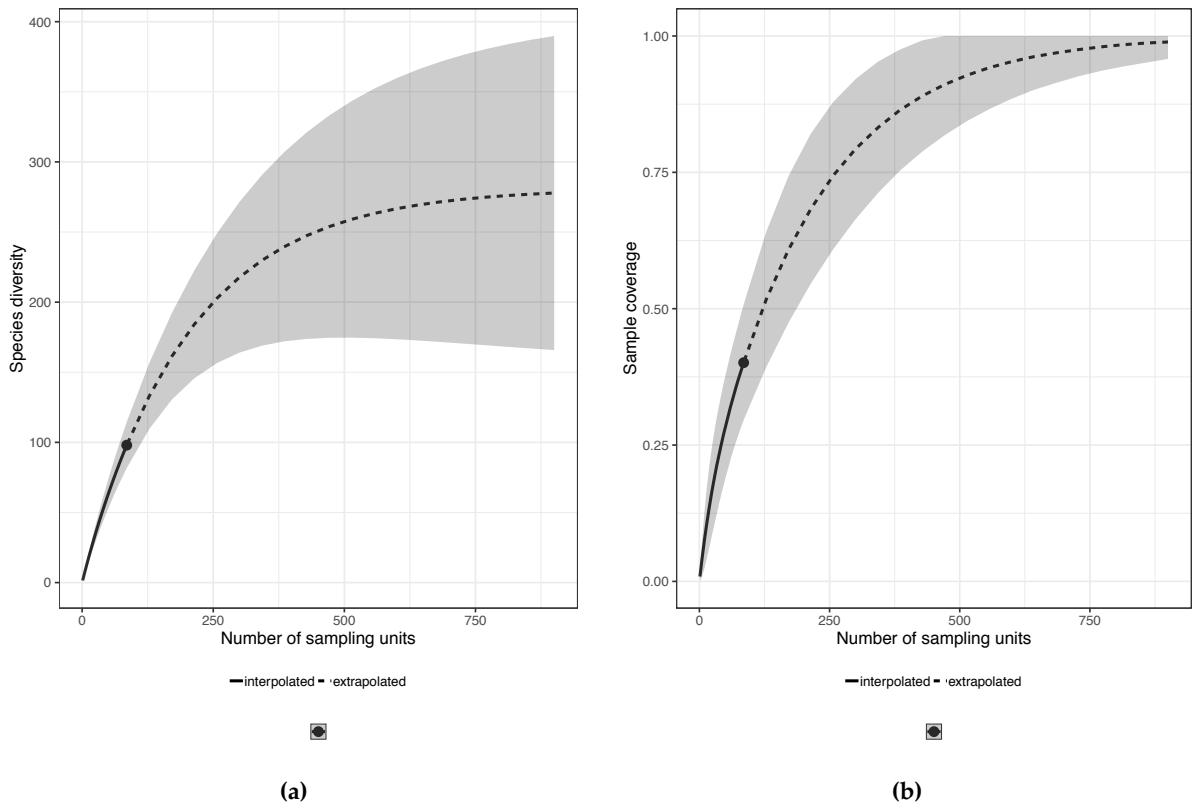
Escherichia Phage	Genome (bp)	ORFs (n)	tRNAs (n)	GC (%)	Novel	Sample	Location	Cluster	Taxonomy	Similarity (%)	Accession number
Sortkaff	42538	61	-	59,5	*	39B	Hårslev	-	Unclassified	89,8	MN850581
Sortsne	41912	62	-	60	*	40A	Odense NV	-	Unclassified	67,6	MK651787
Aldrigsur	42379	55	-	55,7	*	44B	Ålborg V	XIV	Autographvirinae	71,9	MN850592
Altidsur	42197	53	-	55,7	*	33D	Sommersted	XIV	Autographvirinae	71,8	MN850568
Forsur	42476	56	-	55,4	*	48D	Viby	XIV	Autographvirinae	72	MN850617
Glasur	42507	56	-	55,4	*	40D	Odense NV	XIV	Autographvirinae	72,3	MN850583
Lidtsur	42291	56	-	54,6	*	30B	Over Jerstal	XIV	Autographvirinae	69	MK629528
Megetsur	42132	54	-	55,8	*	31B	Skovby	XIV	Autographvirinae	73,1	MN850608
Mellemsur	40770	50	-	55,8	*	34D	Vojens	XIV	Autographvirinae	76,4	MN850570
Smaasur	41110	50	-	55,4	*	11C	Varde	XIV	Autographvirinae	93,3	MN850625
Usur	41906	51	-	55,4	*	27A, 27D	Halk	XIV	Autographvirinae	73,3	MN850624

137

138

139 The high-throughput screening method favours easily culturable plaque-forming lytic phages. Still, we identified 104 unique Escherichia phages of
140 which only 16% were $\geq 95\%$ similar (BLAST) to already published phages (Table 2). Phages were identified in wastewater samples from 43 of the 48
141 investigated treatment facilities. From the majority of positive samples (58) a single phage was sequenced, however in some samples the lysate held
142 more than one phage. Twenty-five of the lysates held two phages, eight lysates held three phages and one had as many as four phages. Of the 104
143 unique phages, 91 represent novel species (Table 1, 2). Of these, 51 differed by $\geq 10\%$ from published phage genomes and some have NT similarities as
144 low as 29% (Table 2).

145 These newly sequenced phages represent a substantial quota of divergent lytic *Escherichia* phages in Danish
146 wastewater, but are still far from disclosing the true diversity hereof (Figure 2). An extrapolation of species
147 richness ($q = 0$) predicts a total of 292 distinct species (requiring a sample size of ~900 phages). The relatively
148 small sample-size in this study ($n = 136$), may subject the estimation to a large prediction bias. The sampling-
149 method also introduces a bias by selecting for abundance and burst size, thereby potentially underestimating
150 diversity. Nonetheless, the results provide an indication of the minimal diversity of lytic *Escherichia* (MG1655,
151 K-12) phages in Danish wastewater, estimated to be as a minimum be in the range of 160 to 420 unique phage
152 species (Figure 2b). The novelty and diversity of these wastewater phages is truly remarkable and verifies our
153 hypothesis, as well as the efficiency of the *High-throughput Screening Method* for exploring diverse phages of a
154 single host (not published).



155 **Figure 2 (a)** Sample-size-based rarefaction and extrapolation curve with confidence intervals (0.95). **(b)** Sample
156 completeness curve with confidence intervals (0.95).

157

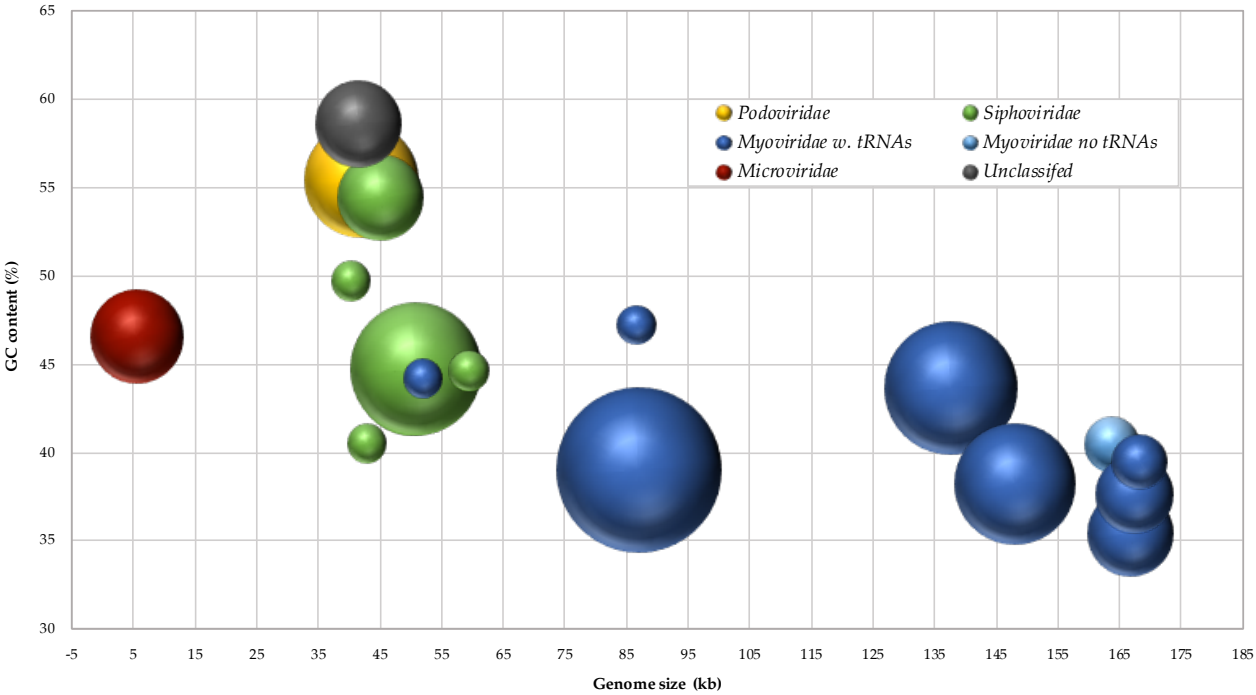
Table 2 Taxonomic distribution of phages identified in 94 Danish wastewater samples, based on similarity to closest related and the ICTV Master Species list. 1. $\leq 95\%$ similarity to other phages in the dataset. 2. $\leq 95\%$ similarity to other phages in the dataset and to published phage genomes.

Taxonomy*	total	unique ¹	novel ²
<i>Caudovirales; Myoviridae; Ounavirinae; Felixounavirus</i>	33	19	19
<i>Caudovirales; Myoviridae; Ounavirinae; Suspvirus</i>	1	1	1
<i>Caudovirales; Myoviridae; Tevenvirinae; Krischvirus</i>	2	2	1
<i>Caudovirales; Myoviridae; Tevenvirinae; Dhakavirus</i>	3	2	2
<i>Caudovirales; Myoviridae; Tevenvirinae; Mosigvirus</i>	4	4	2
<i>Caudovirales; Myoviridae; Tevenvirinae; Tequatrovirus</i>	6	5	5
<i>Caudovirales; Myoviridae; Vequintavirinae; Vequintavirus</i>	15	12	9
<i>Caudovirales; Myoviridae</i>	15	11	10
<i>Caudovirales; Podoviridae; Autographivirinae</i>	10	9	9
<i>Caudovirales; Siphoviridae; Dhillonvirus</i>	6	5	5
<i>Caudovirales; Siphoviridae; Guernseyvirinae; Jerseyvirus</i>	1	1	1
<i>Caudovirales; Siphoviridae; Seuratvirus</i>	1	1	1
<i>Caudovirales; Siphoviridae; Tunavirinae; Hanriovirus</i>	10	8	8
<i>Caudovirales; Siphoviridae; Tunavirinae</i>	15	12	8
<i>Caudovirales; Siphoviridae</i>	1	1	1
<i>Microviridae; Bullavirinae; Alphetvirus</i>	1	1	1
<i>Microviridae; Bullavirinae; Gequatrovirus</i>	7	5	4
Unclassified bacterial viruses	5	5	4
Total	136	104	91

3.2. Phage genome characteristics

The sequenced phage genomes range in sizes from 5 342 bp of the *Microviridae* lilleen to a span in the *Caudovirales* from 38 200 bp of the unclassified sortregn to 170 817 bp of the *Dhakavirus* dhaeg (Figure 3, Table 2). GC contents also vary greatly, from only 35.3% (*Tequatrovirus* teqhad) and up to 60.0% (the unclassified sortsne) (Figure 3, Table 2), heavily diverging from the host GC content of 50.79%. Phages often have a lower GC contents than their host [49], as observed for the majority (81%) of the wastewater phages (Table 2). A lower GC content of phage genomes tends to correlate with an increase in genome size [50], as is also the case for the *Caudovirales* in this study (Table 2). Rocha & Danchin (2002) hypothesized that phages, along with plasmids and insertion sequences, can be considered intracellular pathogens, and like host dependent bacterial pathogens and symbionts, they experience competition for the energetically expensive and less abundant GTP and CTP and as a consequence develop genomes with comparatively higher AT contents [49]. However, the

174 differences in AT richness reported by Rocha & Danchin (2002) for dsDNA and isometric ssDNA phages were
175 merely 4.2% and 5% [49], while the dsDNA and isometric ssDNA phages in this study deviate from their
176 isolation host by having up to 31% higher and up to 19% lower AT content (Table 2). Accordingly, the
177 considerable differences in GC/AT contents may instead primarily be a reflection of past host relations [14].
178 The sequencing of lysates and not only plaquing phages, may have contributed in enabling the capture of such
179 a broad GC-content and overall diversity.



180
181 **Figure 3** Bubble-diagram of the 104 unique *Escherichia* phages displaying genome size- and GC-content distribution. Area
182 of bubbles indicates number of phages. The yellow bubble is *Podoviridae* phages, green ones are *Siphoviridae* phage clusters,
183 dark-blue ones are *Myoviridae* phage clusters with tRNAs, the light-blue bubble is *Myoviridae* phages without tRNAs, the
184 red bubble is *Microviridae* phages and the grey one is unclassified phages.

186 The genome screening algorithms identified no sequences coding for homologs of known virulence or
187 antibiotic resistance genes. Though not a definitive exclusion, this interprets as a reduced risk of presence, a
188 preferable trait for phage therapy application. Currently available tools for AMG screening of viromes did not
189 provide a comprehensive and exclusive assessment of the AMG pool in the dataset. The majority of genes
190 identified are not AMGs, but code for phage DNA modification pathways (Table S1). The function of some of

the suggested AMGs is unknown, these include a nicotinamide phosphoribosyltransferase (NAMPT) present in *cluster I* (*Felixovirus* and *Suspovirus*) and *VII* (unclassified *Myoviridae*), a 3-deoxy-7-phosphoheptulonate (DAHP) synthase present in *cluster X* (*Mosigvirus*) and a complete dTDP-rhamnose biosynthesis pathway present in *cluster VI* and *VII* (unclassified *Myoviridae*). The presence of a dTDP-rhamnose biosynthesis pathway in the DNA metabolism region of phage genomes is peculiar, one possible explanation is, that these phages utilize rhamnose for glycosylation of hydroxy-methylated nucleotides in the same manner as the T4 generated glucosyl-hmC [51]. Dihydrofolate reductase, identified in six of the *Myoviridae* clusters, is involved in thymine nucleotide biosynthesis, but is also a structural component in the tail baseplate of T-even phages [52]. Multiple verified DNA modification genes were identified. Methyltransferases, some putative, were detected in all *Myoviridae* of *cluster III* (*vequintavirus*), *VI* (unclassified), *VII* (unclassified), *VIII* (*Krischvirus*), in all *Siphoviridae* of *cluster IV* (*Hanrivervirus*) and in the unclassified phages of *cluster XIII*. Indeed, *cluster XIII* and the singletons *skarpretter*, *sortsne*, and *sortkaff* code for both DNA N-6-adenine-methyltransferases (*dam*) and DNA cytosine methyltransferases (*dcm*). Finally, two novel epigenetic DNA hypermodification pathways were identified. The *mosigviruses* of *cluster X* code for arabinosylation of hmC [51], while the novel *seuratvirus* *Skure* codes for the recently verified complex 7-deazaguanine DNA hypermodification system [53,54].

3.3. Forty-eight novel *Myoviridae* phages species

The *Myoviridae* genomes (56 unique, 49 novel) represents the greatest span in genome sizes in this study, from the unclassified flopper of 52092 bp to the *dhakavirus* *dhaeg* of 170817 bp), all, except the *krischviruses*, code for tRNAs (Figure 3, Table 2). The *Myoviridae* group into nine distinct clusters and four singletons, representing at least three subfamilies; *Tevenvirinae*, *Vequintavirinae* and *Ounavirinae* in addition to 11 unclassified *Myoviridae* (*cluster VI & VII*) (Figure 1, Table 2). The eleven novel *Tevenvirinae* distributes into four distinct genera, two of the *Krischvirus* (*cluster VIII*, 164kb, 40.5% GC, no tRNAs), two of the *Dhakavirus* (*cluster IX*, 166-170 kb, 39.4-39.5% GC, 3 tRNAs), two of the *Mosigvirus* (*cluster X*, 169 kb, 37.6-37.7% GC, 2 tRNAs) and five of the *Tequatrovirus* (*cluster XI*, 165-168 kb, 35.3-35.5% GC, 6-11 tRNAs), while the nine novel

215 *Vequintavirinae* are all vequintaviruses (*cluster III*, 136-142 kb, 43.6-43.7% GC, 5 tRNAs) closely related (91.1-
216 93.8%, BLAST) to classified species. The vequintaviruses were identified in samples from 12 treatment
217 facilities. Only reads from two samples of a dataset of human gut viromes based on timeseries of faecal
218 samples from 10 healthy persons [55] mapped to the wastewater phages, and only to vequintaviruses. These
219 reads covered 43-54% and 13-26% of all the *Vequintavirus* genomes except pangalan and navn, respectively.
220 This suggests that the vequintaviruses are related to entero-phages. All but one of the *Ounavirinae* (*cluster I*,
221 82-89 kb, 38.9-39.2% GC, 17-22 tRNAs) are felixounaviruses (89.7-93.9%, BLAST) with an intra-Gegenees score
222 of 71-86% (Figure 1). The *Felixounavirus* is a relatively large genus with 17 recognized species. In this study,
223 felixounaviruses were identified in samples from 23 of the 48 facilities, indicating that they are both ubiquitous
224 in Danish wastewater, numerous and easily cultivated. The last ounavirus, mistaenkt (86.7 kb, 47.2% GC, 22
225 tRNAs) is a *Suspovirus*. The five *cluster VI* phages (144.9-151.5 kb, 39±0.1% GC, 10-11 tRNAs) belong to a
226 phylogenetic distinct clade, the recently proposed '*Phapecoctavirus*', and have substantial similarity (86-90%,
227 BLAST) with the anticipated type species *Escherichia* phage phAPEC8 (JX561091) [22,56]. *Cluster VII* have
228 significantly ($p = 0.038$) larger genomes (145.8-147.5 kb) with marginally, though not significantly ($p = 0.786$),
229 lower GC contents of 37.4-37.5%, all have 13 tRNAs (Table 2) and code for NAMPT not present in *cluster VI*
230 (Table S1). As a group, *cluster VII* are even more homogeneous than *cluster VI* and all are closely related (92-
231 95%, BLAST) to the same five unclassified *Enterobacteriaceae* phages vB_Ecom_PHB05 (MF805809),
232 vB_vPM_PD06 (MH816848), ECGD1 (KU522583), phi92 (NC_023693) and vB_vPM_PD114 (MH675927)
233 [57,58], with whom they represent a novel unclassified genera. The distinctive and novel singleton flopper
234 only shares NT similarity (36-87%, BLAST) with six other phages, the *Escherichia* phages ST32 (MF04458) [59]
235 and phiEcoM_GJ1 (EF460875) [60], the *Erwinia* phages Faunus (MH191398) [61] and vB_EamM-Y2
236 (NC_019504) [62], and the *Pectobacterium* phages PM1 (NC_023865) [63] and PP101 (KY087898). This group of
237 unclassified phages all have genomes of 52.1-56.6 kb with 43.6-44.9% GC and while only flopper, phiEcoM_GJ1
238 and PM1 code for a single tRNA, all of them have exclusively unidirectional coding sequences (CDSs) and
239 code for RNA polymerases, characteristics of the *Autographivirinae* of the *Podoviridae* [64]. However, the

verified morphology of ST32, phiEcoM_GJ1, PM1 and PP101, icosahedral head, neck and a contractile tail with tail fibres, classifies them as myoviruses [59,60,62,63]. Based on NT similarity and genome synteny, these seven phages belong to the same, not yet classified, peculiar lineage first described by Jamalludeen *et al.*, (2008) [60].

3.4. Five novel Microviridae phages species

The singleton lilleven (6.1 kb, 44.4% GC, no tRNA) and the five (four novel) *cluster II* phages (5.3-5.5 kb, 46.9-47.2% GC, no tRNAs) are all *Microviridae*, a family of small ssDNA non-enveloped icosahedral phages (Table 2). Lilleven is closely related to (93.9%, BLAST), have pronounced gene pattern synteny and high AA similarities (89-90%), with the Alphatremicrovirus Enterobacteria phage St1 (Figure S2), and is a novel species of the genus *Alphatremicrovirus*, subfamily *Bullavirinae*. The *cluster II* phages comprise four distinct species, only differing by single nucleotide polymorphisms and in non-coding regions (Figure S2). They have high intra-Gegenees score (88-92%) and share genomic organisation and extensive NT similarity (92.6-94.5%, BLAST) with the unclassified microvirus Escherichia phage SECphi17 (LT960607), primarily differing by single nucleotide polymorphisms and in noncoding regions (Figure S2, Table 2). *Cluster II* are only related to one recognised phage species *Escherichia virus ID52* (63%, BLAST), genus *Gequatrovirus*, subfamily *Bullavirinae*, with whom they predominantly differ in the region in and around the major spike protein (*gpG*), a distinctive marker of the subfamily *Bullavirinae* involved in host attachment. The phylogenetic analysis separates *cluster II* from the gequatroviruses, and both the Gegenees scores (<22%) and NT similarities (<65% BLAST) are moderately low (Figure S2). Nonetheless, *cluster II* are still, based on pronounced genome organisation synteny and a conserved AA similarity (62-64%, Gegenees) considered to be gequatroviruses (Figure S2).

The sequencing of the microviruses is peculiar, as library preparation with the Nextera® XT DNA kit applies transposons targeting dsDNA. However, during microvirus infection the host polymerase converts the viral ssDNA into an intermediate state of covalently closed dsDNA, which is then replicated in rolling circle by viral replication proteins transcribed by the host RNA polymerase [65]. This intermediate state may have enabled the library preparation. The presence of host DNA in the sequence results indicates an

insufficient initial DNase1 treatment, which can be attributed to chemical inhibition or inactivation of the enzyme by adhesion to the sides of wells. Hence, it is reasonable to assume that the extracted microvirus DNA was captured as free dsDNA inside host cells during ongoing infections.

3.5. Twenty-five novel Siphoviridae phage species

In spite of similar genome sizes, the large group of *Siphoviridae*, is the most diverse (28 unique, 24 novel) in this study, with GC contents ranging from 43.9-54.6% (Figure 3, Table 2). The majority, *clusters IV-V* and singleton jahat, are of the subfamily *Tunavirinae*, while the remaining are allocated into at least three divergent genera, *Dhillonvirus*, *Jerseyvirus* and *seuratovirus*.

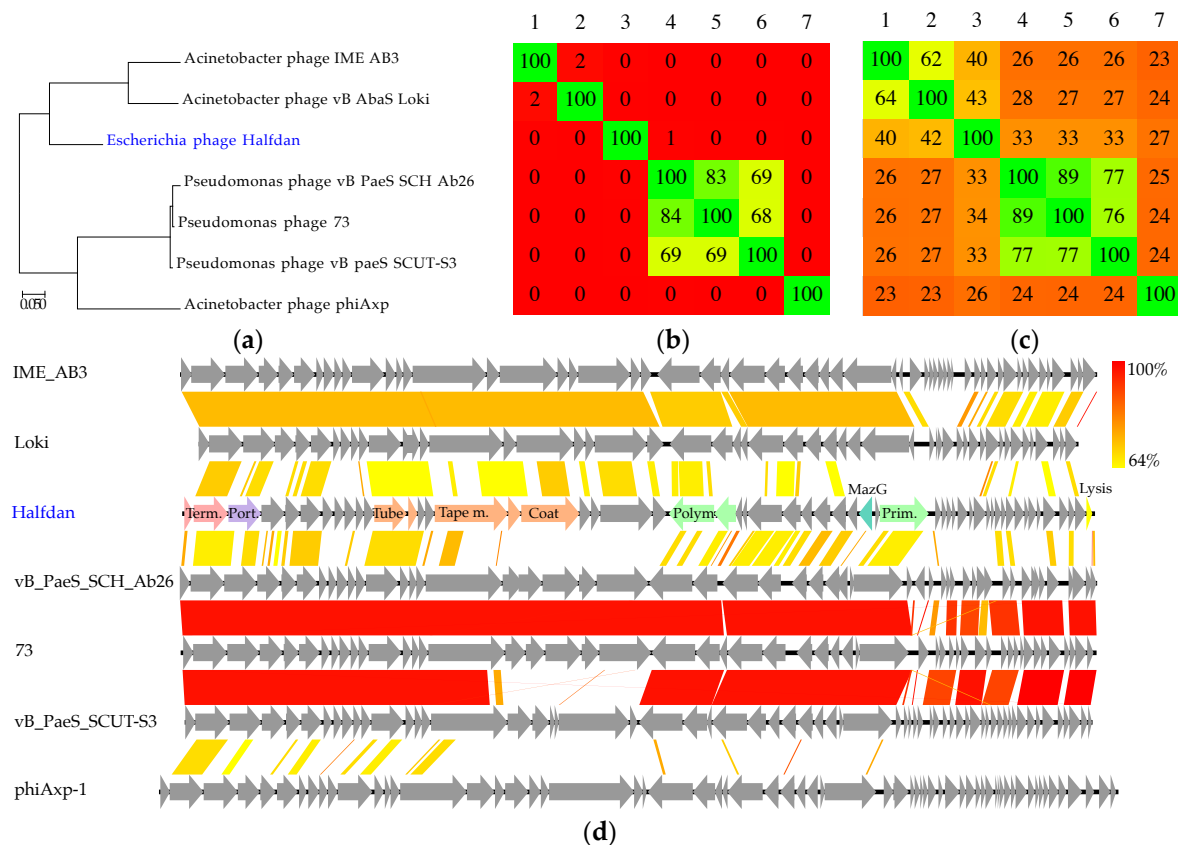
Cluster IV (46.7-51.6 kb, 43.9-44.1% GC, no tRNAs), *V* (49.8-51.6 kb, 44.6-44.8% GC, no tRNAs) and jahat (51.1 kb, 45.7% GC, no tRNAs) have low inter-Gegenees scores (7-20%) (Figure 1), yet they form a monophyletic clade and are all closely related (85-94%, BLAST) to published *Tunavirinae*. The *Cluster IV* phages are notable, as they represent a significant increase to the small genus *Hanrivervirus* comprising only the type species *Shigella virus pSf-1* (51.8 kb, 44% GC, no tRNAs) isolated from the Han River in Korea [66]. The common ancestry of *Cluster IV* and pSf-1 is evident by comparable genome sizes, GC contents, genomic organization and substantial NT (86-90%, BLAST) and AA (77-85%, Gegenees) similarities (Table 2, Figure S3). During their differentiation, many deletions and insertions of small hypothetical genes have occurred, most notable is a unique version of a putative tail-spike protein in seven of the *cluster IV* hanriverviruses, indicating divergent host ranges (Figure S3). All the hanriverviruses code for (putative) *dam* and Psf-1 is resistant towards at least six restriction endonucleases [66], suggesting they employ DNA methylation as a defence strategy.

The five novel dhillonviruses of *Cluster XII* (44.4-46.0 kb, 54.2-54.6% GC, no tRNAs) have substantial NT similarities with a group of unclassified dhillonviruses (80-89%, BLAST) (Table 2) and the type species *Escherichia virus HK578* (77-80%, BLAST). As with the hanriverviruses, and as observed by Korf *et al.*, (2019) [22], their genomes mainly differ in minor hypothetical genes and in putative tail-tip proteins, indicating divergent host ranges (Figure S4). Based on NT similarity and the presence of the canonical 7-deazaguanine

operon the singleton skure (59.4 kb, 44.6% GC, no tRNAs) is a seuratvirus, while the singleton Buks (40.3 kb, 49.7% GC, no tRNAs) is assigned to the genus *Jerseyvirus*, subfamily *Guernseyvirinae*.

3.5.1 Escherichia phage halfdan, a novel lineage within the *Siphoviridae*

Interestingly, the singleton halfdan (42.8 kb, 53.7% GC, no tRNAs) has only miniscule similarity with published phages (12-29%, BLAST). These entail two *Pseudomonas* phages vB_PaeS_SCUT-S3 (MK165657) and Ab26 (HG962376) [67] both *Septimatreviruses*, two *Acinetobacter* phages of the *Lokivirus* IMEAB3 (KF811200) and type species *Acinetobacter virus Loki* [68], and to a lesser degree the unclassified *Achromobacter* phage phiAxp-1 (KP313532) [69]. They have a common genomic organization, yet their intra-Gegenees score is $\leq 1\%$ and NT similarity is negligent in roughly a third of halfdan's 57 CDSs (Figure 4b, d). The phylogeny and AA similarities (Gegenees) also indicate a distant relation, although grouping halfdan closer with the lokiviruses (40-43%) than the septimatreviruses (33-34%) (Figure 4a, c). The genome of halfdan is mosaic, resembling the lokiviruses in the structural region and the septimatreviruses in the replication region (Figure 4d). Remarkably, halfdan has no NT similarity with known *Escherichia* phages, which could be an indication of *E. coli* not being the natural host, and that halfdan may have transcended species barriers. However, host range, morphology and other physical characteristics are beyond the scope of this study, and will be determined in future lab-based studies. Notably, both halfdan, Loki and Ab26 code for a (putative) MazG, although there is negligible sequence similarity (Figure 4d). Phage encoded MazG is hypothesized to be involved in restoring protein synthesis in a starved cell, in order to keep the cell alive, and ensure optimal replication conditions [70]. The phylogeny, whole genome alignments and low NT and AA similarities suggests that halfdan is distinct from other known phages (Figure 4). Accordingly, as per the ICTV guidelines, halfdan is the first phage sequenced of a novel *Siphoviridae* genus. Yet, halfdan is also, by its mosaicism, an indicator of the genetic continuum of phages questioning the validity of taxonomic interpretations.



311 **Figure 4** (a) Phylogenetic tree (Maximum log Likelihood: -7678.71, large terminase subunit), scalebar: substitutions per
 312 site. (b) Phylogenomic nucleotide distances (Gegenees, BLASTn; fragment size: 200, step size: 100, threshold: 0%). (c)
 313 Phylogenomic amino acid distances (Gegenees, tBLASTx; fragment size: 200, and step size: 100, threshold: 0%). (d)
 314 Pairwise alignment of phage genomes (Easyfig, BLASTn), the colour bars between genomes indicate percent pairwise
 315 similarity as illustrated by the colour bar in the upper right corner (Easyfig, BLASTn). Gene annotations are provided for
 316 phages identified in this study and deposited in GenBank.

317 3.6. Nine novel Podoviridae phage species

318 The nine novel *Podoviridae* (cluster XIV, 40.7-42.5 kb, 55.4-55.7% GC, no tRNAs), all with the hallmarks of
 319 the *Autographvirinae* i.e. unidirectionally encoded genes and RNA polymerases [64], have conserved genome
 320 organisation with the type species of the *Phikmvvirus*, *Pseudomonas virus phiKMV* (42.5 kb, 62.3% GC, no
 321 tRNAs) [71], but almost no NT similarity (<1%, BLAST). Besides the unclassified *Autographvirinae*
 322 *Enterobacteria phage J8-65* (NC_025445) (40.9 kb, 55.7% GC, no tRNAs), with which *Cluster XIV* has
 323 considerable NT similarity (69-93%, BLAST) and similar GC content, *Cluster XIV* only share >5% nucleotide
 324 similarity (40-42%, BLAST) with the *Phikmvvirus Pantoea virus Limezero* (43.0 kb, 55.4% GC, no tRNAs) [72]

(Figure 5, Table 2). The *Phikmvvirus* consists of four species, *phiKMV*, *Limezero*, *Pantoea virus Limelight* (44.5 kb, 54% GC, no tRNAs) [72] and *Pseudomonas virus LKA1* (41.6 kb, 60.9% GC, no tRNAs) [73].

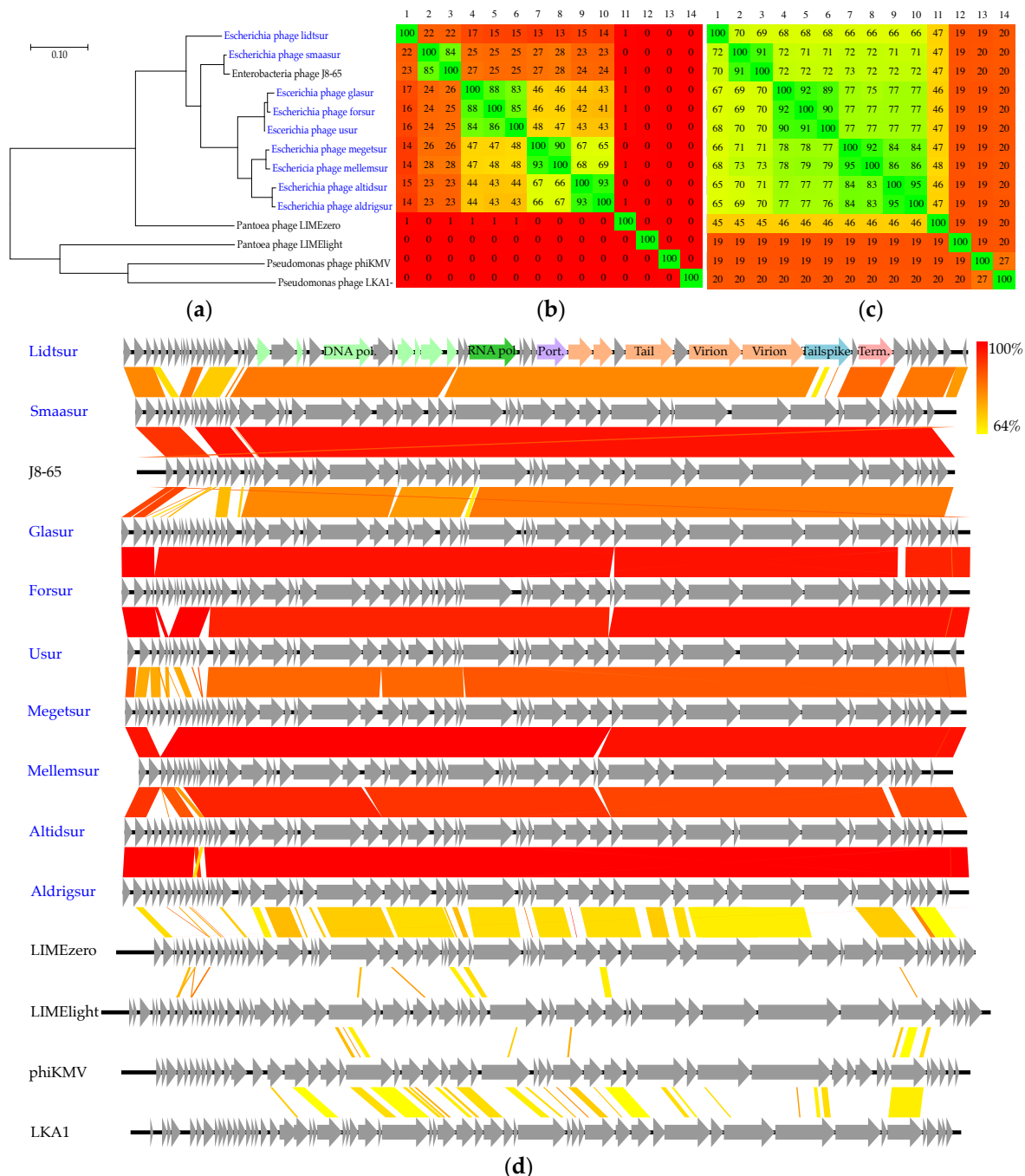


Figure 5 (a) Phylogenetic tree (Maximum log Likelihood: -11728.26, large terminase subunit), scalebar: substitutions per site. (b) Phylogenomic nucleotide distances (Gegenees, BLASTn: fragment size: 200, step size: 100, threshold: 0%). (c) Phylogenomic amino acid distances (Gegenees, tBLASTx: fragment size: 200, and step size: 100, threshold: 0%). (d) Pairwise alignment of phage genomes (Easyfig, BLASTn), the colour bars between genomes indicate percent pairwise similarity as illustrated by the colour bar in the upper right corner (Easyfig, BLASTn). Gene annotations are provided for phages identified in this study and deposited in GenBank.

Cluster XIV and J8-65 form a diverse monophyletic clade, with a substantial amount of deletions and insertions between them, subdividing into three sub-clusters with intra-Gegenees scores $\leq 28\%$ (Figure 5a, d). Phage lidtsur is singled-out and also codes for a unique version of tailspike colanidase, smaasur resembles J8-65 phage and the rest group together (Figure 5). Still, the three sub-clusters have for a large part conserved AA sequences (66-72%, Gegenees) (Figure 5b, c, d). Interestingly, this also applies to *Limezero*, with whom they have a Gegenees NT score $\leq 1\%$, but an AA similarity of 45-48%, supporting the phylogenetic grouping of *Limezero* and *cluster XIV* (Figure 5a, b, c). Based on phylogeny, limited NT and low AA similarities it is evident that there is a very distant relation between *cluster XIV* (and J8-65) and the phikmvviruses (Figure 5). Hence, *cluster XIV* (and J8-65) are not immediately, according to the ICTV guidelines, considered phikmvviruses. However, the *Phikmvvirus* already includes phages with minuscule DNA homology infecting divergent hosts [72,73]. The genus delimitation is based on overall genome architecture with the location of a single-subunit RNA polymerase gene adjacent to the structural genes, and not in the early region as in T7-like phages [16]. Another characteristic feature of the *phikmvvirus* is the presence of direct terminal repeats, though not yet verified in all members [72]. Read abundance in non-coding regions in genomic ends of *cluster XIV* genomes suggests they also have direct terminal repeats of a few hundred bp. In conclusion, we consider the *cluster XIV* phages (and J8-65) to be of the same lineage as the phikmvviruses, but contemplate that this genus may at some point be divided into at least two independent genera, as we learn more of the nature of the genes which distinguish *cluster XIV* from the other phikmvviruses on both NT and AA level and account for more 50% of their genomes.

3.7. Unclassified bacterial viruses represent two novel lineages

The four novel phages sortsyn of *cluster XIII* and the three singletons sortsne, sortkaff (41.9-42.5 kb, 59-60% GC, no tRNAs) and skarpretter (42.0 kb, 55.8% GC, no tRNAs) all have small genomes and high GC contents (Figure 3) and are suggested to represent two distinct novel lineages. They only share considerable ($\geq 6\%$, BLAST) NT similarity with four phages, Enterobacteria phage IME_EC2 (KF591601)(41.5 kb, 59.2% GC,

no tRNAs) isolated from hospital sewage [74], Salmonella phage lumpael (MK125141)(41.4 kb, 59.5% GC, no tRNAs) isolated from wastewater of the same sample-set in a previous study, Klebsiella phage vB_KpnS_IME279 (MF614100) (42.5 kb, 59.3% GC, no tRNAs) and Escherichia phage C130_2 (MH363708)(41.7 kb, 55.4% GC, no tRNAs) isolated from cheese [75].

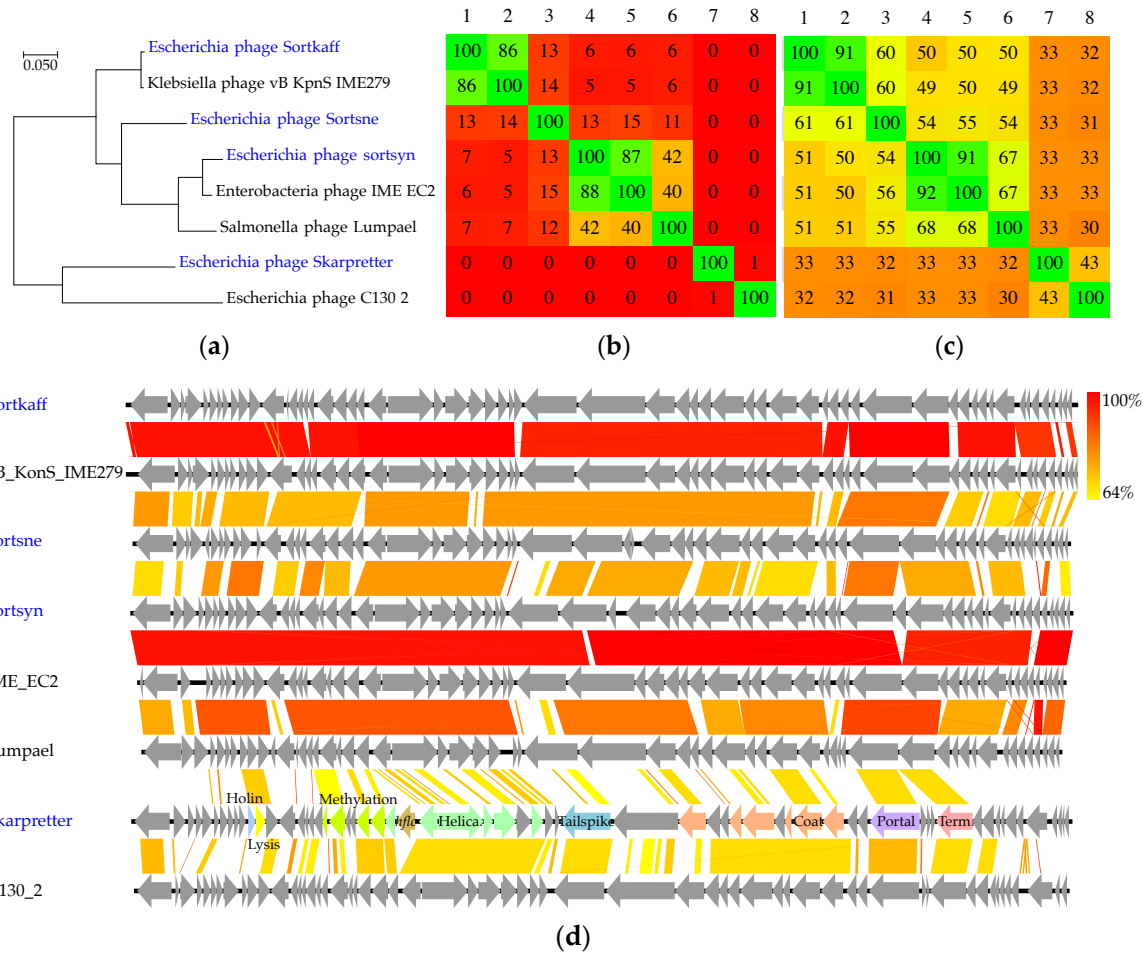


Figure 6 (a) Phylogenetic tree (Maximum log Likelihood: -8023.43, large terminase subunit), scalebar: substitutions per site. (b) Phylogenomic nucleotide distances (Gegenees, BLASTn: fragment size: 200, step size: 100, threshold: 0%). (c) Phylogenomic amino acid distances (Gegenees, tBLASTx: fragment size: 200, and step size: 100, threshold: 0%). (d) Pairwise alignment of phage genomes (Easyfig, BLASTn), the colour bars between genomes indicate percent pairwise similarity as illustrated by the colour bar in the upper right corner (Easyfig, BLASTn). Gene annotations are provided for phages identified in this study and deposited in GenBank.

Curiously, IME_EC2 is a confirmed (TEM) *Podoviridae*, with a short non-contractile tail [74], while C130_2 is a confirmed (TEM) *Myoviridae* with a long contractile tail [75]. Lumpael and vB_KpnS_IME279 are uncharacterised, yet both are assigned as *Podoviridae* in GenBank [17]. Although sortsne, sortkaff, sortsyn,

vB_KpnS_IME279, IME_EC2 and lumpael form a monophyletic clade, the Gegenees score (5-15%) between sub-clusters is surprisingly low (Figure 6a, b). Still, these six phages have comparable genome sizes (41.5-42.5 kb) and organisation, including similar relatively small-sized structural proteins, *dam* and *dcm* genes, equally high GC contents (59.5±0.5%) and relatively high AA similarities (49-91%) (Figure 6a, c, Table 2). Hence, they are clearly of the same lineage and likely to resemble IME_EC2 in having *Podoviridae* morphology. This group is also clearly distinct from all other known phages (<5%, BLAST) and as such constitute a novel genus, with a delimitation to be determined by future physical characterisation. Skarpretter and C130_2 form a monophyletic clade and both have slightly, though not significantly ($P = 0.38$), lower GC contents (55.6±0.2%), indicating a common ancestry and host. The conserved genome organisation and clustering of skarpretter with the other *Podoviridae* (Figure 1, 6d), suggests skarpretter also has *Podoviridae* morphology. Still, its closest relative C130_2 is described as a *Myoviridae* [75], leaving the true morphology of skarpretter a conundrum only resolved by TEM imaging. According to ICTV guidelines both skarpretter and C130_2 are representatives of novel lineages, as they are both clearly distinct from all other known phages. Skarpretter and C130_2 have very low NT (1% Gegenees, 38% BLAST) and AA similarities (43%) (Figure 6b, c). Indeed, skarpretter has ≤20% NT similarity (BLAST) with all other published phages. In addition, skarpretter codes for a putative *hflc* gene possibly inhibiting proteolysis of essential phage proteins, located in the replication region (Figure 6d). This gene, has to our knowledge never before been observed in phages, but occurs in almost all proteobacteria, including *E. coli* MG1655 [76].

4. Conclusion

By screening Danish wastewater, we identified no less than 104 unique *Escherichia* MG1655 phages, but predict the species richness to be at least in the range of 160-420, and even higher if including phages infecting other *Escherichia* species, though it is expected to fluctuate drastically over time and both within and between treatment facilities. Among the unique phages 91 represent novel phage species of at least four different families *Myoviridae*, *Siphoviridae*, *Podoviridae* and *Microviridae*. The diversity of these phages is striking, they

vary greatly in genome size and have a very broad GC-content range - possibly prompted by former or current alternate hosts. These findings add to our growing understanding of phage ecology and diversity, and through classification of these many phages we come yet another step closer to a more refined taxonomic understanding of phages. Furthermore, the numerous and diverse phages isolated in this study, all lytic to the same single strain, serve as an excellent opportunity to learn important phage-host interactions in future studies. These include, but are not limited to lysogen induced phage immunity, host-range and anti-RE systems. Finally, apart from substantial contributions to known genera, we came upon several unclassified lineages, some completely novel, which in our opinion constitute novel phage genera. We consider sortreg, sortkaff, sortsyn and sortsne together with lumpael, vB_kpnS_IME278 and IME_EC2 to constitute a novel genus within the *Podoviridae*. The *Myoviridae* flopper joins an interesting group of not yet classified phages with *Myoviridae* morphology and *Autographvirinae* characteristics, just as *cluster VII* together with five unclassified phages represent a novel unclassified genus within the *Myoviridae*. Lastly, we consider the *Siphoviridae* halfdan and the unclassified bacterial virus skarpretter to be the first sequenced representatives of each their novel genus. In conclusion, this study shows that uncharted territory still remains for even well-studied phage-host couples.

Supplementary Materials: Figure S1: Microviridae, Figure S2: Hanriversviruses, Figure S3: Dhillonviruses, Table 1 Auxiliary metabolism genes.

Funding: This research was funded by Villum Experiment Grant 17595, Aarhus University Research Foundation AUFF Grant E-2015-FLS-7-28 to Witold Kot and Human Frontier Science Program RGP0024/2018

Acknowledgments: This study had not been possible without the much-appreciated contribution by the many members of the Danish Water and Wastewater Association (DANVA), who kindly supplied us with time-series of wastewater samples from their treatment facilities. A special thanks to the Billund and Grindsted treatment facilities of Billund Vand & Energi, Lynetten, Avedøre and Damhusåen treatment facilities of BIOFOS, Kolding treatment facility of BlueKolding, Esbjerg Øst, Esbjerg Vest, Ribe, Varde og Skovlund treatment facilities of DIN Forsyning, Drøbro, Hadsten, Hammel, Hinnerup and Voldum treatment facilities of Favrskov Forsyning, Haerning treatment facility of Herning Vand, Hillerød

420 treatment facility of Hillerød Forsyning, Lemvig treatment facility of Lemvig Vand og Spildevand, Bevtøft, Gram,
 421 Haderslev, Halk, Jegerup, Nustrup, Over Jerstal, Skovby, Skrydstrup, Sommersted, Vojens and Årøsund treatment
 422 facilities of Provas, Marselisborg, Egå and Viby treatment facilities of Aarhus vand, Hedensted, Juelsminde and Tørring
 423 treatment facilities of Hedensted Spildevand, Ejby Mølle, Bogense, Hofmanskave, Hårslev, Nordvest, Nordøst, Otterup
 424 and Søndersø treatment facilities of VandCenterSyd, Øst and Vest treatment facilities of Aalborg Forsyning and finally
 425 Helsingør treatment facility of Forsyning Helsingør.

426 **Competing interests:** The authors declare no competing interests.

427 References

- 428 1. Weinbauer, M.G.; Rassoulzadegan, F. Are viruses driving microbial diversification and diversity? *Environ.*
 429 *Microbiol.* **2003**, *6*, 1–11.
- 430 2. Weitz, J.S.; Wilhelm, S.W. Ocean viruses and their effects on microbial communities and biogeochemical cycles.
 431 *F1000 Biol. Rep.* **2012**, *4*, 17.
- 432 3. Crummett, L.T.; Puxty, R.J.; Weihe, C.; Marston, M.F.; Martiny, J.B.H. The genomic content and context of
 433 auxiliary metabolic genes in marine cyanomyoviruses. *Virology* **2016**, *499*, 219–229.
- 434 4. Boyd, E.F. Bacteriophage-Encoded Bacterial Virulence Factors and Phage–Pathogenicity Island Interactions. *Adv.*
 435 *Virus Res.* **2012**, *82*, 91–118.
- 436 5. Fortier, L.-C.; Sekulovic, O. Importance of prophages to evolution and virulence of bacterial pathogens.
 437 <https://doi.org/10.4161/viru.24498> **2013**.
- 438 6. Volkova, V. V; Lu, Z.; Besser, T.; Gröhn, Y.T. Modeling the infection dynamics of bacteriophages in enteric
 439 *Escherichia coli*: estimating the contribution of transduction to antimicrobial gene spread. *Appl. Environ.*
 440 *Microbiol.* **2014**, *80*, 4350–62.
- 441 7. Bearson, B.L.; Allen, H.K.; Brunelle, B.W.; Lee, I.S.; Casjens, S.R.; Stanton, T.B. The agricultural antibiotic

- 442 carbadox induces phage-mediated gene transfer in Salmonella. *Front. Microbiol.* **2014**, 5, 52.
- 443 8. Grose, J.H.; Casjens, S.R. Understanding the enormous diversity of bacteriophages: the tailed phages that infect
444 the bacterial family Enterobacteriaceae. *Virology* **2014**, 468–470, 421–443.
- 445 9. Dykhuizen, D. Species Numbers in Bacteria. *Proc. Calif. Acad. Sci.* **2005**, 56, 62.
- 446 10. Hatfull, G.F.; Pedulla, M.L.; Jacobs-Sera, D.; Cichon, P.M.; Foley, A.; Ford, M.E.; Gonda, R.M.; Houtz, J.M.;
447 Hryckowian, A.J.; Kelchner, V.A.; et al. Exploring the Mycobacteriophage Metaproteome: Phage Genomics as an
448 Educational Platform. *PLoS Genet.* **2006**, 2, e92.
- 449 11. Hatfull, G.F. Mycobacteriophages: Windows into Tuberculosis. *PLoS Pathog.* **2014**, 10, e1003953.
- 450 12. Hatfull, G.F.; Jacobs-Sera, D.; Lawrence, J.G.; Pope, W.H.; Russell, D.A.; Ko, C.C.; Weber, R.J.; Patel, M.C.;
451 Germane, K.L.; Edgar, R.H.; et al. Comparative Genomic Analysis of 60 Mycobacteriophage Genomes: Genome
452 Clustering, Gene Acquisition, and Gene Size. *J. Mol. Biol.* **2010**, 397, 119–143.
- 453 13. Dedrick, R.M.; Jacobs-Sera, D.; Bustamante, C.A.G.; Garlena, R.A.; Mavrich, T.N.; Pope, W.H.; Reyes, J.C.C.;
454 Russell, D.A.; Adair, T.; Alvey, R.; et al. Prophage-mediated defence against viral attack and viral counter-
455 defence. *Nat. Microbiol.* **2017**, 2, 16251.
- 456 14. Jacobs-Sera, D.; Marinelli, L.J.; Bowman, C.; Broussard, G.W.; Guerrero Bustamante, C.; Boyle, M.M.; Petrova,
457 Z.O.; Dedrick, R.M.; Pope, W.H.; Science Education Alliance Phage Hunters Advancing Genomics And
458 Evolutionary Science Sea-Phages Program, S.E.A.P.H.A.G. and E.S. (SEA-P.; et al. On the nature of
459 mycobacteriophage diversity and host preference. *Virology* **2012**, 434, 187–201.
- 460 15. Hatfull, G.F. The Secret Lives of Mycobacteriophages. *Adv. Virus Res.* **2012**, 82, 179–288.
- 461 16. Lefkowitz, E.J.; Dempsey, D.M.; Hendrickson, R.C.; Orton, R.J.; Siddell, S.G.; Smith, D.B. Virus taxonomy: the

- 462 database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic Acids Res.* **2018**, 46, D708–D717.
- 463 17. Benson, D.A.; Cavanaugh, M.; Clark, K.; Karsch-Mizrachi, I.; Lipman, D.J.; Ostell, J.; Sayers, E.W. GenBank.
464 *Nucleic Acids Res.* **2017**, 45, D37–D42.
- 465 18. Lawrence, J.G.; Hatfull, G.F.; Hendrix, R.W. Imbrolios of viral taxonomy: genetic exchange and failings of
466 phenetic approaches. *J. Bacteriol.* **2002**, 184, 4891–905.
- 467 19. Aiewsakun, P.; Adriaenssens, E.M.; Lavigne, R.; Kropinski, A.M.; Simmonds, P. Evaluation of the genomic
468 diversity of viruses infecting bacteria, archaea and eukaryotes using a common bioinformatic platform: steps
469 towards a unified taxonomy. *J. Gen. Virol.* **2018**, 99, 1331–1343.
- 470 20. Nelson, D. Phage taxonomy: we agree to disagree. *J. Bacteriol.* **2004**, 186, 7029–31.
- 471 21. Adriaenssens, E.; Brister, J.R. How to Name and Classify Your Phage: An Informal Guide. *Viruses* **2017**, 9.
- 472 22. Korf; Meier-Kolthoff; Adriaenssens; Kropinski; Nimtz; Rohde; van Raaij; Wittmann Still Something to Discover:
473 Novel Insights into Escherichia coli Phage Diversity and Taxonomy. *Viruses* **2019**, 11, 454.
- 474 23. Jurczak-Kurek, A.; Gąsior, T.; Nejman-Faleńczyk, B.; Bloch, S.; Dydecka, A.; Topka, G.; Necel, A.; Jakubowska-
475 Deredas, M.; Narajczyk, M.; Richert, M.; et al. Biodiversity of bacteriophages: morphological and biological
476 properties of a large group of phages isolated from urban sewage. *Sci. Rep.* **2016**, 6, 34338.
- 477 24. Sambrook, J.F.; Russell, D.W.; Editors *Molecular cloning: A laboratory manual, third edition.*; 2nd ed.; Cold Spring
478 Harbor Laboratory, 2000; ISBN 0 87969 309 6.
- 479 25. Kot, W.; Vogensen, F.K.; Sørensen, S.J.; Hansen, L.H. DPS – A rapid method for genome sequencing of DNA-
480 containing bacteriophages directly from a single plaque. *J. Virol. Methods* **2014**, 196, 152–156.
- 481 26. Bankevich, A.; Nurk, S.; Antipov, D.; Gurevich, A.A.; Dvorkin, M.; Kulikov, A.S.; Lesin, V.M.; Nikolenko, S.I.;

- 482 Pham, S.; Pribelski, A.D.; et al. SPAdes: a new genome assembly algorithm and its applications to single-cell
483 sequencing. *J. Comput. Biol.* **2012**, *19*, 455–77.
- 484 27. Brettin, T.; Davis, J.J.; Disz, T.; Edwards, R.A.; Gerdes, S.; Olsen, G.J.; Olson, R.; Overbeek, R.; Parrello, B.; Pusch,
485 G.D.; et al. RASTtk: A modular and extensible implementation of the RAST algorithm for building custom
486 annotation pipelines and annotating batches of genomes. *Sci. Rep.* **2015**, *5*, 8365.
- 487 28. Besemer, J.; Borodovsky, M. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses.
488 *Nucleic Acids Res.* **2005**, *33*, W451–W454.
- 489 29. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**,
490 *215*, 403–410.
- 491 30. Soding, J.; Biegert, A.; Lupas, A.N. The HHpred interactive server for protein homology detection and structure
492 prediction. *Nucleic Acids Res.* **2005**, *33*, W244–W248.
- 493 31. Finn, R.D.; Coghill, P.; Eberhardt, R.Y.; Eddy, S.R.; Mistry, J.; Mitchell, A.L.; Potter, S.C.; Punta, M.; Qureshi, M.;
494 Sangrador-Vegas, A.; et al. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids*
495 *Res.* **2016**, *44*, D279–D285.
- 496 32. Enrique González-Tortuero¹, Thomas David Sean Sutton¹, Vimalkumar Velayudhan¹, 4 Andrey Nikolaevich
497 Shkoporov¹, Lorraine Anne Draper¹, Stephen Robert Stockdale¹, 2, Reynolds 5 Paul Ross¹, 3, Colin Hill¹, 3, 6
498 VIGA: a sensitive, precise and automatic de novo Viral Genome Annotator. *bioRxiv* **2018**, 277509.
- 499 33. Zankari, E.; Hasman, H.; Cosentino, S.; Vestergaard, M.; Rasmussen, S.; Lund, O.; Aarestrup, F.M.; Larsen, M.V.
500 Identification of acquired antimicrobial resistance genes. *J. Antimicrob. Chemother.* **2012**, *67*, 2640–4.
- 501 34. Kleinheinz, K.A.; Joensen, K.G.; Larsen, M.V. Applying the ResFinder and VirulenceFinder web-services for easy

502 identification of acquired antibiotic resistance and E. coli virulence genes in bacteriophage and prophage
503 nucleotide sequences. *Bacteriophage* **2014**, 4, e27943.

504 35. Joensen, K.G.; Scheutz, F.; Lund, O.; Hasman, H.; Kaas, R.S.; Nielsen, E.M.; Aarestrup, F.M. Real-Time Whole-
505 Genome Sequencing for Routine Typing, Surveillance, and Outbreak Detection of Verotoxigenic Escherichia coli.
506 *J. Clin. Microbiol.* **2014**, 52, 1501–1510.

507 36. Roberts, R.J.; Vincze, T.; Posfai, J.; Macelis, D. REBASE—a database for DNA restriction and modification:
508 enzymes, genes and genomes. *Nucleic Acids Res.* **2015**, 43, D298–D299.

509 37. Kieft, K.; Zhou, Z.; Anantharaman, K. VIBRANT: Automated recovery, annotation and curation of microbial
510 viruses, and evaluation of virome function from genomic sequences. *bioRxiv* **2019**, 855387.

511 38. Adriaenssens, E.; Brister, J.R. How to Name and Classify Your Phage: An Informal Guide. *Viruses* **2017**, 9.

512 39. Ågren, J.; Sundström, A.; Håfström, T.; Segerman, B. Gegenees: Fragmented Alignment of Multiple Genomes for
513 Determining Phylogenomic Distances and Genetic Signatures Unique for Specified Target Groups. *PLoS One*
514 **2012**, 7, e39107.

515 40. Kumar, S.; Stecher, G.; Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger
516 Datasets. *Mol. Biol. Evol.* **2016**, 33, 1870–1874.

517 41. Lopes, A.; Tavares, P.; Petit, M.-A.; Guérois, R.; Zinn-Justin, S. Automated classification of tailed bacteriophages
518 according to their neck organization. *BMC Genomics* **2014**, 15, 1027.

519 42. Mercanti, D.J.; Rousseau, G.M.; Capra, M.L.; Quiberoni, A.; Tremblay, D.M.; Labrie, S.J.; Moineau, S. Genomic
520 Diversity of Phages Infecting Probiotic Strains of Lactobacillus paracasei. *Appl. Environ. Microbiol.* **2016**, 82, 95–
521 105.

- 522 43. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*
523 **2004**, 32, 1792–7.
- 524 44. Tamura, K.; Nei, M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial
525 DNA in humans and chimpanzees. *Mol. Biol. Evol.* **1993**, 10, 512–26.
- 526 45. Sullivan, M.J.; Petty, N.K.; Beatson, S.A. Easyfig: a genome comparison visualizer. *Bioinformatics* **2011**, 27, 1009.
- 527 46. Hsieh, T.C.; Ma, K.H.; Chao, A. iNEXT: an R package for rarefaction and extrapolation of species diversity (Hill
528 numbers). *Methods Ecol. Evol.* **2016**, 7, 1451–1456.
- 529 47. Chao, A.; Gotelli, N.J.; Hsieh, T.C.; Sander, E.L.; Ma, K.H.; Colwell, R.K.; Ellison, A.M. Rarefaction and
530 extrapolation with Hill numbers: A framework for sampling and estimation in species diversity studies. *Ecol.*
531 *Monogr.* **2014**, 84, 45–67.
- 532 48. RStudio Team RStudio: Integrated Development for R. 2016.
- 533 49. Rocha, E.P.C.; Danchin, A. Base composition bias might result from competition for metabolic resources. *Trends*
534 *Genet.* **2002**, 18, 291–294.
- 535 50. Motlagh, A.M.; Bhattacharjee, A.S.; Coutinho, F.H.; Dutilh, B.E.; Casjens, S.R.; Goel, R.K. Insights of Phage-Host
536 Interaction in Hypersaline Ecosystem through Metagenomics Analyses. *Front. Microbiol.* **2017**, 8, 352.
- 537 51. Thomas, J.A.; Orwenyo, J.; Wang, L.-X.; Black, L.W. The Odd “RB” Phage-Identification of
538 Arabinosylation as a New Epigenetic Modification of DNA in T4-Like Phage RB69. *Viruses* **2018**, 10.
- 539 52. Purohit, S.; Bestwick, K.; Lasser, G.W.; Rogers, C.M.; Mathews, C.K. T4 Phage-coded Dihydrofolate Reductase.
540 **1981**, 256, 9121–9125.
- 541 53. Hutinet, G.; Kot, W.; Cui, L.; Hillebrand, R.; Balamkundu, S.; Gnanakalai, S.; Neelakandan, R.; Carstens, A.B.;

- 542 Lui, C.F.; Tremblay, D.; et al. 7-Deazaguanine modifications protect phage DNA from host restriction systems.
543 *Nature* **2019**, 1–12.
- 544 54. Carstens, A.B.; Kot, W.; Lametsch, R.; Neve, H.; Hansen, L.H. Characterisation of a novel enterobacteria phage,
545 CAjan, isolated from rat faeces. *Arch. Virol.* **2016**, *161*, 2219–2226.
- 546 55. Shkoporov, A.N.; Clooney, A.G.; Sutton, T.D.S.; Ryan, F.J.; Daly, K.M.; Nolan, J.A.; McDonnell, S.A.; Khokhlova,
547 E. V.; Draper, L.A.; Forde, A.; et al. The Human Gut Virome Is Highly Diverse, Stable, and Individual Specific.
548 *Cell Host Microbe* **2019**, *26*, 527-541.e5.
- 549 56. Tsonos, J.; Adriaenssens, E.M.; Klumpp, J.; Hernalsteens, J.-P.; Lavigne, R.; De Greve, H. Complete genome
550 sequence of the novel Escherichia coli phage phAPEC8. *J. Virol.* **2012**, *86*, 13117–8.
- 551 57. Schwarzer, D.; Buettner, F.F.R.; Browning, C.; Nazarov, S.; Rabsch, W.; Bethe, A.; Oberbeck, A.; Bowman, V.D.;
552 Stummeyer, K.; Mühlenhoff, M.; et al. A multivalent adsorption apparatus explains the broad host range of
553 phage phi92: a comprehensive genomic and structural analysis. *J. Virol.* **2012**, *86*, 10384–98.
- 554 58. Schwarzer, D.; Browning, C.; Stummeyer, K.; Oberbeck, A.; Mühlenhoff, M.; Gerardy-Schahn, R.; Leiman, P.G.
555 Structure and biochemical characterization of bacteriophage phi92 endosialidase. *Virology* **2015**, *477*, 133–143.
- 556 59. Liu, H.; Geagea, H.; Rousseau, G.M.; Labrie, S.J.; Tremblay, D.M.; Liu, X.; Moineau, S. Characterization of the
557 Escherichia coli Virulent Myophage ST32. *Viruses* **2018**, *10*.
- 558 60. Jamalludeen, N.; Kropinski, A.M.; Johnson, R.P.; Lingohr, E.; Harel, J.; Gyles, C.L. Complete genomic sequence
559 of bacteriophage phiEcoM-GJ1, a novel phage that has myovirus morphology and a podovirus-like RNA
560 polymerase. *Appl. Environ. Microbiol.* **2008**, *74*, 516–25.
- 561 61. Thompson, D.W.; Casjens, S.R.; Sharma, R.; Grose, J.H. Genomic comparison of 60 completely sequenced

- 562 bacteriophages that infect *Erwinia* and/or *Pantoea* bacteria. *Virology* **2019**, 535, 59–73.
- 563 62. Born, Y.; Fieseler, L.; Marazzi, J.; Lurz, R.; Duffy, B.; Loessner, M.J. Novel Virulent and Broad-Host-Range
564 *Erwinia amylovora* Bacteriophages Reveal a High Degree of Mosaicism and a Relationship to Enterobacteriaceae
565 Phages. *Appl. Environ. Microbiol.* **2011**, 77, 5945–5954.
- 566 63. Lim, J.-A.; Shin, H.; Lee, D.H.; Han, S.-W.; Lee, J.-H.; Ryu, S.; Heu, S. Complete genome sequence of the
567 *Pectobacterium carotovorum* subsp. *carotovorum* virulent bacteriophage PM1. *Arch. Virol.* **2014**, 159, 2185–2187.
- 568 64. Lavigne, R.; Seto, D.; Mahadevan, P.; Ackermann, H.-W.; Kropinski, A.M. Unifying classical and molecular
569 taxonomic classification: analysis of the Podoviridae using BLASTP-based tools. *Res. Microbiol.* **2008**, 159, 406–
570 414.
- 571 65. Doore, S.M.; Fane, B.A. The microviridae: Diversity, assembly, and experimental evolution. *Virology* **2016**, 491,
572 45–55.
- 573 66. Jun, J.W.; Kim, J.H.; Shin, S.P.; Han, J.E.; Chai, J.Y.; Park, S.C. Characterization and complete genome sequence of
574 the *Shigella* bacteriophage pSf-1. *Res. Microbiol.* **2013**, 164, 979–986.
- 575 67. Essoh, C.; Latino, L.; Midoux, C.; Blouin, Y.; Loukou, G.; Nguetta, S.-P.A.; Lathro, S.; Cablanmian, A.; Kouassi,
576 A.K.; Vergnaud, G.; et al. Investigation of a Large Collection of *Pseudomonas aeruginosa* Bacteriophages
577 Collected from a Single Environmental Source in Abidjan, Côte d’Ivoire. *PLoS One* **2015**, 10, e0130548.
- 578 68. Turner, D.; Wand, M.E.; Briers, Y.; Lavigne, R.; Sutton, J.M.; Reynolds, D.M. Characterisation and genome
579 sequence of the lytic *Acinetobacter baumannii* bacteriophage vB_AbaS_Loki. *PLoS One* **2017**, 12, e0172303.
- 580 69. Ma, Y.; Li, E.; Qi, Z.; Li, H.; Wei, X.; Lin, W.; Zhao, R.; Jiang, A.; Yang, H.; Yin, Z.; et al. Isolation and molecular
581 characterisation of *Achromobacter* phage phiAxp-3, an N4-like bacteriophage. *Sci. Rep.* **2016**, 6, 24776.

- 582 70. Bryan, M.J.; Burroughs, N.J.; Spence, E.M.; Clokie, M.R.J.; Mann, N.H.; Bryan, S.J. Evidence for the intense
583 exchange of MazG in marine cyanophages by horizontal gene transfer. *PLoS One* **2008**, *3*, 1–12.
- 584 71. Lavigne, R.; Burkal'tseva, M. V; Robben, J.; Sykilinda, N.N.; Kurochkina, L.P.; Grymonprez, B.; Jonckx, B.;
585 Krylov, V.N.; Mesyanzhinov, V. V; Volckaert, G. The genome of bacteriophage ϕ KMV, a T7-like virus infecting
586 *Pseudomonas aeruginosa*. *Virology* **2003**, *312*, 49–59.
- 587 72. Adriaenssens, E.M.; Ceyssens, P.-J.; Dunon, V.; Ackermann, H.-W.; Van Vaerenbergh, J.; Maes, M.; De Proft, M.;
588 Lavigne, R. Bacteriophages LIMelight and LIMEzero of *Pantoea agglomerans*, belonging to the "phiKMV-
589 like viruses". *Appl. Environ. Microbiol.* **2011**, *77*, 3443–50.
- 590 73. Ceyssens, P.-J.; Lavigne, R.; Mattheus, W.; Chibeu, A.; Hertveldt, K.; Mast, J.; Robben, J.; Volckaert, G. Genomic
591 analysis of *Pseudomonas aeruginosa* phages LKD16 and LKA1: establishment of the phiKMV subgroup within
592 the T7 supergroup. *J. Bacteriol.* **2006**, *188*, 6924–31.
- 593 74. Hua, Y.; An, X.; Pei, G.; Li, S.; Wang, W.; Xu, X.; Fan, H.; Huang, Y.; Zhang, Z.; Mi, Z.; et al. Characterization of
594 the morphology and genome of an *Escherichia coli* podovirus. *Arch. Virol.* **2014**, *159*, 3249–3256.
- 595 75. Sváb, D.; Falgenhauer, L.; Rohde, M.; Chakraborty, T.; Tóth, I. Complete genome sequence of C130_2, a novel
596 myovirus infecting pathogenic *Escherichia coli* and *Shigella* strains. *Arch. Virol.* **2019**, *164*, 321–324.
- 597 76. Bandyopadhyay, K.; Parua, P.K.; Datta, A.B.; Parrack, P. *Escherichia coli* HflK and HflC can individually inhibit
598 the HflB (FtsH)-mediated proteolysis of λ CII in vitro. *Arch. Biochem. Biophys.* **2010**, *501*, 239–243.

599

600

SUPPLEMENTARY FIGURES AND TABLES

Figure S1 *Microviridae*

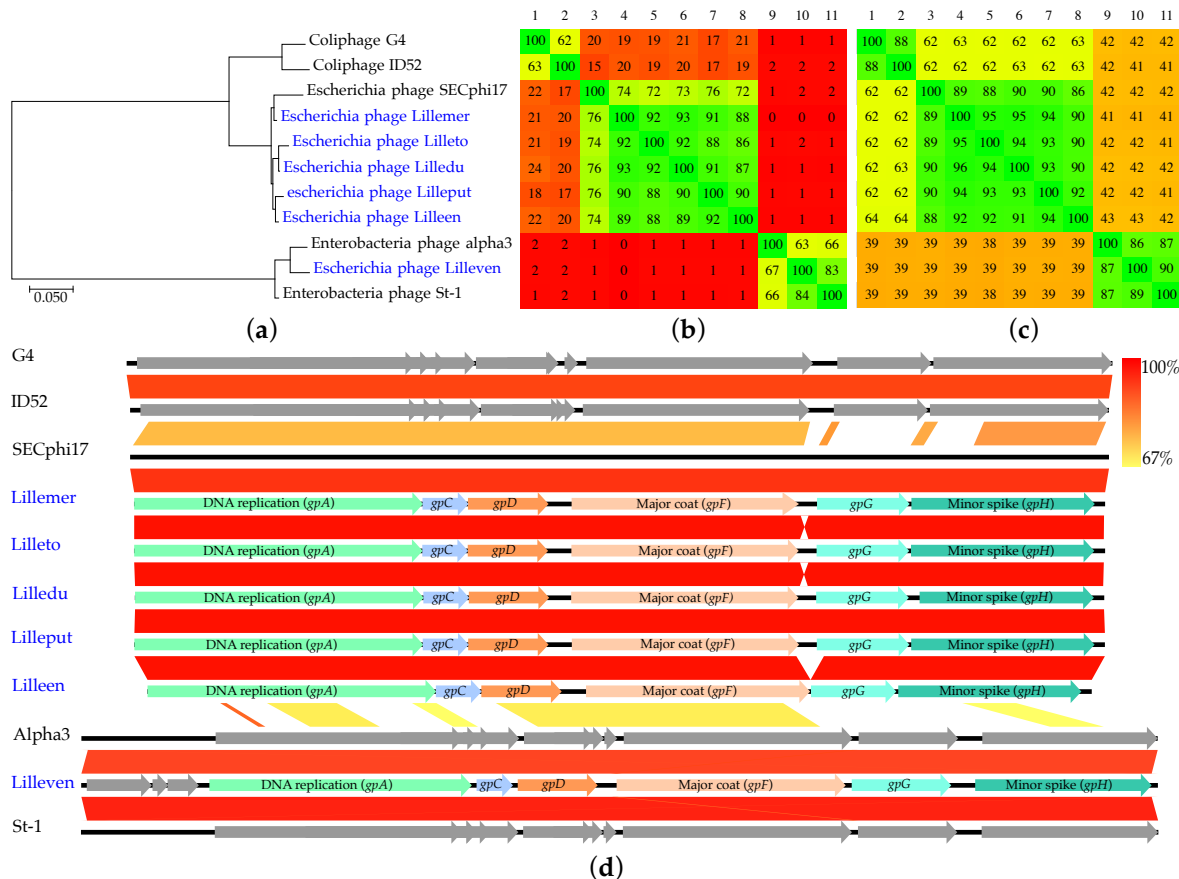


Figure S4 (a) Phylogenetic tree (Maximum log Likelihood: -6065.3, DNA replication protein *gpA*), scalebar: substitutions per site (b) Phylogenomic nucleotide distances (Gegenees, BLASTn: fragment size: 200, step size: 100, threshold: 0%). (c) Phylogenomic amino acid distances (Gegenees, tBLASTx: fragment size: 200, and step size: 100, threshold: 0%). (d) Pairwise alignment of phage genomes, the colour bars between genomes indicate percent pairwise similarity as illustrated by the colour bar in the upper right corner (Easyfig, BLASTn). Gene annotations are provided for phages identified in this study and deposited in GenBank.

Figure S2 *Hanrivervirus*

	1	2	3	4	5	6	7	8	9
1 <i>Damhaus</i>	100	84	85	85	88	79	81	85	83
2 <i>Herni</i>	84	100	86	86	85	82	80	90	85
3 <i>Grams</i>	87	88	100	86	88	81	82	87	84
4 <i>Vojen</i>	85	86	84	100	84	81	82	89	85
5 <i>Aaroes</i>	88	84	85	83	100	79	80	84	81
6 <i>Aalborg</i>	85	88	85	87	86	100	83	86	84
7 <i>Haarsle</i>	85	84	83	84	84	80	100	86	84
8 <i>Egaa</i>	85	89	84	87	84	79	82	100	85
9 <i>Shigella</i>	82	84	81	84	81	77	80	85	100

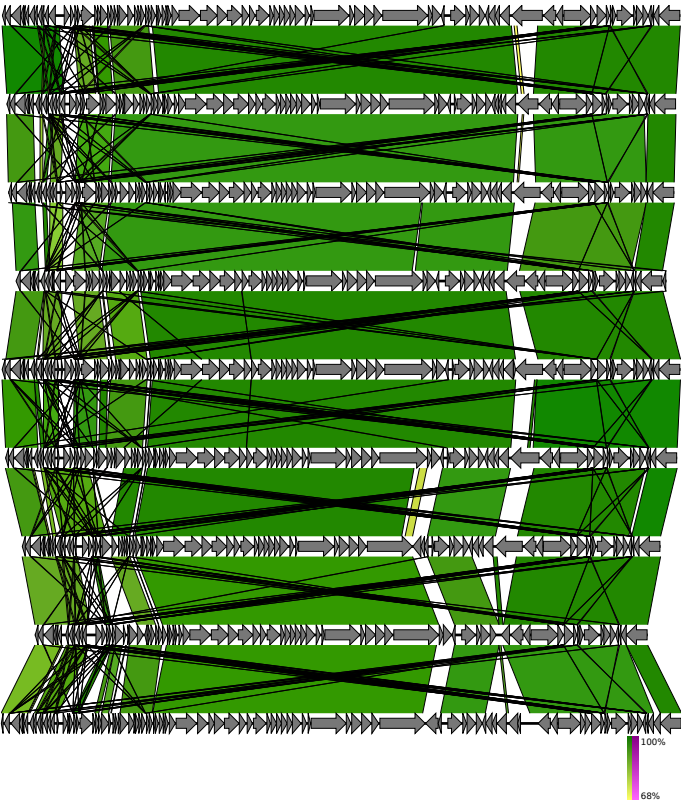


Figure S2 Phylogenomic amino acid distances (Gegenees, tBLASTx: fragment size: 200, and step size: 100, threshold: 0%). and Pairwise alignment of phage genomes (Easyfig, BLASTn), the colour bars between genomes indicate percent pairwise similarity as illustrated by the colour bar in the lower right corner (Easyfig, BLASTn). Genome order is the same in both analyses.

618 Figure S3 *Dhillonvirus*

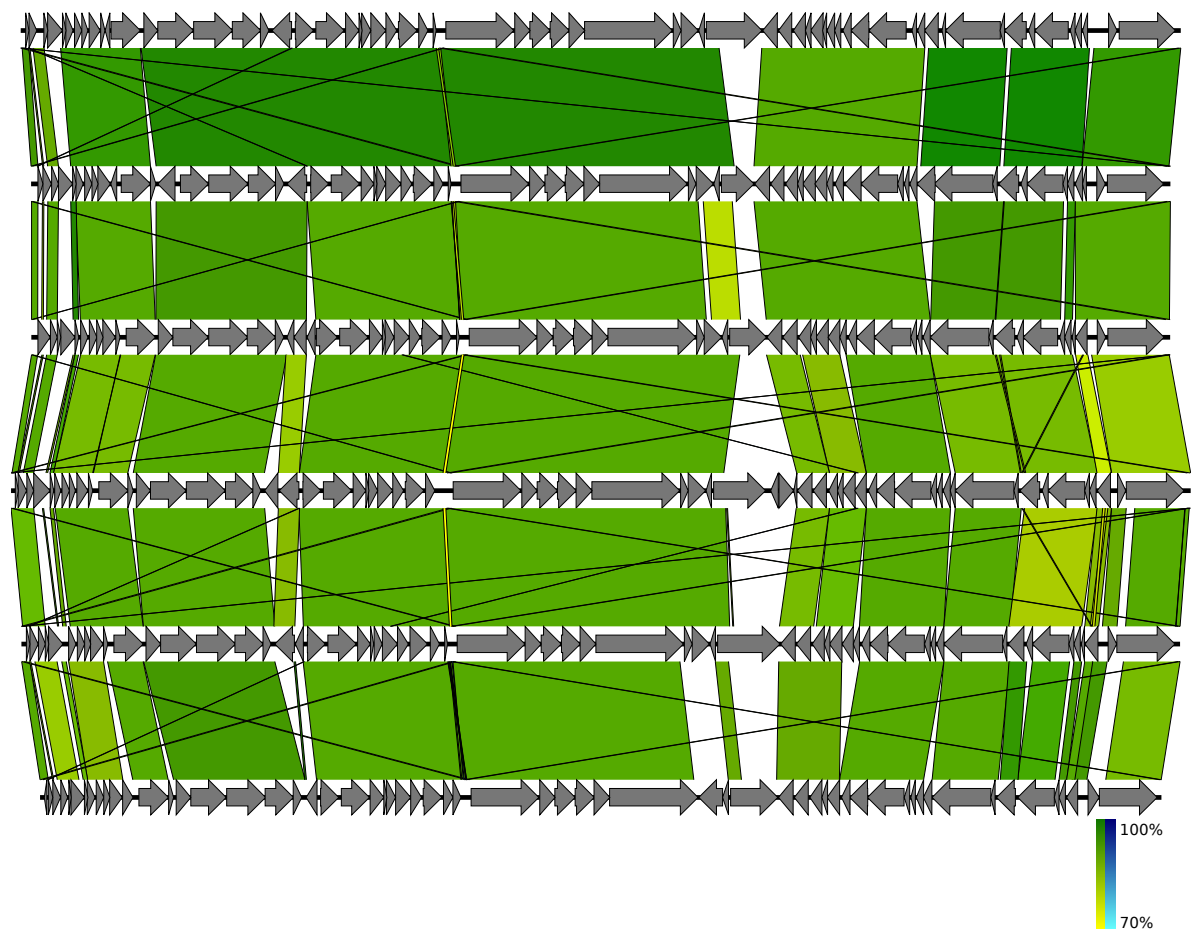


Figure S3 Pairwise alignment of phage genomes (Easyfig, BLASTn), the colour bars between genomes indicate percent pairwise similarity as illustrated by the colour bar in the lower right corner (Easyfig, BLASTn).

623 Table S1 Auxiliary metabolism genes

624

Gene	Protein	KO number	Function in bacteriophages	Count	phages with this gene
<i>rmlC</i>	dTDP-4-dehydorrhamnose 3,5-epimerase	K01790	dtdp rhamnose biosynthesis	10	<i>Cluster VI and VII</i>
<i>rmlA</i>	glucose-1-phosphate thymidyltransferase	K00973	dtdp rhamnose biosynthesis	8	<i>Cluster VII, ukendt and tuntematon</i>
<i>rmlB</i>	dTDP-glucose 4,6-dehydratase	K01710	dtdp rhamnose biosynthesis	8	<i>Cluster VII, ukendt and tuntematon</i>
<i>rmlD</i>	dTDP-4-dehydorrhamnose reductase	K00067	dtdp rhamnose biosynthesis	10	<i>Cluster VI and VII</i>
<i>dcm</i>	DNA (cytosine-5)-methyltransferase	K17398	DNA modification	5	<i>Cluster VII</i>
NAMPT	nicotinamide phosphoribosyltransferase	K03462	unknown	25	<i>Felixounavirus, Suspvirus and Cluster VII</i>
<i>mec</i>	CysO sulfur-carrier protein]-S-L-cysteine hydrolase	K21140	tail fiber protein	20	<i>Hanriverovirus and unclassified Tunavirinae</i>
<i>folA</i>	dihydrofolate reductase	K00287	de novo synthesis of pyrimidines, thymidylic acid, and certain amino acids	33	<i>Felixounavirus, Suspvirus, Krischovirus, Dhakavirus, Mosigovirus and Tequatrovirus</i>
<i>aroAG</i>	3-deoxy-7-phosphoheptulonate synthase	K01626	unknown	3	<i>Three of the four Mosigovirus</i>
<i>AUP1</i>	UDP-N-acetylglucosamine diphosphorylase	K23144	DNA modification	4	<i>Mosigovirus</i>
<i>KdsD</i>	arabinose-5-phosphate isomerase	K06041	DNA modification	4	<i>Mosigovirus</i>
<i>dcm</i>	DNA (cytosine-5)-methyltransferase	K00558	DNA modification	1	<i>Pangalan</i>
<i>queC</i>	7-cyano-7-deazaguanine synthase	K06920	DNA modification	1	<i>Skure</i>
<i>queE</i>	7-carboxy-7-deazaguanine synthase	K10026	DNA modification	1	<i>Skure</i>
<i>folE</i>	GTP cyclohydrolase	K01495	DNA modification	1	<i>Skure</i>
<i>queD</i>	6-carboxy-5,6,7,8-tetrahydropterin synthase	K01737	DNA modification	1	<i>Skure</i>

625