Virology 535 (2019) 59–73

Contents lists available at ScienceDirect

Virology

journal homepage: www.elsevier.com/locate/virology

Genomic comparison of 60 completely sequenced bacteriophages that infect *Erwinia* and/or *Pantoea* bacteria

Daniel W. Thompson[a], Sherwood R. Casjens[b,c], Ruchira Sharma[a], Julianne H. Grose[a,*]

[a] *Department of Microbiology and Molecular Biology, Brigham Young University, Utah, USA*
[b] *Division of Microbiology and Immunology, Department of Pathology, University of Utah School of Medicine, University of Utah, Salt Lake City, UT, 84112, USA*
[c] *School of Biological Sciences, University of Utah, Salt Lake City, UT, 84112, USA*

ARTICLE INFO

Keywords:
Bacteriophages
Erwinia
Pantoea
Bacteriophage
Phage
Cluster
Diversity
Erwiniaceae
Enterobacteriales

ABSTRACT

Erwinia and *Pantoea* are closely related bacterial plant pathogens in the Gram negative *Enterobacteriales* order. Sixty tailed bacteriophages capable of infecting these pathogens have been completely sequenced by investigators around the world and are in the current databases, 30 of which were sequenced by our lab. These 60 were compared to 991 other *Enterobacteriales* bacteriophage genomes and found to be, on average, just over twice the overall average length. These *Erwinia* and *Pantoea* phages comprise 20 clusters based on nucleotide and protein sequences. Five clusters contain only phages that infect the *Erwinia* and *Pantoea* genera, the other 15 clusters are closely related to bacteriophages that infect other *Enterobacteriales*; however, within these clusters the *Erwinia* and *Pantoea* phages tend to be distinct, suggesting ecological niche may play a diversification role. The failure of many of their encoded proteins to have predicted functions highlights the need for further study of these phages.

1. Introduction

Bacteriophages are viruses that infect bacteria. Their virions are comprised of a protein shell containing genetic material that can be dsDNA, ssDNA, dsRNA or ssRNA. Their genomes can contain as few as 3.3 kb or as many as 500 kb (Fiers et al., 1976; Hatfull and Hendrix, 2011). They are the most abundant and diverse biological entities, with an estimate of about 10^{32} tailed bacteriophages on Earth (Suttle, 2007). Since bacteriophages are parasites of bacteria, they have played an important role in the evolution of bacteria. Bacteriophages can have two alternate lifestyles when infecting a bacterium, lytic and temperate. Lytic phages simply replicate to form progeny virions which are released to infect other host cells. Temperate phages can also propagate lytically but may instead enter a semi-dormant “prophage” state in which the phage DNA either replicates as a plasmid or integrates into the host chromosome and replicates passively as part of that replicon. Prophages can be stable indefinitely, but environmental triggers can cause their “induction” into the lytic growth cycle.

Erwinia and *Pantoea* are very closely related Gram negative bacteria in the E*rwiniaceae* family of the *Enterobacteriales* order (F.S. Dworkin et al., 2006) that are often plant pathogens, causing necrosis in the tissues of the infected plant. These pathogens are a large burden on the agricultural community of the United States and are currently listed as possible bio-terrorism agents. For example, *Pantoea agglomerans* is the causative agent of potato blight and has also been documented as an opportunistic human pathogen (Sengupta et al., 2016), and *Erwinia amylovora* is the causative agent of fruit tree fire blight, which is responsible for an average of 100 million US dollars damage annually to apple orchards in the United States (Norelli et al., 2003). Fire blight infections are currently treated with antibiotics; however, up to 70% of these bacteria found in nature are resistant to the currently used antibiotics (Forster et al., 2015). Due to their ability to kill their bacterial hosts, phages are projected to provide an alternative anti-bacterial therapy for these plant diseases.

A number of bacteriophages that infect *Erwinia* or *Pantoea* (*Erwiniaceae* phages) have been isolated by a variety of investigators from several continents, and 60 of their complete genome sequences are available at the National Center for Biological Information (NCBI) GenBank database (Sayers et al., 2019). Thirty of these phages were isolated and characterized in our laboratory (Esplin et al., 2017; Sharma et al., 2018). Host range studies of phages that infect *Erwiniaceae* suggest relatedness between the two host genera in that several phages isolated on *Erwinia* can infect both *Erwinia* and *Pantoea* strains including phages Joad and RisingSun (Arens et al., 2018), Y3 (Buttimer et al., 2018), ØEa2809 (Lagonenko et al., 2015) and CBB (Buttimer et al., 2018). Herein, we compare these 60 bacteriophages and place

[*] Corresponding author.
E-mail address: julianne_grose@byu.ed (J.H. Grose).

https://doi.org/10.1016/j.virol.2019.06.005
Received 16 April 2019; Received in revised form 5 June 2019; Accepted 11 June 2019
Available online 18 June 2019

them in 20 different clusters based on their genomic and proteomic traits. An analysis of each of these clusters is provided, along with comparisons to known phages. The purpose of our analysis is four fold: 1) to gain insight into the relationship and interaction between different bacteriophage types and their host bacterial species, 2) to further understand the relationships among members of the *Enterobacteriales* order by comparing their bacteriophages, 3) to contribute to our understanding of overall bacteriophage diversity, and 4) to provide information that will aid in the treatment of the above plant diseases by development of improved phage therapy cocktail design and safety.

2. Materials and methods

2.1. Isolation, sequencing and assembly of phages

Thirty *Erwinia* phages were isolated by our laboratory at Brigham Young and 28 of these have been previously described (Esplin et al., 2017; Sharma et al., 2018). The genomes of the two previously undescribed phages Rebecca (accession No. MK514281) and Derbicus (MK514282) were sequenced from libraries made with the Illumina TruSeq DNA Nano kit and Illumina HiSeq 2500 sequencing (250-bp paired end). Genomes were assembled with Geneious (Kearse et al., 2012) version 8.1 using *de novo* assembly with medium-low sensitivity as described previously in Sharma et al. (Sharma et al., 2018). Coverage depths were 527–1615 (1015.6 average) for Rebecca and 183–1758 (572.3 average) for Derbicus, both phages circularized their genomes upon assembly.

2.2. Genomic analysis and comparison

Gepard (Krumsiek et al., 2007) was used to generate dot plots that compare nucleotide sequences of multiple genomes. Default settings (word size 10) were used to generate dot plots, however lower and upper color limit were increased in order to allow better image viewing. Geneious (Kearse et al., 2012) was used to align the sequences in an identity matrix using MAFFT plugin and setting parameters to auto-algorithm, a scoring matrix of 200PAM/k = 2, a gap open penalty of 1.53 and an offset value of 0.123. Phamdb (Lamine et al., 2016), an online version of Phamerator (Cresawn et al., 2011), a bioinformatic tool designed to compare bacteriophage genomes was used to visualize both nucleotide and protein similarity using kClust (Hauser et al., 2013). The default settings of PhamDB were used in this comparison. The cluster file generated by Phamerator was aligned using Janus (available on the DNA-master website https://phagesdb.org/DNAMaster/) and then used to generate a phylogenetic tree of the proteins using the SPLITStree program (Dress and Huson, 2004). BLASTp (Boratyn et al., 2013; Madden et al., 1996) was used from the NCBI website except for when accessed through Phamerator.

3. Results and discussion

3.1. Genomic and proteomic analyses separate the 60 Erwiniaceae bacteriophages into 20 clusters

A summary of the 60 *Erwiniaceae* phage genomes available in GenBank as of January 1, 2019 is provided in Table 1. These 60 phages were isolated in 10 countries, and 30 were isolated and characterized by our laboratory (Esplin et al., 2017; Sharma et al., 2018). Three phages, LIMElight, LIMEzero and Vid5, were isolated on *Pantoea* hosts, and 56 were isolated on *Erwinia* hosts (Adriaenssens et al., 2011). One phage, CBB, was isolated on *Pectobacterium* but forms plaques on a strain of *Erwinia* (Buttimer et al, 2017). Among those with a reported isolation location, many were found in infected trees or in the soil around them. Of the 30 phages we isolated, the genomes of only two, Joad and RisingSun, have been fully discussed in the literature (Arens et al., 2018), 26 have been reported in only genome announcements, and two (Rebecca and Derbicus) are first reported here.

The genomes of the *Erwiniaceae* phages range from 378,379 bp (phage CBB) to 29,564 bp (phage ENT90). The average genome length of the 991 other *Enterobacteriales* tailed phage genomes currently in the NCBI database is 81,187 bp, but the *Erwiniaceae* phages have an average genome length of 162,734 bp. Thus, *Erwiniaceae* bacteriophages comprise about five percent of the sequenced *Enterobacteriales* tailed phages, and the average genome size is almost double the overall average. Fig. 1 plots the length of all the *Enterobacteriales* tailed phage genomes and indicates the locations of the *Erwiniaceae* phages. The *Erwiniaceae* phage genome lengths are within the previously known extremes, but it is not known if their large average size is the result of isolation methods used, properties of the hosts or the skew in isolation sources toward trees and the soil around them.

The 60 *Erwiniaceae* phage whole genome nucleotide sequences were compared with Genome Pair Rapid Dotter (Gepard) (Krumsiek et al., 2007) (Fig. 2A). By the criterion of diagonal line strength, these phages fall into 20 clusters that have similarity over 50% of the phage genome as previously described (Grose and Casjens, 2014; Hatfull et al., 2010). The clusters in Fig. 2A are indicated by the founding *Erwiniaceae* phage in the group (the first sequence released in GenBank) unless the phage belongs to a previously-described *Enterobacteriales* cluster, in which case the previously published name for that cluster is used (Grose and Casjens, 2014). An Average Nucleotide Identity (Esplin et al., 2017) matrix was also constructed using Geneious (Kearse et al., 2012), and if phage clusters are defined so that each phage has ≥50% ANI with at least one other phage in the group and ≤24% ANI with phages from other clusters (Supplementary Table S1), the ANI grouping matches the dot plot-defined clusters perfectly. Our clusters correspond in general to genera or subfamilies that have been defined by the International Committee on Virus Taxonomy (ICTV), but a number of our clusters have not yet been formalized by that group.

In addition to genome nucleotide sequence analysis, whole proteome and single protein analyses support these 20 clusters. Whole proteome analysis was performed using Phamerator (Cresawn et al., 2011) to group the phage-encoded proteins into related "Phamilies", and SPLITSTree (Huson and Bryant, 2006) was used to infer relationships based on the Phamily content among the 59 annotated bacteriophages (Fig. 3; phage LS-2018a is not included because it has not been annotated). The SPLITSTree analysis perfectly parallels the cluster assignments generated by whole genome dot plot and ANI analysis above. It also points out the previously observed distant relationship between LIMEzero and LIMElight, which have previously been assigned to separate clusters within the T7 supercluster. Superclusters are groups of related phage clusters that share genome size and synteny (genes that have similar functions and have similar orders) that is not observed at the nucleotide level (Grose and Casjens, 2014). In addition to whole proteome analysis, single protein dot plot analysis was performed using the major capsid (MCP) (Fig. 2B) and large terminase (Fig. 2C) protein sequences, which have been previously used to place phages into related clusters (Grose and Casjens, 2014; Smith et al., 2013). Both of these plots agree with the clustering by the above methods and show similarities within each of the *Erwiniaceae* clusters and differences among them. The fact that all the above analyses give identical phage groupings demonstrates the robustness of such cluster determinations and indicates that the extent of past horizontal exchange of genetic information among these phages was not sufficient to disrupt their overall grouping. Thus, all these methods can be useful tools for

Table 1
Sixty *Erwinia* and *Pantoea* bacteriophages. Phages are organized by clusters (see text for definition of "clusters") which are indicated by different colored cells. The clusters are listed in order of descending genome size, and this group color scheme is carried throughout this report. The first column is the cluster as defined by Grose and Casjens (Casjens and Grose, 2016) when applicable. Pre-existing clusters are named according to the founding *Enterobacteriales* phage, and bold phage names in the second column indicate the first *Erwiniaceae* member of that cluster. N/A (not available) indicates phage genomes in GenBank that are otherwise not published. ǂ LS-2018a has a reported genome length of 59,759 bp, but this appears to be an untrimmed partial concatemer of the true sequence; the genome length given in the table putative properly trimmed sequence.

Cluster	Phage Name	Isolation Location	Isolation Source	Gene Bank Accession	Genome Length	Number of ORF's	Reference
RaK2-like	CBB	Little Island, Ireland	Waste water sludge	KU574722	378,379	605	15
Ea35-70-like	RAY	UT, USA	Leaves and Stem	KU886224	271,182	319	10
	Deimos-minion	UT, USA	Branches and Blossom	KU886225	273,501	326	10
	Special G	UT, USA	Branches and Blossom	KU886222	273,224	324	10
	Simmy50	UT, USA	Bark	KU886223	271,088	322	10
	Ea35-70	Canada	Soil	KF806589	271,084	314	76
	Desertfox	UT, USA	Soil	MG655268	272,485	320	N/A
	Bosolaphorus	UT, USA	Soil	MG655267	272,228	321	N/A
	Rebecca	UT, USA	Tree	MK514281	273,731	320	N/A
	MadMel	UT, USA	Soil	MG655269	275,000	321	N/A
	Mortimer	UT, USA	Unknown	MG655270	273,914	325	N/A
Yoloswag-like	**Yoloswag**	UT, USA	Unknown	KY448244	259,700	334	10
	Y3	Sursee, Switzerland	Soil, Apple tree	KY984068	261,365	333	13
	Alexandra	UT, USA	Unknown	MH248138	266,532	349	N/A
SPN3US-like	Asesino	UT, USA	Branches and Blossom	KX397364	246,291	289	N/A
	phiEaH2	Hungary	Unknown	JX316028	243,050	263	41
	Stratton	UT, USA	Unknown	KX397373	243,953	276	10
	Huxley	UT, USA	Branches and Blossom	KX397368	240,761	271	10
	Machina	UT, USA	Unknown	KX397370	241,654	272	10
	Parshik	UT, USA	Unknown	KX397371	241,050	271	10
	ChrisDB	UT, USA	Unknown	KX397366	244,840	277	10
	Caitlin	UT, USA	Branches and Blossom	KX397365	241,147	271	10
	Phobos	UT, USA	Unknown	KX397372	229,501	247	10
	EarlPhilipIV	UT, USA	Apple tree	KX397367	223,935	241	10
	Derbicus	UT, USA	Pear tree	MK514282	223,950	240	N/A
	Wellington	UT, USA	Unknown	MH426724	244,950	295	11
	Kwan	UT, USA	Unknown	KX397369	246,390	285	10
phiEaH1-like	**phiEaH1**	NCAIM, Hungary	Aerial tissue	KF623294	218,339	244	41
Joad-like	**Joad**	UT, USA	Pear tree	MF459647	235,374	245	10
	RisingSun	UT, USA	Apple tree	MF459646	235,108	243	10
T4-like	**Cronus**	Denmark	Organic waste	MH059636	175,774	295	N/A
Vi01-like	**øEa2809**	Belarus	Leaves of apple tree	KP037007	162,160	145	14
	Bue1	Switzerland	Soil from apple orchard	MG973030	164,037	178	N/A
Felix-O1-like	**øEa21-4**	Canada	Unknown	EU710883	84,576	117	51
	øEa104	Germany	Unknown	FQ482083	84,565	118	69
	M7	Switzerland	Unknown	HQ728263	84,694	117	1
	SunLiRen	USA	Unknown	MH426725	84,559	142	N/A
N4-like	**S6**	Switzerland	Unknown	HQ728266	74,669	115	1
	Frozen	UT, USA	Branches and Blossom	KX098389	75,147	92	10
	Rexella	UT, USA	Branches and Blossom	KX098390	75,448	92	10
	Gutmeister	UT, USA	Apple tree	KX098391	71,173	84	10
	Ea9-2	Canada	Soil	KF806588	75,568	89	N/A
	øEaP-8	South Korea	Unknown	MH160392	75,929	78	56
9g-like	**Vid5**	Lithuania	Thicket shadbush	MG948468	61,437	99	72
PEp14-like	**PEp14**	Korea	Unknown	JN585957	60,714	64	N/A
	Pavtok	UT, USA	Unknown	MH426726	61,401	62	N/A
Gj1-like	**Faunus**	Denmark	Organic waste	MH191398	54,065	78	N/A
	Y2	Switzerland	Unknown	NC019504	56,621	92	55
øEt88-like	**øEt88**	USA	Unknown	FQ482085	47,279	68	69
SP6-like	**Era103**	USA	Unknown	EF160123	45,445	53	33
	øEa100	USA	Unknown	FQ482086	45,554	51	69
	S2	Switzerland	Soil	MG736918	45,495	49	N/A
	øEa1H	USA	Unknown	FQ482084	45,522	50	69
KP34-like	**LIMElight**	Merelbeke, Belgium	Soil from potato	FR687252	44,546	55	24
LIMEzero-like	**LIMEzero**	Merelbeke, Belgium	Soil from potato	FR751545	43,032	57	24
T7-like	**FE44**	Ukraine	T2 phage contamination	KF700371	39,860	47	N/A
	L1	Switzerland	Unknown	HQ728265	39,282	51	1
LS-2018a-likeǂ	**LS-2018a**	MD, USA	Unknown	CP013974	31,798	N/A	N/A
P2-like	**ENT90**	South Korea	Unknown	HQ110084	29,564	60	N/A
	EtG	USA	Cucumber	MF276773	30,413	45	N/A

determining phage relationships, but the fact that all but dot plots do not point out mosaic relationships should not be forgotten, and in situations where horizontally exchanged, mosaically related sequences occur at higher frequency an ANI comparison may be less informative.

A summary of the 20 *Erwiniaceae* phage clusters is provided in Table 2, which shows that they range from eight singleton clusters to two clusters that contain nine or more phages. Phages from all three families of *Caudovirales* (*Podoviradae, Myoviradae* and *Siphoviridae*) have been isolated that infect *Erwiniaceae* bacteria. The number of annotated genes ranges from 47 (phage FE44) to 605 (CBB). The genome length is quite constant within each cluster, varying by at most 9%. As seen with other bacteriophages, *Erwiniaceae* phage genes are tightly packed with an average gene density of 1.2 ORFs (open reading frames)/kb. In Fig. 4 we plot the number of ORFs against the genome size of the founding *Erwiniaceae* phage of each group. Most lie close to the trend line, and we note that since this analysis is dependent on the annotation practices of different research groups, phages furthest from the line may not be as different as their locations suggest. A genomic map comparison of the founding phage members of each cluster is provided in Supplementary Fig. S1. It clearly shows the densely packed genomes of all 19 clusters that have annotated members.

The average G + C content of the *Erwiniaceae* tailed phage genomes is 48.5% and individual phages range from 38.4% to 55.4%. *Erwinia amylovora* is the most common host species for these phages, and its G + C content is 53.6%. *Pantoea agglomerans,* the most common *Pantoea* host is 55.1% G + C. Three phage clusters that have substantially lower G + C content than their host. Cronus, øEa21-4 and Faunus have G + C contents well below their host, with Cronus having the lowest at 38.4% percent. With a few exceptions, the G + C content of bacteriophage genomes is closely related to their target host (Bahir et al., 2009), making this drastic difference interesting. We note that phage Cronos belongs to the T4-like cluster (see below) in which other members are known to have substantially lower G + C contents than their hosts (Limor-Waisberg et al., 2011). Although the purpose for alternate G + C content is unknown, it has been suggested by some authors that lower G + C phages differ from their host in order to introduce their own set of tRNA's which favor the viral genome and the associated preferred codons (Limor-Waisberg et al., 2011).

3.2. Protein function among the Erwiniaceae phages

We selected one representative bacteriophage from each of the 19 annotated *Erwiniaceae* phage clusters and examined their predicted protein functions. Table 3 shows that of the 2667 genes annotated in these 19 phage genomes only 793 (30%) have a predicted function. Since BLASTp detected homology is commonly used to identify putative function, this means that 70% of the annotated genes have no database match or match a protein whose function is unknown. Phage Era103 (Vandenbergh et al., 1985) had the highest percent (63%) of genes called with a putative protein function, which may in part be due to its smaller genome. In most of the *Erwiniaceae* jumbo phages only 20–30% of the encoded proteins have predicted functions. Among those with putative functions, DNA replication and recombination genes are most abundant (35–52% of total proteins with function). Phage structural proteins were also commonly annotated, with the major capsid and large terminase proteins being identified in all 19 clusters.

3.3. Lifestyles of Erwiniaceae tailed phages

We attempted to determine whether the *Erwiniaceae* phages in this study are lytic or temperate by bio-informatic means. Most appear not to carry genes such as integrase that might be indicative of the temperate lifestyle, but since prophages may not be integrated this lack does not prove a lytic lifestyle. In our 2016 study (Casjens and Grose, 2016) we showed that the number of bacterial genome sequences that are available in the extant database is high enough that virtually all known prophage types are represented in those sequences. Therefore, if a newly identified phage is temperate it should have close relatives in extant bacterial genome sequences, especially in genomes of its host species or close relatives. Indeed, nearly all previously examined temperate *Enterobacteriales* phage clusters encode MCP relatives with ≥97% amino acid sequence identity to proteins encoded by prophages in *Enterobacteriales* bacterial host genomes, while no purely lytic phages had such closely related homologues (Casjens and Grose, 2016). We therefore searched for MCP genes similar to those of the 20 clusters described here in bacterial host genomes (Table 4). Of the 20 clusters, only three have homologues with >80% identity in bacterial genomes. *Erwinia* phage ENT90's MCP was 100% identical to a gene (locus_tag C2E16_18005) in a similar prophage in a *Pantoea sp.* PSNIH2, suggesting it is most likely temperate in nature which is consistent with its similarity to temperate *E. coli* phage P2. *Erwinia* phage øET88 MCP has a 97% identical homologue (locus_tag SAMN05216522_1056) as a putative prophage in the genome of *Rosenbergiella nectarea* strain 8N4; this species is a close relative of *Pantoea* and *Erwinia* (Halpern et al., 2013). In addition, Müller et al. (King et al, 2011) reported that øET88 was isolated after mitomycin C treatment of an *Erwinia tasmaniensis* strain, a treatment that often results in prophage induction. Finally, we have previously argued from genomic analysis that øET88 should be considered a member of the phage lambda supercluster (Grose and Casjens, 2014), and all other members of this large group are temperate. The putative phage LS-2018a MCP has 94% identical homologues encoded by the genomes of several *Yersinia pestis* isolates (*e.g.,* strain I-2638). At least one of these genes is present on a circular 34 kb plasmid (Acc. No. KT020860) that is largely homologous to the LS-2018a genome. Thus, we suggest that LS-2018a is very likely a temperate phage with a circular plasmid prophage. In addition, phage PEp-14 encodes a protein with some similarity to phage integrases, suggesting that it could be temperate in spite of the fact that its closest MCP matches in the reported bacterial genome sequences are ≤88% identical and are in very distantly related bacteria; however, it is possible that by chance no host genomes with PEp-14-like prophages have been sequenced. We also note that the *Burkholderia* phages BcepIL02 and Bcep22 have substantial genome synteny with PEp-14 and also carry an apparent integrase gene. They have been reported not to form stable lysogens but may be able to form a transient benign association with the host (Gill et al, 2011); on the other hand, in a single counter-example we find a protein that is 93% identical to Bcep22 MCP encoded by a gene (locus_tag WS71_20305) in an integrated prophage that is quite similar to Bcep22 in the genome of *Burkholderia* sp. DU8 (Acc. No. CP0013389). Thus, definitive determination whether phage PEp-14 is lytic or temperate awaits further study, but we conclude ENT90, øET88 and LS-2018a are almost certainly temperate, and the other 16 clusters discussed here most likely contain lytic phages. Bacterial matches included in Table 4 are all clearly inserted in the bacterial chromosome or in a known plasmid (it is possible that finding a fragment of a phage genome in a bacterial draft genome can be a result of a lytic phage infection at the time of sequencing).

3.4. The 20 clusters of Erwiniaceae tailed phages

Since we have shown that in the *Enterobacteriales* clusters MCP

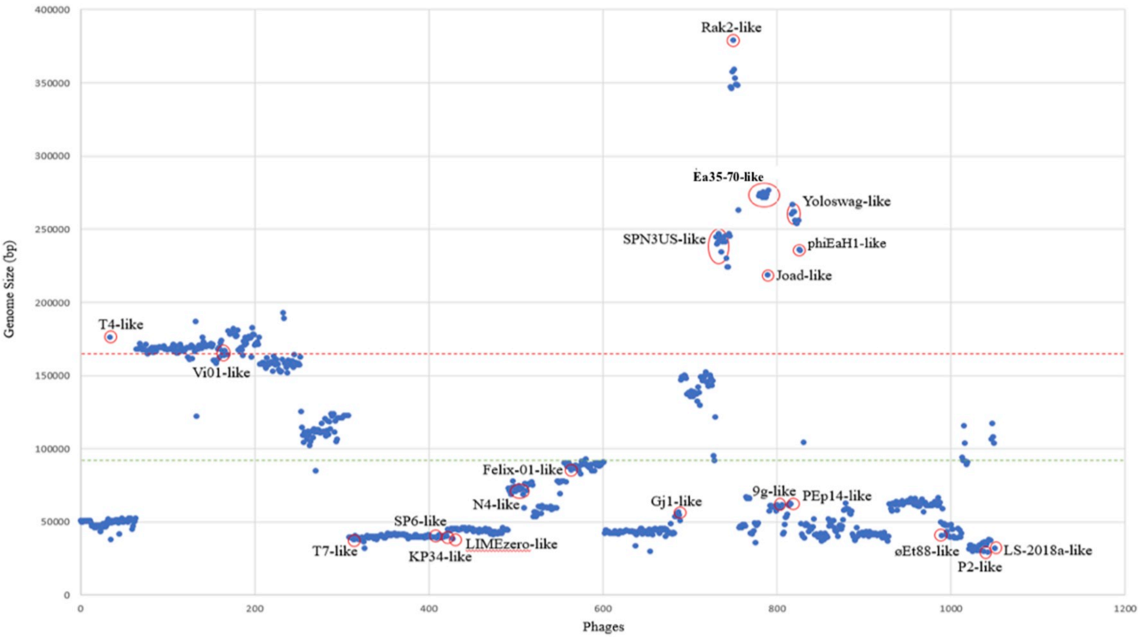

Fig. 1. Comparison of *Enterobacteriales* bacteriophage average genome size with the average *Erwiniaceae* phage genome reveals large *Erwiniaceae* phage genomes. Phage genome size is plotted on the y-axis for each of 1134 *Enterobacteriales* phages on the x-axis. The green dashed line represents the average genome length of all *Enterobacteriales* phages, and the red dashed line represents the average of all *Erwiniaceae* phage genome lengths. The red circles mark *Erwiniaceae* clusters.

sequence clustering nearly always reflects whole genome clustering (Grose and Casjens, 2014), BLASTp searches with MCPs from each of these clusters were first used to identify the most closely related non-*Erwiniaceae* phages. These results and subsequent whole genome nucleotide comparisons showed that 17 of the *Erwiniaceae* clusters can be placed in previously defined *Enterobacteriales* phage clusters (summarized in Table 2). Fig. 5A and B shows nucleotide sequence dot plots that compare phages from each of the 17 non-singleton clusters with their most closely related *Enterobacteriales* phages. Subcluster designations, indicating closer relationships, are provided in Table 1 (see Grose and Casjens for *Enterobacteriales* cluster/sub-cluster assignments). Three *Erwiniaceae* phage clusters typified by phages Yoloswag, Joad and LS-2018a represent novel *Enterobacteriales* tailed phage clusters that have not been previously described. The following paragraphs examine the molecular lifestyles of the 20 phage clusters with members that infect the *Erwiniaceae*:

3.4.1. Jumbo phages with genomes larger than 200 kb

(1) CBB was originally isolated on *Pectobacterium* but forms plaques on an *Erwinia* strain (Buttimer et al., 2017). It is the largest *Erwiniaceae* phage reported to date (Buttimer et al., 2017). CBB fits in the RaK2-like *Enterobacteriales* phage cluster and is most similar to *Cronobacter* phage GAP32. The nine known phages in the RaK2-like cluster form three subclusters and a representative from each cluster is shown in Fig. 5A. The RaK2-like phages are jumbo *Myoviridae* phages, and many or all are "hairy" with unusual whisker-like structural proteins along the contractile tail. The Phamerator map (Supplementary Fig. S1 indicates that the terminal regions of the CBB genome (as it is currently oriented in GenBank) share some similarity to other *Erwiniaceae* phages, specifically with the Cronus and ØEa2809 clusters. These related regions encode proteins annotated as hypothetical proteins and structural proteins. (For more information see reference (Abbasifar et al., 2014) for phage GAP32 characterization).

(2) The phiEaH2-like *Erwiniaceae* group fits into the previously defined *Enterobacteriales* SPN3US-like phage cluster (Grose and Casjens, 2014). This cluster consists of jumbo myoviruses with genomes in the 229–247 kb range (note that an error in Table 1 of reference (Grose and Casjens, 2014) places the SPN3US-like and Rak2-like clusters inside the rV5 supercluster, but this is incorrect). The SPN3US-like cluster also includes phages that infect *Salmonella, Escherichia* and *Cronobacter* hosts, and the dot plot in Fig. 5A shows that the 16 phages currently in this cluster separate into 9 subclusters, of which only subcluster A includes phage from multiple host genera (*Escherichia* and *Salmonella* including phages *SPN3US, SEGD1, NAFV-136*). The 13 *Erwinia* phages form 7 different subclusters that contain no other phages, highlighting the strong correlation between phage subclusters and host genus.

One of the noteworthy features of phage SPN3US is that it encodes a five subunit RNA polymerase that is packaged into the virion and injected into the host cell with the phage DNA and the Erwinia members carry similar genes. This group also shares a number of gene homologies with *Pseudomonas aeruginosa* phage øKZ (91 genes), including 61 virion structural genes (Aliet al, 2017; Thomas et al., 2016). (For more information see reference (Weintraub et al, 2019) for phage SPN3US characterization).

(3) The jumbo *Erwinia Siphoviridae* phage phiEaH1 has a 218 kb genome and is the prototypical member of the *Enterobacteriales* phiEaH1-like phage cluster (Meczker et al., 2014; Grose and Casjens, 2014). The only other phage in this cluster is *Serratia* phage 2050HW. These two phages are moderately distant relatives

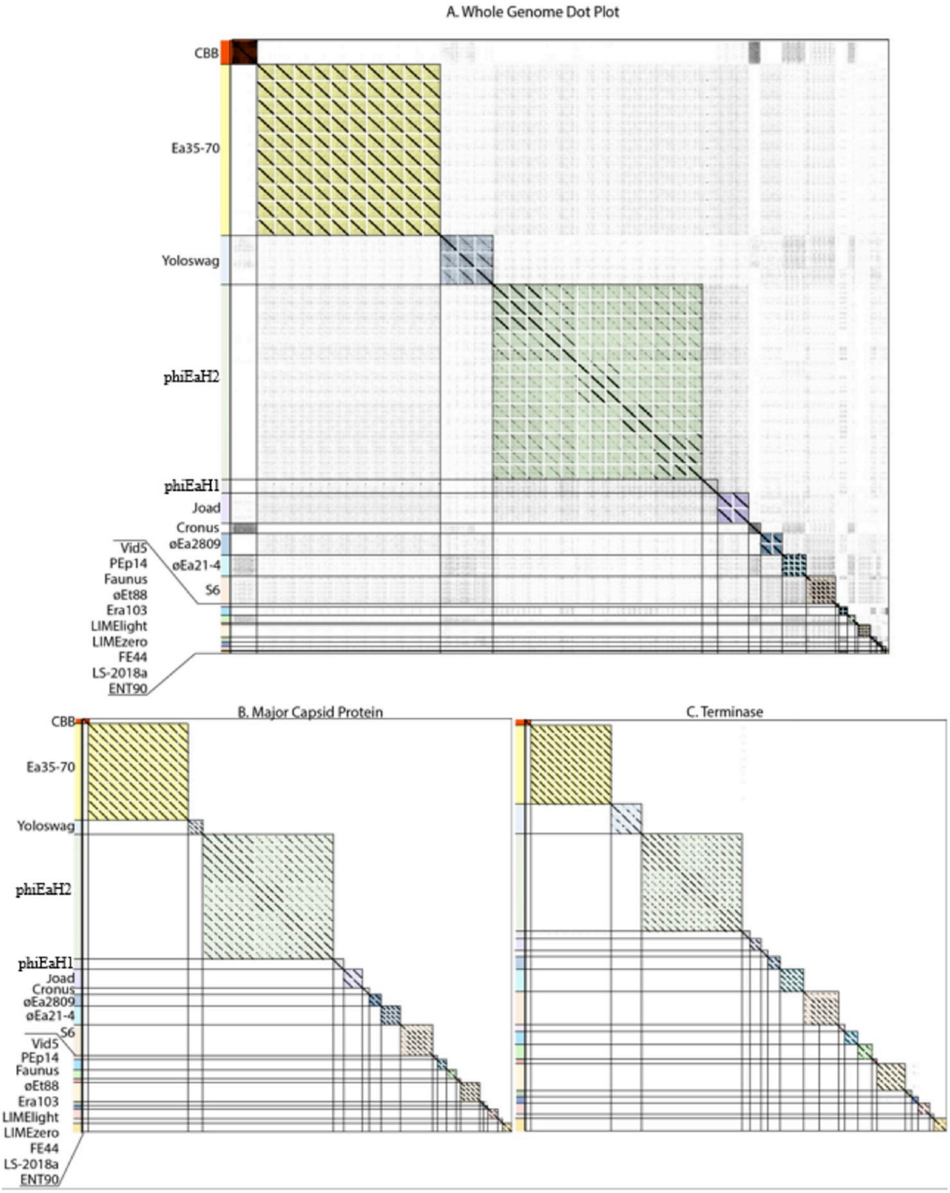

Fig. 2. Dot plots that compare the *Erwiniaceae* tailed phages reveal 20 clusters of related phages. (A) Whole genome nucleotide sequence dot plot. Sequences were reoriented to make parallel genome alignments within each cluster; the founding phage of each cluster (bold in Table 1) labels each whole cluster. (B) Major capsid protein (MCP) amino acid sequence dot plot. (C) Large terminase amino acid sequence dot plot. Horizontal and vertical black lines separate clusters, and white lines within the colored cluster boxes mark the ends of each phage genome. Dot plots were constructed using Gepard (Sharma et al., 2018). Note that the phage LS-2018a sequence was not annotated, but putative MCP and terminase were identified using tBLASTn (Boratyn et al., 2013).

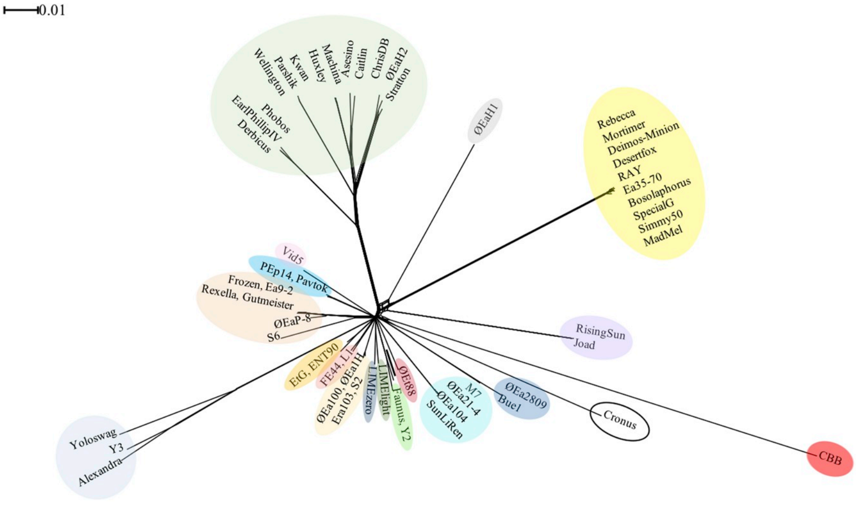

Fig. 3. A proteome phylogeny of 59 of *Erwiniaceae* tailed phages reveals 19 clusters of phages. Phamerator (NC_015249) was used to group phage proteins into phams of related proteins. SPLITStree software (Dress and Huson, 2004) was used to generate the tree from each pham's absence or presence in each phage genome. The phage LS-2018a genome has not been annotated and was therefore not used in this analysis.

Table 2

A summary of the 20 clusters of *Erwiniaceae* phages. The columns contain the group's name (given by founding phage from that group), the number of phages within a group, the average genome length within the group (with standard deviation), the average number of ORF's (with standard deviation), the ORF's/Genome Length (calculated from the average and supplied in ORF's/kb), the average GC content of each cluster (with standard error), the reported morphology, and closest non-*Erwiniaceae* phage relative of the cluster (well-known phages are selectively provided when available as a relative, otherwise less well-known phages are given). Note that phage LS-2018a has not been annotated, however we determined morphology bioinformatically. None – has no close relatives (*i.e.*, defines a novel cluster).

Erwiniaceae Group Name	Number of *Erwiniaceae* phages included	Average genome length	Number of ORF's	Number of ORF's/Genome length*1000	GC content	Morphology	Close Outside Relative of Cluster
CBB	1	378,379	605	1.6 ± 0.2	36.0	*Myoviridae*	RaK2
Ea35-70	10	272,744	321	1.2 ± 0.2	49.7	*Myoviridae*	None
Yoloswag	3	262,532	339	1.3 ± 0.2	48.1	*Myoviridae*	JA11
phiEaH2	13	239,344	269	1.1 ± 0.2	50.9	*Myoviridae*	SPN3US
phiEaH1	1	218,339	244	1.1 ± 0.2	52.3	*Myoviridae*	2050HW
Joad	2	235,241	244	1.0 ± 0.2	48.3	*Myoviridae*	None
Cronus	1	175,774	295	1.7 ± 0.2	38.4	*Myoviridae*	T4
øEa2809	2	163,099	162	1.0 ± 0.2	50.3	*Myoviridae*	Vi01
øEa21-4	4	84,599	124	1.5 ± 0.2	41.8	*Myoviridae*	Felix-01
S6	6	74,656	92	1.3 ± 0.2	47.8	*Podoviridae*	N4
Vid5	1	61,437	99	1.6 ± 0.2	48.8	*Siphoviridae*	9 g
PEp14	2	61,058	63	1.0 ± 0.2	50.0	*Podoviridae*	SopranoGao
Faunus	2	55,343	85	1.5 ± 0.2	43.9	*Myoviridae*	EcoM-GJ1
øEt88	1	47,279	68	1.4 ± 0.2	47.3	*Myoviridae*	T1
Era103	4	45,504	51	1.1 ± 0.2	49.8	*Podoviridae*	SP6
LIMElight	1	44,546	55	1.3 ± 0.2	54.0	*Podoviridae*	KP34
LIMEzero	1	43,032	57	1.3 ± 0.2	55.4	*Podoviridae*	J8-65
FE44	2	39,571	49	1.2 ± 0.2	50.3	*Podoviridae*	T7
LS-2018a	1	31,798	–	–	51.0	*Siphoviridae*	None
ENT90	2	29,989	53	1.8 ± 0.2	55.0	*Myoviridae*	P2
Overall average:	**3**	**162,734**	**198**	**1.2**	**48.5**		

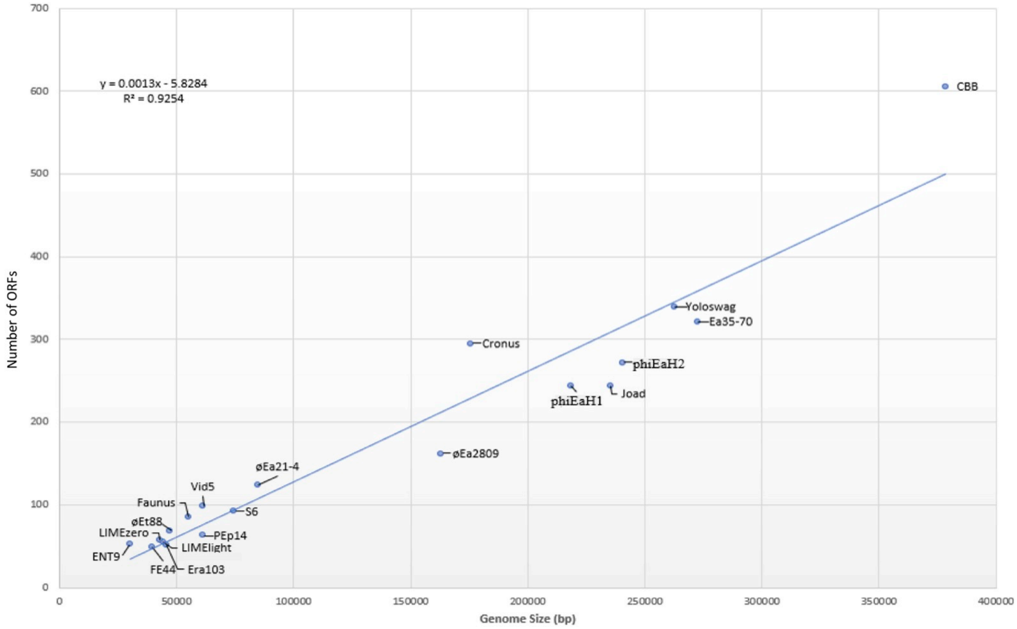

Fig. 4. Open reading frame density in *Erwiniaceae* bacteriophage clusters. Each cluster is labeled by the founding phage but represents the whole cluster's average. Equation and R2 value are displayed on the chart. The line represents a linear regression model of the average number of ORFs per phage compared to the average genome size of 19 *Erwinaceae* clusters.

sharing syntenic proteomes (with their MCP's sharing 56% identity and 71% similarity) and are only very distantly related at the nucleotide level (see Fig. 5A). (For more information see reference (Tian et al., 2019) for phage 2050HW characterization).

3.4.2. Myoviridae with genomes between 50 and 180 kb

(4) *Erwinia* phage Cronus forms a singleton subcluster in the *Enterobacteriales* T4-like *Myoviridae* cluster. Its genome size of 175 kb is typical of phages in this cluster, and like many other phages in this cluster its DNA has a substantially lower G + C content than its host. This cluster currently contains 169 completely sequenced genomes of phages that infect 14 host genera from six of the families within the *Enterobacteriales* order (Supplementary Table S2). The dot plot in Supplementary Fig. S2 Part B shows that there are 21 subclusters (A through U) in this cluster, one of which

Table 3

Putative gene functions reported from representative phages of each of the 19 annotated *Erwiniaceae* phage clusters. One representative phage from each of the 19 clusters was selected to analyze the protein function annotation. Protein function was sorted into four sections shown in different colors: structural proteins are in blue, DNA replication and recombination are in orange, cell lysis genes are in yellow, and host related genes are in green. Numbers refer to the number of proteins annotated for that function. We are aware of some possible overlap among protein function categories, this is due to the use of original annotations. LS-2018a is not represented in this table since it had no annotation.

	CBB	RAY	Yoloswag	Huxley	phiEaH1	Joad	Cronus	øEa2809	øEa21-4	S6	Vid5	PEp14	Faunus	øEt88	Era103	LIMElight	LIMEzero	FE44	ENT90
Head protein	3	1	1	1	1	5	9	3	1	2	3	2	1	3	2	2	2	4	4
Tail Fiber	5	1	11	3	3	4	6	11	5		5	2	1	2	3	3	5	4	7
Baseplate	3	1					8	3	2				2	2					3
Putative virion structural protein	60	26		38	29	26	1		3		1		2		1	2	2		
Neck/whisker	1						2	2			1								
Procapsid							2	3	1										
Terminase	2	1	1	2	1	1	2	2	1	1	2	2	1	2	2	1	1	1	2
DNA Polymerase	3	1	4	1		2	3	3	2	2	2		1		1	2	1	1	
RNA Polymerase	1	7	1	7	5	7	3			1			1		1	1	1	2	
Helicase	3	3	2	2	3	3	2	4	1	1	1		1	1	1	1	1	1	
Nuclease	9	5	6	3	3	5	9	8	2		3		2	2	5	2	2	4	
Hydrolase	4	2	2	2	3	3		2		3	2	1	2		1			2	
Recombination/ Repair	4	1	2		1	2	3	3						3					
Thymidine kinase/synthase	2	3	2	2	2	4	2	2	1	1			1						
Nucleotide reductase	5	2		1		1	4	2	3										
Topoisomerase	1		2				1	2											
Ligase	3		2			1	2	1	1		1		1		1	1		1	
Primase	2		1				1	1	1		1					1	1	1	
DNA-binding protein	1		5	1	1		3	4		1	1	2	1	1				2	
Lysin			2	1		2		1	1	1	1	3	2	1	2	1	1	2	2
Lysozyme	3	1	2				2								1				1
Holin							1		1		1	2	1	1	1	1			
Lysis inhibitor/regulator							4		1		1								
Lytic transglycosylase		2	2		2	2													
Integrase												1		1					1
Transcriptional/ Translational repressor protein	2		3	1		1	1	1										2	3
Nucleoid disruptor protein							1												
Secretion systems			4																
EPS		1	1	1	1			1			1	1			1	1			

Table 4

The closest tBLASTn match to the MCP of 20 *Erwiniaceae* bacteriophage clusters. The MCP of the founding phage for each group (see Table 2) was used in a tBLASTn search for closest relatives in bacterial genomes that were greater than 1 megabase. Erwiniaceae bacteriophage clusters that are not represented in this table had no significant tBLASTn hits.

Phage	Best tBLASTn Bacterial Match	Accession Number	Identity
øEa21-4	*Polyangium brachysporum* strain DSM 7029	CP011371	32%
S6	*Alteromonas sp.* RKMC-009	CP031010	56%
Vid5	*Nitrosomonas ureae* strain Nm10	CP013341	47%
	[a]*Enterococcus faecalis* strain TY1	CP031027	35%
PEp14	*Martelella* sp AD-3	CP014275	75%
Faunus	*Rhizobiales* strain PAMC 29148	CP036515	29%
	[a]*Enterobacter cloacae* strain 20710	CP030076	28%
øEt88	*Rosenbergiella nectarea* strain 8N4	CP009706	97%
Era103	*Pandoraea faecigallinarum* strain DSM 23572	CP011807	30%
LIMElight	*Cronobacter sakazakii* strain ATCC 29544	CP011047	41%
LIMEzero	*Enterobacter kobei* strain DSM 13645	CP017181	52%
LS-2018a	*Yersinia pestis* strain I-2638	CP013974	94%
ENT90	*Pantoea sp.* PSNIH2	CP009866	100%

[a] Closest *Enterobacteriales* bacteria.

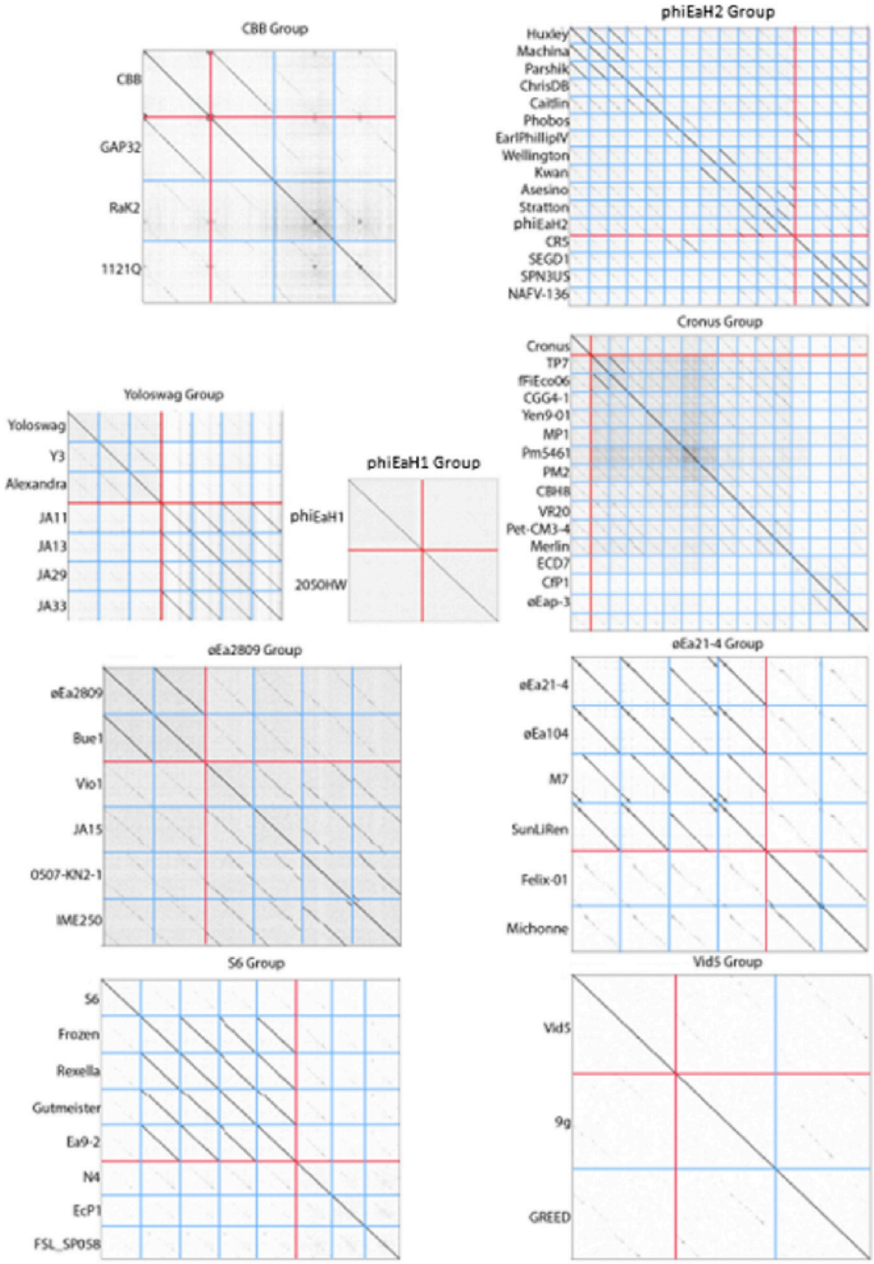

Fig. 5. Dot plots of 17 *Erwinia*ceae phage clusters with their relatives. Red lines separate *Erwiniaceae* bacteriophage clusters and homologous *Enterobacteriales* phage genomes. Blue lines indicate the ends of each genome. Parts A and B depict nine and eight phage clusters, respectively. Due to the large number of phages in some of the phage clusters only representative phages are shown.

is defined by phage Cronus. Six of the subclusters are singletons, but of the 15 subclusters with more than one member, 11 contain members that all infect the same host genus (assuming that *Escherichia* and *Shigella* are actually one genus (Zuo et al., 2013); and all but one has members that infect a single host family. Thus, subcluster membership is far from random, with many genus-specific or family-specific subclusters at this level of analysis. We also note that diversity within this cluster is still quite incompletely understood (NC_015249), since (i) the almost 30% singleton subclusters implies the existence of numerous undiscovered subclusters, (ii) individual genera are often infected by multiple phage subclusters, and (iii) phages of a number of *Enterobacteriales* families and genera remain unexplored.

(5) The *Erwinia* phages øEa2809 (Lagonenko et al., 2015) and Bue1 (accession No. MG973030) share similarity to the *Enterobacteriales* Vi01-like cluster of *Myoviridae*, which is currently comprised of 51 *Enterobacteriales* phages, including *E. coli* phage CBA120 and *Salmonella enterica* phage Det7, that typically have genomes in the 150–165 kb range. These phages have virion structural genes that are moderately distant relatives of those of phage T4, but their

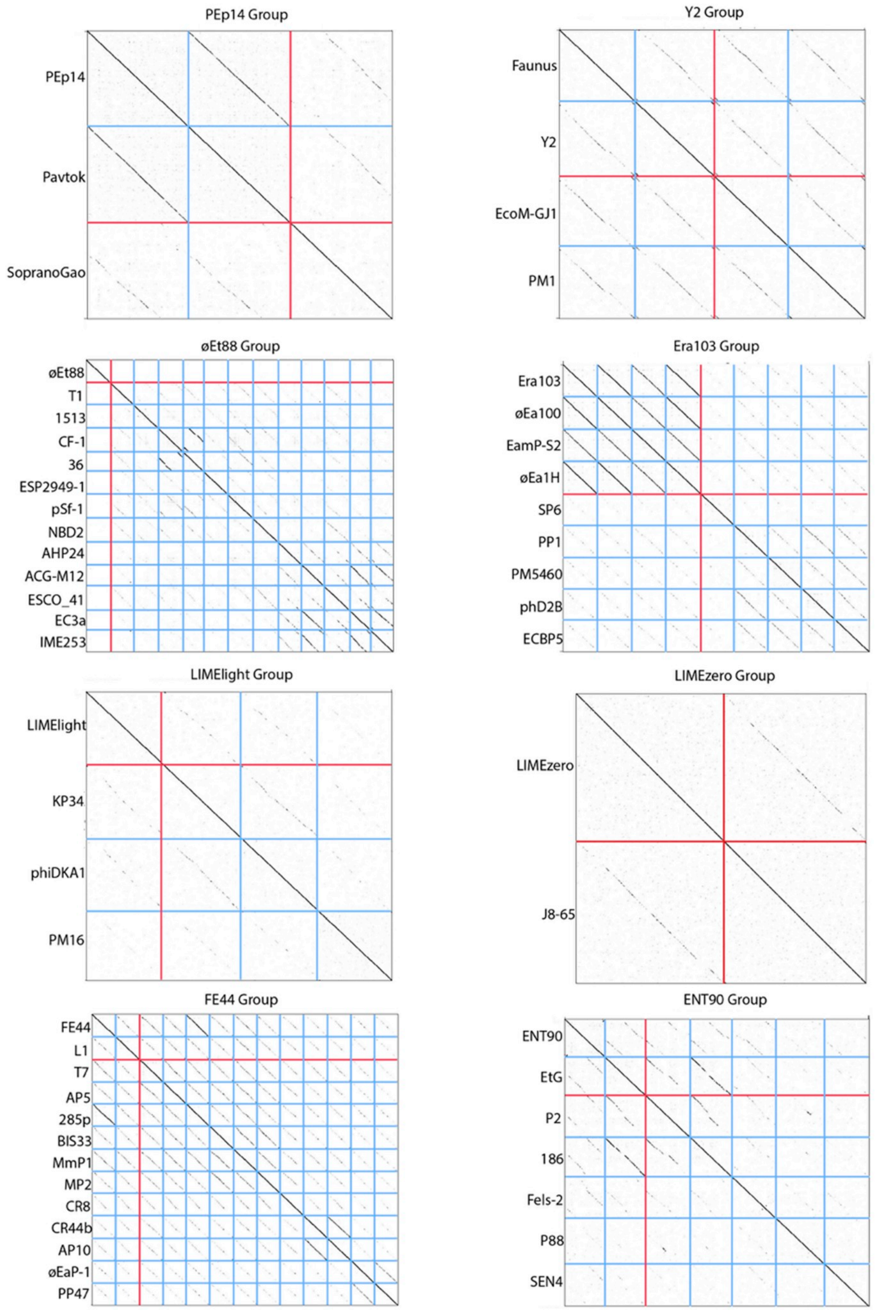

Fig. 5. (*continued*)

virion heads are isomorphic rather than elongated, and their homologous genes are not syntenic with the T4-like phages. They encode a thymidylate synthase that suggests they may incorporate hydroxymethyldeoxyuracil into their DNA (Kutter et al., 2011), and they encode multiple tailspikes that allow them to adsorb to several different hosts (Adriaenssens et al., 2012; Chen et al., 2014; Casjens et al., 2015). This cluster was previously separated into at least six subclusters, one of which is comprised of only the two *Erwinia* phages øEa2809 (Lagonenko et al., 2015) and Bue1, phages that were isolated in Belarus and Switzerland, respectively (Fig. 5A). (For more information see reference (Kutter et al., 2011) for phage CBA120 characterization).

(6) The øEa21-4-like *Erwinia* phage group lies within the previously defined Felix-O1-like *Enterobacteriales* phage cluster (Grose and Casjens, 2014; Magill et al., 2017; Lehman et al., 2009). This cluster of contractile tailed phages have genome sizes that range from 82 to 91 kb and carry a number tRNA genes which are highly conserved across the øEa21-4 group. The Felix-O1-like cluster currently contains 46 completely sequenced *Enterobacteriales* phages that fall into three subclusters (Grose and Casjens, 2014), and the four *Erwinia*

phages in this cluster form one of these subclusters (Fig. 5A). The known phages in this cluster infect six different *Enterobacteriales* host genera, and there are fairly close relatives that infect *P. aeruginosa* in the *Pseudomonadales* order of Gamma-Proteobacteria (Magill et al., 2017). (For more information see reference (Cowley et al., 2015) for phage TP1 characterization).

(7) The phage Y2-like *Erwinia* group has similarity to the previously defined *Enterobacteriales* lytic *Myoviridae* phage øEcoM-Gj1-like cluster (Grose and Casjens, 2014), currently containing four subclusters. This cluster is currently comprised of two *Escherichia* phages, øEcoM-Gj1 (Jamalludeen et al., 2008) and ST32, two *Pectobacterium* phages, PM1 (Kalischuk et al., 2015) and PP101, and two *Erwinia* phages, Faunus and Y2 (Born et al., 2015). The last two are sufficiently different that they each form a distinct singleton subcluster (Fig. 5B). These phages have genomes in the 52–57 kb range and encode a single subunit RNA polymerase like phage T7 (Jamalludeen et al., 2008). (No phages in this cluster have been extensively characterized).

3.4.3. Lytic Podoviridae phage

(8) The *Erwinia* S6-like group fits into the previously defined *Enterobacteriales* N4-like cluster of *Podoviriadae* phages. The 26 currently known completely sequenced members of this cluster fall into six subclusters, three of which, typified by phages Ea9-2, S6 (Born et al, 2011) and øEaP-8 (Park et al., 2018), are made up by the seven *Erwinia* phages and no others; the last two are singleton subclusters (Fig. 5A). The larger group of N4-like phages appears to be a very successful group of phages whose members infect other Gamma-Proteobacteria orders as well as Beta-Proteobacteria hosts (*e.g.,* N4-like phage JWDelta infects the Beta-Proteobacteria *Achromobacter xylosoxidans* (Wittmann et al., 2014)*)*. A unique feature of this group is its large (about 3500 amino acid) single subunit RNA polymerase that is present in the virion and is injected with the DNA into the host cell (Falco et al., 1977).

(9) Phages PEp-14 and Pavtok define an *Erwinia* group that expands the previously defined *Enterobacteriales* PEp-14-like *Podoviridae* singleton cluster (Grose and Casjens, 2014). *Klebsiella* phage SopranoGao is also a recently sequenced member of this cluster, but the two *Erwinia* phages form a unique subcluster (Fig. 5A). As discussed above it remains unclear whether these phages are temperate or lytic. A striking feature of these phages, that have genomes about 61 kb long, is that they encode an exceptionally large putative protein that is 4915, 5007 and 4369 amino acids long in the PEp-14 (Acc. No. YP005098431), Pavtok (AXF51455) and SopranoGao (ASV45029) homologues, respectively. These single genes occupy about a quarter of their genomes, and their products are the longest bacteriophage encoded proteins that we are aware of. Other classes of large phage proteins are the virion RNA polymerases of the N4-like phages (above) (Kazmierczak et al., 2002) and a possible tail fiber of øKO2 at 3433 AA (Casjen et al., 2004). BLASTp searches with the large Pavtok protein (locus_tag PAVTOK_25) have shown that it shares patches of convincing similarity to large proteins in the following phages that infect diverse hosts: ≥50% identity to *Vibrio* phage VvAW1 (3640 AA; Gamma-Proteobacteria host), *Pseudomonas* phage Skulduggery (3695 AA; Gamma-Proteobacteria host), *Agrobacterium* phage atu_ph08 (4877 AA; Alpha-Proteobacteria), and *Sinorhizobium* phage PBC5 (2849 AA; Alpha-Proteobacteria host), as well as 35% identity to proteins from several Beta-Proteobacteria phages including *Burkholderia* phages Bcep22 (4602 AA). The function of these large proteins has not been studied directly, but two sequence matches are informative. First, amino acids 70–170 of all three of the PEp-14-like cluster phages' large protein contain a lysozyme motif and are 33% identical to a section of phage T7 gene *16* protein. There are a small number of molecules of *16* protein in the T7 virion, and they are released into the host with the DNA (Kemp et al., 2005; Chang et al., 2010). Many tailed phage that infect Gram negative bacteria are thought to inject proteins with lysozyme activity that cleave the peptidoglycan so that DNA can pass through it to reach the cytoplasm during injection (Casjens and Molineux, 2012; Moak and Molineux, 2004), and the T7 gene *16* protein has been shown to have such an activity (Moak and Molineux, 2000). Second, a region between amino acids 2700 and 3200 of the PEp-14-like large proteins have weak but convincing similarity to parts of *E. coli* phage P1 DarB protein (2255 AA; accession No. YP_006479), which has also been shown to be injected with the DNA (Piya et al., 2017) and is involved in defense against host restriction endonucleases (Iida et al., 1987). We conclude that it is very likely that these large PEp-14-like cluster proteins are present in the virions and are injected into the host with the DNA. Gill et al. (Gill et al, 2011) have made a similar argument with the homologous large gp75 protein of *Burkholderia* phage Bcep22 (Acc. No. NP944303), which has been shown to be a virion protein. Why are these PEp-14-like phage proteins and their homologues so large? We speculate that when a phage "finds" a new protein function that is advantageous to inject from the virion, it may be evolutionarily simplest to fuse it to an existing protein that is injected. Thus, such proteins may accumulate new polypeptide sections and become large multidomain proteins over time. This would also explain the patchy nature of the relationships between such proteins in different phages. We note that the distantly related phages mentioned above all have similarity to the leftmost approximately 37 kb of the PEp-14-like phages (in the Pavtok GenBank orientation), a region that contains the putative virion assembly genes (Supplementary Material Fig. S3 shows a comparison of phage Pavtok with *Burkholderia* phage DC1/Bcep22). (No phage in this cluster has been extensively characterized).

(10-13) The four *Erwiniaceae* phage clusters discussed in this section fall into the previously defined *Enterobacteriales* T7-, SP6-, KP34- and LIMEZERO-like clusters (Grose and Casjens, 2014), which in turn all reside within the T7 supercluster (classified by the International Committee on Virus Taxonomy as the *Autographivirinae* subfamily of the *Podoviridae*). They all have apparent lytic life cycles similar to phage T7 which infects *E. coli* (Demerec and Fano, 1945) and is one of the best characterized and most prolific tailed bacteriophages. It has many known relatives that infect a wide variety of bacterial hosts, even outside of the *Enterobacteriales.* Hallmarks of these phages include a phage encoded single subunit RNA polymerase.

Erwinia phage ERA103 fits into the Enterobacteriales SP6-like cluster, where it, along with Erwinia phages øEa100 (Muller et al., 2011), øEa1H and S2, form the Erwinia specific subcluster D. Pantoea phage LIMElight belongs to the KP34-like cluster where it forms the singleton subcluster B. Pantoea phage LIMEzero is the prototype phage for the LIMEzero-like cluster, which also contains Escherichia phage J8-65. These two phages define different subclusters. Finally, Erwinia phage FE44 shares its highest overall nucleotide sequence identity of

91–94% to Escherichia phages 285P, BA14 and S523 (Xu et al., 2014; Michalewicz et al., 1991) and is a member of the T7-like cluster. FE44, along with phages that infect the Escherichia, Yersinia, Salmonella, Kluyvera and Pectobacterium genera, form subcluster C of this Enterobacteriales cluster (Fig. 5B).

3.4.4. Lytic Siphoviridae phage

(14) *Pantoea* phage Vid5 is a member of the *Enterobacteriales* 9 g-like cluster of lytic phages (26, 72). This *Siphoviridae* cluster's founding member phage 9 g has deoxy-archaeosine (modified guanosine) nucleotides in its DNA that make it resistant to many restriction endonucleases (Kulikov et al., 2014). Vid5 has a similar but not identical set of genes predicted to be involved in this or a similar DNA modification, and its DNA is similarly resistant to such nucleases (Simoliunas et al., 2018). The 15 phages with available complete genomes in this cluster fall into three subclusters, two of which have been called the *Nonagvirus* and *Seuratvirus* genera (Sazinas et al., 2018), and the third is Vid5 which forms a singleton subcluster. The dot plot in Fig. 5A compares representatives of these three subclusters; subclusters A and B are all *Escherichia* phages except for one *Salmonella* phage (phage SE1; accession No. KY926791) in subcluster A. (For more information see reference (Kulikov et al., 2014) for phage 9 g characterization).

3.4.5. Temperate Myoviridae phage

(15) *Erwinia* phage EtG is quite closely related to *Escherichia* phage 186 and *Salmonella* phage PsP3, and although ENT90 is more distantly related, both of these *Erwinia* phages are clearly members of the P2-like *Enterobacteriales* temperate phage cluster (Grose and Casjens, 2014) (Fig. 5B) (Kalionis et al., 1986). Phages in this cluster are widely distributed with phages that infect many different types of *Enterobacteriales* (Casjens and Grose, 2016), and EtG belongs to subcluster B that also contains phages that infect *Escherichia* and *Salmonella*, while ENT90 defines singleton subcluster D.

3.4.6. Clusters that currently contain only Erwinia/Pantoea phages

Although most of the currently known *Erwiniaceae* tailed phages fall into to one of the over 70 previously defined *Enterobacteriales* phage clusters (Casjens and Grose, 2016; Grose and Casjens, 2014), five of 20 *Erwiniaceae* phage-containing clusters contain only phages that infect the *Erwinia* and/or *Pantoea* genera (the Ea35-70-, Yoloswag-, Joad-, LS-2018a- and øET88-like phages). Three of these five clusters (Yoloswag-, Joad-, and LS02018a-like) form novel *Enterobacteriales* clusters that have not been previously described.

(16) The *Erwiniaceae* phages within the previously defined Ea35-70-like *Enterobacteriales* phage cluster (Grose and Casjens, 2014) form the most highly conserved of all of the *Erwiniaceae* clusters we analyzed, with less than 3% ANI variance among the phages within this cluster. It is comprised of jumbo *Myoviridae* phages typified by phage Ea35-70 that was isolated from soil beneath a fire blight-infected pear tree in Ontario, Canada (Yagubi et al., 2014). No similar phages are known that infect other host species. More than 60% of their 271–275 kb genomes are made up of novel genes without significant BLASTp matches in the current database, and like other jumbo phages their small fraction of genes with predicted functions encode mainly virion structural proteins and DNA metabolism proteins. (No phage in this cluster has been extensively characterized).

(17) Phage Yoloswag represents an *Erwinia* jumbo phage group that includes two closely related phages, Alexandra and Y3. Five *Dickeya* phages have recently been described whose putative MCPs are about 74% identical to that of Yoloswag, and Fig. 5A shows a dot plot analysis of these eight phages. Long weak diagonal similarity lines confirm that they all have similar genome organization and belong in this previously undefined *Enterobacteriales* phage cluster which we call the Yoloswag-like cluster. It also shows that substantial diversity is present within the cluster, and we define three subclusters, A, B and C. Subclusters A and B are quite different from C, and the three phages within B are more diverse (weaker diagonal similarity line) than those within C. Interestingly, the clusters do not correlate perfectly with host genus, since subcluster B contains phages with *Erwinia* and *Dickeya* hosts, and these two genera have recently been placed in the two different but rather closely related families, *Erwiniaceae* and *Pectobacteriaceae.* This suggest that one of these phages, perhaps AD1, has switched hosts in the relatively recent distant past. The proteins that are expressed by the conserved core-genome of the Yoloswag-like *Erwinia* phages are mostly virion structural proteins and DNA replication and repair proteins. Recent publications describing the phage Y3 genome sequence (Buttimer et al., 2018) and the four *Dickeya* relative genomes (Day et al., 2018) have presented various aspects of this group of phages, so we will not discuss them in detail here but will only briefly mention some of this cluster's salient features. Our Phamerator (Cresawn et al., 2011) analysis shows that there are 176 protein Phamilies conserved among the three *Erwinia* members of this cluster, with only 46 of these proteins having a predicted function, including four secretion system proteins (products of the phage Yoloswag genes *88, 107, 152* and *154*) that are not present in any of the other *Erwinia* phages. A conserved Cas4-like protein is also encoded by all three *Erwinia* phages. Cas4 has no well-defined function but is known to be present in CRISPR/Cas gene clusters. It could be carried by the phage to modify CRISPR systems in its hosts. The virions of this cluster have large isometric heads about 130 nm in diameter and a contractile tail about 190 nm long. An interesting reported feature of at least Y3 and the *Dickeya* members of this cluster is the presence of unusual curly hair-like fibrils of unknown function extending from the sheath along the length of the tail similar to those seen in the RaK2-like phages (above). (No phage in this cluster has been extensively characterized).

(18) The jumbo *Myoviridae* phages Joad and RisingSun represent a new *Enterobacteriales* tailed phage cluster with 235 kb genomes (see Fig. 2 of ref (Arens et al., 2018).). These two *Erwinia* phages share 96.6% whole genome ANI, with Joad encoding two genes not present in RisingSun (a putative HNH endonuclease and a hypothetical protein). The RisingSun genome encodes 243 predicted proteins, ~43% of which have no significant BLASTP database match (e-value of $\geq 10^{-7}$); another 24% of its genes have no known function but do have BLASTP matches to hypothetical proteins (Arens et al., 2018). This novel cluster shares some homology with *Pseudomonas* phages EL and OBP as well as *Vibrio* phages P4B and pTD1, with 112 genes that had corresponding BLASTp hits with these *Pseudomonas* phages, indicating these phages are clearly related.

(19) *Erwinia* phage LS-2018a also represents (as a singleton) a new *Enterobacteriales* tailed phage cluster. Its sequence in GenBank contains very large terminal redundancy (if a small amount of

sequence imprecision is allowed), and we believe it very likely has a circular genome that is 31,789 bp long. Its sequence in GenBank is unannotated, but we find a 97% identical homologue of its putative MCP (bp 29319–30479 of accession No. CP013974) and a similar terminase encoded by several isolates of *Yersinia pestis.* In *Y. pestis* biovar Medievalis strain I-2638 the 33,778 bp long circular plasmid pTP33 (MCP encoded between bp 11182 and 12342 of accession No. KT020860 (Kutyrev et al., 2018)) encodes such an MCP homologue (the other matches are on *Yersinia* sequence contigs that are the same size or smaller but are not annotated as plasmids). The dot plot in Supplementary Material Fig. S2 shows that LS-2018a and pTP33 share considerable syntenic similarity and that (with some, not unexpected, mosaicism) they have nearly identical genome organizations. We conclude that pTP33 is very likely a circular plasmid prophage and that LS-2018a may have a similar prophage (although we note that both carry a possible integrase gene). The *Yersinia* genus is a member of the newly defined *Yersiniaceae* family in the *Enterobacteriales* (Adeolu et al., 2016), and LS-2018a and pTP33 represent two singleton subclusters, each of which infects a different host family.

(20) *Erwinia* phage øET88 is the singleton representative of its *Enterobacteriales* cluster (Grose and Casjens, 2014). Although its MCP is up to 49% identical to some phages in the T1-like lytic phage cluster, it is likely a temperate member of the phage lambda supercluster (Grose and Casjens, 2014) (see above).

Eighteen of the 20 *Erwiniaceae* clusters contain more than one authentic phage member (øEt88 and LS-2018a comprise singleton clusters). Within each of these clusters, most of the *Erwiniaceae* phages are more closely related to one another than to phages that infect other host genera and so form distinct subclusters. Nonetheless, eight of the 33 *Erwiniaceae* phage-containing subclusters also contain phages that infect other genera in addition to *Erwinia* and *Pantoea* (Fig. 5), indicating a few possible examples of relatively recent host switching.

4. Conclusions

The purpose of this analysis was to gain insight into the relationships among the 60 *Erwiniaceae* bacteriophages that have completely sequenced genomes and to further understanding of their host interactions. We found that on average *Erwiniaceae* phages have much larger genomes than the average *Enterobacteriales* phage, which may be due to their isolation source (trees and the soil surrounding them) or may be driven by the bacterial host. We note that only dsDNA tailed phages infective of *Erwinia* and *Pantoea* have been isolated to date. When the nucleotide and protein sequences of these 60 phages are compared, they naturally separate into 20 clusters, or 3 phages/cluster on average. This ratio highlights the diversity present in these phages in spite of the fact that they share highly related hosts. In comparison, 472 *E. coli* phages currently in GenBank fall into 50 clusters with an average of 9.4 *E. coli* phages/cluster. The lower phages/cluster ratio for *Erwiniaceae* phages (3 phages/cluster) may not be due to the decreased number of total phages isolated because *Paenibacillus* phages have a comparable number of isolates, but a phages/cluster ratio more similar to *E. coli* (9.6 phages/cluster). Comparison of the 60 *Erwiniaceae* phage genomes with all the other *Enterobacteriales* phage genomes, showed that 17 of the *Erwiniaceae* clusters belong to previously defined *Enterobacteriales* phage clusters that include phages with hosts outside this family, and three form clusters whose known members infect only *Erwiniaceae*. The *Erwiniaceae* phages in the 17 *Enterobacteriales* clusters tend form their own subcluster within their clusters. This latter distinction is perhaps due to the plant-based ecological niche of *Erwinia* and *Pantoea*. A majority of the proteins encoded by the *Erwiniaceae* phages (~70% or 1874 proteins) have unknown function, highlighting the need for further characterization of these phages. Each of the 19 analyzed *Erwiniaceae* bacteriophage clusters encodes unique proteins, including tail fibers, lysins, holins, and CRISPER proteins, which likely contribute to the phage host range and will be important considerations in the development and improvement of phage therapy cocktail design and safety.

Declarations

Competing interests

We declare that there were no competing interests with any author or institution responsible for this manuscript.

Funding

This research was funded by the department of Microbiology and Molecular Biology and the College of Life Sciences at Brigham Young University.

Authors contributions

DWT, SRC and JHG all performed analysis of the phage genomes and wrote the manuscript. RS annotated phages Rebecca and Derbicus and performed analysis of the Ea35-70 phage group.

Acknowledgements

The authors extend a special thanks to the Science Education Alliance – Phage Hunters Advancing Genomics and Evolutionary Science (SEA-PHAGES) for support and training on phage analysis, Steven Cresawn training on Phamerator, Edward Wilcox from the BYU Sequencing Center, and David Parady and Brandon Ekins from Life Science IT at BYU for providing DNAmaster and PhamDB on BYU servers. We would also like to thank the following undergraduates from the BYU Phage hunters program for their contributions to isolation of the phages Derbicus and Rebecca: Nolan Beatty, Jordan Berg, Minsey Choi, Steven Duncan, Rebecca Eardley, Christopher Fajardo, Hannah Ferguson, Jared Kruger, Ashley Payne and C.J. Webb, as well as Professors Donald Breakwell and Sandra Hope from the BYU Phage Hunters program.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.virol.2019.06.005.

References

Abbasifar, R., et al., 2014. Supersize me: *Cronobacter sakazakii* phage GAP32. Virology 460–461, 138–146.

Adeolu, M., Alnajar, S., Naushad, S., S.G.R, 2016. Genome-based phylogeny and taxonomy of the *'Enterobacteriales'*: proposal for *Enterobacterales* ord. nov. divided into the families *Enterobacteriaceae, Erwiniaceae* fam. nov., *Pectobacteriaceae* fam. nov., *Yersiniaceae* fam. nov., *Hafniaceae* fam. nov., *Morganellaceae* fam. nov., and *Budviciaceae* fam. nov. Int. J. Syst. Evol. Microbiol. 66, 5575–5599.

Adriaenssens, E.M., et al., 2011. Bacteriophages LIMElight and LIMEzero of *Pantoea agglomerans*, belonging to the "ØKMV-like viruses" Appl. Environ. Microbiol. 77, 3443–3450.

Adriaenssens, E.M., et al., 2012. A suggested new bacteriophage genus: "Viunalikevirus"

Arch. Virol. 157, 2035–2046.

Ali, B., et al., 2017. To Be or not to Be T4: evidence of a complex evolutionary pathway of head structure and assembly in giant *Salmonella* Virus SPN3US. Front. Microbiol. 8, 2251.

Arens, D.K., et al., 2018. Characterization of two related Erwinia myoviruses that are distant relatives of the ØKZ-like Jumbo phages. PLoS One 13, e0200202.

Bahir, I., Fromer, M., Prat, Y., Linial, M., 2009. Viral adaptation to host: a proteome-based analysis of codon usage and amino acid preferences. Mol. Syst. Biol. 5, 311.

Boratyn, G.M., et al., 2013. BLAST: a more efficient report with usability improvements. Nucleic Acids Res. 41, W29–W33.

Born, Y., Bosshard, L., Duffy, B., Loessner, M.J., Fieseler, L., 2015. Protection of *Erwinia amylovora* bacteriophage Y2 from UV-induced damage by natural compounds. Bacteriophage 5, e1074330.

Born, Y., et al., 2011. Novel virulent and broad-host-range *Erwinia amylovora* bacteriophages reveal a high degree of mosaicism and a relationship to *Enterobacteriaceae* phages. Appl. Environ. Microbiol. 77, 5945–5954.

Buttimer, C., et al., 2017. Things are getting hairy: enterobacteria bacteriophage vB_PcaM_CBB. Front. Microbiol. 8, 44.

Buttimer, C., et al., 2018. *Erwinia Amylovora* Phage vB_EamM_Y3 Represents Another Lineage of Hairy Myoviridae. Res Microbiol.

Casjen, S.R., et al., 2004. The pKO2 linear plasmid prophage of *Klebsiella oxytoca*. J. Bacteriol. 186, 1818–1832.

Casjens, S.R., Grose, J.H., 2016. Contributions of P2- and P22-like prophages to understanding the enormous diversity and abundance of tailed bacteriophages. Virology 496, 255–276.

Casjens, S.R., Molineux, I.J., 2012. Short noncontractile tail machines: adsorption and DNA delivery by podoviruses. Adv. Exp. Med. Biol. 726, 143–179.

Casjens, S.R., Jacobs-Sera, D., Hatfull, G.F., Hendrix, R.W., 2015. Genome sequence of *Salmonella enterica* phage Det7. Genome Announc. 3.

Chang, C.Y., Kemp, P., Molineux, I.J., 2010. Gp15 and gp16 cooperate in translocating bacteriophage T7 DNA into the infected cell. Virology 398, 176–186.

Chen, C., et al., 2014. Crystal structure of ORF210 from *E. coli* O157:H1 phage CBA120 (TSP1), a putative tailspike protein. PLoS One 9, e93156.

Cowley, L.A., et al., 2015. Analysis of whole genome sequencing for the *Escherichia coli* O157:H7 typing phages. BMC Genomics 16, 271.

Cresawn, S.G., et al., 2011. Phamerator: a bioinformatic tool for comparative bacteriophage genomics. BMC Bioinf. 12, 395.

Day, A., Ahn, J., Salmond, G.P.C., 2018. Jumbo bacteriophages are represented within an increasing diversity of environmental viruses infecting the emerging phytopathogen, *Dickeya solani*. Front. Microbiol. 9, 2169.

Demerec, M., Fano, U., 1945. Bacteriophage-resistant mutants in *Escherichia coli*. Genetics 30, 119–136.

Dress, A.W., Huson, D.H., 2004. Constructing splits graphs. IEEE ACM Trans. Comput. Biol. Bioinform 1, 109–115.

Esplin, I.N.D., et al., 2017. Genome sequences of 19 novel *Erwinia amylovora* bacteriophages. Genome Announc. 5.

Falco, S.C., Laan, K.V., Rothman-Denes, L.B., 1977. Virion-associated RNA polymerase required for bacteriophage N4 development. Proc. Natl. Acad. Sci. U. S. A. 74, 520–523.

Fiers, W., et al., 1976. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. Nature 260, 500–507.

Forster, H., McGhee, G.C., Sundin, G.W., Adaskaveg, J.E., 2015. Characterization of streptomycin resistance in isolates of *Erwinia amylovora* in California. Phytopathology 105, 1302–1310.

K.C.I, 2006. Erwinia and related genera. In: In: F.S. Dworkin, M., Rosenberg, E., Schleifer, K.H., Stackebrandt, E. (Eds.), The Prokaryotes Springer, New York, NY.

Gill, J.J., et al., 2011. Genomes and characterization of phages Bcep22 and BcepIL02, founders of a novel phage type in *Burkholderia cenocepacia*. J. Bacteriol. 193, 5300–5313.

Grose, J.H., Casjens, S.R., 2014. Understanding the enormous diversity of bacteriophages: the tailed phages that infect the bacterial family Enterobacteriaceae. Virology 468–470, 421–443.

Halpern, M., Fridman, S., Atamna-Ismaeel, N., Izhaki, I., 2013. *Rosenbergiella nectarea* gen. nov., sp. nov., in the family *Enterobacteriaceae*, isolated from floral nectar. Int. J. Syst. Evol. Microbiol. 63, 4259–4265.

Hatfull, G.F., Hendrix, R.W., 2011. Bacteriophages and their genomes. Curr. Opin. Virol. 1, 298–303.

Hatfull, G.F., et al., 2010. Comparative genomic analysis of 60 Mycobacteriophage genomes: genome clustering, gene acquisition, and gene size. J. Mol. Biol. 397, 119–143.

Hauser, M., Mayer, C.E., Soding, J., 2013. kClust: fast and sensitive clustering of large protein sequence databases. BMC Bioinf. 14, 248.

Huson, D.H., Bryant, D., 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 23, 254–267.

Iida, S., Streiff, M.B., Bickle, T.A., Arber, W., 1987. Two DNA antirestriction systems of bacteriophage P1, darA, and darB: characterization of darA- phages. Virology 157, 156–166.

Jamalludeen, N., et al., 2008. Complete genomic sequence of bacteriophage φEcoM-GJ1, a novel phage that has myovirus morphology and a podovirus-like RNA polymerase. Appl. Environ. Microbiol. 74, 516–525.

Kalionis, B., Dodd, I.B., Egan, J.B., 1986. Control of gene expression in the P2-related template coliphages. III. DNA sequence of the major control region of phage 186. J. Mol. Biol. 191, 199–209.

Kalischuk, M., Hachey, J., Kawchuk, L., 2015. Complete genome sequence of phytopathogenic *Pectobacterium atrosepticum* bacteriophage Peat1. Genome Announc. 3.

Kazmierczak, K.M., Davydova, E.K., Mustaev, A.A., Rothman-Denes, L.B., 2002. The phage N4 virion RNA polymerase catalytic domain is related to single-subunit RNA polymerases. EMBO J. 21, 5815–5823.

Kearse, M., et al., 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28, 1647–1649.

Kemp, P., Garcia, L.R., Molineux, I.J., 2005. Changes in bacteriophage T7 virion structure at the initiation of infection. Virology 340, 307–317.

King, R.A., et al., 2011. Newly discovered antiterminator RNAs in bacteriophage. J. Bacteriol. 193, 5784–5792.

Krumsiek, J., Arnold, R., Rattei, T., 2007. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. Bioinformatics 23, 1026–1028.

Kulikov, E.E., et al., 2014. Genomic sequencing and biological characteristics of a novel *Escherichia coli* bacteriophage 9g, a putative representative of a new Siphoviridae genus. Viruses 6, 5077–5092.

Kutter, E.M., et al., 2011. Characterization of a ViI-like phage specific to Escherichia coli O157:H7. Virol. J. 8, 430.

Kutyrev, V.V., et al., 2018. Phylogeny and classification of *Yersinia pestis* through the lens of strains from the plague foci of commonwealth of independent states. Front. Microbiol. 9, 1106.

Lagonenko, A.L., Sadovskaya, O., Valentovich, L.N., Evtushenkov, A.N., 2015. Characterization of a new ViI-like *Erwinia amylovora* bacteriophage ØEa2809. FEMS Microbiol. Lett. 362.

Lamine, J.G., DeJong, R.J., Nelesen, S.M., 2016. PhamDB: a web-based application for building Phamerator databases. Bioinformatics 32, 2026–2028.

Lehman, S.M., Kropinski, A.M., Castle, A.J., Svircev, A.M., 2009. Complete genome of the broad-host-range *Erwinia amylovora* phage ØEa21-4 and its relationship to *Salmonella* phage felix O1. Appl. Environ. Microbiol. 75, 2139–2147.

Limor-Waisberg, K., Carmi, A., Scherz, A., Pilpel, Y., Furman, I., 2011. Specialization versus adaptation: two strategies employed by cyanophages to enhance their translation efficiencies. Nucleic Acids Res. 39, 6016–6028.

Madden, T.L., Tatusov, R.L., Zhang, J., 1996. Applications of network BLAST server. Methods Enzymol. 266, 131–141.

Magill, D.J., et al., 2017. Pf16 and ØPMW: expanding the realm of Pseudomonas putida bacteriophages. PLoS One 12, e0184307.

Meczker, K., et al., 2014. The genome of the *Erwinia amylovora* phage PhiEaH1 reveals greater diversity and broadens the applicability of phages for the treatment of fire blight. FEMS Microbiol. Lett. 350, 25–27.

Michalewicz, J., Hsu, E., Larson, J.J., Nicholson, A.W., 1991. Physical map and genetic early region of the T7-related coliphage, BA14. Gene 98, 89–93.

Moak, M., Molineux, I.J., 2000. Role of the Gp16 lytic transglycosylase motif in bacteriophage T7 virions at the initiation of infection. Mol. Microbiol. 37, 345–355.

Moak, M., Molineux, I.J., 2004. Peptidoglycan hydrolytic activities associated with bacteriophage virions. Mol. Microbiol. 51, 1169–1183.

Muller, I., Kube, M., Reinhardt, R., Jelkmann, W., Geider, K., 2011. Complete genome sequences of three Erwinia amylovora phages isolated in north America and a bacteriophage induced from an *Erwinia tasmaniensis* strain. J. Bacteriol. 193, 795–796.

NC_015249.

Norelli, J.L., Jones, A.L., Aldwinckle, H.S., 2003. Fire blight management in the twenty-first century: using new technologies that enhance host resistance in apple. Plant Dis. 87, 756–765.

Park, J., Lee, G.M., Kim, D., Park, D.H., Oh, C.S., 2018. Characterization of the lytic bacteriophage ØEaP-8 effective against both *Erwinia amylovora* and *Erwinia pyrifoliae* causing severe diseases in apple and pear. Plant Pathol. J. 34, 445–450.

Piya, D., Vara, L., Russell, W.K., Young, R., Gill, J.J., 2017. The multicomponent antirestriction system of phage P1 is linked to capsid morphogenesis. Mol. Microbiol. 105, 399–412.

Sayers, E.W., et al., 2019. Database resources of the national center for biotechnology information. Nucleic Acids Res. 47, D23–D28.

Sazinas, P., et al., 2018. Comparative genomics of bacteriophage of the genus *Seuratvirus*. Genome Biol. Evol. 10, 72–76.

Sengupta, M., Banerjee, S., Das, N.K., Guchhait, P., Misra, S., 2016. Early onset neonatal septicaemia caused by *Pantoea agglomerans*. J. Clin. Diagn. Res. 10, DD01–02.

Sharma, R., et al., 2018. Genome sequences of nine *Erwinia amylovora* bacteriophages. Microbiol. Resour. Announc. 7.

Simoliunas, E., et al., 2018. *Pantoea* bacteriophage vB_PagS_Vid5: a low-temperature siphovirus that harbors a cluster of genes involved in the biosynthesis of archaeosine. Viruses 10.

Smith, K.C., et al., 2013. Phage cluster relationships identified through single gene analysis. BMC Genomics 14, 410.

Suttle, C.A., 2007. Marine viruses–major players in the global ecosystem. Nat. Rev. Microbiol. 5, 801–812.

Thomas, J.A., et al., 2016. Identification of essential genes in the Salmonella phage SPN3US reveals novel insights into giant phage head structure and assembly. J. Virol. 90, 10284–10298.

Tian, C., et al., 2019. Identification and molecular characterization of *Serratia marcescens* phages vB_SmaA_2050H1 and vB_SmaM_2050HW. Arch. Virol. 164, 1085–1094.

Vandenbergh, P.A., Wright, A.M., Vidaver, A.K., 1985. Partial purification and characterization of a polysaccharide depolymerase associated with phage-infected *Erwinia amylovora*. Appl. Environ. Microbiol. 49, 994–996.

Weintraub, S.T., et al., 2019. Global proteomic profiling of *Salmonella* infection by a giant phage. J. Virol. 93.

Wittmann, J., et al., 2014. First genome sequences of *Achromobacter* phages reveal new members of the N4 family. Virol. J. 11, 14.

Xu, B., Ma, X., Xiong, H., Li, Y., 2014. Complete genome sequence of 285P, a novel T7-like polyvalent E. coli bacteriophage. Virus Gene. 48, 528–533.

Yagubi, A.I., Castle, A.J., Kropinski, A.M., Banks, T.W., Svircev, A.M., 2014. Complete genome sequence of *Erwinia amylovora* bacteriophage vB_EamM_Ea35-70. Genome Announc. 2.

Zuo, G., Xu, Z., Hao, B., 2013. *Shigella* strains are not clones of *Escherichia coli* but sister species in the genus *Escherichia*. Genom. Proteom. Bioinform. 11, 61–65.