INSTITUT PASTEUR

Research in Microbiology 162 (2011) 737–746

www.elsevier.com/locate/resmic

# A modular view of the bacteriophage genomic space: identification of host and lifestyle marker modules

Gipsi Lima-Mendez [a], Ariane Toussaint [b,*], Raphael Leplae [c]

[a] *Department of Applied Biological Sciences, Faculty of Sciences, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium*
[b] *Laboratoire de Bioinformatique des Génomes et des Réseaux (BiGRe), Université Libre de Bruxelles, CP263, Bvd du Triomphe, 1050 Bruxelles, Belgium*
[c] *Département d'informatique, Centre de Calcul, Univeristé Libre de Bruxelles, Campus du Solbosch, CP197, Avenue F.D. Roosevelt 50, 1050 Bruxelles, Belgium*

Received 10 February 2011; accepted 25 May 2011
Available online 28 June 2011

**Abstract**

Bacteriophage genomes can be regarded as an ensemble of modules which are accessible to the whole phage population via recombination. The time spent by prophages in the bacterial host provides them with the opportunity to exchange modules with other prophages or infecting phages. Here we analyze the modular structure of a set of 457 phages and 760 prophages extracted from completely sequenced bacterial genomes using the ACLAME database and its associated tools. We identified 91 modules of proteins with similar phylogenetic profiles. Of these, 25 and 6 are associated with temperate and virulent phages, respectively; 57 are restricted to a host or small group of hosts; and 55 could be annotated with a phage function. We use the transposable phages as a study case and show how the inclusion of prophages allows us to unveil new types of genome organization (i.e. novel module combinations) and obtain insight into the host range for this particular group, highlighting the utility of prophage prediction to better characterize phage diversity.
© 2011 Institut Pasteur. Published by Elsevier Masson SAS. All rights reserved.

*Keywords:* Bacteriophage; Functional module; Transposable phage; Prophage; Phage evolution; Phage classification

## 1. Introduction

Originally observed on lambdoid phages, i.e. phages able to recombine with paradigm phage λ to produce viable hybrids, the mosaic nature of phage genomes has been confirmed for other phage families and lively discussed in the literature [(Nelson, 2004) and references therein]. A modular theory of phage evolution was first proposed many years ago (Botstein, 1980). It states that the units of selection are functional modules which are shuffled by recombination, resulting in new combinations of modules and thus potentially novel and viable phages. Along the same line, it has been suggested that any double-stranded (ds) DNA phage may draw a given gene from another ds DNA phage, although the frequency of recombination events would depend on the phylogenetic distance between the respective hosts (Hendrix et al., 1999). Because there is no single gene shared by all phages (Rohwer and Edwards, 2002) or even by phages with similar morphology (Lavigne et al., 2008), phage population studies and phage classification are difficult tasks. The reference phage taxonomy is maintained by the International Committee on the Taxonomy of Viruses (ICTV), based on the nature of the phage genetic material and morphology and on other molecular features such as the packaging mechanism, presence of RNA polymerase, etc. However this classification has yet to take into account genome sequence relationships. Recently, sequence information was used to define genera within some of the ICTV *Podoviridae* and *Myoviridae* phage families (Lavigne et al., 2008; Lavigne et al., 2009). The classification procedure uses a hierarchical clustering method that relies on overall sequence similarity and is capable of reproducing ICTV grouping for a substantial fraction of the

* Corresponding author. Tel.: +32 (0) 26505499; fax: +32 (0) 26505425.
*E-mail addresses:* gipsi.lima@vib-vub.be (G. Lima-Mendez), ariane@bigre.ulb.ac.be (A. Toussaint), Raphael.Leplae@ulb.ac.be (R. Leplae).

0923-2508/$ - see front matter © 2011 Institut Pasteur. Published by Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.resmic.2011.06.006

sequenced phages. However, in order to define boundaries in the phage population, genome comparisons have to be narrowed to the most conserved and major phage features used in the ICTV classification.

In 2002, Lawrence et al. (Lawrence et al., 2002) proposed overcoming the limitations of hierarchical classification of phage genomes by introducing the notion of "modus", i.e. a group of modules or phenotypes that typifies a group of phages. Modi defined a reticulate classification within which one phage can belong to several modi, introducing the idea of a network of relationships. This concept was implemented into a phage network (Lima-Mendez et al., 2008a), represented by a weighted graph, where nodes are phage genomes and edges represent phage–phage similarities in terms of the number of shared genes/proteins. Alternatively, the network and reticulate classification can be based on sharing evolutionarily conserved modules (ECMs). These ECMs consist of groups of proteins that have a similar phylogenetic profile (i.e. they co-occur in genomes), which is an indication of a common function (Pellegrini et al., 1999). Most of the ECMs derived from 300 phages in a previous study (Lima-Mendez et al., 2008b) contained proteins with unknown function. With the aim of increasing the informational content of the ECMs, we extend this analysis to a larger set: 457 phages and 760 prophages all accessible in the ACLAME database version 0.4 (Leplae et al., 2010). Many of the ECMs obtained with this data set are composed of different functional modules, as inferred from the function of their constituent proteins. Some are constrained to a group of hosts and/or are preferentially found in either temperate or virulent phages. Finally, we focus on the subset of transposable phages and prophages to illustrate how, by starting from ECMs, one can apprehend the very mosaic nature of these genomes.

## 2. Materials and methods

### 2.1. Building the prophage set

The ACLAME database provides a list of 1058 prophages predicted with the Prophinder prediction tool (Lima-Mendez et al., 2008a,b) for 727 complete bacterial genomes (http://aclame.ulb.ac.be/perl/Aclame/Prophages/view_prophage.cgi). Proteins encoded by 1058 predicted prophages and 457 phage genomes (translated CDS) are available in the ACLAME database as protein families. A network was built based on the number of shared families and clustered using the MCL algorithm as described in (Lima-Mendez et al., 2008b). Predicted prophages that did not cluster with any phage genome were analyzed manually; the presence of insertion sequence (IS) transposases and the absence of 'bona-fide' phage proteins (e.g. terminase, portal, major capsid) were used as criteria to remove them from the prophage set. This procedure resulted in a set of 760 prophages.

The ACLAME classification and analysis procedures (Leplae et al., 2010) were applied to this new set of phage and prophage genomes to produce the protein families, the reticulate classifications and the ECMs. They are accessible in the categories 'Prophages' and 'Viruses and prophage' at http://aclame.ulb.ac.be/ and were used in this study.

### 2.2. Statistical analysis

All analyses in this work consisted of defining elements shared within the phage and prophage populations studied. To assess the probability of such common elements occurring by chance alone, we measured statistical significance via hypergeometric distribution using a threshold for significance of one false-positive among all tests (Bonferroni multiple-test correction, (Abdi, 2007). Phage, host, lifestyle and module were considered as separate classes to which sets of elements were assigned. Initially, module and phage classes were described as collections of protein families, and host and lifestyle classes as a collection of phages. The composition of classes made of the same type of element was compared to measure the association of the corresponding classes. The comparisons performed were as follows:

Module – phage: when the number of protein families in common between a given module and a given phage was statistically significant, the phage was considered to have the module. The modules were then described as a collection of phages, which allowed testing the module-host and module-lifestyle associations.

Module – host: when the number of phages in common between a module and a host was statistically significant, we considered the module to be associated with the host.

Module – lifestyle: when looking for modules associated with the virulent or temperate character of a phage, we excluded prophages from analysis. Because temperate phages are predominant in the phage set (202 vs. 110 virulent and 56 unassigned), a statistically significant association between an ECM and the temperate character required a higher number of phages having the ECM than for virulent phages. Therefore, in addition to statistically significant associations, we chose to report all associations where at least four temperate phages had the ECM (and no virulent phage had it) because these may provide interesting biological information. These cases are clearly distinguished in the text and in Table 1.

## 3. Results and discussion

### 3.1. Phage and prophage diversity

#### 3.1.1. Phage and prophage distribution among bacterial hosts

Among the 285 bacterial hosts in ACLAME version 0.4 and either supporting a phage infection or hosting a predicted prophage, 50% had only one associated phage or prophage (data not shown). Fig. 1A shows the number of phages and prophages associated with these bacteria. The bacteriophage host was retrieved from the GenBank sequence record, which does not specify the bacterial strain in all cases, explaining the high number of phages observed for some hosts (phages from all strains are pooled). Interestingly, 20 bacteria for which no phage had been sequenced hosted predicted prophages. At the

Table 1
Evolutionarily cohesive modules.

| Module ID | No. of families | Function | Host | %DNA identity | Life style | Phage type |
|---|---|---|---|---|---|---|
| 1 | 94 | lytic cycle | ***Aeromonas***, ***Prochlorococcus*** | 14, 0 | V | T4 |
| 2 | 49 | ru, t, L, UF | ***Staphylococcus*** | 14 | T | 80alpha |
| 3 | 24 | H, t, UF | ***Bacillus***, *Brucella*, ***Rhodobacter***, ***Staphylococcus*** | 5, 92, 5, 25 | T | HK97, P27 |
| 4 | 23 | Re, Ru, H, TM, UF | | | T | HP1 |
| 5 | 18 | h, t, P | ***Streptococcus*** | 14 | T | |
| 6 | 17 | Rc, Ru, L, UF | ***Escherichia***, ***Salmonella*** | 11, 10 | T | P22, λ, HK97 |
| 7 | 17 | B, TM, F, L | ***Bartonella*** | 37 | T | P2, BcepMu |
| 8 | 16 | H, UF | | | | |
| 9 | 16 | H, UF | | | T | D3112, BcepMu, Mu,B3 |
| 10 | 15 | re, H, UF | ***Streptococcus*** | 34 | | |
| 11 | 14 | H, TS | ***Streptococcus***, ***Staphylococcus*** | 14, 23 | T | 80alpha, SPP1 |
| 12 | 13 | P, UF | *Staphylococcus* | 50 | t | |
| 13 | 13 | T, L | ***Escherichia*** | 10 | | λ, HK97 |
| 14 | 12 | Re, Rc, H, A | | | V | T7 |
| 15 | 11 | B, TM | *Yersinia* | 92 | t | P27, Mu |
| 16 | 11 | H | *Streptococcus* | 45 | | A118 |
| 17 | 10 | IE, Re, Ru | ***Streptococcus***, ***Staphylococcus***, ***Lactobacillus***, ***Lactococcus*** | 9, 16, 2, 10 | T | 80alpha, P22, A118, HK97, P27 |
| 18 | 10 | Re, H, L | ***Escherichia*** | 30 | | |
| 19 | 10 | L, UF | ***Clostridium*** | 2 | t | |
| 20 | 9 | H, UF | | | | λ, P22 |
| 21 | 9 | NA | *Streptococcus* | 70 | | |
| 22 | 9 | H, f, L | | | V | T7 |
| 23 | 9 | H, T, f | | | | P22 |
| 24 | 8 | h, t, UF | *Staphylococcus* | 64 | t | |
| 25 | 8 | ru, UF | *Escherichia* | 65 | t | |
| 26 | 8 | Re | ***Aeromonas***, ***Prochlorococcus*** | 15, 0 | V | T4 |
| 27 | 8 | Re, h, t | ***Mycobacterium*** | 16 | T | L5 |
| 28 | 8 | h, UF | *Yersinia* | 47 | | |
| 29 | 7 | Re, UF | ***Streptococcus*** | 15 | | |
| 30 | 7 | h, t, UF | *Brucella* | 80 | | |
| 31 | 7 | Rc | ***Escherichia*** | 23 | T | λ |
| 32 | 7 | t, f, L | ***Mycobacterium*** | 3 | | L5 |
| 33 | 7 | H, T | ***Lactobacillus***, ***Lactococcus*** | 4, 32 | | |
| 34 | 6 | UF | ***Staphylococcus*** | 24 | T | SPP1 |
| 35 | 6 | Re, UF | *Staphylococcus* | 48 | | |
| 36 | 6 | Rc, UF | ***Streptococcus*** | 29 | | |
| 37 | 6 | h, UF | | | T | |
| 38 | 6 | t | ***Pseudomonas*** | 11 | | D3112, B3 |
| 39 | 6 | UF | | | t | |
| 40 | 5 | t, UF | ***Mycobacterium*** | 7 | | |
| 41 | 5 | H | ***Escherichia*** | 25 | | λ |
| 42 | 5 | UF | ***Bacillus*** | 14 | | |
| 43 | 5 | Re | | | T | BcepMu, B3 |
| 44 | 5 | UF | *Streptococcus* | 54 | | |
| 45 | 4 | UF | | | V | |
| 46 | 3 | Rc | | | | |
| 47 | 3 | Re | | | T | D3112 |
| 48 | 3 | Re, UF | ***Escherichia*** | 34 | t | |
| 49 | 3 | IS transposition | ***Shigella*** | 17 | | |
| 50 | 3 | L | ***Burkholderia*** | 32 | t | |
| 51 | 3 | UF | ***Bacillus*** | 22 | | |
| 52 | 3 | UF | | | | |
| 53 | 3 | Re | ***Vibrio*** | 17 | | |
| 54 | 3 | Ru | ***Lactococcus*** | 16 | | |
| 55 | 3 | UF | ***Lactococcus*** | 27 | t | |
| 56 | 3 | UF | | | | BcepMu |
| 57 | 3 | H | | | | |
| 58 | 3 | UF | | | | |
| 59 | 3 | UF | ***Streptococcus*** | 69 | | |
| 60 | 3 | UF | *Staphylococcus* | 57 | t | |

(*continued*)

Table 1 (*continued*)

| Module ID | No. of families | Function | Host | %DNA identity | Life style | Phage type |
|---|---|---|---|---|---|---|
| 61 | 2 | UF | | | | |
| 62 | 2 | UF | *Yersinia* | 96 | | |
| 63 | 2 | UF | *Salmonella* | 45 | | |
| 64 | 2 | UF | ***Lactococcus*** | 25 | | |
| 65 | 2 | UF | | | | |
| 66 | 2 | UF | | | | |
| 67 | 2 | UF | ***Haemophilus*** | 11 | | |
| 68 | 2 | Re | ***Shewanella*** | 15 | | |
| 69 | 2 | T | | | | |
| 70 | 2 | UF | *Bartonella* | 63 | | |
| 71 | 2 | UF | ***Haemophilus*** | 17 | | |
| 72 | 2 | Ru | | | | |
| 73 | 2 | f,uf | *Rhodobacter* | 22 | | |
| 74 | 2 | L | | | | |
| 75 | 2 | UF | *Neisseria* | 53 | | |
| 76 | 2 | UF | *Xylella* | 44 | | |
| 77 | 2 | UF | | | | |
| 78 | 2 | UF | *Yersinia* | 56 | | |
| 79 | 2 | Re, uf | | | | |
| 80 | 2 | UF | | | | |
| 81 | 2 | UF | | | | |
| 82 | 2 | CWM | | | | |
| 83 | 2 | UF | | | | |
| 84 | 2 | UF | | | | |
| 85 | 2 | UF | ***Burkholderia*** | 14 | | |
| 86 | 2 | UF | | | | |
| 87 | 2 | P | | | | |
| 88 | 2 | UF | ***Escherichia*** | 0 | | Phi29 |
| 89 | 2 | UF | | | V | |
| 90 | 2 | UF | | | | |
| 91 | 2 | UF | | | | |
| 92 | 2 | UF | | | | |
| 93 | 2 | UF | *Chlamydophila* | 56 | | |
| 94 | 2 | UF | | | | |

A function was assigned to a module when ≥50% of the families in the module had been assigned a function. A lower case letter indicates that a small number of protein families in the module were annotated with that function or one of its children's functions in MeGO ontology (http://aclame.ulb.ac.be/Classification/mego.html). A stands for adsorption, Re for replication, Ru for regulation, Rc for recombination, H for head, B for baseplate, TM for long contractile tail, TS for long non-contractile tail, F for fiber, L for lysis, IE for integration/excision, P for pathogenesis, CWM for cell wall modification and UF for unknown (all modules containing ≥50% protein families of unknown function).

The phage type column lists a selection of paradigm phages that have the module.

The "Host" column lists hosts that are significantly associated with modules. Bold characters point to cases where DNA sequence identity among the genomes associated with that host is below 40%.

The "Lifestyle" column lists modules significantly associated with virulent (V) and temperate (T) phages; t was used for associations that are not statistically significant, but where all phages (≥4) having the module are temperate (see text for more details).

The presence of several values in the % DNA identity column indicates corresponding hosts in the "Host" column.

other end of the profile, none of the prophages predicted in Mycobacteria (see http://aclame.ulb.ac.be/perl/Aclame/Prophages/view_prophage.cgi) were retained in the selected prophage set. Overall, not surprisingly, the number of phages sequenced per host did not correlate with the abundance of prophages in those hosts.

### *3.1.2. New protein families*

Within the new categories 'Prophages' and 'Viruses and prophages', ACLAME version 0.4 contains 760 prophages predicted in 199 bacterial genomic sequences (see Materials and methods for details on the procedure used to select the prophage set).

Of the 16,057 families of proteins obtained by the combined all vs. all comparison and clustering of the phage and prophage proteins, 2213 contain both phage and prophage proteins. The contribution of these 2213 families to prophages is shown in Fig. 1B as the number of proteins in these families out of the number of proteins in the prophages. It gives an idea of the "novelty" of the predicted prophages; even though prophages are predicted based on similarity with known phage sequences, we observed that approximately 200 prophages had at least 4 prophages-only families. Prophages outnumbered phages, but there were 9262 phage-only protein families and only 4582 prophage-only protein families, which thus contained potentially novel (phage) protein sequences. Possible explanations for this lower number of prophage-only families are: i) the different size distribution of phage genomes and prophages (phage median size is 39,5 kb; prophage median size is 27,3 kb; $P$-value = $3.2e^{-12}$, Mann–Whitney $U$-test),

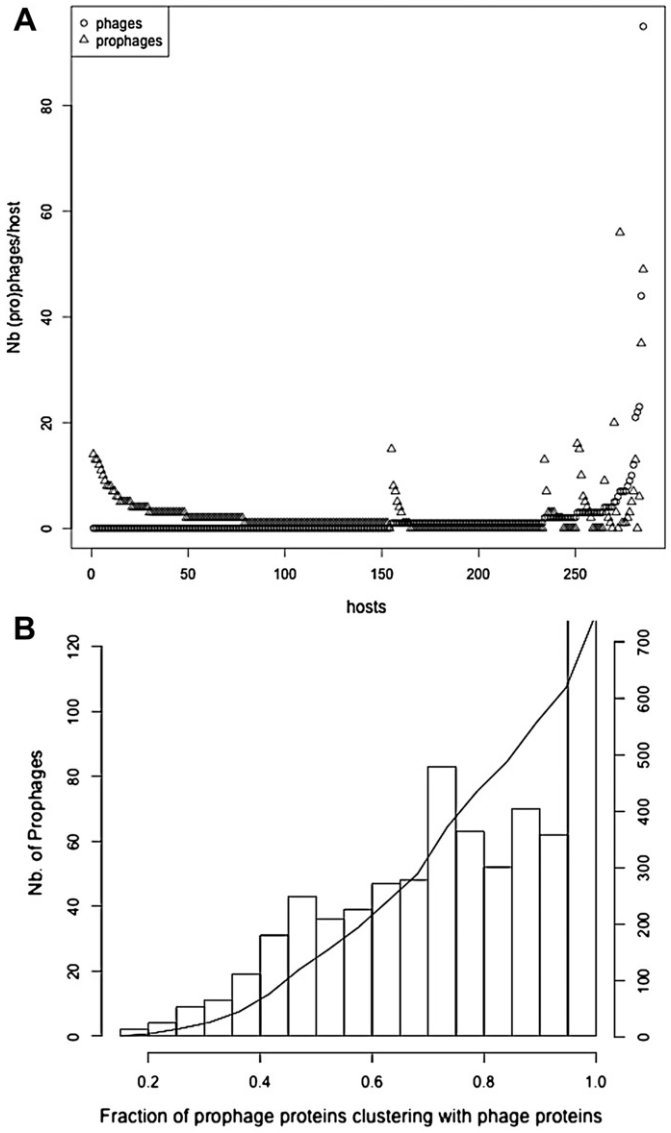


Fig. 1. A: Distribution of phages and prophages among the bacterial genomes.The left side of the plot represents genomes for which there are predicted prophages in spite of the lack of sequenced phages. Genomes for which there are more known phage sequences (right side) have a tendency to have more (predicted) prophages.B: Similarity of prophages to known phages.The fraction of proteins shared with phage proteins relative to the number of proteins encoded by the prophage shows that approximately 200 prophages have at least 40% of novel proteins (sharing, at most, 60% with phages). Likewise, there are 200 prophages that have more than 90% of their encoded proteins classified into a family together with phage proteins.

which may be due to imprecise prophage boundary predictions (ORFans at the boundaries, for instance, will inevitably be missed); ii) the presence of cryptic, and therefore shorter, prophages; and iii) the presence in the phage sequence set of large (non-temperate) phages (e.g. T4). The occurrence of phage proteins in families followed a long-tailed skewed distribution. (Lima-Mendez et al., 2007) (Hatfull et al., 2006); three quarters of the families harbored a single protein (3475 out of 4,582, i.e. 75% for prophage-only families; 6871 out of 9,262, i.e. 74% for phage-only families).

### *3.1.3. Contribution of prophages to characterization of the phage genomic space*

Examining predicted prophages for the presence of genes thus far restricted to virulent phage genomes appeared useful to obtain a preliminary assessment of potentially new information provided by the prophage genome set. In *Desulfovibrio desulfuricans* G20, predicted prophage prophinder:44770 harbored 37 coding sequences (CDS) between the genomic coordinates 939853 and 974604. According to the protein families containing CDS from this prophage (family:vir_proph:389 with T4 gp6 wedge baseplate, family:vir_proph:38 with T4 gp25 wedge baseplate, and family:vir_proph:390 with T4 gp18 tail sheath proteins), a contractile tail similar to that of T4-like phages was encoded. However, this prophage had a CDS in family vir_-proph:41 that contained the late control protein gpD from temperate phage P2 and with no T4 protein. A putative DNA maturation module (protein:proph:172014 and protein:-proph:172015) in that prophage appeared to be related to that of large phage 0305phi8-36 from *Bacillus thuringiensis* only (family:vir_proph:2025, large terminase and family:-vir_proph:1396, portal protein). Two other prophages were similar to prophinder:44770. One (prophinder:44772, coordinates 1914433–1964628) was inserted in an opposite orientation in *D.desulfuricans* as well, while the other (prophinder:44011) was in the other δ-proteobacterium *Syntrophobacter fumaroxidans* MPOB (Fig. 2). These predictions thus represent a possible novel association of phage modules.

### *3.2. (Pro)phage evolutionary cohesive modules*

Protein families were clustered according to the similarity of their phylogenetic profiles (Pellegrini et al., 1999) to produce ECMs (Lima-Mendez et al., 2008b). Using a sig value

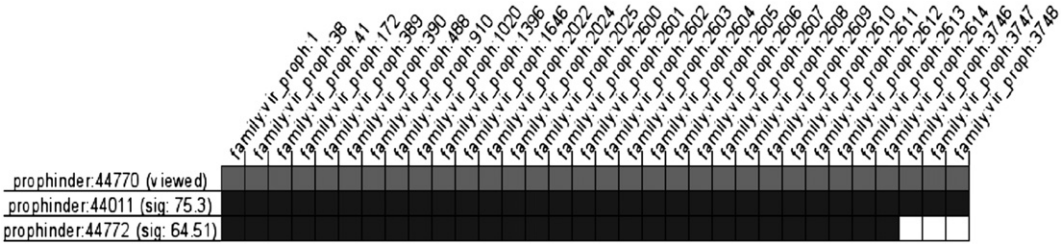


Fig. 2. Three predicted prophages with a new combinations of modules.See text for detailed description. The three predicted prophages can be viewed in detail at http://aclame.ulb.ac.be/perl/Aclame/Prophages/view_prophage.cgi?type=genome&bactGen=NC_007519 for prophinder:44770 and 44772 from *D.desulfuricans* and http://aclame.ulb.ac.be/perl/Aclame/Prophages/view_prophage.cgi?type=genome&bactGen=NC_008554 for prophinder:44011 from *S.fumaroxidans*.

threshold of 10, we found 94 ECMs, ranging in size from 2 to 94 protein families. The family and function compositions of the modules were analyzed manually and, when possible, one or several functions were assigned to the ECMs. A possible association of ECMs with phage lifestyle or with a restricted set of hosts was also investigated.

### 3.2.1. Life-style specific modules

Whether or not a phage is capable of being established and maintained as a latent prophage is of special interest, especially in the context of phage therapy, since only virulent phages are eligible. Presently, there is no automatic way to catalogue phages according to their lifestyle, nor is this information easily retrievable from the literature. Literature search (extension from (Lima-Mendez et al., 2007)) enabled us to classify phages into 110 virulent and 202 temperate phages, while 56 remained unassigned.

We looked for a possible relationship between functional modules and lifestyle by counting the number of phages with a given module and a given lifestyle, and assessing whether this number was statistically significant. Modules restricted to virulent or temperate phages are marked as such in Table 1.

Twenty-three modules were associated with temperate phages only. Thirteen had significant statistical association and 10 were only found in temperate phages, but with lower statistical significance. ECM17 represents the strongest association with temperate phages. It brings together integrase and repressor protein families, robust markers of temperate phages integrating their genome by site-specific recombination, as well as families of proteins involved in genome replication. Interestingly, from the 101 phages represented by this ECM, we defined 8 as virulent and 3 remained unassigned. Inspection of the 8 corresponding genomes showed that several have features suggesting a temperate ancestor (Mikkonen et al., 1996) (Tuohimaa et al., 2006) (Labrie and Moineau, 2002). Another important limitation of ECM17 for predictive purposes is that it is more representative of (pro)phage from Gram(+) bacteria (over 80% of the genomes in ECM17).

ECM43 and 47 include families of transposase and transposase-activator proteins. They are associated with transposable phages, all of which are temperate. This (pro) phage group is further described below.

### 3.2.2. Phage-type specific modules

Using the functional MeGO/GO annotation of the protein families (Toussaint et al., 2007) composing the modules, these were assigned to one or several functional categories (adsorption, replication, regulation, recombination, head, baseplate, contractile tail, long non-contractile tail, tail fiber, lysis, integration, excision, pathogenesis) when they contained over 50% of protein families with an assigned function (see Table 1).

Virulent phage genomes hardly break down into functional modules. One illustrative example is module ECM1 composed of 94 families having all T4 and T4-related proteins, most of which constitute what has been called the core and quasi-core of T4-like viral genomes (Petrov et al., 2010). The functions of these families encompass nearly the whole lytic cycle (DNA replication, recombination and repair, auxiliary metabolism, gene expression and morphogenesis). Part of this core genome does, however, reside in another marker module of virulent phages, ECM26. This module also harbors protein families involved in DNA metabolism such as deoxyribunucleases, helicases and ribonucleotide reductases, but contrary to ECM1, it is present not only in most T4-like phages, but also in other phages and prophages, most of which infect Gram + bacterial hosts.

The cohesion of T7-like phages is also illustrated by the composition of their modules. Only two modules, ECM 14 and ECM 22, together make up most of the genomes of these phages (T7 T3, K1F, Berlin, phiA1122). Several hypotheses have been proposed to account for this tendency of virulent phages to be less prone to reorganization compared to temperate phages. The absence of a prophage state may prevent transition through inactive combinations of genes awaiting the occurrence of new functional (pro)phages (Hendrix, 2003). In the case of T4, functional constraints imposed by the mode of replication have been suggested to prevent major reorganization within the core genome, leaving the variable part of the genome most prone to recombination (see e.g. (Petrov et al., 2010) and references therein). As a result of this cohesion, ECM1 and, with one exception, ECM26, are conserved at a higher taxonomic range within the ICTV classification, in the genus T4-like phages. Similarly, ECM14 is conserved in the subfamily Autographivirinae (T7 and relatives, see suppl. Table 2S, ECMs in ICTV families). However, in both cases, the second conserved module (ECM26 and ECM22) is either shared with phages in other families or absent from some phages in the subfamily.

In most other (pro)phages, replication, head and tail morphogenesis and lysis genes are found in separate modules, each of which represents different subsets of genomes (suppl. Tables 1S and 2S). Except for ECM 6 and 23, which are present in all P22-like viruses available in our analysis, no other module behaves as marker of ICTV subfamilies. However, there are nearly perfect overlaps that deserve further attention. ECM4 is present in 14 out of 15 Peduovirinae viruses (family Myoviridae; it is absent from Burkholderia phage KL3). ECM45 is present in all 6 Myoviridae SPO1-like viruses and in one unclassified Myoviridae; Thermus phage phiYS40. ECM89 is present in 8 of the 10 Picovirinae in the analyzed set (it is absent from *Bacillus* phage Φ29 and *Staphylococcus* phage SAP-2). ECM1, discussed above as a marker of T4-like phages, is actually present in two unclassified Myoviridae phages, *Rhodothermus* phage RM378 and Enterobacteria phage Felix 01. The relationship of these unclassified phages to existing ICTV subfamilies could thus contribute to their classification. Another interesting example is ECM39, which encompasses two Podoviridae subfamilies, the Epsilon15-like and Bpp-1-like viruses. The absence of defined subfamilies within the Siphoviridae families prevents more complete assessment of ECM conservation across ICTV taxonomy levels.

The few paradigm phage types listed in Table 1 highlight long-known reticulate relationships such as those between λ and P22 with similar heads and different tails, or HK97 and λ with different heads and similar tails (Juhala et al., 2000). This is further illustrated in the next section by a description of the relevance of the modules for helping to capture phage genome mosaicism in the particular case of transposable/mutator (pro)phages.

### 3.2.3. Host-specific modules

Host-specific modules appeared of interest because they could include proteins involved in phage-host interactions. Out of the 94 modules, 57 were significantly associated with a restricted set of hosts (see Table 1). The host specificity observed (considered here at the genus taxonomical level) is not due to redundancy, since the average genome sequence identity among most genomes sharing such a module is below 40%. When not unknown, the functions assigned to the families represented by these modules relate to processes such as replication, lysis, pathogenesis, modulation of host functions, etc., all of which involve close interaction between phage and host proteins. Host-specific modules with unknown function appear to be good targets for further experimental characterization, since they point to new proteins of potential medical relevance, as in phage therapy and, more generally, in phage–host interaction/recognition.

## 3.3. Using phage-specific modules to analyze specific phage groups

As mentioned above, ECM43 and 47 appeared specific to transposable/mutator phages. ACLAME version 0.4 contains 7 completely sequenced mutator phage genomes. Four are *Myoviridae* (B3, BcepMu, Mu and PhiE255) and 3 are *Siphoviridae* (D3112, DMS3, MP22). The 7 phages have one among three genetic organizations (Mu-type for Mu, D3112, DMS3 and MP22; B3-type for B3 and BcepMu; the third type was seen so far only in PhiE255, see Fig. 3).

### 3.3.1. From modules to protein families to predicted prophages

Mutator phages stand apart from other phages by their mode of replication and packaging. Replicative transposition spreads viral genome copies in the host genome and these are packaged with short adjacent host DNA sequences (for reviews see (Symonds et al., 1987)). ECM9 is the largest module shared by all seven mutator phages. It is associated with either of the two 'transpositional replication' modules ECM43 (B3, BcepMu and PhiE255) and 47 (the 4 Mu-like phages). Together, these three ECMs are specific to the 7 mutator phages.

The largest families in ECM9 include orthologs of Mu gpF, G and the portal gpH involved in head assembly, gp36, a probable head protein and the GemA semi-essential protein. The largest protein family in ECM9 is family:vir_proph:112 with 49 proteins related to Mu gpG. A set of predicted mutator prophages was derived from the prophages listed in that family (see Fig. 4). Two prophages have two proteins in the family and actually consist of 2 identical prophages that were included in a single prophage prediction (prophinder:44692 and prophinder:47441).

ECM43 and 47 are the expected hallmarks for mutator phages with families of transposases (gpA in phage Mu in family:vir_proph:353 and family:vir_proph:662) and transposase activator proteins (gpB in phage Mu, in family:-vir_proph:355 and family:vir_proph:661 and/or 901)–thus a 'replication-lysogeny' function. The total number of either transposase or transposase activator proteins in these families does not add up to 49, pointing towards either incomplete prophage predictions or prophages missing the replication-lysogeny module. Manual inspection of the bacterial genomes on both sides of the predictions that had no transposase protein showed that at least 10 predictions were incomplete. They missed *A* and *B* genes, which are present and would have fit into ECM43 or 47, and often a portion of the semi-essential region, including *gemA*. In total, once completed, 30 predicted prophages (Fig. 4) had all genes required to be functional. One previously predicted Mu-like prophage (FluMu, (Cunningham et al., 2002) prophinder:46299) was not present in the selected set of 760 prophages, showing that our selection procedure discarded some significant predictions. Two other mutator prophages, Pnm1 in *Neisseria meningitidis* Z2491 ((Klee et al., 2000), here prophinder:44621) and MuMenB in *Neisseria meningitides* B58 ((Masignani et al., 2001), here prophinder:44629) were previously described, but not tested experimentally. Both were imperfectly predicted. Prophinder:

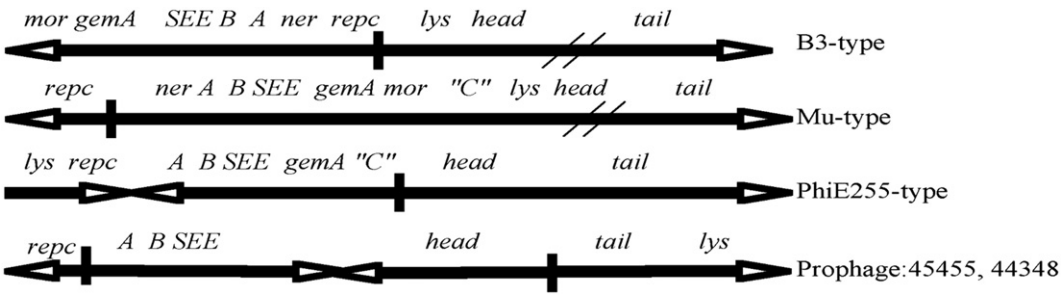


Fig. 3. Schematic view of the general genomic organization of different mutator phages and prophages.B3-type genomes, left end, which include early transposase *A*, positive regulator of transposition *B* and lytic cycle negative regulator *ner* genes, are in opposite orientation vs. late lysis, head and tail genes. On Mu-type genomes, only repressor gene *repc* is in such opposite orientation. PhiE255 and the predicted prophages in Campylobacteriacae (prophage:45455 from *A. butzleri* RM4018 and prophage:44348 from *C. jejuni* RM1221) have still another genome organization, but maintain the same set of functions, including a region that resembles the Mu "semi-essential" region (SEE). Genome organization is not related to morphotype (see text for more details).

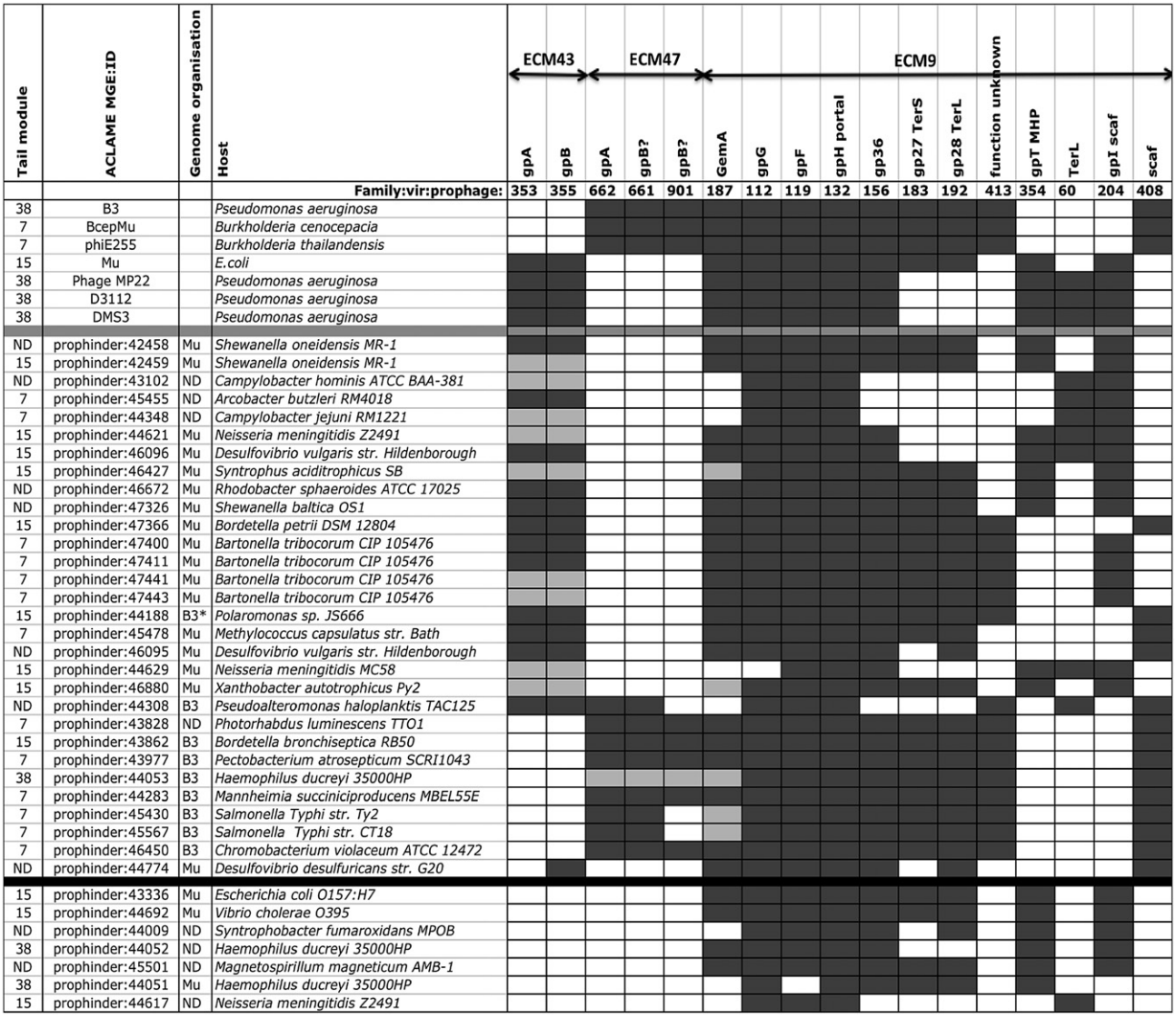


Fig. 4. A set of 37 predicted mutator prophages The 7 sequenced mutator genomes are shown above 37 predicted related prophages. Predicted prophages below the black line are probably cryptic. Black and white boxes mark the presence and absence of proteins. Lines are phage and predicted prophage genomes. Gray boxes mark proteins that could be seen by manual inspection of the host genome sequences of incomplete prophage predictions. Column A: tail modules; ECM7 and 15 contain sheath and tube proteins and hence correspond to contractile tails; ECM38 is a non-contractile tail module; column B: phage and prophages; column C: host; column D–T: ACLAME family IDs and function. As detailed in the text, the family:vir:prophage:60 of TerL proteins is not part of ECM9. Functions: family:vir:prophage:204 and family:vir:prophage:408: scaffolding protein; family:vir:prophage:354 and possibly family:vir:prophage:413: major head protein; family:vir:prophage:353 and family:vir:prophage 662: transposase gpA protein; family:vir:prophage:355 and family:vir:prophage 661/901: transposition activator protein gpB. All the protein families, phage and prohage genomes can be browsed in ACLAME version 0.4.Note that while the four prophages in *B.tribocorum* CIP 105476 are almost identical at the nucleotide level (data not shown), the two prophages in the same strain of *D.vulgaris* display alternative TerL, MHP and scaffolding proteins and also differ in their contractile tail.

44348 corresponds to the previously described Mu-like prophage CJIE1_CMLP1 (Fouts et al., 2005), which has several more or less close relatives in many *Campylobacter jejuni* strains (Parker et al., 2006). Some of these prophages generate chromosomal rearrangements (Scott et al., 2007) and hence are functional. Other predictions correspond to previously recognized putative mutator prophages (e.g. MuMc02, here prophinder:45478) that were not necessarily correctly predicted either. Four predictions in *Bartonella tribocorum* CIP 105476 (prophage:47400, prophage:47411, prophage:47441 and prophage:47443) are copies of the same prophage spread over the host chromosome, supporting the prophage functionality. Prophinder:44692 in *Vibrio cholerae* is an incomplete prediction, which brought together two apparently complete and identical prophages. They have a Mu-like structure, are in inverted orientations and flank a chromosomal segment, likely resulting from an inversion in the host chromosome mediated by a replicative transposition of one prophage copy. That prophage thus also appears functional.

Three incomplete predicted prophages (prophage:45455 in *Arcobacter butzleri* RM4018, prophage:44348/CJIE1/CMLP1 in *C. jejuni* RM1221 and prophage:43102 in *Campylobacter*

*hominis* ATCC BAA-381, have a genetic organization that slightly differs from that in either Mu, B3 or PhiE255 (see Fig. 3). All three predicted prophages, when completed, have all the genes needed to replicate, make virions and lyse.

Overall, the present set of mutator (pro)phage-host range thus extends from β and γ to α,δ and ϵ branches of the proteobacteria.

*3.3.2. From (pro)phage genomes to modular structure*

ECM9 also contains families of terminase subunits TerS and TerL (family:vir_proph:183 and 192), major head proteins (family:vir_proph:354) and two mutually exclusive families of scaffolding proteins (family:vir_proph:204 and family:-vir_proph:408). Some mutator (pro)phages, however, have their TerL protein in family:vir_proph:60, which is part of another head module (ECM8) found in a large variety of phages infecting either Gram + or Gram- hosts.

The heads of mutator (pro)phages are thus mosaics. They are combined not only with alternative sets of transposition genes in ECM43 and 47, but also with three 'tail modules', two of the '*Myoviridae*' contractile type and one of the '*Siphoviridae*' non-contractile type. Relatives of the contractile tails are found in a large number of non-mutator phages including P2 (ECM7 as for BcepMu and PhiE255) and P27 (ECM15 as for Mu) respectively. ECM38 corresponds to a non-contractile tail class and is found in B3, DMS3, D3112 and MP22, a few prophages (see Fig. 4) and only a small number of non-mutator phages (*P.aeruginosa* phages M6 and P73 and *B.cepacia* BcepNazgul).

*3.3.3. From modules to functional predictions*

The B3 major head protein has not been characterized experimentally (M.M. Howe pers. comm.). It could belong to family:vir_proph:413, since it is the only conserved protein family with unknown function in ECM9 that is exclusive to family:vir_proph:354, which contains the Mu gpT-related major head proteins (Fig. 4)

The GemA protein is dispensable for Mu growth in the laboratory. Its function is unknown. Some GemA mutants synchronize the division of infected cells and relax DNA supercoiling, although there is no direct interaction between GemA and the host gyrase (Abbes et al., 2001). The promiscuity and conservation of GemA among 37 (pro)phages analyzed strongly suggests that *gemA* does not belong to the poorly conserved SEE region, as considered thus far. Since it is an early-transcribed gene, it is unlikely to participate in head assembly like the other proteins in ECM9. Rather, its position on the genomes next to the activator of late (or in some case, middle) transcription points towards a role in a switch between replicative transposition and expression of the capsid proteins.

## 4. Conclusions

Although the Prophinder prophage prediction is based on a comparison with existing phage sequences, predicted prophages contribute new protein families and a better global view of the relationships among and between some (pro)phage genome types. Moreover, a significant number of ECMs extracted from the combined set of phage and prophage protein families could be assigned a function (e.g. head or tail assembly, genome replication, recombination) based on the annotation of the families belonging to the modules. As proposed by Lawrence et al. (Lawrence et al., 2002) most (pro)phage genomes can be described as assemblies of several "modi"/ECMs, i.e. sets of genes, not necessarily adjacent on the genomes, that tend to remain associated despite the recombinant nature of phage genomes.

ECMs should contribute experimentally testable hypotheses. ECMs with a defined function may provide hints as to a possible function for proteins of unknown function within that ECM. ECMs with the same function mutually excluding each other in a set of related phage genomes also provide hints on the function of uncharacterized phage proteins (as suggested above for the B3-type major head protein).

Some ECMs appear specific to certain phage types (as detailed above for some virus taxonomy subfamilies and mutator phages), to some host groups (*Streptococcus* and *Staphylococcus* among others) and, in the case of ECM17, to lifestyle. However, the latter module is a composite of two functional (sub)modules, the integration and regulatory switch, which are the true markers of integrative phages, and a replication module found in phages infecting Gram(+) bacteria. In the future, refining methodology may lead to deriving modules that are truly specific for lysogeny and could be used for automatic recognition of temperate phages.

Means of measuring relationships in the phage world remain under debate. The present mosaic representation is not meant to replace classical taxonomy such as that maintained by ICTV, especially for phages like T4 and T7 and their relatives, which have most of their conserved genes in a single or very low number of modules. However, modular representation should provide a more rational description of genomes where features such as sequence similarity, mode of replication and packaging are not congruent with morphotypes, as illustrated above for transposable phages. In addition, statistical values provided by our method could contribute to establishing new phage groups.

Our analysis of predicted mutator prophages brings out specific features (conservation of the *gemA-mor* or $rep_{comp}$-*ner-A-B* gene order at one end of B3-like and Mu-like prophages respectively, high conservation of the portal and other head proteins) that may be useful for improving the performance of prophage prediction methods. More generally, combining prediction of prophages in newly sequenced bacterial genomes with a similarity search in phage protein families and ECM analysis should greatly improve the quality of prophage gene annotation and may, over a longer range, be adapted to identification of (pro)phage genes in metagenomic sequences.

## Acknowledgments

Work on the ACLAME project in the BiGRe laboratory was financed by the European Space Agency (ESA-PRODEX) and the Belgian Science Policy (Belspo) through the EXANAM project (PRODEX agreements No. C90358), the

Fonds de la Recherche Scientifique Médicale (FRSM), the Belgian Program on Interuniversity Attraction Poles, initiated by the Belgian Federal Science Policy Office, project P6/25 (BioMaGNet) and the BioSapiens Network of Excellence funded under the Sixth Framework Program of the European Communities (LSHG-CT-2003-503265).

## Appendix. Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.resmic.2011.06.006

## References

Abbes, C., Joseleau-Petit, D., Liebart, J.C., D'Ari, R., Sezonov, G., 2001. The GemA protein of phage Mu and the GyrB gyrase subunit of Escherichia coli: the search for targets and interactions leading to the reversion of Mu-induced mutations. Biochimie 83, 261–267.

Abdi, H., 2007. Bonferroni and Sidak corrections for multiple comparisons. In: Salkind, N.J. (Ed.), Encyclopedia of Measurement and Statistics. Sage, Thousand Oaks (CA), pp. 103–107.

Botstein, D., 1980. A theory of modular evolution for bacteriophages. Ann N Y Acad Sci., 484–490.

Cunningham, J.T., Bruno, S.B., Higgs, K.A., Sullivan, M.J., 2002. Intrapericardial procaine affects volume expansion-induced fos immunoreactivity in unanesthetized rats. Exp. Neurol. 174, 181–192.

Fouts, D.E., Mongodin, E.F., Mandrell, R.E., Miller, W.G., Rasko, D.A., Ravel, J., Brinkac, L.M., DeBoy, R.T., et al., 2005. Major structural differences and novel potential virulence mechanisms from the genomes of multiple campylobacter species. PLoS Biol. 3, e15.

Hatfull, G.F., Pedulla, M.L., Jacobs-Sera, D., Cichon, P.M., Foley, A., Ford, M.E., Gonda, R.M., Houtz, J.M., et al., 2006. Exploring the mycobacteriophage metaproteome: phage genomics as an educational platform. PLoS Genet. 2, e92.

Hendrix, R.W., 2003. Bacteriophage genomics. Curr. Opin. Microbiol. 6, 506–511.

Hendrix, R.W., Smith, M.C., Burns, R.N., Ford, M.E., Hatfull, G.F., 1999. Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. Proc Natl Acad Sci USA, 2192–2197.

Juhala, R.J., Ford, M.E., Duda, R.L., Youlton, A., Hatfull, G.F., Hendrix, R.W., 2000. Genomic sequences of bacteriophages HK97 and HK022: pervasive genetic mosaicism in the lambdoid bacteriophages. J. Mol. Biol. 299, 27–51.

Klee, S.R., Nassif, X., Kusecek, B., Merker, P., Beretti, J.L., Achtman, M., Tinsley, C.R., 2000. Molecular and biological analysis of eight genetic islands that distinguish Neisseria meningitidis from the closely related pathogen Neisseria gonorrhoeae. Infect. Immun. 68, 2082–2095.

Labrie, S., Moineau, S., 2002. Complete genomic sequence of bacteriophage ul36: demonstration of phage heterogeneity within the P335 quasi-species of lactococcal phages. Virology 296, 308–320.

Lavigne, R., Darius, P., Summer, E.J., Seto, D., Mahadevan, P., Nilsson, A.S., Ackermann, H.W., Kropinski, A.M., 2009. Classification of Myoviridae bacteriophages using protein sequence similarity. BMC Microbiol. 9, 224.

Lavigne, R., Seto, D., Mahadevan, P., Ackermann, H.-W., Kropinski, A.M., 2008. Unifying classical and molecular taxonomic classification: analysis of the Podoviridae using BLASTP-based tools. Res. Microbiol, 406–414.

Lawrence, J., Hatfull, G., Hendrix, R., 2002. Imbroglios of viral taxonomy: genetic exchange and failings of phenetic approaches. J. Bacteriol., 4891–4905.

Leplae, R., Lima-Mendez, G., Toussaint, A., 2010. ACLAME: a CLAssification of Mobile genetic Elements, update 2010. Nucleic Acids Res. 38, D57–D61.

Lima-Mendez, G., Toussaint, A., Leplae, R., 2007. Analysis of the phage sequence space: the benefit of structured information. In: Virology, pp. 241–249.

Lima-Mendez, G., van Helden, J., Toussaint, A., Leplae, R., 2008a. Prophinder: a computational tool for prophage prediction in prokaryotic genomes. Bioinformatics, 863–865.

Lima-Mendez, G., van Helden, J., Toussaint, A., Leplae, R., 2008b. Reticulate representation of evolutionary and functional relationships between phage genomes. Mol. Biol. Evol, 762–777.

Masignani, V., Giuliani, M.M., Tettelin, H., Comanducci, M., Rappuoli, R., Scarlato, V., 2001. Mu-like Prophage in serogroup B Neisseria meningitidis coding for surface-exposed antigens. Infect. Immun. 69, 2580–2588.

Mikkonen, M., Dupont, L., Alatossava, T., Ritzenthaler, P., 1996. Defective site-specific integration elements are present in the genome of virulent bacteriophage LL-H of Lactobacillus delbrueckii. Appl. Environ. Microbiol. 62, 1847–1851.

Nelson, D., 2004. Phage taxonomy: we agree to disagree. J. Bacteriol, 7029–7031.

Parker, C.T., Quinones, B., Miller, W.G., Horn, S.T., Mandrell, R.E., 2006. Comparative genomic analysis of Campylobacter jejuni strains reveals diversity due to genomic elements similar to those present in C. jejuni strain RM1221. J. Clin. Microbiol. 44, 4125–4135.

Pellegrini, M., Marcotte, E., Thompson, M., Eisenberg, D., Yeates, T., 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc. Natl. Acad. Sci. USA, 4285–4288.

Petrov, V.M., Ratnayaka, S., Nolan, J.M., Miller, E.S., Karam, J.D., 2010. Genomes of the T4-related bacteriophages as windows on microbial genome evolution. Virol. J. 7, 292.

Rohwer, F., Edwards, R., 2002. The Phage Proteomic Tree: a genome-based taxonomy for phage. J. Bacteriol, 4529–4535.

Scott, A.E., Timms, A.R., Connerton, P.L., Loc Carrillo, C., Adzfa Radzum, K., Connerton, I.F., 2007. Genome dynamics of Campylobacter jejuni in response to bacteriophage predation. PLoS Pathog. 3, e119.

Symonds, N., Toussaint, A., van de Putte, P., Howe, M.M., 1987. Phage Mu New York.

Toussaint, A., Lima-Mendez, G., Leplae, R., 2007. PhiGO, a phage ontology associated with the ACLAME database. Res. Microbiol. 158, 567–571.

Tuohimaa, A., Riipinen, K.A., Brandt, K., Alatossava, T., 2006. The genome of the virulent phage Lc-Nu of probiotic Lactobacillus rhamnosus, and comparative genomics with Lactobacillus casei phages. Arch. Virol. 151, 947–965.