environmental microbiology
environmental microbiology reports

Genomic and proteomic analysis of phage E3 infecting the soil-borne actinomycete Rhodococcus equi

Journal:	*Environmental Microbiology and Environmental Microbiology Reports*
Manuscript ID:	Draft
Manuscript Type:	EMIR - Brief report
Journal:	Environmental Microbiology Reports
Date Submitted by the Author:	n/a
Complete List of Authors:	foley, sophie; Edinburgh Napier University, School of Life. Sport & Social Sciences Vazquez-Boland, Jose; University of Edinburgh,
Keywords:	Rhodococcus, micolic acid containing Actinobacteria, bacteriophages, genomics/functional genomics/comparative genomics, soil-borne actinomycetes, pathogen ecology, environmental biocontrol

Resubmission EMI-2012-0269

Genomic and proteomic analysis of phage E3 infecting the soil-borne actinomycete *Rhodococcus equi*

Samson P. Salifu [1‡], Ana Valero-Rello [2‡], Samantha A. Campbell [1], Neil F. Inglis [3], Mariela Scortti [2], Sophie Foley[1*], and José A. Vazquez-Boland [2,4*]

[1] *School of Life, Sport and Social Sciences, Edinburgh Napier University, Edinburgh EH11 4BN, UK*
[2] *Microbial Pathogenesis Unit, Centres for Infectious Diseases and Immunity, Infection & Evolution, University of Edinburgh, Edinburgh EH9 3JT, UK*
[3] *Moredun Proteomics Facility, Moredun Research Institute, Pentlands Science Park, Bush Loan, Penicuik, IEH26 0PZ, UK*
[4] *Grupo de Patogenómica Bacteriana, Facultad de Veterinaria, Universidad de León, 24701 Leon, Spain.*

Running Title: *Rhodococcus* bacteriophage E3

‡ Contributed equally to this work.

*For correspondence. Email s.foley@napier.ac.uk, Tel. +44 (0)131 455 2626; Email v.boland@ed.ac.uk, Tel. +44 (0)131 651 3619.

Summary

We report on the characterisation and genomic analysis of bacteriophage E3 isolated from soil and propagating in *Rhodococcus equi* strains. Phage E3 has a circular genome of 142,563 bp and is the first *Myoviridae* reported for the genus *Rhodococcus* and for a non-mycobacterial mycolic acid-containing actinomycete. Phylogenetic analyses placed E3 in a distinct *Myoviridae* clade together with *Mycobacterium* phages Bxz1 and Myrna. The highly syntenic genomes of this myoviridal group comprise vertically evolving core phage modules flanked by hyperplastic regions specific to each phage and rich in horizontally acquired DNA. The hyperplastic regions contain numerous tRNA genes in the mycobacteriophages which are absent in E3, possibly reflecting bacterial host-specific translation-related phage fitness constraints associated with rate-limiting tRNAs. A structural proteome analysis identified 28 E3 polypeptides, including 15 not previously known to be virion-associated proteins. The E3 genome and comparative analysis provide insight into short-term genome evolution and adaptive plasticity in tailed phages from the environmental microbiome.

Introduction

The genus *Rhodococcus* is a group of ubiquitous *Actinobacteria* with more than 40 species widely distributed in the environment. The rhodococci are mycolata actinomycetes, characterised by a lipid-rich cell envelope containing branched-chain mycolic acids and conferring protection to environmental agressions. *Rhodococcus* spp. are environmentally and biotechnologically important due to their extraordinary metabolic versatility and biodegradative properties (Larkin *et al*., 2005). The genus also contains an animal pathogen, *Rhodococcus equi*, a soil-dwelling organism that can cause pyogranulomatous infections in different species. Young foals are especially susceptible and develop severe purulent pneumonia associated with a high mortality (Prescott, 1991; Muscatello *et al*., 2007; Vázquez-Boland *et al*., 2010). In humans, it is an emerging opportunistic pathogen causing life-threatening infections reminiscent to pulmonary tuberculosis (Weinstock and Brown, 2002). *R. equi* propagates in soil rich in herbivore manure and is common in the farm environment worldwide. There is no effective vaccine available and the prophylactic administration of long antibiotic courses is the current strategy to limit the occurrence of foal rhodococcosis in endemic farms (Dawson *et al.,* 2010; Giguere *et al*., 2011). However, *R. equi* is intrinsically refractory to many antimicrobials (Letek *et al*., 2010) and there is risk of emergence and dissemination of acquired resistance to the currently used drugs (rifampin and macrolides/azalides) (Giguere *et al*., 2011).

Due to their lytic properties and host specificity, bacteriophages offer an alternative tool against bacterial pathogens and could be used to contain *R. equi* populations in the farm environment. Preliminary experiments conducted by Summer *et al*. (2011) using inoculated soil samples demonstrate the potential for phages in the biocontrol of *R. equi*. Prior to their exploitation in this way, the phages require to be extensively characterised. Furthermore, considering the contribution of phages to bacterial genome evolution and acquisition of nicheadaptive traits (Brussow *et al*., 2004), characterisation of phages may complement and enhance our basic knowledge of the host organism.

Of the mycolata group of *Actinobacteria*, which includes the genera *Corynebacterium, Dietzia, Gordonia, Mycobacterium, Nocardia, Rhodococcus* and *Tsukamurella* amongst others, only the phages infecting *Mycobacterium* spp. have received significant attention to date. There is a paucity of genome sequences available for phages infecting the genus *Rhodococcus,* with only four recently characterised *R. equi* phages, all belonging to the *Siphoviridae* (Summer *et al.,* 2011). This study reports on the extensive genomic and proteomic analysis of *R. equi* phage E3, isolated from soil. The E3 genome sequence is the first to be described for a *Myoviridae* infecting the environmentally ubiquitous genus *Rhodococcus*.

Results and Discussion

Phage isolation and preliminary characterisation

R. equi-infecting phages were isolated from topsoil samples using *R. equi* NCIMB 10027 as propagating host. Phages could be detected directly (i.e. without enrichment) by spotting a soil aqueous extract on a lawn of *R. equi* bacteria. Of nine soil samples tested, seven yielded phage titres ranging from 1.2×10^3 to 6.7×10^5 pfu g^{-1} of soil. Phage E3 was selected for further analysis on the basis of its broad host range. Using a global collection of *R. equi* isolates (Ocampo-Sosa *et al*., 2007), E3 was capable of infecting a wide variety of strains from different sources (environmental, clinical including equine, porcine, bovine and human isolates) and geographical origins (data not shown). No plaques were observed on non-*equi Rhodococcus* spp. (*R. erythropolis, R. rhodochrous*, *R. ruber*, *R. opacus, R. fascians*) and other related bacteria such as *Gordonia* spp. or *Mycobacterium* spp. Electron microscopy revealed a member of the *Myoviridae* family in the order *Caudovirales* (Fig. 1).

General genome features, organisation and comparative analysis

The phage E3 genome consists of 142,563 bp of double stranded (ds) DNA with an average GC content of 67.65%, similar to that of the host species (68.76%; Letek *et al.*, 2010). Manual sequence gap joining by PCR yielded a circular genome, also supported by restriction analysis of E3 DNA and the failure to identify the presence of cohesive (*cos*)-ends. E3 has a tightly packed genome with 221 ORFs covering 92.9% of the sequence (coding density 1.59 genes per Kb, average gene length 650 bp) (see Table S1 for complete genome annotation). The genome is transcribed in a single direction with the exception of four ORFs, three of which span a discrete 2.5 kb region that includes a putative helicase gene (locus *E3_1340-60*) (Fig. 2). No tRNA or transfer-messenger tRNA genes could be identified in the E3 genome sequence.

BLASTp homology searches showed E3 genome products to be most similar to proteins from mycobacterial *Myoviridae* of the Bxz1-like group (Bxz1 plus six nearly identical phages: Catera, Cali, ET08, LRRHood, Rizal, ScottMcG) and Myrna, all of which also have circular genomes (Hatfull *et al.*, 2010). Pairwise genome alignments showed that E3 and the mycobacterial Bxz1 and Myrna phages are closely related. The highly syntenic genomes share a similar modular arrangement, with identically located conserved housekeeping regions and four interspersed sections of highly divergent DNA or "hyperplastic regions" (HPR 1 to 4) (Fig. 2). A high degree of conservation is observed not only for the morphogenesis modules, generally similarly configured across ds-DNA tailed phages, but also the DNA replication/recombination module, which in the *Myoviridae* tends to appear in different locations disseminated along the genome. Considering the generally extensive structural genetic divergence and mosaicism among phage genomes (Pedulla *et al.*, 2003; Casjens and Thuman-Commike, 2011), these observations suggest that E3, Bxz1 and Myrna have recently diverged from a common ancestor.

Phylogenetic analysis

Phylogenetic trees were constructed based on the terminase large subunit (TerL), prohead protease and DNA polymerase proteins from E3 (*E3_0050*/gp5, *E3_0770*/gp77 and *E3_1540*/gp154, respectively) and representative *Caudovirales*, with the reference phage for each accepted genus and a selection of phages infecting *Actinobacteria*, including those recently described for *R. equi* (Figs. 3 and S1). The genes encoding these three proteins are within the 20 most widely distributed orthologues in phage genomes (Liu *et al.*, 2006) and have been previously used in phage phylogenetic studies (Monier, 2008; Hatfull *et al.*, 2010). E3 grouped in all cases with the *Mycobacterium smegmatis* Bxz1-like and Myrna phages in a robust monophyletic cluster, at short distance with respect to their most recent common ancestor. A new myovirus genus has recently been proposed for the mycobacterial Bxz1 and Myrna (Lavigne *et al.*, 2009) and our data support this proposal and the inclusion of E3 within this group. Indeed, the global nucleotide similarity between the three phages, 50.1 to 50.6%, is consistent with the typical values for phages belonging to a same genus (39.6 to 69.4%, median 50.9%). Moreover, members of a specific phage genus tend to infect phylogenetically related bacteria (Glazko *et al.*, 2007), as is the case for the E3/Bxz1/Myrna cluster (hosts are all mycolata within suborder *Corynebacterineae* of the *Actinomycetales*).

Proteomic analysis

A detailed proteomic characterisation of virion particles by SDS-PAGE and liquid chromatography-electrospray ionization-tandem mass spectrometry (LC-ESI-MS/MS) identified 28 E3-encoded products (Table S2). These included 15 polypeptides initially annotated as hypothetical/uncharacterised proteins, for which we can now establish they are virion-associated proteins. Of the 28 proteins identified, 24 were encoded in one discrete region of the genome encompassing 54.4 kb, which corresponds to the conserved morphogenesis module (Fig. 2).

Core modules

Morphogenesis. This module is highly conserved in the E3/Bxz1/Myrna myoviruses and is interrupted by a horizontally acquired HPR (HPR-2, see below), which in E3 encodes a number of structural proteins of unknown function and two tail fibre proteins. It begins with a head assembly unit. Based on secondary structure similarities and synteny with the well-characterised enterobacteriophage HK97 (Juhala *et al.*, 2000), we identified gp72, gp77 and gp79 as the putative portal, prohead protease and major capsid proteins, respectively. Except for several gene insertions/deletions, the synteny is perfectly maintained with Bxz1 and Myrna (Fig. S2).

The tail morphogenesis unit lies immediately downstream and encompass *E3_1100* to *E3_1160* encoding a minor tail protein (gp110), as suggested by synteny and similarity in the structural fold with the tail terminating protein (TrP) gpU of phage λ; a tail sheath protein (gp111); a tail tube protein (gp112); and several hypothetical proteins (gp113 to 116). In many tailed phages, folding of the tail proteins is mediated by a chaperonin produced by a programmed translational frameshift of two overlapping ORFs, *G* and *T* genes in the case of bacteriophage λ (Xu *et al.*, 2004). These ORFs, typically located downstream of the major tail ORFs, share an overlapping region containing a slippery sequence. In phage E3, ORFs *E3_1140* and *E3_1150* encode a potential “G/T – like” fusion protein (gp114/115) with ribosomal slippage at 5’ GGGAAAA 3’ near the 3’ end of *E3_1140* conserving the protein sequence. This heptanucleotide sequence is also found in gp114/115 homologues amongst Bxz1 (gp127/128) and Myrna (gp126/127) mycobacteriophages. Downstream of the genes expressed via a programmed translational frameshift usually lies a phage tail tape measure protein (TMP) gene (Xu *et al.*, 2004). Although the annotation of Bxz1 and Myrna locates the TMP gene within the head morphogenesis unit, our analyses indeed predict TMP to be encoded by *E3_1160* downstream from the putative fusion protein *E3_1140/1150.* This is supported by the gene size (2,550 bp) and other typical features of TMPs, such as a high alanine-glycine content, absence of cysteine residues, an N-terminus containing α-helices immediately followed by a region of random coils, and similarity to a conserved core region of the TMP of the *Siphoviridae* phage TP901 family (Hatfull, 2006, Pedersen *et al*., 2000). The tail morphogenesis unit ends with a baseplate assembly region *E3_1190* to *_1250*, also conserved in Bxz1 and Myrna (Fig. 2).

DNA processing and packaging. A 24.6 kb region from *E3_1360* to *E3_1640* is largely devoted to DNA replication, repair and recombination (Fig. 2). Comparison with Bxz1 and Myrna identifies two sections: variable on the left, which begins with the divergently transcribed helicase gene *E3_1360* (conserved in the three phages), is part of an HPR (HPR-3, see below) and is identified as HGT DNA; and conserved on the right, encoding putative helicase loader DnaC, ATP-dependent helicase DnaB, DNA primase DnaG, chaperonin protein DnaJ, and DNA polymerase IIIα subunit, plus two putative Holliday resolvases gp156 (*E3_1560*) and gp158 (*E3_1580*). Interestingly, in contrast to gp156, gp158 shows no homology to *Myoviridae* products but is closely related to *Siphoviridae* proteins (*R. equi* ReqiPine5 gp08, *Tsukamurella* phage TPA2 gp15 and *Nocardia* phage NBR1 gp65). Apart from the homology to Bxz1 and Myrna, the closest homologues of the DNA processing region are mostly of bacterial origin.

DNA packaging is predicted to be mediated by gp5/TerL, a member of the terminase 1 family (PF03354) similar to the T4 large terminase, gp17 (Sun *et al.*, 2008), and distantly related to the putative large terminase of *R. equi* phage ReqiDocB7 (Summer *et al.,* 2011). Large terminases are characterised by the presence of an ATP-binding Walker A motif, for which a putative deviant motif (GRRASKG) (Mitchell and Rao, 2004) was identified in gp5 and its mycobacteriophage homologues.

Lysis

Tailed ds-DNA phages typically possess a lysis cassette encoding holin and endolysin, which together are responsible for degrading the host cell wall during the lytic infection cycle. Although not identified for Bxz1 and Myrna, a putative holin gene (*E3_0020*) was found close to *terL* (*E3_0050*). With no similarity to any *R. equi* phage protein reported to date, its closest homologue is the *Lactococcus* phage r1t holin (Sanders *et al.*, 1997). Structural analysis revealed that E3 gp2 is related to the class 3 holins, which includes the *S* gene product of phage λ. The putative E3 endolysin gene (*E3_0980*), with again no homologue in Bxz1 and Myrna, is located 56.4 kb downstream of the holin gene in HPR-2. Gp98 shares similarity to LysA of mycolata *Siphoviridae* including the mycobacteriophages ms6 and Ch8 (Garcia *et al.*, 2002; Payne *et al.*, 2009), *R. equi* phage ReqiDocB7 (Summer *et al.*, 2011) and *Tsukamurella* phage TPA2 (Petrovski *et al.*, 2011a). Significantly, no phage E3 LysA homologues were found in other *Myoviridae*. Gp98 possesses two domains: an N-terminal amidase domain (PF01510) and a C-terminal LGFP (Leu, Gly, Phe and Pro) repeat domain (PF08310) reported in cell wall-associated proteins of the mycolata and believed to play a role in protein anchoring thereby maintaining cell wall integrity (Adindla *et al.*, 2003).

Interestingly, lytic domains were also predicted for the E3 putative baseplate hub protein gp119, including a soluble lytic transglycosylase (SLT) domain (PF01464) and a g-

D,L-glutamate-specific amidohydrolase NLPC/P60 domain (PF00877). Although related to the key baseplate central protein gp44 of Mu and gp27 of T4, phage Mu gp44 does not possess tail lysozyme activity, while the T4 gp27 interacts with a lysozyme encoding protein (Kanamaru *et al.*, 2002; Kondou *et al.*, 2005). Interestingly, the baseplate protein for *E. coli* O157:H7 phage CBA120 possesses an NLPC/P60 domain flanked by sequences with homology to regions of T4 gp5 responsible for T4 gp5/gp27 interaction (Kutter *et al.,* 2011). The lytic domains in the baseplate protein may constitute a "punching device" to aid penetration of the peptidoglycan layer during the infection process (Kanamaru *et al.*, 2002). The gp119 homologues in the myoviridal mycobacteriophages lack the SLT domain and thus the E3 multidomain gp119 represents a novel arrangement for baseplate proteins.

Phages infecting mycobacteria, which possess lipid-rich cell envelopes, encode auxiliary lysins with lipolytic activity generically designated LysB. Three such LysB enzymes are potentially encoded by E3, all in HPRs: gp84, a putative SGNH lipolytic protein of the serine hydrolase family; gp85 cutinase, homologous to LysB of *Mycobacterium* phages D12 and Ms6, for which lipolytic activity has been experimentally determined (Gil *et al.*, 2008; Payne *et al.*, 2009); and gp167 with structural similarity to mycobacteriophage D29 LysB. E3 appears thus to be particularly well endowed in lipolytic proteins, being the first phage reported with three putative LysB proteins.

Hyperplastic regions

The four HPRs in E3, Bxz1 and Myrna mostly encode hypothetical proteins with no significant similarity to products of the corresponding region in the three phages (or indeed any other phage in protein sequence databases). The HPR genes are typically smaller in size compared to genes in the conserved modules (~390 vs 1074 bp). The presence of clusters of small ORFs, with an average of 100 codons, is relatively common within bacteriophages and tend to be associated with regions subjected to greater genetic flux (Hatfull *et al.*, 2010). The HPRs in the three phages are rich in horizontally acquired (HGT) DNA, indicating that they mainly evolve through lateral exchange, while HGT DNA is generally absent from the conserved gene modules (Fig. 2), consistent with a vertical evolutionary pattern. Some of the E3 HPR products are similar to bacterial or eukaryotic proteins with no (or exceptional) phage homologues, suggesting the possibility of a non-viral origin (see Supplementary text). Many of the HPR products are secreted or transmembrane proteins possibly related to host adaptation/virulence functions. Specific features of two of the E3 HPRs are discussed below.

HPR-2: recent acquisition of structural and infectivity traits. HPR-2 interrupts the conserved morphogenesis module between the head and tail assembly units of E3, Bxz1 and Myrna. In E3 it is significantly larger, with a mosaic of HGT genes encoding hypothetical proteins, the putative amidase/LysA endolysin (gp98), two LysB lipolytic enzymes (gp84, 85), and structural proteins including two putative tail fibre proteins (gp86, gp88). Seven additional HPR-2 products were identified as structural proteins by LC-ESI-MS/MS (gp87, gp89, gp99, gp100, gp106 to 108) (Table S2). These features suggest this HPR may encode products relevant to head/tail assembly and also phage infectivity. Except for the putative tail fibre gene *E3_0880* (gp88), none of the HPR-2 genes have homologues in Bxz1 or Myrna. Interestingly, a second HPR-2-encoded E3 tail fibre protein, gp86 (*E3_0860*), is highly similar to proteins of *R. equi* phages, ReqiPoco6 and ReqiPepy6 of the *Siphoviridae* family (Summer *et al.*, 2011). A phylogenetic analysis confirmed that gp88 is evolutionarily related to its Bxz1/Myrna homologues whereas gp86 shares a common origin with tail fibre proteins from *R. equi* siphoviruses (Fig. S3). An additional putative tail fibre protein (gp204), possessing a phage-related tail fibre domain (COG5301), is phylogenetically related to Bxz1/Myrna products (Fig. S3) and encoded by syntenically conserved genes in a cassette immediately downstream of HPR-4 (Fig. 2). Since tail fibre proteins are involved in the binding of the phage to the surface of the host bacterium, the horizontal acquisition of *E3_0860* by E3 may have been critical to gain tropism towards *R. equi*. The evolutionary pattern of E3 putative tail fibre genes, combining vertical evolution and genetic exchanges with distantly related phages, provides clues about the shaping of host adaptation in the *Caudovirales*.

HPR-3: adaptation to the host genome? In E3, Bxz1 and Myrna, HPR-3 includes the divergently transcribed "helicase" locus, comprising three (E3 and Myrna) to four (Bxz1) genes, all different in the three phages except for the conserved helicase gene (Fig. 2). In the mycobacteriophages, this locus is interrupted by a cluster of 23 (Bxz1) to 32 (Myrna) tRNA genes, which is completely absent in E3. There are two additional small tRNA clusters in the Bxz1 and Myrna genomes, which are also absent in E3 (Fig. 2). While some phage genomes lack or have few tRNA genes, others have as many as the host bacteria, with the number of tRNAs being generally positively associated with phage genome size (Bailly-Bechet *et al.*, 2007). The total absence of tRNAs in E3 and the overabundance in the similarly sized and genomically and evolutionarily closely related mycobacterial myoviruses is therefore intriguing. tRNAs are typical integration sites for mobile DNA elements and it has been suggested they are continually recruited during the course of multiple integration events, with accumulation in the phage genome if providing a selective advantage that counteracts the natural deletion bias of non-essential DNA (Williams, 2002). tRNAs may be important for translation-associated phage fitness by compensating differences in codon usage with the host bacterium, becoming positively selected if the corresponding codons are highly used by the phage and rare in the host genome (Bailly-Bechet *et al.,* 2007). While the bacterial hosts for E3 and the Bxz1 and Myrna mycobacteriophages do not appreciably differ in composition of the corresponding genomic tRNA pools, they do differ significantly in genome size (5.0 Mbp for *R. equi* vs ≈7 Mbp for *M. smegmatis*). If tRNA gene expression is rate-limiting, the larger host genome for Bxz1 and Myrna may necessitate additional tRNAs to support efficient multiplication of the parasitic phage.

Concluding remarks

This study reports the first *Myoviridae* infecting a non-mycobacterial actinomycete and the first myoviridal phage hosted by a member of the genus *Rhodococcus*. There is a paucity of *Myoviridae* isolated to date infecting the mycolata, a group of *Actinobacteria* comprising a number of genera of environmental, industrial and medical relevance. In addition, the distribution of available phage sequences within this bacterial group is clearly skewed towards mycobacteriophages, limiting the significance of comparative genomics and phage evolutionary studies. Our findings therefore contribute to fill an existing gap in the diversity of genome sequences available for *Actinobacteria* phages, in particular those infecting mycolic acid-containing actinomycetes.

In a recent study by Lavigne *et al.* (2009) a classification was proposed for all *Myoviridae* into three sub-families and eight independent genera, one of which is the proposed ‘Bxz1-like’ or ‘I3-like‘ genus consisting of the myoviridal mycobacteriophages. Our findings support a case for redefinition of this bacteriophage genus or grouping as ‘mycolata-infecting *Myoviridae*’, with possibly E3 as the reference member since it now represents the best-characterised example of these phages. The comprehensive bioinformatic and proteomic analysis of *R. equi* phage E3 has contributed to the refinement of the annotation of the related mycobacterial myoviruses.

The genomes within the mycolata-infecting *Myoviridae* group have a modular conserved backbone, encoding the essential machinery for phage life cycle, with interspersed laterally acquired hypervariable regions (HPR) that form the basis for the genetic diversity and specialisation. While unique to each phage, these HPRs are syntenically located, indicating they are conserved hot spots for lateral exchange and genome mosaicism. HPRs would appear to encode important infectivity and host tropism traits including enzymatic activities required for phage penetration and release during the infection cycle. In E3, the endolysins LysA targeting the bacterial peptidoglycan and the three predicted LysB proteins, targeted at the lipid-rich bacterial cell envelopes, are encoded by ORFs located within the HPRs. These gene products are amongst the few of phage E3 bearing significant similarity to proteins from other *R. equi* phages, suggesting lateral acquisition of host-specific infectivity traits via lateral exchanges with other *Rhodococcus* phages. Another example is the tail fibre genes identified in HPR-2, one of which is syntenically conserved in the mycobacteriophages while the other encodes a product homologous to tail fibre proteins of *R. equi Siphoviridae*.

The most important limitation for comparative genomic studies of mycolata phages lies in the fact that complete sequence data within this group are currently limited, with only seven *Siphoviridae* and one *Myoviridae* genome sequences available for the genera *Rhodococcus*, *Tsukamurella* and *Corynebacterium,* compared to in excess of 230 *Siphoviridae* and 23 *Myoviridae* for *Mycobacterium*. Our study highlights the importance of isolating and comparatively analysing the genomes of *Myoviridae* infecting other mycolic acid-containing actinomycetes to gain further insight into the evolutionary history of the mycolata phages and their relationship within the *Caudovirales*. Given the extraordinarily fast evolutionary dynamics and mosaicism of phage genomes, our data with phylogenetically and genomically closely related phages infecting different bacteria provide clues to understand short-term phage genome evolution in connection to host adaptation.

Experimental procedures

Phage isolation and microscopy

Soil samples were screened for the presence of phages following the method described by Dabbs (1998) using *R. equi* NCIMB 10027. Following three rounds of plaque purification, host range was analysed using a spot assay technique. Strain details are provided in Table S3. Caesium chloride-purified phages were observed by transmission electron microscopy (Zeiss 912 energy filtering transmission electron microscope) following dialysis against phage buffer (40 mM Tris-HCl, 100 mM NaCl and 10 mM $MgSO_4$, pH 7.4) and methylamine vanadate staining on S Formvar/Carbon-coated 200 mesh copper grids operating at 120kV.

Genome sequencing and annotation

Phages were concentrated and purified according to Sambrook *et al*. (1989) with the following modifications: phage lysate was incubated with 0.5M NaCl for 1 h at 4^oC, prior to centrifugation at 5,000g for 10 min at 4^oC, and the pellet resuspended in phage buffer prior to loading on CsCl step gradient. DNase I and RNase A were added to the purified phage particles solution at 1 mg ml^{-1} final concentration prior to addition of EDTA and proteinase K, and finally DNA precipitation using isopropanol. Shotgun E3 genome sequencing was carried out using 454 pyrosequencing (Roche). Gaps between contigs of an ~24-fold coverage shotgun assembly were closed manually by PCR. The software and databases used for genome analysis and annotation are shown in Table S4. The complete E3 DNA sequence and genome annotation has been deposited in GenBank under accession no. HM114277.

Phylogenetic analyses

Protein sequences were aligned using ClustalX v2.0 (Larkin *et al*., 2007) under default parameters and Maximum-Likelihood and Neighbor-Joining phylogenetic trees constructed with PhyML v2.4.5 (Guindon and Gascuel, 2003) and MEGA v5.0 (Tamura *et al*., 2011), respectively. The latter programme was used for tree visualization and edition.

Phage proteomics

Approx. 5 µg of double CsCl purified phages were subjected to SDS-PAGE (12% tris/glycine mini-gel, Invitrogen). Protein bands were visualised using SimplyBlue Safe Stain™ (Invitrogen), sliced and subjected to standard in-gel trypsinisation (Shevchenko *et al*., 1996). LC-ESI-MS/MS analysis was performed as described by Batycka *et al.* (2006) using a monolithic reversed phase column (200 mm ID; Dionex-LC Packings). Deconvoluted MS/MS data was submitted to an in-house MASCOT server, searched against a cognate *R. equi* E3 phage genomic database and analysed in accordance with published guidelines (Taylor and Goodlett, 2005).

Acknowledgements

We thank the Genepool facility, University of Edinburgh, for DNA sequencing, L. Tetley of the Integrated Microscopy Facility, University of Glasgow, for electron microscopy, and M. Letek for his input during the initial stages of the project. This work received financial support from the Horserace Betting Levy Board (grant Prj-753 to JV-B). AV-R was supported by an EU FP7 Marie-Curie intra-European fellowship.

References

Abascal, F., Zardoya, R., and Posada, D. (2005) ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21**: 2104-2105.

Adindla, S., Inampuldi, K.K., Guruprasad, K., and Guruprasad, L. (2003) Identification and analysis of novel tandem repeats in the cell surface proteins of archaeal and bacterial genomes using computational tools. *Comp Funct Genomics* **5**: 2-16.

Bailly-Bechet, M., Vergassola, M., and Rocha, E. (2007) Causes for the intriguing presence of tRNAs in phages. *Genome Res* **17**: 1486-1495.

Batycka, M., Inglis, N.F., Cook, K., Adam, A., Fraser-Pitt, D., Smith, D.G.E. *et al*. (2006) Ultra-fast tandem mass spectrometry scanning combined with monolithic column liquid chromatography increases throughput in proteomic analysis. *Rapid Com Mass Spect* **20**: 2074-2080.

Brüssow, H., Canchaya, C., and Hardt, W.-D. (2004) Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiol Mol Biol Rev* **68**: 560-602.

Casjens, S.R., and Thuman-Commike, P.A. (2011) Evolution of mosaically related tailed bacteriophage genomes seen through the lens of phage P22 virion assembly. *Virology* **411**: 393-415.

Dabbs, E.R. (1998) Cloning of genes that have environmental and clinical importance from rhodococci and related bacteria. *Antonie van Leeuwenhoek* **74**: 155-168.

Dawson, T.R., Horohov, D.W., Meijer, W.G., and Muscatello G. (2010) Current understanding of the equine immune response to *Rhodococcus equi*. An immunological review of *R. equi* pneumonia. *Vet Immunol Immunopathol* **135**: 1-11.

Garcia, M., Pimentel, M., and Moniz-Pereira, J. (2002) Expression of mycobacteriophage Ms6 lysis genes is driven by two σ70-like promoters and is dependent on a transcription termination signal present in the leader RNA. *J Bacteriol* **184**: 3034-3043.

Giguère, S., Cohen, N.D., Chaffin, M.K., Slovis, N.M., Hondalus, M.K., Hines, S.A., and Prescott, J.F. (2011) Diagnosis, treatment, control, and prevention of infections caused by *Rhodococcus equi* in foals. *J Vet Intern Med* **25**: 1209-1220.

Gil, F., Catalao, M.J., Moniz-Pereira, J., Leandro, P., McNeil, M., and Pimentel, M. (2008) The lytic cassette of mycobacteriophage Ms6 encodes an enzyme with lipolytic activity. *Microbiology* **154**: 1364-1371.

Glazko, G., Makarenkov, V., Liu, J., and Mushegian, A. (2007) Evolutionary history of bacteriophages with double-stranded DNA genomes. *Biol Direct* **2**: 36.

Guindon, S., and Gascuel, O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52**: 696-704.

Hatfull, G.F. (2006) Mycobacteriophages. In *The Bacteriophages.* Calendar, R. (ed). New York: Oxford University press, pp. 602-620.

Hatfull, G.F., Jacobs-Sera, D., Lawrence, J.G., Pope, W.H., Russell, D.A., Ko, C.C. *et al*. (2010) Comparative genomic analysis of 60 mycobacteriophage genomes: genome clustering, gene acquisition, and gene size. *J Mol Biol* **397**: 119-143.

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

Kanamaru, S., Leiman, P.G., Kostyuchenko, V.A., Chipman, P.R., Mesyanzhinov, V.V., Arisaka, F., and Rossmann, M.G. (2002) Structure of the cell-puncturing device of bacteriophage T4. *Nat* **415**: 553-557.

Kondou, Y., Kitazawa, D., Takeda, S., Tsuchiya, Y., Yamashita, E., Mizuguchi, M., *et al*. (2005) Structure of the central hub of Bacteriophage Mu baseplate determined by X-ray crystallography of gp44. *J Mol Biol* **352**: 976-985.

Kutter, E.M., Skutt-Kakaria ,K., Blasdel, B., El-Shibiny, A., Castano, A., *et al*., (2011) Characterization of a ViI-like phage specific to *Escherichia coli* O157:H7. *Virol J.* **8**:430.

Larkin, M.J., Kulakov, A., and Allen, C.C.R. (2005) Biodegradation and *Rhodococcus* – masters of catabolic versatility. *Curr Opinion Biotechnol* **16:**282–290.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., *et al*. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* **23**: 2947-2948.

Lavigne, R., Darius, P., Summer, E., Seto, D., Mahadevan, P., Nilsson, A., *et al*. (2009) Classification of *Myoviridae* bacteriophages using protein sequence similarity. *BMC Microbiol* **9**: 224.

Letek, M., González, P., MacArthur, I., Rodríguez, H., Freeman, T.C., Valero-Rello, A., *et al*. (2010) The genome of a pathogenic *Rhodococcus*: Cooptive virulence underpinned by key gene acquisitions. *PLoS Genetics* **6**: e1001145.

Liu, J., Glazko, G., and Mushegian, A. (2006) Protein repertoire of double-stranded DNA bacteriophages. *Virus Res* **117**: 68-80.

Mitchell, M.S., and Rao, V.B. (2004) Novel and deviant Walker A ATP-binding motifs in bacteriophage large terminase-DNA packaging proteins. *Virology* **321**: 217-221.

Monier, A., Claverie, J.M., and Ogata, H. (2008) Taxonomic distribution of large DNA viruses in the sea. *Genome Biol* **9**:R106.

Muscatello, G., Leadon, D.P., Klayt, M., Ocampo-Sosa, A., Lewis, D.A., Fogarty, U., *et al*. (2007) *Rhodococcus equi* infection in foals: the science of 'rattles'. *Equine Vet J* **39**:470-478.

Ocampo-Sosa, A.A., Lewis, D.A., Navas, J., Quigley, F., Callejo, R., Scortti, M., *et al*. (2007) Molecular epidemiology of *Rhodococcus equi* based on *traA*, *vapA*, and *vapB* virulence plasmid markers. *J. Infect. Dis.* **196**:763-769.

Payne, K., Sun, Q., Sacchettini, J., and Hatfull, G.F. (2009) Mycobacteriophage lysin B is a novel mycolylarabinogalactan esterase. *Mol Microbiol* **73**: 367-381.

Pedersen, M., Østergaard, S., Bresciani, J., and Vogensen, F.K. (2000) Mutational analysis of two structural genes of the temperate lactococcal bacteriophage TP901-1 involved in tail length determination and baseplate assembly. *Virology* **276**: 315-328.

Pedulla, M.L., Ford, M.E., Houtz, J.M., Karthikeyan, T., Wadsworth, C., Lewis, J.A., *et al*. (2003) Origins of highly mosaic mycobacteriophages genomes. *Cell* **113**: 171 - 182.

Petrovski, S., Seviour, R.J., and Tillett, D. (2011) Genome sequence and characterization of the *Tsukamurella* bacteriophage TPA2. *Appl Environ Microbiol* **77**: 1389-1398.

Prescott, J.F. (1991) *Rhodococcus equi*: animal and human pathogen. *Clin Microbiol Rev* **4:** 20-24.

Sambrook, J., Fritsch, E.F., and Maniatis, T. (1989) *Molecular cloning: a laboratory manual*. Cold Spring Harbor: Cold Spring Harbor Laboratory Press pp 2.73-2.81

Sanders, J.W., Venema, G., and Kok, J. (1997) A chloride-inducible gene expression cassette and its use in induced lysis of *Lactococcus lactis*. *Appl Environ Microbiol* **63**: 4877-4882.

Shevchenko, A., Wilm, M., Vorm, O., and Mann, M. (1996) Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels. *Anal Chem* **68**: 850-858.

Summer, E.J., Liu, M., Gill, J.J., Grant, M., Chan-Cortes, T.N., Ferguson, L., *et al*. (2011) Genomic and functional analyses of *Rhodococcus equi* phages ReqiPepy6, ReqiPoco6, ReqiPine5, and ReqiDocB7. *Appl Environ Microbiol* **77**: 669-683.

Sun, S., Kondabagil, K., Draper, B., Alam, T.I., Bowman, V.D., Zhang, Z., *et al*. (2008) The structure of the phage T4 DNA packaging motor suggests a mechanism dependent on electrostatic forces. *Cell* **135**: 1251-1262.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011) MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* **28**: 2731-2739

Taylor, G.K., and Goodlett, D.R. (2005) Rules governing protein identification by mass spectrometry. *Rapid Commun Mass Spectrom* **19**: 3420.

Vazquez-Boland, J.A., Letek, M., Valero-Rello, A., González, P., and Scortti, M. (2010) *Rhodococcus equi* and its virulence mechanisms. In: Microbiol. Monogr. (H.M. Alvarez, *ed*) The Biology of *Rhodococcus*, pp. 331-360. Springer Verlag.

Weinstock, D.M., and Brown, A.E. (2002) *Rhodococcus equi*: an emerging pathogen. *Clin Infect Dis* 34:1379-1385.

Williams, K.P. (2002) Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies. *Nucleic Acids Res* **30**: 866-875.

Xu, J., Hendrix, R., and Duda, R.L. (2004) Conserved translational framesift in dsDNA bacteriophage tail assembly genes. *Mol Cell* **16**: 11-21.

Figure Legends

Fig. 1. Electron micrograph of E3 phage, a *Myoviridae* typified by the presence of long inflexible, contractile tails with a constriction (“neck”) between head and tail. Capsid diameter is approx. 93.55 + 2.53 nm and tail is of a similar length (94.28 + 2.07 nm) based on measurements taken on 5 individual phage particles

Fig. 2. Genomic maps of *Rhodococcus equi* phage E3 and *Mycobacterium* phages Myrna and Bxz1. To facilitate genetic structure/synteny comparison, the terminase region was arbitrarily chosen to “linearise” the E3, Bxz1 and Myrna circular genomes (*gp243* and *gp236* as first ORFs in the linearised Myrna and Bxz1, respectively). In E3, the nucleotide coordinates start 1,650 bp upstream the 5’ end of the *terL* (terminase) gene. Pale blue shadowed links indicate genes encoding protein homologues based on BLASTclust algorithm. ORFs are colour coded according to predicted functions: red, DNA and RNA metabolism; blue, transcription factor; pale green, membrane and secreted proteins; dark green, morphogenesis; magenta, lysis proteins; yellow, other enzymes; grey, conserved hypothetical proteins. HGT regions are underlined, tRNA clusters and hyperplastic regions (HPRs) are boxed with solid or dashed rectangles, respectively. Vertical arrows indicate proteins with significant similarity to other *Rhodococcus* phages; black dots, virion associated proteins confirmed by LC-ESI-MS/MS proteomic analysis (see Table S2); triangles, tail fibres proteins (see Fig. S3). Pairwise alignments of Bxz1-like phages showed all to be almost identical phage species, therefore Bxz1 was selected as representative of this group. Genes mentioned in the text are labelled. Annotations in Bxz1 and Myrna are based on Hatfull *et al.* (2010), with an indication of revised or newly assigned functions (indicated by a star). Abbreviations: Prohead-P, prohead protease; TMP, tape measure protein; SLT, soluble lytic transglycosylase. Genomic maps were built in XPlasMap v0.96 (http://www.iayork.com).

Fig. 3. Maximum Likelihood tree of gp5 (TerL) and related phage terminase large subunit proteins. Model of protein evolution: Blosum62 with estimated Gamma distribution, proportion of invariable sites and empirical frequencies (Blosum62+G+I+F). The best model of evolution for protein sequence as determined by jProtTest v2.4 (Abascal *et al.*, 2005), according to AIC criterion, was used. Numbers in nodes are percent bootstrap for 100 replicates; values under 50% are not represented. Families according to ICTV and NCBI classification are represented in: green, *Myoviridae*; yellow, *Podoviridae*; non-shaded, *Siphoviridae*. The reference bacteriophages for established (solid boxes) or proposed genus groups (dotted boxes) are indicated by asterisks. Numbers in brackets represent the global nucleotide similarity percentages to the reference genome in the respective genus group. Phylum of bacterial hosts is indicated for each taxon by coloured dots: black, *Actinobacteria*; white, *Firmicutes*; red, *Proteobacteria*. The scale shows the number of amino acid substitutions per site.

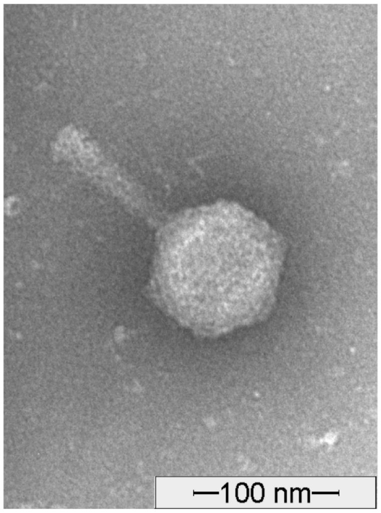

Fig. 1
297x420mm (300 x 300 DPI)

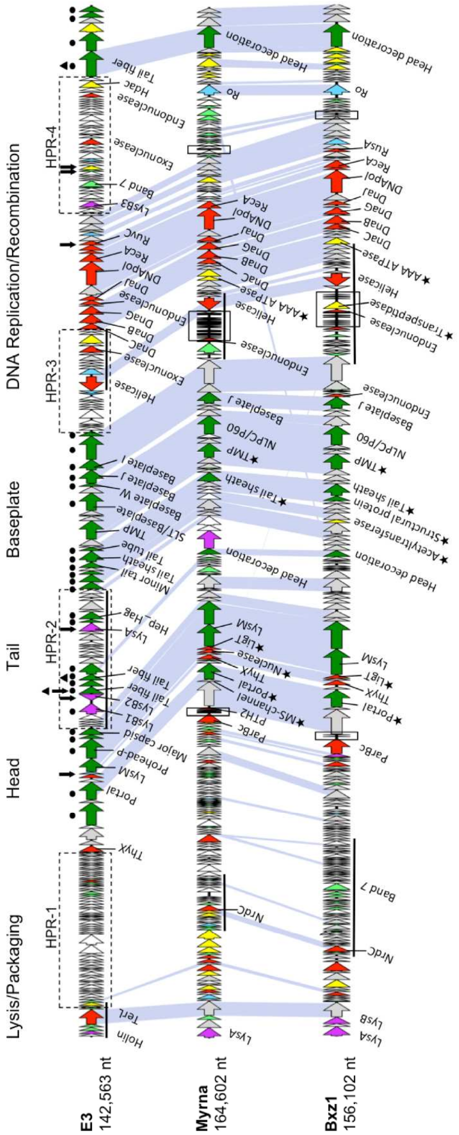

Fig. 2
297x420mm (300 x 300 DPI)

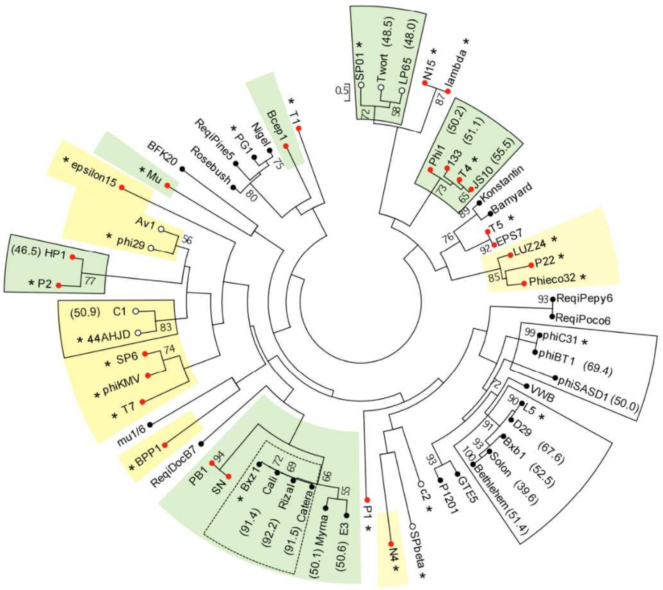

Fig. 3
297x420mm (300 x 300 DPI)

ONLINE SUPPORTING INFORMATION

Genomic and proteomic analysis of phage E3 infecting the soil-borne actinomycete *Rhodococcus equi*

Samson P. Salifu, Ana Valero-Rello, Samantha A. Campbell, Neil F. Inglis, Mariela Scortti, Sophie Foley *, and José A. Vazquez-Boland *

* Email: s.foley@napier.ac.uk, v.boland@ed.ac.uk

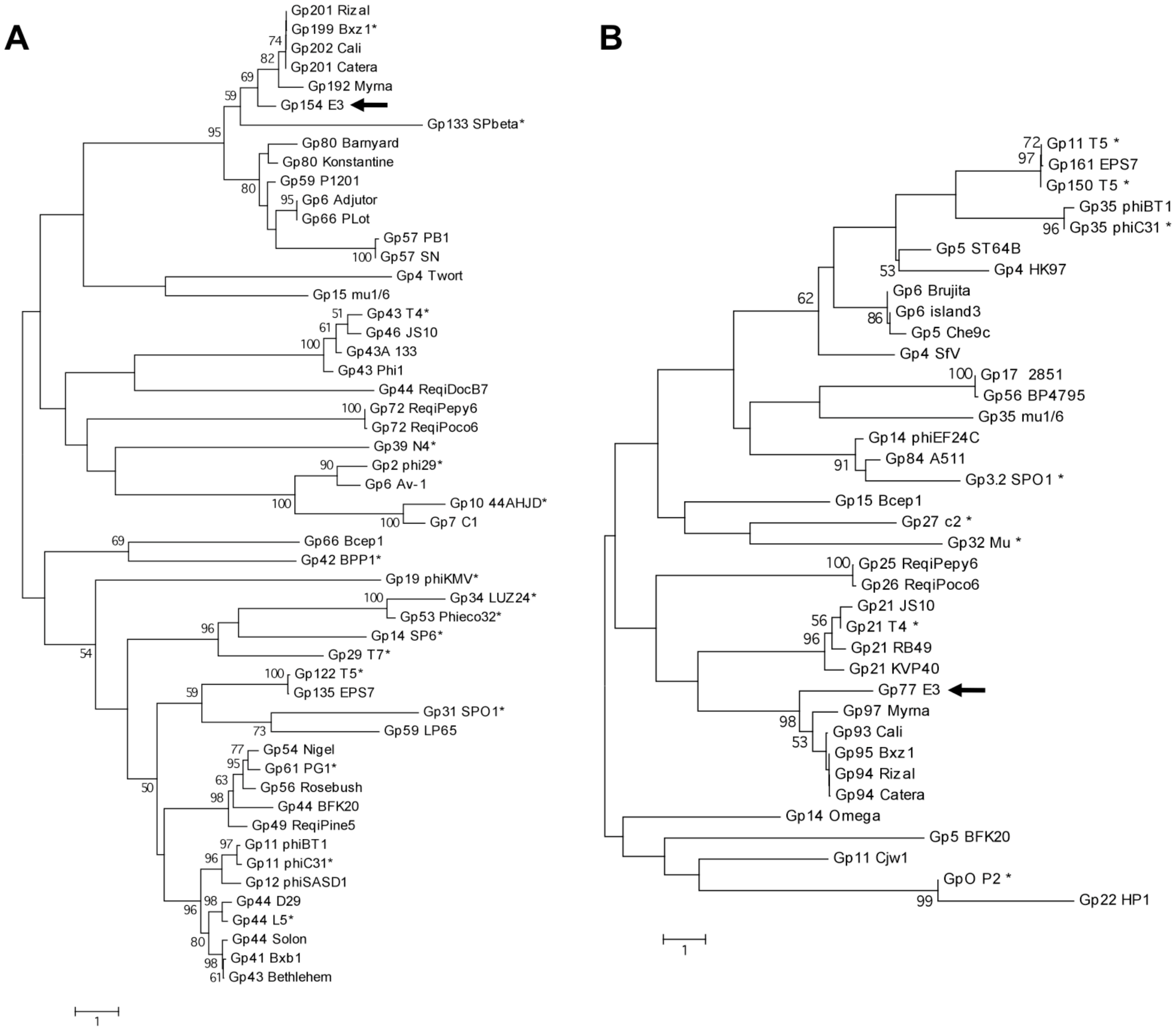

Fig. S1. Neighbor Joining unrooted trees of (A) DNA polymerase (E3 gp154) and (B) prohead protease (E3 gp77). Numbers in nodes are the percent bootstrap values for 1000 replicates; values under 50% are not represented. Reference bacteriophages for accepted genera according to ICTV and NCBI taxonomy are indicated by asterisks. E3 proteins are indicated by arrows. The scale shows the number of amino acid substitutions per site. The topology of the phylogenetic trees (including the TerL tree; see Fig. 3) reproduced the branching pattern of phage phylogenies based on whole genomes (Rohwer and Edwards, 2002; Glazko *et al.*, 2007), and most well-suported clades grouped phages classified within an established genus.

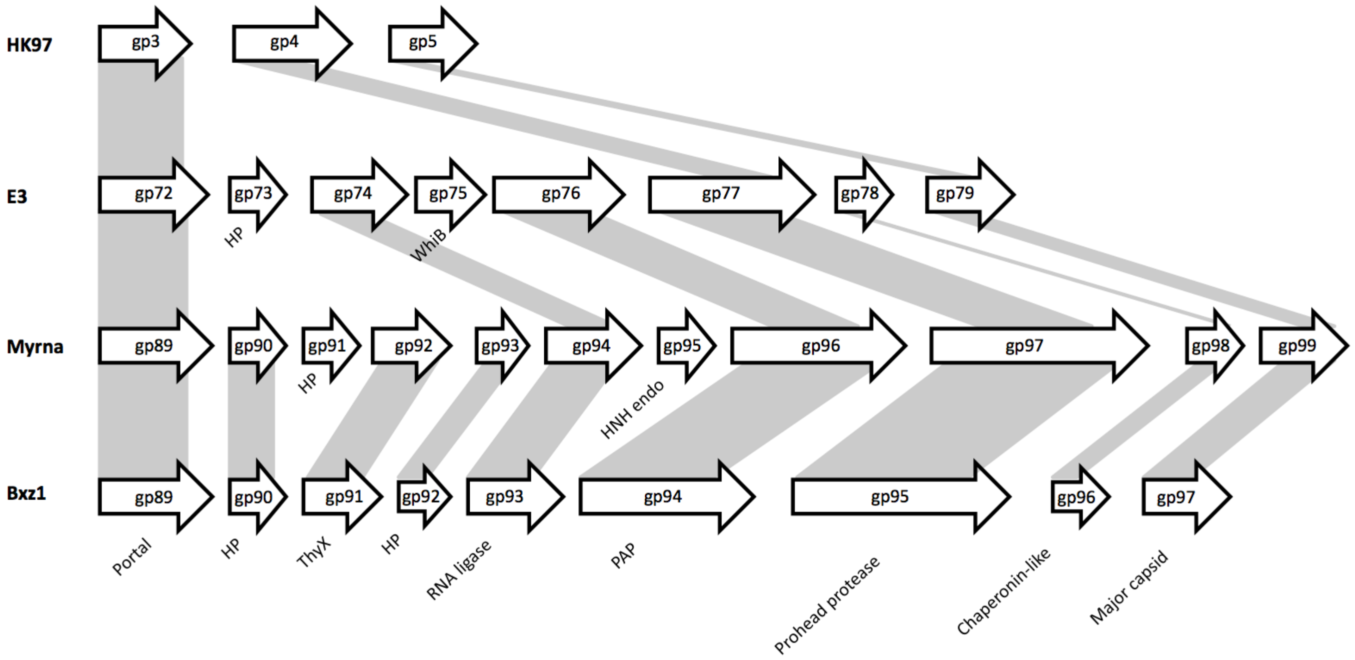

Fig. S2. Alignment of the head morphogenesis module of *R. equi* phage E3, enterobacteriophage HK97, and mycobacteriophages Bxz1 and Myrna. Pairwise sequence similarity between adjacent genomes is indicated by shading. HP: hypothetical protein; PAP: protease-associated protein; ThyX: FAD-dependent thymidylate synthase; HNH endo: HNH endonuclease.

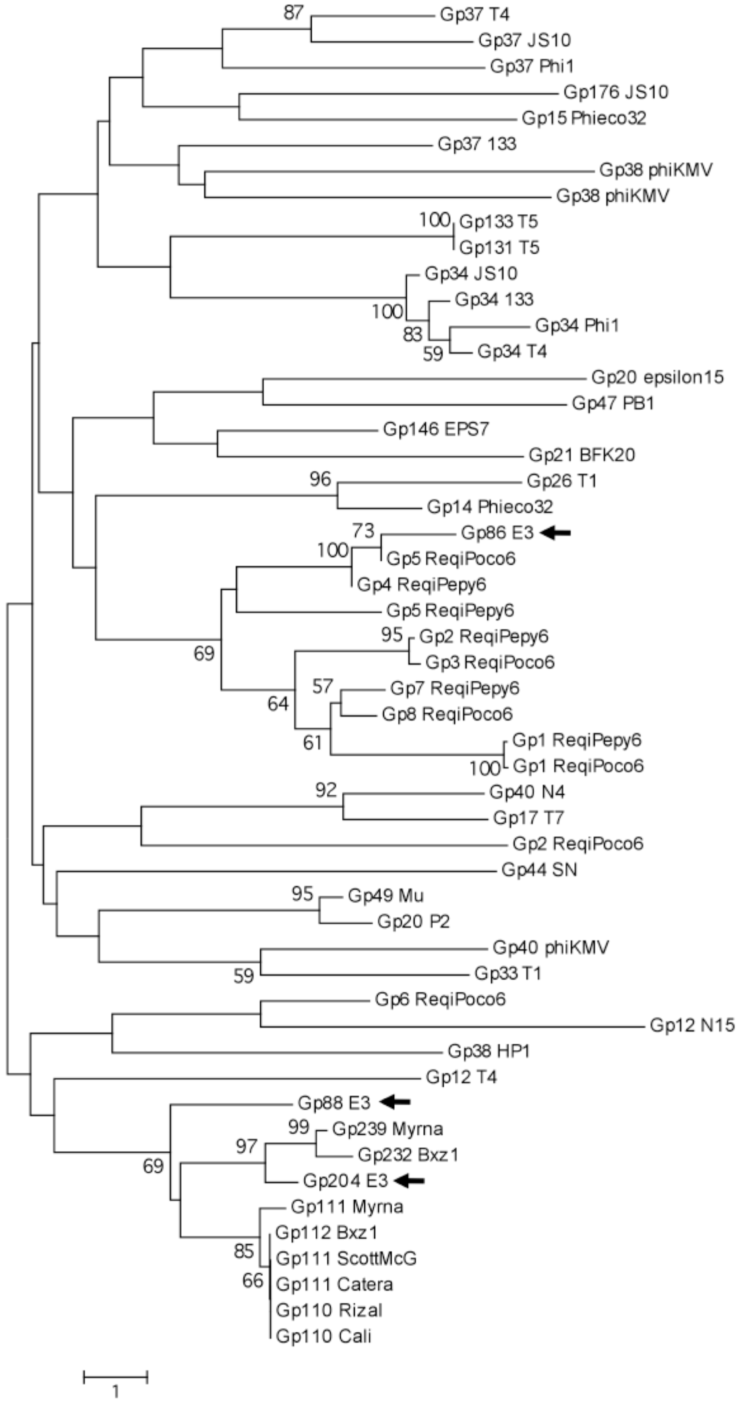

Fig. S3. Neighbor Joining unrooted tree of E3 tail fibre proteins (gp86, gp88 and gp204). The numbers in nodes are the percent bootstrap values for 1000 replicates; values under 50% are not represented. Arrows indicate E3 tail fibre proteins. The scale shows the number of amino acid substitutions per site.

Table S1. Annotation of bacteriophage E3 genome [a].

Locus (strand) [b]	Coordinates [c] (size nt / %GC)	Product (size aa / kDa)	Putative Function	Domain / Motif	Closest Homologue [c]	Acc. no. (E-value <10^{-3})	% Similarity [d] (overlap)
E3_0010	58-366 (309 /63.75)	gp1 (102/11.4)					
E3_0020	363-770 (408 /64.7)	gp2 (135/14.3)	Holin	4 TMDs	HP *Rhodococcus equi*	ZP_06828142 (6e-11)	57 (68/119)
E3_0030	798-1244 (447 /64.87)	gp3 (148/16.1)		4 TMDs			
E3_0040	1312-1650 (339/62.83)	gp4 (112/12.9)		Signal peptide 1 TMD			
E3_0050	1661-3724 (2064/65.93)	gp5 (687/78.5)	Large terminase		gp239 *Mycobacterium* phage Bxz1	NP_818289 (0.0)	65 (428/661)
E3_0060	3779-4009 (231/67.53)	gp6 (76/8.0)		Signal peptide			
E3_0070	4039-4251 (213/65.25)	gp7 (70/8.0)					
E3_0080	4244-4705 (462/65.36)	gp8 (153/17.6)	Polynucleotide dikinase		HP *Saccharopolyspora erythraea*	ZP_06562255 (1e-24)	58 (80/139)
E3_0090	4705-4983 (279/65.94)	gp9 (92/10.5)		Signal peptide 1 TMD			
E3_0100	5068-5391 (324/69.75)	gp10 (107/12.1)					
E3_0110	5427-5819 (393/66.41)	gp11 (130/14.9)					
E3_0120	5895-6152 (258/67.44)	gp12 (85/9.2)					
E3_0130	6179-6376 (198/65.15)	gp13 (65/7.5)					
E3_0140	6379-6642 (264/70.83)	gp14 (87/9.6)					
E3_0150	6639-7127 (489/68.3)	gp15 (162/17.6)		Coiled coil			

Locus (strand) [b]	Coordinates [c] (size nt / %GC)	Product (size aa / kDa)	Putative Function	Domain / Motif	Closest Homologue [c]	Acc. no. (E-value $<10^{-3}$)	% Similarity [d] (overlap)
E3_0160	7127-7486 (360/66.38)	gp14 (119/13.6)					
E3_0165	7486-7668 (183/69.94)	gp 16.5 (60/6.6)					
E3_0170	7669-8121 (453/66.88)	gp17 (150/17.5)					
E3_0180	8198-8452 (255/ 65.09)	gp18 (84/10.0)					
E3_0190	8449-8892 (444/69.36)	gp19 (147/16.4)					
E3_0200	8987-9244 (258/68.21)	gp20 (85/9.2)					
E3_0210	9241-9423 (183/67.21)	gp21 (60/7.1)					
E3_0220	9420-9812 (393/69.21)	gp22 (130/14.4)		Coiled coil UPF0150	HP *Mycobacterium marinum*	YP_001852174 (4e-18)	60 (73/122)
E3_0230	9854-10249 (396/66.41)	gp23 (131/14.1)					
E3_0240	10246-10533 (288/66.31)	gp24 (95/10.6)					
E3_0250	10530-11060 (531/67.79)	gp25 (176/20.0)					
E3_0260	11072-11416 (345/64.92)	gp26 (114/12.9)			gp133 *Mycobacterium* phage Omega	NP_818432 (4e-06)	56 (42/76)
E3_0270	11416-12630 (1215/69.54)	gp27 (404/45.2)					
E3_0280	12642-14021 (1380/67.97)	gp28 (459/51.7)					
E3_0290	14098-14370 (273/67.76)	gp29 (90/9.9)					
E3_0300	14380-14688 (309/66.99)	gp30 (102/11.4)					

Locus (strand) [b]	Coordinates [c] (size nt / %GC)	Product (size aa / kDa)	Putative Function	Domain / Motif	Closest Homologue [c]	Acc. no. (E-value $<10^{-3}$)	% Similarity [d] (overlap)
E3_0310	14685-14942 (258/70.93)	gp31 (85/9.7)					
E3_0320	14939-15238 (300/72.33)	gp32 (99/10.8)					
E3_0330	15238-15555 (318/67.61)	gp33 (105/12.2)					
E3_0340	15555-16058 (504/69.24)	gp34 (167/18.3)					
E3_0350	16055-16240 (186/66.66)	gp35 (61/6.7)					
E3_0360	16237-16512 (276/71.01)	gp36 (91/10.3)					
E3_0365	16512-16634 (123/64.22)	gp36.5 (40/4.4)					
E3_0367	16634-16789 (156/64.1)	gp36.7 (51/5.5)					
E3_0370	16827-17066 (240/66.25)	gp37 (79/9.1)					
E3_0380	17066-17662 (597/69.84)	gp38 (198/22.0)					
E3_0390	17662-17871 (210/66.19)	gp39 (69/8.0)					
E3_0400	17871-18263 (393/70.73)	gp40 (130/14.4)			HP *Burkholderia vietnamiensis*	YP_001119002 (0.001)	56 (30/54)
E3_0410	18260-18676 (417/65.64)	gp41 (138/15.7)					
E3_0420	18669-19067 (399/70.67)	gp42 (132/15.0)					
E3_0430	19067-19393 (327/67.58)	gp43 (108/12.2)					
E3_0440	19393-19662 (270/65.18)	gp44 (89/10.0)		2 TMDs			

Locus (strand) [b]	Coordinates [c] (size nt / %GC)	Product (size aa / kDa)	Putative Function	Domain / Motif	Closest Homologue [c]	Acc. no. (E-value $<10^{-3}$)	% Similarity [d] (overlap)
E3_0450	19655-19999 (345/65.79)	gp45 (114/13.4)		2 TMDs			
E3_0460	20096-20368 (273/67.0)	gp46 (90/10.2)					
E3_0470	20365-20730 (366/70.21)	gp47 (121/13.5)		Ogr/Delta-like			
E3_0475	20730-20876 (147/68.02)	gp47.5 (48/5.1)					
E3_0480	20873-21076 (204/66.66)	gp48 (67/7.9)					
E3_0490	21076-21333 (258/65.89)	gp49 (85/9.8)		2 TMDs			
E3_0500	21333-21623 (291/65.97)	gp50 (96/12.0)	Antidote protein	HTH domain	Plasmid maintenance system *Saccharopolyspora erythraea*	YP_001103117 (1e-22)	74 (68/92)
E3_0510	21620-21835 (216/66.66)	gp51 (71/8.1)					
E3_0520	21828-22199 (372/68.27)	gp52 (123/13.7)			HP *Actinoplanes sp*	AEV86711 (9e-14)	64 (48/75)
E3_0530	22196-22459 (264/72.34)	gp53 (87/9.8)					
E3_0540	22456-22695 (240/67.5)	gp54 (79/9.3)					
E3_0550	22695-23006 (312/69.55)	gp55 (103/12.0)			HP *Bacteroides sp.*	ZP_05761209 (7e-08)	60 (44/74)
E3_0560	23046-23270 (225/65.33)	gp56 (74/8.2)					
E3_0570	23267-23584 (318/65.72)	gp57 (105/11.5)					
E3_0580	23581-23778 (198/68.18)	gp58 (65/6.9)					
E3_0590	23775-23975 (201/66.66)	gp59 (66/7.4)					
E3_0600	23972-24316 (345/67.24)	gp60 (114/12.0)					

Locus (strand) [b]	Coordinates [c] (size nt / %GC)	Product (size aa / kDa)	Putative Function	Domain / Motif	Closest Homologue [c]	Acc. no. (E-value <10^{-3})	% Similarity [d] (overlap)
E3_0610	24313-24513 (201/64.67)	gp61 (66/7.3)					
E3_0620	24510-24896 (387/65.89)	gp62 (128/14.8)			gp82 *Mycobacterium* phage Che8	NP_817420 (1e-08)	56 (64/115)
E3_0630	24893-25126 (234/66.23)	gp63 (77/8.4)					
E3_0640	25123-25353 (231/65.8)	gp64 (76/8.4)					
E3_0650	25350-25502 (153/67.97)	gp65 (50/5.6)					
E3_0660	25499-26308 (810/68.39)	gp66 (269/29.7)	FAD-dependent thymidylate synthase	PF02511	FAD-dependent thymidylate synthase *Mycobacterium tuberculosis*	NP_217270 (6e-953)	60 (163/272)
E3_0670	26459-27037 (579/70.46)	gp67 (192/21.4)					
E3_0680	27170-28744 (1575/70.15)	gp68 (524/56.1)			gp87 *Mycobacterium* phage Myrna	YP_002224998 (1e-04)	54 (47/88)
E3_0690	28741-29160 (420/70.0)	gp69 (139/144.3)					
E3_0700	29160-32319 (3111/69.97)	gp70 (1036/113.0)	Structural		gp86 *Mycobacterium* phage Rizal	YP_002224779 (2e-16)	56 (59/107)
E3_0710	32316-32678 (363/63.36)	gp71 (120/12.8)			gp88 *Mycobacterium* phage Myrna	YP_002224999 (2e-13)	57 (65/115)
E3_0720	32680-35169 (2490/66.95)	gp72 (829/93.2)	Portal		gp89 *Mycobacterium* phage Myrna	YP_002225000 (0.0)	61 (524/866)
E3_0730	35184-35561 (378/70.1)	gp73 (125/13.3)					
E3_0740	35751-36260 (510/66.47)	gp74 (169/18.5)	2’5’ RNA ligase	PF02834	gp94 *Mycobacterium* phage Myrna	YP_002225005 (5e-29)	62 (103/167)
E3_0750	36271-36531 (261/66.28)	gp75 (86/9.7)	WhiB transcription factor	PF02467	Transcriptional regulator *Kineococcus radiotolerans*	BAJ32649 (5e-08)	58 (36/63)
E3_0760	36599-38410 (1812/69.53)	gp76 (603/66.9)	Protease associated protein	LysM	gp96 *Mycobacterium* phage Myrna	YP_002225007 (6e-30)	49 (124/255)
E3_0770	38464-41151 (2688/67.07)	gp77 (895/98.4)	Prohead protease	ZnF_C2H2	gp95 *Mycobacterium* phage Bxz1	NP_818168 (1e-73)	82 (153/188)

Locus (strand) [b]	Coordinates [c] (size nt / %GC)	Product (size aa / kDa)	Putative Function	Domain / Motif	Closest Homologue [c]	Acc. no. (E-value <10^{-3})	% Similarity [d] (overlap)
E3_0780	41180-41701 (522/68.39)	gp78 (173/18.6)	Chaperonin-like		gp96 *Mycobacterium* phage Bxz1	NP_818169 (6e-33)	63 (108/173)
E3_0790	41722-42723 (1002/64.97)	gp79 (333/37.1)	Major capsid		gp97 *Mycobacterium* phage Bxz1	NP_818170 (4e-158)	91 (82/173)
E3_0800	42859-43119 (261/71.64)	gp80 (86/8.9)					
E3_0810	43164-43466 (303/74.58)	gp81 (100/10.5)					
E3_0820	43499-43774 (276/71.01)	gp82 (91/9.9)					
E3_0830	43865-44512 (648/68.51)	gp83 (215/24.1)					
E3_0840	44580-46055 (1476/68.49)	gp84 (491/52.6)	Lipolytic protein (LysB1)	PF13472	Lipolytic *Paenibacillus sp.*	YP_003012269 (4e-11)	47 (98/212)
E3_0850(-)	46429-47322 (894/70.35)	gp85 (297/32.4)	Lipolytic protein (LysB2)		HP *Rhodococcus jostii*	YP_705817 (1e-15)	46 (126/277)
E3_0860	47441-48409 (969/69.24)	gp86 (322/34.3)	Tail fibre		Tail fibre *Rhodococcus* phage ReqiPoco6	ADD81003 (2e-93)	82 (198/242)
E3_0870	48489-49322 (834/68.34)	gp87 (277/30.0)	Structural		HP *Streptococcus* pyogenes	ZP_00366663 (8e-04)	49 (50/103)
E3_0880	49332-50129 (798/68.67)	gp88 (265/27.1)	Tail fibre	COG5301	HP *Aeromicrobium marinum*	ZP_07715597 (4e-38)	57 (152/267)
E3_0890	50133-51623 (1491/68.67)	gp89 (496/49.7)	Structural		HP *Rhodococcus equi*	ZP_06828137 (4e-29)	62 (97/158)
E3_0900	51623-52363 (741/67.47)	gp90 (246/26.5)					
E3_0905	52452-52688 (237/63.71)	gp90.5 (78/8.4)					
E3_0910	52685-53122 (438/65.52)	gp91 (145/15.9)					
E3_0920	53233-53553 (321/64.79)	gp92 (106/12.0)					
E3_0930	53631-53960 (330/65.15)	gp93 (109/12.2)			HP *Rhodococcus erythropolis*	YP_002765948 (8e-11)	58 (61/106)
E3_0940	53960-54166 (207/67.14)	gp94 (68/7.2)					

Locus (strand) [b]	Coordinates [c] (size nt / %GC)	Product (size aa / kDa)	Putative Function	Domain / Motif	Closest Homologue [c]	Acc. no. (E-value $<10^{-3}$)	% Similarity [d] (overlap)
E3_0950	54193-54597 (405/69.62)	gp95 (134/14.8)					
E3_0960	54822-55628 (807/68.4)	gp96 (268/29.3)			gp9 *Mycobacterium* phage Myrna	YP_002224927 (1e-18)	62 (68/110)
E3_0970	55687-56010 (324/68.82)	gp97 (107/12.8)			gp058 *Rhodococcus* phage ReqiDocB7	ADD80844 (3e-08)	54 (50/93)
E3_0980	56068-57180 (1113/68.64)	gp98 (370/41)	Amidase (LysA)	PF01510, PF08310	HP *Rhodococcus opacus*	YP_002781245 (6e-100)	68 (247/368)
E3_0990	57252-57626 (375/68.53)	gp99 (124/12.8)	Structural				
E3_1000	57732-58712 (981/69.82)	gp100 (326/32.8)	Structural	*Yersinia* adhesion	Haemagglutinin family protein *Cyanobium sp.*	ZP_05045923 (3e-07)	56 (57/103)
E3_1010	58723-58938 (216/71.29)	gp101 (71/7.7)					
E3_1020	(59014-602011188/69.27)	gp102 (395/43.8)			HP *Mycobacterium gilvum*	YP_001136526 (6e-28)	48 (193/408)
E3_1030	60278-60553 (276/66.66)	gp103 (91/10.0)					
E3_1040	60684-61229 (546/64.46)	gp104 (181/19.6)			gp115 *Mycobacterium* phage Myrna	YP_002225026 (3e-19)	52 (90/176)
E3_1050	61232-61795 (564/73.04)	gp105 (187/19.6)		Coiled coil			
E3_1060	61890-62816 (927/66.88)	gp106 (308/33.3)	Structural		gp115 *Mycobacterium* phage Rizal	YP_002224808 (9e-32)	59 (105/181)
E3_1070	62826-63836 (1011/68.44)	gp107 (336/37.5)	Structural		gp110 *Mycobacterium* phage ET08	YP_003347789 (2e-63)	58 (192/333)
E3_1080	63836-64429 (594/68.35)	gp108 (197/22.5)	Structural		gp118 *Mycobacterium* phage Myrna	YP_002225030 (1e-49)	66 (129/198)
E3_1090	64426-64863 (438/68.72)	gp109 (145/16.6)			gp119 *Mycobacterium* phage Myrna	YP_002225031 (6e-22)	62 (87/141)
E3_1100	64860-65612 (753/67.59)	gp110 (250/27.2)	Tail minor protein		gp121 *Mycobacterium* phage Myrna	YP_002225032 (2e-34)	56 (132/238)
E3_1110	65682..67124 (1443/67.29)	gp111 (480/50.8)	Tail sheath	PF004984	gp122 *Mycobacterium* phage Myrna	YP_002225033 (1e-133)	69 (327/478)
E3_1120	67184-67660 (477/64.57)	gp112 (158/17.6)	Tail tube		gp123 *Mycobacterium* phage Myrna	YP_002225034 (4e-59)	84 (127/152)

Locus (strand) [b]	Coordinates [c] (size nt / %GC)	Product (size aa / kDa)	Putative Function	Domain / Motif	Closest Homologue [c]	Acc. no. (E-value <10^{-3})	% Similarity [d] (overlap)
E3_1130	67669-67872 (204 /65.19)	gp113 (67/7.1)					
E3_1140	67894-68418 (525/67.42)	gp114 (174/19.2)	λ G-like protein		gp126 *Mycobacterium* phage Catera	YP_656135 (1e-26)	59 (101/173)
E3_1140/E3_1150	67894-68415, 6840-68774 (891/66.55)	gp114/Gp115 (296/33.4)	λ G/T-like		gp119 *Mycobacterium* phage ET08	YP_656134 (7e-43)	58 (165/289)
E3_1160	68771-71320 (2550/68.58)	gp116 (849/87.3)	Tape measure protein		gp129 *Mycobacterium* phage Bxz1	NP_818202 (2e-30)	43 (217/505)
E3_1170	71320-71931 (612/63.56)	gp117 (203/21.9)			gp129 *Mycobacterium* phage Myrna	YP_002225040 (3e-45)	64 (124/194)
E3_1750	71931-72089 (159/63.52)	gp117.5 (52/5.7)			gp131 *Mycobacterium* phage Bxz1	NP_818204 (1e-08)	56 (29/52)
E3_1180	72089-72718 (630/66.03)	gp118 (209/23.0)			gp131 *Mycobacterium* phage Myrna	YP_002225042 (5e-19)	55 (77/142)
E3_1190	72731-75151 (2421/64.92)	gp119 (806/88.1)	Baseplate protein P/SLT	PF01464/ PF00877	gp131 *Mycobacterium* phage Cali	YP_002224604 (6e-89)	59 (280/480)
E3_1200	75148-75987 (840/67.97)	gp120 (279/29.0)			gp133 *Mycobacterium* phage Myrna	YP_002225044 (6e-10)	46 (71/156)
E3_1210	76040-76456 (417/67.14)	gp121 (138/15.5)	Baseplate protein W		gp136 *Mycobacterium* phage Bxz1	NP_818209 (2e-31)	66 (88/134)
E3_1220	76468-78309 (1842/66.72)	gp122 (613/65.8)	Baseplate protein J	PF04865	gp135 *Mycobacterium* phage Cali	YP_002224608 (1e-154)	64 (389/609)
E3_1230	78309-79742 (1434/66.1)	gp123 (477/52.6)	Baseplate protein I		gp138 *Mycobacterium* phage LRRHood	ACU41662 (1e-138)	67 (313/474)
E3_1240	79745-83035 (3291/67.12)	gp124 (1096/120.4)	Structural (SCOP b.18.1.7)		gp138 *Mycobacterium* phage Myrna	YP_002225049 (0.0)	60 (503/839)
E3_1250	83045-83704 (660/69.24)	gp135 (219/22.8)	Structural				
E3_1260	83865-84263 (399/68.67)	gp126 132/14.4)					
E3_1270	84292-84495 (204/69.11)	gp127 (67/7.5)					
E3_1275	84518-84685 (168/66.66)	gp127.5 (55/5.8)					
E3_1280	84713-85261 (549/68.3)	gp128 (182/19.8)					

Locus (strand) [b]	Coordinates [c] (size nt / %GC)	Product (size aa / kDa)	Putative Function	Domain / Motif	Closest Homologue [c]	Acc. no. (E-value <10^{-3})	% Similarity [d] (overlap)
E3_1290	85516-86661 (1146/68.23)	gp129 (381/41.8)					
E3_1300	86722-87051 (330/70.3)	gp130 (109/12.2)					
E3_1310	87048..87383 (336/68.45)	gp131 (111/12.3)					
E3_1315	87380-87700 (321/63.55)	gp131.5 (106/12.0)					
E3_1320	87797-88015 (219/68.49)	gp132 (72/8.0)					
E3_1330	88048-88518 (471/69.0)	gp133 (156/17.4)					
E3_1340 (-)	88720-88917 (252/71.88)	gp134 (83/9.2)	Transcriptional regulator	HTH motif			
E3_1350 (-)	88914-89291 (378/67.19)	gp135 (125/14.2)	Transcriptional regulator	HTH motif			
E3_1360 (-)	89288-91237 (1950/68.36)	gp136 (649/72.9)	Helicase-like	PF00271	gp179 *Mycobacterium* phage Bxz1	NP_818230 (1e-100)	55 (323/592)
E3_1370	91571-92437 (867/68.74)	gp137 (288/31.2)	Transcriptional regulator	HTH motif			
E3_1380	92319-92711 (393/65.13)	gp138 (130/14.4)			gp177 *Mycobacterium* phage Myrna	YP_002225056 (1e-04)	50 (63/125)
E3_1390	92730-93014 (285/71.92)	gp139 (94/10.7)					
E3_1400	93115-93414 (300/65.33)	gp140 (99/10.6)					
E3_1410	93495-94199 (705/66.8)	gp141 (234/26.0)			gp188 *Mycobacterium* phage Catera	YP_656169 (1e-13)	50 (100/203)
E3_1420	94258-94563 (306/63.39)	gp142 (101/11.5)			gp181 *Mycobacterium* phage Myrna	YP_002225060 (4e-14)	63 (65/104)
E3_1430	94614-95510 (897/66.77)	gp143 (298/33.0)	5'3' exonuclease	PF02739, PF01367	DNA polymerase I *Mycobacterium tuberculosis*	ZP_03536625 (5e-29)	50 (137/279)
E3_1440	95507-95842 (336/67.55)	gp144 (111/12.0)					
E3_1450	95839-97035 (1197/70.09)	gp145 (398/43.8)	N-acetyl aminotransferase	PF00202	N-acetylornithine aminotransferase	NP_691999 (2e-17)	44 (172/391)

Locus (strand) [b]	Coordinates [c] (size nt / %GC)	Product (size aa / kDa)	Putative Function	Domain / Motif	Closest Homologue [c]	Acc. no. (E-value <10^{-3})	% Similarity [d] (overlap)
E3_1460	97074-97457 (384/63.28)	gp146 (127/14.1)			gp189 *Mycobacterium* phage Bxz1	NP_818240 (9e-09)	62 (62/101)
E3_1470	97473-97883 (411/62.74)	gp147 (136/15.8)			gp185 *Mycobacterium* phage Myrna	YP_002225064 (7e-17)	67 (72/109)
E3_1480	97870-98664 (795/66.54)	gp148 (264/29.8)	DnaC	PF01695	gp186 *Mycobacterium* phage Myrna	YP_002225065 (6e-70)	63 (171/274)
E3_1490	98673-99938 (1266/68.8)	gp149 (421/47.4)	DnaB	PF03796	gp187 *Mycobacterium* phage Myrna	YP_002225066 (3e-90)	62 (254/410)
E3_1500	99938-101044 (1107/66.93)	gp150 (368 /42.0)	DnaG	PF08275	gp188 *Mycobacterium* phage Myrna	YP_002225067 (1e-63)	55 (202/373)
E3_1505	101044-101217 (174/67.24)	gp150.5 (57/6.1)					
E3_1510	101251-101862 (612/68.79)	gp151 (203/23.3)	HNH endonuclease	PF01844	HP *Thalassomonas* phage BA3	YP_001552315 (5e-08)	52 (45/88)
E3_1520	101849-102472 (624/66.18)	gp152 (207/24.3)	DnaJ	PF00226	gp200 *Mycobacterium* phage LRRHood	ACU41695 (6e-19)	51 (99/197)
E3_1530	102565-103884 (1320/69.54)	gp153 (439/48.9)					
E3_1540	103974-107150 (3177/65.34)	gp154 (1058/119.3)	DNA polymerase IIIα	PF07733, PF02811	gp201 *Mycobacterium* phage Catera	YP_656181 (0.0)	59 (665/1133)
E3_1550	107161-108306 (1146/66.23)	gp155 (381/40.8)	Rec A	PF00154	gp205 *Mycobacterium* phage ScottMcG	YP_002224204 (6e-75)	65 (227/354)
E3_1560	108306-108656 (351/62.39)	gp156 (116/13.6)	Resolvase-like		gp195 *Mycobacterium* phage Myrna	YP_002225074 (3e-12)	56 (61/110)
E3_1570	108661-109473 (813/64.82)	gp157 (270/31.0)	RecB-like		gp204 *Mycobacterium* phage Bxz1	NP_818255 (3e-68)	66 (175/266)
E3_1580	109470-110030 (561/66.48)	gp158 (186/20.4)	Holliday junction resolvase	PF02075	gp8 *Mycobacterium* phage Phlyer	YP_002564106 (7e-24)	60 (106/178)
E3_1590	110027-110746 (720/68.05)	gp159 (239/27.5)			gp200 *Mycobacterium* phage Myrna	YP_002225079 (8e-45)	63 (142/227)
E3_1600	110761-111183 (423/61.7)	gp160 (140/16.1)	Sigma factor 70-like	PF08281	gp207 *Mycobacterium* phage Bxz1	NP_818258 (8e-22)	71 (75/107)
E3_1610	111246-111515 (270/65.92)	gp161 (89/10.1)			gp202 *Mycobacterium* phage Myrna	YP_002225081 (2e-15)	68 (55/82)
E3_1620	111532-112275 (744/69.08)	gp162 (247/27.7)			gp209 *Mycobacterium* phage Bxz1	NP_818260 (3e-19)	52 (91/176)

Locus (strand) [b]	Coordinates [c] (size nt / %GC)	Product (size aa / kDa)	Putative Function	Domain / Motif	Closest Homologue [c]	Acc. no. (E-value <10^{-3})	% Similarity [d] (overlap)
E3_1630	112389-113396 (1008/67.75)	gp163 (335/36.6)		PF06067	gp214 *Mycobacterium* phage LRRHood	ACU41709 (3e-51)	56 (190/340)
E3_1640	113508-113939 (432/64.58)	gp164 (143/16.0)			gp214 *Mycobacterium* phage Myrna	YP_002225091 (7e-07)	54 (66/124)
E3_1650	113969-114532 (564/66.13)	gp165 (187/20.3)					
E3_1660	114529-114771 (243/73.66)	gp166 (80/8.7)					
E3_1670	114768-115532 (765/68.75)	gp167 (254/27.8)	Lipolytic protein (LysB3)		HP *Rhodococcus opacus*	YP_002782668 (5e-08)	41 (101/250)
E3_1680	115529-115918 (390/65.64)	gp168 (129/14.2)			HP *Desulfatibacillum alkenivorans*	YP_002433729 (2e-10)	61 (48/79)
E3_1690	115930-116505 (576/69.44)	gp169 (191/21.1)					
E3_1700	116576-116899 (324/69.75)	gp170 (107/12.1)					
E3_1710	116991-117383 (393/67.93)	gp171 (130/14.4)		1 TMD			
E3_1715	117387-117527 (141/63.82)	gp171.5 (46/4.7)		Signal peptide 1 TMD			
E3_1720	117552-118379 (828/64.73)	gp172 (275/30.7)	Band 7	Signal peptide, PF01145	HP *Streptosporangium roseum*	YP_003337973 (5e-45)	55 (152/277)
E3_1730	118389-118583 (195/68.71)	gp173 (64/6.8)					
E3_1740	118652-118924 (273/67.39)	gp174 (90/9.5)					
E3_1750	118957-119190 (234/69.23)	gp175 (77/8.4)					
E3_1760	119183-119677 (495/67.07)	gp176 (164/18.3)		Coiled coil			
E3_1770	119674-119916 (243/67.07)	gp177 (80/8.7)	Nicotinamide mononucleotide transporter	3 TMDs	HP *Nocardia farcinica*	YP_120153 (4e-17)	74 (58/79)
E3_1780	119909-120457 (549/69.94)	gp178 (182/19.8)	NTPase	PF01503	HP *Methanogenic archaeon*	ADD92914 (2e-04)	56 (43/77)
E3_1790	120454-120771 (318/68.23)	gp179 (105/11.3)	Transcriptional regulator				

Locus (strand) [b]	Coordinates [c] (size nt / %GC)	Product (size aa / kDa)	Putative Function	Domain / Motif	Closest Homologue [c]	Acc. no. (E-value $<10^{-3}$)	% Similarity [d] (overlap)
E3_1800	120848-121060 (213/67.13)	gp180 (70/7.7)					
E3_1810	121084-121449 (366/65.84)	gp181 (121/13.2)	Transcriptional regulator	Winged helix			
E3_1820	121450-122274 (825/67.03)	gp182 (274/31.1)		1 TMD			
E3_1830	122284-122496 (213/67.6)	gp183 (70/7.8)					
E3_1840	122504-122911 (408/68.38)	gp184 (135/15.3)					
E3_1850	122955-123410 (456/67.32)	gp185 (151/17.0)					
E3_1860	123407-124150 (744/67.06)	gp186 (247/27.0)	DNA polymerase IIIε	PF00929	DNA polymerase IIIε *Rhodococcus equi*	ZP_06829322 (3e-34)	56 (136/244)
E3_1870	124301-124552 (252/66.66)	gp197 (82/9.3)					
E3_1875	124703-124864 (162 /67.9)	gp187.5 (53/6.1)					
E3_1880	124861-125292 (432/66.43)	gp188 (143/15.9)		Signal peptide			
E3_1890	125292-125735 (444/68.69)	gp189 (147/16.5)					
E3_1895	125789-125962 (174/68.39)	gp189.5 (57/6.6)					
E3_1900	126014-126658 (645/71.31)	gp190 (214/23.7)					
E3_1910	126687-127466 (780/70.38)	gp191 (259/28.6)					
E3_1920	127532-127831 (300/69.33)	gp192 (99/10.5)					
E3_1930	127828-128352 (525/66.28)	gp193 (174/20.3)			HP *Bacillus thuringiensis*	ZP_04143016 (5e-11)	59 (50/85)
E3_1940	128349-128687 (339/68.43)	gp194 (112/12.9)					
E3_1950	128687-129001 (315/64.76)	gp195 (104/12.1)					

Locus (strand) [b]	Coordinates [c] (size nt / %GC)	Product (size aa / kDa)	Putative Function	Domain / Motif	Closest Homologue [c]	Acc. no. (E-value <10^{-3})	% Similarity [d] (overlap)
E3_1960	128998-129294 (297/70.03)	gp196 (98/10.8)					
E3_1970	129294-129572 (279/69.89)	gp197 (92/10.2)					
E3_1980	129585-129890 (306/66.99)	gp198 (101/11.2)					
E3_1990	129947-130468 (522/66.47)	gp199 (173/19.8)	HNH endonuclease	PF13392	Endonuclease *Clavibacter* phage CMP1	YP_003359141 (5e-10)	50 (59/120)
E3_2000	130562-130786 (225/63.55)	gp200 (74/8.2)					
E3_2010	130852-131274 (423/69.26)	gp201 (140/16.1)					
E3_2020	131323-132225 (903/69.87)	gp202 (300/32.2)	Histone deacetylase	PF00850	Histone deacetylase *Sorangium cellulosum*	YP_001619848 (1e-31)	52 (141/275)
E3_2030	132241-132723 (483/66.87)	gp203 (160/17.7)					
E3_2040	132849-136334 (3486/68.93)	gp204 (1161/118.5)	Tail fiber protein H		gp238 *Mycobacterium* phage Spud	YP_002224457 (1e-151)	51 (607/1212)
E3_2050	136366-138942 (2577/64.68)	gp205 (858/95.6)	Structural		gp102 *Mycobacterium* phage Cali	YP_002224575 (4e-76)	71 (183/258)
E3_2060	138942-140201 (1260/68.65)	gp206 (419/47.0)	Aminotransferase		gp129 *Mycobacterium* phage Pumpkin	ACU42061 (4e-20)	61 (77/127)
E3_2070	140203-140958 (756/64.68)	gp207 (251/26.5)			gp240 *Mycobacterium* phage Myrna	YP_002225117 (2e-05)	47 (93/198)
E3_2080	140968-141756 (789/66.92)	gp208 (262/27.5)	Structural				
E3_2090	141769-142551 (783/69.47)	gp209 (260/26.9)	Structural				

[a] ORFs identified on basis of ATG, GTG or TTG start codons, 40 amino acids minimum coding capacity, and presence of probable Shine-Dalgarno sequences optimally positioned within -15 to -4 nucleotides upstream of the putative start codon. Informational noise was limited using a conservative annotation approach (Letek *et al.*, 2008).

[b] Coordinates of E3 genome according to sequence deposited under GenBank accession no. HM114277; negative strand indicated as (-).

[c] HP, hypothetical protein.

[d] Percentage amino acid similarity retrieved from BLASTp output.

Table S2. Proteomic analysis of E3 virion-associated proteins identified by LC-ESI-MS/MS [a].

Gene product	Mw (kDa)	Size (aa)	NRP [b]	% aa [b]	Putative function	Homologues (E-value $<10^{-3}$)		
						Myrna	Bxz1	Other
gp70	113.0	1036	3	9.5	Structural	gp87	gp87	Bxz1-like phages
gp72	93.2	829	20	32.1	Portal	gp89	gp89	Bxz1-like phages
gp77	98.4	895	3	5.3	Prohead protease	gp97	gp95	Bxz1-like phages
gp78	18.6	173	8	70.5	Chapronin-like protein	gp98	gp96	Bxz1-like phages
gp79	37.1	333	21	87.1	Major capsid	gp99	gp97	Bxz1-like phages
gp84	52.6	491	14	37.9	Lipolytic			Lipolytic *Paenibacillus sp.*
gp86	34.3	322	3	15.8	Tail fibre			Tail fibre proteins of *R. equi* phages ReqiPepy6 (gp004) and ReqiPoco6 (gp005)
gp87	30.0	277	3	15.9	Structural			HP [c] *Streptococcus pyogenes* prophages 10750.2 and 315.5
gp88	27.1	265	4	32.8	Tail fibre	gp111, gp239	gp112, gp232	Bxz1-like phages, HP *Aeromicrobium marinum*
gp89	49.7	496	3	10.5	Structural	gp119, gp239	gp114, gp232	Bxz1-like phages, HP *R. equi*
gp99	12.8	124	7	50.8	Structural			
gp100	32.8	326	9	48.5	Structural			Haemagglutinin protein *Cyanobium* sp.
gp106	33.3	308	5	19.5	Structural	gp117	gp117	Bxz1-like phages
gp107	37.5	336	10	36.6	Structural	gp118	gp119	Bxz1-like phages
gp108	22.5	197	4	21.8	Structural	gp118	gp120	Bxz1-like phages
gp110	27.2	250	2	12	Minor tail	gp121	gp123	Bxz1-like phages
gp111	50.8	480	23	73.5	Tail sheath	gp122	gp124	Bxz1-like phages
gp112	17.6	158	5	39.9	Tail tube	gp123	gp125	Bxz1-like phages
gp119	88.1	806	7	12.3	Baseplate protein	gp132	gp133	Bxz1-like phages
gp121	15.5	138	5	50.0	Baseplate W	gp135	gp136	Bxz1-like phages
gp122	65.8	613	15	35.3	Baseplate J	gp136	gp137	Bxz1-like phages, *Lactobacillus* phage LP65 (gp095), *Staphylococcus* phage Twort (ORF026), *Bacillus* phage SPO1 (gp14.2)
gp123	52.6	477	20	63.3	Baseplate I	gp137	gp142	Bxz1-like phages
gp124	120.4	1096	10	13.4	Structural	gp138	gp143	Bxz1-like phages
gp125	22.8	219	5	37.9	Structural			
gp204	118.5	1161	13	15.9	Tail fibre	gp239	gp232	Bxz1-like phages, *Corynebacterium* phages P1201 (gp40) and BFK20 (gp22)
gp205	95.6	858	9	18.1	Structural	gp102	gp103, gp104	Bxz1-like phages
gp208	27.5	262	2	11.5	Structural			
gp209	26.9	260	4	24.2	Structural			

[a] Data analysed in accordance with published guidelines (Taylor and Goodlett, 2005) with carbamidomethyl (C) and oxidation (M) selected as fixed and variable modifications respectively, and mass tolerance values for MS and MS/MS of 1.5 Da and 0.5 Da respectively. Molecular weight search (MOWSE) scores for individual protein identifications were inspected manually and considered significant if a) two peptides were matched for each protein, and b) each peptide contained an unbroken "*b*" or "*y*" ion series of a minimum of four amino acid residues.

[b] Number of non-redundant peptides and percentage of amino acids identified by mass spectrometry.

[c] Hypothetical protein.

Table S3. Bacterial strains used for host range analysis.

Bacterial strain	Description	Source [a]	E3 susceptibility
Rhodococcus equi [b]			
NCIMB 10027	Equine isolate, type strain	NCIMB	+
103S	Equine isolate, genome strain	Letek *et al.*, 2010	+
CV1	Equine isolate	CVS	+
CV2	Equine isolate	CVS	+
CV3	Equine isolate	CVS	+
VI1	Equine isolate	EVS	+
GV1	Equine isolate	GVS	+
GV2	Equine isolate	GVS	+
Rhodococcus erythropolis			
SQ1	Environmental isolate	Quan and Dabbs, 1993	-
NCIMB 11148	Environmental isolate, type strain	Collection	-
NCIMB 9905	Environmental isolate	NCIMB	-
NCIMB 13065	Chemical storage tank isolate	NCIMB	-
Rhodococcus rhodochrous			
NCIMB 9703	Environmental isolate	NCIMB	-
NCIMB 9160	Environmental isolate	NCIMB	-
NCIMB 1127	Environmental isolate	NCIMB	-
NCIMB 11273	Environmental isolate	NCIMB	-
NCIMB 9259	Environmental isolate	NCIMB	-
NCIMB 13259	Chemical waste isolate	NCIMB	-
Rhodococcus ruber			
NCIMB 11149	Environmental isolate	NCIMB	-
Rhodococcus opacus			
NCIMB10810	Gasworks pipe isolate, type strain	NCIMB	-
Rhodococcus fascians			
IEGM AC170		IEGM	-
ATCC 3318		ATCC	-
Mycobacterium phlei			
NCIMB 8573		NCIMB	-
Gordonia 'australis'			
A554	Environmental isolate	ENU	-

[a] NCIMB, National Collection of Industrial and Marine Bacteria, Aberdeen, UK; UKCVS, Prof Alexander & Lindsay, University of Cambridge Veterinary School; EVS, Dr Smith, University of Edinburgh Veterinary School; GVS, Dr Taylor, University of Glasgow Veterinary School; ENU, Dr Stainsby, Edinburgh Napier University. [b] Most isolates from a selection of strains from different sources and geographical origins of the global *R. equi* collection maintained in JV-B laboratory (Ocampo-Sosa et al. 2007) were susceptible.

Table S4. Software used for genome annotation.

Programs	Purpose	References or websites
Glimmer v2.0 and Prodigal v2.60	ORFs, RBSs and terminators	Delcher *et al*., 1999 Hyatt *et al*., 2010
TMHMM v2.0	Transmembrane domains	Sonnhammer *et al*., 1998
SignalP v3.0	Signal peptide	Bendtsen *et al*., 2004
tRNAscan	tRNA and tmRNA	Laslett and Canback, 2004
ARAGORN	tRNA and tmRNA	Schattner *et al*., 2005
Artemis v12.0	Manual curation and edition of annotation	Rutherford *et al*., 2000
BLASTClust	Cluster of homologue proteins	Altschul *et al*., 1990
Alien Hunter	Horizontal gene transfer (HGT)	http://www.sanger.ac.uk
EMBOSS Stretcher	Global DNA homology	http://www.ebi.ac.uk
Pfam	Functional domains and family proteins	Finn *et al*., 2008
BLASTp	Protein similarity	http://www.ncbi.nlm.nih.gov
NCBI's CDD	Conserved domain database	http://www.ncbi.nlm.nih.gov
InterProScan	Protein signature recognition	Zdobnov and Apweiler, 2001
Phyre v0.2	Protein fold recognition	Kelley and Sternberg, 2009
I-TASSER	Tertiary structure predictions	Roy *et al*., 2010
HHPred	Secondary structure and protein function predictions	Soding *et al*., 2005
ClustalX v2.0	Protein sequence alignment	Larkin *et al*., 2007
MEGA v5.0	Phylogenetic trees using Neighbor Joining (NJ) method	Tamura *et al*., 2011
PhyML v2.4.5	Phylogenetic trees using Maximum Likelihood (ML) method	Guindon and Gascuel, 2003

Supporting Information – Text

E3 products for which phage homologues could not be identified or are exceptional. The coding genes are all in HPRs and highlight the potential lateral exchanges that may occur between phage and non-virus genomes. Examples include gp100 from HPR-2 possessing a Hep_Hag domain typically found in bacterial haemagglutinins, invasins and autotransporters (Tiyawisutsri et al., 2007) but extremely rare in viruses. To date it has been found in the serum resistance immunoglobulin-binding Eib proteins encoded by three *Escherichia coli* prophages (Sandt and Hill, 2000), and in *Bacillus* phage SPO1, encoded in a locus inserted between the terminase and portal genes and containing other bacteria-related genes together with five tRNA genes (Stewart *et al*., 2009). HPR-4 encodes two proteins, gp172 and gp202, for which no phage homologues could be identified. Gp172 contains a Band 7 domain (PF01145) present in eukaryotic integral membrane proteins. Bacterial high frequency lysogenisation proteins also belong to this family, of which HflC has been implicated in temperate phage λ lysogenisation decision making in *E. coli* (Herman *et al*., 1993). Gp202 contains a histone deacetylase domain (PF00850), implicated in stabilising the interaction of histone-like proteins with DNA (Leipe and Landsman 1997). To our knowledge, E3 gp202 is the first histone deacetylase-like protein to be reported in a phage, where it may play a role in regulated host-phage interaction.

Supporting Information – References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic alignment search tool. *JMol Biol* **215**: 403-410.

Bendtsen, J.D., Nielsen, H., von Heijne, G., and Brunak, S. (2004) Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* **340**: 783-795.

Delcher, A.L., Harmon, D., Kasif, S., White, O., and Salzberg, S.L. (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* **27**: 4636-4641.

Finn, R.D., Tate, J., Mistry, J., Coggill, P.C., Sammut, S.J., Hotz, H.R. *et al*. (2008) The Pfam protein families database. *Nucleic Acids Res* **36**: D281-288.

Glazko, G., Makarenkov, V., Liu, J., and Mushegian, A. (2007) Evolutionary history of bacteriophages with double-stranded DNA genomes. *Biol Direct* **2**: 36.

Herman, C., Ogura, T., Tomoyasu, T., Hiraga, S., Akiyama, Y., Ito, K., *et al.* (1993) Cell growth and lambda phage development controlled by the same essential *Escherichia coli* gene, ftsH/hflB. *PNAS* **90**: 10861-10865.

Hyatt, D., Chen, G.-L., LoCascio, P., Land, M., Larimer, F., and Hauser, L. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**: 119.

Kelley, L.A., and Sternberg, M.J. (2009) Protein structure prediction on the Web: a case study using the Phyre server. *Nat Protoc* **4**: 363-371.

Kondou, Y., Kitazawa, D., Takeda, S., Tsuchiya, Y., Yamashita, E., Mizuguchi, M., *et al.* (2005) Structure of the central hub of Bacteriophage Mu baseplate determined by X-ray crystallography of gp44. *J Mol Biol* **352**: 976-985.

Laslett, D., and Canback, B. (2004) ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res* **32**: 11-16.

Letek, M., Ocampo-Sosa, A.A., Sanders, M., Fogarty, U., Buckley, T., Leadon, D.P., *et al.*, (2008) Evolution of the *Rhodococcus equi vap* pathogenicity island seen through comparison of host-associated vapA and vapB virulence plasmids. *J Bacteriol* **190**: 5797-5805.

Leipe, D.D., and Landsman, D. (1997) Histone deacetylases, acetoin utilization proteins and acetylpolyamine amidohydrolases are members of an ancient protein superfamily. *Nucleic Acids Res* **25**: 3693-3697.

Quan S, and Dabbs E.R. (1993) Nocardioform arsenic resistance plasmid characterization and improved *Rhodococcus* cloning vectors. Plasmid **29**:74-79.

Rohwer, F., and Edwards, R. (2002) The Phage Proteomic Tree: a genome-based taxonomy for phage. *J Bacteriol* **184**: 4529-4535.

Roy, A., Kucukural, A., and Zhang, Y. (2010) I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc* **5**: 725-738.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.A., and Barrell, B. (2000) Artemis: sequence visualization and annotation. *Bioinformatics* **16**: 944-945.

Sandt, C.H., and Hill, C.W. (2000) Four different genes responsible for nonimmune immunoglobulin-binding activities within a single strain of Escherichia coli. *Infect Immun* **68**: 2205-2214.

Schattner, P., Brooks, A.N., and Lowe, T.M. (2005) The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res* **33**: W686-689.

Söding, J., Biegert, A., and Lupas, A.N. (2005) The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res* **33**: W244-W248.

Sonnhammer, E.L., von Heijne, G., and Krogh, A. (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc Int Conf Intell Syst Mol Biol* **6**: 175-182.

Stewart, C.R., Casjens, S.R., Cresawn, S.G., Houtz, J.M., Smith, A.L., Ford, M.E., *et al*. (2009) The genome of *Bacillus subtilis* bacteriophage SPO1. *J Mol Biol* **388**: 48-70.

Taylor, G.K., and Goodlett, D.R. (2005) Rules governing protein identification by mass spectrometry. *Rapid Commun Mass Spectrom* **19**: 3420.

Tiyawisutsri, R., Holden, M.T., Tumapa, S., Rengpipat, S., Clarke, S.R., Foster, S.J., *et al*. (2007) *Burkholderia* Hep_Hag autotransporter (BuHA) proteins elicit a strong antibody response during experimental glanders but not human melioidosis. *BMC Microbiol* **7**: 19.

Zdobnov, E,M, and Apweiler, R. (2001) InterProScan - an integration platform for thesignature-recognition methods in InterPro. *Bioinformatics* 17: 847-848.