doi:10.1016/j.jmb.2005.08.075 J. Mol. Biol. (2005) 354, 536–545

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

COMMUNICATION

Genome Comparison of *Pseudomonas aeruginosa* Large Phages

Kirsten Hertveldt[1]*, Rob Lavigne[1], Elena Pleteneva[2], Natalia Sernova[3]† Lidia Kurochkina[3], Roman Korchevskii[3], Johan Robben[1] Vadim Mesyanzhinov[3], Victor N. Krylov[2] and Guido Volckaert[1]

[1]*Laboratory of Gene Technology Katholieke Universiteit Leuven Kasteelpark Arenberg 21 B-3001 Leuven, Belgium*

[2]*State Institute for Genetics and Selection of Industrial Microorganisms, 1st Dorozhnii proezd 1, Moscow 1173545 Russia*

[3]*Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Miklukho-Maklaya Street 16/10, Moscow 117997 Russia*

**Corresponding author*

Pseudomonas aeruginosa phage EL is a dsDNA phage related to the giant φKZ-like Myoviridae. The EL genome sequence comprises 211,215 bp and has 201 predicted open reading frames (ORFs). The EL genome does not share DNA sequence homology with other viruses and micro-organisms sequenced to date. However, one-third of the predicted EL gene products (gps) shares similarity (Blast alignments of 17–55% amino acid identity) with φKZ proteins. Comparative EL and φKZ genomics reveals that these giant phages are an example of substantially diverged genetic mosaics. Based on the position of similar EL and φKZ predicted gene products, five genome regions can be delineated in EL, four of which are relatively conserved between EL and φKZ. Region IV, a 17.7 kb genome region with 28 predicted ORFs, is unique to EL. Fourteen EL ORFs have been assigned a putative function based on protein similarity. Assigned proteins are involved in DNA replication and nucleotide metabolism (NAD^+-dependent DNA ligase, ribonuclease HI, helicase, thymidylate kinase), host lysis and particle structure. EL-gp146 is the first chaperonin GroEL sequence identified in a viral genome. Besides a putative transposase, EL harbours predicted mobile endonucleases related to H–N–H and LAGLIDADG homing endonucleases associated with group I intron and intein intervening sequences.

Keywords: bacteriophage genomics; phage evolution; phage EL; phage φKZ; homing endonuclease

The dsDNA-tailed bacteriophages represent numerically the largest, most widespread and probably the oldest group of bacterial viruses.[1,2] They constitute the order of the Caudovirales and comprise three families: Myoviridae, Siphoviridae and Podoviridae. Myoviridae phages have a contractile tail sheath and they represent about 25% of tailed phages.[2] Bacteriophage T4 of *Escherichia coli* is the best-characterized representative of the Myoviridae family.[3,4] The T4-like phages infect a broad range of different bacteria and can be subdivided into four groups that are increasingly distant from phage T4: the T-evens, the pseudo-T-evens, the schizo-T-evens and the exo-T-evens.[5] The phage T4-like genomes appear to be mosaics containing a large and fixed group of essential genes and a variable set of non-essential genes. The most conserved genes are involved in virion morphology and DNA replication. The non-essential genes are probably important for the adaptation of these phages to their particular lifestyle.[5]

φKZ is a giant Myoviridae phage that efficiently infects many *Pseudomonas aeruginosa* strains.[6] φKZ has a linear, circularly permuted and terminally redundant, A+T-rich (63.2%) genome and a circular genomic map. DNA sequence analysis of the φKZ genome revealed a 280,334 bp sequence

Present address: S. Natalia, Department of Microbiology and Molecular Medecine, University of Geneva, Rue Michel-Servet 1, 1211, Geneva 4, Switzerland.

Abbreviations used: ORF, open reading frame; gps, gene products; NUMOD, nuclease-associated modular DNA-binding domain; CDD, Conserved Domain Database.

E-mail address of the corresponding author: kirsten.hertveldt@biw.kuleuven.be

with 306 predicted ORFs organized into clusters.[7] In all, 10% of the predicted ϕKZ proteins exhibit similarity (Blastp > 111 bits) to proteins of known function from diverse organisms. Most of the conserved gene products are involved in nucleotide metabolism. Neither a viral DNA polymerase nor ssDNA-binding protein were identified based on amino acid similarity, suggesting a novel or host-dependent replication machinery. Limited similarity at the protein level to other Myoviridae phage, including phage T4, classifies ϕKZ as an evolutionary distinctive branch within this family. Other giant phages of the ϕKZ-type (e.g. Lin21, NN, PTB80, EL and RU) that infect *P. aeruginosa* have been isolated.[8] Restriction analysis and DNA hybridization showed a profound difference in genomic background among the ϕKZ-like phages, grouping them into three species: Lin68, ϕKZ (a.o. ϕKZ, Lin21, NN and PTB80) and EL (EL and RU).[9] Phages belonging to the same species are closely related and share significant DNA similarity, whereas ϕKZ and EL phages lack similarity at the DNA level.

ϕKZ-like phages may be valuable in phage therapy or may provide antimicrobial proteins that help to combat *Pseudomonas* infections,[9] since this bacterium has emerged as a major problematic opportunistic human pathogen due to its resistance to antibiotics.

Bacteriophage EL was isolated from natural sources[8] and grows to small irregular plaques on *P. aeruginosa* PAO1. Phage particles are composed of a large icosahedral capsid of ~140 nm in diameter and a long contractile tail of ~200 nm (Figure 1). They appear morphologically similar to ϕKZ phage particles, of which the three-dimensional head structure was determined recently by cryo-electron microscopy.[10] Overall phenotypic characteristics of the ϕKZ-like phages and significant differences observed between both ϕKZ and EL triggered genome sequencing of EL to allow a more in-depth comparison of both giant phages. DNA sequence analysis of phages from both species provides information on their evolutionary relationships, and insight into their genome structure and conserved proteins.

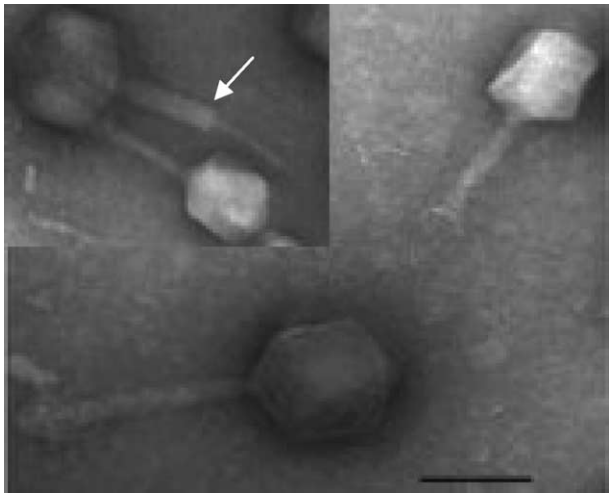

Figure 1. Electron micrograph of the large bacteriophage EL and the smaller phage T4 particles negatively stained (0.75% (w/v) uranyl formate, pH 4.5). The EL tail is 200 nm long and the icosahedral capsid is 140 nm in diameter. An EL phage with a contracted tail (indicated with an arrow) is shown in the upper inset. The scale bar represents 100 nm. Phages were propagated and purified as described.[7]

Genome sequence analysis

A collection of 1525 sequence reads assembled into a circular DNA sequence of 211,215 bp (Figure 2). The average G+C content of the EL genome is 49.3%, which is significantly higher than the average G+C content of ϕKZ (36.8%), but lower than the G+C content of the host genome, which is 65%. *In silico* restriction analysis revealed that PstI, NotI and SpeI do not cut the EL genome sequence, while NotI, PstI, SacI, SmaI, XhoI and XmaIII recognition sites are absent from the ϕKZ genome. The elimination of restriction recognition sites in EL compared to ϕKZ is probably associated with the presence of host endonucleases and adaptation to different natural hosts.[11]

To obtain an objective prediction of putative phage regulatory sequences, the entire genome sequence was scanned using PHIRE (v 1.00).[12] This analysis revealed a conserved consensus sequence (WTTTYAAACCTACATTATY) preceding an ORF at nine locations throughout the genome. These sequences (all in clockwise direction) are putative phage EL promoters. Analysis of the intergenic regions by the EMBOSS program "einverted"[13,14] revealed 46 stem–loop sequences, between 8 bp and 22 bp long, followed by an A+U-rich sequence. Many of these putative rho-independent terminator sequences are located at the end of transcriptional blocks. PHIRE analysis did not reveal any conserved stem–loop sequences in the EL genome sequence, a hallmark feature for ϕKZ.[7,12] The positions of putative promotor and rho-independent terminator sequences are indicated in Figure 2.

A total of 201 ORFs of at least 150 bp length were predicted (Figure 2). The majority of the ORFs (85.6%) are oriented clockwise on the genomic map. The genome is tightly organized (92.4% coding sequences) with small intergenic sequences (from 1 bp to 650 bp) and minor overlapping ORFs (48 ORFs have an overlap between 4 bp and 56 bp). All coding triplets are used. In line with the high G+C content of its predicted ORFs (66.6%), *P. aeruginosa* makes greater use of codons with G or C in the third position, whereas EL (49.3% G+C) also significantly uses codons with U or A in the third position. In general, codon preference is less marked in EL. Though the G+C content of EL differs from that of *P. aeruginosa*, only a single tRNA gene for the Thr codon ACG is predicted in the EL genome, while other phages like ϕKZ,[7] phage T4 and KVP40[15]

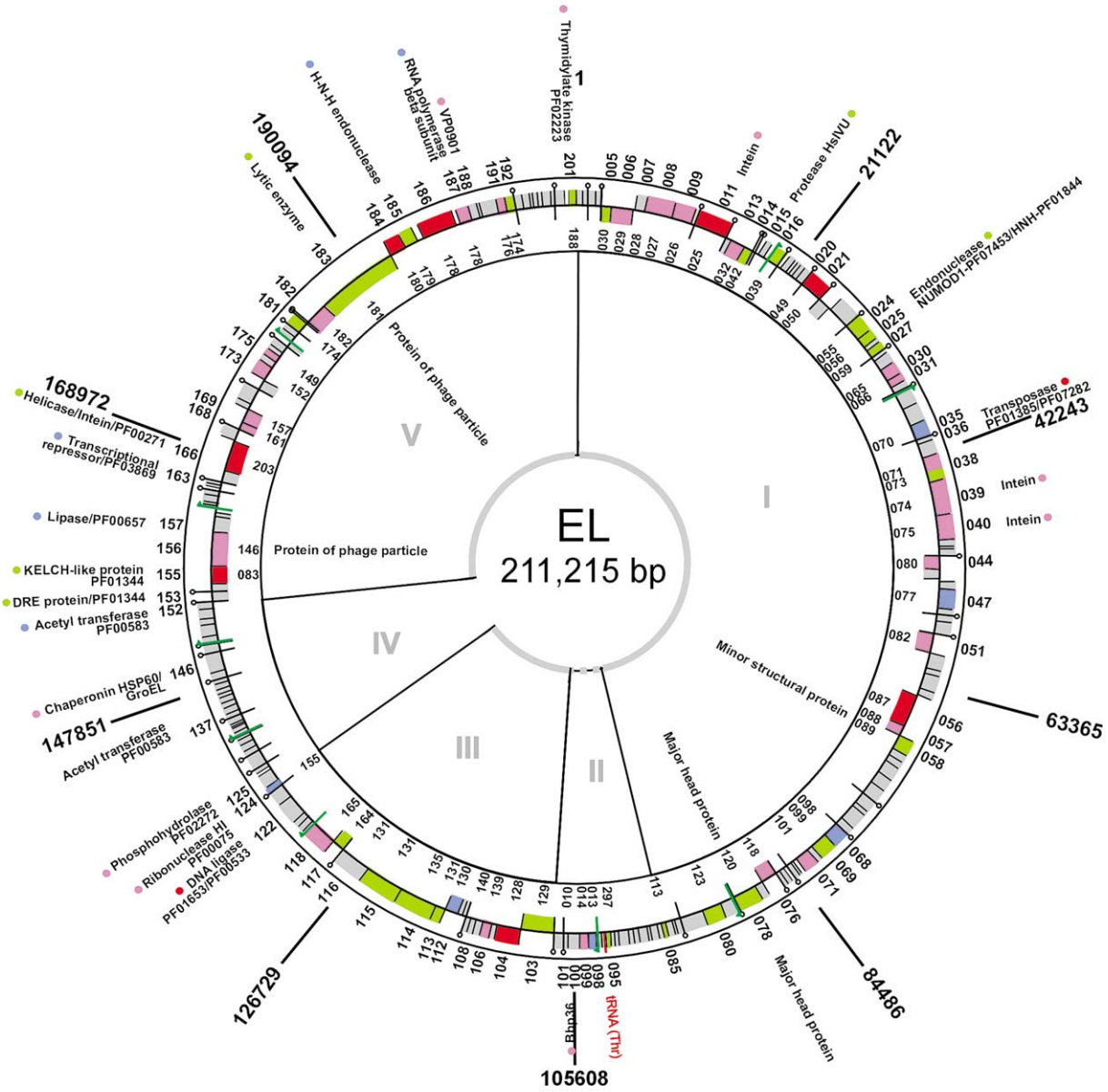

Figure 2. Circular representation of the EL genome map. The outer numbered circle represents the EL ORFs, which are numbered clockwise. EL gene products sharing similarity with ϕKZ gene products (indicated in the inner circle) are coloured according to the level of similarity based on bit scores (grey (<40), blue (40–50), green (50–80), purple (80–200) and red (>200)). Pfam[40]-hits and similarity to other proteins (and putative functions) are mentioned in the outer circle. Level of similarity is indicated by coloured dots based on the bit scores. Putative promoters are indicated as green flags, stem–loop structures (putative terminators) as black lollipops. The grey circle segments and numbers indicate the major genome regions delineated in EL based on relatedness to ϕKZ. The broken grey line indicates a genome region with protein similarity to ϕKZ gene products but with a different genomic localization. Genome sequencing was performed as described[7] with minor modifications. A total of 256 clones were sequenced. Assembly was done by Sequencher version 4.1 software (Gene Codes Corporation) into 63 contigs. The contigs were extended by primer walking with 16-mer oligonucleotide primers on pure EL DNA until all contigs assembled into a single sequence. A total of 1525 sequence reads, with length varying from 400 to 800 nucleotides yielded over 780 kb, resulting in almost fourfold redundancy for the entire genome. Putative ORFs (≥150 bp) were predicted with GeneMarkS[41] and Glimmer 2.0 (probed on the *P. aeruginosa* genome)[42] programs. Additional information based on codon usage, overlapping genes and presence of a Shine–Dalgarno ribosome-binding sequence in front of the initiation codon (AUG, GUG or UUG) were taken into account in the final ORF prediction. The STORM program[43] was used to combine protein analyses of BLAST,[44] Pfam[40] and ProtParam (part of the ExPasy suite[45]) on the batch file of protein sequences. Additionally, conserved domains were searched for in the Conserved Domain Database (CDD).[46] tRNA genes were searched by the tRNAscan-SE program[47] and FAS-tRNA.[48]

provide six, eight and 30 tRNAs, respectively. The relative frequency of the ACG codon for threonine is higher in EL than it is in the host genome, but this observation is also made with other sets of codons. The ϕKZ tRNAs are localized in a 6700 bp region between ϕKZ-gp293 and ϕKZ-gp302, a region largely absent from EL.

Comparative analysis of EL and ϕKZ open reading frames

While EL and ϕKZ lack similarity at the DNA level, one-third of the predicted EL proteins show similarity (Blastp) to ϕKZ proteins. The EL genome roughly encodes proteins similar to ϕKZ-gp10 to ϕKZ-gp203. With the exception of EL-gp95 (similar to ϕKZ-gp297), EL is devoid of gene products similar to ϕKZ-gp204-306, corresponding to 24% of the larger ϕKZ genome. The extent of similarity and the genomic position of similar proteins in both genomes allows insight into their evolutionary relatedness. Based on the position of similar proteins, five major EL genome regions can be delineated (Figure 3(a)). Syntenic genome segments are present in ORF regions I (EL-gp1-85), III (EL-gp103-124) and V (EL-gp153-201), encoding gene products similar to the ϕKZ ORF regions ϕKZ-gp25-123, ϕKZ-gp129-165 and ϕKZ-gp146-203, respectively. Within blocks of similar ϕKZ and EL ORFs, gene order can be identical or reversed, while additional insertions/deletions are observed (Figure 3(b)). A region of consecutive related EL and ϕKZ gene products, interrupted by putative mobile elements EL-gp37 transposase and ϕKZ-gp72 H–N–H endonuclease is shown. Although not all EL gene products indicated in Figure 3(b) share amino acid similarity with their ϕKZ counterparts, ORF position and size suggest a related origin. EL-gp39 and EL-gp40 harbour intein sequences (discussed below). The order of genes encoding similar proteins could be reversed, as is the case for EL-gp5-gp11, compared to ϕKZ-gp30-gp25. EL-gp113 (247 residues), gp114 (1184 residues) and gp115 (1193 residues) share notable similarity and are probably the result of gene duplications.

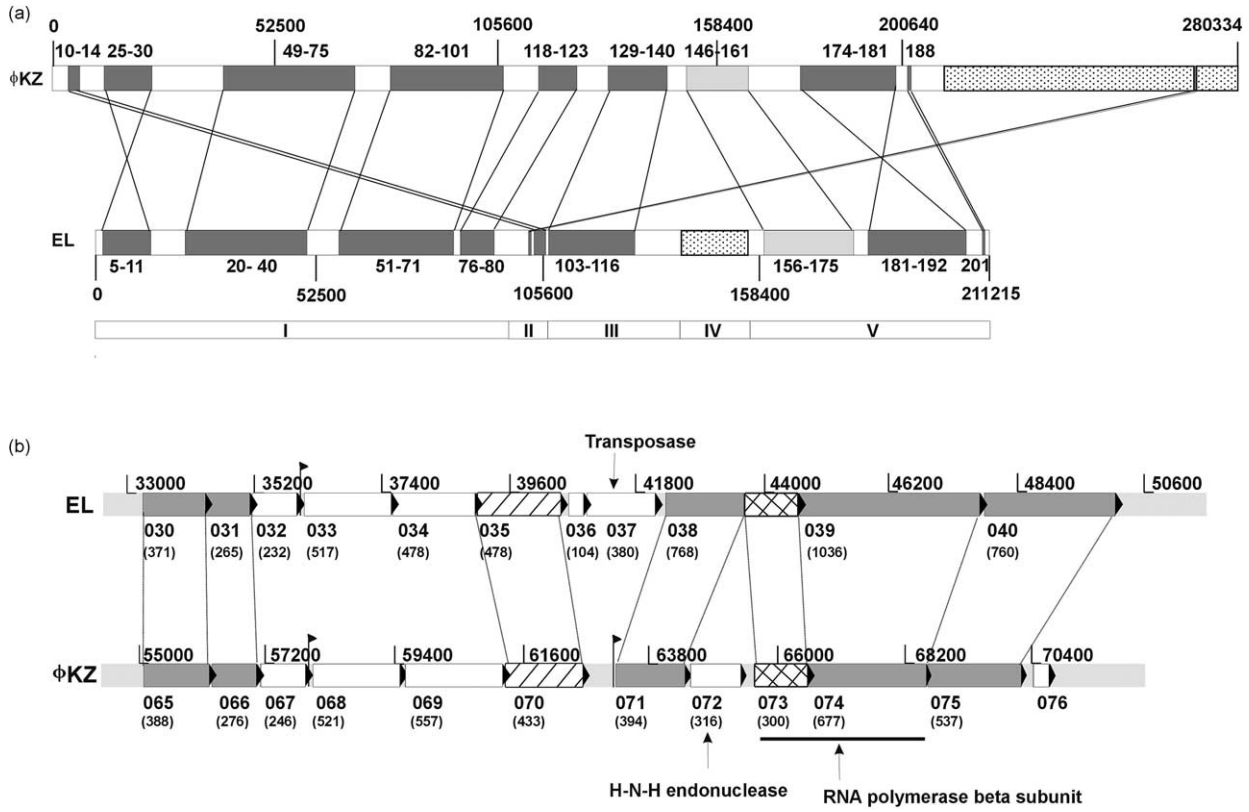

Figure 3. (a) Schematic comparison of the EL and ϕKZ genomes. Regions with blocks of proteins sharing amino acid similarity are indicated in dark grey. The genome regions marked in light grey contain genes of which the ϕKZ and EL gene products share amino acid similarity, but with extensively reshuffled gene order between both phages. Genome regions marked as dotted boxes are unique to either phage. Based on the occurrence of similar gene products, five different genome regions in EL are marked. (b) Schematic alignment of corresponding EL and ϕKZ genome regions encoding EL-gp30-gp40. Gene products sharing similarity are marked, based on bit score values (open boxes (40), hatched (40–50), crosshatched (50–80), grey (80–200)). A transposase sequence is located between EL-gp35 and EL-gp38, and absent between the ϕKZ counterparts, ϕKZ-gp70 and ϕKZ-gp71. The EL-gp38 amino acid sequence is a fusion of ϕKZ-gp71 and ϕKZ-gp73 protein sequences. No ϕKZ-gp72 counterpart, encoding a putative H–N–H endonuclease is present in this EL genome region. ϕKZ-gp73 en ϕKZ-gp74 share similarity with microbial RNA polymerases.

ORF region IV (EL-gp125-152) is unique to EL and encodes gene products related to bacterial proteins, e.g. NAD^+-dependent DNA ligase. ORF region II (EL-gp86-102) encodes four gene products displaying sequence similarity to ϕKZ-gp10, gp13, gp14, gp297, located at a different position on the ϕKZ genome map.

Structural and lytic proteins

Previous analysis of the N-terminal protein sequence after SDS-PAGE separation of proteins of the phage particle identified EL-gp78 as the major capsid protein.[8] Comparison of the predicted EL-gp78 protein sequence and the N-terminal sequence (Gly-Phe-Ser-Met-Gln-Asp-Phe) determined by Edman degradation[8] indicates cleavage of the first 120 N-terminal residues in the final coat protein. In mature ϕKZ particles, the major coat protein gp120 lacks the first 163 amino acid residues.[7] Protein similarity between EL-gp78 and ϕKZ-gp120 is confined to the mature protein region (Blastp alignment) and shows 19% identical residues, which is low compared to the level of homology between major virion proteins of T4-like phages.[16] Examination by electron microscopy, however, does not suggest differences in head structure between ϕKZ and EL. Moreover, recent determination of the three-dimensional structure of the ϕKZ head by cryo-electron microscopy[10] shows that the shape and size of the hexamers is similar to the hexameric building blocks of the bacteriophages T4, ϕ29, P22 and HK97. Post-translational proteolysis in transition from the procapsid to the mature capsid is essential for the morphogenesis of many bacteriophages and has been studied extensively for bacteriophage T4.[17] It is interesting to note that both EL and ϕKZ major protein precursors are cleaved after a Glu residue, which is also the conserved residue in the proteolytic cleavage site of phage T4 head proteins.[17] The only EL gene product with a predicted protease activity is EL-gp15 (182 residues), related to the peptidase subunit of ATP-dependent protease HslVU (ClpYQ) family. According to a recent classification,[18] members of this family have thus far not been associated with prohead maturation proteases.

Besides protein similarity of the major head proteins, EL proteins gp58 (364 residues), gp156 (1046 residues) and gp183 (2543 residues) are similar to the ϕKZ structural proteins gp89 (387 residues), gp146 (1093 residues) and gp181 (2237 residues), respectively (Figure 2). No ϕKZ-gp145 homolog, the second predominant protein in the ϕKZ particle is present in EL. The mentioned ϕKZ proteins were identified as structural proteins after SDS-PAGE separation of phage particle proteins and N-terminal sequencing.[7] EL-gp156 (1046 residues) and its ϕKZ counterpart, ϕKZ-gp146 (1093 residues) contain collagen-type sequence repeats $(Gly\text{-}X\text{-}Y)_n$ which have been reported in genes encoding structural components of bacteriophage virions.[19] The most conserved region (~700–815) in gp156 contains a $(Gly\text{-}X\text{-}Y)_{5+9}$ repeat and shares similarity with amino acid sequences in putative tail fiber proteins and proteins involved in host recognition of a.o. bacteriophage 933W[20] and *Streptococcus thermophilus* bacteriophage DT2.[21]

Protein similarity between EL-gp183 (2543 residues) and ϕKZ structural protein gp181 (2237 residues) is confined to residues 447–1907 of EL and residues 419–1777 of ϕKZ. Both proteins contain a lytic domain responsible for local degradation of the peptidoglycan layer during infection. Activity of both lytic domains, which lack amino acid similarity, was demonstrated experimentally (Y. Briers, personal communication). The presence of unrelated lytic domains in ϕKZ-gp181 and EL-gp183 provides evidence for domain swapping in the evolution of these structural components. Lytic activity was also demonstrated for EL-gp188 (Y. Briers, personal communication). Based on its activity, size and the presence of a peptidoglycan (PG)-binding-1 (Pfam PF01471) and transglycosylase SLT domain (Pfam PF01464), EL-gp188 is assigned as an endolysin, involved in phage-induced degradation of the cell envelope upon release from the host. In contrast to the ϕKZ lytic proteins, EL lytic domains (in gp183 and gp188) do not share sequence similarity. No putative holin gene meeting the characteristics presented by Wang *et al*.[22] could be identified directly up- or down stream of both the ϕKZ and the EL endolysin gene. Possibly, the holin gene has a separate genomic locus as with the lysis components of phage T4.[22] Based on their size and the presence of at least one TM domain, several EL gene products are candidate holins.

Based on *in silico* similarity searches, four additional EL proteins (ϕEL-gp113, gp114, gp115 and gp117) are likely structural proteins. The C-terminal region of EL-gp117 shows amino acid sequence homology to the C-terminal region of predicted baseplate (assembly) proteins and to the N-terminal region of distal tail fiber proteins of various phages (and bacteria). EL-gp113, gp114 and gp115 are related to ϕKZ-gp131, which contains an amino acid sequence region homologous to *Salmonella* phage Felix01 and *Streptococcus pneumoniae* phage EJ-1 putative tail fiber proteins.

Gene products involved in nucleotide metabolism

Several viruses, among which phage T4 and the related Vibriophage KVP40,[15] duplicate several host pyrimidine biosynthesis functions. This duplication appears to enhance growth as judged by mutagenesis experiments.[23] With the exception of EL-gp201, a putative thymidylate kinase (EC 2.7.4.9, ATP:dTMP phosphotransferase), no such enzymes are predicted in the EL genome. Thus, phage EL appears more dependent on host enzymes for dNTP synthesis compared to ϕKZ, which possesses additional genes

for dihydrofolate reductase, deoxycytidine triphosphate deaminase, thymidylate synthase and ribonucleoside-diphosphate reductase α and β subunits.[7] Most of these ϕKZ genes are located in the ϕKZ-gp204-306 genome region, absent from the EL genome (Figure 3(a)).

Gene products involved in DNA replication and transcription

Few EL proteins display similarity to proteins involved in DNA replication, recombination and repair. EL-gp166 (940 residues) shares a C-terminal helicase_C domain with *P. aeruginosa* DNA helicase and is related to the ϕKZ-gp203 putative DNA helicase. In contrast to the large ϕKZ-gp155 (482 residues) putative RNase HI, EL-gp124 (163 residues) is strongly related to *P. aeruginosa* PAO1 ribonuclease HI (148 residues). Based on amino acid sequence similarity there is no evidence for the presence of a phage-encoded DNA polymerase of the known DNA polymerase families. Virulent phages usually possess their own DNA polymerase. Failure to identify a DNA polymerase in both ϕKZ and EL giant phages may imply either that yet another polymerase family exists or that the phages rely on the host replication machinery.

The EL genome encodes proteins sharing amino acid similarity with ϕKZ-gp73, gp74 and gp178, similar to microbial RNA polymerases. Strikingly, the genome regions of EL-gp38/gp39 and their ϕKZ-gp73/gp74 counterparts are somehow reshuffled and interrupted by different putative mobile elements (Figure 3(b)). ϕKZ-gp178 (1451 residues) occurs as two smaller proteins gp186 (1067 residues) and gp187 (357 residues) in EL.

Homing endonucleases

Besides a putative transposase, EL harbours putative homing endonuclease sequences that are commonly associated with phage genomes. At least 15 phage T4 genes belong to two of the four structural families of homing endonucleases

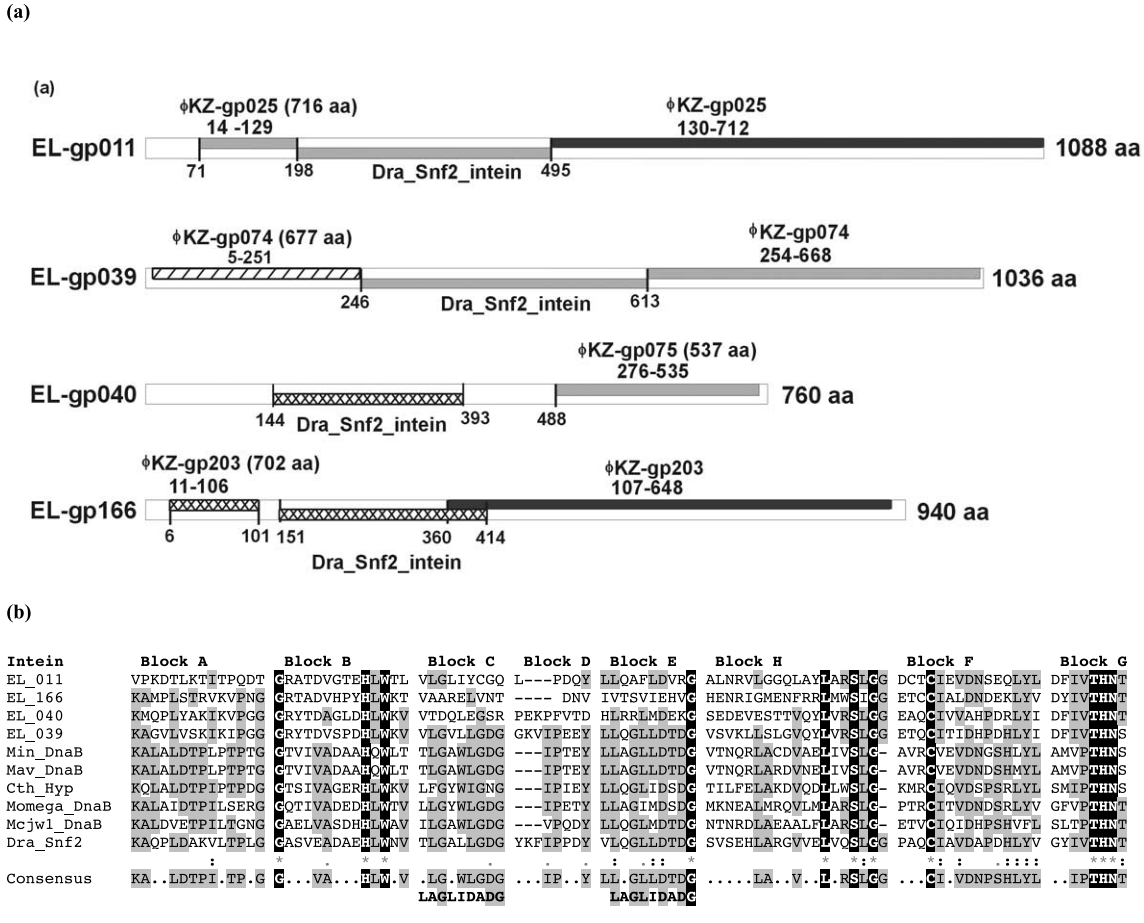

Figure 4. Intein sequences in phage EL. (a) Schematic representation of the presence of intein sequences in EL ORFs. Amino acid similarity to the Snf2 intein sequence and between ϕKZ and EL sequences is marked based on bit score values (hatched (40–50), crosshatched (50–80), grey (80–200), black (>200 bits)). (b) Local alignment of the EL intein sequences (−1 and +1 amino acid residues are included) and inteins in *Mycobacterium intracellulare* (Min), *Mycobacterium avium* (Mav), *Clostridium thermocellum* (Cth), Mycobacteriophage Omega (Momega), Mycobacteriophage CJW1 (Mcjw1), *Deinococcus radiodurans* (Dra). The consensus line represents conserved residues (marked in grey) present in at least five of the ten sequences included in the multiple alignment. Residues conserved in the ten sequences are highlighted in the alignment and consensus. Different intein blocks and LAGLIDADG motifs are indicated as described by Perler *et al*.[36]

defined by the conserved sequences GIY-YIG, LAGLIDADG, H–N–H and His-Cys.[3] Six predicted proteins, EL-gp11, gp25, gp39, gp40, gp166 and gp185, contain sequences related to H–N–H and LAGLIDADG homing endonucleases associated with group I intron and intein intervening sequences (IVS).[24] EL-gp25 and EL-gp185 share the H–N–H motif with ϕKZ proteins gp56, gp72, gp179 and gp269 and other putative H–N–H homing endonucleases in various bacteria and phages. Based on protein similarity, two putative DNA-binding domains can be delineated in ϕEL-gp25. These nuclease-associated modular DNA-binding domains (NUMODs) are structurally independent domains and appear in single or tandem repeats in homing endonucleases.[25] The first NUMOD1 domain in gp25 contains the conserved Asp, Phe and Ala residues.

An internal region in EL-gp11, EL-gp39, EL-gp40 and EL-gp166 aligns with the *Deinococcus radiodurans* Snf2 Ala-type intein sequence, containing two conserved LAGLIDADG homing endonuclease motifs.[24] In EL-gp11, EL-gp39 and EL-gp166, the intein-like sequences clearly interrupt proteins similar to the corresponding ϕKZ proteins gp25, gp74 and gp203 lacking the intein (Figure 4(a)). Increased size of the EL ORFs compared to the respective ϕKZ ORFs is in line with the presence of the intein-like intervening sequence. Several conserved intein motifs (blocks A–H) can be delineated (Figure 4(b)). Smaller EL-gp40 and EL-gp166 intein-like sequences are less conserved in blocks C, D and E, the blocks that are associated with endonuclease activity. The conserved dipeptide His-Asn is present at the C terminus of the putative intein sequences, followed by a Ser or Thr residue. These intein-like sequences could still be mobile and functionally excised or may represent intein remnants. These types of intervening sequences are most frequently located in enzymes that interact with DNA, although they also occur in metabolic enzymes.[26] The LAGLIDADG-type endonucleases in putative EL-gp39 and EL-gp166 inteins interrupt sequences similar to ϕKZ proteins with putative RNA polymerase (ϕKZ-gp74) and helicase (ϕKZ-gp203) function, respectively.

ORFs unique to EL

Ten non-ϕKZ-related gene products (gp15, gp37, gp100, gp122, gp125, gp146, gp152, gp157, gp163, gp188) exhibit similarity (Blastp, bit score >40) to (putative) proteins from bacterial or phage origin. Only one gene product (gp126) exhibits low similarity to a phage T4 gene product (putative 57B protein, Swiss-Prot entry P04533) of unknown function. Pfam and CDD searches for conserved domains provide strong evidence for the presence of transposase, NAD^+-dependent DNA ligase, and GroEL chaperonin activities in gp37, gp122 and gp146, respectively. Other, however less conserved, motifs and protein domains suggest involvement of defined gene products: (1) in nucleic acid binding (DHHA1 in gp125, Arc in gp163, NUMOD1 in gp25); (2) in protein–protein interaction (Kelch in gp153 and gp155, BRCT in gp122); (3) in ATP-binding (Helicase_C in gp166) and (4) in various enzymatic reactions (e.g. acetyltransferase in gp137 and gp152, Lipase_GDSL in gp158).

EL-gp122 is strongly related to the bacterial NAD^+-dependent DNA ligases and not to the typical phage ATP-dependent enzymes. It contains the conserved catalytic Lys114 residue that covalently binds the adenylate group from NAD^+ in the KXDG motif.[27] The EL-gp122 C-terminal region partially aligns with a C-terminal breast cancer suppressor protein, carboxy-terminal (BRCT) type II domain, which is found in many DNA damage repair and cell, cycle checkpoint proteins.[28] One of the few other known viral NAD^+-dependent DNA ligases is coded by phage T5 and is built up from two distinct subunits, A and B, both sharing similarity with corresponding EL-gp122 protein domains.

EL-gp146 (558 residues) is the first chaperonin GroEL ortholog identified in a phage genome. GroEL (Hsp60) proteins together with their cochaperonin GroES constitute a chaperone machine essential for the correct folding of many proteins. The mechanism of chaperonin-mediated polypeptide folding involves the interaction of GroEL with non-native polypeptide, subsequent GroES association, ATP hydrolysis, and release of the native protein. Recent analysis of chaperone networks in bacteria showed that GroEL is the only chaperone that can be encountered in all bacterial species under study[29] (except Mycoplasma). It has been shown that host GroEL is the only host chaperone that is known to be required for phage T4 growth.[30] In phage T4 and phage RB49, GroES function is substituted by phage-encoded cochaperonins gp31[31] and CocO,[32] respectively. Gp31 and CocO are functional orthologs (34% identical), but lack similarity to host GroES. A typical GroELocus, where GroES is followed by GroEL seems to be disturbed in phage EL. No GroES function could be identified in EL based on sequence similarity analysis. Alignment of the EL GroEL sequence with *E. coli* and *P. aeruginosa* GroEL sequences (data not shown) reveals conservation of most residues involved in ATP/ADP and Mg^{2+} binding,[33,34] which, in general, are the best conserved residues.[35] With the exception of two residues (Asn262, Gln267), residues presumably involved in substrate binding have a conserved hydrophobic/aromatic character.[35,36] Preliminary EM and ultracentrifugation data suggest a heptamer-like and 14-mer structure, respectively (V.V.M. *et al.*, unpublished observations). EL GroEL does not complement *E. coli* host defective GroEL44 chaperonin. This was concluded from the lack of phage T4 and lambda phage growth on *E. coli* cells with GroEL44 instead of host GroEL[37] and carrying the plasmid with EL GroEL under inducing conditions (data not shown). We interpret this as inability of phage

T4-gp31 or *E. coli* GroES (necessary for lambda growth) to interact with EL GroEL.

EL-gp37 shares protein similarity with IS605 family transposase B, a protein widespread throughout bacterial species. Transposase B is one of the two ORF products on type IS605 (IS606, IS607, IS608) insertion sequence elements, but also occurs as an integral part of a single protein element encoded on, for example, IS891 in Cyanobacter *Anabaena* sp.[38] There is extensive relatedness to the virulence factor GipA of *E. coli* and the prophage Gifsy-1 putative virulence protein of *S. typhimurium*.[39] Similarity to these virulence factors indicates that the transposase functionality should be investigated thoroughly prior to application of EL in phage therapy. The EL putative transposase, however, is not related to transposases of temperate *P. aeruginosa* phages D3112 and B3.

Conserved domain searches indicate the presence of transposase_2 (residues 8–289) and transposase_35 (residues 301–374) domains. The latter contains four conserved cysteine residues suggestive of DNA-binding as a zinc finger domain.

In summary, current genome sequence analysis of EL provides additional evidence for its classification as a ϕKZ-like Myoviridae. Both giant phages differ profoundly from other Myoviridae representatives. Genome sequence analysis revealed that about one-third of the predicted gene products of EL share similarity with ϕKZ. Comparison of the genome localization of ORFs encoding similar EL and ϕKZ gene products reveals blocks of conserved gene function in large genome regions. Probably, these ORFs encode essential proteins in the phage life-cycle. Besides, both EL and ϕKZ also contain unique genome regions, of 17.7 kb and 67 kb, respectively, encoding different proteins. It is important to understand what the role of these unique regions is, and how these large pieces of DNA have been inserted in the phage genomes. As in other phages, both EL and ϕKZ genomes are quite permissive for the accommodation of homing endonucleases. As for ϕKZ phage, we were not able to identify a DNA polymerase in the EL genome, and other proteins assisting DNA replication, including single-stranded DNA-binding proteins. Either they possess a completely distinct replication machinery or they make use of the host replication system. To understand quantitatively the origin and evolution of these giant phages, a substantially larger set of complete genome sequences of ϕKZ-related phages is needed.

Nucleotide sequence accession number

The phage EL genome sequence has been deposited in GenBank under accession no. AJ697969.

Acknowledgements

The help of Nina N. Sykilinda and Jill Ophoff in initial phage propagation and library construction is gratefully acknowledged. We thank Marleen Voet, Barbara Grymonprez and Ingrid Weltjens for expert technical assistance. The DNA sequencing was financed by the IQR-fund of the K.U.Leuven Laboratory of Gene Technology. Other support was from the Onderzoeksraad K.U.Leuven (grant OT/04/30 to G.V.), the Welcome Trust and Howard Hughes Medical Institute (to V.V.M.), from the Russian Fund for Basic Research (grant 02-04-48152 to V.N.K.) and from the Swiss National Fund (grant FN31-65403 to NS). KH holds a postdoctoral fellowship of the FWO-Vlaanderen (Belgium).

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2005.08.075

References

1. Hendrix, R. W. (2002). Bacteriophages: evolution of the majority. *Theor. Popul. Biol.* **61**, 471–480.
2. Ackermann, H. W. (2003). Bacteriophage observations and evolution. *Res. Microbiol.* **154**, 245–251.
3. Miller, E. S., Kutter, E., Mosig, G., Arisaka, F., Kunisawa, T. & Ruger, W. (2003). Bacteriophage T4 genome. *Microbiol. Mol. Biol. Rev.* **67**, 86–156.
4. Mesyanzhinov, V. V. (2004). Bacteriophage T4: structure, assembly, and initiation of infection studied in three dimensions. *Advan. Virus Res.* **63**, 287–352.
5. Desplats, C. & Krisch, H. M. (2003). The diversity and evolution of the T4-type bacteriophages. *Res. Microbiol.* **154**, 259–267.
6. Krylov, V. N. (2001). Phage therapy in terms of bacteriophage genetics: hopes, prospects, safety, limitations. *Russian J. Genet.* **37**, 869–887.
7. Mesyanzhinov, V. V., Robben, J., Grymonprez, B., Kostyuchenko, V. A., Burkal'tseva, M. V., Sykilinda, N. N. *et al.* (2002). The genome of bacteriophage ϕKZ of *Pseudomonas aeruginosa*. *J. Mol. Biol.* **317**, 1–19.
8. Burkal'tseva, M. V., Krylov, V. N., Pleteneva, E. A., Shaburova, O. V., Krylov, S. V., Volkart, G. *et al.* (2002). Phenogenetic characterization of a group of giant phi KZ-like bacteriophages of *Pseudomonas aeruginosa*. *Genetika*, **38**, 1470–1479.
9. Krylov, V. N., Pleteneva, E. L., Bourkal'tseva, M., Shaburova, O., Volckaert, G., Sykilinda, N. *et al.* (2003). Myoviridae bacteriophages of *Pseudomonas aeruginosa*: a long and complex evolutionary pathway. *Res. Microbiol.* **154**, 269–275.
10. Fokine, A., Kostyuchenko, V. A., Efimov, A. V., Kurochkina, L. P., Sykilinda, N. N., Robben, J. *et al.* (2005). A three-dimensional cryo-electron microscopy structure of the bacteriophage ϕKZ head. *J. Mol. Biol.* In the press.
11. Kessler, C. & Manta, V. (1990). Specificity of restriction endonucleases and DNA modification methyltransferases a review (Edition 3). *Gene*, **92**, 1–248.

12. Lavigne, R., Sun, W. D. & Volckaert, G. (2004). PHIRE, a deterministic approach to reveal regulatory elements in bacteriophage genomes. *Bioinformatics*, **20**, 629–635.
13. Rice, P., Longden, I. & Bleasby, A. (2000). EMBOSS: the European molecular biology open software suite. *Trends Genet.* **16**, 276–277.
14. Letondal, C. (2001). A Web interface generator for molecular biology programs in Unix. *Bioinformatics*, **17**, 73–82.
15. Miller, E. S., Heidelberg, J. F., Eisen, J. A., Nelson, W. C., Durkin, A. S., Ciecko, A. *et al*. (2003). Complete genome sequence of the broad-host-range vibriophage KVP40: comparative genomics of a T4-related bacteriophage. *J. Bacteriol.* **185**, 5220–5233.
16. Tetart, F., Desplats, C., Kutateladze, M., Monod, C., Ackermann, H. W. & Krisch, H. M. (2001). Phylogeny of the major head and tail genes of the wide-ranging T4-type bacteriophages. *J. Bacteriol.* **183**, 358–366.
17. Black, L. W., Showe, M. K. & Steven, A. C. (1994). Morphogenesis of the T4 head. In *Molecular Biology of Bacteriophage T4* (Karam, J., Drake, J. W., Kreuzer, K. N., Mosig, G., Hall, D. H., Eiserling, F. A. *et al*., eds), pp. 218–258, American Society of Microbiology, Washington, DC.
18. Liu, J. & Mushegian, A. (2004). Displacements of prohead protease genes in the late operons of double-stranded-DNA bacteriophages. *J. Bacteriol.* **186**, 4369–4375.
19. Smith, M. C., Burns, N., Sayers, J. R., Sorrell, J. A., Casjens, S. R. & Hendrix, R. W. (1998). Bacteriophage collagen. *Science*, **279**, 1834.
20. Plunkett, G., 3rd, Rose, D. J., Durfee, T. J. & Blattner, F. R. (1999). Sequence of Shiga toxin 2 phage 933W from *Escherichia coli* O157:H7: Shiga toxin as a phage late-gene product. *J. Bacteriol.* **181**, 1767–1778.
21. Duplessis, M. & Moineau, S. (2001). Identification of a genetic determinant responsible for host specificity in *Streptococcus thermophilus* bacteriophages. *Mol. Microbiol.* **41**, 325–336.
22. Wang, I.-N., Smith, D. L. & Young, R. (2000). Holins: the protein clocks of bacteriophage infections. *Annu. Rev. Microbiol.* **54**, 799–825.
23. Drake, J. W. & Kreuzer, K. N. (1994). Transactions in T4-infected *Escherichia coli*. In *Molecular Biology of Bacteriophage T4* (Karam, J., Drake, J. W., Kreuzer, K. N., Mosig, G., Hall, D. H., Eiserling, F. A. *et al*., eds), pp. 11–13, American Society of Microbiology, Washington, DC.
24. Dalgaard, J. Z., Klar, A. J., Moser, M. J., Holley, W. R., Chatterjee, A. & Mian, I. S. (1997). Statistical modeling and analysis of the LAGLIDADG family of site-specific endonucleases and identification of an intein that encodes a site-specific endonuclease of the HNH family. *Nucl. Acids Res.* **25**, 4626–4638.
25. Sitbon, E. & Pietrokovski, S. (2003). New types of conserved sequence domains in DNA-binding regions of homing endonucleases. *Trends Biochem. Sci.* **28**, 473–477.
26. Perler, F. B., Olsen, G. J. & Adam, E. (1997). Compilation and analysis of intein sequences. *Nucl. Acids Res.* **25**, 1087–1093.
27. Aravind, L. & Koonin, E. V. (1999). Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches. *J. Mol. Biol.* **287**, 1023–1040.
28. Thornton, K., Forstner, M., Shen, M. R., West, M. G., Rupp, B. & Thelen, M. P. (1999). Purification, characterization, and crystallization of the distal BRCT domain of the human XRCC1 DNA repair protein. *Protein Expr. Purif.* **16**, 236–242.
29. Wong, P. & Houry, W. A. (2004). Chaperone networks in bacteria: analysis of protein homeostasis in minimal cells. *J. Struct. Biol.* **146**, 79–89.
30. Zeilstra-Ryalls, J., Fayet, O. & Georgopoulos, C. (1991). The universally conserved GroE (Hsp60) chaperonins. *Annu. Rev. Microbiol.* **45**, 301–325.
31. Richardson, A., van der Vies, S. M., Keppel, F., Taher, A., Landry, S. J. & Georgopoulos, C. (1999). Compensatory changes in GroEL/Gp31 affinity as a mechanism for allele-specific genetic interaction. *J. Biol. Chem.* **274**, 52–58.
32. Ang, D., Richardson, A., Mayer, M. P., Keppel, F., Krisch, H. & Georgopoulos, C. (2001). Pseudo-T-even bacteriophage RB49 encodes CocO, a cochaperonin for GroEL, which can substitute for *Escherichia coli's* GroES and Bacteriophage T4's Gp31. *J. Biol. Chem.* **276**, 8720–8726.
33. Boisvert, D. C., Wang, J., Otwinowski, Z., Horwich, A. L. & Sigler, P. B. (1996). The 2.4 A crystal structure of the bacterial chaperonin GroEL complexed with ATP gamma S. *Nature Struct. Biol.* **3**, 170–177.
34. Xu, Z., Horwich, A. L. & Sigler, P. B. (1997). The crystal structure of the asymmetric GroEL-GroES-$(ADP)_7$ chaperonin comples. *Nature*, **388**, 741–750.
35. Brocchieri, L. & Karlin, S. (2000). Conservation among HSP60 sequences in relation to structure, function, and evolution. *Protein Sci.* **9**, 476–486.
36. Fenton, W. A., Kashi, Y., Furtak, K. & Horwich, A. L. (1994). Residues in chaperonin GroEL required for polypeptide binding and release. *Nature*, **371**, 614–619.
37. Georgopoulos, C. P., Hendrix, R. W., Casjens, S. R. & Kaiser, A. D. (1973). Host participation in bacteriophage lambda head assembly. *J. Mol. Biol.* **76**, 45–60.
38. Bancroft, I. & Wolk, C. P. (1989). Characterization of an insertion sequence (IS891) of novel structure from the cyanobacterium *Anabaena* sp. strain M-131. *J. Bacteriol.* **171**, 5949–5954.
39. Stanley, T. L., Ellermeier, C. D. & Slauch, J. M. (2000). Tissue-specific gene expression identifies a gene in the lysogenic phage Gifsy-1 that affects *Salmonella enterica* serovar *typhimurium* survival in Peyer's patches. *J. Bacteriol.* **182**, 4406–4413.
40. Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S. R. *et al*. (2002). The Pfam protein families database. *Nucl. Acids Res.* **30**, 276–280.
41. Besemer, J., Lomsadze, A. & Borodovsky, M. (2001). GeneMarkS: a self-training method for predicition of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucl. Acids Res.* **29**, 2607–2618.
42. Delcher, A. L., Harmon, D., Kasif, S., White, O. & Salzberg, S. L. (1999). Improved microbial gene identification with GLIMMER. *Nucl. Acids Res.* **27**, 4636–4641.
43. Lavigne, R., Sun, W. D. & Volckaert, G. (2003). STORM towards protein function: systematic tailored ORF-data retrieval and management. *Appl. Bioinformatics*, **2**, 177–179.
44. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402.
45. Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I.,

Appel, R. D. & Bairoch, A. (2003). ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucl. Acids Res.* **31**, 3784–3788.
46. Marchler-Bauer, A., Anderson, J. B., DeWeese-Scott, C., Fedorova, N. D., Geer, L. Y., He, S. *et al*. (2003). CDD: a curated Entrez database of conserved domain alignments. *Nucl. Acids Res.* **31**, 383–387.
47. Lowe, T. M. & Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucl. Acids Res.* **25**, 955–964.
48. el-Mabrouk, N. & Lisacek, F. R. (1996). Very fast identification of RNA motifs in genomic DNA. Application to tRNA search in the yeast genome. *J. Mol. Biol.* **264**, 46–55.

Edited by J. Karn

(Received 23 August 2005; accepted 31 August 2005)
Available online 5 October 2005