Nucleotide sequence of the DNA packaging and capsid synthesis genes of bacteriophage P2

Nora A.Linderroth, Rainer Ziermann, Elisabeth Haggård-Ljungquist¹, Gail E.Christie² and Richard Calendar*

Department of Molecular and Cell Biology, University of California, 401 Barker Hall, Berkeley, CA 94720, USA, ¹Department of Microbial Genetics, Karolinska Institutet, Box 60400, S-10401, Stockholm, Sweden and ²Department of Microbiology and Immunology, Virginia Commonwealth University, Box 678, MCV Station, Richmond, VA 23298–0678, USA

EMBL accession no. X61229

ABSTRACT

Overlapping DNA fragments containing the DNA packaging and capsid synthesis gene region of bacteriophage P2 were cloned and sequenced. In this report we present the complete nucleotide sequence of this 6550 bp region. Each of six open reading frames found in the interval was assigned to one of the essential genes (*Q*, *P*, *O*, *N*, *M* and *L*) by correlating genetic, physical and mutational data with DNA and protein sequence information. Polypeptides predicted were: a capsid completion protein, gpL; the major capsid precursor, gpN; the presumed capsid scaffolding protein, gpO; the ATPase and proposed endonuclease subunits of terminase, gpP and gpM, respectively; and a candidate for the portal protein, gpQ. These gene and protein sequences exhibited no homology to analogous genes or proteins of other bacteriophages. Expression of gene *Q* in *E. coli* from a plasmid caused production of a M_r 39,000 Da protein that restored *Qam34* growth. This sequence analysis found only genes previously known from analysis of conditional-lethal mutations. No new capsid genes were found.

INTRODUCTION

Assembly of bacteriophage particles is an ordered, multi-step developmental process amenable to biochemical, genetic and structural analyses. A key step in assembly is encapsidation of the viral genome. This encompasses: (i) construction of empty immature proheads; (ii) maturation of proheads to capsids; (iii) recognition and cleavage of circular or multimeric viral DNA; (iv) energy-dependent DNA packaging (head filling); and (v) capsid completion (reviewed in 1,2). Encapsidation is followed by attachment of tails, which are produced in a separate pathway. Apart from minor variations, these processes appear to be remarkably similar for numerous dsDNA phages.

Bacteriophage P2 is the paradigm and best studied member of the P2 family. It has an icosahedral head of about 60 nm, a contractile tail of 135 nm and multiple tail fibers. The viral genome is a nonpermuted molecule of dsDNA, approximately 33 kb in length, with 19 bp cohesive ends (*cos*). The complete nucleotide sequence of the *cos* region, which includes all *cis*-acting DNA elements required for encapsidation of phage DNA by the P2 packaging factors, has been determined and characterized (3). (See ref. 4 for a recent review of P2 structure and biology.)

P2 proheads and capsids consist primarily of N* [M_r 36,700 daltons (36.7 kDa)] and two minor components, h1 (39.0 kDa) and h2 (38.6 kDa). These polypeptides are derived from full length N protein (40.2 kDa) by amino-terminal processing (5,6; Rishovd and Lindqvist, in prep.). Both the assembly of proheads and the concomitant cleavage of gpN require gene *O*. It is believed that O protein (31.4 kDa) functions as a scaffold during capsid assembly. Findings consistent with this idea are: (i) full length O protein is not found in mature capsids, and (ii) O protein is processed to low molecular mass forms in the presence of the capsid precursor, gpN (5). Scaffolding proteins of other dsDNA phages have similar fates (1).

The products of genes *Q*, *P* and *M* are needed to package DNA into proheads and for the conversion of proheads to capsids. Terminase activity resides with the products of genes *M* and *P* (7,8,9). A 28 kDa protein, gpM, is the leading candidate for the endonuclease that directs *cos* cleavage because extracts from *M* mutant-infected cells package mature but not circular P2 DNA (8). The product of gene *P* (65 kDa) is a DNA-dependent ATPase involved in terminase action and head filling (8,9). The *Q* gene product is also required for head filling but neither its precise role nor site of action are known (5,7). Completion of filled heads is carried out by the product of gene *L*, which acts late in the assembly process and renders the newly packaged DNA in filled heads resistant to DNase (7). L protein is assumed to bind to DNA-filled capsids and might be similar to the F_{II} protein of λ (10).

* To whom correspondence should be addressed

Several features of P2 make it ideal for studying the regulation of DNA packaging and capsid size determination. First, P2 and its satellite P4 are unique among dsDNA phages because they package circular, rather than linear, DNA (11). Second, the P2 capsid assembly pathway can be redirected by P4, which uses its *sid* (size determination) gene and all the capsid genes of P2 to assemble heads having just one-third the volume of P2 heads (12,13). The diminutive heads are more suitable for P4's 11.6 kb genome. Finally, DNA packaging by both phages is regulated in at least two ways, by *cos* recognition and by genome length. Dimeric or trimeric P4 genomes containing more than one *cos* site can be packaged into P2-sized heads (11,13). Similarly, P2 and P4 heads can package multimers of plasmids having a *cos* site, a feature that can be exploited for biotechnology applications (3,14).

To facilitate structural and molecular analyses of P2 capsid assembly and DNA packaging, we cloned and sequenced the genetic determinants for these functions. In this paper we report the DNA and protein sequences of genes *Q*, *P*, *O*, *M* and *L*. This study, along with the recent characterization of gene *N* by Six et al. (6), completes the nucleotide sequence of the DNA packaging and capsid region of P2. An expression vector that overproduces the product of gene *Q* is also described and is used here for positive identification of Q protein.

MATERIALS AND METHODS

Strains, media and growth conditions

All strains used to grow P2 were derivatives of *E. coli* strain C. C-1a is prototrophic *sup*⁺ (15), C-520 carries *supD* and C-1792 carries *supF* (16). Indicator strains for P2 were C-1757 (*supD*), C-1792 (*supF*) and C-1055 (*sup*⁺) (16,17). For cloning, P2 *lg* (18) and P2 *vir1* (19) DNAs were used as wild-type. References for P2 head mutants are: *c5 Oam279* (5); *vir1 Pam24*, *vir1 Mam32*, *vir1 Qam34*, *vir1 Lam79* (16); *Mts52* (20); *vir1 Lam9*, *vir1 Pam137* (21); *c5 Pam253* (22). P2 was grown using media and protocols given in (14). *Mts52* was grown at 30°C; all other phage and bacterial strains were grown at 37°C. Strains JM105 (23) and D-1210 (24) were used for cloning. Ampicillin (50 µg/ml) and kanamycin (25 µg/ml) were added to media as required.

Molecular cloning and recombinant plasmids

Linear DNA from P2 virions was purified by extractions with phenol and CHCl₃. Preparation of plasmid DNA for cloning or analysis and preparation and cloning of restriction fragments in pBR322 (25), pUC18 or pUC19 (26) was carried out using standard techniques (27) and commercial enzymes.

pEE454 contains the 99.0–3.0% *SspI* fragment of P2 [bp –311 to 978, with respect to *cosL*] in pUC18/*SmaI*. pEE481 carries the 0.8–6.7% *BamHI* region of P2 [bp 264–2221] in pUC18/*BamHI*; pMN2 is identical to pEE481 except that *vir1 Qam34* DNA was used. pEE470 contains the 6.7% *BamHI*–13.9% *PstI* interval of P2 [bp 2222–4595] ligated in pUC18/*BamHI*+*PstI*. pGC596-17 is pBR322 with the 5.5–10.5% region of P2 wild-type [bp 1827–3586] (22); pGC17-P24, pGC17-P137 and pGC17-P253 are identical to pGC596-17 but contain restriction fragments from *P* mutants (*vir1 Pam24*, *vir1 Pam137* or *c5 Pam253*, respectively). To make pEE649, PCR amplification and cloning of the 10.5–12.0% region from *c5 Oam279* [bp 3570–4097] into pUC18/*SmaI* was as follows. DNA from 10⁸ plaque forming units was amplified using a GeneAmp kit (Perkin-Elmer). The reaction contained 20

pmol of P2 primers 10.5R and 12.0L (next section), and was subjected to 25 cycles of 1 min at 94°C, 2 min at 42°C and 3 min at 72°C, using a Perkin-Elmer-Cetus thermal cycle unit (PCR1000). pEE488 carries the 13.1% *KpnI*–16.8% *Eco47III* region of P2 [bp 4335–5548] in pUC18/*HincII*+*KpnI*. pRZ12 has the 14.9% *MluI*–18.5% *EcoRV* region of P2 [bp 4908–6115] in pUC19/*HincII*. pM32/1 contains the 17.4% *TaqI*–17.9% *HincII* fragment of *vir1 Mam32* [bp 5744–5898] in pUC19/*HincII*. pM52/1 carries the 15.5% *DdeI*–16.3% *ApaLI* region of *Mts52* [bp 5126–5393] in pUC19/*HincII*. pGC8-5 has the 10.8% *EcoRI*–17.7% *EcoRV* interval of P2 [bp 3585–6116] in pBR322/*EcoRI*+*EcoRV*. pGCRV7 contains the 18.0–25.0% *EcoRV* fragment of P2 [6.3–8.2 kb] in pBR322/*EcoRV*. pGC/Lam9 carries the 17.4–33.4% *HincII* fragment of *vir1 Lam9* [5.9–11 kb] in pUC18/*HincII*. pGC/Lam79 contains the 7.6–17.7% *EcoRV* fragment of *vir1 Lam79* [bp 2505–6116] in pUC18/*HincII*. All other other P2 DNA-containing plasmids used were derived from these sources.

pUHE21-2 was a gift of H. Bujard. For pNL93Q, the 0.0% *BssHII* to 3.7% *BstXI* region of P2 [bp 22–1234] was cloned into pUHE21-2/*HindIII*+*EcoRI* (Fig. 3A). Although not mentioned in the text, pNL93Q was maintained in cells also carrying pRG1, which is pACYC177 with *lacI*^Q (28).

DNA sequence determination and computer analysis

DNA sequence was determined by the method of Sanger et al. (29) or Maxam and Gilbert (30). Double-stranded plasmid templates for Sanger sequencing were prepared as described (31). Forward or reverse universal primers were normally used; P2-specific primers 10.5R (GCACGGAATACCTGGAATTC), 12.0L (GCGGGTTTCCTGGCGCATAG), 15.4L (TCGGTGCCGAGACCCGC), 19.2L (CCATCGCACACACGGCG), 19.4R (GACAGCATTGACAGCAC), 19.8L (TGGCGTCGAGCGTGTCG) and 20.5L (CCGACAACAATCAGCA) were synthesized on a Gene Assembler (Pharmacia), Beckman 200A or Applied Biosystems 380A unit. A Sequenase™ kit (U.S. Biochem. Corp.) and [α -³⁵S]-dATP (Amersham) were used for chain termination reactions. Sequence data were stored and analyzed using Intelligenetics programs, on a Sun Microsystems computer having the UNIX operating system. The comparison method of Pearson and Lipman was employed (32). P2 DNA and protein sequences were compared to GenBank 67 (Mar, 1991), EMBL 26 (Feb, 1991) and SWISS-PROT 17.0 (Feb, 1991) databases.

Marker rescue and complementation

Rescue of wild-type alleles from P2 DNA cloned in plasmids was used to ascertain the approximate locations of mutant genetic markers. Dilutions of P2 *am* or *ts* mutants (10³–10⁶ phages) were spotted onto lawns of C-1a harboring recombinant plasmids. Alternatively, cells and phage were preincubated together for 10 min and added to molten top agar before plating. Plates were incubated overnight at 37°C, or at 42°C in the case of *Mts52*. Plaques or confluent lysis indicated rescue of a wild-type allele by mutant phage. Typical frequencies of rescue were 1–5×10⁻³. Recombinant phages were retested for growth on C-1055 at 37°C (for *am*⁺) or 42°C (for *ts*⁺). For plate complementation, dilutions of P2 *vir1 Qam34* (10¹–10³ phages) were spotted onto lawns of C-1a harboring pNL93Q or pUHE21-2, or on strain C-1757, and were incubated overnight at 37°C. Phage in isolated plaques were replated on C-1055 and C-1757.

Analysis of proteins

Logarithmic cultures were induced with IPTG for 3 h (or uninduced). One ml of cells (6×10^8 cells/ml) was pelleted by centrifugation and lysed at 95°C in protein gel sample buffer. Aliquots representing 3×10^6 cells were run on a 10% polyacrylamide/SDS gel and stained with Coomassie Blue. Protein mass standards (Bio-Rad) were run in parallel.

RESULTS

Nucleotide sequence of the P2 DNA packaging and capsid gene region

The morphopoietic genes of bacteriophage P2 account for two-thirds of its 33 kb genome (Fig. 1). DNA packaging and capsid (DP/C) genes lie in two divergent transcription units in the 0–20% region. In previous studies, we determined the relative positions and approximate boundaries of DP/C genes and identified the start sites of transcription (16,22). The complete sequence of a 6550 bp interval containing the six genetically identified capsid and packaging genes has now been determined. Fig. 1 shows the strategy used to obtain the DNA sequence for the DP/C region and a revised restriction map. We determined the sequence for both DNA strands, including all cloning sites. All of the sequence was determined multiple times. The sequences of *cosL*, the *P* and *O* promoters and the *N* gene (black boxes in Fig. 1) have been published before (3,6,22).

In Fig. 2 we present the complete nucleotide sequence of the DP/C region. Within the interval were six closely spaced but non-overlapping open reading frames (ORFs) commencing with ATG, which were preceded by probable ribosome binding sites (see below). From the left end of the genome, these were: ORF 1, bases 1221–187; ORF 2, bases 2993–1221; ORF 3, bases 3101–4021; ORF 4, bases 4080–5153; ORF 5, bases

5157–5900; and ORF 6, bases 6000–6509 (start and stop codons, inclusive). The bottom DNA strand contained the first two ORFs, oriented from right to left; the four remaining ORFs proceeded left to right on the top strand. The two regions were separated by the divergent *P* and *O* promoters. The directionality of the six ORFs agreed with late transcription patterns (22). A study of the use of synonymous codons revealed no exceptional use of rare codons. Overall, codon usage was very similar to that in highly expressed *E. coli* genes (35). The base composition over the interval was 56.5% GC rich, slightly higher than the P2 genome as a whole, which is 52% G+C (36). We found that 68% of codons in the region ended in either G or C. This could account for its higher G+C content.

Immediately downstream of ORF 1 was a G+C rich dyad preceded by six A residues (bases 160–181). The corresponding mRNA from this region (bottom strand) could be drawn so as to resemble a rho-independent terminator with a predicted free energy value (ΔG°_{37}) of $-14.9 \text{ kcal mol}^{-1}$. In anticipation of its probable role in transcription termination, this region is labeled T_Q in Fig. 2.

Correlation of DP/C region ORFs and P2 genes

Genetic and biochemical data were employed to assign each of the six DP/C region ORFs with its correlate P2 gene (see Fig. 1). ORF 2 and ORF 3 could be aligned with genes *P* and *O*, respectively, whose translation start sites were known (22). ORF 2 contained the published portion of *P* and the final 95% of the gene. Similarly, ORF 3 possessed the known and final portions of gene *O*. It also included bases further upstream that are not part of the *O* transcript (22). Thus, *O* begins with the third Met codon in ORF 3, at bp 3167. The sequence and genetic analysis of ORF 4, the *N* gene, was reported recently (6), and is not discussed here. Identification of the three remaining ORFs was

Figure 1. Genetic map of bacteriophage P2 and genetic, restriction and sequencing map of the DP/C region. (Top) Genetic map of the P2 genome (33). Bold letters indicate structural genes. Arrows denote transcription units. Included are newly identified genes *W* and *I* (E. Haggård-Ljungquist, in prep.) and *F_{II}* (34). (Lower) The DP/C region (enclosed by dotted lines). A partial restriction map shows unique and infrequent cutting sites for the interval reported in this study. White boxes below the gene symbols indicate open reading frames. Bent arrows represent the *P* and *O* promoters, *P_P* and *P_O*, respectively. Straight arrows denote the position, orientation and extent of sequencing reactions. Arrows with an open circle show sequences obtained with uniquely end-labeled DNA. Arrows with a black circle indicate reactions primed by unique P2 primers. All other sequence was obtained using universal primers. Only the minimum number of arrows needed to cover both strands of the sequenced region are shown. The thick black boxes correspond to DNA whose sequence has been reported (3,6,22).

COS	
GGCGAGGCGG·GGAAAGCACT·GCGCGCTGAC·GGTGGTGCTG·ATTGTATTTT·TTCAGCGTCT·CAGCGCGCTCG TGACGGCACT TAGTCTGCCC GTTGAGGCGT	100
TGTGTGTCTG CGGGGTGTTT TGTGCGGTGG TGAGCGTGTG AGGGGGGATG ACGGGGTGTA AAAAGCCGC CCGCAGGCGG CGATGTTTCTG TCGTTGTCTG	200
TQ -----> <----- O D N D T	
TGTCCAGTGA GTAGTTTTTA AAGCGGATGA CCTCCTGACC GAGCCAGCCG TTTATCTCGG GGATCTGTGC CTGTAACGGG ATAAGCTCAT TCGGGACAAA	300
D L S Y N K F R I V E Q G L W G N I E R I R D Q L P I L E N R V F	
GACCTTTGCC ACTTTCTCAA TATCACCAGC GACCCGACG TTCTCCGGCT TGCCACCCAT CAACTGAAAG GGGATGCGGT GCGCGTCCAG CAGGTGACGG	400
V K A V K E I D G L S G V N E P K G G H L N F P I R H A D L L D A	
GCGCTGGCTT TTTTGATATT AAAAAATCG TCCTTCGTGC CCATCTCACT GAGGGGGATA ATTTAATGC CGTCGGCTTT CCCCTGTGGG GCATAGAGAA	500
A S A K K I N F F D D K T A V E S L P I I K I G D A K G Q P A Y L F	
ACAGGTTTT AAAGTTGTTG CGGCTTTTCG ACTTGACCAT GTTTTCGCGA AGCATTTCGA TATCGTTGCG ATCTGTCAGG GCATCGGTGA CATACATGAT	600
L N K F N N R G K S K V N N E R L M E I D N R D Q V A D T V Y M I	
GTATCCGGCA TGTGCGCAT TTTCGTAATA CTTCGCGCGG AACACCTGG CCGACTCATT CAGCCAGGCA GAGTTAAGGG CGCTGAGATA TTCCGGCAGG	700
Y G A H A G N E Y Y K R R F L T A S E N L W A S N L A S L Y E P L	
CCGTACAGCT CCTGATTAAT ATCCGGCTCC AGCAGGTGAA ACACGGAGCC GGGCGCGAAG GCTGTGCGCT CGTTGAAGGA CCGCACCCAC CAGTAACAT	800
G Y L E Q N I D P E L L H F V S G P A F A T P E N F S P V W N Y V D	
CCTCTCCAC GCCACGGCGG GTATATTTTG CCGGTGAGGT TTCCAGTCTG ATGACCTTAC CGTGTGTGCT GTAACGCTTT TCCAGAAACG CATTACCGAA	900
E E V G R R T Y K A P S T E L R I V K G T T S Y R K E L F A N G F	
CACGAGAAA TCCAGCAAA AGCGGTGAA ATCTGCTGG GAAAGCCATG GATGCGGGAT AAATGTCGAG GCCAGAATAT TCGGTTTTCG GTAATTCGGC	1000
V L F D L F Q S L V F R S F D Q Q H P I F T S A L I N R K V Y C P	
GAGCTGTGAT GCACGGCAGC CCGCAGGCTT TTTGCCAGAC CGGTAAAGCT GACCGGTGCG TCATACCATC TGCGGTTACT GATGCACTCG ACGTAATCCA	1100
S S H H V A A R L S K A L G T F S V P P E Y W R G N S I C E V Y D L	
GAATGTCAGC GCGGTGAGT ACCGGCAGCG GCTCACCAAA GGTGAATGCC TCCATTTTCG GGGCGGTGCG GGTCTATTGT TTTGCCGACG GTTGGCGGTG	1200
I D R R D L V P V P E G F T F A E M K P G S A T M T K A A P Q P T	
TTTCCTTTT TTCTTGCTCA TCAGTAAAC TCCAGAATGG TGGATGTCAG CCGGGTGCTG ATACCGGCGG TGAGTGGCTC ATTTAACAGG GCGTGCATGG	1300
O Y F E L I T S T L P T S I G A T L P E N L L A H M T	
K G K K K S M ← [Q] * ***	
TCGCCAGGC GAGGTGCGCG TGGTGGCTT CCTCGTGGC GCTGGCTCA TAGGTGGCGC TCGTCCGCT GCTGTGCTG GTCTTGCGGA TAGCCATAAA	1400
A N A L D A H S A E E S R S A E Y T A S R G S S T M T K R I A N F	
CGAGCTGGTG ATGTCAGTGG CGCTGACGTC GTATTCCAGA CAGCCACGGC GGATAACGTC TTTTGCTTGG AGCACCATTG CGGTTTTCAT TTCCGCGGTG	1500
S S T I D T A S V D Y E L C G R R I V D K A K L V H A T K M E P T	
TAGCGGATAT CACGCGCGGC GGGATAGAAC GAGCGCACGA GCTGGAACAC GCCGACACCG AGGCGCGTGG CATCAATACC GATGTATTCG ACGTTGTATT	1600
Y R I D R A A P Y F S R V L Q F V G V G L G T A D I G I Y E V N Y K	
TTTCGCTGAG TTTGCGGATG GATTCCGCCT GGGTGCGAAA GTCCATGCTT TTCCACTGGT GACGCTCAAG TATTTCTGAAT TTGCCACCGG CCACACCGG	1700
E T L K R I S E A Q T D M G K W Q H R E L I R F K G G A K N F	
CGGTGCCAGT ACCACGCATC CCGCGCTGTC GCCACGGTGT GACGGGTGCT AACCAATCCA TACCGGGCGG GAGCCGAACG GATTGGCGGC AAACGGCGCA	1800
P A L V V C G A S D G R H S P D Y G I W V P R S G F P N A A F P A	
TAGTCTTCCC ATTCTTCCAG CGTGTCGACC ATGCAGCGTT GCAGCTCCTC GAACGGGAAC ACCGATGCTT TGTGCTCAAC AAATTCACAC ATGAACAGGT	1900
Y D E W E E L T D V M C R Q L E E F P F V S A K D D V F E C M F L N	
TTTTAAATC GTCCGCGCTG TTTTCGCGTT TGAGCTGCTC AATGTCGAAC AGCGTGCAGC CGCCTTTCAG GCGCTCTCA ATGGTGACAA TCTGTGCGCA	2000
K F D H A N E R K L Q E I D F T C G G K L A D E I T V I Q R W	
CTGGCGTCC GCACAGAGAA GCCACCGGC AAGGGCGTTA TGACTGACGT CGATTTCAC CCGTTCGCGG CCGCTGGCGG GTCCCCGGTT AAACAGTTCA	2100
Q G D A C L L G G A L A N H S V D I E V R E A A S A R G R N F L E	
CCCGACCAGA ACGGTGAGG GTCTGCGGCC AGCGTGAGC GGGTGAGAA ATAGTTCGAG CCGAGGTGAC TCTGTGAGG CATACCTGAT GCCACCTTAC	2200
G S N F P Y A D H A L T S P T S F Y T S R L H S Q S A M G S A V K R	
GCAGTACCTG AAAATTCGGG ATCCAGAGAA TCTCATCGAC GTACAGGTGG CCGTTATGAC TCTGCGCGGT GTTGAGATTG GTGCCGAGAA AAATCAGTTT	2300
L V Q F N P I W F I E D V Y L D K N H S Q A T N S N T G L F I G N	
TGCGCGTTA TTGCCAGGA CAATCGGGTC ACCGGTCAGG TCAACGTCAA CCAGACGGGC AAAGCGGATG ATGTATTGCG GGAACACATA CGCTGTGTTT	2400
A G N N G L V I P D G T L D V D V L R A F A I I Y E R F V Y A Q T	
TTACTGGCCG ACAGAAAAAT CTGGTTATGA CCGGTTTTCG GGGCAGCGAG CAGCGCTCG CCGGAAAAAT AAAACGTGCG GCCAATCTGG CGGGATTTC	2500
K S A S L F I Q N H G T K L A R L L A E R S F Y F T A G I Q R S K L	
GGATATCGCG GATGCGGTGC TCAAGCCCGG CACGATACCA GTGAGCTGTA TAGTGAAGAG ACTGCTCAA GAAATCTGCG TCCAGCTTTT CGATGGCCTC	2600
I D R I R H E L G A R Y N H L Q Y D F S Q E F F I Q E L K E I A E	
GTCACTGAAA AAATCTTTT TCGGTTTTCG CCGTCCGCCT TTGTTACGCT TAGCGACGTT CCGATTAAAG TCTGCTCGT TGCCGGTCTG GCTGTAGCGG	2700
D S F F N K K P K R R G G K N R N A V N P N L D A E N G T Q S Y R	
TTGACCCGTG CCAGCTGTTT AATCTGGCGT CCCAGCAGGT CAATTCCTTT GAAGTACCG CCGGTTTCTT GTGTTTGTAT GATGAGCTGG GTCAGCCGCG	2800
N V R A L R E I Q R G L L D I E K F D G G T K Q P K I I L Q T L R A	
CTTCCAGACT CATTTGACA CCGCTGATGG GGGCAACGCT GTCCGACCGG TCGCGCTGTT TCCAGCTCTG CACTGTCGGG CGTTTCATCT GCAACATGGC	2900
E L S H E V R S I P A V S D W G R Q K W S Q V T P R K M Q L N A	
GGCAATCTGC GGCACGGAAA ACCCTGCCA GTACAGCAGC GCCCGCTGAC GACCGGGGTC GTGTAAAGA GTGGTGTCTG TGGTGTGGT CATGAATACC	3000
A I Q P V S F G Q W Y L L A A Q R R P D H L L T T D T T I T M ← [P]**	
TCGCGGTGAT GAATACACGG CAAGGCTACT GAOTCGCGCC CCGCGATTGG CTAAGGTGCT GTTGTGTCAG TGATAAGCCA TCCGGGACTG ATGGCGGAGG	3100
** Pp	
Po > **** [O] → N A K K V S K F F R I	
ATGCGCATCG TCGGAAACT GATGCCGACA TGTGACTCCT CTAATCACTA TTCAGGACTC CTGACAAATG CAAAAAAGT CTCAAAATTC TTTGCTATCG	3200
G V E G D T C D G R V I S A Q D I Q E M A E T F D P R V Y G C R I N	
GCTTGAGGG TGACACTGT GACGGCGGTG TCATCACTGC GCAGGATATT CAGGAAATGG CCGAAACCTT TGACCCGCGT GTCTATGGTT GCCGATTAA	3300
L E H L R G I L P D G I F K R Y G D V A E L K A E K I D D D S A L	
CCTGGAACAT CTGCGCGCA TCTGCTGTA CGGTATTTT AAGCGTTATG GCGATGTGGC CCAACTGAAG GCCGAAAAGA TTGACGATGA TTGCGGCTG	3400
K G K W A L F A K I T P T D D L I A M N K A A Q K V Y T S H E I Q	
AAAGGCAAT GGGCGCTGTT TCGAAAAATC ACCCGACCGC ATGACCTTAT CCGATGAAC AAGCGCGCGC AGAAGGTCTA CACCTCAATG GAAATTCAGC	3500

P N F A N T G K C Y L V G L A V T D D P A S L G T E Y L E F C R T A	
CGAAGTTTGC CAACACCGGC AATGTTATC TGGTGGGTCT GGCCTGACCG GATGACCCGG CAAGCCTCGG CACGGAATAC CTGGAATTCT GCCGACGGC	3600
K H N P L N R F K L S P E N L I S V A T P V E L E F E D L P E T V	
AAAACACAC CCCCTGAACC GCTTCAAAAT AAGCCCTGAA AACCTGATT CAGTGGCAAC GCCTGTTGAG CTGGAATTG AAGACCTGCC TGAACCGGTG	3700
F T A L T E K V K S I F G R K Q A S D D A R L N D V H E A V T A V	
TTCACCGCCC TGACGAAAAA GGTGAAGTCC ATTTTGGGCC GCAACACGCG CAGCGATGAT GCCGTCTGA ATGACGTGCA TGAAGCGGTG ACCGCTGTTG	3800
A E H V Q E K L S A T E Q R L A E M E T A F S A L K Q E V T D R A D	
CTGAACATGT CGAGGAAAAA CTGAGCGCCA CTGAGCAGCG CCTCGCTGAG ATGGAACCG CTTTCTCTGC ACTTAAGCAG GAGGTGACTG ACAGGCGGGA	3900
E T S Q A F T R L K N S L D H T E S L T Q Q R R S K A T G G G G D	
TGAACACGCG CAGGCATTCA CCCGCTGAA AACAGTCTC GACCACACCG AAAGTCTGAC CCAGCAGCGC CGCAGCAAAG CCACCGCGCG TGGCGGTGAC	4000
A L M T N C O	
GCCCTGATGA CGAAGTCTG ACCGCGCTCA GTCACTCCGG GAAACCTTC ACGATTAACC CTTAATTTC GGAATAACTA TCGCCAGGA AACCGCTTT	4100
K F N A Y L S R V A E L N G I D A G D V S K K F T V E P S V T Q T	
AAATTTAATG CCTACCTGTC CCGTGTTCGC GAAGTGAACG GCATCGACCG CGGTGATGTG TCGAAAAAT TCACCGTTGA ACCGTCGGTC ACCCAGACCC	4200
L M N T M Q E S D F L T R I N I V P V S E M K G E K I G I G V T G	
TGATGAACAC CATCGAGGAG TCCTCTGACT TTCTGACCG CATCAATATT GTGCCGGTCA GCGAAATGAA AGGGGAAAAA ATTGGTATCG GTGTCAACCG	4300
S I A S T T D T A G G T E R Q P K D F S K L A S N K Y E C D Q I N	
CTCCATCGCC AGCACTACCG ACCTGCGCG TGGTACCGAG CGTCAGCGCA AGGACTTCTC GAAGTGGCG TCAACACAGT ACAGTACGCA CCAGATTAA	4400
F D F Y I R Y K T L D L W A R Y Q D F Q L R I R N A I I K R Q S L	
TTCGATTTT ATATCCGCTA CAAACCGCTG GACCTGTGGG CGCGTTATCA GGATTTCCAG CTCCGTATCC GTAACGCCAT TATCAACGCG CAGTCCCTTG	4500
D F I M A G F N G V K R A E T S D R S S N P M L Q D V A V G W L Q K	
ATTTCATCAT GCGCGGTTTT AACGCGCTGA AGCGTGGCGA AACCTCTGAC CGCAGCAGCA ATCCGATGTT GCAGGATGTG GCGGTGCGGT GGTCTGAGAA	4600
Y R N E A P A R V M S K V T D E E G R T T S E V I R V G K G G D Y	
ATACCGCAAT GAAGCACCAG CGCGCGTAT GAGCAAGGTC ACTGACGAGG AAGCCGCGAC CACCTCTGAG GTTATCCGCG TGGGTAAGGG CGGTGATTAT	4700
A S L D A L V M D A T N N L I E P W Y Q E D P D L V V I V G R Q L	
GCCAGCCTTG ATGCACTGCT GATGGATGCG ACCAACAACC TGATTGAACC GTGGTATCAG GAAGACCCCTG ACCTTGTGGT GATTGTGGGG CGTCAGCTAC	4800
L A D K Y F P I V N K E Q D N S E M L A A D V I I S Q K R I G N L P	
TGGCGGACAA GTATTTCCTC ATCGTCAACA AGGAGCAGGA CAACAGCGAA ATGCTGGCGC CTGACGTGAT CATCAGCCAG AAACGCAATG GTAACTTACC	4900
A V R V P Y F P A D A M L I T K L E N L S I Y Y M D D S H R R V I	
AGCGGTACCG GTCCCGTACT TCCCGCGCGA TGGATGCTC ATCAGCAAGC TGGAAACCT GTCCATCTAC TACATGGATG ACAGCCATCG CCGCGTGATT	5000
E E N P K L D R V E N Y E S M N I D Y V V E D Y A A G C L V E K I	
GAGGAAACCG CGAACTCGA CCGCGTGGAG AACTACGAGT CAATGAACAT TGATTACGTG GTGGAAGACT ACGCCGCGCG TTGTCTGGTG GAAAAATCA	5100
K V G D F S T P A K A T A E P G A O	
AGGTGCGTGA CTTCTCCACA CCGGCTAAGG CGACCGCAGA GCCGGGAGCG TAACCGATGA CGAGTCCCGC ACAGCGCCAC ATGATGCGGG TCTCGGCAGC	5200
M T A Q R E A A P L R H A T V Y E Q M L V K L A A D Q R T L K A I	
GATGAAACCG CAGCGGGAAG CCGCCCGCT GCGACATGCA ACTGCTATG GGTAAAGCTC GCGCGAGACC AGCGCACACT GAAAGCGATT	5300
Y S K E L K A A K K R E L L P F N L P N V N G V L E L G K G A O D	
TACTCAAAG AGCTGAAGGC CGCAAAAAA CGCGAATGCG TGCGTTCTG GTTCCGCTGG GTGAACGGCG TGCTGGAGCT GGGCAAAGT GCACAGGATG	5400
D I L M T V M L N R L D T G D I A G A L E I A R Y A L K Y G L T M P	
ACATTCTGAT GACGGTCATG CTGTGGCGTG TGGATACCG CGATATTGCC GGTGCGCTGG AGATTGCCCG TTATGCCCTG AAGTACGGTC TGACATGCC	5500
G K H R R T P P Y M F T E E V A L A A M R A H A A G E S V D T R L	
GGTAAACAC CGCGGTACCG CCGCGTACAT GTTACCCGAG GAGGTGAGCG TTGCGGCCAT GCGCGCTCAC GCTGCGGTG AGTCTGTGGA TACCCCGCTG	5600
L T E T L E L T A T A D M P D E V R A K L H K I T G L F L R D G G	
CTGACGAGA CCCTGAACT GACCGCCAGG GCTGACATGC CTGATGAAGT GCGCGCAAAG CTGCACAAAA TCACCGGTCT GTTCTGCGT GACGGTGGTG	5700
D A A G A L A H L Q R A T Q L D C Q A G V K K E I E R L E R E L K P	
ATGCCGCCGG TGCGGTGGCG CACCTGCAAC GTGCGACACA GCTGACATGT CAGCGAGCGC TCAAAAAAGA GATTGAACGA CTGAGGCGGG AGCTGAACCC	5800
K P E P O P K A A T R A P R K T R S V T P A K R G R P K K K A S O	
GAAGCGGAG CCGCAGCCCA AAGCGGCCAC CCGCGCCCGC CGTAAGACCC GAGAGCGTAC ACCGGCAAAA CGTGGACGCC CGAAAAAGAA AGCCGCTTAA	5900
CAACCGAATG CGCCCCGCG CAGGCGGCA CGCCGGTCTG TGACGGTGA TCACCTGACA CTGACCCGCC GTCCACCGCC CGACTTTTCA GAGTAGTCA	6000
M M T L I I P R K E A P V S G E G T V V I P Q P A G D E P V I K N T	
TGATGACGCT GATTATTCGG CGAAAGGAGG TCCCGTGTCT CGGTGAGGGT ACGGTGGTCA TCCCGCAACC GGCAGGCGAC GAGCCGGTGA TTAACACAC	6100
F F F P D I D P K R V R E R M R L E Q T V A P A R L R E A I K S G	
GTCTTTTTT CCGGATATCG ACCCGAAGCG CGTCCGGGAA CGTATGCGCC TTGAGCAGAC CGTGCCTCCC GCGCGTCTGC GTGAGGCCAT CAAGTCAGGC	6200
M A E T N A E L Y E Y R E Q K I A A G F T R L A D V P A D D I D G	
ATGGCTGAAA CGAATGCGGA GCTGTACGAG TACCGCAAC AGAAATTGC CGCCGGTTTT ACGCGTCTGG CTGACGTCCC GCGGACGAT ATCGACGGTG	6300
E S I K V F Y Y E R A V C A M A T A S L Y E R Y R G V D A S A K G D	
AAAGCATCAA GGTTTTTTAC TACGAGCGCG CCGTGTGTGC GATGCGGACC GCGTGGCTTT ATGAGCGTTA TCGCGGTGTG GATGCCAGTG CGAAAGCGGA	6400
K K A D S I D S T I D E L W R D M R N A V A R I Q G K P R C I V S	
CAAGAAGGCT GACAGCATTG ACAGCACCAT TGATGAGCTG TGGCGGGATA TGCGTGGGCG GGTGGCGCGC ATCCAGGGCA AGCCGCGCTG CATCGTAGT	6500
Q I O O	
CAATCTGAT GAAGACCTTT GCGCTACAGG GCGACACCTC GACGCCATTT6550	

Figure 2. Nucleotide sequence of the P2 DP/C region. Numbers at the right indicate the nucleotide position with respect to *cosL*. The inferred amino acid sequences of P2 proteins are given in single letter code, with diamonds symbolizing stop codons. Initiating methionine codons are underlined for clarity, and the directions of translation are indicated by small arrows. Asterisks indicate potential ribosome binding sites. For genes *Q* and *P*, the complementary (non-coding) strand is shown. Additional features: *cos*, bases forming the cohesive ends of linear P2 DNA; T_Q —, putative transcriptional terminator following gene *Q*; << and >>, start sites for P_P and P_O transcripts, respectively. P_P differs by one base from previously published sequence (22; D. Anders and G. Christie, unpublished). Underlined residues indicate potential DNA binding domains.

Table 1. Location, DNA sequence change and ORF assignment of DP/C mutations

Mutant allele	Interval [‡]	Base change (bp)*	Codon change (n) [§]	ORF
<i>Qam34</i>	<i>Bam</i> HI - <i>Ase</i> I (bp 262 - 718)	C → A (653)	TCG → TAG (190)	1
<i>Pam24</i>	<i>Sal</i> I - <i>Bam</i> HI (bp 1825 - 2221)	C → T (2210)	CAG → TAG (263)	2
<i>Pam137</i>	<i>Sal</i> I - <i>Bam</i> HI (bp 1825 - 2221)	C → T (2210)	CAG → TAG (263)	2
<i>Pam253</i>	<i>Bgl</i> II - <i>Bsp</i> HI (bp 2689 - 2994)	C → T (2948)	CAG → TAG (16)	2
<i>Oam279</i>	right of <i>Eco</i> RI (right of bp 3585)	C → T (3878)	CAG → TAG (238)	3
<i>Mam32</i>	<i>Taq</i> I - <i>Hinc</i> II (bp 5745 - 5898)	C → T (5751)	CAG → TAG (199)	5
<i>Mts52</i>	<i>Dde</i> I - <i>Apa</i> LI (bp 5127 - 5393)	G → A (5361)	GTG → ATG (69)	5
<i>Lam9</i>	<i>Eco</i> RV - <i>Eco</i> RV (bp 6116 - 6290)	C → T (6240)	CAG → TAG (81)	6
<i>Lam79</i>	left of <i>Eco</i> RV (left of bp 6116)	A → T (6024)	AAG → TAG (9)	6

[‡] The wild-type allele could be rescued from this region.

* Position on coding strand

[§] Number of the affected codon. Amber mutants are presumed to make a truncated polypeptide of length (n-1) residues.

tentatively made by correlating the physical location of each with the P2 genetic map (Fig. 1). Thus, ORFs 1, 5 and 6 were assigned to genes *Q*, *M* and *L*, respectively.

To confirm these ORF-gene assignments, we sequenced amber mutations in genes *Q*, *P*, *O*, *M* and *L*, and a *ts* mutation in gene *M*. The technique of marker rescue was used to localize the mutations to short intervals (see Materials and Methods). The sequence of mutant and wild-type phage DNA for these intervals was determined in parallel. The results obtained (Table 1) confirmed the identity of the assigned ORFs. Each mutant interval differed by a single nucleotide from the wild-type sequence. All amber mutations caused in-frame stop codons that truncated the ORF predicted to contain the assigned

gene. Additionally, the sequence of *Mts52* indicated an amino acid change in ORF 5.

Translation products of the DNA packaging and capsid genes

The amino acid sequences of polypeptides encoded by DP/C genes were deduced from the DNA sequence, and are shown in Fig. 2. Since genes *P* and *Q* are encoded by the complementary DNA strand, the amino acid sequence of their gene products should be interpreted from right to left. Also shown are tetranucleotides (* in Fig. 2), GGAG (genes *Q* and *M*), GAGG (genes *P* and *L*) or AGGA (genes *O* and *N*), which are complementary to the 3' end of 16S rRNA and probably represent ribosome binding sites (37).

Table 2 lists features of the DP/C polypeptides. All polypeptides contained about the same percentage of charged residues (26–31%). The isoelectric points of four predicted acidic proteins (gpP, gpO, gpN, gpL), while gpQ and gpM were basic polypeptides. The hydrophobicity profile of each polypeptide was determined by the method of Kyte and Doolittle (38). The profile of each indicated a soluble protein lacking membrane spanning domains.

Molecular masses predicted for DP/C proteins were compared with those determined experimentally by SDS-PAGE (5,7,39).

Table 2. Predicted and observed characteristics of the DP/C proteins of P2.

Characteristic	polypeptide ^a					
	Q	P	O	N	M	L
Amino acids	344	590	284	357	247	169
% charged amino acids ^b	27%	26%	30%	28%	31%	31%
Isoelectric point (pH)	9.2	6.7	5.0	4.8	10.9	5.5
% hydrophilic ^c	58%	55%	65%	58%	60%	60%
% hydrophobic ^c	32%	35%	26%	32%	31%	34%
Mass, predicted (kDa)	39.1	66.6	31.4	40.2	27.5	19.0
Mass, observed (kDa)	39	65	30	44	28	#

a Full length translation product.

b Charged amino acids are D, E, H, R and K.

c Percent of total residues either hydrophilic or hydrophobic, as predicted by the method of Kyte and Doolittle (38).

not previously detected

The observed masses of gpP, gpO, gpN, and gpM from phage-infected cells agreed closely with sequence predictions (Table 2). Overexpression of *M* and *P* from plasmids also yielded polypeptides of the predicted sizes having the mobility of their phage-encoded counterparts (R. Ziernann, unpublished data). The sequence of the *L* gene predicted a protein of 168 or 169 amino acids having a molecular mass of 18.9–19.0 kDa. The sizes of *L* and gpL are uncertain because the gene initiates with tandem Met codons and either or both starts could be used. The product of gene *L* has not been identified by SDS/polyacrylamide gel electrophoresis. Gene *Q* should encode a protein of 39.1 kDa. A polypeptide reported to be gpQ had a mass of just 32 kDa, which is 18% smaller than is indicated by the gene (39). Inspection of ORF 1 sequence revealed a GTG codon (codon 70) where initiation of translation could yield a polypeptide of about 32 kDa. However, there is no potential ribosome binding site in its vicinity, so its use is unlikely.

Positive identification of Q protein

To positively identify the *Q* gene product, we expressed *Q* in *E. coli* from pNL93Q (Fig. 3A). In the presence of IPTG, C-1a/pNL93Q made a protein of approximately 39 kDa, which could be detected on an SDS/polyacrylamide gel stained with Coomassie Blue (lane 1, Fig. 3B). The apparent mass of the new protein agreed closely with the size predicted for gpQ (39 kDa) from the DNA sequence. The 39 kDa protein was not detectable when cells were uninduced (lane 2) nor when cells carried pUHE21-2, a control plasmid consisting of vector but no *Q* (lanes 3 & 4). To determine if the 39 kDa protein made from pNL93Q could complement a P2 *Q* mutant, a plate complementation test was performed (Materials and Methods). The efficiency of plating by *Qam34* on C-1a (a *sup*^o host) carrying pNL93Q and on C-1757 (a *supD* host) were equal. IPTG stimulated growth of the phage only if C-1a/pNL93Q was the host. Tests of progeny genotype showed that phage growth was due to complementation and not to marker rescue (not shown). No progeny were obtained with strain C-1a/pUHE21-2. Thus, the 39 kDa protein encoded by pNL93Q must be gpQ.

Is the 32 kDa polypeptide in ref. 39 also a product of *Q*? In Fig. 3B, we document the primary translation product of *Q* (gpQ) and show that it migrates according to its DNA-predicted mass. However, because gpQ was overproduced in uninfected *E. coli*,

Figure 3. Identification of the Q gene product. (A) Schematic diagram of the Q expression plasmid, pNL93Q. The Q gene is represented by the thick black arc; P2 sequence is delimited by restriction sites *Bst*XI (5') and *Bss*HII (3'), which were inactivated during cloning. The phage T7 A1 promoter (P_{A1}), *lac*O and *Eco*RI and *Hind*III restriction sites are from vector pUHE21-2. (B) Polypeptides made in C-1a/pRG1 containing pNL93Q (lanes 1 & 2) or pUHE21-2 (lanes 3 & 4), on a 10% SDS/polyacrylamide gel, stained with Coomassie Blue. Cells were induced 3 h with 1 mM IPTG (lanes 1 & 3) or uninduced (lanes 2 & 4). Numbers indicate positions of size standards. The Q gene predicts a 39.1 kDa product (arrow).

any potential processing or modification of this polypeptide by phage-induced factors could not have ensued. Additional experimentation will reveal whether gpQ and the 32 kDa polypeptide are related.

Sequence and protein comparisons

The six DP/C genes and their derived protein sequences were compared to entries in the GenBank, EMBL and SWISS-PROT databases. We found one nucleotide sequence in GenBank having significant (65.8%) homology to the Q gene and the final third of P (P2 bp 187–1886). The matching sequence was part of an approximately 30 kb *E. coli* retron element, Ec67, identified in the laboratory of Dr. Masayori Inouye (Robert Wood Johnson Medical School, Piscataway, N. J.). The homology score would seem to indicate that the retron resides in a defective P2 prophage or the prophage of another P2 family member. Linkage of a retron element to an integrated phage genome is believed to facilitate retron dispersal (c.f. ref. 40).

The amino acid sequences of the terminase subunits, gpM and gpP, were scanned for residues responsible for the DNA binding activity of this enzyme (8,9). Neither sequence predicted site-specific DNA binding motifs (41,42,43), yet both contained stretches rich in basic amino acids (underlined in Fig. 2). Thirteen of the final 34 amino acids in gpM were either Arg or Lys (1 Glu). The basic region in gpP (residues 116–126) had Arg or Lys at 7 of 11 positions. Within the basic stretch of gpM were a KPK motif (KPKP) and a GRP motif (RKTPAKRGRPKKK); both are indicative of proteins that bind DNA with low sequence specificity (see ref. 44 for examples of these DNA binding proteins and motifs). P protein also contained a KPK motif (KPKK). The GRP region in gpM was strikingly similar to one found in a yeast DNA-binding protein, SWI⁵ [KRSPRK-RGRPRK (45)]. The importance of these residues in gpM and gpP for interaction with DNA is presently unknown but DNA binding with low sequence specificity is a property of these and analogous phage proteins (2,9). Whether these amino acids contribute to binding can be investigated using site-specific oligodeoxynucleotide-directed mutagenesis and reverse genetics to generate mutant proteins altered at these positions.

Figure 4. Putative A-type and B-type ATP binding motifs in gpP. Comparison is to the A-type sequence of Walker et al. (46) and to the B-type sequence of Guo et al. (47). Symbols: ↓ and bracket, point out residues that are highly conserved or generally hydrophobic, respectively, in ATP binding motifs; ●, amino acids in gpP fitting the A or B-type consensus; ○, amino acids in gpP that are weakly conserved; +, hydrophobic residues.

In the P protein sequence, we found both A-type and B-type ATP binding motifs (Fig. 4). Inasmuch as the predicted secondary structure at the site of the A-type motif is β -sheet/loop/ β -sheet rather than the more generally observed β -sheet/loop/ α -helix (47,48), it might not represent a true ATP binding site. On the other hand, both B-type motifs occurred in regions predicted to fit the β -sheet consensus, and so are probably genuine. The larger ATPase subunit of many phage terminases [e.g. gp16 of ϕ 29, gp19 of T7, gpA of λ , gp17 of T4, gp2 of P22] also contain ATP-binding motifs (47,49).

DISCUSSION

In this study, we report the sequence of a 6550 bp interval containing the DP/C genes of phage P2 and relate important features of the region. The six genetically identified DP/C genes (Q, P, O, N, M and L) were located and their boundaries were precisely defined. The predicted amino acid sequences and molecular masses of regionally encoded proteins are presented. The DP/C region encodes the phage's major capsid protein (gpN), a scaffolding protein (gpO), its terminase (gpM, gpP), an additional DNA packaging protein (gpQ) and a head completion/stabilization protein (gpL). Noticeably absent from this list is the portal vertex protein, since no P2 gene has been assigned this function.

By deduction, gpQ is a candidate for the capsid-associated portal vertex protein. With respect to mass, Q protein (39.1 kDa) is in the same class as the portal proteins of phages T4, λ , ϕ 29, T3 and P22, which range in size from about 40–80 kDa (1). Although no sequence homology between these portal proteins has been found, all adopt a remarkably similar ring-like structure having 6 or 12-fold rotational symmetry and a central hole of 3–4 nm, through which DNA is believed to pass into a capsid (1). Mutations in portal protein genes typically affect capsid synthesis and DNA packaging. Indeed, *Qam34* mutants make proheads that cannot be converted to capsids (5,9) and they can neither mature nor package their DNA (7,8). An activity from *Q*⁺ extracts capable of rescuing the packaging defect of *Qam34* extracts, co-purifies with empty heads (D. Bowden, 1978, Ph.D. thesis, University of California, Berkeley, California). Thus, in the case of P2, Q mutations cause the types of defects expected if Q encoded its portal protein. Confirmation that gpQ associates with capsids and participates in head assembly would further

support this. Overexpression of the *Q* gene and identification of *Q* protein, work presented here, should pave the way for its purification and a better understanding of its contribution to phage assembly.

One goal of this study was to search in the DP/C region for genes not represented by mutations, e.g. those encoding the minor capsid proteins, h3, h4, h6 and h7 (5). However, sequence analysis revealed only the six essential genes. Several short ORFs nested within these genes were also found but none possessed significant length or optimal codon usage. The region immediately to the right of the sequenced interval contains the lysis gene of P2 (G. Christie, unpublished results). These data suggest that all DP/C genes have been accounted for. A strategy that P2 uses to extend its DNA sequence information is post-translational processing of gpN into multiple forms. Phage λ employs more elaborate polypeptide cleavage plus fusion reactions to generate certain of its capsid proteins (50). Thus, it is possible that the minor proteins in P2 capsids are derivatives of identified translation products. Since the DNA sequence of the DP/C region is now complete, the loci encoding these proteins can be identified when their amino acid sequences become available.

DP/C proteins of P2 lack homology to similar proteins of other phages. The absence of homology between phage proteins with analogous functions is consistent with current knowledge of bacterial, plant and animal viruses. The possibility that similar tertiary structures are adopted by the functionally analogous proteins of unrelated phages needs to be fully explored.

ACKNOWLEDGEMENTS

This publication is dedicated to Paul Berg on the occasion of his 65th birthday! The authors are grateful to Ole Jrgen Marvik for helpful discussions and Doug Anders for communicating unpublished results. Undergraduate student Mehdi Nosrati participated in these experiments. The technical assistance of Virginia Barreiro and Becky Bartlett is gratefully acknowledged. P2-specific sequencing primers were made at: the Department of Oncology, Karolinska Hospital, Radiumhemmet, Sweden [10.5R, 12.0L]; the Department of Biology and Biotechnology Center, University of Oslo, Oslo, Norway [15.4L]; and the Nucleic Acids Core Facility of the Massey Cancer Center, Virginia Commonwealth University, Richmond, VA, USA [19.2L, 19.4R, 19.8L, 20.5L]. These studies were supported by Grant 72 from the Swedish Medical Research Council (to E.H.-L.) and Public Health Service Grants AI08722 (to R.L.C.) and GM34651 (to G.E.C.) from the National Institutes of Health.

REFERENCES

- Casjens, S. and Hendrix, R. (1988) In Calendar, R. (ed.) *The Bacteriophages*, Vol. 2. Plenum Publishing Corp., New York, pp. 15–91.
- Black, L.W. (1988) In Calendar, R. (ed.) *The Bacteriophages*, Vol. 2. Plenum Publishing Corp., New York, pp. 321–373.
- Ziermann, R. and Calendar, R. (1990) *Gene*, **96**, 9–15.
- Bertani, L.E. and Six, E.W. (1988) In Calendar, R. (ed.) *The Bacteriophages*, Vol. 2. Plenum Publishing Corp., New York, pp. 73–143.
- Lengyel, J.A., Goldstein, R.N., Marsh, M., Sunshine, M.G. and Calendar, R. (1973) *Virology*, **53**, 1–23.
- Six, E.W., Sunshine, M.G., Williams, J., Haggard-Ljungquist, E. and Lindqvist, B. (1991) *Virology*, **182**, 34–46.
- Pruss, G.J. and Calendar, R. (1978) *Virology*, **86**, 454–467.
- Bowden, D.W. and Calendar, R. (1979) *J. Mol. Biol.*, **129**, 1–18.
- Bowden, D.W. and Modrich, P. (1985) *J. Biol. Chem.*, **260**, 6999–7007.
- Casjens, S. (1974) *J. Mol. Biol.*, **90**, 1–23.
- Pruss, G., Wang, J.C. and Calendar, R. (1975) *J. Mol. Biol.*, **98**, 465–478.
- Inman, R.B., Schns, M., Simon, L.D., Six, E.W. and Walker, D.H. (1971) *Virology*, **44**, 67–72.
- Shore, D., Deh, G., Tsipis, J. and Goldstein, R. (1978) *Proc. Natl. Acad. Sci.*, **75**, 400–404.
- Kahn, M.L., Ziermann, R., Deh, G., Ow, D.W., Sunshine, M.G. and Calendar, R. (1991) *Methods Enzymol.*, **204**, 264–280.
- Sasaki, I. and Bertani, G. (1965) *J. Gen. Microbiol.*, **40**, 365–376.
- Sunshine, M.G., Thorn, M., Gibbs, W., Calendar, R. and Kelly, B. (1971) *Virology*, **46**, 691–702.
- Wiman, M., Bertani, G., Kelly, B. and Sasaki, I. (1970) *Molec. Gen. Genet.*, **107**, 1–31.
- Bertani, G., Ljungqvist, E., Jagusztyn-Krynicka, K. and Jupp, S. (1978) *J. Gen. Virol.*, **38**, 251–261.
- Bertani, L.E. (1960) *Virology*, **12**, 553–569.
- Lindahl, G. (1969) *Virology*, **39**, 839–860.
- Lindahl, G. (1971) *Virology*, **46**, 620–633.
- Christie, G.E. and Calendar, R. (1983) *J. Mol. Biol.*, **167**, 773–790.
- Yanisch-Perron, C., Vieira, J. and Messing, J. (1985) *Gene*, **33**, 103–119.
- Sadler, J.R., Tecklenburg, M. and Betz, J.L. (1980) *Gene*, **8**, 279–300.
- Bolivar, F., Rodriguez, R.L., Greene, P.J., Betlach, M.C., Heynecker, H.L., Boyer, H.W., Crosa, J.H. and Falkow, S. (1977) *Gene*, **2**, 95–113.
- Norlander, J., Kempe, T. and Messing, J. (1983) *Gene*, **26**, 101–106.
- Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual* (Second Edition). Cold Spring Harbor Laboratory Press, Cold Spring Harbor.
- Griffin IV, T.J. and Kolodner, R.D. (1990) *J. Bacteriol.*, **172**, 6291–6299.
- Sanger, F., Nicklen, S. and Coulson, A.R. (1977) *Proc. Natl. Acad. Sci.*, **74**, 5463–5467.
- Maxam, A.M. and Gilbert, W. (1980) *Methods Enzymol.*, **65**, 499–560.
- Kraft, R., Tardiff, J., Krauter, K.S., Leinwand, L.A. (1988) *BioTechniques*, **6**, 544–547.
- Pearson, W.R. and Lipman, D.J. (1988) *Proc. Natl. Acad. Sci.*, **85**, 2444–2448.
- Haggrd-Ljungqvist, E. (1990) In O'Brien, S.J. (ed.), *Genetic Maps-Locus Maps of Complex Genomes* (Fifth Edition). Cold Spring Harbor Laboratory Press, Cold Spring Harbor, pp. 1.63–1.69.
- Temple, L.M., Forsburg, S.L., Calendar, R. and Christie, G.E. (1991) *Virology*, **181**, 353–358.
- Sharp, P.M. and Li, W.-H. (1986) *Nucleic Acids Res.*, **14**, 7737–7749.
- Skalka, A. and Hanson, P. (1972) *J. Virology*, **9**, 583–593.
- Shine, J. and Dalgarno, L. (1974) *Proc. Natl. Acad. Sci.*, **71**, 1342–1346.
- Kyte, J. and Doolittle, R.F. (1982) *J. Mol. Biol.*, **157**, 105–132.
- Christie, G.E. and Calendar, R. (1985) *J. Mol. Biol.*, **181**, 373–382.
- Inouye, S., Sunshine, M.G., Six, E.W. and Inouye, M. (1991) *Science*, **252**, 969–971.
- Brennan, R.G. and Matthews, B.W. (1989) *J. Biol. Chem.*, **264**, 1903–1906.
- Vallee, B.L., Coleman, J.E. and Auld, D.S. (1991) *Proc. Natl. Acad. Sci.*, **88**, 999–1003.
- Landschulz, W.H., Johnson, P.F. and McKnight, S.L. (1988) *Science*, **240**, 1759–1764.
- Churchill, M.E.A. and Travers, A.A. (1991) *TIBS*, **16**, 92–97.
- Stillman, D.J., Bankier, A.T., Seddon, A., Groenhout, E.G. and Nasmyth, K.A. (1988) *EMBO J.*, **7**, 485–494.
- Walker, J.E., Saraste, M., Runswick, M.J. and Gay, N.J. (1982) *EMBO J.*, **1**, 945–951.
- Guo, P., Peterson, C. and Anderson, D. (1987) *J. Mol. Biol.*, **197**, 229–236.
- Saraste, M., Sibbald, P.R. and Wittinghofer, A. (1990) *TIBS*, **15**, 430–434.
- Eppler, K., Wyckoff, E., Goates, J., Parr, R. and Casjens, S. (1991) *Virology*, **183**, 519–538.
- Hendrix, R.W. and Casjens, S.R. (1974) *Proc. Natl. Acad. Sci.*, **71**, 1451–1455.