

Validated near-atomic resolution structure of bacteriophage epsilon15 derived from cryo-EM and modeling

Matthew L. Baker^a, Corey F. Hryc^{a,b}, Qinfen Zhang^{a,c}, Weimin Wu^{d,1}, Joanita Jakana^a, Cameron Haase-Pettingell^e, Pavel V. Afonine^f, Paul D. Adams^f, Jonathan A. King^e, Wen Jiang^d, and Wah Chiu^{a,b,2}

^aNational Center for Macromolecular Imaging, Verna and Marrs McLean Department of Biochemistry and Molecular Biology and ^bProgram in Structural and Computational Biology and Molecular Biophysics, Baylor College of Medicine, Houston, TX 77030; ^cState Key Lab of Biocontrol, School of Life Sciences, Sun Yat-sen University, Guangzhou 510275, China; ^dMarkey Center for Structural Biology, Department of Biological Sciences, Purdue University, West Lafayette, IN 47907; ^eDepartment of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139; and ^fLawrence Berkeley National Laboratory, Berkeley, CA 94720

Contributed by Wah Chiu, May 30, 2013 (sent for review April 4, 2013)

High-resolution structures of viruses have made important contributions to modern structural biology. Bacteriophages, the most diverse and abundant organisms on earth, replicate and infect all bacteria and archaea, making them excellent potential alternatives to antibiotics and therapies for multidrug-resistant bacteria. Here, we improved upon our previous electron cryomicroscopy structure of *Salmonella* bacteriophage epsilon15, achieving a resolution sufficient to determine the tertiary structures of both gp7 and gp10 protein subunits that form the T = 7 icosahedral lattice. This study utilizes recently established best practice for near-atomic to high-resolution (3–5 Å) electron cryomicroscopy data evaluation. The resolution and reliability of the density map were cross-validated by multiple reconstructions from truly independent data sets, whereas the models of the individual protein subunits were validated adopting the best practices from X-ray crystallography. Some sidechain densities are clearly resolved and show the subunit–subunit interactions within and across the capsomeres that are required to stabilize the virus. The presence of the canonical phage and jellyroll viral protein folds, gp7 and gp10, respectively, in the same virus suggests that epsilon15 may have emerged more recently relative to other bacteriophages.

validation | Pathwalker | PHENIX | EMAN | gold standard

Tailed double-stranded DNA (dsDNA) bacteriophages, the most abundant life forms in the biosphere (1), have become valuable systems for studying not only virus structures and assembly (2–5) but also protein folding (6). Structural and genomic studies have suggested that these tailed bacteriophages are ancient and were present billions of years ago before cellular life diverged into the three domain systems: Eukarya, Eubacteria, and Archaea (7). In addition, bacteriophages also have played critical roles in gene transfer in bacterial host cells, contributing to pathogenicity and population ecology of environmental microorganisms (7).

Both X-ray crystallography and single-particle electron cryomicroscopy (cryo-EM) have been used extensively to determine the structures of various viruses and bacteriophages (3, 4, 8–12). The first near-atomic resolution cryo-EM structure of an infectious bacteriophage was ε15, a T = 7 icosahedral phage belonging to the *Podoviridae* family of *Caudovirales* and infecting *Salmonella anatum* (4). This reported map was barely sufficient to trace the C α backbone of gp7, the 335 amino acid capsid protein. In addition, analysis of that map led to the discovery of another protein, gp10 (111 aa), decorating the outer surface of the capsid shell. However, lower resolvability in this region of the density map precluded a C α backbone trace. Although capsid decoration proteins had previously been seen in other bacteriophages and viruses (13–18), no related structures could be identified for gp10.

The scope of our current work seeks to construct and validate a full-atom model for ε15, including both the gp7 and gp10 capsid proteins. To accomplish this, we devised a strategy of acquiring

better image data, improving both image-processing and model-building protocols, more rigorously measuring map resolution and adopting best practice in model refinement as established in X-ray crystallography. The final ε15 full-atom model clearly demonstrates an intricate set of interactions between the proteins, both within and across capsomeres, contributing to overall viral stability, as well as illuminating new potential interactions during its assembly and maturation.

Results and Discussion

Cryo-EM Data and Map. In the current investigation, we acquired a unique cryo-EM dataset of ε15 (Fig. 1A) using a different electron microscope than the one used in our previous study (4). A substantially larger fraction of high-quality images was recorded with the current electron microscope at liquid nitrogen specimen temperature over the previous instrument, operated at liquid helium temperature. The power spectrum of the boxed-out particles generally shows signals out to 4–5 Å resolution (Fig. S1). The reconstruction was based on a recently implemented protocol (19) (*Materials and Methods*). We adopted the gold standard definition for the resolution estimate (20), whereby we split the particle images into two subsets at the onset of image processing, reconstructed the datasets individually, and combined after determining the resolution estimate (Fig. 1B). Independent initial models were built de novo and used for the subsequent particle refinements in each of the two subsets of particle images (*Materials and Methods*) (Fig. S2). The Fourier Shell Correlation (FSC) between the two independently determined reconstructions was computed and indicated a resolution 4.5 Å using the 0.143 threshold (21) for the combined dataset (Fig. 1C). This map resolution measure differs from the commonly used approach as in the previous report (4), where two maps were computed with particle images split after all of the particles' orientations were determined using a single initial model and common models for all subsequent refinements. This conventional procedure has now been recognized to be a substandard measure of the map resolution (20, 22, 23). Additionally,

Author contributions: W.C. designed research; M.L.B., C.F.H., Q.Z., W.W., J.J., C.H.-P., and W.J. performed research; M.L.B., C.F.H., P.V.A., and P.D.A. contributed new reagents/analytic tools; M.L.B., C.F.H., W.J., and W.C. analyzed data; and M.L.B., C.F.H., J.A.K., W.J., and W.C. wrote the paper.

The authors declare no conflict of interest.

Data deposition: The atomic coordinates and density maps have been deposited in the Protein Data Bank, www.pdb.org (PDB ID code 3J40) and Electron Microscopy Data Bank, www.emdatabank.org (EMDB ID code 5678).

¹Present address: Laboratory of Structural Biology Research, National Institute of Arthritis and Musculoskeletal and Skin Diseases, National Institutes of Health, Bethesda, MD 20892-8025.

²To whom correspondence should be addressed. E-mail: wah@bcm.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1309947110/-DCSupplemental.

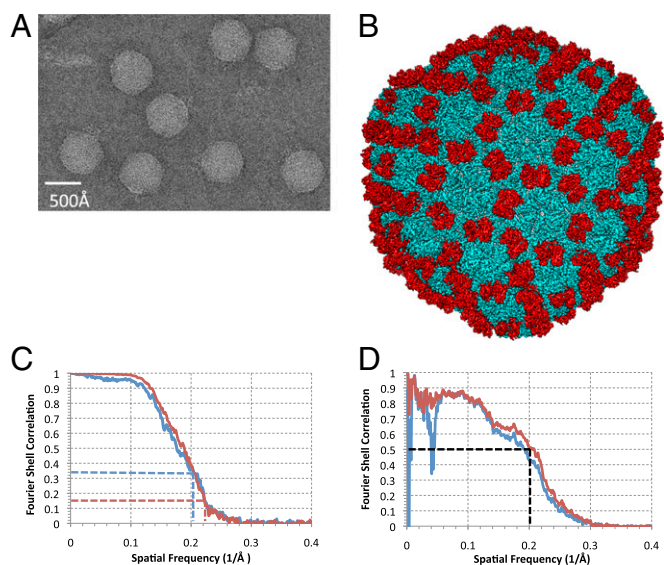


Fig. 1. Structure of $\epsilon 15$. (A) A typical image of $\epsilon 15$ 300 kV at 50,000 \times magnification from the JEM3200FSC electron cryomicroscope. (B) The complete reconstruction of $\epsilon 15$ reveals the capsid protein gp7 (cyan) and the staple protein gp10 (red). (C) The FSC between the two independent datasets used for our current reconstruction is shown in red. The FSC between the previous $\epsilon 15$ density map (EMDB ID code 5003) and our current $\epsilon 15$ density map (EMDB ID code 5678) is shown in blue. The dashed lines correspond to the gold standard criterion for resolution estimate. (D) A comparison of density simulated at 3.5 Å resolution from our all-atom model versus our current density map is shown (red). Application of an inner mask removes the low-resolution dip, which can be attributed to the DNA (blue).

a model for the entire virus was transformed into a 3.5 Å resolution density map, which was compared with the raw cryo-EM density map. Using the 0.5 FSC criterion, a more appropriate measure in map-to-model comparisons (21), estimates their structure match up to ~ 5 Å on average (Fig. 1D).

In line with the new resolution measure and the overall structural features, we computed a FSC between our current map [Electron Microscopy Data Bank (EMDB) ID code 5678] and the previously published map (EMDB ID code 5003). (Fig. 1C). At FSC = 0.333 (21), these two maps are structurally similar up to ~ 4.8 Å on average. This measure suggests the resolution of the previous map (claimed to be 4.5 Å based on 0.5 criterion for FSC computed from maps with split data after they were refined with a conventional common model approach) (4) was slightly overestimated. Nevertheless, the resolvable features in many places within the two density maps are quite comparable (Fig. S3). The resolvability of structural features like the β -strand separation and loop connectivity in the current map is a clear improvement over the previous map, particularly in gp10 (Fig. S3 C and D). Furthermore, as shown in Figs. 2 and 3, the new density map is more uniformly resolved throughout, improving the interpretability of the structural features and the accuracy of the model building.

Model Building. In X-ray crystallography, model building in 4–5 Å resolution density maps is considered to be particularly challenging (24). However, in our previous study of $\epsilon 15$, we developed a manual modeling approach to trace the C α chain of gp7 (4, 25). Briefly, the presence of five distinct helices in gp7 was sufficient to anchor a sequence to structure correspondence, from which we were able to trace the C α backbone aided by a density skeleton connectivity (26, 27). However, in the case of gp10, model building was not possible due to limited resolvability of the structural features. Since our initial success in building the gp7 model, our de novo modeling

protocol has been enhanced with improved algorithms and automation capability (28–30).

The enhanced resolvability along with improved modeling methods allowed us to generate complete all-atom models for gp7 and gp10 from our new density map (Figs. 2 and 3). As illustrated in Fig. S4, we used an iterative model refinement process using ROSETTA (31), COOT (32), and PHENIX (33) to refine mainchain and sidechain fit to density, while simultaneously improving model geometry, bond distances, and inter/intrasubunit interfaces (*SI Materials and Methods*). Best practice procedures from X-ray crystallography were adopted to improve and validate our models (Table 1) based on recent recommendations by the EMDB model validation taskforce (34). Although fully refined and validated, it should be noted that density at this resolution does not provide “perfect” restraints and thus potential errors (i.e., imperfect H-bonding, registry errors, etc.) may be present (24).

Coat Protein, gp7. For constructing and validating the gp7 model, we began with our original models of the seven gp7 subunits [Protein Data Bank (PDB) ID code 3C5B] constructed from the previous $\epsilon 15$ density map (4). We first fit these models to our new density map (Fig. 1B), segmented the subunits in the asymmetric unit, and computed an average gp7 density map (equivalent to Non-Crystallographic Symmetry (NCS) averaging in crystallography) with the six hexon subunits (Fig. 2A). We used our Pathwalker protocol (28) to determine various possible topologies of gp7 constrained by our averaged gp7 subunit density. Our results showed two different populations of models. One model was equivalent to our previous C α trace, whereas the other model had a strand swap in the A-domain and a slight reorientation of the C terminus. Comparing both models to our latest density map, the new gp7 model, with the strand swap, agreed more favorably with the density (Fig. S3 A and B). Seven independent all-atom models were subsequently constructed for the corresponding subunits in the asymmetric unit and refined using this initial C α trace

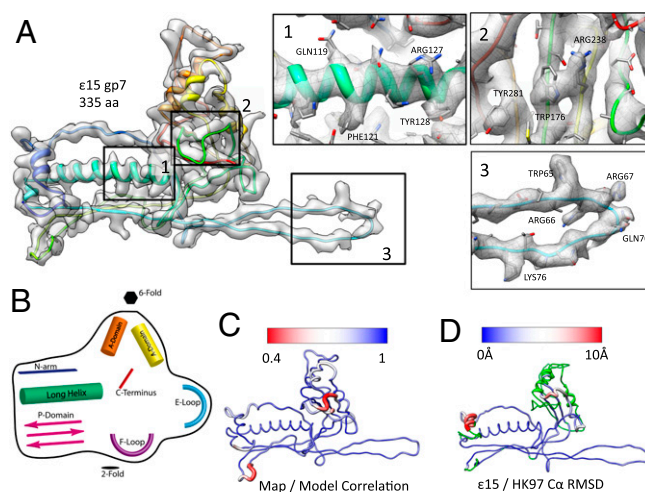


Fig. 2. Structure of the capsid protein, gp7. (A) An averaged gp7 density map is superimposed on the gp7 (Chain C) molecular model (rainbow representation, N–C terminus). Insets (1–3) show zoomed-in views of the typical regions with the unaveraged density and corresponding model of one gp7 subunit (Chain C). (B) A cartoon diagram illustrates the structural domains of gp7 shared with other tailed dsDNA bacteriophage (Fig. S7). (C) Correlation values were computed between the map and model on a per amino acid level, then averaged for all seven subunits in the asymmetric unit. Regions shown in blue had a strong correlation between the map and model, as opposed to regions in red, which had a weak correlation between the map and model. (D) A gp7 model is colored based on the C α rmsd between the capsid proteins from $\epsilon 15$ and HK97. Variations range from low to high rmsd (blue to red); green regions of the model were not used in computing the rmsd.

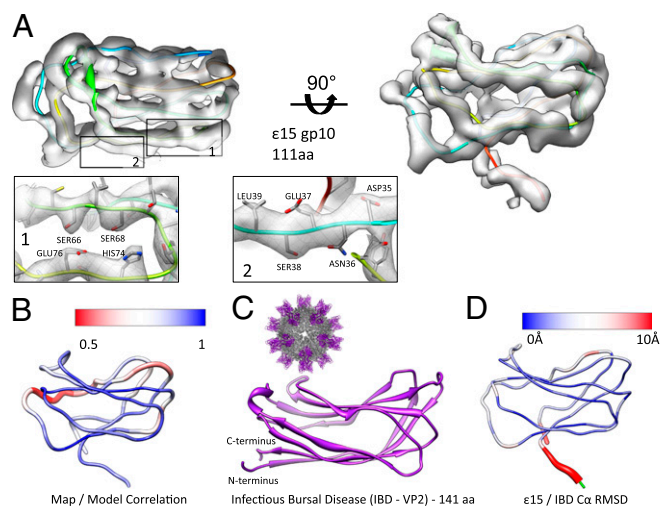


Fig. 3. Structure of the staple protein, gp10. (A) An averaged gp10 density map is shown in two views with the corresponding model (Chain C in rainbow representation). Individual sections of the model are highlighted (insets 1 and 2). (B) Correlation values were computed between the map and model on a per amino acid basis, then averaged for all seven subunits in the asymmetric unit. Regions shown in blue had a strong correlation between the map and model, as opposed to the red regions, which had a weak correlation between the map and model. (C) VP2 of infectious bursal disease virus (PDB ID code 2DF7) shows a similar jellyroll fold with four strands on the top and bottom. A full view of the infectious bursal disease virus is shown (Inset, Upper Left). (D) The gp10 model is colored based on α rmsd between gp10 and the corresponding domain of 2DF7. The largest differences are shown in red. Residues highlighted in green were not used in computing the rmsd.

(Fig. 2A, Fig. S5). In comparison with the previous model of gp7, the overall model structures are similar with a α rmsd of 2.33 Å between the two models over 54% (182 aa) of the protein's structure. The new models have considerably improved fit to the density and are geometrically favorable (Table 1, Fig. S5A).

Several distinct structural features are clearly observed in the model of gp7 and are listed as follows (from N to C termini): N-arm, F-loop, E-loop, P-domain, the long α -helix, A-domain, and the C terminus (Fig. 2B, Fig. S5B). Superimposing the seven subunit models revealed nearly identical structures except for variations at the E-loop and N-arm (Fig. S5B). Furthermore, the penton subunit shows additional variation at the penton opening, with a loop in the A-domain folding inward, narrowing the opening at the penton.

To examine the agreement between each subunit model and the corresponding density, we visualized sidechain placement (Fig. 2A) and calculated a map–model correlation per amino acid as exemplified in one subunit (Fig. 2C, Fig. S5C). In all but three

Table 1. Model statistics for 3,122 residues in 14 chains (one asymmetric unit) in ϵ 15

Statistical measure	Score	MolProbity summary*
Correlation coefficient	0.83	—
Clash score (all atoms)	19.07	96th percentile
Poor rotamers	700	27%
Ramachandran outliers	421	14%
Ramachandran favored	1,843	60%
MolProbity score	3.78	21st percentile
C β deviations	59	2%
Bad backbone bonds	0	0%
Bad backbone angles	11	0.07%

*Percentiles based on structures at similar resolution in PDB.

small regions, the average correlation with the density was above 0.70, although it should be noted that the main source of the low correlation was not in the overall mainchain trace but the absence of well-resolved sidechain density. The poorly correlated regions were primarily located in the A-domain and composed of one strand (155–159) and two loops (246–256 and 267–276) on either end of one helix (Fig. S5C). These regions are surface exposed and likely to be flexible.

The overall structure of gp7 resembles the canonical bacteriophage coat protein fold found in tailed dsDNA bacteriophages (Fig. 2D, Fig. S6) (35, 36). When superimposed on HK97 gp5, the prototypical example of such a fold, the overall structures are remarkably similar despite ϵ 15 gp7 being 52 aa longer. Over the first ~145 aa, the two structures are nearly identical, at which point a domain swap in ϵ 15 gp7 can be seen, where the β -strands in the P-domain occur before the two helices in the A-domain. In the A-domain, the five strands in the central β -sheet are oriented identically to the corresponding domain of the HK97 coat protein, although their ordering is different due to the aforementioned strand swap occurring just before the A-domain. Smaller differences can also be seen in some of the gp7 loops to accommodate the extra residues. Results from the CLICK webserver (37) revealed that 55% of the structure (153 aa) is aligned within 2.19 Å rmsd and has a topology score of 1.0 (i.e., identical topology; topology score of 0 corresponds to irrelevant structures). This high structural similarity with the HK97 bacteriophage coat protein fold provides an additional validation for our gp7 model.

Staple Protein, gp10. As mentioned, ϵ 15 contains a second capsid protein, gp10, which binds to the outer surface of the virus capsid (Fig. 1B). Gp10s from neighboring capsomeres come together to form a dimer at the local twofold axes; visually each gp7 hexon is surrounded by 6 dimers of gp10. With respect to gp7, the gp10s reside adjacent to the A-domain and on top of both the E-loop and long helix from gp7s of neighboring asymmetric units. Due to this arrangement, we previously hypothesized that this protein acted as a molecular staple, performing similar stabilizing functions as the chainmail crosslinks in HK97. However, the exact structural mechanisms necessary for gp10 attachment and its contribution to capsid stability remained unclear.

In our new density map, gp10 appears to contain two sheets consisting of four β -strands each (Fig. 3). Although this map clearly shows β -strand separation, loops connecting the strands at one end of the two sheets were less resolved. A consensus fold model could not be directly established visually. To resolve this, multiple initial models for gp10 were generated using Pathwalker (28). From this pool of possible models, we selected, evaluated, and validated 16 possible topologies based on the density match. Each of these backbone models was converted to an all-atom model using REMO (38) and refined with ROSETTA (39) to evaluate a number of criteria, including density occupancy, model geometry, clashes, interactions, and structural similarities to known folds. Only six of the 16 accounted for all of the gp10 density. Three of these models were eliminated as they displayed clashes and poor protein–protein interfaces when combined with the gp7 model and fit back into the density map. Of the remaining three candidate folds, the top Pathwalker scoring model had the best fit to density. As such, this model was further refined to the density and chosen.

For this model, densities corresponding to aromatic and charged sidechains were generally visible. As with gp7, individual gp10 models in the asymmetric unit were iteratively refined and evaluated with standard X-ray crystallographic modeling tools (40) (Fig. 3, Fig. S7). The final refined model of gp10 clearly showed an eight-stranded jellyroll fold, a common icosahedral virus structural fold found in viruses infecting organisms from all three domains of life (41–44). Using DALI (45), relevant structural similarity to other viruses containing the jellyroll fold were identified, with the top score belonging to the projecting (P) jellyroll domain of the

infectious bursal disease virus VP2 subviral particle (PDB ID code 2DF7, Z-score = 6.3) (Fig. 3 C and D) (43, 46). Other high-scoring jellyroll folds from the DALI search included the coat protein from Satellite Panicum Mosaic Virus (PDB ID code 1STM), the B-domain of the large capsid protein in Cowpea mosaic virus (PDB ID code 2BFU), human rhinovirus 3 coat protein (PDB ID code 1RHI), the single jellyroll domain of the P3 trimer in bacteriophage PRD1 (PDB ID code 1W8X), and the projection domain of the infectious pancreatic necrosis virus VP2 protein (PDB ID code 3IDE). A list of relevant structural similarities is shown in Table S1; a diagram of the gp10 fold with labeled (B–G) strands is shown in comparison with selected jellyroll fold viral proteins (Fig. S8).

Although the infectious bursal disease virus VP2 subviral particle is not a bacteriophage, its crystal structure (PDB ID code 2DF7) (Fig. 3C) (46) offered significant insight into gp10. Visually, gp10 and the 2DF7 domain structure share a remarkable similarity in the overall fold despite a difference in length (141 aa in VP2 versus 111 aa in gp10) and no significant sequence homology; the topology of the two structures is identical (Click Score of 1.0). Over 87 residues in common (81.31%), the two structures have a C α rmsd of 2.44 Å with the majority of the structural difference occurring in the loops joining the β -strands (Fig. 3D). It is interesting to note that the observed jellyroll fold is found in capsids of various triangulation numbers in both animal and plant viruses (Fig. S8). Such structural matches, even in the absence of sequence similarity, serve as a further measure for validating the gp10 model.

Subunit–Subunit Interactions. To maintain capsid stability, $\epsilon 15$ forms an extensive network of interactions between gp7 and gp10, in both their own asymmetric units and across neighboring asymmetric units (Fig. 4 A and B). Based on our models, 14 possible salt bridges were identified using PDBePISA (47), a web-based utility to interactively detect macromolecular interfaces (Table S2). Despite our limited map resolution, corresponding densities were present for some of these modeled salt bridges.

Although $\epsilon 15$ does not have the same covalent chainmail network as HK97 (5), these salt bridges likely play an equivalent role as evidenced of their location when drawing parallels to other bacteriophage capsids, such as P22 and HK97 (9, 12). Of particular interest is a series of three salt bridges found in the gp7 E-loop that tether the long extended loop within and between capsomeres (Table S2). As such, these salt bridges may stabilize the capsid by linking neighboring gp7 subunits in the mature virion, and possibly during capsid maturation.

For gp10, the two subunits, located at the twofold (Fig. 4C), interact along three β -strands (N63–T69, D73–E81, and R31–E37) and are dominated by hydrophobic residues. A number of additional subunit interactions are also seen at the gp7–gp10 interfaces. Like the capsid proteins in HK97 and PSSP-7, the N-arm of gp7 extends parallel to the P-domain; in other bacteriophages like P22 and T4, the N-arm is bent and runs underneath the long helix (Fig. S6). For $\epsilon 15$ gp7, the first ~20 aa of the N-arm takes an abrupt turn toward the capsid surface at Asn21 and forms a small helix and a short strand (blue in Fig. 4C). The first 10 residues of the N-arm from a gp7 subunit in an adjacent capsomere buttress one of the gp10s in the dimer along its exterior surface. Specifically, residues 3–6 of gp7 appear to form a strand parallel to strand C (residues 21–24) of the exposed B-strands in gp10; strands C, H, E, and F are surface exposed whereas strands B, I, D, and G are buried against the gp7 capsid layer (Fig. S8).

Based on the model, additional salt bridges between gp7 and gp10 are predicted with PISA. At the F-loop in gp7, Lys51 appears to form a salt bridge with Asp107 in the C terminus of gp10. Likewise, Lys2 in gp10 appears to coordinate with Ser10 of gp7 in an adjacent capsomere to anchor the N terminus of gp10. Additionally, Asp135 in gp7, located after the long α -helix, interacts with Arg41 of the gp10 model. Based on the location and potential interactions, it is conceivable that gp10 plays a key role in contributing to the capsid stability by joining neighboring capsomeres. This is further

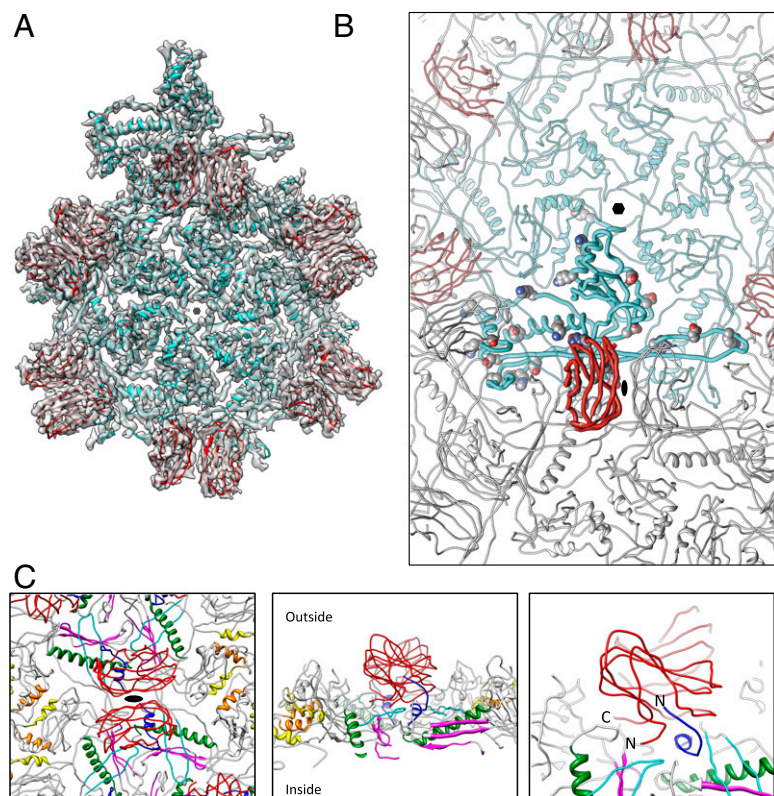


Fig. 4. $\epsilon 15$ subunit interactions. (A) An asymmetric unit for $\epsilon 15$ consists of seven gp7 (cyan) and seven gp10 (red) subunits. For illustration purposes, additional gp10 subunits are highlighted to illustrate the dimeric relationship. (B) One gp10 (bold red subunit) and one gp7 (bold cyan subunit) are shown with potential salt bridges shown as spheres (Table S2). (C) The gp10 dimer is highlighted in red and spans the twofold axis (Left). Four gp7 subunits are directly below a single gp10. The N-arm regions of gp7 are colored in dark blue and flank the outer strands of gp7. The color annotations for gp7 are the same as defined in Fig. 2B. Directly below the gp10 dimers are the gp7 E-loops (cyan). Rotating 90° shows a slice through the capsid with both gp10 termini positioned at the twofold interface, protruding down into the gp7 subunits (Center). A zoomed-in view of this region highlights the positioning of the N-arm of gp7 and the termini of gp10 (Right).

reinforced by the fact that even in 5 M guanidinium chloride (GuHCl), the gp7 and gp10 capsid shell remains intact (4).

Beyond stabilizing the mature $\epsilon 15$ virion, gp10 may play a larger role during capsid maturation. Previous examples of phage maturation in P22 and HK97 show a fairly substantial change in the N-arm conformation (9, 12, 48). In P22, it is postulated that the N-arm binds the C terminus of the scaffolding protein and that the conformational change in the N-arm contributes to scaffolding protein release. If gp10 is present during procapsid assembly, it raises the question if and how the N-arm of gp7 would simultaneously interact with the scaffolding protein (49) and gp10. Alternatively, gp10 may be incorporated in the mature phage, like in lambda and phage L (15, 50). Regardless of when gp10 is introduced into the capsid, the large conformational change during maturation would require dynamic, and possibly transitory, interactions to maintain capsid stability.

Interestingly, two protruding densities corresponding to the cysteine residues (C65) are prominent at the gp10–gp10 interface. Their proximity suggests a possible disulfide bond. However, neither connecting densities nor biochemical evidence exists to support such interaction in the mature form of $\epsilon 15$. Nevertheless, it is possible that if gp10 is present in the procapsid, this disulfide bond may serve as a temporary mechanism to stabilize the capsid during maturation until a full set of inter- and intracapsomeric interactions can be established in the mature virus. Such temporary disulfide bonds are not without precedence as early transient disulfide-dependent steps have been shown in influenza virus nucleocapsid (51) and Simian Virus 40 maturation (52). Additionally, transient interchain disulfide bonds in the P22 tailspike adhesion proteins have been seen (53).

Evolutionary Lineage of $\epsilon 15$ Phage. Tailed dsDNA bacteriophages have demonstrated a highly conserved structural fold as first defined in gp5 of HK97 (35) (Fig. S6). This fold has also been found to exist in the major capsid protein of Herpesvirus (36), suggesting that this fold, despite host, sequence, and structural variations, is a basic template that contributes to the ability of the major capsid proteins to form the primary proteinaceous shell that envelops the genome (Fig. 2B, Fig. S6). The presence of the swapped strands in gp7 further supports the notion that this fold is a template that is conducive for virus formation. As such, the canonical fold found in the major capsid proteins of tailed dsDNA bacteriophages is essentially a scaffolding that can tolerate varied interactions, insertion domains, and decorating proteins, as well as sequence and structural permutations in forming a stable yet flexible capsid that can efficiently accommodate the viral genome.

Likewise, the jellyroll fold is seen in a wide variety of nonrelated viruses including animal viruses, plant viruses, and bacteriophages (Fig. S8) (41, 42, 44), as well as nonviral cellular proteins. Until our structure of $\epsilon 15$, no virus structure has been shown to contain both of these common viral folds. Because *Salmonella*, a bacterial pathogen in higher vertebrates, is the host of $\epsilon 15$ phage, it is conceivable that the $\epsilon 15$ evolved in the vertebrate host, which is a reservoir for both bacteriophage and other viruses. Genetic transfer and recombination within the microbial environment may have resulted in the emergence of $\epsilon 15$ phage.

Materials and Methods

Cryo-EM. Infectious bacteriophage $\epsilon 15$ was purified as described previously (3). The purified infectious phage particles were plunge frozen at liquid N₂ temperature using a FEI Vitrobot device. The images were taken at ~ 17 e/Å² dose in a JEM3200FSC electron cryomicroscope operated at 300 kV and at liquid N₂ specimen temperature. The imaging condition was set with spot size = 2, condenser aperture = 50 μ m, and objective aperture = 120 μ m. The in-column energy filter of this instrument was activated to remove electrons with larger than 25 eV energy loss. The images were recorded on Kodak SO163 films at 50,000 \times nominal magnification. A total of 1,309 films were digitized using Nikon Super CoolScan 9000 ED scanner at 6.35 μ m/pixel and resulted in calibrated sampling of 1.19 Å/pixel for the scanned images.

Image Processing. Individual particles (720 \times 720 pixels) were first automatically selected using the *ethan* method (54) followed by manual screening using EMAN *boxer* program (55). A total of 54,161 particles were selected for initial processing. The selected particles within a micrograph were incoherently averaged to generate 2D power spectrum for contrast transfer function (CTF) parameter determination. CTF parameters were first automatically estimated (56) and then visually verified using the EMAN1 *ctffit* program (55). The defocuses range from 0.5 to 2.5 μ m. The dataset was divided into two data subsets for the following reconstruction steps.

The particle images were first binned 4 \times for initial model building and initial determination of orientation and center parameters. The initial model was built de novo by iterative refinement of a subset of 300 particles randomly selected from the half dataset with randomly assigned initial orientations. The initial orientations of all particles in each of the half datasets were determined using the EMAN1 projection matching program *classesbymr* with an angular projection step size of 3 degrees (55). The orientations were then refined to higher accuracy using the program *align*, which is based on simplex optimization of matching between the particle image and model projections (19). The particle orientation parameters were then transferred to particles binned at 2 \times and ultimately to particles without binning for further refinements. In the last stage of refinement, magnification, astigmatism, and defocus parameters were also included (19). 3D maps with icosahedral symmetry enforcement were reconstructed using a newly developed program *j3dr* using EMAN2 library (57) and parallelized with message passing interface (MPI) to speed up the reconstruction process. These steps were iterated until the refinement converged.

The map for each data subset was reconstructed from $\sim 7,000$ particles by removing particles with poor alignment scores and unstable alignment parameters. The resolution of the map was evaluated using the FSC. Only the icosahedral shell region was included in this FSC analysis by masking out the external background noises and the internal DNA densities using soft masks with a half width of 6 Å. The final map of the entire dataset was then built from $\sim 14,000$ particles by combining these two subsets of particles.

Model Building. For model building, each of the seven individual gp10 and gp7 subunits from one asymmetric subunit was cropped out of the full map using CHIMERA, from the University of California, San Francisco (UCSF) (58). For gp7s, the previous structural model (4) was used as a template to help define subunit boundaries. Once segmented out, the remaining densities were attributed to gp10s, which were then manually segmented out of the map. Individual gp7s and gp10s were then aligned with *Foldhunter* (59) and averaged using *proc3d*, both of which are available in EMAN1 (55). From the initial averaged subunits, a second round of segmentation with UCSF's CHIMERA was performed using the initial subunit averages as templates for cropping. A second round of aligning and averaging resulted in single averaged gp7 and gp10 subunits.

The final averaged gp7 and gp10 subunits were used to build/validate models using *e2pathwalker.py* (28) from EMAN2 (57). With gp10, initial pseudoatom placement for pathwalking was done using *e2segment3d.py*, with minimum and maximum spacings of 2.1 and 4 Å, respectively, resulting in placement of 111 C α atoms within the gp10 density map. Pseudoatoms for gp7 (335 aa) were obtained by fitting the initial C α backbone trace into the averaged density map with UCSF's CHIMERA. For both gp7 and gp10, initial pseudoatom placement was varied ($\sigma = 1$) and 100 models were calculated in *e2pathwalker.py*. A threshold was also set, corresponding to the smallest isosurface value at which the density for a single subunit remained connected. Pathwalking was performed with the default distance constraints using the LKH TSP solver (28). The corresponding sequences were then threaded onto the corresponding models in both the forward and reverse directions.

For gp7, a representative model for each of the two possible fold classes was fit into the $\epsilon 15$ density map at each of the seven positions in the asymmetric unit and evaluated for fit to density. The best model was visually selected and converted to an all-atom model using REMO, a webservice that transforms C α backbone models into all-atom models (38).

For gp10, we identified 16 potential model types from the gallery of Pathwalker models. Each was loaded into GORGON (29) and was interactively adjusted to correct any bond distance and clash problems. These models were then converted to all-atom models using REMO and density refined with ROSETTA (39). Initial evaluation of the models considered (i) geometry, (ii) fit to density, (iii) density occupancy, and (iv) interface between neighboring gp10s and gp7s. Additionally, the models were submitted to the DALI to identify possible structural homologs (45). A final model was selected based on fit to density, model geometry, and structural similarity to known homologs.

Using these all-atom models, slightly more than an entire asymmetric unit (12 gp10s and 7 gp7s) was assembled in the context of the entire capsid

structure. Fitting of these subunits was done manually and refined with the “Fit in Map” option in UCSF’s CHIMERA (58). Individual subunits were iteratively refined with ROSETTA (39) and COOT (32) to optimize sidechain positioning, improve density agreement, and improve the dihedral angles, as well as eliminate subunit clashes. To optimally refine an $\epsilon 15$ entire asymmetric unit using best practices established in X-ray crystallography, PHENIX (33) and COOT (32) were used iteratively.

PHENIX refinement was implemented using *phenix.refine* (60). To use our density map data, a python script was written, converting the phases of the map to HKL coefficients, which could then be used during the refinement process. Refinement was done by applying secondary structure restraints, NCS restraints, and map/stereochemistry weight restraints. Initial rounds of refinement used Ramachandran restraints and a strong geometry weight. To minimize potential overfitting, geometric and density statistics (Table 1)

1. Wommack KE, Colwell RR (2000) Virioplankton: Viruses in aquatic ecosystems. *Microbiol Mol Biol Rev* 64(1):69–114.
2. Hendrix RW, Duda RL (1998) Bacteriophage HK97 head assembly: A protein ballet. *Adv Virus Res* 50:235–288.
3. Jiang W, et al. (2006) Structure of epsilon15 bacteriophage reveals genome organization and DNA packaging/injection apparatus. *Nature* 439(7076):612–616.
4. Jiang W, et al. (2008) Backbone structure of the infectious epsilon15 virus capsid revealed by electron cryomicroscopy. *Nature* 451(7182):1130–1134.
5. Wikoff WR, et al. (2000) Topologically linked protein rings in the bacteriophage HK97 capsid. *Science* 289(5487):2129–2133.
6. Simkovsky R, King J (2006) An elongated spine of buried core residues necessary for in vivo folding of the parallel beta-helix of P22 tailspike adhesin. *Proc Natl Acad Sci USA* 103(10):3575–3580.
7. Hendrix RW (2002) Bacteriophages: Evolution of the majority. *Theor Popul Biol* 61(4):471–480.
8. Fokine A, et al. (2005) Structural and functional similarities between the capsid proteins of bacteriophages T4 and HK97 point to a common ancestry. *Proc Natl Acad Sci USA* 102(20):7163–7168.
9. Gan L, et al. (2006) Capsid conformational sampling in HK97 maturation visualized by X-ray crystallography and cryo-EM. *Structure* 14(11):1655–1665.
10. Liu X, et al. (2010) Structural changes in a marine podovirus associated with release of its genome into *Prochlorococcus*. *Nat Struct Mol Biol* 17(7):830–836.
11. Ionel A, et al. (2011) Molecular rearrangements involved in the capsid shell maturation of bacteriophage T7. *J Biol Chem* 286(1):234–242.
12. Chen DH, et al. (2011) Structural basis for scaffolding-mediated assembly and maturation of a dsDNA virus. *Proc Natl Acad Sci USA* 108(4):1355–1360.
13. Fokine A, et al. (2004) Molecular architecture of the prolate head of bacteriophage T4. *Proc Natl Acad Sci USA* 101(16):6003–6008.
14. Ishii T, Yanagida M (1975) Molecular organization of the shell of the T-even bacteriophage head. *J Mol Biol* 97(4):655–660.
15. Lander GC, et al. (2008) Bacteriophage lambda stabilization by auxiliary protein gpD: Timing, location, and mechanism of attachment determined by cryo-EM. *Structure* 16(9):1399–1406.
16. Pope WH, et al. (2007) Genome sequence, structural proteins, and capsid organization of the cyanophage Syn5: A “horned” bacteriophage of marine synechococcus. *J Mol Biol* 368(4):966–981.
17. Sternberg N, Weisberg R (1977) Packaging of coliphage lambda DNA. II. The role of the gene D protein. *J Mol Biol* 117(3):733–759.
18. Tang L, Gilcrease EB, Casjens SR, Johnson JE (2006) Highly discriminatory binding of capsid-cementing proteins in bacteriophage L. *Structure* 14(5):837–845.
19. Guo F, Jiang W (2013) *Methods in Molecular Biology: Electron Microscopy* (Humana Press, New York).
20. Scheres SH, Chen S (2012) Prevention of overfitting in cryo-EM structure determination. *Nat Methods* 9(9):853–854.
21. Rosenthal PB, Henderson R (2003) Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *J Mol Biol* 333(4):721–745.
22. Grigorieff N (2000) Resolution measurement in structures derived from single particles. *Acta Crystallogr D Biol Crystallogr* 56(Pt 10):1270–1277.
23. Penczek PA (2010) Resolution measures in molecular electron microscopy. *Methods Enzymol* 482:73–100.
24. Karmali AM, Blundell TL, Furnham N (2009) Model-building strategies for low-resolution X-ray crystallographic data. *Acta Crystallogr D Biol Crystallogr* 65(Pt 2):121–127.
25. Ludtke SJ, et al. (2008) De novo backbone trace of GroEL from single particle electron cryomicroscopy. *Structure* 16(3):441–448.
26. Abeyasinghe SS, Baker ML, Chiu W, Ju T (2008) Segmentation-free skeletonization of grayscale volumes for shape understanding. *IEEE International Conference on Shape Modeling and Applications* 63–71.
27. Ju T, Baker ML, Chiu W (2007) Computing a family of skeletons of volumetric models for shape description. *Comput Aided Des* 39(5):352–360.
28. Baker MR, Rees I, Ludtke SJ, Chiu W, Baker ML (2012) Constructing and validating initial C α models from subnanometer resolution density maps with pathwalking. *Structure* 20(3):450–463.
29. Baker ML, et al. (2011) Modeling protein structure at near atomic resolutions with Gorgon. *J Struct Biol* 174(2):360–373.
30. Baker ML, Zhang J, Ludtke SJ, Chiu W (2010) Cryo-EM of macromolecular assemblies at near-atomic resolution. *Nat Protoc* 5(10):1697–1708.
31. Bradley P, et al. (2005) Free modeling with Rosetta in CASP6. *Proteins* 61(Suppl 7):128–134.
32. Emsley P, Lohkamp B, Scott WG, Cowtan K (2010) Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* 66(Pt 4):486–501.
33. Adams PD, et al. (2010) PHENIX: A comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* 66(Pt 2):213–221.
34. Henderson R, et al. (2012) Outcome of the first electron microscopy validation task force meeting. *Structure* 20(2):205–214.
35. Veleser D, Cambillau C (2011) A common evolutionary origin for tailed-bacteriophage functional modules and bacterial machineries. *Microbiol Mol Biol Rev* 75(3):423–433.
36. Baker ML, Jiang W, Rixon FJ, Chiu W (2005) Common ancestry of herpesviruses and tailed DNA bacteriophages. *J Virol* 79(23):14967–14970.
37. Nguyen MN, Tan KP, Madhusudhan MS (2011) CLICK—Topology-independent comparison of biomolecular 3D structures. *Nucleic Acids Res* 39(Web Server issue):W24–W28.
38. Li Y, Zhang Y (2009) REMO: A new protocol to refine full atomic protein models from C-alpha traces by optimizing hydrogen-bonding networks. *Proteins* 76(3):665–676.
39. DiMaio F, Tyka MD, Baker ML, Chiu W, Baker D (2009) Refinement of protein structures into low-resolution density maps using rosetta. *J Mol Biol* 392(1):181–190.
40. Chen VB, et al. (2010) MolProbity: All-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr* 66(Pt 1):12–21.
41. Cheng S, Brooks CL, 3rd (2013) Viral capsid proteins are segregated in structural fold space. *PLoS Comput Biol* 9(2):e1002905.
42. Harrison SC (2007) in *Fields Virology* (Wolters Kluwer Health/Lippincott Williams & Wilkins, Philadelphia), 5th Ed, pp 60–98.
43. Coulbaly F, et al. (2005) The birnavirus crystal structure reveals structural relationships among icosahedral viruses. *Cell* 120(6):761–772.
44. Rossmann MG, Johnson JE (1989) Icosahedral RNA virus structure. *Annu Rev Biochem* 58:533–573.
45. Holm L, Rosenström P (2010) Dali server: Conservation mapping in 3D. *Nucleic Acids Res* 38(Web Server issue):W545–W549.
46. Lee CC, et al. (2006) Crystal structure of infectious bursal disease virus VP2 subviral particle at 2.6Å resolution: Implications in virion assembly and immunogenicity. *J Struct Biol* 155(1):74–86.
47. Krissinel E, Henrick K (2007) Inference of macromolecular assemblies from crystalline state. *J Mol Biol* 372(3):774–797.
48. Gertsman I, et al. (2009) An unexpected twist in viral capsid maturation. *Nature* 458(7238):646–650.
49. Kropinski AM, et al. (2007) The genome of epsilon15, a serotype-converting, Group E1 Salmonella enterica-specific bacteriophage. *Virology* 369(2):234–244.
50. Gilcrease EB, Winn-Stapley DA, Hewitt FC, Joss L, Casjens SR (2005) Nucleotide sequence of the head assembly gene cluster of bacteriophage L and decoration protein characterization. *J Bacteriol* 187(6):2050–2057.
51. Prokudina EN, Semenova NP, Chumakov VM, Rudneva IA (2004) Transient disulfide bonds formation in conformational maturation of influenza virus nucleocapsid protein (NP). *Virus Res* 99(2):169–175.
52. Li PP, Nakanishi A, Clark SW, Kasamatsu H (2002) Formation of transitory intrachain and interchain disulfide bonds accompanies the folding and oligomerization of simian virus 40 Vp1 in the cytoplasm. *Proc Natl Acad Sci USA* 99(3):1353–1358.
53. Kim J, Robinson AS (2006) Dissociation of intermolecular disulfide bonds in P22 tailspike protein intermediates in the presence of SDS. *Protein Sci* 15(7):1791–1793.
54. Kivioja T, Ravanti J, Verkhovaly A, Ukkonen E, Bamford D (2000) Local average intensity-based method for identifying spherical particles in electron micrographs. *J Struct Biol* 131(2):126–134.
55. Ludtke SJ, Baldwin PR, Chiu W (1999) EMAN: Semiautomated software for high-resolution single-particle reconstructions. *J Struct Biol* 128(1):82–97.
56. Jiang W, Guo F, Liu Z (2012) A graph theory method for determination of cryo-EM image focuses. *J Struct Biol* 180(2):343–351.
57. Tang G, et al. (2007) EMAN2: An extensible image processing suite for electron microscopy. *J Struct Biol* 157(1):38–46.
58. Pettersen EF, et al. (2004) UCSF Chimera—A visualization system for exploratory research and analysis. *J Comput Chem* 25(13):1605–1612.
59. Jiang W, Baker ML, Ludtke SJ, Chiu W (2001) Bridging the information gap: Computational tools for intermediate resolution structure interpretation. *J Mol Biol* 308(5):1033–1044.
60. Afonine PV, et al. (2012) Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr D Biol Crystallogr* 68(Pt 4):352–367.