UNIVERSITY OF CALIFORNIA

Los Angeles

Developing Cryo Electron Microscopy Tool for Nanomachines

A dissertation submitted for partial satisfaction of the requirements for the

degree Doctor of Philosophy

in Bioengineering

by

Ke Ding

2019

ABSTRCT OF THE DISSERTATION

Developing Cryo Electron Microscopy Tool for Nanomachines

by

Ke Ding

Doctor of Philosophy in Bioengineering

University of California, Los Angeles, 2019

Professor Z. Hong Zhou, Co-Chair

Professor Ren Sun, Co-Chair

Large biomolecular complexes often work as nanomachines to directly process molecules, which requires these complexes to adapt different conformations at different states. With the recent rapid development of cryo electron microscopy (cryoEM) technique, both data quality and quantity have improved. As a result, more heterogeneities in the dataset can be detected. How to turn these observed heterogeneities to structural dynamics at the molecular level is a pressing topic in the field of structural biology. In this dissertation, the strong relation between unnatural air-water interfaces in the cryoEM sample and heterogeneities found in structures is proposed. To demonstrate that imaging nanomachines *in situ* is a reliable way to understand dynamics, two nanomachines, the vault organelle and reovirus polymerases, are resolved to near-atomic resolution, proving cryoEM is a powerful tool to find heterogeneities existing in the dataset. To further understand the working mechanism of reovirus RNA-dependent RNA polymerase (RdRp),

different states are induced in two different reoviruses and the dynamics of reovirus RdRp are revealed from the conformational changes observed in the assembled, functional transcriptional complexes (with RdRp and genome both inside the capsid). These works also demonstrate that the classification focusing on the area of interest is an effective tool to resolve heterogeneities, much like a divide-and-conquer approach. When the nanomachine's structures are resolved *in situ*, the detected heterogeneities become informative in understanding the nanomachine's function, particularly for polymerases.

The dissertation of Ke Ding is approved.

Timothy J. Deming

Leonard H. Rome

Ren Sun, Co-chair

Z. Hong Zhou, Co-chair

University of California, Los Angeles

2019

I dedicate this Ph.D. dissertation to my beloved grandparents.

Table of Contents

Table of Figures

Table of Tables

Acknowledgements

I would like to thank my dissertation advisor and mentor, Dr. Z. Hong Zhou, for his continuous support and guidance for the past six years. For the first two years, he tolerated my fantasies of academia, and I keep that naïve passion in everyday research. In the next two years, he remodeled me to a calm and rational researcher, shaping me to make solid contributions and share my love of science with audiences in the sciences and beyond. In the last two years, he taught me how to manage time and work with collaborators effectively, showing me the real academia beyond just research. His mentorship is very constructive and inspiring. He has especially emphasized the importance of accuracy and productivity in experiments, academic writing, and even daily life, lessons I deeply appreciate. Beyond supporting my personal growth, he has also supported my research with a plethora of resources: I never worry about microscopy and computation time in his lab. I am truly fortunate to have him as my Ph.D. mentor.

I would like to thank my other committee members, including Dr. Ren Sun, Dr. Leonard Rome and Dr. Tim Deming, for helping me with my dissertation and supporting my career development. I also would like to thank the Bioengineering department at UCLA for awarding me the departmental fellowship and nominating me for the dissertation year fellowship, both fellowships I am grateful to have. Thank you to the molecular biology institute at UCLA and Dr. Philip Whitcome for naming me the Whitcome Fellow: this prestigious fellowship provides fantastic opportunities to international students like me.

Many thanks to my collaborators: Dr. Leonard Rome, Dr. Otto Yang, Dr. Jingchen Sun, Dr. Polly Roy, Dr. Jan Mrazek, Dr. Valerie Kickhoefer and Dr. Cristina Celma. Without them, the papers included in this dissertation would not be published. Their effort and support truly made these inter-lab collaborations efficient and productive.

I also would like to thank Dr. Xing Zhang and Dr. Xuekui Yu for their help with the 2015 Nature paper. They taught me a significant amount of virology when we worked together, and I am incredibly lucky to have written a paper with them. Many thanks for being patient and teaching a not-so-young Ph.D. student like me from scratch.

My special thanks to all my talented undergraduate co-authors in the Zhou lab: Iris Yu, Lisa Nguyen, Mason Lai, Thomas Chang, Wesley Shen, Winston Chang, Justin Bui, Pallavi Chandrasekhar, Jaycob Avaylon and Harold Smith. I am happy for you when you find that you are making solid contributions to our projects. Without your input, our discoveries cannot be thoroughly appreciated.

Many thanks to Dr. Peng Ge, for teaching me cryoEM one-on-one, being always open to answer any questions, and keeping the computation nodes running. I also would like to thank the other senior members in the group for helping me with my questions about microscopy: Ivo Atanasov, Wonghoi Hui, Kang Zhou and Yanxiang Cui. Your 24/7 support (almost) makes accessing the Titan Krios possible for every group member. I also would like to thank Dr. Chi-min Ho, Dr. Kevin Huynh, Dr. Ana Lucia Alvarez-Cabrera and Paul Sieminski for supporting me in many aspects during my graduation season. Thank you to Titania Nguyen for editing my manuscripts and not making fun of my manuscript, but of me.

Thank you to my friends for their understanding and encouragement. Special thanks to Dr. Weiyi Wu, whom I've known for 15 years and who got his Ph.D. in four of those years. Thank you for demonstrating my match-making talent.

My deepest thanks to my family: my parents and parents-in-law have graciously understood my limited chances to visit China for the past six years. Their support always keeps me from depression. Finally, thank you, Dr. Di Liang, my wife. Let's get the firecrackers ready.

Biographical Sketch

Ke Ding

EDUCATION

UCLA, Mechanical and Aerospace Engineering, Micro-Mechanical-Electrical System
M.S. 2011-2013 Overall GPA:4.00/4.00

Peking University, Beijing, P.R. China, Electrical Engineering and Computer Science
B.S. 2007-2011 Major GPA:3.78/4.00

SPECIAL SKILLS

High Resolution Macromolecule Reconstruction by Single-particle Analysis

Sample preparation (Negative stain/Cryo fixation), cryo imaging, solving reconstruction problem for challenging sample case (symmetry variation, large-scale conformational change, etc.).

Dual-beam System Fabrication

Innovative Focused Ion Beam (FIB) fabrication, including clamp-clamp silicon nitride nano-string array, in-plane spinning of membrane driven by FIB induced stress.

Microfluidics and Surface Wettability Engineering

PDMS microfluidics chip fabrication; Microfluidics simulation; Wettability-based micro-channel.

Software simulation & Coding

EMAN, RELION, FREALIGN, IMOD, PEET, COMSOL, MATLAB, C++, Python, LATEX

EMPLOYMENT

Research Assistant, UCLA Electron Imaging Center for Nanomachines (EICN) 2013-2019
Developing cryo electron microscopy tool for nanomachines Advisor: Prof. Hong Zhou

Teaching Assistant
MIMG 105 Biological Microscopy, 2014 Fall, 2015 Fall
MAE 183 Introduction to Manufacturing Processes, 2012 Spring

Research Assistant, UCLA Micro and Nano Manufacturing Lab 2011-2012
Micro-regulator based on super-hydrophobic surface Advisor: Prof. Chang-Jin (CJ) Kim

Research Assistant, Peking University, Department of Microelectronics 2009-2011
Nano-string fabrication with focused ion beam induced fluidization Advisor: Prof. Wengang Wu

HONORS & AWARDS

Dissertation Year Fellowship	UCLA	2017
Whitcome Fellow	UCLA	2015, 2016
Bioengineering Department Fellowship	UCLA	2013
Best Posters, 3rd Annual SoCal cryoEM symposium	TSRI	2018
9th K.H. Kuo symposium travel award	IUPAB	2016
IWAIP cryoEM student grant	IBP, CAS	2013
Panasonic Scholarship	Peking University	2010
Kwang-Hua Scholarship	Peking University	2008

PUBLICATIONS

Ding, K., Celma, C.C., Zhang, X., Chang, T., Shen, W., Atanasov, I., Roy, P. and Zhou, Z.H., 2019. In situ structures of rotavirus polymerase in action and mechanism of mRNA transcription and release. *Nature Communications*, 10(1), 2216

Ding, K., Nguyen, L. and Zhou, Z.H., 2018. In Situ Structures of the Polymerase Complex and RNA Genome Show How Aquareovirus Transcription Machineries Respond to Uncoating. *Journal of virology*, 92(21), pp.e00774-18.

Ding, K., Zhang, X., Mrazek, J., Kickhoefer, V.A., Lai, M., Ng, H.L., Yang, O.O., Rome, L.H. and Zhou, Z.H., 2018. Solution Structures of Engineered Vault Particles. *Structure*, 26(4), pp.619-626.

Zhang, X.*, Ding, K.*, Yu, X.*, Chang, W., Sun, J., and Zhou, Z. H. 2015. In situ structures of the segmented genome and RNA polymerase complex inside a dsRNA virus. *Nature*, 527(7579), 531-534.

Zhang, X., Lai, M., Chang, W., Yu, I., Ding, K., Mrazek, J., Ng, H.L., Yang, O.O., Maslov, D.A. and Zhou, Z.H., 2016. Structures and stabilization of kinetoplastid-specific split rRNAs revealed by comparing leishmanial and human ribosomes. *Nature communications*, 7, p.13223.

Mayle, K.M., Dern, K.R., Wong, V.K., Chen, K.Y., Sung, S., Ding, K., Rodriguez, A.R., Knowles, S., Taylor, Z., Zhou, Z.H, Grundfest, W.S., Wu A.M., Deming, T.J. and Kamei, D.T., 2017. Engineering A11 Minibody-Conjugated,Polypeptide-Based Gold Nanoshells for Prostate Stem Cell Antigen (PSCA)Targeted Photothermal Therapy. *SLAS TECHNOLOGY: Translating Life Sciences Innovation*, 22(1), pp.26-35.

Mayle, K.M., Dern, K.R., Wong, V.K., Sung, S., Ding, K., Rodriguez, A.R., Taylor, Z., Zhou, Z.H., Grundfest, W.S., Deming, T.J. and Kamei, D.T., 2017. Polypeptide-based gold nanoshells for photothermal therapy. *SLAS TECHNOLOGY: Translating Life Sciences Innovation*, 22(1), pp.18- 25.

Wang, J., Feng, B., Li, C., Zhang, F.Q., Ding, K. and Wu, W.G., 2011, June. Ultrasensitive mass sensor using the out-of-phase vibration eigenstate of intercoupled dual-microcantilevers, 2011. 16th International Solid-State Sensors, Actuators and Microsystems Conference (pp. 2010-2013). IEEE.

Li, C., Ding, K., Wu, W.G. and Xu, J., 2011, January. Ultra-fine nanofabrication by hybrid of energetic ion induced fluidization and stress,2011. IEEE 24th International Conference on Micro Electro Mechanical Systems (pp. 340-343). IEEE.

Chapter 1 Introduction

1.1 Structural biology

The correlation between structure and function in biology, seemingly completely divergent at the macroscopic level, strengthens as we approach the microscopic scale: from anatomy/physiology (structure/function studied separately) at organ level, to histology at tissue level, to cellular biology at cell level, and all the way to structural biology at molecular/atomic level.

Structural biology studies the molecular structure and dynamics of biomolecules. Experimental measurement of a biomolecule's structure is an important approach for studying its structure, while the dynamics of a biomolecule can be obtained by resolving the structure of a biomolecule at different states.

1.1.1 Structural biology techniques: averaging

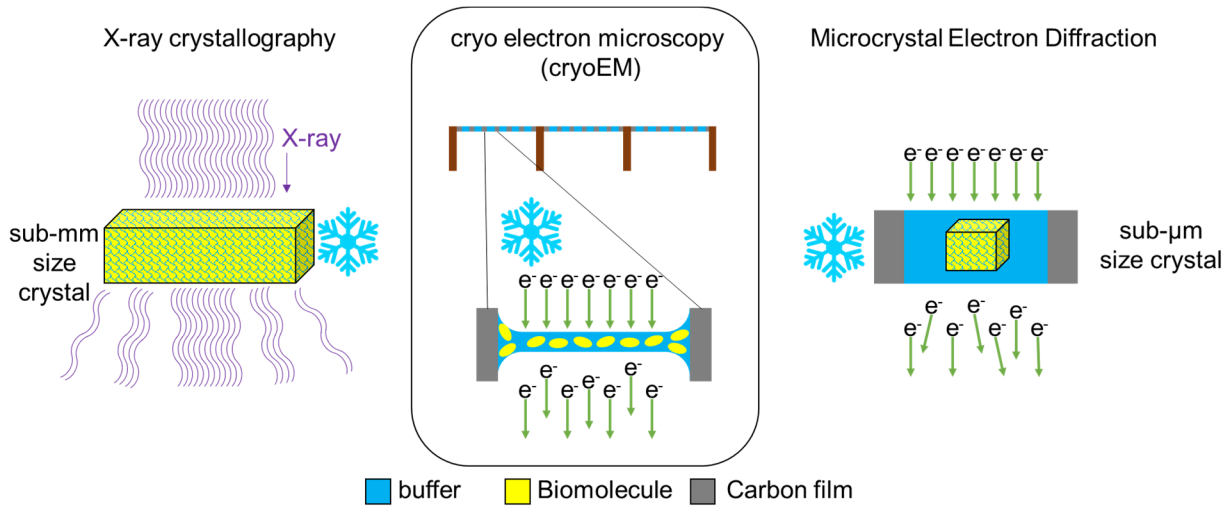


Figure 1.1 Three structural biology approaches to measure biomolecules. X-ray crystallography (left), cryoEM (middle) and Microcrystal Electron Diffraction (right). All of them requires multiple copies of biomolecules and samples are at low temperature during contrast transfer.

To measure a biomolecule experimentally (Figure 1.1), different coherent beams pass through biomolecules, and how the incoming beam is disturbed informs the structural information of these

biomolecule. In X-ray crystallography, the beam is X-rays; in microcrystal electron diffraction (MicroED) and cryo electron microscopy (cryoEM), the beam is accelerated electrons. Biomolecules are vulnerable under beam radiation, so to gather enough information, multiple copies of biomolecules are crystallized in X-ray crystallography and MicroED, while multiple images are taken of different molecules of the same kind in cryoEM. The average dosage each biomolecule endures is minimized so the information in the molecule can be transferred to the beam with minimal damage to the molecule. All these samples need to be kept at low temperature to further minimize radiation damage during imaging.

1.1.2 cryoEM

In the past 40 years, cryoEM has developed into a mainstream technique to determine biomolecular structures. The workflow of cryoEM sample preparation and imaging is described in Figure 1.2. The sample is first purified and transferred to one side of a copper grid with 2.5 mm in diameter, which has a single layer of holey carbon on one side. This holey carbon layer is plasma-cleaned beforehand so that the applied sample (~ 2.5μL) can form a meniscus on the cleaned front side. The grid with sample is then blotted with filter paper with controlled environmental humidity, temperature, blotting force, and time, so that a thin buffer layer can then remain on the holey carbon film, with biomolecule samples embedded. After removing the filter paper, the grid is plunge frozen into liquid ethane to vitrify the buffer. Vitrified buffer film is not crystallized, so the electron beam can pass through it without strong diffraction. This grid with vitrified buffer is then put into a transmission electron microscope (TEM). The electrons coming from the gun will be manipulated by a series of lenses and pass through the biomolecules frozen inside the vitrified buffer to get an image. This image is then magnified by another lens and finally projected on the electron detector, which directly transfers electron signals to a digitalized bitmap file. The vitrified

sample is always kept at liquid nitrogen temperature to prevent the water from crystallizing in the buffer.

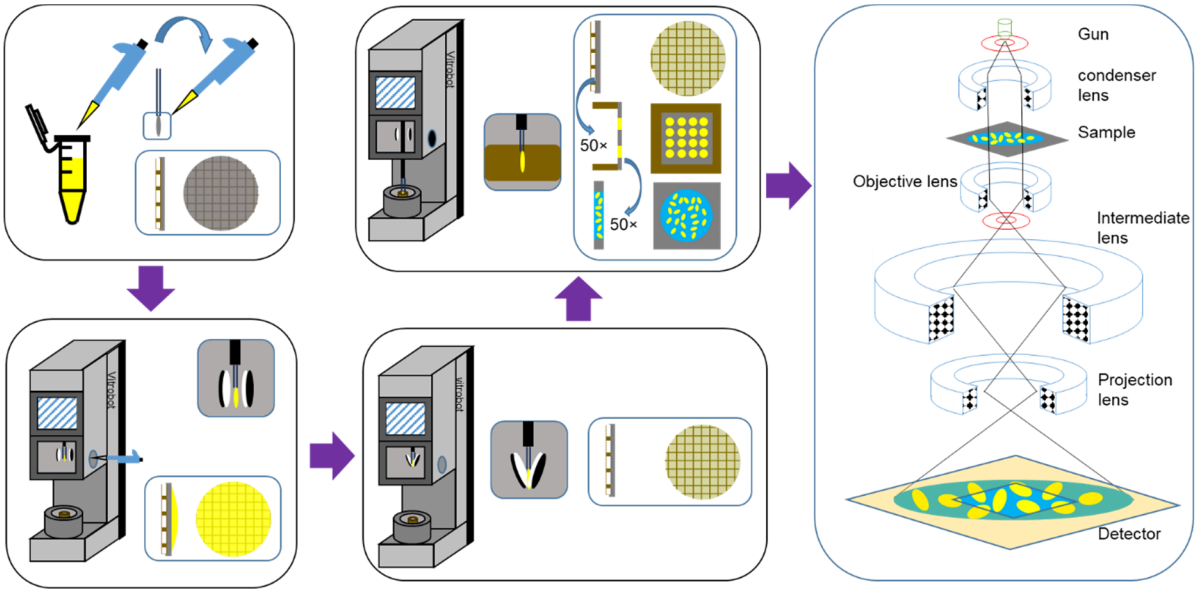


Figure 1.2 cryoEM workflow of sample preparation and imaging

The next step is data processing. The most popular method to obtain high-resolution reconstruction is single particle analysis (Figure 1.3). The bitmap collected from the detector will first be merged into one single micrograph. If the detector collects a movie with multiple frames, the movie will first be combined into a single micrograph, with drifts between frames corrected. In the combined micrograph, each biomolecule or complex is pinpointed and boxed out to be a "particle". Tens of thousands of particles are boxed from thousands of micrographs. These boxed particles are then averaged in 2 dimensions (2D), only searching center and in-plane rotation to align them; the resulting 2D averages are used to estimate data quality. After selecting good particles in 2D averages, those good particles are refined to find a 3-dimensional (3D) orientation. Specific refinement methods include cross common line[1] and projection matching[2]. After determining an 3D orientation for each particle, the particle can be reconstructed based on this specific orientation by inserting the Fourier transform of that particle image into a 3D Fourier

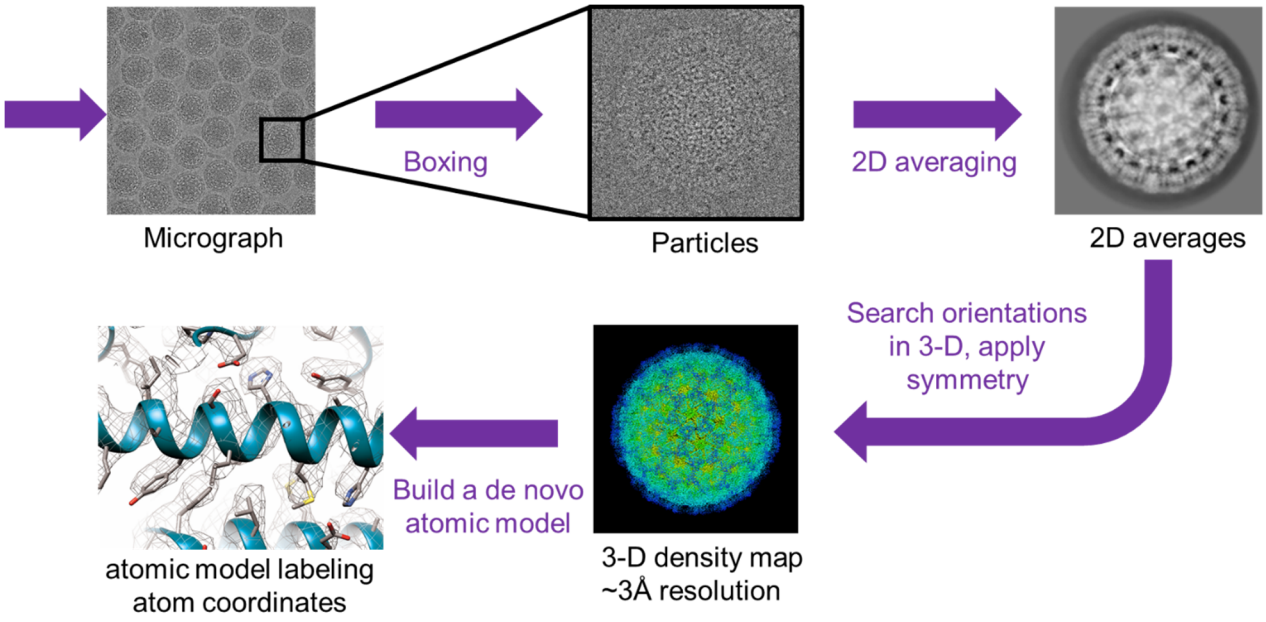


Figure 1.3 The work flow of single particle analysis, from merged micrographs to atomic model

space. After all particles are reconstructed, global symmetry is applied to the 3D Fourier space to obtain a Fourier space as complete as possible. An invert Fourier transform is then conducted on this 3D Fourier space, and a 3D density map is obtained: with enough particles averaged and their orientation accurately determined, this reconstruction can show high resolution features. Such high resolution features include backbone and side chain densities in protein and ligand densities. Based on this accurate density map and the sequence of the biomolecule, atomic models can be built into the regions with strong density and the coordinates of atoms in the biomolecule can be obtained.

Because the density of biomolecules and the water buffer are similar, the contrast in cryoEM is relatively low and the signal-to-noise ratio in each particle is low [3]. In the projection matching method, a particle's orientation is largely dependent on its comparison with averaged, less-noisy results: if a particle matches better with an averaged projection at certain orientation, this orientation will be more likely to be assigned to this particle. If a particle does not match well with the existing projections, there are three possibilities: 1) the imaged biomolecule is not at its physiological state, so its conformation is not the same as other particles; 2) this particle is at its

physiological state but adapts a relative minor conformation; or 3) the current averaged projection is not thoroughly reflecting the structural features in the particle. Mathematically, unmatched particles are "outliers" and have heterogeneity (deviation from averages). The three categories of particulate disparity result in three categories of heterogeneities: the first category of heterogeneity is related to sample preparation, the second one is related to the true dynamics of biomolecules, and the third is usually related to symmetry mismatch phenomena in biomolecules. The rest of this dissertation will discuss these heterogeneities.

1.1.3 Issues in cryoEM sample preparation

Biomolecule samples for cryoEM are usually first checked with a conventional sample examination method such as negative stain, when the sample is first adsorbed on carbon film and then embedded into heavy metal salt; a good negative stain image is often a prerequisite to perform cryoEM. However, when checked with cryoEM, samples often have problems[4]: the biomolecules no longer look like the negative stain result and are not actually sufficient for single particle analysis. Optimizing the cryoEM sample is largely case by case; no reliable guidelines of cryoEM sample preparation currently exist. Remaining questions include how the thin buffer film influences the distribution and conformation of biomolecules and how to consistently optimize samples. Unnatural heterogeneities caused by improper sample preparation will introduce artifacts in the future "averaging-based" data processing.

1.2 Nanomachines

A machine is defined as "an apparatus using mechanical power and having several parts, each with a definite function and together performing a particular task"[5]. As a machine's component parts shrink to a scale of several nanometers and the entire assembly shrinks to a scale of hundreds of nanometers, whether a nanomachine inherits all the key characteristics of a macroscopic machine

remains an open question. When quantum effects play a key role in the determination of electron orbits and bond formation, a molecule-processing nanomachine falls between quantum effects and macroscopic mechanical effects, serving as a hybrid in terms of working principle, and a bridge in terms of transmitting microscopic/macroscopic signals to each other's regime.

The most commonly found nanomachines are biomolecular complexes, constructs of more than one biomolecule, that can be further categorized into protein complexes, RNA/DNA-protein complexes, protein-lipid complexes, etc. These biomolecular complexes are capable of selectively interacting with other molecules and often introduce changes to the nearby environment (morphing a molecule to a different conformation, splicing other molecules, etc.), much like a macroscopic machine. On the other hand, some nanomachines can also form new covalent bonds, such as ribosomes (forming peptide bond) and polymerase complexes (forming phosphodiester bond). These processes (such as reading template information and catalyzing bond formation) require quantum level accuracy and are unique to nanomachines.

To understand the exact working principle of nanomachines, we resolve its structure by building its atomic model. In this model, the position of each individual atoms inside a biomolecule complex is determined. The model not only shows side chain information of amino acids at sub-nanometer accuracy, but also a hierarchical structure (secondary, tertiary and quaternary structures) that reflects the overall mechanism.

However, resolving one single structure of a nanomachine is not sufficient to understand the dynamics and function of the entire nanomachine. Although one single structure of a nanomachine can reveal its components, mechanical movements are involved in functional nanomachines, and it is difficult to guess crucial conformational changes from this static result. As detailed below, although intrinsic heterogeneities can be observed in nanomachines, these

heterogeneities are difficult to resolve, needing more analysis to convert heterogeneities to informative dynamics.

1.2.1 Spatial heterogeneity

There are two kinds of biomolecular complexes: the minor class consists of those with no symmetric components (like ribosomes), and the major class consists of those with symmetric parts[6, 7]. It has been reported that high-order of symmetry can increase coding efficiency[8] and inhibit aggregation by improving stability[6], both of which benefit the assembly of large complexes. A protein complex with a true symmetry of high order is relatively easier to resolve, because this symmetry can be imposed to mathematically boost the dataset's size during data processing. This results in a better signal-to-noise ratio (SNR), which contributes to a more accurate structure with higher resolution.

However, applying symmetry directly is not always a panacea, and the assumption of a true high-order symmetry has greater impact as the model reaches atomic resolution. If monomers in each asymmetric unit are not exactly in a symmetry-related equivalent position, applying this symmetry will smear the structure rather than align the signal contribution from biomolecules. These symmetry-related heterogeneities will disturb the presumed spatial operation, so they are classified into spatial heterogeneities.

Notably, these heterogeneities are part of the complex structure; understanding them is the key step to studying the structure. These spatial heterogeneities can be classified into two categories, conformational changes and symmetry mismatch, described in the following two examples.

1.2.1.1 Conformational Changes − the vault organelle

Larger complexes (greater than 50 nm in length) have more surface area to have extensive environmental interactions, which can potentially influence the overall structure. Also, some complexes with high symmetry stabilize themselves with side chain interactions between neighboring monomers, which allows errors to accumulate in the long range and disrupts the overall symmetry. In some extreme cases, those complexes can be empty inside, introducing even more flexibility to the complex.

With its function remaining unknown 30 years after its first discovery in UCLA[9], the organelle vault has continuously piqued interest from researchers. As biological experiments alone have not yet effectively explained the function of this commonly-found eukaryotic complex, understanding the vault's structure may eventually help unveil the mystery. However, the vault possesses all three problematic features mentioned above: it is roughly 60 by 35 by 35 nm in size, is majorly composed of 78 copies of protein monomer [~900 residues long, in dihedral-39 (D-39) symmetry[10]], and, by definition, is empty inside. Since the vault very likely deviates from ideal D-39 symmetry, these characteristics have hindered the resolving of high-resolution vault structure with good chemical statistics (clashing score, Ramachandran outliers, sidechain outliers, etc). As a result, the ambiguous structures of vaults have lead to numerous explanations of the vault's behavior and function[11-13], especially on whether this hollow barrel-like organelle opens and, if so, how it opens.

The vault's structure can not only imply its function, but also guide its engineering application: the major vault protein (MVP) can form a hollow nanoparticle by itself[14] which is capable of encapsulating drugs. As a natural complex found in humans, the assembled MVP is not neutralized by the immune system. Thus, the vault is not only a good drug delivery vehicle, but

also an effective vaccine platform: it can be processed by antigen-presenting cells (APC) so that the antigen fused with vault proteins is presented on APCs, thus triggering the downstream immune response[15]. However, these add-ons can potentially change the structure of the recombinant vault particle, so it is necessary to study the potential conformational change in these vault particles. These studies will help improve engineering of regions on the vault to carry larger and new categories of payloads.

The vault particles have potential heterogeneities introduced by both natural and engineering sources. A high-resolution structure of the vault in solution with a recombinant payload will improve our understanding of the mechanism and engineering potential of this nanomachine.

1.2.1.2 Symmetry mismatch − reovirus transcription complex

Different portions of a biomolecular complex can have different orders of local symmetry. If we apply a higher order of symmetry to a region with a lower order of symmetry, the lower-order region will be smeared, since there is a "symmetry mismatch" in this biomolecular molecule. Symmetry mismatch is commonly found in nanomachines. For example, GroEL-GroES[16] chaperonin machine can mediate protein folding, with GroEL in dihedral 7 (D7, 14 folds) symmetry and GroES in cyclic 7 (C7, 7 folds) symmetry.

This symmetry mismatching phenomenon occurs more often in viruses, nanomachines that carry their own genomes, which are encapsulated by capsids of very high order of symmetry (icosahedral or helical symmetry). This highly-symmetric genome-encapsulating region is necessary for viruses: the average weight of a DNA/RNA base is around 325 Dalton (Da), and three bases in a codon weigh 975 Da (1950 Da for double stranded genome), while the average weight of an amino acid residue is only 110 Da, making it imperative that minimal DNA/RNA is

used to code the capsid. In contrast with the highly-symmetric genome-encapsulating capsid, the virus only needs to carry one copy of the genome inside. Thus, genome-encapsulating proteins, which encapsulate the genome, have a higher symmetry than genome-processing proteins, which serially interact with the genome, e.g., the portal complexes (driving DNA genome movements, one in each virion) in herpesvirus and RNA-dependent-RNA polymerases (RdRp) in reovirus.

Reoviruses are double-stranded RNA (dsRNA) viruses that process 9-12 segments of genome[17]. Inside its icosahedral capsid, the RdRp is located under the vertices of the icosahedron. In this icosahedral symmetry (60 folds), each vertex is composed ten copies of the capsid shell protein (CSP): five CSPs in conformation A (CSP-A) are close to the central five-fold axis, associating with another five copies of CSP in conformation B (CSP-B), which locate further away from the central five-fold axis. There is a symmetry mismatch between CSP (C5) and RdRp (C1). Previously, an icosahedral symmetry was applied to the entire virus, and as a result, the *in situ* structure of RdRp was not resolved.

Resolving the *in situ* structure of RdRp is the key to understanding the endogenous transcription process of reoviruses: the exposed dsRNA genome will trigger an innate immune response through the Toll-Like Receptor 3/Interferon Regulatory Factor 3 (TLR3/IRF3) signaling pathway[18]. As a result, transcription is conducted endogenously inside the virus, and mRNA is then released through the capsid. Reovirus has a low particle-to-plaque-forming-unit (PFU) ratio[19], suggesting that each reovirus carries the entire segmented genome to reach this high infectivity. Since each reovirus genome has RNA segments that number no more than the number of vertices in the icosahedral virion, it has been suggested that each segment of the dsRNA genome is attached to one RdRp located under each vertex [20, 21]. In this way, each genome segment will

be replicated and transcribed without competing for RdRp[21, 22] and thus each virion is able to carry the entire genome and be infectious.

Resolving the true asymmetric structure of each vertex (relaxing symmetry from C5 to C1, 5 folds) and even the entire virus as an asymmetric virion (relaxing symmetry from icosahedral to C1, 60 folds) remains challenging: the symmetric region will create local minimums in the orientation searching space, resulting in multimodality (multiple peaks in searching). Thus, the resulting reconstruction will be biased towards the initial values of the refinement, and the correct asymmetric features cannot be resolved. As a result, new computational tools/protocols are needed to resolve asymmetric features from the overall symmetric structure. Among all the reoviruses, cytoplasmic polyhedrosis virus (CPV) is the most promising target to test these computation tools: the first cryoEM structure of atomic resolution was from the CPV capsid[23], which possesses a thin capsid (less influence from the capsid) and a relatively stronger asymmetric signal (10 segments of genome out of 12 vertices).

1.2.2 Temporal heterogeneity: study reovirus nanomachine in situ

In a nanomachine, heterogeneities can also be observed in the dimension of time. There are different states in a working nanomachine, and substrates are processed step by step. Thus, similar to a catalyst, a nanomachine, especially an enzyme, is expected to have intermediate states to facilitate overcoming high energy barriers between reactants and products. These intermediate states are not energy favorable but are naturally observed during a dynamic equilibrium, when abundant reactants generate a large influx of the intermediate products, all of which the downstream reaction is not quick enough to consume. Thus, the enzyme, interacting with the intermediate products, will fall into its intermediate states.

Although nanomachines can be paused by adding reactant analogs [*e.g.* adding Fludarabine (9-beta-D-arabinofuranosyl-2-fluoroadenine)[24] to DNA polymerase] or create a reactant depletion [*e.g.* adding no Nucleoside triphosphate (NTP) to RdRp], these paused, energy-favorable states may not accurately resemble the intermediate, energy-unfavorable states in a running nanomachine.

Another major concern about studying nanomachine dynamics is that the conformation of nanomachines is dependent on how it reaches its current state: some nanomachines can only work in the correct temporal context. For example, in reoviruses, adding dsRNA to the assembled core will not trigger transcription since the RdRp is inside the core's capsid: transcription happens only when the dsRNA genome is already packaged inside the assembled core. In reovirus, the assembly

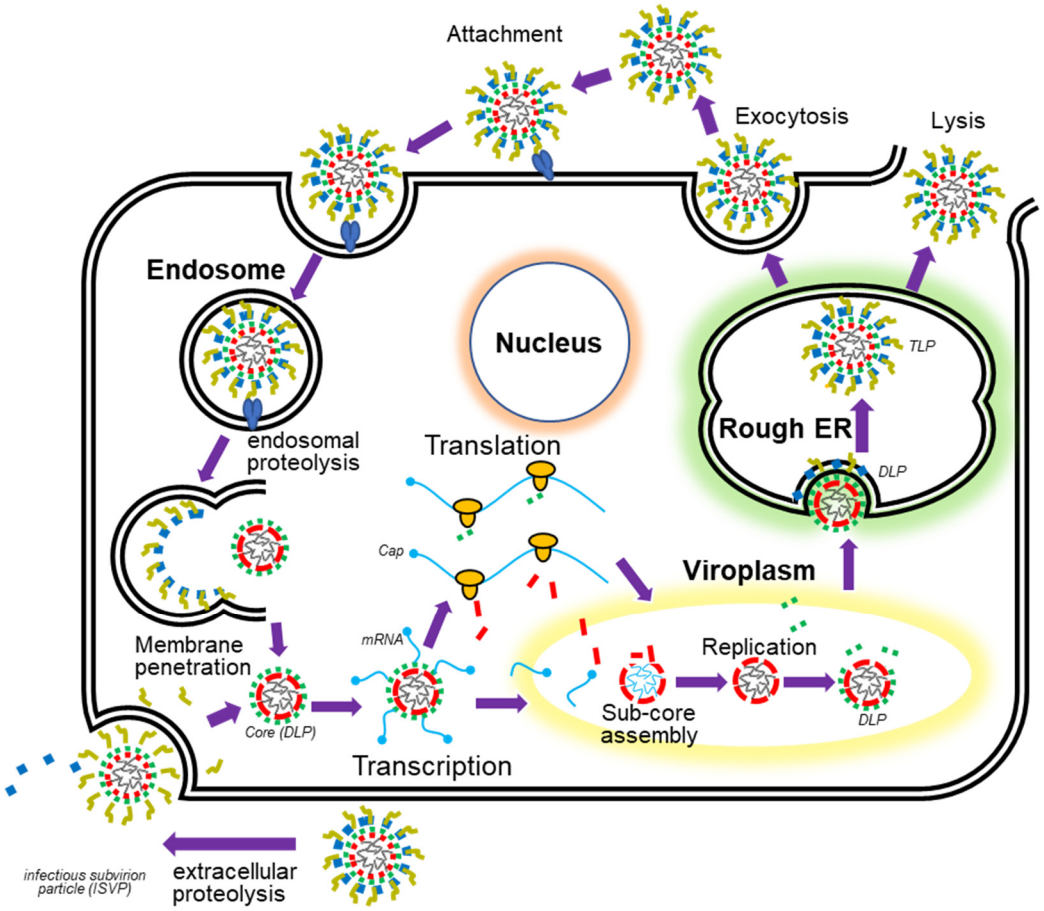


Figure 1.4 Reovirus life cycle

of capsid is a milestone in the viral life cycle and can influence the nanomachine by controlling the reactant/product trafficking through the capsid. Known as "RNA synthesis in a cage"[25], the capsid structure plays a key role in reovirus's state transition and the association of reactants. With some variations among different species, the general viral life cycle in Reoviridae can be summarized (Figure 1.4) as the following.

The first step is cell entry: the virus is first bound to cell surface by a receptor and then enters the cell through endocytosis. The outer layer of the capsid is then removed inside the endosome/lysosome through endosomal proteolysis, allowing the rest of the virion—the inside core—to penetrate the cytoplasm. This mechanism is found in many important mammalian pathogens, such as blue tongue virus (BTV)[26] and rotavirus[27]. In an alternative cell entry mechanism in viruses such as mammalian reovirus[28] and aquareovirus (ARV)[29], the outer layer of capsid is first removed by extracellular proteolysis to form infectious subvirion particle (ISVP). This ISVP then enters the cell through endocytosis. With the outer layer removed, the reovirus cores, located in the cytoplasm, enter the next step: transcription. In most circumstances, these cores are double-layered particles (DLPs). These proteolysis processes, endosomal or extracellular, activate the core[30] and thus can conduct transcription by adding NTP, S-Adenosyl methionine (SAM) and Mg^{2+}, even *in vitro*. The nascent RNA is then capped as messenger RNA (mRNA) and translated by ribosomes.

The second step is replication and capsid-shell assembly: plentiful mRNAs and translated viral proteins form an inclusion body called a viroplasm. In this viroplasm, each mRNA segment associates with RdRp and CSP. The assembly of the capsid shell and the replication of mRNA into the dsRNA genome then takes place in the viroplasm, resulting in a capsid shell with the dsRNA

genome replicated and packaged. Replication and assembly are synchronized so that each reovirus packages the complete set of genome segments. However, the exact mechanism remains unknown.

The third step is the outer layer capsid assembly, which greatly varies among different species. The outer layer assembly defines the two subfamilies of genera in Reoviridae: viruses in Spinareovirinae, such as CPV and ARV, process protein turrets on the inner capsid; while viruses in Sedoreovirinae, such as rotavirus and BTV, process no turrets and are smooth on their second layer. More variation can be found between different species: in CPV, the capsid is single layered, and the virion is further protected by cubic inclusion bodies; in ARV, the virion is assembled into a double layered particle with another coated protein (VP7) on the outer layer (two and a half layers in total) that will later be cleaved by extracellular proteolysis in the cell entry step; in rotavirus and BTV, two more layers of capsid are assembled to make a triple-layered particle (TLP) as a virion (Figure 1.4) in the rough endoplasmic reticulum (ER), and their outer layer can later respond to low pH for BTV[26] and to low Ca^{2+} concentration for rotavirus[31] in the late endosome during the cell entry step.

The last step is viral shedding. Virus-induced apoptosis will lyse the host to release the reoviruses. Exocytosis has also been also reported to be an alternative way to release the virion[32]. Virus release can also occur in the form of an occlusion body, such as in CPV.

Reovirus has several key states in its life cycle: the quiescent state (intact virion), primed state (ISVP, will start transcription with NTP+SAM+Mg), and the transcribing state (synthesizing and releasing nascent RNA). The most interesting biomolecule in the reovirus nanomachine is RdRp, because its conformational change directly reflects its polymerization activity at different states. Isolated from neighboring proteins (inner capsid protein, NTPase, etc.), reovirus RdRp structures have been resolved before[25, 33] to show a “right-handed-shaped” core, similar to

DNA polymerase, and two terminal domains (N- and C-terminal domains). However, previously resolved structures are not *in situ*. These structures not only are insufficient to understand RdRp's own function (terminal domains), but also fail to show how RdRp interacts with the genome and the capsid. Because the purified RdRp can be at an unnatural state, it is difficult to discuss state changes a non-functioning RdRp with significant components missing (CSP and entire genome).

For studying the transition between quiescent state to primed state, aquareovirus is the best model: it is in the quiescent state as a virion and can turn into ISVP through extracellular proteolysis, which can be triggered *in vitro*. Understanding how protein cleaving on the capsid outer layer influences the inner RdRp and genome structure will provide insights on this nanomachine's activation mechanism. Also, elucidating further details on capsid-RdRp interaction and genome conformational changes can potentially help to understand viral assembly mechanism.

To study the transition between primed state and transcribing state, we choose rotavirus as the model: by adding NTP+SAM+Mg *in vitro*, we can trigger the transcribing process *in vitro* and observe the conformational changes of CSP, RNA and RdRp. Seeing this nanomachine running will also provide hints on the function of the mysterious terminal domain, the transcript releasing mechanism, and the replication mechanism.

To study temporal heterogeneities in reovirus nanomachines, spatial heterogeneities (symmetry mismatching) must be resolved: resolving the asymmetrical RdRp structure underneath the vertices allows temporal heterogeneities to be resolved.

1.3 Dissertation outline

To address the possible unnatural heterogeneities introduced by cryoEM sample preparation, we discuss issues and possible solutions in Chapter 2. Next, two projects related to spatial

heterogeneities are presented: the vault organelle's true flexible region and engineering potential are discussed in Chapter 3, and CPV's uniform RNA genome structure is reported in Chapter 4. By overcoming both spatial and temporal heterogeneities, we further study the reovirus nanomachine with two other species: in ARV (Chapter 5), the activation and assembly mechanism is discussed; in rotavirus (Chapter 6), the transcribing and replication mechanism is elaborated. We conclude that all these heterogeneities required new tools to resolve, crucial to learning the structure and function of nanomachines, in Chapter 7.

1.4 References

1. Vainshtein, B.K., et al., *Concerning Angular Reconstitution - a Posteriori Assignment of Projection Directions for 3d Reconstruction.* Ultramicroscopy, 1988. **24**(1): p. 61-61.
2. Penczek, P.A., R.A. Grassucci, and J. Frank, *The Ribosome at Improved Resolution - New Techniques for Merging and Orientation Refinement in 3d Cryoelectron Microscopy of Biological Particles.* Ultramicroscopy, 1994. **53**(3): p. 251-270.
3. Moser, T.H., T. Shokuhfar, and J.E. Evans, *Considerations for imaging thick, low contrast, and beam sensitive samples with liquid cell transmission electron microscopy.* Micron, 2019. **117**: p. 8-15.
4. Lyumkis, D., *Challenges and opportunities in cryo-EM single-particle analysis.* Journal of Biological Chemistry, 2019. **294**(13): p. 5181-5197.
5. *machine, n.1.* Oxford English Dictionary: Oxford University Press, 2019.
6. Goodsell, D.S. and A.J. Olson, *Structural symmetry and protein function.* Annual Review of Biophysics and Biomolecular Structure, 2000. **29**: p. 105-153.
7. Plaxco, K.W. and M. Gross, *Protein Complexes: The Evolution of Symmetry.* Current Biology, 2009. **19**(1): p. R25-R26.

8. Crick, F.H.C. and J.D. Watson, *Structure of Small Viruses.* Nature, 1956. **177**(4506): p. 473-475.

9. Kedersha, N.L. and L.H. Rome, *Isolation and Characterization of a Novel Ribonucleoprotein Particle - Large Structures Contain a Single Species of Small Rna.* Journal of Cell Biology, 1986. **103**(3): p. 699-709.

10. Tanaka, H., et al., *The Structure of Rat Liver Vault at 3.5 Angstrom Resolution.* Science, 2009. **323**(5912): p. 384-388.

11. Mrazek, J., et al., *Polyribosomes Are Molecular 3D Nanoprinters That Orchestrate the Assembly of Vault Particles.* Acs Nano, 2014. **8**(11): p. 11552-11559.

12. Poderycki, M.J., et al., *The vault exterior shell is a dynamic structure that allows incorporation of vault-associated proteins into its interior.* Biochemistry, 2006. **45**(39): p. 12184-12193.

13. Anderson, D.H., et al., *Draft crystal structure of the vault shell at 9-angstrom resolution.* Plos Biology, 2007. **5**(11): p. 2661-2670.

14. Stephen, A.G., et al., *Assembly of vault-like particles in insect cells expressing only the major vault protein.* Journal of Biological Chemistry, 2001. **276**(26): p. 23217-23220.

15. Kar, U.K., et al., *Vault Nanocapsules as Adjuvants Favor Cell-Mediated over Antibody-Mediated Immune Responses following Immunization of Mice.* Plos One, 2012. **7**(7).

16. Hayer-Hartl, M., A. Bracher, and F.U. Hartl, *The GroEL-GroES Chaperonin Machine: A Nano-Cage for Protein Folding.* Trends in Biochemical Sciences, 2016. **41**(1): p. 62-76.

17. *Viral Entry into Host Cells.* Viral Entry into Host Cells, 2013. **790**: p. 1-199.

18. Kawai, T. and S. Akira, *Signaling to NF-kappa B by Toll-like receptors.* Trends in Molecular Medicine, 2007. **13**(11): p. 460-469.

19. Artinger, M., *Meeting Report VLPNPV: Session 10: Rapid total particle quantification.* Human Vaccines & Immunotherapeutics, 2014. **10**(10): p. 3083-3086.

20. Periz, J., et al., *Rotavirus mRNAS are released by transcript-specific channels in the double-layered viral capsid.* Proceedings of the National Academy of Sciences of the United States of America, 2013. **110**(29): p. 12042-12047.

21. Fajardo, T., et al., *Rotavirus Genomic RNA Complex Forms via Specific RNA-RNA Interactions: Disruption of RNA Complex Inhibits Virus Infectivity.* Viruses-Basel, 2017. **9**(7).

22. Fajardo, T., P.Y. Sung, and P. Roy, *Disruption of Specific RNA-RNA Interactions in a Double-Stranded RNA Virus Inhibits Genome Packaging and Virus Infectivity.* Plos Pathogens, 2015. **11**(12).

23. Yu, X.K., L. Jin, and Z.H. Zhou, *3.88 angstrom structure of cytoplasmic polyhedrosis virus by cryo-electron microscopy.* Nature, 2008. **453**(7193): p. 415-U73.

24. Huang, P., S. Chubb, and W. Plunkett, *Termination of DNA-Synthesis by 9-Beta-D-Arabinofuranosyl-2-Fluoroadenine - a Mechanism for Cytotoxicity.* Journal of Biological Chemistry, 1990. **265**(27): p. 16617-16625.

25. Tao, Y.Z., et al., *RNA synthesis in a cage - Structural studies of reovirus polymerase lambda 3.* Cell, 2002. **111**(5): p. 733-745.

26. Forzan, M., M. Marsh, and P. Roy, *Bluetongue virus entry into cells.* Journal of Virology, 2007. **81**(9): p. 4819-4827.

27. Baker, M. and B.V.V. Prasad, *Rotavirus Cell Entry.* Cell Entry by Non-Enveloped Viruses, 2010. **343**: p. 121-148.

28. Chandran, K. and M.L. Nibert, *Protease cleavage of reovirus capsid protein mu 1/mu 1C is blocked by alkyl sulfate detergents, yielding a new type of infectious subvirion particle.* Journal of Virology, 1998. **72**(1): p. 467-475.

29. Zhang, X., et al., *3.3 angstrom Cryo-EM Structure of a Nonenveloped Virus Reveals a Priming Mechanism for Cell Entry.* Cell, 2010. **141**(3): p. 472-482.

30. Faust, M. and S. Millward, *In vitro methylation of nascent reovirus mRNA by a virion-associated methyl transferase.* Nucleic Acids Research, 1974. **1**(12): p. 1739-52.

31. Aoki, S.T., et al., *Structure of Rotavirus Outer-Layer Protein VP7 Bound with a Neutralizing Fab.* Science, 2009. **324**(5933): p. 1444-1447.

32. Patton, J.T., et al., *Rotavirus genome replication and morphogenesis: Role of the viroplasm.* Reoviruses: Entry, Assembly and Morphogenesis, 2006. **309**: p. 169-187.

33. Lu, X., et al., *Mechanism for Coordinated RNA Packaging and Genome Replication by Rotavirus Polymerase VP1.* Structure, 2008. **16**(11): p. 1678-1688.

Chapter 2 Optimizing surface-dominating cryoEM sample

2.1 CryoEM sample are surface-dominating

Because the phase contrast generated in cryoEM is from the interference between elastically scattered electrons and background electrons, inelastically scattered electrons will create noise and decrease the contrast in the image. The inelastic scattering of electrons will increase as electrons travel through thicker samples, so cryo TEM samples are thinner than 500 nm in most cases[1]. For cryoEM, the liquid buffer thickness after blotting is often between 10 nm (Quantifoil carbon thickness) to 200 nm.

On the sample side, 50 kDa is currently considered to be cryoEM's lower limit in resolving structures [2]. Thus, assuming the average density of protein is 1.37 g/cm^3, the current smallest resolvable diameter, of a biomolecule with 50kDa weight, is 4.9 nm. The largest structure resolved by cryoEM with atomic resolution is of nucleocytoplasmic large DNA viruses[3], with 190nm diameter. The sample size (5 nm – 190 nm) fits in the optimal ice thickness of cryoEM (10 nm - 200 nm).

Electrolytes (anions and cations) in buffer will shield charges on protein surface. To quantify interactions between biomolecules, the Debye-Hückel length is used to describe the effective interacting distance between two charged objects in solution, defined by

$$\lambda_D = \left(\frac{\epsilon k_B T}{\sum_{j=1}^{N} n_j^0 q_j^2} \right)^{1/2}$$

where ϵ is the relative static permittivity of solvent, k_B is Boltzmann's constant, T is the temperature, n_j is the mean concentration of charges of the species j and q_j is the charge of the species j. In a typical physiological environment [1X Phosphate Buffered Saline (PBS)] at room

temperature, the Debye length is 0.7 nm[4], which is smaller than the diameter of alpha-helixes (1.2 nm) and much smaller than liquid thickness. This indicate that protein-protein interactions should remain identical to the situation in bulk.

The air-water interface plays a key role in cryoEM samples. The effective distance of air-water interface is described by Knudsen layer thickness, which is approximated by the mean free path l_c of air at room temperature T_s and atmosphere pressure P_s

$$l_c = \frac{k_B T_s}{\pi d^2 p_s}$$

where d is the molecular diameter. The mean free path of air at room temperature and atmosphere is 68 nm[5], and so is the distance of surface effect of the air-water interface at water side[6]. This suggests that for a typical cryoEM sample, the liquid thin layer after blotting is dominated by the influence of the air-water interface.

2.1.1 Experimental evidence

This surface adsorption of protein has been confirmed experimentally by X-ray reflection[7, 8] and spectroscopy[9]. Several electron microscopy results give direct and quantitative evidence of protein surface adsorption on a thin film of buffer.

2.1.1.1 Cryo electron tomography results

Cryo electron tomography (cryoET) is a method to obtain a 3-dimensional reconstruction of samples embedded in vitrified ice. Different from cryoEM, which takes no-tilt images of many particles, cryoET images the same region on the grid many times at different angles, and those images of the same region are reconstructed in 3D. Thus, the position of each biomolecule in the sample can be accurately labeled[10] (Figure 2.1).

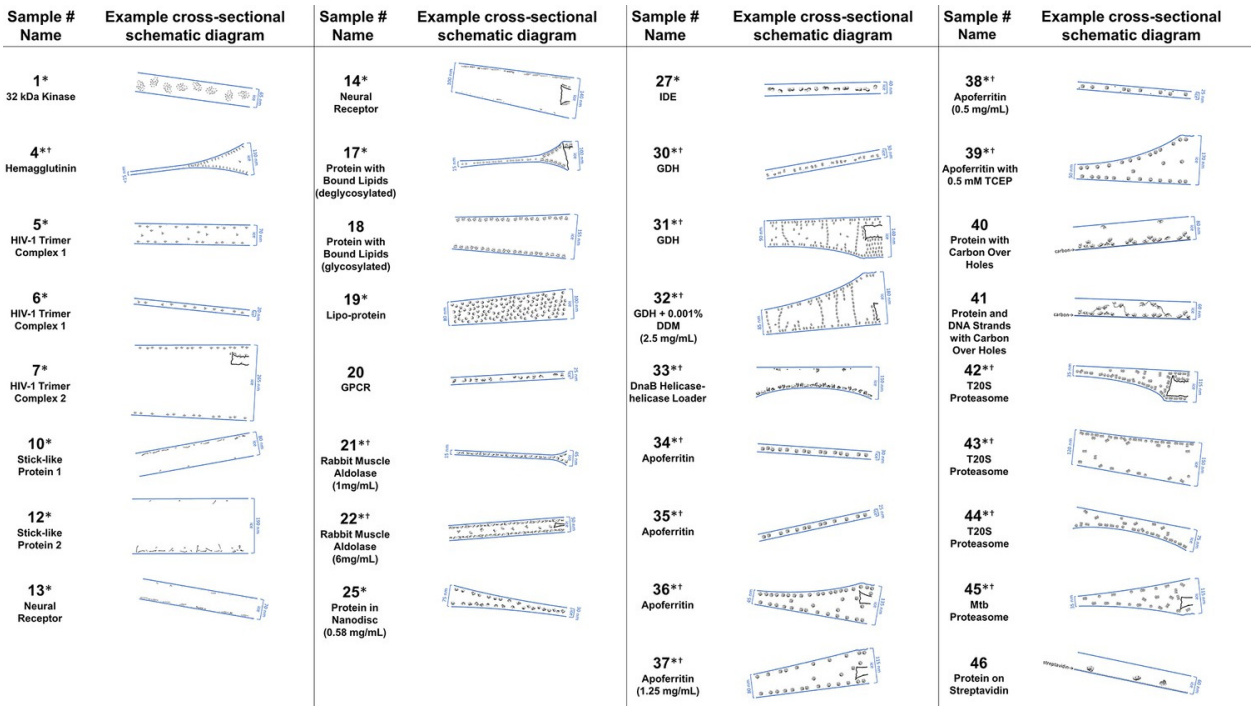


Figure 2.1 A selection of cross-sectional schematic diagrams of particle and ice behaviors in holes as depicted according to analysis of individual tomograms. The relative thicknesses of the ice in the cross-sections are depicted accurately. Each diagram is tilted corresponding to the tomogram from which it is derived; i.e. the depicted tilts represent the orientation of the objects in the field of view at zero-degree nominal stage tilt. If the sample concentration in solution is known, then it has been included below the sample name. Black lines on schematic edges are the grid film. The cross- sectional characteristics depicted here are not necessarily representative of the aggregate [10]

From this result, we can see that for most proteins in thin layer of liquid, the liquid thickness is close to 50 nm. Excepting several samples (no. 19, 31, 32), biomolecules all show strong affinity to the air water interface. There are three typical types of sample distribution and liquid shape: double adsorption, two-sided adsorption, and curved adsorption. The first type is a "double adsorption" type (no. 6, 17, 21, 27, 34, 35, 38), when only a single layer of protein is embedded in a thin layer of buffer, and the thickness of the buffer is close to the dimension of the protein. The second type is "two-sided adsorption," when the ice thickness is usually larger than 50 nm and biomolecules are adsorbed on both ends of the buffer thin film, forming two distinct flat layers of biomolecules (no. 5, 7, 10, 12, 13, 14, 18, 22, 43, 45). In this type, one side can have relatively less particles than the other side (no. 7, 12, 13, 14, 43). The third type is "curved adsorption," when the

buffer is one side curved with the opposite side flat (no. 33, 36, 39, 42, 44). The curved side usually has more adsorbed biomolecules than the flat side (no. 33, 42, 44). Based on these three major types, we can conclude that the biomolecules are indeed adsorbed on the air-water interface in the thin buffer film.

2.1.1.2 Single particle results

To quantify surface adsorption, we analyse three cryoEM samples with measured protein concentration in bulk and particle molecular weight: GroEL[11] (800kDa), β-galactosidase[12] (465kDa) and haemoglobin[13] (64kDa). By assuming the average ice thickness is 50 nm, we calculate the theoretical value of particles in each image and compare it with the actual number of particles in the micrograph (Table 2.1).

	GroEL	**β-galactosidase**	**haemoglobin**
Concentration (mg/mL)	0.1	2.3	1.5
Molecular weight (kDa)	800	465	64
View volume (mL)	9.43E-16	2.75E-15	4.18E-15
Theoretical Number in the view	0.73	8.19	21.13
Actual number in the view	99	110	240
Actual/Theoretical	135.62	13.43	11.36

Table 2.1 Calculated surface adsorption based on cryoEM micrographs. The Actual particle number under microscope is at least ten times more than theoretical value.

From this calculation, we see that surface adsorption can increase the average number of particles boxed in each micrograph by at least 5 times. This means that the majority of the particles we see in the micrograph are under the influence of air-water interface.

2.1.2 Explanations on the distribution of cryoEM sample

The first type of surface adsorption, double adsorption, can happen in two steps. In single adsorption, after blotting, the protein is first adsorbed on one surface. The overall surface energy in bulk buffer (without adsorption) and single adsorption are defined as:

$$E_{bulk} = 2A_1\gamma_{lg} + 2A_2\gamma_{ls}$$

$$E_{single\_adsorption} = 2A_1\gamma_{lg} + A_2\gamma_{ls} + A_2\gamma_{sg} - A_2\gamma_{lg}$$

where γ_{lg}, γ_{ls} and γ_{sg} are surface energy of buffer-air, buffer-protein, and protein-air interface, respectively. A_1 is the size of the hole and A_2 is the top area of the biomolecule. The energy difference between single adsorption and bulk is

$$E_{single\_adsorption} - E_{bulk} = A_2\gamma_{sg} - A_2\gamma_{lg} - A_2\gamma_{ls}$$

which is usually less than 0 because surface tension of water (γ_{lg}) is large (72mN/m). As a result, most protein added to water will create surface excess and decrease overall surface tension.

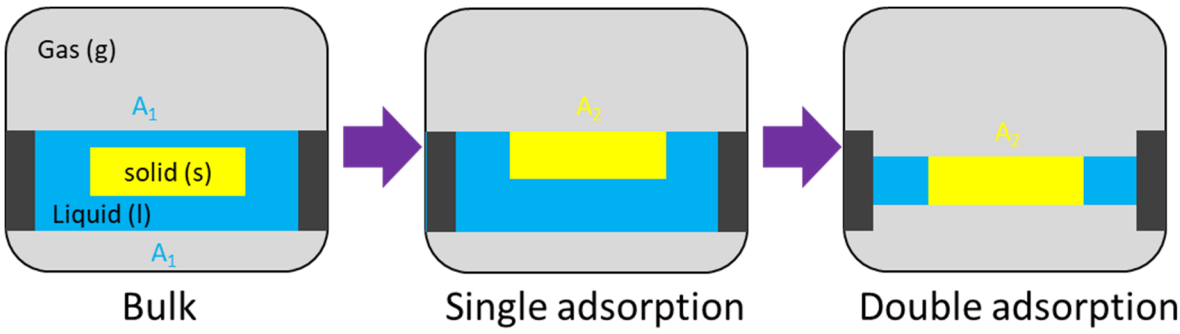


Figure 2.2 Two steps of protein changing the geometry of thin buffer layer

After the first “single adsorption”, the biomolecule can potentially further change the geometry of the thin buffer film by further thinning it in "double adsorption". The energy difference is

$$E_{double\_adsorption}\text{-}E_{single\_adsorption} = A_2\gamma_{sg}\text{-}A_2\gamma_{lg}\text{-}A_2\gamma_{ls}$$

which is also usually less than 0. This "double adsorption" is energetically favored and can be a result of Marangoni flow, which is when buffer is transferred to regions with less surface excess. This process will create a thin layer of buffer film with biomolecules embedded. This first type of adsorption is explained in Figure 2.2.

The difference between the other two types of adsorption, "two-sided adsorption" and "curved adsorption", is primarily the amount of buffer left on the grid. The filter paper usually removes buffer from one side, so the opposite side of the air-water interface is preserved.

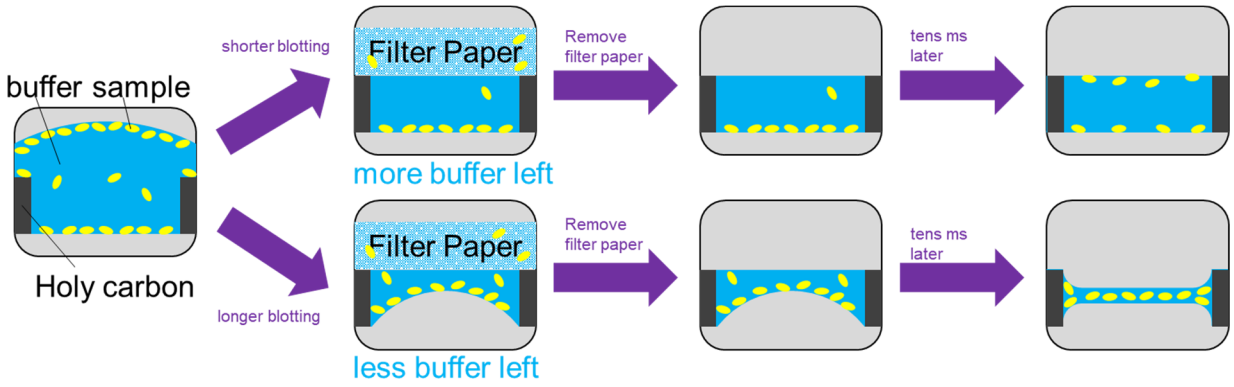


Figure 2.3 Protein adsorption is related to the amount of remaining buffer. More solution left (upper) will lead to "two-sided adsorption", and less solution left (bottom) will lead to "Curved adsorption" and possible "double adsorption" eventually.

After the filter paper is removed, the particles will diffuse, redistributing across the thin buffer film. Based on the formula to calculate the diffusion constant of a spherical object:

$$\frac{\bar{x}^2}{t} = D = \frac{k_B T}{6\pi\eta r}$$

Given $\bar{x} = 200$ nm as the maximum thickness of cryoEM sample, k_B as Boltzmann's constant, $T = 300$ K as room temperature, $\eta = 1$ mPa $\cdot$ S as viscosity, and $r = 7$ nm as average size for a 1.2 MDa protein complex, the resulting $t = 1.3$ mS. The larger the particle, the longer the diffusion

time (e.g., 6.4 ms for rotavirus with 70 nm diameter). Thus, for spherical or near-spherical biomolecules, after blotting, the biomolecule should already have been adsorbed to the air-water interface before plunge freezing. This theoretical value is consistent with previous estimations of the time scale, estimated in the tens of milliseconds [10, 14]. If the protein is already adsorbed on the air-water interface, the free diffusion model will no longer be valid, and proteins will take longer to diffuse to the new air-water interface; this may contribute to the observed uneven particle distribution commonly found in thick samples (Figure 2.1 and 2.3).

In conclusion, surface adsorption is commonly found in cryoEM samples: the apparent concentration is improved after surface adsorption, and the thin buffer layer is influenced by blotting conditions and biomolecule properties.

2.2 CryoEM samples are not in their exact physiological state

There has been a debate[15] about which sample preparation method keeps biomolecules in its natural state: frozen in a thin layer of buffer or tightly packed in a crystal. Although it has been reported that some biomolecules naturally form micro-crystals *in vivo*, most nanomachines are not naturally packed, because packed nanomachines are not likely to have enough access to reactants and room to release products.

However, an air-water interface is rare in cells and will have more influence on biomolecules in a thin film of buffer. Although cryoEM samples are closer to their physiological states in terms of molecular interaction, pH, ionic strength, etc. than crystals are, the air-water interface can introduce dramatic changes to some molecules; thus, we cannot assume that cryoEM samples stay perfectly in their physiological states.

2.2.1 Particle deformation

The first concern of proteins at the surface is deformation: the adsorbed proteins will undergo deformation[8, 16] and usually process a flatter and thinner conformation due to the gradient of

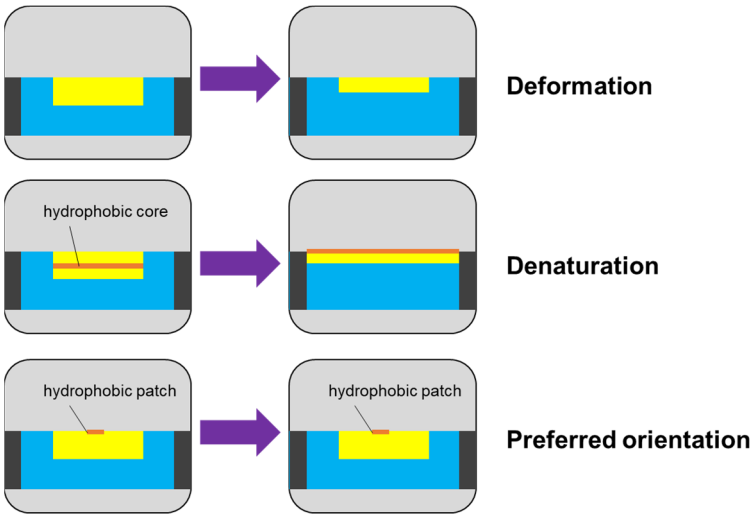


Figure 2.4 Three types of non-physiological phenomena found in cryoEM samples: deformation, denaturation and preferred orientation.

surface tension on the air-water interface (Figure 2.4). It has also been reported that biomolecules with hydrophilic and hydrophobic patches are less deformed[8], suggesting that making a hydrophilic protein surface may not reduce deformation. This deformation introduces anisotropic heterogeneity and can create an issue in subsequent averaging steps.

2.2.2 Biomolecule denaturation

The folding of secondary, super-secondary, and tertiary structure of proteins is largely dependent on the hydrophobic core[17]. It has been proven [8, 18, 19] that proteins attached to the air-water interface will unfold and denature when the hydrophobic core is exposed to the air (Figure 2.4). The scale of denaturing time varies between different proteins but can be within 1 second[7]. Protein adsorption to the air-water interface is quite dynamic; it has been reported[7, 8] that when the kinetics of protein folding are faster than the kinetics of protein adsorption, the protein tends

to denature. The denatured protein irreversibly loses its tertiary structures and changes from the physiological state.

2.2.3 Preferred orientation

Preferred orientation is a non-physiological state specifically related to single particle analysis in cryoEM: the biomolecules will have a preferred orientation when adsorbed on the air-water interface, either a hydrophobic patch on the surface of the protein or a flat surface of the protein turning towards the air.

The issue of preferred orientation in the single particle analysis can be described as the following: each particle is reconstructed in 3D Fourier space based on its orientation; however, because of the preferred orientation issue, particles in the thin buffer film are in similar orientations, and the reconstructed Fourier space will be incomplete. As images from cryoEM are transmissive, the information along the observing direction will be lost.

The preferred orientation issue, a significant challenge in single particle analysis commonly found in variety of samples[20-23], is strongly related to interface adsorption.

2.3 Surfactant application in cryoEM

To overcome these surface adsorption-related issues, surfactants are introduced to change the property of the air-water interface in the thin buffer sample by creating surface excess [21, 23]. A surfactant molecule usually has a hydrophilic head and a hydrophobic tail, making it more energy favorable when adsorbed on the air-water interface. However, adding surfactant before blotting can significantly decrease the surface adsorption of protein, resulting in fewer particles in each micrograph. A relatively higher concentration of protein (2mg/mL[24, 25]) is needed to obtain enough particles in solution (Figure 2.5).

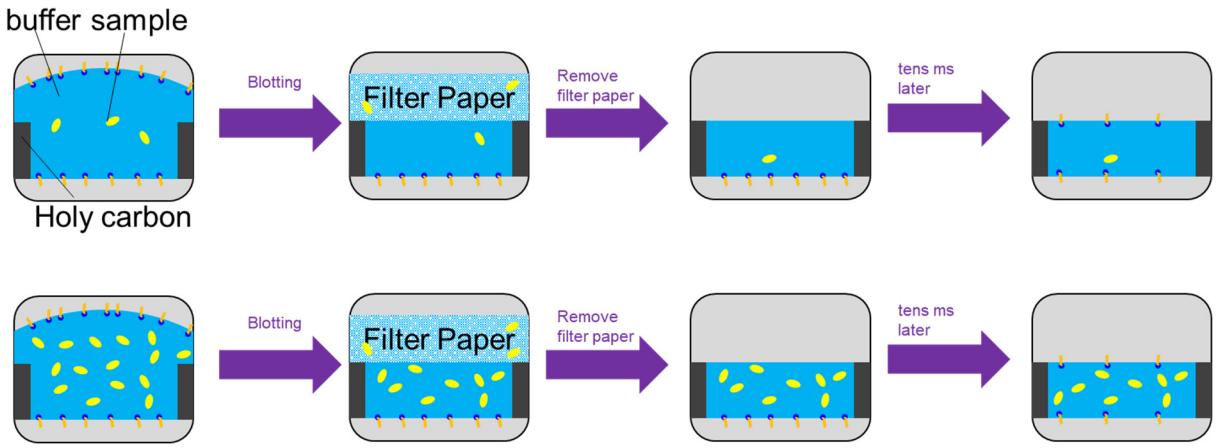


Figure 2.5 Surfactant's effect on low concentration sample (upper) and high concentration sample (lower).

Surfactants are commonly used in biomolecule research: ionic surfactants, such as *sodium dodecyl sulfate* (*SDS*), can denature proteins in SDS-PAGE technique; and non-ionic surfactants[26], such as *octyl-b-D-glucoside* and *5-cyclohexyl-1-pentyl-b-D-maltoside,* can extract membrane proteins from the lipid bilayer. It is important to control the surfactant type and concentration so that the protein structure is not disturbed; notably, ionic surfactants can potentially denature the protein. Another concern is micelles: those with similar sizes to the protein samples in solution can smear the micrograph and increase background noise. For example, the commonly used surfactant *n-Dodecyl β-D-maltoside*'s (DDM) micelle is 72kDa and is highly comparable with small biomolecules in cryoEM (~200kDa) (Figure 2.6). Further, surfactants that

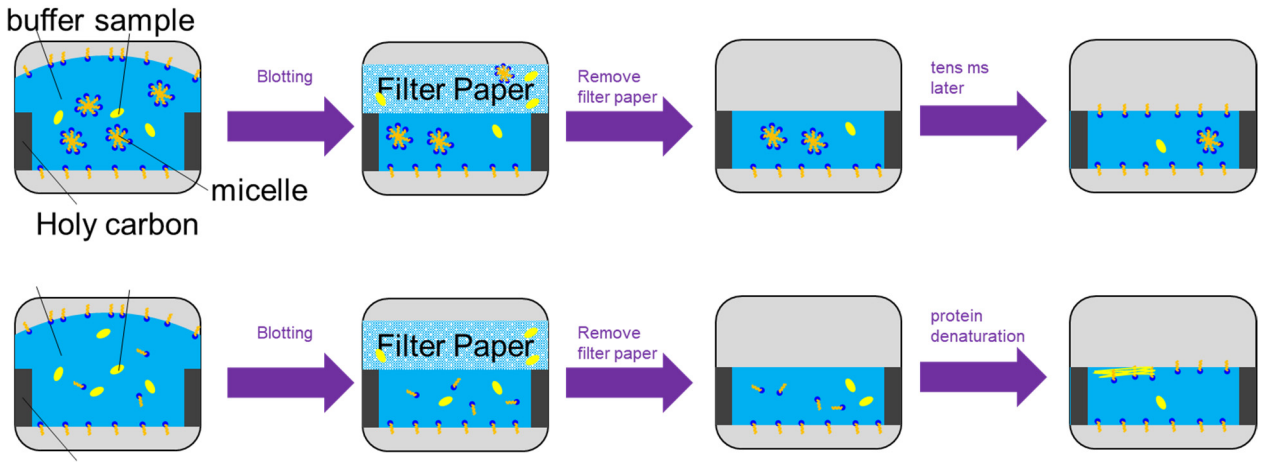


Figure 2.6 Potential artifacts introduced by surfactants. Micelles (upper) and protein denaturation (bottom)

work in some projects may not work in others [5]. The range of surfactants is broad, including DDM[10], Tween 20[27], NP-40[24, 25], etc. Weaker surfactants are not able to create enough surface pressure to block the surface[28]. Thus, it is important to search for the appropriate surfactant for each individual project.

2.3.1 Fluorinated surfactant

Fluorinated fos-choline 8 has a tail that is both hydrophobic and oleophobic. Several publications[29-31] and reviews[28] have shown that this fluorinated surfactant is able to help form a thin buffer layer with proteins embedded (Figure 2.7).

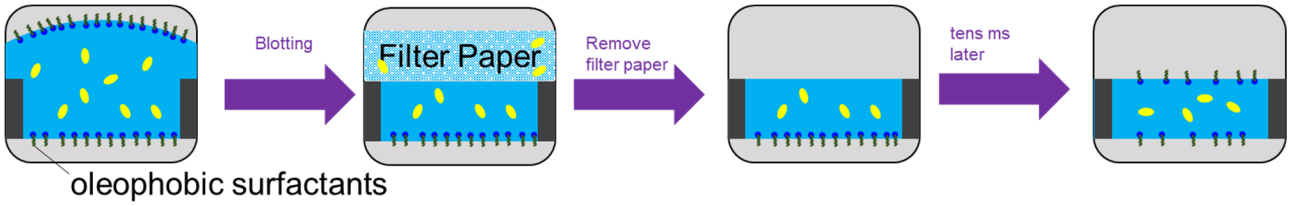


Figure 2.6 Fluorinated surfactant can maintain surface excess after blotting

Fluorinated fos-choline 8 has a high critical micelle concentration (CMC, 2.9mM) compared with DDM (0.15 mM), Tween 20 (0.059 mM) and NP-40 (0.29mM). This high CMC value allows for more surfactant molecules in the buffer without forming micelles. Fluorinated surfactants also have higher surface pressure than hydrocarbon surfactants[32]. This feature helps to block the protein from the air-water interface. Additionally, the oleophobic tail makes fluorinated surfactants "milder" and less likely to denature the biomolecule on its own[33]. However, the exact effect of fluorinated fos-choline 8 on biomolecules remains unknown; further studies are needed.

2.3.2 Gaseous Surfactant

Because the ideal surfactant for blotting and for cryoEM imaging might not be the same compound, surfactants should be introduced to the thin buffer layer after blotting. New surfactants should be evenly added to the entire grid; otherwise, Marangoni flow can deplete regions with higher surfactant concentration.

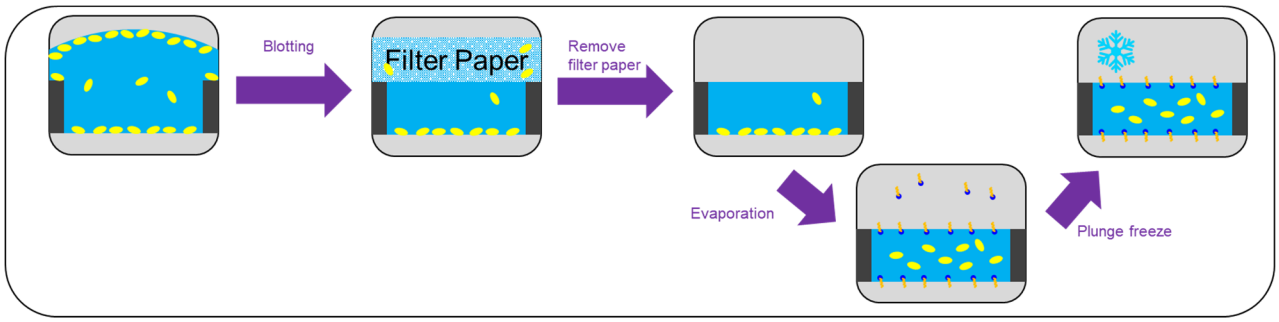


Figure 2.7 Gaseous surfactants to be added to thin buffer film

Gaseous surfactants can be introduced to the thin buffer film evenly after blotting (Figure 2.8). Fatty alcohols are volatile compounds that can work as surfactants at the air-water interface. Potential candidates are *n-amyl* and *n-decyl alcohol*[34]. Gaseous surfactants for industrial use (e.g., *3,5-Dimethyl-1-hexyn-3-ol*) are toxic for biological experiments for cryoEM.

2.4 Conclusion

In this chapter, we describe the typical shapes of thin buffer layer and discuss the particle distribution in cryoEM samples, which are greatly influenced by the air-water interface. We then discuss the principles behind this observation and find that the shape of the thin buffer film is influenced by three factors: blotting, biomolecular properties, and sample concentration. Adsorbed proteins can be in non-physiological states, thus introducing unnatural heterogeneities, which hinder future averaging in single particle analysis. These issues can be addressed by carefully applying surfactants during sample preparation, but it is not a foolproof solution: a relatively higher concentration of sample is needed. Although fluorinated fos-choline 8 is a promising surfactant to

maintain a thin buffer film while keeping the sample inside the film, the exact benefits at molecular level are not fully understood. Thus, there is currently no reliable way to prevent heterogeneities in cryoEM sample preparation.

2.5 References

1. Oikonomou, C.M., et al., *Electron cryotomography.* Imaging Bacterial Molecules, Structures and Cells, 2016. **43**: p. 115-139.
2. Liu, Y.X., D.T. Huynh, and T.O. Yeates, *A 3.8 angstrom resolution cryo-EM structure of a small protein bound to an imaging scaffold.* Nature Communications, 2019. **10**.
3. Fang, Q.L., et al., *Near-atomic structure of a giant virus.* Nature Communications, 2019. **10**.
4. Chu, C.H., et al., *Beyond the Debye length in high ionic strength solution: direct protein detection with field-effect transistors (FETs) in human serum.* Scientific Reports, 2017. **7**.
5. Glaeser, R.M., et al., *Factors that Influence the Formation and Stability of Thin, Cryo-EM Specimens.* Biophysical Journal, 2016. **110**(4): p. 749-755.
6. Gatapova, E.Y., et al., *The temperature jump at water - air interface during evaporation.* International Journal of Heat and Mass Transfer, 2017. **104**: p. 800-812.
7. Yano, Y.F., et al., *Real-time investigation of protein unfolding at an air-water interface at the 1 s time scale.* Journal of Synchrotron Radiation, 2013. **20**: p. 980-983.
8. Yano, Y.F., et al., *Initial Conformation of Adsorbed Proteins at an Air-Water Interface.* Journal of Physical Chemistry B, 2018. **122**(17): p. 4662-4666.
9. de Jongh, H.H.J., et al., *Protein adsorption at air-water interfaces: A combination of details.* Biopolymers, 2004. **74**(1-2): p. 131-135.

10. Noble, A.J., et al., *Routine single particle cryoEM sample and grid characterization by tomography.* Elife, 2018. **7**.

11. Roh, S.H., et al., *Subunit conformational variation within individual GroEL oligomers resolved by Cryo-EM.* Proceedings of the National Academy of Sciences of the United States of America, 2017. **114**(31): p. 8259-8264.

12. Bartesaghi, A., et al., *2.2 angstrom resolution cryo-EM structure of beta-galactosidase in complex with a cell-permeant inhibitor.* Science, 2015. **348**(6239): p. 1147-1151.

13. Khoshouei, M., et al., *Cryo-EM structure of haemoglobin at 3.2 angstrom determined with the Volta phase plate.* Nature Communications, 2017. **8**.

14. Noble, A.J., et al., *Reducing effects of particle adsorption to the air-water interface in cryo-EM.* Nature Methods, 2018. **15**(10): p. 793-+.

15. Shoemaker, S.C. and N. Ando, *X-rays in the Cryo-Electron Microscopy Era: Structural Biology's Dynamic Future.* Biochemistry, 2018. **57**(3): p. 277-285.

16. Wierenga, P.A., et al., *The adsorption and unfolding kinetics determines the folding state of proteins at the air-water interface and thereby the equation of state.* Journal of Colloid and Interface Science, 2006. **299**(2): p. 850-857.

17. Kalinowska, B., et al., *Is the hydrophobic core a universal structural element in proteins?* Journal of Molecular Modeling, 2017. **23**(7).

18. D'Imprima, E., et al., *Protein denaturation at the air-water interface and how to prevent it.* Elife, 2019. **8**.

19. Koepf, E., et al., *Notorious but not understood: How liquid-air interfacial stress triggers protein aggregation.* International Journal of Pharmaceutics, 2018. **537**(1-2): p. 202-212.

20. Glaeser, R.M., *How good can cryo-EM become?* Nature Methods, 2016. **13**(1): p. 28-32.

21. Glaeser, R.M. and B.G. Han, *Opinion: hazards faced by macromolecules when confined to thin aqueous films.* Biophys Rep, 2017. **3**(1): p. 1-7.

22. Tan, Y.Z., et al., *Addressing preferred specimen orientation in single-particle cryo-EM through tilting.* Nature Methods, 2017. **14**(8): p. 793-+.

23. Drulyte, I., et al., *Approaches to altering particle distributions in cryo-electron microscopy sample preparation.* Acta Crystallographica Section D-Structural Biology, 2018. **74**: p. 560-571.

24. Zhu, Y.A., et al., *Structural mechanism for nucleotide-driven remodeling of the AAA-ATPase unfoldase in the activated human 26S proteasome.* Nature Communications, 2018. **9**.

25. Dong, Y.C., et al., *Cryo-EM structures and dynamics of substrate-engaged human 26S proteasome.* Nature, 2019. **565**(7737): p. 49-+.

26. Arachea, B.T., et al., *Detergent selection for enhanced extraction of membrane proteins.* Protein Expression and Purification, 2012. **86**(1): p. 12-20.

27. Gunning, P.A., et al., *Effect of surfactant type on surfactant-protein interactions at the air-water interface.* Biomacromolecules, 2004. **5**(3): p. 984-991.

28. Glaeser, R.M., *Proteins, interfaces, and cryo-EM grids.* Current Opinion in Colloid & Interface Science, 2018. **34**: p. 1-8.

29. Johnson, Z.L. and J. Chen, *Structural Basis of Substrate Recognition by the Multidrug Resistance Protein MRP1.* Cell, 2017. **168**(6): p. 1075-+.

30. Hughes, T.E.T., et al., *Structural basis of TRPV5 channel inhibition by econazole revealed by cryo-EM.* Nature Structural & Molecular Biology, 2018. **25**(1): p. 53-+.

31. Basak, S., et al., *Cryo-EM reveals two distinct serotonin-bound conformations of full-length 5-HT3A receptor.* Nature, 2018. **563**(7730): p. 270-+.

32. Kovalchuk, N.M., et al., *Fluoro- vs hydrocarbon surfactants: Why do they differ in wetting performance?* Advances in Colloid and Interface Science, 2014. **210**: p. 65-71.

33. Park, K.H., et al., *Fluorinated and hemifluorinated surfactants as alternatives to detergents for membrane protein cell-free synthesis.* Biochemical Journal, 2007. **403**: p. 183-187.

34. Malysa, K., K. Khristov, and D. Exerowa, *Surfactant in the Gaseous-Phase .1. Formation of Foams and Thin Liquid-Films.* Colloid and Polymer Science, 1991. **269**(10): p. 1045-1054.

Chapter 3 Solution structures of engineered vault particles

Ke Ding[1,2,3], Xing Zhang[2,3], Jan Mrazek[4,5], Valerie A. Kickhoefer[4], Mason Lai[3], Hwee L. Ng[5], Otto O. Yang[2,3,5,6], Leonard H. Rome [2,4], Z. Hong Zhou[1,2,3,7*]

[1]Department of Bioengineering, University of California, Los Angeles, California 90095, USA

[2]California NanoSystems Institute, University of California, Los Angeles, California 90095, USA

[3]Department of Microbiology, Immunology and Molecular Genetics, University of California, Los Angeles, California 90095, USA

[4]Department of Biological Chemistry, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, California 90095, USA

[5]Division of Infectious Diseases, Department of Medicine, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, California 90095, USA

[6]AIDS Healthcare Foundation, Los Angeles, California 90028, USA

[7]Lead Contact

*To whom correspondence should be addressed

Phone: 310-983-1033, Fax: 310-206-5365; E-mail: Hong.Zhou@UCLA.edu

3.1 Abstract

Prior crystal structures of the vault have provided clues of its structural variability but are non-conclusive due to crystal packing. Here, we obtained vaults by engineering at the N-terminus of rat major vault protein (MVP) an HIV-1 Gag protein segment and determined their near-atomic resolution (~4.8 Å) structures in a solution/non-crystalline environment. The barrel-shaped vaults in solution adopt two conformations, 1 and 2, both with D39 symmetry. From N- to C-terminus, each MVP monomer has three regions: body, shoulder and cap. While conformation 1 is identical to one of the crystal structures, the shoulder in conformation 2 is translocated longitudinally up to 10 Å, resulting in an outward-projected cap. Our structures clarify the structural discrepancies in the body region in the prior crystallography models. The vault's drug-delivery potential is highlighted by the internal disposition and structural flexibility of its Gag-loaded N-terminal extension at the barrel waist of the engineered vault.

3.2 Introduction

First discovered in the mid-1980's [1], vaults are large barrel-shaped ribonucleoprotein complexes existing in most eukaryotic cells. The vault's function remains unknown. Each native vault has a mass of 13 MDa and is composed of multiple copies of at least three different proteins—the major vault protein (MVP, 100 kDa), vault poly(ADP-ribose) polymerase (VPARP), and telomerase-associated protein 1 (TEP1)—and several copies of small untranslated vault-associated RNA (vRNAs). Recombinantly expressed MVP assembles into vault-like nano-particles that can package small molecules and protein antigens, thus can be engineered to deliver drugs and vaccines, respectively. Its expression alone results in an ordered assembly of hollow barrel-shaped structures [2, 3], whose morphology is identical to natural vaults when examined by electron microscopy. Peptides genetically fused to the N-terminus of MVP were found inside the vault at the waist of the barrel [4], while those fused to the C-terminus were outside the vault at its two poles of the barrel [5]. Recombinant vaults have been engineered to enable cell targeting [5], to trigger specific immune responses [6], and to deliver drugs [7], demonstrating great potential for biomedical applications [8, 9]. However, the detailed structures of these recombinant vaults are not known, hampering efforts to engineer MVP for such applications.

Early structural characterization of the vault was accomplished by cryo electron microscopy (cryoEM) and single particle analysis with resolution limited to ~31 Å, mainly due to the low image contrast and the featureless nature of the vault [10]. D48 symmetry was applied to this early cryoEM structure [10] and an X-ray crystal structure at 9 Å resolution [11]. Subsequently, another crystal structure of the rat native vault was solved with D39 symmetry at 3.5 Å resolution (PDB 4V60) [12]. Of the 861-amino-acid (aa) long MVP, PDB 4V60 contains full atom models for aa. 1-427, 449-607,621-813, and C_α-only model for the C-terminal segment (aa 814-845), with

a prominent gap from aa 428-448. From the N- to C-terminus, each MVP monomer consists of a body region containing 9 repeats (domains R1-R9) of an antiparallel β-sheet fold, followed by a shoulder region containing a single domain with 4 α-helices and a 4-stranded β-sheet, and a cap region containing a 155-amino-acid-long cap-helix domain and a cap-ring domain. In this crystal structure, the C_α-only model for the C-terminal segment is encapsulated inside the vault, rather than being exposed outside the vault [5]. Meanwhile, a crystal structure of a truncated MVP monomer (PDB 3GF5, which contains only the first, N-terminal 387 aa residues) was solved to 2.1 Å resolution [13]. These two crystal structures (PDB 4V60 for the vault and PDB 3GF5 for the N terminal segment) differ in the main chain tracing near the N-terminus (R1 and R2 domain). Further model refinement based on the electron density map of PDB 4V60 yielded a new model (PDB 4HL8) [14]. This refined, new model is basically a montage of PDB 3GF5 and 4V60: with its N-terminal domains (R1 and R2) similar to PDB 3GF5 and the following domains similar to those in PDB 4V60. Because PDB 3DF5 was obtained from a crystal containing segmented MVP, which lacked constraints from neighboring MVP subunits as those in the assembled vault, N-terminus domains in PDB 3GF5 are less curved than those in PDB 4HL8. These authors indicated that the backbone-tracing error near the N-terminus in the previous, PDB 4V60 model was a result of weak electron densities at the waist region and suggested that the N-terminal region is the major flexible region of MVP and may undergo large conformational change during assembly [14]. Such conformational changes might result in a leaky vault in solution and thus have bio-engineering significance, when vaults are engineered to package therapeutic compounds, such as hydrophobic all-*trans* retinoic acid [7]. Questions remain whether vaults in solution undergo conformational changes.

Here we have performed cryoEM and single particle analysis on a recombinant vault engineered at MVP N-terminus with a portion of HIV-1 Gag (amino acids 148-214) and obtained structures at near-atomic resolution (~4.8 Å). This highly conserved HIV-1 Gag segment has been shown to trigger immune response in peripheral blood mononuclear cells [15]. Our results clarify previous symmetry and structure discrepancies. Further analysis explains the vault's intrinsic structural flexibility and suggests optimization strategies to engineer MVP-only vaults for vaccine delivery applications.

3.3 Results

3.3.1 The vault has multiple conformations in solution

To enhance image contrast and to help clarify the number of subunits/vault, we recorded movies of recombinant rat vaults embedded in vitreous ice in a Titan Krios 300 kV electron microscope equipped with a K2 Summit direct electron-counting detector. Though vaults appeared mostly in their side views in our movies, occasionally top views of the vault can be spotted, showing features that indicate the separation of individual subunits (*i.e.*, one MVP monomer on the top half of the vault closer to the viewer and the other on the bottom half) of MVP subunits lining along the direction of the view (Figure 3.1A). The number of MVP pairs within one of the four quadrants of the top view (Figure 3.1B) is between 9 and 10, consistent with 39 MVP pairs (*i.e.*, 78 MVP subunits/vault, as in PDB 4HL8 and 4V60) [12, 14], but different from those used in other studies, such as PDB 2QZV (96 MVP subunits/vault) [11].

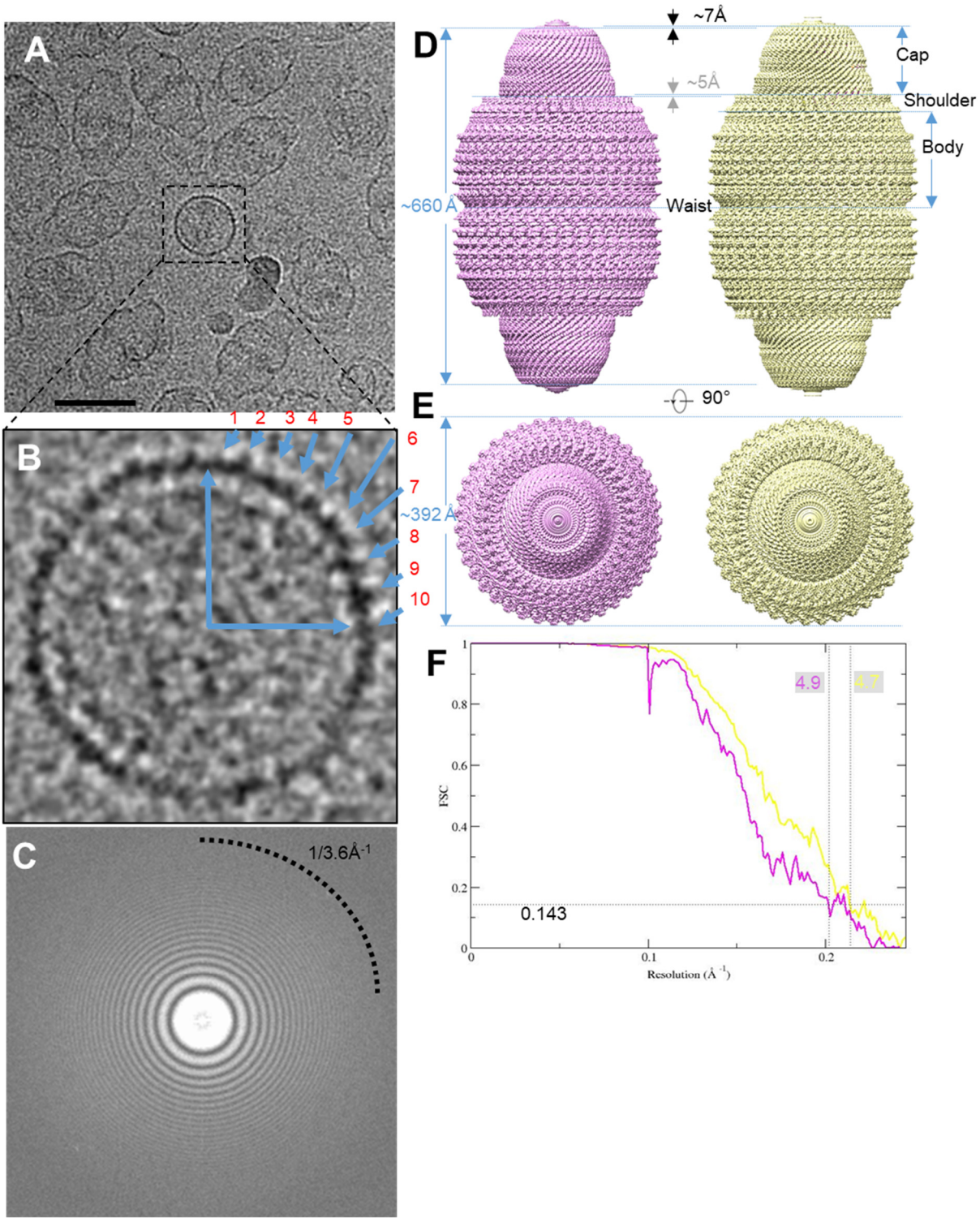


Figure 3.1 ***CryoEM single particle analysis result on engineered MVP-only vault.*** *(A) Aligned sum of rat vault raw image stack, showing this dataset has nice orientation distribution. Typical top views are boxed in black square. The scale bar is 50nm. (B) Magnified raw image of top view to show there are ~10 copies in a quadrant of circle, implying close to 40-fold related symmetry. (C) Fourier transform of a sum micrograph. Thon rings can reach to water signal at close to 3.6Å-1. (D) Density map of two vault conformations refined from a single dataset. Conformation 1 (displayed at 4.4 σ) is in pink and conformation 2 (displayed at 4.5 σ) is in yellow. They are all in D39 symmetry. (E) top view of vault density in (D). No diameter change can be observed. (F) FSC curve showing that the resolution (FSC 0.143) of the two conformations are 4.9 Å and 4.7 Å, respectively.*

The power spectrum of drift-corrected images shows that the Thon rings extend to 1/3.6 $Å^{-1}$ (Figure 3.1C and Figure 3.5), indicating that our images have structural information beyond 3.6 Å resolution. However, the best resolution we achieved after exhaustive attempts to carry out single-particle reconstruction by Frealign [16] was only 13.5 Å. This structure did not resolve

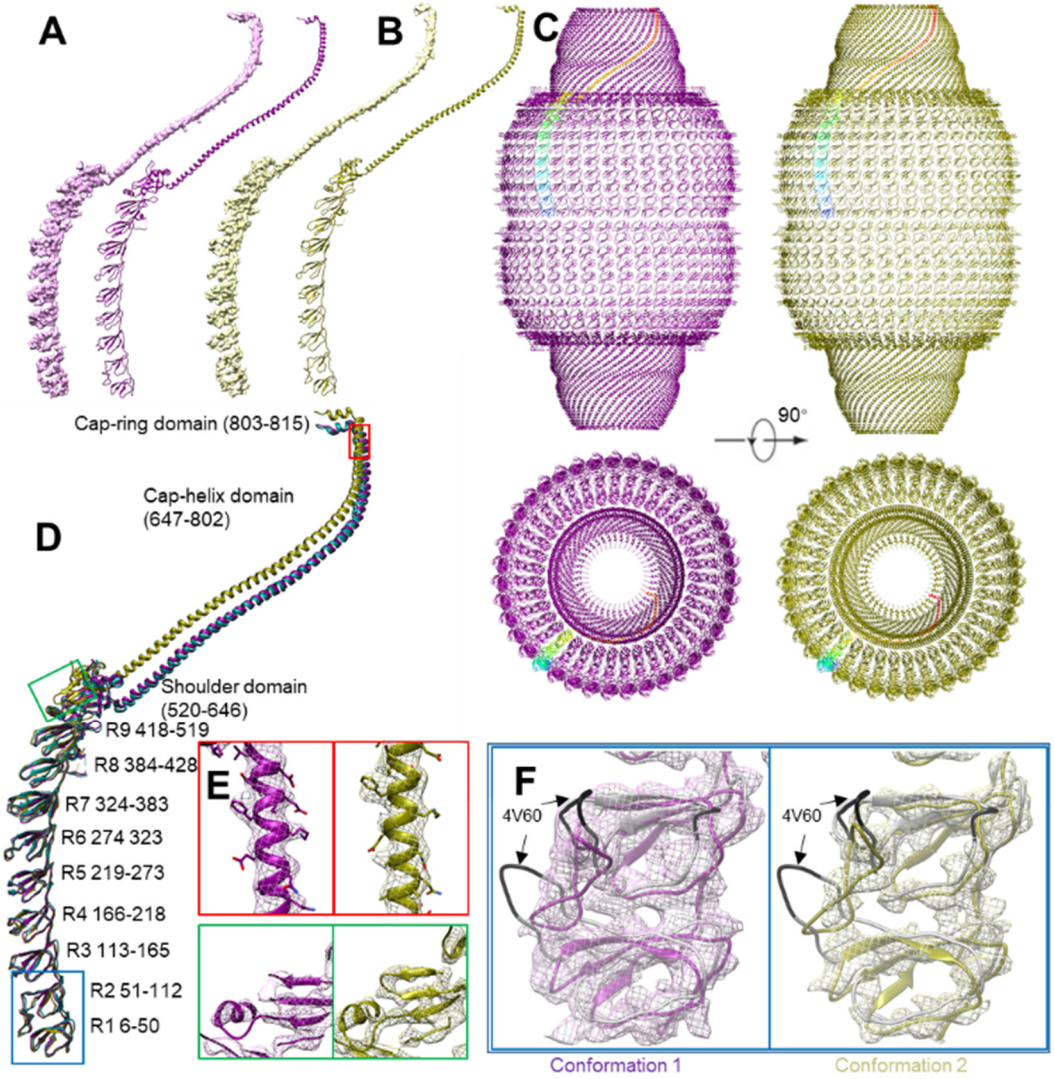


Figure 3.2 ***Structural comparison.*** *(A) Conformation 1 and corresponding model. (B) Conformation 2 and corresponding model. (C) Model comparison between conformation 1 (purple) and conformation 2 (olive). One copy of major vault protein (MVP) is colored in rainbow. The back half is hidden for clarity. (D) Overlapped model comparison. R1-R7 has no major conformational change. PDB 4HL8 is colored in aqua to show similarities between conformation 1 model and PDB 4HL8. (E) Near-atomic resolution feature at shoulder and cap-helix domain in both conformations, including an α-helix (red) and β-sheet (green). Large side chains can be identified and is consistent with current resolution estimation. Position is labeled in (D). Contour displayed at 6.7 σ. (F) Magnified view at R1 and R2 domain of two conformations. 4V60 model (grey) are displayed. The mismatch region in 4V60 is colored with black. No significant flexible region can be found at R1 and R2 domain. The major conformational change of cryoEM vault structure is not at waist region. Mesh contour is displayed at 5 σ for conformation 1 (pink) and 6.4 σ for conformation 2 (yellow)*

individual MVP subunits at the cap region, hence the handedness of the reconstruction could not be established. This observation suggests existence of multiple conformations in the sample. To sort out multiple conformations, we subsequently carried out three-dimensional (3D) classification with Relion [17] following the scheme illustrated in Figure 3.6 (see Method Details). Reconstructions with D38 or D40 symmetry did not converge to structures resolving any detailed features to establish handedness in the cap, even after exhaustive 3D classification and refinement. By contrast, reconstruction with D39 symmetry yielded two structure classes (*i.e.*, conformations) that both converged to near atomic resolution [4.9 Å for conformation 1 and 4.7 Å for conformation 2] (Figures 3.1D, E, and F). Both conformations reveal extensive secondary structures and some bulky side chains of amino acid residues (Figures 3.2A, B and E).

In these two conformations, the two halves of the barrel-shaped vault are joined at the waist. Extending away from the waist to the distal end of the vault are the body, the shoulder and the cap regions of each half (Figure 3.1D). Similar to previously reported vault X-ray structures [12, 14], 39 MVP subunits line in parallel to form half of the vault. The top center of cap region is closed with no discernable features in our D39-symmetry-imposed maps, indicating that the D39 symmetry is not maintained in this location (Figure 3.1E).

While conformation 1 and conformation 2 have the same waist radius, conformation 2 is 14 Å longer than conformation 1 along the D39 symmetry axis direction (Figure 3.1D). This observation suggests that the conformational changes are real and not an artifact of display differences or magnification variation in the microscope. This is direct evidence that multiple vault conformations exist in solution.

3.3.2 Engineered MVP-only vaults can adopt the structure of naturally-occurring vaults

The availability of the crystal structure of naturally-occurring vaults [14] (PDB 4HL8) allowed us to interpret our cryoEM structures of conformation 1 and conformation 2 at moderate resolutions of 4.9 and 4.7 Å, respectively. Consistent with crystal structure PDB 4HL8, the cryoEM densities of vault monomers in both conformation 1 and conformation 2 reveal the characteristic 9 repeats (R1 through R9) of β-sheet domains, followed by a shoulder domain with 4 α-helices and a 4-stranded β-sheet, and a long (~230 Å) cap helix (Figures 3.2A and B). PDB 4HL8 can be fitted only into conformation 1 density as a rigid body, indicating that conformation 1 is very close to crystal structure PDB 4HL8.

Therefore, we chose the PDB 4HL8 crystal structure as the starting model for real-space refinement against conformation 1 monomer. The refinement result shows that it is nearly identical to PDB 4HL8 with a root-mean-square deviation (RMSD) value of 1.3 Å. The traceable region covers most of the vault body and cap side walls, including R1-R9, shoulder, cap-helix and cap-ring domains (Figures 3.2A and D). The traceable sequence ends at P815 in the cap-ring domain. The rest of the density at the cap region is insufficiently resolved for reliable tracing of the C-terminal segment, from aa 816 to 861 (Figures 3.1D and 3.2C). Unlike naturally-occurring vaults, our recombinant vaults do not contain TEP1, VPARP or vRNA, yet the density at the center of the cap top is still solid, suggesting that the density observed at the center of the cap top in the two cryoEM conformations do not exclusively correspond to TEP1, VPARP, or vRNA as previously suggested [10, 12, 18].

At the waist region inside the vault is the Gag 148-214 peptide fused to N-terminus of MVP (Figures 3.4A and C). The sectional view shows that the fused Gag peptide is fully encapsulated

inside the vault, thus not exposed to the external environment (Figures 3.4A, C). The thickness of the MVP shell varies: close to 20 Å from R1 through R7 domains and thicker than 25 Å from R8, R9 and the shoulder domains (Figure 3.4E). R1 through R7 domains are also less structurally complex than R8, R9 and the shoulder domains. For example, in addition to R8 and R9's the antiparallel β-sheet fold sub-domain that resembles the antiparallel β-sheet fold in each of R1 through R7 domains, R8 and R9 domains both contain an inner attachment sub-domain: an α-helix in R8 (R8-helix) and a hairpin in R9 (R9-hairpin) atop R8-helix (Figure 3.3B). The shoulder domain has folding motif different from the antiparallel β-sheet fold of R1 through R9 and can be roughly divided into two sub-domains: a shoulder-helix sub-domain next to R9's antiparallel β-sheet fold and a shoulder-hybrid sub-domain (featuring a combination of α-helix and β-sheet) next to the inward-projecting R9-hairpin (Figure 3.3B). It is the presence of the inner-layer sub-domains in R8, R9 and shoulder domains that contributes to extra thickness of these domains as compared to R1 through R7 domains. The segment from N428 to S449 that connects R8 with R9 missing in X-ray crystal structure [12, 14] remained unresolved in conformation 1 cryoEM structure (Figure 3.3B).

3.3.3 Multiple conformations of the vault in solution

The local resolutions of most regions in the density map of conformation 2 are between 4 Å to 6 Å (Figure 3.4E), with the best resolution in the R1-R9 and the shoulder and the lowest resolution in the folded C-terminal region at the cap and the fused protein (Gag 148-214) region at the waist. The resolution of the entire inner surface of the vault is lower than that of the outer surface. The elongation of MVP monomer towards vault's pole in conformation 2 prevented satisfactory fitting of PDB 4HL8 as a rigid body into the density. R1 through R7 domains in PDB 4HL8 fit well with conformation 2 density, however R8, R9, shoulder, cap-helix and cap-ring domains do not. Thus,

the atomic model for R1 through R7 domains of PDB 4HL8 was fitted into the conformation 2 map. To address the previous discrepancies at the waist region, PDB 4V60 was also docked into the density of conformation 2 to see if this earlier X-ray model could represent conformation 2 of the vault in solution. This docking revealed that R1 and R2 domains in PDB 4V60 do not match

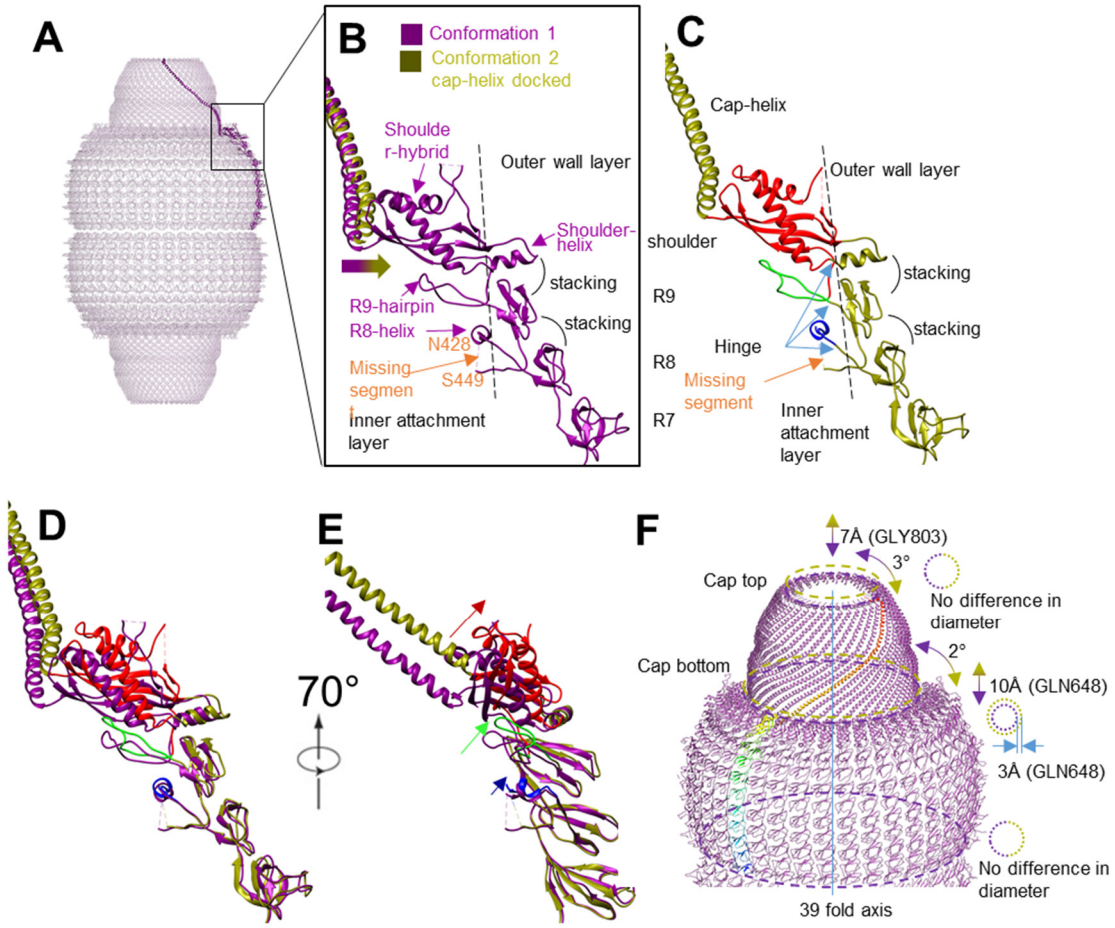


Figure 3.3 ***Conformational change diagram.*** ***(A)*** *Conformation 1 monomer model as a side view in vault over all model.* ***(B)*** *Magnified view of R7 to cap-helix domain. The missing segment between N428 to S449 locates inside vault. The docking of helix-cap domain of conformation 2 into conformation 1 density shows that cap-helix domain in conformation 2 bends outwards comparing with conformation 1. R8 to shoulder domain can all be roughly divided into two parts, separated by the dashed line. The attachment layer locates inside and the wall layer locates outside.* ***(C)*** *R7 to cap-helix domain of conformation 2. Like conformation 1, the R8 to shoulder domain is double layered. The inner-layer is colored from blue to red, from N-terminus to C-terminus. The outer layer is colored in original olive color.* ***(D, E)*** *Direct overlapping of conformation 1 and conformation 2 model. Position shift is magnified from R8 to shoulder domain. The relative movement of attachment layer is labeled with corresponding color and the movement is larger from R8 to shoulder domain.* ***(F)*** *A diagram to show the cap movement and conformational change between conformation 1 (purple) and conformation 2 (olive). In the refinement result applied D39 symmetry, the relative movement freedom of cap is limited to axial (up and down) and rotational (rotation around 39-fold axis). The movement from conformation 1 to conformation 2 of cap region can be described as "being rotated clockwise by 2 degrees and lifted by 10Å". There is minor morphing of cap between the two conformations, when conformation 2 is relatively shorter and more twisted at cap region based on shorter translocation distance along axis and more angular rotation at cap-ring region.*

the cryoEM density of conformation 2 (Figure 3.2F). Recall that conformation 1 of our cryoEM structures adopts the structure of PDB 4HL8, which differs from PDB 4V60 at R1 and R2 domains. Taken together, we conclude that PDB 4V60 does not represent vault's conformation in solution.

The atomic model of individual α-helixes and β-sheets in R8, R9, shoulder domain, cap-helix domain and cap-ring domain of PDB 4HL8 were fitted into their corresponding secondary structure elements visible in the cryoEM density map as rigid bodies. These fitted secondary structures were connected with linkers derived from the MVP sequence to create a full trial atomic model, which was subjected to model refinement against the cryoEM map. The resulting model (Figure 3.2B) has a C_α RMSD value of 5.95 Å when compared with PDB 4HL8 (see Table 3.2). The conformational changes in the refined model of conformation 2 take place at the R8 domain (P420) and become larger as one moves toward the C terminus (Figure 3.2D). R8 and R9 domains undergo minor conformational changes by slightly bending outwards, pivoting around their respective N-termini. Like that in conformation 1, we were unable to model the segment from N428 to S449 in conformation 2. This flexible segment likely occupies some space near the inner surface of R8 and R9 domains (Figure 3.3C, labeled "missing segment"). In general, while the secondary structure elements in R8 and R9 domains are conserved between conformation 1 and conformation 2, the relative positions of these secondary structure elements and their connecting linkers are not. Compared that in conformation 1, the shoulder domain in conformation 2 is twisted further outward, along with the attached cap-helix domain (Figure 3.2D and 3.3D).

Because of the linear arrangement of domains from N- to C-terminus, the large conformational change at the shoulder domain introduces a large translocation of the cap-helix and cap-ring domains (Figure 3.2D). In both conformations, α-helix and β-strands are resolved in the shoulder and cap-helix domains, allowing us to perform detailed comparisons of these high-

resolution features in the two conformations (Figure 3.2E). When viewed along the symmetry axis, the cap of conformation 2 is lifted by 10 Å and rotated by 2° clockwise from the bottom (Figure 3.3F and Movie 3.1, measurement based on C_α of GLN648) as compared to conformation 1. Further alignment shows that the cap top of conformation 2 was further rotated by 1° and shortened along the 39-fold axis by about 3 Å (measurement based on C_α of GLY803). The diameter of the cap top (*i.e.*, the distance from C_α of GLY803 the symmetry axis) is the same in both conformations, while the cap bottom diameter of conformation 2 is 6 Å (*i.e.*, 3%) larger than that of conformation 1. The 4.7 Å resolution of the cryoEM density lends support for these model-based measurements.

3.3.4 Position and structure of the engineered HIV-1 Gag 148-214 peptide inside the vault

Weak densities are observed in both conformations at the waist region close to the N-terminus of MVP. We interpret these densities as the fused HIV-1 Gag 148-214 peptide because of its connection to the N-terminus of MVP and its size matching that expected for the engineered Gag segment (Figures 3.4A, B). The density of the fused peptide is weaker than that of MVP (Figures 4C, D) and the boundary of fused protein at the waist region is at lower resolution. A dimer of Gag 148-214 (PDB 1AFV) [19] can be docked into the donut shaped density seen inside the waist region. However, the docking was non-unique and no secondary structure can be identified in the waist-ring density, suggesting that the fusion peptide is flexible inside the vault with only limited interactions with MVP. This result is consistent with previous observations that peptides fused to N-terminus of MVP tend to extend towards the center of a vault particle [4].

The total number of amino acids for the fused Gag 148-214 and the GFLGL linker is 73. Applying the free chain model assuming a persistence length of 5 amino acids, the engineered peptide would give rise to a maximal end-to-end length of 73 Å ($i.e., 3.8\text{Å} \times 5 \times \sqrt{\frac{73}{5}}$). This maximal length of the engineered segment is much shorter than the 140 Å axial linear distance

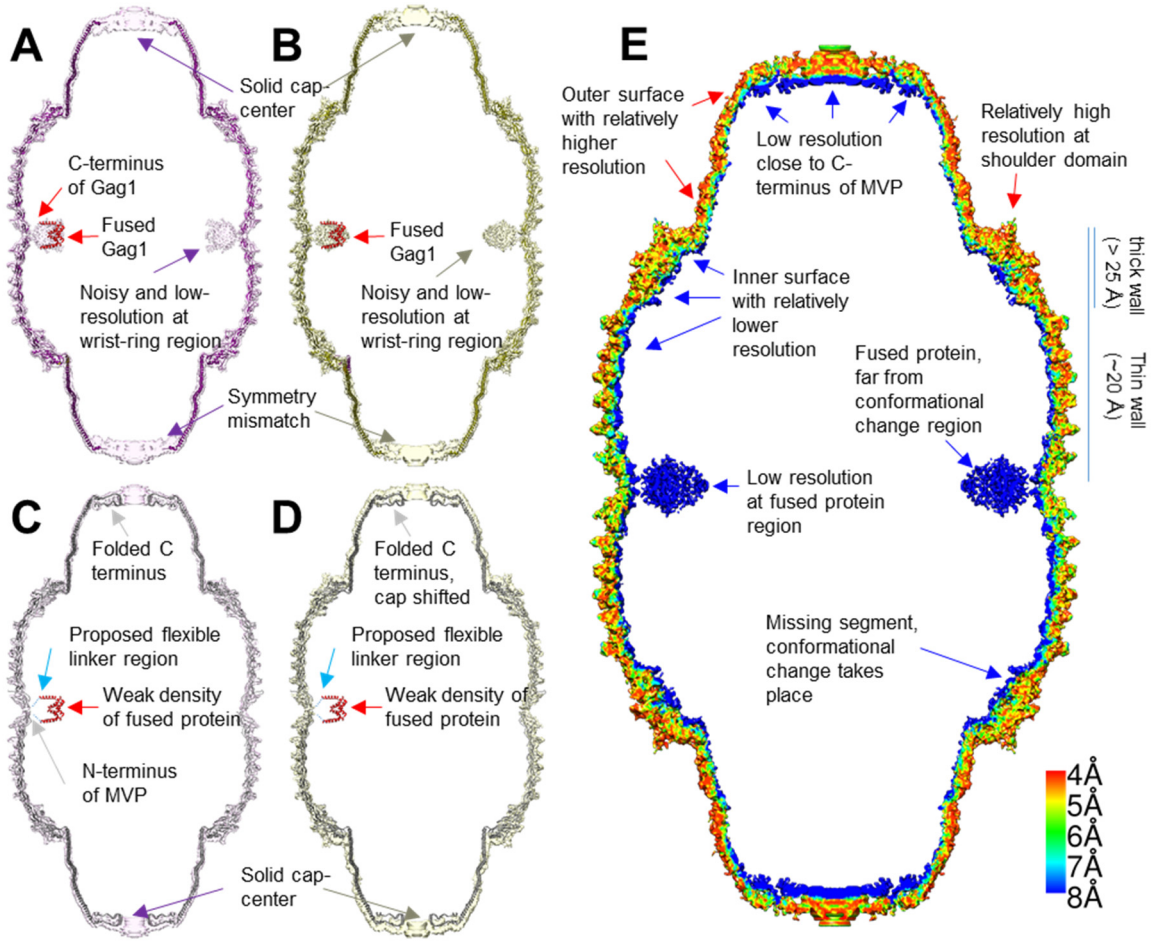


Figure 3.4 ***Model and density comparison among models and densities via longitudinal section.*** ***(A)*** *Conformation 1 model (purple), conformation 1 density (pink, displayed at 3 σ) and docked segmented Gag dimer (PDB 1AFV, red).* ***(B)*** *Conformation 2 model (olive), conformation 2 density (yellow, displayed at 3.7 σ) and docked Gag dimer (red).* ***(C, D)*** *utilizing similar color code with (A) and (B), respectively, with higher visualization threshold (displayed 6.2 σ and 5.1 σ for conformation 1 and 2, respectively). PDB 4V60 (grey) is docked into conformation 1 and conformation 2 density. C-terminus of Gag and N-terminus of MVP is connected by flexible linker, shown as dashed line.* ***(E)*** *Local resolution estimation of conformation 2 density map calculated by Resmap*[26]*. The flexible region is of low resolution and appear blue (fused protein at waist, C terminus near cap and inner surface). The major conformational change takes place at shoulder domain but the resolution is relatively high.*

between the R8 domain and the waist. Therefore, the observed conformational change starting at the R8 domain is unlikely caused by interactions with the fused Gag 148-214 peptide.

3.4 Discussion

Our near-atomic cryoEM structures demonstrate the co-existence of two conformations of the vault in solution. The pole regions of the vault contain solid densities in both conformations. Previous studies proposed that MVP C-terminus (P815-K861) folded back inside the vault, leaving an opening at both poles of the vault [12]. Our MVP-only vault structure suggests that at least some of the C-terminal segment must extend all the way towards the symmetry axis to seal the vault particle, thus different from that as described before (Figures 3.4C, D). This result is consistent with a previous low-resolution cryoEM study showing that an antibody could bind to its antigen fused at MVP's C-terminus [5]. The relative lower resolution of this region of our vault reconstruction suggests that the symmetry at the pole region might deviate from D39.

The comparison between the two conformations suggests varying levels of structural flexibility for different domains of the vault. Targeting flexible region of the vault for engineering applications is desirable in order to reduce possible steric hindrance introduced by peptide insertion. Domains R1 through R7 on MVP contains only antiparallel β-sheet motifs and exhibit no structural changes between the two conformations. Four hydrogen bonds and eight hydrophobic interactions [12] exist between neighboring MVP monomers at the R1-R7 domains, which are believed to play a key role in assembling 78 MVP monomers to an intact vault [14]. The major difference between the outer shell of our recombinant vault and the naturally-occurring vault is the 73 aa Gag and linker peptides fused to the N-terminus of MVP. The folded peptide appears to only directly interact with the R1 domain through a flexible linker. The structures of R1-R7 domains in our cryoEM structures of both conformations in solution are identical to those in the crystal structure

PDB 4HL8, indicating that the fused peptide does not change the conformation of these domains near the site of the engineered peptide and that the cryoEM structures of our recombinant vault can inform about structural dynamics of naturally-occurring vaults in solution.

It is interesting to note that conformation 1 is similar to PDB 4HL8, which was determined from crystals where crystal packing might have constrained the flexible elements of the vault [12]. The cryoEM structures reported here were determined from vaults distributed on carbon film to overcome the preferred orientation problem. Whether this support film influenced the observed structural variability is hard to establish as many cryoEM structures have been determined to near-atomic resolution using similar strategies. R1 through R7 act like rigid plates laced together akin to the arrangement of ancient lamellar armor. Each domain contains antiparallel β-strands and corresponds to a solid plate of the armor. The 12 interactions between these domains [12] are similar to the lace that holds lamellae together. We suggest a "wall-and-attachment" mechanism for the structural dynamics of the vault, whereby the laced R1-R7 plates, the identical antiparallel-β-sheet motif sub-domains in R8-R9, and the shoulder-helix sub-domain together form a rigid *wall* to which flexible regions (R8-helix, R9-hairpin and the shoulder-hybrid sub-domains) *attach*. These flexible regions constitute an "attachment" layer (Figure 3.3B) that undergo conformational changes while holding to a relatively rigid wall (olive in Figure 3.3C). Interestingly, the conformational changes increase from R8 domain to the shoulder domain (Figure 3.3D) and all these domains are more structurally complex than R1 through R7 domains. The protein sequences for the attachment layer are conserved between vaults isolated from different species [20], suggesting its observed structural dynamics is likely an intrinsic property of vaults. Although the vault is closed in both conformations presented here, our observed dynamics may be the structural basis for vault breathing and opening as suggested previously [14].

Engineered vaults have been proposed as bio-compatible vehicles for vaccine delivery [6]. In both conformations, the largest opening on the side wall of the vault is estimated to be of 5 Å by 5 Å. Previous study [21] shows that vault associate protein (TEP1) can enter assembled MVP-only vault, but truncated TEP1 cannot. Our cryoEM observation is consistent with this specific-access model. This feature reduces the chance of cargo contamination and leakage. Our results can also inform how to engineer vaults for carrying peptides and small proteins for therapeutic applications. It is known that the high solubility of vaults under physiological conditions enables packaging of highly potent but often insoluble drugs to improve efficacy and reduce toxicity. Our results show that MVP alone can form a completely sealed enclosure that would protect packaged materials and that the fused Gag 148-214 peptide is located adjacent to the vault waist with size and shape expected for properly folded Gag. This is consistent with previous observation that GFP fused at N-terminal of MVP functioned properly [22]. Therefore, N-terminal fusion is a viable strategy for vaccine delivery though a more complete understanding of the rules governing the position and structure of fused target antigens is needed. To the contrary, antigens fused at the MVP C-terminus might not work because they may fold outwards and become exposed.

One remaining caveat of the N-terminal fusion strategy is potential crowdedness that might interfere with dimerization of the two halves of the vault during assembly [2]. The strong interactions between neighboring MVP monomers in their regions spanning R1-R7 domains leave little room to insert peptide between those repeats. Our structure results reveal other locations inside the vault that can be targeted for peptide fusion to overcome this caveat. For example, the missing segment (N428 to S449) is located inside the vault and is flexible. Fusion at this location would divide the payload into two parts of 39 copies each to be packaged inside the top and bottom

halves away from the waist. Proteins fused near flexible region (R8, R9 and shoulder domain) can

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, Peptides, and Recombinant Proteins		
Recombinant protein: HIV Gag1-M1-GFLGL (MTPRTLNAWVKVVEEKAFSPEVIPMFTALSEGATPSDLNTM LNTIGGHQAAMQMLKDTINEEAAEWDRGFLGL)	Genbank	MH020171
Gibco® Sf-900™ II SFM	Gibco	10902096
Critical Commercial Assays		
In-Fusion® HD Cloning Kit	Takara Bio	638910
Bac-to-Bac® Baculovirus Expression System	Invitrogen	10359-016
Deposited Data		
Vault conformation 1 model	This paper	PDB-6BP8
Vault conformation 2 model	This paper	PDB-6BP7
Vault conformation 1 cryoEM map	This paper	EMDB-7126
Vault conformation 2 cryoEM map	This paper	EMDB-7125
Experimental Models: Cell Lines		
Spodoptera frugiperda: sf9	Gibco	Cat# 11496015
Experimental Models: Organisms/Strains		
MAX Efficiency® DH10Bac™ Competent Cells	Invitrogen	Cat# 10361012
Recombinant DNA		
Plasmid: pFastBac 1	Invitrogen	Cat# 10359016
Plasmid: pFastBac 1-Gag1-M1-GFLGL-MVP	Genbank	MH020171
Software and Algorithms		
Relion	[17]	http://www2.mrc-lmb.cam.ac.uk/relion/index.php/Main_Page
Coot	[23]	http://www2.mrc-lmb.cam.ac.uk/Personal/pemsley/coot/
PHENIX	[24]	http://www.phenix-online.org/
UCSF Chimera	[25]	https://www.cgl.ucsf.edu/chimera/
Resmap	[26]	http://resmap.sourceforge.net/

Table 3.1 Key Resources Table

also potentially reduce impact on the assembly of MVP.

3.5 Methods

3.5.1 Key Resources Table

3.5.2 Contact for Reagent and Resource Sharing

Further information and requests for reagents may be directed to, and will be fulfilled by the Lead Author, Dr. Z. Hong Zhou (Hong.Zhou@UCLA.edu).

3.5.3 Method Details

3.5.3.1 Recombinant vault sample preparation

The HIV Gag148-214-GFLGL fragment was PCR amplified and cloned into pFastBac 1 containing rat MVP. The NcoI cloning site at the MVP 5' end was used as a site of insertion by employing In-Fusion HD cloning kit (cat # 638910) and by strictly following the In-Fusion® HD Cloning kit manual from Takara/Clontech Inc. The resulting pFastBac1-Gag1-M1-GFLGL-rMVP construct was recombined with Bacmid DNA in MAX Efficiency® DH10Bac™ Competent Cells from Invitrogen (cat # 10361012), see manufacturers protocol. The Bacmid containing the Gag 148-214-GFLGL-MVP recombinant DNA was purified following Bac-to-Bac® Baculovirus Expression System manual from Invitrogen (Cat# 10359-016). Generation of Baculovirus expressing the recombinant MVP was accomplished by transfecting sf9 cells in Gibco® Sf-900™ II SFM cells (Cat# 11496015) with Bacmid DNA, using Cellfectin™ II Reagent (Cat# 10362100) and following the user guide. To produce the recombinant vaults, 1x 10^8 sf9 cells were infected with the recombinant Baculovirus in 50 ml Sf-900™ II SFM (Cat# 10902096). The infected cells were shaken for 3 to 4 days at 28 °C, then harvested by centrifuging at 500 x g for 5 min at room temperature. Cell pellet was stored at -80°C or used directly for vault purification.

For Sf9 cell lysis, buffer A (50 mM Tris-Cl buffer, 75 mM NaCl, 0.5 mM MgCl2) containing 2% Triton-X-100, 2% PI (Protease Inhibitor; Sigma-Aldrich P8849-5ML) and 1 mM PMSF (Phenylmethylsulfonyl fluoride) was prepared. 1 mg of RNase A was added to 1 g of Sf9 cells

expressing Gag1-M1-GFLGL-rMVP, then 5 ml of lysis buffer was added and incubated on ice for 15 min. 2 mM DTT was then added and the cell lysate was further incubated on ice for additional 5 min. The lysate was centrifuged at 20,000 x g at 4 °C for 20 min. The supernatant was collected and centrifuged at 40K in Ti 70.1 rotor for 1 h at 4 °C. The pellet was resuspended in 1 ml buffer A supplemented with 2% PI, 2 mM PMSF, and 2 mM DTT. 1 ml Ficoll-sucrose was added and the mixture was further vortexed and centrifuged at 25K in Ti 70.1 rotor at 4 °C for 10 minutes. The supernatant was diluted with 5.5 ml buffer A supplemented with 1% PI, 1 mM PMSF and 1 mM DTT, which was then centrifuged at 40 K in Ti 70.1 rotor for 1h, 30 minutes at 4°C. The pellet was resuspended in 1 ml buffer A containing 1% PI, 1 mM PMSF and 1 mM DTT. 5 µg streptomycin sulfate was added; the mixture was tumbled at room temperature for 30 minutes, then centrifuged 16,100 x g at room temperature for 10 minutes. Clarified supernatant was overlaid on a stepwise sucrose gradient (20%, 30%, 40%, 45%, 50%, 60% sucrose, 1.5 ml each) and then centrifuged at 25K in sw41 rotor at 4 °C for 16 hours. The 40% and 45% sucrose fractions were collected. The fractions were diluted in 4.5 ml PBS, then centrifuged at 40K in Ti 70.1 rotor at 4°C for 2 hours. The pellet was resuspended in 210 µl PBS to serve as cryoEM grid ready sample.

3.5.3.2 Electron microscopy and movie processing

For cryoEM, an aliquot of 2.5 µl of recombinant vault sample was applied to each EM grid with Lacey carbon films. The grid was blotted with Vitrobot in 100% humidity for 10s and then plunged into liquid ethane to vitrify the sample. Movies were obtained in Titan Krios 300kV electron microscope with Gatan K2 direct election detection camera in super-resolution mode with Leginon [27] at ×49000. The pixel size was measured to be 1.036 Å on the specimen scale. We used an electron dose rate of 8 electrons/pixel/second and each movie contains 20 frames recorded in 5 seconds. Image stacks in each movie were aligned with UCSF MotionCorr [28]. The first 16

frames in each stack were averaged to obtain an image sum of 32 $e^-/Å^2$. The whole dataset has 1218 movies.

3.5.3.3 Data processing and 3D reconstruction

Micrographs after alignment were used for contrast transfer function (CTF) determination in CTFFIND3 [27], with defocus values ranging from -1.7 μm to -4.2 μm. A total of 63751 particles were manually picked with 900 × 900 box size in pixel. Particles were first directly refined with Frealign [16] and reported resolution was 13.5 Å with little features showing handedness. Then, all particles were subjected RELION 1.2 [17] for two-dimensional classifications (Class2D). Top views were intentionally excluded from further classification to limit sampling space and to accelerate refinement process. Also, classes with no interpretable features were discarded. 32702 particles were selected for further three-dimensional classifications (Class3D). Particles are classified based on D38, D39 and D40 symmetry in different runs. Classification result with D38 and D40 symmetry also showed little feature with handedness. The following class3D are all conducted applying D39 symmetry with finer searching grid (Figure 3.6). The initial model for Class3D was generated from previously published atomic model (PDB 4HL8) of rat vault to 50 Å resolution to eliminate the potential risk of model bias. Class3D analysis was conducted with D39 fold symmetry applied and 2 distinguish classes with relatively good resolution (~9 Å) were found. These two classes were further refined separately with RELION 1.2 with D39 symmetry. To further enhance signal, mask is generated from cryoEM data to focus the refinement on MVP region. Following 'gold standard' refinement protocol, the two conformations were refined both to near-atomic resolution after RELION post-processing and automatic soft masking [17]. The resolution was determined based on a 'gold standard' Fourier shell correlation (FSC) coefficient of 0.143.

3.5.3.4 Atomic model building, refinement, and visualization

The atomic model of engineered vault was derived from crystal structure PDB 4HL8. By calibrating pixel size from 1.036 Å to 1.000 Å, we achieved an optimal docking of PDB 4HL8 into conformation 1 density. The fitted PDB 4HL8 was subjected to real-space refinement in Phenix [24] using the MVP monomer as density map input. Ramachandran and rotamer outliers were manually corrected with Coot [23] for this conformation 1 model.

The pixel size of conformation 2 density was also adjusted to 1.000 Å accordingly. R1-R7 domains in PDB 4HL8 was first fitted into conformation 2 density. In R8 to cap-helix domains of PDB 4HL8, individual secondary structures were fitted into corresponding densities in conformation 2 map. Those secondary structures were further connected with linker accordingly to create a "morphed" model. Following the same protocol as refining model of conformation 1, this "morphed" PDB 4HL8 was subjected to real-space refinement with segmented density of conformation 2 in Phenix. Ramachandran and rotamer outliers were also manually corrected with Coot [23] for conformation 2.

Visualization and map segment were achieved with UCSF Chimera [25]. Local resolution was calculated by Resmap [26].

3.6 Supplemental information

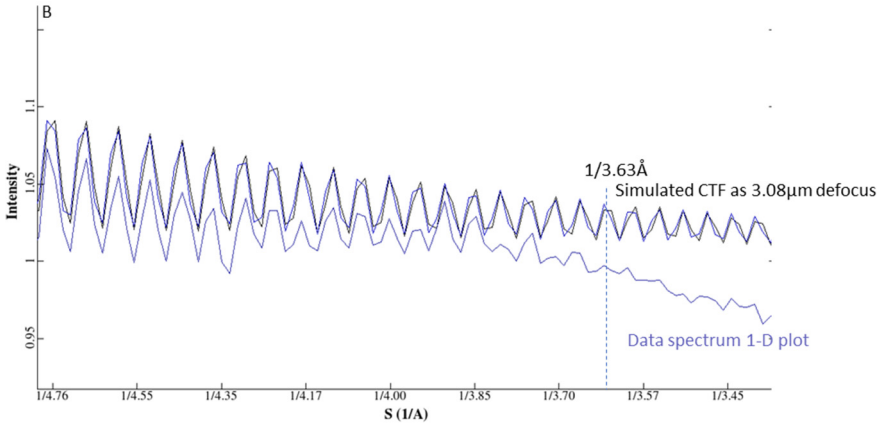


Figure 3.5 1-D plot of a raw micrograph shows that signal is transferred to atomic resolution.

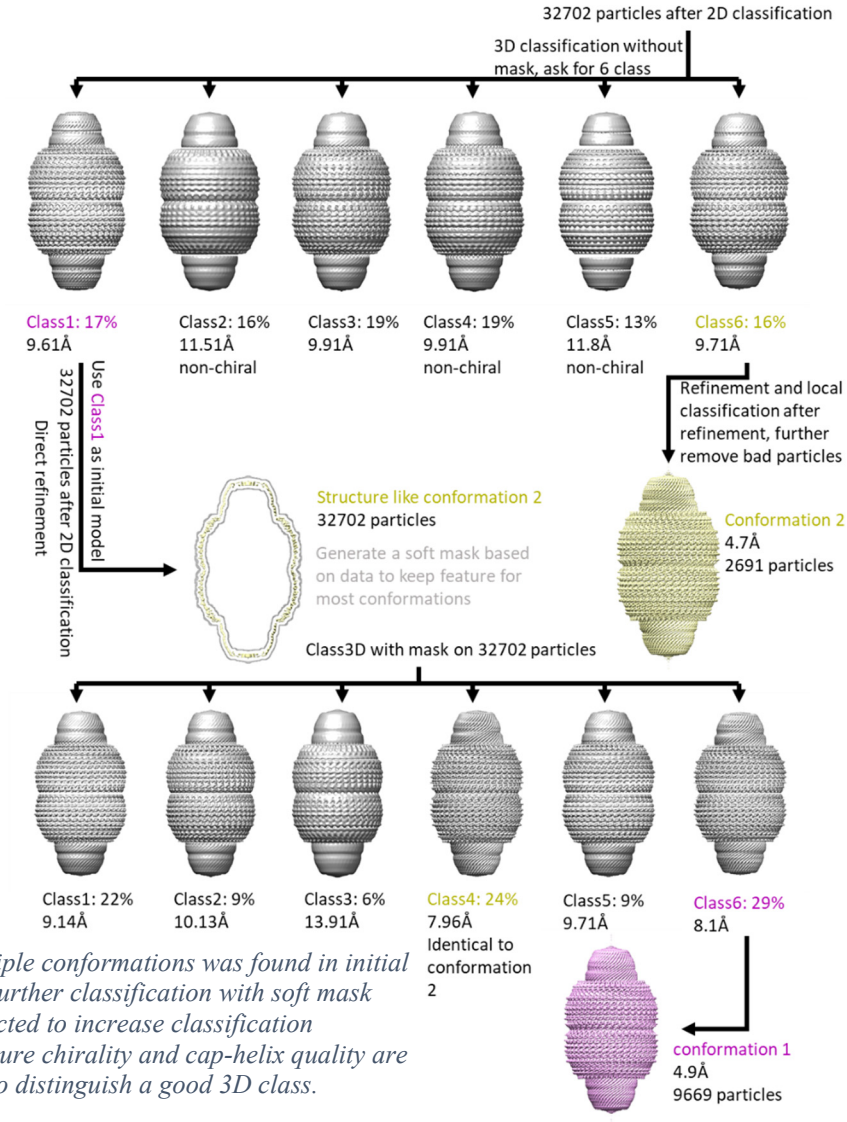


Figure 3.6 Multiple conformations was found in initial classification. Further classification with soft mask was later conducted to increase classification accuracy. Structure chirality and cap-helix quality are major features to distinguish a good 3D class.

Conformation ID		# 1	# 2
B-factor for map (Å^2)		-160	-225.5
MapCC (around atoms)		0.763	0.748
Phenix RMSD	Bond (Å)	0.0026	0.0032
	Angles	0.66	0.77
Ramachandran plot (from Phenix)	Outliers	0.77%	0.78%
	Allowed	4.12%	4.18%
	Favored	95.10%	95.04%
All atom clash score		10.94	12.90
Cα RMSD Value to PDB 4HL8		1.3 Å	5.95 Å
Rotamer outliers		0.00%	0.00%
C-beta deviation		0	0

Table 3.2 Structural statistics of the two conformers.

3.7 Author Contributions

Z.H.Z. and L.H.R. designed and supervised research; O.O.Y. and J.M. designed the vaults; J.M., H.L.N. and V.A.K. prepared samples; K.D., X.Z. and Z.H.Z. recorded and analyzed the cryoEM data; K.D. and M.L. built atomic models; Z.H.Z. and K.D. wrote the paper; all authors reviewed and finalized the paper.

3.8 Acknowledgements

This work was supported in part by grants from the National Institutes of Health (GM071940, AI094386, AI043203, AI106528, and DE025567). We acknowledge the use of instruments at the Electron Imaging Center for Nanomachines supported by UCLA and by grants from NIH (1S10RR23057, 1S10OD018111, U24GM116792) and NSF (DBI-338135). K.D. was supported by UCLA Molecular Biology Whitcome Pre-Doctoral Fellowship.

3.9 Data availability

CryoEM density maps have been deposited in the Electron Microscopy Data Bank under the accession numbers EMD-7126 (conformation 1), EMD-7125 (conformation 2) and their corresponding coordinates of atomic models have been deposited in the Protein Data Bank under the accession number 6BP8 and 6BP7, respectively. The data that support the findings of this study are available from the corresponding author upon request.

3.10 Declaration of Interests

L.H.R., O.O.Y., and V.A.K. declare that they have a financial interest in Vault Nano Inc. and that the Regents of the University of California have licensed intellectual property invented by L.H.R. and V.A.K. to Vault Nano Inc. V.A.K. is currently working for Vault Nano Inc. J.M. declares he has financial interest in Aukera, Inc.

3.11 References

1. Kedersha, N.L. and L.H. Rome, *Isolation and Characterization of a Novel Ribonucleoprotein Particle - Large Structures Contain a Single Species of Small Rna.* Journal of Cell Biology, 1986. **103**(3): p. 699-709.

2. Mrazek, J., et al., *Polyribosomes Are Molecular 3D Nanoprinters That Orchestrate the Assembly of Vault Particles.* Acs Nano, 2014. **8**(11): p. 11552-11559.

3. Stephen, A.G., et al., *Assembly of vault-like particles in insect cells expressing only the major vault protein.* Journal of Biological Chemistry, 2001. **276**(26): p. 23217-23220.

4. Mikyas, Y., et al., *Cryoelectron microscopy imaging of recombinant and tissue derived vaults: Localization of the MVP N termini and VPARP.* Journal of Molecular Biology, 2004. **344**(1): p. 91-105.

5. Kickhoefer, V.A., et al., *Targeting Vault Nanoparticles to Specific Cell Surface Receptors.* Acs Nano, 2009. **3**(1): p. 27-36.

6. Champion, C.I., et al., *A Vault Nanoparticle Vaccine Induces Protective Mucosal Immunity.* Plos One, 2009. **4**(4).

7. Buehler, D.C., et al., *Vaults Engineered for Hydrophobic Drug Delivery.* Small, 2011. **7**(10): p. 1432-1439.

8. Qi, X.Y., et al., *Vault Protein-Templated Assemblies of Nanoparticles.* Nano, 2012. **7**(1).

9. Berger, W., et al., *Vaults and the major vault protein: Novel roles in signal pathway regulation and immunity.* Cellular and Molecular Life Sciences, 2009. **66**(1): p. 43-61.

10. Kong, L.B., et al., *Structure of the vault, a ubiquitous cellular component.* Structure with Folding & Design, 1999. **7**(4): p. 371-379.

11. Anderson, D.H., et al., *Draft crystal structure of the vault shell at 9-angstrom resolution.* Plos Biology, 2007. **5**(11): p. 2661-2670.

12. Tanaka, H., et al., *The Structure of Rat Liver Vault at 3.5 Angstrom Resolution.* Science, 2009. **323**(5912): p. 384-388.

13. Querol-Audi, J., et al., *The mechanism of vault opening from the high resolution structure of the N-terminal repeats of MVP.* Embo Journal, 2009. **28**(21): p. 3450-3457.

14. Casanas, A., et al., *New features of vault architecture and dynamics revealed by novel refinement using the deformable elastic network approach.* Acta Crystallographica Section D-Biological Crystallography, 2013. **69**: p. 1054-1061.

15. Yang, O.O., et al., *Short Conserved Sequences of HIV-1 Are Highly Immunogenic and Shift Immunodominance.* Journal of Virology, 2015. **89**(2): p. 1195-1204.

16. Lyumkis, D., et al., *Likelihood-based classification of cryo-EM images using FREALIGN.* Journal of Structural Biology, 2013. **183**(3): p. 377-388.

17. Scheres, S.H.W., *RELION: Implementation of a Bayesian approach to cryo-EM structure determination.* Journal of Structural Biology, 2012. **180**(3): p. 519-530.

18. Tanaka, H. and T. Tsukihara, *Structural studies of large nucleoprotein particles, vaults.* Proceedings of the Japan Academy Series B-Physical and Biological Sciences, 2012. **88**(8): p. 416-433.

19. Momany, C., et al., *Crystal structure of dimeric HIV-1 capsid protein.* Nature Structural Biology, 1996. **3**(9): p. 763-770.

20. Kedersha, N.L., et al., *Vaults .2. Ribonucleoprotein Structures Are Highly Conserved among Higher and Lower Eukaryotes.* Journal of Cell Biology, 1990. **110**(4): p. 895-901.

21. Poderycki, M.J., et al., *The vault exterior shell is a dynamic structure that allows incorporation of vault-associated proteins into its interior.* Biochemistry, 2006. **45**(39): p. 12184-12193.

22. Slesina, M., et al., *Movement of vault particles visualized by GFP-tagged major vault protein.* Cell and Tissue Research, 2006. **324**(3): p. 403-410.

23. Emsley, P., et al., *Features and development of Coot.* Acta Crystallographica Section D-Biological Crystallography, 2010. **66**: p. 486-501.

24. Adams, P.D., et al., *PHENIX: a comprehensive Python-based system for macromolecular structure solution.* Acta Crystallographica Section D-Biological Crystallography, 2010. **66**: p. 213-221.

25. Pettersen, E.F., et al., *UCSF chimera - A visualization system for exploratory research and analysis.* Journal of Computational Chemistry, 2004. **25**(13): p. 1605-1612.

26. Kucukelbir, A., F.J. Sigworth, and H.D. Tagare, *Quantifying the local resolution of cryo-EMEM density maps.* Nature Methods, 2014. **11**(1): p. 63-+.

27. Mindell, J.A. and N. Grigorieff, *Accurate determination of local defocus and specimen tilt in electron microscopy.* Journal of Structural Biology, 2003. **142**(3): p. 334-347.

28. Li, X.M., et al., *Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM.* Nature Methods, 2013. **10**(6): p. 584-+.

Chapter 4 *In situ* structures of the segmented genome and RNA polymerase complex inside a dsRNA virus

Xing Zhang[1,*], Ke Ding[2,3,*], Xuekui Yu[2,*], Winston Chang[1], Jingchen Sun[2,4,‡], Z. Hong Zhou[1,2,3,‡]

[1] California Nanosystems Institute, [2] Department of Microbiology, Immunology and Molecular Genetics, and [3] Bioengineering, University of California, Los Angeles, CA 90095, USA

[4] Subtropical Sericulture and Mulberry Resources Protection and Safety Engineering Research Center, Guangdong Provincial Key Laboratory of Agro-animal Genomics and Molecular Breeding, College of Animal Science, South China Agricultural University, Guangzhou, Guangdong 510642, China.

*These authors contributed equally.

‡Correspondence and requests for materials should be addressed to Z.H.Z. (email: Hong.Zhou@ucla.edu) and J.S. (cyfz@scau.edu.cn), respectively.

4.1 Abstract

Viruses in the *Reoviridae*, like the triple-shelled human rotavirus and the single-shelled insect cytoplasmic polyhedrosis virus (CPV), all package a genome of segmented dsRNAs inside the viral capsid and carry out endogenous mRNA synthesis through a transcriptional enzyme complex (TEC). By direct electron-counting cryoEM and asymmetric reconstruction, we have determined the organization of the dsRNA genome inside quiescent CPV (q-CPV) and the *in situ* atomic structures of TEC within CPV in both quiescent and transcribing (t-CPV) states. We show that the total 10 segmented dsRNAs in CPV are organized with 10 TECs in a specific, non-symmetric manner, with each dsRNA segment attached directly to a TEC. TEC consists of two extensively-interacting subunits: an RNA-dependent RNA polymerase (RdRP) and an NTPase VP4. We find that the bracelet domain of RdRP undergoes significant conformational change when converted from q-CPV to t-CPV, leading to formation of the RNA template entry channel and access to the polymerase active site. An N-terminal helix from each of two subunits of the capsid shell protein (CSP) interacts with VP4 and RdRP. These findings establish the link between sensing of environmental cues by the external proteins and activation of endogenous RNA transcription by the TEC inside the virus.

4.2 Introduction

Each capsid of viruses in the *Reoviridae* contains 9-12 segmented dsRNAs and up to 12 transcriptional enzyme complexes (TECs). These RNA-containing viruses are fully capable of RNA transcribing and capping[1]. Crystal structures of the RNA-dependent RNA polymerase (RdRP) component of the TEC have been determined for rotavirus and mammalian reovirus (MRV)[2, 3], but no high-resolution *in situ* structure of the TEC is available. Moreover, the organization of TECs with the dsRNA genome and the mechanism of transcriptional activation have remained mysteries, in contrast to the well understood genome organization inside dsDNA viruses[4, 5].

4.3 Results and discussion

With only a single protein shell that encloses 10 different genome segments, CPV is one of the simplest dsRNA viruses[6] and serves as a model system, as highlighted by its contribution to the discovery of RNA capping[7]. To gain insight into the organization of the TEC and segmented dsRNA genome, we have determined CPV structure in a quiescent (q-CPV) state at 5.1Å resolution (see Methods, Figures 4.5 and 4.6). The structure reveals that each CPV contains 10 TECs under 10 specific positions of the 12 icosahedral vertices (Figure 4.1). The two vertices without TECs are occupied by rod-like densities (Figure 4.1a-e, Movie 4.1, Figures 4.7 amd 4.8). The previously ambiguous locations of TECs[8, 9] are now determined to be 10 specific positions in each CPV particle, related by pseudo-D3 symmetry, with only one on a “south tropic” position and three each around the “north tropic”, “north pole” and “south pole” positions (Figure 4.1d, Movie 4.2).

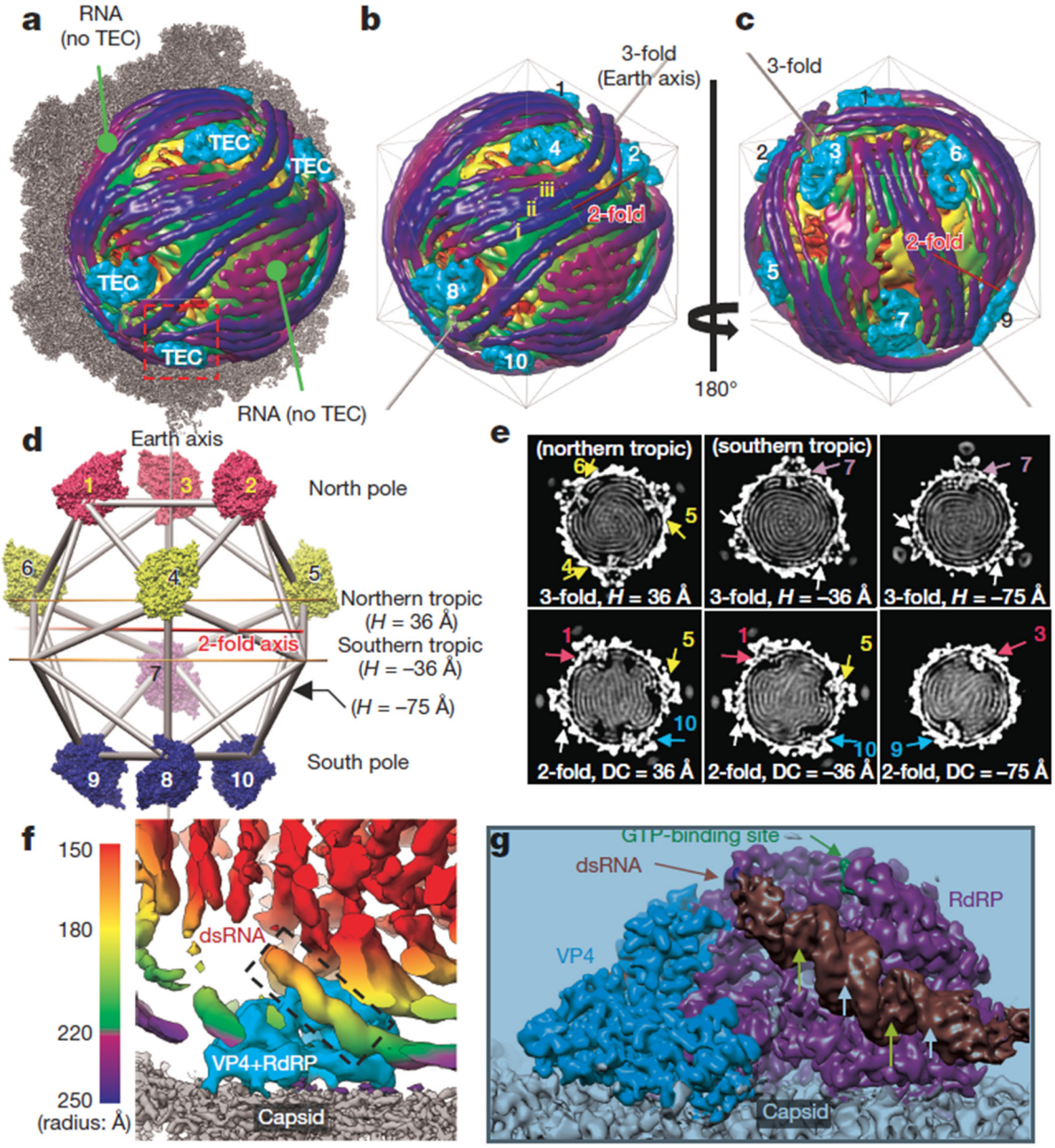


Figure 4.1 ***Transcription enzyme complex (TEC) and dsRNA genome organization inside CPV. a****, Superposition of the high-resolution (3.9 Å) map of half a capsid (grey) and low-resolution (22 Å) map of dsRNA genome (radially colored as in f) and TECs (cyan).* ***b-c****, Front (b) and back (c) views of the dsRNA genome and TECs of (a).* ***d****, Earth-like representation, illustrating the locations of the ten TECs (surface-rendered) with pseudo-D3 symmetry: three on each pole and the northern tropic but only one on the southern tropic.* ***e****, Cross sections of the 22 Å-density map, perpendicular to either the "earth axis" in (d) (top row) or a pseudo-D3 2-fold axis (bottom row). Densities of TECs are numbered as in (d); and the two vertices without TEC but with RNA are indicated by white arrows.* ***f****, Boxed region in (a) containing RNA threads (radially colored as in the bar) and a TEC (cyan) with bound dsRNA (dashed box).* ***g****, Averaged TEC region, filtered to 4.5 Å and viewed as the southern-most TEC of (a). The RdRP-bound dsRNA has the same structure in all TECs and shows major (yellow arrows) and minor (white arrows) grooves.*

Each TEC is surrounded by rod-like densities with lengths up to ~650Å (Figure 4.1a-c, e-f, Figure 4.7a, Movie 4.1). In most regions, these rods form parallel striations with an inter-rod

distance of ~27Å as suggested early ([e.g., 10, 11]). Some of the rods exhibit the characteristic minor and major grooves typical of dsRNA duplex (Figure 4.1g). We therefore interpret these rod-like densities as dsRNA duplexes. Unlike the model of each genome segment spiraling around one TEC (Ref [12]), the duplexes do not spiral locally around TECs (Figure 4.1a-c, Figure 4.8 and Movie 4.1); instead, many extend tangentially from one TEC to another (*e.g.*, duplexes i-iii in Figure 4.1b; Figure 4.8), indicating that each dsRNA segment is organized beyond one TEC. Indeed, the whole RNA genome is organized into 7 to 8 non-concentric layers with visible connections between adjacent layers (Figure 4.1e and Figure 4.7). This extended organization of dsRNA is consistent with the rather long (~620 Å) persistence length of dsRNA[13] and would reduce the energy needed for genome packaging and transcription. One RNA duplex (the brown one in Figure 4.1g) binds to each of the 10 TECs at the same relative position and orientation, suggesting that this RNA duplex is a conserved feature among the 10 dsRNA segments. However, the organization of the remaining RNA duplex differs among the 10 TECs (Figure 4.8g). The two vertices without TECs are occupied only by roughly parallel dsRNA densities (Figure 4.1a-c).

We also obtained a 3.9Å resolution asymmetric reconstruction directly from the raw images of q-CPV and subsequently used non-crystallographic averaging to improve the resolution to 3.3Å for the TEC-containing regions (see Methods and Figure 4.9). The averaged map retains a short (~35Å) RdRP-bound dsRNA density (Figure 4.1g) and resolves the two protein components of the TEC: VP4 and RdRP (Figure 4.2a). We built a backbone model of the RdRP-bound dsRNA and *de novo* atomic models of both VP4 and RdRP (Figure 4.2c-f, Figure 4.10, Movies 4.3-4.8). VP4 and RdRP interact extensively (Figure 4.2a) with a buried interface area of ~2800Å^2.

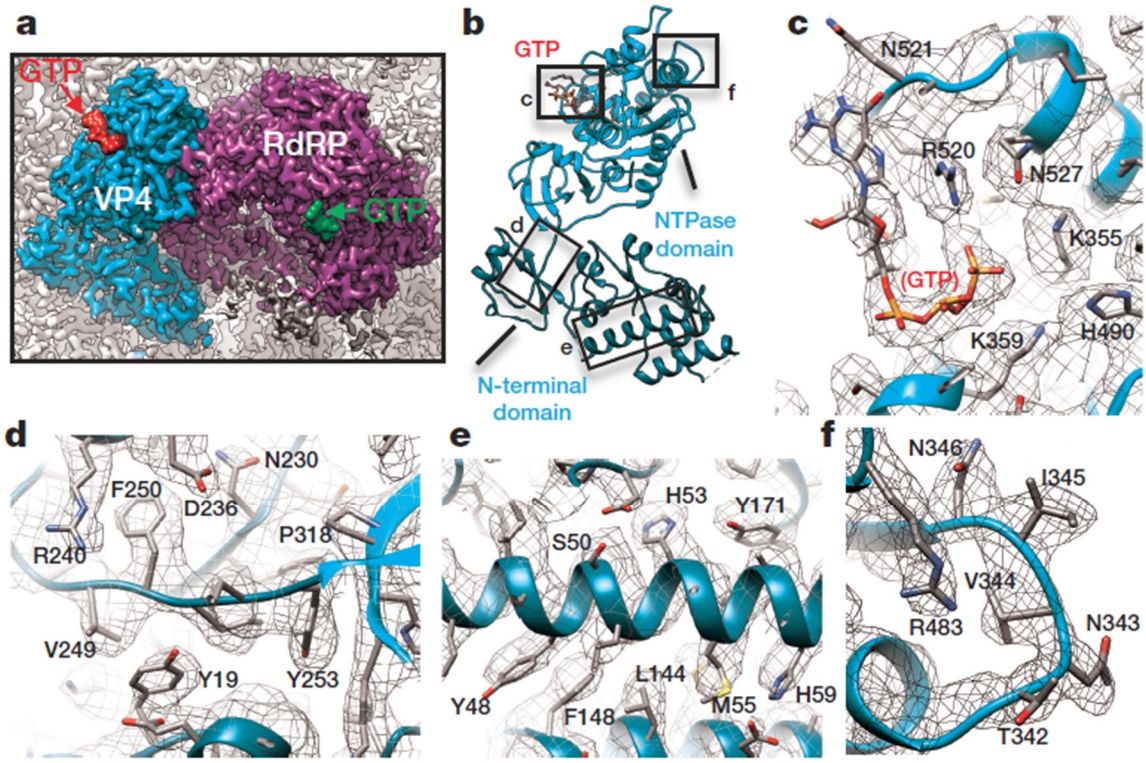


Figure 4.2 ***Averaged TEC map at 3.3Å resolution and de novo modelling of VP4. a****, Averaged map of the TEC region showing VP4 (cyan) and RdRP (purple), both anchored to the inner surface of the capsid (grey).* ***b****, Atomic model of VP4.* ***c-f****, The boxed regions in (b), showing density (meshes) superposed with atomic models of the GTP-binding site (c), a loop (d), a helix (e) and an RdRP-interacting loop (f).*

VP4 appears "L"-shaped and consists of an N-terminal (aa1-252) and a C-terminal (aa 253-561) domain, with two unresolved/flexible segments (aa 23-40 and 86-131) (Figure 4.2a,b, Movie 4.3). The N-terminal domain is formed by two small β-sheets and several α-helices and the main body of the C-terminal domain is a Walker-A α/β motif, a well-known NTP-binding motif found in the P-loop kinase family of proteins. Sequence analysis predicted an NTP binding site in VP4 (Refs [14, 15]). Indeed, the VP4 structure contains a GTP molecule at the predicted NTP binding site of the C-terminal domain (Figure 4.2c, Movie 4.4). We thus rename the C-terminal domain as the NTPase domain (Figure 4.2b,c). A similar fold was also observed in the N-terminal α/β domain of bluetongue virus VP4. But, remarkably, bluetongue virus VP4 is an RNA capping enzyme and its α/β domain does not bind GTP (Ref [16]). CPV VP4 and its homologs in other dsRNA viruses

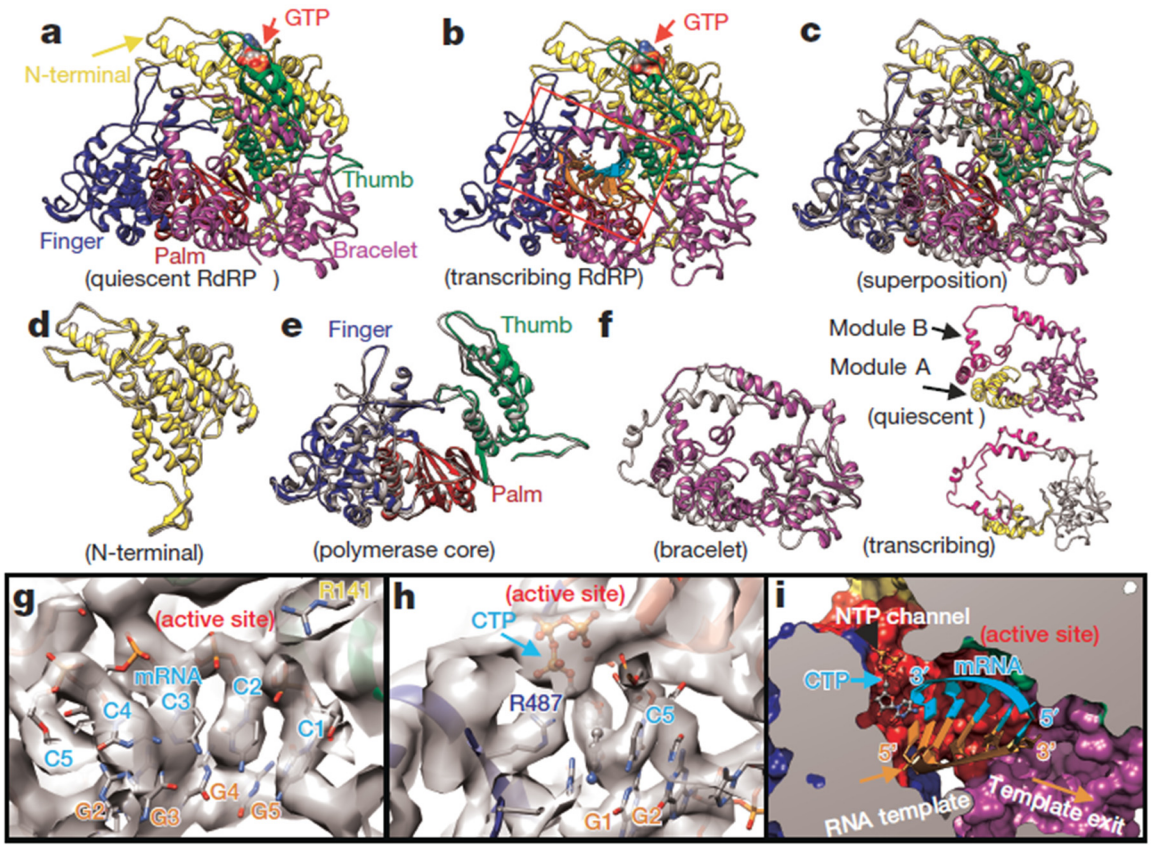


Figure 4.3 ***Comparison of RdRP in quiescent and transcribing states. a-b****, Ribbon models of RdRP in quiescent (**a**) and transcribing (**b**) states. The latter contains fragments of RNA template (orange) and nascent mRNA (cyan) inside the active site (box).* ***c-f,*** *Superpositions of RdRP structures in quiescent (colour) and transcribing (grey) states shown in full (**c**) and as separate domains – N-terminal (**d**), polymerase (**e**), and bracelet (**f**) with Modules A (yellow) and B (magenta) further highlighted on its right panel.* ***g-i****, Densities (grey) and models (ribbons and sticks) of nucleic acids in the active site of RdRP. The fragments of the (-)RNA template and the nascent mRNA in the active site are modelled as a poly-G and poly-C, respectively. In (**h**), a CTP is placed in the NTP-binding site and in (**i**), the template and mRNA form RNA duplex in the active site of RdRP (surface-rendered model).*

have been speculated to function as an NTPase, as an RNA 5'-triphosphatase (RTPase) or as a helicase[14, 17, 18]. Our structure supports VP4 as an NTPase but shows no interaction with dsRNA, suggesting that VP4 is unlikely a helicase. Whether VP4 is the CPV RTPase or an RdRP regulatory factor remains to be determined.

Like other RdRP structures[2, 3, 19], the CPV RdRP contains a polymerase core with finger (aa 349-515, 549-641), thumb (aa 730-863) and palm (aa 516-548, 642-729) subdomains (Figure 4.3a). This polymerase core is sandwiched between the N-terminal (aa 1-348) and C-terminal bracelet (aa 864-1225) domains (Figure 4.3 and Figure 4.10). A GTP is identified (Figures

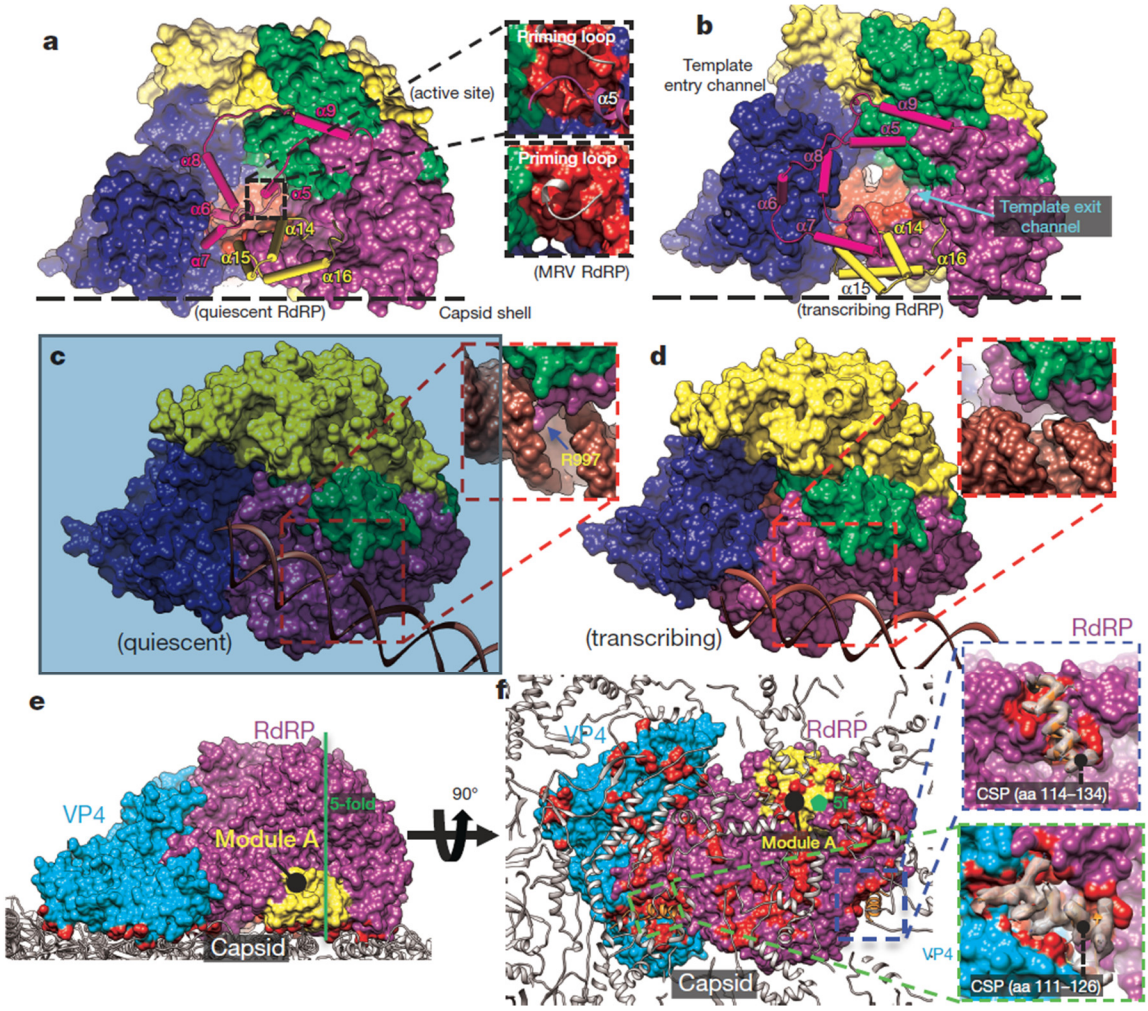


Figure 4.4 ***Interactions between TEC and capsid shell proteins (CSPs). a-b****, Conformational changes of modules A (yellow loops/helices as wires/cylinders) and B (magenta loops/helices as wires/cylinders) in quiescent* ***(a)*** *and transcribing* ***(b)*** *states. Module A interacts with the capsid shell, and the loop-Bα5 fragment of module B blocks the active site (inset) in the quiescent state* ***(a)*** *but retracts to expose the active site in the transcribing* ***(b)*** *state (see Figure 4.11).* ***c-d****, The RdRP-bound dsRNA (ribbon) in the quiescent state* ***(c)*** *is detached from RdRP in the transcribing state* ***(d)****.* ***e-f****, Interactions of CSPs (ribbons) with RdRP (purple and yellow) and VP4 (cyan). Residues of RdRP and VP4 within 4Å distance to the capsid shell are marked in red. An icosahedral 5-fold axis is indicated by a green line in* ***(e)*** *and a green pentagon in* ***(f)****. Insets in* ***(f)*** *indicate two CSP N-terminal helices (white density with ribbon-and-stick models): one (upper) interacts only with RdRP while the other (lower) with both RdRP and VP4.*

4.1g, 4.3a) at the position equivalent to the cap-binding site observed in the MRV RdRP[2]. Interestingly, the bracelet domain of q-CPV RdRP differs from that of MRV significantly, despite close similarities between both their polymerase core and their N-terminal domains. Consequently, the crystal structures of MRV RdRP has an open RNA template entry channel and an accessible polymerase active site[2]; while in the q-CPV RdRP, the polymerase active site is covered by the

bracelet domain and there is no recognizable channel for template entry (Figures 4.3a, 4.4a, Movies 4.5, 4.8). Since q-CPV is incapable of mRNA transcription, we considered that these structural differences might be characteristic of conformational differences between bracelet-containing RdRPs in the quiescent and transcribing states.

To test this hypothesis, we then determined the structure of actively transcribing CPV (t-CPV), obtained an averaged TEC map at 4.0 Å resolution, and built atomic models of VP4 and RdRP (Figure 4.3b, Figure 4.9, Movie 4.9). In t-CPV, the location of TECs remains the same, as do the structures of VP4 and those of the N-terminal and polymerase core domains of RdRP (Figure 4.3a-f, Figure 4.11-4.13, Movie 4.10). By contrast, the RdRP bracelet domain undergoes major conformational change (Figure 4.3d,e). Consistent with the above hypothesis, the *in situ* structure of the t-CPV RdRP is quite similar to the crystal structure of MRV RdRP in its elongation state[2] (Figure 4.14).

The most significant changes of the CPV RdRP between quiescent and transcribing states involve two neighboring structural modules in the bracelet domain, the capsid-proximal Module A (aa 1080-1140 containing helices Bα14-Bα16) and the VP4-proximal Module B (aa 912-1010 containing helices Bα5-Bα9) (Figure 4.4a-b, Figures 4.13 and 4.14). Compared to that in q-CPV, Module A in t-CPV rotates ~40° towards the capsid shell (Figure 4.3f, Figure 4.13, 4.14f-k). Consistent with previous icosahedral reconstructions, our asymmetric reconstructions show that the capsid shell of t-CPV expands outwards from q-CPV, with the maximal (~10 Å) expansion occurring at the vertex region[20, 21], to which Module A of the bracelet domain is attached (Figure 4.4e,f). Likewise, Module B refolds substantially from quiescent to transcribing state, such that a template entry channel is formed (Figure 4.4a,b) and the blockage of the active site by the Bα5-loop-Bα6 fragment is removed (Figure 4.3f, Figure 4.4a-b, Figures 4.13 and 4.14).

In the quiescent state, a helical dsRNA duplex is held inside a shallow cleft formed by Modules A and B (Figures 4.1g, 4.4c, and Figure 4.11d,f) through interaction between a major groove of the RNA duplex and residue Arg979 of Module B (Figure 4.4c inset). In the transcribing state, this RNA duplex becomes detached perhaps as a result of refolding the RdRP bracelet domain (Figure 4.4d, and Figure 4.11e,g). We anticipate that detachment of the RNA duplex would permit RNA to slide towards the template entry channel for RNA synthesis in t-CPV. Indeed, in the catalytic center of the t-CPV RdRP, we observe weak densities (Figure 4.3b,g-i) that match the RNA duplex in the crystal structure of the MRV RdRP elongation complex[2]. We are able to place a 5-basepair RNA backbone model in the active site and a CTP at the NTP binding site (Figure 4.3g-i).

In addition to enclosing the viral genome and anchoring TECs, the capsid shell protein (CSP) also regulates polymerase activity in dsRNA viruses[22-24]. In particular, the CSP N-terminal fragment plays roles in genome replication, mRNA transcription and capping[23, 25, 26]. A CSP N-terminal fragment, unresolved in all previous structures[21, 27-29], is resolved here to form a helix in the two TEC-interacting CSP subunits in both q-CPV and t-CPV (Figure 4.4e,f). The N-terminal helix of one CSP inserts into the interface between the NTPase domain of VP4 and the finger subdomain of RdRP (Figure 4.4f lower inset) and that of the other CSP interacts with the bracelet domain of RdRP (Figure 4.4f upper inset). Notably, the former is in proximity to the NTP-binding site of the VP4 NTPase, suggesting how the N-terminal fragment of CSP is positioned to affect TEC. In addition, the structures reveal that other regions (*i.e.*, vertices area) of CSP also interact with Module A of the RdRP bracelet domain (Figure 4.4e-f). From quiescent to transcribing state, Module A and the CSP regions involved in this interaction both undergo conformational changes. Taken together, these results point to a sequence of conformational

changes that leads to activation of endogenous transcription. Specifically, environmental cues cause the capsid shell to expand[21], which triggers refolding of the RdRP bracelet domain, leading to formation of the entry channel for a RNA template and exposure of the polymerase active site for RNA synthesis.

4.4 Methods

4.4.1 Sample preparation and cryoEM imaging

CPV particles were purified as described previously[29]. Purified polyhedra were treated at pH 10.8 with an alkaline solution (0.2 M Na_2CO_3-$NaHCO_3$) for 1 hour, and then centrifuged at 10,000g for 40 minutes. The supernatant was collected and centrifuged at 80,000g for 60 minutes at 4°C to pellet the CPV virions. The resulting pellet was directly re-suspended in the quiescent buffer (70mM pH 8.0 Tris-Cl, 10 mM $MgCl_2$, 100 mM NaCl and 2 mM GTP). In order to prepare the transcribing CPV (t-CPV) particles, 30µl purified CPV was incubated in a reaction buffer (70mM Tris, pH 8.0, 10 mM $MgCl_2$, 100 mM NaCl, and 1 mM SAM+2mM GTP+2mM UTP+2mM CTP+4 mM ATP) at 31°C for 15 min, and then the reaction was stopped by quenching the reaction tubes on ice.

To prepare cryoEM grids, 2.5µl of purified CPV sample was applied to a Quantifoil grid (2/2), blotted for 15 seconds with an FEI vitrobot in 100% humidity, and then plunged into liquid ethane. CryoEM images of the quiescent CPV (q-CPV) were collected in an FEI Titan Krios cryo electron microscope, operated at 300kV with a nominal magnification of 49,000x (Figure 4.9g). The microscope was carefully aligned and electron beam tilt was minimized by a coma-free alignment procedure. Images were recorded on a Gatan K2 direct electron detection camera with the counting mode, and the pixel size was calibrated as 1.01Å/pixel on the specimen using catalase

crystals. The dose rate of the electron beam was set to ~8e^-/pixel/s, and the image stacks were recorded at 4 frames/sec for 3 seconds. The drift between frames in each image stack was corrected with the UCSF software[30], and the total 12 frames of each stack were merged to generate a final image with a total dose of ~25e^-/Å^2. Contrast transfer function (CTF) parameters, including defocus values and astigmatism, were determined by CTFIND[31] (Figure 4.9g).

Sample grid preparation, cryoEM imaging and drift correction of frames for the transcribing CPV (t-CPV) were performed using the same procedure described above for q-CPV with the exception of the camera used. The t-CPV cryoEM images were recorded on a new Gatan K2 direct electron detection camera attached to a Gatan imaging filter (GIF Quanta) with a pixel size of 1.36Å at the specimen scale (Figure 4.9g).

4.4.2 Asymmetric reconstruction based on original images

A total of 68,526 particles were selected for image processing using *Frealign*[32] and *Relion*[33]. The 2x binned data set was first processed using icosahedral symmetry with *Frealign*[32]. The centers of all particles were then fixed and used for the asymmetrical global search with *Frealign* using 4x binned data set starting at 20 Å resolution.

To generate an initial model, we placed the crystal structure of the MRV RdRP[2] under a previously obtained CPV capsid map[34] at the location corresponding to that in MRV capsid as previously reported[9] and imposed a tetrahedral symmetry (*i.e.*, with 4 three-fold axes, 3 two-fold axes and 12 asymmetric units), resulting in a montage map with an empty CPV capsid containing 12 RdRPs but without any VP4. This montage map was filtered to 30Å resolution and used as the initial model for image processing with *Frealign*. After 9 iterations of global search and 2 iterations of refinement, the resolution of the density map was determined to be 3.9Å. In the final map, only

3 RdRPs (#8-10 in Figure 4.1d) remained at the same locations as in the initial model with the tetrahedral symmetry.

The final map was reconstructed using the top 47,968 (70%) particles of the original unbinned data set. Averaging all TEC densities under different vertices was performed following the procedure described previously[35] to improve the density quality and the resolution. The effective resolution of the asymmetrical and averaged reconstructions were estimated to be 3.9 Å and 3.3 Å, respectively, based on the FSC (≥0.143) and the correlation coefficient (≥0.5) between the density map and atomic model calculated with *Phenix* (Figure 4.9g)[36, 37]. These estimated resolutions are consistent with the observed structural features of the density maps (Figure 4.2, Figure 4.9e, and Movies 4.3-4.8). The averaged map was filtered to the spatial frequency of 1/(3.3 Å) and sharpened with a reverse B-factor of -120Å^2. This B-factor was chosen with a trial-and-error method based on the optimization of noise level, backbone density continuity, and emergence of side-chain densities.

Since there were no densities in the initial montage model at the VP4 locations, the emergence of VP4 densities in the map and the match of side-chain densities to those expected from the VP4 amino acid sequence (Figure 4.2) provide strong internal controls for the validity of the high resolution cryoEM map. Consistent with this assessment, the locations of the RdRP in the final reconstruction are not only different from those in the initial montage model, but also are related by D3 symmetry instead of the tetrahedral symmetry in the initial model. Most convincingly, the density features in the final map agree with the CPV RdRP amino acid sequence but differ from that of the MRV RdRP used in the initial model.

In addition, we also performed independent reconstruction without using the model of the 12 MRV RdRPs, and obtained a nearly identical structure from the same dataset. In this procedure,

we first determined an icosahedral reconstruction without using any initial models. This icosahedral reconstruction was used to restrain refinement without symmetry (i.e., symmetry operator is C1) to search for orientation around the 60 icosahedral-symmetry-related locations with *Relion*[33]. This independent result further validate our TEC structures.

To obtain the 3D structure of the transcribing particles, we low-pass filtered the above 3D map of q-CPV to 30Å resolution and used it as the initial model. After 11 iterations of asymmetrical global search and 2 iterations of local refinement, the density map converged to a resolution of 4.8 Å, and the density quality of the TEC was further improved to ~4.0 Å resolution by aligning and averaging all TEC densities inside the asymmetric reconstruction (Figure 4.9d,f,g).

4.4.3 Asymmetric reconstruction using capsid-subtracted images

To further improve the genome structure, we used the following procedure to carry out asymmetric reconstruction of q-CPV with the same particle image dataset but with capsid contribution subtracted. As illustrated in Figure 4.5, this procedure includes four stages: 1, capsid subtraction in raw particle (orange); 2, initial model generation (green); 3, asymmetric feature emergence in *Relion[33]* refinement (blue); 4, orientation selection (purple).

In the first stage (orange in Figure 4.5), we determined the orientation and center parameters for each particle and obtained an icosahedral reconstruction with *Frealign[32]* from raw particles with an inverse B-factor of -40 $Å^2$ (a-b). Based on these parameters, a CTF-corrected projection (c) with empirical B-factor of 160 $Å^2$ was generated. Next, the capsid contribution to the images was removed by subtracting the 2D projection corresponding to the icosahedral orientation of each image as done before[38-41] with the following improvements. To accurately subtract the contribution from the capsid, we determined a scaling factor between capsid projection (c) and

each raw particle image (a). The projection and raw images were both band-pass filtered between 1/400 $Å^{-1}$ and 1/29 $Å^{-1}$, then radially masked based on the inner and outer diameters of capsid to produce ring-shaped projections (d) and raw (e) images. The standard deviations of these ring-shaped images were calculated and used to normalize both the unmasked and masked (*i.e.*, ring-shaped) projections. The cross-correlation coefficient (0 to1) between the ring-shaped raw image and the normalized ring-shaped projection was computed and used as the probability factor measuring the contribution of capsid signal in the raw particle image. Each raw image was then subtracted by the unmasked projection multiplied by this probability factor to generate a capsid-subtracted particle (f) for the following refinement. Particles with a probability factor less than 0.1 were not included in the subsequent analyses.

In the second stage (green in Figure 4.5), the map from the above *Frealign* asymmetric refinement (g) was low-pass filtered to 60 Å resolution, masked with a 260 Å radius (h), and used to refine the capsid-subtracted particle (f) with *Relion* version 1.2. The Tau2_fudge value (T-factor) in *Relion* was set to 0.5. T-factor is an *ad hoc* value in *Relion* to tune refinement speed, and a value of 0.5 slowed down the refinement progression thus ensuring the priority use of low resolution (up to 20Å, such as dsRNA) data in the refinement. This refinement led to a reconstruction without the capsid (i). This capsid-removed map has 12 TECs with D3 symmetry, which could be classified into two groups: the first group containing six better-resolved TECs close to the 3-fold axis (polar) and the other group containing six less-resolved TECs near the equator (tropical), suggesting potential smear of density due to orientation mis-assignments or TEC flexibility/lower occupancy near the equator.

In order to further eliminate potential orientation mis-assignments, we next conducted the third stage of data processing (blue in Figure 4.5). We first low-pass filtered the capsid-removed

reference (i) to 32 Å resolution (j) and used it to drive *Relion* refinement with the capsid-subtracted particles (f). The T-factor used in this refinement is 0.1, only 2.5% of that used in *Relion* convention, thus ensuring slow progression of the refinement. Slower refinement provides time for asymmetrical feature to emerge. *Relion* global search was carried out with a 3.75 degrees angular interval, followed by local angular search with 1.875 degrees interval and highly-constrained translational search (0.7 pixel in range with 0.5 pixel interval). Asymmetrical RNA density feature with 10 TECs emerged after 10 iterations (k). In our procedure, one way to prevent trapping into local minima in orientation assignment due to symmetric structural elements is to filter the current refinement result back to ~32 Å resolution and refine with T-factor of 0.1 again to remove residual symmetric feature from the working reference. This process is carried out iteratively.

To further improve resolution of the 3D map, we carried out the fourth stage for particle orientation selection (purple in Figure 4.5). From the orientation of each particle determined in the high-resolution (~3Å) icosahedral reconstruction (b), we calculated 60 *icosahedral-related* orientation candidates. The task of the rest of the fourth stage of data processing is to select one out of these 60 orientation candidates to be the asymmetric orientation of the particle as done before[4, 5, 42]. To do this, we continued to run *Relion* refinement for 15 iterations using the above asymmetric map with 10 TECs (k) as initial model and the orientation determined by each iteration was recorded, giving rise to 15 *Relion orientations* for each particle. For each of these 15 Relion orientations, we calculated its angular distances to the 60 icosahedral-related orientation candidates; and the icosahedral-related orientation candidate with the smallest angular distance was selected as the *working* orientation for that iteration, resulting in a total of 15 working orientations for each particle. The particle would be retained if 14 or all of its 15 working

orientations are the same (*i.e.*, the *selected* orientation) and their averaged angular distance was less than 3 degrees. Otherwise, this particle will be discarded. This procedure yielded a total of 11,741 particles with selected orientation. The original raw images of these selected particles were combined to generate an asymmetric reconstruction using *Frealign* and the resolution was determined to be 5.1 Å.

As shown in Figure 4.6, this procedure was repeated by using a Gaussian ball to replace the capsid+TEC model (g) in the initial model generation stage (green in Figure 4.5). The result is the same, confirming our procedure was not influenced by the choice of initial model.

4.4.4 Atomic modeling and visualization

The atomic models of both RdRP and VP4 in the quiescent state were built with *Coot* [43] and refined with *Phenix*[37] as described previously[44].

The atomic model of the VP4 structure was manually built with *Coot*. Because no homology models of VP4 previously existed, the Cα carbon backbone was constructed by matching the VP4 amino acid sequence to the density map. Once the correct placement of each residue was ensured, the backbone was converted to a purely alanine backbone by the function "Mainchain," and mutated to the corresponding amino acids through the function "Mutate Residue Range." With the initial model now completed, the "Density Fit Analysis" validation tool was used to screen for sequences of the model that did not fit the density. When identified, these sequences and the amino acids surrounding them were examined for any other possible conformations that would better fit the density. Due to the high resolution of this structure, this was completed through the refinement tool "Real Space Refine Zone," which optimizes the fit of the model to the mass density while preserving stereochemistry. Additionally, refinement was also performed based on

the Ramachandran plot, an important indicator of three-dimensional protein structure that validates the torsion angles of a protein chain. In the Ramachandran plot, any residues with disallowed values were selected, and the stereochemistry of that residue along with its surrounding residues was optimized with the refinement tool "Regularize Zone." After ideal Ramachandran values were obtained (<1% outliers), the refinement function "Rotamers" was used to select a rotamer that best fit the density.

The atomic model of the polymerase structure was also manually built with *Coot*. However, since an atomic model for the MRV polymerase was available in the Protein Data Bank (accession number 1MUK), this model was used as a template to assist with model building through the identification of the N-terminus, C-terminus, and various secondary structures. Once the Cα carbon backbone was built by matching the polymerase amino acid sequence to the density map and mutated to the appropriate amino acids, the model was refined with "Regularize Zone," "Rotamers," and "Real Space Refine Zone." The model was validated with the Ramachandran plot and the function "Density Fit Analysis." The complex of VP4 and polymerase was then refined with *Phenix*, including the real space refinement[37].

The atomic models of the transcribing state were built by fitting the atomic structures of RdRP and VP4 at quiescent state into the density, manually adjusting the changed residues with *Coot*[43] , and refining the models with *Phenix*[37].

Visualization, segmentation of density maps, and generation of videos were done with UCSF Chimera[30].

4.5 Extended Data

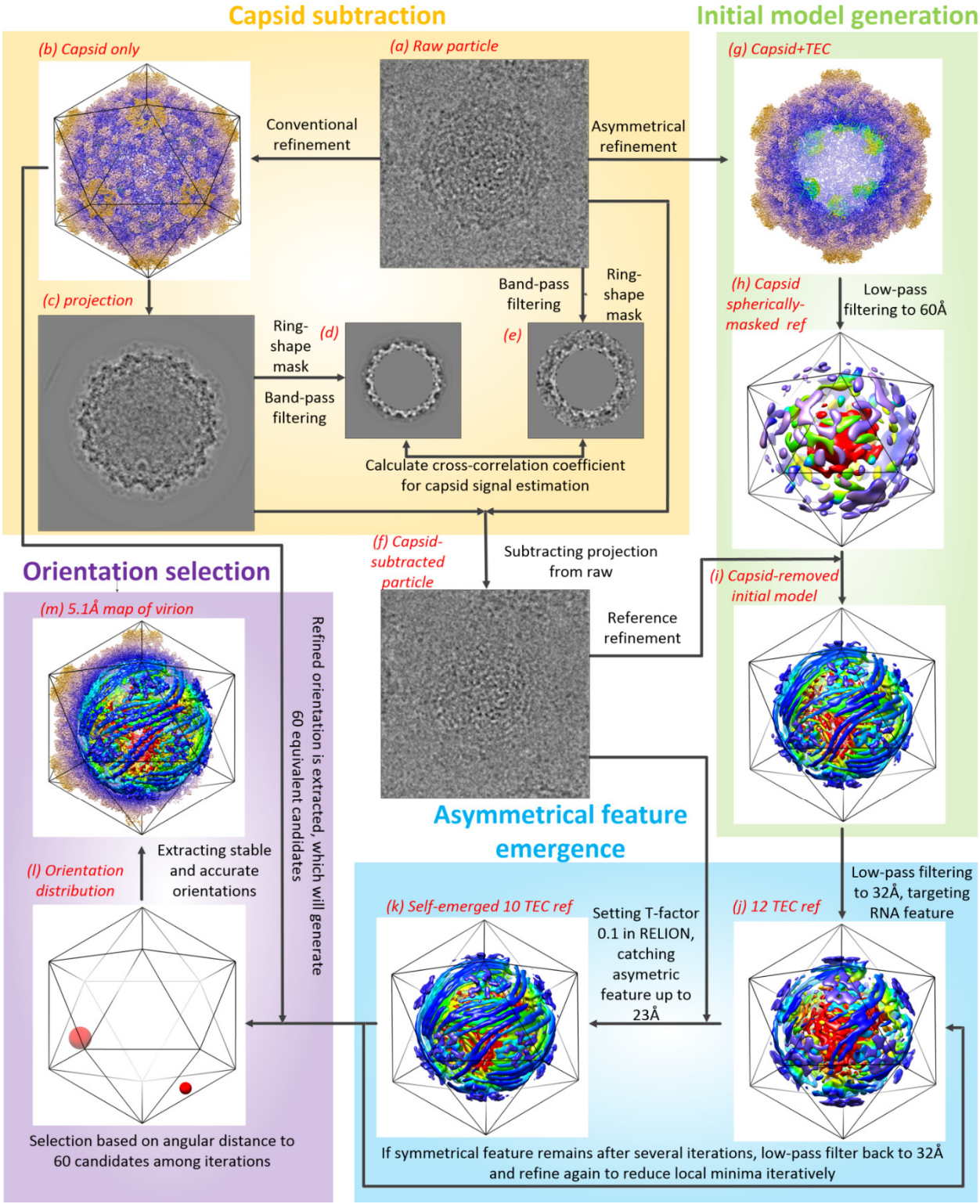


Figure 4.5 ***Illustration of the asymmetric reconstruction procedure using particles with the capsid density subtracted.***

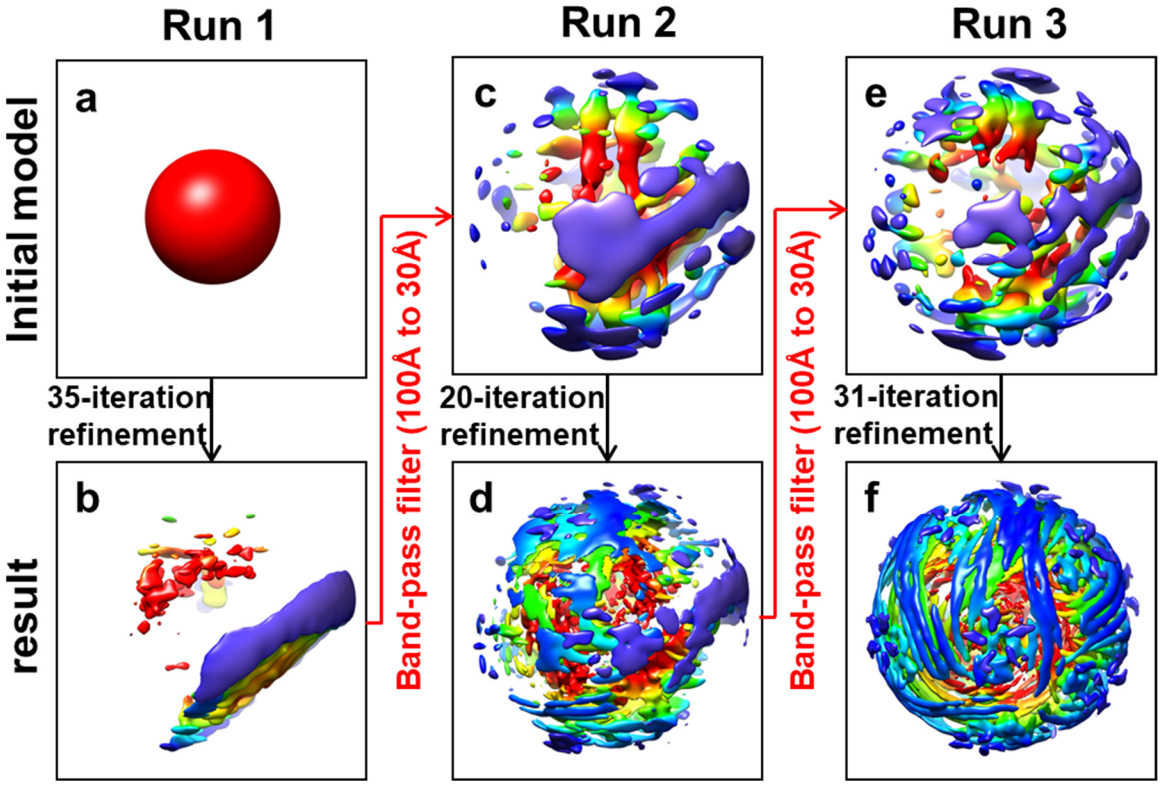


Figure 4.6 ***Validation of asymmetric reconstruction from capsid-subtracted images using a Gaussian ball as the initial model.***

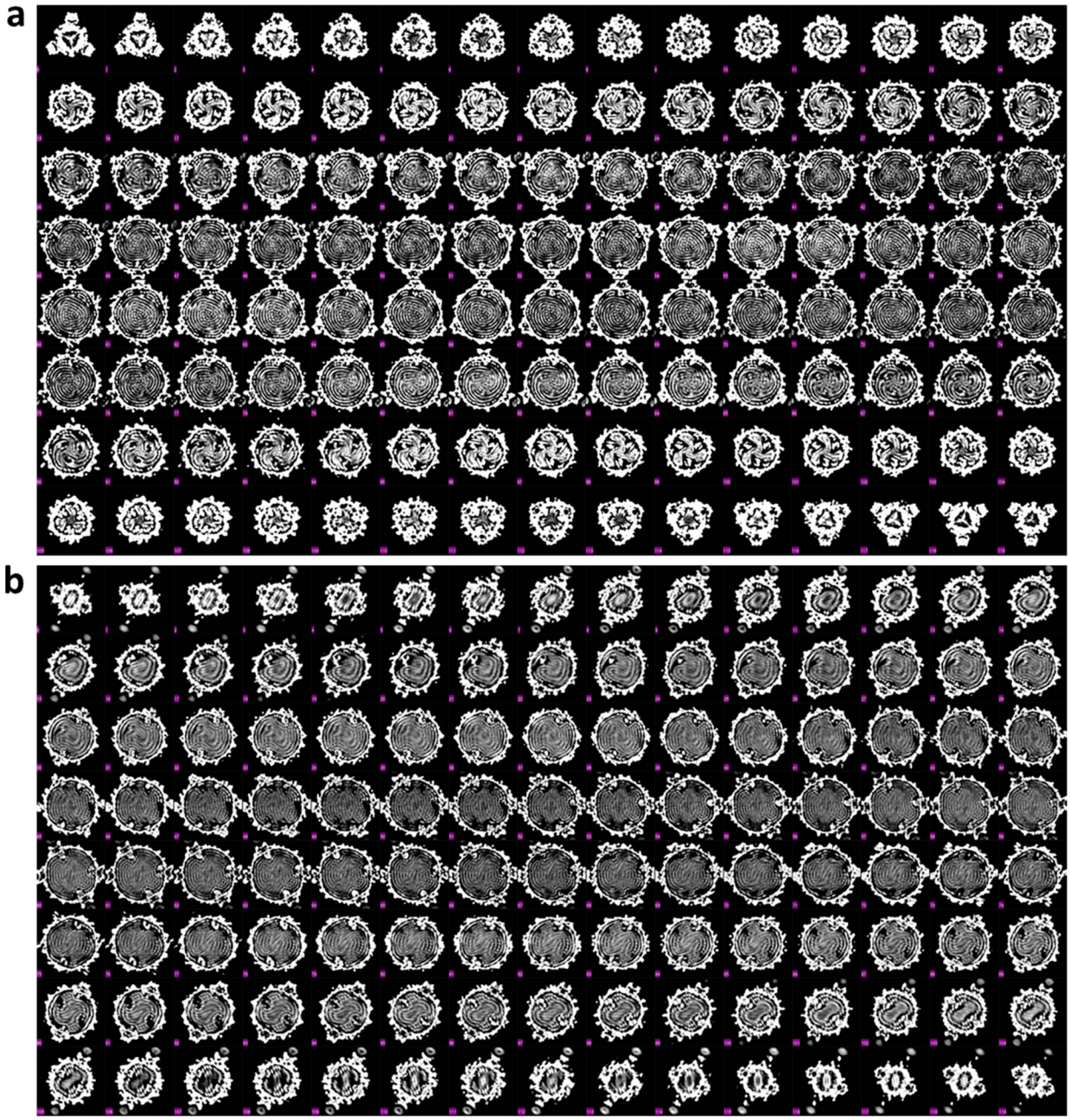


Figure 4.7 ***Sections of the q-CPV density map along the 3-fold (i.e., the earth axis) (a) and 2-fold (b) axes of the pseudo-D3 symmetry.*** *Note: the Lack of 3-fold and 2-fold symmetry in the RNA density in contrast to the perfect symmetry of the capsid shell proteins. Pixel size=4.04Å; Clipped map size=166*166*120 pixels.*

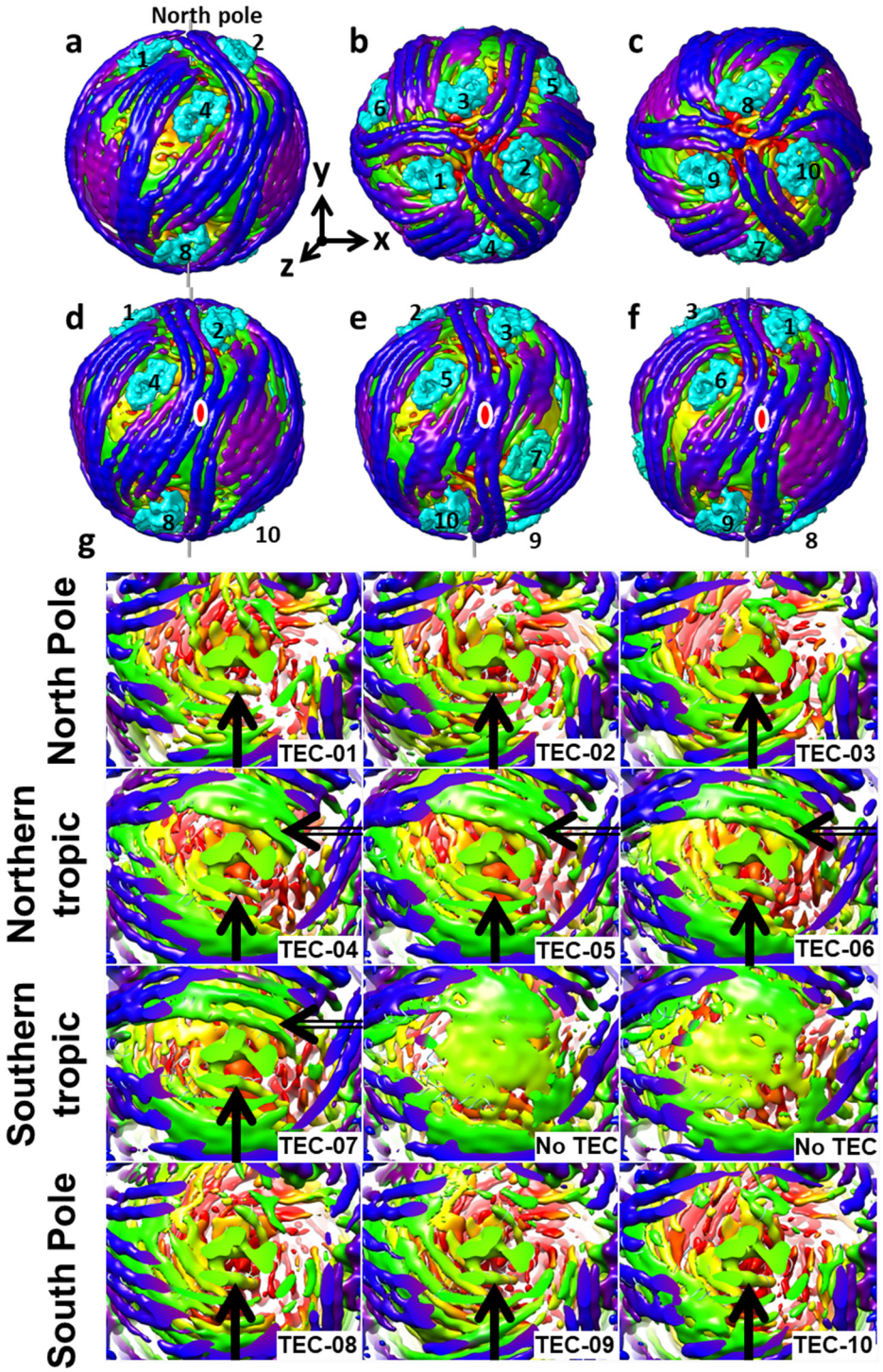
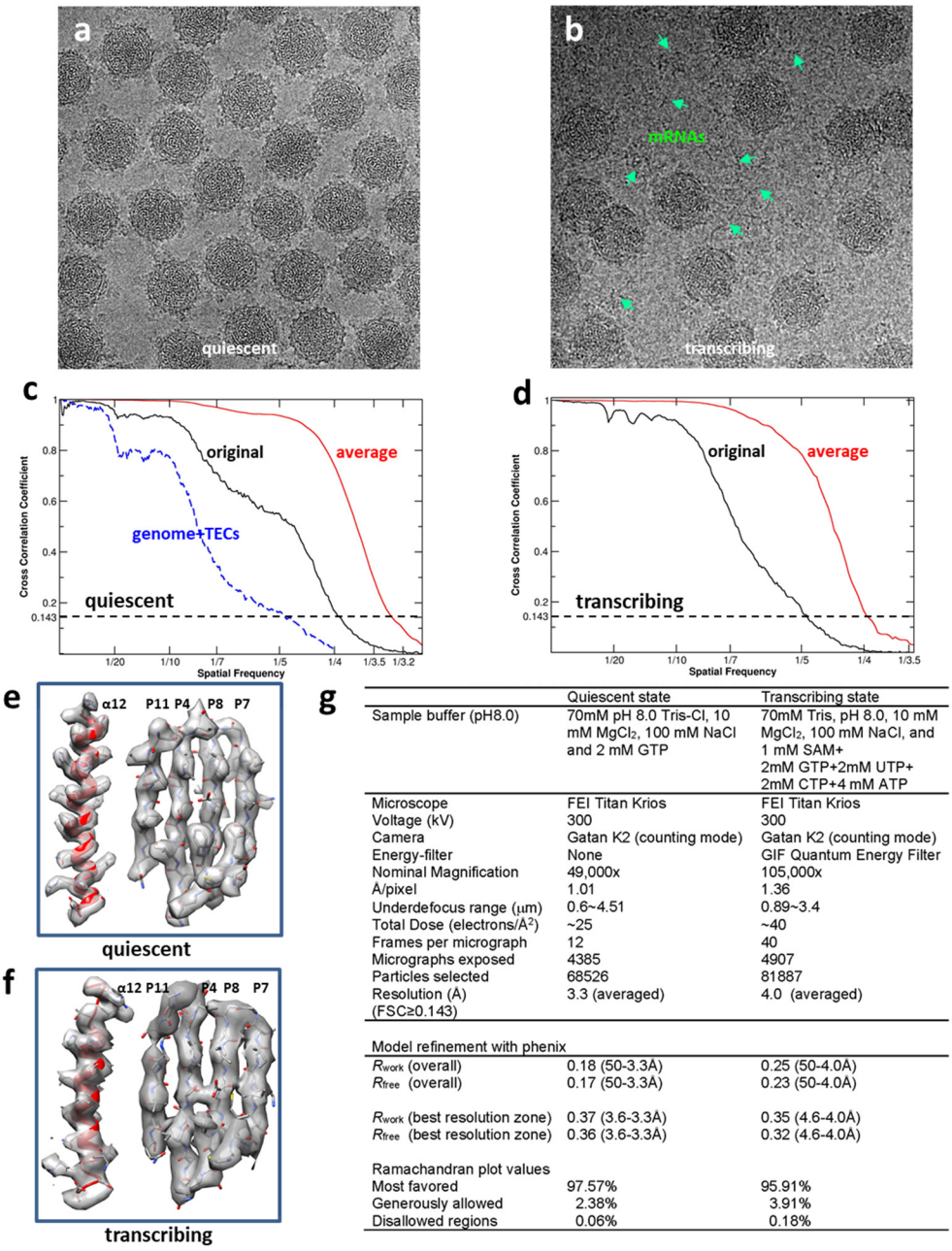


Figure 4.8 ***dsRNA density maps in the quiescent state. a****, View of TEC+RNA densities with the same orientation of Figure 4.1d.* ***b-c****, The same view as in (a) but rotated by +90° (b) or -90° (c) along x axis in (a) to view from either north (b) or south (c) poles.* ***d-f****, Three views from three 2-fold axes on the equator, each is rotated by 120 ° along the y axis from each other.* ***g****. dsRNA density maps at the twelve vertices. TECs are arranged and numbered according to Figure 4.1d. First row: TECs 1, 2, 3; Second row: TECs 4, 5, 6; Third row: TEC 7 and two unoccupied positions; Fourth row: TECs 8, 9, 10. All TECs have a dsRNA segment bonded at the flange, each marked with a black arrow. Compared with polar TECs, all tropical TECs (4-7) have extra piece of dense rods, with locations indicated with an opened black arrow.*



g

	Quiescent state	Transcribing state
Sample buffer (pH8.0)	70mM pH 8.0 Tris-Cl, 10 mM $MgCl_2$, 100 mM NaCl and 2 mM GTP	70mM Tris, pH 8.0, 10 mM $MgCl_2$, 100 mM NaCl, and 1 mM SAM+ 2mM GTP+2mM UTP+ 2mM CTP+4 mM ATP
Microscope	FEI Titan Krios	FEI Titan Krios
Voltage (kV)	300	300
Camera	Gatan K2 (counting mode)	Gatan K2 (counting mode)
Energy-filter	None	GIF Quantum Energy Filter
Nominal Magnification	49,000x	105,000x
Å/pixel	1.01	1.36
Underdefocus range (μm)	0.6~4.51	0.89~3.4
Total Dose (electrons/Å^2)	~25	~40
Frames per micrograph	12	40
Micrographs exposed	4385	4907
Particles selected	68526	81887
Resolution (Å) (FSC≥0.143)	3.3 (averaged)	4.0 (averaged)
Model refinement with phenix		
R_{work} (overall)	0.18 (50-3.3Å)	0.25 (50-4.0Å)
R_{free} (overall)	0.17 (50-3.3Å)	0.23 (50-4.0Å)
R_{work} (best resolution zone)	0.37 (3.6-3.3Å)	0.35 (4.6-4.0Å)
R_{free} (best resolution zone)	0.36 (3.6-3.3Å)	0.32 (4.6-4.0Å)
Ramachandran plot values		
Most favored	97.57%	95.91%
Generously allowed	2.38%	3.91%
Disallowed regions	0.06%	0.18%

Figure 4.9 ***CryoEM reconstructions of CPV in the quiescent and transcribing states. a-b****, CryoEM images of CPV particles in quiescent (**a**) and transcribing (**b**) states. These images were obtained by aligning and averaging frames in direct electron counting image stacks. Fiber-like nascent mRNAs are visible over background in (b) (marked by green arrows), while the background in (a) is clean.* ***c-d****, Fourier shell correlation coefficients (FSCs) as a function of spatial frequency between two half maps for reconstructions in the quiescent (c) and transcribing (d) states. The black and red lines represent FSCs for the asymmetrical reconstructions of capsid + genome and the locally averaged TEC densities, respectively. The effective resolutions of the local averaged maps are ~3.3 Å (c) and ~4.0 Å (d) resolution (FSC ⩾0.143) for maps in the quiescent and transcribing, respectively.* ***e-f****, CryoEM densities (grey surface representations) superimposed with atomic models (ribbons and sticks) for the quiescent (e) and transcribing (f) states. The α-helix (Pα12) and the four-stranded β-sheet (P4, P7-8 & P11) in (e) and (f) are both from the palm subdomain of the polymerase domain at 3.3 Å (e) and ~4.0 Å (f) resolutions.* ***g****, Statistics of CPV reconstructions and atomic model refinement.*

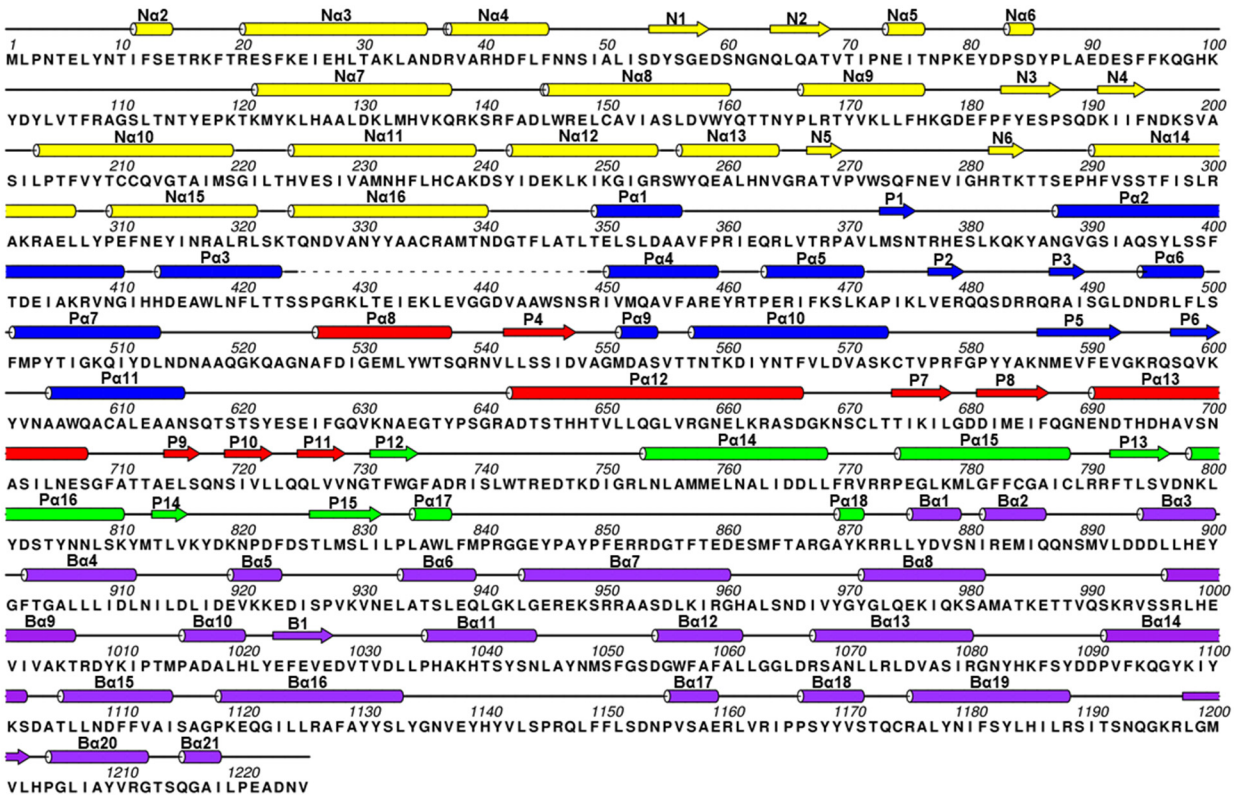


Figure 4.10 ***Sequence and secondary structure assignment of CPV RdRP in the quiescent state.*** *α-helices were marked by cylinders, β-strands by arrows, loops by thin lines, and the flexible tip domain by dashed lines. The colour scheme is the same as Figure 4.3a.*

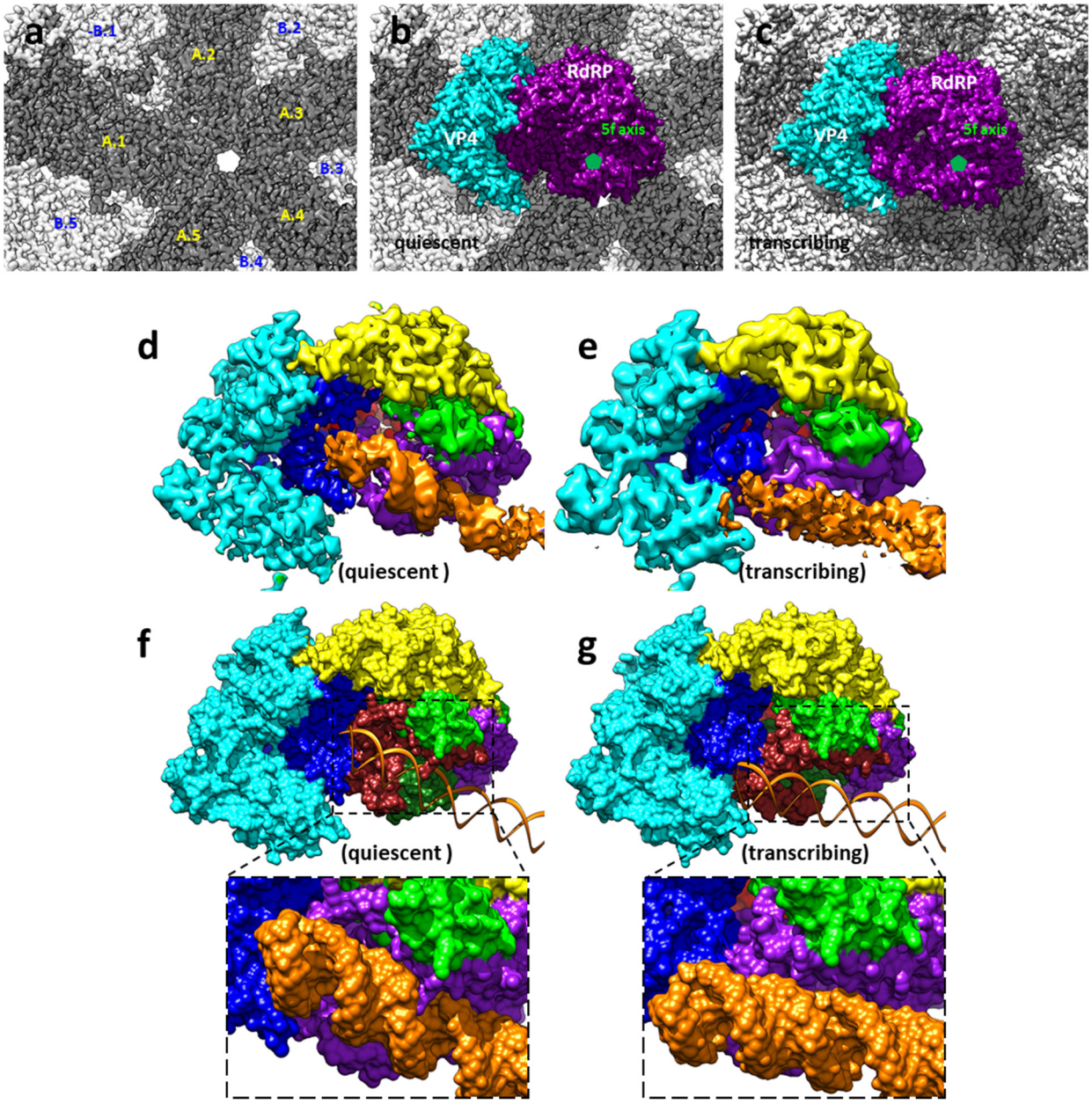


Figure 4.11 ***The RdRP-bound dsRNA in the quiescent and transcribing states. a-c****, Location of a TEC on the inner surface of the capsid shell in the quiescent and transcribing states. The inner surface of the CPV capsid (**a**) with 10 CSPs labelled (CSP-A.1/B.1 ~ CSP-A.5/B.5). Position of a TEC on the inner surface of capsid in the quiescent (**b**) and transcribing (**c**) states. VP4 and RdRP are colored cyan and purple, respectively. An icosahedral 5-fold axis is indicated with a small green pentagon. **d-e**, CryoEM densities of TEC and dsRNA (orange) in the quiescent (**d**) and transcribing (**e**) states. **f-g**, Models of TEC (surface representation) and dsRNA (ribbons) in the quiescent (**f**) and transcribing (**g**) states. Close-up views show the bound dsRNA (surface representation) on RdRP in the quiescent state (f) and its detachment in the transcribing state (g). VP4 is coloured cyan and the RdRP is coloured as in Figure 4.3a. Note: All surfaces displayed in this figure were rendered from models, except for the density maps of RdRP+dsRNA in (d-e).*

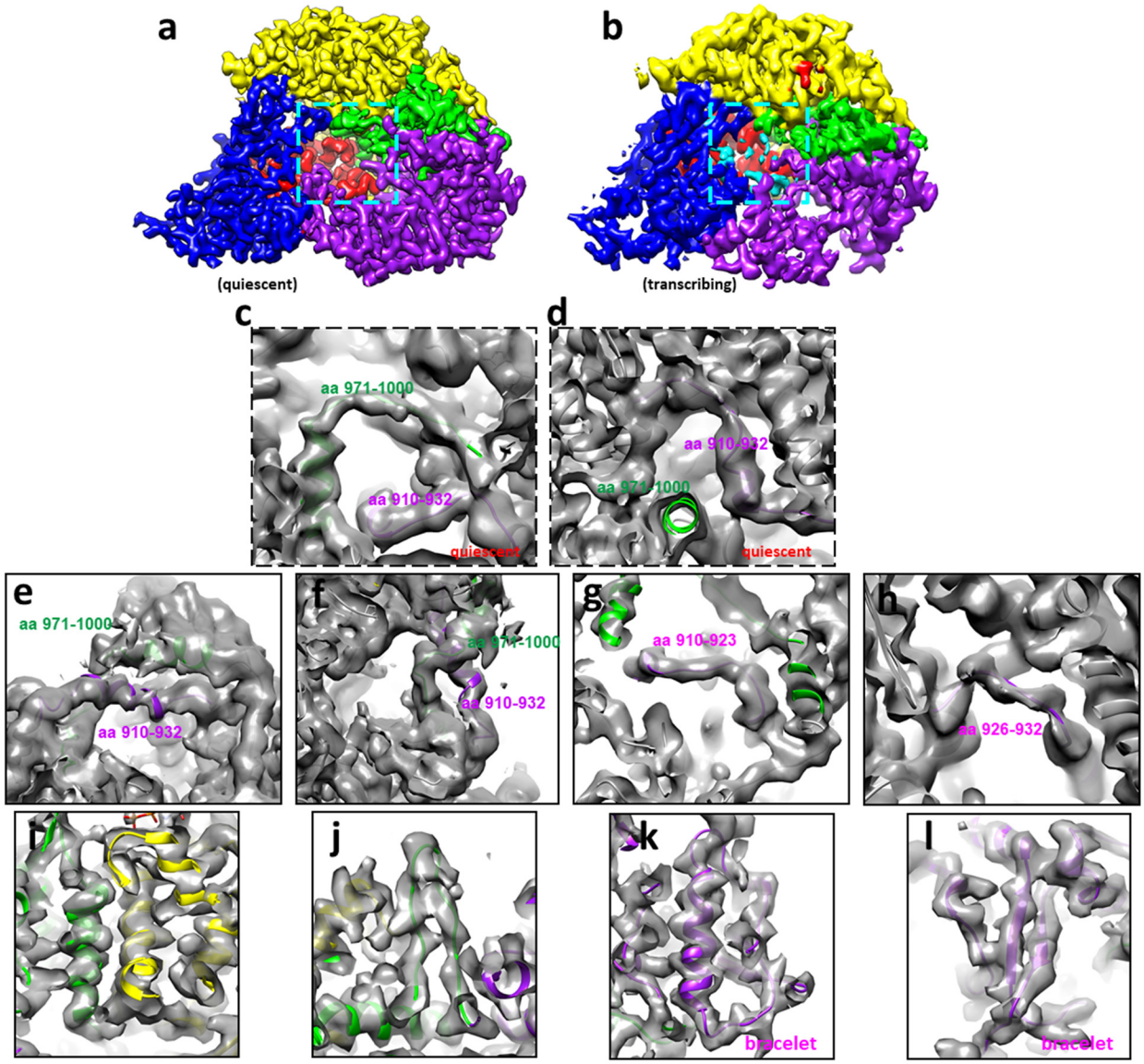


Figure 4.12 ***Tracing amino acid residues 910-932 and 971-1000 of module B of the bracelet domain of RdRP in the quiescent and transcribing states. a-b****, CryoEM densities of RdRP in the quiescent (a) and transcribing (b) states. The locations of the residues 910-932 and 971-1000 are indicated with cyan boxes in (a) and (b). Due to their flexibility, these residues are not readily visible when displayed as in (a) and (b) but become visible when the maps are filtered to a lower resolution (e.g., 4.5Å resolution) as in (c-f). The colour scheme of domains/subdomains is the same as in Figure 4.3a.* ***c****, Trace of the residues 971-1000 (green) and 910-932 (purple) of module B of the bracelet domain of RdRP in the quiescent state.* ***d****, The same as (c) but in a different view.* ***e****, Trace of the residues 971-1000 (green) and 910-932 (purple) of module B of the bracelet domain of RdRP in the transcribing state.* ***f****, The same as (e) but in a different view to show the unambiguous trace of the two peptide fragments.* ***g-h****, Trace of the residues 910-923 (g) (purple) and 926-932 (h) (purple) of the bracelet domain of RdRP in the transcribing state, showing the unambiguous trace of the two peptide fragments.* ***i-j****, CryoEM densities (grey) and model (ribbon) of RdRP in the transcribing state, showing α-helices (i) and a β-hairpin (j). The colour scheme of domains/subdomains is the same as in Figure 4.3a.* ***k-l****, Trace of the residues of the bracelet domain of RdRP in the transcribing state, showing a α-helix (k) and a β-sheet (l).*

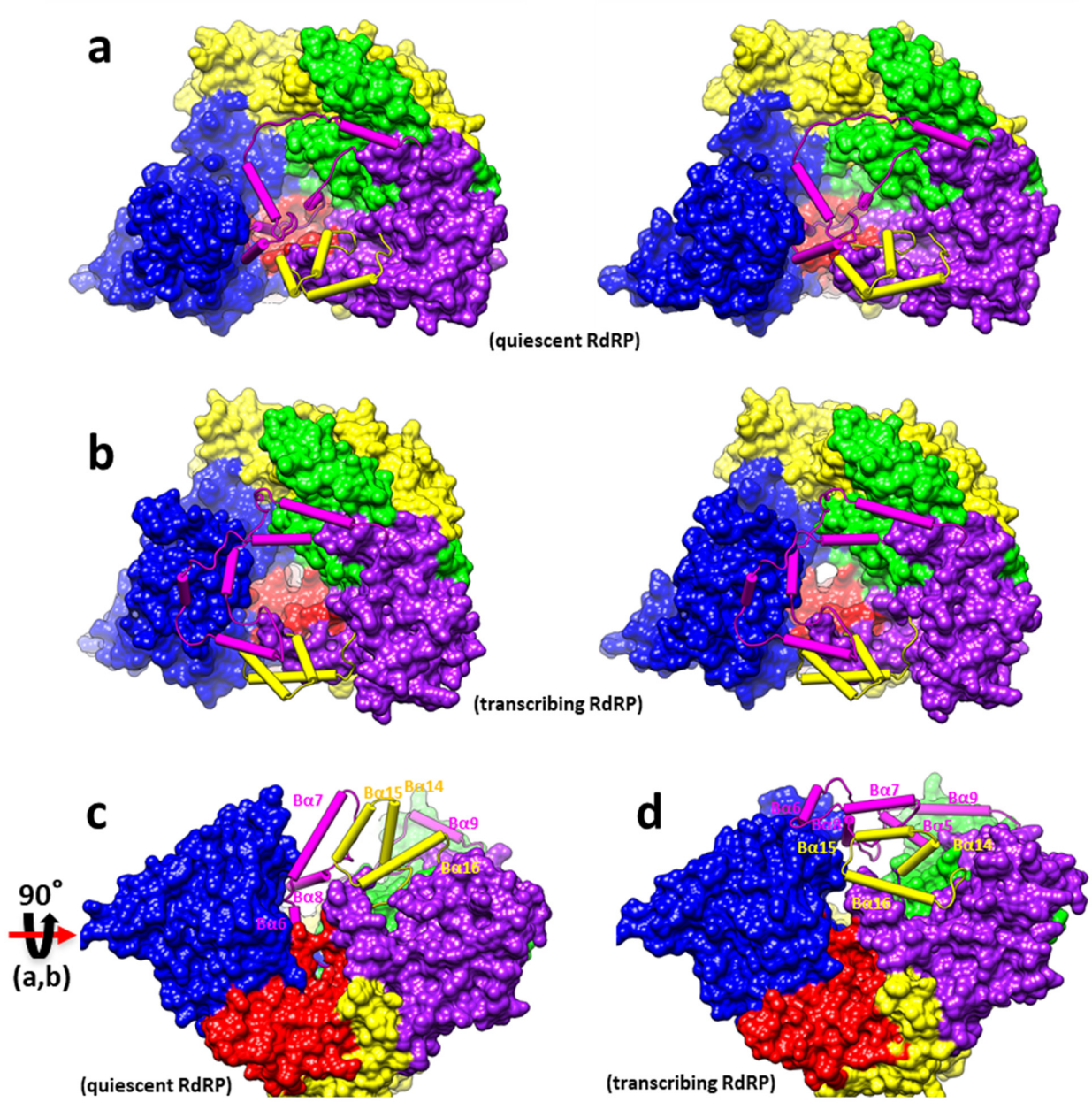


Figure 4.13 ***Stereo and rotated views of Figure 4.4a and 4.4b. a-b****, Stereo views of modules A (yellow cylinders and loops) and B (purple cylinders and loops) of the bracelet domain of RdRP in the quiescent (a) and transcribing (b) states.* ***c-d****, Same as in (a-b) but rotated around the X-axis by 90°. Note: All surfaces displayed in this figure were rendered from models.*

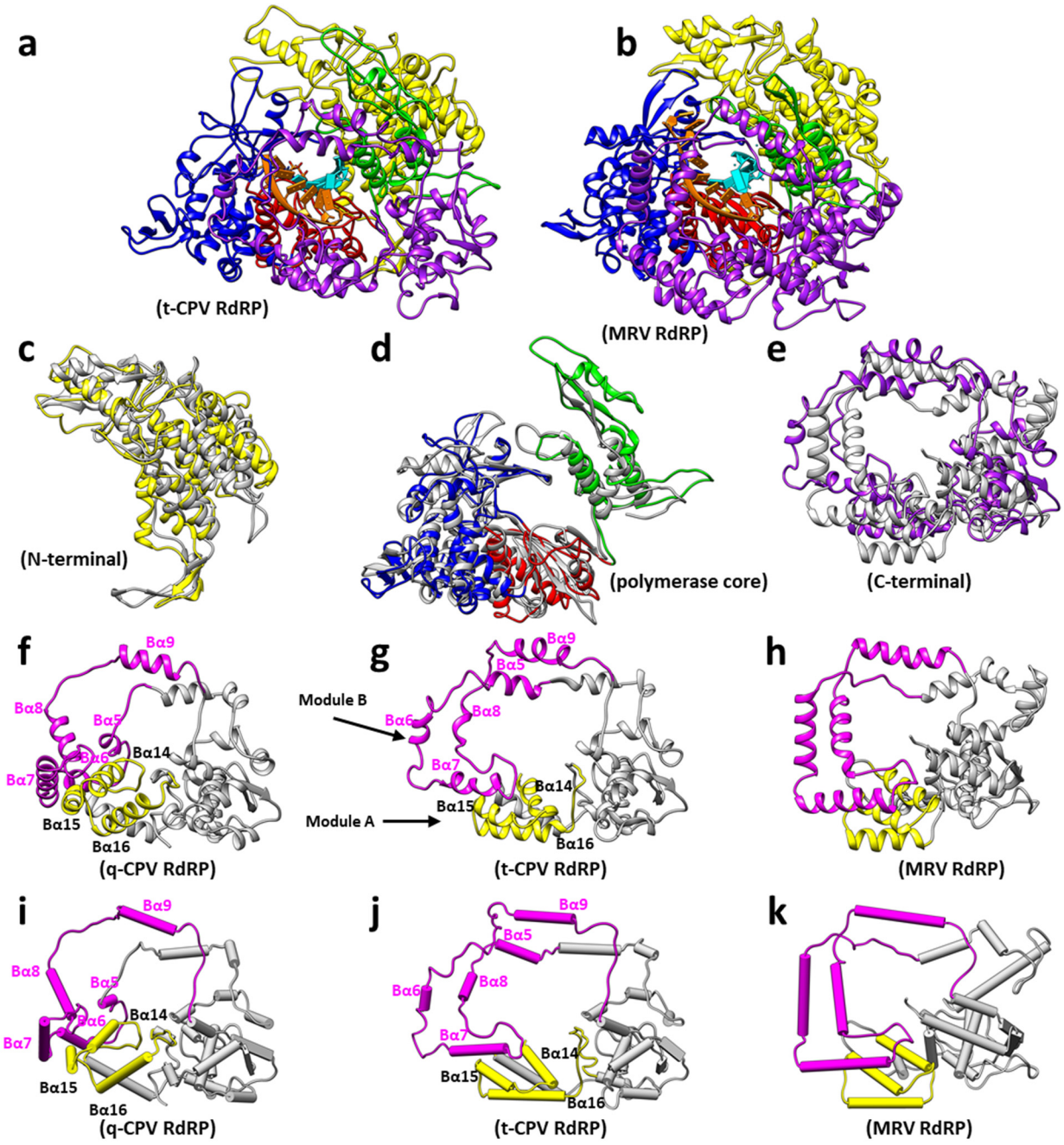


Figure 4.14 ***Comparisons of RdRPs from CPV and MRV. a-b****, CryoEM In situ structure of the RdRP in t-CPV (a) and crystal structure of the MRV RdRP (b), both containing a RNA duplex in the active site.* ***c-e****, Superposition of domains of RdRPs from t-CPV (colour) and MRV (grey): N-terminal (**c**), polymerase (**d**) and bracelet (**e**) domains.* ***f-h****, Comparisons of modules A (yellow) and B (magenta) of the bracelet domain of RdRPs from q-CPV (f) t-CPV (g) and MRV (h).* ***i-k****, The same as in (f-h), but with helices shown as cylinders, as in Figure 4.4 a,b.*

4.6 Acknowledgements

This work was supported in part by grants from the National Institutes of Health (AI094386 and GM071940 to Z.H.Z.), NSFC (31172263 to J.S.) and NSFGD (S2013010016750 to J.S.). We acknowledge the use of instruments at the Electron Imaging Center for Nanomachines supported by UCLA and by instrumentation grants from NIH (1S10RR23057, 1S10OD018111) and NSF (DBI-1338135). We thank Peng Ge for carrying out a *Relion* reconstruction without an initial model as an independent verification step, Stan Schein and Laurie Wang for proof-reading the paper, and Pavel Afonine for model refinement.

4.7 Author Contributions

Z.H.Z. supervised research; X.Z., X.Y. and Z.H.Z. designed and performed the experiments, analyzed and interpreted data and wrote the paper; K.D. wrote programs, analyzed data and prepared figures; W.C. built models; JS prepared reagents. All authors reviewed and finalized the paper.

4.8 Author Information

3D cryoEM density maps have been deposited in the Electron Microscopy Data Bank under the accession numbers EMD-6408 (3.3Å averaged TEC in q-CPV), EMD-6404 (4.0Å averaged TEC in t-CPV), EMD-6407 (3.9Å full q-CPV), EMD-6405 (4.8Å full t-CPV), EMD-6406 (5.1Å asymmetric reconstruction of q-CPV by capsid-subtraction method) and EMD-6409 (filtered 22 Å q-CPV asymmetric reconstruction). The coordinates of atomic models of the TEC in q-CPV and t-CPV have been deposited in the Protein Data Bank under the accession number 3JB6 and 3JB7, respectively. The authors declare no competing financial interests. Correspondence and requests

for materials should be addressed to Z.H.Z. (Hong.Zhou@ucla.edu) and J.S. (cyfz@scau.edu.cn), respectively.

4.9 References

1. Mertens, P.P.C., S. Rao, and Z.H. Zhou, *Cypovirus, Reoviridae*, in *Virus Taxonomy, VIIIth Report of the ICTV*, C.M. Fauquet, et al., Editors. 2004, Elsevier/Academic Press: London. p. 522-533.
2. Tao, Y.Z., et al., *RNA synthesis in a cage - Structural studies of reovirus polymerase lambda 3.* Cell, 2002. **111**(5): p. 733-745.
3. Lu, X., et al., *Mechanism for Coordinated RNA Packaging and Genome Replication by Rotavirus Polymerase VP1.* Structure, 2008. **16**(11): p. 1678-1688.
4. Jiang, W., et al., *Structure of epsilon15 bacteriophage reveals genome organization and DNA packaging/injection apparatus.* Nature, 2006. **439**(7076): p. 612-6.
5. Lander, G.C., et al., *The structure of an infectious P22 virion shows the signal for headful DNA packaging.* Science, 2006. **312**(5781): p. 1791-5.
6. Zhou, Z.H., *Cypovirus*, in *Segmented Double-Stranded RNA Viruses: Structure and Molecular Biology*, J.T. Patton, Editor. 2008, Caister Academic Press: Norfolk, UK. p. 27-43.
7. Furuichi, Y., *"Methylation-coupled" transcription by virus-associated transcriptase of cytoplasmic polyhedrosis virus containing double-stranded RNA.* Nucleic Acids Res, 1974. **1**(6): p. 809-822.
8. Estrozi, L.F., et al., *Location of the dsRNA-Dependent Polymerase, VP1, in Rotavirus Particles.* Journal of Molecular Biology, 2013. **425**(1): p. 124-132.

9. Zhang, X., et al., *Reovirus polymerase lambda 3 localized by cryo-electron microscopy of virions at a resolution of 7.6 angstrom.* Nature Structural Biology, 2003. **10**(12): p. 1011-1018.

10. Nason, E.L., et al., *Interactions between the inner and outer capsids of bluetongue virus.* Journal of Virology, 2004. **78**(15): p. 8059-8067.

11. Xia, Q., et al., *Structural comparisons of empty and full cytoplasmic polyhedrosis virus: protein-RNA interactions and implications for endogenous RNA transcription mechanism.* J Biol Chem, 2003. **278**(2): p. 1094-100.

12. Gouet, P., et al., *The highly ordered double-stranded RNA genome of bluetongue virus revealed by crystallography.* Cell, 1999. **97**(4): p. 481-490.

13. Abels, J.A., et al., *Single-molecule measurements of the persistence length of double-stranded RNA.* Biophysical Journal, 2005. **88**(1): p. 570a-570a.

14. Nibert, M.L. and J.W. Kim, *Conserved sequence motifs for nucleoside triphosphate binding unique to turreted Reoviridae members and coltiviruses.* Journal of Virology, 2004. **78**(10): p. 5528-5530.

15. Zhao, S.L., et al., *Genomic sequence analyses of segments 1 to 6 of Dendrolimus punctatus cytoplasmic polyhedrosis virus.* Arch Virol, 2003. **148**(7): p. 1357-68.

16. Sutton, G., et al., *Bluetongue virus VP4 is an RNA-capping assembly line.* Nature Structural & Molecular Biology, 2007. **14**(5): p. 449-451.

17. Stauber, N., et al., *Bluetongue virus VP6 protein binds ATP and exhibits an RNA-dependent ATPase function and a helicase activity that catalyze the unwinding of double-stranded RNA substrates.* Journal of Virology, 1997. **71**(10): p. 7220-7226.

18. Kim, J., et al., *Orthoreovirus and Aquareovirus core proteins: conserved enzymatic surfaces, but not protein-protein interfaces.* Virus Res, 2004. **101**(1): p. 15-28.

19. Choi, K.H. and M.G. Rossmann, *RNA-dependent RNA polymerases from Flaviviridae.* Curr Opin Struct Biol, 2009. **19**(6): p. 746-51.

20. Yang, C.W., et al., *Cryo-EM structure of a transcribing cypovirus.* Proceedings of the National Academy of Sciences of the United States of America, 2012. **109**(16): p. 6118-6123.

21. Yu, X.K., et al., *A putative ATPase mediates RNA transcription and capping in a dsRNA virus.* Elife, 2015. **4**.

22. Luongo, C.L., et al., *Loss of activities for mRNA synthesis accompanies loss of lambda2 spikes from reovirus cores: an effect of lambda2 on lambda1 shell structure.* Virology, 2002. **296**(1): p. 24-38.

23. Patton, J.T., et al., *Rotavirus RNA polymerase requires the core shell protein to synthesize the double-stranded RNA genome.* Journal of Virology, 1997. **71**(12): p. 9618-9626.

24. Mansell, E.A. and J.T. Patton, *Rotavirus Rna Replication - Vp2, but Not Vp6, Is Necessary for Viral Replicase Activity.* Journal of Virology, 1990. **64**(10): p. 4988-4996.

25. McDonald, S.M. and J.T. Patton, *Rotavirus VP2 Core Shell Regions Critical for Viral Polymerase Activation.* Journal of Virology, 2011. **85**(7): p. 3095-3105.

26. Starnes, M.C. and W.K. Joklik, *Reovirus Protein-Lambda-3 Is a Poly(C)-Dependent Poly(G) Polymerase.* Virology, 1993. **193**(1): p. 356-366.

27. Reinisch, K.M., M. Nibert, and S.C. Harrison, *Structure of the reovirus core at 3.6 angstrom resolution.* Nature, 2000. **404**(6781): p. 960-967.

28. Grimes, J.M., et al., *The atomic structure of the bluetongue virus core.* Nature, 1998. **395**: p. 470-478.

29. Yu, X.K., L. Jin, and Z.H. Zhou, *3.88 angstrom structure of cytoplasmic polyhedrosis virus by cryo-electron microscopy.* Nature, 2008. **453**(7193): p. 415-U73.

30. Pettersen, E.F., et al., *UCSF chimera - A visualization system for exploratory research and analysis.* Journal of Computational Chemistry, 2004. **25**(13): p. 1605-1612.

31. Mindell, J.A. and N. Grigorieff, *Accurate determination of local defocus and specimen tilt in electron microscopy.* Journal of Structural Biology, 2003. **142**(3): p. 334-347.

32. Lyumkis, D., et al., *Likelihood-based classification of cryo-EM images using FREALIGN.* Journal of Structural Biology, 2013. **183**(3): p. 377-388.

33. Scheres, S.H.W., *RELION: Implementation of a Bayesian approach to cryo-EM structure determination.* Journal of Structural Biology, 2012. **180**(3): p. 519-530.

34. Yu, X.K., et al., *Atomic Model of CPV Reveals the Mechanism Used by This Single-Shelled Virus to Economically Carry Out Functions Conserved in Multishelled Reoviruses.* Structure, 2011. **19**(5): p. 652-661.

35. Zhang, X., et al., *Near-atomic resolution using electron cryomicroscopy and single-particle reconstruction.* Proc Natl Acad Sci U S A, 2008. **105**(6): p. 1867-1872.

36. Wolf, M., et al., *Subunit interactions in bovine papillomavirus.* Proc Natl Acad Sci U S A, 2010. **107**(14): p. 6298-303.

37. Adams, P.D., et al., *PHENIX: a comprehensive Python-based system for macromolecular structure solution.* Acta Crystallographica Section D-Biological Crystallography, 2010. **66**: p. 213-221.

38. Huiskonen, J.T., et al., *Structure of a hexameric RNA packaging motor in a viral polymerase complex.* Journal of Structural Biology, 2007. **158**(2): p. 156-164.

39. Briggs, J.A.G., et al., *Classification and three-dimensional reconstruction of unevenly distributed or symmetry mismatched features of icosahedral particles.* Journal of Structural Biology, 2005. **150**(3): p. 332-339.

40. Booy, F.P., et al., *Liquid-crystalline, phage-like packing of encapsidated DNA in herpes simplex virus.* Cell, 1991. **64**(5): p. 1007-15.

41. Zhang, Y., V.A. Kostyuchenko, and M.G. Rossmann, *Structural analysis of viral nucleocapsids by subtraction of partial projections.* Journal of Structural Biology, 2007. **157**(2): p. 356-364.

42. Tao, Y., et al., *Assembly of a tailed bacterial virus and its genome release studied in three dimensions.* Cell, 1998. **95**(3): p. 431-7.

43. Emsley, P. and K. Cowtan, *Coot: model-building tools for molecular graphics.* Acta Crystallogr D Biol Crystallogr, 2004. **60**(Pt 12 Pt 1): p. 2126-32.

44. Zhang, X., et al., *A new topology of the HK97-like fold revealed in Bordetella bacteriophage by cryoEM at 3.5 A resolution.* eLife, 2013. **2**(0): p. e01299.

Chapter 5 *In situ* structures of polymerase complex and RNA genome show how aquareovirus transcription machineries respond to uncoating

Ke Ding[1,2,3], Lisa Nguyen[2,3], Z. Hong Zhou[1,2,3,*]

[1]Department of Bioengineering, University of California, Los Angeles (UCLA), Los Angeles, California 90095, USA

[2]California NanoSystems Institute, UCLA, Los Angeles, California 90095, USA

[3]Department of Microbiology, Immunology and Molecular Genetics, UCLA, Los Angeles, California 90095, USA

*To whom correspondence should be addressed

Phone: 310-983-1033, Fax: 310-206-5365;

E-mail: Hong.Zhou@UCLA.edu

5.1 Abstract

Reoviruses carry out genomic RNA transcription within intact viruses to synthesize plus-sense RNA strands, which are capped prior to their release as mRNA. The *in situ* structures of the transcriptional enzyme complex (TEC) containing the RNA-dependent RNA polymerase (RdRp) and NTPase are known for the single-layered reovirus, cytoplasmic polyhedrosis virus (CPV), but not for multi-layered reoviruses, such as aquareoviruses (ARV), which possess a primed stage that CPV lacks. Consequently, how RNA genome and TEC respond to priming in reoviruses is unknown. Here, we have determined the near-atomic resolution asymmetric structure of ARV at the primed state by cryo electron microscopy (cryoEM), revealing the *in situ* structures of 11 TECs inside each capsid, and their interactions with the 11 surrounding dsRNA genome segments and with the 120 enclosing capsid shell protein (CSP) VP3 subunits. RdRp VP2 and NTPase VP4 associate with each other and with capsid vertices; both bind RNA in multiple locations, including a novel C-terminal domain of VP4. Structural comparison between the primed and quiescent states shows translocation of the dsRNA end from the NTPase to the RdRp during priming. The RNA template channel is open in both states, suggesting that channel-blocking is not a regulating mechanism between these states in ARV. Instead, NTPase's C-terminal domain appears to regulate RNA translocation between quiescent and primed states. Taken together, dsRNA viruses appear to have adapted divergent mechanisms to regulate genome transcription while retaining a similar mechanism to co-assemble their genome segments, TEC, and capsid proteins into infectious virions.

5.2 Introduction

Viruses, divided into 7 classes by the Baltimore Classification system, have various genome replication strategies. RNA viruses (groups III, IV and V) do not rely on host polymerases and instead carry their own RNA-dependent RNA polymerase (RdRp) for genome transcription and replication. However, as RNA regulation is alien to most eukaryotic cells, exposed RNA viral genomes are vulnerable to host antiviral defense mechanisms, such as RIG-I and MDA-5 in humans [1]. While each family of virus has evolved different strategies to avoid host antivirals, reoviruses (belonging to the *Reoviridae* family) remarkably possess the ability to transcribe their own genetic material inside sealed capsids with minimal host involvement, an ability known as endogenous transcription [2]. By incorporating enzymatic functions vital for transcription inside themselves, reoviruses stand transcriptionally self-sufficient. This unique characteristic allows reoviruses to successfully "hide" their genomes from host antivirals, allowing members to infect a wide variety of animal hosts, but it also forces these relatively isolated nano-machines to find very different triggers to convert from the inactive to the infectious state.

Most reoviruses, *e.g.*, aquareovirus (ARV) [3], and bluetongue virus (BTV) [4], conceal their genetic material beneath two or three layers of capsid. The medically significant rotavirus, which causes 215,000 deaths each year [5], is also multilayered. Cytoplasmic polyhedrosis virus (CPV), however, is a unique single-shelled member [6-8]. As the simplest reovirus, CPV has been extensively studied. From it, a basic turreted reovirus infection process has been elucidated as follows: Upon interacting with a host cell, a quiescent virion infects it via endocytosis. Host cell factors such as S-adenosyl methionine (SAM) bind the turret proteins, ultimately remodeling and expanding the virus to make its internal environment more conducive to genome transcription [9]. Transcriptionally active virions use viral RdRp, possibly aided by the viral NTPase, to transcribe

new plus-sense RNA strands, which are capped by the turret proteins prior to expulsion into host cytosol. The capped transcript is translated by host ribosomes to synthesize viral proteins, which are assembled into virions inside cytosolic vesicles (known as inclusion bodies or viral factories). Ultimately, whole progeny virions are released, ready to infect new host cells.

As a turreted reovirus, ARV can be assumed to possess a similar capping mechanism to CPV, but not necessarily a similar overall life cycle. Purified CPV virion is highly infectious [10], but extra treatment is required to achieve high infectivity in ARV virions. This is because CPV lacks an external capsid layer which in ARV consists of penetration protein VP5 and protection protein VP7. ARV VP5 is covered by VP7; removal of this protein allows ARV to transition from a transcriptionally inactive (quiescent) form to a maximally infectious form known as the infectious subvirion particle (ISVP) [11], which uses its newly exposed VP5 penetration proteins to escape the endocytic pathway and invade its host [12]. ARV ISVP, which is primed for yet not actively engaging in genetic transcription, possesses significant surface-level structural differences from quiescent and transcribing CPV [13] and even from transcribing mammalian orthoreovirus (ORV), with which it shares significant sequential homology [14]. That CPV lacks a primed state makes it an inadequate tool for studying its multi-shelled cousins, which may, like ARV, require more complicated structural changes or even protein removal to prepare the virus for transcription.

While most ARV capsid proteins have been resolved to near-atomic resolution by icosahedral reconstruction [12, 15], the structure and location of RdRp and NTPase [together known as the transcribing enzyme complex (TEC)] remain unknown, precluding a full description of the transcription mechanism for multi-shelled reoviruses. Here, we have used cryo electron microscopy (cryoEM) and a novel classification protocol based on our recent asymmetric reconstruction method [13] to obtain asymmetric reconstructions of ARV grass carp reovirus

before and after priming, revealing the *in situ* structures of 11 TECs inside each capsid and their interactions with the 11 surrounding dsRNA genome segments and the 120 enclosing capsid shell protein (CSP) VP3 subunits. Our *de novo* atomic model of NTPase VP4 contains an additional C-terminal domain that both holds the dsRNA end in the quiescent state and translocates it to the RdRp VP2 in the primed state. Our results point to a highly divergent mechanism of genome

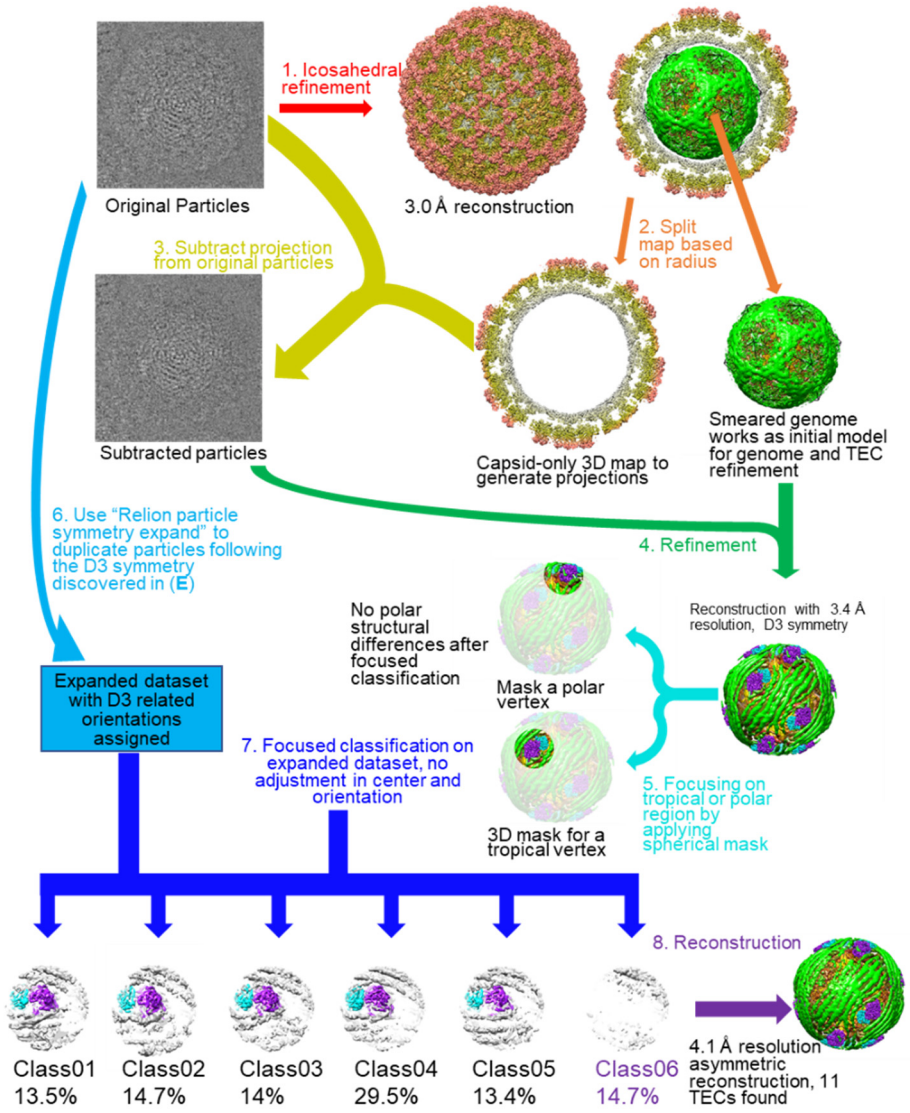


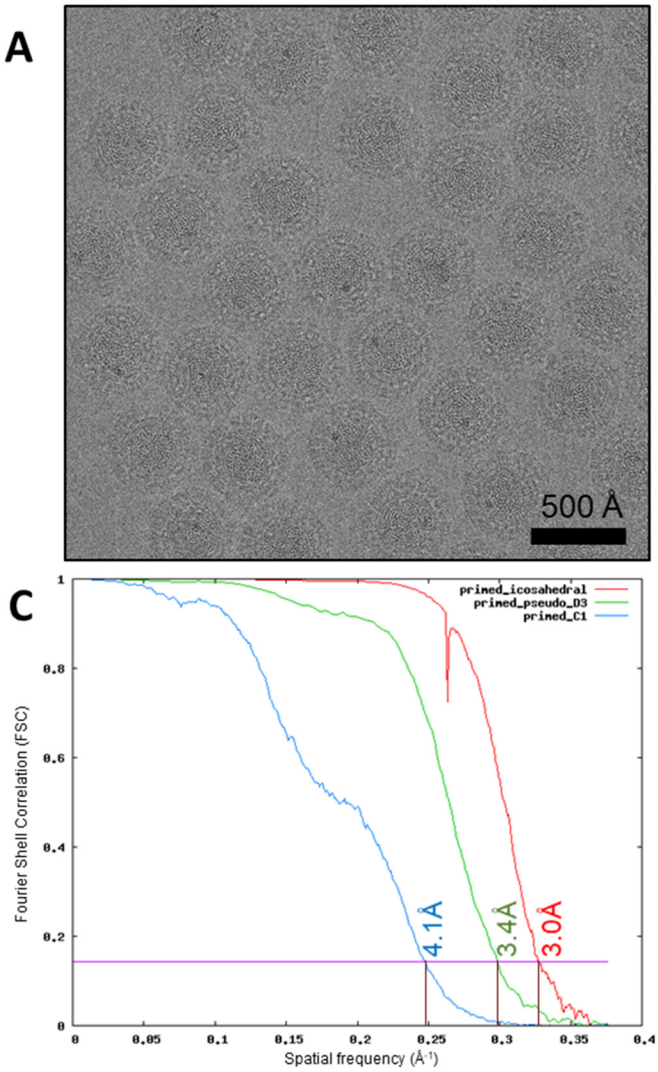
Figure 5.1 ***Asymmetric cryoEM refinement/classification workflow of primed ARV showing ordered genome and 11 associated TECs in each virus.*** *Same-colored arrows represent the same process applying to various data. Black texts describe the properties of the data. Colored texts describe the sequential data processing steps.*

transcription regulation and suggest a conserved co-assembly model among members of the *Reoviridae*.

5.3 Results

5.3.1 11 TECs resolved in the asymmetric reconstruction of the primed ARV.

In order to resolve the structures of the TEC and genome inside primed state ARV, we performed our asymmetric reconstruction by following a new protocol as described in Materials and Methods



Model statistics	RdRp	NTPase
CC around atom	0.804	0.815
Clash score	5.94	6.67
Ramachandran plot favored/allowed/outliers	94.89%/5.04%/ 0.08%	93.97%/6.03%/ 0.00%
Rotamer outliers	0.28%	0.39%
Min nonbonded distance	2.25	2.26

Figure 5.2 ***Raw data, cryoEM reconstruction and model validation****. (A-B) CryoEM micrograph of primed ARV particles (A) and Fourier spectrum (B) of a representative micrograph showing the visibility of Thon rings. (C) Fourier shell correlation (FSC) curves shows the masked icosahedral reconstruction (red), unmasked reconstruction under D3 symmetry (green), and unmasked reconstruction without applying any symmetry (blue). (D) Model quality validation.*

and outlined in Figure 5.1. Grass carp reovirus ISVP was imaged using "super-resolution" electron-counting technology to maximize the image contrast contributed by the internal genome and TEC (Figure 5.2 A and B). Compared to the workflow used to obtain the asymmetric structure of CPV [13], our new procedure contains two additional processes: symmetry expansion and focused classification (*i.e.* processes 6 and 7 in Figure 5.1) under the framework of Relion [16]. Three types of structures were obtained through this procedure: a capsid shell structure with icosahedral symmetry at 3.0 Å resolution, a structure exhibiting D3 symmetry at 3.4 Å resolution, and a genuine asymmetric structure exhibiting pseudo-D3 symmetry at 4.1 Å resolution (Figure

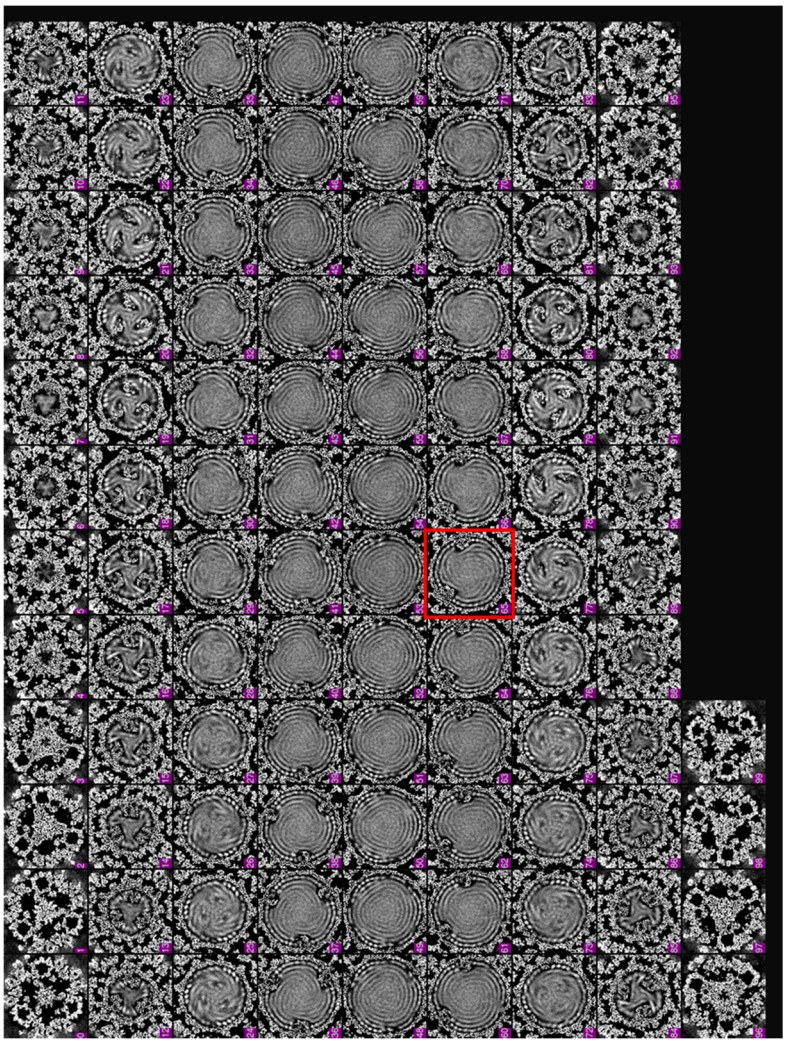


Figure 5.3 ***Density slices perpendicular to the pseudo 3-fold axis of the asymmetric reconstruction of the primed ARV.*** *Note that the pseudo-D3 symmetry breaks at the slice indicated by the red frame (northern tropic).*

5.2 C). Atomic models of TECs were built by imposing D3 symmetry onto density maps, and model statistics are reported (Figure 5.2 D). Our final asymmetric structure of primed state ARV contains 11 TECs, one under each of the virion's 12 vertices, with only the northern tropical vertex lacking a TEC (Figure 5.3, 5.4 A-D). The six TECs on the two poles are related to each other by D3 symmetry and the remaining five TECs (tropical TECs) are related by pseudo-D3 symmetry.

Each TEC is a heterodimer of two protein subunits. The protein closest to the capsid's five-fold axis is RdRp VP2 and the protein further away from the five-fold axis is NTPase VP4 (Figure 5.4 E and F). To facilitate subsequent structural description, we designate the three TEC sides away from the 5-fold axis as front, back and side regions (Figure 5.4F). Our *in situ* TEC structures

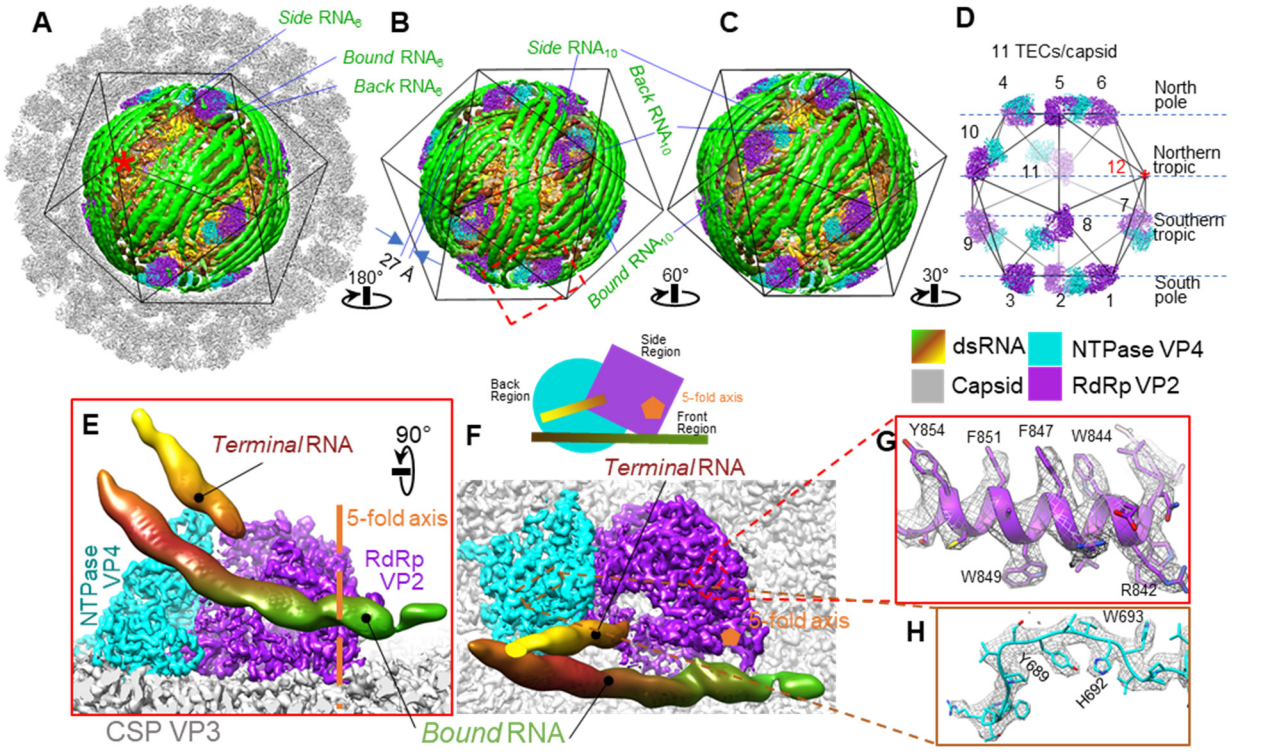


Figure 5.4 ***Asymmetric cryoEM reconstruction of the primed ARV showing ordered genome and 11 associated TECs in each virion.*** *(A-D) Surface representations of the cryoEM density map, showing the full particle (A), the genome and TECs (B,C), and the TECs alone (D). * marks the vertex lacking a TEC. The three segments of RNA under vertex 6 and 10 are labeled as "bound", "back" and "side" RNA. (E-F) The region in the red box of (B) shown in two orthogonal views. For clarity, all but two of the surrounding RNA densities are removed; these are labeled as bound and terminal RNA. The five-fold axis is labeled in orange; a cartoon diagram shows the definition of front, back, and side regions based on TEC's geometry. (G-H) Superposition of our atomic model (color) in the density (grey mesh) extracted from the boxed regions in (F). Color keys for this figure are the same and are indicated under panel D.*

reveal sufficient high-resolution features, such as clearly visible side chains, to support *de novo* atomic modeling for both RdRp and NTPase (Figure 5.4 G and H).

The dsRNA genome is tightly packed inside the capsid surrounding TECs (Figure 5.3, 5.4 A-C). Each dsRNA duplex shows long persistence length and is separated from its neighbors by an average of 27 Å (Figure 5.4B). Several dsRNA strands form stabilizing interactions with each TEC, allowing visualization of their major and minor grooves just as it does for the stem loops of ssRNA viruses [17-19]. Three major dsRNA densities interact with the TEC and are labeled as "*bound*", "*side*" and "*back*" RNA based on their locations relative to the above designated sides of the TEC (Figure 5.4 A-C), with a fourth RNA, labeled "*terminal* RNA", closely approaching the upper region of the TEC (Figure 5.4 E and F). This *terminal* RNA elongates directly towards TEC, differing from other interacting RNAs.

5.3.2 RdRp VP2 has novel ligand interactions and an open template channel.

RdRp VP2 contains 1274 residues organized into three domains: N-terminal domain (a.a. 1-386), core, and C-terminal bracelet (a.a. 902-1274). The core domain is sandwiched between the N-terminal and C-terminal bracelet domains and can be further divided into thumb (a.a. 793-901), fingers (a.a. 387-556 and 595-690), and palm (a.a. 557-594 and 691-792) subdomains, following previously established terminology [20] (Figure 5.5 A-D). The domain arrangement of ARV RdRp follows that of ORV λ3 [21] and CPV RdRp [13, 22, 23]. The fingers and thumb subdomains perform transcript elongation and proofreading whereas the palm catalyzes phosphodiester bond formation between new NTPs and the growing strand via D591, D740, and D741, which are highly

conserved within the *Reoviridae* polymerases (Figure 5.7) [13, 21]. Similar to ORV λ3 [21], ARV RdRp possesses four channels: the template entry, NTP entry, template exit, and transcript exit channels. The template exit channel penetrates the bracelet domain as it does in other reoviruses [13, 24], whereas the NTP entry channel opens near the N-terminal domain [21, 24] (Figure 5.5 C, D, G and H). All four channels intersect at the active site of the palm subdomain, which is

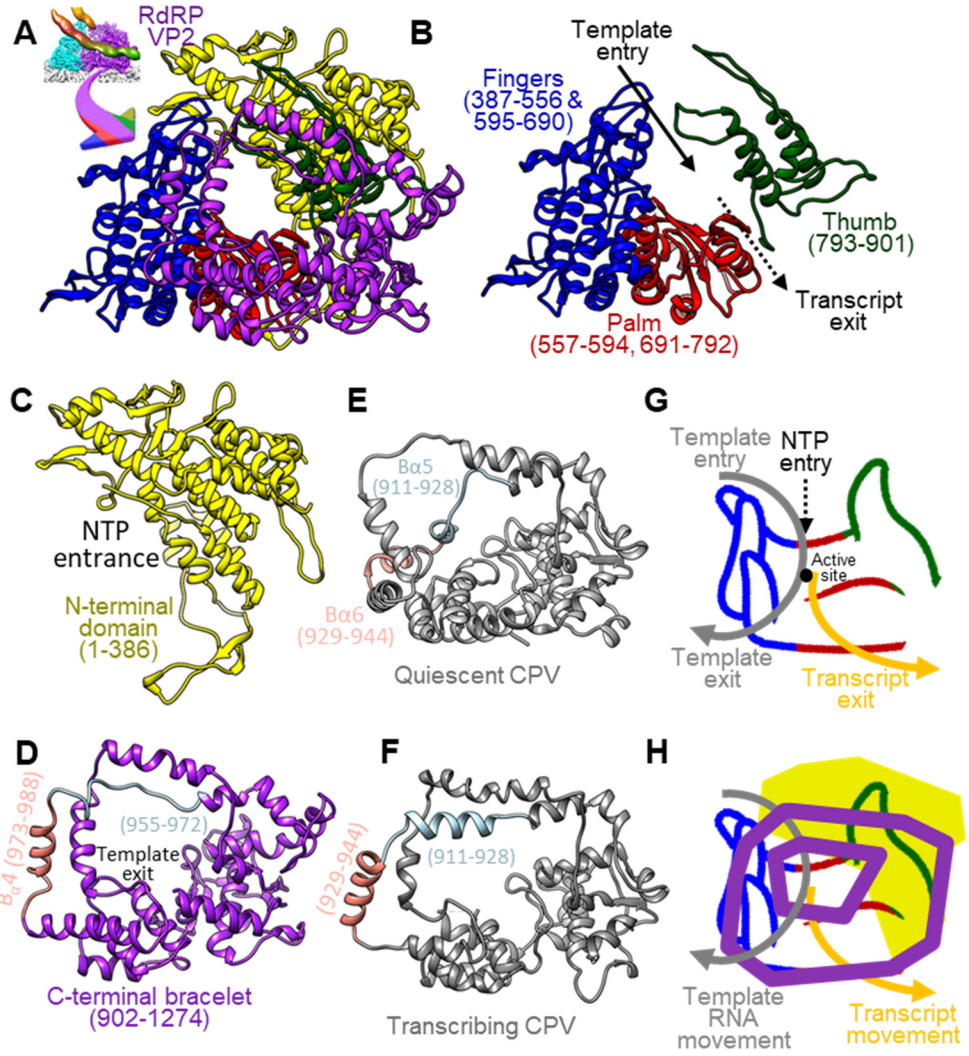


Figure 5.5 ***Structure of ARV RdRp VP2 with bound RNA and comparison of RdRp bracelet domains in ARV and CPV.*** *(A-D) Ribbon diagrams of the atomic model of ARV RdRp VP2 with domains shown all together (A) or separated. The RdRp core domain (B), comprising the fingers, palm, and thumb subdomains, is sandwiched between the N-terminal domain (C) and C-terminal bracelet domain (D). (E-F) The segment blocking the template exit in the CPV RdRp bracelet domain undergoes large conformational changes between the quiescent (E) and transcribing state (F) (42). The equivalent segment in ARV RdRp is similarly colored in (D). (G-H) Cartoons to show the "hand metaphor" of polymerase core structure, with color-coding following (A). Polygons represent the N- and C-terminal domains. Four channels intersecting at the active site are labeled in (G). N-terminal domain and bracelet are displayed on back and front plane, respectively (H).*

unoccupied in our structure of the primed-state ARV RdRp. In notable contrast to the quiescent CPV RdRp, no structures in the primed ARV RdRp bracelet domain occlude the template exit channel or the active site. In the former, these same sites are blocked by the helix-loop structures formed by a.a. 911-928 and a.a. 929-944 respectively (Figure 5.5 D and F). Thus, all four channels in RdRp are open in ARV's primed state and the protein's conformation resembles elongation-state ORV RdRp [21].

5.3.3 NTPase VP4 *has a unique C-terminal domain.*

ARV's NTPase VP4 contains 728 residues arranged into three domains: an N-terminal domain (NTD, a.a. 1-285), an NTPase domain (a.a. 286-602), and a C-terminal domain (CTD, a.a. 603-718; residues 718-728 are flexible) (Figure 5.6 A-D). VP4 contacts the inner capsid at the base of

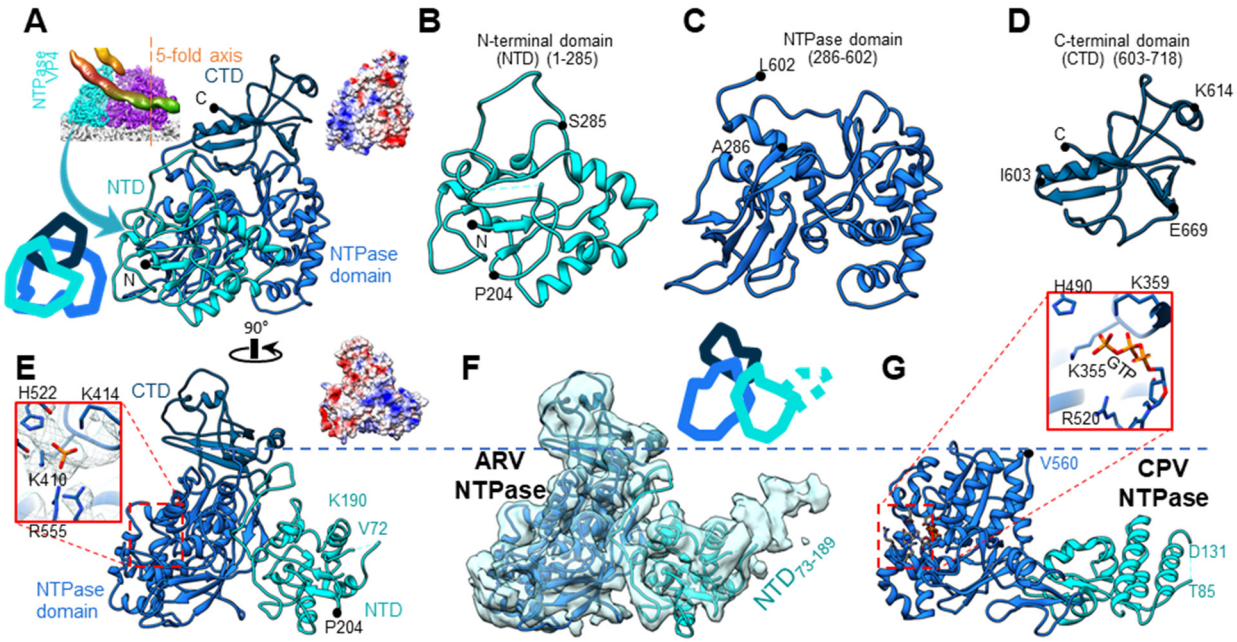


Figure 5.6 ***Structure of NTPase VP4.*** *(A-E) Ribbon diagrams of the atomic model of VP4 shown in two orthogonal views with its three domains in different colors (A and E) or in larger views as separated individual domains (B-D). (F-G) Comparison of NTPases from ARV (F) and CPV (G) (42) demonstrating the existence of an extra CTD (aa 603-718) in ARV. In (F), the atomic model is superimposed with a low-pass filtered density map (semitransparent surface) showing the flexible (and thus unmodeled) NTD_{73-189} subdomain, in contrast to the shorter flexible segment NTD_{86-130} in CPV. The boxed region in (E) contains a phosphate group, whose location is equivalent to the third phosphate in GTP identified in the CPV NTPase at the boxed region in (G). Cartoon diagrams are added to (A) and (F) to show the position of each domain. Surface charge diagrams are added to A and E in the same views.*

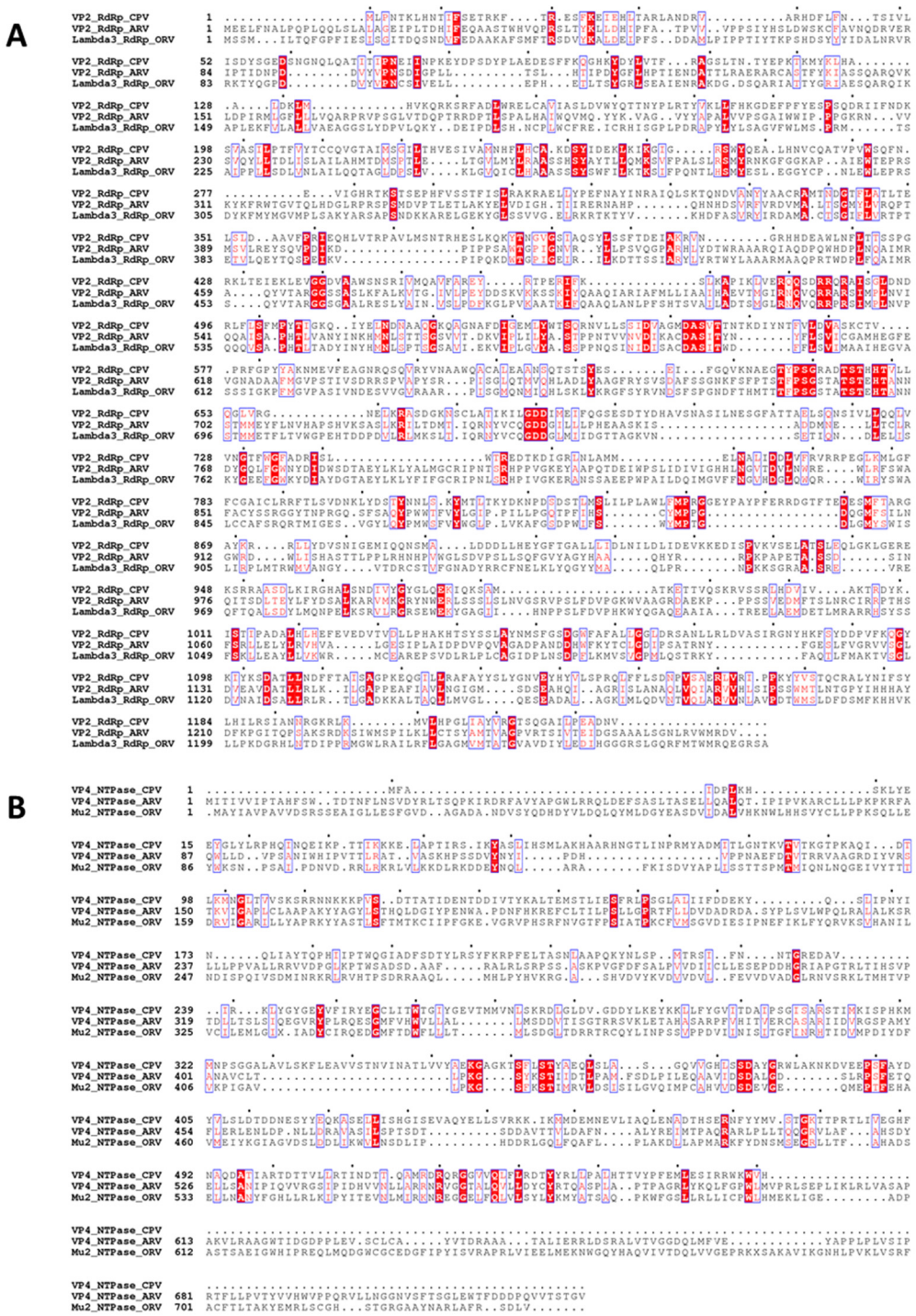


Figure 5.7 ***Sequence alignments of ARV, ORV, and CPV's RdRp (A) and NTPase (B).***

its NTD and NTPase domains, which fold around the end of RdRp opposite to the capsid's five-fold axis.

The surface of VP4's NTD is highly positively charged (Figure 5.6 A, B and E). The domain contains a protruding density that is only visible at low resolution and likely corresponds to its unmodeled residues 73-189, referred to as $NTD_{73\text{-}189}$ subdomain (Figure 5.6F). Within $NTD_{73\text{-}189}$ is a positively charged motif (137-143, RRVAAGR) that potentially interacts with nearby RNA backbones. Our observation that NTPase VP4 interacts with RNA is consistent with a previous study showing that the homologous ORV μ2 is also an RNA-interacting protein [25]. Located at the surface of ARV NTPase VP4 is a conserved residue P204 that may be involved in filamentous inclusion body formation and microtubule-associated virion assembly just as it is in homologous ORV μ2 [14, 26, 27].

Unlike the N-terminal domain, the NTPase domain is mostly negatively charged (Figure 5.6C). Structurally, it resembles the corresponding domain in CPV's NTPase VP4 [13] and binds an additional density at the conserved NTP binding site close to the key residues [K410 and K414 as previously predicted based on homologous ORV μ2 [14]] in the primed state (Figure 5.6E). Thus, we propose that ARV VP4 is also an NTPase. However, this additional density is too small to fit a full NTP molecule and no NTP was added to the PBS buffer to obtain the viral sample. As observed, this small density can potentially be a single phosphate group whose position corresponds to the γ-phosphate group of a bound NTP.

Relative to CPV's NTPase, ARV's NTPase possesses an additional C-terminal domain (CTD, a.a. 603-718) (Figure 5.6 D and F). This domain has a mixed surface charge distribution with two positively charged residues (K614 and R617) forming a positively-charged region at the far surface of VP4, close to the *terminal* RNA (Figure 5.4E, 5.6D, 5.8C). This region is also close to the template entry channel of RdRp, suggesting that the domain may be involved in template RNA regularization between different viral states.

Overall, NTPase VP4's proximity to RdRp VP2 and its affinity for binding cofactors crucial for RNA transcription (e.g. NTP, RNA genome) away from the RdRp suggest that this protein is a key mediator in the polymerase's and therefore the virion's conversion from the transcriptionally inactive (quiescent) to the primed state.

5.3.4 Interactions between RdRp and NTPase.

NTPase VP4 mostly interacts with RdRp through its fingers subdomain (Figure 5.8A), similar to the *in situ* structure of CPV [13]. The shared interface is triangular and occupies a surface area of roughly 1200Å^2 (Figure 5.8B). Although both RdRp and NTPase are highly charged (Figure 5.8C), the two proteins predominantly interact through hydrogen bonds and hydrophobic interactions. We find three strong interactions between these two proteins: (I) hydrogen bonds between the VP4 NTD (near a.a. 57) and the RdRp fingers subdomain (near a.a. 622) (Figure 5.8E), (II) hydrogen bonds between VP4 NTPase domain (a.a. 575-577) and the RdRp fingers subdomain (a.a. 416, 601, 606) (Figure 5.8F), and (III) hydrophobic interactions between the VP4 CTD (a.a. 615, 618)

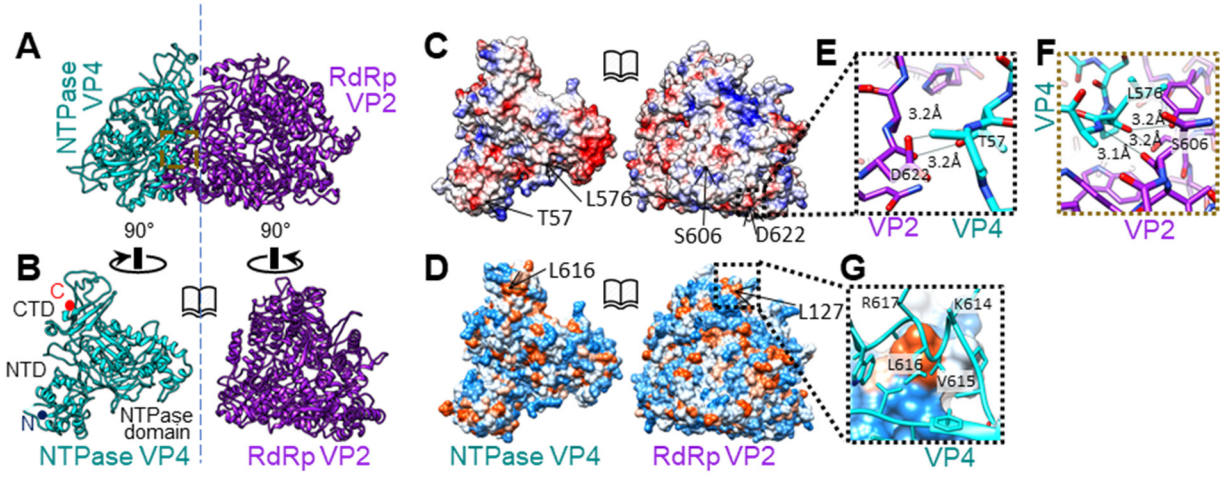


Figure 5.8 ***Interactions between RdRp and NTPase.*** *(A) NTPase interacts with RdRp through RdRp's fingers subdomain, near the template entrance. (B) NTPase and RdRp are rotated 90° in opposite direction to show triangle-shaped interface. (C) Same view as (B), showing positive (blue), neutral (white) and negative (red) coulomb potentials. (D) Same view as (B), showing high hydrophobicity (orange) and high hydrophilicity (dodger blue). (E) NTD-Fingers interaction, region boxed in (C). (F) NTPase-domain-Fingers interaction, region boxed in (A). (G) CTD-N-terminal-domain interaction, region boxed in (D).*

and the RdRp N-terminal domain (near a.a. 127) (Figure 5.8 D and G). The last of these is close to the positively charged residues (K614 and R617) in CTD (Figure 5.8G).

5.3.5 CSP VP3's N-termini anchor TECs and neighboring CSPs.

120 copies of the 1214-residue-long CSP VP3 form the innermost capsid shell with icosahedral symmetry T=2. Ten wedge-shaped VP3 monomers form one vertex in the icosahedron; the five-fold pore at its center lines up with the TEC's RNA exit channel for direct transcript capping and release through turret protein VP1. Icosahedral reconstruction reveals two different conformers of VP3, termed VP3A and VP3B [3, 15]. VP3A conformers form the star that houses the five-fold pore while VP3B conformers wedge themselves between the legs of the star to fill out the pentagonal shape. In our asymmetric reconstruction, we can further differentiate VP3A-VP3B dimers by their position in the vertex relative to the TEC: Looking from the virion center, $VP3_1$ dimer (containing $VP3A_1$ and $VP3B_1$) is mostly positioned behind the TEC, with $VP3_2$ dimer through $VP3_5$ dimer following in a clockwise direction (Figure 5.9A). Under this convention, $VP3_{1-3}$ dimers form the primary seat upon which the TEC is positioned; by contrast, $VP3_4$ dimer and $VP3_5$ dimer make minimal contact with the TEC. Several fragments of VP3 were not resolved in the previous icosahedral reconstruction, most notably the first 187 residues of VP3A and a.a. 501-522 of VP3B [12]. Our asymmetric reconstruction fully resolves VP3B as well as residues 152-191 of VP3A. The latter residues were found to take at least 6 conformations depending on their positions relative to the TEC (Figure 5.9 B and C).

For VP3A, the newly modeled N-terminal residues include an N-terminal helix (N-anchor) (a.a. 152-171) joined to the rest of the molecule by a varying rope-like fragment (N-rope) (a.a. 172-191) (Figure 5.9D). While the N-anchor is folded into an α-helix in all VP3A conformers, the N-rope can fold into a helix ($VP3A_{4-5}$) or partially unfold into extended loops ($VP3A_{1-3}$).

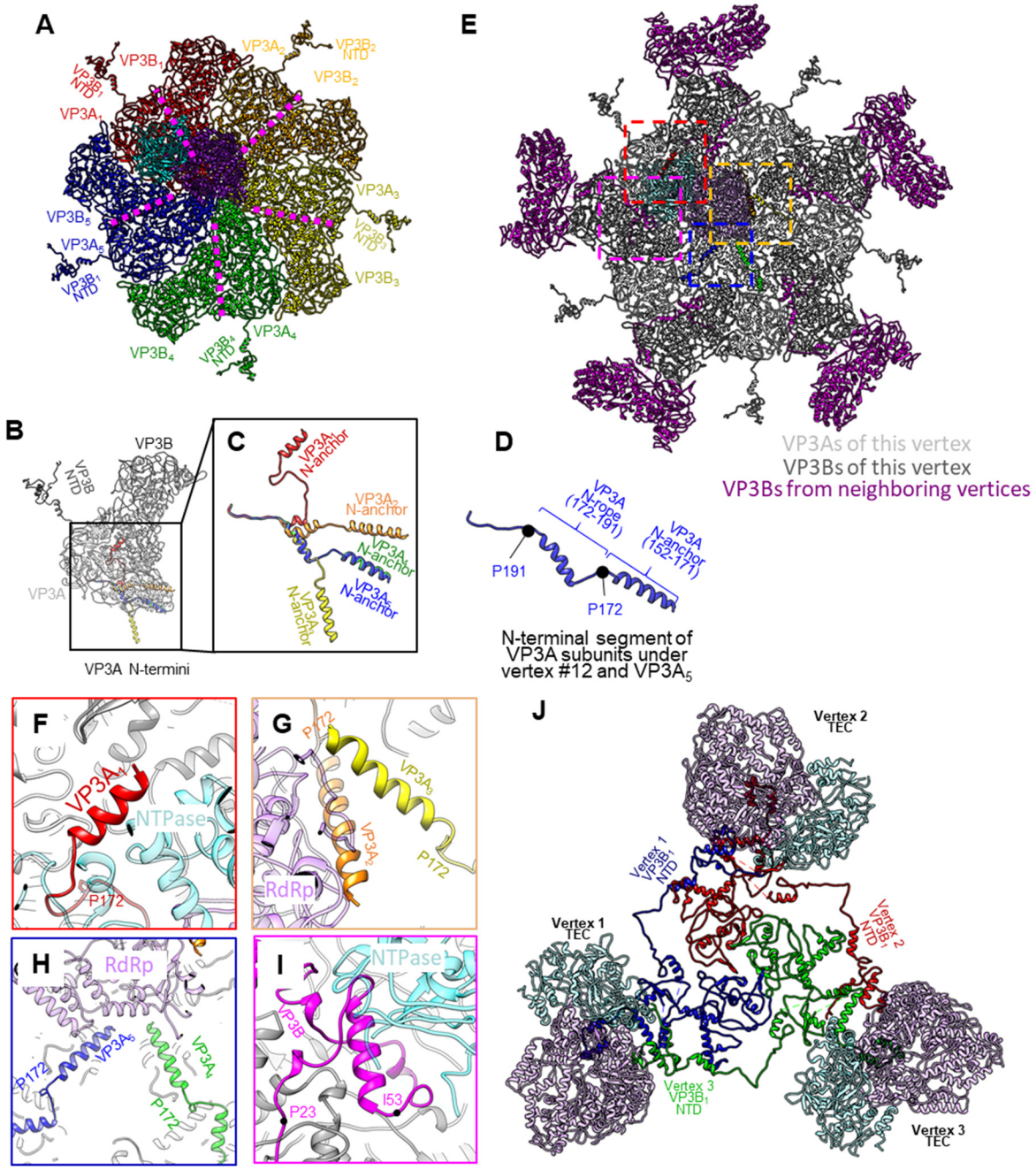


Figure 5.9 ***N-terminal segment of VP3 forms adaptive anchor to interact with TEC.*** *(A) Inside view of a polar vertex showing the position of a TEC relative to five VP3A subunits ($VP3A_{1\text{-}5}$) and five VP3B subunits ($VP3B_{1\text{-}5}$). (B-D) Superposition of the five VP3 dimers from (A), showing dramatic conformational differences among the five subunits, $VP3A_{1\text{-}5}$, in the N-terminal fragment (colored in B and C). In the vertex without TEC (i.e., position 12 in Figure 5.4D), this fragment in the five VP3A subunits has the conformation (D) similar to the $VP3A_5$ subunit (blue in C). (E) The same as (A) with subunits colored as in (B) and five VP3B subunits from neighboring vertices added. (F-I) Zoomed-in views of the regions marked by the four colored boxes in (E). Note that the N-terminal domain of a VP3B of a neighboring vertex also interacts with this TEC (I). (J) Three polar TECs and their associated $VP3B_1$ N-terminal segments (aa 15-518).*

Significant differences in the rope conformations of $VP3A_1$, $VP3A_2$ and $VP3A_3$ allow their

otherwise similar N-anchors to reach and interact with NTPase, RdRp and RdRp respectively (Figure 5.9 B-C, E-G). The N-anchors of $VP3A_{1-2}$ are positioned similarly to the two "N-terminal helices" in the CPV TEC structure, but no N-rope-like features were observed in CPV VP1 [13]. In $VP3A_{4-5,}$ the N-rope is folded into a helix and the N-anchor helix does not interact with the TEC (Figure 5.9 E and H). Notably, the three VP3A subunits whose N-anchors contact the TEC ($VP3A_{1-3}$) are also the ones upon which the TEC is situated. Notably, the newly modeled residues of VP3A subunits at the vertex lacking a TEC (*i.e.*, vertex #12) adopt much simpler conformations. Their N-anchors are structurally identical to those of $VP3A_5$ at the other 11 TEC-associated vertices. This uniform N-anchor conformation creates steric hindrance that prevents the docking of a TEC at this vertex (Figure 5.15 A-C).

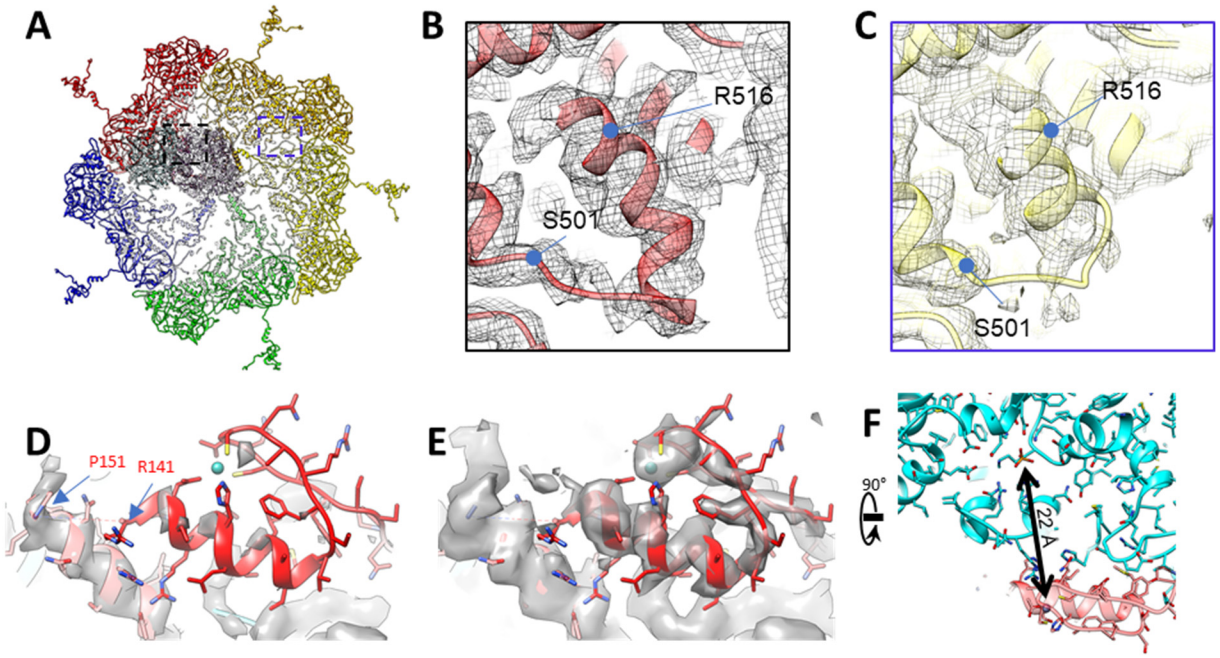


Figure 5.10 ***Asymmetric feature found in primed ARV structure.*** *(A-C) Structural comparisons to show plug-helix conformational change (VP3B 500-525). (A) A polar vertex region. (B-C) Boxed regions from (A), showing one plug-helix region interacting with TEC (B) and one not interacting with TEC (C). The plug-helix interacting with TEC is less flexible. (D and E) Flexible zinc finger structure. Density maps are shown at 2 σ (D) and 1 σ (E), with the "hide dust" function off. The zinc finger at $VP3A_1$ is much weaker than other modeled regions. Location of P151 and R141 is labeled. (F) $VP3A_1$ Zn-finger interacts with the TP-binding motif in VP4. The distance between Zn atom and γ-Phosphate P atom is 22 Å.*

For all VP3A conformers, the remainder of the N-terminus (a.a. 1-147) is relatively flexible and no strong densities can be found in a high-resolution map. However, at a lower display threshold (1σ, Figure 5.10 E), we find a zinc finger structure belonging to $VP3A_1$ (a.a. 115-141) that binds to VP4 with a conformation similar to the zinc finger structure reported in VP3B conformers [15] (Figure 5.14 M and N).

Unlike in VP3A, much more of the N-terminus is modeled in VP3B. Residues 15-190 in all five VP3B subunits forms an extended, largely loop-like structure that inserts itself underneath a flexible loop (a.a. 172-184) of a VP3B in an adjacent vertex and can even interact with the adjacent vertex's TEC (Figure 5.9I). Because the first 200 residues of VP3 generally extend great lengths to bind distant proteins, we rename this region of both VP3 conformers to be the "daisy chain" domain.

Residues 500-525 form the tip of the apical domain of VP3, which in VP3A conformers forms the center of the vertex. In VP3B, these same residues, unmodeled in previous icosahedral reconstructions [12, 15], are now resolved as a loop linked to a short helix. This structure is best resolved in the VP3B conformer interacting with TEC (*i.e.*, $VP3B_1$), where it is wedged between RdRp VP2 and NTPase VP4 (Figure 5.10 B).

5.3.6 TEC's interactions with RNA.

Our *in situ* structure of TEC reveals several RNA binding features. While our asymmetric reconstruction shows that the structure of 11 TECs are the same, the surrounding RNA can adopt different conformations (Figure 5.11 A, B, G, H). Based on the specific residues those RNA strands

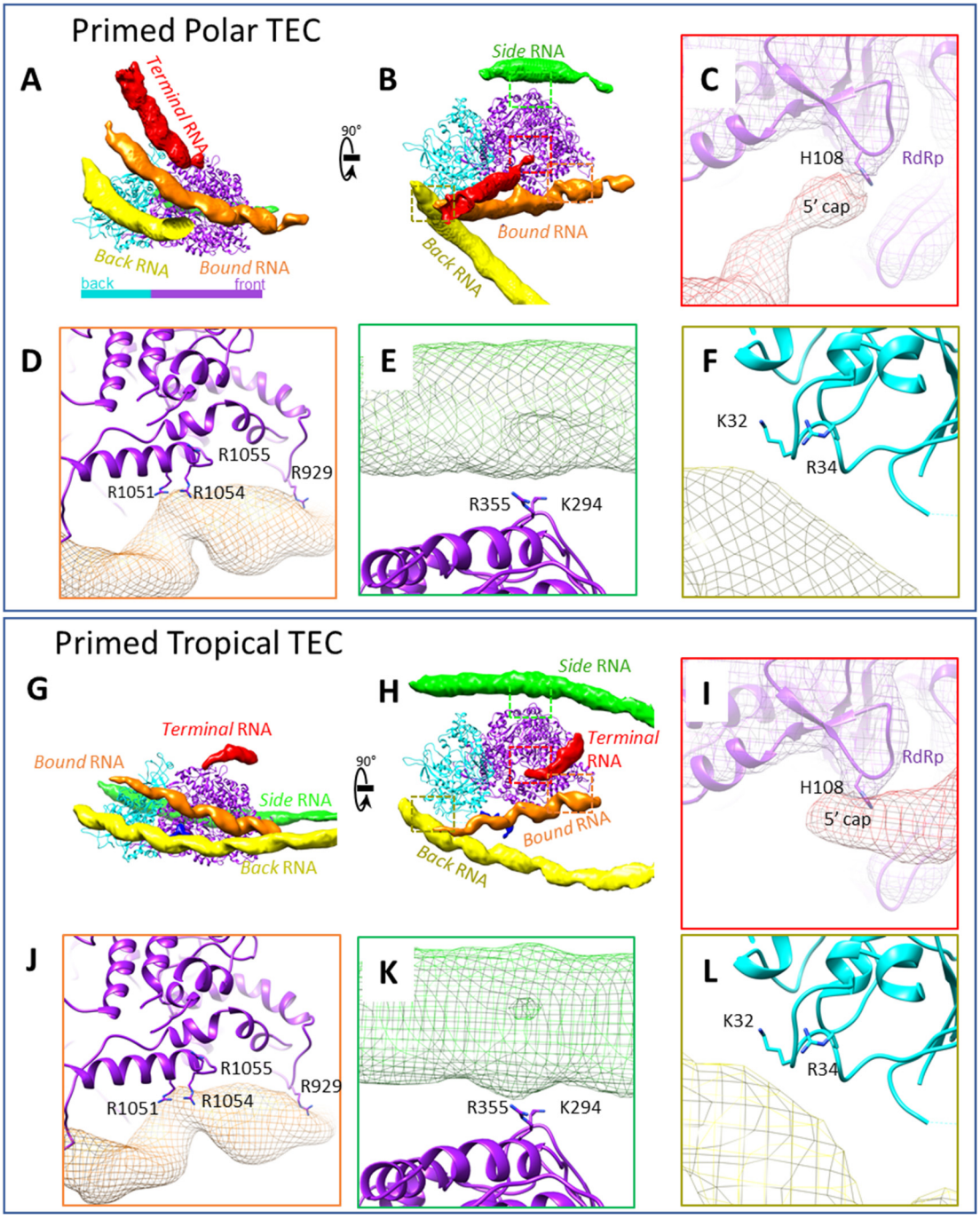


Figure 5.11 ***TEC interacting with RNA.*** *Structural comparison of polar TEC (A-F) and tropical TEC (G-L) in a primed particle. (A-B) Two orthogonal views of polar TEC and four interacting RNA segments: terminal RNA (red), bound RNA (orange), side RNA (green) and back RNA (yellow). (C-F) The boxed regions of (B) showing labeled residues interacting with terminal, bound, side, and back RNA, respectively. (G-L) Same views as (A-F) but on primed tropical TEC. RNA-interacting residues on TEC are conserved between polar TECs and tropical TECs.*

interact with, we can classify them into RdRp-binding RNA (*terminal*, *bound* and *side*) and NTPase-binding RNA (*back*).

For RdRp-binding RNA, we find that the *terminal* RNA 's conformation depends on the position of its bound TEC. Comparison between polar and tropical TECs shows that the terminal RNA approaches RdRp from different directions but share the same skinny geometry and binding site at H108 (Figure 5.11 A, C, G, I). This "thin RNA binding to H108" feature is universal for all 11 TECs (6 polar and 5 tropical). From this, we propose that it is the plus-strand's 5' end we observe binding the TEC in our asymmetric reconstruction. Other RdRp-binding RNA strands are less influenced by the TEC's location. The *bound* RNA is anchored to the RdRp in two locations in the C-terminal bracelet domain. R929 (analogous to CPV's K997) anchors one segment of the *bound* RNA (Figure 5.11 D and J) while another RNA binding region downstream of R929, involving arginine residues 1051, 1054, and 1055, binds another segment (Figure 5.11 D and J). K294 and R355 of the N-terminal domain anchor the *side* RNA (Figure 5.11 E and K).

Back RNA is an NTPase-binding RNA in the primed state which interacts with both K32 and R34 on the positively-charged NTD (Figure 5.11 F and L). This RNA is more curved near polar TECs. The positively charged $NTD_{73\text{-}189}$ subdomain is close to both *back* and *bound* RNA, but no observation reflects that $NTD_{73\text{-}189}$ can anchor RNA in the primed state.

With the exception of R929 (K997 in CPV), these RNA binding features have not been reported in previous asymmetric reconstructions of CPV [13, 22].

5.3.7 Priming introduces changes in RNA binding and protein structure.

The primed ARV described above was produced by removing the outermost protein VP7 from the intact virion [3, 15]. To investigate how the internal components of ARV—the genomic RNA and its associated TECs—respond to this external change, we have also obtained an asymmetric

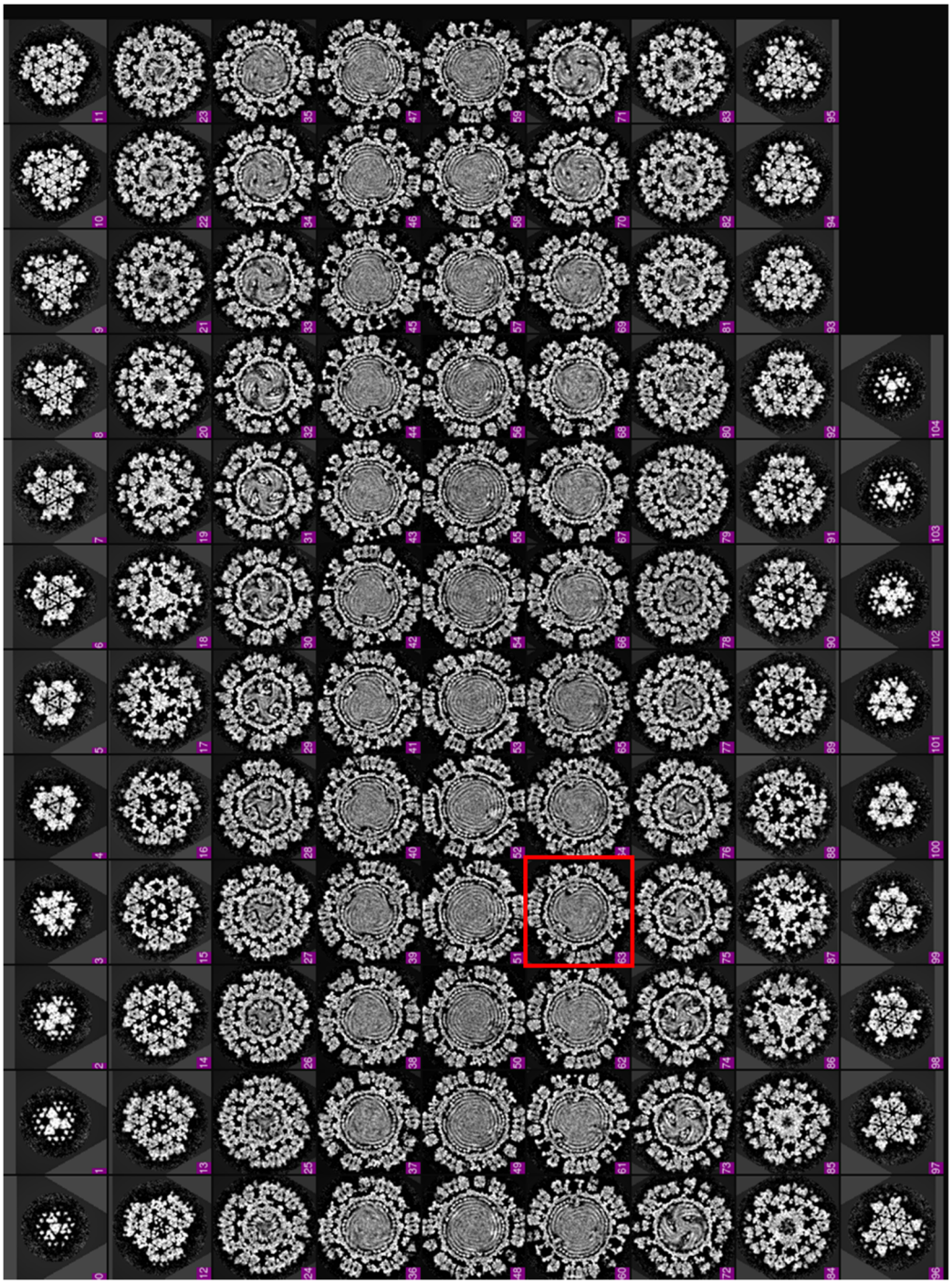


Figure 5.12 ***Density slices perpendicular to the pseudo 3-fold axis of the asymmetric reconstruction of the quiescent ARV.*** *Note that the pseudo-D3 symmetry breaks at the slice indicated by the red frame (northern tropic).*

reconstruction of quiescent-state ARV by reprocessing our old virion images with our new asymmetric reconstruction protocol (Figure 5.12, 5.13 A-C). Externally, the asymmetrically reconstructed quiescent ARV resembles the previous icosahedral reconstruction [15], but

internally, it resembles our new primed-state ARV in that it contains 11 TECs in each capsid, again with the northern tropical vertex lacking a TEC (Figure 5.12). The reported resolution near TEC is 6 Å after averaging all TECs.

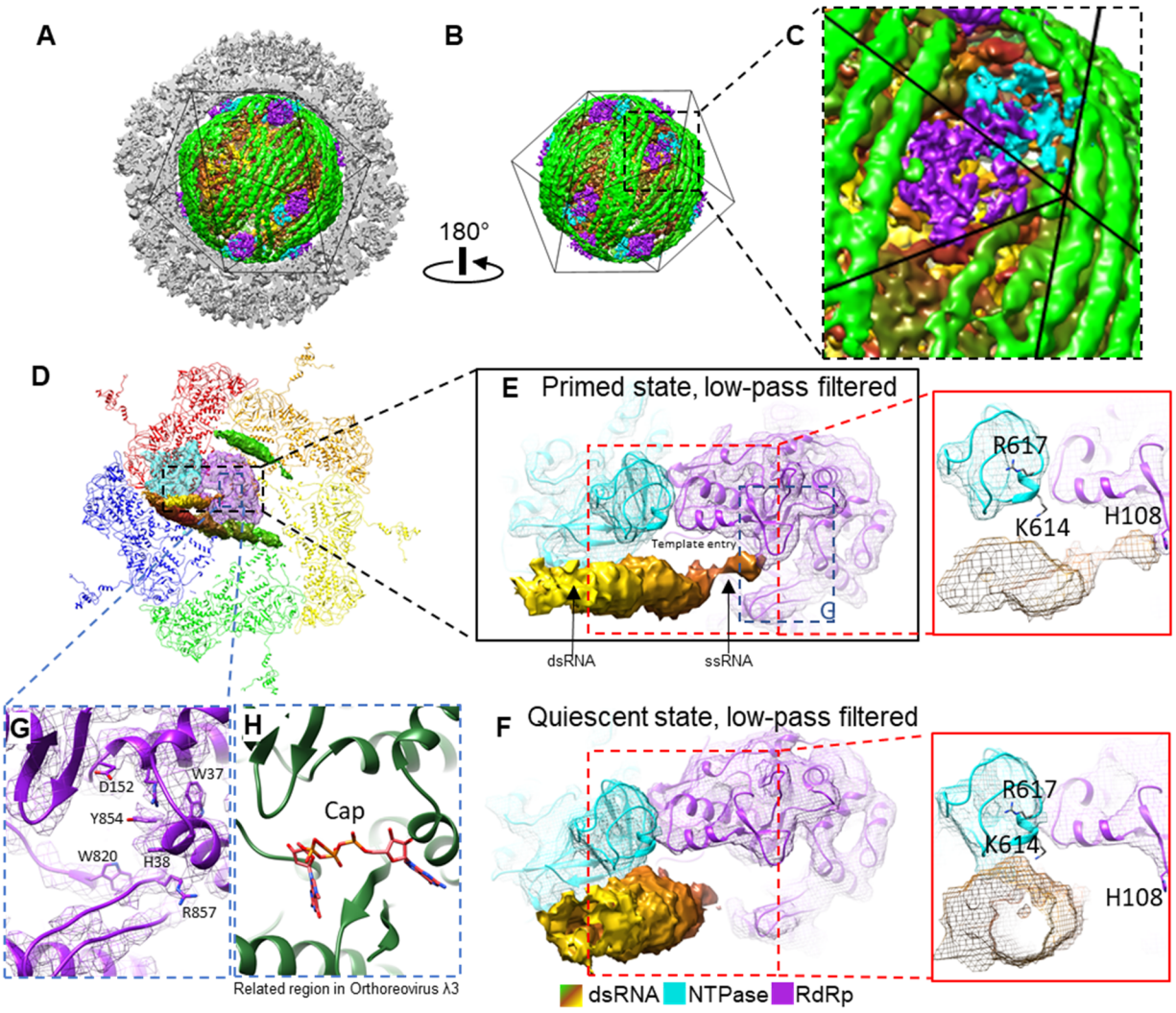


Figure 5.13 ***Reconstruction of quiescent ARV and binding changes in the terminal RNA during priming process.*** *(A-B) Asymmetric reconstruction of quiescent ARV; TEC locations resemble those in primed. (C) Magnified view of the boxed region in B to show a tropical TEC. (D-F) Structure comparison of RNA template entrance in the primed (D and E) and quiescent (F) states. The view in (D) is the same as that of Figure 5.9A, except that the density maps of the terminal, side and bound RNA and the TEC are shown as shaded surfaces and as wireframes superposed with the atomic models, respectively. The densities have been low-pass filtered to facilitate comparison with the quiescent ARV reconstruction. The terminal RNA bound to the CTD of NTPase in the quiescent state (F) detaches from the NTPase to bind H108 of the RdRp in the primed state (E). (G-H) Comparison of the cap-binding sites in primed ARV (G) and in the ORV λ3 crystal structure (PDB 1N1H) (H). Note that in primed ARV, the RNA 5' cap is not bound to the cap-binding site.*

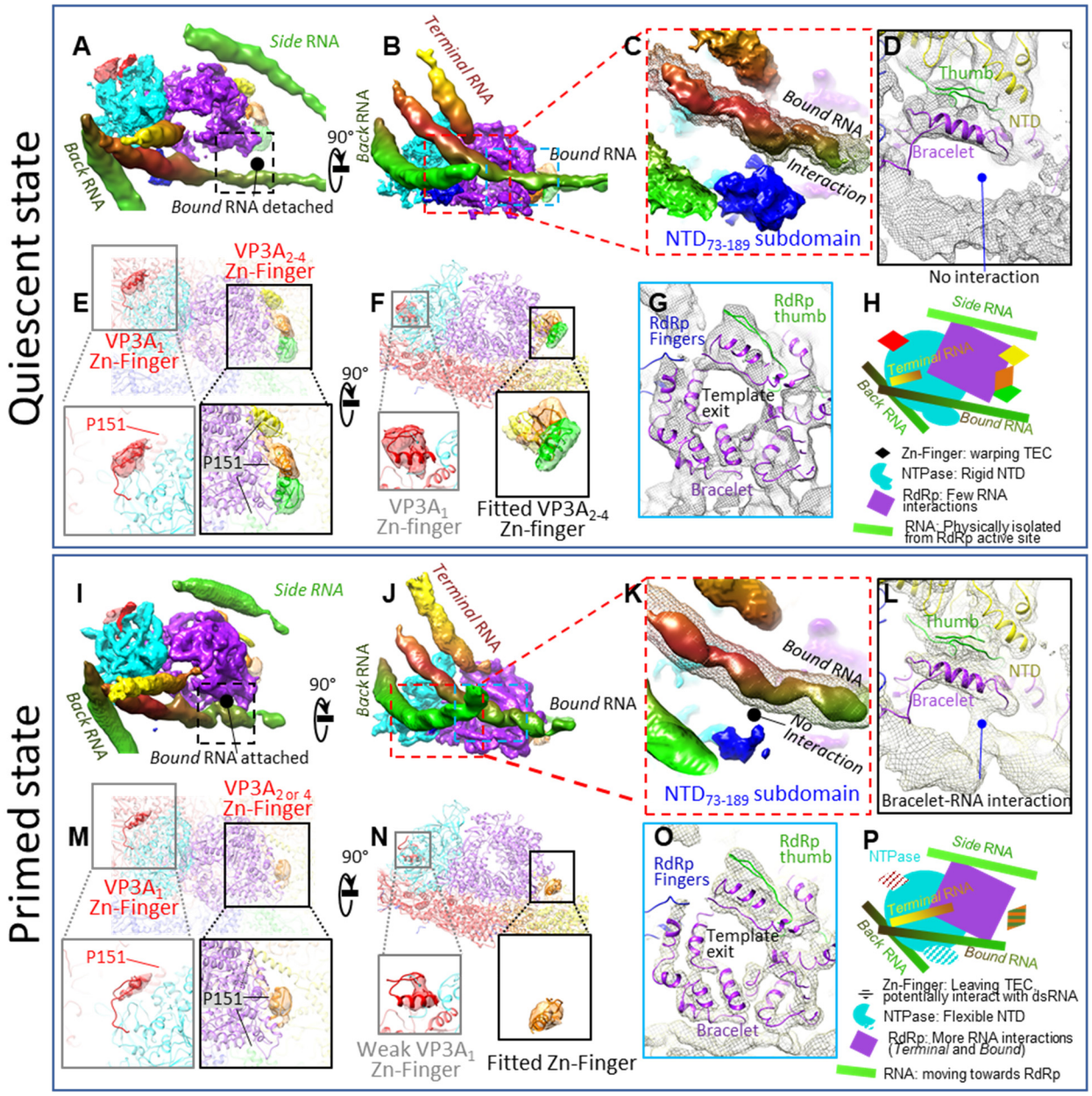


Figure 5.14 ***Comparison between quiescent and primed ARV shows similarities and differences in RNA, TEC and CSP.*** *(A-H) Structural details of a TEC under polar vertex in quiescent state. (A-B) Two orthogonal views show four RNAs interacting with TEC. (C) Zoom-in of (G) showing NTPase NTD_{73-189}'s interaction with bound RNA (mesh). The solid density is at a higher display threshold than the mesh. (D) Zoom-in of black box in (A) shows bound RNA's detachment from RdRp in quiescent state. (E-F) Magnified vertex showing only CSP N-terminal densities, with four structured densities ($VP3A_{1-4}$ Zn-Finger, aa 117-141) labeled. P151 locations in $VP3A_{1-4}$ conformers are labeled to show close proximity between N-anchors and the corresponding Zn-Fingers. (G) Zoom-in of blue box in (B) shows the bracelet domain. (H) Cartoon of the relative positions of RNA and TEC. (I-P) Same view as (A-H) for primed-state structures. Key differences from the quiescent state are: bound RNA attaches to RdRp bracelet domain (A, D, G and L); NTD_{73-189} domain in NTPase becomes more flexible (C and K); and VP3A Zn-Finger becomes more flexible (E, F, M and N). Unlike CPV, no large conformation changes in the bracelet domain can be found (G and O). (P) A cartoon of TEC and RNA structures in the primed state. All map segmentations in this figure are based on the Chimera map segmentation function. No atomic model was used to generate the segmentation.*

The most significant conformational change during the priming process is the movement

of the *terminal* RNA. NTPase's C-terminal domain is unique to ARV, as indicated by comparison with CPV's NTPase (Figure 5.6 E and G). For polar TECs in the quiescent state, K614 and R617 in this domain appear to anchor the *terminal* RNA such that its tip points towards the RNA template entry channel (Figure 5.13F). This dsRNA is released from the NTPase in the primed state and the 5' end of its plus-sense strand extends to bind key residue H108 in the N-terminal domain of RdRp VP2 (Figure 5.13 D and E, Figure 5.11 C and I). The cap density ends at H108, leaving the nearby cap-binding site unoccupied (Figure 5.13 G and H). Although this contrasts previous findings that initiation-state orthoreovirus RdRp λ3's cap-binding site directly binds the plus-sense RNA cap so that the minus-sense strand can enter the polymerase for transcription [21], it is possible that the binding of a single RNA base by H108 is an intermediate step in ARV's priming process. H108 may need to fully anchor the plus-sense ssRNA cap before passing it to the cap-binding site on the thumb subdomain about 10 Å behind it. Although no helicase-related motif can be found near the template entry channel to unwind dsRNA, the 3' end of the minus-sense (template) ssRNA strand is positioned closer to the RdRp active site in the primed state. Thus, the movement and binding of the *terminal* RNA's plus-sense strand's cap seems to facilitate the minus-sense strand's entry into the active site as a template.

The *bound* RNA also undergoes large conformational changes during the priming process. In the quiescent state, the *bound* RNA does not interact with either R929 or the arginine-rich region of the RdRp bracelet domain as it does in the primed state (Figure 5.14 A and D). Instead, it is anchored to the NTPase $NTD_{73\text{-}189}$ subdomain (Figure 5.14C). During priming, the *bound* RNA detaches from the $NTD_{73\text{-}189}$ subdomain and attaches to the RdRp's bracelet domain (Figure 5.14 K and L). Thus, it moves in the same direction as the *terminal* RNA: from NTPase to RdRp.

Inner ARV proteins also change during the transition from quiescent to primed states. The NTPase NTD_{73-189} subdomain is more structured in the quiescent than the primed state (Figure 5.14 C and K); its greater flexibility in the latter state can be attributed to the removal of constraints imposed on it by the *bound* RNA. Quiescent inner capsid proteins, specifically the N-termini zinc finger of VP3A, are also more structured. We can identify four structured regions near $VP3A_{1-4}$'s N-anchor domains, the $VP3A_{1-4}$ Zn-fingers, which like the corresponding N-anchors are strongly associated with NTPase, RdRp, RdRp and RdRp, respectively (Figure 5.14 E and F). Based on the position of these Zn-fingers which bind the TEC in quiescent ARV, we identify another density in the filtered map of primed state ARV (colored orange, Figure 5.14 M and N), which is located near to but not overlapping with the newly identified $VP3A_{2-4}$ Zn-fingers. Based on the distance between P151 and this density, it more likely contains $VP3A_2$'s or $VP3A_4$'s Zn-finger, as $VP3A_3$'s N-anchor is relatively further away. Regardless, the four VP3A Zn-fingers have more defined positions and features in the quiescent state. Greater freedom in the Zn-finger residues in the primed state may correlate with greater flexibility in the neighboring TEC and RNA. This observation suggests that VP3 may help coordinate conformational changes between the outer layer capsid, RNA, and TEC, most likely through directly interacting with all three elements.

Despite significant conformational changes in ARV CSP and NTPase, we do not observe any similarly significant conformational changes in RdRp. In particular, residues 955-988 (roughly equivalent to a.a. 911-928 and a.a. 929-944 in CPV RdRp) of the C-terminal bracelet do not refold to occlude the template exit channel as they do in CPV (Figure 5.14 G and O, Figure 5.5 D-F) [13]. Thus, both domain and template exit channel remain open at all times in ARV, and premature transcription is blocked by a different mechanism, most likely through VP3, VP4, or both proteins in conjunction.

5.4 Discussion

In this study, we found that ARV NTPase VP4 possesses an extra C-terminal domain which is nonexistent in CPV NTPase VP4. Two positively-charged residues in this extra domain anchor the tip of the *terminal* RNA strand outside of RdRp VP2's template entry channel in the quiescent state (Figure 5.13F). This makes the positioning of C-terminal occlusion helices in the active site and in front of the RNA template exit channel, as seen in CPV RdRp, unnecessary in ARV. Both the *bound* and *terminal* RNA strands move from the NTPase to the RdRp when the virus transitions from the quiescent to the primed state. Thus, ARV NTPase may serve to bind RNA segments during viral assembly and prevent premature transcription in the quiescent state, therefore allowing RdRp to focus on its primary function of synthesizing RNA transcript in the transcribing state. From these structural observations, the primed state can be distinguished from the quiescent and transcribing states in that internally stored transcription factors (*e.g.*, genomic dsRNA) are released from their anchoring structures and able to enter but not already entering the viral polymerase. Primed virions may perhaps be prevented from entering the transcribing state by incomplete viral remodeling (*i.e.,* further structural changes are needed to fully expand the capsid to allow sufficient space for transcription) or insufficient externally supplied cofactors (e.g. host NTP and SAM).

Because little external (host-provided) energy is needed to trigger the primed state, transitioning from the quiescent to the primed state must represent a natural transition from a higher- to a lower-energy state. Thus, the quiescent state in ARV is metastable. Similar local energy minimum states achieved prior to cell entry have been reported in many other viruses, such as poliovirus [28] and influenza virus [29]. In order to transition out of this local minimum into the lower-energy primed state and thus "prime" the virion, the energy barrier around the quiescent state must be overcome. We believe the observed increased flexibility in viral proteins in the primed state can facilitate this

process. Greater flexibility in RNA-contacting proteins allows greater freedom of movement in the RNA, which can potentially disturb short-range attractive interactions between positively charged residues and the RNA backbone. This can explain *terminal* RNA's detachment from K614 and R617 in polar VP4 (Figure 5.13 E and F). Movements introduced by protein flexibility can also allow for the establishment of new interactions between positively-charged residues and RNA backbone, such as that observed between R1051, R1054 and R1055 and the RNA genome in the primed state (Figure 5.11 D and J, Figure 5.14 D and L). Conformational changes in the RNA genome were not addressed in previous studies, as conformational changes typically refer to proteins. Here, however, given the more significant observed RNA conformational changes relative to observed protein conformations, it seems that few new protein-protein interactions are formed in the primed state and that proteins are unlikely to power these large RNA changes. It is more likely that the flexible protein structure releases RNA so that it can extend to occupy locations further away from viral center (backbone charge-driven) and/or to establish new interactions with other viral proteins (interaction-driven). This mechanism is more efficient when backbone charges are maximized (stronger repulsion) and the genome has long persistence lengths (rigid genome). Both factors appear more strongly in viruses with a dsRNA genome. While we observe at best minimal conformational changes in the RdRp and the CTD of VP4, we find that the *terminal* RNA does extend towards RdRp and interact with H108 in the primed state of ARV. This observation strongly supports the assumption that energy related to RNA can contribute to gene translocation and state transition. A similar theory has long been accepted for dsDNA viruses such as phages [30].

The Zn-fingers of VP3 may be multifunctional. In addition to stabilizing the capsid in VP3B conformers, they may also play a role in genome regulation because each finger in $VP3A_1$-

$_4$ shows greater flexibility in the primed state. A 10-residue-long linker (a.a. 141-151) separates the N-anchor and the Zn-finger; thus, the detached Zn-finger can potentially interact with surrounding dsRNA. That VP3A$_1$'s Zn-finger density is visible and well-defined but weaker than the other portions of VP3A$_1$ suggest that this zinc finger is not bound to every VP4 and can potentially switch between TEC-binding and TEC-detaching states. The spatial proximity of the VP4 NTPase motif and the VP3A$_1$ Zn-finger suggests that the latter's detachment may potentially change the activity of the former (Figure 5.10 F). Zn-finger detachment in VP3A$_{2\text{-}4}$ can also change RdRp activity through new interactions with RNA. These Zn-fingers, which in ARV contact all other major elements of the virion, are highly conserved in the *Reoviridae* family [31] [32] and can, in conjunction with the rest of the VP3 daisy-chain domain, work as "sensors" to "inform" the TEC of the state of the dsRNA genome and the external environment.

Structural variation among members of the *Reoviridae* appears to be the rule rather than the exception—as one might expect from RNA viruses, given their more rapid mutation and evolution rate relative to DNA viruses. Some members, such as rotavirus and BTV, lack an NTPase protein in the TEC [33], whereas others, such as CPV and ARV, have it. However, the number of domains each member's NTPase contains also varies, with the current study revealing that ARV's NTPase VP4 contains a new domain that binds to genomic RNA which CPV's NTPase lacks. The specific locations of the NTPase's various functions can also vary. Sequence alignment between ORV's μ2 and ARV's VP4 shows that although they have 26% identity overall, only the NTPase domain is highly conserved (Figure 5.7 B). In ARV VP4's CTD, the two key residues for RNA association are K614 and R617. No positively charged residues can be found in the corresponding position of ORV μ2; however, ORV μ2 contains a significant peptide insertion (PRKXSAKAVIKG) in a location equivalent to ARV VP4's E669 (labeled in Figure 5.6D), which

is in turn located close to the fingers subdomain and template entry of RdRp. In ORV, this highly charged peptide insertion may serve to anchor a dsRNA end similar to K614 and R617 in ARV VP4. This variation between ORV and ARV shows that even close evolutionary relatives can have

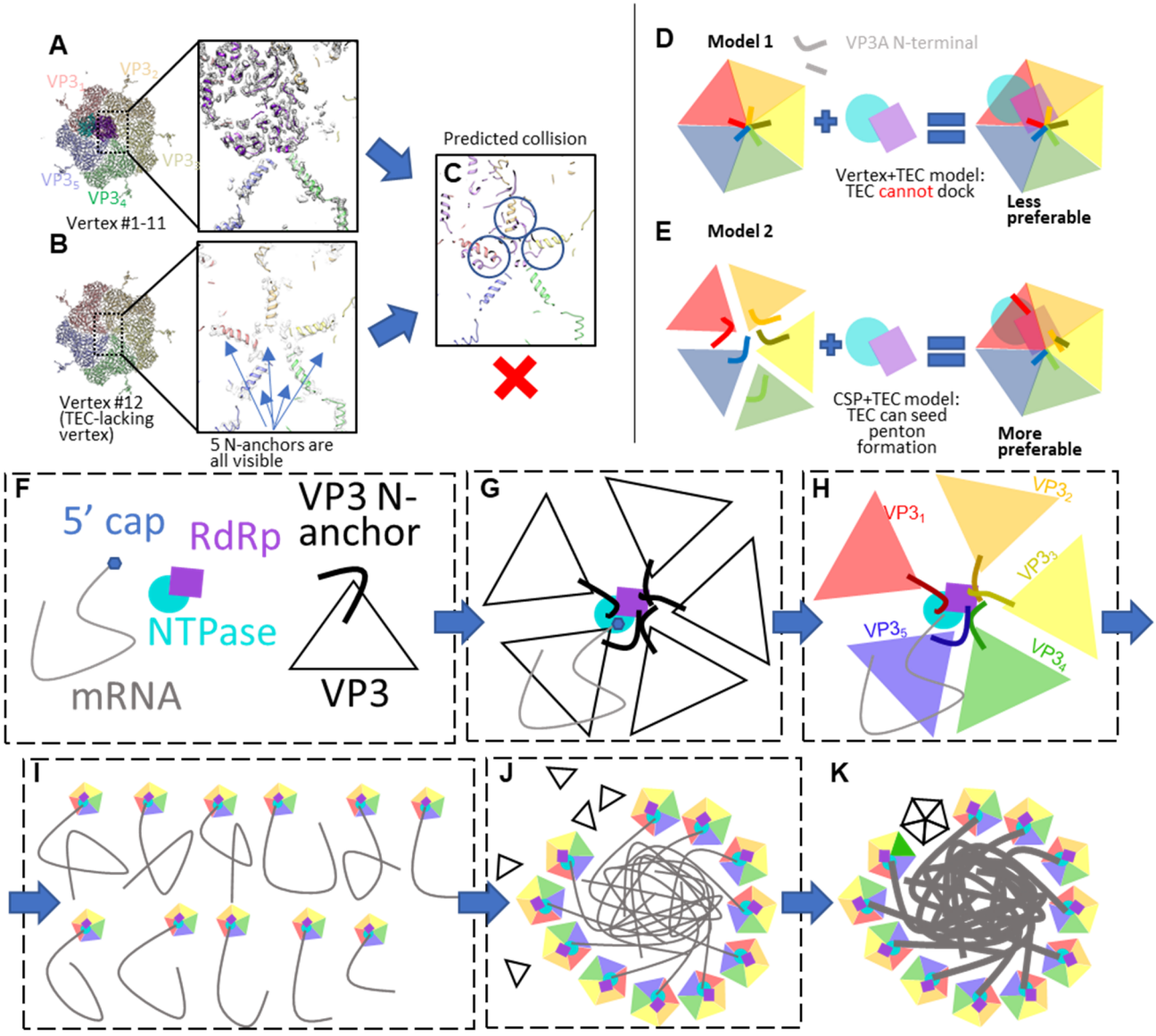


Figure 5.15 ***TEC-lacking vertex #12 and implication to assembly mechanism.*** *Density map and models at vertex with (A) and without (B) a TEC. Note that in icosahedral reconstruction, no N-anchors of VP3A were visible due to averaging. (C) Direct model superposition of TEC in (A) to five VP3As in (B) showing TEC's collision with N-anchors as circled. (D and E) Two proposed models of assembly. In Model 1 (D): vertexes assemble first, followed by attachment of a TEC. TEC would collide with assembled N-anchors. In Model 2 (E): TEC seeds VP3A assembly where VP3As' N-anchors can adapt the observed asymmetric pattern. (F-K) "Wool ball" model for virus assembly. (F) Building blocks of ARV assembly. TECs interact with mRNA and VP3, drawing the latter to their ideal positions (G) leading to properly folded N-anchors around the TEC (H). The 11 genomic RNA segments, each with a vertex (I), coalesce to form a "wool ball", bringing VP3s in close vicinity to accelerate assembly (J). Vertex #12 seals the capsid while RdRp replicates complementary strand from the plus-sense mRNA template in a confined environment (K).*

distinct NTPase structures and differing ways of regulating RNA. This variation holds true even within a single virion: the RNA regulation of polar TECs and tropical TECs are different, especially for *terminal* RNA (Figure 5.11 A, C, G, I).

The numerous TEC interactions with CSP VP3 N-terminal segments and RNA provide evidence for a novel co-assembly model of dsRNA viruses. For many dsRNA viruses, expression of CSP protein alone leads to the formation of core-like particles (CLPs) [34-36], indicating that the constraints to form a capsid with the proper size and shape are fully encoded within CSP-CSP interactions. This notion is consistent with our observation of a TEC-lacking vertex in ARV virions (Figure 5.15B). However, in the presence of TEC, CSP binds to TEC with higher affinity than to itself, as co-expression of polymerase and other capsid proteins leads to the formation of CLPs with TEC under all capsid vertices [34, 35]. Most importantly, viral particles in native conditions [26, 37] or directly purified from infected cells typically contain RNA genomes, but empty particles can be obtained via special treatments, presumably due to loss of the genome through broken vertices [38]. These observations can now be explained by the extensive interactions each TEC has with numerous CSPs [*e.g.*, the extended N-terminal daisy-chains of three VP3A and two VP3B molecules (Figure 5.9 E-J), the inner surfaces of two VP3A and one VP3B molecules (Figure 5.9A)], and the genomic RNA (Figures 5.4, 5.11, 5.13, 5.14). As the number of dsRNA segments of ARV matches the number (*i.e.*, 11) of observed TECs inside each virion (Figure 5.4 A-D)—just as it does in CPV (*i.e.*, 10 segments and 10 TECs) [13]—it is likely that each TEC associates with one and only one segment of the dsRNA genome, perhaps by recognizing *terminal* RNA (Figure 5.15 F-I). These genome-TEC-CSP interactions bring multiple copies of CSP in close vicinity to significantly increase their local concentration in the cytoplasm, improving their chances of finding their interacting interfaces and consequently increasing the yield of infectious

virions containing both TEC and genome (Figure 5.15E). This mechanism is preferable in the assembly process because an independently assembled CSP penton's flexible N-termini (*e.g.*, residues 1-151 residues) can create steric hindrance for the attachment of a TEC (Figure 5.15D). In such a co-assembly model, each TEC binds one genome segment, and interactions among the 11 segments lead to the formation of a "wool ball" (Figure 5.15 I and J). Tethered to the TECs in this wool ball through their N-anchors, CSPs are brought together in approximately correct quantities and distances to coalesce snugly around the genome. As the ability to form a capsid vertex is entirely encoded within the CSP, the formation of a vertex bereft of TEC (*i.e.*, vertex #12 in Figure 5.15B) around the wool ball will happen naturally, ultimately leading to the formation of a complete infectious virion (Figure 5.15K).

5.5 Materials and methods

Sample Preparation and cryoEM Imaging. Primed grass carp reovirus (GCRV, an ARV) was prepared as previously described [12]. Each 2.5 μL sample of purified quiescent and primed GCRV virions was applied to a 400-mesh Quantifoil grid (1.2/1.3 μm) and vitrified in a Vitrobot Mk IV (Thermo Fisher Scientific) after blotting (blot time 7 seconds, 100% humidity and 22ºC). CryoEM samples were imaged in a Titan Krios electron microscope (Thermo Fisher Scientific) operated at 300kV acceleration voltage and recorded on a Quantum LS energy filtering direct electron detecting camera (Gatan Inc.) operated in super-resolution mode with a 20eV slit width around the zero-loss peak. Data collection was facilitated by the Leginion automatic microscopy package [39]. The total dosage was 56e/Å2 and the dosage used for refinement and reconstruction was 25/Å2. The targeted defocus was 1.5μm-2.5 μm. The calibrated pixel size was 1.33 Å/pixel. In total, 5810 movies were collected.

Single Particle Reconstructions. ISVP particles were selected from aligned and averaged movies by UCSF MotionCorr [40]. 94,412 ISVP particles were initially selected. The particles were extracted and initially subjected to icosahedral refinement with RELION [16], resulting in a 2.9 Å resolution map. A symmetry relaxation method as previously described [13] was applied to the refined dataset to obtain a 3.4 Å resolution map of TECs. However, this method failed to resolve the true asymmetric structure (11 TECs). Consequently, the refinement was trapped in a local minimum with pseudo-D3-symmetry, restricting our ability to classify vertices to only 2 groups—6 polar and 6 tropical—in CPV. To precisely locate the unoccupied vertex, a local spherical mask was applied underneath different vertices to focus on areas of interest under vertices (namely, TECs under vertices) and minimize the strong noise introduced by the dsRNA genome in the liquid crystalline state. Each particle was expanded with 6 duplicates related to D3 symmetry, and exhaustive classification was conducted on those duplicated particles with symmetry-related orientations. We found that focusing classification on the northern tropical vertex resolved a class showing a vertex without a TEC, whereas all other tropical vertices were occupied with identical TECs. The above method allowed us to reconstruct GCRV ISVPs in the primed state with 11 TECs.

The structure for ARV in its quiescent state was determined by applying the same method to a previously published GCRV virion dataset [15]. The resolution near the reconstructed TEC is 6 Å.

Atomic Model Building and Refinement. RdRp VP2, capsid shell protein (CSP) VP3, and NTPase VP4 were modeled in Coot [41]. VP2 modeling was guided by a homology model generated by Phyre2 [42] based on ORV λ3 (PDB: 1N1H). VP3 was modeled by combining its previous atomic structure as determined by icosahedral reconstruction [13] with novel asymmetric regions built de novo using Coot. By docking the previously published model of CSP VP3 (PDB: 3IYL, chains X and Y) into our density map and comparing modeled residues with the full protein

sequence, we pinpointed areas of unmodeled density that could be attributed to the missing regions of this protein. These unmodeled regions were filled using the Baton Build function in Coot and combined with the previous model. VP4, a new protein, was modeled entirely de novo using Baton Build. A string of poly-alanines was placed in the density map using Cα bumps and landmark residues (such as Tyr and Trp) as a visual guide; this poly-alanine chain was subsequently mutated to the proper sequence using Coot's Mutate Residue Range tool prior to local (Coot) and global (PHENIX) refinement. Poorly-fitting residues were manually adjusted using the Rotate/Translate, Rotamer Fit, and Real Space Refine Zone tools. Initial models were further refined in PHENIX [43]. The final quality of our models was evaluated based on model geometry, EMRinger score, fit to the density map, and agreement with the Ramachandran plot. Figures were generated using Chimera [44].

5.6 Acknowledgements

This work was supported in part by grants from the National Institutes of Health (AI094386, GM071940 and DE025567). We acknowledge the use of instruments at the Electron Imaging Center for Nanomachines supported by UCLA and by instrumentation grants from NIH (1S10RR23057, 1S10OD018111 and U24GM116792) and NSF (DBI-1338135 and DMR-1548924). K.D. was supported by Whitcome Fellowship. We thank Peng Ge for assistance in atomic modeling of the NTPase, for stimulating discussions during functional interpretation, and for proofreading the paper. We also thank former members of the Zhou group, including Xing Zhang, Qin Fang and Lingpeng Cheng, for their contributions while working on the ARV project at UCLA.

5.7 References

1. Barral, P.M., et al., *Functions of the cytoplasmic RNA sensors RIG-I and MDA-5: Key regulators of innate immunity.* Pharmacology & Therapeutics, 2009. **124**(2): p. 219-234.
2. Estes, M.K., *Rotaviruses and their replication.* Virology, 1990: p. 1353-1404.
3. Cheng, L.P., et al., *Subnanometer-resolution structures of the grass carp reovirus core and virion.* Journal of Molecular Biology, 2008. **382**(1): p. 213-222.
4. Roy, P., *Orbiviruses*, in *Fields Virology*, D.M. Knipe, et al., Editors. 2007, Lippincott Williams & Wilkins: Philadelphia. p. 1976-2000.
5. Tate, J.E., et al., *Global, Regional, and National Estimates of Rotavirus Mortality in Children < 5 Years of Age, 2000-2013.* Clinical Infectious Diseases, 2016. **62**: p. S96-S105.
6. Zhang, H., et al., *Visualization of protein-RNA interactions in cytoplasmic polyhedrosis virus.* J Virol, 1999. **73**(2): p. 1624-9.
7. Zhou, Z.H., *Cypovirus*, in *Segmented Double-Stranded RNA Viruses: Structure and Molecular Biology*, J.T. Patton, Editor 2008, Caister Academic Press: Norfolk, UK. p. 27-43.
8. Hill, C.L., et al., *The structure of a cypovirus and the functional organization of dsRNA viruses.* Nat Struct Biol, 1999. **6**(6): p. 565-568.
9. Yu, X.K., et al., *A putative ATPase mediates RNA transcription and capping in a dsRNA virus.* Elife, 2015. **4**.
10. Payne, C.C. and K.A. Harrap, *Cytoplasmic polyhedrosis viruses*, in *The Atlas of Insect and Plant Viruses*, K. Maramorosch, Editor 1977, Academic Press: New York. p. 105-129.
11. Nason, E.L., S.K. Samal, and B.V.V. Prasad, *Trypsin-induced structural transformation in aquareovirus.* Journal of Virology, 2000. **74**(14): p. 6546-6555.

12. Zhang, X., et al., *3.3 angstrom Cryo-EM Structure of a Nonenveloped Virus Reveals a Priming Mechanism for Cell Entry.* Cell, 2010. **141**(3): p. 472-482.

13. Zhang, X., et al., *In situ structures of the segmented genome and RNA polymerase complex inside a dsRNA virus.* Nature, 2015. **527**(7579): p. 531-+.

14. Kim, J., et al., *Orthoreovirus and Aquareovirus core proteins: conserved enzymatic surfaces, but not protein-protein interfaces.* Virus Research, 2004. **101**(1): p. 15-28.

15. Cheng, L.P., et al., *Backbone Model of an Aquareovirus Virion by Cryo-Electron Microscopy and Bioinformatics.* Journal of Molecular Biology, 2010. **397**(3): p. 852-863.

16. Scheres, S.H.W., *RELION: Implementation of a Bayesian approach to cryo-EM structure determination.* Journal of Structural Biology, 2012. **180**(3): p. 519-530.

17. Dai, X.H., et al., *In situ structures of the genome and genome-delivery apparatus in a single-stranded RNA virus.* Nature, 2017. **541**(7635): p. 112-+.

18. Cui, Z.C., et al., *Structures of Q beta virions, virus-like particles, and the Q beta-MurA complex reveal internal coat proteins and the mechanism of host lysis.* Proceedings of the National Academy of Sciences of the United States of America, 2017. **114**(44): p. 11697-11702.

19. Koning, R.I., et al., *Asymmetric cryo-EM reconstruction of phage MS2 reveals genome structure in situ.* Nature Communications, 2016. **7**.

20. Wang, J., et al., *Crystal structure of a pol alpha family replication DNA polymerase from bacteriophage RB69.* Cell, 1997. **89**(7): p. 1087-1099.

21. Tao, Y.Z., et al., *RNA synthesis in a cage - Structural studies of reovirus polymerase lambda 3.* Cell, 2002. **111**(5): p. 733-745.

22. Liu, H.R. and L.P. Cheng, *Cryo-EM shows the polymerase structures and a nonspooled genome within a dsRNA virus.* Science, 2015. **349**(6254): p. 1347-1350.

23. Li, X.W., et al., *Near-Atomic Resolution Structure Determination of a Cypovirus Capsid and Polymerase Complex Using Cryo-EM at 200 kV.* Journal of Molecular Biology, 2017. **429**(1): p. 79-87.

24. Estrozi, L.F., et al., *Location of the dsRNA-Dependent Polymerase, VP1, in Rotavirus Particles.* Journal of Molecular Biology, 2013. **425**(1): p. 124-132.

25. Brentano, L., et al., *The reovirus protein mu 2, encoded by the M1 gene, is an RNA-binding protein.* Journal of Virology, 1998. **72**(10): p. 8354-8357.

26. Shah, P.N.M., et al., *Genome packaging of reovirus is mediated by the scaffolding property of the microtubule network.* Cellular Microbiology, 2017. **19**(12).

27. Parker, J.S.L., et al., *Reovirus core protein mu 2 determines the filamentous morphology of viral inclusion bodies by interacting with and stabilizing microtubules.* Journal of Virology, 2002. **76**(9): p. 4483-4496.

28. Hogle, J.M., *Poliovirus cell entry: Common structural themes in viral cell entry pathways.* Annual Review of Microbiology, 2002. **56**: p. 677-702.

29. Carr, C.M., C. Chaudhry, and P.S. Kim, *Influenza hemagglutinin is spring-loaded by a metastable native conformation.* Proceedings of the National Academy of Sciences of the United States of America, 1997. **94**(26): p. 14306-14313.

30. Stent, G., *Molecular Biology of Bacterial Viruses.* Royal Society of Health Journal, 1964. **84**(3): p. 151-151.

31. Reinisch, K.M., M. Nibert, and S.C. Harrison, *Structure of the reovirus core at 3.6 angstrom resolution.* Nature, 2000. **404**(6781): p. 960-967.

32. Yu, X.K., et al., *Atomic Model of CPV Reveals the Mechanism Used by This Single-Shelled Virus to Economically Carry Out Functions Conserved in Multishelled Reoviruses.* Structure, 2011. **19**(5): p. 652-661.

33. Bar-Magen, T., E. Spencer, and J.T. Patton, *An ATPase activity associated with the rotavirus phosphoprotein NSP5.* Virology, 2007. **369**(2): p. 389-399.

34. Nason, E.L., et al., *Interactions between the inner and outer capsids of bluetongue virus.* Journal of Virology, 2004. **78**(15): p. 8059-8067.

35. Prasad, B.V., et al., *Visualization of ordered genomic RNA and localization of transcriptional complexes in rotavirus.* Nature, 1996. **382**(6590): p. 471-3.

36. Miyazaki, N., et al., *Transcapsidation and the conserved interactions of two major structural proteins of a pair of phytoreoviruses confirm the mechanism of assembly of the outer capsid layer.* J Mol Biol, 2005. **345**(2): p. 229-37.

37. Chen, J., et al., *Electron tomography reveals polyhedrin binding and existence of both empty and full cytoplasmic polyhedrosis virus particles inside infectious polyhedra.* J Virol, 2011. **85**(12): p. 6077-6081.

38. Zhang, H., et al., *Molecular interactions and viral stability revealed by structural analyses of chemically treated cypovirus capsids.* Virology, 2002. **298**(1): p. 45-52.

39. Lander, G.C., et al., *Appion: An integrated, database-driven pipeline to facilitate EM image processing.* Journal of Structural Biology, 2009. **166**(1): p. 95-102.

40. Li, X.M., et al., *Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM.* Nature Methods, 2013. **10**(6): p. 584-+.

41. Emsley, P., et al., *Features and development of Coot.* Acta Crystallographica Section D-Biological Crystallography, 2010. **66**: p. 486-501.

42. Kelley, L.A., et al., *The Phyre2 web portal for protein modeling, prediction and analysis.* Nature Protocols, 2015. **10**(6): p. 845-858.

43. Adams, P.D., et al., *PHENIX: a comprehensive Python-based system for macromolecular structure solution.* Acta Crystallographica Section D-Biological Crystallography, 2010. **66**: p. 213-221.

44. Pettersen, E.F., et al., *UCSF chimera - A visualization system for exploratory research and analysis.* Journal of Computational Chemistry, 2004. **25**(13): p. 1605-1612.

Chapter 6 In situ structures of rotavirus polymerase in action and mechanism of mRNA transcription and release

Ke Ding[1,2,3], Cristina C. Celma[4], Xing Zhang[2], Thomas Chang[2,3], Wesley Shen[2,3], Ivo Atanasov[2], Polly Roy[4,*], Z. Hong Zhou[1,2,3,*]

[1]Department of Bioengineering, University of California, Los Angeles, California 90095, USA

[2]California NanoSystems Institute, University of California, Los Angeles, California 90095, USA

[3]Department of Microbiology, Immunology and Molecular Genetics, University of California, Los Angeles, California 90095, USA

[4]Department of Pathogen Molecular Biology, London School of Hygiene and Tropical Medicine, United Kingdom

*Correspondence and requests for materials should be addressed to Z.H.Z. (email: Hong.Zhou@ucla.edu) or to P.R. (email: Polly.Roy@lshtm.ac.uk)

6.1 Abstract

Transcribing and replicating a double-stranded genome require protein modules to unwind, transcribe/replicate nucleic acid substrates, and release products. Here we present *in situ* cryo-electron microscopy structures of rotavirus dsRNA-dependent RNA polymerase (RdRp) in two states pertaining to transcription. In addition to the previously discovered universal "hand-shaped" polymerase core domain shared by DNA polymerases and telomerases, our results show the function of N- and C-terminal domains of RdRp: the former opens the genome duplex to isolate the template strand; the latter splits the emerging template-transcript hybrid, guides genome reannealing to form a transcription bubble, and opens a capsid shell protein (CSP) to release the transcript. These two "helicase" domains also extensively interact with CSP, which has a switchable N-terminal helix that, like cellular transcriptional factors, either inhibits or promotes RdRp activity. The *in situ* structures of RdRp, CSP, and RNA in action inform mechanisms of not only transcription, but also replication.

6.2 Introduction

DNA replication and RNA transcription are two of the three steps of Crick's central dogma governing cellular life[1]. The gradual emergence of DNA-based life forms from the RNA world has been hypothesised to be punctuated by major leaps, including RNA replication, RNA-dependent RNA transcription, and RNA reverse transcription to synthesise DNA[2]. Although ribozymes are rare in the modern world, recent discoveries[3] have supported the theory that the first RNA-dependent RNA polymerase (RdRp) was likely a ribozyme[4-6]. In the modern DNA-protein world, proteins have evolved to be the preferred polymerases that catalyze DNA replication and RNA transcription, including RNA-dependent RNA transcription occurring in viruses and cells. The first atomic structure of a polymerase (*Escherichia coli* Polymerase I) revealed a characteristic core shaped like a right hand[7]. Crystal structures of viral RdRps[8, 9], such as those in poliovirus[10], bacteriophage phi6[11], animal reovirus[12], and rotavirus[13], also have cores similar to that of DNA polymerases. A similar core structure also exists in telomerase reverse transcriptase (TERT)[14]. The conserved function of the core is to take a single-stranded nucleotide template and amplify it to a double-stranded product. These polymerases are specialised by both the addition of peripheral domains surrounding the core and the binding of regulatory factors at different time points of polymerization. In the spatial dimension, polymerases that carry out DNA replication (such as DNA polymerase III) contain an exonuclease as a peripheral domain to proofread the dsDNA product; those involved in RNA transcription (such as the viral RdRp of influenza B) possess endonuclease and cap-binding peripheral domains to direct the primer into the active site[15]. In the temporal dimension, this specialization can be further reflected by various regulatory factors, which form various complexes with the polymerase at different stages of

polymerization. For example, the RdRp of bacteriophage Qβ recruits host translation elongation factors to form replicase holoenzyme[16].

In order to fully understand these specialization processes, detailed *in situ* structures of polymerases in its active states are needed. However, there have been issues with obtaining the correct spatial and temporal contexts for these structures. Reoviruses have long served as model organisms for studying viral RdRp and RNA conservative transcription. Structures of Reovirus RdRp with various RNA substrates[12, 13] have been resolved previously by X-ray crystallography, all of which have a cage-like structure with a cap-binding site and four channels: template entry, NTP entry, template exit and transcript exit. However, many purified RdRp only shows binding affinity to RNA/NTP substrates and limited polymerization activity,[17] leaving the spatial context unknown. Additionally, previous studies[18, 19] on active reovirus polymerases also failed to show the complete trajectory of the template or transcript RNA, thus leaving unclear the function of potential RNA-interacting peripheral domains (*i.e.* N- & C-terminal domains in reovirus RdRp). Previous research into these structures has also left unclear the temporal context of these polymerases that undergo conservative transcription (in which the nascent strand is the transcript). Some dsRNA viruses that conduct conservative transcription cannot achieve full polymerase activity by itself. For example, the inner capsid shell protein (CSP) is required for rotavirus' RdRp to be active *in vitro*[20]. On the other hand, for some dsRNA viruses that conduct semi-conservative transcription, in which the nascent strand is part of the dsRNA genome (e.g., bacteriophage φ6[21] and picobirnavirus[22]), RdRp is completely functional for replication *in vitro*. However, exactly how CSP regulates[23] RdRp's activities in rotaviruses remains unknown. Also, unlike other RdRps that conduct semi-conservative transcription, reovirus's RdRp can conduct both replication & transcription and switch between the two states directly after

polymerization. In essence, a virus must be actively running to understand the temporal context, which is very difficult to do through X-ray crystallography.

Cryo electron microscopy (cryoEM) offers opportunities to address both these issues, as it enables the structural characterizations of *in situ* structures in transient, active states. Here, we report the *in situ* near-atomic resolution structures of RdRp before and during transcription in rotavirus double layered particles (DLP). Compared to other viruses in the *Reoviridae* family, rotaviruses are of particular interest for several reasons. In terms of medical significance, they cause diarrhea responsible for up to half a million children deaths annually[24]. Rotaviruses also display significant biochemical simplicity, as their RdRp does not have a separate NTPase protein bound as in other reoviruses; thus, the working mechanisms of rotavirus's RdRp can be studied clearly.

6.3 Results

6.3.1 In situ structures of RdRp in action

To capture RNA transcription in action, we imaged DLPs of rhesus rotavirus (RRV) under active transcribing conditions (Figure 6.6 and Table 6.1). We resolved RdRp and RNA structures following a two-step data analysis procedure (Figure 6.7). First, conventional icosahedral refinement of these particles provided a reconstruction at 3.4 Å resolution. To resolve the RdRp, we carried out localized reconstructions[25]. The final localized reconstruction from sub-particles reached 3.6 Å resolution, which showed RdRp (VP1) interacting with both RNA and inner capsid proteins (VP2) (Figure 6.1a-d). An atomic model was built based on this high-resolution *in situ* structure, with distinct side chain densities and RNA features (Figure 6.1e, Figure 6.8-6.9 and

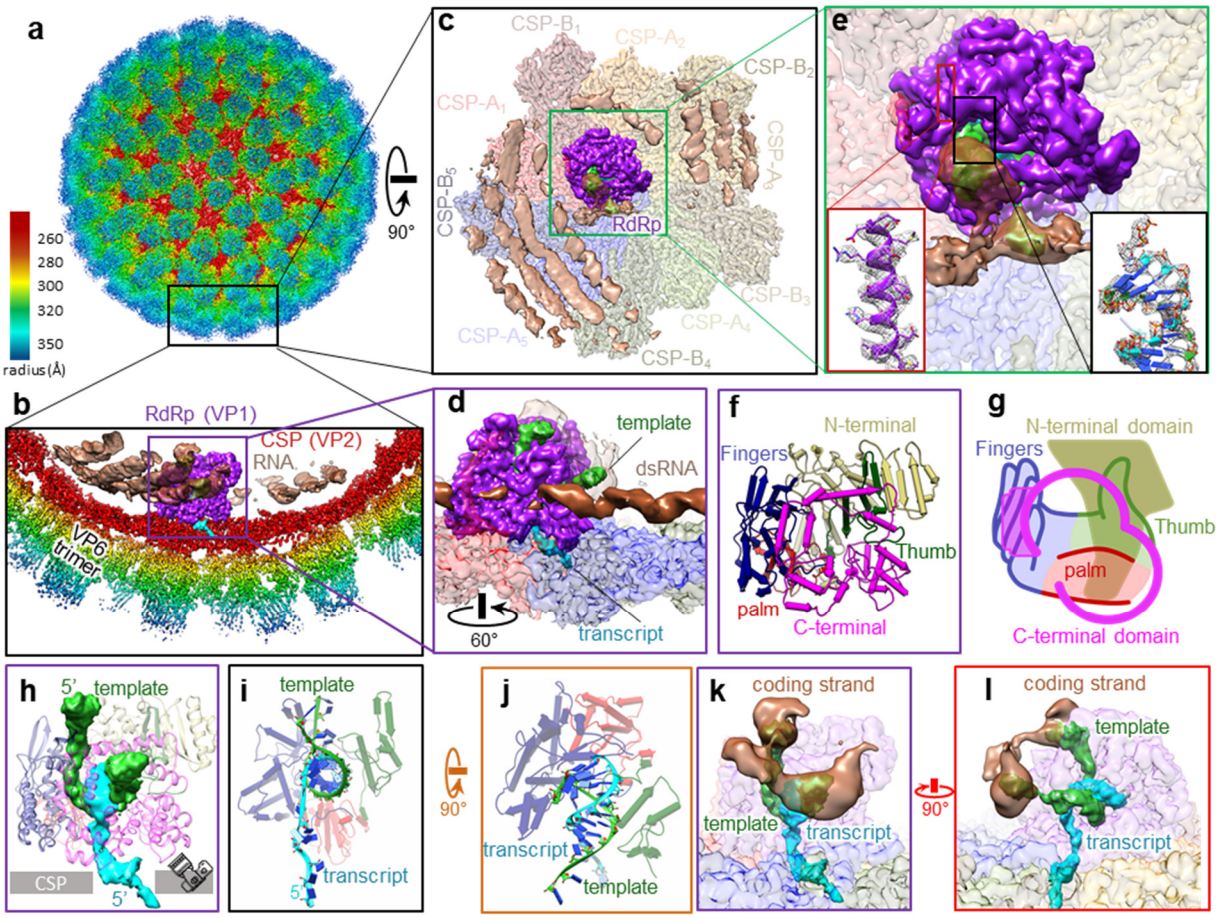


Figure 6.1 ***Visualizing a working polymerase in situ. a,*** *CryoEM reconstruction of rotavirus at 3.4 Å resolution, coloured by radius.* ***b,*** *The RdRp (purple) can be found on the inside of the penton formed by the capsid shell protein (CSP) (red). Genomic RNA density (brown) is packed in the interior of the DLP. The transcript (cyan) is released through the vertex.* ***c,*** *90° rotated view from the boxed region in* ***a****, showing a classic top view with the extended dsRNA genome (with the typical ~27Å distance between its neighboring strands).* ***d,*** *60° rotated view from the boxed region in* ***b****, showing the clear major and minor grooves of dsRNA.* ***e,*** *Magnified view from the boxed region in* ***c****, with additional zoomed-in boxes showing densities (meshes) superimposed upon the atomic models for RNA and RdRp.* ***f, g,*** *Pipes-and-planks representation (****f****) and schematic (****g****) of RdRp in the same classic front view in* ***b****, coloured by domain.* ***h,*** *Ribbon models of RdRp during transcription with RdRp shown as ribbons and RNA densities as coloured surfaces, including the template (lime green) and transcript (cyan).* ***i,*** *View from the camera angle shown in* ***h*** *along the dsRNA axis, shown along with the pipes-and-planks representation for RdRp's core.* ***j,*** *90° rotated view from* ***i*** *showing the 10-base pair-long dsRNA product.* ***k,*** *A transcription bubble is formed by template RNA(green) and genomic RNA (brown).* ***l,*** *90° rotated view from* ***k*** *showing the transcription bubble in near proximity to RdRp.*

Movies 6.1 and 6.2). We determined that the RdRp is attached to CSP decamers at a specific, off-centered location, as previously described[26, 27]. For the ten CSPs in the decamer, we named the five copies close to the decamer center CSP-A$_{1\text{-}5}$, and the others CSP-B$_{1\text{-}5}$, with respect to its relative position to the RdRp (Figure 6.1c, and Figure 6.8). The RdRp has a conserved hand-shaped

core domain (residues 333-778), which is sandwiched between an N-terminal domain (residues 1-332) and a C-terminal domain (residues 779-1088) (Figure 6.1f, g, and Movie 6.3). This core domain can further be divided into the fingers, palm, and thumb subdomains, with the active site located between the fingers and palm. Based on the double-stranded RNA (dsRNA) product density in the active site, we identified two partially-paired single-stranded RNA (ssRNA) strands: the (+)RNA transcript (cyan) and the (-)RNA (lime green) template (Figure 6.1h-j). The 5' end of the transcript extends outside the RdRp, passing through the capsid shell towards the exterior. In contrast, the template strand traverses through the RdRp (parallel to the capsid shell) and reanneals with its complementary coding strand [(+)RNA, brown] to complete a transcriptional bubble within the capsid interior (Figure 6.1k and l). Based on these observations, we conclude that our transcribing DLPs are in a transcript-elongated state (TES) and rotavirus is indeed conducting a conservative transcription.

To further study conservative transcriptional mechanisms, we imaged DLPs at non-transcribing state with the same methods (Figures. 6.6-7, Table 6.1). In the final sub-particle reconstruction at 3.4 Å resolution, we found no RNA density in the active site; however, two ssRNAs that attach to two separate positions on the surface of RdRp were detected. As detailed below, we interpret that these two ssRNAs are the result of an open genomic duplex. Thus, the RdRps in these DLPs existed mainly in a duplex-open state (DOS) (Figure 6.2a) compared to TES (Figure 6.2b). With opened duplex and strands outside the active site, this RdRp structure in DLP is different from all previously reported *in situ* structures of reovirus[18, 19, 28]. In addition to resolving densities of genomic RNA and mRNA in action, our *in situ* structures differ from previous rotavirus's RdRp crystal structures[13] in the following aspects: we resolved two protein

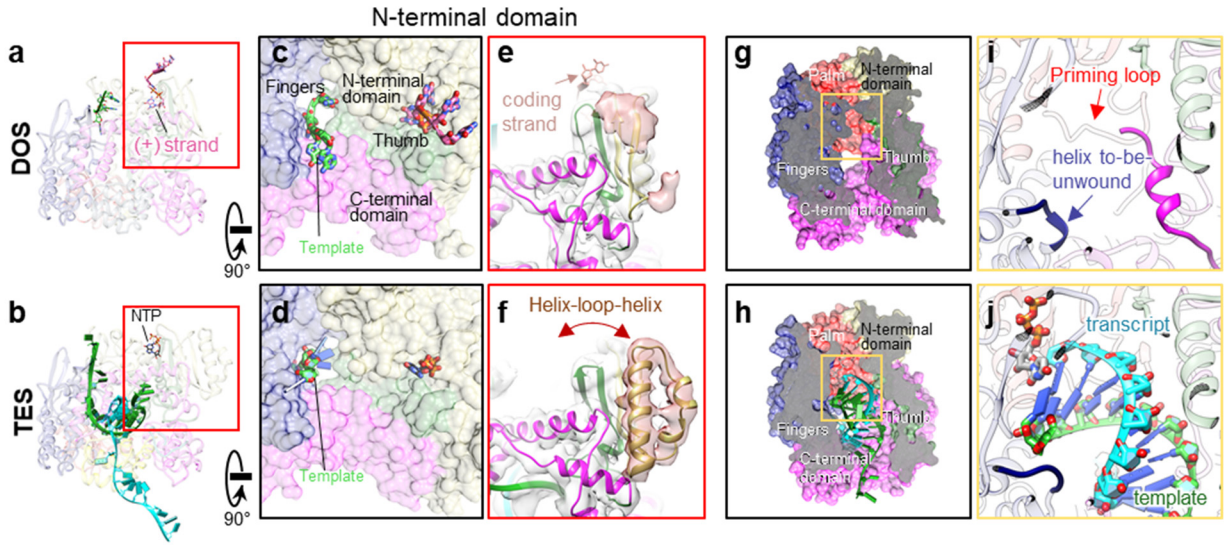


Figure 6.2 ***RNA and RdRp conformational changes between DOS and TES.*** ***a, b,*** *Ribbon models RpRp (pale) and RNA/NTP (bright) of DOS (**a**) and TES (**b**) from the classic front view.* ***c, d,*** *90° rotation from **a** and **b** showing the cap-binding site. The m7G(5')ppp(5')GGC cap (hot pink) binds to the N-terminal cap-binding site in DOS (**c**), and is replaced by an NTP (black) in TES (**d**).* ***e, f,*** *Different conformations of the helix-loop-helix subdomain within the RdRp's N-terminal domain in DOS (**e**) and TES (**f**).* ***g, h,*** *Clipped view of **c** and **d** showing the active site in DOS (**g**) and TES (**h**). The active site contains no RNA in DOS, but is occupied by both the dsRNA product and incoming NTP in TES.* ***i, j,*** *Magnified view from the boxed regions in **g** and **h** shows that the active site is partially blocked by the C-terminal domain in DOS, with the priming loop (residues 489-499) retracted (**i**); in TES (**j**), the active site contains the elongated transcript and the incoming NTP. The priming loop remains retracted in both states.*

fragments (residues 19-21, 346-358) and identified large conformational changes in three fragments (residues 31-69, 923-996, and 1072-1088) (Figs. 9), none of which have been resolved similarly in previous crystallography structures[13, 27]. These new structures are essential to understanding the conservative transcriptional mechanism as detailed below.

6.3.2 RdRp's N-terminal domain splits the genomic dsRNA

Since only the 3' end of a single-stranded template can enter the core, the 5' end of the complementary genomic (+) strand must approach and recognise some region of the RdRp during transcription. In DOS, the cap-binding site of the N-terminal domain (Figure 6.2a, c, Figure 6.10 and Movie 6.4) in RdRp interacts with the conserved terminal m7G(5')ppp(5')GGC residues of the genomic (+) strand in all segments of the rotavirus genome (Figure 6.11). The following bases in all 11 segments of the genome are 6 consecutive bases consisting solely of A and U (Figure 6.11).

In TES, we identified weak densities at the cap-binding site which can only accommodate an NTP molecule (Figure 6.2b, d, Figure 6.10 and Movie 6.4); this cap-binding site has been observed in previous reovirus studies[12, 13, 18]. Compared with other resolved reovirus RdRp structures[18, 28], the rotavirus RdRp' N-terminal domain possesses an additional subdomain that has a helix-loop-helix structural feature (residues 31-69, HLH subdomain) near the cap-binding site. This HLH subdomain extends towards the genomic (+) strand in DOS (Figure 6.2e) and retracts from RNA in TES (Figure 6.2f). The N-terminal domain effectively splits the genome duplex by selectively binding to the 5'-cap-end of the (+) strand RNA, while the HLH subdomain plays a role in further separating the genomic duplex at the downstream AU-box. Later, the (+)RNA bound to the cap-binding site is likely outcompeted by the abundance of NTP in TES.

6.3.3 RdRp's core domain polymerises the complementary RNA

After the dsRNA is split, the unpaired complementary (-)RNA strand traverses the template entrance towards the active site (Figure 6.2g-j). In DOS, the (-)RNA weakly interacts with an ssRNA-binding β-sheet subdomain (residues 400-419) in the fingers (residues 333-488, 524-595) of the core, which can bind ssRNA both specifically and nonspecifically[13]. This strand is then guided by this subdomain through a bottleneck towards the palm (residues 489-523, 596-685) in TES (Figure 6.2h and Figure 6.10). A short helix is unwound (residues 398-401) to accommodate the incoming (-)RNA (Figure 6.2j), confirming its hypothesised role in mediating template RNA entry[13]. The (-)RNA then immediately pairs with complementary NTP in the active site between the fingers and the palm. The incoming NTPs are in position to form a backbone with the 5' end of the nascent RNA (Figure 6.2j). The priming loop (residues 489-499) is slightly offset between the previously published model[13] and our atomic models in the two states, but ultimately stays in a retracted position (away from the active site); it is slightly deformed by CSP but remains

retracted due to the unexpected refolding of neighboring CSP-B1s' N-terminal arm[27] (residues 73-92) outside the RdRp (Figure 6.9, 6.11 and Movie 6.5). Thus, the priming loop does not play the suspected stabilizing role[13] in DOS or TES. Our *in situ* structure shows that the nascent RNA is first stabilised by two conserved positively-charged residues (K679, R680) in the palm (Figure 6.11). The RNA then passes by the thumb (residues 686-778), guided by two other conserved residues (R690, R723). No other charge-based interactions are found that influence the nascent RNA. The dsRNA product is then pushed along by the newly-synthesised nascent RNA backbone until it reaches the C-terminal domain.

6.3.4 RdRp's C-terminal domain splits the dsRNA product

For subsequent translation, the RNA transcript must be split from the template prior to its exit through the capsid. Our structure shows key interactions between the C-terminal bracelet domain and the dsRNA product that facilitate this step (Figure 6.3a-f). A helix-bundle subdomain (residues 923-996, C-HB) blocks the dsRNA's trajectory during elongation; specifically, a conserved I944 residue is responsible for disrupting hydrogen bonds, effectively splitting the dsRNA product (Figure 6.3d and f). Once separated, bases in both strands are immediately flipped to evade the C-HB, and the negatively-charged backbones are further redirected by side-chain-induced electric fields (SCI-EF) (Figure 6.3f-h). As a result, the negatively-charged RNA backbone bends towards the positively-charged surface (blue) and away from the negatively-charged surface (red). The nascent RNA goes towards the capsid through a separate channel between the palm and the bracelet (Movie 6.6). The central subdomain (residues 320-396) of the apical domain (residues

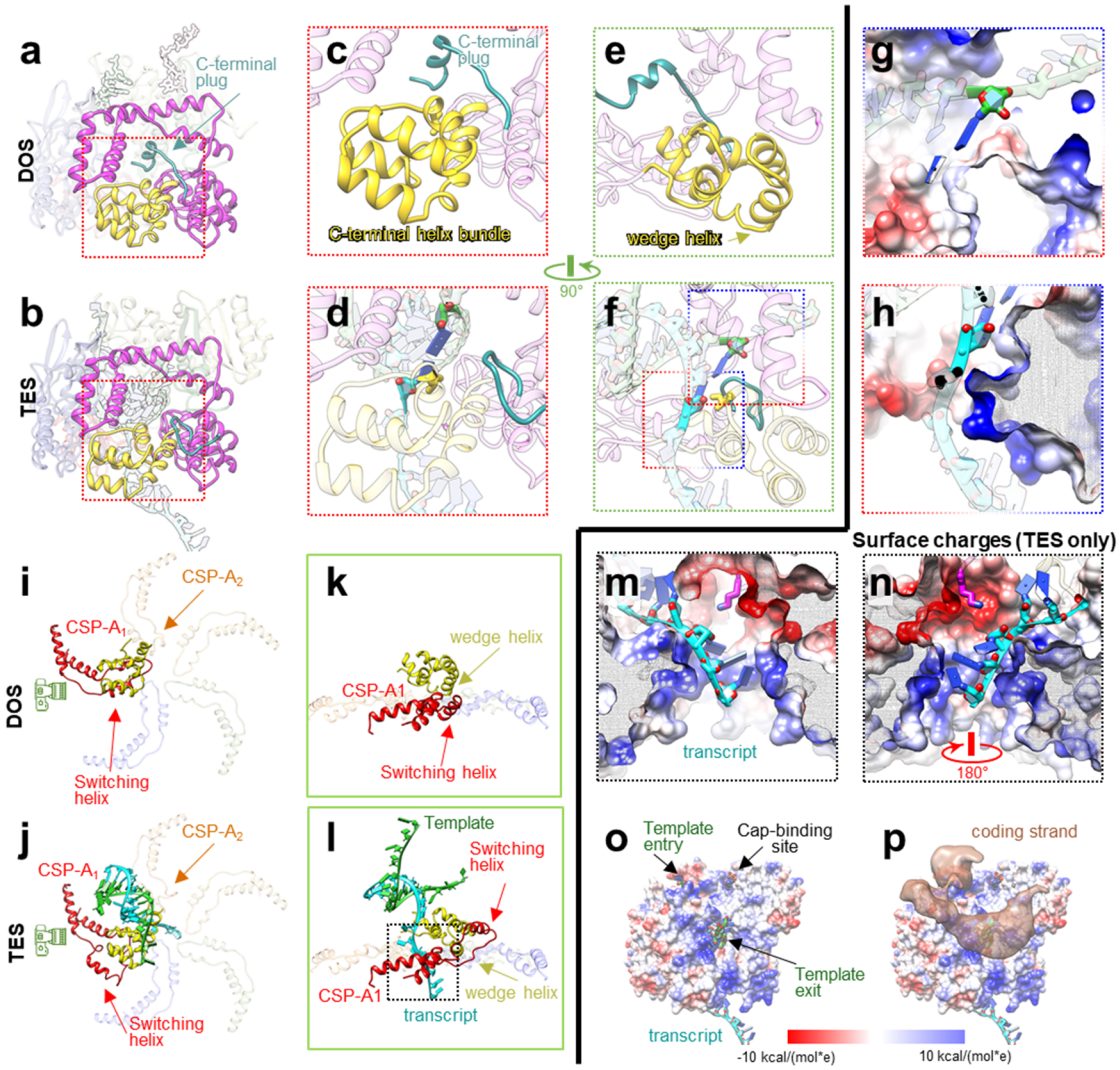


Figure 6.3 ***The RdRp C-terminal bracelet domain splits the dsRNA product.*** ***a, b,*** *Ribbon models of RdRp's C-terminal domain (bright) and nearby RNA (silhouetted) in DOS (**a**) and TES (**b**). **c, d,** Magnified view of the template exit channel in DOS and TES. The C-terminal plug blocks the channel in DOS (**c**). Both the C-terminal plug (dark cyan) and helix-bundle (yellow) undergo dramatic conformational changes in TES (**d**); the helix-bundle is repositioned to aid in duplex base-pair splitting, while the plug is displaced from the exit channel to allow for (-) strand exit. The key residue I944 and the last base pair are highlighted. **e, f,** 90° turn from the classic front view shows the conformational changes of the helix-bundle and plug in DOS (**e**) and TES (**f**). **g, h,** Surface charge representations of C-terminal domain regions show that side-chain-induced electric fields guide the RNA backbone and redirect them towards their respective exit channels. Positively-charged surfaces are colored blue and negatively-charged surfaces are colored red. **i, j,** Ribbon models of RNA and central subdomains within CSP-As in DOS (**i**) and TES (**j**). The transcript can exit a channel through the penton center that is exclusively open in TES. CSP-A_2's apical domain remains lifted away from the penton center in both states (orange arrow). **k, l,** Magnified views from the camera angle in **i** and **j** show that CSP-A_1's apical domain is deformed by the translocated C-HB in TES. **m,** Magnified boxed region (penton center) in **l** in surface charge representations in TES show that the nascent RNA is flipped by side-chain-induced electric fields at the penton center. **n,** A 180-degree rotation from **m** shows the surface charges of the opposite site of the penton center. **o,** Surface charge representations of the C-terminal domain show that it forms a highly-positively charged surface region between the template entry channel, cap-binding site, and the template exit channel. **p,** Superimposing the coding strand's density on **o**. shows that the coding strand density follows the positively-charged RNA track to reanneal with the template. No RNA strand binds to the cap-binding site in TES.

320-596) of five CSP-As is asymmetrically translocated by RdRp (Figure 6.3i-l, and Figure 6.13).

As a result, a pore is formed through the center of the CSP-A penton (Figure 6.3j and l), which processes another SCI-EF to further deflect the nascent RNA (Figure 6.3m and n). This nascent RNA eventually exits the capsid shell through this opening in TES. In DOS, however, the C-HB subdomain retracts from CSP-A_1 and narrows the transcript exit channel (Figure 6.3I, k and Movie 6.7), such that CSP-A_1 returns to a similar conformation as the ones found in CSP-$A_{3\text{-}5}$. Two short helixes [residues 349-360 of CSP-A_1 (switching helix) and residues 968-979 of C-HB (wedge helix)] (Figure 3k and l) compete for a pocket between CSP-A_1 and RdRp in these two states. Seeing that no cleaving of peptide chain is involved, this mechanism is likely reversible: the RNA exit channel can be shut after rotavirus's secondary transcription[29] and reopened upon entering a new host's cytoplasm. In contrast, CSP-A_2's apical domain remains wedged in both states by the neighboring RdRp (Figure 6.3i-j). Simultaneously, the newly isolated (-)RNA exits through the template exit channel located in the center of the C-terminal domain. The C-terminal domain essentially provides a positively-charged ssRNA track on its surface between the template entry and template exit channels; thus, the coding strand can follow this track to reanneal with the template (Figure 6.3o and p) and reform the dsRNA genome. The mechanics in the C-terminal domain not only split the dsRNA product (without utilizing additional NTP like other cellular helicases, crucial for conservative transcription), but also redirects the transcript towards the capsid. These movements create sufficient pressure to selectively open a transcript exit channel on demand.

6.3.5 Two CSP-As' N-terminal: transcriptional factors

As a compact nanomachine, rotavirus RdRps also recruit transcriptional factors to regulate their function, similar to other polymerases. CSP-A's N-terminal regions (residues 62-116) form different transcriptional complexes with RdRp (Figure 6.4a-d) through a tethered amphipathic

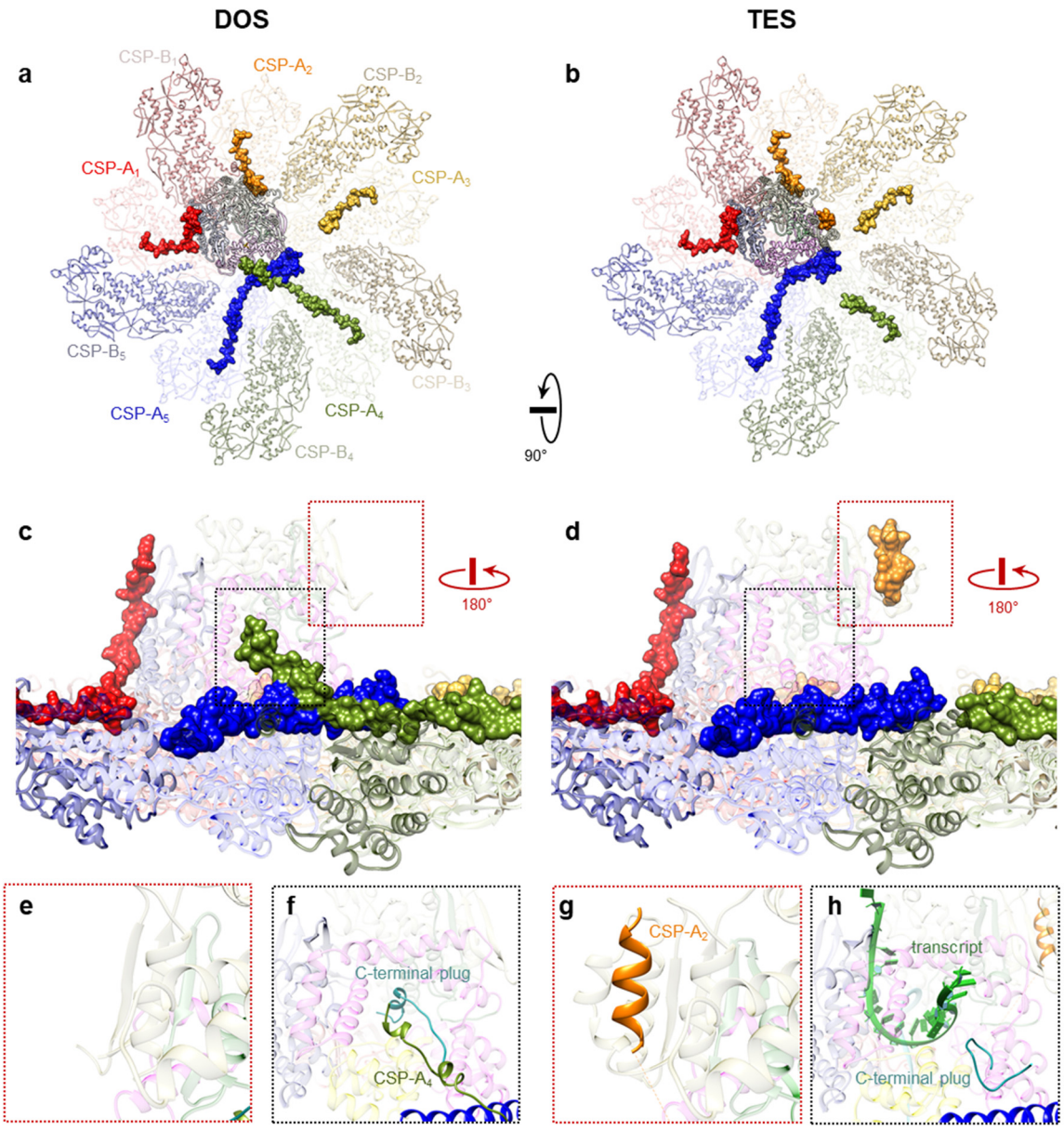


Figure 6.4 ***N-terminal of CSP-As form different transcribing complexes with RdRp. a, b,*** *Ribbon diagrams of CSP-As (pale) and RdRp (silhouette). The N-terminals of CSP-As are highlighted with coloured surfaces in both DOS (**a**) and TES (**b**). **c, d,** N-terminal surfaces rotated 90° degrees from the view in **a** and **b.** The binding sites of CSP-A_4's (green) and CSP-A_2's (orange) N-terminal tethers are boxed in black and red, respectively. **e-h** Detailed detachment/attachment of transcriptional factors on RdRp between the two states from the boxed regions in **c** and **d**. The absence of CSP-A_2's amphipathic helix in DOS (**e**) and its presence in TES (**g**) suggests that this helix in CSP-A_2 stabilises the HLH subdomain in DOS. The presence of CSP-A_4's amphipathic helix in DOS (**f**) and its absence in TES (**h**) suggests that this same amphipathic helix in CSP-A_4 locks C-HB's conformation in DOS.*

helix (residues 78-84, QLLEVLK, Figure 6.4e-h and Figures. 6.14 and 6.15). This tethered amphipathic helix in CSP-A_2 attaches to a hydrophobic pocket next to the structured HLH subdomain in TES but detaches from this pocket as the HLH subdomain becomes flexible in DOS

(Figure 6.4e, g and Movie 6.8). This helix-binding action effectively anchors the HLH subdomain and prevents unfavorable interactions with genomic RNA in TES, thus promoting RdRp activity and RNA release. However, the corresponding amphipathic helix in CSP-A_4 attaches to the C-HB of RdRp in DOS and detaches from RdRp in TES. The association of this helix closes the template exit channel in DOS and opens it in TES (Figure 6.4f and h). In contrast to its counterpart in CSP-A_2, this helix in CSP-A_4 actually inhibits RdRp's activity by locking C-HB's conformation and blocking the template exit channel. Given these observations, we can conclude that CSP's N-terminal regions serve as transcriptional regulating factors for RdRp. Similar regulatory mechanisms can also be found in the structure of the rotavirus RdRp itself. A unique C-terminal plug (residues 1072-1088) inserts into the template exit channel in DOS, but moves away in TES to allow (-)RNA to exit. This C-terminal plug is close to the priming loop in DOS and potentially influences the priming loop's approach to the nascent NTP during initiation (Figure 6.2i). Thus, the C-terminal plug is another example of the regulatory factors present in rotavirus transcription/replication. We also find other minority states in our dataset (Figure 6.15) that potentially reflect the numerous transient states of RdRp.

6.4 Discussion

Because the N and C terminal domains in rotavirus' RdRp play such integral roles in its activity, we infer that these may have evolved as critical extensions to the conserved polymerase core (shared by DNA polymerases, telomerases, and RdRp). Both termini effectively function as minimalistic helicases and are essential for conservative transcription. In DOS, the N-terminus is capable of splitting the dsRNA genome with only around 330 residues; this domain recognises and interacts with 5' consensus bases (GGC) of (+)RNA at the cap-binding site (CBS), so that the subsequent 6-base-long A/U-only box can be more efficiently split by the neighboring HLH

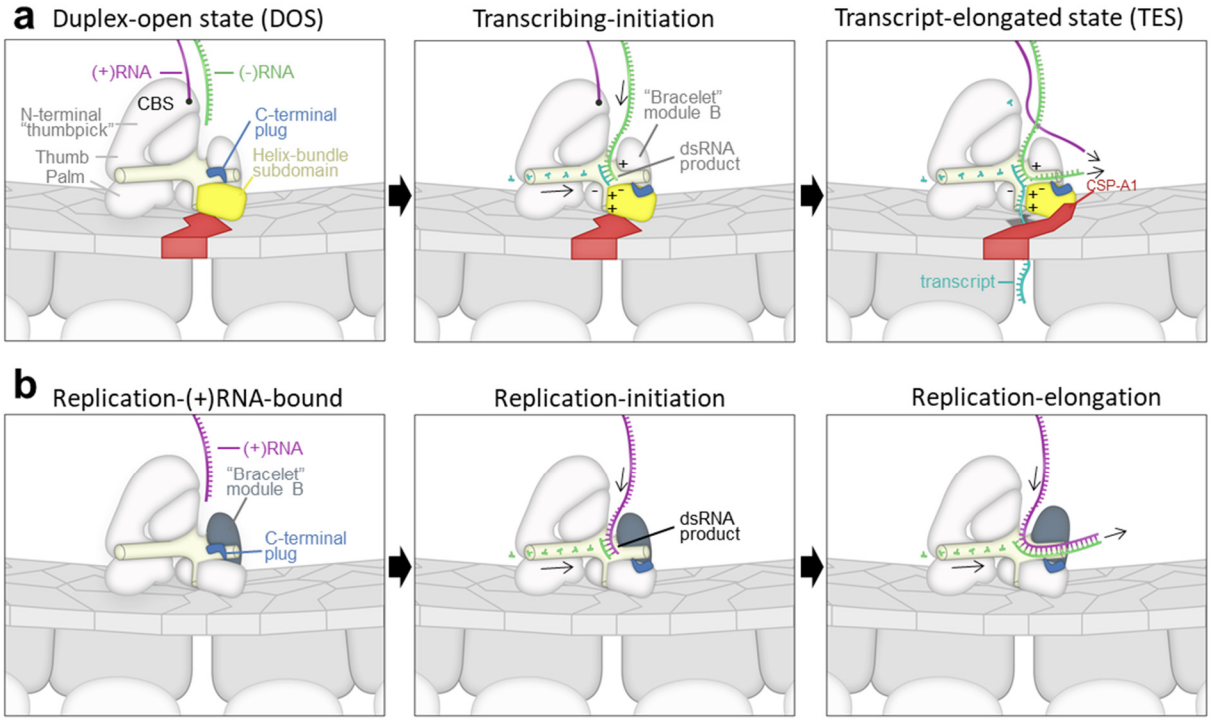


Figure 6.5 ***Transcription and replication mechanism. a,*** *Transcriptional mechanism of rotavirus informed by our in situ RdRp structures in action.* ***b,*** *Possible mechanism of rotavirus RNA replication, deduced from the observed structures of the transcriptional machinery.*

subdomain. As a result, the newly-isolated (-)RNA attaches to the nearby ssRNA recognition site on the fingers (Figure 6.2c). This A/U-only region is similar to the A/T-rich TATA box and Pribnow box, which is easily melted and plays a key role in cellular transcription initiation[30]. Because the RdRp's N-terminal domain interacts with string-like RNA and is close to the thumb, we renamed the N-terminal domain of RdRp the N-terminal "thumbpick" domain. In TES, the C-terminal bracelet not only exhibits functional helicase activity, but also redirects the two RNA strand products to exit through their respective channels. In redirecting RNA strand products, the C-terminal region also helps reorganise the nascent genomes. These peripheral domains allow RdRp to operate in a continuous fashion during transcription (Figure 6.5a). In DOS, the 5' end of genomic (+)RNA binds to CBS, and (-)RNA proceeds to the template entry. The (-)RNA is then transcribed, and the resulting dsRNA product reaches the aforementioned machinery of the C-terminal domain. Specifically, C-HB is needed to split the dsRNA product and isolate the single-

stranded transcript. The C-HB subdomain is pushed by the incoming product and realigned to the center of the product's base pairs in an orientation that allows for effective splitting of the product. As a result, the translocated C-HB subdomain pushes on the CSP-A_1's apical domain to selectively open the transcript exit gate on the capsid shell during ongoing transcription. The (-)RNA undergoes a near U-turn (Figure 6.4h) in RdRp and returns into the capsid interior near the CBS. Under ideal circumstances [abundance of GTP, accumulation of (+)RNA near CBS], elongation results in the displacement of (+)RNA from CBS by a GTP molecule, allowing (+)RNA to reanneal with the nearby exiting (-)RNA, thus completing the transcription bubble in TES. Intriguingly, we did not find the capping enzyme anchored inside the capsid interior as suspected[25]. Our visualization of the nascent RNA transcript through the CSP shell immediately after exiting from the RdRp exit channel would be consistent with the external location of a capping enzyme lining the 5-fold opening, geometrically similar to its location in turreted reoviruses[31].

Not only do the N and C terminal domains regulate the genome, but they may also provide interfaces for potential association of transcription factors. This regulation of transcription factors further specialises the protein's function. In rotavirus, the amphipathic helix in CSP-A_2 locks the HLH subdomain to prevent further undesirable interactions with the genome during elongation; this same amphipathic helix in CSP-A_4 locks C-HB and blocks the template exit channel as an inhibiting factor in DOS. This supports previous findings that rotavirus's RdRp-CSP interactions are crucial for polymerization activity[20, 32]. It is also consistent with previous suggestions[28] that aquareovirus CSP's N-terminal region can form different transcriptional complexes with the polymerase at different time points.

Understanding the polymorphic nature of the C-terminal domain also yields insights into viral replication (Figure 6.5b). Without a complementary strand bound to CBS, the C-terminal

domain is less hindered by RNA on its outer surface. When the duplex pushes the C-HB, the upper part of the C-terminal domain (module B) flaps open to let the duplex enter the capsid interior (without the splitting and guiding aspects it displays in transcription), similar to DNA polymerases. This function is recovered in transcription due to both the presence of bound (+)RNA at the beginning of elongation and a relatively crowded capsid interior.

Based on the observation that the capped end of dsRNA leaves RdRp during TES and re-associates with RdRp at cap-binding site in DOS, we propose that the other end of the dsRNA genome (*i.e.*, the tail end) is close to the capped end in DOS. When elongation starts, the entire dsRNA strand is pulled towards the RdRp so that the tail end will leave RdRp, leaving enough space to accommodate the reannealed capped end. At the end of the elongation step, the capped end follows the tail end and circles back to RdRp again. The capped end can then bind to the nearby cap-binding site and start a new transcription cycle, much like an Ouroboros. In this model, the cap is not always bound to the cap-binding site, so there are no undesirable kinks or sharp U-turns on the dsRNA genome during elongation. This model is also more consistent with other RdRps that conduct semi-conservative transcription (*e.g.*, φ6's RdRp[11]), in which the cap is not bound during transcript elongation. However, φ6 phage's RdRp differs quite drastically from rotavirus' in their terminal domains: the RdRp of φ6 has no N-terminal domain, and its C-terminal domain is shorter (65 a.a.) and is suspected to prime polymerization[11] rather than to split and rearrange RNA products. It is possible that in semi-conservative transcription, the transcript is split from the dsRNA genome by a different mechanism; therefore, in φ6 phage, we do not see N- and C-terminal structures similar to those of rotavirus and other reoviruses that conduct conservative transcription.

In summary, the two *in situ* structures of rotavirus RNA polymerase in action suggest that the peripheral domains organise RNA for the core, thus acting like up-/down-stream nodes on a specialised production line. Similar to other polymerases, viral RdRps have also evolved their core units to recruit other proteins,[18, 28] and we show that the recruited capsid proteins, like cellular transcription factors, form different transcriptional complexes with RdRp. Confined in a crowded viral capsid, the highly specialised rotavirus RdRp has simply co-opted its own N- and C-terminal domains and regions of its capsid protein to regulate transitions between different states. As genome transcription is an essential step in rotavirus infection, the *in situ* structures presented here, as well as those from others[33], will also be informative for ongoing drug discovery efforts, in addition to the above-discussed insights about the fundamental biological processes of transcription and replication (Figure 6.5).

6.5 Methods

6.5.1 Double-layered particle purification

Simian rhesus rotavirus (RRV) double-layered particles were purified from rotavirus-infected cells as described elsewhere[34]. Briefly, MA104 cells infected with RRV at a multiplicity of infection (MOI) of 3 were harvested at 100% cytopathic effect. Cell lysate was generated by freezing and thawing twice. The lysate was treated with 50mM EDTA (pH 8) followed by incubation for 1 h at 37 °C. After centrifugation, the pellet was resuspended in TNC buffer (10 mM Tris·HCl, pH 7.4; 140 mM NaCl; 10 mM $CaCl_2$) supplemented with 0.1% Nonidet P-40, and 50 mM EDTA (pH 8) and trichlorotrifluoroethane was added. The aqueous phase was separated by centrifugation, and DLPs were isolated by equilibrium ultracentrifugation at 100,000 × g in a CsCl gradient for 18 h. A band containing DLPs was collected, diluted in TNC buffer, and pelleted through a sucrose cushion (15% sucrose prepared in TNC buffer) by ultracentrifugation at 110,000 × g for 2 h.

Finally, particles were resuspended in 10 mM Tris·HCl, pH 8 prior to either transcription reaction or plunge-freezing.

6.5.2 Cell-free transcription reaction

For the transcription reaction, purified DLPs were incubated in transcription buffer (10 mM Tris-HCl, pH 8; 4 mM rATP; 2 mM rGTP; 2 mM rCTP and 2 mM rUTP; 0.5 mM S-adenosylmethionine; 6 mM DTT; 9 mM $MgCl_2$) for 5 min at 37°C prior to plunge-freezing for cryoEM.

6.5.3 CryoEM and 3D asymmetric reconstruction by symmetric relaxation

An aliquot of 2.5 micolitres of each sample was applied to plasma-cleaned Quantifoil 1.2/1.3 holey cryoEM grids, which were blotted and plunge-frozen with an FEI Vitrobot Mark IV.

High quality cryoEM images were then collected in an FEI Titan Krios 300kV electron microscope, equipped with a Gatan K2 direct electron detector and a Gatan Quantum energy filter. The microscope was carefully aligned and the coma-free alignment was performed to align the beam tilt immediately before the data collection. As detailed in Figure 6.7, we collected both data sets using the counting mode at a framerate of 8 frames per second without putting in the slit of the energy filter with LEGINON[35] automation. DOS data was collected for 8 seconds with a calibrated pixel size of 1.07 Å, while TES was collected for 10 seconds with a calibrated pixel size of 1.33 Å. The first 25 frames in DOS and first 32 frames in TES were aligned with UCSF MotionCorr software[36] to make micrographs with 22 e per $Å^2$ and 18 e per $Å^2$ dosages, respectively. Contrast transfer function (CTF) parameters were determined with CTFFIND4[37] for both datasets.

For DOS, particles were automatically boxed with ETHAN[38]. Virus particles' center and orientations were refined with Relion[39] with icosahedral symmetry i2 (*i.e.*, the convention with

x, y and z axes along the icosahedral 2-fold axes) applied. The final resolution of the resulting icosahedral reconstruction at FSC >0.143 is 3.6 Å.

To obtain the asymmetric structure of the polymerase, we conducted localized reconstruction[25] to focus on particle vertices (*i.e.*, each vertex treated as a sub-particle). First, the icosahedral reconstruction is rotated to follow an icosahedral symmetry i3 (5-fold axis aligned with z axis) and particle orientations were adjusted accordingly[39] (Figure 6.7I). For each particle in the dataset, we calculated the coordinates (rlnOriginX and rlnOriginY) and orientation parameters for the 12 sub-particles (vertices) using a Python script[40]. These coordinates were then used to box out sub-particles by the relion_preprocess command from the RELION package[39]. Second, each sub-particle was then expanded with the "relion_particle_symmetry_expand" command as 5 entries (Figure 6.7II) in the RELION star file, each having a 5-fold-related orientation around the z axis (*i.e.*: only rotational Euler angle (_rlnAngleRot) differs from each other by an increment of 72 degrees). Third, all sub-particles were then subjected to RELION 3D classification by asking for 16 classes with the "skip_align" option (III in the left panel of Figure 6.7) resulting in 9 "good" classes (*i.e.*, those with densities that can be interpreted as one single RdRp at certain density threshold and are demarcated with colour arrows in III of Figure 6.7) and 7 bad classes (colored in cyan in Figure 6.7). These 9 good classes can be further grouped into 5 groups (coloured red, orange, green, blue and purple for group A,B,C,D and E, respectively in Figure 6.7) based on the orientation of the RdRp in the reconstruction of each good class, while the 7 bad classes were grouped into group X. In the ideal situation, the five consecutive entries of every sub-particle should be sequentially placed into one of the five circular permutations of group list A, B, C, D and E. However, our observed results (Figure 6.7, step III) deviated from such ideal situation for two possible reasons: First, the 5-fold-

related capsid proteins could have obscured the alignment signals during classification; Second, there might be multiple conformations of RdRp.

To make the optimal group placement choice for the sup-particles based on our observed results, we developed a Python script program (Orientation_Selection.py) that processes the RELION star file. By taking the star file from 3D classification as input, this script analyzed order of group (A, B, C, D, E or X) placements of the five entries of each sub-particle and find its best match to the 5 possible circular permutations of the ideal group list. If the best match has less than two outliers out of the five groups, this sub-particle will be retained with permuted orientation; otherwise, this sub-particle will be discarded. For example, the result group list "B,C,D,X,A" best matches one-time permuted ideal list "B,C,D,E,A" with one outlier so this sub-particle would be retained with one rotation of 72°, but result group list "B,C,D,X,E" matches "permuted ideal list "B,C,D,E,A" with two outliers so this sub-particle would be discarded. A new star file was created with the retained sub-particle and their orientation assignments. A RELION local classification with limited range of angle search (relion parameter --sigma_ang 3) was then performed to select the major conformation (Figure 6.16). A RELION gold-standard local refinement was finally conducted and the final sub-particle reconstruction reached 3.4 Å resolution (step IV in Figure 6.7).

For TES, we used a similar method as stated above. The resolutions for the icosahedral reconstruction is 3.4 Å and that for the vertex sub-particle reconstruction is 3.6 Å (Figure 6.7 V-VIII).

6.5.4 Atomic model building and model refinement

The atomic models of RRV's RdRp and CSP were built with Coot[41] and refined with Phenix[42]. We first used the "fit in map" function of UCSF Chimera[43] to dock PDB 4F5X, a previously

published montage model, into the sub particle reconstructions of the two states. There are six kinds of major discrepancies: previously flexible regions in crystallography (residues 19-21, 346-358 in RdRp); backbone tracing error (residues 804-821 in CSP); newly-resolved asymmetric features (residues 62-117, 336-373 in CSP-A); conformational changes introduced by RdRp's docking on CSP (residues 487-510 in RdRp, 73-93 in CSP-B_1); large conformational changes between different states (residues 31-69, 923-996, 1072-1088 in RdRp); and *in situ* RNA features (the template, transcript, coding strand and NTP). For those discrepancies, we manually traced the backbone in all-alanine mode in Coot and then mutated them into the correct sequence. RNA in DOS was built with conserved sequences m7GpppGGC at the 5' end of the coding strand and its complementary strand, while RNA in TES was built with repetitive AU polynucleotides. The models in both states were then refined by the PHENIX real-space refine function and validated by the wwPDB validation server[44].

Visualization of the atomic model, including figures and movies are made with UCSF Chimera[43]. The sequence is visualised by ESPRIPT[45].

6.6 Supplementary information

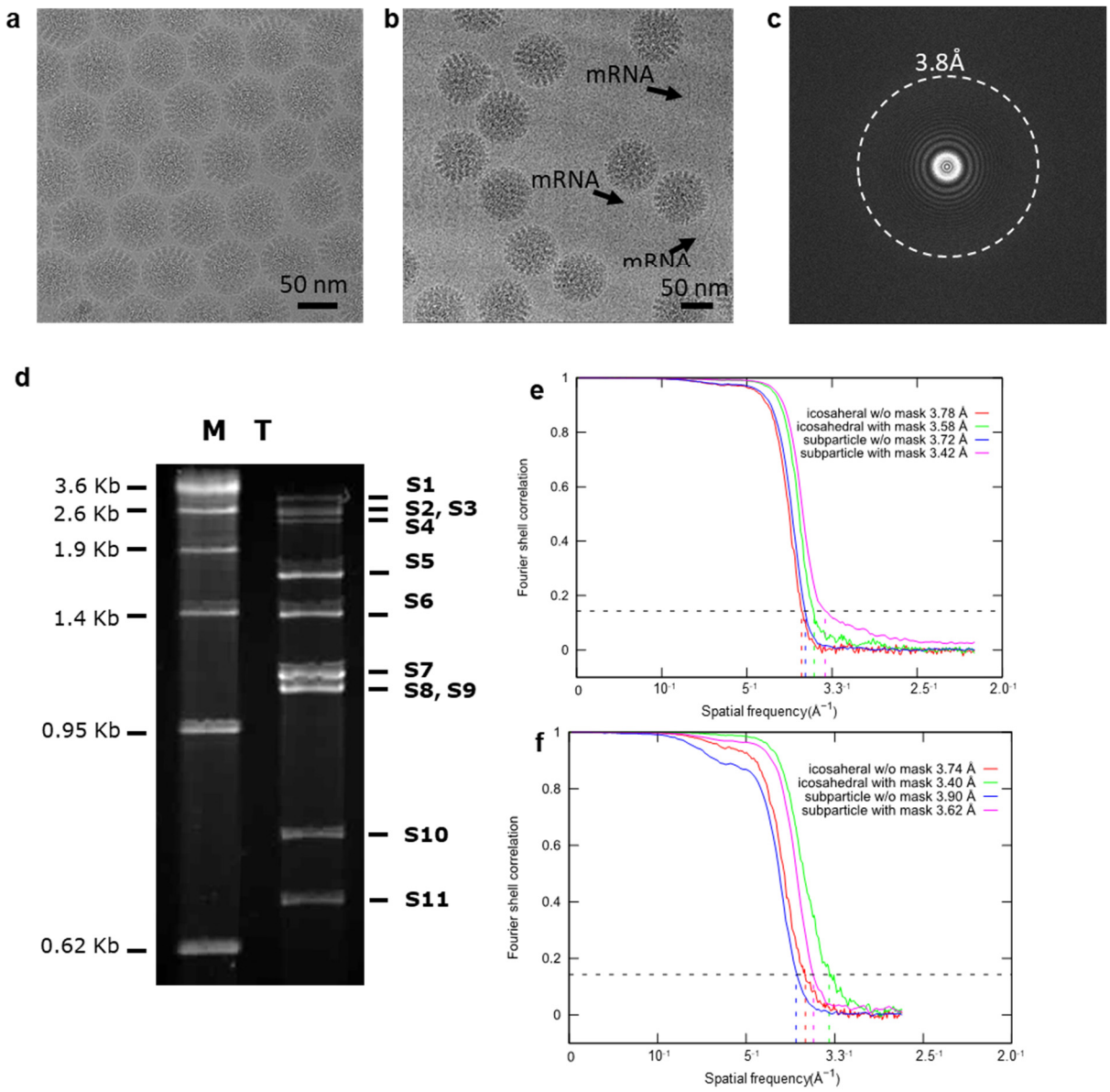


Figure 6.6 ***CryoEM and structure determination statistics. a, b,*** *CryoEM images of rotavirus DLP (**a**) and transcribing DLP (**b**). **c**, Fourier transform of a micrograph of rotavirus DLP, to show the signal limit. **d,** 4% acrylamide- 8M Ura gel to show the transcription product of DLP's in vitro transcription. Lane M: single-stranded RNA marker. Lane T: in vitro transcripts obtained from DLPs. **e, f,** Fourier shell correlation (FSC) coefficient as a function of spatial frequency of DLP (**e**) and transcribing DLP (**f**) reconstructions.*

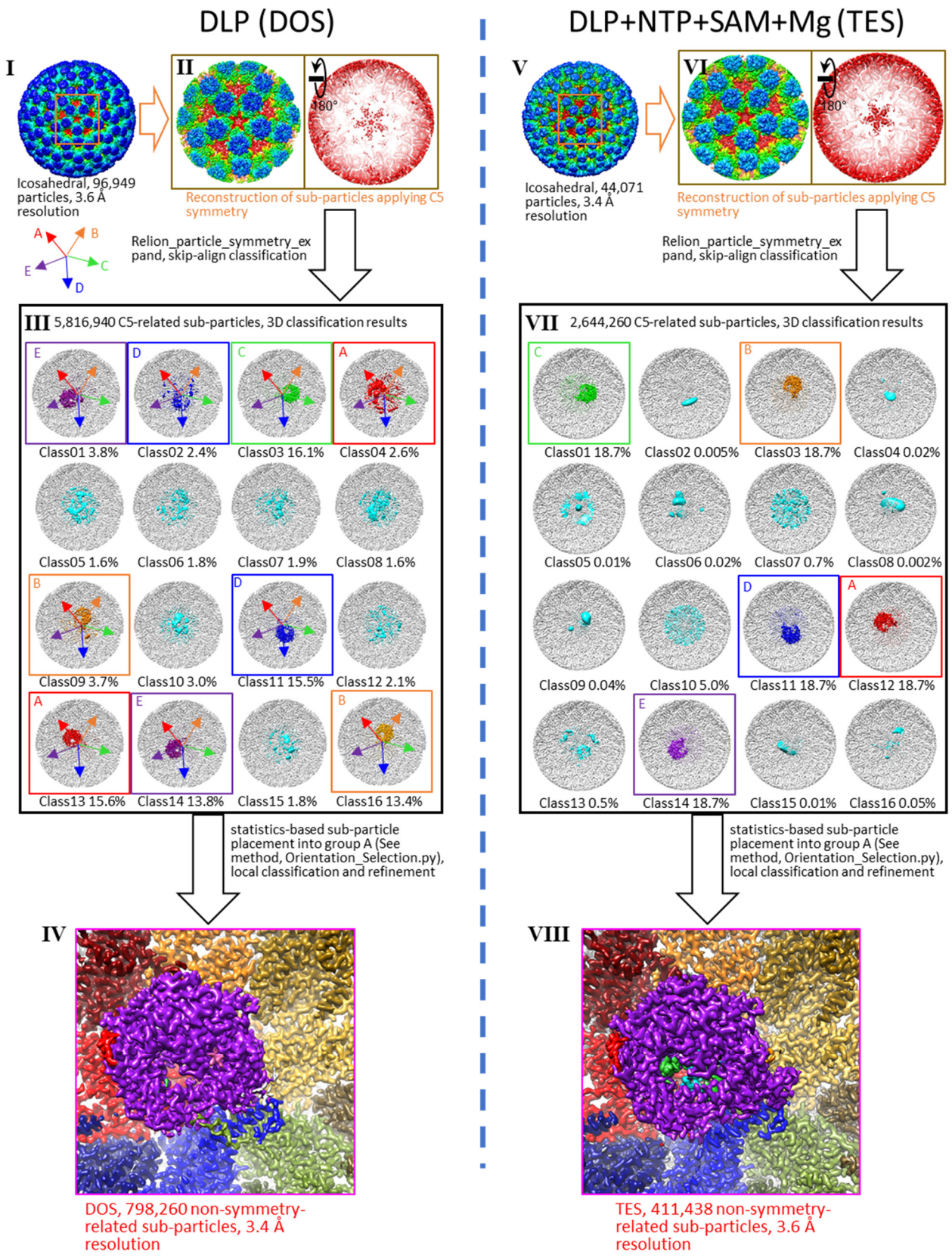


Figure 6.7 ***Data processing workflow for sub-particle reconstructions of DLP (left) and transcribing DLP (right).***

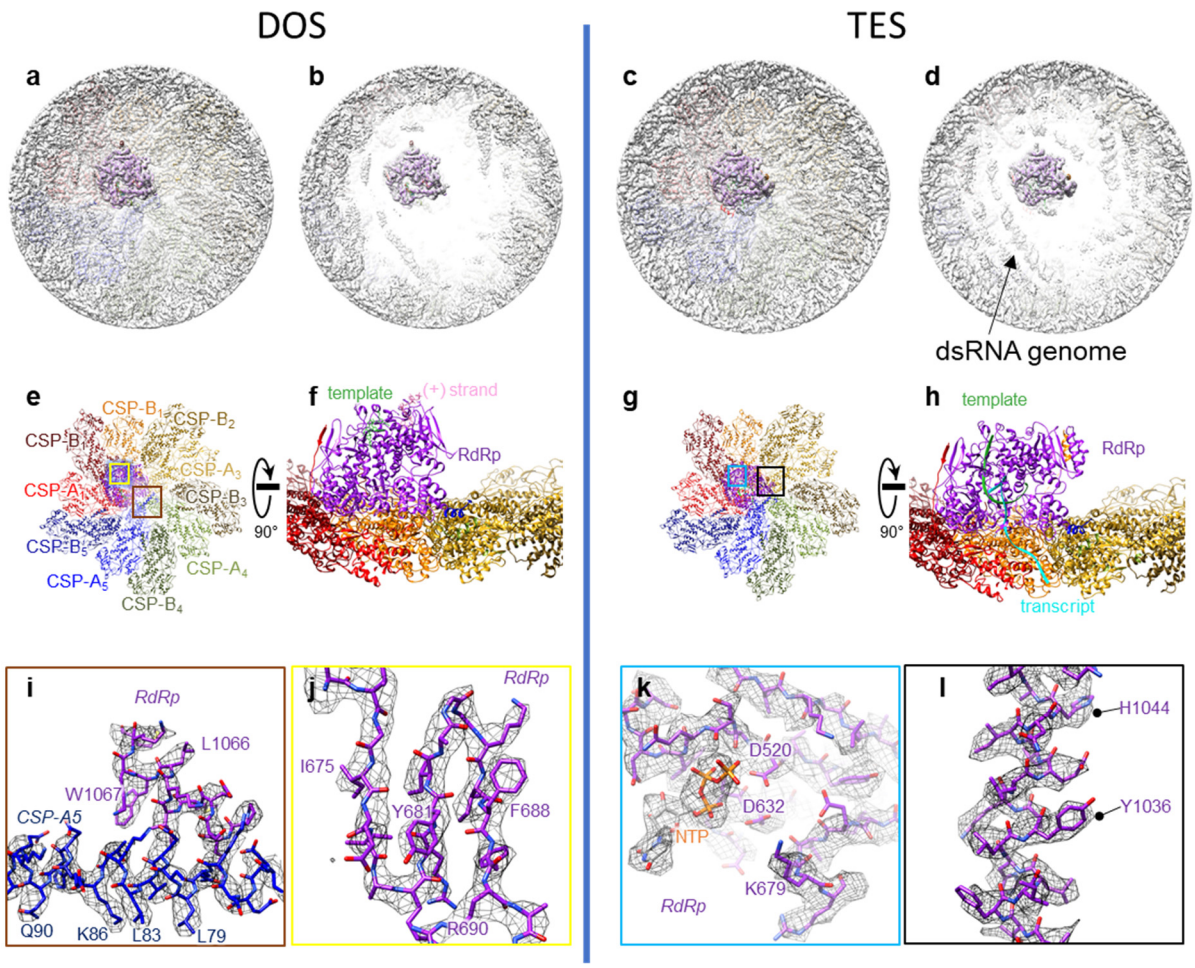


Figure 6.8 ***Sub-particle reconstructions and atomic models of RdRp with associated CSP decamer in two states. a, b,*** *Internal surface view of the DOS sub-particle reconstruction shown fully (**a**) and partially without the rear portion (**b**). **c, d,** Same as (**a, b**) but in TES. Note the surrounding RNA features are better resolved in TES than in DOS. **e-h,** Ribbon diagram of the atomic models of RdRp (purple) and CSP (rainbow) in DOS (**e, f**) and in TES (**g, h**) in two orthogonal views. **i-l** Densities (wires) in the boxed regions of (**e, g**), superposed with atomic models, highlighting some high-resolution features in our structures, including an interaction between RdRp and CSP-A_5 in DOS (**i**), a β-sheet in RdRp in DOS (**j**), the active site with an NTP in TES (**k**), and a helix bundle in RdRp in TES (**l**). See also Movie 6.1.*

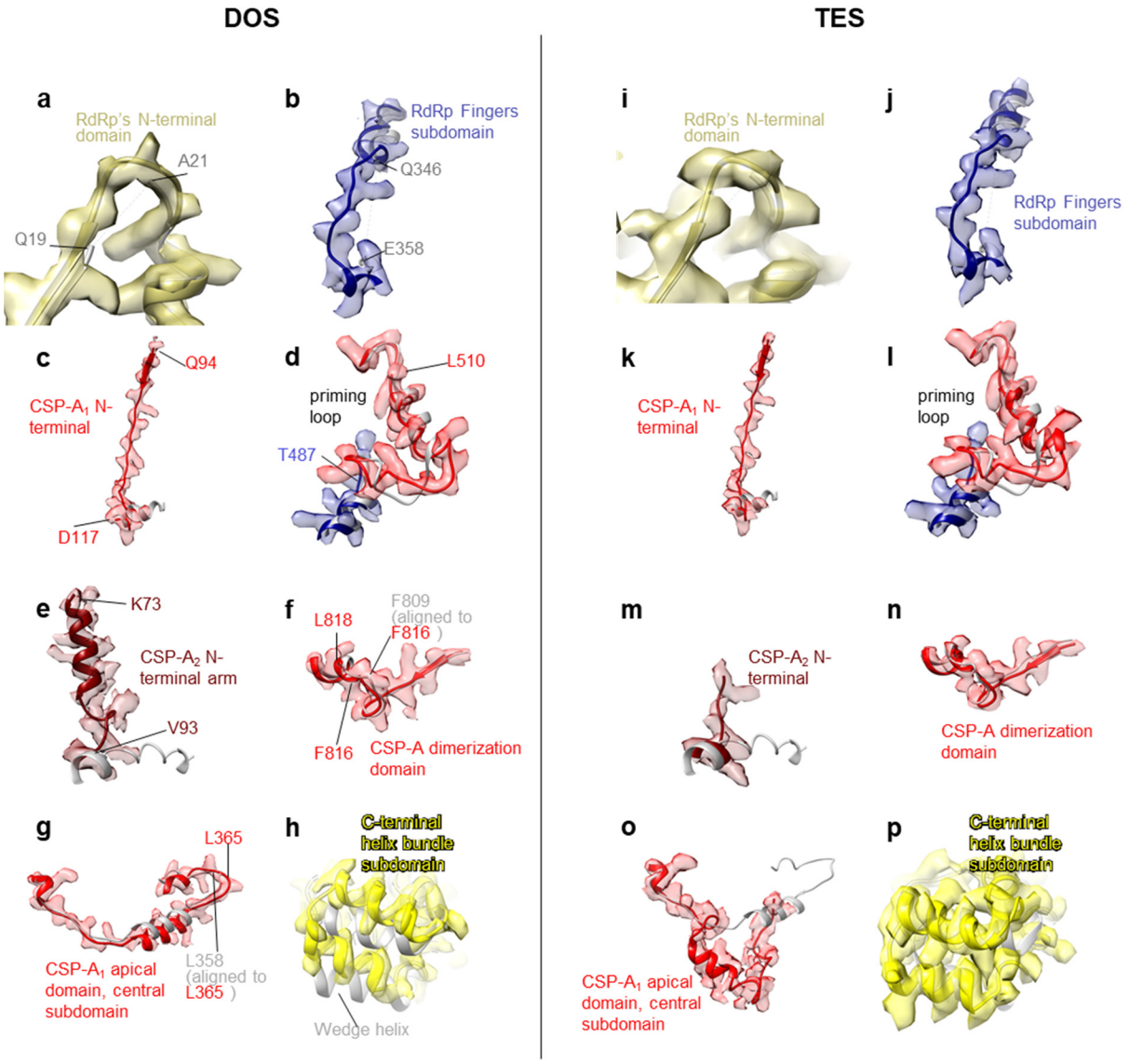


Figure 6.9 ***Key differences between our models and the previous model.*** *Our atomic models (color ribbons) are superposed on the cryoEM density map (semi-transparent) for various fragments (residues indicated) of RdRp and various CSP-A conformers in both DOS (**a-h**) and TES (**i-p**). For comparison, the old model (PDB 2R7Q for RdRp and PDB 4F5X for the capsid) are all shown in grey ribbons, though no previous model existed for the most of the panels shown here.*

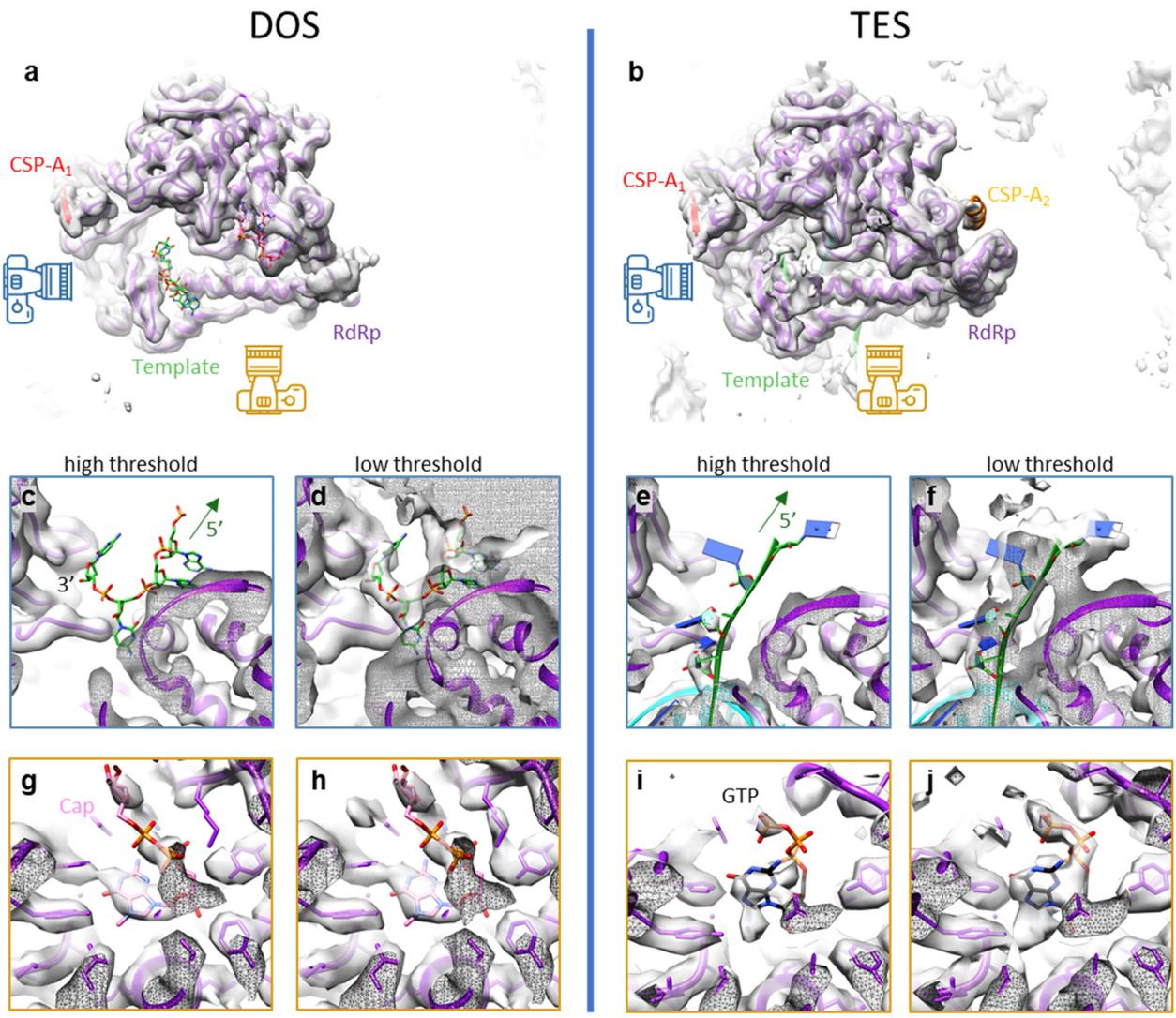


Figure 6.10 ***Template entry and cap-binding site. a, b,*** *RdRp in DOS (**a**) and TES (**b**), shown in the classic top view with models of RdRp (purple) and RNA (ball-and-stick) fit in the Gaussian-filtered map with a 2 Å edge. The camera symbols depict the viewing directions for panels shown in **c-j**. **c-f,** Magnified views (blue camera in **a** and **b**) of the template entrance in DOS (**c**, **d**) and TES (**e**, **f**) at different density thresholds. The template RNA has more flexibility towards 5' end and some features can only be seen at a low density thresholds in the filtered density map. **g-j,** Magnified views (orange camera in **a** and **b**) of the cap-binding site in DOS (**g**, **h**) and TES (**i**, **j**) at two different density thresholds. In DOS, the cap-binding site binds the m7GpppG cap of the (+)RNA strand with high affinity. In TES, the cap-binding site binds a GTP and the feature can be better seen at a lower display threshold. See also Movies 6. 1 and 6.4.*

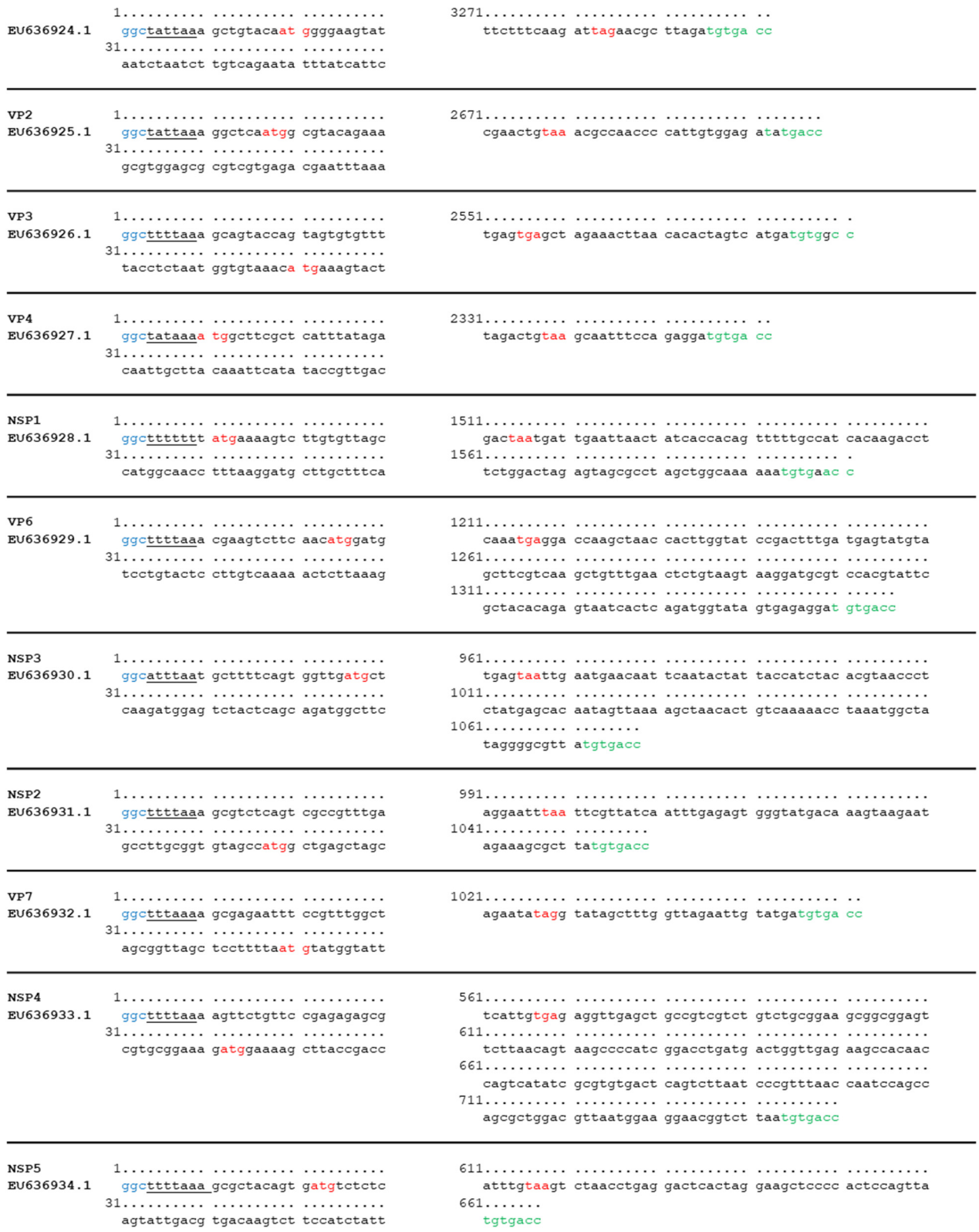


Figure 6.11 ***The 5' and 3' fragments of the 11 genome segments of RRV strain A.*** *The 5' consensus sequence is coloured blue. A/U box is underlined. Start and stop codons are coloured red. The 3' consensus sequence is coloured green.*

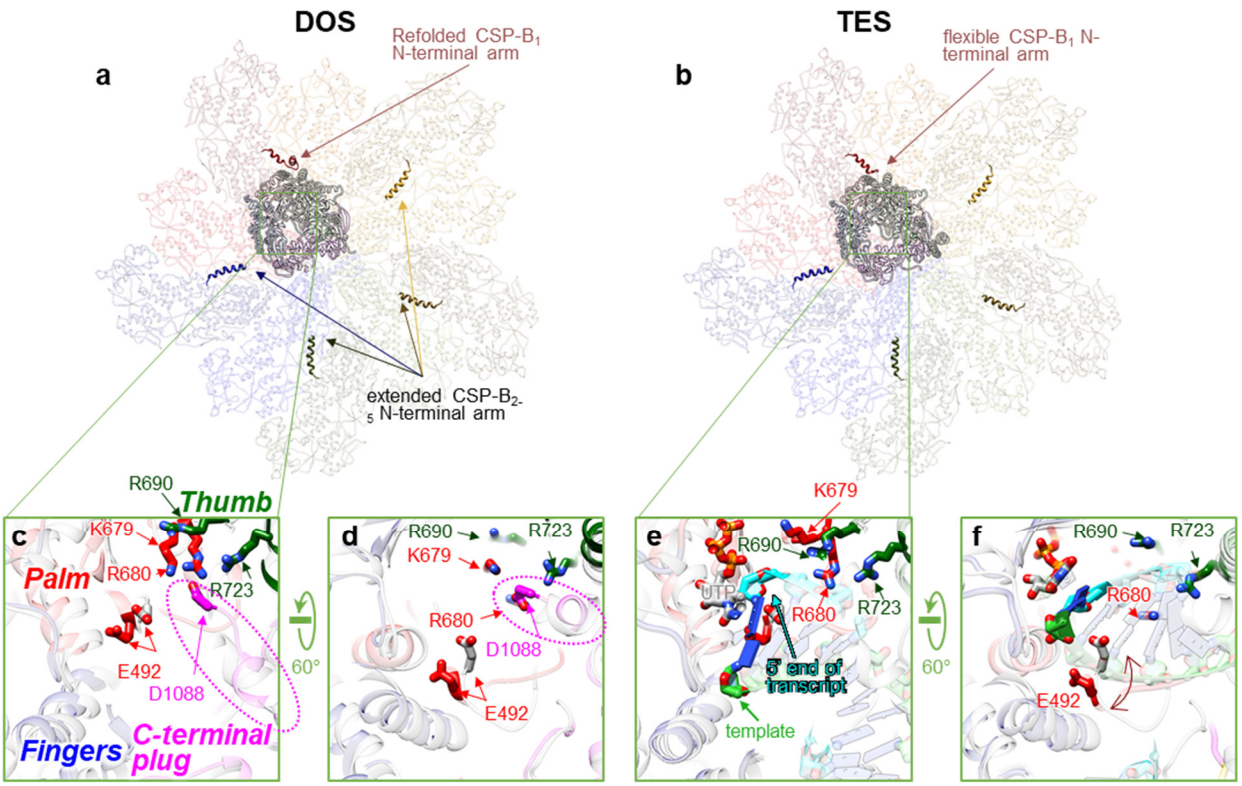


Figure 6.12 ***RdRp's active site and the "priming loop". a,*** *Classical top view of RdRp and CSP decamer at DOS. The 5 N-terminal "tethered arms" (one each from CSP-B$_{1-5}$) are highlighted.* ***b,*** *Classical top view of RdRp and CSP decamer at TES.* ***c, d,*** *Magnified active site in* ***(a).*** *The view in* ***(d)*** *is rotated 60-degrees from the one in* ***(c).*** *A model from crystallography structure (PDB 2R7R, coloured grey) is superimposed. The key residue (E492) on the "priming loop" is labeled to show that the "priming loop" is not extended towards the active site when the RdRp is docked on the capsid. Also, the N-terminal plug (circled, colored in magenta) is inserted nearby.* ***e, f,*** *The same views that* ***(c)*** *and* ***(d)*** *depict, for TES. The C-terminal plug is retracted from the active site while the "priming loop" remains folded at TES. See also Movie 6.5.*

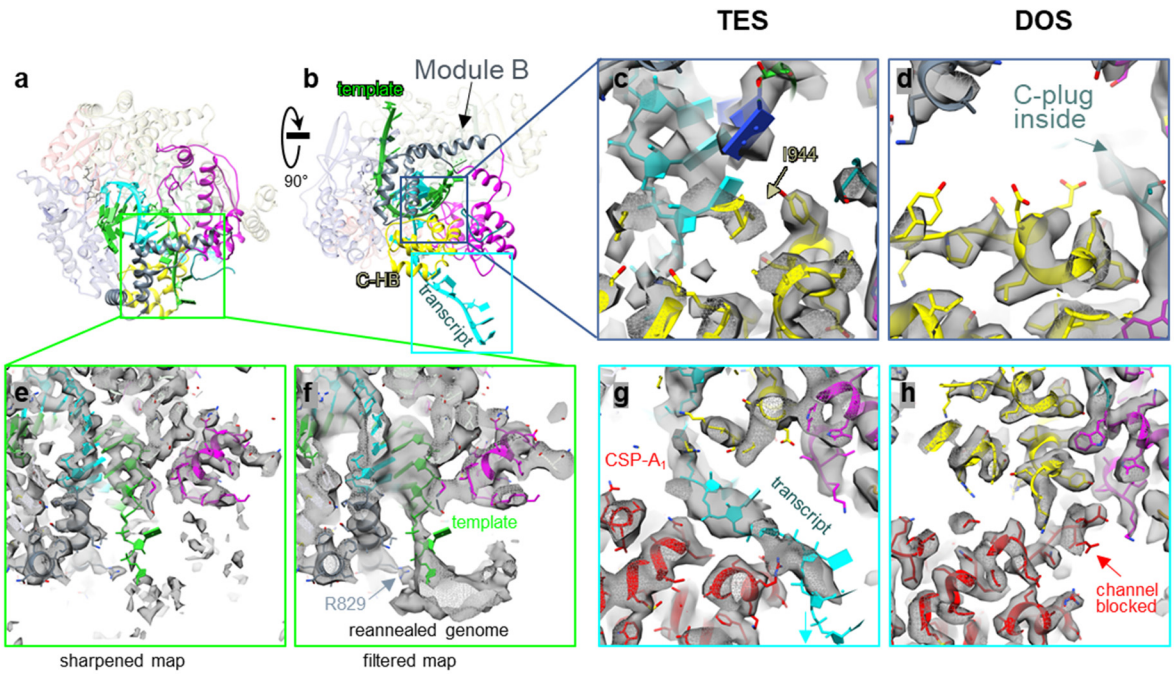


Figure 6.13 ***Quality of the cryoEM densities of RdRp C-terminal domain and RNA in the template and transcript exit channels. a, b,*** *Two orthogonal views of the atomic model of RdRp together with transcript and template RNA strands at TES.* ***c,*** *Magnified from the boxed region in (**b**). Helix bundle subdomain in the C-terminal domain (C-HB) splits the dsRNA product at TES with I944.* ***d,*** *C-HB is retracted, and the C-terminal plug is found inside the template exit channel at DOS.* ***e, f,*** *Details around the template exit in a sharpened map (**e**) and in a low-pass-filtered map (**f**) showing the template strand's reannealing with the coding strand.* ***g, h,*** *Details around the transcript exit in TES (**g**) and DOS (**h**), with the transcript leaving the capsid unobstructed in TES and a CSP-A_1 loop blocking the channel in DOS. See also Movie 6.6.*

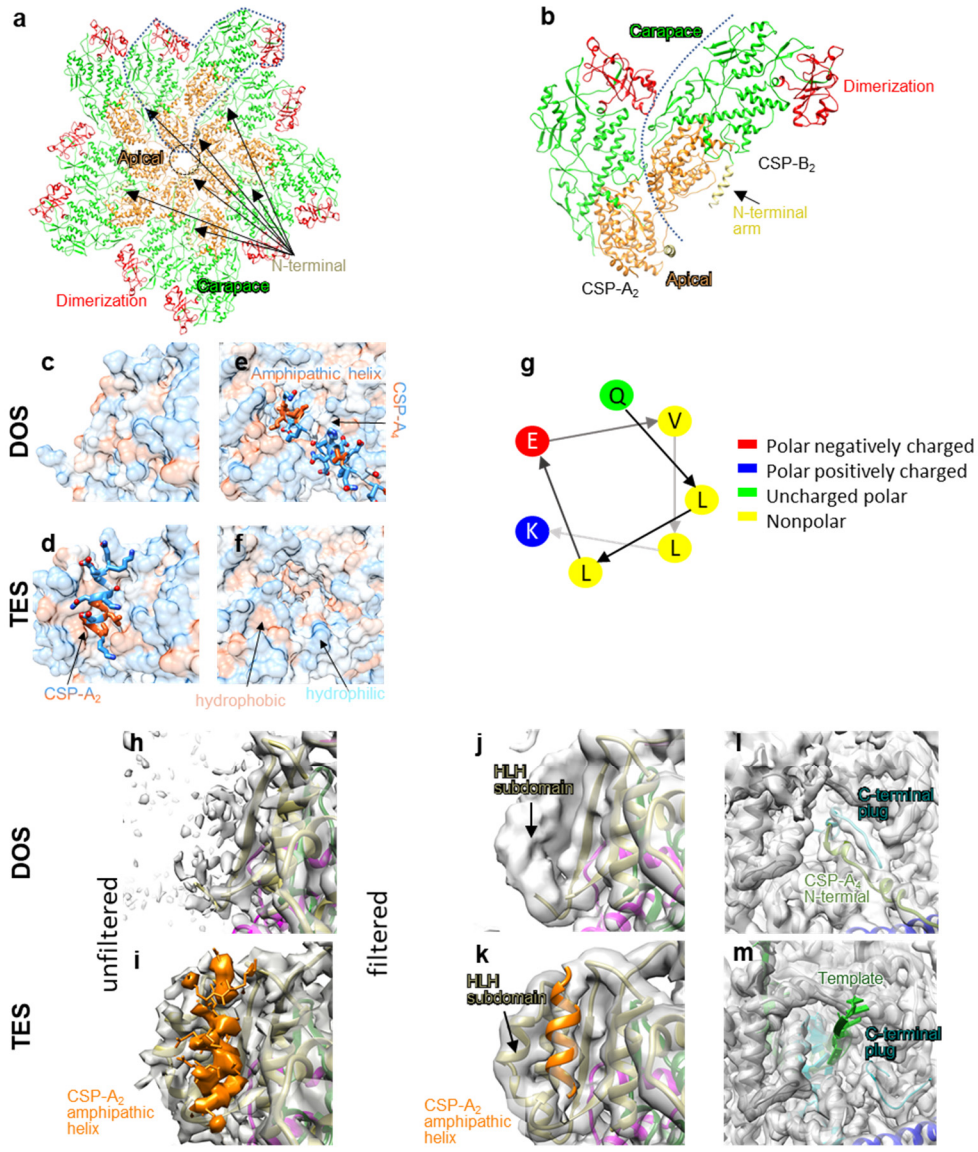


Figure 6.14 ***CSPs and their N-terminal amphipathic helixes acting as transcriptional factors. a,*** *An overview of CSP decamer at TES. Unlike the decamer colour-coded by subunits shown in Figure 6.4**a**, the current display is colour-coded by domains such that the N-terminal domain of CSP and the open hole (circled) at TES can be easily seen.* ***b,*** *Magnified CSP-A_2 and CSP-B_2 from (**a**).* ***c, d,*** *Comparison of the N-terminal domain of RdRp between DOS (**c**) and TES (**d**), showing that CSP-A_2's amphipathic helix binds to a hydrophobic pocket that exists only in TES. This pocket is formed by the helix-loop-helix subdomain at TES. In DOS, this subdomain is flexible and no such hydrophobic pocket exists.* ***e, f,*** *Comparison of the C-HB domain of RdRp between DOS (**e**) and TES (**f**), showing that CSP-A_4's amphipathic helix only binds to C-HB and blocks the template exit channel at DOS. This part of the structure no longer interacts with C-HB at TES and C-HB is translocated.* ***g,*** *Wheel displaying the amphipathic nature of the CSP N-terminal helix.* ***h-m,*** *Detailed detachment/attachment of transcriptional factors on RdRp between the two states, shown with corresponding densities, unfiltered (**h,i**) and filtered (**j-m**). The absence of CSP-A_2's amphipathic helix in DOS (**h,j**) and its presence in TES (**i,k**) suggests that this helix in CSP-A_2 stabilises the HLH subdomain in DOS. The presence of CSP-A_4's amphipathic helix in DOS (**l**) and its absence in TES (**m**) suggests that this same amphipathic helix in CSP-A_4 locks C-HB's conformation in DOS. See also Movie 6.8.*

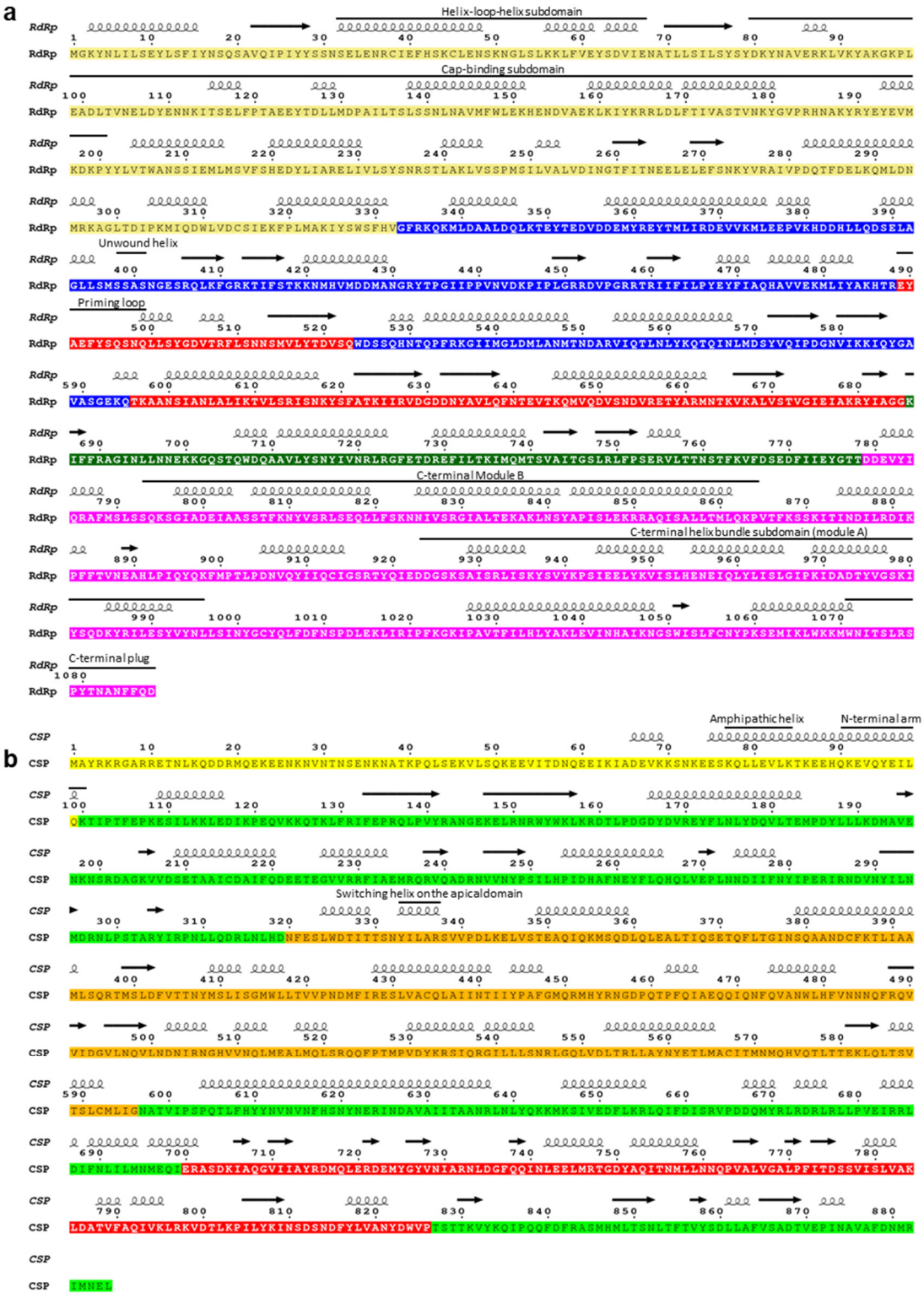


Figure 6.15 ***Secondary structures and domain assignments of RdRp and CSP. a,*** *Sequences of RdRp* ***(a)*** *and CSP* ***(b)*** *along with their secondary structures determined from our cryoEM structures. The sequence is colour-shaped according to domain colours as indicated. See also Movie 6.3 for coloured domains of RdRp.*

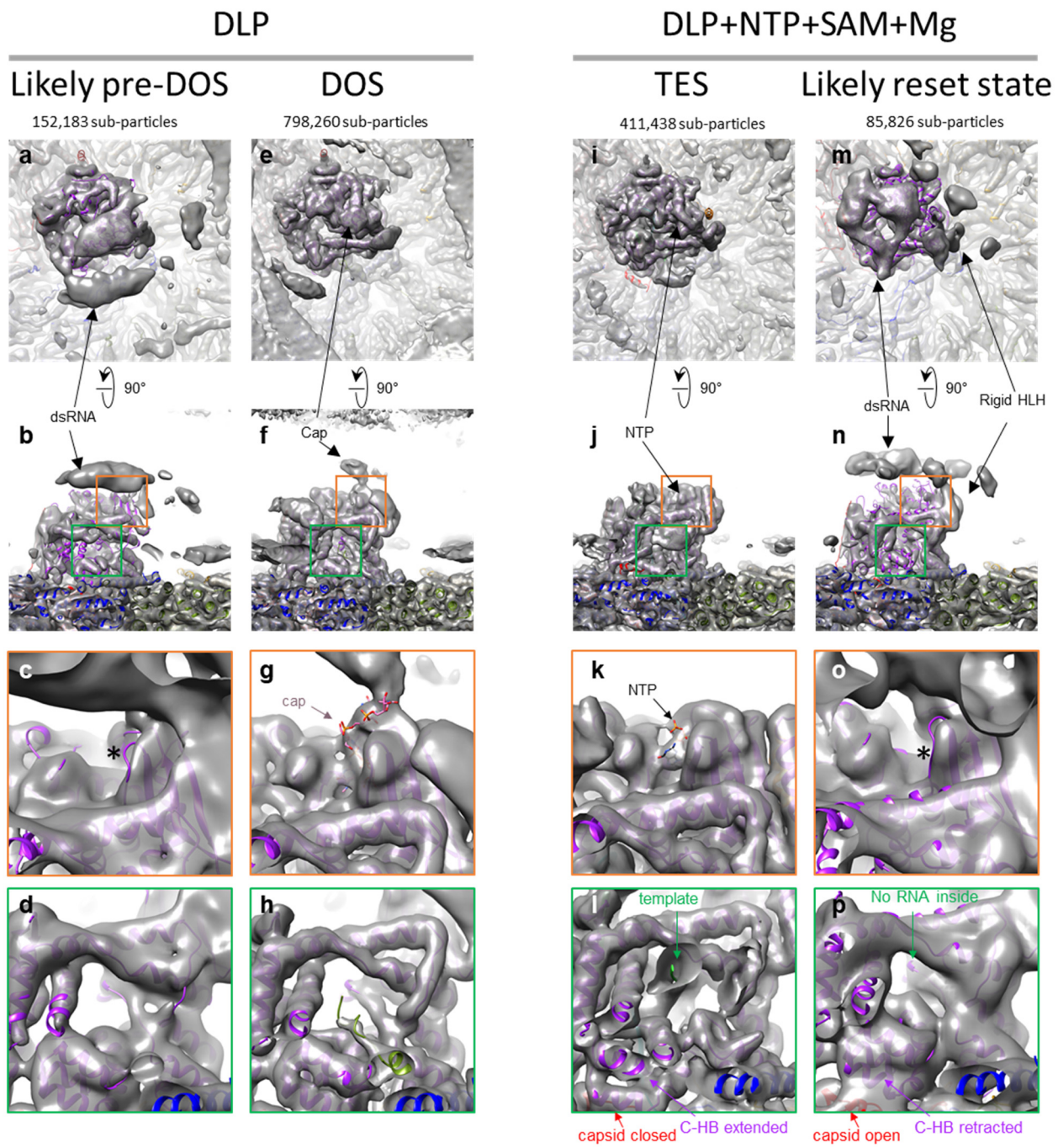


Figure 6.16 ***Identification of other possible states. a, b,*** *Two orthogonal views of a reconstruction (with atomic model superimposed) from classified particle in DLP, likely a pre-DOS.* ***c, d,*** *Magnified cap-binding site (**c**) and template exit channel (**d**) in (**b**). In this state, the RNA remains in the duplex form and there is no NTP/CAP in the cap-binding site [* in (c)].* ***e-h,*** *The same views as (**a-d**), showing corresponding views at DOS.* ***i-l,*** *The same views as (**a-d**), showing corresponding views at TES.* ***m-p,*** *The same views as (**a-d**), showing corresponding views at a state different from TES. This state is likely a reset state. In the putative reset state, there is no RNA density inside RdRp (**p**); the C-terminal helix bundle subdomain of RdRp is retracted from the capsid; the C-terminal plug of RdRp is inserted into the active site; and no NTP/RNA bound to the cap-binding site [* in (**o**)]. Notably, though no NTP/RNA are bound to the cap-binding site, a bulky dsRNA genome density [indicated in (**m**) and (**n**)] similar to that [indicated in (**a**) and (**b**)] also seen in pre-DOS, suggests that this complex is reset to start the transcription again.*

	RdRp in DOS (EMDB-20059) (PDB 6OGY)	RdRp in TES (EMDB-20060) (PDB 6OGZ)
Data collection and processing		
Magnification	130K	105K
Voltage (kV)	300	300
Electron exposure (e–/Å^2)	22	18
Defocus range (μm)	1.4-5.4	1.2-5.4
Pixel size (Å)	1.07	1.33
Symmetry imposed	C1	C1
Initial particle images (no.)	1163388	528852
Final particle images (no.)	798260	411438
Map resolution (Å)	3.4	3.6
FSC threshold 0.143		
Map resolution range (Å)	∞-3.4	∞-3.6
Refinement		
Initial model used (PDB code)	2R7R	2R7R
Model resolution (Å)	3.45	3.63
FSC threshold 0.143		
Map CC	0.8558	0.8616
Model resolution range (Å)	--	--
Map sharpening *B* factor (Å^2)	-160	-160
Model composition		
Non-hydrogen atoms	74364	74962
Protein residues	9081	9081
Nucleotide residues	6	35
Ligands	1	2
B factors (Å^2)		
Protein	56.83	26.96
Nucleotide	131.39	85.24
Ligand	99.28	60.59
R.m.s. deviations		
Bond lengths (Å)	0.006	0.008
Bond angles (°)	0.767	0.882
Validation		
MolProbity score	1.41	1.59
Clashscore	3.56	5.44
Poor rotamers (%)	0.18	0.17
Ramachandran plot		
Favored (%)	96.15	95.77
Allowed (%)	3.84	4.23
Disallowed (%)	0.01	0.00

Table 6.1 Cryo-EM data collection, refinement and validation statistics

6.7 Data availability

The cryoEM density maps have been deposited in the Electron Microscopy Data Bank under accession codes EMD-20059 (DOS) [https://www.ebi.ac.uk/pdbe/entry/emdb/EMD-20059] and EMD-20060 (TES) [https://www.ebi.ac.uk/pdbe/entry/emdb/EMD-20060]. The atomic coordinates have been deposited in the Protein Data Bank under accession codes 6OGY (DOS) and 6OGZ (TES). Other data are available from the corresponding authors upon reasonable request.

6.8 Code availability

Custom-designed programs for particle extraction and orientation selection are deposited in https://github.com/kerichardding/Rotavirus_scripts

6.9 Acknowledgements

We thank P. Ge at the University of California, Los Angeles for initial work in sub-particle averaging. This work was supported in part by grants from the National Institutes of Health (AI094386, GM071940, and DE025567). We acknowledge the use of instruments at the Electron Imaging Center for Nanomachines supported by UCLA and by instrumentation grants from NIH (1S10RR23057, 1S10OD018111, and U24GM116792) and NSF (DBI-1338135 and DMR-1548924). K.D. is a recipient of a Whitcome Fellowship and a Dissertation Year Fellowship from UCLA. We thank Dan Weisman for artistic illustration of Figure 5. We thank Titania Nguyen for editing the manuscript.

6.10 Author Contributions

P.R. and Z.H.Z. initiated the project and designed the experiments. Z.H.Z. supervised the research. C.C.C generated RRV DLP and P.R. prepared the transcribing samples. I.A. made cryoEM grids and X.Z. took cryoEM images. K.D. processed the data and conducted the asymmetric and sub-particle reconstructions. K.D., W.S. and T.L.C. built and refined the atomic models, made figures and movies. K.D. and Z.H.Z. wrote the initial manuscript. W.S. and T.L.C. and P.R. edited the paper. All authors reviewed and approved the paper.

6.11 References

1. Crick, F., *Central Dogma of Molecular Biology.* Nature, 1970. **227**(5258): p. 561-&.
2. de Farias, S.T., et al., *Origin and Evolution of RNA-Dependent RNA Polymerase.* Frontiers in Genetics, 2017. **8**: p. 125.
3. Horning, D.P. and G.F. Joyce, *Amplification of RNA by an RNA polymerase ribozyme.* Proceedings of the National Academy of Sciences of the United States of America, 2016. **113**(35): p. 9786-9791.
4. Gilbert, W., *Origin of Life - the Rna World.* Nature, 1986. **319**(6055): p. 618-618.
5. Joyce, G.F., *The antiquity of RNA-based evolution.* Nature, 2002. **418**(6894): p. 214-221.
6. Hager, A.J., J.D. Pollard, and J.W. Szostak, *Ribozymes: Aiming at RNA replication and protein synthesis.* Chemistry & Biology, 1996. **3**(9): p. 717-725.
7. Ollis, D.L., et al., *Structure of Large Fragment of Escherichia-Coli DNA-Polymerase-I Complexed with Dtmp.* Nature, 1985. **313**(6005): p. 762-766.
8. Jacome, R., et al., *Structural Analysis of Monomeric RNA-Dependent Polymerases: Evolutionary and Therapeutic Implications.* Plos One, 2015. **10**(9).

9. Monttinen, H.A.M., et al., *Automated Structural Comparisons Clarify the Phylogeny of the Right-Hand-Shaped Polymerases.* Molecular Biology and Evolution, 2014. **31**(10): p. 2741-2752.

10. Hansen, J.L., A.M. Long, and S.C. Schultz, *Structure of the RNA-dependent RNA polymerase of poliovirus.* Structure, 1997. **5**(8): p. 1109-1122.

11. Butcher, S.J., et al., *A mechanism for initiating RNA-dependent RNA polymerization.* Nature, 2001. **410**(6825): p. 235-240.

12. Tao, Y.Z., et al., *RNA synthesis in a cage - Structural studies of reovirus polymerase lambda 3.* Cell, 2002. **111**(5): p. 733-745.

13. Lu, X., et al., *Mechanism for Coordinated RNA Packaging and Genome Replication by Rotavirus Polymerase VP1.* Structure, 2008. **16**(11): p. 1678-1688.

14. Gillis, A.J., A.P. Schuller, and E. Skordalakes, *Structure of the Tribolium castaneum telomerase catalytic subunit TERT.* Nature, 2008. **455**(7213): p. 633-U36.

15. Reich, S., et al., *Structural insight into cap-snatching and RNA synthesis by influenza polymerase.* Nature, 2014. **516**(7531): p. 361-+.

16. Gytz, H., et al., *Structural basis for RNA-genome recognition during bacteriophage Q beta replication.* Nucleic Acids Research, 2015. **43**(22).

17. Starnes, M.C. and W.K. Joklik, *Reovirus Protein-Lambda-3 Is a Poly(C)-Dependent Poly(G) Polymerase.* Virology, 1993. **193**(1): p. 356-366.

18. Zhang, X., et al., *In situ structures of the segmented genome and RNA polymerase complex inside a dsRNA virus.* Nature, 2015. **527**(7579): p. 531-+.

19. Liu, H.R. and L.P. Cheng, *Cryo-EM shows the polymerase structures and a nonspooled genome within a dsRNA virus.* Science, 2015. **349**(6254): p. 1347-1350.

20. Patton, J.T., et al., *Rotavirus RNA polymerase requires the core shell protein to synthesize the double-stranded RNA genome.* Journal of Virology, 1997. **71**(12): p. 9618-9626.

21. Makeyev, E.V. and D.H. Bamford, *Replicase activity of purified recombinant protein P2 of double-stranded RNA bacteriophage phi 6.* Embo Journal, 2000. **19**(1): p. 124-133.

22. Collier, A.M., et al., *Initiation of RNA Polymerization and Polymerase Encapsidation by a Small dsRNA Virus.* Plos Pathogens, 2016. **12**(4).

23. Gridley, C.L. and J.T. Patton, *Regulation of rotavirus polymerase activity by inner capsid proteins.* Current Opinion in Virology, 2014. **9**: p. 31-38.

24. Tate, J.E., et al., *Global, Regional, and National Estimates of Rotavirus Mortality in Children < 5 Years of Age, 2000-2013.* Clinical Infectious Diseases, 2016. **62**: p. S96-S105.

25. Ilca, S.L., et al., *Localized reconstruction of subunits from electron cryomicroscopy images of macromolecular complexes.* Nature Communications, 2015. **6**: p. 8843.

26. Rickgauer, J.P., N. Grigorieff, and W. Denk, *Single-protein detection in crowded molecular environments in cryo-EM images.* Elife, 2017. **6**.

27. Estrozi, L.F., et al., *Location of the dsRNA-Dependent Polymerase, VP1, in Rotavirus Particles.* Journal of Molecular Biology, 2013. **425**(1): p. 124-132.

28. Ding, K., L. Nguyen, and Z.H. Zhou, *In Situ Structures of the Polymerase Complex and RNA Genome Show How Aquareovirus Transcription Machineries Respond to Uncoating.* Journal of Virology, 2018. **92**(21).

29. Campagna, M., et al., *RNA interference of rotavirus segment 11 mRNA reveals the essential role of NSP5 in the virus replicative cycle.* Journal of General Virology, 2005. **86**: p. 1481-1487.

30. Juo, Z.S., et al., *How proteins recognize the TATA box.* Journal of Molecular Biology, 1996. **261**(2): p. 239-254.

31. Yu, X.K., et al., *A putative ATPase mediates RNA transcription and capping in a dsRNA virus.* Elife, 2015. **4**.

32. Tortorici, M.A., et al., *Template recognition and formation of initiation complexes by the replicase of a segmented double-stranded RNA virus.* Journal of Biological Chemistry, 2003. **278**(35): p. 32673-32682.

33. Jenni, S., et al., *In situ structure of rotavirus VP1 RNA-dependent RNA polymerase.* bioRxiv, 2019: p. 605063.

34. Patton, J.T., et al., *Virus replication.* Methods Mol Med, 2000. **34**: p. 33-66.

35. Carragher, B., et al., *Leginon: An automated system for acquisition of images from vitreous ice specimens.* Journal of Structural Biology, 2000. **132**(1): p. 33-45.

36. Li, X.M., et al., *Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM.* Nature Methods, 2013. **10**(6): p. 584-+.

37. Rohou, A. and N. Grigorieff, *CTFFIND4: Fast and accurate defocus estimation from electron micrographs.* Journal of Structural Biology, 2015. **192**(2): p. 216-221.

38. Kivioja, T., et al., *Local average intensity-based method for identifying spherical particles in electron micrographs.* Journal of Structural Biology, 2000. **131**(2): p. 126-134.

39. Scheres, S.H.W., *RELION: Implementation of a Bayesian approach to cryo-EM structure determination.* Journal of Structural Biology, 2012. **180**(3): p. 519-530.

40. Liu, Y.T., et al., *CryoEM structures of herpes simplex virus type 1 portal vertex and packaged genome.* Nature, 2019. **in press**.

41. Emsley, P., et al., *Features and development of Coot.* Acta Crystallographica Section D-Biological Crystallography, 2010. **66**: p. 486-501.

42. Adams, P.D., et al., *PHENIX: a comprehensive Python-based system for macromolecular structure solution.* Acta Crystallographica Section D-Biological Crystallography, 2010. **66**: p. 213-221.

43. Pettersen, E.F., et al., *UCSF chimera - A visualization system for exploratory research and analysis.* Journal of Computational Chemistry, 2004. **25**(13): p. 1605-1612.

44. Berman, H., K. Henrick, and H. Nakamura, *Announcing the worldwide Protein Data Bank.* Nature Structural Biology, 2003. **10**(12): p. 980-980.

45. Robert, X. and P. Gouet, *Deciphering key features in protein structures with the new ENDscript server.* Nucleic Acids Research, 2014. **42**(W1): p. W320-W324.

Chapter 7 Conclusion

To understand various heterogeneities found in nanomachines, we first analysed the heterogeneities introduced in cryoEM sample preparation. The surface-dominating nature of the thin buffer film is discussed. It is a blessing and a curse: on one hand, protein adsorbed on the air-water interface are more highly concentrated than in bulk and can reshape the thickness of the buffer film close to the size of the biomolecule, thus increasing both the quantity and quality of the dataset; on the other hand, the interface reshapes the biomolecule and causes deformation, denaturing, and preferred orientation issues, which all increase difficulties for the subsequent averaging steps in the single particle analysis. To alleviate interface effects on the biomolecule, we discussed the effect of surfactants. The surfactant should be strong enough to block biomolecules from the interface, mild enough to keep biomolecules from denaturing, and have a specific molecular structure to reach high CMC. In conclusion, the air-water interface can introduce dramatic unnatural heterogeneities to biomolecules.

To further understand heterogeneities, we first resolved the solution structure of an engineered vault complex, which shows great spatial heterogeneity (flexibility). From the two conformations found on the same grid, we confirmed that cryoEM provides enough contrast to distinguish subtle conformational changes. By applying masks in real space, conformational changes within the area of interest are resolved.

We then resolved another spatial heterogeneity (symmetry mismatching) in CPV by relaxing the symmetry and focusing on a specific spectrum. The resulting asymmetric reconstruction shows that each CPV particle carries 10 RdRps, each one of which corresponds to one of the 10 segments of the genome. This is a very early approach to resolving asymmetric

structures by alleviating the influence of a highly symmetric capsid, a major breakthrough in virus structure.

After confirming our ability to resolve spatial heterogeneity (symmetry mismatching), we introduced different states to other reoviruses to study state transitions in these nanomachines. In ARV, by comparing *in situ* structures of the quiescent and primed states, we proposed a mechanism of virus activation and capsid assembly. In rotavirus, by comparing *in situ* structures of the primed and transcribing states, we proposed a mechanism of transcription and replication. Additionally, we found two more intermediate states in rotavirus through classification, not through pausing the nanomachines in a specific state. By resolving spatial heterogeneities (symmetry match) and temporal heterogeneities (biomolecule at different states) at the same time, the working mechanism of reovirus has been thoroughly studied. Resolving these increased heterogeneities requires more specific data processing tools, such as masked classification and localized reconstruction. Because all the conformational changes observed are inside the capsid, which is away from the air-water interface, nanomachine heterogeneities observed *in situ* are very likely to be natural and biologically relavent.

To summarize, avoiding unnecessary heterogeneities in the cryoEM sample preparation step is preferred but difficult. In order to further resolve heterogeneity, it is important to localize into smaller areas of interest, where a larger SNR can help to distinguish subtle features from noise. By localizing into smaller areas, the number of possible classes in each area will dramatically decrease, since conformational changes in separate areas decouple. Because resolving heterogeneities is largely dependent on which biomolecular area we focus on, and the level of heterogeneity varies among different samples, the data processing tool applied to each biomolecule needs to be adjusted individually and cannot be fully automatic. Our four different examples of

resolving heterogeneities suggest a key feature of automatic heterogeneity resolving tools in the future: locating the area of interest automatically instead of forcing outliers into presumed shared averages. With this tool, the dynamics of nanomachines can be resolved by cryoEM with better accuracy and higher speed.

fin