UNIVERSITY OF CALIFORNIA,
IRVINE

Combination of protein crosslinking mass spectrometry with protein structure prediction for molecular structure modeling of vaccinia virus

DISSERTATION

Submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Biological Sciences

by

Yeva Mirzakhanyan

Dissertation Committee:
Professor Paul D. Gershon, Chair
Professor Bert L. Semler
Assistant Professor Reginald McNulty
Assistant Professor Shane Gonen

2023

DEDICATION

To my parents, Levon and Karine Mirzakhanyan, who left their home, their families, and everything they knew and moved to the USA to give their two daughters (and later son) a chance at a brighter future. I couldn’t have done this without you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Paul Gershon, for the many years of mentorship and support you have shown me. I joined Dr. Gershon’s lab nearly 10 years ago as an undergraduate Bio 199, with the aspiration of studying vaccinia virus and to work towards answering some of the many unresolved questions in the poxvirus field. You have always pushed me to be a better scientist, a more critical thinker, and a better writer. I wouldn’t be the person I am today without this.

I would also like to thank the members of my dissertation committee, Dr. Bert Semler, Dr. Reginald McNulty, and Dr. Shane Gonen for the invaluable support and advice you have given me. I also thank the UCI Center for Virus Research, Rebecca Taylor, and Crystal Spitale for all of their help over the past few years.

Thanks to all of my amazing labmates: Tuan, who taught me how to prepare my first Stagetip. Jun (who is sadly no longer with us) and Jesper, Christine, who has been an incredibly supportive friend over the last 2.5 years (almost 3), Verna, from our neighboring lab, and all our amazing undergrads.

This work presented here was supported in part by the National Institutes of Health grant 1R01GM134144-01, and National Institutes of Health instrumentation grant S10 OD016328, T32AI007319 Virus-host interactions: a multi-scale training program.

VITA

Yeva Mirzakhanyan

Education

2018-2023	Doctor of Philosophy in Biological Sciences. University of California, Irvine Dept. of Molecular Biology and Biochemistry Advisor: Paul D. Gershon, Ph.D.
2017-2018	Master of Science in Biological Sciences. University of California, Irvine Dept. of Molecular Biology and Biochemistry Advisor: Paul D. Gershon, Ph.D.
2010-2014	Bachelor of Science in Biological Sciences. University of California, Irvine Dept. of Molecular Biology and Biochemistry Advisors: Paul D. Gershon, Ph.D., Andrej Lupták, Ph.D.

Research Experience

2018-2023	Graduate Research Assistant, University of California, Irvine
2014-2018	Research Assistant, University of California, Irvine
2013-2014	Undergraduate Student Researcher, University of California, Irvine

Fellowship and Awards

2023	Edward K. Wagner Memorial Award in Virology, School of Biological Sciences, University of California, Irvine
2020-2022	Center for Virus Research T32 Graduate Fellowship
2018-2019	Center for Virus Research T32 Graduate Fellowship
2018	Edward K. Wagner Memorial Award in Virology, School of Biological Sciences, University of California, Irvine
2014	Poster Award, Excellence in Research, School of Biological Sciences, University of California, Irvine

Publications

1. Cayabyab F, Tipirneni J, Chen D, Choi J, Tacto C, Wu J, Wang L, Mirzakhanyan Y, Gershon PD, Fang S, Ipp E, de Aguiar Vallim TQ, Chen LF, Wei Z, Yoshihara E. BET signaling and FXR signaling orchestrate to protect β cells. 2023. Manuscript in preparation.
2. Mirzakhanyan Y and Gershon PD. Addressing structural hierarchies among the membrane and surface proteins of the Vaccinia virion via a combination of deep protein-protein crosslinking and deep learning-based structure prediction. 2023. Under review.
3. Mirzakhanyan Y, Jankevics A, Scheltema RA, Gershon PD. Combination of deep

XLMS with deep learning reveals an ordered rearrangement and assembly of a major protein component of the vaccinia virion. mBio. 2023 Aug 30:e0113523. doi: 10.1128/mbio.01135-23. Epub ahead of print. PMID: 37646531.

4. Ziegler ME, Sorensen AM, Banyard DA, Sayadi LR, Chnari E, Hatch MM, Tassey J, Mirzakhanyan Y, Gershon PD, Hughes CCW, Evans GRD, Widgerow AD. Deconstructing Allograft Adipose and Fascia Matrix: Fascia Matrix Improves Angiogenesis, Volume Retention, and Adipogenesis in a Rodent Model. Plast Reconstr Surg. 2023 Jan 1;151(1):108-117. doi: 10.1097/PRS.0000000000009794. Epub 2022 Oct 11. PMID: 36219861; PMCID: PMC10081826.
5. Weghorst F, Mirzakhanyan Y, Hernandez KL, Gershon PD, Cramer KS. Non-Apoptotic Caspase Activity Preferentially Targets a Novel Consensus Sequence Associated With Cytoskeletal Proteins in the Developing Auditory Brainstem. Front Cell Dev Biol. 2022 Mar 7;10:844844. doi: 10.3389/fcell.2022.844844. PMID: 35330912; PMCID: PMC8940215.
6. Mirzakhanyan Y, Gershon PD. Structure-Based Deep Mining Reveals First-Time Annotations for 46 Percent of the Dark Annotation Space of the 9,671-Member Superproteome of the Nucleocytoplasmic Large DNA Viruses. J Virol. 2020 Nov 23;94(24):e00854-20. doi: 10.1128/JVI.00854-20. PMID: 32999026; PMCID: PMC7925184.
7. Weghorst F, Mirzakhanyan Y, Samimi K, Dhillon M, Barzik M, Cunningham LL, Gershon PD, Cramer KS. Caspase-3 Cleaves Extracellular Vesicle Proteins During Auditory Brainstem Development. Front Cell Neurosci. 2020 Nov 12;14:573345. doi: 10.3389/fncel.2020.573345. PMID: 33281555; PMCID: PMC7689216.
8. Pal S, Mirzakhanyan Y, Gershon P, Tifrea DF, de la Maza LM. Induction of protection in mice against a respiratory challenge by a vaccine formulated with exosomes isolated from Chlamydia muridarum infected cells. NPJ Vaccines. 2020 Sep 18;5:87. doi: 10.1038/s41541-020-00235-x. PMID: 33014435; PMCID: PMC7501220.
9. Liu R, Olano LR, Mirzakhanyan Y, Gershon PD, Moss B. Vaccinia Virus Ankyrin-Repeat/F-Box Protein Targets Interferon-Induced IFITs for Proteasomal Degradation. Cell Rep. 2019 Oct 22;29(4):816-828.e6. doi: 10.1016/j.celrep.2019.09.039. PMID: 31644906; PMCID: PMC6876622.
10. Sandoval R, Boyd RD, Kiszter AN, Mirzakhanyan Y, Santibańez P, Gershon PD, Hayes ML. Stable native RIP9 complexes associate with C-to-U RNA editing activity, PPRs, RIPs, OZ1, ORRM1 and ISE2. Plant J. 2019 Sep;99(6):1116-1126. doi: 10.1111/tpj.14384. Epub 2019 Jun 26. PMID: 31077462; PMCID: PMC6744336.
11. Mirzakhanyan Y, Gershon P. The Vaccinia virion: Filling the gap between atomic and ultrastructure. PLoS Pathog. 2019 Jan 7;15(1):e1007508. doi: 10.1371/journal.ppat.1007508. PMID: 30615658; PMCID: PMC6336343.
12. Mirzakhanyan Y, Gershon PD. Multisubunit DNA-Dependent RNA Polymerases from Vaccinia Virus and Other Nucleocytoplasmic Large-DNA Viruses: Impressions from the Age of Structure. Microbiol Mol Biol Rev. 2017 Jul 12;81(3):e00010-17. doi: 10.1128/MMBR.00010-17. PMID: 28701329; PMCID: PMC5584312.

13. Ngo T, Mirzakhanyan Y, Moussatche N, Gershon PD. Protein Primary Structure of the Vaccinia Virion at Increased Resolution. J Virol. 2016 Oct 14;90(21):9905-9919. doi: 10.1128/JVI.01042-16. PMID: 27558425; PMCID: PMC5068539.

Posters

1. Mirzakhanyan Y, Jankevics A, Scheltema RA, **Gershon PD***. Combination of XLMS with deep learning-based protein structure prediction: Protein structure, dynamics, processing and higher order assembly in Vaccinia Virus. Symposium on Structural Proteomics. [2023].

2. Mirzakhanyan Y*, Gershon PD. Contributions of Atomic Models Towards the Molecular Modeling of a Large DNA Virion. UCI CVR Symposium. [2022] Oral.

3. Mirzakhanyan Y*, Albanese P, Jankevics A, Steigenberger B, Scheltema R.A, Gershon PD. The structural proteome of a large DNA virus Vaccinia through Chemical Crosslinking-Mass Spectrometry. UCI CVR Retreat. [2022] Poster.

4. Mirzakhanyan Y*, Gershon PD. Structural virology: Investigating the Molecular Structure of the Vaccinia Virion. UCI CVR Symposium. [2021] Oral.

5. Mirzakhanyan Y, Albanese P, Jankevics A, Steigenberger B, Scheltema R.A, Gershon PD*. The structural proteome of a large DNA virus Vaccinia. American Society for Mass Spectrometry Annual Conference. [2021] Poster.

7. Mirzakhanyan Y*, Gershon PD. Integrated Structural Proteomics and Dynamics of a Solid-Body Organism by Combined XLMS, Solvent Accessible Surface Modification and QconCAT. American Society for Mass Spectrometry Annual Conference. [2019] Poster.

8. Mirzakhanyan Y, Gershon PD*. To what extent can the fitting of biological structure to biological function be formalized? American Society for Mass Spectrometry Annual Conference. [2018] Poster.

9. Mirzakhanyan Y, Ngo T, Gershon PD*. Exploring the protein-protein chemical crosslinking of a highly non-iso-stoichiometric protein complex. American Society for Mass Spectrometry Annual Conference. [2015] Poster

10. Mirzakhanyan Y*, Misek J, Luptak A, Gershon PD. In vitro selection of single stranded DNA aptamers that bind tyrosine and serine phosphorylated peptides. Excellence in Research Symposium, University of California Irvine. [2014] Poster.

Teaching Experience

2023	Graduate Student Teaching Assistant, University of California, Irvine
2019	Graduate Student Teaching Assistant, University of California, Irvine

ABSTRACT OF THE DISSERTATION

Combination of protein crosslinking mass spectrometry with protein structure prediction for molecular structure modeling of vaccinia virus

by

Yeva Mirzakhanyan

Doctor of Philosophy in Biological Sciences

University of California, Irvine, 2023

Professor Paul D. Gershon, Chair

Poxvirus molecular structure has been a long-standing problem in structural virology. Poxviruses are large, enveloped, double stranded DNA viruses that complete their entire replication cycle in the host cytoplasm. Despite >60 years of investigation into the ultra- and molecular structure of vaccinia virus, the prototypical poxvirus, various factors have rendered its molecular structure refractory to traditional structural and molecular biology approaches. This dissertation describes new insights into the vaccinia virion molecular structure, achieved by protein crosslinking mass spectrometry (XLMS) combined with protein structure prediction by deep learning. To examine the vaccinia molecular structure *in situ*, we developed an XLMS approach implementing a “strategy of variation” by which we were able identify protein-protein interaction interfaces for almost all of the ~75 packaged virion proteins. As a part of this process, we developed an updated virus purification protocol for high yield-high vaccinia virus with proteomic purity, along with new methods for full vaccinia protein solubilization for proteomics, that was able to significantly improve the completeness of digestion of virion proteins to peptides for mass spectrometry. We also implemented bioinformatic and other strategies to maximize the

identification of crosslinked peptides, finally allowing us to achieve what we consider likely to be a saturating XLMS dataset.

Alongside this work, we applied a structural homology prediction approach (HHsuite) to identify homologs of vaccinia proteins and to hopefully generate “placeholder” models that we could integrate with XLMS to generate higher order three dimensional models. We applied structural homology prediction to other viruses from the nucleocytoplasmic large DNA virus (NCLDV) phylum to the identify orthologous genes between the virus families that were undetectable by sequence homology alone. This contributed substantially to expanding annotations of previously uncharacterized proteins from NCLDV proteomes. Confident structural homologs for core structural proteins of the vaccinia virion, however, were not identified. Instead, we eventually pursued protein structure prediction by AlphaFold2 to generate high confidence models of vaccinia virion proteins to combine with XLMS data.

Through the resulting combination of XLMS based structure validation of AlphaFold2 models and crosslink-guided protein docking, we describe previously unidentified higher order structural assemblies of vaccinia virion transmembrane and surface-associated proteins. The density of crosslinks within and between molecules of Vaccinia virion core structural P4a (the major component of the palisade layer of the core wall) allowed us to address in greater depth its structure, maturation, and possible assembly pathway. This work was able to identify a likely order of events in which P4a precursor is first proteolytically processed at a downstream site to release its P4a-3 fragment. Removal of P4a-3 releases a steric block to a major rearrangement between the two domains of P4a-1, resulting in a final conformation that is stabilized by disulfide-locking. Removal of a second proteolytic fragment, P4a-2 and trimerization of P4a-1 then allows the assembly of P4a-1 trimers into a higher order hexagonal lattice as the palisade

layer of the Vaccinia virus core wall. Overall, these findings contributed to our understanding of the Vaccinia mature virion molecular architecture and open the door to future studies of virion dynamics and morphogenesis. This approach was then extended to the 23 transmembrane and virion surface proteins associated with the virion envelope.

CHAPTER 1

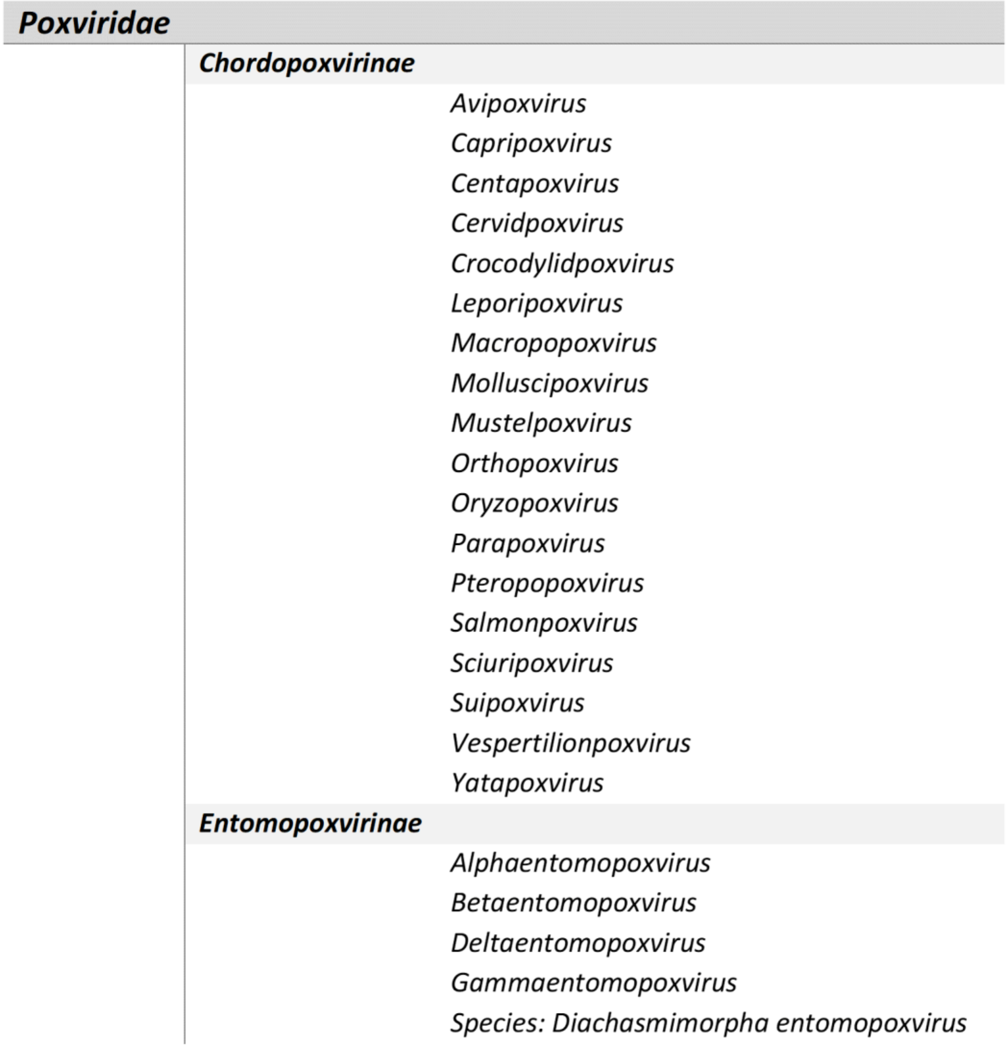
Introduction

Brief overview of poxvirus disease, classification, vaccines and laboratory tools

Members of the poxvirus family have had a profound impact on humanity, as both a cause of great affliction and one of the greatest medical discoveries of all time. The development of a smallpox vaccine from cowpox-like virus by Edward Jenner in 1796 [1, 2] - with subsequent improvements providing safer, more effective, vaccinia virus-based vaccines – culminated in a worldwide vaccination campaign and the complete eradication of smallpox, a disease that has accounted for up to 10% of global annual mortality.

The *Poxviridae* are large viruses with double-stranded DNA genomes that complete their entire replication cycle within the cytoplasm of the host cell. They are classified into two subfamilies, namely the *Chordopoxvirinae*, which infect vertebrates (mammals, birds, fish, and reptiles), and *Entomopoxvirinae*, which infect arthropods [3]. The two subfamilies decompose into multiple genera (**Table 1.1**). Four of the eighteen chordopoxvirus genera (**Table 1.1**) - *Orthopoxvirus*, *Yatapoxvirus*, *Molluscipoxvirus*, and *Parapoxvirus* [4, 5] - contain known causative agents of human disease [6]. Disease caused by members of the latter three genera (orf virus, yaba monkey tumor virus, and molluscum contagiosum virus) is typically only mild and dermatological, and typically resolves without medical intervention [5, 7-9]. Among the orthopoxviruses, however, are key pathogens and immunogens of humans, namely variola virus, monkeypox virus, vaccinia virus, and cowpox virus. The former two cause mild-to-fatal illness depending on the virus species and strain. By contrast, the latter two have proven among the most effective vaccines ever used in humans, leading to the only documented eradication, to date, of a human disease (smallpox): Following a global vaccination campaign with vaccinia, smallpox was declared eradicated in 1980 [10]. Prior to this, variola virus had been the single deadliest human virus in terms of global mortality with an estimated 400 million deaths in the 20th century alone [1, 2, 10]. "Ordinary-type" smallpox ("variola major") represented 89% of reported cases with a case fatality rate (CFR) of 30% [1, 10]. Prior immunization with vaccinia

Table 1.1. Taxonomy of the *Poxviridae**



Poxviridae		
	Chordopoxvirinae	
		Avipoxvirus
		Capripoxvirus
		Centapoxvirus
		Cervidpoxvirus
		Crocodylidpoxvirus
		Leporipoxvirus
		Macropopoxvirus
		Molluscipoxvirus
		Mustelpoxvirus
		Orthopoxvirus
		Oryzopoxvirus
		Parapoxvirus
		Pteropopoxvirus
		Salmonpoxvirus
		Sciuripoxvirus
		Suipoxvirus
		Vespertilionpoxvirus
		Yatapoxvirus
	Entomopoxvirinae	
		Alphaentomopoxvirus
		Betaentomopoxvirus
		Deltaentomopoxvirus
		Gammaentomopoxvirus
		Species: Diachasmimorpha entomopoxvirus

* 2023 summary profile of the *Poxviridae* family by the International Committee on Taxonomy of Viruses [3]

virus lowered smallpox CFR to 0 - 3% depending on the potency of the particular vaccine administered. Vaccinia also provided immunity to variola minor, a less virulent strain of variola prevalent in the Americas [2, 11]. Vaccination was ineffective, however, against rare smallpox types (“flat” or “hemorrhagic”), which almost always resulted in death (95 - 100% CFR; [1]).

Following its discovery in 1970, zoonotic monkeypox virus has emerged as a significant epidemiological concern. Largely since the eradication of smallpox, monkeypox disease has risen steadily, concurrent with waning global smallpox immunity [12, 13]. Monkeypox outbreaks have been largely restricted to central and west Africa where it would arise locally via animal-to-human transmission. However, in the 2022-23 outbreak (which is still not entirely resolved) monkeypox became truly global for the first time with more than 90,000 confirmed cases and clear human-to-human transmission. Fortuitously, however, the 2022-23 outbreak clade showed a 10-fold lower mortality rate than the earlier outbreaks [14, 15].

While first generation vaccinia virus vaccine (used in the eradication of smallpox) was based on animal-derived virus, after 9/11/2001 a second generation vaccinia virus-based vaccine (“ACAM2000”) was developed and stockpiled, based on plaque-purified vaccinia generated in cell culture. A highly attenuated third generation vaccinia virus vaccine (“MVA-BN”) is based on the non-replicating (in humans) MVA strain of vaccinia. Vaccination with second and third generation vaccines can prevent 85% of monkeypox virus infections [16-19], slowing the spread of monkeypox virus among susceptible individuals, and assisting in the treatment of infected individuals [18, 20, 21]. Having a wider safety profile than ACAM2000, the 3rd generation vaccine extends protection to even immunocompromised individuals [22-24].

The post-eradication era has also been marked by the challenge of “feral" (wild) vaccinia virus itself rising to fill the immunity gap. Vaccinia virus infections, either naturally occurring (through animal-to-human transmission) or by vaccination with second generation live-replicating vaccines (e.g., ACAM2000), are typically self-limiting with only mild flu-like symptoms and vesicular lesions at the site of infection or inoculation that resolve over the course of a month [25]. Serious adverse events from vaccinia virus are rare but occur more frequently in immunocompromised individuals [26, 27].

Due to its safety as a globally proven vaccine (above) vaccinia virus has become the model of choice for laboratory studies of poxvirus replication and basic biology. This role is underscored by the high degree of structural and genomic conservation among the *Poxviridae* [1, 28], with all poxviruses sharing, for example, a core set of 49 conserved genes [29, 30]. Vaccinia's relative safety, broad host range and adaptability to tissue culture makes it also a valuable research tool for recombinant gene expression and recombinant vaccine development, oncolytic cancer therapy, and study of the host-pathogen interface [31-35].

Evolutionary history of poxviruses: The NCLDV

The evolutionary origin of the *Poxviridae* has been linked to the switch from protist to animal hosts which occurred approximately 500 million years before present [36-38], around the time of the Cambrian explosion. This would have been followed by the separation of the *Entomopoxvirinae* from the *Chordopoxvirinae* [36]. Further diversification within the *Chordopoxvirinae* would have led to divergence of the orthopoxviruses from their most recent common ancestor [39, 40].

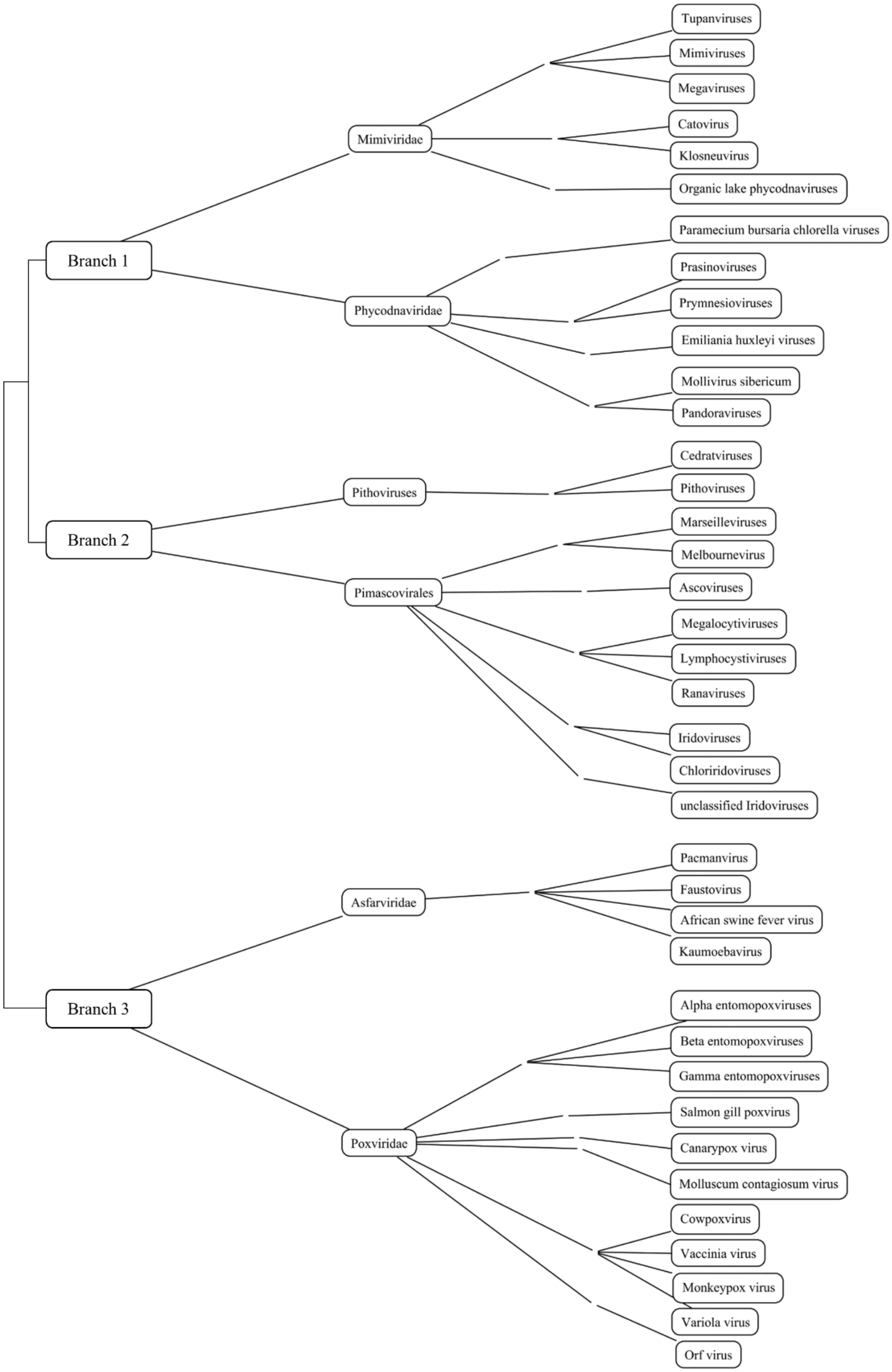
The *Poxviridae* can be placed within an even broader context, namely that of the Nucleocytoviricota phylum [41, 42] – commonly referred to as the nucleocytoplasmic large DNA viruses (NCLDV). The NCLDV comprise a group of highly diverse dsDNA viruses which share a core set of 40 recognizably conserved genes required for genome replication, transcription, and virion morphogenesis [36, 42, 43]. The NCLDV have been suggested to have a common evolutionary origin, speculated to be rooted in an ancient virus that infected protists [44]. Many NCLDV have clinical, agricultural, and environmental significance. Notably, the NCLDV do not encompass all large dsDNA viruses: The herpesviruses and baculoviruses, for example, are evolutionarily distinct and thus are not included [44].

The NCLDV radiate as three discrete branches (Table 1.2). Branch 3 comprises the *Poxviridae* and the *Asfarviridae* [36], the latter comprising African swine fever virus (a highly infectious and deadly virus of domestic pigs and wild boar [45]) along with three unclassified protist-infecting viruses: faustovirus, pacmanvirus, and kaumoebavirus [46-48]. The latter three having moderate to large genomes, with faustovirus and pacmanvirus approaching giantism (currently arbitrarily defined as genomes > 500 kilobase pairs in length [36]).

Branch 1 of the NCLDV includes the virus families *Mimiviridae* and *Phycodnaviridae* [42] (Table 1.2). Viruses from these two families exclusively infect protists [44] and their genomes vary in size between 350 - 400 kb (the Organic Lake phycodnaviruses) and 1.5 – 2.5 Mb genomes (tupanvirus deep ocean and pandoravirus salinus respectively) [49, 50], the latter being considered “giants”. Discovery of these giant viruses at the start of the 21st century overturned the classical assumption that all viruses were “filterable agents”, as many of these giant viruses are larger in overall dimensions than some small bacteria [50, 51]. Some members of the *Phycodnaviridae* are highly ecologically significant in being the primary cause of algal bloom collapse [52, 53].

Branch 2 of the NCLDV has the greatest diversity among the three branches and includes the *Pithoviridae* along with viruses in the Pimascovirales order, namely the *Marseilleviridae*,

Table 1.2. Viral families and representative members of the Nucleocytoviricota phylum.



Ascoviridae, and *Iridoviridae* families. [42, 54]. The latter two families having apparently switched from protist to insect, fish, amphibian, and crustacean hosts [55, 56], where they can often cause fatal disease [57-59]. Despite substantial morphological similarity between pithoviruses and the pandoraviruses from branch 1 [54], their genomes belie only a distant evolutionary relationship [42, 44, 54].

Branches 1 and 2 viruses are grouped phylogenetically relatively closely, based on sequence similarities within the 40 core genes, while branch 3 viruses remain phylogenetically distinct under this scheme [44]. The primary drivers of adaptation among the NCLDV appear to be gene duplication and genome contraction events [60]. Gene gain and loss is most notable in the context of the viral multisubunit DNA-directed RNA polymerases that are encoded by the majority of NCLDV viruses. We found that, in contrast to cellular RNA polymerases (whose regulation and adaptation have been driven by complex arrays of more peripheral protein factors), the viral enzymes have undergone significant refinement within the core enzyme itself showing structural and functional specialization within individual viral families [61]. These adaptations to the viral DNA-dependent RNA polymerases will be discussed in Chapter 2 (in which we find many more virus-encoded subunits than previously known, by searching for structural as opposed to sequence homology) with later updates (the finding of additional subunits and structural homologs of basal transcription factors) in Chapter 3.

Analysis of gene localization in NCLDV genomes revealed an interesting trend: Regardless of genome size (100 kb to 2.5 Mb), essential genes encoding core replicative functions appear to cluster within the central region of the genome, while genes mediating virus-host interactions (including host range and immune evasion) localize towards the outer regions and show little or no conservation between NCLDV families [43]. The poxviruses represent a microcosm of this: Approximately 50% of the 190 - 200 genes encoded by vaccinia virus are conserved across all chordopoxviruses, and 49 of them are conserved across all poxviruses [29, 30]. These genes cluster toward the center of the viral genome while the more peripherally located genes show less conservation across poxviruses in terms of gene presence/absence [60] and are sometimes referred to as “unique” genes. Gene duplication alone does not sufficiently explain the acquisition of unique genes, suggesting that poxviruses, and other NCLDV families, employ multiple mechanisms of genome diversification.

A high incidence of horizontal gene transfer has been reported in poxviruses [43]. Poxvirus genes involved in immune evasion, for example, have almost exclusively evolved from captured host genes, which appear to have been acquired by poxviruses in an RNA-mediated manner, despite poxviruses not encoding or carrying a reverse transcriptase [37]. Genes apparently acquired by horizontal gene transfer have also been identified in other NCLDV species [43, 62], suggesting a common driver of diversification within the NCLDV. Prior to 2022 no mechanism for RNA-mediated horizontal gene transfer had been found within the NCLDV. More recent studies identified the mediation horizontal gene transfer in poxviruses by long interspersed nuclear element-1 (LINE-1) retrotransposon elements from the host genome [63, 64]. LINE-1 retrotransposons encode two proteins (ORF1 and ORF2, collectively referred to as “LINE-1 elements"), one with mRNA binding activity and the other with endonuclease/reverse transcriptase activity, which together are responsible for the reintegration of LINE-1 transcripts into the host genome. LINE-1 elements can also bind host cell mRNAs and appear to be to be transported to sites of viral replication in the cytoplasm during active poxvirus infection, albeit infrequently, where they can mediate the reverse transcription of host mRNA and integration of the resulting DNA product into the poxvirus genome. Reverse transcription and integration is mediated by the reverse transcriptase and endonuclease activity of ORF2 [65, 66]. Complementation and recombination events between co-infecting poxvirus genomes during subsequent rounds of infection then lead to the repair of any interrupted genes resulting from new integrations [44, 67].

Diversification within NCLDV families, and the move towards giantism, may be driven by multiple horizontal gene transfer events and gene duplication events within a viral genome [43, 44]. However, the full extent to which this occurs has been difficult to assess as the completeness of annotations of NCLDV proteomes has been highly variable, ranging from 7.4% annotation for paramecium bursaria chlorella virus, to 89% for vaccinia virus. This has been the limit imposed by sequence homology: Chapter 3 builds upon our prior work in Chapter 2, using a structural (as opposed to sequence) homology prediction method (HHsuite) to provide first-time annotations for uncharacterized proteins of the type-species from 20 distinct NCLDV families. This more comprehensive analysis revealed that Ankyrin repeat, MORN repeat and F-box domain containing proteins appeared to have undergone the most frequent paralogization events. We also found that gene diversification correlated with genome size up to a certain size limit, beyond which gene duplication replaced gene diversification as a driver of genome evolution among the giant viruses.

The structural homology work (above) was initiated, in large part, as an evolution of attempts to initiate crosslink-guided structure prediction as an application of our XLMS data. Here, crosslink information informs alternate theoretical pathways of protein folding, de-emphasizing pathways leading to folds that do not fit the intrachain crosslinking distance restraints. We thought this would be useful for virion structural proteins, for which no sequence homologs were available outside of the poxviruses. A key requirement of available crosslink-guided folding tools (e.g. online Robetta server), was the prior identification of domain boundaries and structural homologs. For this reason, then, we used homology prediction tools to find such homologs within the NCLDV. Crosslink-guided protein structure prediction tools were unsuccessful in our hands, but the body of knowledge generated in the area of structural homology led to two publications (Chapters 2 and 3). Later, AlphaFold2 provided a strong structural prediction tool, and we then switched to using XLMS data to validate models rather than to guide their generation.

Poxviruses at the molecular level

Considerable efforts have been made to understand poxviruses since at least the 1880s, when vaccinia virus particles were first identified [68]. At the molecular level, the cytoplasmic site of poxviral replication has raised questions of how DNA is maintained, replicated and transcribed independently of the nucleus. Furthering our understanding within the past 60 years: The viral genome has been sequenced [69] and roles for over half of vaccinia's ~200 genes have been elucidated at least in outline. Vaccinia's host-independent transcription system and its regulation are now quite well understood [70-72] and many of the viral gene products involved in the virus replication cycle including attachment and entry into host cells [29] and the morphogenesis of nascent virions have also been identified. The remainder of this chapter will discuss what is currently known about vaccinia virus biology and will highlight the approaches we have implemented to advance our understanding of the molecular architecture of the vaccinia virion.

Vaccinia virus ultrastructure, molecular structure, and morphogenesis

Vaccinia virion ultrastructure

Vaccinia virion ultrastructure has been extensively investigated over the past 80 years. The mature virion is ellipsoid or brick shaped, approximately 360 nm x 270 nm x 250 nm in overall dimensions, with a surface envelope that has an average thickness of ~5 - 6 nm [29, 73-75] and a biconcave internal core wall with an overall thickness of 18 - 19 nm. The core wall has two distinct layers – "palisade" layer, located immediately beneath the envelope, above a continuous 'inner' protein layer [73, 74]. The 195 kb linear dsDNA viral genome is packaged within the central portion of the core along with enzymes and protein factors required for early promoter binding and mRNA synthesis [29]. Positioned between the core wall concavities and the virion envelope are two "lateral bodies" which likely contain viral immunomodulatory proteins and oxidoreductases [76, 77].

Initial attempts at visualization in 1942 (not long after invention/commercialization of the electron microscope) revealed a brick-shaped overall geometry with an electron dense center [78]. A complementary study 6 years later revealed the presence of virus DNA within the core by imaging virions following pepsin and pepsin/deoxyribonuclease treatment of virions (subsequent higher resolution studies have suggested that the viral genome may be packaged as a nucleoprotein complex [79, 80]) and identified dense structures attached to opposing surfaces of the virion [81], later to be termed "lateral bodies". The virion surface (envelope) could be recognized as a lipid layer with surface elements or ridges with the appearance of closely packed macromolecules [81]. More comprehensive characterization of virion ultrastructure was achieved with progressive improvements in instrumentation, staining techniques and virus preparation methods, starting with the first thin section images of the virion in 1952 and 1954 [82, 83]. Virions were/are typically imaged with lateral bodies perpendicular to the grid ("vertically" oriented) or parallel to the grid, with only one lateral body visible ("horizontally" oriented). With the advent of negative stain transmission electron microscopy two distinct forms of the virion could be visualized, namely the “mulberry” and “capsular” forms (Fig. 1.1) depending upon the extent of penetration of stain (Fig 1.1). Mulberry-type images arose if the integrity of the MV surface envelope was maintained and stain was excluded from the virion interior [84-88], allowing surface features of the envelope to be imaged. These surface features could also be visualized by TEM following gold shadowing [81]. The surface ridges, first identified in 1948 [81], could be resolved as distinct cylindrical “rods” or “tubules” [84-86, 89] that ranged in diameter from 50 - 80 Å and in length from 200 - 1000 Å [85, 86] (Fig 1.1A). These “surface tubular elements” were organized in a semi-regular arrangement across the surface of the virion, albeit with many localized areas of disruption [84]. Notably, these surface features only appeared

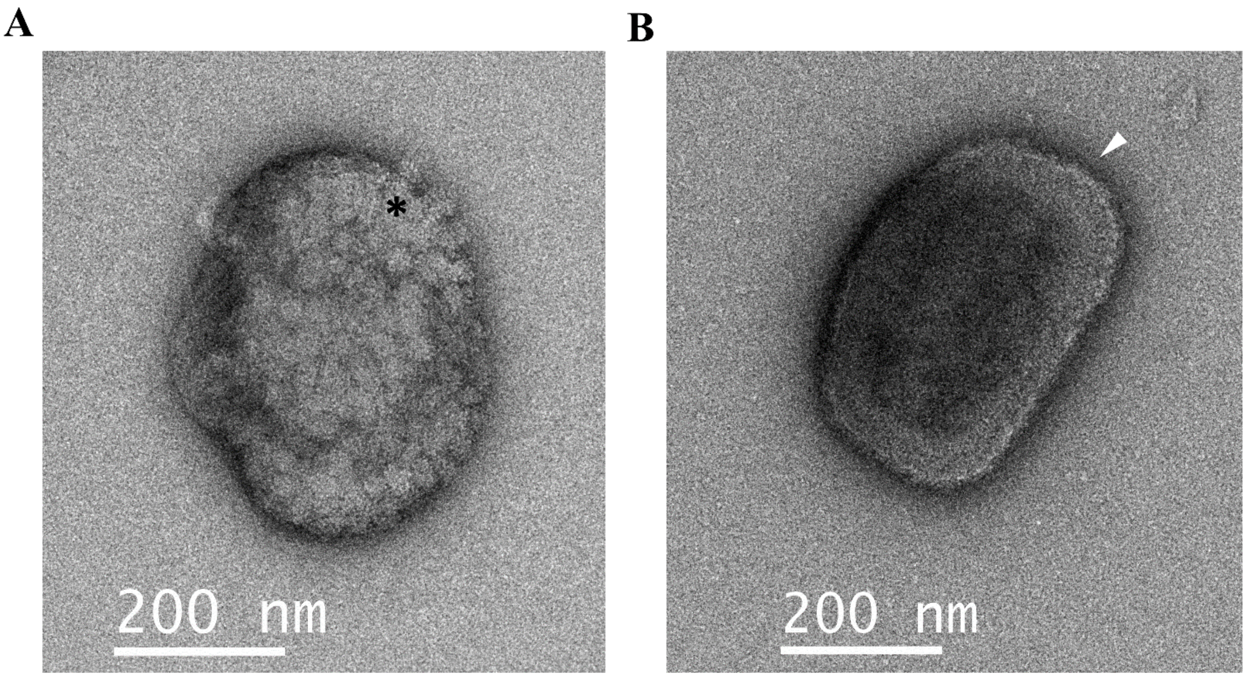


Figure 1.1. Two forms of Vaccinia virion particles by TEM. (A) Mulberry form of a horizontally oriented particle. An asterisk labels one of the surface tubular elements. (B) Capsular form of a horizontally oriented particle. The white arrowhead points out one of the virion “tips”.

under dehydrating imaging conditions [80, 90]. These surface elements appear to be conserved across chordopoxviruses, some of which (i.e., orf virus), exhibit a far more uniform, criss-crossing distribution of tubules [86]. No distinct features beneath the virion envelope could be visualized for the “mulberry” form of the virus, apart from impressions of a raised central region corresponding to a lateral body situated above the core [78, 81, 84-86].

The capsular form, which exposed the internal morphology of the vaccinia virion, was interpreted as stain penetrating beneath the envelope (Fig 1.1B) and could be reproduced by virion pretreatment with non-ionic detergent to remove the envelope entirely [84, 87]. When particles were oriented “horizontally”, a rectangular core was visible beneath the envelope. In this orientation, lateral bodies could not be seen under standard imaging conditions but could be visualized as centrally located above the surface of the core in freeze-fracture preparations of virus particles [88]. The biconcave or dumbbell shape of the core wall was apparent in “vertically” oriented particles, with lateral bodies nestled between the concavities of the core wall and the MV envelope [83-85, 87, 91]. Brief treatment of non-ionic detergent-treated virion cores with protease led to a degradation of the lateral bodies. Moreover, images of sectioned infected cells immediately after virion entry – the time at which lateral bodies separate from the virion core - showed that the core, minus lateral bodies, expands with loss of the pair of depressions on the core wall [85, 87, 90-92]. Later, upon development of cryoEM and cryoET techniques, the core wall could be better resolved revealing a “palisade” layer with cylindrical pegs that were in places organized into hexagonal patches [90, 93]. The palisade layer could also be visualized in some thin section preparations [87, 92, 93].Elements of the core wall surface have been imaged at higher resolution by cryoET [94]. Within the electron dense cores, a nucleoprotein complex could be imaged as three annular structures with dense central regions

[95, 96] or as a dense fibrous layer situated immediately beneath the core wall [73, 93]. The three annular structures have been interpreted as a single Z- or S- shaped structure viewed in a cross-sectional plane.

Other approaches to study the structure of the vaccinia virion have included controlled degradation experiments [73, 97] to "see" beneath the surface, atomic force microscopy of hydrated and dehydrated virus [79, 80], immunogold labeling and stochastic optical reconstruction microscopy (STORM). These studies have revealed general localizations for some of the packaged proteins to virion compartments, such as polarization of the entry fusion complex (EFC) proteins to the "tips" of the virion [30, 77, 98] (**Fig 1.1B**).

Virion molecular structure

Despite 60+ years of ultrastructural imaging, fundamental gaps remain in our understanding of how the virion is organized at the molecular level. The protein composition of the Vaccinia virion has been characterized [73, 99-103] suggesting approximately 75 packaged gene products in the mature virion [101]. Some/all of these are given in Table 1.2. As of this writing, atomic-level structures have been resolved for or parts of 32 of these proteins by X-ray crystallography or cryoEM, including the multisubunit DNA-dependent RNA polymerase and its associated transcription factors [71, 72]. Advances in electron microscope technology have led to atomic or near-atomic level structural models for a number of whole viruses including the capsids of three key members of the NCLDV, namely African swine fever virus [104, 105], paramecium bursaria chlorella virus [106, 107], and Melbournevirus [108]. For various reasons, however (outlined in the general discussion, Chapter 8), the standard approaches of structural biology have not been successfully applied to the vaccinia virion. These reasons include the enveloped, polymorphic, and asymmetric nature of the vaccinia virion [29, 30] which has rendered it refractory to X-ray crystallography and, thus far, to high resolution cryoEM/cryoET.

We felt compelled therefore, to develop an alternative approach to interrogate the higher order assemblies of packaged virion proteins. Protein-protein chemical crosslinking with mass spectrometry (XLMS) (Figure 1.2) is a minimally invasive approach for studying protein-protein interactions *in situ* that is quite agnostic to the interior vs. exterior of the particle, is not destructive to the virus particle, does not require reconstruction of protein complexes, and is not limited to binary complexes or surface proteins. The remaining central (data) chapters of the thesis explore and implements XLMS for vaccinia: **Chapter 4** explores the molecular arrangement of structural and membrane protein complexes resolved by XLMS. **Chapter 5** summarizes strategies that can be implemented to maximize crosslinked peptide identification for the purposes of achieving a saturating dataset of crosslinked proteins for vaccinia virus.

Vaccinia virus life cycle

Attachment to host cells

Attachment of vaccinia MV to the host cell is mediated by any or all of five vaccinia proteins: Four major attachment proteins A26, A27, H3, and D8, and protein L1 which is associated with the "EFC" discussed more fully below [109-114] (**Table 1.3**). Proteins A27, H3, and D8 each bind cell surface glycosaminoglycans (GAGs), with A27 and H3 binding heparin and heparan sulfate [110, 111, 115, 116] and D8 binding to chondroitin sulfate [113]. A26, in turn, attaches to the extracellular matrix protein laminin [109]. The cell surface receptor for L1 has not been elucidated: While L1 is capable of binding soluble GAGs *in vitro*, it can also facilitate vaccinia virus attachment to GAG-deficient cells [114].

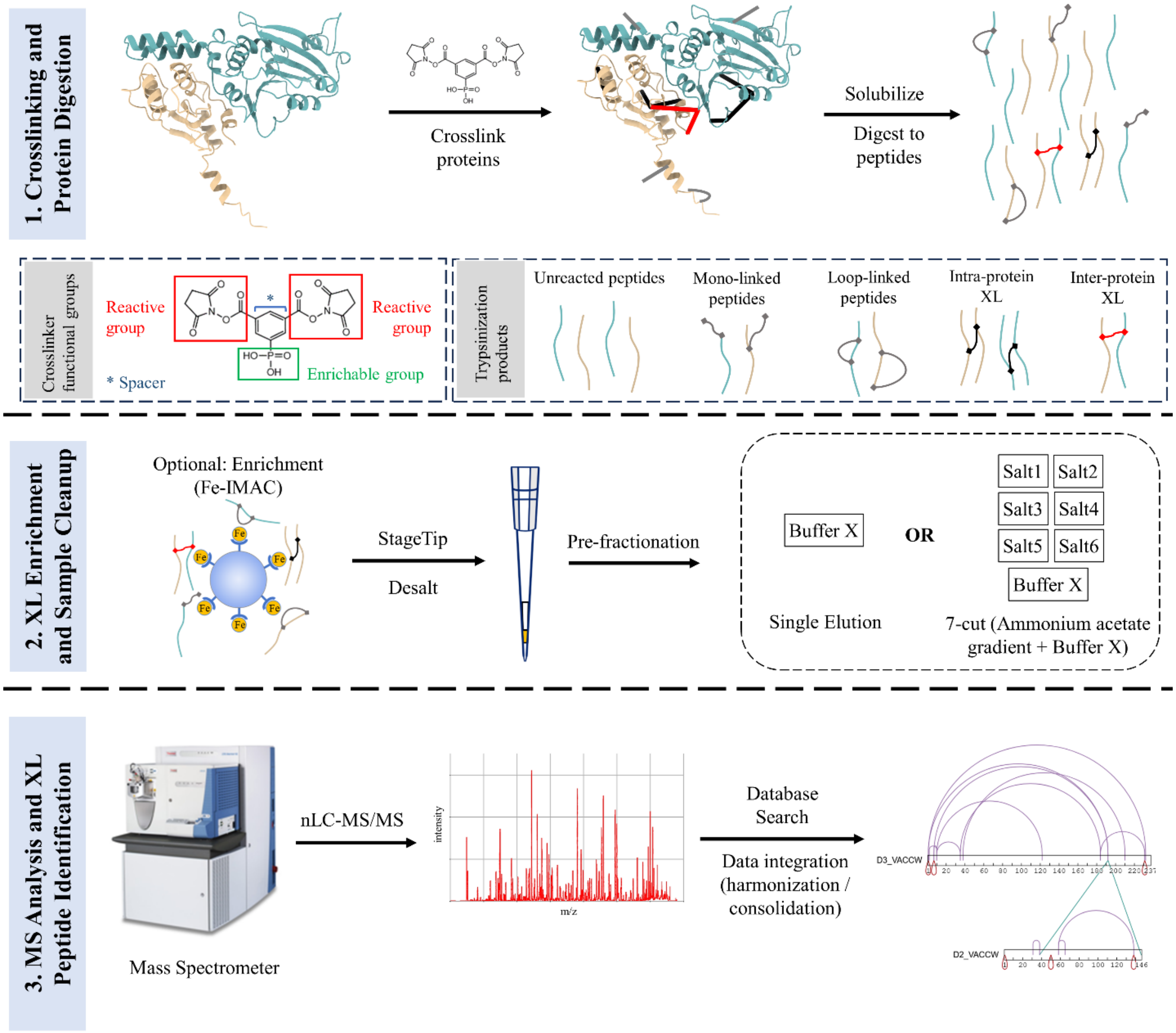
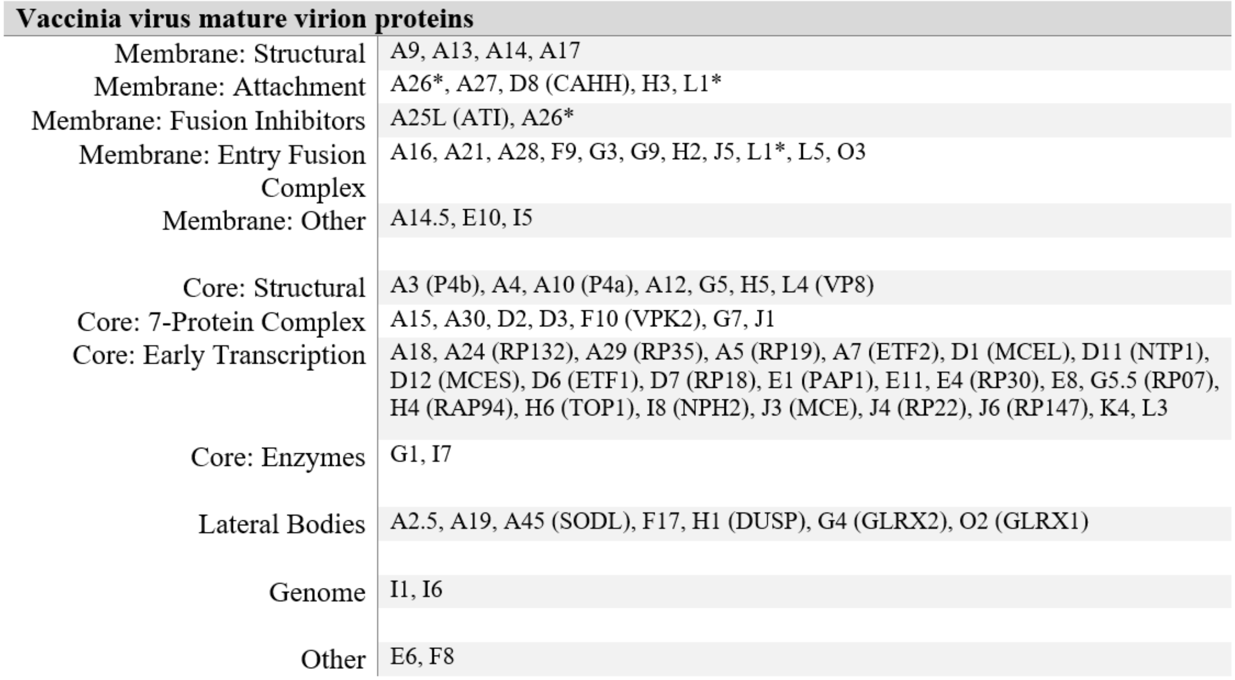


Figure 1.2. Typical workflow of protein-protein crosslinking mass spectrometry. Proteins are crosslinked with crosslinking reagents with fixed, known length spacer arms, two reactive groups, and an optional enrichable handle. Crosslinked proteins are then solubilized and trypsinized, resulting in a mixture of trypsinization products. Where an enrichable crosslinker is used, crosslinked peptides can be enriched from unreacted (regular) peptides. Peptides are then desalted and sometimes fractionated by strong cation exchange and then analyzed by nanoLC-MS/MS. The acquired spectrums are then analyzed by crosslinking search engines to identify crosslinked peptides, and the results of these searches are harmonized into a universal format. The consolidated dataset is used to identify protein-protein interactions, validate predicted protein structures, guide docking of proteins, and build larger protein assemblies and networks.

Repression of either A27, H3 or D8 results in reduced infectivity or attachment, depending on the cell type [113, 117].

H3, D8 and L1 are transmembrane proteins that are C-terminally anchored to the MV envelope [118-120]. A27 is anchored to the virion surface through interaction of its C-terminus with the N-terminal ectodomain of envelope transmembrane protein A17 [111, 121-124]. A26, in turn, is attached to the virion surface through inter-protein disulfide-bonds between its C-terminal domain and A27 [121, 125]. Partial atomic structures are available for all five proteins [115, 118, 126-128], however the manner in which these proteins plug into the molecular architecture of the virion has remained largely a mystery.

Table 1.3. 73 proteins known to be packaged in the vaccinia virus mature virion. Proteins are grouped by function. Proteins with multiple roles are listed in each applicable category with an (*). Where the protein common name differs from the gene name, the common name in parenthesis follows the gene name.



Vaccinia virus mature virion proteins	
Membrane: Structural	A9, A13, A14, A17
Membrane: Attachment	A26*, A27, D8 (CAHH), H3, L1*
Membrane: Fusion Inhibitors	A25L (ATI), A26*
Membrane: Entry Fusion Complex	A16, A21, A28, F9, G3, G9, H2, J5, L1*, L5, O3
Membrane: Other	A14.5, E10, I5
Core: Structural	A3 (P4b), A4, A10 (P4a), A12, G5, H5, L4 (VP8)
Core: 7-Protein Complex	A15, A30, D2, D3, F10 (VPK2), G7, J1
Core: Early Transcription	A18, A24 (RP132), A29 (RP35), A5 (RP19), A7 (ETF2), D1 (MCEL), D11 (NTP1), D12 (MCES), D6 (ETF1), D7 (RP18), E1 (PAP1), E11, E4 (RP30), E8, G5.5 (RP07), H4 (RAP94), H6 (TOP1), I8 (NPH2), J3 (MCE), J4 (RP22), J6 (RP147), K4, L3
Core: Enzymes	G1, I7
Lateral Bodies	A2.5, A19, A45 (SODL), F17, H1 (DUSP), G4 (GLRX2), O2 (GLRX1)
Genome	I1, I6
Other	E6, F8

With the exception of A26, which is dispensable, and L1 which is required for viral infectivity due to its interactions with EFC proteins (below) [129, 130], repression of any one of the remaining three attachment protein individually is not prohibitive to morphogenesis but can result in reduced infectivity or cell attachment, depending on the cell type [113, 117].

The entry fusion complex

Vaccinia has two known pathways for host cell entry, namely (a) micropinocytosis followed by fusion of the virion envelope with the pinocytic vacuolar membrane, and (b) direct fusion of virion envelope with the plasma membrane. In the pinocytic pathway (a, above), the phosphatidylserine-rich outer leaflet of the virion envelope triggers pinocytosis - a process also referred to as “apoptotic mimicry” – in which phosphatidylserine exposure on the outer leaflet of a cell serves as a signal that it is apoptotic debris in need of clearance as such by phagocytosis. This endocytic internalization pathway closely resembles macropinocytosis [131-133]. Once inside the resulting vacuole, vacuolar acidification leads to low-pH-triggered fusion of the virion envelope with endosomal membranes, pore formation, and release of the virion core into the cytoplasm [133-135]. In the direct fusion mechanism (b, above), depending on the virus strain, host cell type, and infection conditions, fusion, pore formation and entry occur directly at the host cell plasma membrane [75, 133, 136-140].

The fusion step (above) proceeds with lipid mixing between the outer leaflets of the apposed lipid bilayers of virion envelope and relevant host cell membrane, leading to a hemifusion intermediate [134]. This is followed by completion of the fusion process, pore formation, and release of the virion core into the host cell cytoplasm. The above process, irrespective of entry pathway, is reliant on the concerted action of the 11 proteins termed the “entry-fusion complex (“EFC”), comprising proteins A16, A21, A28, F9, G3, G9, H2, J5, L1, L5, and O3 [134] (**Table 1.3**). These proteins are conserved across all poxvirus species [141, 142]. All 11 proteins contain N- or C-terminal transmembrane domains by which they are inserted into the virion envelope [141, 142]. Nine of the proteins, A16, A21, A28, G3, G9, H2, J5, L5, and O3 form the stable core of the complex while F9 and L1 are peripherally associated [143].

The mechanism by which these 11 proteins facilitate virus entry is not well defined. By STORM imaging, EFC proteins localize to the "tips" of vaccinia virions (**Figure 1.1B**), whereas attachment proteins appear distributed across the longer sides of the brick shaped virions [98]. Virion attachment appears to occur initially along the virion long sides, followed by fusion and pore formation at virion "tips", thus allowing the release of virion cores into the host cell cytoplasm [98]. Interestingly, repression of attachment protein A27 disrupts the above localization of EFC proteins such that they are dispersed across the entire surface instead of localizing to the virion tips – yet virion attachment and hemifusion can still proceed [113] albeit there is a delay in fusion but has no further effect on infection kinetics or virus production [98].

Protein features that may be involved in the fusion process have been identified: Protein H2, for example, contains a highly conserved fusion motif "LGYSG" that is comparable to fusion motifs of other viruses [144-147]. Although not a member of the EFC, the N-terminal ectodomain of A17, which anchors protein A27 to the envelope surface, may also be involved in fusion, as it can insert into host membranes after acidification and has been shown *in vitro* to induce syncytium formation in HEK293T cells [111].

The loss or repression of any one EFC protein (with the exception of O3, which can be partially compensated for by mutations in the F9 transmembrane domain [148]), results in virions that are morphologically normal but unable to facilitate virus entry [130, 134, 149-157].

Virions lacking A16, A21, F9, G3, G9, H2, J5 or O3 are also deficient in hemifusion [158], [98, 159]. STORM imaging of virus-bound cells has shown that the block in virus entry is due, in part, to disruption of the localization of EFC proteins to the virion tips [98]. However, this alone is insufficient to block fusion, as A27 deficient virions also show a dispersal of EFC proteins across the virion surface but are capable of completing the entry process (above).

Although interactions between EFC proteins (and the formation of subcomplexes within the EFC) have previously been noted [142, 143], how the EFC assembles into larger subcomplexes and a discrete, comprehensive complex has not been identified. Until recently, atomic-level structural models were only available for proteins L1 and F9 [118, 160]. In the past year, X-ray crystal structures were reported for the ectodomains of an A16:G9 subcomplex [161] and a G3:L5 subcomplex [162]. **Chapter 6** explores the predicted structures and higher order assemblies for the 21 known virion envelope and virion surface proteins, including interactions between the virion attachment proteins and assembly of EFC subcomplexes into larger subassemblies.

Fusion suppression

One of the factors that determine whether virus entry occurs at the plasma membrane or after macropinocytosis/endosomal acidification is the packaging status of vaccinia surface proteins A26 and A25L (ATI) [127, 163, 164] (**Table 1.3**). A25L is a C-terminally truncated vaccinia WR ortholog of the cowpox virus A-type inclusion body protein ATIp. Due to this truncation, vaccinia A25L does not form inclusion bodies and is instead incorporated into the virion in an A26-dependent manner [165]. In addition to the role of A26 in attachment of virions to host cells (above), A26 and A25L can suppress virus-host cell fusion at the plasma membrane instead directing viral entry to the endosomal pathway. Fusion suppression by A26 occurs through its interaction with EFC proteins A16 and G9 [163, 164]. The mechanism by which A25L blocks entry at the plasma membrane has not been characterized. Brief acid treatment of MVs can inactivate A26 and A25L, allowing fusion to occur at the plasma membrane [127, 133-135, 163, 164, 166].

Early infection and genome replication

Membrane fusion and release of virion cores into the host cell cytoplasm coincide with the dissociation of lateral bodies from virion cores. Lateral bodies can be visualized, immediately after entry, to be associated with the fused virion envelope [136, 137, 167] and are rapidly disassembled following proteosome-dependent degradation of lateral body scaffold protein F17 [76, 136, 137, 167]. This disassembly process allows the release of viral immunomodulatory proteins that are packaged within the lateral bodies, including vaccinia dual specificity phosphatase VH1 which is able to dephosphorylate p-STAT1 and block the expression of STAT1-dependent IFNγ-stimulated gene expression [76, 168, 169].

Once inside the host cell cytoplasm virion cores become activated and expand from biconcave structures to larger ovoids [170]. Cores are then transported along microtubules to the perinuclear space, where replication sites are established [171]. Early gene expression occurs solely within virion cores at these replication sites [172] using the packaged viral DNA-dependent RNA polymerase, mRNA capping enzyme, poly(A) polymerase, transcription factors, and other enzymes [173]. During this time, permeability of the core wall increases, allowing rNTPs to enter the core and nascent mRNA to exit and be translated on host ribosomes [170]. Up to 100 viral genes (approximately half of the vaccinia virus WR genome) are transcribed within virion cores within the first 2 - 3 hours post infection [29, 174], including genes required for genome replication, transcription factors for virus intermediate gene expression and modulators of the host immune response [29]. Expression of cellular mRNA is shut down. Early gene expression concludes with dissolution of the virion core wall, which releases the contents of the core, including the viral genome, into the host cytoplasm [172]. Pre-replication foci are rapidly formed by association of the viral genome and early proteins with endoplasmic reticulon (ER) cisternae proximal to the nucleus, generating the sites within which genome replication begins [175].

As genome replication proceeds over the course of infection, ER cisternae are continuously recruited to the prereplication foci, allowing these replication sites to expand to accommodate the increasing number of genome copies [172, 175], eventually transforming into fully enclosed viral factories [172]. Replication of the viral genome within these factories appears to provide some protection against host immune sensors of cytoplasmic DNA. The vaccinia genome encodes at least 12 proteins, across different time points of infection, to combat cytoplasmic DNA sensing and consequent shutdown of genome replication [176]. Included in this group is serine/threonine protein kinase B1, which localizes to viral factories during infection and phosphorylates cytoplasmic cellular dsDNA binding protein BAF, preventing its attachment to and cross-binding of the viral genome [171]. The lateral body scaffolding protein F17, which is produced during late gene expression, also prevents detection of the viral DNA by suppressing the activation of cGAS-STING by binding to and sequestering mTOR regulatory proteins Raptor and Rictor [177, 178].

Intermediate/late gene expression and virion assembly

Intermediate gene expression begins after viral DNA replication has initiated, during which time proteins required for DNA binding and packaging, various core associated proteins (including enzymes), and late gene transcription factors are synthesized [179, 180]. Late gene expression, in turn, begins generally 6 - 8 hours post infection and all remaining proteins necessary for virion assembly, genome packaging, and enzymes required for early gene expression (which are packaged in virion cores) are produced [179-181].

Virion assembly begins inside viral factories with the cotranslational insertion of vaccinia transmembrane proteins A17 and A14 into the ER cisternae. Viral membrane assembly proteins (A6, A11, A30.5, H7, and L2), in conjunction with cellular proteins, rupture the ER membrane or stabilize naturally occurring transient breaks in the membrane and prevent them from reforming. These breaks allow pre-assembled trimers of the vaccinia D13 scaffolding protein to enter the ER lumen, where they associate with A17 N-termini to form curved open membrane structures called "crescents". In this manner, the inner leaflet of the ER membrane becomes the outer surface of the crescent membrane [182]. Repression of any of the above eight proteins results in a complete block to virion morphogenesis [122, 182-188].

Individual crescent structures continue to grow towards a closed spherical structure by accumulating additional pieces of modified ER membrane until they form a spherical immature virion (IV), with D13 trimers assembled as a honeycomb-like external scaffold across the IV surface [94, 189-191]. As crescents develop into IV, they associate with "viroplasm" (electron dense granular pre-viral matter), resulting in IVs filled with the viral proteins that will form the mature virion [30]. Crescent formation, association with viroplasm, and formation of filled IVs is reliant, in part, on a complex of seven vaccinia virus encoded proteins A15, A30, D2, D3, G7, J1 and VPK2 (**Table 1.3**), collectively described as the "seven-protein complex" (7PC). The 7PC is conserved across poxviruses, with no known sequence homologs elsewhere. Repression or mutation of any one 7PC member results in a comparable phenotype – disruption of crescent formation and/or of the subsequent attachment of viroplasm to nascent crescents [192-195].

These defects are accompanied by an abrogation of morphogenic protein phosphorylation and processing events, and no production of MV [192, 196, 197].

The newly replicated viral genome is then packed into IVs, forming in IVNs (IVs with nucleoids) [171]. If morphogenesis is interrupted at a stage prior to genome packaging, viral DNA accumulates in the cytoplasm as DNA crystalloids, which can be visualized by TEM [198]. The exact mechanism of genome packaging is not well understood, but it is thought that viral envelope protein A13, possibly with the assistance of protein H3 [117], facilitates the attachment of vaccinia genome packaging ATPase A32 to IVs [171]. A32 is a member of the HerA/FtsK superfamily of ring shaped hexameric ATPase motors that have roles in phage genome packaging, chromosome segregation during bacterial cell division, bacterial conjugation and DNA repair [199]. Similar to other HerA/Ftsk superfamily ATPase motors, the predicted structure of A32 shows the classical Walker A, Walker B, and Arg finger motifs (Mirzakhanyan and Gershon, unpublished data) required for nucleotide binding, exchange and ATP hydrolysis [200] and is predicted to form a hexamer (Mirzakhanyan and Gershon, unpublished data). In conjunction with A13 and vaccinia telomere binding protein I6, A32 is thought to drive DNA translocation and packaging [171].

The final stage of vaccinia virus assembly is the transition from IV to MV. The exact mechanism of this maturation process is poorly understood, but a critical step is the proteolytic processing of core structural and membrane proteins at consensus AG|X sites by vaccinia cysteine protease I7 [201-203]. In the absence of I7, normal morphogenesis is abrogated and MV do not form [201, 202, 204]. Maturation of IVs to MVs begins with the remove the N-terminal 16 aa of A17 by I7, leading to release and disassembly of the external D13 scaffold [203] followed by the incorporation of proteins H3, A27, A26, and A25L at the virion surface. This is followed by processing of virion proteins VP8 and precursor P4a, P4b, A12, and G7 [205, 206].

Proteolytic cleavage of core structural proteins P4a, P4b, and VP8 (**Table 1.3**) is essential for normal morphogenesis and is contextually restricted to IVNs. P4B and VP8 both encode N-terminal signal sequences, with AG|X cleavage sites at residues 61 and 32 respectively, which appear to play a role in targeting of P4B and VP8 to IVs [205, 207]. Expression of P4B or VP8 with the N-terminal signal sequence removed results in a block to morphogenesis and the production of abnormal virus particles [205, 208, 209]. Unlike P4b and VP8, precursor P4a does not encode an N-terminal signal sequence. The precursor P4a N-terminal AG|X site (residue 95) is not processed by I7 and instead appears to be involved in multimerization of the maturation product P4a-1 [210, 211]. Instead, precursor P4a undergoes proteolytic processing events by I7 at two downstream AG|X sites (residues 614 and 697), producing three products: P4a-1, P4a-2, and P4a-3 [207, 209, 211]. P4a-1 and P4a-3 are packaged and become major components of the palisade and core wall [210, 212, 213], while P4a-2 is discarded, possibly at the proteasome [214]. The processing, trimerization and assembly of P4a-1 into the core wall palisade layer are discussed in **Chapter 7**.

Morphogenesis concludes with the rearrangement of core structural proteins, resulting in formation of the palisade layer and core wall within which the genome is packaged. Intramolecular disulfide bonds are formed, assisted by the vaccinia virus encoded redox system (proteins A2.5, E10 and G4). Lateral body proteins condense, and virus particles transition from spherical to brick shaped [29, 30]. The resulting infectious particles are disseminated by cell lysis or are transported across the trans-Golgi stacks (acquiring two additional “wrapping” membranes) to the cell surface [215] where they are released [216].

CHAPTER 2

Multisubunit DNA-dependent RNA polymerases from vaccinia virus and other nucleocytoplasmic large-DNA viruses: impressions from the age of structure

Abstract

The past 17 years have been marked by a revolution in our understanding of cellular multisubunit DNA-dependent RNA polymerases (MSDDRPs) at the structural level. A parallel development over the past 15 years has been the emerging story of the giant viruses, which encode MSDDRPs. Here we link the two in an attempt to understand the specialization of multisubunit RNA polymerases in the domain of life encompassing the large nucleocytoplasmic DNA viruses (NCLDV), a superclade that includes the giant viruses and the biochemically well-characterized poxvirus vaccinia virus. The first half of this chapter surveys the recently determined structural biology of cellular RNA polymerases for a microbiology readership. The second half discusses a reannotation of MSDDRP subunits from NCLDV families and the apparent specialization of these enzymes by virus family and by subunit with regard to subunit or domain loss, subunit dissociability, endogenous control of polymerase arrest, and the elimination/customization of regulatory interactions that would confer higher-order cellular control. Some themes are apparent in linking subunit function to structure in the viral world: as with cellular RNA polymerases I and III and unlike cellular RNA polymerase II, the viral enzymes seem to opt for speed and processivity and seem to have eliminated domains associated with higher-order regulation. The adoption/loss of viral RNA polymerase proofreading functions may have played a part in matching intrinsic mutability to genome size.

Introduction

The multisubunit DNA-directed/dependent RNA polymerases (MSDDRPs) lie at the heart of the central dogma, are key enzymes of living systems, and are universally found in cellular organisms. While eubacteria and archaea possess a single such enzyme, eukaryotes possess at least three, namely, RNA polymerase I (pol I), pol II, and pol III, which specialize in the synthesis of rRNA, mRNA, and a collection of smaller RNAs such as tRNA and 5S rRNA, respectively. MSDDRPs are also encoded by DNA viruses that have a cytoplasmic transcriptional phase, such as the poxviruses. In this chapter, we address the viral MSDDRPs in the context of many recent advances in the structural and biochemical understanding of their cellular counterparts, along with the growing numbers of giant viruses encoding RNA polymerase subunits.

Cellular MSDDRPs

To paraphrase Dobzhansky, very little in transcription makes mechanistic sense except in the light of structural biology. Over the past 18 years, and ongoing, impressive efforts in X-ray crystallography, transitioning to cryo-electron microscopy (cryoEM), have provided deep insights into the structural biology of the cellular MSDDRPs, including their architecture, evolution, structure-function relationships, and dynamics. Growing numbers of such studies have covered the enzymes from bacteria [217, 218], archaeal species [219-221], and eukaryotes [222-251], in various interaction states.

Functionally, all MSDDRPs share a set of common features, namely, the copying of a DNA template to newly synthesized RNA, stepwise enzyme translocation on the template during RNA synthesis, utilization of nucleoside triphosphate (NTP) substrates, pairing of the incoming nucleotide with the DNA template strand via Watson-Crick base pairing, and catalysis of nucleotide transfer via a metal-dependent mechanism. Architecturally, all MSDDRPs comprise two large subunits and a collection of smaller ones, in which the two large subunits have remained remarkably conserved across all domains of life, namely, the three cellular domains and the nucleocytoplasmic large DNA viruses (NCLDV) [252]. Structural and biochemical studies have focused largely on two prototypical MSDDRPs, the bacterial enzyme and pol II from the budding yeast *Saccharomyces cerevisiae* [253, 254]. While bacterial MSDDRP has just 4 distinct subunits, named α (two copies), β, β', and ω, *S. cerevisiae* pol II has either 10 or 12. While the 10-subunit core enzyme comprising subunits Rpb1, -2, -3, -5, -6, and -8 to -12 is competent in transcription elongation, two additional, dissociable subunits, the Rpb4/7 heterodimer, are conditionally required for transcription initiation [255]. Other eukaryotes always have the 12-subunit form. Of the two prototypical enzymes, the viral MSDDRPs more closely resemble the one from *S. cerevisiae*, so this will be described initially.

Structure-based models of nonbacterial transcription: yeast pol II

The earliest X-ray crystallographic structures of yeast pol II [222] showed the core (10-subunit) enzyme with an architecture comprising the two large subunits, RPB1 and -2, flanking a cleft that was deep and wide enough to accommodate a double-stranded DNA helix and which contained the catalytic center on its floor. An opening, or "pore," in the floor immediately beneath the catalytic center exposed the DNA-RNA hybrid to an inverted funnel-shaped cavity on the outside of the enzyme, allowing incoming NTPs accesses to the active site (Fig. 2.1A). One end of the cleft opened at the downstream face of the polymerase (the "front" of the enzyme during its translocation along duplex DNA), while the other end was blocked by a "wall" structure positioned just beyond the catalytic center but prior to the upstream face of the polymerase (facing the promoter [Fig. 2.1A]). One side of the cleft formed a clamp which in the open state is wide enough for double-stranded DNA to enter but in the closed state is only wide enough for a single DNA strand [256] (Fig. 2.1A). Close to the downstream face of the enzyme, just within the cleft, structural features termed "jaws" were present, composed of yeast subunits Rpb5 (lower jaw) and Rpb1/9 (upper jaw). At the upstream face of the enzyme could be found a subassembly of four relatively small subunits (Rpb3, -10, -11, and -12), with the Rpb3/11 subunits corresponding, approximately, to subunits α/α' in the bacterial MSDDRP. The above description accounts for 8 of the 10 core subunits. Additionally, Rpb6 formed a clamp across the cleft, while Rpb8 was located near to the Rpb3-10-11-12 subassembly [222]. In the 12-subunit pol II holoenzyme, the Rpb4/7 heterodimer formed a "stalk" structure toward the enzyme's upstream face [225, 235].

As might be anticipated, pol I and pol III are closely related variants of pol II: five of the 10 subunits of yeast core pol II are identical across all three forms of eukaryotic MSDDRP, while another five are highly conserved between the three enzymes.

RNA polymerase function in the context of structure

Transcriptional elongation can be regarded as a continuous, dynamic "production line" in which the template enters the polymerase cleft at the downstream face and becomes unwound, with the template strand passing over the catalytic center where the 3' end of a complementary nascent RNA transcript comprises the downstream terminus of a 7- to 8-base-pair DNA-RNA hybrid minihelix (Fig. 2.1A). This nascent transcript becomes extended with a complementary incoming ribonucleotide. At the upstream end of the minihelix, the nascent RNA is peeled away from the DNA template strand and channeled away from the polymerase [238]. The template and

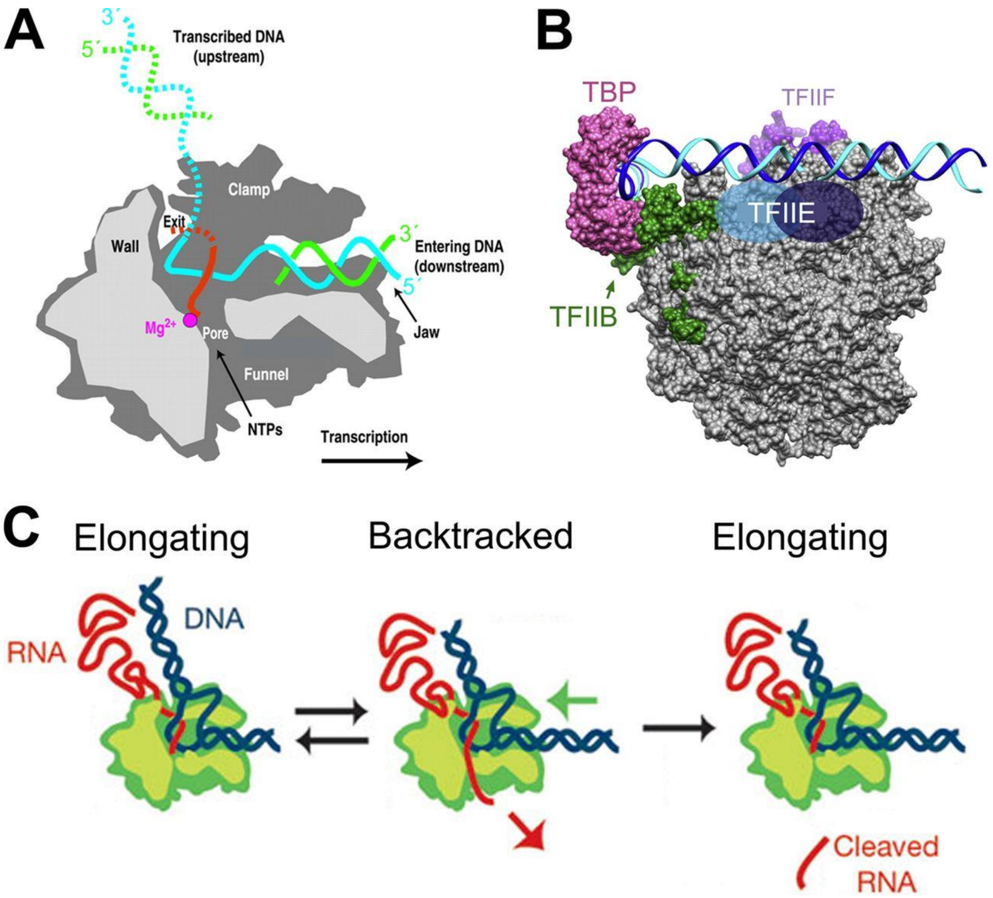


Figure 2.1. Structure and function of yeast RNA pol II. (A) Cutaway section schematic of a pol II transcribing complex (polymerase moving left to right). From the direction of view ("side" of pol II), the cutaway plane exposes nucleic acids and functional elements of the enzyme. Light gray, cut surfaces of the protein (front). Dark gray, receding surfaces. Right side, template and non-template strands (cyan and green, respectively) of the entering DNA duplex (the unwound portion of the non-template strand is not shown). Red, 3' end of nascent RNA within the DNA-RNA hybrid minihelix. Magenta, catalytic metal. Of the two jaws, the cutaway reveals only the lower one. The far wall of the DNA-binding cleft forms the clamp structure [222, 234]. For other details, see the text. (Adapted from reference [234] with permission [copyright 2002 National Academy of Sciences].) (B) Side view (as in panel A) of the surface-rendered (not cutaway) basal preinitiation complex showing pol II (gray), TBP (dark pink) and TFIIB (green) [257], TFIIF (purple), and the position of TFIIE (from cross-linking studies [258]) (blue). The closed-promoter DNA duplex encompassing the transcriptional start site is modeled, suspended above the pol II cleft via general transcription factors TBP, TFIIB, and TFIIE, in which TBP and TFIIB hold the upstream promoter DNA [259]. (Adapted from reference [259] with permission from Elsevier.) (C) Backtracking. In the left and center panels, RNA polymerase can move forward (left) or backtrack with extrusion of the nascent transcript's 3' end through the NTP entry pore leading to an arrested state (center). Transcript 3' end cleavage (right) restores an elongation-competent complex. (Adapted from reference [260] by permission from Nature Publishing Group.)

non-template DNA strands are then free to reanneal and fully exit the enzyme assembly in an upstream direction.

A model for preinitiation complex formation has been developed in a minimal system comprising polymerase, DNA template, and the two general transcription factors (GTFs) TATA-binding protein (TBP) and TFIIB [229, 243-245]. Here, promoter DNA is initially bound at the upstream face of the polymerase, with TBP binding both the promoter's minimal TATA element and factor TFIIB, which is in turn attached to the polymerase (Fig. 2.1B). The DNA duplex bends around the polymerase such that the region downstream of the promoter tracks above the enzyme's cleft. At this stage, however, DNA makes no direct contacts with the polymerase (Fig. 2.1B). As a result of breathing of the duplex, either via natural supercoiling or as induced by the factor TFIIF, an open state of the duplex is captured by TFIIB, and the flexible template strand subsequently descends into the cleft adjacent to the "wall" structure, where, upon reaching the catalytic center, RNA synthesis can commence in the presence of NTPs. Via its lowering into the cleft, the partially melted DNA has been reconfigured so that the downstream duplex can now enter the cleft from the downstream side with the nascent DNA-RNA hybrid helix climbing the wall behind the catalytic center and out of the cleft at the upstream side at an angle of approximately 90 to 105° to the incoming downstream duplex [224, 226] (Fig. 2.1A). During initiation, as downstream DNA enters the cleft and RNA synthesis proceeds to a 5-nucleotide (nt)-long transcript, a steric clash of the 5' end of the nascent RNA with a finger domain of TFIIB that reaches into the cleft forces a decision between the abortion of transcription with mini transcript release or the destabilization of bound TFIIB (and of the whole initiation complex) followed by unhindered transcriptional elongation. At this point, transcripts of 7 nt and longer are able to interact with an unwinding site on the polymerase for the hybrid minihelix, leading to single-stranded RNA exit and DNA duplex rewinding at the upstream side [238]. Elongation can now proceed.

The intricate world of RNA polymerase backtracking

During mRNA synthesis, pol II moves forward along the DNA template as a “Brownian ratchet” [240, 261] (Fig. 2.1C). However, at certain DNA sequences the enzyme may pause, providing opportunities for “stop-go” transcriptional regulation at the level of elongation. In one notable example, partially elongated transcripts of the cellular heat shock gene pause for prolonged periods after synthesis of their 5’ ends, poised for rapid, factor-dependent reactivation later, in response to stress [262]. Pausing also has roles in co-transcriptional RNA folding and processing, transcription termination, and genome stability. In addition, protein roadblocks such as nucleosomes or DNA-bound transcription factors can render transcriptional pausing unavoidable. Even on naked DNA under near-optimal conditions, pol II may persistently pause [263] at sequences where, for example, the DNA-RNA hybrid is weak [247].

During pausing as described above, or if a mismatched nucleotide has been misincorporated at the 3’ position of the transcript, the nascent RNA 3’ end may become “frayed,” i.e., disengaged from the DNA template strand at the polymerase catalytic center. In this case, the polymerase may move backwards on the template for a short distance (“backtracking”) (Fig. 2.1C). Though backtracking by just one or two residues may be reversible via one-dimensional forward diffusion of the enzyme [264], if backtracking continues for 8 or 9 nucleotides or more, the polymerase is likely to become arrested (incapable of spontaneously resuming forward elongation without the assistance of additional factors (reference [242] and references therein) (Fig. 2.1C). The arrested or backtracked enzyme is generally stable but inactive [261]. In eukaryotes, it can be reactivated by transcription factor TFIIS (see below).

Structural correlate of backtracking. As revealed from elegant structural studies, during pol II backtracking the “frayed” RNA 3’ end is extruded from the polymerase active site and through the NTP entry pore in the floor of the enzyme (Fig. 2.1A) and then into the “funnel” immediately below the pore on the outside surface of the polymerase [247]. After backtracking for eight or more nucleotides, the extruded RNA is sufficiently long to bind a conserved “backtrack site” located along one side of the pore and into the funnel [247]. Trapping of the RNA 3’ end at this site strongly inhibits further movement and is the basis for transcriptional arrest. The backtracked state is not equivalent to the forward (transcribing) state simply displaced backward along the nucleic acid scaffold but is instead a distinct and stable off-pathway state that involves structural changes leading to an inhibition of catalytic competency [261] and active retrograde movement [265]. These changes are beyond the scope of this chapter. Where backtracking is less extensive, however, RNA interactions with the backtrack site may be partial and weak, and pol II may then spontaneously diffuse forward [247].

Pol II reactivation by factor TFIIS. Arrested pol II can be reactivated by the cellular transcription factor TFIIS via a mechanism involving cleavage of backtracked RNA at the catalytic center [266, 267] (Fig. 2.1C). TFIIS has three independently folding domains (references [267] and [268] and references therein). Domain 1 (amino acids [aa] 1 to 130, *S. cerevisiae* numbering) is not required for anti-arrest functions (reference [268] and references therein) (see below). Domain II (aa 130 to 240) is tethered to domain III (aa 260 to 309) via a short linker, with domain II and the linker being responsible for pol II binding. Domain III is essential for the anti-arrest activity of polymerase-bound TFIIS and for transcript cleavage [269].

TFIIS binds pol II near the rim of the funnel, extending domain III into the NTP entry pore so that a β-hairpin loop within a “Zn ribbon” region of domain III reaches the active site

[227, 247]. Transcript cleavage by TFIIS probably involves three charged residues within this hairpin [247] which complement the pol II active site and may help catalyze the necessary proton transfers [247]. TFIIS also weakens pol II's grip on backtracked RNA at the backtrack site in the enzyme's funnel as follows. Via direct competition at the backtrack site, bound TFIIS displaces the backtracked RNA, which moves into a region of the pore that remains unblocked after TFIIS domain III insertion [247]. The cleaved 3' portion of the RNA, being already displaced from the polymerase, is released, leaving the enzyme poised for further NTP addition [261].

Intrinsic reactivation: pol II. The pol II reactivation story does not quite end here: in addition to exploiting transcript cleavage factor TFIIS, pol II also has a very weak intrinsic cleavage activity arising from the nonessential [270] intrinsic subunit Rpb9 [264, 271, 272]. Rpb9 has two Zn ribbons, one at each protein terminus. However, these are too distantly located and/or too tightly packed against the core enzyme to readily reach the NTP entry pore [271]. It is unknown whether Rpb9's weak intrinsic cleavage activity arises from the vestigial activity of one or both of these suboptimally positioned ribbons directly [271] or through an ability of Rpb9 to allosterically reconfigure a key catalytic loop in the polymerase active site for transcript cleavage instead of polymerization [273]. Whichever is the true mechanism, while TFIIS's *in vivo* role may be in the reversal of strong arrest (see above), Rpb9's role may be in the proofreading of nucleotide incorporation errors immediately after they occur, increasing pol II's transcriptional fidelity [272, 274, 275].

Intrinsic reactivation: pol I and pol III. The jury is out on whether pol I and pol III also exploit TFIIS [276]. Perhaps more importantly, however, pol I has a very effective intrinsic activity for the transcript 3' end cleavage and the reactivation of backtracked complexes ([277]; see reference [264] and references therein), which is much more potent than the intrinsic activity of pol II [278]. Pol I's intrinsic cleavage activity arises from subunit A12.2 (Fig. 2.2), a homolog of Rpb9 [278]. Like Rpb9, A12.2 possesses a Zn-binding β-ribbon at either protein terminus (being a much smaller subunit than Rpb9, A12.2 has just a flexible linker connecting the two ribbons). The C-terminal ribbon's hairpin includes counterparts of TFIIS's three catalytic residues for transcript cleavage (see above). In pol I "apo" structures (lacking nucleic acid), A12.2's N ribbon was positioned equivalently to that of Rpb9 in pol II (Fig. 2.2), but the C ribbon was positioned inside the NTP entry pore almost perfectly equivalent to the position of the TFIIS C ribbon in pol II structures [230, 231, 279]. As in the TFIIS/pol II complex, the "catalytic" hairpin of A12.2 reached the polymerase active center [230], providing compelling structural evidence for involvement of A12.2's C ribbon in pol I's strong, intrinsic transcript 3' cleavage activity through transient insertion into the NTP entry pore (Fig. 2.2). In the "transcribing" structure of pol I [232], i.e., in the presence of template DNA and partially elongated RNA, the A12.2 N ribbon remained unmoved, but the C ribbon was now displaced from the pore and invisible in the structure (Fig. 2.2). This underlined the apparently transitory nature of pore entry by A12.2's C ribbon.

Like pol I, pol III has a strong, intrinsic transcript cleavage activity [280]. The pol III counterpart to A12.2 is subunit C11 which also has an N ribbon positioned in a similar way to that of Rpb9 [233] (Fig. 2.2). The C ribbon of C11 [281], however, in pol III apo structures [233] was far away from the position of A12.2's C ribbon in pol I. In the pol III elongating structure, the C ribbon was not visible at all [233] (Fig. 2.2). As with pol I, this suggested that the intrinsic C ribbon is mobile and is only temporarily recruited to the catalytic center (above, Fig. 2.2). In an elegant experiment, pol II's weak intrinsic transcript cleavage activity was rescued to strong pol III-type activity by substituting Rpb9's C ribbon with its counterpart from C11 [271].

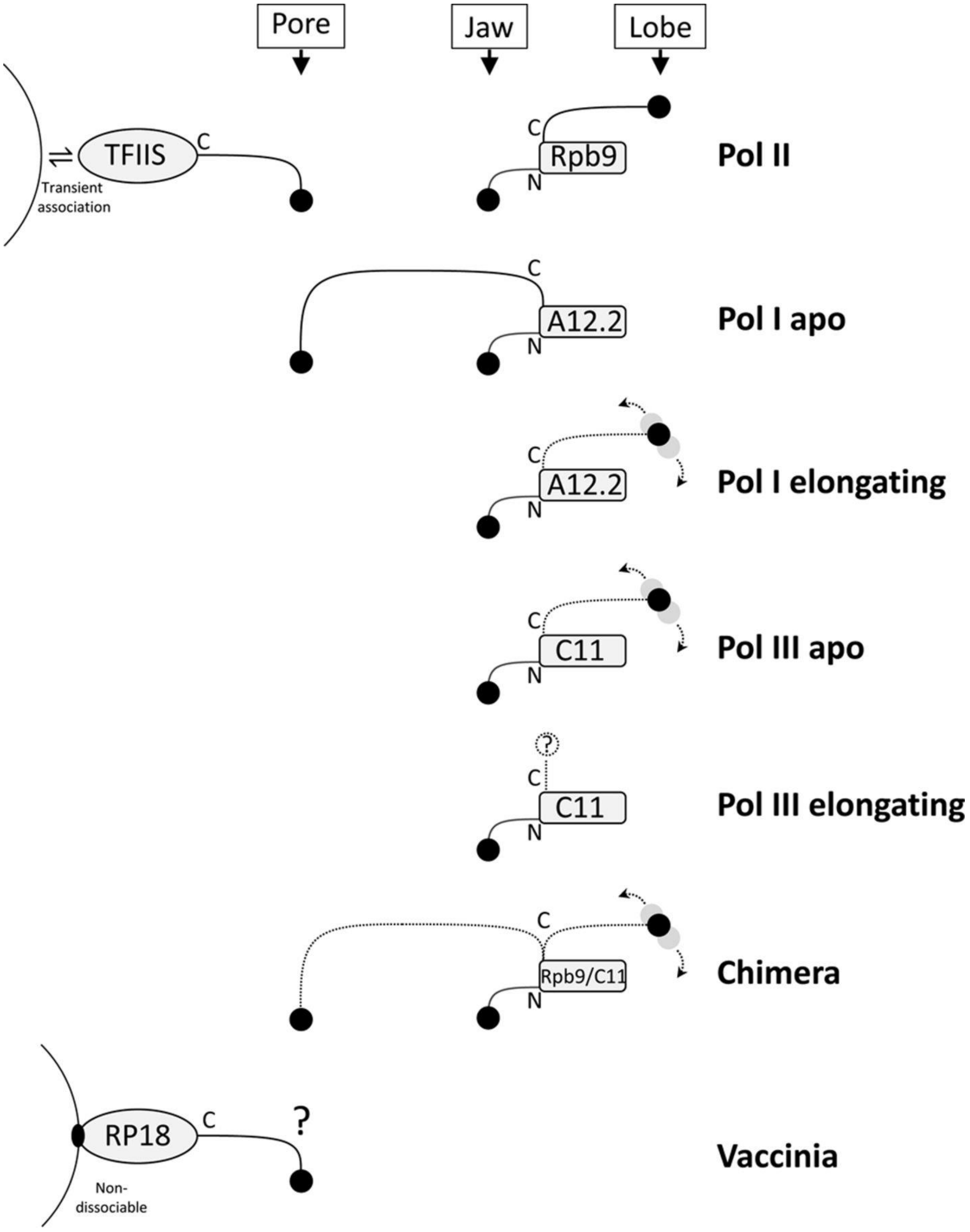


Figure 2.2. Schematic showing occupancy, by TFIIS and Rpb9 subunits and their equivalents, of three sites (NTP entry pore, jaw, and lobe, after reference [271]) in coordination with transcript cleavage activities of pol I, pol II, pol III, Rpb9-C11 chimeric pol II (51), and vaccinia virus RNA polymerase. Subunit names are as given in the text. N- and C-terminal Zn ribbons of the subunits are shown as stalks. In the chimera, the C-terminal ribbon of pol II subunit Rpb9 is replaced with the equivalent ribbon pol III subunit C11 [271]. Dotted arrows and gray circles denote mobility. For details, see the text.

Structural analysis of the resulting chimera showed that the transplanted C ribbon was detached from the site occupied by Rpb9's native C ribbon on the surface of pol II and was mobile. Mutagenesis of "catalytic" residues in the transplanted C ribbon was consistent with its hairpin transiently inserting into the pol II NTP entry pore to complement residues at the active center (Fig. 2.2).

Thus, while evolution may have rendered the pol II system controllable by the dissociable factor TFIIS for regulated pausing, the endogenous cleavage activities of pol I and pol III would tend to favor the rapid, pause-free synthesis of their abundantly required transcripts [278, 282].

The TFIIS N-terminal domain has regulatory roles in transcriptional initiation and elongational "Stop-Go"

Among pol II transcription factors, some crossover between initiation and elongation activities is now recognized. For example TFIIF, long considered a transcription initiation factor, also has a role in transcriptional elongation [265, 283]. In contrast, TFIIS, initially considered a pol II elongation factor, is now recognized to have a role in pol II initiation as indicated by yeast genetic analysis [284] and the finding of TFIIS in pol II preinitiation complexes [268]. While TFIIS's domain III is entirely dispensable for initiation activity [268], TFIIS's domain I (see above) is centrally associated with initiation and also with the higher-order regulation of transcript cleavage for rescue from transcriptional arrest (reference [268] and references therein). Examples of such regulation would include the rescue activity of the multifunctional transcriptional regulator Ccr4-Not, which docks to TFIIS domain I [285]. Other transcription elongation factors likely have comparable interactions with TFIIS ([285-289]. Consistent with its role in higher-order regulation, domain I is the most phylogenetically divergent portion of TFIIS and is also the most variable region of tissue-specific TFIIS isoforms and paralogs [290, 291].

Archaeal MSDDRPs

Archaeal transcription systems appear to be a hybrid of the eukaryotic and bacterial systems [292], with the basal transcription apparatus being more eukaryote-like [293, 294] while the transcriptional regulatory factors are more bacterial [295, 296]. Consistent with this, archaeal RNA polymerase subunit numbers and assignments are quite similar to those of yeast [297, 298]. Three-dimensional similarities were borne out quite dramatically via X-ray crystallography of the archaeal enzymes from *Sulfolobus solfataricus* and *Sulfolobus shibatae* (two species of a thermoacidophile genus from the kingdom Crenarchaeota) in the presence and absence of DNA [219, 220, 299] and the enzyme from *Thermococcus kodakarensis* (from the kingdom Euryarchaeota [221]). Some differences between eukaryotic and archaeal enzymes include a missing domain in archaeal Rpb5 forming the lower "jaw," the distant structural relationship between archaeal Rpo8 and yeast Rpb8 [300], and the unique Rpo13 subunit in archaea [301]. Indeed, Rpo8 and -13 are prominent in distinguishing MSDDRPs from different archaeal species and phyla [299]. Recent studies have shown how virology and RNA polymerase structural biology in the archaea have the capacity to cross-inform [302].

MSDDRPs across all domains of life: the NCLDV

The elaboration of a broad clade of large DNA viruses termed the NCLDV is a recent development [303] arising from the paradigm-shifting discovery, in 2003 and since, of giant viruses [51] (also termed "megavirales" [304]; for reviews, see references [305] and [306]). In addition to the giant viruses, the NCLDV "superclade" includes several of the established large-DNA virus families that feature a cytoplasmic stage, namely, the *Poxviridae*, *Iridoviridae*, *Asfarviridae*, *Ascoviridae*, and *Phycodnaviridae* [303] (Table 2.1).

Table 2.1 Current families within the NCLDV superclade (proposed order *Megavirales* [304])[a]

Family	Year Discovered	Host(s)	Replication site	Assembly site	Genome (kb)
Poxviridae	1798?	Vertebrates, insects	Cytoplasm	Cytoplasm	Linear (130–380)[c]
Asfarviridae	1921	Pigs, warthogs, insects	Cytoplasm	Cytoplasm	Linear (170–190) [c]
Iridoviridae	1966	Fish, frogs, snakes, insects	Nucleus	Cytoplasm	Linear (102–212) [d]
Ascoviridae	1983	Insects, moths	Nucleus	Cytoplasm	Circular (157–186)
Phycodnaviridae	1981	Algae	Nucleus	Cytoplasm	Linear (100–560)
Mimiviridae	**2003**	**Amoebae, zooplankton**	**Cytoplasm**	**Cytoplasm**	**Linear (~1,200)**
Marseilleviridae	**2009**	**Amoebae**	**Cytoplasm**	**Cytoplasm**	**Circular (368)**
Megaviridae	**2010**	**Amoebae**	**Cytoplasm**	**Cytoplasm**	**Linear (1,208–1,259)**
Pandoraviridae	**2013**	**Amoebae**	**Cytoplasm**	**Cytoplasm**	**Linear (1,900–2,500)**
Pithoviridae	**2014**	**Amoebae**	**Cytoplasm**	**Cytoplasm**	**Linear (610)**[d]
Faustovirus	**2015**	***Vermamoeba vermiformis***[b]	**Cytoplasm**	**Cytoplasm**	**Circular (455-470)**[e]

[a] Those families considered to be giant viruses, discovered starting in 2003, are shown in bold. Classification and tree topology are still developing, with, for example, the recently discovered dinodinavirus, faustovirus, cedratvirus, kaumoembavirus, and mollivirus also being considered members of the NCLDV superclade.

[b] A protist.

[c] Has covalently cross-linked ends and inverted terminal repeats.

[d] Circularly permuted and terminally redundant. The upper size limit is 303 kb if redundancy is included.

[e] Eight out of nine *Faustovirus* genomes were circular [307, 308].

Among all of these viruses, the best studied is arguably vaccinia virus, a prototypical member of the *Poxviridae*. Having a cytoplasmic site of replication, the poxviruses encode their own transcription and RNA modification apparatus [29, 70], including a biochemically purified and characterized 8-subunit MSDDRP [309, 310]. A 9th subunit, named RAP94, confers on the polymerase specificity for vaccinia virus early gene promoters via the heterodimeric vaccinia virus early gene transcription factor [311-313]. At the protein sequence level, the two largest subunits of the vaccinia virus MSDDRP are unequivocally orthologous to the two large subunits of cellular enzymes [314-316], although the vaccinia virus largest subunit lacks a counterpart to the repeating C-terminal domain (CTD) found in eukaryotes. The smallest subunit of the vaccinia virus polymerase, RP07, shares sequence homology with the Rpb10 subunit of yeast pol II [317] (Table 2.2), and the vaccinia virus RP30 subunit has sequence homology to eukaryotic transcription elongation factor S-II [318], referred to here as TFIIS (see above).

Table 2.2. RNA polymerase subunit orthologs across all four domains of life, with emphasis on the NCLDV. Y, ortholog found in a virus (among the NCLDV, only for vaccinia virus are orthologs named). Pink background, same polypeptide found in all three eukaryotic polymerases. Blue background, dissociable or not associated with the core polymerase. Khaki background, fused subunits. Yellow background, all other subunits. The *Iridoviridae* are shown by individual genera instead of family because of their divergence at the level of genus. Archaeal nomenclature is from reference [220]. TFIIS is shown in gray font for eukaryotic pol I and pol III because it is unclear what role(s) this protein may play for these enzymes [276]. All rows of the table are supported by complete proteomes. Table rows are in descending order of number of identified subunits. *, composite pattern over all phycodnaviruses; for individual viruses, see Table 2.3.

	1	2	3	4	5	6	7	8	9	10	11	12	13	TFIIS	RP35	RAP94	TFIIF-like		TFIIE-like		
Eukaryotic pol II	Rpb1	Rpb1	Rpb3	Rpb4	Rpb5	Rpb6	rpb7	Rpb8	Rpb9	Rpb10	Rpb11	Rpb12		TFIIS			Tfg1/Rap74	Tfg2/Rap30			
Eukaryotic pol I	A190	A135	AC40	A14	Rpb5	Rpb6	A43	Rpb8	A12.2	Rpb10	AC19	Rpb12		TFIIS			A49	A34.5			
Eukaryotic pol III	C160	C128	AC40	C17	Rpb5	Rpb6	C25	Rpb8	C11	Rpb10	AC19	Rpb12		TFIIS			C37	C53	C82	C34	C31
Archaea	Rpo1	Rpo2	Rpo3	Rpo4	Rpo5	Rpo6	Rpo7	Rpo8	RpoM	Rpo10	Rpo11	Rpo12	Rpo13	TFS							
Phycodnavirus *	Y	Y	Y		Y	Y	Y		Y	Y	Y			Y							
Megaviridae	Y	Yi	Y		Y	Y	Y		Yi	Y	Y			Y							
Mimiviridae	Y	Yi	Y		Y	Y	Y		Yi	Y	Y			Y							
ASFV	Y	Y	Y		Y	Y				Y	Y			Y							
Faustovirus	Y	Y	Y		Y	Y					Y			Y							
Chordopoxvirinae	RP147	RP132			RP22	RP19	RP18			RP07				RP30	RP35	RAP94					
Entomopoxvirinae	Y	Y				Y	Y								Y	Y					
Iridoviridae/Iridovirus	Y	Y			Y		Y			Y				Y							
Pandoravirus	Y	Y			Y					Y				Y							
Pithovirus	Y	Y			Y					Y				Y							
Marseillevirus	Y	Y			Y									Y							
Iridoviridae/Megalocytivirus	Y	Y			Y									Y							
Iridoviridae/Lymphcystis	Y	Y			Y									Y							
Iridoviridae/Chloriridovirus	Y	Y					Y							Y							
Ascoviridae	Y	Y			Y	Y															
Iridoviridae/Ranavirus	Y	Y												Y							

No three-dimensional structures are available for vaccinia virus or other NCLDV MSDDRPs or their subunits. However, the solved structures of yeast RNA polymerase II subunits Rpb5, -6, and -7 allowed them to be matched to the predicted secondary structures of vaccinia virus subunits RP22, RP19, and RP18, respectively [319] (Table 2.2). This left, as orphans, only the RP35 and RAP94 subunits from the vaccinia virus enzyme [319], and they remain so, without detectable orthologs outside the *Poxviridae*.

MSDDRP subunit assignments for the NCLDV

With the growing numbers of giant viruses being characterized, it seems apposite to revisit questions of RNA polymerase subunit assignments among the NCLDV. In our own assessment (Y. Mirzakhanyan and P. D. Gershon, unpublished data) (Tables 2.2 and 2.3), some NCLDV subunits seem to have been misannotated, while others may have been unrecognized in viral genomes. At the current state of reannotation (Mirzakhanyan and Gershon, unpublished data) (Tables 2.2 and 2.3), the enzymes from all NCLDV appear simpler than those of the eukaryotic cell, though some by not very much: among the 12 subunits of the yeast pol II/archaeal holoenzyme found in all eukaryotes and archaea, only Rpb4, -8, and -12 were

Table 2.3. Breakout of phycodnavirus RNA polymerase and transcription apparatus. Numbers 1 to 12 refer to Rpb subunits. Background blue, gray, yellow, and orange refer to RNA polymerase (including TFIIS), RNA polymerase RPB3/11 fusions, cellular transcription factor homologs, and vaccinia virus late transcription factor homologs, respectively. While one subgroup (represented by the prymnesioviruses, chrysochromulina ericina virus [CeV01], and aureococcus anophagefferens virus) seems to encode the most complete RNA polymerases of any NCLDV, another subgroup (phaeoviruses, raphidoviruses, chloroviruses, Yellowstone Lake phycodnavirus, and ostreococcus tauri virus) seems to encode no RNA polymerase at all. All rows of the table are represented by complete proteomes.

	1	2	3	4	5	6	7	8	9	10	11	12	TFIIS	TBP	TFIIB	VLTF2	VLTF3
Phaeovirus																🟧	🟧
Raphidovirus													🟦	🟨		🟧	🟧
Chlorovirus													🟦	🟨		🟧	🟧
Prasinovirus													🟦	🟨	🟨	🟧	🟧
Yellowstone lake phycodnavirus 1, 3													🟦	🟨	🟨	🟧	🟧
Yellowstone lake phycodnavirus 2														🟨	🟨	🟧	🟧
Ostreococcus tauri virus 2													🟦	🟨	🟨	🟧	🟧
Coccolithovirus	🟦	🟦	🟦		🟦					🟦			🟦			🟧	🟧
Organic lake phycodnavirus1 and 2	🟦	🟦	🟫		🟦	🟦	🟦			🟦	🟫		🟦	🟨	🟨	🟧	🟧
Prymnesiovirus	🟦	🟦	🟫		🟦	🟦	🟦		🟦	🟦	🟫		🟦	🟨	🟨	🟧	🟧
Chrysochromulina ericina virus (CeV01)	🟦	🟦	🟫		🟦	🟦	🟦		🟦	🟦	🟫		🟦	🟨	🟨	🟧	🟧
Aureococcus anophagefferens virus	🟦	🟦	🟫		🟦	🟦	🟦		🟦	🟦	🟫		🟦	🟨	🟨	🟧	🟧

universally missing among the viruses. In contrast, apart from the two large subunits (Rpb1 and -2), no subunit was universally conserved (though Rpb5 and TFIIS were almost so). Of the remaining subunits (Rpb3, -6, -7, -9, -10, and -11), although the *Megaviridae*, *Mimiviridae*, and some *Phycodnaviridae* encode representatives of each (Tables 2.2 and 2.3), they seem to be only sporadically present among other virus families. Extreme divergence cannot be ruled out as an explanation for the non-detection of viral homologs to these or of poxviral RP35 and RAP94 (see above). Biochemical isolation would be required to test the complete subunit composition of non-poxviral MSDDRPs.

The major compositional difference between the three major cellular RNA polymerases comprises the possession, by pol I and pol III, of additional subunits as fixed subcomplexes that are distant relatives of the pol II-dissociable GTFs TFIIF and TFIIE [259] (Table 2.2). Counterparts to these subcomplexes were universally absent from the NCLDV, rendering the viral polymerases more akin to pol II than to pol I or pol III in overall architecture. The roles of eukaryotic TFIIE and -F in transcription initiation include clamp opening [256], capture and stabilization of the open promoter complex, and “clearing” of protein obstructions in the cleft for loading of the template strand [320]. This kind of functionality may not be critical for viral templates to achieve promoter opening and template strand loading, due to the A-T richness of viral transcriptional start sites, their genomes being noncircular in most cases (Table 2.1), and their encoding of topoisomerases.

Diversity of the *Phycodnaviridae*

Just one family within the NCLDV superclade, namely, the *Phycodnaviridae*, seemed to represent the full gamut of MSDDRP subunit diversity across all viruses, archaea, and eukaryotes: while one subgroup of phycodnaviruses encodes the most complete RNA polymerase noted in any NCLDV (Table 2.3), another subgroup within the same family seems to be unique among all NCLDV in encoding no RNA polymerase subunits at all (Table 2.3).

The latter group presumably hijacks the host cell enzyme which, speculatively, transcribes the viral genome in combination with viral transcription factors. The above situation (factors encoded but no RNA polymerase) contrasts with the case for the *Poxviridae*, in which a functional MSDDRP is encoded, but there is a tantalizing partial reliance on cellular factors for intermediate- and late-stage transcription [29].

NCLDV with a nuclear phase: the *Iridoviridae* and *Ascoviridae*

The *Iridoviridae*, which are diverse in terms of host range and gene content [321] were, in our annotation (Table 2.2), also diverse in MSDDRP subunit composition. Across *Iridoviridae* genera, numbers of recognizable subunits reflected, approximately, the complexity of the

Table 2.4. *Iridoviridae* by genus.

		No. of ORFs in type species[b]	
Genus	**Type species[a]**	**Total**	**MSDDRP subunit**
Iridovirus	IIV-6	468	6
Lyphocystivirus	Lymphocystis disease virus, isolate China	239	4
Chloriridovirus	IIV-3	126	4
Megalocytivirus	ISKNV	125	4
Ranavirus	FV-3	98	3

[a]IIV, invertebrate iridescent virus; ISKNV, infectious spleen and kidney necrosis virus; FV, frog virus.

[b]The type species is assumed to be representative. The MSDDRP subunits include TFIIS.

virus genome (Table 2.4). The best-characterized iridovirus at the molecular level is perhaps frog virus 3 (FV-3), a ranavirus whose genome contains a total of just 98 open reading frames (ORFs) [322]. FV-3 early transcription is considered to be nuclear and mediated by host RNA polymerase II [323] with the aid of diffusible protein factors from the input virion (reference [324] and references therein). Late viral mRNAs are synthesized in the cytoplasm [325, 326]. Although newly synthesized virus-encoded RNA polymerase is implicated in late transcription [327], only the two large RNA polymerase subunits were recognizable in ranavirus genomes along with TFIIS [322] (Table 2.2). Assuming that all FV-3-encoded subunits have been accounted for, then either a eukaryote-like MSDDRP composed of just two or three subunits is sufficient for transcription, which would be unprecedented, or a host-virus chimera is used (which would also be novel). The *Ascoviridae* encode just four recognizable subunits (Table 2.2). Although the *Ascoviridae* and *Iridoviridae* are considered to be related, very little is known about ascoviral transcription [328]. The *Ascoviridae* and *Iridoviridae* TFIIS homologs are notable in being unusually small (Fig. 2.3).

Viral "Basal" transcription apparatus

In addition to RNA polymerase subunits, key, "basal" transcription factors for the NCLDV were also reannotated (Mirzakhanyan and Gershon, unpublished data) (Table 2.5). These efforts furthered the publicly available annotations substantially, though the task is not guaranteed to be complete. Factors annotated in Table 2.5 include orthologs of the vaccinia virus early transcription factor heterodimer (VETF-L and VETF-S), three of the four vaccinia virus late transcription factors (VLTF-1, VLTF-2, and VLTF-3), the obligate intermediate gene transcription factor heterodimer (VITF3-A8R and VITF3-A23R), and homologs of cellular factors TFIIB and TBP. Late factors VLTF-3 and VLTF-2 were found universally or almost

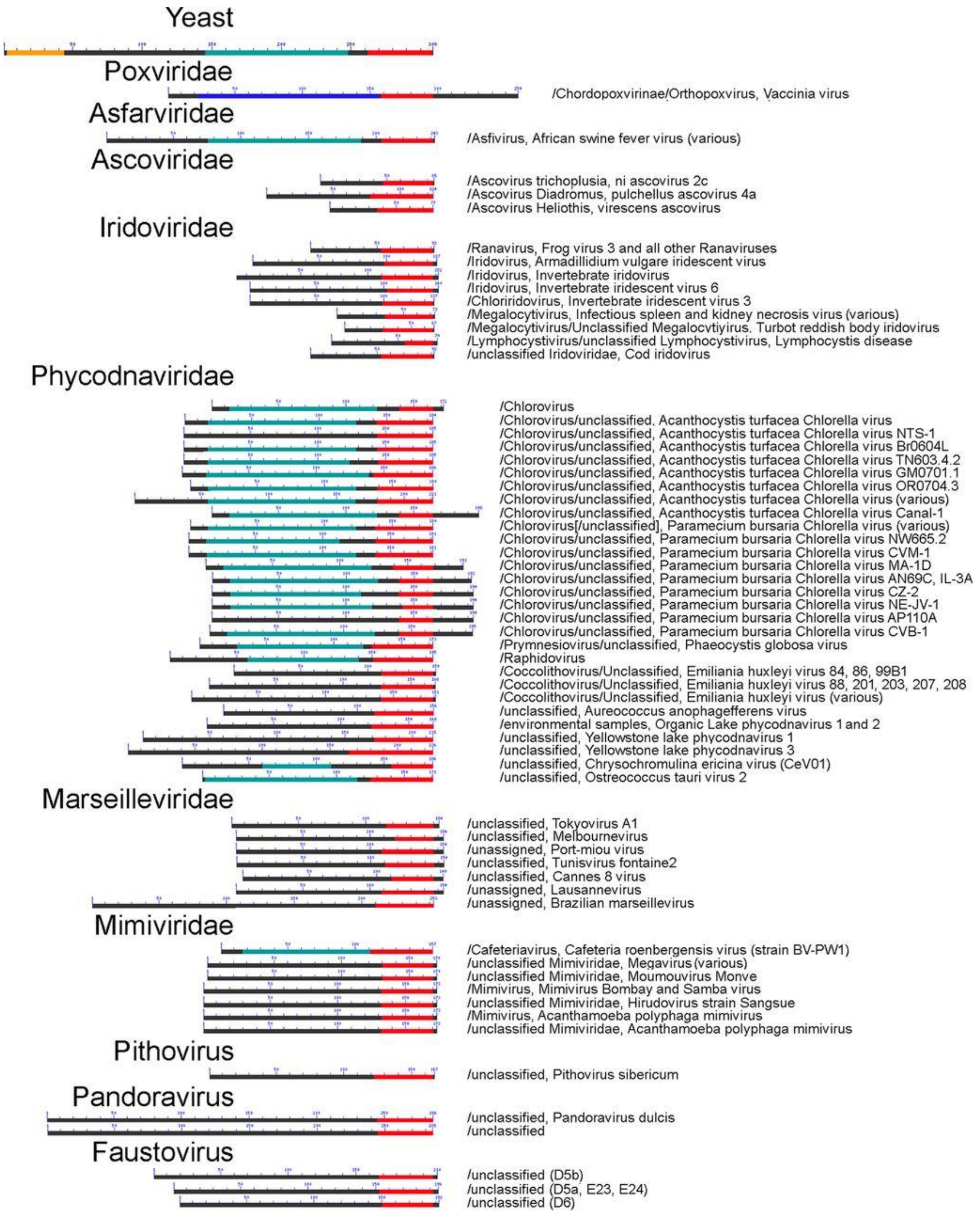

Yeast
Poxviridae
/Chordopoxvirinae/Orthopoxvirus, Vaccinia virus
Asfarviridae
/Asfivirus, African swine fever virus (various)
Ascoviridae
/Ascovirus trichoplusia, ni ascovirus 2c
/Ascovirus Diadromus, pulchellus ascovirus 4a
/Ascovirus Heliothis, virescens ascovirus
Iridoviridae
/Ranavirus, Frog virus 3 and all other Ranaviruses
/Iridovirus, Armadillidium vulgare iridescent virus
/Iridovirus, Invertebrate iridovirus
/Iridovirus, Invertebrate iridescent virus 6
/Chloriridovirus, Invertebrate iridescent virus 3
/Megalocytivirus, Infectious spleen and kidney necrosis virus (various)
/Megalocytivirus/Unclassified Megalocytiyirus, Turbot reddish body iridovirus
/Lymphocystivirus/unclassified Lymphocystivirus, Lymphocystis disease
/unclassified Iridoviridae, Cod iridovirus
Phycodnaviridae
/Chlorovirus
/Chlorovirus/unclassified, Acanthocystis turfacea Chlorella virus
/Chlorovirus/unclassified, Acanthocystis turfacea Chlorella virus NTS-1
/Chlorovirus/unclassified, Acanthocystis turfacea Chlorella virus Br0604L
/Chlorovirus/unclassified, Acanthocystis turfacea Chlorella virus TN603.4.2
/Chlorovirus/unclassified, Acanthocystis turfacea Chlorella virus GM0701.1
/Chlorovirus/unclassified, Acanthocystis turfacea Chlorella virus OR0704.3
/Chlorovirus/unclassified, Acanthocystis turfacea Chlorella virus (various)
/Chlorovirus/unclassified, Acanthocystis turfacea Chlorella virus Canal-1
/Chlorovirus[/unclassified], Paramecium bursaria Chlorella virus (various)
/Chlorovirus/unclassified, Paramecium bursaria Chlorella virus NW665.2
/Chlorovirus/unclassified, Paramecium bursaria Chlorella virus CVM-1
/Chlorovirus/unclassified, Paramecium bursaria Chlorella virus MA-1D
/Chlorovirus/unclassified, Paramecium bursaria Chlorella virus AN69C, IL-3A
/Chlorovirus/unclassified, Paramecium bursaria Chlorella virus CZ-2
/Chlorovirus/unclassified, Paramecium bursaria Chlorella virus NE-JV-1
/Chlorovirus/unclassified, Paramecium bursaria Chlorella virus AP110A
/Chlorovirus/unclassified, Paramecium bursaria Chlorella virus CVB-1
/Prymnesiovirus/unclassified, Phaeocystis globosa virus
/Raphidovirus
/Coccolithovirus/Unclassified, Emiliania huxleyi virus 84, 86, 99B1
/Coccolithovirus/Unclassified, Emiliania huxleyi virus 88, 201, 203, 207, 208
/Coccolithovirus/Unclassified, Emiliania huxleyi virus (various)
/unclassified, Aureococcus anophagefferens virus
/environmental samples, Organic Lake phycodnavirus 1 and 2
/unclassified, Yellowstone lake phycodnavirus 1
/unclassified, Yellowstone lake phycodnavirus 3
/unclassified, Chrysochromulina ericina virus (CeV01)
/unclassified, Ostreococcus tauri virus 2
Marseilleviridae
/unclassified, Tokyovirus A1
/unclassified, Melbournevirus
/unassigned, Port-miou virus
/unclassified, Tunisvirus fontaine2
/unclassified, Cannes 8 virus
/unassigned, Lausannevirus
/unassigned, Brazilian marseillevirus
Mimiviridae
/Cafeteriavirus, Cafeteria roenbergensis virus (strain BV-PW1)
/unclassified Mimiviridae, Megavirus (various)
/unclassified Mimiviridae, Moumouvirus Monve
/Mimivirus, Mimivirus Bombay and Samba virus
/unclassified Mimiviridae, Hirudovirus strain Sangsue
/Mimivirus, Acanthamoeba polyphaga mimivirus
/unclassified Mimiviridae, Acanthamoeba polyphaga mimivirus
Pithovirus
/unclassified, Pithovirus sibericum
Pandoravirus
/unclassified, Pandoravirus dulcis
/unclassified
Faustovirus
/unclassified (D5b)
/unclassified (D5a, E23, E24)
/unclassified (D6)

Figure 2.3. Conserved domain search results for TFIIS across the NCLDV. Orange, turquoise, and red bars represent motif superfamilies for the N-terminal, central, and C-terminal domains (domains I, II, and III), respectively, of TFIIS. Blue bar, poxvirus RP30 superfamily. The region of highest conservation between TFIIS and RP30 is the Zn ribbon-containing domain III [318], which is involved in transcript cleavage. This region is universally found among NCLDV TFIIS/RP30 homologs. Proteins are shown aligned according to the C terminus of this domain. "Yeast" refers to S. cerevisiae.

universally, respectively, among the NCLDV (Mirzakhanyan and Gershon, unpublished data). The small subunit of VETF (VETF-S) was annotated conservatively in the NCLDV due to the presence of a functionally distinct paralog in the *Poxviridae*, namely, NPH I (helicase motifs are present in both). The VETF large subunit (VETF-L) was found in all NCLDV except the phycodnaviruses. Viruses lacking VETF-L tended to possess, instead, a TBP homolog (Table 2.5) suggesting complementarity between these two proteins, presumably in early promoter binding. Indeed, the core poxvirus early promoter region recognized by VETF-L (110) may be considered positionally comparable to the TATA box of core cellular promoters. The TFIIB homolog found in some NCLDV (Table 2.5) may represent a substitute for late factor VLTF-3 or intermediate factor VITF-3, or it may function in an additional (unknown) stage of viral transcription. A substantial overlap was apparent between the presence of TBP and TFIIB in the *Phycodnaviridae* (Table 2.5), consistent with them acting together in this family, in a “basal”-type transcription system. Somewhat counterintuitively, the *Iridoviridae* and *Ascoviridae*, whose early gene expression is considered to be dependent on the host transcription system (see above), encoded homologs of VETF-L, an early factor. Finally, despite the size and apparent sophistication of the pandoraviruses (with the largest genome yet found for any virus, nearly twice the size of its nearest rival [Table 2.1]), only a single homolog of a known viral or basal cellular transcription factor was detected among its genes. The transcription system of this virus may be highly divergent, or there may be a cryptic involvement of cellular proteins.

TFIIS

A TFIIS homolog was found to be nearly universally encoded across the NCLDV (Table 2.2), suggesting a fundamental role in transcription that transcends the deeply divergent evolutionary pathways of the different NCLDV families. Notably, a customized TFIIS is encoded even by those phycodnaviruses lacking an endogenous RNA polymerase (Table 2.3), suggesting that viral TFIIS is an equal-opportunity employer of viral or host cell polymerases and can displace the endogenous, host cell TFIIS from cellular pol II. The only two families/subfamilies lacking a TFIIS homolog were the ascoviruses and entomopoxviruses (Table 2.2). Tantalizingly, ascoviruses and entomopoxviruses are distinct among the NCLDV in infecting, primarily, lepidoptera (butterflies and moths), more specifically the family Noctuidae.

With regard to TFIIS function, the poxvirus homolog, RP30, is reported to have dual roles in transcription: as an intrinsic RNA polymerase subunit within the virion, where free RP30 does not exist [318], and as a free protein in the infected-cell cytoplasm, where it acts as an initiation factor for one of the three transcriptional classes in vaccinia virus, namely, the intermediate class [329]. The intrinsic subunit has a presumed role in anti-arrest [318], consistent with the nascent transcript cleavage activity demonstrated for purified ternary complexes of vaccinia virus early transcription complexes [330]. The second function of RP30, as an initiation factor in vaccinia virus, has parallels with the finding of TFIIS in cellular pol II initiation complexes (see above). However, the absence of key intermediate transcription factors in NCLDV other than the *Poxviridae* (Table 2.5) suggests that this function is poxvirus specific. Interestingly, RP30's C-ribbon is flanked by a 62-aa C-terminal tail which, among the NCLDV,

Table 2.5. Transcription factors across the NCLDV. Background green, blue, orange, and yellow refer to factors corresponding to poxvirus early, intermediate, and late transcriptional stages and unknown transcriptional stage, respectively. VETF-S was annotated conservatively for the NCLDV due to the presence of a functionally distinct paralog NPH-I in the *Poxviridae.* Font colors blue and red indicate *Phycodnaviridae* and *Iridoviridae*, respectively. Horizontal divisions denote patterns of presence/absence. For other details, see the text.

	TBP-like	TFIIB-like	VETF-S	VETF-L	VITF3-A8R	VITF3-A23R	VLTF-1	VLTF2-like	VLTF3-like
Chordopoxvirinae			🟩	🟩	🟦	🟦	🟧	🟧	🟧
Entomopoxvirinae			🟩	🟩	🟦	🟦	🟧	🟧	🟧
ASFV		🟨		🟩				🟧	🟧
Faustovirus		🟨		🟩				🟧	🟧
Marseillevirus		🟨		🟩				🟧	🟧
Megaviridae		🟨		🟩				🟧	🟧
Mimiviridae		🟨		🟩				🟧	🟧
Pithovirus		🟨		🟩				🟧	🟧
Phycodnaviridae/Yellowstone lake phycodnaviruses	🟨	🟨						🟧	🟧
Phycodnaviridae/Prasinovirus	🟨	🟨						🟧	🟧
Phycodnaviridae/Ostreococcus tauri virus	🟨	🟨						🟧	🟧
Phycodnaviridae/Prymnesiovirus	🟨	🟨						🟧	🟧
Phycodnaviridae/Chrysochromulina ericina virus (CeV01)	🟨	🟨						🟧	🟧
Phycodnaviridae/Aureococcus anophagefferens virus	🟨	🟨						🟧	🟧
Phycodnaviridae/Organic lake phycodnaviruses	🟨	🟨						🟧	🟧
Phycodnaviridae/Raphidovirus	🟨							🟧	🟧
Phycodnaviridae/Chlorovirus	🟨							🟧	🟧
Phycodnaviridae/Phaeovirus								🟧	🟧
Phycodnaviridae/Coccolithovirus								🟧	🟧
Ascoviridae				🟩				🟧	🟧
Iridoviridae/Lymphcystis				🟩				🟧	🟧
Iridoviridae/Iridovirus				🟩				🟧	🟧
Iridoviridae/Chloriridovirus				🟩				🟧	🟧
Iridoviridae/Ranavirus				🟩				🟧	🟧
Iridoviridae/Megalocytivirus				🟩					🟧
Pandoravirus									🟧

is also unique to the *Poxviridae* (Fig. 2.3). In the virion-packaged form of RP30, this Pro/Ser-rich tail was recently shown to be highly phosphorylated [331]. It remains to be proven whether phosphorylation modulates a switch between RP30's two functions specifically in the *Poxviridae*.

Viral subunit function in the context of structure

With the wealth of structural information for cellular MSDDRPs alongside the revised NCLDV subunit annotations (see above), questions of structural and functional specialization in the viral enzymes can be addressed in relation to viral replication strategies. Some general themes are discussed below using, as a model, the best studied viral enzyme, namely, the one from vaccinia virus.

Dissociable subunits in Pol II are integral (non-dissociable) in the vaccinia virus enzyme

For some subunits that are firmly attached to the vaccinia virus RNA polymerase, the cellular pol II equivalents are readily dissociable. Examples include the dissociable Rpb4/7 ("stalk") complex of yeast pol II (see above), whose vaccinia virus equivalent, RP18, is non-dissociable. Similarly, while transcription elongation factor TFIIS only transiently associates with pol II during transcription (see above), the vaccinia virus equivalent, RP30, is an integral subunit of the vaccinia virus enzyme, remaining stubbornly polymerase associated during gradient sedimentation and attempted column-based antibody affinity separation [318]. The non-dissociability of these subunits has parallels in the cellular realm: the pol I and pol III counterparts to pol II's stalk, namely, A14/A43 and C17/C25, respectively (Table 2.2) are, like the corresponding RP18 subunit of vaccinia virus, non-dissociable [332]. Moreover, the pol I and pol III functional equivalents of TFIIS/RP30 (subunits A12.2 and C11, respectively [see above]) are also non-dissociable [230, 231, 233]. Whether the giant virus homologs of TFIIS are integral to their core polymerases or dissociable is unknown.

Why might vaccinia virus RP30 be non-dissociable? Just as pol I and pol III have refined their intrinsic anti-arrest activity for the synthesis of relatively few, general purpose RNAs in large quantities without “traffic jams” of paused or arrested polymerase (see above), so vaccinia virus may have opted for rapid waves of processive viral transcription, maximizing the accumulation of virus proteins and nascent virions before the host either dies or restricts the virus. Moreover, intrinsic anti-arrest activity may have favored greater genomic sequence flexibility during the evolution and diversification of large viral genomes, especially if all transcriptional pausing or arrest sites could not be fully eliminated. Other possible explanations for a tight association of RP30 with vaccinia virus RNA polymerase, as suggested upon RP30’s discovery and characterization [318], were to ensure RP30 packaging during virion assembly and its introduction into infected cells in stoichiometric amounts with the viral RNA polymerase.

Why is the vaccinia virus stalk subunit (RP18) non-dissociable? During pol II initiation, the stalk has a role in transient opening of the clamp structure for entry of the template stand into the pol II cleft via either its own transient dissociation or its recruitment of initiation factor TFIIE that can actively open the clamp [333]. Closure of the clamp around the template is associated with the presence of the stalk [235], specifically, the Rpb7 subunit [235]. Within transcription elongation complexes, the clamp is always observed closed, even in the absence of Rpb4/7 [224]. In pol I, pol III, and archaeal RNA polymerase [301], the stalk subunits (Table 2.2) are non-dissociable, consistent with which, the pol I clamp appears to be permanently closed [231], or at least pol I has not yet been co-crystallized in the presence of putative factors that lead to transient clamp opening. In pol III (the apo enzyme), open and closed conformations of the clamp correspond to two distinct conformations of the non-dissociable stalk [233]. It seems that, as in pol I and pol III, vaccinia virus may have opted for a permanently closed clamp during elongation and the possibly greater processivity this may confer. Whatever mechanism of initiation is employed by pol I and pol III would presumably be reflected in the vaccinia virus enzyme.

Pol II proteins and assemblies are reduced to vestigial stubs in vaccinia virus

While the RNA polymerase cleft and catalytic center are highly conserved in all domains of life, evolutionary plasticity seems to follow the smaller subunits. In the case of the NCLDV, this has included the formation of architectural "stubs," as described below.

Stub 1: vaccinia virus subunit RP30. Like vaccinia virus RP30, NCLDV TFIIS orthologs seem to mostly lack an N-terminal region sufficiently long to reflect the yeast N-terminal domain that mediates higher-order regulation of pausing (Fig. 2.3). The absence of this domain would suggest an unresponsiveness to higher-level transcriptional regulation (see above). This is consistent with the comparatively simple genomes and expression patterns typically found in viruses and with a presumptive need to get proteins made and virions assembled as rapidly as possible while escaping from cellular regulation. A possible exception, however, would be the large and sophisticated pandoraviruses, whose N-terminal domains approach the length of cellular TFIIS (Fig. 2.3). Consistent with this, the pandoraviruses currently having the largest genomes of any virus found, by a factor of 2 [50]. Perhaps, as a viral genome approaches the complexity of the simplest cellular genome, there may be pressure to retain or invoke a more sophisticated regulatory mechanism for transcription elongation.

Stub 2: the "Stalk." As discussed above, eukaryotic and archaeal MSDDRPs possess a "stalk" structure located toward the upstream face and which, in pol II, comprises the dissociable Rpb4/7 heterodimer (see above). Within this heterodimer, Rpb7 alone contacts the core polymerase [225] and is an essential subunit in yeast [225]. Yeast Rpb4, which has regulatory roles in, for example, the stress response [334-336], is nonessential for viability [337]. In pol I, "stalk" subunits (Table 2.2) provide a platform for the binding of initiation factors [278, 338-340]. In all NCLDV, however, the polymerase stalk appears to be a "stub," with a homolog of Rpb4 entirely missing from all viral MSDDRPs characterized thus far (Table 2.2).

Stub 3: the Rpb3-10-11-12 subassembly/subcomplex and the upstream face. Biogenesis of the pol II core likely arises through three independent assembly subpathways, based around the two large subunits Rpb1 and Rpb2 and subunit Rpb3 [332]. The Rpb3 subassembly comprises subunits Rpb3, -10, -11, and -12 [222]. pol I, pol III, and archaeal RNA polymerase have equivalent subunits (Table 2.2). During biogenesis, the Rpb3-10-11-12 subcomplex is considered to nucleate the assembly of the holoenzyme, as the alpha subunit homodimer does in bacterial RNA polymerase [332]. In mature pol II this subassembly is located on the upstream face (facing the promoter). Rpb12 is an essential subunit of pol I, pol II, and pol II in yeast and may have a role in maintaining the open promoter during initiation.

While all viruses lack a homolog to the essential eukaryotic rpb12, some (the *Mimiviridae*, the *Megaviridae*, *Faustovirus*, African swine fever virus, and some *Phycodnaviridae*) encode a fusion of subunits Rpb3 and Rpb11 (Table 2.2), and many encode, in addition, an Rpb10 homolog. The vaccinia virus enzyme, however, which is a good biochemical benchmark for the viral enzymes due to its extensive purification and characterization, is remarkable for the highly vestigial character of this subassembly: the only identifiable member is

RP07, a homolog of the very small eukaryotic subunit Rpb10 (Table 2.2). This represents just 11% of the mass of the subassembly in pol II and is unprecedented among all cellular MSDDRPs (bacterial, eukaryotic, or archaeal).

Why is the Rpb3 subassembly minimal in many of the NCLDV? As with other "stubs" (see above), Rpb3 participates in regulatory interactions (reference [341] and references therein). Loss of subunits from the Rpb3 subassembly may therefore be associated with the elimination of higher-order regulatory interactions. Nonetheless, assuming that Rpb3 is central to polymerase assembly in all domains of life and with Rpb12 being an essential subunit apparently for promoter opening [342], it is unclear how these critical functions may be recapitulated in many or all of the NCLDV. It has been speculated that an "orphan" subunit in vaccinia virus of similar size, namely, RP35, may compensate for the absence of Rpb3 in the *Poxviridae*, although RP35 is clearly structurally unrelated [319]. However, RP35 is not found outside the *Poxviridae*, with many NCLDV lacking both RP35 and Rpb3 (Table 2.2).

Stub 4: Rpb5, a "Lower-Jaw" subunit. The Rpb5 subunit of yeast is a 215-aa two domain protein [343]. However, the corresponding archaeal subunit is only 84 aa in length due to an entirely missing N-terminal domain. Rpb5 is located at the lower "jaw" of pol II (see above). The presence/absence of the N domain is, in fact, not critical for polymerase function insofar as the archaeal and yeast subunits can cross-complement [343]. However, in common with the above-described theme, Rpb5 seems to mediate regulatory interactions [344-346]. In the *Coccolithovirus* genus of *Phycodnaviridae* as well as in two ascoviruses, the Rpb5 homolog is equivalent in size to the archaeal subunit. Although this may provide a means to escape regulatory interactions, the homologs in other NCLDV, including vaccinia virus, are close in size to the yeast protein. Whether the N domains of these homologs are sufficiently divergent to have adopted novel functions is unclear.

Subunit Rpb9: modulation of intrinsic mutability?

The majority of the NCLDV examined, including vaccinia virus, lack an ortholog of pol II subunit Rpb9 (Table 2.2). In pol II, this subunit is implicated in transcriptional fidelity (via proofreading [see above]). In contrast to the delicately regulated and maintained eukaryotic cell, fidelity in virus transcripts may be unimportant, even to the extent that defective whole particles are typically well tolerated in virus biology. Thus, there may be few negative consequences of an occasional transcriptional error. Viruses, which are intrinsically mutagenic, do not have long-term health considerations at the level of the individual organism; only the population matters. Moreover, for mRNA in any living system, translational errors tend to swamp transcriptional ones by a substantial margin, providing an incentive for pol I and pol III to be error proof in the production of their highly recyclable and potentially mutagenic transcripts but not necessarily for pol II to be so in the production of mRNA [282]. The loss of Rpb9 in the majority of NCLDV but its retention in others may have been a selectable property in maintaining a balance between fatal error and evolutionary velocity for genomes of various sizes experiencing various evolutionary pressures.

Conclusion

To summarize, while they retain sophistication, NCLDV RNA polymerases are to an extent stripped and honed with respect to their cellular counterparts during their adaptation and specialization for the purposes of speed, processivity, and escape from higher-order cellular control. From what we know so far of their structure-function relationships, these architectural changes seem rational, and they suggest that the many regions of viral polymerases that have universally survived the brutal journey of virus evolution are central to polymerase function, even if that function is not yet known for the cellular and viral enzymes. There is a lot we do not know about the functions of some of the smaller subunits and domains in polymerases in general, but there is every reason to believe that as structural biochemistry teaches us more, the bigger picture of subunit and domain changes and refinements found in each virus family and genus will begin to make sense also.

Materials and Methods

For each of the viruses and virus families described in this chapter, annotated RNA polymerase subunits from UniProt were searched using the Uniprot BLAST toolkit to identify sequence homologs that were either uncharacterized or where the annotation was incomplete. Proteins were also characterized be sequence based structural homology detection by HHpred on the MPI Bioinformatics Toolkit server [347, 348].

CHAPTER 3

Structure-based deep mining reveals first-time annotations for 46 percent of the dark annotation space of the 9,671 member superproteome of the nucleocytoplasmic large DNA viruses

Summary

We conducted an exhaustive search for three-dimensional structural homologs to the proteins of 20 key phylogenetically distinct nucleocytoplasmic DNA viruses (NCLDV). Structural matches covered 429 known protein domain superfamilies, with the most highly represented being ankyrin repeat, P-loop NTPase, F-box, protein kinase, and membrane occupation and recognition nexus (MORN) repeat. Domain superfamily diversity correlated with genome size, but a diversity of around 200 superfamilies appeared to correlate with an abrupt switch to paralogization. Extensive structural homology was found across the range of eukaryotic RNA polymerase II subunits and their associated basal transcription factors, with the coordinated gain and loss of clusters of subunits on a virus-by-virus basis. The total number of predicted endonucleases across the 20 NCLDV was nearly quadrupled from 36 to 132, covering much of the structural and functional diversity of endonucleases throughout the biosphere in DNA restriction, repair, and homing. Unexpected findings included capsid protein-transcription factor chimeras; endonuclease chimeras; enzymes for detoxification; antimicrobial peptides and toxin-antitoxin systems associated with symbiosis, immunity, and addiction; and novel proteins for membrane abscission and protein turnover.

Importance

We extended the known annotation space for the NCLDV by 46%, revealing high-probability structural matches for fully 45% of the 9,671 query proteins and confirming up to 98% of existing annotations per virus. The most prevalent protein families included ankyrin repeat- and MORN repeat-containing proteins, many of which included an F-box, suggesting extensive host cell modulation among the NCLDV. Regression suggested a minimum requirement for around 36 protein structural superfamilies for a viable NCLDV, and beyond around 200 superfamilies, genome expansion by the acquisition of new functions was abruptly replaced by paralogization. We found homologs to herpesvirus surface glycoprotein gB in cytoplasmic viruses. This study provided the first prediction of an endonuclease in 10 of the 20 viruses examined; the first report in a virus of a phenolic acid decarboxylase, proteasomal subunit, or cysteine knot (defensin) protein; and the first report of a prokaryotic-type ribosomal protein in a eukaryotic virus.

Introduction

The 2003 discovery of the first giant virus, mimivirus [51], proved transformative to virology and added new context to the established large DNA virus families (*Poxviridae*, *Iridoviridae*, and *Chlorellaviridae*). A decade later, a new "giant of giants," pandoravirus, with its 2.7-Mb genome encoding more than 2,500 proteins [50], dwarfed the 800-kb mimivirus genome by more than equal measure. The past decade has seen the characterization of many new large DNA virus genomes via the integration of metagenomics, next-generation nucleic acid sequencing, more proficient sequence alignment algorithms [349], and greater interconnectivity of bioinformatics resources for the fast and automated annotation of genes, proteins, protein folds, and protein domains. There are now as many as nine families of nucleocytoplasmic large

DNA viruses (NCLDV), whose shared characteristics include a greater or lesser degree of cytoplasmic involvement in their replication, independence from the host replication machinery, large DNA genomes, and genes for DNA replication, DNA repair, transcription, and mRNA translation [350].

Although many NCLDV genes have been annotated for function, comprehensive genome annotation is confounded by the minimal or nonexistent conservation of amino acid sequence across a broad swath of evolutionary space. In just one example familiar to the authors, a classical Rossmann fold was revealed within the crystal structure for vaccinia protein VP39 [351], whose existence had been entirely unpredictable on the basis of sequencing despite the well-established nature of this fold and the many proteins containing it. New additions to the BLAST pipeline, including BLASTP, PSI-BLAST, and BLASTCLUST, have helped to some extent in closing the annotation gap [349]. The use of tertiary structural information, however, may be a much more sensitive method for finding matches whose similarities have fully decayed at the protein sequence level. The detection of distant sequence homology has been sensitized by the use of sequence substitution “profiles” treated as hidden Markov models (HMMs) of multiple-sequence alignments (MSAs) of the growing numbers of members of various protein families. Using tools such as PfamScan [352, 353], individual sequence queries can be searched against profile MSA HMMs, driving the expansion of the Pfam database of known protein families [354]. More powerful, although currently lacking in PfamScan, would be an ability to perform profile MSA versus profile MSA searches. Such searches led to the prediction of NCLDV members of the archaeoeukaryotic primase superfamily [355]. In our hands, PfamScan seemed slow to update its profiles and seemed to overlook structural homologs we were otherwise able to find in the pdb70 database (unpublished data).

One powerful package, HHsuite [356], employs profile-profile alignments to identify homologous proteins, starting with the creation of MSAs for query proteins and then embellishing these with secondary structural prediction. HMM profiles of the resulting MSAs are searched against a database of HMMs derived from bona fide experimental protein structures (PDB or SCOP). The combination of sequence and secondary structural alignments and the use of real structures provides a potentially powerful tool for protein families with marginal or absent sequence similarity, and has the potential to harvest the biosphere-wide structural proteomics initiative of the earlier part of the current millennium [357]. HHsuite has been applied in a number of problems, including the prediction of open reading frames [358], analysis of G protein-coupled receptors [359], identification of novel protein repeats [360], prediction of poxviral RNA polymerase homologs [61, 319], and the identification of PH domains in the S. cerevisiae proteome [361]. Here, we have applied the HHsuite toolbox more comprehensively, providing the first exhaustive search of the proteomes of 20 NCLDV-type members, identifying protein superfamily members among previously uncharacterized proteins and filling gaps in the NCLDV core proteome. We have expanded our previously published work of multisubunit DNA-directed RNA polymerase (MSDDRP) subunits and predicted a number of viral protein homologs not previously identified.

Results and Discussion

Here, we have "deep-mined" new protein annotations in a selection of 20 phylogenetically distinct NCLDV chosen to cover all known NCLDV families, key subfamilies, genera, species, and unclassified viruses therein (Table 3.1). Mining was based on tertiary structure homology. Proteomes of the 20 viruses comprised a total of 9,671 proteins, from each of which an HMM was derived via a combination of MSA and predicted secondary structure.

Each resulting HMM was used to query an HMM database generated from actual protein tertiary structures deposited in pdb70. The search output for each query protein, showing all matching pdb70 entries/regions, was thresholded according to a probability parameter calculated by the search engine. An 80% threshold was chosen for the probability parameter based on the initial descriptions of HHsuite [356, 362, 363] and prior literature [361] in which a probability threshold of 80% yielded a false-positive rate of just 0.15%. In the current study, the best-scoring database match exceeded the 80% probability threshold for 45% of the 9,671 query proteins and fell within the topmost (99 to 100%) probability bin for fully 23.8% of proteins (Fig. 3.1A). This provided bootstrap confirmation of our chosen probability threshold. Apparently, our approach could successfully uncover structural homologs for nearly half of all NCLDV proteins - in the vast majority of cases covering most of the length of the query and target proteins (Fig. 3.2).

Where an unknown NCLDV query protein matched a pdb70 entry of known function, this annotation was transferred directly to the NCLDV query protein. Since functions are already known for a substantial proportion of proteins resident in pdb70, there were frequent opportunities for such "annotation transfer."

Prior to the current study, the 20 query proteomes were annotated to a variable extent, the most incompletely and completely annotated being chlorella virus (7.4%) and vaccinia virus (89%), respectively (Fig. 3.1B, total green). The current study confirmed between 20% (iridovirus) and 98% (pithovirus) of existing annotations (Fig. 3.1B, dark green versus total green), validating our structure-based approach. Perhaps more interestingly, our approach provided first-time

Table 3.1. The 20 phylogenetically distinct NCLDV analyzed in this study (listed alphabetically)^b .

Virus (local name)	Family	Subfamily	Genus	Scientific or common name	Strain	UniProt proteome identifier	No. of proteins	UniProt reference proteome [c]
Ascovirus	*Ascoviridae*		Ascovirus	Heliothis virescens ascovirus 3g		UP000232493	194	N
Asfarvirus	*Asfarviridae*		Asfivirus	African swine fever virus Georgia 2007/1		UP000141072	188	N
Chlorella virus [a]	*Phycodnaviridae*		Chlorovirus	Paramecium bursaria Chlorella virus 1 (PBCV-1)		UP000000862	794	Y
Emiliania huxleyi virus [a]	*Phycodnaviridae*		Coccolithovirus	Emiliania huxleyi virus 86 (EhV-86)	Isolate United Kingdom/ English Channel/1999	UP000000863	472	Y
Entomopox alpha	*Poxviridae*	*Entomopoxvirinae*	Alphaentomopoxvirus	Anomala cuprea entomopoxvirus		UP000174145	241	Y
Entomopox beta	*Poxviridae*	*Entomopoxvirinae*	Betaentomopoxvirus	Choristoneura biennis entomopoxvirus (CbEPV)		UP000014934	311	Y
Entomopox unclassified	*Poxviridae*	*Entomopoxvirinae*		Melanoplus sanguinipes entomopoxvirus (MsEPV)	Isolate Tucson	UP000172353	261	Y
Faustovirus	Unclassified viruses			Faustovirus sp.	E12	UP000244833	492	Y
Iridoviridae/Chloriridovirus	*Iridoviridae*	*Betairidovirinae*	Chloriridovirus	Invertebrate iridescent virus 3 (IIV-3) (mosquito iridescent virus)		UP000001358	126	Y
Iridoviridae/Iridovirus	*Iridoviridae*	*Betairidovirinae*	Iridovirus	Invertebrate iridescent virus 6 (IIV-6) (Chilo iridescent virus)		UP000001359	469	Y
Iridoviridae/Lymphocystivirus	*Iridoviridae*	*Alphairidovirinae*	Lymphocystivirus	Lymphocystis disease virus	Isolate China	UP000106699	239	Y
Iridoviridae/Megalocytivirus	*Iridoviridae*	*Alphairidovirinae*	Megalocytivirus	Infectious spleen and kidney necrosis virus (ISKNV)	Isolate Mandarin fish/China/Nanhai/1998	UP000008773	125	Y
Iridioviridae/Ranavirus	*Iridoviridae*	*Alphairidovirinae*	Ranavirus	Frog virus 3 (isolate Goorha) (FV-3)		UP000008770	98	Y
Marseillevirus	*Marseilleviridae*		Marseillevirus	Marseillevirus marseillevirus (GBM)		UP000029780	428	Y
Megavirus	*Mimiviridae*		Mimivirus	Megavirus courdo11 (unclassified Mimivirus)		UP000241137	1217	Y
Mimivirus	*Mimiviridae*		Mimivirus	Acanthamoeba polyphaga mimivirus (APMV)	Rowbotham-Bradford	UP000201519	979	N
Mollivirus	Unclassified viruses			Mollivirus sibericum		UP000202709	514	Y
Pandoravirus			Pandoravirus	Pandoravirus inopinatum		UP000202511	1839	Y
Pithovirus	*Pithoviridae*		Pithovirus	Pithovirus sibericum		UP000202176	467	Y
Vaccinia	*Poxviridae*	*Chordopoxvirinae*	Orthopoxvirus	Vaccinia virus	Western Reserve (WR)	UP000000344	217	Y

[a] Two clades within the *Phycodnaviridae* based on RNA polymerase subunit presence/absence (16). Chlorella virus represents the following genera: *Chlorovirus*, *Phaeovirus*, *Raphidovirus*, *Prasinovirus*, Yellowstone lake phycodnaviruses 1 and 2, and *Ostreococcus tauri virus 2*. *Emiliania huxleyi virus* represents the following genera: *Coccolithovirus*, *Prymnesiovirus*, Organic lake phycodnaviruses 1 and 2, *Chrysochromulina ericina virus*, and *Aureococcus anophagefferens virus*.

[b] A total of 9,671 proteins were analyzed among the 20 viruses. Reviewed proteomes were used where available. All taxonomic rankings (family, subfamily, genus, species, and strain) are according to UniProt's "Lineage" fields. Where multiple alternate species or strains were available, the one with the largest proteome was chosen (with the exception of Vaccinia, where the WR strain was used in place of Tian Tan). This table represents the group of NCDV analyzed previously (16) with the following differences: two classes only of Phycodnavirus are considered (see the text), the three subfamilies of *Entomopoxvirinae* are considered separately, and *Mollivirus* has been added. Although *Megavirus courdo11* is an unclassified member of the *Mimiviridae*, it is referred to here as *Megavirus*. After its initial discovery, *Megavirus* was regarded as a new virus family (99, 100), but after phylogenetic studies of additional new *Mimivirus* and *Megavirus* genomes, *Megavirus* was instead grouped with the *Mimiviridae* (101).

[c] N, no; Y, yes.

annotations for many previously uncharacterized proteins from each of the 20 selected viruses. First-time annotations covered between 15% (entomopoxvirus alpha) and 39% (chloriridovirus) of the previously uncharacterized segments of virus proteomes (Fig. 3.1B, dark red versus total red). Apparently, substantial inroads could be made into the uncharacterized proteomes of the NCLDV via structure-based homology.

NCLDV matches to annotated pdb70 entries were formalized into functional classes by visual inspection of, in each case, the HMM homology region in the pdb70 target in order to find overlapping entries in the Pfam [354] protein domain family database. Pfam tagging in this manner accounted for 87.5% of all of the NCLDV proteins showing structural homologs, covering a total of 429 Pfams (see Appendix1.Figure1; the top 50 Pfams are shown in Fig. 3.3A). Pfams with the greatest overall representation among the 20 viruses comprised the ankyrin repeats (636 proteins), P-loop NTPases (288 proteins), F-box proteins (222 proteins), protein kinases (155 proteins), and membrane occupation and recognition nexus (MORN) repeat-containing proteins (149 proteins) (Fig. 3.3A and Appendix1.Figure1).

Proteins in these five families were particularly prevalent among the giant viruses (megavirus, mimivirus, marseillevirus, pandoravirus, and pithovirus). For the 20 viruses, good correlation ($R^2 = 0.88$) was observed between the number of proteins in the proteome versus the number of distinct protein domain superfamilies represented therein (Fig. 3.3B), suggesting that NCLDV genome expansion has gone hand in hand with the acquisition of novel superfamily functions (with larger genomes being more functionally diverse). Interestingly, there was a single, and quite dramatic outlier to this correlation, namely, pandoravirus, whose proteome is the largest by far among all currently known viruses. Despite its genome being larger than that of its closest neighbor (megavirus) by a factor of 1.5, this was not accompanied by any net increase

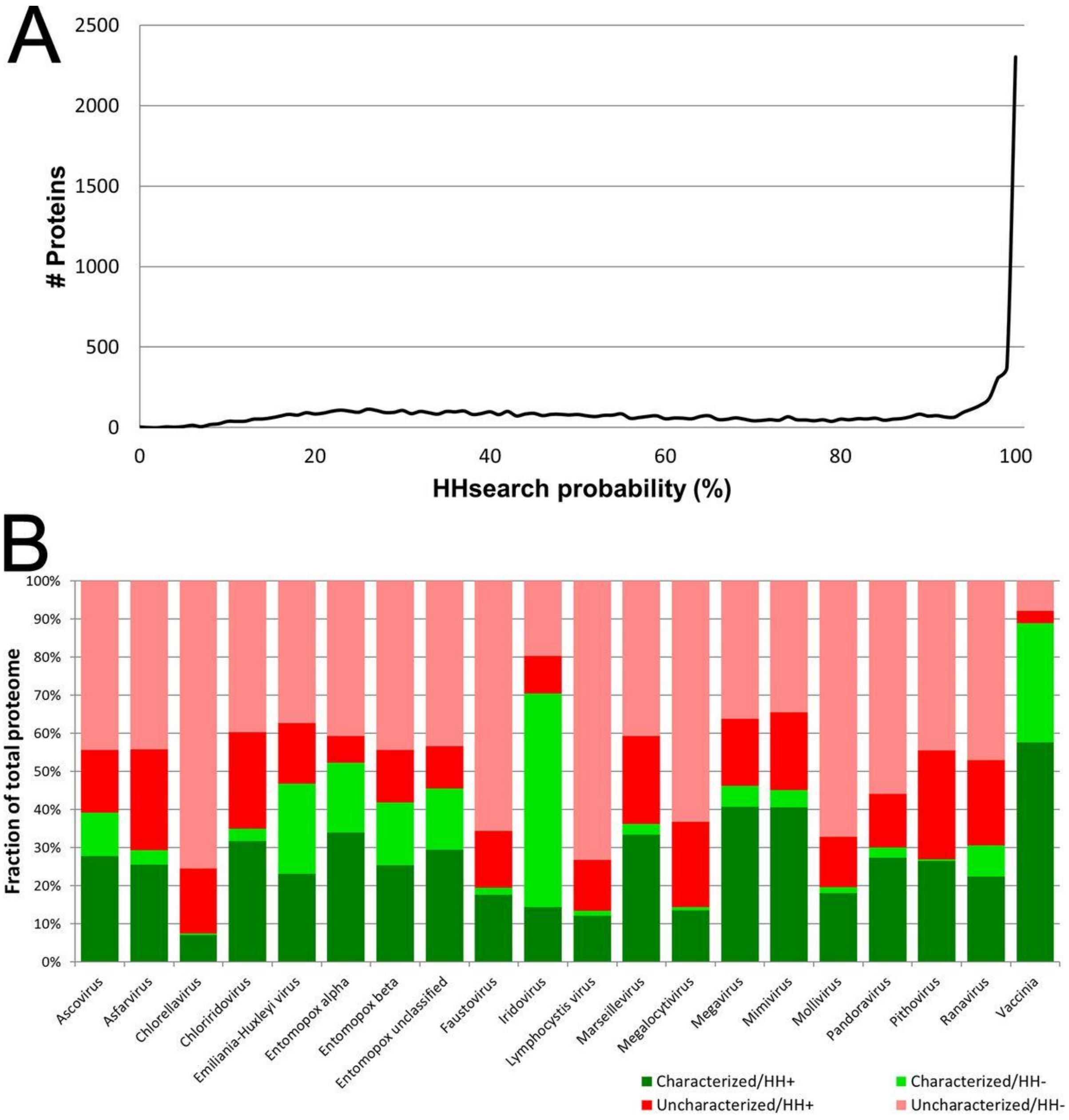
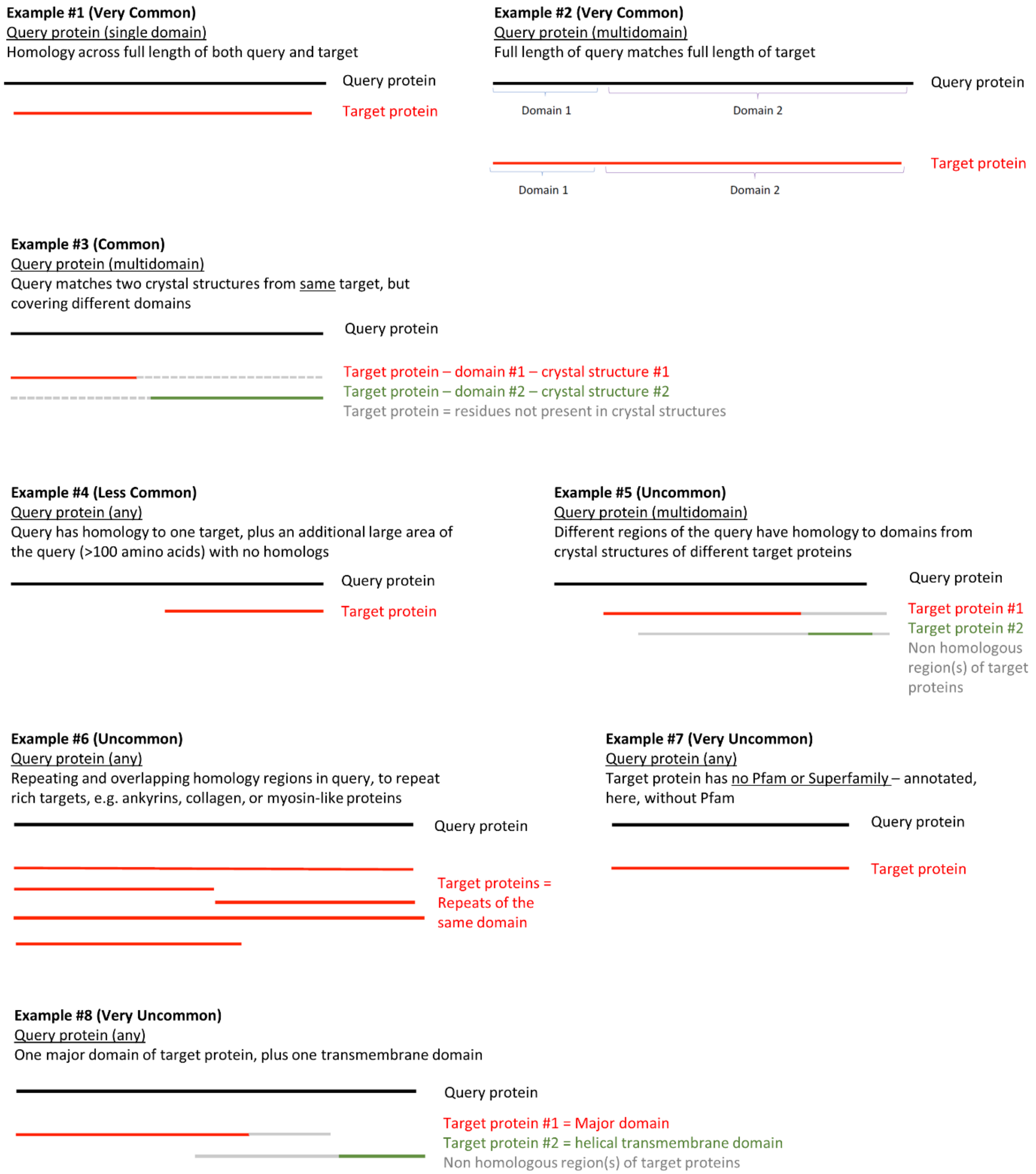


Figure 3.1 Structure-based deep mining of functions for 20 representative NCLDV proteomes. (**A**) Line histogram of HHsearch "probability" score associated with best HHsearch database match to each of the 9,671 query proteins over 20 representative NCLDV proteomes. The x axis shows HHsearch probability parameter (bin width, 1%). The y axis shows number of proteins associated with each bin. (**B**) Extension of existing annotations. Green/red indicates the fraction of virus proteome possessing or lacking, respectively, an associated UniProt functional annotation prior to the current study. Red-group proteins were annotated in UniProt either as "uncharacterized protein" or with an annotation comprising simply the submitted gene name.

Dark/light green indicates the fraction of UniProt-annotated proteomes confirmed or not confirmed, respectively, by HHsearch at or above the 80% probability threshold. Dark/light red indicates the fraction of UniProt-unannotated proteomes for which HHsearch did or did not, respectively, provide a first-time functional annotation at or above the 80% threshold. In generating the dark green region, agreement between annotated NCLDV query proteins and corresponding pdb70 hits was assessed conservatively, as either an identical stated protein function, keyword, or leaf gene ontology (GO) term. Proteomes showing low overall annotation rates prior to the current study (green region below 20% on the y axis; namely, chlorella virus, lymphocystis virus, megalocytivirus, and mollivirus) may have been handicapped by a slow synchronization between UniProt and the Pfam and InterPro databases.



Example #1 (Very Common)
Query protein (single domain)
Homology across full length of both query and target

Query protein
Target protein

Example #2 (Very Common)
Query protein (multidomain)
Full length of query matches full length of target

Query protein
Domain 1
Domain 2
Target protein
Domain 1
Domain 2

Example #3 (Common)
Query protein (multidomain)
Query matches two crystal structures from same target, but covering different domains

Query protein
Target protein – domain #1 – crystal structure #1
Target protein – domain #2 – crystal structure #2
Target protein = residues not present in crystal structures

Example #4 (Less Common)
Query protein (any)
Query has homology to one target, plus an additional large area of the query (>100 amino acids) with no homologs

Example #5 (Uncommon)
Query protein (multidomain)
Different regions of the query have homology to domains from crystal structures of different target proteins

Example #6 (Uncommon)
Query protein (any)
Repeating and overlapping homology regions in query, to repeat rich targets, e.g. ankyrins, collagen, or myosin-like proteins

Example #7 (Very Uncommon)
Query protein (any)
Target protein has no Pfam or Superfamily – annotated, here, without Pfam

Example #8 (Very Uncommon)
Query protein (any)
One major domain of target protein, plus one transmembrane domain

Query protein
Target protein #1 = Major domain
Target protein #2 = helical transmembrane domain
Non homologous region(s) of target proteins

Figure 3.2 Substantial query/target overlap in the overwhelming majority of matches. Overview of the range of match-types encountered**.**

in the numbers of protein domain superfamilies in the pandoravirus proteome - rather, there was actually a 20% decrease in superfamily diversity compared with that of mimivirus (Fig. 3.3B). Apparently, there is a threshold above which the gain of superfamily diversity (new orthologs) has no appreciable selective advantage in relation to the diversification of existing ones (new paralogs). “Paralogization” seems to have taken over as an evolutionary driver at an apparently quite definable point in genome growth. Nonetheless, this conclusion is based on just one data point. Overall, mimivirus and megavirus showed the greatest proteomic diversity in terms of total protein superfamilies represented in their proteomes (Fig. 3.3B, green points). Conversely, extending the linear regression line back to near x-y parity (x = 39 proteins; y = 36 superfamilies), suggested a minimum requirement of around 36 core superfamilies for a viable NCLDV. This would be within range of the 47 NCLDV orthologous (core) genes uncovered by others [36, 44, 303, 350, 364] and discussed further below.

Pan-NCLDV orthologous genes

Protein sequence homology studies have identified a set of nucleocytoplasmic virus orthologous genes/groups (NCVOGs)—genes conserved across the NCLDV [350]. Updated listings have accompanied the discovery of additional NCLDV [36, 49, 303, 364-367], and the current NCVOG count stands at 47 [365], with few changes accompanying more recent virus discoveries [44]. Few of the NCVOGs are universally conserved among the NCLDV. Via our structure-based approach, viral coverage was extended in 44 of the 47 NCVOGs (21 NCVOGs if excluding entomopox beta, which was not included in prior analyses, Table 3.2). For two NCVOGs, namely, RING-finger E3 ligase and the “pfam02902 Ulp1 protease family” (Table 3.2), orthologs were found for the first time in seven distinct viruses. With the finding of

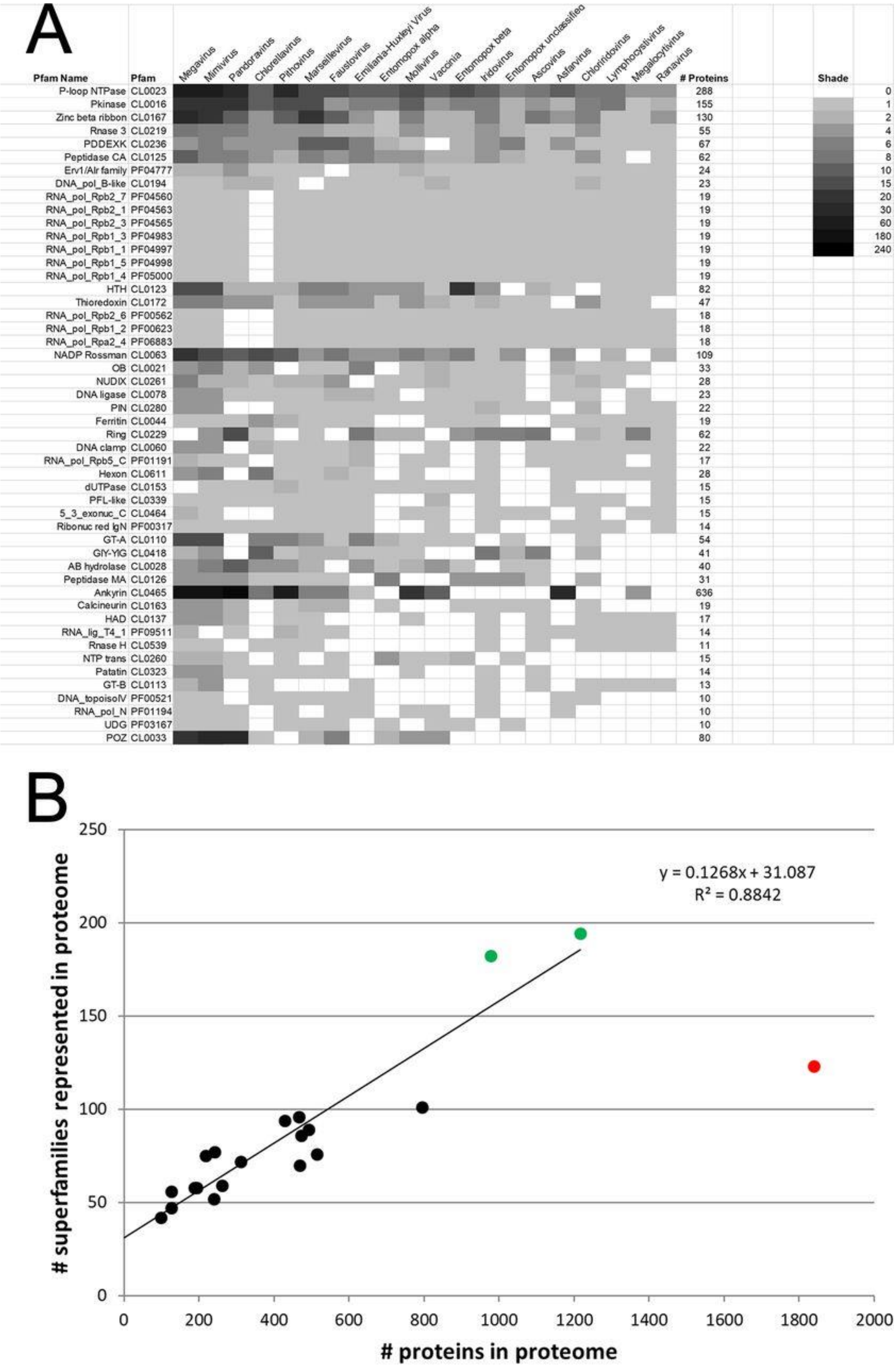

A
Pfam Name
Pfam
Megavirus
Mimivirus
Pandoravirus
Chlorellavirus
Pithovirus
Marseillevirus
Faustovirus
Emiliania-Huxleyi Virus
Entomopox alpha
Mollivirus
Vaccinia
Entomopox beta
Iridovirus
Entomopox unclassified
Ascovirus
Asfarvirus
Chloriridovirus
Lymphocystivirus
Megalocytivirus
Ranavirus
# Proteins
Shade
P-loop NTPase CL0023 288
Pkinase CL0016 155
Zinc beta ribbon CL0167 130
Rnase 3 CL0219 55
PDDEXK CL0236 67
Peptidase CA CL0125 62
Erv1/Alr family PF04777 24
DNA_pol_B-like CL0194 23
RNA_pol_Rpb2_7 PF04560 19
RNA_pol_Rpb2_1 PF04563 19
RNA_pol_Rpb2_3 PF04565 19
RNA_pol_Rpb1_3 PF04983 19
RNA_pol_Rpb1_1 PF04997 19
RNA_pol_Rpb1_5 PF04998 19
RNA_pol_Rpb1_4 PF05000 19
HTH CL0123 82
Thioredoxin CL0172 47
RNA_pol_Rpb2_6 PF00562 18
RNA_pol_Rpb1_2 PF00623 18
RNA_pol_Rpa2_4 PF06883 18
NADP Rossman CL0063 109
OB CL0021 33
NUDIX CL0261 28
DNA ligase CL0078 23
PIN CL0280 22
Ferritin CL0044 19
Ring CL0229 62
DNA clamp CL0060 22
RNA_pol_Rpb5_C PF01191 17
Hexon CL0611 28
dUTPase CL0153 15
PFL-like CL0339 15
5_3_exonuc_C CL0464 15
Ribonuc red lgN PF00317 14
GT-A CL0110 54
GIY-YIG CL0418 41
AB hydrolase CL0028 40
Peptidase MA CL0126 31
Ankyrin CL0465 636
Calcineurin CL0163 19
HAD CL0137 17
RNA_lig_T4_1 PF09511 14
Rnase H CL0539 11
NTP trans CL0260 15
Patatin CL0323 14
GT-B CL0113 13
DNA_topoisoIV PF00521 10
RNA_pol_N PF01194 10
UDG PF03167 10
POZ CL0033 80
0
1
2
4
6
8
10
15
20
30
60
180
240
B
y = 0.1268x + 31.087
R² = 0.8842
# superfamilies represented in proteome
# proteins in proteome
0
50
100
150
200
250
0
200
400
600
800
1000
1200
1400
1600
1800
2000

Figure 3.3. Pfams across the NCLDV. (**A**) From structural homology searches of all 9,671 proteins from the 20 viruses (covering both previously annotated and unannotated proteins), matches passing the 80% probability threshold were assigned to Pfam superfamilies according to Pfam tags mapping to the homology overlap region. A total of 429 protein superfamilies could be assigned, the top 50 of which are shown here. See Appendix1.Figure1 for all 429 superfamily assignments and the ranking method. Grayscale indicates the number of matching proteins per superfamily per virus. Individual query proteins matching multiple superfamilies above the 80% probability threshold were included in the counts for multiple superfamilies according to the rules given in Materials and Methods. Superfamilies (here) are also referred to as “clans” by Pfam. (**B**) Superfamily diversity versus proteome size for the 20 NCLDV. For each virus, the total number of proteins in its proteome (x axis) is charted against the total number of superfamilies found among proteome members (summed from panel a; y axis). Green points indicate the two viruses with the greatest Pfam diversity (megavirus and mimivirus). Red point indicates pandoravirus, a major outlier. The linear regression trendline (extended back to the y axis) applies to all datapoints except that for pandoravirus.

structural homologs in four viruses, coverage of the transcription elongation factor TFIIS was extended to cover all 20 viruses (Table 3.2). The four TFIIS paralogs found in pandoravirus alone (all of which were previously annotated as uncharacterized proteins) supported the expansion of the ultralarge pandoravirus genome by a paralogization mechanism (Fig. 3.3B).

The "major capsid protein" NCVOG (Table 3.2) covers two distinct protein superfamilies: CL0055 ("nucleoplasmin-like viral coat and capsid proteins superfamily"), which is based on the jelly roll fold and includes the "D13-like" external scaffold of the *Poxviridae* [190], and CL0611 ("hexon-like superfamily"), which includes beta-sandwich viral coat proteins such as the recently characterized chlorella virus capsid [106]. Here, domains belonging to CL0611 were found in proteins from mimivirus and megavirus, previously annotated "BTB/POZ-containing," in which the capsid-like domain was fused to the C-terminal side of the BTB/POZ domain in a novel chimeric arrangement (Fig. 3.4). In addition, structural matches were found to a third capsid protein superfamily, CL0605 ("single-stranded DNA [ssDNA] viruses nucleoplasmin-like/VP coat superfamily"). Proteins in the latter superfamily comprise a beta sandwich with two sheets in a jelly roll topology, as found, for example, in coat protein VP2 of parvo-like virus AAV2. Among our 20 viruses, structural matches to CL0605 were found exclusively in the *Entomopoxvirinae*, and all had a TFIIS-type zinc finger fused to the protein N terminus (Fig. 3.4). These orthologs had no BLASTP counterparts in any organism outside the *Entomopoxvirinae* (not even a vaccinia ortholog, for example), and no other member of CL0605 possessed a TFIIS-type fusion. Nonetheless, examples of viral structural proteins incorporating zinc fingers have been reported, include the retrovirus nucleocapsid protein [368] and the reovirus capsid proteins delta 3, sigma 3, and lambda 1 [369-372]. Notably, the entomopoxvirus proteomes were found to possess no TFIIS-like protein other than the TFIIS-capsid protein fusion (Table 3.3).

Structure-based deep mining also revealed two novel "capsid-like" members of superfamily CL0611 from chlorella virus, which had chitin-binding domains fused either centrally or at the C terminus. However, this was described by others in detail while the current article was in preparation [106].

Consistent with sequence-based homology approaches [44], structure-based deep mining revealed no capsid-like protein of any kind in the pandoravirus proteome despite the very large size of its proteome and the apparent central role of capsid-like proteins for viruses in general.

Table 3.2 Expanded coverage by structure-based deep mining of 47 genes previously designated NCVOGs. NCVOGs [44, 350, 364, 365], conserved among NCLDV on the basis of protein sequence homology, are ordered (left to right) by descending coverage among our 20 NCLDV. NCVOGs are named according to additional file 4 in Yutin et al. [365]. "X" (green) indicates prior coverage on the basis of sequence homology (see references [44] and [366] and references therein); "Y" (lilac) indicates NCVOG additions in Koonin and Yutin [44]. "H" (mustard) indicates new coverage via structure-based deep mining. H*, TFIIB structural homolog lacking zinc finger domain. H**, known KilA-N domain protein from literature (vaccinia virus protein p28 [373-375]). H***, known RPB5 homolog from literature (vaccinia virus protein RP22 [319]). "NH" indicates deep mining result supported by UniProt protein name (not present in prior NCVOG analyses). Since prior NCVOG analyses did not include entomopox beta, coverage for this virus was partially elucidated by BLASTP search ("B," yellow). "BH" (brown) indicates combination of BLASTP and deep mining. NK1, NK2 (pink), nucleoside kinase-type (NK) proteins are dually listed between NCVOGs 0319 and 0320 to cover both the original designation and our interpretation. Coverage may appear low for some NCVOGs since they were designated as such on the basis of all known NCLDV [366] as opposed to the 20 representatives considered here.

	NCVOG	Ascovirus	Asfarvirus	Chlorellavirus	Emiliania Huxleyi virus	Entomopox alpha	Entomopox beta	Entomopox unclassified	Faustovirus	Iridoviridae/Chloriridovirus	Iridoviridae/Iridovirus	Iridoviridae/Lymphocystis virus	Iridoviridae/Megalocytivirus	Iridoviridae/Ranavirus	Marseillevirus	Megavirus	Mimivirus	Mollivirus	Pandoravirus	Pithovirus	Vaccinia
A32 packaging ATPase	0249	X	X	X	X	X	BH	X	Y	X	X	X	X	X	X	X	X	Y	X	Y	X
disulfide (thiol) oxidoreductase; Erv1/Alr	0052	X	X	X	X	Y	BH	X	Y	X	X	X	X	X	X	X	X	Y	X	X	X
DNA Pol B	0038	X	X	X	X	X	BH	X	NH	X	X	X	X	X	X	X	X	Y	X	X	X
D5-like helicase	0023	X	X	X	X	X	BH	X	Y	X	X	X	X	X	X	X	X	Y	X	X	X
VLTF3	0262	X	X	X	X	X	B	X	Y	X	X	X	X	X	X	X	X	Y	X	Y	X
TFIIS	0272	H	X	X	X	H	H	H	Y	X	X	X	X	X	X	X	X	Y	X	X	X
Major capsid protein	0022	X	X	X	X	X	BH	X	Y	X	X	X	X	X	X	X	X	Y		X	X
VLTF2	1164	X	X	X	X	X	B	X	Y	X	X	X		X	X	X	X	Y	X	X	X
RPB1	0274	X	X		X	X	BH	X	Y	X	X	X	X	X	X	X	X	Y	X	X	X
RPB2	0271	X	X		X	X	BH	X	Y	X	X	X	X	X	X	X	X	Y	X	X	X
FLAP endonuclease	1060	X			X	X	BH	X	Y	X	X	X	X	X	X	X	X	Y		X	X
DNA or RNA helicases superfamily II	0076		X	X	X	X	BH	X	Y	X	X			X	X	X	Y	Y	X	X	X
RING-finger E3 ligase	0330	X		X	X	H	H	X		H	H	H	X	H	X	X	X	H	X		X
VETF	0261	X	X			X	B	X	Y	X	X	X	X	X	X	X	X			X	X
Nudix	0236		X	H		X	BH	X	Y	X	X	X			X	X	X	NH	X	X	X
Ribonucleotide reductase small	0276		X	X	X		H		Y	Y	Y	Y	X	Y	X	X	X		X	X	X
RPB5	0273	X	X		X				Y	X	X	X		X	X	X	X	Y	X	X	H***
mRNA capping enzyme large	1117		X	Y	Y	X	BH	X	Y				Y		X	X	Y	Y	Y	X	X
Ribonucleoside disphosphate reductase alpha	1353		X	X	X				Y	X	X	X		X	X	X	X		X	X	X
pfam02442; myristylated IMV protein	0211	X	X			X	B	X		X	X	X	X	X		X	X				X
dUTPase	1068		X	X	X	X	H		H		X			X	X			H	X	X	X
TFIIB	1127	H*	X	X					Y	H*	X			H*	X	X	X		H*	X	
RuvC, holliday	0278			X		X	BH	X		X	X				X	X	X		H	H	X
pfam02223 thymidylate kinase	0320	NK1-A	X	H†	X	H	NH			NK1-I	NK1/2-I	NK1-I	NK1-I	NK1-I							X

Vaccinia	Pithovirus	Pandoravirus	Mollivirus	Mimivirus	Megavirus	Marseillevirus	Iridoviridae/Ranavirus	Iridoviridae/Megalocytivirus	Iridoviridae/Lymphocystis virus	Iridoviridae/Iridovirus	Iridoviridae/Chloriridovirus	Faustovirus	Entomopox unclassified	Entomopox beta	Entomopox alpha	Emiliania Huxleyi virus	Chlorellavirus	Asfarvirus	Ascovirus	NCVOG	
	X	H			H	X	X	X	X	X	X	H						H	X	1088	RNA ligase
X		H	H	X	X	X	NK1-I	NK1-I	NK1-I	NK1-I	NK1-I			BH	X			X	NK1-A	0319	Thymidine kinase
	X			X	X	X				X	X	Y				X	X	X		0037	DNA topoisomerase II
H				X	X	H					H	H	H	H	H			X		0246	pfam02902 Ulp1 protease family
X	X	X		X	X	X						H	X	BH	X					1115	uracil-DNA glycosylase
X				X	X		X		X	X	X		X	B	X					1122	Myristylated protein pfam03003; DUF230
				X	X	X				X	X		X	BH	X			X	X	1361	pfam10544; ASPES
		X	NH	X	X							H		H	X		X	X		1192	YqaJ recombinase
	X			X	X	X							X	NH	X			X		0004	AP endonuclease family 2
				X	X					X	X		X	B	X				X	0010	Bro-N; pfam02498
X		X		X	X			X		X	X						X			0040	Dual specificity phosphatase
X	X			X	X								X	BH	X	H				0256	IMV protein p35
X	X					X						H				X	X	X		0034	ATP dependent DNA ligase pfam01068
				X	X					X	X		X	BH	X					0035	NAD+ dependent DNA ligase
X				X	X								X	BH	X			X		0267	RNA-helicase DExH-NPH-II
										X			X	BH	X			X	NH	0009	pfam00653; BIR
X				X	X								X	BH	X					0036	DNA topoisomerase I
		H		X	X							H			X			X		0329	UBCc; Ubiquitin conjugating E2
H**				X	X					X				BH	X					1360	KilA pfam04383
				X	X					X				B	X				X	1424	uncharacterized domain downstream KilA, BRO, MSV199
X																X	H	X		0012	C-type lectin; pfam00059; pfam05966
	X			X		X												X		0024	pfam03121 superfamily II helicase
				X	X							H						X		0059	FtsJ-like methyltransferase pfam01728

For some NCVOG proteins, deep mining revealed additional paralogs within a virus proteome. For example, two copies each of the RPB1, RPB2, and RPB5 subunits of DNA-directed RNA polymerase (below) were found within the megavirus proteome (data not shown).

Multisubunit DNA-dependent RNA polymerase and transcription factor orthologs.

Eukaryotes encode 12-subunit DNA-dependent RNA polymerases (DDRPs) comprising two large subunits (RPB1 and RPB2) and a number of smaller ones. As long established, the poxviruses encode an 8-subunit enzyme comparable in architecture to the eukaryotic one [29]. Subunits of the vaccinia enzyme are orthologous to the eukaryotic subunits (see [29] and references therein) and other NCLDV (see reference [61] and references therein), among which

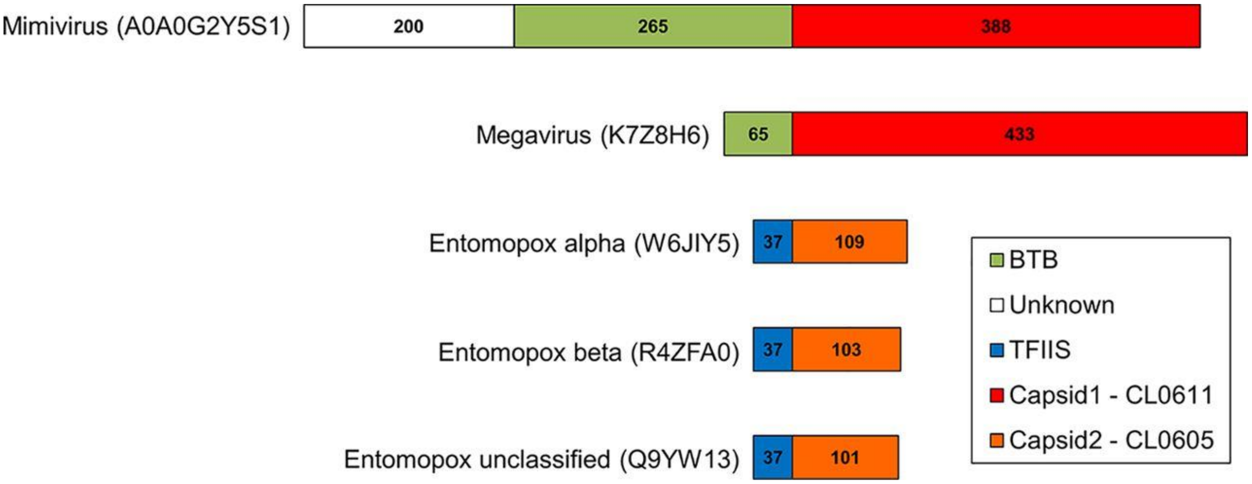


Figure 3.4. Chimeric "capsid-like" proteins in the NCLDV. A BTB/POZ domain is fused N-terminally to the capsid-like domain (Pfam CL0611; see text) of mimivirus and megavirus proteins A0A0G2Y5S1 and K7Z8H6, respectively (both annotated in UniProt as "putative BTB/POZ domain-containing protein"). A TFIIS-type zinc finger is fused N-terminally to the capsid-like domain (Pfam CL0605) of entomopoxvirus alpha, beta, and unclassified proteins W6JIY5, R4ZFA0, and Q9YW13, respectively. These were annotated in UniProt as "capsid protein, polyoma VP1-like," "uncharacterized protein," and "uncharacterized protein MSV079," respectively. These proteins had no BLASTP counterparts outside the *Entomopoxvirinae*. The *Entomopoxvirinae* were unique among the NCLDV examined here in possessing two capsid-like proteins each. The second one, a CL0611 superfamily member, is likely an external scaffold used during virion morphogenesis (see the text).

the two largest, RPB1 and RPB2, are well conserved at the sequence level [61] (Table 3.3). So far, just one NCLDV, namely chlorella virus, has failed to yield any known DDRP subunits at all via any search method, including the structure-based deep mining here (Table 3.3). This failure included a thorough inspection of all chlorella virus search results in the current study for matches below our 80% probability threshold. This nondetection strongly reinforced a conclusion that, perhaps uniquely among the NCLDV, chlorella virus does not encode a DDRP enzyme. It does, nonetheless, encode orthologs of transcription initiation factors TBP and TFIIB and transcription elongation factor TFIIS (Table 3.3).

Viral orthologs of the eukaryotic smaller subunits have proven much more elusive than those of the two large subunits, due to their much weaker protein sequence conservation. Earlier studies [61, 319] demonstrated the potential for structure-based homology searches to find small subunits among the NCLDV and highlighted some instances of their prior misannotation. Here, we have completed the structure-based mining of DDRP small-subunit genes in the NCLDV

Table 3.3. Orthologs of yeast MSDDRP subunits and basal transcription factors in NCLDV found by all methods. RPB12 homologs, which are considered separately in Fig. 3.6a, were omitted. Apart from vaccinia virus, yeast nomenclature is used. Green indicates MSDDRP subunits annotated correctly prior to Mirzakhanyan and Gershon (16). Yellow indicates MSDDRP subunit annotation newly presented in Table 2 of Mirzakhanyan and Gershon [61] via sequence homology searching. Gray indicates the same as yellow, but split gene [61]. Cyan indicates newly identified here by structure-based deep mining. Rows are ordered/underlined according to a phylogenetic tree inferred from a binary trait matrix of subunit/transcription factor presence/absence. Although classified appropriately according to Pfam and InterPro databases, some subunits were not annotated accordingly in UniProt. The yeast subunit nomenclature provides a basis for nomenclature unification across the NCLDV. *, 78% probability. **, TFIIB cyclin domain only. ***, Entomopox TFIIS is fused to the N terminus of a capsid-like protein (Fig. 3.4; see text).

	1	2	3	4	5	6	7	8	9	10	11	12				
	RPB1	RPB2	RPB3	RPB4	RPB5	RPB6	RPB7	RPB8	RPB9	RPB10	RPB11	RPB12	RP35	TFIIS	TFIIB	TBP
					RPABC1	RPABC2	RPA43	RPABC3		RPABC5	RPB3	RPABC4				
												RPC10				
Pandoravirus	🟩	🟩			🟩					🟩				🟨	🟦	
Pithovirus	🟩	🟩			🟩					🟩				🟩	🟨	🟦
Iridoviridae/Iridovirus	🟩	🟩			🟩		🟨			🟨				🟩	🟦	🟦
Entomopox alpha	🟩	🟩				🟩	🟩						🟩	🟦 ***		
Entomopox beta	🟩	🟩				🟩	🟩						🟩	🟦 ***		
Entomopox unclassified	🟩	🟩				🟩	🟩						🟩	🟦 ***		
Vaccinia	🟩 RP147	🟩 RP132			🟩 RP22	🟩 RP19	🟩 RP18			🟩 RP07			🟩 RP35	🟩 RP30		
Marseillevirus	🟩	🟩			🟩									🟩	🟩	🟦
Ascovirus	🟩	🟩			🟨	🟨								🟨	🟦 **	🟦
Mollivirus	🟩	🟩			🟦					🟩				🟩		
Emiliania Huxleyi virus	🟩	🟩	▨		🟩	🟦		🟦		🟩	▨			🟩		
Faustovirus	🟩	🟩	▨		🟩	🟩			🟦	🟦	▨			🟩	🟦	🟦
Megavirus	🟩	🟩	▨		🟩	🟩	🟩		🟩	🟩	▨			🟩	🟩	🟩
Mimivirus	🟩	🟩	▨		🟩	🟩	🟩		🟩	🟩	▨			🟩	🟩	🟩
Asfarvirus	🟩	🟩	▨		🟨	🟨	🟦 *		🟦	🟨	▨			🟨	🟦	🟦
Chlorellavirus														🟨	🟦	🟨
Iridoviridae/Chloriridovirus	🟩	🟩			🟦		🟨							🟩	🟦 **	🟦
Iridioviridae/Ranavirus	🟩	🟩			🟦									🟨	🟦	
Iridoviridae/Megalocytivirus	🟩	🟩												🟨		
Iridoviridae/Lymphocystivirus	🟩	🟩			🟨		🟦							🟩		

(Table 3.3). Yeast RNA polymerase II shows an RPB3-10-11-12 subassembly [222]. The generally coincident presence/absence of RPB3, RPB10, and RPB11 in the NCLDV is now quite clear (Table 3.3), suggesting the coordinated acquisition/loss of this subassembly during viral evolution. Interestingly, the presence/absence of this subassembly seemed partially complementary to that of poxviral subunit RP35. The absence of both RPB3-10-11 and RP35 in ascovirus, the *Iridoviridae*, and the giant viruses marseillevirus, mollivirus, pandoravirus, and pithovirus raises the possibility of an as-yet-undetected complementary subunit or subassembly for these viruses that is unrecognizable in the absence of functional enzyme purification. The entomopoxvirus MSDDRP seems distinct from vaccinia in containing no obvious homolog of RP22 or RP07 by structural homology or BLASTP.

An additional finding was the presence of an apparent RPB8 subunit in EhV-86, with 95% probability (Table 3.3). RPB8 has previously been found only in eukaryotes [376] and some archaea (hyperthermophilic Crenarchaeota and "*Candidatus Korarchaeota*" [300, 377]). The finding of RPB8 in a virus is therefore novel. Using the EhV-86 ortholog as a BLAST query, all additional RPB8 orthologs were from other emiliania huxleyi viruses. The resulting protein cluster showed strong amino acid sequence conservation with no indels (Fig. 3.5). In contrast, emiliania huxleyi virus and *Saccharomyces cerevisiae* RPB8 protein sequences showed only weak sequence similarity, along with substantial truncation of the emiliania huxleyi virus protein relative to yeast (Fig. 3.5). It is not clear why viral representation of RPB8 would be confined to just a single genus (*Coccolithovirus*) of a single NCLDV family (the *Phycodnaviridae*). A prior study [61] suggested two patterns of overall RNA pol subunit representation among the *Phycodnaviridae*, in which the coccolithoviruses could be grouped with the *Prymnesiovirus* genus, chrysochromulina ericina virus (CeV01), aureococcus anophagefferens virus, and unclassified Organic Lake phycodnavirus 1 and 2. However, RPB8 was not detected in any of these relatives.

In eukaryotes, basal promoter utilization by RNA polymerase II is mediated by two basal transcription factors, namely TATA binding protein (TBP), which binds the TATA element of the eukaryotic promoter, and TFIIB, whose C-terminal cyclin domains interact with TBP and whose N-terminal zinc finger interacts with RNA polymerase II. By structure-based deep mining, novel TBP and TFIIB orthologs were predicted in 8 and 10 NCLDV, respectively (Table 3.3), although ascovirus and chloriridovirus TFIIB orthologs possessed only the cyclin domains and not the zinc finger. TFIIB was previously designated an NCVOG (Table 3.2) via protein sequence

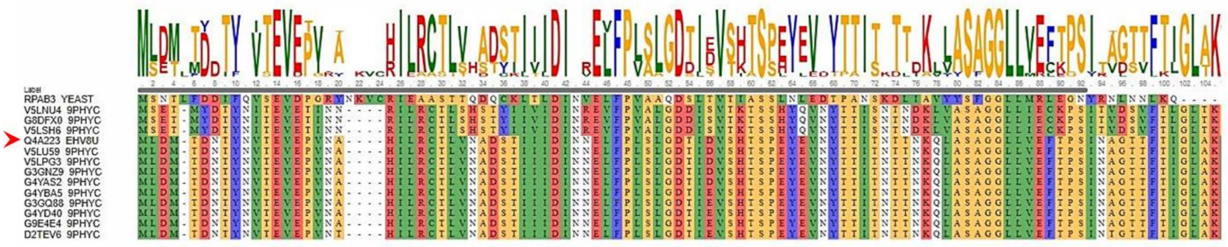


Figure 3.5. Multiple sequence alignment (ClustalW) of a cluster of RNA polymerase subunit RPB8 homologs encoded by *Emiliania-Huxleyi viruses* found using Q4A223_EHV8U (arrowed) as a BLASTP query. BLASTP e-values ranged from 10^{34} to 10^{73}. The alignment includes yeast RPB8 (top) from which amino acids 72 - 107 were removed (since they were absent from all viral sequences). Colors show similarity in amino acid chemical properties, and the consensus sequence is shown above the MSA. EhV orthologs fell into two apparent phylogenetic groups: V5LSH6 (EhV156), G8DFX0 (EhV-202), V5LNU4 (EhV-18) and Q4A223 (EhV-86), V5LU59 (EhV-164), V5LPG3 (EhV-145), G3GNZ9 (EhV-84), G4YAS2 (EhV-88), G4YBA5 (EhV-207), G3GQ88 (EhV-203), G4YD40 (EhV-208), G9E4E4 (EhV-201), D2TEV6 (EhV-99B1). These two groups showed around 71% amino acid similarity to one another, while yeast showed around 35.7% similarity with group 1. EhV orthologs were first aligned against each other by alignment order, then aligned against yeast RPB8 with fixed input order.

similarity [365]. Although TBP has been previously identified in specific virus clades, it has not been designated an NCVOG. Approximately 50% of mimivirus genes contain a conserved, upstream AAAATTGA motif, which may be structurally comparable to the TATA box promoter element [378]. In a cursory analysis, we found similar sequences immediately upstream of the genes of some NCLDV (e.g., megavirus and ascovirus) but not others (e.g., chlorella virus; data not shown).

Novel zinc ribbon protein superfamily

We also noted the presence among NCLDV of structural homologs to the zinc finger region of eukaryotic RNA polymerase subunit RPB12 (Fig. 3.6a). RPB12 is required for RNA polymerase open complex formation [342]. RPB12 orthologs have not been identified previously in viruses, and the family shown in Fig. 3.6a may or may not represent bona fide RPB12. The N-terminal region of eukaryotic RPB12 encompassing the zinc finger region is known to form part of a larger zinc beta ribbon superfamily that includes eukaryotic transcription factors TFIIS and TFIIB and some ribosomal and other proteins [354, 379]. Figure 6a may represent a broader zinc ribbon superfamily for the following reasons: (i) for 26 of the 58 NCLDV proteins shown, the top structural homolog comprised a non-RPB12 C4-type zinc finger containing protein (Fig. 3.6a; marked 1%, 2%, and 5%), although eukaryotic RPB12 was also a structural homolog within the 80% probability threshold; (ii) all proteins of Fig. 3.6a lack the conserved C-terminal region characteristic of eukaryotic RPB12 [379]; (iii) overall sequence conservation among the 58 NCLDV proteins was nonexistent (data not shown); (iv) unlike RPB12, vaccinia protein A19, present within the family (UniProt accession number P68714; Fig. 3.6a), is not a core vaccinia RNA polymerase subunit (although it associates with transcriptional components and is required for vaccinia early gene transcription [380]); (v) whereas RNA pol subunits are typically present in NCLDV proteomes in a single copy, some NCLDV proteomes were found to contain multiple zinc ribbon protein family members (Table 3.4); and (vi) not all NCLDV query proteins matching a C4 zinc finger protein showed eukaryotic RPB12 as a structural homolog (data not shown). Nonetheless, a bona fide RPB12 subfamily seems to exist within the zinc ribbon superfamily of Fig. 3.6a. Supporting this, eukaryotic RPB12 was a specific structural homolog of even NCLDV sequences that entirely lacked a consensus zinc finger: In eight of the RPB12 structural homology regions (Fig. 3.6a), either the first or second CxxC of the finger motif contained a nonconsensus number of "x" residues or was missing a cysteine entirely. In the most dramatic example, the ranavirus RPB12 homolog (UniProt accession number Q6GZX4; Fig. 3.6a) contained no CxxC at all.

Some of the RPB12 structural homologs of Fig. 3.6a contain additional domains or motifs, such as an N-terminal SH3 motif or C-terminal very short patch repair (VSR) endonuclease domain (Fig. 3.6b). In the majority of these proteins, multiple repeating RPB12 structural homology domains were separated by regions with no detectable structural homology. VSR endonucleases have not been previously observed with zinc finger motifs, although some group II HNH endonucleases, which share a similar catalytic core with VSR, contain a C4-type zinc finger domain upstream of the C-terminal HNH endonuclease domain [381]. The RPB12-like repeats observed here may be involved in DNA binding. Overall, it seems likely that RPB12 is a subset of a larger protein superfamily.

Endonucleases

DNA endonucleases fall into several major structural families and superfamilies (Fig. 3.7a) [381-384]. The broadest of these is perhaps "PD-(D/E)xK", which is defined on the basis of a conserved PD-(D/E)xK motif essential for catalysis. It encompasses, functionally, the type I

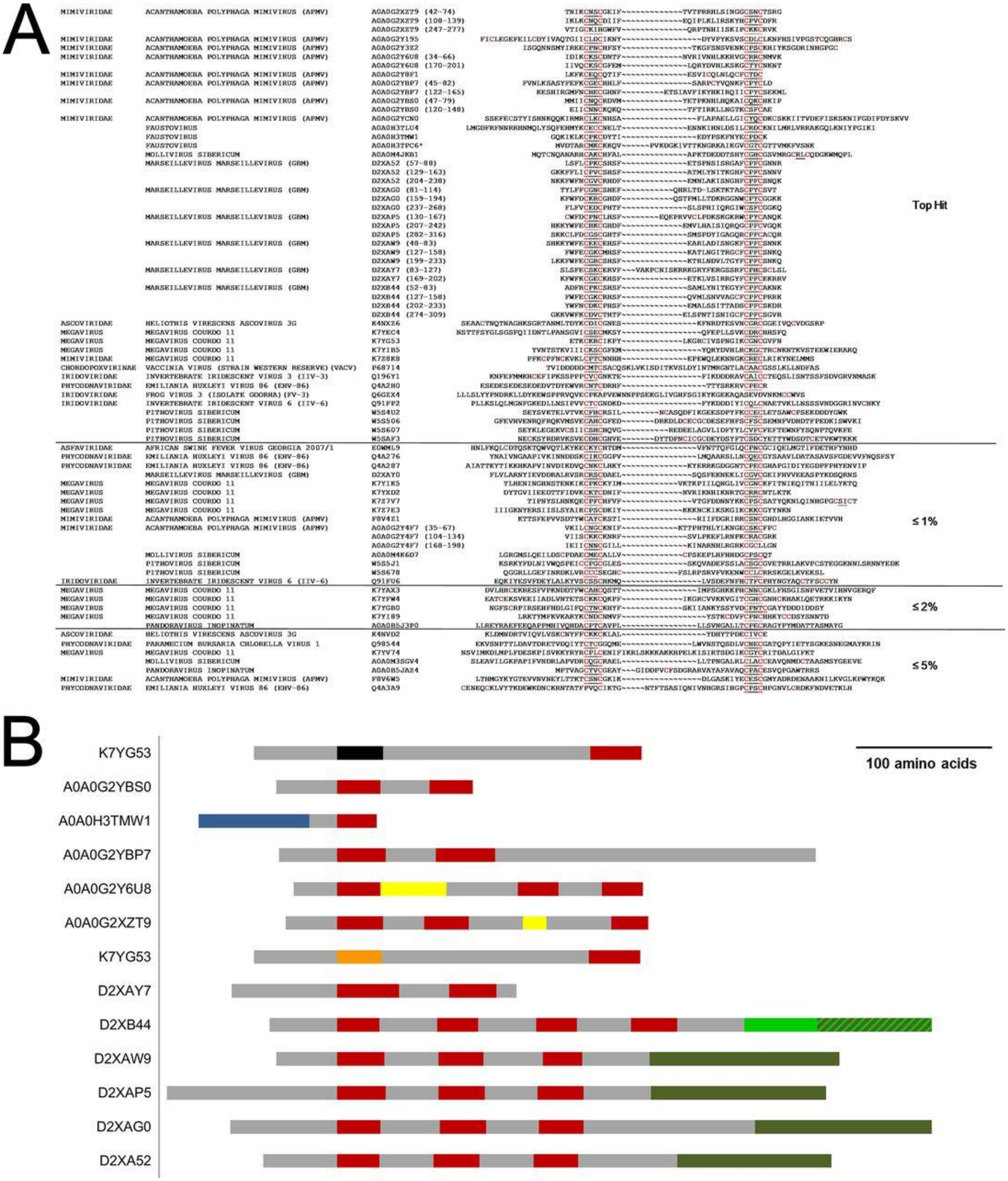
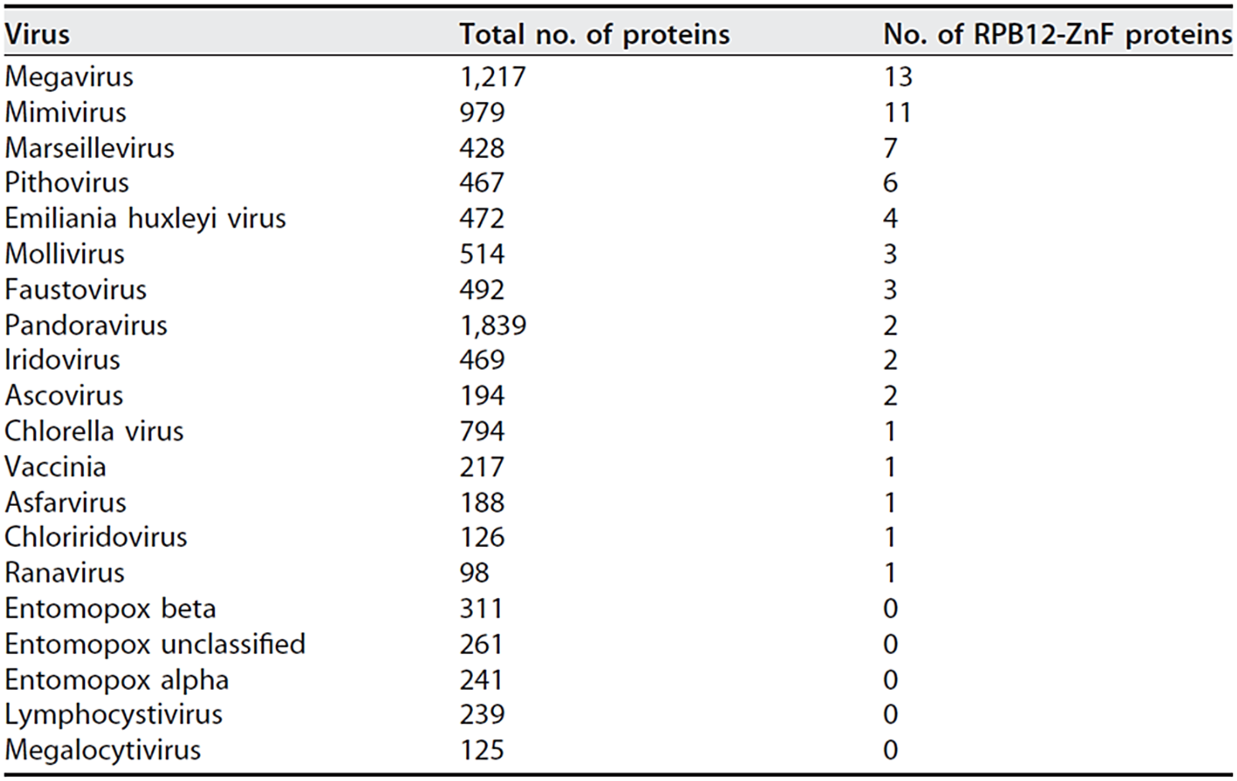


Figure 3.6. Structural homology between NCLDV proteins and eukaryotic RPB12 with scores exceeding the 80% probability threshold. (A) Sequence alignment. The great majority of NCLDV structural homologs possessed a CxxC….CxxC zinc finger (cysteines are highlighted in red). In total, 77 structural homology regions are shown, within a total of 58 proteins. The majority of these proteins were annotated as transcription factors, restriction endonucleases, zinc ribbon-containing proteins, or NAD-glutamate dehydrogenase, or were unclassified. For proteins

containing multiple matching regions, the individual regions are indicated by sequence position in parentheses after the accession number. In only one instance (red star) does the homology region shown comprise the entire NCLDV protein. “Top hit” indicates that eukaryotic RPB12 was the highest scoring structural match to the homology regions in this section. ≤1%, ≤2%, and ≤5% indicate that eukaryotic RPB12 was within 1, 2, or 5 percentage probability points, respectively, of the highest scoring structural homolog (and was not necessarily the number two homolog). This applied to 26 of the 58 proteins shown here. (B) RPB12 orthologs from panel a that contain additional regions of structural homology. Red, RPB12 (PF03604) homology regions; black, SH3 (PF00018) homology; blue, HypA (PF01155) homology; orange, Nab2 (PF11517) homology; dark green, VSR endonuclease (PF03852) homology; light green, type IV restriction endonuclease homology; striped light/dark green, overlapping type IV restriction endonuclease and VSR endonuclease homology; yellow, structural homology with “uncharacterized protein PF0385” (Q8U3S0_PYRFU); gray, uncharacterized region.

to IV restriction endonucleases (REases), which cleave both DNA strands within a specific recognition sequence [385, 386], MMR (mismatch repair)-type and VSR-type nicking and other endonucleases [387, 388] (Fig. 3.7a), with superfamily members existing as either monomers, homodimers, or homotetramers. Of the four types of REase (Fig. 3.7a), type II REases are the best characterized and the most prevalent in the biosphere, with more than 3,500 known members [383, 385, 386]. Although found predominantly in bacteria, they are also encoded by chlorella virus [389], with, for example, two reported in PBCV-1 (R.CviAI [390] and R.CviAII

Table 3.4. Number of RPB12-ZnF proteins versus total number of proteins. Numbers of RPB12-type zinc finger-containing (RPB12-ZnF) proteins per NCLDV proteome by counting proteins shown in Fig. 3.6a (RPB12-ZnF) versus the total number of proteins per proteome (all proteins). Nearly half of the 58 proteins were from mimivirus or megavirus, while several virus taxa (*Entomopoxvirinae*, megavirus, lymphocystivirus, megalocytivirus) had none at all. As a function of proteome size, representation in chlorella virus and pandoravirus was low.



Virus	Total no. of proteins	No. of RPB12-ZnF proteins
Megavirus	1,217	13
Mimivirus	979	11
Marseillevirus	428	7
Pithovirus	467	6
Emiliania huxleyi virus	472	4
Mollivirus	514	3
Faustovirus	492	3
Pandoravirus	1,839	2
Iridovirus	469	2
Ascovirus	194	2
Chlorella virus	794	1
Vaccinia	217	1
Asfarvirus	188	1
Chloriridovirus	126	1
Ranavirus	98	1
Entomopox beta	311	0
Entomopox unclassified	261	0
Entomopox alpha	241	0
Lymphocystivirus	239	0
Megalocytivirus	125	0

[391]), an additional two in chlorella virus NY-2A [392, 393] and one in chlorella virus IL-3A [394]. A large number of additional chlorella virus type II REases can be found in the REBASE database (http://rebase.neb.com/rebase/rebase.html) (43 enzymes from 34 distinct *Chlorella virus* species), most of which remain unpublished. The biological roles of the chlorella virus enzymes remain unverified [389]. Sequence is not very effective as a homology tool for the prediction of endonucleases, particularly those of the PD-(D/E)xK superfamily [395] or the homing endonuclease classes shown in Fig. 3.7a. However, conservation of secondary structure [396, 397] within, for example, the PD-(D/E)xK-containing catalytic domain [385, 395, 397] facilitated our structure-based deep mining approach. Here, a total of 96 new endonucleases were predicted over a number of structural and functional classes (Fig. 3.7a). These supplemented 36 proteins among our 20 NCLDV whose UniProt annotations already included the strings "endonuclease," "restriction endonuclease," or "nuclease" (these 36 included two proteins mentioned in the literature as VSR-type nucleases but missed by UniProt [398]). An updated overall total count of 132 endonuclease/nucleases was therefore yielded by these numbers (Fig. 3.7; see also Appendix1.Table1). For 10 of our 20 NCLDV, deep mining provided the first-time prediction of any endonuclease (Fig. 3.7b), and for many of the remaining viruses prior endonuclease genes numbered just one or two. Totals were highly variable from virus to virus, even among comparable NCLDV (e.g., among the amoebal giant viruses pithovirus, pandoravirus, mollivirus, megavirus, and mimivirus) or between the three entomopoxviruses (Fig. 3.7b), suggesting that their roles may not be central to virus replication. Endonuclease classes predicted in the NCLDV for the first time included type IV/5hmC REases, which were the primary structural homologs of 10 newly predicted endonucleases from seven NCLDV (Fig. 3.7). All of these were annotated by UniProt as uncharacterized, ALI motif, leucinerich repeat, or

N1R/p28-like proteins on the basis of distinct (non-nuclease) domains. Six of them were from one virus alone (entomopox unclassified; Fig. 3.7b)—the largest number within a single NCLDV genome. The type IV/5hmC class of REases recognize modified (typically methylated) DNA, suggesting a specific need to restrict methylated DNA among the NCLDV. In another example, UniProt showed no prior occurrence of an NCLDV VSR endonuclease (the two noted by Aravind et al. [398], above, were reassigned here). In contrast, several major classes of endonuclease remained entirely unrepresented among the NCLDV even after deep mining (Fig. 3.7a). These included the type I and III REases and the LAGLIDADG and His-Cys homing endonucleases (which do not appear to be underrepresented in the PDB), suggesting that modification-coupled DNA restriction and homing are profoundly redundant functions for the NCLDV. Indeed, the finding of NCLDV enzymes that restrict methylated DNA (the type IV/5hmC REases, above) would suggest a reason why methylation may be irrelevant as a self-protection mechanism.

In addition to the newly predicted endonucleases, many of the 36 previously reported nucleases (above) were assigned to a specific class or reassigned based on their primary structural homolog (Appendix1.Table1). For example, nine NCLDV proteins annotated by UniProt as either “restriction endonuclease” (n = 6), “group 1 intron putative endonuclease” (n = 1), “putative nuclease” (n = 1), or “helicase nuclease” (n = 1) were assigned to the VSR subset of PD-(D/E)xK. In another example, protein 069L from IIV-6 (Table 3.1) and protein MSV196 from MSV (Table 3.1), were reassigned from VSR-type endonucleases [398] to type IV/5hmC endonuclease [384, 399, 400], since VSR appeared as only the 5th-ranked structural match for each of the two proteins, with 5hmC versus VSR probabilities of 96.9%/94.4% and 97%/94.8%, respectively. Their UniProt annotations showed them as Bro-N domain (PF02498)-containing

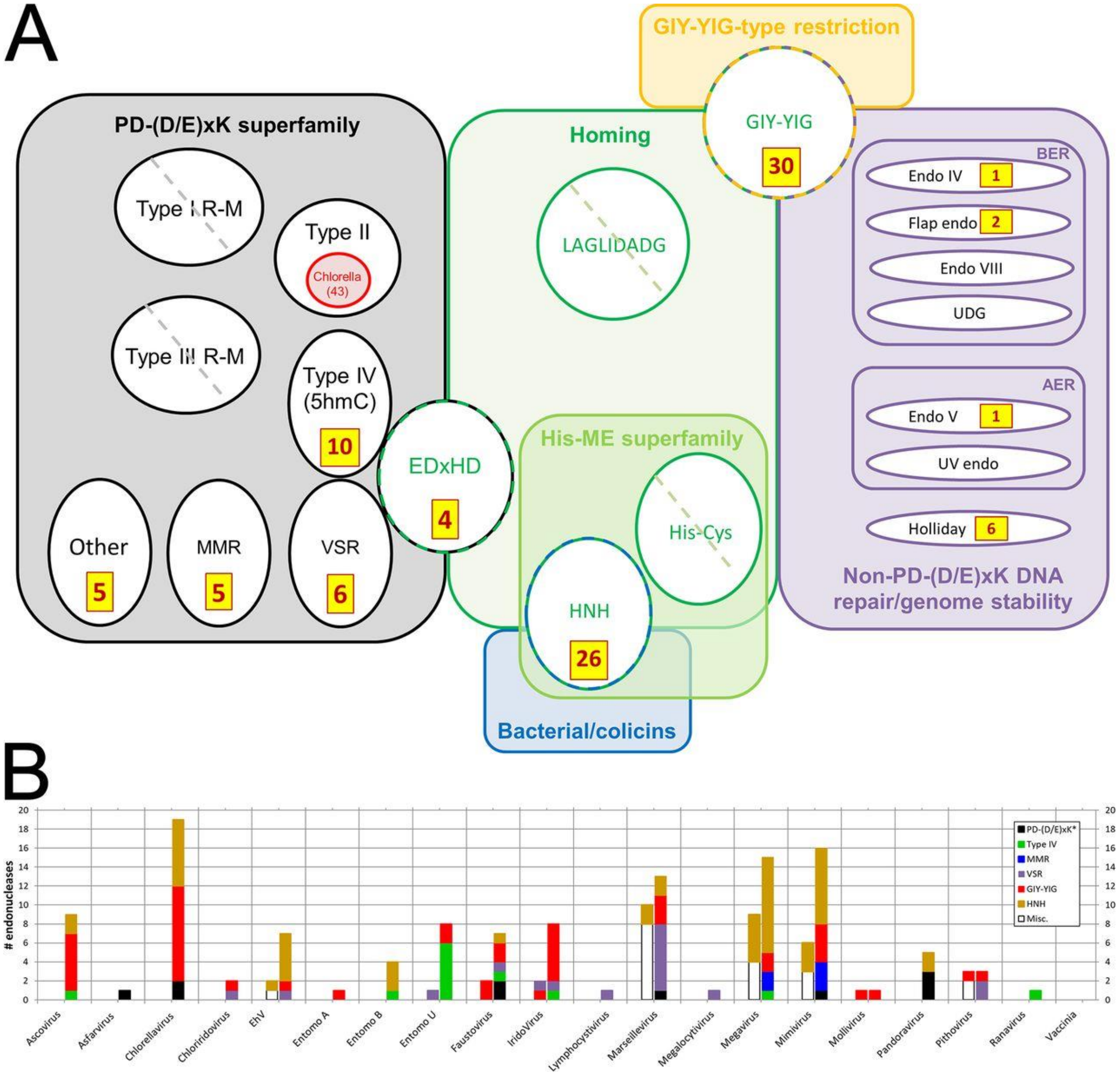


Figure 3.7. Structure-based deep mining markedly elevates numbers of endonucleases identified across the NCLDV. All accession numbers are given in Appendix1.Table1 in the supplemental material. (A) Counts of newly identified NCLDV endonucleases (brown/yellow boxes) mapped onto known DNA endonuclease structural/functional classes (colored boxes/circles/ovals). Circles/ovals with no counts shown indicate major classes with no representatives reported among any NCLDV (diagonal hatch) or none newly identified here (no hatch). PD-(D/E)xK (black), structural superfamily showing the following functional classes: REases (types I to IV), nicking endonucleases for DNA mismatch repair (MMR), and very short patch repair (VSR), and one class of homing endonucleases (EDxHD). The VSR, type IV, and EDxHD groups are shown touching to illustrate their particularly close structural relationship in search results (see the text). Types I and III REase polypeptides are denoted "R-M" due to their

dual function as restriction-modifying enzymes, in which we focused only on the “catalytic site for DNA cleavage” [401] and “endonuclease domain” [402], respectively. Red indicates strain depth dimensionality for chlorella virus type II REases specifically (from REBASE rather than UniProt; see text). Other major colors indicate functional classes encompassing either structural families (green, blue, and orange) or functional classes (purple). In the latter, BER and AER represent base and alternative excision repair, respectively; UDG, uracil DNA glycosylase (an endonuclease); Endo IV, AP-endonuclease. Counts do not include our assignments/reassignments of previously identified/annotated endonucleases (see the text). (b) NCLDV endonuclease counts by virus. Entomopox A, B, and U refer to entomopox alpha, beta, and unclassified, respectively. Bars to the left and right of the central tick represent counts before and after deep mining, respectively. “Before” counts include proteins that matched an endonuclease here and were also “endonuclease” or “nuclease” according to UniProt gene_name. “After” counts represent “before” counts plus endonucleases newly identified here plus reassignments (see the text). Each bar is divided by color according to endonuclease class (see color legend). The PD-(D/E)xK* class refers to PD-(D/E)xK homing plus “Other”” (panel a). The “Misc.” class (“before”) contains, exclusively, members reassigned to other classes in the “after” sections based on primary structural homolog (see the text). Excluded from the graph are all chlorella virus “red ring” (panel a) restriction endonucleases not from PBCV-1 (Table 3.1). For simplicity, the small numbers of repair endonucleases (purple section of panel a) are omitted. Counts represent “top hit only” structural matches).

(069L) or “ALI motif gene family” (MSV196 - the “ALI” motif being a subset of Bro-N). These annotations were based on different domains within the two proteins. No structural homologs were found above the 80% probability threshold for either of two known chlorella virus restriction endonucleases, R.CviAI [390] and R.CviAII [391] (discussed above). Apparently, they did not align well with any structures in the PDB database, in which type II enzymes were well represented. While the HHsearch structural homology tool can find large numbers of authentic PDDExK enzymes, it can fail with PDDExK protein subfamilies with few members or more distant structural homology [395]. Perhaps functional type II REases cover a wider fold space than PD-(D/E)xK alone.

Despite MMR endonucleases (EndoMS and NucS-type) being quite uncommon in the biosphere overall [403, 404], five such endonucleases were predicted here, all from megavirus and mimivirus, all of which were previously annotated as “uncharacterized” or “KilA-N domain-containing” proteins. EndoMS and NucS typically have an N-terminal DNA binding/dimerization domain and C-terminal catalytic domain [405]. While the C-terminal regions of the mimivirus/megavirus homologs matched the catalytic domain, the N-terminal regions comprised an APSES or KilA-N type domain. Both APSES and KilA-N are DNA binding domains commonly found in eukaryotic viruses and in cellular LAGLIDADG endonucleases [373].

GIY-YIG family endonucleases (Fig. 3.7a) are typically encoded by phage and fungi [381], in which their most common function is homing/self-propagation of group 1 homing introns [406, 407]. They are typically small proteins (approximately 100 amino acids) with short “GIY” and “YIG” motifs in the N-terminal region, along with extended recognition sites for their DNA targets. Ten GIY-YIGs had been previously identified in seven NCLDV. Here, we predicted an additional 30 from an additional six viruses (Fig. 3.7), the largest increases being in ascovirus, chlorella virus and iridovirus. Additional domains were found fused to the N- and/or C termini of some of these proteins, such as a NUMOD3, Tc5 transposase DNA-binding domain, CENP-B N-terminal DNA-binding domain, KilA-N, and HIT zinc finger domains (data not shown).

Members of another endonuclease family, HNH, are found as homing enzymes within group 1 and group 2 introns, and also as bacterial restriction endonucleases (e.g., PacI), colicins [408] and/or DNA/RNA nonspecific endonucleases [381, 382]. PacI, a “rare-cutting” REase, cleaves duplex DNA within the sequence 5=-TTAAT^TAA-3= [381, 409]. Group I homing endonucleases such as I-HmuI also have a highly conserved target site, but unlike PacI, they cleave only one DNA strand. DNA/RNA nonspecific endonucleases in the HNH family are extracellular [410] and function in bacterial self-defense against neutrophil extracellular traps [411], among other functions. Here, HNH endonucleases were the primary structural homologs of 26 NCLDV proteins, almost tripling the total known among our 20 NCLDV—the largest increases being observed in chlorella virus and mimivirus (Fig. 3.7b). Some of the newly predicted HNH endonucleases (Appendix1.Table1) showed internal repeats of the I-HmuI or PacI homology regions.

Two faustovirus and two chlorella virus proteins, annotated in UniProt as “uncharacterized,” showed the homing endonuclease I-bth0305I as a top structural hit. I-bth0305I is annotated in UniProt as a “mobile intron protein” of a lineage that has been termed the “EDxHD family” (Fig. 3.7a). Some endonuclease classes, such as the VSR, type IV/5hmC, and EDxHD, which cover very distinct functional roles, were found to be particularly closely related in three-dimensional structure, with members of these classes interleaved in search results for a specific NCLDV query protein. Other NCLDV queries showed only a single structural homolog within the PD-(D/E)xK superfamily. This was probably not due to a paucity of closely related structural choices within the database, since more than 73 of the 142 Pfams within the PD-(D/E)xK superfamily have yielded crystal structures. Instead, individual structural homologs seem to have been selected against quite a fine-grained structural landscape. In yet other cases, cellular REases with a highly conserved type IV/5hmC fold have been found that lack nearly all of the commonly conserved residues [399]. For all of the above reasons, we hesitate to assign functional roles to specific NCLDV endonucleases on the basis of structural homology alone.

Repeat domain proteins

Numerous ankyrin repeat motif-containing proteins were identified in the genomes of the NCLDV, although not all NCLDV were found to encode them. Members of this protein family have recently been shown to target host defense proteins for degradation [412]. Amoeba-infecting NCLDV show a correlation between genome size and the number of encoded proteins containing ankyrin, MORN, and WD40 repeat domains [413], with repeat-containing proteins in megavirus, mimivirus, and pandoravirus comprising a substantial portion of their total proteomes [413, 414]. Here, structure-based deep mining led to the identification of large numbers of additional repeat domain-containing proteins across the NCLDV (Fig. 3.8a), mostly identified with very high probability. The most substantial increases were found among the ankyrin- and MORN-repeat-containing protein families (Fig. 3.8a), with as many as 234 ankyrin repeat-containing proteins found in pandoravirus. Pandoravirus, pithovirus, mimivirus, megavirus, and asfarvirus showed markedly higher proportions of their proteomes devoted to ankyrin repeat proteins than the other viruses (ranging from 11.9% to 14.4% overall; Fig. 3.8b). Although this group includes four amoeba-infecting viruses, three additional amoeba- infecting viruses (faustovirus, marseillevirus, and mollivirus) showed substantially lower numbers (Fig. 3.8b), with only five ankyrin repeat-containing proteins found in faustovirus (Fig. 3.8b) [413]. Perhaps most surprising was the finding of 27 ankyrin repeat-containing proteins in asfarvirus (the highest proportion of all the proteomes; Fig. 3.8b) since UniProt contained no annotated ankyrin repeat proteins at all for this virus family (although three such proteins identified by Pfam were not auto transferred to UniProt). Structural matches comprised a diversity of ankyrin repeat containing proteins in PDB and, as in the PDB structures, ankyrin repeats in the NCLDV queries are scattered throughout the protein.

A total of 147 MORN (membrane occupation and recognition nexus) repeat-containing proteins were also discovered, almost entirely in the amoeba-infecting faustovirus, marseillevirus, and pandoravirus (Fig. 3.8a). Many of the NCLDV query proteins were annotated in UniProt as "unclassified." In contrast to the ankyrin repeat-containing queries (above), the 147 NCLDV queries matched only one MORN-repeat containing protein in PDB, namely, a histone methyltransferase. To the best of our knowledge, only a few MORN repeat-containing proteins have ever been identified in any organism. These include the junctophilins, a group of mammalian proteins found within membrane junctional complexes [415, 416], which serve to bridge membrane pairs (such as the plasma membrane and the membrane of the endoplasmic reticulum or sarcoplasmic reticulum). Within the bridge, the junctophilin's N-terminal MORN motif, comprising eight repeats of a 14 amino-acid sequence [415], interacts with the phosphoinositides of one cell membrane (see Jiang et al. [416] and references therein), while a hydrophobic C-terminal transmembrane region anchors to the other. The NCLDV MORN repeat proteins do not appear to be acting as junctophilins. Of 40 NCLDV MORN repeat proteins examined at random, only one had a transmembrane region, and it was located at the protein N terminus, rather than the C terminus (data not shown). Moreover, the MORN repeat region

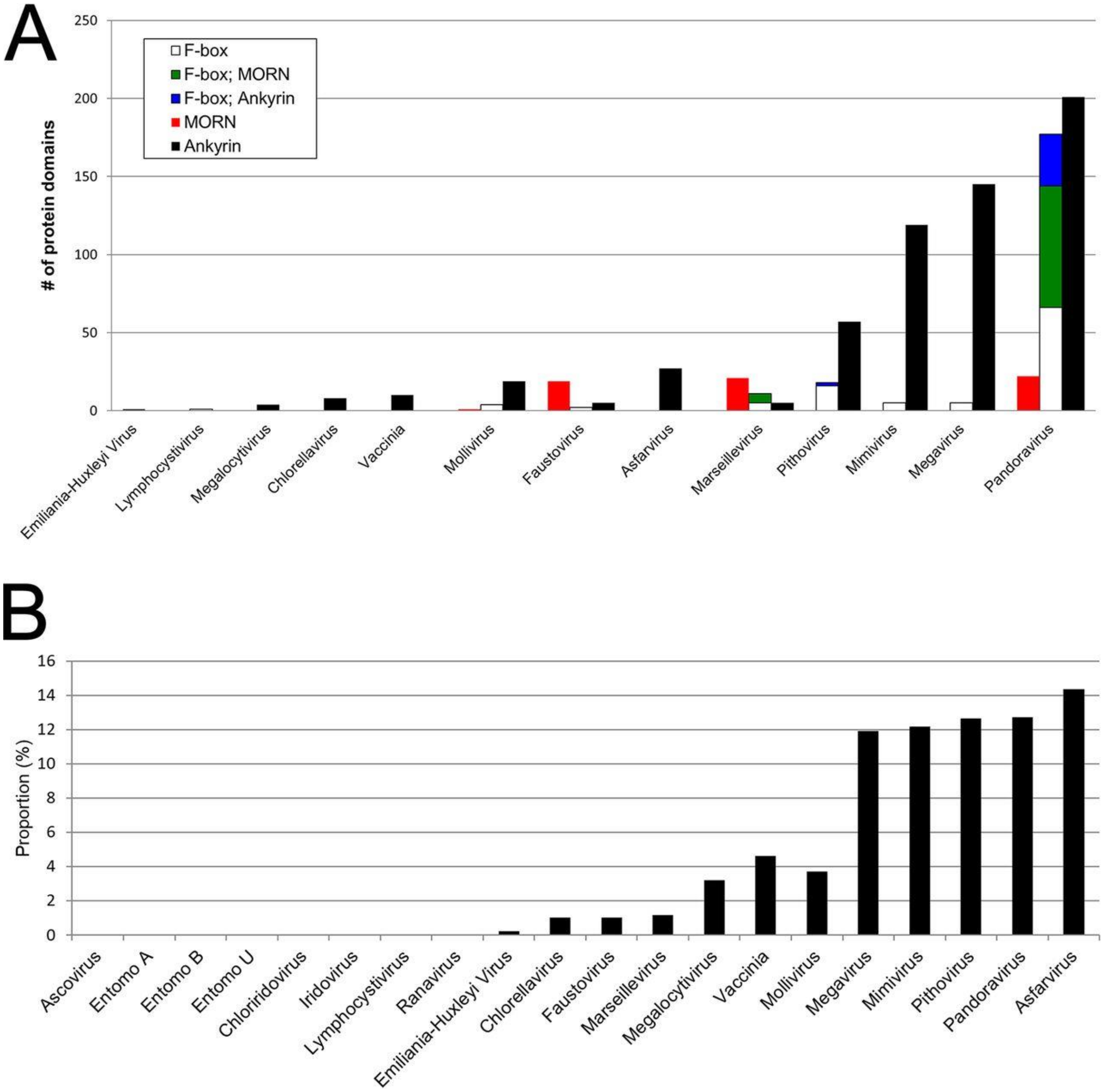


Figure 3.8. Repeat domain proteins. (A) Counts of ankyrin repeat (CL0465), MORN repeat (CL0251), and F-box motif (CL0271) containing proteins per virus for 13 of the 20 representative NCLDV. The remaining seven NCLDV contained none. Blue and green, proteins containing an F-box plus repeat domains. Viruses are ordered (left to right) by overall numbers of such proteins per proteome. (B) Ankyrin repeat-containing protein counts as a proportion of total genes in the NCLDV proteome. Eight of our 20 NCLDV (left) lacked any such proteins. NCLDV are labeled as in panel A and Fig. 3.7B.

tended to fall within the C-terminal halves of most NCLDV proteins. A second group of MORN repeat-containing proteins has been found in unicellular parasites such as *Toxoplasma gondii* and *Toxoplasma brucei* [417, 418]. This group appears to interface membranes with cytoskeletal components. In the NCLDV, many of the MORN repeat-containing proteins also showed structural homology to a TCP10_C family domain (data not shown), which is centriole related. The centriole has two roles, namely, as part of the eukaryotic centrosome (a microtubule organizing center during mitosis) and as the basal body from which cilia and flagella emanate [419, 420]. The NCLDV MORN repeat proteins with a TCP10_C domain may tether viral membranes to cytoskeletal structures such as those found in ciliated or flagellated amoebae.

In pandoravirus, pithovirus, and marseillevirus, many of the proteins possessing C-terminal ankyrin or MORN repeat motifs also contain an N-terminal F-box (Fig. 3.8a). In pandoravirus, 78 of the 100 identified MORN repeat-containing proteins and 33 of the 234 ankyrin repeat-containing proteins showed this arrangement. This orientation is novel for the NCLDV. The F-box domain of poxvirus ankyrin repeat-containing proteins, for example, is located at the protein C terminus [412, 421]. Interestingly, the F-box domains of the poxvirus proteins scored well below our 80% probability threshold for structural homology (data not shown). Instances of N-terminal F-box with C-terminal ankyrin domain proteins have previously been described in *Legionella pneumophila* [422], but sequence alignments of NCLDV proteins with these proteins showed no obvious sequence homology (data not shown).

Structural homologs shared narrowly among NCLDV

In addition to protein families shared broadly among the NCLDV (above), some structural homologs were shared more narrowly (Table 3.5). Structural homologs were from a variety of organisms, which may simply reflect proteins amenable to structural biology or those having some specific interest rather than being a particularly relevant organism for the NCLDV. Nonetheless, the broad representation of microbes among structural homologs (Table 3.5) suggested the possibility of horizontal gene transfer during microbial processes such as phagocytosis. For the most part, structural homology was at the domain or fold level only, so the corresponding protein annotations tended to be structurally oriented (e.g., CHAP domain, winged helix-turn-helix) and therefore unsatisfying in deducing the overall function of the NCLDV protein. In one case, however, the probable function was clear, namely, for NCLDV homologs to herpesvirus glycoprotein B.

Herpesvirus glycoprotein B

NCLDV structural homologs of herpesvirus glycoprotein B (gB) were detected, although the *Herpesviridae* are not considered members of the NCLDV due to their exclusively nuclear sites of replication [423]. gB is an essential, trimeric herpesvirus surface glycoprotein - the most highly conserved member glycoprotein - the most highly conserved member of the 5-protein herpesvirus host cell fusion and virus entry complex [424]. It features an N-terminal signal

Table 3.5. New/expanded trans-NCLDV protein families. For each row, structural homology combined with annotation associated with the structural homolog increased the number of NCLDV covered (among our 20 NCLDV) to two or more. Columns 1, 2, 9: UniProt accessions. Column 2 parentheses: PDB entries. Column 4: UniProt "Protein" field. Columns 5, 6: Pfam(s) covering, or overlapping with, the structural homology region (from the structural homolog's RCSB entry). Column 7: Pfam "Species" field for Pfams with no InterPro link. Otherwise, linked InterPro "Taxonomy" field. Column 9: Any accession from a distinct NCLDV that was previously annotated as in column 6. Column 10: For a homology region covering the entire structural homolog, this is from UniProt's "Function" field. Otherwise, it is from the annotation for the Pfam covering the homology region. Row 5: Pfam clan CL0015 ("Major Facilitator Superfamily") covers PF07690 and PF00854, the Pfam hits to mollivirus (previously) and pandoravirus (here), respectively.

NCLDV query accession(s)	Structural homolog accession(s)	Structural homolog organism	Structural homolog name	Pfam(s) overlapping the homology region	Pfam descriptor(s)	Family phyletic distribution	Match probability (%)	NCLDV: Previously annotated in	Function of structural homolog/homology region
D2XAQ6 (Marseillevirus) O55739 (Iridovirus)	Q9P0M2 (5jj2_A)	Human	A-kinase anchor protein 7 isoform gamma	PF10469	AKAP7 2'5' RNA ligase-like domain	Eukaryotes	97.3 98.1		AKAP7 targets cAMP-dependent protein kinase A to cell membrane/cytoskeleton.
Q84547 (Chlorella virus)	A0A0B8QKV5 (1zuj_A) P0CB53 (2cf7_A)	Lactococcus Streptococcus	DNA protection during starvation protein	PF00210	Ferritin-like domain	Archaea, bacteria, eukaryotes, Pithovirus sibericum, caudovirales	96.7 96.2	W5S6G8 (Pithovirus)	D.P.D.S. protein protects DNA from oxidative damage by sequestering intracellular Fe^{2+} and storing it as Fe^{3+}.
K7Z8N4 (Megavirus) A0A0G2Y127 (Mimivirus)	Q4WZ11 (3w0e_A)	Aspergillus	Elastase inhibitor AFUEI	PF11720	Peptidase inhibitor I78 family	Bacteria, eukaryotes	96.2 95.1		Reduces pathogenicity of *Aspergillus fumigatus*.
K7YW37 (Megavirus) A0A0G2Y3W1 (Mimivirus)	W2SRJ3 (4uet_A)	Nematode	Fatty acid retinoid binding protein	PF05823	Nematode fatty acid retinoid binding protein (Gp-FAR-1)	Bacteria, eukaryotes	85.8 86.9		Binds retinol and fatty acids, assisting in the evasion of plant defense compounds.
A0A0B5JB34, A0A0B5JCF5 (Pandoravirus)	A0A0M3KKZ1 (4w6v_A)	Yersinia	di-/tripeptide transporter	CL0015	Major Facilitator Superfamily	Archaea, bacteria, eukaryotes, unclassified Mimiviridae, Caudovirales, Nudiviridae, Phycodnaviridae, Mollivirus sibericum	97.4, 90.6	A0A0M4JSZ5 (Mollivirus)	One of the two largest families of membrane transporter proteins in the biosphere. Moves small solutes. PF00854 (Pandoravirus) covers H⁺-dependent oligopeptide transporters.
K7Z7J6, K7Z8Q9, K7YFQ8 (Megavirus) A0A0G2Y657, A0A0G2Y9M2, F8V6J0 (Mimivirus)	Q8MQJ9 (4zlr_B)	Drosophila	Brain tumor protein	PF01436	NHL repeat	Archaea, bacteria, eukaryotes, unclassified Mimivirus Satyrvirus	99.4, 99.7, 99.4 99.6, 99.5, 99.5		Translational inhibitor: 6-bladed beta propeller repeat found in a wide variety of proteins.
K7YFR4 (Megavirus) A0A0G2YBR1 (Mimivirus)	Q6Y7T6 (4csh_C)	Staphylococcus	Phage K_071	PF05257	CHAP domain	Archaea, bacteria, eukaryotes, Caudovirales, bacteriophages	96.2 95.7		CHAP domain is found in enzymes with invariant active site Cys and His, eg. Peptidoglycan hydrolases.
A0A0M5KAF0 (Mollivirus) A0A0B5JDI3 (Pandoravirus)	Q9S508 (3ff0_A) Q9GSQ9 (1n1i_A)	Pseudomonas Plasmodium	Phenazine biosynthesis protein B2 Merozoite surface protein 1	PF03284 PF12946 PF12947	Phenazine biosynthesis protein A/B MSP1 EGF domain 1 EGF domain	Archaea, bacteria, eukaryotes	93.0(3ff0_A), 94.4(1n1i_A) 95.7		Synthesis of phenazine, an antifungal and antibacterial antibiotic.
Q98542,Q98543 (Chlorella virus)	P39825 (3d9y_A) Q95VF7 (1acf_A)	Yeast Acanthamoeba	Profilin	PF00235	Profilin	Archaea, bacteria, eukaryotes, poxviridae	80.9 81.7	Q76ZN5 (Vaccinia)	Modulates actin dynamics and intracellular transport of proteins.
Q4A2I6 (Emiliania-Huxleyi virus) D2XAY5 (Marseillevirus)	Q0SDB1 (4u5r_A) P70994 (2opa_A)	Rhodococcus Bacillus	Tautomerase_3 domain-containing protein 2-hydroxymuconate tautomerase	PF01361	4-Oxalocrotonate Tautomerase	Archaea, bacteria, eukaryotes	85.0 90.0		4-Oxalocrotonate tautomerase converts 2-hydroxymuconate to the alpha-beta-unsaturated ketone, 2-oxo-3-hexenedioate.
K7YAA9, K7YUX8, K7YXF3, K7Z9E5 (Megavirus) A0A0G2YCB3, A0A0G2YCB5, A0A0G2Y4L1 (Mimivirus)	E7FCY1 (4BXR_A) Q9VI72 (4MPZ_A)	Zebrafish Drosophila	Centromere protein J Spindle assembly abnormal 4	PF07202	T-complex protein 10 C-terminus	Bacteria, eukaryotes	97.3, 87.4, 96.4, 93.7 97.5, 97.6, 97.5		C-terminal domain of T-complex protein 10, a protein of unknown function.
K7Z767 (Megavirus) E3VYL3 (Mimivirus)	P0A8H8 (1LV3_A)	E. coli	DNA gyrase inhibitor YacG	PF03884	DNA gyrase inhibitor YacG	Euryarchaeota archaeon, bacteria, eukaryotes	83.4 87.9		DNA gyrase inhibitor in bacteria.
W6JPK9, W6JIZ4 (Entomopox alpha) R4ZDQ0, R4ZES4 (Entomopox beta) Q9YVZ3, Q9YW15 (Entomopox unclass.)	P06437 (2gum_A, 5fz2_A	Herpesvirus	Envelope glycoprotein B	PF17416 PF17417 PF00606	Herpesvirus Glycoprotein B Herpesvirus Glycoprotein B PH-like domain Herpesvirus Glycoprotein B ectodomain	Herpesviridae, nematodes, arthropods	98.4, 96.2 98.8, 96.6 98.4, 96.2		Surface glycoprotein of Herpesviruses.
A0A0M5KAC8 (Mollivirus) A0A0B5J3T1 (Pandoravirus)	B6JPK4 (3ub6_A)	Helicobacter	Methyl-accepting chemotaxis transmembrane sensory protein (MCP-like protein)	PF17200	Single Cache domain 2	Archaea, bacteria, eukaryotes	96.0 96.2		Single Cache domain 2: Extracellular protein domain involved in small molecule recognition.

sequence, ectodomain (comprising at least 80% of the 904-residue protein), C-terminal transmembrane anchor, and a relatively short cytoplasmic tail (Fig. 3.9). The crystal structure for the gB ectodomain [425] shows five distinct subdomains connected by flexible linkers, and five intramolecular disulfide bonds [425] (Fig. 3.9). Subdomains III and IV are each discontinuous in the linear sequence and are stabilized by a disulfide bond [425]. A pair of structural homologs was found in each of three entomopoxviruses (Fig. 3.9). Like HSV-1 gB, the six homologs each showed a predicted N-terminal signal sequence, an apparent ectodomain, C-terminal transmembrane anchor, and a short cytoplasmic tail. One of the two protein clusters (Fig. 3.9, center) was highly structurally homologous to HSV-1 gB (98% probability, covering all gB subdomains except subdomain V; Fig. 3.9). The disulfide bonds stabilizing the discontinuous segments of subdomains III and IV were conserved [425], as was the disulfide bond within domain IV (Fig. 3.9).

The other entomopoxvirus protein cluster (Fig. 3.9, lower) showed lower overall structural homology to gB (96% probability), covering only the C-terminal segments of the two discontinuous subdomains—III and IV—and only one of cysteine from each pair of segment-bridging disulfides. The disulfide within domain IV was, however, preserved (Fig. 3.9). The subdomains that were most conserved highly between HSV-1 gB and the six entomopoxvirus proteins, namely subdomains III and IV, lie in the most exposed, membrane-distal region of the protein [425].

Unique structural homologs

Additional structural homologs were identified in individual NCLDV only (Table 3.6). As in Table 3.5, a homology region extending beyond just the single-domain level and accompanied by an explicit Pfam functional description were considered functionally predictive for the

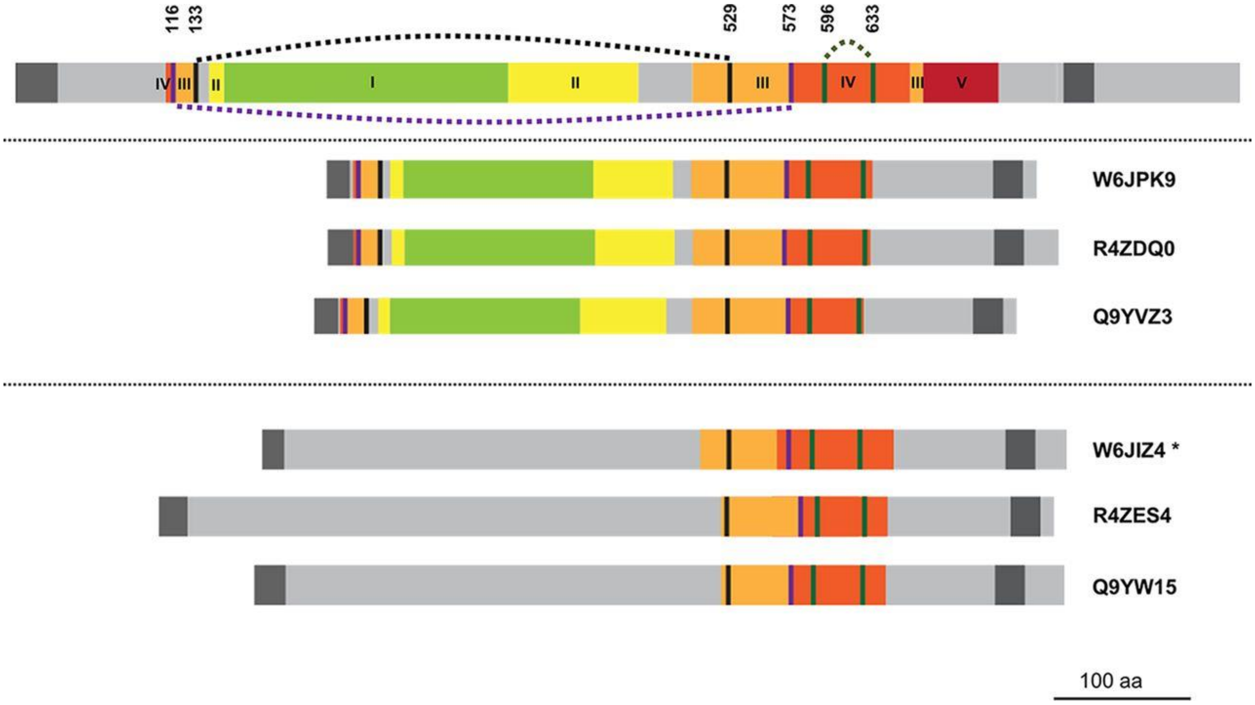


Figure 3.9. HSV-1 envelope glycoprotein B (gB) aligned with six proteins from entomopoxviruses. (Upper) HSV-1 (strain KOS) protein gB (UniProt accession P06437). (Center) Entomopox alpha, beta, and unclassified accessions W6JPK9, R4ZDQ0, and Q9YVZ3, respectively. Lower section: Entomopox alpha, beta, and unclassified accessions W6JIZ4, R4ZES4, and Q9YW15, respectively. Green, yellow, orange, red, and brown indicate subdomains of the HSV-1 gB ectodomain labeled I to V, respectively, in Heldwein et al. [425]. These subdomains were localized within entomopoxvirus proteins by visual inspection of conserved residues identified by multiple sequence alignments and on the basis of secondary structural alignment (data not shown). Intervening light gray regions were not shown in the crystal structure [425]. Dark gray, predicted N-terminal signal peptide and C-terminal transmembrane regions. Of the five disulfide bonds conserved among herpesviruses [425], three are shown (broken curved lines joining pairs of colored vertical lines: 133 to 529 [black]; 116 to 573 [purple]; and 596 to 633 [green]). The two remaining disulfides, in HSV-1 gB domains I and II (not shown), were missing from all six entomopox proteins. In the lower protein cluster, only the 596 to 633 cysteine pair is preserved. The starred accession was annotated “putative glycoprotein B” in UniProt, following the BLASTP homology noted in Table 1 of Mitsuhashi et al. [426], while the others remain “uncharacterized.”

NCLDV query protein. Conversely, where the Pfam descriptor was generic and/or the homology region was only narrowly localized, the result was considered diagnostic of a structural fold only. A number of unique structural homologs were identified (Table 3.6), some of which represented protein classes that were not, apparently, identified previously in any virus.

These included a phenolic acid decarboxylase, a prokaryotic-type ribosomal protein (to our knowledge the first to be reported in a eukaryotic virus), a gasdermin-related apparent molecular decoy, a proteasomal subunit, an HIG1 domain family member (HIG1 being induced by hypoxia), and the first report to our knowledge of cysteine knot proteins in a virus, including an apparent defensin (Table 3.6). These are discussed in greater detail below.

Aromatic acid detoxification. Protein K4NVH5 from ascovirus was structurally homologous to Phenolic acid decarboxylase (PAD, Table 3.6), a class of enzyme that decarboxylates phenolic compounds to their corresponding p-vinyl derivatives via a non-oxidative mechanism [427]. Although PADs have been identified in bacteria, amoeba, protozoa and algae, this appears to be the first report from a virus. Ascoviruses multiply within the larval tissues of Lepidoptera, to which they are eventually fatal [428]. Lepidopteran larvae are voracious herbivores, encountering an array of plant phenolic compounds generated for anti-herbivore defense. These compounds act by covalent inactivation of larval digestive enzymes and covalent reaction with larval gut tissue [429, 430]. They may be detoxified by enzymes secreted into the gut lumen by the larva or the gut microbiota [431, 432].

Ascovirus-filled vesicles accumulate in the larval gut lumen before spreading throughout the larval body [433]. However, the products of some detoxification systems of the larval host, such as quinones produced by the prophenoloxidases, remain toxic to viruses. Baculovirus infectivity, for example, is significantly lowered by the binding of quinones to viral occlusion bodies [434]. Ascovirus protein K4NVH5 may serve to detoxify phenolics in the host's diet while diverting them away from quinone-producing pathways that could remain toxic to the virus.

K4NVH5 had a second structural homolog, *Burkholdia* beta lacatamase, whose structural homology to various aromatic acid decarboxylases (PADs, Ferulic acid decarboxylases and p-Coumaric acid decarboxylase) is apparent in the RCSB database. The relationship between beta lactam ring-opening and aromatic acid decarboxylation is unclear. They both involve hydrolysis at a carbonyl bond, though for beta lactamase this is an N-C bond in a 4-membered ring, while for PAD it is a C-C bond and the carbonyl is part of a terminal carboxyl group. Their catalytic mechanisms may not be related.

Ribosomal protein: The uncharacterized emiliania huxleyi virus accession Q4A2G2 showed structural homology to *Deinococcus* 50S ribosomal protein L19 exceeding the 80% probability threshold (Table 3.6) as well as to the equivalent protein from other prokaryotes, archeae, and the *S. cerevisiae* mitoribosome (data not shown). Eukaryotic cytoplasmic ribosomes possess no homolog to this protein. The large protein sequence family encompassing L19 includes a homolog in the red algae chloroplast (Table 3.6) and, like red algae, the coccolithophore host of emiliania huxleyi virus is photosynthetic. Emiliania huxleyi virus may therefore modulate host photosynthesis via its chloroplast ribosome. A number of other ribosomal proteins have recently been found in viruses (mainly phage [435]), though to our knowledge this is the first report of a prokaryotic L19 homolog or of any prokaryotic type ribosomal protein in genome of a eukaryotic virus. Q4A2G2 is conserved in all emiliania huxleyi virus genomes sequenced to date.

Endocytosis: Ranavirus protein Q6GZV8 is structurally homologous to the 'V-shaped' domain of the human modular protein PDCD6IP/ALIX. PDCD6IP/ALIX functions in the ESCRT pathway for intralumenal endosomal vesicle formation, at the abscission stage of cytokinesis [436]. It is also involved in the abscission and budding of enveloped (lenti)viruses via hijack of the cellular ESCRT machinery (in which a short peptide motif in lentivirus GAG protein interacts with ALIX V-shaped domain [437, 438]). Tiger frog virus (TFV), a member of the *Ranavirus* genus, uses the ESCRT pathway during virus budding, recruiting ALIX and other proteins that bind to the ESCRT protein complex to mediate its release from the host cell [439]. We speculate that Q6GZV8 may be involved in this pathway.

Pandoravirus protein A0A0B5J0R1 was structurally homologous to an SHD1 domain ("SLA1 homology domain 1"). SHD1 domains in yeast protein sla1p act as adaptors during endocytocis: In clathrin-coated vesicles sla1p binds actin while its SHD1 binds cargo proteins containing an NPFX(1,2)D endocytic targeting signal. This signal is found in plasma membrane proteins destined for rapid endocytic internalization [440-443]. Instead of sorting to the lysosomes for complete degradation, however, NPFX(1,2)D-containing proteins are recycled

Table 3.6. Structural homologies found uniquely in individual NCLDV. Almost all had a prior annotation of "Uncharacterized" (column 2). Columns are as in Table 3.5 with two additional ones: "NCLDV query annotation", and "Structural homology region" (residue range in the query protein and % coverage thereof). Column 9: Probability values >99.8% are shown to two decimal places. Column 10: All functional annotation are sourced. Each row represents a distinct query protein or related query family. For rows showing a single structural homolog, this was either the only one exceeding our 80% probability threshold or the highest scoring member of a family of equivalent proteins. The four exceptions are: Rows 1 and 16 (K4NVH5 and E3VYK8): Two distinct families of structural homologs exceeded the 80% threshold, both showing very similar probability values; Row 18 (Q4A2A1): Multiple homology regions were arranged across multiple homologs with similar match probability; Row 9 Q4A223: Albeit 2f3i_A matched with marginally higher probability, 4ayb_G's homology region was much longer demonstrating that RPB8 structural homology extends across the entire length of the query protein (it is therefore included in Table 3.3).

NCLDV query accession(s)	NCLDV query annotation	Structural homolog accession(s)	Structural homolog organism	Structural homolog name	Structural homology region	Pfam(s) overlapping the homology region	Pfam descriptor(s)	Family phyletic distribution	Match probability (%)	Function of structural homolog/homology region
K4NVH5 (Ascovirus)	Uncharacterized protein	O07006 (2p8g_A) A4JJY8 (5hal_A)	Bacillus subtilis Burkholderia vietnamiensis	Phenolic acid decarboxylase PadC Uncharacterized protein]	5-106 (91%) 3-105 (92%)	PF05870 NONE	Phenolic acid decarboxylase (PAD) -	Archaea, bacteria, protozoans, algae -	99.87 99.91	UniProt: Catalyzes decarboxylation of phenolic acids to phenolic derivatives. UniProt: Uncharacterized protein. PDB: Putative beta-lactamase
Q4A2G2 (Emiliania huxleyi virus)	Uncharacterized protein	Q9RWB4 (5dm6_M)	Deinococcus radiodurans	50S ribosomal protein L19	4-56 (85%)	PF01245	Ribosomal protein L19	Archaea, bacteria, eukaryotic organelles	85.4	PDB: From structure of 50S subunit. PDB of RL19_ECOLI: L19 is located at 30S-50S subunit interface, contacting 16S rRNA, may be involved in translocation. PROSITE: Sequence family for this Pfam contains Red Algal chloroplast L19.
Q6GZV8 (Iridovirus/Ranavirus)	Uncharacterized protein 017L	Q8WUM4 (2r03_A)	Human	Programmed cell death 6-interacting protein	141-406 (53%)	PF13949	ALIX V-shaped domain binding to HIV	Archaea, bacteria, Eukaryotes	86.1	PDCD6IP/ALIX is involved in endocytosis, multifvesicular body biogenesis, membrane repair, cytokinesis, apoptosis and maintenance of tight junctions. PDB: A short peptide motif in lentivirus GAG protein interacts with ALIX V-shaped domain to promote abscission and budding. Pfam: Retroviruses thereby hijack cellular ESCRT machinery.
W6JIY4 (Entomopox alpha)	Uncharacterized protein	I6V3Q6 (3zig_A)	Pyrococcus furiosus	Uncharacterized protein	37-104 (61%)	PF04472	Cell division protein SepF	Archaea, bacteria, eukaryotes	82.0	PDB: SepF-like protein. Literature: SepF is involved in the binary fission of gram positive bacteria at the stage of septum formation between vestigial daughter cells. SepF stimulates the bundling of protofilaments of the tubulin-like FtsZ protein into the contractile "Z ring" that constricts during fission.
P26673 (Vaccinia)	Protein A47	Q5Y4Y6 (5b5r_A)	Mus musculus	Gasdermin-A3	61-245 (73%)	PF17708	Gasdermin PUB domain	Eukaryotes	97.3	UniProt: Gasdermins promote pyroptosis. C-terminal domain of gasdermin-A3 protein, may have autoinhibitory role
K7YHS8 (Megavirus)	Uncharacterized protein	Q9LYC2 (1wf9_A)	Arabidopsis thaliana	NPL4-like protein 1	7-82 (90%)	PF11543	Nuclear pore localization protein NPL4	Bacteria, eukaryotes	97.9	PDB: Beta-grasp fold domain of Npl4. UniProt: NPL4 may be part of a complex that binds ubiquitinated proteins and is necessary for transport of misfolded proteins from ER to cytoplasm. Pfam: NPL4 forms a heterodimer with protein Ufd1.
K7Z7B4 (Megavirus)	Uncharacterized protein	P35998 (5l4g_H)	Human	26S proteasome regulatory subunit 7	72-304 (76%)	PF00004 PF17862	AAA proteins or ATPases associated with various cellular activities AAA+ lid domain	Eukaryotes Eukaryotes	99.5	PDB: Belongs to the heterohexameric ring of AAA ATPases that unfold ubiquitinated target proteins and translocate them into the proteasome's proteolytic chamber.
Q98541 (Chlorella virus)	Uncharacterized protein	Q9P298 (2lon_A)	Human	HIG1 domain family member 1B	3-66 (83%)	PF04588	N-terminal transmembrane region found in hypoxic response proteins	Bacteria, eukaryotes	90.9	Uniprot: Unknown. PDB: Protein Structure Initiative (study of multiple integral membrane proteins).
Q4A223 (Emiliania huxleyi virus)	Uncharacterized protein	P52434 (2f3i_A) B8YB59 (4ayb_G)	Human Saccharolobus shibatae B12	DNA-directed RNA polymerases I, II, and III subunit RPABC3 RNA polymerase subunit 8	5-57 (46%) 7-100 (83%)	PF03870 PF16992	RNA polymerase Rpb8 DNA-directed RNA polymerase, subunit G (RpoG/Rpb8)	Archaea, bacteria, eukaryotes Archaea	95.7 95.6	Here: Component of the DNA-directed RNA Polymerase complex
Q197F5 (Iridoviridae/Chlorirido virus)	Uncharacterized protein 005L	P83653 (1q3j_A)	Acrocinus longimanus (Harlequin beetle)	Anti-microbial peptide Alo-3	120-158 (18%)	PF11410	Antifungal peptide	Plants, insects, arthropods, conesnails, sponges	91.0	UniProt: Anti-fungal activity against Candida glabrata. PDB: Insect knottin-type antifungal peptide
A0A0M5KJJ9 (Mollivirus)	Uncharacterized protein	A0A1A9T940 (2n3p_A)	Asteropus (marine sponge)	Asteropsin_G	103-131 (10%) 167-198 (11%) 241-268 (10%)	NONE	-	-	91.0	PDB: Cystine knot peptide. [Structurally homologous to numerous other toxins]. PDBj: Knottin, toxin.
R4ZER6 (Entomopox beta)	Uncharacterized protein	Q7M1F3 (1bk8_A)	Aesculus hippocastanum (Horse-chestnut tree)	Defensin-like protein 1	44-78 (43%)	PF00304	Gamma-thionin family	Bacteria, eukaryotes	86.0	PDB/literature: Member of subfamily A2 of plant defensins, inhibits growth of a range of fungi. SUPFAM: Fold = Knottins
D2XAM0 (Marseillevirus) D2XAC8 (Marseillevirus) D2XAC9 (Marseillevirus)	Uncharacterized protein Small membrane protein Small membrane protein	P23895 (2i68_A)	Escherichia coli	Multidrug transporter EmrE	65-102 (35%) 5-105 (94%) 13-110 (87%)	PF00893	Small Multidrug Resistance protein	Methanococcus maripaludis, bacteria, Rhizophagus irregularis	86.5 97.9 96.6	UniProt: Expels positively charged hydrophobic drugs (e.g. ethidium bromide, acriflavin) across the inner membrane of *E. coli*, coupled to proton influx.
D2XAS7 (Marseillevirus)	Uncharacterized protein	P0AF61 (2llz_A)	Escherichia coli	Endoribonuclease antitoxin GhoS	15-93 (70%)	PF11080	Endoribonuclease GhoS	Caudovirales, bacteria, eukaryotes	83.5	UniProt: Neutralizes toxin GhoT by digesting GhoT transcripts in sequence-specific manner. Similar 3D structure to Cas2 proteins.

Q84630 (Chlorella virus)	Uncharacterized protein	Q8A109 (4acy_A)	Bacteroides thetaiotaomicron	Endo-alpha-mannosidase	173-437 (61%)	PF16317	Glycosyl hydrolase family 99	Guttaviridae, archaea, bacteria, eukaryotes	99.2	PDB: Bacterial ortholog of ER glycan trimming enzymes which can trim alpha-Glc-1-3-Man(9)GlcNAc(2) to alpha-Glc-1-3-Man.
E3VYK8 (Mimivirus)	Uncharacterized protein R118	Q8A921 (5muj_A) D9ZDQ9 (4udg_F)	Bacteroides thetaiotaomicron Uncultured organism	Beta galactosidase Uncharacterized protein	51-352 (85%) 27-352 (91%)	NONE PF04041	- beta-1,4-mannooligosaccharide phosphorylase	- Archaea, bacteria, eukaryotes	99.1 99.1	PDB: Cleaves complex plant polysaccharides. PDB: Involved in N-glycan degradation in human gut bacteria
R4ZE02 (Entomopox beta) R4ZER5 (Entomopox beta)	Uncharacterized protein Uncharacterized protein	Q53591 (1f1s_A)	Streptococcus agalactiae	Hyaluronate lyase	72-290 (34%) 87-303 (33%)	PF08124	Polysaccharide lyase family 8, N terminal alpha-helical domain	Halobacteria, bacteria, eukaryotes	92.0 88.3	PDB: Enzymatic degradation of hyaluronan and chondroitin sulfates in the extracellular matrix of host tissues, acting at beta-1, 4 glycosidic linkages.
Q4A2A1 (Emiliania huxleyi virus)	Uncharacterized protein	Q8Q045 (2kbn_A) P27694 (1jmc_A) O13988 (1qzg_A)	Methanosarcina mazei Human Schizosaccharomyces pombe	Conserved protein Replication protein A 70 kDa DNA-binding subunit Protection of telomeres protein 1	16-89 (16%) 17-210 (43%) 264-331 (15%)	NONE PF01336,PF16900 PF02765	- OB-fold nucleic acid binding domain,Replication protein A OB domain Telomeric single stranded DNA binding POT1/CDC13	- Bacteria, archaea, eukaryotes Eukaryotes	96.3 96.3 96.9	Here: Single-stranded DNA binding domains covering three distinct OB folds
Q677M6 (Iridovirus/Lymphocystivirus)	Uncharacterized protein	Q5L3Q1 (4o8w_A)	Geobacillus kaustophilus	Spore germination protein	298-425 (28%)	PF17898	Spore germination GerD central core domain	Bacteria, eukaryotes	90.1	PDB: Spore inner membrane lipoprotein alpha-helical homotrimer with role in receptor-mediated activation of downstream germination events. Promotes clustering of inner membrane nutrient receptors, promoting rapid and cooperative germination response.
A0A0H3TLY8 (Faustovirus)	Uncharacterized protein	P46003 (3mwh_A)	Escherichia coli	Arsenical resistance operon trans-acting repressor ArsD	26-94 (23%)	PF06953	Arsenical resistance operon trans-acting repressor ArsD	Archaea, bacteria, eukaryotes	86.3	UniProt: Inducer-independent trans-acting repressor of the ars operon. PDB: Arsenic metallochaperone. Literature: Released from DNA by arsenite binding.
A0A0B5J0R1 (Pandoravirus)	Uncharacterized protein	P32790 (2hbp_A)	Saccharomyces cerevisiae	Actin cytoskeleton-regulatory complex protein SLA1	54-93 (11%)	PF03983	SLA1 homology domain 1, SHD1	Archaea, bacteria, eukaryotes	87.5	UniProt: SLA1 is required for endosome internalization during actin-coupled endocytosis. PDB: sla1p is an adaptor during for uptake of transmembrane proteins containing the NPFxD internalization signal. The SHD1 domain within sla1p recognizes the NPFxD signal.

back to the plasma membrane [443]. In pandoravirus, protein A0A0B5J0R1 might be acting as a decoy to block the endocytotic destruction of viral membrane proteins.

Septum formation: Entomopox alpha protein W6JIY4 showed structural homology to *Pyrococcus* protein SepF (Table 3.6). SepF is involved in the binary fission of gram positive bacteria during septum formation between vestigial daughter cells [444-446]. SepF stimulates the bundling of protofilaments of the tubulin-like GTPase FtsZ protein, thereby stimulating formation of the contractile FtsZ "Z ring" that marks the physical site of division of the mother cell followed by ring constriction and fission [447]. The C-terminal portion of SepF contains the FtsZ binding site and is sufficient to promote FtsZ ring formation, while the N-terminal portion contains a transmembrane domain that presumably anchors the Z ring to the dividing bacterial membrane, allowing the membrane to be pulled inwards during contraction. Entomopox W6JIY4 and *Pyrococcus* SepF are comparable in length (109 vs 131 aa, respectively) and the structural homology region covers the C-terminal FtsZ-binding portion of SepF (Table 3.6). A prokaryotic tubulin-binding type domain therefore seems to have been co-opted in entomopox alpha due to its potential for binding the host cell cytoskeleton. Like bacterial SepF, entomopox alpha W6JIY4 has an N-terminal transmembrane region. A number of steps in virion morphogenesis may involve the cytoskeleton driven constriction or remodeling of membranes, an obvious candidate being virus budding and abscission as promoted by cellular ESCRT complexes in many viruses [448]. Speculatively, in entomopox alpha, this role may have adopted a prokaryotic-type constriction/vesicularization mechanism. However, to have such a fundamental role in the virus lifecycle, the protein would likely be conserved among the entomopoxviruses at least, yet no W6JIY4 orthologs or SepF structural homologs were found in the two other entomopoxvirus type species analyzed here (Table 3.1; data not shown) and no proteins with significant sequence similarity to W6JIY4 were identified in BLASTP searches. This lack of conservation may be more typical of a viral defense type protein. Whatever its role, the structural homology detected here, only just exceeding our 80% threshold, may indicate some divergence during its adaptation to a eukaryotic virus.

Gasdermin: Vaccinia virus protein A47 (P26673) showed structural homology to the C-terminal domain of Gasdermin A3. This is a conserved auto-inhibitory domain found in various Gasdermins. Caspase-directed Gasdermin cleavage at a linker region connecting the N- and C-terminal domains unmasks the N-terminal domain from auto-inhibition, allowing it to undergo a conformational change that promotes oligomerization leading to the formation of membrane-spanning pores [449], pyroptotic cell death, and cytokine release [450-453]. A47 may be a molecular decoy for the activated (unmasked) Gasdermin N-terminal domain, thereby suppressing pyroptosis. Protein A47 is expressed early during vaccinia virus infection and contains unusually high numbers of CD8+ T cell epitopes able to prime T cells in vivo [454].

Proteasomal degradation: Megavirus protein K7YHS8 was structurally homologous to the N-terminal beta-grasp fold domain of Arabidopsis NPL4-like protein 1. This fold is found in diverse protein families [455] including the compact globular ubiquitin-like (UbL) domain found in ubiquitin and other proteins. UbL-containing proteins bind substrates destined for degradation and also bind subunits of the proteasome, and thus regulate protein turnover [456]. The beta-grasp fold of NPL4-like protein 1 is likely also a UbL domain [455]. NPL4 interacts with the N-terminal domain of the AAA ATPase VCP/p97 [457] which has diverse functions in the cell mostly centered around ubiquitin-dependent processes [458]: For example, it facilitates the degradation of monoubiquitylated, polyubiquitylated, and non-degradative ubiquitin chain-containing proteins. It also extracts proteins from membranes and other cellular structures for degradation (or activation in the case of transcription factors precursors). It seems possible that megavirus protein K7YHS8 may regulate the degradation of megavirus proteins during infection.

Also in megavirus, protein K7Z7B4 was structurally homologous to a variety of AAA domain containing proteins, with K7Z7B4 residues 135 – 304 showing AAA domain alignment. A small subset of AAA domain containing proteins, namely the AAA domain-containing proteasome regulatory subunits, showed more extensive homology to K7Z7B4. The top homolog, overall, was 26S proteasome regulatory subunit 7 protein (Table 3.6 - also known as PSMC2/RPT1/MSS1) [459]. This appears to be the first finding of an MSS1 homolog in a non-eukaryote. The 26S proteasome comprises a barrel-shaped, proteolytic 20S core with a 19S regulatory "lid" at one or both ends. 19S serves to unfold ubiquitinated target proteins and to translocate them into the 20S proteolytic chamber [460]. 19S contains at least 18 subunits including a hexameric ring of six distinct AAA ATPases - one of which is our top structural homolog, subunit 7. Subunit 7 appeared unique in both the degree and extent of homology with K7Z7B4, being the only protein showing structural homology to the N-terminal side of the AAA ATPase domain (K7Z7B4 residues 72 – 133), a region of unknown function. Other proteasome regulatory subunits show homology within this region (residues 82 – 133) K7Z7B4, which has the Pfam designation "Proteasomal ATPase OB C-terminal domain" (PF16450). However, since these other regulatory subunits lacked RCSB structural data immediately N-terminal to residue 82, it was unclear whether subunit 7's slightly more extensive homology was real or illusory. Interestingly, K7Z7B4 lacks the "AAA+lid" domain present in these proteasome regulatory subunits. K7Z7B4 may serve to modulate the target specificity of the proteasome. In this regard, it would seem to be a reasonable partner for K7YHS8, above. Nonetheless, some proteasomal regulatory subunits (eg. subunit 6A/PSMC3) are multifunctional, with roles in transcriptional tumor suppression and binding to HIV TAT protein [461]. This appears to be the first finding of a proteasomal subunit in a virus of any kind.

Hypoxic response: Chlorella virus protein Q98541, which is currently annotated as an integral membrane protein, was found to be a structural homolog of human HIG1 domain family member 1B, an integral membrane protein induced by hypoxia [462] whose functions remain poorly understood. This may be the first identification of a HIG1 domain family member in a virus.. This may be the first identification of a HIG1 domain family member in a virus.

Antimicrobial peptides: <u>Cystine knot proteins:</u> Cystine knots are highly stable structural motifs comprising four beta sheets crosslinked by three disulfide bridges [463]. One class of cystine knot proteins, the inhibitor cystine knot ("Knottin") class, exhibits toxic, insecticidal or anti-microbial activity [464, 465]. The 217 residue chloriridovirus protein Q197F5 and the 281 residue mollivirus protein A0A0M5KJJ9 contained regions homologous to the Knottin motif. Specifically, Q197F5 (residues 120 - 158) was homologous to the antimicrobial and antifungal peptides Alo-3 (Harlequin beetle, Table 3.6) and antimicrobial peptide 1 (Pokeweed, not shown) along with various conotoxins and other toxin peptides (not shown). A0A0M5KJJ9 contains three adjacent regions of structural homology (between residues 100 and 273) to the pharmacologically inert 32 residue peptide "Asteropsin G" from the marine sponge *Asteropus*. While these regions of A0A0M5KJJ9 matched the general requirements for cystine knots, they did not match the highly specific requirements for Knottin, perhaps consistent with the apparently non-toxic character of Asteropsin G [466]. We are therefore circumspect about whether A0A0M5KJJ9 has actual knottin character. Cystine knots have been identified in many plants and animals, but have not, to our knowledge, been reported in a virus.

Entomopox beta protein R4ZER6 was structurally homologous to Defensin-like protein 1 from horse chestnut, which is a knottin-fold protein. Defensins more generally are arthropod and insect peptides active against Gram-positive bacteria [467, 468]. They are found in many species, including Lepidoptera, the insect host of beta entomopoxviruses [469].

Toxin-antitoxin systems: Marseillevirus contained three proteins comparable in size and structure to the 110-residue *E. coli* multidrug resistance-conferring membrane protein EmrE. EmrE belongs to a family of small multidrug resistance (SMR) transporters driving the efflux of aromatic cationic drugs from the cytoplasm via a drug/H+ antiport mechanism [470]. EmrE's transport substrates have few common structural features [470] [471]. More than 200 SMR genes have been identified in bacteria (plus a few archaea) including bacterial strains with multiple paralogs [471]. All share a critical conserved glutamate (Glu-14) also present in one of the marseillevirus proteins (D2XAC8; residue 15). Their occurrence on plasmids or their proximity in the bacterial chromosome to insertion elements (e.g. EmrE is encoded within the DLP12 cryptic lambdoid prophage region of the E. coli chromosome) suggests a strategy for gene spread via horizontal gene transfer. EmrE homologs were previously found in two Yellowstone Lake phycodnavirus metagenomes [471]. Additional marseillevirus orthologs of the three marseillevirus proteins can be found by BLASTP (data not shown). These maybe the first identifications of members of this SMR protein family in eukaroytes or their viruses, and their roles are not obvious, though paralogous bacterial transporters show substrate complementarity [471]. Drug resistance proteins may occur in NCLDV to promote virus persistence via symbiosis, immunity or addiction [472] or due to their amoebal hosts residing in complex aqueous and phagocytic environments.

Marseillevirus protein D2XAS7 showed structural homology to the short *E. coli* antitoxin GhoS - an endoribonuclease that targets a specific site in a specific *E. coli* mRNA – namely that for toxin GhoT [473]. GhoT functions by damaging the *E. coli* inner membrane via the formation of transient transmembrane pores [473, 474]. Due to the nature of its fold (*E. coli* GhoS shows structural homology to the short CRISPR-associated sequence-specific endoribonuclease CAS2 [473]) we speculate that D2XAS7 acts as an endoribonuclease during marseillevirus infection, though evidence it has antitoxin function therein is lacking.

Glycosylation and oligosaccharide degradation: Chlorella virus protein Q84630 was identified as a structural homolog of bacterial membrane endo-alpha mannosidase, an enzyme required for cell wall biosynthesis [475, 476]. The latter enzyme is a structural prototype for glycan trimming enzymes of the endoplasmic reticulum [476]. Cellular mannosidases function early during the diversification and maturation of protein-attached glycans in the ER and Golgi. Viral surface and secreted proteins are glycosylated [477], and the hijacking of N-glycan synthesis can occur in viral and other diseases [476]. Mannosidases are also implicated in ER-associated protein degradation [478]. Speculatively, Q84630 may serve to redirect the host protein glycosylation machinery to the production of an antigenically distinct pattern of viral protein glycosylation.

Mimivirus protein E3VYK8 was structurally homologous to two enzymes with roles in cleaving oligosaccharides at the glycosidic bond, namely the crystallized N-terminal region of beta galactosidase from *Bacteroides*, and beta-1,4-mannooligosaccharide phosphorylase. Entomopox beta proteins R4ZE02 and R4ZER5, showed structural homology to the N-terminal alpha-helical domain of streptococcal Hyaluronate lyase [479], a secreted enzyme that promotes bacterial tissue invasion by degrading the glycosaminoglycans found in extracellular matrix

[480]. The N-terminal alpha-helical domain in polysaccharide lyase family 8 (PL8) enzymes possesses the catalytic site and contributes one side of a structural cleft that binds substrate [479]. Glycosaminoglycans are found in the insect midgut [481, 482], and the degradative activities of R4ZE02 and R4ZER5 may facilitate the host spread of entomopox beta. Alternatively, in entomopoxvirus this domain may have lost catalytic activity - providing, instead, a viral attachment protein acting in comparable fashion to the glycosaminoglycan binding chordopoxvirus attachment proteins [112, 113, 483].

Structural domains

ssDNA binding domains: Emiliania huxleyi virus protein Q4A2A1 showed three apparent single-stranded DNA binding domains covering three distinct types of OB fold (Table 3.6). It may have a role in virus genome replication and/or the maintenance of virus genome telomeres.

Coiled-coil domain**:** Two overlapping regions of the 447 residue lymphocystivirus protein Q677M6 (residues 298 - 376 and 346 – 425) showed structural homology to a 121-residue core domain of the 155 residue *Bacillus* lipoprotein GerD. GerD is located in the inner membrane of the bacterial spore and functions in its rapid response to external germinants [484]. The 121-residue core peptide forms an alpha helical homotrimer in solution and crystallizes into a neatly twisted superhelical rope [485] that may nucleate the clustering of spore inner membrane proteins. The corresponding triple-helical region in lymphocystivirus could play any number of roles in virus biology. Vaccinia virus attachment protein A27, for example, forms a triple coiled-coiled homotrimer [123, 486, 487].

Ars operon repressor**:** A 69 residue region of the 290 residue faustovirus protein A0A0H3TLY8 is structurally homologous to the 120 residue protein ArsD, a plasmid-encoded trans-acting repressor of the bacterial arsenical resistance (‘ars’) operon (arsRDABC). ArsD represses the operon to basal levels in the absence of trivalent/pentavalent arsenite or antimony metalloids [488] by binding a 24 nt segment of the ars promoter [488] and is released from DNA by arsenite binding. ArsD also sequesters toxic intracellular metalloids [489] and shuttles them to the ATPase component of the arsenical pump (ArsA, encoded within the arsRDABC operon) for reduction and expulsion [490]. It seems unlikely that faustovirus A0A0H3TLY8 has any metal binding role since none of the metal-binding cys of ArsD [491, 492] are conserved (A0A0H3TLY8 is entirely cys-free). However, this fold may have been co-opted for its DNA binding properties or some other role.

Transmembrane domains and potential signal sequences

In addition to structural homology searching, we enumerated predicted transmembrane (TM)-containing [493] and potential secretory signal peptide-containing [494] proteins among the 20 viruses (Fig. 3.10). One of the more unexpected of the predicted TM domains/membrane anchors was located at the N-terminus of the ETF1 subunit of the heterodimeric vaccinia virus transcription factor VETF. However, VETF is considered to be packaged in the virion core, compartmentalized away from the virion envelope by the proteinaceous virion core wall. Since the repression of either ETF1 or ETF2 synthesis during infection is known to lead to a block in virion morphogenesis [495, 496], it seems possible that this TM domain may be a membrane attachment point during virion morphogenesis—perhaps for the packaging of a vaccinia transcriptosome-based assembly [497] or “nucleoid” [498]. In support of such a model, the morphogenic block upon repression of the ETF2 subunit yields immature virions lacking genomic DNA [495]. The structure of ETF1 was subsequently reported, revealing that these residues formed a buried helix, not a transmembrane domain.

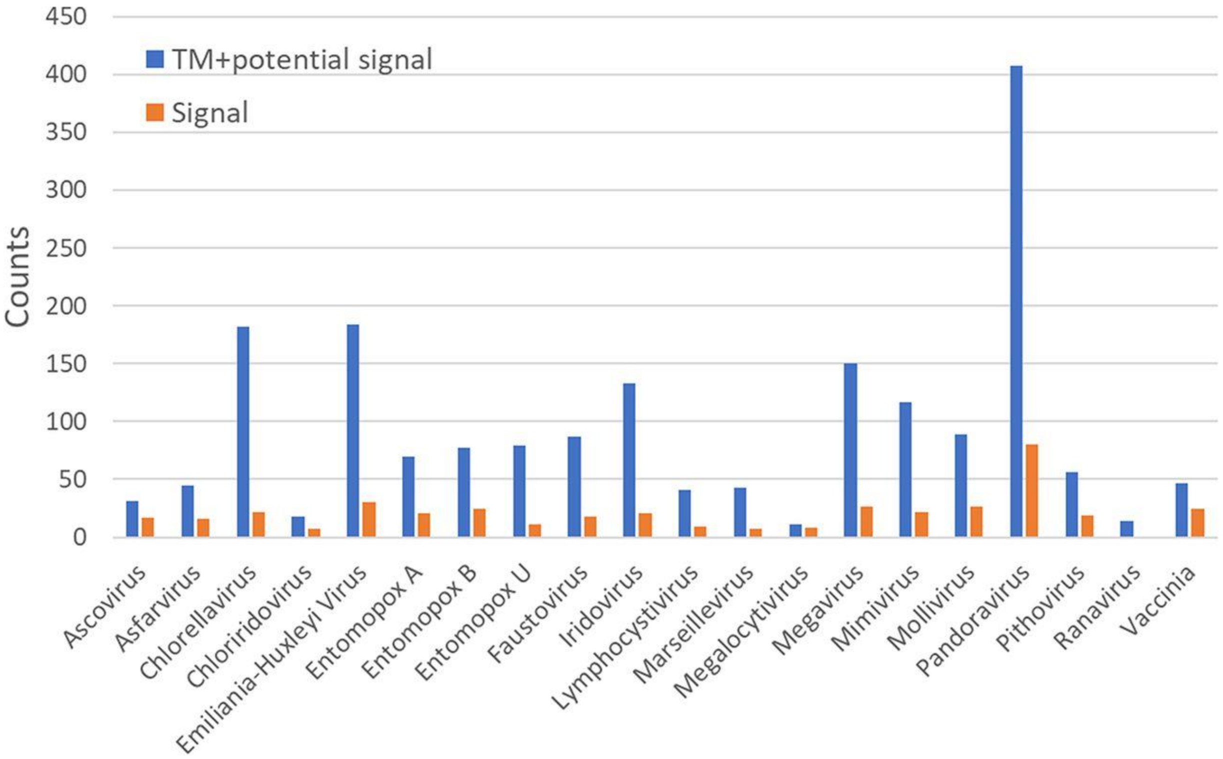


Figure 3.10. Numbers of proteins per viral proteome possessing a predicted transmembrane domain [493] and potential secretory signal peptides [494] enumerated per viral proteome. Some overlap may exist between the two sets of counts.

Conclusions

Here, structural homology was used to expand the annotation of previously unclassified proteins. This approach proved very successful. Gaps among "core" (NCVOG) proteins were filled, and additional RNAP subunits and basal transcription factor homologs were identified, along with many new endonucleases and proteins with functions not previously described in any virus.

In considering the merits of structural over sequence homology, the latter seems challenged in extending protein families with low sequence homology, such as the REases, or those with high sequence homology and therefore already essentially complete, such as the serine/threonine protein kinases. The structural approach will be as powerful as the number of annotated three-dimensional structural models present in the PDB, with the possibility of a bias in structural databases toward proteins of medical and/or economic importance.

Materials and Methods

Version 3.0.0 of the HHsuite package was installed on the High-Performance Computing Cluster at University of California—Irvine/Research Cyber Infrastructure Center. The usage of HHsuite, including the interpretation of results, has been well-described by its developers and earlier users (https://github.com/ soedinglab/hh-suite/wiki and https://toolkit.tuebingen.mpg.de/tools/hhpred) [356, 358, 361-363, 395, 499]. Briefly, for each of the 20 viruses in Table 3.1, the UniProt complete proteome was downloaded and the resulting data set deconstructed to individual FASTA protein sequence files. For each of the resulting query protein sequences, a multiple-sequence alignment (MSA) was generated using HHblits in batch mode against "uniprot20," a database of UniProt sequences clustered at the 20% sequence identity level provided by HHsuite.

The threshold for sequence inclusion in an MSA was an E value of 103. After supplementing MSAs with PSIPRED-generated secondary structural information via the addss.pl tool, profile HMMs combining information from MSAs and their corresponding secondary structure predictions were generated via the tool HHmake. Via the HHsearch tool, the resulting profile HMM were used as sequential queries against a database derived from pdb70 (downloaded from the HHsuite server). Searches were made in local alignment mode with the maximum accuracy alignment algorithm (MAC) “on.” Any initial search terminating with error was rerun using the HHpred server (part of the MPI Bioinformatics Toolkit) with greater numbers of search iterations, modified MAC realignment, or MAC turned off.

One output (.hhr) file was generated per match per query protein and contained extensive header information and a text version of the structural alignment. In-house code was used to extract, from .hhr files, the PDB identifiers and chains of matching structures, statistical match scores, query homology regions, and the target homology regions, then tabulate them on a per query basis. The resulting tables were annotated with query accession number, descriptor, and protein length (from the individual protein FASTA files used as HHblits inputs), then annotated as well with the query protein’s UniProt keyword, gene ontology (GO) biological process, and GO molecular function annotations. The resulting tables were then thresholded at 80% probability in accordance with reports (10, 17) on the high specificity and accuracy of this threshold.

Filtered data were then further annotated manually with motif, domain, and/or other protein information derived from www.rcsb.org by manual lookup via the homology target’s PDB identifier. Manual annotation of homology regions was aided by visual inspection in RCSB’s “full protein feature view” of regions in the target’s primary structure covered by X-ray crystal structures and/or coincident with domains in the Pfam database, transmembrane domains, and/or other features. These annotations (notably all associated Pfams) were then transferred to the query sequence after correction for the differential sequence positions of the homology region in query and target. Query proteins with multiple distinct homology regions were annotated according to all, and query proteins with overlapping homology regions to distinct target proteins were annotated according to the highest probability score. For Pfams within higher order groupings (superfamilies or clans), the former were replaced with the latter (e.g., for heatmap figures).

Rules for the assignment of HHsearch output to multiple superfamilies (heatmap). Five ambiguous situations were handled as follows. (i) A query structurally homologous to distinct superfamilies via distinct regions of the query (e.g., N-terminal F-box, C-terminal ankyrin) was enumerated under both superfamilies. (ii) If a query was structurally homologous to a single target protein with repeats of a superfamily match, each unique superfamily was listed only once per query and counted as a single hit for the heatmap. (iii) If a query was structurally homologous to multiple target proteins in the 80 to 100% probability range that included multiple superfamilies, only the superfamily associated with the highest probability target was enumerated, or the target with greatest coverage if probabilities for both were similar. (iv) Query proteins with highly fragmented homology regions (e.g., collagen-like proteins and query proteins with extended coiled coil regions) were searched again via the HHpred server with a lower MAC realignment threshold or in global realignment mode to yield greater alignment length. (v) Target proteins with no Pfam identifiers across the homology region (Fig. 3.2, example 7) were excluded from heatmaps.

Transmembrane and secretory signal peptide search. FASTA files of the complete UniProt proteome for each of the 20 viruses were searched for putative transmembrane helices and secretory signal peptides using TMHMM v2.0 [493] (https://services.healthtech.dtu.dk/service.php?TMHMM-2.0) and SignalP v5.0 [494] (https://services.healthtech.dtu.dk/service.php?SignalP-5.0), respectively. TMHMM v2.0 was run with a single-line output per protein, then filtered to retain proteins with at least one predicted transmembrane domain (which may also serve as a signal peptide). For SignalP v5.0, proteomes were searched for matches in Eukarya, then filtered to retain proteins with predicted secretory peptides.

Dolpenny. Dolpenny [500] and Consense programs were installed as part of the PHYLIP package from the University of Washington website (http://evolution.genetics.washington.edu/phylip.html). Dolpenny was run using the Dollo parsimony method with species order set to be continually reconsidered. Ancestral states for all RNA polymerase subunits of chlorella virus and for megalocytivirus RPB5 were represented by a question mark ("?"). For all other RNA polymerase subunits, TFIIS, TFIIB, and TBP, they were represented as 1 and 0 for presence and absence, respectively. A rooted consensus tree was built from the Dolpenny output using Consense with the consensus type "Majority rules (extended)."

CHAPTER 4

The vaccinia virion: filling the gap between atomic and ultrastructure

Abstract

We have investigated the molecular-level structure of the vaccinia virion *in situ* by protein-protein chemical crosslinking, identifying 4609 unique-mass crosslink ions at an effective FDR of 0.33%, covering 2534 unique pairs of crosslinked protein positions, 625 of which were inter-protein. The data were statistically non-random and rational in the context of known structures, and showed biological rationality. Crosslink density strongly tracked the individual proteolytic maturation products of P4a and P4b, the two major virion structural proteins, and supported the prediction of transmembrane domains within membrane proteins. A clear sub-network of four virion structural proteins provided structural insights into the virion core wall, and proteins VP8 and A12 formed a strongly-detected crosslinked pair with an apparent structural role. A strongly-detected sub-network of membrane proteins A17, H3, A27, and A26 represented an apparent interface of the early-forming virion envelope with structures added later during virion morphogenesis. Protein H3 seemed to be the central hub not only for this sub-network but also for an ‘attachment protein’ sub-network comprising membrane proteins H3, ATI, CAHH (D8), A26, A27 and G9. Crosslinking data lent support to a number of known interactions and interactions within known complexes. Evidence is provided for the membrane targeting of genome telomeres. In covering several orders of magnitude in protein abundance, this study may have come close to the bottom of the protein-protein crosslinkome of an intact organism, namely a complex animal virus.

Author summary

Vaccinia is one of the most complex virions among the animal viruses, containing 70+ distinct gene products. Although virion ultrastructure has been apparent, at least in outline by electron microscopy since the year 1961 or earlier, its molecular architecture is largely unknown: Vaccinia is resistant to classical structural approaches requiring virus crystallization and moderately resistant to cryoEM. Molecular approaches requiring the maintenance of protein assemblies during virion deconstruction, reconstruction of protein complexes in heterologous or in vitro systems, or internalization of bulky reagents such as antibodies or gold particles may have been already pursued close to exhaustion. Here, protein interfaces within and around the intact virion were identified by virus incubation with bifunctional chemical crosslinkers in situ followed by proteolysis and peptide-level mass spectrometry. This minimally invasive approach revealed the molecular arrangements of structural and membrane protein complexes within the virus, confirming and extending several aspects of virus biology.

Introduction

The virion of vaccinia, the prototypical poxvirus, is one of the largest among the animal viruses. While its ultrastructural characterization is the beneficiary of 60+ years of electron microscopic examination [30, 189, 501] and references therein, attempts to better understand its molecular and atomic architecture have fallen foul of various properties of the vaccinia virion such as asymmetry, polymorphic character, tendency to aggregate, and the general incompatibility of enveloped viruses with X-ray crystallography.

Electron microscopy (EM) and atomic force microscopy (AFM) studies have established clear ultrastructural compartments of the mature virion (MV) [29] including a central, genome-containing ‘core’ that also houses a number of virus-encoded enzymes of mRNA transcription and modification, a proteinaceous wall surrounding the core, a pair of ‘lateral body’ structures flanking the core wall, a single lipid bilayer envelope, and an outer protein-rich coat that appears late during maturation. The virion contains between 58 and 73 distinct gene products [101]. Some of these have been localized at low resolution on the basis of immunogold EM [73, 76, 92, 502, 503], while the compartmental locale of others can be inferred from clearly identifiable transmembrane (TM) domains and other bioinformatics signatures, known function and/or the conditions required for the extraction from the virion. Proteins and visible structures localizing to outer compartments of the virion (outside of the core) have been identified via their fractionation in vivo during virus entry [76, 167] or under pseudo-entry conditions recreated by the gentle, controlled treatment of virions with nonionic detergent or nonionic detergent+disulfide reductant [73, 79, 80, 87, 504]. A number of core enzymes, including the virus-encoded multisubunit DNA-dependent RNA polymerase (RPO), heterodimeric virion capping enzyme (CA), early transcription factor (ETF), poly(A) polymerase (PAP), two protein kinases, at least two proteases and two glutaredoxins have been released from the virion under more harsh conditions (0.2% ionic detergent (sarkosyl) and high salt [309]), retaining solubility, integrity and activity after detergent removal [29]. By contrast, a number of structural proteins of the virion core remain insoluble during virion extraction even in ionic detergent.

Aside from these compartmentalization approaches, little is known of the virion's internal organization at the molecular level. Certainly, the heteromultimeric status of the above core enzymes has long been known [29], and the homomultimeric status of yet other virion proteins has been revealed by X-ray crystallography (e.g., proteins H1 [505, 506] and A27 [128]). Some binary protein-protein interactions have been successfully recapitulated and identified in a yeast two-hybrid system [507]. Other proteins, and fragments thereof, have been co-immunoprecipitated from cell extracts, pulled-out as tagged complexes [30] or inferred by genetic and directed mutational studies.

However, larger macromolecular and ultrastructural assemblies clearly dissociate under the conditions required for full virion disruption. For example, the presence, within the virion core, of a 'transcriptosome' assembly was inferred in studies down-regulating the vaccinia RNA polymerase subunit RAP94. Under non-permissive conditions, virions were morphologically mature but showed low infectivity [497]. Albeit the virus genome was packaged in normal amounts as were ETF and the structural proteins, low or undetectable amounts of RPO, CA, PAP large subunit, and proteins NTP1, RNA helicase and topoisomerase were packaged suggesting the coordinated packaging of the latter components. Such a 'transcriptosome' complex may correspond to the formation, within the core, of a genome-containing tubular ultrastructure [508] that can be resolved by EM under sample preparation conditions that include high pressure freezing [96]. However, no such ultrastructure or any subassembly thereof has been isolated biochemically: Capping enzyme can form a binary complex with RPO in vitro [509], but the soluble fraction from a sarkosyl virion core lysate, for example, even under gentle gradient sedimentation conditions, has yielded no higher order assemblies beyond the sedimentation of RPO as a discrete entity and the partial co-sedimentation of RPO with viral capping enzyme and NTP1 [510]. Other enzymes, including those apparently co-packaged with RAP94 (above) sedimented separately, towards the top of the gradient, suggesting an irreversible disruption of interactions within the transcriptosome upon core rupture. To our knowledge, no comprehensive transcriptosome, or other packaged superstructure has been (re)assembled biochemically as a positive correlate to the subtractive approaches of genetics.

Here, we have taken an approach to the molecular structure of the vaccinia virion that is neither destructive, reconstructive nor exclusively applicable to binary complexes, namely protein-protein crosslinking mass spectrometry (XL-MS). We address the virion in its natural state in situ, with the potential to interrogate multivalent protein complexes. Technical challenges in this approach were not inconsiderable: At the outset of the current study, higher profile XL-MS studies in the literature had focused upon stoichiometric or near-stoichiometric isolated protein complexes, containing around ten or fewer polypeptides, with known crystal structures. Examples of these would include the 26S proteasome [511], multi-ringed TRiC/CCT chaperonin [512, 513], the RNA polymerase II pre-initiation complex [245, 514, 515], RNA polymerase I [516] and RNA polymerase III [517]. By contrast, the vaccinia virion likely contains a variety of protein complexes covering an abundance dynamic range of ~5000 [100] or greater, only a minority of which have yielded X-ray crystallographic structures. Our XL-MS results with vaccinia are described below.

Results

Approach

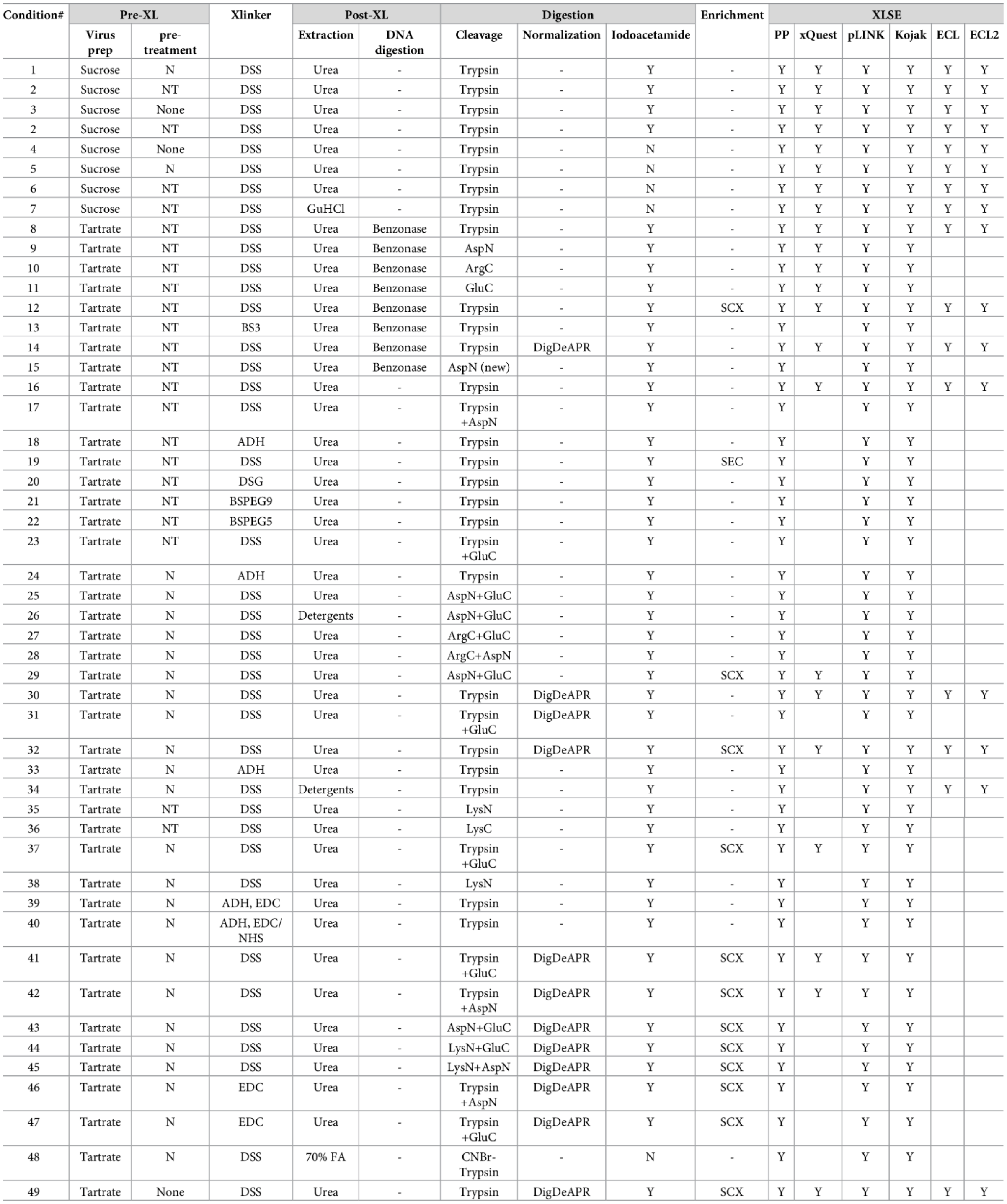
Virions (intact or activated for mRNA transcription) were incubated with bifunctional chemical crosslinkers to impose inter-protein distance restraints. Crosslinked virus was then dissolved and trypsinized to peptides, followed by peptide-level nanoLC-MS/MS and bioinformatics to identify crosslinked peptides. For disuccinimidyl suberate (DSS), the crosslinker used in the majority of experiments, the restraint comprised a lysine Nζ-Nζ distance of 10–11.4 Å with corresponding Cα-Cα distances of 32 Å (give or take molecular dynamics considerations). Crosslinkable lysines thereby sweep a sphere of Cα-Cα distances up to ~6 nm, or ~2% of the diameter of a vaccinia virion for proteins not forming extended, repeating arrays.

Due to the low intrinsic ionizability of crosslinked peptide pairs and the potential for low saturation crosslinking within/between low abundance proteins in the virion, a strategy of variation [101] (Table 1) was implemented to maximize opportunities for the detection of crosslink (XL) ions (Fig 1). This was combined with a total of six distinct XL search engines, used in parallel (Fig 1 and Materials & methods). After data thresholding and filtering, a unique metascore (‘DFscore’, or detection frequency score) was introduced as a guide to the extent of internal confirmation within the dataset.

Overall project dataset (‘crosslinkome’)

The resulting XL dataset yielded a total of 4609 confidently-identified unique-mass ions, each corresponding to a crosslinked peptide pair. Of these, 1486 (32.2%) had a DFscore > 1. The highest DFscore for any ion was 178, and the four top-scoring ions each corresponded to P4a intra-protein XL, of which the two highest scoring were light/heavy versions of the same ion and the third represented a small shift in XL position for one of the two crosslinked peptides.

Table 4.1. Crosslinking experiments and experimental conditions. 49 distinct sets of experimental conditions were sampled as a sparse-matrix through Fig 1.



Condition#	Pre-XL		Xlinker	Post-XL		Digestion			Enrichment	XLSE					
	Virus prep	pre-treatment		Extraction	DNA digestion	Cleavage	Normalization	Iodoacetamide		PP	xQuest	pLINK	Kojak	ECL	ECL2
1	Sucrose	N	DSS	Urea	-	Trypsin	-	Y	-	Y	Y	Y	Y	Y	Y
2	Sucrose	NT	DSS	Urea	-	Trypsin	-	Y	-	Y	Y	Y	Y	Y	Y
3	Sucrose	None	DSS	Urea	-	Trypsin	-	Y	-	Y	Y	Y	Y	Y	Y
2	Sucrose	NT	DSS	Urea	-	Trypsin	-	Y	-	Y	Y	Y	Y	Y	Y
4	Sucrose	None	DSS	Urea	-	Trypsin	-	N	-	Y	Y	Y	Y	Y	Y
5	Sucrose	N	DSS	Urea	-	Trypsin	-	N	-	Y	Y	Y	Y	Y	Y
6	Sucrose	NT	DSS	Urea	-	Trypsin	-	N	-	Y	Y	Y	Y	Y	Y
7	Sucrose	NT	DSS	GuHCl	-	Trypsin	-	N	-	Y	Y	Y	Y	Y	Y
8	Tartrate	NT	DSS	Urea	Benzonase	Trypsin	-	Y	-	Y	Y	Y	Y	Y	Y
9	Tartrate	NT	DSS	Urea	Benzonase	AspN	-	Y	-	Y	Y	Y	Y		
10	Tartrate	NT	DSS	Urea	Benzonase	ArgC	-	Y	-	Y	Y	Y	Y		
11	Tartrate	NT	DSS	Urea	Benzonase	GluC	-	Y	-	Y	Y	Y	Y		
12	Tartrate	NT	DSS	Urea	Benzonase	Trypsin	-	Y	SCX	Y	Y	Y	Y	Y	Y
13	Tartrate	NT	BS3	Urea	Benzonase	Trypsin	-	Y	-	Y		Y	Y		
14	Tartrate	NT	DSS	Urea	Benzonase	Trypsin	DigDeAPR	Y	-	Y	Y	Y	Y	Y	Y
15	Tartrate	NT	DSS	Urea	Benzonase	AspN (new)	-	Y	-	Y		Y	Y		
16	Tartrate	NT	DSS	Urea	-	Trypsin	-	Y	-	Y	Y	Y	Y	Y	Y
17	Tartrate	NT	DSS	Urea	-	Trypsin +AspN	-	Y	-	Y		Y	Y		
18	Tartrate	NT	ADH	Urea	-	Trypsin	-	Y	-	Y		Y	Y		
19	Tartrate	NT	DSS	Urea	-	Trypsin	-	Y	SEC	Y		Y	Y		
20	Tartrate	NT	DSG	Urea	-	Trypsin	-	Y	-	Y		Y	Y		
21	Tartrate	NT	BSPEG9	Urea	-	Trypsin	-	Y	-	Y		Y	Y		
22	Tartrate	NT	BSPEG5	Urea	-	Trypsin	-	Y	-	Y		Y	Y		
23	Tartrate	NT	DSS	Urea	-	Trypsin +GluC	-	Y	-	Y		Y	Y		
24	Tartrate	N	ADH	Urea	-	Trypsin	-	Y	-	Y		Y	Y		
25	Tartrate	N	DSS	Urea	-	AspN+GluC	-	Y	-	Y		Y	Y		
26	Tartrate	N	DSS	Detergents	-	AspN+GluC	-	Y	-	Y		Y	Y		
27	Tartrate	N	DSS	Urea	-	ArgC+GluC	-	Y	-	Y		Y	Y		
28	Tartrate	N	DSS	Urea	-	ArgC+AspN	-	Y	-	Y		Y	Y		
29	Tartrate	N	DSS	Urea	-	AspN+GluC	-	Y	SCX	Y	Y	Y	Y		
30	Tartrate	N	DSS	Urea	-	Trypsin	DigDeAPR	Y	-	Y	Y	Y	Y	Y	Y
31	Tartrate	N	DSS	Urea	-	Trypsin +GluC	DigDeAPR	Y	-	Y		Y	Y		
32	Tartrate	N	DSS	Urea	-	Trypsin	DigDeAPR	Y	SCX	Y	Y	Y	Y	Y	Y
33	Tartrate	N	ADH	Urea	-	Trypsin	-	Y	-	Y		Y	Y		
34	Tartrate	N	DSS	Detergents	-	Trypsin	-	Y	-	Y		Y	Y	Y	Y
35	Tartrate	NT	DSS	Urea	-	LysN	-	Y	-	Y		Y	Y		
36	Tartrate	NT	DSS	Urea	-	LysC	-	Y	-	Y		Y	Y		
37	Tartrate	N	DSS	Urea	-	Trypsin +GluC	-	Y	SCX	Y	Y	Y	Y		
38	Tartrate	N	DSS	Urea	-	LysN	-	Y	-	Y		Y	Y		
39	Tartrate	N	ADH, EDC	Urea	-	Trypsin	-	Y	-	Y		Y	Y		
40	Tartrate	N	ADH, EDC/ NHS	Urea	-	Trypsin	-	Y	-	Y		Y	Y		
41	Tartrate	N	DSS	Urea	-	Trypsin +GluC	DigDeAPR	Y	SCX	Y	Y	Y	Y		
42	Tartrate	N	DSS	Urea	-	Trypsin +AspN	DigDeAPR	Y	SCX	Y	Y	Y	Y		
43	Tartrate	N	DSS	Urea	-	AspN+GluC	DigDeAPR	Y	SCX	Y		Y	Y		
44	Tartrate	N	DSS	Urea	-	LysN+GluC	DigDeAPR	Y	SCX	Y		Y	Y		
45	Tartrate	N	DSS	Urea	-	LysN+AspN	DigDeAPR	Y	SCX	Y		Y	Y		
46	Tartrate	N	EDC	Urea	-	Trypsin +AspN	DigDeAPR	Y	SCX	Y		Y	Y		
47	Tartrate	N	EDC	Urea	-	Trypsin +GluC	DigDeAPR	Y	SCX	Y		Y	Y		
48	Tartrate	N	DSS	70% FA	-	CNBr-Trypsin	-	N	-	Y		Y	Y		
49	Tartrate	None	DSS	Urea	-	Trypsin	DigDeAPR	Y	SCX	Y	Y	Y	Y	Y	Y

3725 of the 4609 unique-mass ions represented intraprotein XL while 884 were inter-protein, consistent with the known tendency for XL to fall within rather than between proteins. 273 of the 884 inter-protein XL ions had a DFscore > 1 among which the highest DFscore was 83 (P4a-position 876 crosslinked to P4b-position 563).

By merging (a) distinct charge states for a crosslinked peptide, (b) identical crosslinked accessions/positions detected within distinct peptide species, (c) light/heavy isotopic forms of the crosslinker and (d) crosslinked peptides with secondary modifications, the 4609 unique XL ion masses collapsed down to 2534 unique pairs of residues within the proteome. 625 of these were inter-protein and, of these, 157 (25.1%) had a DFscore > 1 with the highest DFscore for an inter-protein accession/position pair being 475 (for the P4a-876/P4b-563 XL mentioned above). This accession/position pair was represented by 43 distinct m/z crosslinked peptide ions. Appendix 2.Fig1 shows crosslinking partners among all proteins considered to be packaged in the virion [101] for which XL were detected.

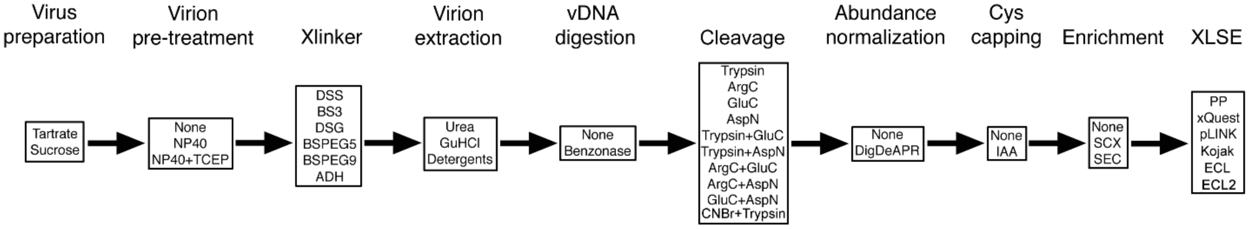


Figure 4.1. Vaccinia MV protein crosslinking multi-threaded workflow and strategy of variation. In total, 49 distinct pathways through the conditions matrix (Table 1) were sampled over 53 experiments. The final step of the workflow ('XLSE', for 'crosslink search engine') was a parallel, rather than a variable element.

Validation

Orthogonal approaches to the validation of in situ–detected protein-protein interactions all seemed less direct than XL-MS itself (involving virion disruption, recapitulation of interactions in vitro, and/or the expression of virus proteins in heterologous systems). We therefore sought to validate the XL dataset via inference criteria, asking four basic questions as follows:

(a) Was reasonable bioinformatic rigor applied (e.g., in program score thresholding)?

All six XL search engines employed a target-decoy approach [518] (Table 2) and primary score thresholding comprised false discovery rate (FDR) or its surrogate, q-value (Materials & methods). For four of the six engines we took the unprecedented step of also applying a second threshold, via the score-type that is native to the engine itself (Table 2). A small fraction of the ions discarded solely on the basis of threshold 2 were then rescued according to the criteria described in Materials & Methods. With a primary threshold alone, namely 5% FDR, around 230 of our 4609 unique-mass ions would have arisen from our decoy database. Via our dual thresholding/rescue approach (see the "Data Assembly" section of "Materials & methods"), only 15 of the 4609 ions involved a decoy accession, representing an effective FDR of just 0.33%—an exceptionally low number. We regard our low effective FDR as a bona fide validation step,

Table 4.2. XL search engine score thresholds. Second thresholds are native to individual search engines. SD-E is described in Materials & Methods, PEP = posterior error probability.

Program	Inbuilt threshold	Primary threshold	Second threshold
Protein Prospector	-	FDR = 6%	SD-E (≥ 5)
pLINK	FDR = 5%	-	e-value (≤ 0.1)
xQuest -> xProphet	-	FDR = 6%	ID-Score (≥ 20)
Kojak -> Percolator	-	q-value ≤ 0.01	PEP (≤ 0.9)
ECL/ECL2	-	q-value ≤ 0.01	-

and an indication of low technical noise in the dataset. All 15 decoy hits had a DFscore of 1 with one exception, whose DFscore was 2.

(b) Did data appear statistically non-random?

Non-randomness was evaluated on the basis of several criteria:

Inter-protein vs. intra-protein XL: For a database of 86 proteins, random partner selection would result in a 1/86 (1.12%) chance of both tryptic peptides in a crosslinked pair arising from the same protein, assuming an equal number of tryptic peptides from each protein in the database. Experimentally, however, far more opportunities exist for efficient crosslinking within a protein than between proteins. Of the 1742 unique accession/position pairs in the dataset, 1294 (74.3%) were intra-protein, conforming to the experimental expectation rather than the random selection of peptides during bioinformatics.

Protein abundance: During MS data acquisition, ions were prioritized for sequencing on the basis of intensity (high-to-low) leading to an expectation of XL detection at a higher frequency for relatively abundant proteins. Consistent with this, the dataset was dominated by XL between the abundant virion structural proteins P4a and P4b. This provided a clear validation of data on the basis of known protein abundance.

Non-random lysine occupancy per protein: If search engines were picking lysine XL sites randomly, then the proportion of lysines occupied with XL would be expected to be fairly constant from protein-to-protein. However, lysine occupancy on a per protein basis covered a broad range, from 32.5% to 100% (Fig 2a). Search engines were therefore not simply picking sites from the database randomly. Some proteins were clearly more ‘detectably crosslinkable’ than others for reasons that presumably included protein abundance, solvent accessibility and lysine basicity for reaction with succinimide-based crosslinkers.

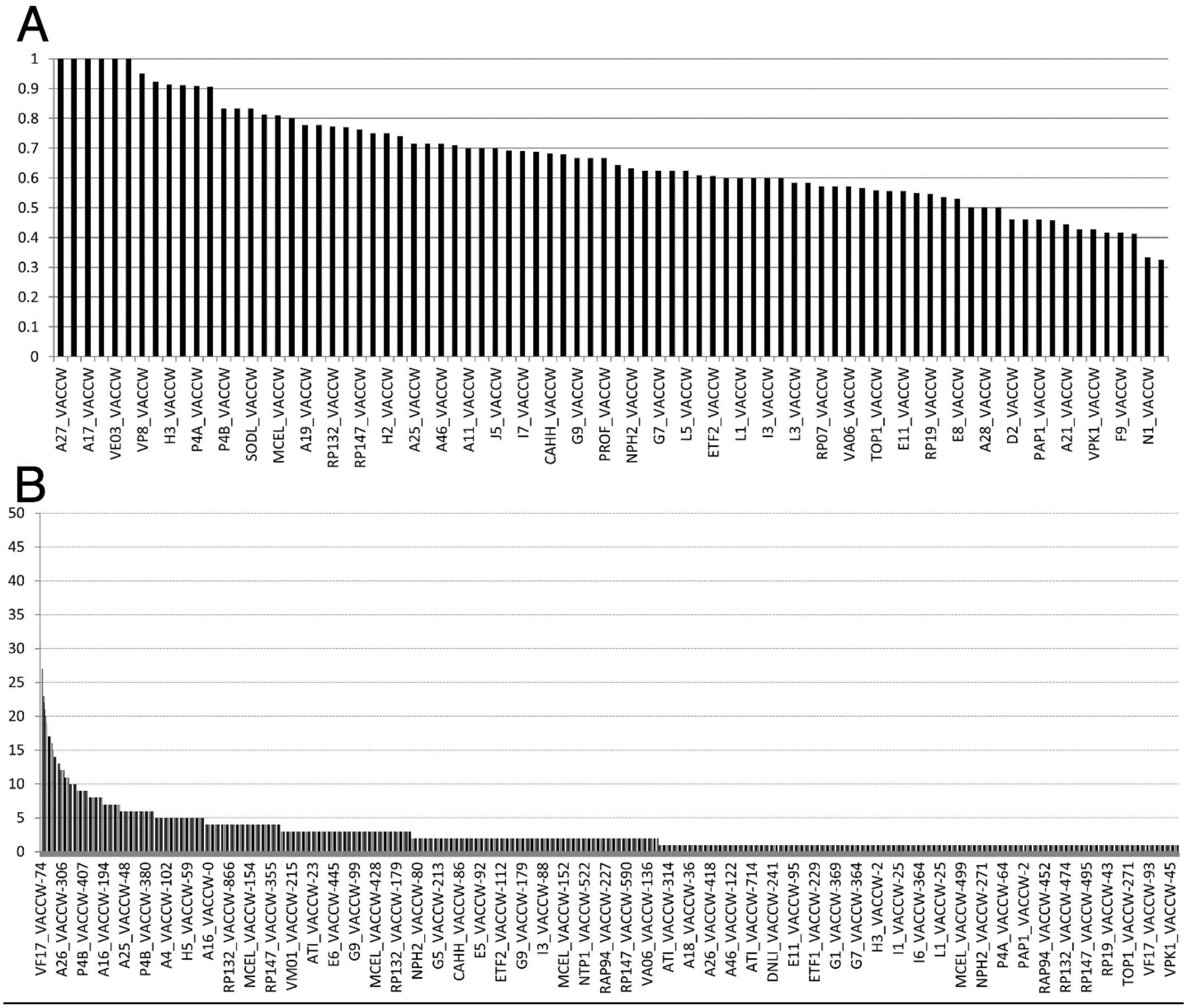


Figure 4.2. Lysine XL sites are non-random. (a) Bar chart showing the extent of lysine occupancy with XL, per protein. X axis: All proteins within which lysine XL were detected (every alternate accession is named). Y axis: Proportion of the protein's lysines found to be crosslinked in the dataset, which ranged from 32.5% to 100%. (b) Bar chart showing the popularity of each unique pair of crosslinking site in the dataset. X axis: Individual lysines in the dataset (there are 1742 bars in total, with only every ~31th bar labeled). Y axis: Number of unique crosslinking sites to which it was attached, ranging from 45 (left) to 1 (right).

Non-random ‘hotspotting’ of lysine XL sites within a protein: Individual XL sites within a protein may vary in exposure, reactivity or flexibility or the number of reactive partners within crosslinking range, resulting in the appearance of crosslinking ‘hotspots’ [519]. The crosslinkability of some protein N-termini in particular (Appendix 2.Fig1) likely arises from their exposure and flexibility, combined with a pKa [520] that promotes chemical reactivity. Consistent with this, individual lysines in our dataset showed substantial variation in predisposition towards XL ‘hotspotting’ (Fig 2b). F17 residue K74, for example, provided a particularly concentrated crosslinking hotspot, appearing in a total of 45 distinct accession/position pairs (Fig 2b) among 15 protein partners (Appendix 2.Fig1). By contrast, many other positions in various accessions appeared just once (Fig 2b, Appendix 2.Fig1).

Non-random coverage of inter-protein XL space: Our 86-protein search database provided a theoretical space of 3655 potential protein-protein pairs from which the XL dataset contained just 449. Despite the depth of analysis (4609 XL ions), this 12.3% coverage of theoretical interprotein crosslinking space suggested a level of specificity.

(c) Did data appear structurally rational, using PDB co-ordinates for known virion protein structure?

At the time of writing, partial or complete X-ray crystallographic structures covered the crosslinked portions of 12 proteins in our XL dataset, with an additional two crystallographic structures from other orthopoxviruses. All possible lysine-lysine through-space (Euclidian) and solvent-accessible surface (SAS) distances within all of these structures [521] were binned, and the resulting two histograms were found to be centered at ~43 and ~54 Å, respectively (Fig 3). By contrast, the Euclidian/SAS distance histograms for all experimental XL found within the 14 proteins was centered at 14.9 and 13.5 Å respectively, with 103 or 114 (SAS/Euclidian) out of the 136 experimental XL distances being structurally rational (32 Å, Cα to Cα distance). Based on the Kolmogorov-Smirnov test, the probability that the "All lys-lys" and "experimental XL" distance histograms (Fig 3) were sampled from a single population was < 10−4, providing 99.99% statistical confidence that the crosslinking dataset was structurally rational.

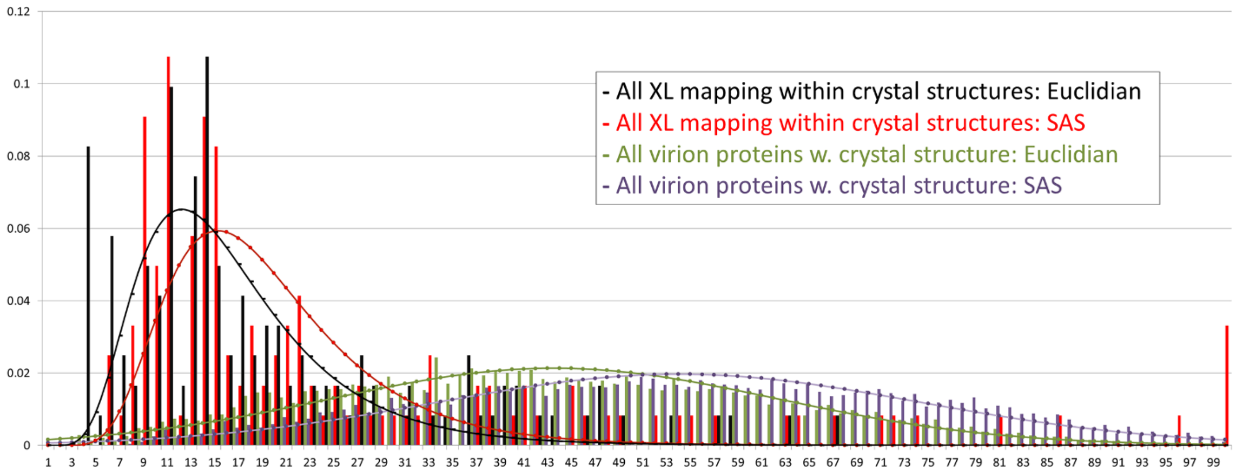


Figure 4.3. Histograms (bars), and fits (lines) to the histograms, showing all lysine-lysine Cα –Cα Euclidian (green) and SAS (purple) distances from within all virion proteins whose crystal structures have been reported to date. Superimposed are Euclidian (black) and SAS (red) distances between those crosslinked residues in our dataset that mapped within these structures (overlay = log-normal distribution). X: Distance (Å). Y: Proportion of total (ie. of all values summed for a histogram or distribution) found within an individual bar or point. XL lengths < 32 Å were considered structurally rational for crosslinkers DSS/BS3. For the longer crosslinkers BSPEG5/9, no XL exceeded an SAS distance of 30 Å. For zero-length XL (EDC), Cα –Cα SAS distances for crosslinked residues of < 18.6 Å were considered structurally rational. The few violators of these restraints likely represent inter-subunit XL within homomultimers as opposed to intra-protein XL. Distances for vaccinia Profilin were based on a homology model of the monkeypox virus ortholog (PROF_MONPZ).

(d) Were data biologically rational, regarding known protein functions and the biology of the virus?

Further assessment of the XL dataset was largely biological, namely, whether the identities of crosslinked protein pairs were consistent with known protein functions. For this analysis, accessions with strong functional annotations were collected into groups (Table 3). Interactions within any group were considered ‘biologically rational’, while the pairing of a

Table 4.3. Six functional groups covering 68 vaccinia virion accessions: ‘7PC’ (seven protein complex), DNA, ‘membrane’ (MV transmembrane and membrane-associated proteins),’ structural’, ‘thiol’ (redox plus an additional glutaredoxin), transcriptosome (mRNA biogenesis). ‘#mem’ = number of members in each group, ‘#comb’ = number of pairwise combinations within a group according to n!/k!(n-k)! (subset of k distinct elements from an n-element set). There are 450 theoretical pairs of ‘membrane’ with ‘transcriptosome’-group proteins (these pairs being designated ‘non-rational’). Some accessions were reassigned during the study (eg. VP8 away from ‘DNA’). The ‘Membrane’ group was chosen to represent all MV proteins with detectable transmembrane domains plus MV proteins considered to be membrane-associated (A26, A27). Since WV-specific proteins were not considered in the current study, VENV (F13L), a membrane-associated WV-specific protein, was included in the search DB in error. A11 (a ‘VMAP’ [522]) is also considered to be not packaged in vaccinia MV [30]. No XL at all for A14 and I2 were detected in this study. The vaccinia stub of the cowpox virus ATI is considered, here, to be a membrane protein. VP8, a virion core protein, is included in the ‘DNA’ group due to its nucleic acid binding properties [523] as opposed to a known role in conjunction with the vaccinia genome.

Group	Proteins	#mem	#comb
7PC	A30,G7,J1,A15,D2,D3,F10	7	21
DNA	DNLI,I1,I3,K4,G5,H5,I6,TOP1,VP8	9	36
Membrane	A9,A11,A13,A14,A16,A17,A21,A28,ATI,CAHH,E8,F9,F14.5,G3,G9,H2,H3,I2,I5,J5,L1,L5,O3,A26,A27,VENV	26	325
Structural	A4,P4A,P4B	3	3
Thiol REDOX	A2.5,E10,GLRX1,GLRX2	5	10
Transcriptosome	MCE,MCES,MCEL,NPH2,NTP1,PAP1,RAP94,RP07,RP18,RP19,RP22,RP30,RP35,RP132,RP147,L3,ETF1,ETF2	18	163
	TOTAL:	**68**	**558**

membrane-group protein with a transcriptosome-group protein was designated ‘biologically nonrational’ since these two groups of proteins are considered, based on controlled degradation studies [73, 309], the most likely among the various groups to occupy distinct virion compartments - separated by the core wall. All other protein-protein pairings were disregarded for the purposes of biological validation as being relatively uninterpretable. Membrane-group proteins showed a moderate, yet unmistakable global positive predilection for other membrane-group proteins as crosslinking partners, and a mild antipathy, globally, for transcriptosome proteins (Fig 4a). Transcriptosome proteins, as a class, showed a mild but unmistakable predilection for other transcriptosome proteins as crosslinking partners and a mild antipathy for the membrane class (Fig 4b). While not absolute, the trends shown in Fig 4 were consistent with accepted compartmentalization models for virion proteins, with the likely location of the transcriptosome within the virion core enclosed by a core wall, and virion TM proteins likely occupying a two-dimensional membrane compartment surrounding the core wall. This provided a suggestion of biological rationality within the XL dataset. Among the top 28 crosslinked protein pairs by DFscore, 12 were ‘rational’ and only 2 were ‘non rational’. The top 28 protein pairs contained 1205 of the 1849 total XL ions and the top 12 “Y” protein pairs represent 92% of all XL ions associated with a “Y” (ie. that were biologically ‘rational’).

Vaccinia virion crosslinkome

Virion structural proteins: Three major structural proteins of the virion, P4a (A10), P4b (A3) and A4 (p39), are thought to comprise the wall of the virion core ([30]and references therein). Among the most clearly discernible interactions in the XL dataset was a connection between the C-terminal portions of P4a and P4b (Fig 5a) providing, perhaps, the first structural

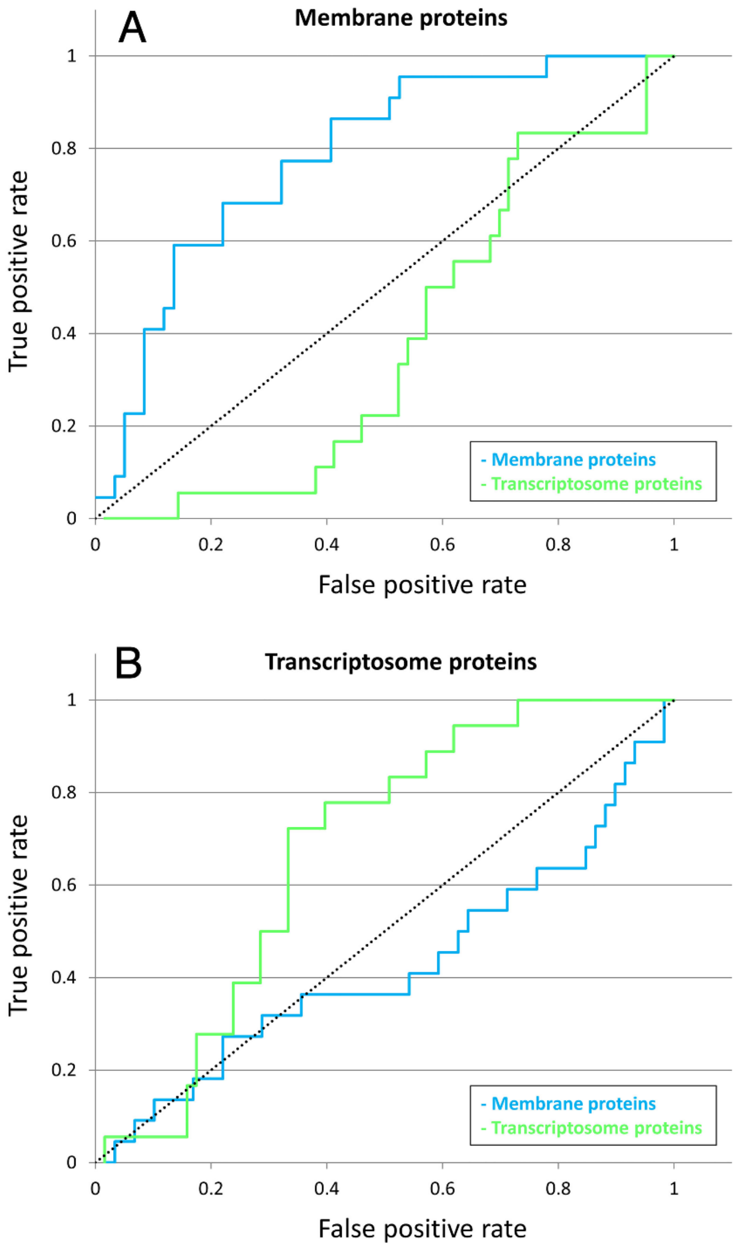


Figure 4.4. 'Biological rationality' in the dataset, as ROC ('receiver operating characteristic') curves. Proteins with strong functional annotations were divided into functional groups (Table 3) and on this basis were subjected to ROC analysis (Materials & methods). Briefly, a listing of all virion proteins with crosslinking partners was ranked by proportion of partners that were classed as: (a) membrane, (b) transcriptosome. ROC curves score, proportionately (0 to 1), positions (y) vs. not(positions) (x) in the ranking that correspond to membrane proteins (blue) or transcriptosome proteins (green). The line of no-discrimination (neutrality) is shown black, dotted. A colored line curving above the diagonal indicates a positive correlation, and vice versa.

information on the mutual arrangement of P4a and P4b in MV. Among the three fragments of P4a arising from proteolytic processing during MV maturation, P4b was most abundantly crosslinked to fragment 3 (the C-terminal proteolytic product), with an additional pair of XL connecting P4b's N-terminal region (around residue 100) to the C-terminal end of P4a fragment 1 (Fig 5a). In this manner, P4b may bring P4a fragments 1 and 3 together after P4a cleavage. The fate of P4a fragment 2 is unknown: The density of XL ions in this fragment was dramatically lower than in fragments 1 or 3 (Fig 5b) suggesting that fragment 2 is discarded or degraded after its excision during virion maturation, with the few residual detectable XL in fragment 2 perhaps representing low level contamination of MV preparations with pre-cleavage viroforms. A similar suppression of XL density was apparent for the N-terminal cleavage product of P4b (Fig 5b). This correlation of XL density with known fragments of P4a and P4b provided additional validation for the dataset as a whole.

Protein A4 interacted with P4a but not P4b (Fig 5a), consistent with prior immunoprecipitation and immunogold EM co-localization studies showing a stable interaction between P4a and A4 [212]. P4a-A4 XL were, with one exception, between the N-terminal ~half of A4 and residues 170–350 of P4a fragment 1. In immunoEM studies, antibodies to A4 decorate a region of MV between the core and outer envelope [30] or stain the surface of the exposed virion core [73], and A4 has been suggested to reside in a 'spike' or 'palisade' layer on the exterior of the core wall [92, 503]. Fig 5c shows a predicted three-domain structure for A4. Short-range intra-protein XL tended to cluster within the predicted N- and C-terminal domains with these two domains donating longer range XL to a third, central domain. The majority of inter-protein XL to A4 were within its N-terminal half (Appendix 2.Fig1),

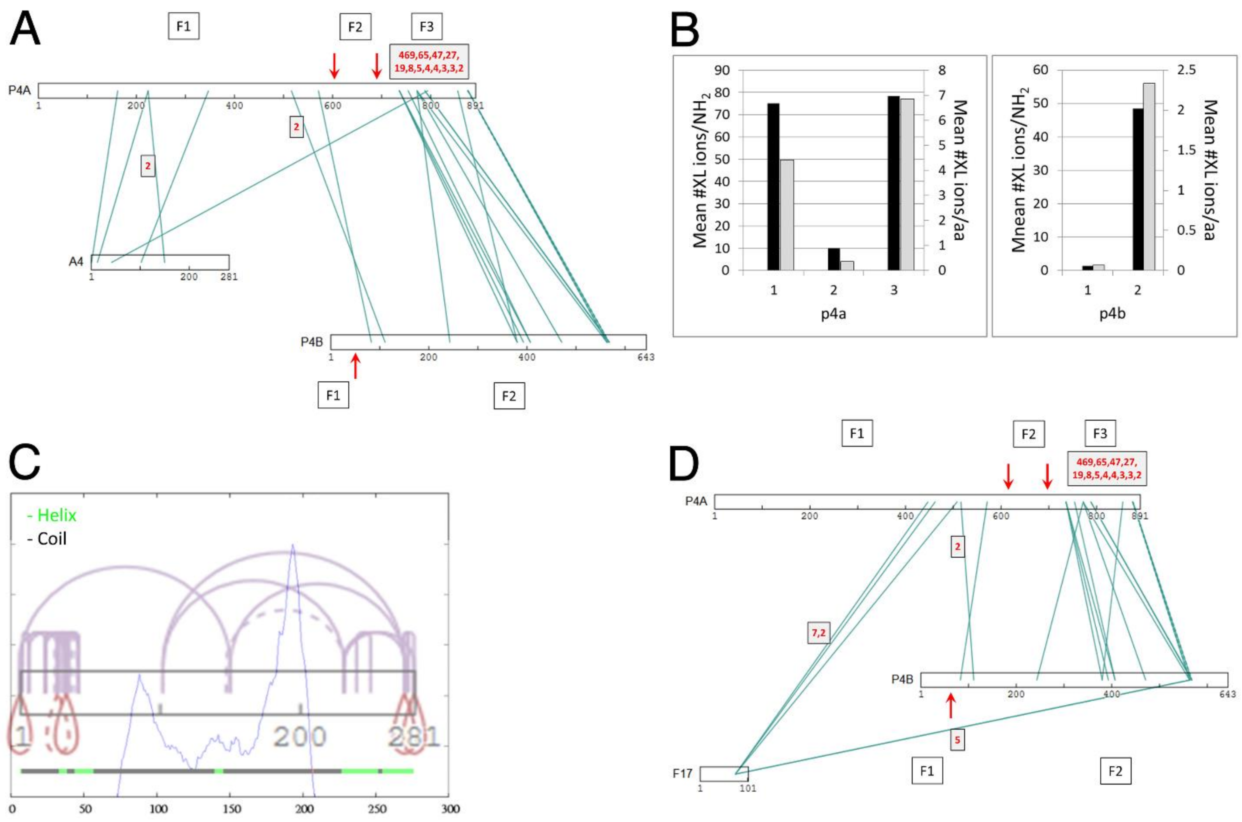


Figure 4.5. Crosslinking of Virion Structural proteins P4a, P4b, and A4. (a) XL interactions between the three major structural proteins P4a, P4b and A4 (p39). F1, F2, F3 (boxed): fragments 1, 2 and 3 of P4a, and fragments 1 and 2 of P4b generated during virion maturation by AG|-specific processing after P4a residues 614 and 697 [211] and P4b residue 61 [209] (red vertical arrows). (b) Mean # of XL ions (intra- and inter-molecular) detected per amino group (lysine sidechain and fragment N-terminus; Y1 axis, black bars) and per residue (Y2 axis, gray bars) in fragments 1, 2 and 3 of P4a (left), and fragments 1 and 2 of P4b (right), as generated by AG|-specific processing at the sites indicated in panel A. Raw numbers are given in Table C in S1 Text. (c) Predicted three-domain structure for protein A4. Lower horizontal line: Predicted helix (green) and coil (black). Blue trace: Predicted domain boundaries at residues 88 and 193 based on an endpoint density profile from 2718 PSIBLAST hits. Mauve: Intra-A4 XL. The density of XL within N- and C-terminal regions is consistent with the presence of discrete N- and C-terminal domains coincident with the green helical regions. XL from both protein termini to positions 100 and 150 (towards the N- and C-terminal ends, respectively of the short central domain located between residues 88 and 193) suggest that the central region of A4 is within crosslinking range of the two terminal domains. (d) XL interactions of structural proteins P4a and P4b with protein F17. Vertical red arrows: As in panel A. In panels A, D: Numbers (red font) in gray squares with black border: DFscore (XL with no red-font numbers had a DF score of 1).

the only exception being the strongly detected XL between A4's C-terminal region and protein F17 (below).

The short (101 amino acid) virion protein F17 formed multiple XL to a localized region of P4a fragment 1 between residues 450 and 510, just upstream of the fragment 1 interaction site of P4b (Fig 5d). All P4a XL to F17 were via a single lysine 'hotspot' of F17, namely K74, the most intensively focused XL hotspot of any found in the current study (see above). K74 may therefore form an anchor point for a number of virion proteins. F17 also formed a very strongly-detected (DFscore = 11) interaction with the C-terminus of A4 (Appendix 2.Fig1) as well as A4's N-terminal region (DFscore = 3). Due to its strong association with P4a and A4, combined with its high abundance in the virion, F17 is considered a good candidate to be a major structural protein of the core wall. F17 has been reported, by immunogold EM, to be a lateral body protein [76], an observation neither inconsistent nor mutually exclusive with its snug fit to the exterior of the core wall exterior shown here. Fig 6 shows a topological arrangement for proteins P4a, P4b, A4 and F17 that satisfies the deduced restraints.

The pairing of virion core proteins VP8 and A12. Another strongly-detected protein pairing was the 251 residue protein VP8 and the 192 residue protein A12, crosslinked via their final ~50 residues and residues 63–88, respectively (Fig 7). The association of these two proteins seems to be a new finding. VP8, a virion core protein with nucleic acid binding properties in vitro [523], is required for the production of infectious, morphologically and transcriptionally normal MV [524, 525]. It is exposed in core material only under the harshest conditions, such as during the use of a virus mutant in P4b along with DNase [73]. A12, a core protein of unknown function, is also essential for the formation of a structurally normal core [30]. Mutants in both VP8 and A12 show morphological defects in IV membrane adhesion to viroplasm during virion

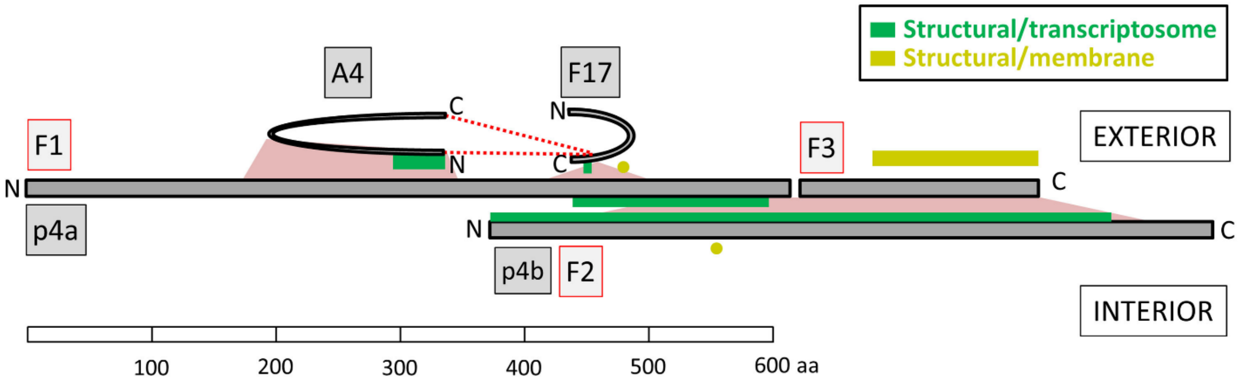


Figure 4.6. Simplest arrangement satisfying topological restraints between structural proteins P4a, P4b, A4, and F17. 'N', 'C', 'F1', 'F3', 'F2' refer, respectively, to protein N- and C-termini, proteolytic fragments 1 and 3 of P4a and fragment 2 of P4b. P4a-fragment 2 and P4b-fragment 1 have been discarded. Pink trapezoids: The P4a/P4b and P4a/A4 interaction regions (from Fig 5a). Pink wedge: The F17(K74)/P4a interaction region (from Fig 5d). Broken red lines: XL from F17(K74) to both ends of protein A4 (from Appendix 2.Fig1). The arrangement shown also emphasizes the absence of P4b interaction with either A4 or F17 and the absence of any interactions within the four-protein complex of protein A4's C-terminal half or P4a's N-terminal region (Fig 5a and 5d). Also accounted for are the three-domain structure of protein A4 (Fig 5c) and a core-wall exterior location for A4 as predicted by immunoEM studies ([92, 503]). Among the membrane proteins, interaction sites for J5, A21, H3, A16, G3, A26 and ATI cluster at the C-terminal region of P4a (yellow line) while A17 interacts further upstream, with P4a fragment 1 (yellow circle; from Appendix 2.Fig1). P4b's only interaction with a bona fide membrane protein is with A17 (yellow circle; from Appendix 2.Fig1). This membrane protein arrangement suggests that P4b may be oriented towards the interior of the core. Green lines: Regions of structural proteins that interact with transcriptosome components (from Appendix 2.Fig1). These presumably extend into the third dimension.

morphogenesis [524, 526] and, during the morphogenic transition from IV/IVN (immature virus; immature virus with nucleoid) to MV, both proteins are N-terminally processed [30, 527] via AG| -specific cleavage immediately after residues 32 and 56, respectively [103, 527, 528]. The A12 precursor may be only incompletely processed at this site [526]. A12 seems to be partially C-terminally processed also, after residue 154, with the C-terminal fragment detectable [103].

A12 showed two emphatically detected crosslinking hotspots, centered at residues 88 and 167 (Appendix 2.Fig1), located within fragments 2 and 3, respectively, of the three-fragment protein if doubly-processed. The two hotspots and the protein N-terminus were strongly connected to one another via intra-protein crosslinking (Appendix 2.Fig1) suggesting that the processed fragments of A12 remain together after proteolytic processing. The strongly detected interaction of A12's N-terminus with protein H3 (DFscore = 6, Appendix 2.Fig1), a virion-resident TM protein, also suggested that the N-terminal fragment is not discarded after A12 cleavage. A12 showed six transcriptosome partners, all crosslinked at the residue 167 hotspot (Appendix 2, Fig1). We speculate that A12 may span the core wall with fragment 3 contacting

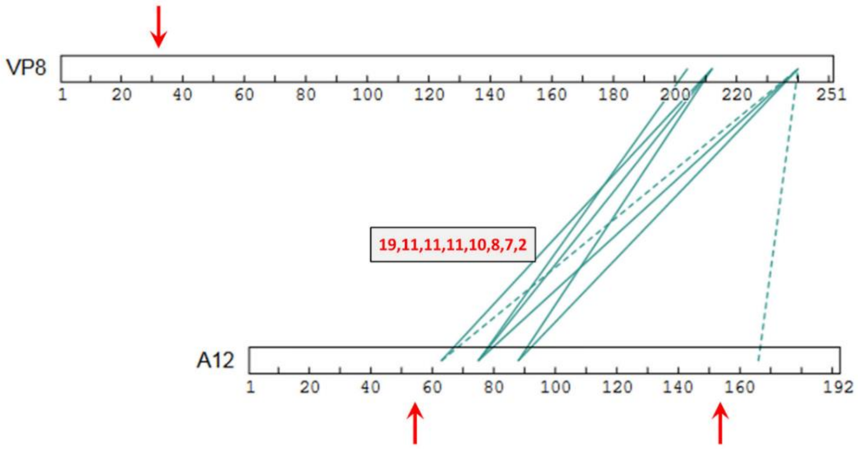


Figure 4.7. Crosslinking between core proteins VP8 and A12. (a) XL interactions between proteins VP8 and A12, showing AG|-specific processing sites (vertical red arrows). Details as in Fig 5.

transcriptosome and P4b interiorly, and fragments 1 and 2 oriented towards the exterior contacting A4, P4a, F17 and TM proteins H3, F9 and H2 (Appendix 2.Fig1). The compromised VP8 encapsidation upon repression of core wall protein P4b [529] seems consistent with a core wall connection to VP8 also, as does the compromised core wall in absence of VP8 [530]. Since, minimally, ~57 amino acids (aa) of linear beta sheet would be required cross a 20 nm core wall, contacts within A12 of the C-terminal 20% of VP8 and some membrane proteins, appear to be more intimate than a wall's-width. We suppose that this apparent intermingling of the VP8 C-terminal region with membrane proteins as well as A12 could be outside, within, or immediately interior to the wall.

A12 also showed connectivity with G1 metalloprotease (Appendix 2.Fig1) but not the I7 cysteine protease, the implications of which are unclear. The finding of protein A19 as a crosslinking partner for A12 (at the central hotspot of A12, Appendix 2.Fig1) was consistent with the A12-A19 interaction detected by yeast two-hybrid (Y2H) analysis [507]. Among VP8's 12 crosslinking partners (Appendix 2.Fig1) was the short (65 aa) F8 protein, of unknown function, supporting the finding of an association of VP8(L4) with F8 by Y2H analysis [507]. F8's XL fell close to the VP8 C-terminus.

A17-H3-A27-A26 membrane protein network. Another clearly discernible virion protein sub-network connected membrane proteins A17, H3, A27 and A26 (Fig 8a), involving the N-terminus of A17, a central region of H3, residues ~300 to 420 of the 500 residue A26 protein and a large portion of the 110 residue A27 protein (Fig 8a). To understand this sub-network requires some consideration of the known functions and properties of the four proteins. Thus, A17, a 203 aa protein, is one of two key proteins acting at the earliest stages of virion morphogenesis (the 'crescent' and IV stages), the other being its partner, the 90 aa protein A14

[30]. During normal infection, A14 and A17 co-localize to ER and ERGIC membranes as well as the earliest assembly structures ('crescents') and IV [121, 522, 531-534], with crescents reportedly forming via the accretion of A17-containing vesicular elements [533]. Unfortunately, no XL were detected in A14, consistent with its three crosslinkable lysines being positioned such that XL to any of them would yield a long tryptic peptide (in the top 4th percentile of peptide lengths for the project). Repression of the gene for A17 leads to a blockade in virion morphogenesis at a very early stage, with membrane tubulovesicular elements accumulating at the periphery of electron-dense virosomes/viroplasm [30, 522]. A17 has four TM domains [522] and appears to use them in a 'reticulon'-like manner to induce membrane curvature [535]. A17's N- and C-termini, which are trimmed in vivo (at residues 17–20 and 185, respectively) by I7 proteinase [103, 201, 536] are both thought to be cytoplasmic. Evidence for this includes their exposure after in vitro expression in the canine microsomal system [531, 532, 537] and, in intact MV, accessibility to antibodies of the N-terminal 60 residue region (prior to the first membrane-spanning region starting at residue 61) [30, 538]. A17 forms disulfide bonded homodimers via Cys178 in the C-terminal tail [531], and the A17 N-terminal region interacts with D13 trimers that assemble to form the honeycomb lattice of the IV external scaffold [67, 189, 533]. Virus is excised from the scaffold in an I7 proteinase-dependent manner [203].

Protein H3 is a heparin sulfate-binding attachment protein. Although not essential for virus replication, it is required for normal plaque size and virus yield [112, 117]. It is immunodominant [539, 540], localizes to the MV surface and can be extracted therefrom with NP40 in the absence of disulfide reducing agent [119]. H3 does not seem to follow a classical protein secretory pathway, but instead seems to be post-translationally anchored to virion membranes [119], via a TM helix that is predicted to lie towards the protein C-terminus (residues

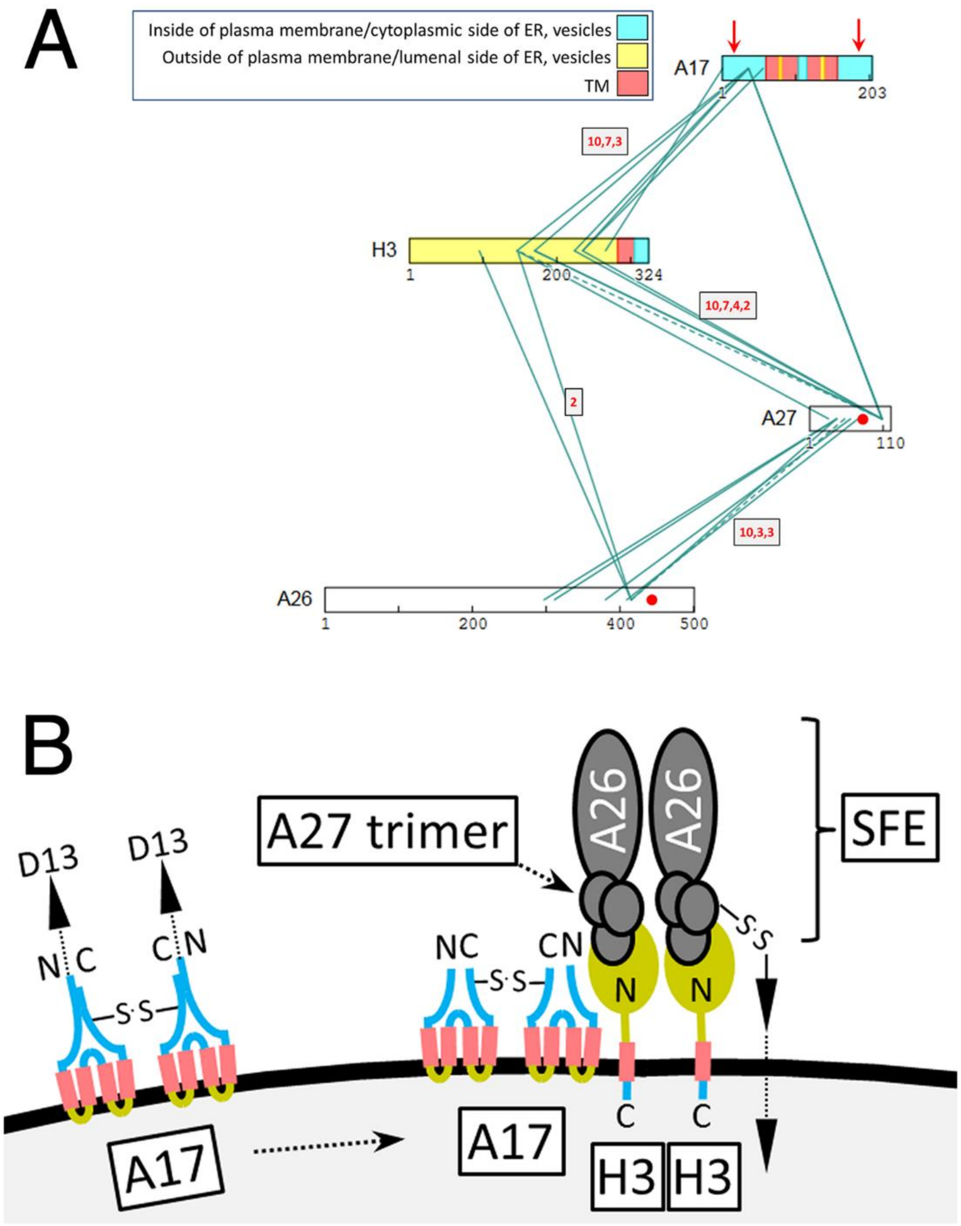


Figure 4.8. XL interactions between major membrane proteins. (a) Crosslinkome for major membrane proteins A17, H3, A27, A26. Yellow, red and cyan fill: ‘Outside’, TM and ‘inside’ domains, respectively, with “Inside” and “Outside” following the convention of the program TMHMM [493] in which “inside” refers to the cytoplasmic side of plasma membrane, ER membrane or vesicles for a classically embedded TM protein, and “outside” refers to the lumenal side or the ER or external side of the plasma membrane. Vertical red arrows: AG|-specific processing sites in protein A17. Red spots (A26, A27): Cysteines that are disulfide bonded to one another. Other details as in Fig 5. (b) Arrangement, at the MV envelope, of proteins shown in panel A satisfying XL, functional and imaging data. Left side: Disulfide bonded homodimers of unprocessed A17 in the IV envelope in reticulon conformation. The unprocessed A17 N-terminus is within crosslinking range of A17’s C-terminal region. The D13 external scaffold is not depicted. Right side: Maturation of the envelope includes A17 N- and C-terminal cleavage

and D13 loss. H3 is now non-canonically tail-anchored in the MV envelope with immunodominant N-terminal domain exposed, and SFEs have been added as stacked trimers of A27 and associated A26. H3's external N-terminal domain mediates most of the interaction between A27 and the processed A17 N-terminal region, with the A17 C-terminus now out of range of H3. The H3/A26 and A27/A17 protein pairs are sufficiently proximal within the A17, H3, A27, A27 network to allow some modest direct crosslinking across their interfaces. SFEs are anchored either directly or indirectly to the core wall beneath the MV envelope by disulfide bonding (arrowed). Proteins CAHH and ATI (not depicted) are candidates for mediating such disulfides. Disulfide bonding does not involve H3 which is detergent-extractable in the absence of disulfide reduction. A27 and A26, although disulfide bonded to one another, interact tightly enough that SFEs show stability in the presence of NP40 plus disulfide reducing agent. 'N' and 'C' denote protein N- and C-termini respectively. TM protein domains are colored according to panel A and Appendix 2.Fig1. Although H3 was designated 'N-outside' by prediction program TMHMM (opposite polarity to A17), rendering the major N-terminal domain yellow, such predictions are presumably invalid for unconventionally added (tail-anchored) TM proteins (see text). Albeit two copies of the H3-A27 trimer-A26 complex are shown to suggest SFE topology, more accurate modeling would require an understanding of protein stoichiometries.

283–305 of the 324 aa protein; [493]) such that the N-terminal region of the protein is cytoplasmically oriented and exposed on the virion surface [119, 541]. H3 is added to MV within the virus factory late during maturation [119] coinciding with the replacement of IV's D13 external scaffold with an antigenically [542] and morphologically [543] distinct surface structure (see below).

XL involving A17 included a very clear connection between the N-terminal domains of A17 and protein H3 (Fig 8a, Appendix 2.Fig1). Albeit both the N- and C-terminal regions of A17 are exposed to the outside of MV [531, 532, 537], XL with H3 were detected only for A17's N-terminal region, with none at all to A17's C-terminal region (Fig 8a). These H3 XL were to the region of A17 remaining after proteolytic removal of A17's extreme N-terminus during maturation. Parenthetically, additional, unrelated XL were detected to the extreme N-terminus of A17 (Appendix 2.Fig1) suggesting the presence of pre-processed viroforms in the MV preparation. Interestingly, one of these XL was to the C-terminal region of A17 itself (residue 180, Appendix 2.Fig1), suggesting juxtaposed A17 termini in the pre-processed form. This in turn suggests that during virus maturation, upon cleavage, and loss of the D13 exoskeleton, A17 may undergo a reconfiguration: Prior to cleavage the unprocessed N- and C-termini mutually interact but after cleavage the only the processed N-terminus can interact with H3 (see Discussion). Finally, a strongly detected (DFscore = 5) inter-subunit XL in the C-terminal region of A17 (between residue 180 and residue 180, Appendix 2.Fig1) was consistent with previously reports of A17 homodimer formation [531]. No XL were detected C-terminal to the A17 C-terminal cleavage site at residue 185.

A27 is an immunodominant [544] disulfide-bonded trimer [486, 487] that can stack into hexamers and higher order multimers in vitro [128, 545]. It functions in virus attachment to cell surface glycosaminoglycans [483], mediating a choice between cell entry pathways [164]. It also functions in the microtubule-dependent transport of MV within the cell [546], the secondary wrapping of MV with Golgi-derived membranes late in infection to form wrapped virus (WV) [547, 548] and in cell-cell fusion [486]. A27 lacks detectable TM domains [493] but is reportedly anchored to the virion membrane via interaction with A17 [536, 545]. A27 can be removed from the MV surface by disulfide reduction [73]. Like H3, A27 is detected more strongly in MV than IV suggesting that it is added to virions during the IV to MV transition [119, 502], maybe at the same time as H3. In the current study, the A17-A27 interaction (above) was represented by just a single direct XL which, at A17 residue 36, occurred within the N-terminal “high affinity” region (32–36) noted in in vitro interaction studies [545]. Far more readily detectable, however, was the crosslinking of both proteins to H3 (Fig 8a) suggesting that H3 mediates a substantial portion of the A17-A27 interaction in MV.

The non-essential A26 protein mediates virus attachment to cell surface laminin [109] and the embedding of cowpox virions in A-type inclusions ([549]; discussed below). A26 is absent from wrapped extracellular virus (EV) suggesting that A26 mediates a choice between wrapping and inclusion formation. Like A27, A26 contains no TM domain [493]. Instead, A26 is anchored to the MV membrane via disulfiding of cysteines 441 and 442 in its C-terminal coiled-coil region with Cys71 and Cys72 towards the C-terminus of A27 [125]. Via these interactions A26 and A27 are tethered to one another and to protein A17 on the virion surface [550]. Here, A26-A27 XL were detected abundantly, at sites in both proteins immediately N-terminal to the abovementioned cysteines (Fig 8a). As with the A17-A27 interaction, some mediation of the A26-A27 interaction by H3 was suggested by the crosslinking pattern (Fig 8a).

The replacement of IV's D13 external scaffold, during late-stage maturation, with a distinct surface protein structure over the MV lipid envelope (above) is supported by substantial evidence. Firstly, a two-domain exterior boundary is visible in thin sections of MV [30]. Second, AFM imaging under ambient conditions in the absence of any virion pre-treatment shows a surface topography described as resembling "surface fibrous elements" (SFEs) [80]. Deep-etch electron microscopy (DEEM) [88, 189], which involves neither fixation nor negative staining, evokes comparable descriptions of the MV surface (disorganized, close-packed, parallel rows of short "railroad tracks" [189]). These patterns also reflect the "Mulberry-like" MV surface features imaged by high-contrast negative staining as described throughout the literature [30, 84, 88, 89, 551-554]. SFEs can be detached from the virion surface via the action of disulfide reducing agent in the presence of NP40 (they remain nominally intact in the presence of both reagents), and appear compellingly similar when imaged by either EM [554] or AFM [80]. They have been described as chain-like, globular protein fibers of uniform size (20 nm diameter x 100–150 nm length) with no obviously hollow interior or helicity [80]. Overall, the late adherence of H3 and A27 to the MV envelope (above) coinciding with the appearance of the mature surface topology, in combination with the crosslinking pattern of Fig 8a suggests that SFEs, comprising or containing the A27/A26 complex, are brought to the MV surface via the late tail-anchoring of H3, and that H3 mediates the interaction of SFEs with A17 already present at the MV envelope. This scheme is depicted in Fig 8b. Although H3 was designated 'N-outside' by prediction program TMHMM, which would polarize H3 with its major N-terminal domain in the ER lumen in the conventional secretory pathway and thence on the inside of the MV envelope, the prediction probability was little better than evens (62% [493]) and moreover, such predictions are presumably invalid for unconventionally added (tail-anchored) TM proteins.

The contrasting crosslinking patterns of proteins H3 and L1. Protein H3 showed 119 distinct contacts with other proteins (Appendix 2.Fig1). This was a remarkable number, suggesting a high degree of connectivity for H3 in comparison with other proteins considered functionally comparable and/or that may be in the same compartment such as L1 (below). Quite dramatically, no inter-protein XL were detected beyond residue 266 of the 324 residue H3 protein, thereby restricting all inter-protein XL to H3's 282 residue N-terminal 'outside' domain. The diversity of contacts suggested that protein H3 may sample multiple, complex environments. H3 also showed an unusually high signal for homomultimer formation (Appendix 2.Fig2A, red loops). Since the published crystal structure for H3, covering residues 1–237 did not show a homomultimer [115], we speculate that either the homomultimer interface is to the C-terminal side of the crystalized region, or H3 forms a mixed multimer (eg. A2B2-type) or H3 crosslinking to itself results from a very dense packing of monomers within virion membranes. Homomultimer formation may be consistent with H3 nucleating SFE formation (above).

Somewhat surprising, and in contrast to the 119 distinct inter-protein contacts involving TM protein H3, was the absence of TM protein interactions observed for the myristoylated [555] immunodominant TM protein L1, whose only detected inter-protein contact was with protein A12 (Appendix 2.Fig1). Like H3, L1 appears on the MV surface later during virion maturation, after departure of the D13 external scaffold, via a C-terminal anchoring domain [556]. Yet L1 appears to be remarkably isolated on the MV surface from other TM proteins and virion proteins in general. Alternatively, there may be a greater difficulty in detecting L1 XL due to, speculatively, a differential abundance of H3 and L1 in MV.

All detected XL between TM proteins. For the majority of virion proteins with predicted TM domains, intra-protein XL were confined to either one or both sides of the predicted TM domain and not within or across it (Appendix 2.Fig2A). This pattern supported independent predictions of TM domain locations within TM proteins and suggested a propensity for crosslinkers to not act across lipid bilayers. Fig 9a shows all detected XL between TM proteins in the 'membrane' protein group (Table 3). TM proteins could be divided into two subsets whose major portions were classified as either 'outside' or 'inside' [493, 557] (Fig 9a, upper and lower regions respectively). Inter-TM protein XL appeared to involve only the major portion of each TM protein (Fig 9a), with extensive inter-protein crosslinking observed within the 'outside' subset, and also between the 'outside' and 'inside' subsets. A major contributor to the latter class was the cluster of A17-H3 XL discussed above. Other clear contacts at the interface of the two subsets mainly involved proteins H3 and ATI (ATI has been referred to by others as 'A25': In our notation, 'A25' refers to protein A2.5), and included the following connections: A28-H3, O3-H3, L5-H3, J5-F9, J5-ATI, F14.5-ATI G3-ATI and A21-CAHH (CAHH has been referred to by others as 'D8'). The only direct XL observed within the 'inside' subset were a very highly detected XL between proteins L5 and G3 and a contact between A17 and A13 (Fig 9a, lower region).

As mentioned above, late in infection, for the purpose of virus dissemination, cowpox virus forms A-type inclusions by the coalescence of the non-essential cowpox virus ATI protein followed by virus embedding in the resulting inclusion via protein A26. Vaccinia ATI is a C-terminally truncated version of the cowpox virus protein which cannot form inclusions. Vaccinia ATI is included in the TM protein set because of its predicted possession of a TM domain with 80% probability, with a 77% probability of 'N-inside' polarity [493] (Appendix 2.Fig2B). The predicted TM domain, located between residues 139 and 161, is flanked to the C-terminal side by two minor ones (20% probability, Appendix 2.Fig2B). Since vaccinia ATI

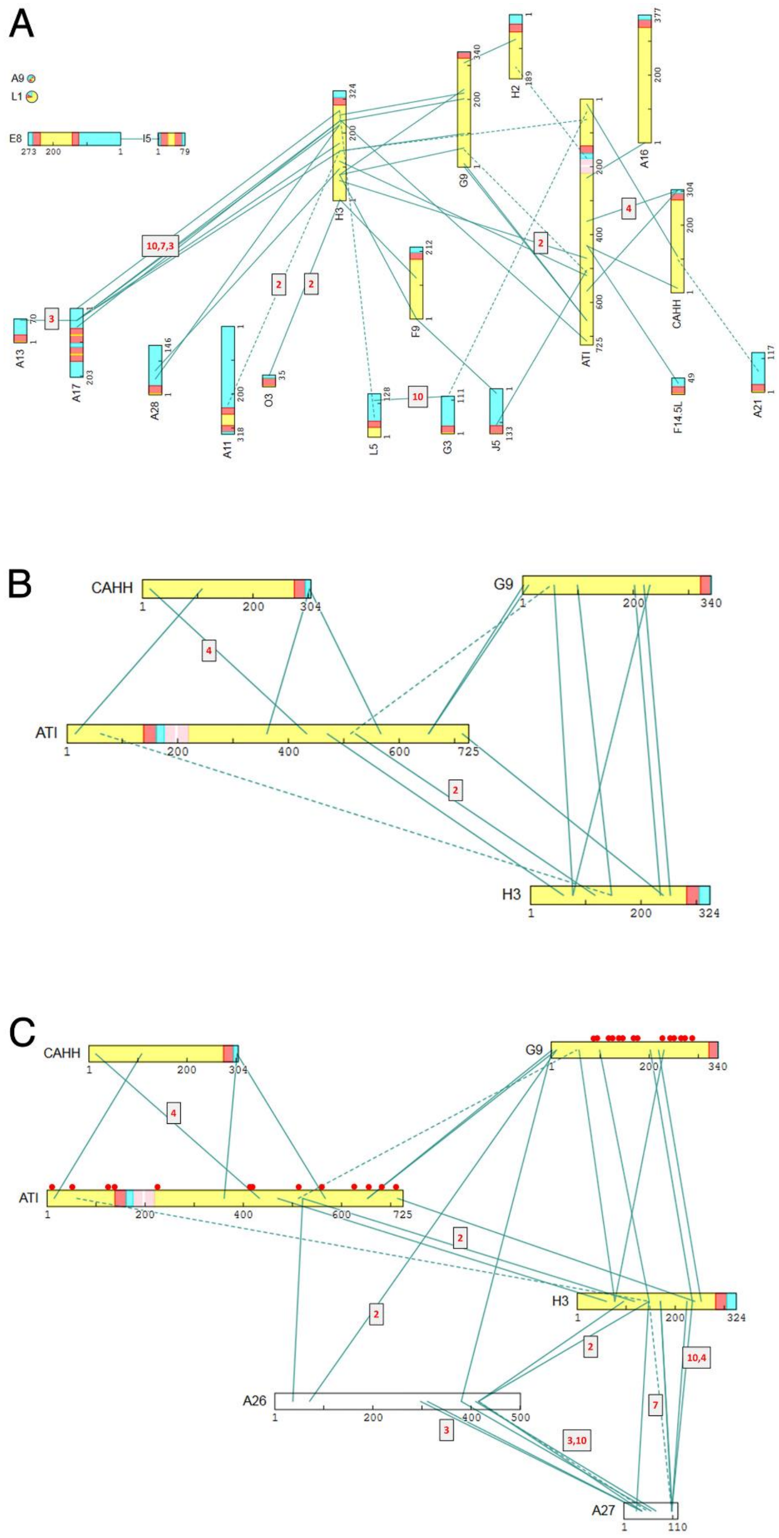


Figure 4.9. XL interactions between TM proteins. (a) All XL detected between virion TM proteins. Coloration as in Fig 8. Upper area: Proteins with XL to predicted 'outside' domains. Lower area: Proteins with XL primarily to predicted 'inside' domains. No XL were detected between protein A9 and L1 (upper left) and other members of the "Membrane" group, no XL at

all were detected for A14 or I2 (not depicted) and protein A26, A27 and VENV(F13) are considered to be membrane-associated only. The remaining 21 TM proteins of the "Membrane" group are shown. 'Minor' predicted TM domains of ATI are colored pink. (b) 'Major outside protein' sub-network (proteins ATI, CAHH, G9, H3). Domain coloration as in panel A. (c) Six-member 'attachment protein subnetwork' ('Major outside protein' plus proteins A26, A27). Domains are colored as in panel A. Red circles: Positions of cysteines within the cysteine-rich proteins ATI and G9. Other details as in Fig 5.

stood out among TM proteins in terms of the many (10 or more) XL spanning the predicted TM domain(s) (Appendix 2.Fig2A), we designated both the C-terminal and the N-terminal region as 'outside' (Appendix 2.Fig2C, Fig 9). This topology may arise if vaccinia ATI were either not a TM protein at all (it may have no known functional requirement to be one), or if it were double-spanning (via the major one noted above plus one or both of the minor predicted TM domains), or if it were membrane-anchored via the multiple (six) acylation/myristoylation sites previously noted within the C-terminal half of the protein [558].

Within the 'outside' subset of TM proteins(above), multiple XL were strongly-detected between ATI and CAHH, H3 and G9 (Fig 9a and 9b). Into this 'major outside protein' subnetwork could be plugged the membrane-associated proteins A26 and A27 to form a clear 6-member 'attachment protein sub-network' (Fig 9c). Of MV's four known attachment proteins, H3, CAHH, A27 and A26 (the first three binding to cell surface glycosaminoglycans (GAGs) [112, 113, 483] and the latter binding extracellular matrix laminin [109]), all four were present in the 'attachment protein sub-network', supplemented by ATI and G9 (Fig 9c). The latter two may play a scaffolding role: Both are acylated/myristoylated and cysteine-rich (13 and 14 cysteines respectively), providing ample potential disulfide anchoring points to the virion infrastructure. Within this sub-network, only CAHH appeared to be disconnected from a direct interaction with H3, whose 'outside' domain otherwise formed a 'hub' for the 6-protein sub-network (Fig 9c). This was consistent with the role of H3's outside domain as a hub for virion membrane proteins more generally (Fig 9a). In terms of previously known interactions among the six proteins: The cowpox virus ATI-A26 interaction has been demonstrated to require the 100–300 aa region of cowpox virus ATI [549]. A corresponding XL was detected, here, between positions 37 and 521 of A26 and vaccinia ATI, respectively (Fig 9c, Appendix 2.Fig1). By co-IP analysis, A26 has been reported to interact with G9 and A16 [163]. The A26-G9 interaction was manifest here via the N-terminus of G9 (Fig 9c, Appendix 2.Fig1). Of the five known TM protein substrates for the vaccinia-encoded redox system namely L1, A28, A21, L5, and H2 [30, 118], all except H2 and orphan L1 protein (above) appeared in the 'outside' subset.

Two previously reported complexes.

EFC: A virion 'entry-fusion complex' (EFC) has been proposed comprising nine central components (proteins A16, A21, A28, G3, G9, H2, J5, L5 and O3) [143]. MV with mutations in these proteins are morphologically normal and transcriptionally active, and can undergo normal membrane wrapping and export from the cell. They can also bind the cell but are defective in penetration [30]. Loss of any one EFC component does not appear to affect the incorporation of others [30]. Two additional EFC-associated proteins (L1 and F9) are also required for cell entry by the virus but are not required for assembly or stability of the core EFC complex [143]. Fig 10 shows the 11 proteins and their detected crosslinking partners. Of three previously reported EFC-EFC protein interactions [143] one (G3-L5) was confirmed by crosslinking (with a high DF score, Fig 10). Crosslinking led to the detection of two additional EFC-EFC protein interactions, namely F9-J5 and G9-H2 (Fig 10) in which the N-terminus of F9 connected directly to the N-terminal region of J5, and the C-terminal region of G9 connected directly to H2. At the current detection depth, other EFC members' connections appeared to be mediated by third-party proteins.

Protein ATI appeared to mediate the connection of five EFC members, namely A16 (N-terminal region), H2 (C-terminal region), G3 (C-terminal region), G9 (N-terminal region) and J5 (C-terminal region). TM protein H3 also appeared to mediate the interaction of five EFC members, namely A28, F9, G9, L5 and O3, three of which (A28, F9, G9), showed multiple connections to H3 (Appendix 2.Fig1) and may therefore have a more intimate H3 connection. Interestingly, EFC protein G9 was common to both the ATI and H3 sets, with ATI crosslinking to G9's N-terminal region, and H3 interacting across a broad swath of the 340 aa G9 protein.

Overall, the EFC may fall into two parts (distinguished in Fig 10 via a red ring), namely, proteins from the 'outside' subset of Fig 9a (A16, F9, H2, G9 plus L1) and those in the inside subset (O3, A28, L5, G3, J5, A21), the former perhaps being tail-anchored and added late during virion maturation.

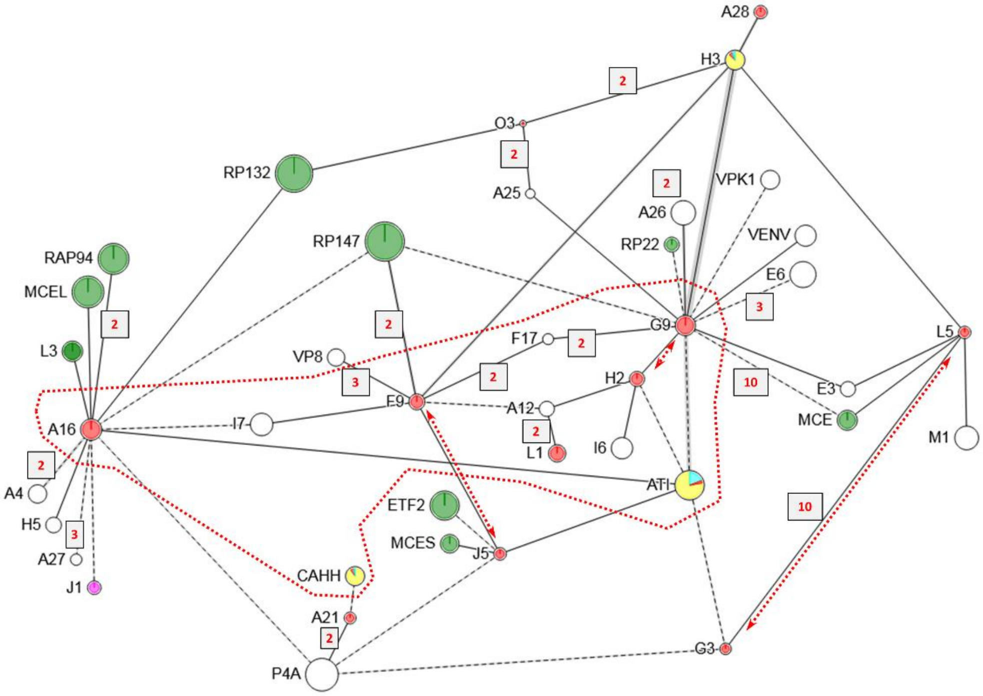


Figure 4.10. XL among EFC proteins and their detected partners. Proteins are depicted as circles whose areas correspond to chain length. Red fill: EFC proteins. Green fill: Proteins of mRNA biogenesis. TM proteins are multi-colored as described under Fig 8. Boxed numbers (red font): XL with DFscore > 1. Red dotted perimeter separates EFC proteins falling in the 'outside subset' and the 'inside subset' (inside and outside the perimeter, respectively). Double-ended red arrows: Three direct EFC-EFC protein interactions detected by crosslinking. Other details as in Fig 5.

7PC: Another reported complex in the virion is the ‘seven protein complex’ (7PC) [192] mutations in whose members (A15, A30, G7, J1, D2, D3 and VPK2) have similar phenotypes relating to the association of viroplasm with growing crescents during early virion morphogenesis, and the appearance of vestigial IV. For example, A30 mutants show enlarged virosomes and empty IV or pseudo-IV with multiple membrane wrappings [559, 560]; G7 is required for the movement of crescents to the periphery of virosomes and the filling of crescents with viroplasm [193, 561]; the repression of J1 phenocopies the repression of A30 and G7 [195, 562], and A15 mutants show characteristic empty IV [192]. Association of the seven proteins in a common complex has been deduced by mutual pullouts from infected cell extract with epitope tagged A15, VPK2, D2 and D3 followed by immunoblotting for the other complex members [192]. Regarding direct interactions, A30 was shown to interact directly with G7 [193], proteins A30 and G7 both become unstable in the absence of J1 at the restrictive temperature [195, 562], and J1 has been shown to self-interact [507, 562].

Fig 11 shows 7PC members and their crosslinking partners. No crosslinking partners were detected for protein D2. The remaining six proteins could be linked via three direct XL (G7-VPK2, G7-A30 and G7-J1) and two mediated contacts, namely via the extreme N-terminus of protein A4 and N-terminal region (aa 54–75) of RP147. Protein G7 appeared as a ‘hub’, directly linking three other 7PC members (VPK2, A30 and J1), and perhaps interfacing an A15/VPK2/A30 sub-complex (nucleated at an N-terminal crosslinking hotspot of G7; Fig 11 left side) with a J1/D3 sub-complex contacting the C-terminal half of G7 (Fig 11 right side). Within the A15/VPK2/A30 sub-complex, structural protein A4 crosslinked to both A15 and VPK2. In addition, P4a fragment 3 crosslinked to G7 and A30. Consistent with the latter observation, P4a mutants show an aberrant and ‘empty’ IV phenotype [213] reminiscent of the 7PC mutants themselves. Contacts with the transcription/mRNA biogenesis apparatus appear to cluster around the J1/D3 sub-complex within which D3 appears to connect to both subunits of the PAP heterodimer (PAP1/MCE, Fig 11). With D3 also contacting protein P4b (which appears to be located predominantly at the interior face of the core wall, see Fig 6 and associated text) it is interesting to consider 7PC as perhaps a core wall-spanning complex. Regarding the J1/D3 sub-complex, the N-terminal region of TM protein H3 crosslinked to sites in J1 and G7 very close to the sites at which these two proteins crosslinked to one another.

Virion proteins of DNA binding/metabolism. Vaccinia DNA ligase (DNLI), a nick sealing protein, showed 22 crosslinking partners (Appendix 2.Fig1)—an unexpectedly large number for a specialized enzyme. Among these were four other members of the DNA binding/DNA metabolism group (Table 3), namely K4—the vaccinia DNA nicking enzyme for genome telomeres [563], I1 –a vaccinia telomere binding protein, I3 –an ssDNA binding protein [564, 565] and vaccinia topoisomerase TOP1. Moreover, direct crosslinking was

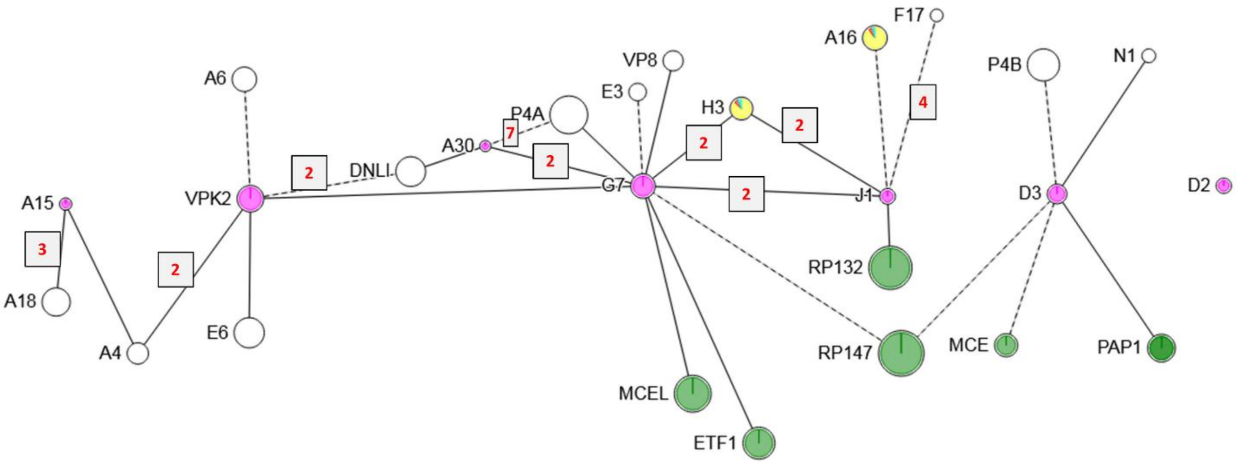


Figure 4.11. XL network among 7PC proteins and their detected partners. Proteins are depicted as circles whose areas correspond to chain length. Magenta fill: 7PC members. Other details as for Figs 10 and 5.

detected between I1 and I6, both of which are known to bind vaccinia genome telomeres ([30, 563] and references therein). Fig 12 shows a crosslinking sub-network encompassing proteins DNLI, I1, I3, I6, K4 and TOP1, along with three proteins that seemed to couple quite well with the above network, namely the two subunits of the vaccinia transcription factor heterodimer (ETF1 and ETF2) and protein E6.

We note that XL were detected very strongly between telomere-binding protein I1 and a centrally-located 'outside' domain of the double membrane-spanning TM protein E8 (Appendix 2.Fig1). E8 is retained with the virion core in the presence of the core-stripping/activating reagent combination NP40/DTT, binds single-stranded DNA in vitro [566], and has been proposed to connect the viral genome with ER and viral membranes ([566, 567] and references therein). The very strongly detected XL between E8 and I1 suggested the targeting of genome

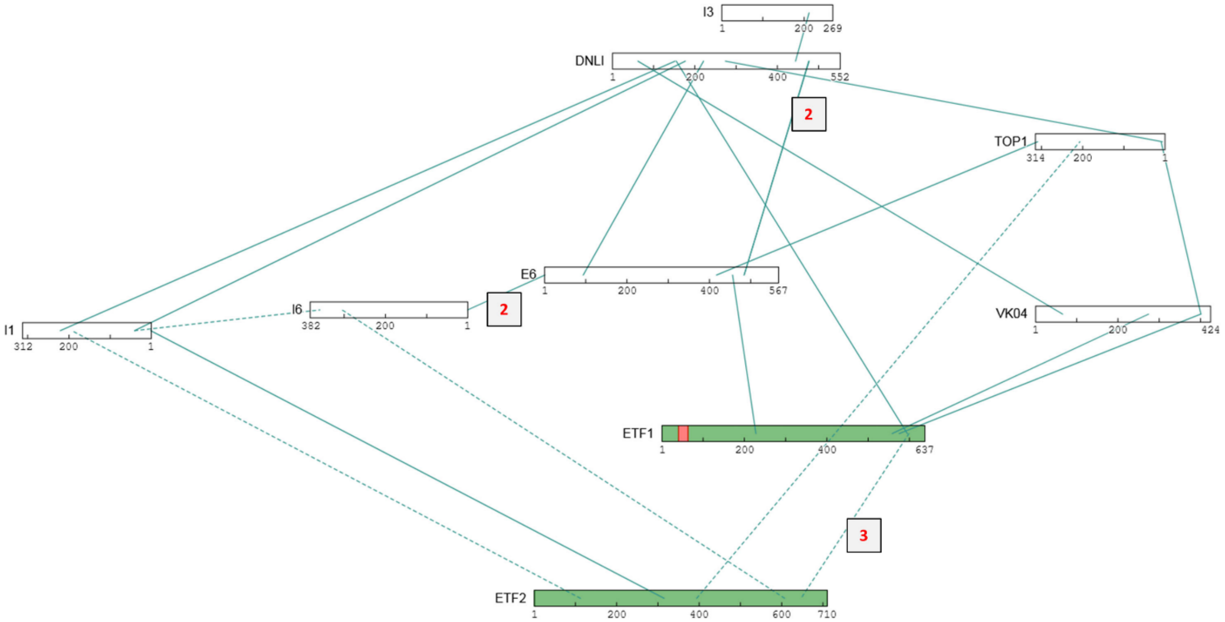


Figure 4.12. XL sub-network that includes 'telomere-binding' and other proteins from the 'DNA' functional group. Two transcriptional DNA binding proteins are also included, namely ETF1 and ETF2 (the early transcription factor heterodimer) along with protein E6. Other details as in Fig 5.

telomeres by a complex containing E8 and I1. Four of E8's 11 detected protein crosslinking partners were transcription-related (Appendix 2.Fig1), all of whose XL were to E8's terminal domains representing the opposite compartment to that targeted by protein I1 (Appendix 2.Fig1). This raised the possibility that, in MV, the genome telomeres may be somehow membrane-compartmentalized away from the transcriptosome plus the bulk of the genome destined to be transcribed by it during early infection. This scenario lacks complete physical clarity, however, since the sub-network of Fig 12 would place the early transcription factor ETF in the telomere compartment while the crosslinking pattern for E8 (Appendix 2.Fig1) would place the transcriptosome outside it. A provocative solution would place the dual-role (morphogenesis and early transcription) ETF in both compartments. Interestingly, virions from the temperature-sensitive mutant tsE8 assemble normally, but exhibit a transcriptionally inactive core [30]. We note that two topological polarities have been proposed for E8 and its interactor I5 as mentioned in the E8 section of Appendix 2.Fig1.

Other enzymes and proteins. DUSP, the vaccinia dual specificity phosphatase (H1) is reportedly resident in the lateral bodies of MV [76]. Although DUSP targets viral and cellular proteins and can dephosphorylate vaccinia proteins A17, A14, and F17 in vitro [30], no interaction with these three substrates was detected here. Instead, XL were detected quite strongly between the C-terminus of DUSP and ATI and CAHH proteins of the attachment complex (Fig 9b and 9c), along with transcriptosome proteins RP132 and ETF2.

Mutants in the I7 cysteine protease are defective in the IV/IVN to MV morphogenic transition and in the processing of major structural proteins P4a, P4b and A4, and the TM protein A17 (A17 being processed earlier during morphogenesis and not subject to a Rifampicin blockade [30]). None of these proteins was represented among the crosslinking partners of I7 detected here (Appendix 2.Fig1) perhaps because they are, in MV, enzymatic products rather than substrates. Among the six TM and membrane-associated XL partners for I7 was the TM protein H3 (discussed above), whose ‘outside’ domain crosslinked, in a strongly-detected fashion, to two positions towards C-terminus of I7. This region of H3 also crosslinked to A17, raising the possibility of H3 providing a scaffold for I7 protease, employed in the AG|-dependent removal of the A17 N-terminus. Crosslinking of I7 to proteins PAP1, F9, A18 and A13 involved the extreme C-terminal regions of each partner; I7 crosslinking to RP35, E6 and CAHH involved the extreme N-termini of these partners. Like H3, none of these proteins contains an AG diamino acid target for I7-dependent proteolysis, and these XL could represent merely the flexing of protein termini.

In virions with repressed G1 metalloprotease, which is defective in the IV to MV morphogenic transition, P4a, P4b, VP8, G7 or A17 precursors are cleaved normally [568], albeit VP8 is cleavable by G1 in a transfection assay [569]. In apparent contrast to this, XL were detected between G1 and each of the three major structural proteins P4a, P4b and A4 though not VP8 (Appendix 2.Fig1).

The C-terminus of the 65 aa F8 protein appeared as a Y2H interactor with VP8(L4) [507]. Consistent with this, crosslinking was clearly detected between the C-terminal regions of F8 and VP8 (Appendix 2.Fig1).

Inter-protein XL partitioning analysis. Appendix 2.Fig3 shows the results of inter-protein XL partitioning analysis (see Materials & methods) for key structural and membrane proteins. For proteins showing higher bars +NP40 and/or +TCEP than -NP40 and/or -TCEP, a barrier to crosslinker access is apparently dissociated by NP40 and/or TCEP. Among the DNA group proteins, inter-protein XL involving telomere-binding protein I1 fell into this category, consistent with a “telomere” compartment in the virion that can be exposed upon treatment with NP40/TCEP. The barrier is not necessarily lipid, since the crosslinker used for the majority of experiments, namely DSS, is membrane-permeable [570]). By contrast, for proteins showing higher bars -NP40 and/or -TCEP, either these proteins or their crosslinking partners are apparently lost by pre-treatment with these reagents. Among the structural proteins, inter-protein XL for VP8 and A12 seemed strongly favored by -NP40 and/or -TCEP as were, to some extent, P4a and P4b suggesting that pre-treatment results in the extraction of either these proteins or their partners. By contrast, A4 and F17 (which are located on the exterior core wall surface, above [73]) were indifferent to pre-treatment (Appendix 2.Fig3). In the membrane protein group, the interactions of proteins A26 and A27 with external partners seemed likewise sensitive to the presence of NP40 and/or TCEP while ATI and CAHH were indifferent. A26/A27 are known to be disulfide-bonded to the virion surface [125]. Overall, these result seem consistent with A26/A27 and VP8/A12 being connected to one another and to P4a/b via reducible disulfide bonds, while A4, F17, ATI and CAHH (D8) seem more resistant to disulfide reduction and/or may be anchored in other ways.

Discussion

We have investigated the molecular structure of the vaccinia virion, a highly non-stoichiometric protein assembly, via XL-MS. Analysis of protein-protein interactions in the virion in situ avoided the need for their preservation during virion extraction with reagents such as deoxycholate, an ionic detergent used for the release of virion core enzymes [309]. There was no requirement to rebuild virus protein complexes de novo, avoiding a need for the correct folding of challenging or insoluble structural proteins in vitro and/or in a heterologous system.

Finally, multivalent/higher order complexes could be addressed that were not accessible via binary assays such as Y2H [507].

As in any XL-MS study, challenges included: The availability and appropriate spacing of crosslinkable sites at protein interfaces; good occupancy of crosslinking sites and robust reaction of both ends of the crosslinker; efficient laboratory digestion of crosslinked proteins (given the tendency of trypsin recognition sites, for example, to become derivatized); the detection of crosslinked peptide pairs against a large excess of non-crosslinked peptides in the same digest; rarity of inter-protein XL (the most informative kind) with respect to other kinds (intra-protein, intra-peptide, and single-ended XL); the tendency of large (more than double-size) crosslinked peptide pairs to ionize less efficiently during MS; inefficient fragmentation and combinatorial complexity of fragment ion mass spectra when simultaneously fragmenting peptide pairs, and the challenge of distinguishing true intra-molecular XL from those that may cross homomultimer interfaces. For vaccinia as a target, the above issues were compounded by: Unknown permeability of the virion core to crosslinker; a protein abundance dynamic range in vaccinia MV of 5000-fold [100] or more; a paucity of existing high resolution protein structures for validation, and the possibility of molecular heterogeneity arising from mixed viroforms in MV preparations and/or mixed proteoforms within a single particle.

Addressing the above challenges (most particularly the abundance range and sensitivity issues) we adopted a “strategy of experimental variation”, as explored initially in our analysis of the MV phosphoproteome [571]. For XL-MS this strategy involved a ‘multithreaded’ workflow (Fig 1) in which experimental steps were matrixed combinatorially (Table 1). In this way, individual XL were placed in a variety of ionic contexts for MS detection, and key interfaces were painted as clusters of alternative XL between closely spaced crosslinking sites. This was combined with the use of diverse XL search engines for the identification of crosslinked peptides, and the use of isotopically coded crosslinkers where available. Our 86-protein search database comprised the maximum set of viral proteins considered likely to be packaged [101]. For all but two of these proteins XL were detected, the exceptions being proteins A14 and I2. These two short proteins (90 aa, 73aa respectively) possess relatively few sites for crosslinking and trypsin cleavage (3 lys/2 arg; 4 lys/0 arg, respectively).

Due largely to the absence of strong corroborating data for our XL-MS dataset such as comprehensive atomic-resolution three dimensional structures, validation relied largely on statistics and trends. The effective FDR of 0.33% for the final dataset as a whole ("Results"), suggested a remarkably low level of bioinformatics noise. Consistent with this, non-target databases from uncorrelated proteomes, namely all human proteins or the non-packaged subset of vaccinia proteins yielded very weak results in preliminary searches.

Alongside the detection of clear crosslinkome sub-networks ("Results") were many single-detect inter-protein XL (DFscore = 1, Appendix 2.Fig1). Notwithstanding the excellent bioinformatic signature for the dataset as a whole (above), it was difficult to ascertain to what extent the single-detect XL were real (from, for example, low abundance proteins, low abundance viroforms, inefficient XL, or poorly ionizing peptides), or represented biochemical noise (eg. virion dissociation pathways during virus preparation or specific experiments). On the one hand, evidence that single-detect XL were true positives included the tendency of single-detect crosslinking patterns within a protein sub-network to conform to patterns of XL with higher DFscore. For example, among the 22 inter-protein XL shown in Fig 5a, 18 were multi-detects vs. 8 single-detects, all contributing to the same overall crosslinking pattern. On the other hand, high DFscoring XL showed a higher ratio of biologically rational:non-rational XL than did single-detect XL, lending greater confidence to former. For example, among the 37 inter-protein XL in the dataset with DFscore > 5 (Table E in Appendix 2.Doc1), the number that were considered biologically rational exceeded the number designated non-rational by a factor of 9.5 while, among the single-detect XL from the same table, rational exceeded non-rational XL with a factor of only 1.5.

Transcriptosome proteins, albeit presumably packaged in relatively low abundance, nonetheless showed a number of strongly detected inter-protein XL. Some of these, including some of the most strongly detected inter-protein XL in the dataset, were between transcriptosome and membrane proteins, including ectodomains of the latter. These XL were considered biologically "non-rational" (above) since the transcriptosome is located within the virion core while the TM proteins surround it according to conventional models. They were strongly supported by their DFscores, were not filterable by raising score thresholds, and their DFscores did not drop when switching between singly- and dually-thresholded filtering (Materials & methods). We were therefore unable to falsify a hypothesis that contacts can occur between transcriptosome components and the ectodomains of membrane proteins, the significance of which is unclear. Possibilities for these resilient, yet 'non-rational' XL may include that: (a) the core wall is not a fundamental barrier to crosslinking (it is porous)—indeed the 7 nm inside-diameter pores that have been imaged in the core wall [73, 93] may be sufficiently large for the majority of vaccinia polypeptides to pass through entirely if they are globular and approximately spherical [572], (b) TM and transcriptosome proteins are both implanted in the barrier (from opposite sites)–a situation, on the transcription side, observed in the cores of turreted reoviruses [573, 574], (c) TM proteins are located in more than one compartment, (d) MV preparations contain developmental viroforms from a time prior to the full emergence of the core wall, (e) they are cryptically artefactual.

The dataset contained evidence for viroforms/proteoforms from proteolytic maturation. Peptides crossing known [30] sites of viral AG| specific proteolytic processing in proteins A17, VP8, G7, P4b and P4a (site2) can be found in tryptic digests of purified MV [101]. These peptides represent pre-cleaved proteoforms. Such peptides were also found in the current study, within crosslinked pairs, from proteins A17, VP8, A12 and G7. XL connecting the N-terminal amino group of pre-cleaved A17 with the C-terminal region of the same protein may be an example of the same phenomenon. In some cases the crosslinker directly spanned an AG| processing site. Apparently, then, MV harvested from Hela cells late in infection followed by 2x sucrose gradient-purification were accompanied by immature viroforms that are detectable by highly sensitive MS. Among crosslinked peptides could be found no trace, however, of a characteristic and abundant marker of IV, namely the external scaffold protein D13 when using XL search databases that included this protein. Apparently, in MV, in which the external scaffold, along with fragment 2 of protein P4a (Fig 5b) are close to or below the detection limit, unprocessed forms of proteins A17, VP8 and A12 are still readily detectable. If, speculatively, MV preparations contain trace viroforms that appear morphologically mature (having already escaped the external scaffold and perhaps received tailanchored and SFE proteins), but which still lack a fully formed interior and/or an impermeable core wall, then this may account for some of the more counter-intuitive XL detected here. Alternatively, some XL may represent structures that appear only transiently in the virion maturation pathway. Another possibility may be that MV particles, albeit fully mature, retain unprocessed proteoforms by design. Within an A17 homodimer, for example, one subunit might be processed and the other not.

Evidence for multiple viroforms/proteoforms also arose from interactions between P4a, P4b and TM proteins: P4a fragment 3 was found to be within crosslinking range of seven distinct membrane proteins (Appendix 2.Fig1, Fig 6) and also within crosslinking range of the C-terminal region of P4b, while none of the seven membrane proteins were apparently within crosslinking range of P4b (Fig 6, Appendix 2.Fig2A and B). Moreover, a crosslinking 'hotspot' in P4a fragment 3 (K736) interacted with three membrane proteins as well as P4b (Appendix 2.Fig1), in the absence of any detectable crosslinking between the latter. While steric factors may allow P4a, P4b and membrane proteins to triangulate in a way that leaves all membrane proteins out of range of P4b, it seems also possible that membrane proteins and P4b may interact with alternate proteoforms of P4a. This could result from distinct and segregated P4a complexes within individual MV, or distinct viroforms in the virus preparation (e.g. the rearrangement of P4a fragment 3 during maturation).

In conclusion: Here, we have covered the crosslinkome of a relatively small whole organism in depth, detecting inter-protein XL for all but two of the 86 proteins that represent the maximal virion proteome. Strategies were developed to detect XL in a proteome covering a wide abundance dynamic range and with minimal pre-existing crystallographic information, allowing the reconstruction of several key virion protein complexes. The challenge of synthesizing the data into an extended understanding of the internal molecular architecture requires some knowledge of intra-particle protein stoichiometry.

Materials & methods

Materials

Vaccinia virus was purified by sucrose or tartrate gradient as described [101] and protein quantitated using BCA (ThermoFisher Inc.), determining concentrations to be between 1 and 3.5 mg ml-1. DSS-H12, DSS-D12, DSG-H6, and DSG-D6 were obtained from Creative Molecules Inc. BS3-H4, BS3-D4, BS(PEG)5, BS(PEG)9, Zeba Spin Desalting Column (7K MWCO), and LysN were obtained from Thermo Scientific. DSS, bis(sulfosuccinimidyl)suberate (BS3), and disuccinimidyl glutarate (DSG) were used as 1:1 mixtures of DSS-H12/DSS-D12 ('DSS-H12/D12'), BS3-H4/BS3-D4 ('BS3-H4/D4'), and DSG-H6/DSG-D6 ('DSG-H6/D6') respectively. Trypsin, dimethyl sulfoxide (DMSO), Benzonase, iodoacetamide, n-LS, adipic acid dihydrazide (ADH), 4-(4,6-dimethoxy-1,3,5-triazin-2-yl)-4-methylmorpholinium chloride (DMTMM), 1-[bis(dimethylamino)methylene]-1H-1,2,3-triazolo[4,5-b]pyridinium 3-oxid hexafluorophosphate (HATU), and cyanogen bromide (CNBr) were from Sigma-Aldrich. GluC, AspN, LysC, LysN and ArgC were from Promega. AspN was from Roche Diagnostics.

C18 and SCX filters were obtained from 3M. N,N-Diisopropylethylamine (DIPEA) was from Alpha Aesar. Centrifugal concentrators (Vivacon, 10kDa MWCO) were from Sartorius Stedim Biotech.

Virus pre-treatment

Prior to crosslinking, virus was washed 5x with phosphate buffered saline (PBS), pH 7.4, by centrifugation and resuspension. For some experiments, washed virus pellets were then resuspended in 10 μL of 0.1 M triethylammonium bicarbonate (TEAB, pH 8.5) and supplemented with an equal volume of 2x 'pre-treatment' buffer comprising either 0.1 M TEAB, 0.1% NP40 (pH 8.5), or 0.1 M TEAB, 0.1% NP40, 80 mM TCEP (pH 8.5), followed by 2 min incubation.

Amine-amine crosslinking

The method of ref. [575] ('xQuest crosslink method') was used with some modifications. Pretreated virus suspension (above), or intact virus suspended in 0.1 M TEAB (pH 8.5), was supplemented with 1/10 volume of 10x crosslinking buffer (0.2 M 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES), KOH, pH 8.2). Crosslinker, dissolved freshly in DMSO, was then added at a final concentration of 7.5 mM. Following 30–60 min incubation at 37˚C, samples were quenched by adding 1 M ammonium bicarbonate (AmBic) to a final concentration of 50 mM followed by 30 min incubation at 37˚C.

Carboxyl group crosslinking

<u>ADH with either HATU/DIPEA or DMTMM</u>: Pre-treated virus suspension (above) was supplemented with 10x ADH-XL buffer (0.2 M HEPES-NaOH, pH 7.2) to 1x ADH-XL buffer (final) then supplemented with ADH and HATU (dissolved separately in 1x ADH-XL buffer) to final concentrations of 6 mM and 9.2 mM respectively. 100% DIPEA was then added to a final concentration of 46 mM. After 120 min incubation at room temperature with continuous shaking, crosslinked virus was exchanged into 50 mM AmBic using a spin desalting column (Zeba, ThermoFisher, Inc.) following the manufacturer's instructions. In some experiments, HATU/DIPEA were replaced with DMTMM, using concentrations of ADH and DMTMM described [576].

<u>ADH/EDC/NHS or EDC/NHS alone</u>: Pre-treated virus suspension (above) was supplemented with 10x ADH-XL buffer (0.2 M HEPES-NaOH, pH 7.2) to 1x ADH-XL buffer (final) then supplemented with ADH (dissolved separately in 1x ADH-XL buffer) to a final concentration of 6 mM. N-(3-dimethylaminopropyl)-N0-ethylcarbodiimide hydrochloride (EDC) and N-

hydroxysuccinimide (NHS), dissolved separately in 1x XL buffer were then added at final concentrations of 8 mM and 10 mM, respectively. After 120 min incubation at room temperature, free crosslinker was removed by spin desalting into 50 mM AmBic (above).

ADH with EDC or EDC alone: Pre-treated virus suspension (above) was supplemented with 10x MES buffer (0.1 M 2-(N-morpholino)ethanesulfonic acid (MES), 20 mM NaCl, pH 4.7) to 1x MES buffer (final), then supplemented with ADH and EDC (dissolved separately in 1x MES buffer) to final concentrations of 6 mM and 2 mM, respectively. After incubation for 120 min at room temperature, free crosslinker was removed by spin desalting into 50 mM AmBic (above). For some experiments (EDC-alone crosslinking) ADH was omitted.

Solubilization of crosslinked virus for protease digestion

Crosslinked virus samples in 50 mM AmBic were disaggregated by supplementing with 0.5 M TCEP, 1 M TEAB and solid urea or guanidine, then diluting to achieve a final formulation of 8 M urea, 0.1 M TEAB, 10 mM TCEP, pH 8.5 (urea buffer) or 6 M GuHCl, 0.1 M TEAB, 10 mM TCEP, pH 8.5 (guanidine buffer). In some experiments, crosslinked virus suspension in 50 mM AmBic was instead supplemented with an equal volume of 2x detergent solution to achieve 0.5% sodium deoxycholate (SDOC), 12 mM n-laurosarcosine (n-LS), 5 mM TCEP, 50 mM TEAB, pH 8.5 (final). After 30 min incubation at 37˚C, some samples were alkylated with iodoacetamide at either 5 mM (if supplemented with urea or GuHCl), followed by 30 min incubation in the dark) or 10 mM (if supplemented with detergents), followed by 15 min incubation in the dark. Some samples were then incubated with Benzonase (250 units) for 60 min. All samples were then diluted with 50 mM AmBic for cleavage, according to the manufacturer’s recommendation for tolerable denaturant (below).

Cleavage

Cleavage employed the following reagents/reagent combinations: Trypsin, ArgC, GluC, AspN, LysN, LysC, or Trypsin+GluC, Trypsin+AspN, ArgC+AspN, ArgC+GluC, AspN+GluC or CNBr+Trypsin.

For digestions containing GluC, samples were diluted to a final urea concentration of 0.5 M. For digestion with LysN, samples were diluted to a final urea concentration of either 1 M or 5 M. For all other proteases, samples were diluted to final denaturant concentrations of either 0.6 M GuHCl, 1 M urea, or 0.1% SDOC/2.4 mM n-LS/1 mM TCEP. With the exception of DigDeApr experiments (below), a protease:substrate ratio of 1:50 or 1:100 was used. With the exception of LysC, which was used for 72 hr at room temperature, all protease digestions were overnight at 37˚C.

For CNBr+Trypsin digestion, quenched amine-amine crosslinking samples (above) were supplemented with 100% formic acid (FA) to 70% (final) followed by the addition of one crystal (~20–100 molar excess) of CNBr and overnight incubation at room temperature in the dark. After evaporation to dryness under vacuum, samples were redissolved in urea buffer (above), followed by 30 min incubation at 37˚C in the dark. Samples were then diluted to 1 M urea with 50 mM TEAB (pH 8.5), and trypsin added to an estimated enzyme-to-substrate ratio of 1:100 followed by overnight incubation at 37˚C. A fresh equivalent of trypsin (same amount) was then added, followed by a further 4 hrs digestion. Undigested material was precipitated by centrifugation at 14,000 g for 2 min followed by resuspension in 70% FA and redigestion with CNBr and trypsin following the same method.

DigDeApr

This was done following ref. [577] with modifications. Briefly, samples were digested with either trypsin alone (enzyme:substrate ratio of 1:2500) or Trypsin+AspN, Trypsin+GluC or AspN+GluC (1:1:2500). After overnight incubation at 37 ˚C, samples were centrifuged into a centrifugal concentrator (10kDa MWCO, Vivacon) at 2500 x g. After collection of flow through, the filter was washed by centrifugation at 2500 g with 8 M urea, 0.1 M TEAB pH 8.5 then with 2 M urea, 0.1 M TEAB pH 8.5 (Wash buffer) for 2 min at 2500 x g. Flow through and wash-throughs were combined. Using a new collection vial, urea buffer was added to the filter which was then inverted and spun for 2 min at 2500 x g. The process was repeated and the combined urea buffer washes were brought to 1 M urea with 0.1 M TEAB then treated again with the same protease combination at an enzyme:substrate ratio of 1:100 (for GluC digestion, samples were diluted to 0.5 M urea, 100 mM TEAB) overnight at 37 ˚C.

C18-SCX

All cleaved samples were acidified with FA to 2% FA final then desalted as described [578] using stacked C18-SCX filters. After washing the filters, peptides were transluted from the C18 to the SCX phase using 80% CH3CN/0.1% FA (translution buffer). Peptides were eluted with 5% NH4OH/80% CH3CN (Buffer X) or with six steps of 20% CH3CN/0.5% FA containing ammonium acetate in the range 160–800 mM followed by a final step of Buffer X. Elutions were dried under vacuum then re-dissolved in 0.1% FA in water for MS.

nanoLC-MS/MS

nanoLC-MS/MS was performed using an LTQ Velos Pro Orbitrap mass spectrometer with Easy-nLC 1000 (ThermoFisher). 2 microL injections were followed by a segmented LC gradient (solvent A = 0.1% FA in water, solvent B = 0.1% FA in CH3CN), progressing from 0 to 10% B over 10 min then to 35% B over 230 min. Some runs used a straight gradient of 0–35% B over 135 min. Precursor spectra were acquired in FT mode at a resolution of 100,000 (centroid) in the range 200–2000 Th. For isotopic pairs with 12 Da mass split (DSS crosslinker), the top 3 most intense ions were selected for HCD activation (above a precursor signal threshold of 150) on the basis of isotopic pairs with m/z spacing of either 4.02524, 3.01893 or 2.41515 (representing +3 to +5 charge-states), and intensity ratio better than 2:1. For a 6 Da mass split (DSG crosslinker), m/z deltas for isotopic pair selection were 2.01456, 1.51092 or 1.20874. For a 4 Da mass split (BS3 crosslinker), m/z deltas were 1.34156 or 1.00616. Both isotopic partners were fragmented. HCD activation used a normalized collision energy (NCE) of 45, activation time of 0.1 mSec and an isolation width of 2 m/z. MS2 spectra were acquired in FT mode with a resolution of 7500 (centroid). The dynamic exclusion list size was 500, exclusion duration was 60 sec, repeat duration was 30 sec and the repeat count was 2, with early expiration enabled. Charge state screening was enabled, with rejection of 1+ and 2+ and unassigned charge states.

Data acquired for xQuest were activated in IT-CID mode instead of HCD. Here, NCE was 35, activation Q = 0.25 and activation time was 10 mSec. For non-isotopic crosslinkers, the 10 most intense ions in each precursor spectrum were subjected to HCD fragmentation, as above, if above a minimum signal threshold of 250 (or 2000 in some early experiments).

Bioinformatics

Protein names used throughout this report follow entry names in the UniProtKB vaccinia WR reference proteome minus the species identifier suffix. They are comprehensively crossreferenced to other naming schemes in Table S1 of ref. [101]. Instrument raw files were converted to mgf, mzXML or mzML using MSConvert by ProteoWizard. Using the resulting data, XL were identified using the following XL search engines: Protein Prospector [579], pLINK [580], xQuest (in combination with xProphet) [575], Kojak [581] (in combination with 'Percolator' [582-584]), ECL [585] and ECL2 [586], as follows:

Protein Prospector: Instrument raw data files were converted to mgf format then uploaded to Protein Prospector via the UCSF online server. Non-standard, PEGylated bis(sulfosuccinimidyl)suberatecrosslinkers (BSPEG5 and BSPEG9) were imputed as user defined parameters. The results file from each run was generated using the program's Search Compare function. Results were sorted by ascending expectation value and "Report type" was set to "crosslinked peptides". 'SD-E' = ScoreDiff–log10(Exp2) ([579], Robert Chalkley, Personal communication) where ScoreDiff is the difference in score between the top- and second-ranked peptide 1 in the search output for a crosslinked pair, Exp2 is the score for peptide 2.

pLink: pLink was downloaded from pFind Studio. A parameter file was configured for each experiment and a folder created, containing mgf files pertaining to that experiment along with the search DB. The 'pLINK.ini' configuration file was modified for each experiment to include the path to the mgf and search DB and search parameters. The enzyme.ini and xlink.ini files were modified for any non-standard cleavage specificities/combinations and crosslinkers, respectively. Results files for loop linked and mono linked peptides were generate using "non-interexport" and "drawpsm", respectively. pLink was run through the flow.exe application.

xQuest: The xQuest VMware package was installed on a Windows PC. Directories were created following instructions provided with xQuest. Files "Xmm.def" and "xquest.def" were modified for the relevant crosslinker isotopic mass, shift and ion charge states. A text file was created containing the mzML file name and parameter files for xProphet. xQuest, then xProphet, were run from the command line. Results were viewed on the xQuest webserver then downloaded. Values reported by xProphet as "FDR" may be Percolator-derived q-values.

Kojak: Kojak and Percolator [587] were installed and run in Linux from the command line. Folders were created for mzML formatted data and search results. The program's configuration file was modified to contain all relevant crosslinkers and the paths to individual data files. Parameters are outlined in Table I in Appendix 2.Doc1. Digestion specificity rules were based on the parameters provided.

ECL/ECL2: ECL and ECL2 were installed on a Java-capable Windows PC and run from the command line. Crosslinker masses were entered manually.

Percolator: For Kojak and ECL, FDR was converted to a q-value using the program 'Percolator' [587], run from the Linux command line. For Kojak, Percolator input comprised "inter", "intra", and "loop" search output files. q-value can be regarded as the expected proportion of false positives among all features as or more extreme than the observed one [585, 588] or, alternatively, the minimal FDR threshold at which a given peptide-spectral match is accepted [583, 584].

Data assembly: Using in-house code, XL search engine/Percolator/xProphet outputs corresponding to various nanoLC-MS/MS runs in various experiments were parsed in their native formats, accepting individual XL to a single unified dataset according to dual score thresholds for each program including in-house-calculated FDR for Protein Prospector (see above and Table 2). The resulting dataset was then sorted by ascending exp_Mr. Groups (blocks) of masses matching to within 25 ppm were annealed, then each block that contained multiple accession/PeptideSeq/ProteinPos was sorted and divided into distinct sub-blocks of ions that were tagged with a common ‘ambig code’ (representing sub-blocks having functionally isomeric mass but had been assigned, by XL search engines, distinct apparent identities). The resulting ‘mature blocks’ each represented a unique combination of exp_Mr and accession/PeptideSeq/ ProteinPos.

This dataset was reformatted/collapsed into a matrix with one row per mature block, and one column for each nanoLC-MS/MS run in the project. The matrix was filled with XL search engine identifiers to indicate all engines identifying a specific mature block member in a specific nanoLC-MS/MS run and the number of times identified. Each row was assigned a DFscore as the sum of search engine identifiers/times identified by that engine that had been assigned to the row. Groups of mature blocks sharing a common ambig code were likelihood-scored against one another as follows: If they all represented intra-protein XL or all represented inter-protein XL, then the ambigscore assigned to those mature blocks was a simple proportion of its DFscore/∑(DFscores for all blocks sharing a common ambig code). If they were a mixture of intra-protein and inter-protein, then intra-protein mature-block(s) were scored 1.0 and inter-protein mature-block(s) 0 (assuming the intra-protein XL to be correct by default). If the ambiguity was simply in choice between multiple lysines within an otherwise identical peptide, both choices were scored 1.0 since both reflect the same approximate position within the same protein partners. Finally, for every specific position in a specific protein represented by multiple rows in the matrix: If the intra-protein XL were discovered by multiple engines and the inter-protein XL were discovered by one only, the latter were annotated as “filterable”. The resulting annotated matrix was written to an Excel worksheet then copied to a second sheet which was re-sorted by protein position then accession.

A ‘discard matrix’ (comparable in structure to the above, ‘passing’ matrix) was generated representing all XL ions in the above assembly that passed threshold1 but were rejected after failing threshold2. Each row of the ‘discard’ matrix was annotated with: (a) DFscore; (b) whether the XL (Accession1/Accession2/ProteinPos1/ProteinPos2) was also present in the passing matrix (above; this criterion being denoted ‘also’ in the following discussion) and (c) ‘biological rationality’ (‘BR’) based on six groups of functionally-related virion proteins (Table 3), annotating “Y”, if the two crosslinked proteins were in same BR group, and “N” if one was from the ‘membrane’ group and the other from the ‘transcription’ group.

Networks and sub-networks for individual accessions or groups of accessions were rendered using CrosslinkViewer [589]. Using in-house code, rows in the passing matrix (above) were picked provided either one or both crosslinked proteins did not match accessions within a user-definable excluded-accession group. ‘Filterable’ rows of the matrix (above) were excluded. The list of picked rows was supplemented with those from the ‘discard’ matrix (above) if DFscore > 1 or ‘also’ = “Y” or BR = “Y”. Rows with common Accession1/Accession2/ProteinPos1/ProteinPos2 were then collapsed summing DFscores, and the resulting dataset reformatted for input to CrosslinkViewer. The resulting DFscores were rendered. If 100% of matrix rows for a given Accession1/Accession2/ProteinPos1/ProteinPos2 had been flagged as ambig (above), then the XL was flagged to be rendered with a broken line. Protein monolinks were ignored in all data operations.

Domain prediction: TM regions were predicted using program TMHMM [493, 557]. Domain boundaries were predicted using DomPred [590, 591]. Output traces show endpoint density profiles for PSI-BLAST alignments generated between a query sequence and a database in which all sequence fragments had been removed.

ROC analysis of the global crosslinking dataset: Each of the 81 proteins in the dataset for which crosslinked partner proteins were found, was flagged according to membership of one of two ‘biological rationality’ groups in Table 3 (‘Membrane’ and ‘Transcription’), and total number of distinct crosslinking partner proteins was printed alongside. In each of two replicates of this listing was printed the # of partners belonging specifically to one of the two groups and the list was sorted (descending) according to proportion of total partners belonging to the specific group. After incrementing four number series at each row in the list that contained either: a membrane protein, not(a membrane protein), a transcription protein and not(a transcription protein), then proportionating each series to a scale from 0 to 1, ROC curves were drawn based on the proportionated values. In ROC space, points above and below the line of no-discrimination (diagonal) represent positive (better than random) and negative correlation, respectively such that a curve representing perfect positive correlation would ascend vertically from (x,y) = (0,0) to (0,1) then travel horizontally to (1,1). Perfect negative correlation would yield the converse curve: (0,0) to (1,0) to (1,1).

Inter-protein XL partitioning analysis: For each XL ion in the global XL dataset representing an inter-protein XL, DFscores from each experiment (column) in which the ion was detected were binned according to whether sample pre-treatment included or excluded NP40 or TCEP (-NP40, +NP40, -TCEP or +TCEP). The resulting tetra-bin DFscore values for individual ions were accumulated on a per-accession basis, according to the accession on each side of the crosslink. The accumulated four DFscore values for each accession were then converted to a proportion of the summed DFscore across the four bins (‘POSD’) for that accession, and the resulting POSD values were finally normalized to the mean POSD per accession for each of the four pre-treatment conditions. See legend to Appendix 2.Fig3 for further details.

Distance measurement

Distance analysis of the form shown in ref. [514]) was generated using 'TopoLink' [521] installed on a computer cluster and run from the command line, in combination with 14 relevant pdb files. Before calculating Euclidean and SAS distances for each experimental XL, "inputfile.inp" was modified to include the crosslinker type, maximum linker distance and reactive residues. All lys-lys Euclidean and SAS distances were also calculated within the 14 structures, setting maximum linker distance to 100 Å. Each of the resulting four distance datasets, in spreadsheet format, was binned for display as a histogram. The mean and standard deviation (SD) from each "all lys-lys distances" histogram informed a normal (Gaussian) curve overlay. For each "experimental XL distances" histogram, Ln(mean) and ln(SD) informed a log-normal curve overlay. "Experimental XL distances" and "all lys-lys distances" datasets were compared to one another via the Kolmogorov Smirnov test, run using the Excel plugin XLSTAT (https://www.xlstat.com/en/).

CHAPTER 5

Structural proteomics (XLMS): workflow optimization for the identification of crosslinked peptide pairs by XLMS from a complex virus

INTRODUCTION

Structural proteomics in the form of chemical crosslinking - mass spectrometry (XLMS) provides a powerful approach for studying protein structures and protein-protein interactions and has become a key component of the structural biology toolkit. Recent advances in mass spectrometry instrumentation and data analysis software have improved the discovery, confident identification, and characterization of crosslinked peptides [592, 593], though some inherent challenges associated with XLMS, namely enrichment and characterization of low abundance crosslinked peptides from a complex peptide mixtures, the confident characterization of both members of a crosslinked peptide pair, and the quadratic expansion of the bioinformatics search space associated with increasing sample complexity [579-581, 594], persist.

The traditional approach to analysis of proteins by mass spectrometry (described as a "bottom-up" approach), requires digestion of proteins to peptides using a defined (predictable) protease such as trypsin, followed by analysis of the peptides by liquid chromatography-tandem mass spectrometry (LC-MS/MS) [595]. This is a reductionist approach, since peptides are far more tractable to ionization, fragmentation and accurate mass analysis, and are far less variable in overall physicochemical properties and modifications than are intact proteins. Peptide sequences are determined from the fragment ion spectra generated by the mass spectrometer and used to establish protein identity via bioinformatic comparison of spectral data to the predicted set of peptides from the target proteome. Sample preparation, through protein extraction, digestion, and enrichment have a substantial influence on the quality and number of peptides and proteins identified [596, 597]. For most sample types, however, the number of peptides present after digestion far exceeds the total number of peptides that can be typically identified by a mass spectrometer in a single run. A high abundance dynamic range within the sample proteome can render the identification of low abundance proteins difficult, as high abundance-high intensity peptides (usually from high abundance proteins) are preferentially selected for fragmentation by the mass spectrometer [595] in standard approaches with modern instruments. Crosslinked peptides are particularly difficult to identify as they typically represent only 1% of the total peptide population of the sample and often have unfavorable ionization and fragmentation properties [581, 598-602].

There are additional challenges associated with crosslinked peptide identification. As the proteome depth of a crosslinked sample increases, the combinatorial search space increases quadratically (because every peptide could theoretically be crosslinked with every other peptide), resulting in fewer peptide identifications and higher false discovery rates (FDR) [594]. Inefficient digestion of proteins to peptides, resulting in peptides with high missed cleavage rates also contributes to a larger search space [597, 600, 603] and may also result in large peptides that fall outside the m/z range of the mass spectrometer and are therefore invisible.

Not only is vaccinia virus a challenging target for molecular structural studies in general (above), but the vaccinia virion is a challenging target for XLMS also, with > 1000 fold dynamic range in protein abundance between the most abundant structural proteins and enzymes that are packaged in trace amounts [101]. The viral structural proteins, by natural design, are quite resistant to disaggregation, solubilization, and proteolysis, resulting in higher rates of missed cleavage sites. Additionally, while only 70 - 80 viral proteins are packaged in the virion [101], purified virus samples often have a high background of contaminating host and viral non-packaged proteins, resulting in a substantially larger bioinformatics search space. In 2019 we implemented a “strategy of variation” to maximize the detection of inter-peptide crosslinks. This involved varying every step in the pipeline shown in (Figure 4.1), resulting in the identification of 4,609 unique crosslinked peptides [604]. However, this paper is regarded as just a proof of concept: Although we successfully established various protein subnetworks, at that time the work had to be done under quite severe resource limitations (e.g. limited funding) and the crosslinking network achieved was of insufficient depth to facilitate a comprehensive structural model of a complex virus. Availability of funding allowed us to really examine the factors required to generate a dataset of the required quality and depth. This involved covering a lot of new ground, since there were no reports in the literature of how to get a truly saturating dataset for an organism. We started with questions of virus yield and purity and bioinformatics. In this chapter, we fully optimize and integrate the premier XLMS search engines available at the time. This level of optimization and integration is unique in the XLMS field.

RESULTS

Crosslinked peptide identification by specialized search engines

Successful identification of crosslinked peptides after mass spectrometry analysis requires specialized software for database interrogation. Standard mass spectrum search engines can predict all theoretical [tryptic] peptides from a target proteome from an organism with sequenced genome, but are incapable of interpreting the complex MS/MS fragmentation spectra generated from a sample in which pairs of peptides are crosslinked to one another, and are limited to the identification of unreacted peptides and those in which just one end of a bifunctional crosslinker has reacted with a peptide to generate a predictable, simple modification (a mono-linked peptide) [579-581, 601]. The abundance dynamic range of proteins in a crosslinked sample, combined with the abundance and ionization dynamic range of peptides of various types (unreacted, mono-linked, loop-linked, and inter-peptide crosslinked) of which inter-peptide crosslinks are the most valuable yet least abundant, adds a further layer of complexity to crosslinked peptide identification [581], not to mention the level of saturation of the original crosslink. Furthermore, as sample complexity increases, the bioinformatics search space undergoes quadratic expansion, along with a corresponding increase in FDR. This is because every peptide could potentially be a partner of every other one, and the corresponding increase in opportunity for false positives leads to an FDR rise.

XLMS search engines have undergone substantial improvements since the start of this project, necessitating a reevaluation of our approach for analyzing the mass spectra of crosslinked samples. In our proof-of-concept work published in 2019 [604], we employed a “strategy of variation” where in addition to diversifying experimental conditions, we subjected all mass spectra to analysis by a total of six search engines. These were: “pLINK” [580], “Protein Prospector” [579], “Kojak” [581], “xQuest” (with xProphet) [598], “ECL” [585], and “ECL2” [586]. Some of these search engines are now no longer supported with updates. From reviewing multiple comparative analyses of XLMS search engines [594, 605-607] and after in-house tests, we identified pLINK2 [594] and Kojak2 [608] (the successors to pLINK and Kojak), the updated Protein Prospector, and the newly released MetaMorpheus search engine [609] as providing the algorithms offering greatest sensitivity and accuracy for our purposes while maintaining complementary algorithmic approaches in their analyses.

All four search engines employ distinct strategies for resolving the complex fragmentation spectra that results from crosslinked peptides, and have developed various approaches to reduce the impact of a large proteome on the bioinformatics search space and search times [579, 594, 608, 609]. While it is still desirable to limit the proteome complexity of the sample via some form of enrichment, and a large search space can still be detrimental for maximizing crosslinked peptide identification, a complex sample is no longer prohibitive for XLMS analysis.

We evaluated the performance of all four search engines to determine if any one of them was superior to the others or if we should maintain our approach of routinely analyzing crosslinking mass spectra with multiple search engines. Purified vaccinia virus was crosslinked with BSPEG5 and solubilized, then digested with LysC and trypsin. Prior to mass spectrometry analysis, the sample was desalted and also pre-fractionated, using a C18/SCX StageTip [578]. The sample fractions were analyzed on an Orbitrap Velos Pro mass spectrometer and the resulting mass spectra were searched using all four XLMS search engines in parallel. Coincidently, we also tightened the thresholding stringency imposed upon all software, from 5% to 1% FDR or better.

The total counts of crosslink spectral matches (CSMs) and unique inter-peptide crosslinks identified by all four search engines at the 1% FDR level, are given in **Figure 5.1**. 408 unique inter-peptide crosslinks (by protein accession, XL position and crosslinked peptide Mr), with a combined CSM count of 5,239, were identified during this analysis, of which 45 unique inter-peptide crosslinks were identified by all four programs, with a combined CSM count of 2398. A further 160 unique inter-peptide crosslinks (82.5% of total identified CSMs) were identified by at least two distinct search engines. The remaining 17.5% of total CSMs, which were identified by individual programs only, accounted for 60% of the unique inter-peptide crosslinks identified (**Fig 5.1**).

The pLINK2 search engine, which is currently ranked as one of the best search engines for XLMS with non-cleavable crosslinkers [593], showed the highest level of agreement with other three engines – 82% of the unique inter-peptide crosslinks identified by pLINK2 (95% of the total CSMs identified by pLINK2) were also identified by at least one other search engine. MetaMorpheus, conversely, showed the least agreement with the other XLMS search engines – only 53% of unique inter-peptide crosslinks identified by MetaMorpheus (75% of total CSMs identified by MetaMorpheus) were identified by other programs (**Fig 5.1**).

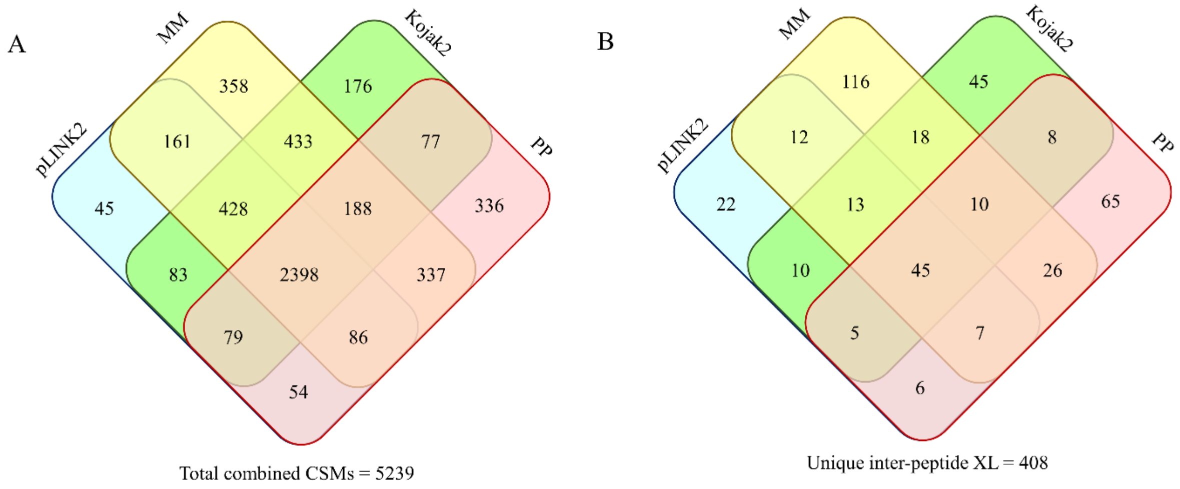


Figure 5.1. Performance evaluation of four crosslinking search engines. Venn diagrams show the overlap in identified crosslinked peptides by pLINK2, MetaMorpheus (MM), Kojak2, and Protein Prospector (PP) based on (**A**) total number of crosslinked spectral matches (CSMs) identified and (**B**) number of unique inter-peptide crosslinked pairs (by protein accession, XL position and crosslinked peptide Mr) identified.

Given that no individual XLMS search engine was vastly superior to the others and the crosslinking proteome depth achieved by incorporating all four search engines compensated for the added computational time and complexity of the analysis, we decided to include all four search engines in our experimental workflow. We discontinued the use of the five outdated search engines – pLINK, Kojak, xQuest, ECL, and ECL2. **Figure 5.2** shows an overview of the bioinformatics pipeline (data flow) and the resulting data files produced by individual search engines from a single crosslinking experiment, which were then consolidated with a “grand unified” crosslinking dataset. pLINK2 and MetaMorpheus can directly process Thermo raw files generated by the mass spectrometer, while Kojak2 and Protein Prospector require files to be converted to mzML and mgf formats respectively. pLINK2, MetaMorpheus, and Kojak2 can all analyze multiple spectral files in one search, reducing the overall complexity of the analysis. However, spectral files must be individually searched on the Protein Prospector online server. Additionally, Protein Prospector and MetaMorpheus can only search for one crosslinking reagent at a time (unless the reagent is available as a preset parameter in the Protein Prospector server), requiring samples crosslinked with stable isotope reagent pairs (equimolar amounts of the deuterated and undeuterated crosslinking reagent) to be searched twice to identify both heavy and light isotope crosslinked peptides. The data output formats and data thresholding by FDR and secondary score also differ for each search engine and a specialized in-house program was written to recompute Mr, filter, and harmonize the data outputs. The latter code was written by my supervisor.

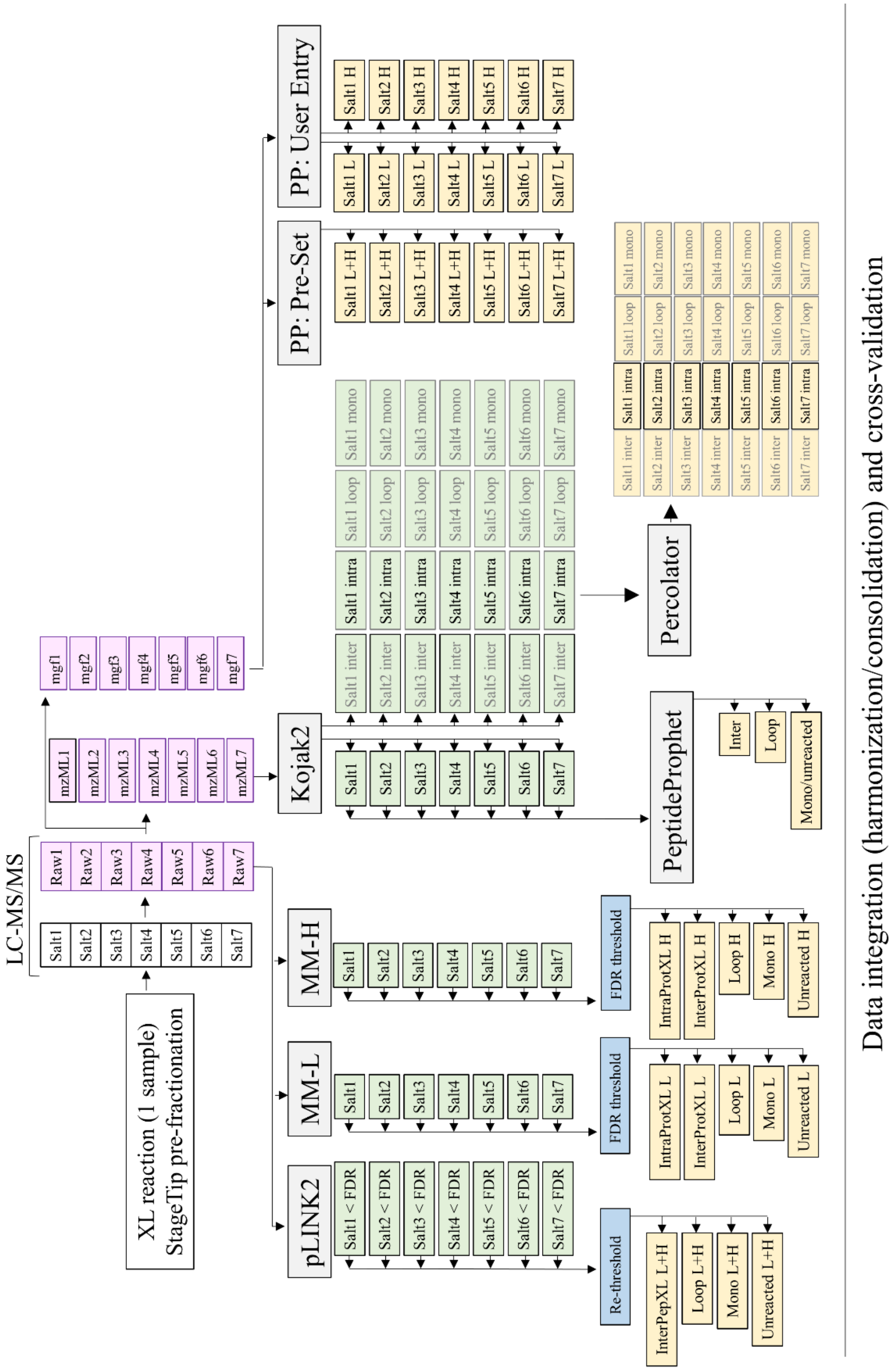


Figure 5.2. Bioinformatics pipeline (data flow) in analysis of a single crosslinked sample fractionated into multiple nLC-MS/MS runs via strong cation exchange pre-fractionation within a StageTip. File types are colored according to category and application: purple = spectral files, green = temporary or pre-thresholding search outputs, yellow = files integrated into

dataset after thresholding. Stagetipping: The crosslinked sample is desalted and pre-fractionated with six steps of 20% CH_3CN/0.5% FA containing increasing concentrations of ammonium acetate followed by a final step of strongly alkaline Buffer "X" (shown here as Salt 1-7). The individual fractions were analyzed by an Orbitrap Velos Pro mass spectrometer. Data flow then depended upon the individual search engine: **pLINK2** could accept instrument .raw input files, which could be searched individually or in batch-mode, allowing me to analyze all instrument files from an experiment in a single run of the program. Multiple crosslinkers could be specified for one search, allowing samples that are crosslinked with stable isotope-encoded reagent pairs (deuterated (H) and undeuterated (L) crosslinker) to be searched simultaneously. The FDR threshold is set prior to the search. pLINK2 consolidates results from all input raw files into one output file per peptide type (unreacted, mono-linked, loop-linked, and inter-peptide crosslinked). We then imposed a secondary threshold upon the pLINK2 output, with a pLINK2 "Score" (essentially a statistical "p" value) threshold of 0.05. **MetaMorpheus** (MM) could also directly accept instrument RAW files, then search files in batch mode. However, only one crosslinking reagent could be searched at a time, therefore samples crosslinked with isotopically labeled crosslinking reagents were searched twice – once with selection of deuterated crosslinking reagent and once with selection of undeuterated crosslinking reagent. MetaMorpheus also consolidates results from each .raw input file into a single output per peptide type. We imposed a Q-value threshold of 0.01 for MetaMorpheus output data after completion of the analysis. **Kojak2** and **Protein Prospector** (PP) require instrument .raw files to be converted to mzML and mgf formats respectively. As with pLINK2, Kojak2 can search files in batch mode with multiple crosslink types selected but returns individual results files for each run file and peptide type, with target and decoy hits reported together. Until recently, scoring and thresholding by Q-value of Kojak2 output was done by a stand-alone program "Percolator", with each result file analyzed separately or manually pooled and analyzed as a single file. Later, after integration of Kojak2 into the Trans Proteomic Pipeline by its authors, thresholding of Kojak2 output was bifurcated (in our lab) between PeptideProphet and Percolator. Mono-linked, loop-linked, and inter-protein crosslinked peptide pairs could by scored and assigned Q-values by PeptideProphet, which combines multiple Kojak2 output files and provide a single output file for each crosslinked peptide type. At the time, the PeptideProphet algorithm was not suited for analysis of intra-protein crosslinked peptide pairs and, as a result, we formulated a method to score and threshold this subset of crosslinked peptides by Percolator. This required separating target and decoy results from individual Kojak2 files and pooling the results into a single target and a single decoy input file for scoring and thresholding by Percolator. A Q-value threshold of 0.01 was imposed on all Kojak2 results. Crosslinked peptide identification by Protein Prospector was run with the Batch-Tag Web application on the Protein Prospector online server. Spectral files were searched individually on Protein Prospector. Some crosslinking reagents are available as preset options on the server, while others have to be manually entered. Isotope reagent pairs that fell into the latter category had to be searched separately. One crosslinked peptide results file is reported per Protein Prospector search. We set a primary FDR threshold of 0.05 and a secondary threshold of SD-E (ScoreDiff–$\log_{10}$(Exp2)) > 5. After thresholding, output files from all programs were integrated via in-house data integration software that harmonized and consolidated the identified crosslinked peptides into an overall crosslinking dataset.

Influence of sample purity on search space

Vaccinia virus purification protocols have largely relied on the absence of organelles and cellular debris, when visualized by transmission electron microscopy (TEM), as a marker for virus purity [100, 102, 504]. They have not changed much over the past 35+ years. After removal of nuclei, infected cell lysates are typically layered over a 36% sucrose cushion. The viral pellet is then resuspended, briefly sonicated, and overlaid on a continuous sucrose, potassium tartrate, or cesium chloride gradient, then purified by rate-zonal centrifugation [102, 610-617]. This two-stage protocol is the standard method for obtaining high yield-high, high biological purity virus. Proteomic purity is typically not reported or required.

We (and others) estimate that 70 – 80 virion proteins are packaged in the virion [101]. One strategy for reducing the bioinformatics search space, and thus increasing confident crosslinked peptide identification, is to reduce the proteomic complexity of the sample. It was therefore necessary to first identify a method that could produce high yield-high proteomic purity vaccinia virus. For budgetary reasons, our "proof of principle" study, above used donated virus from another lab, whose concern was not proteomic purity at all: It was purified over a single potassium tartrate gradient, and was available in comparatively small amounts. Qualitative analysis by mass spectrometry of this material [604] showed a low ratio of packaged : contaminating host and non-packaged viral proteins (**Fig. 5.3A**). Through a series of incremental improvements to the standard vaccinia virus sucrose purification protocol [617], we established an updated purification protocol (**Fig. 5.3B**) that could reproducibly produce high yield/high proteomic purity vaccinia virus. First, after optimizing suspension culture and its infection, and cell breakages to release virus, we started with sufficient viral material to run two sequential sucrose gradients. We optimized the generation, loading, running and harvesting of those gradients. Second, we did MS analysis of each fraction across the viral band harvested from each gradient, and only utilized the fraction with highest purity virus for the subsequent step. This comprised, basically, our optimization approach.

When compared by mass spectrometric analysis to virus prepared by potassium tartrate gradient purification, virus produced by our modified protocol showed a considerably lower numbers of host and viral protein contaminants, with 377 total protein groups identified (302 contaminants, 75 packaged) from the sucrose protocol compared to 1389 protein groups (1314 contaminants, 75 packaged) from potassium tartrate gradient purification (**Fig. 5.3A**). This reduction in ratio of contaminants:packaged proteins from 17.5 to 4.0 represents a 4.4 x 4.4 = 19-fold reduction in bioinformatics search space for every theoretical peptide in the proteome x every other theoretical peptide. Higher proteomic purity also ensures that more time is devoted by the mass spectrometer towards the analysis of packaged virion proteins instead of background contaminants, and is expected to result in an overall increase in detection of crosslinked virion proteins.

Increased crosslinked peptide identification by enrichment

On-line enrichment strategies

Peptide enrichment and depletion strategies are often employed to increase relative concentrations of crosslinked peptides to unreacted peptides prior to data acquisition by the mass spectrometer [577, 604, 618, 619]. By this, we mean that an enrichable chemical group, such as biotin or a phosphate, is incorporated into the crosslinker molecule itself to enable the subsequent enrichment of crosslinker-attached peptides specifically, or higher mass/higher

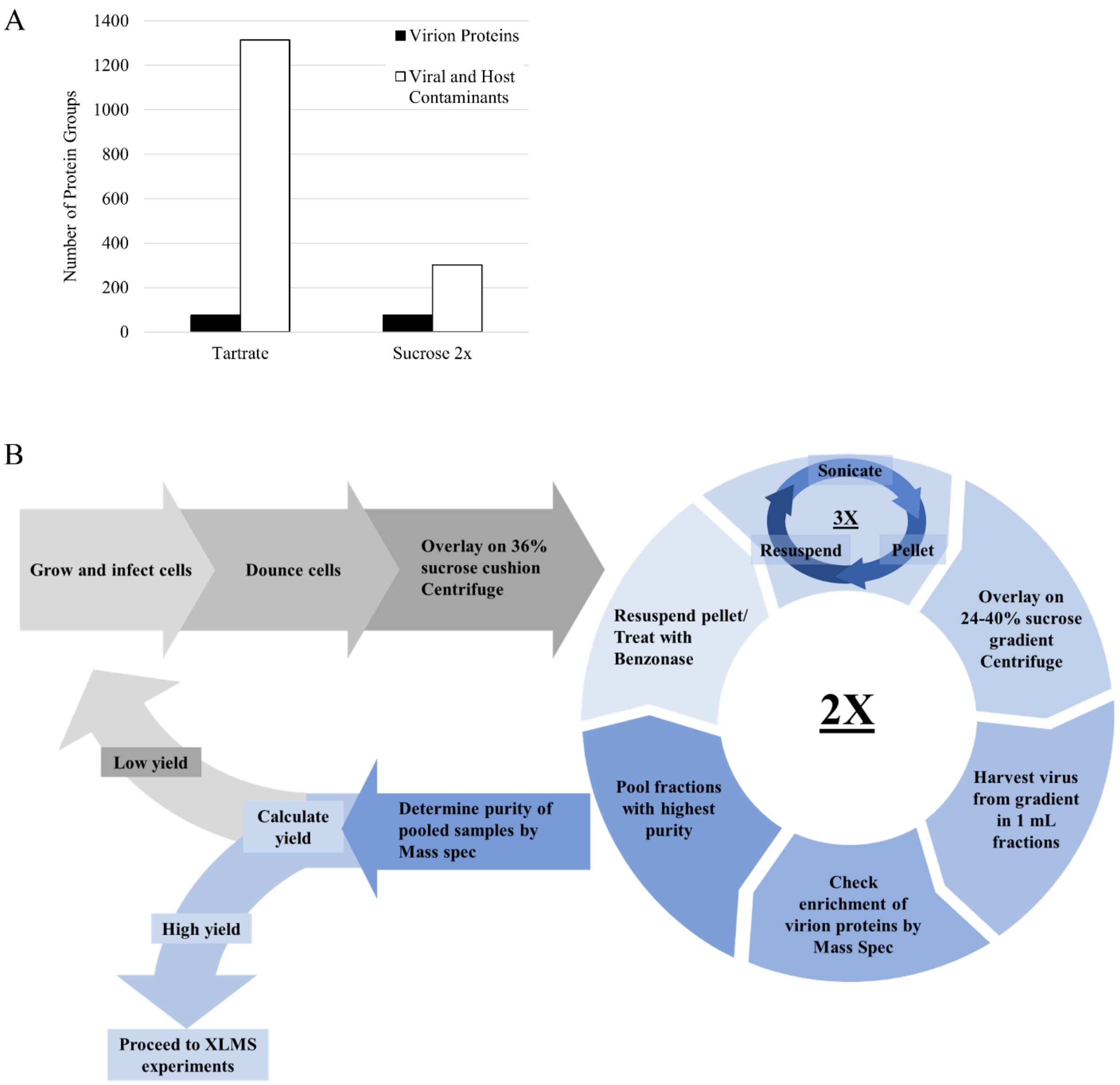

A
1400
1200
1000
800
600
400
200
0
Number of Protein Groups
■ Virion Proteins
□ Viral and Host Contaminants
Tartrate
Sucrose 2x
B
Grow and infect cells
Dounce cells
Overlay on 36% sucrose cushion Centrifuge
Sonicate
3X
Resuspend
Pellet
Resuspend pellet/ Treat with Benzonase
Overlay on 24-40% sucrose gradient Centrifuge
2X
Harvest virus from gradient in 1 mL fractions
Check enrichment of virion proteins by Mass Spec
Pool fractions with highest purity
Determine purity of pooled samples by Mass spec
Calculate yield
Low yield
High yield
Proceed to XLMS experiments

Figure 5.3. Purification of vaccinia virus. (**A**) Number of protein groups identified from vaccinia virus purified by potassium tartrate gradient or after 2 x sucrose gradient purification. Within each category, identified protein counts are separated between known virion proteins and viral and host contaminants. (**B**) Schematic of optimized 2 x sucrose gradient purification protocol for high yield-high proteomic purity vaccinia virus.

charge crosslinked peptides are separated from lower mass/lower charge unreacted, mono-linked and loop-linked peptides by size exclusion or strong cation exchange chromatography.

Standard bottom-up proteomics experiments employ TopN Data-Dependent Acquisition (DDA) strategies, where the top N (number of ions with the highest intensity) from an MS^1 spectra are isolated and fragmented [602, 620]. This approach biases precursor selection towards more abundant and higher intensity peptides, presenting a disadvantage to crosslinked peptides which are typically present in low abundance and often have unfavorable ionization characteristics, leading to an underrepresentation of crosslinked peptides (particularly inter-peptide crosslinked pairs) in the mass spectra [581, 598, 599, 601]. **Figure 5.4** ("Standard XL (DSS)") shows the standard distribution of peptide spectral matches (regular (unreacted), mono-linked, loop-linked, and inter-peptide crosslinked) identified from a crosslinked sample in a single mass spec run. Inter-peptide crosslinks, which include inter-protein and intra-protein crosslinks, represented only 3% of the total peptide spectral matches (PSMs) identified.

Selection strategies can be employed during MS^1 acquisition to increase the number of unique precursors that are selected for fragmentation and reduce bias towards the most abundant ions. One conventional method involves the use of a dynamic exclusion principal, which limits the number of times the same ion can be fragmented within a specified time frame [620, 621]. Dynamic exclusion is routinely used in DDA approaches for proteomic analysis, including in mass spectrometry analysis of crosslinked peptides. Charge state selection (exclusion of +1 charged precursor ions for regular spectra; +2 exclusion for crosslinked spectra) [622] and precursor ion intensity thresholds [620] are also employed in routine MS data acquisition of regular and crosslinked samples.

An additional selection criterion called Mass-Tags can be used during the analysis of crosslinked samples when stable isotope labeled crosslinking reagents are used to crosslink proteins [604, 623]. During data acquisition, a precursor selection strategy is employed to select TopN precursor ions that have a characteristic doublet signal of equally intense peaks, separated by the mass shift represented by isotope labeling, above [575, 604, 623, 624]. This allows for selection of only crosslinked peptides for fragmentation. Ideally, implementing a "Mass-Tags" approach would enrich for the detection of crosslinked peptides by limiting time spent analyzing spectra from non-crosslinked peptides. In practice, while effective at excluding most unreacted peptides from fragmentation, we observed no increase in the number of identified crosslinked peptides when compared to the same sample analyzed without Mass-Tags selection (**Fig. 5.4 "Standard XL (DSS) Mass Tags" vs "Standard XL (DSS)"**) indicating an underlying sensitivity issue that needed to be addressed.

Off-line enrichment strategies

Off-line strategies for the enrichment of crosslinked peptides prior to data acquisition are generally more effective at increasing the numbers of identified crosslinked spectra [577, 601, 618]. We frequently combined sample desalting with strong cation exchange pre-fractionation of crosslinked peptides in StageTips [578] prior to mass spectrometry analysis, which can fortuitously lead to a 2-fold crosslinked peptide enrichment (data not shown). We also routinely applied the DigDeAPR approach to crosslinked samples – this two-step digestion protocol allows for partial pre-depletion of the most abundant proteins, thus reducing the abundance dynamic range of the peptide mixture in the mass spectrometer [577, 604]. DigDeAPR and StageTip pre-fractionation were combined in my hands to further improve separation between unreacted and

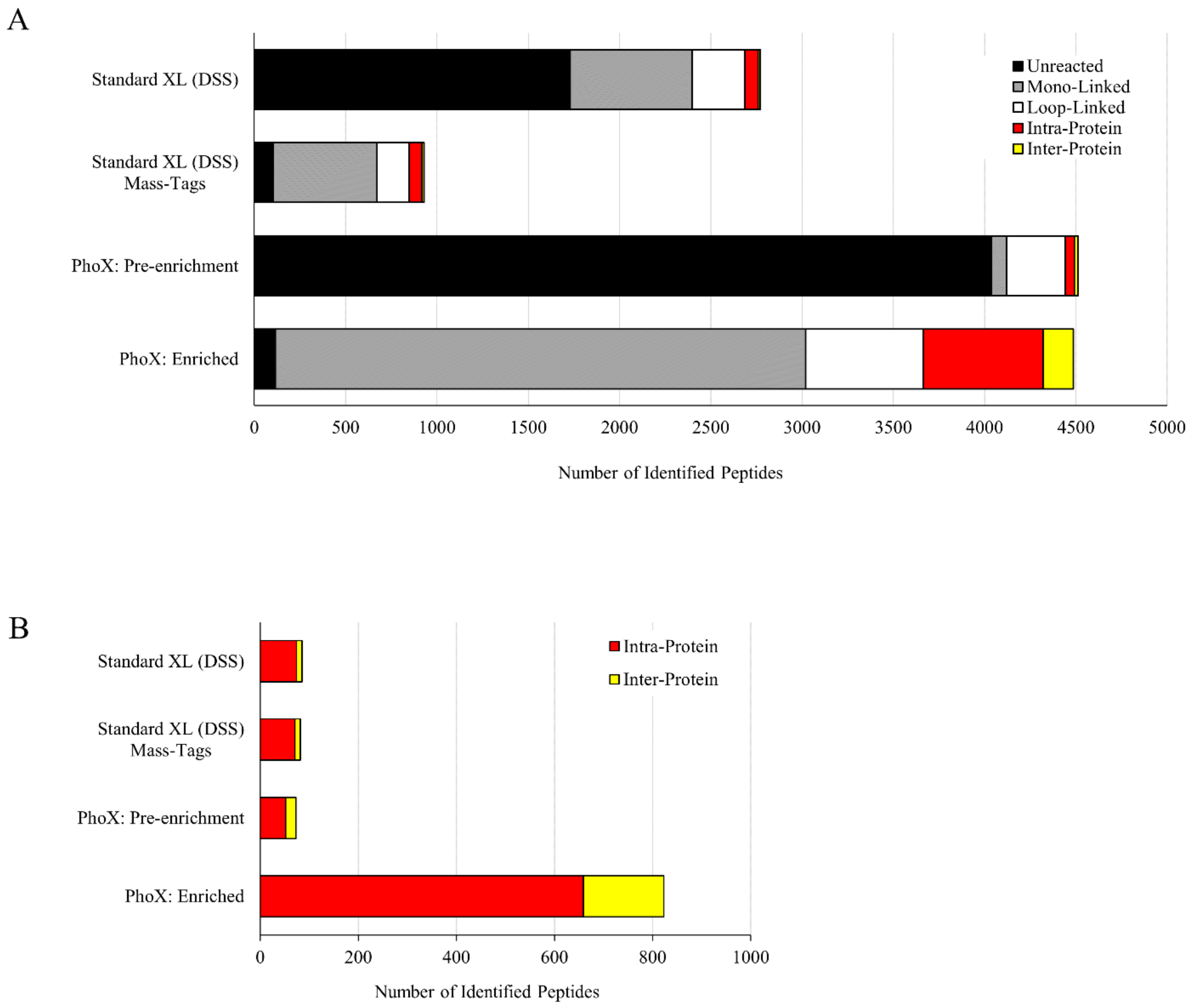


Figure 5.4. Identified crosslinked and unreacted peptides prior to and after on-line and off-line enrichment. Bar graphs showing numbers of peptides identified for (**A**) each peptide type (unreacted, mono-linked, loop-linked, intra-protein crosslinked, and inter-protein crosslinked) and (**B**) intra-protein crosslinked peptides and inter-protein crosslinked peptides only. "Standard XL (DSS)" and "Standard XL (DSS) Mass-Tags" refers to vaccinia virus crosslinked with stable isotope reagent pairs (deuterated and undeuterated DSS crosslinker), analyzed without enrichment and analyzed with on-line (in the mass spectrometer) Mass-Tags enrichment respectively. "PhoX: Pre-enrichment" and "PhoX: Enriched" refer to vaccinia virus crosslinked with PhoX and analyzed without enrichment and after Fe-IMAC enrichment, respectively.

crosslinked peptides. Implementing both approaches does however result in substantial increase in analysis time by increasing numbers of fractions to be analyzed: Pre-fractionation increased the data acquisition time on the mass spectrometer from 4 hours to 28 hours and when combined with DigDeAPR, required up to 56 hours of instrument analysis. Correspondingly increased (**Fig 5.2)** was the number of spectral files to be analyzed - from one instrument .raw file (no pre-enrichment) to seven .raw files with pre-fractionation, and 14 such files with DigDeAPR plus pre-fractionation.

As indicated above, off-line enrichment of crosslinked samples can also be performed using trifunctional crosslinking agents, wherein homobifunctional linkers (typically with NHS-ester chemistry) are synthesized with the addition of an enrichable handle. Historically, enrichable crosslinkers have been relied on biotin:streptavidin or click chemistry approaches, both of which have shown low recovery rates [625-630]. A novel trifunctional crosslinker with NHS-ester chemistry, PhoX, which incorporates an enrichable phosphonic acid tag, was developed recently [631]. Protein crosslinking with PhoX capitalizes on recent advances in high-specificity phosphopeptide enrichment with Fe-NTA immobilized metal affinity chromatography (Fe-IMAC) [619, 632], which also demonstrates a high specificity for phosphonic acid tags, allowing for the enrichment of PhoX crosslinked/derivatized peptides [631]. Pre-treatment of crosslinked peptides with Calf Intestinal Alkaline Phosphatase, which cleaves phospho-ester bonds (present in phosphorylated amino acids) but not carbon-phosphor bonds (present in PhoX), allows for further depletion of unreacted peptides in the sample [631].

We applied this new approach of PhoX crosslinking with Fe-IMAC enrichment to vaccinia virus. Crosslinking followed the standard protocol we employ for all NHS-ester based crosslinkers, with one change - Fe-IMAC enrichment requires a 10-fold greater amount of starting material (200ug vs 20ug). After crosslinking, an aliquot of crosslinked sample was set aside as a pre-enrichment control. After enrichment the pre-and post-enrichment samples were analyzed on an Orbitrap Lumos mass spectrometer (in The Netherlands), and the fragmentation spectra was searched by pLink2. The pre-enrichment control showed no difference in the distribution of identified peptides (unreacted, mono-linked, loop-linked, inter-peptide crosslinked) when compared to a crosslinking experiment carried out with DSS (**Fig. 5.4 "PhoX: Pre-enrichment"**), with inter-peptide crosslinks representing only 1% of identified PSMs. Enrichment with Fe-IMAC resulted in a significant increase in the number of detected crosslinked peptides along with 97% depletion of unreacted peptides (**Fig. 5.4 "PhoX: Enriched"**). The number of inter-peptide crosslinked ions identified from the post-enrichment PhoX sample increased 11-fold compared to both the PhoX pre-enrichment control and compared to samples crosslinked with DSS and analyzed with or without Mass-Tags precursor selection (**Fig. 5.4**), representing 18% of all identified peptides from the post-enrichment sample. The 1.6 fold increase in total number of identified peptides (unreacted and crosslinked) between the "Standard XL (DSS)" sample and the two PhoX samples was due to instrument differences (the DSS sample was analyzed on an Orbitrap Velos Pro while the PhoX samples were analyzed on an Orbitrap Lumos) and did not represent solely the increased inter-peptide crosslink identifications seen in the post-enrichment PhoX sample, particularly as the number of crosslinked peptides identified in the pre-enrichment PhoX sample (also analyzed on an Orbitrap Lumos) was equivalent to the number of crosslinked peptides identified in the DSS sample.

The total number of unique PhoX-crosslinked peptide pairs (by protein accession, XL position and crosslinked peptide Mr) increased from 42 pairs identified pre-enrichment to 325 crosslinked peptide pairs identified post-enrichment (data not shown). By removing unreacted peptides from the sample prior to analysis, the mass spectrometer can reach deeper into the crosslinking proteome, allowing for the identification of lower abundance crosslinked peptide pairs. The proteome depth achieved with an enriched sample could not be reproduced by simply reanalyzing an unenriched sample multiple times by the mass spectrometer, as high intensity-high abundance unreacted and crosslinked peptides will be preferentially selected for fragmentation by the mass spectrometer in every run, resulting in a steady increase in total CSMs detected without a reciprocal increase in unique crosslinked peptide pairs.

Effect of virion protein solubilization strategies on peptide identification and missed cleavages

Complete digestion of proteins to peptides is a prerequisite for improved fragmentation of crosslinked peptides within the mass spectrometer and accurate deconvolution of the resulting fragment spectra. Crosslinked peptides are typically heavier than regular peptides, due to the combined mass of the two peptides and the crosslinker. This is disadvantageous for fragmentation as higher mass species may fall outside the mass range of the mass spectrometer and in Orbitrap mass spectrometers mass accuracy and resolution decrease as m/z increases [581, 600-602]. Bottom-up proteomics experiments typically employ trypsin for sample digestion due to its high specificity for lysine and arginine residues (cleavage occurs C-terminal to the residue), resulting in peptides with average lengths of ~14 amino acids [633] - well within the mass range of the mass spectrometer. LysC, an endoproteinase that cleaves C-terminal to arginine residues, is often used in conjunction with trypsin to improve cleavage efficiency [596, 603, 634, 635]. However, even when both enzymes are employed in tandem, poor protein solubilization leading to inaccessible cleavage sites, the presence of proline in the P1' position, negatively charged amino acids in the P1' position, or protease inhibitors can result in incomplete digestion and produce longer length peptides [596, 636-638].

Crosslinking reagents typically rely on NHS-ester chemistry and react with lysine side chains to form covalent bonds. The presence of the crosslink, however, prevents cleavage by trypsin at the crosslinked lysine residue [600] and thus, the majority of crosslinked peptides (when cleaved by trypsin) have at least one missed cleavage site per peptide, resulting in even larger combined masses for the crosslinked peptide pair. Missed cleavage sites can also have a detrimental impact on the bioinformatics search space [597, 600] with the total possible peptide combinations increasing with the missed cleavage rate. Reducing the overall rate of missed cleavages during sample preparation is critical for maximizing the identification of crosslinked peptides. One strategy towards improving protein digestion and reducing missed cleavage sites is by optimizing protein solubilization/denaturation conditions. This is especially important for a virus – a solid state organism designed by nature to be maximally insoluble and protease-resistant.

The effectiveness of various denaturation methods on solubilizing proteins is therefore often sample, tissue, and organism dependent. In view of this, we first evaluated the effects of different protein solubilization methods on the digestion efficiency and proteome depth achieved by trypsin and LysC digestion of purified vaccinia virus, in the absence of crosslinker. Included in our screen were denaturation methods commonly employed in bottom-up proteomics experiments: denaturation with chaotropes (urea and guanidine hydrochloride (GuHCl)), filter-aided sample preparation (FASP), chaotrope-free solubilization (RapiGest and minimal proteomic sample preparation (mPOP)), new in-house digestion methods (pressure and steam assisted sample preparation (mPOT), pressure assisted denaturation with a barocycler, and on-glass digestion), and various combinations of these approaches (**Table 5.1**). Twice the amount of protein (30 ug vs 15 ug) was used for all experiments that incorporated FASP-based denaturation due to intrinsic sample recovery issues [596]. The effect of reducing agent on digestion efficiency was also evaluated. Denaturation and partial digestion by 70% formic acid with cyanogen bromide [639, 640] with subsequent trypsin and LysC digestion, although a very effective denaturation approach, was excluded from the analysis as it can cause unwanted chemical modifications to peptides.

Purified vaccinia virus was solubilized, digested, and analyzed on an Orbitrap Velos Pro mass spectrometer. All acquired mass spectra were searched using MetaMorpheus under a common set of search parameters (described in **Materials and Methods**). Up to two missed cleavage sites per peptide were allowed. Results were evaluated in two parts: proteome depth achieved based on the total number of identified PSMs and the number of unique non-redundant peptides identified, and completeness of digestion based on the distribution of peptides with 0-missed, 1-missed, and 2-missed trypsin and LysC cleavage sites. Enzyme cleavage rules are described in **Materials and Methods**. Experiments were conducted in duplicate and the averaged results for the 21 experimental conditions are summarized in **Figure 5.5**.

Chaotropic agents increase missed cleavages in peptides

Protein solubilization with chaotropic agents are some of the most commonly employed solubilization methods for proteomics. Solubilization with 8 M urea and reducing agent (10 mM TCEP) has been our primary approach for solubilizing vaccinia virus proteins, particularly for crosslinking experiments [604]. Comparison of the results achieved across all 21 experimental conditions revealed a startling trend – chaotropic agents, when present in the digestion buffer at concentrations reported to be fully compatible with trypsin and LysC, displayed suppressive effects on the cleavage activity of the enzymes (**Figure 5.5, "Urea (TCEP)", "Urea", "Urea (0.5 M)", "GuHCl", "GuHCl+Urea")**. Solubilization with urea and TCEP resulted in low numbers of identified peptides (both total and non-redundant) (**Fig. 5.5 A; "Urea (TCEP)"**) and high missed cleavage rates, with > 60 % of peptides having at least one miss cleaved site (**Fig. 5.5 B, "Urea (TCEP)"**). When vaccinia virus solubilization with 8 M urea was repeated without the addition of reducing agent, ("Urea"), the missed cleavage rate decreased, such that only 40% of peptides had 1-missed or 2-missed cleavage sites (**Fig. 5.5 B, "Urea"**). The number of total PSMs and non-redundant peptides detected nearly doubled (**Fig. 5.5 A, "Urea"**). The effect of reducing agent on digestion of vaccinia virus proteins will be discussed below. Further dilution of urea in the digestion buffer (from 1 M urea to 0.5 M urea) corresponded to a moderate increase in peptide identifications and a moderate decrease in missed cleavages (**Fig. 5.5**, **"Urea (0.5 M)**") and supported our conclusion that cleavage activity of trypsin and LysC was compromised in the presence of reportedly acceptable concentrations of urea.

Protein solubilization with GuHCl also resulted in high rates of miss-cleaved sites (**Fig. 5.5 B, "GuHCl"**). This was further exacerbated when urea was added to the samples as an additional chaotrope after initial solubilization with GuHCl (**Fig. 5.5 B, "GuHCl+Urea"**), with the rate of 0-missed peptides decreasing from 50% to 35%.

Chaotrope-free in-solution solubilization and digestion

Other protein solubilization methods tested, where digestion with trypsin and LysC was carried out in-solution (**Table 5.1**), yielded lower rates of miss-cleaved peptides (**Fig. 5.5 B, "Rapigest", "mPOB", "mPOP", "mPOT", "mPOT-Freeze"**) and overall high rates of peptide identification (**Fig. 5.5 A, "Rapigest", "mPOB", "mPOP", "mPOT", "mPOT-Freeze"**).

Table 5.1. Protein solubilization conditions.

Name	Protein Amount	Solubilization Conditions	Digestion Phase	Digestion conditions
Urea (TCEP)	15 µg	8 M urea, 10 mM TCEP, 100 mM TEAB pH 8.5	In solution	LysC in 6 M urea, trypsin in 1 M urea, 100 mM TEAB pH 8.5
Urea	15 µg	8 M urea, 100 mM TEAB pH 8.5	In solution	LysC in 6 M urea, trypsin in 1 M urea, 100 mM TEAB pH 8.5
Urea (0.5 M)	15 µg	8 M urea, 100 mM TEAB pH 8.5	In solution	LysC in 0.5 M urea, trypsin in 0.5 M urea, 100 mM TEAB pH 8.5
GuHCl	15 µg	6 M GuHCl 100 mM TEAB pH 8.5, 3 x boil/sonicate	In solution	LysC in 3 M GuHCl, trypsin in 0.3 M GuHCl, 100 mM TEAB pH 8.5
GuHCl+Urea	15 µg	6 M GuHCl, 100 mM TEAB pH 8.5, 3 x boil/sonicate; add 8 M urea	In solution	LysC in 3 M GuHCl/4 M urea, trypsin in 0.3 M GuHCl/0.4 M urea, 100 mM TEAB pH 8.5
Rapigest	15 µg	0.2% Rapigest, 100 mM TEAB pH 8.5, 75°C	In solution	LysC, trypsin in 0.1% Rapigest, 100 mM TEAB pH 8.5
mPOB	15 µg	100 mM TEAB pH 8.5, freeze, pressure cycle with a barocycler at 90°C	In solution	LysC, trypsin 100 mM TEAB pH 8.5
FASP (30k)	30 µg	Standard FASP protocol (4% SDS, 100 mM TEAB pH 8.5, wash with 8 M urea in 100 mM TEAB pH 8.5)	Vivacon 30K filter	LysC, trypsin 100 mM TEAB pH 8.5
FASP (10k)	30 µg	Standard FASP protocol, (4% SDS, 100 mM TEAB pH 8.5, wash with 8 M urea in 100 mM TEAB pH 8.5)	Vivacon 10K filter	LysC, trypsin 100 mM TEAB pH 8.5
FASPu (10k)	30 µg	Modified FASP protocol (SDS excluded, urea in 100 mM TEAB pH 8.5)	Vivacon 10K filter	LysC, trypsin 100 mM TEAB pH 8.5
mPOP	15 µg	H2O, freeze, boil	In solution	LysC, trypsin 100 mM TEAB pH 8.5
mPOP+FASP (30k)	30 µg	H2O, freeze, boil, FASPt (30k) protocol	Vivacon 30K filter	LysC, trypsin 100 mM TEAB pH 8.5
mPOP+FASP (10k)	30 µg	H2O, freeze, boil, FASPt (10k) protocol	Vivacon 10K filter	LysC, trypsin 100 mM TEAB pH 8.5
mPOT	15 µg	10 mM TEAB pH 8.5, steam denature under constant pressure (100°C)	In solution	LysC, trypsin 100 mM TEAB pH 8.5
mPOT+FASP (10k)	30 µg	10 mM TEAB pH 8.5, steam denature under constant pressure (100°C), FASPt (10k) protocol	Vivacon 10K filter	LysC, trypsin 100 mM TEAB pH 8.5
mPOT-Freeze	15 µg	10 mM TEAB pH 8.5, freeze, steam denature under constant pressure (100°C)	In solution	LysC, trypsin 100 mM TEAB pH 8.5
mPOT-Freeze+FASP (10k)	30 µg	10 mM TEAB pH 8.5, freeze, steam denature under constant pressure (100°C), FASPt (10k) protocol	Vivacon 10K filter	LysC, trypsin 100 mM TEAB pH 8.5
Glass	15 µg	100 mM TEAB pH 8.5, dry on glass	Glass	LysC, trypsin 100 mM TEAB pH 8.5
Glass (steam)	15 µg	100 mM TEAB pH 8.5, dry on glass. steam denature under constant pressure (100°C),	Glass	LysC, trypsin 100 mM TEAB pH 8.5
Glass (humid)	15 µg	10 mM TEAB pH 8.5, humidified box	Glass	LysC, trypsin 100 mM TEAB pH 8.5
Glass (deactivated)	15 µg	10 mM TEAB pH 8.5, deactivated glass	Glass	LysC, trypsin 100 mM TEAB pH 8.5

Samples digested in Rapigest (a cleavable, mass spectrometry compatible detergent) yielded the highest rates of missed cleavage sites from all five in-solution methods (**Fig. Dig B**). “mPOB”, “mPOT”, and “mPOT-Freeze” methods were adapted from the principles of the “mPOP” method (designed for single cell proteomics) where samples are first frozen in water, then heated to 90°C to solubilize proteins, and then digested at 37°C with trypsin and LysC in 100 mM TEAB [641]. In our “mPOB” method vaccinia virus was resuspended in water and frozen at -80°C, denatured by pressure cycling at 90°C in a barocycler, and then digested at 37°C

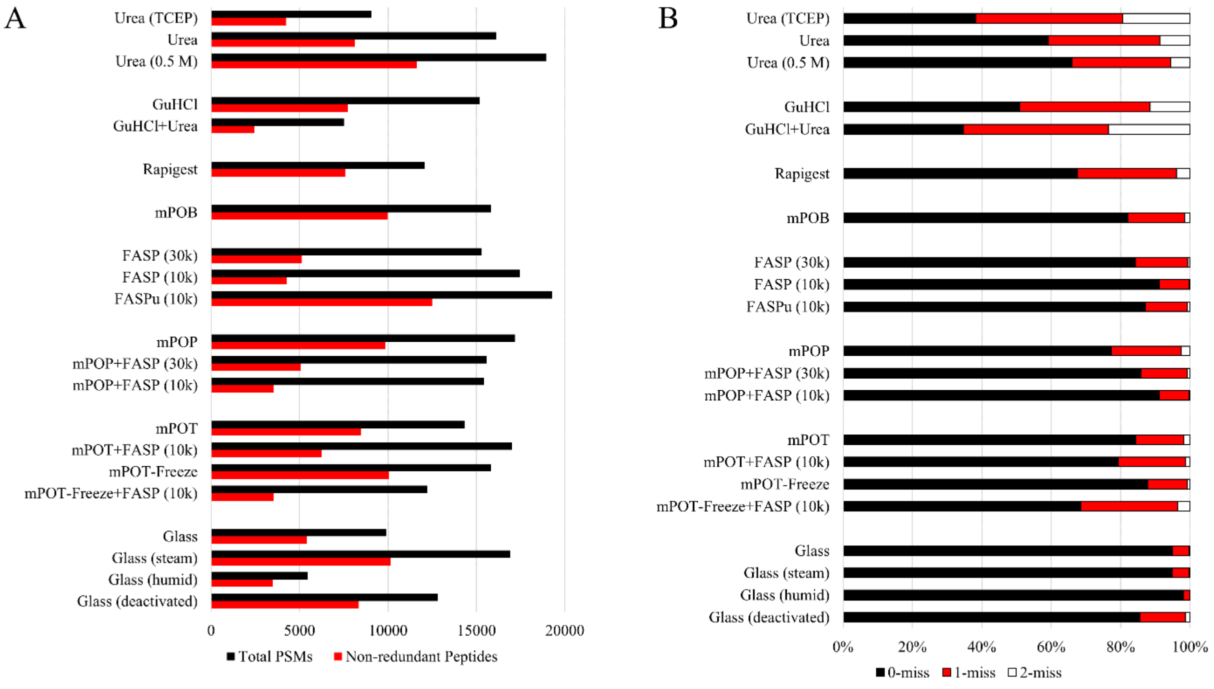


Figure 5.5. Effects of different protein solubilization methods on the number of identified proteins and missed cleavage rate. For each method, samples were prepared in duplicate and averaged results reported. (**A**) Bar graph shows the total number peptide spectral matches (PSMs) and the total number of non-redundant (unique) peptides identified from each method. (**B**) Bar graph shows the distribution of peptides with 0-miss, 1-miss, and 2-miss cleaved sites for each protein solubilization method.

with trypsin and LysC in 100 mM TEAB. Our “mPOT” and “mPOT-Freeze” methods used similar sample preparation techniques, with the exclusion of the freeze step for the former. Samples in 10 mM TEAB (frozen or unfrozen) were steam denatured at 100 °C under constant 10 psi pressure, and then digested at 37°C with trypsin and LysC in 100 mM TEAB. Overall, peptide identifications between the “mPOB”, “mPOP”, “mPOT”, and “mPOT-Freeze” methods were highly comparable for identifications of total PSMs and non-redundant (unique) peptides (**Fig. 5.5 A**) but differed in cleavage efficiency. Solubilization by “mPOT-Freeze” yielded the highest cleavage efficiency – with 88 % of peptides having no missed cleaved sites at all, compared to 77%, 82%, and 84% for “mPOP”, “mPOB”, and “mPOT” respectively (**Fig. 5.5 B**).

Modification of the filter-aided sample preparation protocol to improve peptide identification

The FASP protocol [642], and derivations thereof (eFASP, MED FASP, PVP-FASP [643-645], involves solubilizing of proteins by boiling in 4% sodium dodecyl sulfate (SDS), after which samples are added to 30k cutoff centrifugal ultrafiltration units, washed with 8 M urea, washed with digestion buffer, concentrated and digested in the ultrafiltration units. Peptides are recovered by spin-elution. We typically only employ the standard FASP protocol, with LysC and trypsin digestion, when processing larger amounts of protein for mass spectrometry analysis, as non-specific binding of peptides to the ultrafiltration unit can reduce peptide recovery from samples with low amounts of protein [596, 643]. To account for peptide loss to non-specific binding for the purposes of these experiments, we doubled the amount of protein digested for all solubilization conditions where the digestion step was carried out in Vivacon ultrafiltration units (**Table 5.1**).

Initially, we compared sample preparation with the original FASP protocol "FASP (30k)" to two in-house modifications of the FASP protocol: "FASP (10k)" with a 10k molecular weight cutoff ultrafiltration unit and "FASPu (10k)" where the SDS solubilization step was excluded and proteins were solubilized directly in 8 M urea prior to addition to the 10k ultrafiltration units. All subsequent wash, digestion, and elution steps remained constant between the three methods. Cleavage efficiency was high for all three samples, with "FASP (10k)" slightly outperforming "FASPu (10k)" with 91% of peptides containing no missed cleaved sites compared to 87% (**Fig. 5.5 B**). The original FASP protocol "FASP (30k)" yielded the highest missed cleavage rate of the three samples, with only 84% of peptides containing no missed cleavage sites (**Fig. 5.5 B**). Although total peptide identifications were high for the "FASP (30k)" and "FASP (10k)" methods, unique peptide identifications were low (**Fig. 5.5 A**). This was not the case however, when samples were prepared by "FASPu (10k)". When compared to "FASP (30k)" (which slightly outperformed "FASP (10k)" in peptide identification), we observed a 26% increase in total PSMs and a 143% increase in non-redundant peptides identified by "FASPu (10k)" (**Fig. 5.5 A**).

We also tested combinations of "mPOT" and "mPOP" with FASP protocols but observed no overall significant improvements in results (**Fig. 5.5, "mPOP+FASP (30k)", "mPOP+FASP (10k)", "mPOT+FASP (10k)", "mPOT-Freeze+FASP (10k)"**).

Approaching zero miss cleaved peptides with on-glass denaturation and digestion

We benchmarked the digestion efficiency of trypsin and LysC with a new in-house method of "on-glass" protein solubilization and digestion, targeted towards the preparation of low-abundance samples (**Table 5.1**). This work is still under development but showed promising results with very high trypsin and LysC cleavage efficiency (**Fig. 5.5 B, "Glass", "Glass**

(steam)", "Glass (humid)", "Glass (deactivated)"). When combined with the steam denaturation method used for "mPOT", on-glass denaturation and digestion yielded 95% cleavage efficiency, with peptide identification rates that were competitive with "mPOT-Freeze" and "FASPu (10k)" methods (**Fig. 5.5, "Glass (steam)"**). However, this method requires further optimization to improve reproducibility of sample recovery (data not shown).

Cleavage efficiency is reduced in the presence of TCEP

When comparing vaccinia virus samples solubilized and digested in the presence of urea and urea with TCEP, we observed a substantial decrease in cleavage efficiency and peptide identification when TCEP was included (**Fig. 5.5, "Urea (TCEP)", "Urea"**). This seemed counterintuitive as vaccinia virus encodes three thiol oxidoreductases, intra- and inter-molecular disulfide bond formation is essential for normal virion morphogenesis [646]**,** [143, 210, 550, 647], and almost all standard proteomics sample preparation methods incorporate disulfide bond reducing agents prior to digestion and sample analysis. To determine whether the decrease in digestion efficiency and peptide identification was limited to samples solubilized in urea or was indicative of a systematic suppressive effect on trypsin and LysC digestion of vaccinia virus proteins, we compared results obtained from samples prepared with "mPOP", "Urea", "mPOT-Freeze", and "FASP (10k)" protocols, with and without TCEP. The results of the digestion experiments are summarized in **Figure 5.6**. For the "mPOP", "Urea", and "mPOT-Freeze" sample preparation methods, there was a stark decrease in overall peptide identifications when TCEP was included in the digestion buffer, with a 44-64% decrease in total PSMs identified (**Fig 5.6A**), with a similar decrease in non-redundant (unique) peptide identifications (**Fig 5.6B**). This decrease in peptide identifications correlated with reduced cleavage efficiency, with a 38-55% increase in peptides with at least one missed cleavage site when TCEP was included (**Fig 5.6C**).

Only the “FASP (10k)” samples showed no differences in peptide identification or cleavage efficiency (**Fig 5.6**), since TCEP was removed from the sample after disulfide bond reduction by spinning down and washing the retentate, as is standard for FASP protocols. This suggests that the presence of TCEP in the digestion buffer negatively impacts cleavage efficiency of trypsin and LysC with respect to vaccinia virus proteins and we decided to exclude TCEP from the solubilization and digestion steps of later experiments with vaccinia virus.

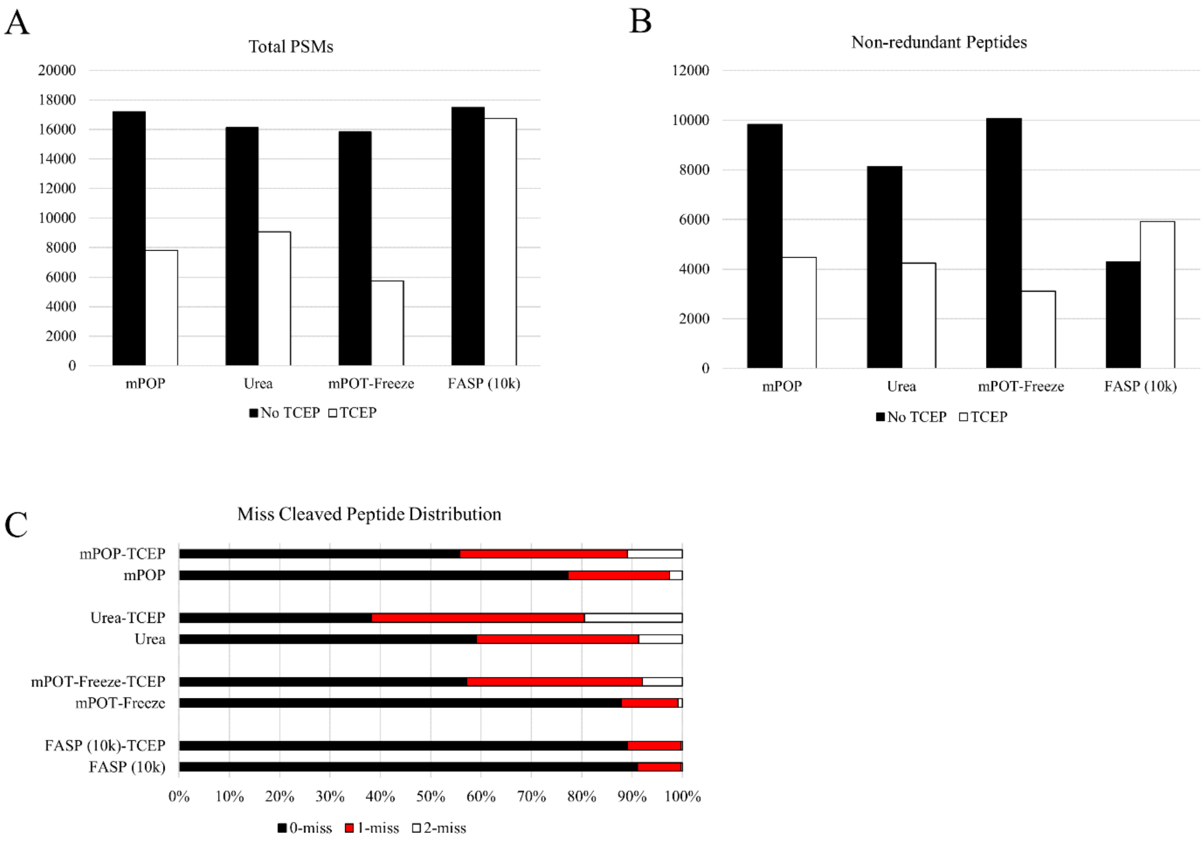


Figure 5.6. Effects of reducing agent (TCEP) on peptide identifications and missed cleavage rates under different protein solubilization conditions. Samples were prepared in duplicate and averaged results are reported. The total number peptide spectral matches (PSMs) (**A**) and the total number of non-redundant (unique) peptides (**B**) identified when TCEP was included or excluded in the samples during solubilization. (**C**) Bar graph shows the distribution of peptides with 0-miss, 1-miss, and 2-miss cleaved sites for each protein solubilization method when TCEP was included or excluded.

Combining optimized solubilization protocols with XLMS

Based on our observations from the solubilization experiments, we decided to compare the effect of sample preparation by the “mPOT-Freeze” and “Urea” protocols on crosslinked peptide identification rates. For each crosslinking reagent tested, purified vaccinia virus was crosslinked then divided into two aliquots and prepared for analysis as described in **Figure 5.7A**. By aliquoting the samples for digestion after crosslinking we could ensure that the differences observed in crosslinked peptide identification were due to the solubilization conditions, not the crosslinking reaction itself. Intact vaccinia virus was crosslinked with DSS or BS3 (**see Table 1.1 Crosslinking Reagents**), solubilized by either protocol, analyzed by the mass spectrometer, and the resulting mass spectra searched using pLINK2. With both BS3 and DSS crosslinked samples, aliquots prepared by the “mPOT-Freeze” protocol yielded substantially greater numbers of inter-peptide crosslink identifications than aliquots prepared by “Urea” denaturation (**Fig 5.7B**). We then extended this analysis to other crosslinking reagents and virus pretreatment conditions. Virions were either crosslinked intact, as in **Figure 5.7B**, or uncoated with NP40/TCEP followed by a brief wash step to remove the uncoating solution and virion membrane proteins. Samples were then prepared as described in **Figure 5.7A.** Sample preparation by the “mPOT-Freeze” protocol yielded substantially greater numbers of crosslinked peptide identifications, regardless of the crosslinking reagent or virus pretreatment condition (**Fig 5.7C**). Overall, this clearly demonstrated that improving protein solubilization and cleavage efficiency resulted in increased crosslinked peptide identification.

Crosslinked peptide identification improves with a more powerful mass spectrometer

Mass spectrometry technologies have undergone significant evolution over the years, with major hardware and software advances allowing for higher mass accuracy, greater speed,

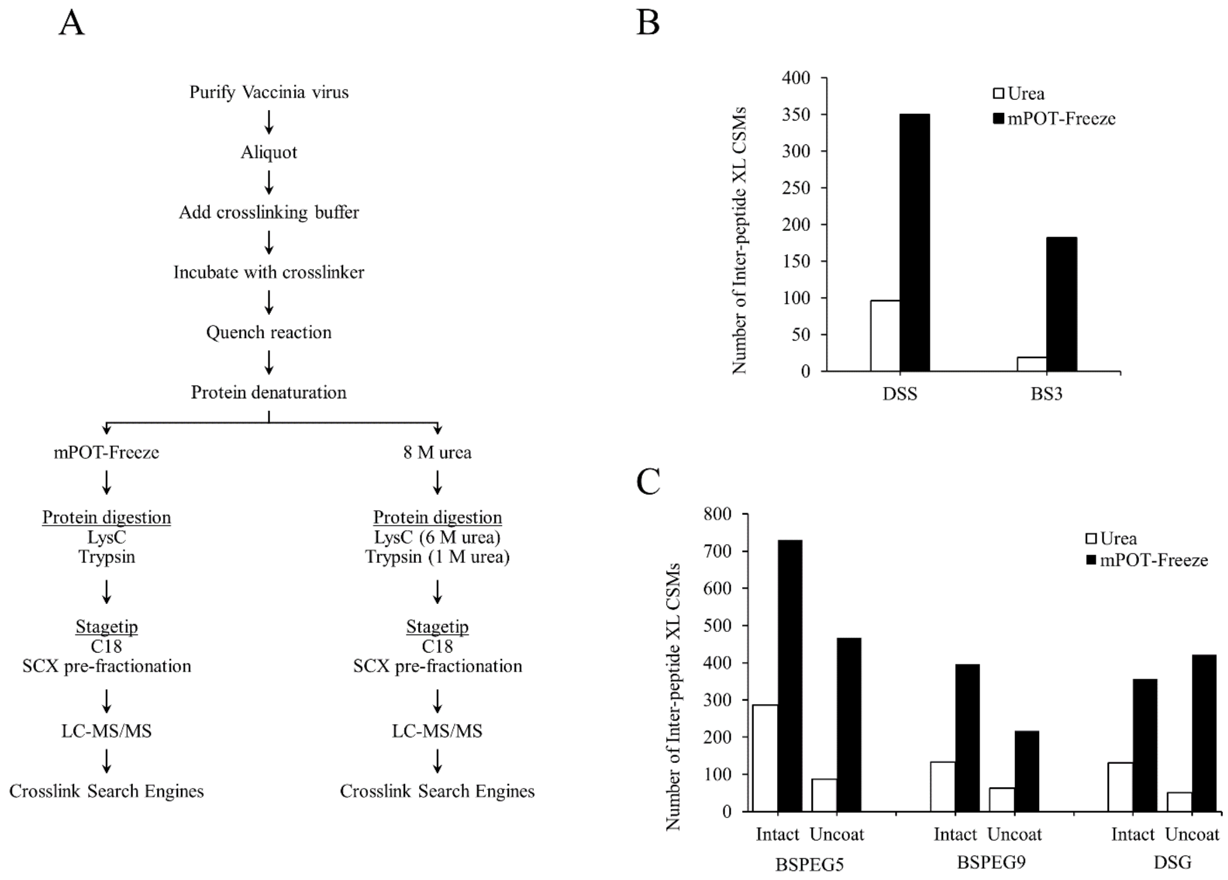


Figure 5.7. Protein solubilization methods affect crosslinked peptide identifications. (**A**) The general workflow for comparing the effects of two different protein solubilization methods, “Urea” and “mPOT-Freeze,” on crosslinked peptide identification rates. Purified vaccinia virus (either intact or after brief uncoating with NP40/TCEP, followed by a wash step to remove the uncoating reagents and solubilized membrane proteins) were crosslinked with various crosslinking reagents (DSS, BS3, BSPEG5, BSPEG9, DSG). After the reaction was quenched, samples were each divided into two aliquots and solubilized by either “Urea” or “mPOT-Freeze” followed by overnight digestion with LysC and then by overnight digestion with trypsin. Peptides were desalted and eluted by strong cation exchange (SCX) pre-fractionation, subjected to mass spectrometry analysis, and crosslinked peptides were identified by pLINK2. (**B**) Number of inter-peptide crosslinked pairs identified by pLINK2 from intact vaccinia virus crosslinked with DSS or BS3 and prepared by “Urea” or “mPOT-Freeze” methods. (**C**) Number of inter-peptide crosslinked pairs identified by pLINK2 from intact or uncoated vaccinia virus crosslinked with BSPEG5, BSPEG9, and DSG and prepared by “Urea” or “mPOT-Freeze” methods.

increased resolution, and sensitivity, all resulting in increased peptide identification rates and more comprehensive proteome coverage [648-652]. Required sample amounts and instrument run times have also decreased, without any loss in peptide identifications. Mass spectrometers in the early 2000s could achieve proteome depths of 1,500-2,000 protein families with 48-68 hours of analysis [653, 654]. After decade of improvement, the same proteome depth could be achieved in a single 4-hour elution with the Orbitrap Velos Pro, the instrument we routinely use for proteomics research (including XLMS).

The current state-of-the-art mass spectrometers, the Orbitrap Eclipse and the newly released Orbitrap Ascend have surpassed the Velos Pro in performance and data acquisition. The Orbitrap Eclipse offers higher sensitivity and specificity, with a 2-fold higher mass range, 9-fold higher resolution, and a 10-fold higher scan speed, resulting in faster, more accurate peptide identification. Higher scan speeds allow for the selection of greater numbers of lower intensity peptides for fragmentation (the TopN limit is higher), while higher resolution and mass accuracy allow for more accurate interpretation of the mass spectra and more accurate analysis of larger peptides with fewer false positives. These features are highly beneficial for XLMS experiments. With method optimization, the Orbitrap Eclipse can confidently identify more than 4,400 or more protein families from a single 2-hour elution [651].

We evaluated the performance of the Orbitrap Eclipse in crosslinked peptide identification compared to the Orbitrap Velos Pro. Intact vaccinia virions were crosslinked with DSS or BSPEG5, digested overnight with LysC and Trypsin, analyzed by both instruments, and the resulting data files were searched by pLINK2. Samples were run on the Orbitrap Velos Pro with 4-hour nLC gradient times with our previously described method [604]. Samples were then run on the Eclipse three times (each) with 2-hour gradients, testing different HCD collision modes – a single collision energy “Fixed” (similar to the Velos Pro), three multiplexed energies “Stepped”, and on-the-fly optimized collision energy “Assisted”.

We observed a 7 - 10 fold increase in the number of MS^1 scans detected by the Eclipse compared to the Velos Pro (in half the time), in line with the 10-fold increase in scan speed, with a 2 - 3 fold increase in MS^2 scans (**Table 5.2**). For both the BSPEG5 and DSS crosslinked samples, this led to a 200 - 400% increase in peptide identifications (**Table 5.2)**. Similar improvements were observed when crosslinked peptide detection was evaluated. Data acquisition on the Eclipse using “Stepped” collision mode resulted in the highest number of crosslinked peptide identifications, with 516 inter-peptide crosslinks identified from the BSPEG5 crosslinked sample, a 153% increase compared to the Velos Pro (**Fig. 5.8A**). More importantly, this correlated to a substantial increase in unique inter-peptide crosslink identifications, with 165 unique inter-peptide crosslink pairs detected by the Eclipse, compared to the 41 unique pairs detected by the Velos Pro (**Fig. 5.8B**). Data acquisition by the Eclipse in “Fixed” and “Assisted” collision modes also resulted in > 3-fold increases in unique crosslinked peptide pair identifications (**Fig. 5.8B**).

Table 5.2. Comparison in MS^1 and MS^2 scans and total identified PSMs by two orbitrap mass spectrometers. The Orbitrap Velos Pro used “Fixed” HCD collision energy. Three different HCD collision modes were tested for the Orbitrap Eclipse.

Instrument	BSPEG5	BSPEG5	BSPEG5	DSS	DSS	DSS
	(MS^1 scans)	(MS^2 scans)	(Total PSMs)	(MS^1 scans)	(MS^2 scans)	(Total PSMs)
Velos Pro	2425	21113	7580	2556	19962	5326
Eclipse-Fixed	17485	60362	28659	19353	55453	27034
Eclipse-Stepped	15766	63335	28140	17779	58332	25271
Eclipse-Assisted	13832	47272	23084	16442	42998	19654

An even greater increase in inter-peptide crosslink identification was seen when the DSS sample was analyzed in “Stepped” HCD collision mode on the Eclipse, with 260% more crosslinked peptides identified compared to when the sample was run on the Velos Pro (**Fig. 5.8A**). Similar to the BSPEG5 sample, there was a 4-fold increase in the number of unique inter-peptide crosslinks identified by the Eclipse (66 peptide pairs vs. 15 peptide pairs by the Velos Pro) (**Fig. 5.8B**). Increases in the numbers of inter-peptide crosslinks detected were also observed when the DSS crosslinked sample was analyzed on the Eclipse with “Fixed” and “Assisted” modes (**Fig. 5.8**).

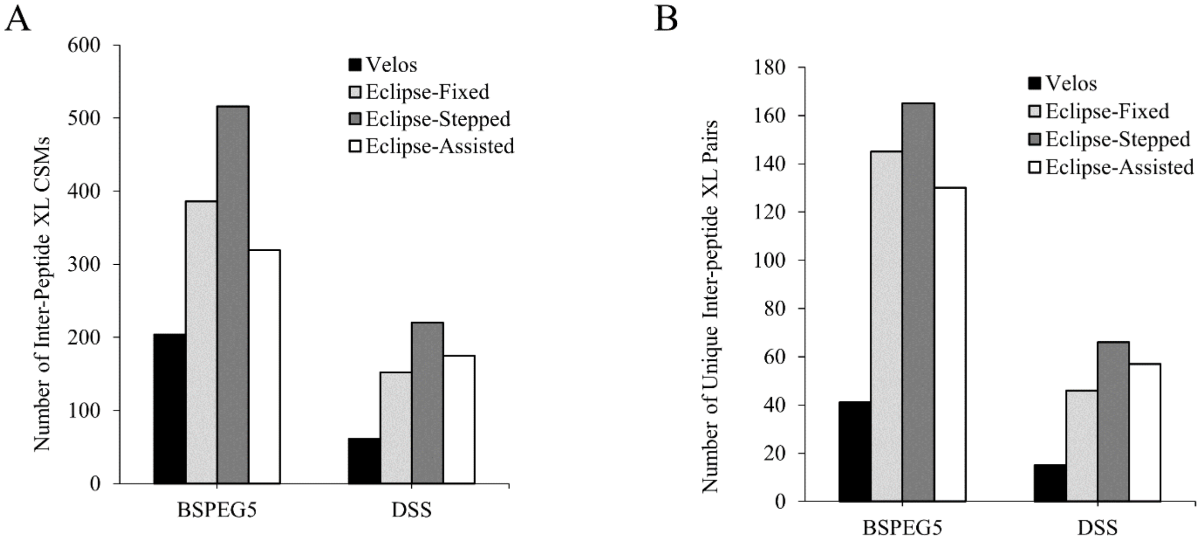


Figure 5.8. Crosslinked peptide identification depends on powerful mass spectrometers and HCD collision mode. Intact vaccinia virus was crosslinked with BSPEG5 or DSS, digested to peptides, desalted and analyzed on an Orbitrap Velos Pro with “fixed” HCD collision energy, followed by analysis of the same samples on an Orbitrap Eclipse with three different HCD collision modes: “Fixed”, “Stepped”, and “Assisted”. Spectral files were analyzed by pLINK2 and the (**A**) total number of identified inter-peptide crosslinked pairs and (**B**) number of unique inter-peptide crosslinked pairs reported.

Overall, we observed substantial increases in both the total number of crosslinked peptides detected and the number of unique crosslinked peptide pairs identified from the mass spectra acquired by the Orbitrap Eclipse compared to the Orbitrap Velos Pro. This demonstrates that one of the major limitations on crosslinking proteome depth achieved is the speed, sensitivity, and resolution of the mass spectrometer.

DISCUSSION

In this study we discuss the various strategies implemented to maximize the fragmentation and identification of crosslinked vaccinia virus peptides, with improvements made to almost every step of the crosslinking workflow. In 2019 we reported a low-resolution crosslinking network of the vaccinia mature virion, derived from a dataset of 4,609 unique crosslinked peptides (by protein accession, XL position and crosslinked peptide Mr), of which 3,667 represented intra-protein and inter-protein crosslinked pairs. The remainder were loop-linked peptides and were structurally uninformative. From this preliminary work it became apparent that it would be necessary to substantially expand the crosslinking dataset to a depth that would be permissive for predicted protein structure validation and crosslink-guided docking of vaccinia virion protein structures.

Our results emphasize the importance utilizing XLMS search engines that combine high sensitivity and accuracy with updated software that can handle the large bioinformatics search space that arises from crosslinked sample analysis. No single, top-ranked XLMS search engine outperforms the others and crosslinked peptide identification benefits from the parallelization of these search engines. We also outlined a strategy for reducing the bioinformatics search space, and thereby improving crosslinked peptide identification and lowering false discovery rates, by optimizing a vaccinia virus purification protocol for the production of high yield-high proteomic purity virus. This resulted in a purified vaccinia virus stock with 82% fewer viral and host contaminants, equating to a 19-fold reduction in the bioinformatics search space.

We demonstrated here how fragmentation and identification of crosslinked peptides can be further improved by modifying various sample preparation steps. Pre-enrichment of crosslinked peptides that contained an enrichable handle allowed for the off-line depletion of unreacted peptides, resulting in substantially higher crosslinked peptide identifications. We also identified multiple protein solubilization methods that reduced missed cleavage rates (and thus reduced the bioinformatics search space) leading to improved peptide detection, resulting in 150 - 400% increases in crosslinked peptide identification. We also showed the benefit of using a more advanced mass spectrometer for data acquisition.

By implementing the various strategies outlined here in our crosslinking experiments, we have achieved a saturating dataset of 22,028 unique intra- and inter-protein crosslinked pairs (by protein accession, XL position and crosslinked peptide Mr), comprising 135,273 total CSMs. In later chapters we will describe how we combined these XLMS results with AlphaFold2 predicted structures to describe the structure and assembly of major structural components of the vaccinia mature virion.

MATERIALS AND METHODS

Vaccinia virus purification

Vaccinia virus WR was purified by sucrose gradient per the protocol described [617], with the following changes.

Preparation of a Crude Infection Stock: HeLa S3 (ATCC CCL-2.2) grown to confluency in a Corning Hyperflask in a 37°C humidified incubator with 5% CO_2, and infected with 0.5-1 pfu/cell of vaccinia virus in complete MEM with 5% FetalGro, 6 mM GlutaMAX, 100 U/ml penicillin and 100 µg/ml streptomycin sulfate. Infection was allowed to proceed for 3 days, after which infected cells were harvested and disrupted by freeze-thaw cycling and ultrasonication in a Sonics VCX-750 cuphorn sonicator (at 100% Amplitude) with an ice-water slurry.

Preparation of High yield-high purity vaccinia virus: HeLa S3 cells were grown in a 1 L spinner flask in a 37°C humidified incubator with 5% CO_2 with complete spinner MEM, 10% FetalGro, 6 mM GlutaMAX (or L-Q), 100 U/ml penicillin and 100 µg/ml streptomycin. Prior to infection 5 x 10^8 HeLa S3 cells were transferred to a 50mL spinner flask with complete spinner MEM, 5% FetalGro, 6 mM GlutaMAX, 100 U/ml penicillin and 100 µg/ml streptomycin sulfate, crude vaccinia virus was added at an MOI of 5-8 pfu/cell and incubated with the HeLa S3 cells to allow for attachment. After 30 minutes, the infected cells were transferred to a 1 L spinner flask that contained 950 mL of warmed complete spinner MEM with 5% FetalGro, 6 mM GlutaMAX (or L-Q), 100 U/ml penicillin and 100 µg/ml streptomycin. Infection was allowed to proceed for 3 days, after which infected cells were transferred to 450 mL Nalgene centrifuge bottles and cells were pelleted for 10 minutes at 1800 x g at 5°C, and the supernatant was discarded. The cells were resuspended in 5 mL of lysis buffer (10 mM TEAB, 2 mM $MgCl_2$, pH 8.5) and homogenized with a glass dounce homogenizer with 10 strokes. Cells were then centrifuged for 5 minutes at 300 x g, at 5°C, and the supernatant was transferred to a microtube and prepared by 5 rounds of ultrasonication in a Sonics VCX-750 cuphorn sonicator with an ice-water slurry, with intermittent vortexing.

36% Sucrose Cushion: The sonicated supernatant was transferred into a Beckman ultraclear centrifuge tube, a 36% sucrose solution in 10 mM TEAB, pH 8.5 was underlaid, and the virus was pelleted through the sucrose cushion for 80 minutes at 32,900 x g, at 4°C in a pre-chilled SW 32Ti rotor. The supernatant was discarded, and the pellet was resuspended in 800 microL of 10 mM TEAB, 2 mM $MgCl_2$, pH 8.5 with 8 microL of Benzonase and incubated at 37°C for 30 minutes.

24 - 40% Linear Sucrose Gradient (#1): The virus sample was cooled on ice and prepared by 5 rounds of ultrasonication as described above, after which the sample was pelleted by centrifugation at 14,000 x g for 30 minutes, at 4 °C. The supernatant was discarded and the pellet was resuspended in 10 mM TEAB, pH 8.5 and the ultrasonication and centrifugation steps were repeated. The virus pellet was then resuspended in 10 mM TEAB, pH 8.5 following the ultrasonication steps described above and overlaid on top of a chilled 24%-40% linear sucrose gradient in 10 mM TEAB, pH 8.5, and banded by ultracentrifugation for 40 minutes at 26,000 x g, 4 °C in a pre-chilled SW 32Ti rotor, with acceleration and brake rates set to "1" and "no brakes" respectively. The virus bands were harvested in 1 mL aliquots.

24 - 40% Linear Sucrose Gradient (#2): The virus band aliquots were prepared similar to how the virus sample was prepared for the first sucrose gradient. Prior to running the second sucrose gradient, 10 - 20 microL of each aliquot was digested by trypsin and LysC and analyzed by mass spectrometry to identify the aliquots with the highest virion protein:background contaminant ratios. These aliquots were then pooled, pelleted by centrifugation at 14,000 x g for 30 minutes, at 4 °C, and resuspended in 1 mL of 10 mM TEAB, pH 8.5 in the method described above. The virus sample was then overlaid on top of a chilled 24% - 40% linear sucrose gradient, banded, and harvested as described above. Aliquots were prepared again as described above and the aliquots with the highest virion protein:background contaminant ratios were pooled for XLMS experiments.

The number of protein families represented in the purified virus sample were identified by the following digestion method: An aliquot of the pooled virus was solubilized by 8 M urea, 100 mM TEAB, pH 8.5, followed by dilution to 6 M urea with 100 mM TEAB and overnight digestion with LysC (1:100 enzyme:substrate ratio), followed by subsequent dilution to 1 M urea with 100 mM TEAB and overnight digestion with trypsin (1:100 enzyme:substrate ratio), and peptides were prepared for analysis by mass spectrometry as described below. The number of protein groups in the vaccinia virus WR sample, purified by potassium tartrate gradient, was also prepared in a similar manner.

Solubilization Assays

15 µg or 30 µg aliquots of purified vaccinia virus (above) were solubilized as described in Table 1, in duplicate. Samples were digested overnight with LysC (1:100 enzyme:substrate ratio), followed by overnight digestion with trypsin (1:100 enzyme:substrate ratio). Samples were acidified with formic acid (FA) to 2% FA final concentration and desalted by C18/SCX and eluted with 5% NH4OH, 80% CH2CN, 0.1% FA Buffer X as described [578]. Samples were dried under vacuum, reconstitute in 0.1% FA in water, and analyzed by nanoLC-MS/MS with HCD fragmentation. Instrument raw files were analyzed individually and peptides identified by MetaMorpheus against a vaccinia virus WR and human protein database. All searches included a decoy database generated by MetaMorpheus. Search settings were kept consistent between all samples. Precursor and product mass tolerances were set to ±20.0000 PPM. Trypsin was selected as the protease and up to 2 missed cleavages were allowed. No fixed modifications were set. Up to two variable modifications were allowed per peptide and were set as follows: Oxidation on M,

Acetylation on K, Acetylation on X, Carbamyl on K, Carbamyl on X, Deamidation on N, Deamidation on Q. A Q-value threshold of 0.01 was set. "Treat modified peptides as different peptides" was enabled. All other settings were left at default values.

Total PSMs and non-redundant peptide counts were identified from the output files and averaged between duplicate samples. Missed cleavage sites were calculated *de novo* (and results averaged), under the following rules:

- KP and RP were not counted as a missed cleavage site
- KK|-, KR|-, RR|-, RK|- (where "|-" refers to the peptide C-terminus) were not counted as a missed cleavage
- -|K and -|R (where the "-|" refers to the peptide N-terminus) when preceded by a K or R in the protein sequence, were not counted as a missed cleavage
- All other uncleaved K or R residues were considered to be missed cleavage sites

Protein Solubilization with TCEP: Samples were prepared in duplicate as described above with the addition of 10 mM tris(2-carboxyethyl)phosphine (TCEP) during the protein solubilization step.

Vaccinia Virus Crosslinking: Samples were prepared as described [604]. Briefly, bis(succinimidyl) penta(ethylene glycol) (BSPEG5) and bis(succinimidyl) nona(ethylene glycol) (BSPEG9), were purchased from ThermoFisher, Inc. Isotopically-coded disuccinimidyl suberate (DSS), bis(sulfosuccinimidyl)suberate (BS3), and disuccinimidyl glutarate (DSG) were purchased from Creative Molecules. 3,5-bis(((2,5-dioxopyrrolidin-1-yl)oxy) carbonyl)phenyl)phosphonic acid (PhoX) was provided by the Scheltema lab and later purchased from Bruker and prepared as described in [631]. Purified vaccinia virus was crosslinked intact in

100 mM TEAB, pH 8.5 or after brief uncoating treatment (0.05% NP40, 40 mM TCEP, 100 mM TEAB, pH 8.5), followed a wash step to remove the uncoating solution and resuspension in 100 mM TEAB, pH 8.5. Appropriate XL concentrations were determined by SDS-PAGE. XL reactions were quenched with ammonium bicarbonate or removed by spin desalting into ammonium bicarbonate.

For assaying protein solubilization methods, crosslinked virus samples were prepared as outlined in Figure 5.7A. After crosslinking, virus samples were separated into two aliquots and solubilized and digested as described in Table 1, with overnight digestion with LysC (1:100 enzyme:substrate ratio), followed by overnight digestion by (1:100 enzyme:substrate ratio). Cleaved samples were acidified with formic acid (FA) to 2% FA final concentration and desalted by C18/SCX as described [578]. Peptides were eluted with a six-step ammonium acetate gradient in 20% CH3CH, 0.5% FA followed by a final elution with Buffer X. Samples were dried under vacuum and reconstitute in 0.1% FA in water for MS and analyzed on an Orbitrap Velos Pro mass spectrometer with the method described in [604].

For evaluating mass spectrometer performance, intact vaccinia virus was crosslinked with BSPEG5 or DSS as described [604] and solubilized and cleaved as described in Figure 5.7A. Cleaved samples for each crosslink type were pooled acidified with formic acid (FA) to 2% FA final concentration, desalted by C18/SCX, and eluted with Buffer X. Samples were dried under vacuum and reconstitute in 0.1% FA in water for MS. The nanoLC-MS/MS method used for the Orbitrap Velos Pro is described in [604]. The samples were then shipped to the ThermoFisher Scientific facility in San Jose and analyzed on the Orbitrap Eclipse.

XLMS search engines: pLINK2 was downloaded from pFind Studio. MetaMorpheus was downloaded from the GitHub repository (https://github.com/smith-chem-wisc/MetaMorpheus).

Kojak2 was initially downloaded and run on a computer running a linux operating system. After Kojak2 was integrated onto the Trans Proteomic Pipeline, we discontinued use of the linux version of Kojak2. Protein Prospector was run on the Protein Prospector online server. Instrument RAW files were converted to mzML and mgf formats for Kojak2 and Protein Prospector using MSConvert by ProteoWizard.

CHAPTER 6

Addressing structural hierarchies among the membrane and surface proteins of the vaccinia virion via a combination of deep protein-protein crosslinking and deep learning-based structure prediction

Abstract

We have combined deep protein-protein crosslinking distance restraint (XLMS) data with deep-learning-based structure prediction for the 21 known envelope or surface-associated proteins of the vaccinia mature virion (MV) and assemblies thereof. MV protein structure predictions, when benchmarked against known experimental structures showed high quality scores, comparable to those of proteins from other taxa. Prediction of the vaccinia reticulon analog A17 with 5 helices and termini on opposing sides of the envelope, which contrasted with classical reticulon models but was supported by XLMS, provided a route for templating the external scaffold geometry to the virion interior during virion morphogenesis. Triple helical surface protein A27 prediction with parallel strands contrasted with an antiparallel strand in the truncated experimental structure but was supported by XLMS. XLMS-guided docking of A17 with A27 and surface protein A26 showed $A17_2A27_3$ and $A17_2A27_326$ stoichiometry and the A27 triplex oriented perpendicular to the envelope, with a 4-helical bundle of A27 triplex plus A26's C-terminal helix, and A26's globular N-terminal head anchored loosely via a > 10 nm flexible linker. Models revealed a 'band' of disulfides stabilizing the quadruplex. Structural homology was noted between fusion suppressor proteins A26 and A25. XLMS-guided docking of short transmembrane protein A13 to membrane-anchored protein H3 suggested A13 as a membrane anchor and/or adaptor for the bulkier H3 after external scaffold release during morphogenesis. The resulting A13-H3 complex could be docked to A17$_2$A27$_3$ and A17$_2$A27$_3$A26. For virion attachment protein D8, XLMS suggested a homotetramer, which could be docked to A13-H3 in XLMS-guided fashion. Via a combination of XLMS and iterative modeling, the 11-component entry-fusion complex (EFC) could be resolved into 5 binary subcomplexes leading to two principal subassemblies (6-member: A21:O3:G3:L5:H2:A28 and 4-member: A16:G9:F9:J5) connected via their membrane spanning helices, with protein L1 associating peripherally. Protein associations within wrapped virions were deduced.

Author Summary

The vaccinia mature virion packages 21 known membrane or membrane-associated surface proteins, the majority of which play essential roles in cell attachment, virion fusion with the plasma membrane, early- and late-stage virion morphogenesis, genome packaging and virion transport to the cell surface and egress. While classical structural biology has made progress in understanding atomic structures of individual proteins and/or their individual domains, questions remain regarding higher order interactions of the holo-proteins with one another and with the virion envelope. Here, we have exploited a deep dataset of structural distance restraints from crosslinking mass spectrometry experiments to inform deep learning-based structural predictions, and to thereby deduce and dock molecular partners among membrane and surface proteins of the vaccinia virion.

INTRODUCTION

The intracellular mature virion (MV) form of the vaccinia virion comprises a 3.2 GDa particle containing approximately 75 vaccinia gene products. Experimental three-dimensional atomic structures are currently available for (all or parts of) 34 of these. How they fit together into the overall molecular architecture of the virion has, however, remained largely a mystery: Although cryoEM has made some recent inroads [71, 94], the vaccinia virion has proven largely refractory to the conventional approaches of structural biology. This can be attributed, at least in part, to its polymorphic, enveloped, and asymmetric nature. To address the virion molecular architecture, we have generated a saturating crosslinking mass spectrometry (XLMS) dataset for the vaccinia virion comprising 135,000 crosslink spectral matches (CSMs) and 22,000 unique XL in terms of protein accession, XL position and crosslinked peptide Mr. Combining XLMS experimental data with AlphaFold2 high confidence predicted structures, we have begun to piece together structural hierarchies within the vaccinia virion. A proof of principle for this approach has been described recently for the structure and assembly of vaccinia major structural protein P4a [210].

Vaccinia morphogenesis follows a pathway in which membranes play a pivotal role. The initial discernable forms (crescents) comprise hemispheres derived from host cell endoplasmic reticulum (ER) that already contains the key vaccinia transmembrane proteins A14 and A17 trafficked to the ER in the classical way. Within virus factories located in the cytoplasm at the periphery of the nucleus, the crescent membrane evolves during vaccinia morphogenesis to form, eventually, the single-membraned envelope of MV which can be released upon cell lysis. The infectivity of MV relies, to a large extent, on its possession of cell-attachment proteins and proteins required for cellular fusion and entry (the “entry-fusion complex” or “EFC”). In a number of vaccinia strains, at least, *en route* from factory to cell periphery the MV particle buds across the trans-Golgi network (TGN), acquiring two additional (wrapping) membranes, The resulting entity is referred to as "Intracellular Enveloped Virus" (IEV). One of the two membranes so-acquired can fuse with the plasma membrane of the host cell upon virus exit leading to externalized virion with two membranes (the original envelope plus one of the two wrapping membranes from the TGN). The resulting dual-membranated, externalized virion is referred to as "extracellular virus" (EV).

Approximately 23 of the ~75 packaged gene products in MV can be regarded as either transmembrane proteins or closely associated with the virion envelope. 19 of them at least can be loosely divided into four distinct functional classes, namely attachment proteins (H3, A27, D8, L1), EFC (A16, A21, A28, F9, G3, G9, H2, J5, L5, O3), fusion suppressors (A26, A25) or virion assembly proteins (A14, A17, A13). The remaining four either cannot be classified definitively (A9, I5) or are redox-related and merely membrane-associated (E10, A2.5) and therefore beyond the scope of the current work. An additional seven proteins are known to be specific to the extracellular virion form. Here, we have combined the approaches of deep XLMS, deep learning structure prediction and deep functional literature review, to deduce and rationalize higher-order structural arrangements for transmembrane and envelope-associated surface proteins in each of the above functional classes. As outlined in the abstract above, we find a number of new molecular interactions, and provide structural correlates for these and previously reported interactors deduced by other means.

RESULTS AND DISCUSSION

Structural modeling: Benchmarking for vaccinia proteins

Here, *de novo* atomic structural modeling of packaged vaccinia proteins was performed using AlphaFold2 [655]. The performance of AlphaFold2 has been benchmarked extensively by comparing *de novo* generated models with experimental structures for the same proteins that were reported subsequently [656] or that were excluded from the AlphaFold2 search databases [655, 657]. We asked whether the vaccinia structural proteome might be a "special case" for accurate structural modeling given the number of vaccinia proteins with no sequence or predicted structural homologs outside of poxviruses [658] and which consequently present challenges for traditional, homology-based comparative modeling [659]. Therefore, we benchmarked AlphaFold2 in the context of vaccinia proteins specifically.

AlphaFold2's per-residue confidence metric (pLDDT, or predicted lDDT-Ca), which provides a per-residue estimate of confidence on a scale from 0 – 100, corresponds to the model's predicted score in the "Local Distance Difference Test" (lDDT) [660]. The latter measures the percentage of correctly predicted interatomic distances, rewarding locally correct structures/domains. However, since lDDT is a superposition-free score - without an alignment of predicted vs. experimental structures, we assessed average pLDDT by making AlphaFold2 structure predictions for the 30 MV-packaged proteins that had yielded experimental structures for the full-length protein or portions thereof, at the outset of this study, followed by domain-level alignment of the predicted vs. experimental structures (Appendix 3.Table 1). For each such polypeptide two distinct prediction modes were employed, described here as "unrestricted" to include, in the AlphaFold2 search database all PDB entries, or "PDB restricted" wherein the PDB template inclusion date was set to at least 1 year prior to the deposition date for the protein's atomic coordinates in PDB. Two alignment algorithms were used, namely Global Distance Test (GDT, also written as GDT_TS to represent "total score") [661] and TM-Align

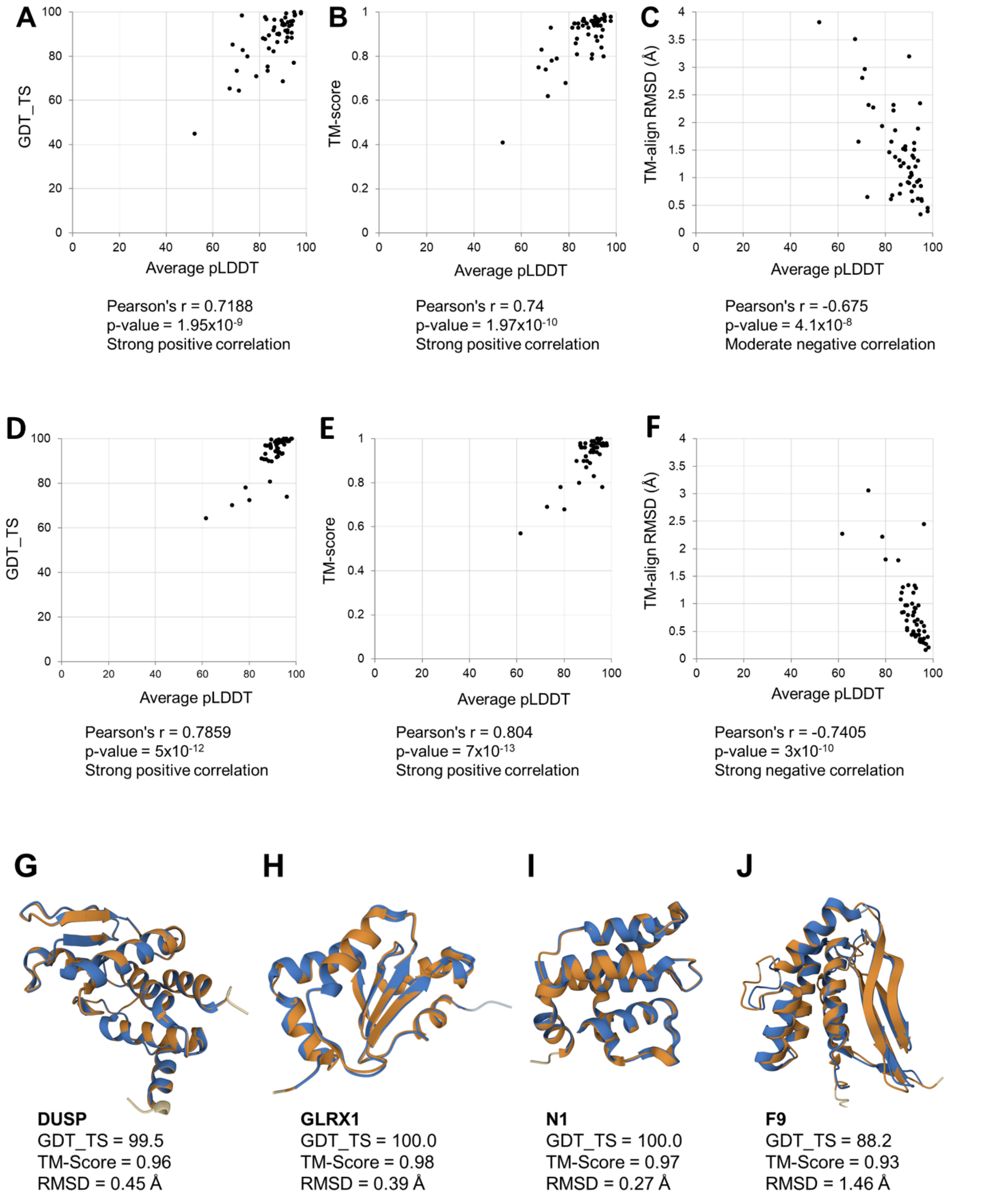


Figure 6.1. Assessment of AlphaFold2 predicted structures for packaged vaccinia virion proteins. (**A-C**) Assessment of AlphaFold2 models ("restricted" mode) for 52 domains in packaged proteins (Appendix 3.Table 1) whose structures have been solved experimentally, by correlation of average pLDDT vs. GDT_TS, TM-score, and TM-align RMSD. Multidomain proteins were split into individual domains for alignment using ChimeraX and the average

pLDDT value for each individual domain was used for the correlation plot. For single domain proteins, if the experimental structure comprised just a portion of the entire protein, the predicted structure was truncated using ChimeraX to just the experimentally determined region for alignment. (**D-F**) Assessment of AlphaFold2 models ("unrestricted" mode) for 52 domains in packaged proteins (Appendix 3.Table 1). Details as described for panels A-C. (**G-J**) Comparison of predicted structures with their experimental structures, in "restricted" mode, for four vaccinia proteins (using TM-Align). Residues in alignment are shown in orange and blue for AlphaFold2 and experimental structures respectively; residues out of alignment are tan and grey for AlphaFold2 and experimental structures respectively. Beneath each alignment are the relevant GDT_TS, TM-Score, and RMSD (Å) values. (**G**) DUSP (PDB: 3CM3). (**H**) GLRX1 (PDB: 2HZE). (**I**) N1 (PDB: 4BBD). (**J**) F9 (PDB: 6CJ6). (**K-N**) AlphaFold2 structural predictions in "restricted" mode for DUSP, GLRX1, N1, and F9 (N-terminal domain) colored by pLDDT. Beneath each alignment are the relevant average pLDDT values.

[662] (Appendix 3.Table 1). In both "restricted" and "unrestricted" prediction modes, the resulting average pLDDT values showed a strong positive correlation with GDT_TS and TM-score, and a moderate negative correlation with RMSD (Fig. 1A-F). This extended, to the vaccinia proteome specifically, prior reports of pLDDT as a reliable indicator for modeling confidence and accuracy [655, 657].

Not only was correlation good - overall scores were also very good: Even in "restricted" mode, AlphaFold2 predictions for the 30 proteins' 52 domains showed median GDT_TS, TM-score and RMSD values of 91.3, 0.94, and 1.29 Å respectively (Appendix 3.Table 2). Such values very closely approach AlphaFold2's high level of accuracy (median GDT_TS of 92.4) for a more generalized set of proteins at CASP14 [657]. Moreover, of the 52 domains (Appendix 3.Table 1), 30 achieved GDT_TS values > 90 and a further 18 scored in the range 70 – 90. In this context, GDT_TS values >70 and >90 are considered to represent, respectively, "high-accuracy" in backbone rotamer placement, and accuracy comparable to that of experimental structures with respect to backbone rotamer placement and by extension sidechain placement [656]. Even values in the range 50 – 70 are considered to represent an accurate backbone fold [656]. Our median TM-Align value of 0.94 (Appendix 3.Table 2) also seemed excellent, with values > 0.5 considered to represent a matching backbone fold and a value of 1.0 representing a perfect match [662].

AlphaFold2 models predicted with high confidence included those for vaccinia proteins DUSP, GLRX1, and N1 (Fig. 1G- N) whose GDT_TS values were 99.5, 100, and 100 respectively. Notably, these highly confident, experimentally accurate structural predictions were achieved in "restricted" mode which excluded in some cases more than two decades of PDB entries from the search space, limiting AlphaFold2's ability to draw upon even homologous non-viral structures deposited within this timeframe. Fig. 2 shows the accurate placement of side chains for protein GLRX1 despite its modeling in "restricted" mode. Structural models with GDT_TS values corresponding to "high accuracy" in backbone rotamer placement, such as the model for membrane protein F9 predicted in "restricted" mode (Fig. 1J), showed an accurate overall structure with only minor deviations in the placement of flexible loop regions. Indeed, such differences may reflect experimental deformations due to crystal lattice contacts [656, 663] wherein predictions that are highly confident represent the *bona fide* solution conformation more authentically than the crystallographic model.

As expected, domain-level alignments of vaccinia polypeptides (modeled in "unrestricted" mode) yielded even better scores than alignments in "restricted" mode, with median values for GDT_TS, TM-score and RMSD of 97.15, 0.96 and 0.71 Å respectively (Appendix 3.Table 2). 45 of the 52 domains were experimentally competitive with GDT_TS values > 90, and a further six were "highly accurate" (GDT_TS values of 70 – 90; Appendix 3.Table 1). Overall, the extraordinary performance of AlphaFold2 in predicting accurate structures for vaccinia proteins in both PDB-restricted and PDB-unrestricted modes provided a confident basis for proceeding with AlphaFold2-based structural predictions of virion proteins whose structures remain unsolved to-date.

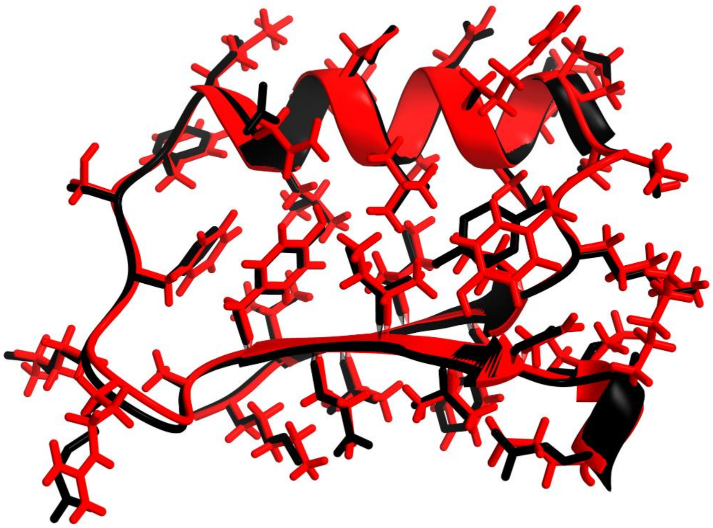


Figure 6.2. Accurate placement of side chains even in "restricted" mode. The AlphaFold2 predicted structure of vaccinia GLRX1 (red) was aligned with the Ectromelia virus GLRX1 (EVM053) X-ray crystal structure (black) (PDB: 2HZE) [663]. The two proteins' sequences differ by only one amino acid. Side chain superpositions are shown for residues 14 - 50.

Major envelope proteins A14, A17, and A27

Vaccinia morphogenesis initiates with the insertion of vaccinia phosphoproteins A14 and A17 into the ER membrane [121, 532, 664], whereupon vaccinia viral membrane assembly proteins (VMAPs) modify the ER by scission, leading to the entry of vaccinia protein D13 which proceeds to assemble a scaffolding cupola resulting in the formation of crescent (hemispherical) structures as the earliest discernible forms in virion morphogenesis [122, 183, 203, 534, 665]. Crescent formation is aided by A17's reticulon activity, which promotes membrane curvature [535] and allows crescents to develop into fully spherical IV within a complete D13 external scaffold, in which the excised patch of ER membrane becomes the nascent virion envelope [203].

Proteins A14 and A17 are interacting partners [121, 196]. Apart from a 9 amino acid stretch of A14 [666], atomic structures for neither A14 nor A17 have been reported to date. Nonetheless, the bulk of each protein has been predicted to comprise membrane-spanning helices: A14 with a pair of such helices and both protein termini on the inner side of the virion envelope [533, 667] and A17 with four membrane spanning helices [535] (Fig. 3A). A14 was modeled by AlphaFold2 with moderate confidence (Fig. 3B; Appendix 3.Table 3) showing the anticipated pair of helices with both protein termini oriented in the same direction. We could detect no intra-protein or inter-protein XL to A14. Though very abundant in virions, the small (90 residue) A14 protein has a suboptimal distribution of lysine residues - the targets of the majority of crosslinkers used in our study.

A17 is AG-processed during maturation at residues 16 [536] and 185 [196] by the vaccinia protease I7. A17 was modeled by AlphaFold2 with high confidence in both unprocessed and fully processed forms (Fig. 4A, B; Appendix 3.Table 3). Predicted structures for the two forms were essentially indistinguishable in the core transmembrane domain (RMSD = 1.23 Å; TM-score = 0.95). No strong structural homologs were identified by DALI for either form of A17. The core domain's predicted structure comprised five alpha helices (Fig. 4A, B) in which the second of the four transmembrane helices expected from TM domain predictions (Fig. 3A) was split at proline 92 and folded back on itself. In this manner, A17's protein termini were placed on opposite sides of the envelope (Fig. 4A, B).

In vivo, A17's N-terminal region is considered exposed on the external side of the virion envelope in both unprocessed [532, 537] and processed [538] forms of the protein [535]. The unprocessed N-terminus anchors the envelope to the interior face of the D13 external scaffold [668], with release of the scaffold and A17's N-terminal peptide achieved via proteolytic processing of A17 by protease I7. The nascent N-terminus so generated, recruits and anchors cell-attachment protein A27 to the envelope exterior [121-123] via C-terminal regions of A27 [111, 545].

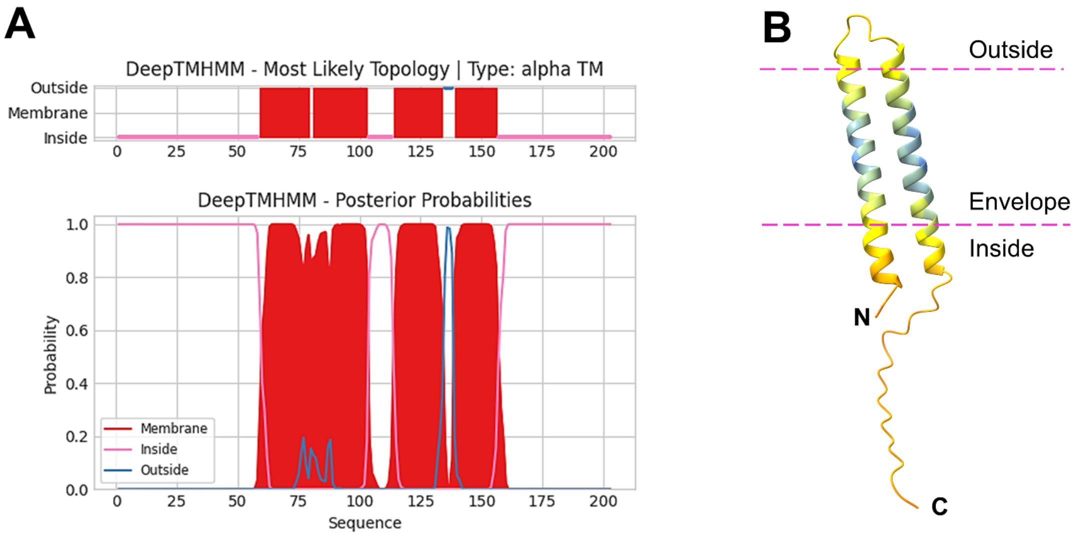


Figure 6.3. Vaccinia virus membrane protein A14. (**A**) Topology prediction for protein A17 by DeepTMHMM, showing the predicted membrane spanning helices. (**B**) AlphaFold2 model of protein A14. Magenta lines ("Outside"/"Envelope"/"Inside") suggest the position of the envelope based on the topology described in [667].

For A17's C-terminus, *in vivo* topology is more ambiguous: In infected cells, residue 178 can be cysteine-labeled (in some molecules at least), suggesting its orientation on the external side of the virion/ER [535]. Immunogold labeling of infected cells with antibody to A17's unprocessed C-terminus (residues 180 - 203), however, places this epitope on the insides of both crescents and IV [196, 203, 532, 534]. Additional circumstantial evidence for an internal orientation of A17's processed C-terminus includes an absence of immunogold surface labeling of MV with antibody to residues 165 - 185 [538] and a failure of this antibody to neutralize MV [538]. Moreover, our XLMS data from intact MV show C-terminal residues of A17 crosslinked to proteins of the core wall (Fig. 4C), in particular P4a, P4b, A4, and A12 while A17's N-terminus crosslinks predominantly to envelope proteins (Fig. 4C), predominantly A27. This crosslinking pattern is consistent with A17 protein termini on opposing sides of the virion envelope and by extension AlphaFold2's 5-helix model. Such a model could provide a mechanism for transmission of the external D13 scaffold template to the virion interior during virion morphogenesis, in which a comparable geometry seems to emerge in the N-terminal fragment of core wall structural protein P4a [210].

Reticulon-like activity has been reported for A17 [535]. Eukaryotic reticulons, including RTN1-4, human DP1 and its yeast homolog YOP1 are responsible for tubular ER curvature [124]. These proteins are characterized by the possession of a ~190 residue C-terminal "reticulon domain" containing two membrane-spanning segments, each originally characterized as ~30 - 35 residues in length (~35 – 45 residues in our analyses). Each span was considered too long to cross the membrane once, and in two prototypical reticulons the first membrane-spanning segment was shown to form a membrane-inserted hairpin while evidence for the second segment also doing so was equivocal [124]. Both spans forming a hairpin would yield a proposed 4-helix

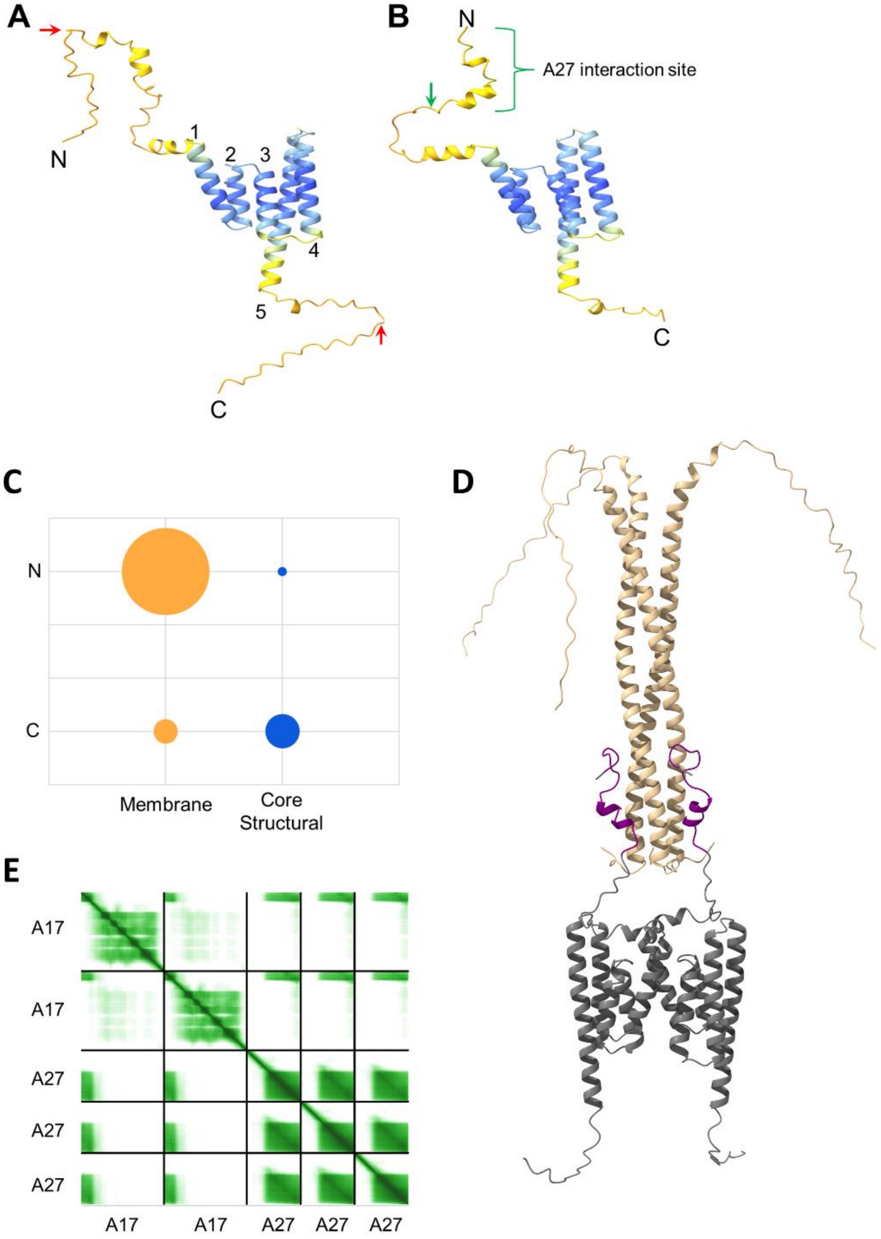


Figure 6.4. Vaccinia membrane protein A17. (**A**) AlphaFold2 predicted structure of full-length A17 colored by pLDDT. Red arrows denote positions of known AG| processing sites. Helices are numbered consecutively (N to C). Helices 2 and 3 comprise the second of four expected helices based on transmembrane predictions (Fig. 3A). (**B**) As panel (**A**) but with AG| processed A17. The green bracket shows A17's N-terminal A27 interaction site, with the green arrow marking its C-terminal end. (**C**) Crosslinking emphasis of the A17 N- and C-termini (indicated by N and C) to other membrane proteins (yellow) or proteins of the virion core wall (blue). Circle area sizes represent total CSM counts. (**D**) A172-A273 complex (processed A17), second-ranked model. Grey: Processed A17; Tan: A27; Purple: A17 residues (18 - 36) that interact with A27. (**E**) PAE plot for the A172-A273 model shown in panel D. The A17-A17 interface shows dark green (confidence) only in the vicinity of the A17-A27 attachment site.

"W" shape [124] with both protein termini on the same side of the membrane. To our knowledge, there is no experimentally determined atomic structure for a complete reticulon. AlphaFold2-predicted structures for prototypical cellular reticulons (Fig. 5) yielded no clear consensus in predicted topology, numbers of predicted helices, or orientation of termini. The same conclusion was drawn from 'reticulon' and reticulon-like predicted structures in the AlphaFold2 database (not shown). There seems no reason, therefore, to consider the current understanding of reticulons as evidentiary either for or against a 5-helix predicted structure for A17.

A17 homodimerizes *in vivo* [531]. The homodimer could be modeled by AlphaFold-multimer for both unprocessed and processed forms of A17, or even just A17's core transmembrane domain lacking the flexible N- and C-termini (Fig. 6A). All three of the above structures showed high confidence in the core transmembrane domain (Fig. 6B). A14 and A17 can be coimmunoprecipitated from infected cells [196] though it is unknown whether this results from a direct interaction. AlphaFold-multimer could model the two proteins together as an A14-A17 heterodimer or an $A14_2A17_2$ tetramer but with low confidence regarding the placement of subunits (data not shown).

A17 can be coimmunoprecipitated with A27 [536]. The A27-A17 interaction leads to anchoring, at the virion envelope surface, of A27 homotrimer as an $A17_2A27_3$ complex [536]. AlphaFold-multimer could successfully model the $A17_2A27_3$ complex for either unprocessed or processed forms of A17 (Fig. 4D) and with high PAE confidence for the A27-A17 interaction (Fig. 4E). $A17_2A27_3$ models showed A27 with a disordered N-terminal 41 amino acid region followed by a homotrimerized region extending from residues 42 to 103 as a parallel triple-helical coiled-coil (Fig. 4D). The C-terminal end of A27's triple helix interacted identically with either processed or unprocessed A17 *in silico*, with the interacting region of unprocessed A17

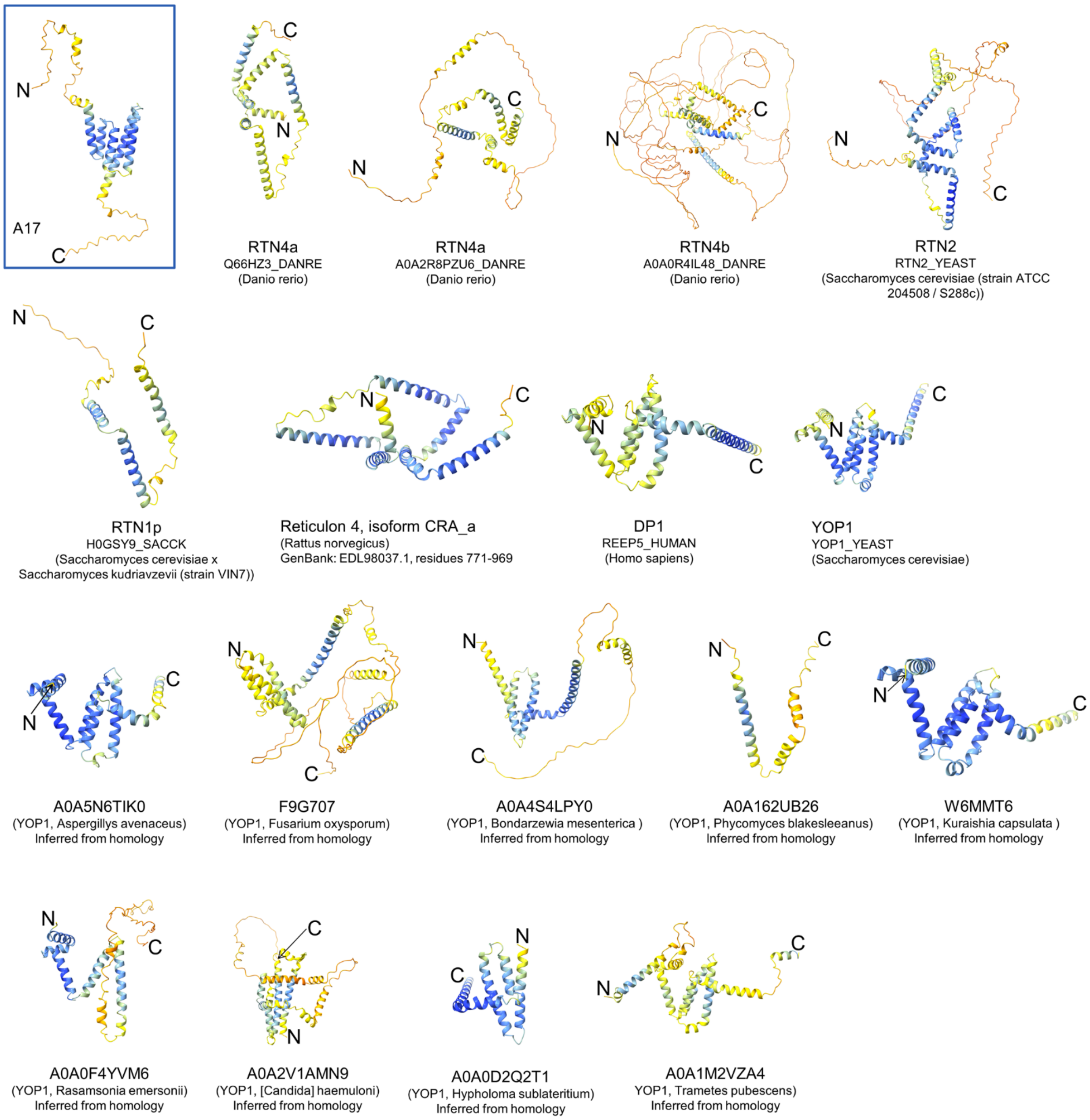


Figure 6.5. AlphaFold2 predicted structure of A17 vs. structures predicted, in house, for cellular reticulons with known reticulon activity or with this function inferred from homology. No clear consensus fold is apparent. Proteins are colored by pLDDT and labeled by UniProt accession.

corresponding to the N-terminus of processed A17. *In vivo*, A27 is synthesized before completion of A17 processing (by immunoblotting, A27 and unprocessed A17 are detectable *in vivo* at high levels 6.5 hours post infection while processed A17 is mostly detectable starting 8 hours post infection [664]). Presumably, newly synthesized A27 does not attach to unprocessed A17 prior to scaffold release due to A17's steric encapsulation within the D13 scaffold. To validate the A17-A27 interaction site, deletion of residues 18 – 36 from unprocessed A17 *in silico* abrogated A17-A27 interaction in AlphaFold2 models entirely (data not shown). Conversely, the 18 – 36 peptide alone was capable of interacting with A27 *in silico* with high confidence (Fig. 7).

The prediction of a parallel triple helix for A27 (above) contrasts with the reported experimental structure for A27 in which a triple-helical coiled coil was reported for residues 42 – 84 but with an antiparallel (flipped) orientation for one of the three helices [128]. The AlphaFold-multimer prediction of parallel strands in the triplex was supported by some lines of evidence, namely: (a) A prior report highlighting interaction between A27's C-terminus and A17 N-terminal residues 20 - 29 and 32 - 36 within a synthetic peptide comprising A17 residues 18 – 50 [545]; (b) A17-A27 crosslinks in our XLMS dataset (Fig. 4C, Fig. 11B), in which A17 N-terminal residue K36 crosslinks strongly to A27 C-terminal residues K98 and K99 but not to A27's N-terminus; (c) extensive crosslinking of A27's C-terminal (but not N-terminal) residues to transmembrane protein A13 and membrane-anchored protein H3, with minimal or no crosslinking near the A27 N-terminus (despite the presence of potential lysine crosslinking sites) or the A27 N-terminal amine (Fig. 11).

The PAE plot for A17$_2$A27$_3$ (Fig. 4E) showed only modest confidence in the orientations of the two A17 core transmembrane domains among the set of models output by AlphaFold-

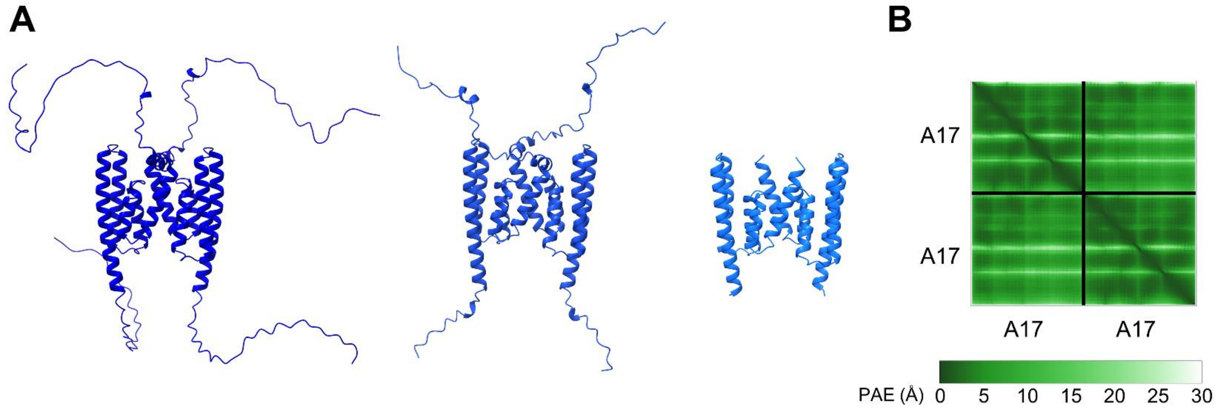


Figure 6.6. **AlphaFold-multimer models of vaccinia protein A17 dimers.** (**A**) Left to right: Full-length, AG|-processed, and core domain only (**B**) PAE plot for the A17 dimer core domain in panel (**A**), strongly supporting the A17-A17 homodimer model. Three homomultimer crosslinks in A17 were identified by XLMS (residues 2-2, 36-36 and 180-180), each with CSM count = 1. All three reside in highly flexible regions of the predicted structure, and so were, overall, regarded as uninformative.

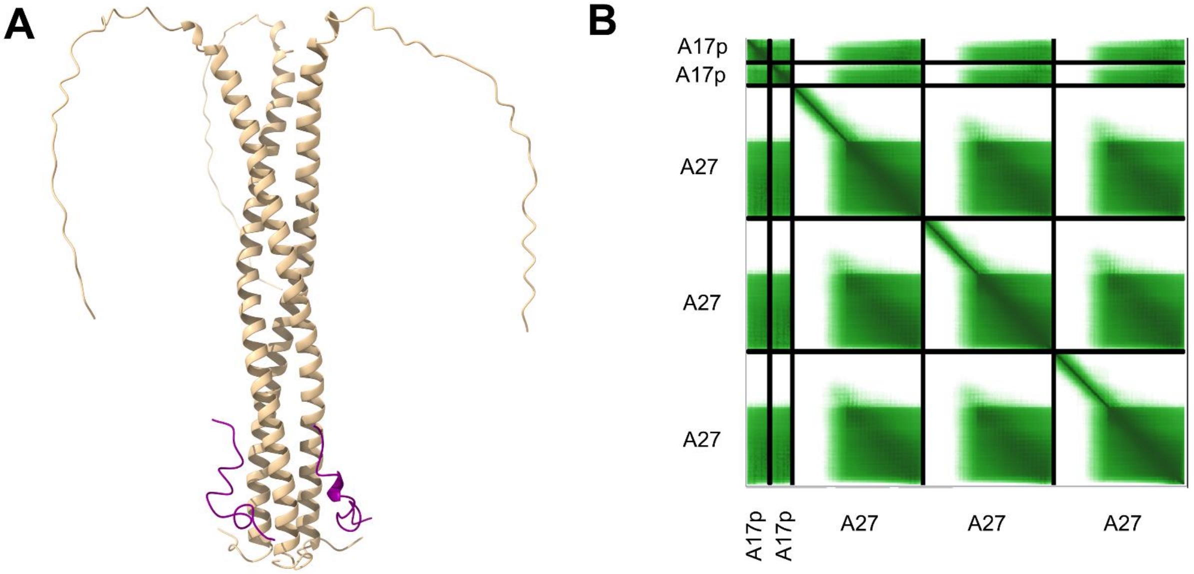


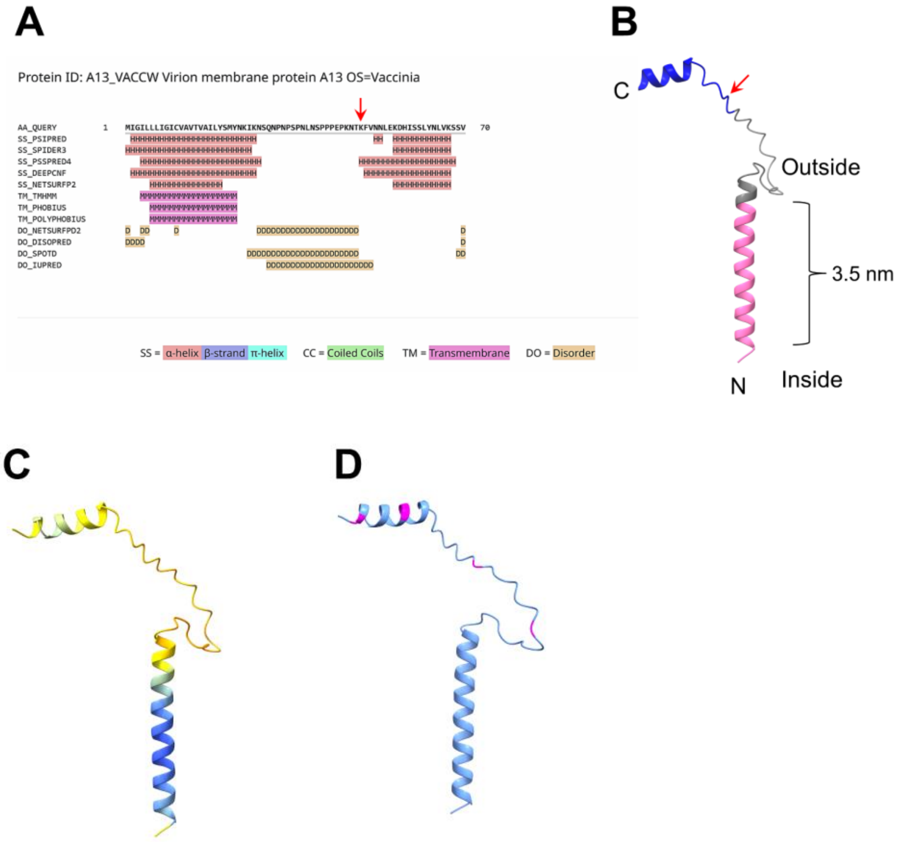
Figure 6.7. A17 residues 18 - 36 mediate the A17-A27 interaction. (**A**) A27 trimer (tan) co-modeled with two peptides covering A17 residues 18 - 36 (purple). (**B**) PAE plot for the complex in panel (A). "A17p" = A17 peptide (residues 18 – 36).

multimer, with no preferred interface between the two A17 chains. Nonetheless, the orientation of A17 in the second-ranked prediction among the set of $A17_2A27_3$ models (Fig. 4D) corresponded to that of the A17 homodimer alone (Fig. 6A), lending support for this orientation of A17 molecules in the $A17_2A27_3$ model of Fig. 4D.

Major envelope proteins H3 and A13

Major envelope proteins A13 and H3 also localize to MV. A13 is a short, abundant integral membrane phosphoprotein with a hydrophobic N-terminal domain that is co-translationally membrane-inserted [198, 669]. By either Quick2D or AlphaFold2, the predicted A13 structure shows an N-terminal helix of sufficient length to span the membrane, followed by a disordered region, then a short C-terminal helical ectodomain (Fig. 8). By immunogold EM the ectodomain has been suggested to lie on the envelope exterior [669]. A13 is essential for normal morphogenesis [198, 669] with a possible role at the stage of viral genome encapsidation by IV [171, 198].

H3 is a 324 aa protein whose N-terminal ectodomain is C-terminally anchored to the MV envelope post-translationally [119]. Its repression, like that of A13, leads to defects in virion morphogenesis beyond the IV stage accompanied by an accumulation of DNA “crystalloids” [117]. Like A27, H3 plays a role in virion attachment to the host cell at the outset of infection via heparan sulfate binding [112, 115]. An X-ray crystal structure for the H3 N-terminal domain (residues 34 – 240 [115]) shows a glycosyltransferase fold, with corresponding UDP-glucose and Mg^{2+} binding activities [115]. The Alphafold2 predicted structure for H3 showed an N-terminal domain with high fidelity to the solved X-ray crystal structure (Fig. 9A).



Protein ID: A13_VACCW Virion membrane protein A13 OS=Vaccinia

```
AA_QUERY       1    MIGILLLIGICVAVTVAILYSMYNKIKNSQNPNPSPNLNSPPPEPKNTKFVNNLEKDHISSLYNLVKSSV   70
SS_PSIPRED           HHHHHHHHHHHHHHHHHHHHHHHHHH                        HH  HHHHHHHHHHHH
SS_SPIDER3          HHHHHHHHHHHHHHHHHHHHHHHHHH                             HHHHHHHHHHHH
SS_PSSPRED4            HHHHHHHHHHHHHHHHHHHHHHHHH                    HHHHHHHHHHHHHHHHHHHH
SS_DEEPCNF           HHHHHHHHHHHHHHHHHHHHHHHHHH                      HHHHHHHHHHHHHHHHHH
SS_NETSURFP2             HHHHHHHHHHHHHHH                                   HHHHHHHHHHHH
TM_TMHMM               MMMMMMMMMMMMMMMMMMMM
TM_PHOBIUS               MMMMMMMMMMMMMMMMMM
TM_POLYPHOBIUS           MMMMMMMMMMMMMMMMMM
DO_NETSURFPD2       D  DD     D                DDDDDDDDDDDDDDDDDDDDD                     D
DO_DISOPRED         DDDD                                                                 D
DO_SPOTD                                     DDDDDDDDDDDDDDDDDDDDDDD                    DD
DO_IUPRED                                        DDDDDDDDDDDDDDDDDDDDDD
```

SS = α-helix β-strand π-helix   CC = Coiled Coils   TM = Transmembrane   DO = Disorder

Figure 6.8. Structure prediction for major transmembrane protein A13. (**A**) Secondary structure prediction by Quick2D, showing predicted N-terminal transmembrane domain, disordered region and C-terminal helix. Red arrow: Crosslinking hotspot (residue 49). Panels (**B**) - (**D**) AlphaFold2 predicted structure of A13: (**B**) Colored by domain based on the secondary structure prediction of panel (A): Transmembrane helix (pink), disordered region (grey), and C-terminal helical region (blue) with the outside/inside of the virion envelope indicated along with A13's crosslinking hotspot (red arrow). A standard hydrated lipid bilayer width of 3.5 nm is marked with pink dotted lines. (**C**) Colored by pLDDT. (**D**) Colored uniformly light blue with phosphorylation sites (as identified in ref. [101]) highlighted magenta.

By DeepTMHMM the region C-terminal to residue 270 is predicted to form the membrane insertion region of H3. Indeed, expressed in isolation, the H3 C-terminal region (residues ~260 - 324) alone, can insert stably into microsomes [119]. The region C-terminal to residue 270 appeared fairly compact covering two short C-terminal alpha helices preceded by a turn-beta-turn motif and approximately half of a third alpha helix (Fig. 9A, pink). By pLDDT (Fig. 9B) and PAE (Fig. 9C), the fold for this region seemed of comparable confidence and likely rigidity to that of the N-terminal ectodomain, and substantially more rigid and reliably predicted than the transmembrane regions of other membrane proteins examined here, reinforcing the likely compact and more globular character of H3's C-terminal membrane-interacting region. This distinction seems consistent with the unusual (post-translational) mechanism for H3

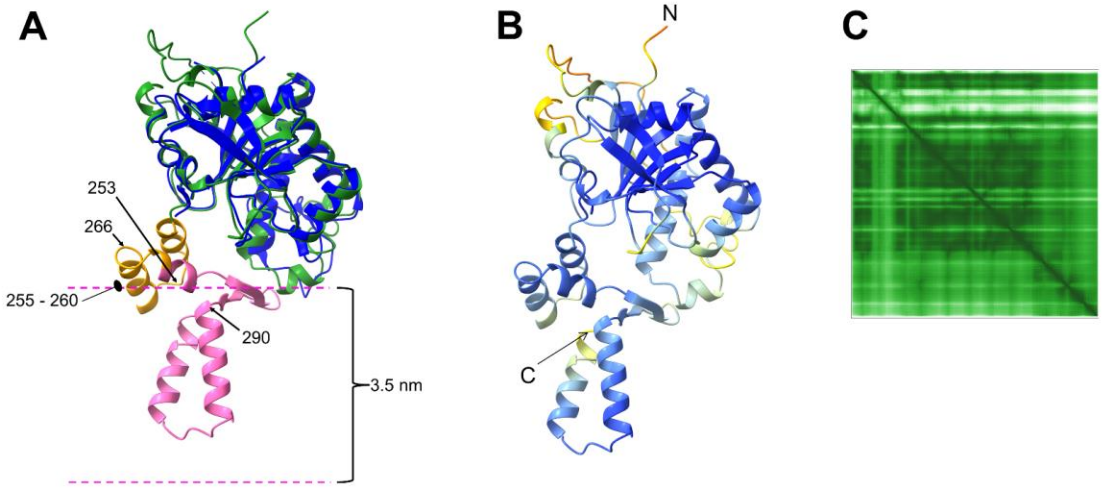


Figure 6.9. Structure prediction for major envelope protein H3. (**A**) Overlay of the H3 X-ray crystal structure (PDB: 5EJ0; blue; residues 34 - 240) with the AlphaFold2 predicted structure for full-length protein. Green: Residues 1 - 240 (ectodomain; covering the X-ray crystal structure [115]). Pink: Residues 271 - 324 (the region predicted by Deep HMTMM to insert into membranes). Orange: The intervening region (residues 241 – 270). This region was mobile and invisible in the X-ray crystal structure but on the other hand lysines 253 and 266 (arrowed) within the orange region were crosslinker-reactive, providing evidence the orange region is bilayer-external. Residues discussed in the text are indicated. A standard hydrated lipid bilayer width of 3.5 nm is indicated with pink dotted lines. The suggested bilayer position is discussed in the text. (**B**) H3 model of panel (A) colored by pLDDT. (**C**) PAE plot for H3.

anchoring to membranes (above). Regarding maximum bilayer insertion depth, our XLMS data showed clear crosslinker reactivity of Lysines 253 and 266, with lysines 380 and 310 (within the two short C-terminal alpha helices) remaining unreactive (data not shown) supporting the placement of the bilayer below lysines 253 and 266 (Fig. 9A). Regarding minimum insertion depth, one of the two orange helices (residues 255 – 260, Fig. 9A) incorporates a hydrophobic patch that therefore likely dips into the bilayer. The bilayer insertion depth suggested in Fig. 9A satisfies the resulting maximum and minimum insertion depths and strongly suggested that the bilayer-occluded region comprises mainly the C-terminal-most two alpha helices, starting at residue 290 (Fig. 9A). It is open to speculation whether the turn-beta-turn motif adjacent to residue 290 sits on the bilayer surface with hydrophobic sidechains that point “downwards” also bilayer-inserted.

Albeit A13 and H3 have comparable phenotypes in virion morphogenesis (above), their direct mutual interaction has not been reported. However, multiple interactions between the two proteins were identified in our expanded XLMS dataset, with a total combined CSM count of 18 (Fig. 10A). Although AlphaFold-multimer was unable to model a confident H3-A13 heterodimer (data not shown), AlphaFold2 models for monomeric H3 and A13 could be docked manually, guided by interprotein crosslink distance restraints (Fig. 10B). The resulting model (Fig. 10B, C) showed A13’s transmembrane (TM) helix extending somewhat deeper into the membrane bilayer than H3’s two short C-terminal helices: While A13’s TM helix can cleanly span the bilayer (Fig. 8B, Fig. 10C), H3’s two C-terminal helices span little more than the outer leaflet (Fig. 9A, Fig. 10C). Albeit the H3 C-terminal region, in isolation, can insert stably into microsomes [119], H3 alone might nonetheless be a weaker, more nonspecific membrane binder, minimizing the energetic requirement for post-translational insertion.

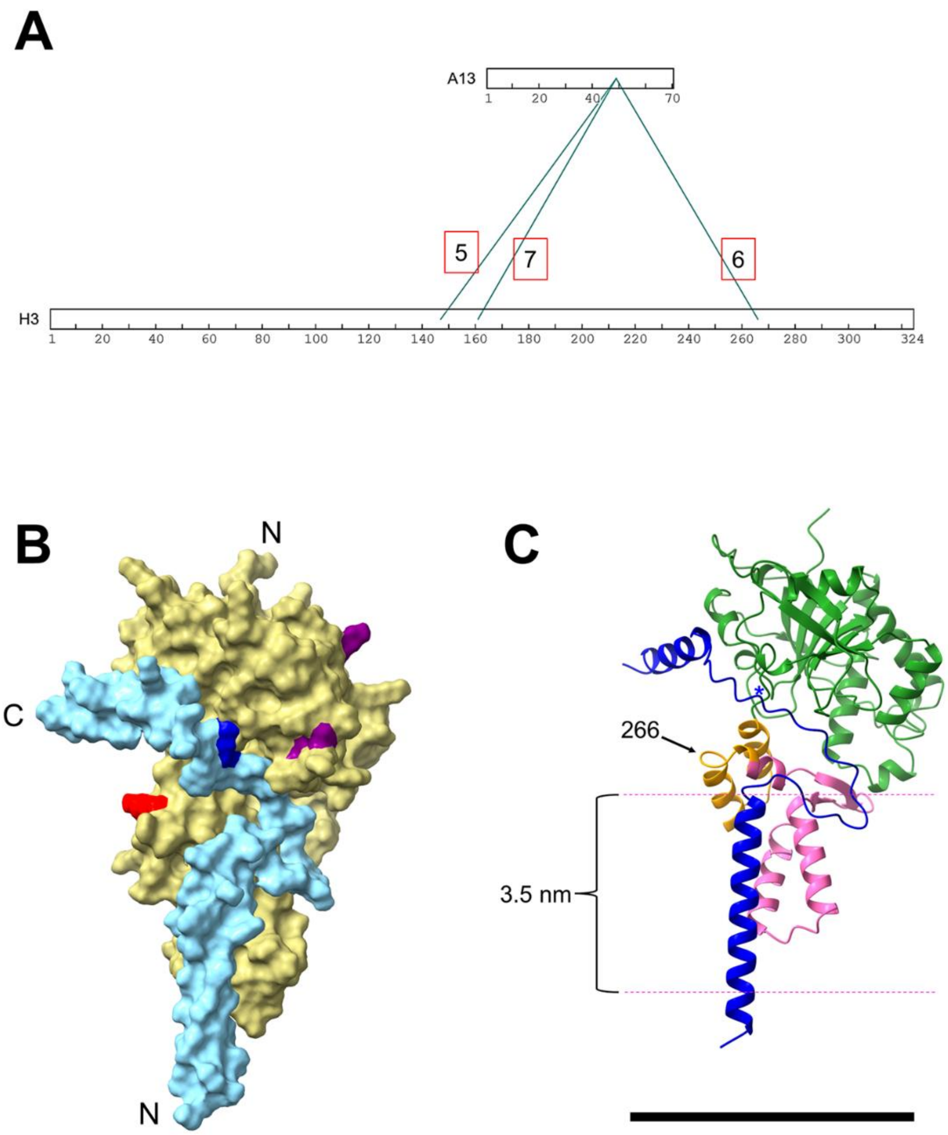


Figure 6.10. H3-A13 interaction. (**A**) Crosslinks between H3 and A13. (**B**), (**C**): Predicted A13 and H3 structures docked manually. Panel (B) shows restraints in manual docking of A13 (light blue) and H3 (yellow), namely A13 residue 49 (dark blue) which crosslinks to H3 residue 266 (red) in intact virions, and to H3 residues 266 (red), 147 and 161 (purple) after brief virion treatment with mild uncoating reagents. Protein N- and C-termini are indicated, where visible. Panel (**C**) shows H3 regions discussed in the text, with H3 residues 1 - 240 (covering the X-ray crystal structure [115]) colored green and residues 271 - 324 (the region predicted by Deep HMTMM to insert into membranes) colored pink. The intervening region (residues 241 – 270) is colored orange, and A13 is colored dark blue. Arrowed: H3 lysine 266 crosslinking site (red in panel B). Blue asterisk = A13 crosslinking hotspot (residue 49). Scale bar = 5 nm. A standard hydrated lipid bilayer width of 3.5 nm is indicated with pink dotted lines. The suggested bilayer position covers A13 residues 5 – 29 (A13's hydrophobic N-terminal region).

In this case, A13 may serve as a targeting anchor or position marker for H3: A13, a short protein of minimal size, could fit under the IV external scaffold's lattice during early virion morphogenesis then after removal of the scaffold A13 would recruit the bulkier H3 protein. This would reflect a likely comparable role for A17 in recruiting and anchoring surface protein A27 (above). Since the docked structure for H3-A13 (Fig. 10B, C) is just a composite of the monomer predictions, it cannot be discounted that A13 docking may stabilize H3 membrane insertion by the deployment of a different conformation in H3's C-terminal domain than the one shown, such as a "straightening" of the two C-terminal helices of H3. Nonetheless, all H3 AlphaFold2 models for H3 showed the same conformation for the entire protein, which contrasts with, for example, core wall processing intermediate P4a-1+2, for which AlphaFold2 models could equally populate two equally probable conformational states of the protein, strongly indicative of triggered conformational change [210]. Recruitment of H3 may be coupled to the known phosphorylated status of A13 in MV, in which all of the known phosphorylation sites of A13 [101] (Fig. 8D) occur within A13's H3-attachment domain. Perhaps A13 phosphorylation is coordinated with H3 anchoring after scaffold release.

A structural model for a complex of the four major MV envelope proteins H3, A13, A17 and A27 was generated (Fig. 11A) that could satisfy the multiple confident crosslinks in our dataset between the four proteins (Fig. 11B). These included the strongly detected crosslinking of A27 residue 98 (located at the A27-A17 interface) to A13 residue 49 (Fig. 12A, C) with a combined CSM count of 10 (Fig. 11B) and to multiple residues in a patch on the H3 surface with a combined CSM combined count of 289 (Fig. 12A, C, Fig. 11B). They also included the crosslinking of A17 residue 36 to H3 and A13 with combined CSMs of 70 and 2 respectively (Fig. 12B, C, Fig. 11B). Thus, in addition to a role in membrane anchoring (above), A13 may

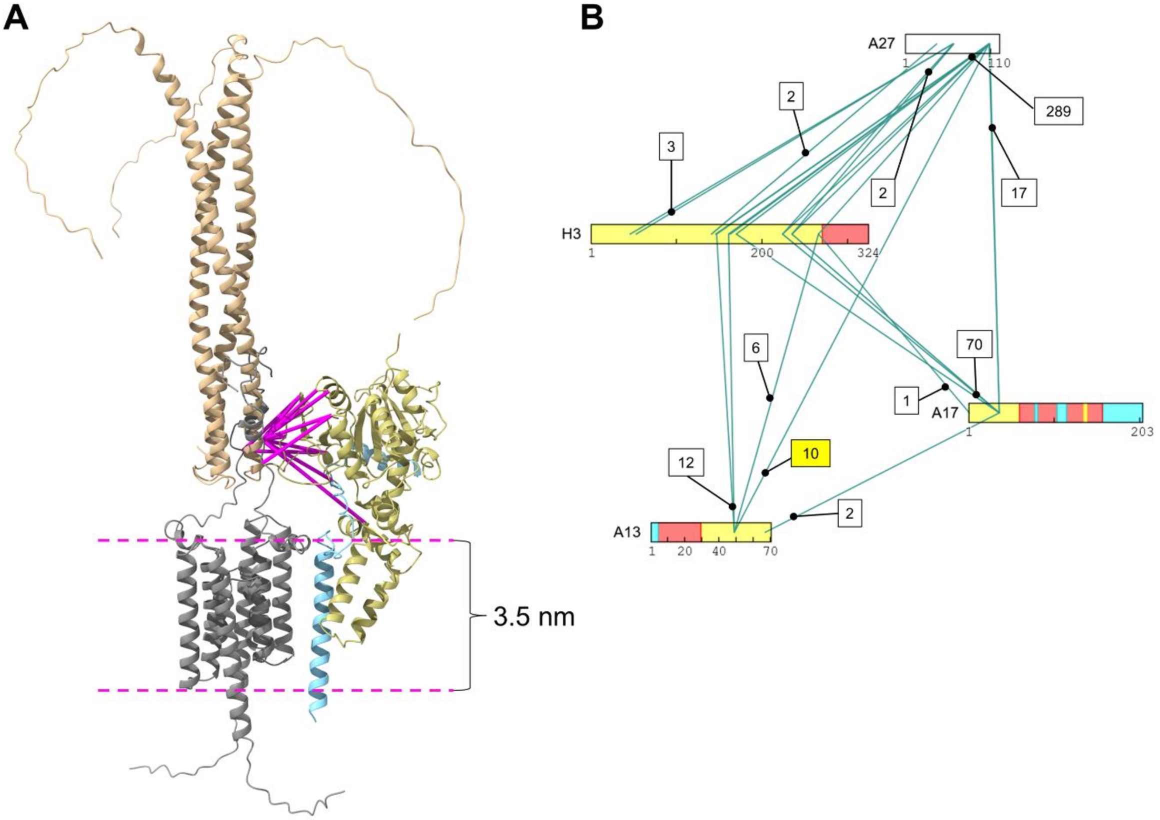


Figure 6.11. H3-A13 docking to the A17$_2$A27$_3$ complex. (**A**) Manual docking on the basis of XLMS distance restraints. H3, A13, A17 (processed), and A27 are colored yellow, light blue, grey, and tan respectively. Crosslinks are shown in magenta. A single docked H3-A13 complex is shown for simplicity, albeit implied symmetry in the A17$_2$A27$_3$ complex reiterates this docking site circumferentially. A standard hydrated lipid bilayer width of 3.5 nm is indicated with pink dotted lines. (**B**) Crosslink network between A17, A27, A13, and H3 supporting the model in panel (A). Box lengths correspond to protein size. H3, A13, and A17 are colored with ectodomains yellow, membrane spanning regions red, and "inside" of the virion envelope cyan. XL were identified from intact virions and virions treated briefly with uncoating reagents (NP-40/TCEP). Boxed numbers: CSM count for each XL or XL cluster. Highlighted yellow: A13-A27 crosslink discussed in the text.

serve as an adaptor protein to draw H3 laterally to the A17-A27 complex (below). The role of the resulting H3-A13-17-A27 complex is open to speculation, but might involve the multivalent or concerted attachment of MV to heparan sulfate at the outset of infection.

Fusion suppressors A25 and A26

During cowpox virus infection, mature virions become recruited/embedded in occlusion bodies composed of viral protein ATIp through an association of the latter with virion surface protein A26 [549, 670, 671]. Vaccinia strain WR encodes a C-terminally truncated ATIp (A25L [165]; Gershon unpublished) which is also incorporated into MV in an A26-dependent manner but does not form occlusion bodies [549, 672]. Consistent with the apparent anchoring mechanism via A26, neither cowpox virus ATIp nor vaccinia A25L possess detectable transmembrane domains (data not shown).

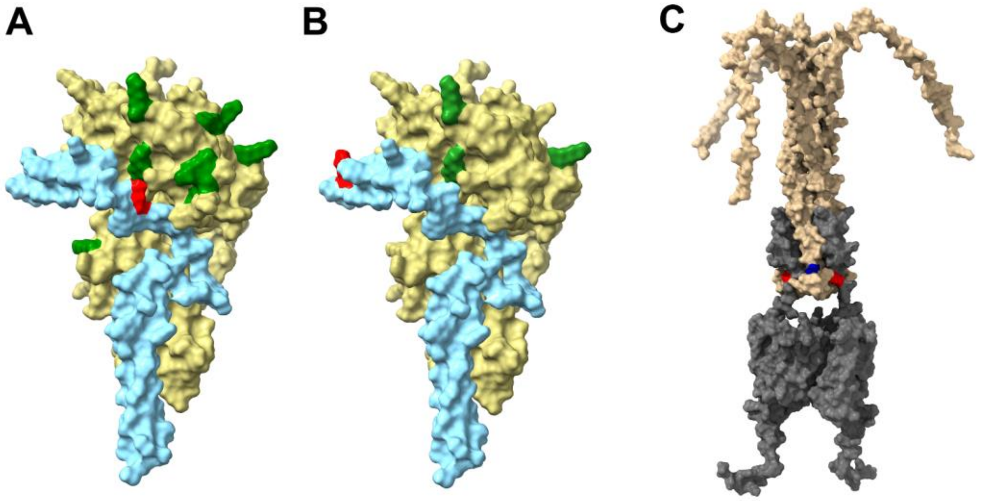


Figure 6.12. Sites of crosslinking between H3-A13 and the A17$_2$A27$_3$ complex. (**A**), (**B**): H3 (yellow) and A13 (light blue) showing crosslink sites (dark green and red) for (**A**) A27 and (**B**) A17, respectively. (**C**) The A17$_2$A27$_3$ complex with A17 (gray) and A27 (tan), showing corresponding H3 and A13 crosslinking sites: A27 residue 98 (blue); A17 residue 36 (red).

A26, in turn, is anchored to the MV surface via an interaction with the A27-A17 complex [550] and acts as a virion attachment protein for cell surface laminin [109]. However, in roles apparently unrelated to cellular attachment, A26 and vaccinia A25L can function also as suppressors of virion-host cell fusion at the plasma membrane, thus directing virus entry through micropinocytosis [127, 163, 164]. Fusion suppression by A26 involves its interaction with EFC proteins A16 and G9 [163] via His-cation pairs in A26's N-terminal region [127]. A26 can be inactivated, and its role in fusion suppression thereby relieved by brief acid treatment of MV [163, 164]. Neither A26 nor A25L are packaged in vaccinia EVs [673].

A26 comprises an N-terminal globular domain, for which an atomic structure is available [127], attached to a C-terminal domain with no experimentally determined structure. AlphaFold2 structural predictions for A26 showed an N-terminal globular domain modeled with high confidence (Fig. 13A, B), that was experimentally competitive in both "PDB restricted" and "PDB unrestricted" modes (Appendix 3.Table 1) and a C-terminal domain, modeled with low confidence (Fig. 13B), comprising an alpha-helical domain joined to the N-terminal domain via a

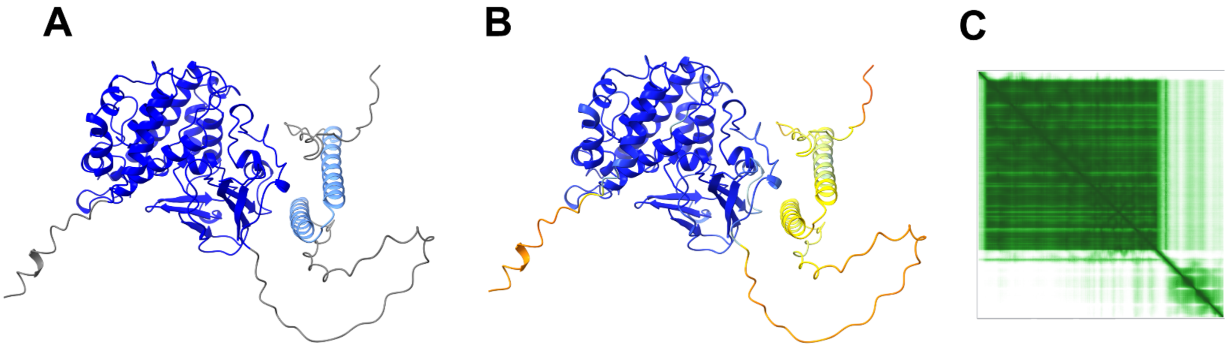


Figure 6.13. AlphaFold2 predicted structure for protein A26. (**A**) Colored by domain. Dark blue: N-terminal domain resolved in the atomic structure (PDB: 6A9S, residues 17 - 364). Light blue: C-terminal helical domain; gray: Unstructured N-terminal residues (1 - 16), inter-domain linker and the C-terminal extension. (**B**) Colored by pLDDT. (**C**) PAE plot of the A26 model shown in panels (A) and (B).

flexible linker (Fig. 13A, B). This C-terminal domain was identified, via HHpred, as structurally homologous to A27's coiled coil domain (data not shown). Via AlphaFold-multimer structure prediction for A26 monomer, flexibility between the N- and C-terminal domains was supported by the PAE plot of A26 (Fig. 13C).

The A26-A27 interaction localizes to A27's triple-helical coiled coil [121] and A26's C-terminal domain [125]. $A27_2A26$ and $A27_3A26$ complexes have been identified by immunoprecipitation, with the latter as the more abundant form [125]. AlphaFold-multimer could successfully model both complexes, with A26's C-terminal "A27-like" domain (data not shown)_co-folding with multiple A27 chains to form either a four-helical or three-helical bundle for $A27_3A26$ or $A27_2A26$ respectively (Fig. 14A, Fig. 15A, respectively). Both models showed highly confident PAE plots between the interacting domains of A26 and A27 (Fig. 14B, Fig. 15B) albeit with variable placement of the A26 N-terminal domain due to flexibility of the inter-domain linker described above, whose overall span is >10 nm (data not shown). In the quadruplex, A26's C-terminal domain is parallel to the three parallel A17 chains (Fig. 14A).

The A26-A27 interaction is dependent upon disulfide bonding involving A27 cysteines 71 and 72 and A26 cysteines 441 and 442 [125] in some manner that was not previously described. The models of Fig. 14A and Fig. 15A highlight the specific A27 and A26 cysteines that form inter-chain disulfide bonds, serving to cross-validate both the study of ref. [125] and the two new models (Fig. 14A and Fig. 15A). Specifically, the two new models both show A27 C71 and A26 C442 within disulfide bonding range, as were A27 C72 and A26 C441 (Fig. 14C, Fig. 15C). The two models also placed A27 cysteines 71 and 72, which are not directly involved

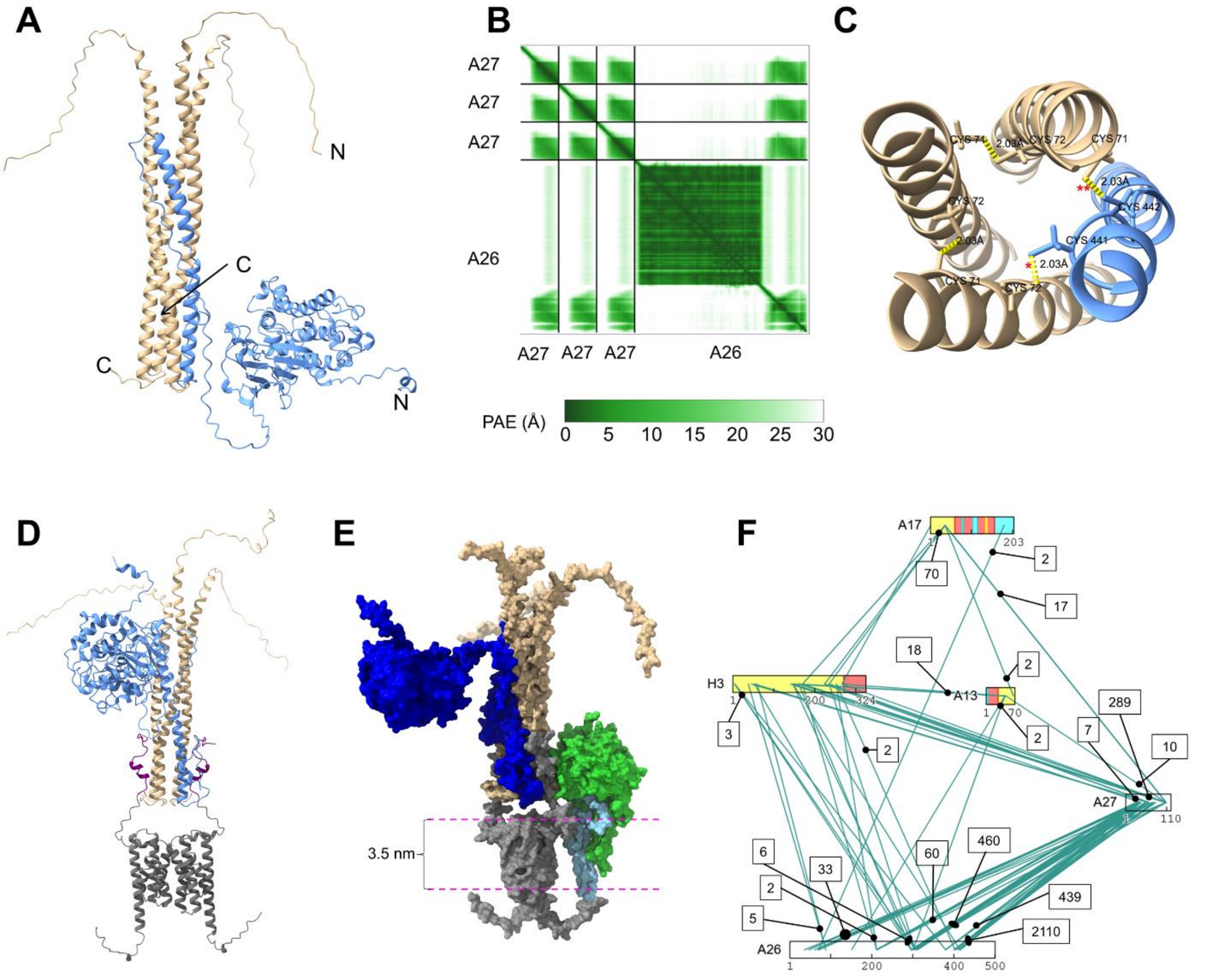


Figure 6.14. A27 trimer interaction with A26. (**A**) A27 trimer (tan) co-models with the A26 (light blue) C-terminal "A27-like" helix (Fig. 13, data not shown) via a four-helix bundle. N- and C- termini are labeled for A26 and one of the three A27 chains. (**B**) PAE plot for the A27$_3$A26 complex. (**C**) Band of disulfide bonds (yellow broken lines) within the four-helix bundle that are required for, and mediate, the A27-A26 interaction (shown are A27 residues 65-89 and A26 residues 435-456). A26 C441 disulfide bonds to A27 C72 (red *), while A26 C442 disulfide bonds to A27 C71 (two red **). Disulfide bond lengths are given for each disulfide pair. (**D**) AlphaFold-multimer model for the A17$_2$A27$_3$A26 complex, with A17 (processed) colored grey (with residues 18 - 36 colored purple), A27 colored tan, and A26 colored light blue. (**E**) Model of the A17$_2$A27$_3$A26 complex docked with H3-A13, showing A17 (processed) grey, A27 tan, A26 dark blue, H3 green, and A13 light blue. A standard hydrated lipid bilayer width of 3.5 nm is indicated with pink dotted lines. (**F**) XL network between A17, A27, A26, A13, and H3 supporting the model in panel (E). Bar lengths correspond to protein Mr. H3, A13, and A17 are colored as in Fig. 11B. XL were identified from intact virions and virions treated briefly with uncoating reagents (NP-40/TCEP). Boxed numbers: CSM count (for combined CSM >1) for each XL or XL cluster.

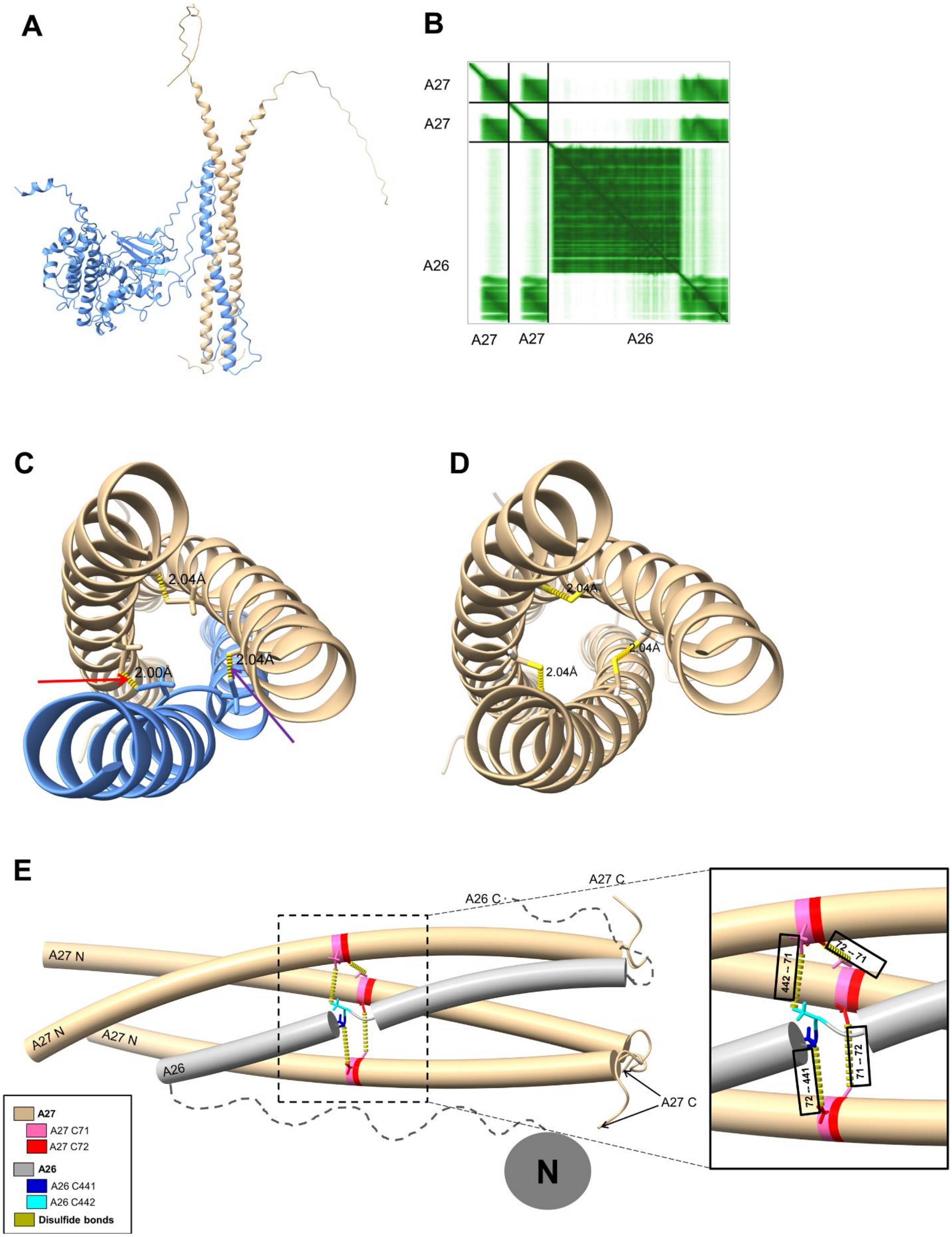


Figure 6.15. A26 interaction with A27 as a 1:2 heterotrimer. (**A**) A27 dimer (tan) co-models with the A26 (light blue) C-terminal "A27-like" helix (Fig. 13) to form a three-helical bundle. (**B**) PAE plot for the A27$_2$A26 complex. (**C**) Disulfide bonds (yellow broken lines) within the three-helical bundle that are required for and mediate the A27-A26 interaction (figure shows A27

residues 55 - 110 and A26 residues 421 - 472). A26 C441 disulfide bonds to A27 residue C72 (red arrow), while A26 residue C442 disulfide bonds to A27 residue C71 (purple arrow). Disulfide bond lengths are given for each disulfide pair. (**D**) Disulfide bonding within the A27 triplex. (**E**) Schematic showing the band of disulfides securing the A27 (tan)-A26 (gray) quadruplex comprising A26 residues 415 - 472 and A27 residues 33 - 110. In ChimeraX, alpha-helical and remaining residues were rendered as tubes and coils respectively after which helices were separated manually for clarity in visualizing the paired cysteines. Manually, A26's N-terminal globular domain and flexible linker were depicted as a circle and dotted line, respectively. The structural prediction split the A26 helix, with residues 441 - 443 modeled as coil. Molecules, residues, and disulfide bonds are colored according to the key. Right side: Zoom of boxed region with disulfide bonded cysteines labeled.

in the modeled A27-A26 interaction, within range to form A27-A27 inter-chain disulfide bonds (Fig. 14C, Fig. 15C - E).

Docking of the A26 C-terminal helical domain with the A27 trimer was supported by A27-A26 crosslinking (Fig. 16A, C), with 98% of XL between A27 and the 26 C-terminal domain falling discretely within the A26 C-terminal helical region (with a combined CSM count = 2070 of 2110 satisfying XL distance restraints and the remaining 2% involving predominantly the apparently flexible N-terminal region of A27). A26-A27 crosslinking also highlighted, by contrast, an enormous promiscuity in positioning of the A26 N-terminal domain with respect to A27 (Fig. 16B, C), in which the diversity of crosslink positions on both molecules was unattributable to any discrete A26-A27 complex: At least five spatial positions/orientations of the A26 N-terminal domain were required (Fig. 16C) to satisfy crosslinks involving A26's N-terminal domain (Fig. 16B). This was apparently attributable to great flexibility within the >10 nm linker connecting A26's N- and C-terminal domains (above). A26 and A27 also co-modeled as an $A27_2A26_2$ complex, with equally high confidence and with cysteine pairs placed within disulfide bonding distance (data not shown). All attempts at modeling higher order multimer combinations failed (data not shown).

We could further extend the $A27_3A26$ complex by co-modeling with A17 dimer. The resulting complex, with stoichiometry $A17_2A27_3A26$ (Fig. 14D) suggested how the three proteins may be arrayed at the MV envelope surface. The A17-A27 interaction described above was maintained within this complex, with the interface of the A17 molecules matching that in the A17 homodimer model (Fig. 6A). As with the $A27_3A26$ complex, A27 and A26 cysteines within $A17_2A27_3A26$ were modeled within disulfide bonding range (data not shown), in the arrangement shown in Fig. 14C and Fig. 15E. The corresponding PAE plot showed a high

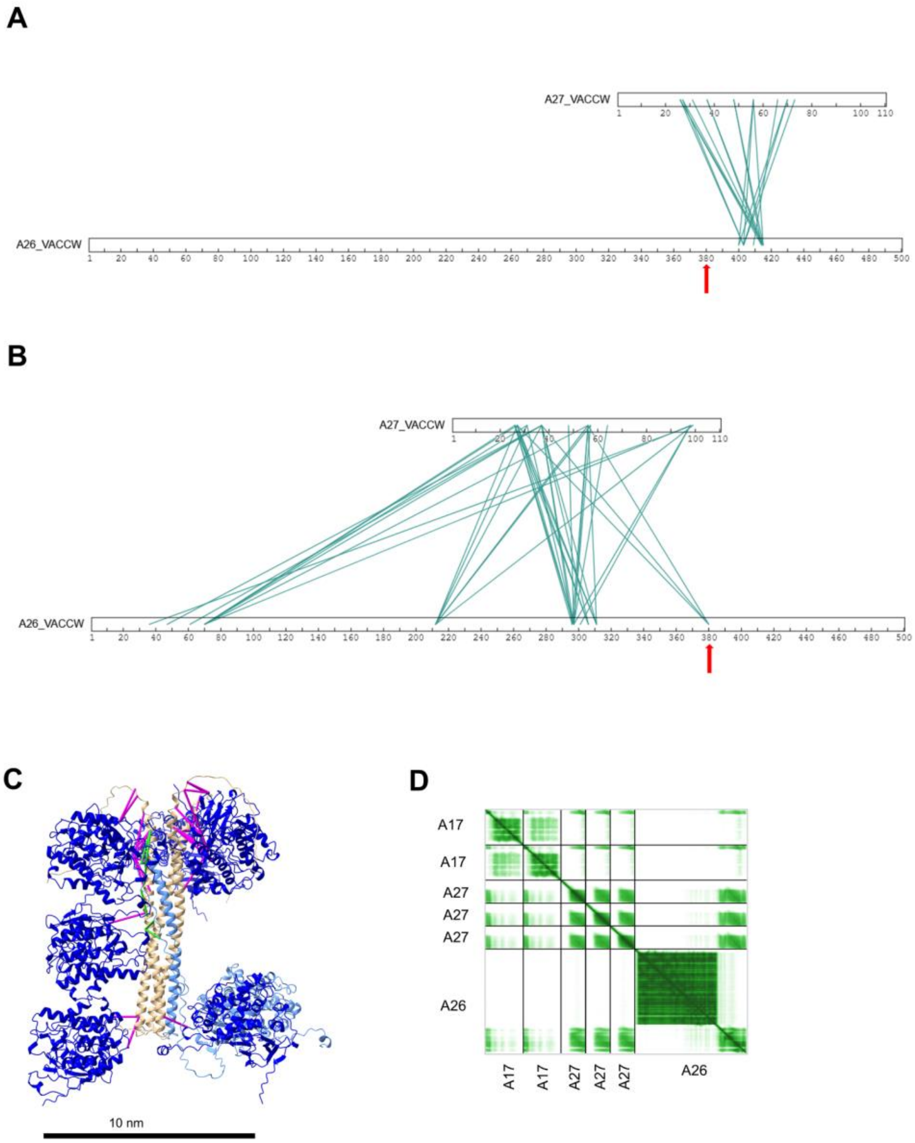


Figure 6.16. **A26-A27 crosslinks highlight the flexibility and broad spatial positioning of the A26 N-terminus with respect to the A27-A26 four-helical bundle.** (**A**) Crosslinks between the A26 C-terminal domain and A27. Total CSM count = 2549. Red arrow: Boundary between A26's N- and C- terminal domains. Crosslinks on A27 are restricted within its triplex region. (**B**) Crosslinks between the A26 N-terminal domain and A27. Total CSMs = 553. Red arrow: Boundary we between A26's N- and C- terminal domains. Crosslinks are promiscuous throughout A27 and across A26's globular 'head". (**C**) The A27$_3$A26 complex modeled by AlphaFold-multimer is colored as in Fig. 13A, with A27 and A26 molecules shown in tan and light blue respectively, and with A26 C-terminal crosslinks to A27 that satisfy XL distance restraints colored green. Overlaid are A26's N-terminal domain models (dark blue), positioned manually to satisfy experimental XL distance restraints between A27-A26 (N-terminal domain) and with XL colored magenta. Scale bar = 10 nm. (**D**) PAE plot for the A17$_2$A27$_3$A26 complex.

structural model for H3, A13, A17 and A27 (Fig. 14E). Fig. 14F shows our XLMS network confidence interaction between the A17 N-terminus and the $A27_3A26$ four-helical bundle (Fig. 16D). Successful modeling of the $A17_2A27_3A26$ complex allowed us to add A26 to the above-described between the five proteins.

A25L, albeit truncated in vaccinia with respect to cowpox virus (above) is nonetheless a multidomain protein, whose atomic structure has not been reported. In an HHpred global alignment search, A25L's N-terminal domain showed very strong structural homology with that of A26 (above; [127]; data not shown). Modeling of A25L by AlphaFold2 supported this, with similar folds evident in the N-terminal domains of A25L and A26 (Fig. 17; Appendix 3.Table 3). This apparent domain duplication in the poxviruses provides a structural correlate to the comparable functions of the two proteins in fusion suppression. Structural alignments show non-retention, in A25L, of the two histidine residues in the A26 N-terminal domain required for His-cation interaction. They seem to be replaced locally with arginine and phenylalanine residues. A25L may employ alternative His-cation pairs or anion-anion pairs to mediate fusions suppression or may rely on an entirely different mechanism for fusion suppression.

Notably, our XLMS dataset showed an extensive XL network between A26, A25L, EFC proteins, and EFC-peripherally associated proteins F9 and L1 (Fig. 18A). The A25L C-terminal domain (residues 332 – 500) was predicted by HHpred to have structural homology with the coiled coils from filamentous proteins (data not shown) [658], with comparable structure in the AlphaFold2 model (Fig. 18B). Interactions of A26 or A25L with the EFC (Fig. 18A and below) did not reveal an obvious mechanism of fusion suppression. Moreover, no models for cowpox virus ATI could be generated to address the structure of the cowpox virus occlusion body. Additionally, we were unable to unearth an A25L-A26 interaction model as a structural correlate to this known interaction with its role in occlusion body-virion embedding (above), or an A27-A25L complex with a comparable 4-helical bundle as the A27-A26 model. Nonetheless, XLMS data from intact MV suggested an interaction between the A26-A27 complex and A25L's C-terminal region between residues 391 and 653 (Fig. 18C)

Virion attachment protein D8L

Vaccinia envelope protein D8 functions as a virion attachment protein in the initial stages of infection, binding chondroitin sulfate on the cell surface [113]. An atomic structure has been reported for C-terminally truncated D8 monomer, covering residues 1 - 261 of the 304-residue protein [126]. D8 homodimerizes by disulfide bond formation at its sole cysteine - C262[126]. The resulting homodimer can, in turn, form higher order homomultimers via non-covalent interactions - showing a homotetramer, reportedly, as the most abundant higher-order form

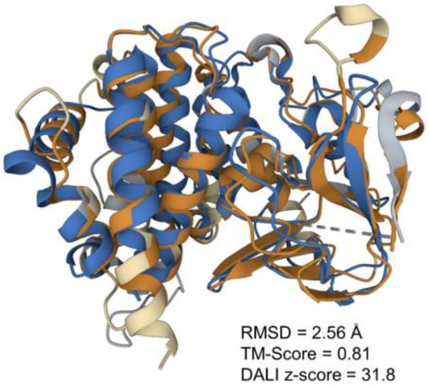


Figure 6.17. **Alignment of the predicted structure for vaccinia protein A25L (residues 1-331) vs. the experimental structure for the A26 N-terminal domain (PDB: 6A9S).** Coloration of the A25L predicted structure was: Orange: Well-aligned regions. Tan: Regions out of alignment. Coloration of the A26 experimental structure was: Blue: Well-aligned regions. Gray: Regions out of alignment. Dotted grey lines: Residues not resolved in the A26 experimental structure. RMSD (Å), TM-Score and DALI z-score (Appendix 3.Table 3) indicated a very close alignment: DALI z-scores > 20 are considered highly significant [674].

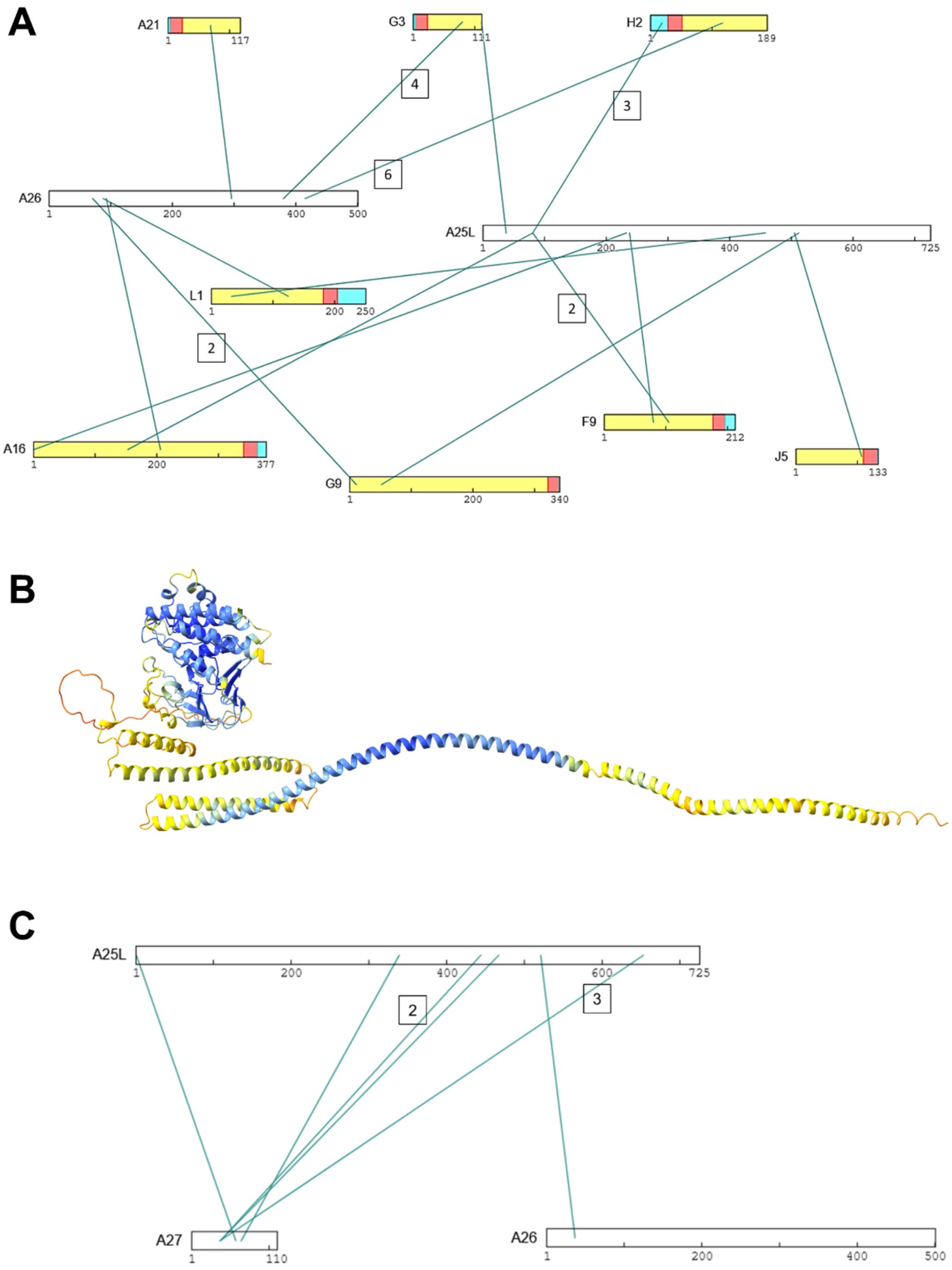


Figure 6.18. **XL interactions of vaccinia fusion suppressors A26 and A25L.** (**A**) XL network of vaccinia fusion suppressors A26 and A25L with EFC proteins and EFC-associated proteins F9 and L1. Bar lengths correspond to protein Mr. EFC proteins, F9, and L1 are colored according to their ectodomains (yellow), membrane spanning regions (red), and "inside" of the virion envelope (cyan). Boxed numbers: As in Fig. 14. (**B**) AlphaFold2 predicted structure for vaccinia A25L colored by pLDDT. (**C**) Crosslinks between A25L, A27, A26, in intact MV. Boxed numbers show XL with a combined CSM >1.

observed *in vitro* [126]. D8 homomultimerization *in vivo* is supported by our XLMS data, with a homomultimer CSM count of 134 from intact and uncoated MV.

A homodimer model for full-length D8 was generated via AlphaFold-multimer (Fig. 19A) with high confidence PAE for the placement of ectodomains (Fig. 19B). However, the C-terminal region was predicted to be highly flexible (Fig. 19B), and the model could not satisfy the inter-subunit disulfide at C262. Upon removal of residues 269 - 304, D8 could be modeled as a homodimer with ectodomains arranged as in the full-length homodimer model (Fig. 19A) but with C262 α-carbons within disulfide bonding range - albeit side chain thiols were out of range for disulfide bond formation (Fig. 20A). The full-length homodimer model (Fig. 19A) satisfied 90% of D8-D8 homomultimer crosslinks (crosslinking of a residue to the same residue within the same protein sequence) with a combined CSM = 121 of 134 (Fig. 20B). However, even this homodimer model satisfied fewer than 25% of D8-D8 inter-protein CSMs, suggesting a higher order structure. Attempts to model higher order D8 homomultimers with AlphaFold-multimer (trimer, tetramer, pentamer, etc.) either failed entirely or showed only very low PAE confidence between domains (data not shown). XL-guided manual docking of the D8 homodimer ectodomains, however, revealed a homotetrameric arrangement comprising two side-to-side homodimers. This arrangement satisfied 98.6 % of inter-protein XL (487 of 494 total intact CSMs; Fig. 19C). Since subunits lie parallel in a plane, we presume the transmembrane domains (removed by truncation in the homotetramer model) project in pairs perpendicularly out of the plane shown. The appealing model of a bundle of four transmembrane domains at the central axis of the homotetramer exhibiting mirror symmetry (Fig. 19D), was incompatible with D8-D8 crosslinking. Equally incompatible was an arrangement of “vertically” oriented ectodomains (Fig. 19D).

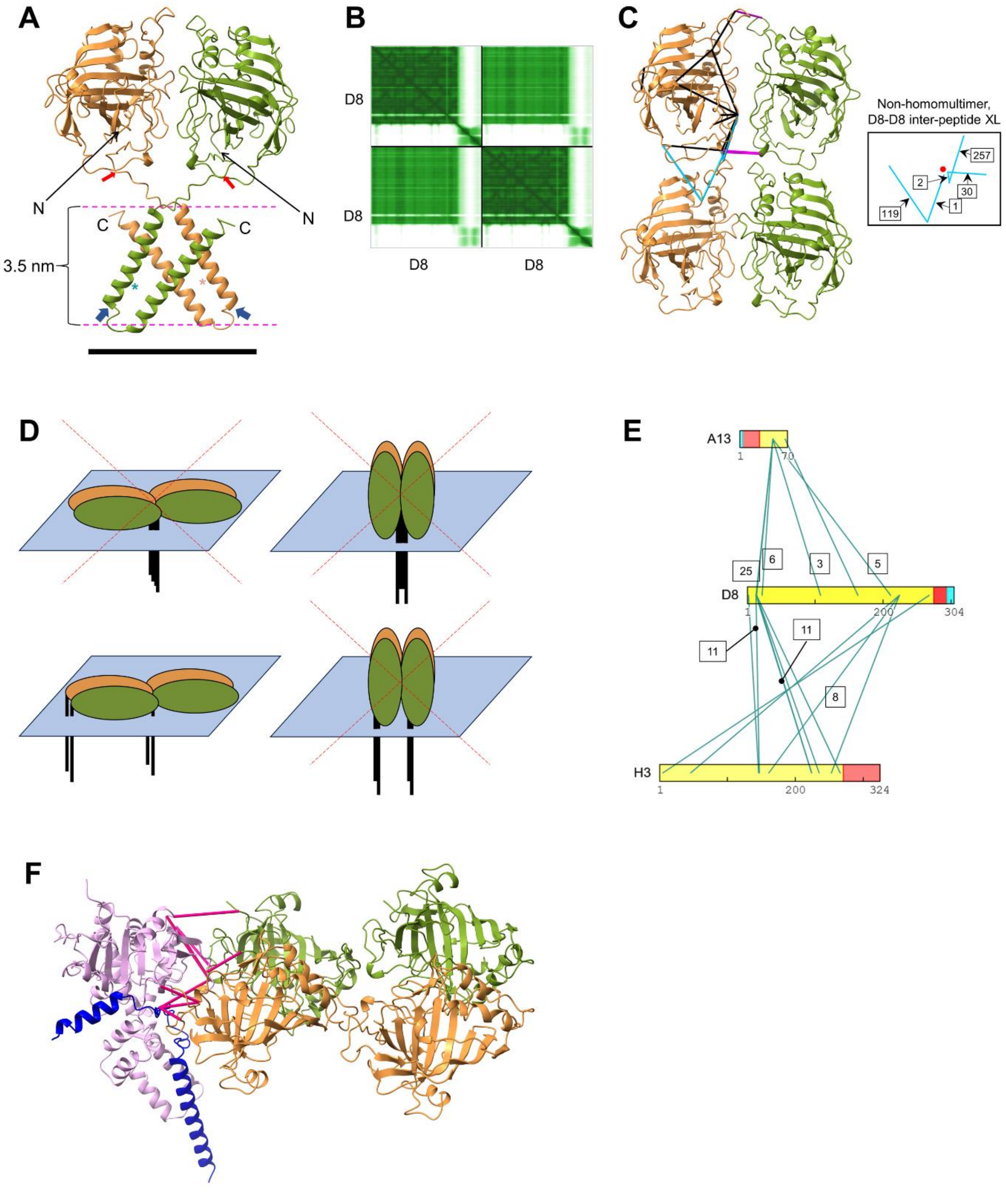


Figure 6.19. D8 multimerization. (**A**) Antiparallel homodimer model of full-length D8, modeled by AlphaFold-multimer. Colors distinguish subunits. Asterisks mark residue C262 in each subunit. Red arrows: Residue 236, the position of truncation in the model of panel (C). Blue arrows: Residue 269, the truncation point for the model of Fig. 17A. Scale bar = 5 nm. Pink dotted lines depict a standard hydrated lipid bilayer width of 3.5 nm. (**B**) PAE plot of the D8 homodimer model in panel (A): Linkers to C-terminal domains were predicted as highly flexible, with AlphaFold-multimer unable to capture the disulfide bond at C262. (**C**) D8 tetramer, by XL-

guided manual docking of two D8 homodimer ectodomains (panel (A); residues 1 - 236). Individual homodimers are colored as in panel (A). Magenta: Homomultimer XL, from Fig. 20. For remaining D8-D8 interpeptide XL, those rationalizable within a chain are shown black (on upper right subunit only albeit they map within each subunit), those not rationalizable within a chain but which could be rationalized upon manual docking of homodimers are shown cyan (on the left-hand monomer pair albeit they map also on the right-hand monomer pair). Since it is the cyan crosslinking network that drives homodimer docking to homotetramer, these are shown alone to the right side of panel (C), with CSM counts boxed. The resulting model satisfied 98.6% of inter-protein XL between D8 ectodomains. (**D**) Schematic of homotetramer theoretical topologies with respect to a horizontal bilayer. Ectodomains are colored as in panel (C), TM domains depicted as black sticks. Left: Ectodomains horizontal. Right: Ectodomains vertical. Upper: TM domains bundled (dimer mirror symmetry). Lower: TM domains parallel (dimer translational symmetry). Light blue: Lipid bilayer surface. The homotetramer model of panel (C) supported the topology on the lower left. (**E**) XL network for proteins H3, A13, and D8. XL were identified from intact virions and virions treated briefly with uncoating reagents (NP-40/TCEP). Boxed numbers: As in Fig. 14. (**F**) H3:A13 (purple and blue respectively) interaction with D8 tetramer (residues 1 - 236, colored as in panel (C)), by XL-guided manual docking. Magenta: Valid inter-protein XL between H3:D8 and A13:D8. These constituted 25 of 26 such CSMs from intact virions (96%) and 43 of 55 from virions briefly treated with NP-40/TCEP (78%).

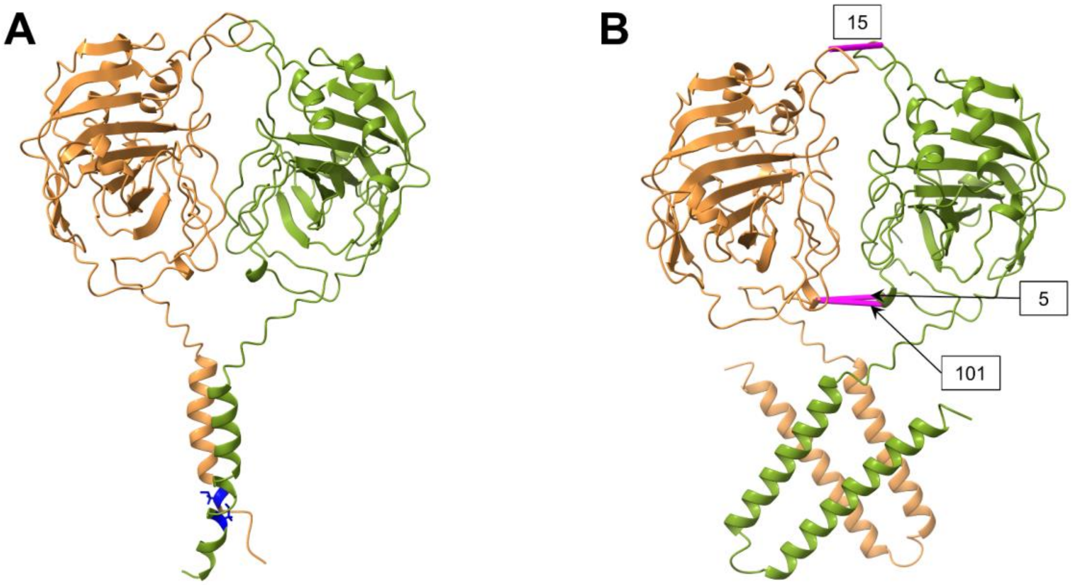


Figure 6.20. **D8 homodimer models.** (**A**) Homodimer of truncated D8 (residues 1 – 268 only), with subunits colored as in Fig. 19A. Upon truncation, C262 (blue) alpha-carbons were within disulfide range. C262 side chains, however, were not within disulfide bonding range. (**B**) D8 homodimer model from AlphaFold-multimer, with those homomultimer crosslinks falling with valid crosslinking range colored magenta. Boxed numbers show the CSM count for each crosslink.

Our expanded XLMS network showed D8 as a crosslinking partner for A13 and H3 (Fig. 19E). By XL-guided manual docking of the H3-A13 heterodimer and the D8 homotetramer, we could identify an arrangement that satisfied 96% of H3:D8 and A13:D8 XL identified from intact virion and 78% of such XL identified from uncoated virions (Fig. 19E, F). Crosslinks between the D8 homotetramer and the H3-A13 heterodimer were identified primarily at the same surface of H3-A13 that we manually docked to the $A17_2A27_3$ subcomplex by XL-guided docking, suggesting that this H3-A13 binding interface is the critical one for interaction of H3 and A13 with other envelope proteins.

Entry-Fusion Complex (EFC)

Poxviruses encode 11 conserved proteins required for virus entry [134]. All 11 EFC proteins can be visualized co-localized at the “tips” (regions of greatest curvature) of the oval or brick shaped MV [98] and can be co-purified. Of these proteins, nine (A16, A21, A28, G3, G9, H2, J5, L5, and O3) are regarded as a core complex, with L1 and F9 associated at the periphery [134, 143]. Mutation of any one of the 11 EFC proteins or repression of any of the nine core proteins limits complex formation to subassemblies only [141]. The resulting MV appear morphologically normal and can proceed through cellular attachment and the initial (hemifusion) step of virus entry, but they are defective in completing the final entry steps of pore formation and core release into the host cell cytoplasm [149, 150, 159, 675-678]. All EFC proteins possess an N- or C-terminal membrane insertion domain.

Experimental atomic structures for L1 and F9 [118, 160] and NMR assignments for truncated A28 have been reported [679]. During preparation of this manuscript, the crystal structures of the ectodomains of the G3/L5 heterodimer [162] and the A16/G9 heterodimer [161] were reported. The structural predictions reported here were constructed unaided by the A16/G9 and G3/L5 atomic models`, since the PDB70 repository used by AlphaFold2 predates the deposition of these structures to PDB. While interactions between individual EFC members and the formation of subcomplexes has been demonstrated by co-immunoprecipitation`, XLMS`, and tripartite split GFP complementation assays [141, 142, 604, 680], unknowns include atomic structures for H2, A28, A21 and O3, complete structures for subcomplexes (with the exception of G3:L5 and A16:G9) and the overall architecture of the EFC. Here, structure prediction attempts for the EFC as a single, comprehensive complex were unsuccessful. However, via a combination of XLMS and iterative modeling, the EFC could be resolved into 5 binary subcomplexes which could be condensed, in turn, into two principal subassemblies connected via their membrane spanning helices, with L1 associating peripherally.

Proteins A16, G9, J5, and F9**:** EFC proteins A16, G9, and J5 are paralogs with low sequence identity. A16 and G9 form a binary complex that can co-immunoprecipitate independently of other EFC members [163]. In MV, this complex interacts with fusion suppressor protein A26 (below), directing virus entry through the endocytic pathway. Within infected cells, interaction between the A16:G9 complex and the complex of vaccinia proteins A56:K2 (below), prevents superinfection of cells expressing A56:K2 on their surface [676, 681].

Here, AlphaFold2 predicted high confidence structures for proteins A16, G9 and J5 (Fig. 21A-C, Appendix 3.Table 3). Although the three proteins have low sequence identity, they showed surprisingly high predicted structural homology across multiple regions. For example, the a-helical region located towards the C-terminal membrane spanning helices was conserved among the three proteins (Fig. 22A, yellow), with good structural alignment between them (Fig. 22B). This domain showed no structural homologs outside of the poxvirus family, and within the poxvirus family the only homologs were these three proteins. The three proteins also showed a pair of b-strands immediately C-terminal to this alpha-helical domain (Fig. 22A). The anti-parallel coiled coil superhelical domain in A16 and G9 (residues 139 – 222 and 85 – 204, respectively; Fig. 22A, purple), also showed strong structural alignment (Fig. 22C). These domains are homologous to proteins with alpha-alpha superhelical folds, including prenyltransferase alpha-subunits, HEAT repeat proteins, and others, with a maximum DALI z-score of 6.8 (Appendix 3.Table 3). This fold encompasses 24 protein superfamilies, many with

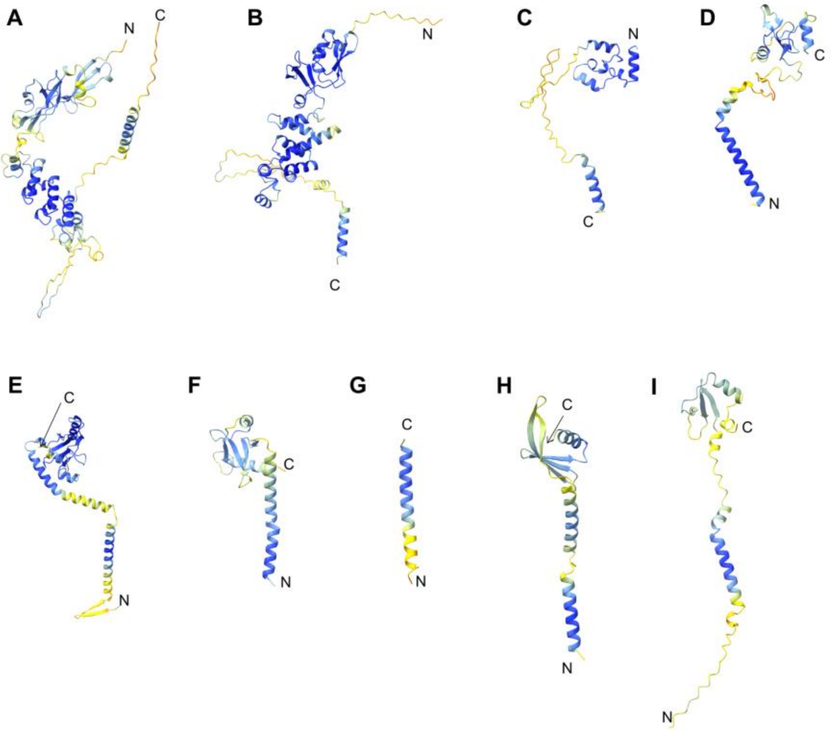


Figure 6.21. AlphaFold2 structure predictions (colored by pLDDT) for individual EFC proteins. (**A**) A16; (**B**) G9; (**C**) J5; (**D**) A28; (**E**) H2, (**F**) A21; (**G**) O3; (**H**) G3; (**I**) L5.

roles in mediating protein-protein interactions [682]. The G9 N-terminal mixed alpha-beta fold (residues 20 – 80) (Fig. 22A, blue) was structurally homologous to the N-terminal domain of human 40S ribosomal protein S15a (Appendix 3.Table 3), with a DALI z-score of 4.4. Given that no nucleic acid binding activity has been demonstrated for G9, and G9 mutations within this motif (particularly H44Y) allow virions to overcome fusion inhibition by A56:K2 [676, 681], we suggest this motif to be an acquired ancestral fold. No structural homologs were identified for the A16 N-terminal domain.

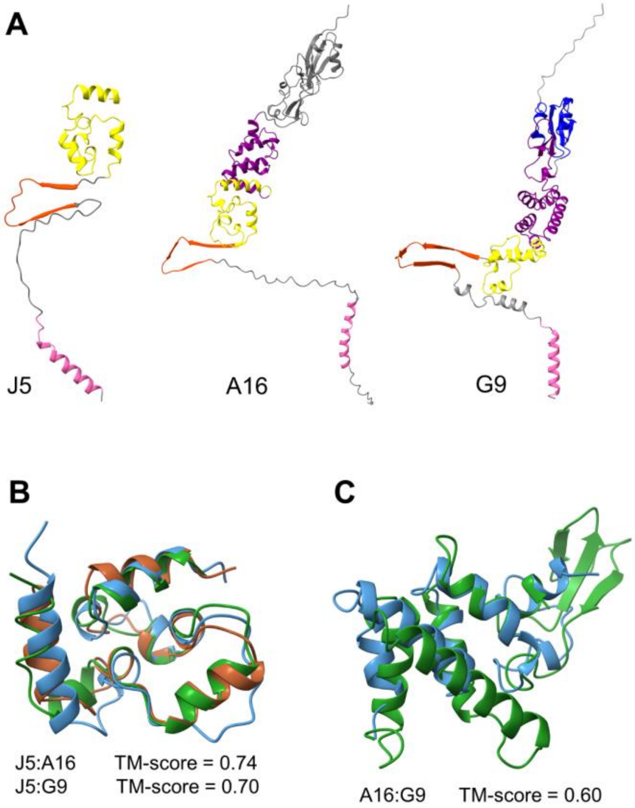


Figure 6.22. **Structurally homologous domains between EFC members A16, G9, and J5.** (**A**) AlphaFold-multimer predicted structures of J5, A16 and G9 colored according to homologous/conserved domains and structural features. Yellow: alpha-helical domain. Orange: Conserved beta strands. Purple: Alpha-alpha superhelical (antiparallel coiled-coil) domain (shared by A16 and G9). Blue: Mixed alpha-beta fold (found in 40S ribosomal protein S15a). Pink: C-terminal membrane insertion helices. Gray: No structural homology within the group or to other proteins. (**B**) Alignment of alpha-helical domains: J5 residues 1 - 65 (brown), A16 residues 223 - 293 (blue) and G9 residues 204 - 267 (green). (**C**) Alignment of alpha-alpha superhelical domains: A16 residues 139 - 222 (blue) and G9 residues 85-203 (green).

J5, A16, and G9 all contain cysteines that are conserved between poxviruses and considered to form intramolecular disulfide bonds [134]. AlphaFold2 predicted structures for all three proteins placed various cysteine pairs within appropriate intramolecular disulfide bonding range (Fig. 23A - C).

Given A16 and G9's known formation of a subcomplex, we modeled this with AlphaFold-multimer. The predicted heterodimer (Fig. 24A) showed a very intimate and extensive interaction between the two proteins including interaction of the a-a superhelical folds, with a highly confident PAE plot (Fig. 25A). The b-strand pair from each protein (above) packed into a four-stranded b-sheet, while the membrane spanning helices were highly flexible based on the corresponding PAE plot (Fig. 25A). An X-ray crystal structure of the A16:G9 heterodimer [161] was released – during the preparation of this manuscript. Retroactively, to compare the accuracy of the AlphaFold-multimer prediction with the reported X-ray crystal structure, we generated new PDB files for the predicted and experimental heterodimers in each of which A16 and G9 coordinates were merged into a single chain, excluding from the AlphaFold-multimer model residues absent in the X-ray crystal structure file. The single chain PDB coordinates were then compared by TM-Align, with very close alignment observed that accurately predicted the heterodimeric arrangement of A16 and G9 (overall TM-score of 0.95 and RMSD of 2.07 Å; Fig. 26A).

We next predicted the remaining EFC subcomplexes. Our earlier XLMS study [604] identified EFC protein F9 as a crosslinkable partner of J5. AlphaFold-multimer predicted a heterodimeric subcomplex of the two proteins (Fig. 24B) with the resulting PAE plot showing high confidence in the prediction and placement of the J5 and F9 ectodomains (Fig. 25B). The F9 ectodomain has been reported to comprise an alpha-helical bundle and a pair of beta sheets [160].

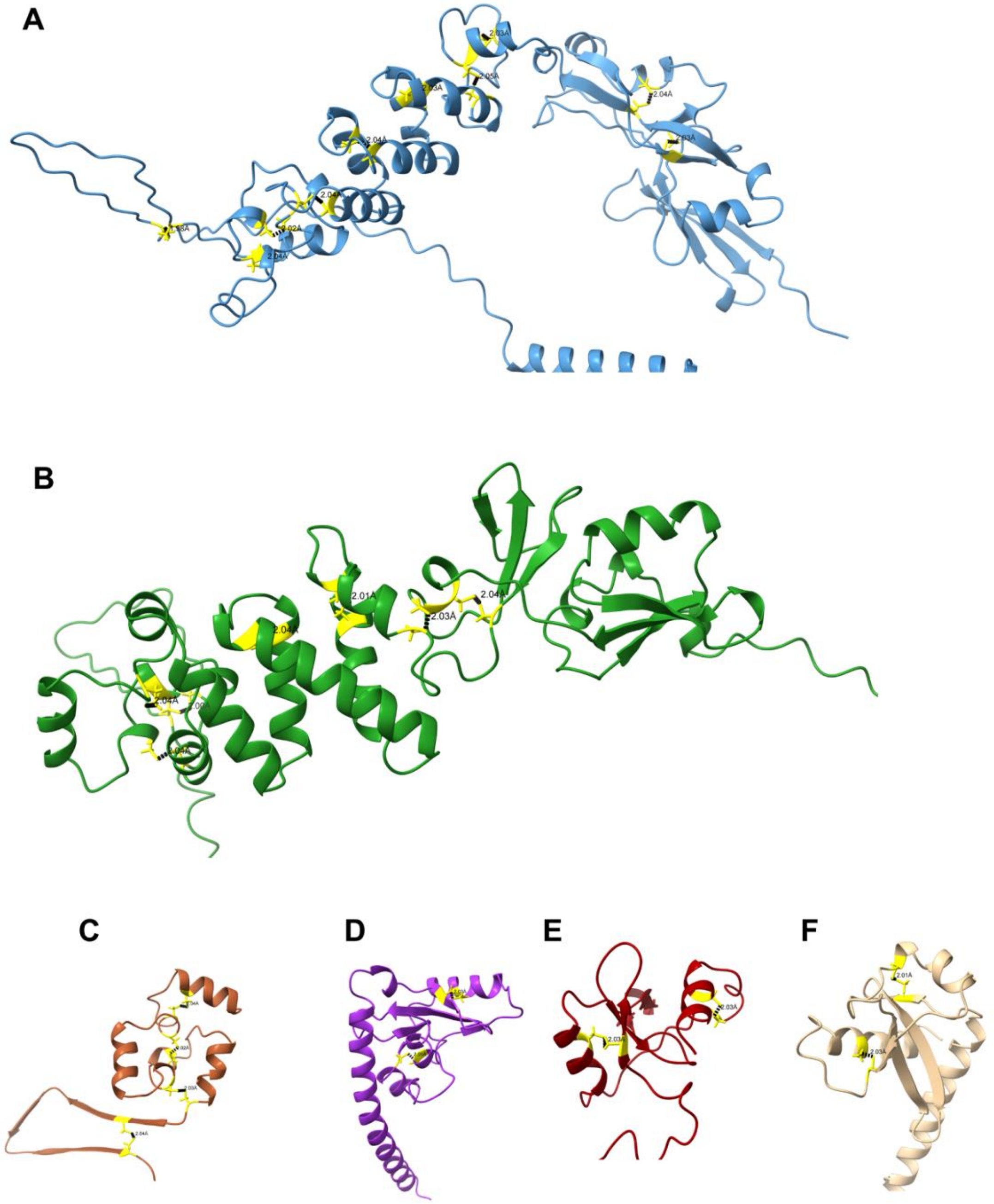


Figure 6.23. **AlphaFold2 predicted structures show the placement of cysteine side chains at close range, supporting intramolecular disulfide bonding.** Cysteine pairs and distances between cysteine sulfur atoms are shown for proteins: (**A**) A16 (C60:C90, C70:C128, C:146:C155, C147:C168, C176:C185, C204:C213, C236:C245, C247:C270, C265:C291 and C296:C316). (**B**) G9 (C88:C117, C89:C127, C135:C145, C177:C186; C223:C248, C243:C267 and C272:C291). (**C**) J5 (C10:C19, C21:C46, C41:C64 and C69:C89). (**D**) H2 (C102:C148 and C162:C182). (**E**) A28 (C75:C112 and C129:C139). (**F**) A21 (C45:C75 and C92:C106).

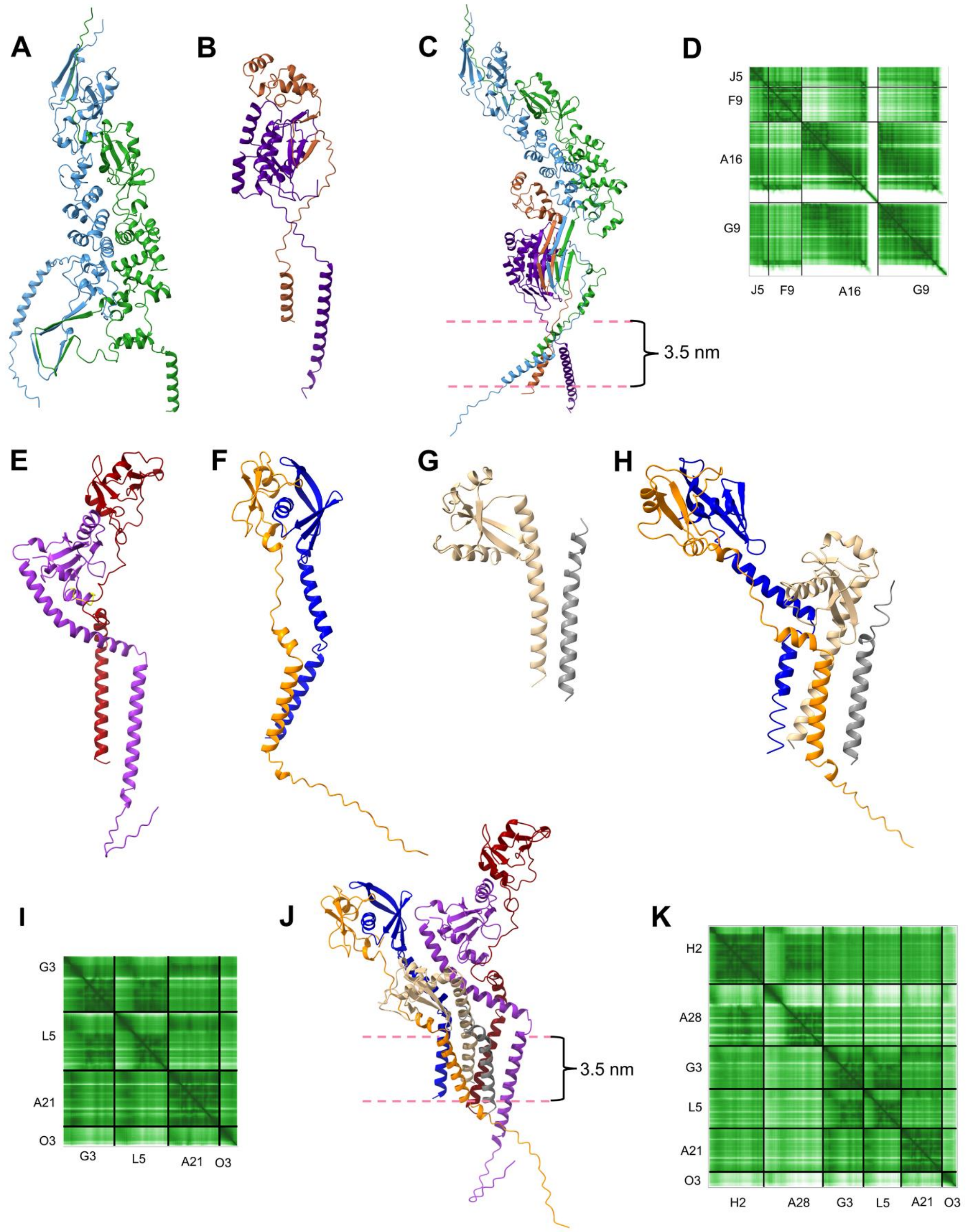


Figure 6.24. **EFC subcomplexes and subassemblies.** (**A**) A16 (blue):G9 (green) subcomplex. (**B**) J5 (brown):F9 (indigo) subcomplex. (**C**) Subassembly of A16:G9 and J5:F9 subcomplexes, with chains colored as in panels (A) and (B). Pink dotted lines indicate the virion envelope thickness. (**D)** PAE plot of the A16:G9:J5:F9 subassembly. Panels (**E**) – (**G**): AlphaFold-multimer predicted heterodimers of (**E**) A28 (red):H2 (purple); (**F**) of G3 (blue):L5 (gold); (**G**) A21 (tan) and O3 (grey). (**H**) Predicted heterotetrameric subassembly of the G3:L5 and A21:O3 subcomplexes, colored as in panels (F) and (G). (**I**) PAE plot of G3:L5:A21:O3 subassembly showing high confidence in both the ectodomain and membrane insertion helix placements. (**J**) A28:H2, G3:L5, and A21:O3 subcomplexes coalesce into a larger subassembly, with chains colored as in panels (**E**) – (**G**). Pink dotted lines indicate virion envelope thickness. (**K**) PAE plot of the H2:A28:G3:L5:A21:O3 subassembly shows high confidence in ectodomain placement.

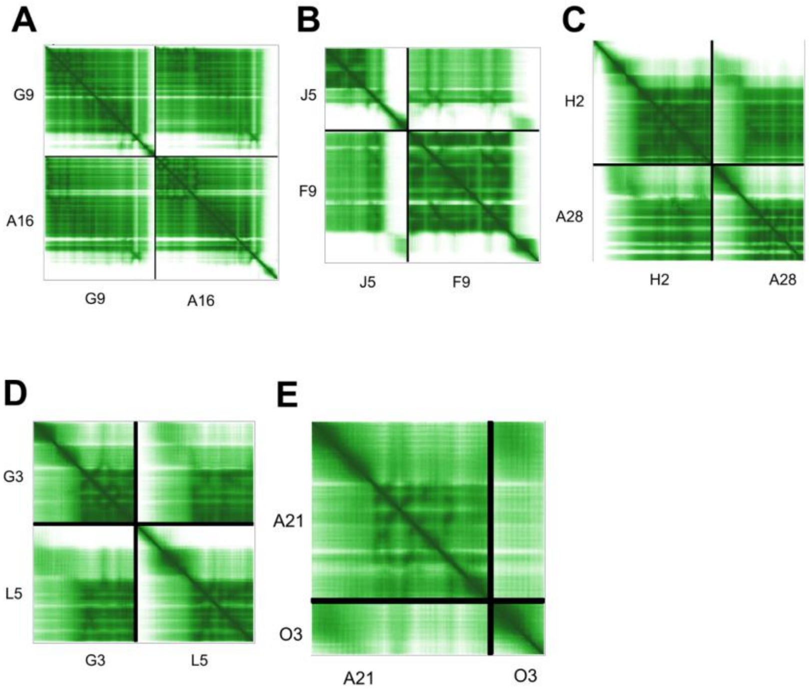


Figure 6.25. **PAE plots of AlphaFold-multimer predicted structures for EFC subcomplexes (Fig. 24), with high confidence in ectodomain modeling and placement.** (**A**) G9:A16 (Fig. 24A). (**B**) J5:F9 (Fig. 24B). (**C**) H2:A28 (Fig. 24E). (**D**) G3:L5 (Fig. 24F). (**E**) A21:O3 (Fig. 24G), showing moderate confidence in the placement of O3 with respect to A21.

Here, we show the F9 beta sheets packing together with J5 beta strands in a manner comparable to the interaction between the A16 and G9 C-terminal beta strands.

Our current, expanded XLMS dataset showed J5 and A16 as crosslinking partners (Fig. 27) suggesting that the J5:F9 and A16:G9 heterodimers may coalesce into a higher order assembly. A direct interaction between these two subcomplexes has not been previously reported. After partial truncation of the membrane spanning helices, we could model this entire subassembly with high confidence (Fig. 24C, D). The truncated membrane spanning helices were added back manually in ChimeraX. The structural model of this heterotetramer shows J5 and A16

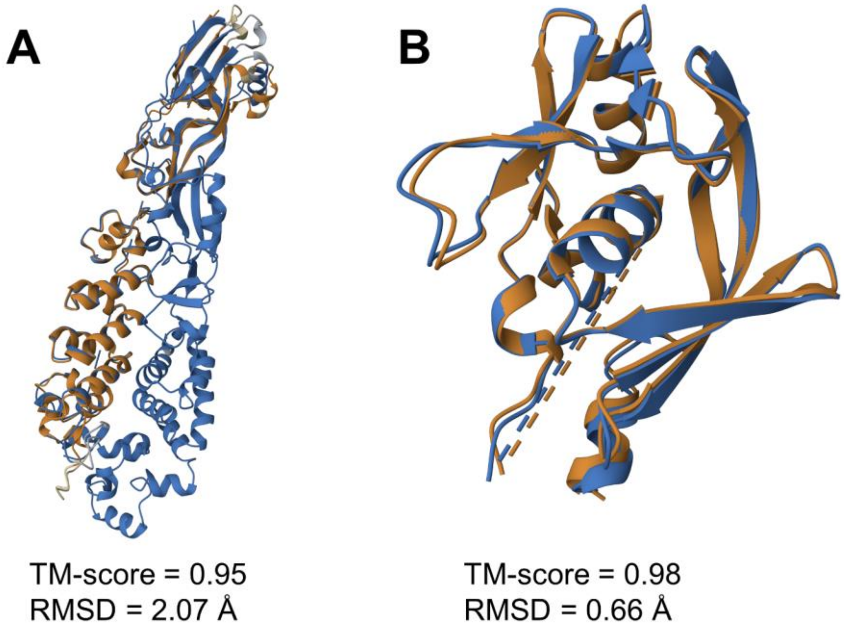


Figure 6.26. **Structural alignment between AlphaFold-multimer predicted structures of EFC subcomplexes and reported crystal structures.** Structures were converted to single-chain coordinate files then aligned via TM-Align. Aligned residues are shown orange/blue while residues not in alignment are colored tan/grey. Dashed lines are an artifact from combining multiple PDB chains into single entries and do not represent any residue coordinates. (**A**) Ectodomain of the A16:G9 subcomplex (orange/tan) predicted by AlphaFold-multimer aligned with its crystal structure (PDB 8GP6) (blue/grey). (**B**) G3:L5 predicted structure (orange/tan) aligned with its crystal structure (PDB 7YTT).

mediating interaction of the two heterodimers consistent with our XLMS data. The J5:A16 interaction occurs through the pair's homologous alpha helical domains and A16's alpha-alpha superhelical domain. The beta strands of all four proteins co-fold into an extensive beta sheet (Fig. 24C, Fig. 28).

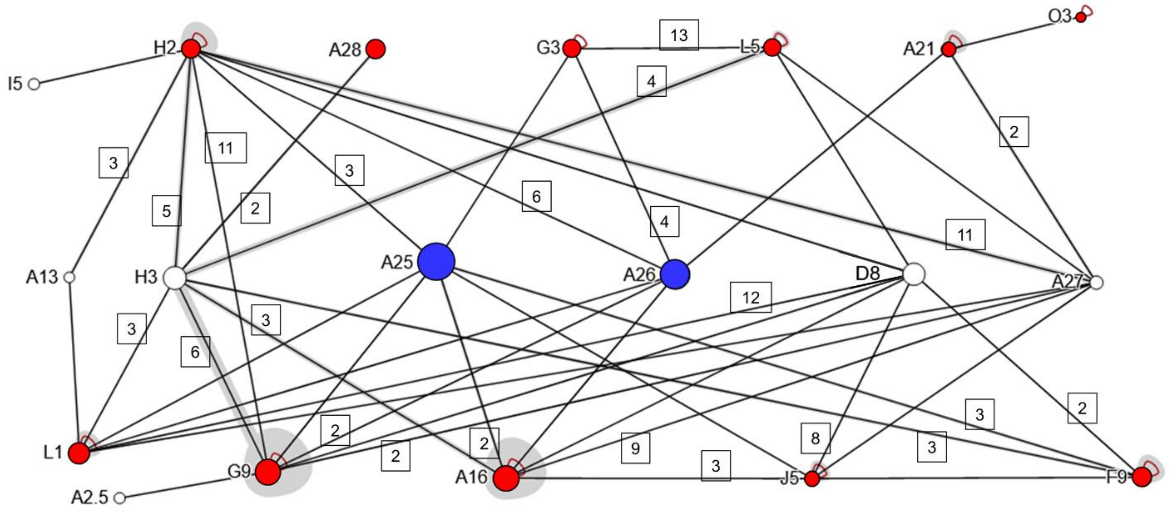


Figure 6.27. **XL network of EFC-EFC protein interactions and EFC interactions with other MV envelope proteins.** Circle areas represent protein size. EFC proteins (red), fusion inhibitors (blue), other envelope proteins (no fill). Red loops on individual proteins represent self-interactions detected as homomultimer-XL. Boxed numbers: As in Fig. 14.

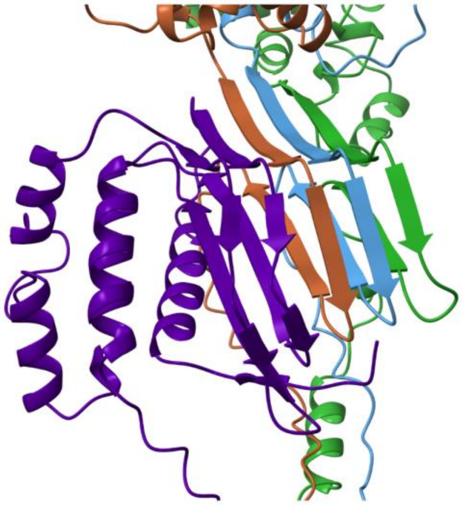


Figure 6.28. **A16, G9, J5, and F9 beta strands co-model to form an extended beta sheet.** Protein chains are colored as in Fig. 24A-C.

Proteins A28, H2: EFC members A28 and H2 form a known subcomplex [142, 157]. Via AlphaFold2, high confidence structural models were obtained for monomeric A28 and H2 (Fig. 21D, E; Appendix 3.Table 3), with the ectodomains of both proteins forming mixed alpha-beta folds. No confident structural homolog was identified for A28. As with A16, G9 and J5 (above), various cysteine pairs in A28 and H2 were within acceptable intramolecular disulfide bonding range (Fig. 23D, E) – within regions of sufficiently high pLDDT to be confident in cysteine sidechain placement. The H2 structure was validated by XLMS from intact MV (Fig. 29A), with a combined CSM score of 42 (out of 43 total crosslinks) agreeing with the structure prediction.

Via AlphaFold-multimer we predicted a structure for the A28:H2 subcomplex (Fig. 24E), with high confidence in the relative orientations of the C-terminal ectodomains of the two proteins (Fig. 25C).

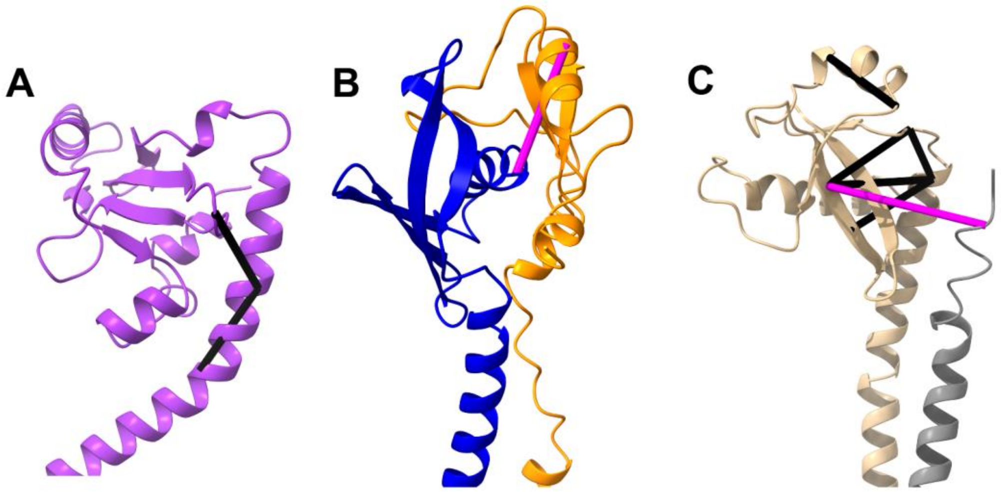


Figure 6.29. **AlphaFold2 and AlphaFold-multimer predicted structures are validated by intra-protein and inter-protein crosslinks.** Black bars represent intra-protein crosslinks. Magenta bars represent inter-protein crosslinks. (**A**) H2, (**B**) G3-L5, (**C**) A21-O3.

H2 contains a highly conserved sequence "LGYSG" comparable to the fusion motifs of fusion proteins from flaviviruses, retroviruses and hepatitis B virus [157]. This sequence is required for the association of H2 with A28. Interaction at these residues may also serve to conceal the putative fusion motif until needed, with its exposure following conformational changes driven by acidification or receptor binding [157]. Consistent with this, A28 residues 42 – 52 in the heterodimer predicted structure, which are conserved between poxviruses (Fig. 30A), appear to interact with H2's "LGYSG" sequence directly, partially obscuring it (Fig. 30B). H2:A28 docking might be mediated by electrostatic interactions according to differences in their calculated surface charge distributions at the docking interface (Fig. 31A, B).

Proteins G3, L5, A21, and O3: EFC proteins G3 and L5 have been reported to associate, as detected by XLMS, co-IP, and a tripartite split GFP complementation assay [142, 604, 678, 680]. AlphaFold2 reported high confidence monomeric structures for the two proteins (Fig. 21H, I, Appendix 3.Table 3), with average pLDDT of 83.4 and 78.0 respectively. Via AlphaFold-multimer, a structure for the G3:L5 subcomplex was predicted with high confidence (Fig. 24F, Fig. 25D). It showed very strong alignment with the G3:L5 ectodomain crystal structure (Fig. 26B) that was reported during preparation of this manuscript [162], which resulted in a TM-score of 0.98 and RMSD of 0.66 Å (Fig. 26B). Both structures agreed with crosslinking data (Fig. 29B) as previously reported [604], with a combined CSM count of 13.

Our expanded XLMS dataset also showed A21 and O3 as crosslinking partners (Fig. 27). This interaction is supported by the tripartite split GFP complementation assay [142]. Although A21 co-purifies with other EFC members, to date it has only been shown to associate directly with EFC protein O3 [142]. Via AlphaFold2, we generated a high confidence structure for A21 (Fig. 21F) with an average pLDDT of 81.9 (Appendix 3.Table 3). We validated the predicted structure against our XLMS data. The A21 structure agreed with all identified intra-protein XL (Fig. 29C) with a combined CSM score of 77. DALI identified partial homology between the A21 C-terminal ectodomain and baculovirus polyhedron envelope protein Orf22 from Cydia

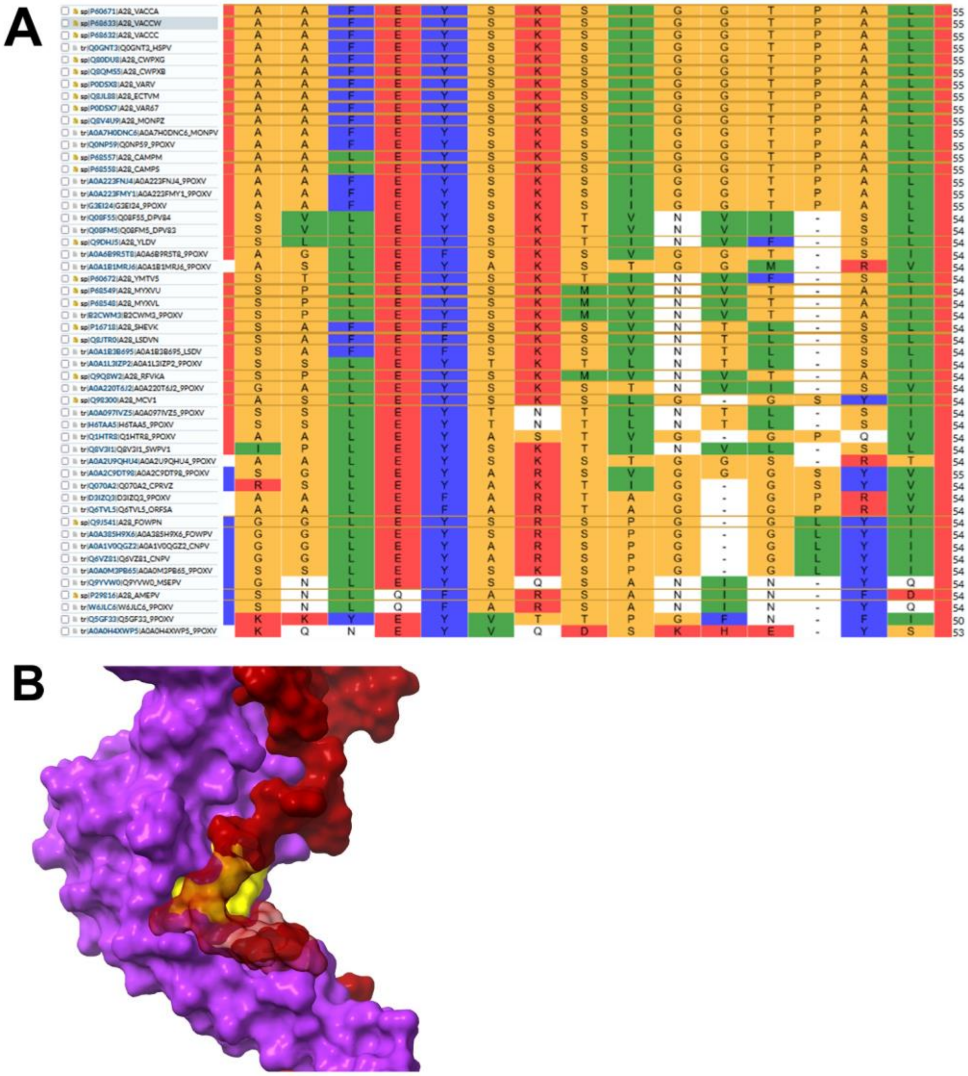


Figure 6.30. **A28 conserved residues 42 - 52 interact directly with the H2 putative fusion motif.** (**A**) Multiple sequence alignment (Clustal Omega) between vaccinia protein A28 and homologous poxvirus proteins, showing residues 41 - 55. Residues are colored by Clustal, showing similarity in amino acid properties. (**B**) H2 putative fusion motif (residues 170 – 174; yellow) are partially obscured by A28 residues 42 - 52 (transparent red).

pomonella granulosis virus (PDB 4YE7), with a z-score of 3.8 (Appendix 3.Table 3). Alignment (TM-Align) showed a TM-score of 0.45, suggesting that A21 and Orf22 have similar features, but do not possess the same fold (data not shown). We also generated a structure for O3 (Fig. 21G) which, as previously suggested [683], comprised a short alpha-helix.

Using AlphaFold-multimer, a heterodimeric structure of the A21:O3 subcomplex was predicted (Fig. 24G), which was validated by XLMS (Fig. 29C). The PAE plot for this subcomplex showed moderate confidence in its overall structure (Fig. 25E). Seeking further associations of A21:O3 and other EFC subcomplexes via iterative modeling, we could confidently co-model A21:O3 with the G3:L5 subcomplex following partial truncation of the membrane spanning helices (Fig. 24H). The truncated helices were manually added back in with ChimeraX. This subassembly showed a notably increased PAE confidence for both subcomplexes along with high confidence in the mutual placement of both (Fig. 24I). When co-modeled, G3, L5, and A21 also showed moderately increased average pLDDT (not shown). The predicted heterotetrameric structure retained paired interactions predicted in the G3:L5 and A21:O3 subcomplexes (above), with the four membrane insertion helices packing together and the A21 C-terminal alpha helix docking against G3 and L5 ectodomain “linkers” (Fig. 32A). As with H2 and A28, electrostatic forces may mediate the A21, G3, and L5 interaction based on the electrostatic charge differential at the G3:L5 and A21:O3 docking interface (Fig. 32B).

Consolidation of subcomplexes into larger subassemblies: The G3:L5, A21:O3, and H2:A28 subcomplexes could be modeled as a larger subassembly via AlphaFold-multimer, following partial truncation of the membrane spanning helices (Fig. 24J, K). The truncated

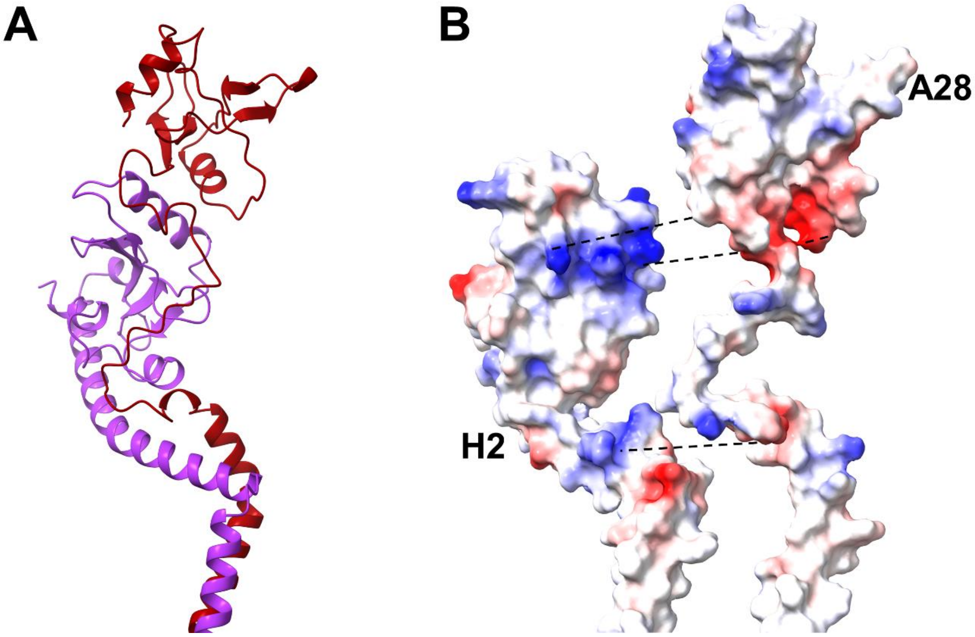


Figure 6.31. Electrostatic interactions between H2 and A28. (**A**) The H2:A28 subcomplex, colored by chain (H2 = purple, A28 = red). (**B**) Space-fill model of the H2:A28 subcomplex, colored by coulombic electrostatic potential (Red: Negative; blue: Positive), to showing complementary surface charge distributions at the H2:A28 docking interface, with H2 and A28 separated manually to view the docking interface. Dotted lines show how the separated molecules would dock.

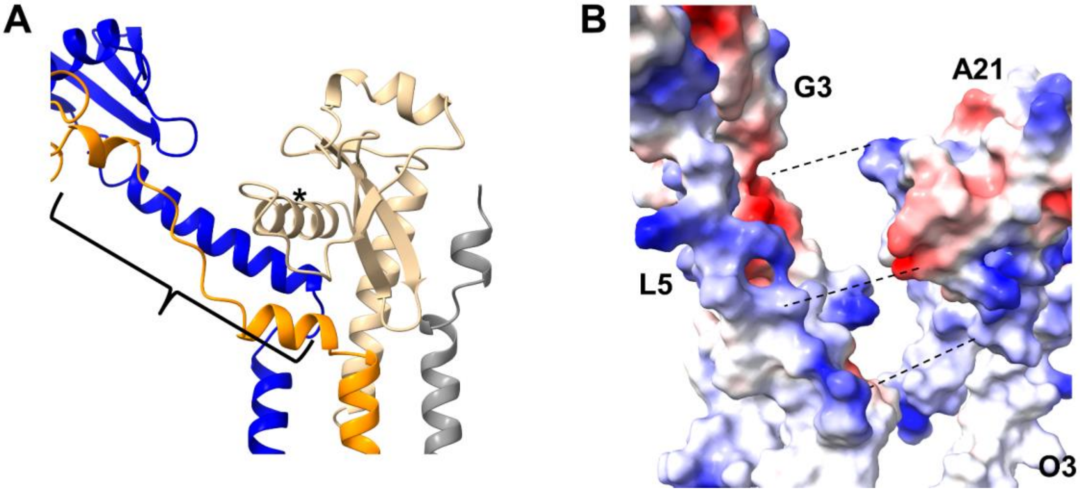


Figure 6.32. **G3:L5 and A21:O3 coalesce to a larger subassembly. (A)** The A21 C-terminal helix (denoted '*****') packs against the G3 and L5 "linkers" (black bracket). Chains are colored as in Fig. 24F-G. **(B)** Space-fill model of the heterotetramer, colored by coulombic electrostatic potential (Red: Negative; Blue: Positive) with G3:L5 and A21:O3 separated manually to view the docking interface. Dotted lines show how the separated molecules would dock.

helices were restored manually using ChimeraX. This subassembly coalesced via interactions between H2, G3, and A21.

We have reported, previously, the interaction of G9 with H2 by XLMS [604]. Our expanded XLMS dataset reaffirmed this interaction (Fig. 27), with a CSM count raised from 1 to 11. Short-range crosslinks between the membrane spanning helices of G9 and H2 (data not shown) suggest that the A21:O3:G3:L5:H2:A28 and A16:G9:F9:J5 subassemblies may interact via their membrane spanning helices, since attempts to co-model A16:G9 with H2:A28 succeeded only in bringing together the membrane spanning helices, with variable placement of the H2:A28 ectodomains with respect to A16:G9 (data not shown).

By XLMS, all EFC members except A28 were found to self-associate, as identified by homomultimer crosslinks (Fig. 27), suggesting that multiple EFC complexes may be very clustered on the surface of MV or that individual EFC complexes coalesce into a higher order assembly.

Other MV membrane proteins (A9, I5)

Vaccinia envelope proteins A9 and I5 are short transmembrane proteins (108 aa and 79 aa respectively). Both are reported to contain dual membrane spanning helices, both with their N- and C- termini oriented towards the outside of the virion envelope [684-686]. AlphaFold2 models (Fig. 33A, B) and the overall crosslinking pattern for both proteins (Fig. 33C, D) supported this topology. Attempts to confidently dock A9 and I5 to other envelope proteins by XL-guided manual docking or AlphaFold-multimer were unfruitful.

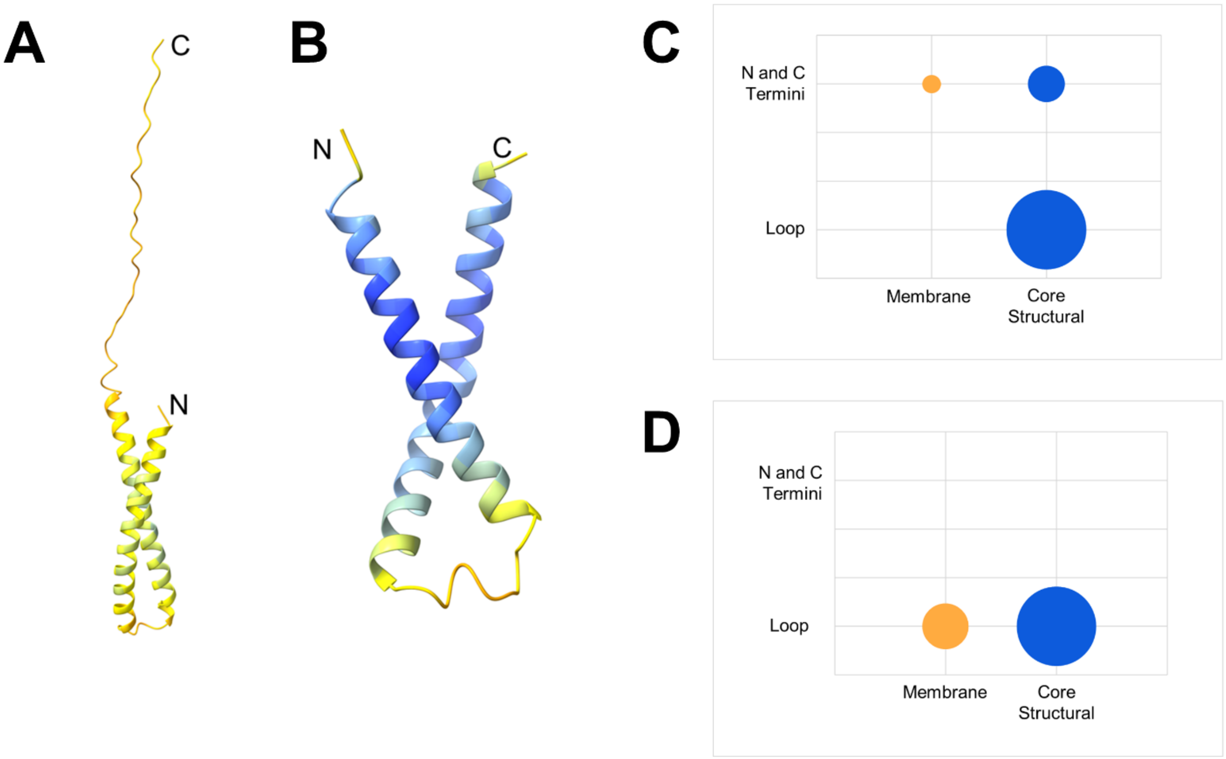


Figure 6.33. **Vaccinia membrane proteins A9 and I5.** (**A-B**) AlphaFold2 models of A9 and I5, colored by pLDDT. (**A**) A9. (**B**) I5. (**C-D**) Crosslinking emphasis of the A9 and I5 N- and C-termini (indicated as N and C termini) vs their hydrophilic loop (indicated as loop) to other membrane proteins (yellow) or proteins of the virion core wall (blue). Circle area sizes represent total CSM counts. (**C**) A9 (**D**) I5. The circle representing the crosslink between the I5 hydrophilic loop and membrane proteins is a rational crosslink that was detected between the loop region of I5 and the protein N-terminus of H2 (Fig. 21E), which is expected to be on the inside of the virion.

MATERIALS & METHODS

Protein Structure Prediction and Validation

Vaccinia virus protein structure predictions were made on a local installation of AlphaFold2 (version 2.2.2) and AlphaFold-multimer, using a non-docker setup (https://github.com/kalininalab/alphafold_non_docker). Statistical analyses of the accuracy of AlphaFold2 structures were performed by adapting the methods described in ref. [687]. For each vaccinia protein with an experimentally resolved atomic structure, the earliest PDB deposition date of its atomic coordinates was identified. Each protein was subsequently modeled, in its entirety, twice by AlphaFold2, first with the maximum template release date set to at least 1 year prior to the PDB deposition date (referred to here as "restricted" mode), then with the maximum template release date set to the date of installation of the AlphaFold2 search databases on our local setup – 2022-06-21 (referred to here as "unrestricted" mode).

Alignment and scoring of the top ranked AlphaFold2 models for each protein against experimental structures were performed initially on a per-domain basis (similar to the description in ref. [687]), since some vaccinia multi-domain proteins show a high degree of inter-domain mobility as seen in the experimental structures of vaccinia MCEL and RAP94. Residue delimits included for each domain split are listed in Appendix 3.Table 1. Experimental structures were scanned to identify unresolved residues. Per-domain PDB files were then written for the AlphaFold2 models based on the domain splits in Appendix 3.Table 1 with experimentally unresolved residues excluded from the new PDB files. Per-domain PDB files were also written for the experimentally resolved structures.

AlphaFold2 models and experimental structures were aligned per-domain and scored by GDT_TS and TM-Align. GDT_TS was run on the AS2TS server (http://linum.proteinmodel.org/)

following the instructions here (https://proteopedia.org/wiki/index.php/Calculating_GDT_TS). TM-Align was run on RCSB-PDB (https://www.rcsb.org/alignment). In all instances, the AlphaFold2 model was specified first, as the query structure, with the experimental structure entered second, as the reference structure.

The per-domain GDT_TS, TM-Scores, and RMSD from TM-Align are reported in Appendix 3.Table 1. The average pLDDT for each domain from the AlphaFold2 models was calculated from pLDDT values for all α-carbon atoms included within the listed domain splits and reported in Appendix 3.Table 1. Average pLDDT values were also calculated for the predicted models, in their entirety or covering just the core domain (excluding flexible N- and C-termini) and are reported in Appendix 3.Table 3. From scatter plots of per-domain average pLDDT values with their respective GDT_TS, TM-Score, RMSD values, Pearson correlation coefficients and p-values were calculated.

Structural models for the other 40 - 45 packaged virion proteins were predicted by AlphaFold2 and AlphaFold-multimer with a maximum template release date of 2022-06-21. Proteins were selected for AlphaFold-multimer prediction based on published known protein complexes and our XLMS networks. For entry fusion complex proteins A16, G9, G3, and L5, for which heterodimeric experimental structures of the ectodomains were released during the preparation of this manuscript (and were deposited to PDB subsequent to our AlphaFold2 database setup), TM-Align was used to validate the AlphaFold-multimer models against the experimental structures. Structures were predicted by AlphaFold-multimer for full length chains of the A16:G9 and G3:L5 heterodimer experimental structures. The top scoring AlphaFold-multimer models were used to generate new PDB files, including only the experimentally resolved residues for both protein subcomplexes. Within each file, chain B was appended to chain A with chain B residues renumbered accordingly. Coordinate files of the A16:G9 and G3:L5 experimental structures were downloaded from RCSB and renumbered to match the AlphaFold-multimer renumbered coordinate files. The predicted and experimental structures were aligned by TM-Align to assess the accuracy of the AlphaFold-multimer predictions.

AlphaFold2 and AlphaFold-multimer predicted structures were visualized and images were created using UCSF ChimeraX [688, 689]. Per-residue pLDDT values were visualized using the ChimeraX AlphaFold palette. PAE plots for AlphaFold-multimer structures were obtained for the top ranked "relaxed_model" structure files (identified by aligning all 25 "relaxed_model" predicted structure files against the "ranked_0.pdb" structure file using ChimeraX MatchMaker) with the AlphaFold Error Plot tool in ChimeraX.

Full-length EFC subcomplex models were generated in ChimeraX by aligning monomer or heterodimer models against the larger subassembly model with MatchMaker. Residues truncated from the larger subassemblies for the purpose of AlphaFold-multimer prediction were re-added based on structural alignments.

Euclidean (through space) crosslink distances were visualized in ChimeraX using the "distance" command or by uploading a pseudobond file (.pb). Euclidean and Solvent Accessible Surface distances were calculated by Topolink [521].

XLMS

Vaccinia virion proteins were crosslinked and analyzed as described [604] with some changes. Briefly, vaccinia virus was grown in HeLa S3 cells from ATCC (CCL-2.2) and prepared as described [690] by purification through a cushion of 36% sucrose in 10 mM triethylammonium bicarbonate buffer (TEAB) pH 8.0 followed by two sequential gradients of 24

- 40% sucrose in 10 mM TEAB pH 8.0. Purified virions were resuspended in 0.1 M TEAB, pH 8.5 and crosslinked intact or after brief uncoating treatment (0.05% NP40, 40 mM TCEP, 0.1 M TEAB, pH 8.5). The XL reaction was quenched after 30 - 60 minutes with ammonium bicarbonate buffer or removed by spin desalting into ammonium bicarbonate buffer. Crosslinking reagents bis(succinimidyl) penta(ethylene glycol) (BSPEG5), bis(succinimidyl) nona(ethylene glycol) (BSPEG9), adipic acid dihydrazide (ADH), 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide hydrochloride (EDC), and n-hydroxysuccinimide (NHS) were purchased from ThermoFisher Scientific. Isotopically-coded disuccinimidyl suberate (DSS), bis(sulfosuccinimidyl)suberate (BS3), disuccinimidyl glutarate (DSG), and disuccinimidyl adipate (DSA) were purchased from Creative Molecules Inc. 3,5-bis(((2,5-dioxopyrrolidin-1-yl)oxy) carbonyl)phenyl)phosphonic acid (PhoX) was provided by the Scheltema lab or purchased (Bruker Corporation). PhoX crosslinked samples were prepared and enriched as described in ref. [631] or were enriched with the PRO-Q Diamond Phosphoprotein Enrichment Kit (ThermoFisher Scientific) instead of IMAC prior to mass spectrometry analysis. Appropriate crosslinker concentrations were determined by SDS-PAGE.

Crosslinked samples were disaggregated by various methods, including urea denaturation and an adapted FASP protocol, and proteins digested to peptides enzymatically or with cyanogen bromide as described [604]. Digests were supplemented with formic acid (FA) to 3% final concentration and desalted by C18/SCX stagetip as described [578]. Peptides were eluted with 5% NH_4OH, 80% CH_3CN, 0.1% FA (Buffer X) or with a six-step ammonium acetate gradient in 20% CH_3CN, 0.5% FA followed by a final elution with Buffer X. Samples were dried under vacuum and reconstituted in 0.1% FA in water for mass spectrometry. nanoLC-MS/MS was performed as described [604].

Instrument raw files were used directly or converted to mgf or mzML formats using MSConvert by ProteoWizard. Crosslinks were identified by the programs described [604] and also with pLink2, Kojak2, MetaMorpheus, and XlinkX. XLMS data were consolidated, and various downstream analyses performed using an upgraded version of the in-house code described [604]. Crosslink networks were rendered using CrosslinkViewer [589].

Sequence based structural homology detection of A25L was conducted by HHpred on the MPI Bioinformatics Toolkit server [347, 348]. Transmembrane domains were predicted by DeepTMHMM (https://doi.org/10.1101/2022.04.08.487609) and Quick2D [347, 348]. Multiple sequence alignment of A28 was performed by BLASTP and Clustal Omega online [691].

CHAPTER 7

Combination of deep XLMS with deep learning reveals an ordered rearrangement and assembly of a major protein component of the vaccinia virion

Summary

Vaccinia virus, the prototypical poxvirus and smallpox/monkeypox vaccine, has proven a challenging entity for structural biology, defying many of the approaches leading to molecular and atomic models for other viruses. Via a combination of deep learning and crosslinking mass spectrometry (XLMS) we have developed an atomic-level model and an integrated processing/assembly pathway for a structural component of the vaccinia virion, protein P4a. Within the pathway, proteolytic separation of the C-terminal P4a-3 segment of P4a triggers a massive conformational rotation within the N-terminal P4a-1 segment that becomes fixed by disulfide-locking while removing a steric block to trimerization of the processing intermediate P4a-1+2. These events trigger the proteolytic separation of P4a-2, allowing the assembly of P4a-1 into a hexagonal lattice that encloses the nascent virion core.

Importance: An outstanding problem in the understanding of poxvirus biology is the molecular structure of the mature virion. Via deep learning methods combined with chemical crosslinking mass spectrometry we have addressed the structure and assembly pathway of P4a, a key poxvirus virion core component.

Introduction

Vaccinia, the prototypical member of the poxvirus family, is a large DNA virus with just under 200 genes. It comprised the vaccine used for smallpox eradication and is the current vaccine against monkeypox. On the basis of genomic similarity, all three of these viruses are likely structurally identical. Over the past three decades, substantial insights have been gained into the biology of vaccinia at the functional level: Broad roles have been deduced for most of the essential virus genes via phenotypic analysis of natural or engineered conditional mutants [29]. The ultrastructure of the vaccinia virion has been apparent in outline since the 1960s comprising an ellipsoid or "brick-shaped" structure whose core contains the genome with various enzymes and structural proteins, enclosed by at least one lipid envelope with envelope proteins. The core wall is perhaps the virion's most tangible ultrastructural feature though it likely has substantially less structural rigidity than the capsid of a non-enveloped virus [76, 79, 80, 167]. Atomic structures for 34 of the 75 – 80 proteins packaged within the mature virion (MV) can be found, currently, in RCSB. A much more limited picture has been achieved of the virion's internal molecular architecture. Reasons for this include the size and complexity of vaccinia, its polymorphic nature and overall asymmetry, and the presence of at least one phospholipid envelope. As an intact entity, vaccinia has largely defied the classical approaches of X-ray crystallography and cryoelectron microscopy/tomography.

Chemical crosslinking with bifunctional reagents (crosslinking mass spectrometry; XLMS) can inform structural biology in multiple ways, for example by constraining atomic models of individual proteins, guiding the docking of proteins of known 3D structure, and providing protein molecular interaction network models. In the latter regard, our preliminary XLMS study of vaccinia revealed intra-virion protein-protein networks in outline [604]. We have fundamentally extended this XLMS dataset, resulting in, after stringent thresholding, a saturating dataset comprising ~135,000 crosslink-spectral matches (CSMs) representing ~22,000 unique crosslinks (XL) on the basis of protein accession, crosslink position and crosslinked peptide Mr. To assemble these into a molecular model, a desirable starting point would be atomic models for at least the major structural protein components of the virion - the very entities for which classical structural biology has been unsuccessful at providing atomic models, perhaps as a result of their higher-order complexity and intrinsic insolubility.

Recent deep learning approaches for the *de novo* prediction of protein structure have been described as having, to a large extent, solved the “protein folding problem”. Initially skeptical, particularly with regard to vaccinia structural proteins – the majority of which are unique to the poxviridae - we generated structural predictions merely as placeholders for XLMS network analysis. However, via exhaustive statistical and other benchmarking studies, we came to regard some structural models as approaching experimental accuracy (manuscript in preparation). In the study described here, a vaccinia major structural protein, P4a, was structurally modeled via deep learning methods. In combination with experimental data (XLMS), the resulting structural models yield insights into processing, conformational change, disulfide rearrangement and higher order assembly to the level of the whole-virion.

Results

Vaccinia protein P4a is one of the virion’s most abundant protein components. It is conserved in all known poxviridae with no identifiable sequence or structural homologs outside this family. Conditional mutants under non-permissive conditions exhibit an interruption of normal virion morphogenesis, reduced virus yield and an accumulation in the cytoplasm of abnormal immature

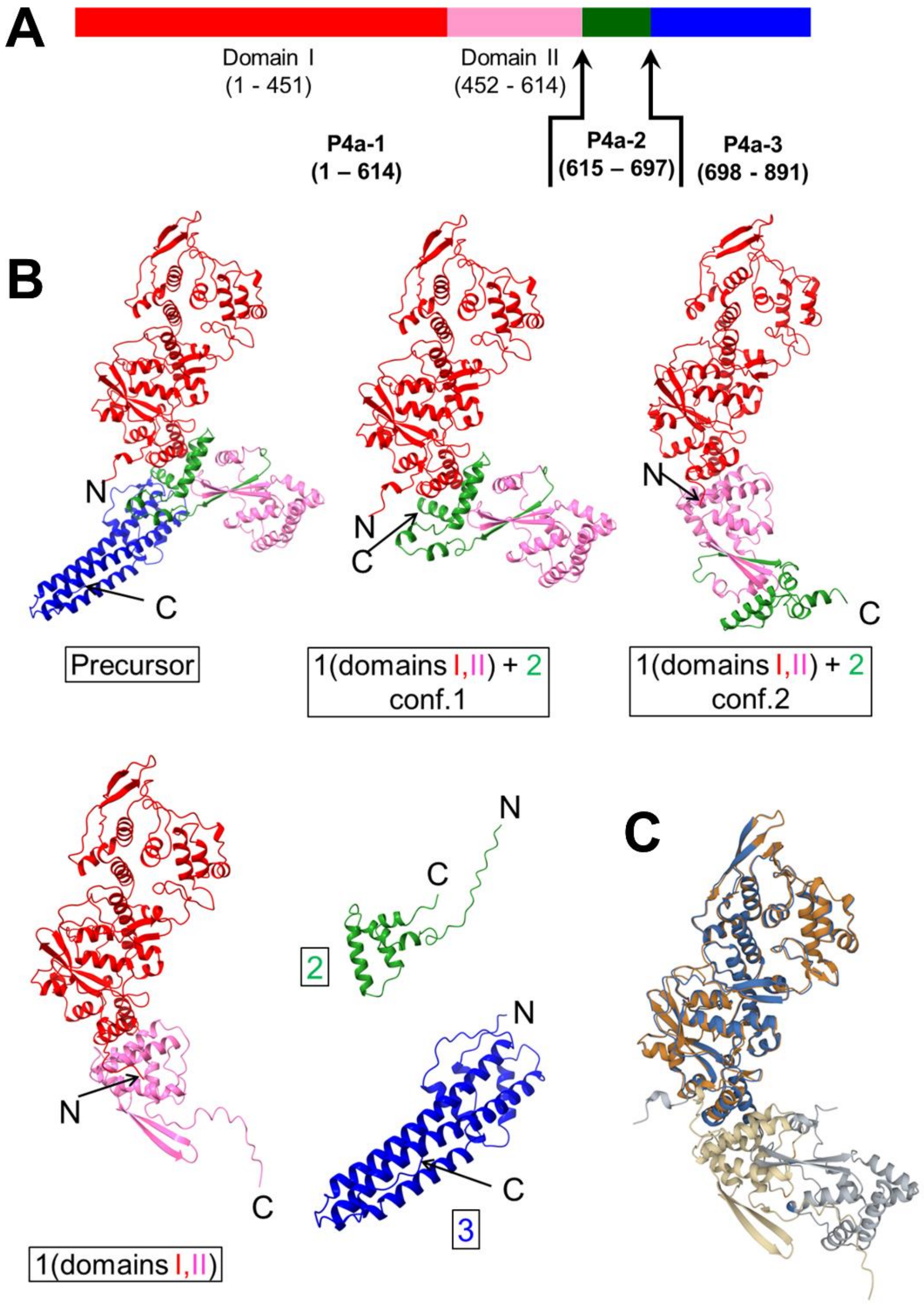


Figure 7.1. Structural models of P4a. (**A**) P4a precursor as a linear chain, colored by processed segments (P4a-1, -2, -3) and domains of P4a-1. Black arrows indicate AG| processing sites. (B) Structural models for P4a precursor and processing products P4a-1, P4a-2, P4a-3, and the two P4a-1+2 intermediate forms. P4a-1 was predicted with two domains denoted here I (red) and II (pink) connected via a hinge. Green and blue: P4a-2 and P4a-3 segments, respectively. Conformational inversion at the hinge corresponded with the removal of P4a-3, the C-terminal processed product of P4a. (**C**) Aligned AlphaFold2 predictions for the P4a-1 segment from P4a-precursor (residues 1 – 614; blue: Aligned residues; gray: Residues not in alignment) and P4a-1 processed product (orange: Aligned residues; tan: Residues not in alignment). P4a-1 domain II has rotated ~180 degrees.

virus (IV)-like particles lacking the normal, dumbbell-shaped core morphology [205, 209, 212, 213, 524-526, 529, 530, 692, 693]. During virion morphogenesis, P4a is proteolytically processed at two sites (at residues 614 and 697) marked by the diamino acid AG| [97, 201, 205, 207, 209, 211, 214, 526, 694]. Processing at these two sites is critical for normal morphogenesis [201, 202, 695-699].

P4a structural models: The three major products of P4a processing are referred to here, in linear order as P4a-1, P4a-2 and P4a-3 (Fig. 7.1A). P4a-1 and -3 are packaged, while P4a-2 is likely discarded and eliminated at the proteasome [214]. Despite an absence of sequence homologs outside of the poxviridae, atomic structures for P4a-precursor, P4a-1 and P4a-3 could be predicted by AlphaFold2 (Fig. 7.1B, Fig. 7.2) with high statistical confidence (Fig. 7.2). An equally confident structure was predicted for the P4a-1 segment of the presumed cleavage intermediate P4a-1+2 (Fig. 7.1B, Fig. 7.2). No significant structural homologs were identified to any of the above forms of P4a (data not shown) supporting the uniqueness of this protein to the poxviruses.

Conformational hinging: The AlphaFold2 model for P4a-1 showed two subdomains (denoted here as P4a-1 domains I and II; Fig. 7.1B) connected by a hinge that could support a substantial (~180^{o}) rotation of domain II with respect to domain I. This rotation was evident by comparison of models for the P4a-1 segment in precursor vs. processed forms (Fig. 7.1C, Movie 7.1). Supporting the authentic modeling of both conformations, precursor and product models of the P4a-1 segment showed equally strong average pLDDT scores (Fig. 7.2). Models for mature P4a-1 predicted a single conformation exclusively (Fig. 7.3), and those for the N-terminal region of P4a precursor spanning segments P4a-1 and P4a-2 showed almost exclusively a single

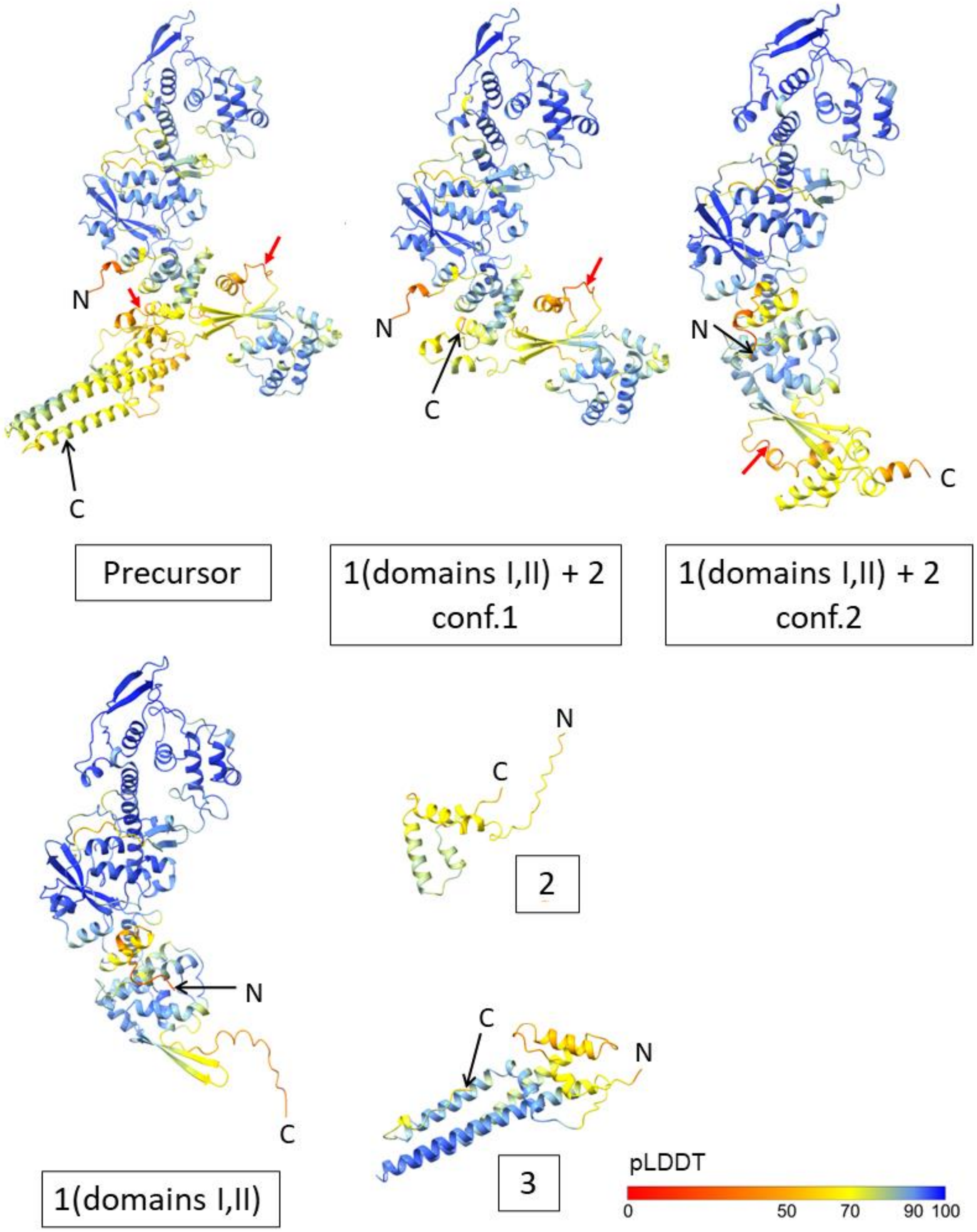


Figure 7.2. Structural models for P4a precursor and processing products and intermediates colored by pLDDT and labeled according to Figure 7.1. Red arrows point to AG| processing sites. Average pLDDT values for P4a-precursor, P4a-1 and P4a-3 were 78.0, 86.5 and 78.2 respectively, with a value of 83.3 for the P4a-1 segment of P4a-precursor. The P4a-1 segment of the cleavage intermediate P4a-1+2 (both forms) had an average pLDDT of 83.5. High confidence extended throughout the structures.

conformation (Fig. 7.3). Models for the P4a-1+2 cleavage intermediate, however, clustered neatly around a pair of distinct conformations which differed at the hinge (Fig. 7.3, Fig. 7.1, Fig. 7.2), strongly suggesting that this intermediate, specifically, was bistable.

The conformation of mature P4a-1 could be verified from experimental intra-protein XLMS data from intact MV: Of the summed XLMS CSM count of 4,009 available for intra-P4a-1 crosslinks, a subset totaling 3,829 (95.5% of the overall total) supported the conformation modeled in mature P4a-1 (Fig. 7.4), while a subset totaling just 2,265 (only 56.5% of the overall total) supported the conformation of P4a-1 in the P4a precursor model (Fig. 7.4). Among the distance-violating crosslinks in the latter model that were structurally rational in mature P4a-1, the majority lay directly across the P4a-1 hinge, strongly supporting the "post-rotation" conformation of P4a-1 in the virion *in vivo*.

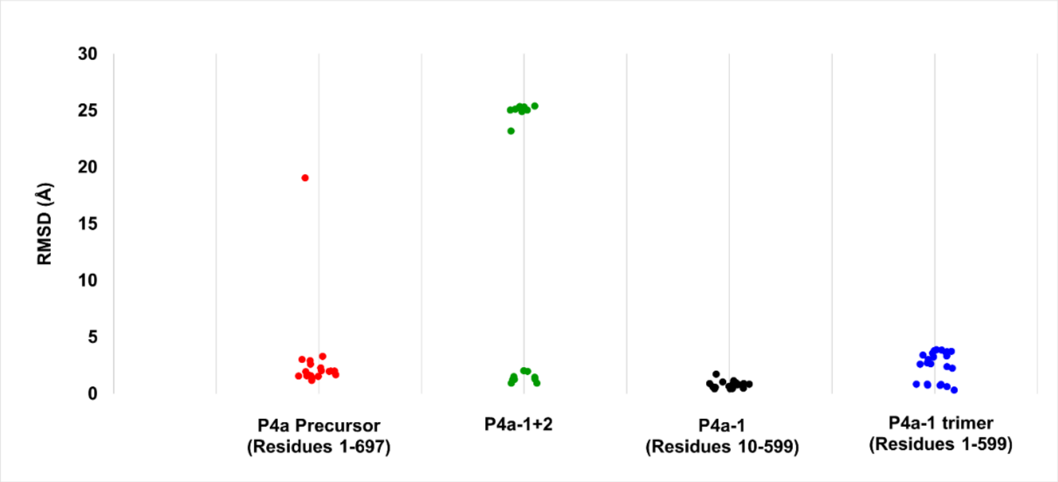


Figure 7.3. Four modeling exercises via AlphaFold-2 or AlphaFold-multimer from each of which the top 20 structural models were considered. Each data point (Y) represents the RMSD (Å) across all α-carbon atom pairs between the top ranked model and one of the remaining 19 models within the same modeling exercise. Models for P4a-1+2 fell into either of two conformers, namely that for precursor (Fig. 7.1B, "1(domains I,II) + 2, conf.1"; lower cluster, < 5 Angstroms RMSD, 10 models) or that for mature P4a-1 (Fig. 7.1B. "1(domains I,II) + 2, conf.2"; upper cluster, > 20 Angstrom RMSD, 10 models). Colors were chosen randomly.

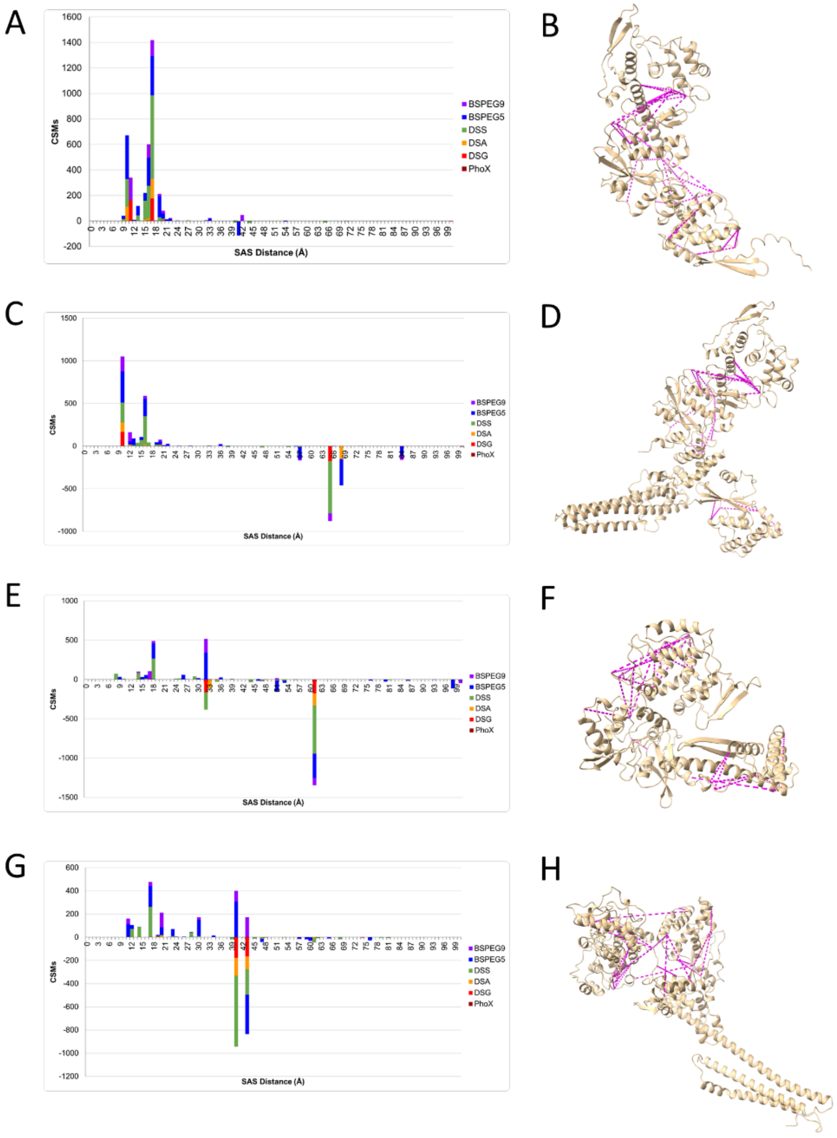


Figure 7.4. P4a-1-localized intramolecular crosslinks in intact MV vs. AlphaFold2 and RosettaFold models of the P4a-1 segment of processed and precursor P4a. (**A, C, E, G**). Solvent-accessible surface (SAS) crosslinking distance vs. CSMs for AlphaFold2 processed P4a-1, AlphaFold2 precursor P4a-1, RosettaFold processed P4a-1, and RosettaFold precursor P4a. Distances are only shown for the P4a-1 segment of precursor P4a models. Colors represent crosslinkers given in the graph legend. Positive and negative values represent, respectively, non-violators and violators for the individual spans of individual crosslinkers. (**B, D, F, G**) Non-violating crosslinks (magenta) mapped on the predicted models. RosettaFold failed to generate a model that could satisfy XLMS data. (**B**) AlphaFold2 processed P4a-1: 95.5% of the total CSM count from intra-protein XL from intact MV were distance non-violators. (**D**) AlphaFold2 precursor P4a-1: Just 56.5% of the total CSM count from intra-protein XL from intact MV were distance non-violators. (**F**) RosettaFold processed P4a-1: 41% of the total CSM count from intra-protein XL from intact MV were distance non-violators. (**H**) RosettaFold precursor P4a: 49% of the total CSM count from intra-protein XL from intact MV were distance non-violators.

P4a-3 removal as the trigger for conformational hinging: In the precursor conformation, P4a-1 domain I and domain II are separated from one another by a wedge-like interjection of the P4a-2 segment into the P4a-1 hinge. This interjection comprises the three C-terminal alpha helices of P4a-2 nestled between two helices of P4a-1 domain I and one helix of P4a-1 domain II, along with a three-stranded beta sheet comprising one strand from the N-terminus of P4a-2 and two strands from P4a-1 domain II (Fig. 7.1B: “Precursor”). This conformation of P4a-2 (and thereby of the overall precursor) seems to be stabilized, externally, by the P4a-3 segment which, while flexible in absolute placement (data not shown), modeled consistently in the vicinity of the hinge in the precursor model (Fig. 7.1B: “Precursor”) and presents a barrier to P4a-2 movement or rearrangement. Cleavage at P4a’s downstream AG| processing site to release the P4a-3 segment removes the external brace, destabilizing the P4a-2 wedge and allowing hinging to the “post-rotational” P4a-1 product conformation (Fig. 7.1B; Movie 7.1).

Disulfide locking: Models were next scanned for cysteine pairs lying within potential intra-protein disulfide bonding range. One such pair was found, namely Cys31 and Cys569. These two cysteines (which are conserved in all known poxvirus P4a sequences) are located within the P4a-1 segment, one each in domains I and II where they directly span the hinge. While they lay far outside disulfide bonding range in P4a precursor (Fig. 7.5) and in the P4a-1+2 intermediate conf.1 (pre-rotation conformation; data not shown), they are within disulfide bonding range in fully processed P4a-1 (Fig. 7.5) and in the P4a-1+2 intermediate conf.2 (post-rotation conformation). After AG| processing and removal of the P4a-3 product, a potential mechanism for locking of the post-rotational conformation is provided by the vaccinia-encoded and packaged oxidoreductases [700-702], whose activity in disulfide bond stabilization is documented [646].

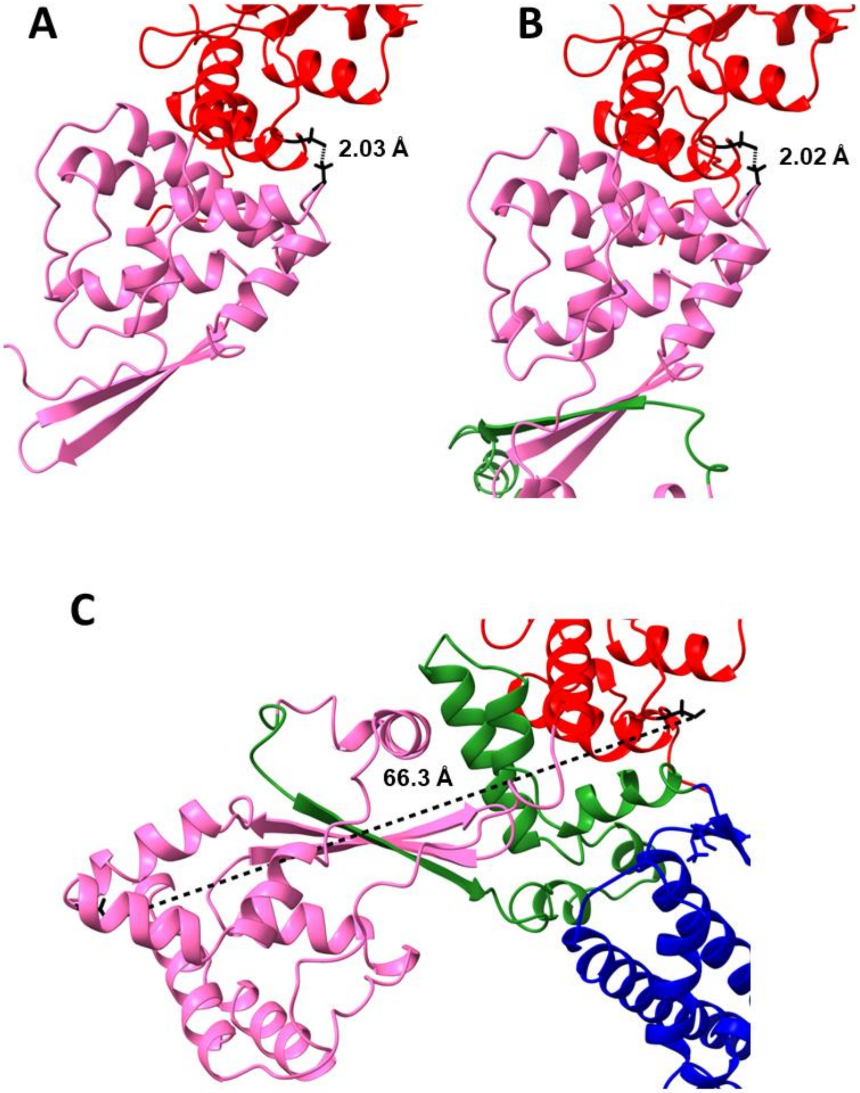


Figure 7.5. Disulfide locking: Distance between the sulfur atoms of Cys31 and Cys569 in P4a models. (A) P4a-1. (B) P4A-1+2 intermediate (conf.2). (C) P4a precursor. Cys31-Cys569 distances are 2.03 Å, 2.02 Å and 66.3 Å, respectively.

P4a-1 trimerization: To investigate the higher order structure of P4a-1 in MV, structural predictions were attempted for P4a-1 dimer, trimer, tetramer, and pentamer using AlphaFold-multimer. Among these, the trimeric model showed a particularly high confidence (Fig. 7.6A, B) and by far the strongest PAE plot (Fig. 7.6C; Fig. 7.7). Moreover, a very high degree of convergence was noted between the 25 trimer models reported by AlphaFold-multimer (Fig. 7.8, Fig. 7.3). The trimer model adopted the approximate shape of a cylindrical "candlestick holder" with "opening petals", with overall dimensions of ~ 6.8 nm diameter (at the waist) x ~11.5 nm height (Fig. 7.6A, B). Around the trimer body the subunits partially enwrap one another at the base (Fig. 7.6D). Views of the trimer from above and below each suggest a hexagonal outline (Fig. 7.6E, F). The upper and lower hexagons superimpose on one another "in phase" as a hexagonal prism shape (Fig. 7.6G), with a ~60 degree twist within each individual subunit.

P4a-3 removal is both necessary and sufficient for P4a-1 trimer formation: All attempts to model a trimer for the full length P4a precursor failed (data not shown), consistent with which the P4a-3 segment presents a fundamental steric block to subunit joining (Fig. 7.1B, data not shown). However, trimer models could be readily obtained for both P4a-1 and the P4a-1+2 intermediate (in both pre- and post-rotation conformations about the P4a-1 hinge; Fig. 7.9). We conclude that removal of the P4a-3 segment is both necessary and sufficient for P4a-1 trimer formation.

Trimer molecular model is consistent with cryoEM imaging: Since P4a-1 is considered to lie on the outer surface of the virion core wall [73], [73] we scrutinized published reports for the dimensions of core wall ultrastructural features that might correlate with P4a-1 trimers. Early electron microscopy (EM) and cryoEM studies of virion cores

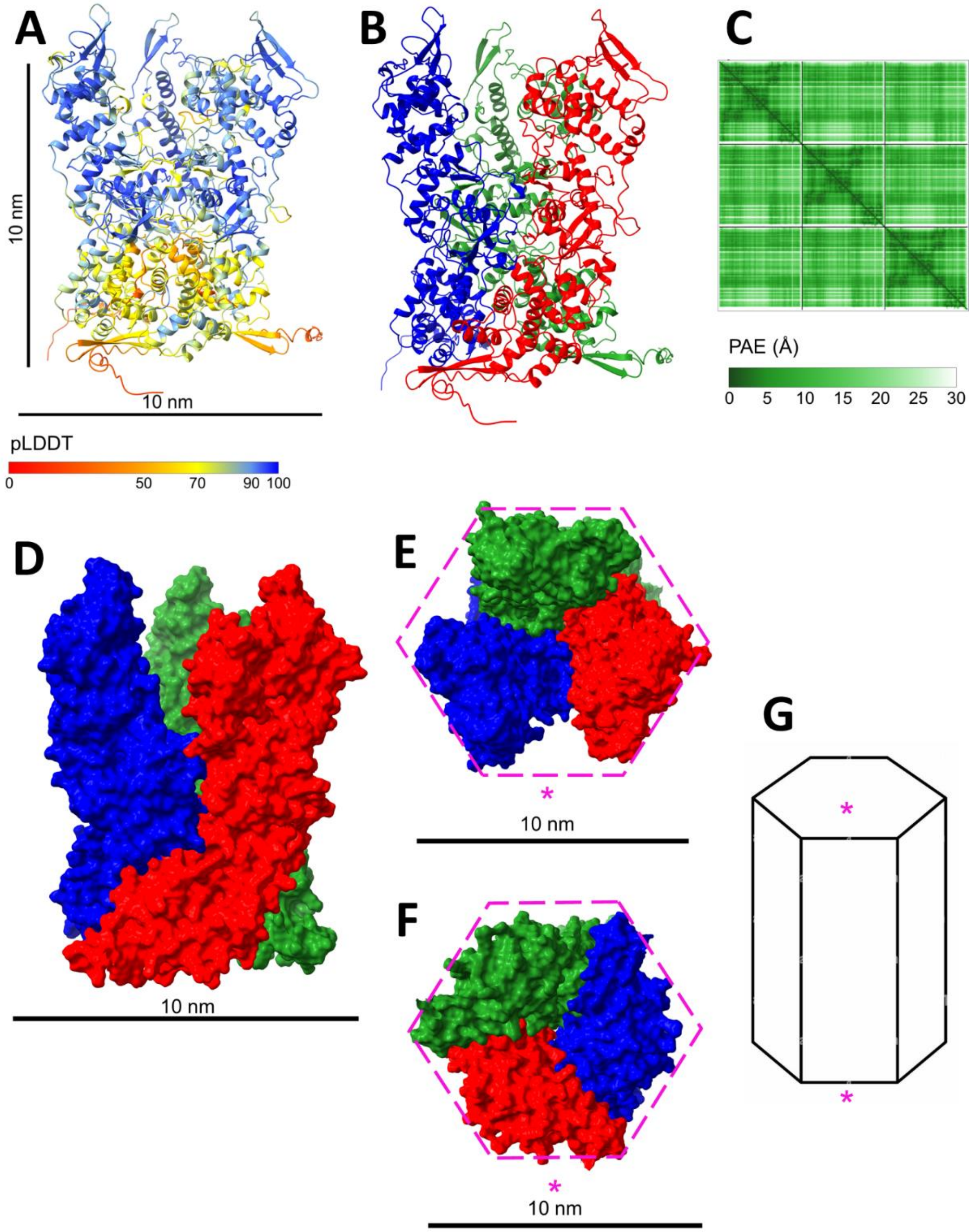


Figure 7.6. **P4a-1 trimer.** (**A**) Colored pLDDT (scale bars: 10 nm), (**B**) Colored by P4a-1 chain within the trimer. (**C**) PAE plot for trimer. Color scale is in Angstrom units. (**D**), (**E**), (**F**) Trimer in spacefill from side, above and below respectively (colored as in panel B) showing the mutual enwrapping of subunits at the base, and an outline approximating to hexagonal geometry (residues 600 - 614 excluded from view). (**G**) The trimer model can be fit within a hexagonal prism geometry (upper and lower hexagons are in phase), with a ~60 degree twist along the body of each subunit such that the starred side, viewed from above (E) corresponds to the starred side viewed from below (F).

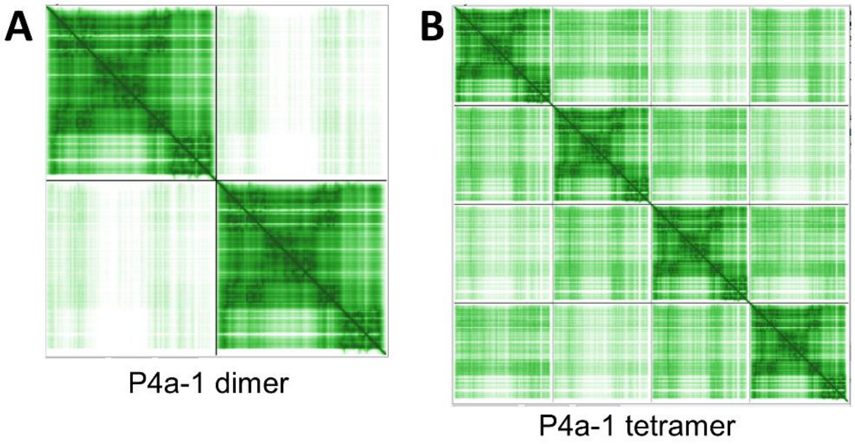


Figure 7.7. Dimer and tetramer predictions of P4a-1 by AlphaFold-multimer: PAE plots. (A-B) P4a-1 dimer and tetramer models respectively. A sharp green/white delineation at subunit boundaries indicates low confidence in subunit condensation and placement. A smooth green/green boundary indicates high confidence. Among the dimer, trimer (Fig. 7.6C), and tetramer models, only the trimer model showed high confidence, comparable to the confidence of the monomer fold.

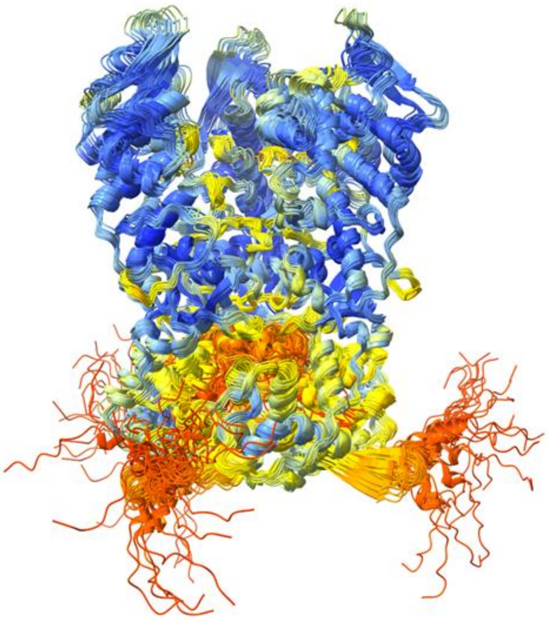


Figure 7.8. Superposition of all 25 AlphaFold-multimer models for P4a-1 trimer, colored by pLDDT. Models showed a high degree of convergence upon a single optimal solution (with exception of the 15 aa flexible C-terminal tail). No inter-subunit disulfide bonds were apparent in the trimer models. RMSD for the top 20 of these models (minus the C-terminal tail) is shown in Fig. 7.3.

reported a “palisade” layer of spikes or “pegs” on the outer core wall surface [84, 87, 89, 90]. Measurements of the spike dimensions varied in these early reports, perhaps due to limitations in instrument capabilities, with spike lengths ranging from 100-200 Å [87] and diameters of 50-100 Å [92]. A very recent cryoEM study (released during the final preparation of this manuscript) of vaccinia IVs and MVs by Hernandez-Gonzalez *et al*. [94] measured the palisade layer thickness at 12.5 nm. These measurements agreed closely with dimensions of our P4a-1 trimer model (Fig. 7.6). Furthermore, the P4a-1 trimer model’s size and shape were comparable with the numerous scattered features visible in Fig. 8 of ref. [90] that could now be identified as probable

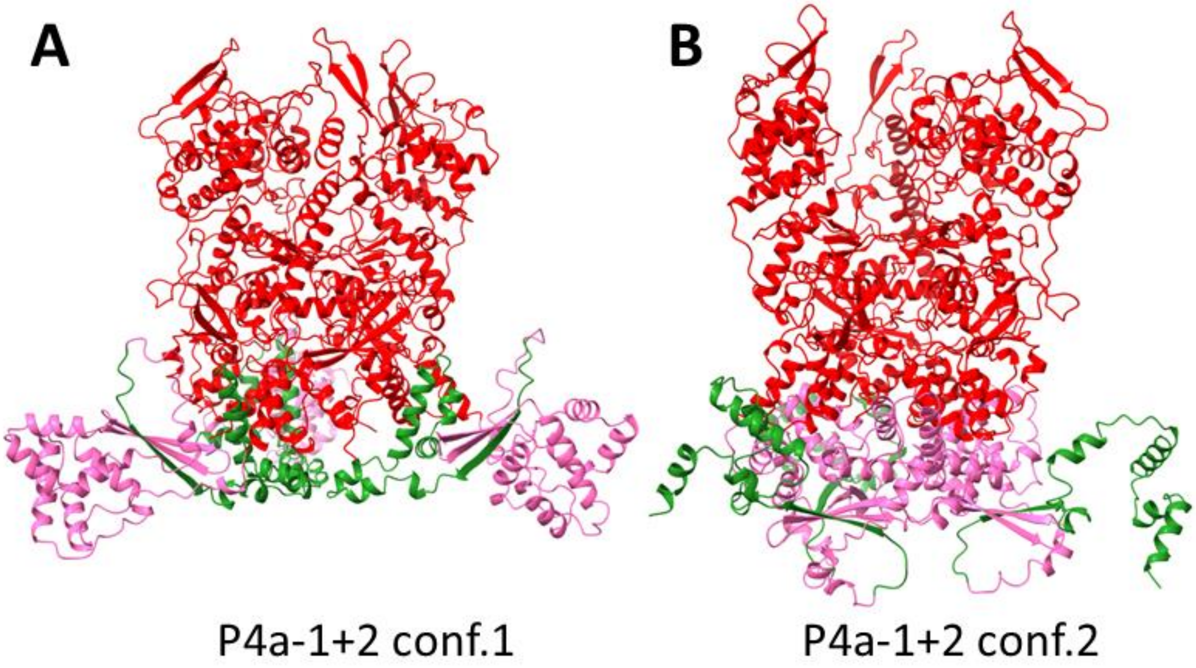


Figure 7.9. Trimer models for the P4a-1+2 intermediate predicted by AlphaFold-multimer fell in two conformations about the hinge joining P4a-1 domains I and II. (A), (B) The two conformations: Pre- and post-rotation about the hinge, respectively. Domain coloration follows Fig. 7.1 (Red: P4a-1 domain I (aa 1 - 451); Pink: P4a-1 domain II (aa 452-614); Green: P4a-2 (aa 615 - 697)).

dissociated P4a-1 trimers (Fig. 7.10). The above data were consistent with P4a-1 trimer being the primary distinguishable component of the "palisade" spikes.

Higher order P4a-1 trimer structure and assembly: We next investigated the higher order organization of P4a-1 homotrimers, using XLMS data for guidance: Identical peptides crosslinked to one another ("homomultimer XL") can have only arisen if the crosslink bridges homomultimer subunits. Our homomultimer XL dataset for P4a-1 from intact MV, showed a summed CSM count of 779. While the discrete homotrimer model (above, Fig. 7.6) satisfied95.5% of the CSM count for all intra-P4a-1 crosslinks (Fig. 7.11), it could satisfy only 44.7% of the 779 homomultimer CSMs.

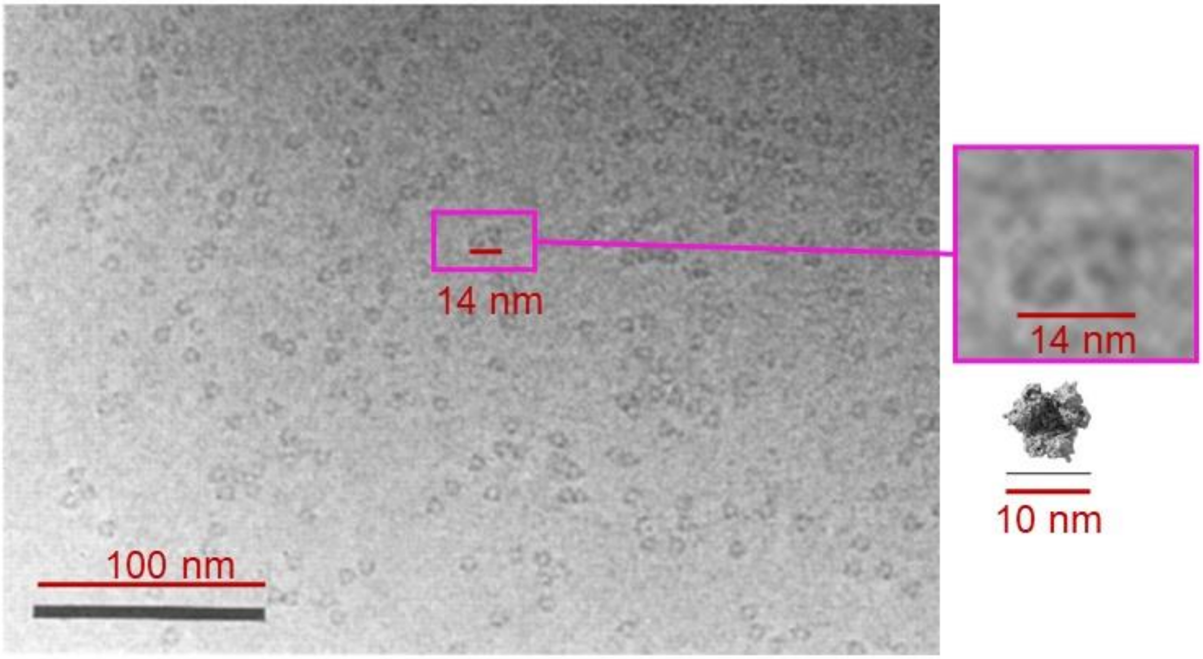


Figure 7.10. Comparison of the P4a-1 AlphaFold-multimer molecular model with prior cryoEM images. (A) Left: A series of identical features scattered on cryoEM grid upon virion uncoating from Dubochet (left) (38), enlarged. Upper right: Zoom of a pair of features with proportionately enlarged scale bar. Lower right: Single trimer model (AlphaFold-multimer) to same scale.

The remaining 55.3% represented, apparently, undefined higher-order assemblies of P4a-1. Most striking among these was crosslink 366-366, which accounted for nearly half of all P4a-1 homomultimer CSMs, and was detected strongly for all crosslinker types in our dataset including those with the shortest crosslinking distances (e.g. PhoX; just 23 Å). Other higher order homomultimer XL detected in intact MV included 508-508 and 557-557 (CSM count = 66 and 6, respectively).

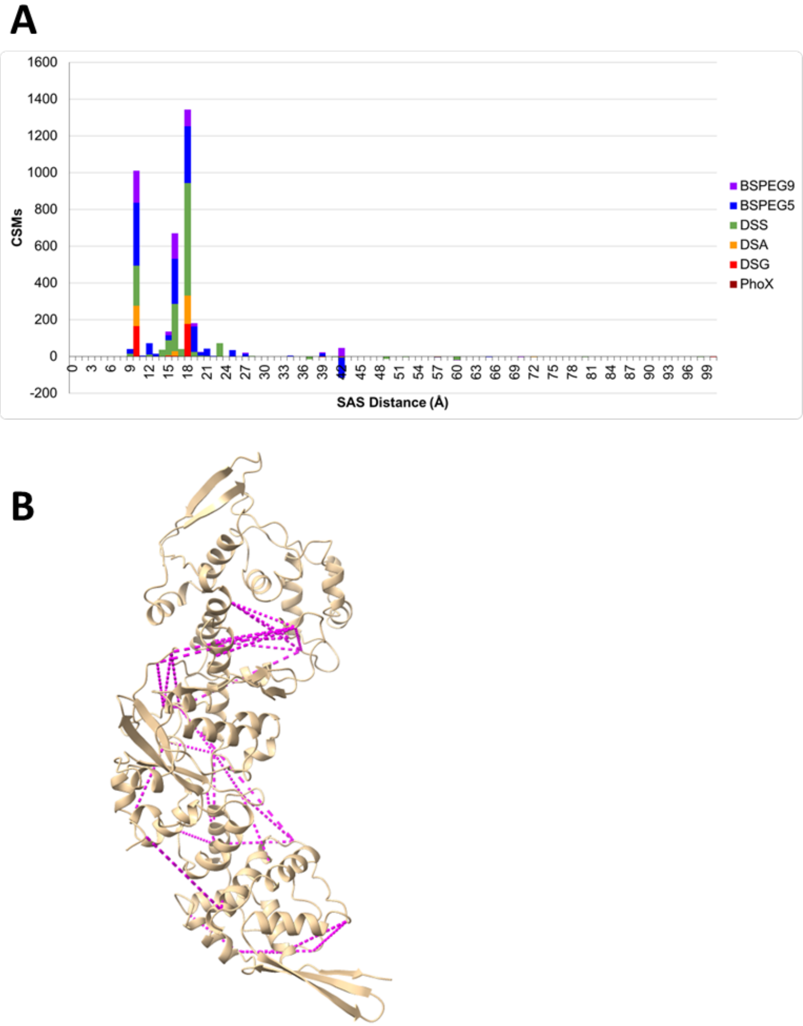


Figure 7.11. P4a-1-localized intramolecular crosslinks in MV vs. one subunit from the AlphaFold-multimer model of P4a-1 trimer. Details as in Fig. 7.4.

The trimer's hexagonal outline (above; Fig. 7.6E-G) suggested a higher order organization comprising some form of hexagonal array. Arranging individual trimers (Fig. 7.12A) within the simplest conceivable hexagonal array yielded three potential rotationally symmetrical trimer-hexamer models (Fig. 7.12B). Of these, Fig. 7.12B(iii) could not satisfy the 366-366 homomultimer crosslink at all. Figure 7.12B(ii) could satisfy neither 366-366 with crosslinkers PhoX or DSG, nor 508-508 with crosslinkers PhoX, DSA, DSS or BSPEG5, presenting a fatal limitation for this model. The model of Fig. 7.12B(i), however, could satisfy 366-366 for all crosslinker types (with a crosslinking distance of just 21.6 Å, Fig. 7.13) as well as 508-508 and all other higher order homomultimer XL. This model (Fig. 7.12B(i)) accounted for 97.2 % of all homomultimer XL observed for P4a-1 from MV (either intact or de-enveloped) and was selected as a basis for higher order modeling (Fig. 7.12C). Higher order models comprised a trimer-hexamer whose central hole was filled by a 7th trimer (trimer-heptamer; Fig. 7.12C(i)) or a tessellation of either abutting or fused trimer-hexamers (Fig. 7.12C(ii), (iii) respectively). Albeit the trimer-heptamer model could not be fully saturated with 366-366 homomultimer XL (Fig. 7.12C(i)), this by itself did not render the model invalid. The trimer-heptamer model did, however, lead to an entirely "closed" core wall rather than the more open lattice with trimer-hexamers (Fig. 7.12C(ii), (iii)), with the former reminiscent of a rigid capsid rather than the flexible, collapsible sac that is more characteristic of the vaccinia core [79, 80].

Between the two trimer-hexamer interface models (abutted or fused, Fig. 7.12C(ii), (iii) respectively), only the fully fused model (Fig. 7.12C(iii)), when extended to a larger lattice (Fig. 7.12D), could fully saturate the lattice with "holes" (Fig. 7.12D) to combine uniform strength in all directions with maximum flexibility. Also unique to this model: Every trimer fell at the vertex of three fused trimer-hexamers, every trimer-trimer interface separated two holes in the lattice,

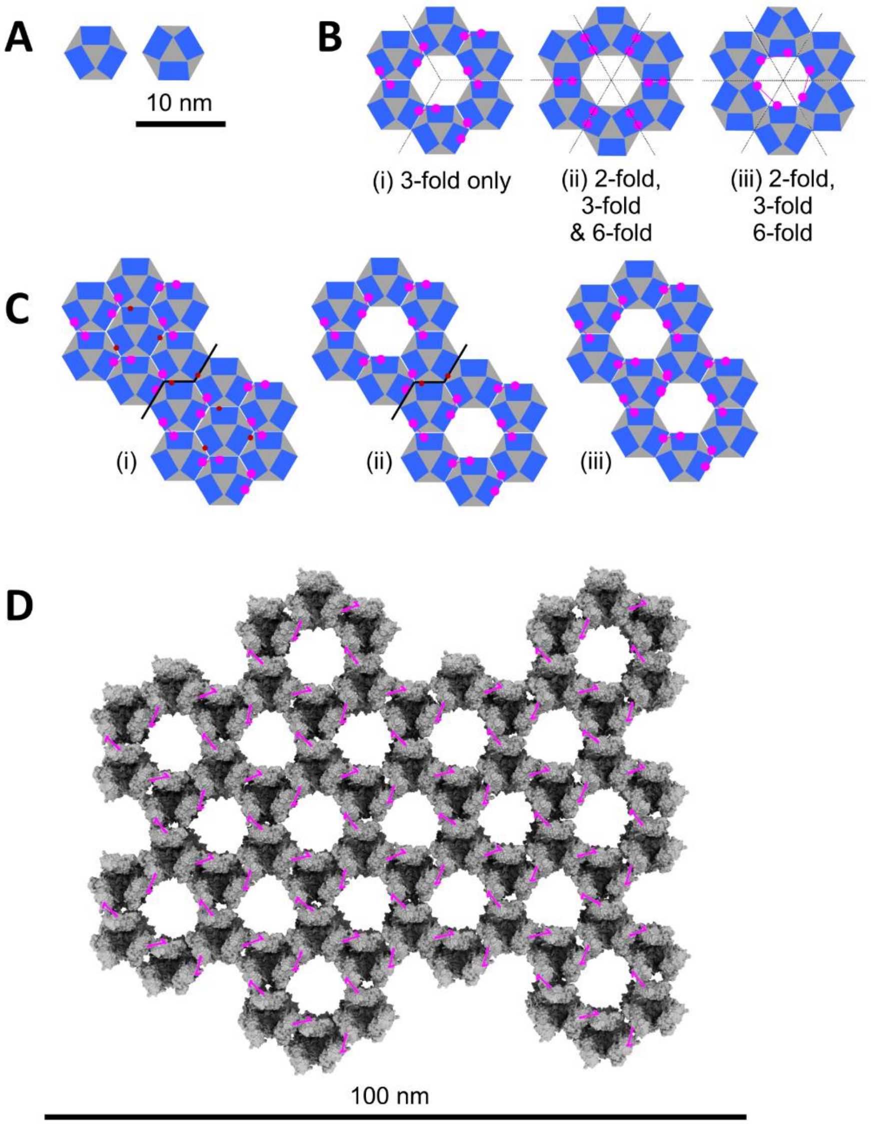


Figure 7.12. Hexamer-of-trimers models: Genesis, in schematic form. (A) The two rotamers of an individual trimer (60 degree difference). (B) Rotationally symmetrical hexamer-of-trimers models based on the trimer's hexagonal outline (symmetries are indicated). In (i), all trimers are in the same rotational orientation while in (ii) and (iii) the two forms in panel (A) alternate. Internal trimer-trimer 366-366 crosslinks are indicated in Magenta (the outward-facing six subunits are unsaturated providing potential hooks to neighboring trimer-hexamers). Crosslink distances favor model (i) over models (ii) and (iii). (C) Trimer-heptamer (i) and trimer-hexamer interface schemes (ii, iii) based on the model in panel B(i). Schemes (i) and (ii) do not saturate all interfacial 366-366 XL (unsatisfied crosslinkable sites shown brown), while (iii) does. (D) Lattice based on the model in panel C(iii). 366-366 crosslinks, which reside below the upper surface of the protein when viewed from above (Fig. 7.14), have been projected over the surface to emphasize their locations. The Euclidian plane tessellation of panel (D) has regular p6m symmetry (1-uniform) with periodic (h,k) = (3,0). The low confidence C-terminal tails (residues 600 – 614) were removed for clarity.

and three sides of every trimer faced a hole. Moreover, the highest order trimer cluster was simply a trimer-dimer: In contrast to C(i) and C(ii) or mixed forms (not shown), there were no higher order clusters. Finally, model C(iii) was unique in providing a 366-366 crosslink for every P4a-1 subunit, uniformly saturating the lattice with this XL (Fig. 7.12D). Model C(iii) was also fully saturated with 508-508 and all other higher order homomultimer XL.

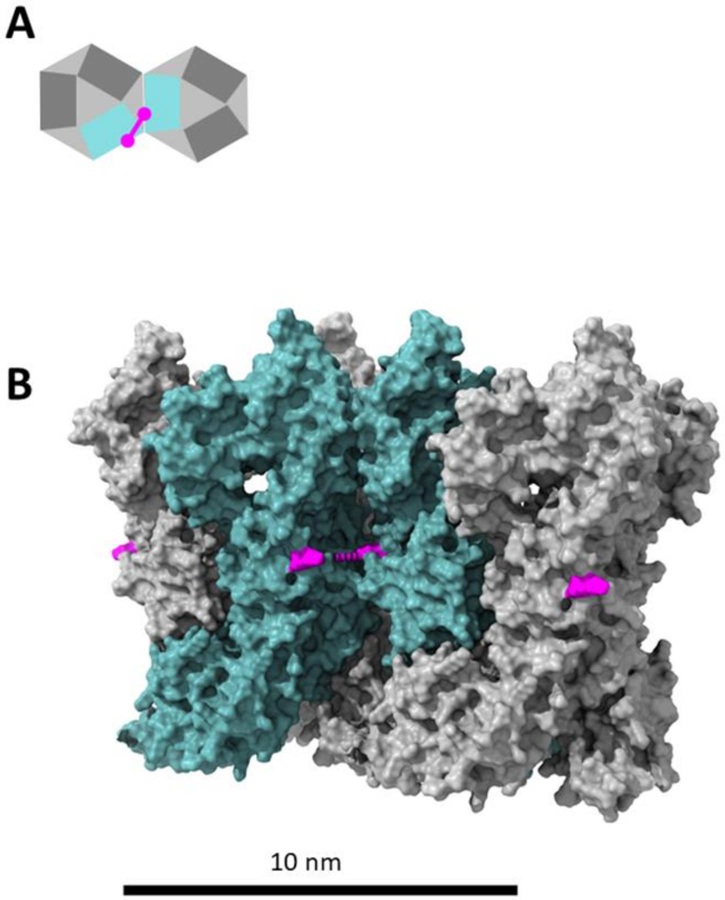


Figure 7.13. P4a-1 trimer-dimers. (A) Two P4a-1 trimers (viewed from above in schematic form) brought into a side-by-side alignment that can rationalize the 366-366 homomultimer crosslink; crosslinked subunits shown in blue. (B) Crosslinked dimer-of-trimers (side view of panel (A), with residues 600-614 excluded from view) – Blue represents crosslinked subunits. Residues 366 from both trimers colored magenta; the pair within crosslinking range is connected as a crosslink (also magenta). Solvent accessible surface distance is 21.6 Å for this pair (median 24.6 Å).

For all the above reasons, C(iii) was the favored building block for a full core wall lattice (Fig. 7.12D).

The higher order lattice model is consistent with cryoEM imaging: We next attempted to reconcile the trimer-hexamer molecular model with cryoEM images of the intact virion core wall. Dubochet [90], Cyrklaff [93], and Hernandez-Gonzalez [94], reported patches of hexagonal features on the surface of the core wall, identified as the "palisade" coated with pegs [93], all of which fit and scaled to our P4a-1 lattice (Fig. 7.14). Pegs visible in relief around the edges of the virion core in cryoEM images seemed to have a periodicity of ~ 8.5 nm, comparable to that of both the hexagonally-arranged features in the same cryoEM images [90, 93] and individual P4a-1 trimers in our lattice molecular model (Fig. 7.14). At a specific imaging orientation (Fig. 7.12D), trimer electron density may be expected to sum through sequential rows of trimers to yield the "palisade" effect.

Published cryoEM images show hexagonal features only discontinuously across the core [90, 93]. This could arise from overlaying features (such as bent or distorted pegs), the technicalities of cryosectioning, or because the P4a-1 lattice is discontinuous. To establish whether the lattice can be reasonably considered to cover the entire surface of the core, we derived the anticipated experimental mass of P4a-1 in the virion based on known virion mass, the proportion of total mass comprising protein, and proportion of total virion protein comprising P4a-1 (described in the Materials and Methods). This calculation indicated that 12.4% of the overall virion mass, or 404 MDa, comprised P4a-1. In an independent calculation for the theoretical mass of P4a-1 in the virion if it were enclosed entirely by our lattice, we calculated the surface area of a virion core (taking the mean core size from a number of published images)

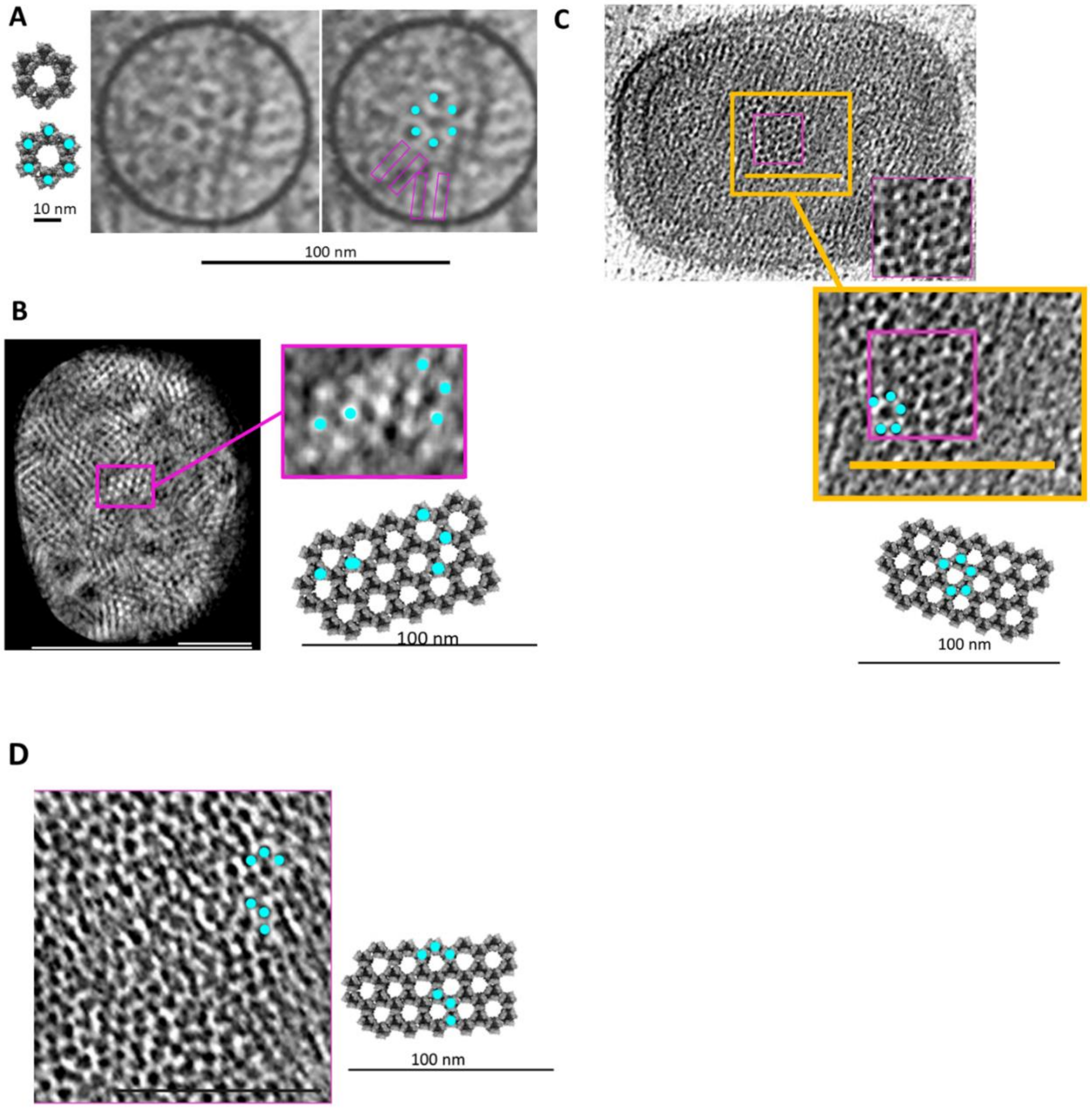


Figure 7.14. CryoEM hexagonal features scale with P4a-1 trimer-hexamer. (A) A hexagonal feature on core wall in Fig 7 of ref. (38) is comparable in scale to a trimer-hexamer modeled here. Right panel same as left but with pegs outlined magenta. The circular outline was on the original image. The EM images show the same periodicity as cyan puncta of hexagonal features. (B) Hexagonal features in Fig 5c of ref. (41), comparable in scale to a trimer-hexamer modeled here. Cyan dots are arbitrarily placed in order to register EM images with molecular model lattice. Based on the original figure legend, the colors in this figure are inverted, so white shows regions of higher electron density. (C and D) Hexagonal features in Fig 5c and 6b respectively of ref. (40), are comparable in scale to the trimer-hexamer modeled here. Scale bars are 100 nm. In panel C, the scale bar was measured and transferred from a different region in Fig 5c (40) is shown in orange. Cyan dots were arbitrarily placed to highlight similarities in spacing of hexagonal features.

and the total number and then mass of P4a-1 molecules that could coat it fully according to the lattice of Fig. 7.12D. This totaled 6,064 molecules or 430 MDa of P4a (see Materials and Methods). From the similarity of the two values (404 and 430 mDa), we conclude that the modeled P4a-1 lattice likely encloses the entire surface of the virion core, albeit with, perhaps, localized regions of disruption. In this regard, for example, "pore-like” features on the surface of the core wall [73, 93, 94] may contribute to the disruption of the P4a-1 lattice.

Correspondence of P4a trimer lattice to the external scaffold: Vaccinia morphogenesis initiates with the insertion of virion envelope proteins A17 and A14 into the ER membrane. This membrane is subsequently fractured, allowing A17 to associate with trimers of vaccinia external scaffold protein D13, resulting in spherical IVs coated with the D13 external scaffold [29, 189]. Deep-etch EM has shown that the external scaffold forms a hexagon-like honeycomb lattice across the entire surface of IV, with pentameric and heptameric defects [189]. The known experimental structure of the external scaffold also shows D13 trimers forming a hexagonal trimer-hexamer lattice [703]. During maturation from IV to MV, the A17 N-terminus is cleaved by vaccinia protease I7, leading to D13 scaffold release. Either simultaneously or soon thereafter, the A17 C-terminus and core structural proteins (including P4a) are processed, resulting in condensation of the core wall and palisade layer [201, 202, 695-699]. Hexagonal features of the palisade layer, visualized by cryoEM *(40, 41)* are reminiscent of the honeycomb D13 lattice. Our P4a-1 trimer-hexamer lattice model was, dimensionally, entirely coincident with the atomic structure of the D13 lattice (Fig. 7.15). In some manner, the external scaffold, through its interaction with A17, may template condensation of the P4a-1 lattice during virion

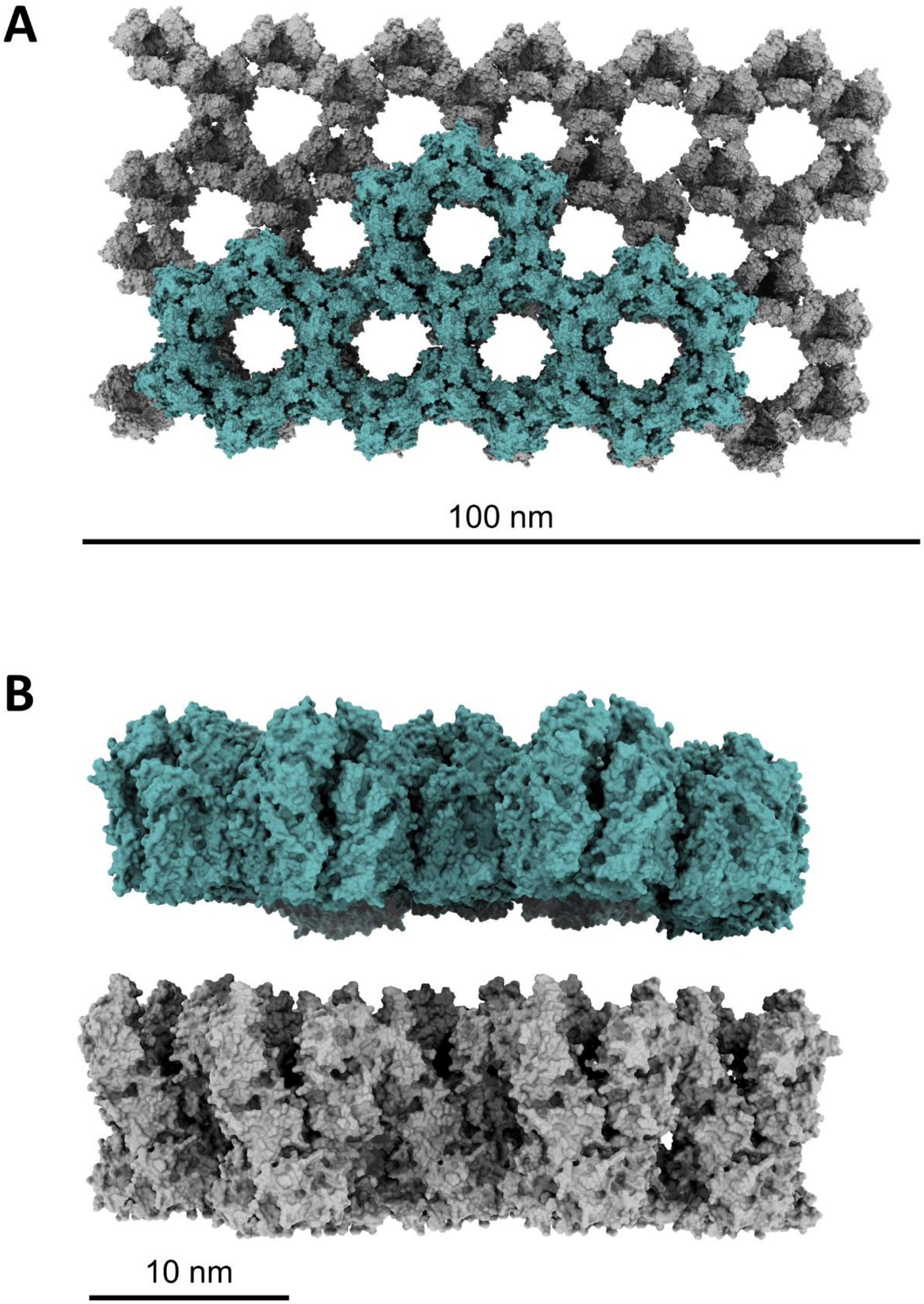


Figure 7.15. **Modeled lattice of P4a-1 is co-dimensional with the D13 external scaffold** [703]. Gray: P4a-1. Turquoise: D13. View from above (**A**) and side in higher zoom (**B**). In panel (B), P4a-1 and D13 lattices are separated by an arbitrary distance of 5 nm.

morphogenesis, for subsequent core wall assembly.

Discussion

There are currently atomic level structures for all or parts of 34 of the ~75 packaged vaccinia gene products. However, with perhaps one exception, namely the packaged multisubunit vaccinia RNA polymerase [71], how these proteins condense into higher order assemblies has eluded analysis. Here we have investigated vaccinia major structural protein P4a, which accounts for 12.4% of the total virion mass. P4a precursor is transported to IVs during assembly where it is proteolytically processed at two AG| sites [214]. P4a-1 and P4a-3 are packaged in MVs while P4a-2 is discarded [214]. The biological and functional significance of the two processing events has not been previously elucidated. Moreover, how P4a-1 forms the core wall and/or palisade layer associated with the core wall has remained unresolved. Here, we have combined deep-learning protein structure prediction with a deep XLMS dataset comprising ~135,000 crosslink spectral matches via 10 distinct chemical crosslinkers to yield ~22,000 unique-mass crosslinked peptide pairs. The predicted structure for P4a-1 monomer was consistent with all intra-protein crosslinks (Fig. 7.1, Fig. 7.4). Homomultimer models were optimal at the trimer level (Fig. 7.6, Fig. 7.7, Fig. 7.13) and the resulting trimer model was consistent with published low resolution cryoEM images (Fig. 7.10) but did not accommodate all crosslinks of the inter-subunit type. Building a hexamer of trimers of a specific topology, however, allowed essentially all inter-subunit crosslinks to be satisfied and they remained satisfied upon building the trimer-hexamer to a full core wall lattice (Fig. 7.12, Fig. 7.13, Fig. 7.14).

We suggest a 5-step ordered/concerted assembly pathway for P4a-1 (Fig. 7.16, Movie 7.2). In the initial P4a precursor-monomer conformation (prior to step 1), the P4a-3 segment presents a steric block to trimerization and the P4a-1 trimerization interface is conformationally split. P4a-3 detachment by AG| processing initiates the pathway (step 1), relieving the block to trimerization, while also destabilizing the precursor conformation of the P4a-1 segment leading to the massive conformational rotation of P4a-1 domain II (Fig. 7.16, steps 2 and 3). We cannot distinguish the temporal order of trimerization and conformational inversion without evidence that either is dependent on the other: Models for both P4a-1+2 monomer and trimer showed both pre- and post-rotation conformations (Fig. 7.16, step2-lower or step3-upper, Fig. 7.9, Fig. 7.3), and both conformations were able to trimerize (Fig. 7.16, step2-upper or step3-lower, Fig. 7.9).

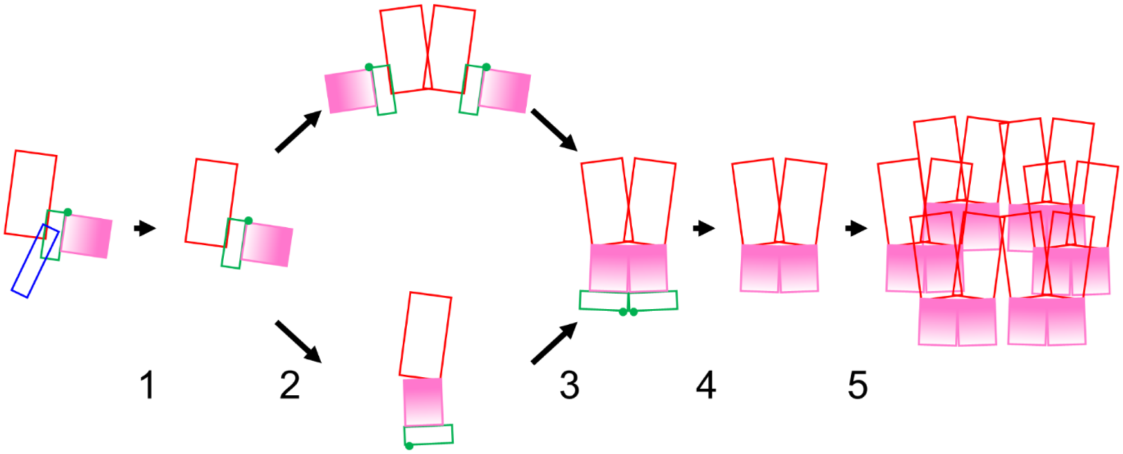


Figure 7.16. Suggested five-step pathway of P4a processing, conformational change and assembly. Coloration (as in Fig. 7.1A): P4a-1 domain I red, P4a-1 domain II pink, P4a-2 green, P4a-3 blue. P4a-1 domain II is shown with gradient fill to emphasize its conformational inversion at steps 2/3. Green spot: P4a-1|2 processing site. In step 1, P4a-3 removal permits both conformational inversion and trimerization of the P4a-1+2 intermediate (steps 2 and 3). The order of steps 2 and 3 is undefined hence the upper and lower loops of the pathway. In step 4, P4a-1|2 processing permits higher order P4a trimer assembly (step 5). For simplicity, only two of the three trimer subunits are shown in multimer schematics. For details see text.

Perhaps the two events occur simultaneously. After conformational inversion, P4a-1 domains I and II pack together (Fig. 7.1B) and become conformationally locked by disulfide bond formation between Cys31 and Cys569 (Fig. 7.5). In addition, P4a-1 domain I and II helices now pack together allowing new interactions between them (Fig. 7.1B). The P4a-2 segment is now isolated at the end of the molecule (Fig. 7.16, steps 2/3). Since a trimer could be modeled for not only fully processed P4a-1 but also the P4a-1+2 intermediate (Fig. 7.16, before/after step 4) albeit with a lower confidence PAE, excision of the P4a-2 segment is not an absolute prerequisite for P4a-1 trimerization. P4a-2 excision is, however, a prerequisite for higher order trimer assembly: After conformational inversion in the P4a1+2 trimer (Fig. 7.16, step3-upper) removes one steric block to higher order assembly (Fig. 7.16, step 3), a steric block remains after step 3 due to presence of the P4a-2 segment on the outside surfaces of P4a-1+2 trimer subunits coincident with sites of higher order assembly. In step 4 (Fig. 7.16) release of the P4a-2 segment via processing at the 1|2 AG| site allows higher order assembly into a trimer-hexamer lattice (Fig. 7.16, step 5). The overall pathway (Fig. 7.16) is animated in Movie 7.2.

In studies with wild-type MV in the infected cell [214], anti-P4a-2 antibody could detect the P4a-1+2 intermediate but not P4a-2+3, suggesting that processing to remove P4a-3 precedes processing to split P4a-2 from P4a-1. Our pathway reflects this and now can be rationalized as the deferral of higher order assembly until after trimerization has occurred. A model in which this order of processing events is regulated sterically (steric occlusion of the 1|2 processing site until after 2|3 processing and the resulting conformational inversion) is not viable since both P4a processing sites appear to be solvent exposed in all conformers and intermediates (Fig. 7.1). We therefore suggest an alternative model in which the three 1|2 processing sites of a trimer must be brought into spatial proximity (at a point in space beneath the trimer’s center; Fig. 7.16, step 3) for 1|2 processing to occur. This would be the case if, for example, the processing protease is spatially restricted instead of freely diffusible, and would then defer P4a-2 removal (the last block to higher order trimer assembly) to the final portion of the pathway.

Movie 7.1. Proposed conformation change in P4a during AG| processing to the three known products.
Download link: https://journals.asm.org/doi/suppl/10.1128/mbio.01135-23/suppl_file/mbio.01135-23-s0003.mp4

Movie 7.2. Proposed overall pathway of P4a processing, conformation change, and assembly.
Download link: https://journals.asm.org/doi/suppl/10.1128/mbio.01135-23/suppl_file/mbio.01135-23-s0004.mp4

Materials and Methods

Protein structure prediction, *de novo*: Monomer and multimer structure predictions for vaccinia protein P4a (precursor and processing products) were run on local installations of AlphaFold2 and AlphaFold-multimer, using a non-docker setup (https://github.com/kalininalab/alphafold_non_docker). Models shown here represent the top-ranked prediction for each protein or protein complex. Side chains were present in models but not rendered except where such rendering was chosen. For generating 20 models from AF2 (e.g. Fig. 7.4) where it would otherwise produce only produce 5 models AlphaFold2 was run four times and the resulting models were pooled, or AlphaFold-multimer was run for the target protein. For each proteoform (P4a precursor, P4a-1+2 intermediate, mature P4a-1 and P4a-1 trimer), the top ranked structure was compared individually against the other 19 models using

ChimeraX MatchMaker. The RMSD (Å) between all α-carbon atom pairs, per comparison, was reported as a data point, with 19 total data points for each model set.

Structure predictions through RoseTTAFold were run on the Robetta server (https://robetta.bakerlab.org/submit.php) using standard parameters.

Scoring of structural models: Average pLDDT scores for structures predicted with AlphaFold2 were obtained by averaging pLDDT values for all α-carbon atoms (representing all atoms) from the top predicted model for each structure (in many cases some top models were either identically or almost identically scored).

Visualization: Reported structures were visualized and measured, and images generated using UCSF ChimeraX [688, 689]. PAE plots of AlphaFold-multimer structures were visualized by the "AlphaFold Error Plot" tool provided by ChimeraX.

XLMS: Virion proteins were crosslinked, prepared for XL-MS, and analyzed as described in greater detail in ref. [604]. Briefly, bis(succinimidyl) penta(ethylene glycol) (BSPEG5), bis(succinimidyl) nona(ethylene glycol) (BSPEG9), adipic acid dihydrazide (ADH), 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide hydrochloride (EDC), and n-hydroxysuccinimide (NHS) were purchased from ThermoFisher, Inc. Isotopically-coded disuccinimidyl suberate (DSS), bis(sulfosuccinimidyl)suberate (BS3), disuccinimidyl glutarate (DSG), and disuccinimidyl adipate (DSA) were purchased from Creative Molecules. 3,5-bis(((2,5-dioxopyrrolidin-1-yl)oxy) carbonyl)phenyl)phosphonic acid (PhoX) was provided by the Scheltema lab and later purchased from Bruker and prepared as described in (*45*). Vaccinia virus was grown in HeLa S3 cells (ATCC CCL-2.2) and purified over a 36% sucrose cushion, followed by two 24-40% sucrose gradients. The harvested virus was washed with 100 mM triethylammonium bicarbonate buffer

(TEAB) pH 8.5 and crosslinked as intact virions or after brief uncoating treatment (0.05% NP40, 40 mM TCEP, 100 mM TEAB, pH 8.5). Appropriate XL concentrations were determined by SDS-PAGE. XL reactions were quenched with ammonium bicarbonate or removed by spin desalting into ammonium bicarbonate.

Crosslinked virus samples were disaggregated by various methods, including an in-lab modified FASP protocol and urea denaturation. Cleavage of solubilized proteins into peptides employed various reagents/reagent combinations: Trypsin, ArgC, GluC, AspN, LysC, Trypsin+LysC, Trypsin+GluC, Trypsin+AspN, ArgC+AspN, ArgC+GluC, AspN+GluC, or CNBr+Trypsin. All cleaved samples were acidified with formic acid (FA) to 2% FA final concentration and desalted by C18/SCX as described [578]. Peptides were eluted with 5% NH4OH, 80% CH2CN, 0.1% FA (Buffer X) or with a six-step ammonium acetate gradient in 20% CH3CH, 0.5% FA followed by a final elution with Buffer X. Samples were dried under vacuum and reconstitute in 0.1% FA in water for MS. The nanoLC-MS/MS method used is described in (6).

Instrument raw files were converted to mgf or mzML using MSConvert by ProteoWizard. Crosslinks were identified by the programs described in [604] with the addition of pLink2, Kojak2, MetaMorpheus and XlinkX. The in-house code described in [604], with substantial upgrades including for large datasets, was used to consolidate and assess XLMS data and calculate CSM (crosslink-spectral match) counts. The latter represent total CSMs (crosslink-spectral matches) for a particular crosslink or set of crosslinks.

Crosslink networks were rendered using CrosslinkViewer [589]. Solvent accessible surface distances between crosslinked residues were calculated using Topolink [521]. Models of crosslinked pairs with through-space (Euclidean) distances indicated, were rendered using ChimeraX.

Feature size measurement (EM): Spike lengths and diameters were measured using the "tape" feature of ChimeraX: Images from publications by these groups were saved as TIFF files, opened in ChimeraX, and spike lengths and diameters were measured with the ChimeraX Tape feature. The provided scale bars were used to convert the ChimeraX measurement values to Å.

Calculation of the abundance and core wall surface coverage of P4a-1

Core surface area
Dumbell core as rectangle: 245 x 62 x 135 nm (mean values of measurements from several publications)
Two faces 245 x 62 = 30,380
Two faces 62 x 135 = 16,740
Two faces 245 x 135 = 66,150
Total: 113,270 nm^2 = ~0.113 mm^2

Virion molecular composition
Assuming 5 x 10^6 lipid molecules per mm^2 of typical biological membrane:
https://www.ncbi.nlm.nih.gov/books/NBK26871/

Vaccinia envelope has 0.113 x 5 x 10^6
= **5.65 x 10^5 lipid molecules**

Lipid composition (plasma membrane):
https://www.britannica.com/science/lipid/Lipids-in-biological-membranes

Average lipid Mr based on plasma membrane composition:

	% composition	Mr	mole%
Phosphatidyl choline	31%	314.25	97.34
Cholesterol	28%	386.7	108.28
Sphingomyelin	16%	493.6	79
Phosphatidylethanolamine	14.3%	299.21	42.76
TOTAL:	89%		327.38

327.38 / 0.89 (since above accounts for 89% of total **lipid): 367.8 (average Mr)**

5.65 x 10^5 x 367.8 = 207.8 MDa of lipid in virion

Genome MW = 190,000 x 333 = 126 mDa DNA in virion.

Virion dry protein mass calculation

Virion dry molecular mass = 3.26 GDa

Molecular mass of protein = virion dry mass – (lipid + DNA) = 3,260 – (207.8 + 126) Mda = 2,926 MDa protein
13.8% (by mass) of total protein = P4a-1 (*1*)
0.138 x 2,926 = 404 MDa of virion Mr = P4a-1

Parenthetically:
Protein content of virion from above calculation = 2,926 / 3,260 = 89.75% protein. Compared to:
The composition of the Vaccinia particle is 90% protein, 5% lipid and 3.2% DNA (*1*).

Molecular counting calculation, assuming the core wall lattice model entirely encloses the core:

Area of 10-hole lattice segment in molecular model
Lattice (10 holes) falls within a 69.5 x 24.2 nm = 1,681 sq. nm box.

Total #holes in core wall
(Surface area of core / surface area of 10 holes) x 10 = (113,270/1681) x10 = 673 holes enclosing the core
Each hole represents addition of 3 unique trimers to lattice = 9 P4a-1 molecules
So, 9 x 673 = 6,064 P4a-1 molecules, if lattice entirely encloses the core.
Given Mr of each P4a-1 = 71 kDa, then mass of 6,064 P4a-1 molecules = 430 MDa

Parenthetically: 430 MDa = 13.2% of total virion dry mass; 6,064 P4a-1 molecules = 62 million atoms.

The similarity of the two numbers (404 vs. 430 MDa) suggests P4a-1 forms a near-complete mesh around the core (as one might expect) with localized areas of disruption.

CHAPTER 8

Summary and conclusions

The architecture of the vaccinia virion at the molecular level has remained a stubborn problem, largely untouched by great strides in imaging, structural and molecular biology technologies. Over the past 60 years, nearly every relevant tool in the toolkits of structural and molecular biology have been implemented to address virion ultra- and molecular structure. At the ultrastructure level, the virion has been well characterized as brick-shaped with an outer envelope, a dumbbell-shaped proteinaceous core wall composed of a “palisade” outer layer enclosing a continuous inner layer within which the viral genome is packaged, and two opposing lateral bodies situated between core wall concavities and the envelope. Negative stain TEM and cryoEM images of vaccinia MV from 1963 [85] and 1996 [90] still provide some of the most enduring representations of the virion envelope and core wall surfaces, respectively. Various approaches (some of which are listed in Table 8.1) have been successfully applied to generate near-atomic structures for other viruses, both enveloped and non-enveloped. The same cannot be said for vaccinia, which has presented a near insurmountable challenge due, most likely, to its enveloped, polymorphic and asymmetric character. Heterogeneity between vaccinia virus particles is problematic for ensemble approaches such as X-ray crystallography and cryoEM that require either a high degree of crystallinity to generate a diffraction pattern [704] or the summing of thousands of images to produce a molecular model. This is compounded by vaccinia’s propensity to aggregate even after ultrasonication. While some viruses approximate to a platonic solid (the icosahedron), vaccinia may be more akin to a deformable, sticky “sack”. Atomic force microscopy (AFM) presents some advantages here in that it is a single-particle technique not reliant on summing or averaging, can image particles in their native environment (hydrated, in biological buffer) and can identify non-symmetrical surface features. However, AFM lacks atomic level resolution and is essentially a topographical approach, limited to surface features only. Complementary to these approaches is XLMS (Table 8.1) which, although not a direct imaging approach and only providing maximum distances ("restraints") between reactive (usually lysine amino) groups, nonetheless can query virion molecular structure in its native state irrespective of the presence of a virion envelope (the most commonly used crosslinking reagent, DSS, being membrane permeable) or its lack of symmetry.

Table 8.1. Comparison of common approaches for virus structure determination with XLMS.

	X-ray crystallography	***CryoEM***	***AFM***	***XLMS***
Structural features	Yes	Yes	Yes	Yes
Resolution	Atomic	Near-atomic	Moderately high	Low
Interrogate membranes	No	Yes, partially	Yes	Yes
Single particle or averaging	Averaging	Averaging	Single particle	Averaging
Symmetry requirement	Required	Ideal	Not required	Not required
Homogeneity requirement	Yes	Yes	No	Most likely
Protein abundance bias	No	No	No	Yes
Exterior bias	Yes	Yes	Yes	No
Sampling environment	Crystal	Vitrified ice	Native/hydrated	Native/hydrated

XLMS is also a powerful molecular approach for interrogating individual proteins and identifying their interacting partners. Table 8.2 provides a comparison of XLMS with molecular biology tools for identifying and characterizing protein-protein interactions. In contrast to X-ray crystallography or cryoEM, XLMS will not produce high-resolution three dimensional structures unassisted , but it can guide and/or validate (a) protein structure prediction, (b) protein docking, (c) the interpretation of cryoEM electron density maps.

Table 8.2. Approaches for identifying protein-protein interactions and protein localization.

	Immunogold-EM	***co-IP***	***Y2H***	***STORM***	***XLMS***
Structural information	No	No	No	No	Yes
Resolution	Low	Low	Low	Low	Low
Sample purity requirements	Moderate	Moderate	-	Moderate	High
Sample amount	Low	Low	Low	Low	Moderate
Investigate "flexible" proteins	Yes	Yes	Yes	Yes	Yes
Expression in heterologous host	No	No	Yes	No	No
Protein localization	Yes	No	No	Yes	Yes
Distance measurement	–	–	–	–	Maximum
Identification of protein - protein interactions	No	Yes	Yes	No	Yes
Membrane proteins	Yes	Yes	Yes	Yes	Yes

While XLMS requires high sample purity for the confident identification of crosslinked peptides (due to factors addressed in **Chapter 5**), XLMS is far more tolerant of contaminants than either X-ray crystallography or cryoEM, which require a far more rigorous degree of sample purity (essentially homogeneity). XLMS can also identify interactions between virion proteins *in situ*, not reliant on the solubilization of virion structural proteins without the disruption of interfaces (solubilization occurs after covalent bond formation in situ), or the reconstruction of complexes during or after heterologous expression, and does not require antibodies and/or bait-tagging with either an affinity handle or a fluorophore to identify proteins within a complex or protein localization within the virion; commercial antibodies are not available for the majority of vaccinia virion proteins and are expensive to produce. In practice, XLMS has also been successful at validating previously reported interactions between packaged proteins and has provided additional validation beyond what could be achieved by yeast 2-hybrid (Y2H) screening [507], which, as an approach, requires heterologous expression in yeast and also only applies to binary complexes. **Chapter 4**, in particular, describes the identification of higher order (ultrastructure-level) complexes of virion proteins.

XLMS is compatible with and can be integrated into a number of structural biology workflows and structure prediction approaches. It is perhaps most informative, however, in the context of authentic atomic structures of individual protein chains, whereby it can potentially link these structures into a higher order assembly, at least as “placeholders” for experimental structures that are not yet available. At the outset of the work described in this dissertation, atomic structures were available for only parts or all of 12 of the ~75 - 85 packaged vaccinia gene products, with two additional atomic structures available from other orthopoxviruses. Atomic models for the major structural proteins of the virion had not been resolved. In this regard, alongside experimental XLMS, described in part in **Chapter 4**, we pursued structural homology predictions for vaccinia virion proteins that have no sequence homologs outside the *Poxviridae*. After initial unsuccessful attempts at identifying structural homologs (and predicting structural models) through the online Robetta server, we asked whether a sequence-based structural homolog approach (HHsuite, with downstream integration with Modeller) could be applied to vaccinia virus proteins. As Vaccinia is a member of the NCLDV, we also investigated whether structural homology could be identified between vaccinia proteins and other member viruses. While the NCLDV share a common set of core genes, identification of additional orthologous genes had been limited by the use of mostly sequence-based homology approaches. However, a significant portion of the proteins encoded by NCLDV families display minimal or nonexistent sequence homology to previously characterized proteins. In **Chapter 2** and **Chapter 3** we have described the first large scale genomic analyses of organisms using the HHsuite software. By implementing this new approach to protein characterization, we were able to provide first-time annotations for 15 - 39% of previously uncharacterized proteins from 20 viruses in the NCLDV and able to confirm 20 - 98% of existing annotations. These findings also revealed many first-time occurrences of unique proteins in eukaryotic viruses. The results of these studies were beneficial to not just the poxvirus field, but also to the larger NCLDV community.

Renewed interest in the vaccinia RNA polymerase complex following directly from our work described in **Chapter 2** (after we requested permission to reprint a published figure from the Cramer lab and sent a copy of our work during that process) resulting, a little later, in published cryoEM structures for the complete vaccinia RNA polymerase/transcriptosome complex with early gene transcription factors (Rap94, ETF1, and ETF2), mRNA processing factors (MCEL and MCES), transcription termination factor (NPH1) and RNA polymerase stabilizing protein (E11) [71]. This essentially doubled the number of atomic structures available for packaged proteins from 14 to 27 (the atomic structures of the mRNA processing factors having been previously reported [705]). The 2022 global monkeypox outbreak led to an upsurge of interest in the orthopoxviruses. Currently, in the Protein Data Bank are atomic structures for parts or all of 32 packaged gene products. Atomic structures however have not yet been reported for the major core wall proteins.

Although the work described in **Chapter 3** provided many new insights into protein structural families within the NCLDV, we were unable to identify non-poxviral homologs of the major structural envelope and core wall proteins of vaccinia virus that could allow us to generate "placeholder" atomic structural predictions necessary to convert the protein-protein interaction networks we identified by XLMS to three dimensional models of molecular complexes, described in **Chapter 4**, and thereby to resolve the molecular architecture of the virion envelope surface and core wall, two of our primary areas of interest. At this point, however, protein structural prediction by deep learning (AlphaFold2) had just "come of age" and, unlike other protein structure prediction approaches, showed remarkable accuracy even when predicting structures *de novo* [655, 657]. We also strove to conduct an extensive review of the literature on virion protein function in order to maximally inform our XLMS work with data on known or suspected protein-protein interactions, members of higher order complexes, protein multimerization status and concordant defects at the genetic level. The work described in **Chapters 6 and 7** combines the protein-protein interaction networks identified by XLMS with high confidence AlphaFold2 and AlphaFold-multimer predicted models and the known structural biology of vaccinia virion proteins to build or extend structural models of the virion envelope and surface proteins and the palisade layer of the core wall. We describe protein-protein interactions among the virion attachment proteins and assembly of the proteins of the 11-member EFC as two distinct protein subassemblies. We also provide a maturation and simultaneous multimerization pathway for core structural protein P4a-1, the major component of the palisade layer of the core wall, and suggest that the D13 external scaffold, via transmembrane protein A17, may template the condensation of P4a-1 trimers. These trimers appear to form a honeycomb lattice reflecting that of D13.

The work described in **Chapters 6 and 7** regarding the structure and higher order assembly of the virion envelope proteins and P4a-1, respectively, is ongoing. We have achieved one of the most comprehensive crosslinking proteomes in the XLMS field, with 22,028 unique intra- and inter-protein crosslink pairs. In terms of protein hierarchy **Chapter 7** we have made maximal use of crosslink-type information to show that a large portion of homomultimer crosslinks between P4a-1 molecules are attributable to higher order organization of P4a-1 trimers (not just between individual P4a-1 chains within a trimer), allowing us to extend P4a-1 trimers into a larger, hexameric assembly. This was facilitated, in part, by our identification of the trimeric nature of the palisade spikes shown in ref. [90], as supported by XLMS data and AlphaFold-multimer models. These findings were later validated by cryoET [94]. This work has highlighted the importance of identifying the stoichiometry and multimerization state of the other major structural proteins of the virion core, namely, P4a-3, P4b, VP8, A4, and A12, which represents ongoing work. We have also modeled all packaged proteins with AlphaFold2 and identified various multimeric assemblies within the virion for which XLMS data were sparse. Continuing work includes the crosslink-guided docking of P4a-3 to core wall protein P4b, characterizing interactions of members of the “7PC”, and structural and functional analyses of

Vaccinia protease I7. The combination of XLMS, protein crystallization, which is being actively pursued by other members of the Gershon lab, and AlphaFold2 opens new avenues for studying protein-protein interactions within the vaccinia virion.

REFERENCES

1. Fenner, F., *Smallpox and Its Eradication* History of International Public Health, No. 6. 1988, Geneva: World Health Organization.
2. Fenner, F., *Smallpox: emergence, global spread, and eradication.* Hist Philos Life Sci, 1993. **15**(3): p. 397-420.
3. McInnes, C.J., Damon, I.K., Smith, G.L., McFadden, G., Isaacs, S.N., Roper, R.L., Evans, D.H., Damaso, C.R., Carulei, O., Wise, L.M., and Lefkowitz, E.J., *ICTV Virus Taxonomy Profile: Poxviridae 2023.* J Gen Virol, 2023. **104**(5).
4. Silva, N.I.O., de Oliveira, J.S., Kroon, E.G., Trindade, G.S., and Drumond, B.P., *Here, There, and Everywhere: The Wide Host Range and Geographic Distribution of Zoonotic Orthopoxviruses.* Viruses, 2020. **13**(1).
5. Buller, R.M. and Palumbo, G.J., *Poxvirus pathogenesis.* Microbiol Rev, 1991. **55**(1): p. 80-122.
6. Diven, D.G., *An overview of poxviruses.* J Am Acad Dermatol, 2001. **44**(1): p. 1-16.
7. Basdag, H., Rainer, B.M., and Cohen, B.A., *Molluscum contagiosum: to treat or not to treat? Experience with 170 children in an outpatient clinic setting in the northeastern United States.* Pediatr Dermatol, 2015. **32**(3): p. 353-7.
8. Moore, R.M., Jr., *Human Orf in the United States, 1972.* J Infect Dis, 1973. **127**(6): p. 731-2.
9. Downie, A.W. and Espana, C., *Comparison of Tanapox virus and Yaba-like viruses causing epidemic disease in monkeys.* J Hyg (Lond), 1972. **70**(1): p. 23-32.
10. Fenner, F., *The global eradication of smallpox.* Med J Aust, 1980. **1**(10): p. 455-5.
11. Meyer, H., Ehmann, R., and Smith, G.L., *Smallpox in the Post-Eradication Era.* Viruses, 2020. **12**(2).
12. Americo, J.L., Earl, P.L., and Moss, B., *Virulence differences of mpox (monkeypox) virus clades I, IIa, and IIb.1 in a small animal model.* Proc Natl Acad Sci U S A, 2023. **120**(8): p. e2220415120.
13. Okwor, T., Mbala, P.K., Evans, D.H., and Kindrachuk, J., *A contemporary review of clade-specific virological differences in monkeypox viruses.* Clin Microbiol Infect, 2023.
14. Organization, W.H., *2022-23 Mpox Outbreak: Global Trends*. 2023, World Health Organization.: World Health Organization, Geneva. p. https://worldhealthorg.shinyapps.io/mpx_global/
15. Africa, W.H.O.R.O.f., *WHO Africa Weekly Bulletin on Outbreaks and Other Emergencies - Week 42 : 12 - 18 October 2020*. 2020, World Health Organization Regional Office for Africa (WHO-AFRO): Brazzaville, Congo.
16. Chakraborty, S., Mohapatra, R.K., Chandran, D., Alagawany, M., Sv, P., Islam, M.A., Chakraborty, C., and Dhama, K., *Monkeypox vaccines and vaccination strategies: Current knowledge and advances. An update - Correspondence.* Int J Surg, 2022. **105**: p. 106869.
17. Fine, P.E., Jezek, Z., Grab, B., and Dixon, H., *The transmission potential of monkeypox virus in human populations.* Int J Epidemiol, 1988. **17**(3): p. 643-50.
18. See, K.C., *Vaccination for Monkeypox Virus Infection in Humans: A Review of Key Considerations.* Vaccines (Basel), 2022. **10**(8).

19. Reina, J. and Iglesias, C., *Vaccines against monkeypox.* Med Clin (Barc), 2023. **160**(7): p. 305-309.

20. Ryckeley, C., Goodwin, G., and Alvarez-Calderon, A., *The Reemerging Condition of Vaccinia: A Case Report and Brief Review of Monkeypox and Vaccinia Vaccines.* Am J Case Rep, 2023. **24**: p. e941006.

21. Torres-Laboy, P., Militello, M., Dykes, R., and Krishnamurthy, K., *Beyond the norm: A case of prolonged mpox virus infection.* JAAD Case Rep, 2023. **39**: p. 139-141.

22. Overton, E.T., Lawrence, S.J., Stapleton, J.T., Weidenthaler, H., Schmidt, D., Koenen, B., Silbernagl, G., Nopora, K., and Chaplin, P., *A randomized phase II trial to compare safety and immunogenicity of the MVA-BN smallpox vaccine at various doses in adults with a history of AIDS.* Vaccine, 2020. **38**(11): p. 2600-2607.

23. Ilchmann, H., Samy, N., Reichhardt, D., Schmidt, D., Powell, J.D., Meyer, T.P.H., Silbernagl, G., Nichols, R., Weidenthaler, H., De Moerlooze, L., Chen, L., and Chaplin, P., *One- and Two-Dose Vaccinations With Modified Vaccinia Ankara-Bavarian Nordic Induce Durable B-Cell Memory Responses Comparable to Replicating Smallpox Vaccines.* J Infect Dis, 2023. **227**(10): p. 1203-1213.

24. Hatch, G.J., et al., *Assessment of the protective effect of Imvamune and Acam2000 vaccines against aerosolized monkeypox virus in cynomolgus macaques.* J Virol, 2013. **87**(14): p. 7805-15.

25. Oliveira, J.S., Figueiredo, P.O., Costa, G.B., Assis, F.L., Drumond, B.P., da Fonseca, F.G., Nogueira, M.L., Kroon, E.G., and Trindade, G.S., *Vaccinia Virus Natural Infections in Brazil: The Good, the Bad, and the Ugly.* Viruses, 2017. **9**(11).

26. Fenner, F., *Risks and benefits of vaccinia vaccine use in the worldwide smallpox eradication campaign.* Res Virol, 1989. **140**(5): p. 465-6; discussion 487-91.

27. Kroon, E.G., Mota, B.E., Abrahao, J.S., da Fonseca, F.G., and de Souza Trindade, G., *Zoonotic Brazilian Vaccinia virus: from field to therapy.* Antiviral Res, 2011. **92**(2): p. 150-63.

28. Upton, C., Slack, S., Hunter, A.L., Ehlers, A., and Roper, R.L., *Poxvirus orthologous clusters: toward defining the minimum essential poxvirus genome.* J Virol, 2003. **77**(13): p. 7590-600.

29. Moss, B., *Poxviridae*, in *Fields Virology*. 2013, Wolters Kluwer | Lippincott, Williams & Wilkins: Philadelphia. p. 2129-2159.

30. Condit, R.C., Moussatche, N., and Traktman, P., *In a nutshell: structure and assembly of the vaccinia virion*, in *Advances in Virus Research*, K. Maramorosch and J. Shatkin, Editors. 2006, Elsevier. p. 31–124.

31. Truong, C.S. and Yoo, S.Y., *Oncolytic Vaccinia Virus in Lung Cancer Vaccines.* Vaccines (Basel), 2022. **10**(2).

32. Samson, A., et al., *Neoadjuvant Intravenous Oncolytic Vaccinia Virus Therapy Promotes Anticancer Immunity in Patients.* Cancer Immunol Res, 2022. **10**(6): p. 745-756.

33. Verardi, P.H., Titong, A., and Hagen, C.J., *A vaccinia virus renaissance: new vaccine and immunotherapeutic uses after smallpox eradication.* Hum Vaccin Immunother, 2012. **8**(7): p. 961-70.

34. Moss, B., *Genetically engineered poxviruses for recombinant gene expression, vaccination, and safety.* Proc. Natl. Acad. Sci. USA, 1996. **93**: p. 11341-11348.

35. Ulaeto, D. and Hruby, D.E., *Uses of vaccinia virus in vaccine delivery.* Curr Opin Biotechnol, 1994. **5**(5): p. 501-4.

36. Koonin, E.V. and Yutin, N., *Multiple evolutionary origins of giant viruses.* F1000Res, 2018. **7**.

37. Koonin, E.V. and Krupovic, M., *A life LINE for large viruses.* Elife, 2022. **11**.

38. Senkevich, T.G., Yutin, N., Wolf, Y.I., Koonin, E.V., and Moss, B., *Ancient Gene Capture and Recent Gene Loss Shape the Evolution of Orthopoxvirus-Host Interaction Genes.* mBio, 2021. **12**(4): p. e0149521.

39. Babkin, I.V., Babkina, I.N., and Tikunova, N.V., *An Update of Orthopoxvirus Molecular Evolution.* Viruses, 2022. **14**(2).

40. Babkin, I.V. and Shchelkunov, S.N., *[The time scale in poxvirus evolution].* Mol Biol (Mosk), 2006. **40**(1): p. 20-4.

41. Woo, A.C., Gaia, M., Guglielmini, J., Da Cunha, V., and Forterre, P., *Phylogeny of the Varidnaviria Morphogenesis Module: Congruence and Incongruence With the Tree of Life and Viral Taxonomy.* Front Microbiol, 2021. **12**: p. 704052.

42. Koonin, E.V., Dolja, V.V., Krupovic, M., Varsani, A., Wolf, Y.I., Yutin, N., Zerbini, F.M., and Kuhn, J.H., *Global Organization and Proposed Megataxonomy of the Virus World.* Microbiol Mol Biol Rev, 2020. **84**(2).

43. Filee, J., *Lateral gene transfer, lineage-specific gene expansion and the evolution of Nucleo Cytoplasmic Large DNA viruses.* J Invertebr Pathol, 2009. **101**(3): p. 169-71.

44. Koonin, E.V. and Yutin, N., *Evolution of the Large Nucleocytoplasmic DNA Viruses of Eukaryotes and Convergent Origins of Viral Gigantism.* Adv Virus Res, 2019. **103**: p. 167-202.

45. Alonso, C., Borca, M., Dixon, L., Revilla, Y., Rodriguez, F., Escribano, J.M., and Ictv Report, C., *ICTV Virus Taxonomy Profile: Asfarviridae.* J Gen Virol, 2018. **99**(5): p. 613-614.

46. Temmam, S., et al., *Faustovirus-Like Asfarvirus in Hematophagous Biting Midges and Their Vertebrate Hosts.* Front Microbiol, 2015. **6**: p. 1406.

47. Andreani, J., Khalil, J.Y.B., Sevvana, M., Benamar, S., Di Pinto, F., Bitam, I., Colson, P., Klose, T., Rossmann, M.G., Raoult, D., and La Scola, B., *Pacmanvirus, a New Giant Icosahedral Virus at the Crossroads between Asfarviridae and Faustoviruses.* J Virol, 2017. **91**(14).

48. Bajrai, L.H., Benamar, S., Azhar, E.I., Robert, C., Levasseur, A., Raoult, D., and La Scola, B., *Kaumoebavirus, a New Virus That Clusters with Faustoviruses and Asfarviridae.* Viruses, 2016. **8**(11).

49. Abrahao, J., et al., *Tailed giant Tupanvirus possesses the most complete translational apparatus of the known virosphere.* Nat Commun, 2018. **9**(1): p. 749.

50. Philippe, N., Legendre, M., Doutre, G., Coute, Y., Poirot, O., Lescot, M., Arslan, D., Seltzer, V., Bertaux, L., Bruley, C., Garin, J., Claverie, J.M., and Abergel, C., *Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes.* Science, 2013. **341**(6143): p. 281-6.

51. La Scola, B., Audic, S., Robert, C., Jungang, L., de Lamballerie, X., Drancourt, M., Birtles, R., Claverie, J.M., and Raoult, D., *A giant virus in amoebae.* Science, 2003. **299**(5615): p. 2033.

52. Schroeder, D.C., Oke, J., Malin, G., and Wilson, W.H., *Coccolithovirus (Phycodnaviridae): characterisation of a new large dsDNA algal virus that infects Emiliana huxleyi.* Arch Virol, 2002. **147**(9): p. 1685-98.

53. Van Etten, J.L., Graves, M.V., Muller, D.G., Boland, W., and Delaroque, N., *Phycodnaviridae--large DNA algal viruses.* Arch Virol, 2002. **147**(8): p. 1479-516.
54. Legendre, M., et al., *Thirty-thousand-year-old distant relative of giant icosahedral DNA viruses with a pandoravirus morphology.* Proc Natl Acad Sci U S A, 2014. **111**(11): p. 4274-9.
55. Yu, H., Xiao, H.Y., Li, N., Yang, C.J., and Huang, G.H., *An Ascovirus Utilizes Different Types of Host Larval Regulated Cell Death Mechanisms To Produce and Release Vesicles.* J Virol, 2023. **97**(1): p. e0156622.
56. Chinchar, V.G., Hick, P., Ince, I.A., Jancovich, J.K., Marschang, R., Qin, Q., Subramaniam, K., Waltzek, T.B., Whittington, R., Williams, T., Zhang, Q.Y., and Ictv Report, C., *ICTV Virus Taxonomy Profile: Iridoviridae.* J Gen Virol, 2017. **98**(5): p. 890-891.
57. Yang, C.J., Ren, G.H., Du, X.X., Li, S.W., Qian, Y.R., Huang, G.H., and Yu, H., *Comparisons of pathogenic course of two Heliothis virescens ascovirus isolates (HvAV-3i and HvAV-3j) in four noctuid (Lepidoptera) pest species.* J Invertebr Pathol, 2022. **189**: p. 107734.
58. Kleespies, R.G., Tidona, C.A., and Darai, G., *Characterization of a new iridovirus isolated from crickets and investigations on the host range.* J Invertebr Pathol, 1999. **73**(1): p. 84-90.
59. Rimmer, A.E., Becker, J.A., Tweedie, A., Lintermans, M., Landos, M., Stephens, F., and Whittington, R.J., *Detection of dwarf gourami iridovirus (Infectious spleen and kidney necrosis virus) in populations of ornamental fish prior to and after importation into Australia, with the first evidence of infection in domestically farmed Platy (Xiphophorus maculatus).* Prev Vet Med, 2015. **122**(1-2): p. 181-94.
60. Elde, N.C., Child, S.J., Eickbush, M.T., Kitzman, J.O., Rogers, K.S., Shendure, J., Geballe, A.P., and Malik, H.S., *Poxviruses deploy genomic accordions to adapt rapidly against host antiviral defenses.* Cell, 2012. **150**(4): p. 831-41.
61. Mirzakhanyan, Y. and Gershon, P.D., *Multisubunit DNA-Dependent RNA Polymerases from Vaccinia Virus and Other Nucleocytoplasmic Large-DNA Viruses: Impressions from the Age of Structure.* Microbiol Mol Biol Rev, 2017. **81**(3).
62. Colson, P. and Raoult, D., *Gene repertoire of amoeba-associated giant viruses.* Intervirology, 2010. **53**(5): p. 330-43.
63. Rahman, M.J., Haller, S.L., Stoian, A.M.M., Li, J., Brennan, G., and Rothenburg, S., *LINE-1 retrotransposons facilitate horizontal gene transfer into poxviruses.* Elife, 2022. **11**.
64. Fixsen, S.M., Cone, K.R., Goldstein, S.A., Sasani, T.A., Quinlan, A.R., Rothenburg, S., and Elde, N.C., *Poxviruses capture host genes by LINE-1 retrotransposition.* Elife, 2022. **11**.
65. Feng, Q., Moran, J.V., Kazazian, H.H., Jr., and Boeke, J.D., *Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition.* Cell, 1996. **87**(5): p. 905-16.
66. Mathias, S.L., Scott, A.F., Kazazian, H.H., Jr., Boeke, J.D., and Gabriel, A., *Reverse transcriptase encoded by a human transposable element.* Science, 1991. **254**(5039): p. 1808-10.

67. Szajner, P., Weisberg, A.S., Lebowitz, J., Heuser, J., and Moss, B., *External scaffold of spherical immature poxvirus particles is made of protein trimers, forming a honeycomb lattice.* J. Cell Biol., 2005. **170**: p. 971-981.

68. Gordon, M., *Virus Bodies: John Buist and the Elementary Bodies of Vaccinia.* Edinb Med J, 1937. **44**(2): p. 65-71.

69. Goebel, S.J., Johnson, G.P., Perkus, M.E., Davis, S.W., Winslow, J.P., and Paoletti, E., *The complete DNA sequence of vaccinia virus.* Virology, 1990. **179**: p. 247-266.

70. Moss, B., Ahn, B.-Y., Amegadzie, B., Gershon, P.D., and Keck, J.G., *Cytoplasmic transcription system encoded by vaccinia virus (minireview).* J. Biol. Chem., 1991. **266**: p. 1355-1358.

71. Grimm, C., et al., *Structural Basis of Poxvirus Transcription: Vaccinia RNA Polymerase Complexes.* Cell, 2019. **179**(7): p. 1537-1550 e19.

72. Hillen, H.S., Bartuli, J., Grimm, C., Dienemann, C., Bedenk, K., Szalay, A.A., Fischer, U., and Cramer, P., *Structural Basis of Poxvirus Transcription: Transcribing and Capping Vaccinia Complexes.* Cell, 2019. **179**(7): p. 1525-1536 e12.

73. Moussatche, N. and Condit, R.C., *Fine structure of the vaccinia virion determined by controlled degradation and immunolocalization.* Virology, 2015. **475**: p. 204-18.

74. Cyrklaff, M., Linaroudis, A., Boicu, M., Chlanda, P., Baumeister, W., Griffiths, G., and Krijnse-Locker, J., *Whole cell cryo-electron tomography reveals distinct disassembly intermediates of vaccinia virus.* PLoS One, 2007. **2**(5): p. e420.

75. Carter, G.C., Law, M., Hollinshead, M., and Smith, G.L., *Entry of the vaccinia virus intracellular mature virion and its interactions with glycosaminoglycans.* J Gen Virol, 2005. **86**(Pt 5): p. 1279-1290.

76. Schmidt, F.I., Bleck, C.K., Reh, L., Novy, K., Wollscheid, B., Helenius, A., Stahlberg, H., and Mercer, J., *Vaccinia virus entry is followed by core activation and proteasome-mediated release of the immunomodulatory effector VH1 from lateral bodies.* Cell Rep, 2013. **4**(3): p. 464-76.

77. Bidgood, S.R., Samolej, J., Novy, K., Collopy, A., Albrecht, D., Krause, M., Burden, J.J., Wollscheid, B., and Mercer, J., *Poxviruses package viral redox proteins in lateral bodies and modulate the host oxidative response.* PLoS Pathog, 2022. **18**(7): p. e1010614.

78. Green, R.H., Anderson, T.F., and Smadel, J.E., *Morphological Structure of the Virus of Vaccinia.* J Exp Med, 1942. **75**(6): p. 651-6.

79. Malkin, A.J., McPherson, A., and Gershon, P.D., *Structure of intracellular mature vaccinia virus visualized by in situ atomic force microscopy.* Journal of Virology, 2003. **77**(11): p. 6332-6340.

80. Kuznetsov, Y., Gershon, P.D., and McPherson, A., *Atomic force microscopy investigation of vaccinia virus structure.* J Virol., 2008. **82**(15): p. 7551-7566.

81. Dawson, I.M. and McFarlane, A.S., Nature, 1948. **161**: p. 464.

82. Gaylord, W.H., Jr., Melnick, J.L., and Bunting, H., *Intracellular development of vaccinia virus.* Proc Soc Exp Biol Med, 1952. **80**(1): p. 24-7.

83. Morgan, C., Ellison, S.A., Rose, H.M., and Moore, D.H., *Structure and development of viruses observed in the electron microscope. II. Vaccinia and fowl pox viruses.* J Exp Med, 1954. **100**(3): p. 301-10.

84. Westwood, J.C., Harris, W.J., Zwartouw, H.T., Titmuss, D.H., and Appleyard, G., *Studies on the Structure of Vaccinia Virus.* J Gen Microbiol, 1964. **34**: p. 67-78.

85. Dales, S., *The uptake and development of vaccinia virus in L strain cells followed with labeled viral deoxynucleic acid.* Journal of Cell Biology, 1963. **18**: p. 51-72.

86. Nagington, J. and Horne, R.W., *Morphological studies of orf and vaccinia viruses.* Virology, 1962. **16**: p. 248-260.

87. Easterbrook, K.B., *Controlled degradation of vaccinia virions in vitro: An electron microscopic study.* Journal of Ultrastructural Research, 1966. **14**: p. 484-496.

88. Medzon, E.L. and Bauer, H., *Structural features of vaccinia virus revealed by negative staining, sectioning, and freeze-etching.* Virology, 1970. **40**(4): p. 860-867.

89. Stern, W. and Dales, S., *Biogenesis of vaccinia: isolation and characterization of a surface component that elicts antibody suppressing infectivity and cell-cell fusion.* Virology, 1976. **75**: p. 232-241.

90. Dubochet, J., Adrian, M., Richter, K., Garces, J., and Wittek, R., *Structure of intracellular mature vaccinia virus observed by cryoelectron microscopy.* Journal of Virology, 1994. **68**(3): p. 1935-1941.

91. Peters, D., *Morphology of resting vaccinia virus.* Nature, 1956. **178**(4548): p. 1453-5.

92. Roos, N., Cyrklaff, M., Cudmore, S., Blasco, R., Krijnse-Locker, J., and Griffiths, G., *A novel immunogold cryoelectron microscopic approach to investigate the structure of the intracellular and extracellular forms of vaccinia virus.* EMBO J. , 1996. **15**: p. 2345-2355.

93. Cyrklaff, M., Risco, C., Fernández, J.J., Jiménez, M.V., Estéba, M., Baumeister, W., and Carrascosa, J.L., *Cryo-electron tomography of vaccinia virus.* Proc Natl Acad Sci U S A, 2005. **102**(8): p. 2772-2777.

94. Hernandez-Gonzalez, M., Calcraft, T., Nans, A., Rosenthal, P.B., and Way, M., *A succession of two viral lattices drives vaccinia virus assembly.* PLoS Biol, 2023. **21**(3): p. e3002005.

95. Peters, D. and Mueller, G., *The Fine Structure of the DNA-Containing Core of Vaccinia Virus.* Virology, 1963. **21**: p. 267-9.

96. Jesus, D.M., Moussatche, N., and Condit, R.C., *An improved high pressure freezing and freeze substitution method to preserve the labile vaccinia virus nucleocapsid.* J Struct Biol, 2016. **195**(1): p. 41-8.

97. Sarov, I. and Joklik, W.K., *Characterization of intermediates in the uncoating of vaccinia virus DNA.* Virology, 1972. **50**(2): p. 593-602.

98. Gray, R.D.M., Albrecht, D., Beerli, C., Huttunen, M., Cohen, G.H., White, I.J., Burden, J.J., Henriques, R., and Mercer, J., *Nanoscale polarization of the entry fusion complex of vaccinia virus drives efficient fusion.* Nat Microbiol, 2019. **4**(10): p. 1636-1644.

99. Sarov, I. and Joklik, W.K., *Studies on the nature and location of the capsid polypeptides of vaccinia virions.* Virology, 1972. **50**(2): p. 579-92.

100. Chung, C.S., Chen, C.H., Ho, M.Y., Huang, C.Y., Liao, C.L., and Chang, W., *Vaccinia virus proteome: identification of proteins in vaccinia virus intracellular mature virion particles.* Journal of Virology, 2006. **80**(5): p. 2127–2140.

101. Ngo, T., Mirzakhanyan, Y., and Gershon, P.D., *Protein primary structure of the Vaccinia virion at increased resolution.* J. Virol., 2016.

102. Manes, N.P., Estep, R.D., Mottaz, H.M., Moore, R.J., Clauss, T.R., Monroe, M.E., Du, X., Adkins, J.N., Wong, S.W., and D., S.R., *Comparative proteomics of human monkeypox and vaccinia intracellular mature and extracellular enveloped virions.* J. Proteome Res., 2008. **7**: p. 960-968.

103. Takahashi, T., Oie, M., and Ichihashi, Y., *N-terminal amino acid sequences of vaccinia virus structural proteins.* Virology, 1994. **202**: p. 844-852.

104. Liu, S., et al., *Cryo-EM Structure of the African Swine Fever Virus.* Cell Host Microbe, 2019. **26**(6): p. 836-843 e3.

105. Andres, G., Charro, D., Matamoros, T., Dillard, R.S., and Abrescia, N.G.A., *The cryo-EM structure of African swine fever virus unravels a unique architecture comprising two icosahedral protein capsids and two lipoprotein membranes.* J Biol Chem, 2020. **295**(1): p. 1-12.

106. Fang, Q., Zhu, D., Agarkova, I., Adhikari, J., Klose, T., Liu, Y., Chen, Z., Sun, Y., Gross, M.L., Van Etten, J.L., Zhang, X., and Rossmann, M.G., *Near-atomic structure of a giant virus.* Nat Commun, 2019. **10**(1): p. 388.

107. Shao, Q., Agarkova, I.V., Noel, E.A., Dunigan, D.D., Liu, Y., Wang, A., Guo, M., Xie, L., Zhao, X., Rossmann, M.G., Van Etten, J.L., Klose, T., and Fang, Q., *Near-atomic, non-icosahedrally averaged structure of giant virus Paramecium bursaria chlorella virus 1.* Nat Commun, 2022. **13**(1): p. 6476.

108. Okamoto, K., Miyazaki, N., Reddy, H.K.N., Hantke, M.F., Maia, F., Larsson, D.S.D., Abergel, C., Claverie, J.M., Hajdu, J., Murata, K., and Svenda, M., *Cryo-EM structure of a Marseilleviridae virus particle reveals a large internal microassembly.* Virology, 2018. **516**: p. 239-245.

109. Chiu, W.L., Lin, C.L., Yang, M.H., Tzou, D.L., and Chang, W., *Vaccinia virus 4c (A26L) protein on intracellular mature virus binds to the extracellular cellular matrix laminin.* J Virol, 2007. **81**(5): p. 2149-57.

110. Vazquez, M.I. and Esteban, M., *Identification of functional domains in the 14-kilodalton envelope protein (A27L) of vaccinia virus.* J Virol, 1999. **73**(11): p. 9098-109.

111. Kochan, G., Escors, D., Gonzalez, J.M., Casasnovas, J.M., and Esteban, M., *Membrane cell fusion activity of the vaccinia virus A17-A27 protein complex.* Cell Microbiol, 2008. **10**(1): p. 149-64.

112. Lin, C.L., Chung, C.S., Heine, H.G., and Chang, W., *Vaccinia virus envelope H3L protein binds to cell surface heparan sulfate and is important for intracellular mature virion morphogenesis and virus infection in vitro and in vivo.* J Virol, 2000. **74**(7): p. 3353-65.

113. Hsiao, J.C., Chung, C.S., and Chang, W., *Vaccinia virus envelope D8L protein binds to cell surface chondroitin sulfate and mediates the adsorption of intracellular mature virions to cells.* J Virol, 1999. **73**(10): p. 8750-61.

114. Foo, C.H., Lou, H., Whitbeck, J.C., Ponce-de-Leon, M., Atanasiu, D., Eisenberg, R.J., and Cohen, G.H., *Vaccinia virus L1 binds to cell surfaces and blocks virus entry independently of glycosaminoglycans.* Virology, 2009. **385**(2): p. 368-82.

115. Singh, K., Gittis, A.G., Gitti, R.K., Ostazeski, S.A., Su, H.P., and Garboczi, D.N., *The Vaccinia Virus H3 Envelope Protein, a Major Target of Neutralizing Antibodies, Exhibits a Glycosyltransferase Fold and Binds UDP-Glucose.* J Virol, 2016. **90**(10): p. 5020-30.

116. Shih, P.C., Yang, M.S., Lin, S.C., Ho, Y., Hsiao, J.C., Wang, D.R., Yu, S.S., Chang, W., and Tzou, D.M., *A turn-like structure "KKPE" segment mediates the specific binding of viral protein A27 to heparin and heparan sulfate on cell surfaces.* J Biol Chem, 2009. **284**(52): p. 36535-36546.

117. da Fonseca, F.G., Wolffe, E.J., Weisberg, A., and Moss, B., *Effects of deletion or stringent repression of the H3L envelope gene on vaccinia virus replication.* J. Virol., 2000. **74**(16): p. 7518-7528.

118. Su, H.P., Garman, S.C., Allison, T.J., Fogg, C., Moss, B., and Garboczi, D.N., *The 1.51-Angstrom structure of the poxvirus L1 protein, a target of potent neutralizing antibodies.* Proc Natl Acad Sci U S A, 2005. **102**(12): p. 4240-5.

119. da Fonseca, F.G., Wolffe, E.J., Weisberg, A.S., and Moss, B., *Characterization of the Vaccinia Virus H3L Envelope Protein: Topology and Posttranslational Membrane Insertion via the C-Terminal Hydrophobic Tail.* J Virol. , 2000. **74**(16): p. 7508–7517.

120. Niles, E.G. and Seto, J., *Vaccinia virus gene D8 encodes a virion transmembrane protein.* Journal of Virology, 1988. **62**: p. 3772-3778.

121. Rodriguez, J.R., Risco, C., Carrascosa, J.L., Esteban, M., and Rodriguez, D., *Characterization of early stages in vaccinia virus membrane biogenesis: implications of the 21-kilodalton protein and a newly identified 15-kilodalton envelope protein.* J Virol, 1997. **71**(3): p. 1821-33.

122. Rodriguez, D., Esteban, M., and Rodriguez, J.R., *Vaccinia virus A17L gene product is essential for an early step in virion morphogenesis.* J Virol, 1995. **69**(8): p. 4640-8.

123. Vazquez, M.I., Rivas, G., Cregut, D., Serrano, L., and Esteban, M., *The vaccinia virus 14-kilodalton (A27L) fusion protein forms a triple coiled-coil structure and interacts with the 21-kilodalton (A17L) virus membrane protein through a C-terminal alpha-helix.* J Virol, 1998. **72**(12): p. 10126-37.

124. Voeltz, G.K., Prinz, W.A., Shibata, Y., Rist, J.M., and Rapoport, T.A., *A class of membrane proteins shaping the tubular endoplasmic reticulum.* Cell, 2006. **124**(3): p. 573-86.

125. Ching, Y.C., Chung, C.S., Huang, C.Y., Hsia, Y., Tang, Y.L., and Chang, W., *Disulfide bond formation at the C termini of vaccinia virus A26 and A27 proteins does not require viral redox enzymes and suppresses glycosaminoglycan-mediated cell fusion.* J Virol, 2009. **83**(13): p. 6464-76.

126. Matho, M.H., Maybeno, M., Benhnia, M.R., Becker, D., Meng, X., Xiang, Y., Crotty, S., Peters, B., and Zajonc, D.M., *Structural and biochemical characterization of the vaccinia virus envelope protein D8 and its recognition by the antibody LA5.* J Virol, 2012. **86**(15): p. 8050-8.

127. Chang, H.W., Yang, C.H., Luo, Y.C., Su, B.G., Cheng, H.Y., Tung, S.Y., Carillo, K.J.D., Liao, Y.T., Tzou, D.M., Wang, H.C., and Chang, W., *Vaccinia viral A26 protein is a fusion suppressor of mature virus and triggers membrane fusion through conformational change at low pH.* PLoS Pathog, 2019. **15**(6): p. e1007826.

128. Chang, T.H., Chang, S.J., Hsieh, F.L., Ko, T.P., Lin, C.T., Ho, M.R., Wang, I., Hsu, S.T., Guo, R.T., Chang, W., and Wang, A.H., *Crystal structure of vaccinia viral A27 protein reveals a novel structure critical for its function and complex formation with A26 protein.* PLoS Pathog, 2013. **9**(8): p. e1003563.

129. Ichihashi, Y. and Oie, M., *Neutralizing epitope on penetration protein of vaccinia virus.* Virology, 1996. **220**(2): p. 491-4.

130. Bisht, H., Weisberg, A.S., and Moss, B., *Vaccinia virus l1 protein is required for cell entry and membrane fusion.* J Virol, 2008. **82**(17): p. 8687-94.

131. Amara, A. and Mercer, J., *Viral apoptotic mimicry.* Nat Rev Microbiol, 2015. **13**(8): p. 461-9.

132. Mercer, J. and Helenius, A., *Vaccinia virus uses macropinocytosis and apoptotic mimicry to enter host cells.* Science, 2008. **320**(5875): p. 531-5.

133. Mercer, J., Knebel, S., Schmidt, F.I., Crouse, J., Burkard, C., and Helenius, A., *Vaccinia virus strains use distinct forms of macropinocytosis for host-cell entry.* Proc Natl Acad Sci U S A, 2010. **107**(20): p. 9346-51.

134. Moss, B., *Membrane fusion during poxvirus entry.* Semin Cell Dev Biol, 2016. **60**: p. 89-96.

135. Townsley, A.C., Weisberg, A.S., Wagenaar, T.R., and Moss, B., *Vaccinia virus entry into cells via a low-pH-dependent endosomal pathway.* J Virol, 2006. **80**(18): p. 8899-908.

136. Armstrong, J.A., Metz, D.H., and Young, M.R., *The mode of entry of vaccinia virus into L cells.* J Gen Virol, 1973. **21**(3): p. 533-7.

137. Chang, A. and Metz, D.H., *Further investigations on the mode of entry of vaccinia virus into cells.* J Gen Virol, 1976. **32**(2): p. 275-82.

138. Doms, R.W., Blumenthal, R., and Moss, B., *Fusion of intra- and extracellular forms of vaccinia virus with the cell membrane.* J Virol, 1990. **64**(10): p. 4884-92.

139. Bengali, Z., Satheshkumar, P.S., and Moss, B., *Orthopoxvirus species and strain differences in cell entry.* Virology, 2012. **433**(2): p. 506-12.

140. Bengali, Z., Townsley, A.C., and Moss, B., *Vaccinia virus strain differences in cell attachment and entry.* Virology, 2009. **389**(1-2): p. 132-40.

141. Senkevich, T.G., Ojeda, S., Townsley, A., Nelson, G.E., and Moss, B., *Poxvirus multiprotein entry-fusion complex.* Proc Natl Acad Sci U S A, 2005. **102**(51): p. 18572-7.

142. Schin, A.M., Diesterbeck, U.S., and Moss, B., *Insights into the Organization of the Poxvirus Multicomponent Entry-Fusion Complex from Proximity Analyses in Living Infected Cells.* J Virol, 2021. **95**(16): p. e0085221.

143. Moss, B., *Poxvirus cell entry: how many proteins does it take?* Viruses, 2012. **4**(5): p. 688-707.

144. Allison, S.L., Schalich, J., Stiasny, K., Mandl, C.W., and Heinz, F.X., *Mutational evidence for an internal fusion peptide in flavivirus envelope protein E.* J Virol, 2001. **75**(9): p. 4268-75.

145. Durell, S.R., Martin, I., Ruysschaert, J.M., Shai, Y., and Blumenthal, R., *What studies of fusion peptides tell us about viral envelope glycoprotein-mediated membrane fusion (review).* Mol Membr Biol, 1997. **14**(3): p. 97-112.

146. Rodriguez-Crespo, I., Nunez, E., Gomez-Gutierrez, J., Yelamos, B., Albar, J.P., Peterson, D.L., and Gavilanes, F., *Phospholipid interactions of the putative fusion peptide of hepatitis B virus surface antigen S protein.* J Gen Virol, 1995. **76 (Pt 2)**: p. 301-8.

147. Pecheur, E.I., Sainte-Marie, J., Bienven e, A., and Hoekstra, D., *Peptides and membrane fusion: towards an understanding of the molecular mechanism of protein-induced fusion.* J Membr Biol, 1999. **167**(1): p. 1-17.

148. Tak, A.I., Americo, J.L., Diesterbeck, U.S., and Moss, B., *Loss of the vaccinia virus 35-amino acid hydrophobic O3 protein is partially compensated by mutations in the transmembrane domains of other entry proteins.* J Virol, 2021. **95**(8).

149. Ojeda, S., Senkevich, T.G., and Moss, B., *Entry of vaccinia virus and cell-cell fusion require a highly conserved cysteine-rich membrane protein encoded by the A16L gene.* J Virol, 2006. **80**(1): p. 51-61.

150. Townsley, A.C., Senkevich, T.G., and Moss, B., *The Product of the Vaccinia Virus L5R Gene Is a Fourth Membrane Protein Encoded by All Poxviruses That Is Required for Cell Entry and Cell-Cell Fusion.* J. Virol., 2005. **79**(17): p. 10988–10998.

151. Townsley, A.C., Senkevich, T.G., and Moss, B., *Vaccinia Virus A21 Virion Membrane Protein Is Required for Cell Entry and Fusion.* J.Virol. , 2005. **79:** : p. 9458-9469

152. Brown, E., Senkevich, T.G., and Moss, B., *Vaccinia Virus F9 Virion Membrane Protein Is Required for Entry but Not Virus Assembly, in Contrast to the Related L1 Protein.* J Virol. , 2006. **80**(19): p. 9455–9464.

153. Senkevich, T.G., Ward, B.M., and Moss, B., *Vaccinia virus entry into cells is dependent on a virion surface protein encoded by the A28L gene.* J Virol, 2004. **78**(5): p. 2357-66.

154. Nelson, G.E., Sisler, J.R., Chandran, D., and Moss, B., *Vaccinia virus entry/fusion complex subunit A28 is a target of neutralizing and protective antibodies.* Virology, 2008. **380**(2): p. 394-401.

155. Senkevich, T.G., Ward, B.M., and Moss, B., *Vaccinia virus A28L gene encodes an essential protein component of the virion membrane with intramolecular disulfide bonds formed by the viral cytoplasmic redox pathway.* J Virol, 2004. **78**(5): p. 2348-56.

156. Senkevich, T.G. and Moss, B., *Vaccinia virus H2 protein is an essential component of a complex involved in virus entry and cell-cell fusion.* J Virol, 2005. **79**(8): p. 4744-54.

157. Nelson, G.E., Wagenaar, T.R., and Moss, B., *A conserved sequence within the H2 subunit of the vaccinia virus entry/fusion complex is important for interaction with the A28 subunit and infectivity.* J Virol, 2008. **82**(13): p. 6244-50.

158. Turner, P.C., Dilling, B.P., Prins, C., Cresawn, S.G., Moyer, R.W., and Condit, R.C., *Vaccinia virus temperature-sensitive mutants in the A28 gene produce non-infectious virions that bind to cells but are defective in entry.* Virology, 2007. **366**(1): p. 62-72.

159. Laliberte, J.P., Weisberg, A.S., and Moss, B., *The membrane fusion step of vaccinia virus entry is cooperatively mediated by multiple viral proteins and host cell components.* PLoS Pathog, 2011. **7**(12): p. e1002446.

160. Diesterbeck, U.S., Gittis, A.G., Garboczi, D.N., and Moss, B., *The 2.1 A structure of protein F9 and its comparison to L1, two components of the conserved poxvirus entry-fusion complex.* Sci Rep, 2018. **8**(1): p. 16807.

161. Yang, F., Lin, S., Chen, Z., Yue, D., Yang, M., He, B., Cao, Y., Dong, H., Li, J., Zhao, Q., and Lu, G., *Structural basis of poxvirus A16/G9 binding for sub-complex formation.* Emerg Microbes Infect, 2023. **12**(1): p. 2179351.

162. Lin, S., Yue, D., Yang, F., Chen, Z., He, B., Cao, Y., Dong, H., Li, J., Zhao, Q., and Lu, G., *Crystal structure of vaccinia virus G3/L5 sub-complex reveals a novel fold with extended inter-molecule interactions conserved among orthopoxviruses.* Emerg Microbes Infect, 2023. **12**(1): p. e2160661.

163. Chang, S.J., Shih, A.C., Tang, Y.L., and Chang, W., *Vaccinia mature virus fusion regulator A26 protein binds to A16 and G9 proteins of the viral entry fusion complex and dissociates from mature virions at low pH.* J Virol, 2012. **86**(7): p. 3809-18.

164. Chang, S.J., Chang, Y.X., Izmailyan, R., Tang, Y.L., and Chang, W., *Vaccinia virus A25 and A26 proteins are fusion suppressors for mature virions and determine strain-specific virus entry pathways into HeLa, CHO-K1, and L cells.* J Virol, 2010. **84**(17): p. 8422-32.

165. Amegadzie, B.Y., Sisler, J.R., and Moss, B., *Frame-shift mutations within the vaccinia virus A-type inclusion protein gene.* Virology, 1992. **186**(2): p. 777-82.

166. Townsley, A.C. and Moss, B., *Two distinct low-pH steps promote entry of vaccinia virus.* J Virol, 2007. **81**(16): p. 8613-20.

167. Pedersen, K., Snijder, E.J., Schleich, S., Roos, N., Griffiths, G., and Locker, J.K., *Characterization of vaccinia virus intracellular cores: implications for viral uncoating and core structure.* Virol., 2000. **74**(8): p. 3525-3536.

168. Mann, B.A., Huang, J.H., Li, P., Chang, H.C., Slee, R.B., O'Sullivan, A., Anita, M., Yeh, N., Klemsz, M.J., Brutkiewicz, R.R., Blum, J.S., and Kaplan, M.H., *Vaccinia virus blocks Stat1-dependent and Stat1-independent gene expression induced by type I and type II interferons.* J Interferon Cytokine Res, 2008. **28**(6): p. 367-80.

169. Najarro, P., Traktman, P., and Lewis, J.A., *Vaccinia virus blocks gamma interferon signal transduction: viral VH1 phosphatase reverses Stat1 activation.* J Virol, 2001. **75**(7): p. 3185-96.

170. Bidgood, S.R. and Mercer, J., *Cloak and Dagger: Alternative Immune Evasion and Modulation Strategies of Poxviruses.* Viruses, 2015. **7**(8): p. 4800-25.

171. Greseth, M.D. and Traktman, P., *The Life Cycle of the Vaccinia Virus Genome.* Annu Rev Virol, 2022. **9**(1): p. 239-259.

172. Mallardo, M., Leithe, E., Schleich, S., Roos, N., Doglio, L., and Krijnse Locker, J., *Relationship between vaccinia virus intracellular cores, early mRNAs, and DNA replication sites.* J Virol, 2002. **76**(10): p. 5167-83.

173. Yang, Z. and Moss, B., *Interaction of the vaccinia virus RNA polymerase-associated 94-kilodalton protein with the early transcription factor.* J Virol, 2009. **83**(23): p. 12018-26.

174. Mallardo, M., Schleich, S., and Krijnse Locker, J., *Microtubule-dependent organization of vaccinia virus core-derived early mRNAs into distinct cytoplasmic structures.* Mol Biol Cell, 2001. **12**(12): p. 3875-91.

175. Tolonen, N., Doglio, L., Schleich, S., and Krijnse Locker, J., *Vaccinia virus DNA replication occurs in endoplasmic reticulum-enclosed cytoplasmic mini-nuclei.* Mol Biol Cell, 2001. **12**(7): p. 2031-46.

176. El-Jesr, M., Teir, M., and Maluquer de Motes, C., *Vaccinia Virus Activation and Antagonism of Cytosolic DNA Sensing.* Front Immunol, 2020. **11**: p. 568412.

177. Meade, N., King, M., Munger, J., and Walsh, D., *mTOR Dysregulation by Vaccinia Virus F17 Controls Multiple Processes with Varying Roles in Infection.* J Virol, 2019. **93**(15).

178. Meade, N., Furey, C., Li, H., Verma, R., Chai, Q., Rollins, M.G., DiGiuseppe, S., Naghavi, M.H., and Walsh, D., *Poxviruses Evade Cytosolic Sensing through Disruption of an mTORC1-mTORC2 Regulatory Circuit.* Cell, 2018. **174**(5): p. 1143-1157 e17.

179. Yang, Z., Reynolds, S.E., Martens, C.A., Bruno, D.P., Porcella, S.F., and Moss, B., *Expression profiling of the intermediate and late stages of poxvirus replication.* J Virol, 2011. **85**(19): p. 9899-908.

180. Broyles, S.S., *Vaccinia virus transcription.* J Gen Virol, 2003. **84**(Pt 9): p. 2293-303.

181. Roberts, K.L. and Smith, G.L., *Vaccinia virus morphogenesis and dissemination.* Trends Microbiol, 2008. **16**(10): p. 472-9.

182. Maruri-Avidal, L., Weisberg, A.S., and Moss, B., *Direct formation of vaccinia virus membranes from the endoplasmic reticulum in the absence of the newly characterized L2-interacting protein A30.5.* J Virol, 2013. **87**(22): p. 12313-26.

183. Traktman, P., Liu, K., DeMasi, J., Rollins, R., Jesty, S., and Unger, B., *Elucidating the essential role of the A14 phosphoprotein in vaccinia virus morphogenesis: Construction and characterization of a tetracycline-inducible recombinant.* Journal of Virology, 2000. **74**(8): p. 3682-3695.

184. Zhang, Y. and Moss, B., *Immature viral envelope formation is interrupted at the same stage by lac operator-mediated repression of the vaccinia virus D13L gene and by the drug rifampicin.* Virology, 1992. **187**(2): p. 643-53.

185. Resch, W., Weisberg, A.S., and Moss, B., *Vaccinia virus nonstructural protein encoded by the A11R gene is required for formation of the virion membrane.* J Virol, 2005. **79**(11): p. 6598-609.

186. Satheshkumar, P.S., Weisberg, A., and Moss, B., *Vaccinia virus H7 protein contributes to the formation of crescent membrane precursors of immature virions.* J Virol, 2009. **83**(17): p. 8439-50.

187. Maruri-Avidal, L., Domi, A., Weisberg, A.S., and Moss, B., *Participation of vaccinia virus l2 protein in the formation of crescent membranes and immature virions.* J Virol, 2011. **85**(6): p. 2504-11.

188. Meng, X., Embry, A., Rose, L., Yan, B., Xu, C., and Xiang, Y., *Vaccinia virus A6 is essential for virion membrane biogenesis and localization of virion membrane proteins to sites of virion assembly.* J Virol, 2012. **86**(10): p. 5603-13.

189. Heuser, J., *Deep-etch EM reveals that the early poxvirus envelope is a single membrane bilayer stabilized by a geodetic "honeycomb" surface coat.* Journal of Cell Biology, 2005. **169**(2): p. 269 - 283.

190. Bahar, M.W., Graham, S.C., Stuart, D.I., and Grimes, J.M., *Insights into the evolution of a complex virus from the crystal structure of vaccinia virus D13.* Structure, 2011. **19**(7): p. 1011-20.

191. Liu, L., Cooper, T., Howley, P.M., and Hayball, J.D., *From crescent to mature virion: vaccinia virus assembly and maturation.* Viruses, 2014. **6**(10): p. 3787-808.

192. Szajner, P., Jaffe, H., Weisberg, A.S., and Moss, B., *A complex of seven vaccinia virus proteins conserved in all chordopoxviruses is required for the association of membranes and viroplasm to form immature virions.* Virology, 2004. **330**(2): p. 447-59.

193. Szajner, P., Jaffe, H., Weisberg, A.S., and Moss, B., *Vaccinia virus G7L protein Interacts with the A30L protein and is required for association of viral membranes with dense viroplasm to form immature virions.* Journal of Virology, 2003. **77**(6): p. 3418-3412.

194. Dyster, L.M. and Niles, E.G., *Genetic and biochemical characterization of vaccinia virus genes D2L and D3R which encode virion structural proteins.* Virology, 1991. **182**(2): p. 455-67.

195. Chiu, W.L. and Chang, W., *Vaccinia virus J1R protein: a viral membrane protein that is essential for virion morphogenesis.* J Virol, 2002. **76**(19): p. 9575-87.

196. Betakova, T., Wolffe, E.J., and Moss, B., *Regulation of vaccinia virus morphogenesis: phosphorylation of the A14L and A17L membrane proteins and C-terminal truncation of the A17L protein are dependent on the F10L kinase.* J. Virol., 1999. **73**(5): p. 3534-3543.

197. Szajner, P., Weisberg AS, and B., M., *Evidence for an essential catalytic role of the F10 protein kinase in vaccinia virus morphogenesis.* J Virol., 2004. **78**(1): p. 257-265.

198. Unger, B. and Traktman, P., *Vaccinia virus morphogenesis: A13 phosphoprotein is required for assembly of mature virions.* J. Virol. , 2004. **78**(16): p. 8885–8901.

199. Burroughs, A.M., Iyer, L.M., and Aravind, L., *Comparative genomics and evolutionary trajectories of viral ATP dependent DNA-packaging systems.* Genome Dyn, 2007. **3**: p. 48-65.

200. Rzechorzek, N.J., Blackwood, J.K., Bray, S.M., Maman, J.D., Pellegrini, L., and Robinson, N.P., *Structure of the hexameric HerA ATPase reveals a mechanism of translocation-coupled DNA-end processing in archaea.* Nat Commun, 2014. **5**: p. 5506.

201. Ansarah-Sobrinho, C. and Moss, B., *Role of the I7 protein in proteolytic processing of vaccinia virus membrane and core components.* J Virol, 2004. **78**(12): p. 6335-43.

202. Kane, E.M. and Shuman, S., *Vaccinia virus morphogenesis is blocked by a temperature-sensitive mutation in the I7 gene that encodes a virion component.* J Virol, 1993. **67**(5): p. 2689-98.

203. Bisht, H., Weisberg, A.S., Szajner, P., and Moss, B., *Assembly and disassembly of the capsid-like external scaffold of immature virions during vaccinia virus morphogenesis.* J Virol, 2009. **83**(18): p. 9140-50.

204. Ericsson, M., Cudmore, S., Shuman, S., Condit, R.C., Griffiths, G., and Locker, J.K., *Characterization of ts 16, a temperature-sensitive mutant of vaccinia virus.* J Virol, 1995. **69**(11): p. 7072-86.

205. Lee, P. and Hruby, D.E., *Analysis of the role of the amino-terminal peptide of vaccinia virus structural protein precursors during proteolytic processing.* Virology, 1995. **207**(1): p. 229-33.

206. Byrd, C.M. and Hruby, D.E., *Vaccinia virus proteolysis--a review.* Rev Med Virol, 2006. **16**(3): p. 187-202.

207. Lee, P. and Hruby, D.E., *Proteolytic cleavage of vaccinia virus virion proteins. Mutational analysis of the specificity determinants.* J Biol Chem, 1994. **269**(11): p. 8616-22.

208. Lee, P. and Hruby, D.E., *trans processing of vaccinia virus core proteins.* J Virol, 1993. **67**(7): p. 4252-63.

209. VanSlyke, J.K., Franke, C.A., and Hruby, D.E., *Proteolytic maturation of vaccinia virus core proteins – identification of a conserved motif at the N termini of the 4b and 25K virion proteins.* Journal of General Virology, 1991. **72**: p. 411-416.

210. Mirzakhanyan, Y., Jankevics, A., Scheltema, R.A., and Gershon, P.D., *Combination of deep XLMS with deep learning reveals an ordered rearrangement and assembly of a major protein component of the vaccinia virion.* mBio, 2023: p. e0113523.

211. VanSlyke, J.K., Whitehead, S.S., WIlson, E.M., and Hruby, D.E., *The multi-step proteolytic maturation pathway utilized by vaccinia virus P4a protein: a degenerate conserved cleavage motif within core proteins.* Virology, 1991. **183**: p. 467-478.

212. Risco, C., Rodriguez, J.R., Demkowicz, W.E., Heljasvaara, R., Carrascosa, J.L., Esteban, M., and Rodriguez, D., *The vaccinia virus 39-kDa protein forms a stable complex with the p4a/4a major core protein early in morphogenesis.* Virology, 1999. **265**(2): p. 375-386.

213. Heljasvaara, R., Rodriguez, D., Risco, C., Carrascosa, J.L., Esteban, M., , and Rodriguez, J.R., *The Major Core Protein P4a (A10L Gene) of Vaccinia Virus Is Essential for Correct Assembly of Viral DNA into the Nucleoprotein Complex To Form Immature Viral Particle.* Journal of Virology, 2001. **75**: p. 5778-5795.

214. Whitehead, S.S., Bersani, N.A., and Hruby, D.E., *Physical and molecular genetic analysis of the multistep proteolytic maturation pathway utilized by vaccinia virus P4a protein.* J Gen Virol, 1995. **76 (Pt 3)**: p. 717-21.

215. Xu, A., Basant, A., Schleich, S., Newsome, T.P., and Way, M., *Kinesin-1 transports morphologically distinct intracellular virions during vaccinia infection.* J Cell Sci, 2023. **136**(5).

216. Smith, G.L., Vanderplasschen, A., and Law, M., *The formation and function of extracellular enveloped vaccinia virus.* J Gen Virol, 2002. **83**(Pt 12): p. 2915-2931.

217. Bae, B., Feklistov, A., Lass-Napiorkowska, A., Landick, R., and Darst, S.A., *Structure of a bacterial RNA polymerase holoenzyme open promoter complex.* Elife, 2015. **4**.

218. Murakami, K.S., *Structural biology of bacterial RNA polymerase.* Biomolecules, 2015. **5**(2): p. 848-64.

219. Hirata, A., Klein, B.J., and Murakami, K.S., *The X-ray crystal structure of RNA polymerase from Archaea.* Nature, 2008. **451**(7180): p. 851-4.

220. Korkhin, Y., Unligil, U.M., Littlefield, O., Nelson, P.J., Stuart, D.I., Sigler, P.B., Bell, S.D., and Abrescia, N.G., *Evolution of complex RNA polymerases: the complete archaeal RNA polymerase structure.* Plos Biology, 2009. **7**(5): p. e1000102.

221. Jun, S.H., Hirata, A., Kanai, T., Santangelo, T.J., Imanaka, T., and Murakami, K.S., *The X-ray crystal structure of the euryarchaeal RNA polymerase in an open-clamp configuration.* Nat Commun, 2014. **5**: p. 5132.

222. Cramer, P., Bushnell, D.A., Fu, J., Gnatt, A.L., Maier-Davis, B., Thompson, N.E., Burgess, R.R., Edwards, A.M., David, P.R., and Kornberg, R.D., *Architecture of RNA polymerase II and implications for the transcription mechanism.* Science, 2000. **288**: p. 640-649.

223. Cramer, P., Bushnell, D.A., and Kornberg, R.D., *Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution.* Science, 2001. **292**(5523): p. 1863-76.

224. Gnatt, A.L., Cramer, P., Fu, J., Bushnell, D.A., and Kornberg, R.D., *Structural basis of transcription: an RNA polymerase II elongation complex at 3.3 A resolution.* Science, 2001. **292**(5523): p. 1876-82.

225. Armache, K.J., Kettenberger, H., and Cramer, P., *Architecture of initiation-competent 12-subunit RNA polymerase II.* Proc Natl Acad Sci U S A, 2003. **100**(12): p. 6964-8.

226. Bernecky, C., Herzog, F., Baumeister, W., Plitzko, J.M., and Cramer, P., *Structure of transcribing mammalian RNA polymerase II.* Nature, 2016. **529**(7587): p. 551-4.

227. Kettenberger, H., Armache, K.J., and Cramer, P., *Complete RNA polymerase II elongation complex structure and its interactions with NTP and TFIIS.* Molecular Cell, 2004. **16**(6): p. 955-65.

228. Spahr, H., Calero, G., Bushnell, D.A., and Kornberg, R.D., *Schizosacharomyces pombe RNA polymerase II at 3.6-A resolution.* Proc Natl Acad Sci U S A, 2009. **106**(23): p. 9185-90.

229. Cheung, A.C. and Cramer, P., *A movie of RNA polymerase II transcription.* Cell, 2012. **149**(7): p. 1431-7.

230. Engel, C., Sainsbury, S., Cheung, A.C., Kostrewa, D., and Cramer, P., *RNA polymerase I structure and transcription regulation.* Nature, 2013. **502**(7473): p. 650-5.

231. Fernandez-Tornero, C., Moreno-Morcillo, M., Rashid, U.J., Taylor, N.M., Ruiz, F.M., Gruene, T., Legrand, P., Steuerwald, U., and Muller, C.W., *Crystal structure of the 14-subunit RNA polymerase I.* Nature, 2013. **502**(7473): p. 644-9.

232. Neyer, S., Kunz, M., Geiss, C., Hantsche, M., Hodirnau, V.V., Seybert, A., Engel, C., Scheffer, M.P., Cramer, P., and Frangakis, A.S., *Structure of RNA polymerase I transcribing ribosomal DNA genes.* Nature, 2016.

233. Hoffmann, N.A., Jakobi, A.J., Moreno-Morcillo, M., Glatt, S., Kosinski, J., Hagen, W.J., Sachse, C., and Muller, C.W., *Molecular structures of unbound and transcribing RNA polymerase III.* Nature, 2015. **528**(7581): p. 231-6.

234. Bushnell, D.A., Cramer, P., and Kornberg, R.D., *Structural basis of transcription: alpha-amanitin-RNA polymerase II cocrystal at 2.8 A resolution.* Proc Natl Acad Sci U S A, 2002. **99**(3): p. 1218-22.

235. Bushnell, D.A. and Kornberg, R.D., *Complete, 12-subunit RNA polymerase II at 4.1-Å resolution: Implications for the initiation of transcription.* Proc Natl Acad Sci U S A, 2003. **100**: p. 6969-6973.

236. Bushnell, D.A., Westover, K.D., Davis, R.E., and Kornberg, R.D., *Structural basis of transcription: an RNA polymerase II-TFIIB cocrystal at 4.5 Angstroms.* Science, 2004. **303**(5660): p. 983-8.

237. Westover, K.D., Bushnell, D.A., and Kornberg, R.D., *Structural basis of transcription: nucleotide selection by rotation in the RNA polymerase II active center.* Cell, 2004. **119**(4): p. 481-9.

238. Westover, K.D., Bushnell, D.A., and Kornberg, R.D., *Structural basis of transcription: separation of RNA from DNA by RNA polymerase II.* Science, 2004. **303**(5660): p. 1014-6.

239. Wang, D., Bushnell, D.A., Westover, K.D., Kaplan, C.D., and Kornberg, R.D., *Structural basis of transcription: role of the trigger loop in substrate specificity and catalysis.* Cell, 2006. **127**(5): p. 941-54.

240. Brueckner, F. and Cramer, P., *Structural basis of transcription inhibition by alpha-amanitin and implications for RNA polymerase II translocation.* Nat Struct Mol Biol, 2008. **15**(8): p. 811-8.

241. Kaplan, C.D., Larsson, K.M., and Kornberg, R.D., *The RNA polymerase II trigger loop functions in substrate selection and is directly targeted by alpha-amanitin.* Molecular Cell, 2008. **30**(5): p. 547-56.

242. Wang, D., Bushnell, D.A., Huang, X., Westover, K.D., Levitt, M., and Kornberg, R.D., *Structural basis of transcription: backtracked RNA polymerase II at 3.4 angstrom resolution.* Science, 2009. **324**(5931): p. 1203-6.

243. Liu, X., Bushnell, D.A., Wang, D., Calero, G., and Kornberg, R.D., *Structure of an RNA polymerase II-TFIIB complex and the transcription initiation mechanism.* Science, 2010. **327**(5962): p. 206-9.

244. Liu, X., Bushnell, D.A., Silva, D.A., Huang, X., and Kornberg, R.D., *Initiation complex structure and promoter proofreading.* Science, 2011. **333**(6042): p. 633-7.

245. Murakami, K., Elmlund, H., Kalisman, N., Bushnell, D.A., Adams, C.M., Azubel, M., Elmlund, D., Levi-Kalisman, Y., Liu, X., Gibbons, B.J., Levitt, M., and Kornberg, R.D., *Architecture of an RNA polymerase II transcription pre-initiation complex.* Science, 2013. **342**(6159): p. 1238724.

246. Robinson, P.J., Trnka, M.J., Bushnell, D.A., Davis, R.E., Mattei, P.J., Burlingame, A.L., and Kornberg, R.D., *Structure of a Complete Mediator-RNA Polymerase II Pre-Initiation Complex.* Cell, 2016. **166**(6): p. 1411-1422 e16.

247. Cheung, A.C. and Cramer, P., *Structural basis of RNA polymerase II backtracking, arrest and reactivation.* Nature, 2011. **471**(7337): p. 249-53.

248. He, Y., Yan, C., Fang, J., Inouye, C., Tjian, R., Ivanov, I., and Nogales, E., *Near-atomic resolution visualization of human transcription promoter opening.* Nature, 2016. **533**(7603): p. 359-65.
249. Sainsbury, S., Niesser, J., and Cramer, P., *Structure and function of the initially transcribing RNA polymerase II-TFIIB complex.* Nature, 2013. **493**(7432): p. 437-40.
250. He, Y., Fang, J., Taatjes, D.J., and Nogales, E., *Structural visualization of key steps in human transcription initiation.* Nature, 2013. **495**(7442): p. 481-6.
251. Cheung, A.C., Sainsbury, S., and Cramer, P., *Structural basis of initial RNA polymerase II transcription.* Embo Journal, 2011. **30**(23): p. 4755-63.
252. Boyer, M., Madoui, M.A., Gimenez, G., La Scola, B., and Raoult, D., *Phylogenetic and phyletic studies of informational genes in genomes highlight existence of a 4 domain of life including giant viruses.* PLoS ONE, 2010. **5**(12): p. e15530.
253. Darst, S.A., *Bacterial RNA polymerase.* Curr Opin Struct Biol, 2001. **11**(2): p. 155-62.
254. Cramer, P., et al., *Structure of eukaryotic RNA polymerases.* Annu Rev Biophys, 2008. **37**: p. 337-52.
255. Sampath, V. and Sadhale, P., *Rpb4 and Rpb7: a sub-complex integral to multi-subunit RNA polymerases performs a multitude of functions.* IUBMB Life, 2005. **57**(2): p. 93-102.
256. Grunberg, S. and Hahn, S., *Structural insights into transcription initiation by RNA polymerase II.* Trends in Biochemical Sciences, 2013. **38**(12): p. 603-11.
257. Kostrewa, D., Zeller, M.E., Armache, K.J., Seizl, M., Leike, K., Thomm, M., and Cramer, P., *RNA polymerase II-TFIIB structure and mechanism of transcription initiation.* Nature, 2009. **462**(7271): p. 323-30.
258. Chen, H.T., Warfield, L., and Hahn, S., *The positions of TFIIF and TFIIE in the RNA polymerase II transcription preinitiation complex.* Nat Struct Mol Biol, 2007. **14**(8): p. 696-703.
259. Vannini, A. and Cramer, P., *Conservation between the RNA polymerase I, II, and III transcription initiation machineries.* Molecular Cell, 2012. **45**(4): p. 439-46.
260. Shaevitz, J.W., Abbondanzieri, E.A., Landick, R., and Block, S.M., *Backtracking by single RNA polymerase molecules observed at near-base-pair resolution.* Nature, 2003. **426**(6967): p. 684-7.
261. Nudler, E., *RNA polymerase backtracking in gene regulation and genome instability.* Cell, 2012. **149**(7): p. 1438-45.
262. Mahat, D.B., Salamanca, H.H., Duarte, F.M., Danko, C.G., and Lis, J.T., *Mammalian Heat Shock Response and Mechanisms Underlying Its Genome-wide Transcriptional Regulation.* Molecular Cell, 2016. **62**(1): p. 63-78.
263. Svejstrup, J.Q., *Contending with transcriptional arrest during RNAPII transcript elongation.* Trends in Biochemical Sciences, 2007. **32**(4): p. 165-71.
264. Lisica, A., Engel, C., Jahnel, M., Roldan, E., Galburt, E.A., Cramer, P., and Grill, S.W., *Mechanisms of backtrack recovery by RNA polymerases I and II.* Proc Natl Acad Sci U S A, 2016. **113**(11): p. 2946-51.
265. Schweikhard, V., Meng, C., Murakami, K., Kaplan, C.D., Kornberg, R.D., and Block, S.M., *Transcription factors TFIIF and TFIIS promote transcript elongation by RNA polymerase II by synergistic and independent mechanisms.* Proc Natl Acad Sci U S A, 2014. **111**(18): p. 6642-7.

266. Fish, R.N. and Kane, C.M., *Promoting elongation with transcript cleavage stimulatory factors.* Biochimica Et Biophysica Acta, 2002. **1577**(2): p. 287-307.

267. Wind, M. and Reines, D., *Transcription elongation factor SII.* Bioessays, 2000. **22**(4): p. 327-36.

268. Kim, B., Nesvizhskii, A.I., Rani, P.G., Hahn, S., Aebersold, R., and Ranish, J.A., *The transcription elongation factor TFIIS is a component of RNA polymerase II preinitiation complexes.* Proc Natl Acad Sci U S A, 2007. **104**(41): p. 16068-73.

269. Awrey, D.E., Shimasaki, N., Koth, C., Weilbaecher, R., Olmsted, V., Kazanis, S., Shan, X., Arellano, J., Arrowsmith, C.H., Kane, C.M., and Edwards, A.M., *Yeast transcript elongation factor (TFIIS), structure and function. II: RNA polymerase binding, transcript cleavage, and read-through.* J Biol Chem, 1998. **273**(35): p. 22595-605.

270. Woychik, N.A., Lane, W.S., and Young, R.A., *Yeast RNA polymerase II subunit RPB9 is essential for growth at temperature extremes.* J Biol Chem, 1991. **266**(28): p. 19053-5.

271. Ruan, W., Lehmann, E., Thomm, M., Kostrewa, D., and Cramer, P., *Evolution of two modes of intrinsic RNA polymerase transcript cleavage.* J Biol Chem, 2011. **286**(21): p. 18701-7.

272. Walmacq, C., Kireeva, M.L., Irvin, J., Nedialkov, Y., Lubkowska, L., Malagon, F., Strathern, J.N., and Kashlev, M., *Rpb9 subunit controls transcription fidelity by delaying NTP sequestration in RNA polymerase II.* J Biol Chem, 2009. **284**(29): p. 19601-12.

273. Cabart, P., Jin, H., Li, L., and Kaplan, C.D., *Activation and reactivation of the RNA polymerase II trigger loop for intrinsic RNA cleavage and catalysis.* Transcription, 2014. **5**(3): p. e28869.

274. Nesser, N.K., Peterson, D.O., and Hawley, D.K., *RNA polymerase II subunit Rpb9 is important for transcriptional fidelity in vivo.* Proc Natl Acad Sci U S A, 2006. **103**(9): p. 3268-73.

275. Koyama, H., Ueda, T., Ito, T., and Sekimizu, K., *Novel RNA polymerase II mutation suppresses transcriptional fidelity and oxidative stress sensitivity in rpb9Delta yeast.* Genes Cells, 2010. **15**(2): p. 151-9.

276. Ghavi-Helm, Y., Michaut, M., Acker, J., Aude, J.C., Thuriaux, P., Werner, M., and Soutourina, J., *Genome-wide location analysis reveals a role of TFIIS in RNA polymerase III transcription.* Genes Dev, 2008. **22**(14): p. 1934-47.

277. Tschochner, H., *A novel RNA polymerase I-dependent RNase activity that shortens nascent transcripts from the 3' end.* Proc Natl Acad Sci U S A, 1996. **93**(23): p. 12914-9.

278. Kuhn, C.D., Geiger, S.R., Baumli, S., Gartmann, M., Gerber, J., Jennebach, S., Mielke, T., Tschochner, H., Beckmann, R., and Cramer, P., *Functional architecture of RNA polymerase I.* Cell, 2007. **131**(7): p. 1260-72.

279. Kettenberger, H., Armache, K.J., and Cramer, P., *Architecture of the RNA polymerase II-TFIIS complex and implications for mRNA cleavage.* Cell, 2003. **114**(3): p. 347-57.

280. Whitehall, S.K., Bardeleben, C., and Kassavetis, G.A., *Hydrolytic cleavage of nascent RNA in RNA polymerase III ternary transcription complexes.* J Biol Chem, 1994. **269**(3): p. 2299-306.

281. Chedin, S., Riva, M., Schultz, P., Sentenac, A., and Carles, C., *The RNA cleavage activity of RNA polymerase III is mediated by an essential TFIIS-like subunit and is important for transcription termination.* Genes Dev, 1998. **12**(24): p. 3857-71.

282. Alic, N., Ayoub, N., Landrieux, E., Favry, E., Baudouin-Cornu, P., Riva, M., and Carles, C., *Selectivity and proofreading both contribute significantly to the fidelity of RNA polymerase III transcription.* Proc Natl Acad Sci U S A, 2007. **104**(25): p. 10400-5.

283. Ishibashi, T., Dangkulwanich, M., Coello, Y., Lionberger, T.A., Lubkowska, L., Ponticelli, A.S., Kashlev, M., and Bustamante, C., *Transcription factors IIS and IIF enhance transcription efficiency by differentially modifying RNA polymerase pausing dynamics.* Proc Natl Acad Sci U S A, 2014. **111**(9): p. 3419-24.

284. Prather, D.M., Larschan, E., and Winston, F., *Evidence that the elongation factor TFIIS plays a role in transcription initiation at GAL1 in Saccharomyces cerevisiae.* Mol Cell Biol, 2005. **25**(7): p. 2650-9.

285. Dutta, A., Babbarwal, V., Fu, J., Brunke-Reese, D., Libert, D.M., Willis, J., and Reese, J.C., *Ccr4-Not and TFIIS Function Cooperatively To Rescue Arrested RNA Polymerase II.* Mol Cell Biol, 2015. **35**(11): p. 1915-25.

286. Kim, J., Guermah, M., and Roeder, R.G., *The human PAF1 complex acts in chromatin transcription elongation both independently and cooperatively with SII/TFIIS.* Cell, 2010. **140**(4): p. 491-503.

287. Palangat, M., Renner, D.B., Price, D.H., and Landick, R., *A negative elongation factor for human RNA polymerase II inhibits the anti-arrest transcript-cleavage factor TFIIS.* Proc Natl Acad Sci U S A, 2005. **102**(42): p. 15036-41.

288. Zhang, C., Yan, H., and Burton, Z.F., *Combinatorial control of human RNA polymerase II (RNAP II) pausing and transcript cleavage by transcription factor IIF, hepatitis delta antigen, and stimulatory factor II.* J Biol Chem, 2003. **278**(50): p. 50101-11.

289. Elmendorf, B.J., Shilatifard, A., Yan, Q., Conaway, J.W., and Conaway, R.C., *Transcription factors TFIIF, ELL, and Elongin negatively regulate SII-induced nascent transcript cleavage by non-arrested RNA polymerase II elongation intermediates.* J Biol Chem, 2001. **276**(25): p. 23109-14.

290. Labhart, P. and Morgan, G.T., *Identification of novel genes encoding transcription elongation factor TFIIS (TCEA) in vertebrates: conservation of three distinct TFIIS isoforms in frog, mouse, and human.* Genomics, 1998. **52**(3): p. 278-88.

291. Liao, J.M., Cao, B., Deng, J., Zhou, X., Strong, M., Zeng, S., Xiong, J., Flemington, E., and Lu, H., *TFIIS.h, a new target of p53, regulates transcription efficiency of pro-apoptotic bax gene.* Sci Rep, 2016. **6**: p. 23542.

292. Bell, S.D. and Jackson, S.P., *Transcription and translation in Archaea: a mosaic of eukaryal and bacterial features.* Trends Microbiol, 1998. **6**(6): p. 222-8.

293. Langer, D., Hain, J., Thuriaux, P., and Zillig, W., *Transcription in archaea: similarity to that in eucarya.* Proc Natl Acad Sci U S A, 1995. **92**(13): p. 5768-72.

294. Zillig, W., Stetter, K.O., and Tobien, M., *DNA-dependent RNA polymerase from Halobacterium halobium.* European Journal of Biochemistry, 1978. **91**(1): p. 193-9.

295. Brinkman, A.B., Ettema, T.J., de Vos, W.M., and van der Oost, J., *The Lrp family of transcriptional regulators.* Molecular Microbiology, 2003. **48**(2): p. 287-94.

296. Ouhammouch, M., *Transcriptional regulation in Archaea.* Curr Opin Genet Dev, 2004. **14**(2): p. 133-8.

297. Huet, J., Schnabel, R., Sentenac, A., and Zillig, W., *Archaebacteria and eukaryotes possess DNA-dependent RNA polymerases of a common type.* Embo Journal, 1983. **2**(8): p. 1291-4.

298. Werner, F., *Structure and function of archaeal RNA polymerases.* Molecular Microbiology, 2007. **65**(6): p. 1395-404.

299. Wojtas, M.N., Mogni, M., Millet, O., Bell, S.D., and Abrescia, N.G., *Structural and functional analyses of the interaction of archaeal RNA polymerase with DNA.* Nucleic Acids Res, 2012. **40**(19): p. 9941-52.

300. Kwapisz, M., Beckouet, F., and Thuriaux, P., *Early evolution of eukaryotic DNA-dependent RNA polymerases.* Trends Genet, 2008. **24**(5): p. 211-5.

301. Wojtas, M., Peralta, B., Ondiviela, M., Mogni, M., Bell, S.D., and Abrescia, N.G., *Archaeal RNA polymerase: the influence of the protruding stalk in crystal packing and preliminary biophysical analysis of the Rpo13 subunit.* Biochem Soc Trans, 2011. **39**(1): p. 25-30.

302. Sheppard, C., Blombach, F., Belsom, A., Schulz, S., Daviter, T., Smollett, K., Mahieu, E., Erdmann, S., Tinnefeld, P., Garrett, R., Grohmann, D., Rappsilber, J., and Werner, F., *Repression of RNA polymerase by the archaeo-viral regulator ORF145/RIP.* Nat Commun, 2016. **7**: p. 13595.

303. Yutin, N. and Koonin, E.V., *Hidden evolutionary complexity of Nucleo-Cytoplasmic Large DNA viruses of eukaryotes.* Virol J, 2012. **9**: p. 161.

304. Colson, P., De Lamballerie, X., Yutin, N., Asgari, S., Bigot, Y., Bideshi, D.K., Cheng, X.W., Federici, B.A., Van Etten, J.L., Koonin, E.V., La Scola, B., and Raoult, D., *"Megavirales", a proposed new order for eukaryotic nucleocytoplasmic large DNA viruses.* Arch Virol, 2013. **158**(12): p. 2517-21.

305. Raoult, D. and Forterre, P., *Redefining viruses: lessons from Mimivirus.* Nat Rev Microbiol, 2008. **6**(4): p. 315-9.

306. Abergel, C., Legendre, M., and Claverie, J.M., *The rapidly expanding universe of giant viruses: Mimivirus, Pandoravirus, Pithovirus and Mollivirus.* FEMS Microbiol Rev, 2015. **39**(6): p. 779-96.

307. Benamar, S., Reteno, D.G., Bandaly, V., Labas, N., Raoult, D., and La Scola, B., *Faustoviruses: Comparative Genomics of New Megavirales Family Members.* Front Microbiol, 2016. **7**: p. 3.

308. Reteno, D.G., Benamar, S., Khalil, J.B., Andreani, J., Armstrong, N., Klose, T., Rossmann, M., Colson, P., Raoult, D., and La Scola, B., *Faustovirus, an asfarvirus-related new lineage of giant viruses infecting amoebae.* J Virol, 2015. **89**(13): p. 6585-94.

309. Baroudy, B.M. and Moss, B., *Purification and characterization of a DNA-dependent RNA polymerase from vaccinia virions.* J. Biol. Chem., 1980. **255**: p. 4372-4380.

310. Spencer, E., Shuman, S., and Hurwitz, J., *Purification and properties of vaccinia virus DNA-dependent RNA polymerase.* Journal of Biological Chemistry, 1980. **255**(11): p. 5388-5395.

311. Ahn, B.-Y., Gershon, P.D., and Moss, B., *RNA polymerase-associated protein Rap94 confers promoter specificity for initiating transcription of vaccinia virus early stage genes.* J. Biol. Chem., 1994. **269**: p. 7552-7557.

312. Ahn, B.Y. and Moss, B., *RNA polymerase-associated transcription specificity factor encoded by vaccinia virus.* Proc Natl Acad Sci U S A, 1992. **89**(8): p. 3536-40.

313. Deng, L. and Shuman, S., *A role for the H4 subunit of vaccinia RNA polymerase in transcription initiation at a viral early promoter.* J Biol Chem, 1994. **269**(19): p. 14323-8.

314. Broyles, S.S. and Moss, B., *Homology between RNA polymerases of poxviruses, prokaryotes, and eukaryotes: Nucleotide sequence and transcriptional analysis of*

vaccinia virus genes encoding 147-kDa and 22-kDa subunits. Proc. Natl. Acad. Sci. USA, 1986. **83**: p. 3141-3145.

315. Amegadzie, B.Y., Holmes, M.H., Cole, N.B., Jones, E.V., Earl, P.L., and Moss, B., *Identification, sequence, and expression of the gene encoding the second-largest subunit of the vaccinia virus DNA-dependent RNA polymerase.* Virology, 1991. **180**(1): p. 88-98.
316. Patel, D. and Pickup, D.J., *The second-largest subunit of the poxvirus RNA polymerase is similar to the corresponding subunits of procaryotic and eucaryotic RNA polymerases.* J. Virol., 1989. **63**: p. 1076-1086.
317. Amegadzie, B.Y., Ahn, B.Y., and Moss, B., *Characterization of a 7-kilodalton subunit of vaccinia virus DNA-dependent RNA polymerase with structural similarities to the smallest subunit of eukaryotic RNA polymerase II.* Journal of Virology, 1992. **66**(5): p. 3003-3010.
318. Ahn, B.-Y., Gershon, P.D., Jones, E.V., and Moss, B., *Identification of rpo30, a vaccinia virus RNA polymerase gene with structural similarity to a eucaryotic transcription elongation factor.* Mol. Cell Biol., 1990. **10**(10): p. 5433-5441.
319. Knutson, B.A. and Broyles, S.S., *Expansion of poxvirus RNA polymerase subunits sharing homology with corresponding subunits of RNA polymerase II.* Virus Genes, 2008. **36**(2): p. 307-11.
320. Plaschka, C., Hantsche, M., Dienemann, C., Burzinski, C., Plitzko, J., and Cramer, P., *Transcription initiation complex structures elucidate DNA opening.* Nature, 2016. **533**(7603): p. 353-8.
321. Eaton, H.E., Ring, B.A., and Brunetti, C.R., *The genomic diversity and phylogenetic relationship in the family iridoviridae.* Viruses, 2010. **2**(7): p. 1458-75.
322. Tan, W.G., Barkman, T.J., Gregory Chinchar, V., and Essani, K., *Comparative genomic analyses of frog virus 3, type species of the genus Ranavirus (family Iridoviridae).* Virology, 2004. **323**(1): p. 70-84.
323. Goorha, R., *Frog virus 3 requires RNA polymerase II for its replication.* J Virol, 1981. **37**(1): p. 496-9.
324. Chinchar, V.G., Hyatt, A., Miyazaki, T., and Williams, T., *Family Iridoviridae: poor viral relations no longer.* Curr Top Microbiol Immunol, 2009. **328**: p. 123-70.
325. Willis, D.B. and Granoff, A., *Macromolecular synthesis in cells infected by frog virus 3. IX. Two temporal classes of early viral RNA.* Virology, 1978. **86**(2): p. 443-53.
326. Goorha, R., Murti, G., Granoff, A., and Tirey, R., *Macromolecular synthesis in cells infected by frog virus 3. VIII. The nucleus is a site of frog virus 3 DNA and RNA synthesis.* Virology, 1978. **84**(1): p. 32-50.
327. Sample, R., Bryan, L., Long, S., Majji, S., Hoskins, G., Sinning, A., Olivier, J., and Chinchar, V.G., *Inhibition of iridovirus protein synthesis and virus replication by antisense morpholino oligonucleotides targeted to the major capsid protein, the 18 kDa immediate-early protein, and a viral homolog of RNA polymerase II.* Virology, 2007. **358**(2): p. 311-20.
328. Oliveira, G.P., Andrade, A.C., Rodrigues, R.A., Arantes, T.S., Boratto, P.V., Silva, L.K., Dornas, F.P., Trindade, G.S., Drumond, B.P., La Scola, B., Kroon, E.G., and Abrahao, J.S., *Promoter Motifs in NCLDVs: An Evolutionary Perspective.* Viruses, 2017. **9**(1).
329. Rosales, R., Harris, N., Ahn, B.-Y., and Moss, B., *Purification and identification of a vaccinia virus-encoded intermediate stage promoter-specific transcription factor that has*

homology to eukaryotic transcription factor SII (TFIIS) and an additional role as a viral RNA polymerase subunit. J. Biol. Chem., 1994. **269**(19): p. 14260-14267.

330. Hagler, J. and Shuman, S., *Nascent RNA cleavage by purified ternary complexes of vaccinia RNA polymerase.* J Biol Chem, 1993. **268**(3): p. 2166-73.

331. Ngo, T., Mirzakhanyan, Y., Moussatche, N., and Gershon, P.D., *Protein Primary Structure of the Vaccinia Virion at Increased Resolution.* J Virol, 2016. **90**(21): p. 9905-9919.

332. Wild, T. and Cramer, P., *Biogenesis of multisubunit RNA polymerases.* Trends in Biochemical Sciences, 2012. **37**(3): p. 99-105.

333. Schulz, S., Gietl, A., Smollett, K., Tinnefeld, P., Werner, F., and Grohmann, D., *TFE and Spt4/5 open and close the RNA polymerase clamp during the transcription cycle.* Proc Natl Acad Sci U S A, 2016. **113**(13): p. E1816-25.

334. Bourbonnais, Y., Faucher, N., Pallotta, D., and Larouche, C., *Multiple cellular processes affected by the absence of the Rpb4 subunit of RNA polymerase II contribute to the deficiency in the stress response of the yeast rpb4(delta) mutant.* Molecular & General Genetics, 2001. **264**(6): p. 763-72.

335. Kimura, M., Suzuki, H., and Ishihama, A., *Formation of a carboxy-terminal domain phosphatase (Fcp1)/TFIIF/RNA polymerase II (pol II) complex in Schizosaccharomyces pombe involves direct interaction between Fcp1 and the Rpb4 subunit of pol II.* Mol Cell Biol, 2002. **22**(5): p. 1577-88.

336. Pillai, B., Verma, J., Abraham, A., Francis, P., Kumar, Y., Tatu, U., Brahmachari, S.K., and Sadhale, P.P., *Whole genome expression profiles of yeast RNA polymerase II core subunit, Rpb4, in stress and nonstress conditions.* J Biol Chem, 2003. **278**(5): p. 3339-46.

337. Woychik, N.A. and Young, R.A., *RNA polymerase II subunit RPB4 is essential for high- and low-temperature yeast cell growth.* Mol Cell Biol, 1989. **9**(7): p. 2854-9.

338. Engel, C., Plitzko, J., and Cramer, P., *RNA polymerase I-Rrn3 complex at 4.8 A resolution.* Nat Commun, 2016. **7**: p. 12129.

339. Peyroche, G., Milkereit, P., Bischler, N., Tschochner, H., Schultz, P., Sentenac, A., Carles, C., and Riva, M., *The recruitment of RNA polymerase I on rDNA is mediated by the interaction of the A43 subunit with Rrn3.* Embo Journal, 2000. **19**(20): p. 5473-82.

340. Blattner, C., Jennebach, S., Herzog, F., Mayer, A., Cheung, A.C., Witte, G., Lorenzen, K., Hopfner, K.P., Heck, A.J., Aebersold, R., and Cramer, P., *Molecular basis of Rrn3-regulated RNA polymerase I initiation and cell growth.* Genes Dev, 2011. **25**(19): p. 2093-105.

341. De Angelis, R., Iezzi, S., Bruno, T., Corbi, N., Di Padova, M., Floridi, A., Fanciulli, M., and Passananti, C., *Functional interaction of the subunit 3 of RNA polymerase II (RPB3) with transcription factor-4 (ATF4).* Febs Letters, 2003. **547**(1-3): p. 15-9.

342. Reich, C., Zeller, M., Milkereit, P., Hausner, W., Cramer, P., Tschochner, H., and Thomm, M., *The archaeal RNA polymerase subunit P and the eukaryotic polymerase subunit Rpb12 are interchangeable in vivo and in vitro.* Molecular Microbiology, 2009. **71**(4): p. 989-1002.

343. Sommer, B., Waege, I., Pollmann, D., Seitz, T., Thomm, M., Sterner, R., and Hausner, W., *Activation of a chimeric Rpb5/RpoH subunit using library selection.* PLoS ONE, 2014. **9**(1): p. e87485.

344. Wei, W., Dorjsuren, D., Lin, Y., Qin, W., Nomura, T., Hayashi, N., and Murakami, S., *Direct interaction between the subunit RAP30 of transcription factor IIF (TFIIF) and*

RNA polymerase subunit 5, which contributes to the association between TFIIF and RNA polymerase II. J Biol Chem, 2001. **276**(15): p. 12266-73.

345. Lin, Y., Nomura, T., Cheong, J., Dorjsuren, D., Iida, K., and Murakami, S., *Hepatitis B virus X protein is a transcriptional modulator that communicates with transcription factor IIB and the RNA polymerase II subunit 5.* J Biol Chem, 1997. **272**(11): p. 7132-9.

346. Cheong, J.H., Yi, M., Lin, Y., and Murakami, S., *Human RPB5, a subunit shared by eukaryotic nuclear RNA polymerases, binds human hepatitis B virus X protein and may play a role in X transactivation.* Embo Journal, 1995. **14**(1): p. 143-50.

347. Zimmermann, L., Stephens, A., Nam, S.Z., Rau, D., Kubler, J., Lozajic, M., Gabler, F., Soding, J., Lupas, A.N., and Alva, V., *A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core.* J Mol Biol, 2018. **430**(15): p. 2237-2243.

348. Gabler, F., Nam, S.Z., Till, S., Mirdita, M., Steinegger, M., Soding, J., Lupas, A.N., and Alva, V., *Protein Sequence Analysis Using the MPI Bioinformatics Toolkit.* Curr Protoc Bioinformatics, 2020. **72**(1): p. e108.

349. Boratyn, G.M., et al., *BLAST: a more efficient report with usability improvements.* Nucleic Acids Res, 2013. **41**(Web Server issue): p. W29-33.

350. Iyer, L.M., Aravind, L., and Koonin, E.V., *Common origin of four diverse families of large eukaryotic DNA viruses.* J Virol, 2001. **75**(23): p. 11720-34.

351. Hodel, A.E., Gershon, P.D., Shi, X., and Quiocho, F.A., *The 1.85Å structure of vaccinia protein VP39: A bifunctional enzyme that participates in the modification of both mRNA ends.* Cell, 1996. **85**: p. 247-256.

352. Mistry, J., Bateman, A., and Finn, R.D., *Predicting active site residue annotations in the Pfam database.* BMC Bioinformatics, 2007. **8**: p. 298.

353. Madeira, F., Park, Y.M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A.R.N., Potter, S.C., Finn, R.D., and Lopez, R., *The EMBL-EBI search and sequence analysis tools APIs in 2019.* Nucleic Acids Res, 2019. **47**(W1): p. W636-W641.

354. Mitchell, A.L., et al., *InterPro in 2019: improving coverage, classification and access to protein sequence annotations.* Nucleic Acids Res, 2019. **47**(D1): p. D351-D360.

355. Iyer, L.M., Koonin, E.V., Leipe, D.D., and Aravind, L., *Origin and evolution of the archaeo-eukaryotic primase superfamily and related palm-domain proteins: structural insights and new members.* Nucleic Acids Res, 2005. **33**(12): p. 3875-96.

356. Hildebrand, A., Remmert, M., Biegert, A., and Soding, J., *Fast and accurate automatic structure prediction with HHpred.* Proteins-Structure Function and Bioinformatics, 2009. **77 Suppl 9**: p. 128-32.

357. Montelione, G.T., *The Protein Structure Initiative: achievements and visions for the future.* F1000 Biol Rep, 2012. **4**: p. 7.

358. Lobb, B., Kurtz, D.A., Moreno-Hagelsieb, G., and Doxey, A.C., *Remote homology and the functions of metagenomic dark matter.* Front Genet, 2015. **6**: p. 234.

359. Nordstrom, K.J., Sallman Almen, M., Edstam, M.M., Fredriksson, R., and Schioth, H.B., *Independent HHsearch, Needleman--Wunsch-based, and motif analyses reveal the overall hierarchy for most of the G protein-coupled receptor families.* Mol Biol Evol, 2011. **28**(9): p. 2471-80.

360. O'Day, D.H., Suhre, K., Myre, M.A., Chatterjee-Chakraborty, M., and Chavez, S.E., *Isolation, characterization, and bioinformatic analysis of calmodulin-binding protein cmbB reveals a novel tandem IP22 repeat common to many Dictyostelium and Mimivirus proteins.* Biochem Biophys Res Commun, 2006. **346**(3): p. 879-88.

361. Fidler, D.R., Murphy, S.E., Courtis, K., Antonoudiou, P., El-Tohamy, R., Ient, J., and Levine, T.P., *Using HHsearch to tackle proteins of unknown function: A pilot study with PH domains.* Traffic, 2016. **17**(11): p. 1214-1226.

362. Soding, J., Biegert, A., and Lupas, A.N., *The HHpred interactive server for protein homology detection and structure prediction.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W244-8.

363. Soding, J., *Protein homology detection by HMM-HMM comparison.* Bioinformatics, 2005. **21**(7): p. 951-60.

364. Koonin, E.V. and Yutin, N., *Origin and evolution of eukaryotic large nucleo-cytoplasmic DNA viruses.* Intervirology, 2010. **53**(5): p. 284-92.

365. Yutin, N., Wolf, Y.I., Raoult, D., and Koonin, E.V., *Eukaryotic large nucleo-cytoplasmic DNA viruses: clusters of orthologous genes and reconstruction of viral genome evolution.* Virol J, 2009. **6**: p. 223.

366. Yutin, N., Wolf, Y.I., and Koonin, E.V., *Origin of giant viruses from smaller DNA viruses not from a fourth domain of cellular life.* Virology, 2014. **466-467**: p. 38-52.

367. Schulz, F., Yutin, N., Ivanova, N.N., Ortega, D.R., Lee, T.K., Vierheilig, J., Daims, H., Horn, M., Wagner, M., Jensen, G.J., Kyrpides, N.C., Koonin, E.V., and Woyke, T., *Giant viruses with an expanded complement of translation system components.* Science, 2017. **356**(6333): p. 82-85.

368. Urbaneja, M.A., McGrath, C.F., Kane, B.P., Henderson, L.E., and Casas-Finet, J.R., *Nucleic acid binding properties of the simian immunodeficiency virus nucleocapsid protein NCp8.* J Biol Chem, 2000. **275**(14): p. 10394-404.

369. Shepard, D.A., Ehnstrom, J.G., Skinner, P.J., and Schiff, L.A., *Mutations in the zinc-binding motif of the reovirus capsid protein delta 3 eliminate its ability to associate with capsid protein mu 1.* J Virol, 1996. **70**(3): p. 2065-8.

370. Lemay, G. and Danis, C., *Reovirus lambda 1 protein: affinity for double-stranded nucleic acids by a small amino-terminal region of the protein independent from the zinc finger motif.* J Gen Virol, 1994. **75 (Pt 11)**: p. 3261-6.

371. Olland, A.M., Jane-Valbuena, J., Schiff, L.A., Nibert, M.L., and Harrison, S.C., *Structure of the reovirus outer capsid and dsRNA-binding protein sigma3 at 1.8 A resolution.* Embo Journal, 2001. **20**(5): p. 979-89.

372. Bartlett, J.A. and Joklik, W.K., *The sequence of the reovirus serotype 3 L3 genome segment which encodes the major core protein lambda 1.* Virology, 1988. **167**(1): p. 31-7.

373. Iyer, L.M., Koonin, E.V., and Aravind, L., *Extensive domain shuffling in transcription regulators of DNA viruses and implications for the origin of fungal APSES transcription factors.* Genome Biol, 2002. **3**(3): p. RESEARCH0012.

374. Huang, J., et al., *The poxvirus p28 virulence factor is an E3 ubiquitin ligase.* J Biol Chem, 2004. **279**(52): p. 54110-6.

375. Nerenberg, B.T., Taylor, J., Bartee, E., Gouveia, K., Barry, M., and Fruh, K., *The poxviral RING protein p28 is a ubiquitin ligase that targets ubiquitin to viral replication factories.* J Virol, 2005. **79**(1): p. 597-601.

376. Briand, J.F., Navarro, F., Rematier, P., Boschiero, C., Labarre, S., Werner, M., Shpakovski, G.V., and Thuriaux, P., *Partners of Rpb8p, a small subunit shared by yeast RNA polymerases I, II and III.* Mol Cell Biol, 2001. **21**(17): p. 6056-65.

377. Koonin, E.V., Makarova, K.S., and Elkins, J.G., *Orthologs of the small RPB8 subunit of the eukaryotic RNA polymerases are conserved in hyperthermophilic Crenarchaeota and "Korarchaeota".* Biol Direct, 2007. **2**: p. 38.

378. Suhre, K., Audic, S., and Claverie, J.M., *Mimivirus gene promoters exhibit an unprecedented conservation among all eukaryotes.* Proc Natl Acad Sci U S A, 2005. **102**(41): p. 14689-93.

379. Treich, I., Carles, C., Riva, M., and Sentenac, A., *RPC10 encodes a new mini subunit shared by yeast nuclear RNA polymerases.* Gene Expr, 1992. **2**(1): p. 31-7.

380. Satheshkumar, P.S., Olano, L.R., Hammer, C.H., Zhao, M., and Moss, B., *Interactions of the vaccinia virus A19 protein.* J Virol, 2013. **87**(19): p. 10710-20.

381. Stoddard, B.L., *Homing endonuclease structure and function.* Q Rev Biophys, 2005. **38**(1): p. 49-95.

382. Stoddard, B.L., *Homing endonucleases: from microbial genetic invaders to reagents for targeted DNA modification.* Structure, 2011. **19**(1): p. 7-15.

383. Roberts, R.J., et al., *A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes.* Nucleic Acids Res, 2003. **31**(7): p. 1805-12.

384. Horton, J.R., Borgaro, J.G., Griggs, R.M., Quimby, A., Guan, S., Zhang, X., Wilson, G.G., Zheng, Y., Zhu, Z., and Cheng, X., *Structure of 5-hydroxymethylcytosine-specific restriction enzyme, AbaSI, in complex with DNA.* Nucleic Acids Res, 2014. **42**(12): p. 7947-59.

385. Pingoud, A. and Jeltsch, A., *Structure and function of type II restriction endonucleases.* Nucleic Acids Res, 2001. **29**(18): p. 3705-27.

386. Pingoud, A., Fuxreiter, M., Pingoud, V., and Wende, W., *Type II restriction endonucleases: structure and mechanism.* Cell Mol Life Sci, 2005. **62**(6): p. 685-707.

387. Hennecke, F., Kolmar, H., Brundl, K., and Fritz, H.J., *The vsr gene product of E. coli K-12 is a strand- and sequence-specific DNA mismatch endonuclease.* Nature, 1991. **353**(6346): p. 776-8.

388. Macintyre, G., Doiron, K.M., and Cupples, C.G., *The Vsr endonuclease of Escherichia coli: an efficient DNA repair enzyme and a potent mutagen.* Journal of Bacteriology, 1997. **179**(19): p. 6048-52.

389. Agarkova, I.V., Dunigan, D.D., and Van Etten, J.L., *Virion-associated restriction endonucleases of chloroviruses.* J Virol, 2006. **80**(16): p. 8114-23.

390. Xia, Y.N., Burbank, D.E., Uher, L., Rabussay, D., and Van Etten, J.L., *Restriction endonuclease activity induced by PBCV-1 virus infection of a Chlorella-like green alga.* Mol Cell Biol, 1986. **6**(5): p. 1430-9.

391. Zhang, Y., Nelson, M., Nietfeldt, J.W., Burbank, D.E., and Van Etten, J.L., *Characterization of Chlorella virus PBCV-1 CviAII restriction and modification system.* Nucleic Acids Res, 1992. **20**(20): p. 5351-6.

392. Chan, S.H., Zhu, Z., Dunigan, D.D., Van Etten, J.L., and Xu, S.Y., *Cloning of Nt.CviQII nicking endonuclease and its cognate methyltransferase: M.CviQII methylates AG sequences.* Protein Expr Purif, 2006. **49**(1): p. 138-50.

393. Zhang, Y., Nelson, M., Nietfeldt, J., Xia, Y., Burbank, D., Ropp, S., and Van Etten, J.L., *Chlorella virus NY-2A encodes at least 12 DNA endonuclease/methyltransferase genes.* Virology, 1998. **240**(2): p. 366-75.

394. Xia, Y.N., Burbank, D.E., Uher, L., Rabussay, D., and Van Etten, J.L., *IL-3A virus infection of a Chlorella-like green alga induces a DNA restriction endonuclease with novel sequence specificity.* Nucleic Acids Res, 1987. **15**(15): p. 6075-90.

395. Laganeckas, M., Margelevicius, M., and Venclovas, C., *Identification of new homologs of PD-(D/E)XK nucleases by support vector machines trained on data derived from profile-profile alignments.* Nucleic Acids Res, 2011. **39**(4): p. 1187-96.

396. Steczkiewicz, K., Muszewska, A., Knizewski, L., Rychlewski, L., and Ginalski, K., *Sequence, structure and functional diversity of PD-(D/E)XK phosphodiesterase superfamily.* Nucleic Acids Res, 2012. **40**(15): p. 7016-45.

397. Venclovas, C., Timinskas, A., and Siksnys, V., *Five-stranded beta-sheet sandwiched with two alpha-helices: a structural link between restriction endonucleases EcoRI and EcoRV.* Proteins-Structure Function and Bioinformatics, 1994. **20**(3): p. 279-82.

398. Aravind, L., Makarova, K.S., and Koonin, E.V., *SURVEY AND SUMMARY: holliday junction resolvases and related nucleases: identification of new families, phyletic distribution and evolutionary trajectories.* Nucleic Acids Res, 2000. **28**(18): p. 3417-32.

399. Kazrani, A.A., Kowalska, M., Czapinska, H., and Bochtler, M., *Crystal structure of the 5hmC specific endonuclease PvuRts1I.* Nucleic Acids Res, 2014. **42**(9): p. 5929-36.

400. Wang, H., Guan, S., Quimby, A., Cohen-Karni, D., Pradhan, S., Wilson, G., Roberts, R.J., Zhu, Z., and Zheng, Y., *Comparative characterization of the PvuRts1I family of restriction enzymes and their application in mapping genomic 5-hydroxymethylcytosine.* Nucleic Acids Res, 2011. **39**(21): p. 9294-305.

401. Loenen, W.A., Dryden, D.T., Raleigh, E.A., and Wilson, G.G., *Type I restriction enzymes and their relatives.* Nucleic Acids Res, 2014. **42**(1): p. 20-44.

402. Wyszomirski, K.H., Curth, U., Alves, J., Mackeldanz, P., Moncke-Buchner, E., Schutkowski, M., Kruger, D.H., and Reuter, M., *Type III restriction endonuclease EcoP15I is a heterotrimeric complex containing one Res subunit with several DNA-binding regions and ATPase activity.* Nucleic Acids Res, 2012. **40**(8): p. 3610-22.

403. Ishino, S., Skouloubris, S., Kudo, H., l'Hermitte-Stead, C., Es-Sadik, A., Lambry, J.C., Ishino, Y., and Myllykallio, H., *Activation of the mismatch-specific endonuclease EndoMS/NucS by the replication clamp is required for high fidelity DNA replication.* Nucleic Acids Res, 2018. **46**(12): p. 6206-6217.

404. Nakae, S., Hijikata, A., Tsuji, T., Yonezawa, K., Kouyama, K.I., Mayanagi, K., Ishino, S., Ishino, Y., and Shirai, T., *Structure of the EndoMS-DNA Complex as Mismatch Restriction Endonuclease.* Structure, 2016. **24**(11): p. 1960-1971.

405. Ren, B., Kuhn, J., Meslet-Cladiere, L., Briffotaux, J., Norais, C., Lavigne, R., Flament, D., Ladenstein, R., and Myllykallio, H., *Structure and function of a novel endonuclease acting on branched DNA substrates.* Embo Journal, 2009. **28**(16): p. 2479-89.

406. Dunin-Horkawicz, S., Feder, M., and Bujnicki, J.M., *Phylogenomic analysis of the GIY-YIG nuclease superfamily.* BMC Genomics, 2006. **7**: p. 98.

407. Kowalski, J.C., Belfort, M., Stapleton, M.A., Holpert, M., Dansereau, J.T., Pietrokovski, S., Baxter, S.M., and Derbyshire, V., *Configuration of the catalytic GIY-YIG domain of intron endonuclease I-TevI: coincidence of computational and molecular findings.* Nucleic Acids Res, 1999. **27**(10): p. 2115-25.

408. Jurica, M.S. and Stoddard, B.L., *Homing endonucleases: structure, function and evolution.* Cell Mol Life Sci, 1999. **55**(10): p. 1304-26.

409. Shen, B.W., Heiter, D.F., Chan, S.H., Wang, H., Xu, S.Y., Morgan, R.D., Wilson, G.G., and Stoddard, B.L., *Unusual target site disruption by the rare-cutting HNH restriction endonuclease PacI.* Structure, 2010. **18**(6): p. 734-43.

410. Moon, A.F., Midon, M., Meiss, G., Pingoud, A., London, R.E., and Pedersen, L.C., *Structural insights into catalytic and substrate binding mechanisms of the strategic EndA nuclease from Streptococcus pneumoniae.* Nucleic Acids Res, 2011. **39**(7): p. 2943-53.

411. Brinkmann, V., Reichard, U., Goosmann, C., Fauler, B., Uhlemann, Y., Weiss, D.S., Weinrauch, Y., and Zychlinsky, A., *Neutrophil extracellular traps kill bacteria.* Science, 2004. **303**(5663): p. 1532-5.

412. Liu, R., Olano, L.R., Mirzakhanyan, Y., Gershon, P.D., and Moss, B., *Vaccinia Virus Ankyrin-Repeat/F-Box Protein Targets Interferon-Induced IFITs for Proteasomal Degradation.* Cell Rep, 2019. **29**(4): p. 816-828 e6.

413. Shukla, A., Chatterjee, A., and Kondabagil, K., *The number of genes encoding repeat domain-containing proteins positively correlates with genome size in amoebal giant viruses.* Virus Evol, 2018. **4**(1): p. vex039.

414. Aherfi, S., Colson, P., La Scola, B., and Raoult, D., *Giant Viruses of Amoebas: An Update.* Front Microbiol, 2016. **7**: p. 349.

415. Takeshima, H., Komazaki, S., Nishi, M., Iino, M., and Kangawa, K., *Junctophilins: a novel family of junctional membrane complex proteins.* Molecular Cell, 2000. **6**(1): p. 11-22.

416. Jiang, J., Tang, M., Huang, Z., and Chen, L., *Junctophilins emerge as novel therapeutic targets.* J Cell Physiol, 2019. **234**(10): p. 16933-16943.

417. Gubbels, M.J., Vaishnava, S., Boot, N., Dubremetz, J.F., and Striepen, B., *A MORN-repeat protein is a dynamic component of the Toxoplasma gondii cell division apparatus.* Journal of Cell Science, 2006. **119**(Pt 11): p. 2236-45.

418. Morriswood, B. and Schmidt, K., *A MORN Repeat Protein Facilitates Protein Entry into the Flagellar Pocket of Trypanosoma brucei.* Eukaryot Cell, 2015. **14**(11): p. 1081-93.

419. Hatzopoulos, G.N., Erat, M.C., Cutts, E., Rogala, K.B., Slater, L.M., Stansfeld, P.J., and Vakonakis, I., *Structural analysis of the G-box domain of the microcephaly protein CPAP suggests a role in centriole architecture.* Structure, 2013. **21**(11): p. 2069-77.

420. Zheng, X., et al., *Conserved TCP domain of Sas-4/CPAP is essential for pericentriolar material tethering during centrosome biogenesis.* Proc Natl Acad Sci U S A, 2014. **111**(3): p. E354-63.

421. Mercer, A.A., Fleming, S.B., and Ueda, N., *F-box-like domains are present in most poxvirus ankyrin repeat proteins.* Virus Genes, 2005. **31**(2): p. 127-33.

422. Price, C.T., Al-Quadan, T., Santic, M., Jones, S.C., and Abu Kwaik, Y., *Exploitation of conserved eukaryotic host cell farnesylation machinery by an F-box effector of Legionella pneumophila.* J Exp Med, 2010. **207**(8): p. 1713-26.

423. Pellett, P.E. and Roizman, B., *Herpesviridae*, in *Fields Virology*, D.M. Knipe, et al., Editors. 2013, Lippincott Williams & Wilkins: Philadelphia, PA, USA.

424. Cai, W.H., Gu, B., and Person, S., *Role of glycoprotein B of herpes simplex virus type 1 in viral entry and cell fusion.* J Virol, 1988. **62**(8): p. 2596-604.

425. Heldwein, E.E., Lou, H., Bender, F.C., Cohen, G.H., Eisenberg, R.J., and Harrison, S.C., *Crystal structure of glycoprotein B from herpes simplex virus 1.* Science, 2006. **313**(5784): p. 217-20.

426. Mitsuhashi, W., Miyamoto, K., and Wada, S., *The complete genome sequence of the Alphaentomopoxvirus Anomala cuprea entomopoxvirus, including its terminal hairpin loop sequences, suggests a potentially unique mode of apoptosis inhibition and mode of DNA replication.* Virology, 2014. **452-453**: p. 95-116.

427. Cavin, J.F., Dartois, V., and Divies, C., *Gene cloning, transcriptional analysis, purification, and characterization of phenolic acid decarboxylase from Bacillus subtilis.* Appl Environ Microbiol, 1998. **64**(4): p. 1466-71.

428. Govindarajan, R. and Federici, B.A., *Ascovirus infectivity and effects of infection on the growth and development of noctuid larvae.* J Invertebr Pathol, 1990. **56**(3): p. 291-9.

429. Bi, J.L. and Felton, G.W., *Foliar oxidative stress and insect herbivory: Primary compounds, secondary metabolites, and reactive oxygen species as components of induced resistance.* J Chem Ecol, 1995. **21**(10): p. 1511-30.

430. Bhonwong, A., Stout, M.J., Attajarusit, J., and Tantasawat, P., *Defensive role of tomato polyphenol oxidases against cotton bollworm (Helicoverpa armigera) and beet armyworm (Spodoptera exigua).* J Chem Ecol, 2009. **35**(1): p. 28-38.

431. Xia, X., Gurr, G.M., Vasseur, L., Zheng, D., Zhong, H., Qin, B., Lin, J., Wang, Y., Song, F., Li, Y., Lin, H., and You, M., *Metagenomic Sequencing of Diamondback Moth Gut Microbiome Unveils Key Holobiont Adaptations for Herbivory.* Front Microbiol, 2017. **8**: p. 663.

432. Wu, K., Zhang, J., Zhang, Q., Zhu, S., Shao, Q., Clark, K.D., Liu, Y., and Ling, E., *Plant phenolics are detoxified by prophenoloxidase in the insect gut.* Sci Rep, 2015. **5**: p. 16823.

433. Bigot, Y., Rabouille, A., Doury, G., Sizaret, P.Y., Delbost, F., Hamelin, M.H., and Periquet, G., *Biological and molecular features of the relationships between Diadromus pulchellus ascovirus, a parasitoid hymenopteran wasp (Diadromus pulchellus) and its lepidopteran host, Acrolepiopsis assectella.* J Gen Virol, 1997. **78 (Pt 5)**: p. 1149-63.

434. Felton, G.W. and Duffey, S.S., *Inactivation of baculovirus by quinones formed in insect-damaged plant tissues.* J Chem Ecol, 1990. **16**(4): p. 1221-36.

435. Mizuno, C.M., Guyomar, C., Roux, S., Lavigne, R., Rodriguez-Valera, F., Sullivan, M.B., Gillet, R., Forterre, P., and Krupovic, M., *Numerous cultivated and uncultivated viruses encode ribosomal proteins.* Nat Commun, 2019. **10**(1): p. 752.

436. Hurley, J.H. and Emr, S.D., *The ESCRT complexes: structure and mechanism of a membrane-trafficking network.* Annu Rev Biophys Biomol Struct, 2006. **35**: p. 277-98.

437. Martin-Serrano, J., Yarovoy, A., Perez-Caballero, D., and Bieniasz, P.D., *Divergent retroviral late-budding domains recruit vacuolar protein sorting factors by using alternative adaptor proteins.* Proc Natl Acad Sci U S A, 2003. **100**(21): p. 12414-9.

438. Fisher, R.D., Chung, H.Y., Zhai, Q., Robinson, H., Sundquist, W.I., and Hill, C.P., *Structural and biochemical studies of ALIX/AIP1 and its role in retrovirus budding.* Cell, 2007. **128**(5): p. 841-52.

439. Mi, S., Qin, X.W., Lin, Y.F., He, J., Chen, N.N., Liu, C., Weng, S.P., He, J.G., and Guo, C.J., *Budding of Tiger Frog Virus (an Iridovirus) from HepG2 Cells via Three Ways Recruits the ESCRT Pathway.* Sci Rep, 2016. **6**: p. 26581.

440. Howard, J.P., Hutton, J.L., Olson, J.M., and Payne, G.S., *Sla1p serves as the targeting signal recognition factor for NPFX(1,2)D-mediated endocytosis.* J Cell Biol, 2002. **157**(2): p. 315-26.

441. Tan, P.K., Howard, J.P., and Payne, G.S., *The sequence NPFXD defines a new class of endocytosis signal in Saccharomyces cerevisiae.* J Cell Biol, 1996. **135**(6 Pt 2): p. 1789-800.

442. Mahadev, R.K., Di Pietro, S.M., Olson, J.M., Piao, H.L., Payne, G.S., and Overduin, M., *Structure of Sla1p homology domain 1 and interaction with the NPFxD endocytic internalization motif.* Embo Journal, 2007. **26**(7): p. 1963-71.

443. Piao, H.L., Machado, I.M., and Payne, G.S., *NPFXD-mediated endocytosis is required for polarity and function of a yeast cell wall stress sensor.* Molecular Biology of the Cell, 2007. **18**(1): p. 57-65.

444. Hamoen, L.W., Meile, J.C., de Jong, W., Noirot, P., and Errington, J., *SepF, a novel FtsZ-interacting protein required for a late step in cell division.* Molecular Microbiology, 2006. **59**(3): p. 989-99.

445. Duman, R., Ishikawa, S., Celik, I., Strahl, H., Ogasawara, N., Troc, P., Lowe, J., and Hamoen, L.W., *Structural and genetic analyses reveal the protein SepF as a new membrane anchor for the Z ring.* Proc Natl Acad Sci U S A, 2013. **110**(48): p. E4601-10.

446. Szwedziak, P., Wang, Q., Freund, S.M., and Lowe, J., *FtsA forms actin-like protofilaments.* Embo Journal, 2012. **31**(10): p. 2249-60.

447. Jekely, G., *Origin and evolution of the self-organizing cytoskeleton in the network of eukaryotic organelles.* Cold Spring Harb Perspect Biol, 2014. **6**(9): p. a016030.

448. Chen, B.J. and Lamb, R.A., *Mechanisms for enveloped virus budding: can some viruses do without an ESCRT?* Virology, 2008. **372**(2): p. 221-32.

449. Ruan, J., Xia, S., Liu, X., Lieberman, J., and Wu, H., *Cryo-EM structure of the gasdermin A3 membrane pore.* Nature, 2018. **557**(7703): p. 62-67.

450. Lin, P.H., Lin, H.Y., Kuo, C.C., and Yang, L.T., *N-terminal functional domain of Gasdermin A3 regulates mitochondrial homeostasis via mitochondrial targeting.* Journal of Biomedical Science, 2015. **22**: p. 44.

451. Ding, J., Wang, K., Liu, W., She, Y., Sun, Q., Shi, J., Sun, H., Wang, D.C., and Shao, F., *Pore-forming activity and structural autoinhibition of the gasdermin family.* Nature, 2016. **535**(7610): p. 111-6.

452. Ding, J., Wang, K., Liu, W., She, Y., Sun, Q., Shi, J., Sun, H., Wang, D.C., and Shao, F., *Erratum: Pore-forming activity and structural autoinhibition of the gasdermin family.* Nature, 2016. **540**(7631): p. 150.

453. Evavold, C.L., Ruan, J., Tan, Y., Xia, S., Wu, H., and Kagan, J.C., *The Pore-Forming Protein Gasdermin D Regulates Interleukin-1 Secretion from Living Macrophages.* Immunity, 2018. **48**(1): p. 35-44 e6.

454. Yuen, T.J., Flesch, I.E., Hollett, N.A., Dobson, B.M., Russell, T.A., Fahrer, A.M., and Tscharke, D.C., *Analysis of A47, an immunoprevalent protein of vaccinia virus, leads to a reevaluation of the total antiviral CD8+ T cell response.* J Virol, 2010. **84**(19): p. 10220-9.

455. Burroughs, A.M., Balaji, S., Iyer, L.M., and Aravind, L., *Small but versatile: the extraordinary functional and structural diversity of the beta-grasp fold.* Biol Direct, 2007. **2**: p. 18.

456. Su, V. and Lau, A.F., *Ubiquitin-like and ubiquitin-associated domain proteins: significance in proteasomal degradation.* Cell Mol Life Sci, 2009. **66**(17): p. 2819-33.

457. Isaacson, R.L., Pye, V.E., Simpson, P., Meyer, H.H., Zhang, X., Freemont, P.S., and Matthews, S., *Detailed structural insights into the p97-Npl4-Ufd1 interface.* J Biol Chem, 2007. **282**(29): p. 21361-9.

458. Meyer, H., Bug, M., and Bremer, S., *Emerging functions of the VCP/p97 AAA-ATPase in the ubiquitin system.* Nat Cell Biol, 2012. **14**(2): p. 117-23.

459. Dubiel, W., Ferrell, K., and Rechsteiner, M., *Peptide sequencing identifies MSS1, a modulator of HIV Tat-mediated transactivation, as subunit 7 of the 26 S protease.* Febs Letters, 1993. **323**(3): p. 276-8.

460. Schweitzer, A., Aufderheide, A., Rudack, T., Beck, F., Pfeifer, G., Plitzko, J.M., Sakata, E., Schulten, K., Forster, F., and Baumeister, W., *Structure of the human 26S proteasome at a resolution of 3.9 A.* Proc Natl Acad Sci U S A, 2016. **113**(28): p. 7816-21.

461. Shibuya, H., Irie, K., Ninomiya-Tsuji, J., Goebl, M., Taniguchi, T., and Matsumoto, K., *New human gene encoding a positive modulator of HIV Tat-mediated transactivation.* Nature, 1992. **357**(6380): p. 700-2.

462. Denko, N., Schindler, C., Koong, A., Laderoute, K., Green, C., and Giaccia, A., *Epigenetic regulation of gene expression in cervical cancer cells by the tumor microenvironment.* Clin Cancer Res, 2000. **6**(2): p. 480-7.

463. Colgrave, M.L. and Craik, D.J., *Thermal, chemical, and enzymatic stability of the cyclotide kalata B1: the importance of the cyclic cystine knot.* Biochemistry, 2004. **43**(20): p. 5965-75.

464. Park, H.G., Kyung, S.S., Lee, K.S., Kim, B.Y., Choi, Y.S., Yoon, H.J., Kwon, H.W., Je, Y.H., and Jin, B.R., *Dual function of a bee (Apis cerana) inhibitor cysteine knot peptide that acts as an antifungal peptide and insecticidal venom toxin.* Dev Comp Immunol, 2014. **47**(2): p. 247-53.

465. Kolmar, H., *Biological diversity and therapeutic potential of natural and engineered cystine knot miniproteins.* Current Opinion in Pharmacology, 2009. **9**(5): p. 608-14.

466. Su, M., Li, H., Wang, H., Kim, E., Kim, H.S., Kim, E.H., Lee, J., and Jung, J.H., *Stable and biocompatible cystine knot peptides from the marine sponge Asteropus sp.* Bioorg Med Chem, 2016. **24**(13): p. 2979-2987.

467. Lambert, J., Keppi, E., Dimarcq, J.L., Wicker, C., Reichhart, J.M., Dunbar, B., Lepage, P., Van Dorsselaer, A., Hoffmann, J., Fothergill, J., and et al., *Insect immunity: isolation from immune blood of the dipteran Phormia terranovae of two insect antibacterial peptides with sequence homology to rabbit lung macrophage bactericidal peptides.* Proc Natl Acad Sci U S A, 1989. **86**(1): p. 262-6.

468. Bulet, P., Cociancich, S., Dimarcq, J.L., Lambert, J., Reichhart, J.M., Hoffmann, D., Hetru, C., and Hoffmann, J.A., *Insect immunity. Isolation from a coleopteran insect of a novel inducible antibacterial peptide and of new members of the insect defensin family.* J Biol Chem, 1991. **266**(36): p. 24520-5.

469. Yi, H.Y., Chowdhury, M., Huang, Y.D., and Yu, X.Q., *Insect antimicrobial peptides and their applications.* Appl Microbiol Biotechnol, 2014. **98**(13): p. 5807-22.

470. Rotem, D. and Schuldiner, S., *EmrE, a multidrug transporter from Escherichia coli, transports monovalent and divalent substrates with the same stoichiometry.* J Biol Chem, 2004. **279**(47): p. 48787-93.

471. Schuldiner, S., *EmrE, a model for studying evolution and mechanism of ion-coupled transporters.* Biochimica Et Biophysica Acta, 2009. **1794**(5): p. 748-62.

472. Villarreal, L.P., *Persistent virus and addiction modules: an engine of symbiosis.* Curr Opin Microbiol, 2016. **31**: p. 70-79.

473. Wang, X., Lord, D.M., Cheng, H.Y., Osbourne, D.O., Hong, S.H., Sanchez-Torres, V., Quiroga, C., Zheng, K., Herrmann, T., Peti, W., Benedik, M.J., Page, R., and Wood, T.K., *A new type V toxin-antitoxin system where mRNA for toxin GhoT is cleaved by antitoxin GhoS.* Nat Chem Biol, 2012. **8**(10): p. 855-61.

474. Kim, J.S., Schantz, A.B., Song, S., Kumar, M., and Wood, T.K., *GhoT of the GhoT/GhoS toxin/antitoxin system damages lipid membranes by forming transient pores.* Biochem Biophys Res Commun, 2018. **497**(2): p. 467-472.

475. Lubas, W.A. and Spiro, R.G., *Golgi endo-alpha-D-mannosidase from rat liver, a novel N-linked carbohydrate unit processing enzyme.* J Biol Chem, 1987. **262**(8): p. 3775-81.

476. Thompson, A.J., et al., *Structural and mechanistic insight into N-glycan processing by endo-alpha-mannosidase.* Proc Natl Acad Sci U S A, 2012. **109**(3): p. 781-6.

477. Watanabe, Y., Bowden, T.A., Wilson, I.A., and Crispin, M., *Exploitation of glycosylation in enveloped virus pathobiology.* Biochim Biophys Acta Gen Subj, 2019. **1863**(10): p. 1480-1497.

478. Pan, S., Cheng, X., and Sifers, R.N., *Golgi-situated endoplasmic reticulum alpha-1, 2-mannosidase contributes to the retrieval of ERAD substrates through a direct interaction with gamma-COP.* Molecular Biology of the Cell, 2013. **24**(8): p. 1111-21.

479. Li, S. and Jedrzejas, M.J., *Hyaluronan binding and degradation by Streptococcus agalactiae hyaluronate lyase.* J Biol Chem, 2001. **276**(44): p. 41407-16.

480. Hynes, W.L. and Walton, S.L., *Hyaluronidases of Gram-positive bacteria.* FEMS Microbiol Lett, 2000. **183**(2): p. 201-7.

481. Dinglasan, R.R., Alaganan, A., Ghosh, A.K., Saito, A., van Kuppevelt, T.H., and Jacobs-Lorena, M., *Plasmodium falciparum ookinetes require mosquito midgut chondroitin sulfate proteoglycans for cell invasion.* Proc Natl Acad Sci U S A, 2007. **104**(40): p. 15882-7.

482. Mathias, D.K., Pastrana-Mena, R., Ranucci, E., Tao, D., Ferruti, P., Ortega, C., Staples, G.O., Zaia, J., Takashima, E., Tsuboi, T., Borg, N.A., Verotta, L., and Dinglasan, R.R., *A small molecule glycosaminoglycan mimetic blocks Plasmodium invasion of the mosquito midgut.* PLoS Pathog, 2013. **9**(11): p. e1003757.

483. Chung, C.S., Hsiao, J.C., Chang, Y.S., and Chang, W., *A27L protein mediates vaccinia virus interaction with cell surface heparan sulfate.* J Virol, 1998. **72**(2): p. 1577-85.

484. Pelczar, P.L., Igarashi, T., Setlow, B., and Setlow, P., *Role of GerD in germination of Bacillus subtilis spores.* Journal of Bacteriology, 2007. **189**(3): p. 1090-8.

485. Li, Y., Jin, K., Ghosh, S., Devarakonda, P., Carlson, K., Davis, A., Stewart, K.A., Cammett, E., Pelczar Rossi, P., Setlow, B., Lu, M., Setlow, P., and Hao, B., *Structural and functional analysis of the GerD spore germination protein of Bacillus species.* J Mol Biol, 2014. **426**(9): p. 1995-2008.

486. Rodriguez, J.F., Paez, E., and Esteban, M., *A 14,000-Mr envelope protein of vaccinia virus is involved in cell fusion and forms covalently linked trimers.* J Virol, 1987. **61**(2): p. 395-404.

487. Lai, C., Gong, S., and Esteban, M., *Structural and functional properties of the 14-kDaenvelope protein of vaccinia virus synthesized in Escherichia coli.* Journal of Biological Chemistry, 1990. **265**: p. 22174-22180.

488. Chen, Y. and Rosen, B.P., *Metalloregulatory properties of the ArsD repressor.* J Biol Chem, 1997. **272**(22): p. 14257-62.

489. Li, S., Rosen, B.P., Borges-Walmsley, M.I., and Walmsley, A.R., *Evidence for cooperativity between the four binding sites of dimeric ArsD, an As(III)-responsive transcriptional regulator.* J Biol Chem, 2002. **277**(29): p. 25992-6002.

490. Lin, Y.F., Walmsley, A.R., and Rosen, B.P., *An arsenic metallochaperone for an arsenic detoxification pump.* Proc Natl Acad Sci U S A, 2006. **103**(42): p. 15617-22.

491. Lin, Y.F., Yang, J., and Rosen, B.P., *ArsD: an As(III) metallochaperone for the ArsAB As(III)-translocating ATPase.* J Bioenerg Biomembr, 2007. **39**(5-6): p. 453-8.

492. Lin, Y.F., Yang, J., and Rosen, B.P., *ArsD residues Cys12, Cys13, and Cys18 form an As(III)-binding site required for arsenic metallochaperone activity.* J Biol Chem, 2007. **282**(23): p. 16783-91.

493. Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E.L., *Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes.* J Mol Biol, 2001. **305**(3): p. 567-80.

494. Almagro Armenteros, J.J., Tsirigos, K.D., Sonderby, C.K., Petersen, T.N., Winther, O., Brunak, S., von Heijne, G., and Nielsen, H., *SignalP 5.0 improves signal peptide predictions using deep neural networks.* Nat Biotechnol, 2019. **37**(4): p. 420-423.

495. Hu, X., Wolffe, E.J., Weisberg, A.S., Carroll, L.J., and Moss, B., *Repression of the A8L gene, encoding the early transcription factor 82- kilodalton subunit, inhibits morphogenesis of vaccinia virions.* Journal of Virololgy, 1998. **72**(1): p. 104-112.

496. Hu, X., Carroll, L.J., Wolffe, E.J., and Moss, B., *De novo synthesis of the early transcription factor 70-kilodalton subunit is required for morphogenesis of vaccinia virions.* Journal of Virology, 1996. **70**(11): p. 7669-77.

497. Zhang, Y., Ahn, B.-Y., and Moss, B., *Targeting of a multicomponent transcription apparatus into assembling vaccinia virus particles requires RAP94, an RNA polymerase-associated protein.* J. Virol, 1994. **68**(3): p. 1360-1370.

498. Holowczak, J.A., Thomas, V.L., and Flores, L., *Isolation and characterization of vaccinia virus "nucleoids".* Virology, 1975. **67**(2): p. 506-519.

499. Steinegger, M., Meier, M., Mirdita, M., Vohringer, H., Haunsberger, S.J., and Soding, J., *HH-suite3 for fast remote homology detection and deep protein annotation.* BMC Bioinformatics, 2019. **20**(1): p. 473.

500. Hendy, M.D. and Penny, D., *Branch and bound algorithms to determine minimal evolutionary trees.* Mathematical Biosciences, 1982. **59**: p. 277-290.

501. Dales, S. and Siminovitch, L., *The development of vaccinia virus in Earle's L strain cells as examined by electron microscopy.* J Biophys Biochem Cytol, 1961. **10**: p. 475-503.

502. Sodeik, B., Cudmore, S., Ericsson, M., Esteban, M., Niles, E.G., and Griffiths, G., *Assembly of vaccinia virus: incorporation of p14 and p32 into the membrane of the intracellular mature virus.* J Virol, 1995. **69**(6): p. 3560-74.

503. Cudmore, S., Blasco, R., Vincentelli, R., Esteban, M., Sodeik, B., Griffiths, G., and Krijnse Locker, J., *A vaccinia virus core protein, p39, is membrane associated.* J. Virol., 1996. **70**(10): p. 6909-6921.

504. Jensen, O.N., Houthaeve, T., Shevchenko, A., Cudmore, S., Ashford, T., Mann, M., Griffiths, G., and Krijnse Locker, J., *Identification of the major membrane and core proteins of vaccinia virus by two-dimensional electrophoresis.* Journal of Virology, 1996. **70**(11): p. 7485-7497.

505. Koksal, A.C., Nardozzi, J.D., and Cingolani, G., *Dimeric quaternary structure of the prototypical dual specificity phosphatase VH1.* J Biol Chem, 2009. **284**(15): p. 10129-37.

506. Koksal, A.C. and Cingolani, G., *Dimerization of Vaccinia virus VH1 is essential for dephosphorylation of STAT1 at tyrosine 701.* J Biol Chem, 2011. **286**(16): p. 14373-82.

507. McCraith, S., Holtzman, T., Moss, B., and Fields, S., *Genome-wide analysis of vaccinia virus protein-protein interactions.* Proc. Natl. Acad. Sci. USA, 2000. **97**(9): p. 4879-4884.

508. McFadden, B.D., Moussatche, N., Kelley, K., Kang, B.H., and Condit, R.C., *Vaccinia virions deficient in transcription enzymes lack a nucleocapsid.* Virology, 2012. **434**(1): p. 50-8.

509. Hagler, J. and Shuman, S., *A freeze-frame view of eukaryotic transcription during elongation and capping of nascent mRNA.* Science, 1992. **255**(21 Feb): p. 983-986.

510. Broyles, S.S. and Moss, B., *Sedimentation of an RNA polymerase complex from vaccinia virus that specifically initiates and terminates transcription.* Mol. Cell Biol., 1987. **7**(1): p. 7-14.

511. Lasker, K., Forster, F., Bohn, S., Walzthoeni, T., Villa, E., Unverdorben, P., Beck, F., Aebersold, R., Sali, A., and Baumeister, W., *Molecular architecture of the 26S proteasome holocomplex determined by an integrative approach.* Proc Natl Acad Sci U S A, 2012. **109**(5): p. 1380-7.

512. Kalisman, N., Adams, C.M., and Levitt, M., *Subunit order of eukaryotic TRiC/CCT chaperonin by cross-linking, mass spectrometry, and combinatorial homology modeling.* Proc Natl Acad Sci U S A, 2012. **109**(8): p. 2884-9.

513. Leitner, A., et al., *The molecular architecture of the eukaryotic chaperonin TRiC/CCT.* Structure, 2012. **20**(5): p. 814-25.

514. Chen, Z.A., Jawhari, A., Fischer, L., Buchen, C., Tahir, S., Kamenski, T., Rasmussen, M., Lariviere, L., Bukowski-Wills, J.C., Nilges, M., Cramer, P., and Rappsilber, J., *Architecture of the RNA polymerase II-TFIIF complex revealed by cross-linking and mass spectrometry.* Embo Journal, 2010. **29**(4): p. 717-26.

515. Herzog, F., Kahraman, A., Boehringer, D., Mak, R., Bracher, A., Walzthoeni, T., Leitner, A., Beck, M., Hartl, F.U., Ban, N., Malmstrom, L., and Aebersold, R., *Structural probing of a protein phosphatase 2A network by chemical cross-linking and mass spectrometry.* Science, 2012. **337**(6100): p. 1348-52.

516. Jennebach, S., Herzog, F., Aebersold, R., and Cramer, P., *Crosslinking-MS analysis reveals RNA polymerase I domain architecture and basis of rRNA cleavage.* Nucleic Acids Res, 2012. **40**(12): p. 5591-601.

517. Wu, C.C., Herzog, F., Jennebach, S., Lin, Y.C., Pai, C.Y., Aebersold, R., Cramer, P., and Chen, H.T., *RNA polymerase III subunit architecture and implications for open promoter complex formation.* Proc Natl Acad Sci U S A, 2012. **109**(47): p. 19232-7.

518. Elias, J.E. and Gygi, S.P., *Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry.* Nat Methods, 2007. **4**(3): p. 207-14.

519. Muhlbacher, W., Sainsbury, S., Hemann, M., Hantsche, M., Neyer, S., Herzog, F., and Cramer, P., *Conserved architecture of the core RNA polymerase II initiation complex.* Nat Commun, 2014. **5**: p. 4310.

520. Mentinova, M., Barefoot, N.Z., and McLuckey, S.A., *Solution versus gas-phase modification of peptide cations with NHS-ester reagents.* J Am Soc Mass Spectrom, 2012. **23**(2): p. 282-9.

521. Ferrari, A.J.R., Clasen, M.A., Kurt, L., Carvalho, P.C., Gozzo, F.C., and Martinez, L., *TopoLink: evaluation of structural models using chemical crosslinking distance constraints.* Bioinformatics, 2019. **35**(17): p. 3169-3170.

522. Moss, B., *Poxvirus membrane biogenesis.* Virology, 2015. **479-480**: p. 619-26.

523. Bayliss, C.D. and Smith, G.L., *Vaccinia virion protein VP8, the 25 kDa product of the L4R gene, binds single-stranded DNA and RNA with similar affinity.* Nucleic Acids Research, 1997. **25**: p. 3984-3990.

524. Wilcock, D. and Smith, G.L., *Vaccinia virus core protein VP8 is required for virus infectivity, but not for core protein processing or for INV and EEV formation.* Virology, 1994. **202**(1): p. 294-304.

525. Wilcock, D. and Smith, G.L., *Vaccinia virions lacking core protein VP8 are deficient in early transcription.* J. Virol., 1996. **70**: p. 934-943.

526. Yang, S.J. and Hruby, D.E., *Vaccinia virus A12L protein and its AG/A proteolysis play an important role in viral morphogenic transition.* Virol J, 2007. **4**: p. 73.

527. Whitehead, S.S. and Hruby, D.E., *Differential utilization of a conserved motif for the proteolytic maturation of vaccinia virus proteins.* Virology, 1994. **200**(1): p. 154-61.

528. Yang, W.P., Kao, S.Y., and Bauer, W.R., *Biosynthesis and post-translational cleavage of vaccinia virus structural protein VP8.* Virology, 1988. **167**(2): p. 585-90.

529. Jesus, D.M., Moussatche, N., McFadden, B.B., Nielsen, C.P., D'Costa, S.M., and Condit, R.C., *Vaccinia virus protein A3 is required for the production of normal immature virions and for the encapsidation of the nucleocapsid protein L4.* Virology, 2015. **481**: p. 1-12.

530. Jesus, D.M., Moussatche, N., and Condit, R.C., *Vaccinia virus mutations in the L4R gene encoding a virion structural protein produce abnormal mature particles lacking a nucleocapsid.* J Virol, 2014. **88**(24): p. 14017-29.

531. Betakova, T. and Moss, B., *Disulfide bonds and membrane topology of the vaccinia virus A17L envelope protein.* J Virol, 2000. **74**(5): p. 2438-42.

532. Krijnse-Locker, J., Schleich, S., Rodriguez, D., Goud, B., Snijder, E.J., and Griffiths, G., *The role of a 21-kDa viral membrane protein in the assembly of vaccinia virus from the intermediate compartment.* J Biol Chem, 1996. **271**(25): p. 14950-8.

533. Unger, B., Mercer, J., Boyle, K.A., and Traktman, P., *Biogenesis of the vaccinia virus membrane: genetic and ultrastructural analysis of the contributions of the A14 and A17 proteins.* J Virol, 2013. **87**(2): p. 1083-97.

534. Wolffe, E.J., Moore, D.M., Peters, P.J., and Moss, B., *Vaccinia virus A17L open reading frame encodes an essential component of nascent viral membranes that is required to initiate morphogenesis.* J. Virol., 1996. **70**(5): p. 2797-2808.

535. Erlandson, K.J., Bisht, H., Weisberg, A.S., Hyun, S.I., Hansen, B.T., Fischer, E.R., Hinshaw, J.E., and Moss, B., *Poxviruses Encode a Reticulon-Like Protein that Promotes Membrane Curvature.* Cell Rep, 2016. **14**(9): p. 2084-2091.

536. Rodriguez, D., Rodriguez, J.-R., and Esteban, M., *The vaccinia virus 14-kilodalton fusion protein forms a stable complex with the processed protein encoded by the vaccinia virus A17L gene.* J. Virol., 1993. **67**: p. 3435-3440.

537. Betakova, T., Wolffe, E.J., and Moss, B., *Membrane topology of the vaccinia virus A17L envelope protein.* Virology, 1999. **261**(2): p. 347-56.

538. Wallengren, K., Risco, C., Krijnse-Locker, J., Esteban, M., and Rodriguez, D., *The A17L gene product of vaccinia virus is exposed on the surface of IMV.* Virology, 2001. **290**(1): p. 143-152.

539. Chertov, O., Telezhinskaya, I.N., Zaitseva, E.V., Golubeva, T.B., Zinov'ev, V.V., Ovechkina, L.G., Mazkova, L.B., and Malygin, E.G., *Amino acid sequence determination of vaccinia virus immunodominant protein p35 and identification of the gene.* Biomed Sci, 1991. **2**(2): p. 151-4.

540. Zinoviev, V.V., Tchikaev, N.A., Chertov, O., and Malygin, E.G., *Identification of the gene encoding vaccinia virus immunodominant protein p35.* Gene, 1994. **147**(2): p. 209-14.

541. Kutay, U., Hartmann, E., and Rapoport, T.A., *A class of membrane proteins with a C-terminal anchor.* Trends Cell Biol, 1993. **3**(3): p. 72-5.

542. Morgan, C., Rifkind, R.A., and Rose, H.M., *The Use of Ferritin-conjugated Antibodies in Electron Microscopic Studies of Influenza and Vaccinia Viruses.* Cold Spring Harbor Symposia on Quantitative Biology, 1962. **27**: p. 57-65.

543. Dales, S. and Pogo, B.G.T., eds. *Biology of poxviruses.* Virology monographs. 1981, Springer-Verlag: Vienna, Austria.

544. Rodriguez, J.F., Janeczko, R., and Esteban, M., *Isolation and characterization of neutralizing monoclonal antibodies to vaccinia virus.* J Virol, 1985. **56**(2): p. 482-8.

545. Wang, D.R., Hsiao, J.C., Wong, C.H., Li, G.C., Lin, S.C., Yu, S.S., Chen, W., Chang, W., and Tzou, D.L., *Vaccinia viral protein A27 is anchored to the viral membrane via a cooperative interaction with viral membrane protein A17.* J Biol Chem, 2014. **289**(10): p. 6639-55.

546. Sanderson, C.M., Hollinshead, M., and Smith, G.L., *The vaccinia virus A27L protein is needed for the microtubule-dependent transport of intracellular mature virus particles.* J Gen Virol, 2000. **81**(Pt 1): p. 47-58.

547. Rodriguez, J.F. and Smith, G.L., *IPTG-dependent vaccinia virus: identification of a virus protein enabling virion envelopment by Golgi membrane and egress.* Nucleic Acids Research, 1990. **18**: p. 5347-5351.

548. Ward, B.M., *Visualization and characterization of the intracellular movement of vaccinia virus intracellular mature virions.* J Virol, 2005. **79**(8): p. 4755-63.

549. Howard, A.R., Weisberg, A.S., and Moss, B., *Congregation of orthopoxvirus virions in cytoplasmic A-type inclusions is mediated by interactions of a bridging protein (A26p) with a matrix protein (ATIp) and a virion membrane-associated protein (A27p).* J Virol, 2010. **84**(15): p. 7592-602.

550. Howard, A.R., Senkevich, T.G., and Moss, B., *Vaccinia virus A26 and A27 proteins form a stable complex tethered to mature virions by association with the A17 transmembrane protein.* J Virol, 2008. **82**(24): p. 12384-91.

551. Dales, S., *An electron microscope study of the early association between two mammalian viruses and their hosts.* Journal of Cell Biology, 1962. **13**: p. 303-322.

552. Noyes, W.F., *The surface fine structure of vaccinia virus.* Virology, 1962. **17**: p. 282-287.

553. Noyes, W.F., *Further studies on the structure of vaccinia virus.* Virology, 1962. **18**: p. 511-516.

554. Wilton, S., Mohandas, A.R., and Dales, S., *Organization of vaccinia envelope and relationship to the structure of intracellular mature virions.* Virology, 1995. **214**: p. 503-511.

555. Franke, C.A., Wilson, E.M., and Hruby, D.E., *Use of a cell-free system to identify the vaccinia virus L1R gene product as the major late myristylated virion protein M25.* J Virol, 1990. **64**(12): p. 5988-96.

556. Wolffe, E.J., Vijaya, S., and Moss, B., *A myristylated membrane protein encoded by the vaccinia virus L1R open reading frame is the target of potent neutralizing monoclonal antibodies.* Virology, 1995. **211**(1): p. 53-63.

557. Moller, S., Croning, M.D., and Apweiler, R., *Evaluation of methods for the prediction of membrane spanning regions.* Bioinformatics, 2001. **17**(7): p. 646-53.

558. Yoder, J.D., Chen, T., and Hruby, D.E., *Sequence-independent acylation of the vaccinia virus A-type inclusion protein.* Biochemistry, 2004. **43**(26): p. 8297-302.

559. Szajner, P., Weisberg, A.S., and Moss, B., *Unique temperature-sensitive defect in vaccinia virus morphogenesis maps to a single nucleotide substitution in the A30L gene.* J Virol, 2001. **75**(22): p. 11222-6.

560. Szajner, P., Weisberg, A.S., Wolffe, E.J., and B., M., *Vaccinia virus A30L protein is required for association of viral membranes with dense viroplasm to form immature virions.* J. Virol., 2001. **75**(13): p. 5752-5761.

561. Mercer, J. and Traktman, P., *Genetic and cell biological characterization of the vaccinia virus A30 and G7 phosphoproteins.* J Virol., 2005. **79**: p. 7146-7161.

562. Chiu, W.L., Szajner, P., Moss, B., and Chang, W., *Effects of a temperature sensitivity mutation in the J1R protein component of a complex required for vaccinia virus assembly.* J Virol, 2005. **79**(13): p. 8046-56.

563. DeMasi, J., Du, S., Lennon, D., and Traktman, P., *Vaccinia virus telomeres: interaction with the viral I1, I6, and K4 proteins.* J Virol, 2001. **75**(21): p. 10090-105.

564. Harrison, M.L., Desaulniers, M.A., Noyce, R.S., and Evans, D.H., *The acidic C-terminus of vaccinia virus I3 single-strand binding protein promotes proper assembly of DNA-protein complexes.* Virology, 2016. **489**: p. 212-22.

565. Greseth, M.D., Boyle, K.A., Bluma, M.S., Unger, B., Wiebe, M.S., Soares-Martins, J.A., Wickramasekera, N.T., Wahlberg, J., and Traktman, P., *Molecular genetic and biochemical characterization of the vaccinia virus I3 protein, the replicative single-stranded DNA binding protein.* J Virol, 2012. **86**(11): p. 6197-209.

566. Doglio, L., De Marco, A., Schleich, S., Roos, N., and Krijnse Locker, J., *The Vaccinia virus E8R gene product: a viral membrane protein that is made early in infection and packaged into the virions' core.* Journal of Virology, 2002. **76**(19): p. 9773-9786.

567. Sodeik, B. and Krijnse-Locker, J., *Assembly of vaccinia virus revisited: de novo membrane synthesis or acquisition from the host?* Trends Microbiol, 2002. **10**(1): p. 15-24.

568. Ansarah-Sobrinho, C. and Moss, B., *Vaccinia virus G1 protein, a predicted metalloprotease, is essential for morphogenesis of infectious virions but not for cleavage of major core proteins.* J Virol, 2004. **78**(13): p. 6855-63.

569. Whitehead, S.S. and Hruby, D.E., *A transcriptionally controlled trans-processing assay: putative identification of a vaccinia virus-encoded proteinase which cleaves precursor protein P25K.* J Virol, 1994. **68**(11): p. 7603-8.

570. Mattson, G., Conklin, E., Desai, S., Nielander, G., Savage, M.D., and Morgensen, S., *A practical approach to crosslinking.* Molecular Biology Reports, 1993. **17**(3): p. 167-83.

571. Matson, J., Chou, W., Ngo, T., and Gershon, P.D., *Static and dynamic protein phosphorylation in the Vaccinia virion.* Virology, 2014. **452-453**: p. 310-23.

572. Erickson, H.P., *Size and shape of protein molecules at the nanometer level determined by sedimentation, gel filtration, and electron microscopy.* Biol Proced Online, 2009. **11**: p. 32-51.

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

574. Zhang, X., Walker, S.B., Chipman, P.R., Nibert, M.L., and Baker, T.S., *Reovirus polymerase lambda 3 localized by cryo-electron microscopy of virions at a resolution of 7.6 A.* Nature Structural Biology, 2003. **10**(12): p. 1011-8.

575. Leitner, A., Walzthoeni, T., and Aebersold, R., *Lysine-specific chemical cross-linking of protein complexes and identification of cross-linking sites using LC-MS/MS and the xQuest/xProphet software pipeline.* Nat Protoc, 2014. **9**(1): p. 120-37.

576. Leitner, A., Joachimiak, L.A., Unverdorben, P., Walzthoeni, T., Frydman, J., Forster, F., and Aebersold, R., *Chemical cross-linking/mass spectrometry targeting acidic residues in proteins and protein complexes.* Proc Natl Acad Sci U S A, 2014. **111**(26): p. 9455-60.

577. Fonslow, B.R., Stein, B.D., Webb, K.J., Xu, T., Choi, J., Park, S.K., and Yates, J.R., 3rd, *Digestion and depletion of abundant proteins improves proteomic coverage.* Nat Methods, 2013. **10**(1): p. 54-6.

578. Rappsilber, J., Mann, M., and Ishihama, Y., *Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips.* Nat Protoc, 2007. **2**(8): p. 1896-906.

579. Trnka, M.J., Baker, P.R., Robinson, P.J., Burlingame, A.L., and Chalkley, R.J., *Matching cross-linked peptide spectra: only as good as the worse identification.* Mol Cell Proteomics, 2014. **13**(2): p. 420-34.

580. Yang, B., et al., *Identification of cross-linked peptides from complex samples.* Nat Methods, 2012. **9**(9): p. 904-6.

581. Hoopmann, M.R., Zelter, A., Johnson, R.S., Riffle, M., MacCoss, M.J., Davis, T.N., and Moritz, R.L., *Kojak: efficient analysis of chemically cross-linked protein complexes.* J Proteome Res, 2015. **14**(5): p. 2190-8.

582. Kall, L., Canterbury, J.D., Weston, J., Noble, W.S., and MacCoss, M.J., *Semi-supervised learning for peptide identification from shotgun proteomics datasets.* Nat Methods, 2007. **4**(11): p. 923-5.

583. Kall, L., Storey, J.D., MacCoss, M.J., and Noble, W.S., *Assigning significance to peptides identified by tandem mass spectrometry using decoy databases.* J Proteome Res, 2008. **7**(1): p. 29-34.

584. Kall, L., Storey, J.D., and Noble, W.S., *Non-parametric estimation of posterior error probabilities associated with peptides identified by tandem mass spectrometry.* Bioinformatics, 2008. **24**(16): p. i42-8.

585. Yu, F., Li, N., and Yu, W., *ECL: an exhaustive search tool for the identification of cross-linked peptides using whole database.* BMC Bioinformatics, 2016. **17**(1): p. 217.

586. Yu, F., Li, N., and Yu, W., *Exhaustively Identifying Cross-Linked Peptides with a Linear Computational Complexity.* J Proteome Res, 2017. **16**(10): p. 3942-3952.

587. The, M., MacCoss, M.J., Noble, W.S., and Kall, L., *Fast and Accurate Protein False Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0.* J Am Soc Mass Spectrom, 2016. **27**(11): p. 1719-1727.

588. Storey, J.D. and Tibshirani, R., *Statistical significance for genomewide studies.* Proc Natl Acad Sci U S A, 2003. **100**(16): p. 9440-5.

589. Combe, C.W., Fischer, L., and Rappsilber, J., *xiNET: cross-link network maps with residue resolution.* Mol Cell Proteomics, 2015. **14**(4): p. 1137-47.

590. Marsden, R.L., McGuffin, L.J., and Jones, D.T., *Rapid protein domain assignment from amino acid sequence using predicted secondary structure.* Protein Sci, 2002. **11**(12): p. 2814-24.

591. Bryson, K., Cozzetto, D., and Jones, D.T., *Computer-assisted protein domain boundary prediction using the DomPred server.* Curr Protein Pept Sci, 2007. **8**(2): p. 181-8.

592. Yu, C. and Huang, L., *New advances in cross-linking mass spectrometry toward structural systems biology.* Curr Opin Chem Biol, 2023. **76**: p. 102357.

593. Nie, M. and Li, H., *Innovation in Cross-Linking Mass Spectrometry Workflows: Toward a Comprehensive, Flexible, and Customizable Data Analysis Platform.* J Am Soc Mass Spectrom, 2023. **34**(9): p. 1949-1956.

594. Chen, Z.L., et al., *A high-speed search engine pLink 2 with systematic evaluation for proteome-scale identification of cross-linked peptides.* Nat Commun, 2019. **10**(1): p. 3404.

595. Switzar, L., Giera, M., and Niessen, W.M., *Protein digestion: an overview of the available techniques and recent developments.* J Proteome Res, 2013. **12**(3): p. 1067-77.

596. Wisniewski, J.R., *Quantitative Evaluation of Filter Aided Sample Preparation (FASP) and Multienzyme Digestion FASP Protocols.* Anal Chem, 2016. **88**(10): p. 5438-43.

597. Verheggen, K., Martens, L., Berven, F.S., Barsnes, H., and Vaudel, M., *Database Search Engines: Paradigms, Challenges and Solutions.* Adv Exp Med Biol, 2016. **919**: p. 147-156.

598. Singh, P., Panchaud, A., and Goodlett, D.R., *Chemical cross-linking and mass spectrometry as a low-resolution protein structure determination technique.* Anal Chem, 2010. **82**(7): p. 2636-42.

599. Chen, Z.A. and Rappsilber, J., *Protein structure dynamics by crosslinking mass spectrometry.* Curr Opin Struct Biol, 2023. **80**: p. 102599.

600. Sinz, A., *Chemical cross-linking and mass spectrometry to map three-dimensional protein structures and protein-protein interactions.* Mass Spectrometry Reviews, 2006. **25**: p. 663 -.

601. Leitner, A., Walzthoeni, T., Kahraman, A., Herzog, F., Rinner, O., Beck, M., and Aebersold, R., *Probing native protein structures by chemical cross-linking, mass spectrometry, and bioinformatics.* Mol Cell Proteomics, 2010. **9**(8): p. 1634-49.

602. Back, J.W., de Jong, L., Muijsers, A.O., and de Koster, C.G., *Chemical cross-linking and mass spectrometry for protein structural modeling.* J Mol Biol, 2003. **331**(2): p. 303-13.

603. Chiva, C., Ortega, M., and Sabido, E., *Influence of the digestion technique, protease, and missed cleavage peptides in protein quantitation.* J Proteome Res, 2014. **13**(9): p. 3979-86.

604. Mirzakhanyan, Y. and Gershon, P., *The Vaccinia virion: Filling the gap between atomic and ultrastructure.* PLoS Pathog, 2019. **15**(1): p. e1007508.

605. Netz, E., Dijkstra, T.M.H., Sachsenberg, T., Zimmermann, L., Walzer, M., Monecke, T., Ficner, R., Dybkov, O., Urlaub, H., and Kohlbacher, O., *OpenPepXL: An Open-Source Tool for Sensitive Identification of Cross-Linked Peptides in XL-MS.* Mol Cell Proteomics, 2020. **19**(12): p. 2157-2168.

606. Mendes, M.L., et al., *An integrated workflow for crosslinking mass spectrometry.* Mol Syst Biol, 2019. **15**(9): p. e8994.

607. Dai, J., Jiang, W., Yu, F., and Yu, W., *Xolik: finding cross-linked peptides with maximum paired scores in linear time.* Bioinformatics, 2019. **35**(2): p. 251-257.

608. Hoopmann, M.R., Shteynberg, D.D., Zelter, A., Riffle, M., Lyon, A.S., Agard, D.A., Luan, Q., Nolen, B.J., MacCoss, M.J., Davis, T.N., and Moritz, R.L., *Improved Analysis of Cross-Linking Mass Spectrometry Data with Kojak 2.0, Advanced by Integration into the Trans-Proteomic Pipeline.* J Proteome Res, 2023. **22**(2): p. 647-655.

609. Lu, L., Millikin, R.J., Solntsev, S.K., Rolfs, Z., Scalf, M., Shortreed, M.R., and Smith, L.M., *Identification of MS-Cleavable and Noncleavable Chemically Cross-Linked Peptides with MetaMorpheus.* J Proteome Res, 2018. **17**(7): p. 2370-2376.

610. Joklik, W.K., *The purification fo four strains of poxvirus.* Virology, 1962. **18**: p. 9-18.

611. Cotter, C.A., Earl, P.L., Wyatt, L.S., and Moss, B., *Preparation of Cell Cultures and Vaccinia Virus Stocks.* Curr Protoc Microbiol, 2015. **39**: p. 14A 3 1-14A 3 18.

612. Earl, P.L. and Moss, B., *Preparation of cell cultures and vaccinia virus stocks*, in *Current protocols in molecular biology*, F.M. Ausubel, et al., Editors. 1991, Wiley Interscience: New York. p. 16.16.1-16.16.7.

613. Sharp, D.G. and McGuire, P.M., *Spectrum of physical properties among the virions of a whole population of vaccinia virus particles.* J Virol, 1970. **5**(3): p. 275-81.

614. Souza, A.R., Luques, M.N., and Damaso, C.R., *Genomic diversity of vaccinia virus strain Cantagalo isolated in southeastern Brazil during the early years of the outbreak, 1999-2006.* Mem Inst Oswaldo Cruz, 2021. **115**: p. e200521.

615. Resch, W., Hixson, K.K., Moore, R.J., Lipton, M.S., and Moss, B., *Protein composition of the vaccinia virus mature virion.* Virology, 2007. **358**(1): p. 233-247.

616. Planterose, D.N., Nishimura, C., and Salzman, N.P., *The purification of vaccinia virus from cell cultures.* Virology, 1962. **18**: p. 294-301.

617. Cotter, C.A., Earl, P.L., Wyatt, L.S., and Moss, B., *Preparation of Cell Cultures and Vaccinia Virus Stocks.* Curr Protoc Protein Sci, 2017. **89**: p. 5 12 1-5 12 18.

618. Leitner, A., Reischl, R., Walzthoeni, T., Herzog, F., Bohn, S., Forster, F., and Aebersold, R., *Expanding the chemical cross-linking toolbox by the use of multiple proteases and enrichment by size exclusion chromatography.* Mol Cell Proteomics, 2012. **11**(3): p. M111 014126.

619. Ye, J., Zhang, X., Young, C., Zhao, X., Hao, Q., Cheng, L., and Jensen, O.N., *Optimized IMAC-IMAC protocol for phosphopeptide recovery from complex biological samples.* J Proteome Res, 2010. **9**(7): p. 3561-73.

620. Kreimer, S., Belov, M.E., Danielson, W.F., Levitsky, L.I., Gorshkov, M.V., Karger, B.L., and Ivanov, A.R., *Advanced Precursor Ion Selection Algorithms for Increased Depth of Bottom-Up Proteomic Profiling.* J Proteome Res, 2016. **15**(10): p. 3563-3573.

621. Aebersold, R. and Mann, M., *Mass spectrometry-based proteomics.* Nature, 2003. **422**: p. 198.

622. Wang, N. and Li, L., *Exploring the precursor ion exclusion feature of liquid chromatography-electrospray ionization quadrupole time-of-flight mass spectrometry for improving protein identification in shotgun proteome analysis.* Anal Chem, 2008. **80**(12): p. 4696-710.

623. Seebacher, J., Mallick, P., Zhang, N., Eddes, J.S., Aebersold, R., and Gelb, M.H., *Protein cross-linking analysis using mass spectrometry, isotope-coded cross-linkers, and integrated computational data processing.* J Proteome Res, 2006. **5**(9): p. 2270-82.

624. Muller, D.R., Schindler, P., Towbin, H., Wirth, U., Voshol, H., Hoving, S., and Steinmetz, M.O., *Isotope-tagged cross-linking reagents. A new tool in mass spectrometric protein interaction analysis.* Anal Chem, 2001. **73**(9): p. 1927-34.

625. Gillet, L.C., Leitner, A., and Aebersold, R., *Mass Spectrometry Applied to Bottom-Up Proteomics: Entering the High-Throughput Era for Hypothesis Testing.* Annu Rev Anal Chem (Palo Alto Calif), 2016. **9**(1): p. 449-72.

626. Fujii, N., Jacobsen, R.B., Wood, N.L., Schoeniger, J.S., and Guy, R.K., *A novel protein crosslinking reagent for the determination of moderate resolution protein structures by mass spectrometry (MS3-D).* Bioorg Med Chem Lett, 2004. **14**(2): p. 427-9.

627. Chu, F., Mahrus, S., Craik, C.S., and Burlingame, A.L., *Isotope-coded and affinity-tagged cross-linking (ICATXL): an efficient strategy to probe protein interaction surfaces.* J Am Chem Soc, 2006. **128**(32): p. 10362-3.

628. Makepeace, K.A.T., Mohammed, Y., Rudashevskaya, E.L., Petrotchenko, E.V., Vogtle, F.N., Meisinger, C., Sickmann, A., and Borchers, C.H., *Improving Identification of In-organello Protein-Protein Interactions Using an Affinity-enrichable, Isotopically Coded, and Mass Spectrometry-cleavable Chemical Crosslinker.* Mol Cell Proteomics, 2020. **19**(4): p. 624-639.

629. Petrotchenko, E.V., Serpa, J., and Borchers, C.H., *An isotopically coded CID-cleavable biotinylated cross-linker for structural proteomics.* Mol Cell Proteomics. , 2011. **Feb;10(2):M110.001420. Epub 2010 Jul 9.**

630. Chowdhury, S.M., Du, X., Tolic, N., Wu, S., Moore, R.J., Mayer, M.U., Smith, R.D., and Adkins, J.N., *Identification of cross-linked peptides after click-based enrichment using sequential collision-induced dissociation and electron transfer dissociation tandem mass spectrometry.* Anal Chem, 2009. **81**(13): p. 5524-32.

631. Steigenberger, B., Pieters, R.J., Heck, A.J.R., and Scheltema, R.A., *PhoX: An IMAC-Enrichable Cross-Linking Reagent.* ACS Cent Sci, 2019. **5**(9): p. 1514-1522.

632. Sanford, E.J. and Smolka, M.B., *Fe-NTA Microcolumn Purification of Phosphopeptides from Immunoprecipitation (IP) Eluates for Mass Spectrometry Analysis.* Bio Protoc, 2021. **11**(15): p. e4113.

633. Burkhart, J.M., Schumbrutzki, C., Wortelkamp, S., Sickmann, A., and Zahedi, R.P., *Systematic and quantitative comparison of digest efficiency and specificity reveals the impact of trypsin quality on MS-based proteomics.* J Proteomics, 2012. **75**(4): p. 1454-62.

634. Glatter, T., Ludwig, C., Ahrné, E., Aebersold, R., Heck, A.J.R., and Schmidt, A., *Large-scale quantitative assessment of different in-solution protein digestion protocols reveals superior cleavage efficiency of tandem Lys-C/trypsin proteolysis over trypsin digestion.* Journal of Proteome Research, 2012. **11**(11): p. 5145-5156.

635. Schmudlach, A., Felton, J., Cipolla, C., Sun, L., Kennedy, R.T., and Dovichi, N.J., *Sample preparation protocol for bottom-up proteomic analysis of the secretome of the islets of Langerhans.* Analyst, 2016. **141**(5): p. 1700-6.

636. Pan, Y., Cheng, K., Mao, J., Liu, F., Liu, J., Ye, M., and Zou, H., *Quantitative proteomics reveals the kinetics of trypsin-catalyzed protein digestion.* Anal Bioanal Chem, 2014. **406**(25): p. 6247-56.

637. Gershon, P.D., *Cleaved and missed sites for trypsin, lys-C, and lys-N can be predicted with high confidence on the basis of sequence context.* J Proteome Res, 2014. **13**(2): p. 702-9.

638. Simpson, R.J., *Fragmentation of protein using trypsin.* CSH Protoc, 2006. **2006**(5).

639. Simpson, R.J., *Cleavage at met-x bonds by cyanogen bromide.* CSH Protoc, 2007. **2007**: p. pdb prot4704.

640. Kaiser, R. and Metzka, L., *Enhancement of cyanogen bromide cleavage yields for methionyl-serine and methionyl-threonine peptide bonds.* Anal Biochem, 1999. **266**(1): p. 1-8.

641. Specht, H.V., Harmange, G., Perlman, D.H., Emmott, E., Niziolek, Z., Budnik, B., and Slavov, N. *Automated sample preparation for high-throughput single-cell proteomics*. 2018; doi: https://doi.org/10.1101/399774]. Available from: https://www.biorxiv.org/content/10.1101/399774v1.

642. Wisniewski, J.R., Zougman, A., Nagaraj, N., and Mann, M., *Universal sample preparation method for proteome analysis.* Nat Methods, 2009. **6**(5): p. 359-62.

643. Erde, J., Loo, R.R., and Loo, J.A., *Enhanced FASP (eFASP) to increase proteome coverage and sample recovery for quantitative proteomic experiments.* J Proteome Res, 2014. **13**(4): p. 1885-95.

644. Wisniewski, J.R. and Mann, M., *Consecutive proteolytic digestion in an enzyme reactor increases depth of proteomic and phosphoproteomic analysis.* Anal Chem, 2012. **84**(6): p. 2631-7.

645. Tremblay, T.L. and Hill, J.J., *Adding polyvinylpyrrolidone to low level protein samples significantly improves peptide recovery in FASP digests: An inexpensive and simple modification to the FASP protocol.* J Proteomics, 2021. **230**: p. 104000.

646. Senkevich, T.G., White, C.L., Koonin, E.V., and Moss, B., *Complete pathway for protein disulfide bond formation encoded by poxviruses.* Proc Natl Acad Sci U S A, 2002. **99**(10): p. 6667-6672.

647. Locker, J.K. and Griffiths, G., *An unconventional role for cytoplasmic disulfide bonds in vaccinia virus proteins.* Journal of Cell Biology, 1999. **144**: p. 267-269.

648. Hu, Q., Noll, R.J., Li, H., Makarov, A., Hardman, M., and Graham Cooks, R., *The Orbitrap: a new mass spectrometer.* J Mass Spectrom, 2005. **40**(4): p. 430-43.

649. Espadas, G., Borras, E., Chiva, C., and Sabido, E., *Evaluation of different peptide fragmentation types and mass analyzers in data-dependent methods using an Orbitrap Fusion Lumos Tribrid mass spectrometer.* Proteomics, 2017. **17**(9).

650. Yu, Q., Paulo, J.A., Naverrete-Perea, J., McAlister, G.C., Canterbury, J.D., Bailey, D.J., Robitaille, A.M., Huguet, R., Zabrouskov, V., Gygi, S.P., and Schweppe, D.K., *Benchmarking the Orbitrap Tribrid Eclipse for Next Generation Multiplexed Proteomics.* Anal Chem, 2020. **92**(9): p. 6478-6485.

651. He, Y., Shishkova, E., Peters-Clarke, T.M., Brademan, D.R., Westphall, M.S., Bergen, D., Huang, J., Huguet, R., Senko, M.W., Zabrouskov, V., McAlister, G.C., and Coon, J.J., *Evaluation of the Orbitrap Ascend Tribrid Mass Spectrometer for Shotgun Proteomics.* Anal Chem, 2023. **95**(28): p. 10655-10663.

652. de Godoy, L.M., Olsen, J.V., Cox, J., Nielsen, M.L., Hubner, N.C., Frohlich, F., Walther, T.C., and Mann, M., *Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast.* Nature, 2008. **455**(7217): p. 1251-4.

653. Washburn, M.P., Wolters, D., and Yates, J.R., 3rd, *Large-scale analysis of the yeast proteome by multidimensional protein identification technology.* Nat Biotechnol, 2001. **19**(3): p. 242-7.

654. de Godoy, L.M., Olsen, J.V., de Souza, G.A., Li, G., Mortensen, P., and Mann, M., *Status of complete proteome analysis by mass spectrometry: SILAC labeled yeast as a model system.* Genome Biol, 2006. **7**(6): p. R50.

655. Jumper, J., et al., *Highly accurate protein structure prediction with AlphaFold.* Nature, 2021. **596**(7873): p. 583-589.

656. Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., and Moult, J., *Critical assessment of methods of protein structure prediction (CASP)-Round XIV.* Proteins, 2021. **89**(12): p. 1607-1617.

657. Jumper, J., et al., *Applying and improving AlphaFold at CASP14.* Proteins, 2021. **89**(12): p. 1711-1721.

658. Mirzakhanyan, Y. and Gershon, P.D., *Structure-Based Deep Mining Reveals First-Time Annotations for 46 Percent of the Dark Annotation Space of the 9,671-Member Superproteome of the Nucleocytoplasmic Large DNA Viruses.* J Virol, 2020. **94**(24).

659. Kahraman, A., Herzog, F., Leitner, A., Rosenberger, G., Aebersold, R., and Malmstrom, L., *Cross-link guided molecular modeling with ROSETTA.* PLoS One, 2013. **8**(9): p. e73411.

660. Mariani, V., Biasini, M., Barbato, A., and Schwede, T., *lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests.* Bioinformatics, 2013. **29**(21): p. 2722-8.

661. Zemla, A., *LGA: A method for finding 3D similarities in protein structures.* Nucleic Acids Res, 2003. **31**(13): p. 3370-4.

662. Zhang, Y. and Skolnick, J., *TM-align: a protein structure alignment algorithm based on the TM-score.* Nucleic Acids Res, 2005. **33**(7): p. 2302-9.

663. Bacik, J.P. and Hazes, B., *Crystal structures of a poxviral glutaredoxin in the oxidized and reduced states show redox-correlated structural changes.* J Mol Biol, 2007. **365**(5): p. 1545-58.

664. Rodriguez, D., Risco, C., Rodriguez, J.R., Carrascosa, J.L., and Esteban, M., *Inducible expression of the vaccinia virus A17L gene provides a synchronized system to monitor sorting of viral proteins during morphogenesis.* J Virol, 1996. **70**(11): p. 7641-53.

665. Moss, B., *Origin of the poxviral membrane: A 50-year-old riddle.* PLoS Pathog, 2018. **14**(6): p. e1007002.

666. Liu, J., Xiao, Z., Ko, H.L., Shen, M., and Ren, E.C., *Activating killer cell immunoglobulin-like receptor 2DS2 binds to HLA-A*11.* Proc Natl Acad Sci U S A, 2014. **111**(7): p. 2662-7.

667. Mercer, J. and Traktman, P., *Investigation of structural and functional motifs within the vaccinia virus A14 phosphoprotein, an essential component of the virion membrane.* J Virol, 2003. **77**(16): p. 8857-71.

668. Garriga, D., Headey, S., Accurso, C., Gunzburg, M., Scanlon, M., and Coulibaly, F., *Structural basis for the inhibition of poxvirus assembly by the antibiotic rifampicin.* Proc Natl Acad Sci U S A, 2018. **115**(33): p. 8424-8429.

669. Salmons, T., Kuhn, A., Wylie, F., Schleich, S., Rodriguez, J.R., Rodriguez, D., Esteban, M., Griffiths, G., and Locker, J.K., *Vaccinia virus membrane proteins p8 and p16 are cotranslationally inserted into the rough endoplasmic reticulum and retained in the intermediate compartment.* J Virol, 1997. **71**(10): p. 7404-20.

670. McKelvey, T.A., Andrews, S.C., Miller, S.E., Ray, C.A., and Pickup, D.J., *Identification of the orthopoxvirus p4c gene, which encodes a structural protein that directs intracellular mature virus particles into A-type inclusions.* J Virol, 2002. **76**(22): p. 11216-25.

671. Kastenmayer, R.J., Maruri-Avidal, L., Americo, J.L., Earl, P.L., Weisberg, A.S., and Moss, B., *Elimination of A-type inclusion formation enhances cowpox virus replication in mice: implications for orthopoxvirus evolution.* Virology, 2014. **452-453**: p. 59-66.

672. de Carlos, A. and Paez, E., *Isolation and characterization of mutants of vaccinia virus with a modified 94-kDa inclusion protein.* Virology, 1991. **185**(2): p. 768-78.

673. Ulaeto, D., Grosenbach, D., and Hruby, D.E., *The vaccinia virus 4c and A-type inclusion proteins are specific markers for the intracellular mature virus particle.* J Virol, 1996. **70**(6): p. 3372-7.

674. Holm, L., *Dali server: structural unification of protein families.* Nucleic Acids Res, 2022. **50**(W1): p. W210-5.

675. Ojeda, S., Domi, A., and Moss, B., *Vaccinia virus G9 protein is an essential component of the poxvirus entry-fusion complex.* J Virol, 2006. **80**(19): p. 9822-30.

676. Cotter, C.A. and Moss, B., *Mutations Near the N Terminus of Vaccinia Virus G9 Protein Overcome Restrictions on Cell Entry and Syncytium Formation Imposed by the A56/K2 Fusion Regulatory Complex.* J Virol, 2020. **94**(10).

677. Zajac, P., Spehner, D., and Drillien, R., *The vaccinia virus J5L open reading frame encodes a polypeptide expressed late during infection and required for viral multiplication.* Virus Res, 1995. **37**(2): p. 163-73.

678. Izmailyan, R.A., Huang, C.Y., Mohammad, S., Isaacs, S.N., and Chang, W., *The envelope G3L protein is essential for entry of vaccinia virus into host cells.* J Virol, 2006. **80**(17): p. 8402-10.

679. Wu, D., Lou, Y.C., Chang, W., and Tzou, D.M., *NMR assignments of vaccinia virus protein A28: an entry-fusion complex component.* Biomol NMR Assign, 2021. **15**(1): p. 117-120.

680. Wolfe, C.L. and Moss, B., *Interaction between the G3 and L5 proteins of the vaccinia virus entry-fusion complex.* Virology, 2011. **412**(2): p. 278-83.

681. Hong, G.C., Tsai, C.H., and Chang, W., *Experimental Evolution To Isolate Vaccinia Virus Adaptive G9 Mutants That Overcome Membrane Fusion Inhibition via the Vaccinia Virus A56/K2 Protein Complex.* J Virol, 2020. **94**(10).

682. Fox, N.K., Brenner, S.E., and Chandonia, J.M., *SCOPe: Structural Classification of Proteins--extended, integrating SCOP and ASTRAL data and classification of new structures.* Nucleic Acids Res, 2014. **42**(Database issue): p. D304-9.

683. Satheshkumar, P.S. and Moss, B., *Sequence-divergent chordopoxvirus homologs of the o3 protein maintain functional interactions with components of the vaccinia virus entry-fusion complex.* J Virol, 2012. **86**(3): p. 1696-705.

684. Yeh, W.W., Moss, B., and Wolffe, E.J., *The vaccinia virus A9L gene encodes a membrane protein required for an early step in virion morphogenesis.* J Virol, 2000. **74**(20): p. 9701-11.

685. Sood, C.L., Ward, J.M., and Moss, B., *Vaccinia virus encodes I5, a small hydrophobic virion membrane protein that enhances replication and virulence in mice.* J Virol, 2008. **82**(20): p. 10071-8.

686. Unger, B., Nichols, R.J., Stanitsa, E.S., and Traktman, P., *Functional characterization of the vaccinia virus I5 protein.* Virol J, 2008. **5**: p. 148.

687. Kryshtafovych, A., et al., *Computational models in the service of X-ray and cryo-electron microscopy structure determination.* Proteins, 2021. **89**(12): p. 1633-1646.

688. Goddard, T.D., Huang, C.C., Meng, E.C., Pettersen, E.F., Couch, G.S., Morris, J.H., and Ferrin, T.E., *UCSF ChimeraX: Meeting modern challenges in visualization and analysis.* Protein Sci, 2018. **27**(1): p. 14-25.

689. Pettersen, E.F., Goddard, T.D., Huang, C.C., Meng, E.C., Couch, G.S., Croll, T.I., Morris, J.H., and Ferrin, T.E., *UCSF ChimeraX: Structure visualization for researchers, educators, and developers.* Protein Sci, 2021. **30**(1): p. 70-82.

690. Cotter, C.A., Earl, P.L., Wyatt, L.S., and Moss, B., *Preparation of Cell Cultures and Vaccinia Virus Stocks.* Curr Protoc Mol Biol, 2017. **117**: p. 16 16 1-16 16 18.

691. Sievers, F. and Higgins, D.G., *Clustal Omega for making accurate alignments of many protein sequences.* Protein Sci, 2018. **27**(1): p. 135-145.

692. Rodriguez, D., Barcena, M., Mobius, W., Schleich, S., Esteban, M., Geerts, W.J., Koster, A.J., Griffiths, G., and Locker, J.K., *A vaccinia virus lacking A10L: viral core proteins accumulate on structures derived from the endoplasmic reticulum.* Cell Microbiol, 2006. **8**(3): p. 427-37.

693. Williams, O., Wolffe, E.J., Weisberg, A.S., and Merchlinsky, M., *Vaccinia Virus WR Gene A5L Is Required for Morphogenesis of Mature Virions.* Journal of Virology, 1999. **73**: p. 4590-4599.

694. Yang, S.J., *Characterization of vaccinia virus A12L protein proteolysis and its participation in virus assembly.* Virol J, 2007. **4**: p. 78.

695. Katz, E. and Moss, B., *Formation of a vaccinia virus structural polypeptide from a higher molecular weight precursor: inhibition by rifampicin.* Proc Natl Acad Sci U S A, 1970. **66**(3): p. 677-84.

696. Moss, B. and Rosenblum, E.N., *Letter: Protein cleavage and poxvirus morphogenesis: tryptic peptide analysis of core precursors accumulated by blocking assembly with rifampicin.* J Mol Biol, 1973. **81**(2): p. 267-9.

697. Miner, J.N. and Hruby, D.E., *Rifampicin prevents virosome localization of L65, an essential vaccinia virus polypeptide.* Virology, 1989. **170**(1): p. 227-37.

698. Villarreal, E.C., Roseman, N.A., and Hruby, D.E., *Isolation of vaccinia virus mutants capable of replicating independently of the host cell nucleus.* J Virol, 1984. **51**(2): p. 359-66.

699. Child, S.J., Franke, C.A., and Hruby, D.E., *Inhibition of vaccinia virus replication by nicotinamide: evidence for ADP-ribosylation of viral proteins.* Virus Res, 1988. **9**(2-3): p. 119-32.

700. Senkevich, T., White, C., Weisberg, A., Granek, J., Wolffe, E., Koonin, E., and Moss, B., *Expression of the vaccinia virus A2.5L redox protein is required for virion morphogenesis.* Virology, 2002. **300**(2): p. 296-303.

701. White, C.L., Senkevich, T.G., and Moss, B., *Vaccinia virus G4L glutaredoxin is an essential intermediate of a cytoplasmic disulfide bond pathway required for virion assembly.* Journal of Virology, 2002. **76**: p. 467-472.

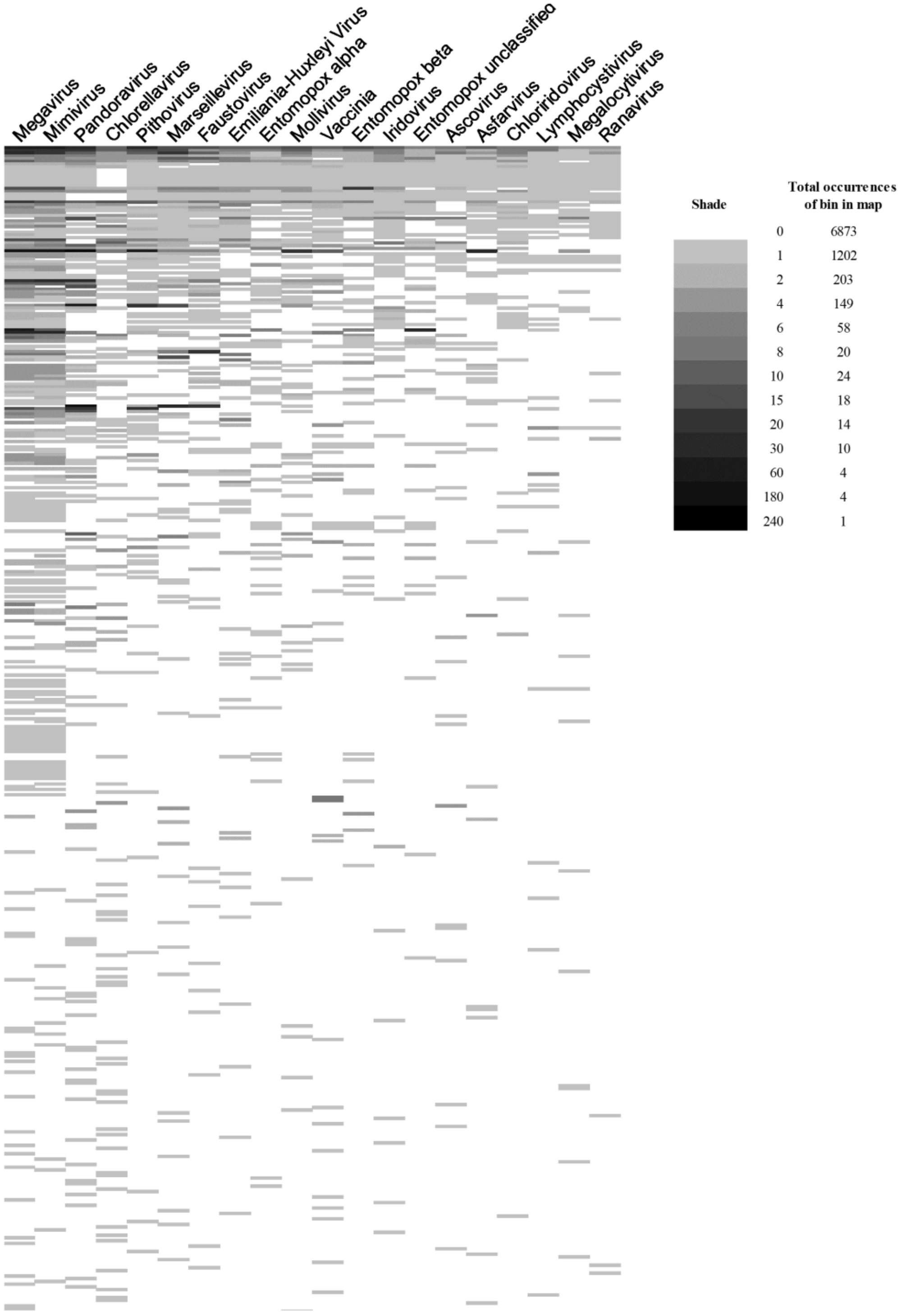
702. White, C.L., Weisberg, A.S., and Moss, B., *A glutaredoxin, encoded by the G4L gene of vaccinia virus, is essential for virion morphogenesis.* J Virol, 2000. **74**(19): p. 9175-83.

703. Hyun, J., Matsunami, H., Kim, T.G., and Wolf, M., *Assembly mechanism of the pleomorphic immature poxvirus scaffold.* Nat Commun, 2022. **13**(1): p. 1704.

704. Lawrence, R.M., et al., *Serial femtosecond X-ray diffraction of enveloped virus microcrystals.* Struct Dyn, 2015. **2**(4): p. 041720.

705. Kyrieleis, O.J., Chang, J., de la Pena, M., Shuman, S., and Cusack, S., *Crystal structure of vaccinia virus mRNA capping enzyme provides insights into the mechanism and evolution of the capping apparatus.* Structure, 2014. **22**(3): p. 452-65.

Appendix 1. Figure 1. All-superfamily heatmap. Superfamilies were ranked by number of viruses in which matches were found to the superfamily (high to low), then by total number of protein matches within each #viruses block. Viruses are sorted (left to right) by total superfamilies. The sum of all matches in the heatmap was 6873.



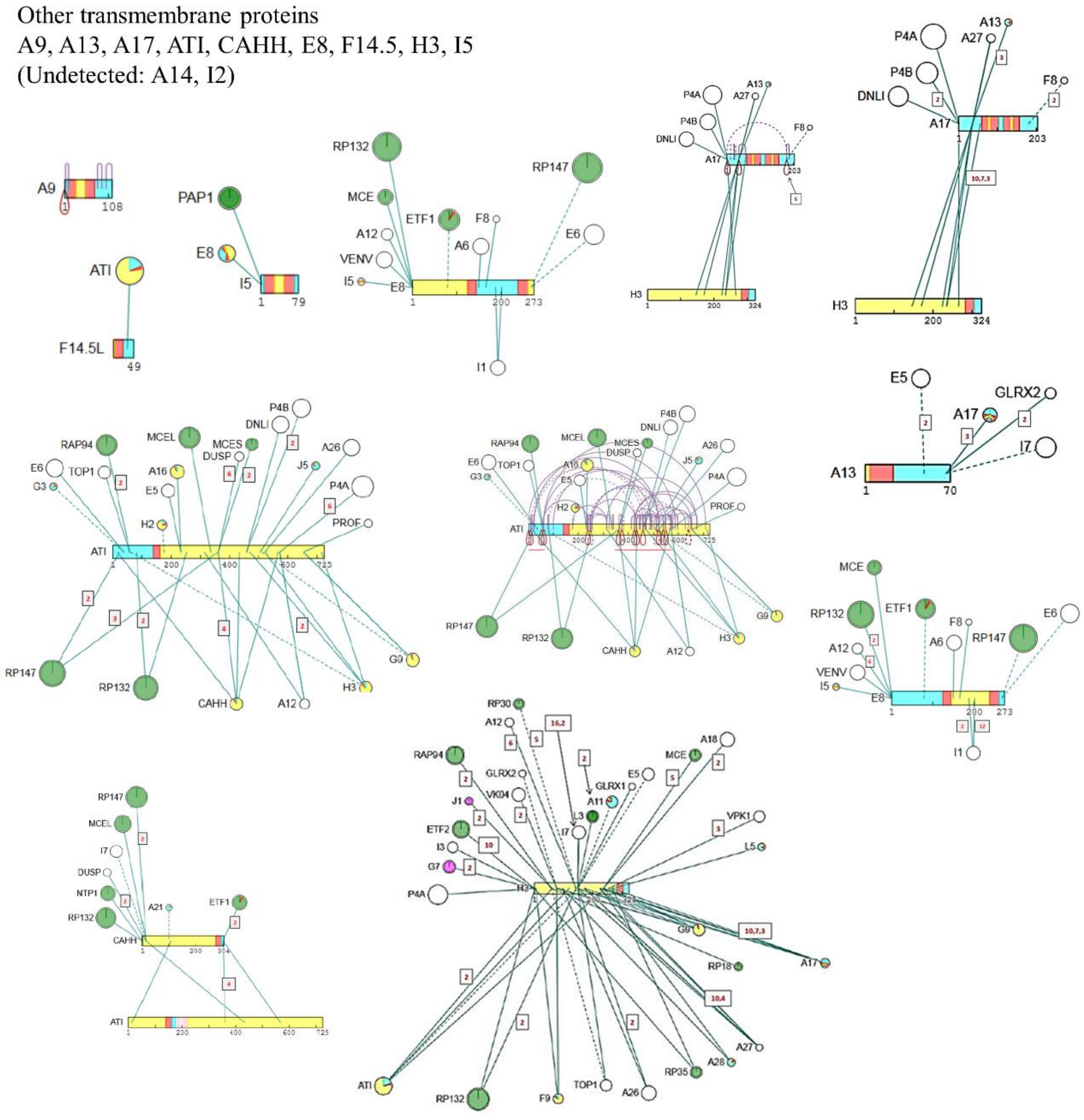
Appendix 1. Table 1. Endonuclease genes among the 20 NCLDV

Virus	Gene	Protein Annotation	HHsearch Annotations	Endonuclease type
Ascovirus	K4NYD4_9VIRU	Uncharacterized protein	5hmC endonuclease	Type IV
Entomopox beta	R4ZF06_CBEPV	N1R/p28-like protein	5hmC endonuclease	Type IV
Entomopox unclassified	Q9YVL3_MSEPV	ORF MSV229 leucine rich repeat gene family protein	5hmC endonuclease	Type IV
Entomopox unclassified	Q9YVN8_MSEPV	ORF MSV204 ALI motif gene family protein	5hmC endonuclease	Type IV
Entomopox unclassified	Q9YVP6_MSEPV	ORF MSV196 ALI motif gene family protein	5hmC endonuclease	Type IV
Entomopox unclassified	Q9YW64_MSEPV	Uncharacterized protein MSV028	5hmC endonuclease	Type IV
Entomopox unclassified	Q9YW67_MSEPV	ORF MSV026 ALI motif gene family protein	5hmC endonuclease	Type IV
Entomopox unclassified	Q9YW68_MSEPV	ORF MSV024 ALI motif gene family protein	5hmC endonuclease	Type IV
Faustovirus	A0A0H3TM84_9VIRU	Uncharacterized protein	5hmc Endonuclease	Type IV
Iridovirus	069L_IIV6	Putative Bro-N domain-containing protein 069L	5hmc endonuclease	Type IV
Megavirus	K7YIM3_9VIRU	Uncharacterized protein	5hmC endonuclease	Type IV
Ranavirus	094L_FRG3G	Uncharacterized protein 094L	5hmC endonuclease	Type IV
Asfarvirus	E0WMJ7_ASF	EP364R	XPF endonuclease	PDDEXK
Chlorella virus	Q84474_PBCV1	Uncharacterized protein	I-Bth0305I homing endonuclease catalytic domain	PDDEXK
Chlorella virus	Q84449_PBCV1	Uncharacterized protein	I-Bth0305I homing endonuclease catalytic domain	PDDEXK
Chloriridovirus	VF307_IIV3	Uncharacterized protein 033L	VSR endoncuclease	PDDEXK
Emiliania huxleyi virus	Q4A3A9_EHV8U	Uncharacterized protein	VSR endonuclease + rubredoxin or RPB12 like domain	PDDEXK
Faustovirus	A0A0H3TLX1_9VIRU	Uncharacterized protein	I-Bth0305I homing endonuclease catalytic domain	PDDEXK
Faustovirus	A0A0H3TLI2_9VIRU	Uncharacterized protein	I-Bth0305I homing endonuclease catalytic domain	PDDEXK
Faustovirus	A0A0H3TLQ9_9VIRU	Uncharacterized protein	VSR endonuclease	PDDEXK
Iridovirus	VF307_IIV6	Uncharacterized protein 307L	VSR endonuclease	PDDEXK
Lymphocystivirus	Q677W0_9VIRU	Uncharacterized protein	VSR endonuclease	PDDEXK
Marseillevirus	D2XB63_GBMV	Vsr/MutH/archaeal HJR family endonuclease	endonuclease I T7	PDDEXK
Marseillevirus	D2XA52_GBMV	Restriction endonuclease	RPB12 zinc finger + VSR endonuclease	PDDEXK
Marseillevirus	D2XAG0_GBMV	Restriction endonuclease	RPB12 zinc finger + VSR endonuclease	PDDEXK
Marseillevirus	D2XAP5_GBMV	Restriction endonuclease	RPB12 zinc finger + VSR endonuclease	PDDEXK
Marseillevirus	D2XAW9_GBMV	Restriction endonuclease	RPB12 zinc finger + VSR endonuclease	PDDEXK
Marseillevirus	D2XAY9_GBMV	Restriction endonuclease	RPB12 zinc finger + VSR endonuclease	PDDEXK
Marseillevirus	D2XB44_GBMV	Restriction endonuclease	RPB12 zinc finger + VSR endonuclease	PDDEXK
Marseillevirus	D2XAT6_GBMV	Putative nuclease	VSR endonuclease	PDDEXK
Megalocytivirus	Q8QUM4_ISKNN	ORF086L	VSR endonuclease	PDDEXK
Mimivirus	F8V6I3_MIMIV	Uncharacterized protein R641	PDDEXK endonuclease	PDDEXK
Pandoravirus	A0A0B5IYU7_9VIRU	Uncharacterized protein	PDDEXK endonuclease	PDDEXK
Pandoravirus	A0A0B5J343_9VIRU	GRF zinc finger motif-containing protein	PDDEXK endonuclease	PDDEXK
Pandoravirus	A0A0B5JB16_9VIRU	Uncharacterized protein	PDDEXK endonuclease	PDDEXK
Pithovirus	W5S678_9VIRU	Group I intron putative endonuclease	RPB12 zinc finger + VSR endonuclease	PDDEXK
Pithovirus	W5S5L2_9VIRU	Helicase nuclease	VSR endonuclease	PDDEXK
Megavirus	K7YID2_9VIRU	Putative KilA-N domain-containing protein	KilA-N + NucS C-terminal catalytic domain	MMR
Megavirus	K7Z8Z6_9VIRU	Putative KilA-N domain-containing protein	KilA-N + NucS C-terminal catalytic domain	MMR
Mimivirus	A0A0G2Y6T8_MIMIV	Putative KilA-N domain-containing protein	KilA-N + NucS C-terminal catalytic domain	MMR
Mimivirus	A0A0G2Y7Z1_MIMIV	Putative Kila-N domain-containing protein	KilA-N + NucS C-terminal catalytic domain	MMR
Mimivirus	A0A0G2Y4C6_MIMIV	Putative Kila-N domain-containing protein	KilA-N + EndoMS c-terminal catalytic domain	MMR
Ascovirus	K4NXW6_9VIRU	Uncharacterized protein	DNA/RNA non-specific endonuclease	HNH
Ascovirus	K4P9N6_9VIRU	Uncharacterized protein	DNA/RNA non-specific endonuclease	HNH
Chlorella virus	Q84554_PBCV1	Uncharacterized protein	HNH endonuclease	HNH

Chlorella virus	Q84668_PBCV1	Uncharacterized protein	I-HmuI homing endonuclease (two I-HmuI domains)	HNH
Chlorella virus	Q98474_PBCV1	Uncharacterized protein	I-HmuI homing endonuclease (two I-HmuI domains)	HNH
Chlorella virus	Q84408_PBCV1	Uncharacterized protein	NUMOD4 + HNH homing endonuclease	HNH
Chlorella virus	Q84584_PBCV1	Uncharacterized protein	PacIR endonuclease	HNH
Chlorella virus	Q98528_PBCV1	Uncharacterized protein	PacIR endonuclease (two PacIR domains)	HNH
Chlorella virus	Q98540_PBCV1	Uncharacterized protein	PacIR endonuclease (two PacIR domains)	HNH
Emiliania huxleyi virus	Q4A2Z6_EHV8U	Uncharacterized protein	HNH endonuclease	HNH
Emiliania huxleyi virus	Q4A3C0_EHV8U	Uncharacterized protein	HNH endonuclease	HNH
Emiliania huxleyi virus	Q4A340_EHV8U	HNH endonuclease family protein	HNH endonuclease	HNH
Emiliania huxleyi virus	Q4A346_EHV8U	Uncharacterized protein	I-HmuI homing endonuclease (two I-HmuI domains)	HNH
Emiliania huxleyi virus	Q4A2V0_EHV8U	Putative DNA-binding protein	I-HmuI homing endonuclease (two I-HmuI domains)	HNH
Entomopox beta	R4ZFH0_CBEPV	Uncharacterized protein	DNA/RNA non-specific endonuclease	HNH
Entomopox beta	R4ZDL8_CBEPV	Uncharacterized protein	DNA/RNA non-specific endonuclease	HNH
Entomopox beta	R4ZF90_CBEPV	Uncharacterized protein	DNA/RNA non-specific endonuclease	HNH
Faustovirus	A0A0H3TMV2_9VIRU	Uncharacterized protein	PacIR endonuclease	HNH
Marseillevirus	D2XAU1_GBMV	HNH-family endonuclease	AP2 domain + HNH endonuclease + AP2 domain	HNH
Marseillevirus	D2XB23_GBMV	HNH endonuclease	HNH endonuclease	HNH
Megavirus	K7Y8Q0_9VIRU	Uncharacterized protein	HNH endonuclease	HNH
Megavirus	K7YH83_9VIRU	Uncharacterized protein	HNH endonuclease	HNH
Megavirus	K7Y946_9VIRU	Uncharacterized protein	HNH endonuclease + PacIR endonuclease	HNH
Megavirus	K7YEM3_9VIRU	Putative intron HNH endonuclease	I-HmuI homing endonuclease (two I-HmuI domains)	HNH
Megavirus	K7YVN2_9VIRU	Putative intron encoded HNH endonuclease	I-HmuI homing endonuclease (two I-HmuI domains)	HNH
Megavirus	K7YGX2_9VIRU	Putative HNH endonuclease	I-Hmul HNH homing endonuclease	HNH
Megavirus	K7YHH0_9VIRU	Putative HNH endonuclease	I-Hmul HNH homing endonuclease	HNH
Megavirus	K7Z7Q9_9VIRU	Putative intron encoded HNH endonuclease	I-Hmul HNH homing endonuclease	HNH
Megavirus	K7YF08_9VIRU	Putative intron encoded nuclease	PacIR endonuclease	HNH
Megavirus	K7YW13_9VIRU	Putative intron encoded nuclease	PacIR endonuclease	HNH
Mimivirus	A0A0G2Y7N7_MIMIV	Uncharacterized protein R328	HNH endonuclease	HNH
Mimivirus	A0A0G2Y929_MIMIV	Uncharacterized protein R424	HNH endonuclease	HNH
Mimivirus	A0A0G2Y7F9_MIMIV	Uncharacterized HNH endonuclease	I-HmuI homing endonuclease (two I-HmuI domains)	HNH
Mimivirus	A0A0G2YCQ8_MIMIV	Uncharacterized HNH endonuclease	I-HmuI homing endonuclease (two I-HmuI domains)	HNH
Mimivirus	A0A0G2Y5U3_MIMIV	Uncharacterized HNH endonuclease	I-Hmul HNH homing endonuclease	HNH
Mimivirus	A0A0G2Y3X0_MIMIV	Putative nuclease	PacIR endonuclease	HNH
Mimivirus	A0A0G2Y5H6_MIMIV	Uncharacterized protein R423	PacIR endonuclease	HNH
Mimivirus	A0A0G2Y097_MIMIV	Putative prophage protein	zinc binding + PacIR endonuclease	HNH
Pandoravirus	A0A0B5IZ79_9VIRU	Uncharacterized protein	PacIR endonuclease	HNH
Pandoravirus	A0A0B5J5M8_9VIRU	Uncharacterized protein	PacIR endonuclease	HNH
Ascovirus	K4NW91_9VIRU	Bro25	GIY-YIG	GIY-YIG
Ascovirus	K4NXX1_9VIRU	Bro1	GIY-YIG	GIY-YIG
Ascovirus	K4NY04_9VIRU	Bro4	GIY-YIG	GIY-YIG
Ascovirus	K4NYK4_9VIRU	Bro24	GIY-YIG	GIY-YIG
Ascovirus	K4P976_9VIRU	Bro3	GIY-YIG	GIY-YIG
Ascovirus	K4P9M5_9VIRU	Uncharacterized protein	GIY-YIG	GIY-YIG
Chlorella virus	Q66213_PBCV1	Uncharacterized protein	GIY-YIG	GIY-YIG
Chlorella virus	Q84430_PBCV1	Uncharacterized protein	GIY-YIG	GIY-YIG
Chlorella virus	Q84454_PBCV1	Uncharacterized protein	GIY-YIG	GIY-YIG
Chlorella virus	Q98431_PBCV1	Uncharacterized protein	GIY-YIG	GIY-YIG
Chlorella virus	Q99169_PBCV1	Uncharacterized protein	GIY-YIG	GIY-YIG
Chlorella virus	Q84665_PBCV1	Uncharacterized protein	GIY-YIG + CENP-B N-terminal DNA-binding domain (homeodomain-like) + Tc5 transposase DNA-binding domain	GIY-YIG
Chlorella virus	O41133_PBCV1	Uncharacterized protein	GIY-YIG I-Tevl homing endonuclease (full length protein, including NUMOD3 domain)	GIY-YIG

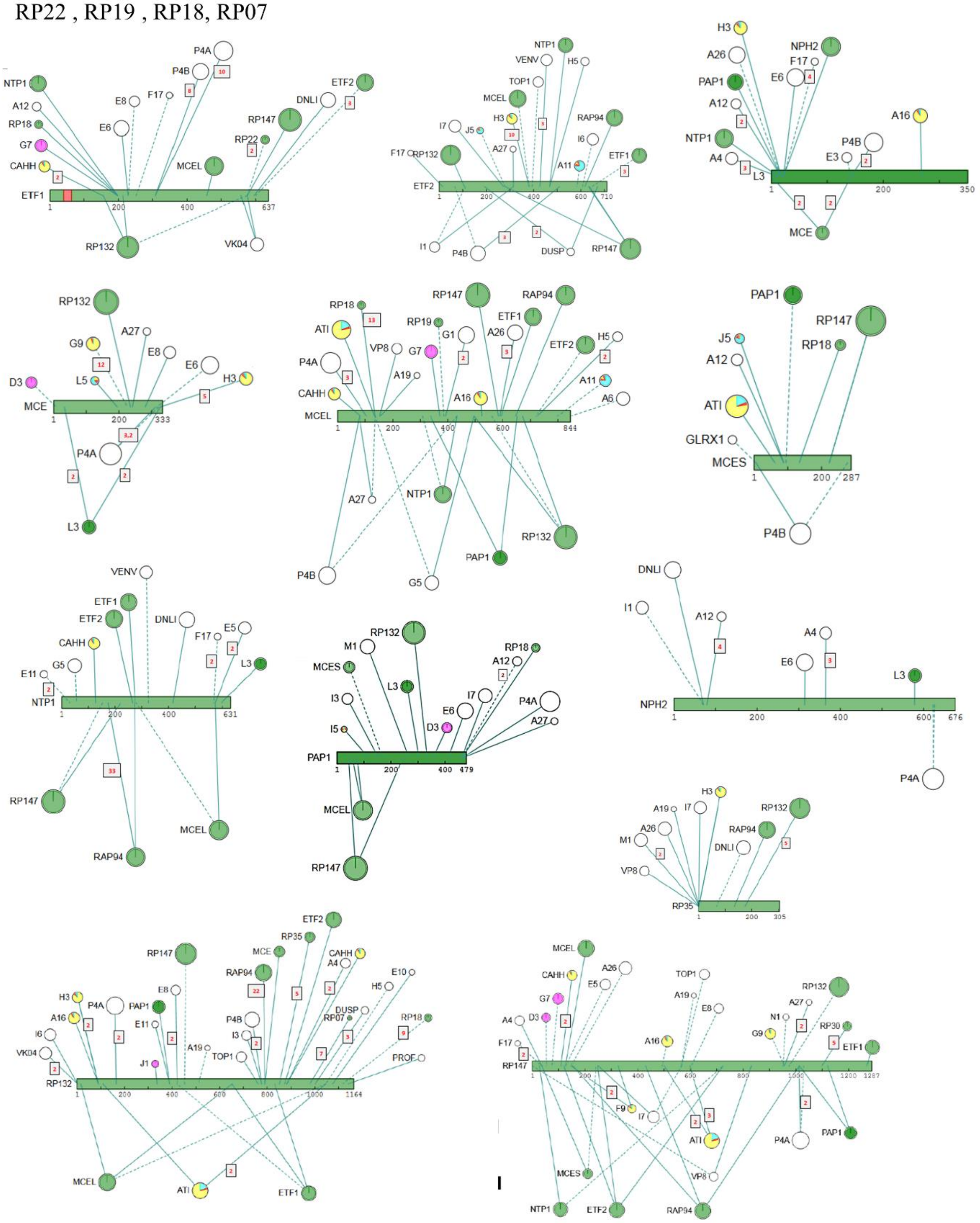
Chlorella virus	Q84603_PBCV1	Uncharacterized protein	GIY-YIG I-Tevl homing endonuclease (full length protein, including NUMOD3 domain)	GIY-YIG
Chlorella virus	Q89820_PBCV1	Uncharacterized protein	GIY-YIG I-Tevl homing endonuclease (full length protein, including NUMOD3 domain)	GIY-YIG
Chlorella virus	Q98545_PBCV1	Uncharacterized protein	GIY-YIG I-Tevl homing endonuclease (full length protein, including NUMOD3 domain)	GIY-YIG
Chloriridovirus	VF201_IIV3	Putative Bro-N domain-containing protein 019R	GIY-YIG	GIY-YIG
Emiliania huxleyi virus	Q4A392_EHV8U	Putative endonuclease	GIY-YIG	GIY-YIG
Entomopox alpha	W6JIM3_9POXV	ALI motif gene family protein	GIY-YIG	GIY-YIG
Entomopox unclassified	Q9YVP4_MSEPV	ORF MSV198 MTG motif gene family protein	GIY-YIG	GIY-YIG
Entomopox unclassified	Q9YVP8_MSEPV	ORF MSV194 ALI motif gene family protein	GIY-YIG	GIY-YIG
Faustovirus	A0A0H3TP79_9VIRU	GIY-YIG catalytic domain-containing endonuclease	GIY-YIG	GIY-YIG
Faustovirus	A0A0H3TMT7_9VIRU	GIY-YIG catalytic domain-containing protein	GIY-YIG + AP2	GIY-YIG
Iridovirus	146R_IIV6	Putative MSV199 domain-containing protein 146R	GIY-YIG	GIY-YIG
Iridovirus	242L_IIV6	Putative GIY-YIG domain-containing protein 242L	GIY-YIG	GIY-YIG
Iridovirus	VF201_IIV6	Putative Bro-N domain-containing protein 201R	GIY-YIG	GIY-YIG
Iridovirus	VF019_IIV6	Uncharacterized protein 019R	GIY-YIG	GIY-YIG
Iridovirus	460R_IIV6	Uncharacterized protein 460R	GIY-YIG + Bro-C	GIY-YIG
Iridovirus	315L_IIV6	Putative KilA-N domain-containing protein 315L	KilA-N + GIY-YIG	GIY-YIG
Marseillevirus	D2XAI4_GBMV	Uncharacterized protein	GIY-YIG	GIY-YIG
Marseillevirus	D2XAW7_GBMV	Uncharacterized protein	GIY-YIG + HIT zinc finger	GIY-YIG
Marseillevirus	D2XA40_GBMV	Uncharacterized protein	Helicase associated domain + giy-yig	GIY-YIG
Megavirus	K7Y8T8_9VIRU	Putative intron encoded endonuclease	GIY-YIG	GIY-YIG
Megavirus	K7YFF5_9VIRU	Putative endo/excinuclease amino terminal domain protein	GIY-YIG	GIY-YIG
Mimivirus	A0A0G2Y1Y3_MIMIV	Uncharacterized protein L5	GIY-YIG	GIY-YIG
Mimivirus	A0A0G2Y3M7_MIMIV	Uncharacterized endo/excinuclease amino terminal domain protein	GIY-YIG	GIY-YIG
Mimivirus	A0A0G2Y597_MIMIV	Uncharacterized protein R9	GIY-YIG	GIY-YIG
Mimivirus	A0A0G2YDI2_MIMIV	Putative intron encoded endonuclease	GIY-YIG	GIY-YIG
Mollivirus	A0A0M4JAS0_9VIRU	Putative GIY-YIG endonuclease	GIY-YIG	GIY-YIG
Pithovirus	W5S5I6_9VIRU	Group I intron GIY-YIG endonuclease	GIY-YIG	GIY-YIG
Pandoravirus	A0A0B5J8X8_9VIRU	Uncharacterized protein	Mitochondrial Holliday junction resolvase Ydc2	Non-PD-(D/E)xK DNA repair/genome stability
Asfarvirus	E0WM83_ASF	EP296R	Endonuclease IV	Non-PD-(D/E)xK DNA repair/genome stability
Emiliania huxleyi virus	Q4A2Q2_EHV8U	Putative endonuclease	Endonuclease V	Non-PD-(D/E)xK DNA repair/genome stability
Emiliania huxleyi virus	Q4A3A7_EHV8U	Putative endonuclease	Flap endonuclease	Non-PD-(D/E)xK DNA repair/genome stability
Iridovirus	273R_IIV6	Uncharacterized protein 273R	Holliday Junction resolvase	Non-PD-(D/E)xK DNA repair/genome stability
Lymphocystivirus	Q677U4_9VIRU	Uncharacterized protein	Flap endonuclease	Non-PD-(D/E)xK DNA repair/genome stability
Megavirus	K7YFI7_9VIRU	Uncharacterized protein	Holliday Junction resolvase	Non-PD-(D/E)xK DNA repair/genome stability
Megavirus	K7Z7Y8_9VIRU	Uncharacterized protein	Holliday Junction resolvase	Non-PD-(D/E)xK DNA repair/genome stability
Mimivirus	A0A0G2Y961_MIMIV	Uncharacterized protein L451	Holliday Junction resolvase	Non-PD-(D/E)xK DNA repair/genome stability
Pithovirus	W5S5K8_9VIRU	Uncharacterized protein	Holliday Junction resolvase	Non-PD-(D/E)xK DNA repair/genome stability

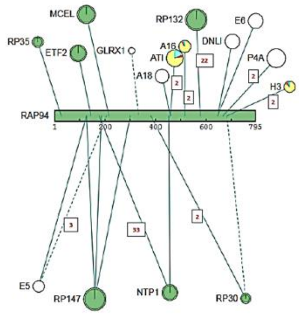
Appendix 2**. Figure 1. Crosslinked partners (circles whose areas correspond to chain length) for each protein (rectangle whose length corresponds to chain length) in the crosslink dataset, one protein per page.** Above the target protein are partners with a single XL, and below are those with two or more distinct crosslinks. Green fill: Proteins in the transcriptosome group. Magenta fill: Proteins in the '7PC' group. TM proteins—yellow, red and cyan fill: 'Outside', TM and 'inside' domains, respectively. Green lines: Inter-protein XL. Numbers (red font) in gray squares with black border: DFscore (crosslinks showing no DFscore had a DFscore of 1). Dashed crosslinks are 'ambiguous', ie. they are members of a group of distinct crosslinked peptide pairs whose experimental ion masses differed by less than the annealing mass tolerance. Although the isomeric group members likely represent distinct crosslinks (with distinct sequencing ions), they are flagged here simply to note that they would not be discriminatable on the basis of parental ion mass alone. Some dotted ('ambig') XL overlay non-dotted ones that are inapparent. For proteins A9, A46, D2 and SODL, for which no XL partners were found, intra-protein XL are shown.

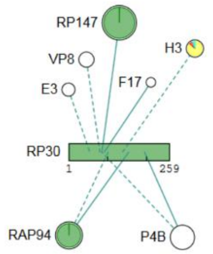


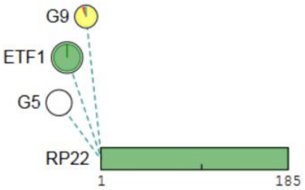
Other transmembrane proteins
A9, A13, A17, ATI, CAHH, E8, F14.5, H3, I5
(Undetected: A14, I2)

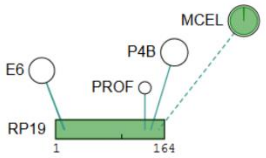
Transcriptosome (mRNA biogenesis) proteins
ETF1, ETF2, L3, MCE, MCEL, MCES, NTP1, PAP1, NPH2, RP147, RP132, RAP94 , RP35 , RP30 , RP22 , RP19 , RP18, RP07

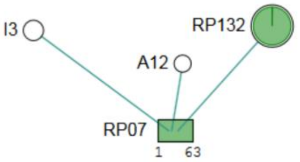


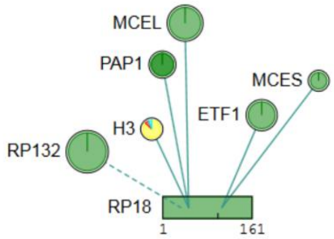




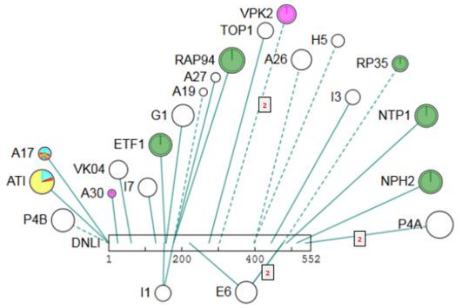


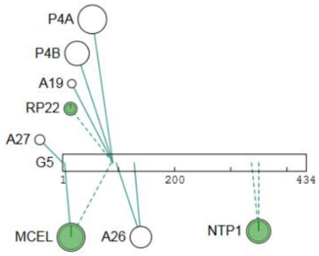


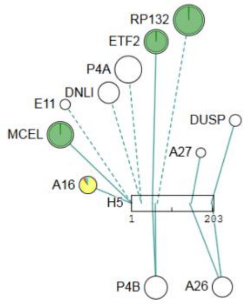


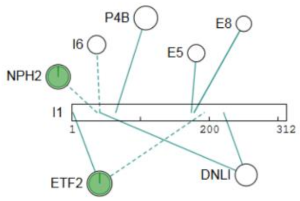


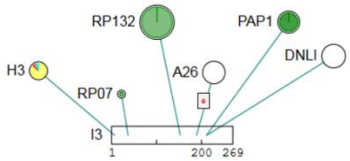
DNA-related proteins: DNLI, G5, H5, I1, I3, I6, K4, TOP1, VP8

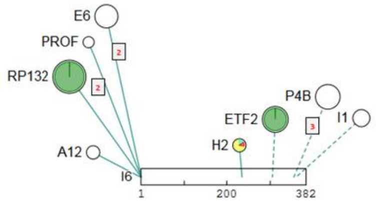


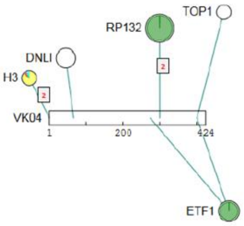


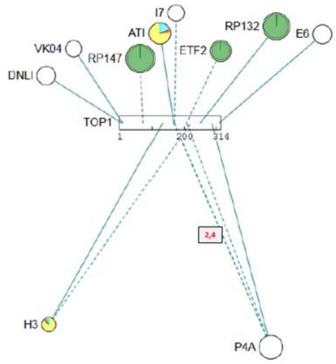


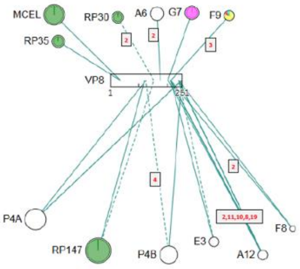




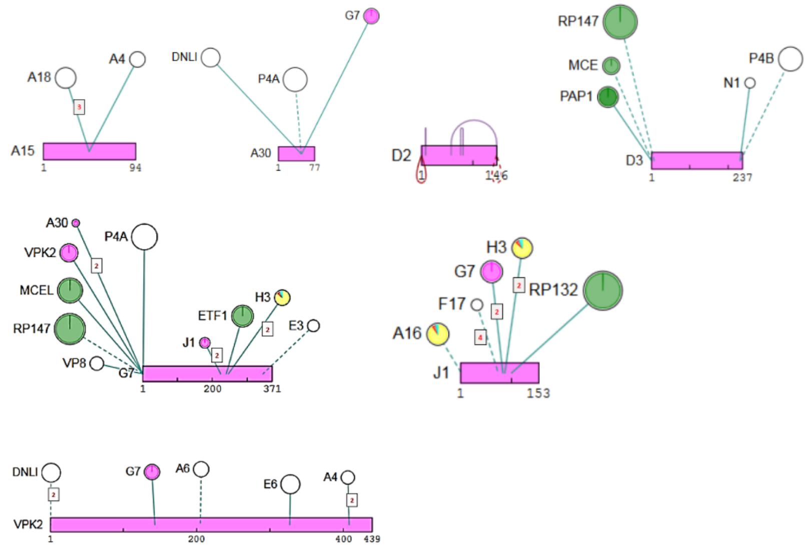




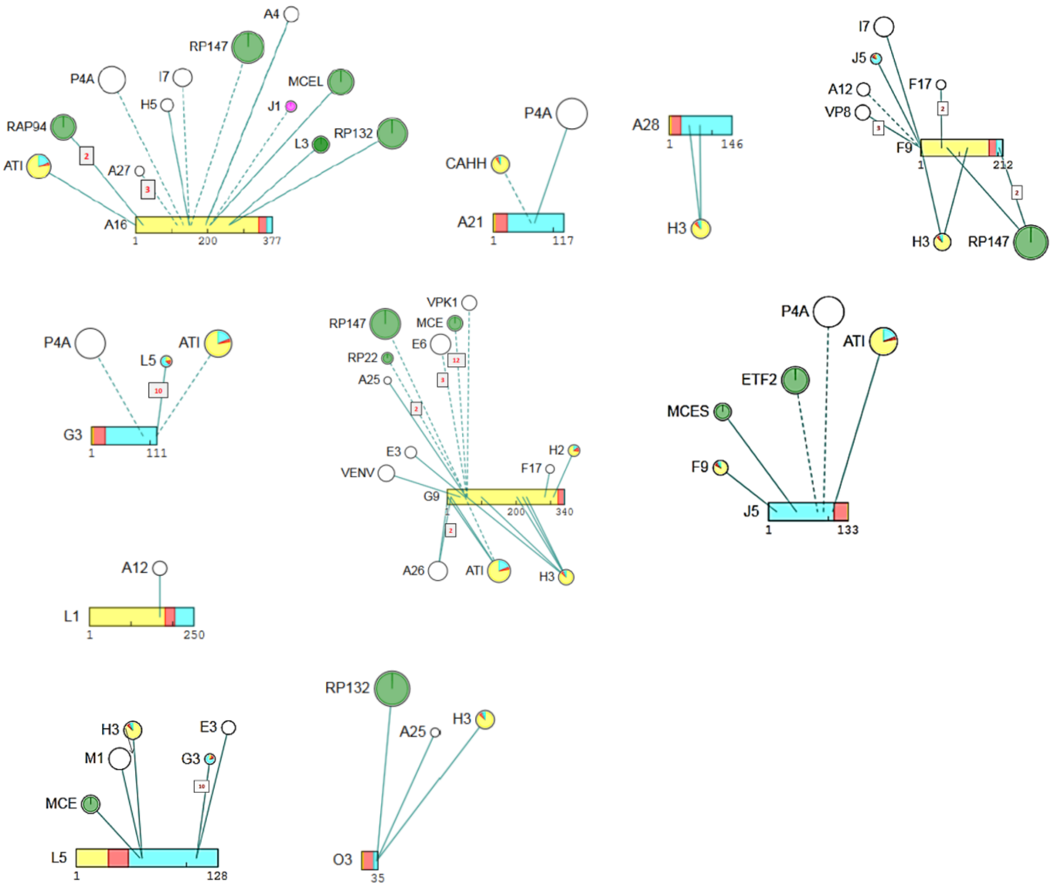




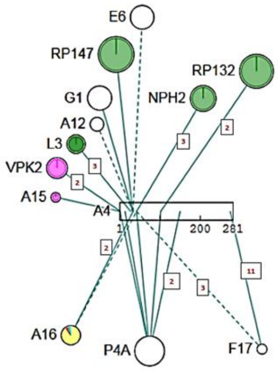
7PC: A15, A30 , D2, D3, G7, J1, VPK2

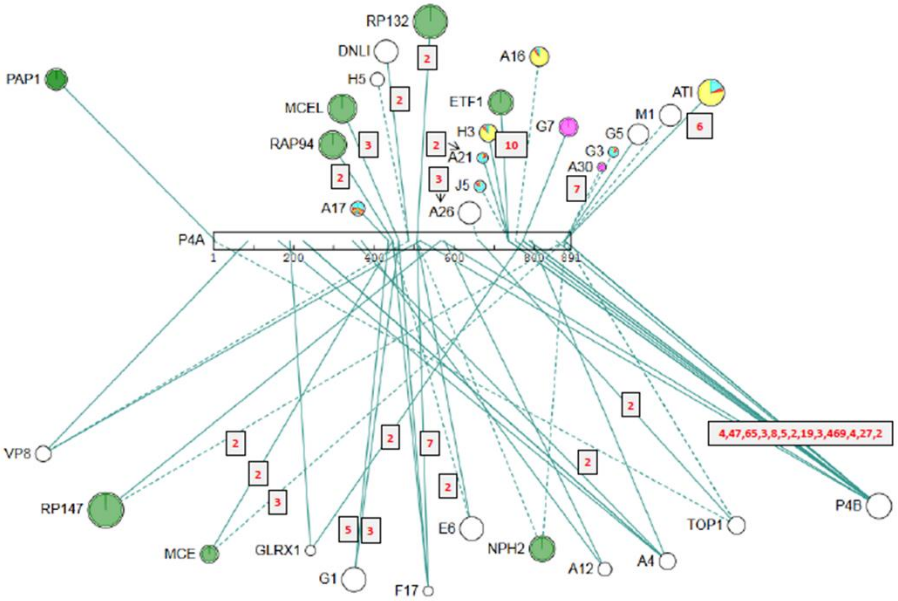


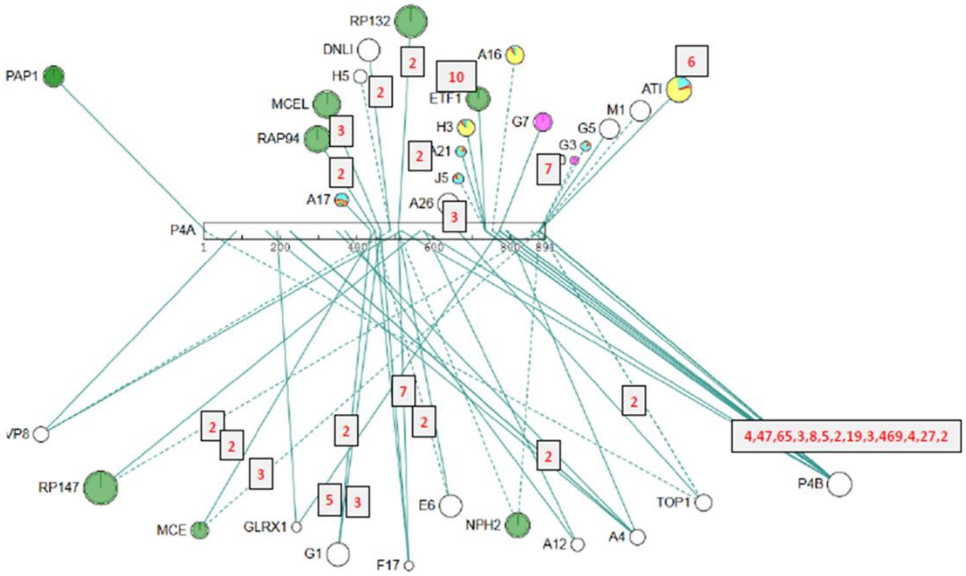
EFC/EFC-associated: A16, A21, A28, F9, G3, G9, H2, J5, L1, L5, O3



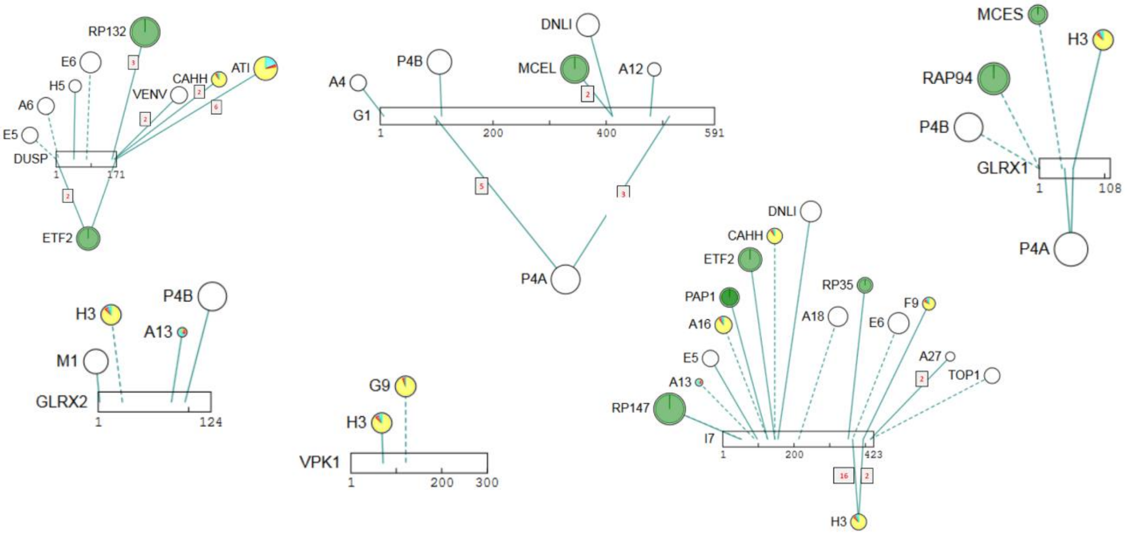
‘Structural’ proteins: A4, p4a, p4b



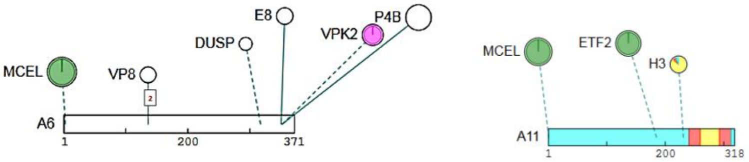




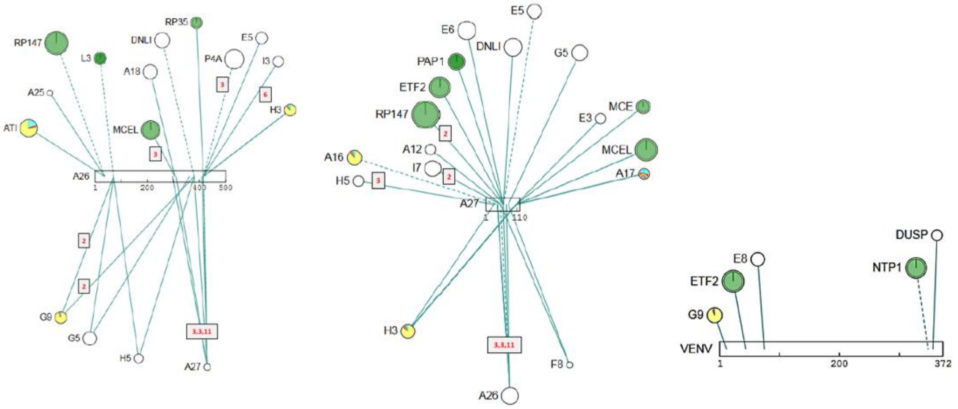
Other enzymes: DUSP, G1, GLRX1, GLRX2, I7, VPK1



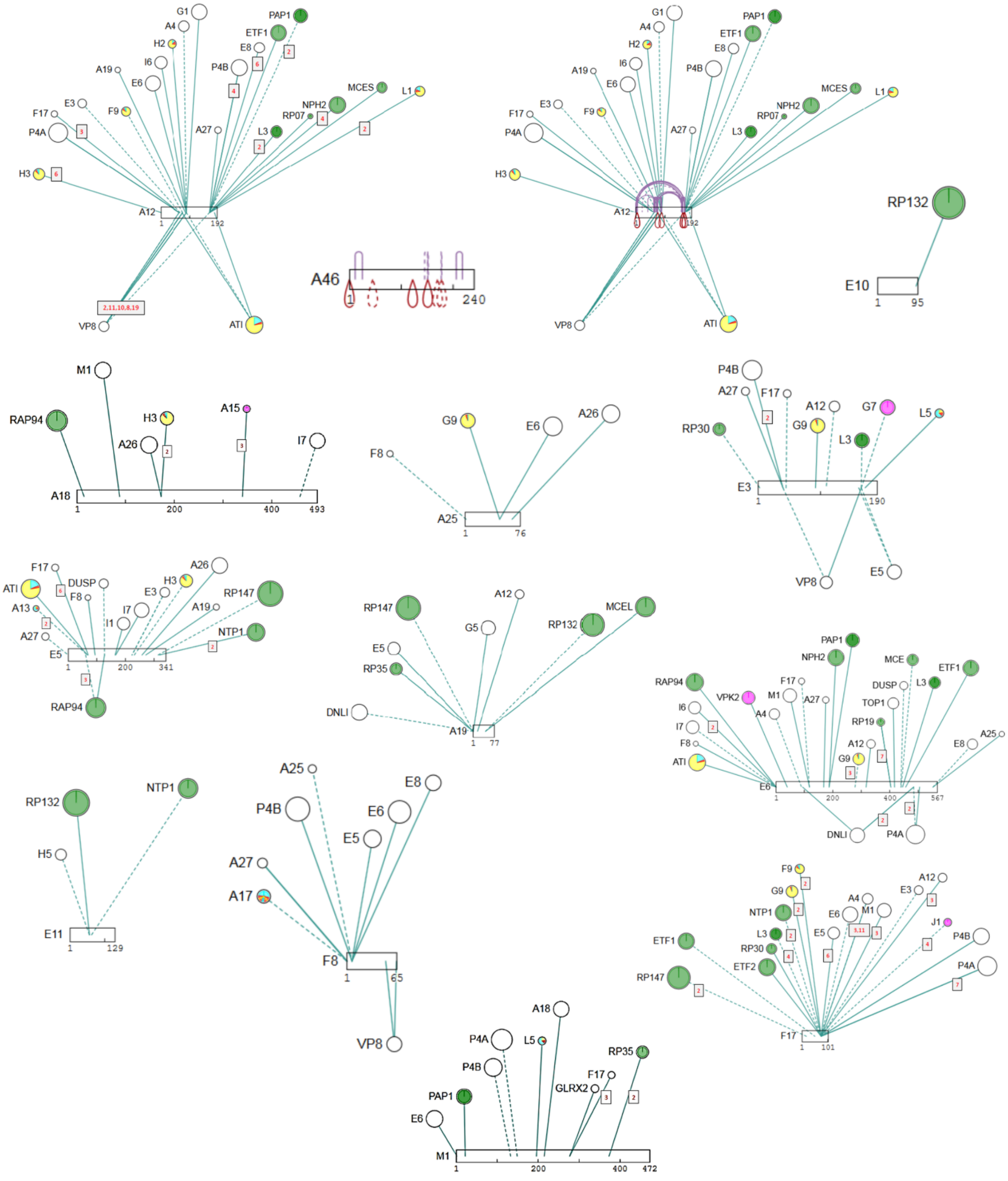
VMAPs : A6, A11



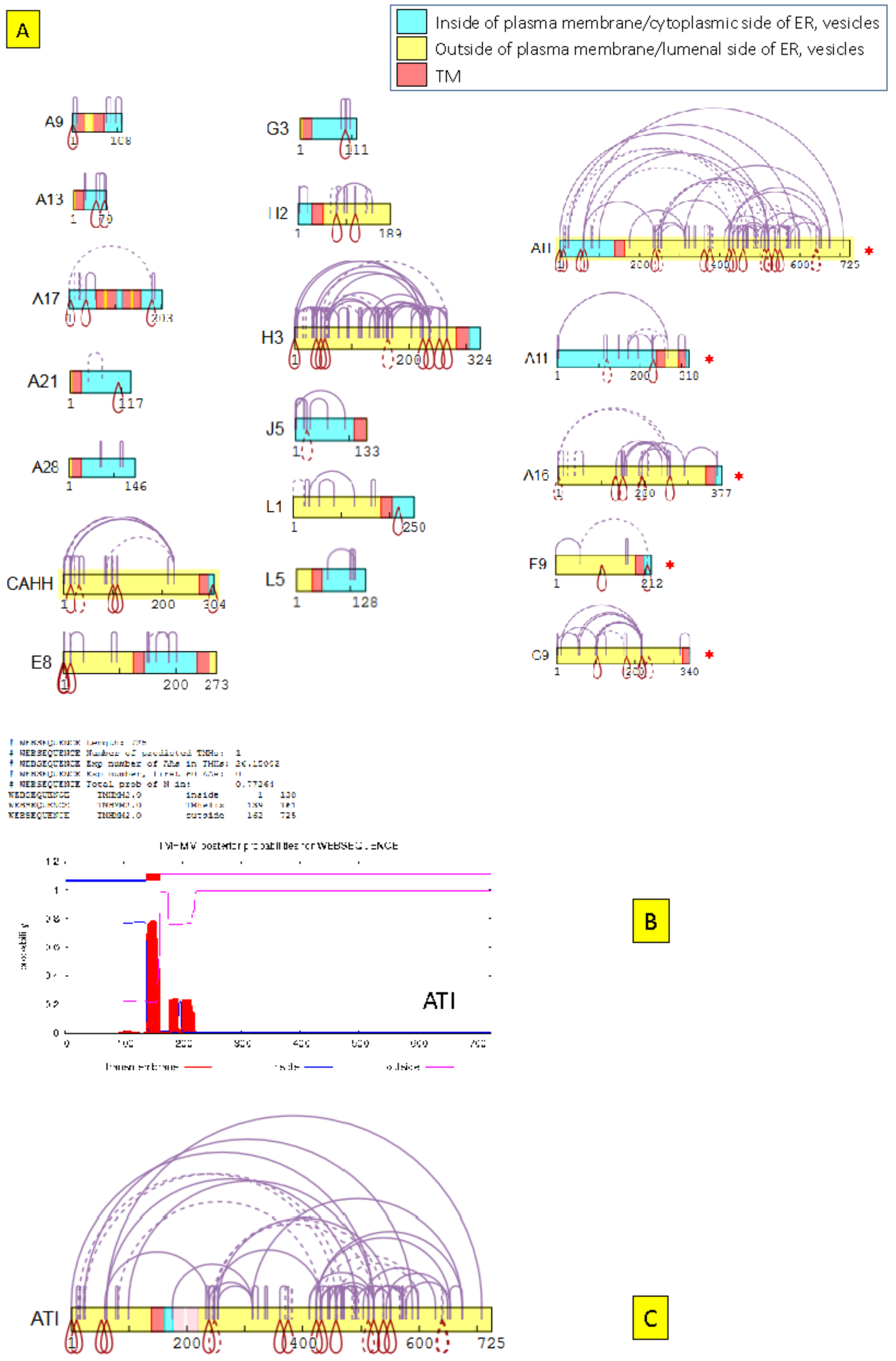
Membrane-associated non-TM proteins: A26, A27, VENV(F13)



Other proteins: A12, A18, A19, A25(A2.5), A46, E3, E5, E6, E10, E11, F8, F17, M1, N1, PROF, SODL, A46



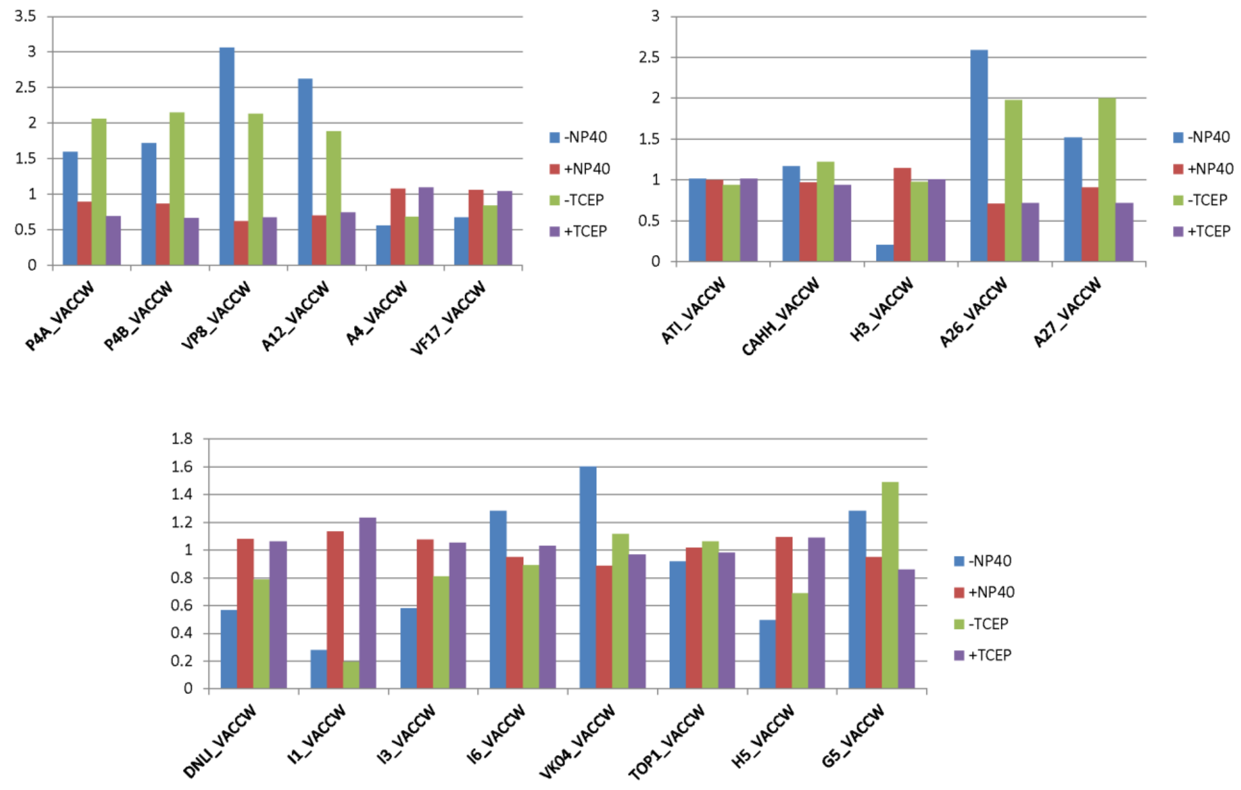
Appendix 2. Figure 2. Intra-protein crosslinks within all TM proteins (a) Yellow, red, cyan fill: ‘Outside’, TM, ‘Inside’ domains, respectively. Mauve loops: Crosslinked peptides from the same protein sequence. Red loops: Homomultimer crosslinks. Vaccinia ATI protein is included in the TM protein group due to its predicted possession of a TM domain with 80% probability (albeit this was a lower probability than for the other TM proteins, see below). (b) TMHMM prediction of TM domain(s) in Vaccinia protein ATI. (c) ATI shown with the two minor TM domains from panel B colored pink.



```
# WEBSEQUENCE Length: 725
# WEBSEQUENCE Number of predicted TMHs:  1
# WEBSEQUENCE Exp number of AAs in TMHs: 26.15092
# WEBSEQUENCE Exp number, first 60 AAs:  0
# WEBSEQUENCE Total prob of N-in:        0.77261
WEBSEQUENCE     TMHMM2.0        inside       1   138
WEBSEQUENCE     TMHMM2.0        TMhelix    139   161
WEBSEQUENCE     TMHMM2.0        outside    162   725
```

Appendix 2. Figure 3. Crosslink partitioning analysis (Materials & methods) for inter-protein XL.

S3 Fig



Appendix 3. Table1. Assessment of AlphaFold2 structures. List of packaged virion proteins with experimentally resolved structures used to benchmark AlphaFold2 performance for Vaccinia proteins, along with domain splits, pLDDT values and alignment scores for the AlphaFold2 models vs. experimental structures. Column A ("Protein"): Protein accession with the corresponding experimental structure's PDB entry ID in column B ("Experimental Structures"). For experimental protein structures from poxviruses other than Vaccinia, virus species is listed in parentheses. For proteins with multiple PDB entries, the entry listed in column B was used for analysis. Column C ("Domain Splits") defines domain boundaries and residues used in each comparison. For GLRX2, residues 31 - 69 were excluded from the analysis since AlphaFold2 predicted the "closed" form as opposed to the "open" form (as reported in the experimental structure). Columns D, E: Average pLDDT values for individual domains of AlphaFold2 models (predicted in "PDB restricted" and "unrestricted" modes respectively). Columns F, G: GDT_TS scores for the alignment of AlphaFold2 models (predicted in "PDB restricted" and "unrestricted" modes, respectively) with experimental structures, for the residues defined in Column C. Columns H – K: TM-score and RMSD values calculated by TM-Align for alignment of AlphaFold2 models (predicted in "PDB restricted" and "unrestricted" mode respectively) with experimental structures, for residues defined in Column C.

Protein	Experimental Structures	Domain Splits	Average pLDDT (restricted)	Average pLDDT (unrestricted)	GDT_TS (restricted)	GDT_TS (unrestricted)	TM-score (restricted)	TM-score (unrestricted)	RMSD Å (restricted)	RMSD Å (unrestricted)
A26_VACCW	6A9S	17-364	90.3	96.9	94.2	98.9	0.95	0.97	1.01	0.16
A27_VACCW	3VOP	45-84	93.7	92.5	94.4	93.7	0.84	0.83	0.62	0.92
A46_VACCW	5EZU	1-76	52.1	61.6	45	64.5	0.41	0.57	3.82	2.27
	4LQK	90-221	95.2	97	98.9	99.4	0.96	0.97	0.58	0.27
A6_VACCW	6CB7	1-118	95	93.5	98.5	99.2	0.97	0.98	0.85	0.71
	6BR8 (Fowlpox virus)	152-374	89.9	94.7	68.6	100	0.81	0.98	3.2	0.31
CAHH_VACCW	4E9O	2-234	95.1	96.2	98.7	99.5	0.98	0.98	0.61	0.37
DUSP_VACCW	3CM3	5-168	97.7	98.1	99.5	100	0.96	0.97	0.45	0.21
E11_VACCW	6RFG	1-122	89.7	86.3	94.3	90.8	0.79	0.8	1.19	1.08
ETF1_VACCW	7AMV_W	1-224	86.3	86.8	93.2	93.2	0.97	0.97	1.21	1.2
		233-480	87.5	87.2	90.1	90.9	0.95	0.96	1.52	1.3
		500-613	90.8	91.1	94.3	94.5	0.94	0.94	1.09	1
ETF2_VACCW	7AMV_K	7-165	91	91.7	91.7	91.7	0.96	0.95	1.05	1.2
		179-272	74.8	78.5	80	78.2	0.79	0.78	2.27	2.22
		275-352	86	88.8	82.2	80.8	0.96	0.96	0.71	0.7
		362-465	86.4	87.6	96.6	96.8	0.96	0.96	0.87	0.85
		469-605	91.1	91.6	97.6	97.3	0.98	0.97	0.75	0.79
		606-710	70.2	72.7	73.5	70.3	0.74	0.69	2.81	3.06
F9_VACCW	6CJ6	2-166	81.8	86.9	88.2	97	0.93	0.98	1.46	0.84
GLRX1_VACCW	2HZE (Ectromelia virus)	1-107	97.7	97.8	100	100	0.98	0.98	0.39	0.4
GLRX2_VACCW	2G2Q	1-37;70-117	93.4	92.3	95.8	95.8	0.94	0.94	0.93	0.93
H3_VACCW	5EJ0	4-240	85.9	85.3	91.8	91.3	0.94	0.9	1.32	1.79
L1_VACCW	1YPY	4-185	67.2	89	65.5	97.6	0.75	0.98	3.51	0.52
MCE_VACCW	1AV6	3-297	94.2	96.2	96.1	99	0.92	0.97	0.96	0.6
MCEL_VACCW	4CKB_A	1-527	90	91.9	95	96	0.96	0.97	0.9	0.85
		550-844	89.4	91.5	96.3	99.2	0.96	0.97	0.92	0.46
MCES_VACCW	4CKB_B	2-287	92.7	92.7	93.7	94	0.96	0.96	1.2	1.29
N1_VACCW	4BBD	1-114	94.8	95.9	100	100	0.99	0.97	0.34	0.29
NTP1_VACCW	6RLF_Y	4-234	92	94.3	90.9	93.4	0.97	1	1.51	0.32
		253-533	92	88.3	88.5	90	0.87	0.9	1.63	0.97
		573-631	82.8	90.9	96.8	98.6	0.86	0.89	0.68	0.44
PAP1_VACCW	2GA9	12-479	92.1	92.1	95.4	95.5	0.94	0.94	0.85	0.85
RAP94_VACCW	6RFL_I	1-80	68.5	89.1	85.4	95.9	0.83	0.92	1.65	0.97
		104-292	72.8	89.4	82.9	99.7	0.78	0.87	2.32	0.8
		325-375	72.4	95	98.5	97.5	0.93	0.93	0.65	0.67
		407-571	82.3	94	97.7	98.5	0.95	0.96	0.61	0.51
		638-795	84	92.6	89.6	99.4	0.93	0.99	1.38	0.41
RP07_VACCW	6RFL_J	2-62	78.6	80	71	72.5	0.68	0.68	1.94	1.81
RP132_VACCW	6RFL_B	1-1021	87.6	93.1	86.4	98.5	0.95	0.96	1.26	0.62
		1026-1143	83.4	94.6	73.5	98.2	0.81	0.99	2.22	0.32
RP147_VACCW	6RFL_A	4-278	82.4	91.7	87.9	99.4	0.95	0.99	1.65	0.5
		284-442	88.3	94.6	90.3	99.8	0.93	0.99	1.57	0.36
		442-970	84	94	83.6	99.6	0.95	1	1.86	0.44
		976-1242	83.4	92.6	75.4	99.7	0.88	0.99	2.32	0.44
RP18_VACCW	6RFL_G	2-159	91.4	92.3	86.6	98	0.89	0.95	1.4	0.68
RP19_VACCW	6RFL_F	62-164	93.7	93.8	88.3	93.2	0.89	0.94	1.89	0.97
RP22_VACCW	6RFL_E	1-184	91.9	92.2	91.9	92.2	0.95	0.95	1.36	1.33
RP30_VACCW	6RFL_S	36-150	71.2	89.1	64.4	96.5	0.62	0.92	2.97	0.56
RP35_VACCW	6RFL_C	2-305	93.7	95.5	90.6	99.7	0.97	1	1.31	0.35
TOP1_VACCW	1VCC	1-77	91.3	96.4	99.4	99.4	0.97	0.98	0.58	0.5
	1A41	81-218	94.6	96.1	77.1	74	0.8	0.78	2.35	2.45
		219-310	88.3	89.6	89.9	89.9	0.9	0.9	1.51	

Appendix 3. Table2. Overall Alignment Scores. Summary of AlphaFold2 performance (in "PDB restricted" and "unrestricted" mode) for all 52 protein domains benchmarked. Median values are calculated from the individual values reported in Appendix3. Table1.

AlphaFold2 mode (restricted/unrestricted)	Median pLDDT	Median GDT_TS	Median TM-score	Median RMSD (Å)
AlphaFold2 "restricted" structures	89.55	91.3	0.94	1.29
AlphaFold2 "unrestricted" structures	92.25	97.15	0.96	0.71

Appendix 3. Table3. Structural homologies of Vaccinia proteins. List of Vaccinia envelope proteins as well as the proteins used for AlphaFold2 benchmarking (Table S1), with protein-level pLDDT values and the top scoring structural homologs identified for each protein by HHpred and DALI. Protein accessions are given in Column A. For proteins discussed by gene name in the text, gene name is provided in parentheses. For proteins with experimentally resolved structures, the corresponding PDB entry ID is given in Column B ("Experimental structure"). Where the experimental structure showed < 90% protein coverage, % coverage of the experimental structure is provided in parentheses. For experimental structures resolved from poxvirus species other than Vaccinia, virus species is given also (Column B). Entries in Column B with a ** refer to structures deposited to PDB and published during preparation of this manuscript. Column C ("HHpred 80%, local alignment)": Highest probability structural homolog identified (with probability ≥ 80%) by HHpred in local alignment mode with MAC realignment "on". Column D ("Average pLDDT"): Average pLDDT for the AlphaFold2 model of the full-length protein. For proteins with corresponding experimental structures, the given average pLDDT was calculated from AlphaFold2 structures predicted in "unrestricted" mode. Column E: Average pLDDT for the AlphaFold2 model of the protein core domain, excluding N- and C-terminal non-structured regions (where applicable). For protein A46, AlphaFold2 provided higher confidence and more accurate structures when N- and C-terminal domains were predicted separately. Average pLDDT and "average pLDDT (core domain)" values are provided for each A46 domain, separated by a comma in columns D and E. Columns F and G: Top structural homolog identified by DALI and corresponding z-score for structural similarity. "-" denotes the absence of a confident structural homolog. For proteins A46 and G9, DALI homology results are provided for the individual domains, separated by a comma.

Protein Accession	Experimental structures	HHpred (80%, local alignment)	Average	Average pLDDT	DALI homolog	DALI z-score
A14_VACCW	4N8V (10%)	Alpha helical transmembrane protein	67.4	67.4	Various alpha helical proteins	8.3
A16_VACCW	8GP6_A (75%)**		82.8	86.8	Alpha-alpha superhelix fold containing proteins	5.9
A17_VACCW			71.1	87.8	-	-
A21_VACCW			85.3	85.3	Orf22 from Cydia pomonella granulosis virus (Baculovirus polyhedron envelope (associated) protein ORF22	3.8
A25_VACCW (A2.5L)			69.8	69.8	-	-
A26_VACCW	6A9S (70%)	Vaccinia A26	85.7	97.0	Vaccinia A26	61.9
A27_VACCW	3VOP (58%), 3U59 (9%)	Vaccinia A27	77.8	91.7	Various alpha helical proteins	5.2
A46_VACCW	5EZU (35%), 4LQK (60%)	Vaccinia A46	86.1,91.1	90.1, 97.2	Vaccinia A46 N-terminal domain, Vaccinia A46 C-terminal domain	7.3, 26.6
A6_VACCW	6CB7 (33%) - Vaccinia, 6BR8 (68%) - Fowlpox virus	Vaccinia A6; Fowlpox virus A6	91.2	91.3	Fowlpox virus A6	29.8
A9_VACCW			64.6	64.6	Tail fiber protein and various alpha helical proteins	6.9
ATI_VACCW (A25L)		Tropomyosin	76.7	91.7	Vaccinia A26	31.8
CAHH_VACCW (D8L)	4E9O (88%)	Vaccinia D8L	90.9	96.3	Vaccinia CAHH	41.5
DUSP_VACCW	3CM3	Vaccinia DUSP	97.3	97.3	Dual specificity protein phosphatase	32.2
E10_VACCW		FAD-linked sulfhydryl oxidase	94.9	94.9	FAD-linked sulfhydryl oxiidase from ASFV	10.6
E11_VACCW	6RFG, 6RFL_Q	Vaccinia E11	84.3	86.5	Vaccinia E11	21.6
ETF1_VACCW	7AMV_W	Vaccinia ETF1	86.0	87.4	Vaccinia ETF1	42.6
ETF2_VACCW	7AMV_K	Vaccinia ETF2	85.3	85.3	Vaccinia ETF2	36.3
F9_VACCW	6CJ6 (83%)	Vaccinia F9	87.7	81.3	Vaccinia F9	29.7
G3_VACCW	7YTT_A (64%)**	Alpha helical transmembrane protein	83.4	83.4	αβββ proteins	4.6
G9_VACCW	8GP6_B (76%)**		86.4	92.3	40S ribosomal protein S15a; alpha-alpha superhelix fold containing proteins	4.4; 6.8
GLRX1_VACCW	2HZE - Ectromelia Virus	Ectromelia virus GLRX1	97.5	97.5	Ectromelia GLRX1	22.9
GLRX2_VACCW	2G2Q	Vaccinia GLRX2	91.5	91.5	Vaccinia GLRX2	15.5
H2_VACCW			85.6	92.8	AP2 domain	4.5
H3_VACCW	5EJ0 (73%)	Vaccinia H3	85.8	86.1	Vaccinia H3	29.3
I5_VACCW			83.6	83.6	Various helix-turn-helix proteins	7.3
J5_VACCW			80.7	94.0	-	-
L1_VACCW	1YPY (72%)	Vaccinia L1	78.7	89.2	Vaccinia L1	31.5
L5_VACCW	7YTT_B (53%)**		78.0	78.0	-	-
MCE_VACCW	1AV6	Vaccinia MCE	90.2	96.1	Vaccinia MCE	45.8
MCEL_VACCW	4CKB_A, 6RFL_O	Vaccinia MCEL	90.8	90.8	Vaccinia MCEL	46.5
MCES_VACCW	4CKB_B, 6RFL_L	Vaccinia MCES	92.7	92.7	Vaccinia MCES	42.9
N1_VACCW	4BBD	Vaccinia N1	95.3	95.3	Vaccinia N1	24.9
NTP1_VACCW	6RLF_Y	Vaccinia NTP1	90.7	90.7	Vaccinia NTP1	48
O3_VACCW			80.0	80.0	Various membrane coiled coil proteins	4.6
PAP1_VACCW	2GA9, 3ERC	Vaccinia PAP1	91.0	91.0	Vaccinia PAP1	52.6
RAP94_VACCW	6RFL_I	Vaccinia RAP94	90.5	90.5	Vaccinia RAP94	30.9
RP07_VACCW	6RFL_J	Vaccinia RP07	79.2	79.2	Vaccinia RP07	7.7
RP132_VACCW	6RFL_B	Vaccinia RP132	93.1	93.1	Vaccinia RP132	50.9
RP147_VACCW	6RFL_A	Vaccinia RP147	93.1	83.9	Vaccinia RP147	53.8
RP18_VACCW	6RFL_G	Vaccinia RP18	91.9	91.9	Vaccinia RP18	23
RP19_VACCW	6RFL (63%)	Vaccinia RP19	76.6	94.6	Vaccinia RP19	17.4
RP22_VACCW	6RFL_E	Vaccinia RP22	92.0	92.0	Vaccinia RP22	28.9
RP30_VACCW	6RFL (61%)	Vaccinia RP30	81.7	81.7	Vaccinia RP30	14.2
RP35_VACCW	6RFL_C	Vaccinia RP35	95.4	95.4	Vaccinia RP35	41.5
TOP1_VACCW	1VCC (25%), 1A41 (75%)	Vaccinia TOP1	93.8	93.8	Variola TOP1	40.1