Homologous Capsid Proteins Testify to the Common Ancestry of Retroviruses, Caulimoviruses, Pseudoviruses, and Metaviruses

Mart Krupovic, Eugene Koonin

Homologous capsid proteins testify to the common ancestry of retroviruses, caulimoviruses, pseudoviruses and metaviruses

Mart Krupovic[a*] and Eugene V. Koonin[b]

a – Unité Biologie Moléculaire du Gène chez les Extrêmophiles, Institut Pasteur, 25 rue du Docteur Roux, 75015
b – National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD 20894, USA

* – Correspondence
E-mail: krupovic@pasteur.fr

Text

Reverse-transcribing viruses are classified into 5 different families, *Retroviridae*, *Metaviridae*, *Pseudoviridae*, *Caulimoviridae* and *Hepadnaviridae* (1). Retroviruses, hepadnaviruses and caulimoviruses are full-fledged viruses, whereas metaviruses and pseudoviruses are more often referred to as retrotransposons. Nevertheless, similar to other reverse-transcribing viruses, certain members of the families *Metaviridae* and *Pseudoviridae* form icosahedral or irregularly-shaped particles that play important roles in their infection cycles (2, 3). Phylogenetic analyses reveal monophyly of the reverse transcriptases (RT) of all these viruses, to the exclusion of the RTs of other retroelements (1, 4). By contrast, the evolutionary relationships between the capsid proteins (CP) of these viruses remain obscure.

Evolutionary connections between viruses that have diverged in a distant past are often difficult to trace due to high mutation and recombination rates in the viral genomes, which abrogate their taxonomic classification, especially at the level of higher taxa. However, it has been argued that comparison of sequences and structures of the major virion proteins provides a deeper reach into the evolutionary

history of the virosphere (5-8). High-resolution CP structures are available for several retroviruses and hepadnaviruses. Members of the *Retroviridae* encode their CP as part of a larger polyprotein precursor known as Gag, which is proteolytically processed by the virus-encoded protease to release several major proteins, three of which, namely matrix, capsid (CA), nucleocapsid (NC), are conserved in all retroviruses (9). The CA protein is composed of two α-helical domains, the N-terminal domain (NTD) and the C-terminal domain (CTD) (10, 11). The CP of hepadnaviruses also adopts an α-helical fold, but does not bear recognizable similarity to retroviral CA (12). By contrast, members of the family *Metaviridae*, including Ty3/Gypsy retrotransposons, encode a Gag polyprotein, which contains the CA and NC domains related to those of retroviruses (13, 14). Additionally, the Zn-binding module, known as 'Zn-knuckle', that is found in many retrovirus and metavirus NC domains is also present in the CPs of caulimoviruses and members of the family *Pseudoviridae*, such as Ty1/Copia retrotransposons (15). However, the CPs from the latter two groups of viruses are generally considered to be unrelated to the CA of retroviruses or metaviruses (1, 16), suggesting that different groups of reverse-transcribing viruses could have evolved from non-viral retroelements via acquisition of capsid-encoding genes from different sources. As a case in point, envelope proteins responsible for host recognition and membrane fusion have been recruited by reverse-transcribing viruses on several independent occasions (1, 17).

To investigate the provenance of the pseudovirus and caulimovirus CPs, we performed sensitive profile-profile searches using HHpred (18) seeded with sequences of CPs of pseudoviruses, caulimoviruses and metaviruses. In all cases, we obtained highly significant matches (>98% HHpred probability) to the Gag of retroviruses. The region of similarity encompasses the end of the CA-NTD and extends throughout the CA-CTD and NC (data not shown). To validate these matches, we aligned the retroviral CA-NC sequences to the corresponding protein sequences from the three other families of reverse-transcribing viruses. Inspection of the alignment showed that the sequence conservation among the aligned proteins

encompasses not only CA-CTD and NC domains but also CA-NTD, except for two short α-helixes (α5 and α6) that have no equivalents in CPs of viruses other than retroviruses (Figures 1 and 2). Importantly, non-polar residues, which stabilize the hydrophobic core of the α-helical bundles of both CA-NTD and CA-CTD (10, 19), are conserved in all CP/CA proteins. Thus, patterns of sequence and secondary structure conservation strongly support the homology of the CA/CP-NC module of all reverse-transcribing viruses, except for hepadnaviruses. This result, combined with the monophyly of viral RTs (4), strongly suggests that retroviruses, caulimoviruses, pseudoviruses and metaviruses evolved from a common viral ancestor, rather than from distinct capsid-less retrotransposons. Such an ancestral virus encoding the CA/CP-NC, protease (14) and the RT, including the RNase H domain (20), is likely to have existed prior to the divergence of plants and opisthokonts (fungi and animals) ~1.6 billion years ago (21). Furthermore, given that, in the RT phylogeny, hepadnaviruses cluster with pseudoviruses (1, 4), it appears most likely that the conserved CP/CA was acquired by the common ancestor of all reverse-transcribing viruses, followed by replacement with an unrelated protein in hepadnaviruses. Finally, our results seem to justify the creation of a single, high-order viral taxon unifying all families of reverse-transcribing viruses (or alternatively, four families, excluding hepadnaviruses), consistent with previous suggestions based on the shared mechanisms of genome replication employed by these viruses (22).

References

1. **Koonin, E. V., V. V. Dolja, and M. Krupovic.** 2015. Origins and evolution of viruses of eukaryotes: The ultimate modularity. Virology **479-480:**2-25.
2. **King, A. M. Q., M. J. Adams, E. B. Carstens, and E. J. Lefkowitz.** 2011. Virus Taxonomy. Ninth Report of the International Committee on Taxonomy of Viruses. Elsevier Academic, London.
3. **Palmer, K. J., W. Tichelaar, N. Myers, N. R. Burns, S. J. Butcher, A. J. Kingsman, S. D. Fuller, and H. R. Saibil.** 1997. Cryo-electron microscopy structure of yeast Ty retrotransposon virus-like particles. J Virol **71:**6863-8.
4. **Gladyshev, E. A., and I. R. Arkhipova.** 2011. A widespread class of reverse transcriptase-related cellular genes. Proc Natl Acad Sci U S A **108:**20311-6.
5. **Krupovic, M., and D. H. Bamford.** 2010. Order to the viral universe. J Virol **84:**12476-9.
6. **Sinclair, R. M., J. J. Ravantti, and D. H. Bamford.** 2017. Nucleic and amino acid sequences support structure-based viral classification. J Virol.
7. **Krupovic, M., and E. V. Koonin.** 2017. Multiple origins of viral capsid proteins from cellular ancestors. Proc Natl Acad Sci U S A **114:**E2401-E2410.
8. **Abrescia, N. G., D. H. Bamford, J. M. Grimes, and D. I. Stuart.** 2012. Structure unifies the viral universe. Annu Rev Biochem **81:**795-822.
9. **Mattei, S., F. K. Schur, and J. A. Briggs.** 2016. Retrovirus maturation-an extraordinary structural transformation. Curr Opin Virol **18:**27-35.
10. **Ganser-Pornillos, B. K., A. Cheng, and M. Yeager.** 2007. Structure of full-length HIV-1 CA: a model for the mature capsid lattice. Cell **131:**70-9.
11. **Ganser-Pornillos, B. K., M. Yeager, and W. I. Sundquist.** 2008. The structural biology of HIV assembly. Curr Opin Struct Biol **18:**203-17.
12. **Steven, A. C., J. F. Conway, N. Cheng, N. R. Watts, D. M. Belnap, A. Harris, S. J. Stahl, and P. T. Wingfield.** 2005. Structure, assembly, and antigenicity of hepatitis B virus capsid proteins. Adv Virus Res **64:**125-64.
13. **Larsen, L. S., M. Zhang, N. Beliakova-Bethell, V. Bilanchone, A. Lamsa, K. Nagashima, R. Najdi, K. Kosaka, V. Kovacevic, J. Cheng, P. Baldi, G. W. Hatfield, and S. Sandmeyer.** 2007. Ty3 capsid mutations reveal early and late functions of the amino-terminal domain. J Virol **81:**6957-72.
14. **Llorens, C., M. A. Fares, and A. Moya.** 2008. Relationships of gag-pol diversity between Ty3/Gypsy and *Retroviridae* LTR retroelements and the three kings hypothesis. BMC Evol Biol **8:**276.
15. **Covey, S. N.** 1986. Amino acid sequence homology in gag region of reverse transcribing elements and the coat protein gene of cauliflower mosaic virus. Nucleic Acids Res **14:**623-33.
16. **Pachulska-Wieczorek, K., S. F. Le Grice, and K. J. Purzycka.** 2016. Determinants of Genomic RNA Encapsidation in the *Saccharomyces cerevisiae* Long Terminal Repeat Retrotransposons Ty1 and Ty3. Viruses **8:**E193.
17. **Malik, H. S., S. Henikoff, and T. H. Eickbush.** 2000. Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. Genome Res **10:**1307-18.
18. **Söding, J.** 2005. Protein homology detection by HMM-HMM comparison. Bioinformatics **21:**951-60.
19. **Ball, N. J., G. Nicastro, M. Dutta, D. J. Pollard, D. C. Goldstone, M. Sanz-Ramos, A. Ramos, E. Mullers, K. Stirnnagel, N. Stanke, D. Lindemann, J. P. Stoye, W. R. Taylor, P. B. Rosenthal, and I. A. Taylor.** 2016. Structure of a spumaretrovirus Gag central domain reveals an ancient retroviral capsid. PLoS Pathog **12:**e1005981.
20. **Malik, H. S., and T. H. Eickbush.** 2001. Phylogenetic analysis of ribonuclease H domains suggests a late, chimeric origin of LTR retrotransposable elements and retroviruses. Genome Res **11:**1187-97.

21. **Hedges, S. B., J. E. Blair, M. L. Venturi, and J. L. Shoe.** 2004. A molecular timescale of eukaryote evolution and the rise of complex multicellular life. BMC Evol Biol **4:**2.
22. **Hull, R.** 2001. Classifying reverse transcribing elements: a proposal and a challenge to the ICTV. International Committee on Taxonomy of Viruses. Arch Virol **146:**2255-61.
23. **Llorens, C., R. Futami, D. Bezemer, and A. Moya.** 2008. The Gypsy Database (GyDB) of mobile genetic elements. Nucleic Acids Res **36:**D38-46.
24. **Pei, J., and N. V. Grishin.** 2014. PROMALS3D: multiple protein sequence alignment enhanced with evolutionary and three-dimensional structural information. Methods Mol Biol **1079:**263-71.

Figure legend

Figure 1. Multiple sequence alignment of CA/CP proteins of reverse-transcribing viruses belonging to families *Retroviridae*, *Metaviridae*, *Pseudoviridae* and *Caulimoviridae*. Representative sequences from all four virus groups were downloaded from the Gypsy Database (23), supplemented with sequences of retroviral CA proteins for which X-ray structures are available and were aligned using PROMALS3D (24). Secondary structure elements above the alignment are indicated for the Rous sarcoma virus CA (PDB: 5A9E), which was retrieved as the best hit to CP proteins of metaviruses, pseudoviruses and caulimoviruses in HHpred search (data not shown). Red and blue asterisks indicate the conserved residues in the single or tandem Zn-knuckle motifs of the nucleocapsid (NC) domain. Abbreviations: HIV-1, human immunodeficiency virus 1; BLV, bovine leukemia virus; JSRV, Jaagsiekte sheep retrovirus; LPDV, lymphoproliferative disease virus; MMTV, mouse mammary tumor virus; PyERV, *Python molurus* endogenous retrovirus; EIAV, equine infectious anemia virus; BSGFV, banana streak goldfinger virus; BSOLV, banana streak OL (badna)virus; CaMV, cauliflower mosaic virus; CERV, carnation etched ring virus; DBV, dasheen bacilliform virus; DrMV, Dracaena mottle virus; MiMV, Mirabilis mosaic virus; TaBV, taro bacilliform virus.

Figure 2. Conservation of the amino acid residues contributing to the stabilization of the hydrophobic core of the α-helical bundles of CA-NTD and CA-CTD. Sequence conservation derived from the alignment shown in Figure 1 was mapped onto the CA structure of Rous sarcoma virus (PDB: 5A9E). The least conserved regions are colored red, whereas those most conserved are in blue.