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

Mart Krupovic,^a Eugene V. Koonin^b

Unité Biologie Moléculaire du Gène Chez les Extrémophiles, Institut Pasteur, Paris, France^a; National Center for Biotechnology Information, National Library of Medicine, Bethesda, Maryland USA^b

KEYWORDS *Caulimoviridae*, structural polyprotein Gag, *Metaviridae*, *Pseudoviridae*, *Retroviridae*, nucleocapsid proteins, reverse-transcribing viruses, Ty3/Gypsy and Ty1/Copia LTR retrotransposons, virus evolution

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, like 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 (RTs) of all these viruses, to the exclusion of the RTs of other retroelements (1, 4). In contrast, the evolutionary relationships between the capsid proteins (CPs) of these viruses remain obscure.

Evolutionary connections between viruses that have diverged in the distant past are often difficult to trace due to high rates of mutation and recombination 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 for several retroviruses and hepadnaviruses are available. 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), and 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 CPs of hepadnaviruses also adopt an α -helical fold but do not bear recognizable similarity to retroviral CA (12). In contrast, members of the family *Metaviridae*, including Ty3/Gypsy retrotransposons, encode a Gag polyprotein which contains CA and NC domains related to those of retroviruses (13, 14). Additionally, the Zn-binding module known as Zn-knuckle, which 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 last 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 might have evolved from nonviral 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).

Accepted manuscript posted online 29 March 2017

Citation Krupovic M, Koonin EV. 2017. Homologous capsid proteins testify to the common ancestry of retroviruses, caulimoviruses, pseudoviruses, and metaviruses. *J Virol* 91:e00210-17. <https://doi.org/10.1128/JVI.00210-17>.

Editor Wesley I. Sundquist, University of Utah

Copyright © 2017 American Society for Microbiology. All Rights Reserved.

Address correspondence to Mart Krupovic, krupovic@pasteur.fr.

(Continued on next page)

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

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 domain (data not shown). To validate these matches, we aligned the retroviral CA-NC sequences with 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 the CA CTD and NC domain but also the CA NTD, with the exception of two short α -helices ($\alpha 5$ and $\alpha 6$) that have no equivalents in CPs of viruses other than retroviruses (Fig. 1 and 2). Importantly, nonpolar residues, which stabilize the hydrophobic core of the α -helical bundles of both the 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 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 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).

FIG 1 Legend (Continued)

of retroviral CA proteins for which X-ray structures are available, and aligned using PROMALS3D (24). Secondary-structure elements above the alignment are indicated for Rous sarcoma virus CA (PDB accession number [5A9E](#)), which was retrieved as the best match to CPs of metaviruses, pseudoviruses, and caulimoviruses in an 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 type 1; BLV, bovine leukemia virus; JSRV, Jaagsiekte sheep retrovirus; LPDV, lymphoproliferative disease virus; MMTV, mouse mammary tumor virus; MPMV, Mason-Pfizer monkey virus; PyERV, *Python molurus* endogenous retrovirus; EIAV, equine infectious anemia virus; DRM, *Danio rerio* Mag element; BSGFV, banana streak Goldfinger virus; BSOLV, banana streak OL (badna) virus; CaMV, cauliflower mosaic virus; CERV, carnation etched ring virus; CSSV, cacao swollen-shoot virus; DBV, dasheen bacilliform virus; DrMV, *Dracaena* mottle virus; MiMV, *Mirabilis* mosaic virus; TaBV, taro bacilliform virus.

ACKNOWLEDGMENTS

E.V.K. is supported by intramural funds of the U.S. Department of Health and Human Services (to the National Library of Medicine).

REFERENCES

- Koonin EV, Dolja VV, Krupovic M. 2015. Origins and evolution of viruses of eukaryotes: the ultimate modularity. *Virology* 479-480:2–25. <https://doi.org/10.1016/j.virol.2015.02.039>.
- King AMQ, Adams MJ, Carstens EB, Lefkowitz EJ. 2011. Virus taxonomy. Ninth report of the International Committee on Taxonomy of Viruses. Elsevier Academic, London, United Kingdom.
- Palmer KJ, Tichelaar W, Myers N, Burns NR, Butcher SJ, Kingsman AJ, Fuller SD, Saibil HR. 1997. Cryo-electron microscopy structure of yeast Ty retrotransposon virus-like particles. *J Virol* 71:6863–6868.
- Gladyshev EA, Arhipova IR. 2011. A widespread class of reverse transcriptase-related cellular genes. *Proc Natl Acad Sci U S A* 108:20311–20316. <https://doi.org/10.1073/pnas.1100266108>.
- Krupovic M, Bamford DH. 2010. Order to the viral universe. *J Virol* 84:12476–12479. <https://doi.org/10.1128/JVI.01489-10>.
- Sinclair RM, Ravanti JJ, Bamford DH. 25 January 2017. Nucleic and amino acid sequences support structure-based viral classification. *J Virol* <https://doi.org/10.1128/JVI.02275-16>.
- Krupovic M, Koonin EV. 2017. Multiple origins of viral capsid proteins from cellular ancestors. *Proc Natl Acad Sci U S A* 114:E2401–E2410. <https://doi.org/10.1073/pnas.1621061114>.
- Abrescia NG, Bamford DH, Grimes JM, Stuart DI. 2012. Structure unifies the viral universe. *Annu Rev Biochem* 81:795–822. <https://doi.org/10.1146/annurev-biochem-060910-095130>.
- Mattei S, Schur FK, Briggs JA. 2016. Retrovirus maturation—an extraordinary structural transformation. *Curr Opin Virol* 18:27–35. <https://doi.org/10.1016/j.coviro.2016.02.008>.
- Ganser-Pornillos BK, Cheng A, Yeager M. 2007. Structure of full-length HIV-1 CA: a model for the mature capsid lattice. *Cell* 131:70–79. <https://doi.org/10.1016/j.cell.2007.08.018>.
- Ganser-Pornillos BK, Yeager M, Sundquist WI. 2008. The structural biology of HIV assembly. *Curr Opin Struct Biol* 18:203–217. <https://doi.org/10.1016/j.sbi.2008.02.001>.
- Steven AC, Conway JF, Cheng N, Watts NR, Belnap DM, Harris A, Stahl SJ, Wingfield PT. 2005. Structure, assembly, and antigenicity of hepatitis B virus capsid proteins. *Adv Virus Res* 64:125–164. [https://doi.org/10.1016/S0065-3527\(05\)64005-5](https://doi.org/10.1016/S0065-3527(05)64005-5).
- Larsen LS, Zhang M, Beliakova-Bethell N, Bilanchone V, Lamsa A, Nagashima K, Najdi R, Kosaka K, Kovacevic V, Cheng J, Baldi P, Hatfield GW, Sandmeyer S. 2007. Ty3 capsid mutations reveal early and late functions of the amino-terminal domain. *J Virol* 81:6957–6972. <https://doi.org/10.1128/JVI.02207-06>.
- Llorens C, Fares MA, Moya A. 2008. Relationships of gag-pol diversity between Ty3/Gypsy and *Retroviridae* LTR retroelements and the three kings hypothesis. *BMC Evol Biol* 8:276. <https://doi.org/10.1186/1471-2148-8-276>.
- Covey SN. 1986. Amino acid sequence homology in gag region of reverse transcribing elements and the coat protein gene of cauliflower mosaic virus. *Nucleic Acids Res* 14:623–633. <https://doi.org/10.1093/nar/14.2.623>.
- Pachulska-Wieczorek K, Le Grice SF, Purzycka KJ. 2016. Determinants of genomic RNA encapsidation in the *Saccharomyces cerevisiae* long terminal repeat retrotransposons Ty1 and Ty3. *Viruses* 8:E193. <https://doi.org/10.3390/v8070193>.
- Malik HS, Henikoff S, Eickbush TH. 2000. Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. *Genome Res* 10:1307–1318. <https://doi.org/10.1101/gr.145000>.
- Söding J. 2005. Protein homology detection by HMM-HMM comparison. *Bioinformatics* 21:951–960. <https://doi.org/10.1093/bioinformatics/bti125>.
- Ball NJ, Nicastro G, Dutta M, Pollard DJ, Goldstone DC, Sanz-Ramos M, Ramos A, Mullers E, Stirnagel K, Stanke N, Lindemann D, Stoye JP, Taylor WR, Rosenthal PB, Taylor IA. 2016. Structure of a spumaretrovirus Gag central domain reveals an ancient retroviral capsid. *PLoS Pathog* 12:e1005981. <https://doi.org/10.1371/journal.ppat.1005981>.
- Malik HS, Eickbush TH. 2001. Phylogenetic analysis of ribonuclease H domains suggests a late, chimeric origin of LTR retrotransposable elements and retroviruses. *Genome Res* 11:1187–1197. <https://doi.org/10.1101/gr.185101>.
- Hedges SB, Blair JE, Venturi ML, Shoe JL. 2004. A molecular timescale of eukaryote evolution and the rise of complex multicellular life. *BMC Evol Biol* 4:2. <https://doi.org/10.1186/1471-2148-4-2>.
- Hull R. 2001. Classifying reverse transcribing elements: a proposal and a challenge to the ICTV. *International Committee on Taxonomy of Viruses. Arch Virol* 146:2255–2261.
- Llorens C, Futami R, Bezemer D, Moya A. 2008. The Gypsy Database (GyDB) of mobile genetic elements. *Nucleic Acids Res* 36:D38–D46. <https://doi.org/10.1093/nar/gkm697>.
- Pei J, Grishin NV. 2014. PROMALS3D: multiple protein sequence alignment enhanced with evolutionary and three-dimensional structural information. *Methods Mol Biol* 1079:263–271. https://doi.org/10.1007/978-1-62703-646-7_17.