# The Role of Mini Zinc-Finger Protein in the Replication of the AIDS Virus

# The Role of Mini Zinc-Finger Protein in the Replication of the AIDS Virus

Edited by

Jean-Luc Darlix

Cambridge
Scholars
Publishing

The Role of Mini Zinc-Finger Protein in the
Replication of the AIDS Virus

Edited by Jean-Luc Darlix

This book first published 2021

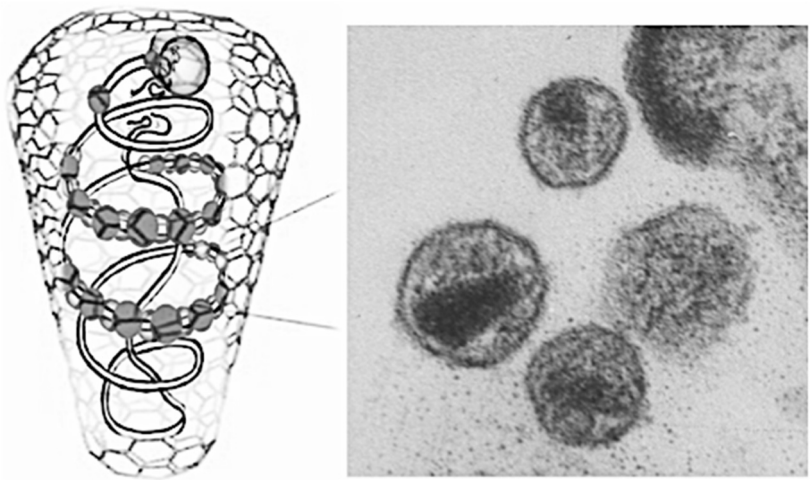
Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library



ISBN (10): 1-5275-6230-1
ISBN (13): 978-1-5275-6230-1



**Schematic Representation of the HIV-1 Particle:**

Right panel. Mature HIV-1 particles as observed by electron microscopy (EM), with a dense conical core and a diameter of about 115 nm. Note the heterogeneity of the particles.

Left panel. Illustration of the mature HIV-1 core resulting from the step-by-step protease-mediated processing of the Gag and Gag-Pol polyproteins (adapted from Campbell and Hope, 2015). This model has been called 'fullerene cone model' in which hexamers of the capsid protein are associated to form an hexagonal surface lattice that is closed by incorporating 12 capsid-protein pentamers (Mattei et al., 2016). The cellular metabolite inositol hexaphosphate (IP6) participates in Core stability by interacting with arginine residues as well as in the reverse transcription during the early phase of virus replication (Dick et al. 2018; Mallery et al., 2018; David et al., 2016)

The viral RNP is encaged in the cone structure, and results from the association of two genomic RNA molecules, 9630 nt in length, called the dimeric RNA genome (two parallel black lines), coated by about 1500-1800 molecules of the mature nucleocapsid protein (NC, multiple circles), while the reverse transcriptase enzyme (RT) is at the initation site for viral DNA synthesis (twin circles at the top).

# TABLE OF CONTENTS

# EDITOR

**Jean-Luc Darlix grew up in Morocco and got his diploma** in Natural Sciences, Virology and Biochemistry and ultimately his PhD in Natural Sciences at the university of Paris in May 1970. He worked in Tchad for his military period in 1970-1971; next he worked on the role of Rho, RNase III and RNase H on the polarity of gene expression in Bacteriophages T7 and T4 at the CEA Saclay. He went to Geneva at Sciences II (Prof. PF Spahr) to investigate the process of reverse transcription in retroviruses ASLV and MoMuLV (1975-1985); he was among the first to carry out RNA sequencing of ASLV and SV40 T-antigen bound microRNAs. He discovered the role of NC in cDNA synthesis and virus assembly. Back to France at the university of Toulouse, he set up a lab working on the role of NC in the replication of MuLV and HIV and on the translation of the retroviral RNA. He then moved to Lyon at ENS to set up a department of human virology with a high safety facility (BSL3). His lab was involved in the development of retroviral and lentiviral vectors for transgenesis in collaboration with Transgene (Strasbourg). He has been member of the INSERM, SIDACTION, and Fondation Recherche Médicale scientific councils as well as an expert for the French government (1995-2015), and for the Canadian government.

# FOREWORD ON RNA BINDING PROTEINS

## The expanding Universe of RNA and RNA-binding proteins.

In the last three decades, RNA has emerged as a family of highly diverse molecules both in length (from tiny such as micro-RNAs to very long non-coding RNAs such as circular RNAs, circRNA) and activity (from RNA ligation to ribozyme-directed cleavage and terminal transferase activity). The molecular identification of a large diversity of RNA molecules and many RNA functions, together with the functional annotation of RNA transcripts in genomes represent an exciting challenge for modern biology (Higgs and Lehman 2015; Ralser 2014; Nam et al., 2016; Dolja and Koonin 2018). The latest developments in interactome technologies allow the profiling of RNA-protein interactions at the genome-wide scale and have led to the identification of new RNA classes associated with globally-acting RNA-binding proteins (RBPs) (Smirnov et al., 2017; A, Schneider et al., 2018; Garcia-Moreno et al., 2018M; Beckmann et al., 2016). RBPs participate in many key biological functions, notably transcription, translation, splicing and translation, and present RS and RG-rich sequences in intrinsically disordered regions of low complexity (Järvelin et al.2016; Tompa 2002; Uversky et al. 2014) that allow them to interact dynamically with their target RNAs. The consequences of these dynamic interactions span from the control of RNA biogenesis and turnover (Kato and McKnight, 2018), to the assembly of higher-order structures that exert different key physiological functions in the cells.

## The Nucleocapsid protein (NC) of HIV-1.

This Zinc Finger miniprotein represents a remarkable example of an RBP, because it plays essential roles in all steps of the virus replication cycle, starting from its structural presence in virion particles, in which about 1500-1900 molecules of NC coat the entire, dimeric genomic RNA within viral cores. When the virion particle attaches to the target cell, during the early phases of the viral life cycle, NC chaperones the reverse

transcription (RTion) process whereby a double-stranded DNA is synthesized from the viral genomic RNA. Once the viral DNA integrates in the host cell's DNA under the form of a provirus (Temin, 1976) viral genes are expressed by the RNA polymerase II transcription machinery as spliced and unspliced RNAs that are then exported into the cell's cytoplasm. The genomic RNA, also named full length viral RNA (FL RNA), resembles a canonical cellular mRNA and exhibits similar 5' cap and 3' polyA structures. The FL RNA is then translated into the major structural proteins and enzymes of the virus, in the form of the polyprotein precursors Pr55 Gag and Pr160 GagPol, that assemble at the site of virion assembly together with the viral genomic RNA (Rein 2019; Hellmund and Lever, 2016; Keane and Summers, 2016).

Again, it is the NC domain, within the yet unprocessed Gag polyprotein, that directs genomic RNA selection and dimerization, which in turn drives Gag oligomerization and final polyprotein processing once the virion assembly process is engaged. As a result of this process, NC finds itself within viral cores, coating the entire genomic RNA, ready for the next cycle of infection when either cell-free, or cell-associated virus particles are ready to infect a new cell (Sourisseau et al. 2007; Alvarez et al. 2014).

## References

Alvarez RA, Barría MI, Chen BK. Unique features of HIV-1 spread through T cell virological synapses. PLoS Pathog. 2014 Dec 18;10(12):e1004513.

Beckmann BM, Castello A, Medenbach J. The expanding universe of ribonucleoproteins: of novel RNA-binding proteins and unconventional interactions. Pflugers Arch. 2016 Jun;468(6):1029-40.)

Dolja VV, Koonin EV. Metagenomics reshapes the concepts of RNA virus evolution by revealing extensive horizontal virus transfer. Virus Res. 2018 Jan 15;244:36-52.

Higgs PG, Lehman N. The RNA World: molecular cooperation at the origins of life. Nat Rev Genet. 2015 Jan;16(1):7-17.

Garcia-Moreno M, Järvelin AI, Castello A. Unconventional RNA-binding proteins step into the virus-host battlefront. Wiley Interdiscip Rev RNA. 2018 Nov;9(6):e1498.

Hellmund C, Lever AM Coordination of Genomic RNA Packaging with Viral Assembly in HIV-1.Viruses. 2016 Jul 14;8(7).)

Järvelin et al. Cell Communication and Signaling (2016) 14:9

Kato M, McKnight SL. A Solid-State Conceptualization of Information Transfer from Gene to Message to Protein. Annu Rev Biochem. 2018 Jun 20;87:351-390

Keane SC, Summers MF NMR Studies of the Structure and Function of the HIV-1 5'-Leader.Viruses. 2016 Dec 21;8(12). pii: E338. doi:

Ralser M.The RNA world and the origin of metabolic enzymes. Biochem Soc Trans. 2014 Aug;42(4):985-8;

Rein A. RNA Packaging in HIV. Trends Microbiol. 2019 Aug;27(8):715-723.

Nam JW, Choi SW, You BH. Incredible RNA: Dual Functions of Coding and Noncoding. Mol Cells. 2016 May 31;39(5):367-74;

Sourisseau M, Sol-Foulon N, Porrot F, Blanchet F, Schwartz O. Inefficient human immunodeficiency virus replication in mobile lymphocytes. J Virol. 2007 Jan;81(2):1000-12.

Tompa P. Intrinsically unstructured proteins. Trends Biochem Sci. 2002;27:527–33.

Uversky VN, Davé V, Iakoucheva LM, Malaney P, Metallo SJ, Pathak RR, et al. Pathological unfoldomics of uncontrolled chaos: intrinsically disordered proteins and human diseases. Chem Rev. 2014;114:6844–79.

Smirnov A, Schneider C, Hör J, Vogel J. Discovery of new RNA classes and global RNA-binding proteins. Curr Opin Microbiol. 2017 Oct;39:152-160. doi: 10.1016/j.mib.2017.11.016;

Temin HM. The DNA provirus hypothesis. Science. 1976 Jun 11;192(4244):1075-80

The following chapters aim at describing the key roles of NC protein in the structure of the HIV viral particle and in its replication, in particular in the process of viral DNA synthesis by the viral RNA/DNA dependent DNA polymerase (RT) and in particle assembly. This scenario, as we see it in 2020, is described in a stepwise manner mainly according to *in vitro* and *ex vivo* investigations**.**

*Keywords*: Retroviruses, HIV-1, AIDS, Virus structure, Zinc Finger Protein, Dimeric RNA genome, Reverse transcription, Recombination, Genetic diversity, Resistance to anti-retroviral compounds, Retrotransposon, origin of the Zinc Finger Protein

# FOREWORD ON RETROVIRUSES AND NC PROTEIN

During the XXth century retroviruses were discovered and recognized as pathogenic agents that can cause cancers, immunodeficiencies and diseases of the central nervous system, bone and joints in animals such as horses, birds, and rodents. In the early eighties two human retroviruses, HTLV and HIV, were found to cause leukaemia and AIDS, respectively.

Retroviruses are widespread in vertebrates and have been classified into two subfamilies that are the Orthoretroviruses represented by 6 genera (Alpha-, Beta-, Delta-, Gamma-, Epsilon-retrovirus, and Lentivirus), and the Spumaretroviruses. Spuma- or Foamy viruses are considered to be the most ancient retroviruses that appeared about 500 million years ago with the first vertebrates.

Retroviral NC proteins have been studied for more than thirty years, starting from the observation that NC is found in the inner structure of the viral particle and the notion that NC proteins are non-specific nucleic acid binding proteins (NABP). Today our view of NC drastically changed since NC has been revealed to be a multifunctional protein that interacts with both viral and cellular RNAs, as well as with the viral enzymes reverse transcriptase (RT) and integrase (IN), as described below.

## NC Protein in the Retroviral Particle

The structure and composition of retroviral particles of rodent and human origins have been extensively studied, and data show that particles are globular with a mean diameter of 115 +/- 10 nm (Fig. 2), composed of three visible structures, namely the outer envelope, the matrix and the inner core (Fig. 3 shows a simple schematic representation of an HIV-1 particle).

The genome is incorporated inside viral cores as a positive strand RNA, resembling a cellular mRNA and possessing a 5’ cap and a 3’ polyA structure. This genome is incorporated as a dimer, so that two identical RNA molecules are associated in the form of a 60-70S complex, together with a specific cellular tRNA annealed to the 5’ends. Furthermore, the genomic RNA is coated by hundreds of NC protein molecules forming a

nucleocore structure, shown to be resistant, at least in part, to nucleases and host defences.

The viral enzymes protease (PR), reverse transcriptase (RT) and integrase (IN), all essential for virus replication, are also found in the inner core of the retroviral particle.

The genomic RNA with the associated NC protein molecules, together with RT and IN, constitute the viral replicative molecular machinery, which functions in a unique copy-and-paste mode that converts the genomic RNA into a double stranded DNA, which is later inserted into the host genome by IN (see below).

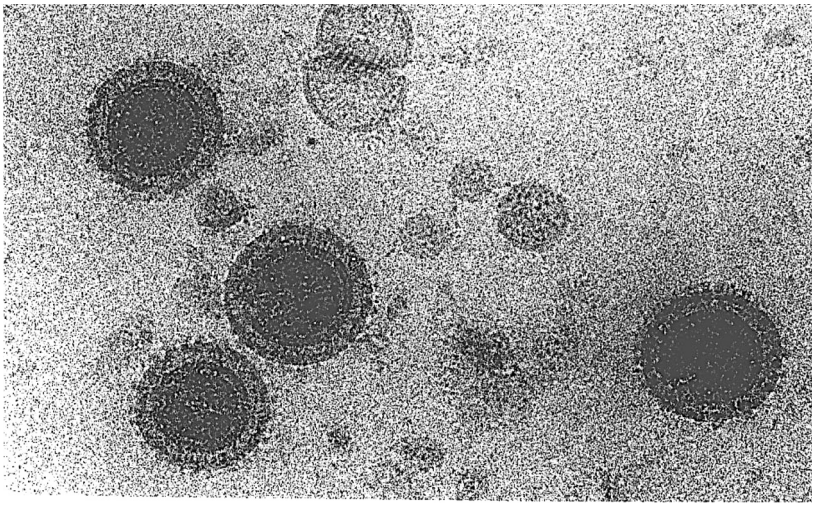


**Figure 2. MuLV virions as seen by cryo-electron microscopy.**

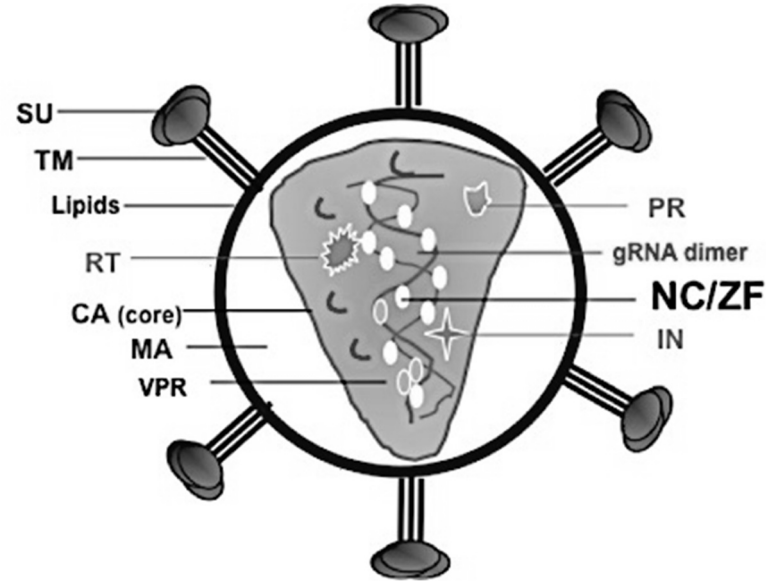



**Figure 3. Simple representation of the HIV-1 viral particle.**
The envelope formed of trimers of surface (SU) and transmembrane (TM) glycoproteins (9-15 molecules) is embedded in the lipid bilayer of cellular origin; the matrix (MA, 1500 molecules); the core with a core shell of capsid (CA, 1500 molecules); in the interior, the nucleocore with a genomic RNA in a dimeric form (gRNA dimer), coated by NC protein (NC/ZF, 1500 molecules), together with reverse transcriptase (RT), integrase (IN), protease (PR) and VPR (80 molecules each). Cellular RNA such as ribosomal RNA and tRNAs are also present, but not shown here. Of note, cellular proteins are also found inside the virion and at the surface of the particle, notably actin.

The NC protein is a basic protein of 55 residues in its fully processed form, with two mini-zinc fingers flanked by basic sequences (Fig. 4). It shows a high degree of sequence conservation in all isolates of HIV-1. Similar NC proteins with either one or two ZF structures are found in all Orthoretroviruses. NC proteins closely related to that of HIV-1 are encoded by all lentiviruses and alpharetroviruses, while gammaretroviruses code for a highly basic protein with a single ZF structure. In addition, a large fraction of mobile retroelements found in all eukaryotes code for a similar basic protein with a unique ZF (see Chapter I).

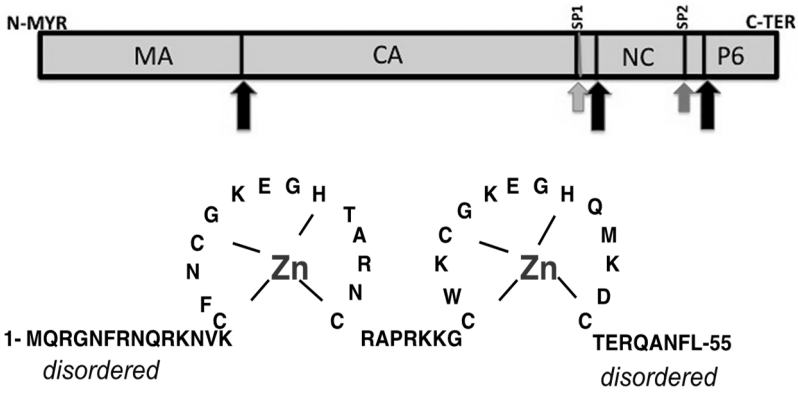



**Figure 4. NC protein of the AIDS virus, HIV-1.**
TOP. Simple representation of the Pr55 Gag polyprotein precursor where the matrix (MA), capsid (CA), nucleocapsid (NC) and p6 domains are indicated. The vertical arrows indicate protease-mediated cleavages. Small arrows indicate that the cleavage is slow, notably at the NC-p6 junction.
BOTTOM. The sequence of the NC protein is shown here using the one-letter amino acid code (positions 1-55). The extended form (NC-SP2) is also found in newly formed virions, which can be processed by the viral protease (PR) into the 1-55 NC and spacer peptide SP2. Note the ZFs where the $Zn^{2+}$ ion is coordinated by the CCHC residues. The ZF flanking sequences are basic. The NC protein is highly conserved in all isolates of HIV-1, including those showing resistance to highly active antiretroviral therapy (HAART).

According to H1 NMR analyses, NC protein has a flexible 3D conformation, where the disordered basic sequences can freely rotate around the structured ZFs (Fig. 4); as such the ZF protein belongs to the intrinsically unstructured protein class (IUPs) of proteins. NMR-based studies of the 3D conformation of NC protein of murine leukemia virus (MuLV), a gammaretrovirus, show that the N- and C-terminal sequences can also rotate around the unique structured ZF. In the absence of $Zn^{2+}$, both HIV and MuLV ZF proteins are completely unstructured. The conformation of the ZF protein of ancient retrotransposons, such as TYs of yeast, has also been investigated by H1 NMR, and the protein was found to be totally disordered.

As summarized in figure 5, NC is involved in almost all steps of the replication cycle of HIV-1, from reverse transcription to virus formation, core condensation, stability and structure.

This is highlighted by selected mutations, either in the zinc fingers (ZF), or in the flanking basic residues (see figure 6). As shown, mutations

in NC such as deletion of one ZF or point mutations H23C and H44C result in the formation of defective particles, the lack of a condensed core and the presence of DNA in place of RNA (Darlix et al., 2001 and references therein).

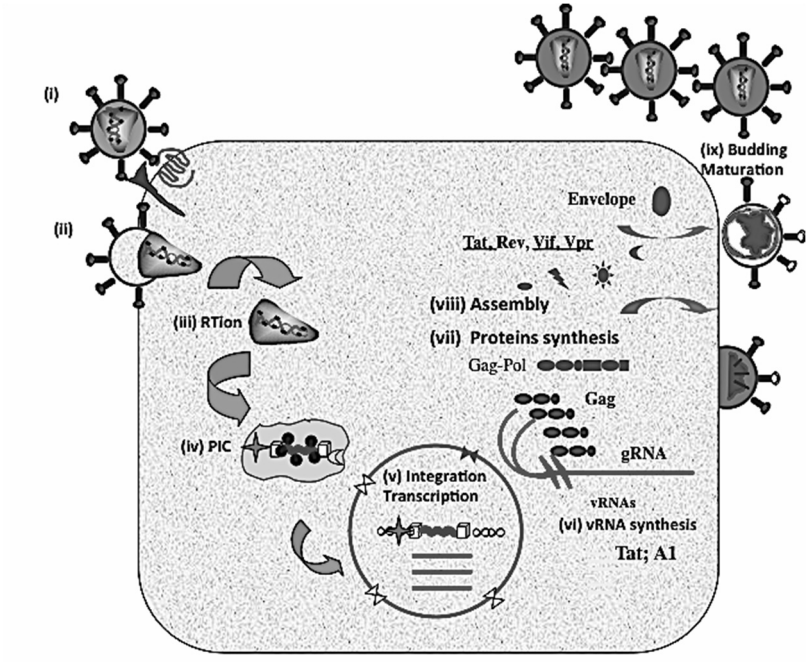
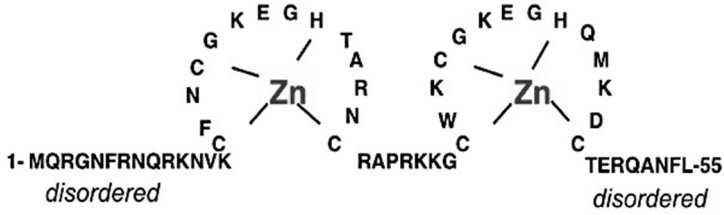



**Figure 5. Schematic representation of the replication cycle of HIV-1.**
Binding of the virus to the CD4 receptor and CCR5 coreceptor (i), upon fusion of the virus envelope with the plasma membrane, the core is delivered into the cytoplasm (ii), reverse transcription takes place in the core by RT chaperoned by NC and requires from 4 to 16 hours, depending on the cell type (A Cimarelli et al., 2016) (iii), this causes the formation of the preintegration complex (PIC), followed by interaction with the nucleopore complex, import of the viral DNA into the nucleus and integration by the integrase enzyme boosted by NC, a process that requires from 8 to 24 hours, also according to the cell type (Cimarelli et al; 2016) (v); viral RNA synthesis (vi); viral protein synthesis (vii); accumulation of the viral proteins under the plasma membrane and assembly, promoted by interactions between the 5' leader of the full length viral RNA (FL RNA) and Gag-NC (viii); completion of assembly under the plasma membrane;

virus budding with the concomitant maturation of Gag and Gag-Pol precursors by the viral protease, followed by maturation and core condensation, a series of steps that are considered to necessitate about 45 min (ix). Apart from being a constituent of the viral particle, NC is involved in the early steps (i to v), and in the late steps of the viral life cycle (vii, viii and ix).




| WT (pNL4.3) Deletion | | | (Del ZF1) | | | | (Del ZF2) | |
|---|---|---|---|---|---|---|---|---|
| Poi,v Mutation | R7R10K11S | | K14D | | H23C | P31L | | H44C |
| DATA | | | | | | | | |
| Infectivity | + | +/- | +/- | (-) | - | - | (-) | - |
| G RNA | + | + | + | (+) | + | + | (+) | + |
| vDNA | + | +/- | + | (+) | + | + | (+) | + |
| LTR | + | +/- | +/- | (-) | - | - | | - |
| Integ | + | +/- | +/- | | - | - | | - |
| Gag cond | + | + | + | (+) | + | + | (+) | _ |
| Core | + | + | + | (-) | - | - | (-) | |
| Stability | | | | | | | | |
| Virion DNA | | - - | - | (+) | + | | (+) | + |
| In vitro annealing | + | +/- | + | (+) | + | + | (+) | + |

**Figure 6: selected mutations in NC, what they tell us on its roles in virus production, reverse transcription, integration and ultimately on Gag assembly and virus production (see text).**

The above presentation of the HIV-1 viral particle and its replication cycle is most probably overly simplified for several reasons, as listed below:

(i) Particles produced by infected cells are heterogeneous in size and morphology and contain large amounts of cellular RNAs and proteins. Only 1 in 100 or 1000 virion particles are infectious *ex vivo*, so what are the true morphological differences between an infectious particle versus a non infectious one?

(ii) Particles tend to cluster and can massively infect naïve cells via virological synapses.

(iii) The genomic RNA is shown in a dimeric form coated by NC molecules. However, RNA complexes of very high molecular weight are found, raising the possibility that that trimeric and multimeric RNAs might coexist with the dimeric RNA in the particle.

(iv) NC molecules are believed to coat the genomic RNA, thus forming the nucleocapsid structure. However, it is entirely possible that such a nucleocapsid structure is more complex with molecular extension in the virus envelope. Indeed, peptides called SEVI found in the seminal fluid are prone to aggregation, a process that drastically enhances virus infectivity. Interestingly, these fibrils interact with NC *in vitro*, favouring the notion that SEVI stabilizes the nucleocapsid structure via direct interactions in a physiological milieu.

In conclusion, our view of the HIV viral particle in 2020 is more like a global vision of a cloud-like particle, far from being a well-organized molecular assembly of RNA, proteins, glycoproteins and enzymes. Yet do we have a clear view of the structure and functions of NC protein?

As the reader will see in the next chapters, genetic and virological analyses show that NC is a multifunctional factor; but how NC carries out such functions in virion structure, formation, reverse transcription, integration and assembly is far from being clear and the molecular mechanisms that could explain the roles of NC are still poorly understood.

Lastly, beside retroviruses and retroelements, NC-like functions also exist in many, if not all small RNA/DNA viruses, such as hepadnaviruses, caulimoviruses, orthomyxoviruses, picornaviruses, and flaviviruses.

# CHAPTER I

# FROM RNA TO DNA (AND BACK AGAIN): REVERSE TRANSCRIPTION AND REVERSE TRANSCRIBING VIRUSES

ROLAND IVANYI-NAGY

NTU Institute of Structural Biology, Nanyang Technological University
59 Nanyang Drive, Singapore 636921
Correspondence: roland.ivanyi-nagy@ntu.edu.sg

Short biography: Throughout his career, Roland Ivanyi-Nagy has always been interested in the characterization of medically important non-coding RNAs in different model systems. He earned his Ph.D. in virology from the Ecole Normale Supérieure de Lyon, France, where he described and characterized the nucleic acid chaperone activity of various viral RNA-binding proteins. He later carried out post-doctoral research at the University of Oxford, where he studied the role of non-coding RNAs in the regulation of antigenic variation in malaria parasites. Roland is currently a Research Fellow at Nanyang Technological University, Singapore, working on various aspects of telomere biology and RNA-RNA interaction networks.

## Reverse transcribing viruses

While all extant cellular life forms are derived from a single common ancestor and use double-stranded DNA as their genetic material, viruses display a remarkable diversity in their mechanism of information storage and expression (Baltimore 1971). Indeed, either single-stranded or double-stranded RNA or DNA can be incorporated into virions to transmit the viral genome between susceptible host cells, and the genetic information is expressed to produce viral progeny using enzymatic activities frequently unique to the virus world.

In particular, RNA viruses use two fundamental strategies for messenger RNA (mRNA) expression. While the majority of them encode an enzyme with RNA-dependent RNA polymerase (RdRp) activity, a handful of virus families (collectively referred to as reverse transcribing viruses) first copy their genetic information from RNA to an intermediate double-stranded DNA form, which is then transcribed by the canonical cellular transcription machinery to generate mRNAs for viral protein expression.

The process of reverse transcription—the transformation of genetic material from RNA to DNA—is mediated by the enzyme reverse transcriptase (RT), an RNA-dependent DNA polymerase (Baltimore 1970; Temin and Mizutani 1970). Although RTs are encoded by a wide variety of selfish genetic elements present in all kingdoms of life (Nakamura and Cech 1998; Gladyshev and Arkhipova 2011), RT activity apparently has not independently evolved in cellular life forms. Indeed, the cellular telomerase reverse transcriptase (TERT) enzyme, responsible for the maintenance of linear eukaryotic chromosome ends (Blackburn and Collins 2011), is itself believed to be derived from mobile genetic elements (Nakamura and Cech 1998).

Retroviruses—members of the medically important *Retroviridae* virus family—are reverse transcribing viruses that infect all vertebrate classes and cause a wide variety of serious disease, including malignancies, autoimmune diseases, immunodeficiencies, and neurological disorders in numerous animal species. Importantly, repeated cross-species transmissions from non-human primates, followed by viral adaptation, gave rise to human diseases of retroviral origin. Human immunodeficiency virus 1 (HIV-1) (Barré-Sinoussi et al. 1983), the causative agent of acquired immunodeficiency syndrome (AIDS), and human T-lymphotropic virus 1 (HTLV-1) (Poiesz et al. 1980), a virus causing adult T-cell leukemia and neurological disease in a subset of infected individuals, are members of the lentivirus and deltaretrovirus genera of retroviruses, respectively.

Most retroviruses package their genome as a dimeric complex containing two copies of positive-sense single-stranded RNA that, upon infection of a new host, is reverse transcribed into a dsDNA form using host cell-derived tRNA as a primer for DNA synthesis, followed by integration into the host genome as a provirus.

Retroviruses show a common genetic organization (Petropoulos 1997) (Fig. 1-1), with the canonical gag (group-specific antigen), pro (protease), pol (polymerase), and env (envelope) genes flanked on both ends by non-coding long terminal repeat (LTR) sequences that regulate proviral transcription. Gag structural proteins are required for viral particle formation and in most retrovirus genera are proteolytically processed by the virus-encoded protease to release several smaller proteins, including matrix (MA), capsid (CA), and nucleocapsid (NC) (Fig. 1-2). NC is usually formed of one or two CCHC-type zinc fingers, flanked by small, non-structured basic domains, and is packaged in the virion where ~1500 molecules of the protein cover and protect the viral genomic RNA. In addition, NC proteins are endowed with nucleic acid chaperone activities that are essential throughout virus replication (Darlix et al. 1995; 2011). Pol encodes the RT and RNase H activities necessary for the generation of DNA from the viral RNA, and the integrase domain mediating the integration of the reverse-transcribed genetic material into the host genome. Finally, the envelope gene codes for glycoproteins that mediate the recognition of susceptible host cells and viral entry (membrane fusion). Importantly, the env gene has been repeatedly lost and (re)acquired during evolution, leading to transitions between infectious, exogenous retroviruses (XRVs) and env-less retrotransposons. Besides these canonical factors, additional gene products—frequently involved in virus-host interactions—might be encoded by retroviruses in a lineage- or virus-specific manner (Fig. 1-1).

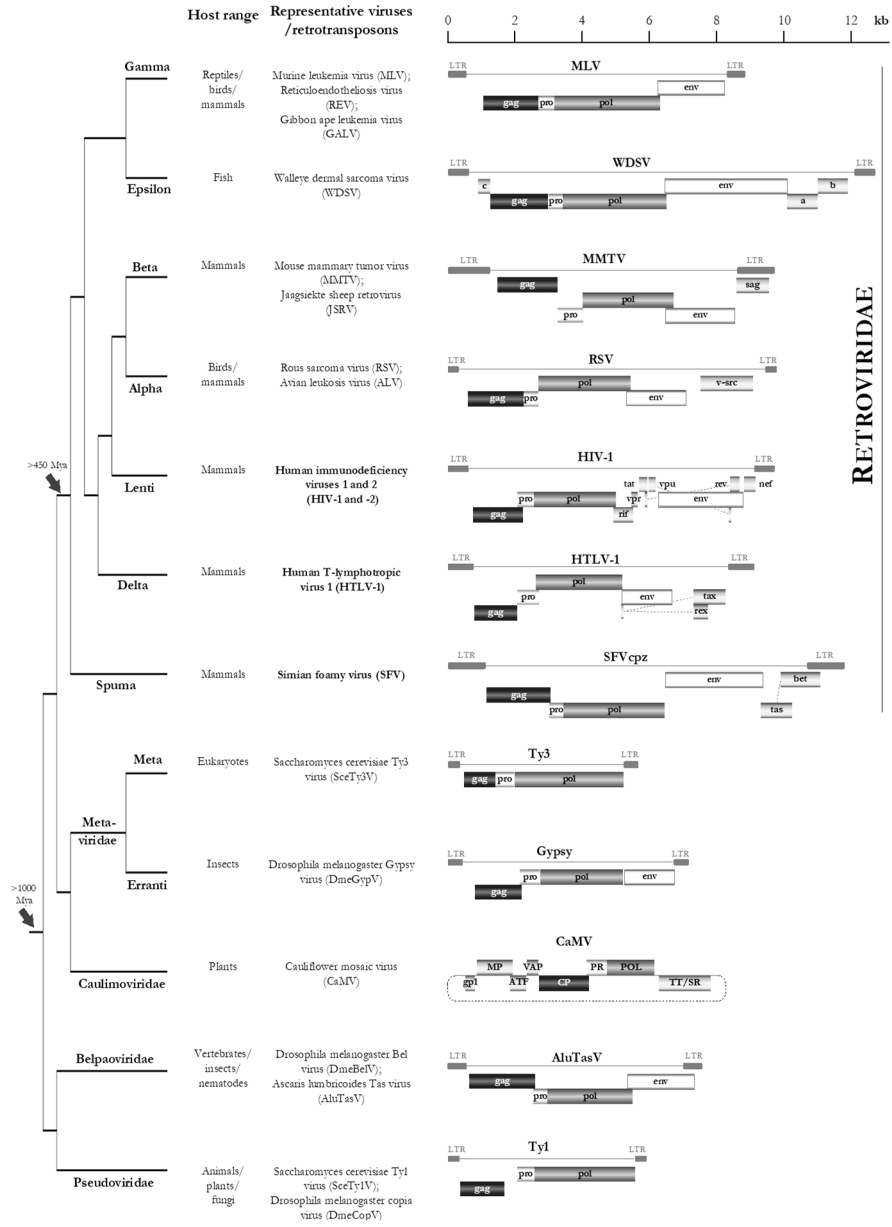

Host range
Representative viruses /retrotransposons
0 2 4 6 8 10 12 kb
Gamma
Reptiles/ birds/ mammals
Murine leukemia virus (MLV); Reticuloendotheliosis virus (REV); Gibbon ape leukemia virus (GALV)
MLV
Epsilon
Fish
Walleye dermal sarcoma virus (WDSV)
WDSV
Beta
Mammals
Mouse mammary tumor virus (MMTV); Jaagsiekte sheep retrovirus (JSRV)
MMTV
Alpha
Birds/ mammals
Rous sarcoma virus (RSV); Avian leukosis virus (ALV)
RSV
>450 Mya
Lenti
Mammals
Human immunodeficiency viruses 1 and 2 (HIV-1 and -2)
HIV-1
Delta
Mammals
Human T-lymphotropic virus 1 (HTLV-1)
HTLV-1
Spuma
Mammals
Simian foamy virus (SFV)
SFVcpz
RETROVIRIDAE
Meta
Eukaryotes
Saccharomyces cerevisiae Ty3 virus (SceTy3V)
Ty3
Meta-viridae
Erranti
Insects
Drosophila melanogaster Gypsy virus (DmeGypV)
Gypsy
>1000 Mya
Caulimoviridae
Plants
Cauliflower mosaic virus (CaMV)
CaMV
Belpaoviridae
Vertebrates/ insects/ nematodes
Drosophila melanogaster Bel virus (DmeBelV); Ascaris lumbricoides Tas virus (AluTasV)
AluTasV
Pseudoviridae
Animals/ plants/ fungi
Saccharomyces cerevisiae Ty1 virus (SceTy1V); Drosophila melanogaster copia virus (DmeCopV)
Ty1
LTR
gag
pro
pol
env
sag
v-src
tat
vpu
rev
nef
vpr
rif
tax
rex
bet
tas
MP
VAP
PR
POL
gpl
ATF
CP
TT/SR

**Figure 1-1**

**Genome organization of viruses and retrotransposons in the *Ortervirales* order.**
Viruses in the order share the major structural genes (gag/CP) and enzymatic activities (pro-pol), while many also encode taxon- or virus-specific accessory factors. The viral envelope gene was repeatedly lost and (re)acquired during evolution, and the viral genera can include both env-containing and env-less members. The phylogenetic relationship shown is based on (Krupovic et al. 2018). Note that the exact evolutionary relationship between the deeper-branching taxa remains somewhat uncertain.

The origin of the *Ortervirales* order is believed to reach back more than a billion years, while the first retroviruses appeared more than 450 million years ago (events indicated by arrows).

Reverse transcribing viruses are currently classified into six families, namely *Retroviridae*, *Metaviridae* (also known as Ty3/Gypsy retrotransposons), *Caulimoviridae*, *Belpaoviridae* (aka Bel/Pao retrotransposons), *Pseudoviridae* (aka Ty1/Copia retrotransposons), and *Hepadnaviridae* (Lefkowitz et al. 2018) (Fig. 1-1). Although members of the *Caulimoviridae* and *Hepadnaviridae* families (collectively known as pararetroviruses) differ in significant aspects of their genome organization and their replication strategy from viruses in the other families (i.e., they encapsidate circular, dsDNA genomes and they do not routinely integrate into the host genomic DNA), phylogenetic analyses based on RT sequences suggest that reverse transcribing viruses are monophyletic, derived from a common ancestor (Xiong and Eickbush 1990). While hepadnaviruses encode a unique capsid protein of unrelated origin, all the other families share significant homology in their capsid and nucleocapsid genes as well (Krupovic and Koonin 2017b), including the near-ubiquitous presence of Zn-knuckle(s) in Gag (Fig. 1-2). Common descent and conserved genetic organization, together with other shared features—such as a common tRNA-based priming mechanism—resulted in the recent creation of a new viral order, *Ortervirales*, regrouping the *Retroviridae*, *Metaviridae*, *Pseudoviridae*, *Belpaoviridae* and *Caulimoviridae* clades (Krupovic et al. 2018). Based on the ubiquitous phylogenetic distribution of reverse transcribing viruses and their integrated forms (Fig. 1-1), the (hypothetical) ancestral virus giving rise to the *Ortervirales* order—already encoding capsid and zinc-knuckle-containing nucleocapsid modules, as well as protease, reverse transcriptase, and RNase H enzymatic activities—might have existed prior to the divergence of the plant, fungi, and animal kingdoms ~1-2 billion years ago (Krupovic and Koonin 2017b).

## Paleovirology of LTR-containing retroelements

Occasionally, retroviruses infect and integrate their genetic material into germ-line cells (oocytes, sperm cells, or their precursors) (Boeke and Stoye 1997; Stoye 2012). The integrated provirus (termed endogenous retrovirus – ERV) is vertically transmitted and inherited by subsequent generations in a classical Mendelian fashion. Following the initial endogenization process, ERV copy numbers may be further amplified through several distinct mechanisms, such as by reinfection of neighbouring cells, retrotransposition within the same cell, or through amplification of the genome region where the integration had occurred. Although recently acquired proviruses might retain their infectious abilities and can spread both vertically and horizontally, on evolutionary timescales most ERVs will eventually acquire and accumulate inactivating mutations, deletions, or truncations. Specifically, loss of the env gene is associated with loss of infectious potential, and at the same time results in increased retrotransposition capacity (Magiorkinis et al. 2012).

Since integration is an obligate step in their replication cycle, LTR-containing reverse transcribing viruses are uniquely positioned to give rise to endogenous viruses. Surprisingly however, a large number of RNA viruses that have no DNA stage in their life cycle (such as filoviruses and bornaviruses) have also left fossils (called endogenous viral elements – EVEs) in their host genomes (Katzourakis and Gifford 2010; Holmes 2011), possibly by hijacking a reverse transcriptase from cellular retroelements or from a co-infecting retrovirus. EVEs in general (and ERVs in particular) provide unprecedented insights into evolution, serving as molecular fossils documenting ancient viral infections (Hayward, Grabherr, and Jern 2013; Hayward, Cornwallis, and Jern 2015) that allows us to reconstruct the evolutionary history of virus families and their co-evolution with their hosts (Emerman and Malik 2010). Importantly, if an endogenous (retro)virus is found in a conserved (orthologous) position in the genome of two or more species, then we can confidently conclude that the integration event must have happened in a common ancestor of these species, and can infer a minimum age for the virus group as that of the last common ancestor of the hosts.

Indeed, analysis of ERVs derived from the spumaretrovirus genus of *Retroviridae* suggests that retroviruses originated—presumably from a lineage of fish retrotransposon family—more than 450 million years ago in the Ordovician period, coinciding with (Aiewsakun and Katzourakis 2017), or even pre-dating (Hayward, Cornwallis, and Jern 2015; Hayward 2017), the appearance of jawed vertebrates. Since then, retroviruses have

diverged into seven genera with unique differences in their genome organization and host range (Fig. 1-1), infecting and occasionally colonizing the genome of all vertebrate classes.

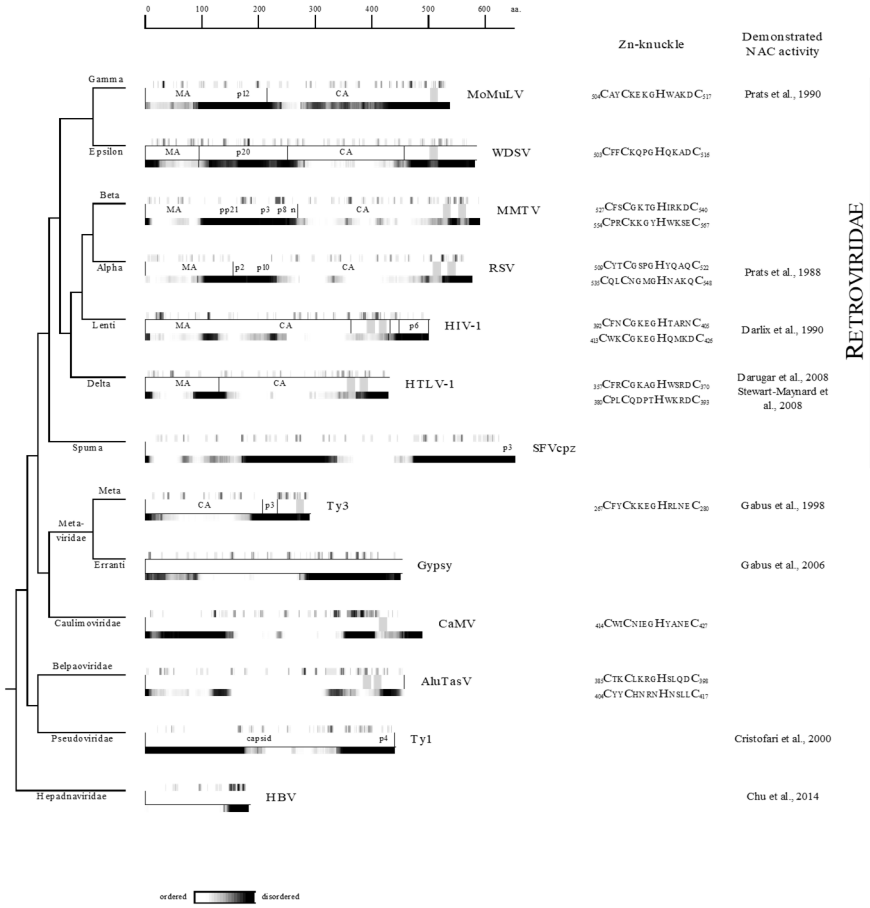



**Figure 1-2.**
**Common themes in the domain organization of the structural proteins of reverse transcribing viruses.**
Gag proteins are frequently further processed to yield several mature products, including matrix (MA), capsid (CA), and nucleocapsid (NC) proteins. The charge distribution along the proteins is depicted on top (red: basic regions; blue: acidic regions). Disordered segments—shown below the protein domains—were

predicted using the MFDp webserver (Mizianty et al. 2010). Highly flexible protein regions are shown in red, while well-folded segments are in green. Note that regions associated with nucleic acid chaperone (NAC) activity frequently show a highly basic character and high flexibility, and usually contain one or two CCHC-type Zn-knuckle(s) (mauve rectangles). The position and sequence of the Zn-knuckles–bearing the consensus Cx2Cx4Hx4C sequence–is shown on the right. In certain taxa, such as spumaretroviruses (Hamann et al. 2014) and hepatitis B virus (HBV) (Chu et al. 2014), NC-like functions are carried out by short, arginine-rich protein domains, without the involvement of the canonical Zn-knuckles.
Virus name abbreviations are explained in Figure 1-1.

The last two decades—following the sequencing of the first eukaryotic organism (Goffeau et al. 1996) and the initial report on the first draft of the human genome (Lander et al. 2001)—have witnessed major advances in genome sequencing technologies and informatics, resulting in unprecedented insights into biological function and evolution. At the time of this writing, the genome sequence of thousands of species has already been deciphered and efforts are under way to catalogue and sequence ~1.5 million eukaryotic species within the framework of the Earth BioGenome Project (EBP) (Lewin et al. 2018). Analyzing the genomes of life forms in all their diversity will undoubtedly contribute—among other important goals—to a better understanding of (retro)virus evolution and virus-host co-evolution.

## Endogenous retroelements as a source of evolutionary innovation

Colonization of vertebrate genomes by LTR-containing retroelements has been extremely successful. Indeed, it is estimated that ~8% of the human genome is derived from identifiable LTR-elements (Lander et al. 2001), substantially more than the fraction contributed by protein coding genes. In some vertebrates, colonization by transposable elements has resulted in bloated genome sizes, such as in salamanders, whose giant genome can reach 120 billion base pairs (40 times that of the human genome size), partly due to expansion of LTR-retrotransposon classes (Sun and Mueller 2014).

Although long considered “junk” or “selfish” DNA with mostly deleterious or at best neutral effects, we now realize that endogenous retroelements may also serve as raw material for evolutionary innovation, and have contributed to the emergence of new functions and regulatory potential, increasing the fitness of host organisms (Stoye 2012; Kaneko-Ishino and Ishino 2012; Naville et al. 2016).

In the simplest scenario, integration of ERVs near coding regions can result in the modulation of the expression patterns of human genes by proviral LTRs that can provide tissue-specific or alternative promoter, enhancer, insulator, or repressive activities for transcription regulation (Cohen, Lock, and Mager 2009; Chuong, Elde, and Feschotte 2017). In addition, genes of retroviral origin have been frequently "hijacked" during evolution to carry out functions adaptive for the host organism.

A fascinating example for the exaptation of ancestrally endogenized sequences is the placenta-specific expression of two human ERV-derived envelope genes with intact coding potential, syncytin 1 (Mi et al. 2000) and syncytin 2 (Blaise et al. 2003). The outer cellular layer of the placenta—the syncytiotrophoblast—is formed by the fusion of trophoblast cells and plays essential roles in mediating nutrient and gas exchange between the mother and the developing fetus. By virtue of their cell-fusion activities, syncytins orchestrate syncytiotrophoblast formation and placenta development. In addition, due to the presence of an immunosuppressive domain in its transmembrane region, syncytin 2 might be also involved in the protection of the fetus against rejection from the maternal immune system (Mangeney et al. 2007). Importantly, domesticated Env expression has been shown to be absolutely required for normal placenta development and successful reproduction in both mice and sheep (Dunlap et al. 2006; Dupressoir et al. 2009).

As a remarkable example of convergent evolution, independently domesticated retroviral envelope genes regulating placenta development and function have been described in several placental (eutherian) mammalian clades, in marsupials (who develop a short-lived placenta), and in the viviparous *Mabuya* lizard (Dupressoir et al. 2005; Lavialle et al. 2013; Cornelis et al. 2015; 2017). The independent capture of retroviral envelope at the transition to viviparity in both mammals and reptiles suggests that Env domestication might have been essential for the emergence of placentation (Cornelis et al. 2017).

Besides Env gene products, retroviral Gag proteins have also been co-opted to carry out important physiological functions. Indeed, more than 80 genes putatively derived from the Gag domain of *Metaviridae* retrotransposons have been identified bioinformatically in mammalian genomes (Campillos et al. 2006). Although the function – if any – for most of these sequences remains to be determined, there is overwhelming evidence supporting the importance of the Arc (Activity-Regulated Cytoskeleton-Associated) protein in neuronal function (Plath et al. 2006; Shepherd 2018).

The above-mentioned examples illustrate the important contributions of retroelement-derived sequences to various host functions, including to fetal development and memory formation, features that we consider to be at the heart of what is essentially human. Due to the abundance of retroelements in our genomes, undoubtedly many more similar examples will be uncovered in the future.

## The ultimate origin of viruses and viral nucleic acid chaperones

Due to the unique molecular fossils left in our genome by ERVs, we now know that retroviruses have infected, colonized, and co-evolved with our ancestors from the dawn of vertebrate evolution and that reverse transcribing viruses are more than a billion years old (Hayward 2017; Krupovic and Koonin 2017b).

What can we tell about the ultimate origin of reverse transcription and the provenance of the key viral genes? Generally, three models for the origin of viruses have been considered: (*i*) descent from a primordial virus/replicon world pre-dating the occurrence of cellular life, (*ii*) descent from obligate intracellular parasitic (cellular) life forms by reductive evolution, and (*iii*) escape and gain of autonomy/infectivity of genes from cellular life forms (Krupovic, Dolja, and Koonin 2019).

Importantly, homologues of viral replication proteins are notably absent in extant organisms, with a few exceptions—such as the telomerase reverse transcriptase—that are clearly derived from mobile elements. In addition, the evolution of RdRp-like and RT-like activities must have been essential for the existence of an RNA-peptide/protein world and for the transition to the current DNA-RNA-protein world, respectively. Thus, it has been suggested that the viral replication machinery is a remnant of a primordial, pre-cellular replicon pool (Krupovic, Dolja, and Koonin 2019). In contrast, proteins with significant structural homology to viral capsid/nucleocapsid proteins are abundant in extant prokaryotic and eukaryotic species, suggesting that viral structural genes might have been captured from infected cells (Krupovic and Koonin 2017a). For example, the HIV-1 NCp7 protein was suggested to be structurally closely related to cellular Zn-knuckle containing proteins, such as Lin28 and Air2p (Krupovic and Koonin, 2017b). Thus, the different functional modules of current viruses (for replication vs structure formation) might have arisen independently, at different times during their evolution (Koonin, Dolja, and Krupovic 2015; Krupovic, Dolja, and Koonin 2019).