A short biased history of RNA viruses

DANIEL KOLAKOFSKY
Department of Microbiology and Molecular Medicine, University of Geneva Medical School, 1211 Geneva, Switzerland

Among its many roles, RNA can also act as a viral genome. In the majority of cases these genomes are single-stranded RNA, and both the form that can be directly translated by ribosomes (plus strands) and that which is the complement of the mRNA (minus strands) are used. As a mark of RNA's versatility as genetic material, there are also "ambisense" genomes where the same ssRNA has regions that are (+) strand or (−) strand, as well as dsRNA genomes.

RNA viruses were discovered about the same time as a series of momentous events in molecular biology; mRNA was discovered as the immediate carrier of genetic information for protein synthesis, the genetic code was quickly deciphered with Khorana's synthetic trinucleotides, and a modified form of methionine (N-formyl-methionine) that minimally resembled a peptide was found to initiate protein synthesis in bacteria. To study protein synthesis in more detail, RNAs of random sequence made with template-independent polynucleotide phosphorylase were used as ersatz mRNAs. Natural mRNAs at the time were hard to get hold of, especially in intact form, until it was appreciated that some recently-found bacteriophages contained RNA genomes that were also *bona fide* mRNAs. In contrast to cellular mRNAs like reticulocyte β-globin, phage RNA genomes came enclosed in "cast-iron" capsids that were easy to purify intact, free of RNase.

Although the first RNA phages discovered (f2/MS2/R17) provided *bona fide* mRNAs for protein synthesis, and, remarkably, were also infectious upon transfection into protoplasts, their RNA-dependent RNA polymerases (RdRp) were next-to-impossible to purify from infected bacteria in active form. Fortunately, the RdRp of a closely-related bacteriophage, Qβ, found in Japan, could be purified in a highly active form. Qβ RdRp opened the study of RNA-dependent RNA synthesis in vitro, and the role of the RNA genome in the virus life cycle. This RdRp was found to be composed of 4 subunits: only one, the polymerase subunit itself, was encoded by the virus. The 3 others were host proteins; S1, a ribosome-associated protein, Tu, the factor which carries aa-tRNA to the ribosome and aligns it in the A site during protein synthesis and Hfq (host factor Q) which is now known to bind to a number of regulatory RNAs in bacteria. The ability to carry out both protein and RNA synthesis in vitro resolved one of the obvious dilemmas of (+)RNA virus replication. This viral replication requires the prior translation of the polymerase subunit. However, if ribosomes translate the viral genome in the 5′ to 3′ direction, and the synthesis of the viral (−)antigenome must start at the genome 3′ end and proceed in the opposite direction, what happens when the RdRp runs into the translating Rb? It turns out that the viral RdRp does not simply resort to brute force and knock the translating Rb out of the way; it doesn't need to. The viral genome is folded such that translation begins mostly, if not exclusively, at the coat protein gene, and all further translation depends on coat synthesis. Once the polymerase subunit is synthesized and associates with its host proteins, its S1 subunit binds to the genome near its 3′ end as well as at a site just upstream of the coat gene, where it acts to repress further initiation of the coat protein and subsequently all protein synthesis. Once the genome is thus freed of ribosomes, RdRp can initiate at the genome 3′ end and continue antigenome synthesis unhindered. This automatically leads to a self-regulatory system, where the relative use of the genome as template for translation or replication is determined by the RdRp concentration; when RdRp is sufficient, further synthesis of the polymerase subunit is repressed; when insufficient, more polymerase subunit is synthesized. Moreover, the (−)antigenome RNA does not anneal to its (+)genome template during its synthesis, but is released as free ssRNA, as these viral RNAs apparently fold on themselves as they are being made, presumably as part of a folding program. RNA viruses are nothing if not good examples of intelligent design.

(+)RNA virus genomes like those of Qβ, polio, and HCV are infectious as pure RNA because their genomes are directly translated upon infection to initiate the virus life cycle. Vesicular stomatitis virus (VSV), one of the first animal viruses to be studied, because like polio, it grew to high titers in cell culture and "plaqued well," was found by Baltimore's group to have evolved a very different replication program. VSV genomes and those of other (−)RNA viruses like influenza and Ebola, are the complements of the mRNAs. Viral replication must thus begin here by mRNA synthesis from the (−)genome, and hence the viral RdRp must be pre-packaged within

Corresponding author: daniel.kolakofsky@unige.ch
Article and publication date are at http://www.rnajournal.org/cgi/doi/10.1261/rna.049916.115. Freely available online through the *RNA* Open Access option.

the (−)virion. This simple fact was an eye-opener to this observer, because unlike Qβ RdRp whose purification was difficult and tedious, producing small amounts of enzyme that were quickly consumed, some (−)RNA viruses like VSV were easy to purify and contained a highly active RdRp. The task of studying this RNA synthesis was thus much easier. It also paved the way for the discovery of reverse transcriptase.

A hallmark of (−)RNA genomes is that they are never free, but always found within helical nucleocapsids (NCs, also called RNPs), a unique structure in nucleic acid biology where the genome RNA is sheathed within multiple copies of the nucleocapsid protein (N). For (+)RNA viruses where the (+)genome is also the mRNA, there is generally 100 times more (+)genomes than (−)antigenomes, as the only function of the latter is to act as a template for the former. In the case of (−)RNA viruses, this disparity in the relative abundance of complementary RNAs is not possible as large amounts of both (−)genomes and mRNAs are required. It is the confinement of (−)genomes within NCs that prevents these complementary viral RNAs from annealing and neutralizing each other. Just how RNA synthesis takes place on a structure whose ssRNA template is so tightly enclosed that it is resistant to RNase digestion is unclear, but presumably the N subunits are transiently displaced as the RdRp traverses the template. (−)RNA viruses (NSV) themselves come in two flavors, those with a single nonsegmented genome (nsNSV), and those with 2 to 8 genome segments (sNSV). sNSV transcribe a single mRNA from each segment, or two mRNAs from ambisense segments, and these are initiated with a capped primer snatched from a host mRNA, and mostly polyadenylated and terminated by RdRp repetitively copying (stuttering on) a stretch of template uridylates. Ambisense mRNAs are a special case, and terminate on strong stem-loops. nsNSV contain 5–10 genes in tandem separated by intergenic regions containing polyadenylation-stop signals followed by gene start signals. RdRp enters the nsNSV template at its 3′ end and each mRNA is initiated with ATP in turn as RdRp traverses the genome, and then enzymatically capped by the vRdRp itself, when the nascent mRNA is 40–50 nt long. Capping at this point is essential for continued mRNA synthesis, and the mRNAs are polyadenylated and terminated as for non-ambisense sNSV. During (−)genome replication, RdRp disregards the start/stop signals, producing an exact complementary (+)antigenome copy fully assembled with N protein, presumably because synthesis of this RNA and its assembly with N protein are coupled.

Given their limitation on genome size, many RNA viruses pack as much coding info into their genomes as possible using overlapping ORFs, and have unusual mechanisms to gain access to them. Many of these mechanisms occur during translation, such as leaky Rb scanning to alternate start codons, or programmed ribosomal frameshifting (PRF) to gain access to downstream ORFs. For some nsNSV, like those of the *Paramyxovirinae* subfamily (e.g., measles, mumps, and Sendai virus) and Ebola filovirus, there is also a form of programmed transcriptional frameshifting (PTF). The most egregious example of this is parainfluenza virus type 3 (PIV3), where a stretch of 300-odd nucleotides in the middle of the P gene is translated in all 3 frames. The nsNSV P gene codes for the P protein, an essential vRdRp cofactor common to all these viruses and all but one of these viruses also codes for a V protein that neutralizes host defenses. P and V use the same start codon and thus share N-terminal regions, but a fixed fraction of the P gene mRNA will contain one or more additional G residue, inserted at an mRNA-editing site in the middle of the gene. This transcriptional frameshifting is as programmed as its ribosomal counterpart, as the pattern of G insertions clearly reflects the ORF possibilities of each virus; e.g., for measles and Sendai viruses where the uninserted mRNA codes for P and there is only a single alternate ORF downstream, 30% of the mRNA contain a single G residue (to switch to the V ORF) and mRNA with further G insertions are rare; for PIV3 where there are two alternate ORFs downstream, one to six Gs are inserted roughly equal frequency, and for mumps-like viruses where the uninserted mRNA codes for V rather than P, two Gs are inserted at high frequency to access the downstream P ORF.

This programmed pattern of G insertions occurs by the initial reiterative copying of a single template C (or CC for mumps-like viruses) within a "slippery sequence" (e.g., 3′ UUUUUUC**C**C 5′ for SeV). The mechanism seems to be that the slippery sequence and the immediate sequence upstream induces RdRp to pause while copying the critical template Cs. This allows the 3′ end of the nascent mRNA to realign itself upstream on the (−)genome allowing for G:U pairs, such that the critical template C (or CC for mumps-like viruses) can be copied a 2nd time before vRdRp moves on, thus shifting the ORF at the editing site. Presumably the length of the pause determines whether this process will repeat itself. Most of the information for this PTF is contained within a very limited sequence, as SeV (3′ $UugU_6C_3$; +1 at high frequency) can be induced to add 1 to 6 Gs at roughly equal frequency similar to PIV3 simply by placing the PIV3 sequence (3′ $UaaU_6C_3$) in SeV. Hexamer phase of the ectopic PIV3 editing site (see below), however, also plays a role, as the pattern of G insertions reverts to that of SeV in two of the six hexamer phases. N apparently remains closely associated with RdRp as it is transiently displaced from the template during mRNA synthesis, and its hexamer phase presumably also informs the length of the RdRp pause.

All parainfluenza viruses that use PTF also strictly adhere to the "rule of 6." To wit; genome length must be precisely a multiple of 6 (no exceptions are tolerated), presumably because each N protein is associated with precisely 6 nucleotides, and the hexamer phases of all promoters and other regulatory signals are in fact conserved. One explanation for this association is that this is a way to avoid these G insertions being fixed as hard copies of the genome should they occur, even at a reduced frequency, during genome replication. Such genomic insertions would strongly reduce P

protein synthesis from this gene to levels incompatible with virus viability (in some cases from 70% to 2%). PTF in Ebola virus (EBOV), in contrast, occurs within its GP gene, where the uninserted mRNA codes for a truncated soluble form of GP, and the mRNA with a single additional A inserted within a stretch of 7 adenosines codes for the full length attachment protein. Given that (i) the slippery sequence here is a homogenous stretch of uridines, the nascent mRNA 3′ end at the editing site can realign itself both upstream and downstream, creating both insertions and deletions, and (ii) the precise length of the Ebola genome is not constrained, the Ebola virus genome in fact exists in two forms, containing either 7 or 8 uridines. Passage of EBOV/7U in cell culture selects for EBOV/8U, and passage of EBOV/8U in guinea pigs back-selects for EBOV/7U. This selection is presumably due to environmental constraints, a property that may be advantageous for a virus infecting different hosts.