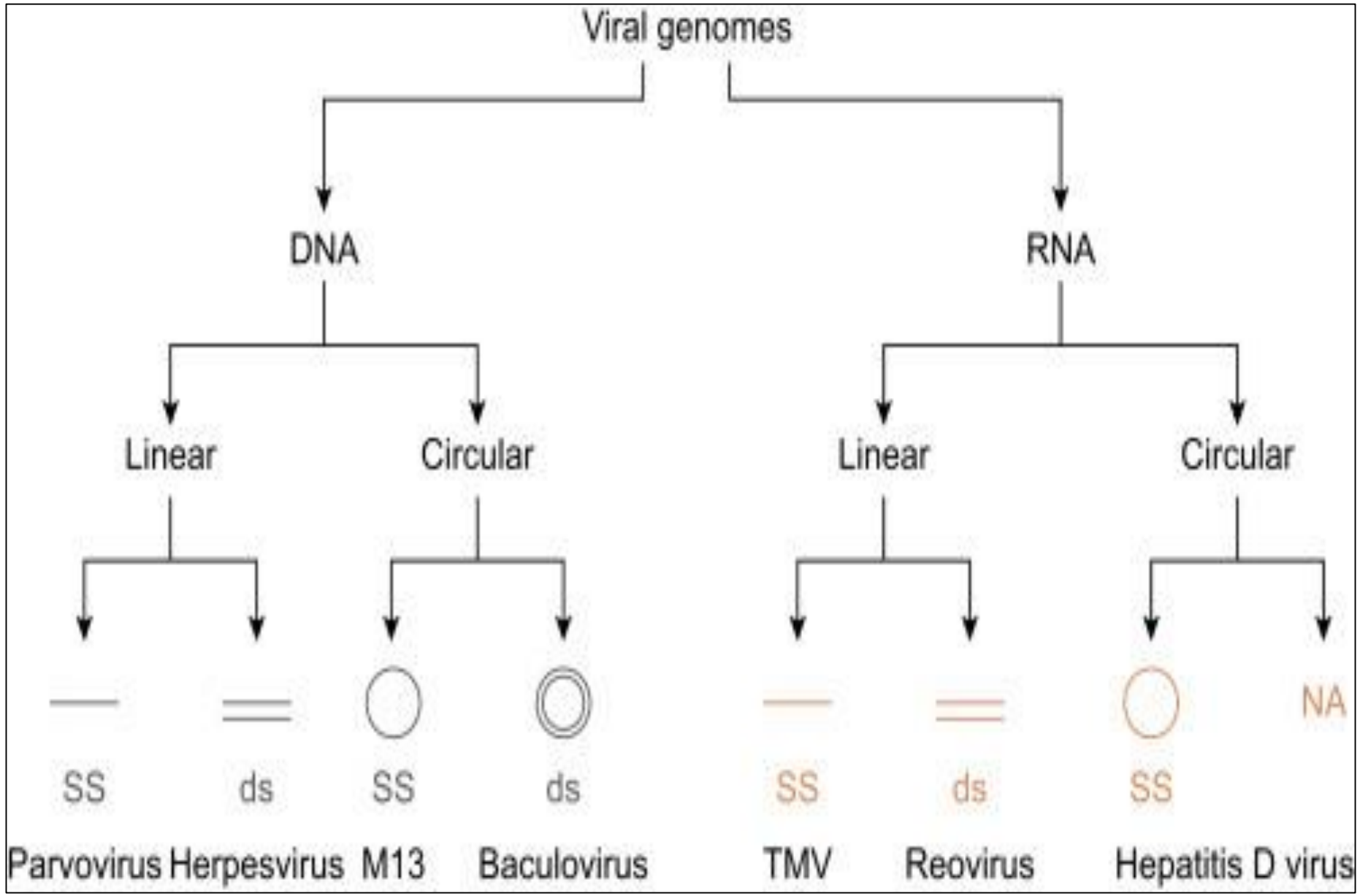
Salient features of viral genome

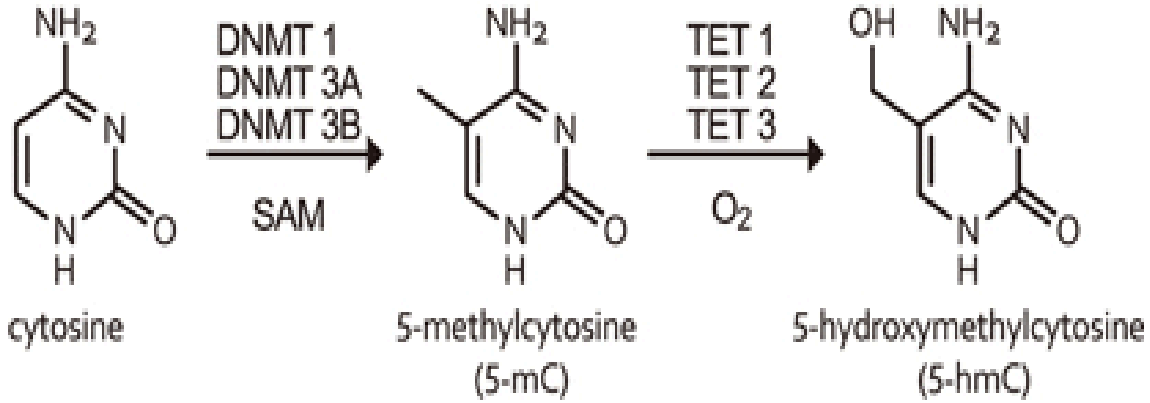
Viral Genomes

Viral genomes consist of DNA or RNA only, never both. DNA and RNA molecules can be double stranded or single stranded, linear or circular, segmented (composed of multiple pieces of nucleic acid) or nonsegmented. Strictly speaking, a genome segment is an individual and unique piece of nucleic acid among multiple pieces comprising one whole viral genome. For example, the influenza A virus has segmented genome comprised of eight ssRNA segments . Herpesviruses, which have nonsegmented genomes composed of one linear dsDNA molecule, have the so-called UL (unique long) and US (unique short) segments, corresponding to regions/portions of their genomes flanked by repeats.
The HIV genome can be somewhat confusing too with each virion carrying two copies of the same ssRNA molecule. The HIV genome is considered nonsegmented, and the two ssRNA molecules are called copies not segments. In many viruses the genome ends contain repeated sequences, chemical modifications, or secondary structures, which often have regulatory functions. Genomes are tightly packed inside the capsids and frequently the genome and the capsid are collectively called nucleocapsid. Amazingly, viruses are able to execute productive infection and of course make us sick with very limited genetic information. The flu virus genome, for example, contains only 15,000 nucleotides. For comparison the human genome is 3,200,000,000 nucleotides or approximately 200,000 times longer.


Viral genomes
DNA
RNA
Linear
Circular
Linear
Circular
SS
ds
SS
ds
SS
ds
SS
NA
Parvovirus
Herpesvirus
M13
Baculovirus
TMV
Reovirus
Hepatitis D virus

Unusual Bases in DNA: *The T4 Genome*

Although the bases most commonly present in DNA are adenine, guanine, cytosine and thymine, other bases have also been found. In some viruses (e.g., PBS 1 and PBS 2) uracil occurs in place of thymine in DNA. Also, in some bacterial viruses (bacteriophages) cytosine is replaced by 5-hydroxymethyl cytosine (HMC). The variations of C, A and G in DNA can be considered to be the result of methylation of these bases.



The DNA is specially modified HMC-DNA, meaning that (16%) of the cysteine bases are chemically modified into glucosylated hydroxymethyl cytosine (HMC). This makes the DNA resistant to endonucleases (nucleic acid degrading enzymes), such as host endonucleases which digest foreign DNA or the T4 endonucleases which digest host DNA.

TMV genome

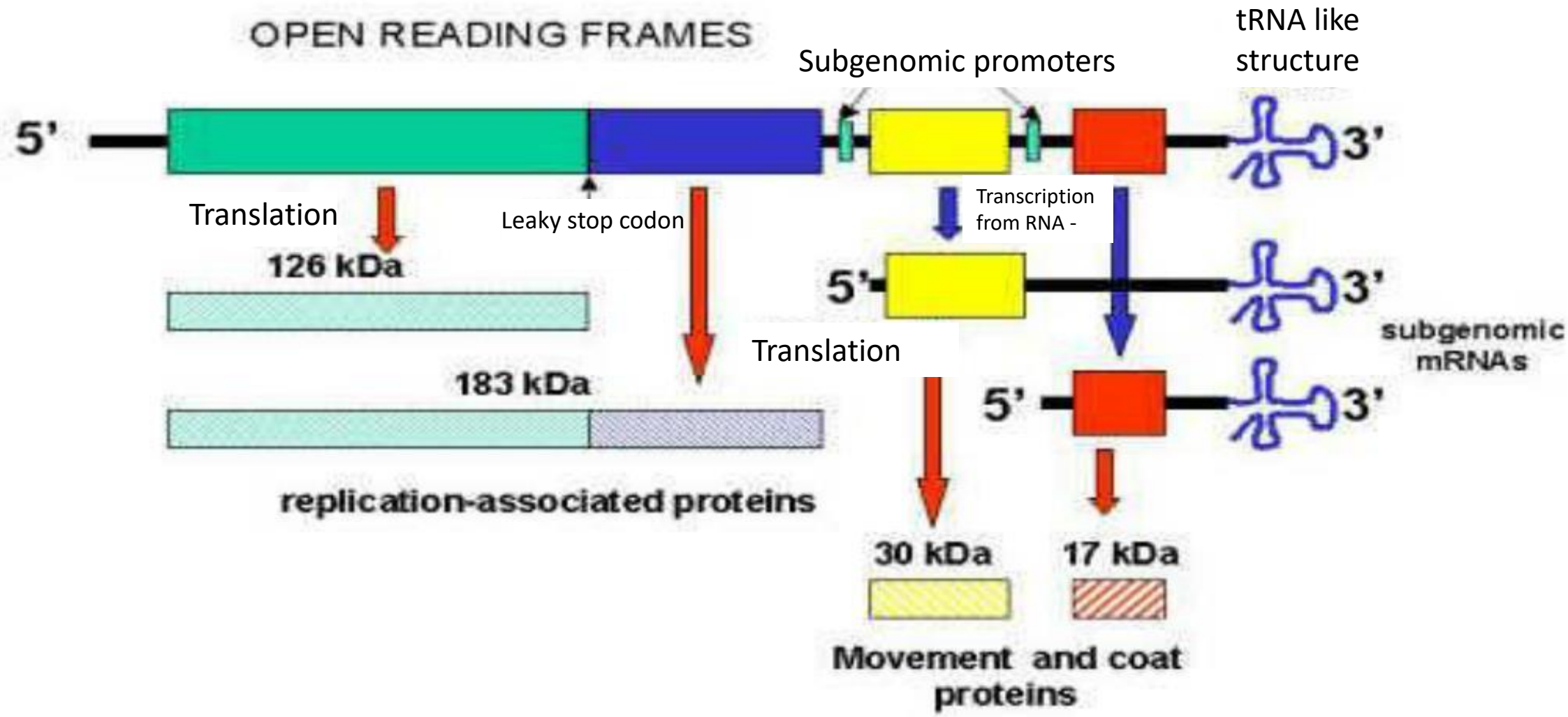
- The TMV genome consists of a 6.3–6.5 kb single-stranded (ss) RNA. The 3’-terminus has a tRNA-like structure, and the 5’ terminus has a methylated nucleotide cap. (m7G5’pppG).

- The sequence of the 5' noncoding region of tobacco mosaic virus RNA has been determined. The noncoding region is 68 nucleotides long and is unusual in that it contains no internal guanosine residues

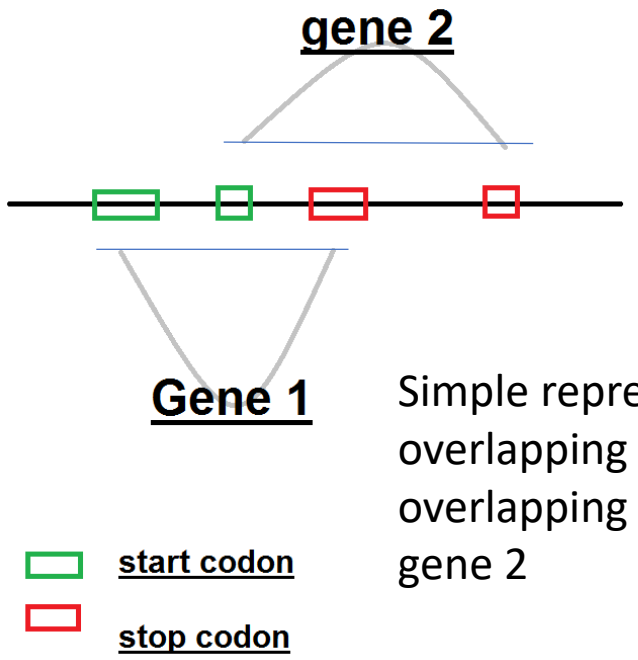
- The guanosine-free tract is terminated by the first potential initiation codon in the RNA molecule and several lines of evidence suggest that this AUG triplet is operational in initiating viral protein synthesis

Nucleotide Sequence at the 5' Extremity of Tobacco-Mosaic-Virus RNA 1. The Noncoding Region (Nucleotides 1 - 68) Kenneth RICHARDS, Hubert GUILLEY, Gerard JONARD, and Leon HIRTH Laboratoire des Virus, Institut de Biologie Moleculaire et Cellulaire du Centre National de la Recherche Scientifique, Strasbourg

Genes encoded by *Tobacco mosaic virus*



Overlapping genes



Simple representation of overlapping genes: partial overlapping of gene 1 and gene 2

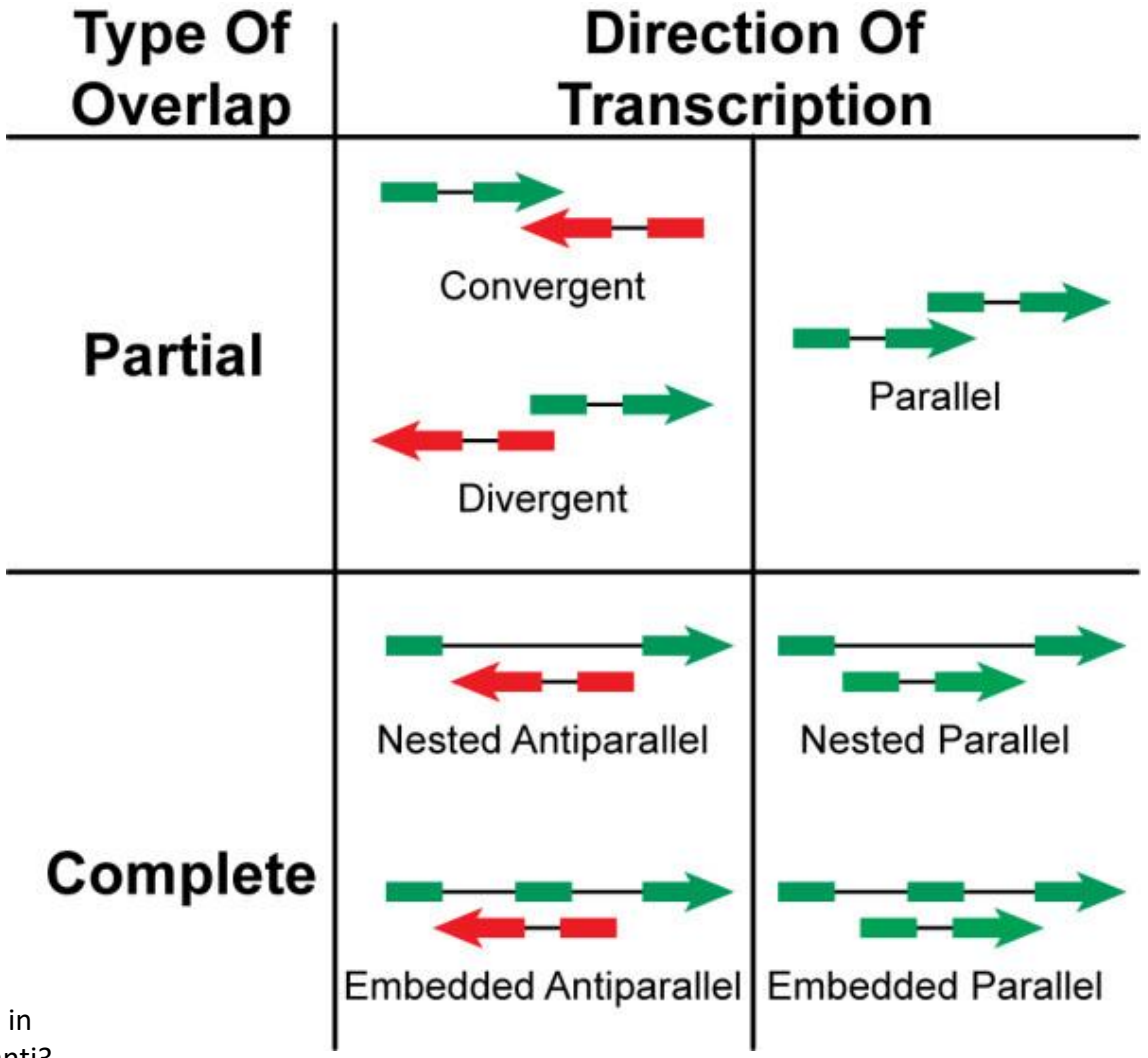
Overlapping genes are defined as a pair of adjacent genes whose coding regions are partially overlapping. In other words, a single stretch of DNA codes for portions of two separate proteins. Such an arrangement of genetic code is ubiquitous. Many overlapping genes have been identified in the genomes of **prokaryotes**, **eukaryotes**, **mitochondria**, and viruses.

Overlapping genes enable the production of more proteins from a given region of DNA than is possible if the genes were arranged sequentially. Indeed, for the bacteriophage PhiX174, overlapping of genes is necessary. The amount of DNA present in the circular, single-stranded DNA genome of this virus would not be sufficient to encode the eleven bacteriophage proteins if transcription occurred in a linear fashion, one gene after another.

overlapping gene clusters (OGC)

Classes of overlapping genes. OGC classification was based on the overlap extent (complete or partial) and on the reciprocal direction of transcription of the involved genes (same or opposite strand).

Convergent overlaps involve the 3' termini of both genes, while **divergent overlaps** involve the 5' ends (UTR and/or CDS).
Complete overlap occurs when the entire sequence of one gene is contained within another gene.
In nested OGCs one gene lies completely within an intron of the other, while **embedded genes** can share more than one intron or exon



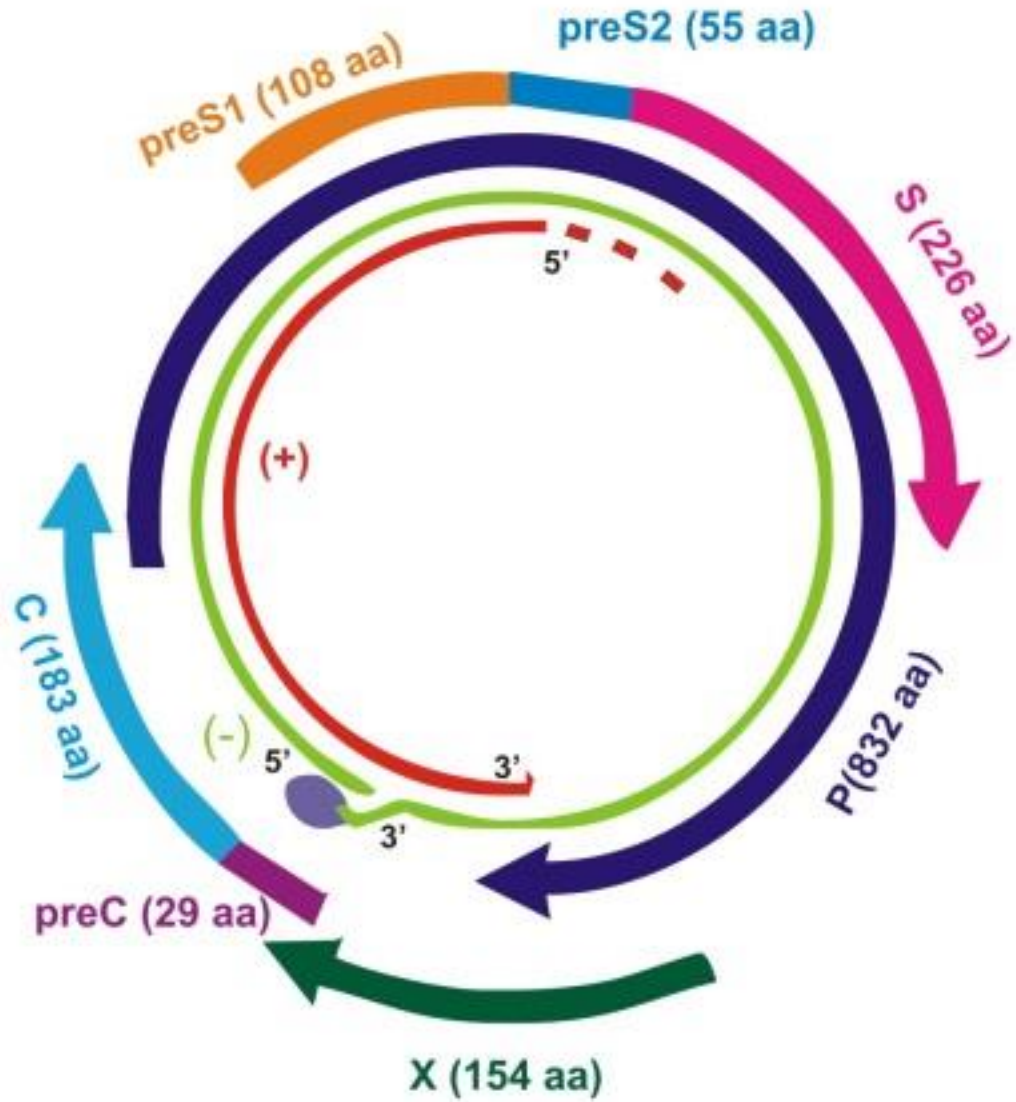
Research article Open Access Non-random retention of protein-coding overlapping genes in Metazoa Giulia Soldà†1, Mikita Suyama†2, Paride Pelucchi3, Silvia Boi1, Alessandro Guffanti3, Ermanno Rizzi3, Peer Bork4, Maria Luisa Tenchini1 and Francesca D Ciccarelli*5,6

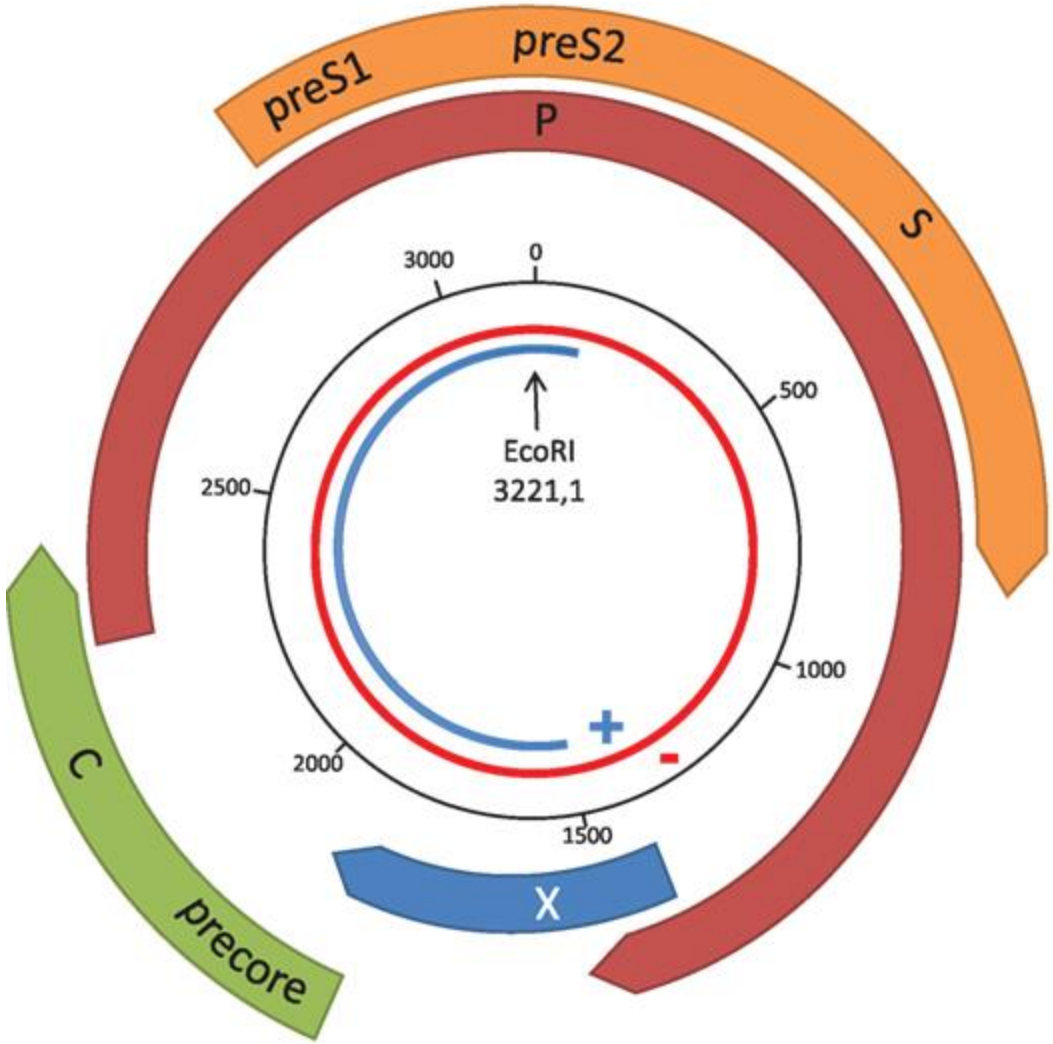
Overlapping genes HBV

- HBV is one of the smallest DNA viruses infecting humans, and its genome is a relaxed circular, partially doublestranded DNA of around 3200 bp. The genome contains four partially overlapping ORFs encoding the P (polymerase), C (core), S (surface) and X proteins, organized in order, resulting in about two-thirds of the viral genome encoding multiple proteins. From an evolutionary point of view, this genomic organization has a striking importance, as a synonymous nucleotide substitution in one ORF can potentially result in a non-synonymous nucleotide substitution in the overlapping ORF. In this way, it is believed that HBV genome evolution is constrained in order to maintain essential protein functions (Mizokami et al., 1997; Yang, 2007).

- Relative to other double-stranded DNA viruses capable of independent replication, HBV possesses the smallest genome of any virus known to infect man. Therefore, it is not surprising that HBV utilizes its genetic material economically. This is accomplished by two rare genetic arrangements: proteins are encoded from overlapping translation frames, and all regulatory signal sequences reside within protein-encoding sequences. Thus, HBV obtains multiple use from many regions of its genome.

Hepatitis B Virus genome organization





Gene S → HBs protein

- Outer envelope protein
- Non-infectious subviral HBsAg particles outnumber infectious virions by x100–10 000
- HBsAg promotes T cell exhaustion.
- *Immune/diagnostic escape mutants*

Gene P → Polymerase

4 domains, with 3 enzymatic activities:

- Terminal Protein (TP) domain: protein-priming function
- Non-conserved spacer domain (no enzymatic activity)
- Reverse Transcriptase domain: RNA-dependent DNA polymerase (RT) and DNA-dependent DNA polymerase
- RNase H domain: ribonuclease H activity
- *Drug resistance mutations*

Gene C → HBc/HBe proteins

- HBc antigen: icosahedral nucleocapsid
- HBe antigen: immunoregulatory roles
    - viral persistence, suppresses anti-viral T cell responses against HBcAg by stimulating T-reg cells
    - drive pro-inflammatory T cell polarization during viral clearance
    - may induce immunologic tolerance *in utero*
- *Pre-core mutants truncates precore/core protein*
- *Core promoter mutants – associated with fulminant hepatitis*

Gene X → HBx protein

- HBx regulates viral replication and host functions eg. transcription, cell cycle progression, DNA damage repair, apoptosis
- *Truncation mutants associated with oncogenesis*

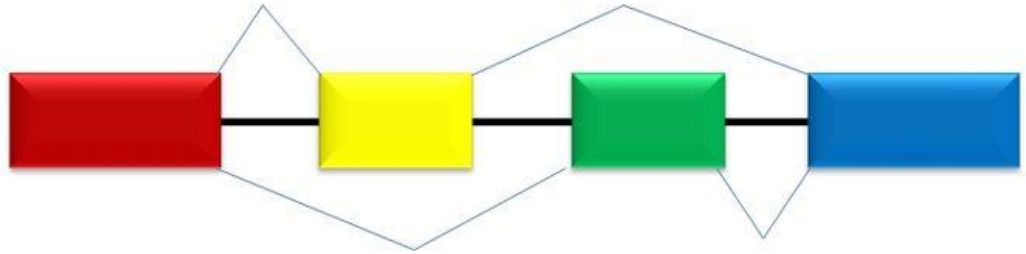
https://doi.org/10.3389/fimmu.2018.01561

Alternative splicing

Alternative splicing

Alternative splicing is a process that enables a messenger RNA (mRNA) to direct synthesis of different protein variants (isoforms) that may have different cellular functions or properties. It occurs by rearranging the pattern of intron and exon elements that are joined by splicing to alter the mRNA coding sequence.

Alternative splicing (AS) is a process by which exons can be either excluded or included in or from a pre-mRNA resulting in multiple mRNA isoforms.
From: Methods in Enzymology, 2013



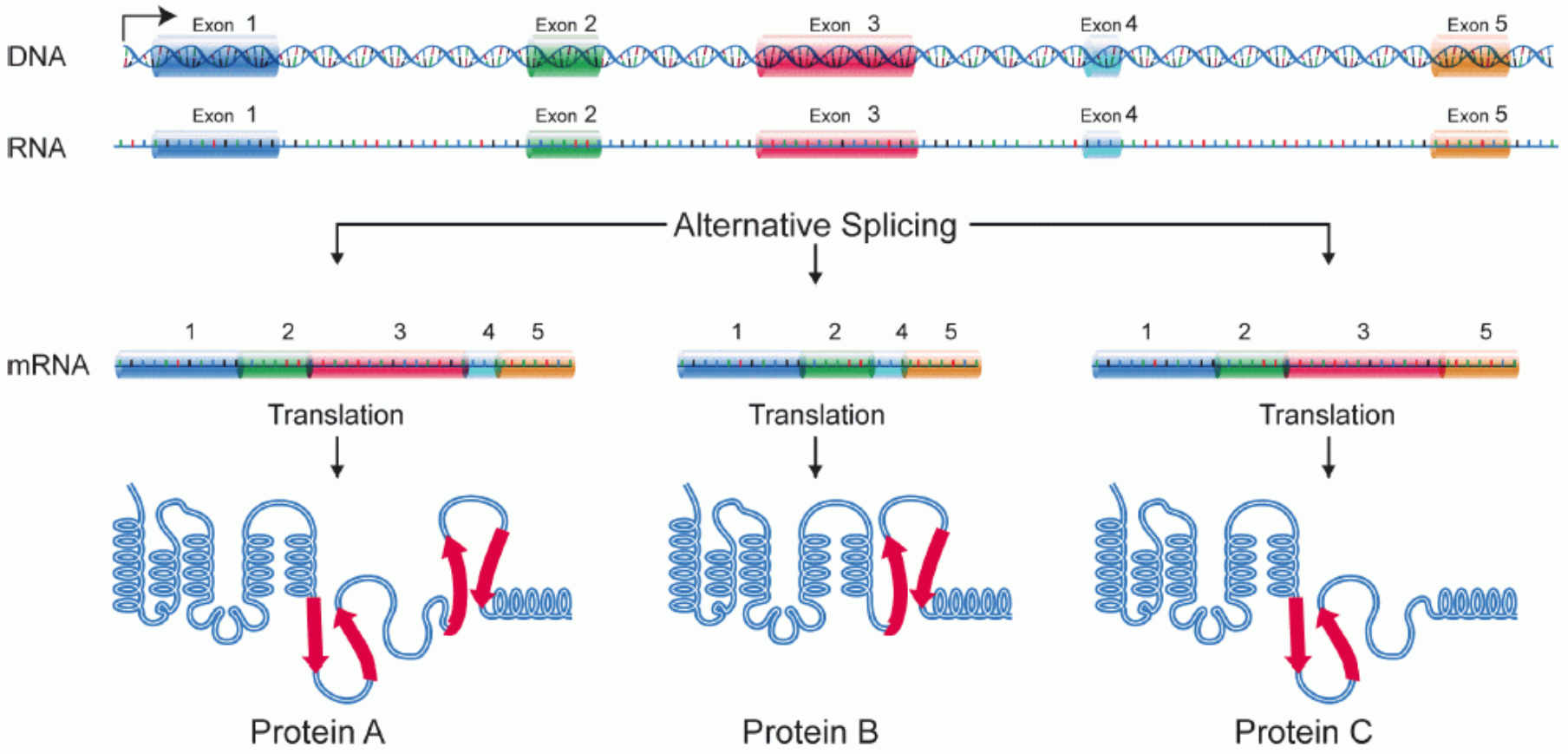
pre-mRNA

splicing

alternatively spliced mRNAs

translation

protein isoforms

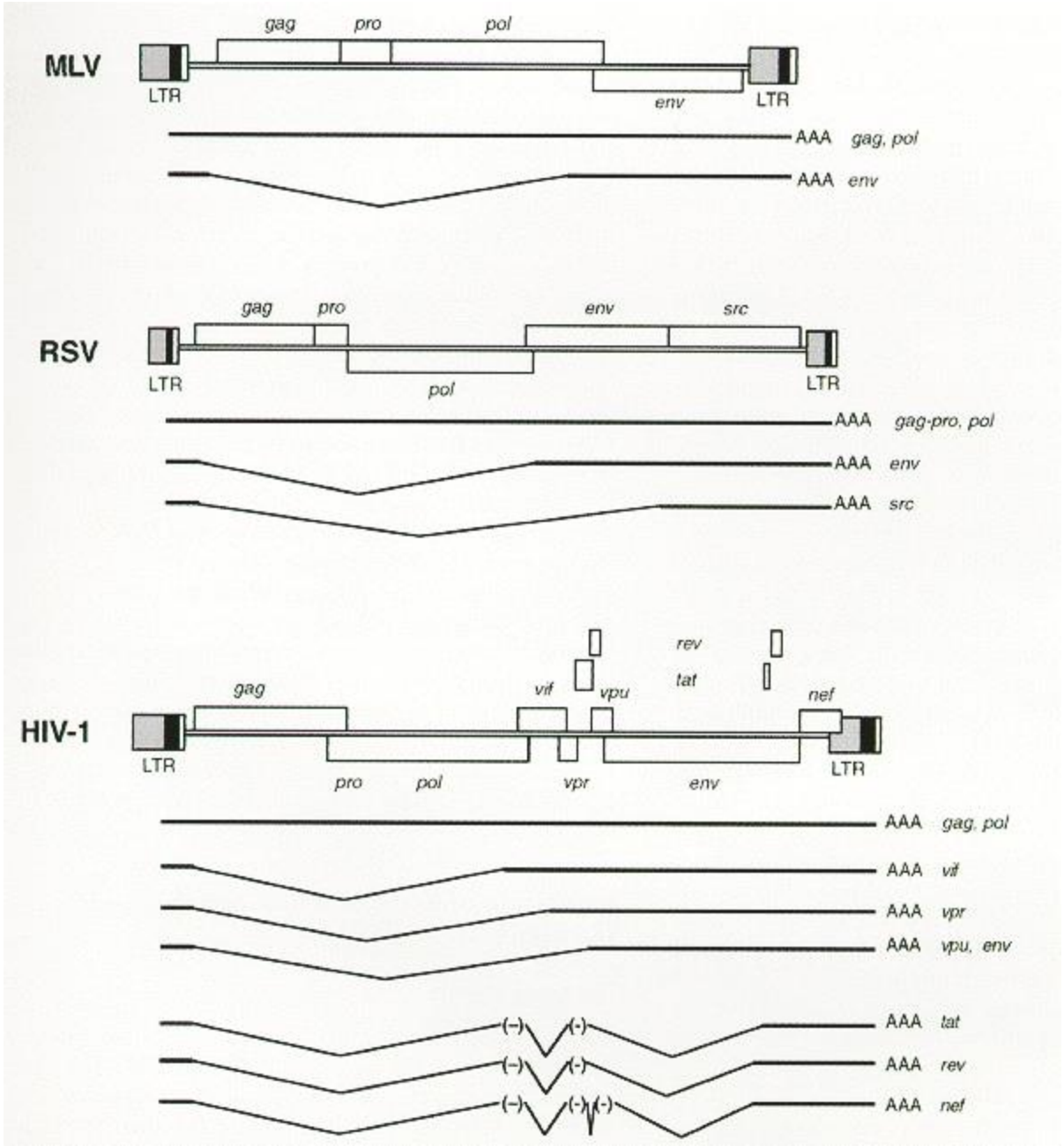


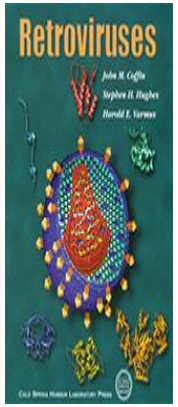
Alternative splicing produces three protein isoforms protein A, protein B and protein C.

Alternative splicing, or **alternative RNA splicing**, or **differential splicing**, is a regulated process during gene expression that results in a single gene coding for multiple proteins. In this process, particular exons of a gene may be included within or excluded from the final, processed messenger RNA (mRNA) produced from that gene.[1] Consequently, the proteins translated from alternatively spliced mRNAs will contain differences in their amino acid sequence and, often, in their biological functions (see Figure). Notably, alternative splicing allows the human genome to direct the synthesis of many more proteins than would be expected from its 20,000 protein-coding genes.

Control of mRNA splicing during retrovirus infection

During retrovirus infection of cells, splicing is also regulated in an interesting way. Now, retroviruses are RNA viruses that synthesize a DNA copy of their genome, which is then integrated into the host cell, and the cell then makes mRNAs for the virus by transcription. And one of the viral mRNAs produced is a pre-mRNA shown up here, which is unspliced. It's translation leads to the production of the Gag and Pol viral proteins, but the virus also needs to make a 3rd protein called Envelope (Env) which is encoded down here and to get at that coding region, the Gag and the Pol regions have to be removed by splicing. So the unspliced mRNA gives you Gag and Pol proteins. The spliced mRNA gives you the envelope protein.





Alternative splicing patterns in retroviruses.

Examples of different patterns of retroviral splicing include the single splicing event that generates the MLV *env* RNA, alternative splicing of RSV responsible for *env* and *src* RNAs, and the multiple alternative splicing events characteristic of complex retroviruses such as HIV-1. HIV-1 splicing complexity is increased by the alternative use of small, noncoding, central exons (denoted in parentheses).

From: Processing of Retroviral RNA

Retroviruses.
Coffin JM, Hughes SH, Varmus HE, editors.
Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 1997.

Terminal redundancy

A linear DNA molecule with the same sequence (genetic information) at each end.
If genetic information is represented by
ABCDEFGH
then a terminally redundant sequence can be for instance
ABCDEFGHAB.
Terminal redundancy is seen in some phages (eg T2) and is generated because a phage head is capable of containing a DNA molecule larger than the complete genome and packaging of DNA into phage heads is determined by the headful. These phages also show circular permutation.

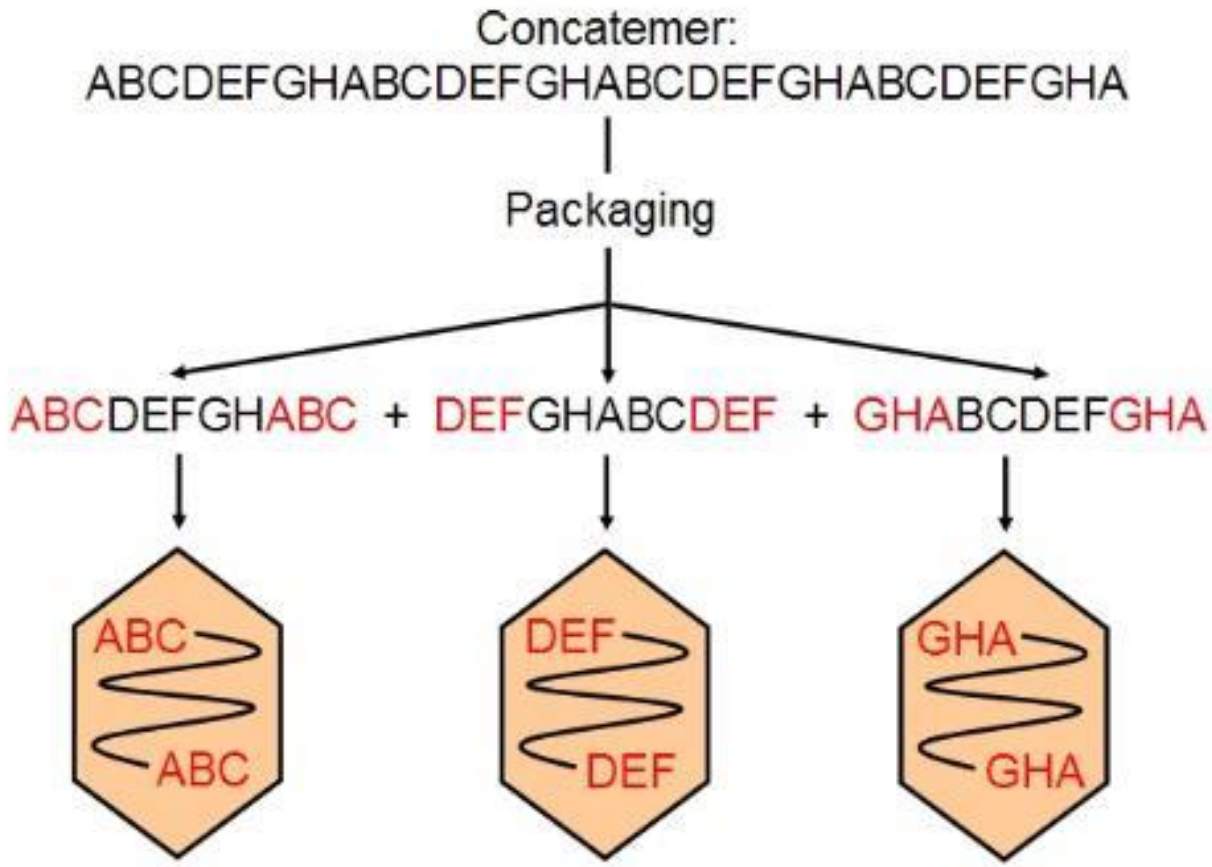
Terminally redundant DNA is DNA that contains repeated sequences at each end called terminal repeats. These ends are used (e.g. in virus T4) to join the ends of the linear DNA to form a cyclic DNA.[1] The term was first coined by Dr. Michael London in 1964.

Madigan M, Martinko J, eds. (2006). *Brock Biology of Microorganisms* (11th ed.). Prentice Hall. ISBN 0-13-144329-1.

Terminal redundancy: the repetition of sequences at both ends of a **DNA MOLECULE**.

Terminal redundancy is exhibited by **BACTERIOPHAGE** T4 **DNA** where the ends are **DIRECT REPEATS**.

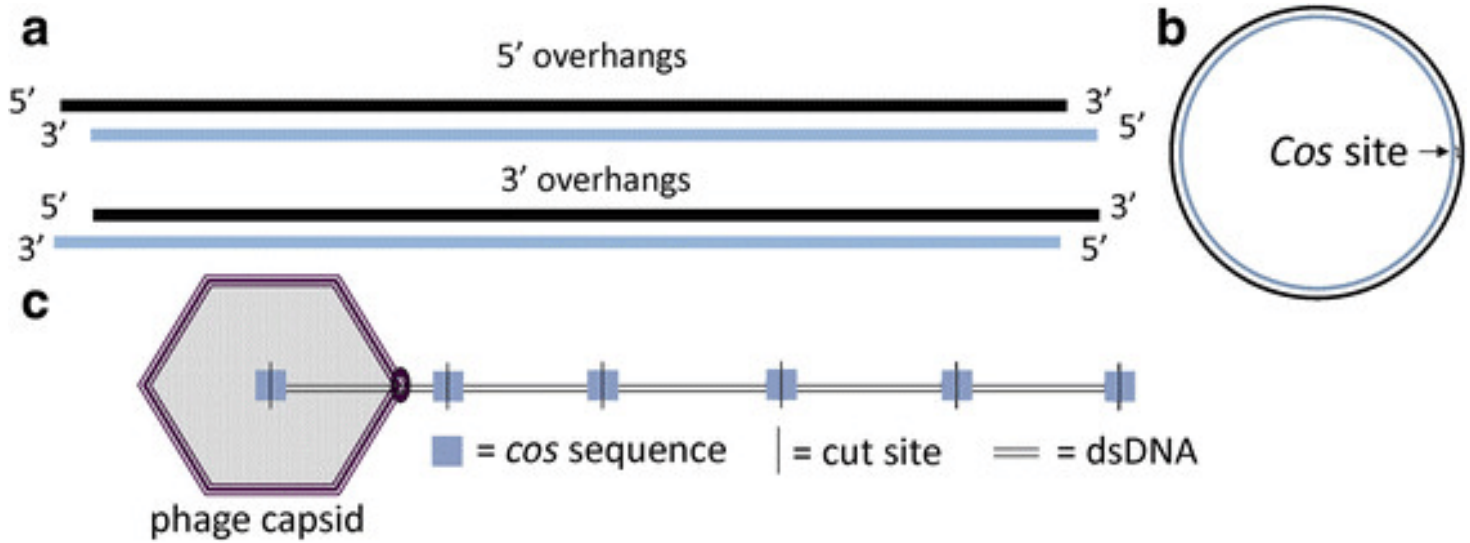
T4 phage has a circular genetic map, it's DNA is linear dsDNA and remains so in the host and replicates to produce linear dsDNA copies with '**sticky ends**' - single-stranded regions and concatemers are formed by enzymes joining together several single copies of the genome by zipping together their single-stranded sticky ends. These are then cut-up during packaging, as explained, with one phage head taking up slightly more than one complete genome's worth and then the concatemer is cut and packaging moved on to the next phage head. This creates linear dsDNA in which the two ends repeat and as there is some spare DNA this is called **terminal redundancy**.



Packaging of DNA into the head of T4 phage

Cohesive ends (lamda phage)

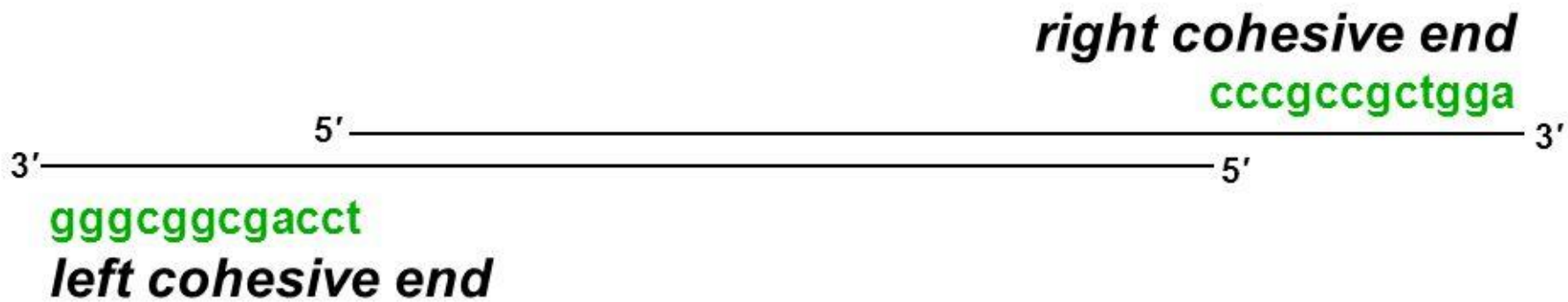
- Lambda is a medium size **E.coli bacteriophage**. The DNA molecule of **48502 basepairs** is linear and except for the extreme ends double-stranded. At each end the 5' strand overhangs the 3' strand by 12 bases. The sequences of the ends are complementary. At ambient temperatures, in a solution containing purified Lambda-DNA these so-called 'cos ends' may pair and form the so-called '**cos-site**'. As a consequence, the DNA is (partly) circularised or have formed concatemers.



Genomic organisation of lambda

- genome = 49 kb **linear double-stranded** DNA

- DNA present in phage particles as linear double-stranded molecule with **single-stranded complementary termini** of 12 nt ("cohesive ends")



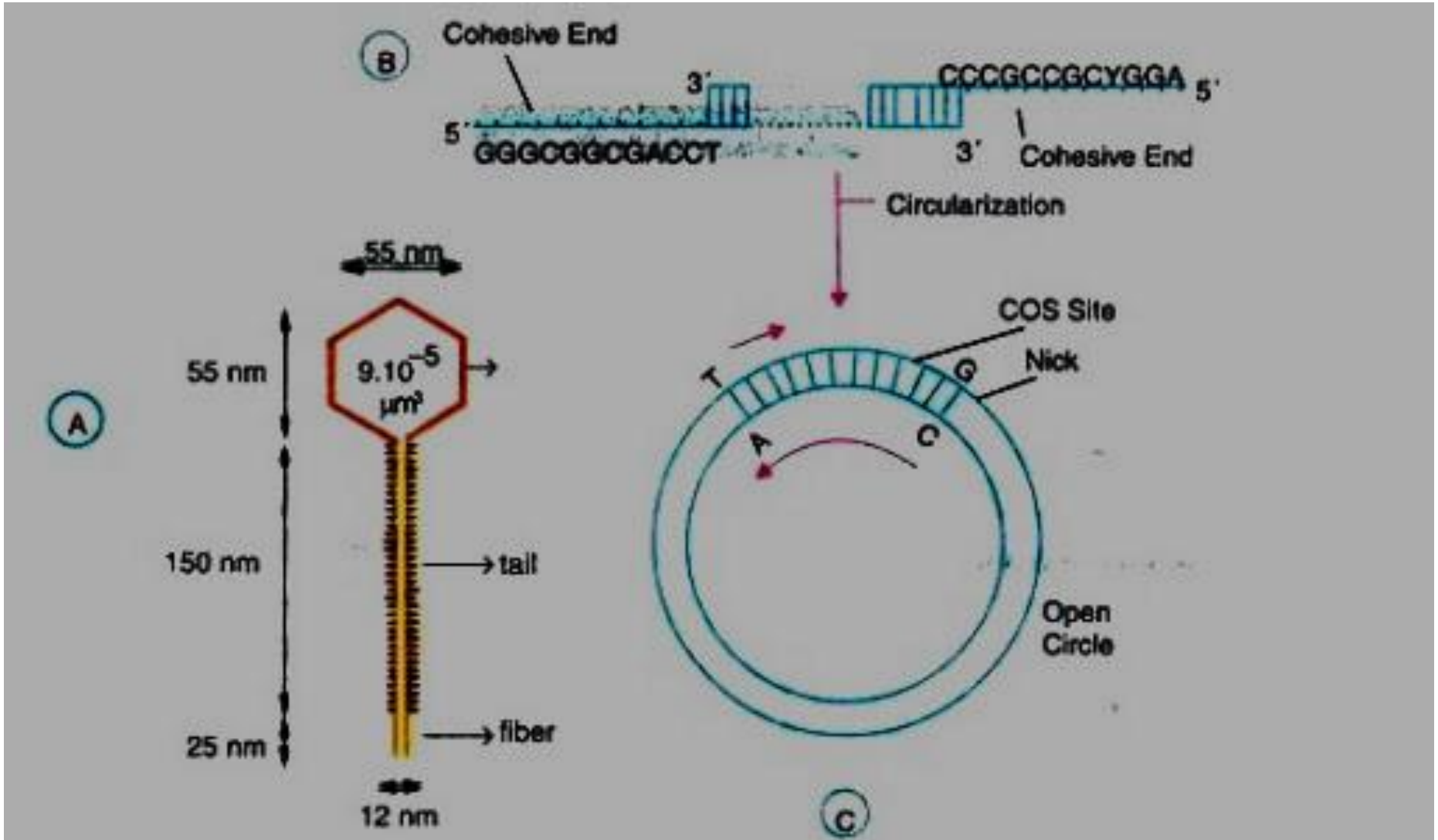


Fig. 18.11 : Structure of bacteriophage lambda (λ). (A), single standard cohesive ends of 12 bases (B), and circularization of complementary DNA to form *cos* site (C).

Ambisense genome

- **Ambisense** is a situation in which both the **genome** and its complement contain some coding information. Remember, translation always occurs in the 5' to 3' direction, so the two strands are being translated in opposite directions. Strictly speaking, each strand has regions of + and - polarity, hence **ambisense**.

Single-stranded RNA viruses and RNA Sense

- RNA viruses can be further classified according to the sense or polarity of their RNA into negative-sense and positive-sense, or ambisense RNA viruses.
- Positive-sense viral RNA is similar to mRNA and thus can be immediately translated by the host cell.
- Negative-sense viral RNA is complementary to mRNA and thus must be converted to positive-sense RNA by an RNA-dependent RNA polymerase before translation.
- As such, purified RNA of a positive-sense virus can directly cause infection though it may be less infectious than the whole virus particle.
- Purified RNA of a negative-sense virus is not infectious by itself as it needs to be transcribed into positive-sense RNA; each virion can be transcribed to several positive-sense RNAs.
- **Ambisense RNA viruses resemble negative-sense RNA viruses, except they also translate genes from the negative strand.**[5]

Ambisense transcription in negative stranded RNA viruses

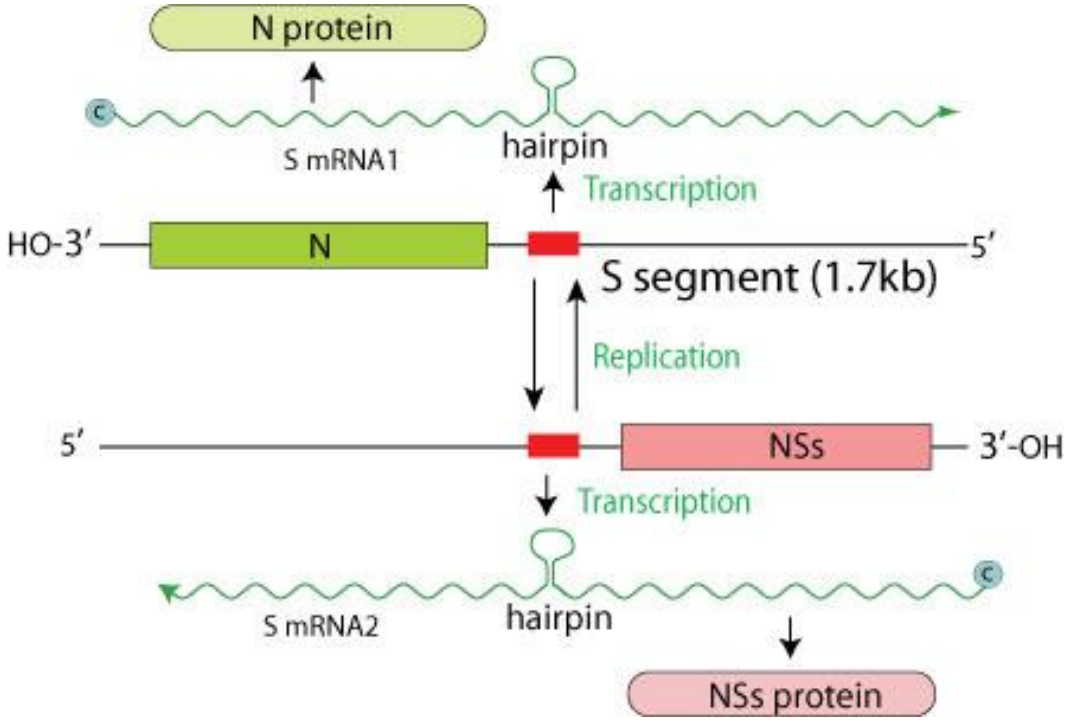
- An ambisens genome is a genome which both nucleic acid strands encode for proteins

.

- This expression strategy is found in four genera of segmented negative stranded RNA viruses: Arenavirus, Phlebovirus, Tospovirus, and Tenuivirus.

- It depends on solid transcription termination signals to avoid creating dsRNA, a pattern recognized by all cells to trigger antiviral response.

- All the ambisens negative standed RNA viruses indeed encode a hairpin to stop transcription. Non-segmented negative stranded RNA viruses, and orthomyxoviridae have leaky termination signal, preventing them to use this transcription strategy



Ambisense RNA viruses

- Among the negative RNA viruses, ambisense RNA viruses or ‘ambisense viruses’ occupy a distinct niche. Ambisense viruses contain at least one ambisense RNA segment, i.e. an RNA that is in part of positive and in part of negative polarity. Because of this unique gene organization, one might expect ambisense RNA viruses to borrow expression strategies from both positive and negative RNA viruses. However, they have little in common with positive RNA viruses, but possess many features of negative RNA viruses.

Expression strategies of ambisense viruses, Author links open overlay panelMarieNguyenAnne-LiseHaenni

Partial double stranded genome: Hepatitis B

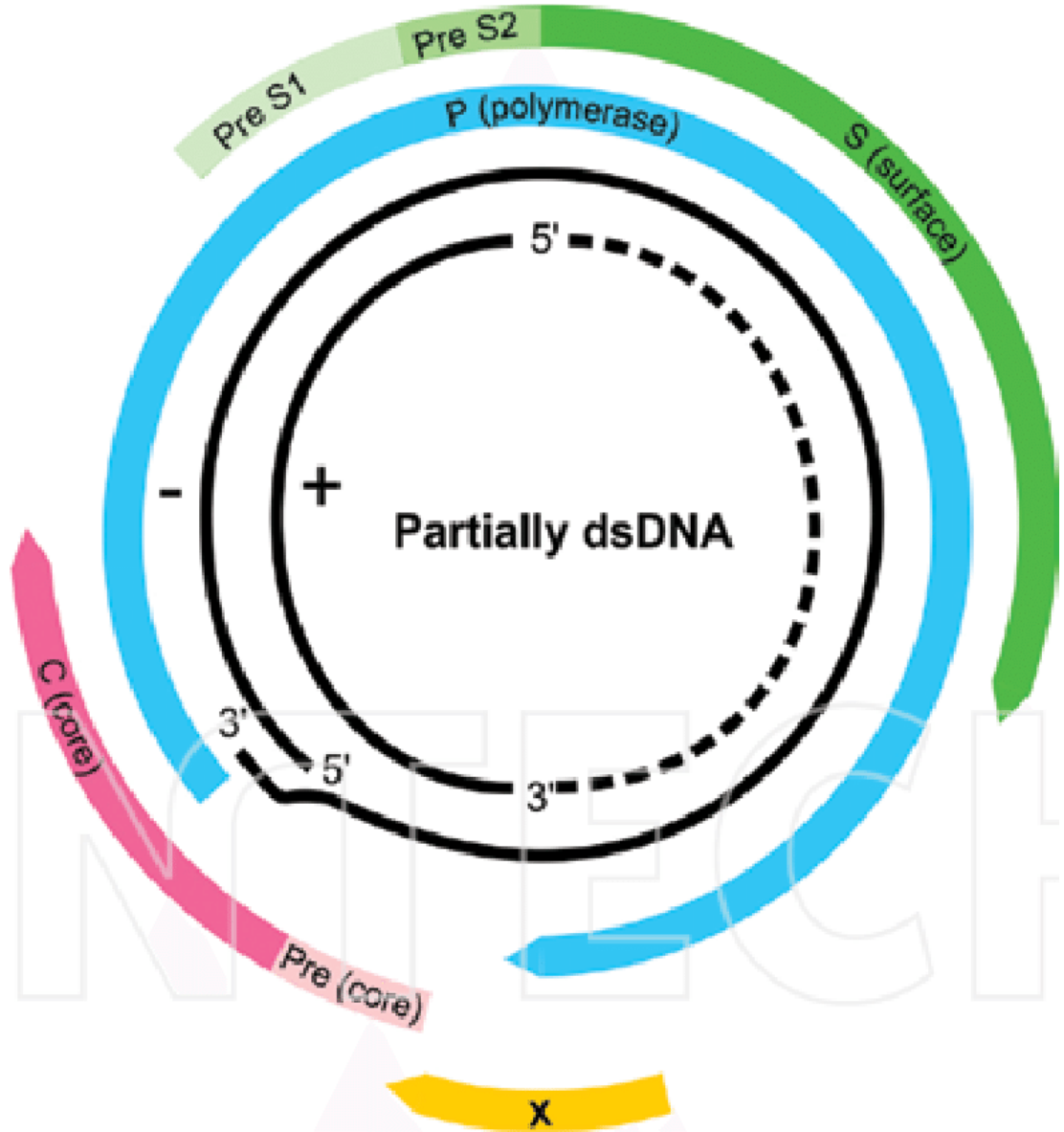
Hepadnaviridae[a] is a family of viruses. Humans, apes, and birds serve as natural hosts. There are currently over 12 species in this family, divided among 5 genera. Its best-known member is hepatitis B virus.

The genome organisation of HBV; the genes overlap.

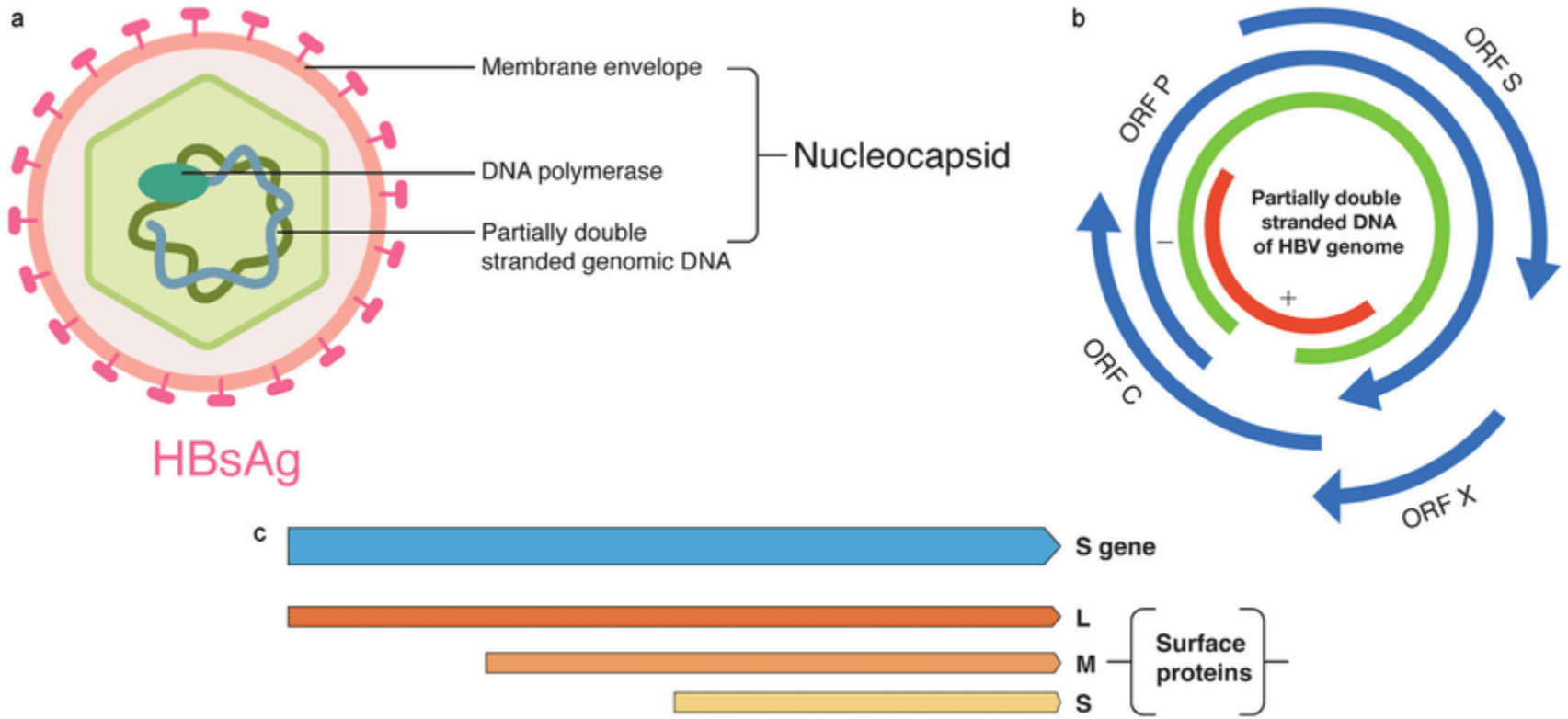
Hepadnaviruses have very small genomes of partially double-stranded, partially single stranded circular DNA (pdsDNA). The genome consists of two strands, a longer negative-sense strand and a shorter and positive-sense strand of variable length. In the virion these strands are arranged such that the two ends of the long strand meet but are not covalently bonded together. The shorter strand overlaps this divide and is connected to the longer strand on either side of the split through a direct repeat (DR) segment that pairs the two strands together. In replication, the viral pdsDNA is converted in the host cell nucleus to covalently-closed-circular DNA (cccDNA) by the viral polymerase.

As it is a group 7 virus, replication involves an RNA intermediate. Four main open reading frames are encoded (ORFs) and the virus has four known genes which encode seven proteins: the core capsid protein, the viral polymerase, surface antigens—preS1, preS2, and S, the X protein and HBeAg. The X protein is thought to be non-structural. Its function and significance are poorly understood but it is suspected to be associated with host gene expression modulation.

The genome organisation of HBV; the genes overlap.



The genome organisation of HBV; the genes overlap.



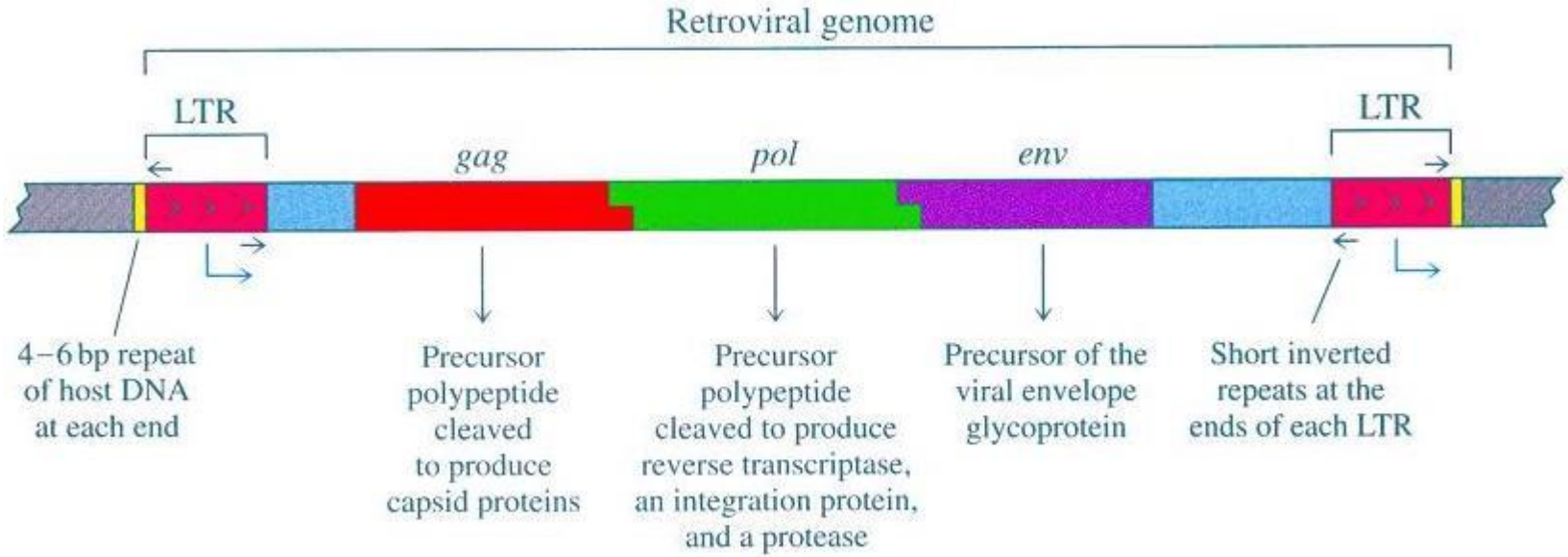
Structural and genetic organization of HBV.
a, structure of HBV;
b, compact organization of the HBV genome showing multiple overlapping open reading frames;
c, expression of three envelope proteins from three in-frame start codons from the S-gene of the HBV genome.

Long terminal repeats (LTRs)

- In eukaryotes, there is a large class of RNA viruses known as *retroviruses*. They have an obligatory stage where the RNA is reverse transcribed into DNA and the DNA is inserted into the genome where it resides as a provirus.

- **Long terminal repeats** (**LTRs**) are identical sequences of DNA that repeat hundreds or thousands of times found at either end of retrotransposons or proviral DNA formed by reverse transcription of retroviral RNA.[1] They are used by viruses to insert their genetic material into the host genomes.[1]

- The long terminal repeat (LTR) is the control center for gene expression. As may be expected because of the integrated phase of their life cycle, retroviruses have somewhat typical eucaryotic promoters with transcriptional enhancers and some also have regulatory elements responsive to either viral or specialized cellular (e.g. hormonal) trans-activating factors (HIV, MMTV). Enhancer functions have also been mapped to the gag (ALSV) and gag-pol (SIV and HIV) regions of some viruses but their role in the virus life cycle has yet to be established.



The structure of the integrated retrovirus genome is shown above. The ends of the viral genome contain long terminal repeat (LTR) sequences of several hundred base pairs. Both LTRs are arranged in the same orientation and the outside ends of each are flanked by short inverted repeats.