VIROLOGY 2

SALIENT FEATURES OF VIRAL GENOMES

Dr. Arpita Mandal

SEM 3, CC5, Unit 3,

TMV unusual bases

- Many messenger and viral RNAs contain a substantial region of restricted base composition, such as the **poly(A) at the 3'** end of many viral and most cellular mRNAs and the **poly(C) tract** found in many picornaviruses.
- The 5' end of tobacco mosaic virus (TMV) genomic RNA is **capped with 7- methylguanosine**.
- Tobacco mosaic virus (TMV) RNA bears no such homopolymer tract but it is known to contain a sequence of about **70 nucleotides devoid of guanosine residues**. This guanosine-free (G-free) segment can be isolated from a total T, ribonuclease digest of TMV RNA as part of very long T, oligonucleotide, which has been **designated Ω**.
- This **Ω** is situated immediately **after the capped guanosine residue at the 5' extremity of the RNA** molecule, the 5' terminal sequence being **$m^7G^5ppp^5G$ Ω-**.
- It is evident that the 5'-terminal G-free segment cannot form part of the coding portion of TMV RNA since **it is not preceded by an A-U-G or G-U-G initiation codon**. The first potential translation-initiation codon in the RNA molecule comes at the end of the G-free tract since **Ω** is known to have the sequence **C-A-A-U-G at its 3' extremity**.
- The 3'-terminal **A-U-G of Ω** is in fact the **starting point for viral protein synthesis**.
- The oligonucleotide **Ω** has several unusual features which may be binding sites **for proteins involved in replication, transport or regulatory processes.**
- The run of **five U near the 5' terminus** seem to be **associated with termination of transcription** in bacterial messengers.

T4 bacteriophage unusual bases

- **The T4 bacteriophage** is a complex and highly evolved virus which infects the bacterium *Escherichia coli*. The T4 phage is quite large for a virus: the head (capsid) is 119.5 nm tall and 86 nm in diameter, whilst the tail is 100 nm long and 21 nm in diameter.
- The virion is composed of over 2000 protein subunits, which are the products of over 50 different genes.
- Viruses have genome with certain unusual bases. Replacement of thymine by uracil in DNA bacteriophages (T4).
- Thymine is also replaced by hydroxy methyl uracil.
- In T4 phages, **cytosine is replaced by 5-hydroxy methyl cytosine** to **protect phage DNA** from **host restriction** and **hydroxy methyl** group is **Glucosylated** to protect phage DNA from **host Mcr (modified cytosine restriction**.
- 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.
- The glucose molecules added to the cysteine residues also increases the stability of the DNA since the - OH and -H groups of the glucose can hydrogen-bond to neighbouring bases.
- This may be especially important as the genome **is low in G+C base pairs at 34.5% G+C**.
- In some regions both strands are transcribed and both may be translated into proteins.
- The T4 dsDNA approximates D-form DNA (poly(dA-dT)) which is overwound with only 8 bp per turn and a wider and shallower major groove and a deeper and narrower minor groove.
- This form of DNA is possibly transcribed and replicated faster as it may unzip more easily.

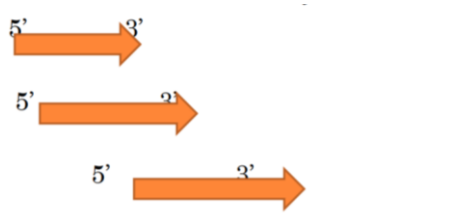
Definition of overlapping genes

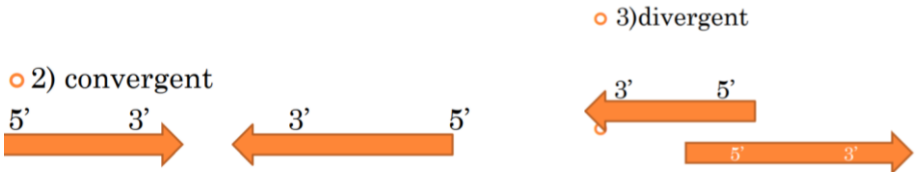
Overlapping genes are defined as a pair of adjacent genes whose coding regions are partially overlapping. In other words, a single stretch of DNA codes for portions of two separate proteins. Such an arrangement of genetic code is ubiquitous. Many overlapping genes have been identified in the genomes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

ORFs in different reading frames may overlap each other.

These are classified into 3 categories:

1) Unidirectional It is found most commonly

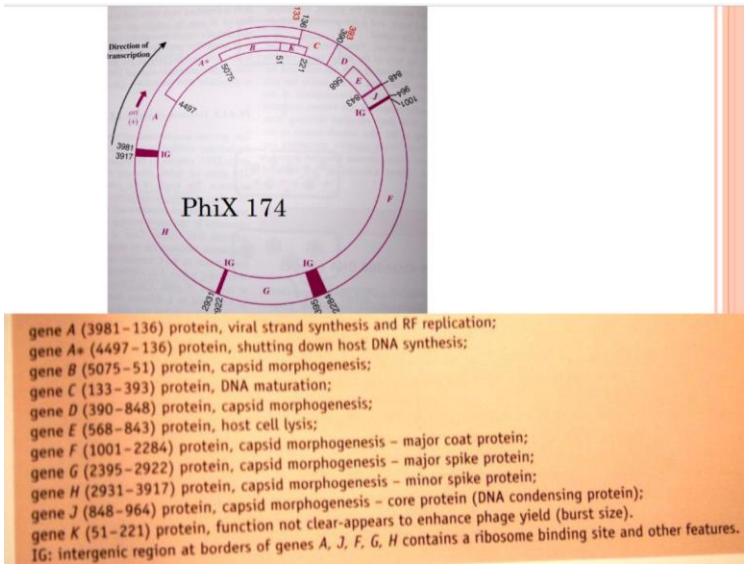


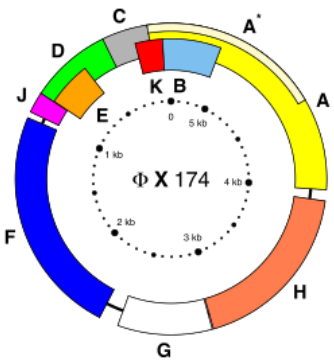


Example : ø×174

The **phi X 174** (or **ΦX174**) bacteriophage is a single-stranded DNA (ssDNA) virus that infects *Escherichia coli*, and the first DNA-based genome to be sequenced.

Figure 1.





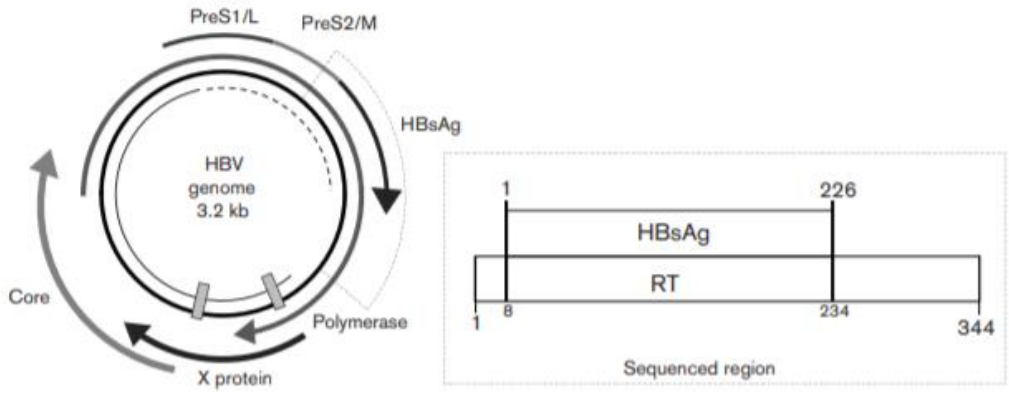
ΦX174 encodes **11 genes**, named as consecutive letters of the alphabet in the order they were discovered, with the exception of **A* which is an alternative start codon** within the **large A genes**.

Hepatitis B virus

HBV is one of the smallest DNA viruses infecting humans, and its genome is a relaxed circular, partially double stranded DNA of around 3200 bp. The genome contains **four partially overlapping open reading frames (ORF- part of reading frame which have the ability to be translated) 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.**

Figure 2. A and 2.B



Schematic representation of the HBV genome and its ORFs.

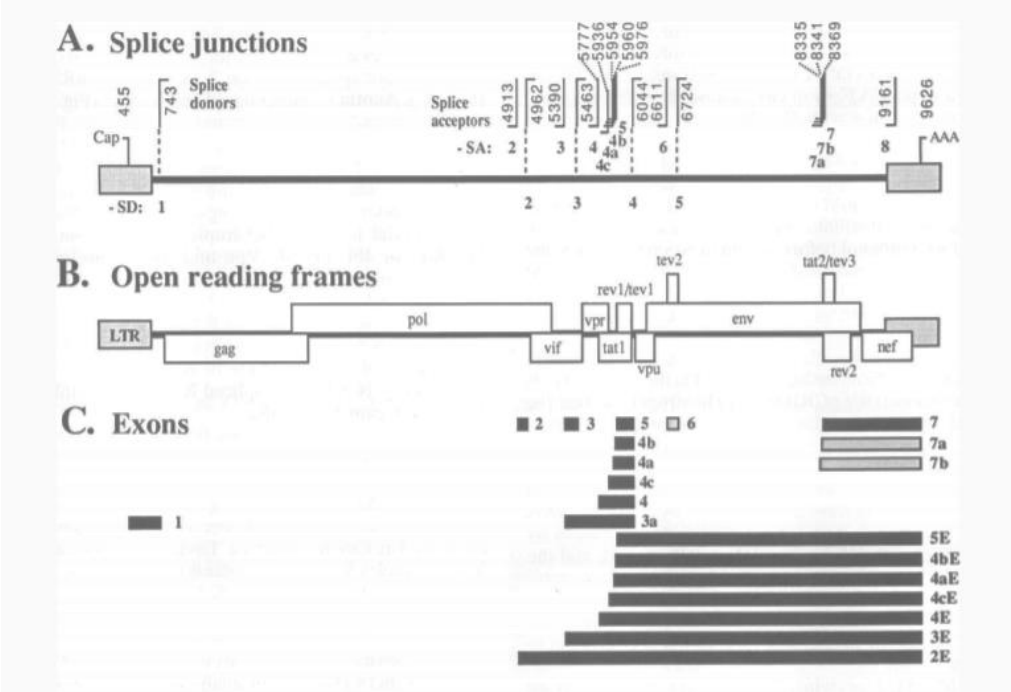
- The HBV genome comprises a 3.2 kb circular positive strand complementary to a negative shorter incomplete strand (1700–2800 nt).
- The viral genome encodes four overlapping open reading frames (ORFs: S, C, P, and X) (Fig. 2A).
- The **S ORF encodes** the viral **surface envelope** proteins, the HBsAg, and can be structurally and functionally divided into **the pre-S1, pre-S2, and S regions**.
- The core or **C** gene **has the pre core and core regions**.
- Multiple in-frame translation initiation codons are a feature of the S and C genes, which give rise to related but functionally distinct proteins.
- The C ORF encodes either the viral nucleocapsid **HBcAg** or hepatitis B e antigen (**HBeAg**) depending on whether translation is initiated from the core or pre core regions, respectively (Fig. 2B).
- The **polymerase (pol**) is a large protein (about 800 amino acids) encoded by the **P ORF** and is functionally divided into **three domains**:
  - the **terminal protein domain**, which is involved in **en-capsidation** and **initiation of minus-strand synthesis**;
  - **the reverse transcriptase (RT) domain**, which **catalyzes genome synthesis**;
  - and the **ribonuclease H domain**, which **degrades pre genomic RNA and facilitates replication**.
- The **HBV X ORF** encodes a 16.5-kd protein (**HBxAg**) with multiple functions, **including signal transduction, transcriptional activation, DNA repair, and inhibition of protein degradation**.

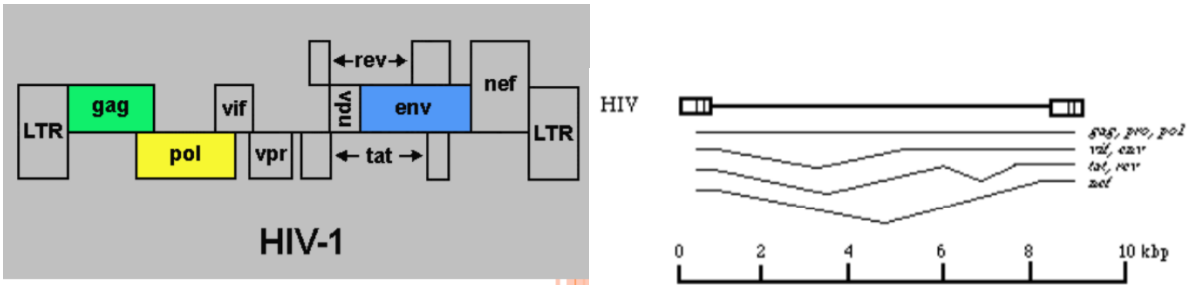
Alternative Splicing of Human Immunodeficiency Virus

Alternative splicing (or differential splicing) is a process by which the coding regions of the RNA produced by transcription of a gene (a **primary gene transcript or pre-mRNA**) are **reconnected in multiple ways during RNA splicing**. The resulting different mRNAs may be translated into different protein; thus, **a single gene may code for multiple proteins.**

Alternative splicing occurs as a **normal phenomenon in eukaryotes**, where it greatly increases the diversity of proteins that can be encoded by the genome; in humans, ~95% of multiexonic genes are alternatively spliced.

- The most common **method** is **exon skipping**. In this mode, a particular exon may be included in mRNAs under some conditions or in particular tissues, and omitted from the mRNA in others.
- **Multiple RNA splicing sites** exist within human immunodeficiency virus type 1 (**HIV-1) genomic RNA**, and these sites enable the synthesis of many mRNAs for each of several viral proteins.
- **The biological significance of the alternatively spliced mRNA** species during productive HIV-1 infections of peripheral blood lymphocytes and human T-cell lines to determine the potential role of alternative RNA splicing in the regulation of HIV-1 replication and infection.
- The redundant RNA splicing signals in the HIV-1 genome and alternatively spliced mRNAs provides a mechanism for regulating the relative proportions of HIV-1 proteins and, in some cases, viral infectivity.
- All retroviruses require RNA splicing to remove upstream gag and pol coding sequences from the env mRNA. In addition, HIV-1 generates a distinctly complex pattern of spliced RNA to encode the essential regulatory proteins, Tat and Rev, as well as several other proteins (Vif, Vpr, and Nef) needed for successful replication in vivo.
- The HIV-1 Rev protein binds viral RNA species that contain the Rev-responsive element (RRE), located in the env gene, thereby promoting the export, and possibly the stability and translation, of partially spliced and unspliced RNAs from the nucleus into the cytoplasm for its translation and/or packaging into progeny virions.
- The Rev-RRE system alleviates the paradoxical requirement for both spliced and unspliced HIV-1 RNA for successful virus replication.
- Rev protein also regulates the temporal change from multiply spliced HIV-1 RNAs to partially spliced or unspliced RNAs during productive virus infection (27, 29).
- The splicing of HIV-1 RNA is extremely complex because of the presence of both constitutive and alternatively used 5' RNA splice donor (SD) and 3' splice acceptor (SA) motifs. Numerous weak SA motifs, located toward the center of the genomic RNA, are competing points of ligation for splicing, and their alternate selection usually determines the protein encoded by the mature RNA.
- However, some of these mRNAs are multi-cistronic, encoding more than one protein.
- Increased diversity of spliced mRNAs for several HIV-1 proteins results from the alternative cassetting of two noncoding exons into a proportion of transcripts.
- In addition, the use of several cryptic SA and SD sites may lead to the synthesis of novel chimeric viral proteins.
- The varied use of these diverse splicing signals results in the synthesis of several sets of structurally different RNAs that serve as alternative templates for the translation of the same protein, including the viral envelope, regulatory, and accessory proteins.
- Because this complex pattern of RNA expression is maintained among many HIV-1 isolates of diverse origins, it is likely that this complexity is critical for the successful completion of the HIV-1 infectious cycle and not simply an inherent redundancy in viral RNA processing.





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 (e.g. T2, T4**) 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 headfull. These phages also show circular permutation. In T4 about 500 base pair are terminally redundant. T4 phage is linear and show circular permutation- starting point. The genome differs for various members of a particular viral population.

- Results from rolling circle replication
- Packaging is seq. independent and occur by head full mechanism
- Terminal redundancy also seen in HERPES Virus.

For further reading please see figure a-d

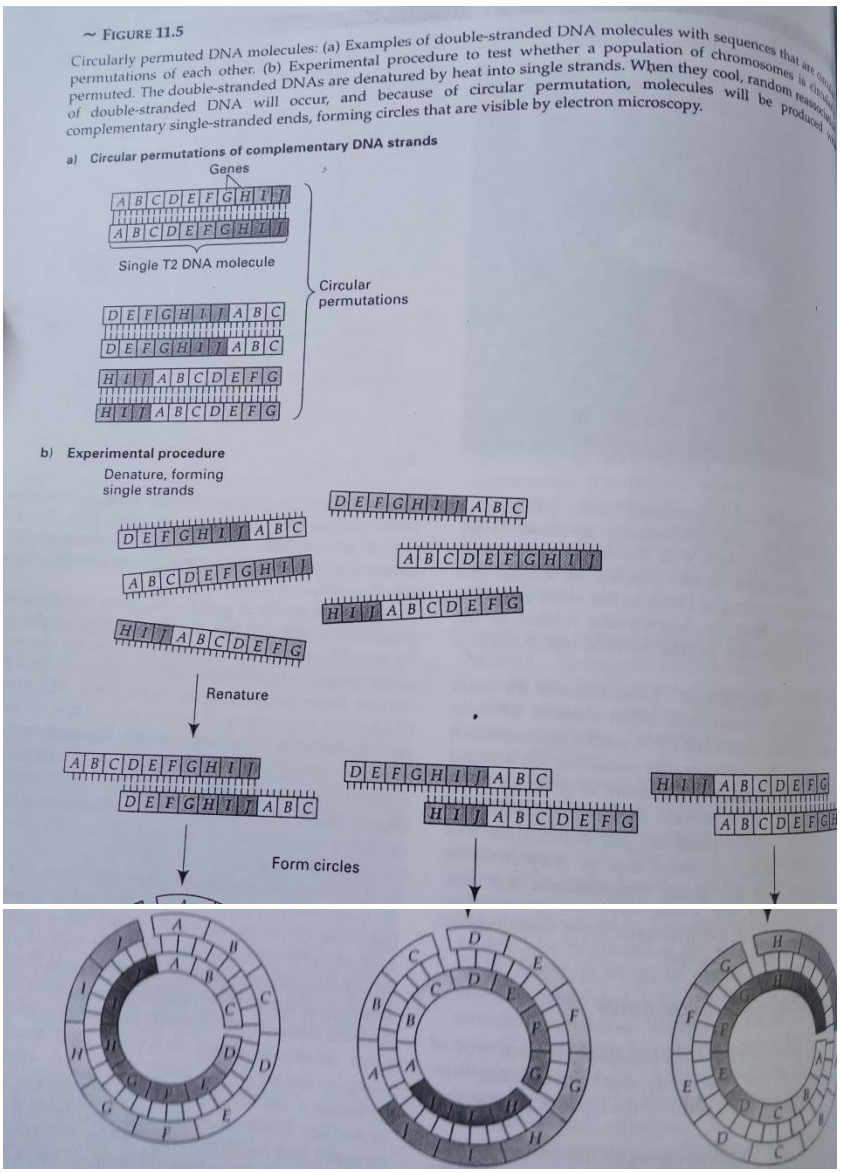

~ FIGURE 11.5
Circularly permuted DNA molecules: (a) Examples of double-stranded DNA molecules with sequences that are circular permutations of each other. (b) Experimental procedure to test whether a population of chromosomes is circularly permuted. The double-stranded DNAs are denatured by heat into single strands. When they cool, random reassociation of double-stranded DNA will occur, and because of circular permutation, molecules will be produced with complementary single-stranded ends, forming circles that are visible by electron microscopy.
a) Circular permutations of complementary DNA strands
Genes
Single T2 DNA molecule
Circular permutations
b) Experimental procedure
Denature, forming single strands
Renature
Form circles

Figure a

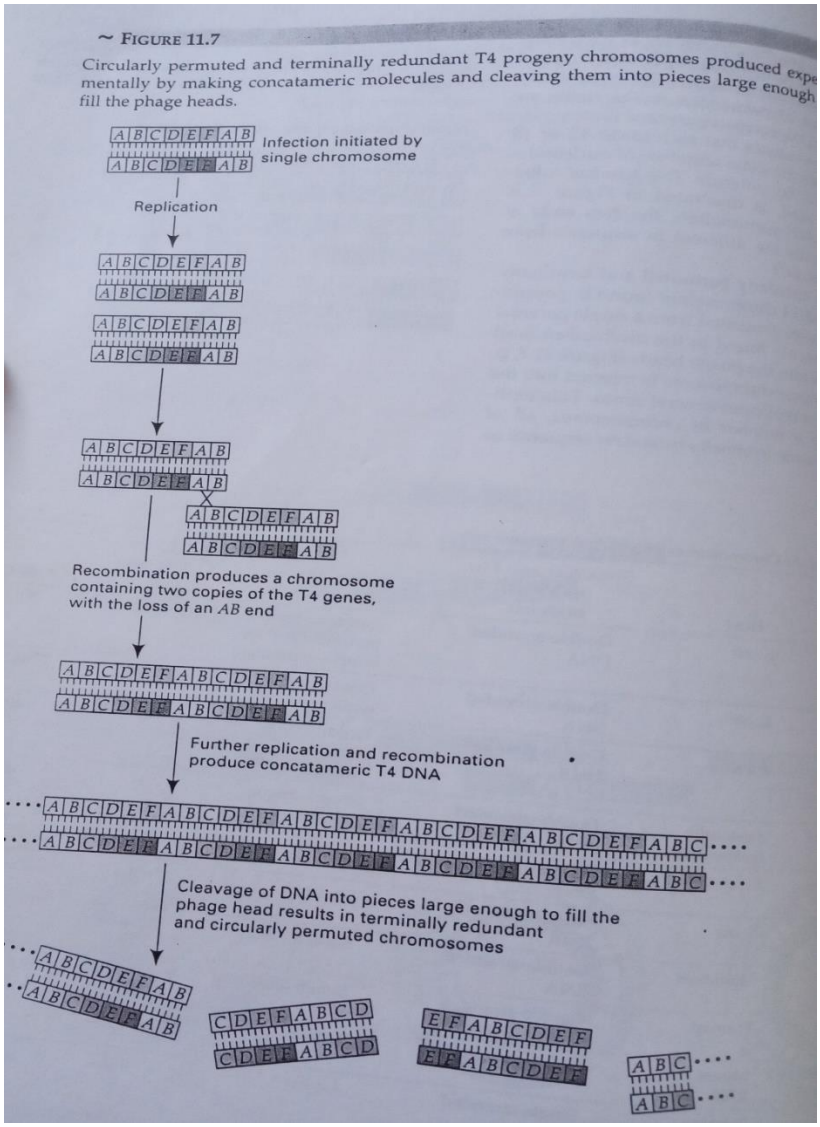


Figure b

There is also evidence that each linear T2 or T4 DNA molecule has the same sequence of nucleotides at the two ends of the molecule. This *terminal redundancy*, as it is called, is illustrated in Figure 11.6. (Because of circular permutation, the two ends of each DNA molecule are different in sequence from molecule to molecule.)

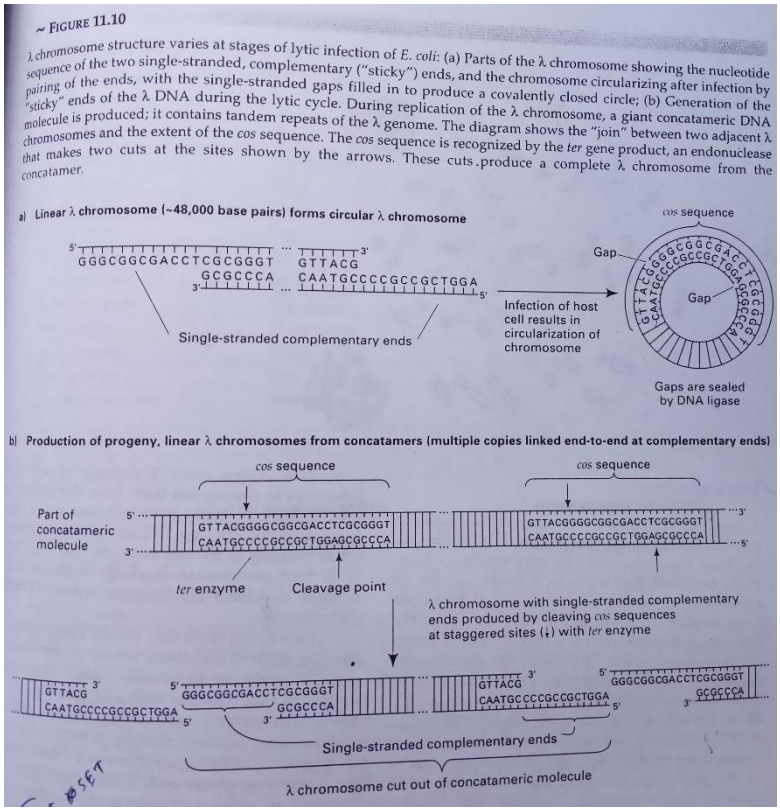
How are the circularly permuted and terminally redundant T2 and T4 chromosomes found in populations of those phages generated from a single parental phage? The answer is found in the mechanism used to package DNA into the phage heads (Figure 11.7, p. 320). After the phage chromosome is injected into the host bacterium, it replicates several times. This replication produces a number of chromosomes, all of which have the same *terminally redundant* sequence as

Figure c

the parental DNA. Molecular recombination occurs between the DNA molecules at the terminally redundant ends, splicing the chromosomes together into very long molecules, called *concatamers*. These structures repeat the base sequence of the original unit phage chromosome in a tandem fashion. The concatamers undergo several rounds of replication, and late in infection the DNA is packaged into the phage heads that have been produced concurrently. The packaging is done by the *headful*; the length of DNA that is put into each head is determined by the volume of the head. Since the head can hold a little more than a genome's worth of DNA, we see how each phage chromosome contains a terminally redundant region. In addition, the successive clipping of the concatamer molecules into headful lengths leads to the circular permutation of the population of progeny phage chromosomes. In other words, each phage contains the same amount of DNA, with the same sequences represented, and with terminal redundancy. In a population of phage chromosomes, individual chromosomes will have different terminal sequences as a result of different permutations of the same sequence.

Figure d

Terminal cohesive ends



- The phage chromosome is a linear double-stranded DNA molecule ca. 49 kilobase pairs (kbp) in length.
- At the 5' ends are single-stranded segments, 12 bases long, that are complementary; they are called cohesive ends.
- Rolling circle synthesis generate a concatemeric genome cleaved during packaging into molecules of the correct length possessing the same cohesive ends as the linear genome which initiated infection.
- The presence of the cohesive ends results in cyclization of the DNA molecule after injection. The resulting nicks are then sealed by host ligase.
- At early times after infection, replication produces progeny rings that are not a substrate for packaging; at later times, rolling circle replication produces the linear multi-chromosomal lengths of DNA, called **concatemers**, that are the substrate for the packaging process.
- DNA packaging for phage involves specific recognition of phage DNA from a pool that also includes bacterial DNA.

Partially double stranded genomes complete negative strand and incomplete positive strand hepatitis b

♣ Genome is partially dsDNA that forms a covalently closed circle with 5' end of the full length minus strand which is linked to viral DNA polymerase.

♣ The genome sequence has termini with cohesive ends that matched the uniquely located 5' ends of the two strands which overlaps by ~240 nucleotide.

♣ Complementary to the viral mRNA, the –ve strand or noncoding strand is full length; the viral +ve sense strand is shorter than full length.

WHY THE (+) DNA IN THE VIRION IS INCOMPLETE?

- During (+)DNA synthesis a nucleocapsid can either migrate to the nucleus to increase the pool of DNA
- It can undergo a maturation event that enables it to bud through a membrane containing virus envelope proteins.
- DNA synthesis ceases on budding, as the nucleocapsid is cut off from the pool of nucleotides in the cytoplasm.

Example- HPV Nucleic Acid

The HBV DNA is a relaxed, circular, partial double strand, approximately 3.2 kb long.

The two strands are asymmetric, a feature exclusive to the Hepadnaviruses (Figure 2).

The minus strand is complete but contains a ‘nick’ at a unique site, while the plus strand is incomplete.

The genome sequence has termini with cohesive ends that match the distinctively located 5’ ends of the two strands, and maintain the circular configuration of the DNA.

The 5’ends of both the strands incorporate direct repeats (DRs), regions of short repeat sequences, 11 nucleotides long, which are crucial in priming viral replication.

The 5’end of the negative DNA strand encodes the first DR, termed ‘DR1’, while the positive DNA strand starts with another direct repeat, ‘DR2’. The negative strand also has a terminal protein, which is a part of the viral polymerase, covalently linked to its 5’ end.

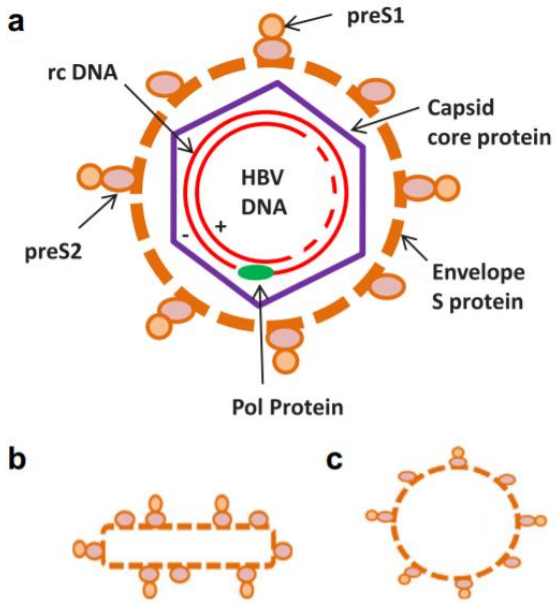
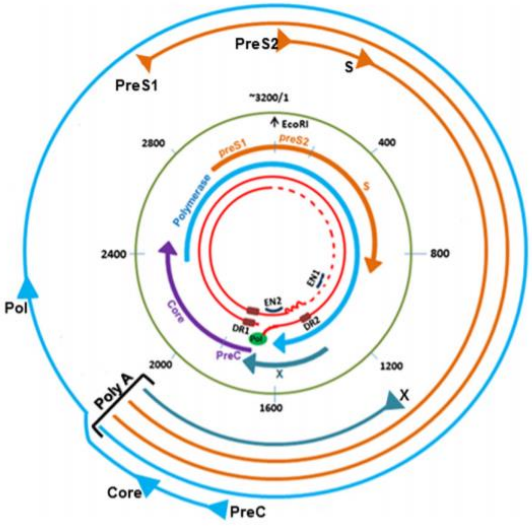


Figure 1 Schematic representation of different forms of infectious and non infectious hepatitis B virus particles. (a) Complete viral particle, (the 42–47 nm Dane particle), (b & c) two species (filamentous and spherical respectively) of non-infectious 20 nm surface antigen particles. ‘Pol’ indicates HBV polymerase protein. Figure not according to the scale.

Viral Genomic Organization



- Schematic diagram of the genome and translational map of HBV.
- The innermost red circles depict the circular DNA with the reverse transcriptase/ polymerase (Pol) attached to the 5' end of the complete minus strand DNA (solid red circle) and a capped RNA oligomer (wavy red line) attached to the 5' end of the incomplete plus strand DNA (dotted red circle).
- The positions of the direct repeats (DR1 and DR2) and enhancers (EN1 and EN2) are also indicated. The green line indicates the viral genome positions in nucleotides (approximate).
- The four protein-coding regions are shown between the green and red circles by coloured semi-circular arrows. They include the pre-core (PC) and core genes (violet), the polymerase gene (blue), the X gene (aqua) and the envelope genes preS1, preS2, and S (orange). Genome positions may change depending upon the genotype of the HBV genome (modified from Nassal 2008). The outermost semi-circular lines with arrowheads represent the 4 RNAs (genomic and sub-genomic) corresponding to the ORFs. The arrowheads indicate the positions of different initiation codons within each ORF.

LONG TERMINAL REPEATS RETROVIRUS

¬ They are found in retroviral DNA

- ¬ Many linear virus genomes have repeat sequences at the ends(termini), in which case the sequences are known as terminal repeats.
- ¬ example: **LTR**—PBS—PSI—gag—pol—env—**LTR**.
- ¬ If the repeats are in the same orientation they are known as **direct terminal repeats** (DTRs).
- ¬ If they are in the opposite orientation they are known as **inverted terminal repeats** (ITRs). ITRs in single stranded nucleic acids are not repeated until the second strand is synthesized during replication.
- ¬ partially transcribed into an RNA intermediate, followed by reverse transcription into RNA and ultimately dsDNA with full LTRs.
- ¬ The LTRs mediate **integration of the retroviral DNA via LTR specific integrase** into another region of the host chromosome
- ¬ HIV use this basic mechanism.
- ¬ Strong promoters are present within LTRs
- ¬ Do not encode proteins'

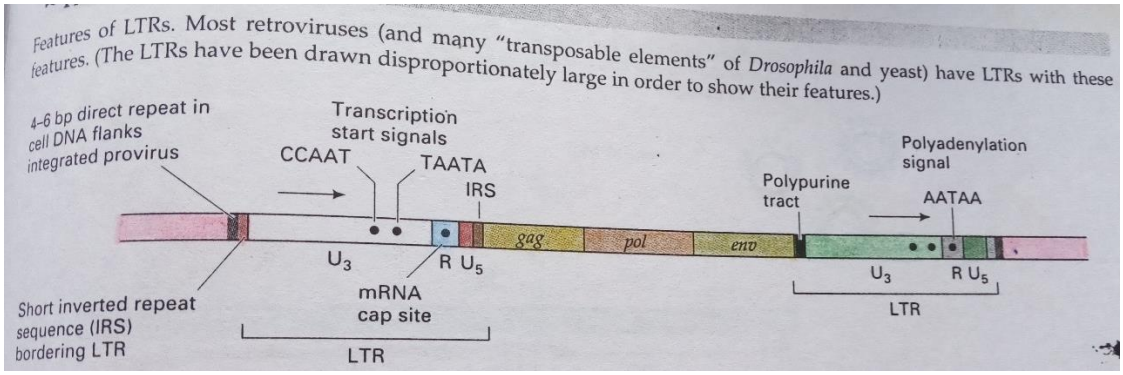


Figure a

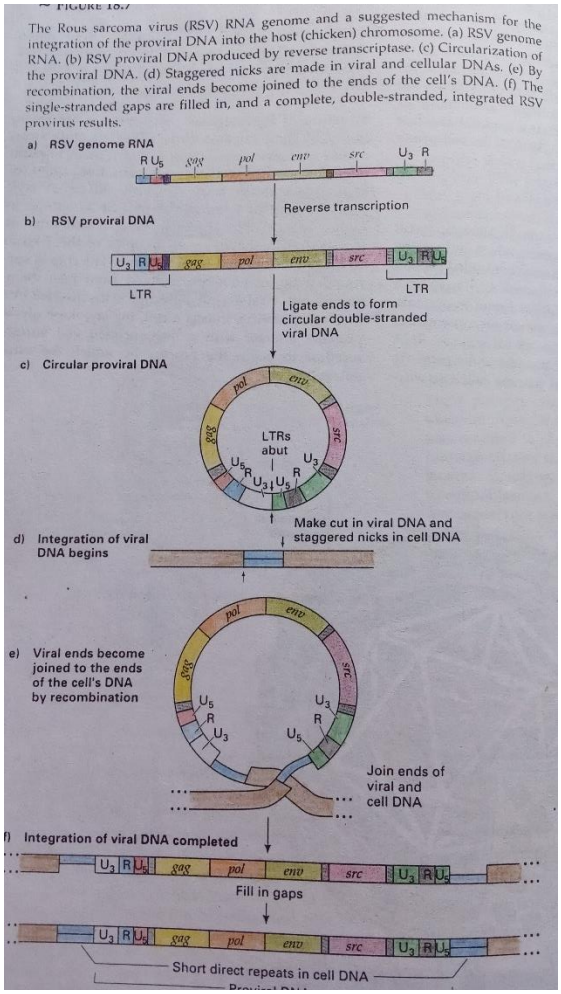


Figure b

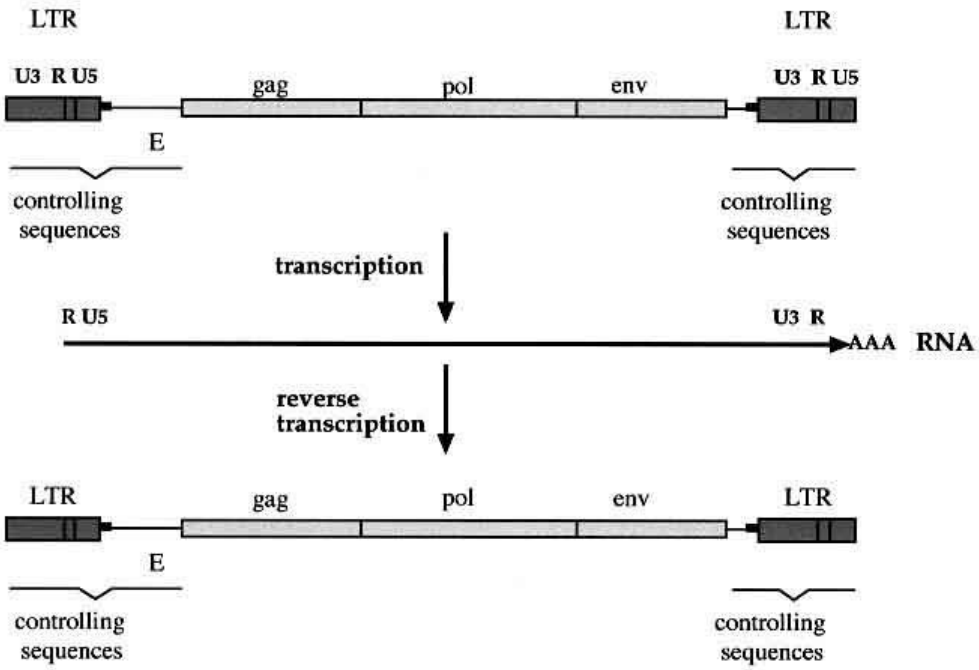


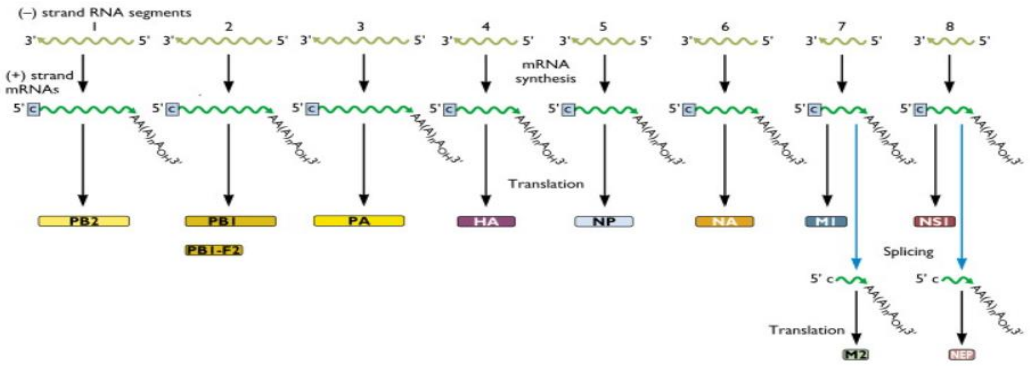
Figure c

Function of LTR

All infectious retrovirus DNAs end with a series of nucleotides derived from the 5' and 3' ends of viral RNA and called the long terminal repeat (LTR). The LTR also has been called the large terminal repeat, the long terminal redundancy and the terminal repeat sequence.

- Retroviruses must have sequences **that will ensure synthesis of progeny viral RNA** and
- **viral mRNA**, **encapsidation of progeny viral RNA in virion proteins**,
- **reverse transcription of progeny viral RNA by reverse transcriptase** and
- **integration of progeny viral DNA into cell DNA** so that the cycle can be repeated.
- The LTR appears to be a region that contains **the control sequences** for almost all these functions and that is **separated from the coding sequences for virion proteins**.
- They then **act as a primer for copying the U3** region at the 3' end of viral RNA, making **one copy** of the **LTR**. This **copy of the LTR then acts as a template** for synthesis of **another copy of the LTR**, which is **transferred to the 5' end** of the virus (reviewed by Temin, Cell 27, 1-3, 1981).
- The LTR has sequence homologies to the **5' noncoding and 3' noncoding regions** of **eukaryotic genes**, which indicates that these sequences **may have a role in RNA transcription**.
- The LTR also has other **sequence homologies to the ends of cellular movable genetic elements**, which indicates that these sequences **may have a role in integration** (and/or transposition) (see Temin, op. cit.).

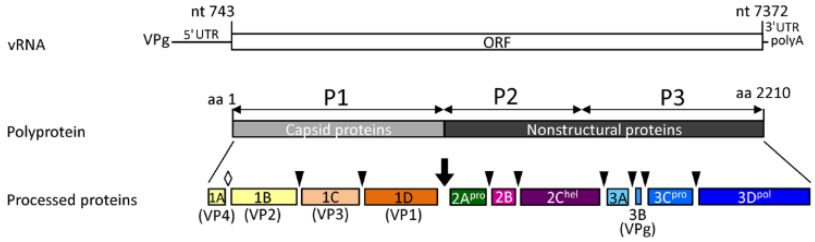
SEGMENTED GENOME THE GENOME IS COMPOSED OF SEPARATE SEGMENTS. INFLUENZA VIRUS



- The influenza A and B virus genomes each comprise **eight negative-sense, single-stranded viral RNA (vRNA) segments**, while **influenza C virus has a seven-segment genome**.
- The eight segments of influenza A and B viruses (and the seven segments of influenza C virus) are numbered in order of decreasing length.
- In influenza A and B viruses, segments 1, 3, 4, and 5 encode just one protein per segment: the PB2, PA, HA and NP proteins.
- All influenza viruses encode the polymerase subunit PB1 on segment 2.
- segment 6 of the influenza A virus encodes only the NA protein, while that of influenza B virus encodes both the NA protein and the NB matrix protein.
- Segment 7 of both influenza A and B viruses code for the M1 matrix protein.
- Finally, both influenza A and B viruses possess a single RNA segment, segment 8, from which they express the interferon-antagonist NS1 protein and, by mRNA splicing, the NEP/NS2, which is involved in viral RNP export from the host cell nucleus.
- The ends of each RNA segment form a helical hairpin, which is bound by the heterotrimeric RNA polymerase complex

NON-SEGMENTED GENOME PICONARVIRUS

- Consists of linear positive sense ssRNA of 7-8kb size.
- Genome is unsegmented.
- There is an Untranslated Region (UTR)at both ends of piconarvirus genome.
- The 5'UTR is longer compared to that of the 3'UTR.
- The 5'UTR is important in translation and the 3'UTR in negative strand synthesis.
- The UTR regions contains information for regulation of translation and mRNA stability.



- Picornavirus virions are nonenveloped and the **+ssRNA nonsegmented genome** is encapsulated in an **icosahedral protein structure made from four capsid proteins encoded by the virus**.
- Virions contain **one molecule of positive-sense ssRNA of 6.7–10.1 kb** and possessing a **single long ORF**.
- A poly(A) tail, heterogeneous in length, is located after the 3′-terminal heteropolymeric sequence. A small protein, VP-g (c. 2.2 to 3.9 kDa), is linked covalently to the 5′-terminus.
- The untranslated regions (UTRs) at both termini contain regions of secondary structure which are essential to genome function.

CAPPING TOBACCO MOSAIC VIRUS (TMV)

- TMV is a **positive sense ssRNA** virus that infects plants(tobacco).
- Exhibit capping and tailing of genome.

CAPPING

- Tobacco mosaic virus (TMV) is a **plus-strand RNA plant virus**. The 5' end of the genomic RNA is terminated by **7- methylguanosine in a 5' → 5' linkage giving** a cap structure.
- The mechanism whereby the cap structure is attained by the viral RNA transcript has not been determined.
- A virus-coded polypeptide with **guanylyltransferase activity** has been investigated.
- This enzyme is responsible for forming the 5'→5' linkage of guanosine 5'-monophosphate to the 5'-diphosphate of an acceptor RNA, thereby forming the cap.
- A critical step in the mechanism for cap formation in the eukaryotic nucleus is for guanylyltransferase to bind covalently to guanosine 5'- monophosphate with the hydrolysis of pyrophosphate when guanosine 5'- triphosphate is the substrate.
- The TMV 126-kilodalton protein, which is most probably a component of the TMV replicase, was found to have this activity
- The capping enzymes adds a GTP molecule generating **a phosphate link-TO phosphate**. The linkage is **5' to 5' rather than the normal 5' to 3'.**
- Next a methyl group is added by methyl transferase to nitrogen no 7 in the guanine base.
- The function of this cap structure seems to **be to protect the mRNA from degradation**.
- It also seems to be necessary for **binding the mRNA to the ribosome during translation initiation**.

TAILING

- **Adenyl AMP** residues are added to the **new 3' end of the pre mRNA** by an enzyme called **poly A polymerase.**
- This generates a poly A tail at the 3' end of the pre mRNA
- Poly A tails are present at the 3' end of all mRNAs except the mRNAs for histones.
- Function is unknown.
- The poly A tail may have something to do with protecting the 3'end of the mRNA from degradation. A cap and a poly A tail on a genome RNA may indicate that the molecule is ready to function as mRNA, but neither structure is essential for translation