Structure and Organization of Virus Genomes

1

Abstract

This chapter provides an in depth study on the structure, composition, and organization of viral genomes, their classification into double stranded and single stranded DNA viruses, positive and negative stranded RNA viruses with and their genome diversity. Segmentation and re-assortment of viral genomes have been discussed along with the multipartite virus genomes. Genome details of 13 different viruses have been provided as type studies for better understanding of these topics. Concepts of viral genome evolution have also been discussed.

1.1 Introduction

A Virus is a small electron microscopic parasite, incapable of reproducing by its own, survives by directing the host cell machinery for the production of more viruses, which emerges from their respective host cell through lysis. Most of these viral organisms contain either double stranded or single stranded DNA as well as RNA in their genomes, which may be either single stranded or double stranded. After the purification and partial crystallization of Tobacco Mosiac Virus in 1935 by Wendell Stanley, the study of viruses has inspired many scientists, which lead to identification and characterization of plant, bacteria, archaea, and animal viruses. Since viruses are capable of infecting a large number of various cell types, genetically modified viruses are being considered for the gene therapy. All these factors and applications make the virus an important organism for its capability to infect any living organism on this planet.

Viruses are small submicroscopic, obligate intracellular parasites, which contains either DNA or RNA as genome protected by a virus-encoded protein coat called capsid. Viruses are mobile genetic elements, depends on metabolic and biosynthetic machinery of host cells for their propagation. Viruses cannot carry out their life sustaining functions outside the host cell. They cannot synthesize proteins as they lack ribosomes and uses the host cell ribosome machinery for translating their

K. V. Chaitanya, *Genome and Genomics*,
https://doi.org/10.1007/978-981-15-0702-1_1

mRNA into proteins. Viruses can neither generate nor store ATP but derive the necessary energy and other metabolic functions by parasitizing the host cell for the basic materials such as amino acids, nucleotides, and lipids. Even though the viruses are speculated as a form of proto-life, their inability to survive outside living organisms does not make them as living organisms in the strict sense and it is highly unlikely that they preceded cellular life during the evolution of the earth. Some scientists even speculate that viruses have begun as rogue segments of genetic code which got adapted to a parasitic form of existence. Virions are the complete virus particles that are produced by the assembly of the pre-formed viral components, whose main function is to deliver its genome into the host cell for its expression (transcription and translation). The icosahedral virion belonging to the *icosahedral* or *quasi-spherical* structural class of viruses is made of 20 identical triangular faces, each face is constructed with three identical capsid protein units making 60 subunits per capsid, with five subunits symmetrically contacting each of the 12 vertices, thus making all proteins in equivalent interaction with each other. The viral genome is packed inside a symmetric protein capsid, composed of either a single or multiple proteins, each of them is encoding a single viral gene. Due to this symmetric structure, viruses could encode all the necessary information for constructing a large capsid using a small set of genes. A capsid along with the enclosed nucleic acid is called a nucleocapsid. In enveloped viruses, the nucleocapsid is surrounded by a lipid bilayer and is studded with a layer of glycoproteins. Often, the nucleic acid is associated with a protein called nucleoprotein. Viruses possess a great diversity with respect to their size. *Mimivirus* is the largest virus reported with 400 nm in diameter, bigger than the *Mycoplasma* bacteria, which is ~200 to 300 nm long. They also exhibit a wide diversity in shapes and forms, such as spherical, rod-like, etc.

There are few subviral entities which are highly similar, pathogenic and are possessing the properties similar to that of viruses. These entities are viroids, virusoids, and prions. Viroids are a short stretch of highly complementary, single stranded (200–400 nucleotides), circular RNA molecules possessing a rod-like secondary structure without any capsid or envelope. These viroids are associated with plant and human diseases such as hepatitis D. They are obligate intracellular parasites, with replication strategies similar to viruses. Virusoids are a satellite, circular single-stranded RNAs (1000 nucleotides) dependent on plant viruses for replication and encapsidation. They are packed into virus capsids as passengers. Their genome encodes only structural proteins. Prions are anomalous infectious agents that cause fatal neurodegenerative diseases mediated by contemporary mechanisms. Prions consist of a single type of protein molecule without any nucleic acid component. The protein is a modified isoform of prion protein (PrP) designated PrPSc. The normal cellular PrPC is converted into PrPSc by a structural transition of its α-helical and coil structure is refolded into β-sheet. The prion protein and its encoding gene are often found in normal uninfected cells and are associated with viral diseases such as Creutzfeldt–Jakob disease in humans, scrapie in sheep and bovine spongiform encephalopathy (BSE) in cattle.

1.2 General Features of Viral Genomes and their Classification

It has been estimated that there are 10^{31}–10^{32} viruses in the earth's atmosphere, which exceeds the number of host cells fairly by an order of magnitude. As a consequence, every organism on the planet or even every living cell is under constant attack from viruses, even viruses are highly responsible for the greatest selection pressure on the living organisms. In spite of their small size, viruses play an important role as obligate intracellular parasites, modulating their host cells for energy and reproduction leading to adverse effects. The main emphasis of virology is focused on the identification and control of pathogenic viruses that invade humans, domestic animals, and plants. But, origin and organization of viruses, their evolution is the deep questions which are fundamental to molecular virology. The comparative genomics has allowed closely related viruses to be compared and classified. In addition, the sequencing of eukaryotic genomes has revealed that 5–10% of their DNA encodes information for these organisms. A large fraction of the remainder is thought to be composed of mobile retrovirus-like elements (retrotransposons), which may have played a considerable role in shaping these complex genomes. Bacterial genomes do not have such extra genetic material. But, the genomes of certain bacteriophages have a close resemblance with bacterial plasmids in their structure and in the way of their replication, revealing that the relationship between viruses and other living organisms is perhaps more complex than what was previously thought.

1.2.1 Classification of Viral Genomes

Currently, over 4000 viruses have been described, classified into 71 families. Even though viruses possess small genomes, they exhibit enormous diversity compared with plants, animals and even bacteria. With respect to the genome, viruses are broadly divided into DNA viruses and RNA viruses. Both DNA and RNA viruses can either single stranded or double stranded, with a circular, linear or segmented arrangement. DNA and RNA viruses are distinguished by their features, such as monopartite or multipartite. In monopartite, their genome is having a single nucleic acid molecule. All double stranded DNA genomes contain only a single nucleic acid molecule and few of the viruses with single stranded genomes are reported to have multiple segments. In contrast, RNA viral genomes are generally multipartite, with more frequency for single stranded RNA viruses. Additionally, single stranded virus genomes may be either positive sense (+) where the RNA present in the genome will of the same polarity as mRNA and will encode the genes or negative sense (−), where the entire ssRNA genome must be copied and the copied strand is transcribed. Some single stranded viruses are ambisense (a mixture of + and − sense). Bacteriophages such as MS2, Qb, and Mimivirus belonging to family Leviviridae consists of ambisense viral genomes. Size of the DNA viruses is larger than that of RNA viruses. Few DNA viruses can be as large as 305,000 nucleotides. DNA viruses are called large viruses and RNA viruses are small. Size of a few single

stranded RNA genomes is up to 31,000 nucleotides. The small size of the single-stranded virus might be limited due to the fragility of the RNA, which provides the tendency of the large RNA strands to break and also due to the fact that RNA viruses are more susceptible to mutations than DNA viruses. Single stranded DNA and RNA viruses are fragile than double stranded viruses.

The genomes of the largest double-stranded DNA viruses such as herpesviruses and poxviruses are quite complex, resembling their host cells. Genomes of Polyomavirus are complexed with cellular histone proteins, which form a chromatin-like structure inside the virus particle. After infecting the host cell, these genomes behave like miniature satellite chromosomes inside the host cell following the dictates of cellular enzymes and the cell cycle. mRNAs of Vaccinia virus were polyadenylated at their 3¢. The genome of Adenovirus consists of split genes with non-coding introns, protein-coding exons, and spliced mRNAs. In order to streamline the replication cycle, the Adenovirus takes control of many biological processes in the cell, such as alternative RNA splicing and polyadenylation for expanding the coding potential of the limited viral genome.

Phage genomes are simple and least complex. Introns in prokaryotes were first discovered in the genome of T4 phage in 1984. The genome of T4 phage is 160 kbp double-stranded DNA, highly compressed with promoters and sequences that control translation are nested within the coding regions of overlapping upstream genes. It consists of three self-splicing group I introns, located in the genes that codes for thymidylate synthase (td), aerobic ribonucleotide reductase (nrdB) small subunit and the anaerobic ribonucleotide reductase (nrdD). Like any other group I introns, the td and nrdD introns each contain an open reading frame encodes for a homing endonuclease that renders the introns mobile and they can be inserted into new phage genomes. The nrdB intron is non-mobile due to a deletion in the intron-borne homing endonuclease gene. It was speculated that the presence of in these introns may confer a selective advantage to the phage by offering the possibility of regulating the expression of the intron-containing genes by the regulation of splicing. It was also believed that horizontal transfer of the introns between phages through homing has played a significant role in the evolution of group I introns in T-like phages.

The size of viral genomes also depends on the type of host cell. Viruses with prokaryotic host cells tend to replicate quickly to keep up with their host cells, which are reflected in the compact nature with overlapping genes of many bacteriophages, leading to the minimum genome size. Viruses with eukaryotic cells as hosts show tremendous compression while the core is getting packed into the capsid so that only optimum amount of genome can be packed. Viruses pack their genomes into protective protein contained capsids. Packaging strategies of viruses are related to the type of genome. Viruses with double stranded DNA genomes use a molecular motor with a spool like structure to pack their genomes into the capsid. In contrast, viruses with single stranded DNA or RNA genomes employ a co-operative mechanism, in which package and assembly of the genome into the capsid will occur simultaneously, enhancing the assembly efficiency of the capsid. In bacteriophage MS2, genomic RNA sequence will form a short term loop, for its interaction with the

Table 1.1 General features of viral genomes sequenced

S. No.	Class	Sequenced genomes	Size (Nt)	Proteins
1	DsDNA	414	4697–335,593	6–240
2	SsDNA	230	1360–10,958	6–11
3	DsRNA	61	3090–29,174	2–13
4	SsRNA (+)	421	2343–31,357	1–11
5	SsRNA (−)	81	8910–25,142	5–6

capsid proteins. This interaction will trigger conformational changes that convert symmetric protein dimers into asymmetric form, needed to build the capsid. Some bacteriophages of the family Myoviridae, such as T4 contain relatively large genomes, up to 170 kbp. The largest viral genome currently is known that of Mimivirus, ~1.2 Mbp, consisting of 1200 open reading frames, with only 10% of them showing the similarity with proteins of known function. Among the eukaryotic viruses, herpesviruses and poxviruses possess relatively large genomes, up to 235 kbp. These genomes contain genes involved in their own replication, particularly enzymes concerned with nucleic acid metabolism. Therefore, these viruses can partially escape the restrictions of the host cell biochemistry by encoding additional biochemical apparatus with the penalty of encoding all information necessary for a genome to pack into the capsid, which also causes upward pressure on its size. Whatever might be the composition of a genome, all viruses are obligate intracellular parasites, which can replicate only inside the appropriate host cells, their encoded genomes must be recognized by a specific host cell, which is getting parasitized. For that, the genetic code employed by the virus must either match or recognized by the host cell. Similarly, the signals that regulate the expression of viral genes must be inappropriate to the host. The general features of viral genomes sequenced are displayed in Table 1.1.

1.3 Size, Structure, and Composition of Double Stranded DNA Virus Genomes

Cellular life forms of all categories possess genomes with double stranded DNA and utilize the same as a standard scheme for their replication and expression. In contrast, viruses and other pathogenic and selfish elements exploit all possible interconversions of nucleic acids, with their genomes that can be either DNA or RNA, which are single-stranded or double-stranded. Viral genomes that have been sequenced and annotated are compared with genomes of cellular life forms, which were small with unknown gene functions. But, in the past few years, discovery of giant viruses has rapidly expanded the size of viral genomes whose range spans up to 3 orders of their magnitude, from ~2 kb to ~2 Mb. Surprisingly, the genomes of giant viruses are larger than the genomes of many bacteria and archaea, obliterating the gulf between cells and viruses in terms of genome size and complexity. A number of viral groups possess double-stranded DNA as their genomes are of

considerable complexity, classified into small and large genomes depending on the type of replication. Small genomes use a host DNA polymerase for replication, In contrast, large genomes encode a virus-specific DNA polymerase, responsible for their genome replication. These viruses are genetically similar to the host cells they infect. Large DNA viruses encode more proteins than small DNA viruses.

There are two major viral groups representing the double stranded DNA genomes ie. members of the Adenoviridae and Herpesviridae families. Human adenovirus is one of the most common pathogens that causes minor, self-limiting illness to most of the patients. Adenovirus is one of the very well understood viruses, whose basic biology has been extensively studied over the past 60 years. Human Adenovirus has been isolated in the early 1950s from the adenoid tissue causing respiratory infections. This virus has a remarkable capacity to spread among patients contacting from as few as five virus particles. Adenovirus-induced acute respiratory disease is the most common infection in confined populations of daycare centers, hospitals, retirement homes, and military training venues, accounting for ~8% of all childhood respiratory tract infections, which can lead to bronchitis, bronchiolitis, or pneumonia, requiring hospitalization in ~25% of diagnosed cases. It also causes other localized diseases, such as colitis, hemorrhagic cystitis, hepatitis, nephritis, encephalitis, myocarditis and disseminated disease with multiorgan failure. A few such diseases can be more serious in pediatric and geriatric populations, especially in the individuals with suppressed immune systems, such as transplant recipients or patients suffering from AIDS. Many aspects of Adenoviral life cycle have been completely elucidated in great detail, which has allowed the development of adenoviral vectors that are highly efficient in delivering genes into the mammalian systems, especially human cells for the transgene expression and for delivering therapeutic genes in human gene therapy. Adeno viruses specifically consist of a nucleoprotein core with a 30–40 kb linear double-stranded DNA, surrounded by an icosahedral, non-enveloped capsid of 70 to 100 nm diameter. These viral genomes contain 30–40 genes. The terminal sequence of each DNA strand has an inverted repeat of 100–140 bp. The denatured single strands form a '**panhandle**' like structures, which are important for the DNA replication. A 55 kDa protein known as the terminal protein is covalently attached to the 5¢ end of each strand. During the genome replication, this protein might acts as a primer, for initiating the synthesis of new DNA strands. The expression of the genes is rather more complex in Adenoviruses with clusters of genes expressed from a limited number of shared promoters. Multiply spliced mRNAs and alternative splicing patterns are used to express a variety of polypeptides from each promoter.

Apart from the differences in host and tissue tropism, a very less amount of variation is found in the Adenoviral genomes and their structural parameters. Human adenoviral serotype 5, is one of the highly characterized adenovirus, consisting of ~36 kb genome. Its coding region is divided into early (E1–E4) and late (L1–L5) transcripts based on their stage of expression. Essential early E1 region is deleted in most adenoviral vectors, rendering their incapability of replicating in most cell lines. Numerous studies have shown that after the deletion of E1, Adenoviral vectors are more ideal for the in vivo and in vitro studies requiring short-term transgene

expression. The Second generation Adenoviral vector constructs are made after deletions in the essential E2 or E4 regions, facilitating the prolonged transgene expression. Helper-dependent Adenoviral vectors (hdAd) are generated by deleting all genes that code for viral proteins for increased cloning capacity. Removal of protein coding sequences in these vectors allows the overall reduction in their genome size from 36 kb for the wildtype to 30 kb in E1/E3-deleted Adenoviruses and ~500 bp for helper-dependent Adenoviral vectors.

1.3.1 Analysis of Adenovirus 2(AD2) Genome

The genome of adenovirus 2(Ad2) was the first adenoviral genome to be fully sequenced, which is of the size ~36 kb, encoding over 40 proteins. Adenoviral coding regions are designated as early or late depending on their expression before or after the DNA replication. The early genes E1A, E1B, E2, E3, and E4 are the first ones to get transcribed. They encode for the proteins involved in activating transcription of other viral regions, altering the cellular environment to promote viral production. E1A proteins also induce mitogenic activity in the host cell, which stimulates the expression of other viral genes. E2 proteins regulate viral DNA replication, while E3 and E4 proteins are involved in altering the host immune responses and cell signaling. Activation of the major late promoter (MLP) followed by the start of viral DNA synthesis, allowing the expression of late genes encoding primarily virion structural proteins. L1–L5 of the late regions are transcribed from an alternatively spliced transcript. The regions encoding the L4-22 K and L4-33 K proteins are initially expressed at low levels from a novel promoter located within the L4 region and these proteins functions in fully activating the major late promoter (MLP). Four small proteins including the structural protein IX (pIX) and the IVa2 protein produced at intermediate/late stages of infection, helps in packing of viral DNA into immature virions. The late products VA RNA I and II inhibits the activation of the interferon response, impede cellular micro-RNA processing and also influence the expression of host genes. There are 100 bp inverted terminal repeats (ITRs), located at both ends of the genome, which act as the sites for the origin of replication, with the ~200 bp viral packaging sequence positioned next to the left ITR (Fig. 1.1). Even though Adenovirus is being studied in great detail for more than 60 years, our knowledge of the genes encoded by this virus is still expanding. In 2007, a new open reading frame (ORF) has been identified to be located between the fiber ORF and E3, which is termed as U exon. The U exon protein (UXP) is expressed from a unique promoter during later stages of infection and is hypothesized to play a significant role in the transcription.

Most of the transcription processes in the adenovirus result in more than two alternatively spliced mRNAs. Keeping the compactness of the adenovirus genome, regulatory events that take place at the level of RNA processing are of great importance for controlling the lytic cycle of the virus. Splicing of large introns results in the production and accumulation of shorter mRNAs, at the later stages of viral infection. Except for E1A, L4-33Kand the U exon protein, viral introns do not

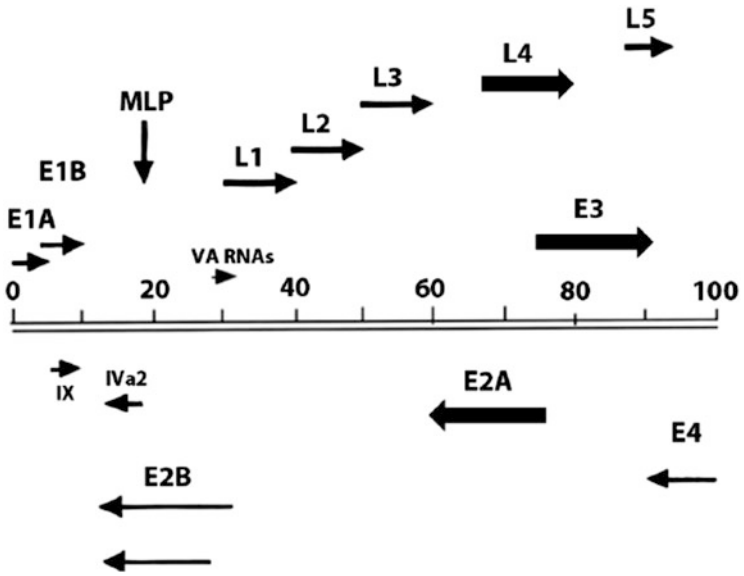

Fig. 1.1 Organisation of the Adenovirus genome (Courtesy Russel WC, School of Biology, University of St Andrews, UK)

interrupt the ORF of the gene. Further, Adenoviral genes contain very few introns compared to that of cellular genes. Viral mRNAs mature by removing one to three introns. As the virus has to compress much of its genetic information into a small genome, the selection pressure appears to have favored few and small introns. Deep cDNA sequencing of Adenovirus genome has identified many novel alternatively spliced transcripts, suggesting that there may be numerous new or altered polypeptides produced by this virus in the infected host cell, reflecting that this genome still has many secrets that remain to be uncovered.

1.3.2 Herpesviridae

Herpesviridae belongs to a large family of ~100 members with at least one for each of the animal species that have been examined till date. There are eight human herpesviruses, all of them share a common overall genome structure, but differs in the fine details of genome organization and at the level of the nucleotide sequence. Herpesviridae is a family of enveloped, DNA viruses with complex genomes. They replicate in the nucleus of a wide range of vertebrate and invertebrate hosts, such as humans, horses, cattle, mice, pigs, chickens, turtles, lizards, fish, and invertebrates such as oysters. There are eight human Herpesviruses divided into three subfamilies. Herpesviruses possess very large complex genomes composed of large complex virus particles up to 235 kbp of linear, double-stranded DNA and ~35 virion polypeptides. Members of Herpesviridae are highly diversified in terms of their genome sequence and protein synthesis but show a great similarity in terms of

structure, genome organization and almost all of their genomes encode the enzymes involved in nucleic acid metabolism, DNA synthesis, and protein processing. Few Herpesviral genomes consist of two covalently joined sections named as a unique long (UL) and a unique short (US) regions, each bounded by inverted repeats. These repeats allow structural rearrangements of the unique regions, facilitating the existence of genomes as a mixture of four functionally equivalent isomers. These genomes also reported contain multiple repeated sequences and depending on their number, size of the genome may vary up to 10 kbp.

Integration of viral genome into the chromosomes of the host cell is mandatory for the successful completion of the virus lytic cycle. In contrast, Herpesviruses maintains their genomes as extrachromosomal circular episomes in the nuclei of infected cells without the need for integration. There are also reports of chromosomal integration of Herpesviral DNA, suggesting that Herpesviruses are also capable of integrating into the host's chromosomes under few circumstances. It was hypothesized that the replication of non-integrated Herpesviral DNA occurs through the rolling-circle mechanism, yielding long DNA concatemers that are subsequently cleaved into single genome equivalents during nucleic acid encapsidation. In addition, Human Herpesvirus 6 (HHV-6) is found to be integrated into the germlines of approximately 1% of the world's population. But how the replication of linear CIHHV DNA occurs still remains unknown. Chromosomal insertions of α-Herpesvirus DNA, Herpes simplex viruses and Equine Herpesvirus have been detected following infection with defective interfering particles or transfection of sheared or subgenomic viral DNA fragments. The integrated viral genome consists of mostly subgenomic fragments without any possibility for the production of infectious viral particles to occur. Many of the cells carrying integrated viral DNA displayed a transformed phenotype, fueling the hypotheses on the oncogenic nature of these viruses.

1.3.2.1 Analysis of Herpes Simplex Virus Genome

The prototype member of the Herpesviridae family is Herpes Simplex Virus (HSV), whose genome is ~152 kbp, composed of double-stranded DNA. The complete nucleotide sequence of the Herpes Simplex Virus has now been determined. This virus contains about 80 genes, densely packed and with overlapping open reading frames. Each gene is expressed under its own promoter. HSV-1 is perhaps the most intensively studied complex virus genome. Before the development of nucleotide sequencing, the HSV genome has been extensively mapped by conventional genetic analysis and mutant analysis. The HSV-1 genome is composed of a single linear double stranded DNA of ~152,261 base pairs in length. The genome is divided into two segments called Unique Long (UL) and Unique Short (US). Small regions of repeated sequence occur at the genome ends between the L and S segments. As DNA is replicated, the inversion of L and S segments takes place at a very high rate, creating four genome isomers, occurring at equal frequencies in most of the HSV-1 wild type populations. This virus genome synthesizes 75 genes encoding for known proteins. Among them, 69 genes are found to exist in a single copy and three genes in two copies each. Among the 75 genes with identified functions, 43 are considered to be the core genes common to α, β and γ Herpesviruses located in the UL segment of

the genome. These genes are involved in many vital functions including the entry of the core DNA into the host cell and its replication, assembly into the capsid and its dispersal. All genes located in the unique short segment are non-core and are highly divergent. These are mainly found at the ends of the segment. Proteins encoded by the non-core genes are involved in both lineage and species-specific functions such as transcriptional transactivation, immune evasion, and host cell recognition.

Type 1 Herpes simplex virus is a member of the α-Herpesvirinae subfamily of the Herpesviridae family, whose infection results in cold, ocular, genital sores and encephalitis. Several strains of HSV-1 have been isolated varying in virulence, which might be due to base substitutions resulting in amino acid or *cis*-regulatory changes. One of the HSV-1 strains, KOS isolated from a human labial lesion and is frequently used as a marker to investigate the HSV-1 gene function and pathogenesis. KOS is less virulent compared with other HSV-1 strains, such as McCrae and 17, which has raised the curiosity for the comparative genomics. KOS genomic DNA was isolated from infected African green monkey (Vero) kidney cells and an unpaired 42-bp Illumina library was generated and run at Genome Technology Access Center, Washington University. Since viral DNA was isolated from Vero cells, potential contaminating host reads that matched the *Rhesus macaque* and/or human genomes were removed using Bowtie. The remaining reads of 16,494,831 bp were assembled into contigs using the Velvet de novo assembler against the reference HSV-1 strain 17 genome (GenBank accession number NC_001806) with SeqMan Pro (DNASTAR, Inc.). Because the HSV-1 genome includes two sets of inverted repeat regions TRL/IRL and IRS/TRS, contigs assembling into one of the repeat units were reverse complemented and also placed into the other repeat unit. The final KOS genome is a linear double stranded DNA of 152,011 bp, consisting of 80 genes and has 13 gaps, exclusively at VNTR regions, totaling up to 1582 bp in length. In the Gen Bank annotation, the sequence and length of each VNTR were copied from strain 17. Using Bowtie for aligning the filtered reads against the de novo assembly, the average sequence coverage per base pair for the KOS genome was determined to be 4257 (Fig. 1.2).

To identify nucleotide variants between the genomes of strains KOS and 17, the genomes were aligned using fast statistical alignment (FSA) and applied custom Perl and R scripts. KOS differs from strain 17 by 1024 SNPs, 320 of them are non-synonymous changes in 65 of 77 HSV-1 open reading frames. The two genomes also differ by 172 indels, most of them are insertions or deletions of single bases in non-coding regions. However, 26 indels are in frame additions or deletions of codons. Further analyses are in progress for the comparative genomics of KOS with other genomes of HSV-1 strains. Such studies will increase the scope for better identification of the genetic attributes of KOS and its contributions to its pathogenesis.

Fig. 1.2 Organisation of the Herpes Simplex Virus-1 genome

1.4 Genomes of Single Stranded DNA Viruses and their Mosaicism

Viruses with single stranded DNA genomes infect hosts that belong to all three domains of life and are considered to be economically, medically and environmentally important pathogens. Recent studies have shown that these single stranded DNA viruses exist in great numbers in highly diverse habitats, ranging from extreme geothermal springs to the gut of humans and other animals. International Committee on Taxonomy of Viruses currently classified single stranded DNA viruses into 10 different taxa. However, several viruses that can be classified into additional groups have been isolated and many of their genomes were sequenced. All single stranded DNA viruses are pathogenic on eukaryotes, possess non-enveloped, icosahedral capsids, along with Microviridae family members, which infects bacteria. Single stranded DNA viruses pathogenic on other prokaryotes have filamentous (Inovirus), rod-shaped (Plectrovirus), coil-shaped (Spiraviridae), or pleomorphic (proposed family"Pleolipoviridae") virions (Table 1.2).

Single stranded DNA viruses are the group comprising of smallest viruses and their genomes are as small as 1–2 kb, encoding two proteins; one for capsid formation and the other for genome replication. Such irreducible simplicity of single stranded DNA viruses epitomizes their essence of being a virus and makes them an attractive model for investigating virus origins and evolution. Numerous metagenomic studies have revealed a high range of genetic diversity existing in single stranded DNA viruses in the environment, suggesting a highly dynamic interaction between these viruses and their respective hosts. Also, single stranded DNA viruses with the smallest genomes and simplest proteomes were found to be widespread in cellular chromosomes, providing new important insight into the evolution of these viral.

Table 1.2 Morphological diversity of single stranded DNA viruses

Host virus taxon	Virion morphology	Genome topology	Genome size
Microviridae	Icosahedral	Circular	4.4–6.1
Inoviridae			
Inovirus	Filamentous		5.8–12.4
Plectrovirus	Rod-shaped		4.5–8.2
Pleolipoviridae	Pleomorphic	Circular	7–10.6
Spiraviridae	Coil-shaped	Circular	24.9
Anelloviridae	Icosahedral	Circular	2–4
Bidnaviridae	Icosahedral	Linear, segmented, 6.5 per segment	
Circoviridae	Icosahedral	Circular	1.7–2.3
Geminiviridae	Icosahedral	Circular, segmented 3 per segment	
Nanoviridae	Icosahedral	Circular, segmented 0.98–1.1per segment	
Parvoviridae	Icosahedral	Linear	4–6.3

1.4.1 Genomes of Bacteriophages

Bacteriophages are the smallest viruses with simple genomes. Since their discovery in 1915 and 1917 by Fredrick Twort and Felix d'Herelle respectively, bacteriophages have been studied in many laboratories and are being used in a variety of practical applications. The Density of phage viruses present in the oceans is 10^6–10^7 particles per ml. It was estimated that the total population of the bacteriophages is 10^{31} particles and the ratio of environmental virus and bacteria are 5–10:1, after the validation of 10^{30} bacterial cells in the biosphere. Altogether, the prokaryotic population is highly dynamic, with an estimated number of ~10^{23} global infections per second. It has been hypothesized that oceanic bacteriophages infect bacterial cells at the rate of 10^{29} phage infections per day, which releases over 10^{11} kg of carbon from the biological pool per day. Over the past three decades, research on bacteriophages has revealed their abundance in nature, genome diversity, impact on the evolution of microbial diversity, their utilization in control of infectious diseases and their influence in regulating the microbial balance in the ecosystem has been explored, leading to a resurgence of interest in the phage research. Research on phages has played a pivotal role in the most significant discoveries, that were made in biological sciences right from the identification of DNA as the genetic material, in the elucidation of the genetic code, leading to the development of the molecular biology. Research on phages has continuously broken new grounds in our understanding of the basic molecular mechanisms of gene expression and their structure. In recent times, phage genomics has revealed novel biochemical mechanisms for replication, maintenance, and expression of the genetic material and is providing new insights into the origins of infectious diseases, utilization of phage gene products and even whole phage as an agent for the gene therapy.

In addition to the killing of bacterial cells, temperate phage genomes also carry toxins and other critical virulence factor genes that are important for many bacterial pathogens to infect human beings. Phages also contribute to the diversity of the bacterial community by serving as vectors for the transduction of different genetic alleles, such as antibiotic resistance genes, between bacterial cells. Phages also have great medical and nanotechnological potential. Strategies for using tailed phages for detecting bacteria, curing bacterial diseases through phage therapy or decontaminating surfaces have been implemented for almost 100 years in Russia and Georgia. These phages are currently being used to treat agricultural diseases as well as in the prevention of food contamination in western countries. Phage virions are being developed as nanocontainers for specific chemical cargoes that can be delivered to specific targets.

Small size and the simplicity of isolation have made bacteriophages as the primary choice for the complete genome sequencing. Phage φX174 is the first organism with the complete genome sequence of 5386 bases of single stranded DNA and λ phage genome is the first organism with double stranded DNA of 48,502 bp, followed by phage T7 genome of 39,936 bp. dsDNA tailed mycobacteriophage L5 is the first among non-*E. coli* phage genomes to be fully sequenced. Further, the sequencing of the bacteriophage genomes are propelled exponentially with two main objectives;

1. To understand the relationship between the phage genomes the evolutionary mechanisms that shaped these bacteriophage populations.
2. For increased utilization of bacteriophages in the development of tools, utilities, and techniques related to genetics and biotechnology.

Phage genomes display a considerable amount of variation in their size, varying from *Leuconostoc* phage L5 (2435 bp) to *Pseudomonas* phage 201 (316,674 b). Tailed phages with double stranded DNA genomes vary in their size from >10 kbp to <15 kbp, consistent with their overall virion structure and gene assembly, which encompass up to 15 kbp of the genome space. *Siphoviruses* of the genome size 1.5–6 kbp are characterized by a long flexible non-contractile tail with a tape measure protein gene, whose length corresponds to the phage tail length Many phages with the morphologies similar to *Siphoviruses* have genomes longer than 20 kbp. Contrastingly, *Myoviruses* with contractile tails are the phages with larger genomes of >125 kbp and the *Bacillus* phage SPBc2, is the largest Siphoviral genome of the length 134,416 bp. The main reason for the absence of large *Siphoviruses* is still unknown.

1.4.2 Phage Genome Sequence Diversity

Bacteriophages are estimated to be the most widely distributed biological entity of the biosphere. They are found in all habitats of the world, where bacteria proliferate. Most of the viral population is dominated by bacteriophages, with double stranded DNA tailed phages, or Caudovirales, accounting for 95% of all the phages, possibly making up the majority of phages on the planet. However, phages belonging to other groups also occur abundantly in the biosphere, such as phages with different virions, genomes, and lifestyles. Two key approaches were made for studying the viral diversity are metagenomics of total concentrated phage samples collected from the environment and a genome-by-genome strategy of individually isolated phages. These two approaches are compatible, having distinct outcomes. Metagenomics generates a large amount of sequence data, which provides a good insight into their diversity. Sequencing and analysis of individually isolated phages generate small data sets, which are structured into whole genomes. As phage genomes are architecturally mosaic, the availability of complete genomes contextualizes the complexities of their relationships. The nucleotide sequences of phage genomes with non-overlapping hosts rarely share sequence similarity, as noticed in the published genomes of four *Streptomyces* phages and available collection of 50 mycobacteriophage genomes. Phages infecting a common bacterial host are in genetic contact with each other, and they share common nucleotide sequences. Genomes of over 30 phages with common host have been isolated and sequenced from *Pseudomonas*, *Staphylococcus,* and *Mycobacterium* containing related sequences, with a few exceptions. Most of these phages share a very low or no sequence similarity, as illustrated by the nucleotide sequence comparisons of *mycobacteriophages* and *Pseudomonas phages*.

1.4.3 Genome Mosaicism of Phages

Phages were evolved not only by the accumulation of mutations but also through the recombination events, during which they exchanged genetic material with other phages. These events have been suggested to explain the mosaic structure of the phages, arisen by comparison of two or more phage genomes. During the comparison of the genomes, nearly identical sequences alternate with merely similar sequences or completely divergent sequences. Such type of exchanges in bacteriophages was obtained by heteroduplex mapping in the early 1990s. Since then, numerous mosaics have been identified by sequence comparison, and the mosaic structure of bacteriophages is now a well-documented phenomenon. This mosaicism is also found to be ubiquitous among bacteria, where the genes are acquired through horizontal genetic exchange mostly through transduction, transformation, and conjugation. But, the extent of mosaicism is highly remarkable in phage genomes as evidenced by the increasing number of genomes available for comparative genomics analysis.

The mechanism of genome mosaicism in bacteriophages can be understood at two levels; 1. by comparing nucleotide sequence through DNA heteroduplex mapping, 2. by comparing their DNA sequences. There are two models which explain the recombination mechanisms that are responsible for these patterns. Model 1 describes the role of short conserved boundary sequences that are located at gene junctions in targeting various exchange events that are catalyzed by homologous recombinations, by using the recombinases synthesized by either host-or phages. Model 2 attempts to explain that the homologous recombination events are not specifically targeted and occur randomly with the preference of a few short sequences so that most of the events results in non-functional genomic trash. Comparison of the predicted amino acid sequences encoding phage gene products is an alternative manifestation of mosaicism. This is an informative approach, since many phages including those that infect common hosts may not share any nucleotide sequence information. In that case, protein sequence data reveals genes that share much older ancestry.

1.4.4 Genomes of Enterobacteria Phage M13 and λ Phages

M13 Enterobacteria phage infects *E. coli*. The genome of M13 phage consists of 6.4 kb single-stranded, (+) sense, circular DNA, which encodes for 10 genes. Unlike most icosahedral virions, the capsid of M13 phage is filamentous, which can be expanded by the addition of further protein subunits. Hence, the genome size can also be increased by the addition of extra sequences in the nonessential intergenic region without becoming incapable of being packaged into the capsid (Fig. 1.3).

In λ phage, the packaging constraints are much more rigid with DNA of ~46–54 kbp of the normal genome size can be packaged into the virus capsid and the substrate packaged into the phage heads during assembly consists of long concatemers of phage DNA that are produced during the later stages of

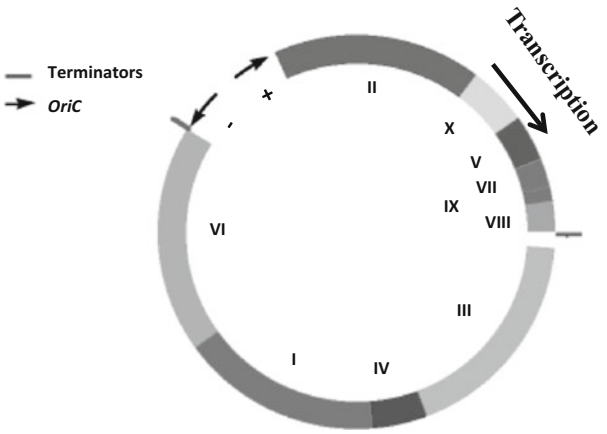

Fig. 1.3 Genome map of M13 phage

vegetative replication. The DNA is apparently reeled into the phage head and after the incorporation of the complete genome, DNA is cleaved at a specific sequence by a phage-coded endonuclease, leaving a 12-bp 5¢ overhang on the end of each of the cleaved strands, known as the *cos* site. Hydrogen bond formation between these 'sticky ends' can result in the formation of a circular molecule (Fig. 1.4).

In a newly infected cell, the gaps on either side of the *cos* site are closed by DNA ligase, and resulting circular DNA undergoes vegetative replication and integration into the bacterial chromosome.

1.4.5 The Genome of T4 Phage

Bacteriophages T2 and T4 are the model organisms playing an instrumental role in the development of modern genetics and molecular biology since the 1940s. They were involved in the development of many salient concepts related to biological sciences, including the recognition of nucleic acids as genetic material, identification of a gene through structural, mutational, recombinational, and functional analyses, in the demonstration triplet genetic code, in the identification of mRNA and establishing the importance of recombination in the replication of DNA, in the light-dependent and light-independent DNA repair mechanisms, restriction and modification of DNA, self-splicing introns in prokaryotes, etc. The main advantage of using T4 phage as a model system is its capability of totally inhibiting its host's gene expression, permitting the investigators to identify the differences between host specific and phage specific macromolecular syntheses. Analysis of the T4 capsid assembly and functioning of its nucleotide-synthesizing complex, replisome, and recombination complexes has led to important insights into

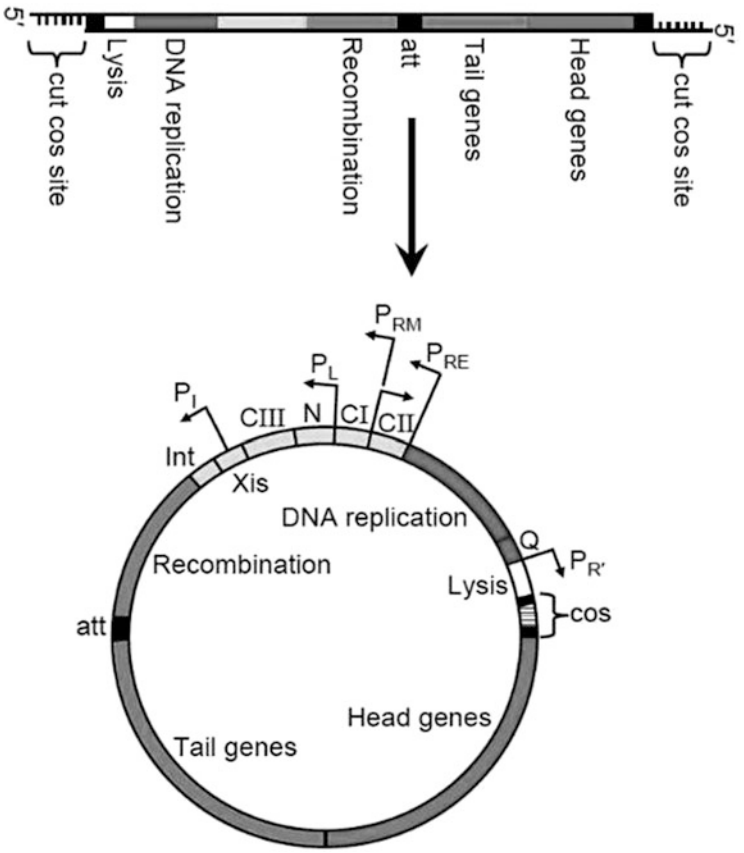

Fig. 1.4 Genetic map of λ phage

macromolecular interactions, substrate channeling, and co-operation between phage and host proteins within such complexes.

The genome of T4 phage is considered as the best avenue for understanding and evaluating the complete genome of a well organized biological system. On the basis of all available information, T4 phage genome comprises of ~300 probable genes, packed into a 168,903 bp genome. This genome comprises 289 expressing genes, 8 tRNA genes, and a minimum of 2 genes that encodes small, stable RNAs with unknown function. Genes 16, 17, and 49 contains multiple coding regions that encode more than one protein. T4 phage genome is four times higher than that of Herpesviruses and yeast, two times higher than *E. coli*. A very small number of genes contains non-coding regions of ~9 kb, accounting for 5.3% of the genome. Regulatory regions in this phage genome are compact, occasionally with overlapping coding regions. Another significant feature of this genome is the overlap of one gene's termination codon with the start codon of the next one. T4

phage has several groups of nested genes. It was found that only 62 genes in this organism are absolutely essential under standard laboratory conditions (rich medium, aeration, 30 to 37 °C). Mutants generated by altering a few other genes produced very small plaques under similar standard laboratory conditions. Many of the 62 essential genes are larger than an average T4 gene, occupying half of the genome. Essential genes encode proteins of the replisome and nucleotide precursor complex, transcriptional regulatory factors, and proteins involved in the structure and assembly of the phage particle. The genome of T4 phage illustrates another rare molecular feature of certain linear viral genomes, terminal redundancy. Replication this phage genome produces long concatemers of DNA, which are cleaved by a specific endonuclease, gets incorporated into the particle with the length exceeding its complete genome due to the repetition of some genes at each end of the genome. Resulting T4 phage genome containing reiterated information is packed into the phage head.

Three T4 phage genes that encode for thymidylate synthase (td), subunit of the aerobic ribonucleotide reductase (nrdB) and the anaerobic ribonucleotide reductase (nrdD) are found to contain introns that are later spliced out of these transcripts. A possibility of an unusual relationship between the nucleic acid sequence and protein sequence occurring through translational bypassing is demonstrated in gene 60 of the T4 phage genome. A 50 bp mRNA segment in the coding region of this gene is not translated by the regular mechanism. This mRNA segment is the only known and unique high-efficiency translational bypass site in the entire T4 phage genome.

DNA in the genome of this phage contains only 34.5% GC, compared with its host genome of *E. coli* consisting of 50% GC. In the genome, 18 of the known or predicted genes containing less than 60% AT and 4 predicted genes have less than 58%. Capsid proteins, which are the most widely conserved among the T4-related phages have the lowest AT contents. Gene 23, which encodes for the major head protein, has the lowest AT content of 55%. A substantial decrease in the pairing of G against C in the coding strands of translated regions has been identified. 4 genes having more than 20% C in the coding strand, while more than 130 genes have more than 20% G and 37 genes have more than 22% G. A and T are equally divided between the coding strands. However, some AT bias has been identified in the T4 phage genome, which is stronger in the third position of codons, as expected in genomes with a high amount of AT-rich regions.

1.5 Positive and Negative Stranded RNA Viral Genomes

The ultimate size of the RNA viral genomes is affected by the fragility of RNA and the tendency of their long strands to break. In addition, RNA genomes tend to have higher mutation rates than those composed of DNA because they are copied less accurately. This tendency might have tended to drive RNA viruses towards smaller genomes. Genomes of RNA viruses encode for a limited number of proteins. RNA viral genomes are broadly divided into double stranded RNA, positive and negative strand single stranded RNAs, monopartite and multipartite RNA viruses. One of the

primary proteins encoded by all these RNA genomes is RNA dependent RNA polymerase, essential for their replication. A major difference between + and − strand ssRNA viruses is that the RNA polymerase can be immediately translated from ss(+) RNA, whereas it is contained in ss(−) RNA. In monopartite ssRNA viruses, the genome encodes a single polyprotein, which is further processed into a number of small molecules, critical for the completion of the viral life cycle. In multipartite ssRNA genomes, each segment encodes for a single gene.

1.5.1 Positive Stranded RNA Viral Genomes

Among the viruses possessing RNA genomes, single stranded plus sense (+) RNA genomes represent an important subgroup including many pathogenic plant, animal and human viruses. These genomes contain cis-acting RNA elements that direct different viral processes, such as protein translation, genome replication, and subgenomic mRNA transcription mRNAs. Single stranded RNA genomes vary in size from coronaviruses (∼30 kb) to those of phages such as MS2 and Qb(∼3.5 kb). Although members of distinct families, most (+) sense RNA viruses share common features in terms of their genomes. Importantly, purified (+) sense virus RNA is capable of infecting the host cells in the absence of any viral proteins.

1.5.1.1 Picornavirus Genome

Picornaviruses are the etiologic agents of numerous diseases with medical and veterinary importance such as Poliomyelitis, common cold, flu, hepatitis, foot-and-mouth disease all are caused by picornaviruses. These viruses have a single-stranded RNA genome of positive polarity in the size of 7200 nucleotides in human rhinoviruses to 8500 nucleotides in foot and mouth disease virus, containing a number of features conserved in all picornaviruses. There is a long 5′ untranslated region (UTR) of 600–1200 nucleotides, which is important for translation, virulence, possibly encapsidation as well as a shorter 3′ UTR of 50–100 nucleotides, necessary for the (−) strand synthesis during replication. 5′ UTR contains a 'clover-leaf' secondary structure known as the internal ribosomal entry site (IRES). The rest of the genome encodes a single polyprotein between 2100 and 2400 amino acids. Both ends of the genome are modified, with the 5′ end by a covalently attached small, basic VPg protein of the 23 amino acids length and the 3′ end by poly(rA) tail. Genome replication occurs in a process that uses the (+) strand as a template for the (−) strand synthesis, which, in turn, is used as a template for the production of an excess of (+) strands. Initiation of both (+) and (−) strand RNA synthesis is thought to be primed by a uridylylated form of VPg, VPg-pUpU.

1.5.1.2 Toga Virus Genome

Togaviridae consists of two genera, alphaviruses, and rubiviruses. Alphavirus genus has 27 members, many of which can be transmitted *via* insect vectors. Rubella virus is the only known member of the rubivirus. The genome of togavirus consists of a single stranded, non-segmented, + sense RNA of ∼11.7 kb. Capsids of these viruses

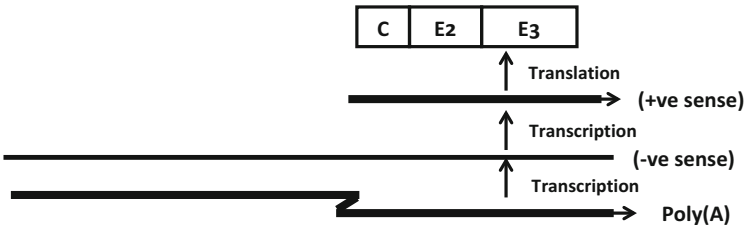

Fig. 1.5 Translation of togavirus subgenomic RNA

are composed of 240 copies of a single capsid protein of ~264 amino acids. The envelope contains 2 virus-encoded glycoproteins, E1 and E2. The genome resembles an mRNA with a cap at the 5′ end and polyadenylation at the 3′ end. The 5′ of the genome encodes the nonstructural proteins required for transcription and replication and the 3′ encodes the structural proteins (capsid, E1, and E2). Translation of the non-structural proteins will be done from genomic RNA, resulting in the synthesis of a polyprotein, which is cleaved into the matured proteins. A subgenomic mRNA is derived from the 3′ end of the genome synthesized using an internal promoter on the (−) strand. Translation of the subgenomic mRNA gives three structural proteins, capsid, E1, and E2 respectively (Fig. 1.5). The capsid protein is synthesized on free cytoplasmic ribosomes. It is cleaved off co-translationally, exposing a signal sequence which directs the ribosome to the ER membrane where the translation of the remaining E1 and E2 proteins gets completed. E1 and E2 are synthesized in association with the rough ER membrane and are then processed through the Golgi apparatus before being transported to the plasma membrane.

1.5.1.3 Flavi Virus Genome

The family *Flaviviridae* consisting of three genera, (1) Flaviviruses (yellow fever virus), (2) Pestiviruses (Bovine viral diarrhea virus) and (3) Hepatitis C Viruses. Many of the Flaviviruses are transmitted *via* insects. These viruses are enveloped and have a + strand genome of 10.5 kb capped at the 5′ and non-polyadenylated 3′ end. 5′ end of the genome encodes for structural proteins and 3′ end encodes for non-structural proteins. The entire viral genome is translated as a single polyprotein, which is cleaved into the mature proteins. No subgenomic RNA is formed.

1.5.1.4 The genome of Coronavirus

The family *Coronaviridae* is comprised of *2 genera*, coronaviruses, and toroviruses. Inclusion of arteriviruses into this family is recent and is not still widely accepted. Coronaviruses are enveloped with a large 27–30 kb + strand RNA genome, which is capped at the 5′ end and is polyadenylated at the 3′ end. Approximately the first 60% of the genome from the 5′ end consists of two overlapping open reading frames (ORF1a and ORF1b) that encode the viral RNA-dependent RNA polymerase, proteases, and other non-structural proteins, along with a 60–80 nucleotide-long "leader" RNA, followed by 200–500 untranslated nucleotides. ORF 1b is translated

via ribosome frame-shifting. The replication or gene expression in these viruses is done by the translation of ORF 1a or 1b to make the viral polymerase and other non-structural proteins, the transcription of (−) strand RNA using the viral polymerase and the synthesis of both full-length viral RNA and subgenomic mRNAs, using (−) strand RNA as template.

All subgenomic mRNAs contain similar leader sequence, derived from the 5′ end of the genome, followed by different regions of the genome, forming a nested set of transcripts, each of them is polyadenylated at the same site. Each of these transcripts is monocistronic and contains a single translation unit, which starts at the first AUG after the leader. The mechanism of the nested transcript synthesis is not completely clear, but it is not *via* RNA splicing since it occurs in the nucleus and coronaviruses replicate in the cytoplasm. Translation of the subgenomic mRNAs yields structural proteins along with some additional non-structural proteins. The new full-length + strands can either be translated or packaged into new virions. These viruses have a unique crown-like structure as identified in the electron microscope. The envelopes contain two viral glycoproteins; M protein (membrane protein), which binds the viral nucleocapsid to the viral envelope during budding and S (spike protein), which facilitates receptor binding and cell fusion. Entire (+) strand is coated with a nucleocapsid protein, directed to intracellular membranes bearing the M protein. The new virions bud through these membranes and are transported to the cell surface through the Golgi smooth-walled vesicles, which then fuse with the plasma membrane, releasing the virus from the cell, without lysis.

1.5.2 Negative Stranded RNA Viral Genomes

The life cycle of negative-strand RNA viruses differs from that of the other RNA viruses in many ways. Specifically, the genome of (−) RNA viruses is not infectious, and infectious virus particles must also deliver their own RNA-dependent RNA polymerase into the infected cell to start the first round of virus-specific mRNA synthesis. These viral genomes are more diverse than the (+) stranded viruses, possibly because of the difficulties in expression. These organisms tend to have large genomes encoding more genetic information. Because of this, most of these viral genomes are segmented. None of these genomes are infectious in its purified RNA. Although the gene encoding for RNA-dependent RNA polymerase has been found in some eukaryotic cells, most of the uninfected cells do not contain enough RNA-dependent RNA polymerase activity to support virus replication. As the (−) stranded RNA genome cannot be converted into mRNA without the activity of viral polymerase packed in each particle, these genomes are effectively inert.

The non-segmented (−) stranded RNA viruses belongs to the order Mononegavirales having linear, single stranded (−) sense RNA as their genetic material. This order includes four families: Rhabdoviridae, Paramyxoviridae, Filoviridae and the Bornaviridae, comprising a wide variety of human, animal and plant pathogens such as rabies virus, measles virus, canine distemper virus, Rinderpest virus and human, bovine respiratory syncytial viruses as well as the lethal Ebola

and Marburg viruses and the recently described Nipah and Hendra virus. These four virus families hold one common factor, that the genetic information in their (−) sense RNA genomes is expressed *via* transcription of a series of discrete monocistronic mRNAs. Transcription originates from a single polymerase entry site near the 3′ end of the genome and is obligatorily sequential, but attenuation occurs specifically at each gene junction, resulting in a progressive reduction in the transcription of genes that are located further from the promoter. This simple method of regulation supports the order of core genes in the Mononegavirales, which are highly conserved among these four families and the genes whose products are required in large amounts are located proximally to the promoter, whereas those needed in catalytic amounts are distally located. The complete genome sequences of almost all genera belonging to all four families of the Mononegavirales have been determined, which ranges in size from the ∼8.9 kb of Borna disease virus to ∼18.9 Kb of Ebola virus genome, which is twice the size and contains 5–10 genes. The genomes of all four virus families contain 4 core genes encoding for a nucleocapsid protein gene (N), phosphoprotein gene (P), matrix protein gene (M) and an RNA-dependent RNA polymerase gene (L). There is a single additional G gene in Rabies virus encoding for the transmembrane attachment and glycoprotein entry. In a few members of this order, three additional genes encoding for transmembrane glycoproteins were also identified. Additionally, a small hydrophobic gene encoding for a protein of unknown function was found in the genus Rubulavirus. Many members of this order encode nonstructural proteins, which are encoded either individually as separate genes or by multiple ORFs within a single gene. The pneumoviral genomes have separate genes encoding for two non-structural proteins involved in evading the host response. These genes are not found in the genomes of avian pneumoviruses. Some rhabdoviruses have a gene between the G and L genes encoding a small nonstructural protein. In the Paramyxovirinae the P gene coding capacity is extended to give rise to a surprising number of polypeptides by utilizing multiple overlapping open reading frames on a single transcript or by co-transcriptional editing. In Pneumovirinae, an M2 gene encoding for an additional transcription factor M2–1 protein has been reported with a second overlapping ORF encoding the M2–2 protein. Some of the viruses are not strictly (−) sense but are ambisense, ie., they are partly (−) sense and partly (+) sense. Ambisense coding strategies occur in both plant viruses (Tospovirusgenus of the bunyaviruses) and animal viruses (the Phlebovirusgenus of bunyaviruses and arenaviruses).

1.5.2.1 Genome of Bunyavirus

Viruses in the family Bunyaviridae possess three distinct linear, single-stranded, negative or ambisense RNA segments in their genome named as small (S), medium (M) and large (L). RNA of the S segment is 0.9 kb, codes for the nucleocapsid protein and a non-structural protein NSs, which interferes with innate immunity. M segment RNA is 5.7 kb, encodes for a polyprotein, eventually giving rise to two glycol proteins Gn and Gc. The large segment (L) RNA is 8.5 kb, codes for the transcriptase, replicase p, and the large RNA-dependent RNA polymerase proteins.

In addition to polymerase activity, Bunyaviral L proteins have an endonuclease activity which cleaves cellular messenger RNAs for the production of capped primers used to initiate transcription of viral messenger RNAs which is known as cap snatching. Exceptionally, in the members of the Phlebovirus and Tospovirus genera, S segment is rather larger than that of M and L. All the three segments of the genome have the same basic structure with the coding region flanked by untranslated regions (UTRs) at the 5′ and 3′ ends. In common with all (−) sense RNAs, the 5′ ends are not capped and the 3′ ends are not polyadenylated. Even though bunyaviruses are categorized into (−) strand viruses, some of the members have genome segments with an ambisense coding strategy, such as Phlebovirus and Tospovirus with 5′ end of each segment is (+) sense, but the 3′ end is (−) sense.

Bunyaviral genomes are remarkably flexible. UTRs of the three segments can be exchanged or can be drastically shortened, the ORFs within an ambisense segment can be swapped around, the genomes can be lengthened through insertions of epitope tags and additional ORFs and the tripartite genome can even be converted into two-segmented or four-segmented versions. The promoters for replication and transcription of the genome segments are located in the terminal sequences of the UTRs that are largely complementary and form panhandle like structures.

1.5.2.2 Genome of Rhabdovirus

Rhabdovirus is a member of a special class of viruses with linear single-strand (−) RNA genome, which is completely non-infectious and is complementary to functional, virus-specific, (+) sense messenger RNAs (mRNAs). These viruses are bullet-shaped, ubiquitous in nature with a uniquely broad and highly diversified range of host system comprising of vertebrates, invertebrates, and plants. Two most popular and frequently studied rhabdoviruses are the animal pathogenic *Vesicular Stomatitis Virus* (VSV) and the human pathogen *Rabies Virus*. Their genomes are non-segmented and are up to 11 kb. There are 60 nucleotides UTR at the 5′ end and a leader region of approximately 50 nucleotides at the 3′ end of the genome. It contains five genes nucleoprotein (N), phosphoprotein (P), matrix protein (M), glycoprotein (G) and polymerase (L). Each gene is terminated with a conserved polyadenylation signal with short intergenic regions between these five genes. There are two structural components in Rhabdoviral genomes: a helical ribonucleoprotein core (RNP) and a surrounding envelope. In the ribonucleoprotein, genomic RNA is tightly encased by the nucleoprotein. Two viral proteins, the phosphoprotein and the large protein (L-protein) are associated with the ribonucleoprotein. The glycoprotein forms approximately 400 trimeric spikes which are tightly arranged on the viral surface. Matrix protein is associated both with the envelope and the ribonucleoprotein and may function as the central protein during the rhabdovirus assembly.

Rabies virus, a member of the genus Lyssavirus in the family Rhabdoviridae, is a neurotropic virus that causes fatal encephalitis in warm-blooded animals. This virus has a non-segmented, single-stranded, negative-sense RNA genome that is approximately 12 kb, comprising of the same five genes nucleoprotein (N), phosphoprotein (P), matrix protein (M), glycoprotein (G) and polymerase (L). encoding for five proteins (Fig. 1.6). Gene encoding for the G protein is known to play a predominant

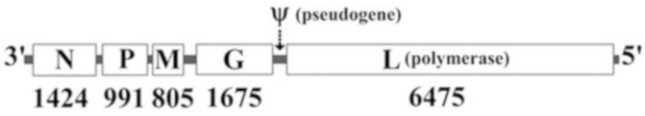

Fig. 1.6 Rabies virus genome

role in the pathogenesis of rabies virus. The genome of this virus is tightly encapsidated by a viral nucleoprotein. RNA-N complex will act as the template for the process of transcription and replication performed by the viral specific RNA-dependent RNA polymerase and its cofactor, phosphoprotein. The domestic dog is a primary reservoir and vector of rabies transmission along with other animal species, such as cat, ferret—Badger, fox, pig, cattle, donkey. An outbreak of pig rabies emerged in a rural pig farm in Sihui of southern China's Guangdong Province in March 2011, resulted in the death of 14 pigs. A virulent wild RABV strain, GD-SH-01, was isolated from the brain tissue of a rabid pig, and its complete genomic nucleotide sequence has been determined recently.

1.6 Segmentation in Viral Genomes

Segmented virus genomes are those that are divided into two or more physically separate molecules of nucleic acid, all of which are then packaged into a single virus particle. There are three viral families Arenaviridae, Bunyaviridae and Orthomyxoviridae containing a (−) sense RNA genome and does not exist as a continuous molecule, but is divided into several segments. These segmented genomes require the presence of RNA-dependent RNA polymerase enzyme to perform the synthesis and replication of mRNA, which reaches the cell along with other viral components during infection.

Segmentation of the viral genome consists of a wide number of advantages and disadvantages. There is an upper limit to the size of a non-segmented virus genome due to the physical properties of nucleic acids, specifically the shearing tendency of long molecules and the maximum length of the nucleic acid chain of each particular virus that can be packed into the capsid. The problem of shearing is particularly relevant for single-stranded RNA, which is much more fragile than the double-stranded DNA. The Physical breakage of the genome results in its biological inactivation, as it cannot be completely transcribed, translated even or replicated. With genome segmentation, there is a risk of shearing by an increase in the total potential coding capacity of the entire genome. Segmentation of the genome enables the virus to generate reassortants. During this process, RNA molecules of different viral strains get reshuffled in doubly infected cells during replication and morphogenesis. As a result, progeny viruses can obtain new combinations of RNA segments and thus gain novel properties. This mechanism, referred to as an antigenic shift, is more common and well-studied in influenza-A viruses. The main disadvantage of the segmentation strategy is that all the individual genome segments must be packed

into each virus particle or the virus will be defective as a result of the loss of genetic information. In general, it is not understood how this control of packaging is achieved.

1.6.1 Influenza Virus

Influenza is the most commonly circulated viruses which cause flu in humans by infecting their respiratory tract. Compared with other respiratory infections such as common cold, flu often causes severe illness. World health organization in February 2014 has divided the influenza viruses into three sections; seasonal influenza, pandemic influenza and zoonotic or variant influenza. Seasonal influenza viruses, categorized into A, B and C type infect humans and cause flu every year, especially in the winter months, spreading from person to person through sneezing, coughing, or touching the contaminated surfaces. These seasonal viruses are capable of mild to severe infection, which might even lead to death, particularly in some high-risk individuals. Type-A influenza viruses are further divided into subtypes according to the specific variety and combinations of two proteins that occur on the surface of the virus, the hemagglutinin (H protein) and the neuraminidase (N protein). Currently, influenza A(H1N1) and A(H3N2) are the prominent seasonal influenza A virus subtypes. There are two type B viruses, which are named as Victoria lineage and Yamagata lineage after the areas where they were first identified. Type C influenza poses much less of a disease burden than influenza A and B causing mild infections, associated with sporadic cases and minor localized outbreaks.

Pandemic influenza occurs when this virus has not previously circulated among humans and to people not having proper immunity. These viruses may emerge, circulate and cause large outbreaks outside of the normal influenza season. As the majority of the population has no immunity to these viruses, a large proportion of a population is getting infected. In 2009, A(H1N1) virus which was not reported before, emerged and spread across the world causing the 2009 H1N1 pandemic. Since then, this virus is established in human populations as a seasonal influenza virus, as discussed above. Currently, there is no pandemic virus reported to be circulating in the world.

Zoonotic or variant influenza are the viruses that are circulating in animals, such as avian influenza virus subtypes A(H5N1) and A(H9N2) and swine influenza virus subtypes A(H1N1) and A(H3N2). Domestic animals such as horses and dogs also have their own influenza viral types. Even though these viruses may be named as the same subtype as viruses found in humans, all of these animal viruses are distinct from human influenza viruses and do not easily transmit to humans. Occasionally, some may infect humans, causing mild conjunctivitis to severe pneumonia and even leads to death. Usually, zoonotic influenza infections are acquired through direct contact with infected animals or contaminated environments, which do not spread rapidly among humans. If such a virus acquires the capacity to spread easily among people either through adaptation or acquisition of certain genes from human viruses, it would be considered an epidemic or pandemic, such as A(H3N2). When viruses of

influenza A(H3N2) circulating in swine also began to infect people of the United States in 2011, they were isolated and labeled as variant (v) in order to distinguish them from human viruses of the same subtype and are named as swine flu, which refers to the influenza viruses infecting swine, and is never used when such viruses infect people. Other animal viruses such as A(H5N1), A(H7N7), A(H7N9), and A (H9N2) infecting humans are named as avian influenza viruses. When animal influenza viruses infect their natural animal host, they are named for that host, like avian influenza viruses, swine influenza viruses, equine influenza viruses, etc.

1.6.1.1 The Genome of Influenza Virus

The genomes of influenza-A and B virus are composed of eight segments and influenza-C is of seven. All segments are single stranded, (−) sense RNA. Proteins that are encoded by each segment of the genome were determined by their electrophoretic mobilities and mutant analysis. The size and the list of proteins encoded by each segment are listed in Table 1.3. All these eight segments have common nucleotide sequences at the 5′ and 3′ ends, necessary for the replication of the genome. These sequences are complementary to one another and the ends of the genome segments are held together by base-pairing to form a panhandle like structure that might be involved in the replication. The genome segments are not packed as naked nucleic acid but in association with the nucleoprotein which is visible in electron micrographs as helical structures, which shows some interaction between the genome segments and could be the reason for the ability of influenza virus particles to pack the genome segments within each viral particle with a low error rate. The exact mechanism of the interaction of genome sequences and proteins involved in this subtle mechanism is yet completely elucidated.

Influenza A virus is responsible for acute and highly contagious viral infection in the respiratory tracts of humans. Infection of this seasonal virus has been reported to cause approximately five million cases of severe illness and 250,000 to 300,000 deaths according to the World Health Organization (WHO). Reassortant influenza A strain arises when influenza virus variants are coincident in an animal population such as avian or swine reservoirs and genes from one variant become incorporated into the other. Zoonotic transmission of avian or swine influenza directly to humans as well as transmission of the reassortant virus has caused notable human pandemics over the last decade. The genome of the influenza A virus (family Orthomyxoviridae) consists of eight single stranded (−) sense RNA molecules spanning approximately 13.5 kilobases (kb). The length of the segments ranges from 890 to 2341 nucleotides and encodes up to 11 proteins (Table 1.3). Haemagglutinin and Neuraminidase genes encode the viral envelope proteins responsible for viral attachment and spreading the infection from cell to cell. The other six viral genes, PB2, PB1, PA, NS, M, and NP encode viral internal proteins required for viral replication and assembly. Among the Influenza A viral proteins, the replicative complex such as polymerases PB2, PB1, and PA together with nucleoprotein play a crucial role in viral infectivity in a different range of host environments. NS gene is found to be important in regulating the host cell response with an additional function of optimizing the viral replication inside the host. The

Table 1.3 Size of influenza virus segments and the proteins they encode

RNA segment	No. of nucleotides	Encoding protein	No. of amino acids
1	2341	Polymerase PB2	759
2	2341	Polymerase PB1	757
3	2233	Polymerase PA	716
4	1778	Haemagglutinin HA	566
5	1565	Nucleoprotein NP	498
6	1413	Neuraminidase NA	454
7	1027	Matrix protein M1	252
		Matrix protein M2	97
8	890	Non-structural protein NS1	230
		Non-structural protein NS2	121

C-terminus of the NS protein has been shown to play a pivotal role in the inhibition of interferon expression.

1.6.1.2 Reassortment of Influenza Viral Genomes

Viral reassortment is a type of genetic recombination, very specific to the segmented RNA viruses. In this process, co-infection of multiple viruses into the same host cell might leads to the shuffling of gene segments to generate a progeny of viruses with novel genome combinations. Reassortment is found to occur in all viral families whose genomes are segmented. But, this reassortment is most prominently studied in influenza viruses as a primary mechanism for interspecies transmission and the emergence of pandemic viral strains. The emergence of new influenza virus genes in humans and their subsequent establishment to cause pandemics have been consistently linked with the reassortment of novel and previously circulating viruses. Reassortment of the segmented viral genomes takes place by the entry of multiple virus particles into a single host cell, followed by the concomitant production of genome segments. Models related to two alternative mechanisms have been proposed for the reassortment.

1. Random packaging model posits the incorporation of the viral RNA into the virions without any discrimination, permitting the likelihood of forming viable reassortants with an entire genome set occurs by chance.
2. Selective packaging model states the packaging of a virus particle into eight unique viral RNA segments through specific packaging signals.

Experimental evaluation of RNA interactions during the assembly of viral segments has revealed an epistatic interaction of packaging signals among viral segments, which might play an imperative role in directing the reassortment. Through the experimental swapping of packaging signals between influenza viruses of different types, Essere et al. were able to overcome the bias observed towards specific genotypes. Baker et al. have shown that the swapping of packaging signals

of two different influenza species has enabled the reassortment of viable particles that have not been naturally observed, indicating the central role of packaging signals in reassortment. Despite the lack of a mechanistic understanding of the function of packaging signals, these observational studies highlight some important implications for viral evolution through epistatic interaction between gene segments and the emergence of novel reassortants.

Influenza virus exhibits a very high level of mixed infections in its hosts, ranging up to 25% of all its infections in avian hosts involving multiple influenza subtypes. Large-scale genomic analysis has identified various levels of restrictions on random reassortment between co-circulating influenza viruses, which differ from the host, subtype and preferential genetic combinations. The highest frequency of influenza reassortment was observed in their natural reservoir, wild aquatic birds, where viruses of different subtypes will frequently exchange their gene segments leading to a tremendous genomic novelty. Reassortment leads to a significant increase in transient amino acid mutations, primarily on the surface glycoprotein hemagglutinin, leading to antigenic change. However, evolutionary change in the reassortment might also be due to the selection pressure induced by herd immunity. The emergence of drug-resistant mutations might have acquired the reassortment, as shown in the emergence of amantadine-resistant H3N2 viruses and oseltamivir-resistant seasonal H1N1 viruses, suggesting that reassortment con-founds the available methods of virus control.

1.7 Multipartite Viral Genomes

Genome segmentation is one of the most common traits in a broad variety of viruses, which is particularly striking in the case of multipartite viruses as their genome segments achieve complete independence at the apparent cost of reducing their infectivity. Multipartite viruses have their genomes fragmented up to eight segments; each of them packed into a separate capsid and contains at least one gene essential for the virus to complete its infection cycle. These viruses require complementation, in order to complete each genomic segment with the rest of segments for producing viral offsprings. The complementation requirements for multipartite viruses have a strong impact on the way they are transmitted. Multiple viral particles may have to enter into each cell so that at least one representative of each segment will be present. The multiplicity of infection (MOI) is a key component in the biology of multipartite viruses. Another significant feature of multipartite viruses is their asymmetry in host distribution. All multipartite viruses identified till date are found to infect only plants and none of these viruses infecting animals have been described. Reason for this asymmetry could be the complementation requirement of the segments, acts as a limiting factor leading to a bottleneck of viral extinction.

1.7.1 The Genome of Gemini Virus

Geminiviruses are characterized by the particles of size 20 × 30 nm, which are twinned, isometric containing circular, single stranded DNA genomes of approximately 2.7 kb. These viruses are classified into two major subgroups based on host plants, insect vector and number of DNA components. One subgroup contains viruses possessing a single DNA component, which infect monocotyledonous plants and are transmitted through leafhoppers. The second sub-group consists of viruses possessing a bipartite genome, which is designated as DNA A and B, infecting dicotyledonous plants and are transmitted by whiteflies. Beet curly top virus (BCTV) demonstrates the characteristics of both the groups. The genomes of Mastrevirus and Curtovirus genera consists of a single stranded (+) sense, DNA of 2.7 kb packaged into the virions. But both (+) and (−) sense strands are found in the infected cells containing protein-coding sequences. The genome of genus Begmovirusis is bipartite consists of two circular, single-stranded DNA molecules of 2.7 kb. Both the strands differ from each other completely in the DNA sequence, except for a common 200 nucleotides non-coding sequence involved in DNA replication. Each strand is packed into entirely separate capsids. The establishment of a productive infection requires both parts of the genome, it is necessary for a minimum of two viral particles bearing at least one copy each of the genome segments for infecting a new host cell. Both DNA strands of the virus found in infected cells contain coding information, present in overlapping open reading frames. Regulating the expression of such a high density of genetic information could be the possible reason for the multipartition of the genomes.

1.8 Evolution of Viral Genomes

Viruses are the most abundant biological entities on this planet. Virology studies show the occurrence of at least 10^{33} viruses on our planet, which is roughly 10 times higher than the bacterial number. They are found everywhere including the oceans, soil, abundantly in plants and in the human body. A healthy human being consists of 10^{13} cells, which harbors 10^{14}–10^{18} bacteria and an unknown number of viruses. These numbers pronounce that the genetic complexity of bacteria is 100 times greater than that of our own genome, making the bacteria as our second genome. With respect to the total genetic information of the human body, we are 99% bacteria, harboring more than 1.5 kg of 1500 different kinds of bacteria in our guts. Viruses form our third genome with over two hundred viruses have been identified in human gut samples based on similarities to known viruses. Archaea and fungi are also present in our guts, making humans a superorganism as well as a complicated ecosystem. Bacteriophages are viruses that can lyse bacteria, which gave them their name. It all started with RNA as the main building block of life on this planet, which is widely accepted today. It is unknown how the RNA has been formed, but it is hypothesized that hydrothermal vents with extreme temperatures of 400 °C to freezing temperatures at the bottom of the oceans might have lead to the

synthesis of RNA with the catalytic support from the clay and energy supplied from the chemical reactions of the rocks composed of metal-rich granite helped life to evolve. RNA has evolved into catalytic oligonucleotides known as ribozymes. This catalytic RNA is capable of cleaving or joining the RNA molecules, which can replicate and mutate. Plant pathogens such as viroids reflect the properties of the early RNA world. Viroids are the ribozymes, which are widespread in plants proficient of damaging them. They look like remnants of a pre-protein world, consisting of non-protein-coding naked RNA without any protein coat. They are small with a few hundred nucleotides of single stranded RNA folded to form a hairpin-loop like structure, which might protect against environmental threats. It has been suggested that a viroid may have entered the human body and evolved into a human virus by acquiring genetic information for a protein from the host such as hepatitis delta virus (HDV), as HDV antigens are related to a human protein and HDV is the only known virus to be a catalytic ribozyme and pathogen in humans.

The next step in the viral evolution might be the plant viruses, such as tobacco mosaic virus (TMV). These viruses are extremely stable with rod-shaped structures. Many of the plant viruses do not synthesize their own replicase, instead use their host cellular RNA replicases, which are considered to be one of the first and oldest polymerases. Almost all plant viruses are small with immense genetic information. The limitation of a simple small RNA molecule to further increase in its length would have made it more unstable. This might have lead to the accumulation of several similar molecules in some kind of protective compartment, as reflected in the present day viruses with segmented genomes. RNA viruses developed their strategies to protect the ends or their RNA with some of them having tRNA like structures, which could fold back and bind to the target RNA for the transfer of individual amino acids. This might be the beginning of peptide synthesis.

The transition of DNA from RNA is a matter of speculation. DNA is much less multifunctional when compared with RNA but is having long-term memory and stabilizes the genetic information. In contrast, RNA is highly fragile and variable. The transition of RNA to DNA occurs in telomeres of chromosomes catalyzed by telomerase, which is a complex of reverse transcriptase, RNA-dependent DNA polymerase, and an endogenous RNA, consisting of sole repeats of a few ribonucleotides. Enzyme closely related to the telomerase is reverse transcriptase (RTase), which copies RNA into DNA, which is the hallmark of retroviruses. RTases are also been isolated from bacteria, which could have been the leftover from ancient retrovirus infections. Alternatively, RTase could also be a precursor of retroviruses. Sequences homologous to RTase have been identified in phage genomes, with the probable function of increasing the diversity of phages by allowing their transmission into the new hosts. Finally, it was the viruses, who invented the DNA, which preceded the formation of the three domains of life and the enzyme RTase appears to be an important evolutionary link between RNA and DNA.

1.9 Conclusions

Viral genomes are considered as one of the most rapidly evolving organisms in biology due to their short replication time and a large amount of the offspring's released from the infected host cell. Viral genomes play a significant role in our increased understanding of epidemiology, diagnosis, surveillance and their evolution. As the basic structure of the virus genomes are highly conserved, the determination of genomes in most of the viruses, understanding their pathogenicity and identification of new viruses is being conveniently done. These genome sequences have been deposited in the databases such as NCBI and Gene Bank and are being maintained. As the new viruses are being identified, the amount of the sequenced viral genomes is going to get increased enormously which might need the invention of next generation sequencing systems and accurate annotation and assemblies.