Viral Genome Organization:
A General Overview

Phil McClean
September 2004

All biological organisms have a genome. As we have seen previously, the genome can be either DNA or RNA. This nucleic acid used to encode functions necessary for it to complete its life cycle and its interaction with its environments. There is great variation in the nature of these genomes. Genome sequencing projects are uncovering many unique features of these that had been previously known.

GENERAL FEATURES

Currently over 4000 viruses have been described. These are classified into 71 families. Although viruses are generally the smallest genomes, as a collection of biological genomes they exhibit the greatest variation. The major difference is that some of the genomes are DNA whereas others are RNA. In addition, both DNA and RNA genomes can be either double- or single-stranded (ds or ss). Finally, some ssRNA viruses use the RNA present in the genome to encode genes. These are called positive-strand ssRNA genomes. Alternatively, negative strand ssRNA genomes must be copied, and the copy (or negative strand) is used for transcription.

DNA and RNA viral genomes have several distinguishing features. First viral genomes can be monopartite or multipartite (Table 1). If they are multipartite, they can have several segments. All dsDNA genomes sequenced to date contain only a single nucleic acid molecule. A few of the ssDNA genomes have multiple segments. In contrast, multipartite genomes are much more frequent for RNA viruses. In particular, the negative strand ssRNA viral genomes are generally multipartite.

DNA viruses tend to be larger in size than RNA viruses (Table 1, Figs. 1 and 2). It is hypothesized that single-stranded virus are smaller because that type of molecule is more fragile than the double stranded molecule. This is generally true for both ssDNA and ssRNA viruses. Some ssDNA genomes can be as small as 1300 nt in length, whereas the minimal genome size of a sequenced ssRNA virus is 2300 nt. Some ssRNA viruses are as large as 31,000 nt, though. The fact that RNA viruses are more susceptible to mutation is thought to have driven these to smaller genome sizes. The largest viruses are the dsDNA viruses, which can be as large as 305,000 nt.

DNA VIRUSES

DNA viruses are typically classified as either ‘small’ or ‘large’ genomes (Fig. 1). As the term implies, the difference between these is the size of the genome. Typical of all viruses, these genomes use nearly all of the sequences for genes that encode proteins. The only difference is the actual number of proteins.

Table 1. General features of sequenced viral genomes. The data was collected from NCBI (http://www.ncbi.nlm.nih.gov/genomes/VIRUSES/viruses.html) and represents the September 27, 2004 release from NCBI.

Viral class (# of segments; range of # protein) **Examples**	**# Completed genomes**	**Size range (nt)**	**# proteins**
dsDNA (1; 5-698)	414		
Bovine polyomavirus		4,697	6
Ectocarpus siliculosus virus		335,593	240
ssDNA (1,2, 8,10,11; 0-15)	230		
Coconut foliar decay virus		1,360	6
Milk vetch dwarf virus		10,958	11
dsRNA (1,2,3,4,10,11,12; 1`-16)	61		
Mycovirus FusoV		3,090	2
Colorado tick fever virus		29,174	13
ssRNA negative strand (1,2,3,4,6,8; 3-12)	81		
Borna disease virus		8,910	5
Rice grassy stunt virus		25,142	6
ssRNA positive strand (1,2,3,4,5; 1-12)	421		
Ophiostoma novo-ulmi mitovirus 6-Ld		2,343	1
Murine hepatitis virus		31,357	11
Retroid (1; 0-10)	83		
Avian carcinoma virus		2,630	1
Human foamy virus		13,242	5
Satellites (1; 0-2)	64		
Rice yellow mottle virus satellite		220	0
Honeysuckle yellow vein mosaic virus-associated DNA beta		1,432	1

SV40 (Fig. 2) is a good example of how a small genome can be extensively utilized. Both strands of the 5243 nt dsDNA contains genes. Five of the six genes overlap with another gene. One genomic region is important for the early events of the development of new SV40 virions. The two early genes are encoded from the negative strand. These genes are actually represented in a single mRNA that is alternatively spliced to encode the small and large T-antigens, proteins critical for launching the replication of the genome. Alternatively splicing is also key for the late genes that encode the structural components of the virus. In this case, VP2 and VP3 are products of the alternative splicing of the same mRNA. Finally, the VP1 gene overlaps the VP2/VP3 gene region. This feature of maximixing genomic information through gene clustering is a common feature of 'small' genome viruses.

Another feature that distinguishes the 'small' and 'large' genomes is the manner in which they are replicated. The 'small' genomes use a host DNA polymerase for replication. The

genome typically encodes proteins that prepare the DNA for replication. In contrast, the 'large' genomes encode a DNA polymerase that is responsible for the genomes's replication.

As the term implies, 'large' genome DNA viruses can also encode many more proteins. Table 2 and Figure 1 both show the breadth of genes that can be encoded. The example here is the human adenovirus A. In general, the genome organization is less complex. The plus strand encodes all but five of the genes. Another feature is that few of the genes share sequences with another gene.

Table 2. Human adenovirus A genes and gene locations.

Genes	**nt location**
E1A (HAdVAgp01)	503-1099
E1B, small T-antigen (HAdVAgp02)	1542-2033
E1B large T-antigen (HAdVAgp03)	1847-3395
Hexon-associated protein (HAdVAgp04)	3374-3808
Maturation protein (HAdVAgp05)	3844-5202
DNA polymerase (HAdVAgp06)	4953-8138
DNA binding protein (HAdVAgp07)	7602-8219
DNA terminal protein (HAdVAgp08)	8312-10131 (complement)
52 KD protein (HAdVAgp09)	10428-11549
Hexon-associated protein (HAdVAgp10)	11570-13318
Penton protein (HAdVAgp11)	13394-14887
Minor core protein (HAdVAgp12)	15500-16543
11 KD core protein (HAdVAgp13)	16568-16786
Protein VI (HAdVAgp14)	16843-17640
Hexon protein, Late protein 2 (HAdVAgp15)	17740-20499
Virus encoded endoprotease, Late protein 3 (HAdVAgp16)	20525-21145
Early E2A (HAdVAgp17)	21215-22669 (complement)
Late 100 KD protein (HAdVAgp18)	22695-25043
33 KD phophoprotein fragment (HAdVAgp19)	25202-25558
Hexon-associated protein precursor (HAdVAgp20)	25612-26313
E3 12.1 KD protein (HAdVAgp21)	26313-26630
E3B 10.4 KD protein (HAdVAgp22)	28207-28482
E3B 14.5 KD protein (HAdVAgp23)	28479-28811
E3B 14.7 KD protein (HAdVAgp24)	28804-29190
Fiber protein (HAdVAgp25)	29368-31131
Early E4 17 KD protein (HAdVAgp26)	31183-31407
Early E4 34 KD protein (HAdVAgp27)	31436-32311 (complement)
E4 13 KD protein (HAdVAgp28)	32244-32606 (complement)
E4 11 KD protein (HAdVAgp29)	32613-32963 (complement)

RNA VIRUSES

In general, genomes of RNA viruses encode a limited number of proteins. One protein that is very often encoded by these genomes is a RNA-dependent RNA polymerase (RdRp). These polymerases are essential for the replication of both positive and negative strand ssRNAs. as well as dsRNAs. This is true for both monopartite and multipartite RNA viruses that show a range of 1-13 proteins. A major difference between + and – strand ssRNA viruses is that the polymersase is contained within ss (-) virion whereas it is immediately translated from the RNA of the ss(+) RNA.

As with DNA viruses, the genome is an example of maximizing it size. For monopartite ssRNA viruses, the genome encodes a single polyprotein. This protein is then processed into a number of small molecules, each of them critical for the completion of the life cycle of the virus. In multipartite ssRNA genomes, each segment normally contains a single gene. Examples of the two human viruses are depicted in Fig. 3.

Retroviruses provide another example of how viruses utilize their minimal sizes. All retroviruses contain *gag*, *pol*, and *env* genes that, respectively, encode structural proteins, the reverse transcriptase, and proteins embedded in the viral coat. These genes generally exist as polyproteins that are processed into several protein products. The reverse transcriptase is the most unique gene for retroviruses. It converts RNA into a DNA copy. This DNA is then used to replicate the genome.

In addition, to this basic gene set, retroviruses can encode other genes (Fig. 3). The simplest case is the addition of a single gene. Oncogenic retroviruses contain the *gag/pol*/*env* genes, but also contain a fourth gene called the oncogene. For the Rous Sarcoma Virus, this additional gene is the *src* gene that encodes a tyrosine kinase that modifies protein by adding phosphate groups to tyrosine residues. Many oncogenic retroviruses encode a unique protein that is tied to the cell division process. These genes are similar to other genes found in the host. The retrovirus version, though, is mutated relative to the host gene. This mutation general leads to uncontrolled cell growth, a phenotypic feature of cancer cells.

Human immunodeficiency virus I is the causative agent of AIDS (Fig. 4). As with all other retroviruses it contains the *gag/pol*/*env* suite of genes. These genes are also expressed as polyproteins that are processed. The additional genes affect other processes including viral infectivity (*vif*), transcription activation (tat), replication (*vpr*, *vpu*, *nef*), and regulation of virion protein expression (*rev*). HIV 1 is an example of a retrovirus that has accumulated multiple genes beyond the basic set. Therefore it is a good model of retroviral evolution.

Table 3. Human immunodeficiency virus 1 control regions, genes, mature proteins and genetic locations.

Control region or gene **Mature protein(s)**	**Location (nt)**
polyA signal	73-78
5' UTR	97-181
Primer binding	182-199
Gag-pol gene (HIV1gp1) Gag-pol transframe peptide (p6) Pol (unprocessed Pol polyprotein) Protease Reverse transcriptase Reverse transcriptase p51 subunit Integrase	336-4642 1632-1637, 1637-1798 1655-4639 1799-2095 2096-3775 2096-3415 3776-4639
Gag (HIV1gp2) Matrix (p17) Capsid (p24) p2 Nucleocapsid (p7) p1 p6	336-1838 336-731 732-1424 1425-1466 1467-1631 1632-1679 1680-1835
Vif (HIV1gp3) Vif (p23)	4587-5165 4587-5165
Vpr (HIVgp4) Vpr (p15)	5105-5396 5105-5396
Tat (HIV1gp5) Tat (p14)	5377-7970 5377-5591,7925-7970 (spliced)
Rev (HIV1gp6) Rev (p19)	5516-8199 5516-5592, 7925-8199 (spliced
Vpu (HIV1gp4) Vpu (p16)	5608-5856 5608-5856
Env (HIV1gp8) Envelope polyprotein Env signal peptide Envelope surface glycoprotein (gp120) Envelope transmembrane glycoprotein (gp41)	5771-8341 5771-8341 5771-5854 5855-7303 7304-8338
Nef (HIVgp9) Nef (p27)	8343-8963 8343-8963
3' UTR	8631-9085

Viral Genome Organization

All biological organisms have a genome

- The genome can be either DNA or RNA
- Encode functions necessary
    - to complete its life cycle
    - interact with the environments

Variation

- Common feature when comparing genomes

Genome sequencing projects

- Uncovering many unique features
- These were previously known.

General Features of Viruses

Over 4000 viruses have been described

- Classified into 71 taxa
- Some are smaller than a ribosome

Smallest genomes

- But exhibit great variation

Major classifications

- DNA vs RNA

Sub classifications

- Single-stranded vs. double-stranded

Segments

- Monopartitie vs. multipartite

ssRNA virus classifications

- RNA strand found in the viron
    - positive (+) strand
        - most multipartite
    - negative (-) strand
        - many are multipartite

Virus replication by

- DNA viruses
    - DNA polymerase
        - Small genomes
            - Host encoded
        - Large genomes
            - Viral encoded
- RNA viruses
    - RNA-dependent RNA polymerase
    - Reverse transcriptiase (retroviruses)

Genome sizes

- DNA viruses larger than RNA viruses

ss viruses small than ds viruses

- Hypothesis
- ss nucleic acids more fragile than ds nucleic
- drove evolution toward smaller ss genomes

Examples:

ssDNA viral genomes

- As small as 1300 nt in length

ssRNA viral genomes

- Smallest sequenced genome = 2300 nt
- Largest = 31,000 nt

Mutations

- RNA is more susceptible to mutation during transcription than DNA during replication
- Another driving force to small RNA viral genomes

Largest genome

- DNA virus
- *Paramecium bursaria* Chlorella virus 1
    - dsDNA
    - 305,107 nt
    - 698 proteins

Table 1. General features of sequenced viral genomes. The data was collected from NCBI (http://www.ncbi.nlm.nih.gov/genomes/VIRUSES/viruses.html) and represents the September 27, 2004 release from NCBI.

Viral class (# of segments; range of # protein) **Examples**	**# Completed genomes**	**Size range (nt)**	**# proteins**
dsDNA (1; 5-698)	414		
Bovine polyomavirus		4,697	6
Ectocarpus siliculosus virus		335,593	240
ssDNA (1,2, 8,10,11; 0-15)	230		
Coconut foliar decay virus		1,360	6
Milk vetch dwarf virus		10,958	11
dsRNA (1,2,3,4,10,11,12; 1`-16)	61		
Mycovirus FusoV		3,090	2
Colorado tick fever virus		29,174	13
ssRNA negative strand (1,2,3,4,6,8; 3-12)	81		
Borna disease virus		8,910	5
Rice grassy stunt virus		25,142	6
ssRNA positive strand (1,2,3,4,5; 1-12)	421		
Ophiostoma novo-ulmi mitovirus 6-Ld		2,343	1
Murine hepatitis virus		31,357	11
Retroid (1; 0-10)	83		
Avian carcinoma virus		2,630	1
Human foamy virus		13,242	5
Satellites (1; 0-2)	64		
Rice yellow mottle virus satellite		220	0
Honeysuckle yellow vein mosaic virus-associated DNA beta		1,432	1

DNA Viruses

Classified as

- ‘small’ genome
- ‘large’ genomes

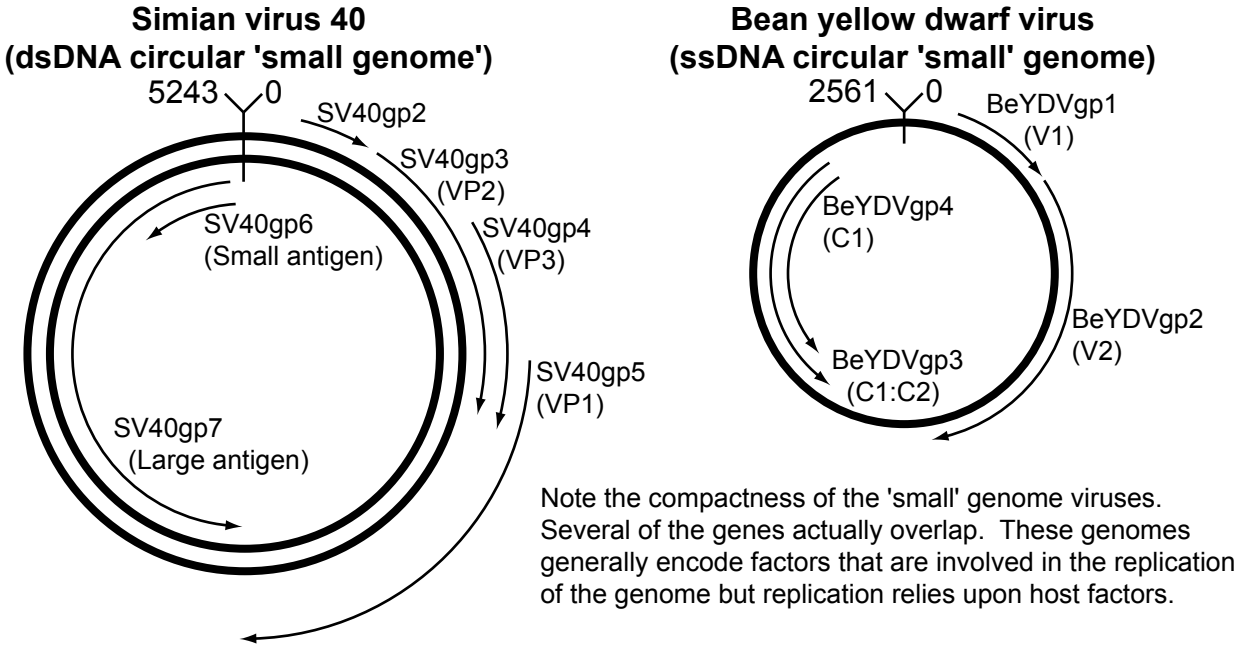
Typical observation of all viruses

- Nearly all DNA used for genes
- Different genomes differ in the number of proteins encoded

DNA Viral Genomes

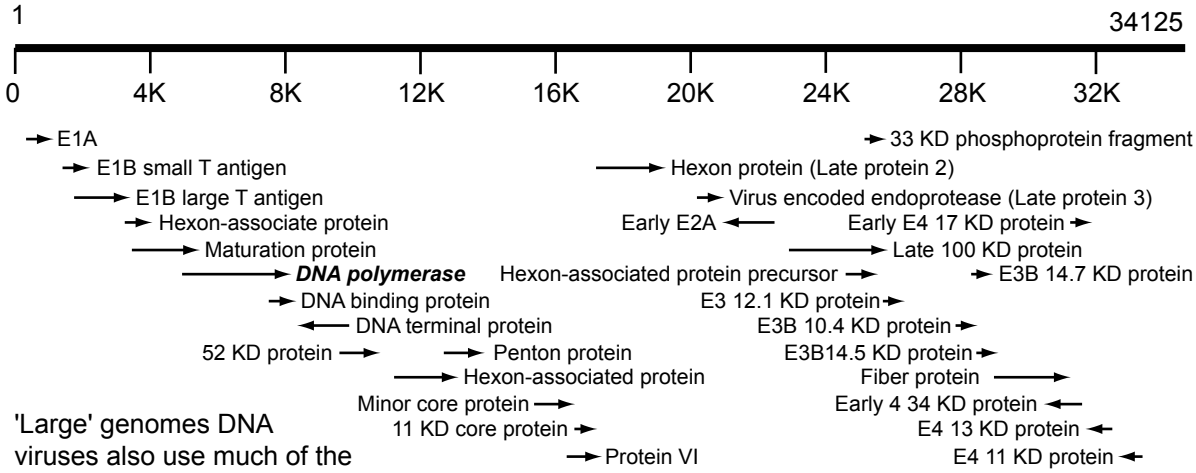
A. The Concept

DNA viruses are a major class of this biological entity. The viruses can be either ***double- or single-stranded***. In general, the single stranded genomes are smaller than those that are double-stranded. Among the double-stranded genomes, these can either have ***'small' or 'large' genomes***. One major difference between the two genomes is the mechanism of DNA replication. Small genomes use ***host polymerase activities***, whereas large genomes ***encode a DNA polymerase***.



Note the compactness of the 'small' genome viruses. Several of the genes actually overlap. These genomes generally encode factors that are involved in the replication of the genome but replication relies upon host factors.

Human adenovirus A (dsDNA linear 'large genome')



'Large' genomes DNA viruses also use much of the genome. A major difference is that these genomes do not have extensive overlaps between the genes. These genomes also encode a DNA polymerase (italicized above) that is used for genome replication.

Figure 1. Organization of DNA viral genomes.

Clustering is a common feature of 'small' genome viruses

Simian Virus 40

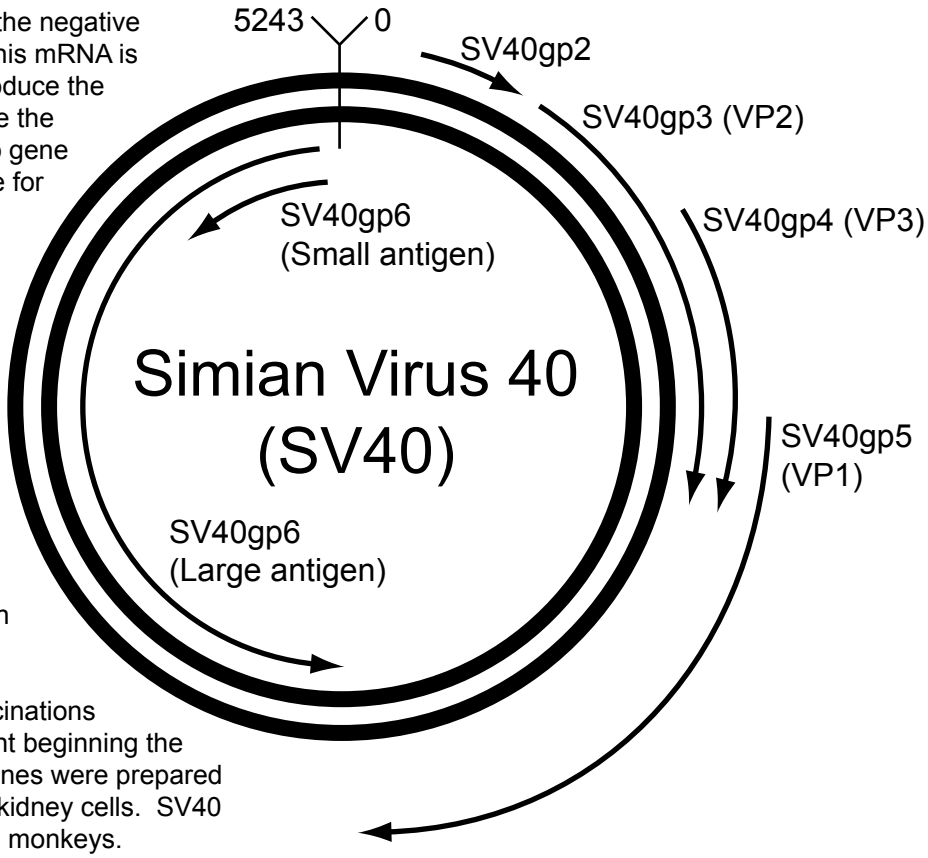
- Good example of a small genome
- Shows how a small genome can be extensively utilized
- Features:
    - 5243 nt dsDNA genome
        - Both strands contains genes
            - Five of the six genes overlap
        - "Life-cycle specific" regions
            - Early genes
                - Negative strand genes important for early development of new virions
                - Genes transcribed in a single mRNA
                - mRNA is alternatively spliced
                - Encode the small and large T-antigens
                - Proteins critical replication of the genome
            - Late genes
                - Encoded on the positive strand
                - Two genes (VP2 and VP3) overlap
                - A single mRNA for these genes
                - Alternative splicing produces unique mRNAs
                - Proteins critical for the structure of the virion

DNA Viruses: The SV40 Example

A. The Concept

DNA viruses genomes range from very small to very (1,360 nt) to very large (305,107 nt). These genomes encode a few (one) to many proteins (698). They also show a complex pattern of gene organization and gene expression. A good example is the ds DNA Simian Virus 40. This monkey pathogen has a small genome (5243 nt), encodes seven protein products, cotains overlapping genes, some of which are the result of alternate splicing.

1. Early gene expression. The first SV40 genes to be expressed are SV40pg6 and SV40gp7. Thes encolde the small and large T antigens, respectively. A single mRNA is transcribed counterclockwise using the negative strand as a template. This mRNA is alternately spliced to produce the mRNAs used to translate the two proteins. These two gene products are responsible for unwinding the DNA and making it ready for replication.



2. Late gene expression. The products of the SV40gp3, SV50gp4, and SV40gp5 genes produce the three proteins found in the viral capsid. These genes encode the VP2, VP3, and VP1 proteins, respecitvely. Alternate splcing of a single mRNA produces the mRNA used for VP2 and VP3 translation.SV40pg5 overlaps the coding region for SV40gp3 and SV40gp4. The transcription of these genes is in a clockwise orientation using the positive strand.

3. Note of interest. SV40 has received much attention because of its assocaition with polio vaccinations. Polio vaccinations were a major social event beginning the mid 1950s. These vaccines were prepared using Rhesus monkeys kidney cells. SV40 is a pathogen of Rhesus monkeys. Subsequents studies showed that vaccines developed from these cells were contaminated with SV40. This was of concern because SV40 can cause cancer. From 1955-1963, between 10 and 30 million individuals recevied the contaminated vaccines. These indiiduals may be infected with the virus.

Figure 2. Organization of SV40 viral genomes.

Large vs small DNA genomes

- Major difference
    - Small genomes
        - Use host factors for replicatioin
    - Large genomes
        - Encode a DNA polymerase

Large genomes

- More genes encode more proteins
- Genome organization is less complex
- Overlap of genes is less frequent

Human adenovirus A

- Genome size
    - dsDNA
- 34,125 nt
    - Number of genes
- 29 genes
    - Overlapping genes
        - Five
    - Complementary strand genes
        - Five

Table 2. Human adenovirus A genes and gene locations.

Genes	**nt location**
E1A (HAdVAgp01)	503-1099
E1B, small T-antigen (HAdVAgp02)	1542-2033
E1B, large T-antigen (HAdVAgp03)	1847-3395
Hexon-associated protein (HAdVAgp04)	3374-3808
Maturation protein (HAdVAgp05)	3844-5202
DNA polymerase (HAdVAgp06)	4953-8138
DNA binding protein (HAdVAgp07)	7602-8219
DNA terminal protein (HAdVAgp08)	8312-10131 (complement)
52 KD protein (HAdVAgp09)	10428-11549
Hexon-associated protein (HAdVAgp10)	11570-13318
Penton protein (HAdVAgp11)	13394-14887
Minor core protein (HAdVAgp12)	15500-16543
11 KD core protein (HAdVAgp13)	16568-16786
Protein VI (HAdVAgp14)	16843-17640
Hexon protein, Late protein 2 (HAdVAgp15)	17740-20499
Virus encoded endoprotease, Late protein 3 (HAdVAgp16)	20525-21145
Early E2A (HAdVAgp17)	21215-22669 (complement)
Late 100 KD protein (HAdVAgp18)	22695-25043
33 KD phophoprotein fragment (HAdVAgp19)	25202-25558
Hexon-associated protein precursor (HAdVAgp20)	25612-26313
E3 12.1 KD protein (HAdVAgp21)	26313-26630
E3B 10.4 KD protein (HAdVAgp22)	28207-28482
E3B 14.5 KD protein (HAdVAgp23)	28479-28811
E3B 14.7 KD protein (HAdVAgp24)	28804-29190
Fiber protein (HAdVAgp25)	29368-31131
Early E4 17 KD protein (HAdVAgp26)	31183-31407
Early E4 34 KD protein (HAdVAgp27)	31436-32311 (complement)
E4 13 KD protein (HAdVAgp28)	32244-32606 (complement)
E4 11 KD protein (HAdVAgp29)	32613-32963 (complement)

RNA Viruses

General Features

- Minimal Genome Size
- Encode a limited number of proteins
    - Often encodes a RNA-dependent RNA polymerase (RdRp)
        - Essential for the replication
        - Both positive and negative strand ssRNAs use such an enzyme
        - A function of dsRNA genomes also
        - A gene in both monopartite and multipartite genomes
- Number of proteins
    - Range from 1-13 proteins
- + vs – strand ssRNA
    - Polymersase is contained within ss (-) virion
    - Polymerase immediately translated from the RNA of the ss(+) RNA

Monopartite ssRNA viruses

- Genome can encode a single polyprotein
- Processed into a number of small molecules
- Each critical to complete the life cycle of the virus

Multipartite ssRNA Viruses

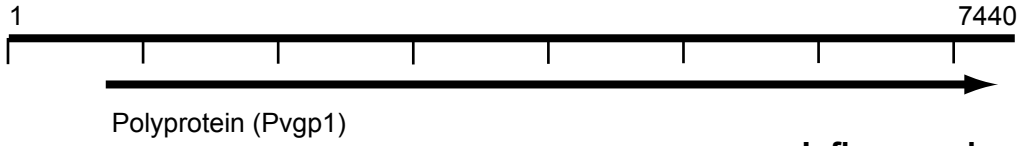
- Each segment generally contains a single gene

RNA Viral Genomes

A. The Concept

RNA viral genomes can be either ***single- or double-stranded***. In addition, these can be ***multipartite***, meaning they consist of several RNA molecules. The ssRNA molecules are also classified as ***positive- or negative-strand or retrovirues***. The + and - strand ssRNA genomes are replicated by a RNA-dependent RNA polymerase that is encoded by their genomes. Retroviruses are replicated as DNA following the conversion of the RNA into DNA by a ***reverse transcriptase***.

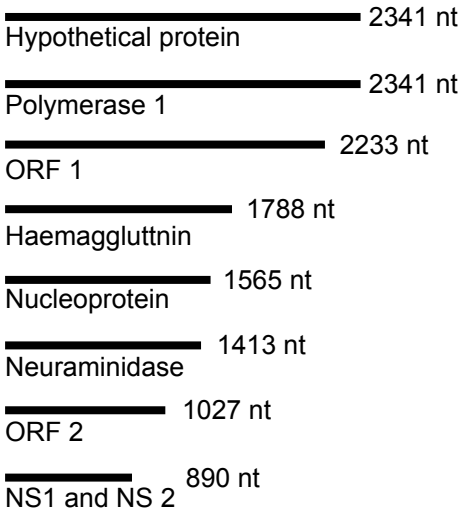
Poliovirus A (monopartite positive strand ssRNA)



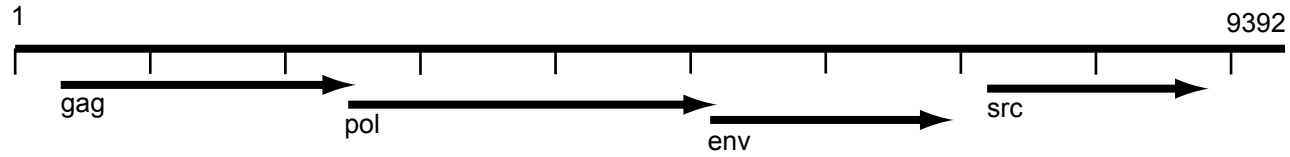
ssRNA viruses are compact. Monopartitie genomes generally have only a few genes. This masks the actually coding capacity. The Poliovirus A mRNA is used to translate a single **polyprotein**. This polyprotein is then cleaved into the 11 proteins necessary for the function of the virus. Multipartite ssRNA genomes partion the protein genes to several RNAs as seen here for Influenza Virus A. Both of these viruses encode a **RNA-dependent RNA polymerase**.

Retroviruses are also ssRNA viruses. The basic gene set for these viruses is *gag* (that encode the structural proteins, *pol* that (encodes the reverse transcriptase), and *env* (proteins that attach to the viron surface). In addtion, they can encode other genes. Oncogenes, such as Rous Sarcoma Virus *src* gene, can induce the cancerous state. Many of the retroviruses genes also encode polyproteins. The eight HIV I genes actually encode 22 different protiens. Retroviruses are replcated via a DNA intermediate. The reverse transcriptase creates a DNA copy of the genome that is used for replication purposes.

Influenza virus A (multipartite negative strand ssRNA)



Rous sarcoma virus (retrovirus)



Human immunodeficiency virus I (retrovirus)

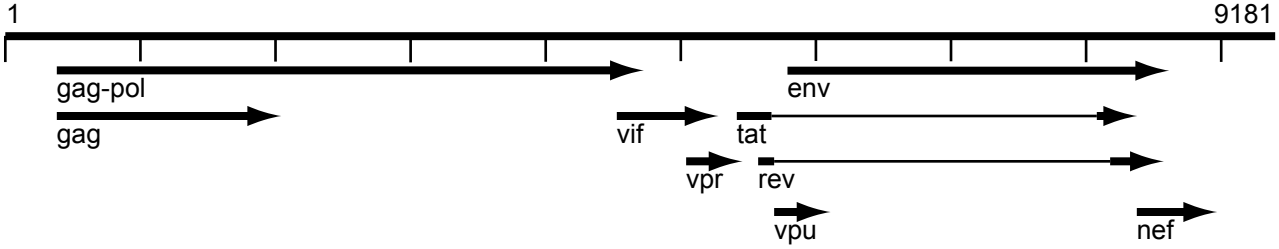


Figure 3. Organization of RNA viral genomes.

Retroviruses

- Minimal genome sizes
- Basic gene set
    - *Gag*
        - encode structural proteins
    - *Pol*
        - reverse transcriptase
    - *Env*
        - proteins embedded in the viral coat

Reverse transcriptase

- Converts RNA into a DNA copy
- Used to replicate the genome

Other Retroviral Genes

- Oncogenic retroviruses
    - Rous sarcoma virus
        - Fourth gene
            - Tyrosine kinase
            - Viral gene a mutated version of host gene
            - Gene causes uncontrolled cell growth
- Oncogenes generally are involved in cell growth and division

Human Immunodeficiency Virus I

- Causative agent of AIDS
- Contains the *gag/pol*/*env* suite of genes
- Example of a retrovirus that accumulated multiple genes
- A good model of retroviral evolution
- Genes expressed as polyproteins that are processed
- Additional genes
    - Affect other processes
        - Viral infectivity (*vif*)
        - Transcription activation (*tat*)
        - Replication (*vp*r, *vpu*, *nef*)
        - Regulation of virion protein expression (*rev*)

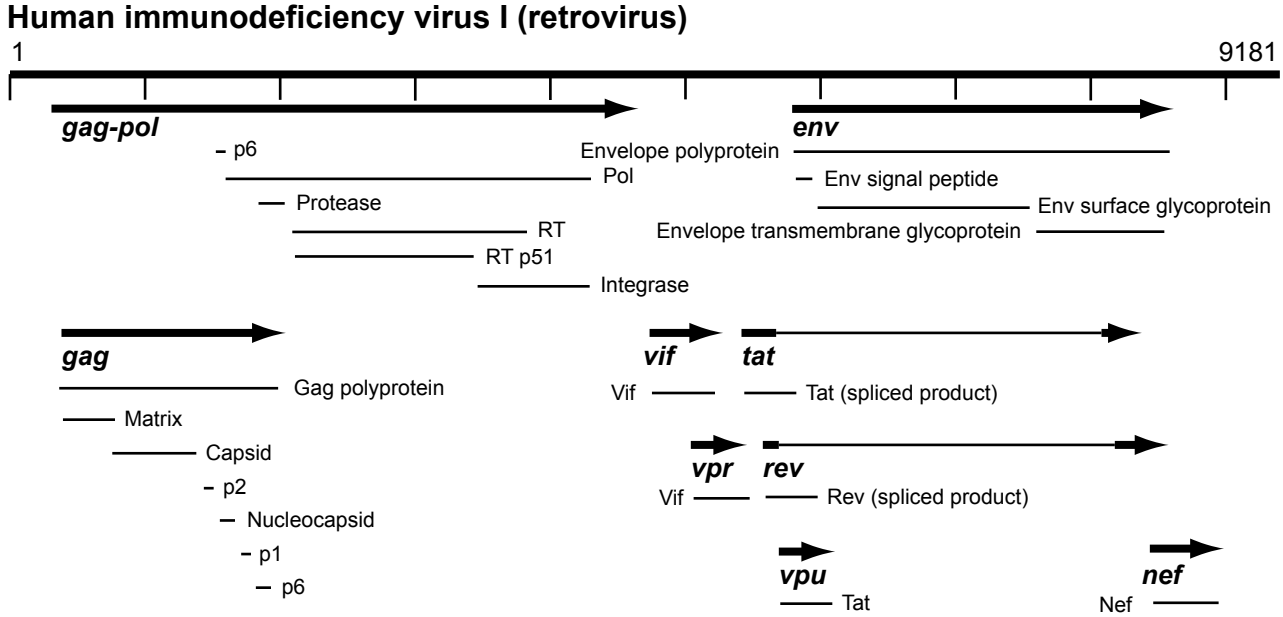
Table 3. Human immunodeficiency virus 1 control regions, genes, mature proteins and genetic locations.

Control region or gene **Mature protein(s)**	**Location (nt)**
polyA signal	73-78
5' UTR	97-181
Primer binding	182-199
Gag-pol gene (HIV1gp1) Gag-pol transframe peptide (p6) Pol (unprocessed Pol polyprotein) Protease Reverse transcriptase Reverse transcriptase p51 subunit Integrase	336-4642 1632-1637, 1637-1798 1655-4639 1799-2095 2096-3775 2096-3415 3776-4639
Gag (HIV1gp2) Matrix (p17) Capsid (p24) p2 Nucleocapsid (p7) p1 p6	336-1838 336-731 732-1424 1425-1466 1467-1631 1632-1679 1680-1835
Vif (HIV1gp3) Vif (p23)	4587-5165 4587-5165
Vpr (HIVgp4) Vpr (p15)	5105-5396 5105-5396
Tat (HIV1gp5) Tat (p14)	5377-7970 5377-5591,7925-7970 (spliced)
Rev (HIV1gp6) Rev (p19)	5516-8199 5516-5592, 7925-8199 (spliced
Vpu (HIV1gp4) Vpu (p16)	5608-5856 5608-5856
Env (HIV1gp8) Envelope polyprotein Env signal peptide Envelope surface glycoprotein (gp120) Envelope transmembrane glycoprotein (gp41)	5771-8341 5771-8341 5771-5854 5855-7303 7304-8338
Nef (HIVgp9) Nef (p27)	8343-8963 8343-8963
3' UTR	8631-9085

HIV I Genome and Proteins

A. The Concept

Viruses pack a lot of genetic information into a small amount of genomic space. A good example is the human immunodeficiency virus I. HIV I is the causal agent of AIDS. The small virus has eight gene which encode 22 proteins.



The HIV I virus is a good example of how a small genome can encode a large amount of genetic information. These eight HIV I genes encode 22 different genes. Three of the genes, gag-pol, gag, and env each encode a polyprotein that is processed into a collection proteins. It is relatively common for the *gag-pol*, *gag*, and *env* genes of retroviruses to encode multiple proteins. The most important protein for replication of the genome, the reverse transcriptase (RT), is part of the Pol polyprotein. The *gag-pol* polyprotein is cleaved by the protease encoded by the *gag-pol* gene. p2 cleaves the *gag* polyprotein.

One way to make the most from a limited size genome is to used the same sequence for multiple genes. Four of the genes, *ta*t, *rev*, *vpu*,and *nef* each share sequences with the *env* gene.

HIV 1 also has a feature in common with other retroviruses. It contains the required gag, pol and genes. It has functions,though, are defined by a series of other genes. These genes are necessary for viral infectivity (*vif*), transcription activation (*ta*t), replication (*vpr*, *vpu*, *nef*), and regulating virion protein expression (*rev*),

Figure 4. The genes and proteins of the human immunodeficiency virus I genome.