Viral Genome Organization

General Features of Viruses

Over ~14,000 viruses sequences (mostly whole genome) have been described

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

Smallest genomes in biological world

- But exhibit great variation

Major classifications

- DNA vs RNA

Segmental classifications

- Monopartitie vs. multipartite
- One vs multiple nucleic acids

Sub classifications

- Single-stranded vs. double-stranded

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 transcriptase (retroviruses)***

Genome sizes

- DNA viruses larger than RNA viruses

ss viruses smaller than ds viruses

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

Why RNA Viruses are Smaller

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

"Viral-Related" Genomes

- ***Virus Satellite***
  - Require helper virus for replication
  - Only a single-stranded RNA or DNA
  - Only encode proteins that coat the nucleic acid
    - Rice Yellow Mottle Virus Satellite
      - 220 nt in length
- ***Viroid***
  - Naked, single-stranded circular RNA
  - Only a single stranded RNA acid
  - Do not encode proteins
    - Coconut cadang-cadang viroid
      - 229 nt in length

Smallest Viral Genome

- Cocoa necrosis virus
  - Linear, single-stranded DNA genome
  - 229 nucleotides
    - No envelope
  - One CDS (**C**o**D**ing **S**equence ~ proteins)

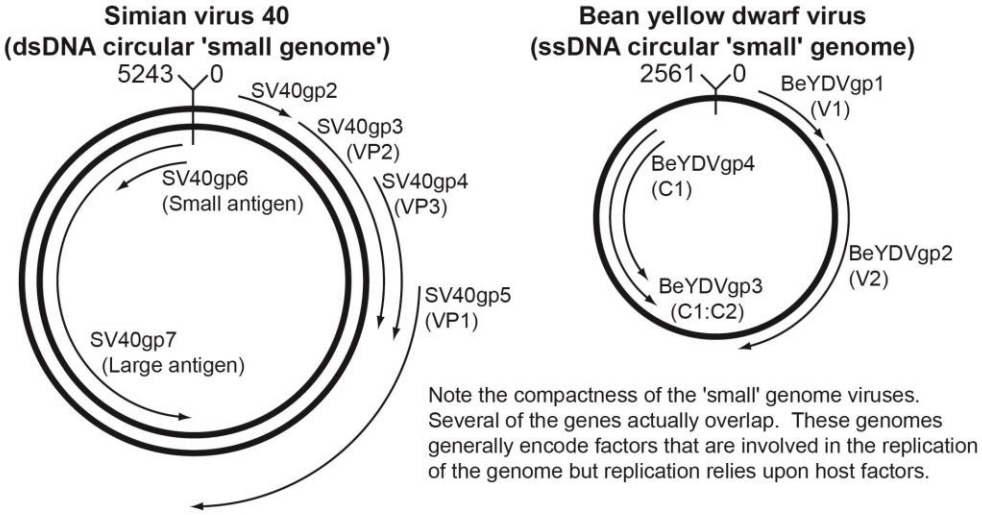
Largest Viral Genome

- Pandoravirus salinus
  - Member of Megaviridae
    - Originally mistaken as bacteria
      - Infect amoebae living in sediment
  - DNA virus
    - 2,473,870 nt (2.48 Mb)
    - 1,430 CDS

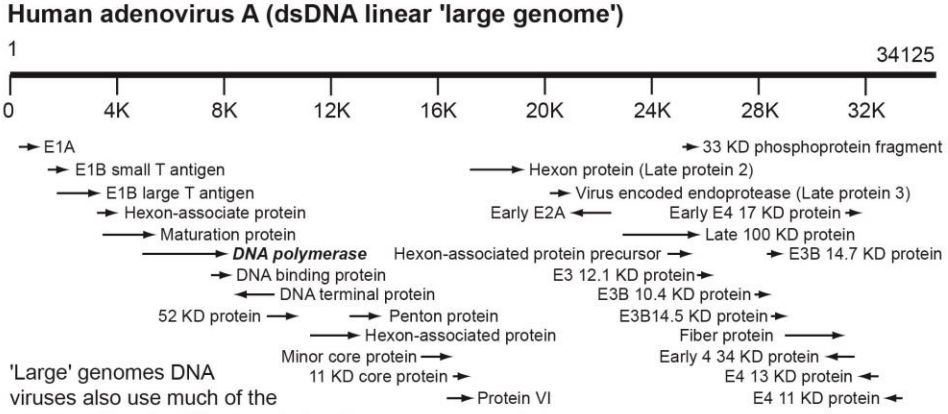
DNA Viral Genomes

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.



'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

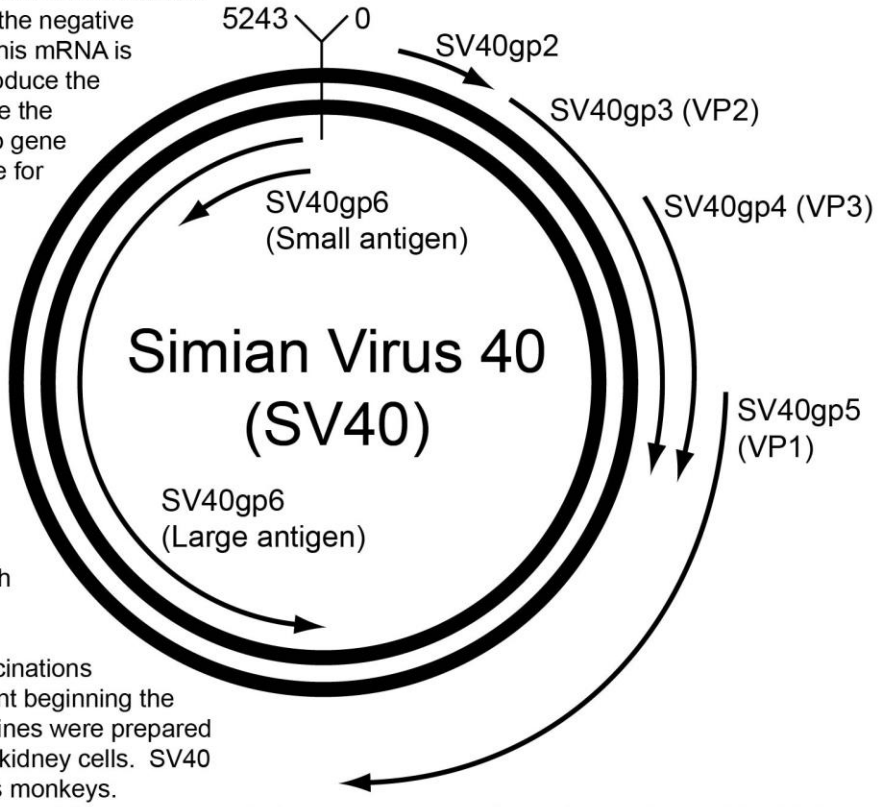
- Good example of a small genome
- Shows how a small genome can be extensively utilized
- Features:
    - 5243 nt dsDNA genome
        - Both strands contains genes
            - Five of the six genes overlap
        - "Life-cycle specific" regions
            - Early genes
                - Negative 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 replication
    - 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

RNA Viruses

General Features

- ***Minimal Genome Size***
- ***Encode a limited number of proteins***
    - **Often encodes a RNA-dependent RNA polymerase (RdRp)**
        - Found in positive and negative strand ssRNAs genomes
            - Essential for the replication
        - A function of dsRNA genomes also with
            - Encoded 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

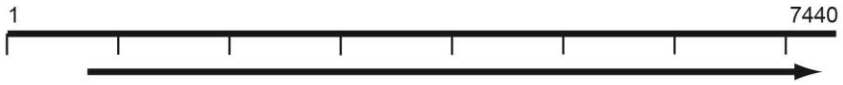
- Each segment generally contains a single gene

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***.

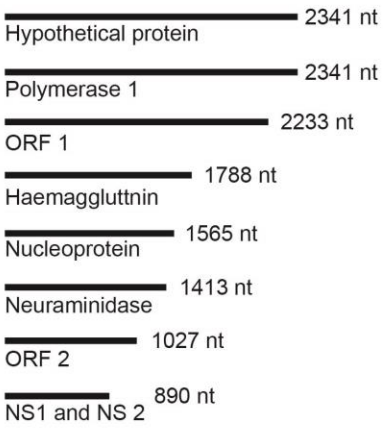
Poliovirus A (monopartite positive strand ssRNA)



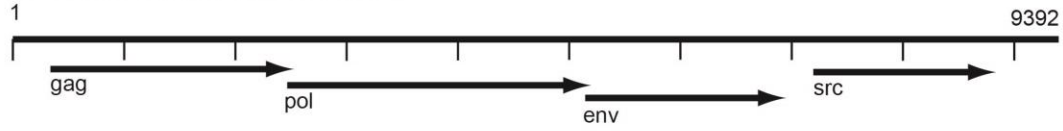
ssRNA viruses are compact. Monopartitie genomes generally have only a few genes. This masks the actually coding capacity. The Poliovirus A mRNA is used to translate a single **polyprotein**. This polyprotein is then cleaved into the 11 proteins necessary for the function of the virus. Multipartite ssRNA 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)

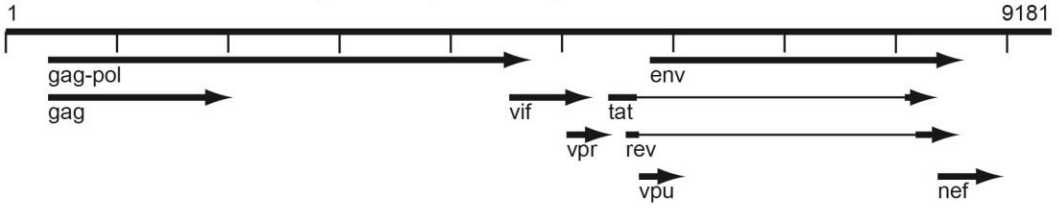


Figure 3. Organization of RNA viral genomes.

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
            - Cancer causing 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***
    - Protein product cleaved into individual proteins
  - Additional HIV genes
    - ***Affect other processes***
      - Viral infectivity (*vif*)
      - Transcription activation (*tat*)
      - Replication (*vpr*, *vpu*, *nef*)
      - Regulation of virion protein expression (*rev*)

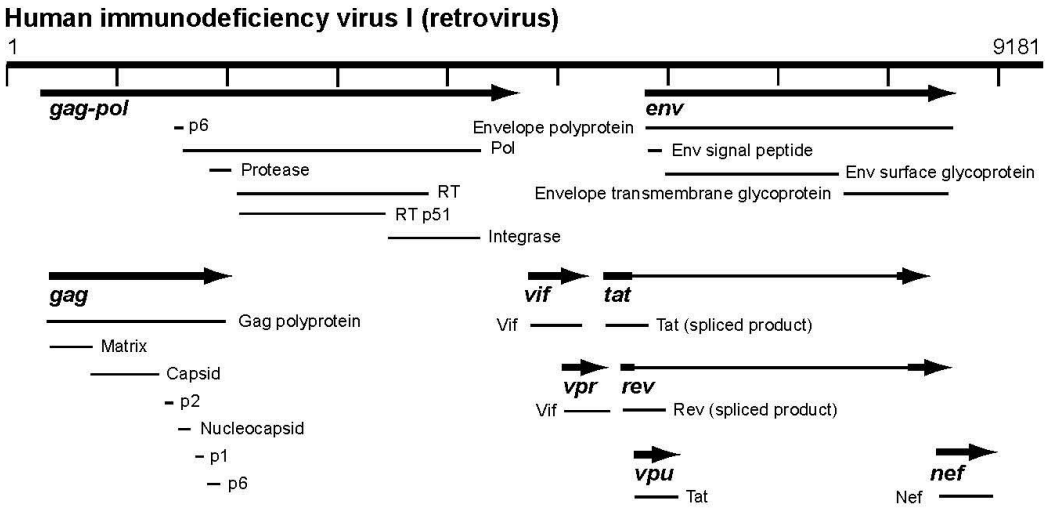
Table 3. Human immunodeficiency virus 1 control regions, genes, mature proteins and genetic locations.

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, *tat*, *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 (*tat*), replication (*vpr*, *vpu*, *nef*), and regulating virion protein expression (*rev*),

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