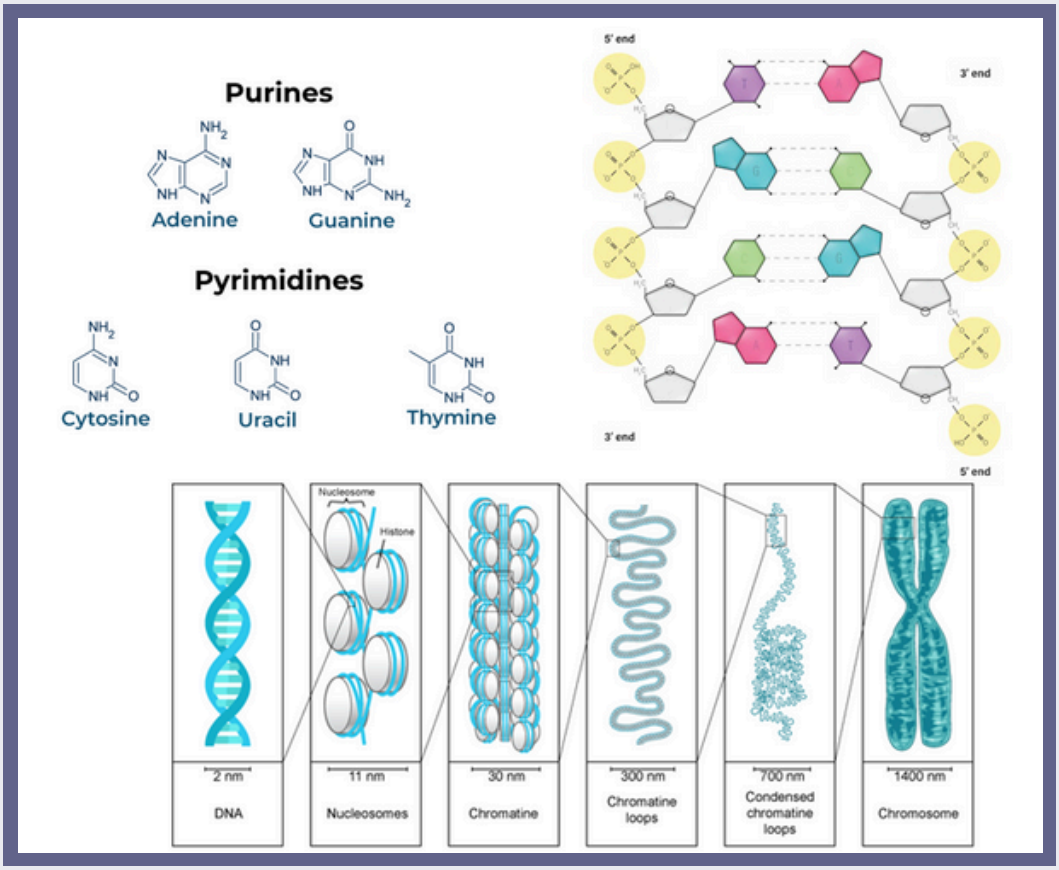
Pathogen Genomics

I. General Genomics Primer

To begin, we will review concepts in genetics and molecular biology as they relate to viruses and bacteria. Deoxyribonucleic Acid (DNA) contains nucleic acids composed of pyrimidine bases (cytosine and thymine) and purine bases (adenine and guanine). These nucleotides are bound together with phosphodiester bonds along the backbone, and two DNA strands are held together with hydrogen bonds. In a cell's nucleus, DNA is wrapped tightly around proteins called histones which assemble into larger structures called chromosomes.



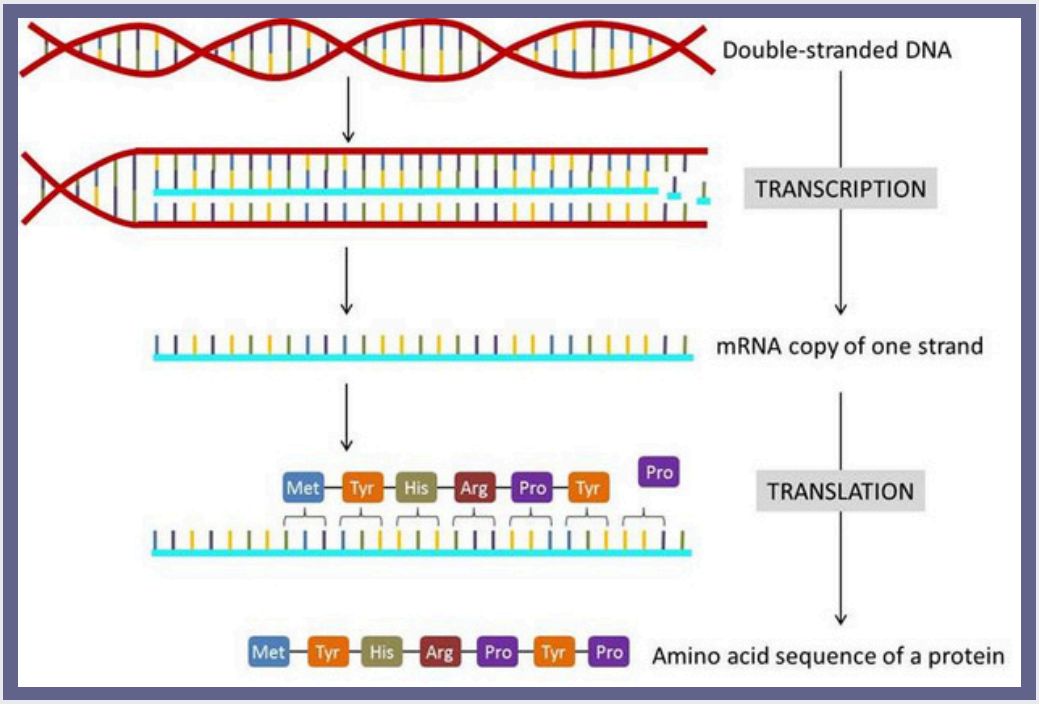
There are different kinds of genomic variation that can occur. The most common is nucleotide variation. These are often called single nucleotide poolymorphisms (SNPs) and are mutations in the genome leading to a change in the nucleotide composition of the DNA sequence. There are several types of mutations which cause nucleotide variation:

- **Point mutations:** mutations that are single base pair substitutions
  - Missense mutation: an amino acid to be replaced with a similar amino acid
  - Nonsense mutation: a regular codon to be replaced with a stop codon

    - Silent mutation: a change in codon but to the same amino acid
- **Insertion:** an addition of one or more nucleotides into the sequence
- **Deletion**: a removal of one or more nucleotides into the sequence
- **Inversion:** when a segment of DNA is reversed end to end
- **Amplification:** when a segment of DNA is duplicated
- **Translocation:** when a segment of DNA is removed and attached to a different location in the sequence

Another kind is copy-number variation where there are large structural variations in the genome that lead to different copies of sections of DNA. There are also tandem repeats, where short sequences of DNA are repeated one after the other several times. Transposons also occur where short segments of DNA can jump around the genome and can cause mutations or structural alterations to the sequence.

Genes are expressed via transcription of DNA, and the subsequent translation of mRNA. To begin, DNA is copied, or transcribed into a messenger RNA molecule (mRNA). This mRNA is then translated in three-nucleotide chunks (codons) which code for amino acids. These amino acids are then assembles into cellular proteins set off to perform whatever duty they are designed for.



II. Bacterial Genomics

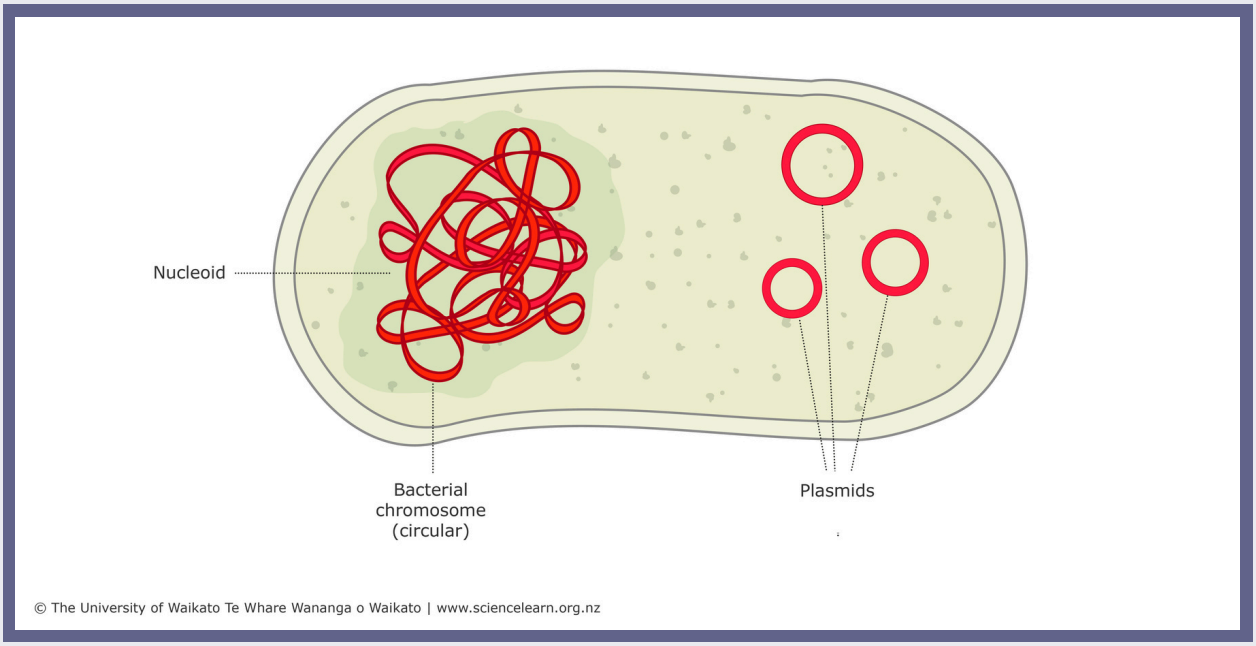
Defining Terms

Bacterial genomes are different from eukaryotic genomes. Here, we define some important terms to help our understanding:

- **Open Reading Frame (ORF):** A sequence of DNA or NRA that could be translated into a polypeptide.
- **Operon:** A cluster of genes that are transcribed into a single RNA and under the control over a single regulatory site.

Bacterial genomes typically consist of one circular chromosome in addition to plasmids. **Plasmids** are linear or circular DNA strands that are much smaller than chromosomes and have the ability to replicate independently. They can also be transferred from one bacterium to another

The chromosome is integrated into a **nucleoid** which is an area of the cell that contains most of the bacterial genome, but it is not delimited. The DNA is wrapped tightly by nucleoid-associated proteins which are similar to histones.

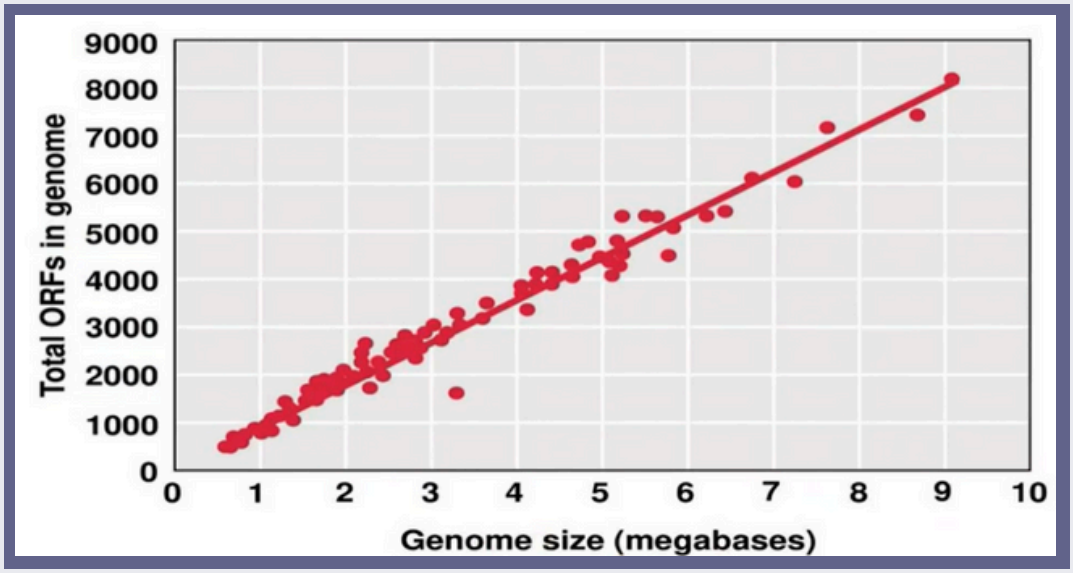


Structure and Function

There are several assumptions that can be made about bacterial genomes:

Bacterial genomes vary in size and number of open reading frames.

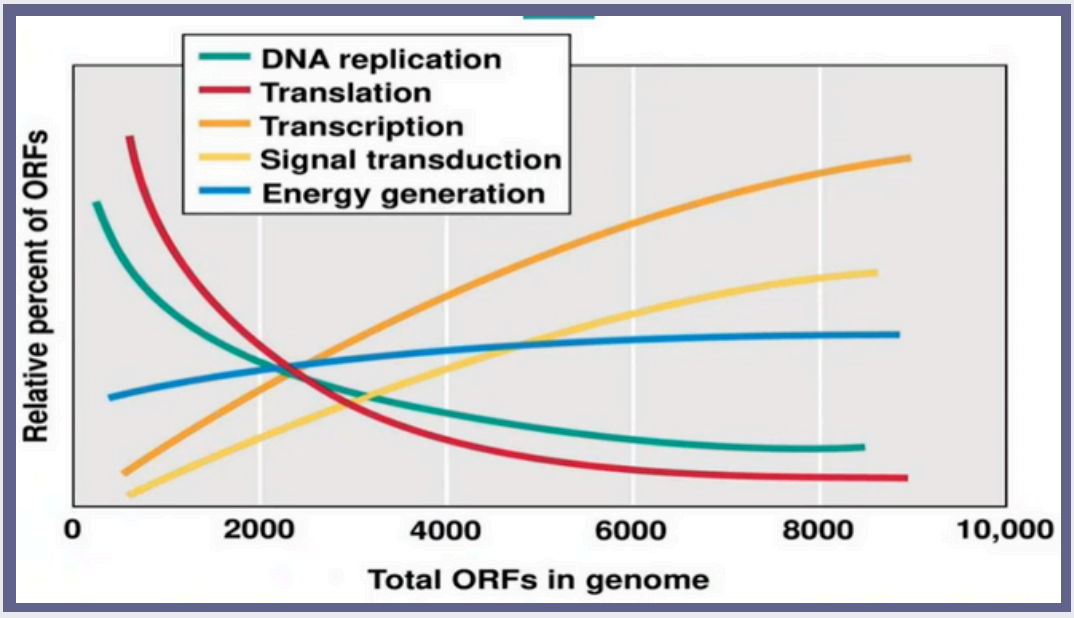
Generally, as the size of the genome increases, the number of ORFs increases as well, indicating that larger genomes encode more proteins. Larger genomes are typically seen in bacterial that need to adapt to more conditions or are more free living bacteria.



Bacterial genomes contain essential genes with conserved function and have limited non-coding regions.

Essential genes include genes required for cellular function. Non-coding regions include regulatory elements, mobile genetic elements (plasmids, phages, transposons), and repetitive elements. Below are further descriptions of these genomic elements:

- **Prophage:** When the genome of a temperate virus is replicating synchronously with the bacterial host.
- **Transposon:** A transposable element that carries genes to other parts of the genome, often genes that confer selectable phenotypes.
- **Chromosomal Island:** A region of a bacterial chromosome of foreign origin that contains genes for extra properties (like virulence).
- **Pathogenicity Island:** A region of a bacterial chromosome of foreign origin that contains genes for virulence.



Bacterial genomes have organized structure and arrangement which allows for gene regulation and coordinated control of gene expression.

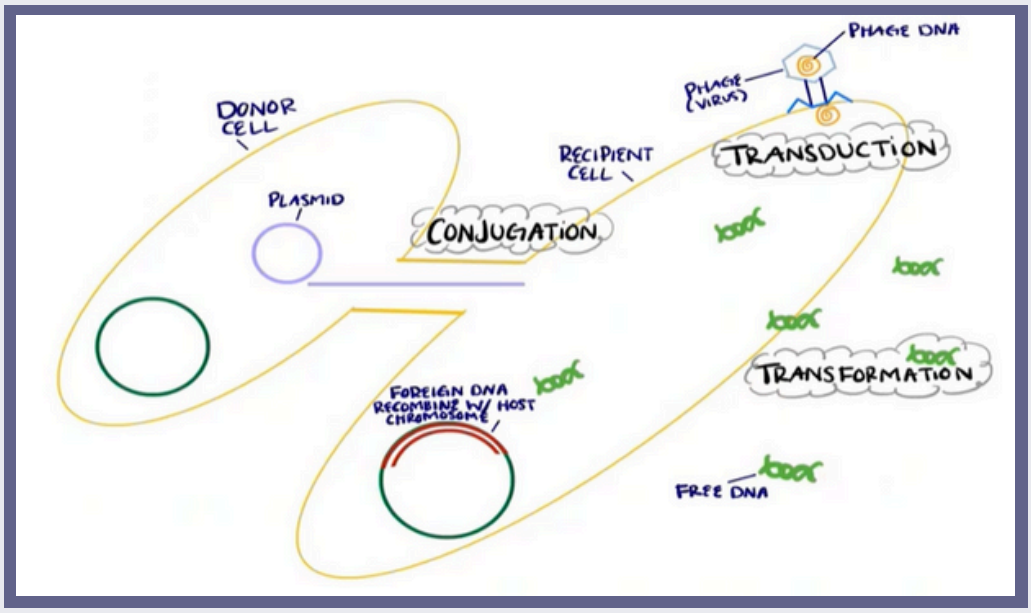
Since it is energetically costly to make proteins that are not always in use, bacterial genomes contain regions for signal transduction to regulate gene expression.

Bacteria undergo replication and cell division.

Bacteria can perform horizontal gene transfer.

This contributes the most diversity to the genome and can happen in three different ways:

- **Conjugation:** Genetic material transferred between bacteria through cell-to-cell contact.
- **Transduction:** When genetic material is injected into a bacterial cell via a bacteriophage.
- **Transformation:** When free floating genetic material is taken up by a bacterial cell.



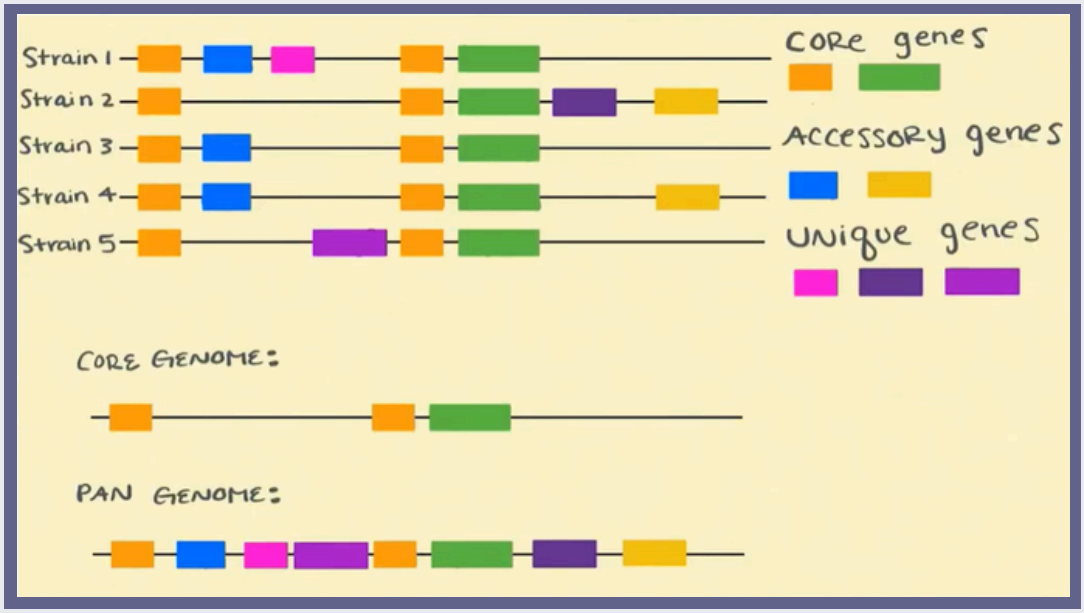
Once DNA enters the cell, it may be degraded by restriction enzymes, it may replicate itself if possible, or it may recombine with the host genome. There are two different methods of recombination that can occur:

- **Nonhomologous Recombination:** Recombination between two DNA sequences that are not necessarily similar (transposition, site specific recombination)
- **Homologous Recombination:** Recombination between DNA molecules with the same sequence in the crossover region.

Bacterial genomes are plastic (can undergo changes) and heterogeneous (contain diversity and variation).

This variation can be represented in different ways:

- **Core Genome:** The set of genes present in all strains of a bacterial species.
- **Accessory Genome:** Genes that may not be present in all strains, but are present in come individual bacteria.
- **Pan Genome:** The complete set of genes found within a bacterial species, composed of the core and accessory genomes.

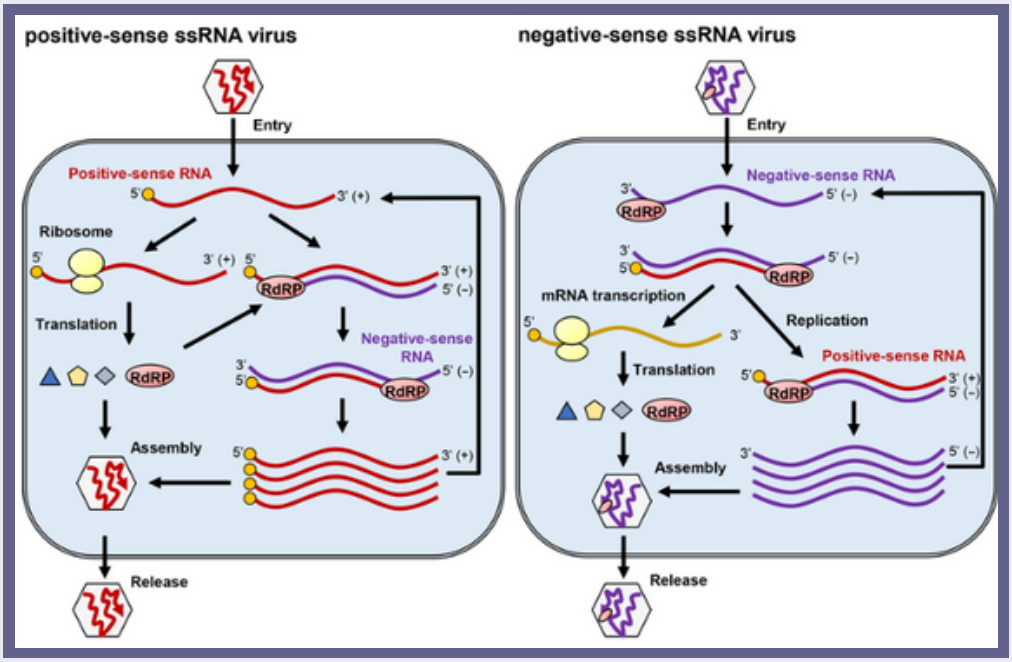


The size of the core and pan genome can vary between species. Additionally, the size of the core genome is expected to decrease as evolutionary distance between strains increases while the pan genome is expected to increase with each new genome sequenced.

II. Viral Genomics

Defining Terms

Understanding polarity is critical to understanding viral genomics.



- Ribosome ready mRNA is always + strand
- Equivalent polarity of DNA is also + strand
- The compliments of RNA and DNA + strands are - strands. Viral genomes needs to make mRNA that can be read by host ribosomes

Structure and Function

There are several assumptions that can be made about viral genomes:

Viral genomes are composed of genetic material that differs in complexity and structure

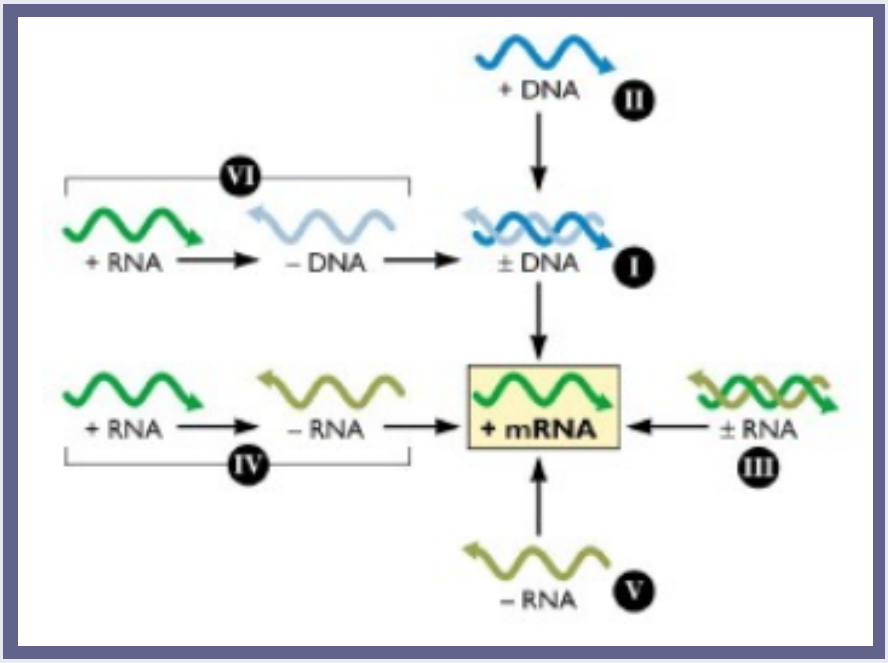
Viruses can either be DNA or RNA, single stranded or double stranded, linear, circular, segmented, or gapped. They also have the potential adopt secondary and tertiary structures to conduct different molecular interactions.

Viral genomes can be classified by virus replication strategy

Based on the Baltimore classification scheme, there are seven different kinds of viral genomes.

- dsDNA
- gapped dsDNA
- ssDNA
- dsRNA
- ss -RNA
- ss +RNA

- ss +RNA with DNA intermediate
- gapped DNA *Hepadnaviridae*



Viral genomes vary in length and size

Viral genomes range from <2kb to <2500k

Viral genomes encode proteins essential for replication assembly as well as proteins required for host entry

Viruses do not replicate the same way eukaryotic cells or bacterial cells do, they must build new viral units which then exit cells to continue the infection. Part of this means that viruses with RNA genomes must encode an RNA-dependent RNA polymerase (RdRp) or a reverse transcriptase to replicate its genome.

Viral genomes are essentially a "parts list" and tells us little about how the virus interacts with its host

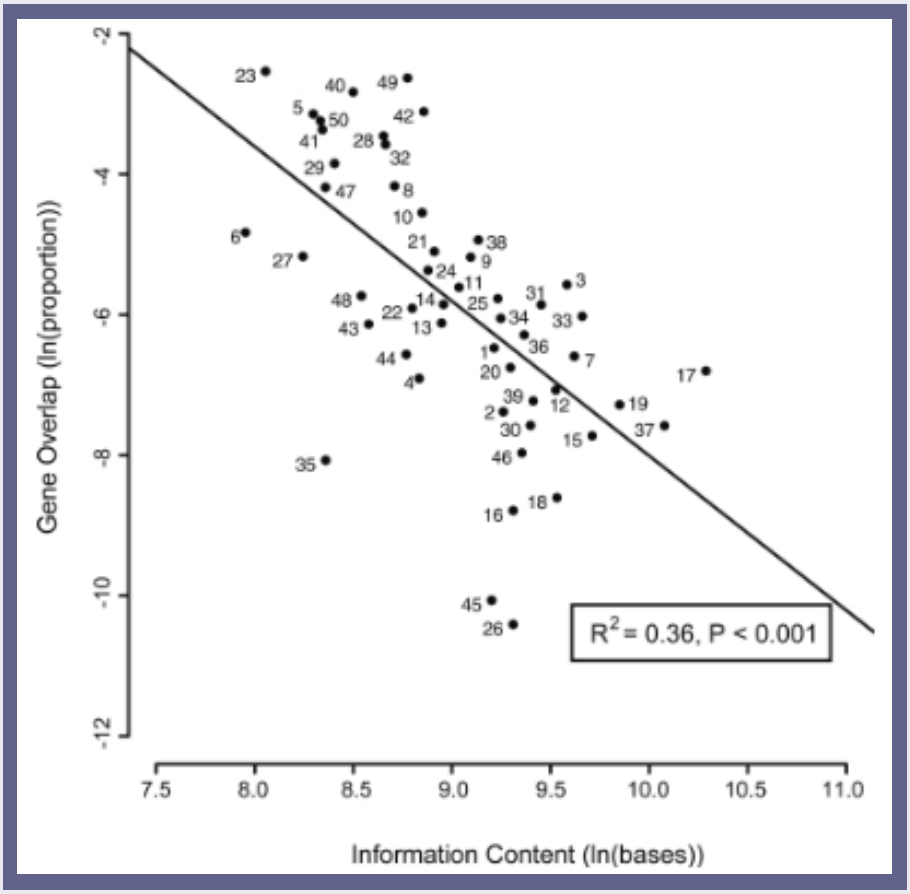
Various things are encoded in the viral genome, including proteins to contribute to:

- Genome replication
- Genome expression
- Assembly and packaging of the genome
- Regulation of replication
- Modulation of host defenses

This means that the viral genome does not tell us a lot about how a virus interacts with the host. The viral genome does not encode translation proteins or proteins involved in metabolism, which the virus must obtain from host cells. The viral genome also does not have telomeres.

Viral genomes are compact and efficient: usually only carry essential genetic information

Information encoded in viral genomes is optimized for compaction. The smaller the genome, the greater the compression of genetic information.

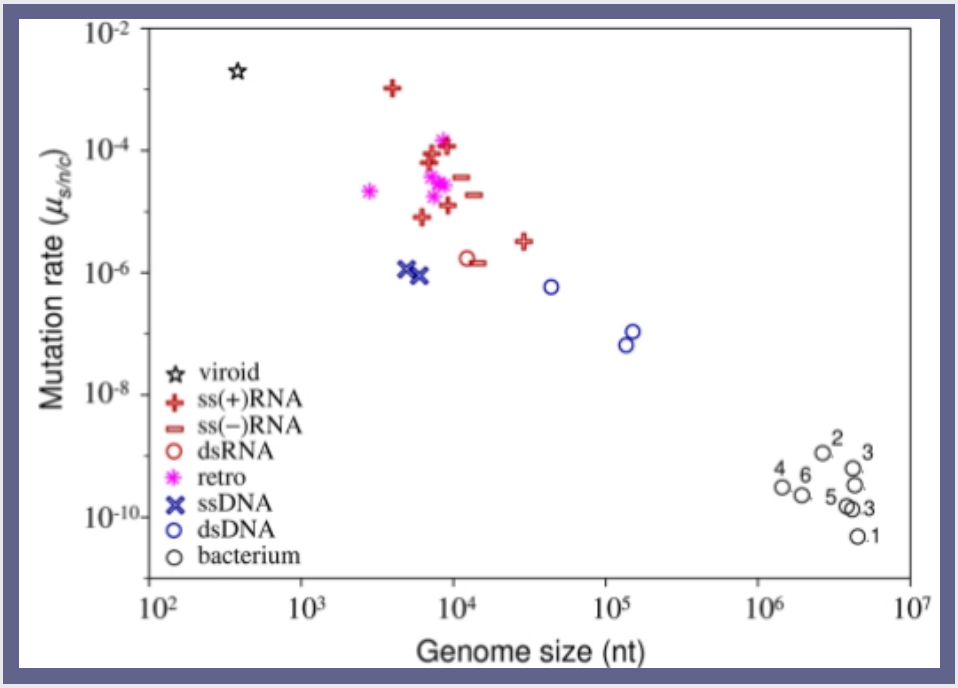


Viral genomes contain genetic variation due to factors like mutations, recombination, and re-assortment

Mutations contribute to genetic variation by creating heritable changes in the genome. Recombination exchanges nucleotide sequences among different DNA strands, and reassortment is the exchange of entire RNA molecules between genetically related viruses. This means that when two or more viruses infect a single host, they can package each other's genomes into a new virion, creating a hybrid virus. Each of these contributions of genetic variation are how new strains of virus are created.

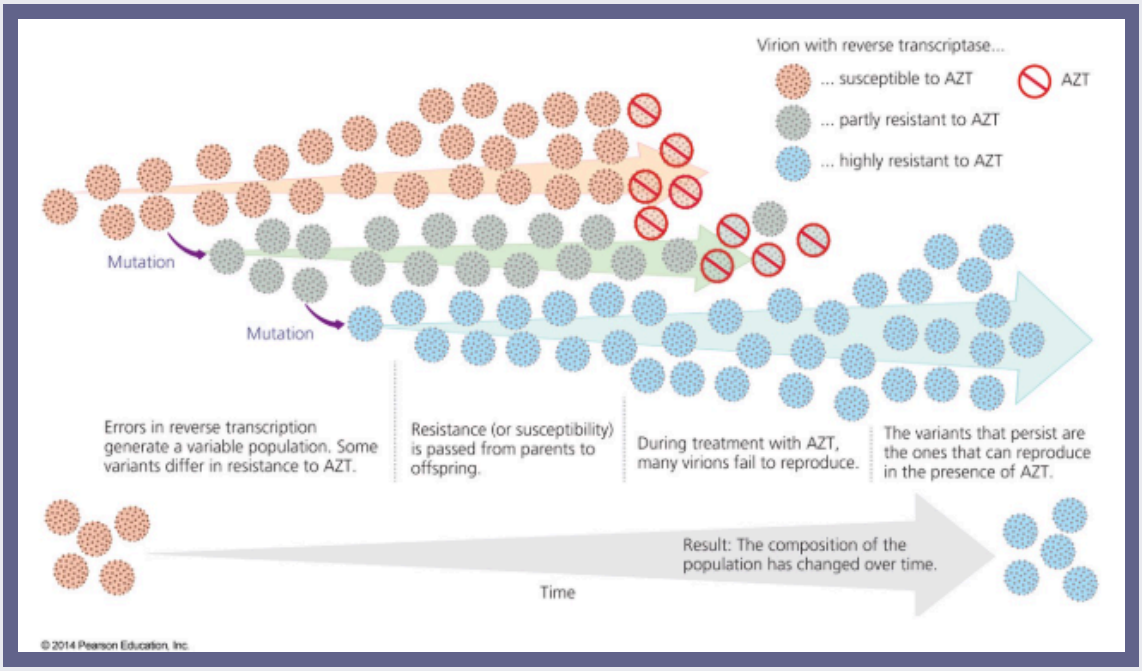
Viral genomes possess a mutation rate

The mutation rate reflects the probability of occurrence of each of the above contributions to genetic variation. Only those mutations in lineages that persist contribute to the mutation rate that is measured by whole-genome sequencing. Mutation rate typically differs by virus groups.



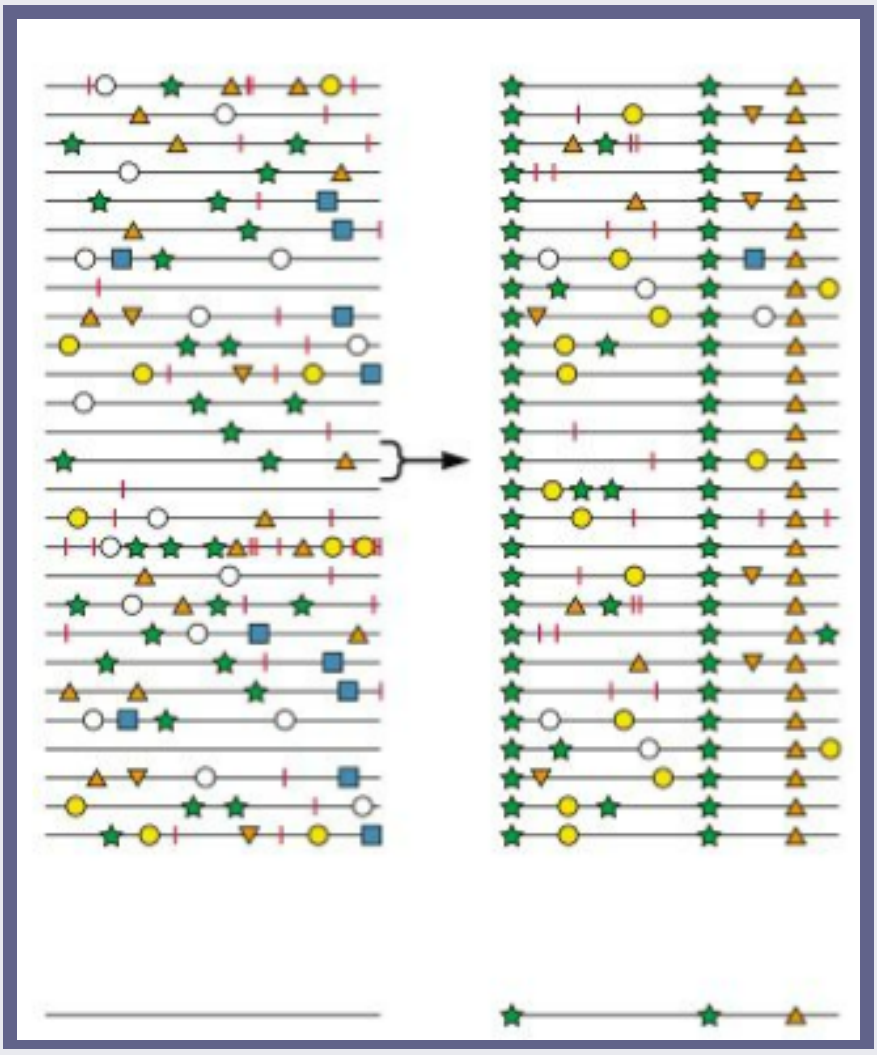
Viral genomes evolve in response to selective pressure and adapt to host interactions

The host environment is always changing and acts as a selective pressure on the virus. Over time we see that viral lineages with mutations that increase fitness increase in frequency.



Viral genomes are diverse in genome composition, structure, and reproduction

Viruses can be positive sense, negative sense, or ambisense. Each of these have their own features when impact host interaction. Viral genomes also contain differing genomic elements (ie. regulatory elements, non-coding regions, repeated elements, GC/AT rich sequences). Whole genome sequencing captures consensus genomes, which only captures a fraction of viral diversity.



III. Sources

Penguin Random House. (2015). Chapter 4: Molecular Biology. In MCAT Biology Review (2nd ed.). essay.

Lunn, Stephanie 2023. Fundamentals of Bacterial Genomics. https://nwpage.org/node/19

Lunn, Stephanie 2024. Fundamentals of Viral Genomics. https://docs.google.com/document/d/1wvG74E8yQK7NWqMUFgcJT4V8cn7T2QFh57LnflCMo-0/edit?tab=t.0

Racaniello, V. (2025, February 1). Virology Lectures 2025 3: Genomes and Genetics. MicrobeTV.