https://www.nlm.nih.gov/ncbi/workshops/2022-08_intro-to-pathogen-data/

Exercise 1: Identification of my pathogen based on a nucleotide sequence

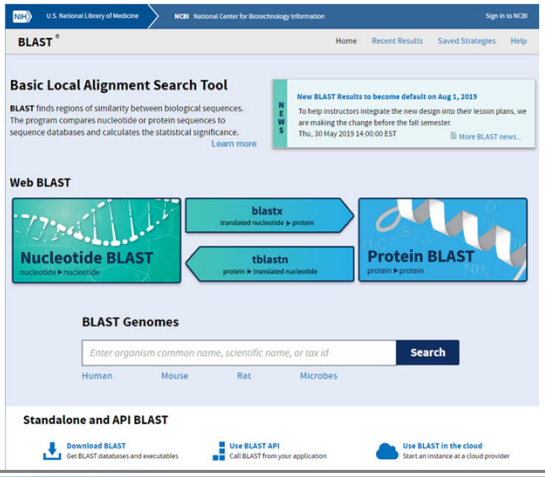
- Perform a BLAST search
- For **viral** isolates, compare your sequence to the Virus Reference Genomes database

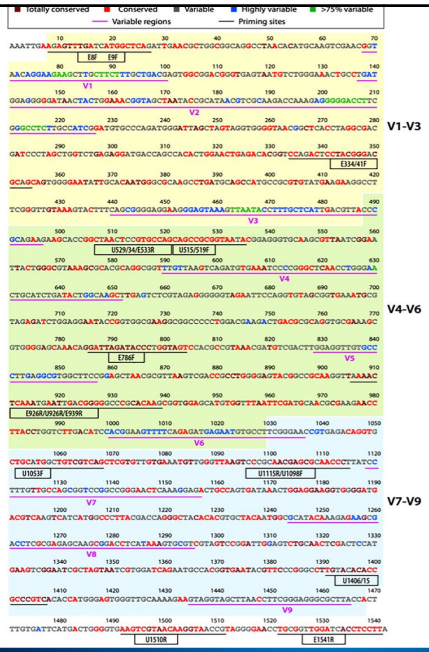
  The RefSeq Virus Genome crew is researching ways to more effectively identify viral strains and key features (genotypes, critical genetic variants).
- For **bacterial or fungal** isolates, compare your sequence to a Targeted Loci database

  The RefSeq Genome and Pathogen Project teams have automated systems to determine the identity of pathogenic samples, potentially related isolates, and characterize key phenotypes such as antimicrobial resistance.
- ***More advanced resources are at the end**, such as resources for designing your own primers!*

NCBI BLAST

https://blast.ncbi.nlm.nih.gov/



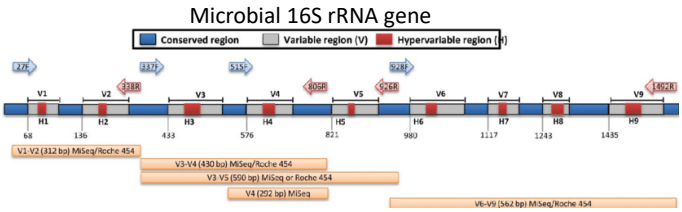


NCBI RefSeq Targeted Loci Project

https://www.ncbi.nlm.nih.gov/refseq/targetedloci/

Taking advantage of regions of:

- conservation for binding sequencing primers
- variability for sequence-based identification



https://www.nlm.nih.gov/ncbi/workshops/2022-08_intro-to-pathogen-data/

A common practice to identify pathogens… targeted loci sequencing & sequence comparison

Curated BLAST databases include selected RefSeq records and validated GenBank sequences.

- **Bacteria and Archaea: 16S rRNA gene**

  curated full length 16S ribosomal RNA sequences that correspond to bacteria and archaea type materials
- **Bacteria and Archaea: 23S rRNA gene**

  complete and near full length GenBank sequences
- **Fungi: Internal transcribed spacer (ITS) regions**

  ITS1, 5.8S gene and ITS2 sequences - near full length to complete
- **Fungi: 28S (LSU) rRNA gene**

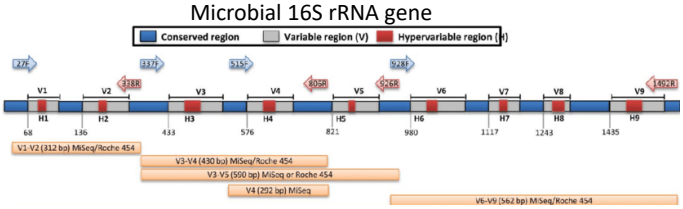
  at a minimum the sequences contain the hyper variable D1/D2 region
- **Fungi: 18S (SSU) rRNA gene**

  at a minimum the sequences contain most of the variable V4 region and part of the V5 region

NCBI RefSeq Targeted Loci Project

https://www.ncbi.nlm.nih.gov/refseq/targetedloci/

Microbial 16S rRNA gene



Fungal rRNA gene region

A common practice to identify **viral** pathogens… amplification & sequencing of key gene regions

Regions for sequencing can vary from virus-to-virus:

- **Influenza A**

  Hemagglutinin (18 subtypes): surface glycoprotein responsible for docking and membrane fusion for entry into host cells

  Neuraminidase (11 subtypes): surface protein promotes release of the virus from the host cell
- **HIV**

  Integrase portion of the polymerase gene
- **Dengue**

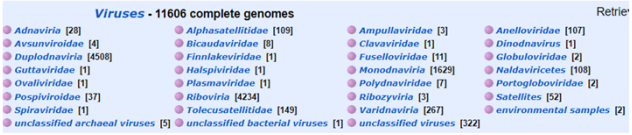
  4 serotypes – Non-structural peptide 5 (NS5)
- **SARS-CoV-2**

  CDC: Nucleocapsid (N) gene

  ORF1ab

NCBI RefSeq Viral Genomes

https://www.ncbi.nlm.nih.gov/genome/viruses/



https://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?taxid=10239

NCBI Virus Sequences for discovery **RefSeq Genome (11,598)**