

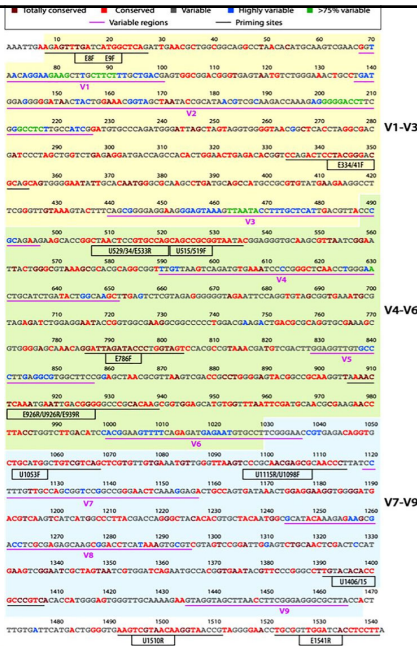
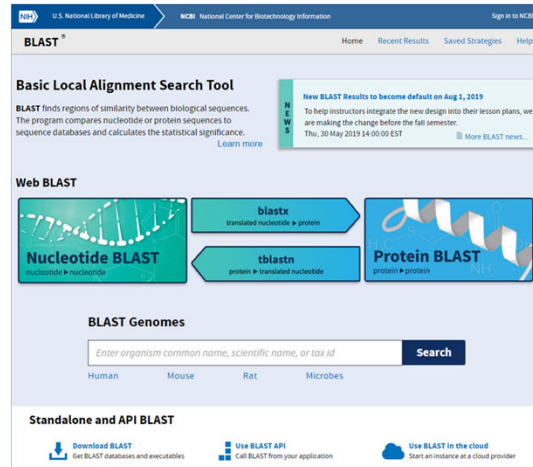
https://www.nlm.nih.gov/ncbi/workshops/2022-08_intro-to-pathogen-data/

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

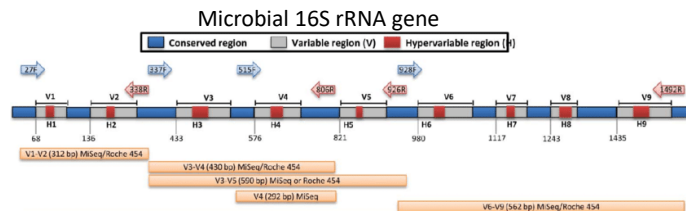
<https://blast.ncbi.nlm.nih.gov/>



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



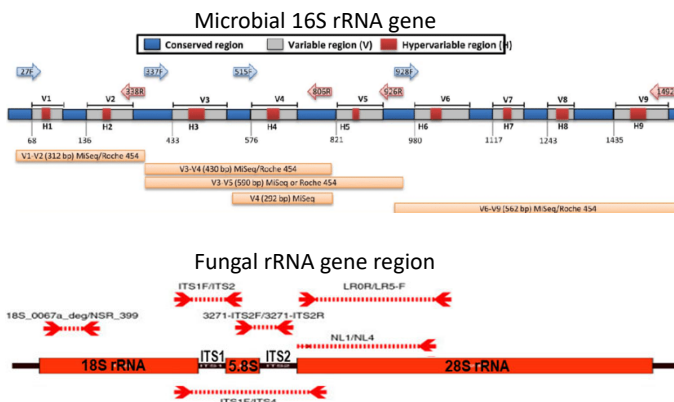
A common practice to identify pathogens... targeted loci sequencing & sequence comparison

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/>



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/>

Viruses - 11606 complete genomes

<ul style="list-style-type: none"> Adenoviridae [28] Arcanoviridae [4] Duplodnaviria [4508] Guttaviridae [1] Ovalviridae [1] Pospitiviridae [37] Spirnaviridae [1] unclassified archaeal viruses [5] 	<ul style="list-style-type: none"> Alphacaudoviridae [109] Bicaudoviridae [8] Finnlakeviridae [1] Halspiviridae [1] Plasmaviridae [1] Riboviria [4234] Tolucaviridae [149] unclassified bacterial viruses [1] 	<ul style="list-style-type: none"> Ampullaviridae [3] Clavoviridae [1] Fuselloviridae [11] Monodnaviria [1629] Polydnaviridae [7] Ribozyviria [1] Vairidnaviria [267] unclassified viruses [222] 	<ul style="list-style-type: none"> Anelloviridae [107] Dinodnavirus [1] Globuloviridae [2] Naldaviricetes [108] Portogloboviridae [2] Satellites [32] environmental samples [2]
--	---	--	--

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

NCBI Virus RefSeq Genome (11,598)

Choose Search Set

Database: Standard databases (nr etc.) rRNA/ITS databases Genomic + transcript databases Betacoronavirus

RefSeq Genome Database (refseq_genomes)

Organism: Viruses (taxid:10239) exclude

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

Title: Refseq viruses representative genomes
Molecule Type: mixed DNA
Update date: 2022/06/08
Number of sequences: 15002