Exercise 2: Find high-quality sequence information for my pathogen

NCBI Genome & Assembly Databases

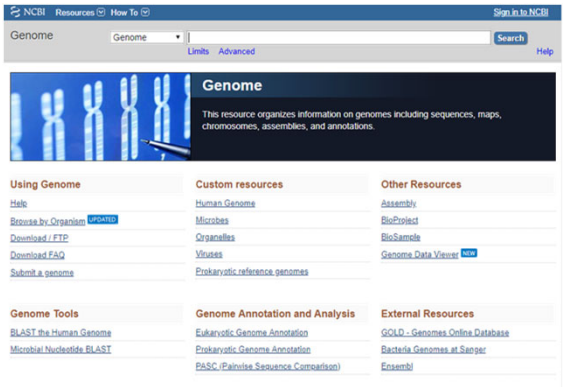
https://www.ncbi.nlm.nih.gov/genome

https://www.ncbi.nlm.nih.gov/assembly

- Understand how reference genomes are selected and annotated
- Identify key reference genomes with assembly/annotation statistics
- Find information about other related genomes
- Download genome sequences or annotations
- Access gene & protein sequences

It is often convenient to link to or start in Genome to find genome assembly & annotation information.

(or start in the Gene database for gene or protein-specific information)

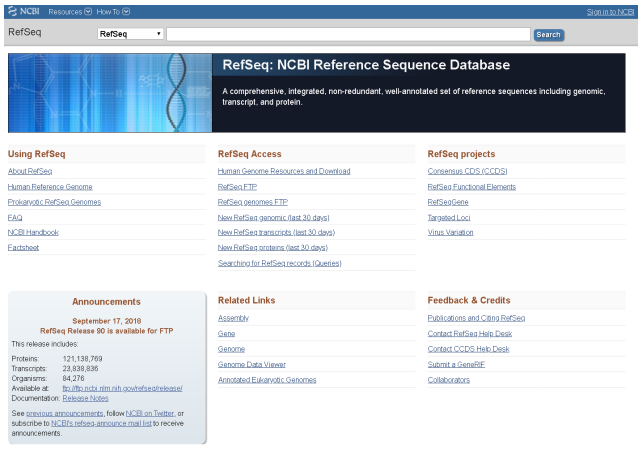


International Nucleotide Sequence Database Collaboration (INSDC)

https://www.insdc.org/

NCBI's Reference Sequences Project (RefSeq)

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



Sources & Types of Sequences

Primary sequences: Submitted nucleotide sequences with protein translations "owned" by the submitters

Stored in the Nucleotide & Protein databases

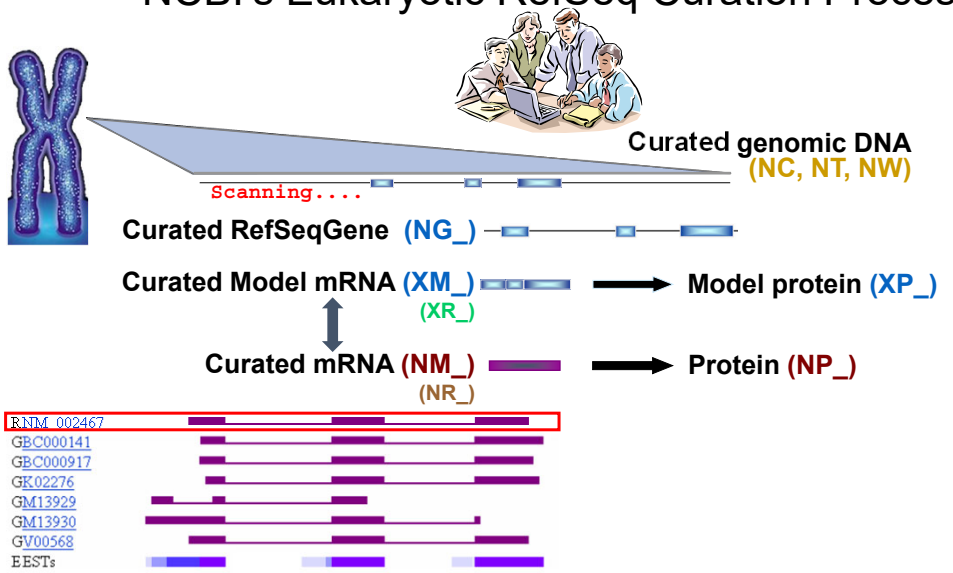
International Sequence Database Collaboration (INSDC)

- **GenBank** – U.S. Sequence Database at NCBI
- **European Nucleotide Archive** at EBI
- **DNA Databank of Japan** at NIG

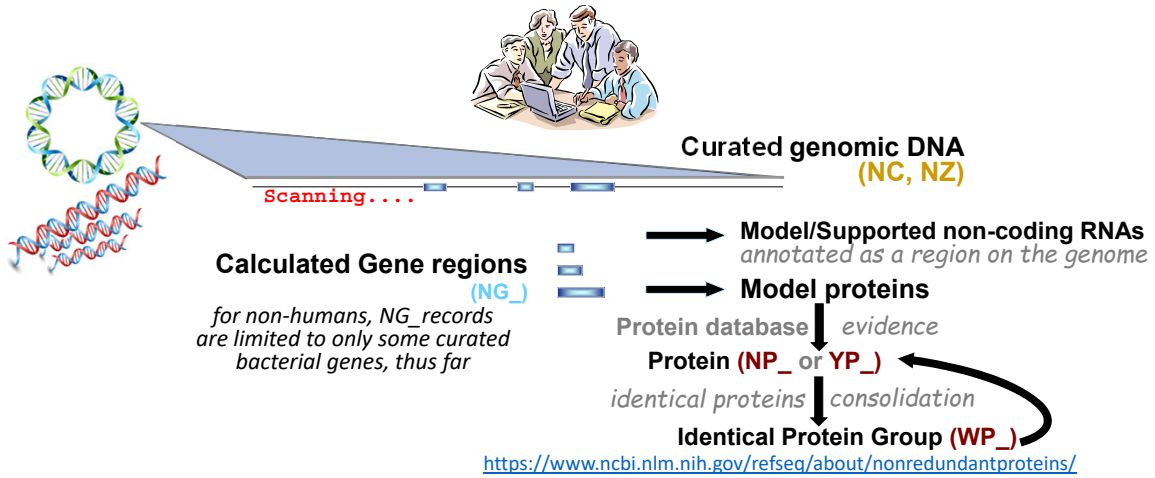
Stored in the Sequence Read Archive (SRA) database

- A repository for NextGen sequence data: including whole genome, metagenome and transcriptome data
- Supporting data are in BioProjects, BioSample, & GEO and the Pathogen Detection Project
- Data are shared with the other INSDC repositories

NCBI's Eukaryotic RefSeq Curation Process



NCBI's Microbial RefSeq Curation Process



Sources & Types of Sequences

Primary sequences: Submitted nucleotide sequences with protein translations "owned" by the submitters

Stored in the Nucleotide & Protein databases

International Sequence Database Collaboration (INSDC)

- **GenBank** – U.S. Sequence Database at NCBI
- **European Nucleotide Archive** at EBI
- **DNA Databank of Japan** at NIG

Stored in the Sequence Read Archive (SRA) database

- A repository for NextGen sequence data: including whole genome, metagenome and transcriptome data
- Supporting data are in BioProjects, BioSample, & GEO and the Pathogen Detection Project
- Data are shared with the other INSDC repositories

Curated high-quality nucleotide and protein sequences
Produced, "owned" and updated by NCBI.

NCBI Reference Sequences (RefSeq)

- Provide reference standards
- Records represent all molecules in the central dogma
  - Eukaryotes: genomic, mRNA & ncRNA, proteins
  - Prokaryotes and Viruses: genomic, ncRNA & protein *(no mRNA records)*
- Distinct accessions with a "prefix and underscore (_)"
  - genomic: NC_, AC_, NG_, NZ_
  - RNA: NM_, NR_, XM_, XR_
  - protein: NP_ (YP_), XP_, **WP_**
- Developing issue: redundant, redundant, redundant proteins
  - We have over 200,000 RefSeq bacterial assemblies - *many of them have identical protein sequences*
  - A Solution: Make one copy of a "shared protein sequence" to link all annotations in an **Identical Proteins Report**

    Example: AMR RefSeq Gene NG_049253.1

    Annotated on 4,793 Bacterial genomic assemblies

    WP_004199234.1, MULTISPECIES ***Taxonomic Group*** carbapenem-hydrolyzing class A beta-lactamase KPC-2 [Bacteria]

Curated Genome data at NCBI: Selection criteria

Viruses may have one or more ***reference genomes*** per species.
International Committee on Taxonomy of Viruses (ICTV) designates exemplar(s) for selection of reference genome assembly(ies).

Prokaryotes may have more than one ***reference or representative genomes*** per species.

- RefSeq **reference genomes** - selected based on assembly and annotation quality, existing experimental support, and recognition as a ***community standard*** (ex: Escherichia coli str. K-12 substr. MG1655) or of ***clinical importance*** (ex: Escherichia coli O157:H7 str. Sakai or Mycobacterium tuberculosis H37Rv).
- RefSeq **representative genomes** - assigned to type strain assemblies if there is no current ***reference*** genome or another one if it is scientifically significant and exhibits strong sequence ***diversity*** as compared to the assigned reference genome(s) (such as Mycobacterium avium subsp. paratuberculosis K-10 or Streptococcus thermophilus JIM 8232)

Eukaryotes (incl. fungi & helminths) - ***no more*** **than one** ***reference or representative genome*** per species.

- RefSeq **reference genomes** – selection based on assembly and annotation quality, existing experimental support, and recognition as a community standard or of clinical importance (ex: Aspergillus fumigatus Af293)
- If there are no assemblies in RefSeq for a particular eukaryotic species, then RefSeq will select a **representative genome** from the highest quality GenBank assembly (ex: Schistosoma mansoni (ASM23792v2))

For more information: https://www.ncbi.nlm.nih.gov/assembly/help/

Viral vs. Bacterial Genome Annotation

Up till now

Manual Curation of Viral Genomes

Scan GenBank for new sequences → Genome selection by ICTV → Reference creation → link related genomes → "Index" genomes

Provisional: direct copy of a "good quality" GenBank genome record

Reviewed: manually-reviewed record with reviewed sequence & annotation information, publications, and BLAST-enabled annotations

Influenza, Norovirus, Dengue: *automated "wizards" for GenBank Submissions*

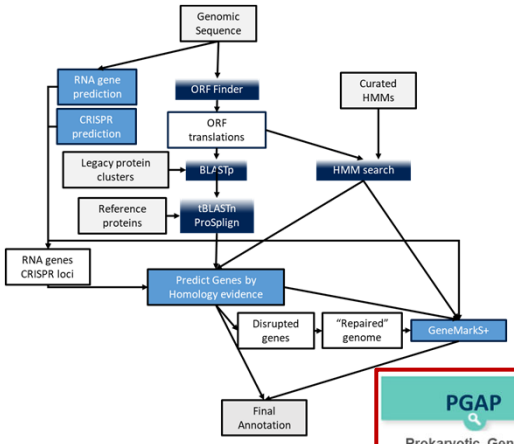
Were working to add a Flavivirus pipeline and others….

and then SARS-CoV-2 hit!

VADR
Viral Annotation DefineR
https://github.com/ncbi/vadr
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7245624

Automated Process : **PGAP** pipeline

https://www.ncbi.nlm.nih.gov/genome/annotation_prok/process/



PGAP
Prokaryotic Genome Annotation Pipeline
https://github.com/ncbi/pgap