Lessons: Sequence Formats and Online Databases

Student name: Ruth Gómez Graciani

1. Nucleotide database and its divisions.

Localize all of the rabbit genomic records at the NCBI nucleotide database coming from a whole genome shotgun (WGS) project.

- *Nucleotide:* (Rabbit[Primary Organism]) AND "wgs"[Properties]

2. Cross-references at nucleotide database.

How to find at NCBI nucleotide database all human codifying mRNA sequences with link to the OMIM database?

- *Nucleotide:* ((Human[Primary Organism] AND "cds"[Feature key]) AND "biomol mrna"[Properties]) AND nucleotide_OMIM[Filter]

3. Bacterial genomes.

How many Escherichia coli genomes are there at the NCBI regardless of their status?

- *Genome:* "Escherchia coli"[Organism]

4. Viral genomes.

There are several genomes at the NCBI without annotation. That means none of the genetic elements like CDSs, genes or any other features annotated as "misc_feature", have been indicated in the sequence. List all the records at the NCBI nucleotide database for HIV-1 unannotated complete genomes. How many entries have you found? Indicate here the accesion.version number of all of them.

- *Nucleotide:* "hiv 1"[Primary Organism] AND "complete genome"[Title] NOT ("cds"[Feature key] OR "gene"[Feature key] OR "misc feature"[Feature key])
- 5 entries were found. Accession list: JN571034.1, AY781128.1, AY781127.1, AY781126.1, AY781125.1

5. Taxonomy database.

How do I find all the "validated" bacterial genera with nucleotide sequences at the NCBI?

Note: "validated" means not including "candidate" or "candidatus" genera.

- *Taxonomy:* ("bacteria"[Subtree] AND "genus"[Rank] NOT ("candidate"[Text Word] OR "candidatus"[Text Word])) AND taxonomy_nucleotide[Filter]

6. Protein database.

There are several entries at NCBI containing human hemoglobin sequences. However, the following search strategy gives no results at NCBI protein database. Why?

(human[porgn] AND hemoglobin[protein name] AND refseq[filter])

- This query is serching RefSeq entries that have exactly 'Hemoglobin' in their protein name, and this is impossible in humans because it is composed by different subunits, so the protein name will always be 'hemoglobin subunit' and something else. The query can be improved by modifying the [protein name] filter to be `hemoglobin subunit *[Protein Name]` and adding `hemoglobin subunit [Title]`

7. Cross-references at protein database

The protein sequences at the NCBI Protein database come from several different sources, such as UniProt. How to list with a simple search all the proteins at the NCBI coming from the Swiss-prot?

- *Protein:* "srcdb swiss prot"[Properties]

8. Redundancies at protein database.

Look for record WP_003094337.1 at the protein NCBI database. What kind of entry is it? What species it belongs? From what nucleotide entry this protein sequence was obtained? Why the NCBI has created this sort of records?

- It is a non-redundant protein sequence from multiple *Pseudomonas* species, and its source nucleotide entry is undetermined, as different organisms and sequences were used to create this protein entry. This way they avoid reduncdancies in protein entries when that protein is widely used by a genus.

9. The Entrez Programming Utilities

Create an URL using E-utilities to search in PubMed for papers with the term "cancer" in the title and with publication date in 2017; retrieve the first 100 PMIDs (IDs at PubMed).

Tips here https://dataguide.nlm.nih.gov/eutilities/utilities.html

- https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=cancer%5Btitle%5D+AND+2017%5BPDat%5D&retmax=100