Practical Session #2: Primary nucleotide sequence resources

Help here: https://www.ncbi.nlm.nih.gov/books/NBK49540/

I. Organization of information in the NCBI nucleotide database.

What taxonomical division do these sequences belong to?

NM_001727: **9606 homo sapiens → PRI**

JFFV01000040: **1280 Staphylococcus aureus → BCT**

Z12018: **6239 Caenorhabditis elegans → INV**

What functional division, high-throughput or sub-database do these sequences belong to?

NM_001727: -

JFFV01000040: **WGS**

Z12018: **HTG**

II. Identify the function of the following indexed field when added to any term during a query to a sequence database

[porgn] **Filter for primary organism**

[GENE] **Gene names on database records**

[PROP] **Properties of the term (molecular type, database sources, …)**

[filter] **Filter data subsets**

[Protein name] **Gives names of protein products**

III. At NCBI nucleotide database what does this query search do?

Txid9031[porgn] AND biomol_mrna[PROP] AND refseq[filter] AND 0:1000[SLEN]

Gives the properties of messenger RN from chickens up to position 1000, filtering only from the RefSeq database.

IV. How many mice (*Mus musculus*) ribosomal RNA (rRNA) sequences are there in the nucleotide database at the NCBI?

82

How many of these sequences belong to the sub-database RefSeq? What is the accession.version number of the longest sequence of this filtered set?

66 belong to sub-database Ref Seq

NR_046233.2 is the accesion version

V. Find out the nucleotide entries at NCBI which meet the following statements.

- The human reference mRNA sequence encoding for the epidermal growth factor variant 1.

  ~~**EGF[GENE] AND refseq[filter] AND Txid9606[Organism] → 7**~~

  EGF[GENE] AND variant 1[title] AND biomol_mrna[PROP] AND Txid9606[Organism] → NM_001963

- The largest human mRNA sequence.

  Txid9606[Organism] AND mrna[filter] → NM_001267550.2

  ---------------------------------biomol_mrna[PROP]

- It is a DNA sequence from Amoeba proteus that contains the gene AQP.

  "Amoeba proteus"[Organism] AND "AQP"[GENE]

  ---AND biomol_genomic[PROP]

- The longest genomic sequence from the bacterium *Burkholderia pseudomallei* bearing among others the genes *recA* and *recX*.

  Burkholderia pseudomallei[Primary Organism] AND recA[Gene Name] AND recX[Gene Name] *+sort by seq length*

  Enter first option>the sequence is identical to the next one

- An mRNA sequence greater than 10 kb from the Asian tiger mosquito without an annotated protein coding region (take a look at the **exercise VI** to solve this)

  N

V.1 In the entry for the reference mRNA sequence for the human epidermal growth factor variant 1 what are the coordinates of the codifying region or CDS? What is the name for this gene?

VI. The Search Field tag [Feature Key] or [fkey] is used to retrieve records annotated with a given biological feature.

Select what the result would be for the following search strategies (match each strategy with the expected result):

Query

human[porgn] AND CDS[fkey]

human[porgn] AND gene[fkey]

Txid9606[porgn] NOT CDS[fkey]

Homo sapiens[porgn] NOT intron[fkey]

Result

A list of human sequences without an annotated coding region

A list of human sequences without an annotated intron

A list of human sequences with at least one annotated protein codifying segment

A list of human nucleotide sequences with an indicated gene region

Tips: https://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html

VII. Linking two databases (hint: use filter key: *currentdatabase_linkeddatabase*).

7.1 List all mouse nucleotide sequences codifying for at least one protein indexed in NCBI.

7.2 List all bacterial nucleotide sequences that have a reference to a scientific paper.

7.3 List all human nucleotide sequences related with at least one phenotype or disease(human genes and genetic disorders).