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 PRI

JFFV01000040 BCT

Z12018 INV

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

The answer is in the field "KEYWORDS".

NM_001727 Refseq

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] It is relevant when there is more than one source organism.

[GENE] It retrieves records with the term annotated in the field for gene names.

[PROP] Molecular type, source database, and other properties of the sequence record.

[filter] It can filter records according to database source or molecular type.

[Protein name] It uses the names of protein products as annotated on sequence records.

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

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

It lists all chicken reference mRNA sequences less than 1kb

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

"Mus musculus"[Primary Organism] AND "biomol rrna"[Properties]

How many of these sequences belong to the sub-database RefSeq?

See here Refseq prefixes:
https://www.ncbi.nlm.nih.gov/books/NBK21091/table/ch18.T.refseq_accession_numbers_and_mole/?report=objectonly

All of them start with a "_" so we can download the results in "summary" format and:

```
grep "_" III-nuccore_result.txt |less -NS
```

or use filter in the left margin or
(("Mus musculus"[Primary Organism] AND "biomol rrna"[Properties])) AND "refseq"[Filter]

66

What is the accession.version number of the longest sequence of this filtered set?

Sort by Sequence Length:
Mus musculus 45S pre-ribosomal RNA (Rn45s), ribosomal RNA
13,400 bp linear rRNA
NR_046233.2 GI:577019615

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. NM_001963.6
  ((epidermal growth factor variant 1[Title] AND "biomol_mrna"[Properties]) AND "Homo sapiens"[Primary Organism] AND refseq[filter]) NOT receptor[Title]
- The largest human mRNA sequence. NM_001267550.2
  human[Primary Organism] AND "biomol_mrna"[Properties]
  (and sort by sequence length)
- It is a DNA sequence from Amoeba proteus that contains the gene AQP. AB292479.1
  Amoeba proteus [porgn] AND AQP[gene] AND biomol_genomic[prop]
  Notice: No biomol_dna property exists.
- The longest genomic sequence from the bacterium *Burkholderia pseudomallei* bearing among others the genes *recA* and *recX*. NZ_KN150939.1
  "Burkholderia pseudomallei"[Primary Organism] AND recA[Gene Name] AND recX[Gene Name]
  Can you answer the question: Why the GeneBank entry from which this RefSeq entry comes from is not listed in the search results?
- An mRNA sequence greater than 10 kb from the Asian tiger mosquito without an annotated protein coding region (look at the **exercise VI** to solve this). JO860042.1
  txid7160[porgn] AND biomol_mrna[prop] NOT CDS[fkey] NOT 0:10000[slen]

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? From 454 to 4077. What is the name for this gene? EGF
((epidermal growth factor variant 1[Title] AND "biomol_mrna"[Properties]) AND "Homo

sapiens"[Primary Organism] AND refseq[filter]) NOT receptor[Title]
NM_001963.6

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):

human[porgn] AND CDS[fkey]

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

human[porgn] AND gene[fkey]

A list of human nucleotide sequences with an indicated gene region

Txid9606[porgn] NOT CDS[fkey]

A list of human sequences without an annotated coding region

Homo sapiens[porgn] NOT intron[fkey]

A list of human sequences without an annotated intron

VII. Linking two databases.

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

Txid10090[porgn] AND nucleotide_protein[Filter]

Mus musculus[porgn] AND nucleotide_protein[Filter]

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

Txid2[porgn] AND nucleotide_pubmed[Filter]

THIS ONE DOES NOT WORK: bacteria[porgn] AND nucleotide_pubmed[Filter] Why?

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

Txid9606[porgn] AND nucleotide_OMIM[Filter]

human[porgn] AND nucleotide_OMIM[Filter]