

<b>Started on</b>	Thursday, 5 October 2023, 9:26 AM
<b>State</b>	Finished
<b>Completed on</b>	Monday, 9 October 2023, 11:08 PM
<b>Time taken</b>	4 days 13 hours
<b>Marks</b>	14.43/20.00
<b>Grade</b>	7.21 out of 10.00 (72.14%)

### Question 1

Correct

Mark 1.00 out of 1.00

**Question 1.** Select the description best corresponding to each [NCBI](#) accession code. (No penalty)

**Note:** Select a different commentary for each accession number.

**Tips:** The revision history format shows the various accession numbers, version numbers, and update dates for sequences that appeared in a specific GenBank record.

- |                |   |   |
|----------------|---|---|
| GFSJ01000014.1 | ✓ | The entry contains a sequence from a transcriptome shotgun assembly (TSA) ▾                 |
| AC275307.1     | ✓ | The entry contains a working draft sequence from a high-throughput genomic (HTG) approach ▾ |
| CP018631.1     | ✓ | The current version of a nuccore genbank entry for a circular DNA molecule ▾                |
| M60741.1       | ✓ | An old version of a nuccore genbank entry ▾   |
| LKPB02002952.1 | ✓ | The entry contains a whole genome shotgun sequence (WGS) ▾                                  |
| NM_000517.6    | ✓ | The current version of a refseq entry ▾   |

The correct answer is: GFSJ01000014.1 → The entry contains a sequence from a transcriptome shotgun assembly (TSA), AC275307.1 → The entry contains a working draft sequence from a high-throughput genomic (HTG) approach, CP018631.1 → The current version of a nuccore genbank entry for a circular DNA molecule, M60741.1 → An old version of a nuccore genbank entry, LKPB02002952.1 → The entry contains a whole genome shotgun sequence (WGS), NM\_000517.6 → The current version of a refseq entry

**Question 2**

Correct

Mark 2.00 out of 2.00

**Question 2.** What sequencing technology was used to obtain sequence LKPB02002952.1?

- ☐ Illumina HiSeq2500.
- ☒ PacBio. ✓
- ☐ Ion torrent.
- ☐ Oxford nanopore.
- ☐ Don't know/no answer. (without penalty)

Mark 1.00 out of 1.00

The correct answer is: PacBio.

What was the aim of the sequencing project in which this sequence was obtained?

- ☐ To sequence only the X and Y chromosomes in a family.
- ☐ To study the expression of human genes.
- ☐ To sequence the human genome for the very first time.
- ☒ To add diverse allelic variation to the reference human genome. ✓
- ☐ Don't know/no answer (without penalty)

Mark 1.00 out of 1.00

The correct answer is: To add diverse allelic variation to the reference human genome.

GenBank files usually contains this information.

<https://www.ncbi.nlm.nih.gov/nuccore/LKPB02002952.1>

Sequencing Technology :: PacBio
---------------------------------

Sequence from this project will be used to improve the contiguity of the human reference sequence and add diverse allelic variation
---

You can also find the second answer in the bioproject link:

<https://www.ncbi.nlm.nih.gov/bioproject/PRJNA288807>

For the 1st one you will have to go to ASSEMBLY from the "related information" link in the bioproject page:

[https://www.ncbi.nlm.nih.gov/assembly?LinkName=bioproject\\_assembly\\_all&from\\_uid=288807](https://www.ncbi.nlm.nih.gov/assembly?LinkName=bioproject_assembly_all&from_uid=288807)And select "[NA19240\\_prelim\\_3.0](#)" which will send you to the entry in the Assembly database:[https://www.ncbi.nlm.nih.gov/assembly/GCA\\_001524155.4](https://www.ncbi.nlm.nih.gov/assembly/GCA_001524155.4)

**Question 3**

Correct

Mark 1.00 out of 1.00

**Question 3.** Two of these search tags will help you finding sequences that are codifying at [nucleotide](#) database (CDS). Which ones?

**Tips:**

Learn about Search Field Descriptions for Sequence Database in this [link](#).

Select one or more:

- ☐ a. protein\_nucleotide[filter]
- ☐ b. All[protein]
- ☒ c. nucleotide\_protein[filter] ✓
- ☒ d. CDS[fkey] ✓
- ☐ e. protein[fkey]
- ☐ f. protein[filter]

nucleotide\_protein[filter] will deliver sequences that are in nucleotide (AKA: nuccore) and have a link to protein.

CDS[fkey]: sequences that contain one or more CDS features annotated

however you must take into account that not all annotated CDSs have a link to the protein DB

The correct answers are: nucleotide\_protein[filter], CDS[fkey]

**Question 4**

Partially correct

Mark 2.00 out of 2.00

**Question 4.** Localize all of the records in the NCBI nucleotide database containing rabbit genomic sequences that are codifying. How many have you found?  ✓

Optionally write here your search strategy

✗ (no rate)

¿How many of these sequences come from the mitochondrial compartment?  ✓

Here the question asked for genomic sequences, then, biomol\_genomic[PROP].

rabbit[porgn] AND biomol\_genomic[PROP] AND CDS[fkey]

This is not the right answer:

rabbit[PORGN] AND biomol\_genomic[PROP] AND nucleotide\_protein[filter]

the fact that their products don't have a link to the protein DB does not mean that they are not there.

In the left margin you can see that 159 are in the Mithocondrion

OR use this search:

rabbit[porgn] AND biomol\_genomic[PROP] AND CDS[fkey] AND mitochondrion[filter]

**Question 5**

Partially correct

Mark 1.43 out of 5.00

**Question 5.** Do you think you have found in Question 4 the NCBI nucleotide reference entry for the complete rabbit mitochondrial chromosome? Yes

If yes, what is its refseq access number? AJ001588.1

How many proteins does it code? 13

Between what coordinates do we find the tRNA for tryptophan? From 4971 to 5037

Click on the filter "Mithochondrion" in the question 4 results and select the only refseq result referring to the Mithochondrion complete genome "Oryctolagus cuniculus mitochondrion, complete genome". NC\_001913.1 ([https://www.ncbi.nlm.nih.gov/nuccore/NC\\_001913.1](https://www.ncbi.nlm.nih.gov/nuccore/NC_001913.1)). You can recognize it is a code from RefSeq database due to its starting characters "NC\_".

Or perform a search like this:

rabbit[porgn] AND biomol\_genomic[prop] AND CDS[fkey] AND refseq[filter] AND mitochondrion[filter]

Use "Related information" box and click in protein: ([https://www.ncbi.nlm.nih.gov/protein?LinkName=nuccore\\_protein&from\\_uid=5835526](https://www.ncbi.nlm.nih.gov/protein?LinkName=nuccore_protein&from_uid=5835526)) to obtain results for the proteins encoded in this entry.

In the gbff file ([https://www.ncbi.nlm.nih.gov/nuccore/NC\\_001913.1](https://www.ncbi.nlm.nih.gov/nuccore/NC_001913.1)) search for "tRNA-Trp"

```
tRNA      4971..5037
          /product="tRNA-Trp"
```

**Question 6**

Partially correct

Mark 1.00 out of 1.00

**Question 6.** ¿How many complete mammalian mitochondrial chromosomes can be found at the NCBI nucleotide database? 77718

Optionally write here your search strategy

mammals[filter] AND "complete genome"[title] also select mitochondrion in (no rate)

**Tips:** Mitochondria has its own genome. Researchers and/or curators use to add to the title of entries containing complete genomic sequences the phrase “complete genome” for fully sequenced chromosomes or genomes of eukaryotic organelles.

mammalia[porgn] AND mitochondrion[filter] AND "complete genome"[TI]

mitochondrion[filter] and gene\_in\_mitochondrion[PROP] both give the same result.

txid40674 and mammalia are synonyms.

"complete genome" should be added using [title] or [word] as search tag or simply nothing. This is an approximation to the best way to get the most results. There is not a search tag to filter for the status of a genome sequencing project at NCBI nucleotide.

**Question 7**

Correct

Mark 1.00 out of 1.00

**Question 7.** One of the proteins encoded in the mitochondrial chromosomes is cytochrome b. The best strategy to search for sequences encoding for cytochrome b proteins in NCBI nucleotide database is to include in a query the expression:

Select one:

- ☒ a. AND (cytochrome b[protein name]) ✓
- ☐ b. AND "cytochrome b"[title]
- ☐ c. NOT cytochrome\*
- ☐ d. AND cytochrome\*[title]
- ☐ e. Don't know/no answer (without penalty)

in this case you should evaluate what would happen if you add each of the phrases in a search strategy:

AND "cytochrome b"[title]; with this phrase you are including other proteins like cytochrome b reductase (e.g. NM\_001011954.1).

AND (cytochrome b[protein name]); this strategy will retrieve all nucleotide sequences encoding for a protein named exactly as "cytochrome b", including complete bacterial genomes and eukaryotic mitochondrial genomes.

AND cytochrome\*[title]; this strategy includes other variants like cytochrome c.

NOT cytochrome\*; this strategy will exclude cytochrome b from your search

The correct answer is: AND (cytochrome b[protein name])

**Question 8**

Correct

Mark 2.00 out of 2.00

**Question 8.** Find out the NCBI genomic reference sequence encoding for human (us) cytochrome b protein.

Indicate the NCBI accession numbers for both the genomic sequence  ✓ and the encoded protein  ✓ .

Strategy

"human"[Primary Organism] AND "cytochrome b"[protein name] AND refseq[filter]

Sequence with accession number NC\_012920.1 is the reference sequence in refseq for the Homo sapiens mitochondrion complete genome. This is the reference DNA sequence encoding, among other proteins, for the human cytochrome b protein (YP\_003024038.1).

AC\_000021 is an obsolete version of NC\_012920, however, if you have answered "NC\_012920, AC\_000021" this is counted as correct. Please notify me if you have introduced a different variation

**Question 9**

Correct

Mark 1.00 out of 1.00

**Question 9.** By querying the NCBI nucleotide database investigate if bacteria genomes encode for cytochrome b protein.

**Tips:**

[Enterobacteria](#) and [Cyanobacteria](#) are two different groups of [bacteria](#).

Select one:

- ☐ a. Yes, cytochrome b is exclusive of enterobacteria.
- ☐ b. No, bacteria do not have cytochrome b protein.
- ☐ c. Don't know/no answer (without penalty).
- ☐ d. Yes, but only in the group of Cyanobacteria.
- ☒ e. Yes, several bacterial groups encode for a protein named cytochrome b. ✓

bacteria[porgn] AND "cytochrome b"[protein] You will get more than 128000 entries in almost all bacterial groups.

Enterobacteria and Cyanobacteria are two groups of bacteria. If you indicate the following search strategies you will find fewer number of entries.

cyanobacteria[porgn] AND "cytochrome b"[protein]

enterobacteria[porgn] AND "cytochrome b"[protein]

The correct answer is: Yes, several bacterial groups encode for a protein named cytochrome b.

**Question 10**

Partially correct

Mark 2.00 out of 2.00

**Question 10.** Complete microbial genomes at NCBI.**10.1.** Open entry with accession number NC\_001802.1. What does this sequence represent?

- ☐ A partial sequence for a viral genome.
- ☐ A complete genomic sequence for a human gene.
- ☒ A reference sequence for a complete viral genome. ✓
- ☐ The reference sequence for the human genome.
- ☐ Don't know/no answer (without penalty)

Mark 1.00 out of 1.00

The correct answer is: A reference sequence for a complete viral genome.

**10.2.** It is possible to find complete genome sequences in the NCBI nucleotide database for any organism. Create a search strategy trying to find at the nucleotide database all the complete or nearly complete genomic sequences for Human immunodeficiency virus type 1.

How many entries have you found? 6723 ✓

Optionally write here your search strategy

txid11676[Organism] AND "complete genome"[Title] ✗ (no rate)

**Tips:**

- Use the taxid instead of the full name for the Human immunodeficiency virus type 1
- Researchers and/or curators use to add to the title of entries containing complete or nearly complete genome sequences the phrase "complete genome"

[https://www.ncbi.nlm.nih.gov/nuccore/NC\\_001802.1](https://www.ncbi.nlm.nih.gov/nuccore/NC_001802.1)

Human immunodeficiency virus 1, complete genome

Strategy at nucleotide database:

Usually you will go with:

txid11676[Primary organism] AND "complete genome"[title]

But it is not bad to check this:

txid11676[Primary organism] AND "complete genome"

you get 10 more results. Some of them are partial, but some of them are good like this one:

<https://www.ncbi.nlm.nih.gov/nuccore/AJ508595.1>

**Question 11**

Incorrect

Mark 0.00 out of 2.00

**Question 11.** There are several genomes at the NCBI without annotations. That means none of the genetic elements like CDSs, genes or any other features annotated as “misc\_feature”, have been indicated in the sequence. Search the records in the NCBI nucleotide database for HIV-1 (Human immunodeficiency virus type 1) complete genomes without any annotation.

How many entries have you found?  ❌

Indicate here the accession.version number of only one of them.  ❌

Optionally indicate here your search strategy

❌ (no rate)

Option 1.

`txid11676[porgn] AND \"complete genome\"[title] NOT CDS[fkey] NOT Gene[fkey] NOT misc_feature[fkey]`

Only 5 entries fulfill these criteria. JN571034.1, AY781127.1, AY781125.1, AY781128.1 and AY781126.1.

UPDATE: now it is 10!

JN571034.1, AY781127.1, AY781125.1, AY781128.1, AY781126.1, ON245430.1, ON245428.1, ON245429.1, ON245427.1, ON245431.1

so both answers are considered right.

However, there is a second option:

`txid11676[Primary organism] AND complete genome NOT CDS[fkey] NOT Gene[fkey] NOT misc_feature[fkey]`

Here we are not forcing the words "complete" and "genome" neither to the title nor to be together so we can view more results

We use to get 21 good entries (2LDL\_A is not a complete genome) with accession numbers:

```
JX503079.1
M93259.1
M93258.1
JX503077.1
JX503078.1
JX503080.1
JX503073.1
JX503074.1
JX503082.1
JX503071.1
JN571034.1
JX503072.1
JX503075.1
JX503081.1
JX503076.1
JX503083.1
AY781127.1
AY781125.1
AY781128.1
AY781126.1
MK457954.1
```

UPDATE: now is 26 (again 2LDL\_A is not a complete genome)



JX503079.1  
M93259.1  
M93258.1  
JX503077.1  
JX503078.1  
JX503080.1  
JX503073.1  
JX503074.1  
JX503082.1  
JX503071.1  
JN571034.1  
JX503072.1  
JX503075.1  
JX503081.1  
JX503076.1  
JX503083.1  
AY781127.1  
AY781125.1  
AY781128.1  
AY781126.1  
ON245430.1  
ON245428.1  
MK457954.1  
ON245429.1  
ON245427.1  
ON245431.1