Solutions exercises

- **Transform these SQL statements to mongo statements**

INSERT INTO users(user_id, age, status) VALUES ("bcd001", 45, "A");
db.users.insert({ user_id: "bcd001", age: 45, status: "A" })

SELECT * FROM users WHERE status != "A";
db.users.find({ status: { $ne: "A" } })

SELECT * FROM users WHERE age > 25;
db.users.find({ age: { $gt: 25 } })

SELECT * FROM users WHERE status = "A" ORDER BY user_id DESC;
db.users.find({ status: "A" }).sort({ user_id: -1 })

SELECT COUNT(*) FROM users;
db.users.count()
db.users.find().count()

SELECT COUNT(user_id) FROM users;
db.users.count({ user_id: { $exists: true } }
db.users.find({ user_id: { $exists: true } }).count()

SELECT DISTINCT(status) FROM users;
db.users.distinct("status")

SELECT * FROM users LIMIT 1;
db.users.findOne
db.users.find().limit(1)

SELECT * FROM users LIMIT 5 SKIP 10;
db.users.find().limit(5).skip(10)

UPDATE users SET age = age + 3 WHERE status = "A";
db.users.update({ status: "A" } , { $inc: { age: 3 } }, { multi: true })

DELETE FROM users WHERE status = "D";
db.users.remove({ status: "D" })

- **Import the protein coding genes from file protCodingGenes.bson**

$ mongorestore --collection protCodingGenes --db test protCodingGenes.bson

Return the first 10 genes (alphabetically) on chromosome 22
db.protCodingGenes.find({"Chromosome Name":22},{"Associated Gene Name":1}).sort({ "Associated Gene Name": 1 }).limit(10)

Return the last but one group of 10 genes (by position) on chromosome 12
db.protCodingGenes.find({"Chromosome Name":12},{"Associated Gene Name":1}).sort({ "Gene Start (bp)": -1 }).limit(10).skip(10)

Return the number of unique gene names
db.protCodingGenes.distinct("Associated Gene Name").length

Return the 50 most common genes and their number of occurences

db.protCodingGenes.aggregate([{$group:{_id:"$Associated Gene Name",count:{$sum:1}}},{ $sort : { count : -1} },{ $limit : 50 }])

Return a sorted list of the number of genes per chromosome

db.protCodingGenes.aggregate([{$group:{_id:"$Chromosome Name",count:{$sum:1}}},{ $sort : { count : -1} }])

- **Databases in bioinformatics**

How many nucleotide sequences for collagen genes from nematode worms are there in the NCBI Database?

Nematoda[ORGN] AND collagen 5685

How many mRNA sequences for collagen genes from nematode worms are there in the NCBI Database?

Nematoda[ORGN] AND collagen AND "biomol mRNA"[PROP] 1383

How many protein sequences for collagen proteins from nematode worms are there in the NCBI database?

Nematoda[ORGN] AND collagen 8298

What is the accession number for the Trypanosoma cruzi genome in NCBI?

"Trypanosoma cruzi"[ORGN] NZ_AAHK00000000

How many fully sequenced nematode worm species are represented in the NCBI Genome database?

Nematoda[ORGN] 111

Find the accession number of human beta-globin mRNA sequence. What is the accession number of the encoded protein? How many amino acids does it contain?

NM_000518
NP_000509
147

Find the publication in PubMed with the following ID: 8663200

Use the search record of this publication in PubMed to obtain this data entry in the Nucleotide database

Download the GenBank formatted flatfile

sequence.gb

Find the corresponding record from the previous exercise in the ENA database

Compare the EMBL format to the GenBank format you downloaded before

sequence.txt

How many alternative transcripts are known for Drosophila melanogaster eIF-4E

alternative transcripts 9

Find the human hemoglobin alpha protein in UniprotKB. What is the entry name?

P69905 (HBA_HUMAN) Hemoglobin subunit alpha

How many genes are associated with Huntington Disease (HD), with Alzheimer's disease (AD) and with Parkinson's disease (PD)?

Huntington 1
Alzheimer 6
Parkinson 7

How many transcripts does Ensembl predict for the human gene ACHE?

ACHE 14

Find the mouse orthologue of the human SSBP4. Does this gene have paralogues

Orthologue	*ENSMUSG00000070003*
Paralogues	*ENSMUSG00000061887*
	ENSMUSG00000003992