Lab Exercise 9

1)

A)

Download all **GRCh38** gene annotations from **GENCODE** (you already worked with this dataset in Homework Exercise 3).

Use Pandas to parse it into a DataFrame. Filter only the records of type **transcript**. Extract the Ensembel gene ID of each record, and keep only the gene ID, chromosome and strand fields. Use Pandas's *drop_duplicates* function to keep only one record for each *(gene_id, chr, strand)* combination.

You are supposed to get a DataFrame that looks like:

	index	gene_id	chr	strand
0	1	ENSG00000223972	chr1	+
1	13	ENSG00000227232	chr1	-
2	26	ENSG00000278267	chr1	-
3	29	ENSG00000243485	chr1	+
4	37	ENSG00000284332	chr1	+
5	40	ENSG00000237613	chr1	-
6	48	ENSG00000268020	chr1	+
7	51	ENSG00000240361	chr1	+
8	58	ENSG00000186092	chr1	+
9	77	ENSG00000238009	chr1	-
10	100	ENSG00000239945	chr1	-

B)

Use Pandas grouping to count the number of transcripts in each <chromosome, strand> pair. Sort the chromosome by their relative discrepancy between the two strands (i.e. sort them by $\frac{|N_+ - N_-|}{N_+ + N_-}$ where N_+ and N_- are the number of genes on the positive and negative strands of the chromosome, respectively). Use Pandas to do all the required operations (avoid using Python native data structures).

2)

A)

Download again the **TiGER** database for tissue-specific expression of genes (you worked with it on the last Homework Exercise). Parse the downloaded file into a DataFrame of two columns, **refseq** and **tissue**, where each record is a mapping between one RefSeq ID and one tissue name. If a row in the original file contains multiple tissues, break it down to multiple rows in your DataFrame.

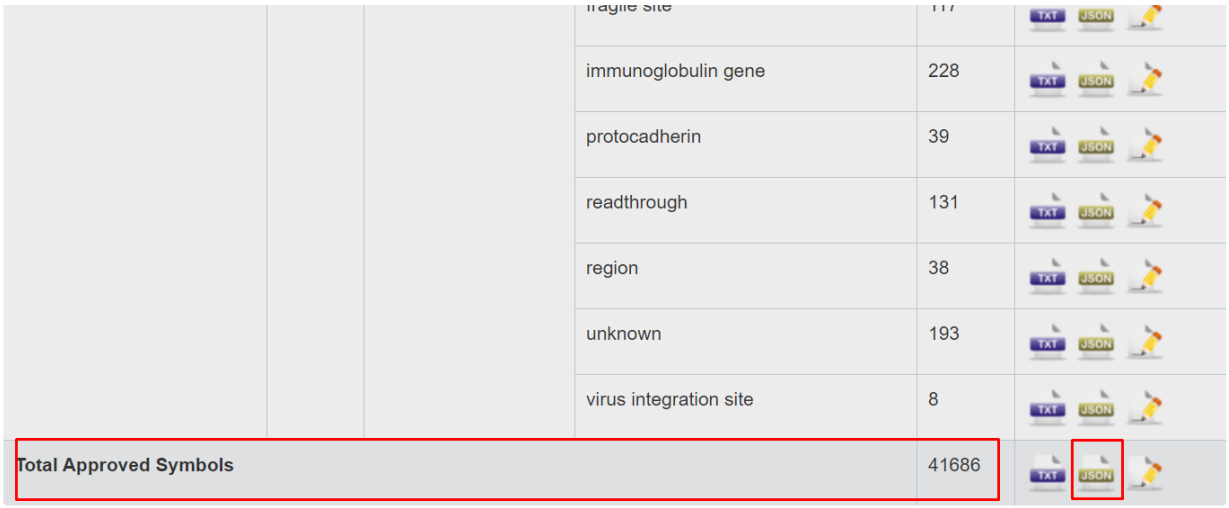
You are supposed to get a DataFrame that looks like:

	refseq	tissue
0	NM_033169	bladder
1	NM_000253	liver
2	NM_000253	small_intestine
3	NM_033168	bladder
4	NM_000252	liver

B)

Go to the **HGNC** database of human gene names: https://www.genenames.org/download/statistics-and-files/.

Download their dataset of all human gene names in JSON format:



This file contains a mapping between various IDs, names and symbols of human genes. In particular it has Ensembel and RefSeq IDs (thus it can be used to map between the two datasets you obtained earlier). It also has UniProt IDs, gene symbols and names, and gene groups.

Parse it into a DataFrame with the following columns:

- symbol
- name
- ensembel_id
- refseq
- uniprot_ids
- gene_groups

If a record in the original file contains multiple RefSeq IDs, break it to multiple rows in your DataFrame. Note that each record can also contain multiple UniProt IDs and groups, so these two columns need to store lists of values.

By the end of the parsing, you are supposed to get a DataFrame that looks like:

	symbol	name	ensembel_id	refseq	uniprot_ids	gene_groups
0	A1BG	alpha-1-B glycoprotein	ENSG00000121410	NM_130786	[P04217]	[Immunoglobulin like domain containing]
1	A1BG-AS1	A1BG antisense RNA 1	ENSG00000268895	NR_015380	[]	[Antisense RNAs]
2	A1CF	APOBEC1 complementation factor	ENSG00000148584	NM_014576	[Q9NQ94]	[RNA binding motif containing]
3	A2M	alpha-2-macroglobulin	ENSG00000175899	NM_000014	[P01023]	[C3 and PZP like, alpha-2-macroglobulin domain...
4	A2M-AS1	A2M antisense RNA 1	ENSG00000245105	NR_026971	[]	[Antisense RNAs]

C)

Merge the three datasets you parsed into a single DataFrame. This DataFrame should map between tissues and chromosomes associated with genes.

It should look like:

	symbol	name	ensembel_id	refseq	uniprot_ids	gene_groups	tissue	index	chr	strand
0	A1BG	alpha-1-B glycoprotein	ENSG00000121410	NM_130786	[P04217]	[Immunoglobulin like domain containing]	liver	2491117	chr19	-
1	A1CF	APOBEC1 complementation factor	ENSG00000148584	NM_014576	[Q9NQ94]	[RNA binding motif containing]	liver	1339593	chr10	-
2	A1CF	APOBEC1 complementation factor	ENSG00000148584	NM_014576	[Q9NQ94]	[RNA binding motif containing]	stomach	1339593	chr10	-
3	A2M	alpha-2-macroglobulin	ENSG00000175899	NM_000014	[P01023]	[C3 and PZP like, alpha-2-macroglobulin domain...	liver	1596480	chr12	-
4	A2ML1	alpha-2-macroglobulin like 1	ENSG00000166535	NM_144670	[A8K2U0]	[C3 and PZP like, alpha-2-macroglobulin domain...	colon	1595751	chr12	+
5	A2ML1	alpha-2-macroglobulin like 1	ENSG00000166535	NM_144670	[A8K2U0]	[C3 and PZP like, alpha-2-macroglobulin domain...	tongue	1595751	chr12	+

D)

Calculate the enrichments for the associations between each chromosome and each tissue, using the formula:

$$enrichment\ factor = \frac{number\ of\ observations}{number\ of\ expected\ observations}$$

In our case, the number of observations is the number of genes shared by the tissue and chromosome. The number of expected observations is the number we would expect to observe if the two were independent (think how to calculate this number). An enrichment factor of 1 indicates no enrichment, an enrichment factor greater than 1 indicates positive enrichment (i.e. more shared genes than expected at random), and an enrichment factor smaller than 1 indicates negative enrichment (i.e. less shared genes than expected at random).

Draw a heatmap showing the associations between all tissues and chromosomes. For the sake of visibility, it is recommended to use a log scale (this means that the threshold between positive and negative enrichments would be 0 instead of 1).

Looking at the resulted figure, which tissues appear enriched in expressing genes on chromosome X?