MET586/MET588 ASSIGNMENT - Python

1. Create a tab-delimited file with the following column headings (in the following order):
   - Geneid
   - dmso. br1
   - dmso.br2
   - lbet151.br1
   - lbet151.br2
   - sgc0946.br1
   - sgc0946.br2

   for the 6 files given, Geneid is taken from the first column and it is the same for each file. The other columns are the last (7^{th}), in each file where the column name is the name of the file with the appendix ".count". Only include genes where the length in column 6 is greater than 1000 bp. An example of what the resulting file would look like:

Geneid	dmso.br1	dmso.br2	lbet151.br1	lbet151.br2	sgc0946.br1	sgc0946.br2
ENSG00000228463	4	2	7	8	7	6
ENSG00000237094	4	1	3	2	3	6
ENSG00000233653	0	0	0	0	0	0
ENSG00000230021	5	5	9	7	10	10
ENSG00000229376	0	0	1	0	0	1
ENSG00000228327	47	46	43	25	64	69
ENSG00000237491	17	9	17	8	23	12
ENSG00000228794	54	62	75	71	85	109
ENSG00000227775	324	202	348	253	304	347
ENSG00000215790	126	114	149	149	159	155
ENSG00000008130	1294	1339	986	1387	1241	1475
ENSG00000078369	6383	6390	5251	6498	4603	5989
ENSG00000169885	4	1	5	3	5	6
ENSG00000142609	4	0	0	0	9	8
ENSG00000158747	4	3	2	7	2	2
ENSG00000162542	244	201	65	72	189	252
ENSG00000169914	258	182	400	306	271	277
ENSG00000127472	0	0	0	0	0	0
ENSG00000117215	4	2	0	1	1	8
ENSG00000158786	1	0	0	1	1	2
ENSG00000227066	1	5	0	0	2	4

2. Create a tab-delimited file with the columns as mentioned above. Only include genes where the counts are greater than 0 (non-zero counts) in all columns and length in column 6 is greater than 2000 bp.

3. Create a tab-delimited file with the following column headings (in the following order):
   - Geneid
   - Gene Name
   - Chromosome
   - Start
   - End
   - Strand

- Description

  from the 6 files given. Retrieve the “protein_coding” information from the ENSEMBL database. Do this for all genes with non-zero counts and length greater than 1000 bp.

The assignment should be submitted electronically to deshpandes1@cardiff.ac.uk with mscbioinformatics@cardiff.ac.uk cc-ed by **5 pm on 16th May 2022**.

This assignment will contribute 40% of the assessment for the module.