Main Assignment 2: This assignment is to test your ability to use UNIX command piping to solve a task.

1) Download GTF file ' Homo_sapiens.GRCh38.87.gtf.gz' from ftp://ftp.ensembl.org/pub/release-87/gtf/homo_sapiens/

2) Make Gene-wise table format with the following information,

Geneid	GeneSymbol	Chromosome	Class	Strand	Length
ENSG00000223972	DDX11L1	chr1:11869-14409	transcribed_unprocessed_pseudogene	+	2540
ENSG00000227232	WASH7P	chr1:14404-29570	unprocessed_pseudogene	-	15166
ENSG00000278267	MIR6859-1	chr1:17369-17436	miRNA	-	67
ENSG00000243485	MIR1302-2	chr1:29554-31109	lincRNA	+	1555
ENSG00000237613	FAM138A	chr1:34554-36081	lincRNA	-	1527
ENSG00000268020	OR4G4P	chr1:52473-53312	unprocessed_pseudogene	+	839
ENSG00000240361	OR4G11P	chr1:62948-63887	unprocessed_pseudogene	+	939
ENSG00000186092	OR4F5	chr1:69091-70008	protein_coding	+	917
ENSG00000238009	RP11-34P13.7	chr1:89295-133723	lincRNA	-	44428
ENSG00000239945	RP11-34P13.8	chr1:89551-91105	lincRNA	-	1554

Main Assignment 3: This assignment will test whether you are able to download a simple bioinformatics tool and explore its usage.

GeneSCF serves as command line tool for pathway or biological process enrichment analysis based on available functional annotation (Geneontology, KEGG, REACTOME and NCG). It requires gene list in the form of Entrez Gene ID (UIDs) or Official gene symbols as a input.

Part I: Using GeneSCF individually on your gene lists

1) Download GeneSCF tool, http://genescf.kandurilab.org/ftp/geneSCF-master-v1.1-p2.tar.gz

2) Explore the help section of the tool and check how to use it

3) Consider number of background protein coding genes for the analysis as 20,000 genes

4) Use the tool to check pathways enriched in two gene lists containing gene symbols provided in the practice session (*practice/gene_lists/H0.list* and *practice/gene_lists/H12.list*) by using KEGG as reference database

5) Use the tool to check pathways enriched in two gene lists containing gene symbols provided in the practice session (*practice/gene_lists/TumorNormal_fc2.list*) by using KEGG as reference database

6) Filter pathways enriched in all three lists using p-value < 0.05 from the GeneSCF output (.TSV file).

Part II: Using GeneSCF on all the list in one-go (geneSCF_batch usage)

Please read the documentation carefully to run GeneSCF on multiple gene lists at once.

Tip:

1) http://genescf.kandurilab.org/faqs.php

2) https://www.biostars.org/p/108669/