1. **Take a Chip-seq library of a transcription factor(TF) and its matching input control. Find peaks and make wiggle track using Dfilter, upload the wiggle track to UCSC genome browser. Call motifs with Homer (findMotifs function) using top 1000 peaks of TF ChIP-seq and perfrom gene ontology analysis using GREAT gene ontology server**

In your folder created in /storage/DSG/your_account_name, create a folder ChIPseq

a) In the course server, the Chip-seq profile of TF and input contol is kept at /storage/vibhor/ChIPseq

ChiP-seq libraries are : GSM288349_E2f1.bed   GSM288356_c-Myc.bed
GSM288355_Esrrb.bed
Matching input control for all: GSM288358_GFP.bed

In your folder just copy one chip-seq library and the input control.
type on console
$source /storage/vibhor/addpath.sh   **(very important to get access to tools installed)**
Then use command
$run_dfilter.sh
Give input chip-seq as
These are TF chip-seq so the parameter would be –bs=50 -ks=20 –nonzero -wig

See the instructions here https://reggenlab.github.io/DFilter/tutorial.html
It will create wiggle track, upload the wiggle track in mm8 version in ucsc browser to visualise Chip-seq data
https://genome-asia.ucsc.edu/cgi-bin/hgGateway?redirect=manual&source=genome.ucsc.edu

a) Paste the screen shot, show certain regions with repeats in mm8 genome using input control or chip-seq track

b. Select the top 1000 peaks or peaks with -log(P-value) > 5. For motif calling
The P-value is in 7th column, you can use awk command

```
$ awk '$7 > 5' peaks.bed > selectedPeaks.bed
```

Using the selected peaks run motif finder of Homer using the command after removing first line in selectedPeaks.bed

```
$findMotifsGenome.pl  selectedPeaks.bed mm8 motif
```

Visualise the output of motif calling, it will give you both denovo motif and known motif enriched in peaks.
Validate that the known motif that you get on top is correct for the give the TF.

c. For gene ontology enrichment use http://great.stanford.edu/public/html/

However here you have to upload only first columns i.e only the genomic locations of selected peaks

Notice these are in mouse genome version mm8 you can lift over the genomic location of mm8 locations to mm9 or mm10 using the lift over tool

https://genome.ucsc.edu/cgi-bin/hgLiftOver

Check if the enriched functions are correct for the used TF using literature

2**. Given a single-cell data-set, visualize it after tSNE based dimension reduction, once you perform this task using all the genes (features). Perform feature selection using approach taught in class. It is the same data-set as was used for practice in class. Do feature selection as taught in the class i.e top genes with high CV2 in each bin where genes were binned according to mean level. Ignore genes with low mean expression.**

Some hints are given here
Download data from

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE81861
Download file : GSE81861_Cell_Line_COUNT.csv.gz

Load the file in R using command

```
 data1 =  read.csv(gzfile('GSE81861_Cell_Line_COUNT.csv.gz’))
```

Get the count data

```
 data = data1[ , 2:562] ;
 rownames(data) = data[,1]  ;       # get the gene names
gme = apply(data, 1, mean) ;  gsd = apply(data,1, sd) ;
plot( log(gme  + 1), log((gsd /gme)*(gsd/gme) + 1)  ) ;  # visualize the mean Vs CV2 plot,
```

Get Principal component decomposition

```
 stepP:  pcs = svd( data, nv = 30) ;
```

Run tSNE, first use library (tSNE) ;

```
ts = tsne(pcs$v) ;
```

Plot tsne coordinates

```
plot( ts[,1] , ts[,2] )
```

Try putting color

```
 cnames = strsplit(colnames(data1) , "__") ;

df = matrix("blue", 561, 1) ;

 for( i in 2:561)
 {
 df[i-1] = (cnames[i])[[1]][3] ;
 }
```

Plot with color by cell type :

```
plot( ts[,1], ts[,2], col=df) ;
```

Now do feature selection to pick highly variant genes repeat from stepP
Do you find a difference

3. Take the single-cell profiles used in question 2 and predict gene-network between top 7000 expressed genes and find gene with highest centrality. What dos this mean. You can use ant method to infer network but keep only 30000-50000 edges.

4. What is the difference and similiarity between linear discriminant analysis (LDA) and ROC-AUC maximising detection filter (hotelling observer , taught in class). Explain where they can be applied with example of type of genomic data?

5. Explain 3 source of bias in ChiP-seq profiles ? How they can be removed during peak calls ?

6. What is the difference between binomial distribution and negative binomial distribution? Explain why negative binomial distribution could be more accurate than Poisson and Gaussian distribution in model RNA-seq read-count of gene across samples.

Does it also apply to ChIP-seq read-count on any genomic site.

Q7. Explain an algorithm for heirarchial clustering.

Q8. Explain one approach for biclustering. What is the cost function which is minimised in biclustering ?