HW4

October 21, 2021

0.1 HW4. RNA and Clustering

0.1.1 In this homework, we will analyze breast cancer RNA-seq data generated by TCGA in order to re-discover the clinically used BRCA subtypes defined by the PAM50 assay. We will be making use of the following files

- EBPlusPlusAdjustPANCAN_IlluminaHiSeq_RNASeqV2.geneExp_BRCAsubset.tsv.gz. This file contains a gene x sample matrix of normalized RNA expression quantifications subset to BRCA samples download from the GDC website https://gdc.cancer.gov/about-data/publications/pancanatlas. See here for a description of their quantification pipeline: https://gdc.cancer.gov/about-data/data-harmonization-and-generation/genomic-data-harmonization/high-level-data-generation/rna-seq-quantification.
- BRCA_clinical.txt. This file gives us clinical information about the patients from which tumors were samples, as reported in the following TCGA paper https://www.cell.com/cancer-cell/pdfExtended/S1535-6108(18)30119-3. Of particular interest is the column "BRCA_Subtype_PAM50" which categorizes BRCA into the following subtypes with distinct expression, clinical properties, and cell of origin
  - Basal
  - Her2
  - LumA
  - LumB
  - Normal-like

Using PCA and the clustering techniques in class, we will characterize the main sources of variation in this dataset.

[1]:
```
import pandas as pd
```

[2]:
```
X = pd.read_csv('EBPlusPlusAdjustPANCAN_IlluminaHiSeq_RNASeqV2.
↪geneExp_BRCAsubset.tsv.gz',sep='\t',index_col=0)
```

[3]:
```
X.iloc[0:5,0:5]
```

[3]:
```
         TCGA-3C-AAAU-01A-11R-A41B-07  TCGA-3C-AALI-01A-11R-A41B-07  \
gene_id
A1BG                          197.090                      237.3840
A1CF                            0.000                        0.0000
A2BP1                           0.000                        0.0000
```

```
A2LD1                         102.963                       70.8646
A2M                          5798.370                     7571.9800

         TCGA-3C-AALJ-01A-31R-A41B-07  TCGA-3C-AALK-01A-11R-A41B-07  \
gene_id
A1BG                         423.2370                      191.0180
A1CF                           0.9066                        0.0000
A2BP1                          0.0000                        0.0000
A2LD1                        161.2600                       62.5072
A2M                         8840.4000                    10960.2000

         TCGA-4H-AAAK-01A-12R-A41B-07
gene_id
A1BG                         268.8810
A1CF                           0.4255
A2BP1                          3.8298
A2LD1                        154.3700
A2M                         9585.4400
```

1 Question 1. Data Exploration

1.1 (1a) Calculate the average expression of genes across the cohort and make a histogram (Hint: for better visualization you may want to apply a simple transform). Based on the observed distribution, choose a cutoff and filter "non-expressed" genes in the peak near zero expression.

1.2 (1b) Using the PAM50 subtype annotations provided in the clinical data, visualize the expression of the following genes stratified by subtype to confirm they are indeed marker genes for the stated subtype

- ERBB2 (HER2)
- ESR1 (LumA/B)
- KRT5 (Basal)

2 Question 2. Dimensionality reduction

2.1 (2a) Run PCA on the (log(x+1)-transformed) data. You can use an implementation such as sklearn, but explain which parts of the result object are the eigenvectors and values of the data's covariance matrix we described in class. Create a scatterplot of the data in PC-space for the first two dimensions colored by PAM50 classification. Do you see a separation of the subtypes?

2.2 (2b) Explore the gene loadings (i.e. weightings in the eigenvector) of the first two components. For each, look up the few highest and lowest scored genes on a site like genecards.org and hypothesize the biological signal underlying the variation in each dimension. Do they have to do with tumor subtype? Or is there some confounding source of variation?

2.3 (2c) Make a plot of the variance explained by each for the first 50 principal components. Using the "elbow method", how many PCs is reasonable to take for downstream analysis?

2.4 (2d) Run one of the non-linear embedding techniques discussed in class (e.g. t-SNE) on your PCA-transformed data, using the number of PCs you determined in 2c. Plot the data colored by PAM50.

3 Question 3. Clustering

3.1 (3a) Cluster the data using one of the techniques discussed in class (k-means, graph-based, etc) using your PCA-transformed data. Is there a number of clusters/resolution that reflects the PAM50 subtypes? Create a visualization to demonstrate the extent of the agreement.

3.2 (3b) Find and plot marker genes for each cluster: Using a t-test, test for each gene whether expression is significantly different between samples that are inside vs outside of the cluster. If you have K clusters and G genes, you should end up with $K * G$ p-values. Find the top 10 marker genes for each cluster (10xK total genes) and plot a heatmap of the expression of these across all the samples. Group samples by their cluster, so you should see blocks of the heat map light up (e.g. the first 10 genes should light up in samples of the first cluster and be dark in the rest, the next 10 high in the second cluster, etc). If you have clusters not corresponding to the PAM50 subtypes, what signal do you think drives them?

[ ]:

```

```