Q2-R

August 15, 2024

1 Q2 - R section

In this section, we use R to perform differential expression analysis for feature selection. We need to determine whether the clinical data are included among our selected features. Additionally, we can also check the position of each clinical data point in the DEGs table to find out if they are really important or not.

[1]:
```
# Import needed libraries
library(limma)
library(edgeR)
```

[2]:
```
# Load Data
normal_counts <- read.csv("train_normal_counts.csv")
meta_data <- read.csv("train_meta_data.csv")
```

[3]:
```
head(normal_counts)
```

A data.frame: 6 × 134

	X35 <dbl>	X80 <dbl>	X190 <dbl>	X187 <dbl>	X129 <dbl>	X12 <dbl>	X78 <dbl>
1	5.820135	6.5462994	6.6040504	6.480745	6.550016	6.5692529	5.9652146
2	-1.060061	0.5821648	-0.8650363	-1.083676	-1.222374	0.7672549	-0.2056643
3	4.388400	3.7520898	4.3514891	4.361634	4.534941	4.1504701	3.1093774
4	4.080172	4.6451746	4.0721368	4.313540	4.370763	4.1660389	4.8092861
5	2.564430	3.8408991	3.1431376	3.120196	3.512952	3.7570108	3.9527651
6	3.552685	3.2010747	4.0374758	1.941859	2.517867	3.2536535	2.7857226

[4]:
```
dim(normal_counts)
```

1. 17400 2. 134

[5]:
```
head(meta_data)
```

A data.frame: 6 × 1

	Simplified_class <chr>
1	Normal
2	Advanced_fibrosis
3	Normal
4	Normal
5	Non_advanced_Fibrosis
6	Normal

[6]:
```
dim(meta_data)
```

1. 134 2. 1

[7]:
```
labels <- factor(meta_data$Simplified_class)
```

[8]:
```
print(labels)
```

```
  [1] Normal                Advanced_fibrosis     Normal
  [4] Normal                Non_advanced_Fibrosis Normal
  [7] Advanced_fibrosis     Non_advanced_Fibrosis Non_advanced_Fibrosis
 [10] Normal                Advanced_fibrosis     Advanced_fibrosis
 [13] Advanced_fibrosis     Non_advanced_Fibrosis Advanced_fibrosis
 [16] Advanced_fibrosis     Normal                Normal
 [19] Normal                Normal                Advanced_fibrosis
 [22] Advanced_fibrosis     Non_advanced_Fibrosis Non_advanced_Fibrosis
 [25] Advanced_fibrosis     Non_advanced_Fibrosis Normal
 [28] Non_advanced_Fibrosis Normal                Advanced_fibrosis
 [31] Advanced_fibrosis     Advanced_fibrosis     Advanced_fibrosis
 [34] Normal                Normal                Non_advanced_Fibrosis
 [37] Non_advanced_Fibrosis Advanced_fibrosis     Advanced_fibrosis
 [40] Non_advanced_Fibrosis Advanced_fibrosis     Advanced_fibrosis
 [43] Advanced_fibrosis     Advanced_fibrosis     Non_advanced_Fibrosis
 [46] Non_advanced_Fibrosis Normal                Advanced_fibrosis
 [49] Advanced_fibrosis     Advanced_fibrosis     Non_advanced_Fibrosis
 [52] Advanced_fibrosis     Normal                Non_advanced_Fibrosis
 [55] Normal                Advanced_fibrosis     Advanced_fibrosis
 [58] Non_advanced_Fibrosis Normal                Normal
 [61] Normal                Non_advanced_Fibrosis Advanced_fibrosis
 [64] Advanced_fibrosis     Normal                Normal
 [67] Advanced_fibrosis     Normal                Advanced_fibrosis
 [70] Advanced_fibrosis     Advanced_fibrosis     Advanced_fibrosis
 [73] Normal                Advanced_fibrosis     Non_advanced_Fibrosis
 [76] Normal                Advanced_fibrosis     Advanced_fibrosis
 [79] Non_advanced_Fibrosis Non_advanced_Fibrosis Advanced_fibrosis
 [82] Normal                Advanced_fibrosis     Non_advanced_Fibrosis
 [85] Non_advanced_Fibrosis Advanced_fibrosis     Normal
 [88] Non_advanced_Fibrosis Advanced_fibrosis     Non_advanced_Fibrosis
 [91] Advanced_fibrosis     Normal                Normal
 [94] Advanced_fibrosis     Non_advanced_Fibrosis Non_advanced_Fibrosis
 [97] Non_advanced_Fibrosis Advanced_fibrosis     Advanced_fibrosis
[100] Advanced_fibrosis     Normal                Normal
[103] Non_advanced_Fibrosis Non_advanced_Fibrosis Normal
[106] Normal                Advanced_fibrosis     Non_advanced_Fibrosis
[109] Non_advanced_Fibrosis Normal                Advanced_fibrosis
[112] Non_advanced_Fibrosis Normal                Normal
[115] Advanced_fibrosis     Normal                Normal
[118] Non_advanced_Fibrosis Advanced_fibrosis     Non_advanced_Fibrosis
```

```
[121] Normal                 Advanced_fibrosis     Non_advanced_Fibrosis
[124] Non_advanced_Fibrosis Non_advanced_Fibrosis Normal
[127] Advanced_fibrosis     Non_advanced_Fibrosis Advanced_fibrosis
[130] Normal                 Normal                Normal
[133] Normal                 Normal
Levels: Advanced_fibrosis Non_advanced_Fibrosis Normal
```

Let's perform DE analysis

[9]:
```
# Create a design matrix
design <- model.matrix(~0 + labels)
colnames(design) <- levels(labels)
```

[10]:
```
fit <- lmFit(normal_counts, design)
```

[11]:
```
contrast.matrix <- makeContrasts(
    AdvancedFibrosis_vs_Normal = `Advanced_fibrosis` - Normal,
    Fibrosis_vs_Normal = Non_advanced_Fibrosis - Normal,
    AdvancedFibrosis_vs_Fibrosis = `Advanced_fibrosis` - Non_advanced_Fibrosis,
    levels = design
)

# Apply contrasts to the fit
fit2 <- contrasts.fit(fit, contrast.matrix)

# Empirical Bayes moderation to get p-values
fit2 <- eBayes(fit2)
```

Now, we are going to extract the DEGs for each pair of classes and save them

[12]:
```
# Get the top DEGs for the Advanced Fibrosis vs Normal comparison
top_genes_adv_vs_norm <- topTable(fit2, coef = "AdvancedFibrosis_vs_Normal",␣
 ↪adjust.method = "BH", number = Inf)

# Get the top DEGs for the Fibrosis vs Normal comparison
top_genes_fib_vs_norm <- topTable(fit2, coef = "Fibrosis_vs_Normal", adjust.
 ↪method = "BH", number = Inf)

# Get the top DEGs for the Advanced Fibrosis vs Fibrosis comparison
top_genes_adv_vs_fib <- topTable(fit2, coef = "AdvancedFibrosis_vs_Fibrosis",␣
 ↪adjust.method = "BH", number = Inf)

# View the top DEGs
head(top_genes_adv_vs_norm)
head(top_genes_fib_vs_norm)
head(top_genes_adv_vs_fib)
```

A data.frame: 6 × 6

	logFC <dbl>	AveExpr <dbl>	t <dbl>	P.Value <dbl>	adj.P.Val <dbl>	B <dbl>
10728	-1.3278914	1.902497	-10.235955	1.733518e-18	3.016320e-14	31.41965
13385	1.0088994	6.051398	9.886435	1.300002e-17	9.022952e-14	29.46381
10694	-1.3132140	3.432308	-9.855214	1.555681e-17	9.022952e-14	29.28951
16113	-3.4233202	-1.392599	-9.734869	3.105810e-17	1.351027e-13	28.61830
16278	-2.8530739	-0.172656	-9.680411	4.244891e-17	1.477222e-13	28.31497
6969	0.4379282	4.207436	9.600175	6.723208e-17	1.949730e-13	27.86851

A data.frame: 6 × 6

	logFC <dbl>	AveExpr <dbl>	t <dbl>	P.Value <dbl>	adj.P.Val <dbl>	B <dbl>
13623	0.7946468	5.255822	11.73977	2.832278e-22	4.928163e-18	40.01856
10970	0.6321228	5.361546	11.47198	1.340291e-21	1.022312e-17	38.49799
5442	-1.4931333	2.180779	-11.38733	2.190866e-21	1.022312e-17	38.01721
17075	0.6202635	7.265926	11.37524	2.350142e-21	1.022312e-17	37.94855
4461	0.6782287	6.569435	11.29232	3.803564e-21	1.323640e-17	37.47748
6563	0.6004911	5.545244	11.13142	9.679073e-21	2.806931e-17	36.56355

A data.frame: 6 × 6

	logFC <dbl>	AveExpr <dbl>	t <dbl>	P.Value <dbl>	adj.P.Val <dbl>	B <dbl>
16863	1.223013	2.9208664	9.097264	1.182032e-15	2.056735e-11	24.90951
3296	1.594761	1.9955120	8.122844	2.755791e-13	1.021895e-09	19.70536
673	1.485822	1.1568774	8.121638	2.774145e-13	1.021895e-09	19.69902
14913	-1.153029	2.6115748	-8.115173	2.874656e-13	1.021895e-09	19.66504
12060	1.675825	0.1951384	8.111307	2.936480e-13	1.021895e-09	19.64473
3227	1.279852	4.3441375	8.016340	4.947725e-13	1.434840e-09	19.14666

[13]:
```
write.csv(top_genes_adv_vs_norm, "DEGs_AdvancedFibrosis_vs_Normal.csv")
write.csv(top_genes_fib_vs_norm, "DEGs_Fibrosis_vs_Normal.csv")
write.csv(top_genes_adv_vs_fib, "DEGs_AdvancedFibrosis_vs_Fibrosis.csv")
```

We have filtered the top 200 DEGs for each pair. The choice of =200 appears to be optimized based on our greedy search, which has not been included in this notebook.

[14]:
```
filtered_genes_adv_vs_norm <- top_genes_adv_vs_norm[1:200,]
filtered_genes_fib_vs_norm <- top_genes_fib_vs_norm[1:200,]
filtered_genes_adv_vs_fib <- top_genes_adv_vs_fib[1:200,]

# View filtered DEGs
head(filtered_genes_adv_vs_norm)
head(filtered_genes_fib_vs_norm)
head(filtered_genes_adv_vs_fib)
```

A data.frame: 6 × 6

	logFC <dbl>	AveExpr <dbl>	t <dbl>	P.Value <dbl>	adj.P.Val <dbl>	B <dbl>
10728	-1.3278914	1.902497	-10.235955	1.733518e-18	3.016320e-14	31.41965
13385	1.0088994	6.051398	9.886435	1.300002e-17	9.022952e-14	29.46381
10694	-1.3132140	3.432308	-9.855214	1.555681e-17	9.022952e-14	29.28951
16113	-3.4233202	-1.392599	-9.734869	3.105810e-17	1.351027e-13	28.61830
16278	-2.8530739	-0.172656	-9.680411	4.244891e-17	1.477222e-13	28.31497
6969	0.4379282	4.207436	9.600175	6.723208e-17	1.949730e-13	27.86851

A data.frame: 6 × 6

	logFC <dbl>	AveExpr <dbl>	t <dbl>	P.Value <dbl>	adj.P.Val <dbl>	B <dbl>
13623	0.7946468	5.255822	11.73977	2.832278e-22	4.928163e-18	40.01856
10970	0.6321228	5.361546	11.47198	1.340291e-21	1.022312e-17	38.49799
5442	-1.4931333	2.180779	-11.38733	2.190866e-21	1.022312e-17	38.01721
17075	0.6202635	7.265926	11.37524	2.350142e-21	1.022312e-17	37.94855
4461	0.6782287	6.569435	11.29232	3.803564e-21	1.323640e-17	37.47748
6563	0.6004911	5.545244	11.13142	9.679073e-21	2.806931e-17	36.56355

A data.frame: 6 × 6

	logFC <dbl>	AveExpr <dbl>	t <dbl>	P.Value <dbl>	adj.P.Val <dbl>	B <dbl>
16863	1.223013	2.9208664	9.097264	1.182032e-15	2.056735e-11	24.90951
3296	1.594761	1.9955120	8.122844	2.755791e-13	1.021895e-09	19.70536
673	1.485822	1.1568774	8.121638	2.774145e-13	1.021895e-09	19.69902
14913	-1.153029	2.6115748	-8.115173	2.874656e-13	1.021895e-09	19.66504
12060	1.675825	0.1951384	8.111307	2.936480e-13	1.021895e-09	19.64473
3227	1.279852	4.3441375	8.016340	4.947725e-13	1.434840e-09	19.14666

[15]:
```
dim(filtered_genes_adv_vs_norm)
dim(filtered_genes_fib_vs_norm)
dim(filtered_genes_adv_vs_fib)
```

1. 200 2. 6

1. 200 2. 6

1. 200 2. 6

[16]:
```
genes_adv_vs_norm_names <- rownames(filtered_genes_adv_vs_norm)
genes_fib_vs_norm_names <- rownames(filtered_genes_fib_vs_norm)
genes_adv_vs_fib_names <- rownames(filtered_genes_adv_vs_fib)
```

then we combined the filtered DEGs to create a new feature space

[17]:
```
combined_gene_names <- unique(c(genes_adv_vs_norm_names,
                                genes_fib_vs_norm_names,
                                genes_adv_vs_fib_names))
```

[18]:
```
length(combined_gene_names)
```

527

[19]:
```
common_genes <- intersect(rownames(normal_counts), combined_gene_names)
selected_normal_counts <- normal_counts[common_genes, ]
head(selected_normal_counts)
```

A data.frame: 6 × 134

	X35 <dbl>	X80 <dbl>	X190 <dbl>	X187 <dbl>	X129 <dbl>	X12 <dbl>	X7 <d
10	4.5895546	5.4821690	5.01315395	5.0660709	4.6558168	4.29907767	4.7
57	-0.7190239	0.5821648	-1.15454295	-1.0836762	-0.2894877	-0.00646928	-1.1
265	0.5876374	-2.2251901	-0.07654044	0.1964317	0.2723912	2.52315019	1.4
275	1.6624052	2.5296974	1.50842206	1.9916119	1.4679420	1.93415217	2.0
278	5.5299023	5.8462723	5.55358296	5.3592673	5.7745493	4.96266190	5.2
297	7.5494873	7.7592284	7.61253797	8.0700567	7.4394046	7.67164807	7.4

[20]:
```
dim(selected_normal_counts)
```

1. 527 2. 134

We extracted a subset from the data based on selected features. Let's save it and continue the analysis in Python Jupyter Notebook and check if clinical data have been selected or not

[21]:
```
write.csv(selected_normal_counts, "subset_data.csv")
```