

# Embedding Visualizations

Vivek Sriram

2023-06-07

```
library(ggplot2)

## Warning: replacing previous import 'ellipsis::check_dots_unnamed' by
## 'rlang::check_dots_unnamed' when loading 'tibble'

## Warning: replacing previous import 'ellipsis::check_dots_used' by
## 'rlang::check_dots_used' when loading 'tibble'

## Warning: replacing previous import 'ellipsis::check_dots_empty' by
## 'rlang::check_dots_empty' when loading 'tibble'

## Warning: replacing previous import 'ellipsis::check_dots_unnamed' by
## 'rlang::check_dots_unnamed' when loading 'pillar'

## Warning: replacing previous import 'ellipsis::check_dots_used' by
## 'rlang::check_dots_used' when loading 'pillar'

## Warning: replacing previous import 'ellipsis::check_dots_empty' by
## 'rlang::check_dots_empty' when loading 'pillar'

library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##   filter, lag

## The following objects are masked from 'package:base':
##
##   intersect, setdiff, setequal, union

library(umap)
library(gridExtra)

##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##   combine

library(ggpubr)

## Warning: replacing previous import 'ellipsis::check_dots_unnamed' by
## 'rlang::check_dots_unnamed' when loading 'hms'

## Warning: replacing previous import 'ellipsis::check_dots_used' by
```

```
## 'rlang::check_dots_used' when loading 'hms'

## Warning: replacing previous import 'ellipsis::check_dots_empty' by
## 'rlang::check_dots_empty' when loading 'hms'
```

## Read in input data files

Includes k-means clustering results with k=5

```
embeddingMat_m = read.csv("/Users/viveksriram/Desktop/embeddingMat_m_fixed.csv")
embeddingMat_m$kmeansLabel = as.factor(embeddingMat_m$kmeansLabel)

embeddingMat_f = read.csv("/Users/viveksriram/Desktop/embeddingMat_f_fixed.csv")
embeddingMat_f$kmeansLabel = as.factor(embeddingMat_f$kmeansLabel)

#interestingDiseases = c("C64-C68", "F20-F29", "I20-I25", "L40-L45", "M05-M14", "M15-M19", "M45-M49")
interestingDiseases = c("M05-M14", "G35-G37", "C43-C44", "E70-E90")

embeddingMat_f$toBeLabeled = ifelse(embeddingMat_f$Trait %in% interestingDiseases, 1, 0)
embeddingMat_m$toBeLabeled = ifelse(embeddingMat_m$Trait %in% interestingDiseases, 1, 0)
```

Female M05 - M14: Inflammatory polyarthropathies (rheumatoid arthritis) M15 - M19: Osteoarthritis M80 - M85: Osteoporosis L80 - L99: Scleroderma G35 - G37: Multiple sclerosis

Male M45 - M49: Ankylosing spondylitis I20 - I25: Ischemic heart diseases C43 - C44: Melanoma E70 - E90: Metabolic disorders L40 - L45: Psoriasis

## PCA

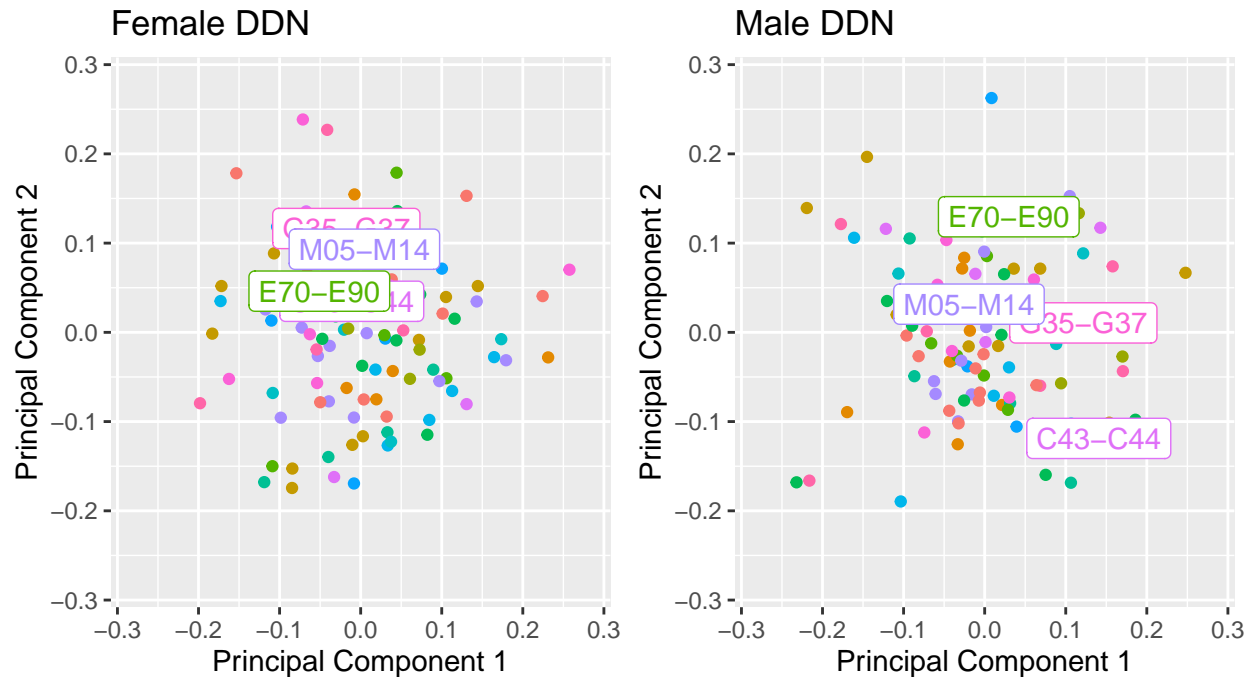
```
pca_category_m = ggplot(embeddingMat_m,
  aes(x=pca.2d.one,
      y=pca.2d.two,
      color=Disease.Category,
      label=Trait)) +
  geom_point() +
  geom_label(aes(pca.2d.one, pca.2d.two, label=Trait),
    data = embeddingMat_m[embeddingMat_m$Trait %in% interestingDiseases,]) +
  labs(title="Male DDN",
    x="Principal Component 1",
    y = "Principal Component 2") +
  xlim(-0.28, 0.28) +
  ylim(-0.28, 0.28)
```

```
pca_category_f = ggplot(embeddingMat_f,
  aes(x=pca.2d.one,
      y=pca.2d.two,
      color=Disease.Category)) +
  geom_point() +
  geom_label(aes(pca.2d.one, pca.2d.two, label=Trait),
    data = embeddingMat_f[embeddingMat_f$Trait %in% interestingDiseases,]) +
  labs(title="Female DDN",
    x="Principal Component 1",
    y = "Principal Component 2") +
  xlim(-0.28, 0.28) +
```

```
ylim(-0.28, 0.28)
```

```
ggarrange(pca_category_f, pca_category_m,  
          common.legend = TRUE)
```

a Circulatory System	a Ear Disorders	a Genitourinary	a Mental and Behavioral Disorders
a Dermatologic	a Endocrine/Metabolic	a Hematopoietic	a Musculoskeletal
a Digestive	a Eye Disorders	a Infectious Diseases	a Neoplasms



```
pca_kmeans_m = ggplot(embeddingMat_m,  
  aes(x=pca.2d.one,  
       y=pca.2d.two,  
       color=kmeansLabel)) +  
  geom_point() +  
  geom_label(aes(pca.2d.one, pca.2d.two, label=Trait),  
             data = embeddingMat_m[embeddingMat_m$Trait %in% interestingDiseases,]) +  
  labs(title="Male DDN",  
       x="Principal Component 1",  
       y = "Principal Component 2") +  
  xlim(-0.28, 0.28) +  
  ylim(-0.28, 0.28)
```

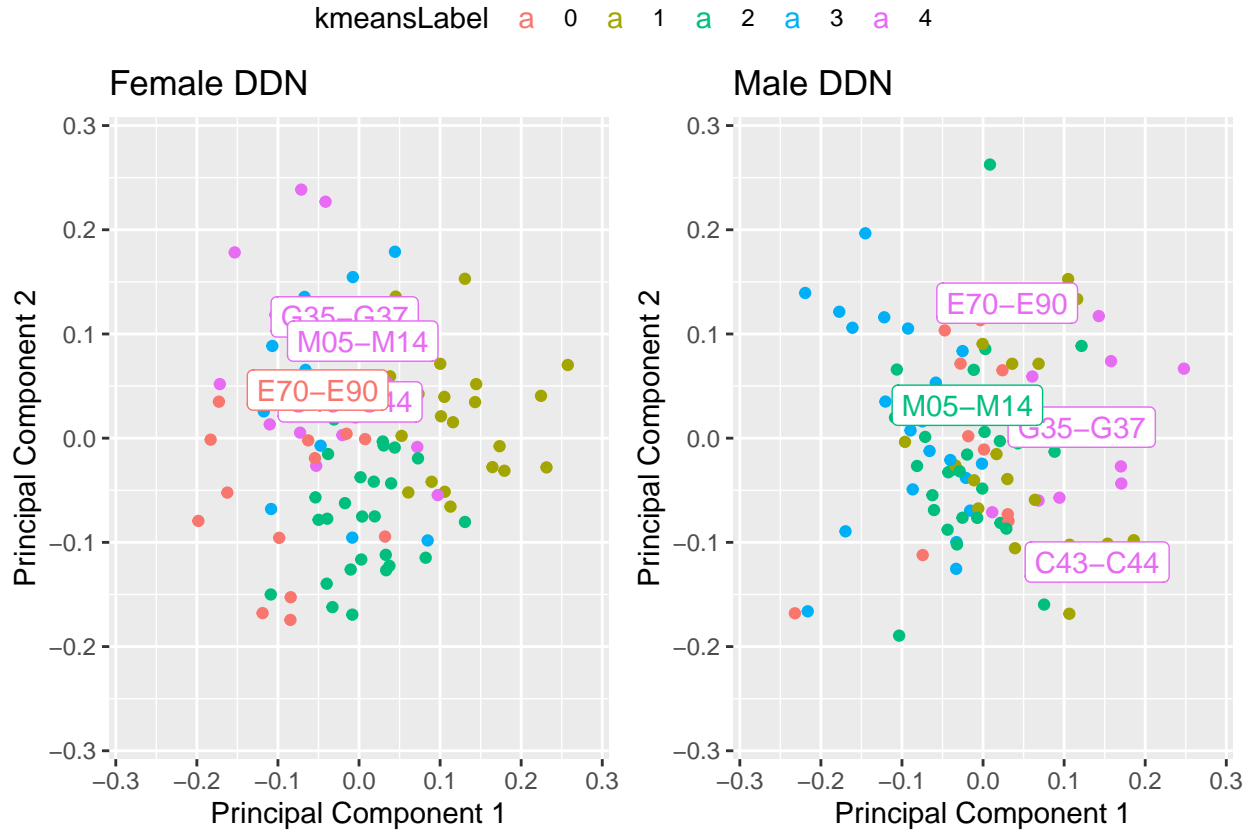
```
pca_kmeans_f = ggplot(embeddingMat_f,  
  aes(x=pca.2d.one,  
       y=pca.2d.two,  
       color=kmeansLabel)) +  
  geom_point() +  
  geom_label(aes(pca.2d.one, pca.2d.two, label=Trait),  
             data = embeddingMat_f[embeddingMat_f$Trait %in% interestingDiseases,]) +  
  labs(title="Female DDN",  
       x="Principal Component 1",
```

```

y = "Principal Component 2") +
xlim(-0.28, 0.28) +
ylim(-0.28, 0.28)

```

```
ggarrange(pca_kmeans_f, pca_kmeans_m, common.legend = TRUE)
```



## UMAP

```

embeddingData_f = embeddingMat_f %>%
  select(X0, X1, X2, X3, X4, X5, X6, X7, X8, X9, X10, X11, X12, X13, X14, X15)

```

```

embeddingData_m = embeddingMat_m %>%
  select(X0, X1, X2, X3, X4, X5, X6, X7, X8, X9, X10, X11, X12, X13, X14, X15)

```

```

embeddingDataM.umap <- umap(embeddingData_m, n_neighbors = 3)
embeddingDataF.umap <- umap(embeddingData_f, n_neighbors = 3)

```

```

embeddingMat_m$umapDim1 = embeddingDataM.umap$layout[,1]
embeddingMat_m$umapDim2 = embeddingDataM.umap$layout[,2]

```

```

embeddingMat_f$umapDim1 = embeddingDataF.umap$layout[,1]
embeddingMat_f$umapDim2 = embeddingDataF.umap$layout[,2]

```

```

umap_category_f = ggplot(embeddingMat_f,
  aes(x=umapDim1,
    y=umapDim2,

```

```

    color=Disease.Category)) +
  geom_point() +
  geom_label(aes(umapDim1, umapDim2, label=Trait),
    data = embeddingMat_f[embeddingMat_f$Trait %in% interestingDiseases,]) +
  labs(title="Female DDN",
    x="UMAP Dimension 1",
    y = "UMAP Dimension 2") +
  xlim(-5, 5) +
  ylim(-5,5)

```

```

umap_category_m = ggplot(embeddingMat_m,
  aes(x=umapDim1,
    y=umapDim2,
    color=Disease.Category,
    label=Trait)) +
  geom_point() +
  geom_label(aes(umapDim1, umapDim2, label=Trait),
    data = embeddingMat_m[embeddingMat_m$Trait %in% interestingDiseases,]) +
  labs(title="Male DDN",
    x="UMAP Dimension 1",
    y = "UMAP Dimension 2") +
  xlim(-5, 5) +
  ylim(-5,5)

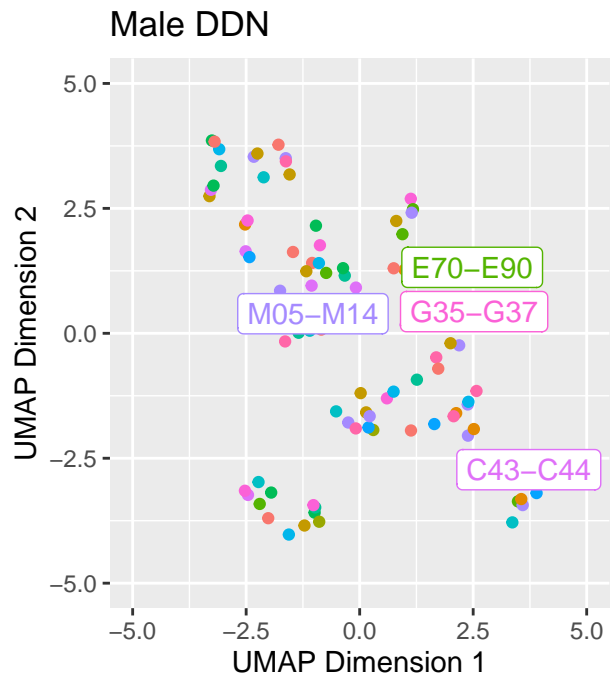
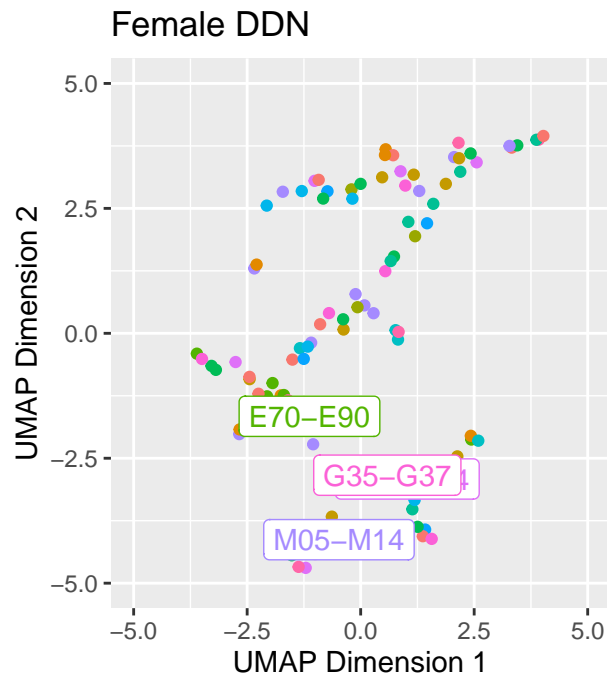
```

```

ggarrange(umap_category_f, umap_category_m, common.legend = TRUE)

```

a Circulatory System	a Ear Disorders	a Genitourinary	a Mental and Behavioral Disorders
a Dermatologic	a Endocrine/Metabolic	a Hematopoietic	a Musculoskeletal
a Digestive	a Eye Disorders	a Infectious Diseases	a Neoplasms



```
umap_kmeans_f = ggplot(embeddingMat_f,
  aes(x=umapDim1,
      y=umapDim2,
      color=kmeansLabel)) +
  geom_point() +
  geom_label(aes(umapDim1, umapDim2, label=Trait),
    data = embeddingMat_f[embeddingMat_f$Trait %in% interestingDiseases,]) +
  labs(title="Female DDN",
    x="UMAP Dimension 1",
    y = "UMAP Dimension 2") +
  xlim(-5, 5) +
  ylim(-5,5)
```

```
umap_kmeans_m = ggplot(embeddingMat_m,
  aes(x=umapDim1,
      y=umapDim2,
      color=kmeansLabel)) +
  geom_point() +
  geom_label(aes(umapDim1, umapDim2, label=Trait),
    data = embeddingMat_m[embeddingMat_m$Trait %in% interestingDiseases,]) +
  labs(title="Male DDN",
    x="UMAP Dimension 1",
    y = "UMAP Dimension 2") +
  xlim(-5, 5) +
  ylim(-5,5)
```

```
ggarrange(umap_kmeans_f, umap_kmeans_m, common.legend = TRUE)
```

kmeansLabel   a   0   a   1   a   2   a   3   a   4

