Large Super SOM

```
library(ggplot2)
library(reshape)
library(plyr)
```

```
##
## Attaching package: 'plyr'
##
## The following objects are masked from 'package:reshape':
##
##     rename, round_any
```

```
library(kohonen)
library(goseq)
library(GO.db)
```

```
mostDEgenes <- read.csv("../data/allGeneListBothGenotypes_analysis5b.csv")

mostDEgenes <- mostDEgenes[c(7, 2, 1, 4)] #keep only needed columns (gene, genotype, type, mean)

head(mostDEgenes)
```

```
##                 gene genotype   type    mean
## 1 Solyc00g014800.1.1      tf2   Ambr  75.159
## 2 Solyc00g014800.1.1       wt   Ambr   8.643
## 3 Solyc00g014800.1.1      tf2 Aother  15.792
## 4 Solyc00g014800.1.1       wt Aother   3.723
## 5 Solyc00g014800.1.1      tf2   Bmbr 124.304
## 6 Solyc00g014800.1.1       wt   Bmbr  57.467
```

```
#Change from long to wide data format
mostDEgene.long <- cast(mostDEgenes, genotype + gene ~ type, value.var = mean, fun.aggregate = "mean")
```

```
## Using mean as value column.  Use the value argument to cast to override this choice
```

```
head(mostDEgene.long)
```

```
##   genotype               gene   Ambr  Aother    Bmbr Bother   Cmbr Cother
## 1      tf2 Solyc00g005050.2.1  9.526  1.2970   3.964 11.025  9.458  6.843
## 2      tf2 Solyc00g005070.1.1 16.175 14.2026 158.811  4.480 11.542  3.108
## 3      tf2 Solyc00g005080.1.1 11.796  7.7876  15.482  8.519 11.464  7.041
## 4      tf2 Solyc00g005840.2.1 13.585 44.2406   7.508 19.328 10.452 29.709
## 5      tf2 Solyc00g005870.1.1  6.110  0.5291  37.612  1.456  2.344  1.564
## 6      tf2 Solyc00g005880.1.1  1.840  1.1236  61.639  2.036  5.711  1.787
```

```
mostDEgene.long <- as.data.frame(mostDEgene.long)
names(mostDEgene.long)
```

```
## [1] "genotype" "gene"     "Ambr"     "Aother"   "Bmbr"     "Bother"
## [7] "Cmbr"     "Cother"
```

```
scale_data <- as.matrix(t(scale(t(mostDEgene.long[c(3:8)]))))
head(scale_data)
```

```
##      Ambr  Aother    Bmbr  Bother    Cmbr   Cother
## 1  0.6682 -1.5250 -0.8142  1.0678  0.6501 -0.04691
## 2 -0.3039 -0.3363  2.0338 -0.4956 -0.3799 -0.51810
## 3  0.4554 -0.8054  1.6148 -0.5755  0.3509 -1.04014
## 4 -0.5191  1.6854 -0.9562 -0.1061 -0.7444  0.64040
## 5 -0.1488 -0.5336  2.0229 -0.4697 -0.4085 -0.46226
## 6 -0.4346 -0.4642  2.0366 -0.4265 -0.2746 -0.43675
```

```
#Principle Component Analysis
pca <- prcomp(scale_data, scale=TRUE)

summary(pca)
```

```
## Importance of components:
##                          PC1   PC2   PC3   PC4   PC5      PC6
## Standard deviation     1.306 1.125 1.036 1.021 0.956 2.89e-15
## Proportion of Variance 0.284 0.211 0.179 0.174 0.152 0.00e+00
## Cumulative Proportion  0.284 0.495 0.674 0.848 1.000 1.00e+00
```

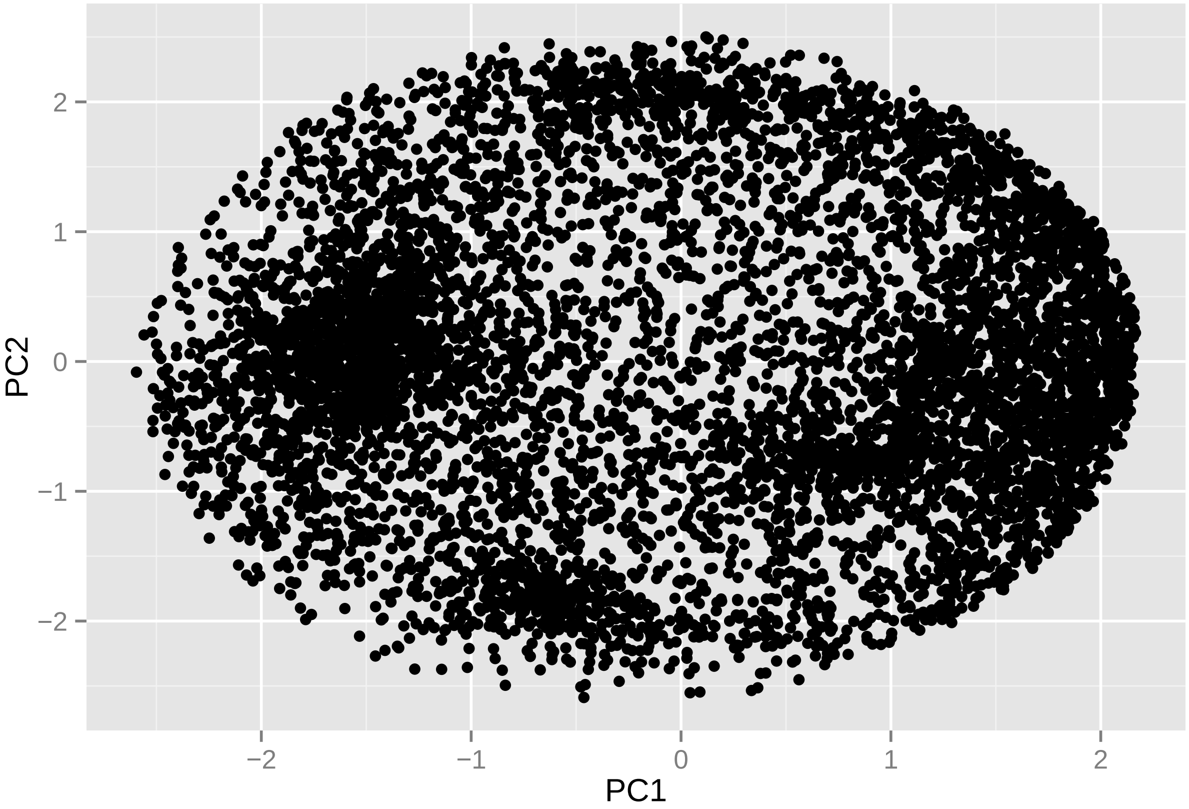
```
pca.scores <- data.frame(pca$x)

data.val <- cbind(mostDEgene.long, scale_data, pca.scores)

head(data.val)
```

```
##   genotype              gene   Ambr  Aother    Bmbr Bother   Cmbr Cother
## 1      tf2 Solyc00g005050.2.1  9.526  1.2970   3.964 11.025  9.458  6.843
## 2      tf2 Solyc00g005070.1.1 16.175 14.2026 158.811  4.480 11.542  3.108
## 3      tf2 Solyc00g005080.1.1 11.796  7.7876  15.482  8.519 11.464  7.041
## 4      tf2 Solyc00g005840.2.1 13.585 44.2406   7.508 19.328 10.452 29.709
## 5      tf2 Solyc00g005870.1.1  6.110  0.5291  37.612  1.456  2.344  1.564
## 6      tf2 Solyc00g005880.1.1  1.840  1.1236  61.639  2.036  5.711  1.787
##      Ambr  Aother    Bmbr  Bother    Cmbr   Cother     PC1      PC2
## 1  0.6682 -1.5250 -0.8142  1.0678  0.6501 -0.04691  0.2265 -1.62837
## 2 -0.3039 -0.3363  2.0338 -0.4956 -0.3799 -0.51810 -1.6467 -0.09637
## 3  0.4554 -0.8054  1.6148 -0.5755  0.3509 -1.04014 -2.2767 -0.74388
## 4 -0.5191  1.6854 -0.9562 -0.1061 -0.7444  0.64040  1.0796  1.71915
## 5 -0.1488 -0.5336  2.0229 -0.4697 -0.4085 -0.46226 -1.6906 -0.23662
## 6 -0.4346 -0.4642  2.0366 -0.4265 -0.2746 -0.43675 -1.5295 -0.32074
##       PC3     PC4     PC5        PC6
## 1  0.3867  1.5784  1.0599  1.485e-15
## 2 -0.8791 -1.3966  0.8451 -1.665e-15
## 3 -0.1000 -0.4977  0.9484 -2.887e-15
## 4  0.2205 -0.2686 -1.0926  1.110e-16
## 5 -1.0235 -1.2020  0.8960 -1.499e-15
## 6 -0.8679 -1.4646  0.8819 -1.332e-15
```

Visualizing the PCA

Looks to be three major clusters.

```
p <- ggplot(data.val, aes(PC1, PC2))
p + geom_point()
```



SuperSOM

```
set.seed(6)
names(data.val)
```

```
##  [1] "genotype" "gene"     "Ambr"     "Aother"   "Bmbr"     "Bother"  
##  [7] "Cmbr"     "Cother"   "Ambr"     "Aother"   "Bmbr"     "Bother"  
## [13] "Cmbr"     "Cother"   "PC1"      "PC2"      "PC3"      "PC4"     
## [19] "PC5"      "PC6"
```

```
superSomData <- data.val[,c(1:8)]

tf2 <- subset(superSomData, genotype == "tf2", select = 3:8)
wt <- subset(superSomData, genotype == "wt", select = 3:8)

wt <- as.matrix(wt)
tf2 <- as.matrix(tf2)
```

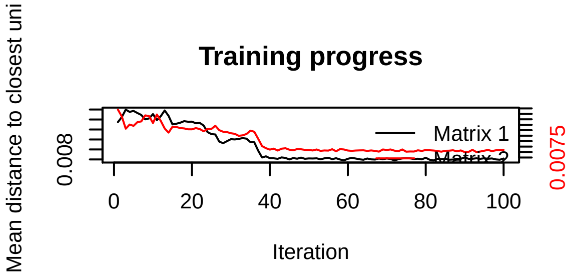
```
sc.wt <- t(scale(t(wt)))
sc.tf2 <- t(scale(t(tf2)))

all.data <- list(sc.wt,sc.tf2)

ssom <- supersom(all.data, somgrid(6, 6, "hexagonal"),weights=c(0.5,0.5))

summary(ssom)
```

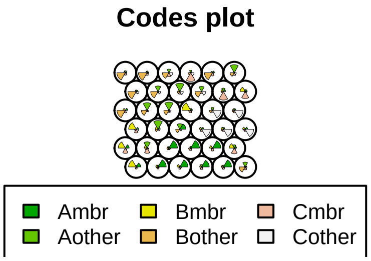
```
## supersom map of size 6x6 with a hexagonal topology.
## Training data included of  3580 objects
## The number of layers is 2
## Mean distance to the closest unit in the map: 0.03723
```

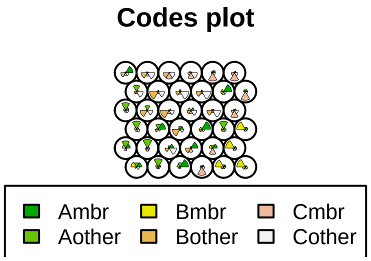
```
par(mfrow = c(3, 2))
plot(ssom, type ="changes")
plot(ssom, type = "codes")
plot(ssom, type = "counts")
plot(ssom, type = "quality")

data.val <- cbind(data.val,ssom$unit.classif,ssom$distances)

head(data.val)
```

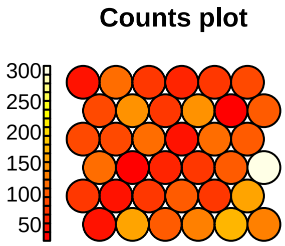
```
##   genotype               gene   Ambr  Aother    Bmbr Bother   Cmbr Cother
## 1      tf2 Solyc00g005050.2.1  9.526  1.2970   3.964 11.025  9.458  6.843
## 2      tf2 Solyc00g005070.1.1 16.175 14.2026 158.811  4.480 11.542  3.108
## 3      tf2 Solyc00g005080.1.1 11.796  7.7876  15.482  8.519 11.464  7.041
## 4      tf2 Solyc00g005840.2.1 13.585 44.2406   7.508 19.328 10.452 29.709
## 5      tf2 Solyc00g005870.1.1  6.110  0.5291  37.612  1.456  2.344  1.564
## 6      tf2 Solyc00g005880.1.1  1.840  1.1236  61.639  2.036  5.711  1.787
##      Ambr  Aother    Bmbr  Bother    Cmbr   Cother     PC1      PC2
## 1  0.6682 -1.5250 -0.8142  1.0678  0.6501 -0.04691  0.2265 -1.62837
## 2 -0.3039 -0.3363  2.0338 -0.4956 -0.3799 -0.51810 -1.6467 -0.09637
## 3  0.4554 -0.8054  1.6148 -0.5755  0.3509 -1.04014 -2.2767 -0.74388
## 4 -0.5191  1.6854 -0.9562 -0.1061 -0.7444  0.64040  1.0796  1.71915
## 5 -0.1488 -0.5336  2.0229 -0.4697 -0.4085 -0.46226 -1.6906 -0.23662
## 6 -0.4346 -0.4642  2.0366 -0.4265 -0.2746 -0.43675 -1.5295 -0.32074
##       PC3     PC4     PC5        PC6 ssom$unit.classif ssom$distances
## 1  0.3867  1.5784  1.0599  1.485e-15                33       0.070370
## 2 -0.8791 -1.3966  0.8451 -1.665e-15                18       0.002183
## 3 -0.1000 -0.4977  0.9484 -2.887e-15                18       0.039053
## 4  0.2205 -0.2686 -1.0926  1.110e-16                25       0.014707
## 5 -1.0235 -1.2020  0.8960 -1.499e-15                18       0.008249
## 6 -0.8679 -1.4646  0.8819 -1.332e-15                18       0.012013
```

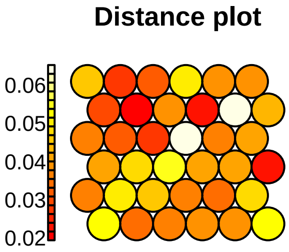
```
write.table(data.val, file="../data/ssom.data.analysis5d.txt")
```











Visualization

Use the file you wrote out above with the `superSOMtutorial.Rmd` script to look at clusters further.