Self Organizing Maps (SOM): Example using RNAseq reads

Part 3a: Visualizing superSOMs

Getting the SOMs is easy, the hard part is interpreting. You can approach these types of analyses in many ways, but below if how I approached understanding my work. I wrote a bunch of functions, that likely only work with my data, but you can go into `sSOM_functions.R` to see how these functions are made to modify for your data.

Required Libraries and source code

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.3.2
```

```
library(reshape)
library(goseq)
```

```
## Loading required package: BiasedUrn

## Loading required package: geneLenDataBase

##
```

```
library(knitr)
```

```
## Warning: package 'knitr' was built under R version 3.3.2
```

```
library(GO.db)
```

```
## Loading required package: AnnotationDbi

## Warning: package 'AnnotationDbi' was built under R version 3.3.1

## Loading required package: stats4

## Loading required package: BiocGenerics

## Loading required package: parallel

##
## Attaching package: 'BiocGenerics'

## The following objects are masked from 'package:parallel':
##
##     clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
##     clusterExport, clusterMap, parApply, parCapply, parLapply,
##     parLapplyLB, parRapply, parSapply, parSapplyLB
```

```
## The following objects are masked from 'package:stats':
##
##     IQR, mad, xtabs

## The following objects are masked from 'package:base':
##
##     anyDuplicated, append, as.data.frame, cbind, colnames,
##     do.call, duplicated, eval, evalq, Filter, Find, get, grep,
##     grepl, intersect, is.unsorted, lapply, lengths, Map, mapply,
##     match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,
##     Position, rank, rbind, Reduce, rownames, sapply, setdiff,
##     sort, table, tapply, union, unique, unsplit

## Loading required package: Biobase

## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.

## Loading required package: IRanges

## Loading required package: S4Vectors

## Warning: package 'S4Vectors' was built under R version 3.3.1

##
## Attaching package: 'S4Vectors'

## The following objects are masked from 'package:reshape':
##
##     expand, rename

## The following objects are masked from 'package:base':
##
##     colMeans, colSums, expand.grid, rowMeans, rowSums
```

```
source("sSOM_functions.R")
```

Read in data from previous analysis.

```
plot.data <- read.table("../data/ssom.data.txt",header=TRUE)
names(plot.data)
```

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

```
## [13] "Cmbr.1"            "Cother.1"          "PC1"
## [16] "PC2"               "PC3"               "PC4"
## [19] "PC5"               "PC6"               "ssom.unit.classif"
## [22] "ssom.distances"
```

```
head(plot.data)
```

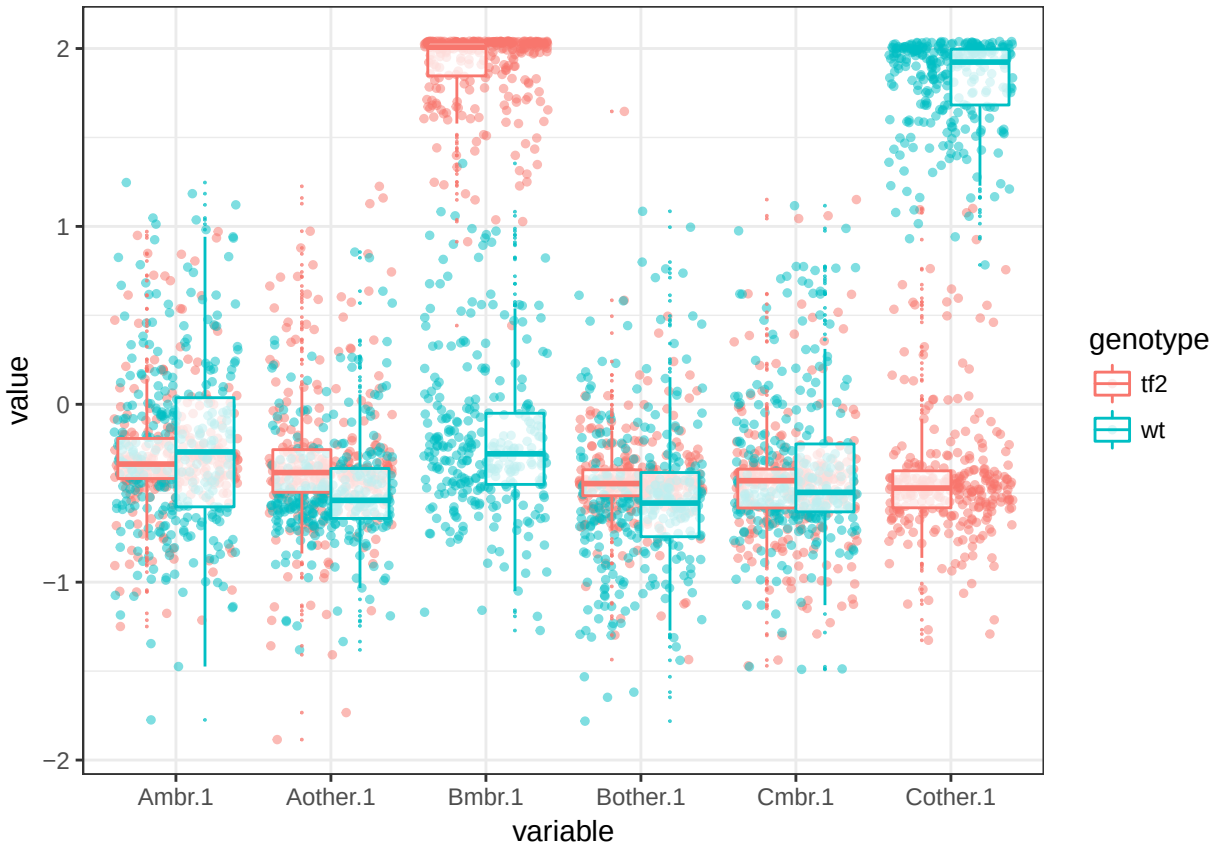
```
##   genotype               gene      Ambr     Aother       Bmbr    Bother
## 1      tf2 Solyc00g005050.2.1  9.525735  1.2969785   3.963779 11.024831
## 2      tf2 Solyc00g005070.1.1 16.174774 14.2025962 158.811326  4.479819
## 3      tf2 Solyc00g005080.1.1 11.796048  7.7875692  15.482330  8.518527
## 4      tf2 Solyc00g005840.2.1 13.584917 44.2405788   7.507857 19.327679
## 5      tf2 Solyc00g005870.1.1  6.110472  0.5291395  37.612382  1.456090
## 6      tf2 Solyc00g005880.1.1  1.839943  1.1235811  61.638800  2.035812
##        Cmbr    Cother     Ambr.1   Aother.1     Bmbr.1   Bother.1
## 1  9.457630  6.842589  0.6682306 -1.5249843 -0.8142005  1.0677854
## 2 11.542347  3.107667 -0.3039380 -0.3362605  2.0337657 -0.4956094
## 3 11.463872  7.041336  0.4553713 -0.8054291  1.6148298 -0.5755183
## 4 10.452300 29.708776 -0.5191322  1.6854463 -0.9561593 -0.1061458
## 5  2.344326  1.564099 -0.1488347 -0.5336037  2.0228620 -0.4697011
## 6  5.711226  1.787418 -0.4345752 -0.4641784  2.0365771 -0.4264810
##       Cmbr.1    Cother.1        PC1         PC2        PC3        PC4
## 1  0.6500786 -0.04690986  0.2265119 -1.62836740  0.3866619  1.5784047
## 2 -0.3798599 -0.51809787 -1.6467167 -0.09637224 -0.8791275 -1.3966283
## 3  0.3508907 -1.04014435 -2.2767329 -0.74387704 -0.1000315 -0.4976597
## 4 -0.7444119  0.64040284  1.0796206  1.71914907  0.2204601 -0.2686421
## 5 -0.4084674 -0.46225509 -1.6906182 -0.23661604 -1.0234687 -1.2019845
## 6 -0.2745967 -0.43674578 -1.5294584 -0.32073888 -0.8679183 -1.4646347
##          PC5           PC6 ssom.unit.classif ssom.distances
## 1  1.0599396  1.484923e-15               11    0.072268688
## 2  0.8451023 -1.665335e-15                1    0.001603610
## 3  0.9484308 -2.886580e-15                1    0.036801305
## 4 -1.0926003  1.110223e-16               23    0.009979580
## 5  0.8959657 -1.498801e-15                1    0.007301491
## 6  0.8818916 -1.332268e-15                1    0.010737197
```

Visualization

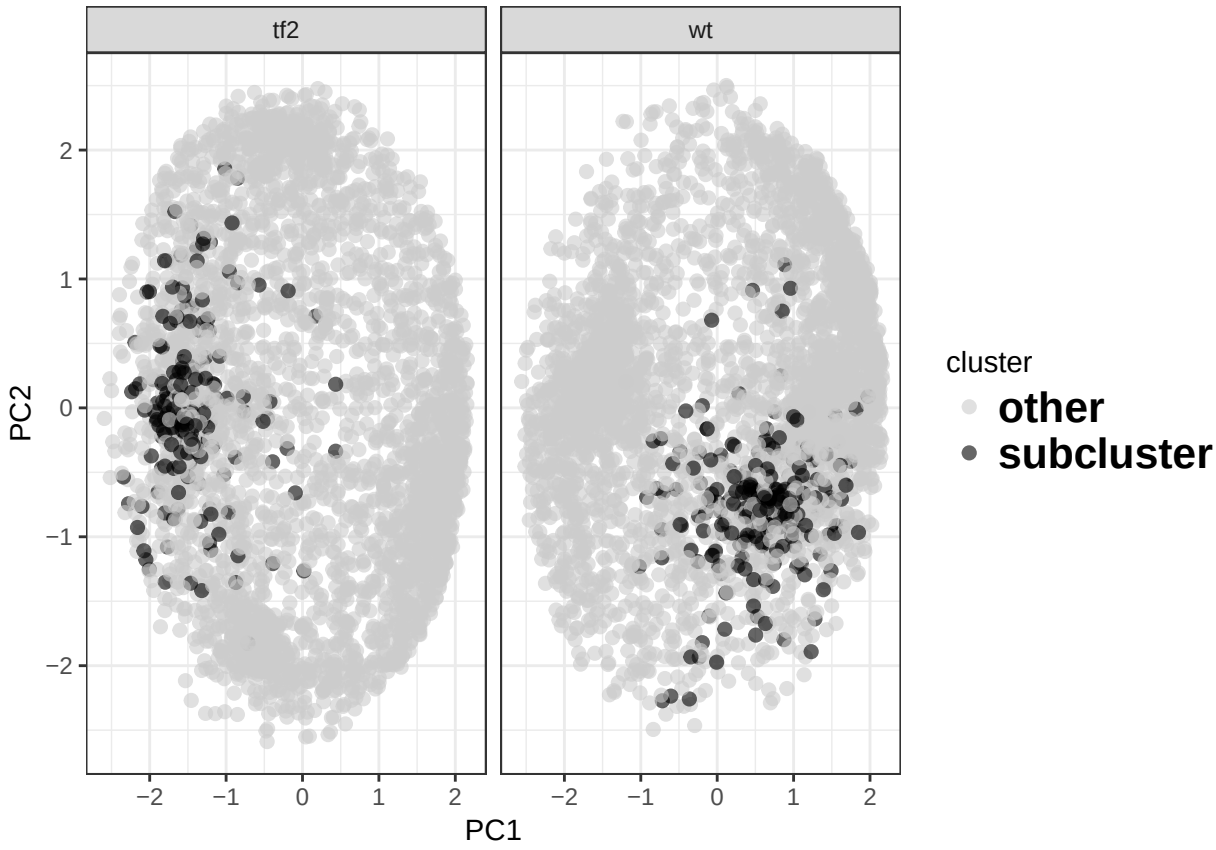
You can examine each cluster with these functions, just change the number in the function.

```
clusterVis(1)
```

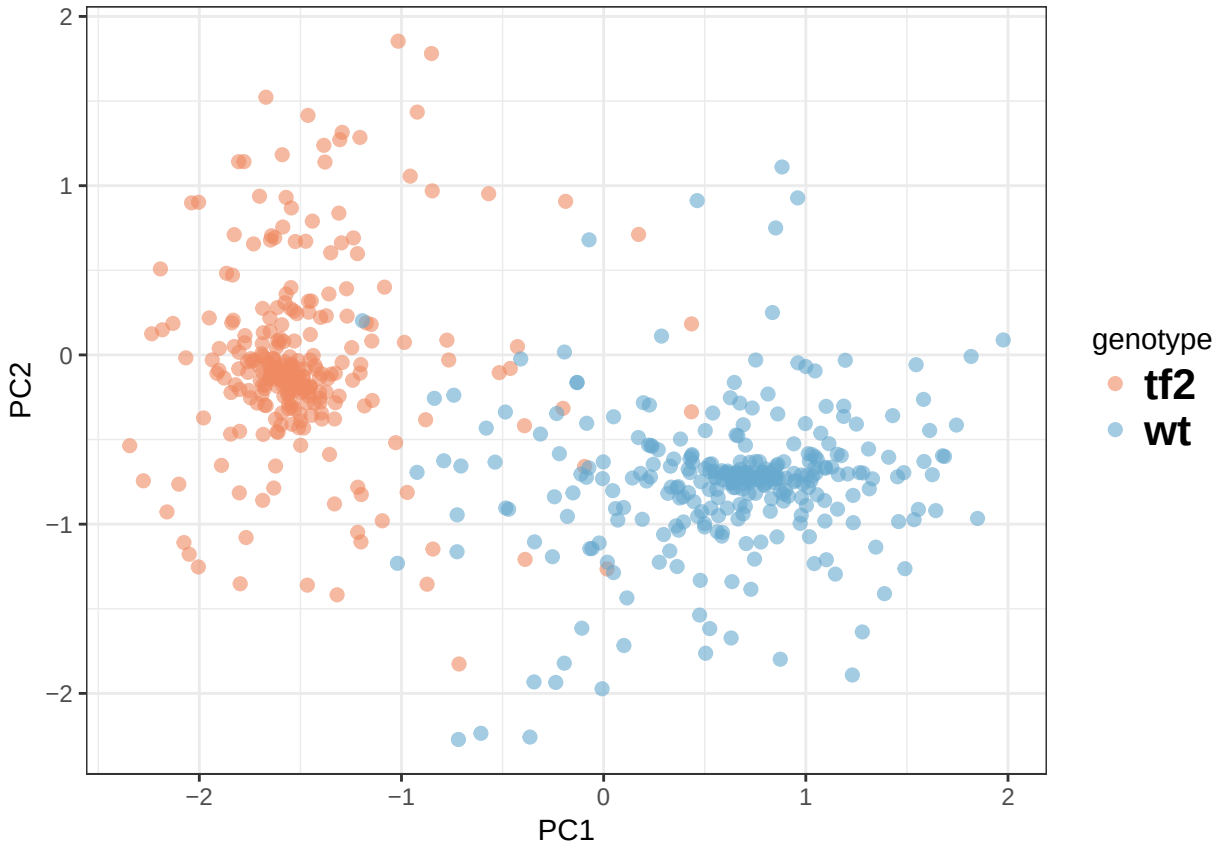
```
## Using genotype as id variables
```




```
clusterVis_PCA(1)
```

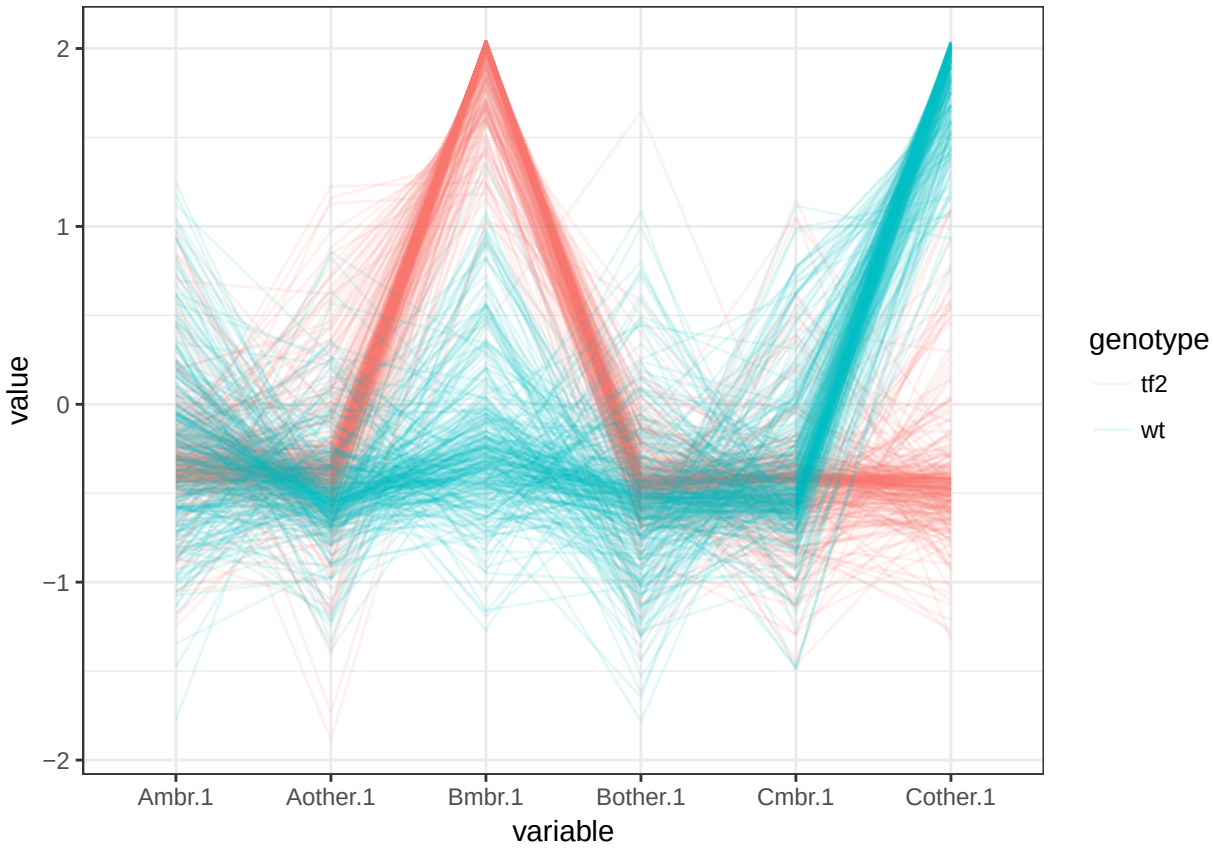



```
clusterVis_PCAsub(1)
```




```
clusterVis_line(1)
```

```
## Using genotype, gene as id variables
```




```
y <- genesInClust(1, plot.data, annotation)
```

```
## [1] 295
```

```
#kable(y, format = "latex", booktabs = TRUE)
clusterGO(1)
```

```
## Using manually entered categories.

## For 2936 genes, we could not find any categories. These genes will be excluded.

## To force their use, please run with use_genes_without_cat=TRUE (see documentation).

## This was the default behavior for version 1.15.1 and earlier.

## Calculating the p-values...

## 'select()' returned 1:1 mapping between keys and columns

##            [,1]
## GO:0015074 "DNA integration"
## GO:0003964 "RNA-directed DNA polymerase activity"
## GO:0006278 "RNA-dependent DNA replication"
## GO:0006333 "chromatin assembly or disassembly"
```

```
## GO:0000785 "chromatin"
## GO:0003682 "chromatin binding"
## GO:0008270 "zinc ion binding"
## GO:0003723 "RNA binding"
## GO:0003677 "DNA binding"
## GO:0043229 "intracellular organelle"
## GO:0003676 "nucleic acid binding"
## GO:0031969 "chloroplast membrane"
## GO:0004190 "aspartic-type endopeptidase activity"
## GO:0006310 "DNA recombination"
## GO:0006259 "DNA metabolic process"
## GO:0004518 "nuclease activity"
## GO:0006508 "proteolysis"
## GO:0016779 "nucleotidyltransferase activity"
## GO:0043170 "macromolecule metabolic process"
## GO:0032549 "ribonucleoside binding"
## GO:0003899 "DNA-directed RNA polymerase activity"
## GO:0008233 "peptidase activity"
## GO:0006915 "apoptotic process"
```