Analysis 1 - Top 25% of coefficient of variation Large SOM

Author: Ciera Martinez Date: October 23 - 31, 2014

Overview

Purpose:

I used the top 25% of genes based on co-efficient of variation, then proceeded to Self Organizing Map (SOM) clustering of gene co-expression across tissue. From discussions with Neelima, the only aspect that we are really interested in co-expression of genes through time in each tissue seperatley. We are not interested in the interaction between these tissues at this time. Ideally we are looking for genes that have co-expression patterns of up-regulation through time or down regulation through time. That was the aim of Analysis1 (for details see dclcmSOM_analysis1_102314 files).

Tissue Key:

SAM: Refers to shoot apices, likely with P0 - P4. Leaf: Likley P5

The plants were allowed to grow to 5 different ages (still need to talk with Yasu about specifics), the same tissue (SAM & Leaf), were extracted from plant of the five different ages (a1, b2, c3, d4, e5).

```
##      0%     25%     50%     75%    100%
## 0.00000 0.09877 0.25478 0.61264 3.16228
```

PCA

```
#write.csv(allGenes25, "../data/analysis4.top25.csv") #to write out data if needed.
scale_data <- as.matrix(t(scale(t(allGenes25[c(2:11)]))))  #scale data

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

summary(pca)
```

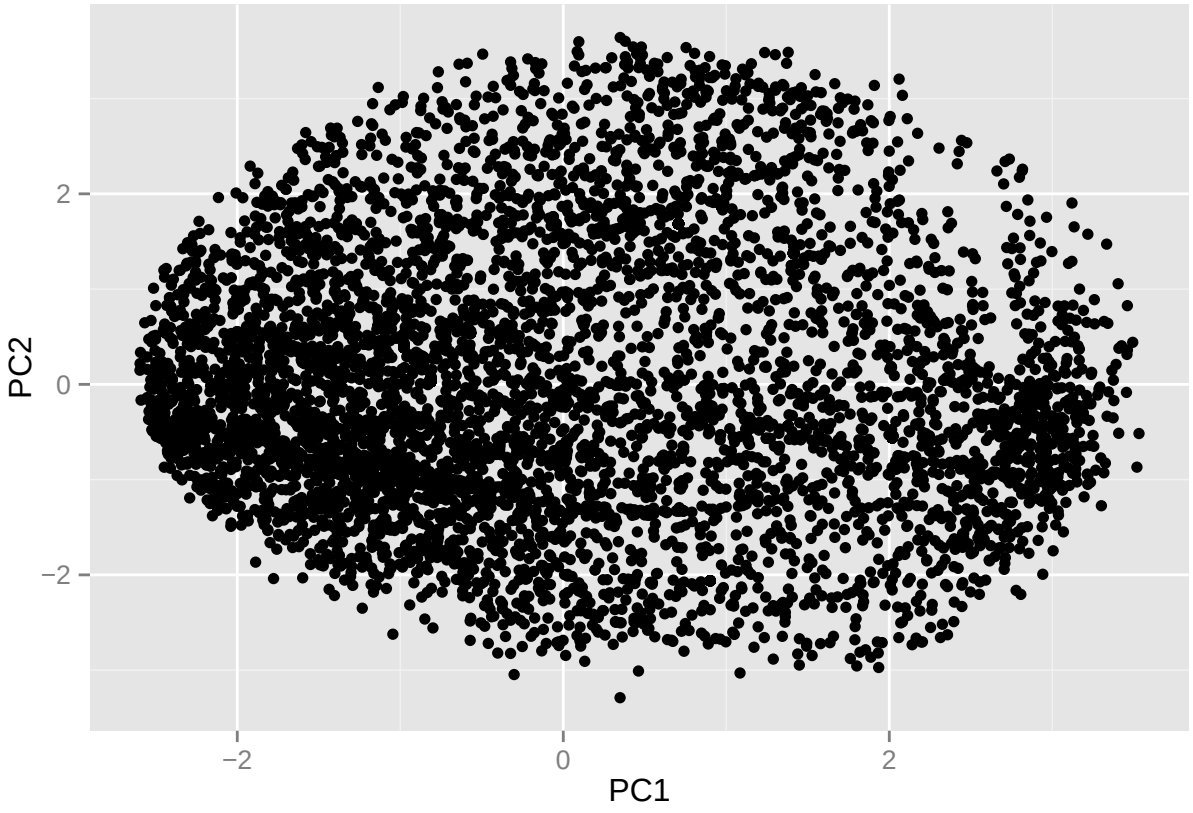
```
## Importance of components:
##                           PC1   PC2   PC3   PC4    PC5    PC6    PC7    PC8
## Standard deviation     1.453 1.337 1.089 1.008 0.9302 0.9175 0.8876 0.8735
## Proportion of Variance 0.211 0.179 0.119 0.102 0.0865 0.0842 0.0788 0.0763
## Cumulative Proportion  0.211 0.390 0.508 0.610 0.6966 0.7808 0.8596 0.9359
##                           PC9     PC10
## Standard deviation     0.8009 4.27e-15
## Proportion of Variance 0.0641 0.00e+00
## Cumulative Proportion  1.0000 1.00e+00
```

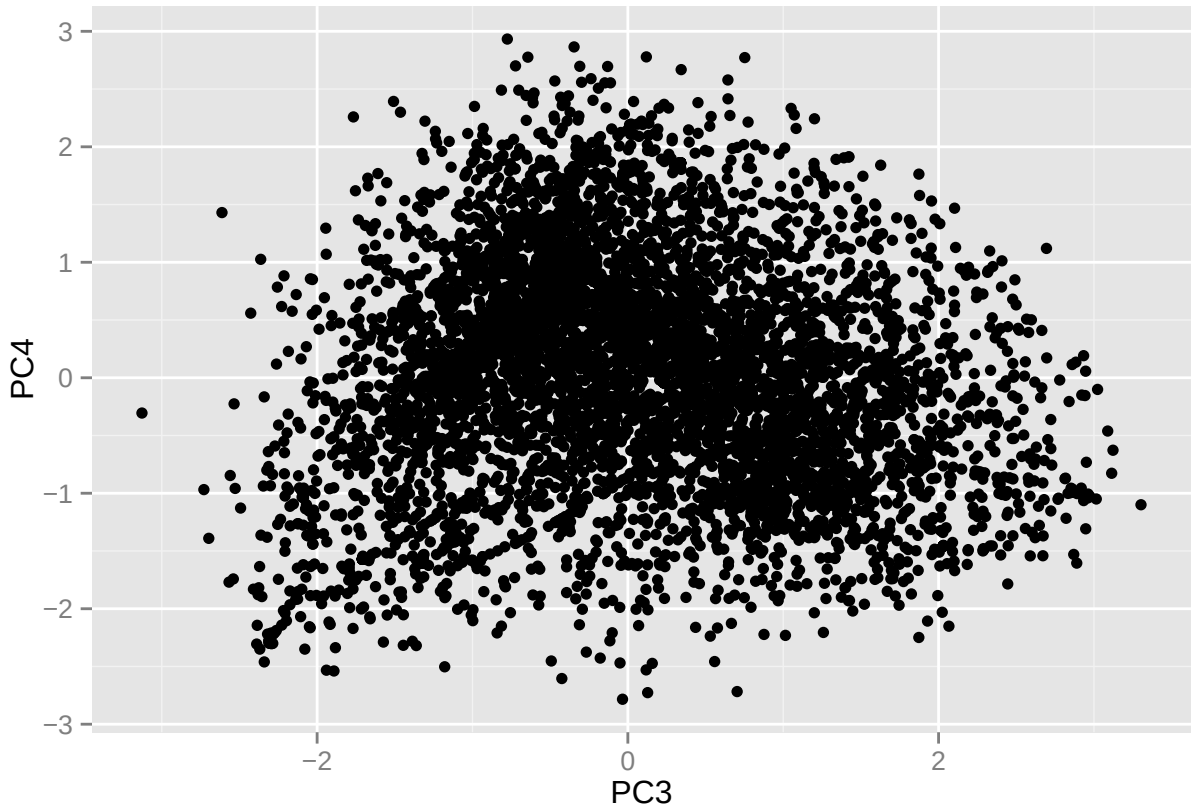
```
pca.scores <- data.frame(pca$x)

data.val.allGenes25 <- cbind(allGenes25, scale_data, pca.scores)
```

Summary of Analysis 1

Visualizing the PCA



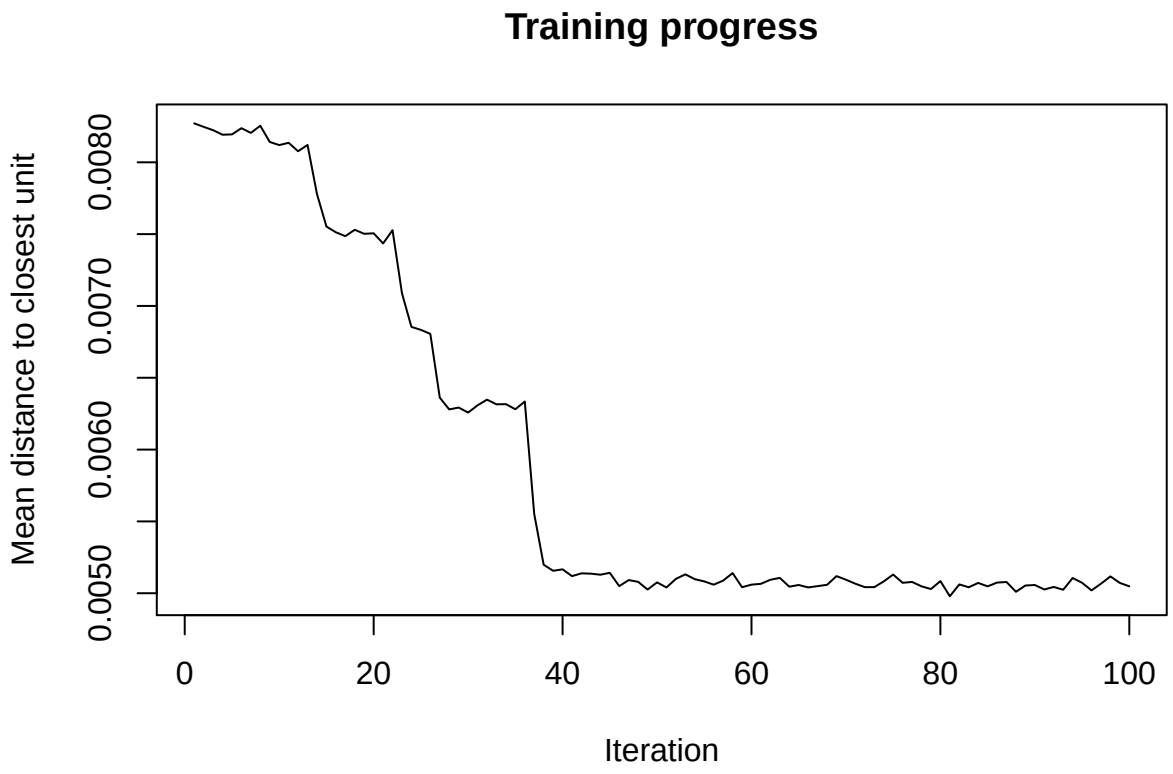


There doesn't appear to be clear clustering. There are these swooping lines. Not sure what they are. Aashish informs me that they happen often, but I don't really understand what causes them. Maybe you know Dan?

Self Organizing Map - (6,6) Large

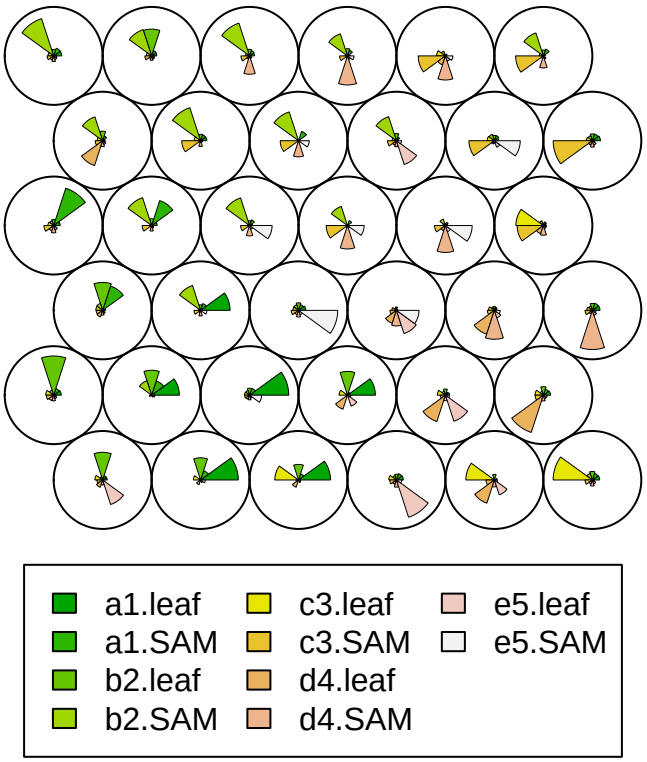
Since we are interested in particular co-expression pattern (up or down through time), I did a large SOM to explicitly find clusters with these trends.

```
## som map of size 6x6 with a hexagonal topology.
## Training data included; dimension is 6935 by 10
## Mean distance to the closest unit in the map: 1.771
```



Influence of tissue on clusters

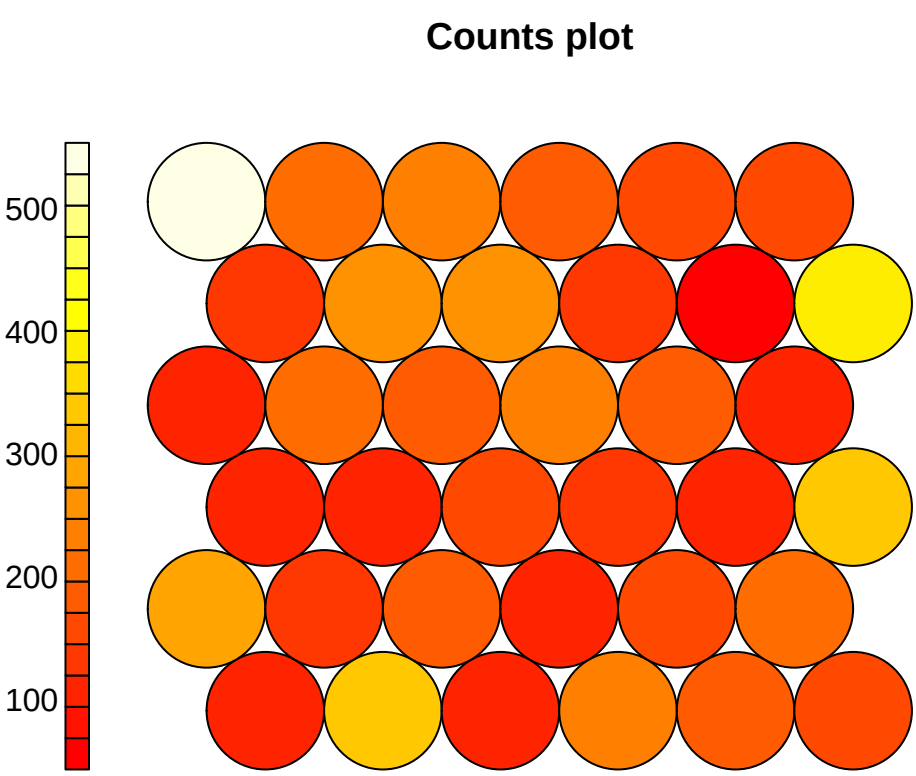
```
plot(som, type = "codes")
```



Count Plot - Large

Pretty uniform clustering, with the exception of cluster 36 which has over 500, but most clusters have around 100-200. I performed a larger SOM in analysis 2. I chose this size because it was the smallest SOM that was still able to do GO enrichment and identify gene clusters that show trends of up or down regulation through plant ages. Some clusters may not be large enough to see any significant GO enrichment in the interesting clusters.

```
plot(som, type = "counts")
```



Visualize by Cluster (Only interesting Clusters)

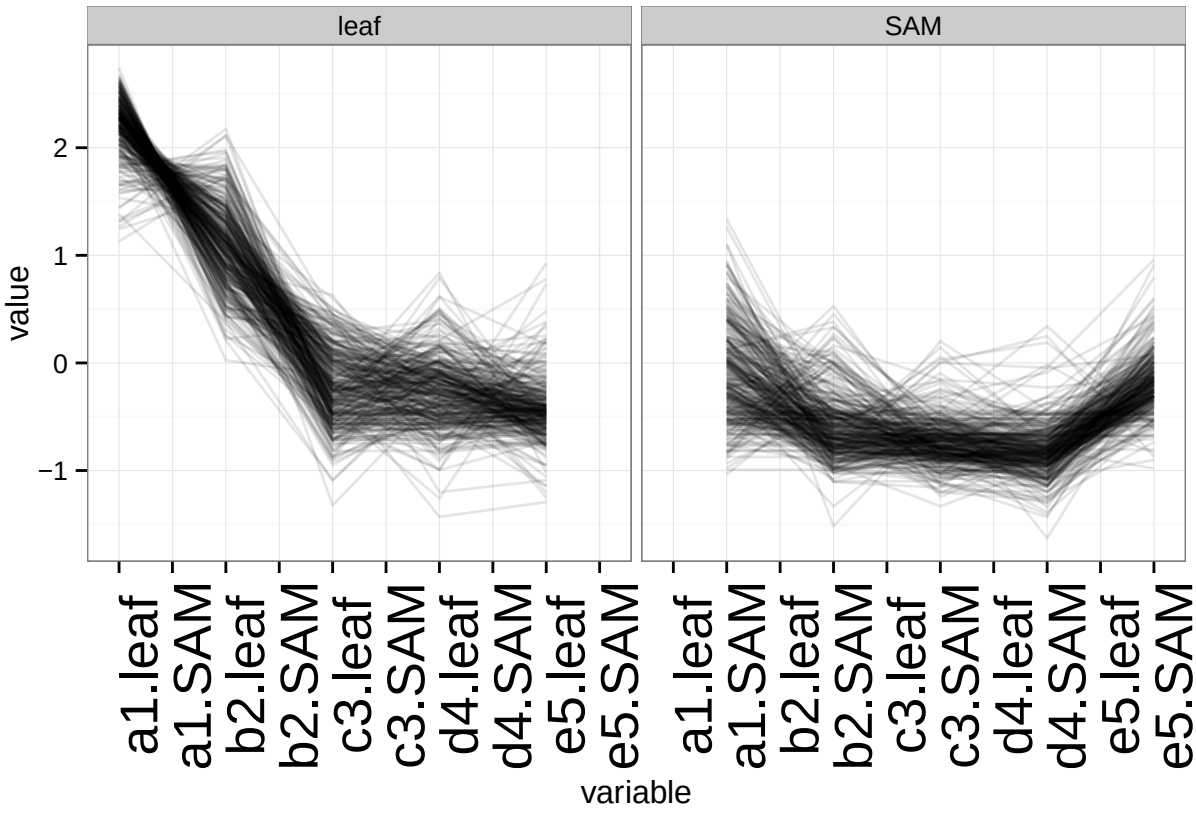
I went through each cluster and tried to identify clusters that have the co-expression patterns we are interested in. The clusters that were the most interesting are clusters 2, 11, 16. You can find .csv tables of these clusters in the clusterTables folder.

CLUSTER 2

344 genes. Trend of down regulation in Leaf tissue through plant age. GO Enrichment of photosynthetic genes.

```
clusterVis_line(2) #down through age in leaf
```

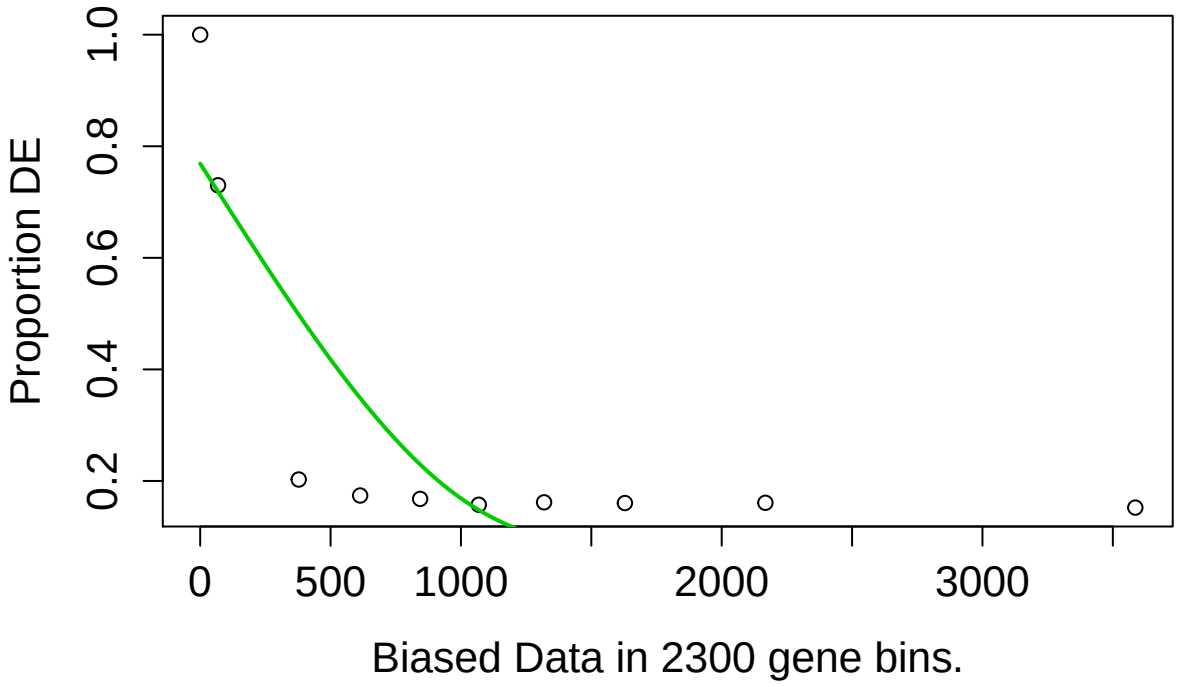
```
## Using gene as id variables
```



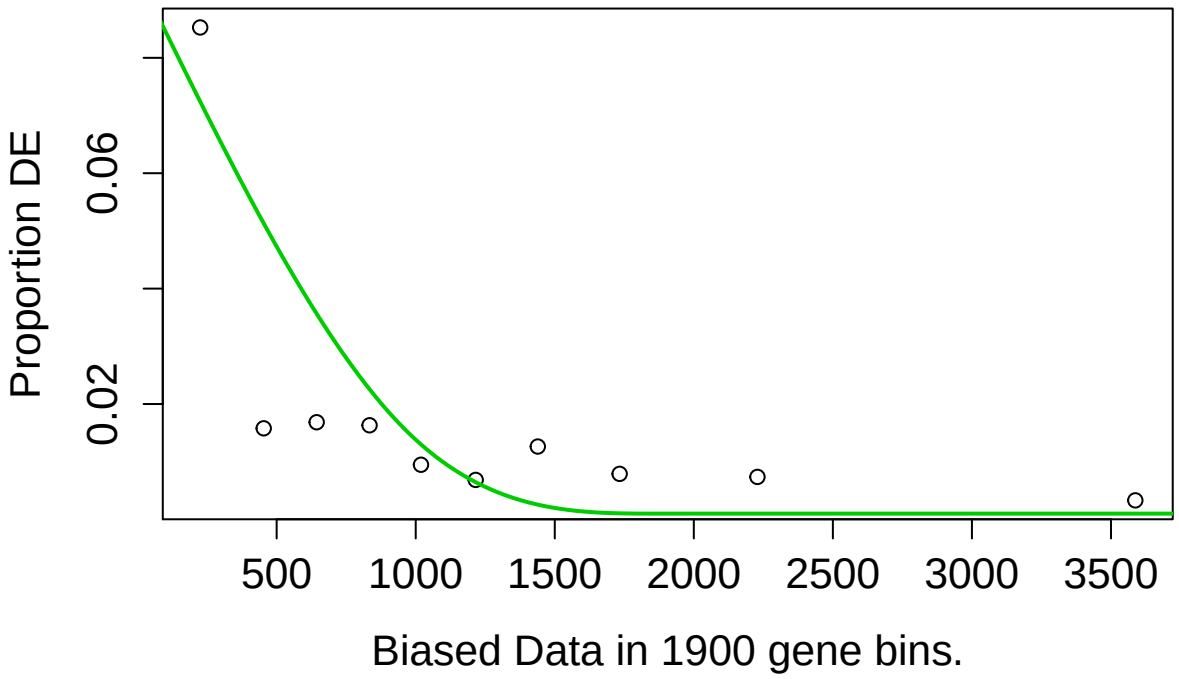

```
#What's in this cluster?
y <- genesInClust(2, plot.data, annotation)
```

```
## [1] 344
```

```
write.csv(y, "../clusterTables/analysis1.cluster2.csv")
clusterGO(2)
```




```
## Using manually entered categories.
## For 2965 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...
```




```
##            [,1]
## GO:0009570 "chloroplast stroma"
## GO:0010598 "NAD(P)H dehydrogenase complex (plastoquinone)"
```

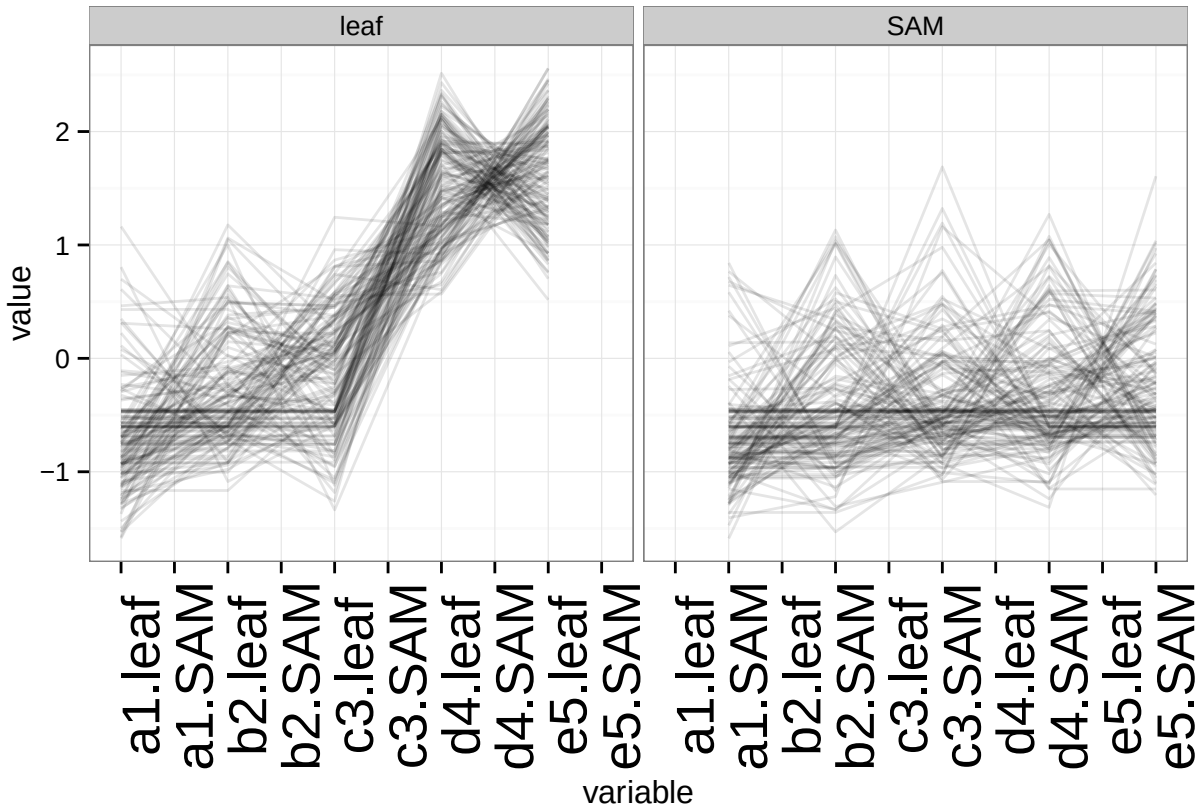
```
## GO:0048046 "apoplast"
## GO:0008152 "metabolic process"
## GO:0009535 "chloroplast thylakoid membrane"
## GO:0070009 "serine-type aminopeptidase activity"
## GO:0055114 "oxidation-reduction process"
## GO:0009941 "chloroplast envelope"
## GO:0008116 "prostaglandin-I synthase activity"
## GO:0009773 "photosynthetic electron transport in photosystem I"
## GO:0034046 "poly(G) binding"
## GO:0010319 "stromule"
## GO:0005975 "carbohydrate metabolic process"
## GO:0010258 "NADH dehydrogenase complex (plastoquinone) assembly"
## GO:0016655 "oxidoreductase activity, acting on NAD(P)H, quinone or similar compound as acceptor"
```

CLUSTER 11

147 genes. Trend of up regulation of these genes in Leaf tissue through plant age. GO enrichment in defense response and cell wall.

```
clusterVis_line(11)#up through time in leaf
```

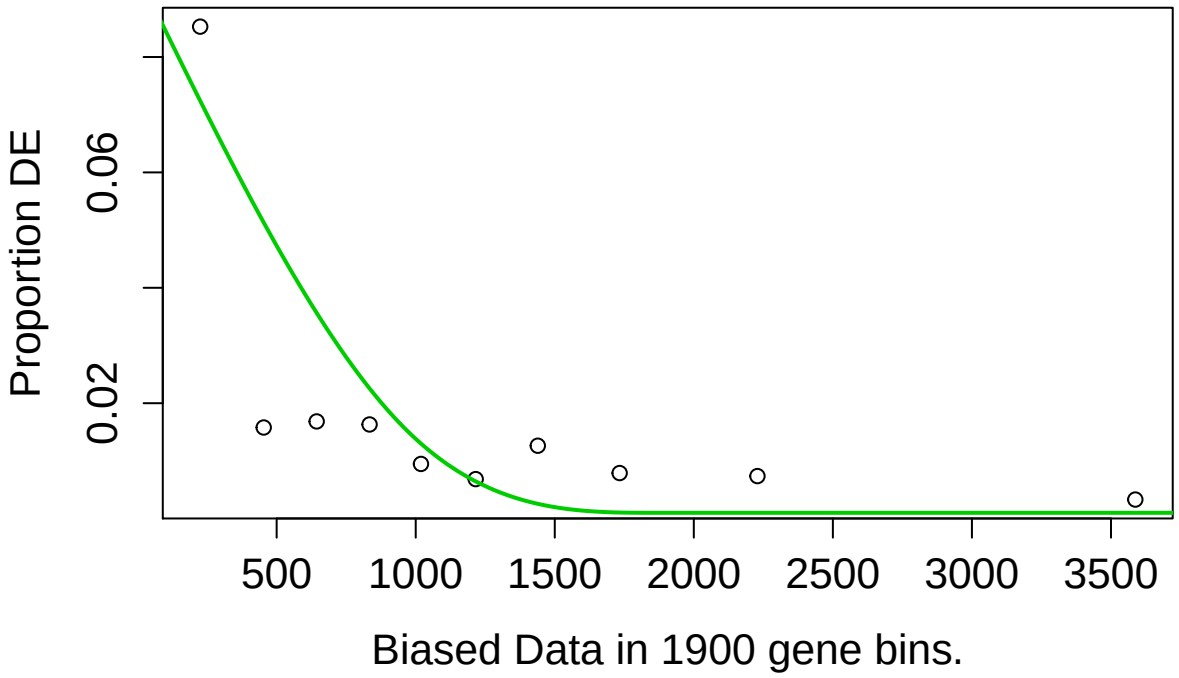
```
## Using gene as id variables
```



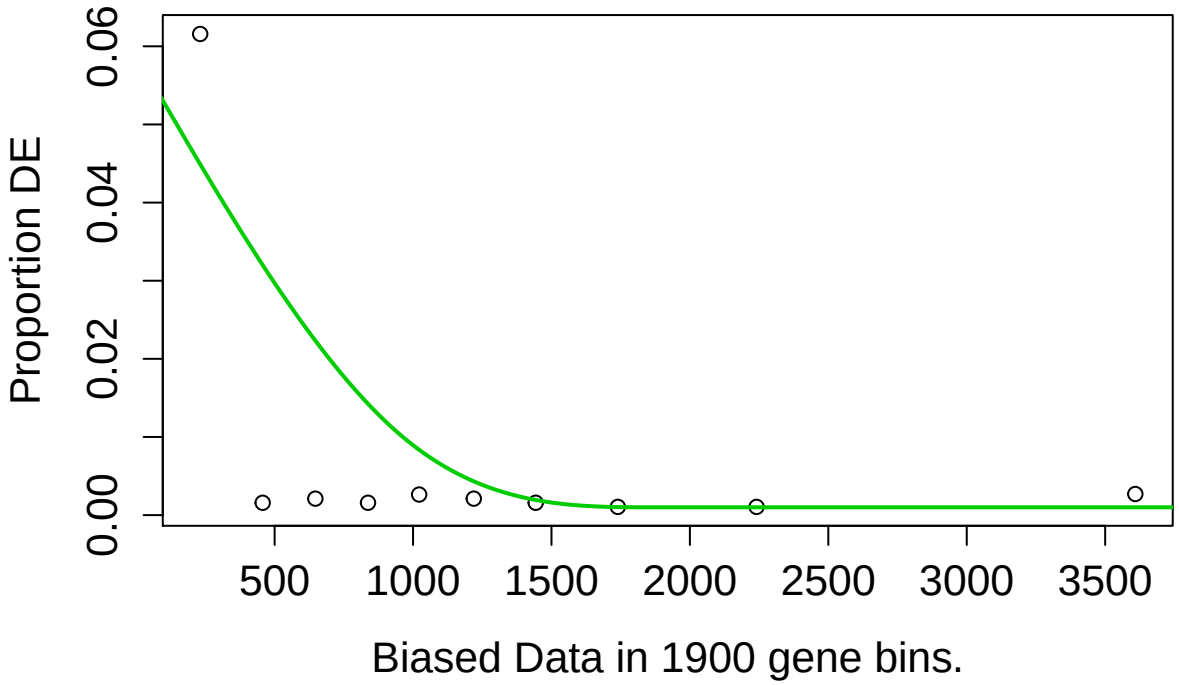

```
#what's in this cluster?
y <- genesInClust(11, plot.data, annotation)
```

```
## [1] 147
```

```
write.csv(y, "../clusterTables/analysis1.cluster11.csv")
clusterGO(11)
```




```
## Using manually entered categories.
## For 2962 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...
```



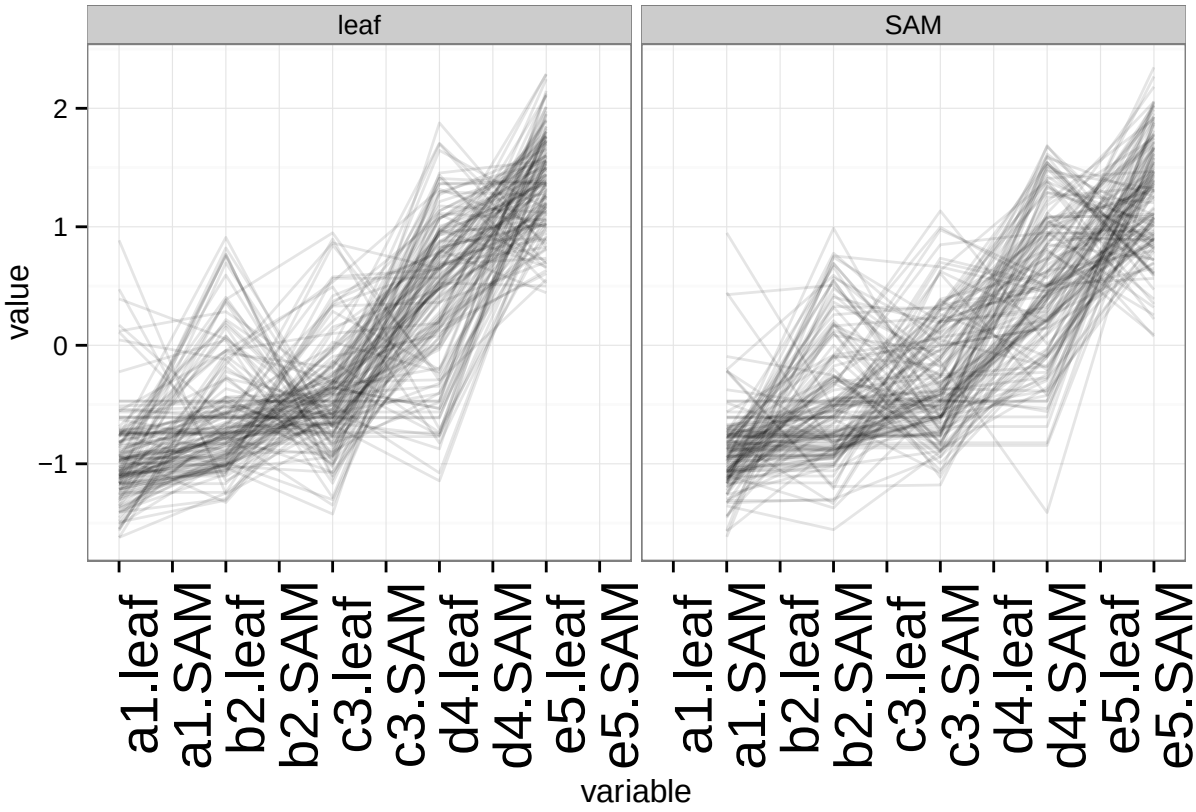
```
##            [,1]
## GO:0006952 "defense response"
## GO:0005618 "cell wall"
```

CLUSTER 16

141 Genes. Trending of up-regulated through age in both SAM and Leaf. Go enrichment doesn't mean anything to me.

```
clusterVis_line(16) #up through age in both SAM and leaves
```

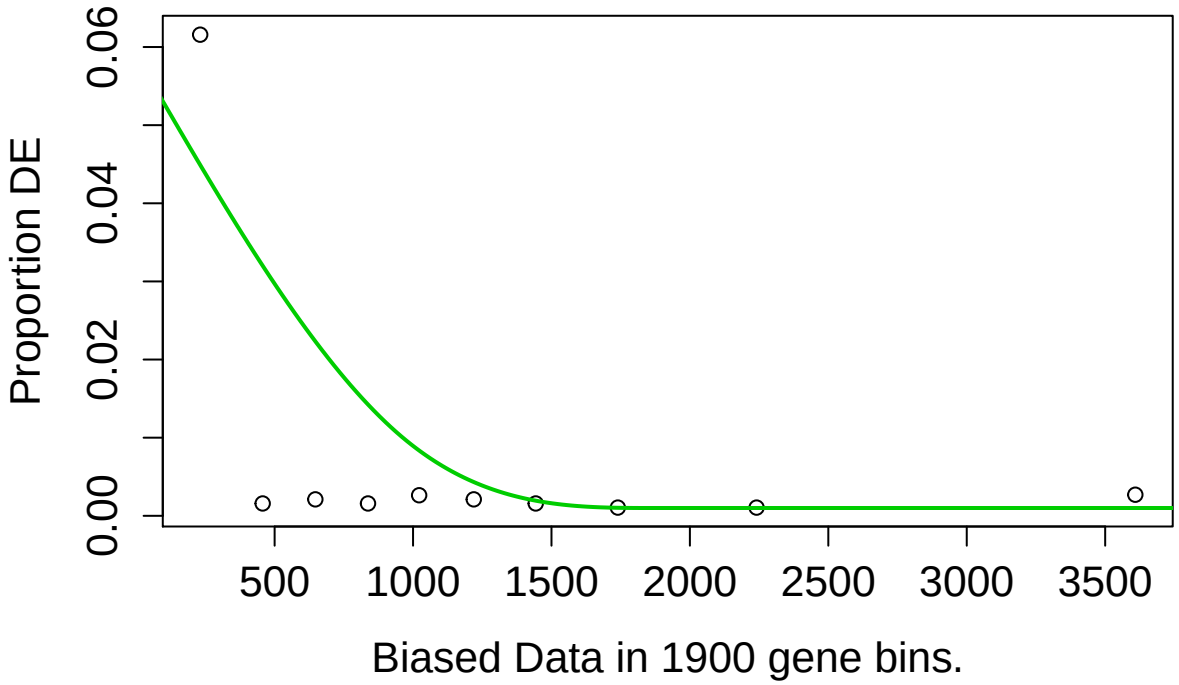
```
## Using gene as id variables
```



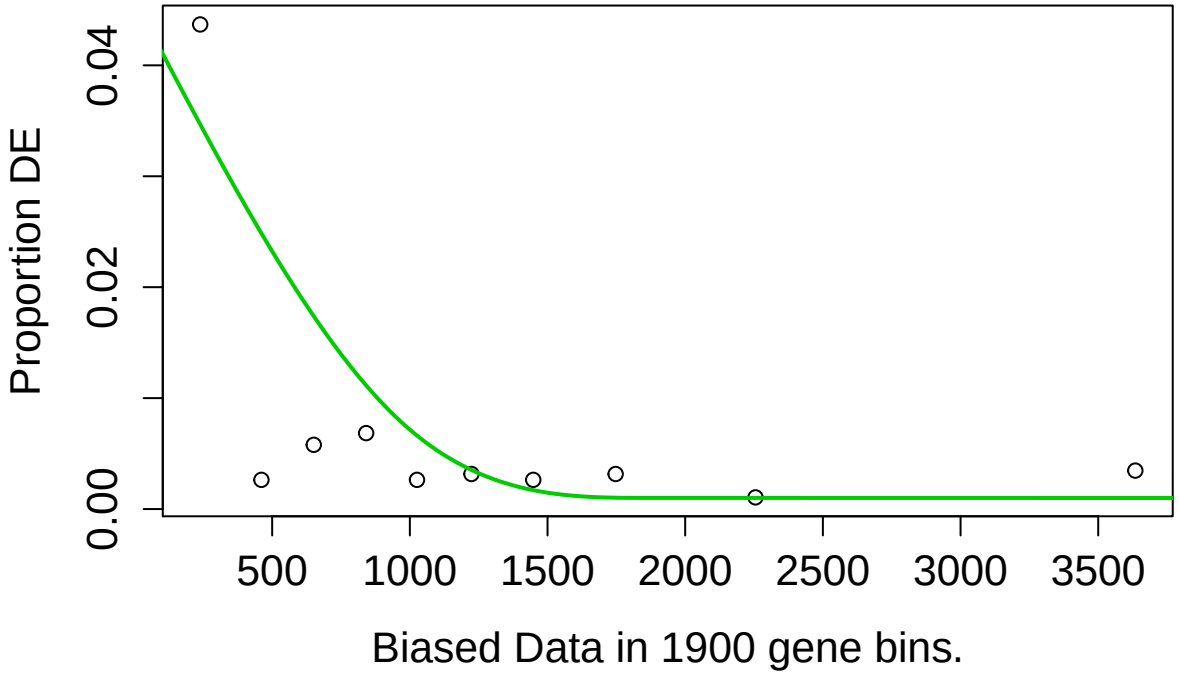

```
y <- genesInClust(16, plot.data, annotation)
```

```
## [1] 141
```

```
write.csv(y, "../clusterTables/analysis1.cluster16.csv")
clusterGO(16)
```




```
## Using manually entered categories.
## For 2953 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...
```




```
##            [,1]
## GO:0010466 "negative regulation of peptidase activity"
## GO:0004867 "serine-type endopeptidase inhibitor activity"
```

```
## GO:0005576 "extracellular region"
## GO:0009611 "response to wounding"
```

Aim 2: Specific Genes

Talking to Dan Chitwood: we need to look into specific genes. Which clusters do they fall into? From Dan via email:

*The idea behind these experiments is a bit abstract, but let me try to convey it simply. 1) KNOXs are up in the leaf primordium in foliar shade. 2) As you would expect from this, leaves are statistically more complex in shade. 3) But shade also modulates the heteroblastic series. There is lots of classical literature on this. 4) Leaf complexity in tomato increases across the heteroblasty series already.

What we didn't know is whether KNOX gene expression increases in the primordia of successive leaves across the heteroblastic series or not. If so, it suggests a mechanism by which shape, heteroblasty, and environmental response are integrated. If not, it suggests that increases in KNOX expression in shade affect leaf shape more than heteroblasty per se for shade, and that mechanisms modulating increases in leaf complexity across the series are not mediated through KNOX genes (a recent commentary Neelima and I wrote on a piece by Detlef suggests that actually TCPs/CUCs mediate heteroblasty more than KNOXs in Arabidopsis).

For starters, how do the following Knotted-like genes behave in your dataset?

Solyc04g077210.2.1 Solyc05g005090.2.1 Solyc01g100510.2.1 Solyc11g069890.1.1 Solyc02g081120.2.1

Other genes to consider are the most significant in Dataset S2, which are those differentially expressed between constant sun and 28hr shade swapped leaf primordia.**

Method:

I first extracted genes that are differentially expressed between constant sun and 28 hr shade swapped leaf primordia via Supplemntary Material v9_DatasetS2.xls. There are 645 genes in this list. When I merged with the top 25% of genes, there were 212 that overlapped.

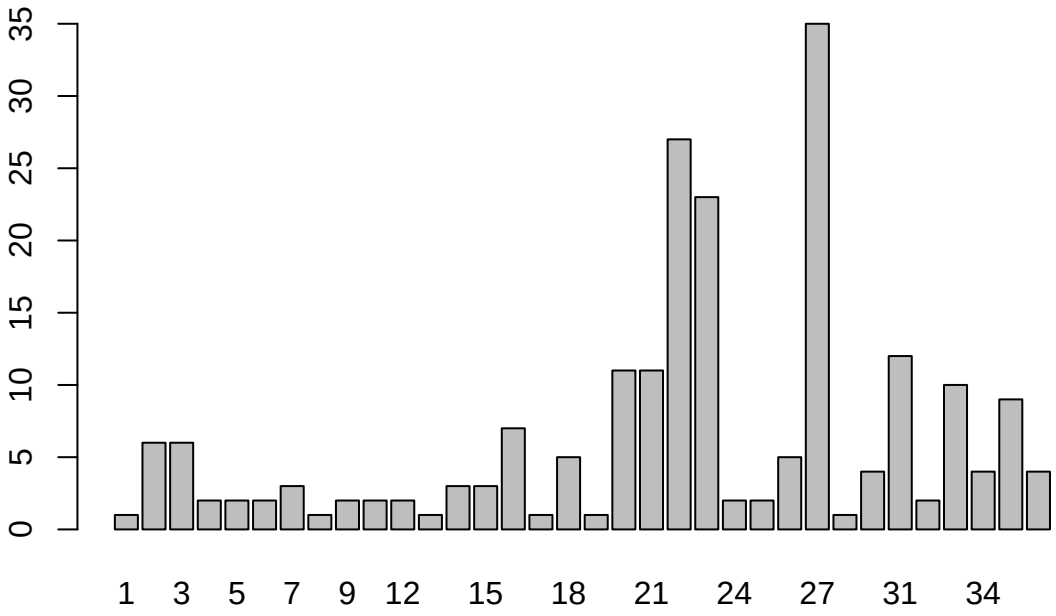
```
## 'data.frame':    212 obs. of  3 variables:
##  $ ITAG         : Factor w/ 645 levels "Solyc00g005050.2.1",..: 15 16 21 22 32 34 35 43 46 48 ...
##  $ cluster      : int  6 3 22 7 35 27 24 21 15 26 ...
##  $ som$distances: num  0.237 1.118 0.603 0.141 2.069 ...
```

I then visualized where if these 212 genes are over-represented in specific clusters.

```
summary(v9.clusterIDs$cluster)
```

```
##  1  2  3  4  5  6  7  8  9 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
##  1  6  6  2  2  2  3  1  2  2  2  1  3  3  7  1  5  1 11 11 27 23  2  2  5
## 27 28 30 31 32 33 34 35 36
## 35  1  4 12  2 10  4  9  4
```

```
plot(v9.clusterIDs$cluster)
```



Possible enriched in cluster #27, 22, and 23? Is there a way to statistically test this?

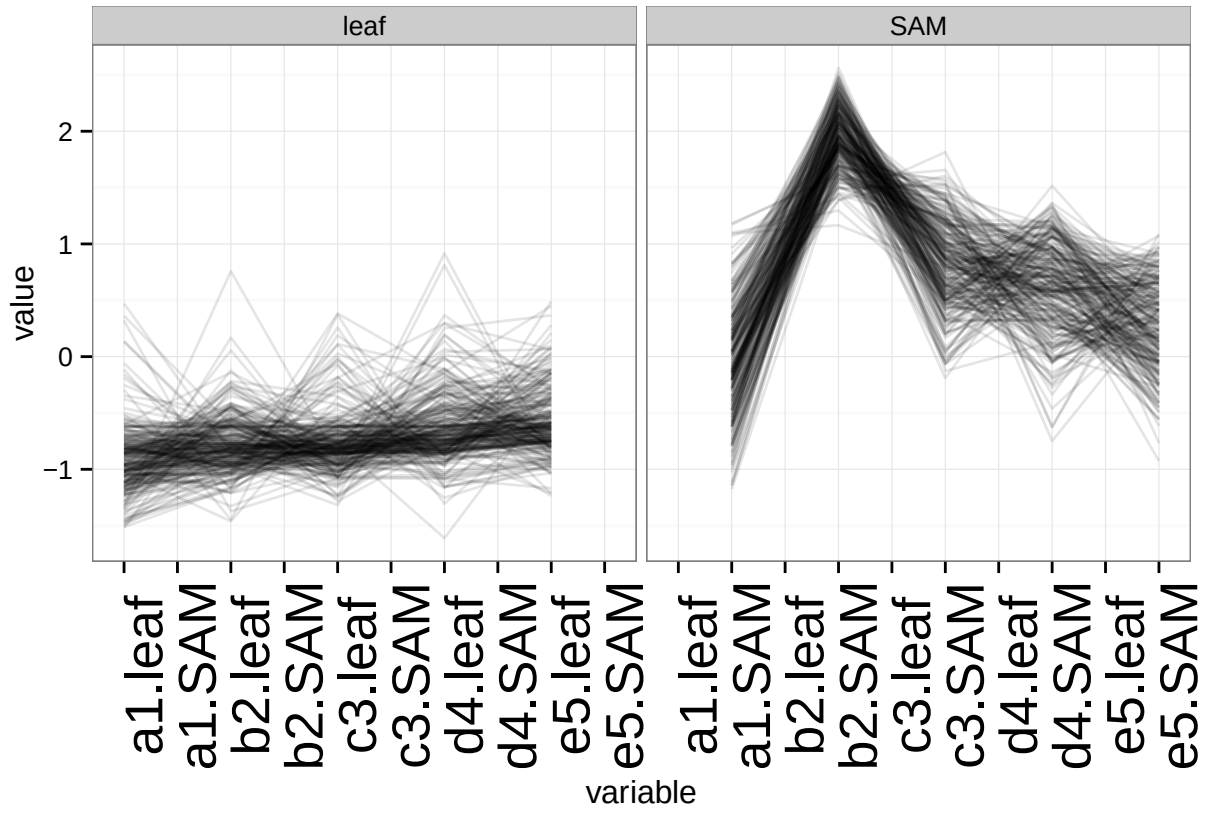
CLUSTER 27

There are 35 genes that are in cluster 27, but is this due to cluster size? What genes are in these clusters?

The gene co-expression pattern found in cluster 27 is a spike after the first plant age with down regulation in trend following plant age. I don't know what that gene pattern would mean. Cluster 27 is enriched in three GO categories:1. Regulation of transcription, DNA templates, 2. sequence-specific DNA binding transcription factor activity, and 3. maintence of inflorescence meristem identity. If we ignore the first time point, it could show that the older the plant, the less it maintains meristematic idenity in SAM and early leaf tissue. Remember "SAM" likely includes SAM and P0-P4. Cluster 27 is a large cluster at 263 genes, but not enough to explain this much overlapp. Again, is there a way to statistically test this?

```
clusterVis_line(27)
```

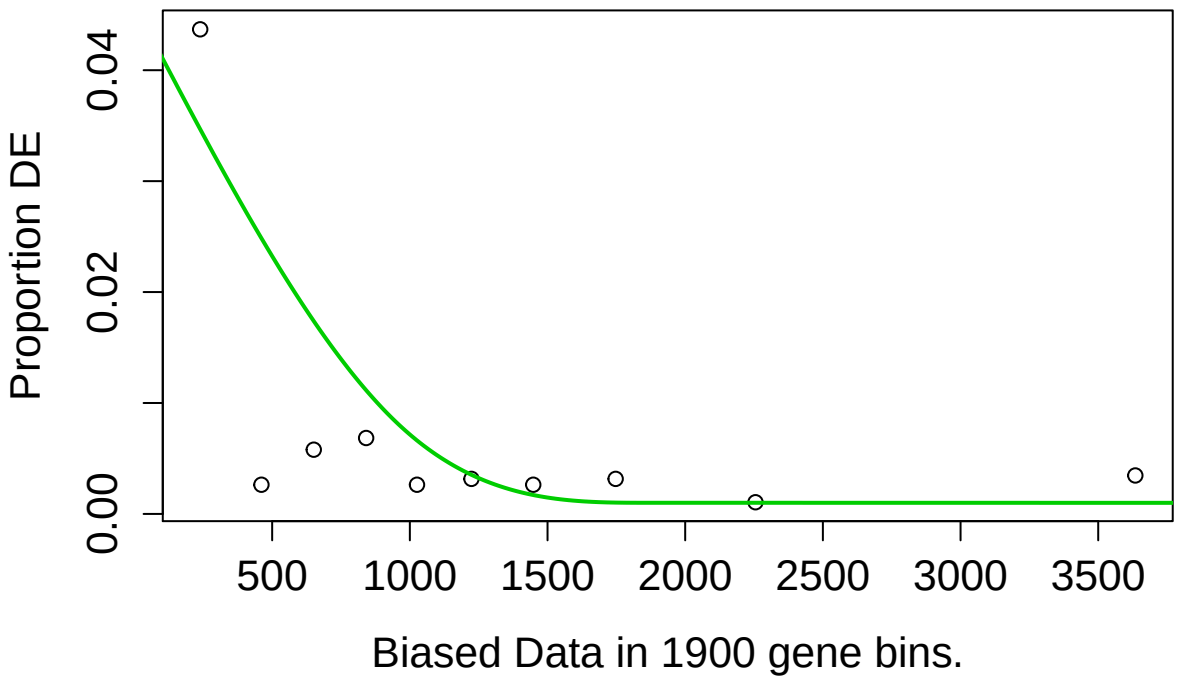
```
## Using gene as id variables
```




```
y <- genesInClust(27, plot.data, annotation)
```

```
## [1] 263
```

```
write.csv(y, "../clusterTables/analysis1.cluster27.csv")
clusterGO(27)
```



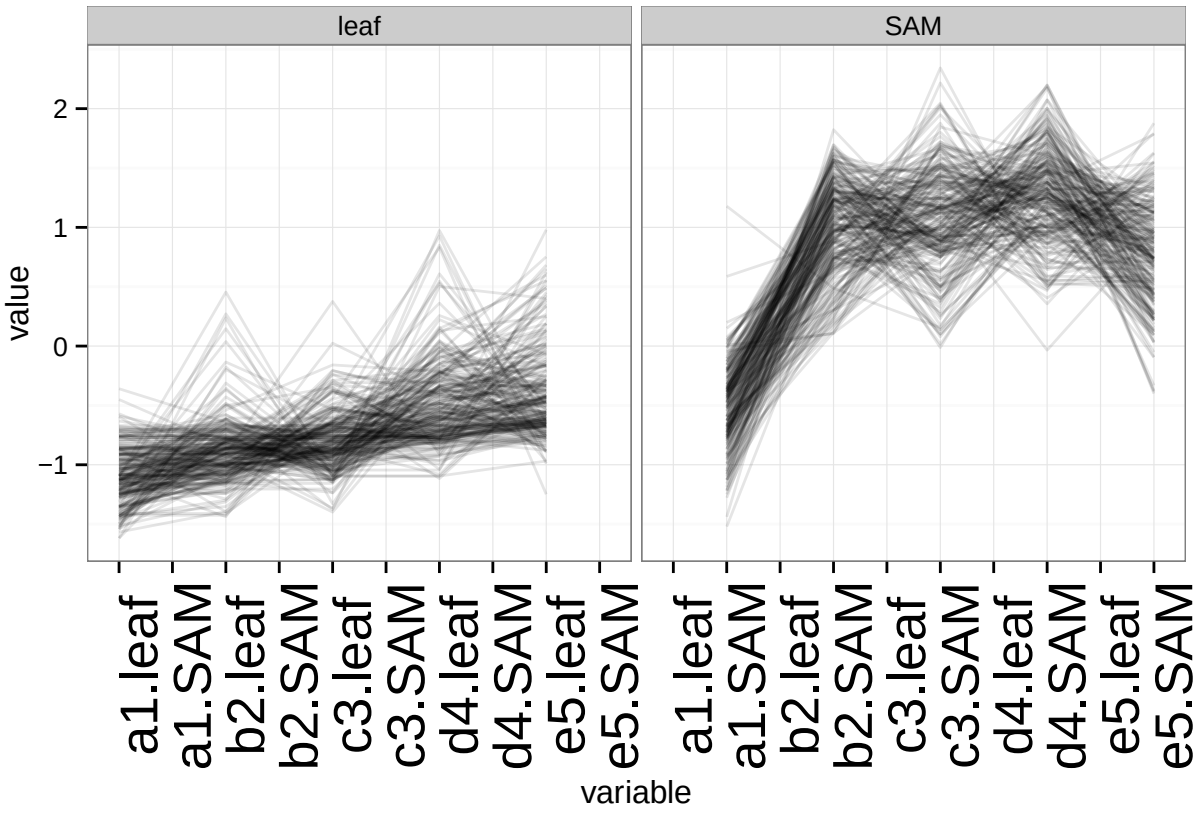
```
## Using manually entered categories.
## For 2958 genes, we could not find any categories. These genes will be excluded.
## To force their use, please run with use_genes_without_cat=TRUE (see documentation).
## This was the default behavior for version 1.15.1 and earlier.
## Calculating the p-values...
```


```
##            [,1]
## GO:0006355 "regulation of transcription, DNA-templated"
## GO:0003700 "sequence-specific DNA binding transcription factor activity"
## <NA>       NA
## GO:0010077 "maintenance of inflorescence meristem identity"
```

CLUSTER 22 The gene expression pattern is maybe up-regulation through plant age (?). GO enrichment of sequence-specific DNA binding transcription factor activity and transcription factor complex. Not very clear though.

```
clusterVis_line(22)
```

```
## Using gene as id variables
```

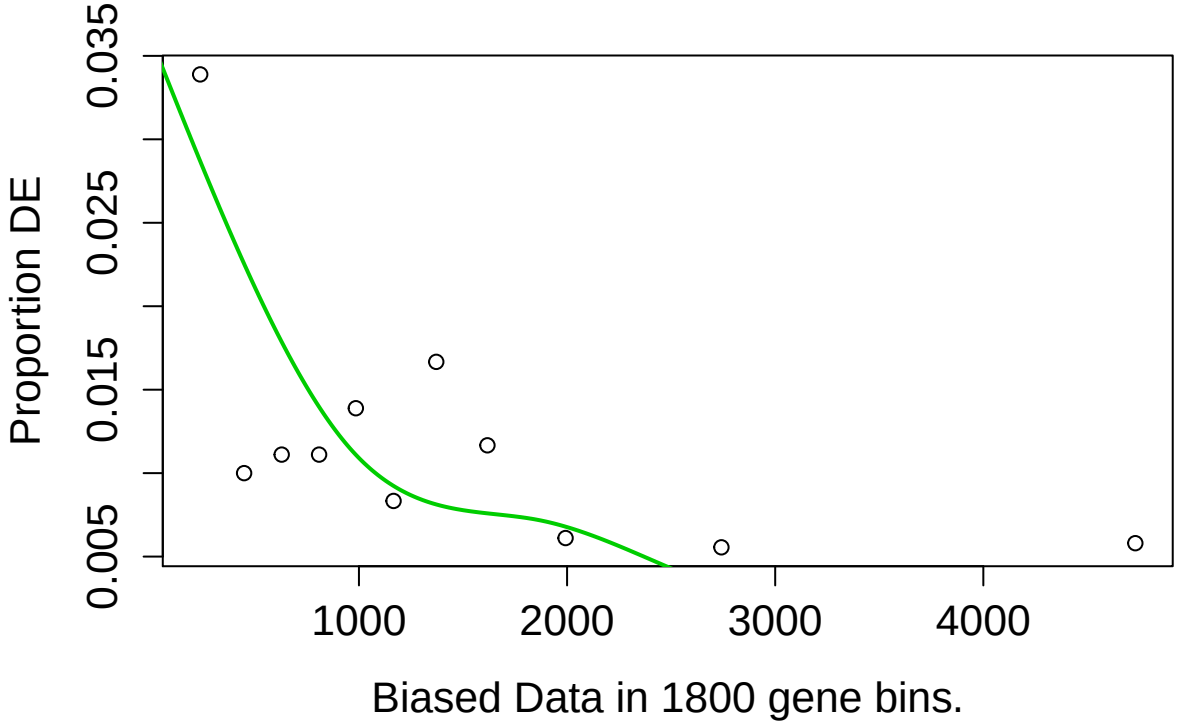



```
y <- genesInClust(22, plot.data, annotation)
```

```
## [1] 234
```

```
write.csv(y, "../clusterTables/analysis1.cluster22.csv")
clusterGO(22)
```


```
## Using manually entered categories.
## For 2954 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...
```




```
##           [,1]
```

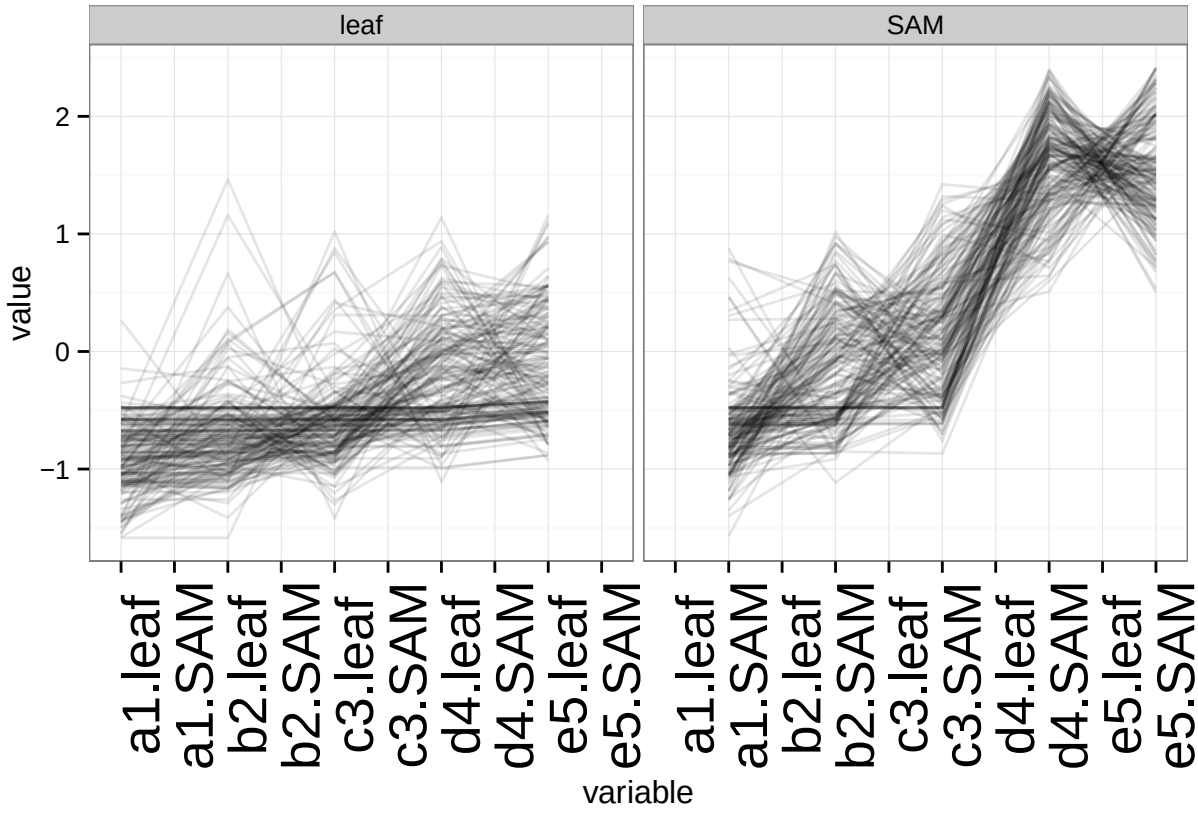
```
## <NA>       NA
## GO:0003700 "sequence-specific DNA binding transcription factor activity"
## GO:0005667 "transcription factor complex"
```

Cluster 23

This cluster auctually shows a clear trend of up-regulation through plant age in the SAM. There is no GO enrichment. Only 188 genes. Look through table output for individual genes that could be interesting?

```
clusterVis_line(23)
```

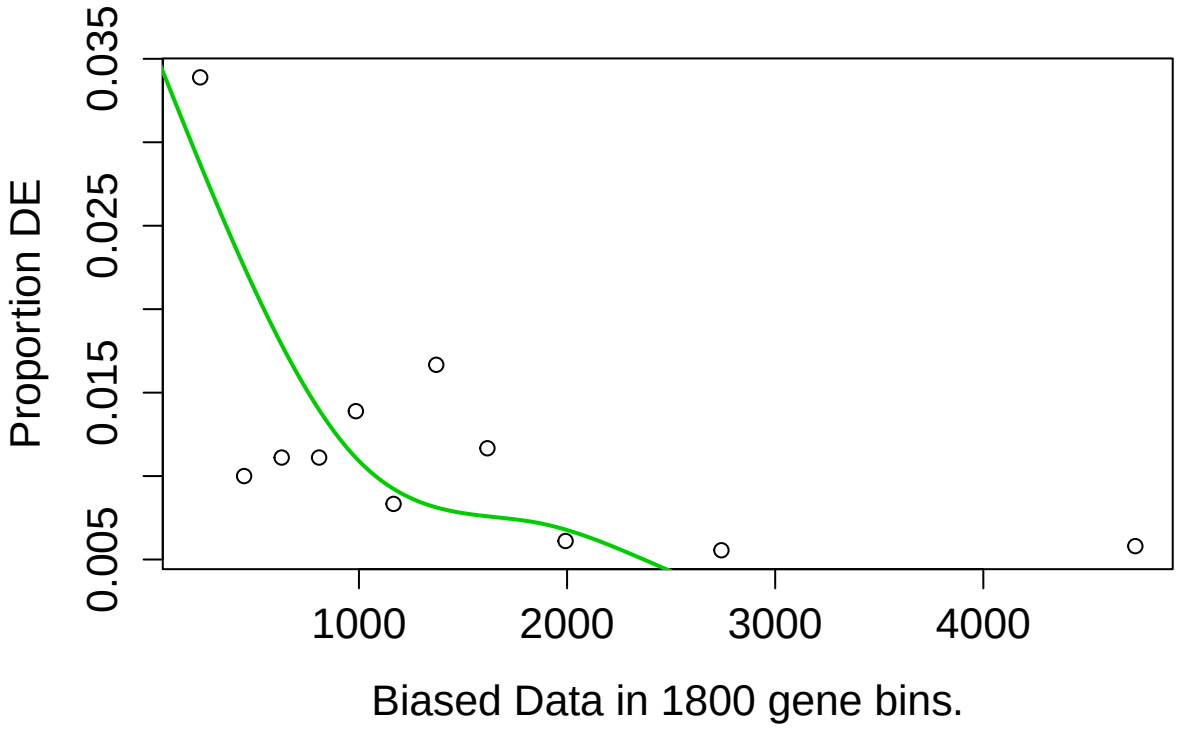
```
## Using gene as id variables
```



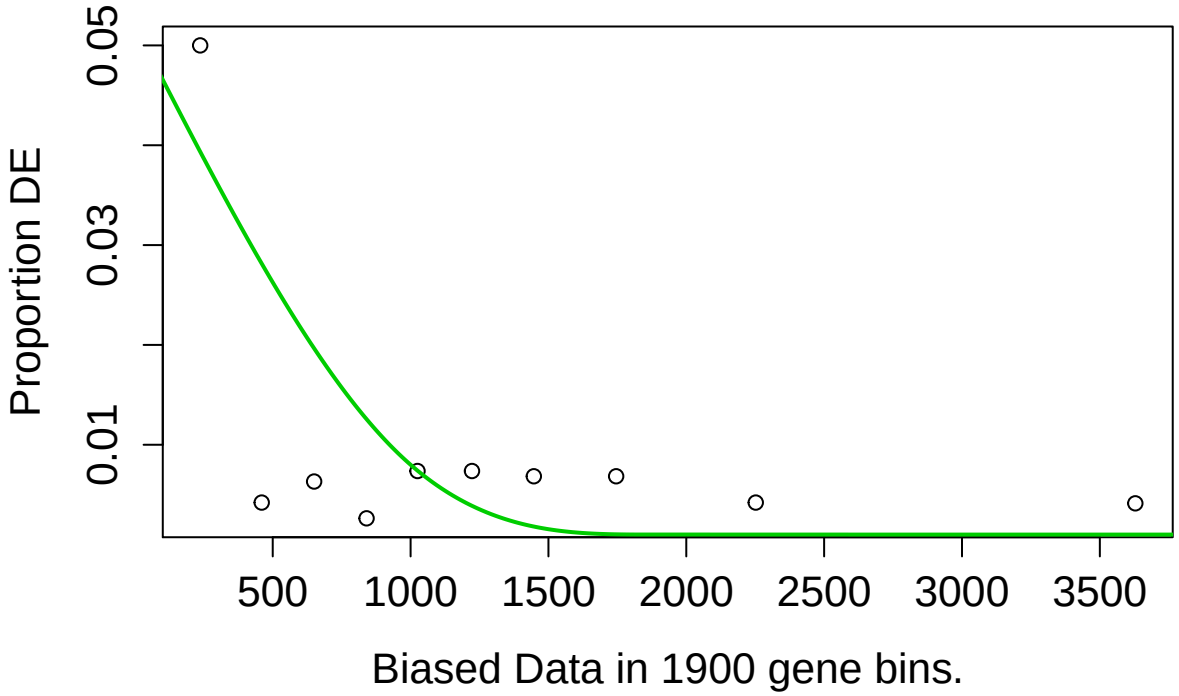

```
y <- genesInClust(23, plot.data, annotation)
```

```
## [1] 188
```

```
write.csv(y, "../clusterTables/analysis1.cluster23.csv")
clusterGO(23)
```



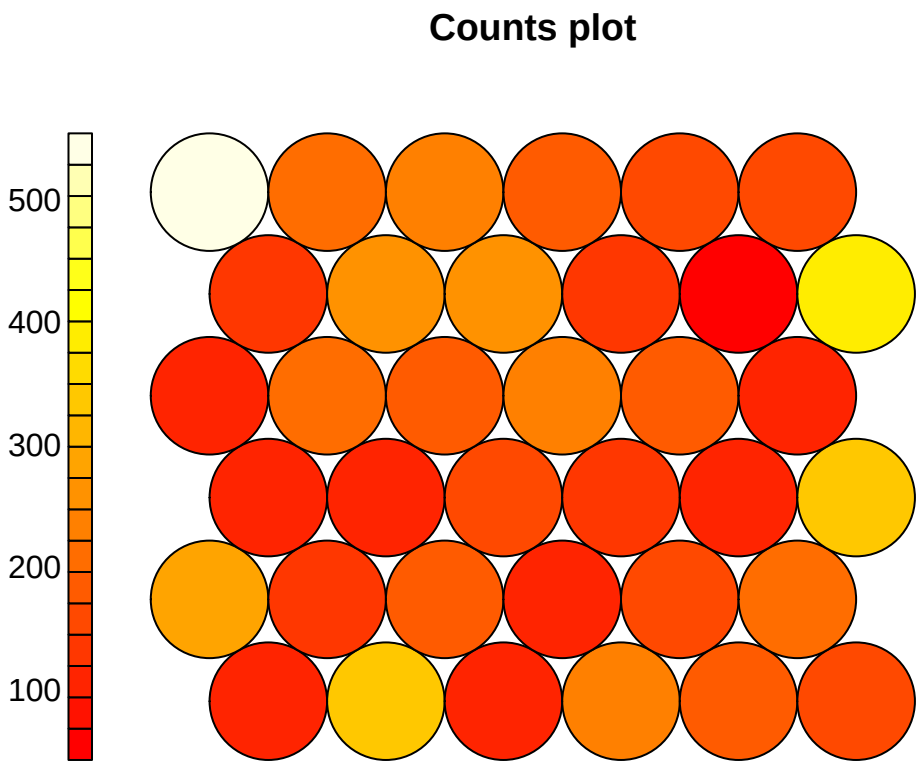

```
## Using manually entered categories.
## For 2963 genes, we could not find any categories. These genes will be excluded.
## To force their use, please run with use_genes_without_cat=TRUE (see documentation).
## This was the default behavior for version 1.15.1 and earlier.
## Calculating the p-values...
```




```
##      [,1]
```

The size of these clusters are about average, if not less.

```
plot(som, type = "counts")
```



Knotted - like genes

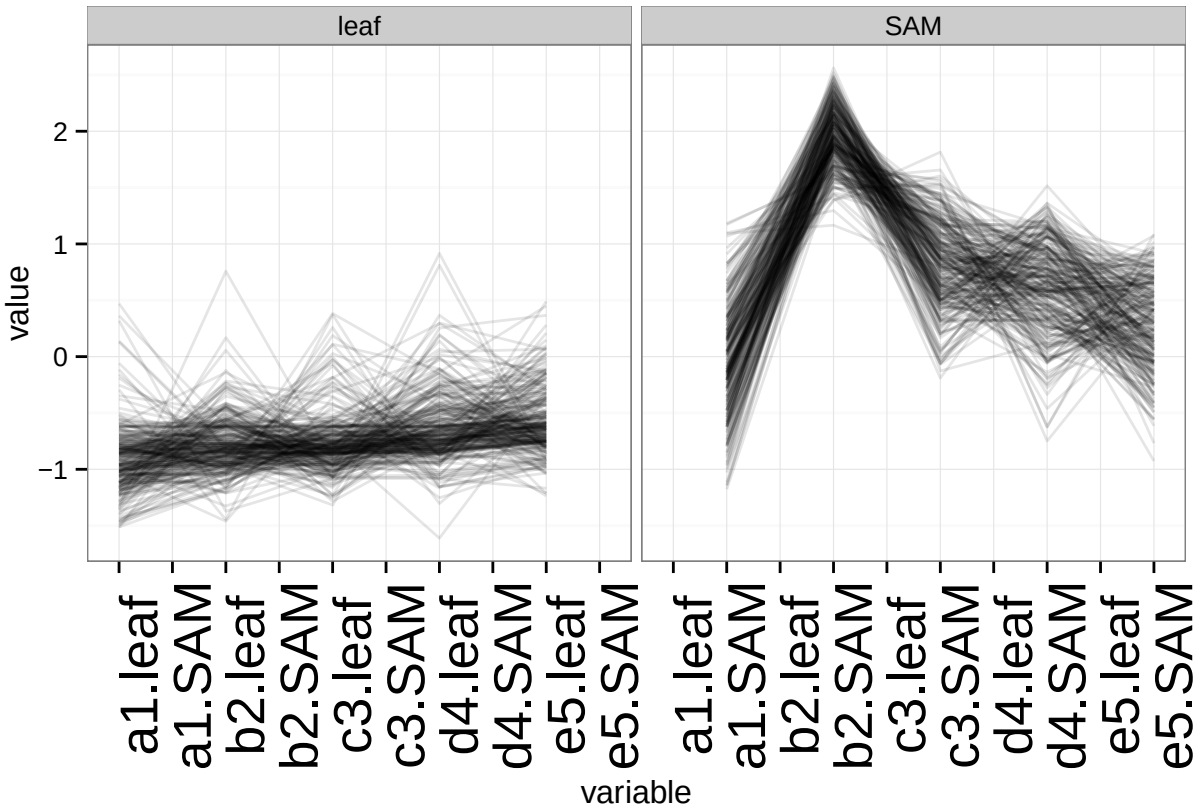
Isolated the genes and checked to see which clusters they belonged to.

```
##                  ITAG som$unit.classif som$distances
## 1 Solyc01g100510.2.1                22        0.5608
## 2 Solyc02g081120.2.1                21        1.6558
## 3 Solyc04g077210.2.1                27        0.3147
## 4 Solyc05g005090.2.1                27        0.4463
## 5 Solyc11g069890.1.1                27        0.5256
```

Three out of five of them are cluster 27.

```
clusterVis_line(27)
```

```
## Using gene as id variables
```



Overall Results and Future Analysis:

There are several clusters that could be looked at more closely. These are in the `clusterTables` directory. The clusters that were picked out for up or down regulation trends per tissue are clusters 2, 11 and 16. The clusters that were identified for "enrichment" of v9.supplementary genes are 27, 22, and 23. Cluster 27 not only had the most gene overlapp of the v9.supplementary genes, but also contained 3 out of 5 of the knotted-like genes. The expression pattern in this cluster is somewhat confusing though.

The clustering may be confounded between SAM and leaf tissue being forced into same cluster. I think looking at each of these tissues seperatley could be useful. I performed this analysis but did not yield much though, for some reason the clustering was weird, even on a small SOM map, the genes all clustered together. See `dclcmSOM_analysis2_102814.Rmd` for more details.

Also, varying SOM sizes could yield more explicit gene expression patterns if larger SOM map or allow the ability to do GO-enrichment/promoter enrichment if smaller SOM.

For Larger SOM map, see Analysis4. Basically it just makes smaller SOMs with more specific co-expression patterns. I did the same analysis as I did here but with a larger SOM map. I wasn't able to to GO enrichment on these clusters, but gives smaller gene lists, the interesting ones I printed out to the clusterTable folder.