# Analysis 7 - Getting a closer look at the clusters from the orignal clusters.

## Purpose

To get start to understand the differences in GO categories between the clusters. This is the data from analysis1D.

### Part 1

This will look into the number of genes that are the same between the clusters and the genotypes. With some basic visualization.

Required Libraries

```
library(VennDiagram)
library(ggplot2)
library(reshape)
library(kohonen)
library(goseq)
library(GO.db)
```

### Visualize by Cluster

Read in data used for GO enrichment analysis

```
geneLength <- read.csv("../../../07GO_enrichment/requisiteData/normalized_genes_length.csv")
cate <- read.table("../../../07GO_enrichment/requisiteData/melted.GOTable.txt",header=TRUE)
```

Read in data produced from analysis1D.

```
plot.data <- read.table("../data/analysis1.som.data.small.ALLD.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"              "som.unit.classif"
## [22] "som.distances"
```

## Cluster Specific analysis

Now I want to take a look at what are is going on exactly in these clusters. The clusters start with the bottom left, which is cluster number 1.

This is a function that makes a boxplot showing the transformed values of expression in the clusters.

```
#clusterVis Function
#displays transformed data in a box plot and
clusterVis <- function(clustNum){

  sub_cluster <- subset(plot.data, som.unit.classif==clustNum)
  sub_data <- sub_cluster[,9:14] # just the sample types
  m.data <- melt(sub_data)
  p <- ggplot(m.data, aes(x=variable, y=value))
  p + geom_point(alpha=0.5, position="jitter", size=1) + geom_boxplot(alpha=0.75, outlier.size=0)
}
```

Number of genes function, which gives you some basics about the clusters between

```
clusterNum <- function(clustNum){

  sub_cluster <- subset(plot.data, som.unit.classif==clustNum)
  print(paste("total number of genes in sub cluster is ",
              nrow(sub_cluster)
              )
        )

  scwt <- subset(sub_cluster, genotype == "wt")
  print(paste("total number of genes in wt cluster is ",
              nrow(scwt)
              )
        )

  sctf2 <- subset(sub_cluster, genotype == "tf2")
  print(paste("total number of genes in tf2 cluster is ",
              nrow(sctf2)
              )
        )

  scIntersect <- as.data.frame(intersect(scwt$gene, sctf2$gene))
   print(paste("There are",
               length(intersect(scwt$gene, sctf2$gene)),
               " that are the same between wt and tf2"
               )
         )

  ##Venn Diagram part
  grid.newpage()
  venn.plot <- draw.pairwise.venn(area1 = nrow(scwt),
                              area2    = nrow(sctf2),
                              cross.area = length(intersect(scwt$gene, sctf2$gene)),
                              scaled       = F,
                              category     = c("Wildtype", "tf2"),
                              fill         = c("blue", "red"),
                                alpha        = 0.3,
                                lty          = "blank",
                                cex          = 2,
                                cat.cex      = 2,
                                cat.pos      = c(315, 25),
                                cat.dist     = 0.09,
                                cat.just     = list(c(-1, -1), c(1, 1)),
                                ext.pos      = 30,
                                ext.dist     = -0.05,
                                ext.length   = 0.85)
  grid.draw(venn.plot)

}
```

```
clusterGO <- function(clustNum){
##GO Enrichment on the catergories

#we need to first get the data in the right format.
#First get the list of ITAG,

#sub_cluster
sub_cluster <- subset(plot.data, som.unit.classif==clustNum)
scwt <- subset(sub_cluster, genotype == "wt")
sctf2 <- subset(sub_cluster, genotype == "tf2")
scIntersect <- as.data.frame(intersect(scwt$gene, sctf2$gene))

itag.sc <- as.data.frame(sub_cluster$gene)
colnames(itag.sc)[1] <- "itag"
itag.sc$sc <- 1

#scwt
itag.scwt <- as.data.frame(scwt$gene)
colnames(itag.scwt)[1] <- "itag"
itag.scwt$wt <- 1

#sctf2
itag.sctf2 <- as.data.frame(sctf2$gene)
colnames(itag.sctf2)[1] <- "itag"
itag.sctf2$tf2 <- 1

#Intersect
itag.scIntersect <- as.data.frame(scIntersect[1])
colnames(itag.scIntersect)[1] <- "itag"
itag.scIntersect$intersect <- 1

#Merge all by itag
ITAGmerge <- merge(itag.scIntersect, itag.scwt, by = "itag", all= TRUE)
ITAGmerge <- merge(ITAGmerge, itag.sctf2, by = "itag", all = TRUE)
matrixGO <- merge(ITAGmerge, geneLength, by = "itag", all = TRUE)
matrixGO[is.na(matrixGO)] <- 0
pat <- matrixGO
```

```r
#Now that we have the data in the right format we can proceed with GO enrichment.

#First specify vector to loop over for each column

sigType <- c("intersect", "wt", "tf2")

  for(type in sigType) {

    genes = as.integer(pat[,type])
    names(genes) = pat$itag
    table(genes)
    length(genes)

    pwf = nullp(genes,bias.data=pat$length)

    GO.wall = goseq(pwf,gene2cat = cate)
    head(GO.wall)

  #This is going to correct for multiple testing.  You can specify the p-value cut-off of GO categories

    enriched.GO = GO.wall$category[p.adjust(GO.wall$over_represented_pvalue, method = "BH") < 0.05]

    enriched.GO

    my.GO <- as.character(enriched.GO)
    my.GO.table <- Term(my.GO)
    my.GO.table
    t <- as.matrix(my.GO.table)

    print(type) #this is for the knitr document
    print(t) #this is for the knitr document
  }
}
```

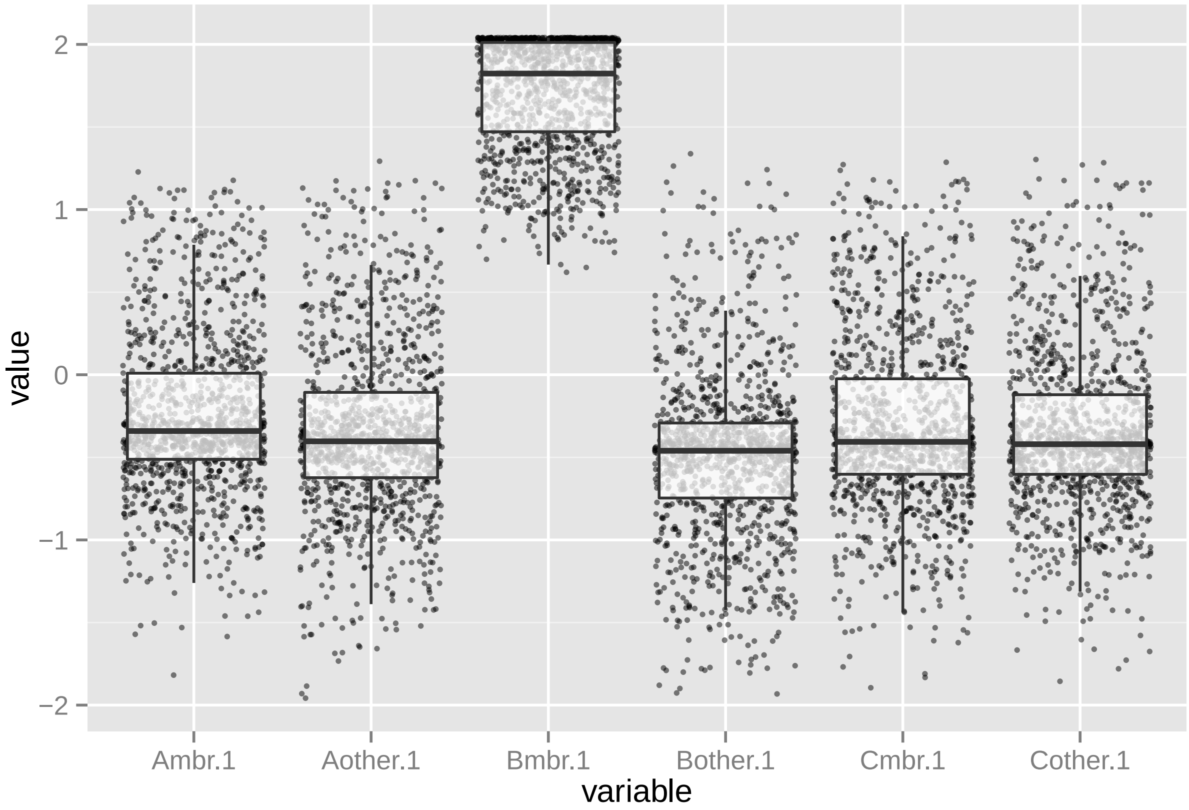
vennDiagram Function:

### Cluster 1

Sub cluster 1 is defined by up regulation of genes in Bmbr, which is the early leaflet region of the terminal leaflet.

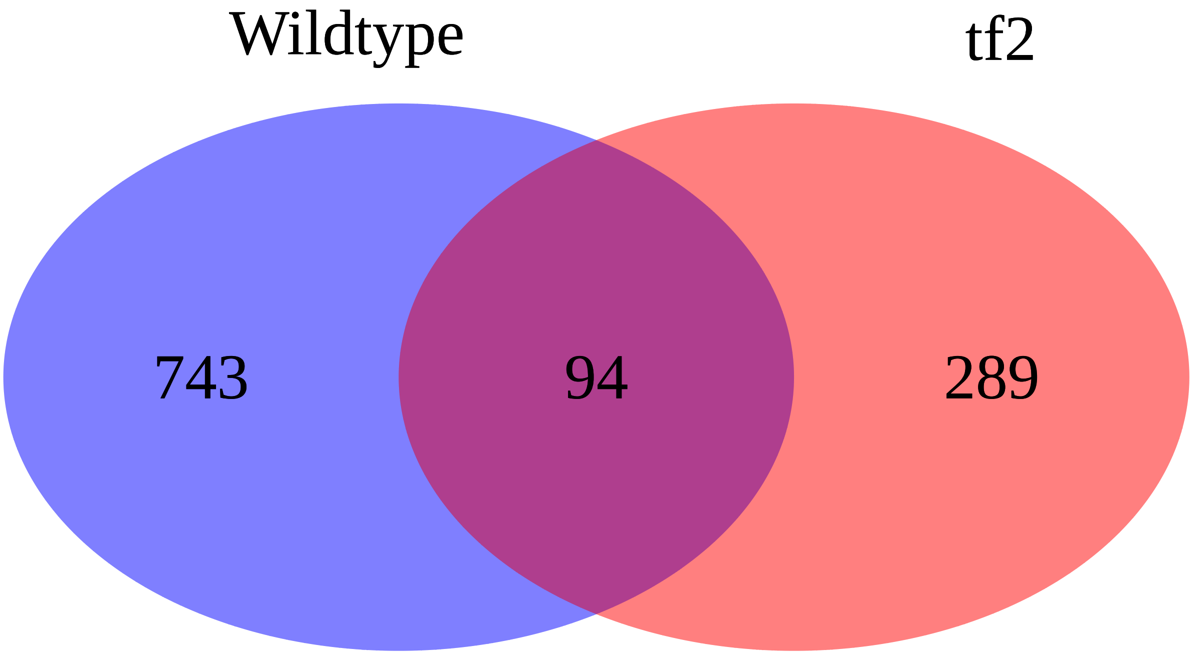
```r
clusterVis(1)
```

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



```
clusterNum(1)
```

```
## [1] "total number of genes in sub cluster is  1220"
## [1] "total number of genes in wt cluster is  837"
## [1] "total number of genes in tf2 cluster is  383"
## [1] "There are 94  that are the same between wt and tf2"
```



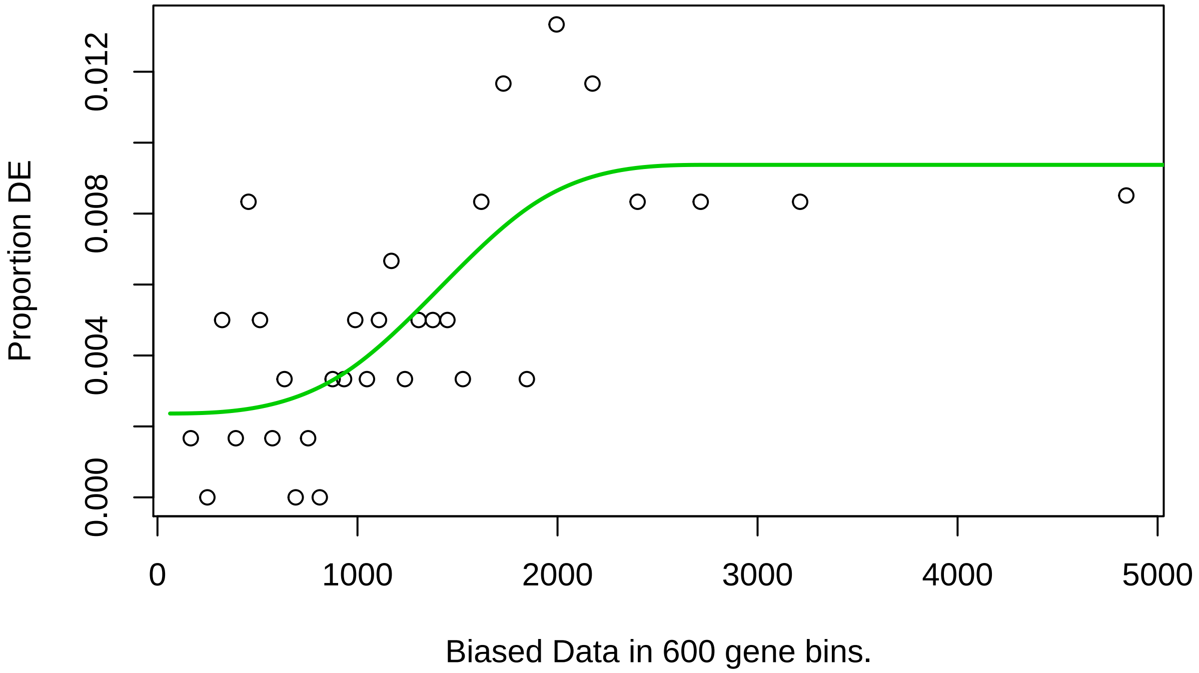
```
clusterGO(1)
```

```
## Warning: initial point very close to some inequality constraints
```

```
## 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...
```

```
## [1] "intersect"
##      [,1]
```

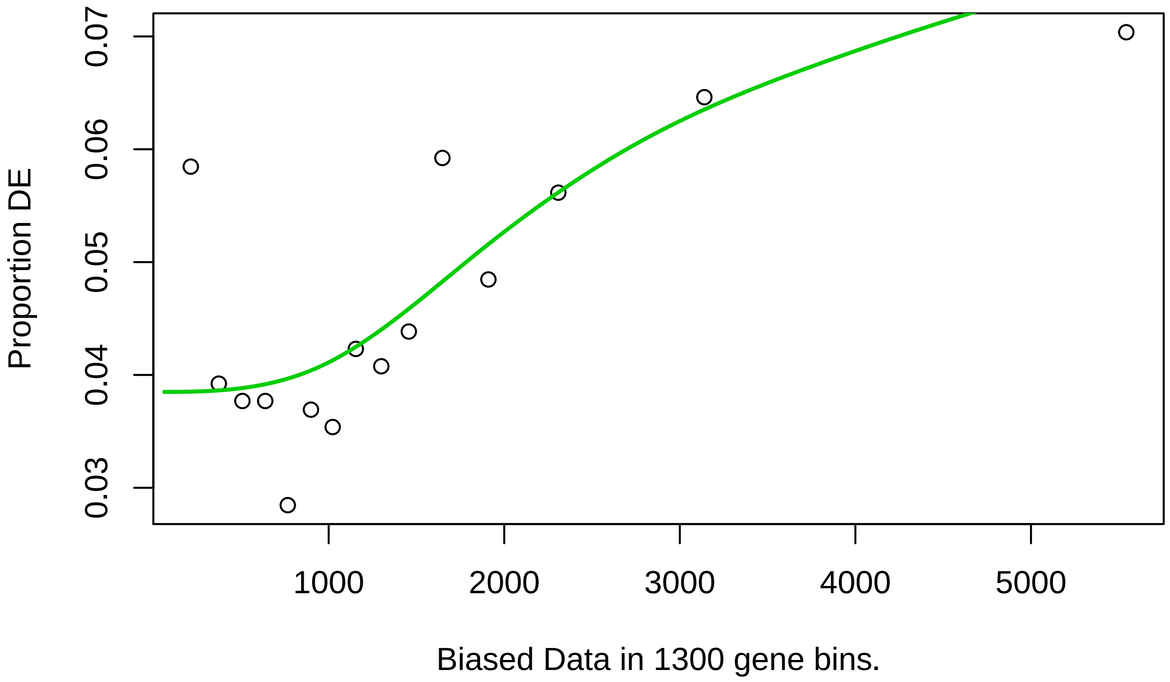
```
## Warning: initial point very close to some inequality constraints
```



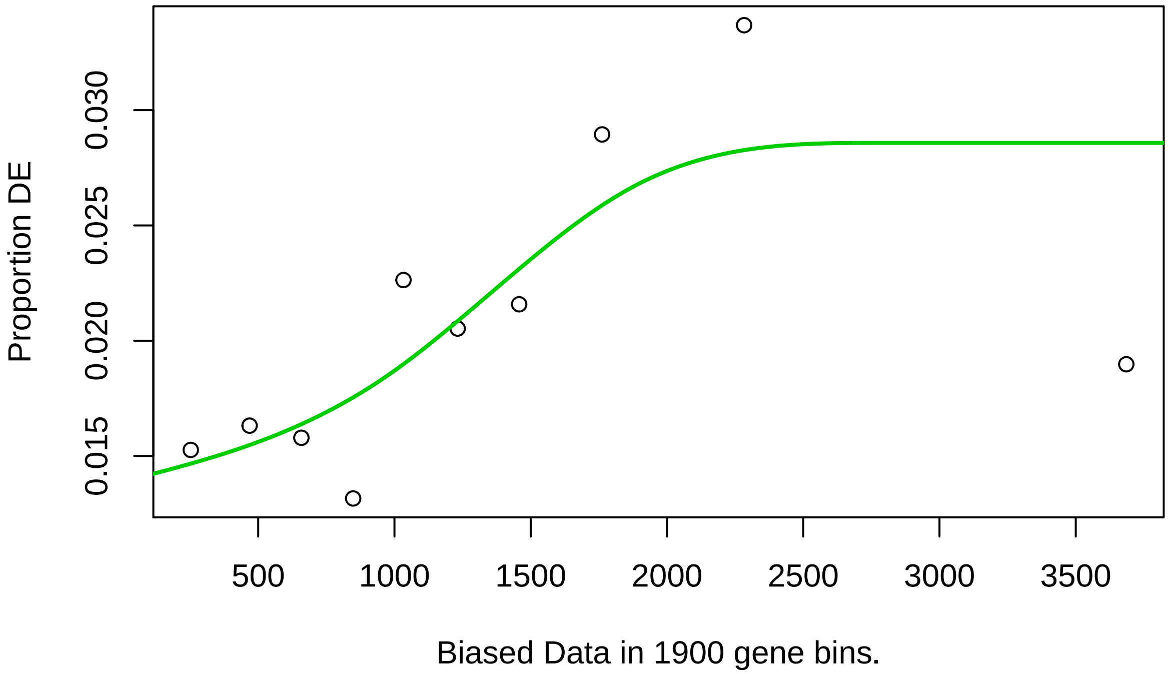
```
## Using manually entered categories.
## For 2936 genes, we could not find any categories. These genes will be excluded.
## To force their use, please run with use_genes_without_cat=TRUE (see documentation).
## This was the default behavior for version 1.15.1 and earlier.
## Calculating the p-values...
```

```
## [1] "wt"
##            [,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:0003682 "chromatin binding"
## GO:0000785 "chromatin"
## GO:0016651 "oxidoreductase activity, acting on NAD(P)H"
## GO:0031969 "chloroplast membrane"
## GO:0043229 "intracellular organelle"
## GO:0006310 "DNA recombination"
## GO:0009575 "chromoplast stroma"
## GO:0003899 "DNA-directed RNA polymerase activity"
## GO:0003723 "RNA binding"
## GO:0003677 "DNA binding"
## GO:0004190 "aspartic-type endopeptidase activity"
## GO:0048038 "quinone binding"
## GO:0006351 "transcription, DNA-templated"
## GO:0008270 "zinc ion binding"
## GO:0032549 "ribonucleoside binding"
## GO:0009926 "auxin polar transport"
## GO:0003676 "nucleic acid binding"
## GO:0005030 "neurotrophin receptor activity"
```

```
## Warning: initial point very close to some inequality constraints
```



```
## 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...
```



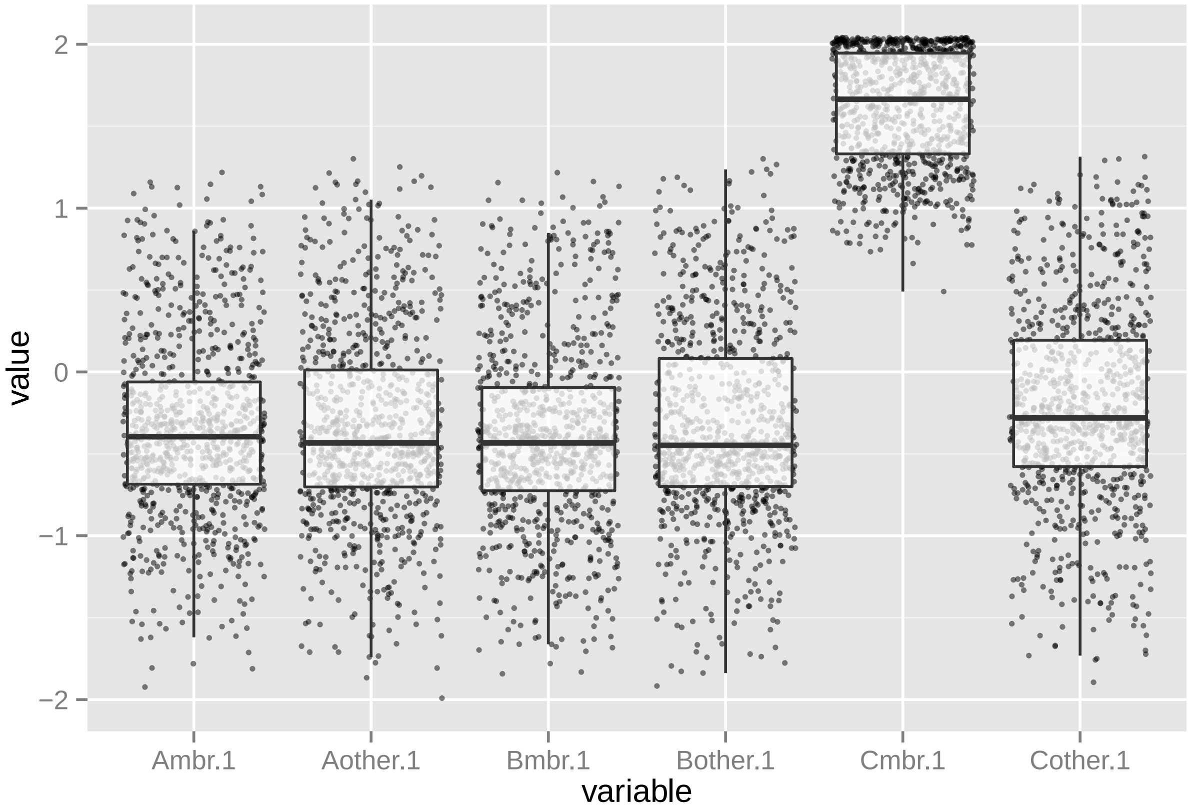
```
## [1] "tf2"
##      [,1]
```

### Cluster 2

Sub cluster 2 is defined by up regulation of genes in Cmbr, which is the base “marginal blastozone” region, which should be the most pluripotent in WT.

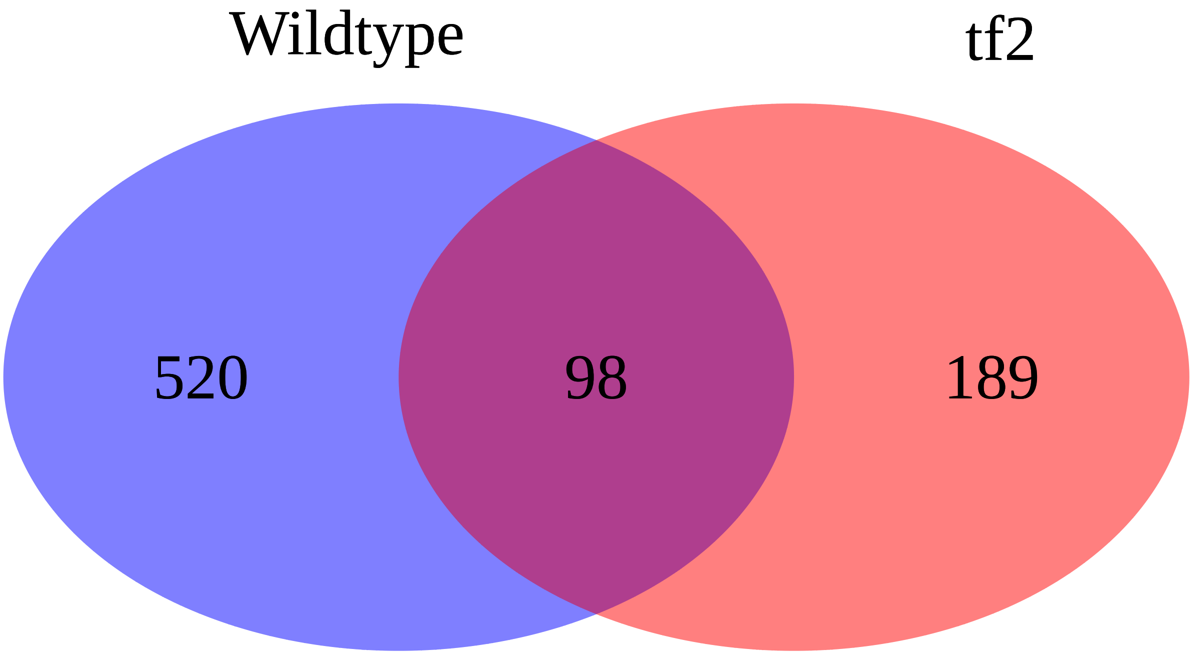
```
clusterVis(2)
```

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



```
clusterNum(2)
```

```
## [1] "total number of genes in sub cluster is  905"
## [1] "total number of genes in wt cluster is  618"
## [1] "total number of genes in tf2 cluster is  287"
## [1] "There are 98  that are the same between wt and tf2"
```



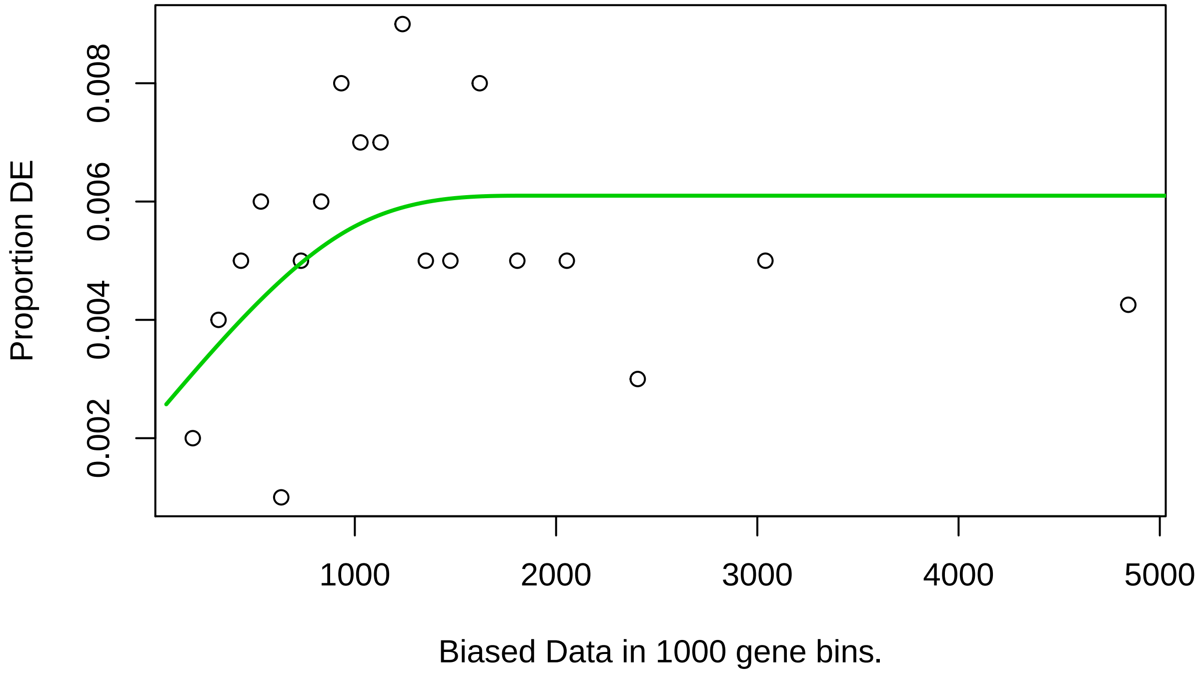
```
clusterGO(2)
```

```
## Warning: initial point very close to some inequality constraints
```

```
## 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...
```

```
## [1] "intersect"
##            [,1]
## GO:0003700 "sequence-specific DNA binding transcription factor activity"
## GO:0005667 "transcription factor complex"
## <NA>       NA
```

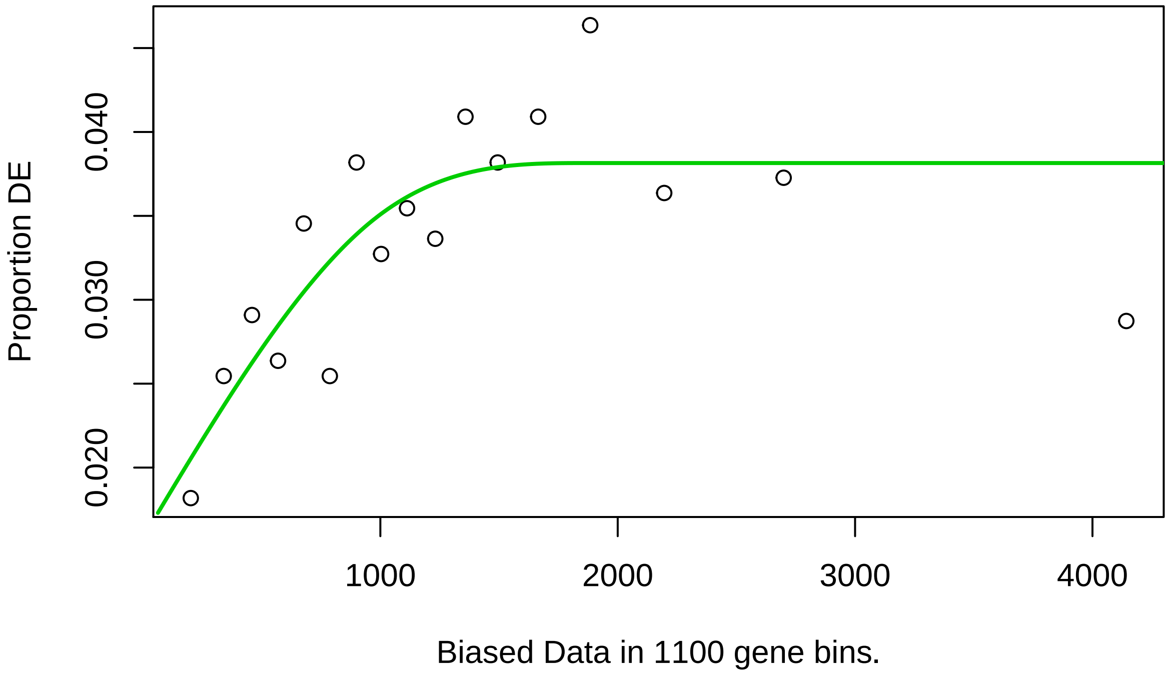
```
## Warning: initial point very close to some inequality constraints
```



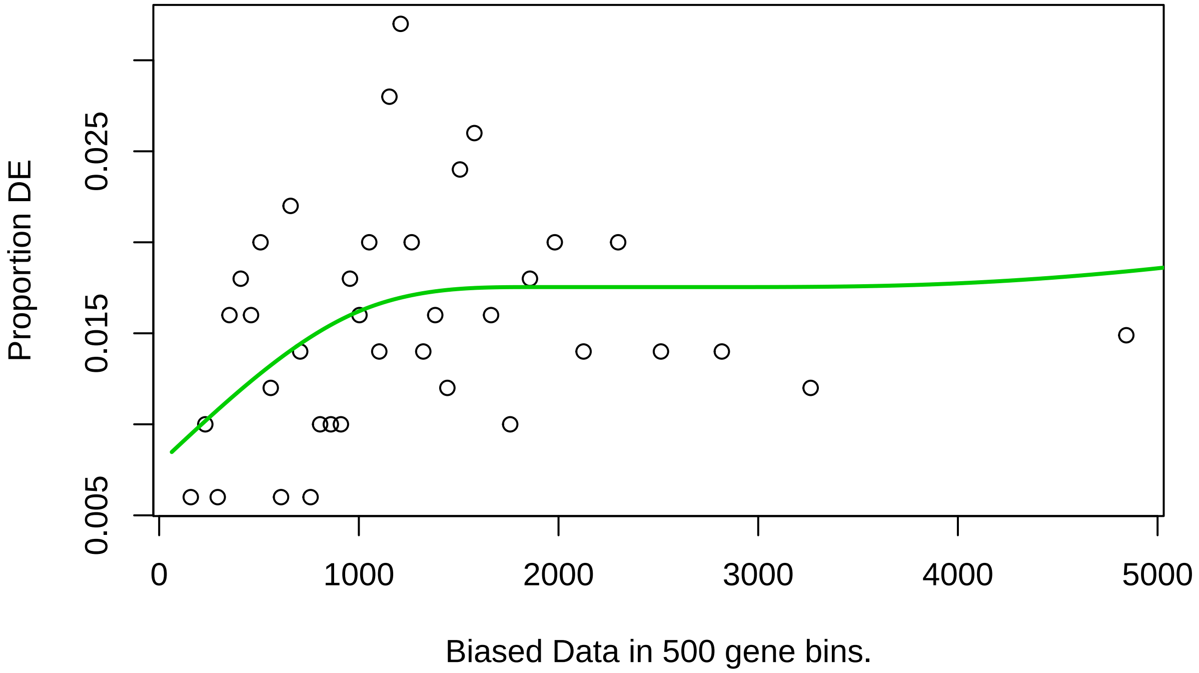
```
## Using manually entered categories.
## For 2936 genes, we could not find any categories. These genes will be excluded.
## To force their use, please run with use_genes_without_cat=TRUE (see documentation).
## This was the default behavior for version 1.15.1 and earlier.
## Calculating the p-values...
```

```
## [1] "wt"
##            [,1]
## GO:0003700 "sequence-specific DNA binding transcription factor activity"
```

```
## Warning: initial point very close to some inequality constraints
```



```
## 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...
```



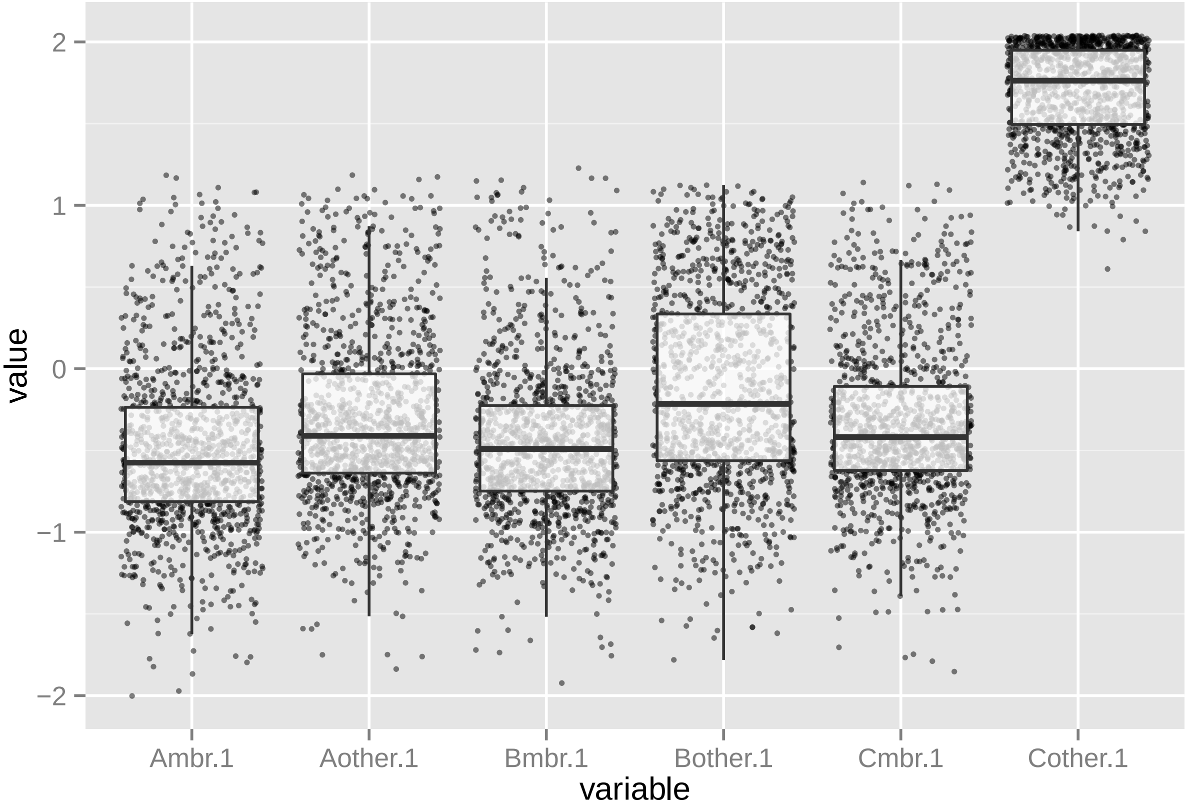
```
## [1] "tf2"
##            [,1]
## GO:0003700 "sequence-specific DNA binding transcription factor activity"
```

### Cluster 3

This cluster is specific to Cother, which is specific to the rachis region at the base.

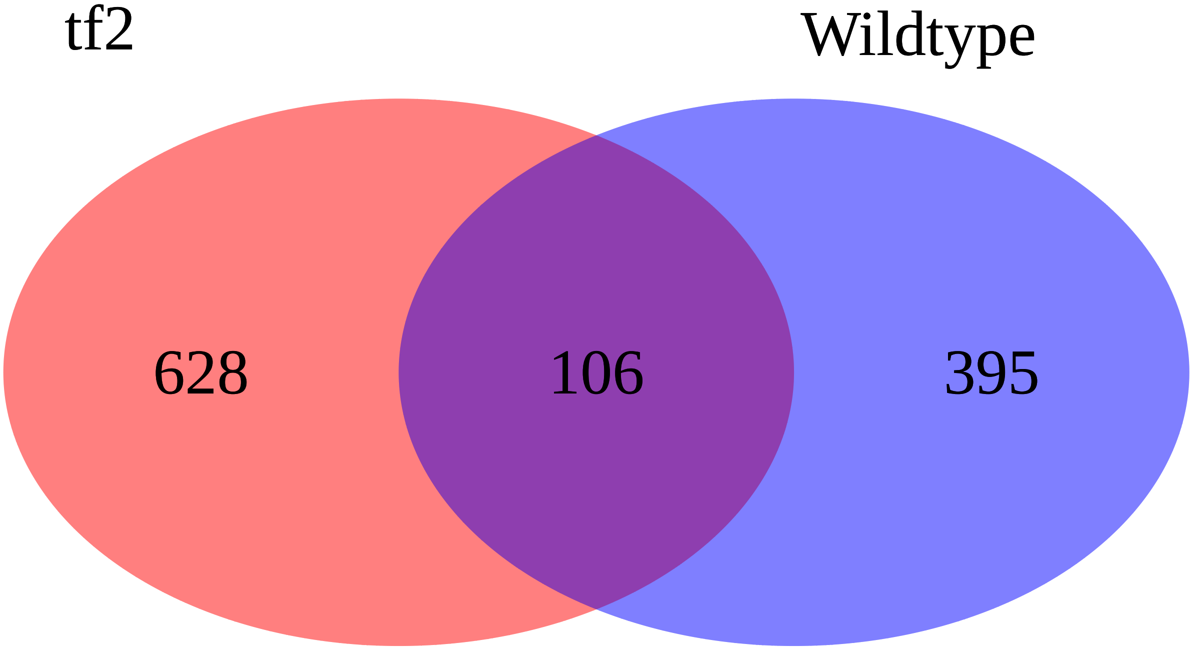
```
clusterVis(3)
```

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



```
clusterNum(3)
```

```
## [1] "total number of genes in sub cluster is  1235"
## [1] "total number of genes in wt cluster is  501"
## [1] "total number of genes in tf2 cluster is  734"
## [1] "There are 106  that are the same between wt and tf2"
```



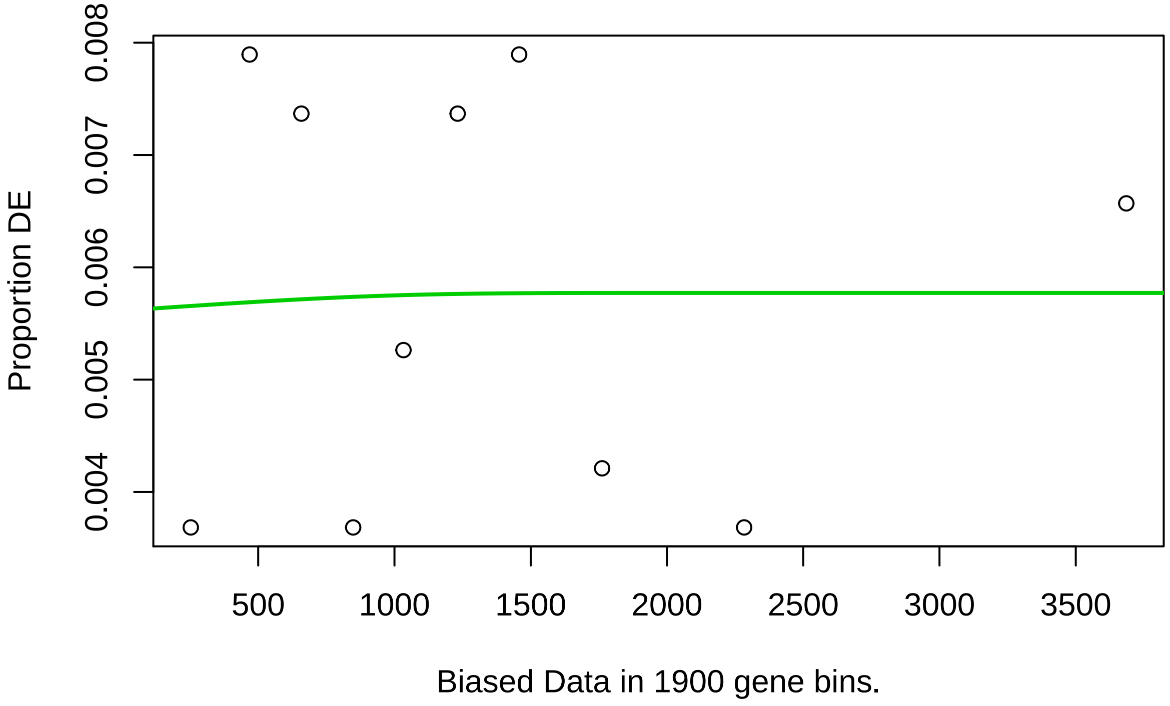
```
clusterGO(3)
```

```
## Warning: initial point very close to some inequality constraints
```

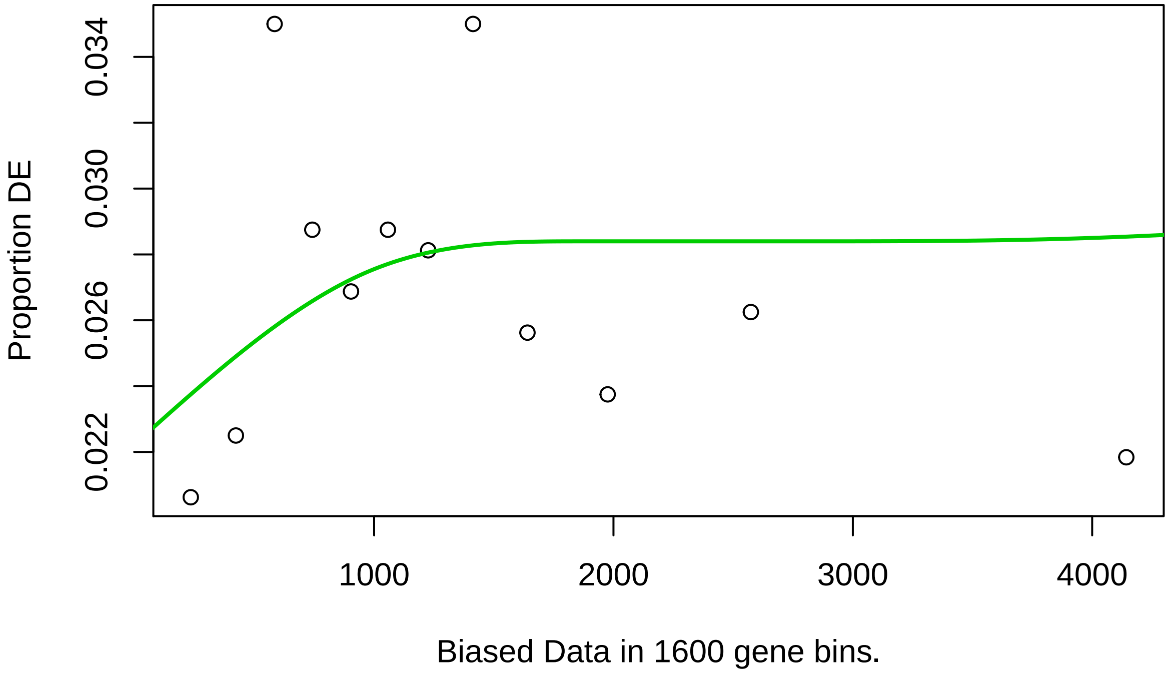
```
## Using manually entered categories.
## For 2936 genes, we could not find any categories. These genes will be excluded.
## To force their use, please run with use_genes_without_cat=TRUE (see documentation).
## This was the default behavior for version 1.15.1 and earlier.
## Calculating the p-values...
```

```
## [1] "intersect"
##      [,1]
```

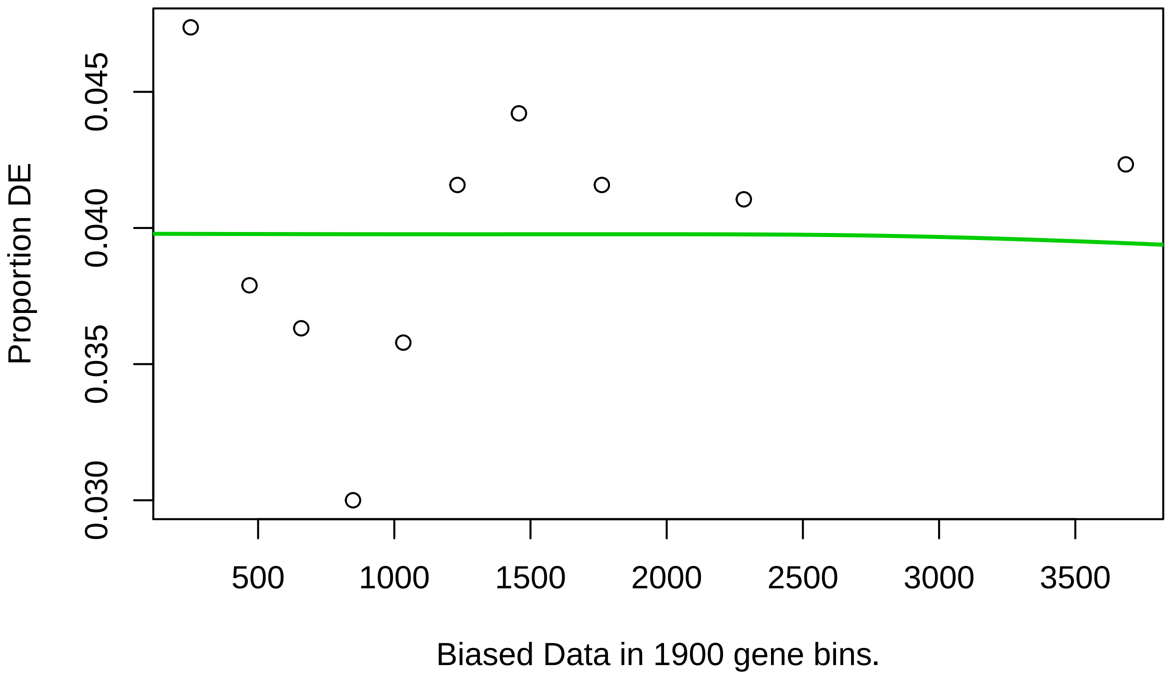
```
## Warning: initial point very close to some inequality constraints
```



```
## 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...
```



```
## [1] "wt"
##      [,1]
## <NA> NA
## 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...
```



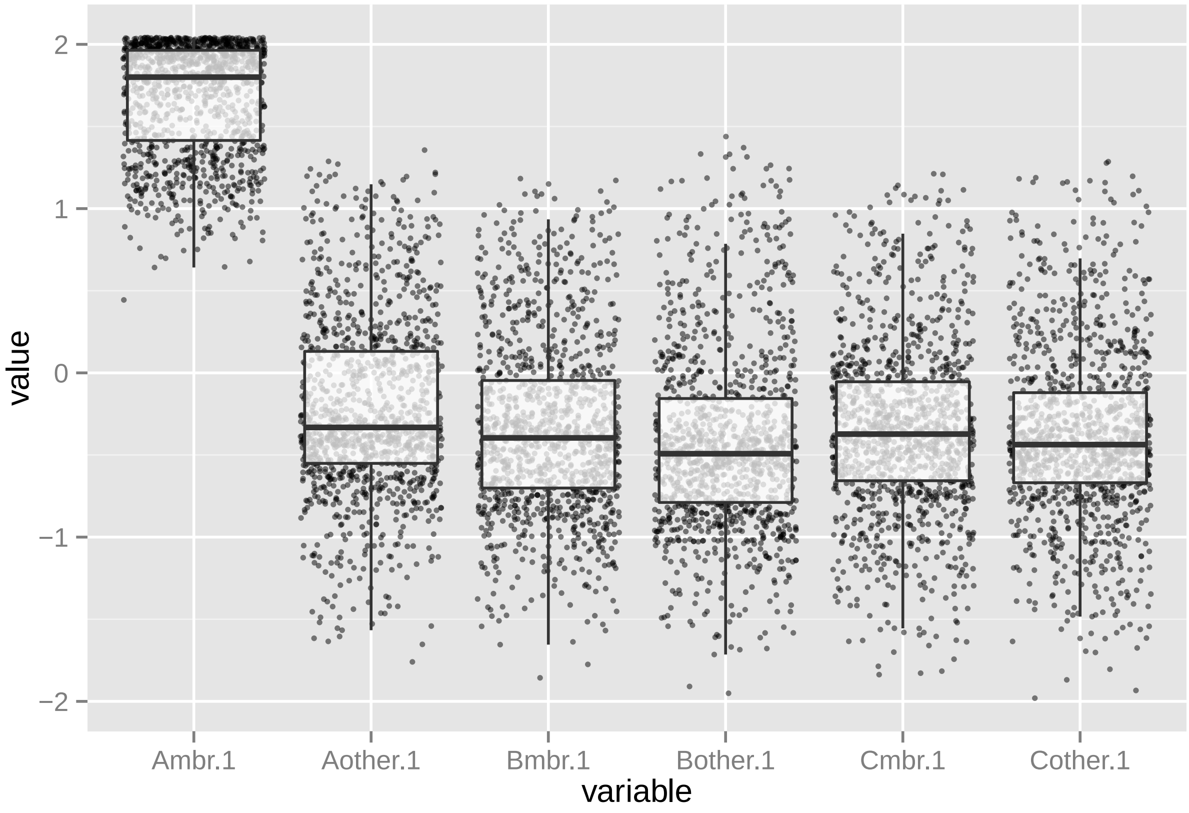
```
## [1] "tf2"
##            [,1]
## GO:0015074 "DNA integration"
## GO:0003964 "RNA-directed DNA polymerase activity"
## GO:0006278 "RNA-dependent DNA replication"
## GO:0006333 "chromatin assembly or disassembly"
## GO:0000785 "chromatin"
## GO:0003682 "chromatin binding"
## GO:0008270 "zinc ion binding"
## GO:0043229 "intracellular organelle"
## GO:0003677 "DNA binding"
## GO:0004190 "aspartic-type endopeptidase activity"
## GO:0003723 "RNA binding"
## GO:0031969 "chloroplast membrane"
```

### Cluster 4

This cluster has genes that are preferentially up-regulated in Ambr, which is the tip most region that becomes the terminal leaflet. This is the terminal leaflet blade region

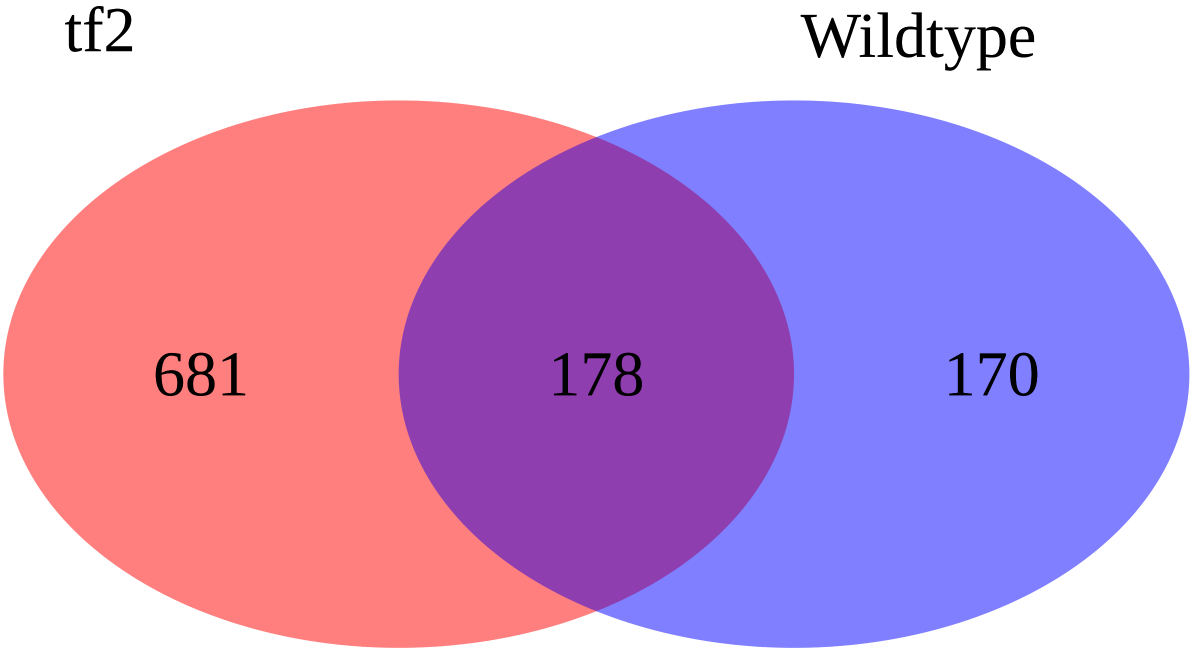
```
clusterVis(4)
```

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



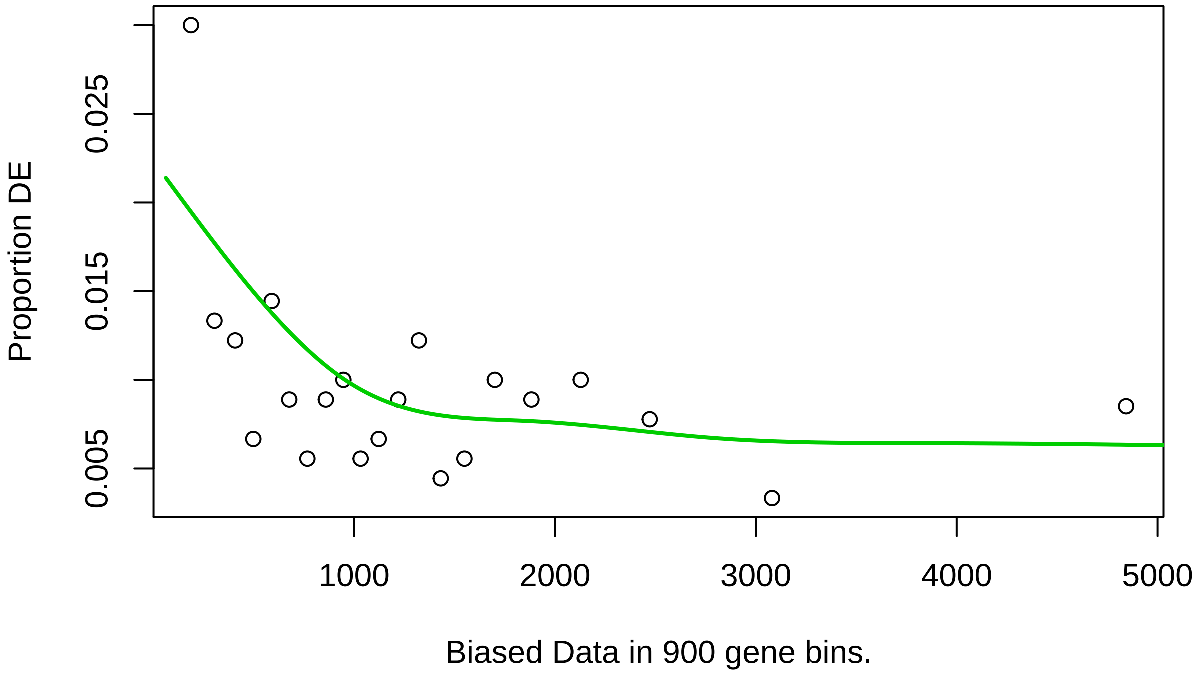
```
clusterNum(4)
```

```
## [1] "total number of genes in sub cluster is  1207"
## [1] "total number of genes in wt cluster is  348"
## [1] "total number of genes in tf2 cluster is  859"
## [1] "There are 178  that are the same between wt and tf2"
```



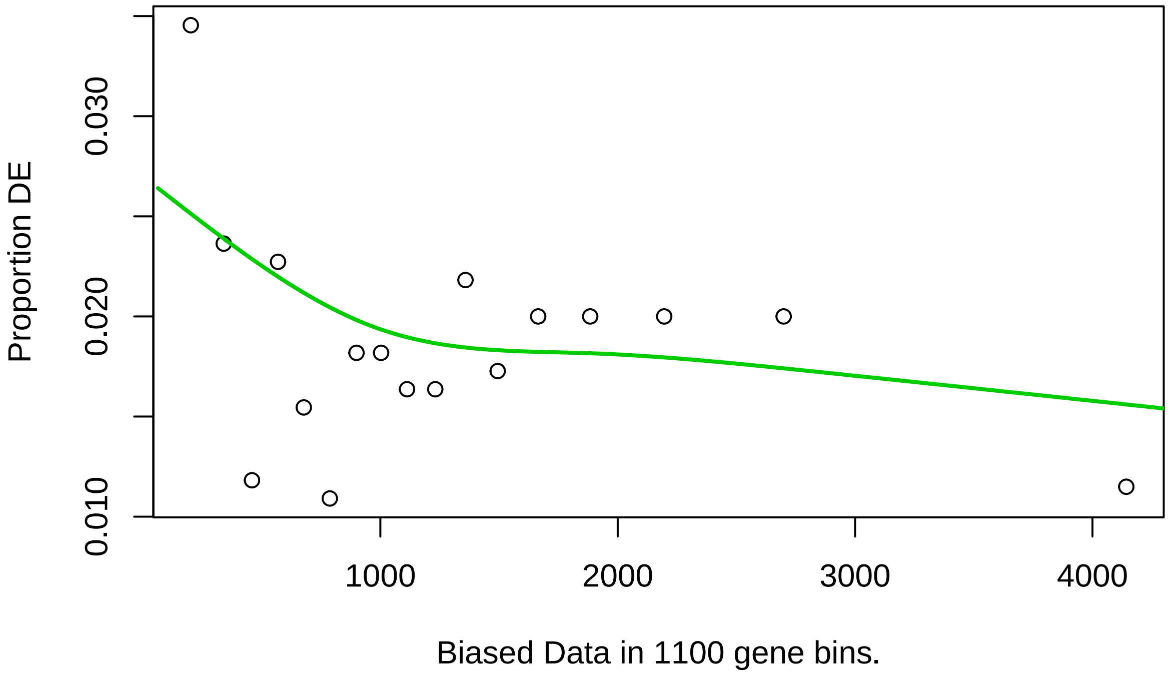
```
clusterGO(4)
```

```
## 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...
```



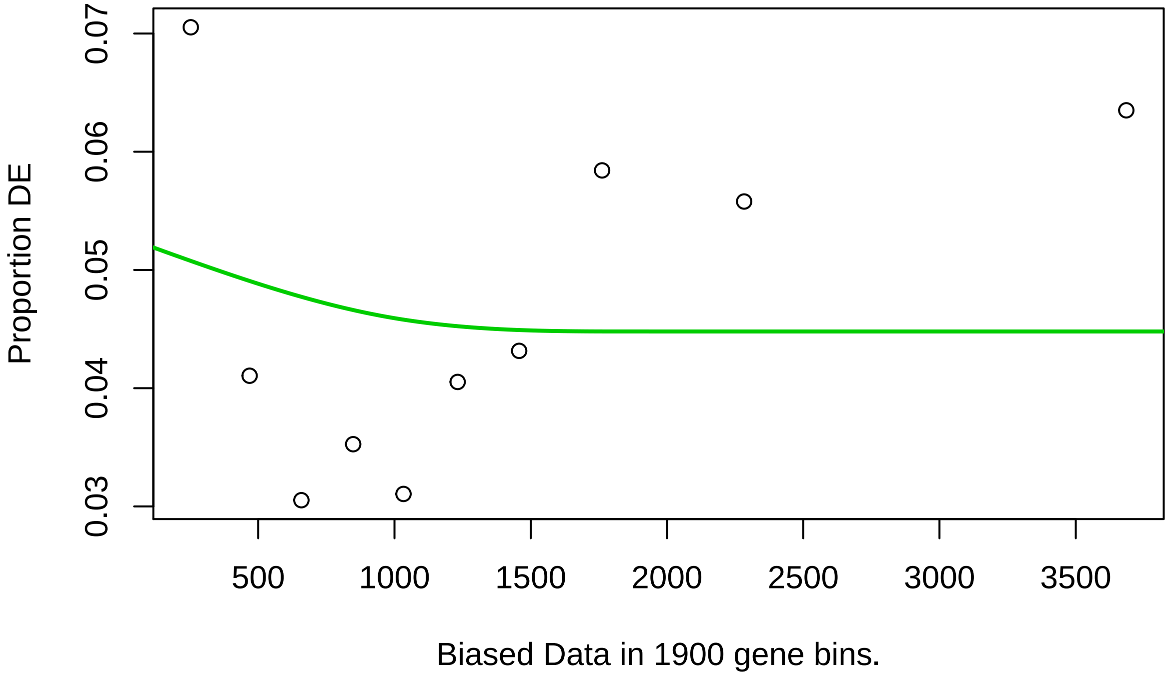
```
## [1] "intersect"
##            [,1]
## GO:0004397 "histidine ammonia-lyase activity"
## GO:0045548 "phenylalanine ammonia-lyase activity"
## GO:0009813 "flavonoid biosynthetic process"
## GO:0016841 "ammonia-lyase activity"
## <NA>       NA
```

```
## 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...
```



```
## [1] "wt"
##      [,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...
```



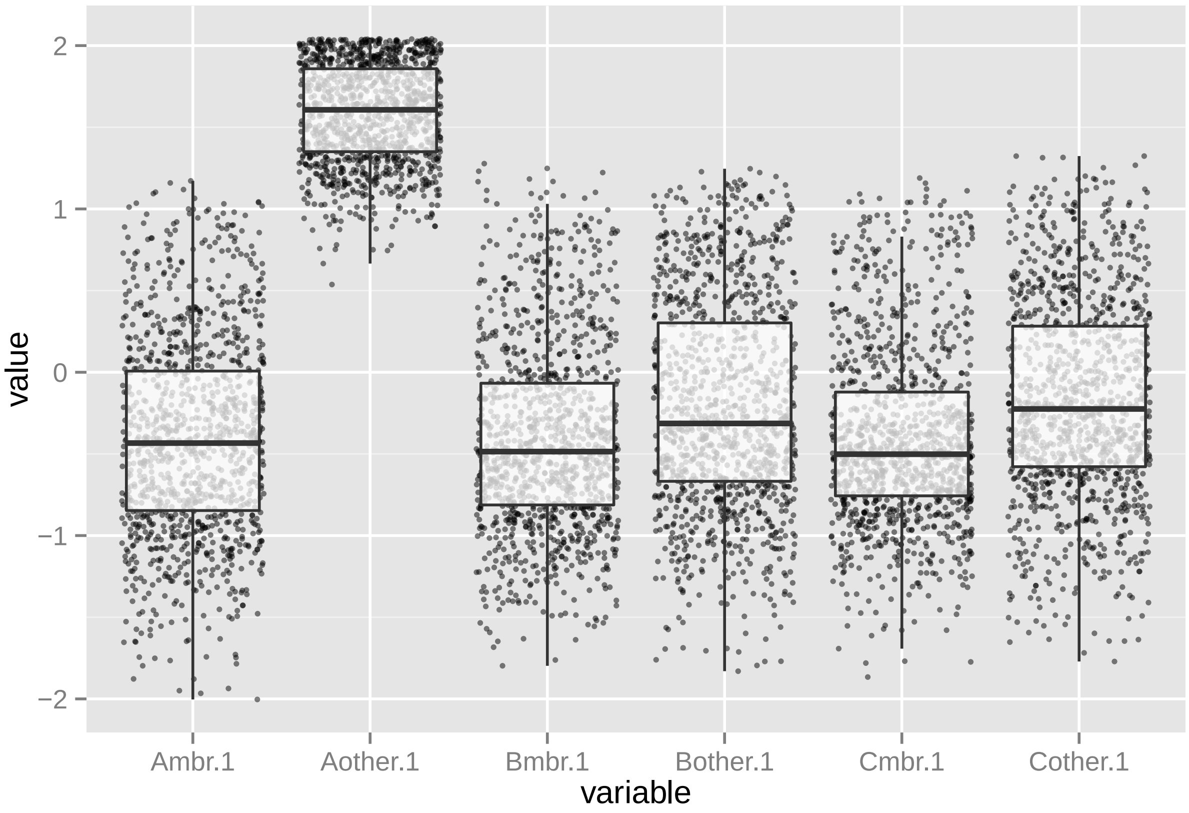
```
## [1] "tf2"
##            [,1]
## GO:0009575 "chromoplast stroma"
## GO:0004397 "histidine ammonia-lyase activity"
## GO:0045548 "phenylalanine ammonia-lyase activity"
## GO:0016841 "ammonia-lyase activity"
## GO:0010466 "negative regulation of peptidase activity"
## GO:0006559 "L-phenylalanine catabolic process"
## GO:0004867 "serine-type endopeptidase inhibitor activity"
## GO:0009698 "phenylpropanoid metabolic process"
```

### Cluster 5

The cluster ihas genes that are preferentially up-regulated in Aother, which is the rachis region at the tip; what will eventually become the midvien of the terminal leaflet.

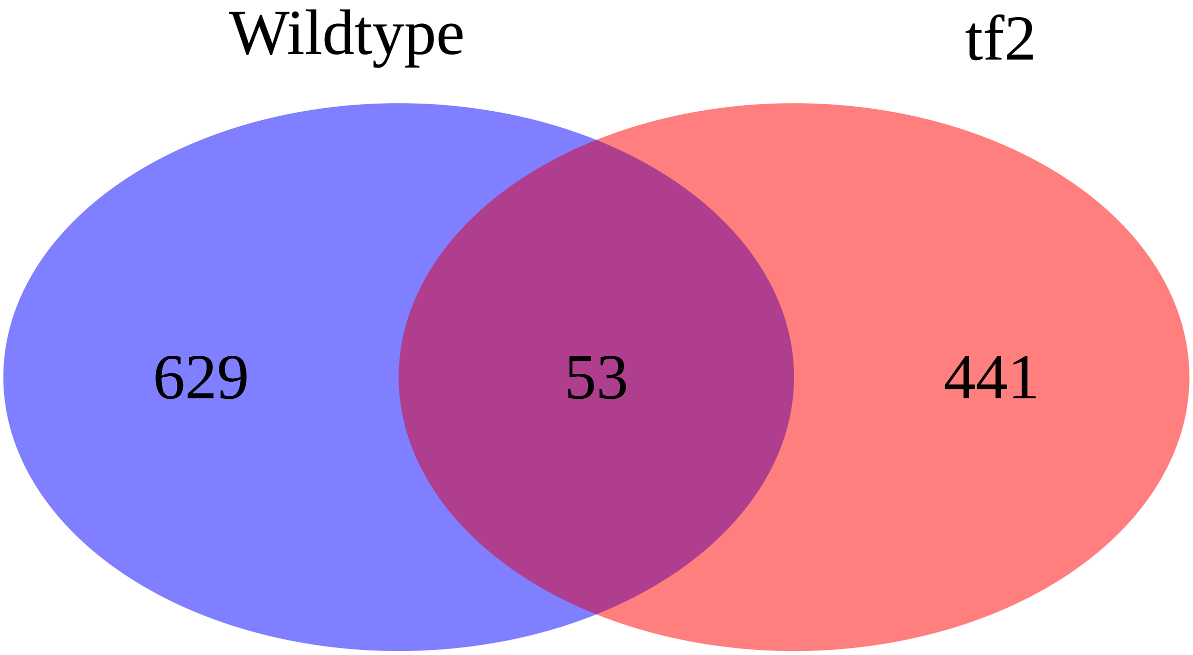
```
clusterVis(5)
```

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



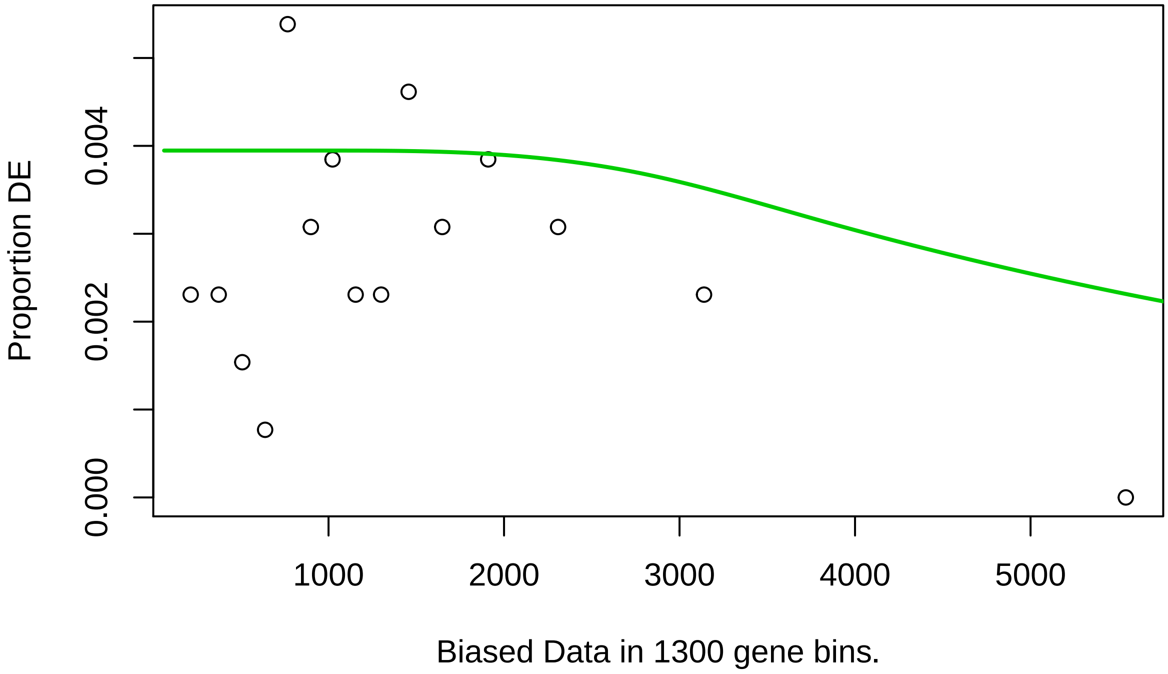
```
clusterNum(5)
```

```
## [1] "total number of genes in sub cluster is  1176"
## [1] "total number of genes in wt cluster is  682"
## [1] "total number of genes in tf2 cluster is  494"
## [1] "There are 53  that are the same between wt and tf2"
```



```
clusterGO(5)
```

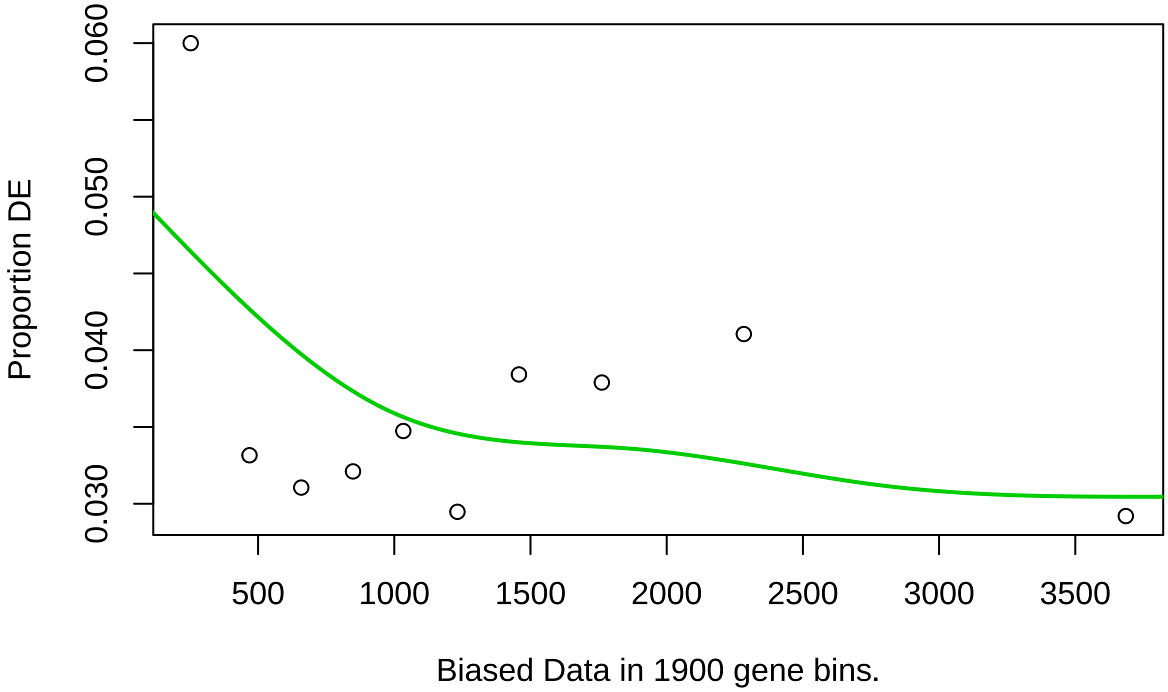
```
## Using manually entered categories.
## For 2936 genes, we could not find any categories. These genes will be excluded.
## To force their use, please run with use_genes_without_cat=TRUE (see documentation).
## This was the default behavior for version 1.15.1 and earlier.
## Calculating the p-values...
```



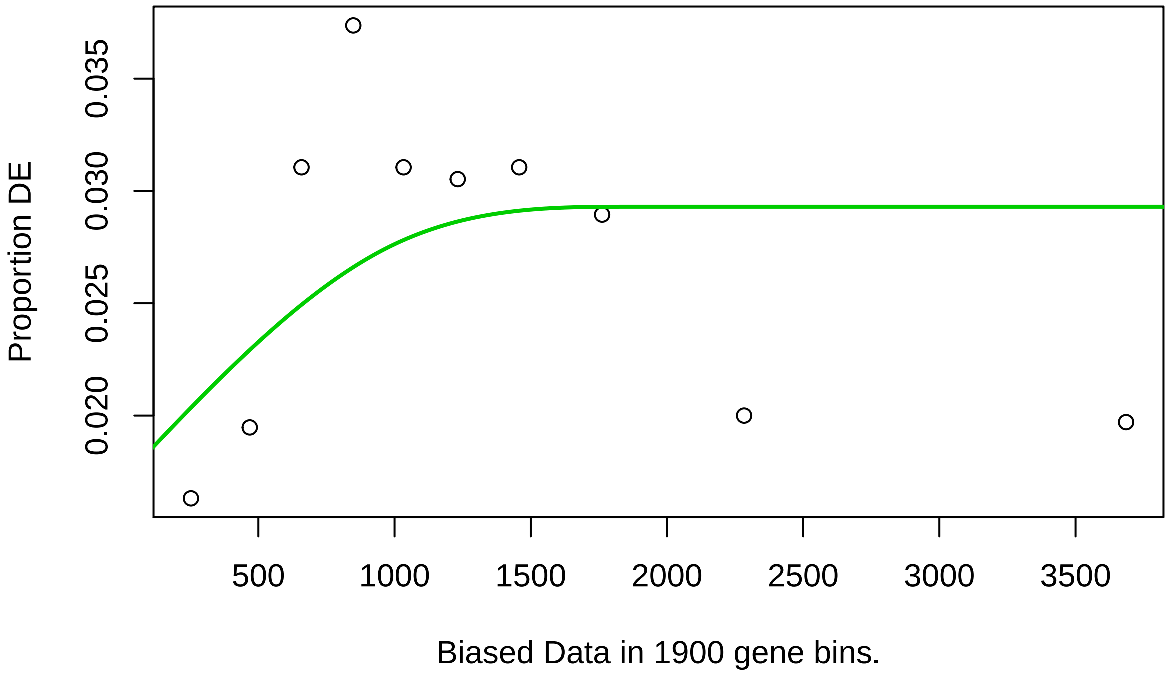
```
## [1] "intersect"
##      [,1]
## Using manually entered categories.
## For 2936 genes, we could not find any categories. These genes will be excluded.
## To force their use, please run with use_genes_without_cat=TRUE (see documentation).
## This was the default behavior for version 1.15.1 and earlier.
## Calculating the p-values...
```

```
## [1] "wt"
##      [,1]
```

```
## Warning: initial point very close to some inequality constraints
```



```
## 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...
```



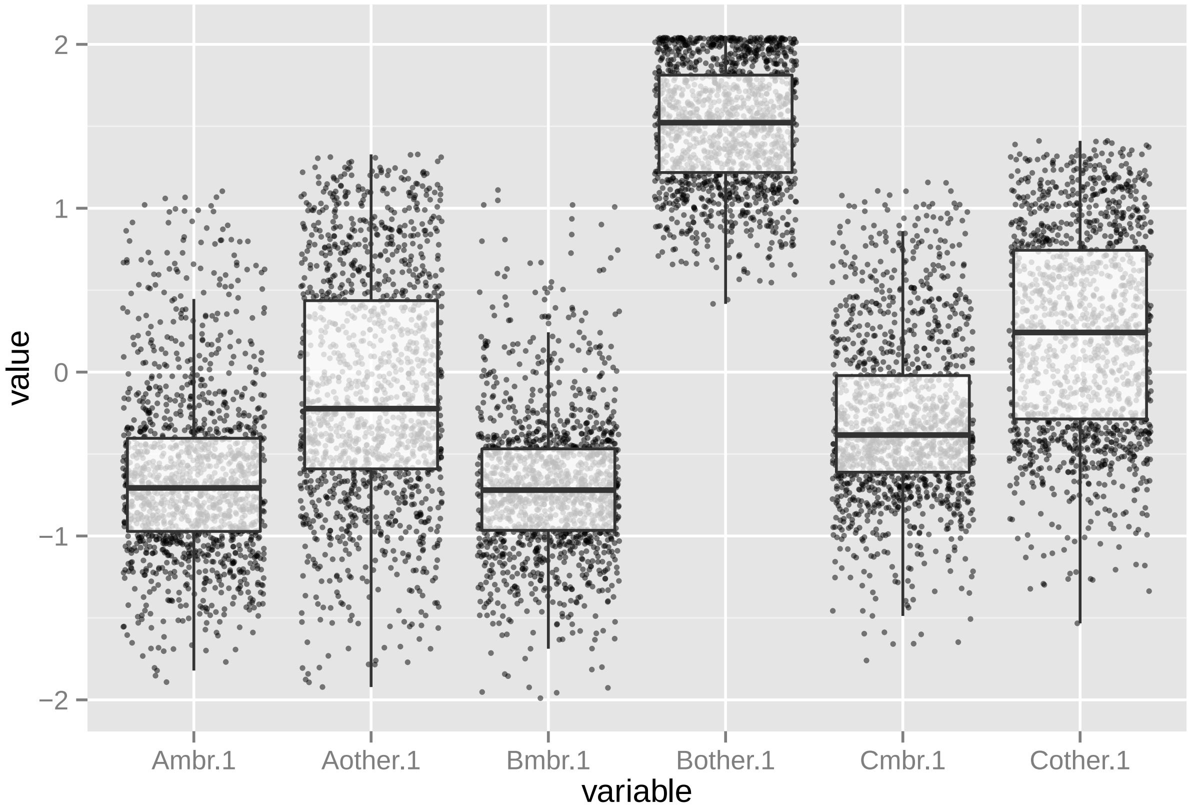
```
## [1] "tf2"
##            [,1]
## GO:0009523 "photosystem II"
## GO:0009765 "photosynthesis, light harvesting"
## GO:0016168 "chlorophyll binding"
## GO:0018298 "protein-chromophore linkage"
## GO:0009772 "photosynthetic electron transport in photosystem II"
## GO:0009522 "photosystem I"
## GO:0045156 "electron transporter, transferring electrons within the cyclic electron transport pathway
## GO:0030077 "plasma membrane light-harvesting complex"
## GO:0009535 "chloroplast thylakoid membrane"
## GO:0030076 "light-harvesting complex"
## GO:0005985 "sucrose metabolic process"
## GO:0005982 "starch metabolic process"
## GO:0016021 "integral component of membrane"
## GO:0042973 "glucan endo-1,3-beta-D-glucosidase activity"
```

### Cluster 6

The cluster ihas genes that are preferentially up-regulated in Bother, which is the rachis region at site of leaflet initiation.

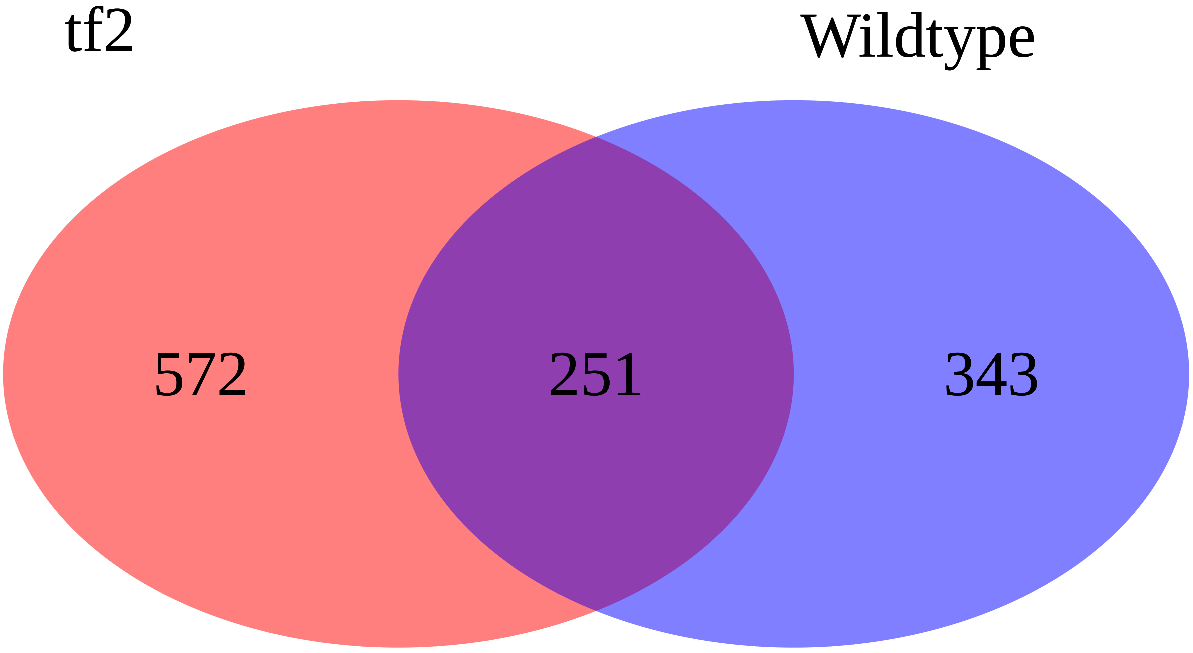
```
clusterVis(6)
```

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



```
clusterNum(6)
```

```
## [1] "total number of genes in sub cluster is  1417"
## [1] "total number of genes in wt cluster is  594"
## [1] "total number of genes in tf2 cluster is  823"
## [1] "There are 251  that are the same between wt and tf2"
```



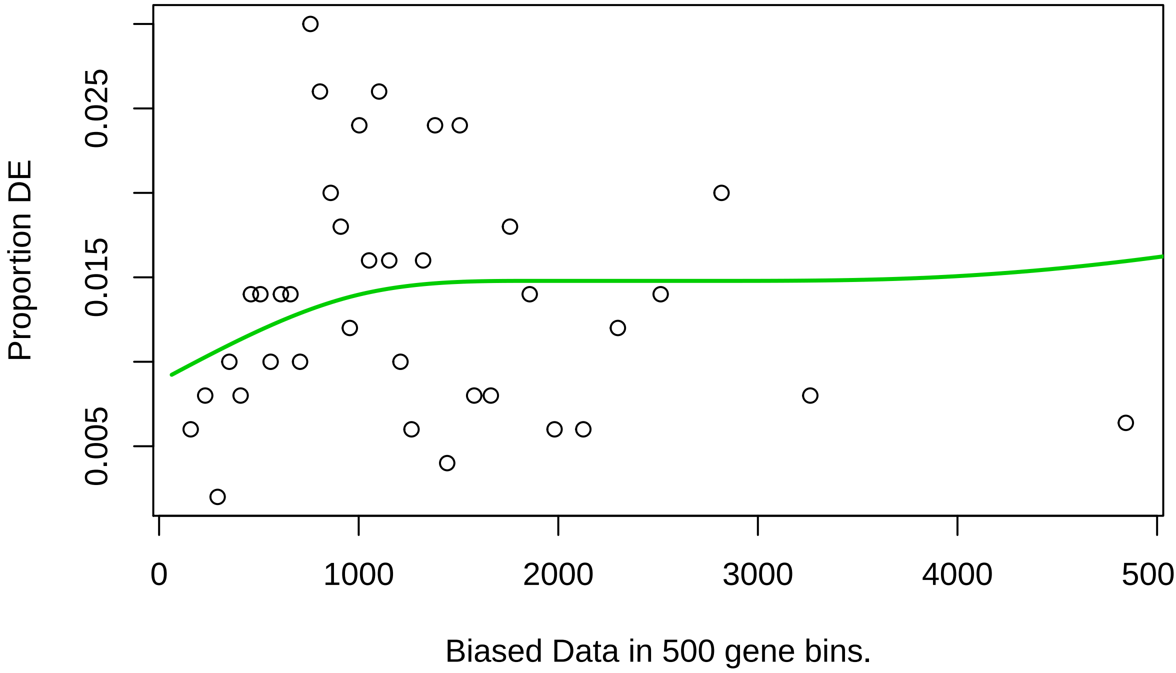
```
clusterGO(6)
```

```
## Warning: initial point very close to some inequality constraints
```

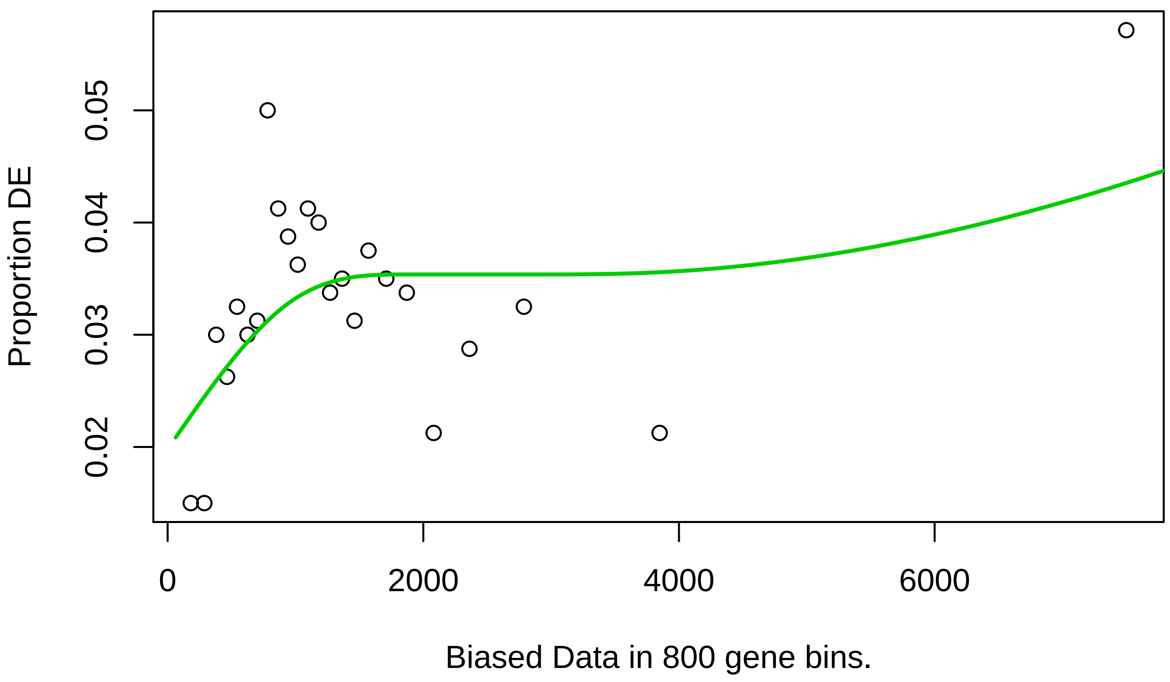
```
## Using manually entered categories.
## For 2936 genes, we could not find any categories. These genes will be excluded.
## To force their use, please run with use_genes_without_cat=TRUE (see documentation).
## This was the default behavior for version 1.15.1 and earlier.
## Calculating the p-values...
```

```
## [1] "intersect"
##            [,1]
## GO:0015250 "water channel activity"
## GO:0009535 "chloroplast thylakoid membrane"
## GO:0010067 "procambium histogenesis"
## GO:0006833 "water transport"
## GO:0009768 "photosynthesis, light harvesting in photosystem I"
## GO:0009523 "photosystem II"
## GO:0030076 "light-harvesting complex"
## GO:0016021 "integral component of membrane"
```

```
## Warning: initial point very close to some inequality constraints
```

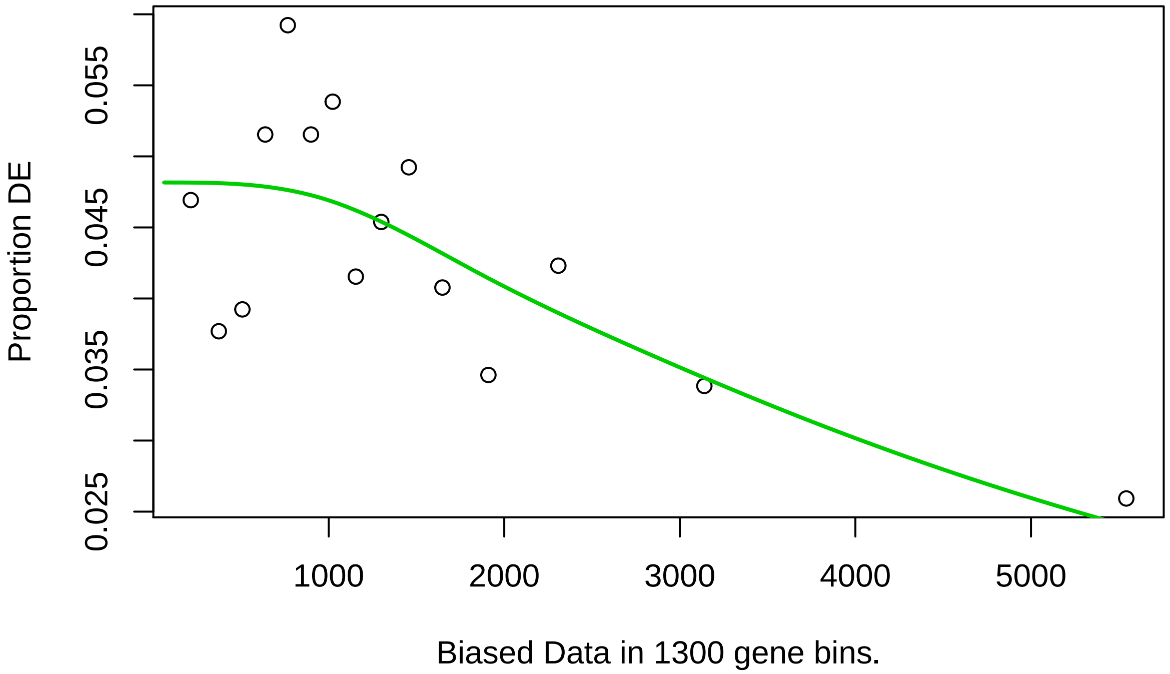


```
## Using manually entered categories.
## For 2936 genes, we could not find any categories. These genes will be excluded.
## To force their use, please run with use_genes_without_cat=TRUE (see documentation).
## This was the default behavior for version 1.15.1 and earlier.
## Calculating the p-values...
```



```
## [1] "wt"
##            [,1]
## GO:0009523 "photosystem II"
## GO:0009765 "photosynthesis, light harvesting"
## GO:0018298 "protein-chromophore linkage"
## GO:0009522 "photosystem I"
## GO:0009535 "chloroplast thylakoid membrane"
## GO:0016168 "chlorophyll binding"
## GO:0016021 "integral component of membrane"
## GO:0005985 "sucrose metabolic process"
## GO:0030076 "light-harvesting complex"
## GO:0005982 "starch metabolic process"
## GO:0046872 "metal ion binding"
## GO:0006833 "water transport"
## GO:0055085 "transmembrane transport"
## GO:0015250 "water channel activity"
## GO:0016020 "membrane"
## GO:0030077 "plasma membrane light-harvesting complex"
## GO:0042807 "central vacuole"
## GO:0010067 "procambium histogenesis"
## GO:0009768 "photosynthesis, light harvesting in photosystem I"
## GO:0010287 "plastoglobule"
```

```
## 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...
```



```
## [1] "tf2"
##            [,1]
## <NA>       NA
## GO:0009535 "chloroplast thylakoid membrane"
## GO:0003700 "sequence-specific DNA binding transcription factor activity"
## GO:0009523 "photosystem II"
```