Super SOM - Large

Purpose

The purpose of this analysis is to make a superSOM. This is a 6 x 6 hexogonal plot.

```
library(ggplot2)
```

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

```
library(reshape)
```

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

```
library(plyr)
```

```
##
## Attaching package: 'plyr'
```

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

```
library(kohonen)
```

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

```
source("../r/clusterFunctions.R")
```

PCA

Upload that dataset:

```
genes25 <- read.csv("../data/output/analysis4.top25_19Oct2017.csv")

genes25 <- genes25[,c(2:14)]
m.genes25 <- melt(genes25)
```

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

```
# head(m.genes25)

names(m.genes25) <- c("gene", "sample", "mean")

#set genotype

m.genes25$genotype <- ifelse(grepl("wt", m.genes25$sample, ignore.case = T), "wt",
           ifelse(grepl("tf2", m.genes25$sample, ignore.case = T), "tf2", "unknown"))

#set tissue

m.genes25$tissue <- ifelse(grepl("other", m.genes25$sample, ignore.case = T), "other",
           ifelse(grepl("mbr", m.genes25$sample, ignore.case = T), "mbr", "unknown"))

  #Set Region
m.genes25$region <- ifelse(grepl("a", m.genes25$sample, ignore.case = T), "A",
           ifelse(grepl("c", m.genes25$sample, ignore.case = T), "C", "B"))
```

```r
#Set type

m.genes25$type <- paste(m.genes25$region, m.genes25$tissue,  sep = "")

m.genes25.sub <- m.genes25[,c(1,7,4,3)]
# head(m.genes25.sub)

#Change from long to wide data format
m.genes25.long <- cast(m.genes25.sub, genotype + gene ~ type, value.var = mean, fun.aggregate = "mean")
```

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

```r
m.genes25.long <- as.data.frame(m.genes25.long)
```

Scaling the Data seperately

```r
# head(m.genes25.long)
wt <- subset(m.genes25.long, genotype == "wt")
tf2 <- subset(m.genes25.long, genotype == "tf2")

#transformation.
scale_data.wt <- as.matrix(t(scale(t(wt[c(3:8)]))))
scale_data.tf2 <- as.matrix(t(scale(t(tf2[c(3:8)]))))

scale_data_sep <- rbind(scale_data.tf2,scale_data.wt)
```

Continuing on with PCA

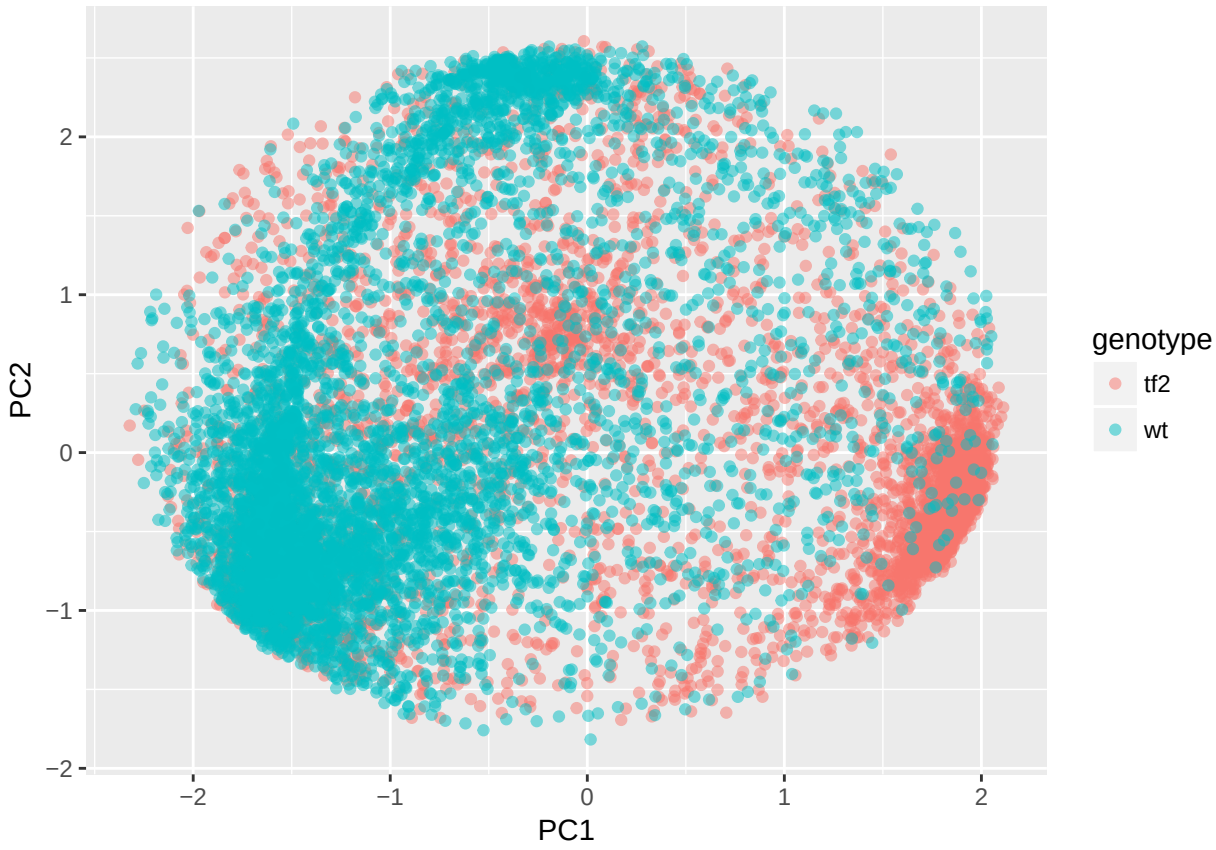
```r
pca_sep <- prcomp(scale_data_sep)

summary(pca_sep)
```

```
## Importance of components:
##                           PC1    PC2    PC3    PC4     PC5       PC6
## Standard deviation     1.4430 0.9236 0.9121 0.7283 0.59891 1.169e-15
## Proportion of Variance 0.4472 0.1832 0.1787 0.1139 0.07704 0.000e+00
## Cumulative Proportion  0.4472 0.6304 0.8090 0.9230 1.00000 1.000e+00
```

```r
pca.scores_sep <- data.frame(pca_sep$x)

data.val_sep <- cbind(m.genes25.long, scale_data_sep, pca.scores_sep)
```

Visualizing the PCA

```r
p <- ggplot(data.val_sep, aes(PC1, PC2, color = genotype))
p + geom_point(alpha = 0.5)
```



Scaling the data together

```
#transformation
scale_data_tog <- as.matrix(t(scale(t(m.genes25.long[c(3:8)]))))
```

Continuing on with PCA

```
pca_tog <- prcomp(scale_data_tog)

summary(pca_tog)
```

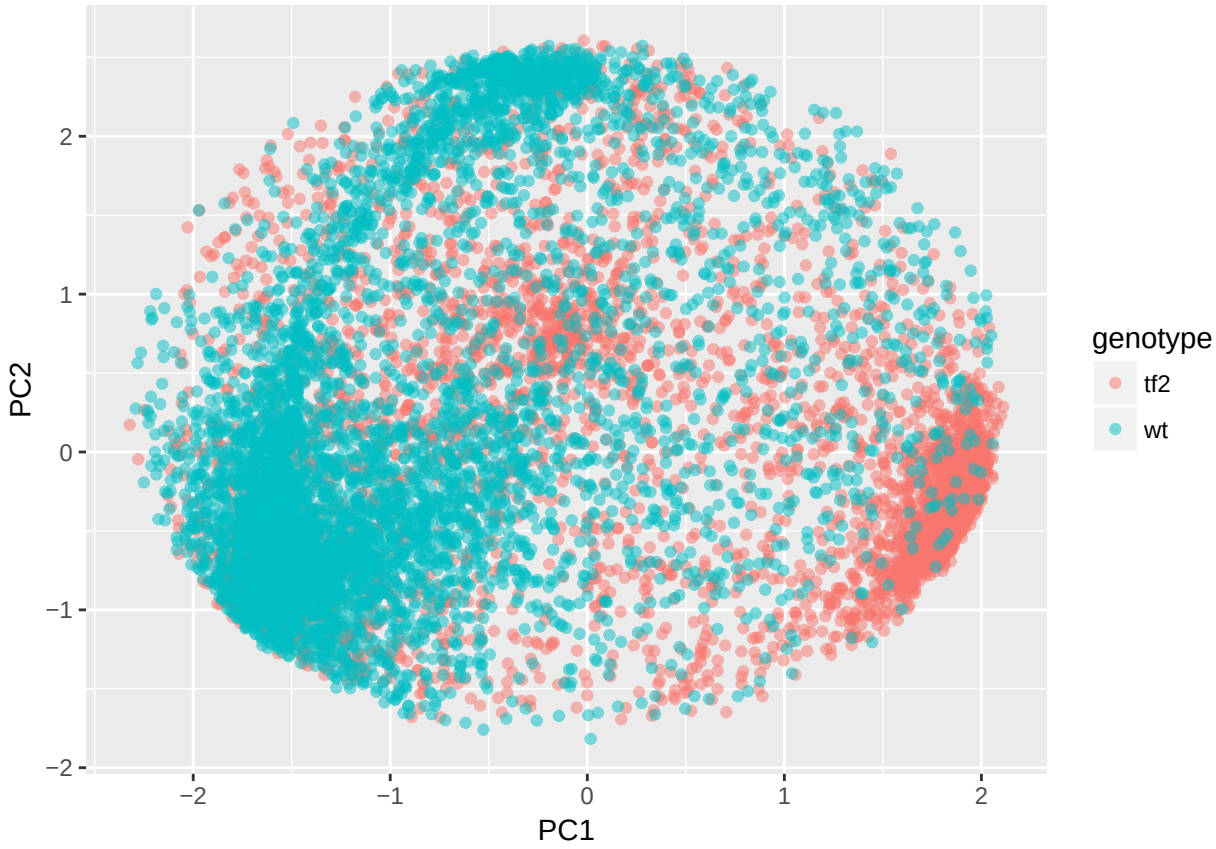
```
## Importance of components:
##                           PC1    PC2    PC3    PC4     PC5       PC6
## Standard deviation     1.4430 0.9236 0.9121 0.7283 0.59891 1.169e-15
## Proportion of Variance 0.4472 0.1832 0.1787 0.1139 0.07704 0.000e+00
## Cumulative Proportion  0.4472 0.6304 0.8090 0.9230 1.00000 1.000e+00
```

```
pca.scores_tog <- data.frame(pca_tog$x)

data.val_tog <- cbind(m.genes25.long, scale_data_tog, pca.scores_tog)
```

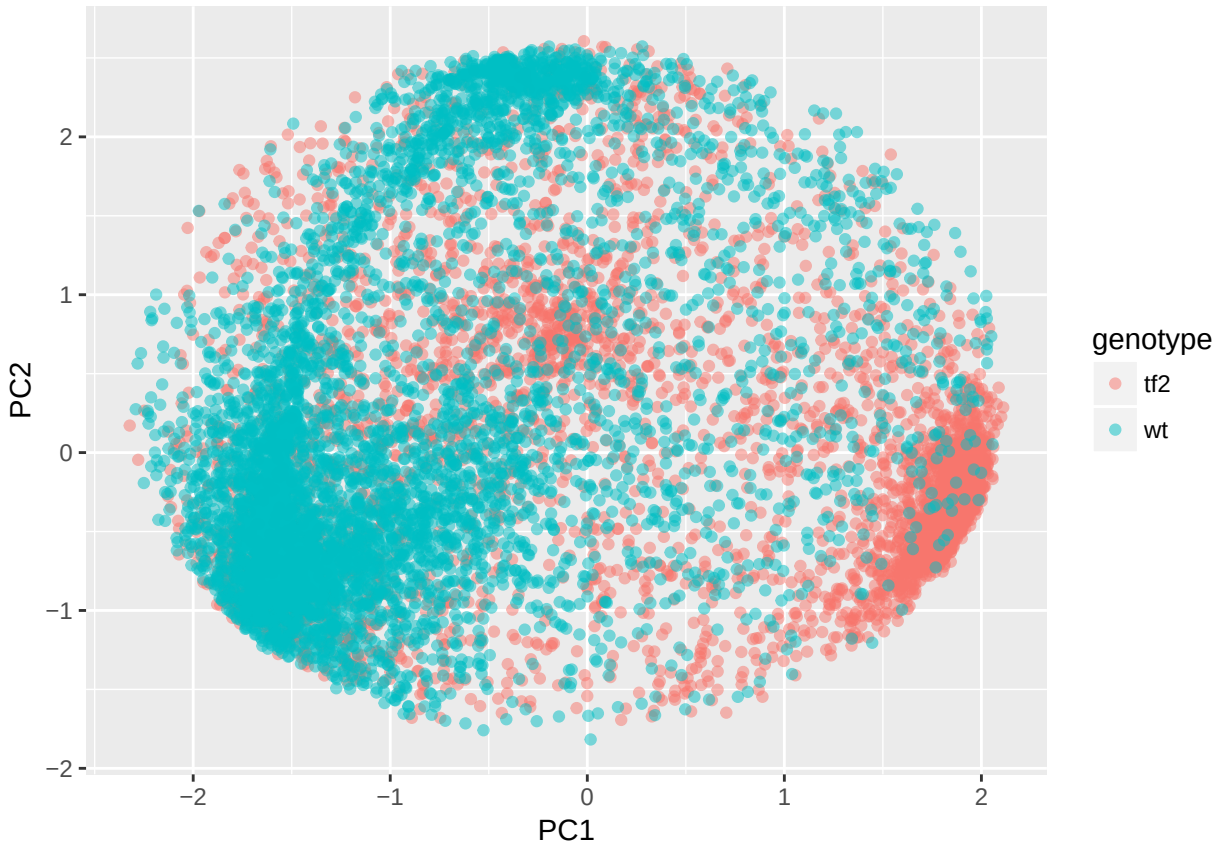
Visualizing the PCA

What is the difference if you scale the genes seperately?

```
p <- ggplot(data.val_tog, aes(PC1, PC2, color = genotype))
p + geom_point(alpha = 0.5)
```




```
p <- ggplot(data.val_sep, aes(PC1, PC2, color = genotype))
p + geom_point(alpha = 0.5)
```



This doesn't seem to affect the analysis very much.

SuperSOM

```
## Using the the version where the values were scaled seperatley.
# head(data.val_sep)
data.val <- data.val_sep

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"
```

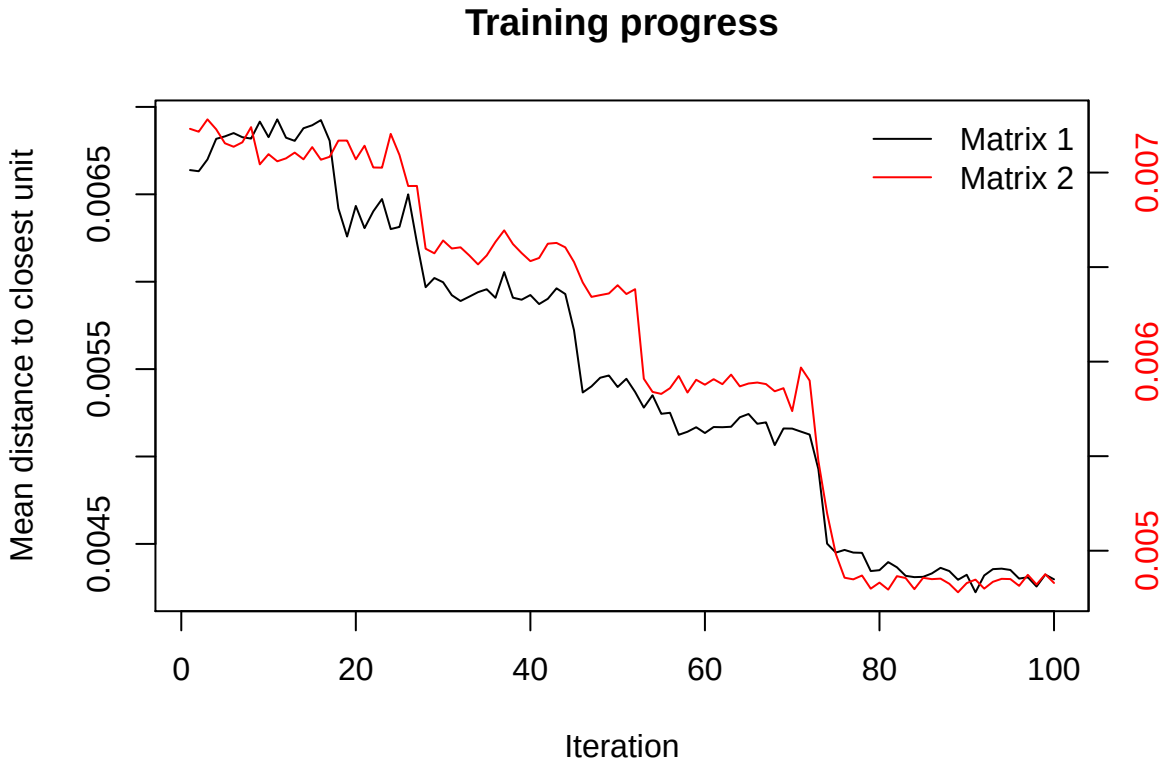
```
# head(data.val)

## Isolate only the scaled values as matrices
tf2 <- as.matrix(subset(data.val, genotype == "tf2", select = 9:14))
wt <- as.matrix(subset(data.val, genotype == "wt", select = 9:14))

# Make sure they are in proper order
all.data <- list(tf2, wt)
# head(all.data)
```

SOM

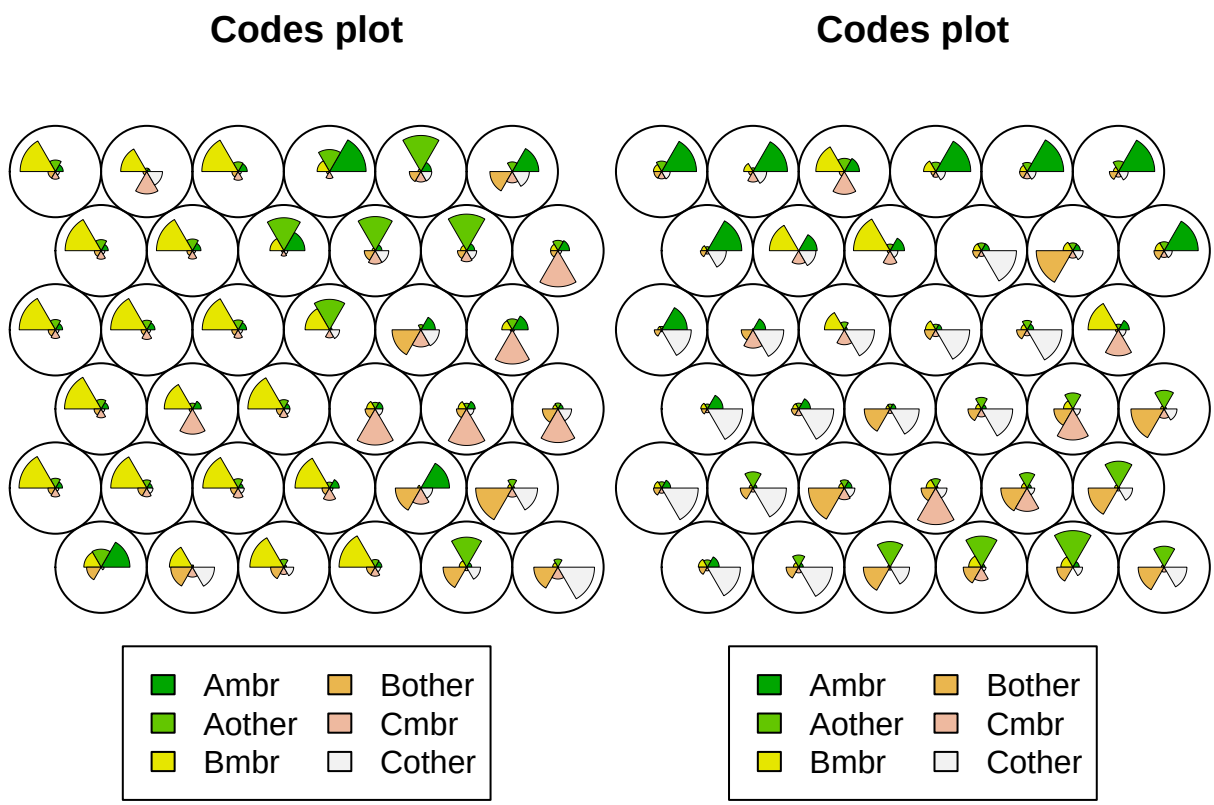
```
## Making the SOM map
ssom <- supersom(all.data, somgrid(6, 6, "hexagonal"))

summary(ssom)
```

```
## SOM of size 6x6 with a hexagonal topology and a bubble neighbourhood function.
## Training data included of 6582 objects
## The number of layers is 2
## Mean distance to the closest unit in the map: 0.8167551
```

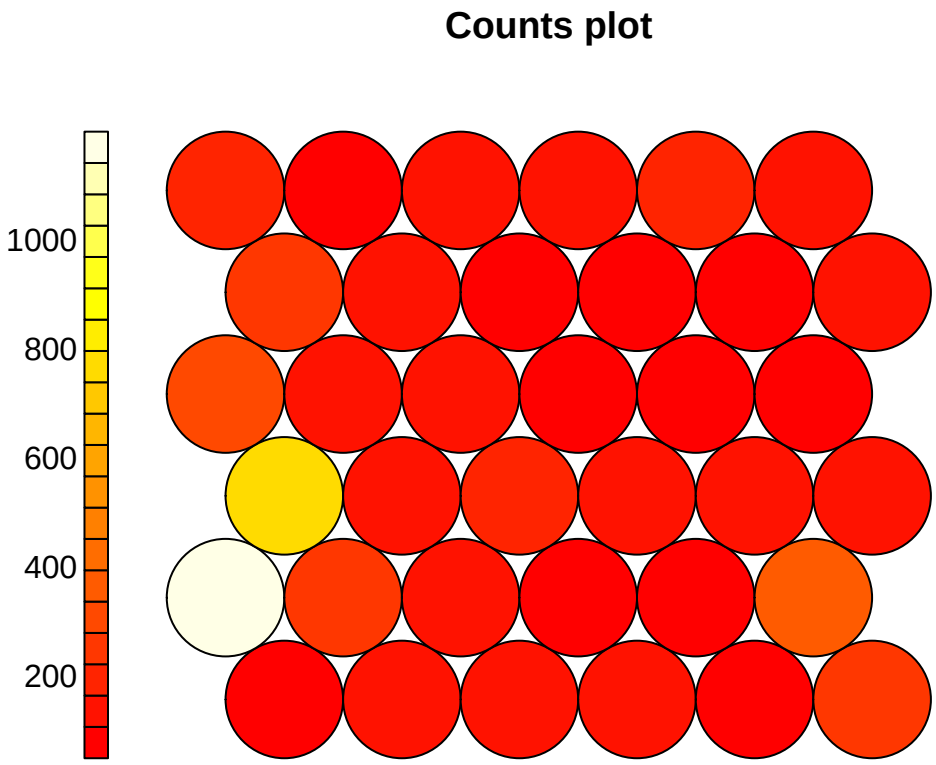
```
#par(mfrow = c(3, 2))
plot(ssom, type = "changes")
```



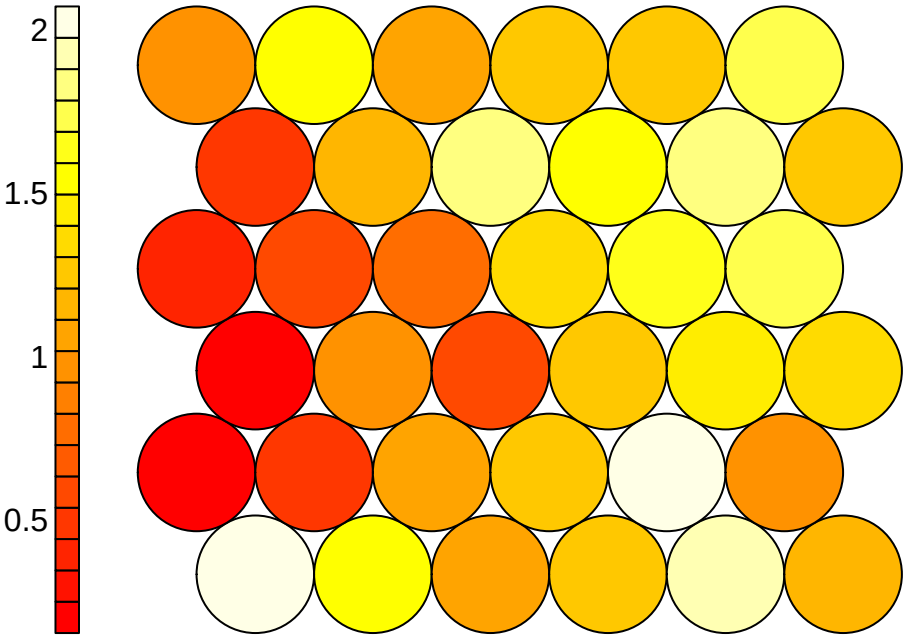

```
plot(ssom, type = "codes")
```




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




```
plot(ssom, type = "quality")
```




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

# head(data.val)

## write.table(data.val, file = "../data/output/ssom.data.analysis5c_05Nov2017_large.txt")
```

Visualization

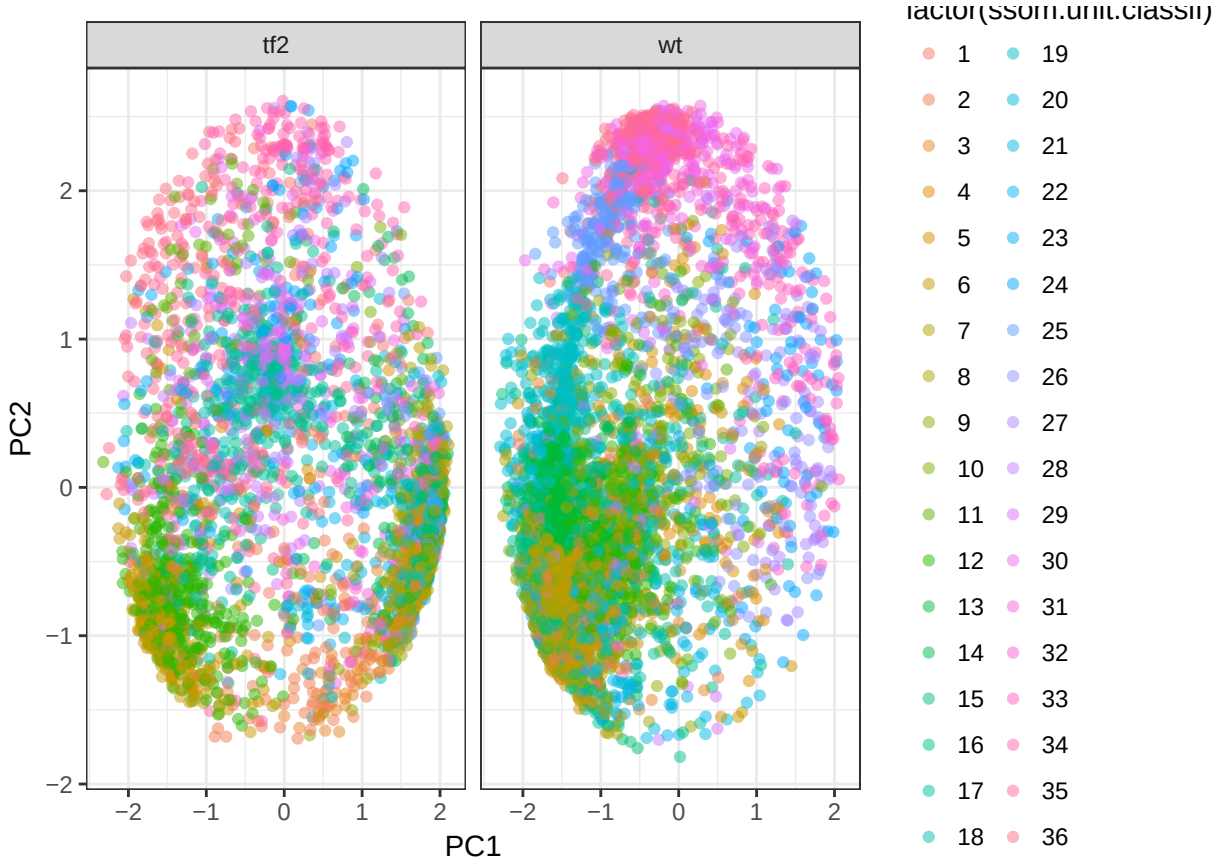
```
## Read in Data from previous section
plot.data <- read.table("../data/output/ssom.data.analysis5c_05Nov2017_large.txt", header = TRUE)
names(plot.data)
```

```
##  [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"      "ssom.unit.classif.1" "ssom.distances.1"
```

```
dim(plot.data)
```

```
## [1] 13164    24
```

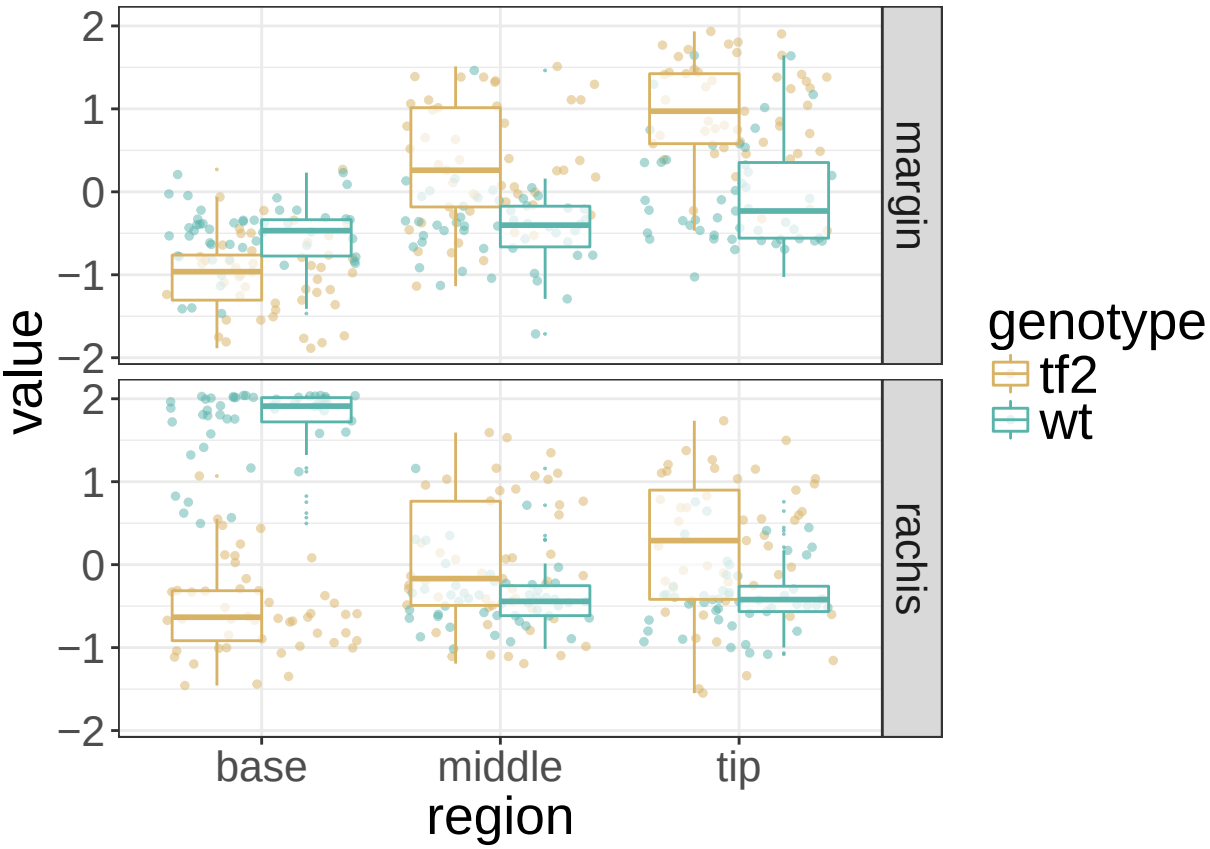
```
## Princliple components colored by clusters
p <- ggplot(plot.data, aes(PC1, PC2, colour = factor(ssom.unit.classif)))
p + geom_point(alpha = .5) +
  theme_bw() +
  facet_grid(.~genotype)
```



Each of the clusters

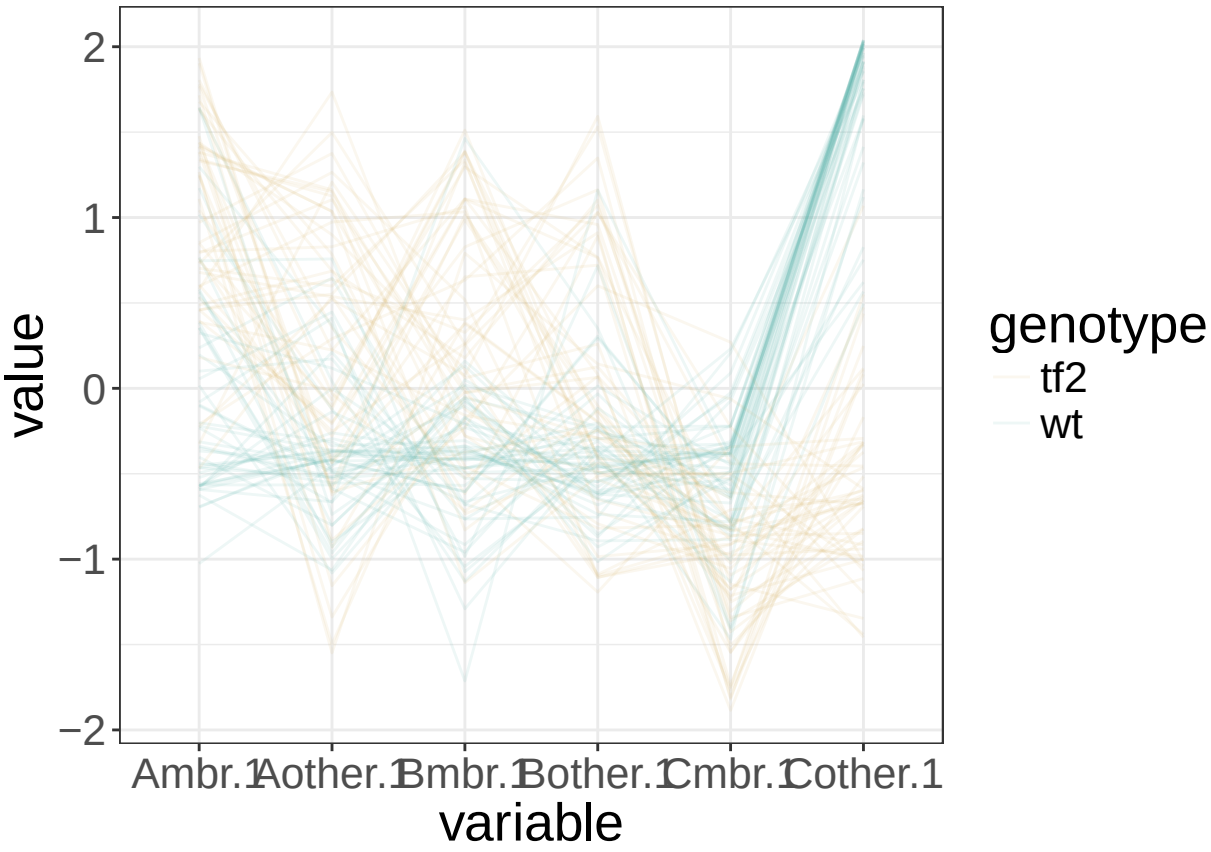
```
data.val2 <- read.table("../data/output/ssom.data.analysis5c_05Nov2017_large.txt", header = TRUE)

clusterVis_region_ssom(1)
```

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



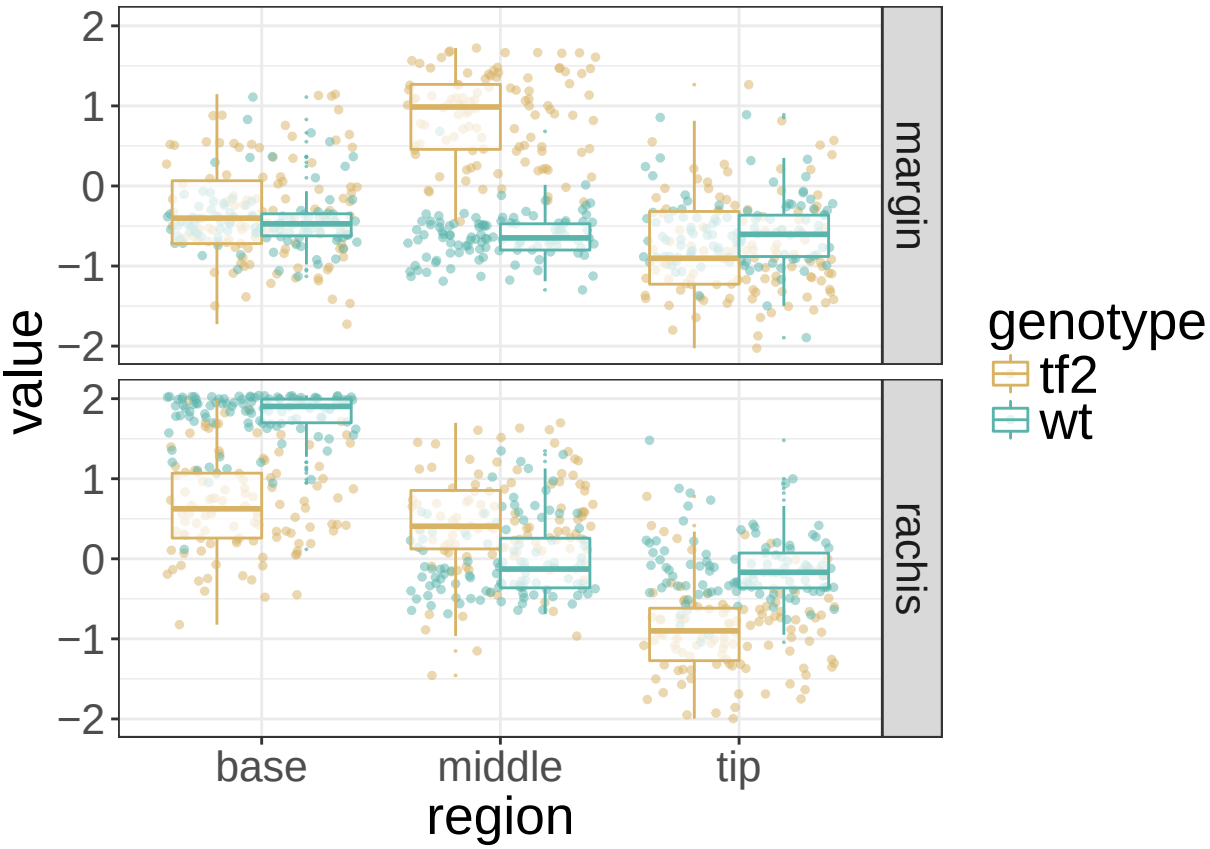

```
clusterVis_line_ssom(1)
```

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



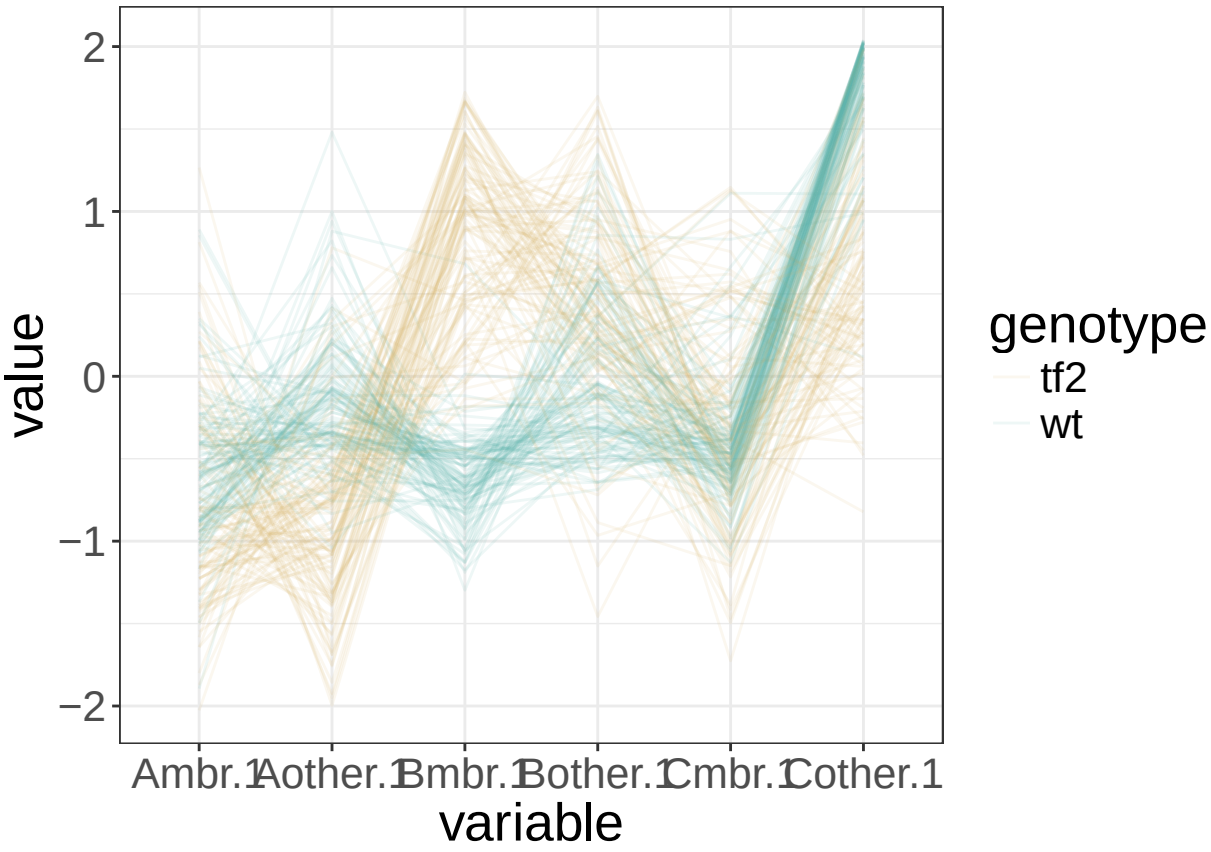

```
clusterVis_region_ssom(2)
```

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



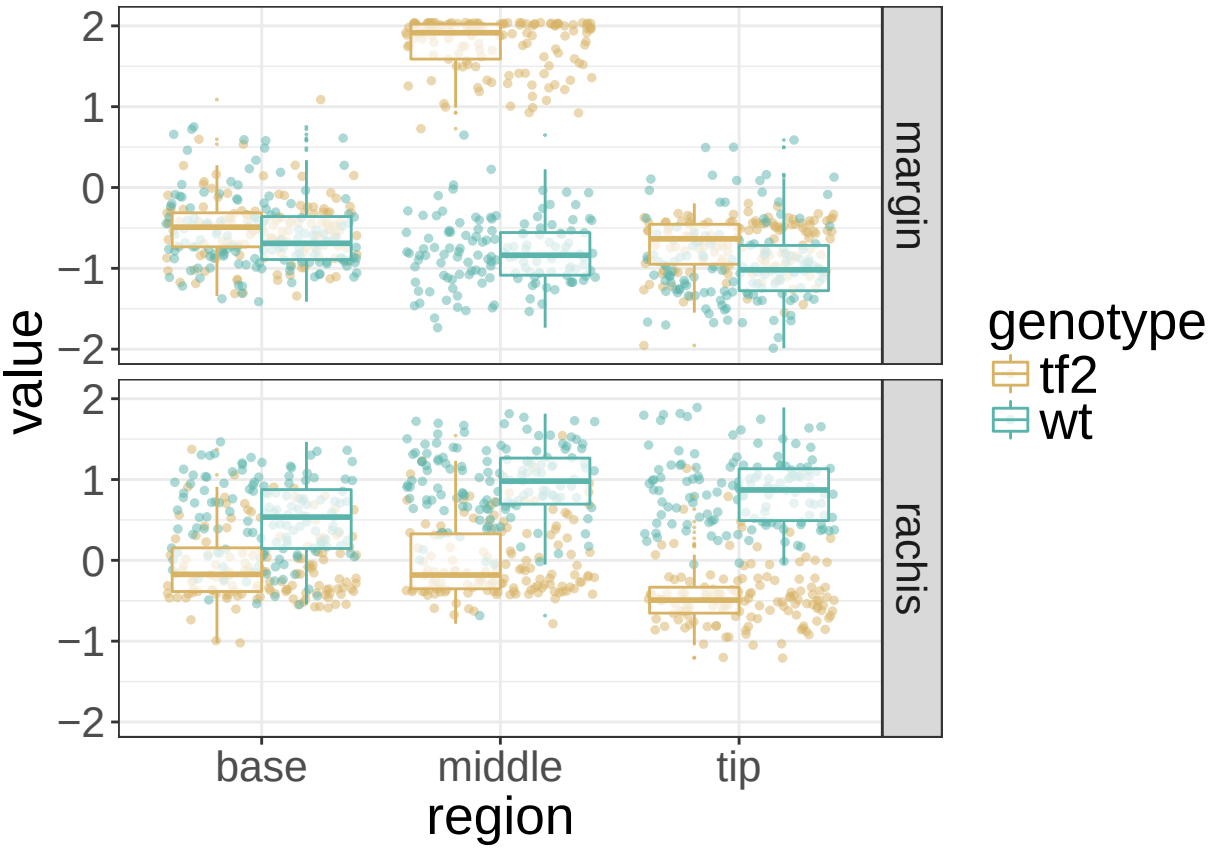

```
clusterVis_line_ssom(2)
```

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



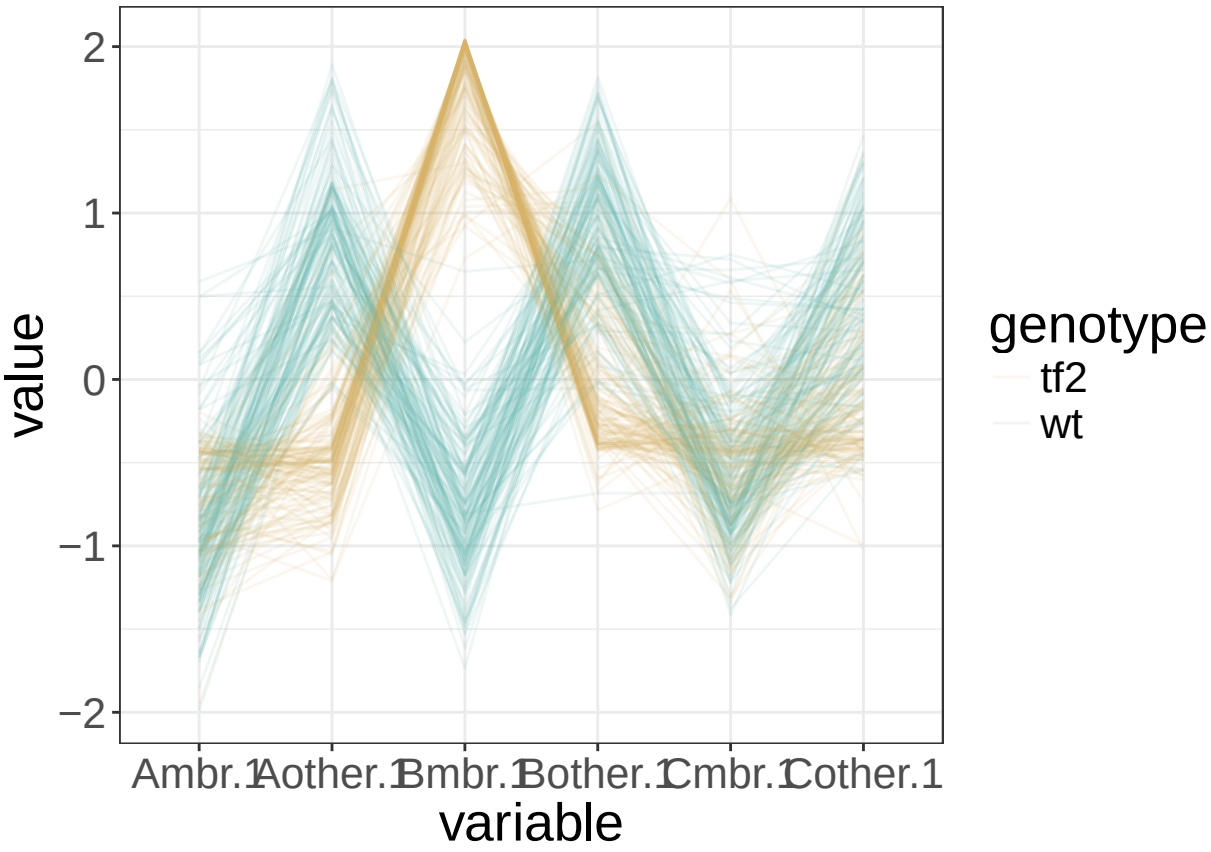

```
clusterVis_region_ssom(3)
```

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



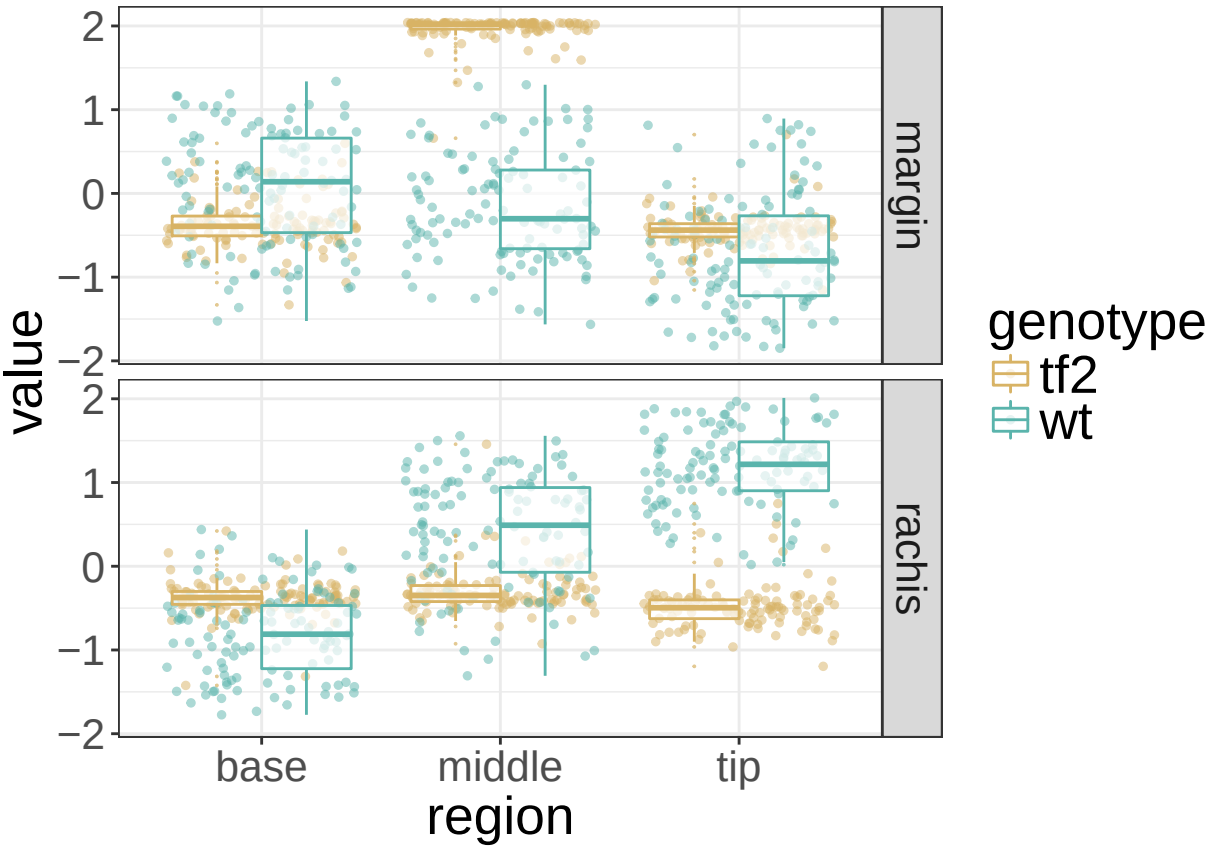

```
clusterVis_line_ssom(3)
```

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



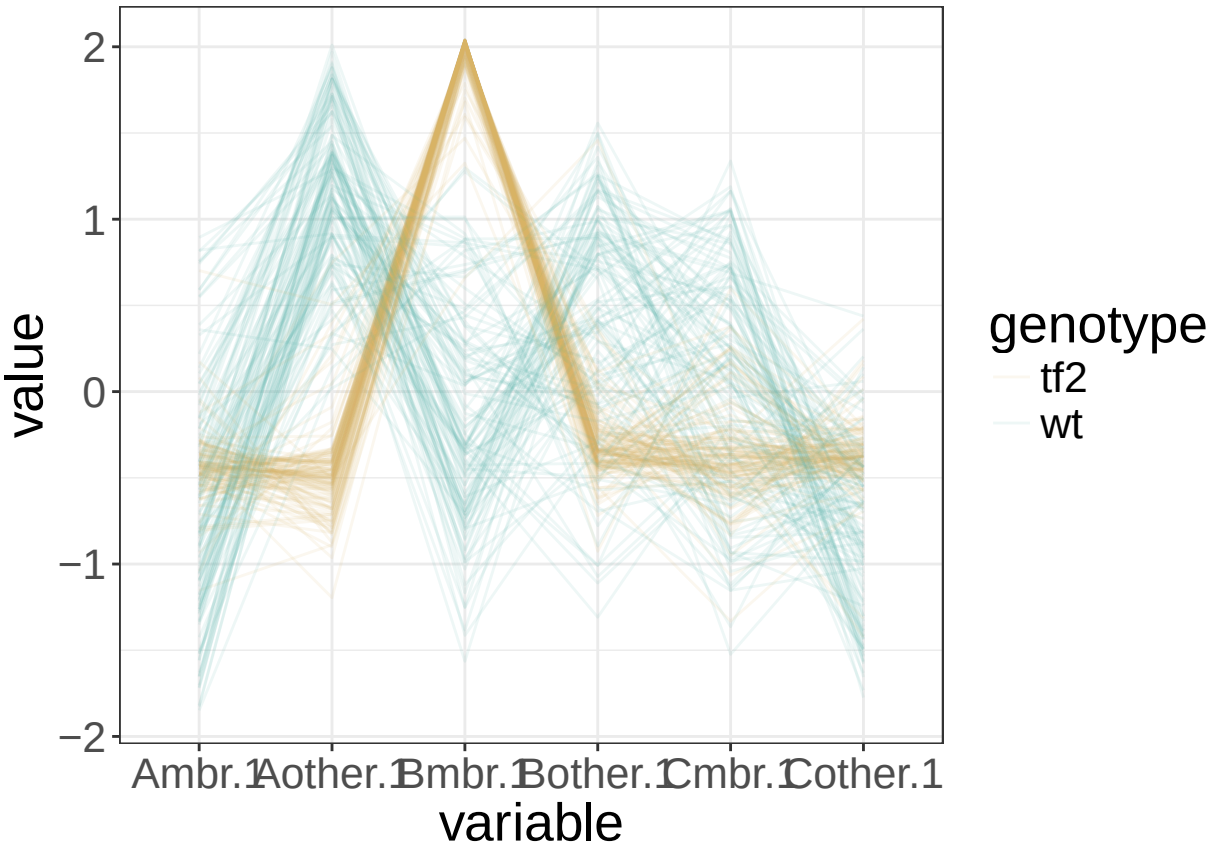

```
clusterVis_region_ssom(4)
```

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



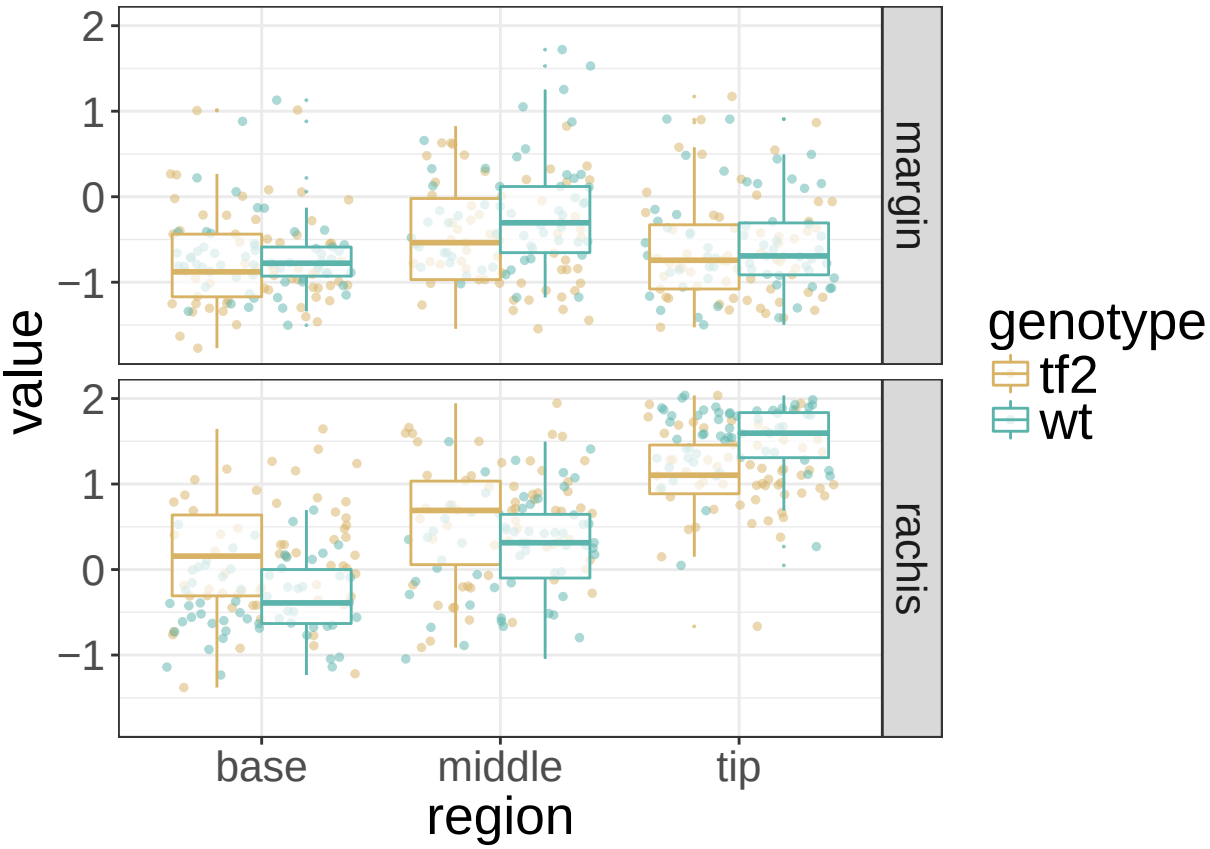

```
clusterVis_line_ssom(4)
```

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



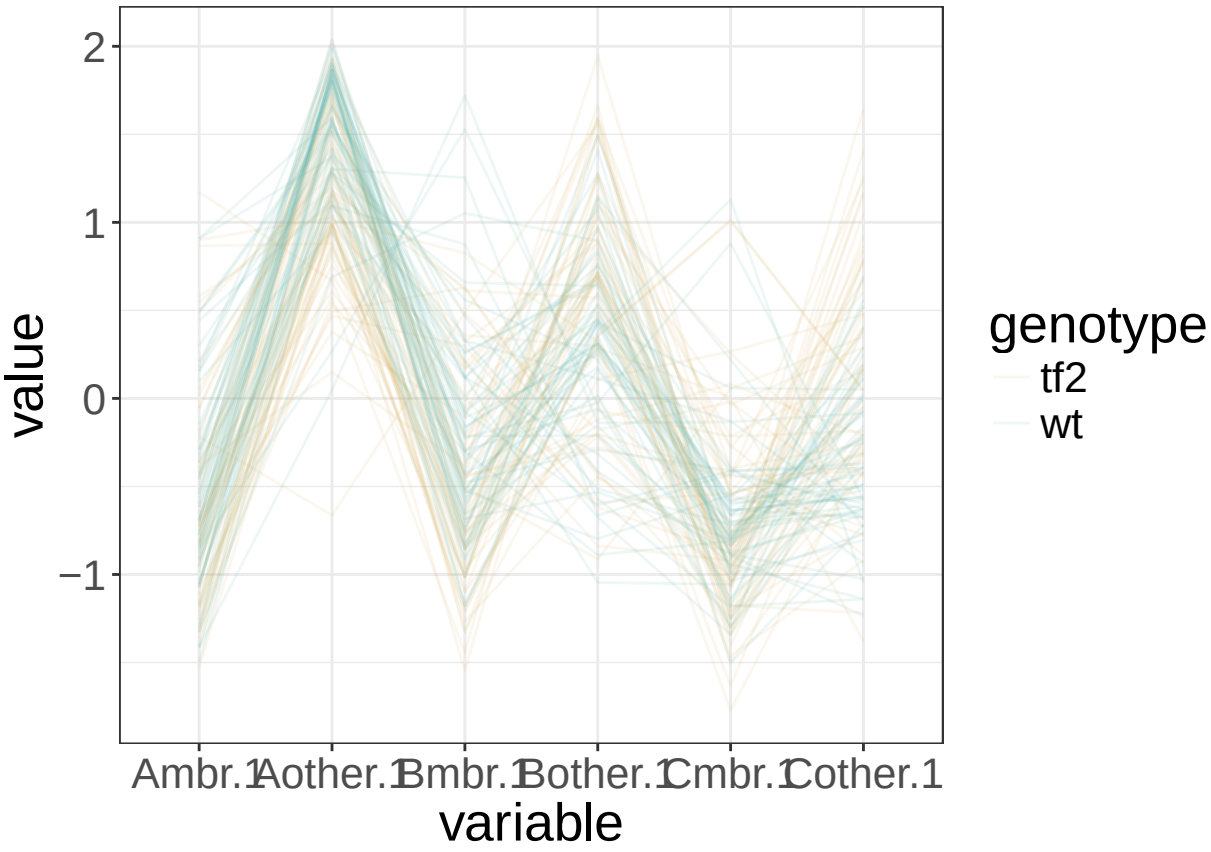

```
clusterVis_region_ssom(5)
```

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



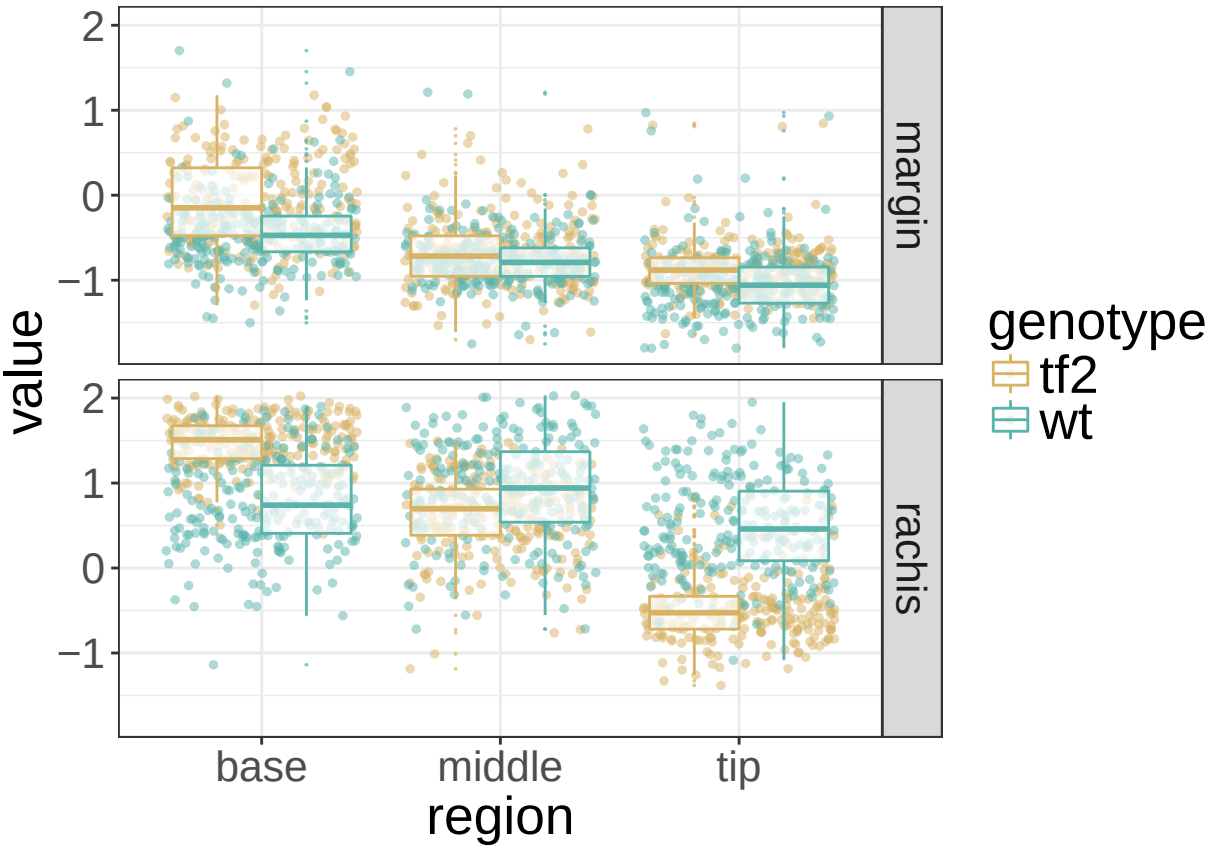

```
clusterVis_line_ssom(5)
```

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



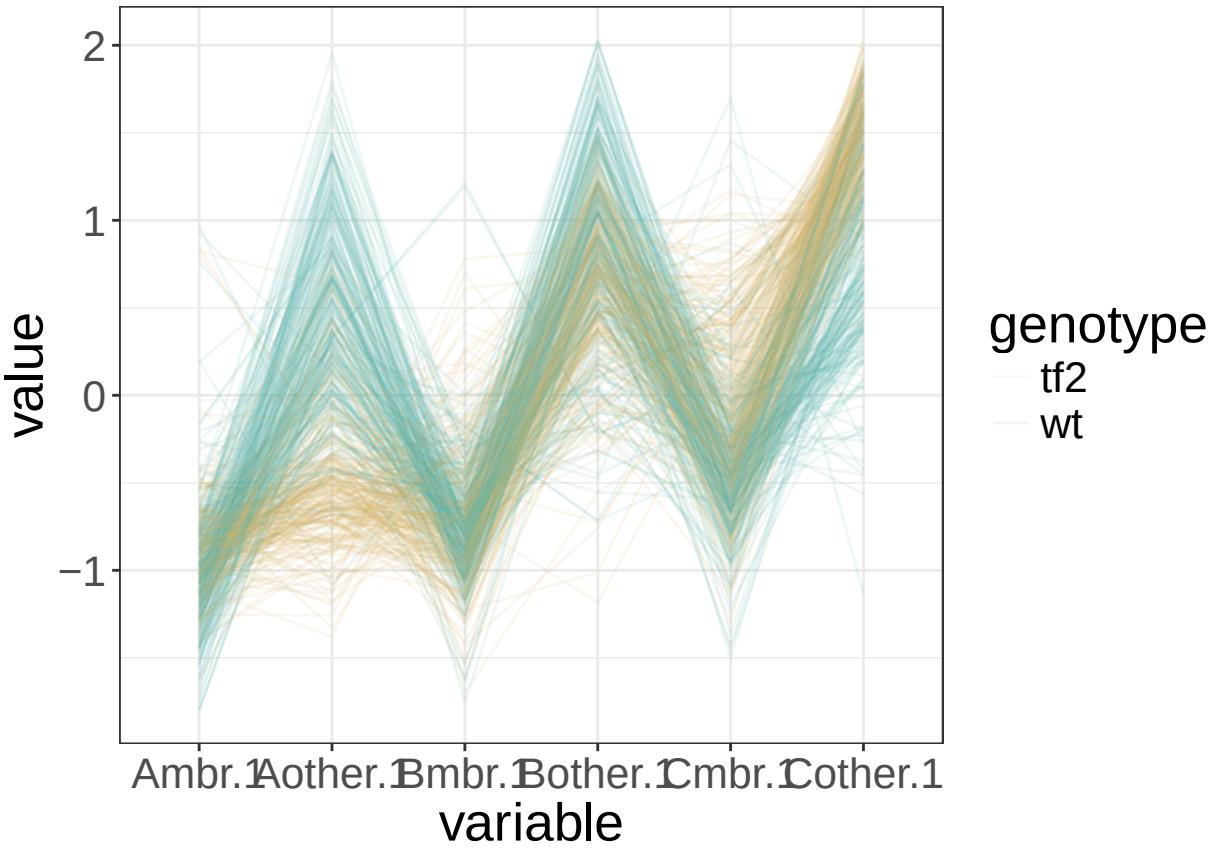

```
clusterVis_region_ssom(6)
```

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



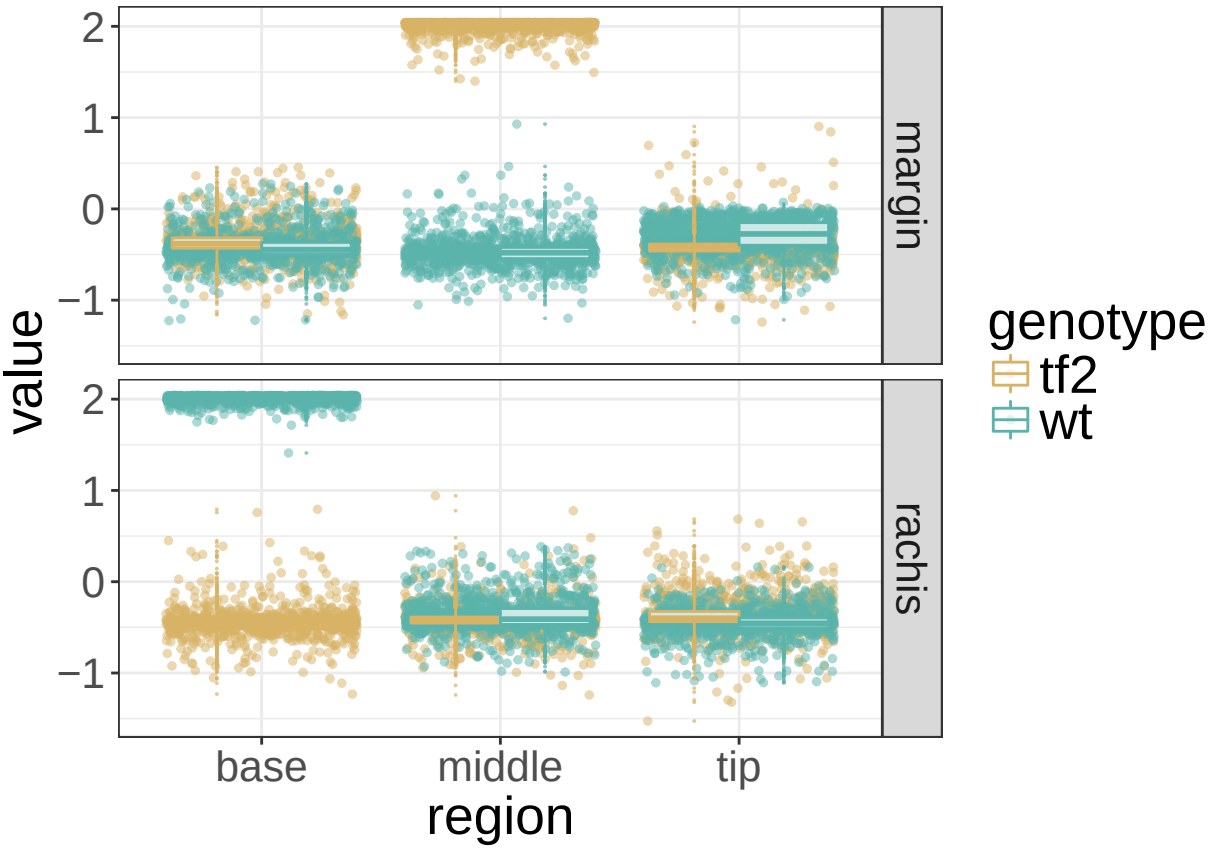

```
clusterVis_line_ssom(6)
```

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



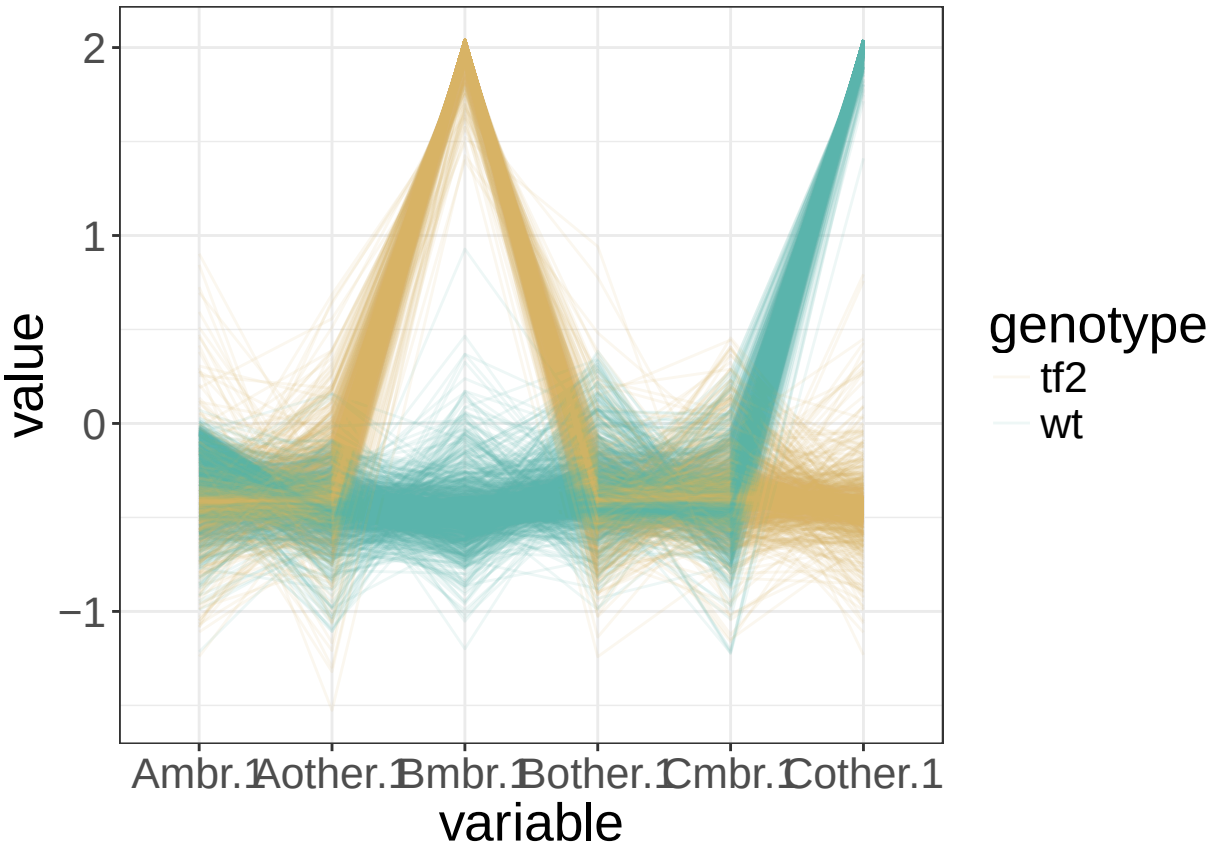

```
clusterVis_region_ssom(7)
```

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



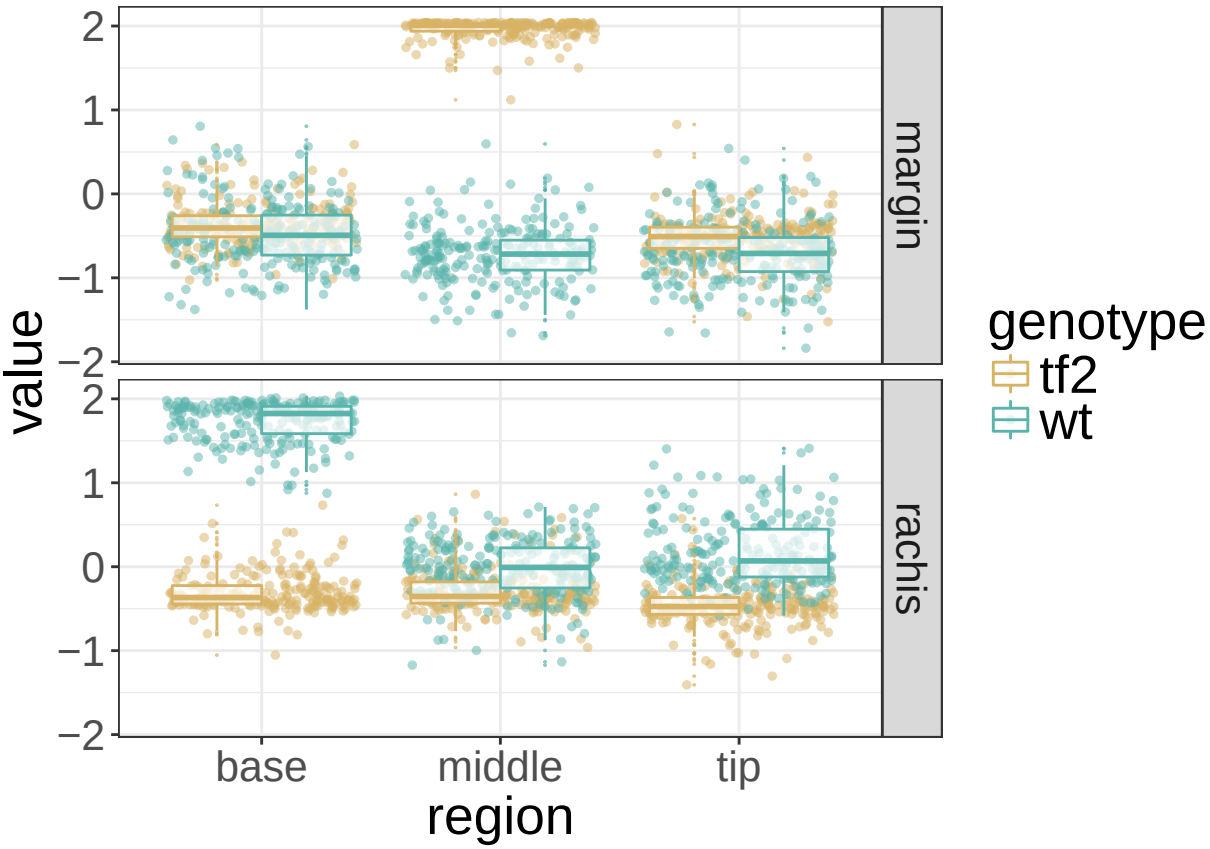

```
clusterVis_line_ssom(7)
```

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



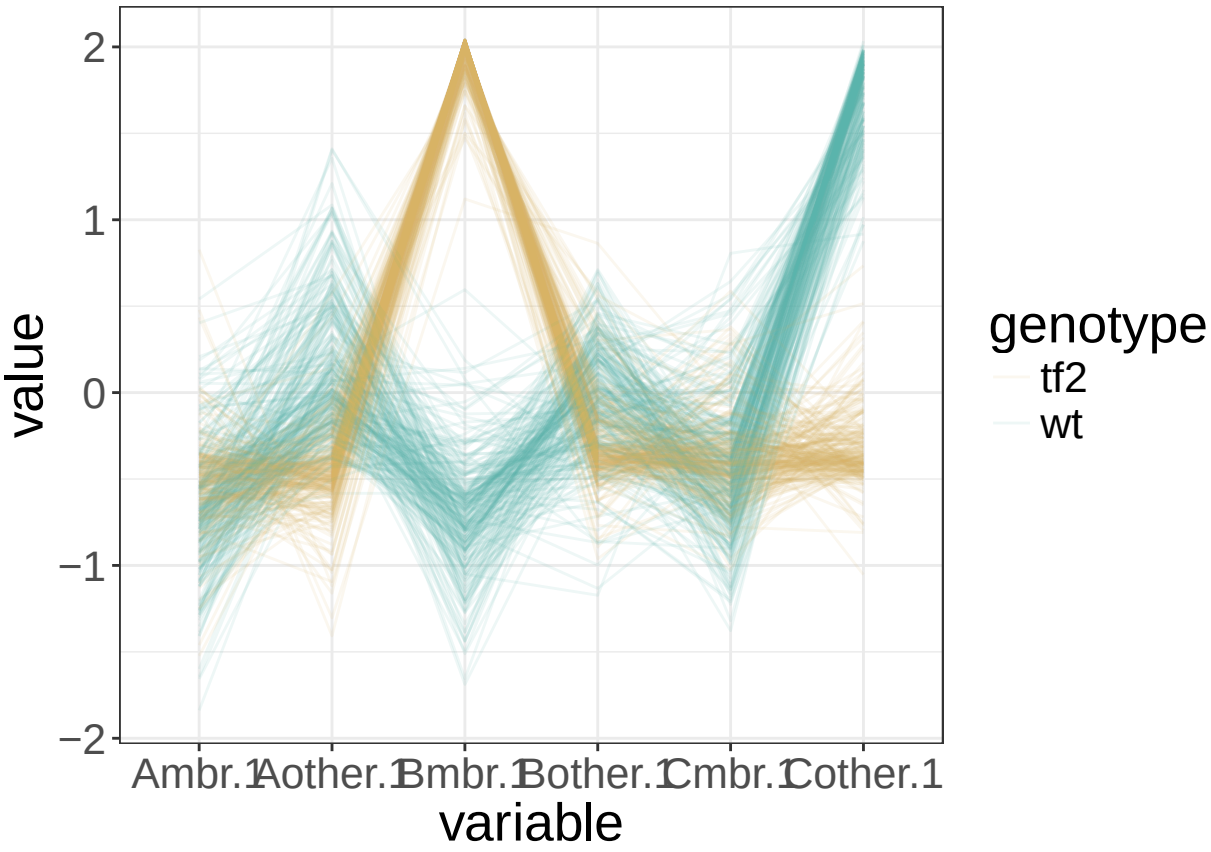

```
clusterVis_region_ssom(8)
```

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



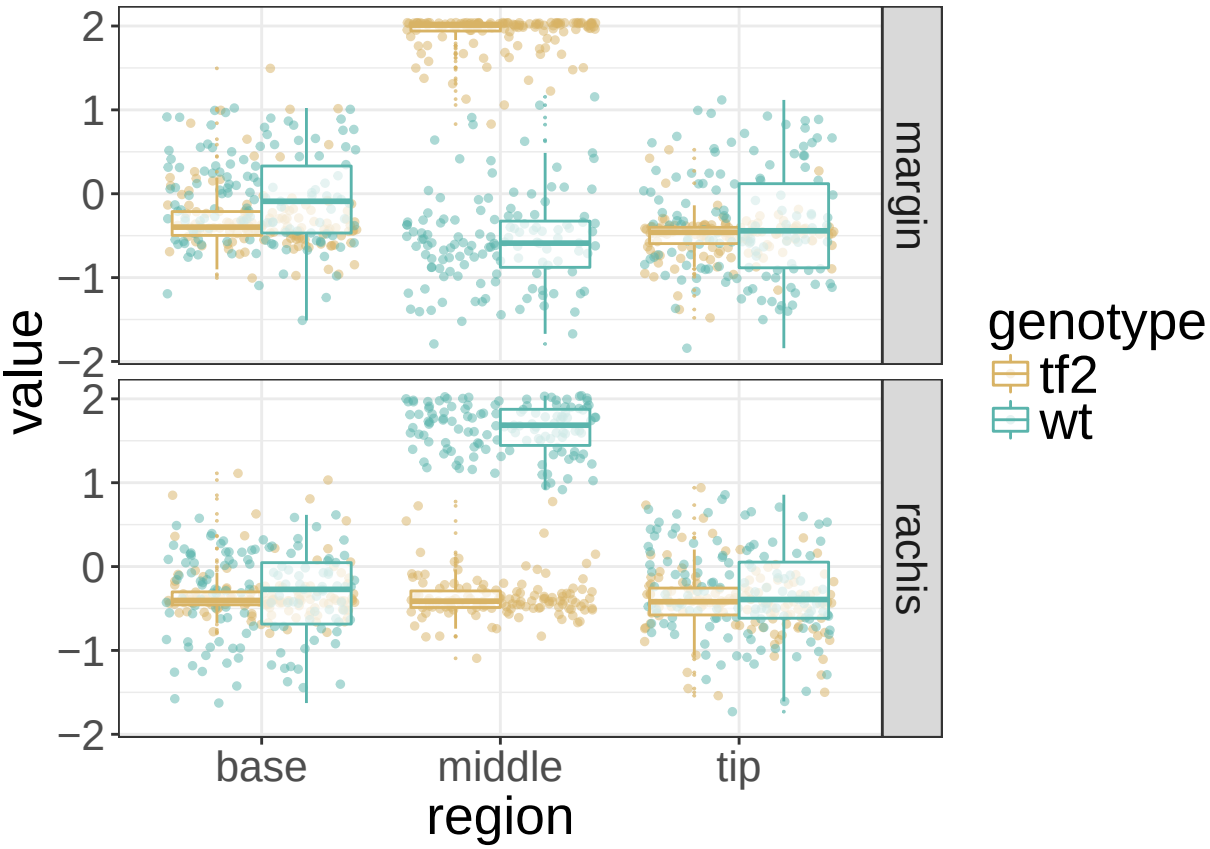

```
clusterVis_line_ssom(8)
```

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



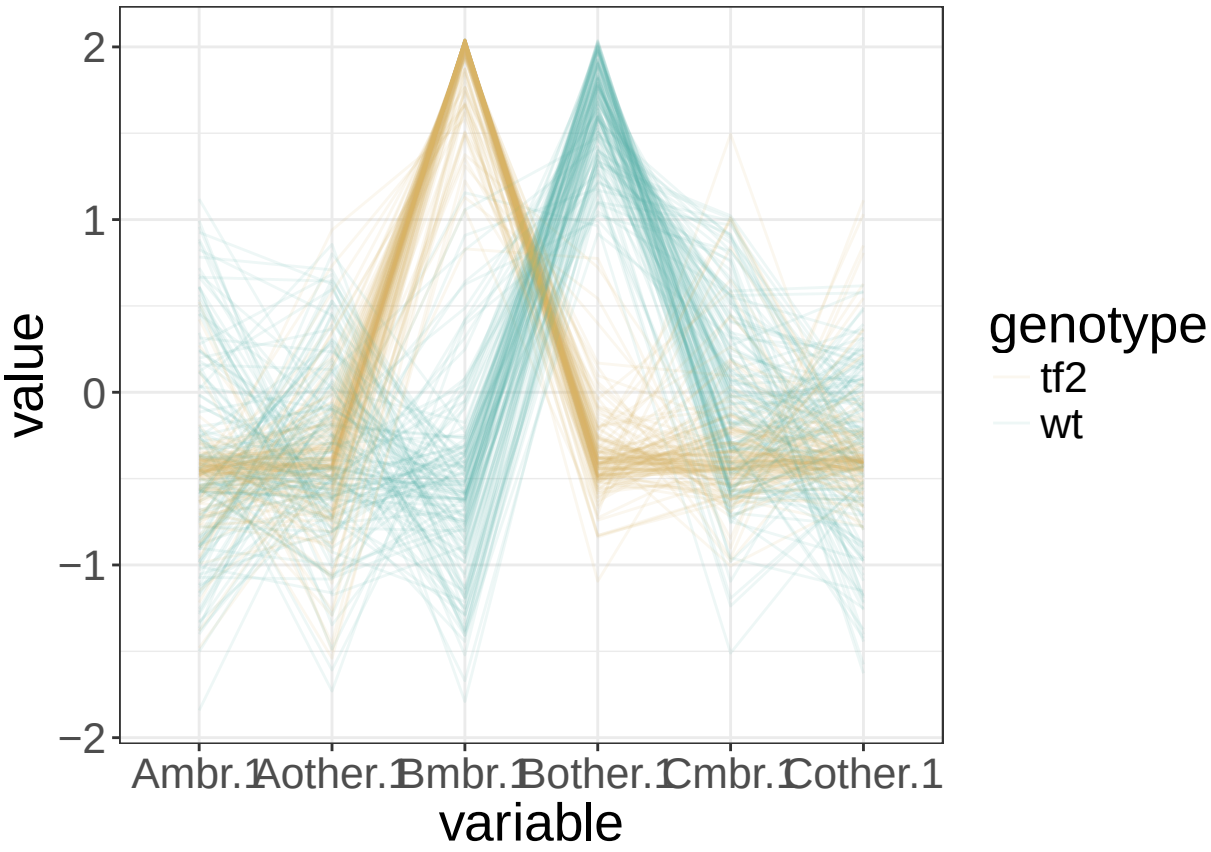

```
clusterVis_region_ssom(9)
```

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



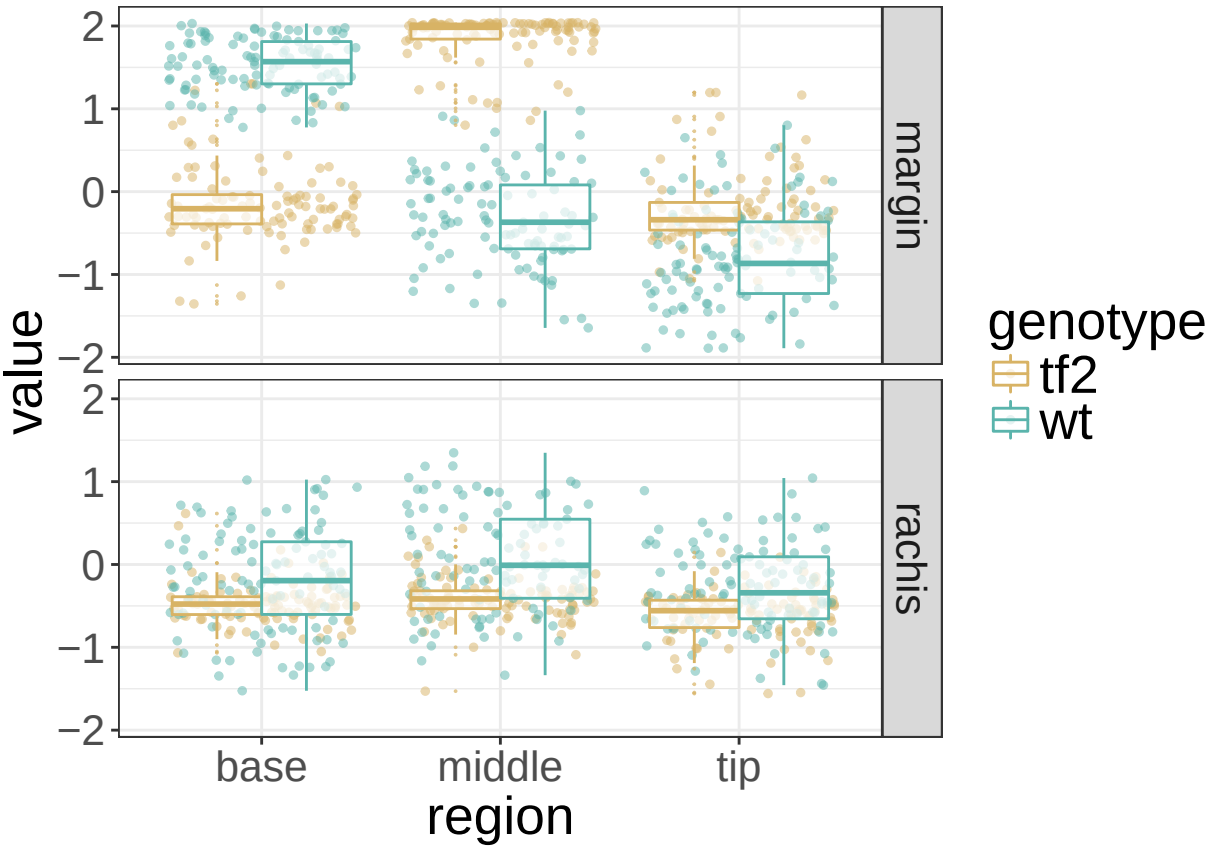

```
clusterVis_line_ssom(9)
```

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



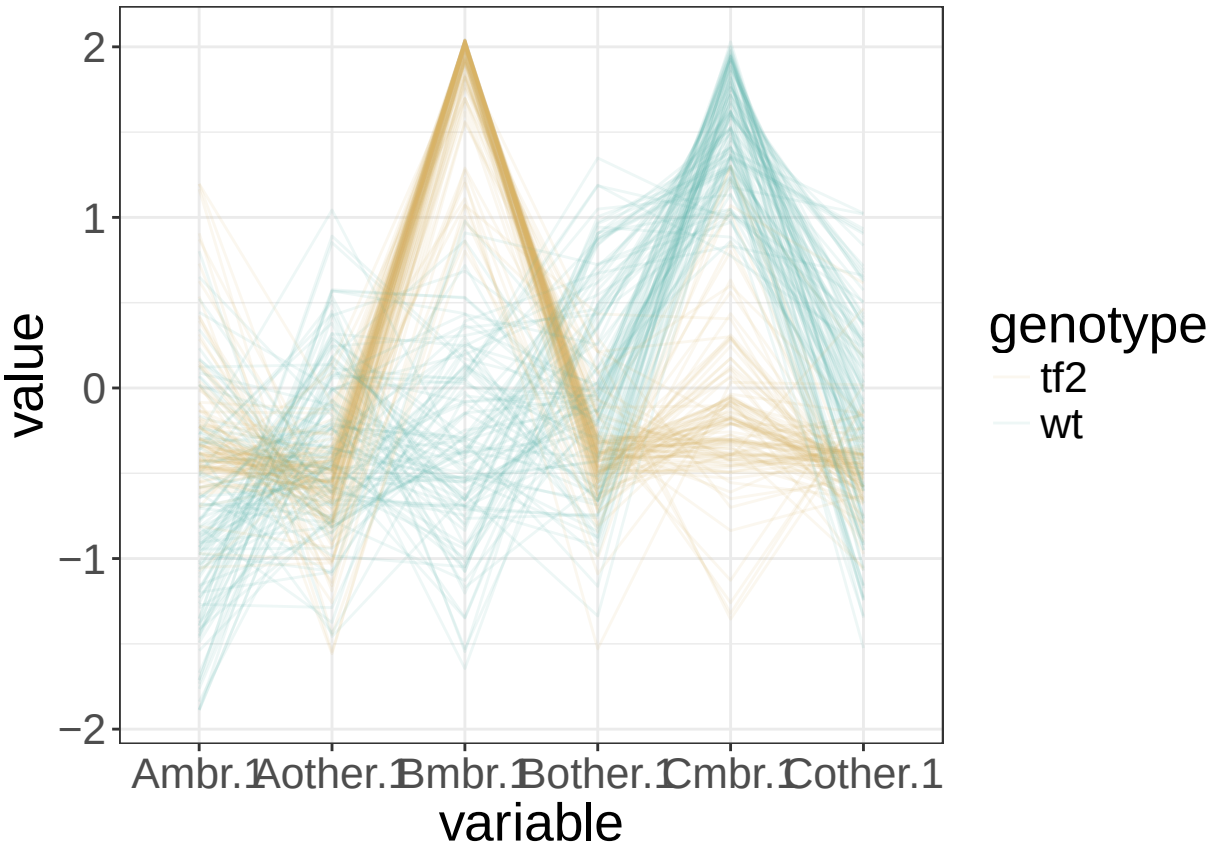

```
clusterVis_region_ssom(10)
```

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



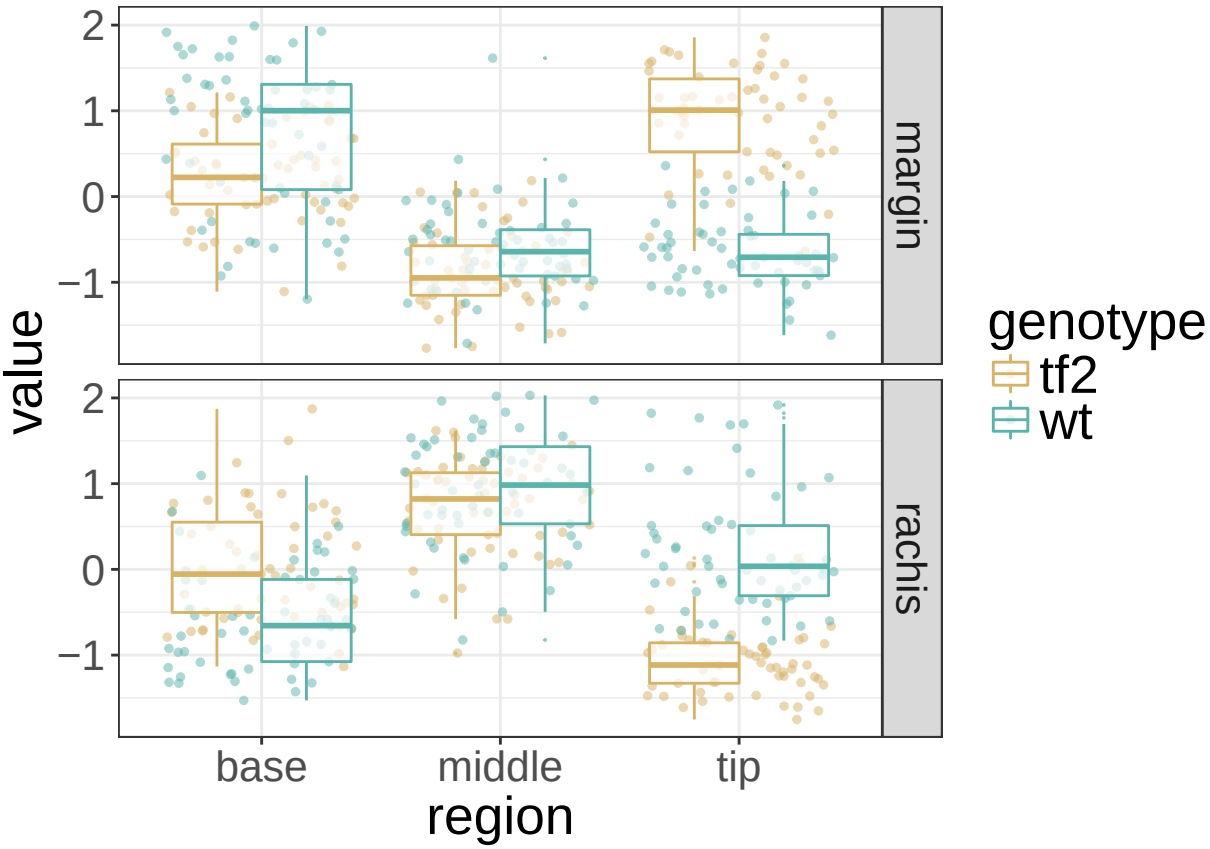

```
clusterVis_line_ssom(10)
```

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



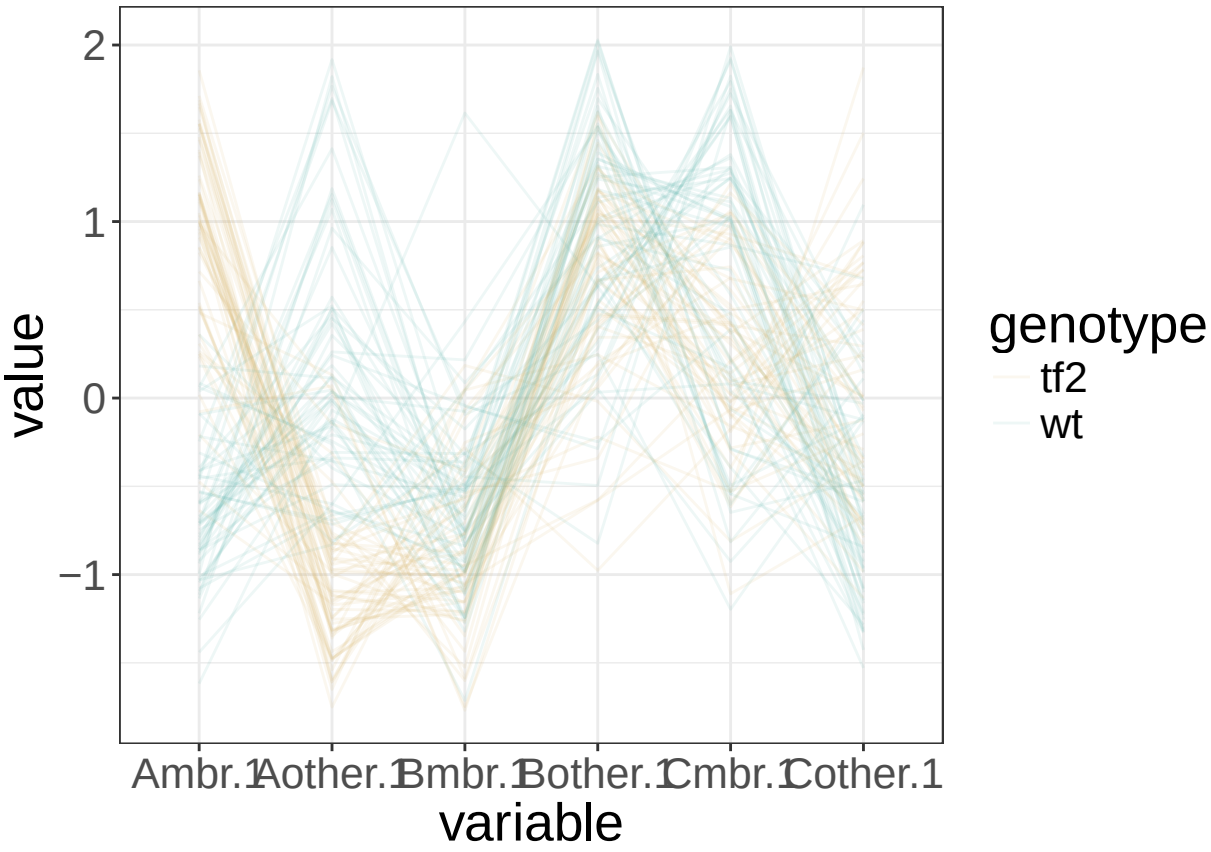

```
clusterVis_region_ssom(11)
```

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



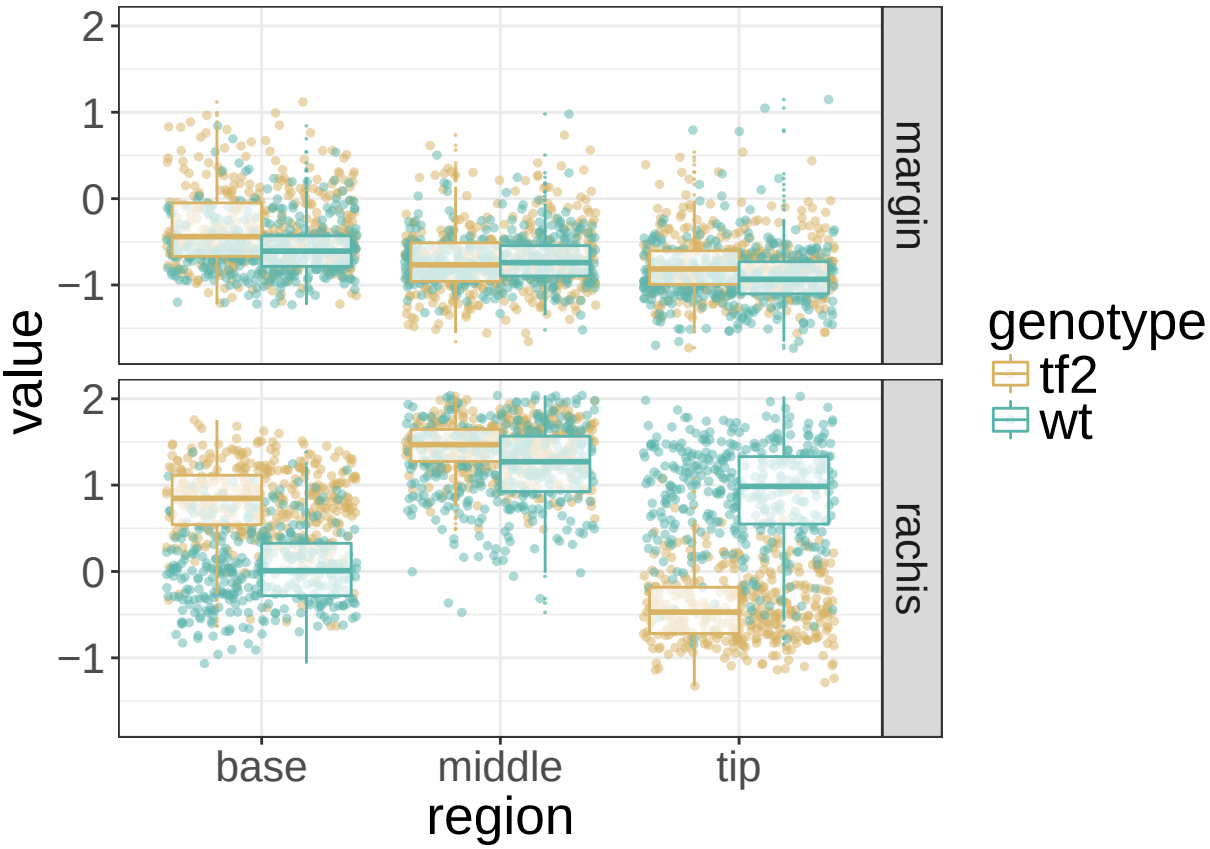

```
clusterVis_line_ssom(11)
```

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



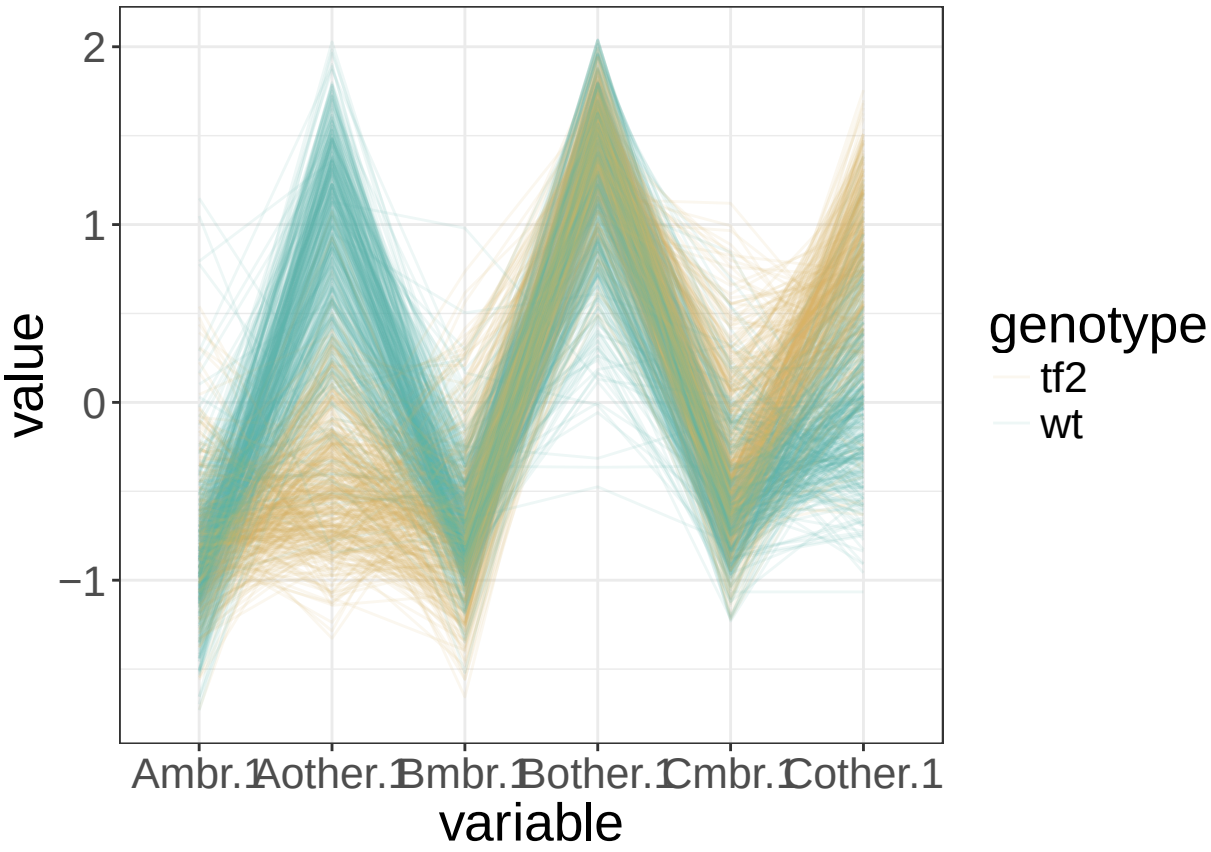

```
clusterVis_region_ssom(12)
```

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



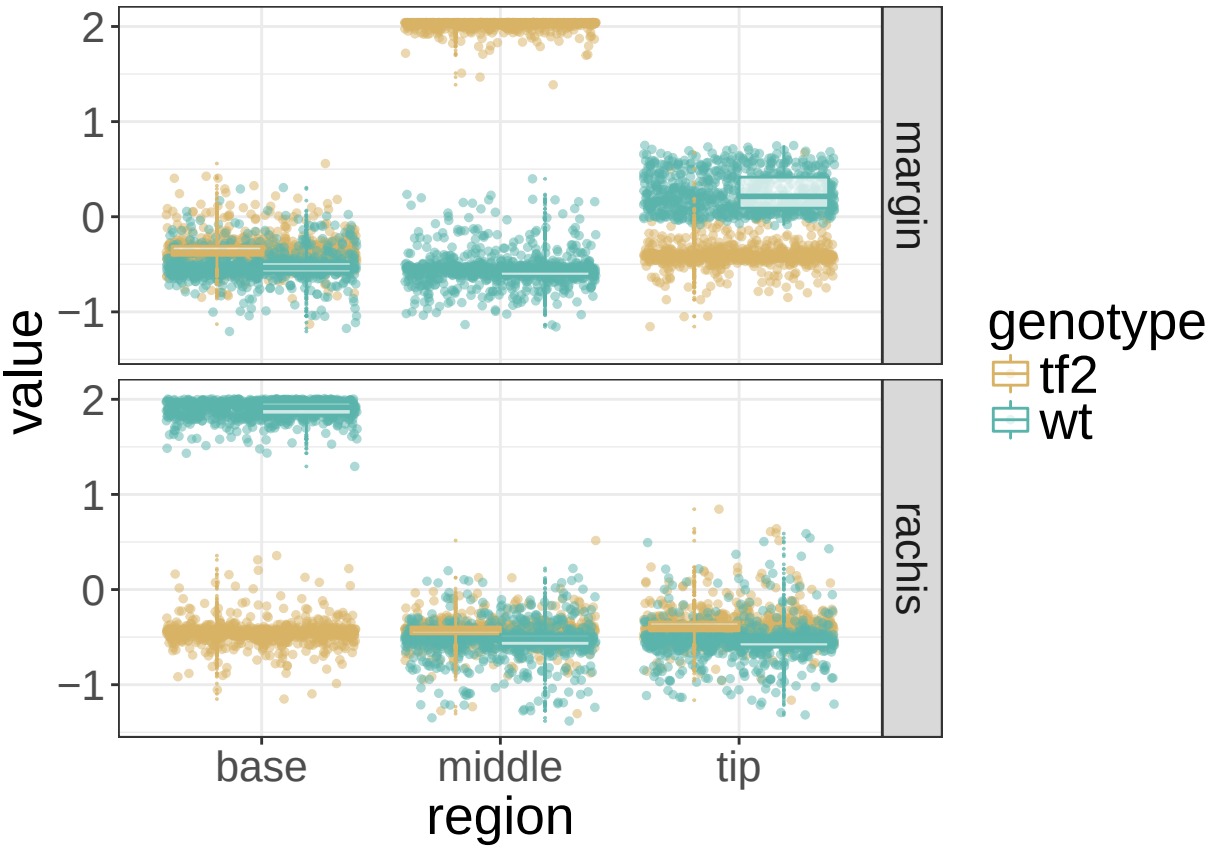

```
clusterVis_line_ssom(12)
```

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



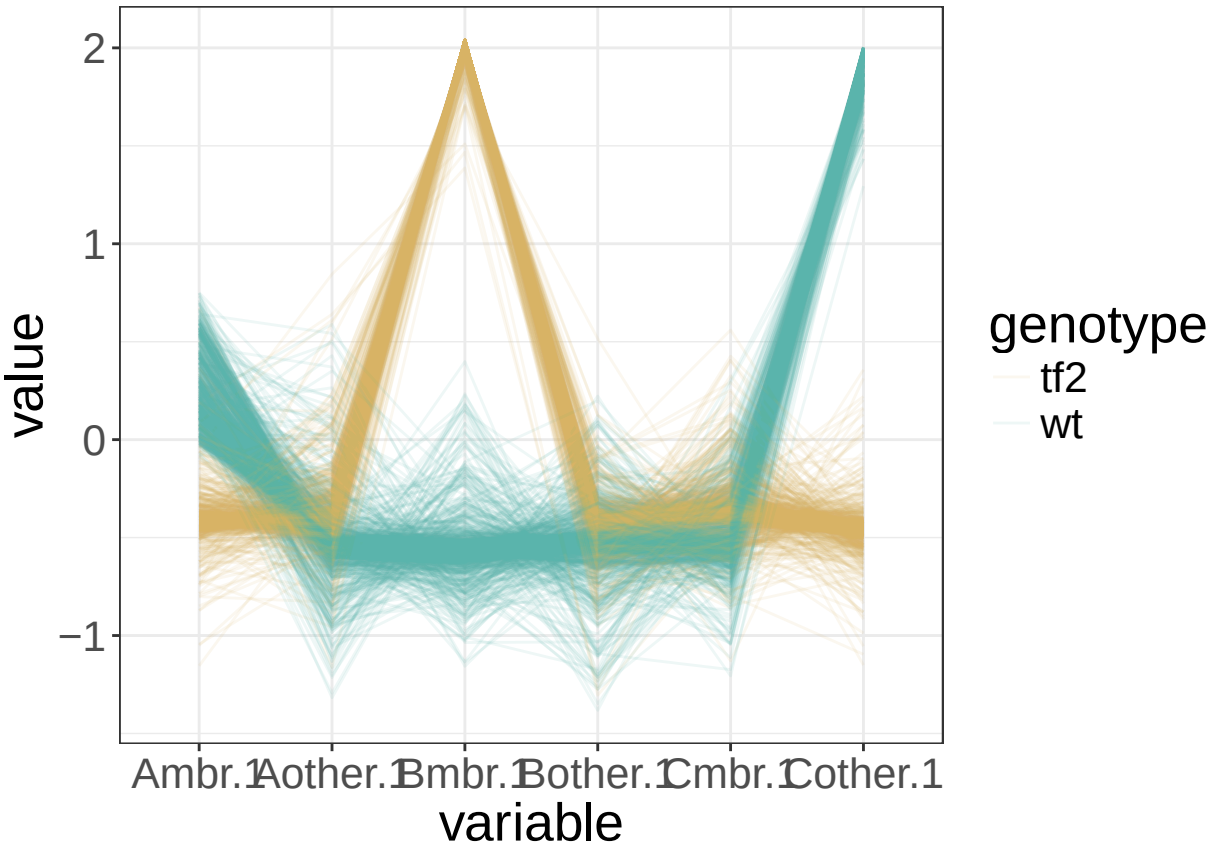

```
clusterVis_region_ssom(13)
```

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



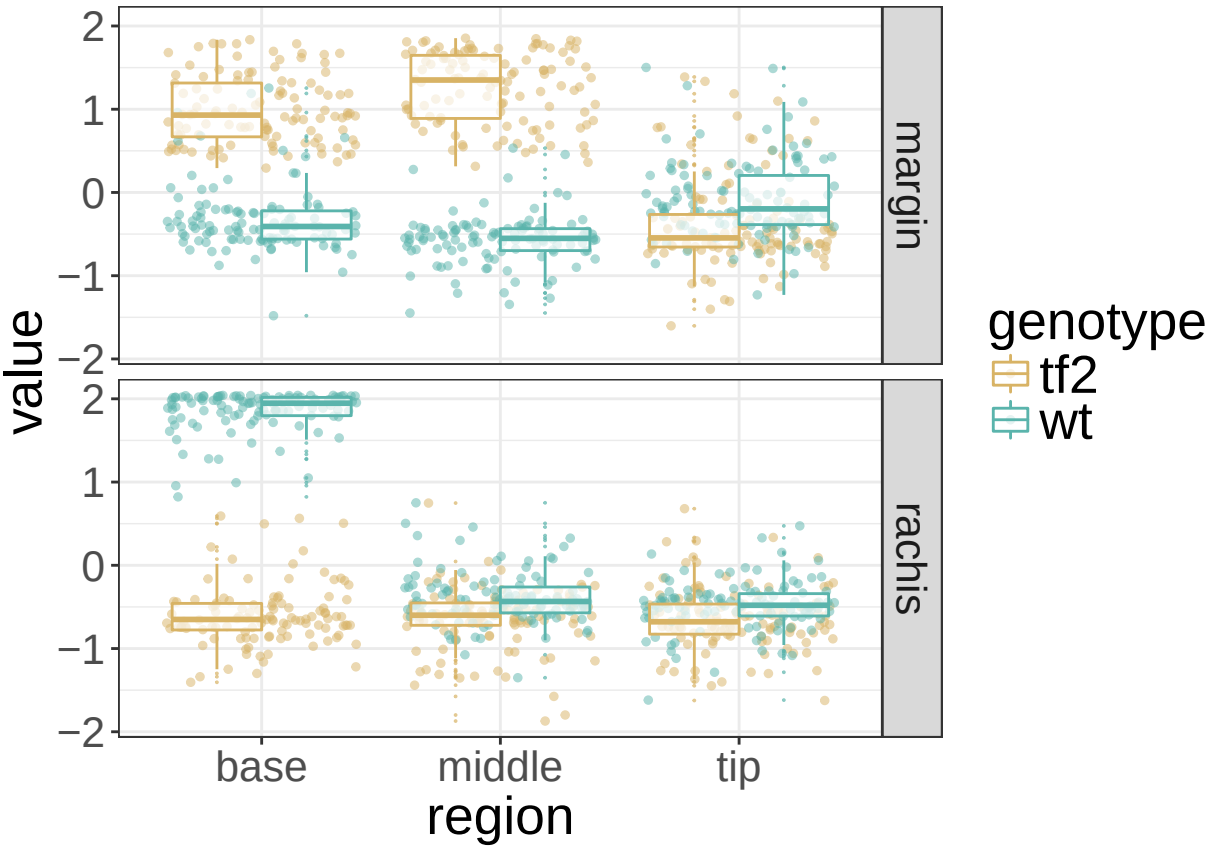

```
clusterVis_line_ssom(13)
```

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



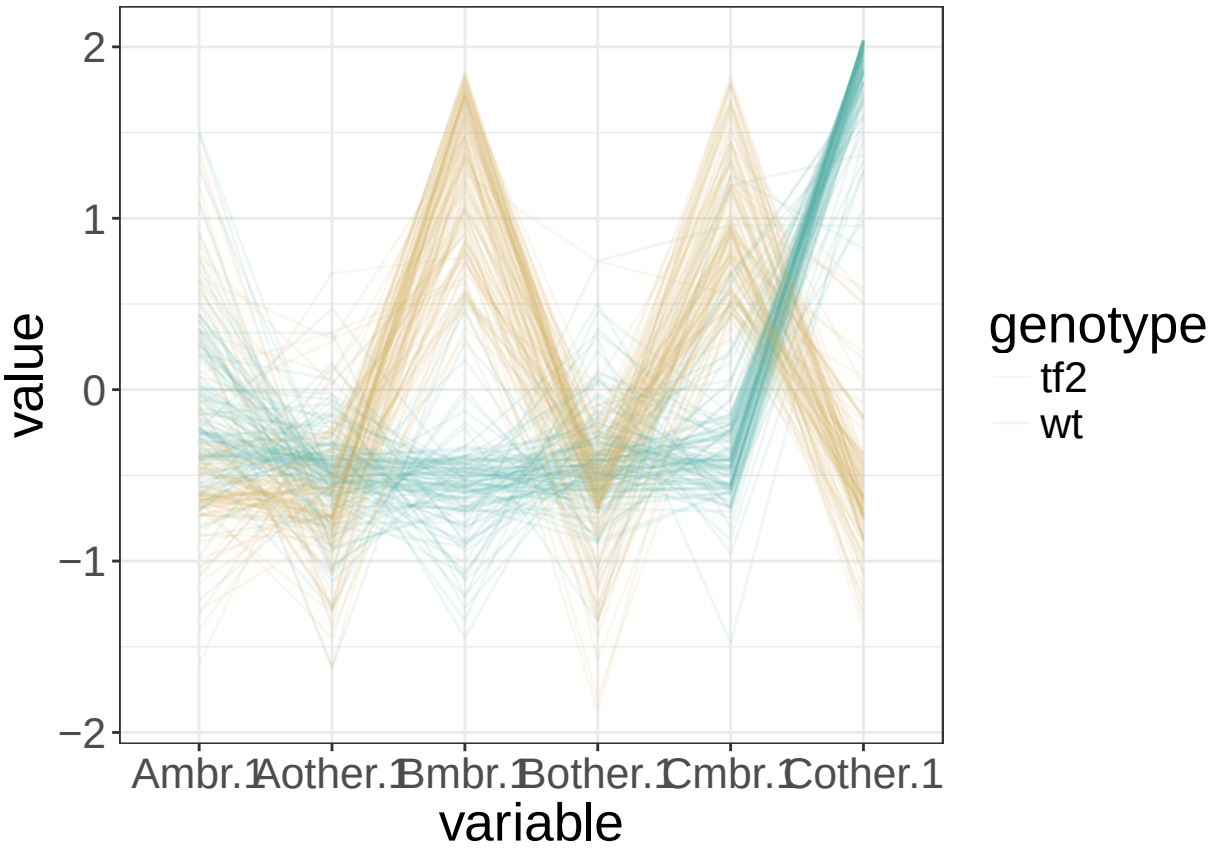

```
clusterVis_region_ssom(14)
```

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



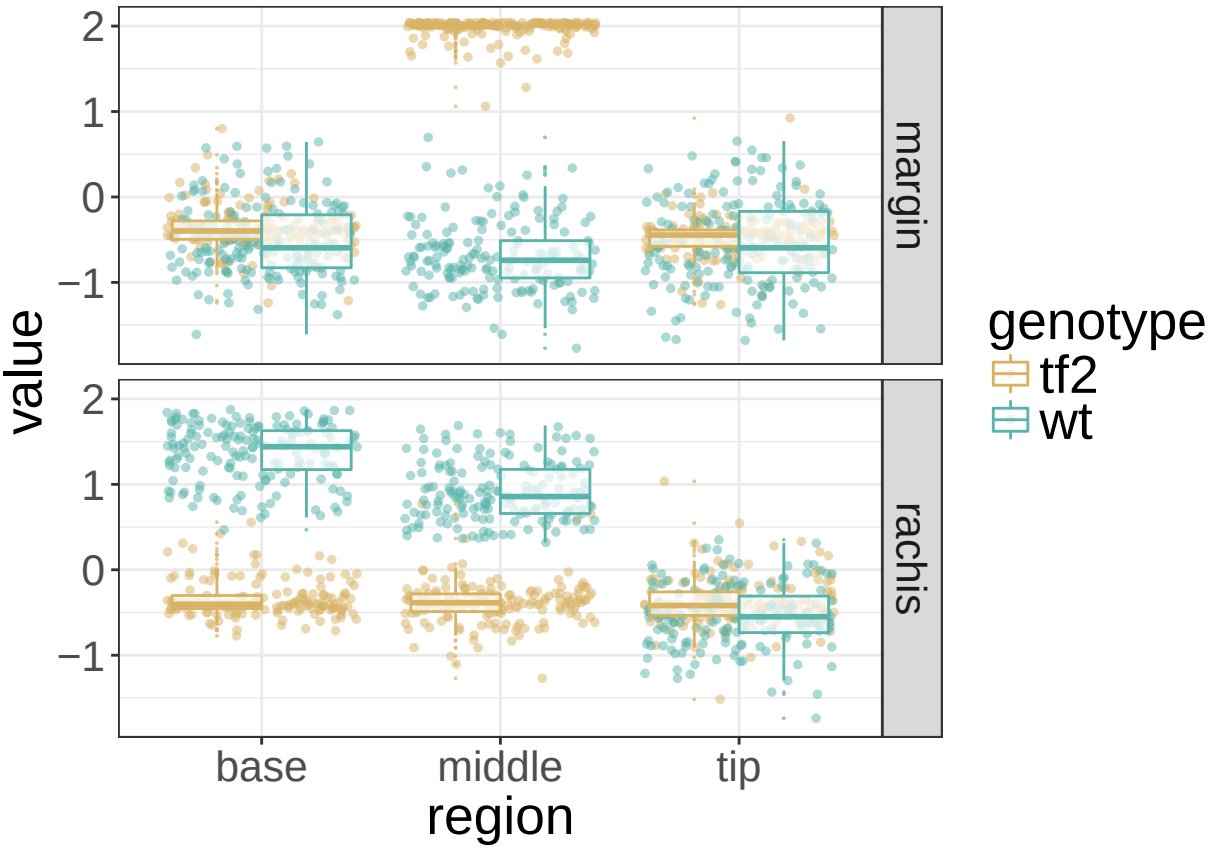

```
clusterVis_line_ssom(14)
```

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



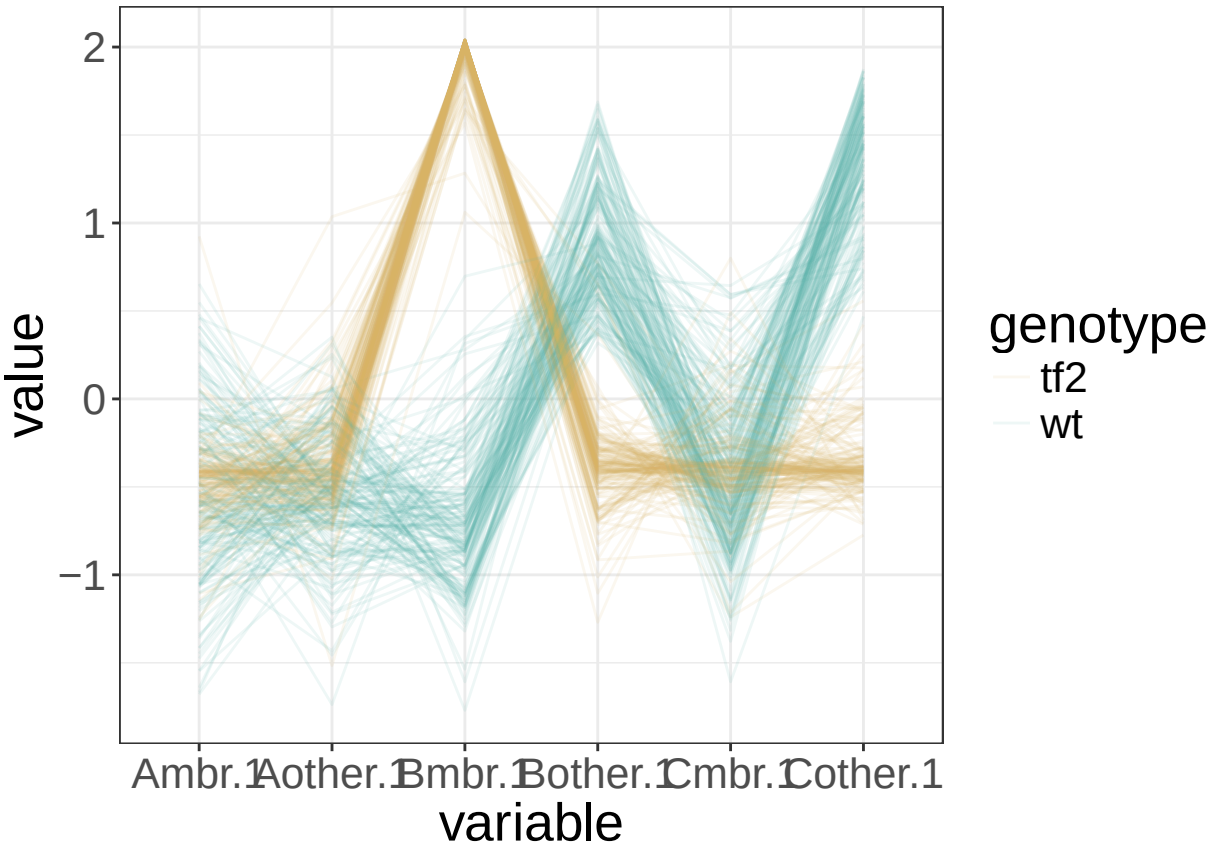

```
clusterVis_region_ssom(15)
```

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



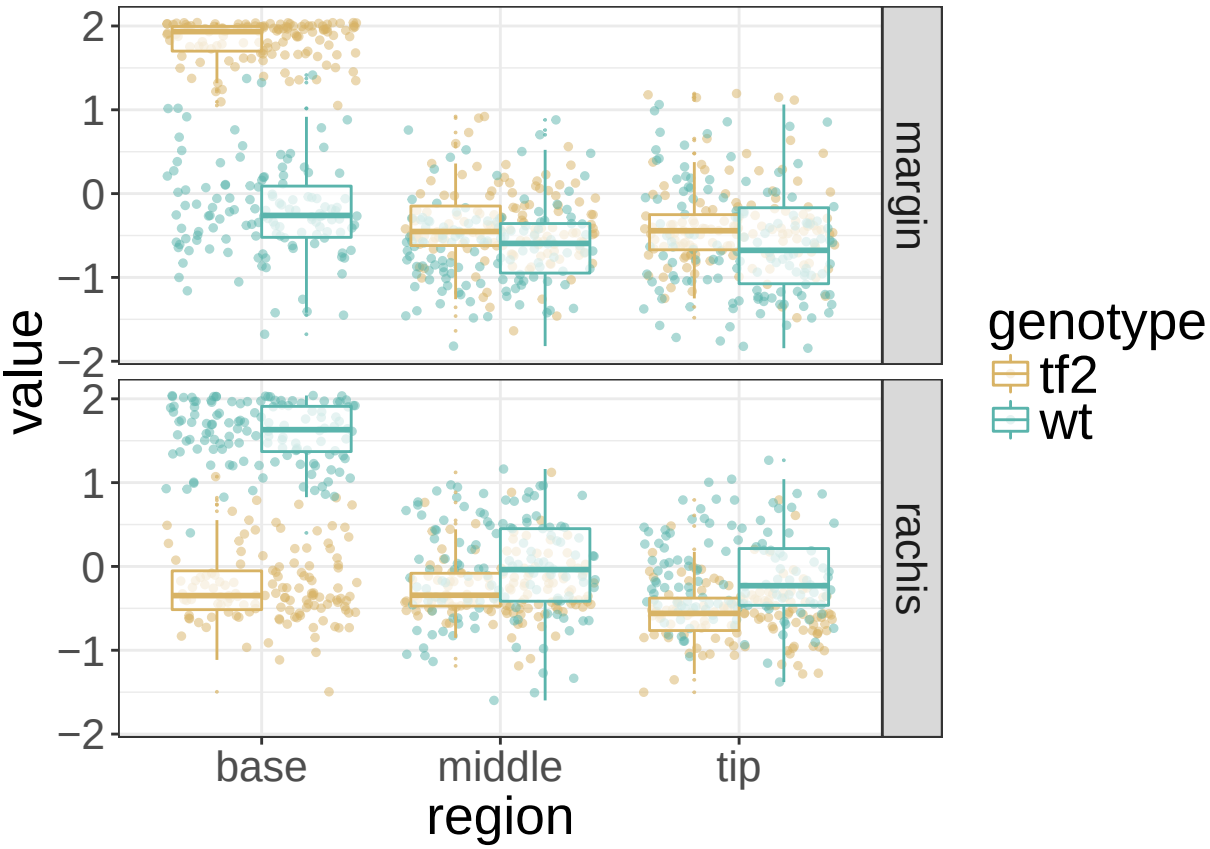

```
clusterVis_line_ssom(15)
```

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



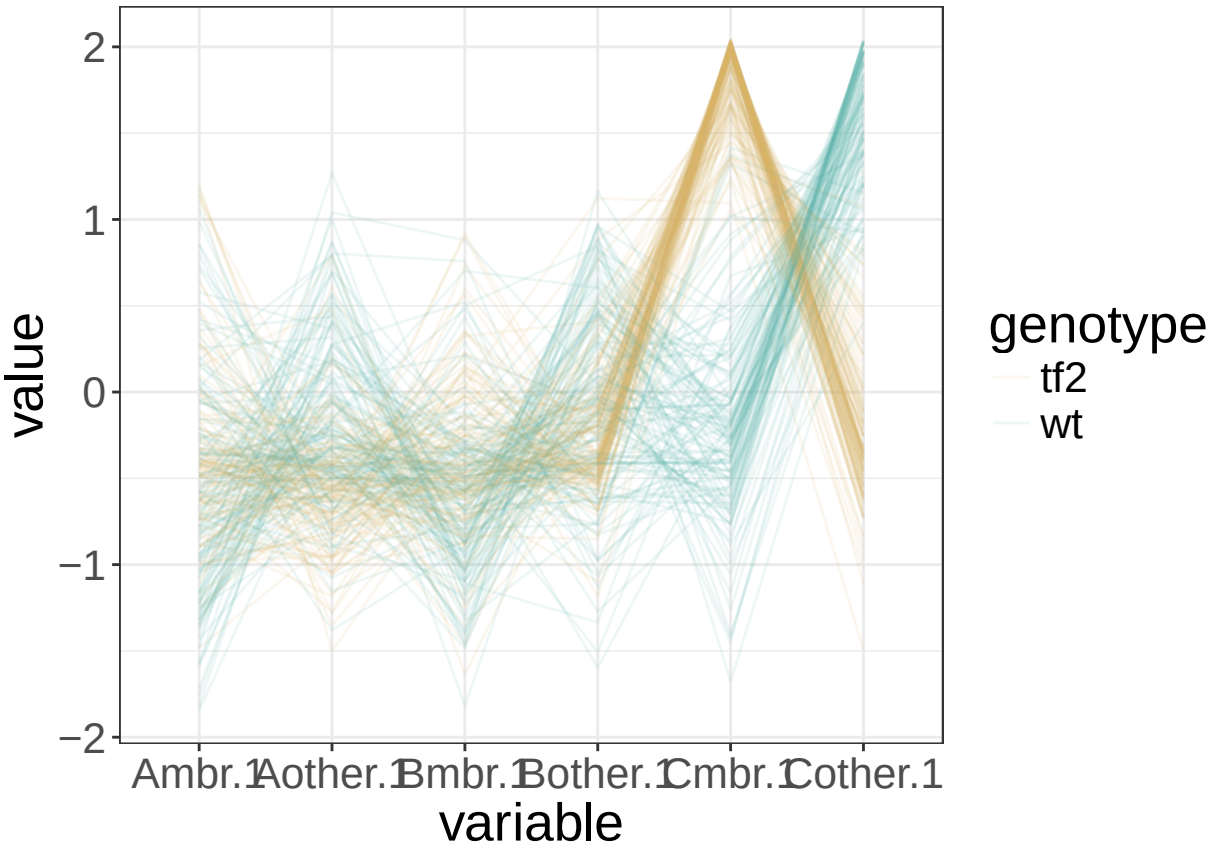

```
clusterVis_region_ssom(16)
```

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



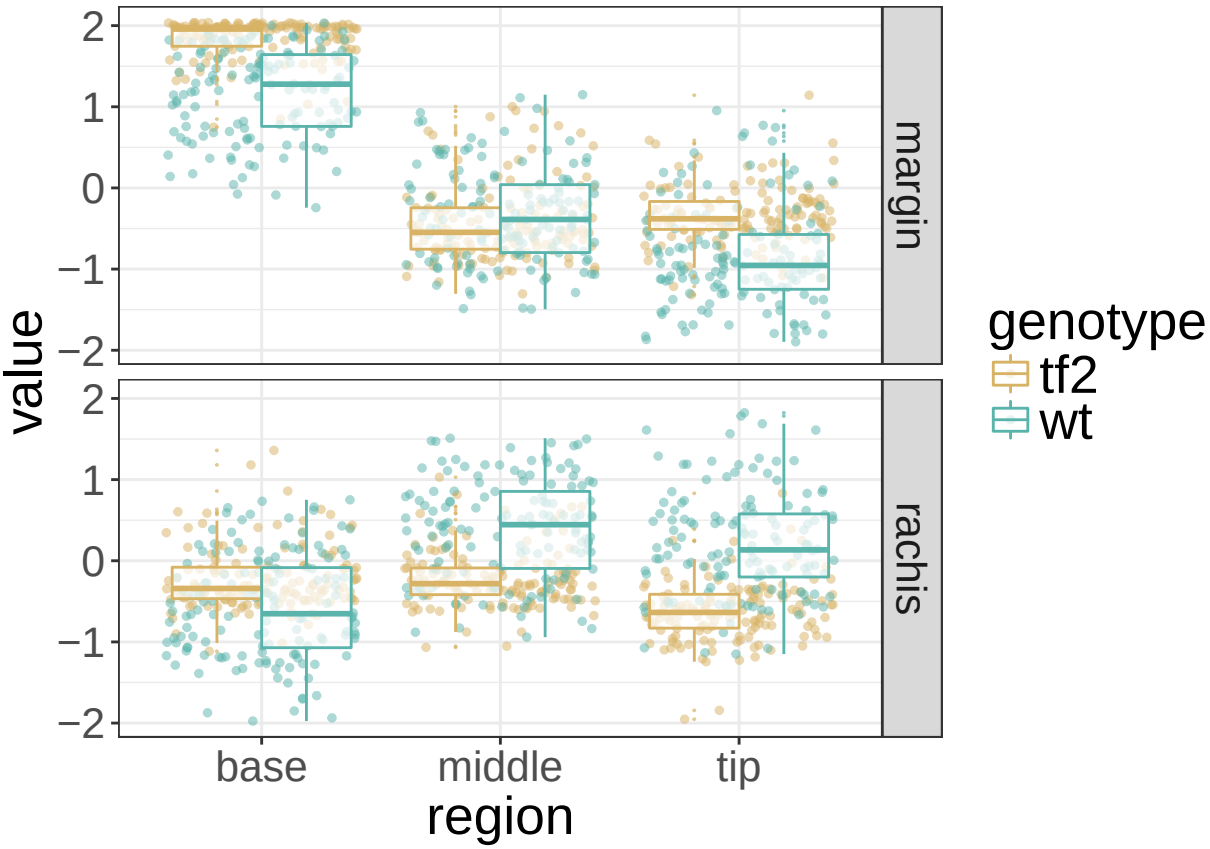

```
clusterVis_line_ssom(16)
```

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



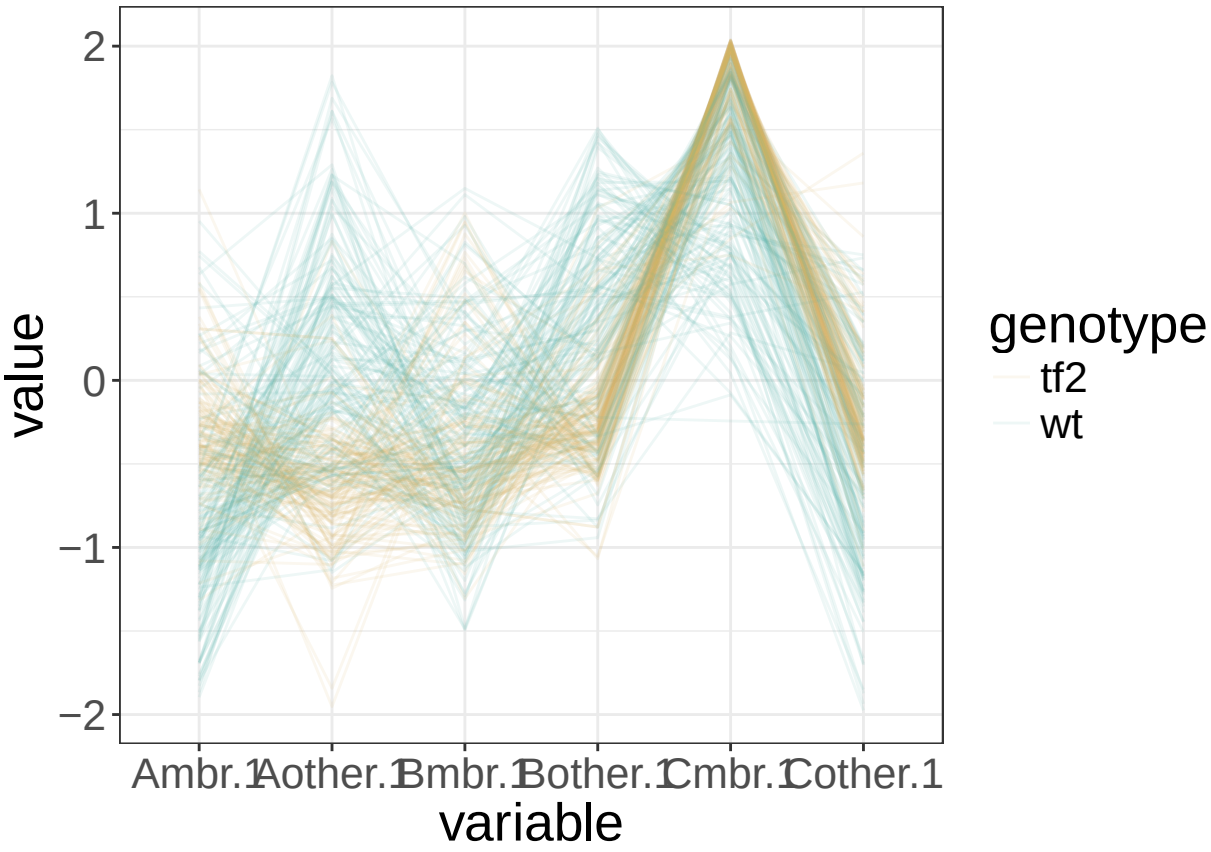

```
clusterVis_region_ssom(17)
```

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



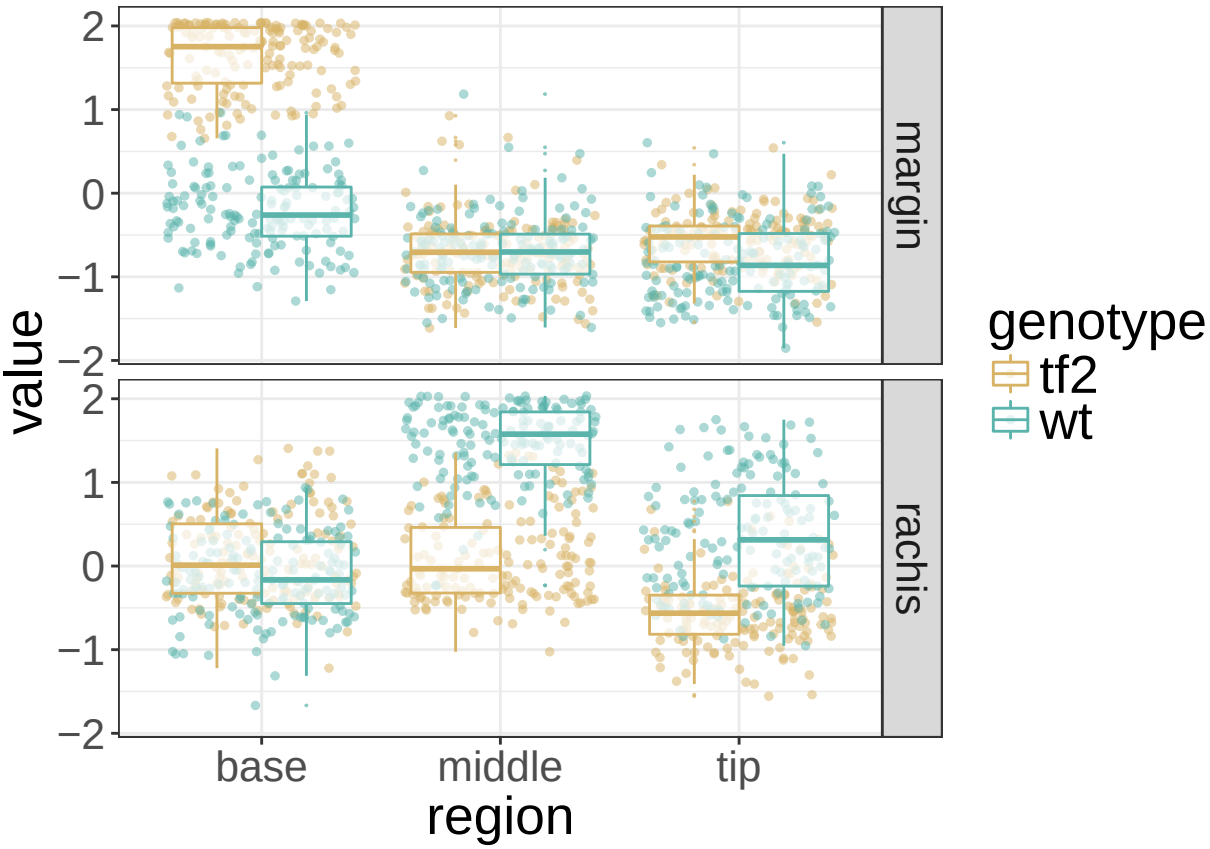

```
clusterVis_line_ssom(17)
```

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



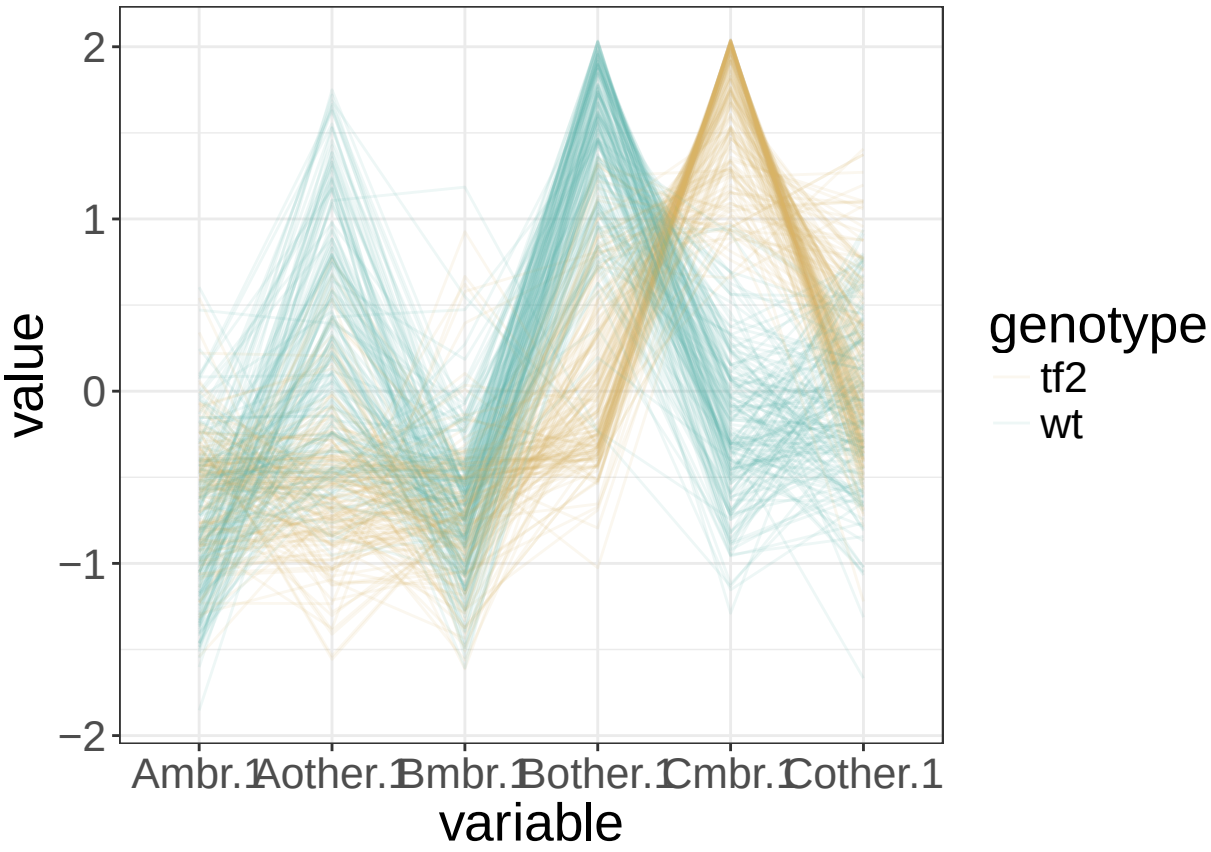

```
clusterVis_region_ssom(18)
```

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



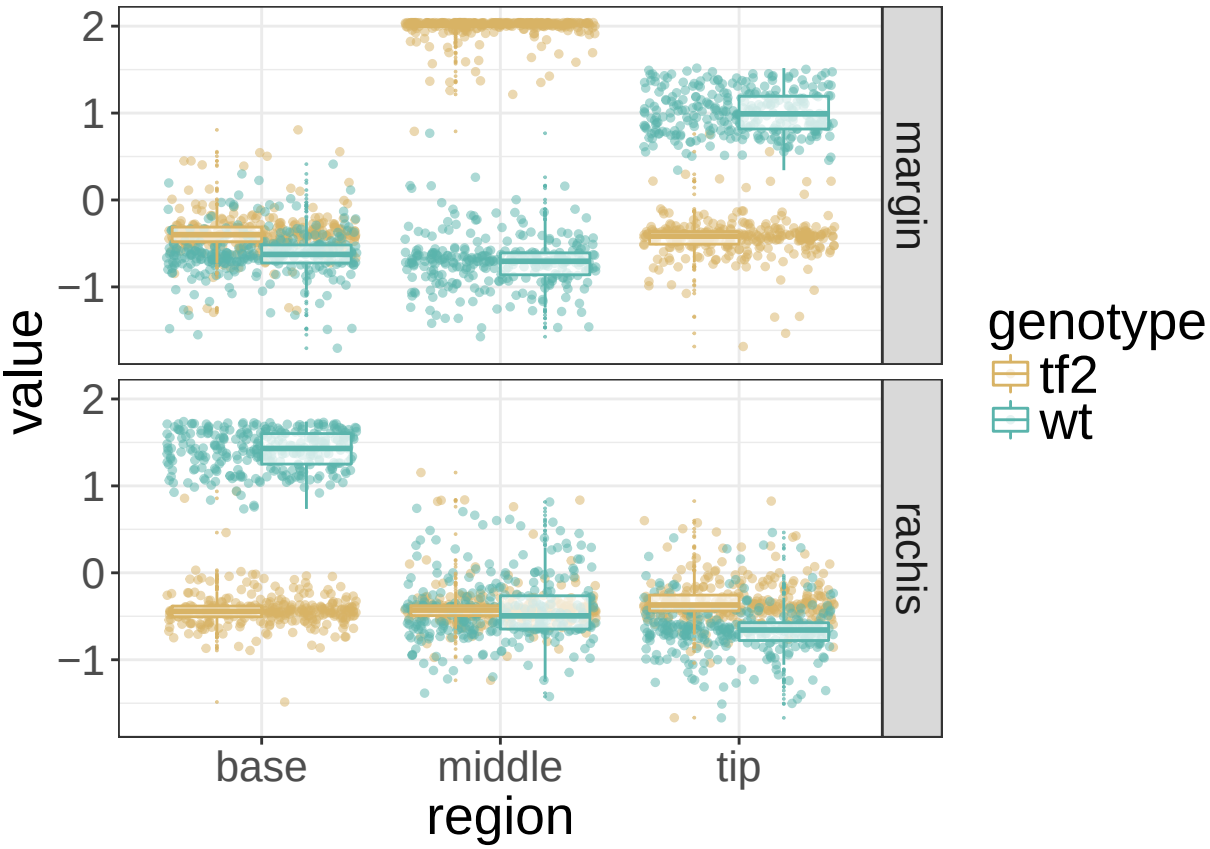

```
clusterVis_line_ssom(18)
```

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



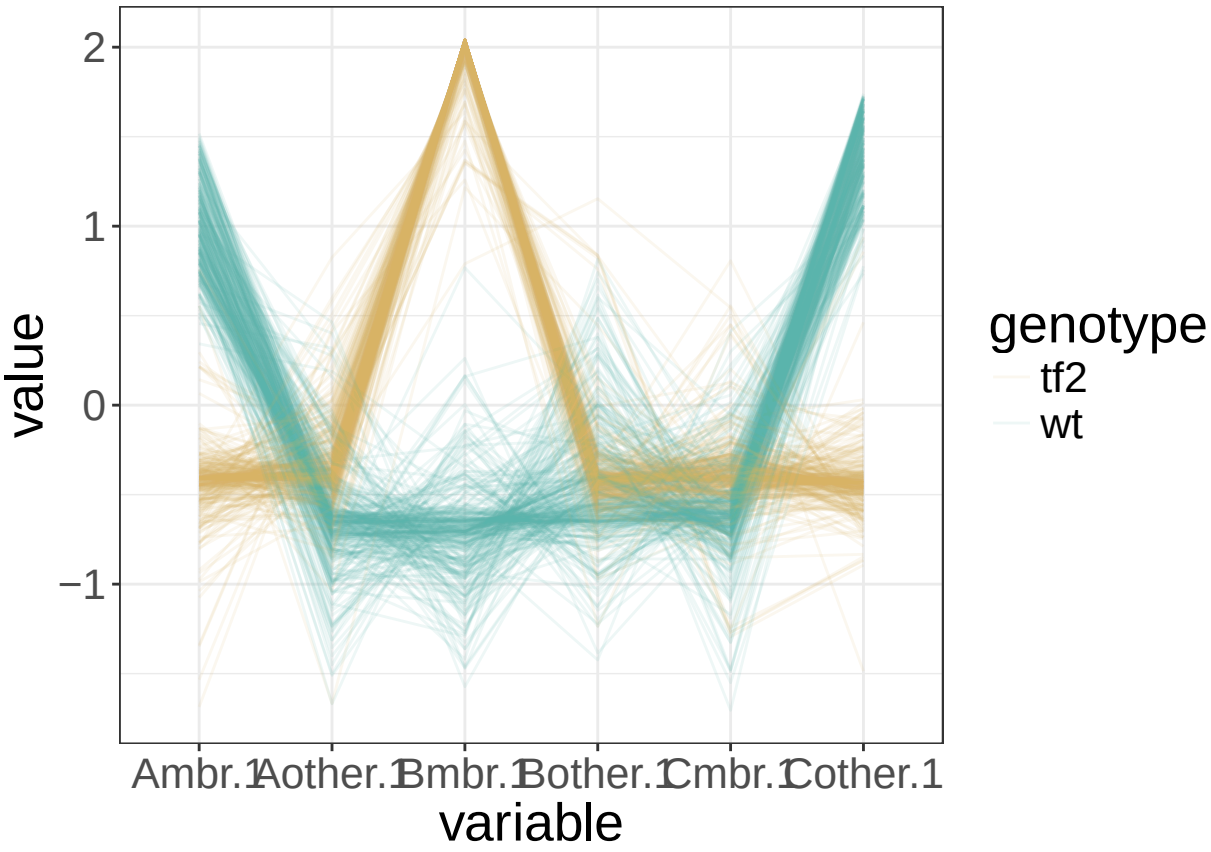

```
clusterVis_region_ssom(19)
```

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



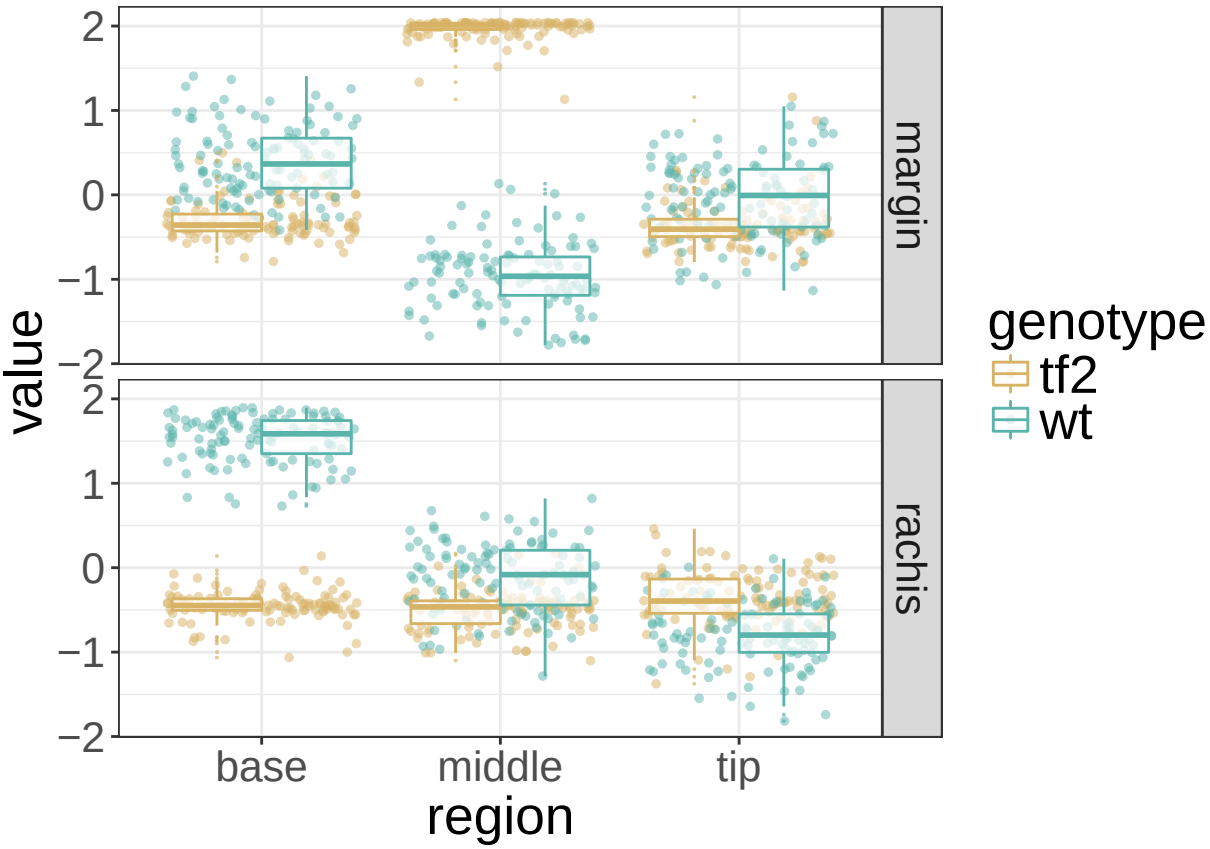

```
clusterVis_line_ssom(19)
```

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



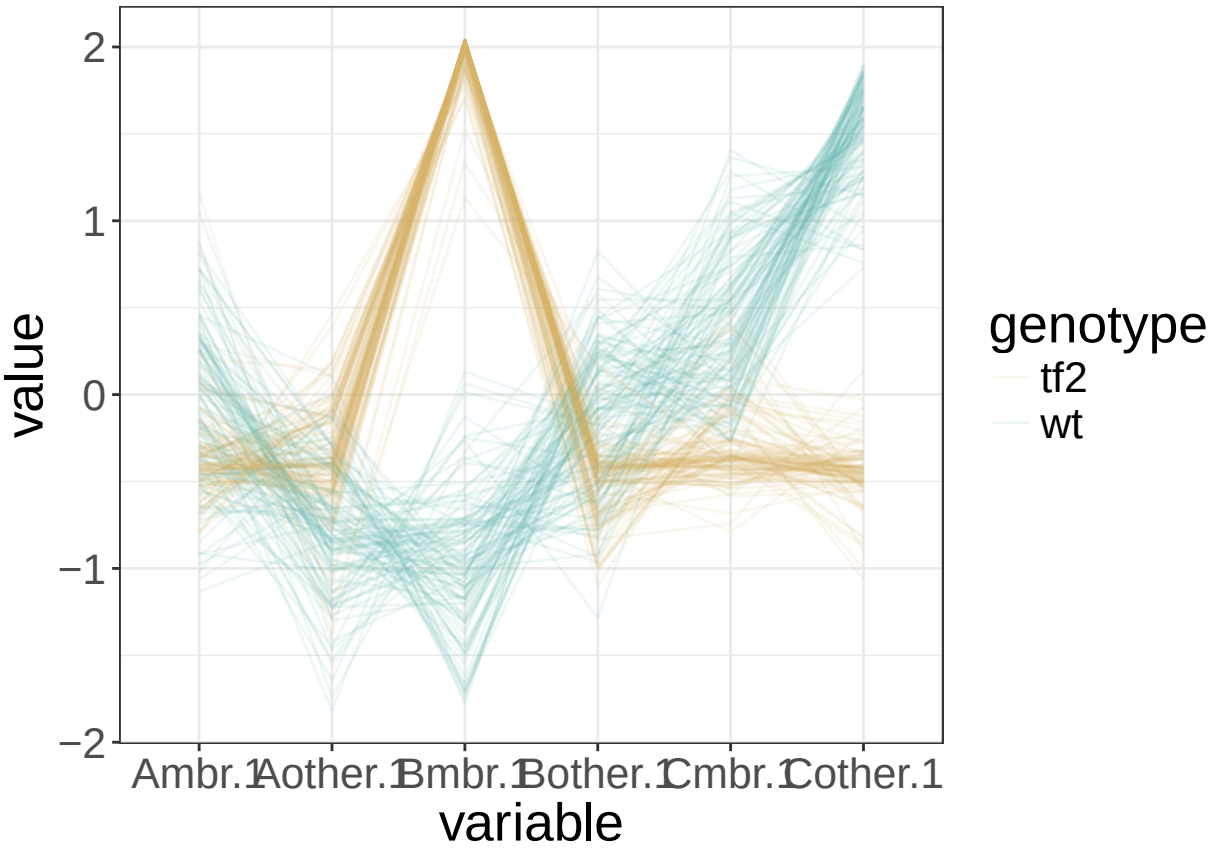

```
clusterVis_region_ssom(20)
```

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



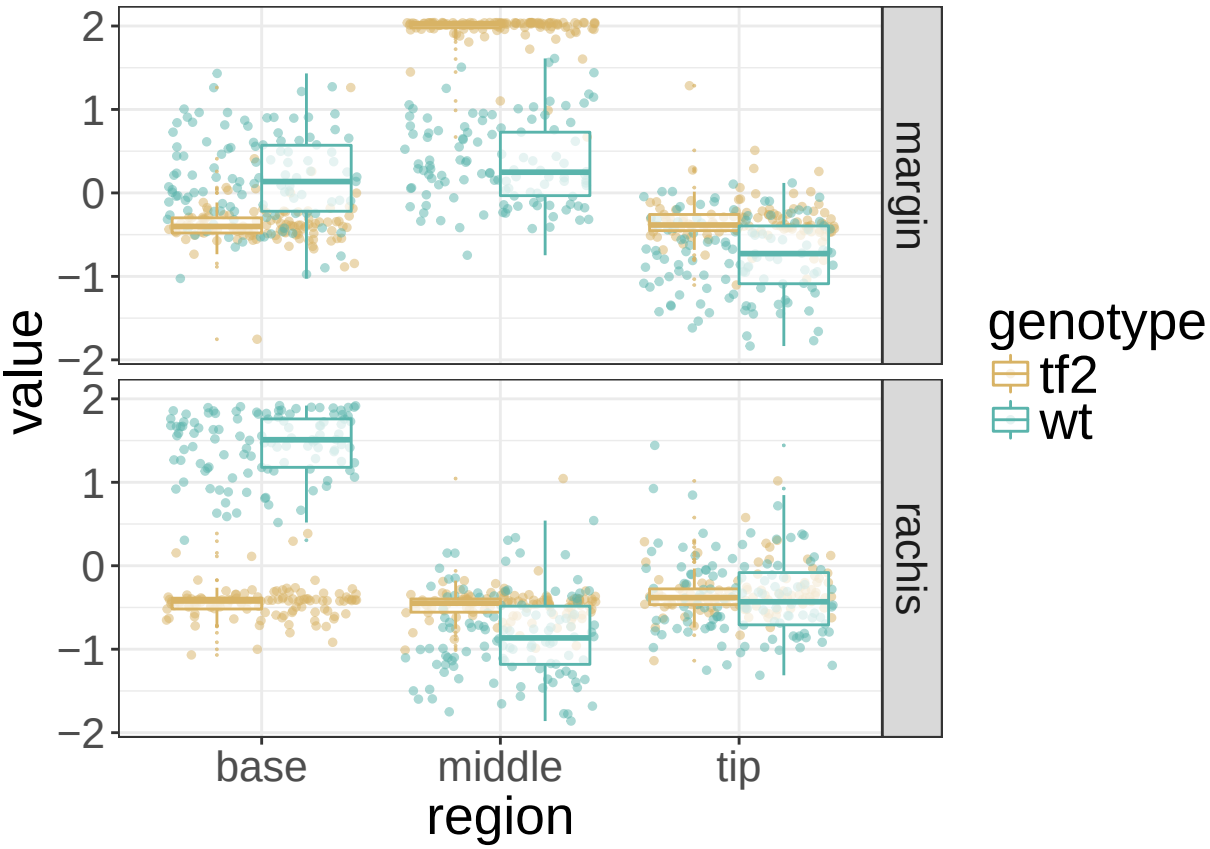

```
clusterVis_line_ssom(20)
```

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



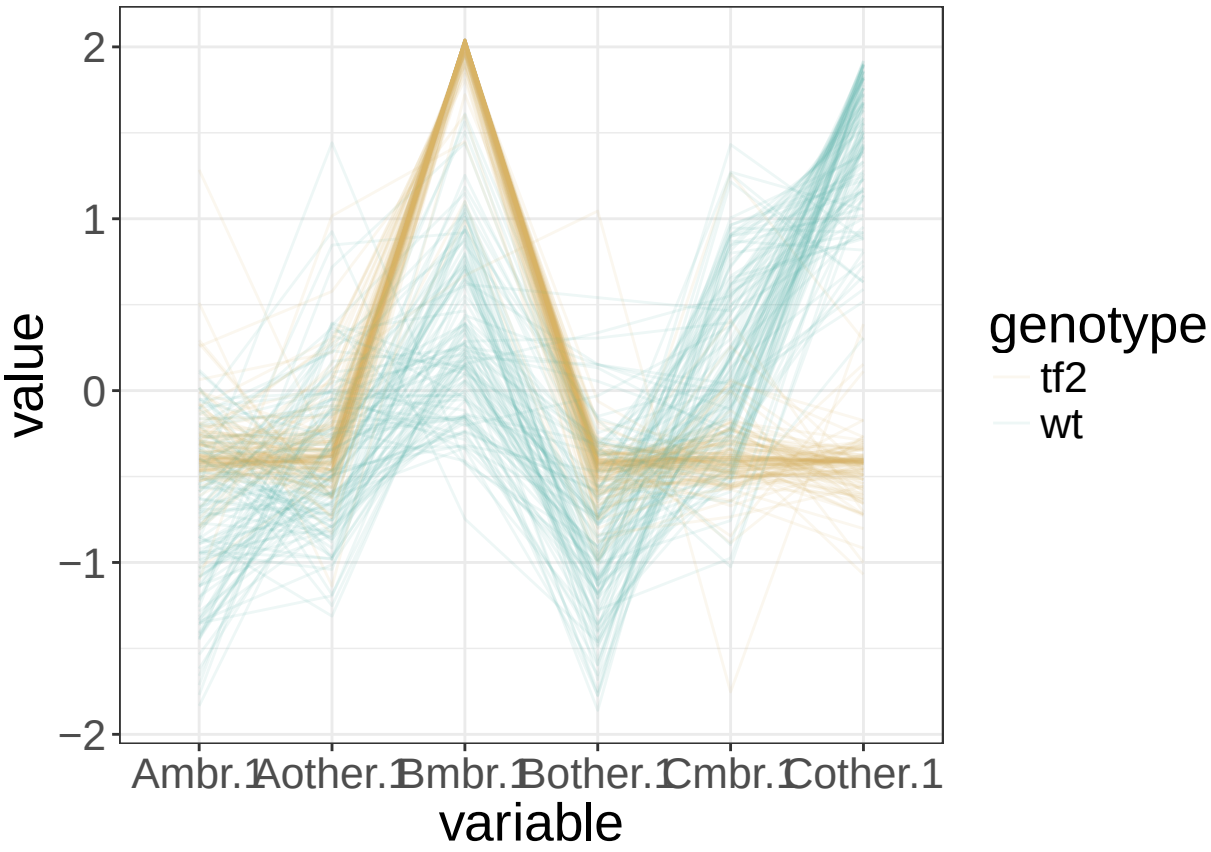

```
clusterVis_region_ssom(21)
```

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



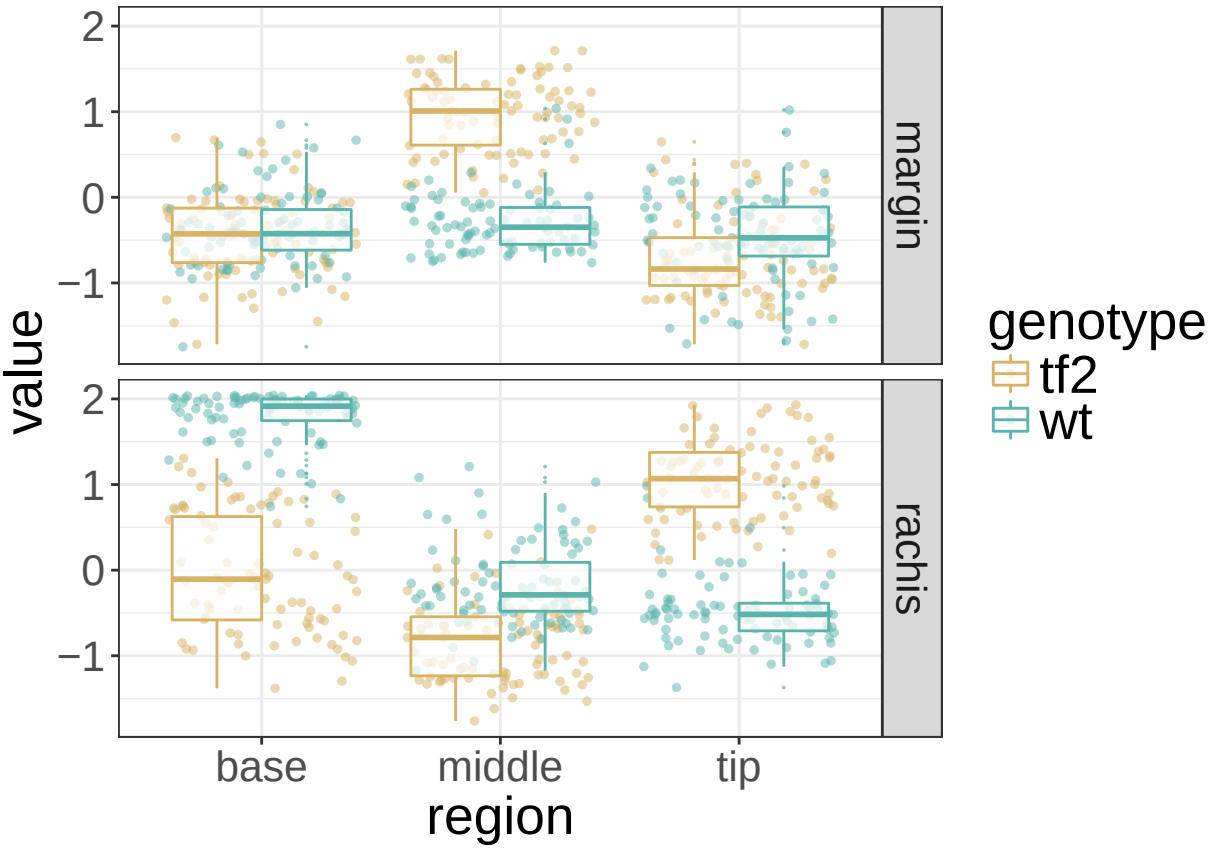

```
clusterVis_line_ssom(21)
```

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



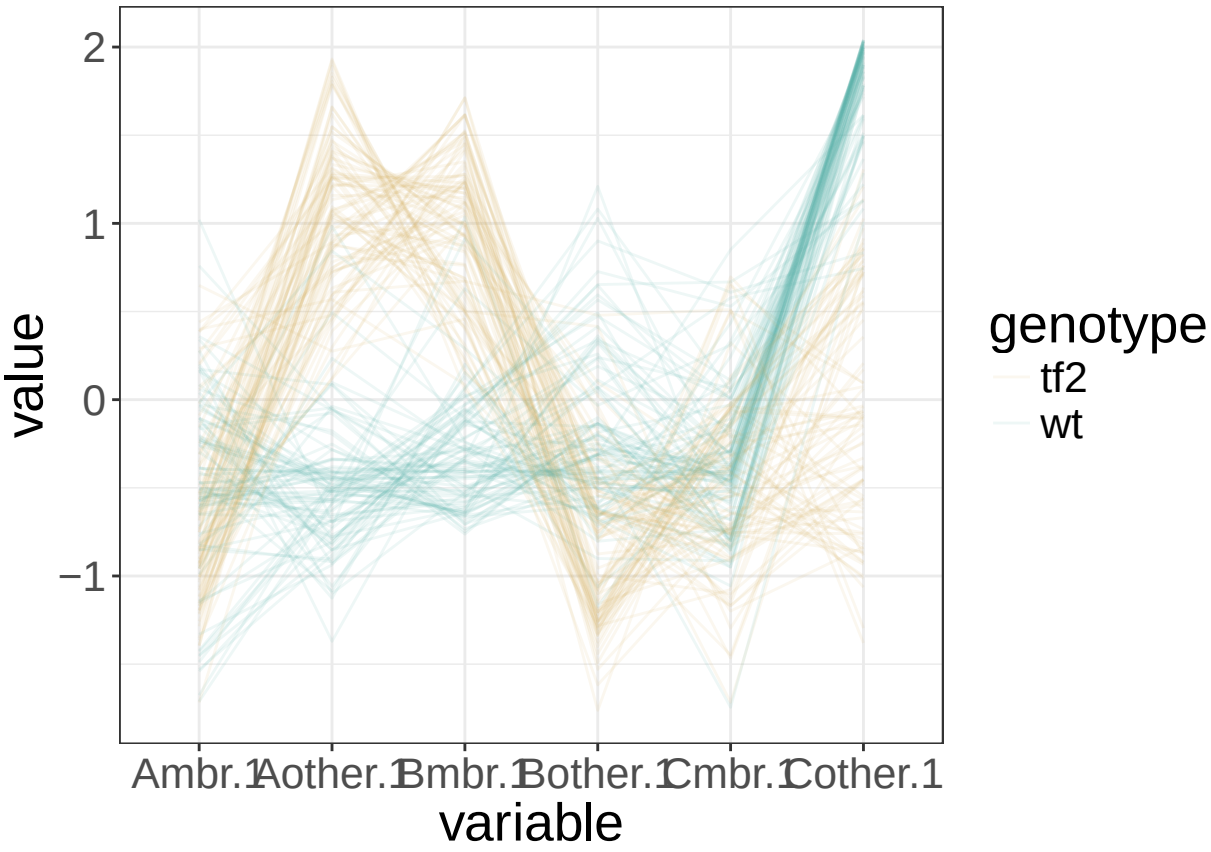

```
clusterVis_region_ssom(22)
```

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



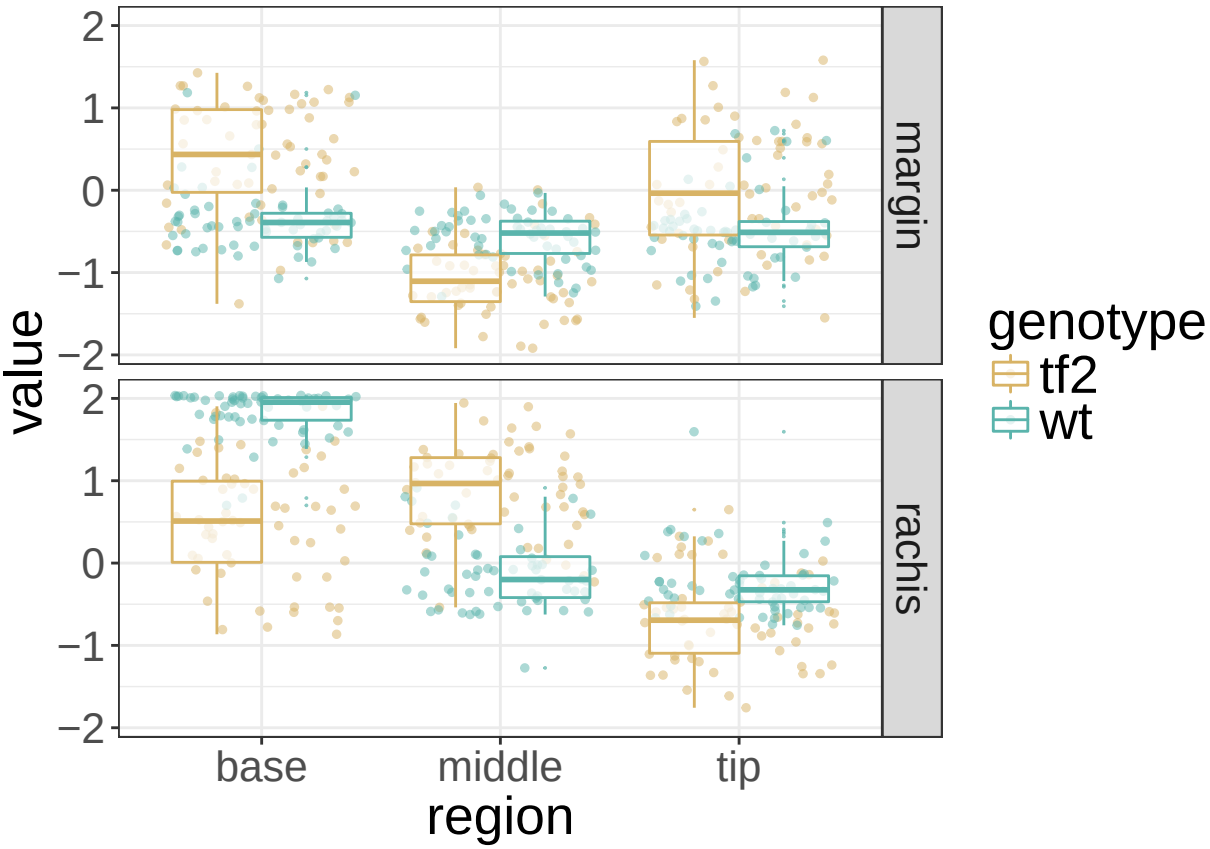

```
clusterVis_line_ssom(22)
```

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



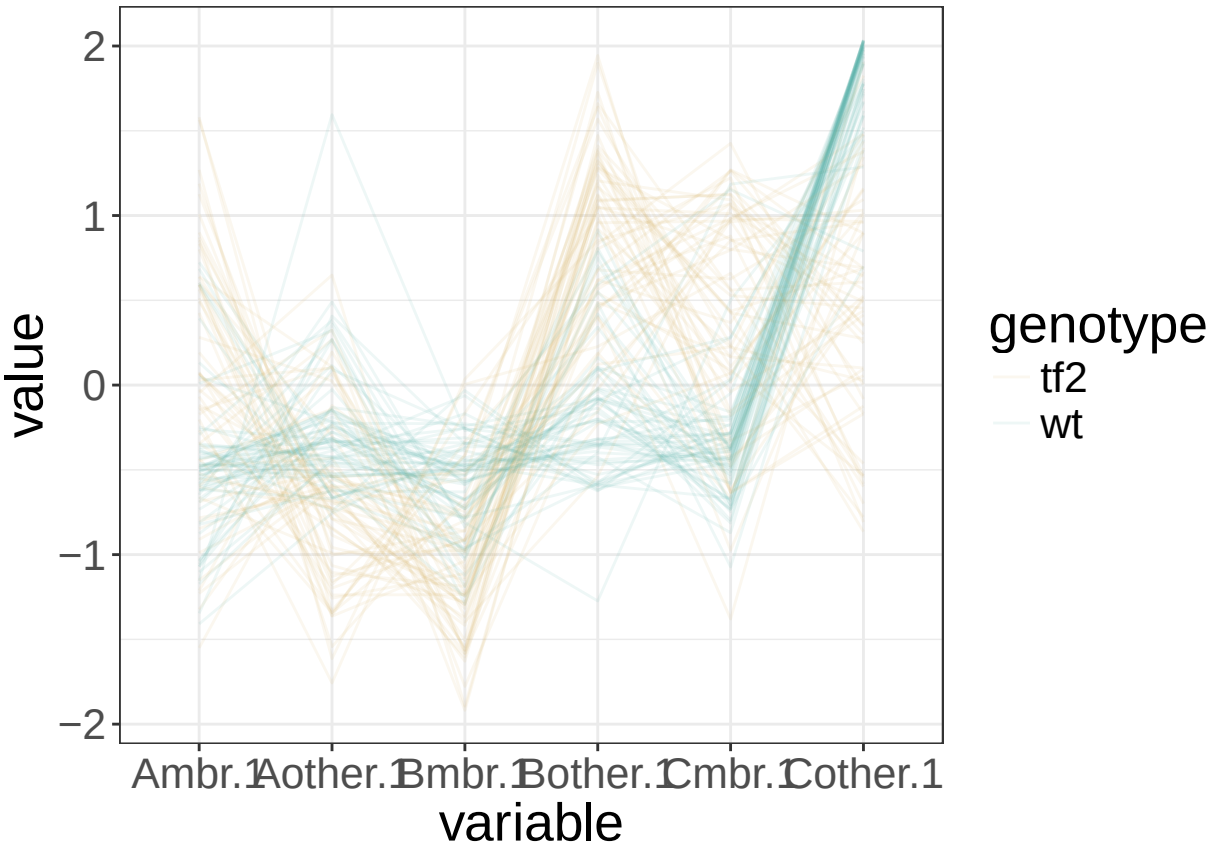

```
clusterVis_region_ssom(23)
```

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



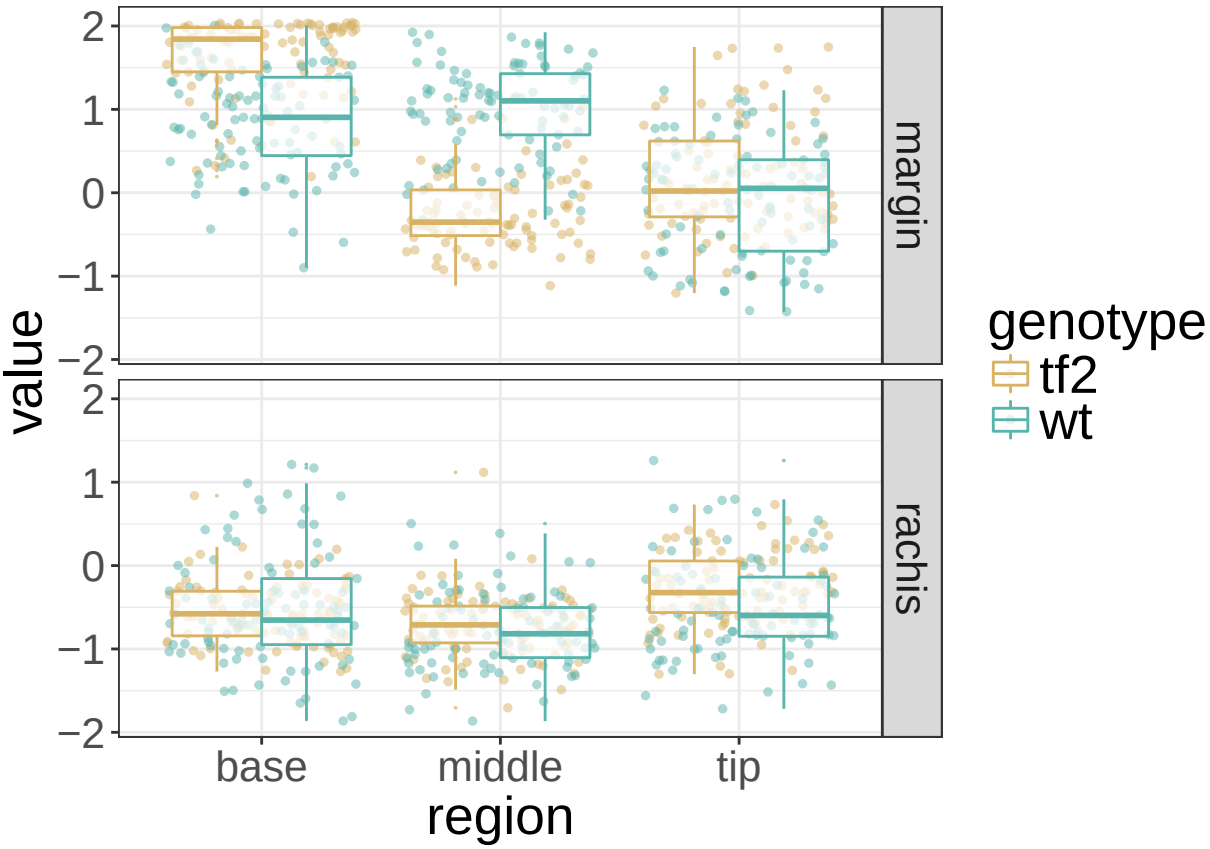

```
clusterVis_line_ssom(23)
```

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



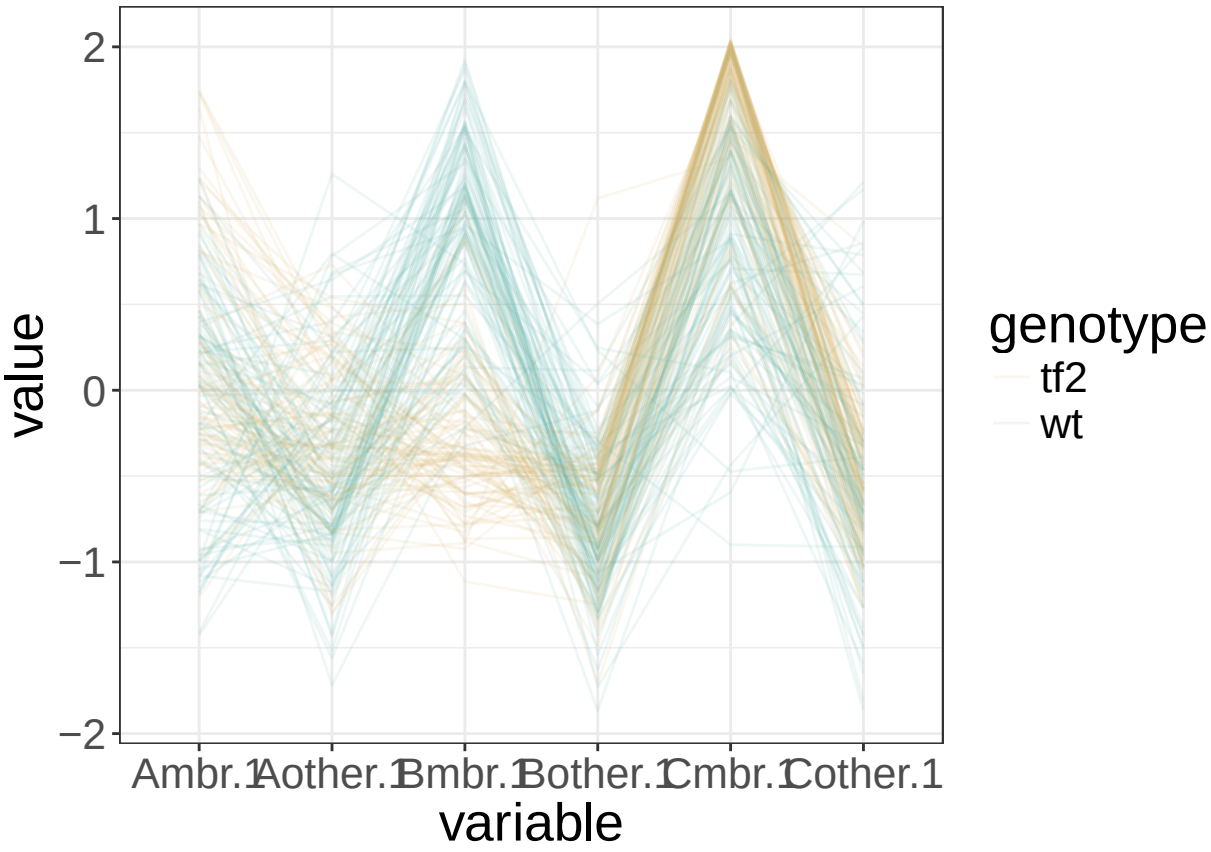

```
clusterVis_region_ssom(24)
```

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



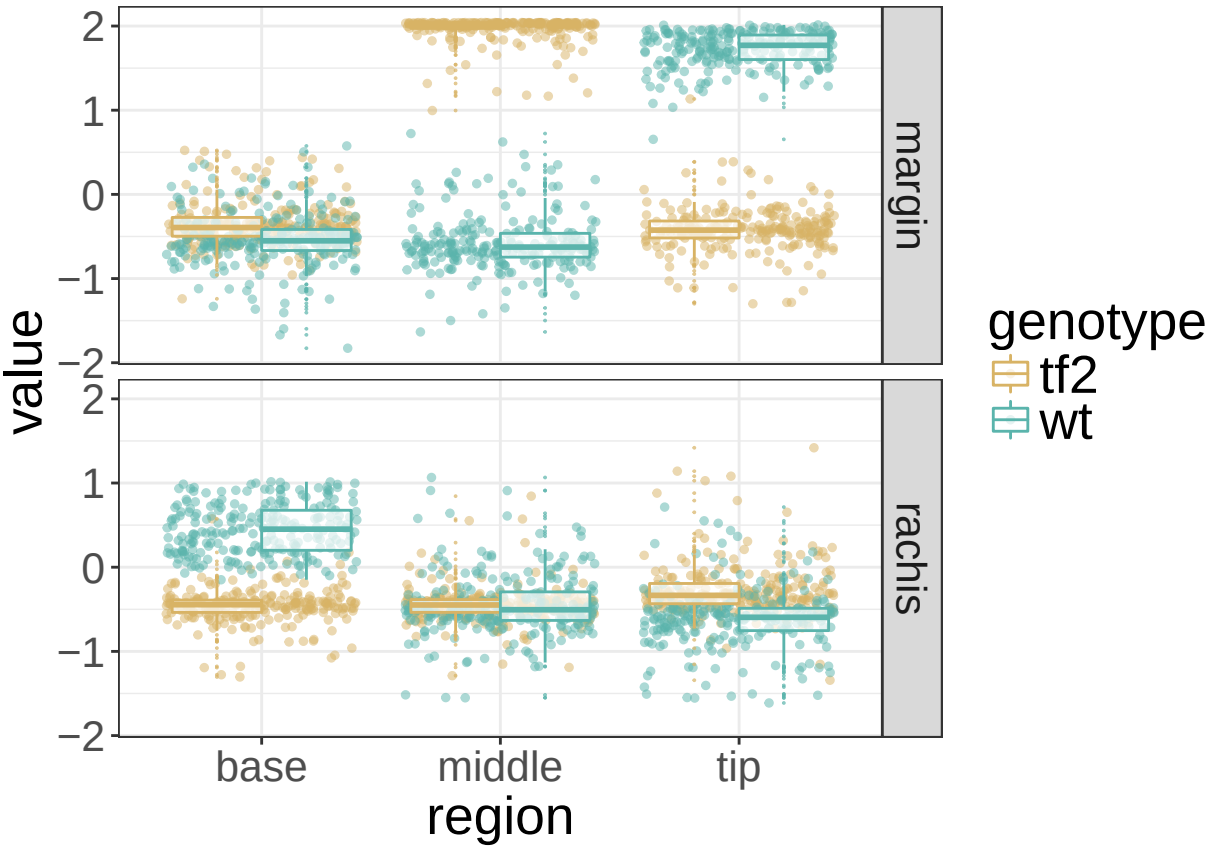

```
clusterVis_line_ssom(24)
```

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



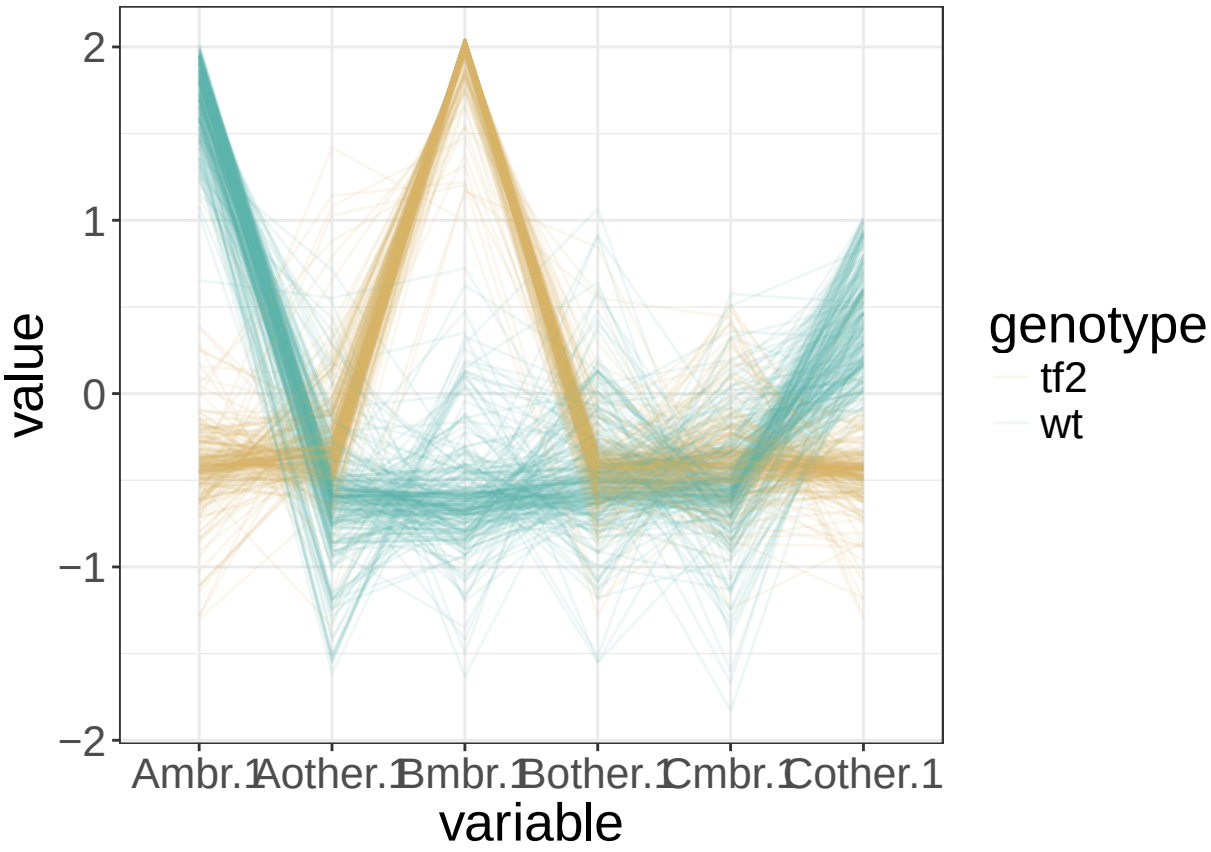

```
clusterVis_region_ssom(25)
```

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



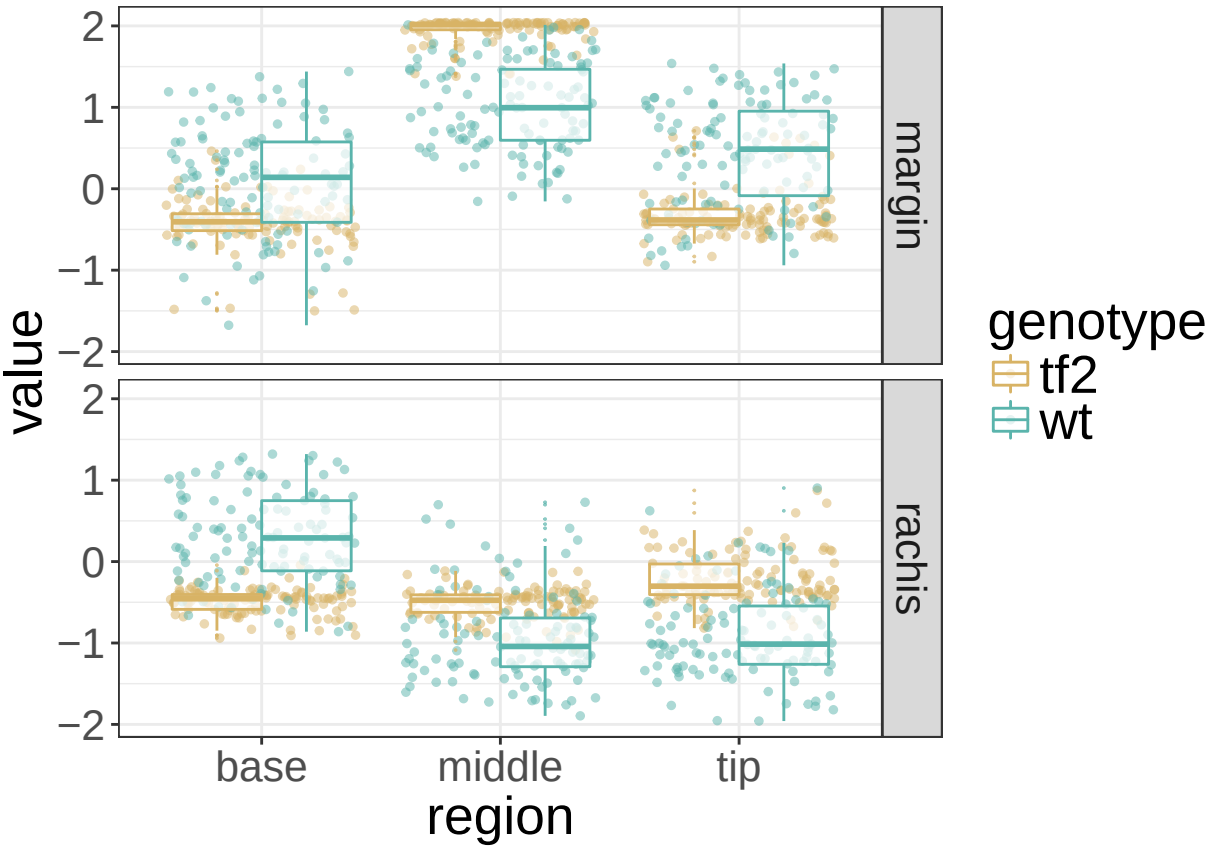

```
clusterVis_line_ssom(25)
```

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



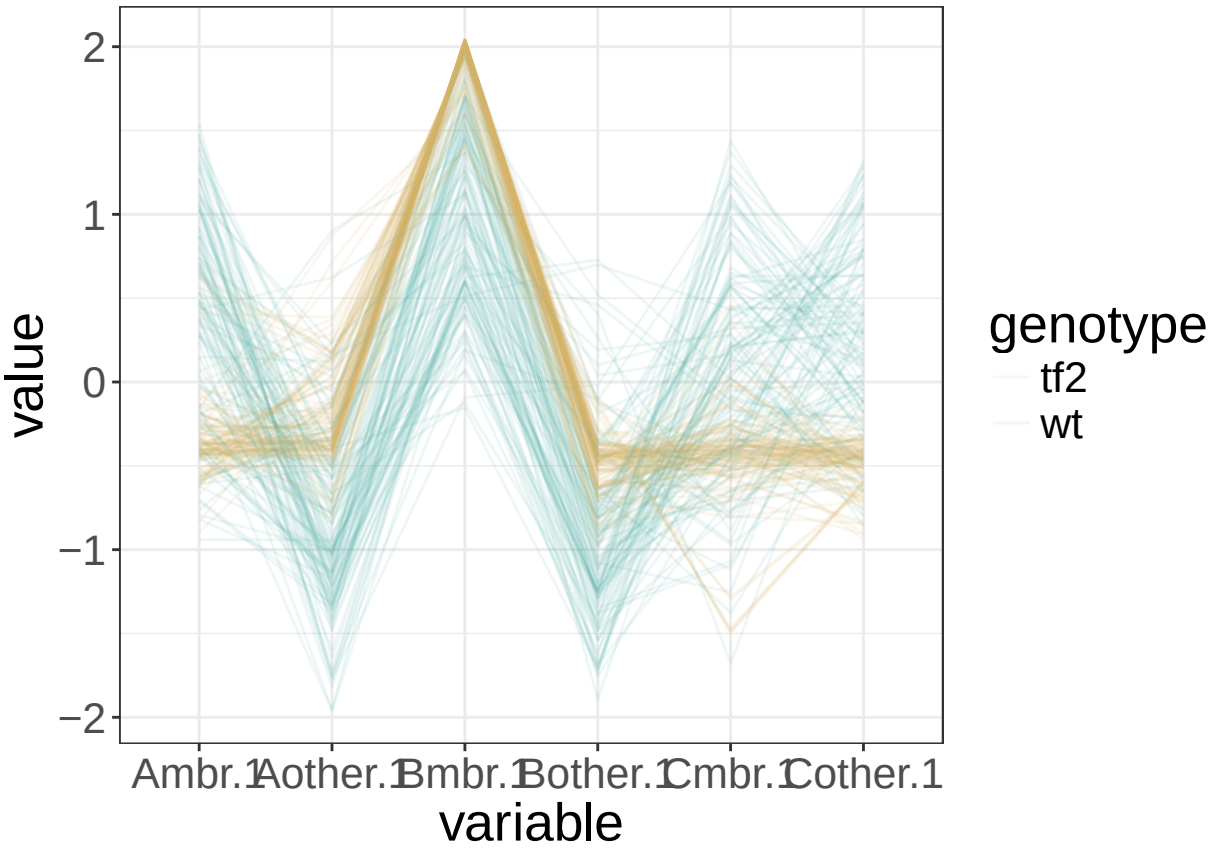

```
clusterVis_region_ssom(26)
```

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



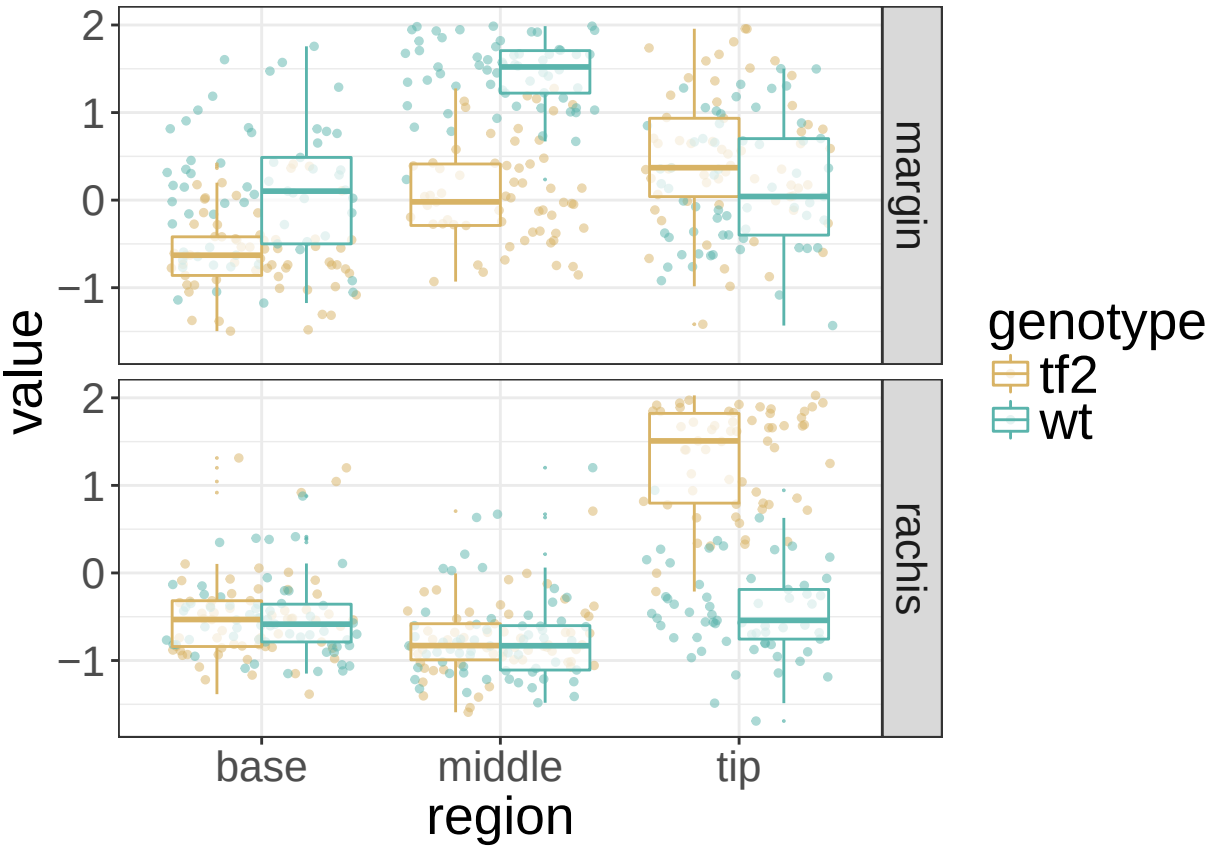

```
clusterVis_line_ssom(26)
```

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



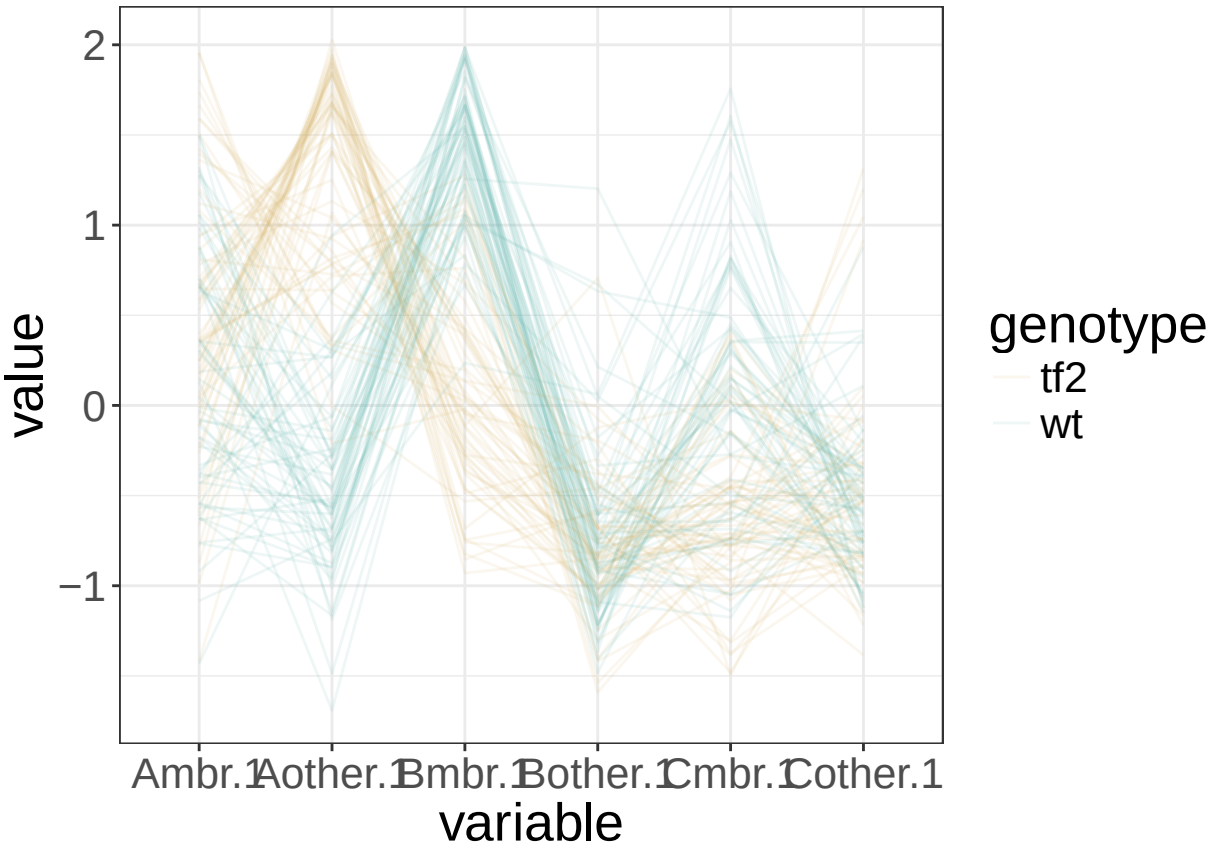

```
clusterVis_region_ssom(27)
```

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



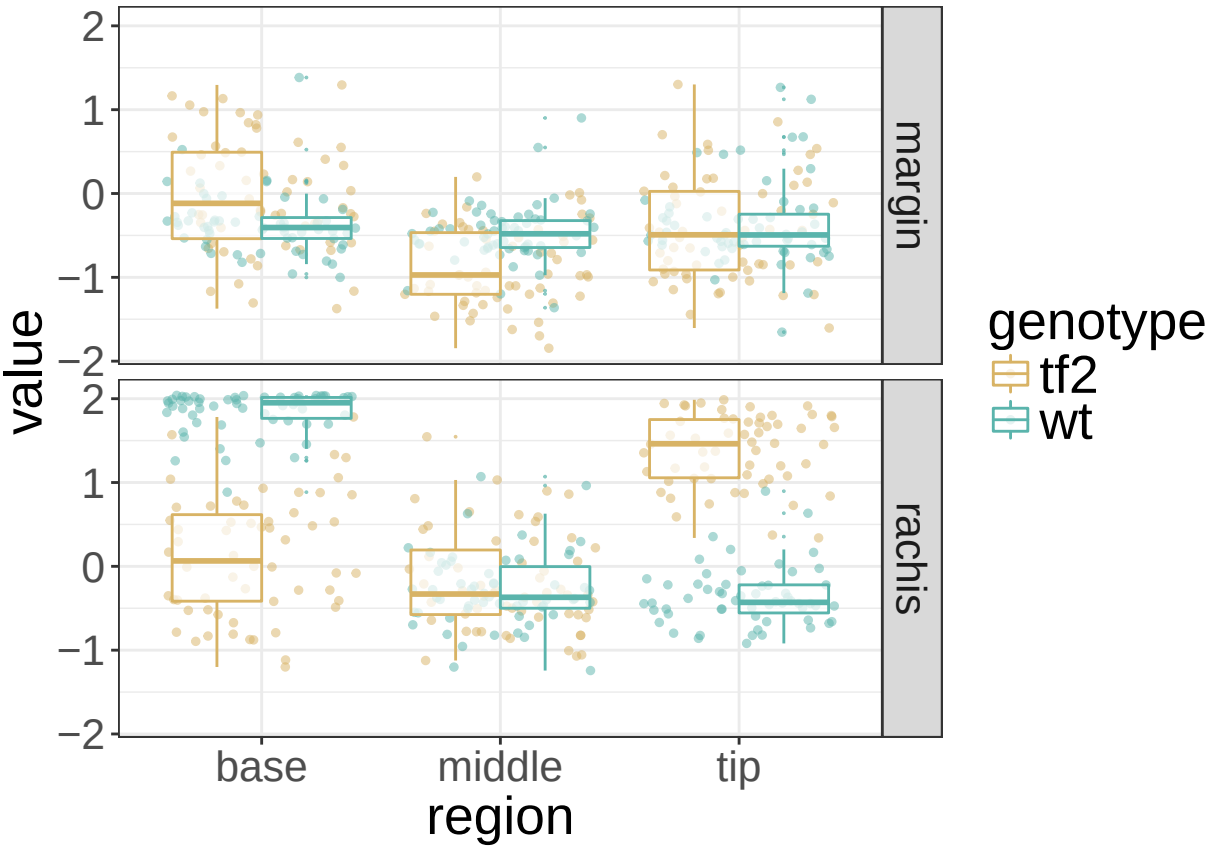

```
clusterVis_line_ssom(27)
```

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



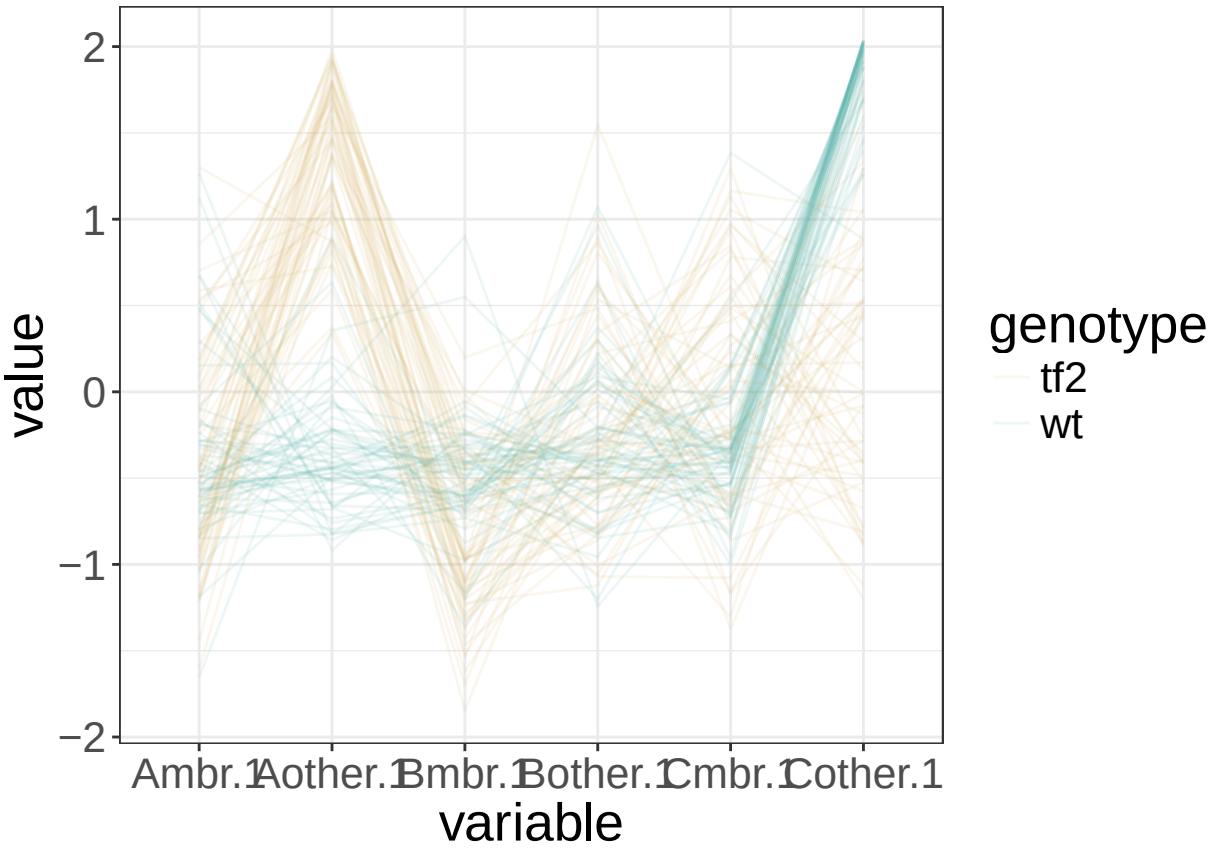

```
clusterVis_region_ssom(28)
```

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



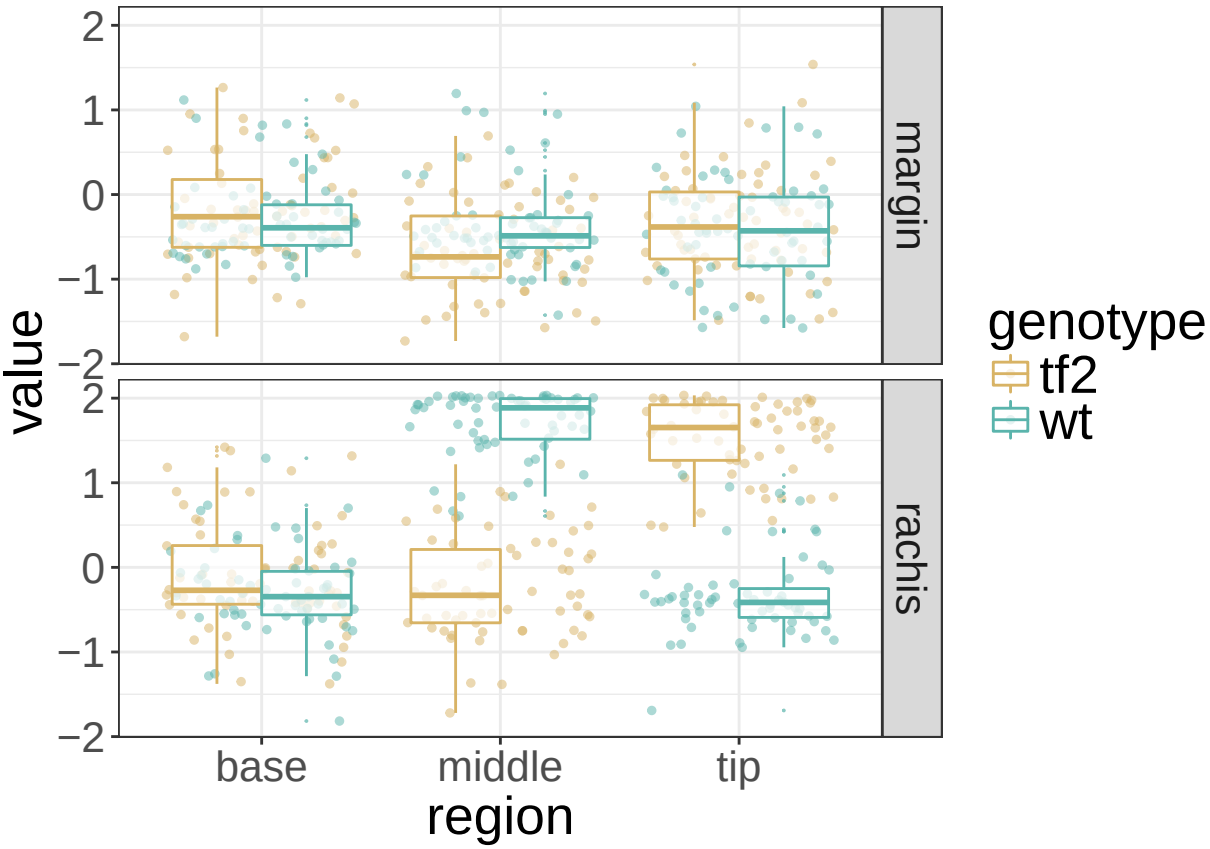

```
clusterVis_line_ssom(28)
```

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



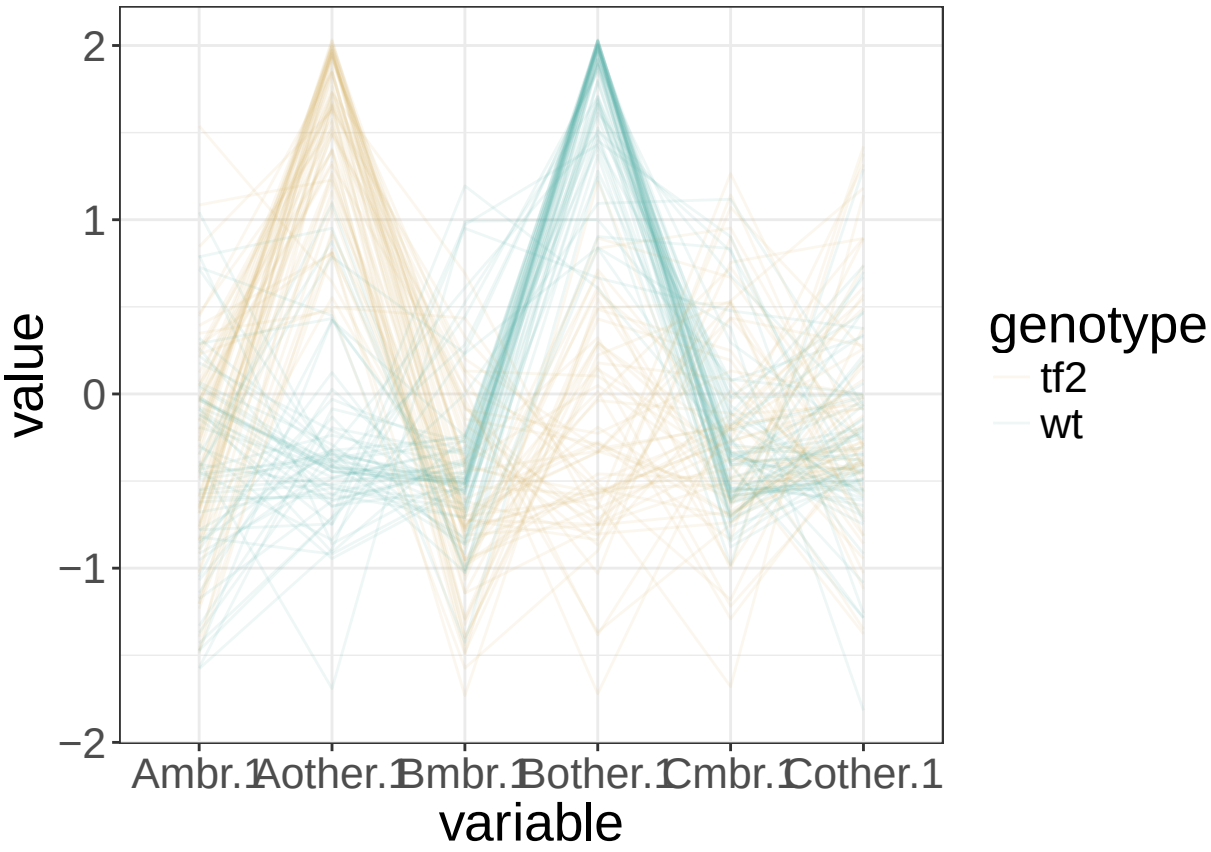

```
clusterVis_region_ssom(29)
```

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



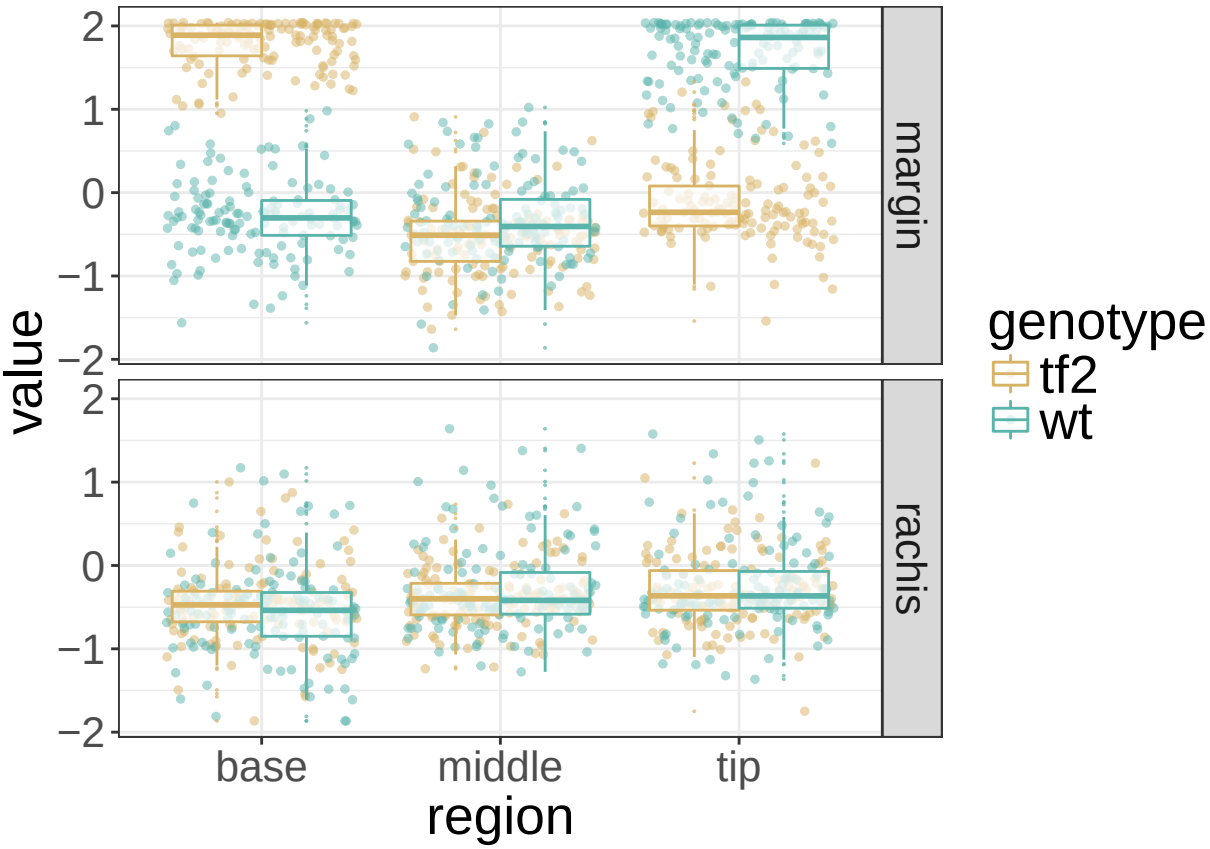

```
clusterVis_line_ssom(29)
```

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



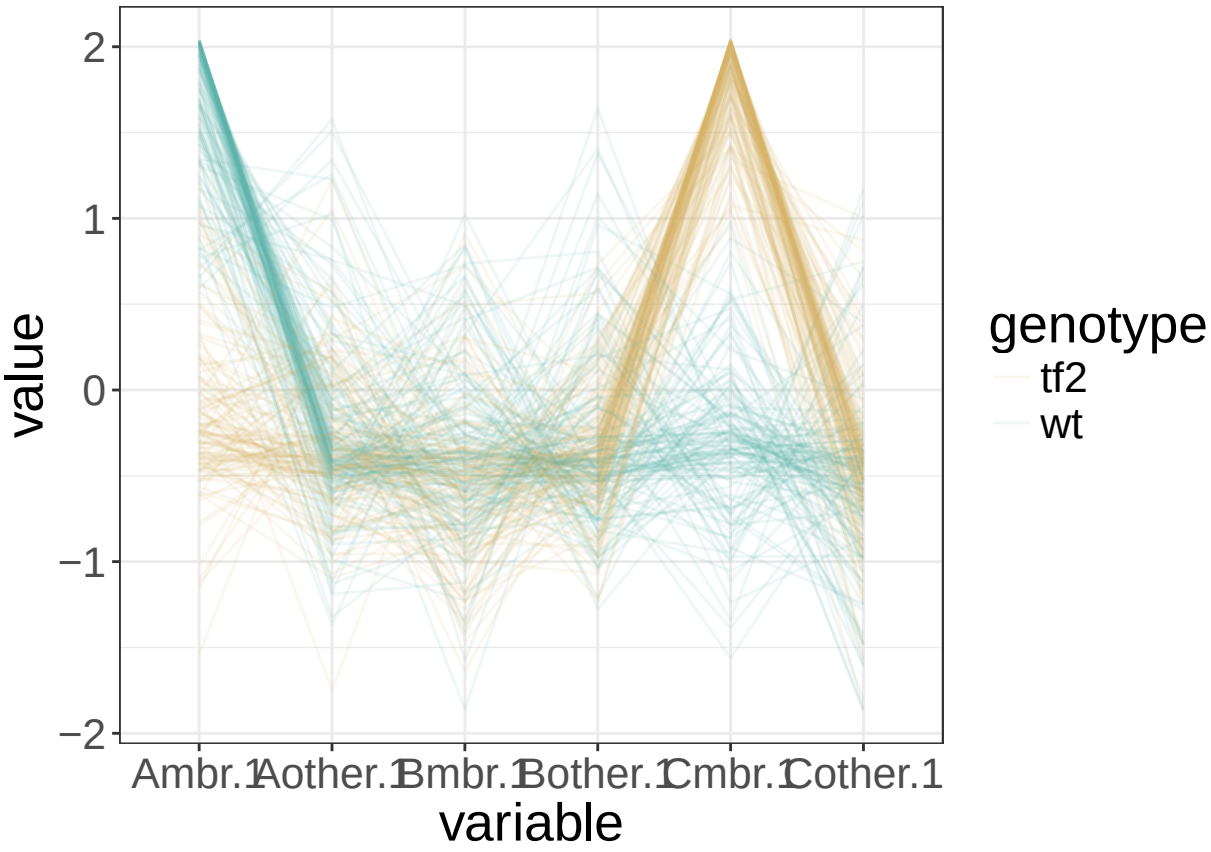

```
clusterVis_region_ssom(30)
```

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



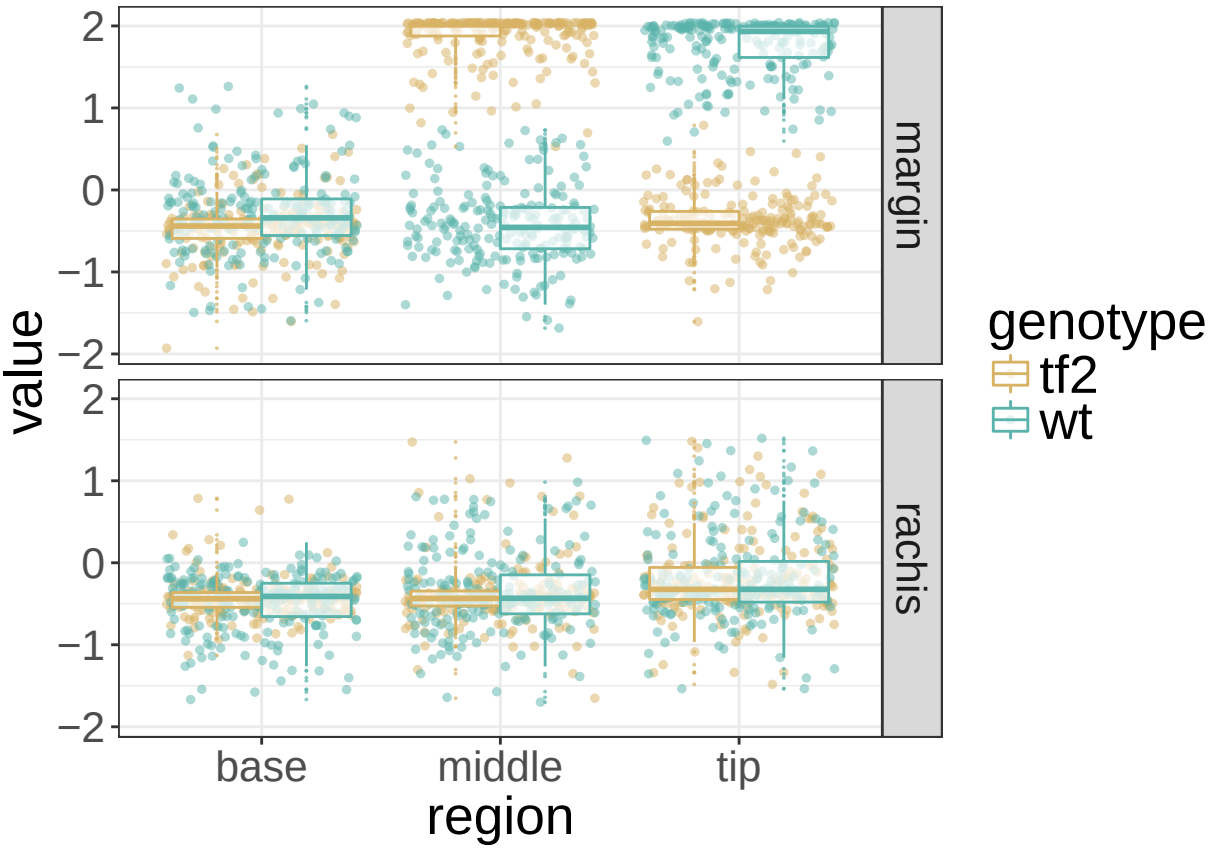

```
clusterVis_line_ssom(30)
```

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



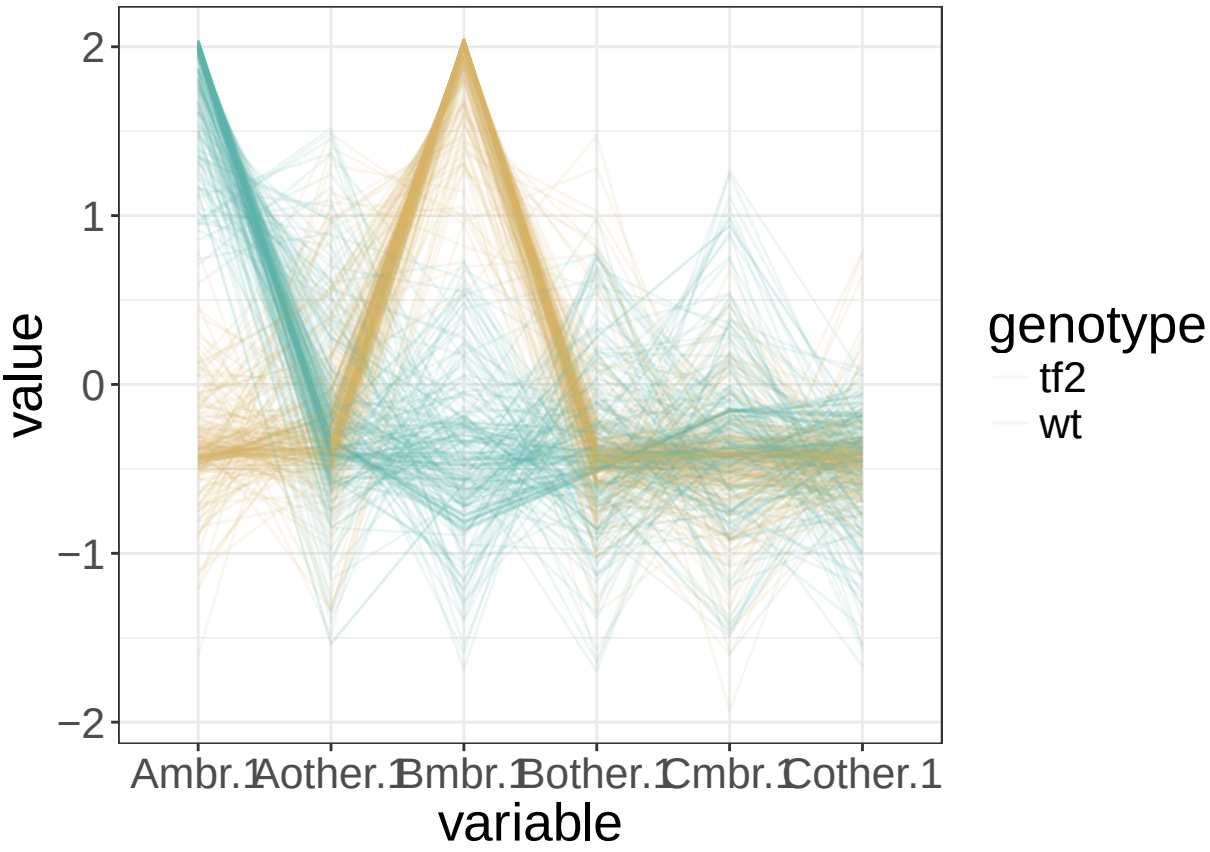

```
clusterVis_region_ssom(31)
```

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



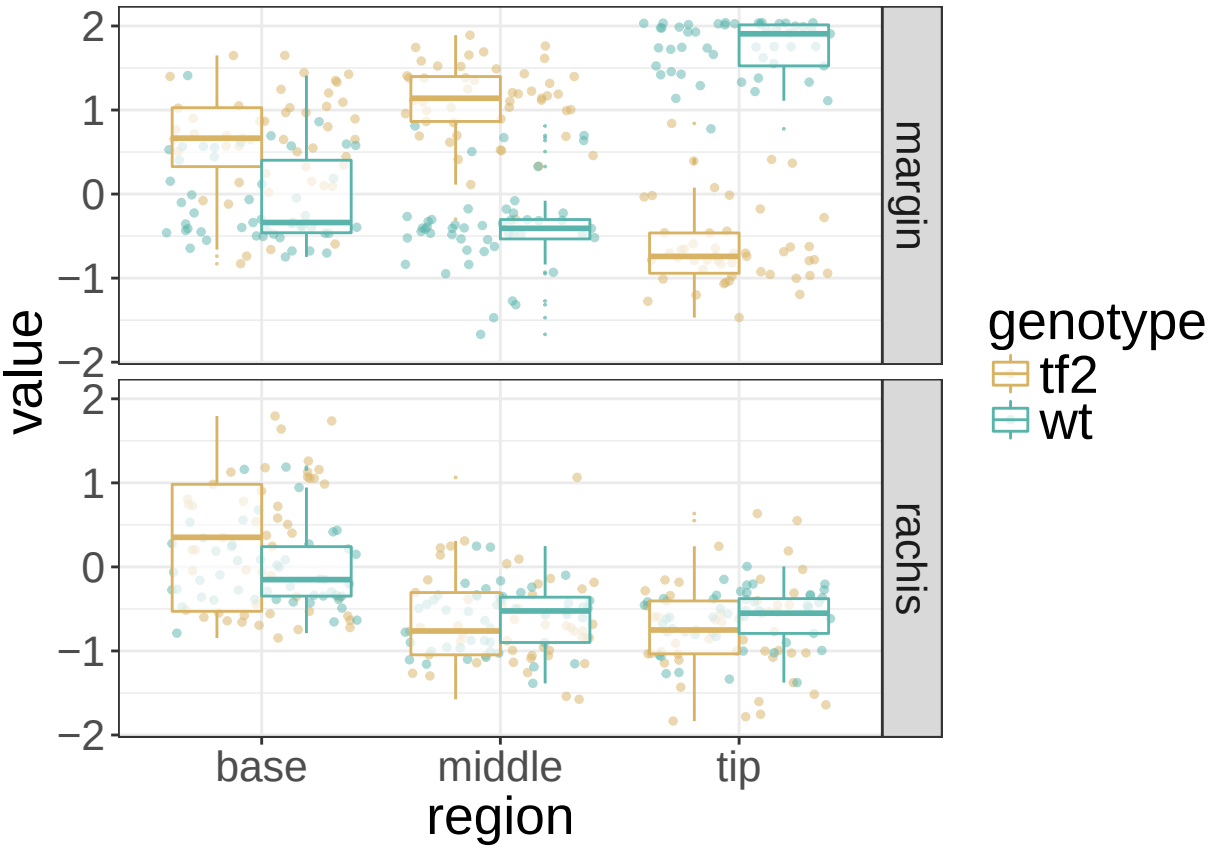

```
clusterVis_line_ssom(31)
```

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



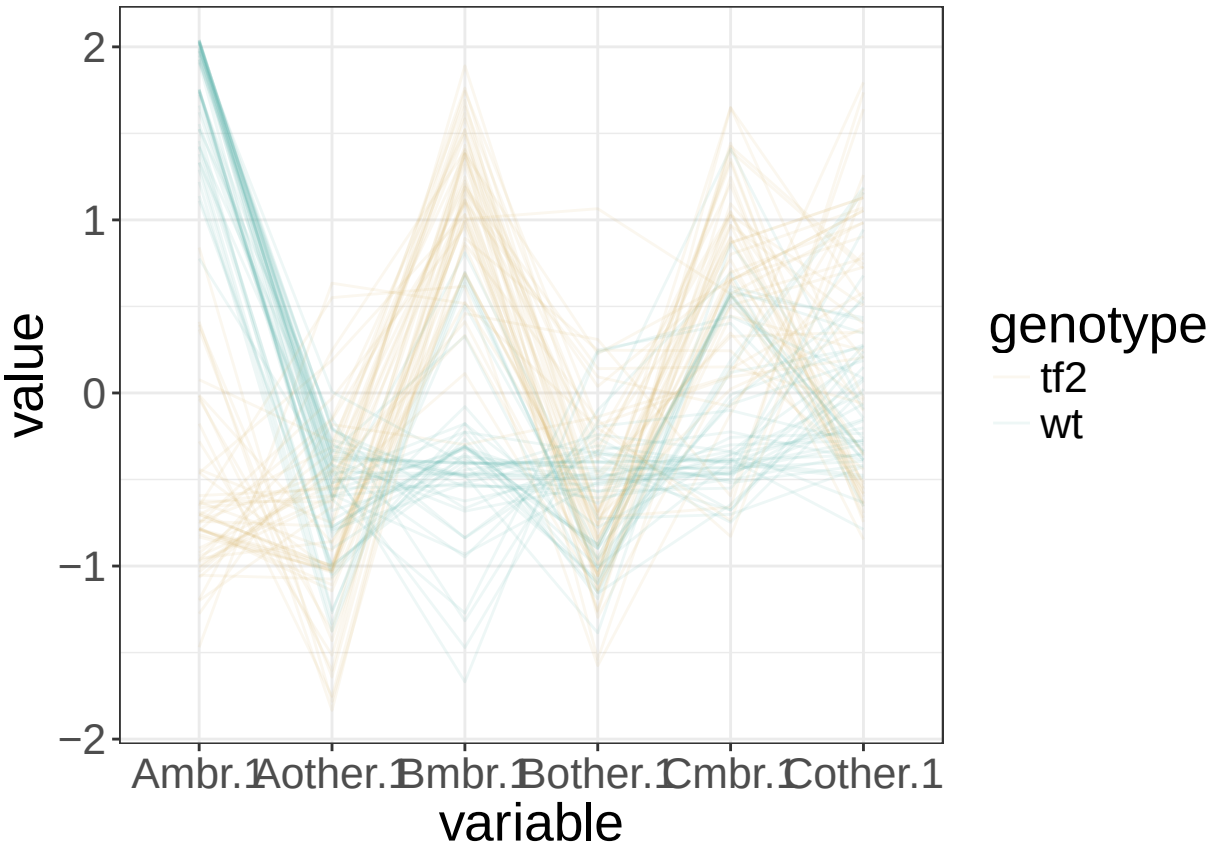

```
clusterVis_region_ssom(32)
```

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



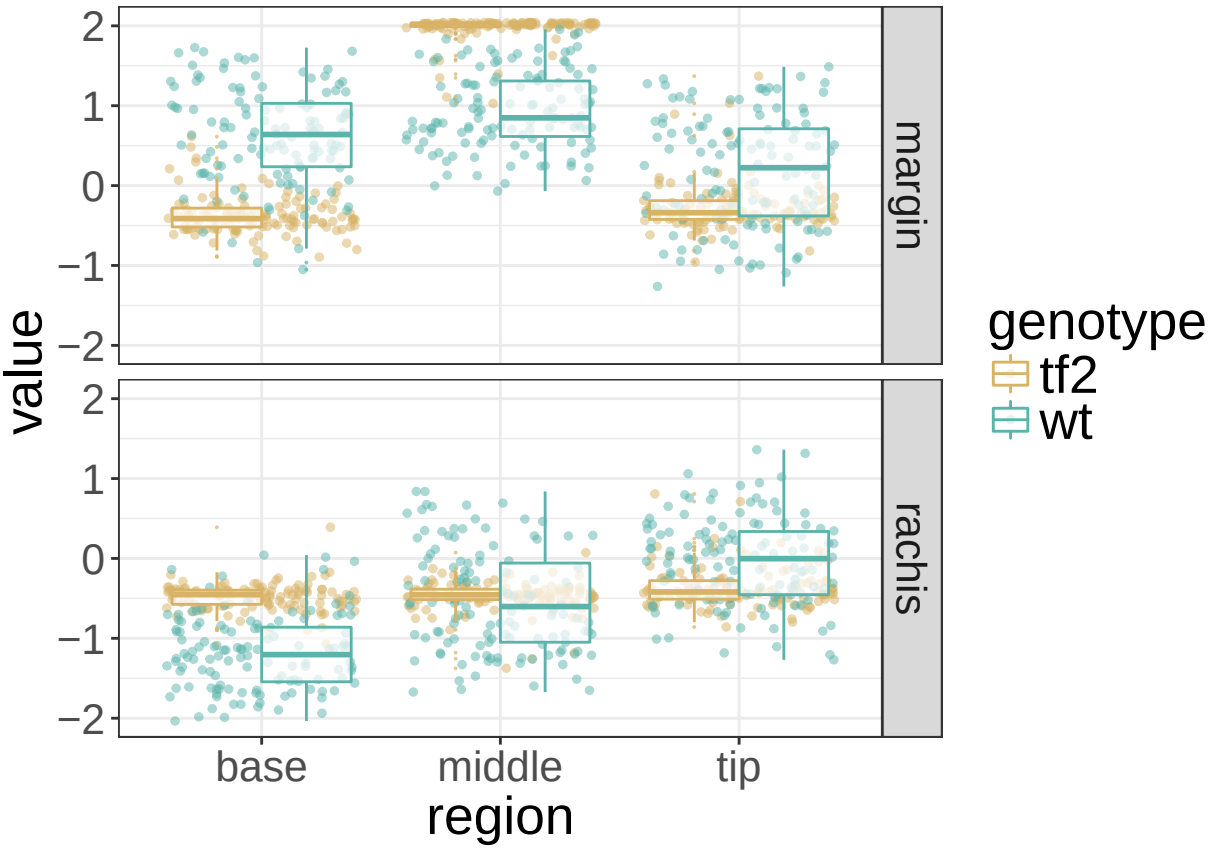

```
clusterVis_line_ssom(32)
```

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



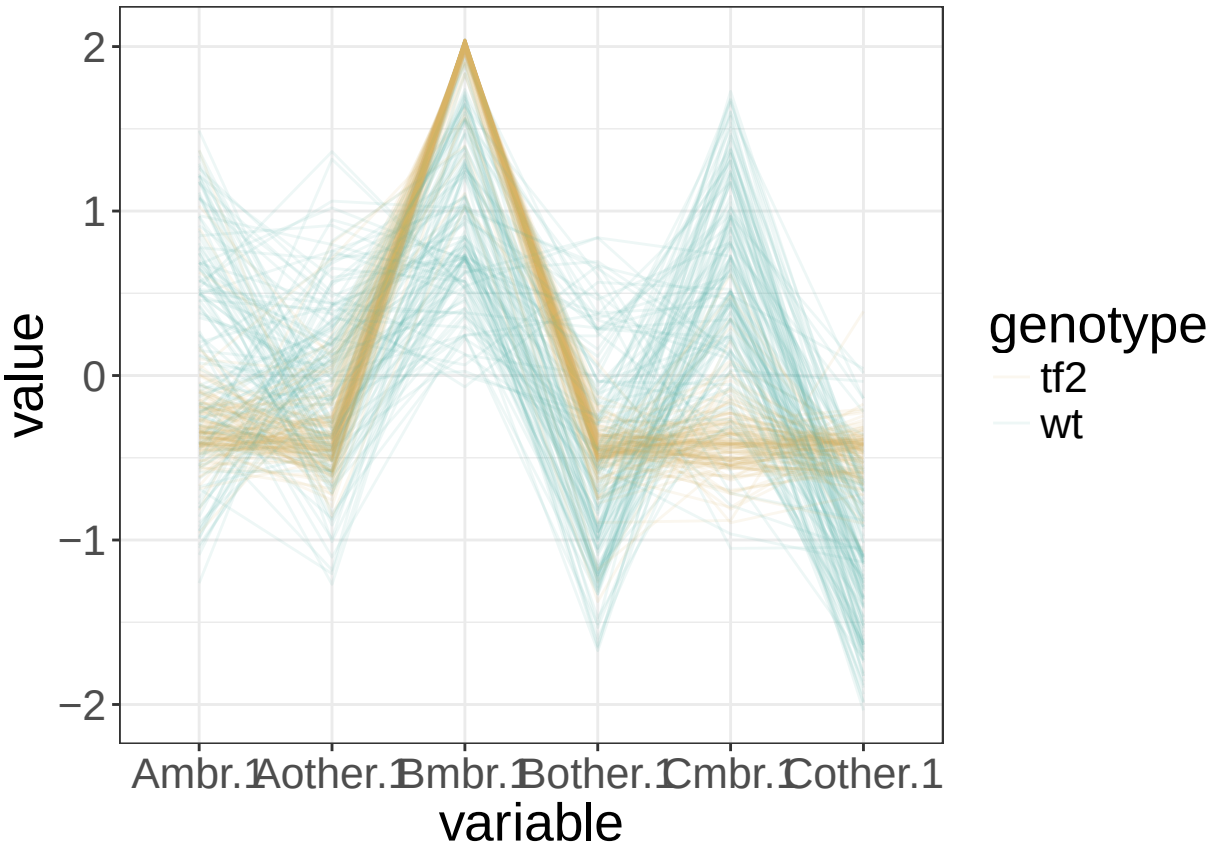

```
clusterVis_region_ssom(33)
```

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



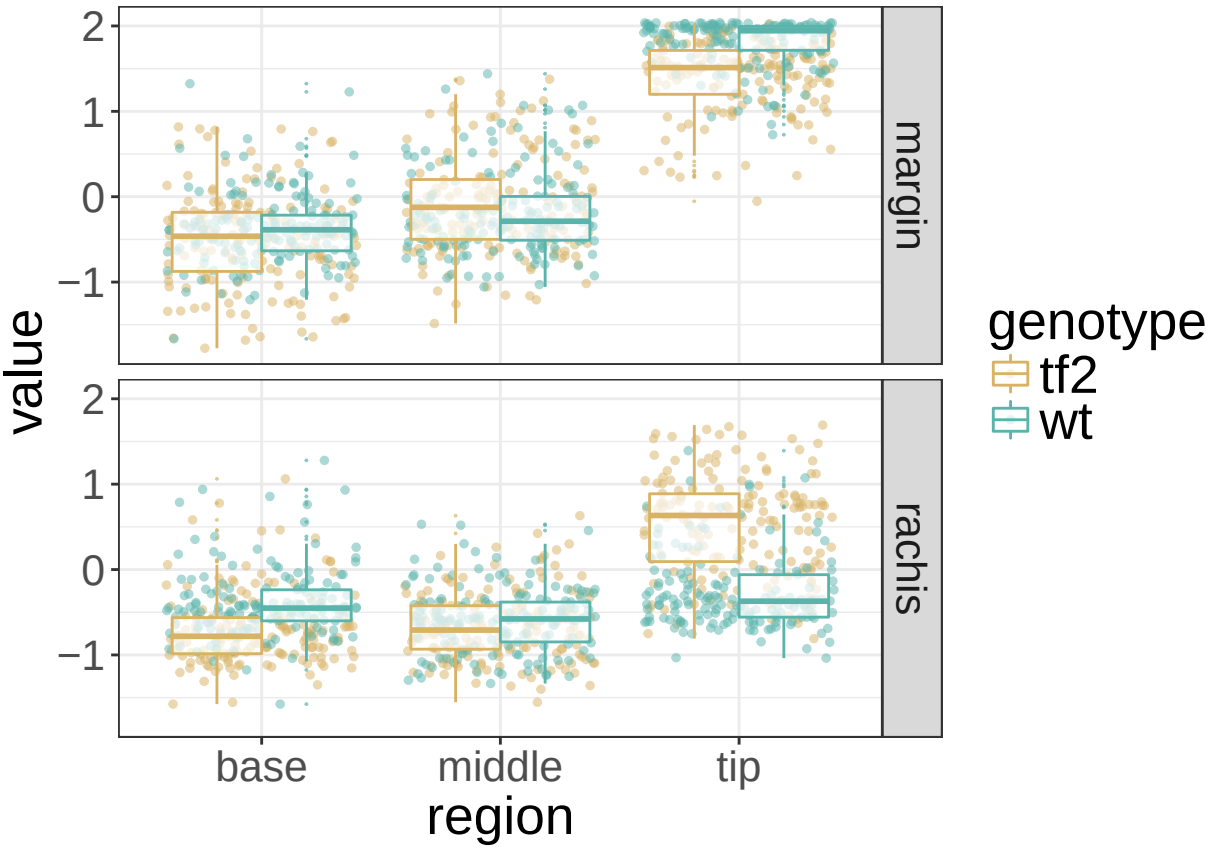

```
clusterVis_line_ssom(33)
```

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



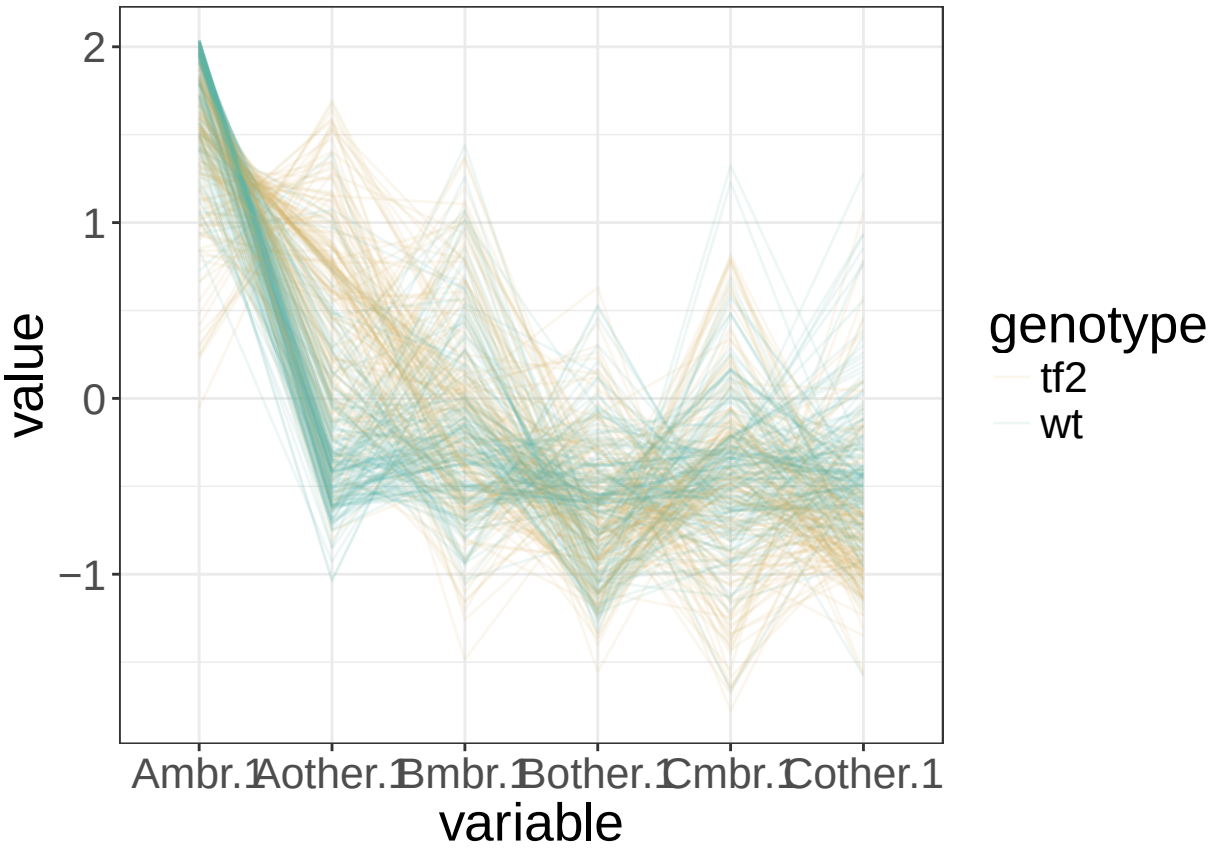

```
clusterVis_region_ssom(34)
```

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



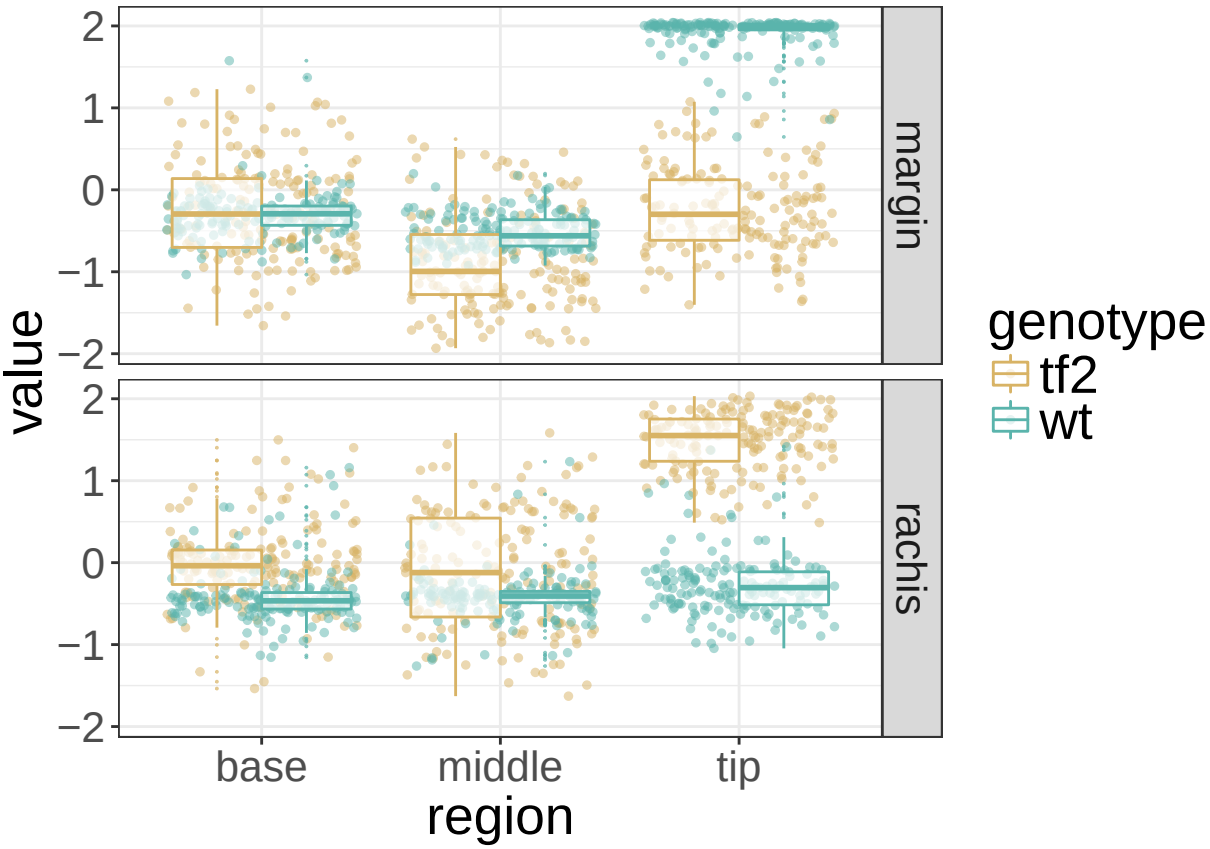

```
clusterVis_line_ssom(34)
```

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



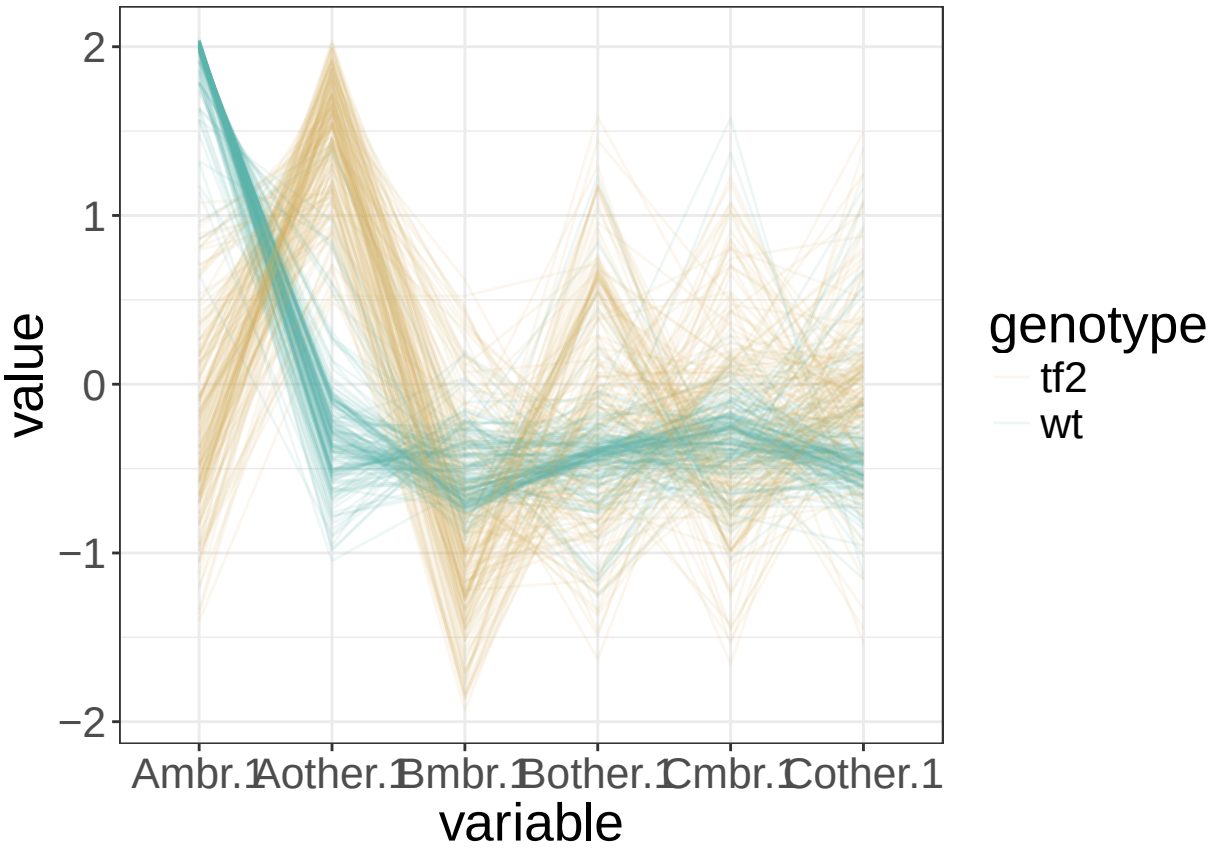

```
clusterVis_region_ssom(35)
```

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



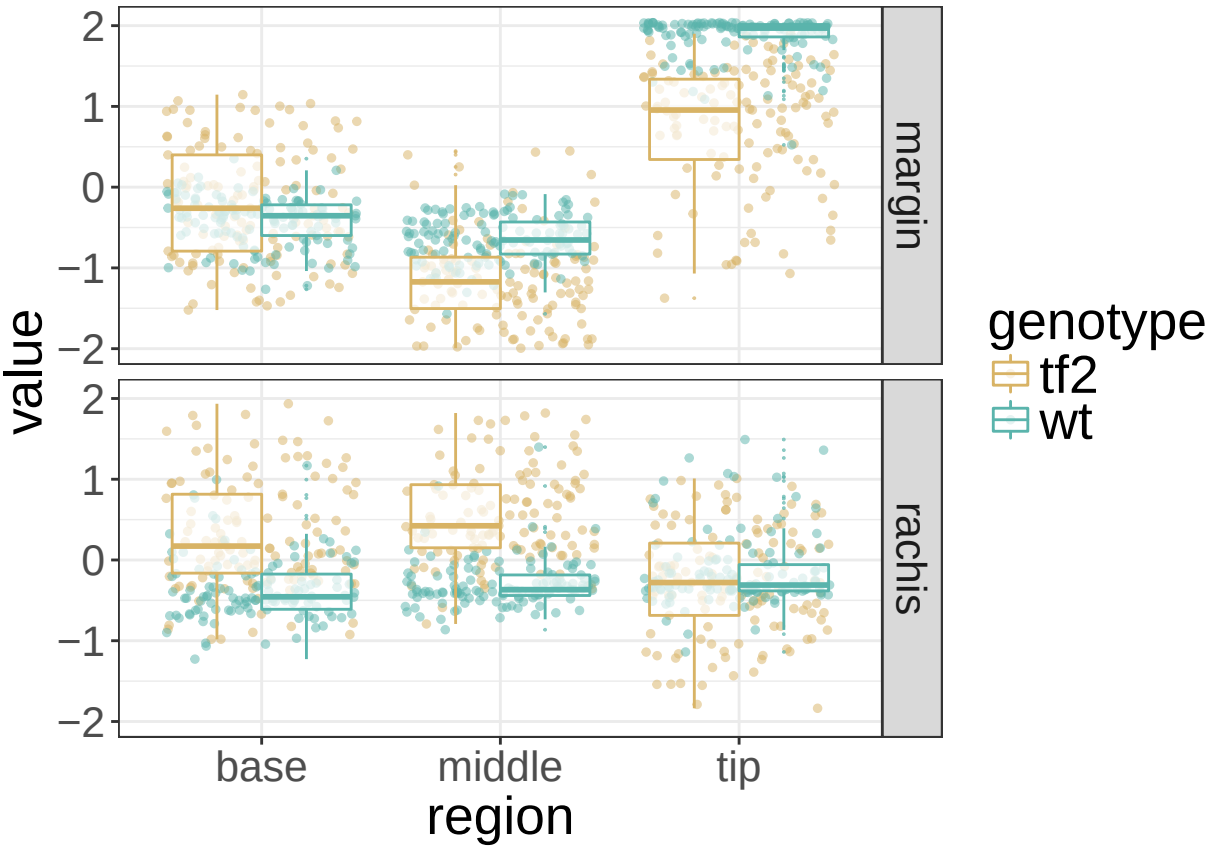

```
clusterVis_line_ssom(35)
```

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



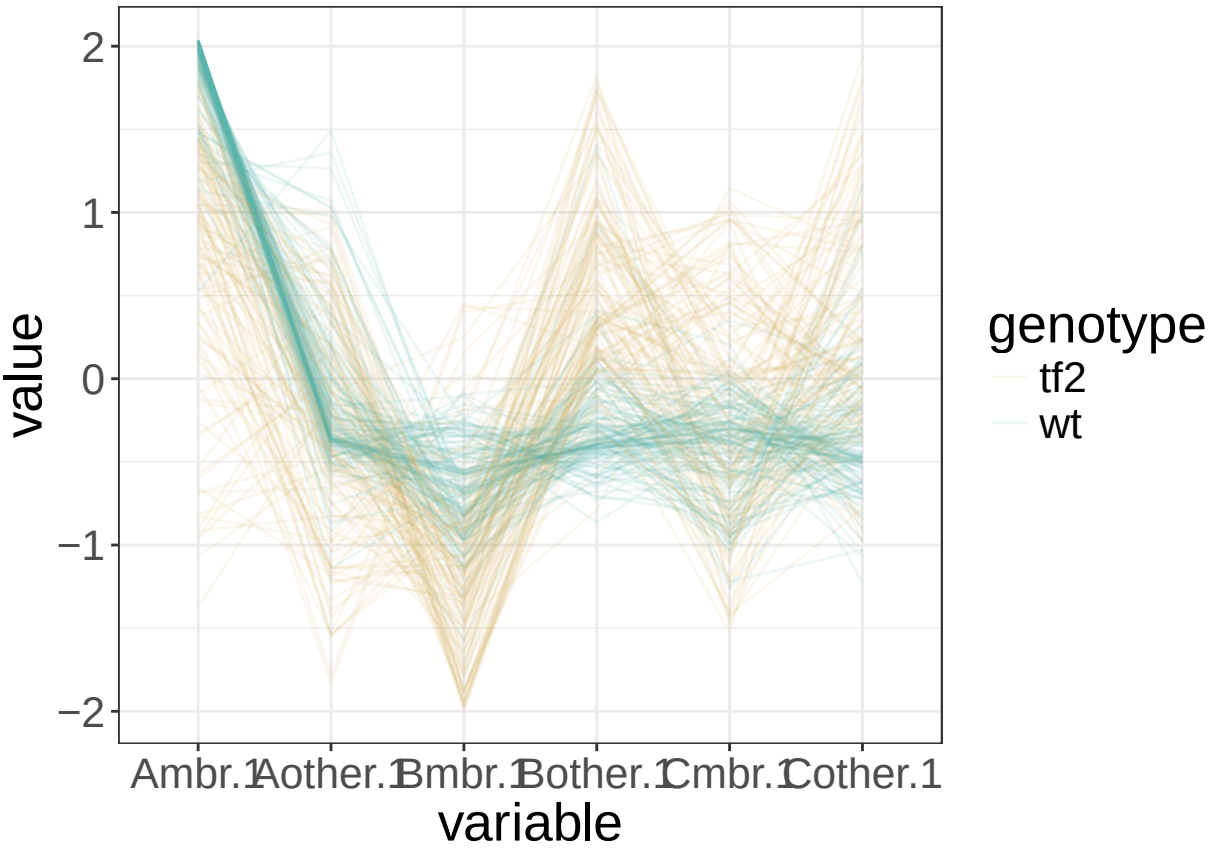

```
clusterVis_region_ssom(36)
```

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




```
clusterVis_line_ssom(36)
```

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


2
1
0
−1
−2
value
Ambr.1
Aother.1
Bmbr.1
Bother.1
Cmbr.1
Cother.1
variable
genotype
tf2
wt