Super SOM - Large

Purpose

The purpose of this analysis is to make a superSOM. The difference from lcmSOM-analysis_5c is that I am now playing with different SOM sizes.

```
library(ggplot2)
library(reshape)
library(plyr)
library(kohonen)
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"))

#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
```

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

Scaling the Data seperately

```
# 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

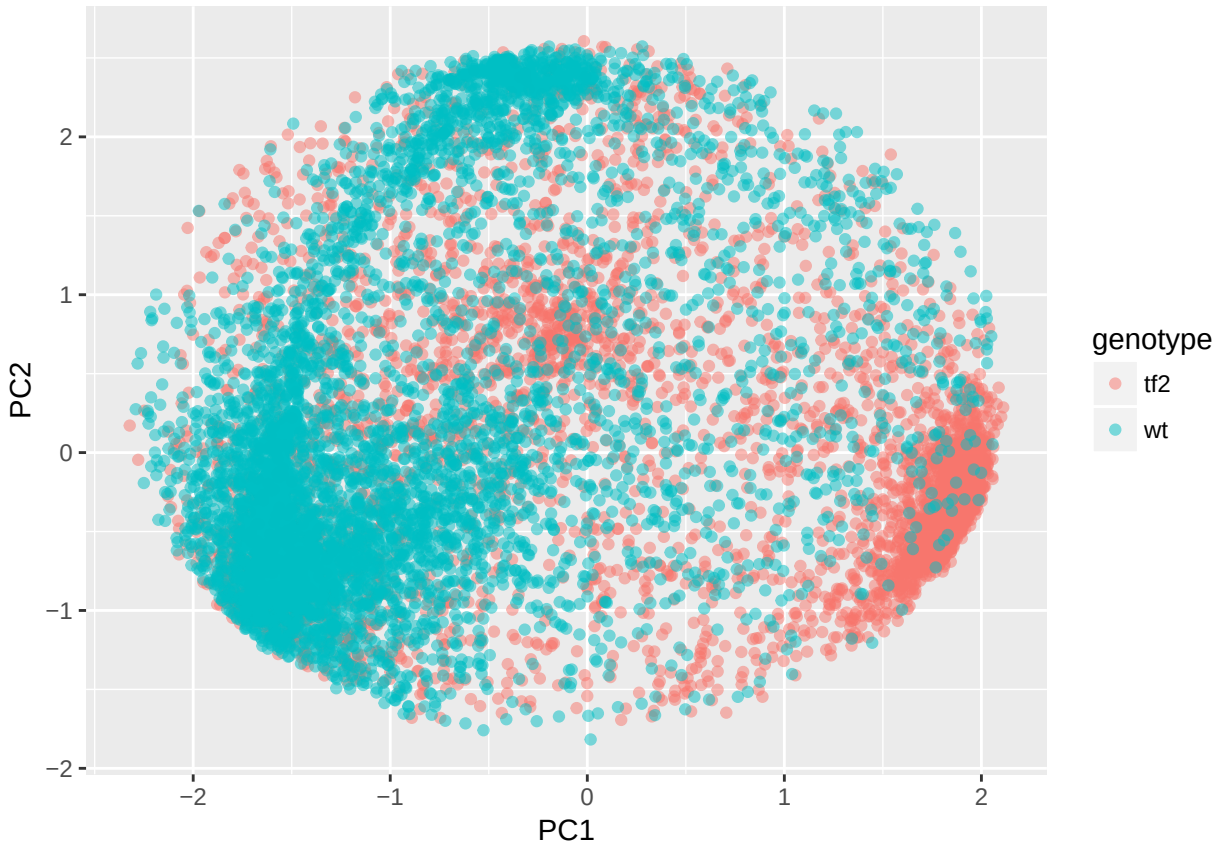
```
pca_sep <- prcomp(scale_data_sep)

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
```

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

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

Visualizing the PCA

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



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"
```

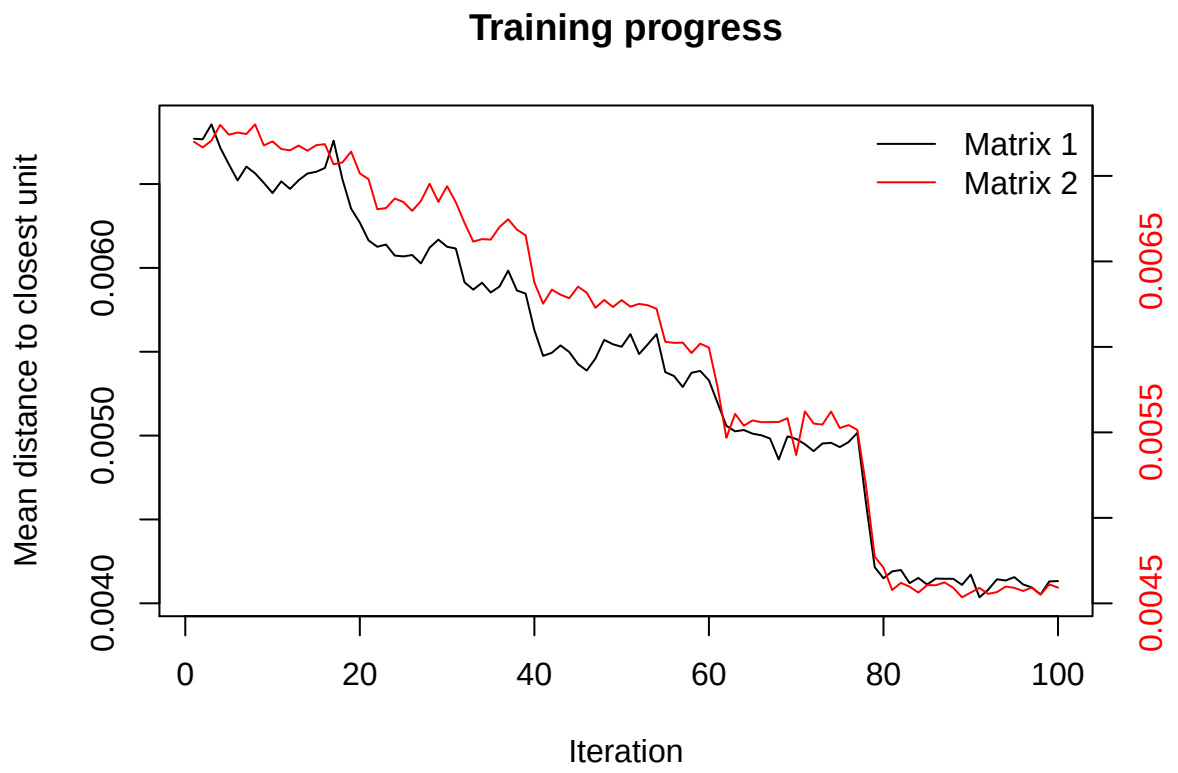
```
# head(data.val)

## 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

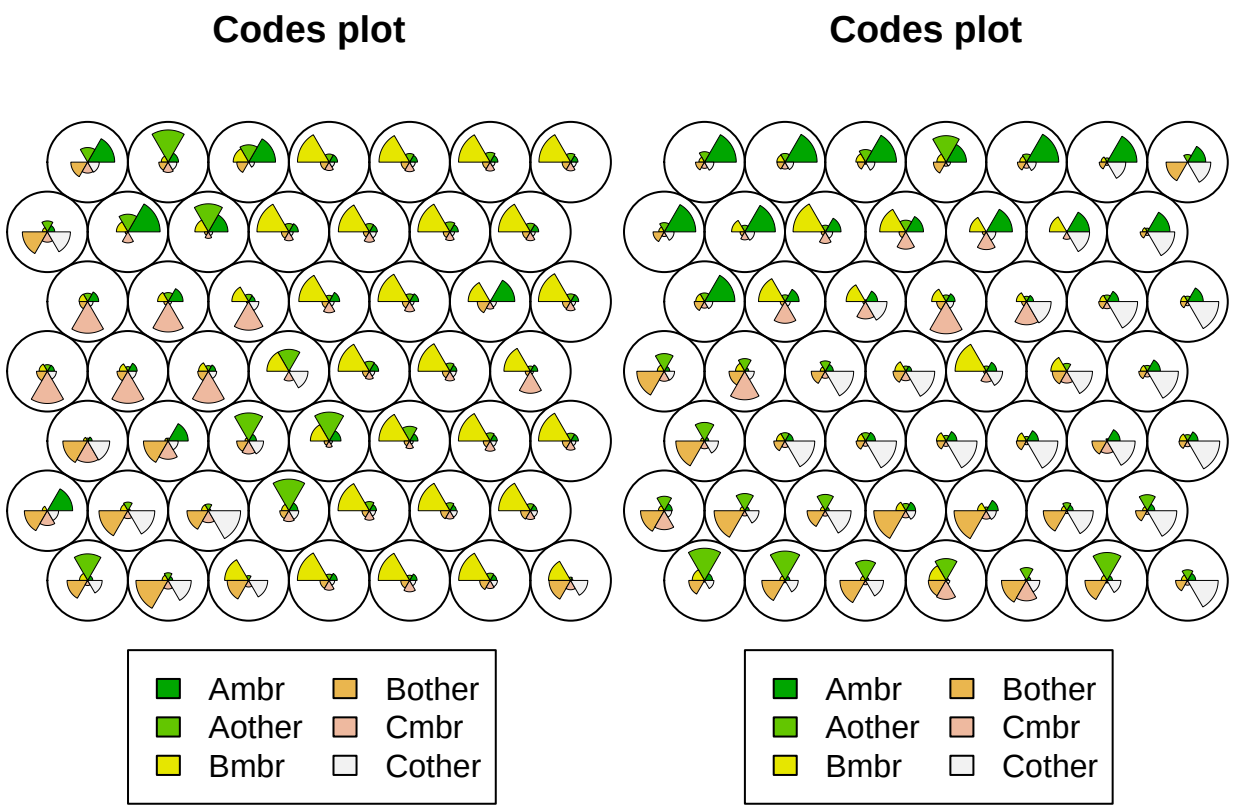
```r
## Making the SOM map
ssom <- supersom(all.data, somgrid(7, 7, "hexagonal"))

summary(ssom)
```

```
## SOM of size 7x7 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: 0.7344724
```

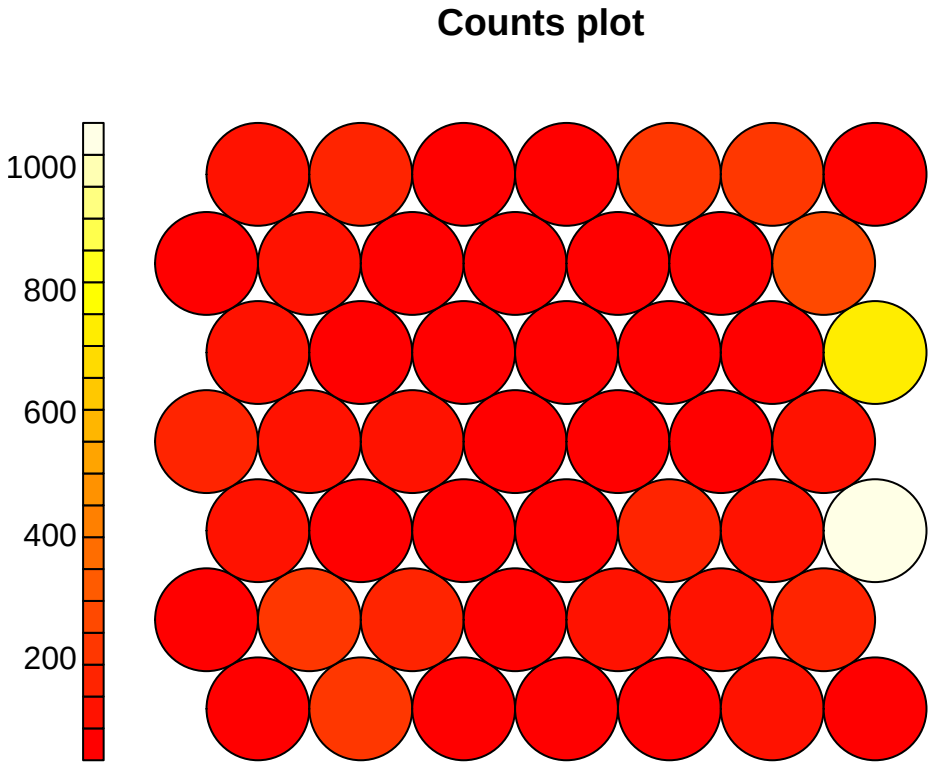
```r
#par(mfrow = c(3, 2))
plot(ssom, type = "changes")
```



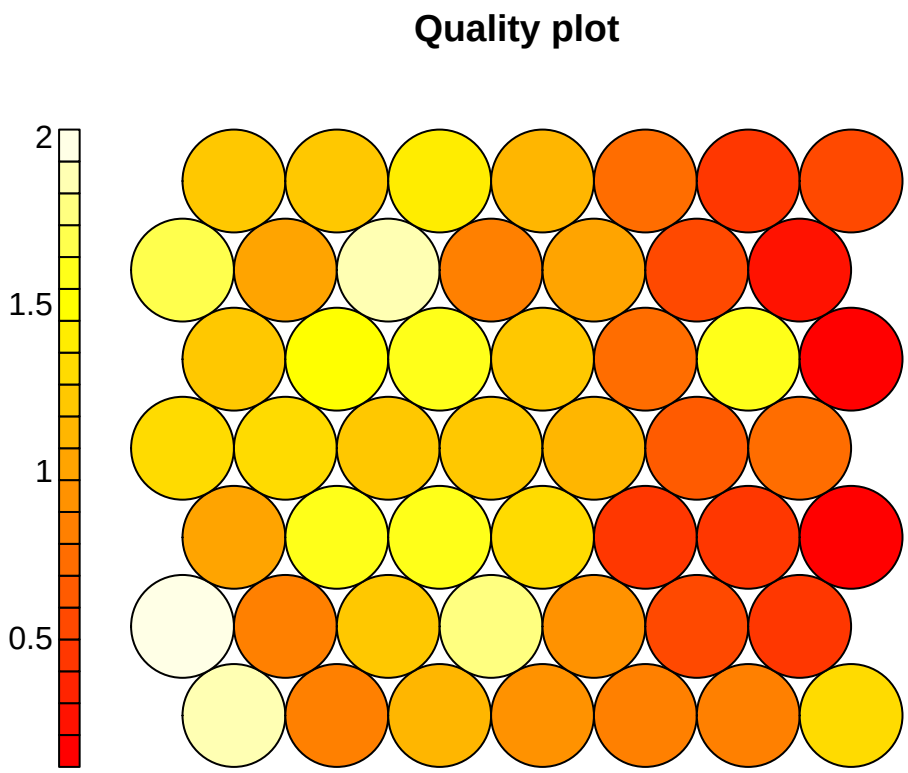

```r
plot(ssom, type = "codes")
```




```
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.analysis5d_02Jan2017_larger.txt")
```

Visualization

```
## Read in Data from previous section
plot.data <- read.table("../data/output/ssom.data.analysis5d_02Jan2017_larger.txt", header = TRUE)
names(plot.data)
```

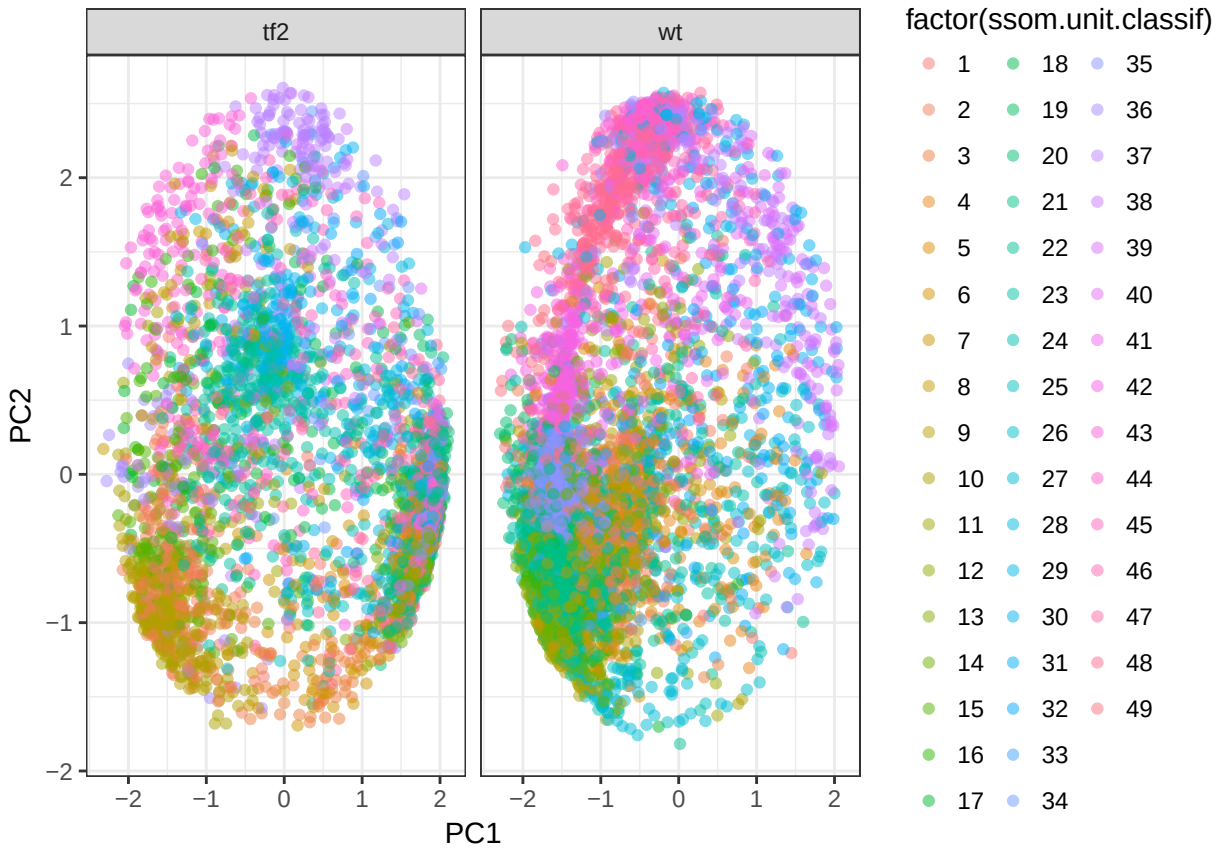
```
##  [1] "genotype"          "gene"              "Ambr"
##  [4] "Aother"            "Bmbr"              "Bother"
##  [7] "Cmbr"              "Cother"            "Ambr.1"
## [10] "Aother.1"          "Bmbr.1"            "Bother.1"
## [13] "Cmbr.1"            "Cother.1"          "PC1"
## [16] "PC2"               "PC3"               "PC4"
## [19] "PC5"               "PC6"               "ssom.unit.classif"
## [22] "ssom.distances"
```

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

```
## [1] 13164    22
```

```
## 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

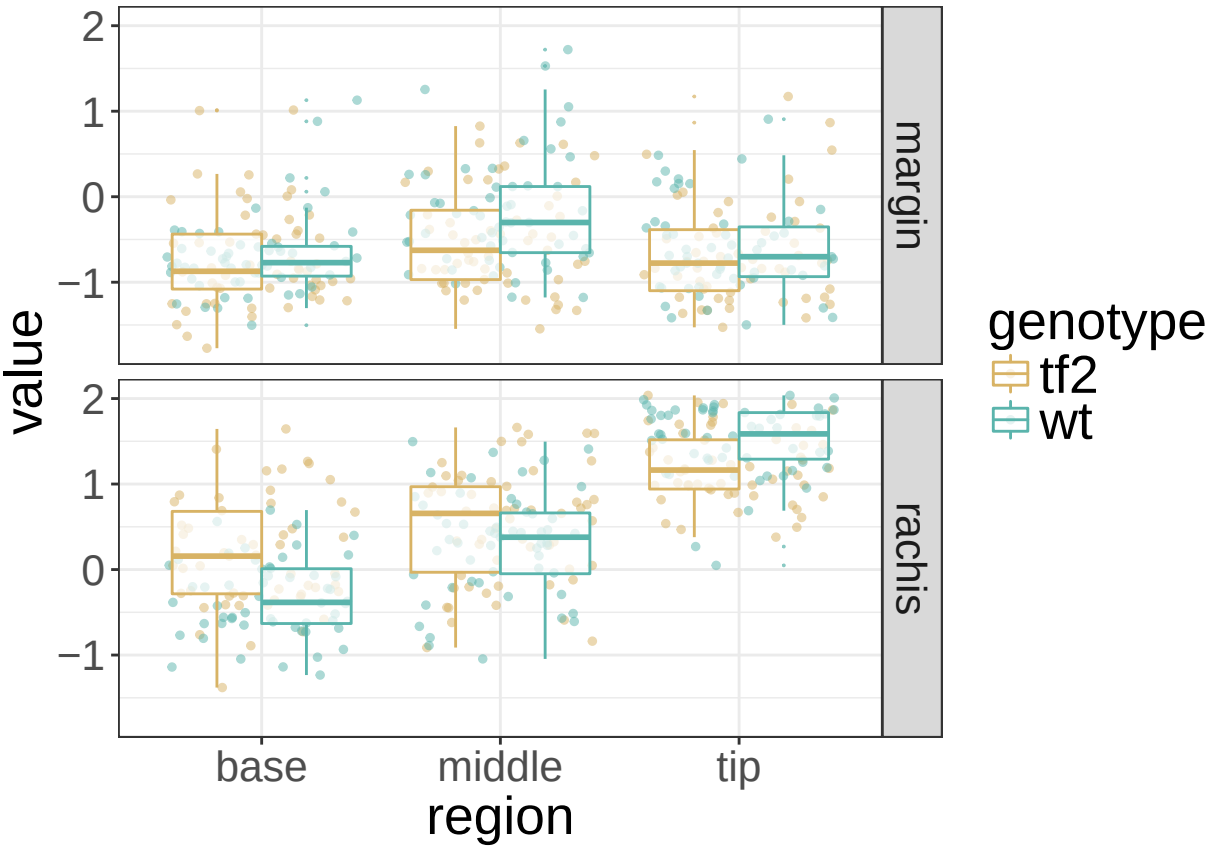
```
data.val2 <- read.table("../data/output/ssom.data.analysis5d_02Jan2017_larger.txt", header = TRUE)
```

Cluster 1

All the same.

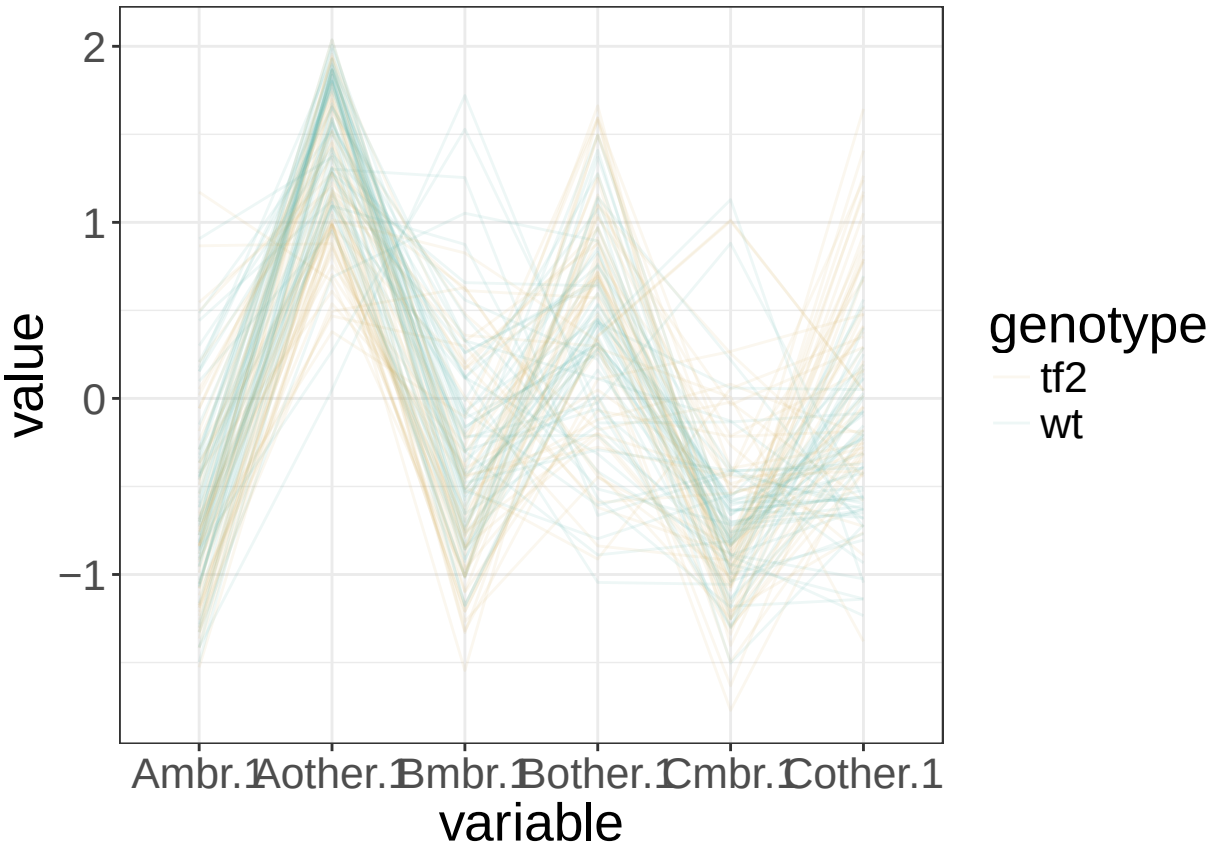
```
clusterVis_region_ssom(1)
```

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




```
clusterVis_line_ssom(1)
```

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



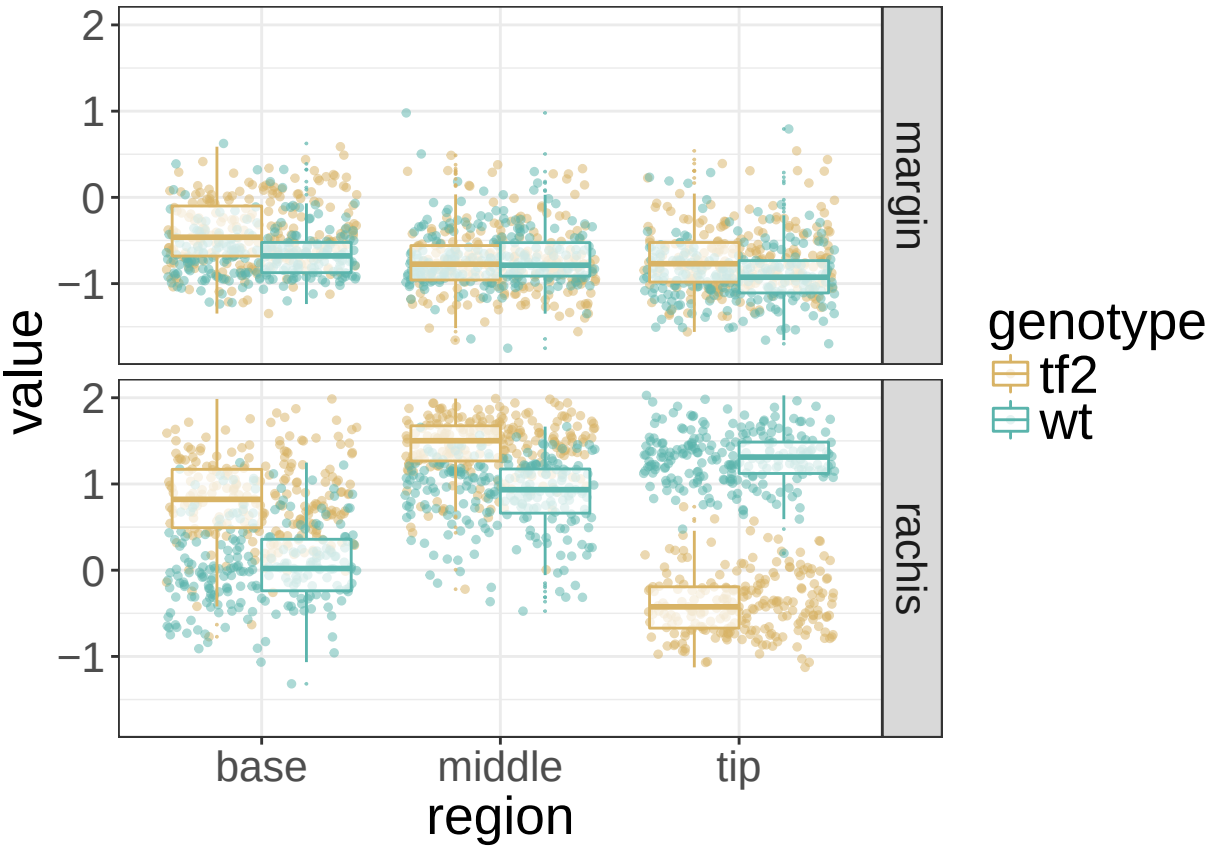

```
# genesInClust_ssom(1)
```

Cluster 2

Zig Zag across tissue in both genotypes. There are a lot of photosynthetic genes. In the tip wt gene expression up and wt down in base rachis. Do photosynthetic genes need to be down-regulated for leaflet inititation?

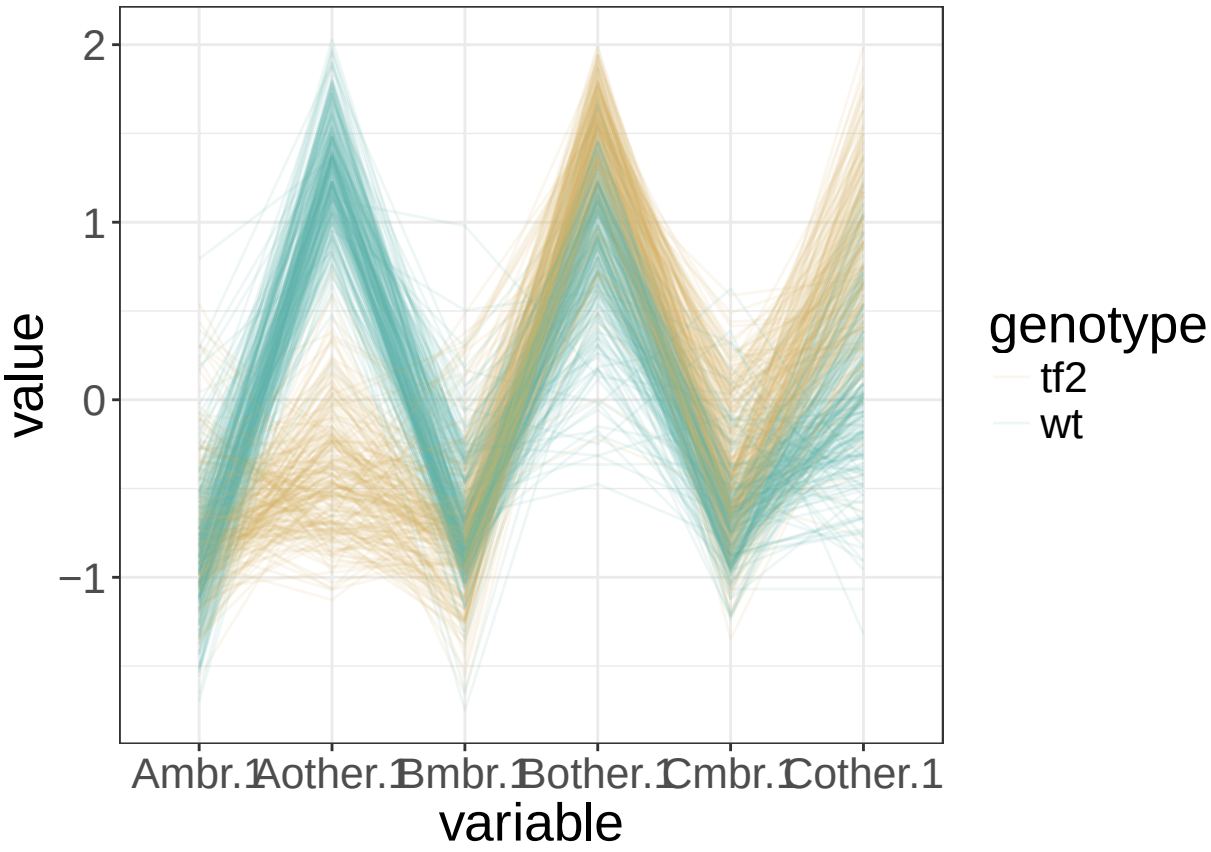
```
clusterVis_region_ssom(2)
```

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




```
clusterVis_line_ssom(2)
```

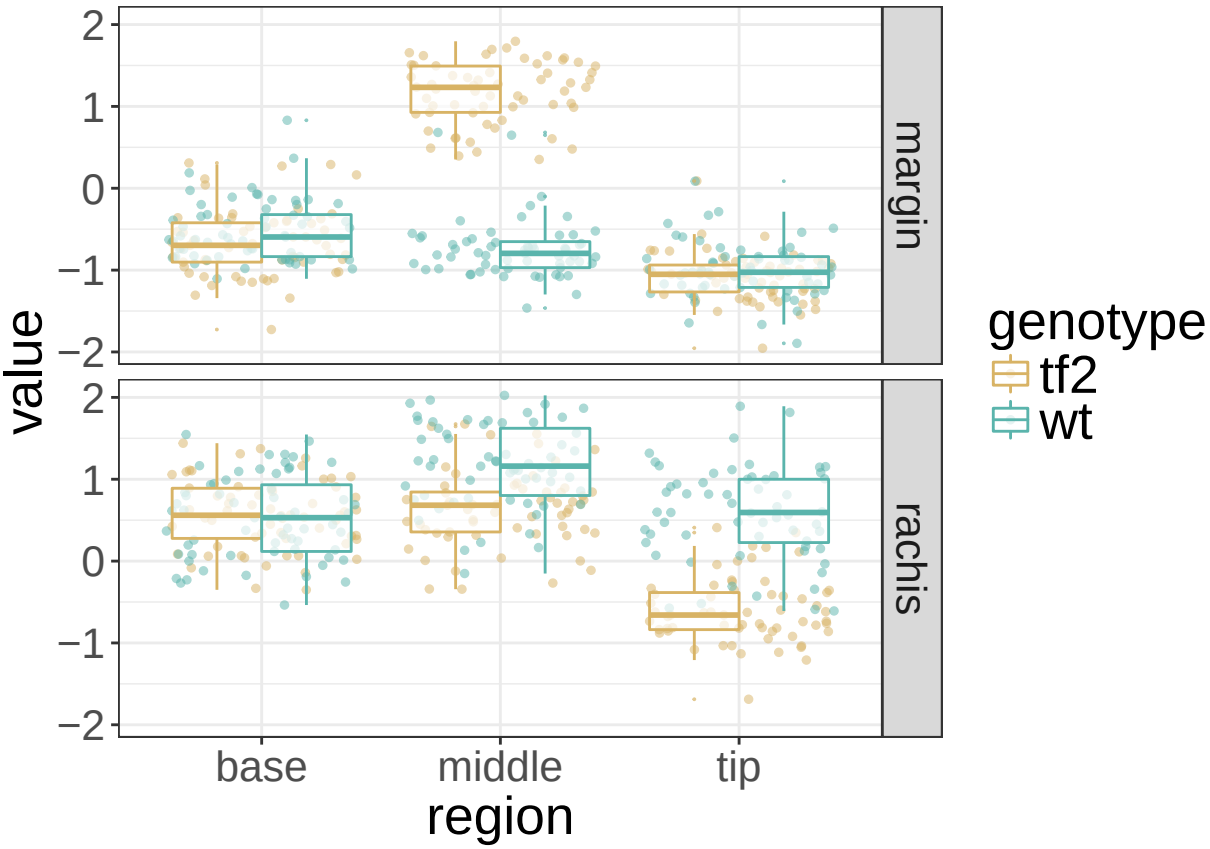
```
## Using genotype, gene as id variables
```




```
# genesInClust_ssom(2)
```

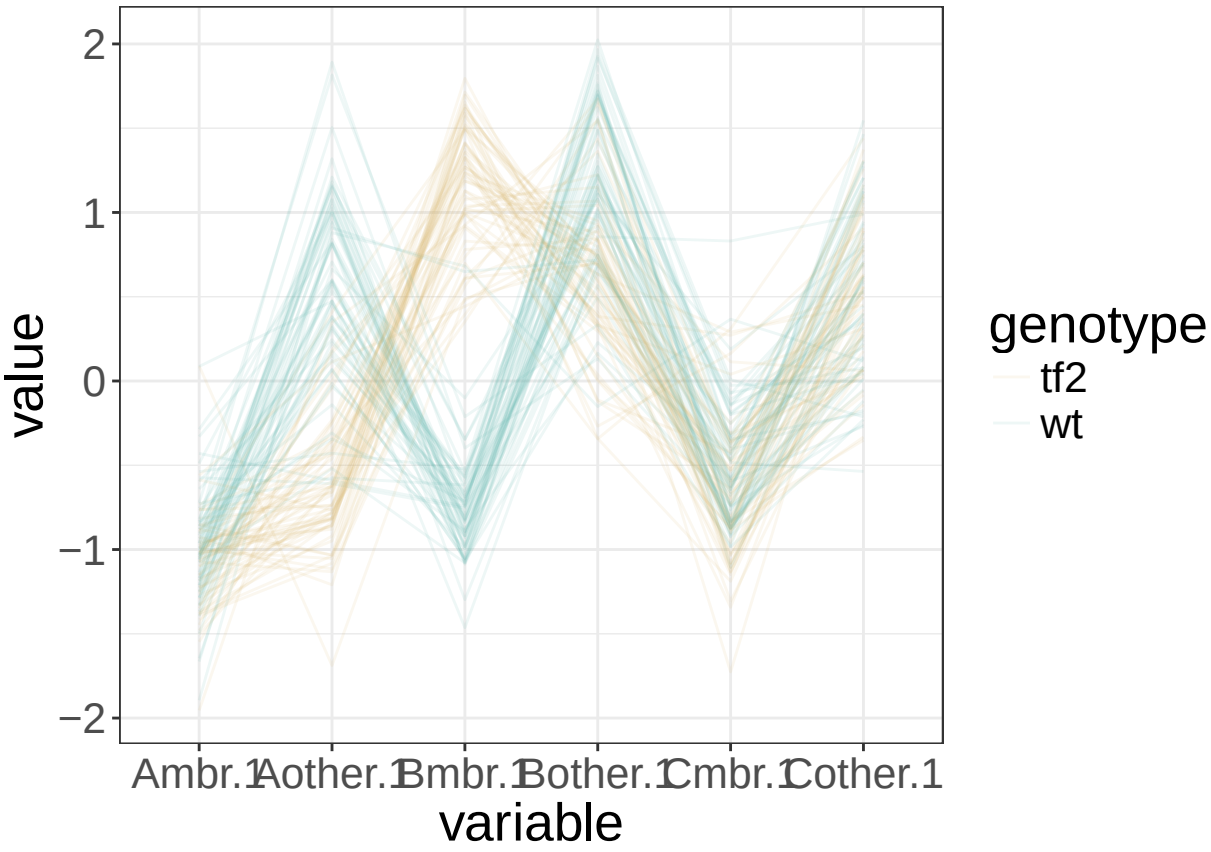
```
clusterVis_region_ssom(3)
```

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



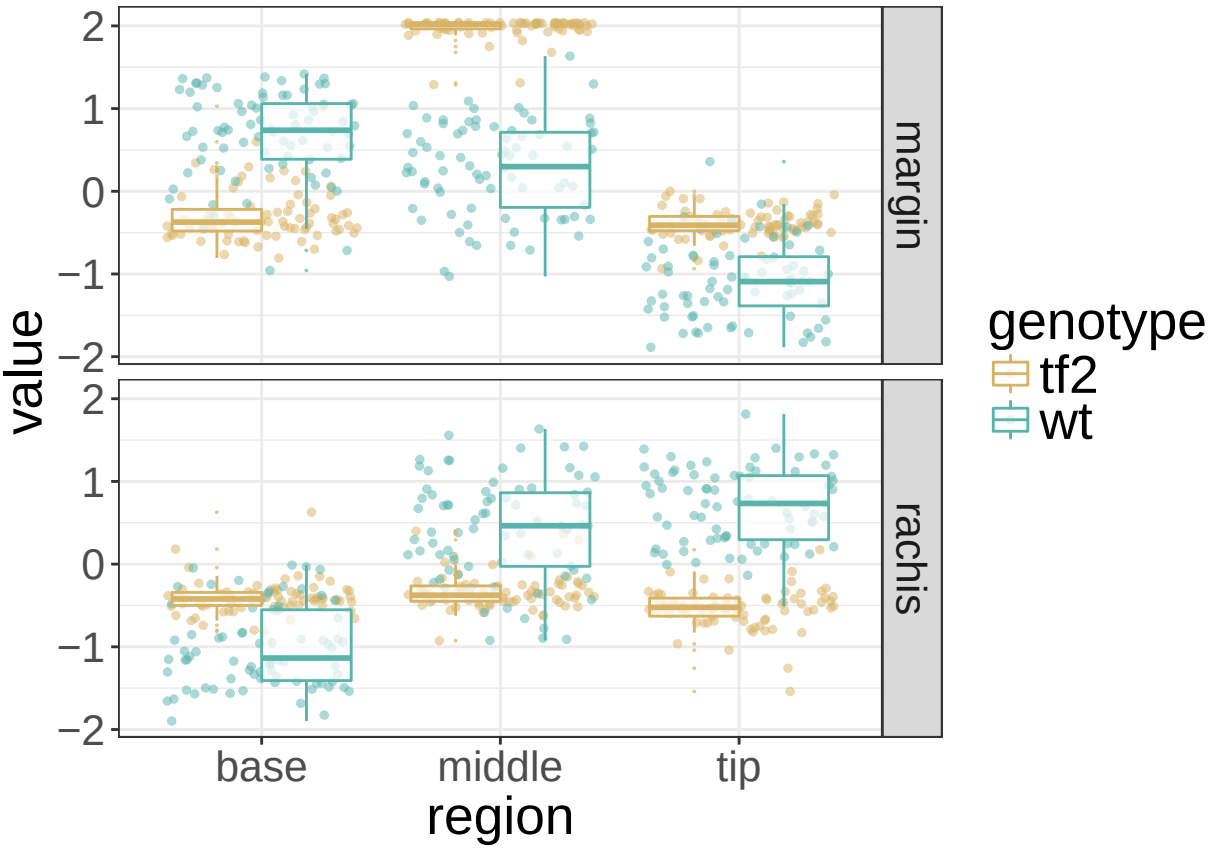

```
clusterVis_line_ssom(3)
```

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



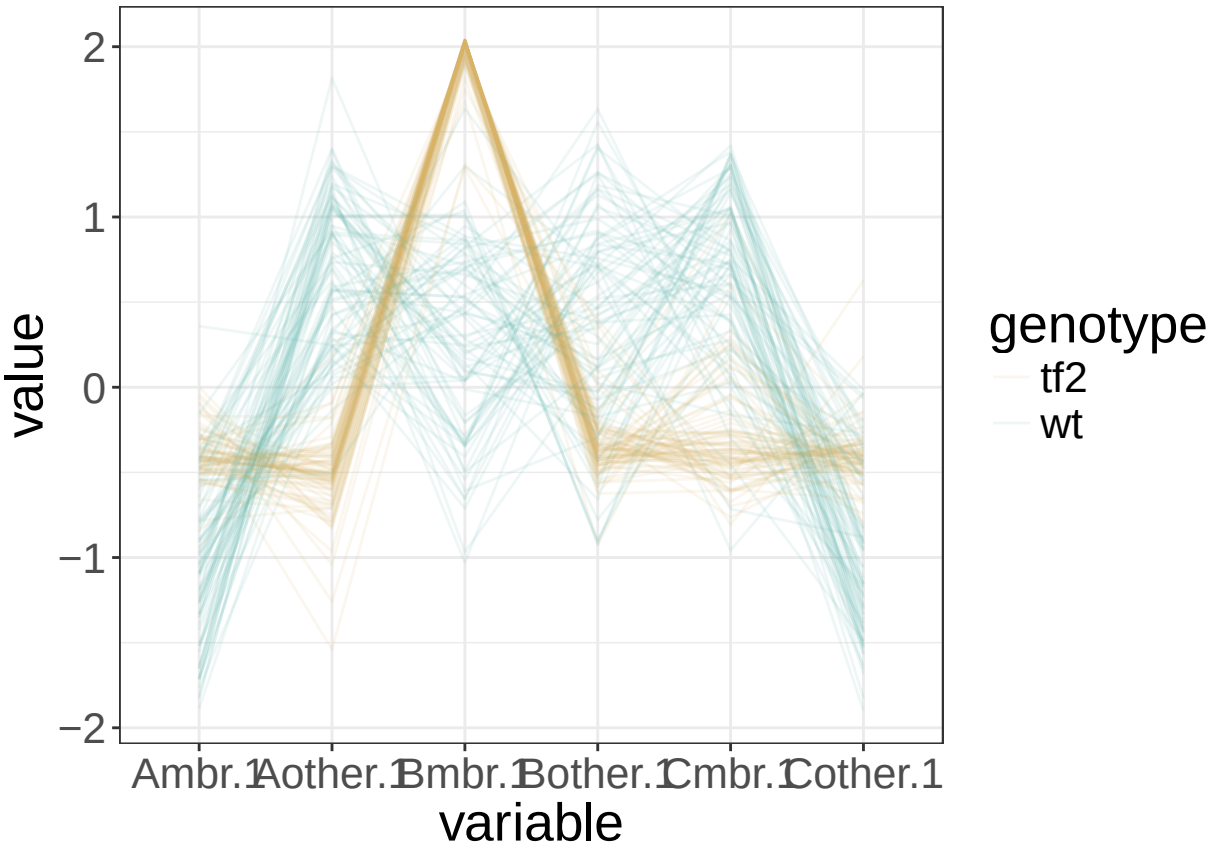

```
# genesInClust_ssom(3)


clusterVis_region_ssom(4)
```

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



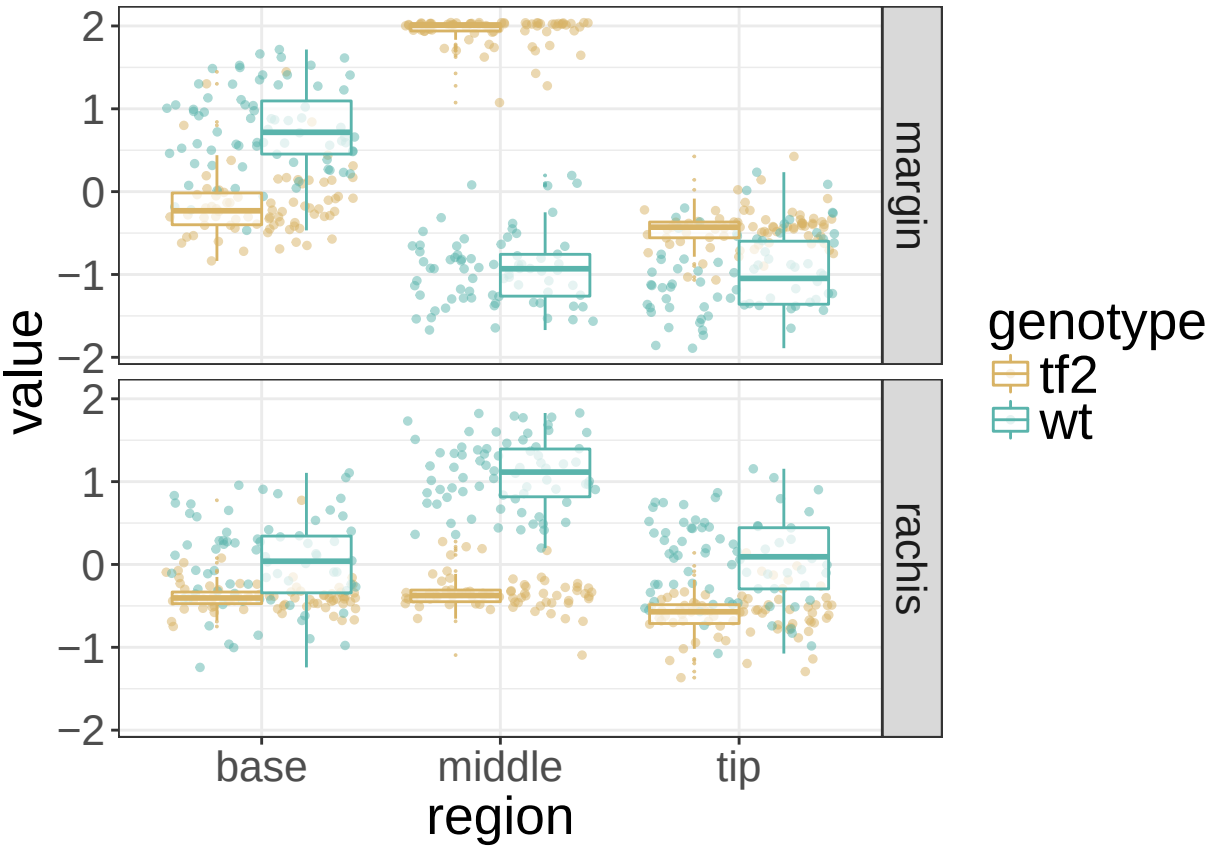

```
clusterVis_line_ssom(4)
```

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



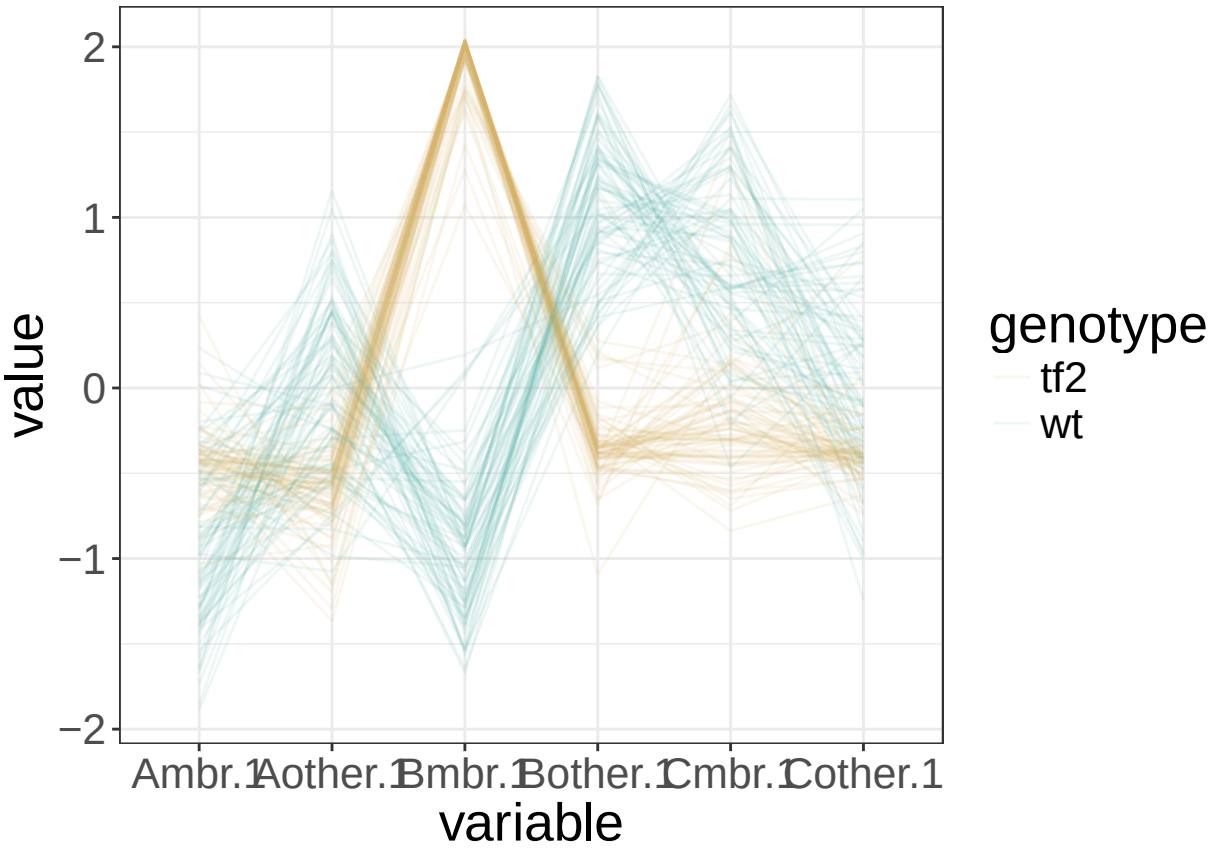

```
# genesInClust_ssom(5)

clusterVis_region_ssom(5)
```

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



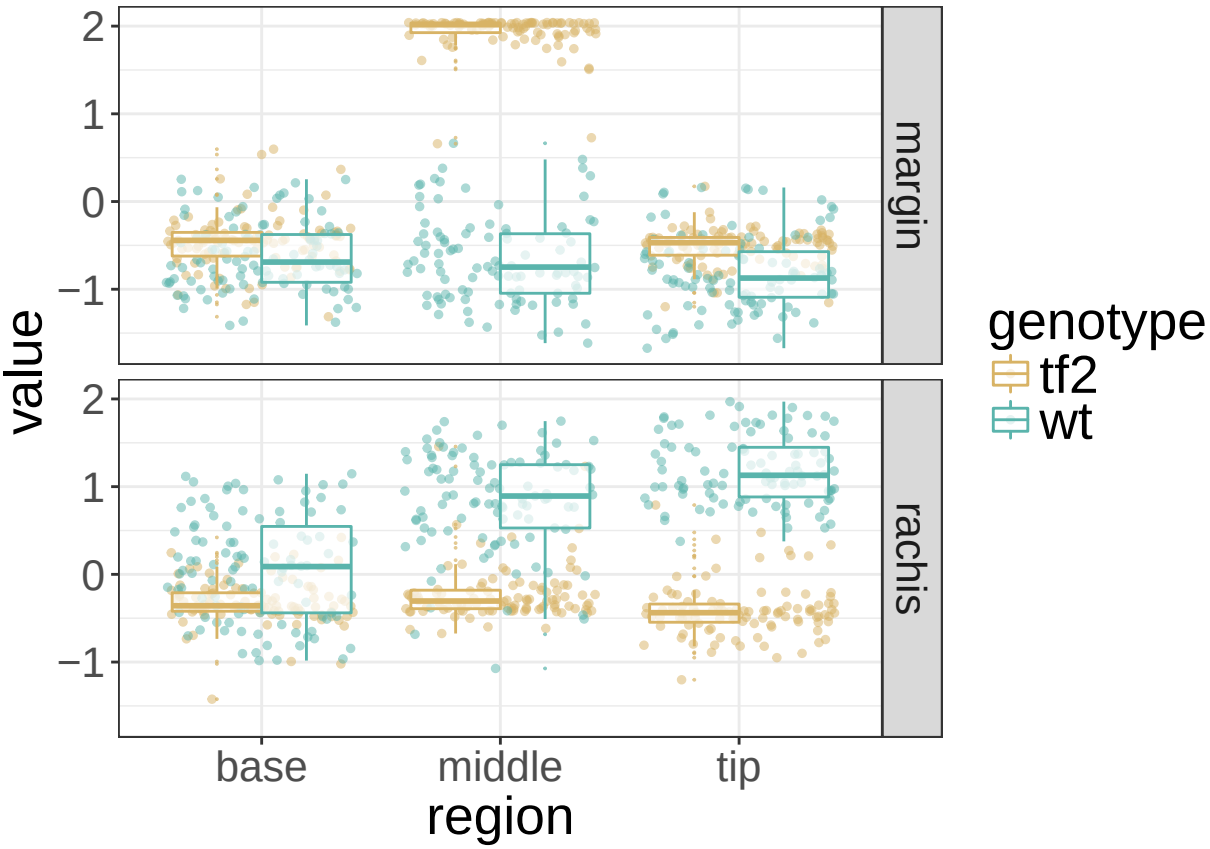

```
clusterVis_line_ssom(5)
```

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



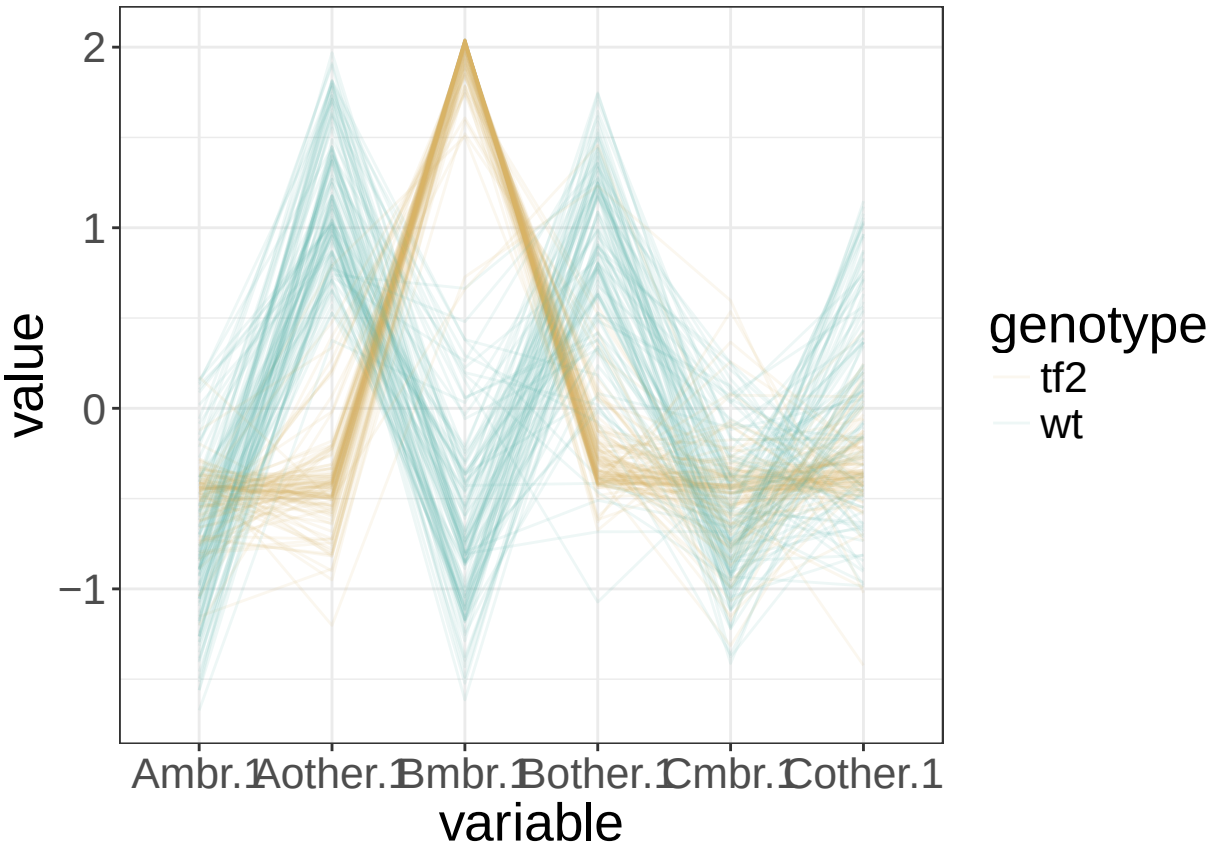

```
# genesInClust_ssom(5)

clusterVis_region_ssom(6)
```

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



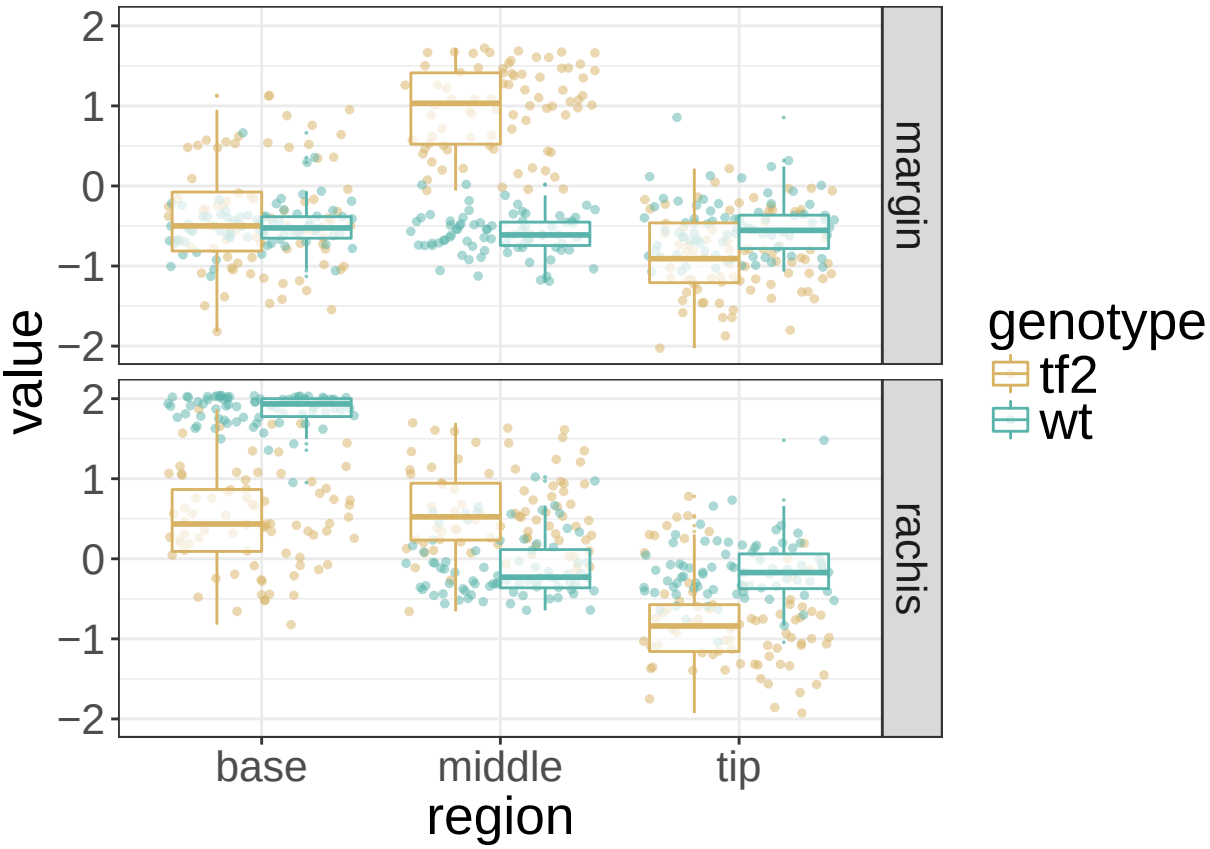

```
clusterVis_line_ssom(6)
```

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



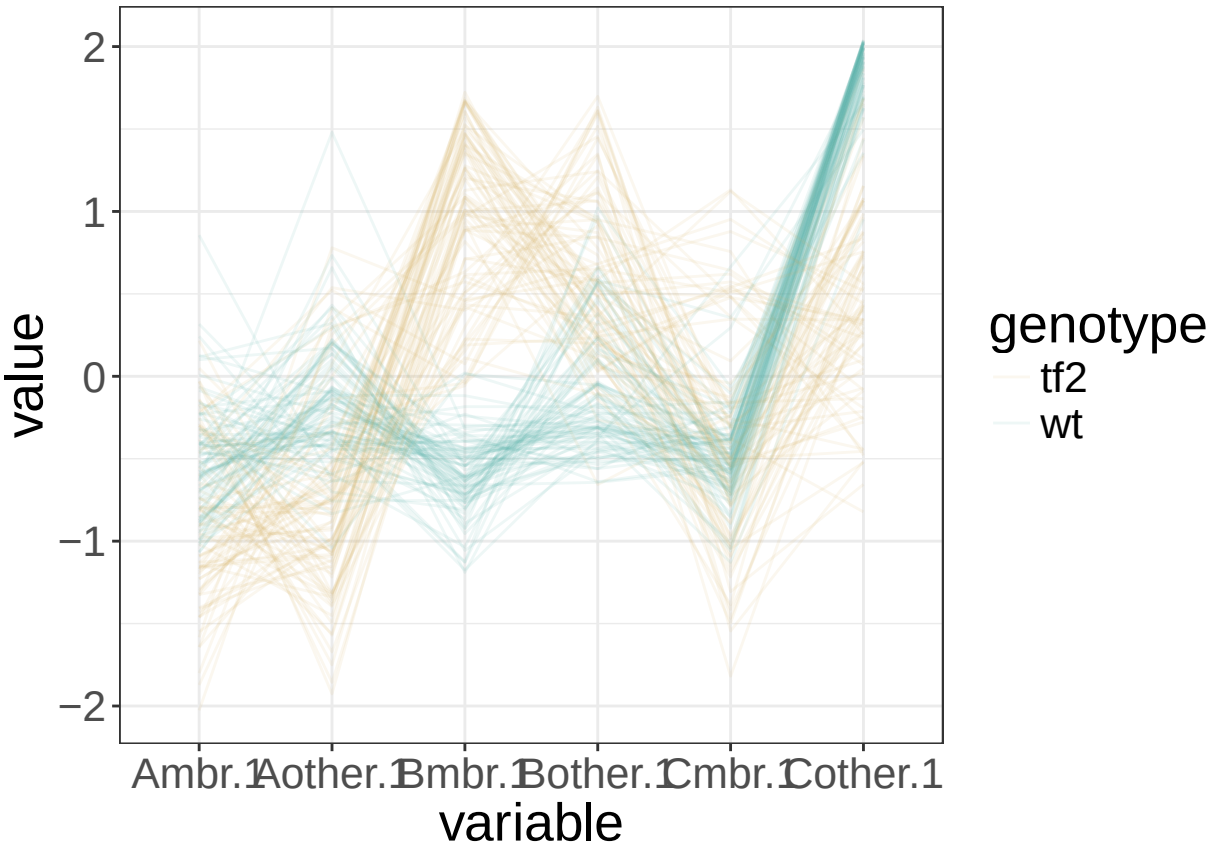

```
# genesInClust_ssom(6)

clusterVis_region_ssom(7)
```

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



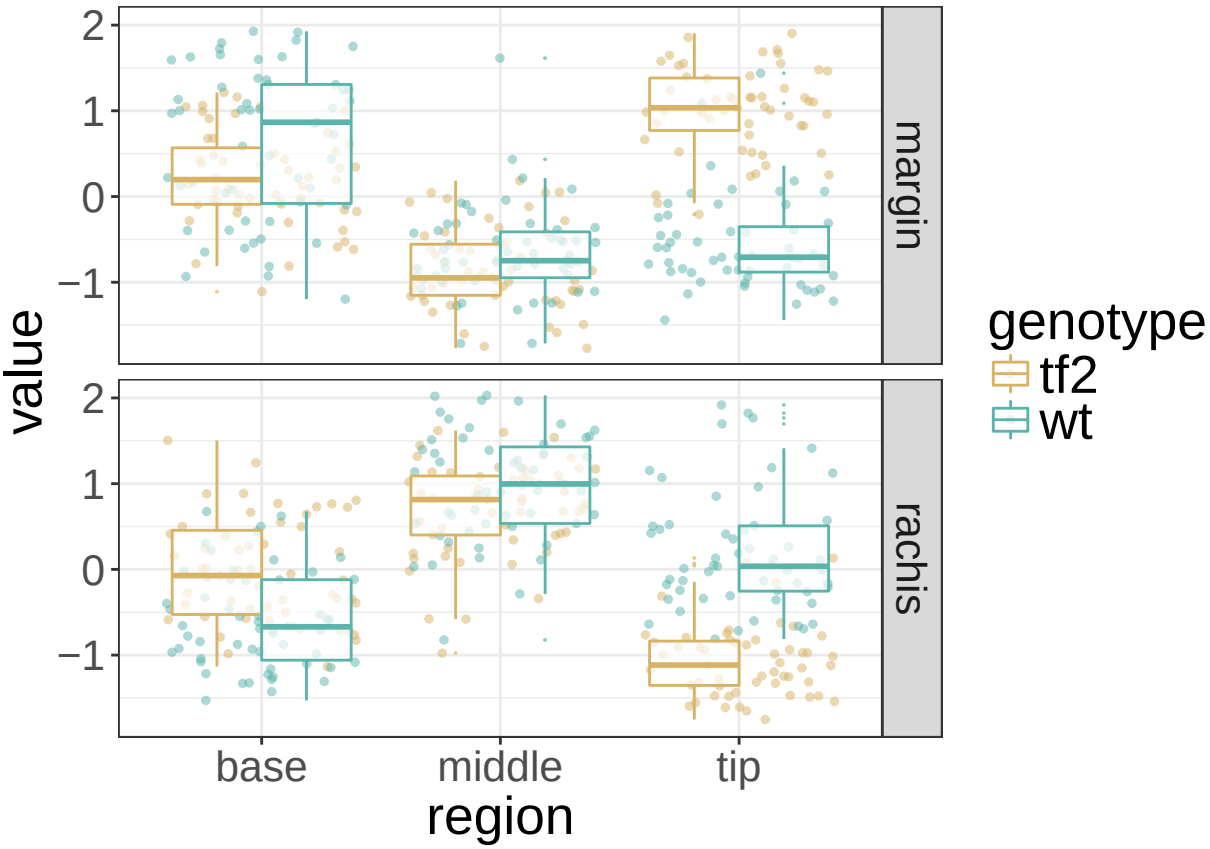

```
clusterVis_line_ssom(7)
```

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



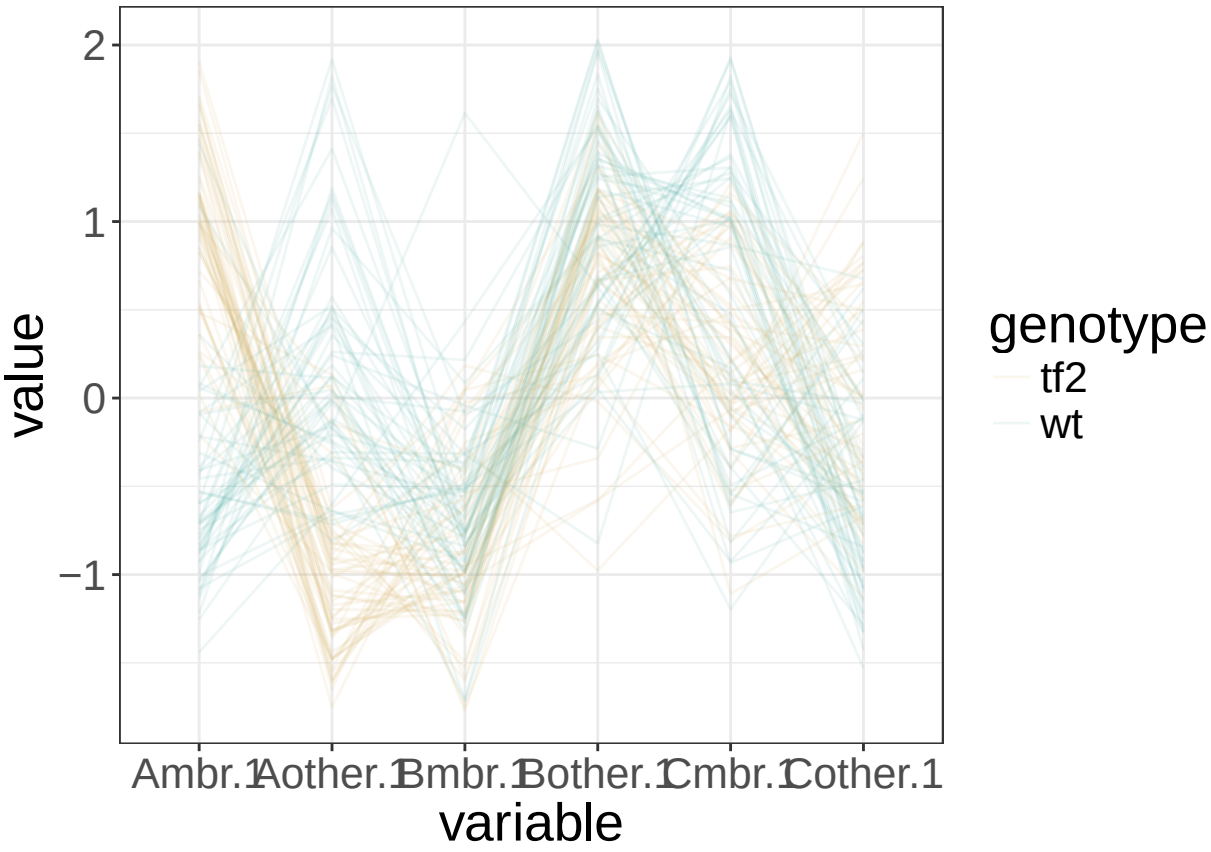

```
# genesInClust_ssom(7)

clusterVis_region_ssom(8)
```

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




```
clusterVis_line_ssom(8)
```

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




```
# genesInClust_ssom(8)
```

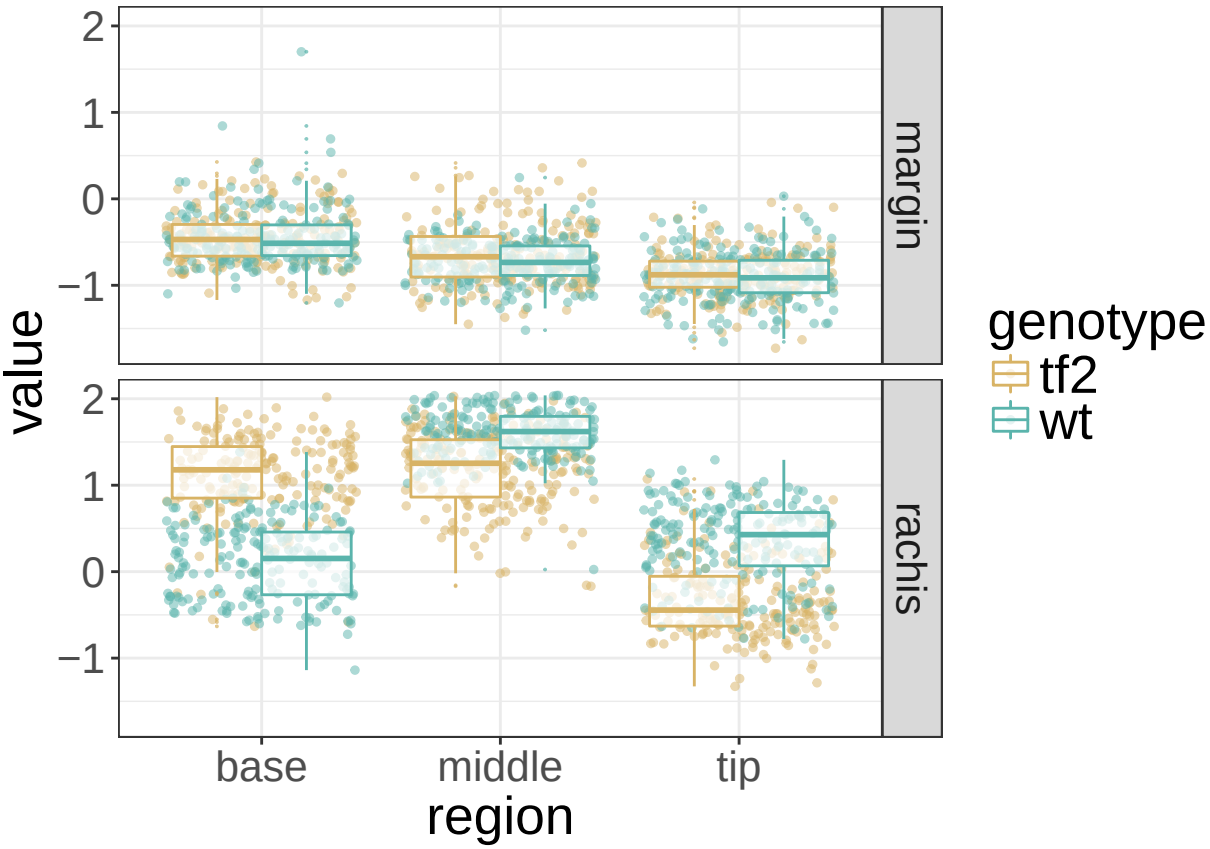
Cluster 9

Difference in base margin. Up regulation in tf2. Zig Zag.

PIN5 Membrane anchored cell wall Cylin F-Box CLAVATA HAT22 Homeodomain protein LOB Class III HD ZIP ATHB22 RDR6- RNA polymerase Cellulose synthase HDZIP I FIN219 - Auxin Induced Gene ARR17 -Response Regulator ATHB2 - HOmeobox AGO10- Translation initiation factor. Required to establish the central-peripheral organization of the embryo apex. Along with WUS and CLV genes, controls the relative organization of central zone and peripheral zone cells in meristems. F-Box Protein ARF4 MATE Family ?

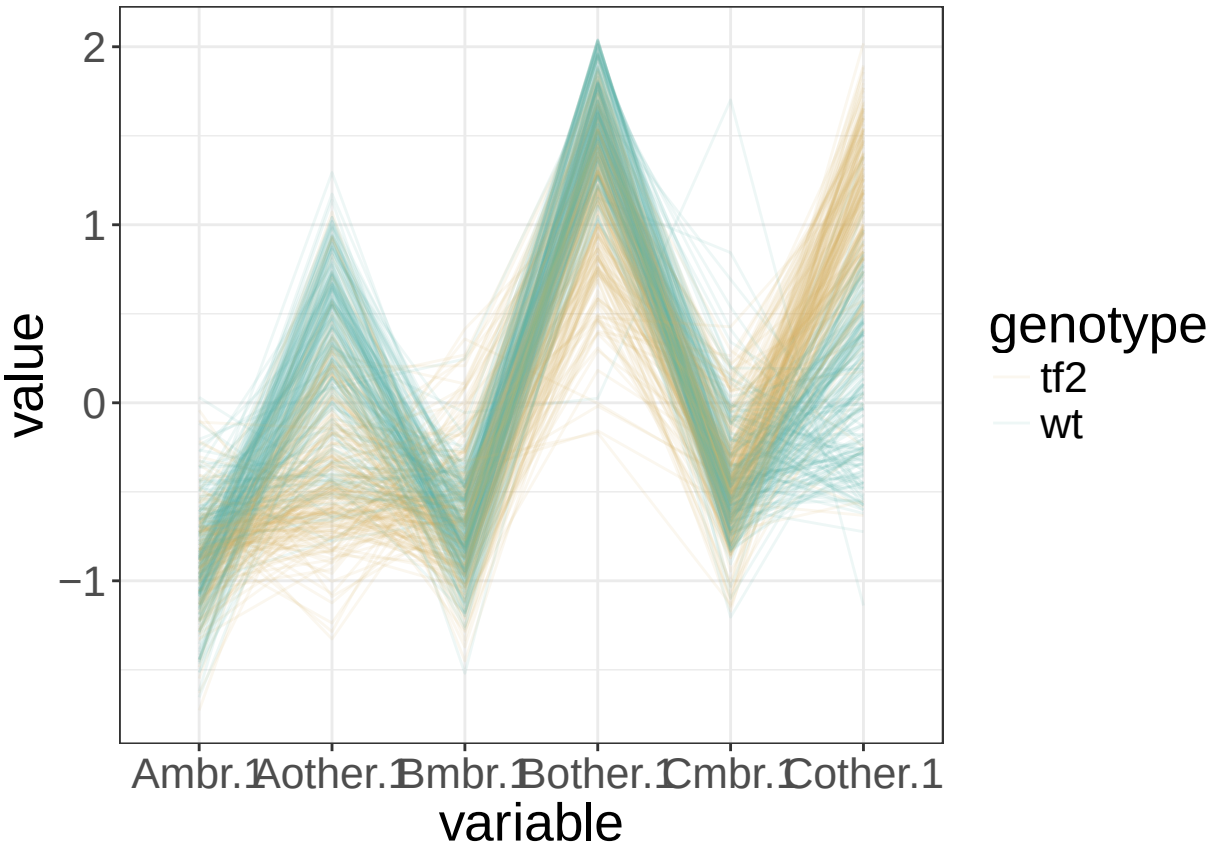
```
clusterVis_region_ssom(9)
```

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




```
clusterVis_line_ssom(9)
```

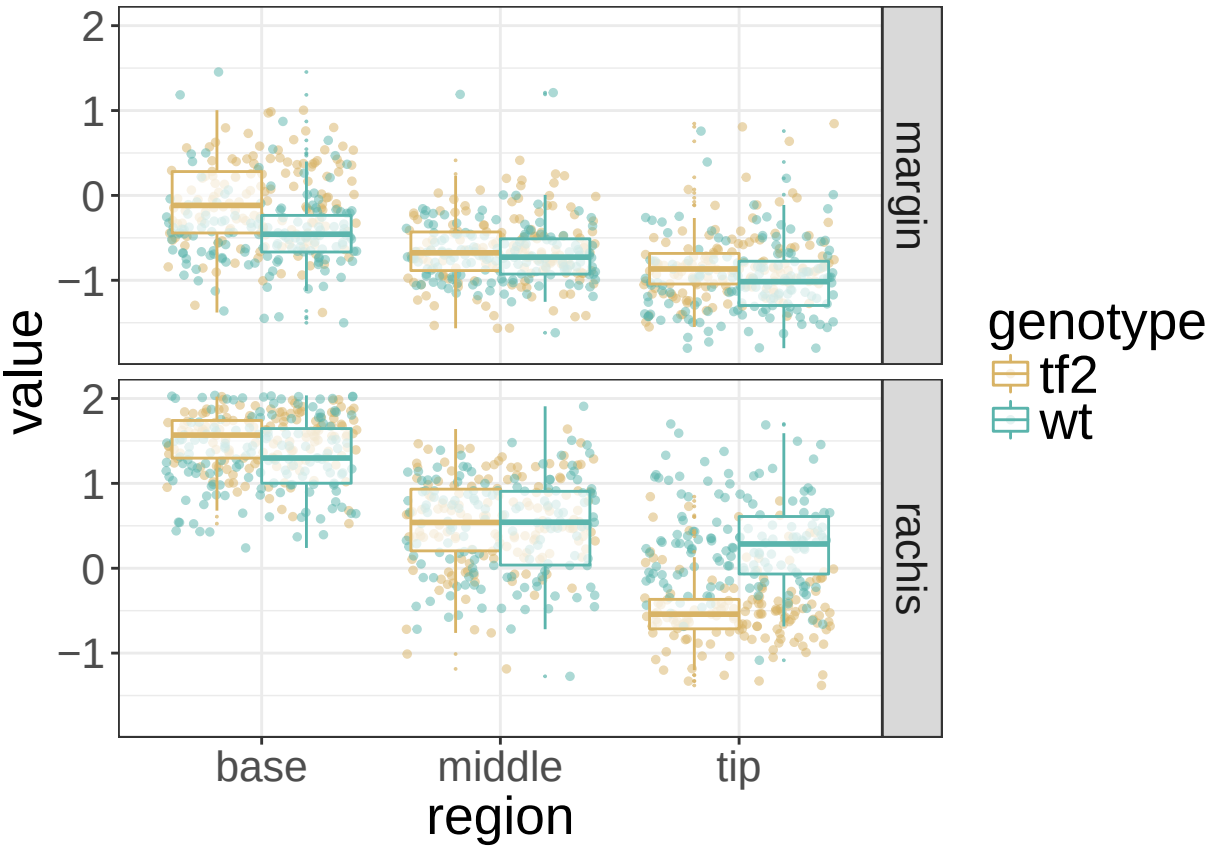
```
## Using genotype, gene as id variables
```




```
# genesInClust_ssom(9)
```

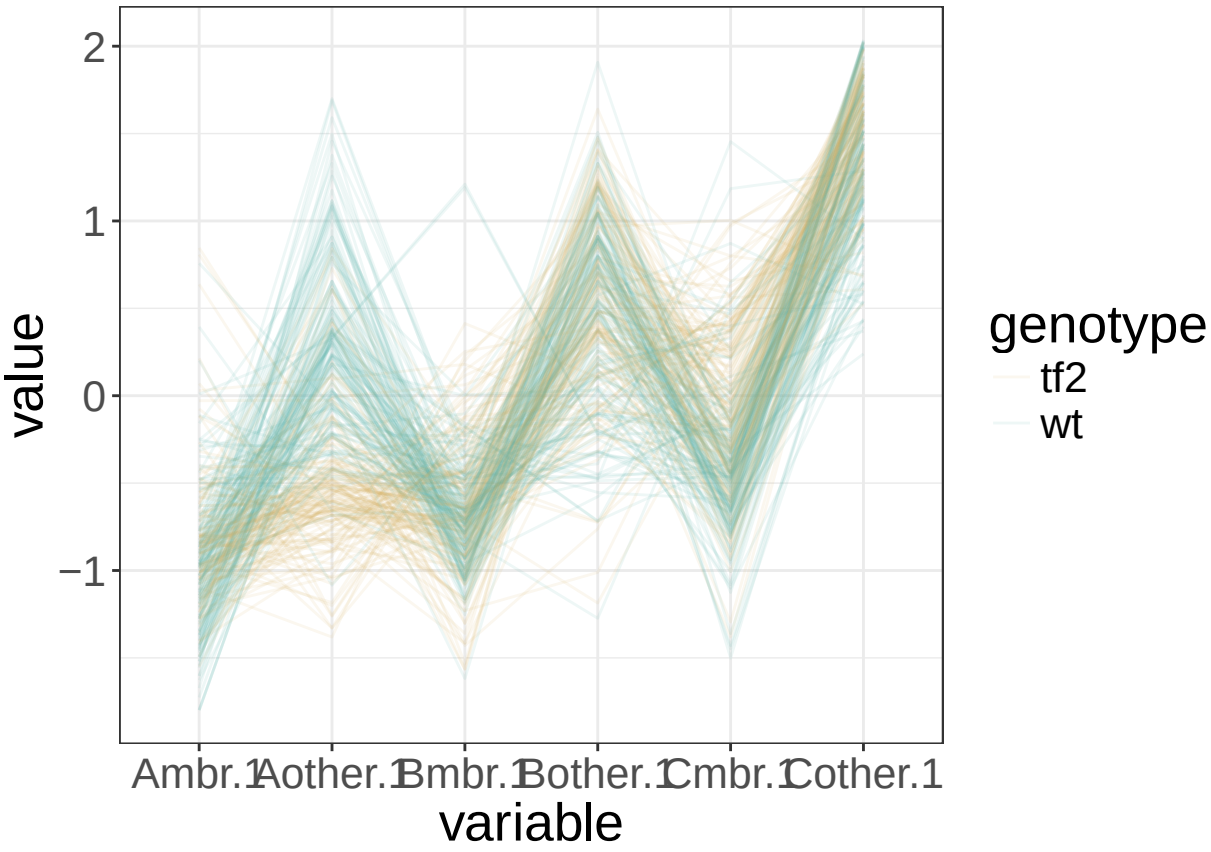
```
clusterVis_region_ssom(10)
```

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



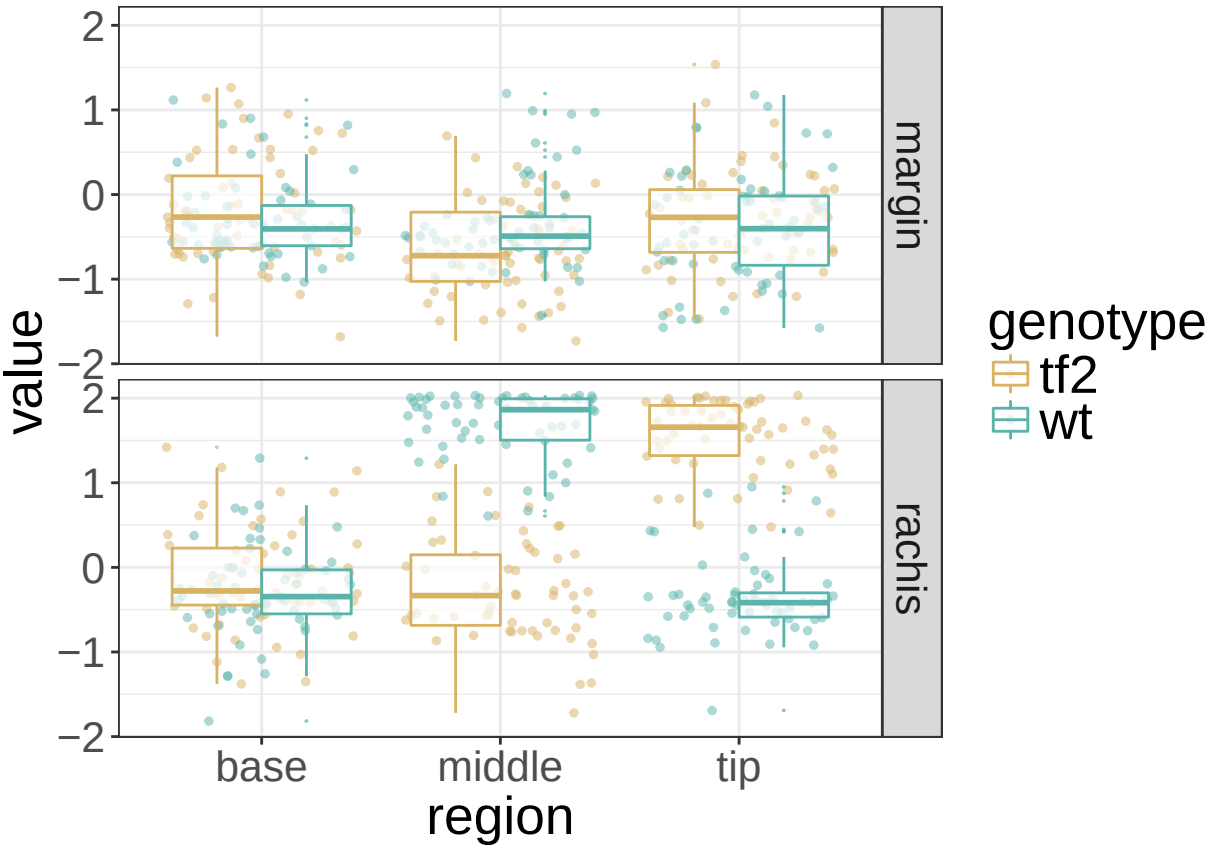

```
clusterVis_line_ssom(10)
```

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



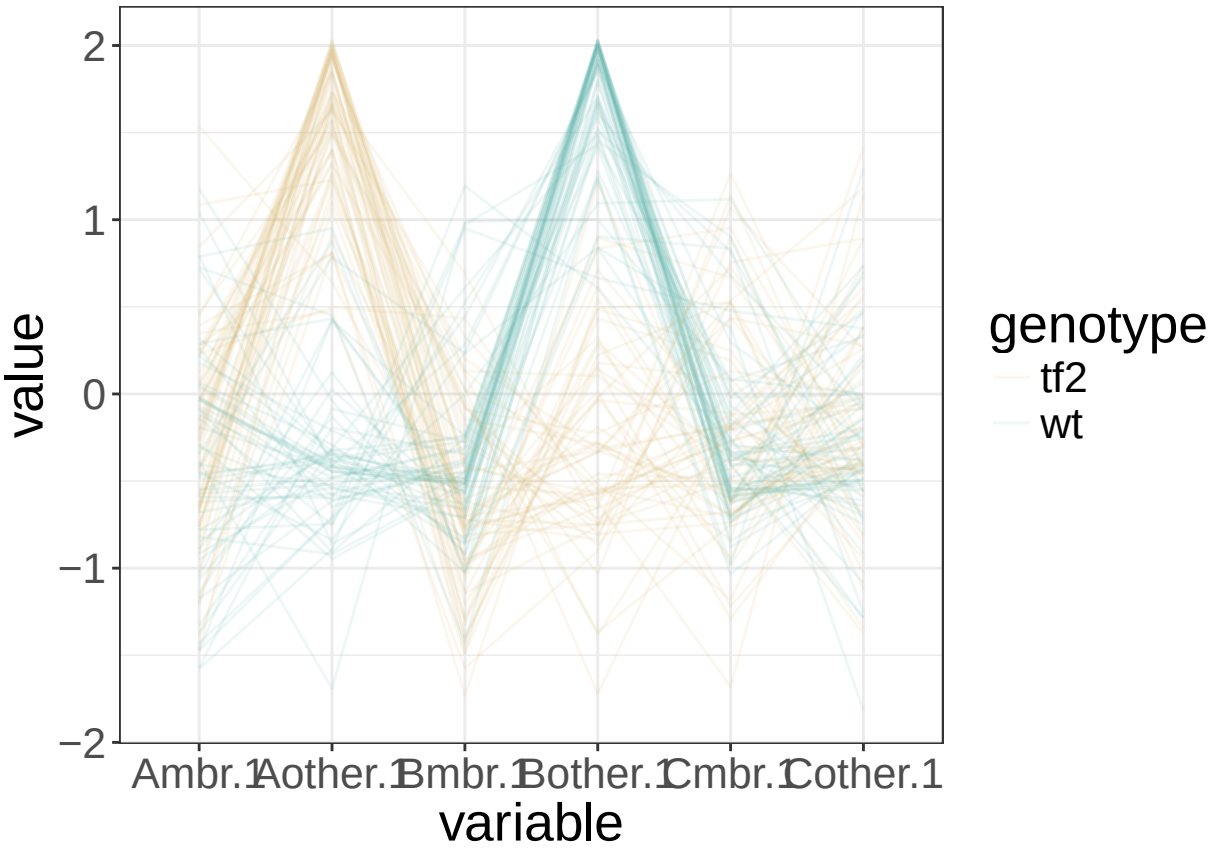

```
# genesInClust_ssom(10)

clusterVis_region_ssom(11)
```

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



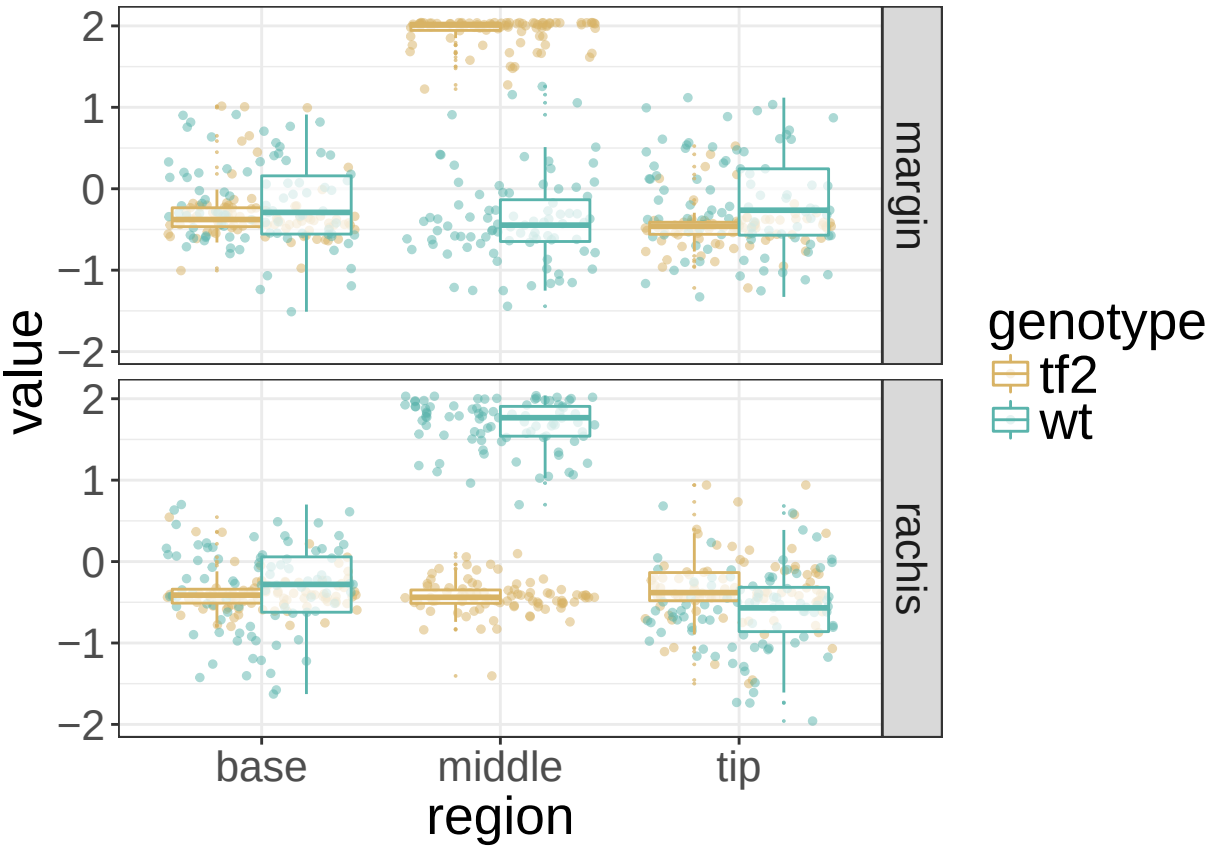

```
clusterVis_line_ssom(11)
```

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



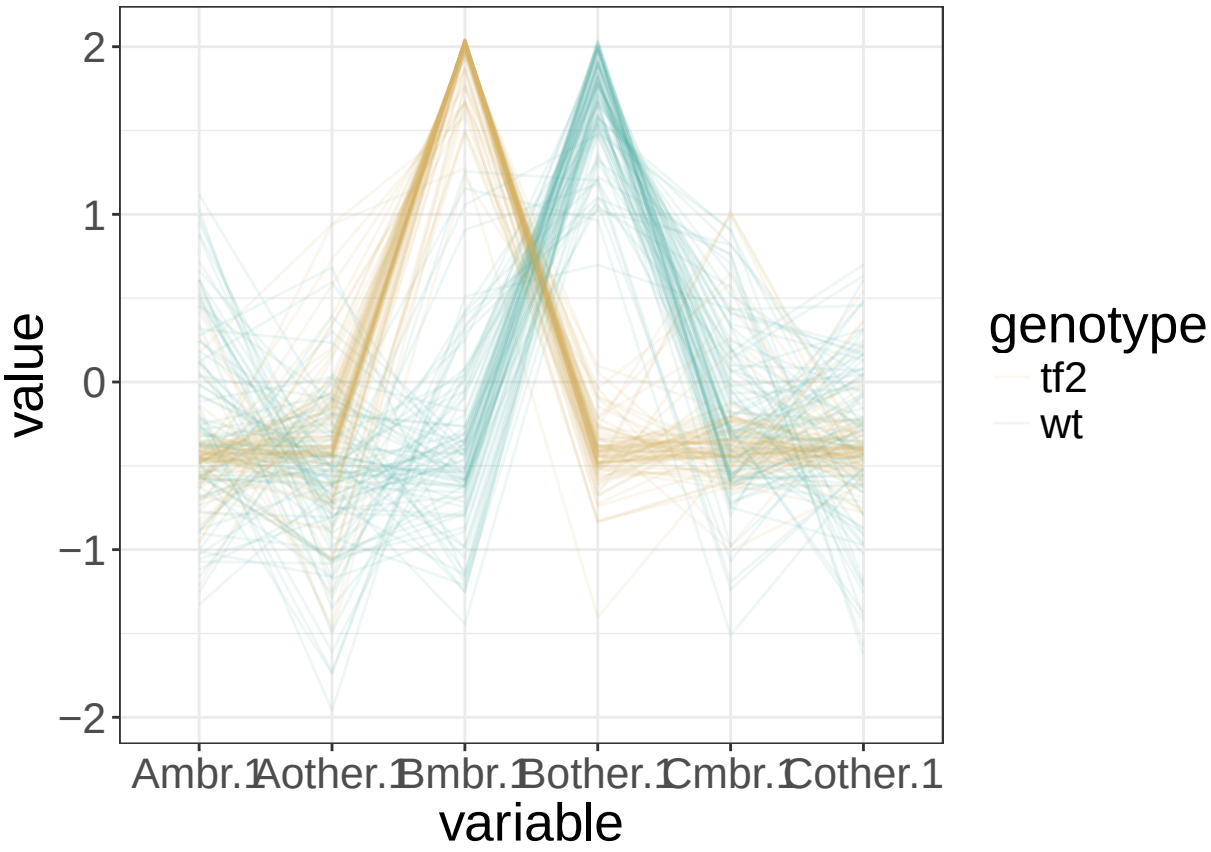

```
# genesInClust_ssom(11)

clusterVis_region_ssom(12)
```

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



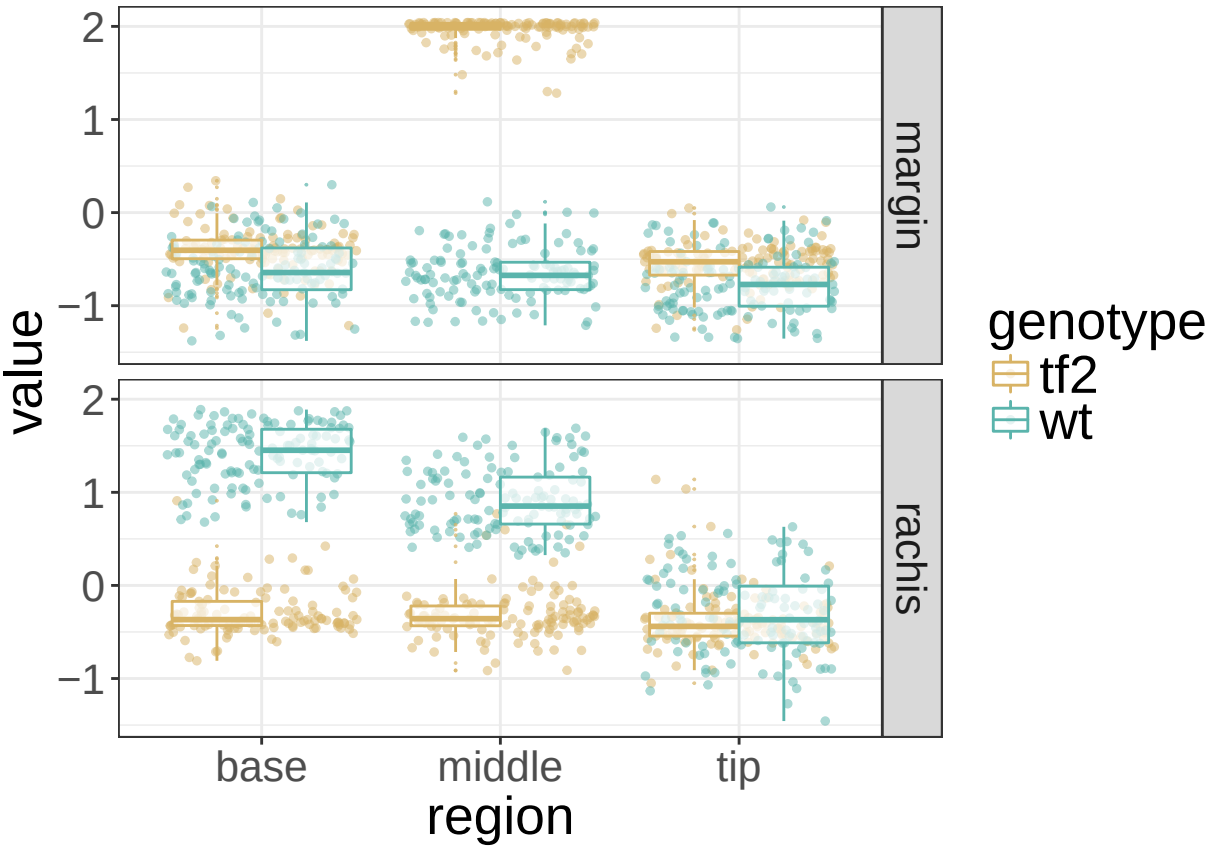

```
clusterVis_line_ssom(12)
```

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



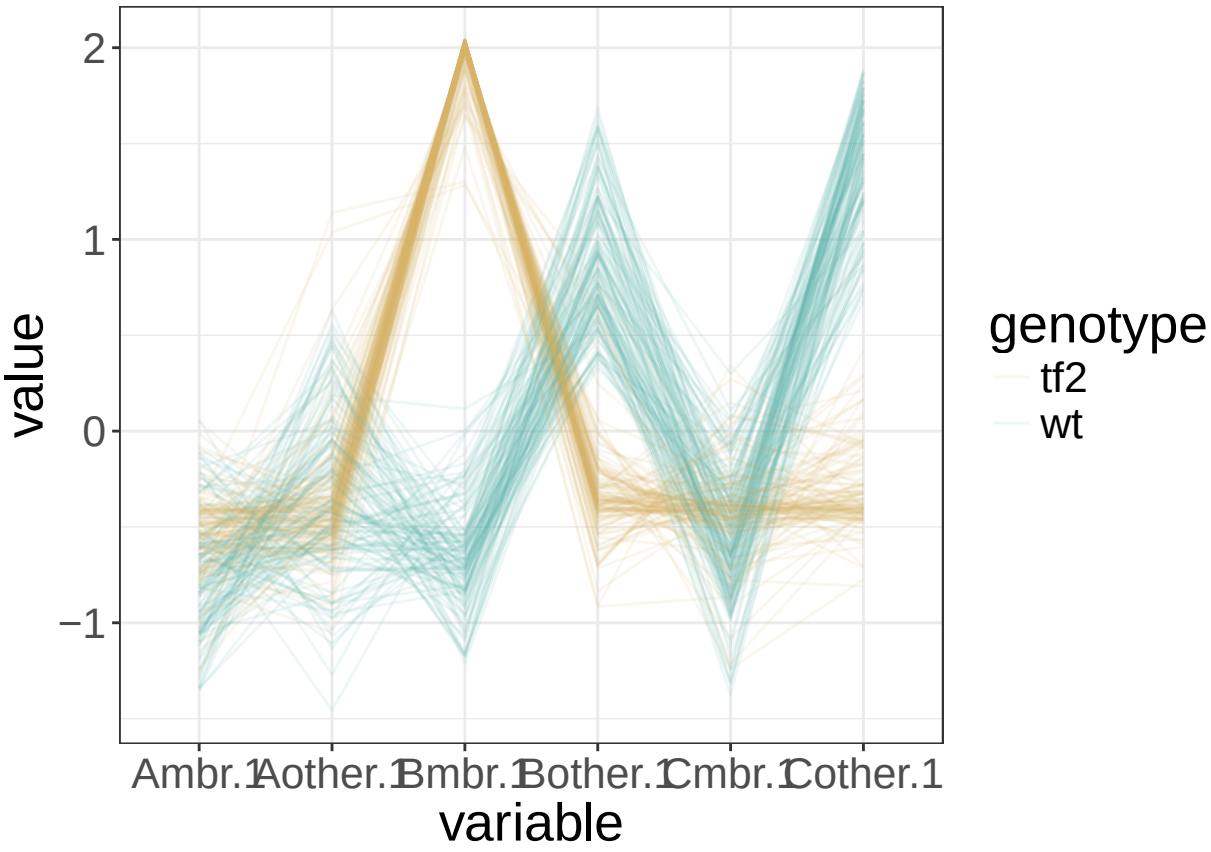

```
# genesInClust_ssom(12)

clusterVis_region_ssom(13)
```

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



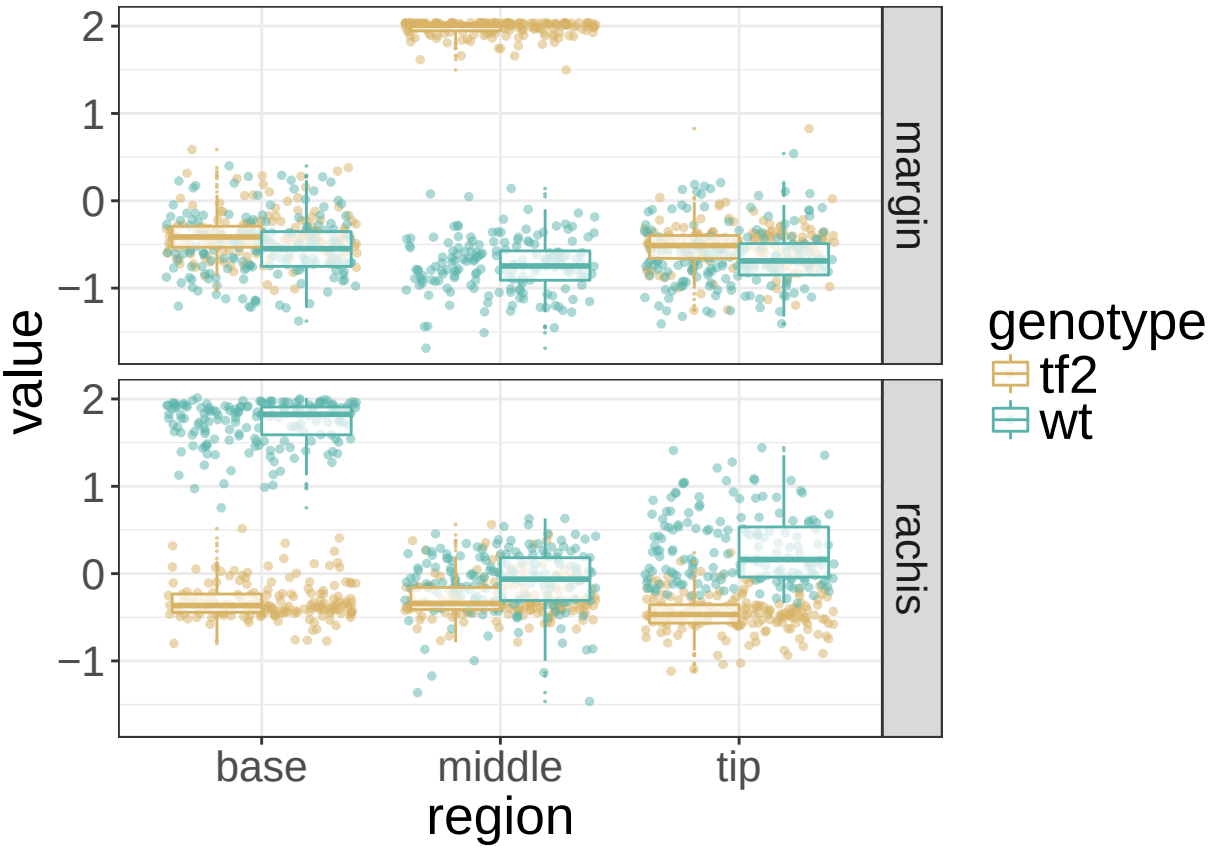

```
clusterVis_line_ssom(13)
```

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



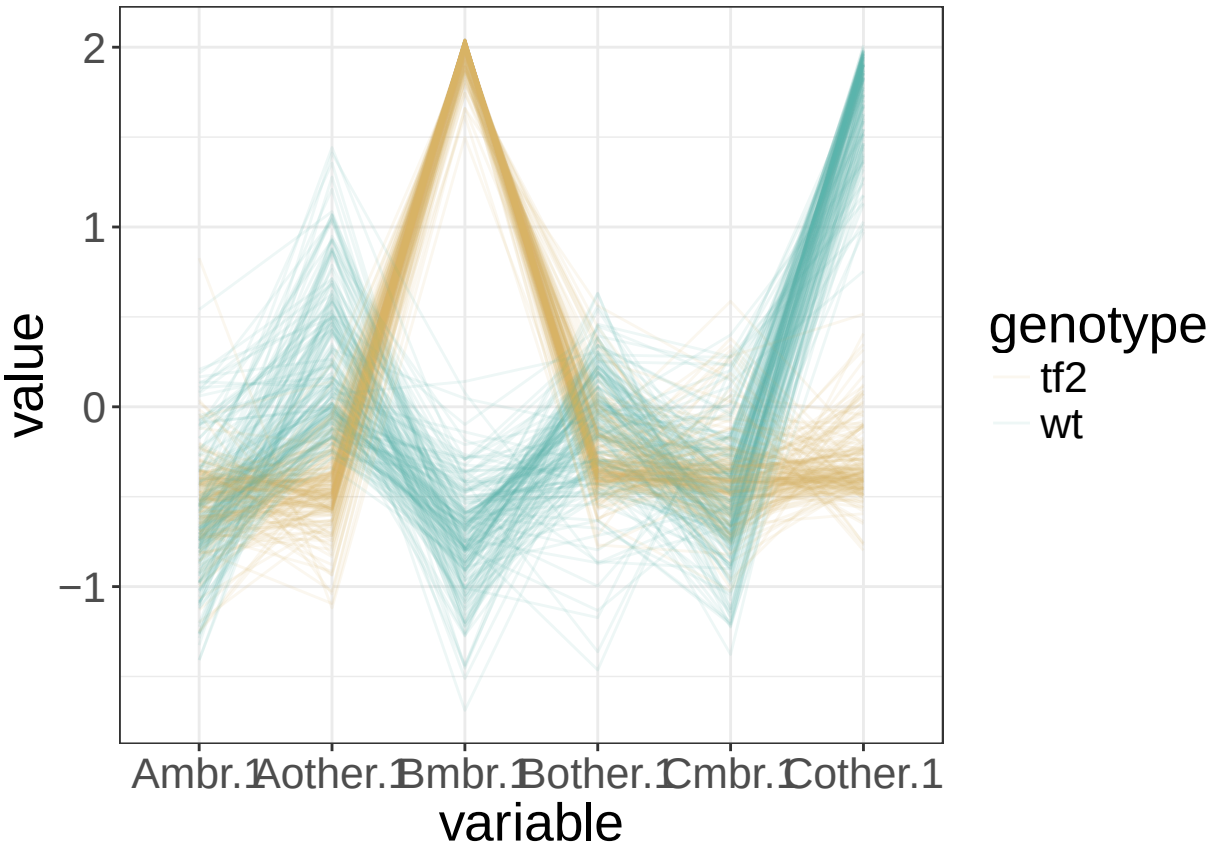

```
# genesInClust_ssom(13)

clusterVis_region_ssom(14)
```

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




```
clusterVis_line_ssom(14)
```

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




```
# genesInClust_ssom(14)
```

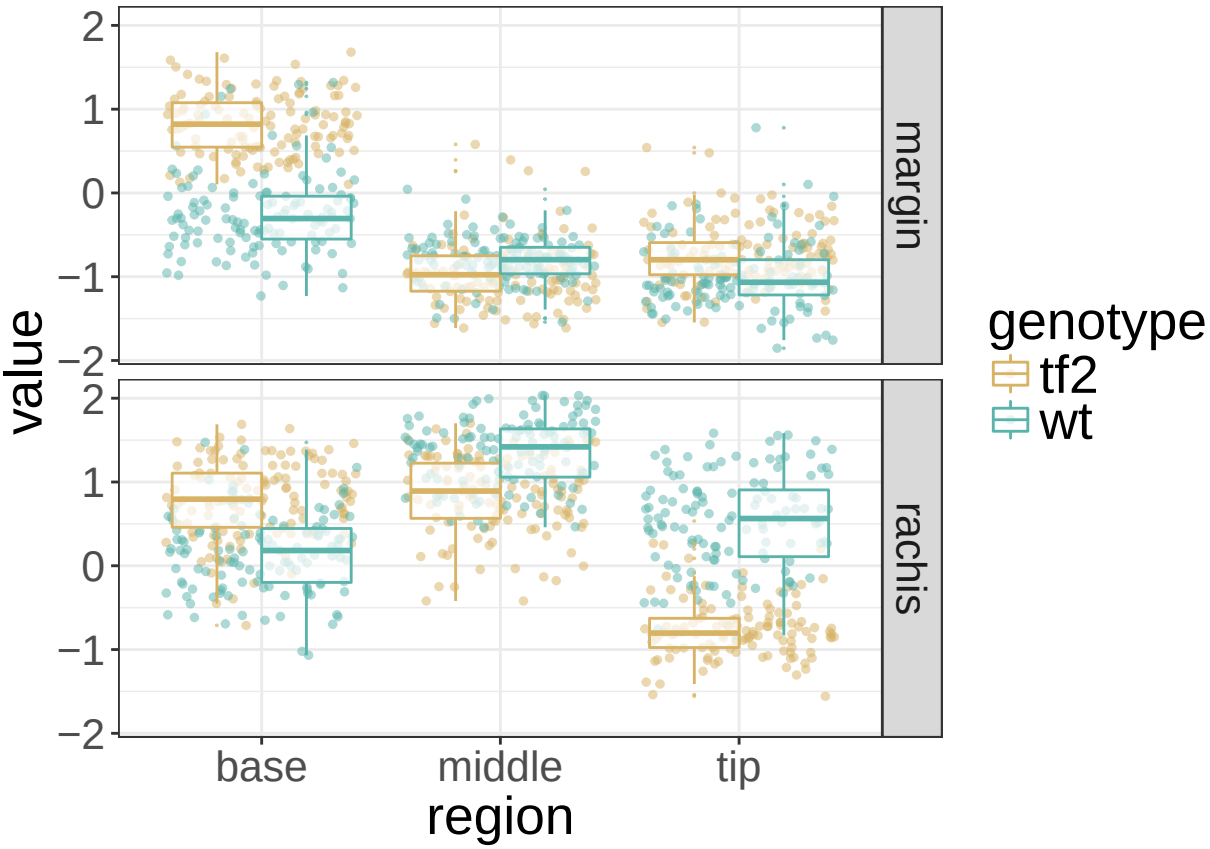
Cluster 15

Up regulated in the base region (rachis and margin) in tf2 compared to wt.

ARF8 Auxin inducible promoter

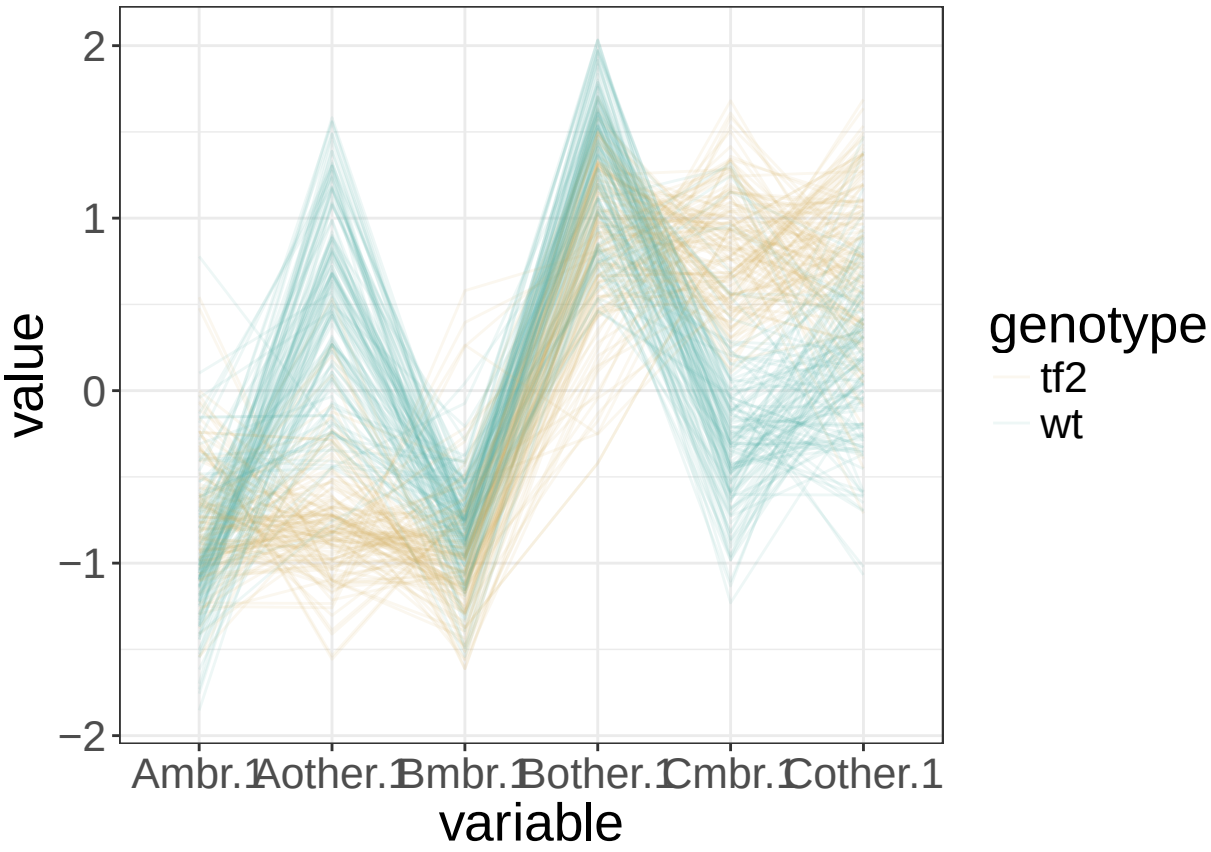
```
clusterVis_region_ssom(15)
```

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



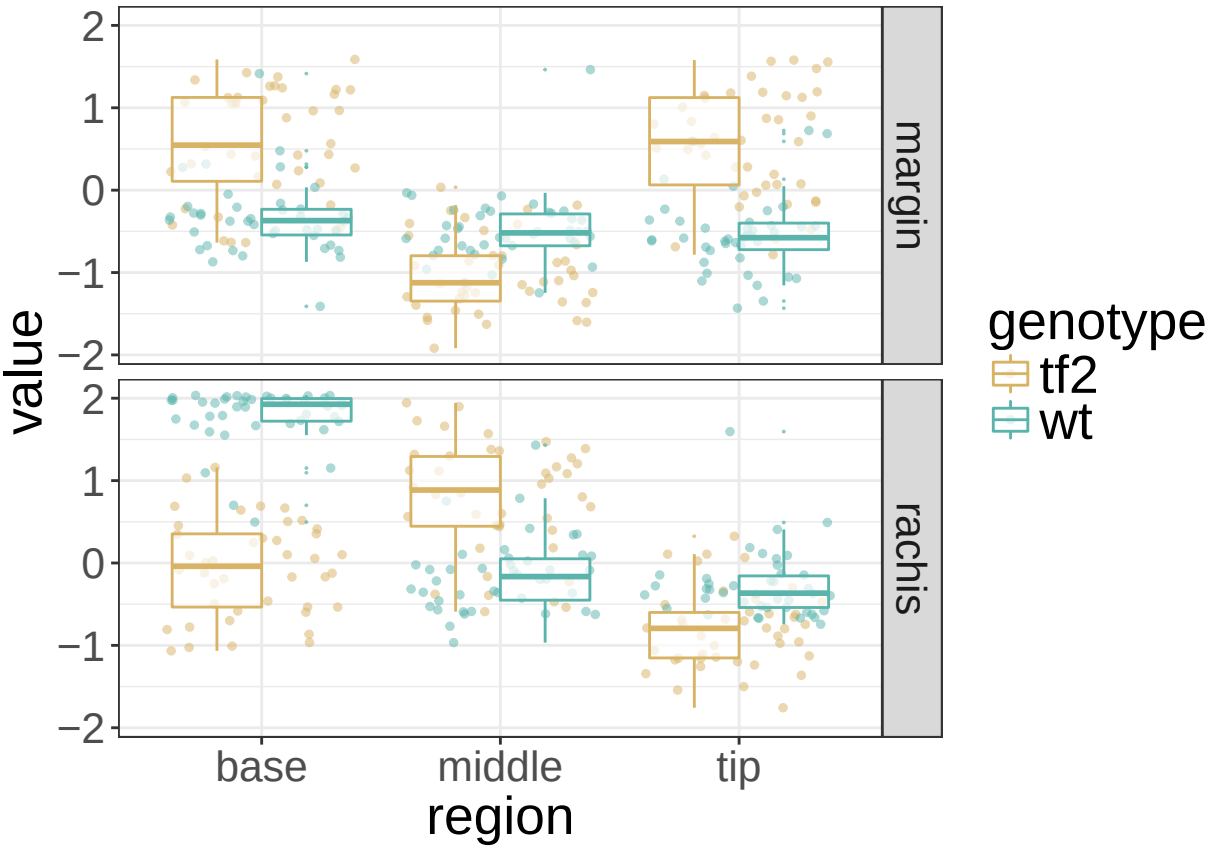

```
clusterVis_line_ssom(15)
```

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



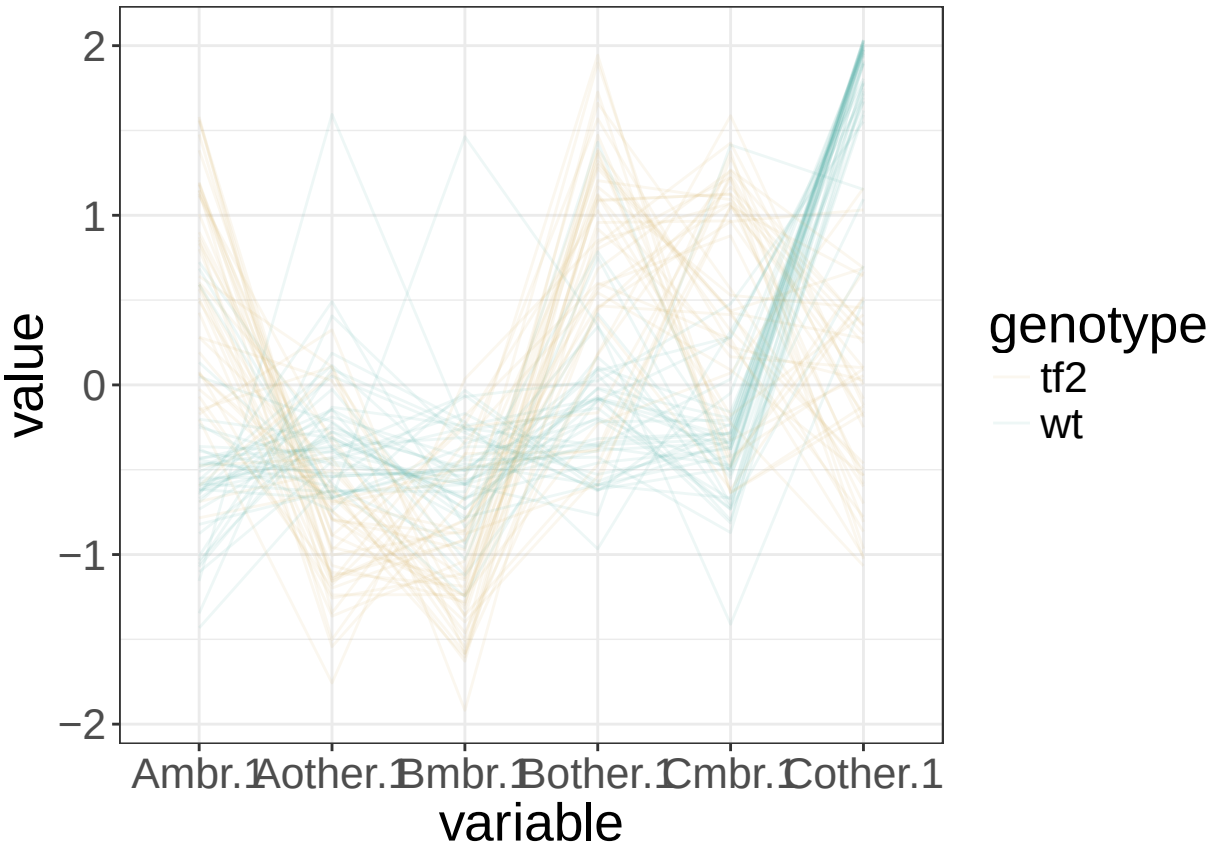

```
# genesInClust_ssom(15)

clusterVis_region_ssom(16)
```

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



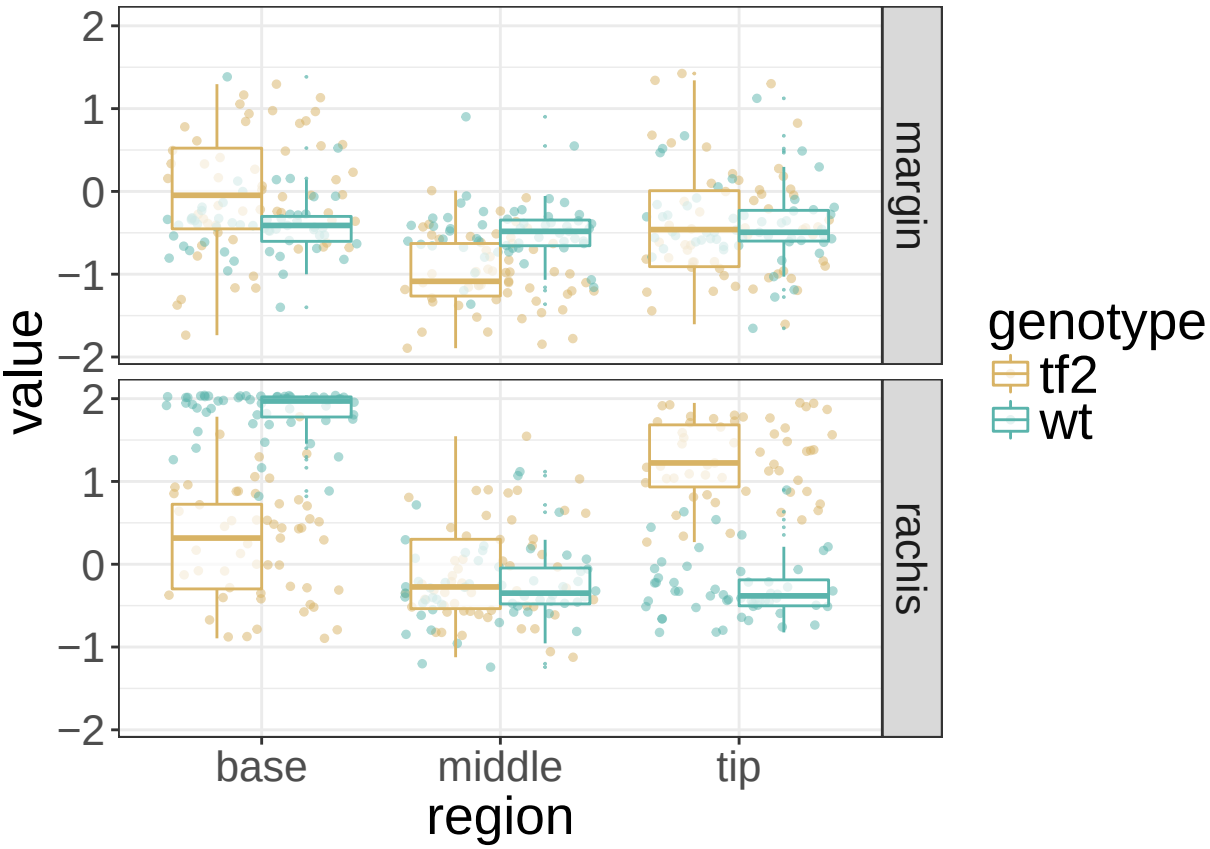

```
clusterVis_line_ssom(16)
```

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



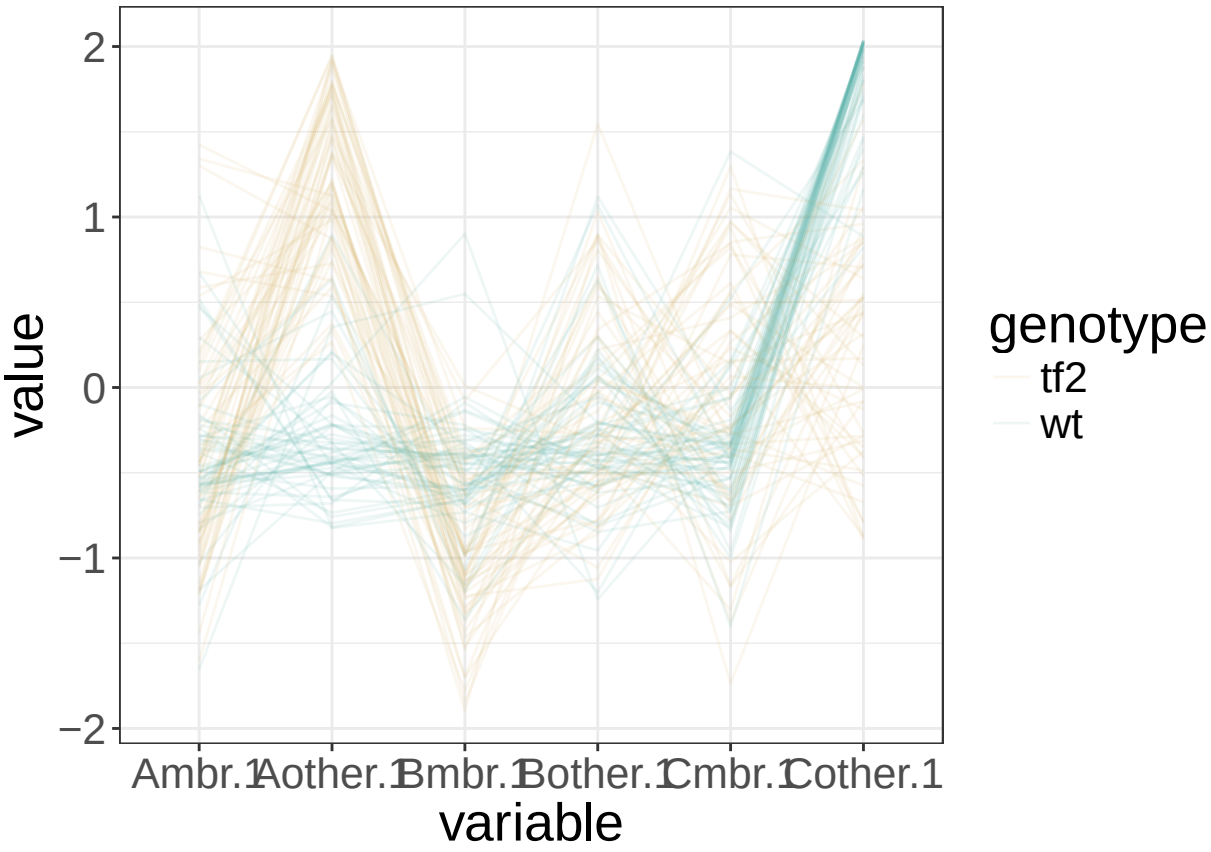

```
# genesInClust_ssom(17)

clusterVis_region_ssom(17)
```

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



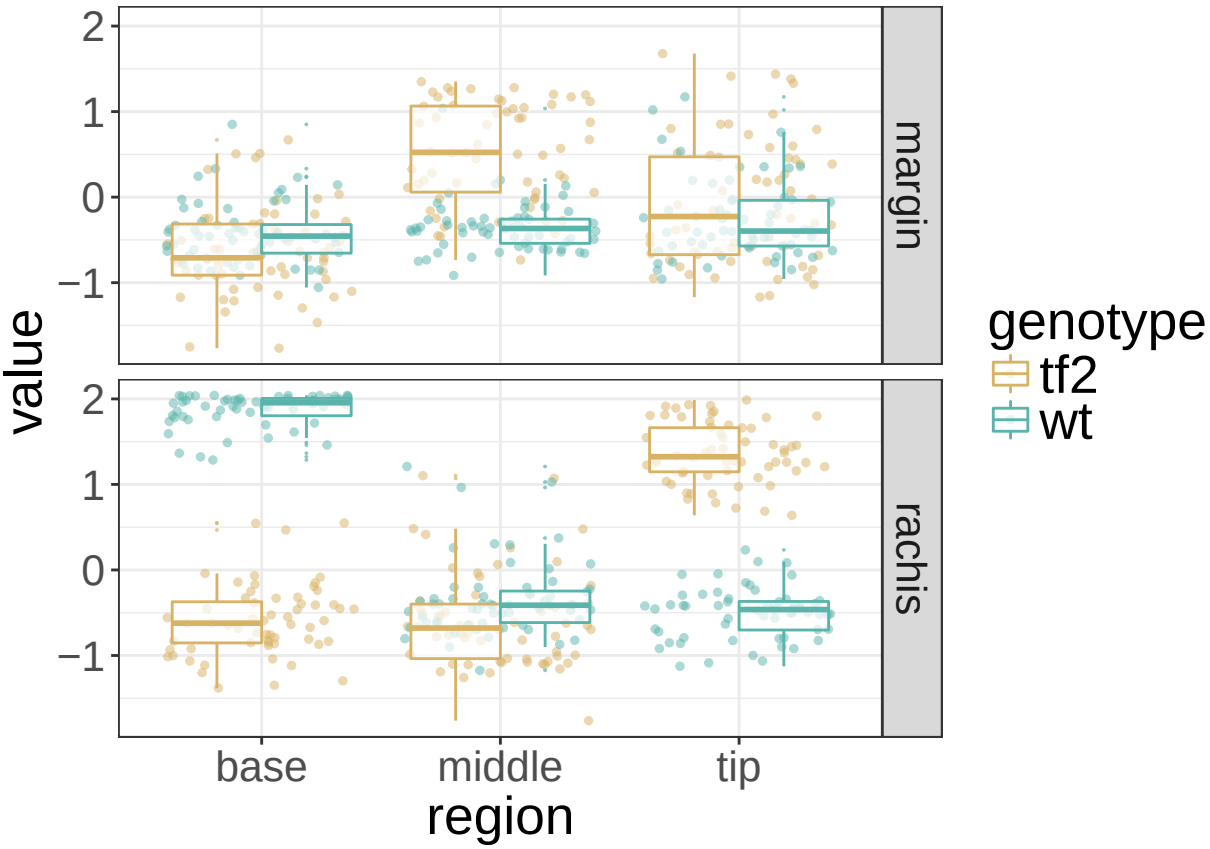

```
clusterVis_line_ssom(17)
```

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



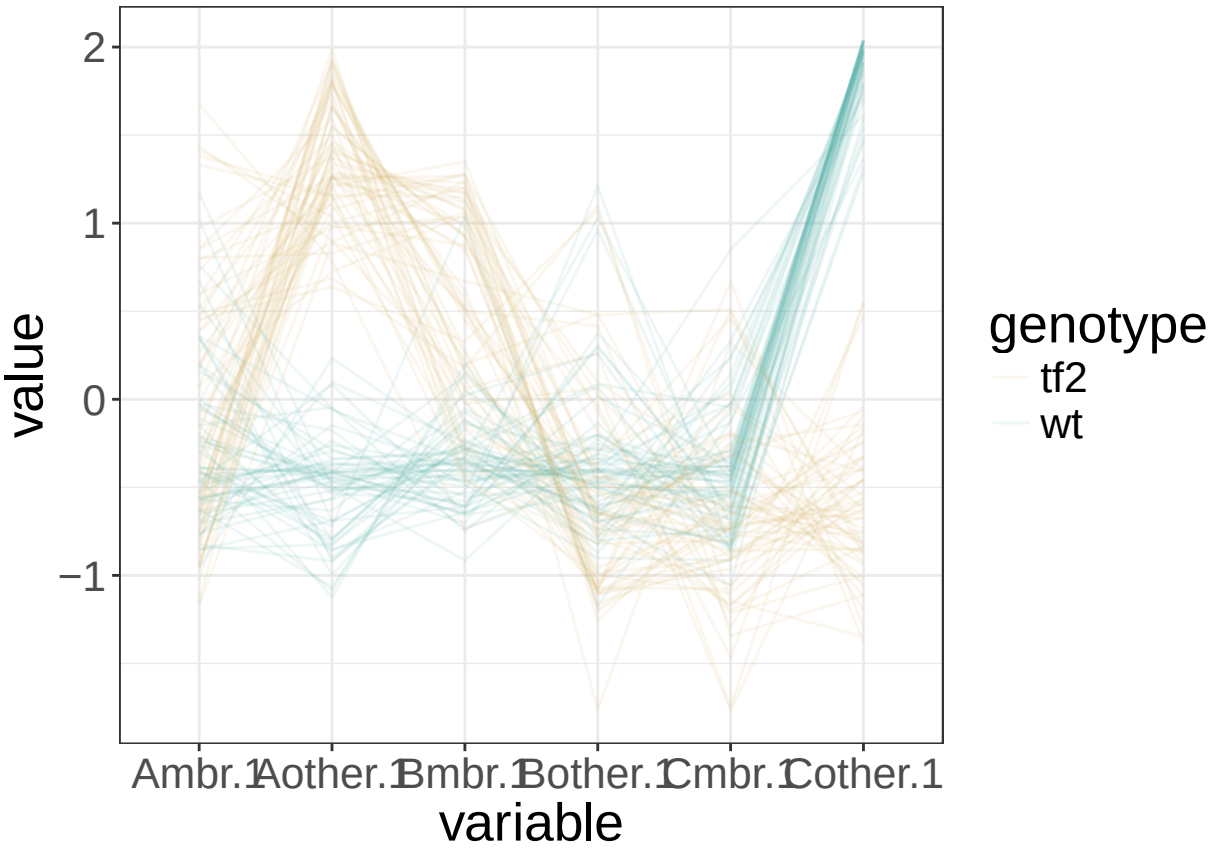

```
clusterVis_region_ssom(18)
```

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



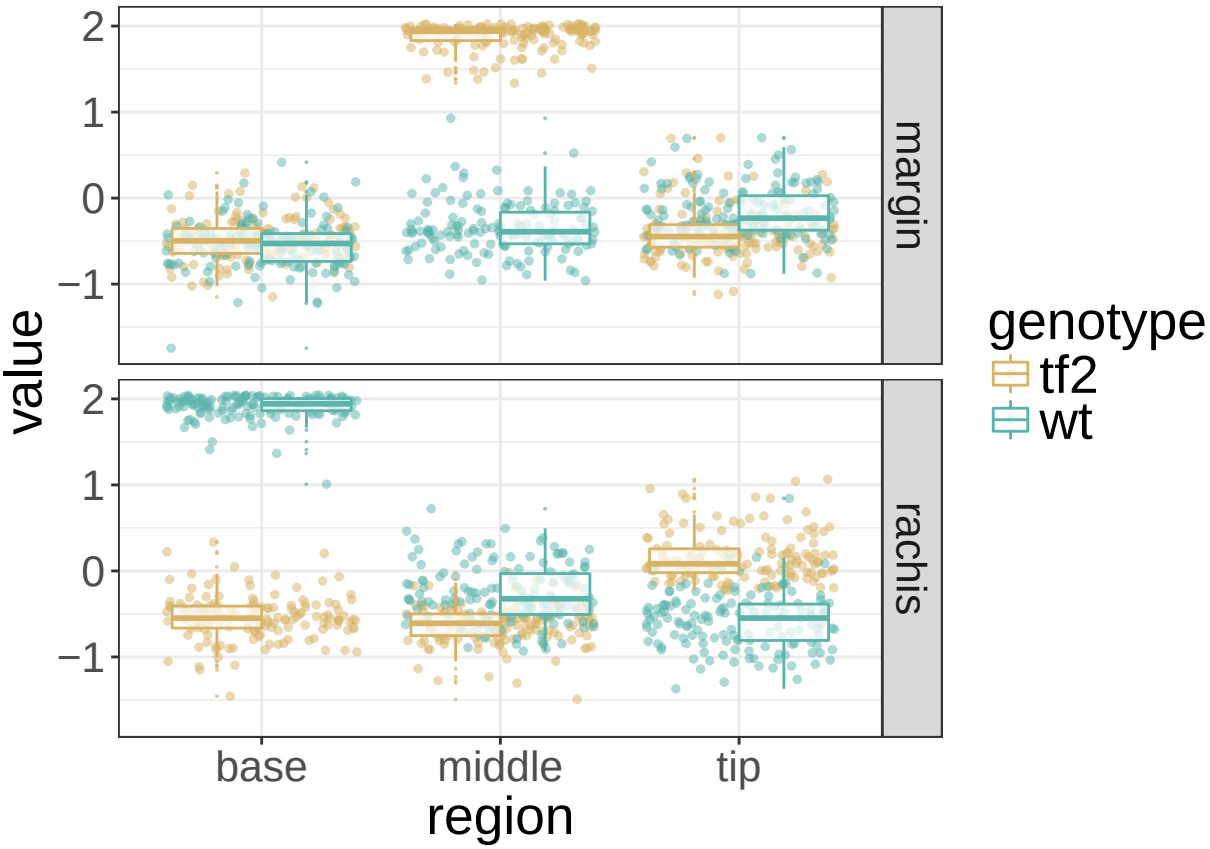

```
clusterVis_line_ssom(18)
```

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



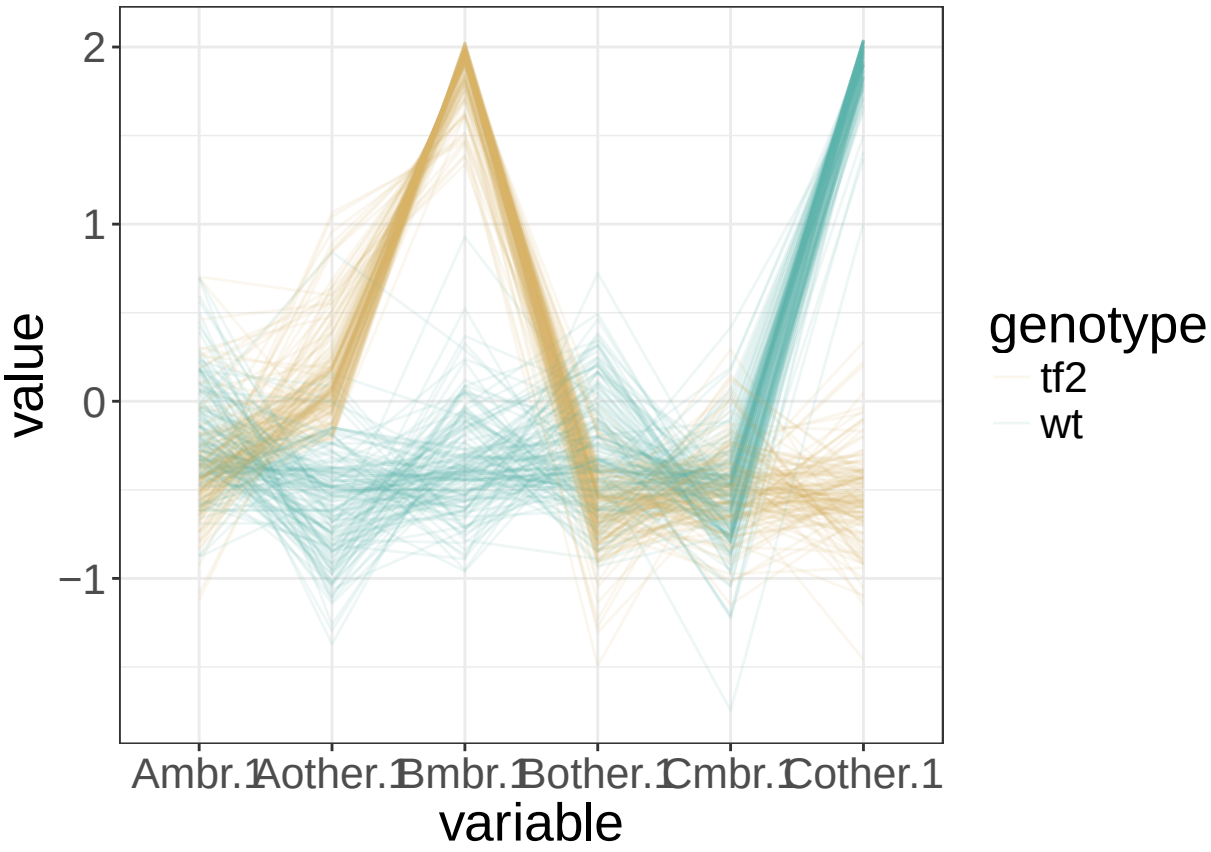

```
clusterVis_region_ssom(19)
```

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



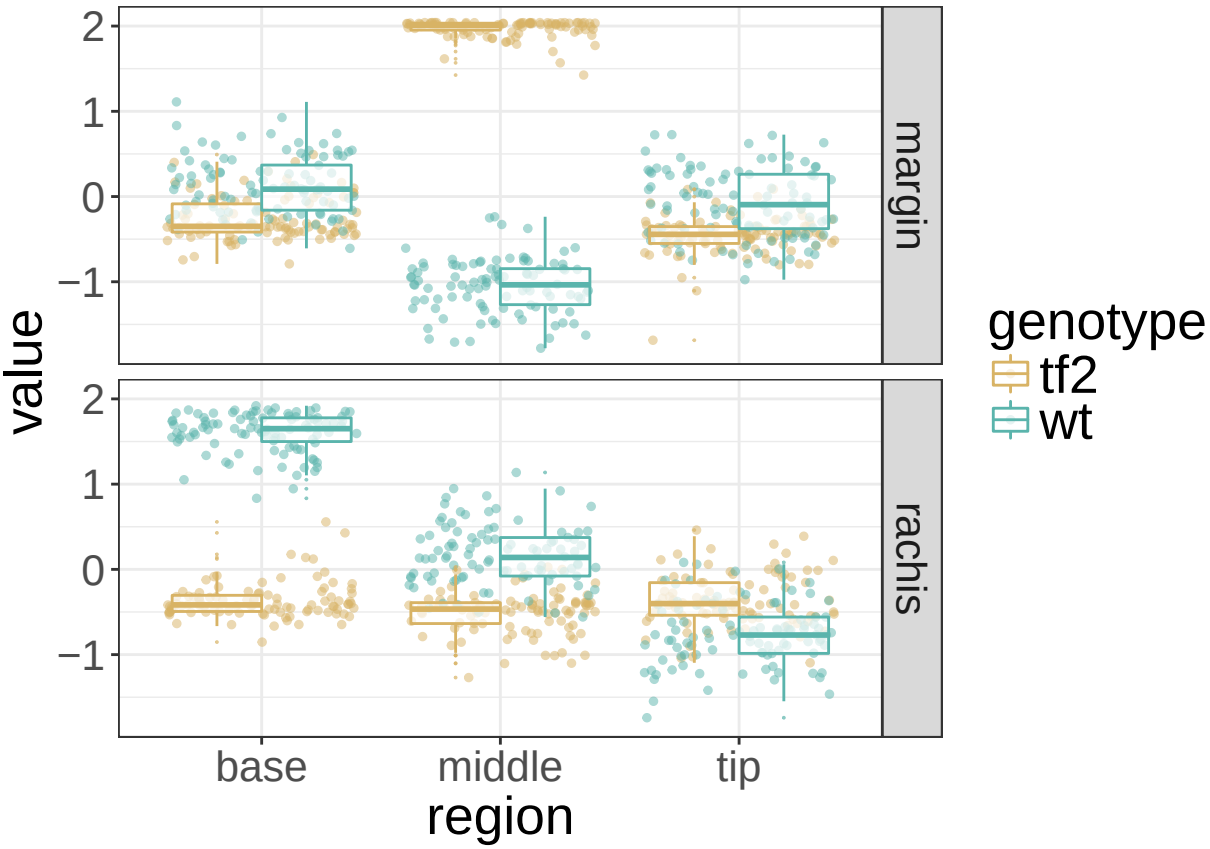

```
clusterVis_line_ssom(19)
```

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



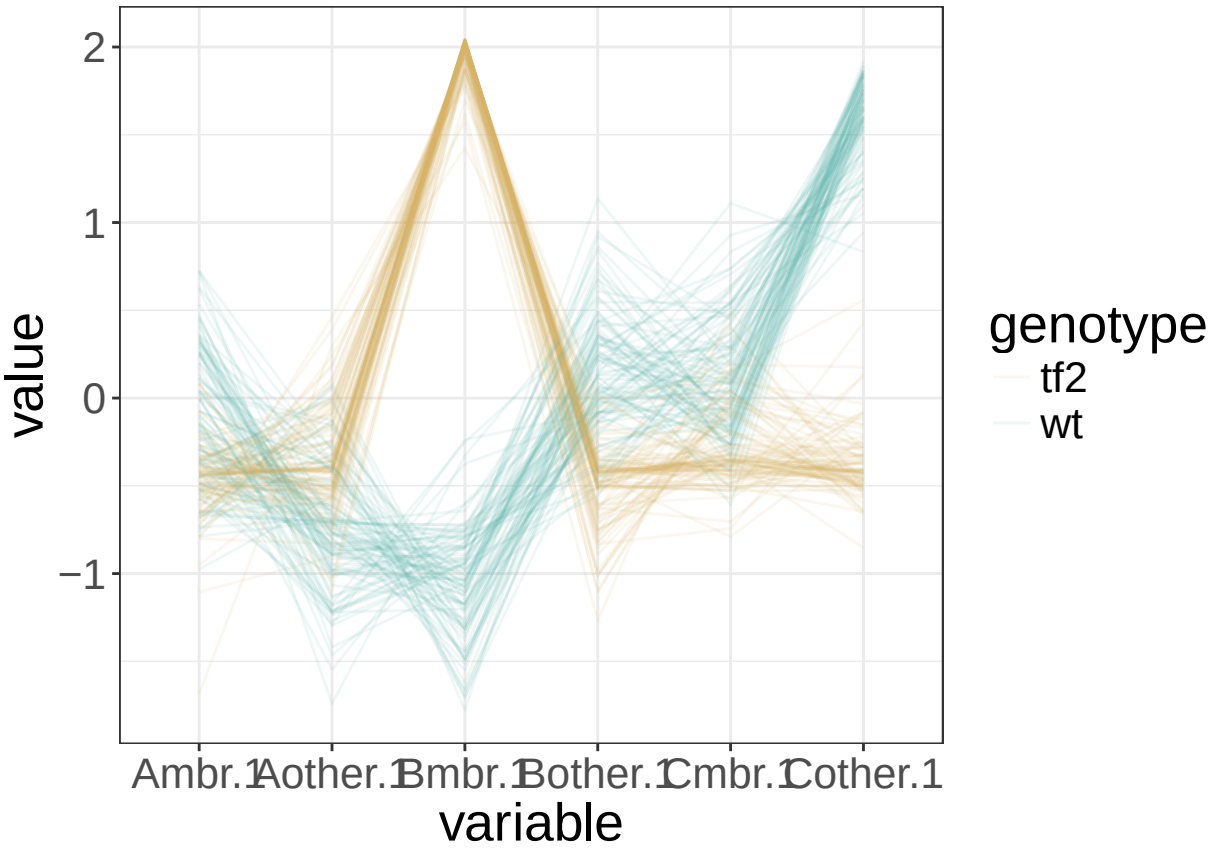

```
clusterVis_region_ssom(20)
```

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



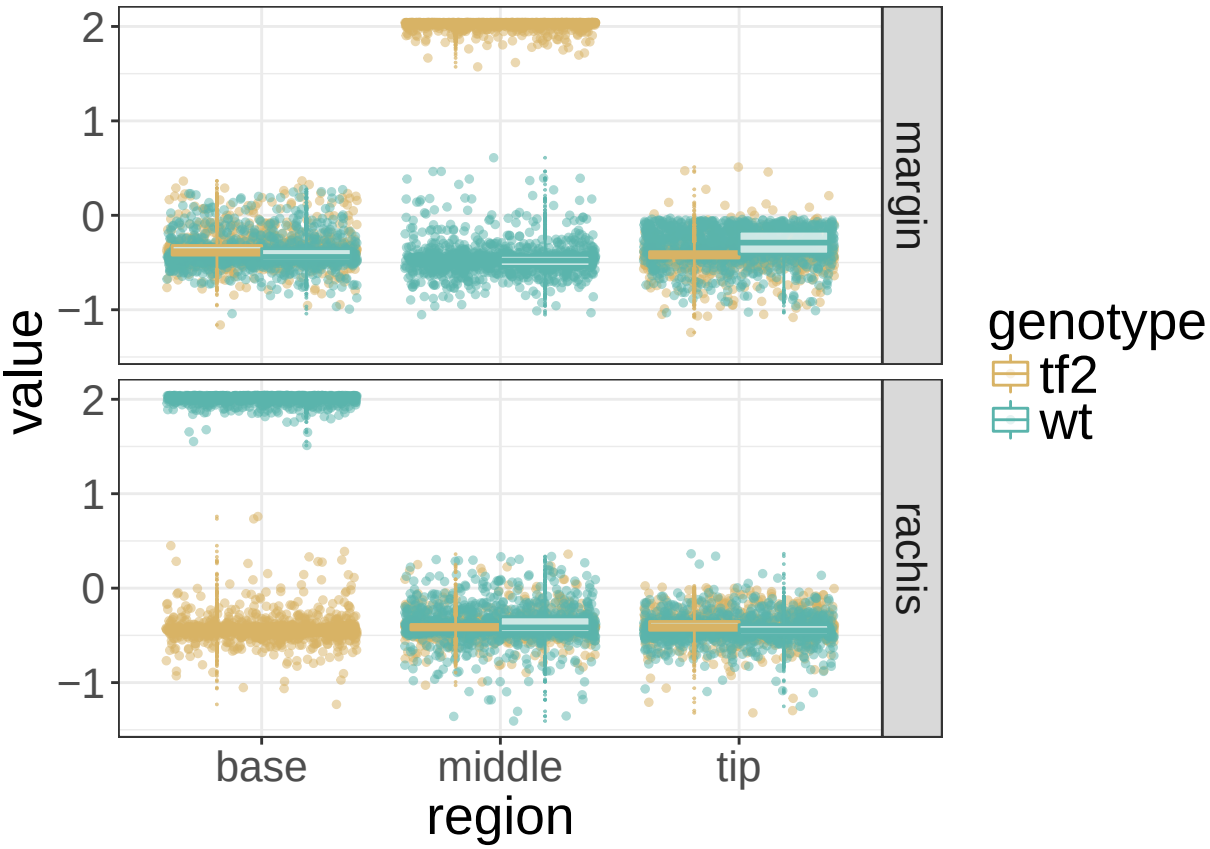

```
clusterVis_line_ssom(20)
```

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



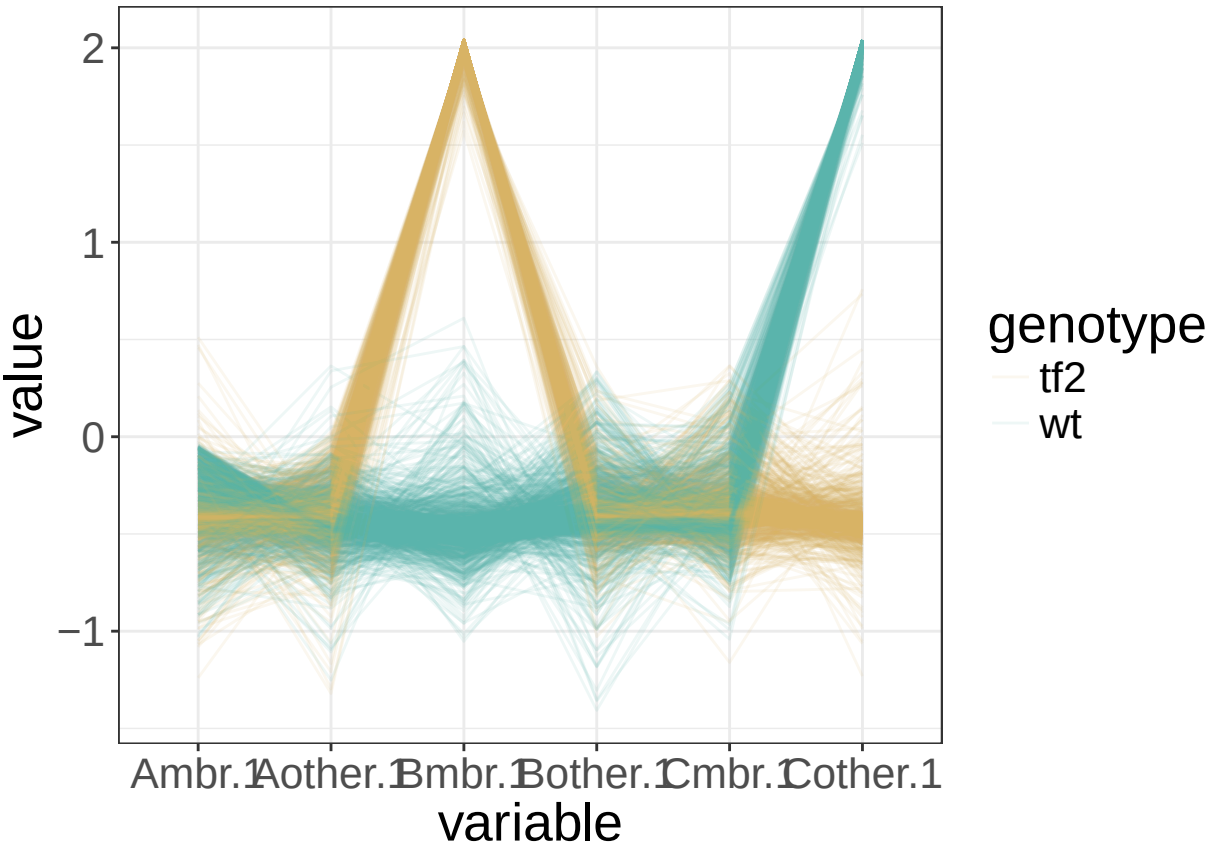

```
clusterVis_region_ssom(21)
```

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




```
clusterVis_line_ssom(21)
```

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



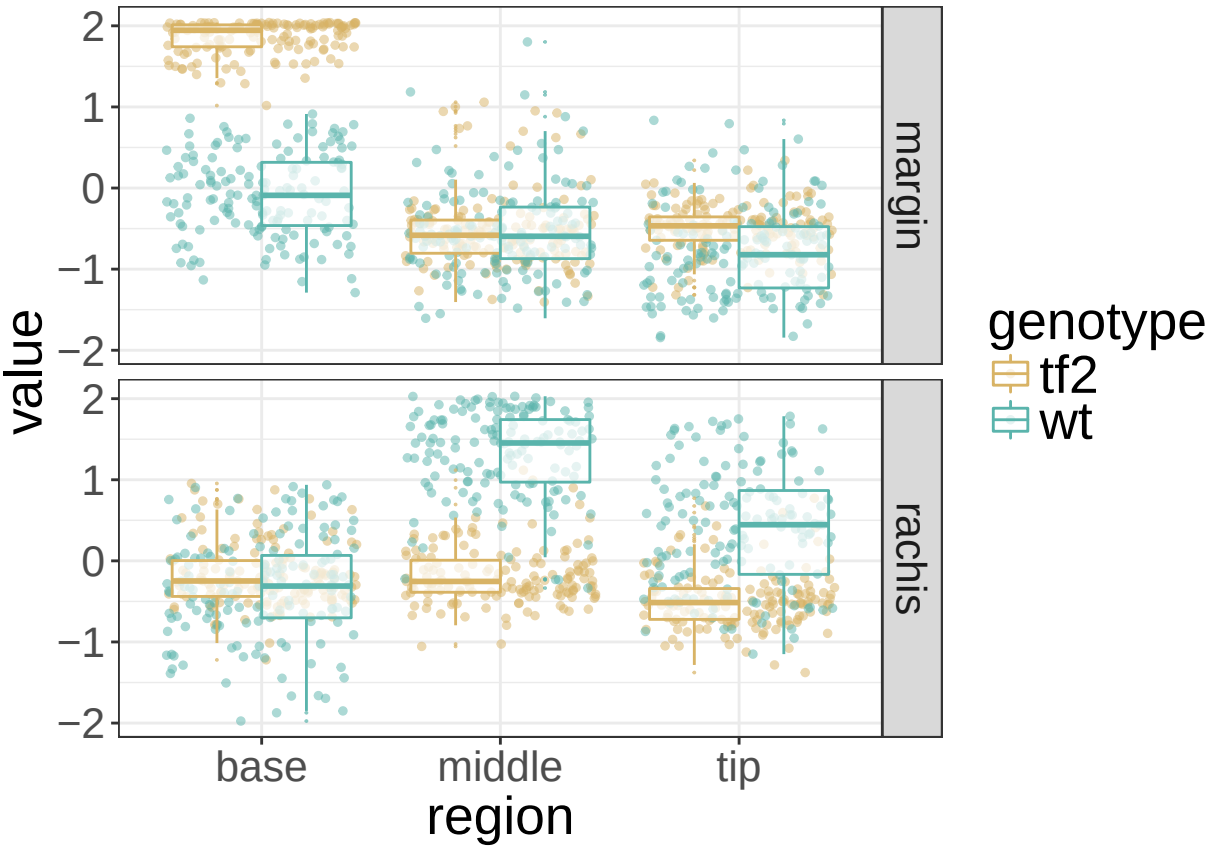
Cluster 22

Up-regulated in margin in tf-2 compared with WT.

IAA9 ATHB-2 ARF9 MYB family transcription factor

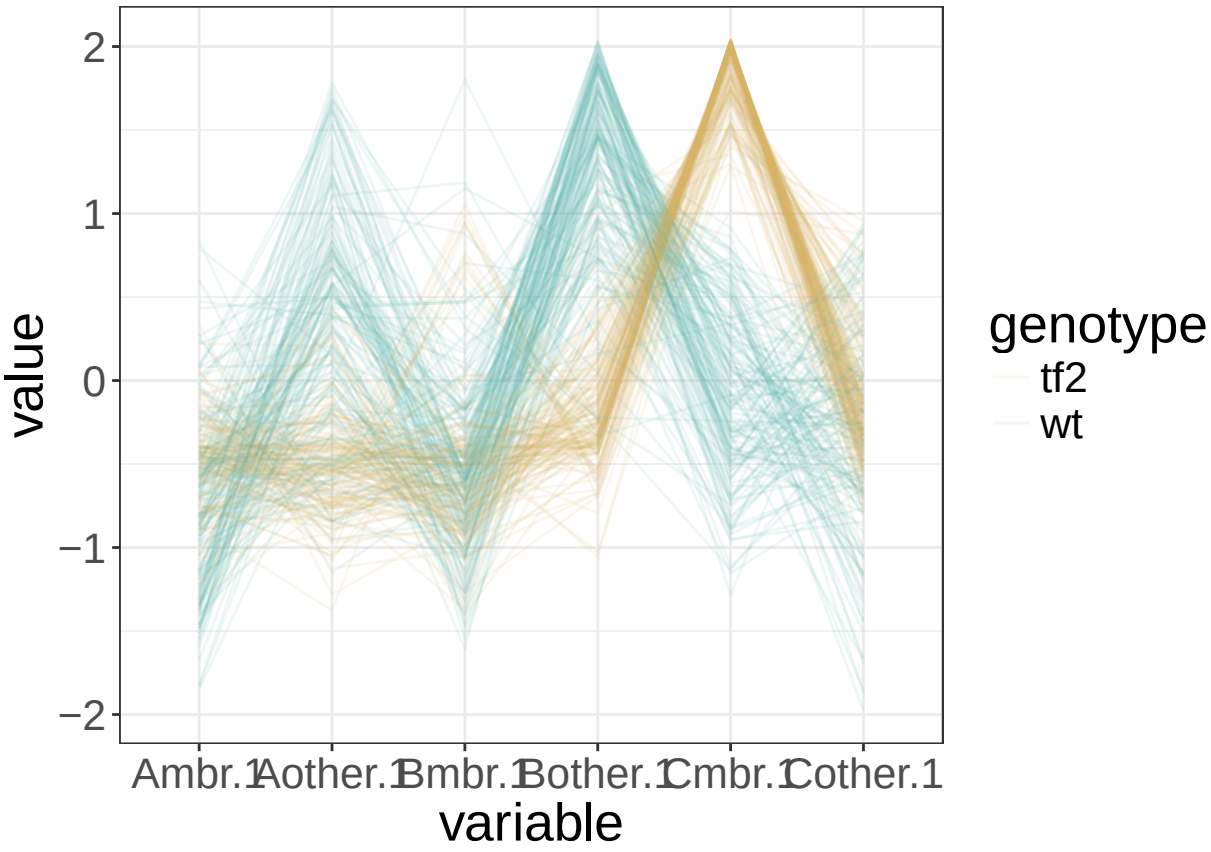
```
clusterVis_region_ssom(22)
```

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



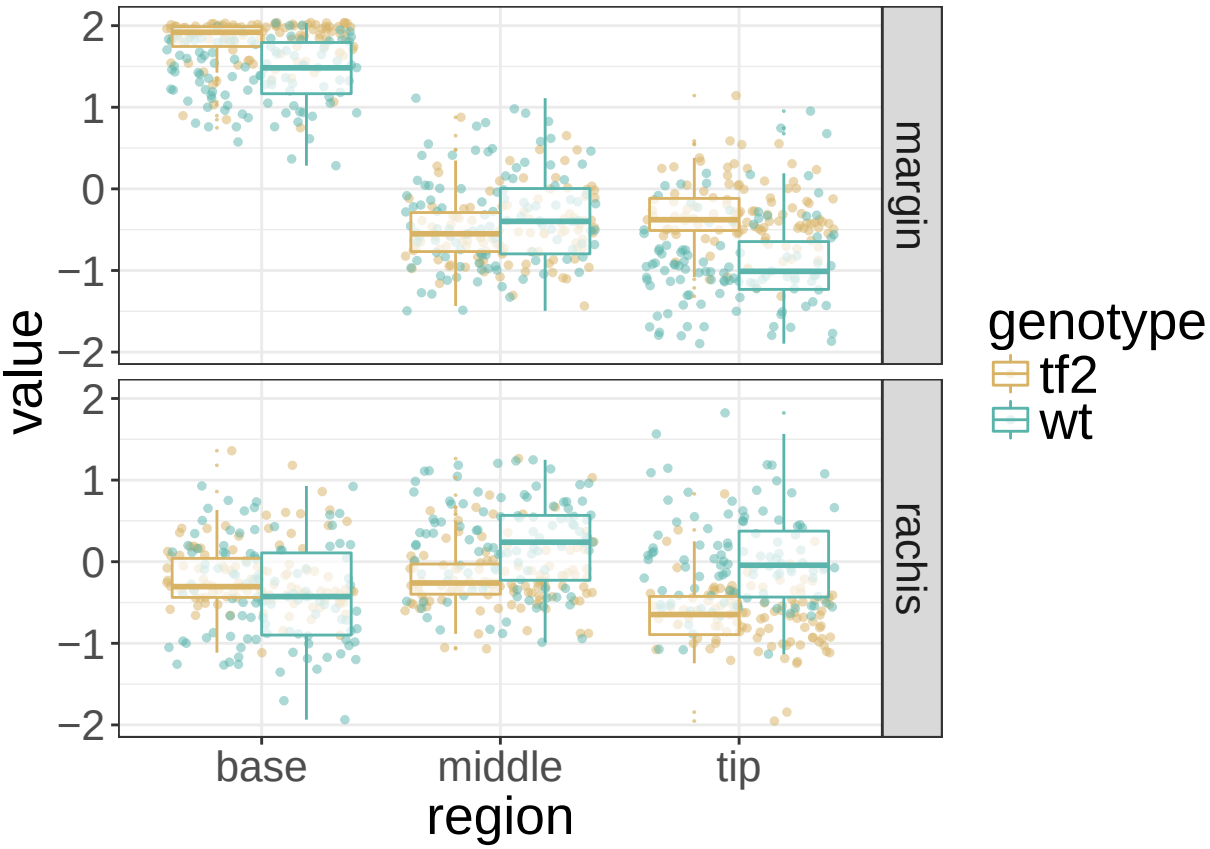

```
clusterVis_line_ssom(22)
```

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



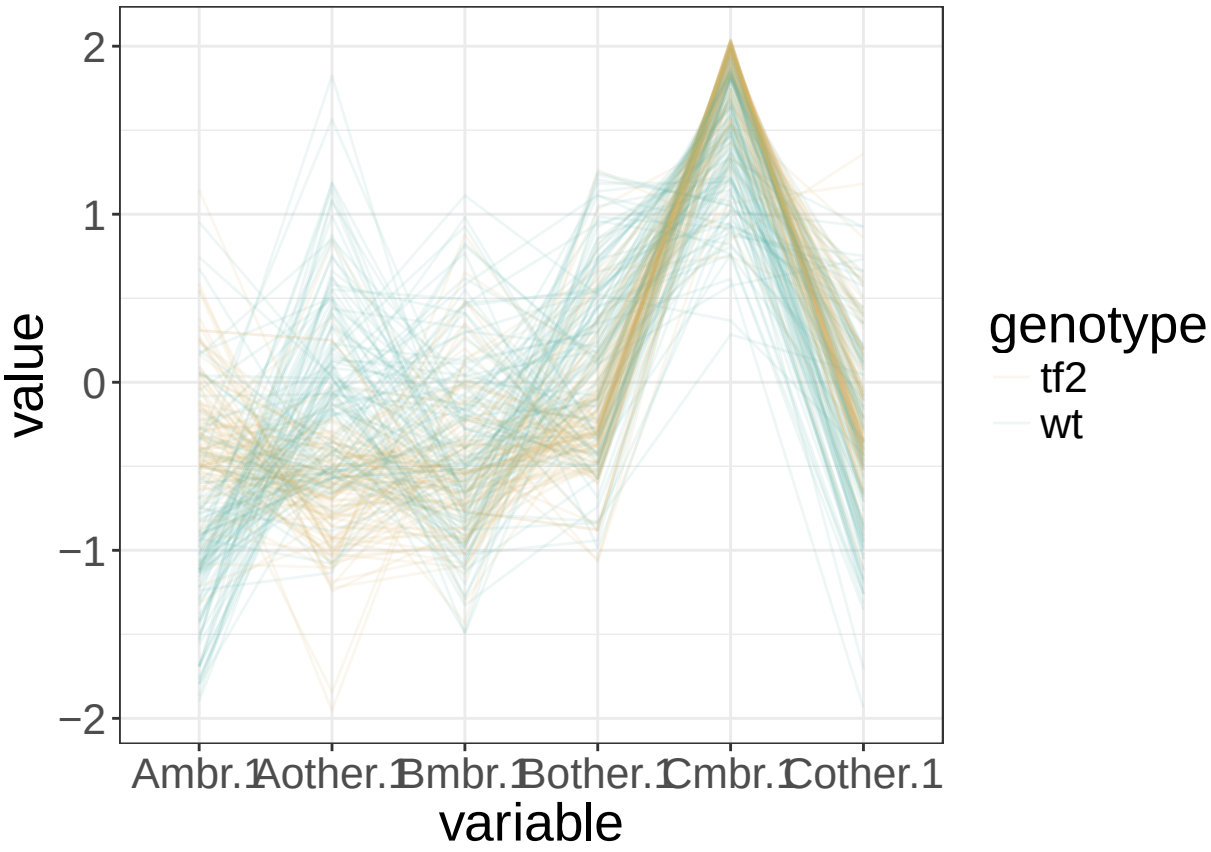

```
clusterVis_region_ssom(23)
```

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




```
clusterVis_line_ssom(23)
```

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



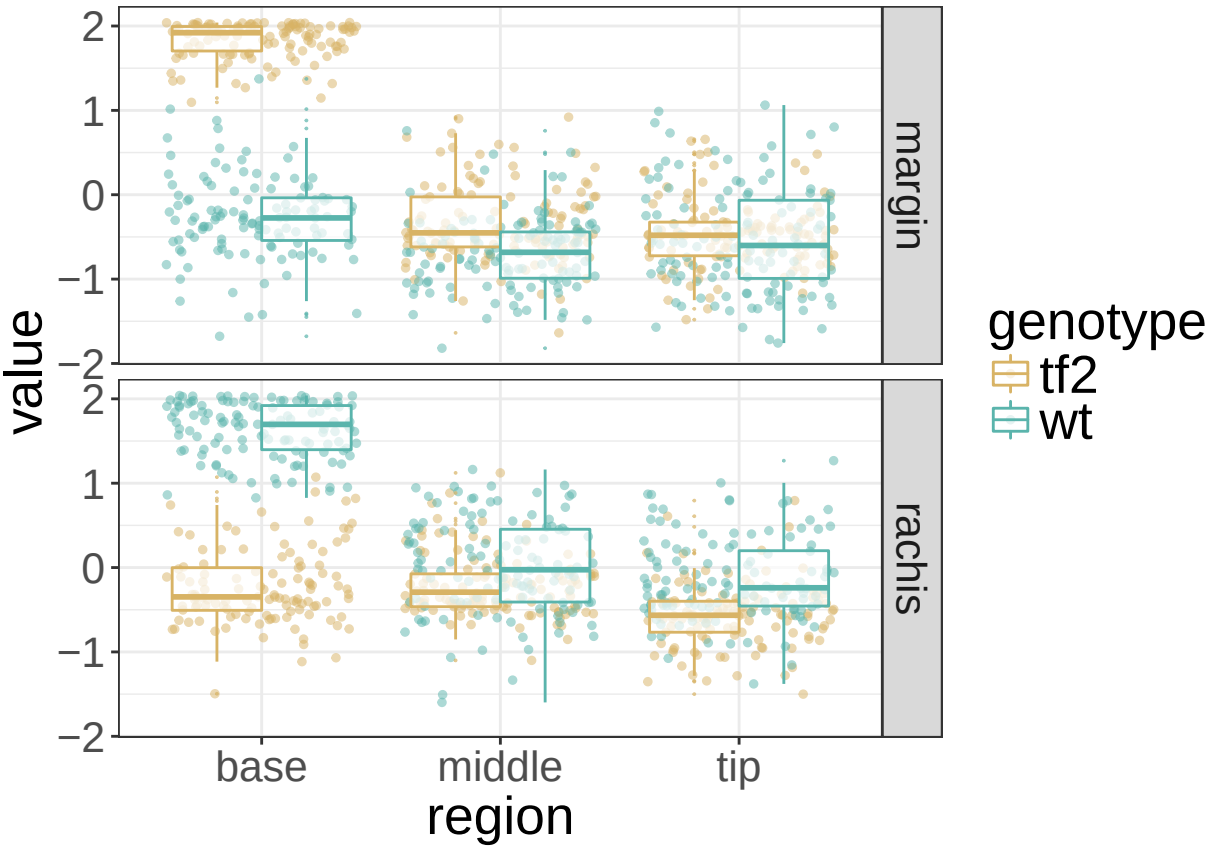
Cluster 24

Up

ACS8 Encodes an auxin inducible ACC synthase. **NA** similar to unknown protein [Arabidopsis thaliana] (TAIR:AT1G11990.1); similar to putative auxin-independent growth promoter [Oryza sativa (japonica cultivar-group)] (GB:BAD37877.1); **NA** transporter-related; similar to carbohydrate transporter/ sugar porter [Arabidopsis thaliana]

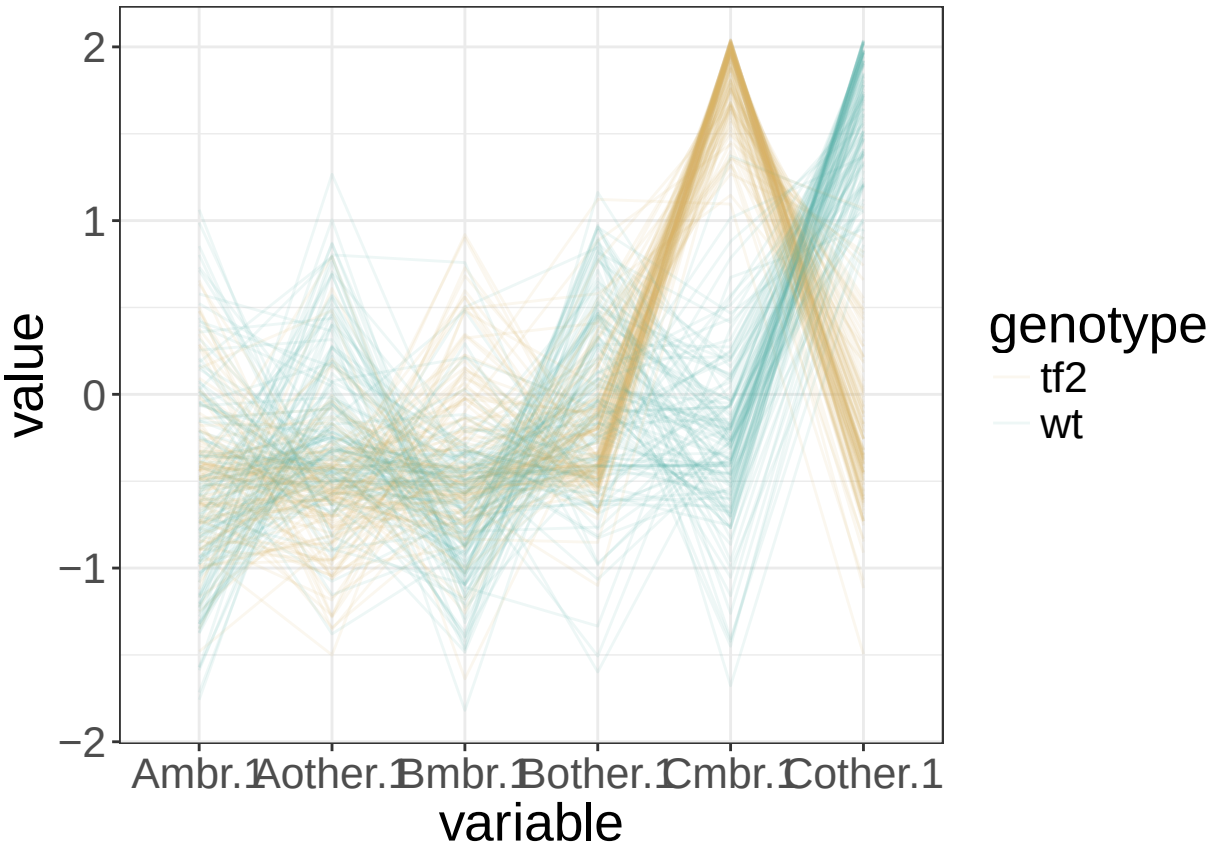
```
clusterVis_region_ssom(24)
```

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



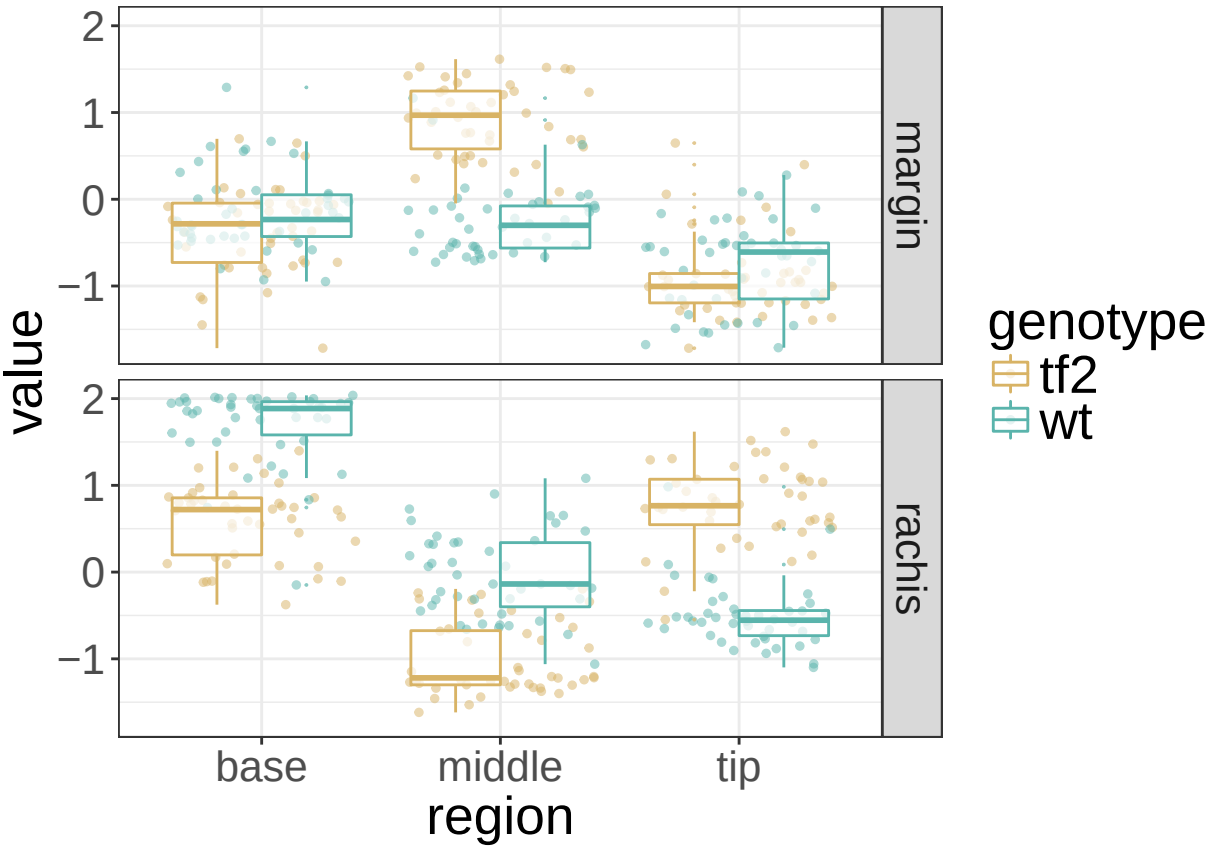

```
clusterVis_line_ssom(24)
```

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



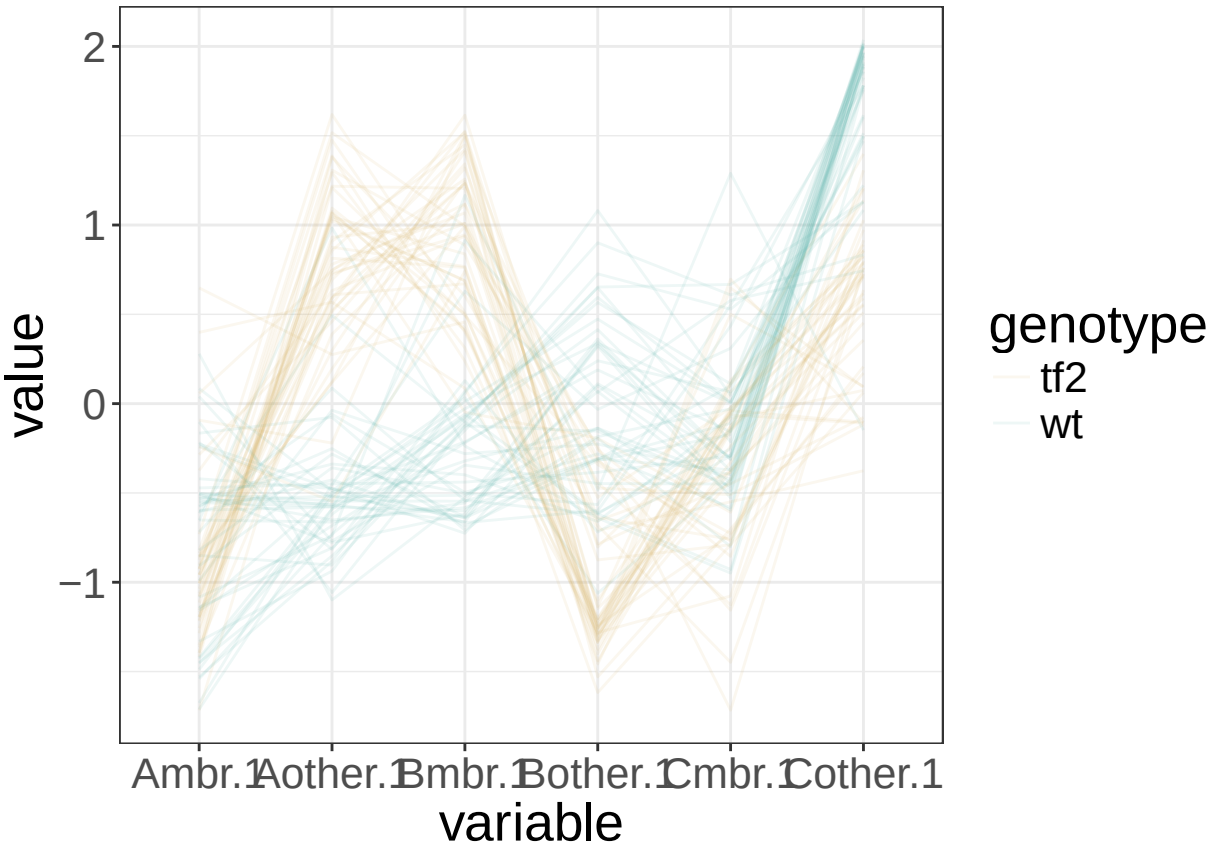

```
clusterVis_region_ssom(25)
```

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



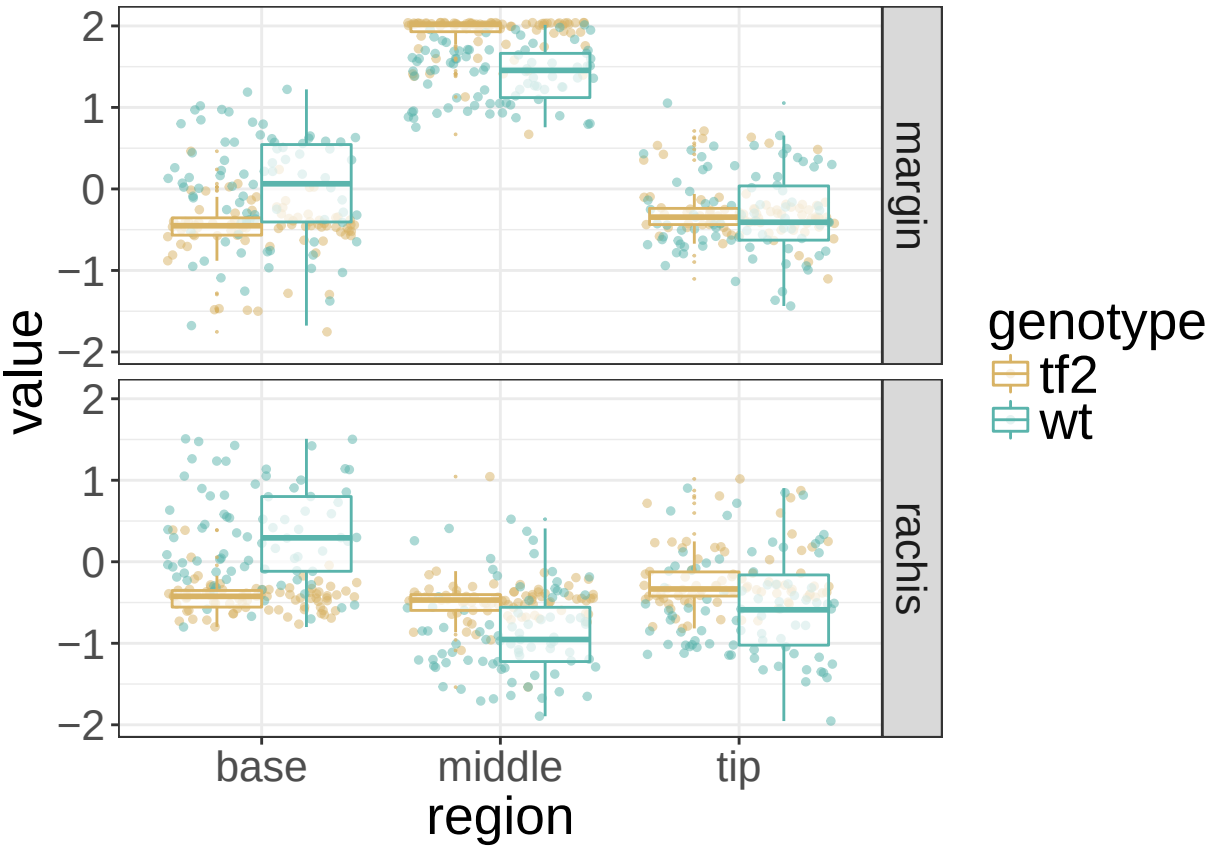

```
clusterVis_line_ssom(25)
```

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



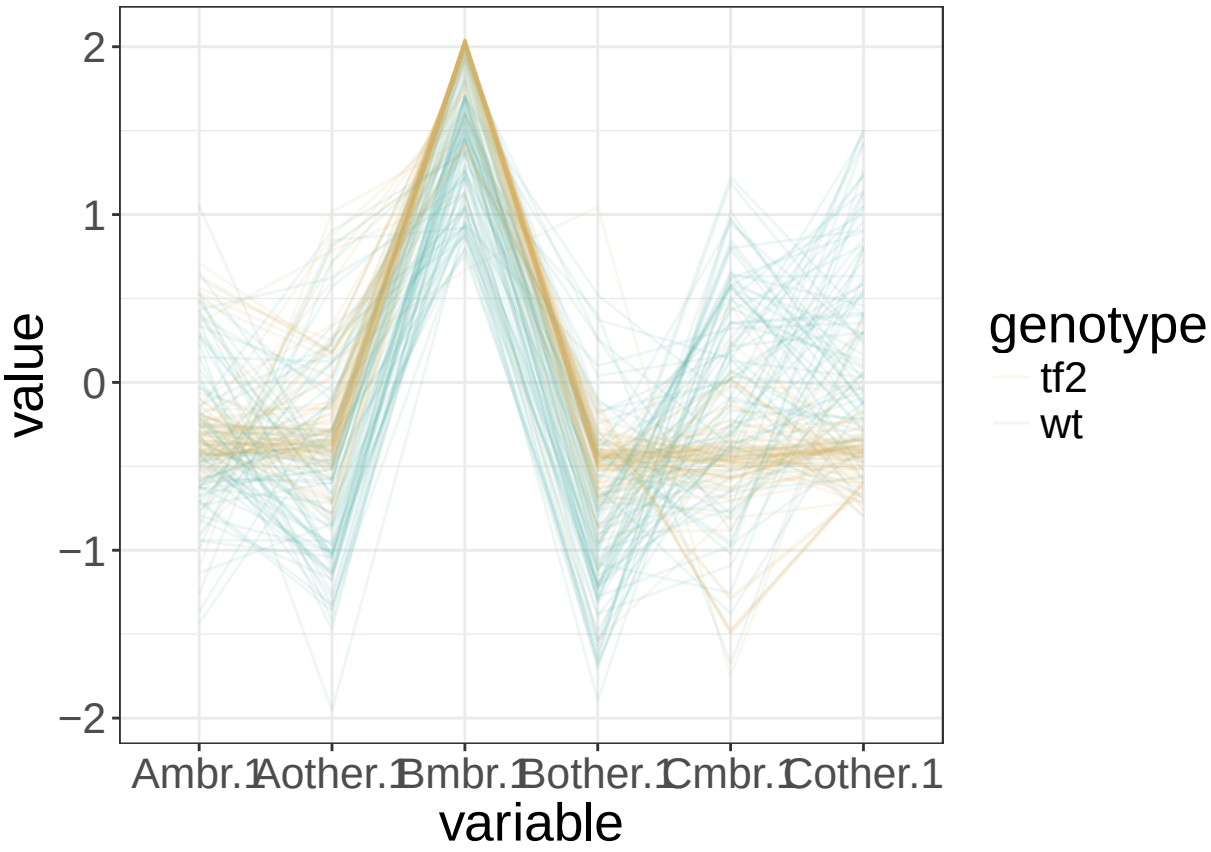

```
clusterVis_region_ssom(26)
```

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



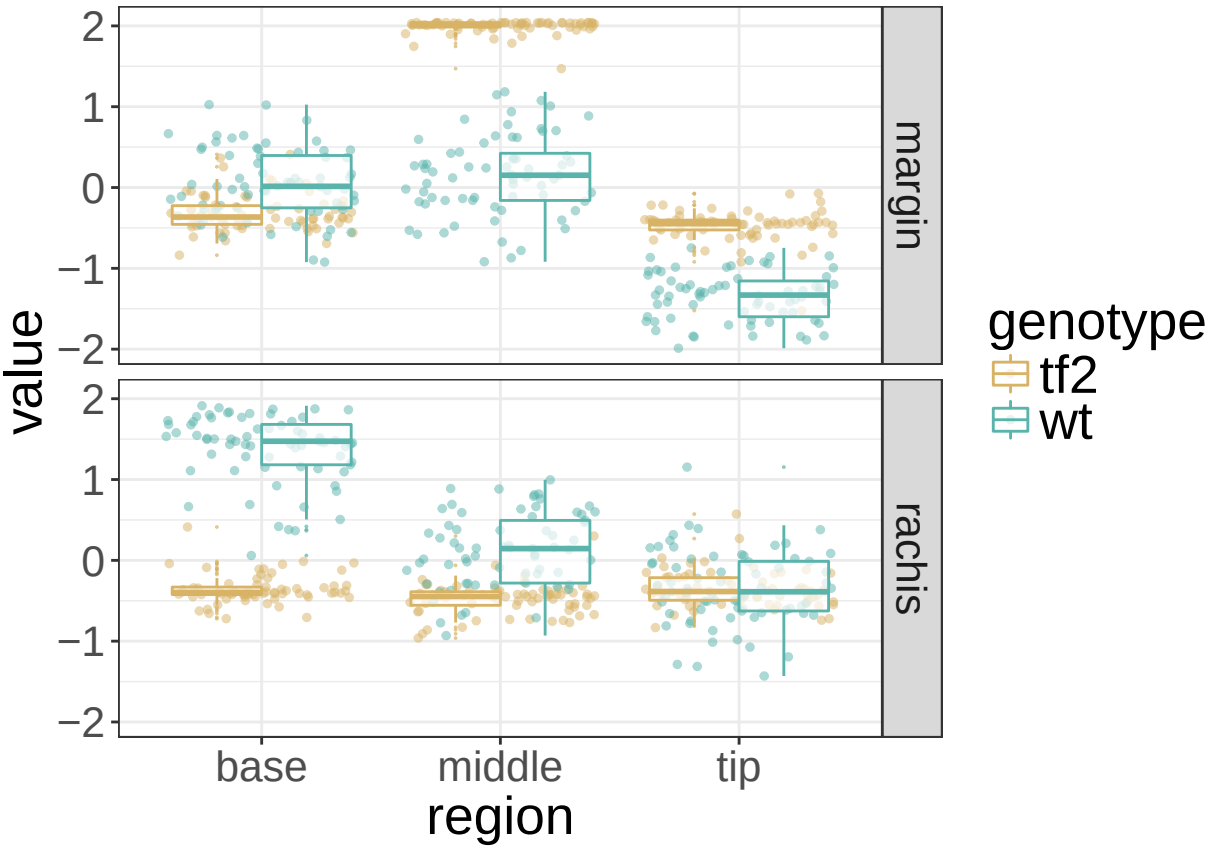

```
clusterVis_line_ssom(26)
```

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



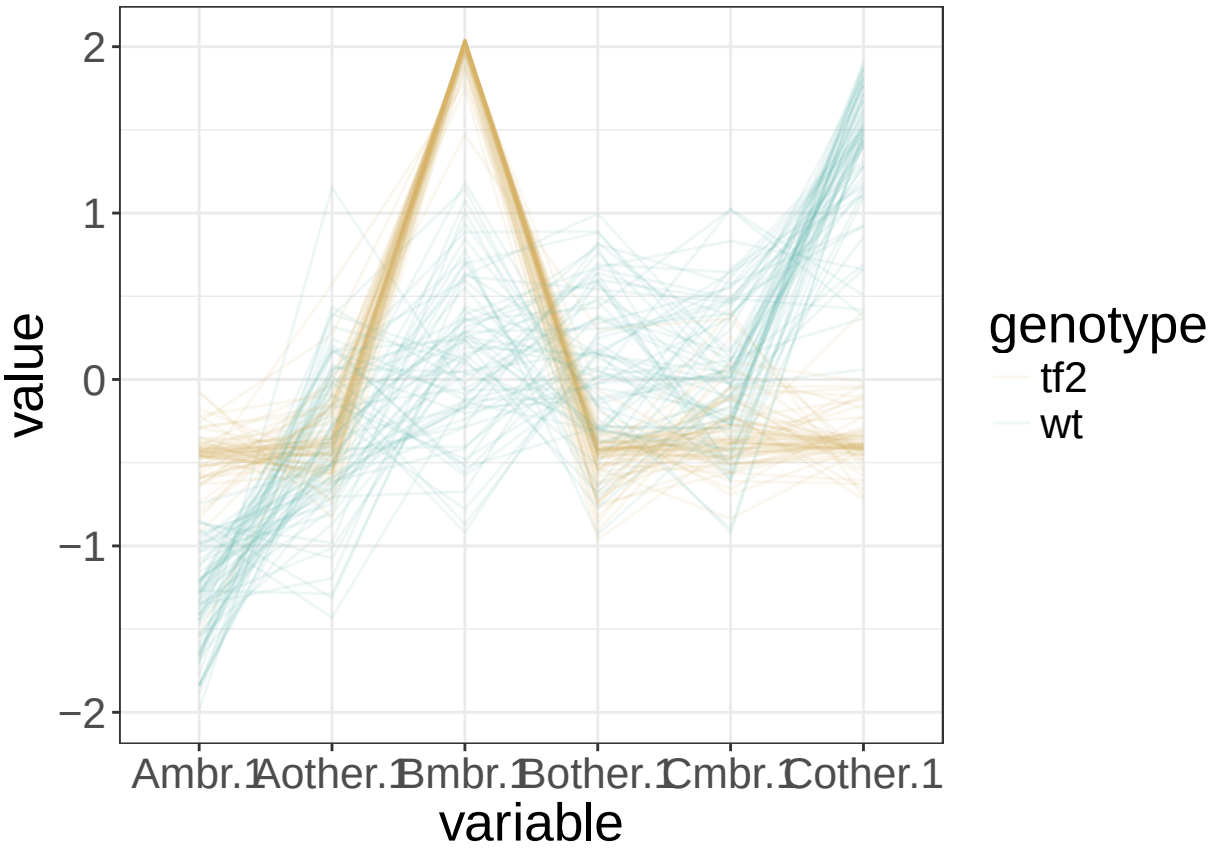

```
clusterVis_region_ssom(27)
```

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



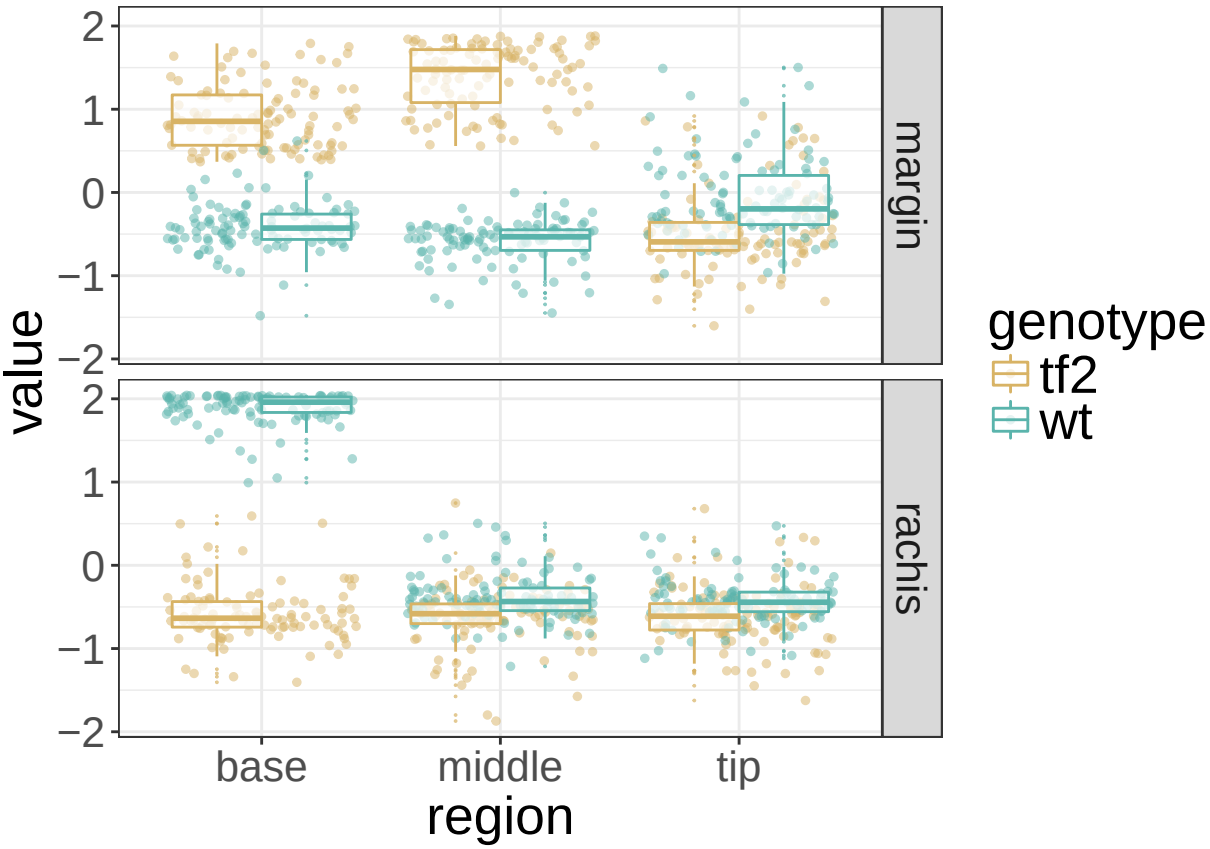

```
clusterVis_line_ssom(27)
```

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



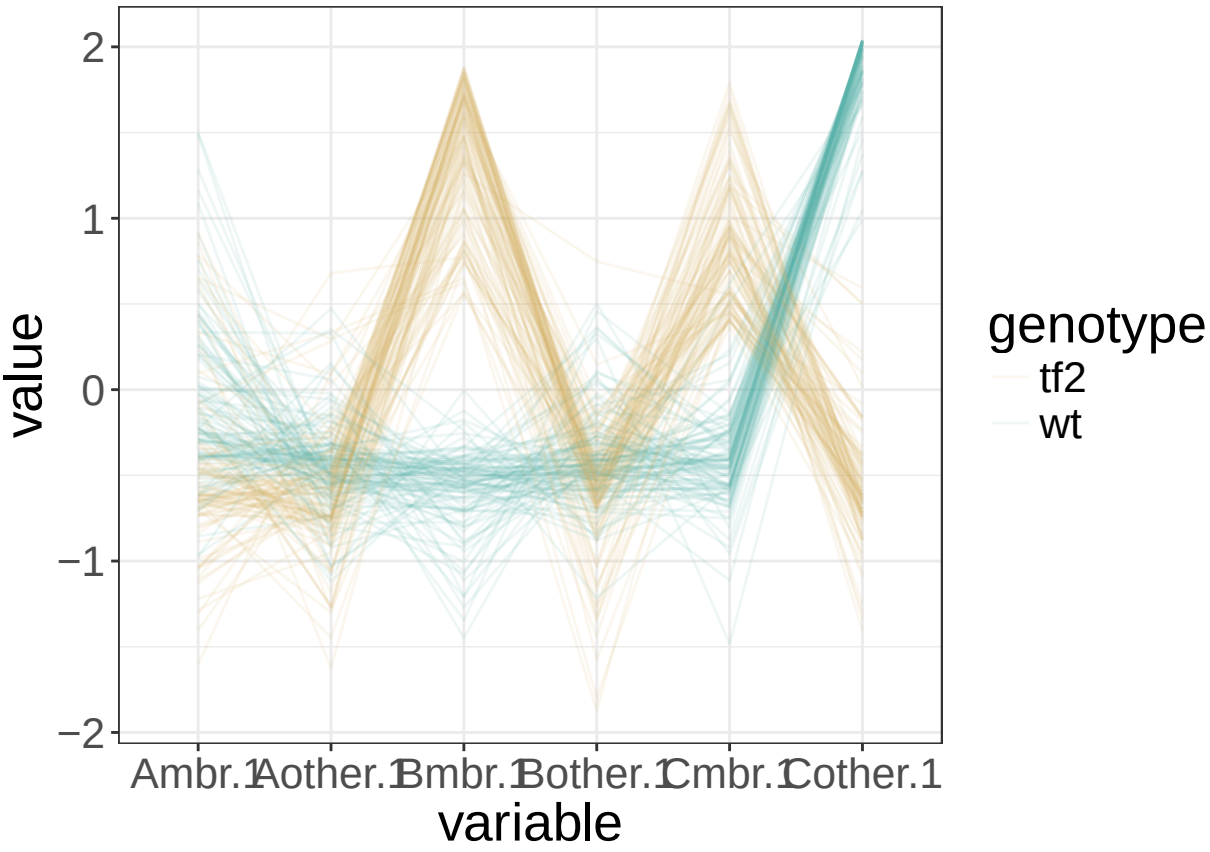

```
clusterVis_region_ssom(28)
```

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




```
clusterVis_line_ssom(28)
```

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

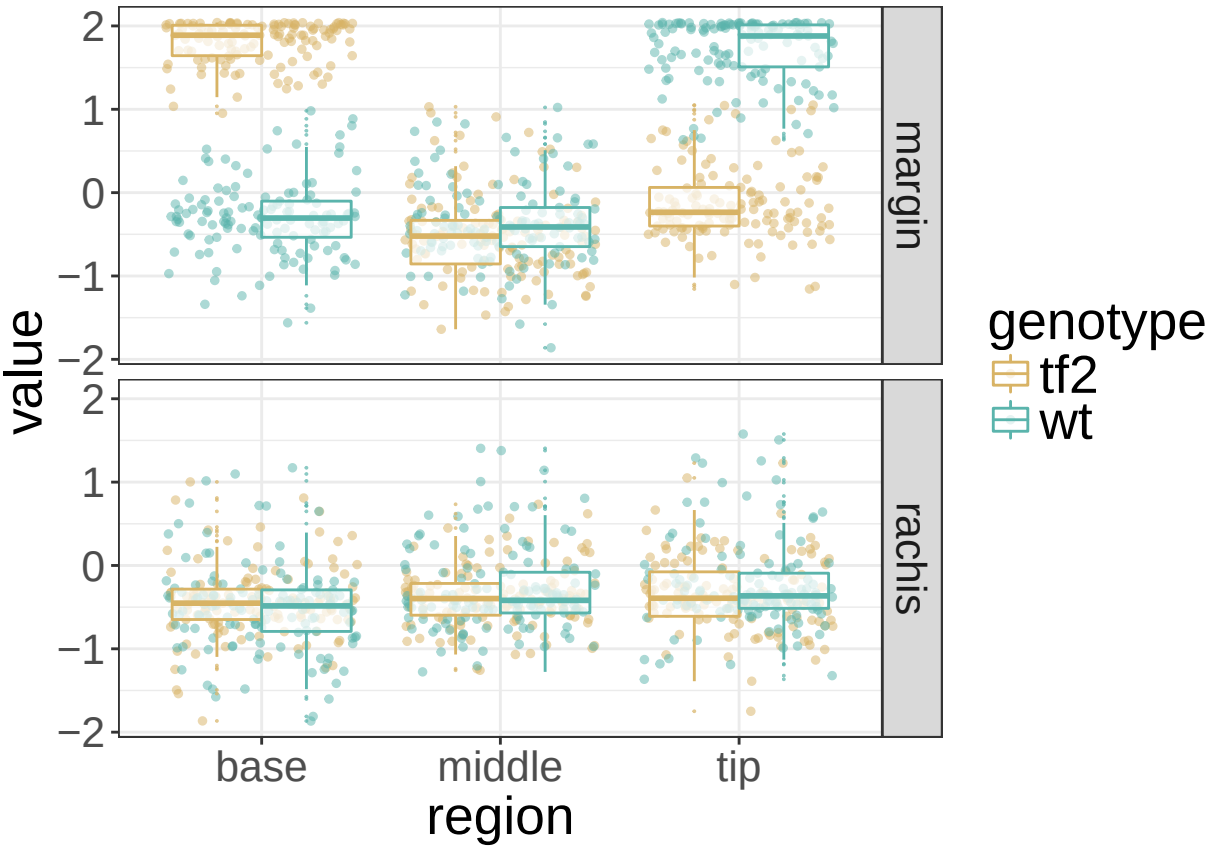


Cluster 29

AtMYB93 Member of the R2R3 factor gene family. ATAUX2-11 Auxin inducible protein similar to transcription factors.

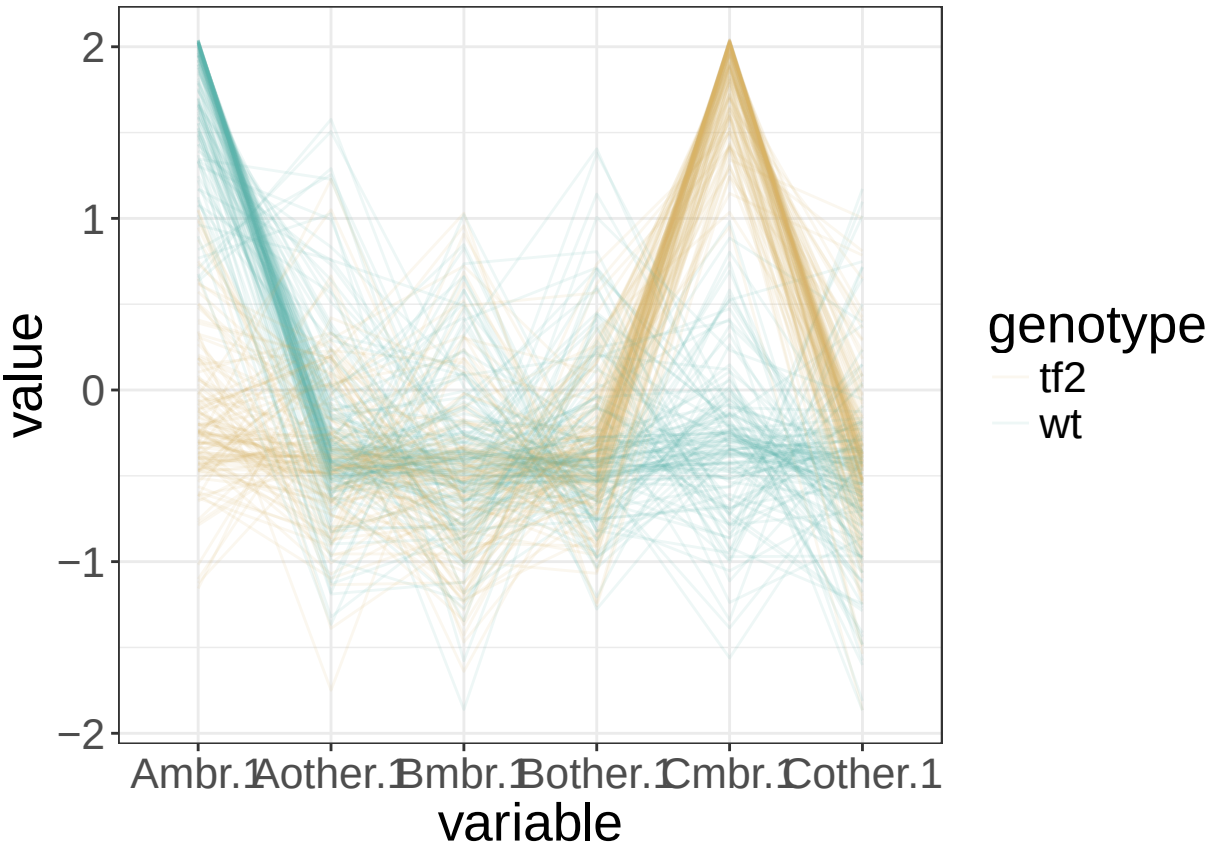
```
clusterVis_region_ssom(29)
```

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




```
clusterVis_line_ssom(29)
```

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



Cluster 30

AGO7 Encodes ARGONAUTE7, a member of the ARGONAUTE family, characterised by the presence of PAZ and PIWI domains. Involved in the regulation of developmental timing. Required for the accumulation of TAS3 ta-siRNAs but not for accumulation of miR171, miR173, miR390 or mi391. Localized in mature rosette leaves and floral buds.

ATGA2OX4 Encodes a gibberellin 2-oxidase. AtGA2OX4 expression is responsive to cytokinin and KNOX activities.

ATDTX35 MATE efflux family protein; similar to MATE efflux family protein [Arabidopsis thaliana] (TAIR:AT4G00350.1);

AP2 Encodes a floral homeotic gene, a member of the AP2/EREBP (ethylene responsive element binding protein) class of transcription factors and is involved in the specification of floral organ identity, establishment of floral meristem identity, suppression of floral meristem indeterminancy, and development of the ovule and seed coat. AP2 also has a role in controlling seed mass. Dominant negative allele I28, revealed a function in meristem maintenance-mutant meristems are smaller than normal siblings. AP2 appears to act on the WUS-CLV pathway in an AG independent manner.

NA F-box family protein; similar to F-box family protein [Arabidopsis thaliana] (TAIR:AT5G51380.1); similar to Os02g0658500 [Oryza sativa (japonica cultivar-group)] (GB:NP_001047636.1); similar to Os11g0641200 [Oryza sativa (japonica cultivar-group)] (GB:NP_001068351.1); similar to Leucine Rich Repeat family protein, expressed [Oryza sativa (japonica cultivar-group)] (GB:ABA95013.1); contains InterPro domain Cyclin-like F-box; (InterPro:IPR001810)

EMB1006 EMB1006 (EMBRYO DEFECTIVE 1006); binding; similar to pentatricopeptide (PPR) repeat-containing protein [Arabidopsis thaliana] (TAIR:AT5G02860.1); similar to Putative indole-3-acetate beta-

glucosyltransferase [Oryza sativa (japonica cultivar-group)] (GB:AAM15782.1); similar to Os05g0294600 [Oryza sativa (japonica cultivar-group)] (GB:NP_001055108.1); contains InterPro domain Pentatricopeptide repeat; (InterPro:IPR002885); contains InterPro domain Protein prenyltransferase; (InterPro:IPR008940); contains InterPro domain Tetratricopeptide-like helical; (InterPro:IPR011990)

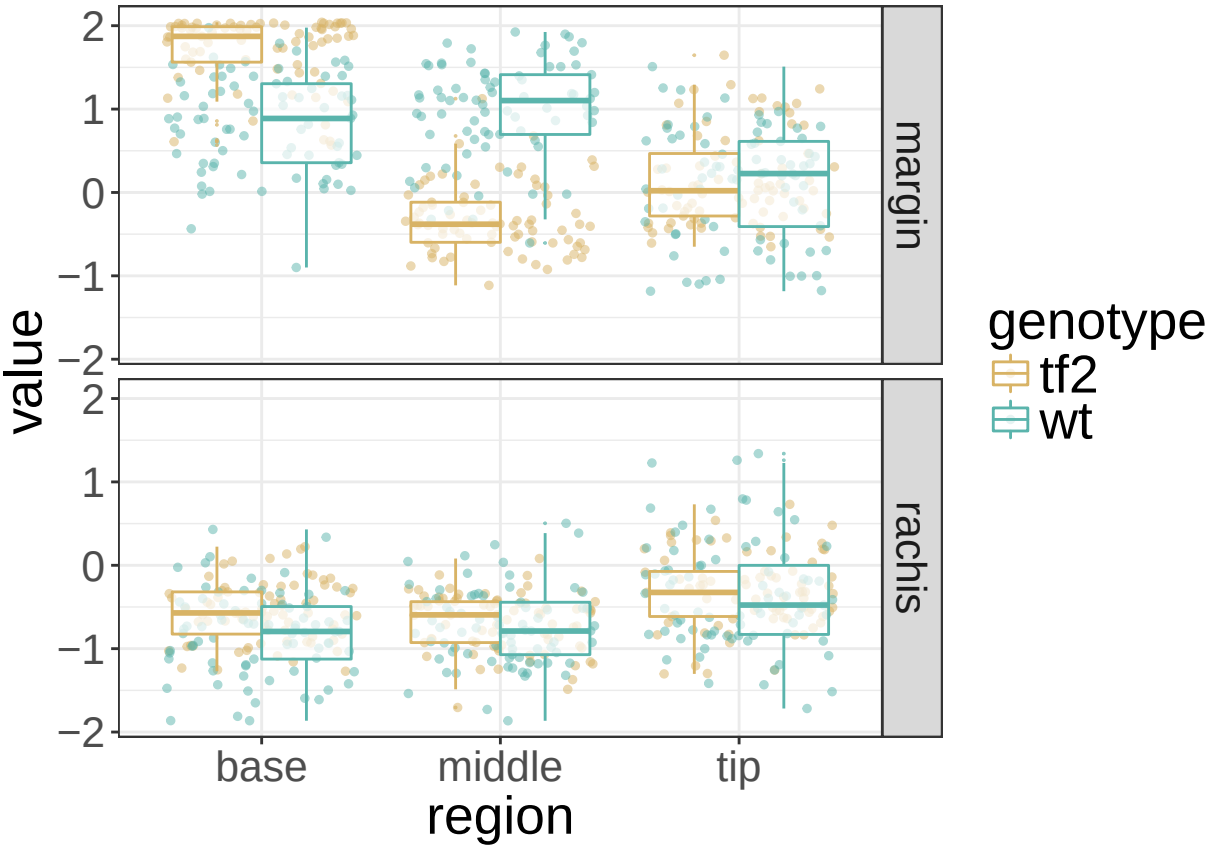
ARP3 Mutant has defect in trichome cell expansion and actin organization resulting in a distorted trichome phenotype.

YUC4 Belongs to the YUC gene family. Encodes a predicted flavin monooxygenase YUC4 involved in auxin biosynthesis and plant development.

ATGA2OX2 Encodes a gibberellin 2-oxidase. AtGA2OX2 expression is responsive to cytokinin and KNOX activities.

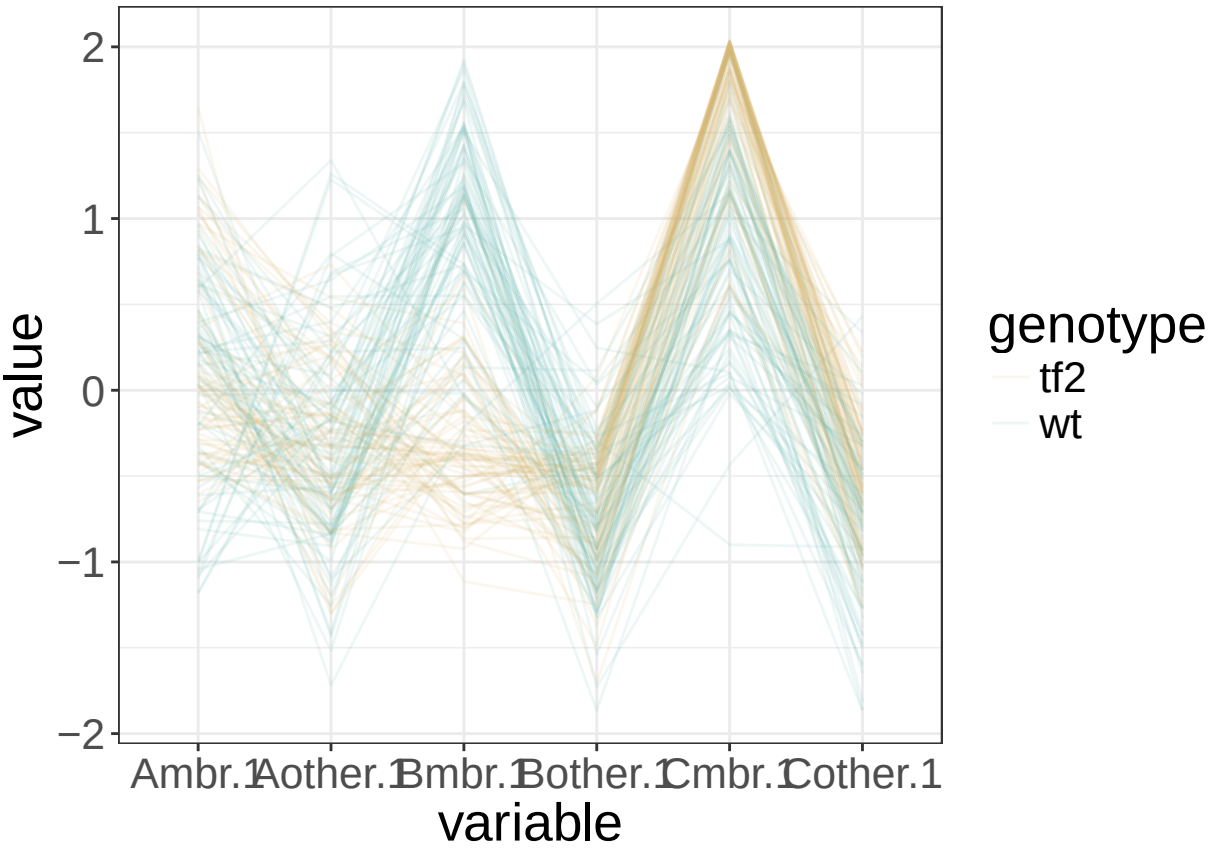
```
clusterVis_region_ssom(30)
```

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




```
clusterVis_line_ssom(30)
```

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



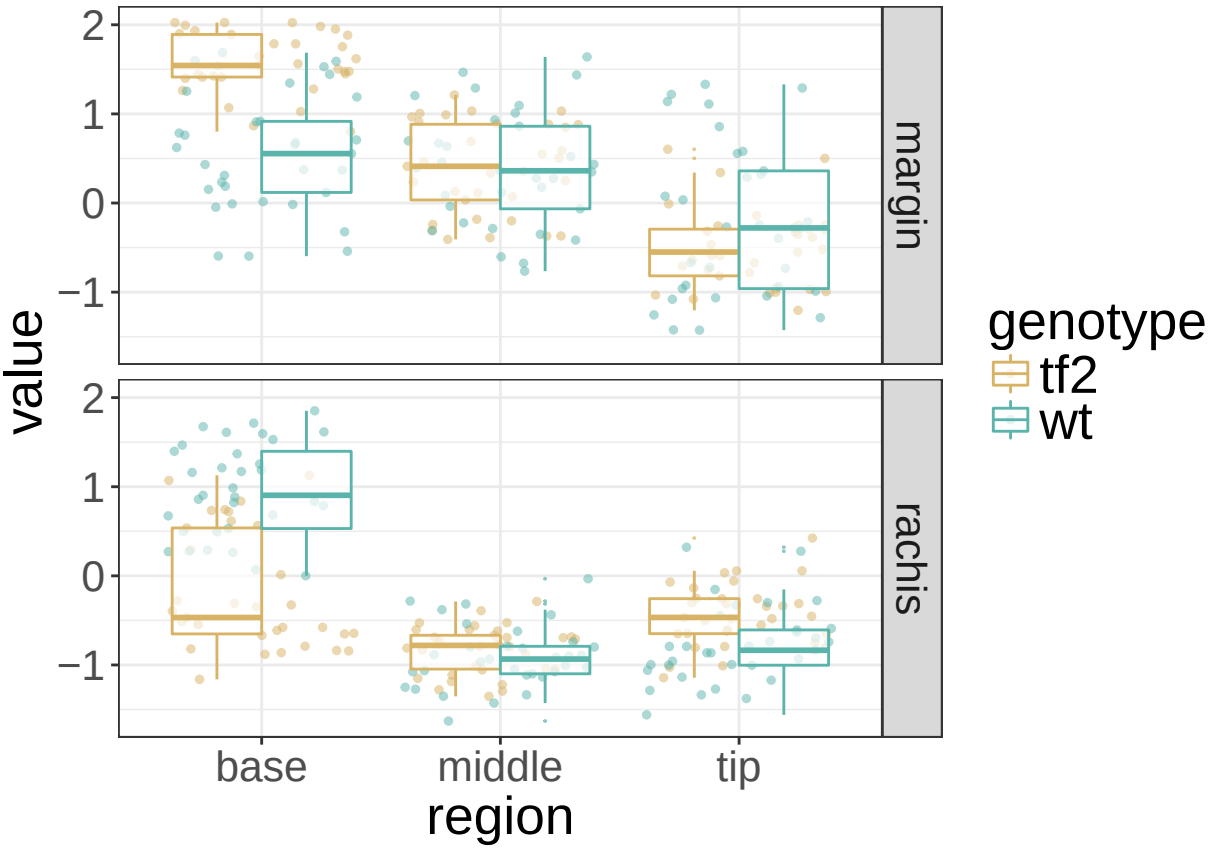
Cluster 31

SAP A recessive mutation in the Arabidopsis STERILE APETALA (SAP) causes severe aberrations in inflorescence and flower and ovule development.

AIL6 AIL6 (AINTEGUMENTA-LIKE 6); DNA binding / transcription factor; similar to AIL7 (AINTEGUMENTA-LIKE 7), DNA binding / transcription factor [Arabidopsis thaliana] (TAIR:AT5G65510.1); similar to 117M18_31 [Brassica rapa] (GB:AAZ66950.1); contains InterPro domain Pathogenesis-related transcriptional factor and ERF; (InterPro:IPR001471)

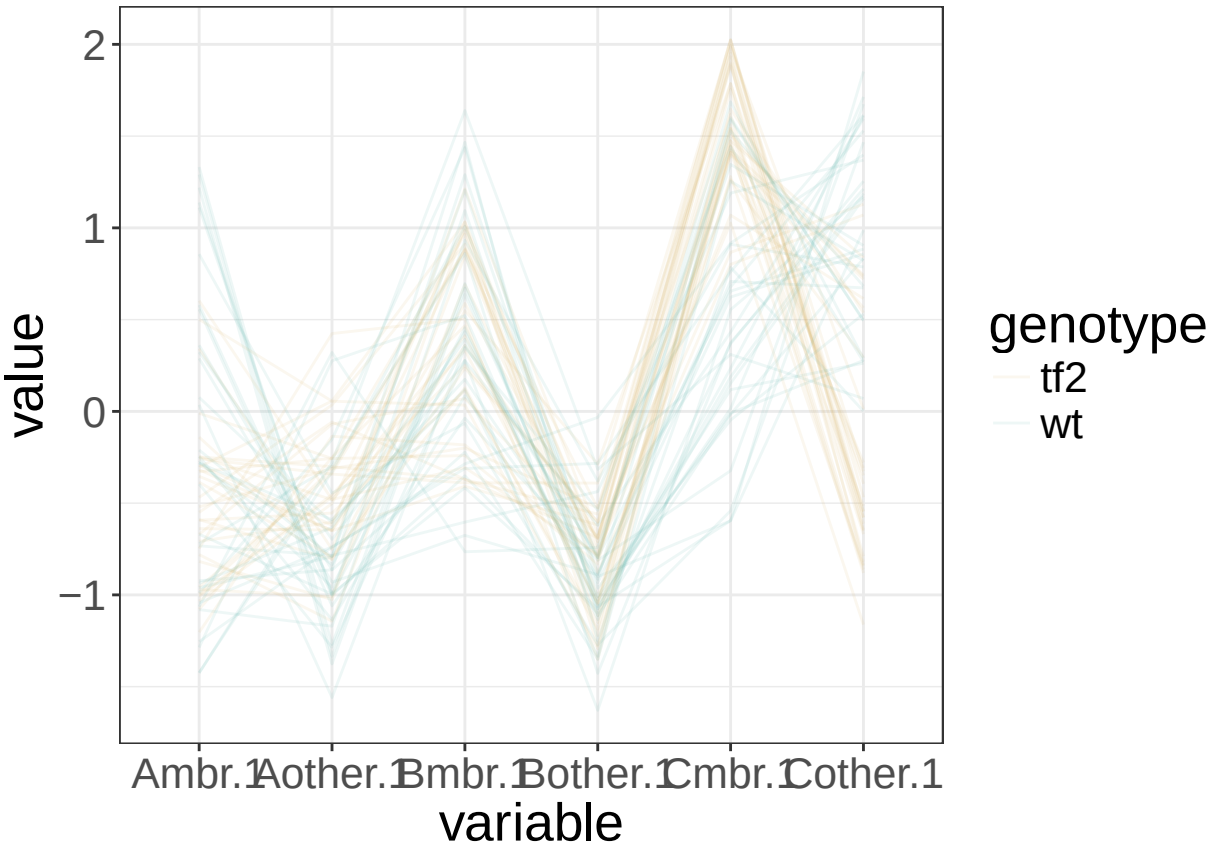
```
clusterVis_region_ssom(31)
```

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



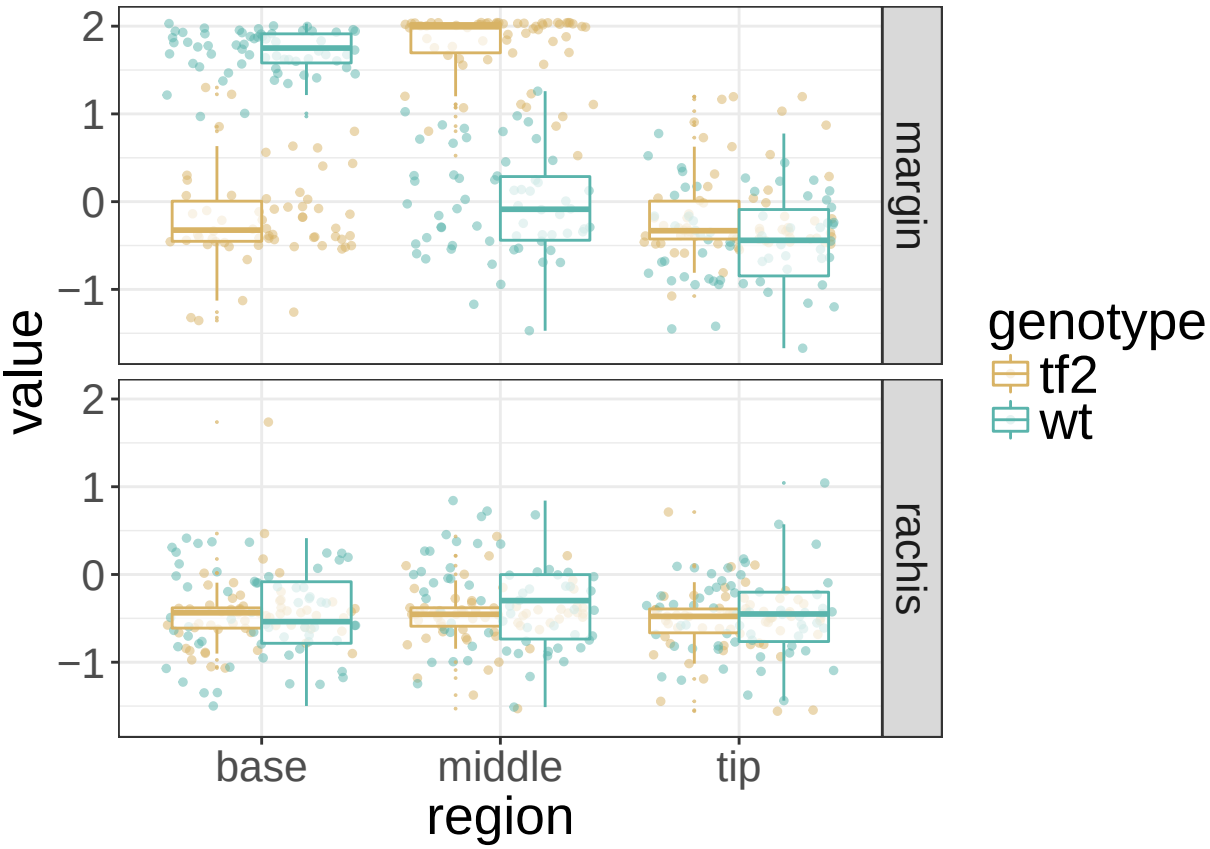

```
clusterVis_line_ssom(31)
```

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



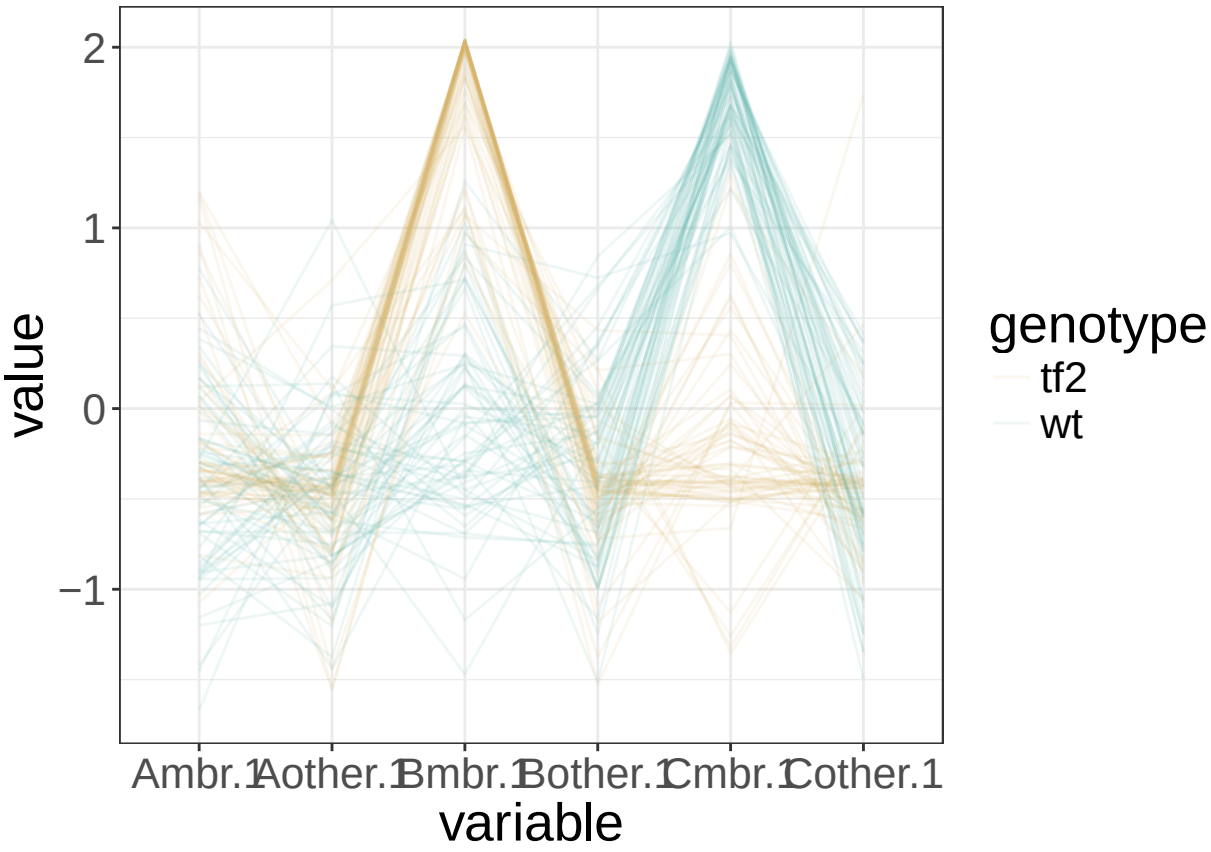

```
clusterVis_region_ssom(32)
```

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



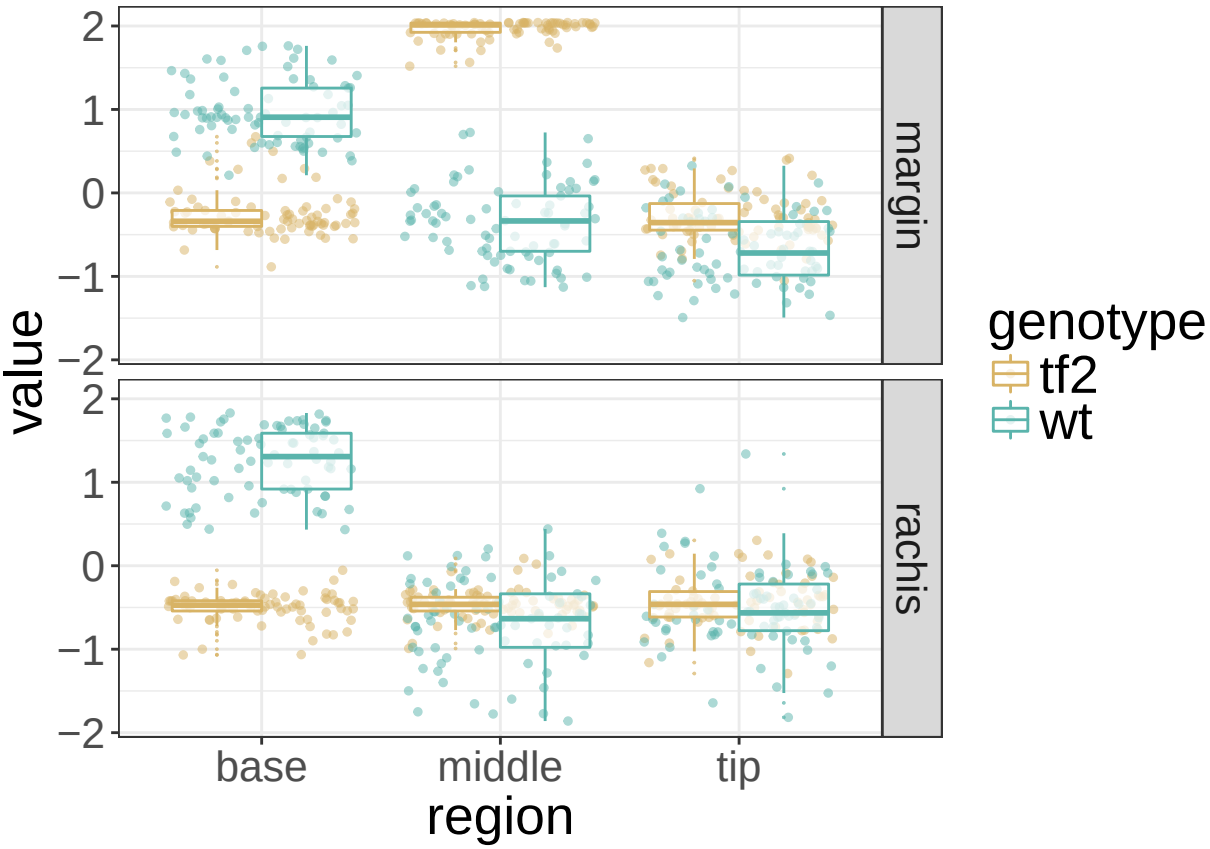

```
clusterVis_line_ssom(32)
```

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



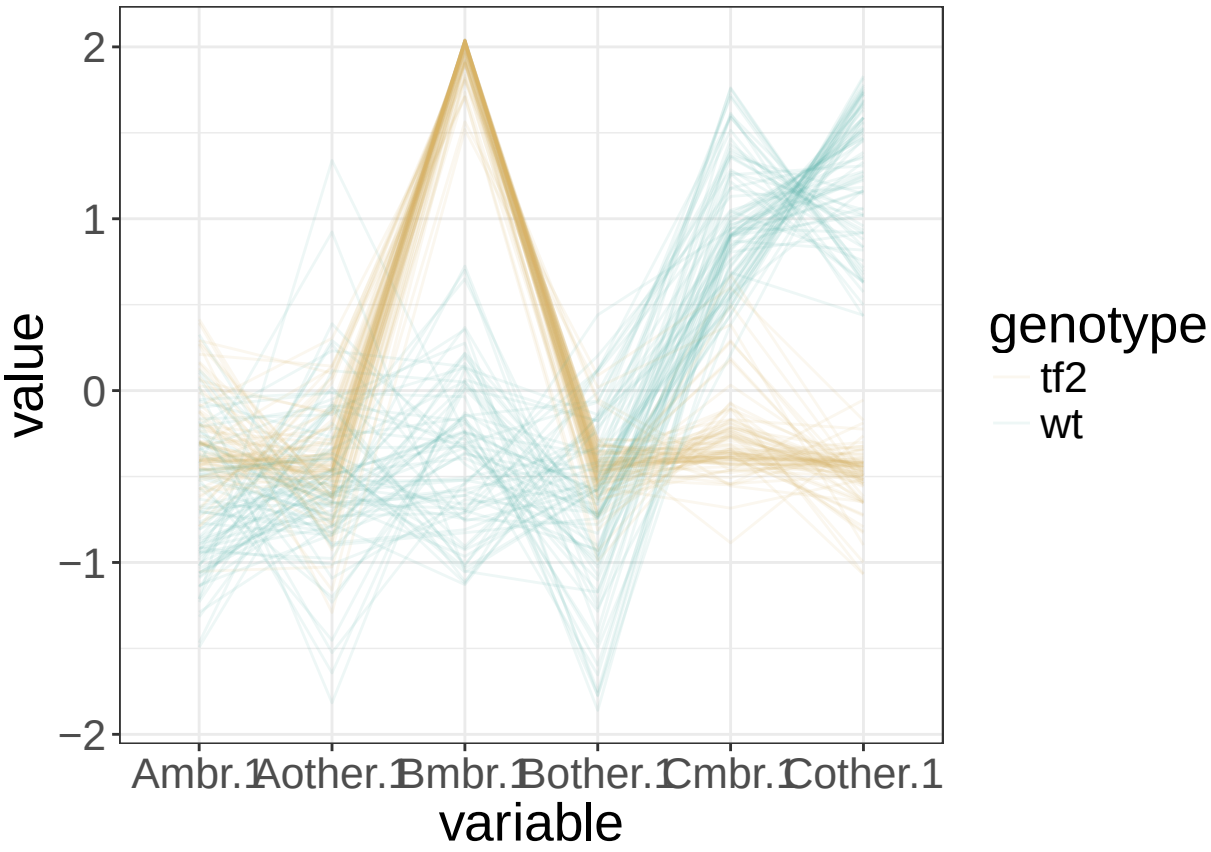

```
clusterVis_region_ssom(33)
```

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



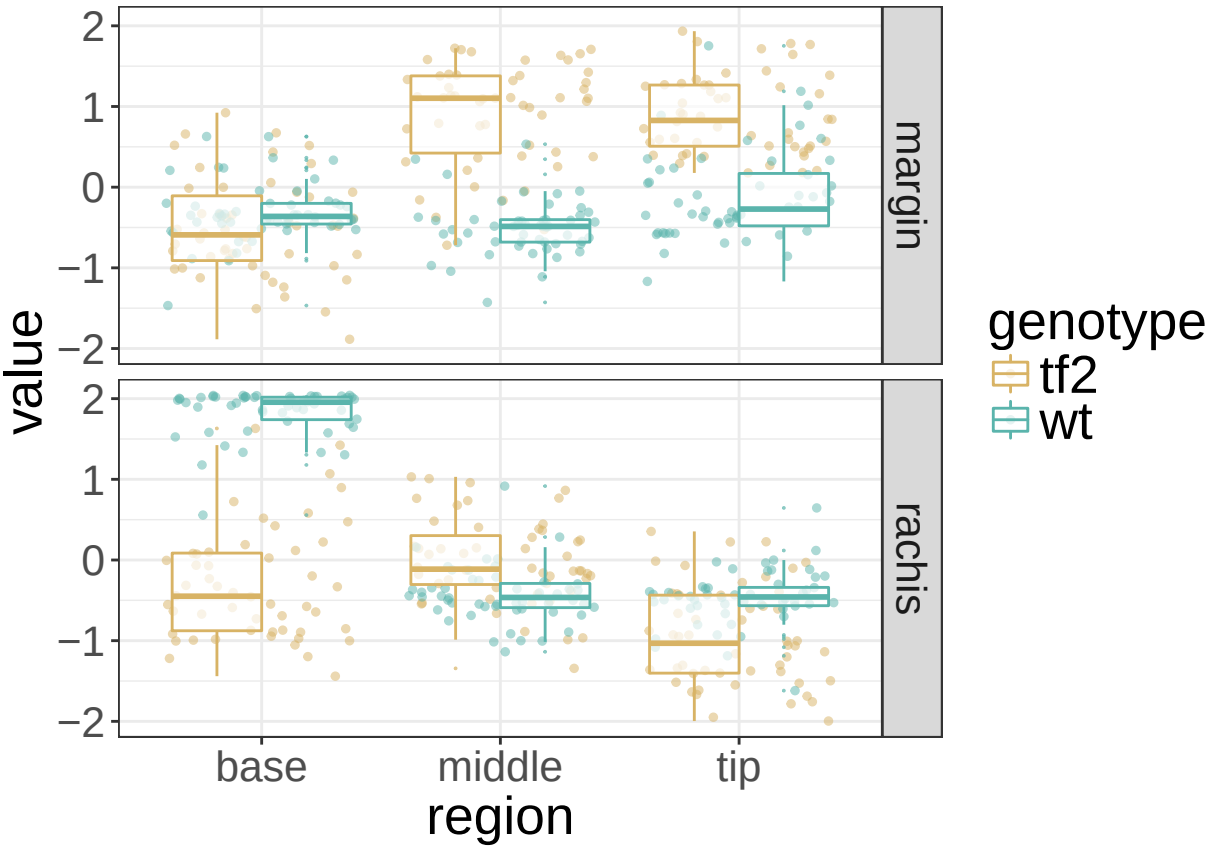

```
clusterVis_line_ssom(33)
```

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



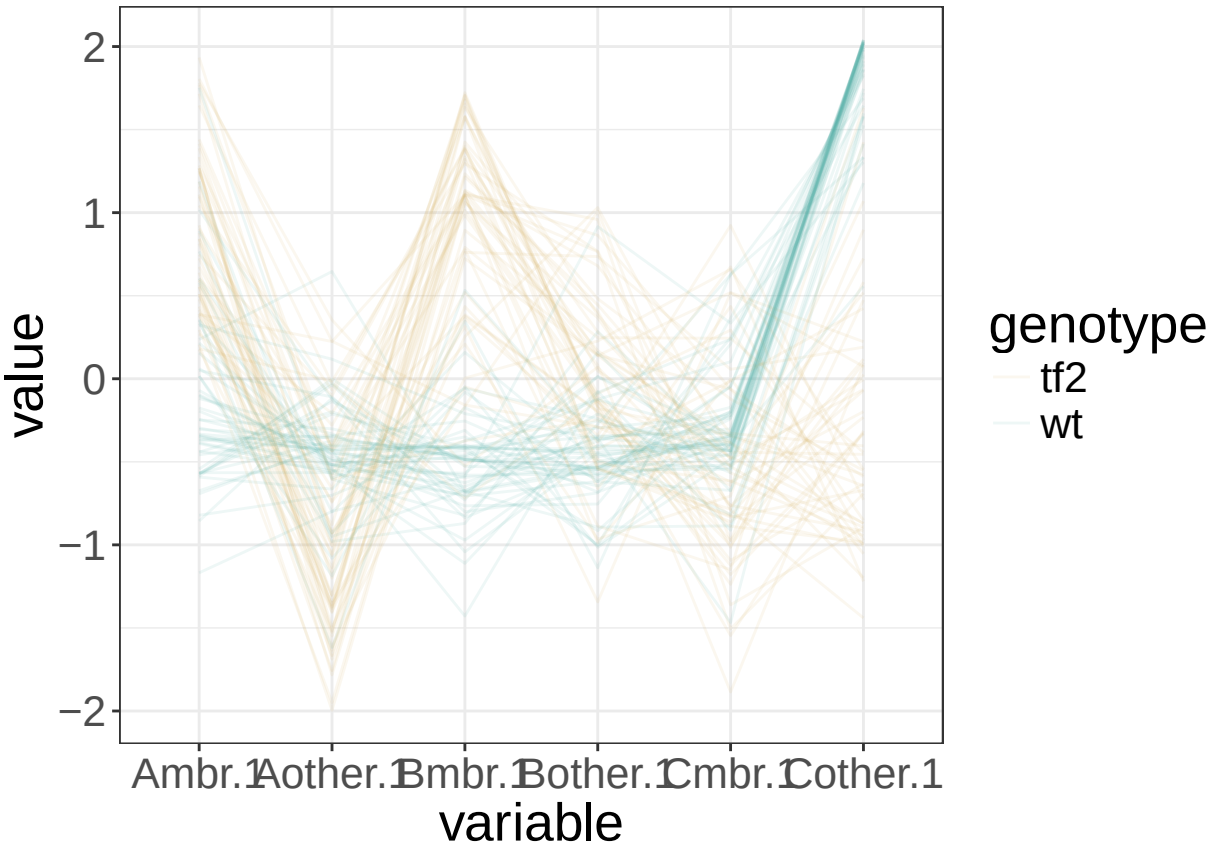

```
clusterVis_region_ssom(34)
```

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



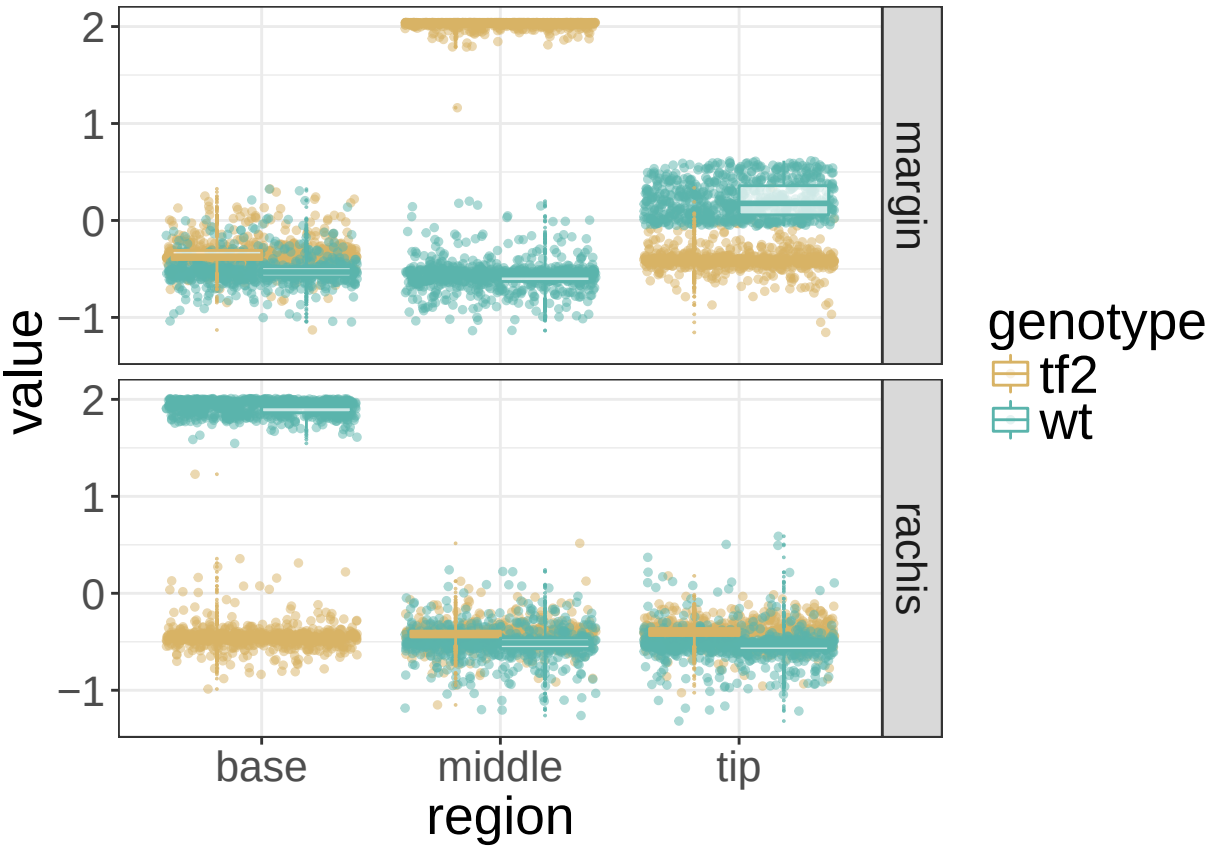

```
clusterVis_line_ssom(34)
```

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



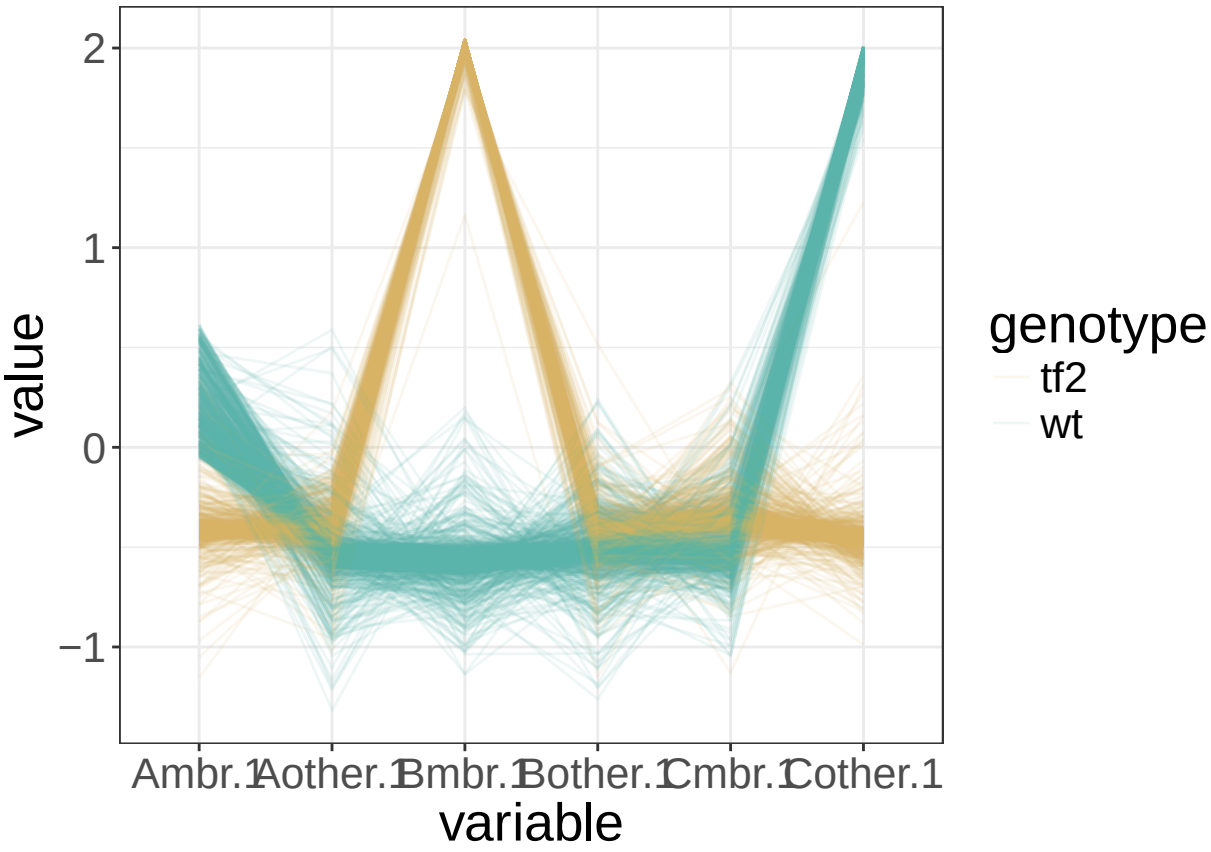

```
clusterVis_region_ssom(35)
```

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



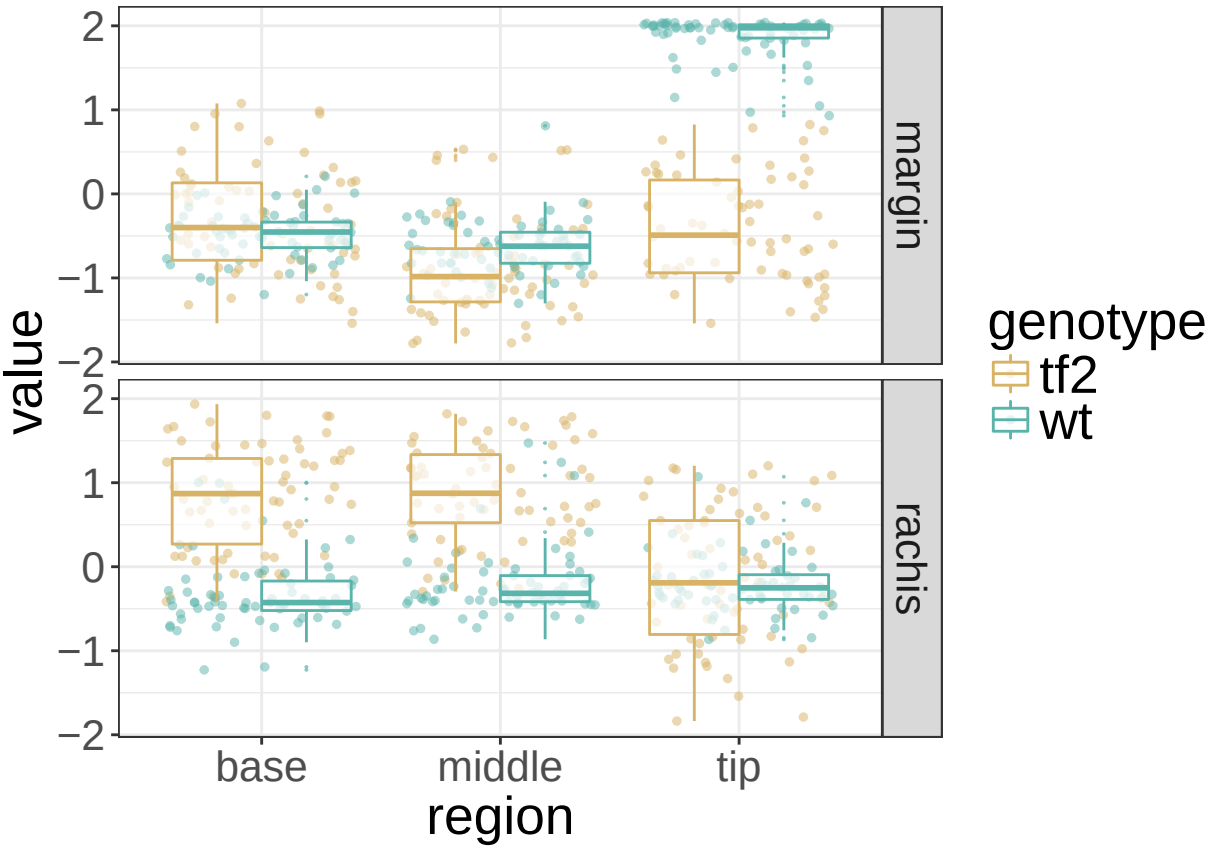

```
clusterVis_line_ssom(35)
```

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



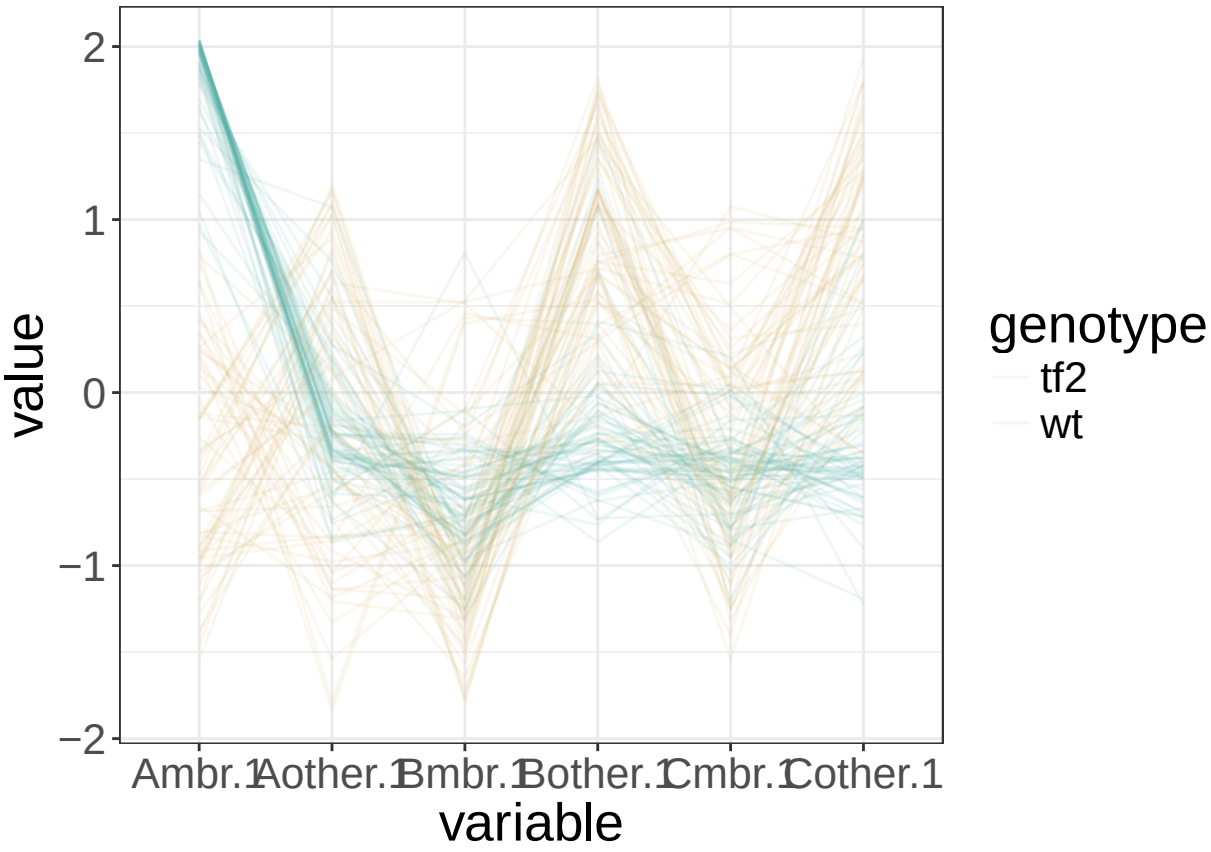

```
clusterVis_region_ssom(36)
```

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



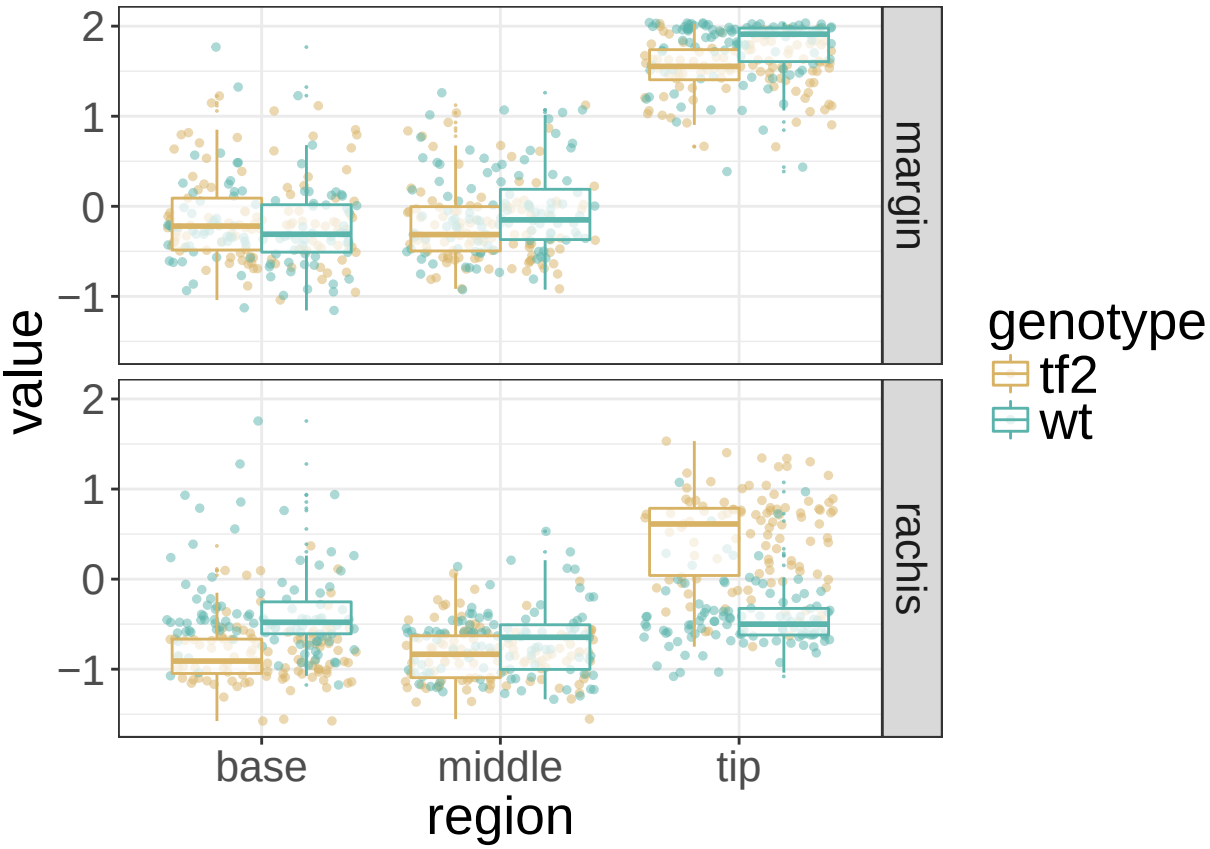

```
clusterVis_line_ssom(36)
```

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



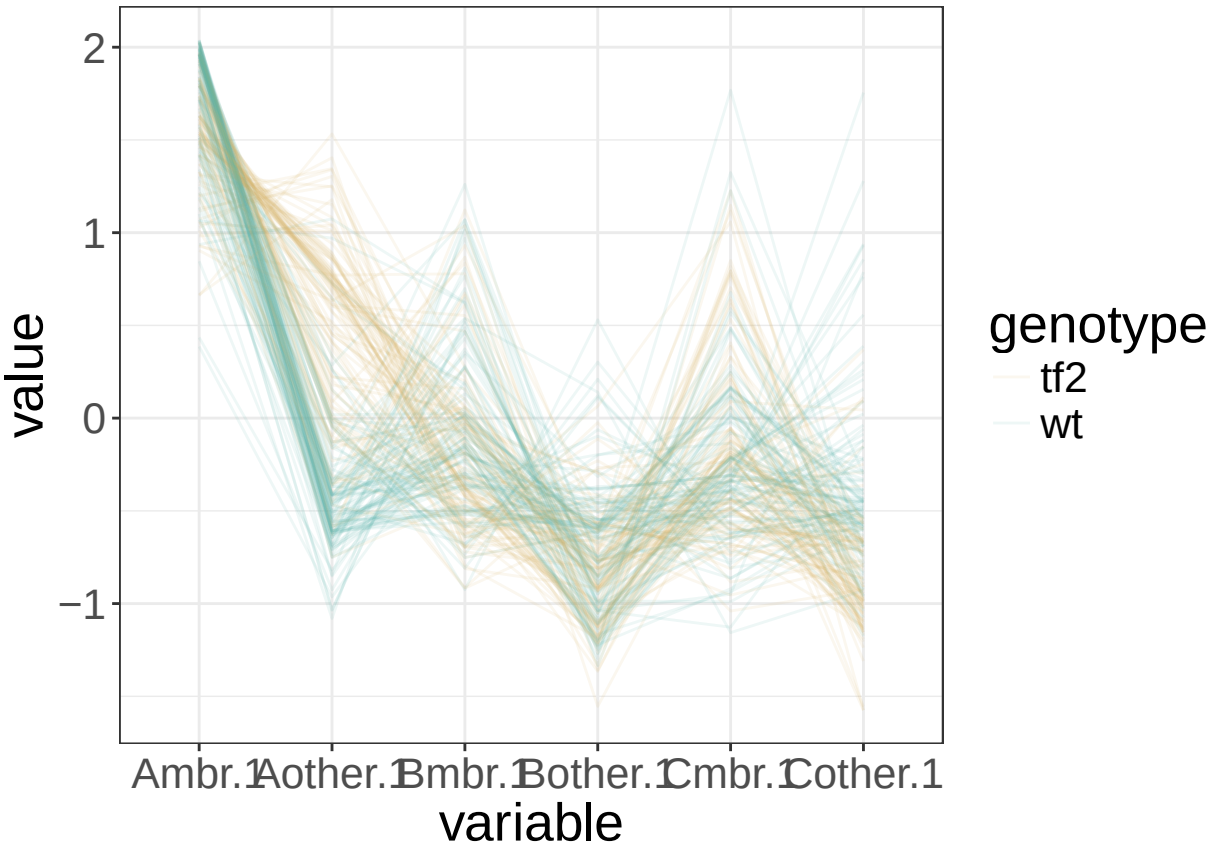

```
clusterVis_region_ssom(37)
```

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



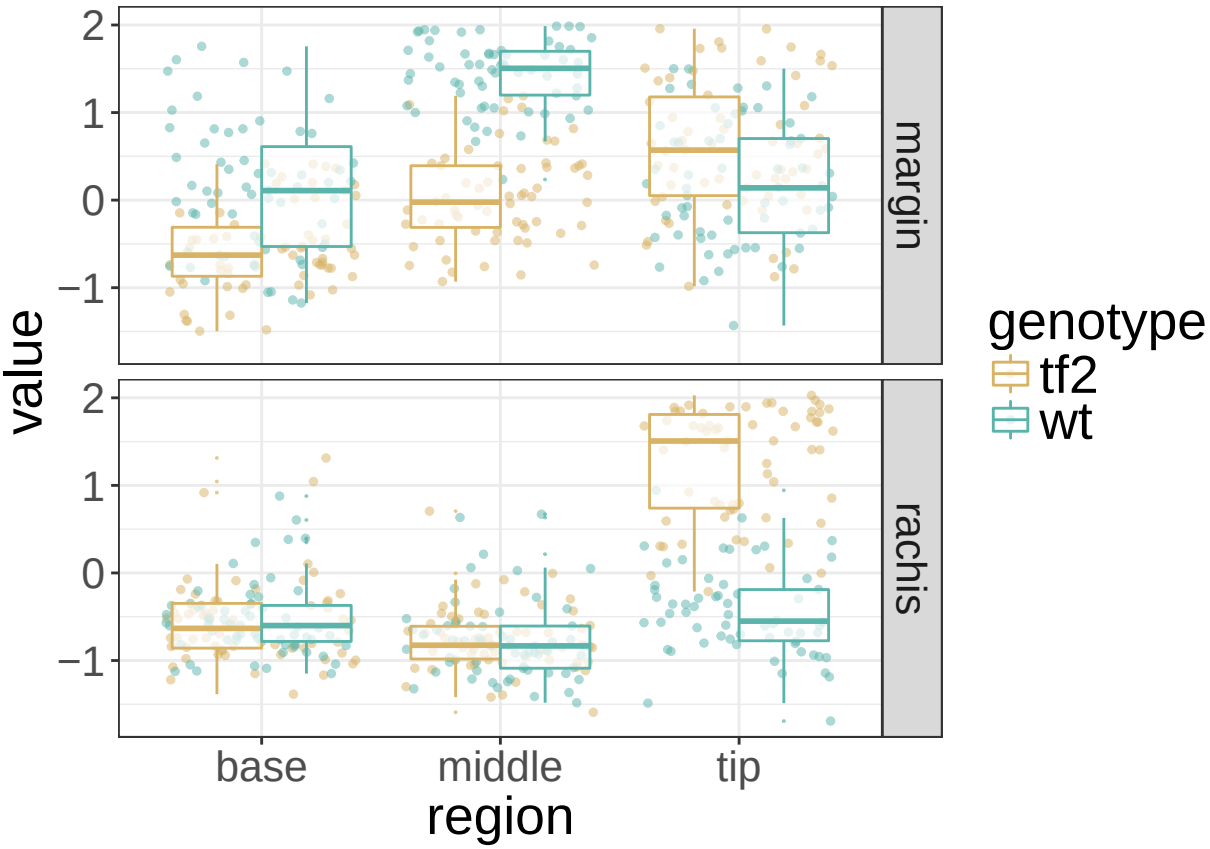

```
clusterVis_line_ssom(37)
```

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



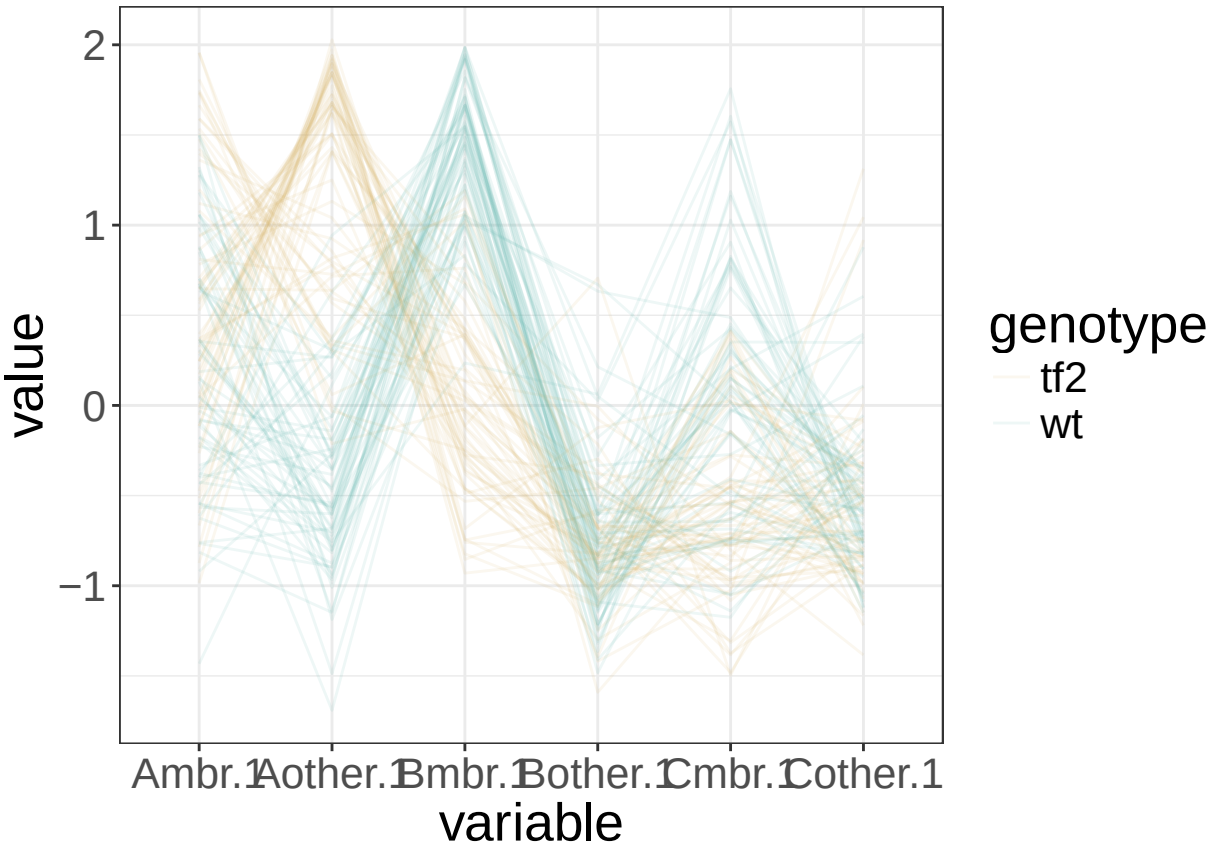

```
clusterVis_region_ssom(38)
```

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



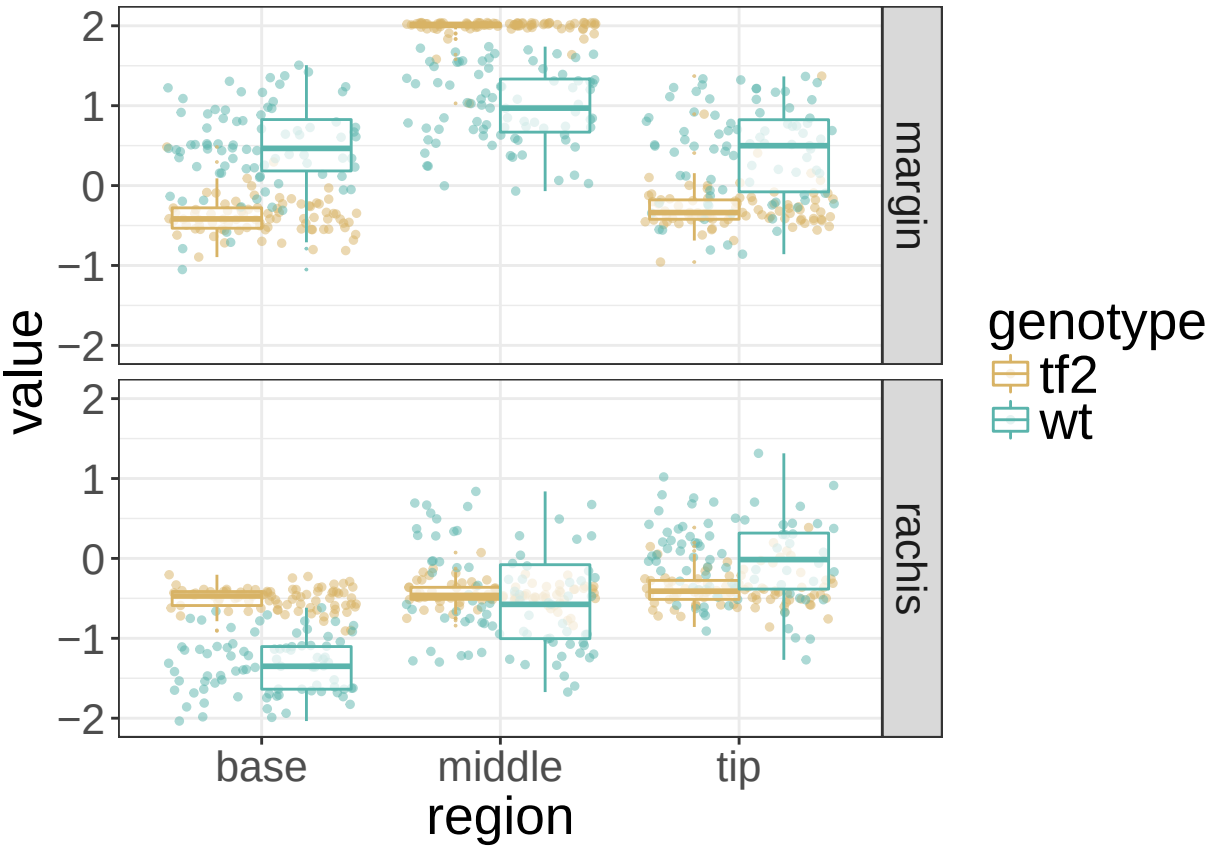

```
clusterVis_line_ssom(38)
```

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



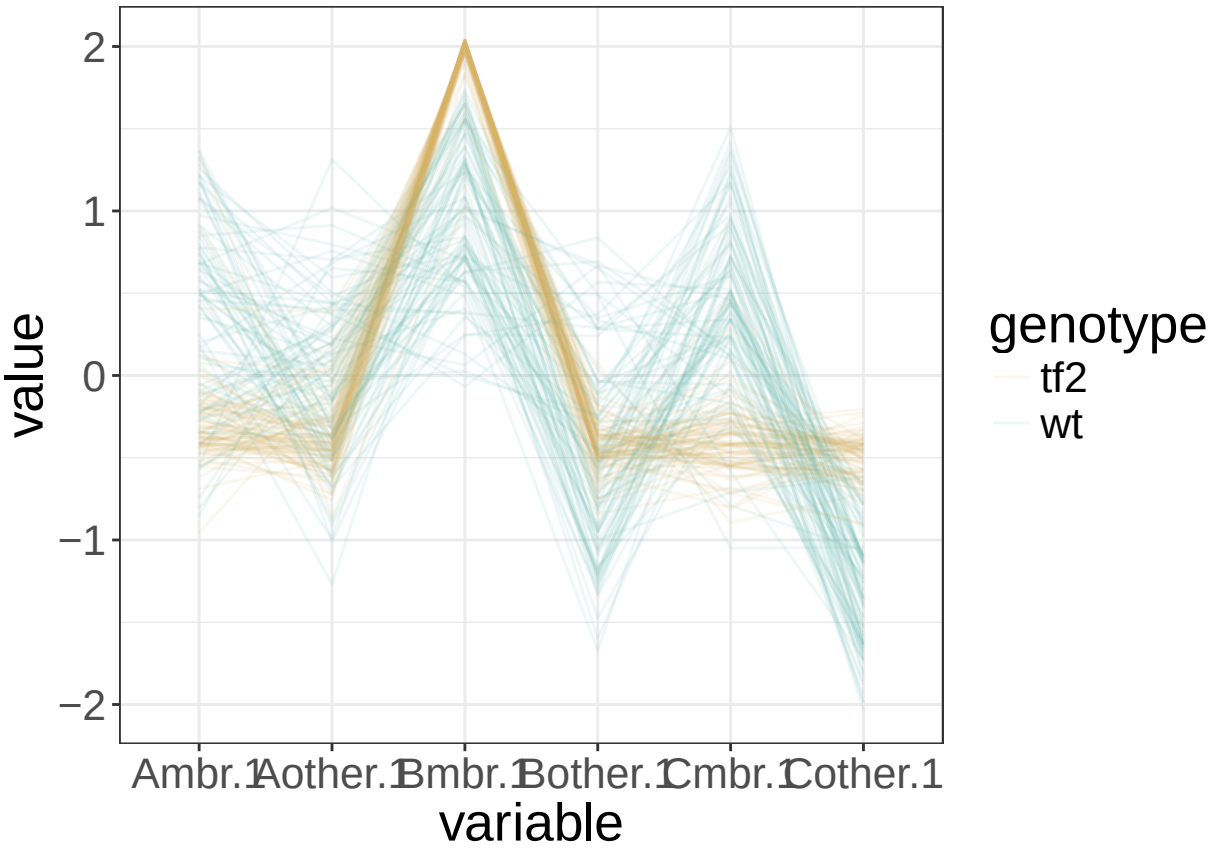

```
clusterVis_region_ssom(39)
```

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



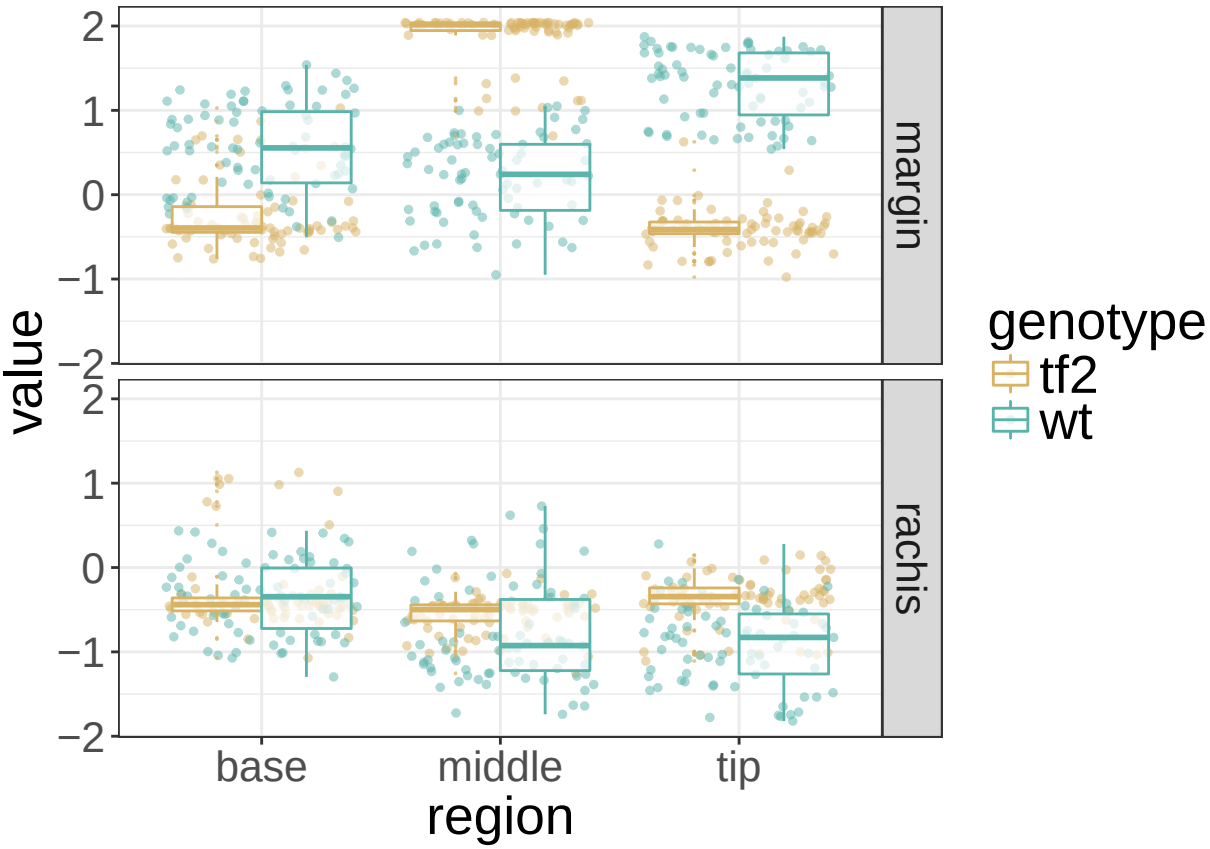

```
clusterVis_line_ssom(39)
```

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



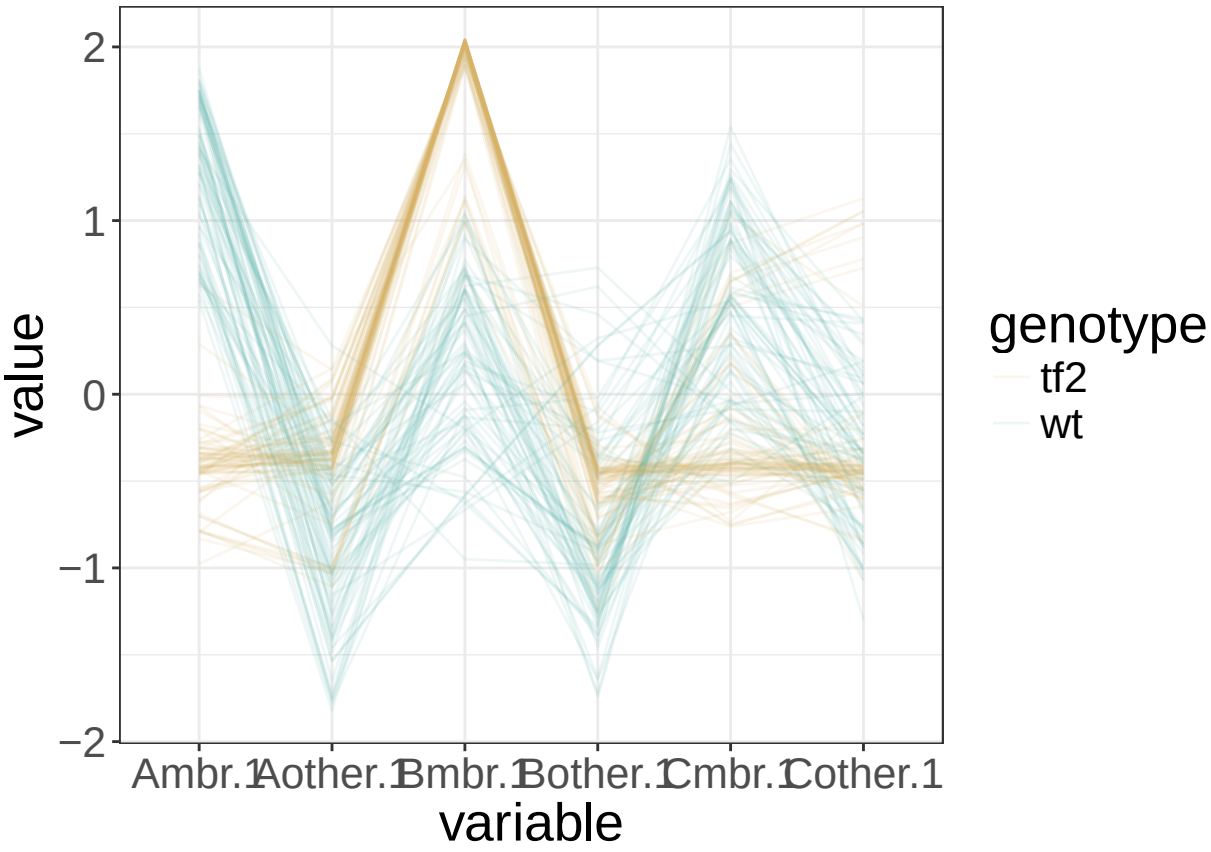

```
clusterVis_region_ssom(40)
```

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



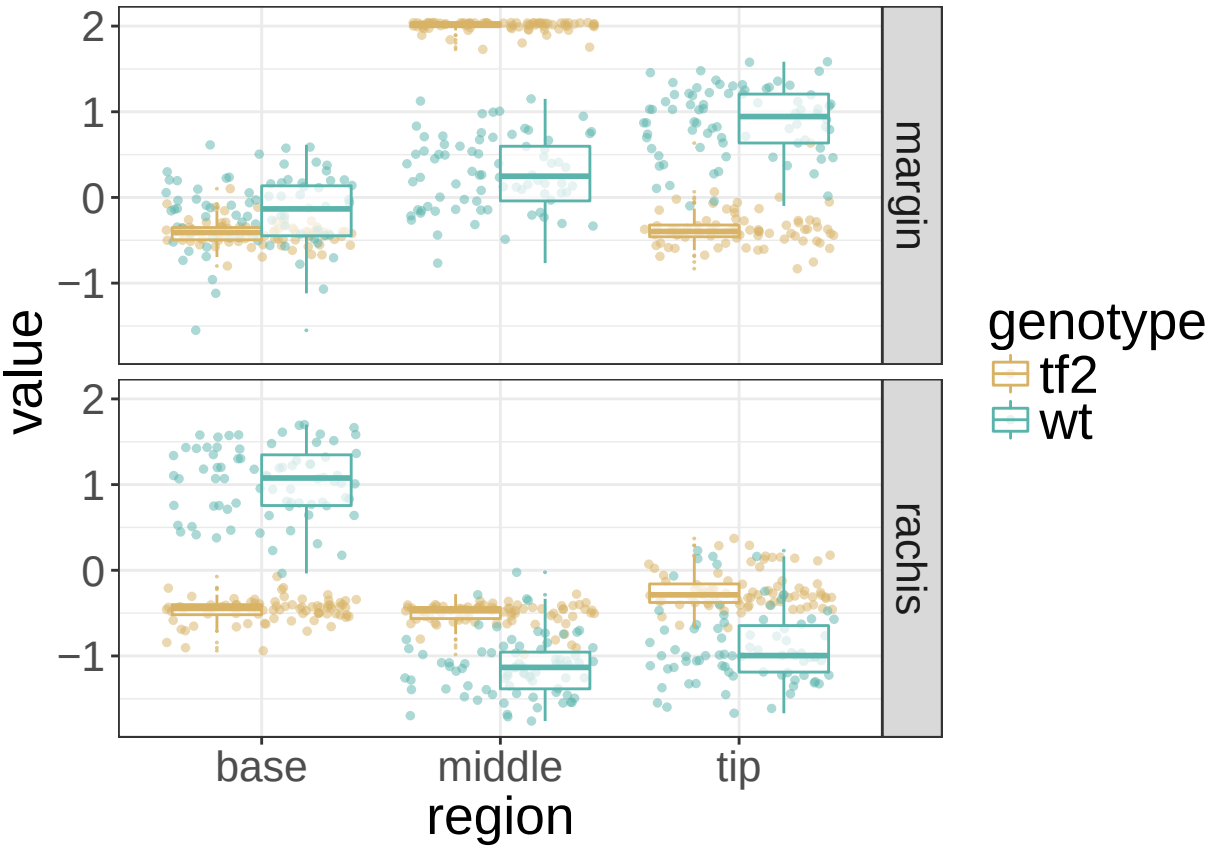

```
clusterVis_line_ssom(40)
```

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



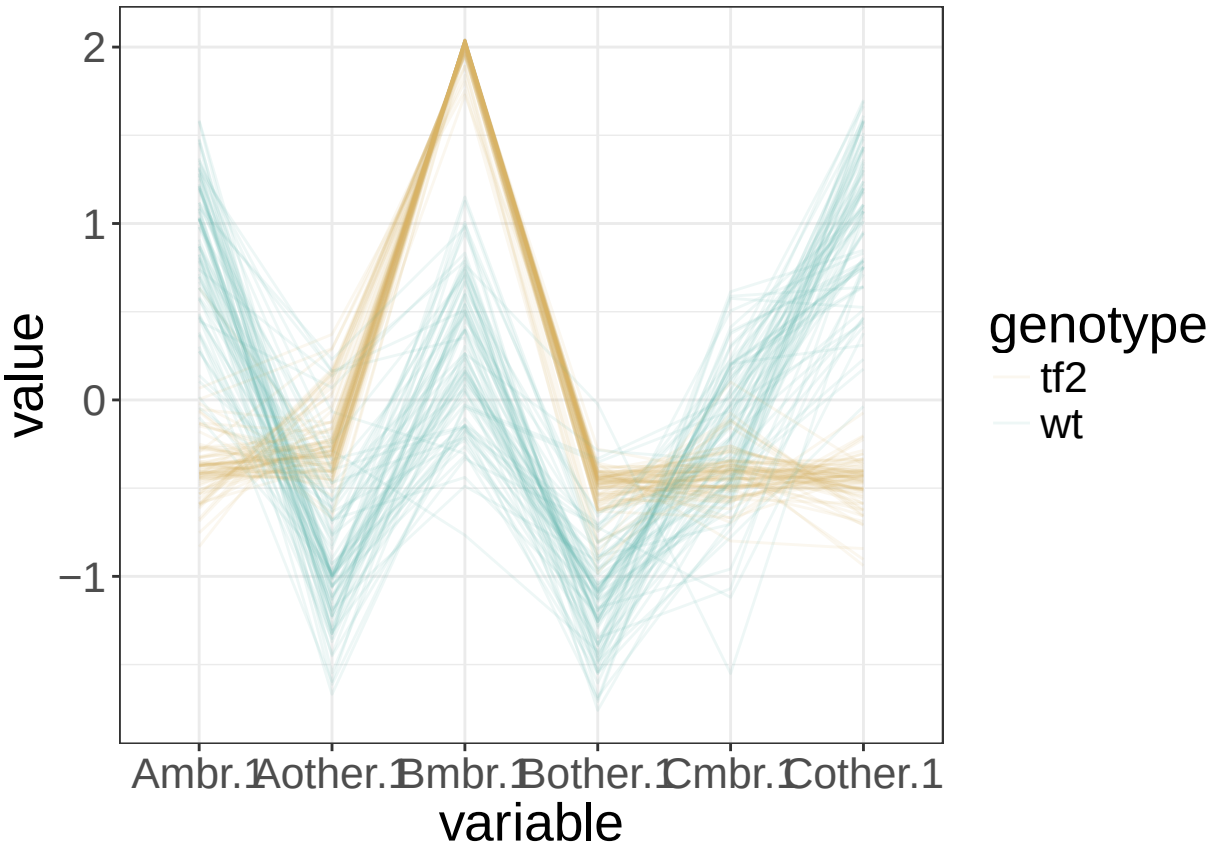

```
clusterVis_region_ssom(41)
```

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



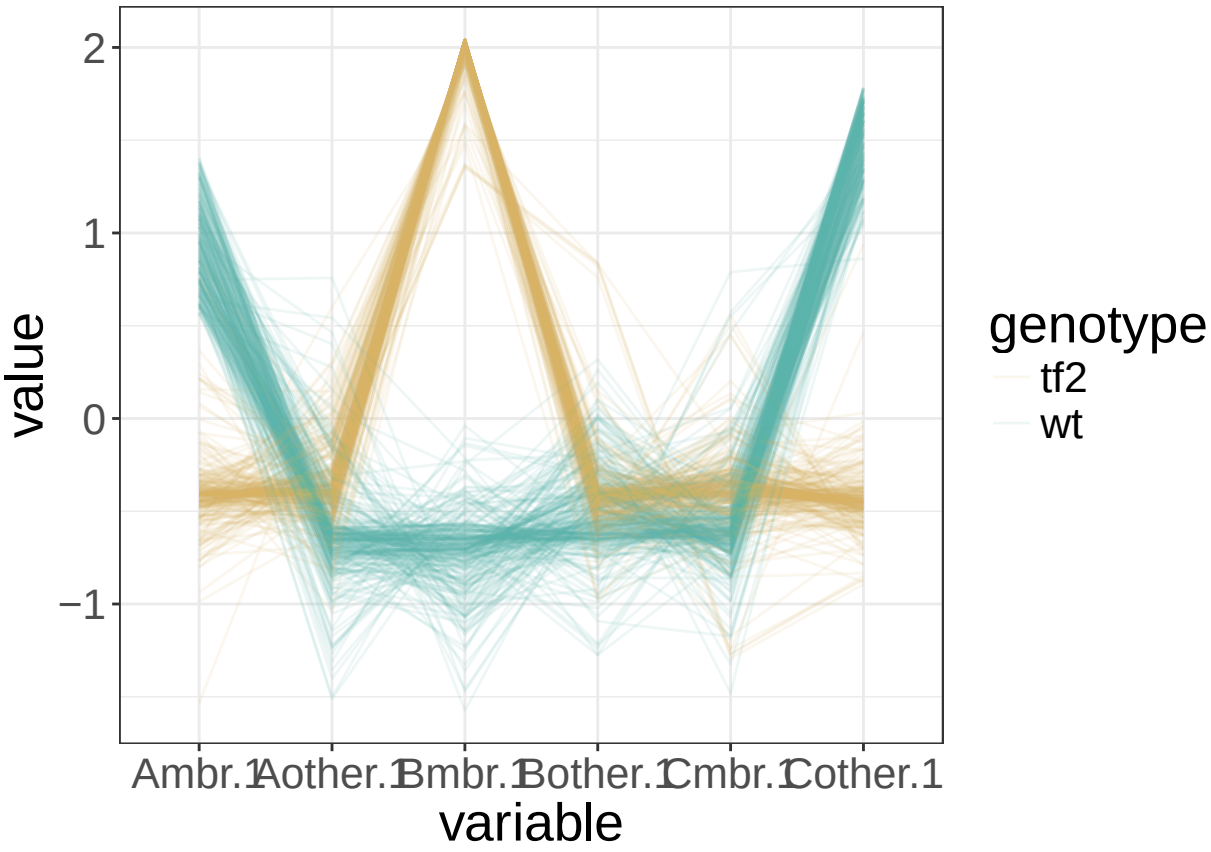

```
clusterVis_line_ssom(41)
```

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



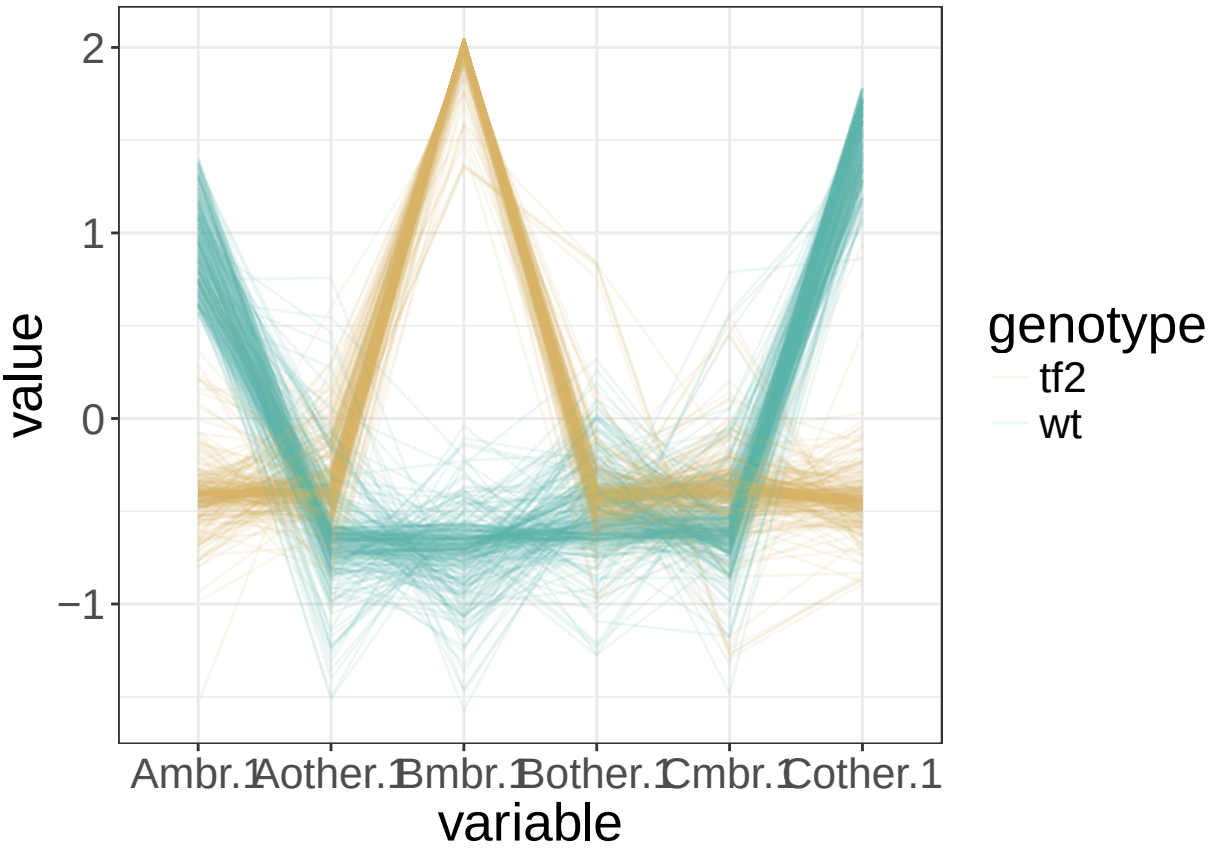

```
clusterVis_line_ssom(42)
```

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




```
clusterVis_line_ssom(42)
```

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



Make data table with gene names

```
#Prereq annotation files for function
annotation1<- read.delim("../../../06diffGeneExp/analysis1_2014/DE/DE_analysis/beforeAnalysis/requisitel
colnames(annotation1) <- c("ITAG", "SGN_annotation")
annotation2<- read.delim ("../../../06diffGeneExp/analysis1_2014//DE/DE_analysis/beforeAnalysis/requisi
annotation <- merge(annotation1,annotation2, by = "ITAG")

#Only Gene Name and ITAG
names(annotation)
```

```
##  [1] "ITAG"                "SGN_annotation"      "AGI"
##  [4] "symbol"              "gene_name"           "X..identity"
##  [7] "alignment.length"    "e.value"             "bit.score"
## [10] "percent.query.align"
```

```
annotation <- annotation[,c(1,4,5)]

#fix to one regex
annotation$ITAG <- gsub("^(.*)[.].*", "\\1", annotation$ITAG)
annotation$ITAG <- gsub("^(.*)[.].*", "\\1", annotation$ITAG)

data.val3 <- data.val2
colnames(data.val3)[2] <- "ITAG"
everything <- merge(data.val3, annotation, by = "ITAG", all.x = TRUE)

write.csv(everything, file = "../data/output/ssom.data.analysis5d_02Jan2017_larger_geneList.csv")
```