Data mining exercise

Names: Rikke Stausholm, Stefano Pellegrini

Group: 5

Introduction

Your collaborator has made a CAGE experiment using 7 different tissues. As a remainder, CAGE tags are 20-21 nt long tags mapping to the genome. We often cluster CAGE tags that are close to one another on the genome to a "tag cluster". A tag cluster can then have several tags from one or several tissues.

One can view these CAGE tag clusters as being "core promoters" in the sense that they are measuring the activity and location of a core promoter.

The data file `htbinf_cage_tpms` shows the CAGE tag clusters as rows, and tissues as columns. The cell values are the TPMs from the given tissue in the cluster. There are three additional columns: the tag cluster ID, the location of the cluster in mm8 and the strand of the cluster. The collaborator now wants to know:

1. How many types of core promoters are there in terms of tissue expression patterns, and what expression patterns are these?

```
library(tidyverse)
library(pheatmap)
library(viridis)

# Load data
cage_data <- read_tsv("htbinf_cage_tpms.txt")

# Columns
colnames(cage_data)
```

```
##  [1] "tc_id"    "location" "strand"   "cer"      "emb"      "liv"
##  [7] "lun"      "mac"      "som"      "vis"
```

```
dim(cage_data)
```

```
## [1] 1000   10
```

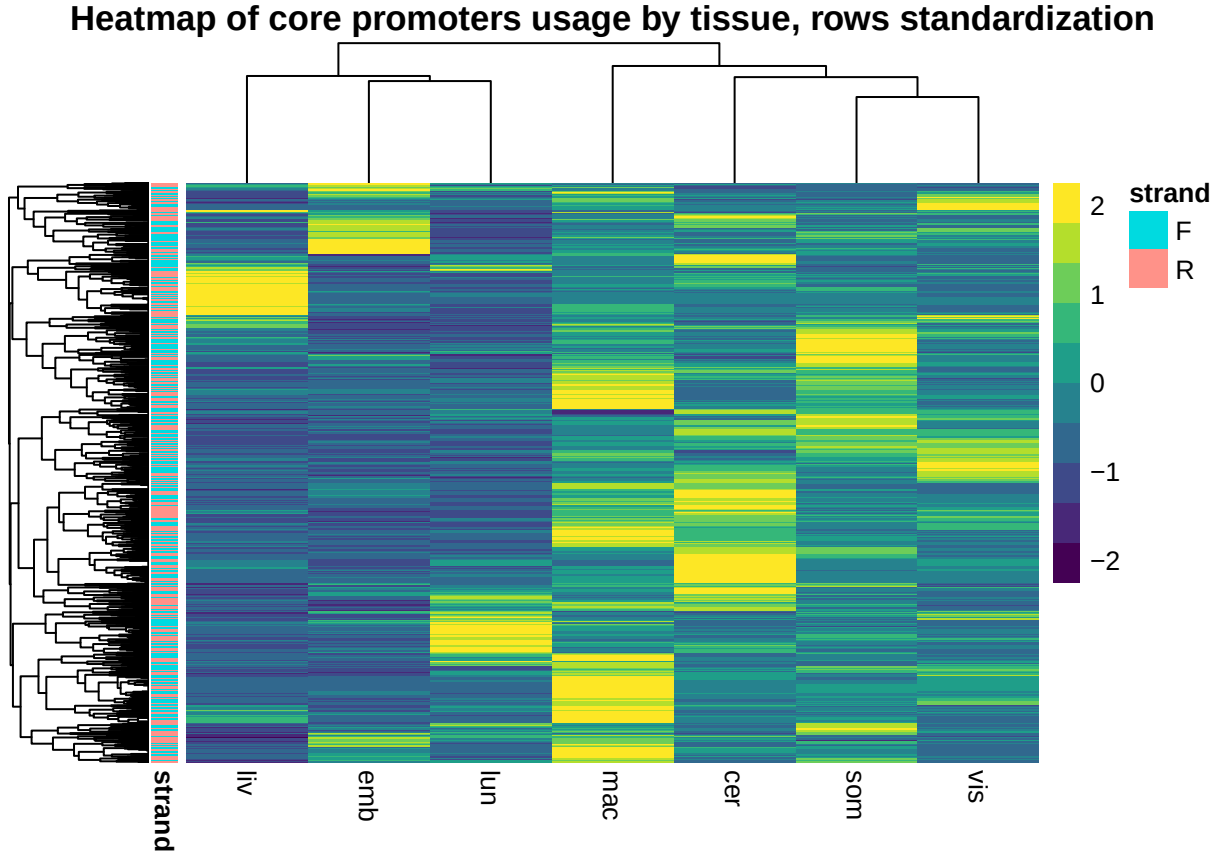
```
# Tissues: cerebellum, whole embrio, liver, lung, macrophages, somato sensory cortex, visual cortex
tissues <- colnames(cage_data[,4:10])

# TPM for each tissues (omits ID, location and strand)
TPM <- cage_data[,4:10]
```

The more tags we have, the higher is the activitity of the promoter.

Seven tissues: cerebellum, whole embrio, liver, lung, blood macrophages, somatosensory cortex, visual cortex.

Generate an heatmap with normalized rows

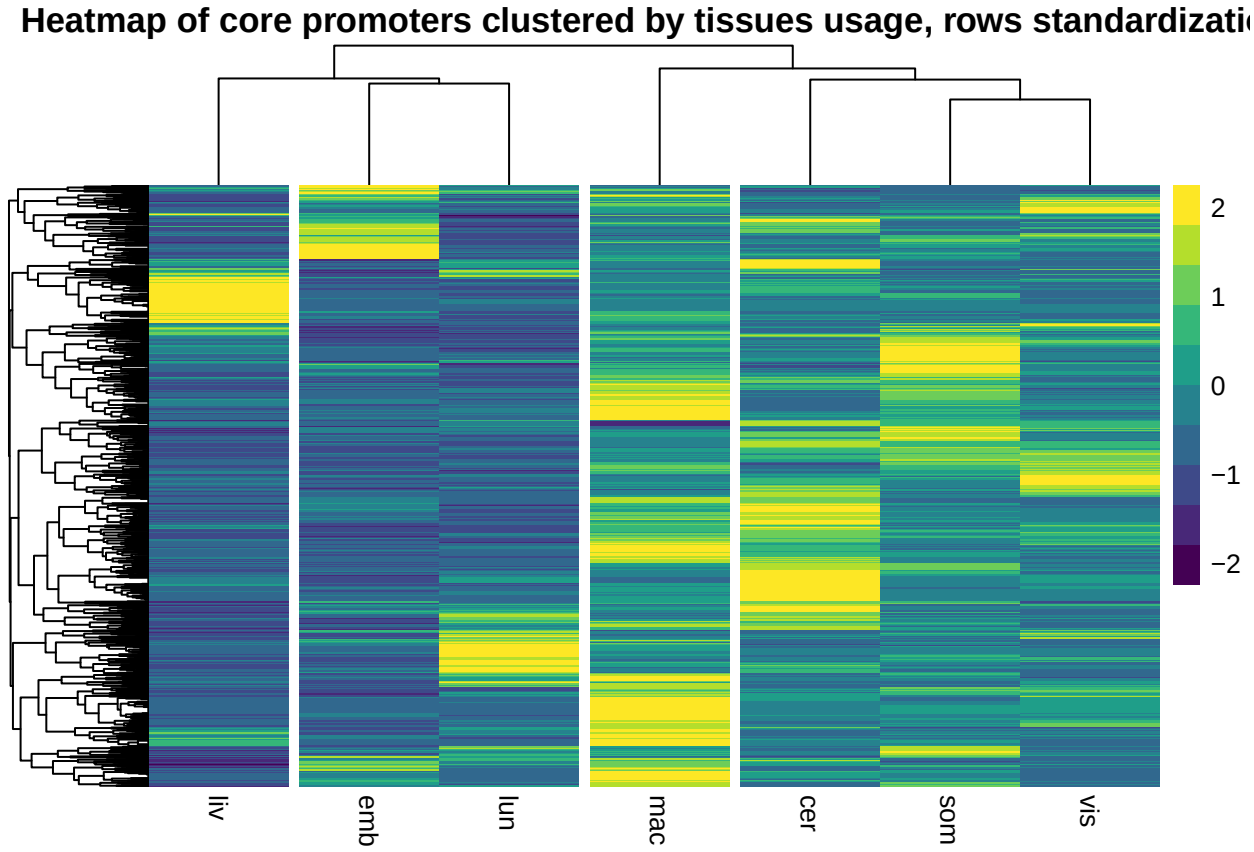
```r
# Convert to matrix to use annotation row
cage_matrix <- cage_data %>%
  select(-location, -strand) %>%
  as.data.frame() %>%
  column_to_rownames("tc_id") %>%
  as.matrix()

# Genererate heatmap, with normalizization by rows and annotation row
annotation_df <- cage_data %>%
  select(tc_id, strand) %>%
  as.data.frame() %>%
  column_to_rownames("tc_id")

pheatmap(cage_matrix,
         scale= "row",
         show_rownames = FALSE,
         color=viridis(10),
         annotation_row = annotation_df,
         main = "Heatmap of core promoters usage by tissue, rows standardization")
```



Normalizing by rows, we can see the usage of each core promoter in the different tissues. From the heatmap hierarchical clustering, it seems that we have 4 clusters of tissues. The hierarchical clustering use the Euclidean distance to cluster both core promoters (rows) and tissues (columns) and build the trees. The information about the strand doesn't seem to be important.

Divide the tissues in 4 clusters

```
# Divide the tissues in 4 clusters
pheatmap(TPM,
         scale= "row",
         show_rownames = FALSE,
         cutree_cols = 4,
         color=viridis(10),
         main = "Heatmap of core promoters clustered by tissues usage, rows standardization")
```



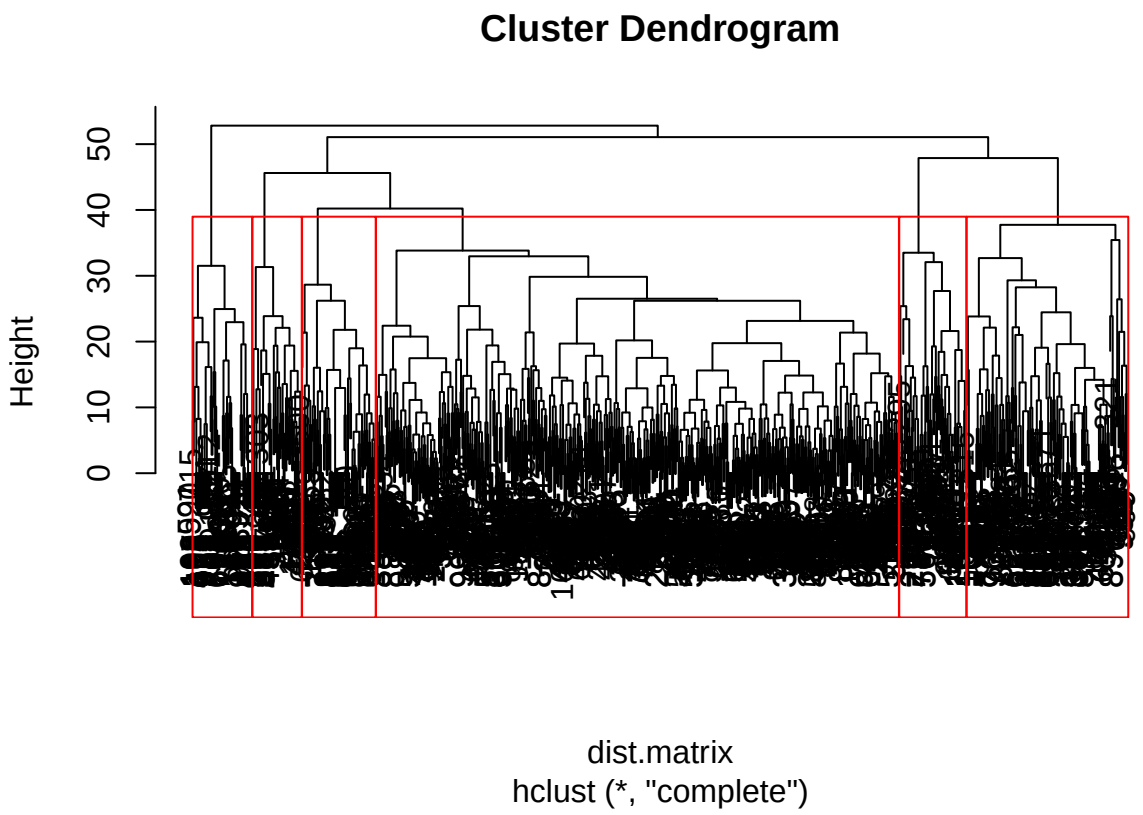
The hierarchical clustering on the heatmap shows that we have 4 types of tissue expression patterns, that uses different core promoters.

The expression patterns are:

- Liver tissue
- Embryonic and lungs tissues
- Blood macrophages tissue
- Cerebellum, somatosensory cortex and visual cortex tissues

From an other perspective we could try to find optimal promoters clusters, by dividing the core promoters

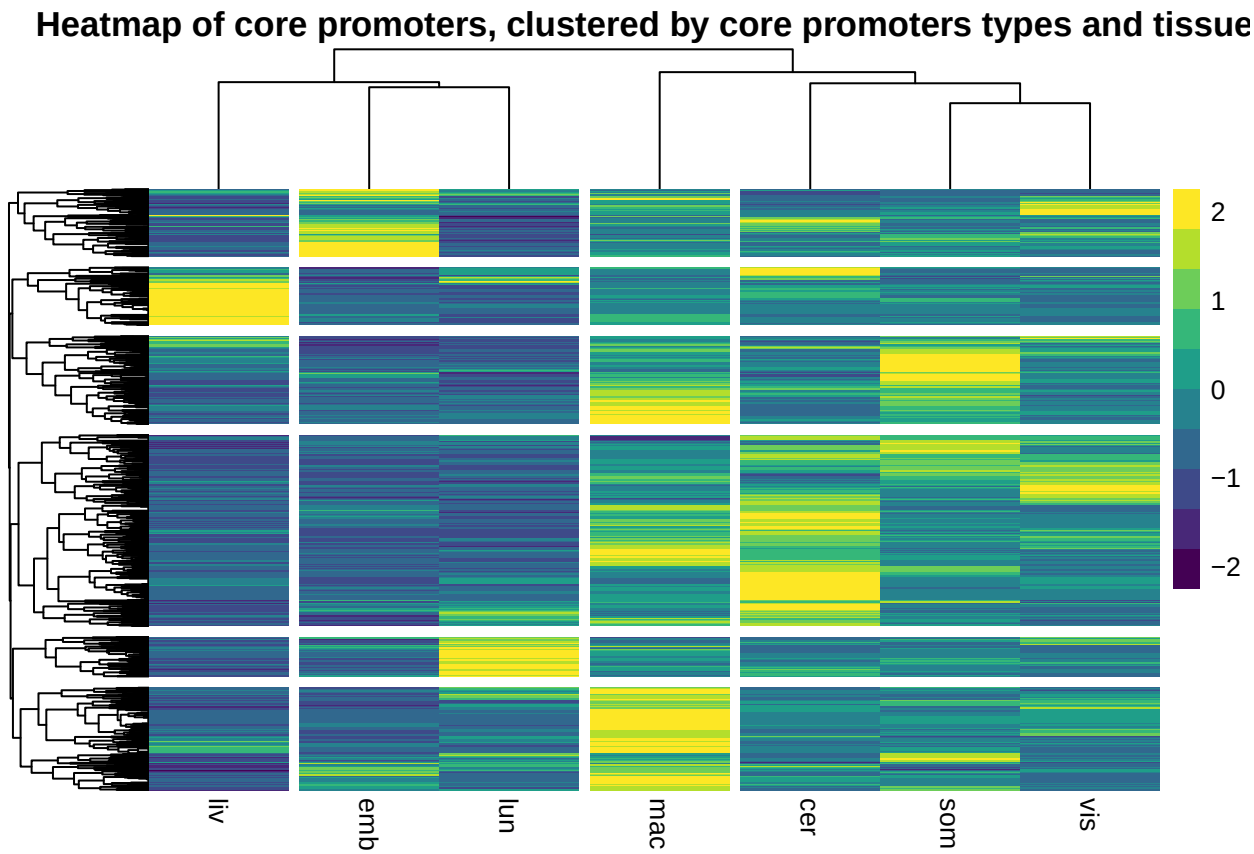
```r
# Find optimal promoters clusters
dist.matrix <- dist(TPM)
htree <- hclust(dist.matrix)
plot(htree)
rect.hclust(htree, k=6, border="red")
```



From hierarchical tree division, it seems that we have 6 clusters of core promoters.

Heatmap of the different types of core promoters, divided by tissues expression pattern

```r
# Heatmap of the different types of core promoters
pheatmap(TPM,
         scale= "row",
         show_rownames = FALSE,
         cutree_rows = 6,
         cutree_cols = 4,
         color=viridis(10),
         main = "Heatmap of core promoters, clustered by core promoters types and tissues")
```



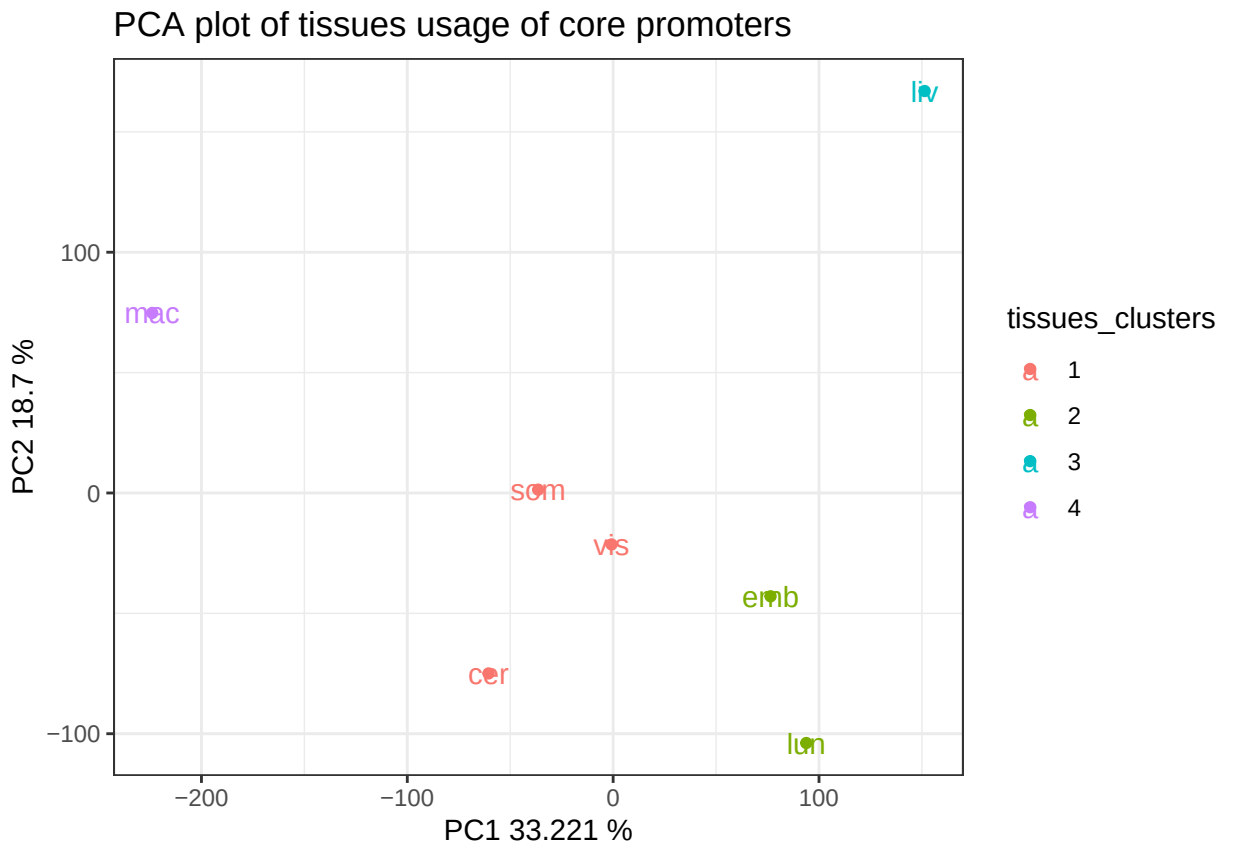
We have 6 clusters of core promoters based on their usage from the 4 clusters of tissues.

2. What tissues are similar to each other in terms of promoter usage? They would really like to have this as a picture and not just "values"

PCA plot

```
# Convert tissues' TPM to matrix and transpose
pca_cage_matrix <- TPM %>% as.matrix %>% t

# Singular value decomposition
pca_cage <- prcomp(pca_cage_matrix, center=TRUE)

# Extract % variance
percent_variance <- summary(pca_cage)$importance["Proportion of Variance",] * 100

# Create a vector for the tissues clusters color
tissues_clusters <- as.factor(c(1, 2, 3, 2, 4, 1, 1))

# PCA plot
as_tibble(pca_cage$x) %>%
  ggplot(aes(x=PC1, y=PC2, label=tissues, col=tissues_clusters)) +
  geom_point() +
  geom_text() +
  xlab(label = paste("PC1", percent_variance[1], "%")) +
  ylab(label = paste("PC2", percent_variance[2], "%")) +
```

```
  labs(title = "PCA plot of tissues usage of core promoters") +
  theme_bw()
```



We can see the differences between the core promoters usage of the tissues by both, the heatmaps shown in the previous exercise and in the above PCA plot. Overall, we can confirm that we have the 4 mentioned different tissues patterns, in terms of core promoters usage.

3. How many tissue-specific promoters are there, per tissue and just allover? It would be very helpful to calculate two "specificity scores": one for each tissue, and one summary score across all tissues for each promoter.

4. They want to have genome browser examples of the most tissue-specific promoter for each tissue

5 Lastly, they want a list with the 10 most tissue-specific promoters, taking all tissues into account.

Use of variance and Gini coefficient to find the tissue-specific promoters

```
library(reldist)
```

```
## reldist: Relative Distribution Methods
## Version 1.6-6 created on 2016-10-07.
## copyright (c) 2003, Mark S. Handcock, University of California-Los Angeles
##  For citation information, type citation("reldist").
##  Type help(package="reldist") to get started.
```

```
# Find most specific promoter using using variance
cage_data %>%
  mutate(variance=apply(cage_data[tissues], 1, var)) %>%
  arrange(desc(variance)) %>%
  head(10)
```

```
## # A tibble: 10 x 11
##    tc_id  location   strand   cer   emb    liv    lun   mac   som   vis variance
##    <chr>  <chr>      <chr>  <dbl> <dbl>  <dbl>  <dbl> <dbl> <dbl> <dbl>    <dbl>
##  1 C17F1~ chr17:236~ F       27.6  31.6   2.46   1.54   5.89  4.79  4.30     162.
##  2 C15RB~ chr15:123~ R       27.6  31.6   2.46   1.54   5.89  4.79  4.30     162.
##  3 C3R58~ chr3:9257~ R       27.6  30.1   0.821  0.770 17.7   4.79  8.60     152.
##  4 C6F44~ chr6:4489~ F       27.6  30.1   0.821  0.770 17.7   4.79  8.60     152.
##  5 C18R2~ chr18:409~ R       27.6  30.1   0.821  0.770 17.7   4.79  8.60     152.
##  6 C17F2~ chr17:346~ F       27.6  30.1   0.821  0.770 17.7   4.79  8.60     152.
##  7 C9F5E~ chr9:9926~ F       27.6   3.16  0.821 31.6    5.89 14.4   8.60     146.
##  8 C15F2~ chr15:375~ F       27.6   3.16  0.821 31.6    5.89 14.4   8.60     146.
##  9 C14F6~ chr14:111~ F        7.90 30.1   0.821  9.24  29.4   9.57  4.30     139.
## 10 C1F14~ chr1:2124~ F       19.7   1.58 32.8    0.770 11.8   4.79  4.30     138.
```

```
# Find most specific promoter using gini index
cage_data %>%
  mutate(gini=apply(cage_data[tissues], 1, reldist::gini)) %>%
  arrange(desc(gini)) %>%
  head(10)
```

```
## # A tibble: 10 x 11
##    tc_id    location      strand   cer   emb   liv   lun   mac   som   vis  gini
##    <chr>    <chr>         <chr>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1 C8R1C4F~ chr8:2968631~ R       7.90  1.58 0.821 0.770 29.4   4.79 12.9  0.572
##  2 C15F637~ chr15:651674~ F       7.90  1.58 0.821 0.770 29.4   4.79 12.9  0.572
##  3 C3R7943~ chr3:1271545~ R       7.90  1.58 0.821 1.54  29.4   9.57  4.30 0.561
##  4 C17F37F~ chr17:586912~ F      27.6   1.58 0.821 0.770 17.7   4.79  8.60 0.560
##  5 C13F574~ chr13:915362~ F      11.8   0    3.28  0.770  0     4.79  8.60 0.554
##  6 C4F7668~ chr4:1241605~ F       3.95  3.16 0.821 0.770 11.8   4.79 25.8  0.552
##  7 C16F2C1~ chr16:462168~ F       3.95  1.58 1.64  0.770 11.8  23.9   4.30 0.551
##  8 C15RCAA~ chr15:132811~ R       3.95  1.58 2.46  0.770 23.5   4.79 25.8  0.551
##  9 C5R58E5~ chr5:9321284~ R      27.6   1.58 0.821 2.31   5.89  4.79 25.8  0.550
## 10 C7F37D0~ chr7:5852617~ F       3.95  1.58 0.821 3.85  29.4  19.1   4.30 0.550
```

The variance, since it is a measure of the spread of a distribution, it could be used to measure the promoter specificity across all tissues, but the Gini coefficient should be a more specific metric for this task. The Gini coefficient is a common measure of inequality within a distribution. It goes from 0 and 1, where 0 is the complete equality and 1 is the complete inequality. It can be used to compute the tissue specificity.