PCA documentation

Franziska Heinkele

8 Juni 2019

Load data:

```
library(readr)
  Untreated_notnormalized = readRDS(paste0(wd,"/data/NCI_TPW_gep_untreated.rds"))
  Treated_notnormalized = readRDS(paste0(wd,"/data/NCI_TPW_gep_treated.rds"))
  Metadata = read_tsv(paste0(wd,"/data/NCI_TPW_metadata.tsv"))
```

Data need to be normalized:

```
Untreated <- apply(Untreated_notnormalized, 2, function(x){
  (x - mean(x)) / sd(x)
})

Treated <- apply(Treated_notnormalized, 2, function(x){
  (x - mean(x)) / sd(x)
})

FC_notnormalized <- Treated_notnormalized - Untreated_notnormalized

FC <- apply(FC_notnormalized, 2, function(x){
  (x - mean(x)) / sd(x)
})
```

BROAD ANALYSIS

PRINCIPAL COMPONENT ANALYSIS

Description of this file:

This file documents the PCA analysis which was performed for two matrices:

- Treated data
- Fold change data

For both PCAs, the celllines were colored by 2 different features:

- tissue-type
- drug-type

Treated data PCA

Execute the PCA for Treated data:

```
treated.pca = prcomp(Treated, center=T, scale. = T)
```

Hereinafter, we want to use **information from Metadata** to color different celllines in the PCA. Therefore we need to check, if the celllines in the sample-column of Metadata are in the same order as in the Treated matrix. First of all we test, if the number of samples is equal.

```
identical(nrow(Metadata), ncol(Treated))
```

```
## [1] FALSE
```

```
nrow(Metadata)
```

```
## [1] 1638
```

```
ncol(Treated)
```

```
## [1] 819
```

Metadata consists of twice as much celllines as the Treated matrix since Metadata contains information for treated and untreated celllines. We want to print those rows from Metadata which do not contain a zero concentration because they belong to the treated samples.

```
TreatedrowsMetadata <- grep(Metadata$sample, pattern = "_0nM_", invert = TRUE)
```

Check, if the sample order is equal in the Treated-matrix and in Metadata:

```
Metadata <- as.data.frame(Metadata)
Metadatasamples <- Metadata[TreatedrowsMetadata,"sample"]
identical(colnames(Treated), Metadatasamples)
```

```
## [1] TRUE
```

Consequently the drug information of the Metadata-matrix can be assigned to the samples in the Treated-matrix sequentially. For better readability, we assign the column of interest to the name “Metadatadrugs”:

```
Metadatadrugs <- Metadata[TreatedrowsMetadata,"drug"]
```

Add Metadatadrugs as a new row to the Treated-matrix:

```
Treatedwithdrugs <- rbind(Treated, Metadatadrugs)
```

Save drug information as factors so it can be used for coloring:

```
drugfactor <- as.factor(Treatedwithdrugs["Metadatadrugs",])
```

Now we can go on with coloring!

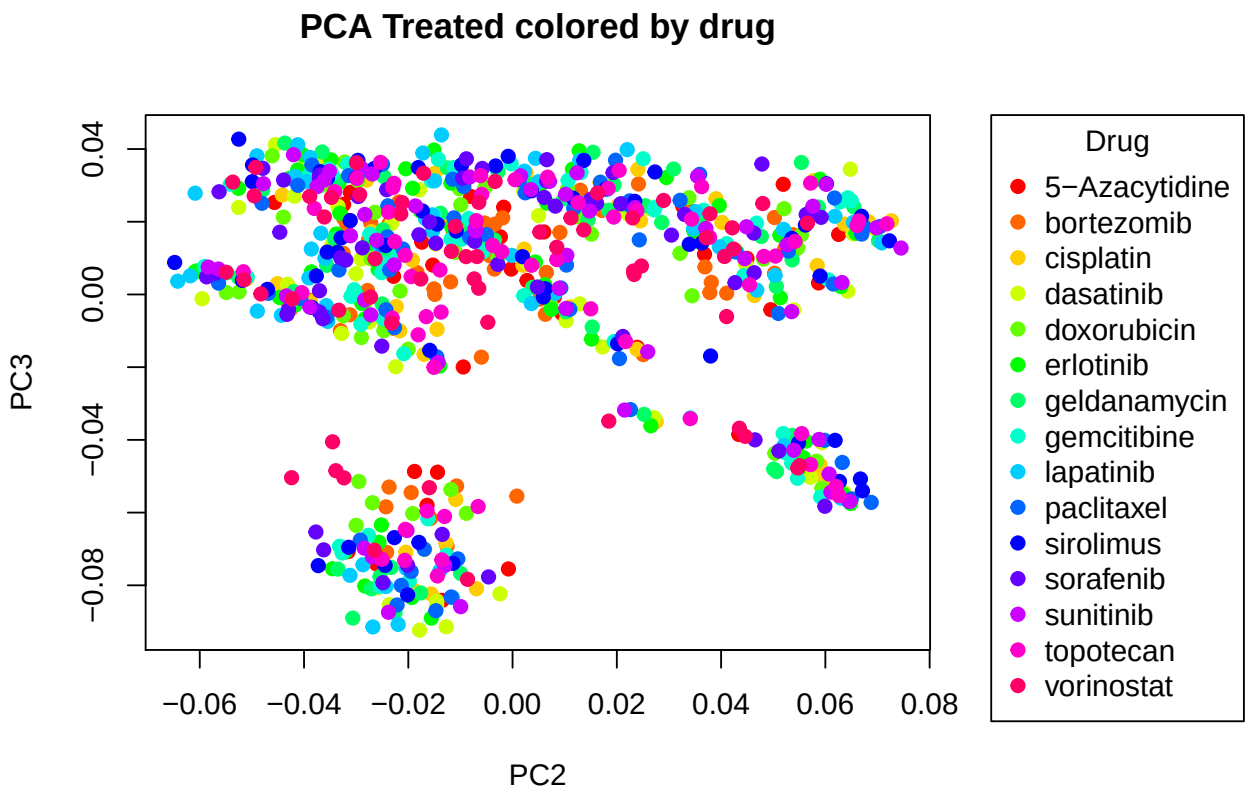
PLOT PCA & COLOR ACCORDING TO DRUGS:

Since we have 15 different drugs we need 15 different colors:

```
palette(rainbow(15))
```

Plot Principal component 1 and 2 and add a legend to the plot. To see the PCA plot and the legend next to each other, the graphical parameters are setted by the par() function.

```
par(mar=c(5, 4, 5, 9))
plot(treated.pca$rotation[, 2], treated.pca$rotation[, 3], pch = 19, xlab = "PC2",
     ylab = "PC3", col=drugfactor, main = "PCA Treated colored by drug")
levels <- as.factor(levels(drugfactor))
legend("topright", inset = c(-0.4,0),levels(drugfactor), xpd = TRUE, pch=19,
       col = levels, title = "Drug")
```



-> We do not see, that samples treated with the same drug form groups in the plot. However, we did not expect that, since we only look at the final expression and not at the expression change.

PLOT PCA & COLOR ACCORDING TO TISSUE:

The information which is needed for coloring is summarized as Metadatatissue:

```
Metadatatissue <- Metadata[TreatedrowsMetadata,"tissue"]
```

Bind Metadatatissue as a new row to the Treated matrix:

```
Treatedwithtissue <- rbind(Treated, Metadatatissue)
```

Save tissue information as factors so it can be used for coloring:

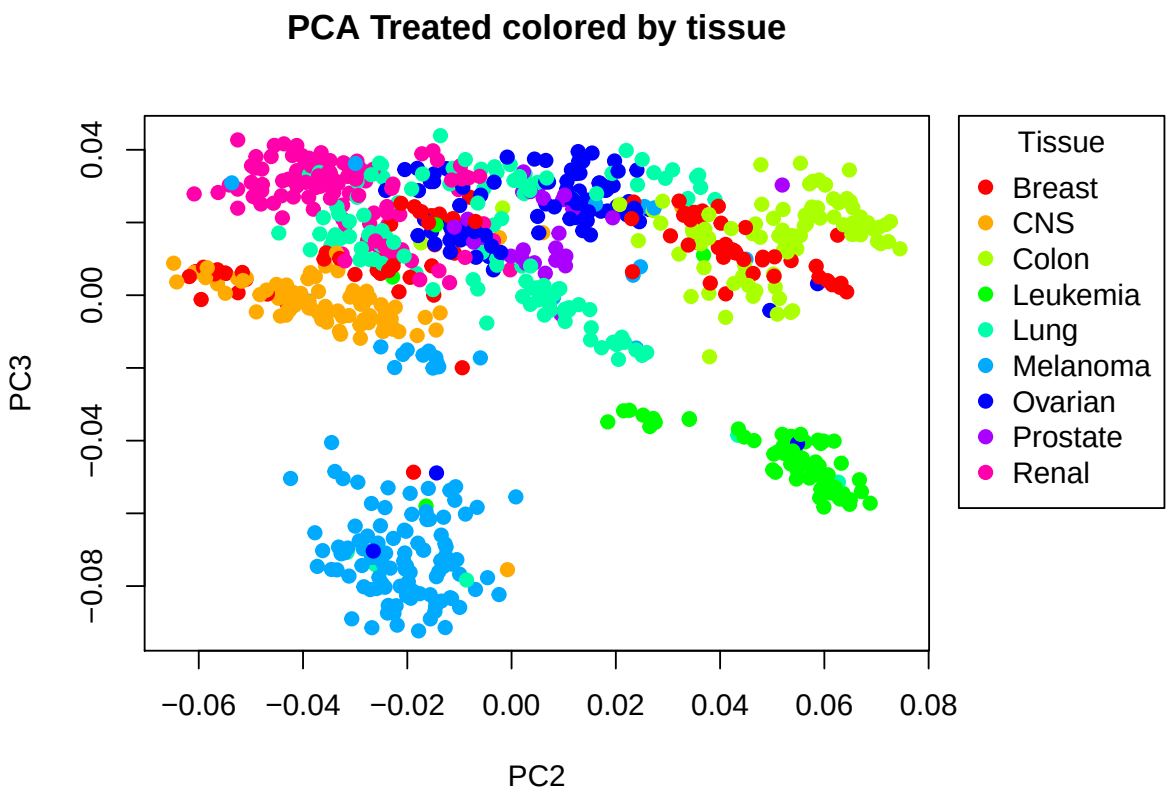
```
tissuefactor <- as.factor(Treatedwithtissue["Metadatatissue",])
```

Since we have 9 different tissue types we need 9 different colors:

```
palette(rainbow(9))
```

Plot PC 2 and PC 3 and add a legend:

```
par(mar=c(5, 4, 5, 9))
plot(treated.pca$rotation[, 2], treated.pca$rotation[, 3], pch = 19, xlab = "PC2",
     ylab = "PC3", col=tissuefactor, main= "PCA Treated colored by tissue")
levels <- as.factor(levels(tissuefactor))
legend("topright", inset = c(-0.3,0), levels(tissuefactor), xpd = TRUE, pch=19,
       col = levels, title = "Tissue")
```



-> PC 2 and PC 3 group the treated celllines as well as other PC combinations. Thus, most of the celllines of the same tissue-type seem to have similarities regarding their gene expression.

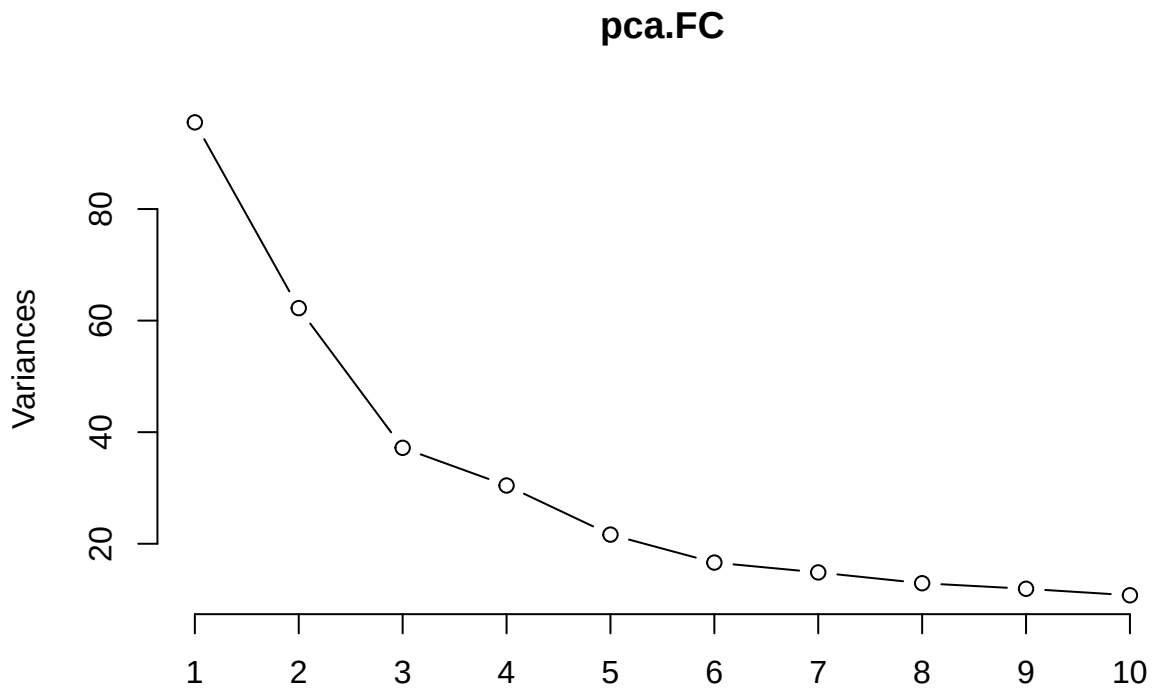
FC data PCA

We execute the PCA with FC data:

```
pca.FC = prcomp(FC, center = T, scale. = T)
```

We want to see how much variance is explained by each principle component:

```
plot(pca.FC, type = "l")
```



We can interpret, that **PC 1-3** explain most of the variance because an "elbow" can be seen after the third PC. Nevertheless, we should not exclude other PCs from our further analysis.

PLOT PCA & COLOR ACCORDING TO TISSUE

Bind the tissue-information as a new row to the FC matrix:

```
FCwithtissue <- rbind(FC, Metadatatissue)
```

Save tissue information as factors so it can used for coloring:

```
tissuefactorFC <- as.factor(FCwithtissue["Metadatatissue",])
```

9 different tissue types require 9 different colors:

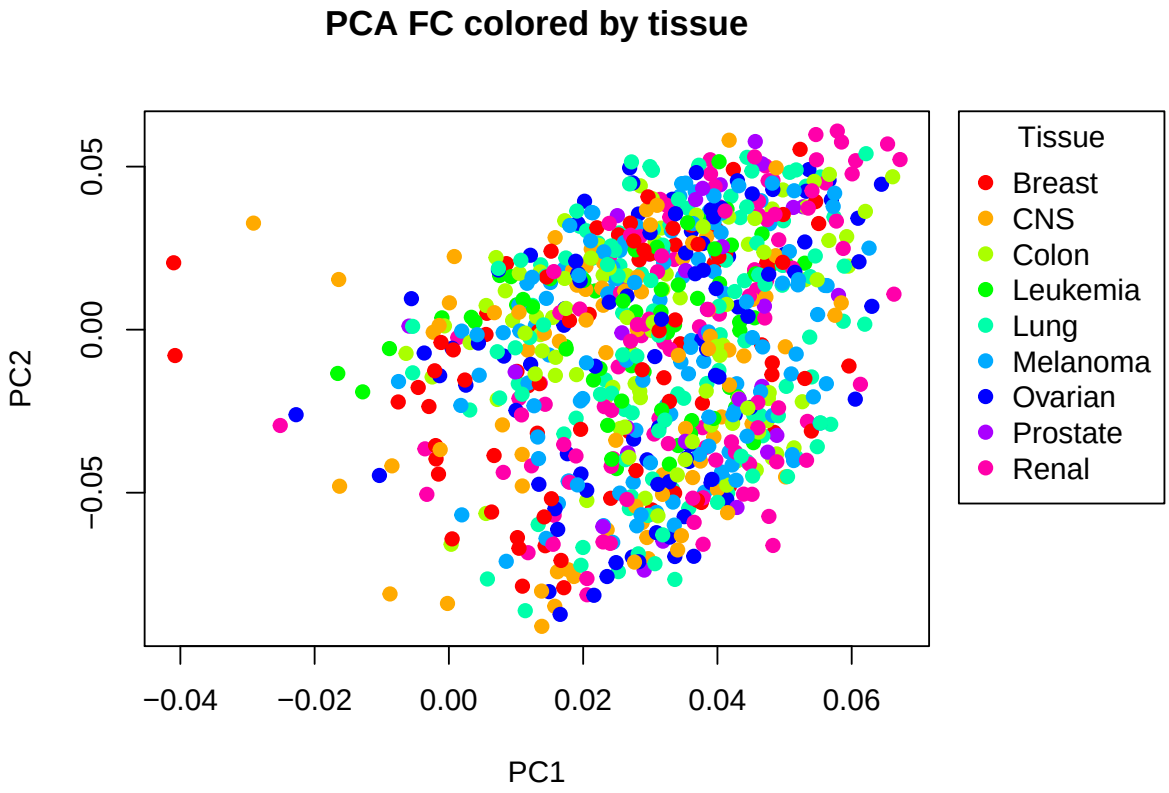
```
palette(rainbow(9))
```

Different PCs are plotted to see which combination groups the samples best. However, different tissues do not seem to group the points in any of the following PC combinations:

```
   PC1PC2 <- plot(pca.FC$rotation[, 1], pca.FC$rotation[, 2], col= tissuefactor, pch = 19,
                  xlab = "PC1", ylab = "PC2")
   PC2PC3 <- plot(pca.FC$rotation[, 2], pca.FC$rotation[, 3], col= tissuefactor, pch = 19,
                  xlab = "PC2", ylab = "PC3")
   PC3PC4 <- plot(pca.FC$rotation[, 3], pca.FC$rotation[, 4], col= tissuefactor, pch = 19,
                  xlab = "PC3", ylab = "PC4")
```

Example: PC1 and PC2 do not group celllines of same tissue-type:

```
par(mar=c(5, 4, 5, 9))
plot(pca.FC$rotation[, 1], pca.FC$rotation[, 2], col = tissuefactor, pch = 19, xlab = "PC1",
     ylab = "PC2", main = "PCA FC colored by tissue")
levels <- as.factor(levels(tissuefactorFC))
legend("topright", inset = c(-0.3,0), levels(tissuefactorFC), xpd = TRUE, pch=19,
       col = levels, title = "Tissue")
```



-> Since we are not able to identify groups of celllines of the same tissue, fold changes might be not very tissue-specific.

PLOT PCA & COLOR ACCORDING TO DRUGS

Create a new matrix ("FCwithdrugs") where the druginformation is added as a new row to the FC-matrix:

```
FCwithdrugs <- rbind(FC, Metadatadrugs)
```

Save drug information as factors so it can used for coloring:

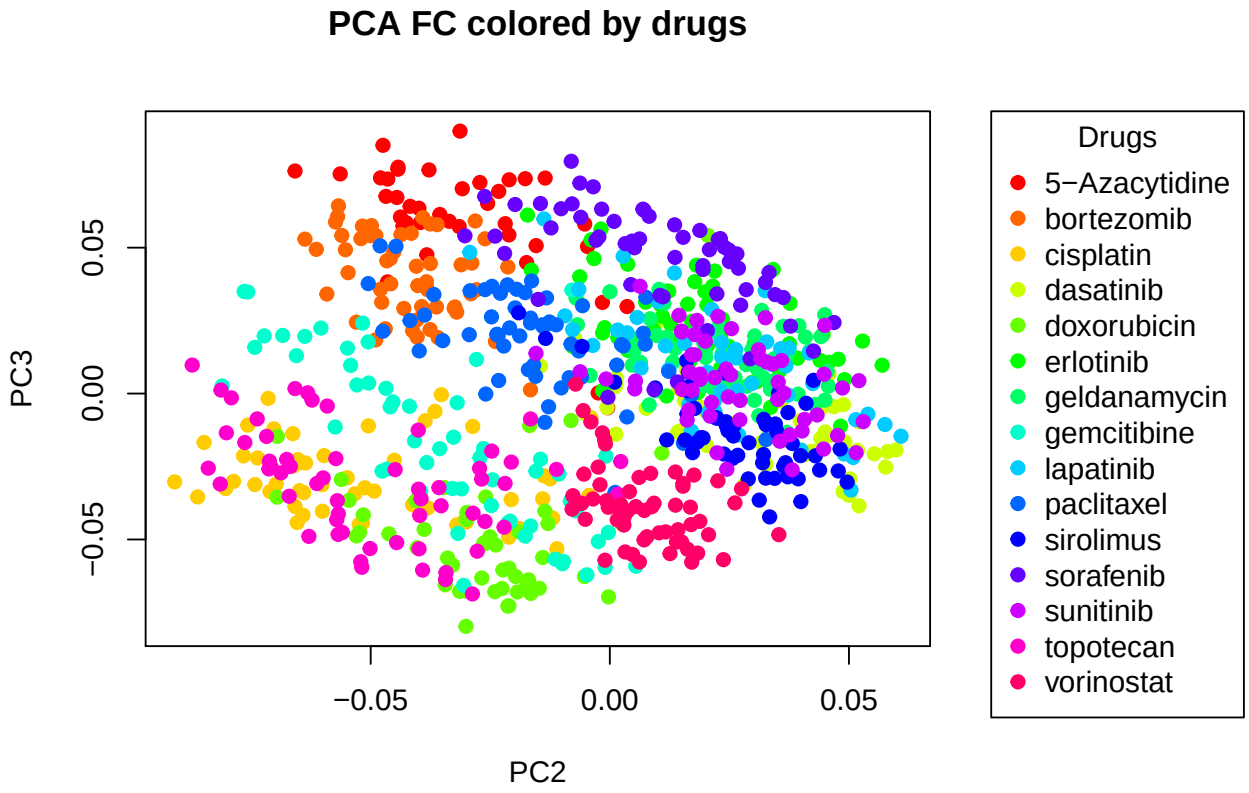
```
drugfactorFC <- as.factor(FCwithdrugs["Metadatadrugs",])
```

According to 15 different tissue types we need 15 different colors:

```
palette(rainbow(15))
```

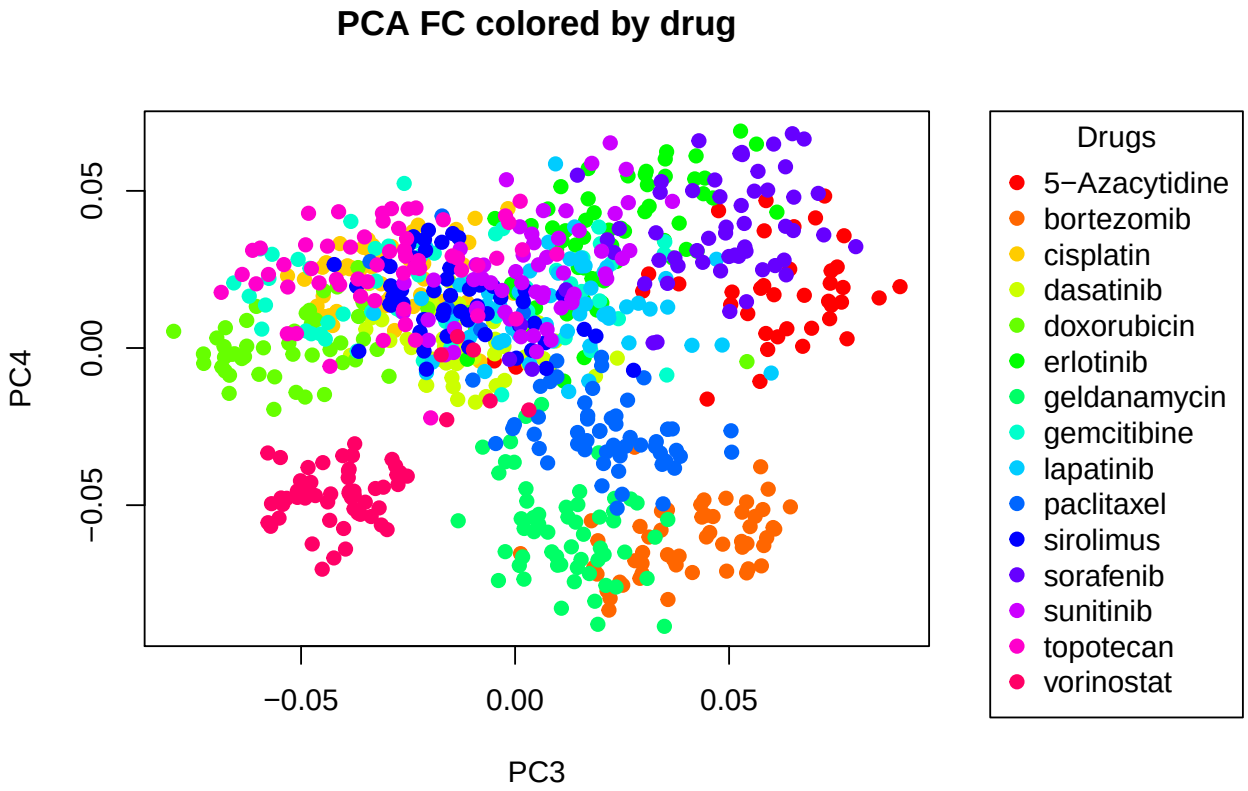
Plot PC 2 & PC 3:

```
par(mar=c(5, 4, 5, 9))
plot(pca.FC$rotation[, 2], pca.FC$rotation[, 3], col = drugfactor , pch = 19, xlab = "PC2"
         , ylab = "PC3", main = "PCA FC colored by drugs")
levels <- as.factor(levels(drugfactorFC))
legend("topright", inset = c(-0.4,0), levels(drugfactorFC), xpd = TRUE, pch=19,
       col = levels, title = "Drugs")
```



->Plot PC 3 and PC 4:

```
par(mar=c(5,4,5,9))
plot(pca.FC$rotation[, 3], pca.FC$rotation[, 4], col = drugfactor , pch = 19, xlab = "PC3",
     ylab = "PC4", main = "PCA FC colored by drug")
levels <- as.factor(levels(drugfactorFC))
legend("topright", inset = c(-0.4,0), levels(drugfactorFC), xpd = TRUE, pch=19,
       col = levels, title = "Drugs")
```



-> Many combinations of Principal Components clearly group celllines treated with the same drug. Consequently, the FC of celllines seems to be drug-specific. ***

HIGHLIGHT VORINOSTAT

Since we are going to analyze the effects of Vorinostat in our specific analysis we want to plot a PCA that highlights exclusively those celllines which belong to Vorinostattreatment.

Therefore we use the ifelse-function:

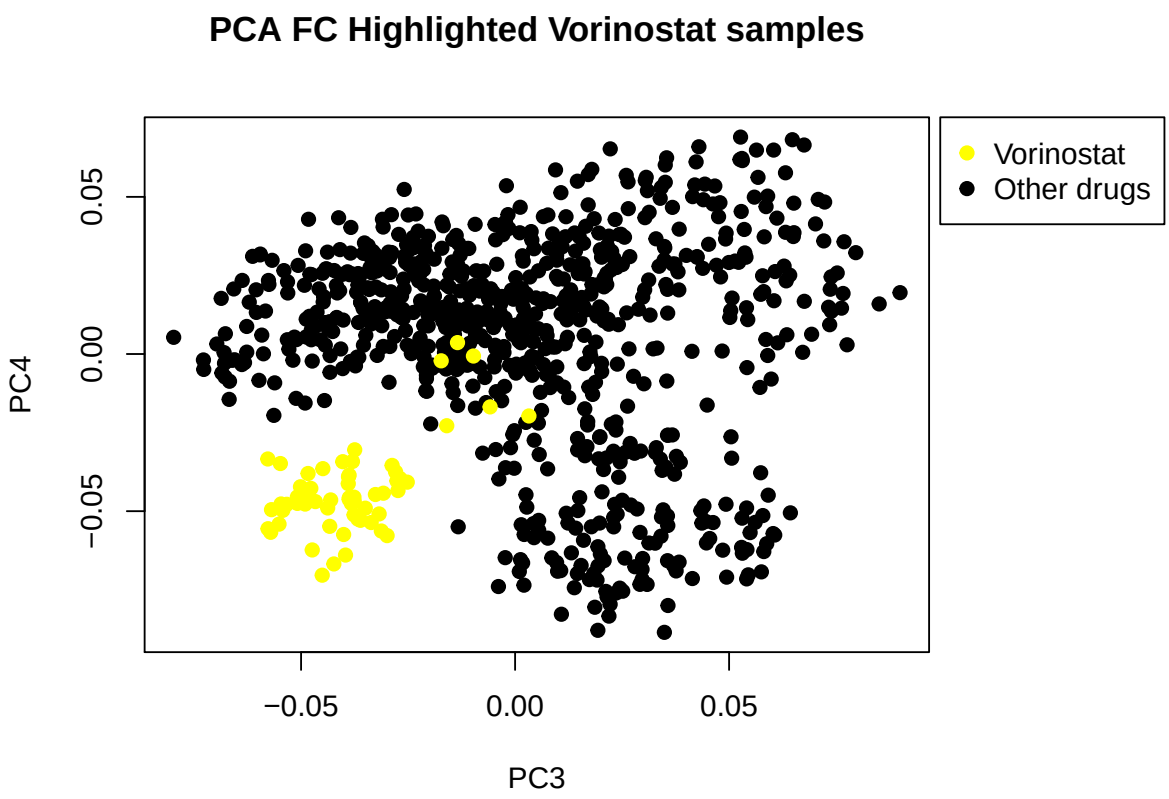
```
Metadata <-as.data.frame(Metadata)
Marking <- ifelse(Metadata$drug == "vorinostat", "yellow", "black")
```

Add the information, whether samples belong to Vorinostat, to the FC matrix:

```
HighlightVorinostat <- cbind(`FC` = Marking)
```

Plot PC 3 and PC 4:

```
par(mar=c(5, 4, 5, 9))
plot(pca.FC$rotation[, 3], pca.FC$rotation[, 4], col = HighlightVorinostat, pch = 19,
     xlab = "PC3", ylab = "PC4", main = "PCA FC Highlighted Vorinostat samples")
legend("topright", inset = c(-0.3,0), legend = c("Vorinostat","Other drugs"),
       xpd = TRUE, pch=19, col = c("yellow", "black"))
```


PCA FC Highlighted Vorinostat samples
Vorinostat
Other drugs
PC4
0.05
0.00
−0.05
−0.05
0.00
0.05
PC3