Transcriptomic and physiological responses to seasonal and diurnal cycles in Ostreococcus

Francisco J. Romero-Campero & Ana B. Romero-Losada

August 11, 2020

Experimental design

In this project we study the transcriptomic and physiological responses to seasonal and diurnal cycles in the picoeukaryote *Ostreococcus tauri*. Our favourite microalgae was grown in 1.8L column photochemostats under long day conditions (16 hours light : 8 hours dark) representing a summer day and under short day conditions (8 hours light : 16 hours dark) simulating a winter day. After four weeks of entrainment under each condition samples were collected for three days every four hours. Then the program controlling the light in the photochemostats was set to free running conditions consisting on continuous light and samples were again collected every four hours for two days.

Load data and principal component analysis.

The matrix containing the gene expression data analyzed in this study can be downloaded from **this link from the GEO data base**. Make sure to uncompress this file and rename it to **gene_expression.tsv**. First, the gene expression data is loaded and converted into a numeric matrix setting the rownames to gene ids.

```
gene.expression <- read.table(file ="gene_expression.tsv",sep="\t",header=T,as.is=T)
head(gene.expression[,1:7])
```

```
##               X  ld_zt00_1  ld_zt04_1 ld_zt08_1 ld_zt12_1 ld_zt16_1 ld_zt20_1
## 1 ostta01g00010   9.664395  25.045279  30.81788  44.97020  32.61381  30.13931
## 2 ostta01g00020  15.688867   9.913202  10.36669  14.64151  11.66096  21.09974
## 3 ostta01g00030  16.108133  11.134813  13.85848  38.54982  24.16540  16.82981
## 4 ostta01g00040  59.247765  32.837433  27.26293  42.82092  52.59695  87.45710
## 5 ostta01g00050  27.909069  14.945981  12.04561  20.26003  28.25942  25.53805
## 6 ostta01g00060 248.044205 145.486374  68.38238  66.96495 224.52632 246.65002
```

```
gene.ids <- gene.expression$X

gene.expression <- as.matrix(gene.expression[,2:ncol(gene.expression)])
rownames(gene.expression) <- gene.ids
head(gene.expression[,1:6])
```

```
##                ld_zt00_1  ld_zt04_1 ld_zt08_1 ld_zt12_1 ld_zt16_1 ld_zt20_1
## ostta01g00010   9.664395  25.045279  30.81788  44.97020  32.61381  30.13931
## ostta01g00020  15.688867   9.913202  10.36669  14.64151  11.66096  21.09974
## ostta01g00030  16.108133  11.134813  13.85848  38.54982  24.16540  16.82981
## ostta01g00040  59.247765  32.837433  27.26293  42.82092  52.59695  87.45710
## ostta01g00050  27.909069  14.945981  12.04561  20.26003  28.25942  25.53805
## ostta01g00060 248.044205 145.486374  68.38238  66.96495 224.52632 246.65002
```

The current version of *Ostreococcus tauri* genome available from **here** identifies 7668 genes. In our experiment only 8 genes were never expressed and 40 genes never presented an expression level greater than 1 FPKM. This shows that practically the entire transcriptome of Ostreococcus is expressed under the seasonal and diurnal cycles studied in this project.

```
number.genes <- nrow(gene.expression)
number.genes
```

```
## [1] 7668
```

```
length(which(apply(X = gene.expression,MARGIN = 1,FUN = max) == 0))
```

```
## [1] 8
```

```
length(which(apply(X = gene.expression,MARGIN = 1,FUN = max) < 1))
```

```
## [1] 40
```

We focus on the data generated under long and short days cycles by extracting them from the gene expression matrix. The resulting matrix has 7668 rows representing genes and 36 columns. This number of columns correspond to 36 different data points, each day is represented by 6 data points and we took samples for three days under both long and short day conditions.

```
ld.zt <- paste("ld",paste0("zt",sprintf(fmt = "%02d",seq(from=0,to=20,by=4))),sep="_")
ld.zt.i <- sapply(X = ld.zt,FUN = function(x){ paste(x,1:3,sep="_")})
sd.zt <- paste("sd",paste0("zt",sprintf(fmt = "%02d",seq(from=0,to=20,by=4))),sep="_")
sd.zt.i <- sapply(X = sd.zt,FUN = function(x){ paste(x,1:3,sep="_")})

ld.sd.gene.expression <- gene.expression[,c(ld.zt.i,sd.zt.i)]
head(ld.sd.gene.expression[,1:6])
```

```
##                ld_zt00_1  ld_zt00_2  ld_zt00_3  ld_zt04_1  ld_zt04_2 ld_zt04_3
## ostta01g00010   9.664395   8.753878  21.967098  25.045279  16.192572  29.56268
## ostta01g00020  15.688867  14.411269  15.126039   9.913202   9.555948  17.26569
## ostta01g00030  16.108133  18.037844   5.178177  11.134813  18.246208  13.43679
## ostta01g00040  59.247765  62.436951  50.582394  32.837433  20.027222  27.73791
## ostta01g00050  27.909069  27.673790  28.704229  14.945981  18.921703  23.65231
## ostta01g00060 248.044205 304.855804 101.986931 145.486374 134.537064 113.76479
```

```
dim(ld.sd.gene.expression)
```

```
## [1] 7668   36
```

We perform **Principal Component Analysis** and a **Hierarchical clustering** in order to uncover the underlying structure in our data. We use the packages FactoMineR and factoextra and reformat the data as needed for the function PCA.

```
library(FactoMineR)
library(factoextra)
```

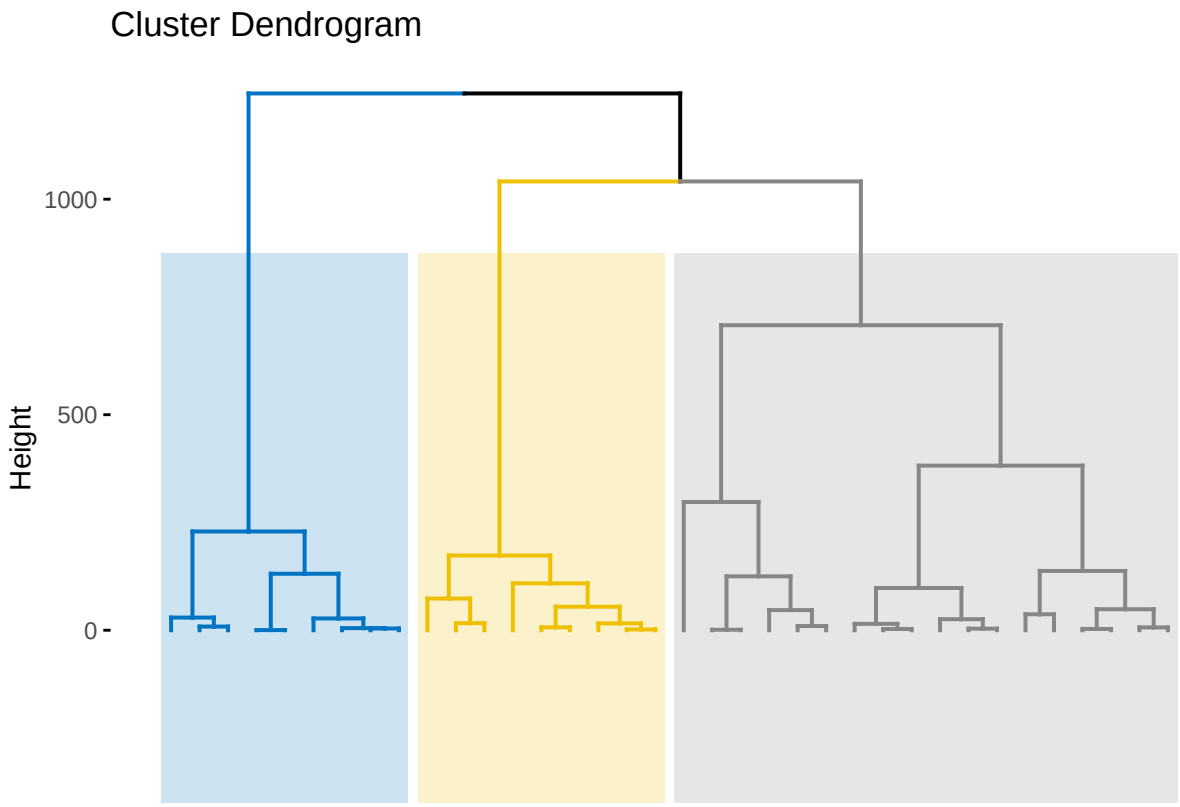
```
## Loading required package: ggplot2
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
pca.gene.expression.ld.sd <- data.frame(colnames(ld.sd.gene.expression),t(ld.sd.gene.expression))
colnames(pca.gene.expression.ld.sd)[1] <- "Time point"

res.pca.ld.sd <- PCA(pca.gene.expression.ld.sd, graph = FALSE,scale.unit = TRUE,quali.sup = 1 )
res.hcpc.ld.sd <- HCPC(res.pca.ld.sd, graph=FALSE)
fviz_dend(res.hcpc.ld.sd,k=3,
```

```
          cex = 0,                          # Label size
          palette = "jco",                  # Color palette see ?ggpubr::ggpar
          rect = TRUE, rect_fill = TRUE,    # Add rectangle around groups
          rect_border = "jco",              # Rectangle color
          type="rectangle",
          labels_track_height = 400         # Augment the room for labels
)
```



The transcriptomes corresponding to the same time during the three different days under both LD and SD conditions tend to cluster together. This indicates a high circadian synchronization in our cultures. Using hierarchical clustering, the 36 transcriptomes under LD and SD conditions assemble together into three different groups. The first cluster corresponds to **midday**. The transcriptomes at time points ZT4 and ZT8 under LD and ZT4 under SD constitute this cluster. These time points correspond to the moments of maximal incident light irradiance under both LD and SD conditions. The second cluster conforms the **dusk** group. Here the transcriptomes at time points ZT12 and ZT16 under LD and ZT8 under SD are found. These time points coincide with the end of the light period in both LD and SD conditions when light irradiance is low. The third cluster represents **night/dawn** and comprises the transcriptomes at time points ZT20, ZT0 under LD and ZT12, ZT16, ZT20 and ZT0 under SD. The transcriptomes at time points in the LD and SD nights or dark periods constitute two distinct groups suggesting noticeable differences in the transcriptomic responses during the night under LD and SD conditions. It is also noteworthy the higher similarity between the dusk, night/dawn transcriptomes when compare to the midday ones.