Code ▾

Case Study 1: Exploratory Data Analysis Of DNA, RNA, environment, and fitness

Course: Data Analysis and Visualization in R

Authors:

- Farag, Salma
- Gomes, Guilherme
- Holmberg, Olle
- Zafar, Atiqa

This notebook contains Exploratory Data Analysis for Case Study 1. It summarizes the methods we have used to explore the genetics data set and investigate relationships between fitness of organism, genotype, gene expression and environment.

Published: http://rpubs.com/atiqazafar/EDA_1 (http://rpubs.com/atiqazafar/EDA_1)

About the Genotype and Fitness Dataset

For the Genotype analysis, we are given genetic and fitness data for 185 yeast strains. These strains are offsprings, or segregants, of a cross between two parental strains "Lab strain" and "Wild isolate".

Load and examine the Genotype and Fitness Dataset

Hide

```
options(warn=-1)
#load packages
library(data.table)
library(magrittr)
library(ggplot2)
library(gridExtra)
library(grid)
library(plotly)
library(dplyr)
#set data dir
DATA_DIR <- c('C:/Users/User/Desktop/DATR-CaseStudy-1/')
marker_file <- file.path(DATA_DIR, 'marker.txt')
growth_file <- file.path(DATA_DIR, 'growth.txt')
genotype_file <- file.path(DATA_DIR, 'genotype.txt')
#read in data
marker <- read.delim(marker_file)
growth <- read.delim(growth_file)
genotype <- read.delim(genotype_file)
#set style for plotting
outlook <- theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
                 panel.background = element_blank(), axis.line = element_line(colour = "black"), plot.title = element_text(h
just = 0.5))
```

Let's have a first look at a summary of the columns in the genotype file:

Hide

```
#str(genotype)
dim(genotype)
```

```
[1]  158 1001
```

Some key things to notice here are:

- There are 158 rows or cases in the data set representing the yeast strains.
- The data set has 1001 variables or columns representing the genetic markers.
- At each marker, the genotype values either "Lab strain" or "Wild isolate".

Next, the markers file shows the genomic coordinates of the markers (chromosome, start, and stop).

Hide

```
str(marker)
```

```
'data.frame':	1000 obs. of  4 variables:
 $ id   : Factor w/ 1000 levels "mrk_1","mrk_1000",..: 1 284 393 501 618 726 834 951 60 168 ...
 $ chrom: Factor w/ 16 levels "chr01","chr02",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ start: int  1512 29161 38275 47695 56059 61507 68448 71737 80562 86129 ...
 $ end  : int  2366 29333 38317 47695 56059 61645 68448 71746 80562 86236 ...
```

The fitness dataset contains the growth rate of each strain for 5 growth media. The growth rate is measured as the nnumber of Generations per day.

Hide

```
str(growth)
```

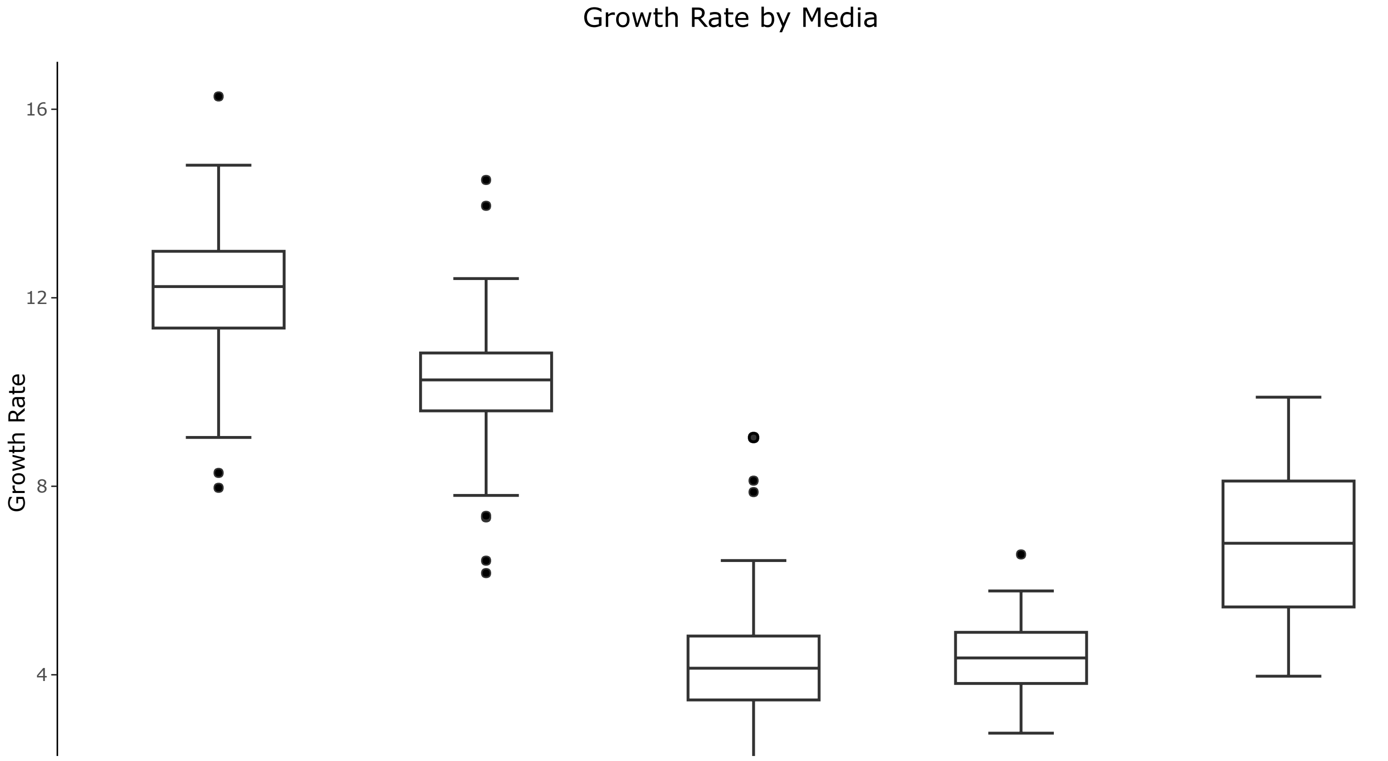
```
'data.frame':	158 obs. of  6 variables:
 $ strain  : Factor w/ 158 levels "seg_01B","seg_01C",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ YPD     : num  12.6 10.8 12.8 10.3 11.1 ...
 $ YPD_BPS : num  10.46 11.63 10.42 9.1 9.26 ...
 $ YPD_Rapa: num  2.5 NA 3.14 4.31 3.55 ...
 $ YPE     : num  5.27 5.37 5.58 3.26 3.82 ...
 $ YPMalt  : num  6.72 7.43 6.91 4.92 4.41 ...
```

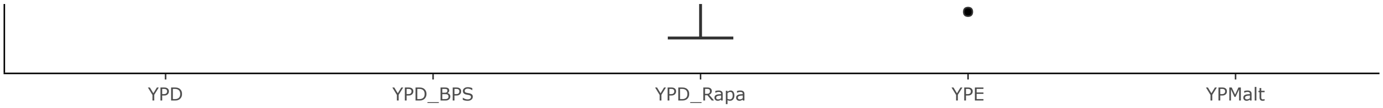
We can visualize the growth rate of the yeast strains for the 5 media using box-plots to understand how the growth rate is distributed in each medium (environment) for strains.

Hide

```
#melt growth for plotting
growth.m <- melt(growth, id.var = "strain")
#plot boxplot by media
p <- ggplot(growth.m, aes(x=variable, y=value)) +
  geom_boxplot() +
  xlab("") + ylab("Growth Rate") +
  ggtitle("Growth Rate by Media") + outlook
ggplotly(p)
```

```
We recommend that you use the dev version of ggplot2 with `ggplotly()`
Install it with: `devtools::install_github('hadley/ggplot2')`
Removed 42 rows containing non-finite values (stat_boxplot).
```





Examining these box plots, we note that **the median growth rate of yeast strains in YPD(glucose) is more than the the median growth rate in other media.**

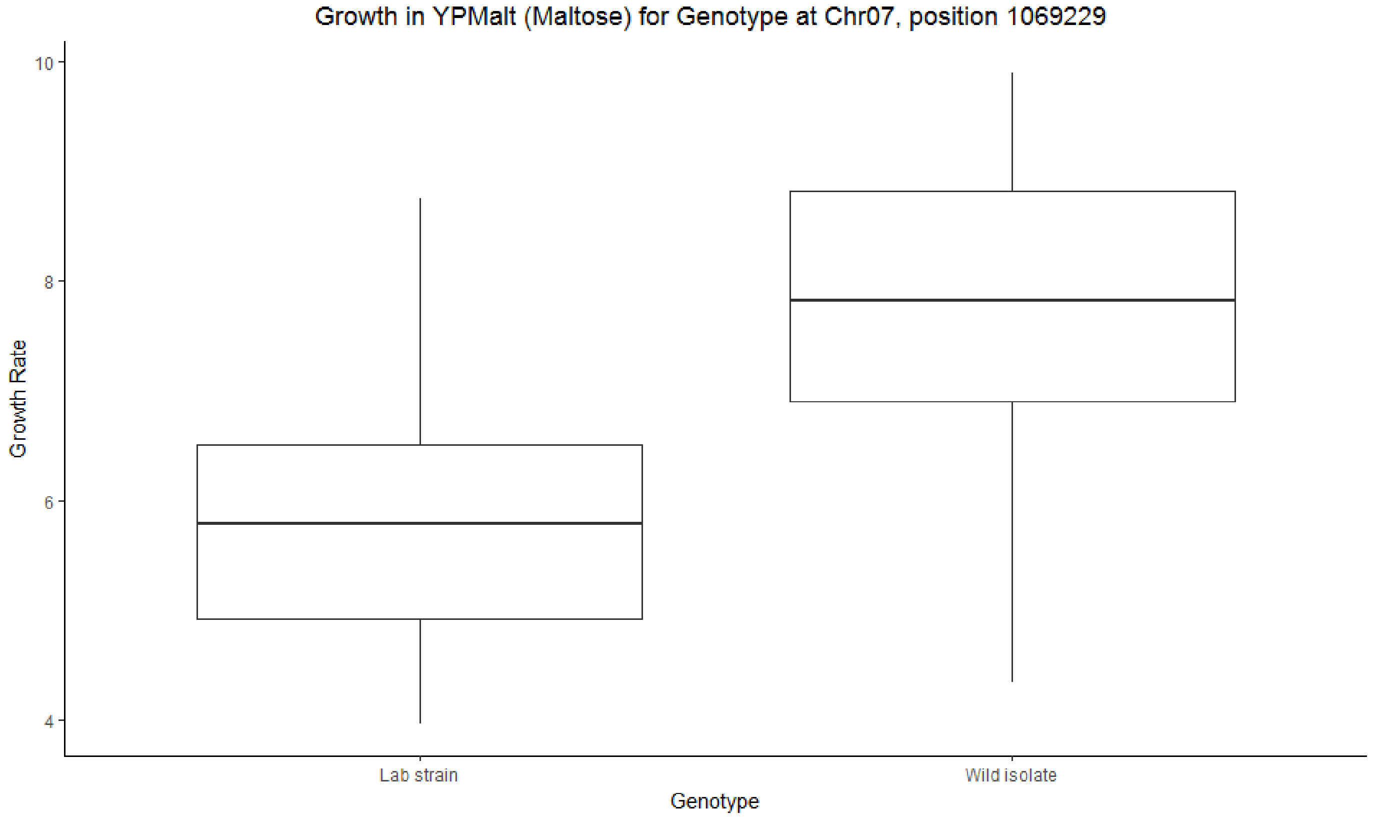
Does Genotype affect Fitness?

Next, we can investigate if genotype affects fitness to understand how good a particular genotype is at leaving offspring in the next generation in relation to how good the other genotype is.

Let's begin by plotting the distribution of Cellular growth in maltose (YPMalt) for genotype at chr07, postion 1069229.

Hide

```
# genotype at chr07, postion 1069229
mygeno <- genotype[, which(marker$chrom=="chr07" & marker$start== 1069229)]
names(mygeno) <- genotype$strain
# Growth in YPMalt (Maltose)
ggplot(data = growth , aes(x=mygeno, y=YPMalt)) +
  geom_boxplot()   +
  xlab("Genotype") + ylab("Growth Rate") + outlook +
  ggtitle("Growth in YPMalt (Maltose) for Genotype at Chr07, position 1069229 ")
```



We can note that this particular genetic marker has a strong association between the genotype and growth rate.

To see an overall picture, we can measure the growth averaged over yeast strains and all markers to investigate the overall cellular growth for each genotype in the different media. We also plot the standard deviation.

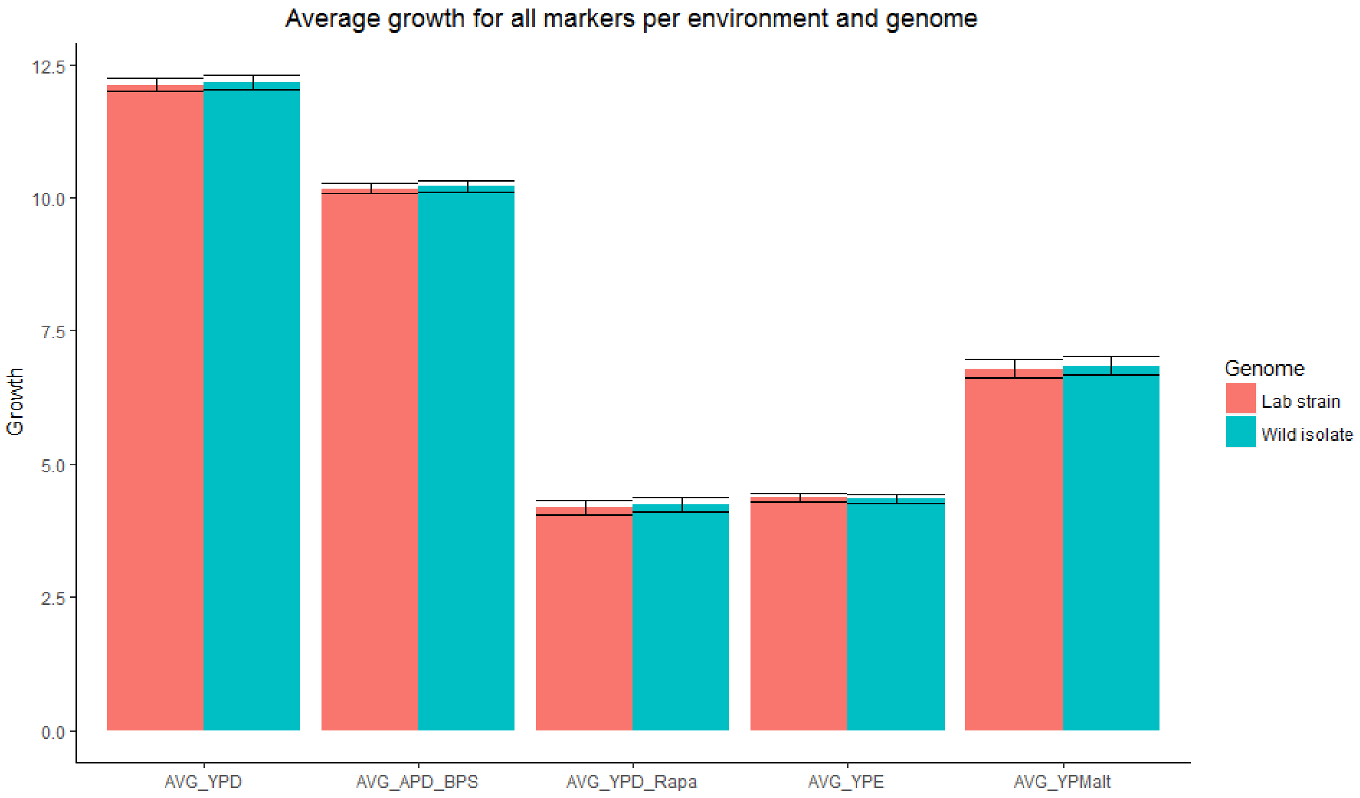
Hide

```
#merge for data for later plotting
genotype_dt <- as.data.table(genotype)
growth_dt <- as.data.table(growth)
base <- merge(genotype_dt, growth_dt)
base_long <- melt(base, id.vars = c("strain","YPD", "YPD_BPS", "YPD_Rapa", "YPE" ,"YPMalt"))
#adding all avarages in one table
avg <- base_long[,mean(YPD, na.rm = TRUE), by = .(variable, value)]
avg[,"AVG_APD_BPS" := base_long[,mean(YPD_BPS, na.rm = TRUE), by = .(variable, value)][,3]]
avg[,"AVG_YPD_Rapa" := base_long[,mean(YPD_Rapa, na.rm = TRUE), by = .(variable, value)][,3]]
avg[,"AVG_YPE" := base_long[,mean(YPE, na.rm = TRUE), by = .(variable, value)][,3]]
avg[,"AVG_YPMalt" := base_long[,mean(YPMalt, na.rm = TRUE), by = .(variable, value)][,3]]
#set names for data columns
setnames(avg, old = c("variable","value", "V1"), new=c("ID","Genome","AVG_YPD"))
#melt table
avg_long <- melt(avg,id.variable = "Nutrition", measure.name = 3:7)
```

```
To be consistent with reshape2's melt, id.vars and measure.vars are internally guessed when both are 'NULL'. All non-numeri
c/integer/logical type columns are conisdered id.vars, which in this case are columns [ID, Genome]. Consider providing at le
ast one of 'id' or 'measure' vars in future.
```

Hide

```
#avg_long[, mean(value, na.rm = TRUE), by = Genome]
#derive marker avarages and standard deviationn for each nutrition
avg_nutrition_markers <- avg[,.(mean(AVG_YPD),mean(AVG_APD_BPS),mean(AVG_YPD_Rapa), mean(AVG_YPE), mean(AVG_YPMalt)), by = G
enome]
sd_nutrition_markers <- avg[,.(sd(AVG_YPD),sd(AVG_APD_BPS),sd(AVG_YPD_Rapa), sd(AVG_YPE), sd(AVG_YPMalt)), by = Genome]
setnames(sd_nutrition_markers, old = 2:6, new = c("AVG_YPD","AVG_APD_BPS","AVG_YPD_Rapa", "AVG_YPE", "AVG_YPMalt"))
setnames(avg_nutrition_markers, old = 2:6, new = c("AVG_YPD","AVG_APD_BPS","AVG_YPD_Rapa", "AVG_YPE", "AVG_YPMalt"))
#plot
avg_n_merkers_long <- melt(avg_nutrition_markers, id.vars = "Genome")
sd_n_merkers_long <- melt(sd_nutrition_markers, id.vars = "Genome")
#plot
ggplot(data = avg_n_merkers_long, aes(x=variable, y=value,fill = Genome)) + geom_col(position = "dodge") +
  geom_errorbar(aes(ymax=value + sd_n_merkers_long$value, ymin=value - sd_n_merkers_long$value), position = "dodge") +
  outlook +ylab("Growth") + xlab("") + ggtitle("Average growth for all markers per environment and genome")
```

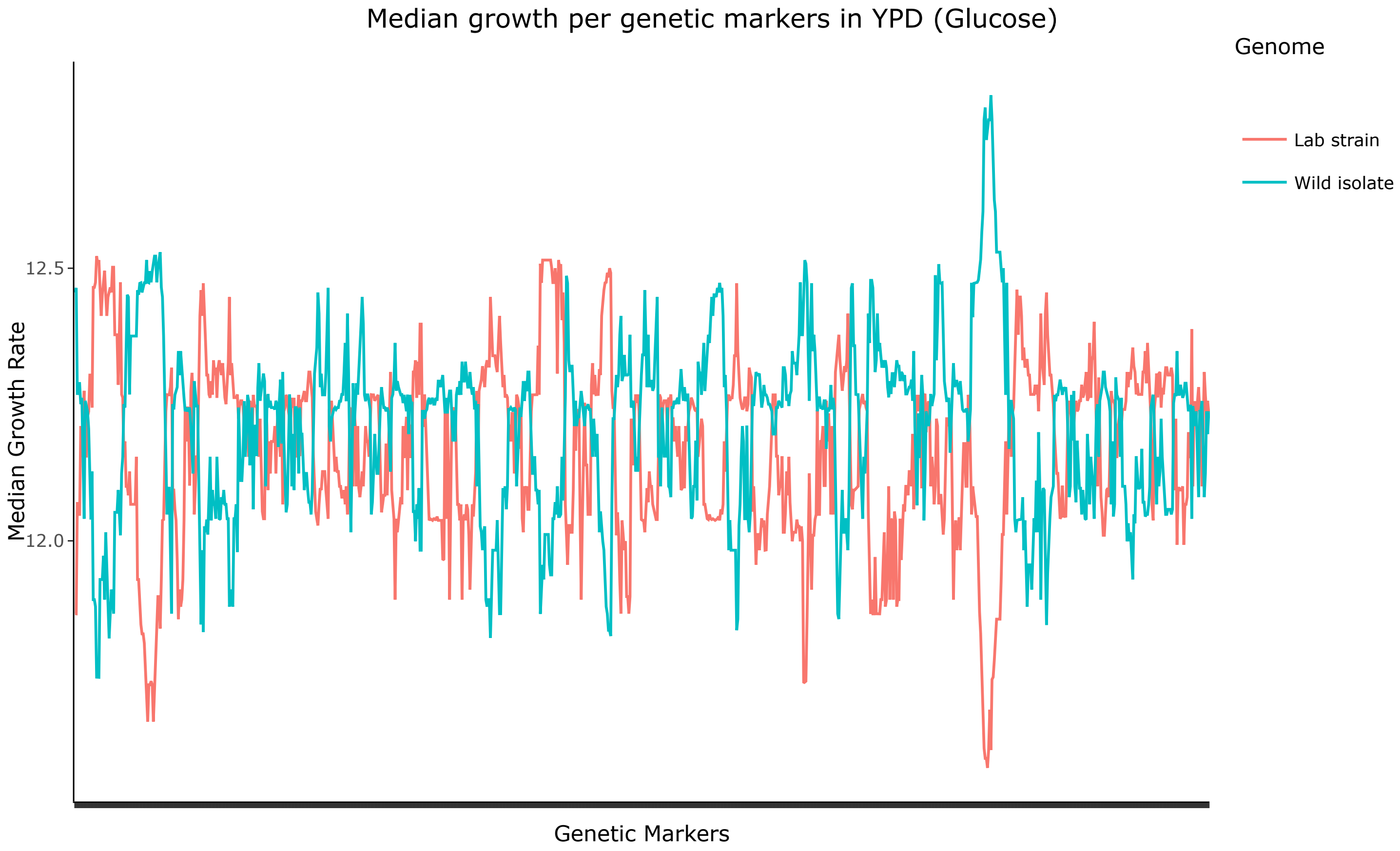


It’s clear to us that this plot shows only the growth variation per environment and misses out the variation due to genotype.To follow up, we plot lineplots for the median growth per marker in YPD (Glucose) for each Genotype.

Hide

```
#adding all median in one table
med <- base_long[,median(YPD, na.rm = TRUE), by = .(variable, value)]
med[,"MEDIAN_APD_BPS" := base_long[,median(YPD_BPS, na.rm = TRUE), by = .(variable, value)][,3]]
med[,"MEDIAN_YPD_Rapa" := base_long[,median(YPD_Rapa, na.rm = TRUE), by = .(variable, value)][,3]]
med[,"MEDIAN_YPE" := base_long[,median(YPE, na.rm = TRUE), by = .(variable, value)][,3]]
med[,"MEDIAN_YPMalt" := base_long[,median(YPMalt, na.rm = TRUE), by = .(variable, value)][,3]]
#set names for data columns
setnames(med, old = c("variable","value", "V1"), new=c("ID","Genome","MEDIAN_YPD"))
#Plotting median growth per marker in YPD
p <- ggplot(data = med, aes(x = ID, y = MEDIAN_YPD, color = Genome)) +
  geom_line(aes(group = Genome)) +
  ylab("Median Growth Rate") + xlab("Genetic Markers") +
  theme(axis.text.x = element_blank())+ outlook +
  ggtitle("Median growth per genetic markers in YPD (Glucose)")
ggplotly(p)
```

```
We recommend that you use the dev version of ggplot2 with `ggplotly()`
Install it with: `devtools::install_github('hadley/ggplot2')`
```



Here, we can look at the lineplots for the Wild Genotype and Lab Genotype respectively to visualize how the median growth per marker varies for each Genotype in the Glucose medium. To visualize these variations further, we can plot the whole distribution of Cellular growth for each genetic markers using box-plots. Note that we're keeping the environment/medium constant (maltose/YPMalt) but varying the marker position.

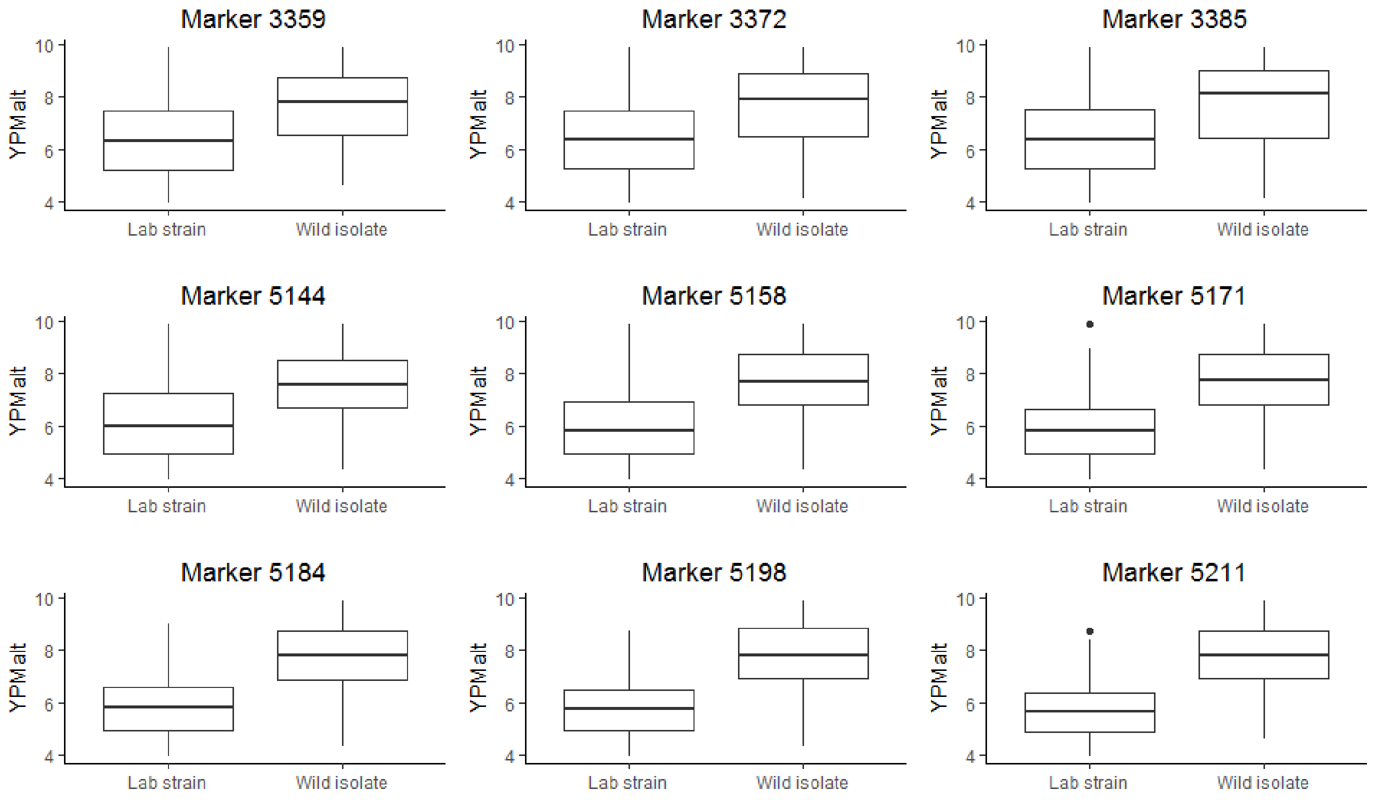
Hide

```
# markers/genome medians
medians <- base_long[, lapply(.SD, median, na.rm = TRUE),
                       .SDcols = c("YPD", "YPD_BPS", "YPD_Rapa", "YPE", "YPMalt"),
                       by = .(marker = variable, genome = value)]
medians$genome <- gsub(" ", "_", medians$genome)
ypmalt_diff <- m<- dcast(medians, marker ~ genome, value.var = "YPMalt") %>%
  mutate(Diff = Wild_isolate - Lab_strain) %>%
  select(marker, Diff)
ypmalt_diff$Diff = abs(ypmalt_diff$Diff)
ypmalt_diff <- as.data.table(ypmalt_diff)
# compute affected markers
markers_affected_by_genotype <- ypmalt_diff[Diff > 4.5*mean(Diff)]
setnames(markers_affected_by_genotype, "marker", "id")
query_markers <- inner_join(marker, markers_affected_by_genotype, by="id")
```

```
Column `id` joining factors with different levels, coercing to character vector
```

Hide

```
# markers grid plot
plots = lapply(query_markers$id,
               function(.x) ggplot(growth, aes(( genotype[, .x])[strain], YPMalt)) + geom_boxplot() + outlook + xlab("") + g
gtitle(sub("mrk_", "Marker ",.x)))
do.call(grid.arrange,  plots)
```



Here, we can note that **wild isolates** show a clear advantage over **lab strains** in some markers and contribute to more growth in a strain, specially in **YPMalt** than any other enviroment.

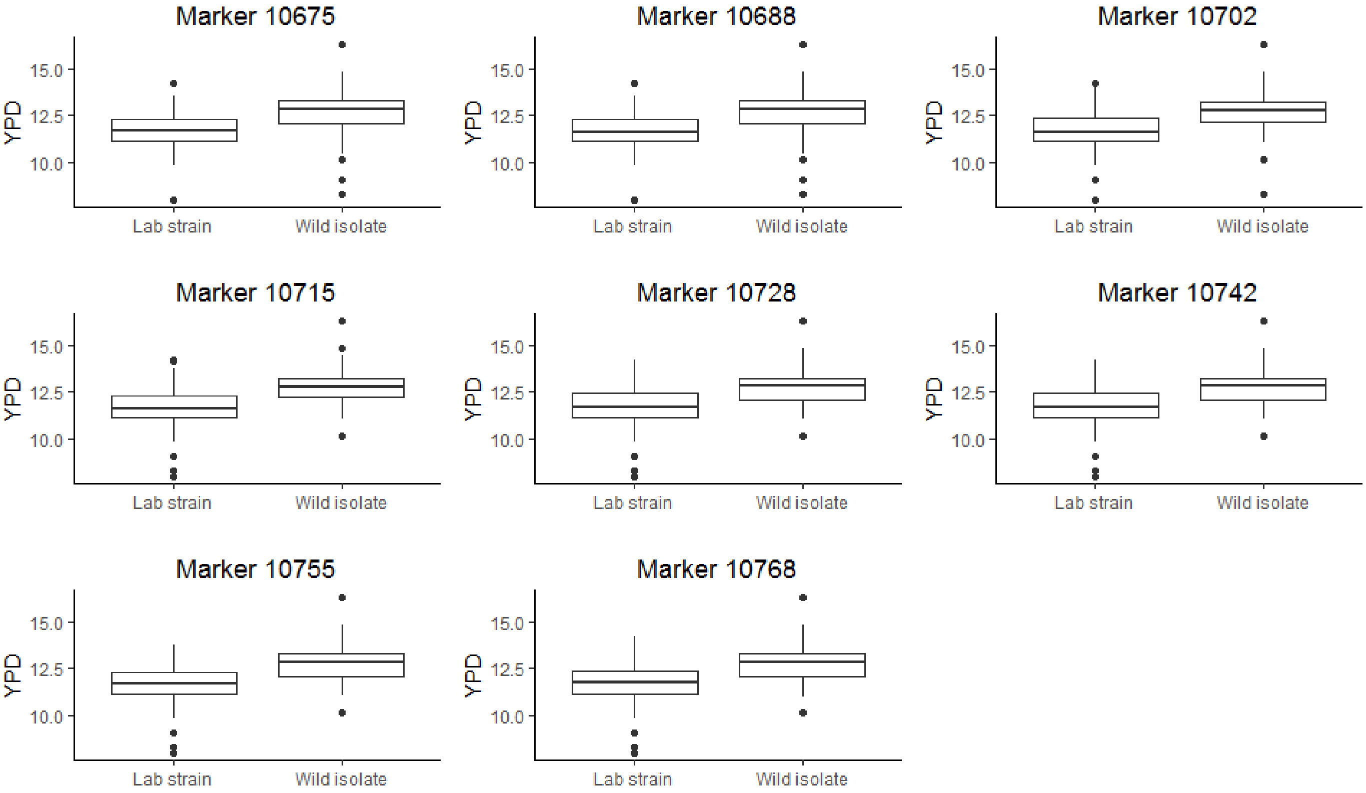
Hide

```
ypd_diff <- dcast(medians, marker ~ genome, value.var = "YPD") %>%
  mutate(Diff = Wild_isolate - Lab_strain) %>%
  select(marker, Diff)
ypd_diff$Diff = abs(ypd_diff$Diff)
ypd_diff <- as.data.table(ypd_diff)
ype_diff <- dcast(medians, marker ~ genome, value.var = "YPE") %>%
  mutate(Diff = Wild_isolate - Lab_strain) %>%
  select(marker, Diff)
ype_diff$Diff = abs(ype_diff$Diff)
ype_diff <- as.data.table(ype_diff)
# compute affected markers
markers_affected_by_genotype <- ypd_diff[Diff > 4.1*mean(Diff)]
setnames(markers_affected_by_genotype, "marker", "id")
query_markers <- inner_join(marker, markers_affected_by_genotype, by="id")
```

```
Column `id` joining factors with different levels, coercing to character vector
```

Hide

```
# markers grid plot
plots = lapply(query_markers$id,
               function(.x) ggplot(growth, aes(( genotype[, .x])[strain], YPD)) + geom_boxplot() +  outlook + xlab("") + ggt
itle(sub("mrk_", "Marker ",.x)))
do.call(grid.arrange,  plots)
```



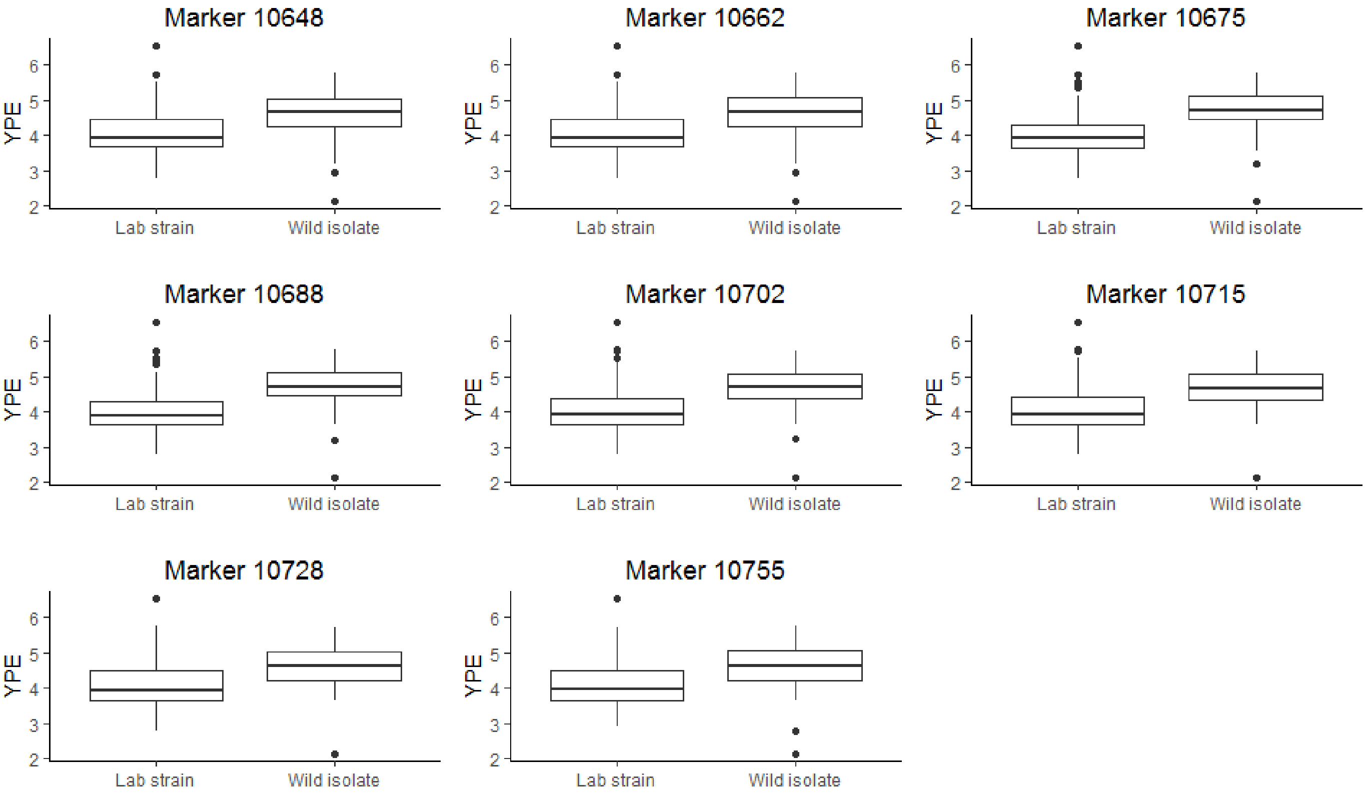
Hide

```
# compute affected markers
markers_affected_by_genotype <- ype_diff[Diff > 3.5*mean(Diff)]
setnames(markers_affected_by_genotype, "marker", "id")
query_markers <- inner_join(marker, markers_affected_by_genotype, by="id")
```

```
Column `id` joining factors with different levels, coercing to character vector
```

Hide

```
# markers grid plot
plots = lapply(query_markers$id,
               function(.x) ggplot(growth, aes(( genotype[, .x])[strain], YPE)) + geom_boxplot() +  outlook + xlab("") + ggt
itle(sub("mrk_", "Marker ",.x)))
do.call(grid.arrange,  plots)
```



When querying to get the markers with the highest performance difference between the wild isolates and lab strains, it came to our attention that the environments **YPD** and **YPE** have a very similar effect on growth and correlate positively with markers like: **10675**, **10688**, **10702**, **10715**, **10728** and **10755** being affected by the genotype the most and by roughly the same amount.

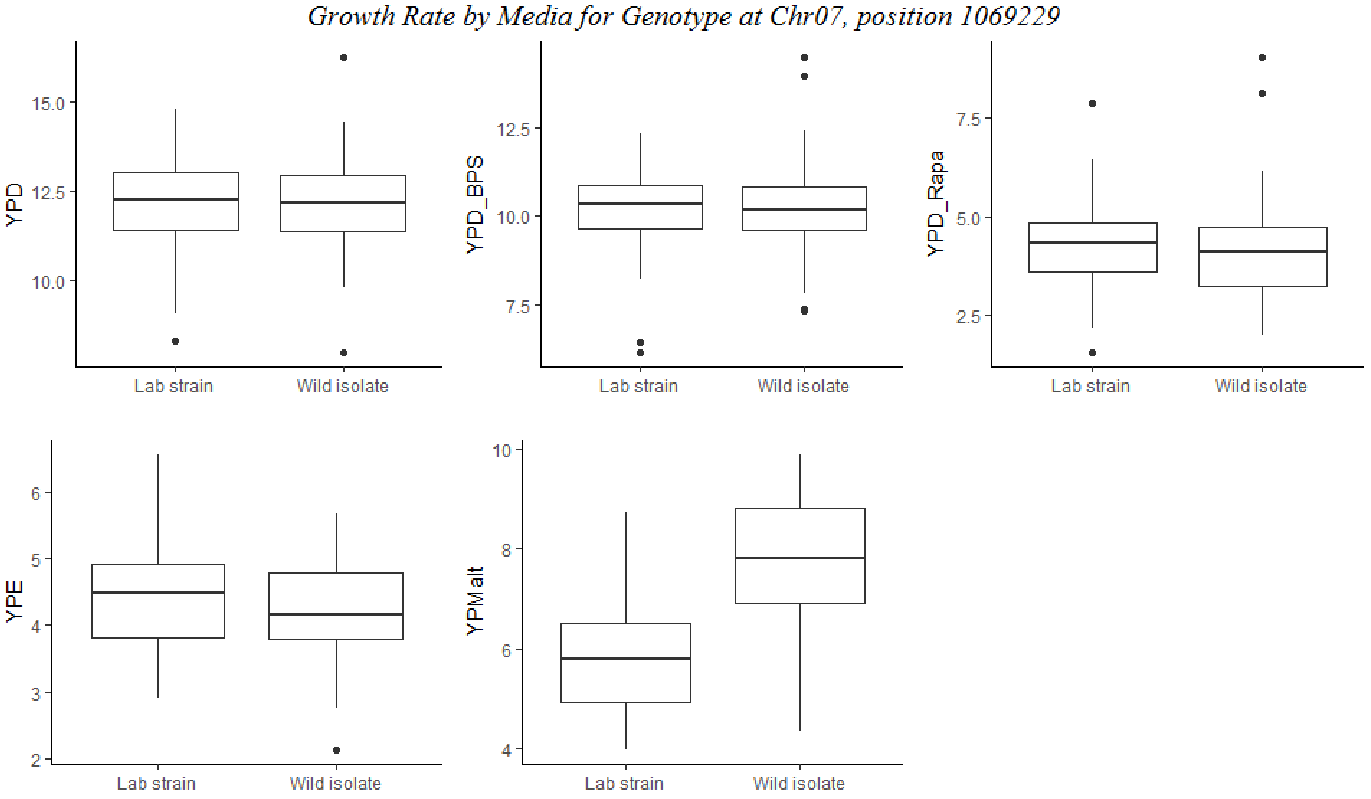
Does genotype affects fitness independently of Environment?

Next, we can investigate if genotype affects fitness independently of Environment to understand how good a particular genotype is at leaving offspring in the next generation in relation to how good the other genotype is, irrespective of the surrounding environmental influences.

Let’s begin by plotting the distribution of Cellular growth for genotype at chr07, postion 1069229.

Hide

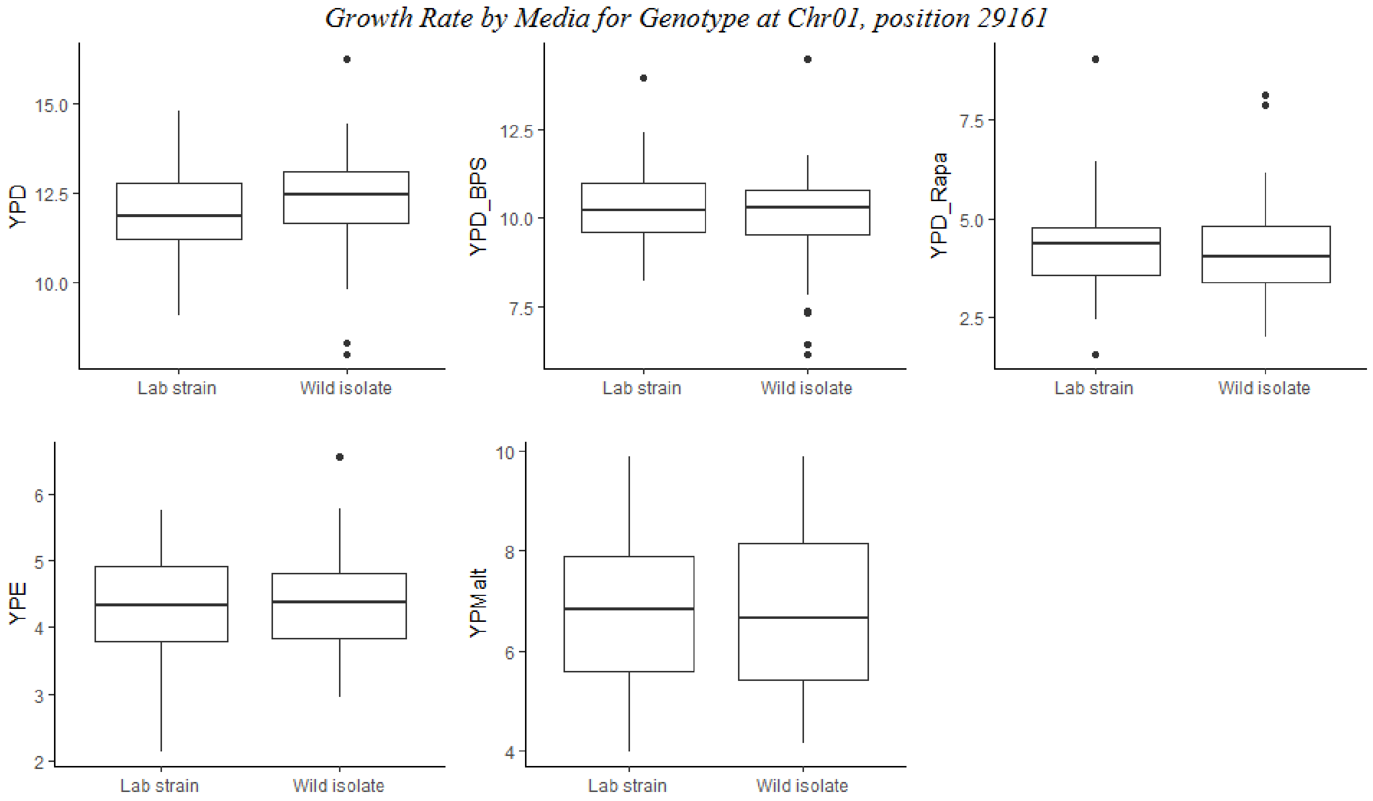
```
mygeno <- genotype[, which(marker$chrom=="chr07" & marker$start== 1069229)]
names(mygeno) <- genotype$strain
#convert growth to a data table
growth_dt <- as.data.table(growth)
#style specifications
pbox <- geom_boxplot()
x_lab <- xlab("")
##create plot for grid
p1 <- ggplot(data = growth , aes(x=mygeno, y=YPD)) + pbox  + x_lab +outlook
p2 <- ggplot(data = growth , aes(x=mygeno, y=YPD_BPS)) + pbox  + x_lab +outlook
p3 <- ggplot(data = growth , aes(x=mygeno, y=YPD_Rapa)) + pbox + x_lab +outlook
p4 <- ggplot(data = growth , aes(x=mygeno, y=YPE)) + pbox + x_lab +outlook
p5 <- ggplot(data = growth , aes(x=mygeno, y=YPMalt)) + pbox + x_lab +outlook
#assemble plot on grid
grid.arrange(p1,p2,p3,p4,p5, nrow=2, top=textGrob("Growth Rate by Media for Genotype at Chr07, position 1069229", gp=gpar(fo
ntsize=15,font=8)))
```



We note that there is a variation in cellular growth in maltose (YPMalt) for genotype at chr07, postion 1069229 but the pattern is not the same in other media. Let's look at another genetic marker.

Hide

```
mygeno <- genotype[, which(marker$chrom=="chr01" & marker$start== 29161)]
names(mygeno) <- genotype$strain
#convert growth to a data table
growth_dt <- as.data.table(growth)
#style specifications
pbox <- geom_boxplot()
x_lab <- xlab("")
##create plot for grid
p1 <- ggplot(data = growth , aes(x=mygeno, y=YPD)) + pbox  + x_lab +outlook
p2 <- ggplot(data = growth , aes(x=mygeno, y=YPD_BPS)) + pbox  + x_lab +outlook
p3 <- ggplot(data = growth , aes(x=mygeno, y=YPD_Rapa)) + pbox + x_lab +outlook
p4 <- ggplot(data = growth , aes(x=mygeno, y=YPE)) + pbox + x_lab +outlook
p5 <- ggplot(data = growth , aes(x=mygeno, y=YPMalt)) + pbox + x_lab +outlook
#assemble plot on grid
grid.arrange(p1,p2,p3,p4,p5, nrow=2, top=textGrob("Growth Rate by Media for Genotype at Chr01, position 29161", gp=gpar(fontsize=15,font=8)))
```


Growth Rate by Media for Genotype at Chr01, position 29161

Again, we note that the pattern of cellular growth per Genotype varies in each medium, showing that the growth is strongly environment-specific and that the genotype does not affect fitness independently of environment.

About the Gene Expression Dataset

The second dataset provided in this case study is about Gene Expressions (the production of RNA for a given gene). RNA is produced during the molecular process of transcription and depends on the environment or the genotype.

Load and examine the Gene Expression Dataset

Hide

```
#set path
gene_file <- file.path(DATA_DIR, 'gene.txt')
expression_file <- file.path(DATA_DIR, 'expression.txt')
#read in data
gene <- read.delim(gene_file)
expression <- fread(expression_file)
```

Let’s have a first look at a summary of the columns in the expression file:

Hide

```
#str(expression)
dim(expression)
```

```
[1] 8382  183
```

Some key things to notice here are:

- There are 8382 rows or cases in the data set representing the genes.
- The data set has 183 variables or columns representing the RNA expression level (on a logaritmic scale) for 8,382 genes in 183 samples, where one sample is one segregant grown in one media. Unlike for growth rate, obtaining transcription level is expensive. Note that we don’t have the data for every segregant grown in every media.

Next, the genes file shows the genomic coordinates of the genes (chromosome, start and stop, strand, type, source, novel, name).

Hide

```
str(gene)
```

```
'data.frame':	8382 obs. of  9 variables:
 $ chrom     : Factor w/ 16 levels "chr01","chr02",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ start     : int  6601 9392 11378 28205 29136 31860 32340 33720 34120 35638 ...
 $ end       : int  9080 9954 11715 28501 29934 32320 33710 34053 35616 35786 ...
 $ strand    : Factor w/ 2 levels "-","+": 2 2 2 2 2 2 2 2 2 2 ...
 $ name      : Factor w/ 8157 levels "CUT004","CUT006",..: 990 261 991 992 993 2691 2691 994 2690 995 ...
 $ commonName: Factor w/ 8155 levels "AAC1","AAC3",..: 5286 4525 5287 5288 5289 1493 1493 5290 373 5291 ...
 $ type      : Factor w/ 4 levels "CUT","ORF-T",..: 4 4 4 4 4 2 2 4 2 4 ...
 $ source    : Factor w/ 4 levels "genenv","SGD",..: 1 3 1 1 1 3 2 1 2 1 ...
 $ novel     : logi  TRUE FALSE TRUE TRUE TRUE FALSE ...
```

Note that we have four types of genes in total with 2 different novelity levels and 4 sources.

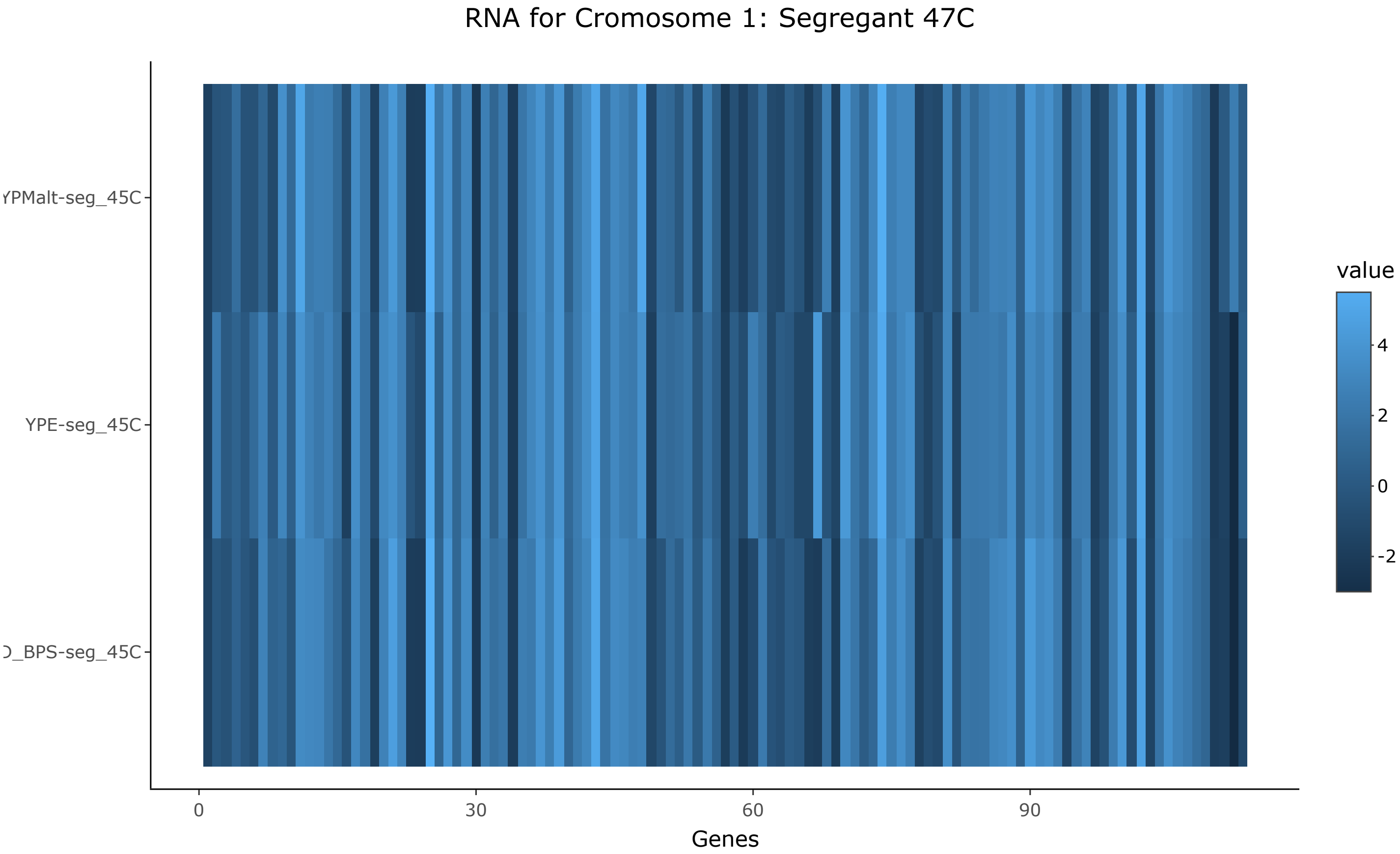
Is the gene expression under genetic or environmental control?

Firstly, Let's investigate if the gene expression under genetic or environmental control, and if it is more influenced by one or the other. For this, we can plot the RNA/Gene Expression for a particular segregant at a chromosome in different media. Take segregant 'seg_45C' in YPMalt (Maltose), YPE (Ethanol) and YPD_BPS (low iron) at Chromosome 01.

Hide

```
#select columns
indx <- grepl('seg_45C', colnames(expression))
#data for plot
heat_map <- expression[which(gene$chrom=="chr01")][,indx, with = FALSE]
heat_map[,rown := rownames(heat_map)]
heat_map_long <- melt(heat_map, id.vars = "rown")
p<- ggplot(heat_map_long, aes(x=as.integer(rown) , y=variable)) +
  geom_tile(aes(fill = value) , color = "white") +
  outlook  +ylab("") + xlab("Genes") +
  ggtitle("RNA for Cromosome 1: Segregant 47C")
ggplotly(p)
```

```
We recommend that you use the dev version of ggplot2 with `ggplotly()`
Install it with: `devtools::install_github('hadley/ggplot2')`
```



Note how the color (RNA value) changes over Genes, showing that the amount of RNA produced at each gene differ and is under genetic control. The color (RNA Value) also changes from one medium to another for any given gene, and is therefore under environment control as well. We conclude from the plot that the RNA value differs more per gene than per medium, showing that **transcription is affected more by genetic than environmental influences for chromosome 01 of segregant 47C**.

CONCLUSION

We can conclude that **genotype certainly affects fitness** but there are only certain genetic markers that associate with variations in cellular growth. Moreover, these growth variations have a **strong environment-specific pattern**, that is, some markers show growth variation in one environment but not in others due to genotype. Our Exploratory Data Analysis with the Genotype dataset shows

- YPD (peptone dextrose) is the most suitable environment for growth.
- The Wild Genotype has significant growth in YPMalt (Maltose) compared to all other media (including YPD (peptone dextrose))
- Some markers have noticable growth in both YPD (peptone dextrose) and YPE (ethanol). The Causes/Reasons behind such a pattern are definitely worth investigating.

We further investigate the overall cellular growth as the outcome of complicated molecular processes occurring in cells. Through our analysis of gene expression dataset, we have found out that the molecular process **transcription is under both the genetic and environmental influence**. For any given environment, the amount of RNA produced will vary drastically from gene to gene due to genotype. And for any given gene, this amount will also vary from one environment to another.