# PASS1A DMAQC data: analysis by BIC

```r
# Set the working directory to the folder with the data
# older:
# dmaqc_data_dir = "/Users/David/Desktop/MoTrPAC/data/pass_1a/dmaqc_pheno/v1"
# dictionary path
# dmaqc_dict_dir = "/Users/David/Desktop/MoTrPAC/data/pass_1a/dmaqc_pheno/dictionary/"
# official rlease:
dmaqc_data_dir = "/Users/David/Desktop/MoTrPAC/data/pass_1a/dmaqc_pheno/official/3-Data_Sets/"
# dictionary path
dmaqc_dict_dir = "/Users/David/Desktop/MoTrPAC/data/pass_1a/dmaqc_pheno/official/1-Data_Dictionary/"
all_csvs = list.files(dmaqc_data_dir,full.names = T) # get all files in dir
all_csvs = all_csvs[grepl(".csv$",all_csvs)] # make sure we take csv only
# read all files
csv_data = list()
for(fname in all_csvs){
  csv_data[[fname]] = read.csv(fname,stringsAsFactors = F)
}# sapply(csv_data,dim) # check the dimensions of the different datasets

all_dict_csvs = list.files(dmaqc_dict_dir,full.names = T) # get all files in dir
all_dict_csvs = all_dict_csvs[grepl(".csv$",all_dict_csvs)] # make sure we take csv only
# read all files
dict_data = list()
for(fname in all_dict_csvs){
  dict_data[[fname]] = read.csv(fname,stringsAsFactors = F)
}
#sapply(dict_data,dim)
```

## 1 Sanity check: Acute tests basic statistics

```r
# Get the acute test data
ac_test_data = csv_data[[which(grepl("Acute.Test",names(csv_data)))]]
dim(ac_test_data)
```

```
## [1] 108  23
```

```r
# check the time differences between start and end
test_times = as.difftime(ac_test_data$t_complete) - as.difftime(ac_test_data$t_start)
# table of the values: all except for on are 0.5 hours
table(test_times)
```

```
## test_times
## 0.466666666666667               0.5
##                 1               107
```

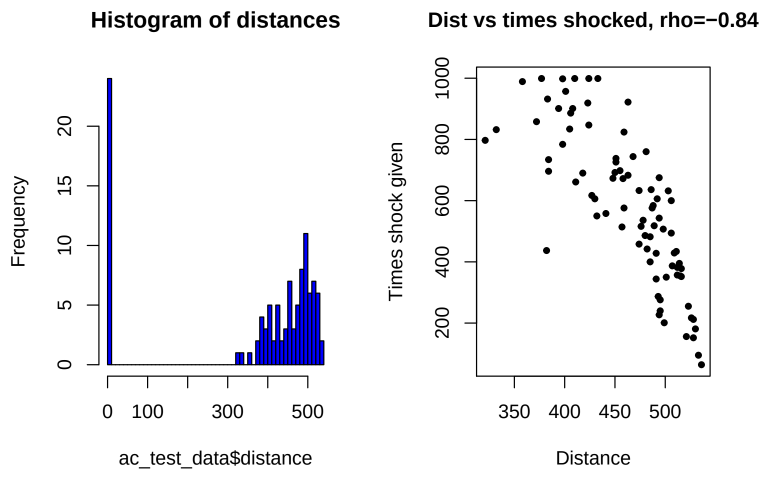
```r
# Get the comment of the sample that is not 0.5h
ac_test_data[test_times!=0.5,"comments"]
```

```
## [1] "Treadmill stopped 28:49 (mm:ss) into the acute bout due to problems with the other rat on the s
```

```r
ac_test_data$formatted_test_time = test_times
```

Next, we analyze the distances. We illustrate how these are a function of the shocks and sex/weight.

```r
# convert the shock lengths to numbers (seconds)
parse_shocktime<-function(x){
  arr = strsplit(x,split=":")[[1]]
  if(length(arr)<2){return(NA)}
  return(as.numeric(arr[1])*60+as.numeric(arr[2]))
}
tmp_x = ac_test_data$howlongshock
tmp_x = sapply(tmp_x, parse_shocktime)
ac_test_data$howlongshock = tmp_x
rm(tmp_x)

par(mfrow=c(1,2))
# histogram of distances
hist(ac_test_data$distance,col="blue",breaks=50,main = "Histogram of distances")

# Correlation between distance and number of shocks
# Get the indices of the samples with shock information -
# these the animals that did the acute test
timesshock_inds = !is.na(ac_test_data$timesshock)
# create a new dataframe with the selected animals
trained_animals_data = ac_test_data[timesshock_inds,]
sp_corr = cor(trained_animals_data$distance,
              trained_animals_data$timesshock,method="spearman")
plot(trained_animals_data$distance,trained_animals_data$timesshock,
     main=paste("Dist vs times shocked, rho=",format(sp_corr,digits = 2),sep=""),
     pch=20,ylab="Times shock given",xlab="Distance",cex.main=1.1)
```



```r
# A "smarter" analysis: regression of the distance using shock info
dist_lm  = lm(distance~timesshock+howlongshock+weight+days_start,
              data=trained_animals_data)
# Summary of the model, points to take: high R^2, significance of
# the features
summary(dist_lm)
```

```
##
## Call:
## lm(formula = distance ~ timesshock + howlongshock + weight +
##     days_start, data = trained_animals_data)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.1248  -1.0430   0.6867   2.4416   8.8814
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  562.127804   2.427681 231.549   <2e-16 ***
## timesshock     0.003171   0.003464   0.916    0.363
## howlongshock  -0.296415   0.005700 -52.004   <2e-16 ***
## weight        -0.147827   0.006563 -22.526   <2e-16 ***
## days_start     0.032731   0.024534   1.334    0.186
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.623 on 79 degrees of freedom
## Multiple R-squared:  0.9921, Adjusted R-squared:  0.9917
## F-statistic:  2466 on 4 and 79 DF,  p-value: < 2.2e-16
```

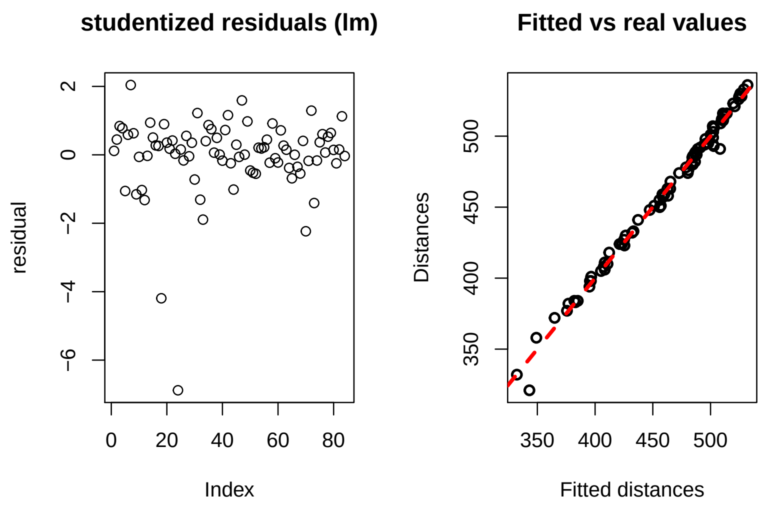
```
# We have some clear outliers:
library(MASS)
par(mfrow=c(1,2))
plot(studres(dist_lm),main="studentized residuals (lm)",ylab="residual")
# Select the top outliers and look at their comments
outliers = abs(studres(dist_lm)) > 2
# how many outliers have we selected?
sum(outliers)
```

```
## [1] 4
```

```
# their comments:
trained_animals_data[outliers,"comments"]
```

```
## [1] "Increased shock at 20 min."
## [2] "Treadmill stopped 28:49 (mm:ss) into the acute bout due to problems with the other rat on the sa
## [3] "Shock grid increased to 1.0 mA at 22 minutes. Treadmill bout stopped at 28:49 (mm:ss) due to ani
## [4] ""
```

```
# Plot the fitted values of the linear regression vs.
# the true distances
plot(dist_lm$fitted.values,trained_animals_data$distance,lwd=2,
     main="Fitted vs real values",ylab="Distances",xlab="Fitted distances")
abline(0,1,col="red",lty=2,lwd=3)
```



## 2 Site comparison

In some versions of the DMAQC data there is a single site. In this case this section will not result in an output.

```
# Load additional information about the animals
registr_data = csv_data[[which(grepl("Regist",names(csv_data)))]]
rownames(registr_data) = as.character(registr_data$pid)
# make the rownames in the test data comparable
rownames(trained_animals_data) = trained_animals_data$pid
# add sex to the trained animal data data frame
sex_key = c("Female","Male")
trained_animals_data$sex = sex_key[registr_data[rownames(trained_animals_data),"sex"]]

# Map site Ids to their names
site_names = c("910"="Joslin","930"="Florida")
trained_animals_data$site = site_names[as.character(trained_animals_data$siteID)]

# Sanity check: the numbers should be the same for both sites
table(ac_test_data$siteID)
```

```
##
## 910
## 108
```

```
table(trained_animals_data$site,trained_animals_data$sex)
```

```
##
##          Female Male
##   Joslin     42   42
```

```
run_wilcox<-function(x1,x2){
  return(wilcox.test(x1[x2==x2[1]],x1[x2!=x2[1]])$p.value)
}
# Compare the distances, shocks, and weight (if we have multiple site)
if (length(unique(ac_test_data$siteID))>1){
  par(mfrow=c(1,3),mar=c(10,4,4,4))
  # Site only
  p_dist = run_wilcox(trained_animals_data$distance,trained_animals_data$site)
  boxplot(distance~site,data=trained_animals_data,col="cyan",ylab="Distance",
        main=paste("Site vs. distance, p<",format(p_dist,digits = 2)),
        cex.main=1,las=2)
  p_timesshock = run_wilcox(trained_animals_data$timesshock,trained_animals_data$site)
  boxplot(timesshock~site,data=trained_animals_data,col="red",ylab="Times shocked",
        main=paste("Site vs. times shocked, p<",format(p_timesshock,digits = 3)),
        cex.main=1,las=2)
  p_w = run_wilcox(trained_animals_data$weight,trained_animals_data$site)
  boxplot(weight~site,data=trained_animals_data,col="cyan",ylab="Weight",
        main=paste("Site vs. weight, p=",format(p_w,digits = 2)),
        cex.main=1,las=2)
  # Site and sex
  par(mfrow=c(1,3),mar=c(10,4,4,4))
  boxplot(distance~site+sex,data=trained_animals_data,col="cyan",ylab="Distance",
        main="Site vs. distance",cex.main=1,las=2)
  boxplot(timesshock~site+sex,data=trained_animals_data,col="red",ylab="Times shocked",
        main="Site vs. times shocked",cex.main=1,las=2)
  boxplot(weight~site+sex,data=trained_animals_data,col="cyan",ylab="Weight",
        main="Site vs. weight",cex.main=1,las=2)

  # Regress time shocked and distance vs. site and sex
  summary(lm(timesshock~site+sex,data=trained_animals_data))
  summary(lm(distance~site+sex,data=trained_animals_data))
  }
```

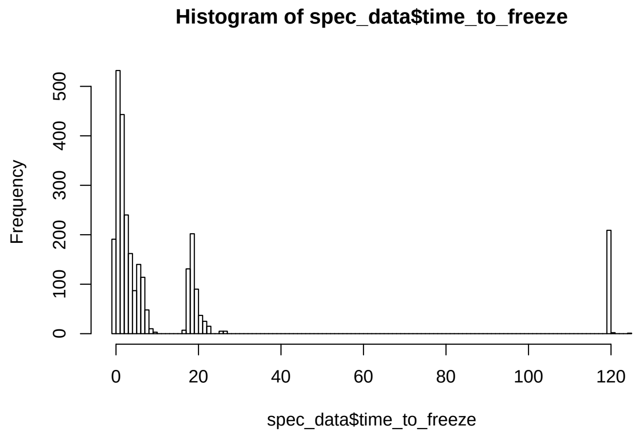
## 3 Sanity checks: Biospecimen data

```
# Analysis of biospecimen data
spec_data = csv_data[[which(grepl("Specimen.Processing.csv",names(csv_data)))]]
rownames(spec_data) = spec_data$labelid
# Parse the times and compute the difference between the freeze time and
# the collection time
time_to_freeze1 = as.difftime(spec_data$t_freeze,units = "mins") -
  as.difftime(spec_data$t_collection,units="mins")
# For some samples we have the edta spin time instead of the collection
# time, use these when there are no other options
time_to_freeze2 = as.difftime(spec_data$t_freeze,units = "mins") -
  as.difftime(spec_data$t_edtaspin,units="mins")
time_to_freeze = time_to_freeze1
# Fill in the NAs by taking the time between the edta spin and the freeze
table(is.na(time_to_freeze1),is.na(time_to_freeze2))
```

```
##
##         FALSE TRUE
##   FALSE     0 2182
##   TRUE    517    0
```

```
time_to_freeze[is.na(time_to_freeze1)] = time_to_freeze2[is.na(time_to_freeze1)]
spec_data$time_to_freeze = as.numeric(time_to_freeze)
spec_data$time_to_freeze_from_collection = as.numeric(time_to_freeze1)
spec_data$time_to_freeze_from_edta_spin = as.numeric(time_to_freeze2)
```

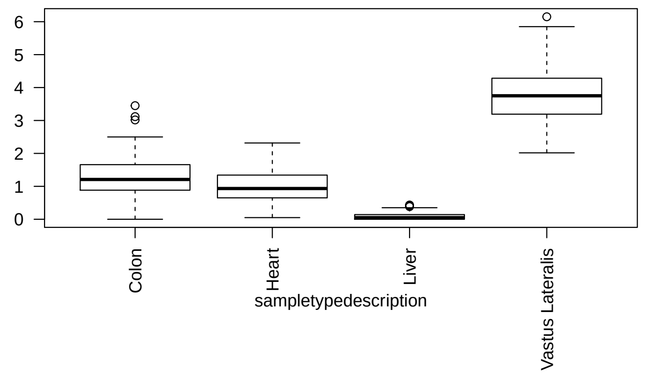
```r
hist(spec_data$time_to_freeze,breaks = 100)
```



```r
# Add site by name
site_names = c("910"="Joslin","930"="Florida")
spec_data$site = site_names[as.character(spec_data$siteid)]
table(spec_data$site)
```

```
##
## Joslin
##   2699
```

```r
inds = !is.na(time_to_freeze1)
inds = grepl("adipose",spec_data$sampletypedescription,ignore.case = T)
inds = grepl("heart",spec_data$sampletypedescription,ignore.case = T) |
  grepl("liver",spec_data$sampletypedescription,ignore.case = T) |
  grepl("colon",spec_data$sampletypedescription,ignore.case = T) |
  grepl("vastus",spec_data$sampletypedescription,ignore.case = T)
# Using site info:
# Here we use an interaction term and not addition as the R^2 is >2 times
# greater this way
if (length(unique(spec_data$site))>1){
  par(mar=c(10,2,2,2))
  boxplot(time_to_freeze~site:sampletypedescription,data=spec_data[inds,],
        ylab="Time to freeze",las=2)
  summary(lm(time_to_freeze~sampletypedescription:site,data=spec_data[inds,]))
}
# A single site
if (length(unique(spec_data$site))==1){
  par(mar=c(10,2,2,2))
  boxplot(time_to_freeze~sampletypedescription,data=spec_data[inds,],
        ylab="Time to freeze",las=2)
  summary(lm(time_to_freeze~sampletypedescription,data=spec_data[inds,]))
}
```



```
##
## Call:
## lm(formula = time_to_freeze ~ sampletypedescription, data = spec_data[inds,
##     ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.78765 -0.33731 -0.05216  0.27994  2.34568
##
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                            1.32052    0.05478  24.106  < 2e-16
## sampletypedescriptionHeart            -0.30046    0.07747  -3.878 0.000122
## sampletypedescriptionLiver            -1.23503    0.07747 -15.942  < 2e-16
## sampletypedescriptionVastus Lateralis  2.48380    0.07747  32.061  < 2e-16
##
## (Intercept)                           ***
## sampletypedescriptionHeart            ***
## sampletypedescriptionLiver            ***
## sampletypedescriptionVastus Lateralis ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5693 on 428 degrees of freedom
## Multiple R-squared:  0.8548, Adjusted R-squared:  0.8538
## F-statistic: 839.8 on 3 and 428 DF,  p-value: < 2.2e-16
```

## 4 Format the metadata table according to vial ids

We now use DMAQC's mapping of label ids to vial ids and use it to generate a single metadata table that we can share with other sites.

```
# Helper function for merging columns from data2 into data1
# The function makes sure there is no column duplications when
# adding information from data2 into data1
merge_avoid_col_dup<-function(data1,data2,by_col){
  data2_cols = c(by_col,setdiff(colnames(data2),colnames(data1)))
  res = merge(data1, data2[,data2_cols], by=by_col)
  return(res)
}
# Note that Specimen.Processing is intentionally the last added dataset
```

```r
# We merge by PIDs so all data before that are animal-level data
formnames = c("Acute.Test","Animal.Familiarization",
              "Animal.Key","Animal.Registration",
              "Specimen.Collection","Specimen.Processing")
merged_dmaqc_data = c()
for(currname in formnames){
  curr_data = csv_data[[which(grepl(currname,names(csv_data)))]]
  colnames(curr_data) = paste(currname,colnames(curr_data),sep=".")
  colnames(curr_data)[grepl(".pid$",colnames(curr_data))]="pid"
  colnames(curr_data)[grepl(".bid$",colnames(curr_data))]="bid"
  colnames(curr_data)[grepl(".vialid$",colnames(curr_data))]="vialid"
  colnames(curr_data)[grepl(".viallabel$",colnames(curr_data))]="viallabel"
  colnames(curr_data)[grepl(".labelid$",colnames(curr_data))]="labelid"
  colnames(curr_data) = tolower(colnames(curr_data))
  by_col = "pid"
  if(length(merged_dmaqc_data)==0){
    merged_dmaqc_data = curr_data
  }
  else{
    merged_dmaqc_data = merge_avoid_col_dup(merged_dmaqc_data,curr_data,by_col)
  }
}
print("Merged animal and biospecimen data tables, dim is:")
```

```
## [1] "Merged animal and biospecimen data tables, dim is:"
```

```r
print(dim(merged_dmaqc_data))
```

```
## [1] 2699  106
```

```r
# Now map DMAQC's label ids to vialids
# Sort to make the most up to date file the first in the order
mapping_files = sort(all_csvs[grepl("BICLabelData",all_csvs)],decreasing = T)
mapping_info = csv_data[[mapping_files[1]]]
colnames(mapping_info) = tolower(colnames(mapping_info))
# Not all samples in the specimen data are necessarily covered in the mapping
# file. The mapping file contains info only about samples that were shipped
# to CAS. As can be seen here:
table(is.element(merged_dmaqc_data$labelid,set=mapping_info$labelid))
```

```
##
## FALSE  TRUE
##  1061  1638
```

```r
# We therefore need to extract the intersection:
shared_labelids = intersect(merged_dmaqc_data$labelid,mapping_info$labelid)
merged_dmaqc_data = merged_dmaqc_data[
  is.element(merged_dmaqc_data$labelid,set = shared_labelids),]
mapping_info = mapping_info[
  is.element(mapping_info$labelid,set = shared_labelids),]
print("Merged animal and biospecimen data tables, new dim is:")
```

```
## [1] "Merged animal and biospecimen data tables, new dim is:"
```

```r
print(dim(merged_dmaqc_data))
```

```
## [1] 1638  106
```

```r
# We also have a many to one mapping from vial ids to labels, we
# merge the tables to avoid information loss
merged_dmaqc_data = merge_avoid_col_dup(merged_dmaqc_data,mapping_info,"labelid")
print("Merged animal and biospecimen data tables, after adding vialids, new dim is:")
```

```
## [1] "Merged animal and biospecimen data tables, after adding vialids, new dim is:"
```

```r
print(dim(merged_dmaqc_data))
```

```
## [1] 8616  109
```

```r
# Now put the dictionary in one file as well
merged_column_dictionary = c()
cols_to_take = c("Field.Name","Data.Type","Categorical.Values",
                 "Categorical.Definitions")
for(currname in formnames){
  tmp_dict_data = dict_data[[which(grepl(currname,names(dict_data)))]]
  tmp_dict_data = tmp_dict_data[,cols_to_take]
  tmp_dict_data[,1] = paste(currname,tmp_dict_data[,1],sep=".")
  tmp_dict_data[grepl(".pid$",tmp_dict_data[,1]),1]="pid"
  tmp_dict_data[grepl(".bid$",tmp_dict_data[,1]),1]="bid"
  tmp_dict_data[grepl(".labelid$",tmp_dict_data[,1]),1]="labelid"
  tmp_dict_data[grepl(".vialid$",tmp_dict_data[,1]),1]="vialid"
  tmp_dict_data[grepl(".viallabel$",tmp_dict_data[,1]),1]="viallabel"
  tmp_dict_data[,1] = tolower(tmp_dict_data[,1])
  tmp_dict_data = cbind(tmp_dict_data,rep(currname,nrow(tmp_dict_data)))
  merged_column_dictionary = rbind(merged_column_dictionary,tmp_dict_data)
}

# Add the calculated features
formnames = c(formnames,"Calculated.Variables")
currname = "Calculated.Variables"
# Data
curr_data = csv_data[[which(grepl(currname,names(csv_data)))]]
colnames(curr_data) = paste(currname,colnames(curr_data),sep=".")
colnames(curr_data)[grepl(".labelid$",colnames(curr_data))]="labelid"
colnames(curr_data) = tolower(colnames(curr_data))
merged_dmaqc_data = merge_avoid_col_dup(merged_dmaqc_data,curr_data,"labelid")
# Dictionary
tmp_dict_data = dict_data[[which(grepl(currname,names(dict_data)))]]
tmp_dict_data = tmp_dict_data[,cols_to_take]
tmp_dict_data[,1] = paste(currname,tmp_dict_data[,1],sep=".")
tmp_dict_data[grepl(".pid$",tmp_dict_data[,1]),1]="pid"
tmp_dict_data[grepl(".bid$",tmp_dict_data[,1]),1]="bid"
tmp_dict_data[grepl(".labelid$",tmp_dict_data[,1]),1]="labelid"
tmp_dict_data[,1] = tolower(tmp_dict_data[,1])
tmp_dict_data = cbind(tmp_dict_data,rep(currname,nrow(tmp_dict_data)))
merged_column_dictionary = rbind(merged_column_dictionary,tmp_dict_data)

# Final checks of the data
dim(merged_dmaqc_data)
```

```
## [1] 8616  116
```

```r
merged_column_dictionary = merged_column_dictionary[is.element(
  merged_column_dictionary[,1],set=colnames(merged_dmaqc_data)
  ),]
dim(merged_column_dictionary)
```

```
## [1] 155   5
```

```
merged_column_dictionary = unique(merged_column_dictionary)
```

## 5 Compare to the DMAQC computed scores

As requested by Ashley (email from June 7 2019), the following computed fields were added by DMQAQC:

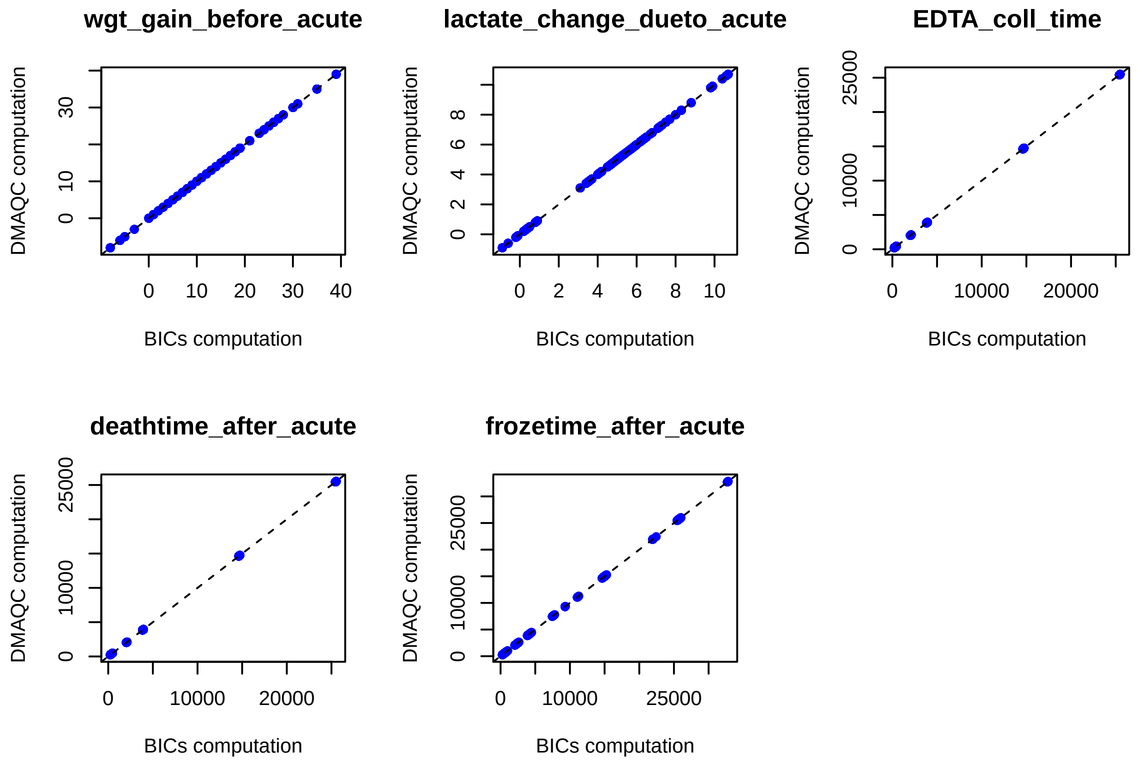
1. Weight gain before acute test: (Animal_Acute_Test.weight – Animal_Registration.weight)
2. Lactate changes due to acute exercise: (Aminal_Acute_Test.endblood - Aminal_Acute_Test.beginblood)
3. EDTA sample collection time: (Animal_Specimen_Collection.t_edtafill - Aminal_Acute_Test.t_complete)
4. Time of death after acute test: (Animal_Specimen_Collection.t_death - Aminal_Acute_Test.t_complete)
5. Sample frozen time after acute test: (Animal_Sample_Processing.t_freeze - Aminal_Acute_Test.t_complete)

Below, we show that our merged table and computations in R result in the same numbers.

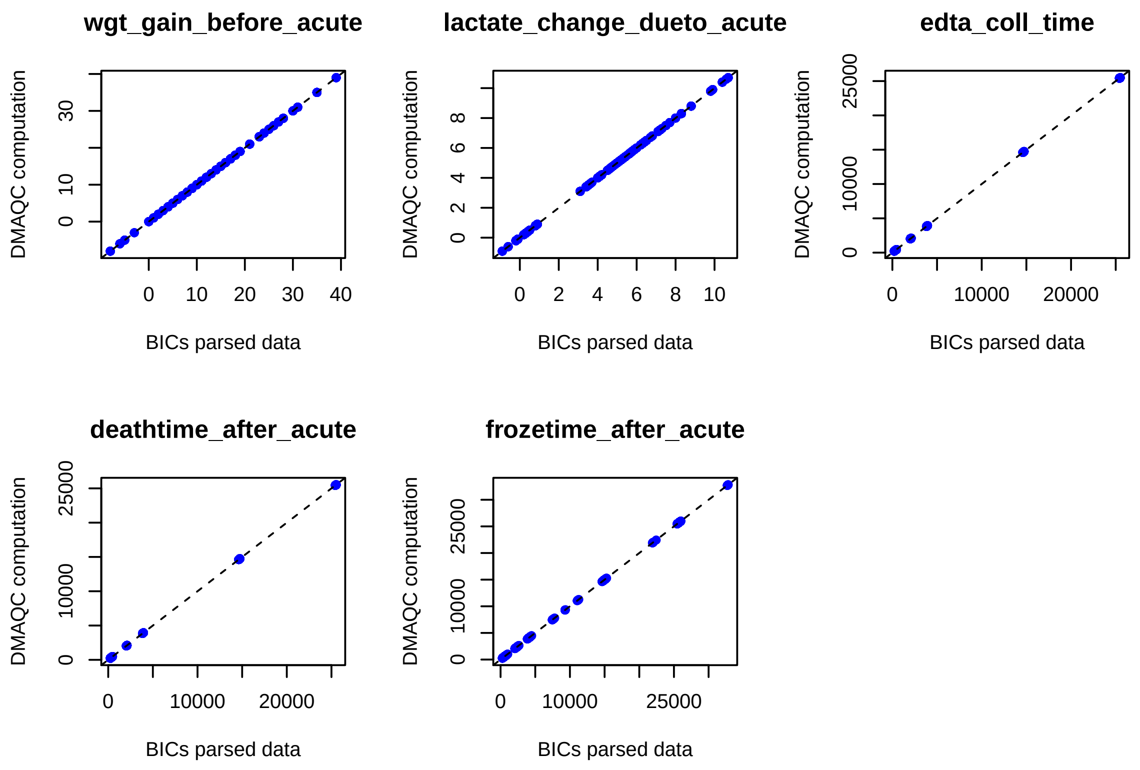
```
# Read the DMAQC calculated fields (do not use the prev ones from the merge
# for an extra QC)
calc_data_file = all_csvs[grepl("Calculated.Variables",all_csvs)]
calc_data = read.csv(calc_data_file)
rownames(calc_data) = calc_data$labelid
# Extract the relevant columns from our merged dataset
cols_for_analysis = c("labelid",
                      "acute.test.weight","animal.registration.weight",
                      "acute.test.endblood","acute.test.beginblood",
                      "specimen.collection.t_edtafill","acute.test.t_complete",
                      "specimen.collection.t_death","acute.test.t_complete",
                      "specimen.processing.t_freeze","acute.test.t_complete")
# table(is.element(cols_for_analysis,set=colnames(merged_dmaqc_data))) # sanity

# Go over each score and compare the two versions
par(mfrow=c(2,3))
for(j in seq(2,length(cols_for_analysis),by=2)){
  bic_version = unique(merged_dmaqc_data[,cols_for_analysis[c(1,j,j+1)]])
  rownames(bic_version) = bic_version[,1]
  dmaqc_version = calc_data[rownames(bic_version),c(3,3+j/2)]
  if(mode(bic_version[,2])=="character"){
    bic_version_score = as.difftime(bic_version[,2])-as.difftime(bic_version[,3])
    bic_version_score = as.numeric(bic_version_score)*60*60
  }
  else{
    bic_version_score = bic_version[,2]-bic_version[,3]
  }
  plot(bic_version_score,dmaqc_version[,2],pch=20,cex=1.2,col="blue",
       xlab="BICs computation",ylab="DMAQC computation",
       main = colnames(dmaqc_version)[2])
  abline(0,1,lty=2)
}

print("Now go over the columns, but this time take our version from the merged data")
```

```
## [1] "Now go over the columns, but this time take our version from the merged data"
```

```
colnames(calc_data) = tolower(colnames(calc_data))
par(mfrow=c(2,3))
```



```
for(feature_name in colnames(calc_data)[-c(1:3)]){
  bic_feature_name = paste("calculated.variables.",feature_name,sep="")
  bic_version = unique(merged_dmaqc_data[,c("labelid",bic_feature_name)])
  rownames(bic_version) = as.character(bic_version[,1])
  dmaqc_version = calc_data[rownames(bic_version),c("labelid",feature_name)]
  plot(bic_version[,2],dmaqc_version[,2],pch=20,cex=1.2,col="blue",
       xlab="BICs parsed data",ylab="DMAQC computation",
       main = colnames(dmaqc_version)[2])
  abline(0,1,lty=2)
}
```



## 6 Correlations with time points

Based on the analyses above we know that the distances are mostly correlated with the shock length and weight/sex. We now plot the achieved distances as a function of the shock data but colored by the time point of each animal in the exercise group.

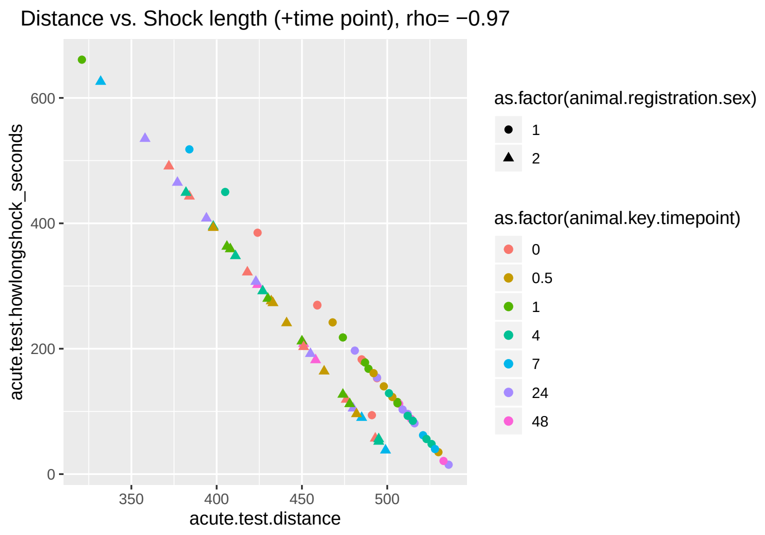
```
library(ggplot2)
parse_timepoint<-function(x){
  arrs = strsplit(x,split=" ")
  tps = sapply(arrs,function(x)x[3])
  tps = as.numeric(tps)
  tps[is.na(tps)]=0 # IPEs are marked as 0
  # tps[grepl("IPE",x)] = 0
  return(tps)
}

# colnames(merged_dmaqc_data)[grepl("sex",colnames(merged_dmaqc_data))]
merged_dmaqc_data$animal.key.timepoint = parse_timepoint(
  merged_dmaqc_data[,"animal.key.anirandgroup"])
```

```
## Warning in parse_timepoint(merged_dmaqc_data[, "animal.key.anirandgroup"]):
## NAs introduced by coercion
```

```
merged_dmaqc_data$animal.key.is_control = grepl("control",
      merged_dmaqc_data[,"animal.key.anirandgroup"],ignore.case = T)

# Reduce the data by label ids to avoid duplications
merged_dmaqc_data$acute.test.howlongshock_seconds = sapply(
  merged_dmaqc_data$acute.test.howlongshock,
  parse_shocktime)
inds = !is.na(merged_dmaqc_data$acute.test.howlongshock_seconds)
df = merged_dmaqc_data[inds,c("bid","acute.test.distance",
                              "acute.test.howlongshock_seconds",
                              "animal.key.timepoint",
                              "animal.registration.sex")]
df = unique(df)
print(paste("Number of bids in the reduced data.",nrow(df)))
```

```
## [1] "Number of bids in the reduced data. 72"
```

```
# Marginal correlation
rho = cor(df$acute.test.howlongshock_seconds,
          df$acute.test.distance)
rho = format(rho,digits = 3)

# A simple 2D plot
ggplot(df,
       aes(x=`acute.test.distance`, y=acute.test.howlongshock_seconds,
           shape=as.factor(animal.registration.sex), color=as.factor(animal.key.timepoint))) +
  geom_point(size=2) + ggtitle(paste("Distance vs. Shock length (+time point), rho=",rho)) +
  theme(plot.title = element_text(hjust = 0.5))
```



```
# Look at the linear regression, do we see a correlation between time
# and distance?
dist_lm2 = lm(acute.test.distance~acute.test.howlongshock_seconds+
                as.factor(animal.key.timepoint)+animal.registration.sex,data=df)
print("No significant linear association between the time points and distance:")
```

```
## [1] "No significant linear association between the time points and distance:"
```

```
# summary(dist_lm2)
```

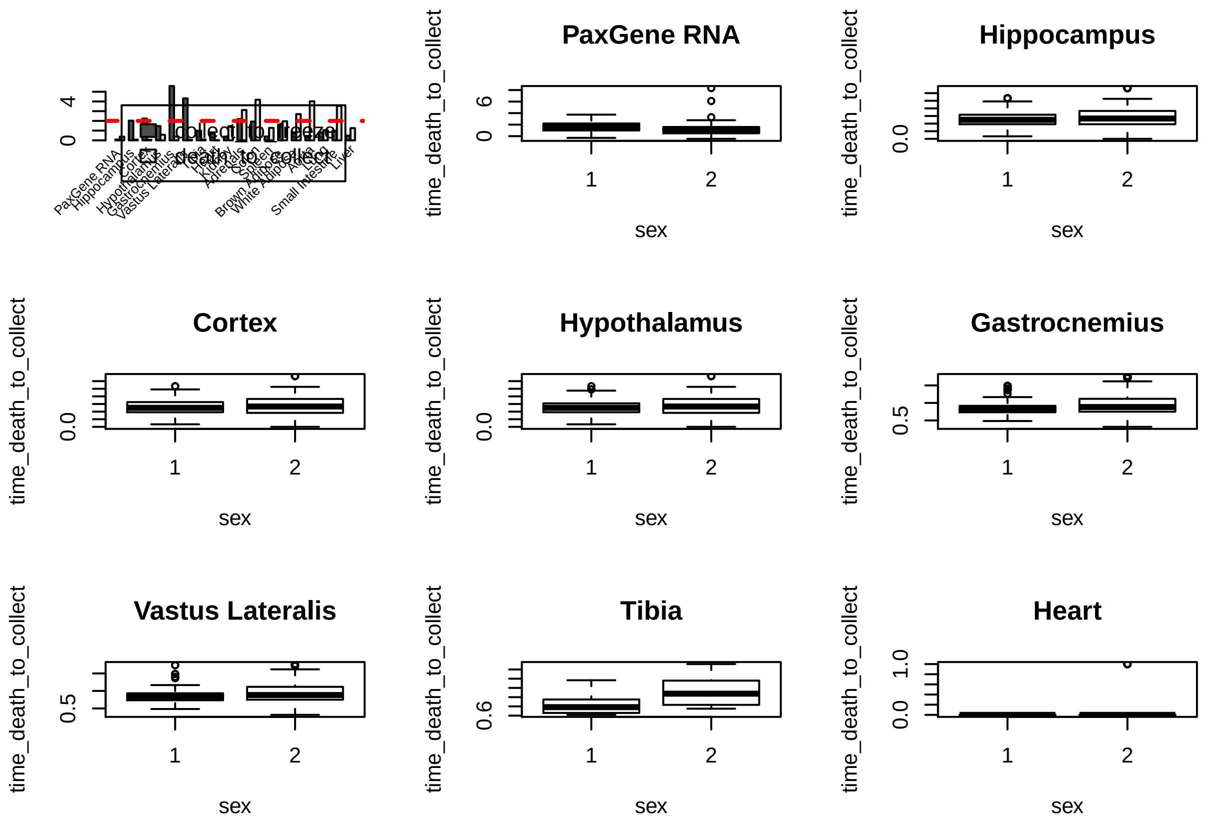
## 6.1 QC tests and time definitions

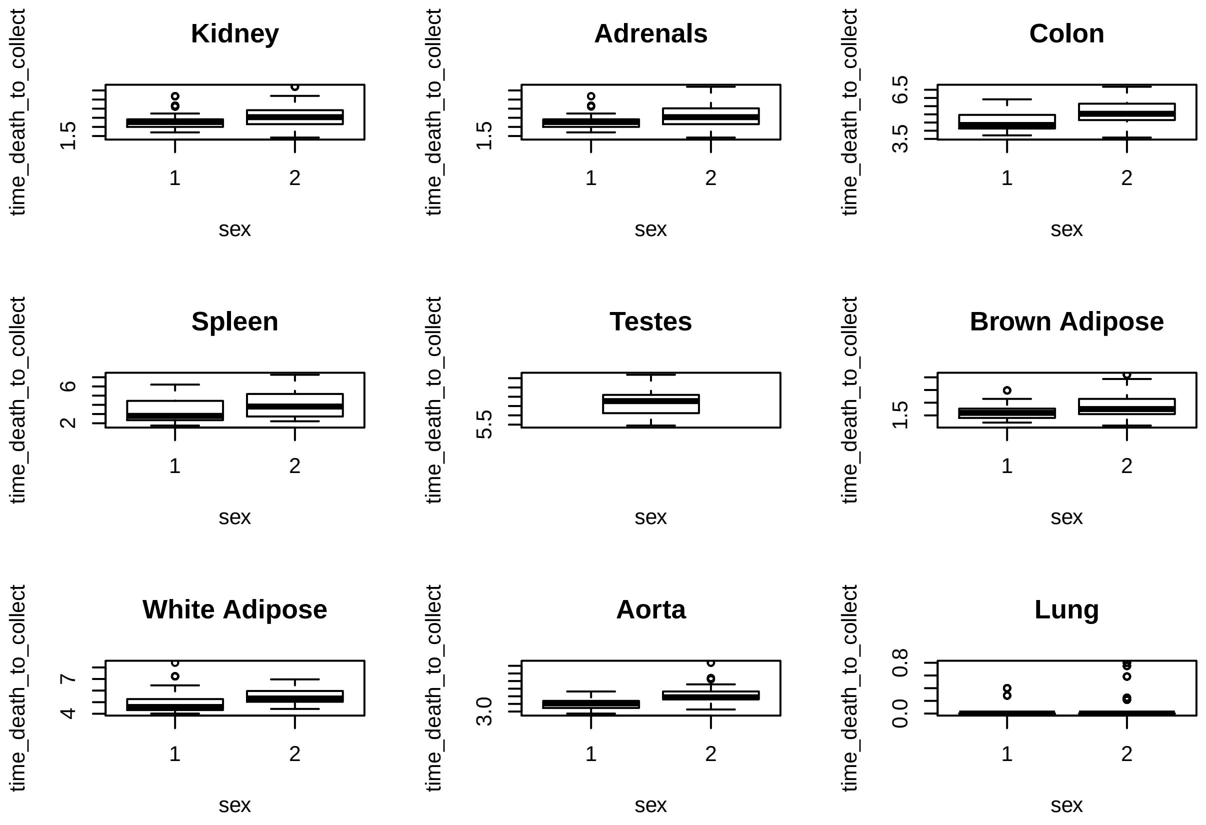
```
merged_dmaqc_data$tissue = merged_dmaqc_data$sampletypedescription
```

```r
# Define three time intervals:
# I1: complete to death
# I2: death to collection
# I3: collection to freeze
# These are defined in hours:
I1 = as.difftime(merged_dmaqc_data$specimen.collection.t_death) -
    as.difftime(merged_dmaqc_data$acute.test.t_complete)
I2 = as.difftime(merged_dmaqc_data$specimen.processing.t_collection) -
    as.difftime(merged_dmaqc_data$specimen.collection.t_death)
I3 = as.difftime(merged_dmaqc_data$specimen.processing.t_freeze) -
    as.difftime(merged_dmaqc_data$specimen.processing.t_collection)
I1 = as.numeric(I1);I2 = as.numeric(I2);I3 = as.numeric(I3)
daysdiff = merged_dmaqc_data$acute.test.days_visit -
    merged_dmaqc_data$acute.test.days_start
I1 = I1 + 24*daysdiff
# Change to minutes
I1 = I1*60
I2 = I2*60
I3 = I3*60
merged_dmaqc_data$calculated.variables.time_complete_to_death_min = I1
merged_dmaqc_data$calculated.variables.time_death_to_collect_min = I2
merged_dmaqc_data$calculated.variables.time_collect_to_freeze_min = I3

# look at time vs sex diffs in each tissue
tmpx = cbind(
  merged_dmaqc_data$labelid,
  merged_dmaqc_data$specimen.processing.sampletypedescription,
  merged_dmaqc_data$calculated.variables.time_complete_to_death_min,
  merged_dmaqc_data$calculated.variables.time_death_to_collect_min,
  merged_dmaqc_data$calculated.variables.time_collect_to_freeze_min,
  merged_dmaqc_data$animal.registration.sex,
  merged_dmaqc_data$animal.key.timepoint,
  merged_dmaqc_data$acute.test.weight

)
colnames(tmpx) = c(
  "labelid",
  "tissue",
  "time_complete_to_death",
  "time_death_to_collect",
  "time_collect_to_freeze",
  "sex",
  "timepoint",
  "weight"
)
tmpx = data.frame(tmpx)
for(j in 3:ncol(tmpx)){tmpx[[j]]=as.numeric(as.character(tmpx[[j]]))}
par(mfrow=c(3,3))
tissue_sex_pvals = c()
for(tissue in unique(tmpx$tissue)){
  df = tmpx[tmpx$tissue==tissue,]
  df = unique(df)
  print(dim(df))
  df$sex = df$sex-1

  if(any(is.na(df$time_collect_to_freeze))){next}
  if(length(unique(df$sex))<2){next}

  curr_lm = summary(glm(
    sex~time_collect_to_freeze + time_death_to_collect,data=df,
    family = "binomial"))
  pval = curr_lm$coefficients[2,4]
  beta = curr_lm$coefficients[2,2]

  tissue_sex_pvals = rbind(tissue_sex_pvals,
     c(curr_lm$coefficients[2,4],curr_lm$coefficients[3,4],nrow(df)))
  rownames(tissue_sex_pvals)[nrow(tissue_sex_pvals)] = tissue
  print(paste(tissue,beta,pval))
}
```

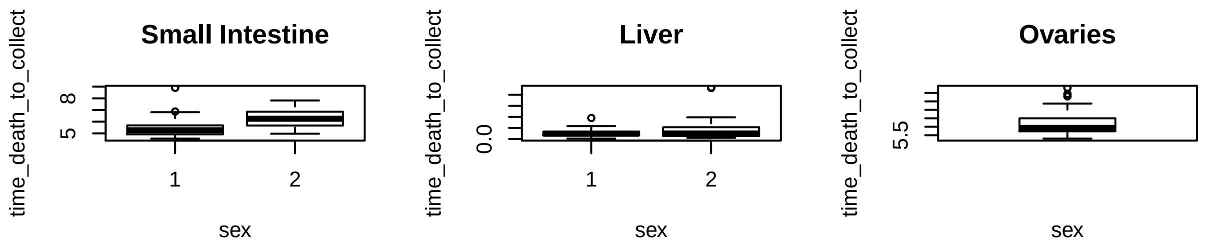
```
## [1] 78  8
## [1] "PaxGene RNA 0.790279848326075 0.817156550332186"
## [1] 156   8
## [1] 78  8
## [1] "Hippocampus 0.321660571954462 0.00944810189048159"
## [1] 78  8
## [1] "Cortex 0.318839241451852 0.00572878866462239"
## [1] 78  8
## [1] "Hypothalamus 0.378654221482666 0.0339555538040965"
## [1] 94  8
## [1] "Gastrocnemius 0.914589614953042 2.58817867587478e-06"
## [1] 78  8
## [1] "Vastus Lateralis 0.498770220233718 4.79418104086723e-05"
## [1] 16  8
## [1] "Tibia 1.67557140261054 0.103317278345256"
## [1] 92  8
## [1] "Heart 0.466171696137944 0.199060221467731"
## [1] 78  8
## [1] "Kidney 0.322552892032564 0.418355934840108"
## [1] 78  8
## [1] "Adrenals 0.425404017880244 0.00645760047457231"
## [1] 78  8
## [1] "Colon 0.54604928815029 0.0116198748828277"
## [1] 78  8
## [1] "Spleen 13.0548069891023 0.403120930656265"
## [1] 39  8
## [1] 78  8
## [1] "Brown Adipose 0.702405652020479 0.0232419073391203"
## [1] 94  8
## [1] "White Adipose 0.959424347900243 0.194632437895066"
## [1] 78  8
## [1] "Aorta 0.823315161101129 0.35965475499893"
## [1] 78  8
## [1] "Lung 0.656236769773995 0.152635830757761"
## [1] 78  8
## [1] "Small Intestine 0.857967259568141 0.104642959884655"
## [1] 94  8
## [1] "Liver 1.92900712951959 0.32997414426506"
## [1] 39  8
```

```r
tissue_log_ps = -log(tissue_sex_pvals[,1:2],base=10)
colnames(tissue_log_ps) = c(
  "collect_to_freeze",
  "death_to_collect"
)
plt = barplot(t(tissue_log_ps),beside = T,xaxt="n",legend=T)
text(colMeans(plt), par("usr")[3], labels = rownames(tissue_log_ps),
     srt = 45, adj = c(1.1,1.1), xpd = T, cex=0.6)
abline(h = 2,lwd=2,col="red",lty=2)

for(tissue in unique(tmpx$tissue)){
  df = tmpx
  df = df[grepl(tissue,df$tissue,ignore.case = T),]
  if(all(is.na(df$time_collect_to_freeze))){next}
  boxplot(time_death_to_collect~sex,data=df,main=tissue)
}
```





```
# # reproduce the results directly from the specimen data
# spec_data = read.csv("~/Desktop/MoTrPAC/data/pass_1a/dmaqc_pheno/QC_EXPORT_TXFRO.01_1A.6M_DS_MoTrPAC..
# spec_collection = read.csv("~/Desktop/MoTrPAC/data/pass_1a/dmaqc_pheno/QC_EXPORT_TXFRO.01_1A.6M_DS_Mo
# animal_data = read.csv("~/Desktop/MoTrPAC/data/pass_1a/dmaqc_pheno/QC_EXPORT_TXFRO.01_1A.6M_DS_MoTrPA
# acute_data = read.csv("~/Desktop/MoTrPAC/data/pass_1a/dmaqc_pheno/QC_EXPORT_TXFRO.01_1A.6M_DS_MoTrPAC
# animal_data2 = read.csv("~/Desktop/MoTrPAC/data/pass_1a/dmaqc_pheno/QC_EXPORT_TXFRO.01_1A.6M_DS_MoTrP
# # Limit to gastroc data
# spec_data = spec_data[grepl("gastroc",spec_data$sampletypedescription,ignore.case = T),]
# # Make all ids PIDs
# rownames(spec_data) = as.character(spec_data$pid)
# rownames(animal_data) = as.character(animal_data$pid)
# rownames(spec_collection) = as.character(spec_collection$pid)
# rownames(acute_data) = as.character(acute_data$pid)
# rownames(animal_data2) = animal_data2$pid
# # Intersection
# inds = intersect(rownames(spec_data),rownames(animal_data))
# spec_data = spec_data[inds,]
# animal_data = animal_data[inds,]
# acute_data = acute_data[inds,]
# spec_collection = spec_collection[inds,]
# animal_data2 = animal_data2[inds,]
#
# # three time intervals:
# # I1: complete to death
# # I2: death to collection
# # I3: collection to freeze
# I1 = as.difftime(spec_collection$t_death) - as.difftime(acute_data$t_complete)
# I2 = as.difftime(spec_data$t_collection) - as.difftime(spec_collection$t_death)
# I3 = as.difftime(spec_data$t_freeze)-as.difftime(spec_data$t_collection)
# I1 = as.numeric(I1)
# I2 = as.numeric(I2)
# I3 = as.numeric(I3)
# sex = animal_data$sex
# daysdiff = acute_data$days_visit - acute_data$days_start
# I1 = I1 + 24*daysdiff
# # Change to minutes
# I1 = I1*60
# I2 = I2*60
# I3 = I3*60
#
# group = animal_data2$ANIRandGroup
# group = gsub("Control","C",group)
# group = gsub("Exercise","E",group)
#
# # Sex vs times
# par(mfrow=c(1,3))
# boxplot(log(I1)~sex,main = "I1: death-complete",ylab = "minutes(log)",xlab = "sex")
# boxplot(I2~sex,main = "I2: collection-death",xlab = "sex",ylab="minutes")
# boxplot(I3~sex,main = "I3: freeze-collection",xlab = "sex",ylab="minutes")
# # Study group vs times
# par(mfrow=c(3,1))
# boxplot(log(I1)~group,main = "I1: death-complete",las=2,xlab="",ylab="minutes(log)")
# boxplot(I2~group,main = "I2: collection-death",las=2,xlab="",ylab="minutes")
# boxplot(I3~group,main = "I3: freeze-collection",las=2,xlab="",ylab="minutes")
#
# kruskal.test(I1,group)$p.value
# kruskal.test(I2,group)$p.value
# kruskal.test(I3,group)$p.value
```



## 7 Save the merged datasets in the cloud

```
# To see the bucket list
# gsutil ls -p motrpac-portal-dev
currdate = Sys.Date()
txtname = paste("merged_dmaqc_data",currdate,".txt",sep="")
rdataname = paste("merged_dmaqc_data",currdate,".RData",sep="")
dictfname = paste("merged_column_dictionary",currdate,".txt",sep="")
write.table(merged_dmaqc_data,file=txtname,
            quote = F,sep="\t",row.names = F)
save(merged_dmaqc_data,file=rdataname)
write.table(merged_column_dictionary,file=dictfname,
            quote = F,sep="\t",row.names = F)
system(paste("~/google-cloud-sdk/bin/gsutil cp", txtname,
             "gs://bic_data_analysis/pass1a/pheno_dmaqc/"))
system(paste("~/google-cloud-sdk/bin/gsutil cp", rdataname,
             "gs://bic_data_analysis/pass1a/pheno_dmaqc/"))
system(paste("~/google-cloud-sdk/bin/gsutil cp", dictfname,
             "gs://bic_data_analysis/pass1a/pheno_dmaqc/"))
system(paste("~/google-cloud-sdk/bin/gsutil cp",
             "~/Desktop/repos/motrpac/animal_data/README.txt",
             "gs://bic_data_analysis/pass1a/pheno_dmaqc/"))

system("rm merged_dmaqc_data.txt")
system("rm merged_dmaqc_data.RData")
system("rm merged_column_dictionary.txt")
```