Preliminaries

The initial set-up for filtering was similar to the Cosinor code appendix. The `CircadianTools` package was loaded. The count data (Laurasmappings) was read in and the CT18.4 column was removed due to this column showing no activity. Any genes which show zero activity for all readings were removed: as is customary.

```
library(CircadianTools)
Laurasmappings <- read.csv("~/MEGA/Uni/Masters/Diss/Stats/Raw_Data/Laurasmappings.csv",
                           stringsAsFactors = FALSE) # Read in count data
Laurasmappings$CT18.4 <- NULL # Remove column of zeroes
Laurasmappings <- GeneClean(Laurasmappings) # Remove genes which show no activity
```

For this code appendix, an additional preliminary steps was taken: `circadian.csv`, which contains the names of genes found to be circadian via BLAST, was read in.

```
circadian <- read.csv("~/MEGA/Uni/Masters/Diss/Stats/circadian.csv", sep="",
                      stringsAsFactors=FALSE)
```

ANOVA Filtering

Threshold of 5%

Firstly, the genes were filtered using ANOVA with a signficance level of 5%.

```
system.time(a.filter <- AnovaFilter(dataset = Laurasmappings, threshold = 0.05))
```

```
##    user  system elapsed
##  35.170   3.937  81.870
```

```
anova.no <- nrow(a.filter) # Number of genes in the reduced dataset
circadian.subset <- GeneSub(circadian, a.filter)
circadian.no <- nrow(circadian.subset)



cat(paste("There are", anova.no,
          "genes in the dataset filtered via ANOVA with a threshold of 5%.\n",
          circadian.no, "of the", nrow(circadian),
         "genes found to be circadian via BLAST are in the reduced dataset.\n
         These genes are: \n")
    )
```

```
## There are 11186 genes in the dataset filtered via ANOVA with a threshold of 5%.
##  5 of the 18 genes found to be circadian via BLAST are in the reduced dataset.
##
##          These genes are:
```

```
circadian.subset$sample
```

```
## [1] "comp97405_c0_seq1"  "comp99101_c0_seq3"  "comp102279_c0_seq7"
## [4] "comp102609_c0_seq3" "comp939723_c0_seq1"
```

Threshold of 2.5%

As 11,186 genes is still too many genes for some computational methods, a threshold of 2.5% was considered.

```
system.time(a.filter <- AnovaFilter(dataset = Laurasmappings,
                                    threshold = 0.025)
            )
```

```
##    user  system elapsed
##  34.934   3.679  81.315
```

```
anova.no <- nrow(a.filter) # Number of genes in the reduced dataset
circadian.subset <- GeneSub(circadian, a.filter)
circadian.no <- nrow(circadian.subset)



cat(paste("There are", anova.no,
          "genes in the dataset filtered via ANOVA with a threshold of 2.5%.\n",
          circadian.no, "of the", nrow(circadian),
         "genes found to be circadian via BLAST are in the reduced dataset.\n
         These genes are: \n")
    )
```

```
## There are 7124 genes in the dataset filtered via ANOVA with a threshold of 2.5%.
##  3 of the 18 genes found to be circadian via BLAST are in the reduced dataset.
##
##          These genes are:
```

```
circadian.subset$sample
```

```
## [1] "comp97405_c0_seq1"  "comp99101_c0_seq3"  "comp102279_c0_seq7"
```

The reduced dataset was saved as a csv file for use in later chapters.

```
write.csv(a.filter,"a_filter.csv", row.names = FALSE )
```

T-Test Filtering

Filtering via t-tests, as presented in the main document was then used to filter the count data.

```
system.time(t.filter <- TFilter(Laurasmappings, maxdifference = 1,
                                minchanges = 2, psignificance = 0.05)
            )
```

```
##    user  system elapsed
##  34.960   4.088  82.536
```

```
circadian.subset <- GeneSub(circadian, t.filter)
circadian.no <- nrow(circadian.subset)

t.filter.no <- nrow(t.filter) # Number of genes in the reduced dataset
circadian.no <- nrow(GeneSub(circadian, t.filter))

cat(paste("There are",t.filter.no,
          "genes in the dataset filtered via t-tests. \n",
          circadian.no, "of the", nrow(circadian),
         "genes found to be circadian via BLAST are in the reduced dataset.\n
         These genes are: \n"))
```

```
## There are 6294 genes in the dataset filtered via t-tests.
##  2 of the 18 genes found to be circadian via BLAST are in the reduced dataset.
```

```
##
##          These genes are:
```

```
circadian.subset$sample
```

```
## [1] "comp100937_c0_seq1" "comp939723_c0_seq1"
```

The reduced dataset was saved as a csv file for use in later chapters.

```
write.csv(t.filter,"t_filter.csv", row.names = FALSE )
```

T-Test Experimentation

As filtering by using the t-tests methood involves three important paramters, these parameters were each varied whilst fixing the other two in order to see the effect of varying the paramters.

p

```
p.values <- c(0.05, 0.025, 0.01)

p.value.results <- data.frame()
 for (i in p.values){
  filtered <- TFilter(Laurasmappings, maxdifference = 1, minchanges = 2,
                      psignificance = i)

  p.value.results <- rbind (p.value.results,
                            data.frame(p = i, genecount = nrow(filtered)
                                       )
                            )
}

print(p.value.results)
```

```
##       p genecount
## 1 0.050      6294
## 2 0.025      2643
## 3 0.010       805
```

Minimum Significant Changes

```
changes <- 1 : 5

sig.change.results <- data.frame()

for (i in changes){
  filtered <- TFilter(Laurasmappings, maxdifference = 1, minchanges = i,
                      psignificance = 0.05)
  sig.change.results <- rbind(sig.change.results,
                              data.frame(min.changes = i,
                                         genecount = nrow(filtered)
                                         )
                              )
}

print(sig.change.results)
```

```
##   min.changes genecount
```

```
## 1           1     22614
## 2           2      6294
## 3           3      1775
## 4           4       259
## 5           5        55
```

Maximum Difference Between Signficant Changes

```
max.diff <- 0 : 5

max.diff.results <- data.frame()

for (i in max.diff){
  filtered <- TFilter(Laurasmappings, maxdifference = i, minchanges = 2,
                      psignificance = 0.05)
  max.diff.results <- rbind(max.diff.results,
                            data.frame(max.diff = i,
                                       genecount = nrow(filtered)
                                       )
                            )
}


print(max.diff.results)
```

```
##   max.diff genecount
## 1        0      4724
## 2        1      6294
## 3        2      7297
## 4        3      7312
## 5        4      7312
## 6        5      7312
```

Comparing T-Test and ANOVA Filtering

The number of genes found in both the ANOVA filtered and t-test filtered datasets were then found.

```
shared.no <- nrow(GeneSub(a.filter, t.filter))
cat(paste( "There are", shared.no,
           "genes which can be found in both reduced datasets.\n"))
```

```
## There are 3097 genes which can be found in both reduced datasets.
```