Workbook 3 - Transcriptomics

MAM5220 - Statistical Techniques for Computational Biology

Note: see code/images: https://github.com/sap218/R/tree/master/mam5220/w3

Question 1

Crab_transcriptomics.csv

a)

i)

```
> length(which(crab$Condition_A == 0))
[1] 183
```

ii)

```
> length(which(crab$Condition_B == 0))
[1] 1920
```

b)

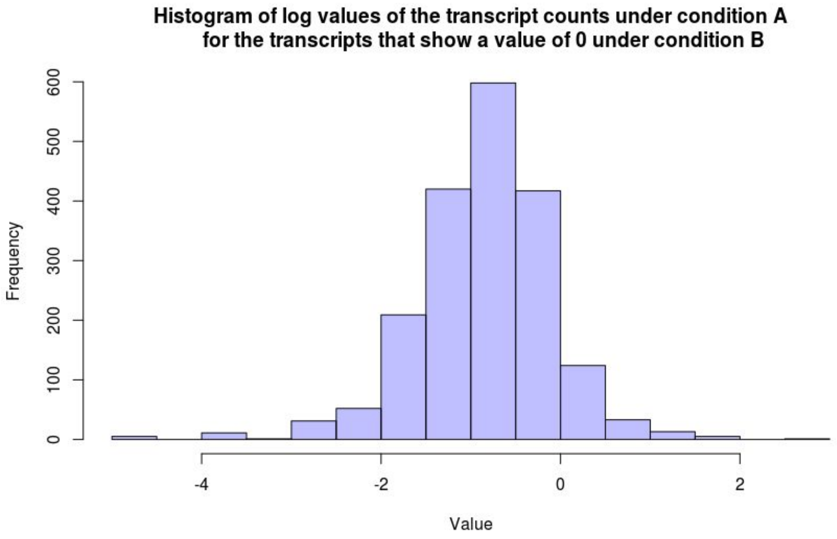


Figure: a histogram of the $\log_2$ values of the transcript counts under condition A for the transcripts that show a value of 0 under condition B.

c)

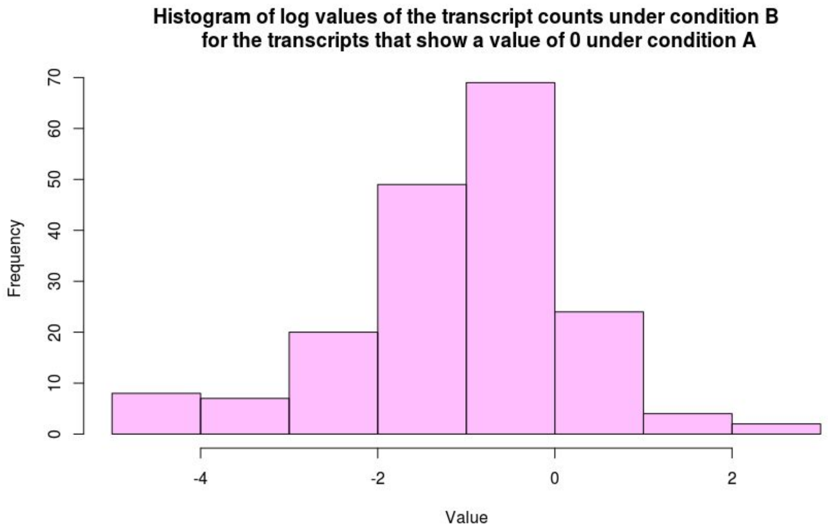


Figure: a histogram of the $\log_2$ values of the transcript counts under condition B for the transcripts that show a value of 0 under condition A.

d)

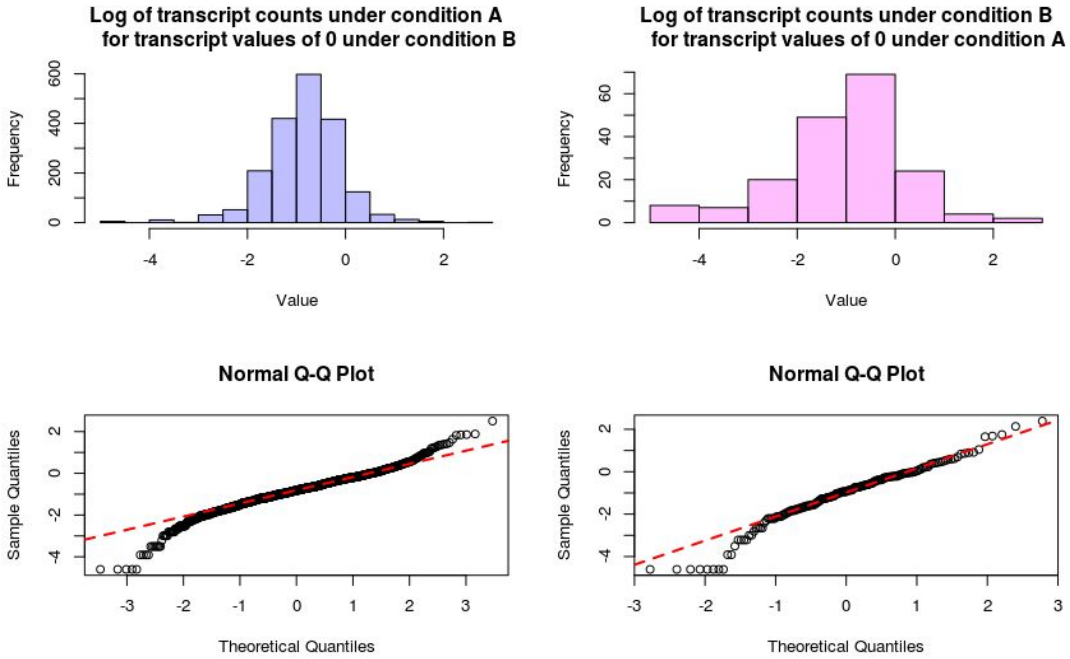


Figure: the histogram of transcript counts under condition A for values of 0 under condition B is a bell-shaped curve, concluding it is Normal, this can be backed up by the Q-Q plot. On the other hand, the histogram of transcript counts under condition B for values of 0 under condition A is less ‘normal’ due to a less curved shape: there being a smaller data set (more 0 values in condition A rather than B).

e)

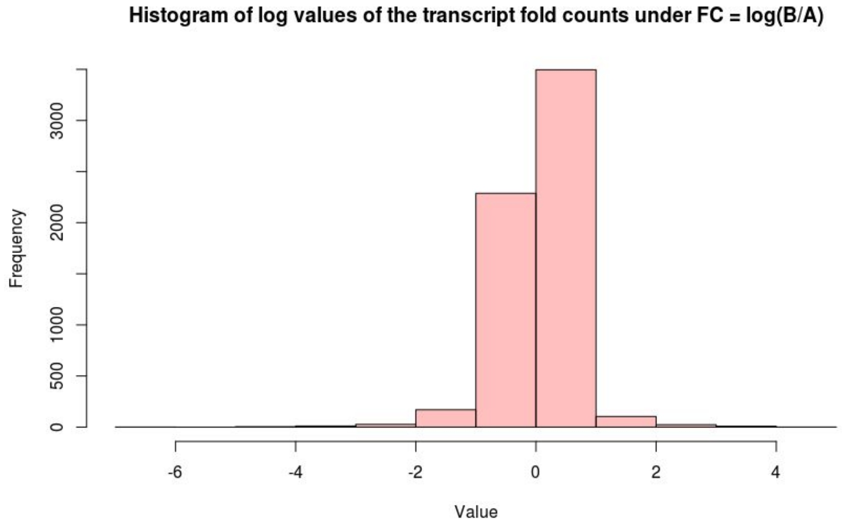


Figure: a histogram of the fold changes in a subsetted data frame of no zero values.

f)

The five transcripts that show the largest positive fold-change between condition A and condition B, plus the five transcripts that show the largest negative fold-change.

Transcript ID	Fold-change
4786	61
2918	39.6
2919	39
5071	30.6666666667
306	28
4667	0.0012936611
5548	0.0085470085
4674	0.0119047619
1403	0.0149812734
5661	0.0157480315

Question 2

LR.transcriptomics.csv

a)

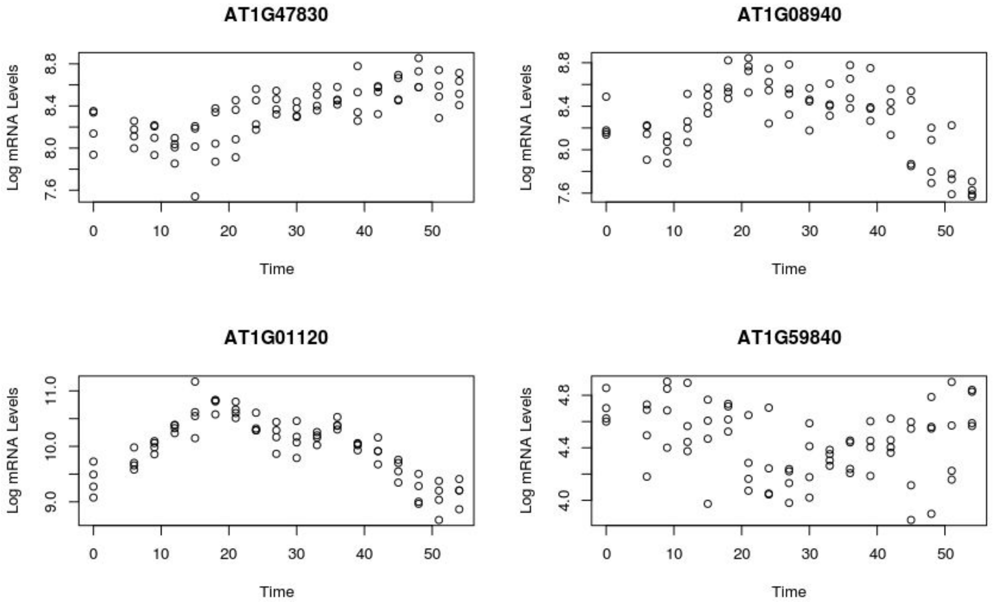


Figure: $\log_2$ mRNA levels by time of four genes; AT1G47830, AT1G08940, AT1G01120 and AT1G59840. Observing the first three plots, AT1G47830, AT1G08940, and AT1G01120, they profile a distinctive response with time, hypothesising that these will have a significant p-value (small) due to the defined pattern. On the other hand, the fourth plot, AT1G59840, shows a little change over time, predicting the p-value being large: least significant.

b)

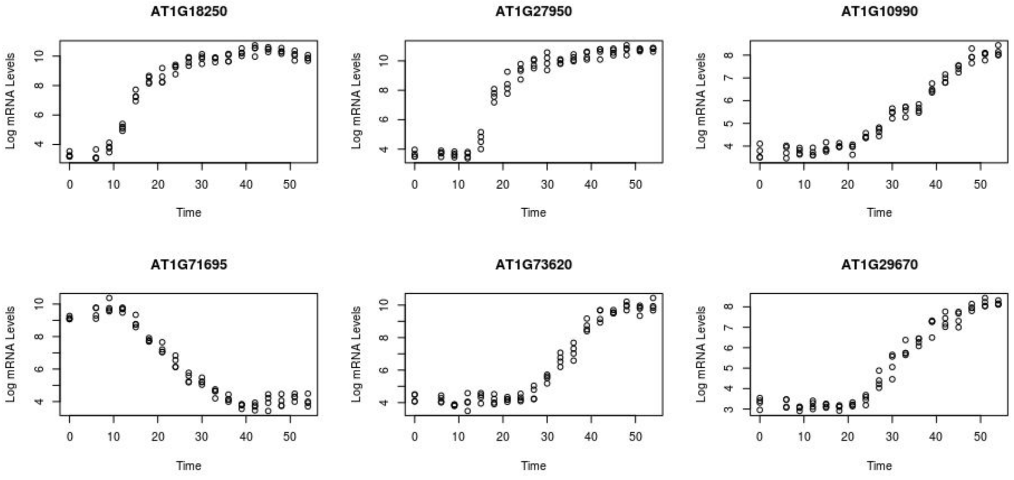


Figure: time-profile plots of the 6 genes with the most significant p-values.

c)

Proportions of genes with corrected p-values, using the Bonferroni correction, less than or equal to: (0.05, 0.025, 0.01, 0.005, 0.001).

```
> (sum(p.v.adj <= 0.05))/5000
[1] 0.5566
> (sum(p.v.adj <= 0.05))/5000
[1] 0.5566
> (sum(p.v.adj <= 0.025))/5000
[1] 0.5456
> (sum(p.v.adj <= 0.01))/5000
[1] 0.5278
> (sum(p.v.adj <= 0.005))/5000
[1] 0.515
> (sum(p.v.adj <= 0.001))/5000
[1] 0.4914
```

d)

i)

Proportions of null hypothesis rejected with the false discovery rate threshold is set at each of the values in the vector (1%, 5%, 10%, 20%, 25%).

```
> (sum(p.v.fdr <= 0.01))/5000
[1] 0.6834
> (sum(p.v.fdr <= 0.05))/5000
[1] 0.7386
> (sum(p.v.fdr <= 0.1))/5000
[1] 0.7696
> (sum(p.v.fdr <= 0.2))/5000
[1] 0.8112
> (sum(p.v.fdr <= 0.25))/5000
[1] 0.8284
```

ii)

Values at which the false discovery rate threshold should be set to ensure that 20%, 30% and 40% of the genes respectively are rejected in the analysis.

```
> sum(p.v.fdr <= 0.6)/5000
[1] 0.9204
> sum(p.v.fdr <= 0.7)/5000
[1] 0.9416
> sum(p.v.fdr <= 0.8)/5000
[1] 0.9624
```

e)

i)

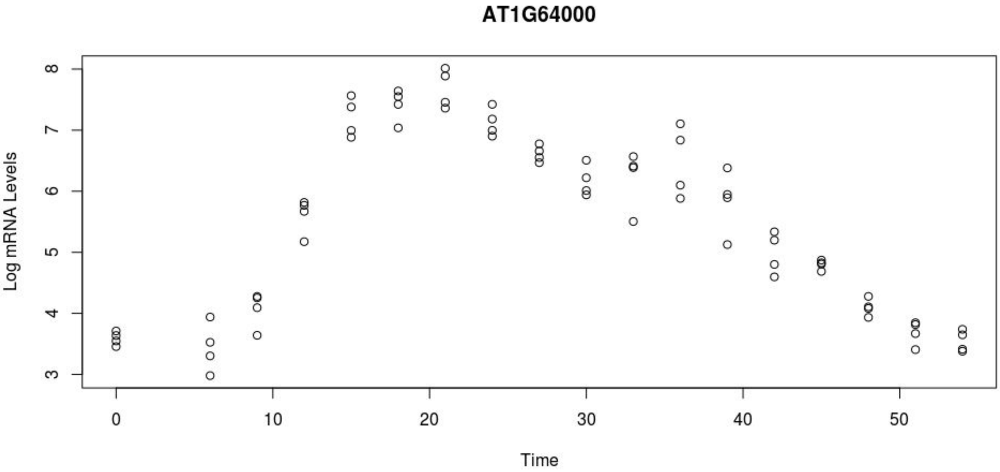
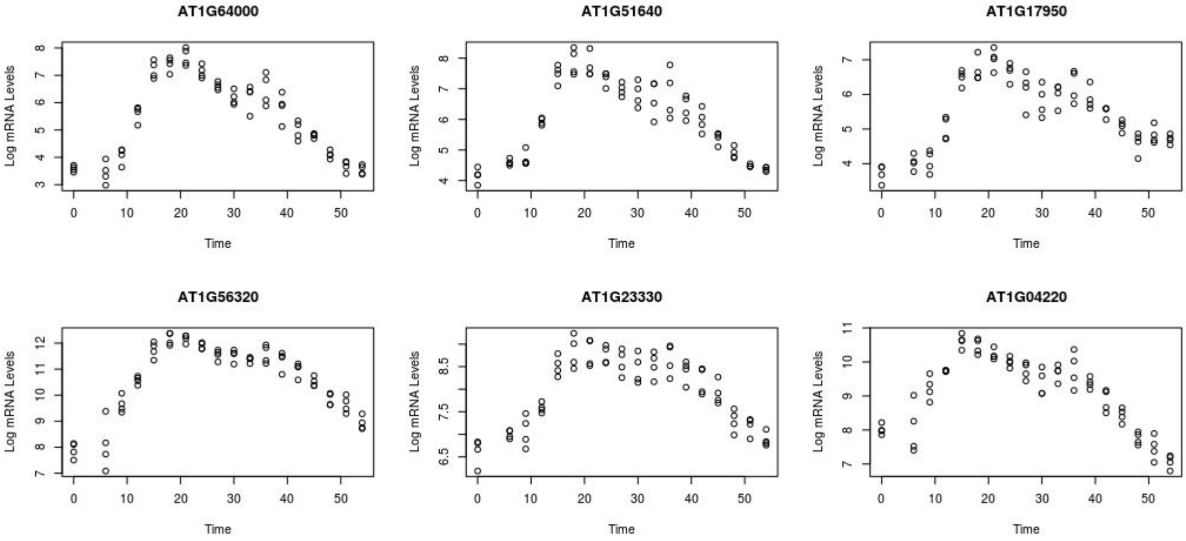

AT1G64000

Figure: plot of $\log_2$ mRNA levels against time for gene, AT1G64000, member of the WRKY transcription family of genes. It’s profile is a distinctive response with time, hypothesising that these will have a significant p-value (small) due to the defined pattern.

ii)


AT1G64000

AT1G51640

AT1G17950

AT1G56320

AT1G23330

AT1G04220

Figure: correlations between AT1G6400 and the top five highest correlation results. We can identify that the top left box, AT1G6400, gene is the one we are studying. The other five are very similar in profile.

iii)

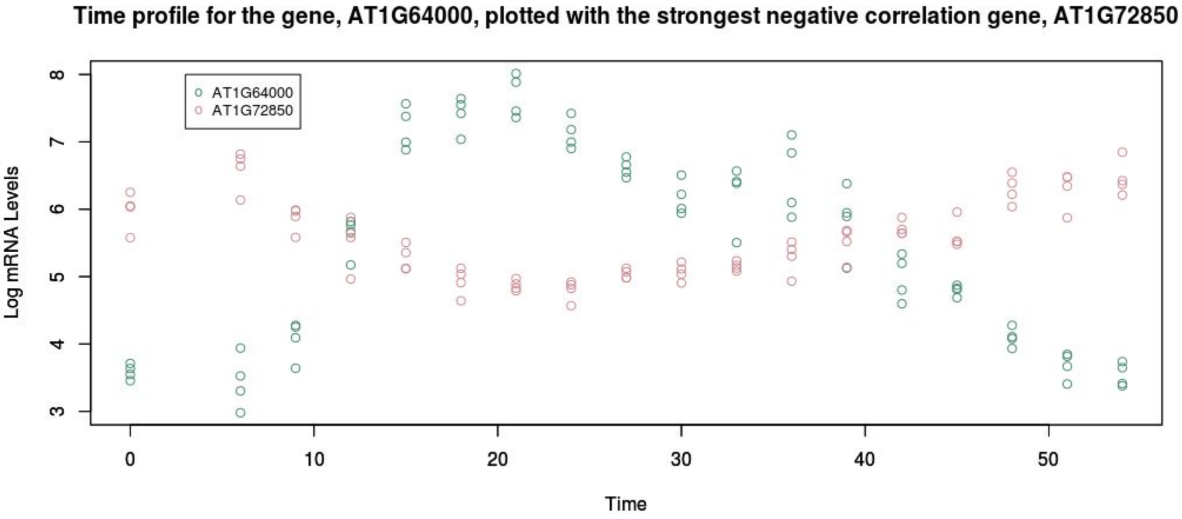


Figure: plot of the time profile for the gene, AT1G64000, against the strongest negative correlation gene, AT1G72850.

Question 3

LR.transcriptomics.csv

a)

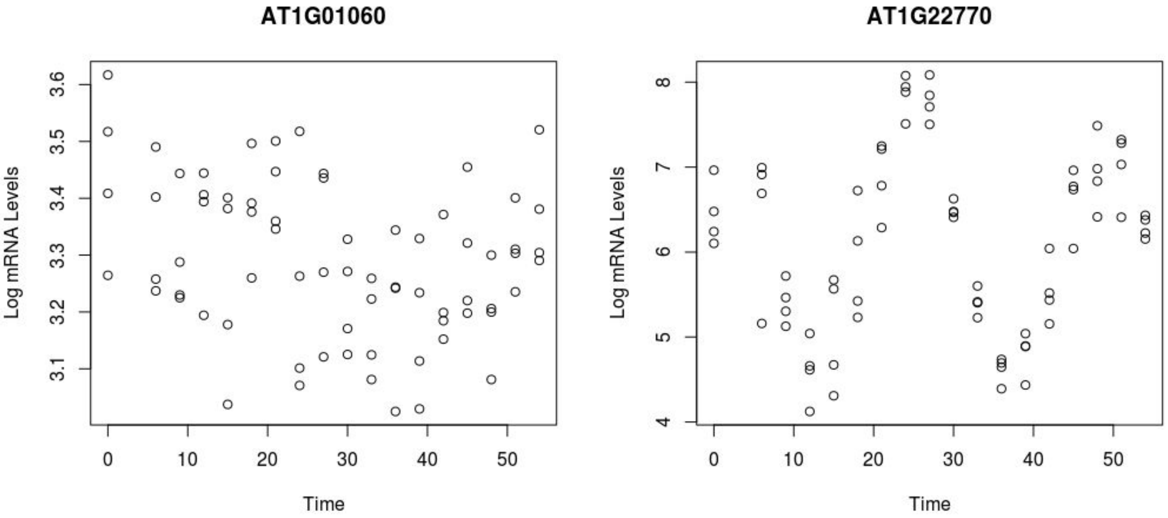


Figure: time profile plots of genes AT1G01060 (LHY) and AT1G22770 (GI).

b)

LHY: AT1G01060

Positive

```
> sorting.common.g1[["ix"]][2:6] # index
[1] 2210  174 1891 4368   38
> LR$X1[(sorting.common.g1[["ix"]][2:6])] # Gene
[1] "AT1G28630" "AT1G03070" "AT1G23730" "AT1G67550" "AT1G01430"
> round((sorting.common.g1[["x"]][2:6]),2) # Value
[1] 0.52 0.47 0.47 0.47 0.46
```

Negative

```
> sorting.common.g1.neg[["ix"]][2:6] # index
[1] 2196 3213 1143 2053 3135
> LR$X1[(sorting.common.g1.neg[["ix"]][2:6])] # Gene
[1] "AT1G28470" "AT1G50700" "AT1G14680" "AT1G26730" "AT1G49780"
> round((sorting.common.g1.neg[["x"]][2:6]),2) # Value
[1] -0.44 -0.43 -0.43 -0.43 -0.42
```

GI: AT1G22770

Positive

```
> sorting.common.g2[["ix"]][2:6] # index
[1] 4572  428 2881 2593 3130
> LR$X1[(sorting.common.g2[["ix"]][2:6])] # Gene
[1] "AT1G69830" "AT1G06040" "AT1G44050" "AT1G33840" "AT1G49720"
> round((sorting.common.g2[["x"]][2:6]),2) # Value
[1] 0.91 0.63 0.62 0.58 0.58
```

Negative

```
> sorting.common.g2.neg[["ix"]][2:6] # index
[1] 4468 1787 1151  947 4273
> LR$X1[(sorting.common.g2.neg[["ix"]][2:6])] # Gene
[1] "AT1G68620" "AT1G22550" "AT1G14730" "AT1G12240" "AT1G66330"
> round((sorting.common.g2.neg[["x"]][2:6]),2) # Value
[1] -0.69 -0.67 -0.64 -0.61 -0.61
```

c)

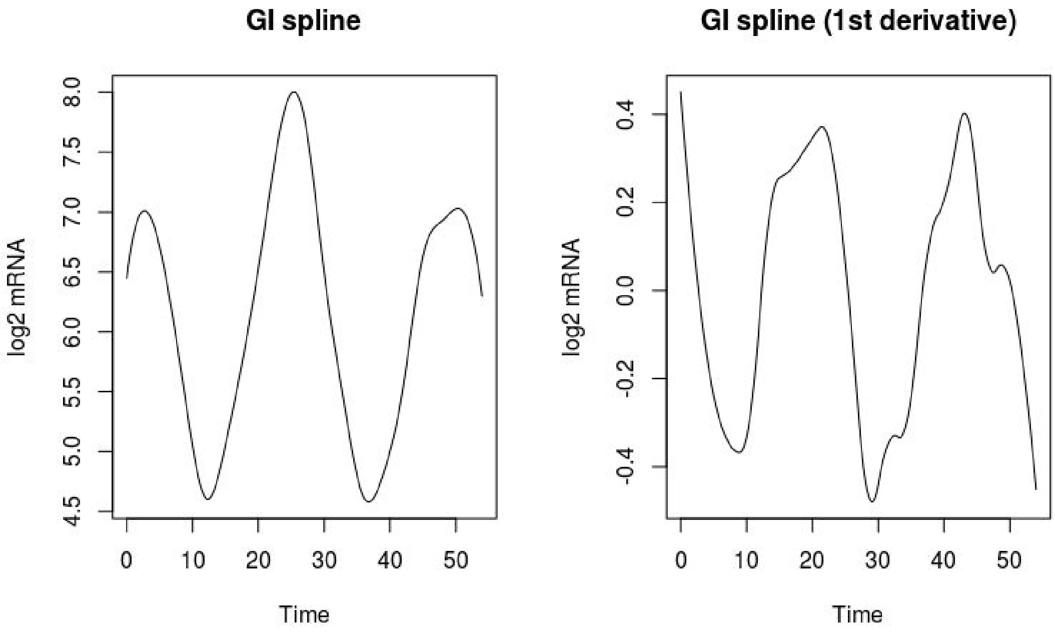


Figure: the spline fit, plus first derivative, of the gene, GI (AT1G22770).

d)

```
> LR[which(means < 13 & means > 11 & s.d <= 3),]$X1

[1] "AT1G02780" "AT1G09200" "AT1G09470" "AT1G10480" "AT1G11130"
"AT1G13420" "AT1G14180" "AT1G15570" "AT1G17140" "AT1G18250"
[11] "AT1G18370" "AT1G33320" "AT1G44900" "AT1G53070" "AT1G54340"
"AT1G64060" "AT1G64650" "AT1G67510" "AT1G70210" "AT1G74520"
```

e)

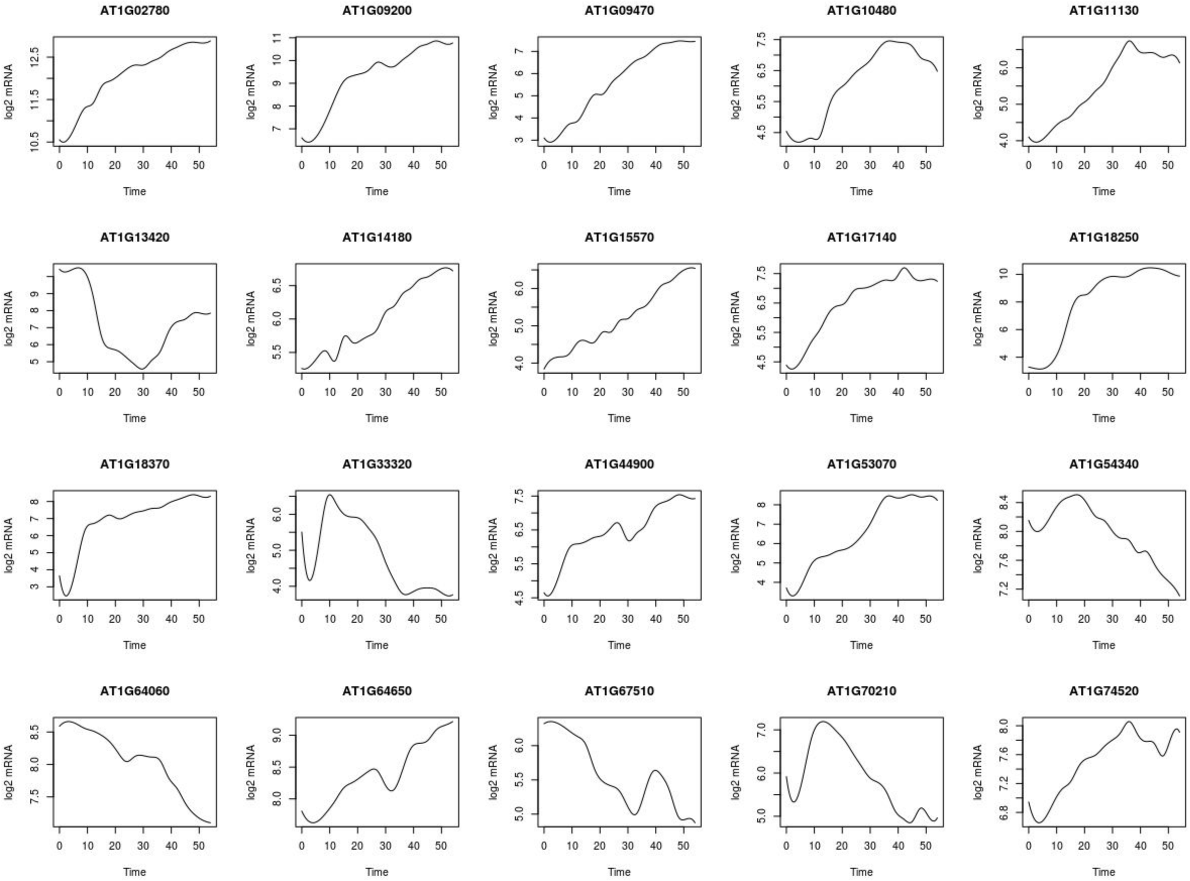


Figure: plots of the transcriptomics profiles, with fitted spline of the genes.