Expression of six clustered genes is proposed to be up-regulated by transcription of an adjacent lncRNA, *hotsix*. Below are relative quantity measures for each of the six gene from explicitly paired samples of animals that have or do not have *hotsix*.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| animal | pair | *hotsix* | *lukewarm* | *tepid* | *mild1* | *athermal* | *bathwater* | *coldshower* |
| 3011 | 1 | - | 0.4202 | 1.0750 | 0.3974 | 0.9355 | 0.4695 | 0.1688 |
| 3022 | 2 | - | 0.4718 | 1.0288 | 1.2747 | 0.6859 | 0.5616 | 0.4862 |
| 3026 | 3 | - | 0.5351 | 0.9428 | 1.0208 | 0.5423 | 0.1972 | 0.2022 |
| 3045 | 4 | - | 0.4955 | 0.7869 | 0.9073 | 0.6276 | 0.1854 | 0.2918 |
| 3052 | 5 | - | 0.4299 | 0.9417 | 1.3223 | 0.8431 | 0.3982 | 0.3084 |
| 3059 | 6 | - | 0.5609 | 1.2320 | 0.4473 | 1.5524 | 0.2859 | 0.1451 |
| 3066 | 7 | - | 1.4554 | 1.0419 | 0.6562 | 1.4554 | 0.4248 | 0.1613 |
| 3080 | 8 | - | 0.8959 | 0.7659 | 0.4807 | 1.1754 | 0.2734 | 0.1762 |
| 3082 | 9 | - | 0.3992 | 0.7921 | 1.4045 | 0.9383 | 0.4720 | 0.3975 |
| 3110 | 10 | - | 0.3526 | 0.7998 | 1.2606 | 0.8413 | 0.4511 | 0.3654 |
| 3118 | 11 | - | 0.4693 | 1.0066 | 1.1801 | 1.0622 | 0.4872 | 0.2672 |
| 3012 | 1 | + | 0.4440 | 1.4432 | 0.4406 | 1.6114 | 0.4515 | 0.2432 |
| 3020 | 2 | + | 0.6421 | 1.0284 | 1.3161 | 0.8921 | 0.6262 | 0.5400 |
| 3023 | 3 | + | 0.4760 | 0.7981 | 1.0000 | 0.5934 | 0.2310 | 0.3009 |
| 3040 | 4 | + | 0.6719 | 0.8398 | 1.2740 | 0.7259 | 0.2996 | 0.3784 |
| 3049 | 5 | + | 0.5112 | 0.9713 | 1.3513 | 0.7823 | 0.3387 | 0.4030 |
| 3056 | 6 | + | 0.6103 | 1.3715 | 0.5100 | 1.7307 | 0.3617 | 0.1874 |
| 3079 | 7 | + | 1.6036 | 1.1495 | 0.7285 | 1.3375 | 0.4124 | 0.2952 |
| 3081 | 8 | + | 1.1222 | 0.8942 | 0.5130 | 1.2949 | 0.4346 | 0.2583 |
| 3085 | 9 | + | 0.4161 | 0.9395 | 1.1992 | 0.9451 | 0.6665 | 0.4637 |
| 3107 | 10 | + | 0.5402 | 0.9351 | 1.3654 | 1.0016 | 0.5347 | 0.5092 |
| 3121 | 11 | + | 0.4242 | 1.2581 | 1.3110 | 1.1429 | 0.5469 | 0.4098 |

A. How would you analyze the data to test the hypothesis that *hotsix* + animals express a higher level than *hotsix* - for each of these thermally-named genes? Perform this analysis in R and report appropriate p-values for each gene.

> hot<-read.table("hot.txt",head=T)

> hot

animal pair hotsix lukewarm tepid mild1 athermal bathwater coldshower

1 3011 1 - 0.4202 1.0750 0.3974 0.9355 0.4695 0.1688

2 3022 2 - 0.4718 1.0288 1.2747 0.6859 0.5616 0.4862

…

21 3107 10 + 0.5402 0.9351 1.3654 1.0016 0.5347 0.5092

22 3121 11 + 0.4242 1.2581 1.3110 1.1429 0.5469 0.4098

> attach(hot)

*# Since the samples are explicitly paired, the normality that matters for the t test is the difference between samples, e.g.:*

> lm<-lukewarm[1:11]

> lp<-lukewarm[12:22]

> luke<-cbind(lm,lp)

> luke

lm lp

[1,] 0.4202 0.4440

[2,] 0.4718 0.6421

[3,] 0.5351 0.4760

[4,] 0.4955 0.6719

[5,] 0.4299 0.5112

[6,] 0.5609 0.6103

[7,] 1.4554 1.6036

[8,] 0.8959 1.1222

[9,] 0.3992 0.4161

[10,] 0.3526 0.5402

[11,] 0.4693 0.4242

> lm-lp

[1] -0.0238 -0.1703 0.0591 -0.1764 -0.0813 -0.0494 -0.1482 -0.2263 -0.0169 -0.1876 0.0451

# test normality of the difference between hotsix+ and hotsix- for each pair

> shapiro.test(lp-lm)

Shapiro-Wilk normality test

data: lp - lm

W = 0.9304, p-value = 0.4147

# and equality of variance:

> var.test(lm,lp)

F test to compare two variances

data: lm and lp

F = 0.7721, num df = 10, denom df = 10, p-value = 0.6904

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

0.2077404 2.8698384

sample estimates:

ratio of variances

0.7721278

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | *lukewarm* | *tepid* | *mild1* | *athermal* | *bathwater* | *coldshower* |
| Shapiro.test | 0.4147 | 0.7611 | 0.06518 | 0.0129 | 0.9437 | 0.402 |
| var.test | 0.6904 | 0.2439 | 0.9675 | 0.7063 | 0.8103 | 0.939 |

# Depending on what one believes about the evidence for/against normality in *mild1* and *athermal* one could argue for either a parametric approach on the grounds that the lone “significant” departure is no more than one would expect by chance in 6 assays, or one may argue that at least two genes are not persuasively normal and a non-parametric approach should be more conservative and not too much less powerful. Each test provides evidence against the null hypothesis given certain assumptions and one may explain uncertainty in the assumptions and present both approaches, provided the rationale for doing so is clear and specified in advance. I would also accept an argument that, for this data structure, one might expect skewing of the distribution to be a specific function of the gene or assay and thus use *t* for most, but use the non-parametric test for *athermal* and possibly *mild1*. However one chooses to view it, the test or tests must be specified before performing either – one should even specify how possible outcomes of the normality and equality of variance tests will lead to the analysis before testing the data distributions. Any test must be paired-sample as this is explicit in the data collection. The question also specifies a directional hypothesis, so one tail is appropriate; e.g.:

> t.test(lukewarm~hotsix, paired=T, alternative="less")

Paired t-test

data: lukewarm by hotsix

t = -2.983, df = 10, p-value = 0.006868

alternative hypothesis: true difference in means is less than 0

95 percent confidence interval:

-Inf -0.03481709

sample estimates:

mean of the differences

-0.08872727

# Possible nominal p-values:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *lukewarm* | *tepid* | *mild1* | *athermal* | | *bathwater* | | *coldshower* | |
| paired t | 0.006868 | 0.01013 | 0.08552 | 0.03468 | 0.01015 | | 2.429e-06 | |
| Wilcoxon signed rank | 0.01221 | 0.01611 | 0.02686 | 0.01611 | 0.009277 | | 0.0004883 | |
| Mixed by normality | 0.006868 | 0.01013 | 0.02686 | 0.01611 | 0.01015 | | 2.429e-06 | |

B. To quantify the support for the independent hypotheses at each gene, one should control for type I error. Adjust the p-values from A using Hommel’s method.

> pt<-c(0.006868, 0.01013, 0.08552, 0.03468, 0.01015, 2.429e-06)

> p.adjust(pt,method="hommel",6)

[1] 2.0604e-02 3.0390e-02 8.5520e-02 6.9360e-02 3.0450e-02 1.4574e-05

# Possible adjusted p-vlaues:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| paired t | 2.060e-02 | 3.039e-02 | 8.552e-02 | 6.936e-02 | 3.045e-02 | 1.457e-05 |
| Wilcoxon signed rank | 0.02442 | 0.02686 | 0.02686 | 0.02686 | 0.02416 | 0.00293 |
| Mixed by normality | 2.060e-02 | 2.417e-02 | 2.686e-02 | 2.686e-02 | 2.417e-02 | 1.457e-05 |

C. How would you use this data to test the hypothesis that *hotsix* has the predicted effect on expression of these six genes as a group? Perform the test in R and report an appropriate p-value.

*# I show two possible approaches, others approaches were also acceptable for full credit if they were explained and fit the data structure.*

*.*

*# 1. Simple, back-of-the-envelope approach using methods shown in class: pool data for single test.*

> detach(hot)

> pool<-rbind(luke,tep,mil,ath,bat,cold)

> shapiro.test(lm-lp) #hotsix- samples retain header lm and hotsix+ lp, from luke

Shapiro-Wilk normality test

data: lm - lp

W = 0.8761, p-value = 8.551e-06

> var.test(lm,lp)

F test to compare two variances

data: lm and lp

F = 0.8857, num df = 65, denom df = 65, p-value = 0.6262

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

0.5423572 1.4465246

sample estimates:

ratio of variances

0.8857387

> wilcox.test(lm,lp,paired=T,alternative="less")

Wilcoxon signed rank test with continuity correction

data: lm and lp

V = 245, p-value = 1.967e-08

alternative hypothesis: true location shift is less than 0

*# 2. Meta-analysis, which you would have had to find on your own (here is one example, using Fisher’s classic test):*

> Fisher.test<-function(p){

+ Xsq<- -2\*sum(log(p))

+ p.val<-pchisq(Xsq, df=2\*length(p),lower.tail=FALSE)

+ return(c(Xsq=Xsq,p.value=p.val))

+ }

> Fisher.test(p=pt) #meta-analysis based on paired t-tests

Xsq p.value

6.582409e+01 1.918954e-09

> Fisher.test(p=pw) #meta-analysis based on Wilcoxon rank sum tests

Xsq p.value

5.716809e+01 7.370639e-08

> Fisher.test(p=pm) #meta-analysis based on assay-specific tests.

Xsq p.value

6.967376e+01 3.686145e-10

# For either/any approach, since each gene was individually significant, one should expect support for the broader statement about the group to have substantially stronger statistical support.

D. What do you conclude about the effect of *hotsix* on these six genes?

The data provide strong evidence that loss of *hotsix* results in lower expression of these six genes.