Exploratory Data Analysis: Arenosa Case Study

Chapter 1, Lab 4: Solutions

OpenIntro Biostatistics

This lab presents the details of conducting the analysis discussed in Section 1.7.3 of *OpenIntro Biostatistics*. A reader interested in applied data analysis may benefit from working through this lab and reviewing the solutions instead of reading the section in the text. Refer to the section in the text for a brief description of RNA sequencing technology.

Background information

Arabidopsis arenosa populations exist in different habitats, and exhibit a range of differences in flowering time, cold sensitivity, and perenniality. Sensitivity to cold is an important trait for perennials, plants that live longer than one year. It is common for perennials to require a period of prolonged cold in order to flower. This mechanism, known as vernalization, allows perennials to synchronize their life cycle with the seasons such that they flower only once winter is over. Plant response to low temperatures is under genetic control, and mediated by a specific set of cold-responsive genes.

In a recent study, researchers used RNA sequencing (RNA-Seq) to investigate how cold responsiveness differs in two populations of *A. arenosa*: TBG (collected from Triberg, Germany) and KA (collected from Kasparstein, Austria).[1] TBG grows in and around railway tracks, while KA is found on shaded limestone outcrops in wooded forests. As an annual, TBG has lost the vernalization response and does not require extended cold in order to flower; in the wild, TBG plants usually die before the onset of winter. In contrast, KA is a perennial plant, in which vernalization is known to greatly accelerate the onset of flowering.

Winter conditions can be simulated by incubating plants at 4 °C for several weeks; a plant that has undergone cold treatment is considered vernalized, while plants that have not been exposed to cold treatment are non-vernalized. Expression data were collected for 1,088 genes known to be cold-responsive in TBG and KA plants that were either vernalized or non-vernalized; the expression data were obtained from three specimens from each population that were exposed to cold treatment, and three that were not.[2] The data are in the arenosa dataset in the oibiostat package.

Each row corresponds to a gene; the first column indicates gene name, while the rest correspond to expression measured in a particular plant sample. Three individuals of each population were not exposed to cold (non-vernalized, denoted by nv) and three were exposed to cold (vernalized, denoted by v). Expression was measured in gene counts; as a result of normalization between samples, the counts are not integers. A high number of transcripts indicates a high level of gene expression.

---

[1] Baduel P, et al. Habitat-Associated Life History and Stress-Tolerance Variation in *Arabidopsis arenosa*. *Plant Physiology* 2016: **171**: 437-451.

[2] The data have been normalized between samples to allow for comparisons between gene counts.

Exploring the data

The first part of this lab focuses on using numerical and graphical methods to explore the overall picture of how expression levels differ by population (TBG and KA) and vernalization conditions (V and NV). For simplicity, we will work with a sample from the complete dataset.

1. Take a sample of 100 genes (without replacement) from the 1,088 genes in the dataset, using '5011' in the set.seed() command. Name the sample arenosa.sample.

```
#load the data
library(oibiostat)
data(arenosa)

#create arenosa.sample
arenosa.rows = 1:nrow(arenosa)
set.seed(5011)
arenosa.sample.rows = sample(arenosa.rows, size = 100, replace = FALSE)
arenosa.sample = arenosa[arenosa.rows, ]
```

2. Print out the first five rows and first seven columns of arenosa.sample; briefly describe the data matrix shown. Does expression of these genes seem higher in vernalized or non-vernalized plants?

The data matrix shows expression data for the first five genes in arenosa.sample in the six KA specimens. For these gens, expression seems higher in vernalized plants (columns 5-7).

```
arenosa.sample[1:5, 1:7]
```

```
##    gene.name  ka.nv.1  ka.nv.2  ka.nv.3    ka.v.1    ka.v.2   ka.v.3
## 1       PUX4 288.1956 322.5500 305.3483 1429.2920 1408.2546 1487.085
## 2        TZP  79.3582  93.3387  73.4382 1203.4021 1230.4892 1214.032
## 3       GAD2 590.5921 492.6863 458.0224 2639.4170 2645.0475 2705.318
## 4      GAUT6  86.8764  99.2462  57.9775  586.2383  590.0297  579.711
## 5         FB 791.0760 912.1195 746.9437 3430.0320 3680.1209 3467.064
```

3. The three measured individuals in a particular group represent biological replicates: individuals of the same type grown under identical conditions. Collecting data from multiple replicates captures the inherent biological variability between organisms. Thus, averaging expression levels across replicates provides an estimate of the typical expression level in the larger population.

   Using the apply() function as shown in the template, calculate mean expression level across the three replicates for each type of sample: non-vernalized KA, vernalized KA, non-vernalized TBG, and vernalized TBG.

```
ka.nv.mean.sample = apply(arenosa.sample[2:4], 1, mean)
ka.v.mean.sample = apply(arenosa.sample[5:7], 1, mean)
tbg.nv.mean.sample = apply(arenosa.sample[8:10], 1, mean)
tbg.v.mean.sample = apply(arenosa.sample[11:13], 1, mean)
```

4. Using graphical methods, compare expression levels of cold-responsive genes between non-vernalized and vernalized KA, and between non-vernalized and vernalized TBG. How does

gene expression differ between non-vernalized and vernalized plants?

For both KA and TBG plants, expression of cold-responsive genes tends to be higher in vernalized plants than non-vernalized plants; vernalized plants have a higher median expression level relative to non-vernalized plants.

```
#graphical summaries
par(mfrow = c(1, 2))
boxplot(ka.nv.mean.sample, ka.v.mean.sample, names = c("KA NV", "KA V"))
boxplot(tbg.nv.mean.sample, tbg.v.mean.sample, names = c("TBG NV", "TBG V"))
```
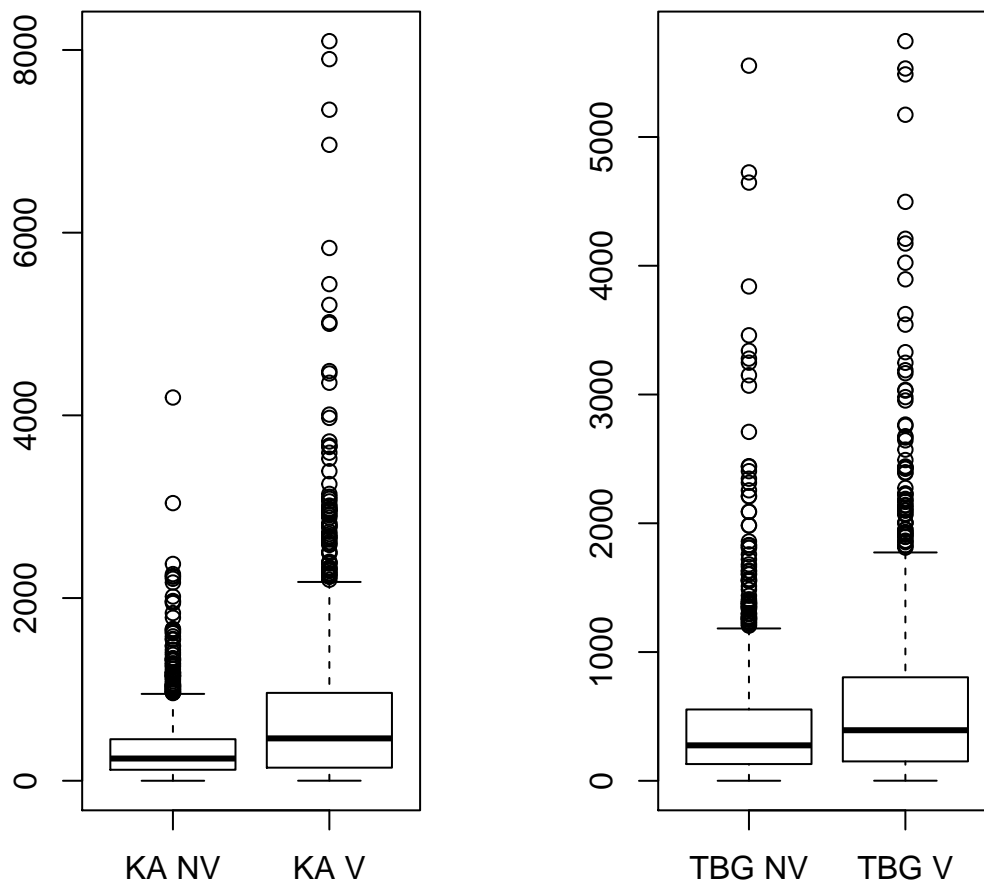

5. Using graphical and numerical methods, compare expression levels of cold-responsive genes between non-vernalized KA and non-vernalized TBG, and between vernalized KA and vernalized TBG. How does gene expression differ between non-vernalized KA and TBG plants? How does gene expression differ between vernalized KA and TBG plants?

Expression level of cold-responsive gense in non-vernalized TBG is slightly higher than in

non-vernalized KA; median expression in NV TBG is 275 versus 243 in NV KA. In contrast, vernalized KA plants show higher expression of cold-responsive genes; median expression in vernalized KA is at 464 versus 393 in vernalized TBG.

```
#numerical summaries
summary(ka.nv.mean.sample)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   120.3   243.4   357.2   454.8  4195.4
```

```
summary(tbg.nv.mean.sample)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    0.544  130.297  275.489  439.203  553.449 5553.907
```

```
summary(ka.v.mean.sample)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    0.448  143.332  463.808  728.380  961.393 8096.570
```
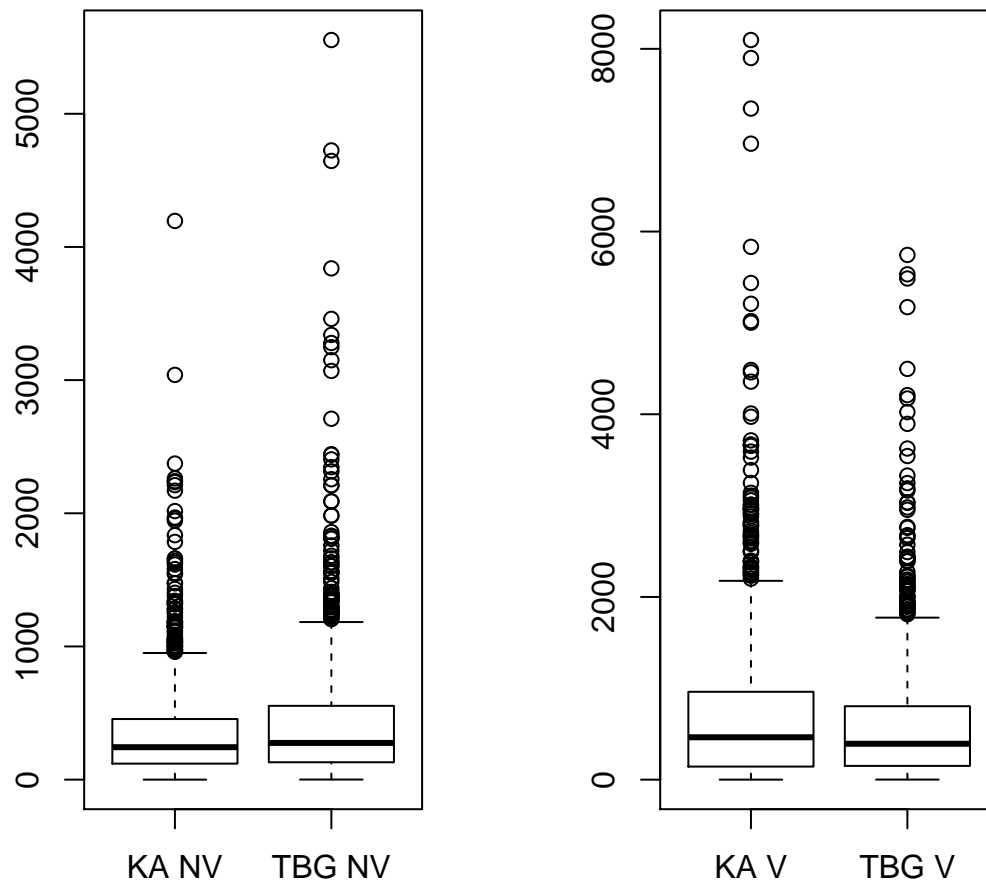
```
summary(tbg.v.mean.sample)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    0.983  150.933  392.933  604.717  803.762 5744.169
```

```
#graphical summaries
par(mfrow = c(1, 2))
boxplot(ka.nv.mean.sample, tbg.nv.mean.sample, names = c("KA NV", "TBG NV"))
boxplot(ka.v.mean.sample, tbg.v.mean.sample, names = c("KA V", "TBG V"))
```

6. Based on the observations made in Questions 4 and 5, does vernalization appear to trigger a stronger change in gene expression in KA plants or TBG plants?

Vernalization appears to trigger a stronger change in gene expression in KA plants. Non-vernalized KA plants start out at overall lower levels of expression in comparison to TBG plants, but vernalized KA plants have higher overall levels of expression than vernalized TBG plants. The difference in median expression level for KA between vernalized and non-vernalized means is $464 - 243 = 221$; for TBG, the difference is $393 - 275 = 118$.

Identifying outliers for responsiveness

7. A more quantitative way to explore the data is to use a gene-level approach. Let the ratio of expression under vernalized conditions to expression under non-vernalized conditions represent the 'responsiveness' of a gene to vernalization.

Using the expression means defined in Question 3, calculate the responsiveness of the genes in `arenosa.sample` for TBG and for KA. Examine the responsiveness for the first three genes in the sample. Interpret the meaning of responsiveness values of 1, less than 1, and greater than 1.

A responsiveness value of 1 occurs when the mean expression level under vernalized conditions is equal to the mean expression level under non-vernalized conditions. A value less than 1 indicates higher expression under non-vernalized conditions, such as with gene 3, in TBG plants. A value greater than 1 indicates higher expression under vernalized conditions. For example, mean expression of gene 2 is almost 15 times higher in vernalized KA than non-vernalized KA.

```
#calculate responsiveness
ka.resp.sample = ka.v.mean.sample/ka.nv.mean.sample
tbg.resp.sample = tbg.v.mean.sample/tbg.nv.mean.sample

#view responsiveness for genes 1-3
ka.resp.sample[1:3]
```

```
##         1         2         3
##  4.720729 14.820819  5.183792
```

```
tbg.resp.sample[1:3]
```

```
##         1         2         3
## 2.0628668 2.7985374 0.9535514
```

8. Create a plot to compare the responsiveness of genes in KA versus TBG for the genes in `arenosa.sample`; a transformation may be helpful. Describe what you see.

A log transformation is especially helpful for clarifying features of the boxplot. On a log scale, values close to 0 are indicative of low responsiveness, while large values in either direction correspond to high responsiveness. The magnitude of response to vernalization in TBG is smaller than in KA. The spread of responsiveness in KA is larger than for TBG, as indicated by the larger IQR and range of values; this indicates that more genes in KA are differentially expressed between vernalized and non-vernalized samples. Additionally, median responsiveness in KA is higher than in TBG.

```
par(mfrow = c(1, 2))
boxplot(tbg.resp.sample, ka.resp.sample, names = c("TBG resp", "KA resp"))
boxplot(log(tbg.resp.sample), log(ka.resp.sample),
    names = c("log(TBG resp)", "log(KA resp)"))
```
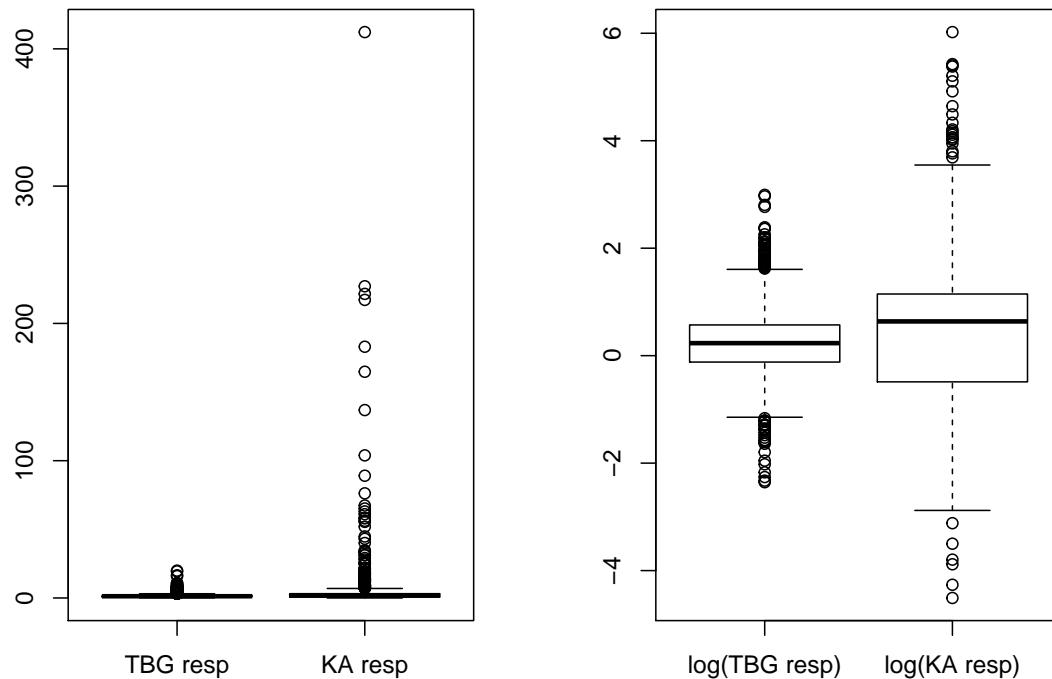
9. Among the 1,088 genes in the complete dataset, which seem to have unusually high or low response to cold treatment; in other words, which genes have a response that is an outlier? Conduct separate analyses for TBG and KA plants, on log-transformed responsiveness.

   a) Report the genes that are high outliers for both KA and TBG, and the genes that are low outliers for both KA and TBG.

There are 17 high outliers for both KA and TBG (see below for the list), and 3 small outliers for both KA and TBG (485694, 921327, HR3).

```
#calculate log responsiveness for all genes
ka.nv.mean = apply(arenosa[2:4], 1, mean)
ka.v.mean = apply(arenosa[5:7], 1, mean)
tbg.nv.mean = apply(arenosa[8:10], 1, mean)
tbg.v.mean = apply(arenosa[11:13], 1, mean)

log.ka.resp = log(ka.v.mean/ka.nv.mean)
log.tbg.resp = log(tbg.v.mean/tbg.nv.mean)


#define outlier boundaries for ka
ka.quart.3 = quantile(log.ka.resp, 0.75, na.rm = TRUE)
ka.quart.1 = quantile(log.ka.resp, 0.25, na.rm = TRUE)
ka.iqr = ka.quart.3 - ka.quart.1
```

```
ka.lb.outlier = ka.quart.1 - 1.5*ka.iqr
ka.ub.outlier = ka.quart.3 + 1.5*ka.iqr

#define outlier boundaries for tbg
tbg.quart.3 = quantile(log.tbg.resp, 0.75, na.rm = TRUE)
tbg.quart.1 = quantile(log.tbg.resp, 0.25, na.rm = TRUE)
tbg.iqr = tbg.quart.3 - tbg.quart.1
tbg.lb.outlier = tbg.quart.1 - 1.5*tbg.iqr
tbg.ub.outlier = tbg.quart.3 + 1.5*tbg.iqr

#identify large outliers
arenosa.response = data.frame(arenosa$gene.name, log.tbg.resp, log.ka.resp)

which.tbg.pos.out = log.tbg.resp > tbg.ub.outlier
tbg.pos.out = as.matrix(arenosa.response[which.tbg.pos.out, ])
order.pos.tbg.out = order(log.tbg.resp, decreasing = TRUE)
ordered.pos.tbg.out = as.matrix(arenosa.response[order.pos.tbg.out, ])
ordered.pos.tbg.out[1:10, ] #show the first 10 outliers
```

```
##      arenosa.gene.name log.tbg.resp    log.ka.resp
## 477 "918965"          " 2.992054867" " 4.4893456"
## 850 "493317"          " 2.964571823" " 5.2095451"
## 165 "887585"          " 2.805206081" " 4.1439593"
## 329 "949566"          " 2.769357198" "       Inf"
## 634 "DEAR5"           " 2.388950946" " 2.6559185"
## 12  "AtPDC1"          " 2.367914186" " 6.0215064"
## 93  "ELIP2"           " 2.359683277" " 2.8414385"
## 453 "LHY"             " 2.257776984" " 4.1111471"
## 20  "919141"          " 2.203273814" " 4.9191118"
## 8   "CYP75B1"         " 2.195286715" " 5.4005191"
```

```
which.ka.pos.out = log.ka.resp > ka.ub.outlier
ka.pos.out = as.matrix(arenosa.response[which.ka.pos.out, ])
order.pos.ka.out = order(log.ka.resp, decreasing = TRUE)
ordered.pos.ka.out = as.matrix(arenosa.response[order.pos.ka.out, ])
ordered.pos.ka.out[1:10, ] #show the first 10 outliers
```

```
##      arenosa.gene.name log.tbg.resp    log.ka.resp
## 329  "949566"          " 2.769357198" "       Inf"
## 12   "AtPDC1"          " 2.367914186" " 6.0215064"
## 18   "AtGolS3"         " 2.080005953" " 5.4252137"
## 8    "CYP75B1"         " 2.195286715" " 5.4005191"
## 22   "316086"          " 1.011232750" " 5.3805349"
## 850  "493317"          " 2.964571823" " 5.2095451"
## 37   "RAP2.1"          " 1.775293835" " 5.1046451"
## 20   "919141"          " 2.203273814" " 4.9191118"
## 1055 "ATSPS4F"         "-0.109118341" " 4.6432735"
## 477  "918965"          " 2.992054867" " 4.4893456"
```

```
#identify common large outliers
##using intersect()
tbg.pos.out = arenosa[which.tbg.pos.out, 1]
ka.pos.out = arenosa[which.ka.pos.out, 1]
intersect(tbg.pos.out, ka.pos.out)
```

```
##  [1] "VIN3"     "CYP75B1"  "AtPDC1"   "485985"   "AtGolS3"  "919141"
##  [7] "ATGRP9"   "RAP2.1"   "ELIP"     "ADH"      "887585"   "949566"
## [13] "LHY"      "918965"   "476817"   "BETA-VPE" "493317"
```

```
##using (merge)
merge(arenosa.response[which.tbg.pos.out, ], arenosa.response[which.ka.pos.out, ])
```

```
##    arenosa.gene.name log.tbg.resp log.ka.resp
## 1             476817     1.675762    3.802519
## 2             485985     2.128377    3.949113
## 3             493317     2.964572    5.209545
## 4             887585     2.805206    4.143959
## 5             918965     2.992055    4.489346
## 6             919141     2.203274    4.919112
## 7             949566     2.769357         Inf
## 8                ADH     1.789649    4.015278
## 9            AtGolS3     2.080006    5.425214
## 10            ATGRP9     1.791859    4.210084
## 11            AtPDC1     2.367914    6.021506
## 12          BETA-VPE     1.888220    4.177205
## 13           CYP75B1     2.195287    5.400519
## 14              ELIP     2.045299    4.333618
## 15               LHY     2.257777    4.111147
## 16            RAP2.1     1.775294    5.104645
## 17              VIN3     2.175460    4.063880
```

```
#identify small outliers
which.tbg.neg.out = log.tbg.resp < tbg.lb.outlier
tbg.neg.out = as.matrix(arenosa.response[which.tbg.neg.out, ])
order.neg.tbg.out = order(log.tbg.resp, decreasing = FALSE)
ordered.neg.tbg.out = as.matrix(arenosa.response[order.neg.tbg.out, ])
ordered.neg.tbg.out[1:10, ] #show the first 10 outliers
```

```
##      arenosa.gene.name log.tbg.resp   log.ka.resp
## 1018 "HR3"             "-2.355093192" "-4.5095199"
## 770  "KCS12"           "-2.325655481" "-2.6250338"
## 1012 "921327"          "-2.259238757" "-4.2637559"
## 508  "PME35"           "-2.171382098" "-2.0818175"
## 988  "488795"          "-2.021821893" "-2.2089513"
## 961  "494886"          "-1.953940569" "-2.2603057"
## 1052 "921527"          "-1.796035333" "-1.8909080"
## 1063 "GAMMA-TIP"       "-1.639803059" "-1.5362122"
## 119  "AtBRN1"          "-1.608860587" "-1.6654878"
```

```
## 646  "479015"            "-1.586085617" "-2.1743573"
```

```
which.ka.neg.out = log.ka.resp < ka.lb.outlier
ka.neg.out = as.matrix(arenosa.response[which.ka.neg.out, ])
order.neg.ka.out = order(log.ka.resp, decreasing = FALSE)
ordered.neg.ka.out = as.matrix(arenosa.response[order.neg.ka.out, ])
ordered.neg.ka.out[1:10, ] #show the first 10 outliers
```

```
##       arenosa.gene.name log.tbg.resp   log.ka.resp
## 1018 "HR3"               "-2.355093192" "-4.5095199"
## 1012 "921327"            "-2.259238757" "-4.2637559"
## 1069 "318143"            "-0.172750230" "-3.8824674"
## 753  "485694"            "-1.356991157" "-3.7962238"
## 764  "AtGH9B13"          "-1.124686334" "-3.4990328"
## 758  "481373"            "-0.729807028" "-3.1177908"
## 1060 "898023"            "-0.107203964" "-2.8802413"
## 843  "GIG1"              "-0.409852643" "-2.8765108"
## 888  "474144"            "-0.957172973" "-2.8121210"
## 875  "SKS1"              "-0.610543918" "-2.7779473"
```

```
#identify common small outliers
##using intersect()
tbg.neg.out = arenosa[which.tbg.neg.out, 1]
ka.neg.out = arenosa[which.ka.neg.out, 1]
intersect(tbg.neg.out, ka.neg.out)
```

```
## [1] "485694" "921327" "HR3"
```

```
##using (merge)
merge(arenosa.response[which.tbg.neg.out, ], arenosa.response[which.ka.neg.out, ])
```

```
##   arenosa.gene.name log.tbg.resp log.ka.resp
## 1            485694    -1.356991    -3.796224
## 2            921327    -2.259239    -4.263756
## 3               HR3    -2.355093    -4.509520
```

b) Why might it be interesting to further investigate cold-responsive genes that are either high outliers or low outliers for both KA and TBG?

These highly cold-responsive genes likely play a role in how plants cope with colder temperatures; they could be involved in regulating freezing tolerance, or controlling how plants detect cold temperatures.