Exploratory Data Analysis: Arenosa Case Study

Chapter 1, Lab 4

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`.

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?

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.[3]

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?

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?

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?

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.

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.

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?

[3] Recall that the `apply()` function was introduced in Lab 3 of this chapter; refer to the Lab Notes for an explanation of the function syntax.

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.

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?