Integrated Assignment, CBW 2013

Nicholas Harding
OICR

June 24-25, 2013

Let's see if we can put all of today's material together into an analysis pipeline for a microarray experiment. This experiment compares gene expression in a rat strain and a mouse strain subjected to a dose of the dioxin TCDD. We are interested in genes that show differential expression between the treatment group (TCDD) and the control group (corn oil).

1 DATA

Download the CEL files from the integrated assignment section of the wiki. You will need to make your own phenotype data file, copy the format used in the examples in the lecture earlier. Double check your CEL files are indentified correctly!

2 PRE-PROCESSING

QA/QC the data as before and determine if any samples should be omitted. Next, pre-process the data, do you believe that an RMA pre-processing approach is sufficient, or are you more comfortable using MAS5?

3 STATISTICAL ANALYSIS

Perform differential testing using t-tests to compare:

1. TCDD versus corn oil (control) in rat.
2. TCDD versus corn oil (control) in mouse.

Calculate the *fold change* for each probe between control and experimental conditions.
Use a multiple-testing adjustment to control your p-values. You may find `https://en.wikipedia.org/wiki/Multiple_comparisons` helpful if you are unclear what the multiple testing problem is.
R has a built in function that handles p-value correction. Access the help by entering `?p.adjust()`. You can use the *FDR* adjustment method, or try alternative approaches. If you are curious, try to find which correction method is the most conservative.

4 VISUALIZATION

Now let's visualize the results in several different ways:

1. Make a density plot of your normalized intensity values.
2. Make a histogram of your p values before and after multiple testing correction. What is the correction doing?
3. Select a probe (perhaps your highest fold change), and create a boxplot of TCDD expression vs control.
4. Make a scatterplot of the fold-changes of your probes versus the associated p-values, this is a volcano plot!

5 INTERPRETATION

Repeat steps 1-4 with the appropriate alternative CDF. Does the CDF change alter your results in any meaningful way? How many more/less differentially expressed probes do you observe? From these data alone, are you convinced that you should be using one?

6 BOOTSTRAPPING P-VALUES

OPTIONAL: IF YOU HAVE TIME
T tests are subject to assumptions- most pertinently that your values are normally distributed. It is a *parametric* test, as opposed to a *non-parametric* test. For more details see `http://changingminds.org/explanations/research/analysis/parametric_non-parametric.htm` or `http://www.vassarstats.net/textbook/parametric.html`.

1. From what you have seen- does this assumption of normality seem reasonable?
2. As an alternative to a t-tests perform a non parametric test (remember to perform a p value adjustment).
3. Compare the p values you generate with those from the t-test. What do you notice?
4. Alternatively, we can estimate our own p values using *Bootstrapping*.

Ensure you are comfortable with what bootstrapping is, ask for help if you are unsure! You may find the page at `http://en.wikipedia.org/wiki/Bootstrapping_(statistics)` helpful.
Bootstrapping approach:

1. For each gene, draw 3 samples from all samples (i.e. TCDD and control). Calculate the median. Hint: use the `sample()` function.

2. Is this value higher or lower than the median of the TCDD only? If it is lower, we take this as a positive result. Otherwise negative. The ratio of +ve to -ve results provides an estimate of our p-value, as the random sampling is an estimate of our null distribution.

3. If you think it's necessary, perform a p value adjustment as above.

The more times we do this the more accurate your estimate! How many iterations are appropriate? How can you tell when you have performed sufficient iterations?
What are the limitiations of the bootstrapping approach? What other alternatives are there to t-tests? (Ask a member of CBW if you need help).

7 COMPARING HOMOLOGUES

OPTIONAL: IF YOU HAVE TIME

1. Visit `http://orthologene.org/`

2. Try to download mapping information between mouse and rat Entrez gene IDs

3. Look for common gene effects between species.

4. How should we present this data?