Softwarewerkzeuge der Bioinformatik

Prof. Dr. Volkhard Helms
PD Dr. Michael Hutter, Markus Hollander,
Andreas Denger, Marie Detzler, Larissa Fey

Saarland University
Department of Computational Biology

Winter semester 2020/2021

Tutorial 8

January 7, 2021

Gene Expression

In this tutorial, you are going to perform two types of microarray analysis on two different types of microarray datasets. All datasets are taken from patients or cell lines with *Acute Lymphoblastic Leukemia* (T-ALL), both before and several hours after being given a potential treatment.

Exercise 8.1: Preparation

Typically, this type of analysis is performed with the *Bioconductor* packages that are available for the R programming language. Luckily for us, the webserver *CARMAWeb* (https://carmaweb.genome.tugraz.at/) provides a frontend for these packages, so we don't need to do any programming ourselves.

(a) Visit the website and create a user account. You don't need to provide your email adress, just a username and a password. Use your credentials to log in.

(b) In this exercise, we will use the test datasets provided by the webserver. Click on the menu entry called **Data directory** on the left. In this directory, you will find a button that loads the test datasets into your data directory. The files in this directory can then be used for the remaining exercises.

Exercise 8.2: Fold change analysis for two-color microarrays

First, we will calculate the fold changes between the red and the green signals of a two-color microarray. The green signals represents the expression of genes from a T-ALL cell line without the treatment, the red signal shows the gene expressions of the same cell line after being exposed to the treatment for 6 hours.

(a) Preprocessing

(1) Click on *New Analysis*, and select *Perform a two color microarray analysis*. Now we need to add the table with gene expression data. Add the *GenePix* file with the name *Nr026004.gpr* and proceed to the next step.

(2) CARMAWeb has already deducted from the file-name that we want to analyze a *GenePix* file, and has selected the correct channels for the green and red signals. The test files also include a GAL file, which maps the spots on the microarray to gene names, as well as other annotations. Select the file *Batch08_modUG.GAL* in the dropdown menu.

(3) Next, perform the preprocessing. Choose *normexp* for background correction, *printtiploess* for within-array-normalization and *quantile normalization* for between-array-normalization, then proceed to the next step.

(4) We can skip the replicate handling, as we are only looking at one array.

(b) Analysis

(1) On the next page, select **Fold change analysis to define differentially expressed genes**.

(2) Now it is time for calculating the fold changes. Perform one comparison between the red channel and the green channel of *Nr026004.gpr*. Make sure that red vs. green is selected.

(3) Next, we need to set a log fold change (LFC) threshold, to narrow down the results to just the genes with the highest change in expression. We only want to look at genes with a LFC of 1.5 or more, and -1.5 or less (*Hint: LFC is called "M (log ratio)" here.*). Also let the program draw an MA plot of the comparison.

(4) Since we are only performing one comparison, there are no comparisons to combine. Therefore, we can now finally start the analysis.

(c) Interpretation of results

(1) Open the PDF file. How many genes were defined as up or down regulated according to our fold change cutoff? Interpret the MA plot on the last page.

(2) The leukemia cells were treated with *Glucocorticoids* (GC), a class of drugs commonly used for treatment of accute lymphoblastic leukemia. Their cytotoxic effect is mediated through their binding to the glucocorticoid receptor (GR), which is encoded by the gene **NR3C1**. Was the expression of GR affected by the presence of GC? (*Hint: open the .txt file*)

Exercise 8.3: Differential gene expression analysis for single-channel microarrays

Next, we will perform a differential gene expression analysis using a t-test, this time on single-color Affymetrix *hgu133plus2* microarrays from accute lymphoblastic leukemia (T-ALL) patients. Tissue samples were taken both before, as well as 6-8 hours after receiving the glucocorticoid (GC) treatment for T-ALL.

(a) Preprocessing

(1) Start an Affymetrix GeneChip analysis, and add the six files ending with *.CEL.gz*

(2) The GeneChips are of type *hgu133plus2*, so we need to select *conventional 3' array*. For preprocessing, we will choose *robust multiarray average* (RMA), since it can be calculated faster than the Affymetrix standard method, which is called *MAS5*. Let the program draw a histogram before and after normalization.

(3) Again, we can skip replicate handling. We need the replicates to have enough samples for the t-test.

(b) Analysis

(1) Select **Test statistics to detect differentially expressed genes**.

(2) Define two groups: The samples without treatment (0h) are group 0, the samples taken after treatment (6h or 8h) are in group 1.

(3) Next, we need to select a test statistic. There are two samples for each patient that were taken at different points in time. So a *paired t-test* is the most appropriate for this purpose. Select *paired moderated t-statistic (limma)* from the drop-down menu, which is suggested for small group sizes. Make sure that the pairs align with the patient identifier (patient 2 is pair 1, patient 20 is pair 2, patient 25 is pair 3).

(4) When testing many hypotheses at the same time, multiple testing correction needs to be applied. Select Bonferroni, as well as Benjamini-Hochberg (BH) as your multiple hypothesis testing methods.

(5) Finally, we want the program to give us an additional file with the top 100 genes according to their p-value. To easily find out which spot on the microarray corresponds to which gene identifier, it is also a good idea to let CARMAweb include the gene annotations. Let it draw a volcano plot of the raw p-values.

(c) Interpretation of results

(1) The histograms taken before and after normalization should be in your results folder, as pdf files starting with "analysis". Interpret them in relation to each other. Did the normalization algorithm do a good job?

(2) Interpret the volcano plot. What do the axes stand for? Where would a significantly differentially expressed gene with a high fold change be located in the plot?

(3) Take a look at the top 100 genes according to p-value. Which gene has the highest mean log fold change (meanM)?

(4) Now look at the raw p-value of the gene, as well as the p-values after multiple testing correction. Why is fold change alone not enough for finding significantly D.E. genes? Can you explain the difference between the Bonferroni- and Benjamini-Hochberg correction?

(5) Did the GC treatment have a significant effect on the gene expression in T-ALL patients, according to this analysis?

Have fun!