Lab: Two-Color Microarray Quality Assessment and Quality Control

Nolwenn Le Meur

10 January, 2007

For this lab, we make use of a two-color microarray experiment from B. Knudsen's lab in collaboration with P. Nelson's lab (Fred Hutchinson Cancer Research Center, Seattle WA). The experimental design is as follows:

- 5 cell cultures (biological samples) PEC32, PEC34, PEC36, PEC39, PEC40.
- 2 drugs (SFN and HGF) each present at two levels (`Low` and `High`).
- each chip has 15488 spots representing 2 technical replicates of each of 7744 features.
- reference samples are labeled with Cy5.

You will use different methods to assess and control the quality of the data before and after different pre-processing analysis. You will need the packages *limma*, *arrayMagic*, *marray*, *arrayQuality*, *convert*. and the data package *HGFSNF*.

1 Quality assessment of raw data

We begin by loading up the appropriate libraries and the data.

```
> library("Biobase")
> library("HGFSNF")
> library("limma")
> library("arrayMagic")
> library("arrayQuality")
> library("marray")
> library("convert")
```

Locate the data directory in the *HGFSNF* package and setup a working directory.

```
> dataDir <- system.file("extdata", package = "HGFSNF")
> workDir <- "/home/nlemeur/"
```

Exercise 1: Reading the files
The files were processed with GenePix as image analysis software producing a .gpr file for each array. A targets file ("ArrayDesc") contains the names of the .gpr files and further annotation

about the experiment.

Read the file using the *limma* package and flag the negative spots if any. Have a look at the data structure and save the object as `RG`.

```
> targets <- readTargets("ArrayDesc", path = dataDir)
> RG <- read.maimages(targets$FileName, path = dataDir, source = "genepix",
+     wt.fun = wtflags(weight = 0, cutoff = -50))
> save(RG, file = file.path(workDir, "RG.rda"))
```

Read the annotation file named "PEDB_ARRAY_annotations.csv" and have look at a few rows. Some spots are empty and some have failed the sequencing analysis ("EMPTY"and "Failed sequencing" probes, respectively). Identify those features and set their weight to zero.

```
> annotation <- read.csv(file.path(dataDir, "PEDB_ARRAY_annotations.csv"),
+     as.is = TRUE)
```

Exercise 2: Quality assessment *via* visualization plots
An essential step in the analysis of any microarray data is to check the quality of the data from the arrays. You can use diffferent plots such as image plots, MA-plots or boxplots. Those plots will give you different views and will show different properties of the data.

As an example, draw the image plot for the background intensities in Cy5 and Cy3 for the first 3 slides. What can you tell? The `plotMA3by2` function gives an easy way to produce MA-plots for all the arrays in a large experiment. this function writes plots to disk as .png files, 6 plots to a page.

Exercise 3: Technical replicates correlation
A slide is composed of 15488 features arranged in 32 blocks. The layout of each block is 22 columns by 22 rows. The slides are designed such as half of the slide (the first 7744 features) is the technical replicate of the other half (the last 7744 features).

Verify that it is actually true by matching the spot names from one half of the slide to the other half. Are there spot names present more that twice?

Since half of the slide is the replicate of the other half, we should have a high correlation between probe intensities. Calculate the coefficient of correlation between the technical replicates within a slide per channel to evaluate the level of correlation.

Exercise 4: Diagnostic plots
Using the *arrayQuality* package, generate diagnostic plots (`maQualityPlots`) for a qualitative assessment of 1 slide of your choice. The results are saved as .png files in the working directory.

2 Pre-processing: background correction, normalization and quality control

Exercise 5: Background correction
Background correction is more important than often appreciated because it impacts markedly on the variability of the log-ratios for low intensity spots. Use *limma* and the `backgroundCorrect` function to compare the effect of different background corrections. Visualize the results as MA plots.

Exercise 6: Normalization
Use *limma* to perform a variance stabilization normalization (`vsn`) taking into account the technical replicates. Verify the result of the normalization by plotting the standard deviation versus the rank of the mean (`meanSdPlot` function).

Apply on the same raw dataset the default `normalizationWithinArrays` function proposed by *limma*. Compare the duplicate correlation after the two different normalization methods using the `duplicateCorrelation` function.

What can you tell? Can you explain the differences?

Save the `ExpresssionSet` issued from the *vsn* normalization (for convenience use the filename *M.rda*).

Exercise 7: QC/QA of the pre-processing steps
Using the *vsn* normalized data, filter out the "bad" features (i.e."EMPTY" and "Failed Sequencing"), divide and combine the dataset so that you have the technical replicates as individual element. Draw different heatmaps using the different information available in the targets file to label the samples (*e.g.* Treatment, Date of data acquisition). For computational efficiency and visualization purposes use a sample of 500 genes and do not display the gene names. What can you tell?

```
> sessionInfo()

R version 2.5.0 Under development (unstable) (--)
x86_64-unknown-linux-gnu

locale:
LC_CTYPE=en_US;LC_NUMERIC=C;LC_TIME=en_US;LC_COLLATE=en_US;LC_MONETARY=en_US;LC_MESSAGES=en_US;LC_PAPER=

attached base packages:
 [1] "grid"      "splines"   "tools"     "stats"     "graphics"  "grDevices"
 [7] "utils"     "datasets"  "methods"   "base"

other attached packages:
     statmod arrayQuality RColorBrewer     gridBase       hexbin   colorspace
     "1.2.4"     "1.11.0"      "0.2-3"      "0.4-2"      "1.9.0"        "0.9"
     lattice      convert       marray   arrayMagic   genefilter     survival
   "0.13-10"      "1.9.1"     "1.13.0"     "1.13.0"     "1.13.7"       "2.29"
       limma       HGFSNF          vsn      Biobase
     "2.9.8"        "0.1"      "2.0.1"    "1.13.30"
```