Lab 2: Analyzing gene expression microarray data

In this lab we will go through the typical microarray data analysis procedure using some popular Bioconductor packages. The workflow of microarray data analysis usually follows the steps of (1) reading in data (often from binary files), (2) normalization, (3) differential expression detection and (4) generate report. We will show that using Bioconductor packages these tasks can be easily achieved in a few lines of R codes.

The Bioconductor packages to be used in this lab are: oligo (for reading in data and normalization), limma and siggenes (for differential expression), pd.hg.u133a, pd.hg.u133.plus.2 (for annotation and generating reports).

Data:

We will use the microarray data generated by MAQC (MicroArray Quality Control) project phase I, which was initiated by FDA to assess the qualities of microarray technologies and data analysis methods. For details of the project please visit the project webpage at http://www.fda.gov/ScienceResearch/BioinformaticsTools/MicroarrayQualityControlProject. Raw data can be obtained from GEO database (http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE23906. The single tar file (42M) contains 12 zipped CEL files, which were generated from 12 microarrays (2 types of microarrays × 2 samples × 3 replicates). Three replicates for each of the two MAQC samples (A: Universal Human Reference RNA, and B: Human Brain Reference RNA) were profiled by using both the U133A and U133Plus2 microarrays from Affymetrix. The file names are

	U133A array	U133Plus2 array
Sample A	GSM589506, GSM589507, GSM589508	GSM589512, GSM589513, GSM589514
Sample B	GSM589509, GSM589510, GSM589511	GSM589515, GSM589516, GSM589517

We want to explore the expression profiling results from these two platforms and assess their concordances. Analysis results of these data were published earlier at http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-S6-S10, we want to reproduce some of the findings of that paper.

Steps:

1. Install following Bioconductor packages: oligo, limma, siggenes, pd.hg.u133a, pd.hg.u133.plus.2, and hgu133a.db.
2. Do following to download and decompress the raw data from the tar file.
    a. Go to GEO database homepage at http://www.ncbi.nlm.nih.gov/geo/.
    b. Put "GSE23906" in the dataset query box and click "search". This will bring you to the summary page for GSE23906. You can briefly read the descriptions.
    c. Go down to the bottom of the page, under "Supplementary file" table, download GSE23906_RAW.tar (42.1Mb) by clicking the ftp or http link.
    d. Decompress the tar file to your lab3 folder using a software. The individual files contained in the tar are gz files (compressed). You don't need to decompress them because the function in oligo can read in without decompressing.
3. Follow the R scripts to:
    - Compare log fold changes of gene expressions generated from different platforms.
    - Compare the log fold changes of gene expressions to the gold standard generated by RT-PCT (Taqman).
    - Using limma to detect DE genes, and compare to gold standard.