

# Module 3: microarray/pathway analysis

Perry Moerland

April 27, 2010

☞ Information on how to log on to a PC in the exercise room and the UNIX server can be found here: <http://bioinformaticslaboratory.nl/twiki/bin/view/BioLab/EducationBioinformaticsII>. Don't forget to enable X11 forwarding when starting up PuTTY.

☞ These exercises depend on the chapters from (Gentleman, Carey, Huber, Irizarry, and Dudoit 2005) used in module 2.

## Signatures of Oncogenic Pathway Deregulation in Human Cancers

The ability to define cancer subtypes, recurrence of disease, and response to specific therapies using DNA microarray-based gene expression signatures has been demonstrated in multiple studies. Such data is also of substantial importance to the analysis of cellular signaling pathways central to the oncogenic process. With this focus, Bild et al. (2006) have developed a series of gene expression signatures that reliably reflect the activation status of several oncogenic pathways. When evaluated in several large collections of human cancers, these gene expression signatures identify patterns of pathway deregulation in tumors, and clinically relevant associations with disease outcomes. Combining signature-based predictions across several pathways identifies coordinated patterns of pathway deregulation that distinguish between specific cancers and tumor sub-types. Clustering tumors based on pathway signatures further defines prognosis in respective patient subsets, demonstrating that patterns of oncogenic pathway deregulation underlie the development of the oncogenic phenotype and reflect the biology and outcome of specific cancers. Furthermore, predictions of pathway deregulation in cancer cell lines are shown to coincide with sensitivity to therapeutic agents that target components of the pathway, underscoring the potential for such pathway prediction to guide the use of targeted therapeutics. In this module, you will analyze part of the data from (Bild et al. 2006).

## Getting the data

All data from (Bild et al. 2006) is available from the NCBI Gene Expression Omnibus (GEO): <http://www.ncbi.nlm.nih.gov/projects/geo/query/browse.cgi?mode=series&submitter=2133>. Data in GEO comes in various forms and the Bild data are stored as GEO series (GSE). A GSE defines a set of related GEO samples (GSM), how the samples are related, and if and

how they are ordered. A GSE can be quite heterogeneous, for example, it might contain samples run on different array platforms. The package `GEOquery` contains the function `getGEO` that transfers data from GEO and then parses the data into useful data structures.

The first GSE that we will study is GSE3151. For this series of arrays, RNA was extracted from human mammary epithelial cells expressing oncogenes (10 Myc, 9 E2F3, 10 Ras, 7 Src, 9  $\beta$ -catenin samples) or control (10 GFP samples), so 55 samples in total. The platform used is the Affymetrix GeneChip Human Genome U133 Plus 2.0 Array which consists of 54675 probesets. We will call this dataset the *oncogene signature dataset*. See (Downward 2006) for a short introduction into the biology behind the Bild paper.

The second GSE we will study is GSE3158. This is a mouse tumor dataset of 28 samples expressing oncogenes (5 Myc, 6 Rb knockout (=E2F3 activation in human), 3 Ras, 7 HER2) versus control (7 Wild type). It turns out that the 28 samples were all done on two platforms: Affymetrix GeneChip Murine 11K SubA Array Mu11K-A consisting of 6584 probesets and the Affymetrix GeneChip Murine 11K SubB Array Mu11K-B consisting of 6595 probesets. We will call this dataset the *mouse tumor dataset*.

**Exercise 1:** (1 point)

The document `GEOquery.pdf` on the website describes how to convert a GSE to an expression set using the `GEOquery` package. Write a function to do so for the two GSEs mentioned above. Perform some quality control on the resulting expression sets. Is any additional pre-processing necessary?

**Exercise 2:** (0 points)

As said above the mouse tumor dataset was run on two different platforms. Merge the expression sets you created for this dataset in the previous exercise into a single expression set of 28 samples.

## Data analysis

Figure 1A in (Bild et al. 2006) gives an image intensity display of the expression of the genes most highly weighted in a classification model differentiating between each of the oncogenic pathways and the GFP controls.

**Exercise 3:** (1 point)

Make similar pictures for each of the pairwise comparisons by first using  $t$ -tests (using the function `rowttests` from the `genefilter` package, for example) to determine how well each gene separates the two classes (oncogene versus GFP), and then selecting the 200 most differentially expressed genes. Save the figures in png format.

**Exercise 4:** (1 point)

Write the results for each of the pairwise comparisons to a tab-delimited file with columns for Affymetrix probesetID, gene symbol, and  $t$ -statistic.

Compare the files you just generated to the results in Supplementary Table 1 (see website for the link to the Supplementary Information of (Bild et al. 2006)). Which could be the reasons

for the differences you observe?

**Exercise 5:** (1 point)

In this experiment five key oncogenic pathways have been activated by mutational activation of MYC, E2F3, HRAS, SRC or  $\beta$ -catenin (CTNNB1). Investigate the mRNA expression for these specific genes in detail. Do the results correspond to what you would expect? Explain.

**Exercise 6:** (1 point)

Figure 1B in (Bild et al. 2006) shows scatter plots of the different samples. As explained in the clustering lecture, principal component analysis can be used to find such a low-dimensional representation. Use the function *prcomp* to make similar figures following the description in the caption of Figure 1B (singular value decomposition = principal component analysis). Save the figures in png format.

In the tutorial on meta-data and pathways, you have learned how to make sets of gene sets and to test for gene set enrichment. Now apply these techniques to the oncogene signature dataset.

**Exercise 7:** (0 points)

Make KEGG gene sets for the Affymetrix GeneChip Human Genome U133 Plus 2.0.

As you already know, the *limma* package offers a simple way of testing whether a set of genes is enriched for differential expression. It is based on the location tests explained during the lecture and implemented in *geneSetTest*.

**Exercise 8:** (1 point)

Adapt your function from the meta-data exercises to a function that tests which of the KEGG pathways is enriched for differential expression in the *t*-test between each of the oncogenes separately and GFP. Since you are testing for many different pathways, don't forget to adjust for multiple testing using the Benjamini-Hochberg correction. What are the IDs and names of the ten most enriched pathways? Write the results for each of the pairwise comparisons to a tab-delimited file with columns for KEGG ID, KEGG pathway name, and FDR value. Compare and contrast the pathway results for all five comparisons of an oncogene versus GFP. Do the results make sense biologically?

## Classification and validation in mouse models and human cancers

In Exercise 3 you determined the 200 most differentially expressed genes in the oncogene signature dataset. Now we will construct classifiers for this dataset and apply these classifiers to the mouse tumor dataset.

**Exercise 9:** (1 point)

Write a function that builds two classification models of your choice (from packages *e1071* and *MLInterfaces*, for example) for all five comparisons of an oncogene versus GFP separately.

Evaluate the resulting classifiers and report the confusion matrices of false/true positives and false/true negatives.

To map the probe sets across various generations of Affymetrix GeneChip arrays, (Bild et al. 2006) utilized an in-house program, Chip Comparer (<http://tenero.duhs.duke.edu/genearray/perl/chip/chipcomparer.pl>). Each probeset ID in a given Affymetrix gene chip was mapped to the corresponding gene ID. Second, probesets from different gene chips are matched by sharing the same gene ID or orthologous pair of gene IDs in the case of mapping gene chips across species.

**Exercise 10:** (1 point)

Map the Affymetrix GeneChip Human Genome U133 Plus 2.0 to Affymetrix GeneChip Murine 11K SubA and SubB. Merge the two resulting text files. How many unique human probesets remain? Then retrain the classifiers of the previous exercise using only those human probesets that can be mapped to a mouse probeset. How would you generate the mapping from Affymetrix GeneChip Human Genome U133 Plus 2.0 to Affymetrix GeneChip Murine 11K SubA and SubB using only Bioconductor packages?

For three of the human oncogenic pathway signatures, there are matching mouse models that could be used for validation: Myc, Ras and E2F3. The matching mouse models are Myc, Ras, and Rb knockout.

**Exercise 11:** (1 point)

Redo the analysis leading to Figure 2A in (Bild et al. 2006) where the Myc, Ras and E2F3 classifiers from the previous exercises are used to classify the mouse tumor data `eset3158`. Save the figures in png format.

The third GSE that we will study is GSE3141. For this series of arrays RNA was extracted from human primary lung tumors, 111 samples in total. The platform used is again the Affymetrix GeneChip Human Genome U133 Plus 2.0 Array which consists of 54675 probesets.

**Exercise 12:** (1 point)

First generate the expression set from the GSE object. Then redo the analysis of Figure 3A in (Bild et al. 2006) using the five classifiers from exercise 9 and hierarchical clustering. Can you also discern a cluster of patients with poor survival? Hint: clinical annotation for this experiment is stored in `Lung_clinical_summary.xls` that you can find at `/data/home/stud00/Module3`.<sup>1</sup>

## References

Bild, A. H., G. Yao, J. T. Chang, J. R. Nevins, and et al (2006). Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* 439(7074), 353–357.

Downward, J. (2006). Signatures guide drug choice. *Nature* 439, 274–275.

---

<sup>1</sup>This and other files can also be downloaded from <http://data.cgt.duke.edu/oncogene.php>.

Gentleman, R., V. Carey, W. Huber, R. Irizarry, and S. Dudoit (Eds.) (2005). *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. New York: Springer.