# ANALYZING DNA MICROARRAYS WITH UNDERGRADUATE STATISTICIANS

Johanna Hardin, Laura Hoopes, and Ryan Murphy
Pomona College, United States
jo.hardin@pomona.edu

*With advances in technology, biologists have been saddled with high dimensional data that need modern statistical methodology for analysis. DNA microarrays are able to simultaneously measure thousands of genes (and the activity of those genes) in a single sample. Biologists use microarrays to trace connections between pathways or to identify all genes that respond to a signal. The statistical tools we usually teach our undergraduates are inadequate for analyzing thousands of measurements on tens of samples. The project materials include readings on microarrays as well as computer lab activities. The topics covered include image analysis, filtering and normalization techniques, and statistical methods. The course materials are designed for someone with little or no statistical background, but due to the novel concepts covered, they could easily be adjusted to accommodate students with practically any background.*

## BACKGROUND

It is apparent from headlines in national newspapers and magazines that recent discoveries in genetics and molecular biology are changing the way we think about medicine, health, pharmaceuticals, and human life. The results of the human genome project have allowed biologists to study the way different organisms work at the genetic level. One recent technological innovation is the microarray: a laboratory chip designed to simultaneously measure activity of thousands of genes in a single sample simultaneously. By comparing multiple samples, we can identify which genes act differently between types of samples (e.g., which genes are different in healthy vs. cancerous tissue samples.)

Measuring thousands of genes on tens or hundreds of samples, however, induces statistical and computational difficulties with which biologists (or statisticians, to some extent) have not previously had to deal. New statistical techniques are constantly being developed to address the issues associated with microarrays, but there is still a gap in the implementation of such techniques by biologists. In particular, there is little or no work being done on educating undergraduate students of *statistics* in topics such as microarray analysis or bioinformatics.

It seems as though biologists are aware of the need for quantitative methods, and they are working toward educating themselves and their students. Hack and Kendall (2005) report, "If biosciences are to evolve from a predominantly descriptive discipline to an information science, practitioners will require enhanced skills in mathematics, computing, and statistical analysis." Bialek and Botstein (2004) also address the need for biologists to improve their quantitative skills in the face of 21st-century biology. And biologists are paying attention to said advice. At Drake University, Jerry Honts (2003) is introducing undergraduate biology students to software and databases in 3 courses. At Davidson College, Malcolm Campbell (2002; Brewster *et al*., 2004) has had undergraduate students perform microarray experiments along with image analysis and clustering techniques.

Statisticians also think that statistics is playing an increasingly important role in biology. From a recent workshop at NSF, Lindsay *et al*. (2004) reported that "the large amounts of data produced by modern biological experiments and the variability in human response to medical intervention produce an increasing demand for statisticians who can communicate with biologists and devise new methods to guide experimental design and biological data analysis." However, there is a conspicuous absence of programs designed to make undergraduate statistics students knowledgeable about the issues facing modern biology.

## MICROARRAY COURSE MODULES

### *Goals*

The materials in this paper are designed for an undergraduate course for quantitatively inclined biology students or biologically inclined statistics students. There are no prerequisites,

but the materials could easily be modified to incorporate prerequisites of introductory statistics, regression, or genetics. In putting together course materials, our goals are,

- To introduce modern statistical techniques to undergraduates (who wouldn't be exposed to them elsewhere)
- To communicate important links between biology and statistics
- To improve the literacy of the students in basic methods and applications of bioinformatics

*Structure of Modules*

The microarray project modules that we have created can be used together as a major part of a course or individually as an add-on to a statistics or biology course. The modules address different aspects of analyzing microarray data and do not depend on the previous section. However, the modules are built with the same structure so that they can easily flow together.

| Modules | Respective educational goals | Pedagogical Components |
|---|---|---|
| a. Analyzing images | a. Collecting data *well* is important! | 1. Educational topics<br>2. Articles and other reading |
| b. Normalizing data | b. How to compare apples and oranges | 3. Computer lab assignments (both in and out of class) |
| c. Class comparison and Class prediction | c. Basic, novel, and fancy statistical techniques | 4. Homework |

The first module is designed to communicate the inherent difficulty in measuring gene activity, even with microarray technology. The reading will be DeRisi *et al*. (1997) and chapters 1-3 from Draghici (2003). The computer lab and homework assignments are based primarily on work designed by Laurie Heyer at Davidson College. She uses *MagicTool* (Heyer *et al*., 2005, http://www.bio.davidson.edu/projects/magic/magic.html), an exploratory data analysis program written entirely by undergraduates, to analyze .tif files of yeast data on adaption during shifting from glucose to ethanol as a carbon source, from DeRisi *et al*. (1997). Heyer has made her labs publicly available (http://gcat.davidson.edu/GCAT/workshop2/derisi_lab.html).

*Normalizing Data*

The second module is designed to convey the importance of normalizing data, filtering data, and identifying outlying values. The lecture topics include discussions of scale vs. location normalization, normalizing across all samples or within a sample, log transformations (their mathematical results as well as their usefulness in practice), investigating flagged spots, and scaling to decrease bias due to dye color (a technical aspect of the microarray.) The topics we cover seem to be full of jargon and high level biology, but the concepts are quite straight forward and as easy to convey to a group of students as any similar statistical concept (e.g., constant variance across groups when performing ANOVA.) The primary motivation of normalization in the micorarray context has to do with the novelty of the technology and its lack of ability to measure as precisely as we might hope.

In addition to covering the above topics in class, the students will read articles (Schuchhardt *et al*., 2000, and Yang *et al*., 2002, chapters 12 and 13 of Draghici, 2003), perform lab activities, and have assigned homework to reinforce the ideas. The computer lab activities will be done using *BRB ArrayTools* (written by Richard Simon and Amy Peng Lam, http://linus.nci.nih.gov/BRB-ArrayTools.html), a software program written for analyzing microarrays and free for non-commercial use. The students will work with data that is freely available (or possibly the data they processed in the analyzing images module). We work through different normalization techniques as well as creating graphical displays of the data which help to both understand the qualities of the data and communicate the results to biologists.

One particularly useful plot ("MA-plot") describes the relationship between relative gene activity (denoted "Median-centered M") and absolute total signal (denoted "A"). We would hope that the amount of total signal for a given gene would not be related to the relative signal (e.g., how highly expressed is a cancerous sample *relative* to a healthy sample). However, we often see

that there are artifacts due to total amount of signal on the microarray. The plots in Figure 1 give two examples; the first where the relative signal is not dependent on the absolute signal, and the second where it is. The line is a smoothing spline. The first plot also indicates, not surprisingly, that the relative signal at very low absolute signals is quite variable.

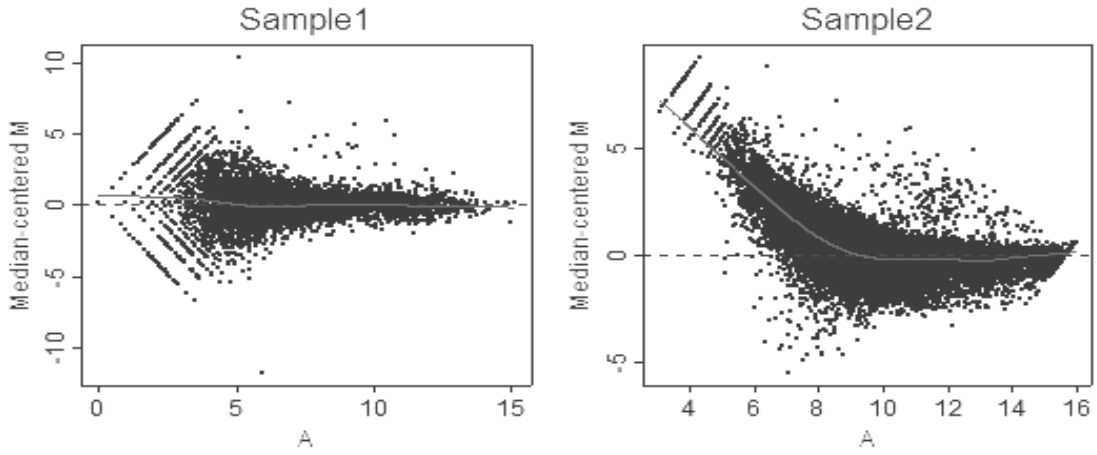


Figure 1: MA – Plot for two different microarray samples. The left hand panel gives a sample that is not dependent on the overall signal; the right hand panel gives a sample that is highly dependent on the overall signal.

*Class Comparison and Class Prediction*

The last module will be the most comfortable for statisticians. The techniques are a combination of established statistical methodologies with a new twist (for example $p > n$) and novel techniques developed specifically to address issues with microarray analyses. Class prediction differs from class comparison in that the former builds models that can be used to predict class membership, and the later answers the question of whether the classes are significantly different.

Some class comparison methods include a review or introduction (depending on the background of the students) of *t*-tests and *F*-tests. Along with discussing their usefulness, however, we will cover ideas of multiple comparisons and permutation tests. Ideas of multiple comparisons are extremely important in microarray analyses because of the large number of tests of significance usually performed. We plan to have the students analyze data sets with varying numbers of microarrays to drive home the value of replication in achieving significance. Permutation tests resolve issues of distributional assumption that are not often valid with these data.

We will also introduce a test called Significance Analysis for Microarrays (SAM) (Tusher *et al.*, 2001) which builds on *t*-tests by (a) using a slightly more robust test statistics (though very similar to the *t*-test statistic), and (b) using permutations / False Discovery Rate (FDR) instead of the *t*-distribution / level of significance (α) to determine significance. In Figure 2 you see a scatterplot of observed test statistic vs. expected test statistic (under permutations). The dotted lines represent a cutoff of 0.1% false discoveries. In our plot, we have 1303 genes, so, on average, we will find 1.3 significant genes which are not, in fact, statistically different across groups. The dots above the dotted line represent genes that are over-expressed in the experimental group; the dots below the dotted line represent genes that are under-expressed in the experimental group. The large number of significant genes comes from comparing two groups that are genetically quite different.

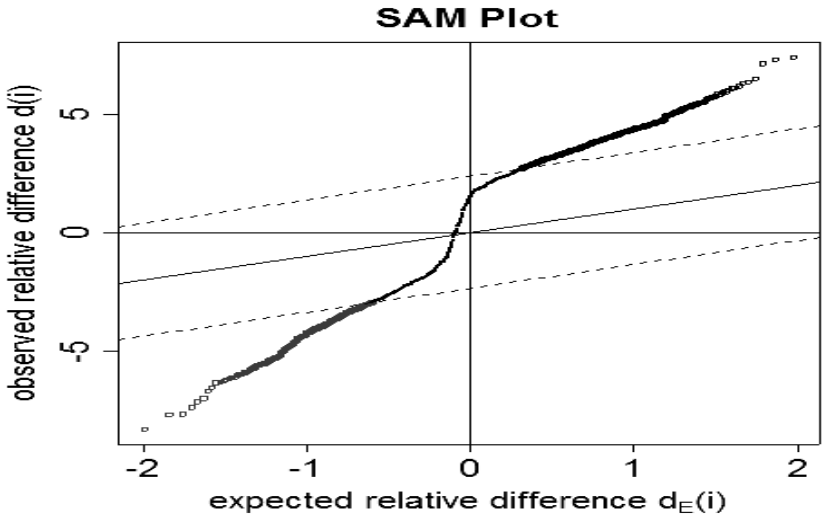


Figure 2: SAM plot identifying genes (above and below the dotted lines) which are significantly different across groups

For class prediction techniques we will discuss three techniques: nearest neighbor classification, compound covariate predictor (Radmacher *et al*., 2002), and a procedure related to SAM, Prediction Analysis for Microarrays (PAM) (Tibshirani *et al*., 2002). Again, we discuss multiple comparisons, and we bring up ideas of cross validation and sensitivity vs. specificity. For both class comparison and class prediction, the students will work with computer software (*ArrayTools*) to produce results and compare different techniques applied to the same data set.

EXAMPLE MODULE (for *Class Comparison and Class Prediction*)

- Reading:
  - Radmacher, M., McShane, L., and Simon, R. (2002). A paradigm for class prediction using gene expression profiles, *Journal of Computational Biology*, 9, 505-511.
  - Hesterberg, T., Moore, D., Monaghan, S., Clipson, A., and Epstein, R. (2005). *Bootstrap Methods and Permutation Tests* (2nd edition). New York: W. H. Freeman.
  - Draghici, S. (2003). *Data Analysis Tools for DNA Microarrays*. London: Chapman and Hall, chapters 6 and 9.
- Lecture Topics:
  - Class Comparison (*t*-tests, *F*-tests(ANOVA), permutation tests, significance analysis for microarrays, multiple comparisons)
  - Class Prediction (prediction analysis for microarrays, nearest neighbor classification, compound covariate predictor, cross validation)
- Lab:
  - Using *ArrayTools*, apply the class comparison and class prediction techniques to publicly available data
- Sample Homework Problems:
  - Explain the difference between a supervised and an unsupervised analysis.
  - Write out the algorithm for 10-fold cross validation. How would your algorithm change if you used leave-one-out cross validation?
  - Why is it important to cross validate to assess model validity?
  - Under what circumstances (e.g., hypotheses and/or assumptions) would you use each of the class comparison techniques we've discussed?
  - Under what circumstances (e.g., hypotheses and/or assumptions) would you use each of the class prediction techniques we've discussed?
  - Describe the differences and similarities in class comparison and class prediction techniques.
  - Explain why having more genes than samples can be a problem.

## CONCLUSION

We have argued that there is a great need (for both statisticians and biologists) of undergraduate courses and ideas in the field of bioinformatics. Biologists seem to be embracing bioinformatics at the undergraduate level, and we believe that statisticians can and should do the same thing. We have introduced a series of course modules that could be used with undergraduates in a standard introductory statistics course, an introductory biostatistics course, a biostatistics seminar, or as a course on their own. Though there are still spaces in the above course projects to be filled in, we hope that we have provided enough momentum to convince you that (a) these types of topics are essential when training the next set of scientists, and (b) you can introduce pieces of bioinformatics easily into a statistics curriculum.

## ACKNOWLEDGEMENTS

This work supported by a grant from the Howard Hughes Medical Institute to Pomona College, an NIH-AREA grant (#1 R15 AG021907-01A1) to JH and LH, and an NSF MRI grant (#0318944) to Pomona College.

## REFERENCES

Bialek, W. and Botstein, D. (2004). Introductory science and mathematics education for 21st-century biology. *Science*, 303, 788-790.

Brewster, J., Beason, K. B., Eckdahl, T., and Evans, I. (2004). The microarray revolution. *Biochemistry and Molecular Biology Education*, 32, 217-227.

Campbell, A. M. (2002). Meeting report: Genomics in the undergraduate curriculum – Rocket science or basic science? *Cell Biology Education*, 1, 70-72.

DeRisi, J., Iyer, V., and Brown, P. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278, 680-686.

Draghici, S. (2003). *Data Analysis Tools for DNA Microarrays*. London: Chapman and Hall.

Hack, C. and Kendall, G. (2005). Bioinformatics: Current practice and future challenges for life science education. *Biochemistry and Molecular Biology Education*, 33, 82-85.

Hesterberg, T., Moore, D., Monaghan, S., Clipson, A., and Epstein, R. (2005). *Bootstrap Methods and Permutation Tests* (2nd edition). New York: W. H. Freeman.

Heyer, L., Moskowitz, D., Abele, J., Karnik, P., Choi, D., Campbell, M.A., Oldham, E., and Akin, B. (2005). MAGIC Tool: Integrated microarray data analysis. *Bioinformatics*, 21, 2114-2115.

Honts, J. (2003). Evolving strategies for the incorporation of bioinformatics within undergraduate cell biology curriculum. *Cell Biology Education*, 2, 233-247.

Radmacher, M., McShane, L., and Simon, R. (2002) A paradigm for class prediction using gene expression profiles. *Journal of Computational Biology*, 9, 505-511.

Schuchhardt, J., Beule, D., Malik, A., Wolski, E., Eickhoff, H., Lehrach, H., and Herzel, H. (2000). Normalization strategies for cDNA microarrays. *Nucleic Acids Research*, 28, e47.

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, C. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99, 6567-6572.

Tusher, V., Tibshirani, R., and Chu, C. (2001). Significance analysis of microarrays applied to ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98, 5116-5121.

Yang, Y., Dudoit, S., Luu P., Lin, D., Peng, V., Ngai, J., and Speed, T. (2002). Normalization for cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, 4, e15.