# STATISTICS COLLOQUIUM

MONDAY, DECEMBER 07, 2009
TALK: 4:00 PM — SCIENCE CENTER, ROOM 309
RECEPTION: 5:15 PM — SCIENCE CENTER, 7TH FLOOR

## "Cross-study differential gene expression"

**Giovanni Parmigiani**
**Dana Farber Cancer Institute**
**Harvard University**

ABSTRACT

In this lecture I will present statistical issues associated with combining microarray data across studies. I will cover exploratory data analysis approaches to visualize the extent of agreement of gene expression measurements across studies, drawing from examples molecular profiling of cancer. I will then discuss the role of hierarchical Bayesian models in constructing useful rules to shrink across both genes and studies, and to classify genes based on the patterns of concordance across studies. I will describe in detail a model we call XDE, and evaluate its performance in a comprehensive fashion, using both artificial data, and a "split-sample" validation approach that provides an agnostic assessment of the model's behavior not only under the null hypothesis but also under realistic alternatives. Compared to a more direct combination of t- or SAM-statistics, the 1 − AUC values for the Bayesian model is roughly half of the corresponding values for the t- and SAM-statistics. In small studies, XDE generally outperforms other methods when evaluated by AUC, FDR, and MDR across a range of simulation parameters, and this di erence diminishes for larger sample sizes in the individual studies. Finally, I will illustrate our model using four breast cancer studies employing di erent technologies (cDNA and A ymetrix) to estimate di erential expression in estrogen receptor positive tumors versus negative ones. Software and data for reproducing our analysis are publicly available. A technical report can be obtained from http://www.bepress.com/jhubiostat/paper158/