

Statistical methods for analysis of genomics data
Fall 2008

Hour/location: Tue and Thur 12-1:30pm, Blockley Hall 940

Instructor: Professor Hongzhe Li

Office: Blockley Hall 920

Office hour: Make appointment by email (hongzhe@mail.med.upenn.edu)

Requirements: Good understanding of statistics and probability. Ability to program in R (<http://cran.r-project.org>).

Materials to be Covered: This course focuses on statistical methods for analysis of genomic data, including (mainly) Affymetrix microarray gene expression, array CGH data and Chip-chip data. I will introduce in details some of the most commonly used statistical methods for analyzing different types of genomic data. I will cover the following materials:

- (1) Pre-processing of Affymetrix gene expression data, gene expression index by dChip, data normalization
- (2) Methods for identifying differentially expressed genes, multiple testing, and SAM and empirical Bayes methods.
- (3) Sample classification by nearest shrunken centroids.
- (4) Regularized regression for linking gene expression data to clinical phenotypes, ridge regression and lasso.
- (5) Structured analysis of microarray data, gene ontology and functions, pathways, and Hidden Markov random field models, gene set enrichment analysis.
- (6) Identification of transcription factors, motifs, linking sequences data to gene expression data.
- (7) Analysis of array CGH data, detection of multiple change points, Finite states Markov chain models.
- (8) Analysis of ChIP-chip data.
- (9) (time permits): Bayes networks, graphical models and MCMC learning.

I will present methods together with real examples.

Homework (those who have registered for credits):

(1) Type 1 (30%): for a given topic summarize the course materials using Latex, add more relevant materials and review of literatures.

(2) Type 2 (60%): data analysis projects. Learn R and Bioconductor. You can only learn these techniques by doing your own data analysis. About six data analysis projects.

<http://cran.r-project.org>

<http://www.bioconductor.org>

(3) Attending lectures (10%): participating lectures.

Datasets for Homework projects:

(1). Affymetrix data normalization and pre-processing, identification of differentially expressed genes

Dr. Thomas Cappola's gene expression study of detection of cardiac allograft rejection and response to immunosuppressive therapy with peripheral blood gene expression.

<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE5967>

(2). High-dimension regression, lasso and Lars, classification, nearest shrunken centroids. Glioblastoma survival time from UCLA.

<http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/ASPMgene/>

(3). Array CGH data

171 primary breast tumors from Nottingham City Hospital and 49 breast cancer cell-lines were hybridised with male DNA as reference.

<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE8757>

(4) More data sets to add later.