

Statistical analysis of gene expression data

Presenters:

Terry Speed, UC Berkeley and WEHI

Jean Yee Hwa Yang, University of Sydney

Benjamin Bolstad, UC Berkeley

James Wettenhall, WEHI.

Abstract

This short course will be an introduction to the statistical analysis of gene expression data from cDNA, long-oligonucleotide and high-density short oligonucleotide microarrays, and serial analysis of gene expression (SAGE) data. We will begin with a short biological introduction, and then turn to brief descriptions of representative microarray platforms and the SAGE process.

Preprocessing and QC issues will be addressed, as will issues related to probe or tag identity and data visualization. Following this introductory material, attention will turn to the gene expression measurements themselves, and the identification of differentially expressed (d.e.) genes using replicated data. With this background, it is possible to discuss design issues underlying the search for d.e. genes, including the different types of replicates, and the dependencies inherent in microarray gene expression data. The next topic to be discussed is the annotation of d.e. genes using the Gene Ontology, and the search for d.e. sets of genes. The remainder of the short course will be devoted to more specialized topics including classification, clustering, and the analysis of time course data.

Experience with statistical methods in data analysis is essential, but no previous knowledge of gene expression data is required. The course will include the opportunity to apply statistical methods to several data sets that will be provided.

Participants are expected to bring their own laptops and to have downloaded the specified software and datasets prior to arriving in Minneapolis, or to purchase CDs with the same on site. The course will be a series of 8 30-45 minute lectures followed by 45-60 minute lab sessions, with lengthy breaks that we hope will keep the pace leisurely.

SCHEDULE

Day 1

8.30-10.00: Introduction to gene expression studies.

Microarray technology and platforms. Pre-processing for spotted two-channel arrays: image analysis, data exploration, within-array normalization, between-array normalization.

Examine GenePix and SPOT data using limmaGUI. Setup targets and spot-types files. Input data. Examine effects of different background corrections on MA-plot. Command-line data entry and normalization using limma.

10.00-10.30: Break

10:30-12:00: Practical aspects of assessing two-color spotted array quality.

12.00-1.30: Lunch.

1.30-3.00: High-density short oligonucleotide (Affymetrix GeneChip®) arrays.
Pre-processing for Affy arrays: normalization, data exploration.
QA/QC for Affy chip data.
Introduction to the affy package. CEL files, CDF package, AffyBatch and exprSet objects.
Image plots of robust regression weights.

3.00-3.30: Break

3.30-5.00: Differential expression.
Linear models with Affymetrix and two-channel data.
Analysis of time-series expression data.
Differential expression and linear modelling using limma.

Day 2

8.30-10.00 Design of microarray experiments.
General considerations for all platforms. Special issues for two-channel arrays.
Varieties of replication. Pooling.
More lab work on differential expression and time-series data.

10.00-10.30: Break

10.30-12.00: Clustering and classification.
Basic approaches including gene selection and cross-validation.
Illustrated with breast cancer gene expression data. The lecture will focus on two-sample classification, but more complex scenarios will also be discussed.

12.00-1.30: Lunch

1.30-3.00: Genes and set of gene.
Probe and tag identity and reliability. Gene annotation.
Gene Ontology. Pathways. Statistical analyses for gene sets.

3.00-3.30: Break

3.30-5.00: Other technologies for measuring gene expression.
Introduction to other gene expression platforms
 a) Quantitative real-time PCR;
 b) Serial Analysis of Gene Expression;
Recent trends and technologies.
Assessing and comparing methods.