Microarray Data Analysis in Functional Genomics

PJ "Asa" Wirapati

Bioinformatics
Core Facility

Swiss Institute
of Bioinformatics

PhD Course in Functional Genomics, 22 January 2009

Outline

- Introduction
- Structure of microarray data
- Preprocessing
- Exploratory data analysis
- Statistical analysis: correlation, association and prediction
- Integrative analysis of multiple datasets

Omics Biology

DNA	genomics, genetics	microsatellite, SNP arrays, UHTS
RNA	transcriptomics	expression arrays, RT-PCR, UHTS
Protein	proteomics	mass spec, 2D-PAGE, ...
Everything else	phenomics, metabolomics,...	...

Functional genomics is concerned with how variations and changes at the molecular level related to phenotypes, in contrast to "static" characterization/cataloging of sequences and structures.

Variations, *changes* $\Rightarrow$ many samples/cases/subjects

Generated by either controlled experimental conditions or naturally present variability

Data Analysis Challenges

Large number of *variables* (elements of the omics: genes, transcripts, proteins, . . .) present some challenges:

Informatics Storage, handling and computation of large datasets

Statistics Large number of variables, but small (or even smaller than usual) sample size: overfitting, false positives, confounding/bias

Presentation How to visualize/summarize in a compact form

Interpretation What the analysis output means, biologically?

Results Integration Meta-analysis, reproducibility, validation (between independent studies of the same design/approach, and between modes of analysis (DNA/RNA/protein/etc).

Gene Expression Microarray

We'll focus on gene expression microarray as an example of omics data analysis

Similar data analysis principles apply to others.

- Brief history, $\approx$10 years (still young but maturing...)
- Wide range of applications (experimental biology, population-based medical sciences, evolution, ...)
- Extensive datasets are publicly available (often annotated by the phenotype, experimental conditions, clinical data, ...)

Why study microarray data analysis?

1. You want to use the technology in your research
2. You want to use data produce by others to support your research
3. You need to review/assess results from your peers based on microarrays

Typical steps in microarray projects

- Experimental design: what to compare? sample size?
- Generate biological materials: collect tissues, conduct experiments
- Laboratory analysis: RNA extraction, amplification, hybridization and scanning
- Data Analysis
    1. Preprocessing: spot quantitation, normalization, transformation, summarization, QC
    2. Exploratory Data Analysis: PCA, clustering
    3. Downstream analysis: differential expression, prediction/classification, coexpression network, ...
    4. Integrative analysis: compare/combine with other microarray dataset, match with "gene sets" knowledge (ontology, pathways, signatures)

(Re)Using Public Microarray Datasets

- Define study (dataset) inclusion criteria, survey availability
- Data download, curation, clean up of annotations (experimental design, clinical data, ...)
- Preprocessing: expression data (re)normalization, probe (re)mapping and matching
- Downstream analysis: differential expression, prediction, coexpression, pathway/ontology analysis, ...

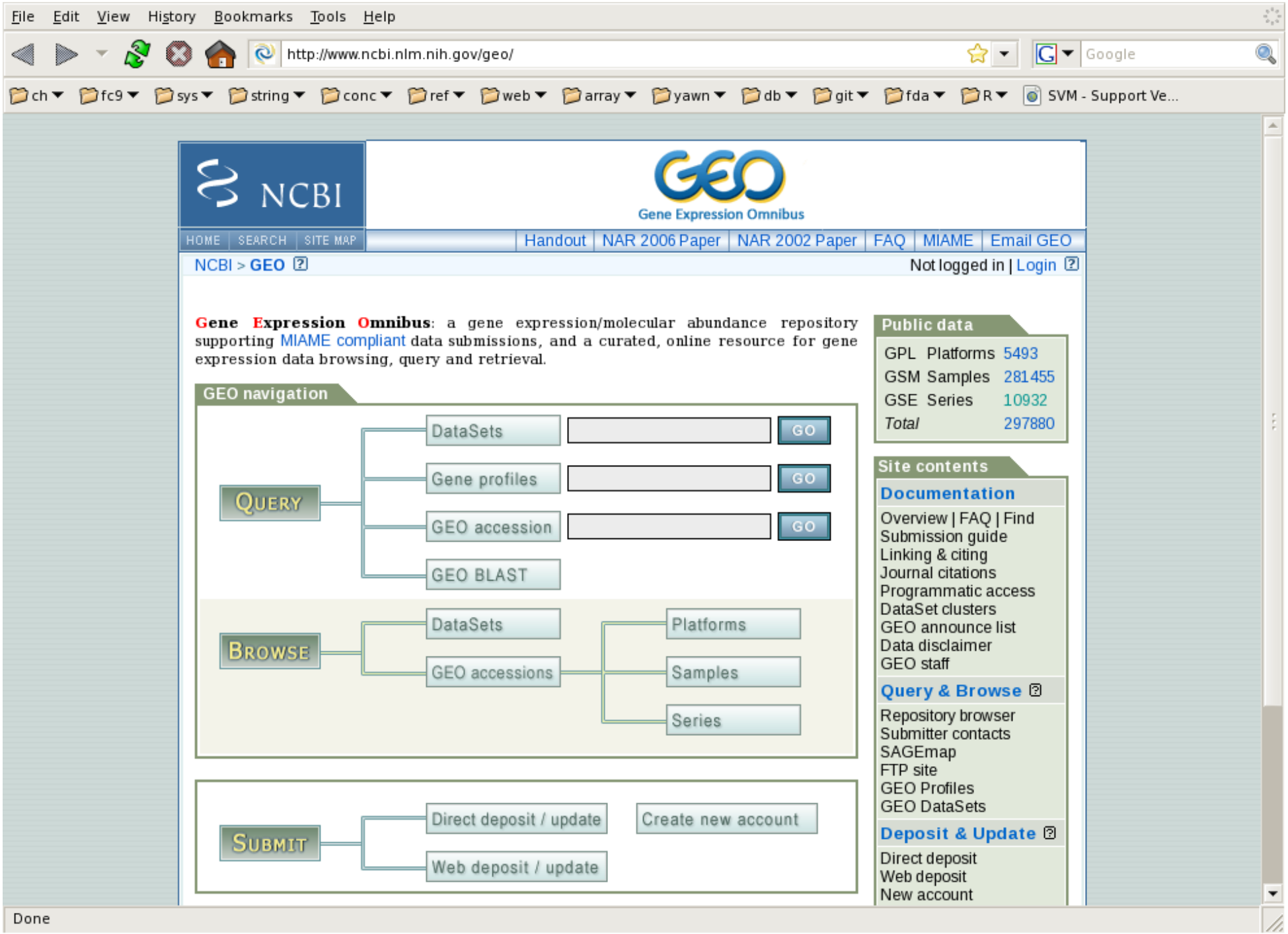
  Similar to single-dataset analysis, *but* we need to account for between-study heterogeneity in microarray platforms, study designs, systematic biases, other methodological differences

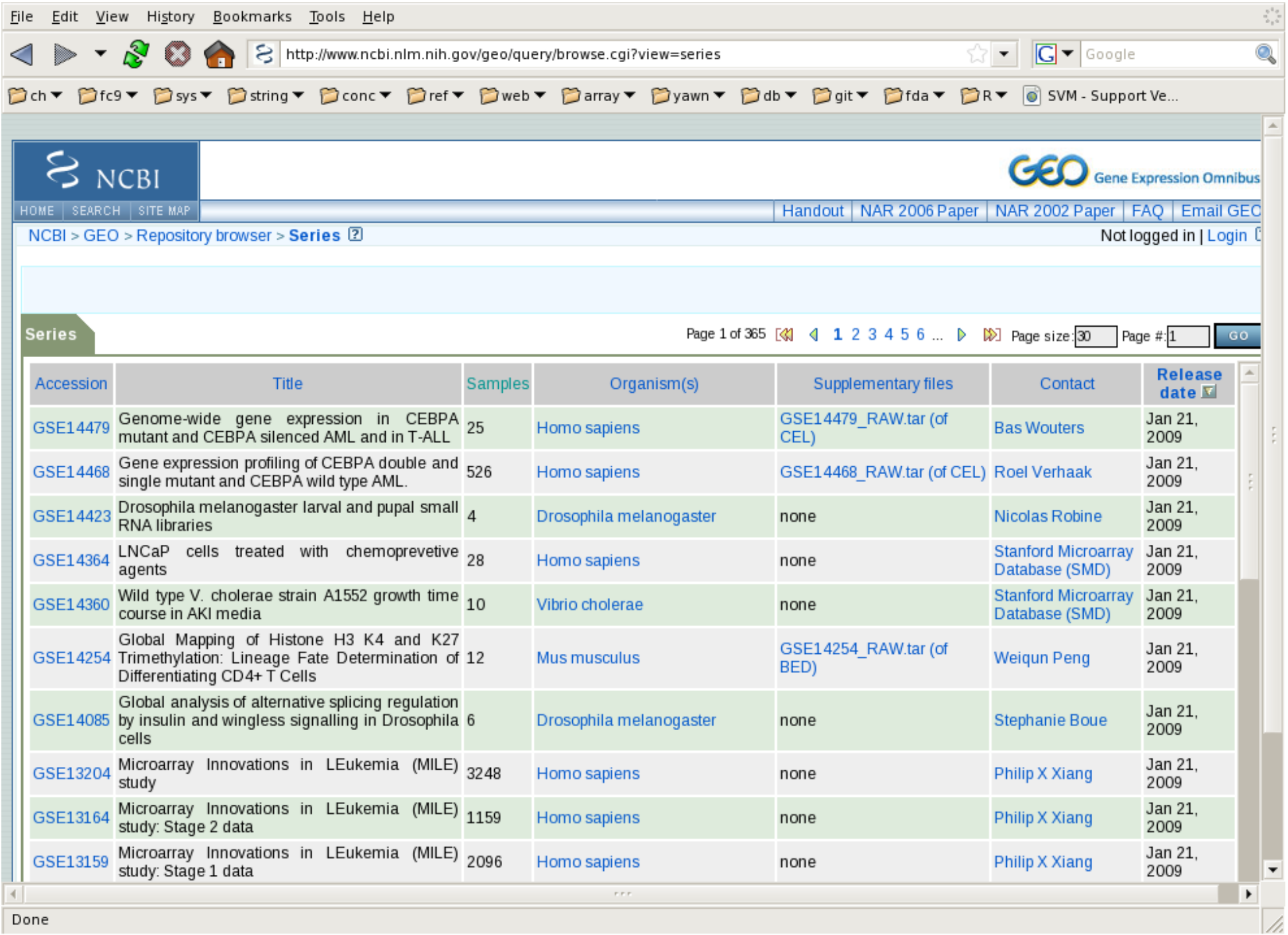
Outline

- Introduction
- Structure of microarray data
- Preprocessing
- Exploratory data analysis
- Statistical analysis: correlation, association and prediction
- Integrative analysis of multiple datasets

Structure of Microarray Data

- A bundle ("series", "dataset") of multiple arrays from the same experiment or study
- Probe definition (GPLxxxx in GEO), depends on the platform/model (typically just one platform per study, but can be more)
- Expression data for each sample (GSMxxxx in GEO), can be "raw" data (CEL, GPR, etc.) or processed/normalized.
- Sample annotation: experimental design ("design matrix": treatment, condition, batches), or observational variables (e.g. clinical data of patient characteristics)
- Database or compendium of studies/datasets, such as GEO itself, or more specialized ones, such as Oncomine for cancer studies


File Edit View History Bookmarks Tools Help
http://www.ncbi.nlm.nih.gov/geo/
Google
ch fc9 sys string conc ref web array yawn db git fda R SVM - Support Ve...
NCBI
Gene Expression Omnibus
HOME SEARCH SITE MAP
Handout NAR 2006 Paper NAR 2002 Paper FAQ MIAME Email GEO
NCBI > GEO
Not logged in | Login
Gene Expression Omnibus: a gene expression/molecular abundance repository supporting MIAME compliant data submissions, and a curated, online resource for gene expression data browsing, query and retrieval.
GEO navigation
QUERY
DataSets
Gene profiles
GEO accession
GEO BLAST
GO
BROWSE
DataSets
GEO accessions
Platforms
Samples
Series
SUBMIT
Direct deposit / update
Web deposit / update
Create new account
Public data
GPL Platforms 5493
GSM Samples 281455
GSE Series 10932
Total 297880
Site contents
Documentation
Overview | FAQ | Find
Submission guide
Linking & citing
Journal citations
Programmatic access
DataSet clusters
GEO announce list
Data disclaimer
GEO staff
Query & Browse
Repository browser
Submitter contacts
SAGEmap
FTP site
GEO Profiles
GEO DataSets
Deposit & Update
Direct deposit
Web deposit
New account
Done


File Edit View History Bookmarks Tools Help
http://www.ncbi.nlm.nih.gov/geo/query/browse.cgi?view=series
Google
ch fc9 sys string conc ref web array yawn db git fda R SVM - Support Ve...
NCBI
HOME SEARCH SITE MAP
GEO Gene Expression Omnibus
Handout NAR 2006 Paper NAR 2002 Paper FAQ Email GEO
NCBI > GEO > Repository browser > Series
Not logged in | Login
Series
Page 1 of 365 1 2 3 4 5 6 ... Page size: 30 Page #: 1 GO
Accession Title Samples Organism(s) Supplementary files Contact Release date
GSE14479 Genome-wide gene expression in CEBPA mutant and CEBPA silenced AML and in T-ALL 25 Homo sapiens GSE14479_RAW.tar (of CEL) Bas Wouters Jan 21, 2009
GSE14468 Gene expression profiling of CEBPA double and single mutant and CEBPA wild type AML. 526 Homo sapiens GSE14468_RAW.tar (of CEL) Roel Verhaak Jan 21, 2009
GSE14423 Drosophila melanogaster larval and pupal small RNA libraries 4 Drosophila melanogaster none Nicolas Robine Jan 21, 2009
GSE14364 LNCaP cells treated with chemoprevetive agents 28 Homo sapiens none Stanford Microarray Database (SMD) Jan 21, 2009
GSE14360 Wild type V. cholerae strain A1552 growth time course in AKI media 10 Vibrio cholerae none Stanford Microarray Database (SMD) Jan 21, 2009
GSE14254 Global Mapping of Histone H3 K4 and K27 Trimethylation: Lineage Fate Determination of Differentiating CD4+ T Cells 12 Mus musculus GSE14254_RAW.tar (of BED) Weiqun Peng Jan 21, 2009
GSE14085 Global analysis of alternative splicing regulation by insulin and wingless signalling in Drosophila cells 6 Drosophila melanogaster none Stephanie Boue Jan 21, 2009
GSE13204 Microarray Innovations in LEukemia (MILE) study 3248 Homo sapiens none Philip X Xiang Jan 21, 2009
GSE13164 Microarray Innovations in LEukemia (MILE) study: Stage 2 data 1159 Homo sapiens none Philip X Xiang Jan 21, 2009
GSE13159 Microarray Innovations in LEukemia (MILE) study: Stage 1 data 2096 Homo sapiens none Philip X Xiang Jan 21, 2009
Done

[DEMO]

- Examples of a series from GEO (downloaded files): GSE*, GPL*, GSM*, raw files
- Example of clinical sample annotations (original form and compiled)
- "Ideal" data format for statistical analysis: table of samples versus genes (and other variables)
- Querying specific genes and clinical variables
- Exploratory analysis: scatter plots, boxplots, histograms, heatmaps and clustering, principal component analysis
- An example paper (Urban *et al.* 2006 *J Clin Oncol* 24:4245)

Outline

- Introduction
- Structure of microarray data
- Preprocessing
- Exploratory data analysis
- Statistical analysis: correlation, association and prediction
- Integrative analysis of multiple datasets

Generalized Linear Models

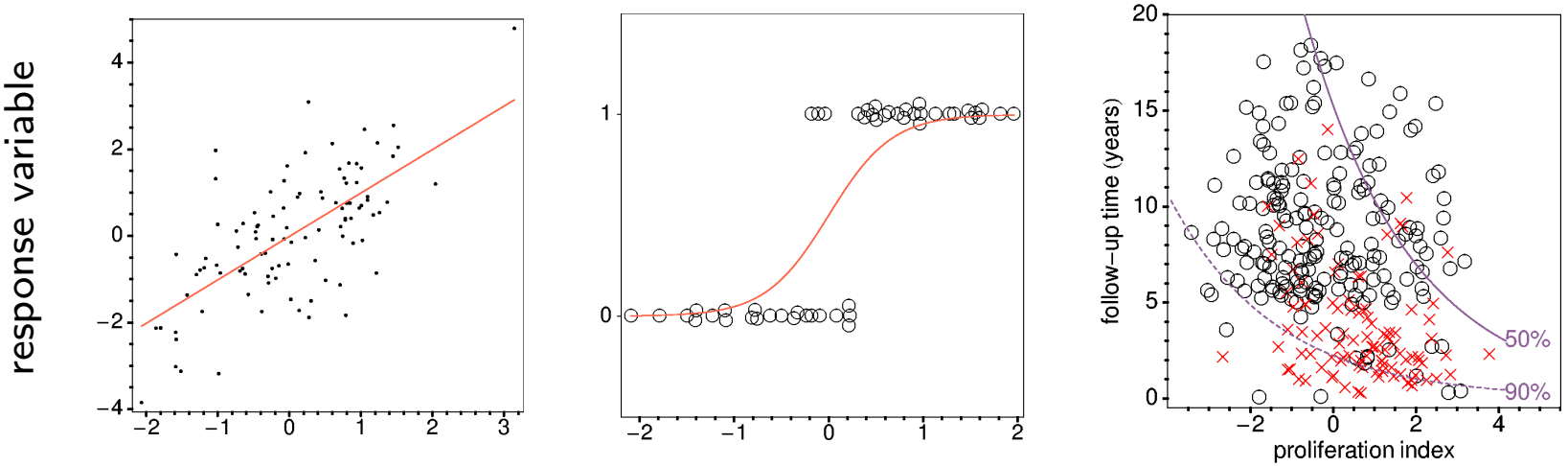
Relate response/outcome and explanatory variables

Different types of response variables
$\Rightarrow$ use appropriate error models and link functions

continuous
(Gaussian/Normal)

binary
(Binomial)

censored (survival)
(Weibull or Cox)



linear combination of explanatory variables

Multiple explanatory factors (those of interests, as well as nuisance or baseline factors) are considered simultaneously

Basic concept of regression analysis

[DEMO]

Linear models between genes, and clinical variables

Adjustment and multiple regression ("multivariate" analysis)

Survival analysis using Cox regression

Regression models in genome-wide data

- Gene-by-gene scanning: differential expression analysis (t-test or ANOVA) vs experimental conditions, searching for genes that can explain patient survival,. . .

  Results are ranked; top gene list usually called a "signature"

- Correlation between all pairs of genes: coexpression analysis

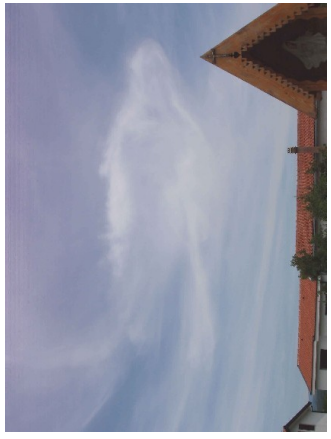
  Results are organized into networks, hierarchical clusters, etc.

- More complex network: graphical models (conditional independent graph)

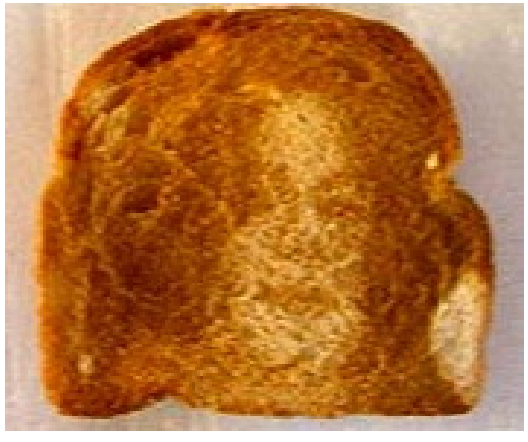
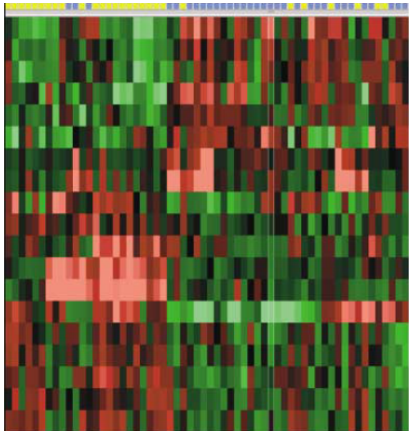
  Results are parsimonious networks (can distinguish between direct and indirect correlations), if time dimension are present in the data, may even infer *causality*

Issues: computation, multiple testing (a.k.a. multiple selection)

Multiple Selection Problem





When searching through large number of possibilities, interesting patterns may occur by chance

The larger the number of genes, the higher the false positive rate.

This can be mitigated by having larger sample size.

The usefulness of genome-wide data is measured by the ratio of number of variables to sample size.

[DEMO]

Prediction

When Y is associated/correlated with X, then if we observe only X, we can predict Y (with some uncertainty).

Model complexity is important for performance in future data. If it's too complex (roughly, too many explanatory variables or terms) it will fit current data better, but fail to predict future data.

Also subject to overfitting in genome-wide context, due to variable selection. This is the main cause of overfitting.

Ideally, we would like to validate using independent samples.

If sample size is not enough, assessment of future performance maybe done using cross-validation. (Not perfect!)

Further topics

Analysis of multiple gene expression datasets:
`http://www.isrec.isb-sib.ch/~pwirapat/embnet/`

Other BCF/SIB teaching materials:
`http://bcf.isb-sib.ch/Teaching.html`

Upcoming course:
`http://bcf.isb-sib.ch/teaching/2009-microarrays/`