Initial Case Study

(Gene Expression in Parkinson’s Disease)

Utah State University – Spring 2014
STAT 5570: Statistical Bioinformatics
Notes 1.0

Gene Expression Crashcourse

- Technology takes a snap-shot of the activity of all genes at once; estimate each gene’s expression level in each sample
- Want to identify genes that behave differently in one group (treatment, diseased) compared to another (control, healthy)
- Many statistical methods proposed
- Goal: If know which genes affect disease progression, maybe develop drug to stop their activity.
(Or, identify predictive / prognostic genes.)

Typical Statistical Analysis Process

- Obtain an array image for each subject, and convert image to quantitative measures of genes’ expression - or – sequence mRNA fragments and map to genes (to quantify genes’ expression)
- For each gene –
  - compare expression level between treatment conditions (healthy vs. diseased, for example)
  - determine whether gene’s expression values predict clinical outcome
- For groups of genes –
  - find combinations (profiles or signatures) that significantly predict clinical outcome
  - find similarities (molecular function, e.g.) among significant genes
- “Validate” genes – qRT-PCR, for example.

(Similar process for newer technologies)

Example: Parkinson’s Disease

- Scherzer et al., Jan. 2007 PNAS

- Whole blood samples from 105 subjects
  - 50 Parkinson’s disease (PD) patients
  - 23 Alzheimer’s disease (AD) patients
  - 10 other neurodegenerative (ND) patients
    - progressive supranuclear palsy (PSP)
    - multiple system atrophy (MSA)
    - corticobasal degeneration (CBD)
    - essential tremor (ET)
  - 22 healthy controls (H)

- Goal:
  - find a set of genes (out of 22,000+) whose expression levels (from a laboratory blood test) can reliably predict PD status

- Results:
  - identified 8 genes

A brief statistical view

- For gene k on array (or subject) i:

Estimate by sharing info. across genes

$$\underbrace{Y_{i,k}}_{\substack{\text{expression}\\\text{level}\\\text{(log scale)}}} = \beta_{k,0} + \underbrace{\beta_{k,1}}_{\substack{\text{treatment}\\\text{effect}\\\text{(gene-specific differential expression, DE)}}}\underbrace{T_i}_{\text{indicator (0/1) of "treatment" level}} + \varepsilon_{i,k}, \quad Var[\varepsilon_{i,k}] = \sigma_k^2$$

- What if there are more covariates than just “treatment”?

$$Y_{i,k} = \beta_{k,0} + \beta_{k,1}T_i + \beta_{k,sex}S_i + \varepsilon_{i,k}$$

- Analysis: usually some variant of ANOVA or t-test

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y$$

(inference often based on permutation tests)

A brief statistical view, continued:

- Another perspective:

$$p_i = P(T_i = 1) \longleftarrow \text{probability of disease status}$$

$$\log \frac{p_i}{1-p_i} = \alpha + \underbrace{\sum_{k=1}^{m} Y_{i,k}\beta_k}_{\text{cumulative effect of [possibly multiple] genes}} + \beta_{sex}S_i$$

- Find significant subset of genes k=1,…,m

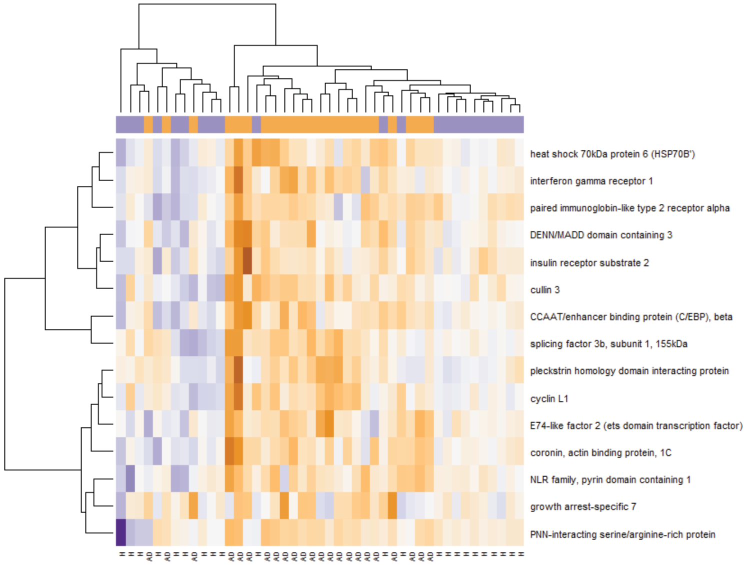
- Analysis: usually some variant of logistic regression (sharing info. across genes)

Example – Alzheimer’s Disease

- Partial data from Scherzer et al. (2007)

- Top 15 genes in predicting AD vs. H

  (but what do these genes have in common?)



[Expression value on scale: dark orange (low) to light (mid) to dark purple (high)]

Statistical Issues

- “Preprocessing” data (technology output to useful data)
  - Microarrays
  - Next-generation sequencing
  - Mass Spectrometry
- Distribution of data (& appropriate tests)
  - Continuous → Normal / Nonparametric
  - Count → Poisson / Negative Binomial
- Multiple hypothesis testing (individual & groups)
- Effective Communication
  - Visualization
  - Interactive Reports
  - “Characterization” of results → more statistical issues