

Statistical models and methods in case–control 5'-nuclease assays, for identification of differentially expressed genes. (with Schizophrenia application)

Rolf Sundberg, Stockholm Univ.
A. Castensson, E. Jazin, Uppsala Univ.

Situation

- Patients and controls (55 of each, =110)
- Several (brain) samples per individual (2)
- Put on plates with < 96 wells per plate
- Fluoresc. measurements of mRNA by Real Time PCR combined with TaqMan assay: One master-plate => many replica plates, one per gene

Statistical aspects

- Design: Balanced incomplete design on plates
- Basic model: MRANCOVA, i.e. multivariate nested random effects analysis of covariance model
- Inference:
 - (1) Reference genes for increased precision
 - (2) Prediction aspects
 - (3) Minor problems: plate effect estimation
left-censoring for low-expressing genes, outliers, non-constant variances, multiple testing, etc.

Modelling

- Basic \approx MRANCOVA model, for *controls*:
 $Y = \log(\text{fluoresc})$ vector (gene \Leftrightarrow comp. y)
 $y_{hij} = \mu + \alpha_h + \beta' u_{hi} + \gamma_{k(hij)} + \delta_{hi} + \epsilon_{hij}$
 h = stratum index (brain bank, sex),
 i = individuals within stratum h ,
 j = samples within individual,
 k = plate number allocation,
 u = individ. covariate (age, time post mort.)
Nested variance components from μ and α

For patients

- Either the same model with constant shift, or
- a random individual disease effect.

Test for absence of effect, and under significance, explore effect distribution (affected group, interactions)

Multivariate aspects

- Nested components Σ and σ are multivariate, i.e. represented by covariance matrices, dimension=#genes
- Correlations btw components were high in Σ and even higher in σ .
- Motivates use of unaffected *reference genes*, for statistical efficiency.
- Predict candidate gene values from ref-genes, adjusting for other covariates

For candidate genes

- With x like y , but for ref-gene, fit $E(y|x)$,

$$y_{hij} = \text{as before} + \sigma x_{hij},$$

or correspondingly for averages y_{hi} .

Note: parameters have new interpretations, and some are no longer needed in model

Prediction aspects

- Alternative interpretation of $E(y|x)$:
Predict candidate gene values from ref- genes, for each individual, adjusting for other factors.
- Predict patient values via model fitted to the unaffected controls, to explore non-constant disease effects
Varying disease effect => loss of power in standard two-sample tests

Plate effects and averaging

- Incomplete design motivates plate effect estimation within individuals, for statistical efficiency
- But regression on x ‘within individuals’ will be different from regression on x ‘between individuals’
- => sacrifice ‘within’ plate effect estimates, and average over samples from individual

Results

- Gain from use of reference genes:
Std error typically reduced by factor 2 – 3,
crucial for obtaining significant effects.
- 2 out of 16 genes were found significant,
see box-plots etc (not included)
- Their individual effects were correlated,
see scatter-plot (not included)

rolfs@math.su.se

www.math.su.se/~rolfs/Publications.html