Bayesian Analysis of Differential Gene Expression

Biostat Journal Club

Chuan Zhou

chuan.zhou@vanderbilt.edu

Department of Biostatistics
Vanderbilt University

Lewin et al., 2006

- Goals and Data
- Model specification
- Model checking
- Integrated vs. non-integrated
- Application to mouse data
- Discussion

Goals

- Difficulties with differential gene expression analysis
  - Fold change not comparable between genes
  - Small numbers of replicates
  - Multiple sources of variability
- Proposed approach: model biological variability, systematic array effects and differential expression simultaneously
- A fully Bayesian model, take into account uncertainty in parameter estimates
- Decision rules based on posterior distributions
- Use Bayesian FDR estimate to select cutoff point

Data

- Three wild-type mice and three mice with Cd36 gene removed
- Hybridized to Affymetrix U74A, U74B and U74C chips, total 18 microarrays
- U74A chip data: three repeated measurements of two conditions for each of the 12,488 genes
- Data pre-processed by Affymetrix MAS 5.0 software
- Skewed data – use log transformation

Bayesian hierarchical model

- Intuition

  Observed = Gene effect + Differential effect + Array effect

- Notations: y_{gsr} = log-expression of gene g, condition $s = 1, 2$, and replicate r.

- An ANOVA model

$$y_{g1r} \sim N(\alpha_g - \frac{1}{2}\delta_g + \beta_{g1r}, \sigma^2_{g1})$$

$$y_{g2r} \sim N(\alpha_g + \frac{1}{2}\delta_g + \beta_{g2r}, \sigma^2_{g2})$$

- Identifiability constraint: $\bar{\beta}_{gs\cdot} = 0, \forall g, s$

The Model

- Model array effect as a function of the expression level

$$\beta_{gsr} = f_{sr}(\alpha_g).$$

For example, a quadratic spline

$$\begin{aligned}\beta_{gsr} =& b^{(0)}_{sr0} + b^{(1)}_{sr0}(\alpha_g - a_0) + b^{(2)}_{sr0}(\alpha_g - a_0)^2 \\ &+ \sum_{k=1}^{K} b^{(2)}_{sr0}(\alpha_g - a_0)^2 I[\alpha_g \geq a_{srk}],\end{aligned}$$

- Assume variances are exchangeable within condition

$$\sigma^2_{gs} \sim \text{logNorm}(\mu_s, \eta^2_s).$$

- Gene effects α_g and knots a_{srk} are uniform between (a_0, a_{K+1}) which are pre-specified bounds.

The Model

- Confounding: normalizing across replicates and conditions in a preprocessing step implicitly assume there is no differential effects
- Implementation: MCMC using WinBUGS, code available online, 74,922 data points, 1000 iterations took approximately 3 hours on a dual processor 2.4 GHz machine
- Rules for selecting genes

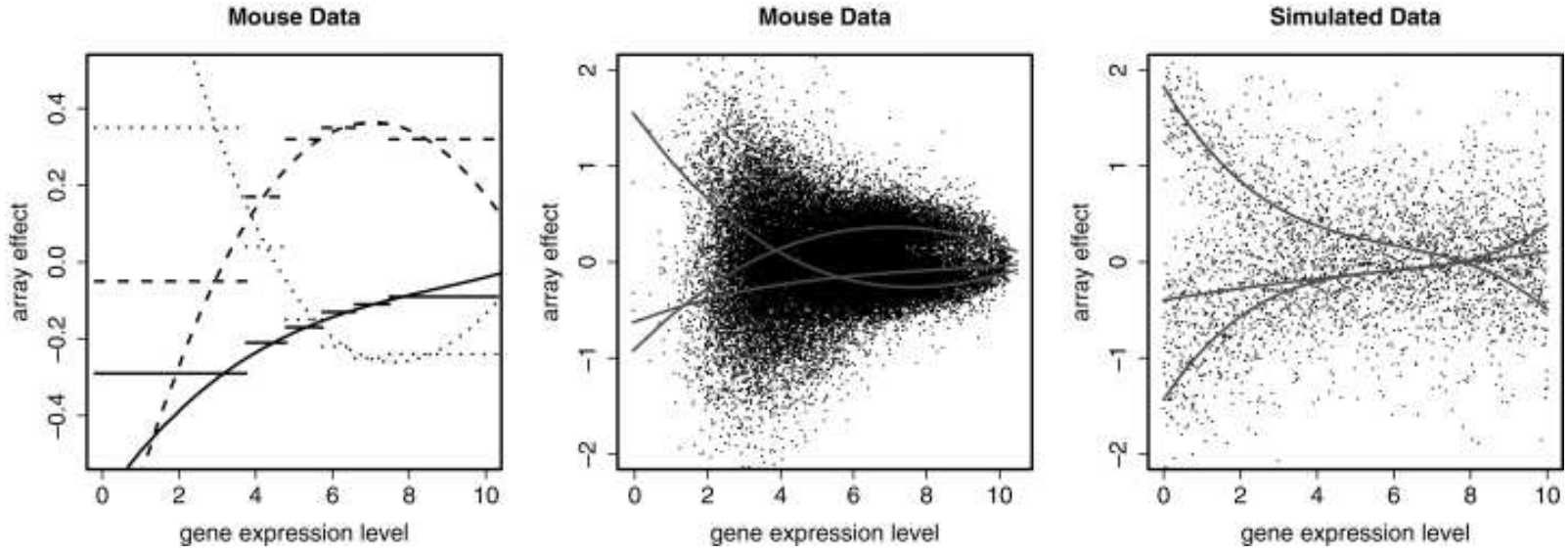
$$p_g \equiv P(|\delta_g| > \delta_{cut} \text{ and } \alpha_g > \alpha_{cut} \,|\, \text{data})$$

Genes are selected if $p_g \geq p_{cut}$. The choice of p_{cut} is determined by the evaluation of FDR.

Model checking

- Use biological replicate data
- Exploratory analysis of array effects
  - Divide genes into J groups with similar expression levels
  - $y_{g1r} \sim N(\alpha_g + \beta_{j1r}, \sigma^2_{j1})$
  - Try various functional forms $f(\cdot)$ for array effects
  - determined by DIC
- Clearly see non-linear relationship between gene effects and array effects

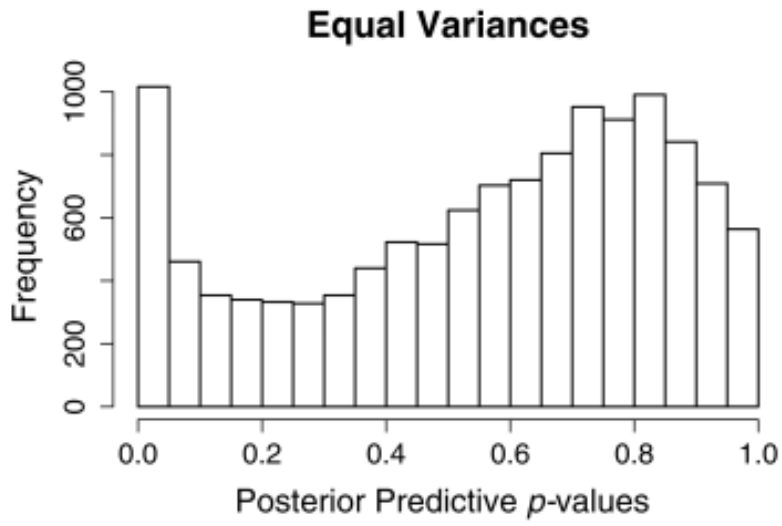
Non-linear array effects

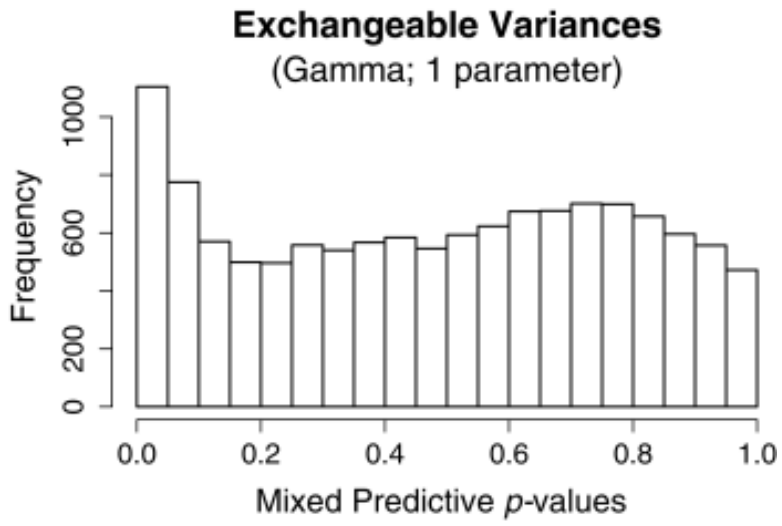


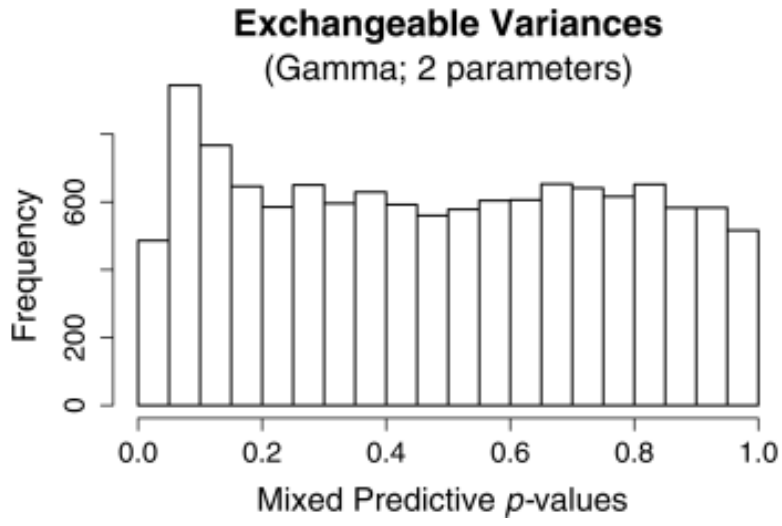
Predictive checks on prior for variance

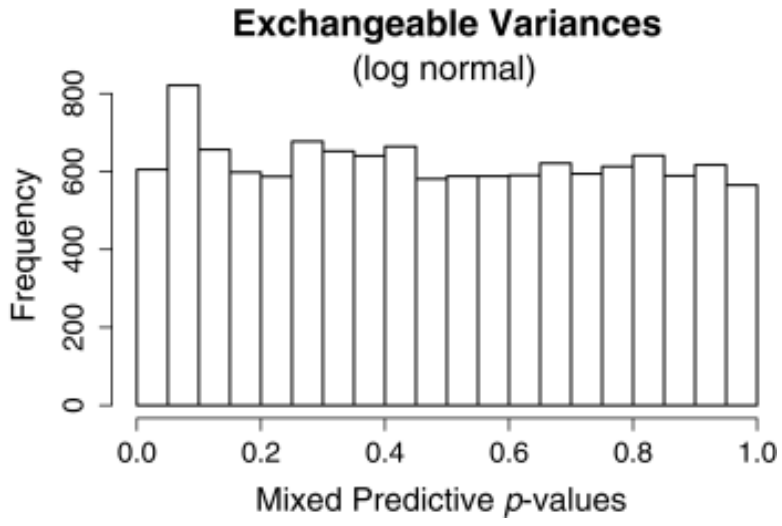
- Four possibilities: "equal variance model" $\sigma^2_{gs} \equiv \sigma^2_s$, $\sigma^2_s \sim \text{logNorm}(0, 10^4)$; exchangeable log-normal variance model; exchangeable with 1-parameter gamma $\sigma^{-2}_{gs} \sim \text{Gam}(2, \beta^{\text{prior}}_s)$, $\beta^{\text{prior}}_s \sim \text{Gam}(10^{-2}, 10^{-2})$; exchangeable model with 2-parameter gamma $\sigma^{-2}_{gs} \sim \text{Gam}(\alpha^{\text{prior}}_s, \beta^{\text{prior}}_s)$, $\alpha^{\text{prior}}_s \sim \text{Gam}(10^{-2}, 10^{-2})$ and $\beta^{\text{prior}}_s \sim \text{Gam}(10^{-2}, 10^{-2})$.
- Use predictive p-values under the model
- Simulate $y^{(\text{pred})}_{gs} \sim N(\alpha_g + f(\alpha_g), \sigma^2_{gs})$
- $p_{gs} \equiv \mathcal{P}(S^{2(\text{pred})}_{gs} > S^{2(\text{obs})}_{gs})$
- Under the null hypothesis of the model being "true", the distribution of p-values is almost uniform.

Non-linear array effects





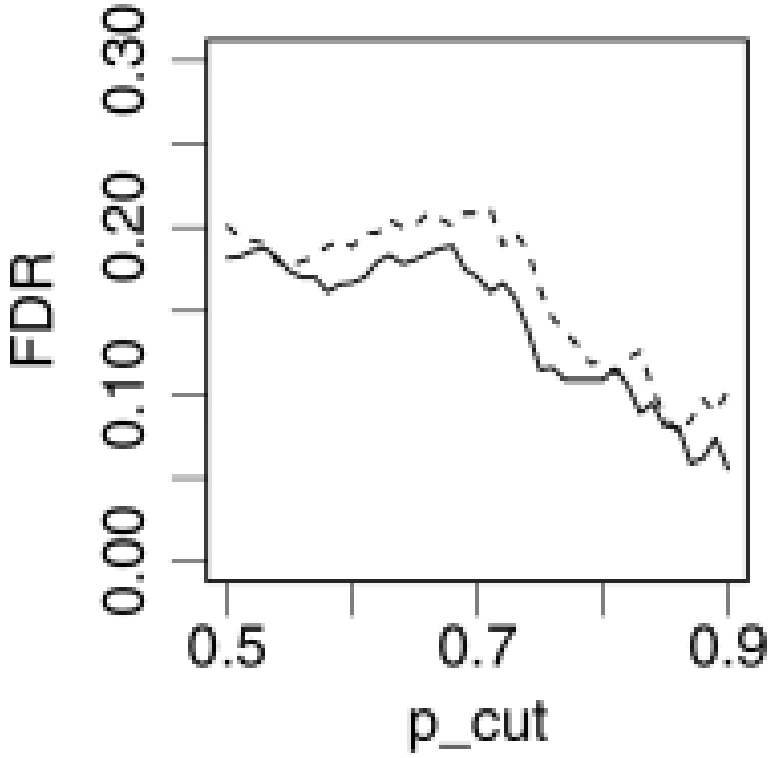


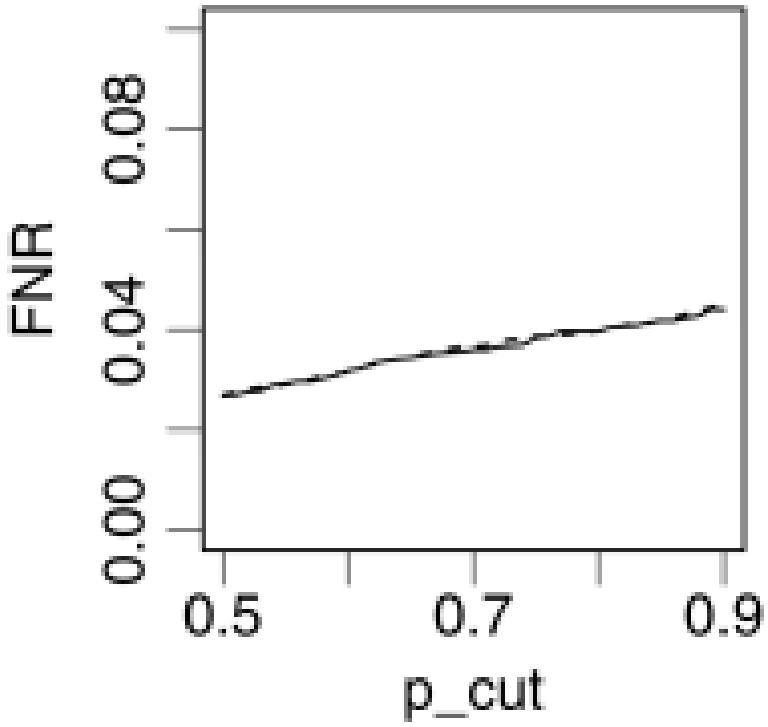


Integrated vs. non-integrated analysis

- Expect to obtain biased estimates of the array effects if they are estimated in a preprocessing step, similar to a measurement error problem
- A simulation study compare to pre-normalization using "loess' smoothing
- The ratios of MSE of array effects (loess vs. full model) are 1.5, 1.3, 1.2, 1.2, 1.4, 1.3 (averaged for 5 simulation each chip).
- Lower estimated FDR with full model, but no difference in FNR
- The larger the magnitude of array effects, the larger the difference between the pre-normalized and integrated models

Non-linear array effects

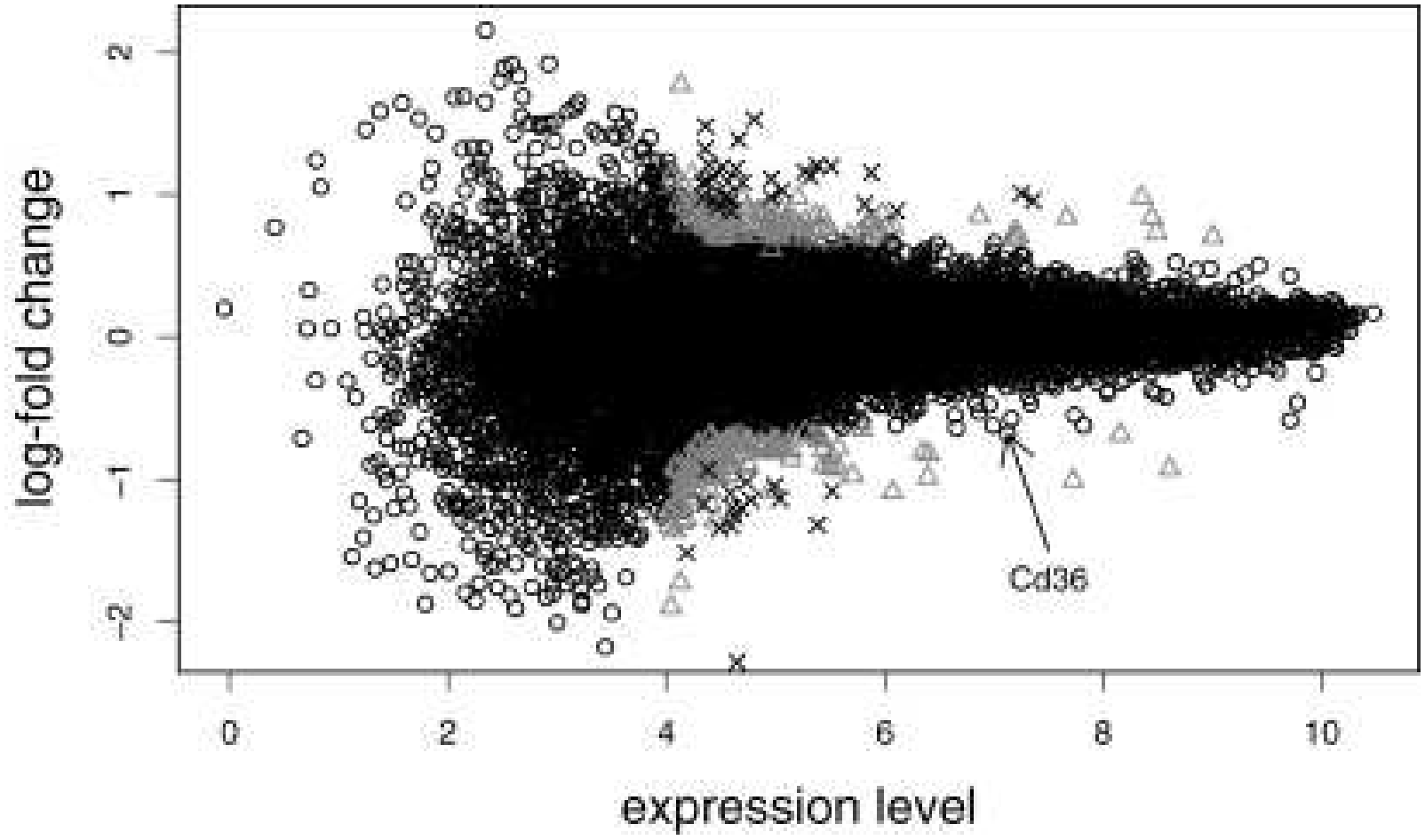




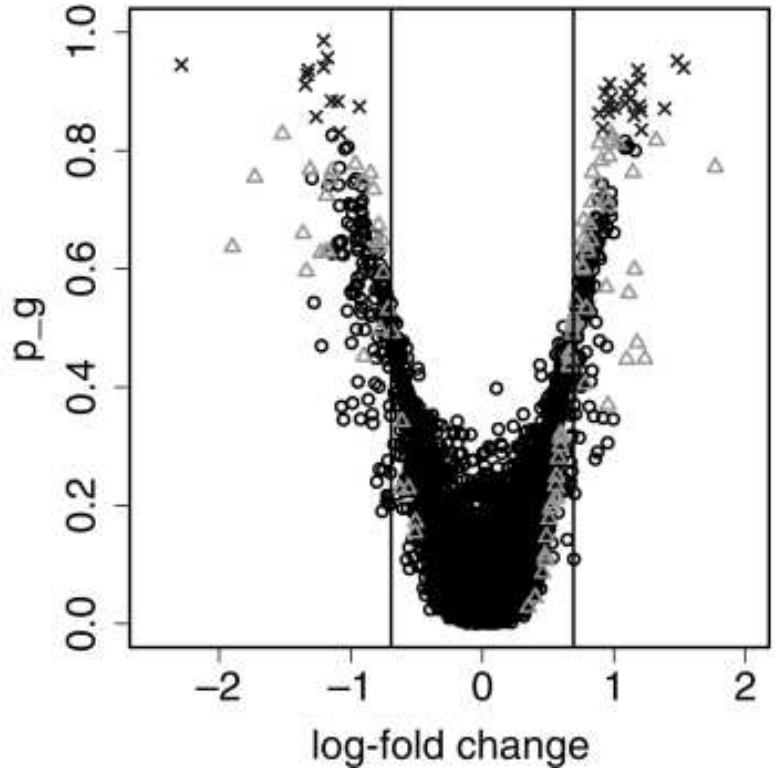
Application to mouse data

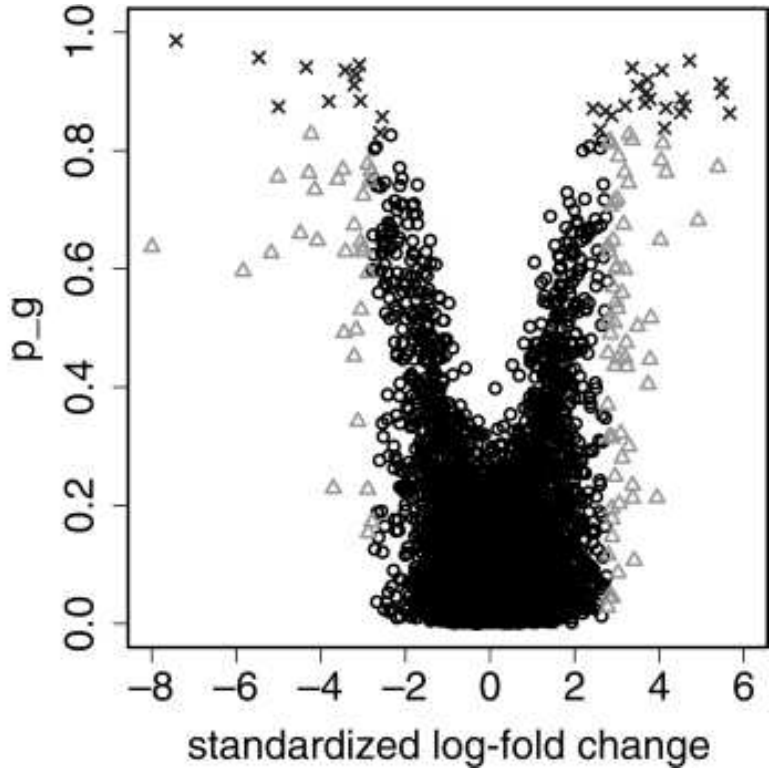
- Choose $\alpha_{cut} = 4$, $\delta_{cut} = \log(2)$ a priori
- $\widehat{FDR} = [1/|S(p_{cut}]| \sum_{g \in S(p_{cut})} (1 - p_g)$, where $S(p_{cut})$ is the group of genes with $p_g > p_{cut}$ and $|S(p_{cut})|$ is its cardinality.
- standardized log-fold difference $t_g \equiv E(\delta_g / [\sigma_{g1}^2 + \sigma_{g2}^2/3]^{1/2} \,|\, \text{data})$

Non-linear array effects



Non-linear array effects





Discussion

- Unified Bayesian hierarchical model
- Justify functional choice by exploratory analysis
- Joint estimation of differential effects and array effects
- Richer output
- Expert opinions expressed in pre-selected cutoffs