

**A Comparison of Direct and
Sequential FDR Algorithms:
*Computational Experiments for Exploratory
DNA Microarray Studies***

Danh V. Nguyen

Division of Biostatistics

University of California, Davis

May 27, 2004

Interface 2004, Baltimore, Maryland

Introduction

- Context: exploratory DNA microarray studies
- Primary goal: identify genes for follow-up studies, control for false positives
- Example: identify genes differentially expressed (DE) in breast cancer patients w/ mutations in the *BRCA1* relative to patients with *BRCA2* gene
- Notation: testing m hypotheses (one per gene):

	Accept	Reject	Total
Null true	U	V	m_0
Alternative true	T	S	m_1
Total	W	R	m

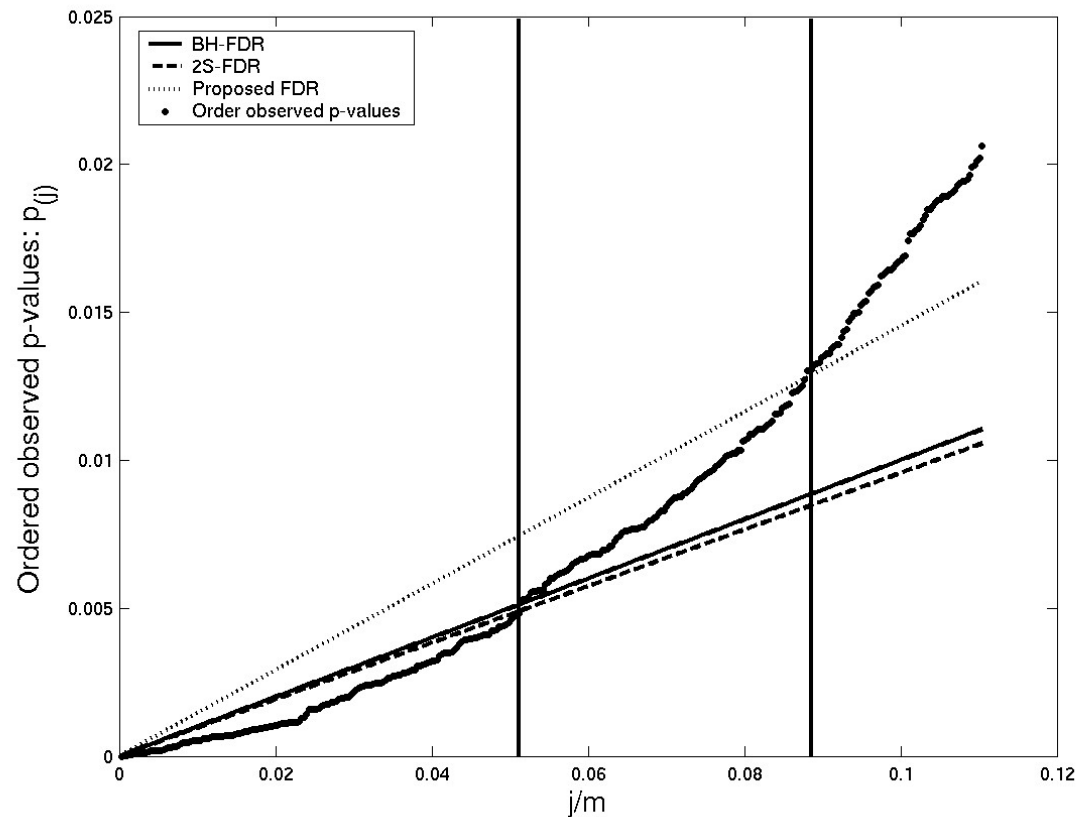
- $\text{FDR} = E\left(\frac{V}{R}I_{\{R>0\}}\right)$ (Benjamini & Hochberg 1995: BH95)
- The proportion of true null hypotheses is $\pi_0 \equiv m_0/m$
- Ordered observed p -values: $p_{(1)}, \dots, p_{(m)}$
- “Two” FDR algorithms (operational procedures)
 1. **Sequential FDR method:** (1) fix FDR control level α , (2) estimate rejection region (BH95)

2. **Direct FDR method:** (1) fix RR, then (2) estimate the corresponding FDR

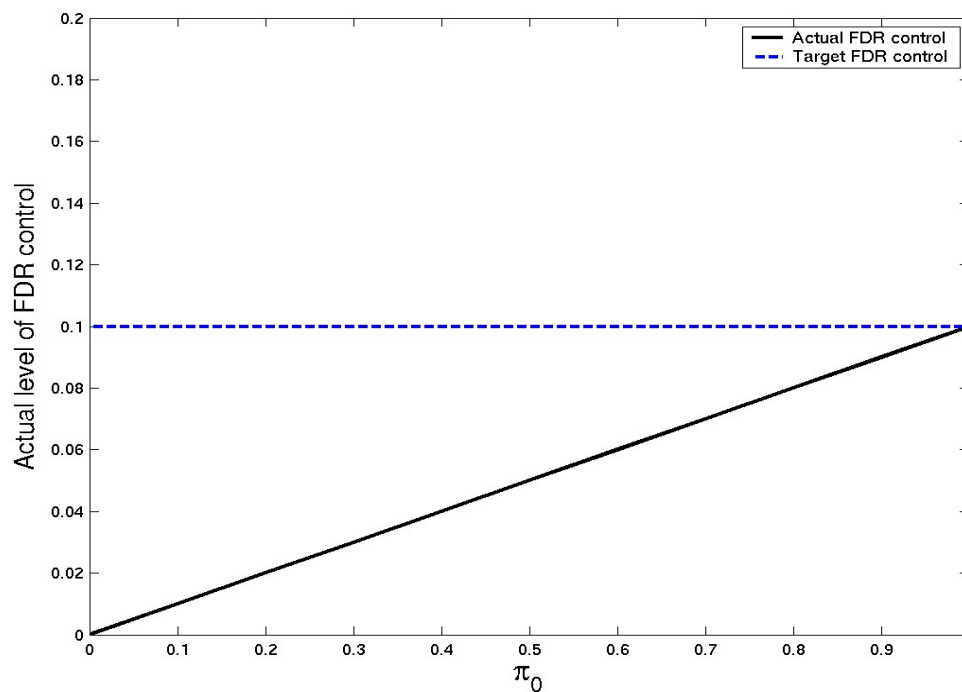
- Differences, i.e. operationally & power, can be *unified* via estimation $\pi_0 \longrightarrow$ computational framework

Sequential FDR & π_0

- *BH-FDR (original)*: $\hat{k}_{\text{BH}} = \max \{j : p_{(j)} \leq \frac{j}{m}\alpha\}$ and reject $p_{(1)}, \dots, p_{(\hat{k}_{\text{BH}})}$, $\alpha \in (0, 1)$



- $FDR \leq \pi_0 \alpha$ for $0 \leq m_0 \leq m$ (BH95);
 $FDR = \pi_0 \alpha$ (Finner & Rotter 2001)
- Therefore, conservative by $\pi_0 \alpha$



- \longrightarrow Decreased power

Least Conservative Sequential FDR & π_0

- Correction: run BH-FDR at $\alpha' = \alpha/\pi_0 \rightarrow \text{FDR} = \alpha$

- $\hat{k}_{\text{LC}} = \max \left\{ j : p_{(j)} \leq \frac{j}{m} \left(\frac{\alpha}{\pi_0} \right) \right\}$

- A class of sequential FDR controlling algorithms:

$$\hat{k} = \max \left\{ j : p_{(j)} \leq \frac{j}{m} \left(\frac{\alpha^*}{\hat{\pi}_0} \right) \right\}, \text{ reject } p_{(1)}, \dots, p_{(\hat{k})}$$

- $\hat{k}_{\text{BH}} = \max \left\{ j : p_{(j)} \leq \frac{j}{m} \left(\frac{\alpha}{\hat{\pi}_0(\text{BH})} \right) \right\}, \hat{\pi}_0(\text{BH}) \equiv 1$

Multi-Stage/Adaptive Sequential FDR & π_0

- Two-stage (2S-FDR): (Benjamini, Krieger, Yekutieli 2001: BKY)
 - (1) Estimate π_0 , $\hat{\pi}_0(\text{BKY}) = (m - r_1)/m$; $r_1 = \#$ rejections from BH-FDR procedure at $\alpha' = \alpha/(1+\alpha)$ level
 - (2) $\hat{k}_{\text{BKY}} = \max \left\{ j : p_{(j)} \leq \frac{j}{m} \left(\frac{\alpha'}{\hat{\pi}_0(\text{BKY})} \right) \right\}$
- Modified 2S-FDR (2SM-FDR):
 - (1) since choice of α' is more strict than α (i.e. $\alpha' < \alpha$), get estimate of π_0 at level $\alpha \rightarrow \hat{\pi}_0(\text{BKY-M})$
 - (2) $\hat{k}_{\text{BKY-M}} = \max \left\{ j : p_{(j)} \leq \frac{j}{m} \left(\frac{\alpha'}{\hat{\pi}_0(\text{BKY-M})} \right) \right\}$

Direct FDR Estimation & π_0 (Storey 2002, 2003)

- Specify RR: $\gamma = 0.005$ say (\rightarrow reject for p -values $< \gamma$)

- Estimate FDR (for chosen γ):

$$\widehat{\text{FDR}}_{\lambda}(\gamma) = \hat{\pi}_0(\lambda) \left(\gamma / \widehat{\text{Pr}}(P \leq \gamma) \right)$$

$$\widehat{\text{Pr}}(P \leq \gamma) = \#\{p_j \leq \gamma\} / m$$

- Key is estimation of π_0 (prop. of true nulls):

$$\hat{\pi}_0(\text{UB}) = \frac{\#\{\text{Null } p_j > \lambda\}}{m(1-\lambda)}: \text{ unbiased for } \pi_0$$

- Computable (conservatively biased) estimate:

$$\hat{\pi}_0(\lambda) = \frac{\#\{p_j > \lambda\}}{m(1-\lambda)}: E(\hat{\pi}_0) \geq \pi_0$$

“Direct is sequential”, via $\hat{\pi}_0(\lambda)$

- “Direct is sequential”:

$$\hat{k}_\lambda = \max \left\{ j : p_{(j)} \leq \frac{j}{m} \left(\frac{\alpha}{\hat{\pi}_0(\lambda)} \right) \right\}$$

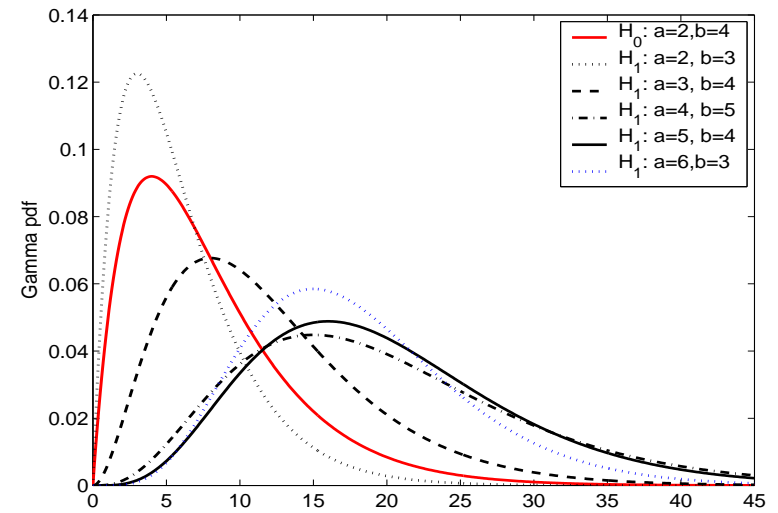
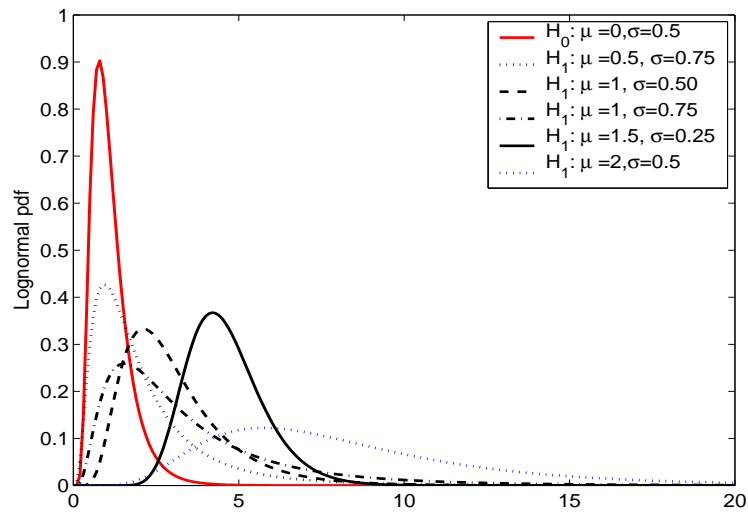
- Under this framework, direct and sequential methods (algorithms), both just approximate the least conservative (“optimal”) sequential FDR algorithm:

$$\hat{k}_{\text{LC}} = \max \left\{ j : p_{(j)} \leq \frac{j}{m} \left(\frac{\alpha}{\pi_0} \right) \right\}$$

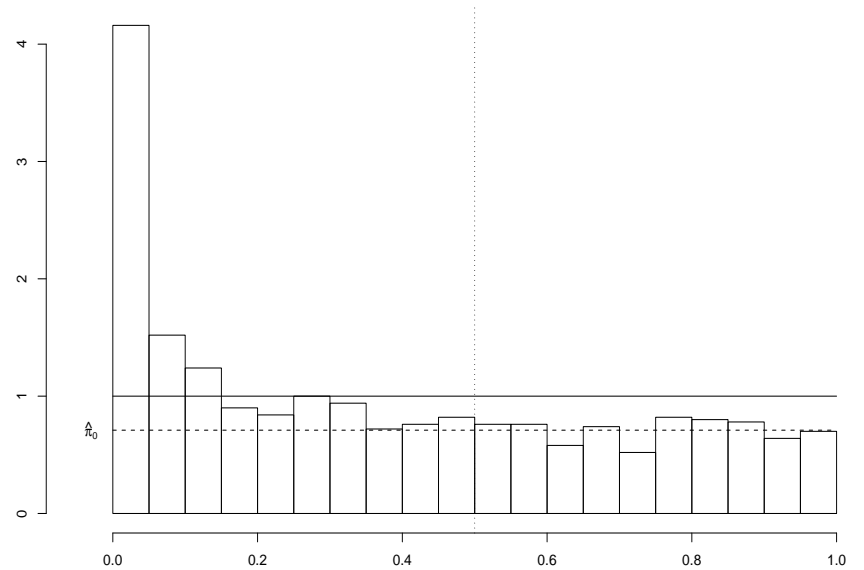
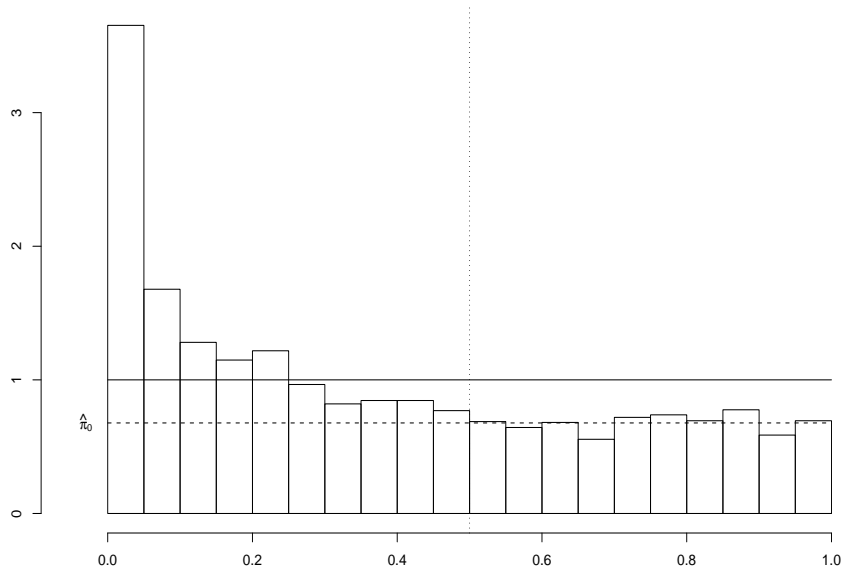
Less ideal simulation conditions

- E.g. $z_i \sim N(\mu, 1)$, π_0 w/ $N(0, 1)$, π_1 w/ $N(2, 1)$;
 p -values: $p_i = \Pr(Z \geq z_i)$ $i = 1, 2, \dots, m$ (*exact*)
- Microarray data, at a "minimum" (H_0 & H_1):
 - ★ variance: non-constant; σ_0^2, σ_1^2 a mixture
 - ★ mean expression for each gene different; μ_0, μ_1 also a mixture
 - ★ now sampling distribution of test statistics already unknown; \longrightarrow sample size matter (p -values not "exact")
 - ★ measurement error (additive and multiplicative) (e.g. Rocke & Durbin 2001)
 - ★ independence? (coordinated pathways \longrightarrow cluster/groups of co-expressed genes/family)

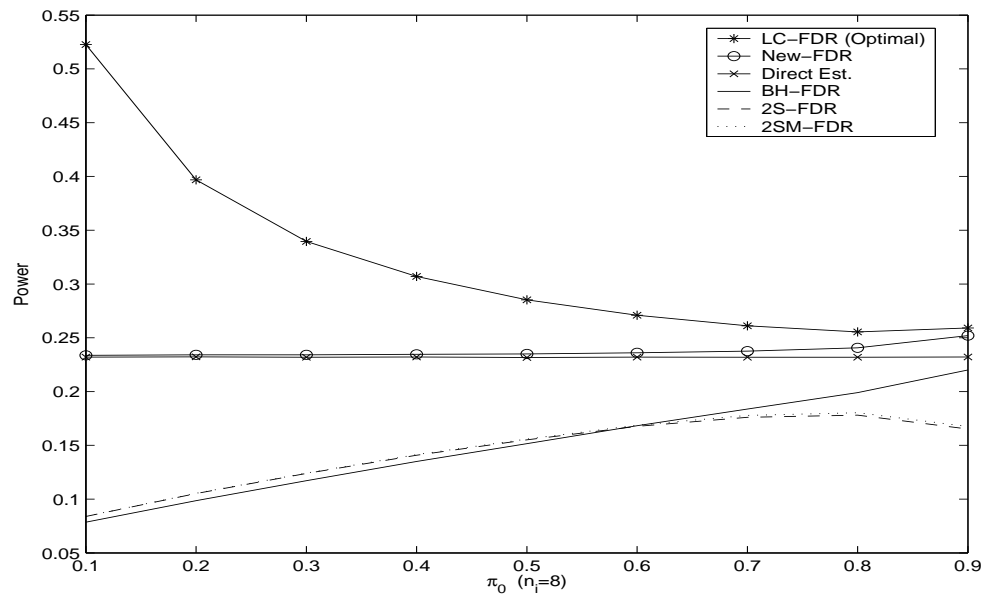
Simple example:



BRCA-mutation & simulated (“sample”) data:



Power: an example



Summary of simulation results (power)

- Two sample comparison
- General: Poor estimation of $\pi_0 \rightarrow$ substantial power lost
- General: n (per group) large (e.g. 64) all converge to optimal/benchmark
- *Small to moderate n : 4, 8, 16 per group*
 - ★ 2S-FDR gain is substantial for $n = 16$ over BH-FDR
 - ★ Direct, and its sequential “equivalent” algorithm (\hat{k}_λ) negligible difference w/ optimal algorithm

- *Larger n*: 32, 64; Adaptive algorithms still far from benchmark (also reflect in poor estimate of π_0)
- Other cases: changing effect size, unequal group sample sizes, variation of two-sample t-statistics (e.g., penalized t-statistics), p from t-distribution under constant variance assumption, permutation p -values, violation of independence (block dependence), different distribution model for expression (normal, lognormal, gamma etc.)