



Gene Identification Using True Discovery Rate Degree of Association Sets and Estimates Corrected for Regression to the Mean

ASA San Francisco Bay Area Chapter
Symposium on Statistical Genetics, Genomics and Proteomics
14 May 2011
Michael R. Crager

Turning the promise of genomics
into the practice of medicine™

Crager MR (2010). Gene identification using true discovery rate degree of association sets and estimates corrected for regression to the mean. *Statistics in Medicine* **29**:33-45.

Background: Gene Identification at Genomic Health, Inc.



- Use genomic information from tumor tissue to
 - Estimate risk of cancer recurrence
 - Estimate effectiveness of preventative therapy
- Goal: Help patients and physicians decide on treatment options

- Continuous measure
- Assessed from amount of gene's RNA in tumor tissue sample
 - Reverse transcriptase polymerase chain reaction (RT-PCR)
 - Log scale

Oncotype DX[®] Recurrence Score[®] Risk of Recurrence of Breast Cancer

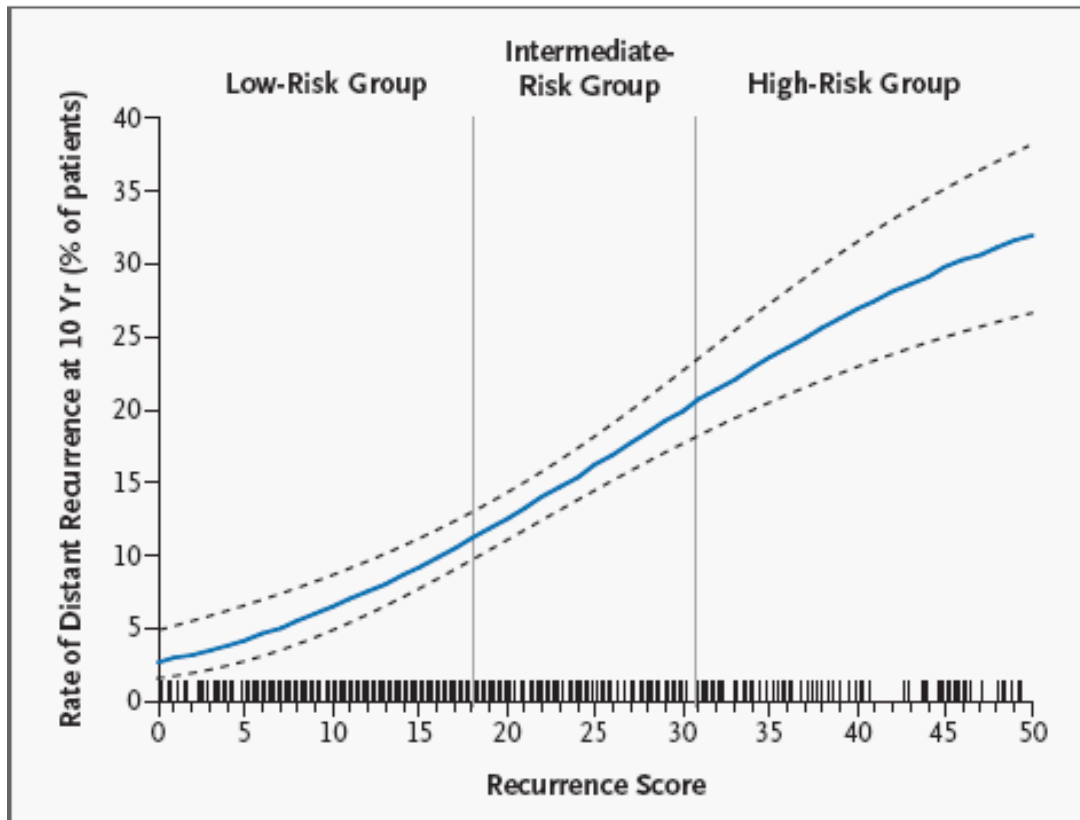


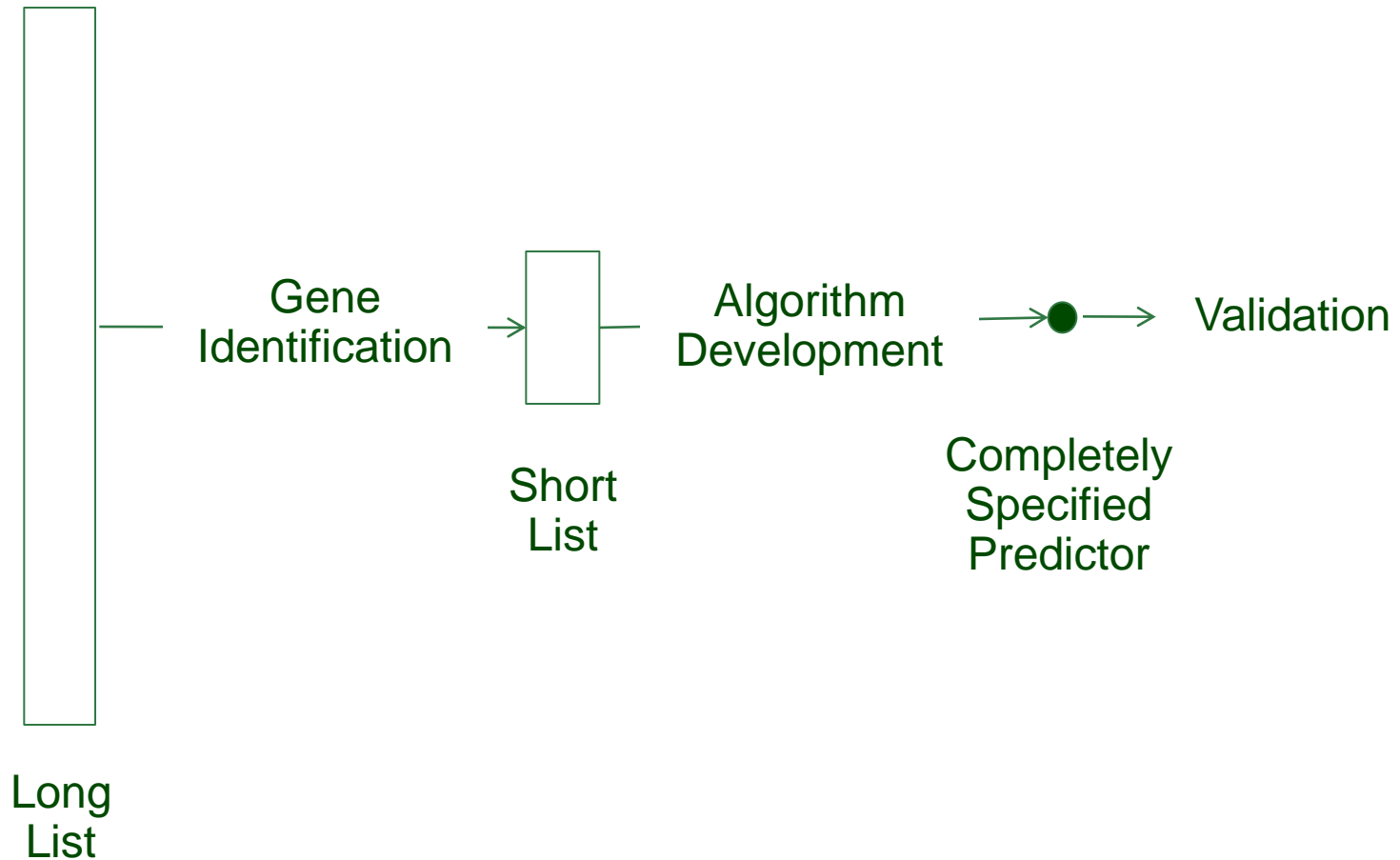
Figure 4. Rate of Distant Recurrence as a Continuous Function of the Recurrence Score.

The continuous function was generated with use of a piecewise log-hazard-ratio model.²⁸ The dashed curves indicate the 95 percent confidence interval. The rug plot on top of the x axis shows the recurrence score for individual patients in the study.

NSABP B-14

Paik, Shak, Tang *et al.*, 2004
NEJM 24:3726-3734

Clinical Development Process Synopsis



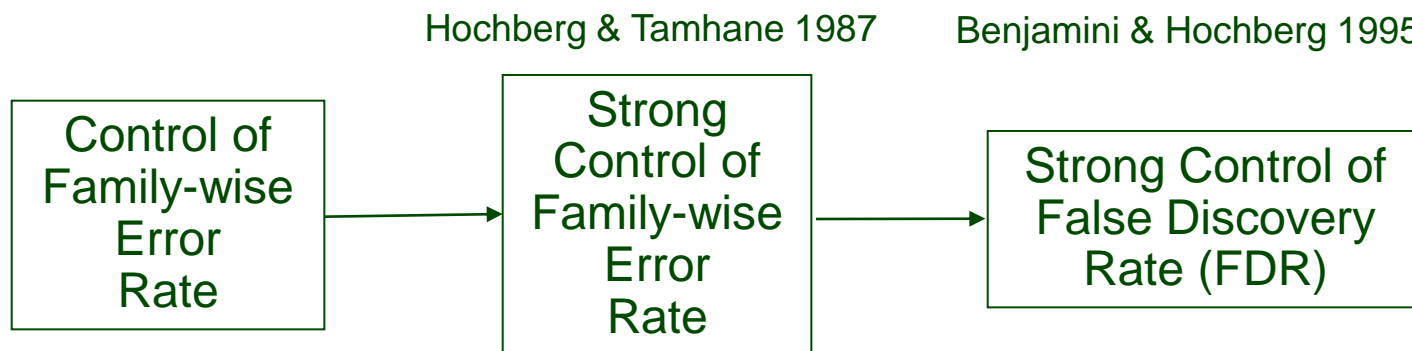
- Quick Review of False Discovery Rate Concepts and Methods
- From “Any Association” to “Substantial Association” of Genes with Clinical Outcome
 - True Discovery Rate Degree of Association (TDRDA) Sets
 - Estimates Corrected for Regression to the Mean



Dan C.
Oncotype DX Patient

False Discovery Rate Concepts

Turning the promise of genomics
into the practice of medicine™



What: $P(\geq 1 \text{ false rejection})$

$P(\geq 1 \text{ false rejection})$

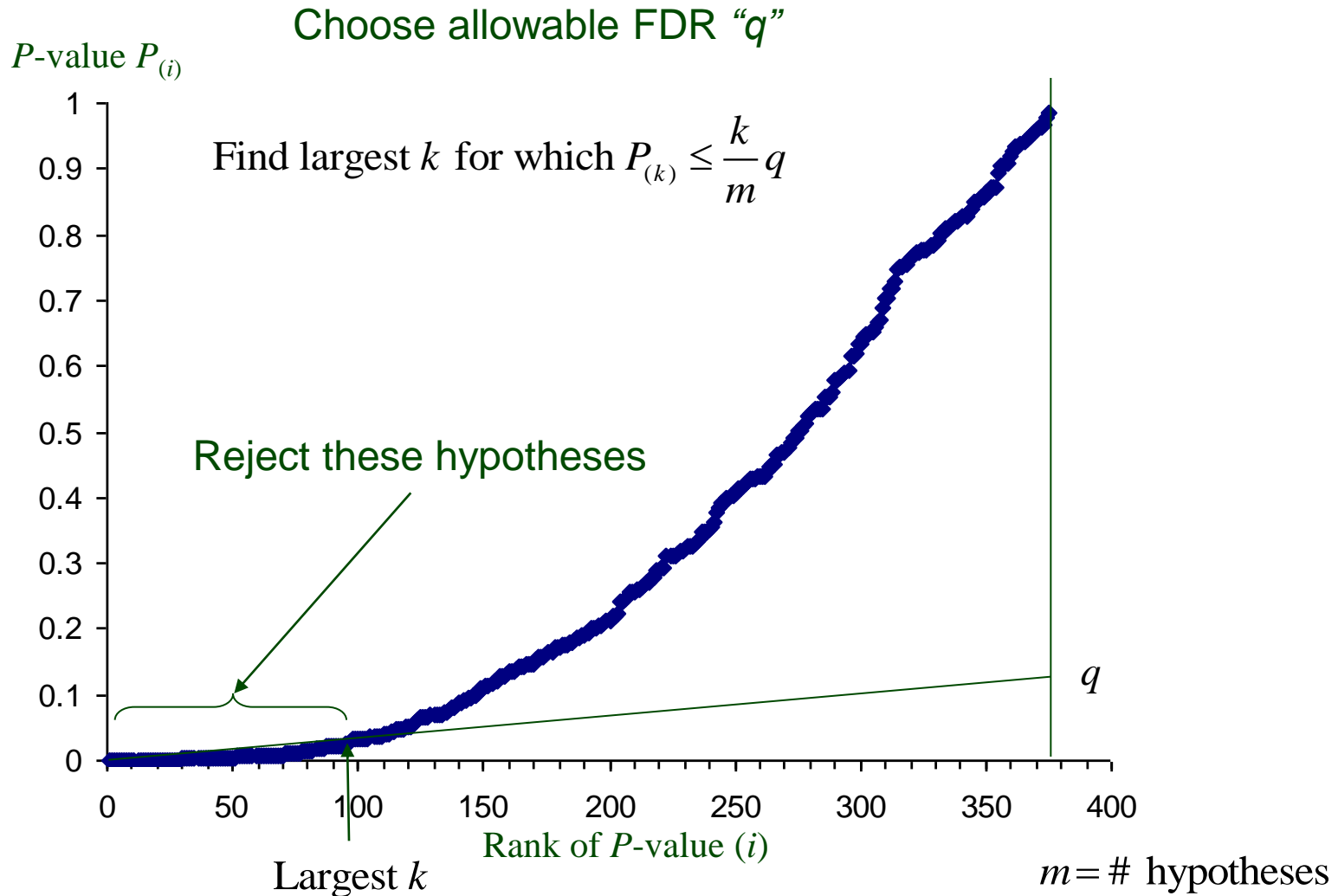
Expected proportion of rejections that are false

When: When all null hypotheses are true

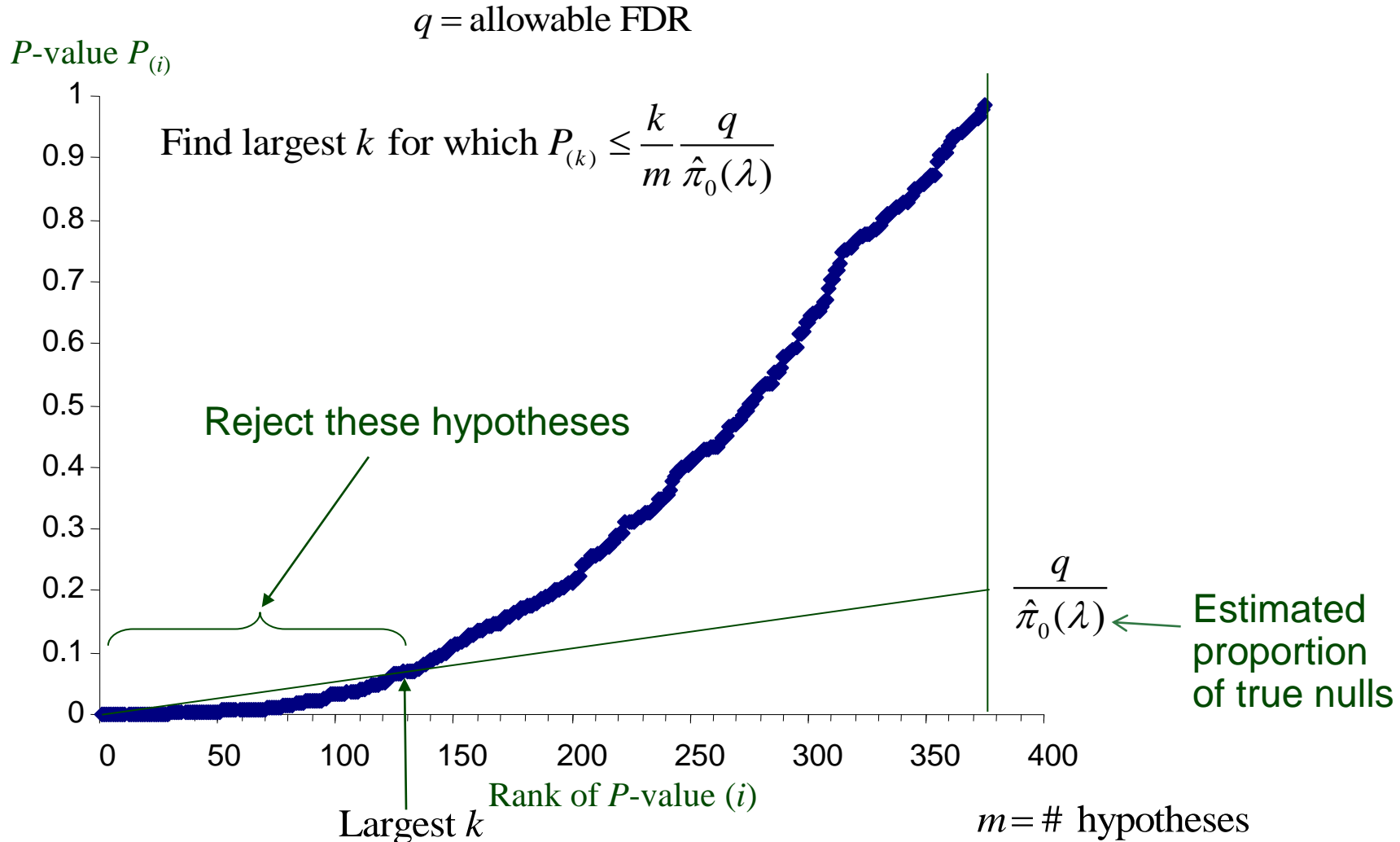
Under all combinations of true, false nulls

Under all combinations of true and false nulls

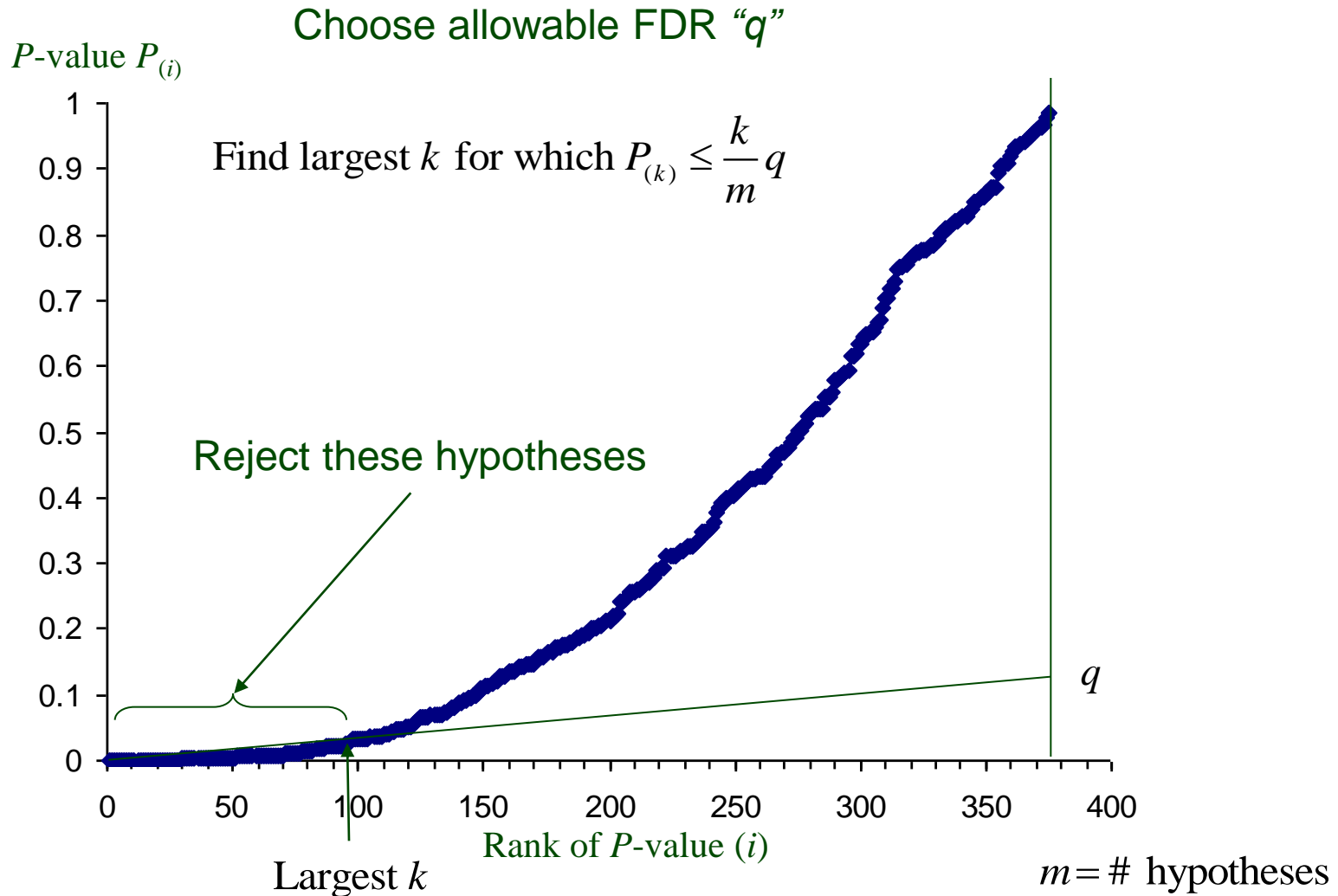
Benjamini-Hochberg (1995) Method



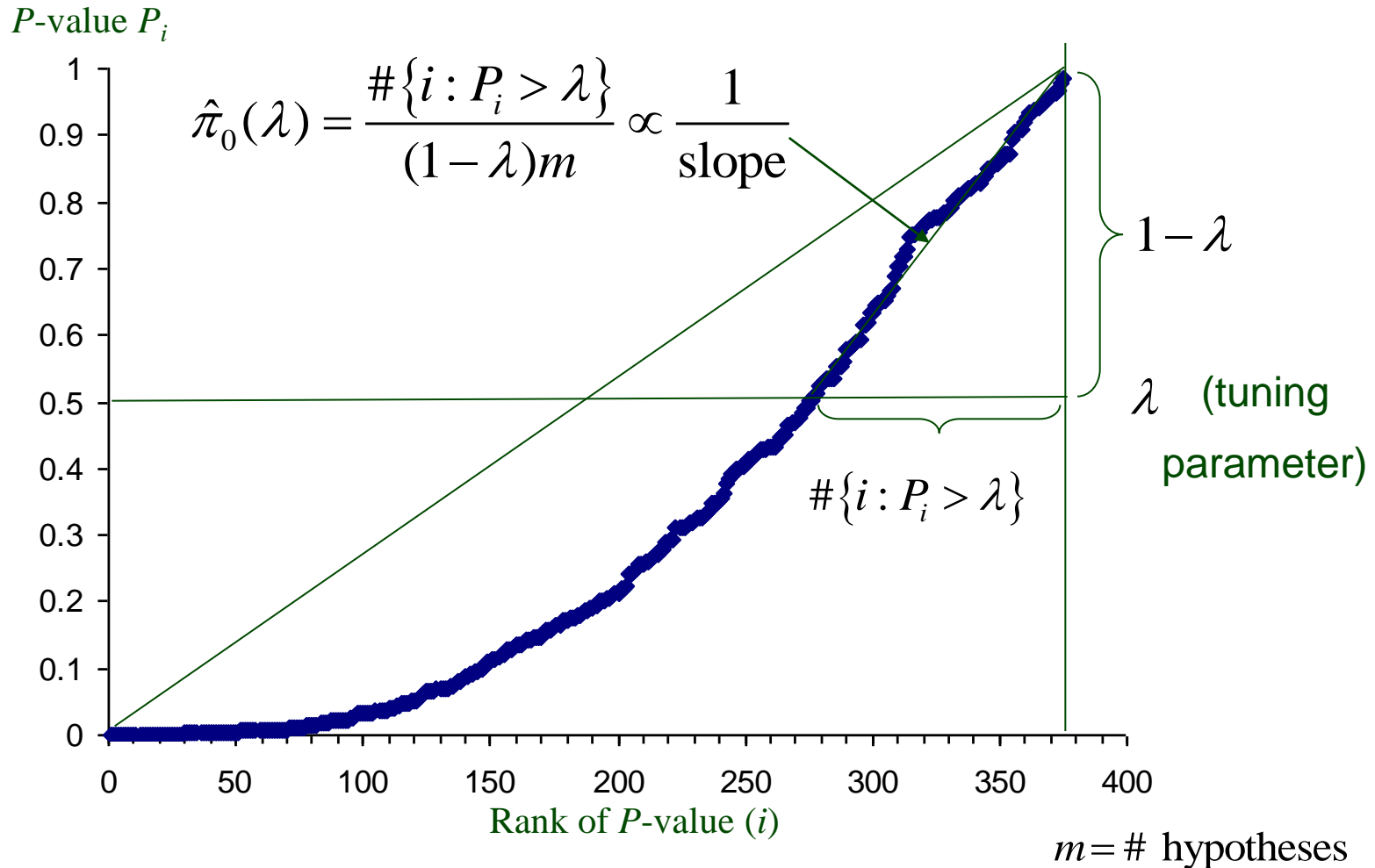
Storey's (2002) Method



Benjamini-Hochberg (1995) Method

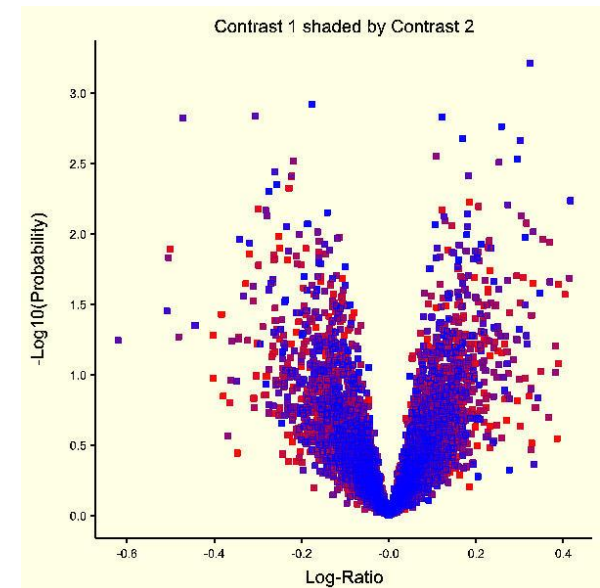


Storey's Method: Estimation of Proportion of True Null Hypotheses

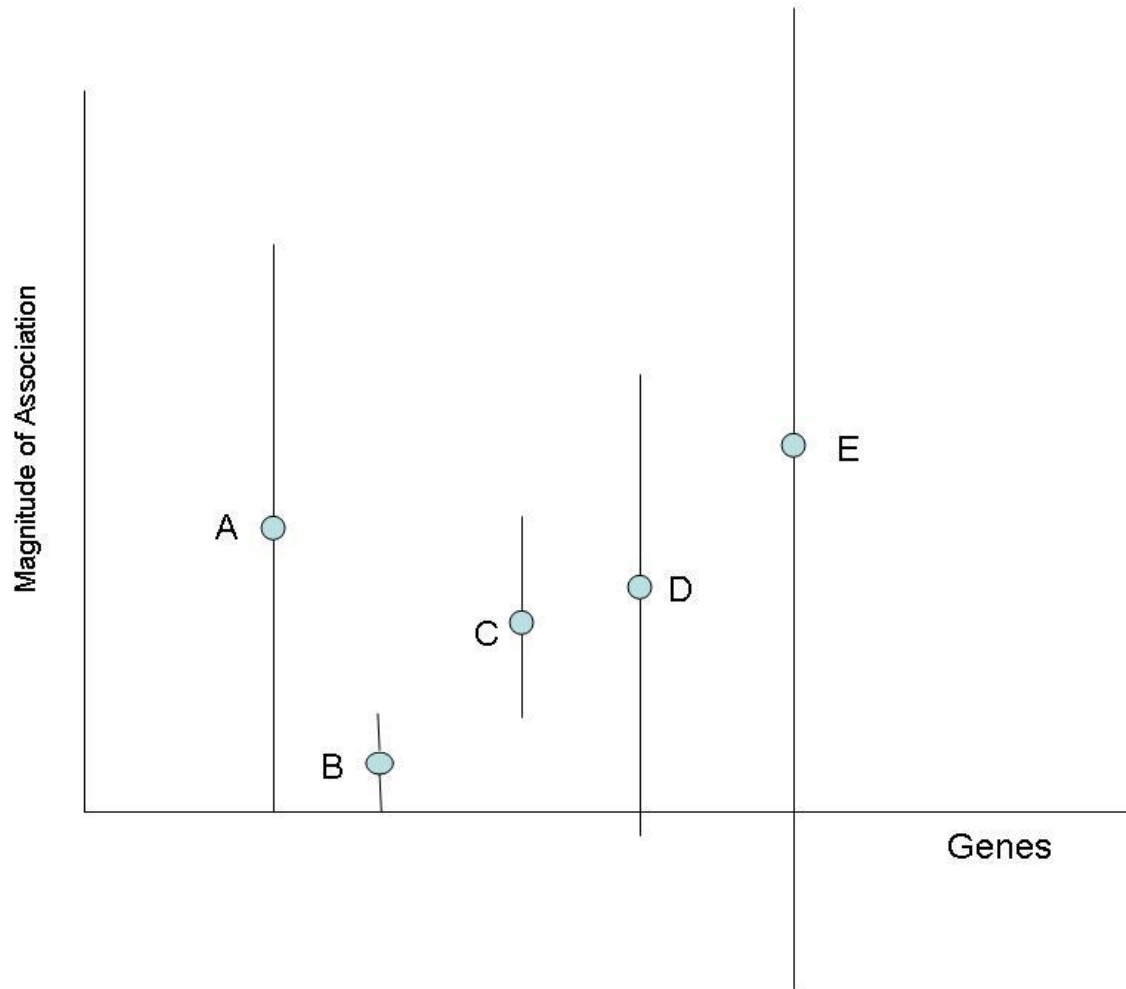


- P -values from tests of point null hypothesis of *no association at all* with gene expression
 - FDR control based on cutoff criterion for p -values
- Point estimates of degree of association with gene expression
 - Estimated hazard ratios, for example

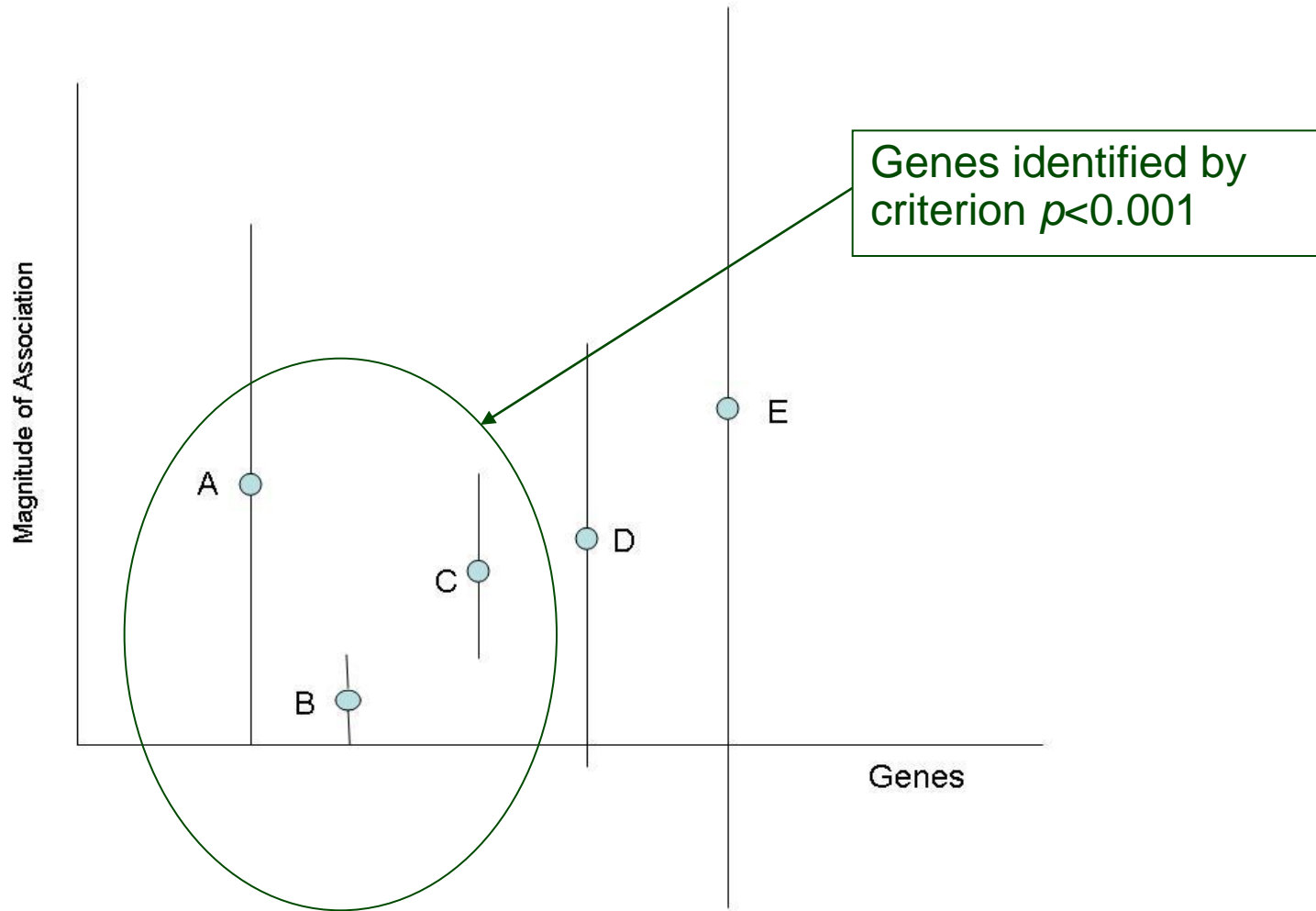
Volcano Plots
Combine p -values with
Estimates



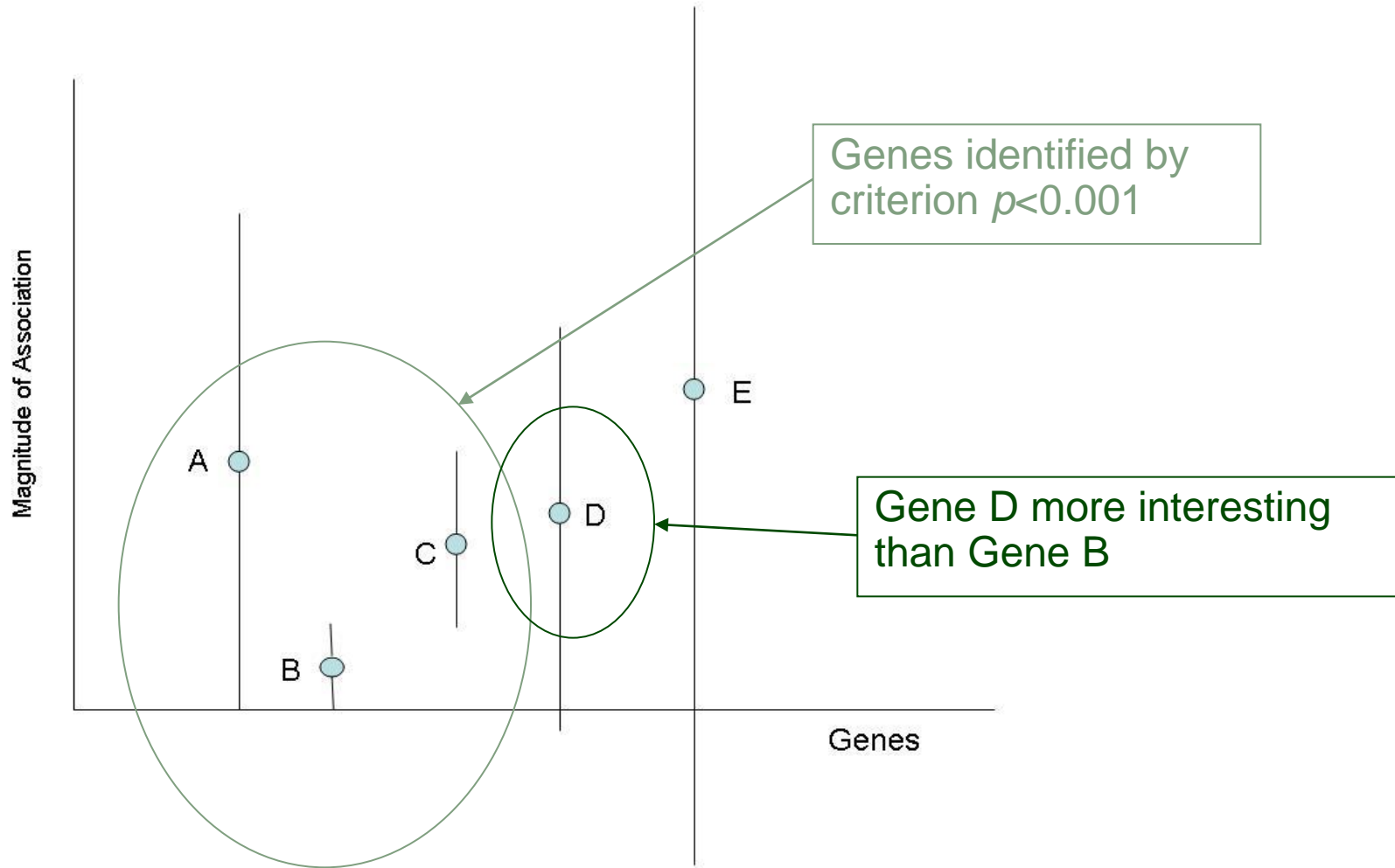
Hypothetical standardized degree of association estimates and 99.9% confidence intervals for 5 genes



Hypothetical standardized degree of association estimates and 99.9% confidence intervals for 5 genes



Hypothetical standardized degree of association estimates and 99.9% confidence intervals for 5 genes



Identify genes with ~~any~~ substantial association with clinical outcome . . .

. . . while controlling false discovery rate



True Discovery Rate
Degree of Association Analysis

Turning the promise of genomics
into the practice of medicine™

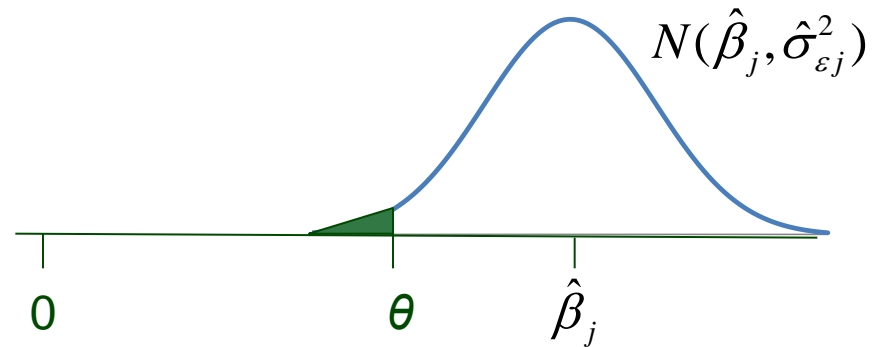
Interval Null Hypotheses (Van de Miel and Kim, 2007)

Usual Methods	Van de Wiel-Kim
Point null hypotheses $H_0: \beta=0$ vs. $H_1: \beta \neq 0$	Interval null hypotheses $H_0: \beta \leq \theta$ vs. $H_1: \beta > \theta$
Estimate proportion of true nulls and FDR using known distribution of test statistic under H_0	Estimate proportion of true nulls and FDR using nonparametric deconvolution
Mathematical demonstration that FDR estimate is conservative	Heuristic argument that FDR estimate is sensible, validation by simulation study

FDR Control Using Interval Null Hypotheses

- Minimal “interesting” standardized absolute log hazard ratio $\theta = |\ln \gamma|$
- For $j=1,2,\dots,m$ genes, log hazard ratio estimate $\hat{\beta}_j$
standard error of estimate $\hat{\sigma}_{\varepsilon j}$
- Estimate is approximately normal \Rightarrow size alpha test of null hypothesis $H_j^\theta : |\beta_j| \leq \theta$ versus alternative $|\beta_j| > \theta$ is to reject null if

$$\frac{|\hat{\beta}_j| - \theta}{\hat{\sigma}_{\varepsilon j}} > \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$



Storey's method applied to interval null hypothesis tests

1. Fix λ

2. Fix θ

3. Set acceptable FDR q

4. Compute p -values $P_j(\theta) = \min \left\{ 2 \left[1 - \Phi \left(\frac{|\hat{\beta}_j| - \theta}{\hat{\sigma}_{\varepsilon j}} \right) \right], 1 \right\}$

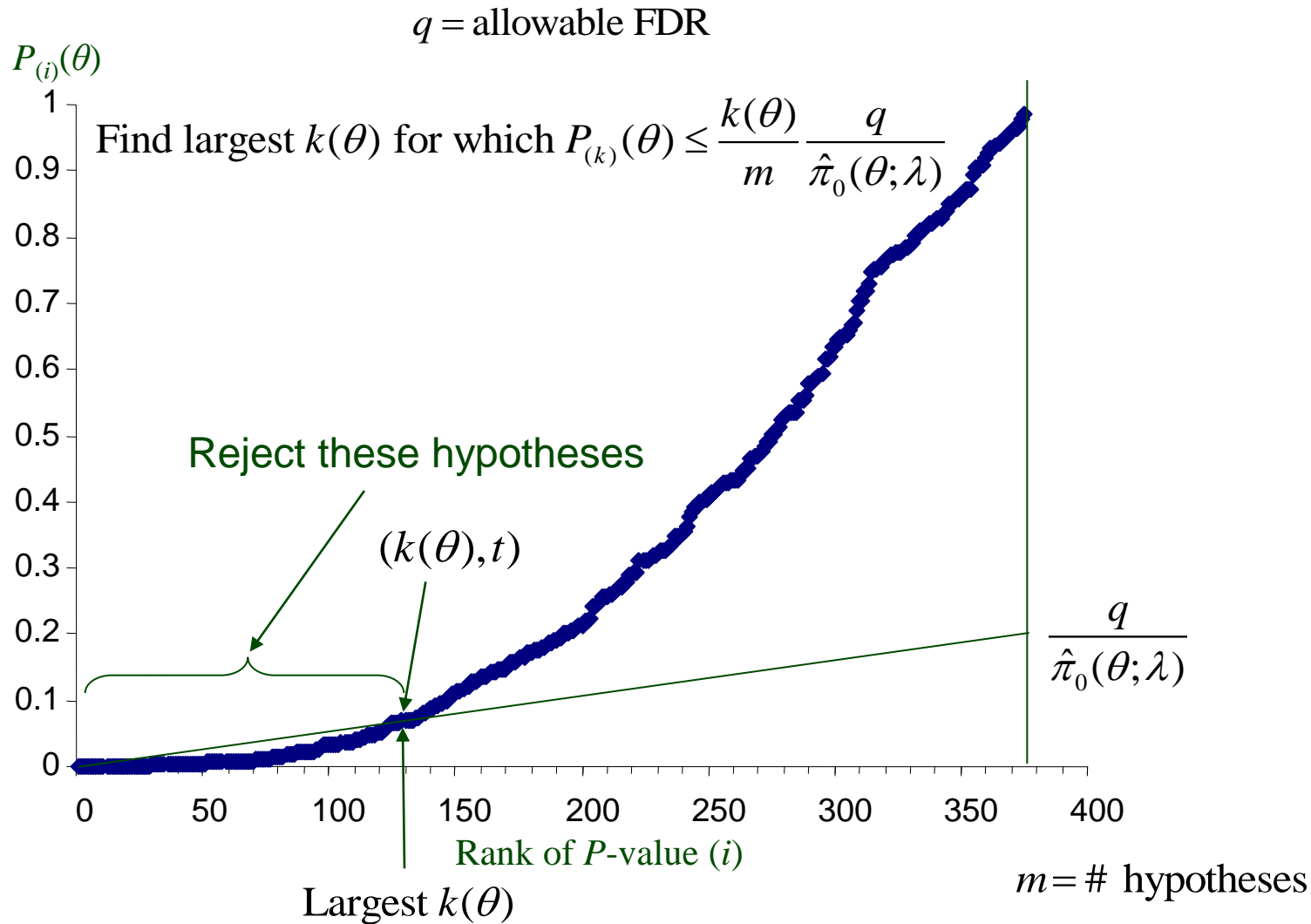
5. Order p -values $P_{(j)}(\theta)$, $j = 1, 2, \dots, m$

6. Find largest value $k(\theta)$ for which

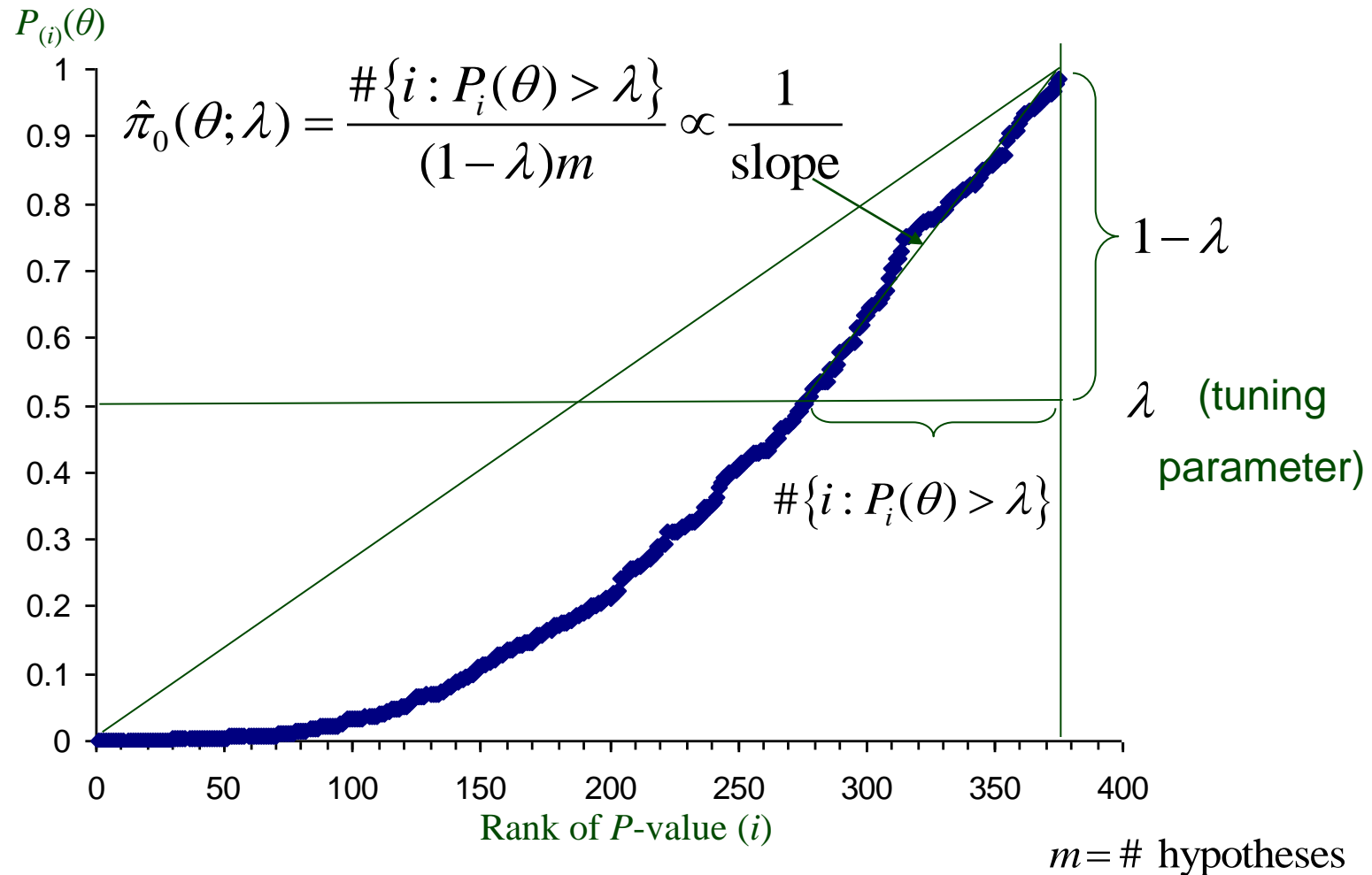
$$P_{(k(\theta))}(\theta) \leq \frac{k(\theta)}{m} \frac{q}{\hat{\pi}_0(\theta; \lambda)} \quad \text{where} \quad \hat{\pi}_0(\theta; \lambda) = \frac{\#\{i : P_i(\theta) > \lambda\}}{(1 - \lambda)m}$$

7. Reject hypotheses $H_{(1)}^\theta, H_{(2)}^\theta, \dots, H_{(k)}^\theta$ and identify associated genes } TDRDA($\theta; 1-q$)

Storey's method applied to interval null hypothesis tests



Storey's Method: Estimation of Proportion of True Interval Null Hypotheses



True Discovery Rate Degree of Association (TDRDA) Sets

- We can expect $100(1-q)\%$ of identified genes $\text{TDRDA}(\theta; 1-q)$ truly have absolute log hazard ratio $> \theta$
- If minimal “interesting” θ not known:
 - Vary θ and generate all the sets $\text{TDRDA}(\theta; 1-q)$
 - Sort genes by maximum lower bound (MLB) θ for which each is included in $\text{TDRDA}(\theta; 1-q)$

$$\theta_j^{\max} = \max \{ | \theta | : H_j^\theta \text{ is rejected} \}$$

Theorem If $\theta_1 < \theta_2$ then $\text{TDRDA}(\theta_1; 1-q) \supseteq \text{TDRDA}(\theta_2; 1-q)$

Note: Requires fixed λ

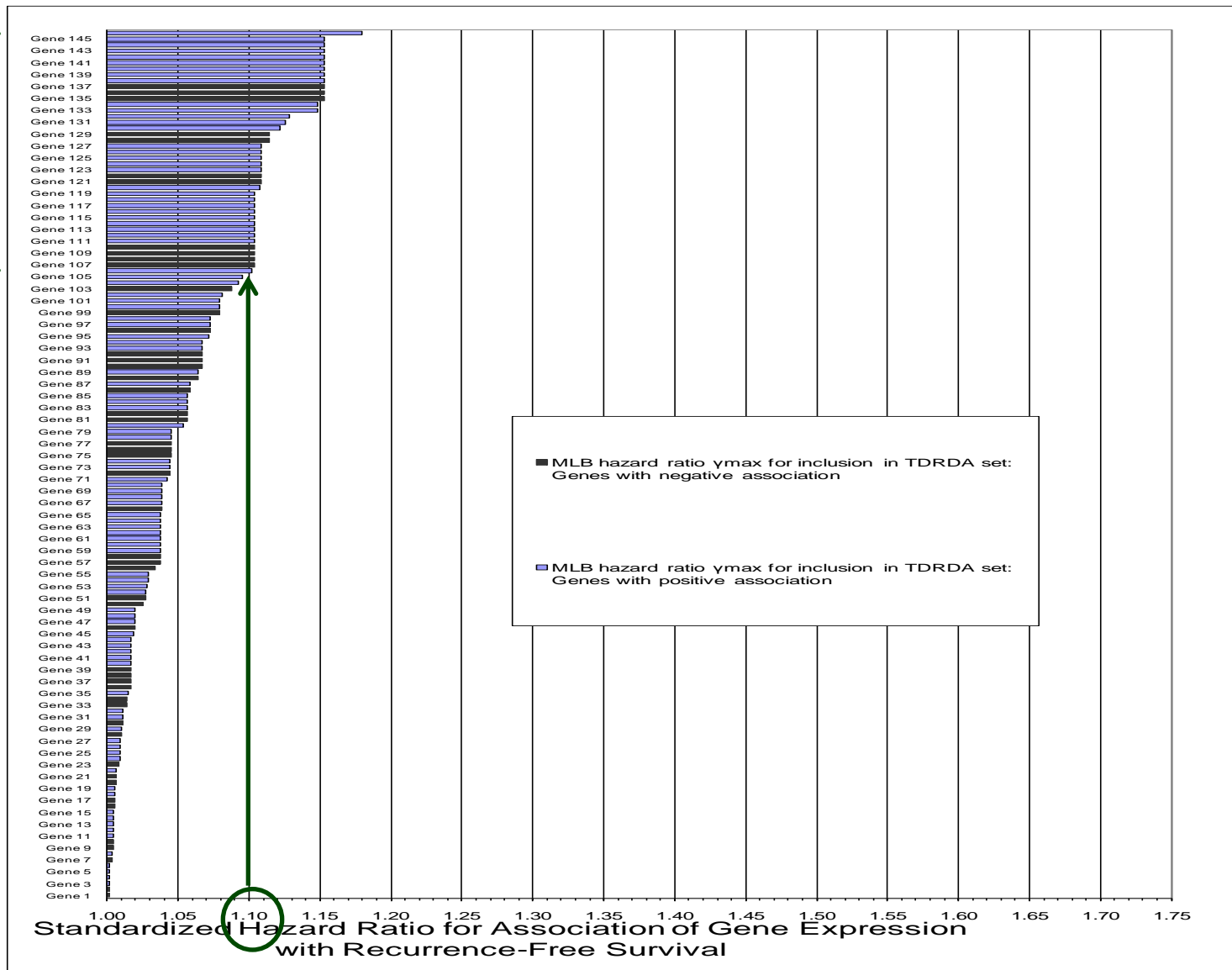
Corollaries

- Gene j will be included in every set $\text{TDRDA}(\theta; 1-q)$ for every $\theta \in [0, \theta_j^{\max}]$
- $\text{TDRDA}(\theta; 1-q) \subseteq \text{TDRDA}(0; 1-q)$ for all $\theta > 0$
 - TDRDA sets refine the set of genes identified by Storey's procedure with Wald test of point null hypothesis

- Case series of 136 node-negative, ER-positive breast cancer patients
- Follow-up up to 12 years
- Endpoint: breast cancer recurrence
 - 26 events
- 363 genes
 - Reference-gene-normalized expression by PCR
 - Standardized to 1 SD

Example TDRDA Set Plot (TDR = 80%) Hazard Ratios for Recurrence of Breast Cancer

TDRDA Set(min. HR=1.10)



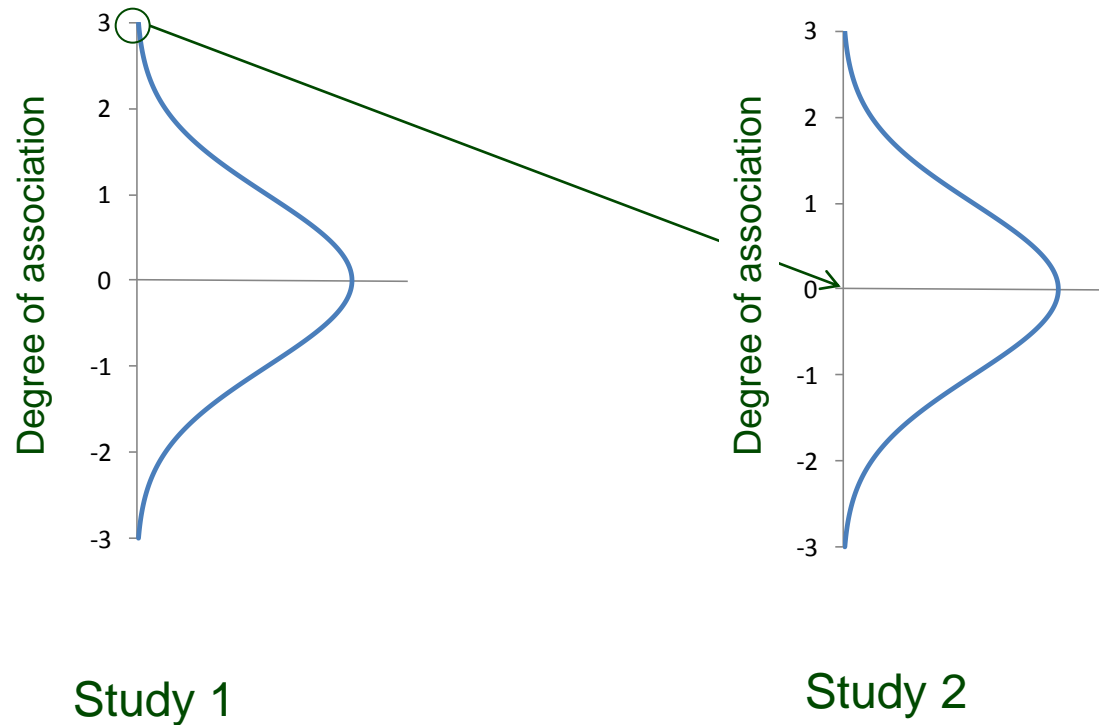


Estimates Corrected for Regression to the Mean

Turning the promise of genomics
into the practice of medicine™

Key Problem: Regression to the Mean

Hypothetical example: *No gene is truly associated with outcome*



Key problem: Regression to the mean

Idea: model “true” log hazard ratios $\beta_j, j = 1, 2, \dots, m$

as (not necessarily independent) sample from distribution $N(\mu_\beta, \sigma_\beta^2)$

Log HR estimates $\hat{\beta}_j$ with independent error $N(0, \sigma_{\varepsilon j}^2)$

=> True log HR and estimate have bivariate normal distribution

$$\begin{pmatrix} \hat{\beta}_j \\ \beta_j \end{pmatrix} \sim N \left(\begin{pmatrix} \mu_\beta \\ \mu_\beta \end{pmatrix}, \begin{bmatrix} \sigma_\beta^2 + \sigma_{\varepsilon j}^2 & \sigma_\beta^2 \\ \sigma_\beta^2 & \sigma_\beta^2 \end{bmatrix} \right)$$

which implies

$$E(\beta_j | \hat{\beta}_j) = \mu_\beta + \frac{\rho \sigma_\beta}{\sqrt{(\sigma_\beta^2 + \sigma_{\varepsilon j}^2)}} (\beta_j - \mu_\beta) = \mu_\beta + \frac{\sigma_\beta^2}{\sigma_\beta^2 + \sigma_{\varepsilon j}^2} (\hat{\beta}_j - \mu_\beta)$$

Unknown

Regression-to-the-mean-corrected estimate of log hazard ratio

By the linearity of expectation (independence of the $\hat{\beta}_i$ not required)

$$E\left(\hat{\beta}_i - \mu_\beta\right)^2 = \sigma_\beta^2 + \sigma_{\varepsilon i}^2 \Rightarrow \sum_{i=1}^m E\left(\beta_i - \mu_\beta\right)^2 = m\sigma_\beta^2 + \sum_{i=1}^m \sigma_{\varepsilon i}^2$$

Estimate of the variance of the true log hazard ratios

$$\hat{\alpha}_\beta^2 = \frac{1}{m-1} \sum_{i=1}^m \left(\hat{\beta}_i - \bar{\beta}\right)^2 - \bar{\sigma}_\varepsilon^2 \quad \text{where} \quad \bar{\alpha}_\varepsilon^2 = \frac{1}{m} \sum_{i=1}^m \sigma_{\varepsilon i}^2 \quad \bar{\beta} = \frac{1}{m} \sum_{i=1}^m \beta_i$$

RM-corrected estimate of log hazard ratio

$$\hat{\beta}_j^* = \bar{\beta} + \frac{\hat{\sigma}_\beta^2}{\hat{\alpha}_\beta^2 + \sigma_{\varepsilon j}^2} \left(\hat{\beta}_j - \bar{\beta}\right)$$

Regression-to-the-mean-corrected estimate of log hazard ratio

By the linearity of expectation (independence of the $\hat{\beta}_i$ not required)

$$E\left(\hat{\beta}_i - \mu_\beta\right)^2 = \sigma_\beta^2 + \sigma_{\varepsilon i}^2 \Rightarrow \sum_{i=1}^m E\left(\beta_i - \mu_\beta\right)^2 = m\sigma_\beta^2 + \sum_{i=1}^m \sigma_{\varepsilon i}^2$$

Estimate of the variance of the true log hazard ratios

$$\hat{\alpha}_\beta^2 = \frac{1}{m-1} \sum_{i=1}^m \left(\hat{\beta}_i - \bar{\beta}\right)^2 - \bar{\sigma}_\varepsilon^2 \quad \text{where} \quad \bar{\alpha}_\varepsilon^2 = \frac{1}{m} \sum_{i=1}^m \sigma_{\varepsilon i}^2 \quad \bar{\beta} = \frac{1}{m} \sum_{i=1}^m \beta_i$$

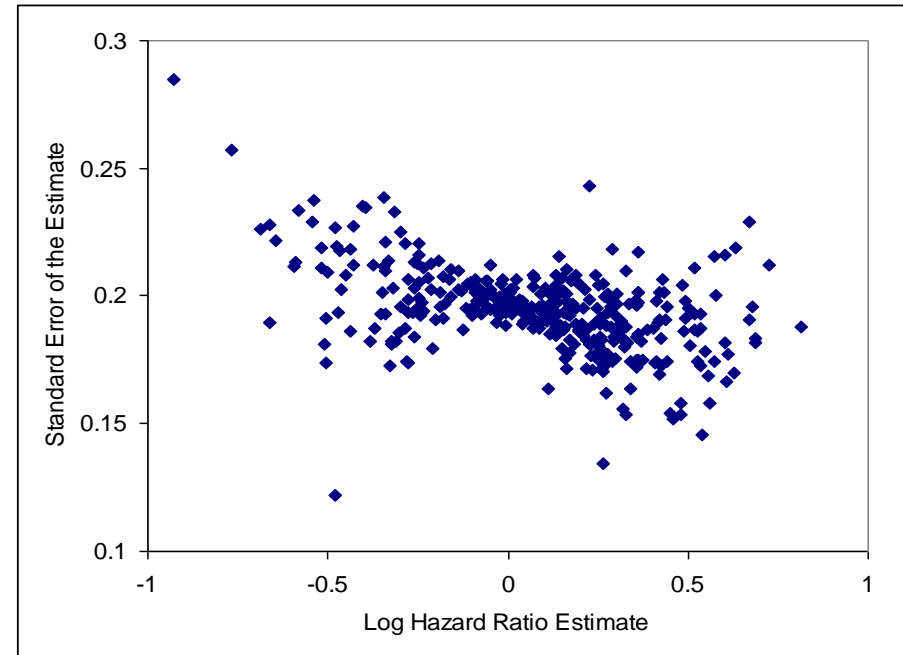
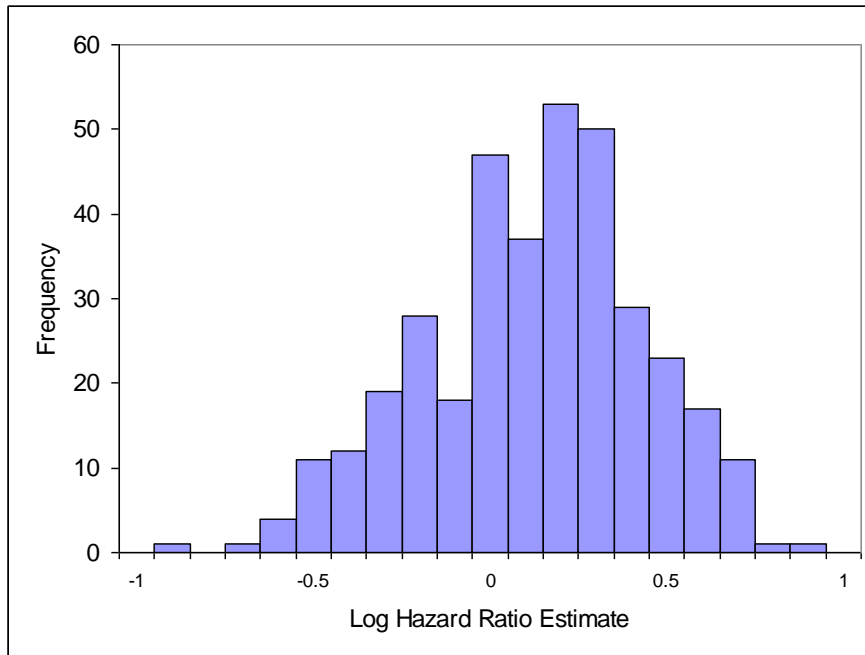
RM-corrected estimate of log hazard ratio

Uses information from all genes

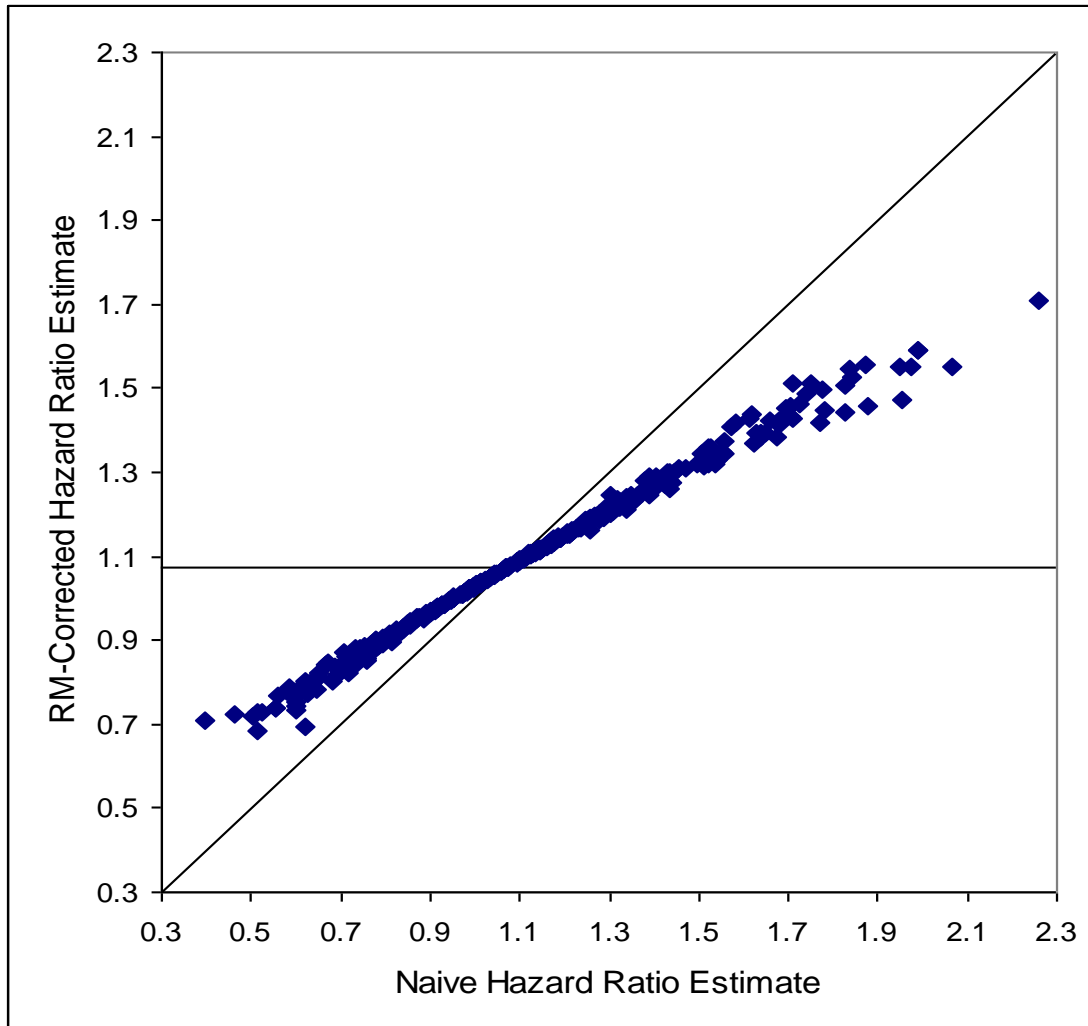
$$\hat{\beta}_j^* = \bar{\beta} + \frac{\hat{\sigma}_\beta^2}{\hat{\alpha}_\beta^2 + \sigma_{\varepsilon j}^2} \left(\hat{\beta}_j - \bar{\beta}\right)$$

Regression-to-the-mean adjustment based on individual gene estimate variability

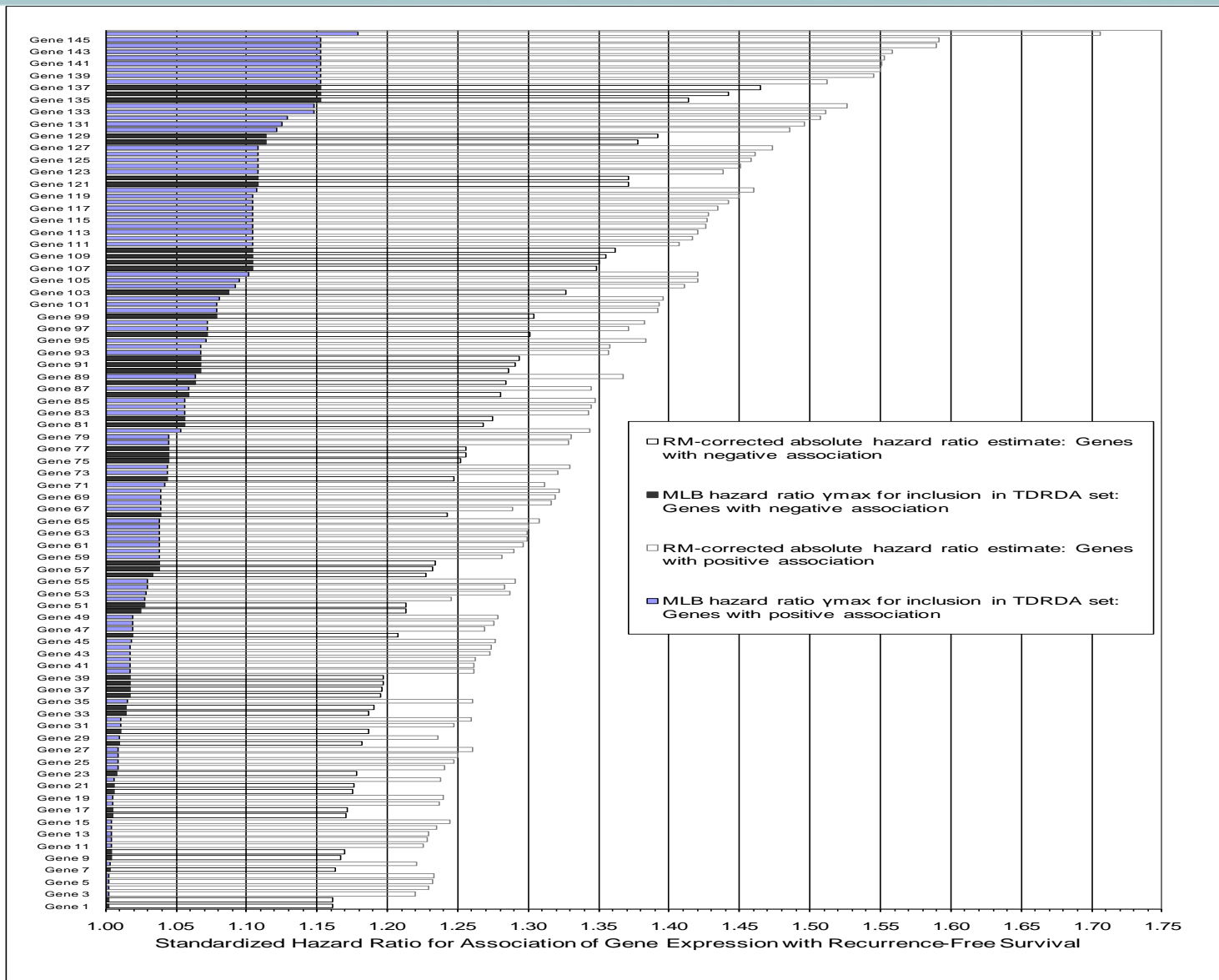
Distribution of Estimated Log Hazard Ratios Breast Cancer Study



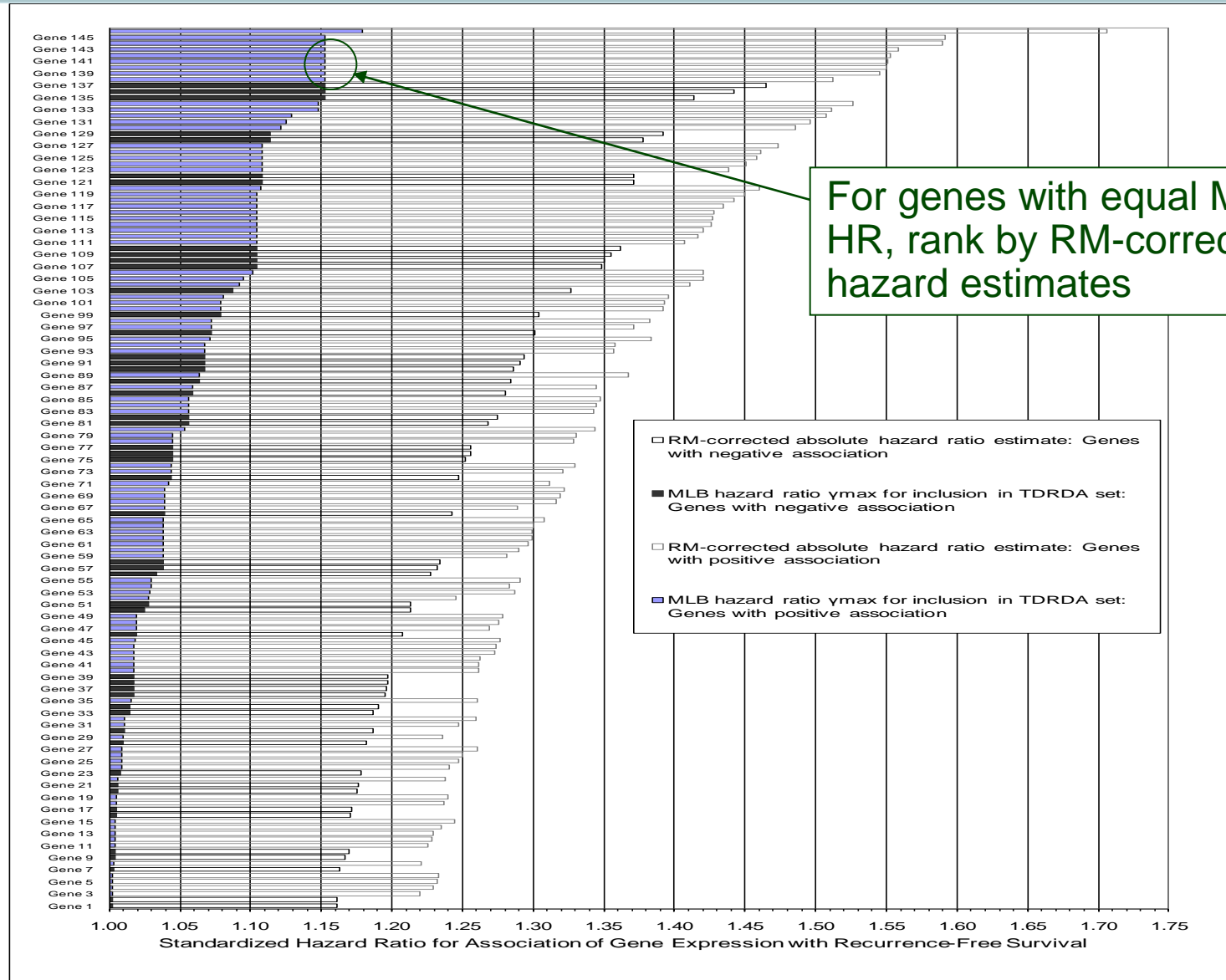
Comparison of Naïve and RM-Corrected HR Estimates from Breast Cancer Study



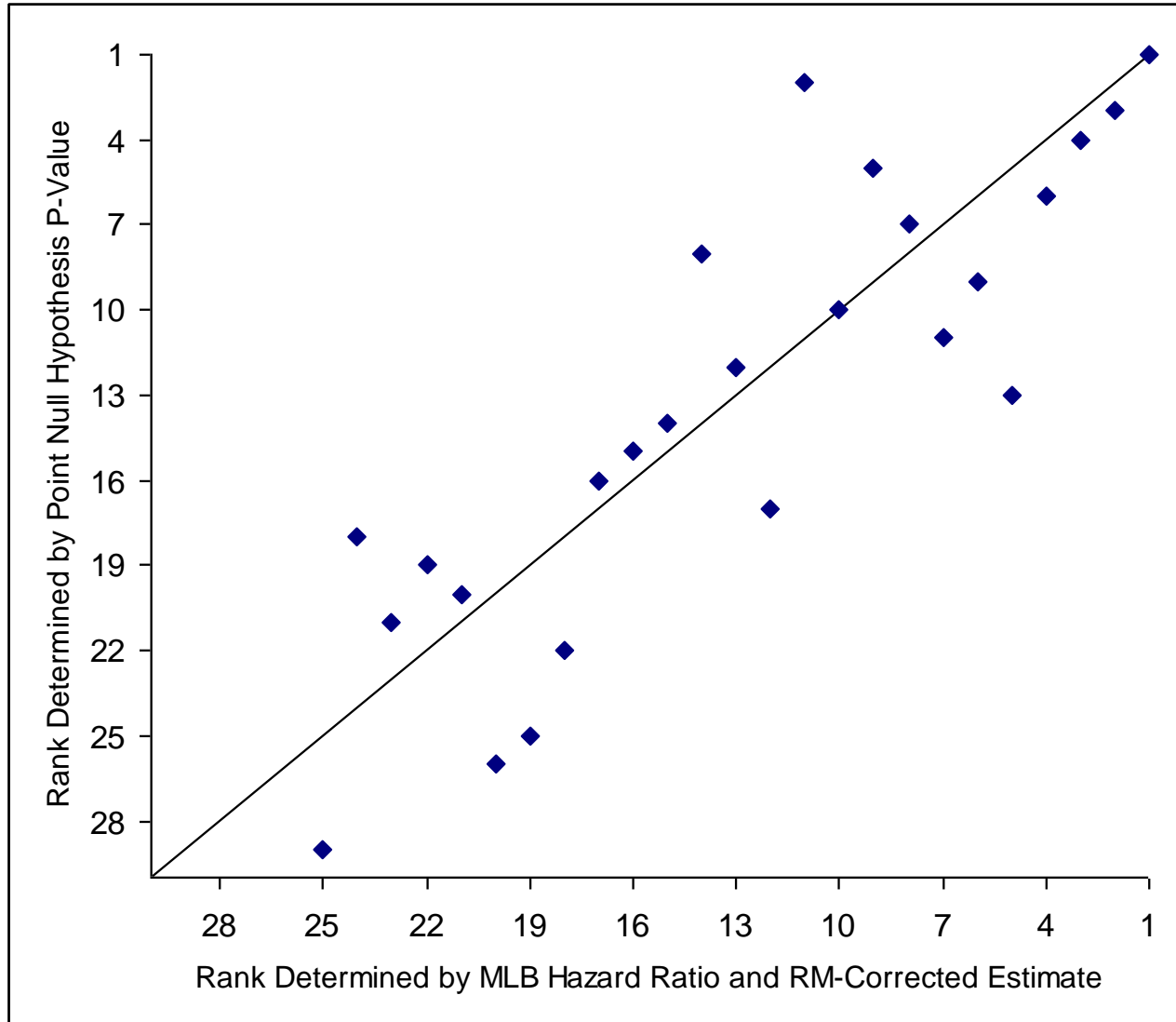
Example TDRDA Set Plot (TDR = 80%) Hazard Ratios for Recurrence of Breast Cancer



Example TDRDA Set Plot (TDR = 80%) Hazard Ratios for Recurrence of Breast Cancer



Rank by Point Null Hypothesis p -Value vs. Rank by MLB and RM-Corrected Estimate Breast Cancer Study



- TDRDA method allows rationale choice of *number* of genes selected
- Standardization of degree of association is important
 - Make scale-invariant
 - Divide each covariate by its SD
 - Divide each covariate by IQ range

- TDRDA method uses Wald tests
 - Avoid including covariates highly correlated with gene expression
- TDRDA method can be used with
 - Log hazard ratios from proportional hazard regression
 - Log odds ratios from logistic regression
 - Means from linear models
 - Any (asymptotically) normally-distributed estimate of degree of association