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
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
Gene Expression as a Predictor

- 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

![Graph showing rate of distant recurrence as a continuous function of recurrence score.](image)

**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. 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

Gene Identification → Algorithm Development → Validation

Long List

Short List

Completely Specified Predictor
Overview

- 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
False Discovery Rate Concepts
Paradigm Shift

Control of Family-wise Error Rate → Strong Control of Family-wise Error Rate → Strong Control of False Discovery Rate (FDR)

Hochberg & Tamhane 1987

Benjamini & Hochberg 1995

What: P(≥1 false rejection) → P(≥1 false rejection) → Under all combinations of true, false nulls

Expected proportion of rejections that are false

When: When all null hypotheses are true → Under all combinations of true and false nulls

Choose allowable FDR “q”

Find largest $k$ for which $P_{(k)} \leq \frac{k}{m} q$

Reject these hypotheses

$Largest \quad k$  \hspace{1cm} $Rank \quad of \quad P-value \quad (i)$

$m = \# \quad hypotheses$

$P-value \quad P_{(i)}$
Storey’s (2002) Method

\[ q = \text{allowable FDR} \]

Find largest \( k \) for which

\[ P_{(k)} \leq \frac{k}{m} \frac{q}{\hat{\pi}_0(\lambda)} \]

Reject these hypotheses

Estimated proportion of true nulls

\[ \frac{q}{\hat{\pi}_0(\lambda)} \]

\( m = \# \) hypotheses

Choose allowable FDR “q”

Find largest $k$ for which $P_{(k)} \leq \frac{k}{m} q$

Reject these hypotheses

$Largest k$  
$Rank of P-value (i)$  

$m = \# \text{ hypotheses}$
Storey’s Method: Estimation of Proportion of True Null Hypotheses

\[ \hat{\pi}_0(\lambda) = \frac{\#\{i : P_i > \lambda\}}{(1 - \lambda)m} \propto \frac{1}{\text{slope}} \]

- \( P_i \): P-value
- \( \hat{\pi}_0(\lambda) \): Estimation of proportion of null hypotheses
- \( \lambda \): Tuning parameter

The rank of the P-values is plotted against the P-values themselves. The proportion of P-values greater than \( \lambda \) is estimated as the slope of the linear regression line.
Usual Ranking Strategies for Gene Discovery

- *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

Genes identified by criterion $p<0.001$
Hypothetical standardized degree of association estimates and 99.9% confidence intervals for 5 genes

Genes identified by criterion $p<0.001$

Gene D more interesting than Gene B
Identify genes with any substantial association with clinical outcome . . .

. . . while controlling false discovery rate
True Discovery Rate
Degree of Association Analysis
### Interval Null Hypotheses (Van de Miel and Kim, 2007)

<table>
<thead>
<tr>
<th>Usual Methods</th>
<th>Van de Wiel-Kim</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Point null hypotheses</strong></td>
<td><strong>Interval null hypotheses</strong></td>
</tr>
<tr>
<td>$H_0$: $\beta = 0$ vs. $H_1$: $\beta \neq 0$</td>
<td>$H_0$: $</td>
</tr>
<tr>
<td>Estimate proportion of true nulls and FDR</td>
<td>Estimate proportion of true nulls and FDR using nonparametric</td>
</tr>
<tr>
<td>using known distribution of test statistic</td>
<td>deconvolution</td>
</tr>
<tr>
<td>under $H_0$</td>
<td></td>
</tr>
<tr>
<td>Mathematical demonstration that FDR estimate is conservative</td>
<td>Heuristic argument that FDR estimate is sensible, validation by simulation study</td>
</tr>
</tbody>
</table>
FDR Control Using Interval Null Hypotheses

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

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

$N(\hat{\beta}_j, \hat{\sigma}_{\epsilon_j}^2)$
Storey’s method applied to interval null hypothesis tests

1. Fix $\lambda$
2. Fix $\theta$
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 $P_j(\theta)$, $j = 1, 2, \ldots, m$
6. Find largest value $k(\theta)$ for which
   \[ P_{(k(\theta))}(\theta) \leq \frac{k(\theta)}{m} \frac{q}{\hat{\pi}_0(\theta; \lambda)} \]
   \[ \text{where } \hat{\pi}_0(\theta; \lambda) = \frac{\# \{ i : P_i(\theta) > \lambda \}}{(1 - \lambda)m} \]
7. Reject hypotheses $H^\theta_{(1)}, H^\theta_{(2)}, \ldots, H^\theta_{(k)}$ and identify associated genes
   \[ \text{TDRDA}(\theta; 1-q) \]
Storey’s method applied to interval null hypothesis tests

$q = \text{allowable FDR}$

Find largest $k(\theta)$ for which $P_{(k)}(\theta) \leq \frac{k(\theta)}{m} \frac{q}{\hat{\pi}_0(\theta; \lambda)}$

Reject these hypotheses

$Largest k(\theta)$

$m = \# \text{ hypotheses}$

$Largest (k(\theta), t)$

$P_{(i)}(\theta)$
Storey’s Method: Estimation of Proportion of True Interval Null Hypotheses

\[ P_{(i)}(\theta) \]

\[ \hat{\pi}_0(\theta; \lambda) = \frac{\#\{i : P_i(\theta) > \lambda\}}{(1 - \lambda)m} \propto \frac{1}{\text{slope}} \]

\[ m = \# \text{ hypotheses} \]
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” $\theta$ not known:
  - Vary $\theta$ and generate all the sets $\text{TDRDA}(\theta;1-q)$
  - Sort genes by maximum lower bound (MLB) $\theta$ for which each is included in $\text{TDRDA}(\theta;1-q)$

$$\theta^\text{max}_j = \max \{ |\theta| : H^\theta_j \text{ is rejected} \}$$
Monotonicity Property of TDRDA Sets

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

**Note:** Requires fixed $\lambda$

**Corollaries**

- Gene $j$ will be included in every set $\text{TDRDA}(\theta;1-q)$ for every $\theta \in [0, \theta_j^{\text{max}}]$

- $\text{TDRDA}(\theta;1-q) \subseteq \text{TDRDA}(0;1-q)$ for all $\theta>0$
  - TDRDA sets refine the set of genes identified by Storey’s procedure with Wald test of point null hypothesis
Example Calculation
Breast Cancer Study

- 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)

MLB hazard ratio $\gamma_{\text{max}}$ for inclusion in TDRDA set:

- Genes with negative association
- Genes with positive association
Estimates Corrected for Regression to the Mean
Key Problem: Regression to the Mean

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

- Study 1
- Study 2
Key problem: Regression to the mean

Idea: model “true” log hazard ratios $\beta_j$, $j = 1, 2, \ldots, m$ as (not necessarily independent) sample from distribution $N(\mu_\beta, \sigma^2_\beta)$

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

=>$\text{True log HR and estimate have bivariate normal distribution}$

$$
\begin{pmatrix}
  \hat{\beta}_j \\
  \beta_j
\end{pmatrix}
\sim
N
\begin{pmatrix}
  \mu_\beta \\
  \mu_\beta
\end{pmatrix},
\begin{bmatrix}
  \sigma^2_\beta + \sigma^2_\varepsilon & \sigma^2_\beta \\
  \sigma^2_\beta & \sigma^2_\beta
\end{bmatrix}
$$

which implies

$$
E(\beta_j | \hat{\beta}_j) = \mu_\beta + \frac{\rho \sigma_\beta}{\sqrt{\left(\sigma^2_\beta + \sigma^2_\varepsilon\right)}} (\beta_j - \mu_\beta) = \mu_\beta + \frac{\sigma^2_\varepsilon}{\sigma^2_\beta + \sigma^2_\varepsilon} (\hat{\beta}_j - \mu_\beta)
$$
Regression-to-the-mean-corrected estimate of log hazard ratio

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

\[
E\left( \beta_i - \mu_\beta \right)^2 = \sigma_\beta^2 + \sigma_{\epsilon i}^2 \implies \sum_{i=1}^m E\left( \beta_i - \mu_\beta \right)^2 = m\sigma_\beta^2 + \sum_{i=1}^m \sigma_{\epsilon i}^2
\]

Estimate of the variance of the true log hazard ratios

\[
\sigma_\beta^2 = \frac{1}{m-1} \sum_{i=1}^m \left( \beta_i - \bar{\beta} \right)^2 - \bar{\sigma}_\epsilon^2 \quad \text{where} \quad \bar{\sigma}_\epsilon^2 = \frac{1}{m} \sum_{i=1}^m \sigma_{\epsilon i}^2 \quad \bar{\beta} = \frac{1}{m} \sum_{i=1}^m \beta_i
\]

RM-corrected estimate of log hazard ratio

\[
\hat{\beta}_j = \bar{\beta} + \frac{\hat{\sigma}_\beta^2}{\sigma_\beta^2 + \sigma_{\epsilon j}^2} \left( \beta_j - \bar{\beta} \right)
\]
Regression-to-the-mean-corrected estimate of log hazard ratio

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

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

Estimate of the variance of the true log hazard ratios

$$\hat{\alpha}_\beta = \frac{1}{m-1} \sum_{i=1}^{m} \left(\hat{\beta}_i - \bar{\beta}\right)^2 - \bar{\sigma}_\epsilon^2$$

where

$$\bar{\alpha}_\epsilon = \frac{1}{m} \sum_{i=1}^{m} \sigma_{\epsilon_i}^2 \quad \bar{\beta} = \frac{1}{m} \sum_{i=1}^{m} \beta_i$$

RM-corrected estimate of log hazard ratio

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

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

Uses information from all genes
Distribution of Estimated Log Hazard Ratios
Breast Cancer Study

![Histogram of Log Hazard Ratio Estimate](image1)

- Frequency
- Log Hazard Ratio Estimate

![Scatter plot of Standard Error of the Estimate](image2)

- Standard Error of the Estimate
- Log Hazard Ratio Estimate
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

Standardized Hazard Ratio for Association of Gene Expression with Recurrence-Free Survival
Example TDRDA Set Plot (TDR = 80%)
Hazard Ratios for Recurrence of Breast Cancer

For genes with equal MLB HR, rank by RM-corrected hazard estimates.
Rank by Point Null Hypothesis $p$-Value vs. Rank by MLB and RM-Corrected Estimate

Breast Cancer Study

Rank Determined by MLB Hazard Ratio and RM-Corrected Estimate

Rank Determined by Point Null Hypothesis P-Value

![Graph showing the relationship between ranks determined by MLB hazard ratio and RM-corrected estimate vs. ranks determined by point null hypothesis p-value.](image)
Discussion

• 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
Discussion

- 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