# Looking at the Other Side of Bonferroni

Caitlin McHugh

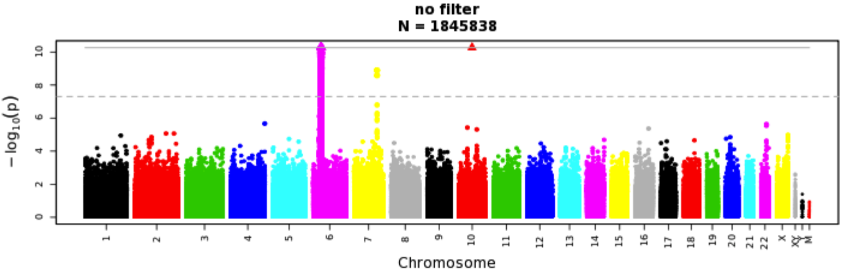
Department of Biostatistics
University of Washington

24 May 2012

# Multiple Testing: Control the Type I Error Rate

- When analyzing genetic data, one will commonly perform over 1 million (and growing) hypothesis tests.
- In categorical data analysis, one may want to test all pairwise combinations.
- How do we ensure we are properly controlling for the number of false rejections?

# 2.5 Million Hypothesis Tests



# Recall: error rates

type I error $\mathbb{P}(\text{reject } H_0 | H_0 \text{ is true}) \leq \alpha$

family-wise error rate $\text{FWER} = \mathbb{P}(\# \text{ false pos} \geq 1)$
This is the probability of one or more false positives.

per family error rate $\text{PFER} = \mathbb{E}(\# \text{ false pos})$
This is the expected number of false positives.

false discovery rate $\text{FDR} = \mathbb{E}(\# \text{ false pos}/\text{total } \# \text{ rejected})$
This can be thought of as the average proportion of null hypotheses that are falsely rejected.

# How it all fits together

| | decide true | decide false | |
|---|---|---|---|
| $H_0$ true | $U$ | $V$ | $m_0$ |
| $H_0$ false | $R$ | $S$ | $m - m_0$ |
| | $m - T$ | $T$ | $m$ |

- $V$ denotes a type I error.
- The FWER is $\mathbb{P}(V \geq 1)$.
- The PFER is $\mathbb{E}(V)$.
- The FDR is $\mathbb{E}(V/T)$.

# Bonferroni and Benjamini-Hochberg (BH) procedures

- Bonferroni correction calculates

$$\alpha^* = \alpha/m$$

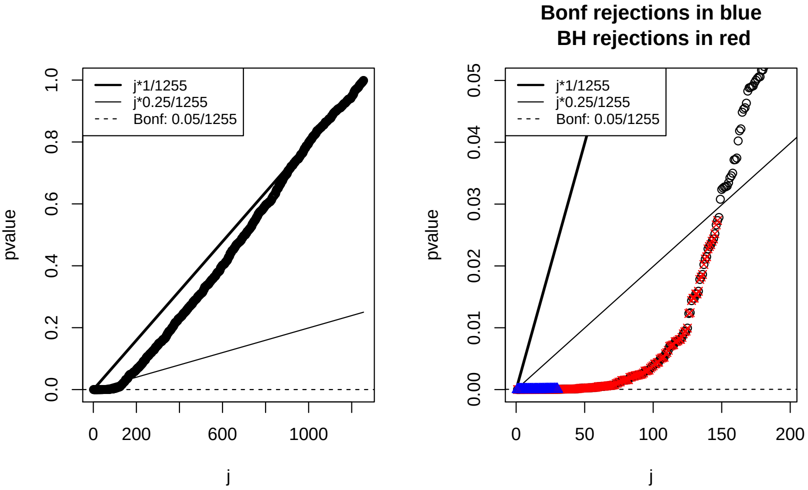
  and controls the FWER or PFER.

- BH correction orders the $p$-values in decreasing order, and for each $i$ starting at the largest value, finds the point at which

$$p_{(i)} \leq \frac{\alpha i}{m}$$

  and this set of decisions controls the FDR.

# A BH example




at $\alpha$=0.25, reject $H_{0j}$ for all $j \leq 147$ using BH

for 'usual' Bonferroni correction, reject 30 hypotheses

# When test statistics are correlated

- Under the most extreme case, with perfect correlation, it is as if one test is performed $m$ times.
- With Bonferroni correction, for any $i \in 1 \ldots m$

$$\begin{aligned} \mathbb{P}(p_i \leq \alpha/m) &= \mathbb{P}(p_1 \leq \alpha/m) \\ &= \alpha/m \end{aligned}$$

which is more stringent than if we just used $\alpha$ in the presence of this correlation.

# FWER, Bonferroni and FDR

- With FWER, 5 and 1000 false positives are equally ‘bad.’
- With FDR, the ‘badness’ depends on the number of rejections made.
- Using Bonferroni to control the FWER is a conservative measure in terms of controlling the presence of any type I errors.
- Could we use Bonferroni to control the expected false positives?

# Bonferroni can control the PFER

Applying the Bonferroni correction to the desired PFER threshold, $\gamma$, when performing $m$ hypothesis tests, we get

$$\begin{aligned} \text{PFER} &= \mathbb{E}(\#\text{ false positives}) \\ &= \mathbb{E}(\sum_{i \in \mathcal{T}} \mathbb{I}_{p_i \leq \gamma/m}) \\ &= \sum_{i \in \mathcal{T}} \mathbb{P}(p_i \leq \gamma/m) \\ &\leq m_0 \frac{\gamma}{m} \\ &\leq \gamma \end{aligned}$$

where $\mathcal{T}$ is the set of $m_0$ true null hypotheses and $p_i$ are calculated $p$-values.

- This is robust to dependence of test statistics.
- The last line is less dramatic when $m_0 \approx m$.

# Simulation Studies: Goal

- With simulated data, I (and Gordon et al) show that the Bonferroni and BH procedures are comparable, for intelligently chosen PFER and FDR thresholds.

# Simulation Studies: The Data

- Simulate 1255 gene expression values, measured for 50 individuals.
- 2 measurements per individual where 125 of the 1255 genes have a different mean.
- Generate a *p*-value for each gene from a standard t-test; 125 of them should be significant.
- Count the number of true and false rejections when using the Bonferroni and BH procedures, at various thresholds.

# Equating Error Rates

How can we make the Bonferroni and BH procedures comparable?

- Define initial thresholds $\gamma_i$ ranging from 0 to 100 and thresholds $\beta_i = \frac{\gamma_i}{125+\gamma_i}$.
- Find FDR and PFER using Bonferroni$^{\gamma_i}$.
- Find FDR and PFER using BH$^{\beta_i}$.
- Do this 500 times over and define the means as $\hat{\text{FDR}}_{BH^{\beta_i}}, \hat{\text{FDR}}_{Bonf^{\gamma_i}}, \hat{\text{PFER}}_{BH^{\beta_i}}, \hat{\text{PFER}}_{Bonf^{\gamma_i}}$.

# Equating Error Rates

- For 'equalized FDR' define

$$\gamma_j^* = \underset{1 \le i \le 280}{\text{argmin}} |\hat{\text{FDR}}_{BH^{\beta_j}} - \hat{\text{FDR}}_{Bonf^{\gamma_i}}|$$

for $j = 1, \ldots, 280$.

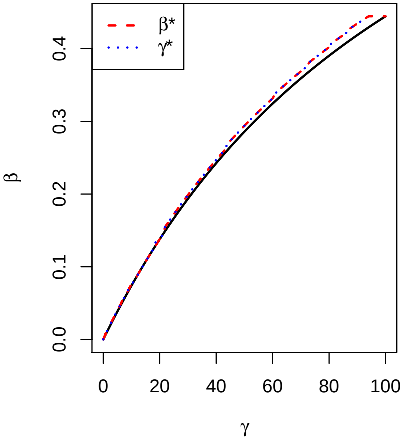
- For 'equalized PFER' define

$$\beta_j^* = \underset{1 \le i \le 280}{\text{argmin}} |\hat{\text{PFER}}_{Bonf^{\gamma_j}} - \hat{\text{PFER}}_{BH^{\beta_i}}|$$

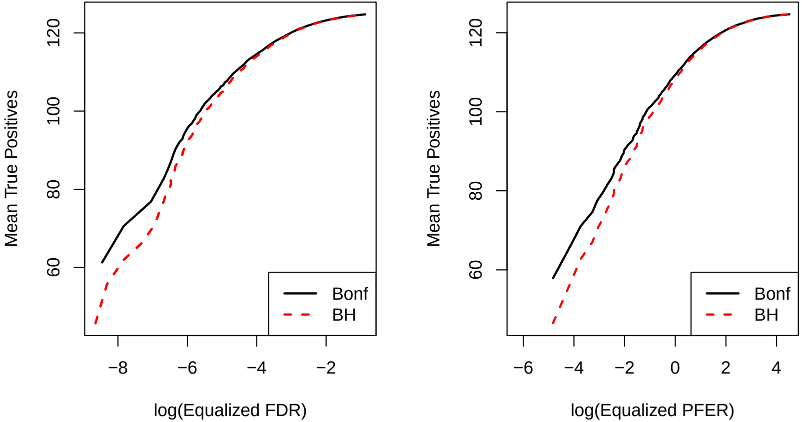
for $j = 1, \ldots, 280$.

# Equating Error Rates

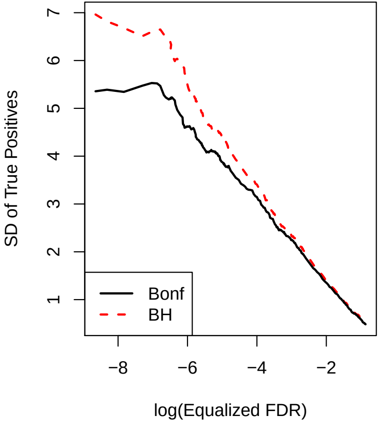
- With FDR as equalizer, use Bonferroni$^{\gamma^*}$ and BH$^{\beta}$.
- For PFER as equalizer, use Bonferroni$^{\gamma}$ and BH$^{\beta^*}$.

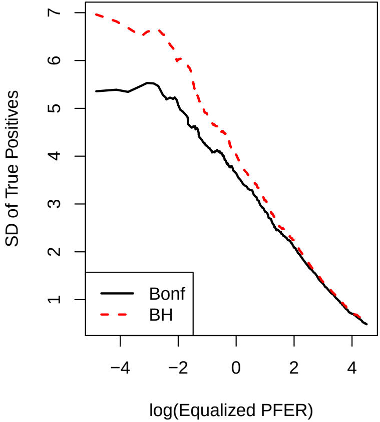


# Simulation Results: Power

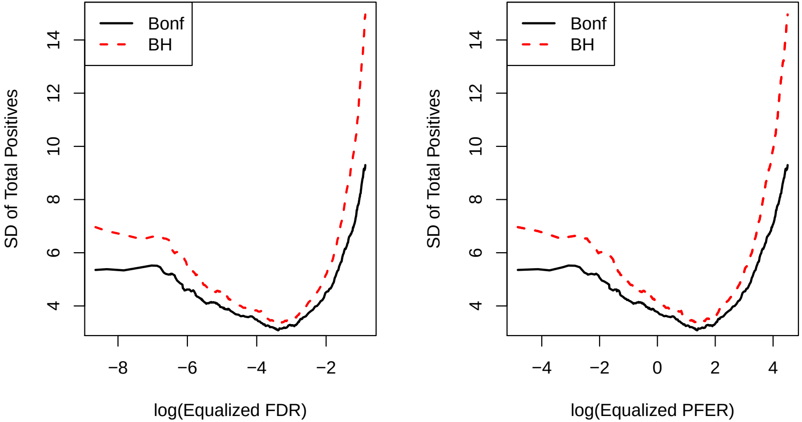


# Simulation Results: Stability



# Simulation Results: Stability



# Simulation Thoughts

- With thresholds chosen correctly, the MTPs look quite similar.
- The number of outcomes rejected are highly correlated among the two procedures.
- Bonferroni is more stable when looking at the standard deviation of either the true positives or total positives.
- Bonferroni is more powerful than the BH procedure, here.

# In Conclusion

- Choose the rate you want to control; do you have the budget to follow up a fixed number of 'hits?' Or can you only follow up those with a $p^{exciting*}$ result?
- Choose your favorite MTP from the Bonferroni or BH procedure and rest assured your results will be in line with your expectations.

* borrowing Ken's jargon

# Final Steps

- Simulate *correlated* data, and calculate the same metrics as presented here.