Looking at the Other Side of Bonferroni

Caitlin McHugh

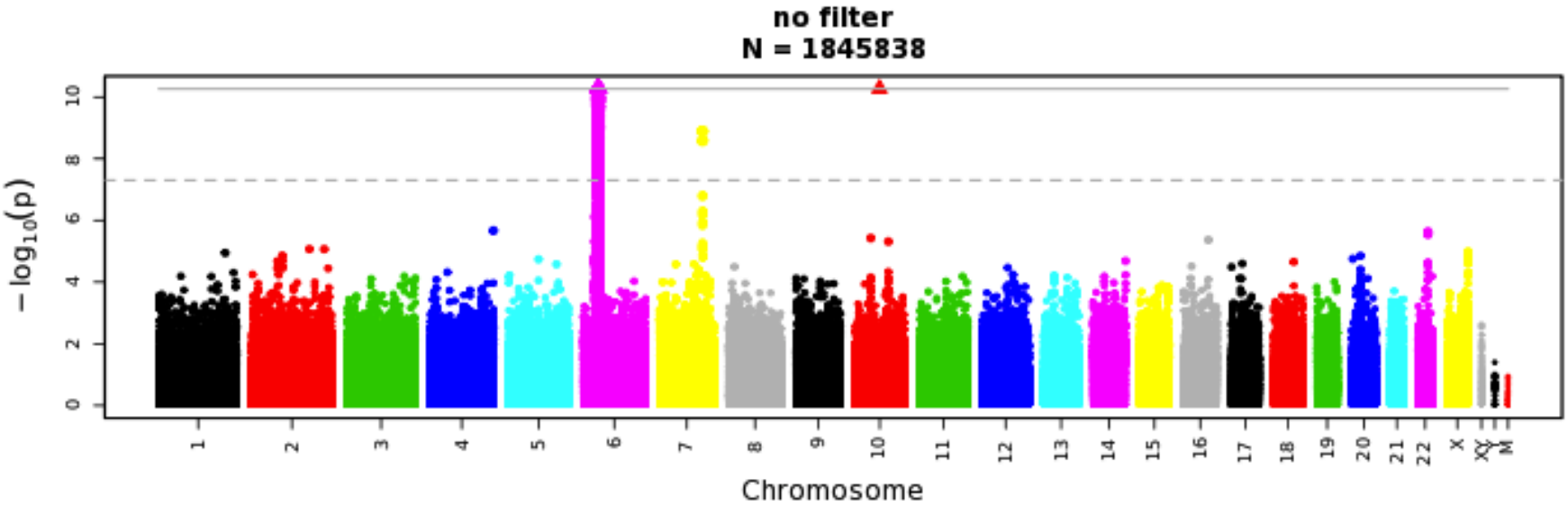
Department of Biostatistics
University of Washington

1 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 $\mathbb{E}(V/T)$.

FWER vs. PFER

- Define $\mathcal{A}$ to be the acceptance interval for some summary statistic V_g that gives the desired α risk for one hypothesis test.
- Then, define a rule for all $1 \leq g \leq m$ such that

$$\begin{aligned} &\text{if } V_g \in \mathcal{A}, && \text{can't reject } H_0 \\ &\text{if } V_g \notin \mathcal{A}, && \text{reject } H_0 \end{aligned}$$

FWER vs. PFER

- For some test where we know the null hypothesis is true, the probability that our summary measure V_g will fall in the acceptance interval $\mathcal{A}$ is

$$\begin{aligned} \mathbb{P}(V_g \in \mathcal{A} | g \in \mathcal{T}) &= \int_{\mathcal{A}} f_0(v) dv \\ &= 1 - \alpha \end{aligned}$$

  for all g, where f_0 is the null pdf, and $\mathcal{T}$ is the set of indices of the true null hypotheses.
- We choose $\mathcal{A}$ to be the smallest interval that satisfies the above equation.

FWER vs. PFER

- We can specify the type I error probability for any g to be

$$\alpha = \mathbb{P}(V_g \notin \mathcal{A} | g \in \mathcal{T})$$

- Then, we can also see that

$$\begin{aligned} \text{FWER} &= \mathbb{P}(V_g \notin \mathcal{A} \text{ for at least one } g \in \mathcal{T}) \\ &\leq \sum_{g \in \mathcal{T}} \mathbb{P}(V_g \notin \mathcal{A} | g \in \mathcal{T}) \\ &= m_0 \mathbb{P}(V_g \notin \mathcal{A} | g \in \mathcal{T}) \\ &= m_0 \alpha \end{aligned}$$

- $\text{FWER}/m_0 \leq \alpha$.
- The unadjusted significance level can be defined by the desired family wise error rate divided by the total number of tests for which the null hypothesis is true.

FWER vs. PFER

- If we define $\mathcal{A}$ as shown on the previous slide such that FWER$\leq m_0\alpha$ then
- FWER $\leq \alpha$, so our definition is consistent with the desired objective.

FWER vs. PFER

- Finally, we get that

$$\begin{aligned} \text{PFER} &= \mathbb{E}(V) \\ &= \alpha m_0 \\ &\geq \text{FWER} \\ \text{PFER}/m_0 &= \alpha \end{aligned}$$

  where V is the number of false positives, α is the overall significance level and m_0 is the number of true null hypotheses.
- The expected number of false positives, the PFER, is equal to the type I error rate divided by the true number of null hypotheses.

Bonferroni and Benjamini-Hochberg (BH) procedures

- Bonferroni correction calculates

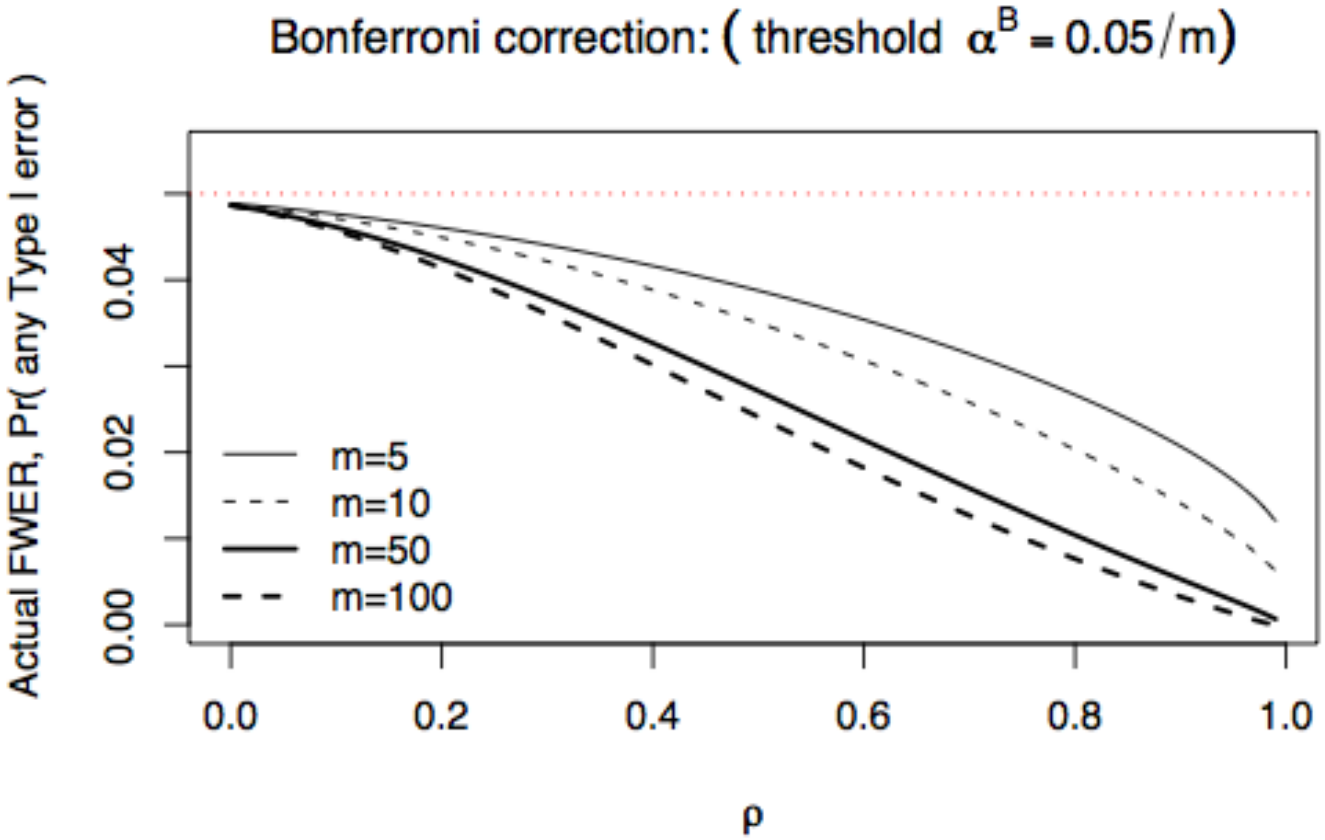
$$\alpha^* = \alpha/m$$

  and controls the type I error rate.
- 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 controls the FDR.

How does dependence change the FWER?



- Dependence makes the actual type I error less than desired.
- As m and/or ρ increases, this becomes worse.

Plot courtesy of Ken Rice

Simulation Studies

- 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 rejections when using the both the Bonferroni and BH procedures.

Next Steps

With these results, I aim to show when choosing the error rate thresholds appropriately, the Bonferroni and BH procedures are comparable. This will be reproducing the results from the paper [1].

References

Alexander Gordon, Galina Glazko, Xing Qiu, and Andrei Yakovlev.
Control of the mean number of false discoveries, Bonferroni and stability of multiple testing.
The Annals of Applied Statistics, 1(1):179–190, 2007.