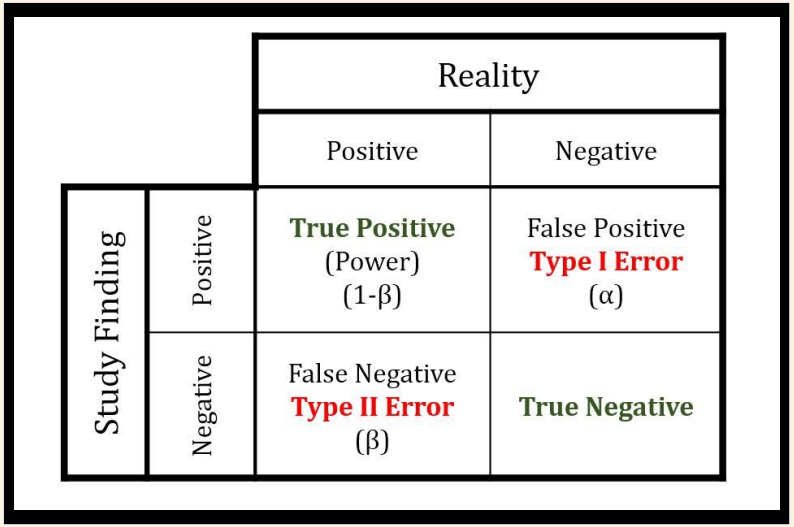
Multiple Testing in Genetic Research

Key Concepts:

- A **Type 1 Error (α)** occurs when we reject the null if it is true.
- Valid tests satisfy P(T1 Error) <= α
- The **power (β)** of a test is P(Reject the null when the alternative is true).



Setting The Stage

H_0^1 = Gene 1 not related to hair color

H_0^2 = Gene 2 not related to hair color

$\vdots$

H_0^{12000} = Gene 12000 not related to hair color

H_1^1 = Gene 1 is related to hair color

H_1^2 = Gene 2 is related to hair color

$\vdots$

H_1^{12000} = Gene 12000 is related to hair color

Suppose that H_0^k holds for all k e.g.hair color not related to any gene

P(T1 Error for test k) $\leq \alpha$ for all k

However, P(any Type 1 Error for all 12,000 tests) $\approx 100\%$

If every hypothesis corresponds to one gene in a genetic study, conclusions may be incorrect because there are so many tests.

Why is this bad?

- If we present all test results together, we cannot say anything meaningful about any single one.
- If 12,000 hypothesis tests are ran- are we rejecting a null because we discovered something meaningful, or is it because we ran so many tests?
- A correction for multiple testing is a process by which we can present the results of many tests together in a “meaningful” way. *(In this talk: control false discovery rate)*

False Discovery Rate (FDR)

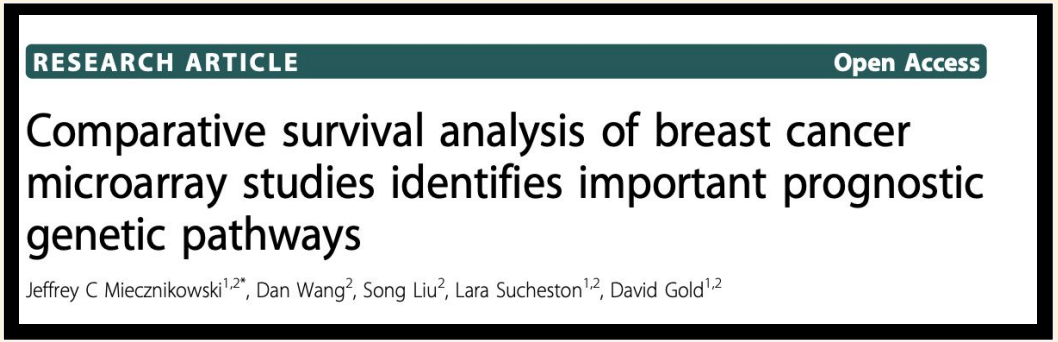
- Proportion of false discoveries (Type 1 Errors) among all discoveries (any rejection of a null).
- For example, controlling the FDR at 0.2 means that on average, no more than 20 out of 100 significant results will be false positives.

Benjamini-Hochberg:

1. Rank p-values corresponding to each gene from smallest to largest.
2. $\text{Threshold}_i = \frac{i}{m} * Q$ where Q=FDR, i= rank, m=total tests
3. Find largest p-value where $p_i \leq \text{Threshold}_i$ and name it p_k
4. All genes with $p_i \leq p_k$ declared significant.

Genetic Study Design/Background

- Investigating predictors of breast cancer from 5 breast cancer related datasets (Miecznikowski, et al, 2010).
- Focus on one dataset with 12,649 genes.
- Tested associations between genes and survival time.
- Controlled False Discovery Rate using Benjamini Hochberg at 0.2


RESEARCH ARTICLE Open Access

Comparative survival analysis of breast cancer microarray studies identifies important prognostic genetic pathways

Jeffrey C Miecznikowski[1,2*], Dan Wang[2], Song Liu[2], Lara Sucheston[1,2], David Gold[1,2]

Results of the study

- 3,246 genes were found to be significant after controlling the FDR.
- Roughly 650 may be false positives.

Interpretation of Results After FDR Control

- Recall: Our goal was to identify potential genes that could be linked to breast cancer, but we were worried about the large number of tests we were doing.
- The interpretation of our results after controlling FDR is that after focusing our attention to all of the genes seemingly related with breast cancer survival time, there is a 20% chance that any given gene is not actually related with breast cancer survival time.

Conclusions

- Simultaneously testing tens of thousands of genes increases the risk of false positives.
- FDR controls expected proportion of false positives.
  - Less conservative than other measures.
- FDR balances discovery with high reliability.