Multiple testing

Mikhail Dozmorov

2021-03-17

Multiple testing problem

- With thousands of genes, we're not testing one hypothesis, but many hypotheses – one for each gene
- Analysis of 20,000 genes using commonly accepted significance level $\alpha = 0.05$ will identify 1,000 differentially expressed genes simply by chance
- If probability of making an error in one test is 0.05, probability of making at least one error in ten tests is

$$(1 - (1 - 0.05)^{10}) = 0.40126$$

Naomi Altman & Martin Krzywinski "Points of significance: P values and the search for significance", Nat. Methods 2016, http://www.nature.com/nmeth/journal/v14/n1/full/nmeth.4120.html

Multiple Hypothesis Testing for differential expression detection

- The test statistics and hence the p-values are likely correlated due to co-regulation of the genes.
- Would like multiple testing procedures that take into account the dependence structure of the genes.
- This could be accomplished by estimating the joint null distribution of the unadjusted, unknown p-values.

Permutation based methods

Permutation based adjusted p-values

- Under the H_0, the joint distribution of the test statistics can be estimated by permuting the columns of the gene expression matrix
- Permuting entire columns creates a situation in which membership to the groups being compared is independent of gene expression but preserves the dependence structure between genes

Permutation based methods

- Permutation algorithm for the b^{th} permutation, $b = 1, ..., B$

1. Permute the n columns of the data matrix X
2. Compute test statistics $t_{j,b}$ for each hypothesis (gene, $j = 1, ..., g$)

- The permutation distribution of the test statistic T_j for hypothesis H_j is given by the empirical distribution of $t_{j,1}, ..., t_{j,B}$

Permutation based methods

- For two-sided alternative hypotheses, the permutation p-value for hypothesis H_j is

$$p_j^* = \frac{\sum_{b=1}^{B} I(|t_{j,b}| \geq |t_j|) + 1}{B + 1}$$

where $I()$ is the indicator function, equaling 1 if the condition in parentheses is true and 0 otherwise.

Permutation based methods

- Permutation method permits estimation of the joint null distribution of the unadjusted unknown p-values.
- Dependency structure between the genes is preserved.
- May suffer from a granularity problem (when two groups, should have 6 arrays in each group to use permutation based method). $\frac{n!}{n_1!n_2!}$ ways of forming two groups
- The permutation p-value cannot be smaller than $1/(B+1)$, meaning that a large number of permutations are needed

Question: How many samples per group you need to get ~1,000 distinct permutations?

Increasing permutation efficiency: Adaptive scheme

- Run permutations until a given number X (typically, 100) of null correlations stronger than the observed one are found
- Limit permutations to no more than Y (typically, 100,000) permutations in total
- The adjusted p-value is calculated as

$$p_{adj} = \frac{min(I, X) + 1}{min(B, Y) + 1}$$

- Still unable to provide adjusted p-values below $1/(M + 1)$

Increasing permutation efficiency: beta approximation

- Order statistics of independent uniformly distributed random variables are beta-distributed
- The permutation p-values are also beta-distributed (Dudbridge and Koeleman, 2004)
- Perform Y permutations, estimate the parameters of beta distribution using maximum likelihood
- Approximate permutation p-value from the best p-value obtained during Y permutations

Ongen, Halit, Alfonso Buil, Andrew Anand Brown, Emmanouil T. Dermitzakis, and Olivier Delaneau. “Fast and Efficient QTL Mapper for Thousands of Molecular Phenotypes.” Bioinformatics (Oxford, England) 32, no. 10 (May 15, 2016): 1479–85. https://doi.org/10.1093/bioinformatics/btv722.

Results of Multiple hypothesis testing

Assume we are testing $H_1, H_2, ..., H_m$. m_0 - number of true null hypotheses

	Declared non-significant	Declared significant
# True null hypothesis	**U**	**T**
# Non-true null hypothesis	**V**	**S**
	m0	m-m0

- **U, S** - True negatives/positives - unobservable random variable
- **V** - False positives [Type I errors] - unobservable random variable
- **T** - False negatives [Type II errors] - unobservable random variable
- **R** - All positives (# of rejected null hypotheses) - observable

Error rates

- False Discovery rate (FDR)

$$E\left[\frac{False\ Discoveries}{True\ Discoveries}\right]$$

- Family wise error rate (FWER)

$$Pr(Number\ of\ False\ positives \geq 1)$$

- Expected number of false positives

$$E[Number\ of\ False\ positives]$$

Interpretation

Suppose 550 out of 10,000 genes are significant at $\alpha = 0.05$

P-value < 0.05

- Expect $0.05 * 10,000 = 500$ false positives

False Discovery Rate < 0.05

- Expect $0.05 * 550 = 27.5$ false positives

Family Wise Error Rate < 0.05

- The probability of at least 1 false positive is ≤ 0.05

Multiple Hypothesis Testing: FWER

- Given p is the probability of error, $1 - p$ is the probability of correct choice in one test
- $1 - (1 - p)^g$ is the probability of one error in g tests

Multiple Hypothesis Testing: FWER

- Given p is the probability of error, $1 - p$ is the probability of correct choice in one test
- $1 - (1 - p)^g$ is the probability of one error in g tests

Sidak single step

- Testing g null hypotheses
- Reject any H_i with $p \leq 1 - \sqrt[g]{1 - \alpha}$
- When testing 22,283 genes for differential expression, use the following cutoff:

$$1 - \sqrt[22,283]{1 - 0.05} = 0.000002302$$

Multiple Hypothesis Testing: FWER

Bonferroni procedure

- Testing g null hypothesis
- Reject any H_i with $p_i \leq \alpha/g$
- $0.05/22{,}283 = 0.0000022$

Multiple Hypothesis Testing: FWER

Bonferroni procedure

- Testing g null hypothesis
- Reject any H_i with $p_i \leq \alpha/g$
- $0.05/22{,}283 = 0.0000022$
- Controls the FWER to be $\leq \alpha$ and to be equal to α if all hypotheses are true.
- As the number of hypotheses increases, the average power for an individual hypothesis decreases
- Very conservative; no attempt to incorporate dependence between tests

Multiple Hypothesis Testing: FWER

Holm step-down procedure

1. Order the p-values and hypotheses $P_1 \geq ... \geq P_g$ corresponding to $H_1, ..., H_g$
2. Let $i = 1$
3. If $P_{g-i+1} > \alpha/(g - i + 1)$ then accept all remaining hypotheses H_{g-i+1} and STOP
4. If $P_{g-i+1} \leq \alpha/(g - i + 1)$ then reject H_{g-i+1} and increment i, then return to step 3.

Multiple Hypothesis Testing: FWER

Sidak step down

1. Order the p-values and hypotheses $P_1 \geq ... \geq P_g$ corresponding to $H_1, ..., H_g$
2. Let $i = 1$
3. If $P_{g-i+1} > 1 - \sqrt[g-i+1]{1-\alpha}$ then accept all remaining hypotheses H_{g-i+1} and STOP
4. If $P_{g-i+1} \leq 1 - \sqrt[g-i+1]{1-\alpha}$ then reject H_{g-i+1} and increment i, then return to step 3.

Multiple Hypothesis Testing: FWER

Hochberg step up

1. Order the p-values and hypotheses $P_1 \geq ... \geq P_g$ corresponding to $H_1, ..., H_g$
2. Let $i = 1$
3. If $P_i \leq \alpha/i$ then reject all remaining hypotheses $H_i, ..., H_g$ and STOP
4. If $P_i > \alpha/i$ then accept H_i and increment i, then return to step 3.

Considerations for controlling the FWER

- Control over FWER is only appropriate in situations where the intent is to identify only a small number of genes that are truly different.
- Otherwise, the severe loss in power in controlling FWER is not justified.

Considerations for controlling the FWER

- Approaches that set out to control the FWER seek to control the probability of at least one false positive regardless of the number of hypotheses being tested.
- When the number of hypotheses N is very large, this may be too strict = too many missed findings.

False discovery rates: FDR

- It may be more appropriate to emphasize the proportion of false positives among the differentially expressed genes.
- The expectation of this proportion is the false discovery rate (FDR) (Benjamini & Hochberg, 1995)

False discovery rate | Benjamini and Hochberg 1995

Definition: FDR is the proportion of false positives among all positives

$$FDR = E\left[\frac{V}{V+S}\right] = E\left[\frac{V}{R}\right]$$

- Select the desired proportion q, e.g., 0.1 (10%)
- Rank the p-values $p_1 \leq p_2 \leq ... \leq p_m$.
- Find the largest rank i such that $p_i \leq \frac{i}{m} * q$
- Reject null hypotheses corresponding to $p_1, ..., p_i$

False positive vs. False discovery rates

False positive rate is **the rate at which truly null genes are called significant**

$$FPR \approx \frac{\mathit{false\ positives}}{\mathit{truly\ null}} = \frac{V}{m_0}$$

False discovery rate is **the rate at which significant genes are truly null**

$$FDR \approx \frac{\mathit{false\ positives}}{\mathit{called\ significant}} = \frac{V}{R}$$

False Discovery Rates

Two procedures for controlling FDR:

- Fix the acceptable FDR level σ *a priori*, then find a data-dependent threshold so that the $FDR \geq \sigma$. (Benjamini & Hochberg)
- Fix the threshold rule and then form an estimate of the FDR whose expectation is $\geq$ the FDR rule over the significance region. (Storey)

Storey's positive FDR (pFDR)

$$BH:\ FDR = E\left[\frac{V}{R}|R > 0\right]p(R > 0)$$

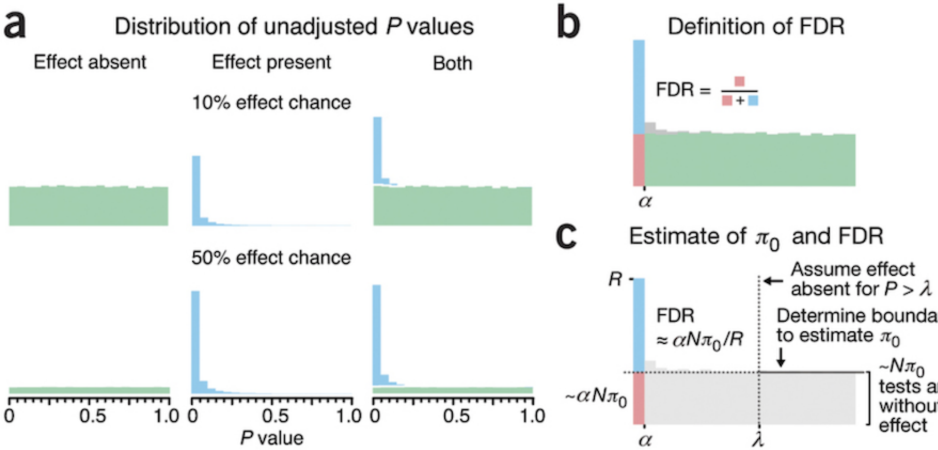
$$Storey:\ pFDR = E\left[\frac{V}{R}|R > 0\right]$$

- Since P(R $>$ 0) is $\sim$ 1 in most genomics experiments, FDR and pFDR are very similar
- Omitting P(R $>$ 0) facilitated development of a measure of significance in terms of the FDR for each hypothesis

Q-value

- Storey & Tibshirani, "**Statistical significance for genomewide studies**", PNAS, 2003 http://www.pnas.org/content/100/16/9440.full
- Empirically derived – uses the p-value distribution
- Storey's method first estimates the fraction of comparisons for which the null is true, π_0, counting the number of P values larger than a cutoff λ (such as 0.5) relative to $(1 - \lambda) * N$ (such as $N/2$), the count expected when the distribution is uniform
- Multiply the Benjamini & Hochberg FDR by π_0, thus less conservative

Q-value

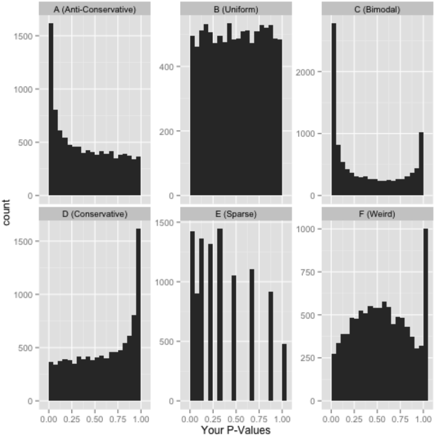


Martin Krzywinski & Naomi Altman "**Points of significance: Comparing samples—part II**" *Nature Methods* 2016
http://www.nature.com/nmeth/journal/v11/n4/full/nmeth.2900.html

Q-value

- q-value is defined as the minimum FDR that can be attained when calling a “feature” significant (i.e., expected proportion of false positives incurred when calling that feature significant)
- The estimated q-value is a function of the p-value for that test and the distribution of the entire set of p-values from the family of tests being considered
- Thus, in an array study testing for differential expression, if gene X has a q-value of 0.013 it means that 1.3% of genes that show pvalues at least as small as gene X are false positives

Check p-value distribution!



http://varianceexplained.org/statistics/interpreting-pvalue-histogram/

Increase power - Filtering

Three filtering methods

- Mean filtering
- Variance filtering
- Threshold filtering

All three filtering methods reduce the number of hypothesis tests to be performed.

Mean Filtering

- Removes the genes with low mean gene signal values
- The genes with mean signal less than a fixed cut-off value C are filtered out.
- The cut-off C is chosen based on background noise level.
- Removes non-expressed genes or genes with low signal values at background noise level.

Mean Filtering

Issues

- Ignores the treatment effect by comparing single mean expression value to a cut-off value.
- Differentially expressed genes with moderate expression in one group and low expression in the other group filtered out.

Variance Filtering

- Removes the genes with low variances across samples
- Genes are sorted in ascending order based on their sample variance estimates and the first X percent of genes are filtered out.
- The cut-off percentage X is arbitrarily determined by the investigator.
- Removes genes at different expression levels

Variance Filtering

Issues

- Gene-specific variance estimates are unreliable in small sample size studies
- Non-expressed genes with higher variances being retained for the analysis and consequently higher number of false positives.
- Differentially expressed genes with low variances estimates being filtered out and a lower number of true positives.
- Uses total gene variance rather than between/within group variance

Threshold Filtering

- Threshold Filtering Method aims to filter out only non- expressed genes.
- For sample size up to 5 per group, genes are filtered out only if one or no samples (across groups) have a signal greater than the background cut-off value.
- For sample size greater than 5 per group, genes are filtered out only if 20% or less samples in each group have signal values greater than the background cut-off value.