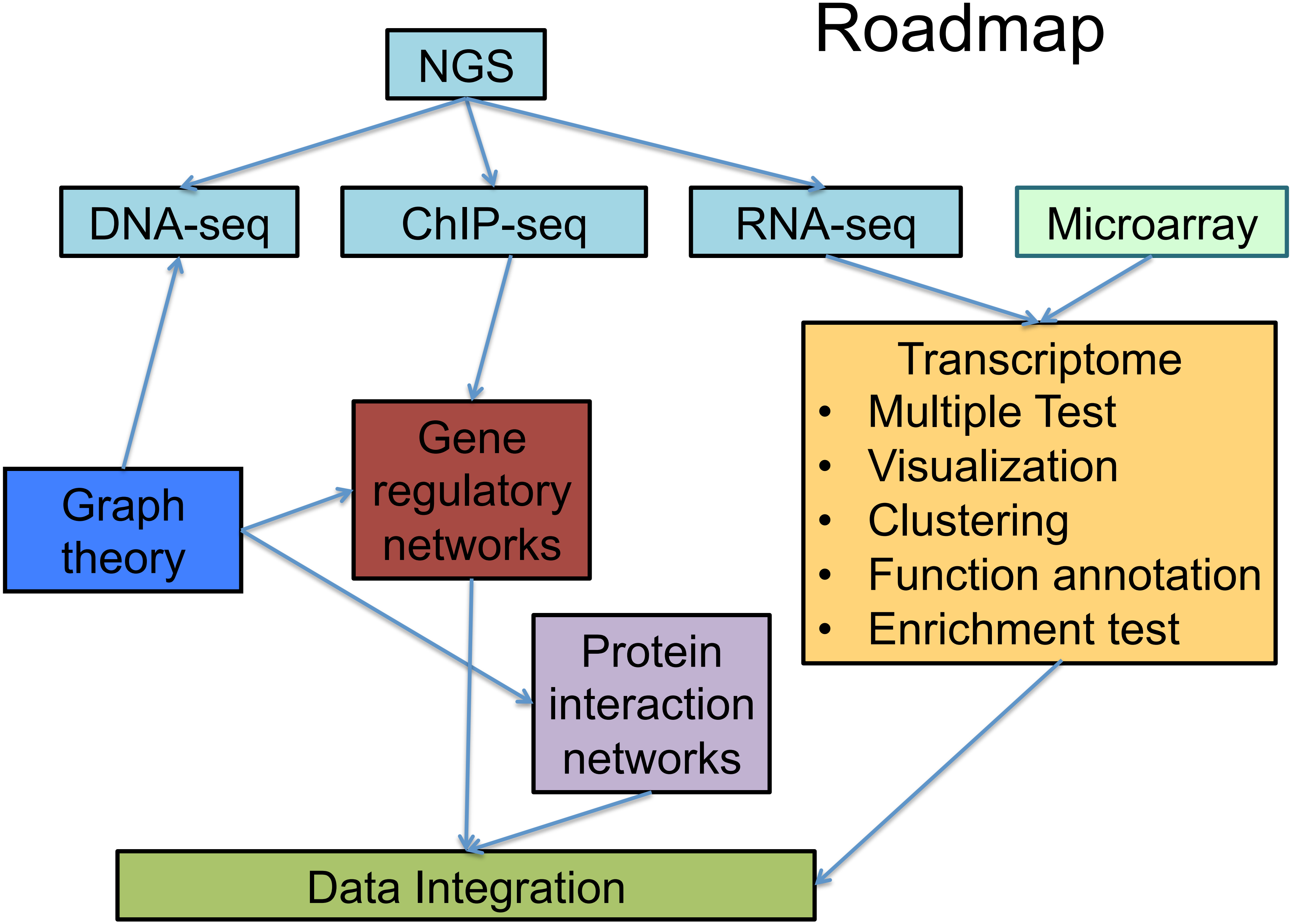
Presentation:

- ~15 minutes
- On 11/24, 12/1, and 12/3
- 11/24: Baral, Daharsh, DasGupta, Du, Kesinger
- 12/1: Kumar, Levine, Mao, Mediratta, Moore
- 12/3: Neal Payne, Ronish, Yang

Roadmap



NGS

DNA-seq

ChIP-seq

RNA-seq

Microarray

Graph theory

Gene regulatory networks

Protein interaction networks

Transcriptome
- Multiple Test
- Visualization
- Clustering
- Function annotation
- Enrichment test

Data Integration

Transcriptome

Lecture 1

Outline

- Multiple Testing Procedures
- Data Visualization, Distance Measures
- Clustering
- Gene Annotation and Enrichment Analysis

The problem

- After differential expression testing (from RNA-seq or Microarray assay), a list of P-values can be obtained, one for each gene.

- Most investigators want to
  - Identify the genes that are differentially expressed
  - Estimate the proportion of errors in the list of selected "differentially expressed genes"

A single gene example (small scale case)

- Suppose you are only interested in a single gene.
- You want to compare the expression level (the level of transcription) of this gene between two conditions (control and treatment).
- For each conditions, there are three replicates.
- Experiments are performed on each sample to measure gene expression levels (e.g., quantitative PCR, gel blot).
- A t-test is performed and a p-value is obtained.
- Declare there is differential expression if p-value is below some threshold (e.g., 0.05).

Extreme parallel hypothesis testing

- With high throughput technology, we can and often perform the same hypothesis test on each and every gene.
- Thus, tens of thousands of hypotheses are tested in parallel.

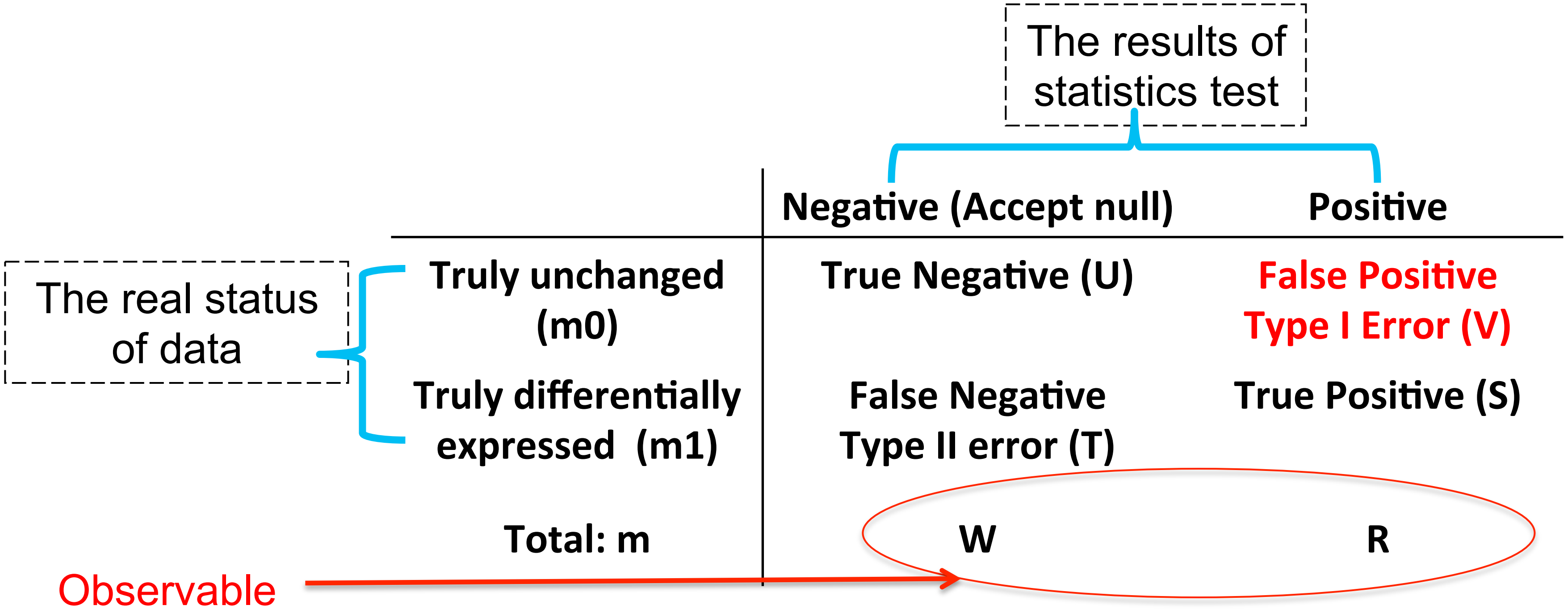
A naïve solution

- Since genes with small p-values are likely to be differentially expressed, why don't we just use the traditional (pre-specified) $\alpha = 0.05$ to decide?
  - ❑Yes?
  - ☒No? But why?

What is P-value?

- P-value is the probability of obtaining a test result as extreme as the one you are getting under the null hypothesis (i.e., area in both tails of the distribution).
  - Null hypothesis: The difference in average expression between the two groups is zero.
- The *lower* the p-value, the *less* probable the result is. (assuming the null hypothesis is true).
- Interpretation: if you repeat the same experiment many times (i.e., computing a T-statistics for each gene on a microarray), the p-value represents the proportion of times that you would expect to see a T-statistic this extreme.

What is P-value?
A more rigorous interpretation



The results of statistics test

	Negative (Accept null)	**Positive**
Truly unchanged (m0)	**True Negative (U)**	**False Positive Type I Error (V)**
Truly differentially expressed (m1)	**False Negative Type II error (T)**	**True Positive (S)**
Total: m	**W**	**R**

The real status of data

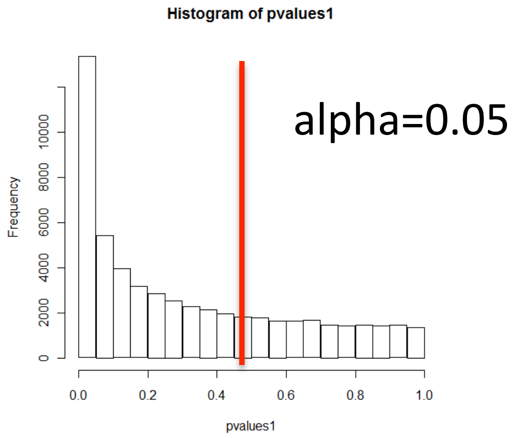
Observable

- P-value = Prob(Type I Error) <- describe the false positive rate
- New Interpretation: if you repeat the same experiment many times (i.e., computing a t-statistic for each gene on a microarray), the p-value represents the proportion of times that you would commit a type I error (i.e., false positive call).

What does this mean to RNA-seq/ microarray data?

- The result is that we obtain one p-value for each gene

	T 1	T 2	T 3	N 1	N 2	N 3	T-statistics	P-value
G 1							T1	P1
G 2							T2	p2
…							…	….
G 20000							T20000	P20000



- 20,000 p-values…
- If we use alpha=0.05 to decide differentially expressed genes, 5% of the 20,000 genes would then be selected by chance
- That means 1000 genes would be false positives…

A naïve solution

- Since genes with small p-values are likely to be differentially expressed, why don't we just use the traditional (pre-specified) $\alpha = 0.05$ to decide?
  - ❑Yes?
  - ❑No!   20,000x0.05 = 1000 false positives!
    - If the investigator is interested in selecting 100 genes for downstream analysis, they could all be false positives by chance!
  - ❑ Other solutions?

The Multiple Testing Problem

- Suppose one test of interest has been conducted for each of *m* genes in a RNA-seq experiment.

- Let $p_1, p_2, \ldots , p_m$ denote the *p*-values corresponding to the *m* tests.

- Let $H_{01}, H_{02}, \ldots , H_{0m}$ denote the null hypotheses corresponding to the *m* tests.

The Multiple Testing Problem

- H_{0i}: *no differential expression for gene I*
- H_{1i}: *differential expression for gene i*

- Let one single c serve as a cutoff for significance:

  - Reject H_{0i} if $p_i \leq c$ (declare significant)

  - Fail to reject (or accept) H_{0i} if $p_i > c$ (declare non-significant)

- $i=1,2,\ldots.m$

The solutions

- To select differentially expressed genes, we need to do multiple testing (multiplicity) corrections
  - Familywise Error Rate (FWER), such as Bonferroni correction and Holm's method: adjust the p-value threshold from alpha to alpha/(number of genes)
  - Control False Discovery Rate: algorithm proposed by Benjamini & Hochberg
  - Re-sampling techniques (i.e., Permutation P-values)

Familywise Error Rate (FWER)

- Traditionally statisticians have focused on controlling FWER when conducting multiple tests.
- FWER is defined as the probability of one or more false positive results:

$$\text{FWER}=P(V>0).$$

- Controlling FWER amounts to choosing the significance cutoff *c* so that FWER is less than or equal to some desired level α.

The Bonferroni Method

- The Bonferroni Method is the simplest way to achieve control of the FWER at any desired level α.

- Simply choose $c = \alpha / m$.

- With this value of *c* for each individual test, the FWER will be no larger than α for any family of *m* tests.

Bonferroni correction

	T 1	T 2	T 3	N 1	N 2	N 3	P-value
G 1	y1	y2	y3	y4	y5	y6	0.012
G 2	y1	y2	y3	y4	y5	y6	0.045
…							
G 20000							

- using α = 0.05 we reject the null hypothesis that the expression of gene 1 (2) is not changed in tumor versus normal tissue.
- In the other words, gene 1 (2) is differentially expressed genes between tumor and normal tissues.

Bonferroni correction

- However, the probability that either the expression difference observed for gene 1 (p=0.012) or the expression difference observed for gene 2 (p=0.045) under null hypothesis is 0.012+0.045 = 0.057 (>0.05!).

- Using an overall p-value alpha = 0.05, we have no evidence to reject the null hypothesis that the expression of either gene 1 or gene 2 has no change in the comparison between tumor versus normal tissues.
  - Here overall p-value is the probability of making at least 1 mistake in the two performed tests.
  - Hence, the α=0.05 is not stringent enough for each test.

Bonferroni correction

- The Bonferroni rule
  - To guarantee that the probability of making at least 1 mistake in the two performed tests is not larger than alpha, we need to use for each test α/2 as significance level

  - To guarantee that he probability of making at least 1 mistake in the ten performed tests is not larger than alpha, we need to use for each test α/10 as significance level

Bonferroni correction

Genome wide gene expression profiles

	T 1	T 2	T 3	N 1	N 2	N 3	P-value
G 1							0.012
G 2							0.045
...							...
G 20000	y1	y2	y3	y4	y5	y6	P20000

- 20,000 p-values need to be combined to give an overall conclusion of how many genes are differentially expressed.

Bonferroni correction

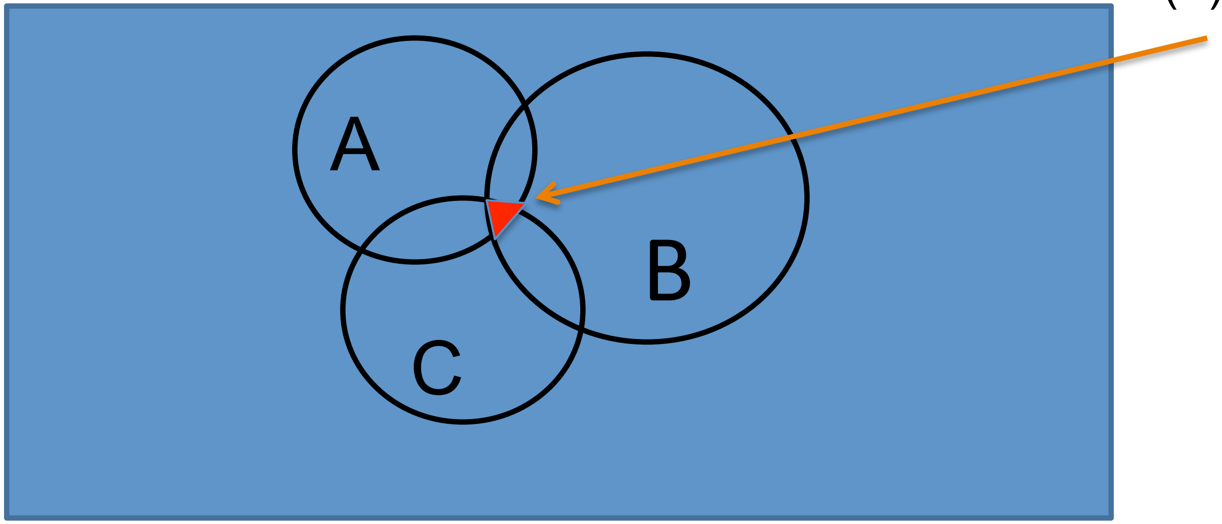
- Hence, under Bonferroni rule, we need to use a significance level of alpha/20000 for each gene .
  - Simply choose $c = \alpha / m$.
  - $\alpha = 0.05 \Rightarrow c = \alpha/20000 = 0.0000025$
  - In other words, under Bonferroni rule, we will select a gene as differentially expressed if its P-value < 0.0000025. This will guarantee the probability of making at least 1 mistake in the 20000 performed tests is not larger than 0.05.
    - More specifically, out of the genes selected, there is only very small chance (5%) that at least one of them is a false positive
  - Is this too tough (stringent, conservative)?
    - ❑Yes (if few genes'p-values are less than α/200000: Game Over…)

Weak Control vs. Strong Control

- A method provides *weak control* of an error rate for a family of *m* tests if the FWER control at level α is guaranteed **only** when all null hypotheses are true (i.e. when $m=m_0$ so the global null hypothesis is true).

- A method provides *strong control* of an error rate for a family of *m* tests if the FWER control at level α is guaranteed for **any** configuration of true and non-true null hypotheses (including the global null hypothesis)

Bonferroni’s method can achieve strong control



P(*A*)+P(*B*)+P(*C*)

Assuming the rectangle has probability 1, the three circles, A, B, C, represents three events. The probability P(*A*U*B*U*C*), i.e., the probability of A or B or C, is smaller than P(*A*)+P(*B*)+P(*C*).

Holm's Method for Controlling FWER at Level α

- Let $p_{(1)}, p_{(2)}, \ldots, p_{(m)}$ denote the m p-values ordered from smallest to largest. (need to sort all P-values first)

- Find the largest integer k so that

  $$p_{(i)} \leq \alpha / (m-i+1) \text{ for all } i=1,\ldots,k.$$

  (when you see it first time)

- set $c = p_{(k)}$ (reject the nulls corresponding to the smallest k p-values).

- If no such k exists, set $c = 0$ (declare nothing significant).

An Example

- Suppose we conduct 5 tests and obtain the following *p*-values for tests 1 through 5.

Test	1	2	3	4	5
p-value	0.042	0.001	0.031	0.014	0.007

- Which tests' null hypotheses will be rejected if you wish to control the FWER at level 0.05?

- Use both the Bonferroni method and the Holm method to answer this question.

Solution

Test	T1	T2	T3	T4	T5
P-value	0.042	0.001	0.031	0.014	0.007

- The cutoff for significance is c = 0.05/5=0.01 using the Bonferroni method. Thus we would reject the null hypothesis for tests 2 and 5 with the Bonferroni method.

T2: $0.001 \leq 0.05/(5-1+1)=0.01$
T5: $0.007 \leq 0.05/(5-2+1)=0.0125$
$\underline{\text{T4: } 0.014 \leq 0.05/(5-3+1)=0.0167}$
T3: $0.031 > 0.05/(5-4+1)=0.025$
T1: $0.042 \leq 0.05/(5-5+1)=0.05$

These calculations indicate that Holm’s method would reject null hypotheses for tests 2, 5, and 4.

Adjusted p-value

- P-value: the probability to observe more or equally extreme data under the null hypothesis.

- Alternatively, a *p*-value for an individual test can be defined as the smallest significance level (tolerable type 1 error rate) for which we can reject the null hypothesis. For example, if p-value is 0.045, this null hypothesis will be rejected if α=0.05 but note rejected if α=0.04. The smallest α to reject this null hypothesis is 0.45 (p-value).

- The ***adjusted p*-value** for one test in a family of tests is the smallest significance level for which we can reject the null hypothesis for that one test and all others with smaller *p*-values.

Adjusted p-values

- FWER: the *adjusted p*-value for one test in a family of tests is the smallest FWER (α) for which we can reject the null hypothesis for that one test and all others with smaller *p*-values.
- Bonferroni: the null hypothesis will be rejected if unadjusted *p-value* ≤ *α/m*. So the smallest α that can lead to rejection will be *m* x *p-value*, i.e., the adjusted p-value is the raw p-value times *m*.
- Holms: adjusted p-value for *i*-th ordered p-value is

$$p_{(i)} \times (m - i + 1)$$

- The advantage of adjusted p-values: they can be compared directly with α.

Example

Test	T1	T2	T3	T4	T5
Raw P-value	0.042	0.001	0.031	0.014	0.007
Bonferroni adjusted	0.21	0.005	0.155	0.07	0.035

Reject hypotheses 2 and 5 for Bonferroni’ s method

Holms

0.001*(5-1+1)=0.005

0.007*(5-2+1)=0.028

0.014*(5-3+1)=0.042 $\alpha<0.05$

0.031*(5-4+1)=0.062

0.042*(5-5+1)=0.042

These calculations indicate that Holm’s method would reject null hypotheses for tests 2, 5, and 4.

The solutions with R

> results=topTable(fit2, number=20, adjust.method=”xxx”)

> results=topTags(fit2, number=20, adjust.method=”xxx”)

adjust.method: “holm”, “hochberg”, “hommel”, “bonferroni”, “BH”, “BY”, “fdr”, “none”

The solutions

- To select differentially expressed genes, we need to do multiple testing (multiplicity) corrections
  - Familywise Error Rate (FWER), such as Bonferroni correction and Holm's method: adjust the p-value threshold from alpha to alpha/(number of genes)
  - Control False Discovery Rate: algorithm proposed by Benjamini & Hochberg
  - Re-sampling techniques (i.e., Permutation P-values)

FDR (False Discovery Rate)

- The investigators, after spending thousands of dollars, want to obtain a list of selected genes
- As Bonferroni correction is very strict, only a few genes might be selected
- As an alternative solution, we can choose to control the proportion of false positives out of selected genes.
- FDR is an alternative error rate that can be useful for high throughput experiments.

FDR (False Discovery Rate)

The results of statistics test

The real status of data

	Negative	Positive	Total
Truly unchanged	True Negative (U)	False Positive (V) Type I Error	M0
Truly differentially expressed	False Negative (T) Type II error	True Positive (S)	M-M0
Total	W=M-R	R	M

- U: number of true negatives; S: number of true positives
- T: number of false negatives; V: number of false positives
- In our RNA-seq/Microarray experiment, M could be 20,000 genes
- R is known (i.e., how many genes are called positive by statistics tests)

FDR (False Discovery Rate)

The results of statistics test

The real status of data

	Negative	Positive	Total
Truly unchanged	**True Negative (U)**	**False Positive (V) Type I Error**	**M0**
Truly differentially expressed	**False Negative (T) Type II error**	**True Positive (S)**	**M-M0**
Total	**M-R**	**R**	**M**

- FDR is defined as the expected proportion of false positives (type I errors) among all rejected null hypotheses

$$FDR = E(Q) \quad \text{with} \quad \begin{array}{ll} Q = V / R & \text{if } R > 0 \\ Q = 0 & \text{if } R = 0 \end{array}$$

False Discovery Rate (FDR)

- FDR was introduced by Benjamini and Hochberg (1995) and is formally defined as

  FDR=E(Q)

  Q=V/R=False Positive/(True Positive + False Positive)

- Controlling FDR amounts to choosing the significance cutoff *c* so that FDR is less than or equal to some desired level α.

- More specifically, if we want to control at most 5% false positives, which genes should be selected?

FDR (False discovery rate): How?

- The Benjamini & Hochberg procedure to control FDR :
    - For each gene (out of a total of n), perform one test
    - Obtain m P-values: $p_1, p_2, \ldots, p_m$
    - Sort the obtained P-values: $p_{(1)}, p_{(2)}, \ldots, p_{(m)}$
    - To control the FDR at q, we will reject all genes with p-values $p \leq p_{(j)}$, where j is the largest index for which

$$p_{(j)} \leq \frac{qj}{m}$$

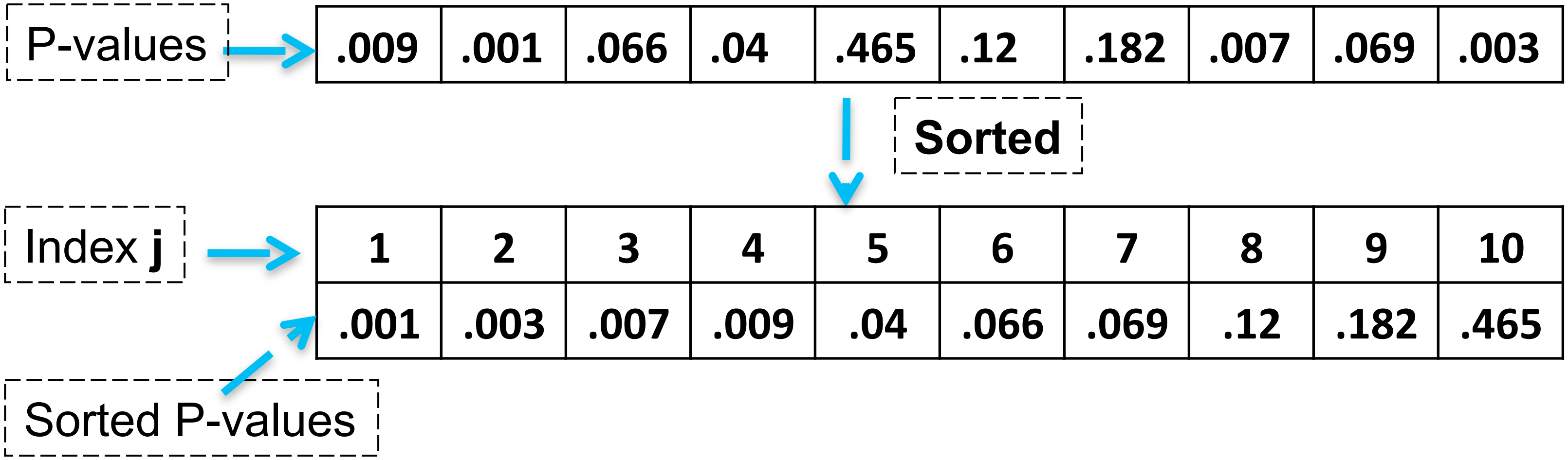
FDR (False Discovery Rate): An Example of 10 genes

- Aim: To control the FDR at level of 5%

P-values	.009	.001	.065	.04	.454	.123	.172	.007	.68	.003

FDR (False Discovery Rate): An Example of 10 genes

- Aim: To control the FDR at 5% (q =0.05)



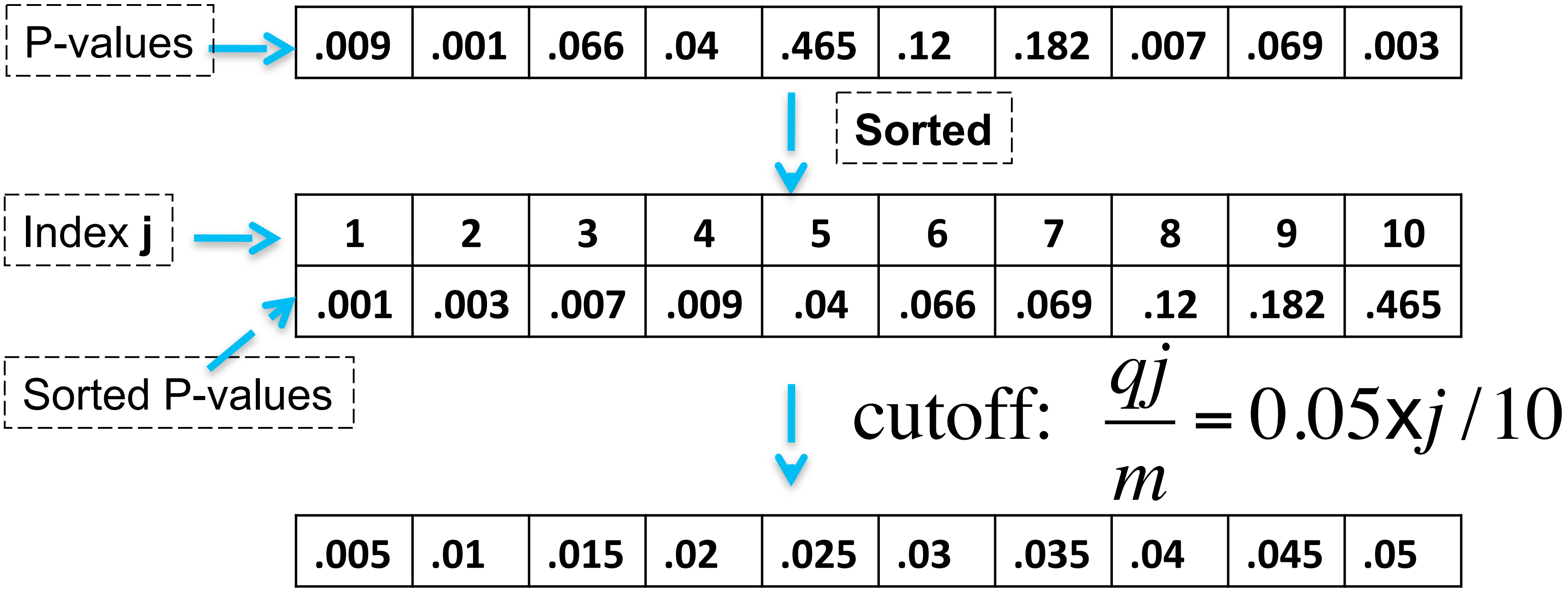
P-values									
.009	.001	.066	.04	.465	.12	.182	.007	.069	.003

Sorted

Index **j**	1	2	3	4	5	6	7	8	9	10
Sorted P-values	.001	.003	.007	.009	.04	.066	.069	.12	.182	.465

FDR (False Discovery Rate): An Example of 10 genes

- Aim: To control the FDR at 5% (q =0.05)



P-values

.009	.001	.066	.04	.465	.12	.182	.007	.069	.003

Sorted

Index **j**

Sorted P-values

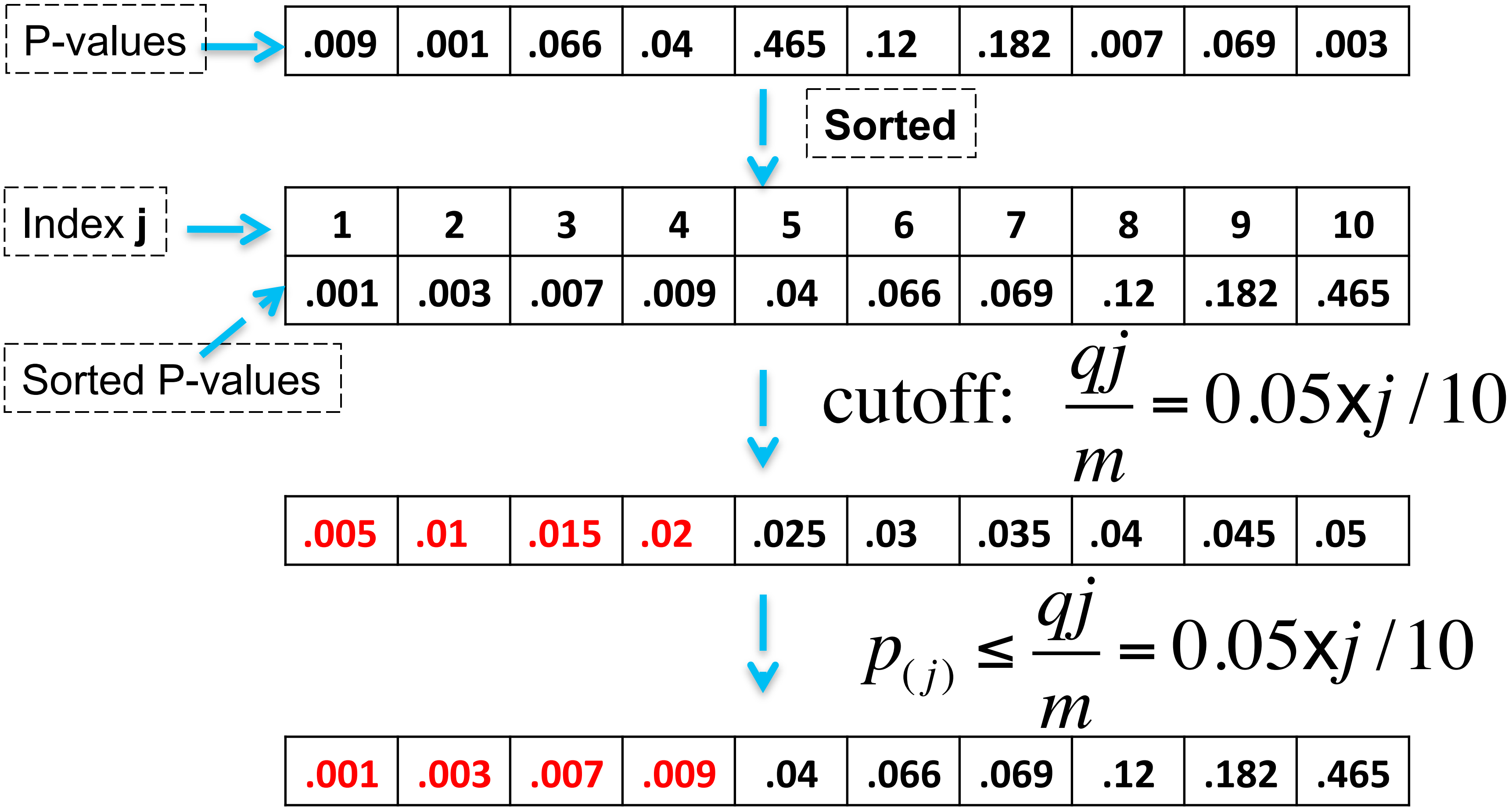
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
.001	.003	.007	.009	.04	.066	.069	.12	.182	.465

cutoff: $\frac{qj}{m} = 0.05 \text{x} j / 10$

.005	.01	.015	.02	.025	.03	.035	.04	.045	.05

FDR (False Discovery Rate): An Example of 10 genes

- Aim: To control the FDR at 5% (q =0.05)



P-values

.009	.001	.066	.04	.465	.12	.182	.007	.069	.003

Sorted

Index **j**

Sorted P-values

1	2	3	4	5	6	7	8	9	10
.001	.003	.007	.009	.04	.066	.069	.12	.182	.465

cutoff: $\frac{qj}{m} = 0.05\text{x}j/10$

.005	.01	.015	.02	.025	.03	.035	.04	.045	.05

$$p_{(j)} \le \frac{qj}{m} = 0.05\text{x}j/10$$

.001	.003	.007	.009	.04	.066	.069	.12	.182	.465

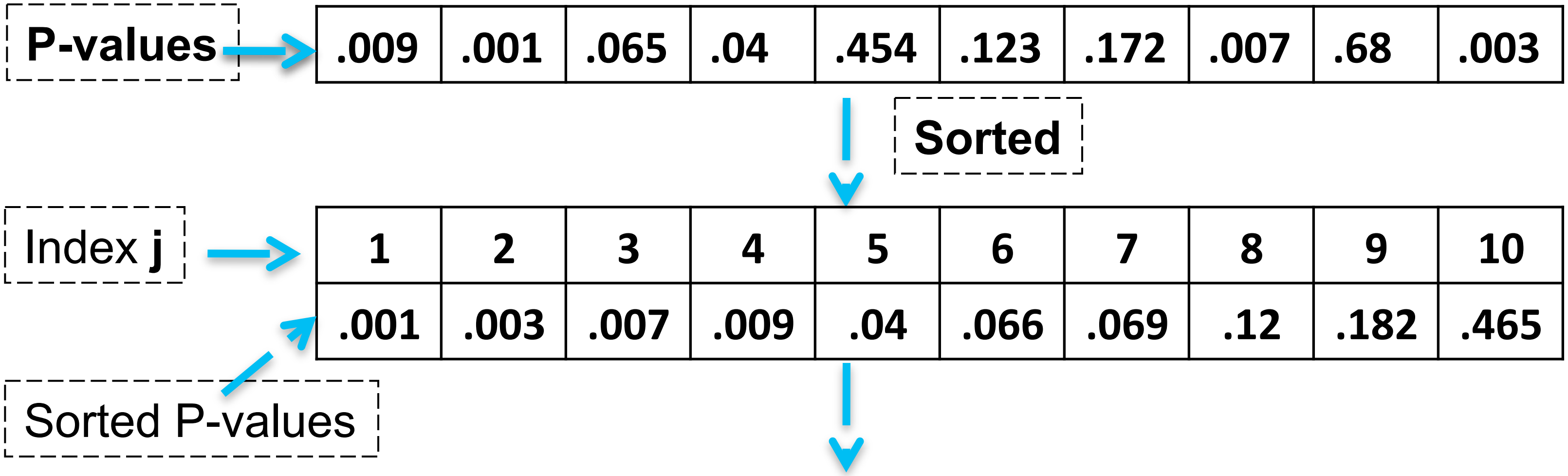
How about Bonferroni correction? The Same Example of 10 genes

- Aim: Use Bonferroni correction, α=0.05

P-values	.009	.001	.065	.04	.454	.123	.172	.007	.68	.003

How about Bonferroni correction? The Same Example of 10 genes

- Aim: Use Bonferroni correction, α=0.05



P-values									
.009	.001	.065	.04	.454	.123	.172	.007	.68	.003

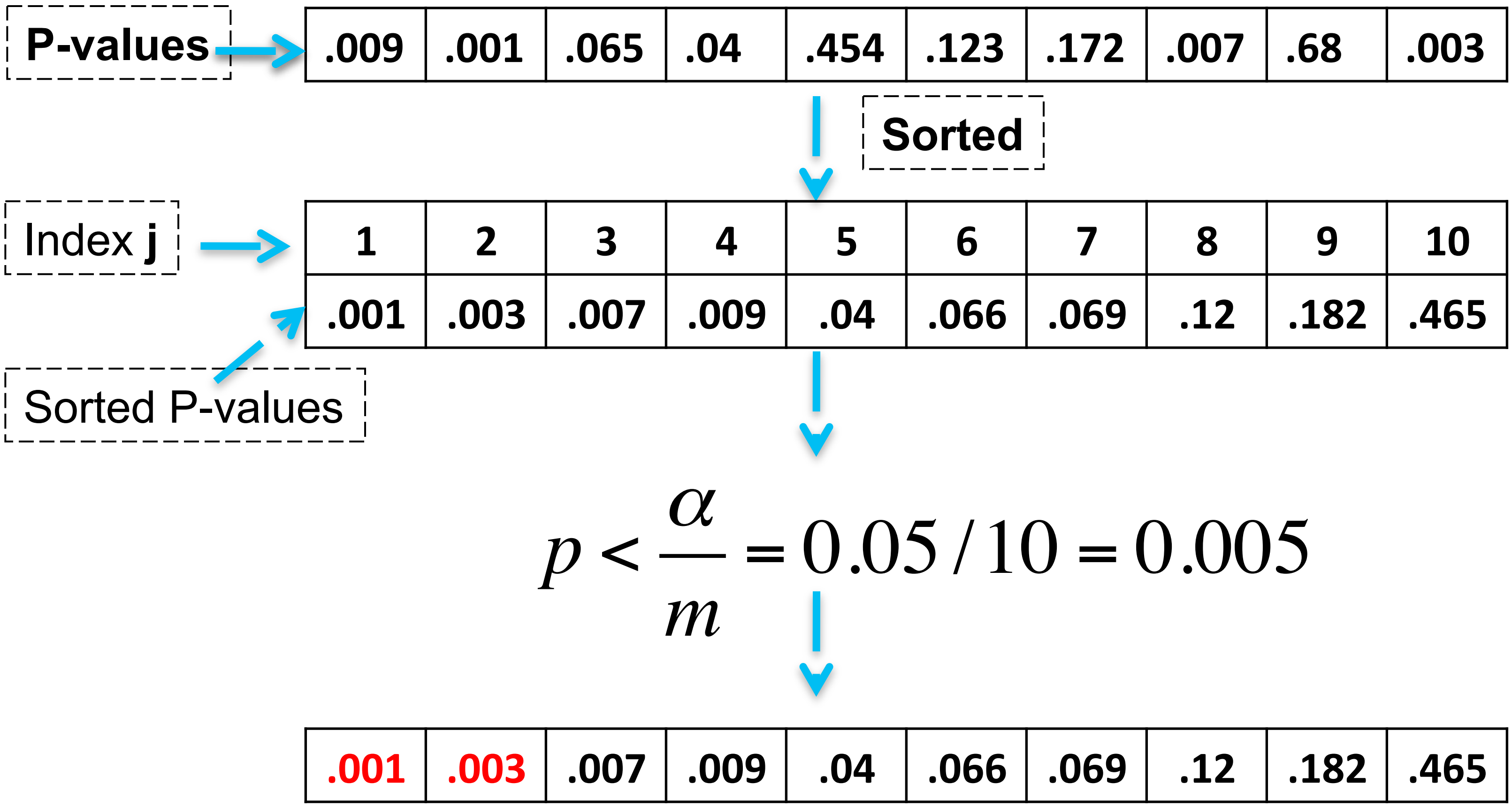
Sorted

Index **j**	1	2	3	4	5	6	7	8	9	10
Sorted P-values	.001	.003	.007	.009	.04	.066	.069	.12	.182	.465

$$\text{cutoff:} \quad \frac{\alpha}{m} = 0.05/10 = 0.005$$

How about Bonferroni correction? The Same Example of 10 genes

- Aim: Use Bonferroni correction, α=0.05



P-values	.009	.001	.065	.04	.454	.123	.172	.007	.68	.003

Sorted

Index **j**	1	2	3	4	5	6	7	8	9	10
Sorted P-values	.001	.003	.007	.009	.04	.066	.069	.12	.182	.465

$$p < \frac{\alpha}{m} = 0.05 / 10 = 0.005$$

.001	.003	.007	.009	.04	.066	.069	.12	.182	.465

Adjusted p-values (q-values)

- If we use FDR as the significance threshold, the adjusted *p*-value for one test in a family of tests is the smallest FDR for which we can reject the null hypothesis for that one test and all others with smaller *p*-values.

- In FDR setting, adjusted p-values are also called q-values. q-value is derived in an empirical Bayes setting, but it is equivalent to adjusted p-value in practice.

The adjusted p-value or *q*-value for a given test fills the blanks in the following sentences:

- "If I set my cutoff for significance *c* equal to this *p*-value, I must be willing to accept a false discovery rate of ______."

- "To reject the null hypothesis for this test and all others with smaller *p*-values, I must be willing to accept a false discovery rate of _______."

- "To include this gene on my list of differentially expressed genes, I must be willing to accept a false discovery rate of _____."

Computation and Use of *q*-values

- Let $q_{(i)}$ denote the *q*-value that corresponds to the ith smallest *p*-value $p_{(i)}$.

- $q_{(i)} = \min \{ p_{(k)}\, m / k : k = i, ..., m \}$.

The solutions with R

```
> results=topTable(fit2, number=20,
  adjust.method="fdr", lfc=1)
> results=topTags(fit2, number=20,
  adjust.method="fdr”)
```

adjust.method: "holm", "hochberg", "hommel", "bonferroni", "BH", "BY”, "fdr", "none”

Outline

- Multiple Testing Procedures
- Data Visualization, Distance Measures
- Clustering
- Gene Annotation and Enrichment Analysis