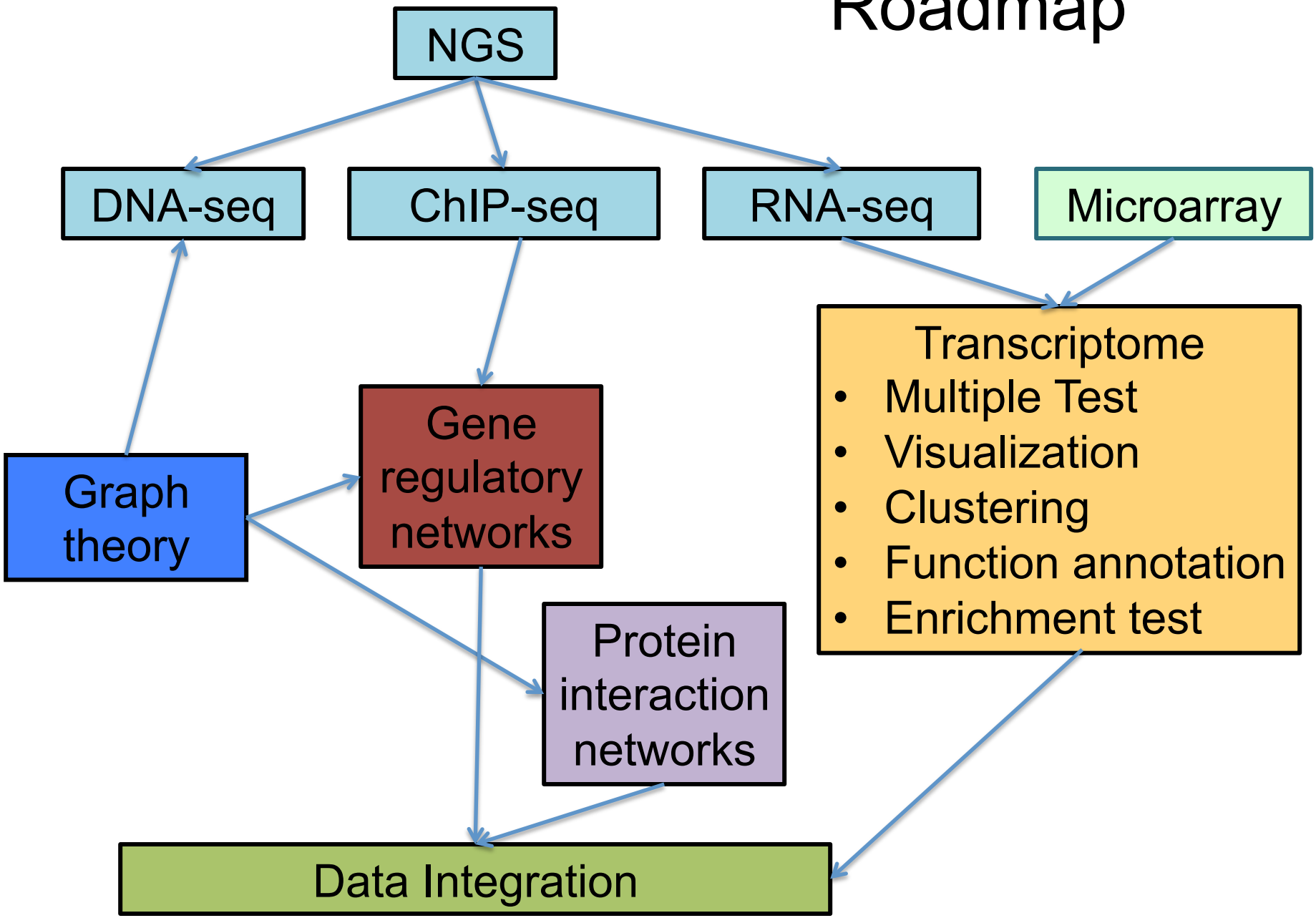
1. Find a research article for your presentation and final exam
2. Prepare your presentation
3. No class for the week of Midterm exam
4. http://sysbio.unl.edu/Teaching/BIOS497897_2014/

Roadmap


NGS
DNA-seq
ChIP-seq
RNA-seq
Microarray
Transcriptome
• Multiple Test
• Visualization
• Clustering
• Function annotation
• Enrichment test
Graph theory
Gene regulatory networks
Protein interaction networks
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-value is 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 (low throughput 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 treatment, there are three biological samples.
- Experiments are performed on each sample to measure gene expression levels (e.g., quantitative PCR, gel blot).
- There are three independent treatment per condition.
- 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?

The Multiple Testing Problem

- Suppose one test of interest has been conducted for each of *m* genes in a microarray experiment. For example, through two-sample t-test.
- 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

- Usually,
- H_{0i}: *no differential expression for gene i*
- H_{1i}: *differential expression for gene i*
- $i=1,2,\ldots.m$

The Multiple Testing Problem (continued)

- Suppose m_0 of the null hypotheses are true and m_1 of the null hypotheses are false.
- Let c denote a value between 0 and 1 that will serve as a cutoff for significance:

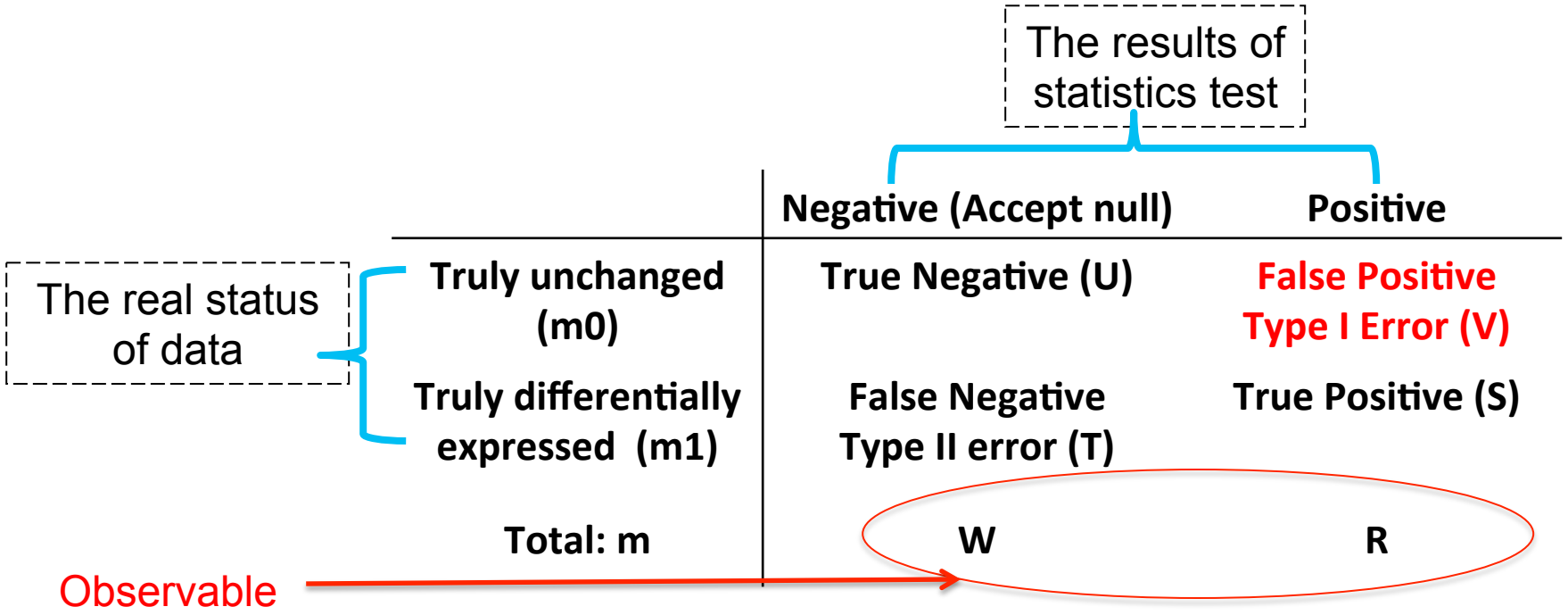
  - Reject H_{0i} if $p_i \leq c$ (declare significant)

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

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



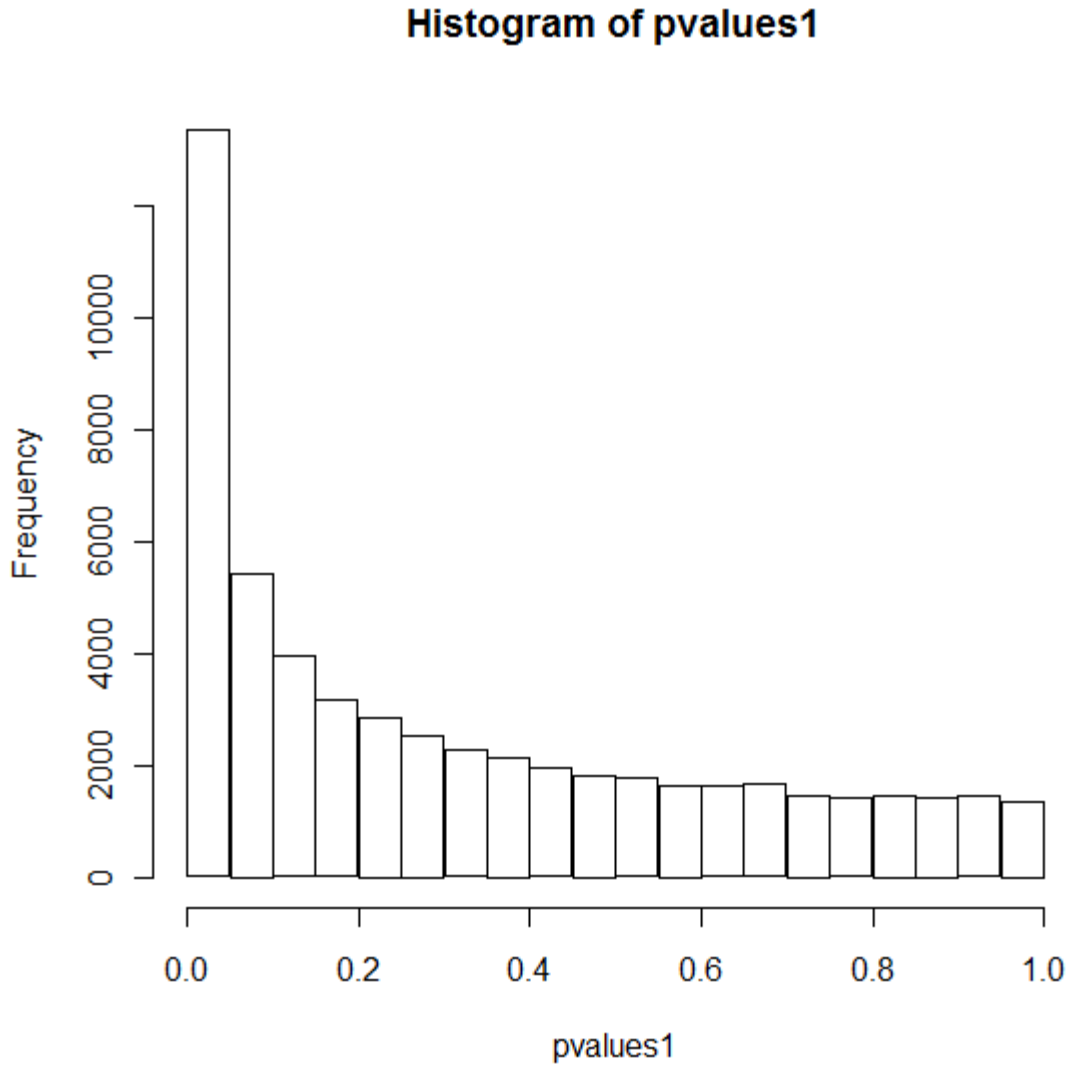
- P-value = Prob(Type I Error) <- describe the false positive rate
- New Interpretation: if you repeat the same experiment many times (i.e., computing a t-statistic for each gene on a microarray), the p-value represents the proportion of times that you would commit a type I error (i.e., false positive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

What does this mean to 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

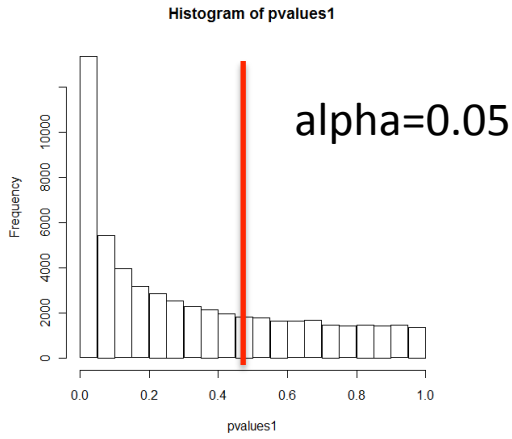
- 20,000 p-values…


Histogram of pvalues1
Frequency
0
2000
4000
6000
8000
10000
0.0
0.2
0.4
0.6
0.8
1.0
pvalues1

What does this mean to 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 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)