6.874/… Recitation 1

Courtesy of an MIT Teaching Assistant.

Separate 6.874 recitation

- Teaching duties shared with Charlie + 1 guest lecture
- Cover extra AI material in recitation
  - Usually topics complementing lecture
  - Extra problem set/exam problems
  - 6.874 will start exams early
- Other recitation sections will review lecture

Reminders

- Pset 1 posted – due Feb 20th (no AI problem)
- Pset 2 posted soon – Due Mar 13th
  - Programming problem
- Python tutorial – Feb 10th (Monday) 4-5pm.
- Project interests due – Feb 11th
  - Name, program, previous experience, interest in computational biology
  - We'll post these next week for you to find groups for project
- Office hours posted soon

Today: Statistics Review/Multiple Testing

- Basic probability: motif representation/scanning
- Basic statistics
- Multiple hypothesis testing in context of motif scanning
  - Bonferroni/Benjamini-Hochberg

Nature Biotechnology **27**, 1135 - 1137 (2009)
doi:10.1038/nbt1209-1135

How does multiple testing correction work?

William S Noble[1]

Minimal biology review

- DNA is composed of 4 nucleotides: A, C, G, T
- DNA is *transcribed* into mRNA which is *translated* into protein
- A *gene* is a said to be *expressed* when it is transcribed
- *Transcription factors (TF)* are proteins that bind DNA and affect (promote/repress) gene expression
- A *DNA sequence motif* can be a sequence where specific TFs bind (others too – eg. splicing signals for mRNA)

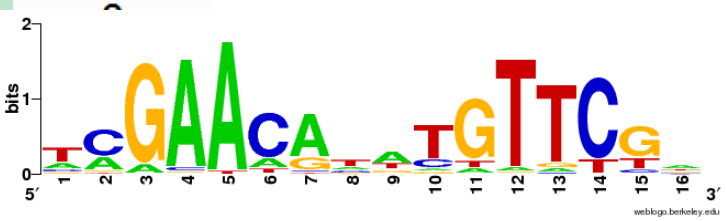
DNA sequence motif representation

- Proteins (TFs) bind to motifs that are not fully specified
- Consensus sequence: TCGAACATATGTTCGA
- Collection of k-mers:
  - TCGAACATATGTTCGA
  - TCGAAAATATGTTCGA
  - TAGAACATATCTTCGA …
- Probabilistic model (PWM/PSSM)

Position Weight Matrix (PWM)

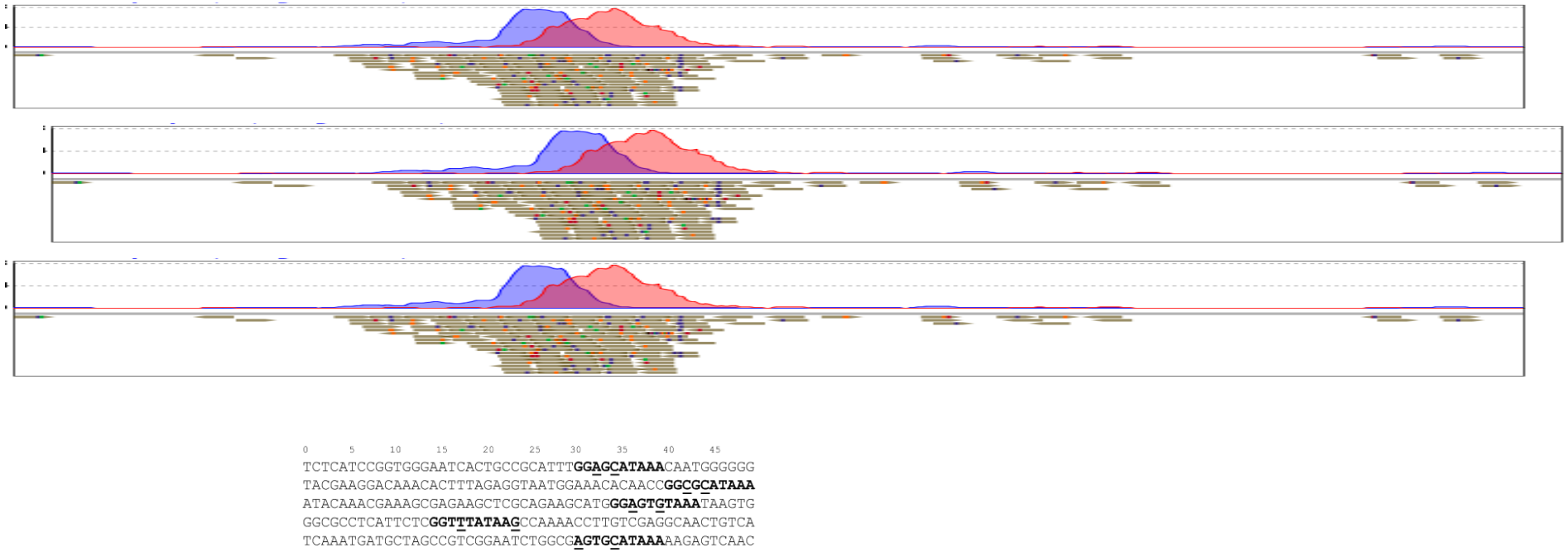
- Proteins (TFs) bind to motifs that are not fully specified
- Matrix of probabilities
  - Each column (position) is a multinomial distribution over the nucleotides – sums to 1
  - Each column (position) is independent of other columns

	1	2	3	4
A	0.6	0.25	0.1	1
G	0.4	0.25	0.1	0
T	0	0.25	0.4	
C	0	0.25	0.4	



Aside: How to get a PWM?

- Motif finding on ChIP-seq data for a particular TF



S = GCAA

	1	2	3	4
A	0.6	0.25	0.1	1
G	0.4	0.25	0.1	0
T	0	0.25	0.4	0
C	0	0.25	0.4	0

What do we do with PWM?

- Evaluate probability that a sequence was generated by the motif (does this TF bind this sequence?) S = GCAA

$$P(S|M) = 0.4 \times 0.25 \times 0.1 \times 1.0 = 0.01$$

	1	2	3	4
A	0.6	0.25	0.1	1
G	0.4	0.25	0.1	0
T	0	0.25	0.4	0
C	0	0.25	0.4	0

What do we do with PWM?

- Evaluate probability that a sequence was generated by the motif (does this TF bind this sequence?) S = GCAA

$$P(S|M) = 0.4 \times 0.25 \times 0.1 \times 1.0 = 0.01$$

- Evaluate probability that a sequence was generated by background

$$P(S|B) = 0.4 \times 0.4 \times 0.1 \times 0.1 = 0.0016$$

	1	2	3	4
A	0.6	0.25	0.1	1
G	0.4	0.25	0.1	0
T	0	0.25	0.4	0
C	0	0.25	0.4	0

A	0.1
G	0.4
T	0.1
C	0.4

What do we do with PWM?

- Using Bayes' rule compute posterior probability that motif generated the sequence
  - Assume prior probability of $\mathrm{P(M)}$ = .1
  - $\mathrm{P(S|M)}$ = 0.01; $\mathrm{P(S|B)}$ = .0016 (from previous slide)

$$\mathrm{P(M|S)} = \frac{\mathrm{P(S|M)} \times \mathrm{P(M)}}{\mathrm{P(S)}} = \frac{\mathrm{P(S|M)} \times \mathrm{P(M)}}{\mathrm{P(S|B)P(B)} + \mathrm{P(S|M)P(M)}}$$

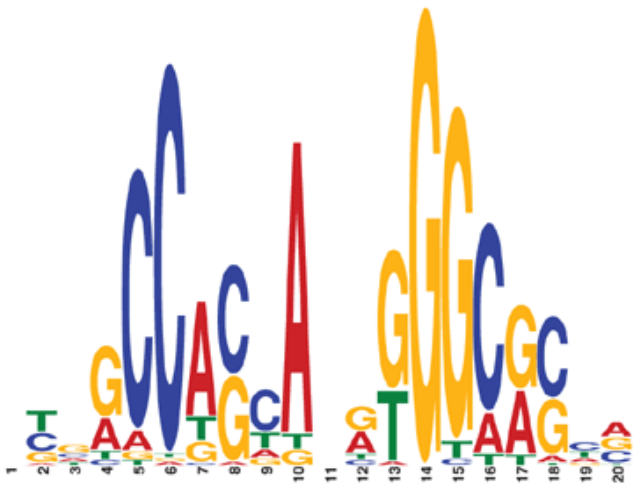
$$= \frac{0.01 \times 0.1}{0.0016 \times 0.9 + 0.01 \times 0.1} = 0.41$$

Assigning significance

- We just scanned to test if one sequence was an instance of a motif
  - …. 3 billion to go
  - Like BLAST example in lecture – slide it along the genome
- Out of these 3 billion, how do we decide which ones we think are bound?

Nature Biotechnology example

a



b

Position	Str	Sequence	Score
19390631	+	TTGACCAGCAGGGGGCGCCG	26.30
32420105	+	CTGGCCAGCAGAGGGCAGCA	26.30
27910537	–	CGGTGCCCCCTGCTGGTCAG	26.18
21968106	+	GTGACCACCAGGGGGCAGCA	25.81
31409358	+	CGGGCCTCCAGGGGGCGCTC	25.56
19129218	–	TGGCGCCACCTGCTGGTCAC	25.44
21854623	+	CTGGCCAGCAGAGGGCAGGG	24.95
12364895	+	CCCGCCAGCAGAGGGAGCCG	24.71
13406383	+	CTAGCCACCAGGTGGCGGTG	24.71
18613020	+	CCCGCCAGCAGAGGGAGCCG	24.71
31980801	+	ACGCCCAGCAGGGGGCGCCG	24.71
32909754	–	TGGCTCCCCCTGGCGGCCGG	24.71
25683654	+	TCGGCCACTAGGGGGCACTA	24.58
31116990	–	GGCCGCCACCTTGTGGCCAG	24.58
29615421	–	CTCTGCCCTCTGGTGGCTGC	24.46
6024389	+	GTTGCCACCAGAGGGCACTA	24.46
26610753	–	CACTGCCCTCTGCTGGCCCA	24.34
26912791	–	GGGCGCCACCTGGCGGTCAC	24.34
20446267	+	CTGCCCACCAGGGGGCAGCG	24.22
21872506	–	TGGCGCCACCTGGCGGCAGC	24.22

Null distribution

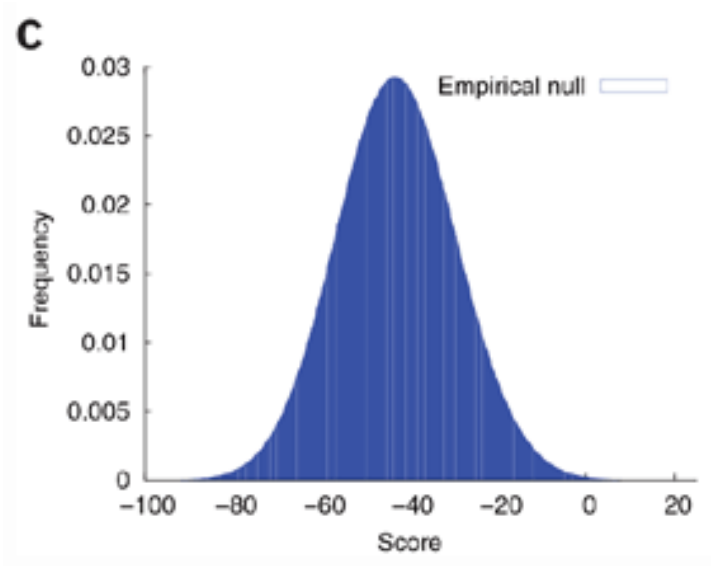
- How biologically meaningful are these scores?
- Assess probability that a particular score would occur by random chance
  - How likely is it that 20 random nucleotides would match CTCF motif?

b

Position	Str	Sequence	Score
19390631	+	TTGACCAGCAGGGGGCGCCG	26.30
32420105	+	CTGGCCAGCAGAGGGCAGCA	26.30
27910537	–	CGGTGCCCCCTGCTGGTCAG	26.18
21968106	+	GTGACCACCAGGGGGCAGCA	25.81
31409358	+	CGGGCCTCCAGGGGGCGCTC	25.56
19129218	–	TGGCGCCACCTGCTGGTCAC	25.44
21854623	+	CTGGCCAGCAGAGGGCAGGG	24.95
12364895	+	CCCGCCAGCAGAGGGAGCCG	24.71
13406383	+	CTAGCCACCAGGTGGCGGTG	24.71
18613020	+	CCCGCCAGCAGAGGGAGCCG	24.71
31980801	+	ACGCCCAGCAGGGGGCGCCG	24.71
32909754	–	TGGCTCCCCCTGGCGGCCGG	24.71
25683654	+	TCGGCCACTAGGGGGCACTA	24.58
31116990	–	GGCCGCCACCTTGTGGCCAG	24.58
29615421	–	CTCTGCCCTCTGGTGGCTGC	24.46
6024389	+	GTTGCCACCAGAGGGCACTA	24.46
26610753	–	CACTGCCCTCTGCTGGCCCA	24.34
26912791	–	GGGCGCCACCTGGCGGTCAC	24.34
20446267	+	CTGCCCACCAGGGGGCAGCG	24.22
21872506	–	TGGCGCCACCTGGCGGCAGC	24.22

Null distribution

- Empirical null
  - Shuffle bases of chr21 and rescan
  - Any high scoring CTCF instances occur due to random chance, not biology
  - Histogram of scores in empirical null distribution



b

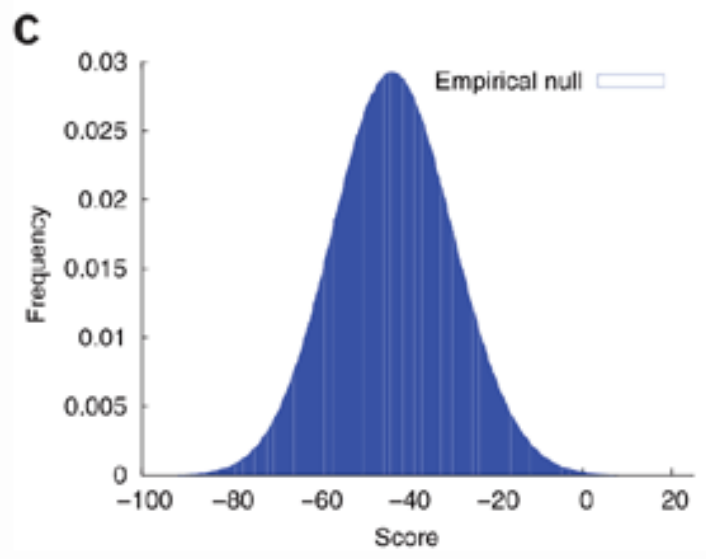
Position	Str	Sequence	Score
19390631	+	TTGACCAGCAGGGGGCGCCG	26.30
32420105	+	CTGGCCAGCAGAGGGCAGCA	26.30
27910537	−	CGGTGCCCCCTGCTGGTCAG	26.18
21968106	+	GTGACCACCAGGGGGCAGCA	25.81
31409358	+	CGGGCCTCCAGGGGGCGCTC	25.56
19129218	−	TGGCGCCACCTGCTGGTCAC	25.44
21854623	+	CTGGCCAGCAGAGGGCAGGG	24.95
12364895	+	CCCGCCAGCAGAGGGAGCCG	24.71
13406383	+	CTAGCCACCAGGTGGCGGTG	24.71
18613020	+	CCCGCCAGCAGAGGGAGCCG	24.71
31980801	+	ACGCCCAGCAGGGGGCGCCG	24.71
32909754	−	TGGCTCCCCCTGGCGGCCGG	24.71
25683654	+	TCGGCCACTAGGGGGCACTA	24.58
31116990	−	GGCCGCCACCTTGTGGCCAG	24.58
29615421	−	CTCTGCCCTCTGGTGGCTGC	24.46
6024389	+	GTTGCCACCAGAGGGCACTA	24.46
26610753	−	CACTGCCCTCTGCTGGCCCA	24.34
26912791	−	GGGCGCCACCTGGCGGTCAC	24.34
20446267	+	CTGCCCACCAGGGGGCAGCG	24.22
21872506	−	TGGCGCCACCTGGCGGCAGC	24.22

P-value

- Probability that a score at least as large as the observed score would occur in the data drawn according to the null hypothesis

- $P(S > 26.30) = \frac{1}{68 \text{ million}} = 1.5 \times 10^{-8}$
- $P(S > 17) = \frac{35}{68 \text{ million}} = 5.5 \times 10^{-7}$

- Compare to confidence threshold
  - $\alpha = 0.01$ or 0.051
- Analytical null



Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Noble, William S. "How does Multiple Testing Correction Work?." *Nature Biotechnology* 27, no. 12 (2009): 1135.

Multiple testing problem

- P-values are only valid when a single score is computed – we are computing 68 million (or 3 billion!)
- Even though $\mathrm{P}(\mathrm{S} > 17) = 5.5 \times 10^{-7}$ is a small p-value, the large number of tests makes it more likely that a significant score could occur by random chance alone

Multiple testing example

- Coin is biased if in 10 flips it landed heads at least 9 times
- Null hypothesis that coin is fair
- P(fair coin would come up heads at least 9 out of 10 times) = .0107
- We want to test 100 coins using this method
- P(all 100 fair coins are identified as fair) =

http://en.wikipedia.org/wiki/Multiple_comparisons

Multiple testing example

- Coin is biased if in 10 flips it landed heads at least 9 times
- Null hypothesis that coin is fair
- P(fair coin would come up heads at least 9 out of 10 times) = $(10 + 1) \times (1/2)^{10} = 0.0107$
- Very unlikely. We would reject null hypothesis - coin is unfair

- We want to test 100 coins using this method
- Given above probability, flipping 100 fair coins ten times each to see a *pre-selected coin* come up heads 9 or 10 times would still be very unlikely
- But, seeing *any coin* behave that way, without concern for which one, would be more likely than not
- P(all 100 fair coins are identified as fair) = $(1 - 0.0107)^{100} \approx 0.34$
- Application of our single-test coin-fairness criterion to multiple comparisons would be more likely to falsely identify at least one fair coin as unfair

http://en.wikipedia.org/wiki/Multiple_comparisons

Bonferroni correction

- Simple method
- Makes each individual test more stringent
- Controls family-wise error rate (FWER)
- FWER is the probability of at least one false rejection
- In order to make the FWER equal to at most α, reject H_{0j} if $\mathrm{p_j} \leq \frac{\alpha}{\mathrm{M}}$
  - M is number of tests performed

Table 18.5 summarizes the theoretical outcomes of M hypothesis tests. Note that the family-wise error rate is $\Pr(V \geq 1)$. Here we instead focus

TABLE 18.5. *Possible outcomes from M hypothesis tests. Note that V is the number of false-positive tests; the type-I error rate is $\mathrm{E}(V)/M_0$. The type-II error rate is $\mathrm{E}(T)/M_1$, and the power is $1 - \mathrm{E}(T)/M_1$.*

	Called Not Significant	Called Significant	Total
H_0 True	U	V	M_0
H_0 False	T	S	M_1
Total	$M - R$	R	M

on the *false discovery rate*

$$\mathrm{FDR} = \mathrm{E}(V/R). \tag{18.43}$$

The Elements of Statistical Learning

Bonferroni correction applied to CTCF motif

- Can be useful if M is relatively small, but for large M it is too conservative - calls too few significant
- $\alpha = 0.05$
- Bonferroni adjustment deems only $p < \frac{0.01}{68 \times 10^6} = 1.5 \times 10^{-10}$ significant
- Lower than smallest observed p-value
- No scores are significant

- With Bonferroni, α = 0.01 means we can be 99% sure that NONE of the scores would be observed by chance when drawn according to the null hypothesis
- Relax – instead let's control the percentage of scores drawn according to the null

Controlling the False Discovery Rate (FDR)

- Expected proportion of tests that are incorrectly called significant, among those that are called significant

Table 18.5 summarizes the theoretical outcomes of M hypothesis tests. Note that the family-wise error rate is $\Pr(V \geq 1)$. Here we instead focus

TABLE 18.5. *Possible outcomes from M hypothesis tests. Note that V is the number of false-positive tests; the type-I error rate is $\mathrm{E}(V)/M_0$. The type-II error rate is $\mathrm{E}(T)/M_1$, and the power is $1 - \mathrm{E}(T)/M_1$.*

	Called Not Significant	Called Significant	Total
H_0 True	U	V	M_0
H_0 False	T	S	M_1
Total	$M - R$	R	M

on the *false discovery rate*

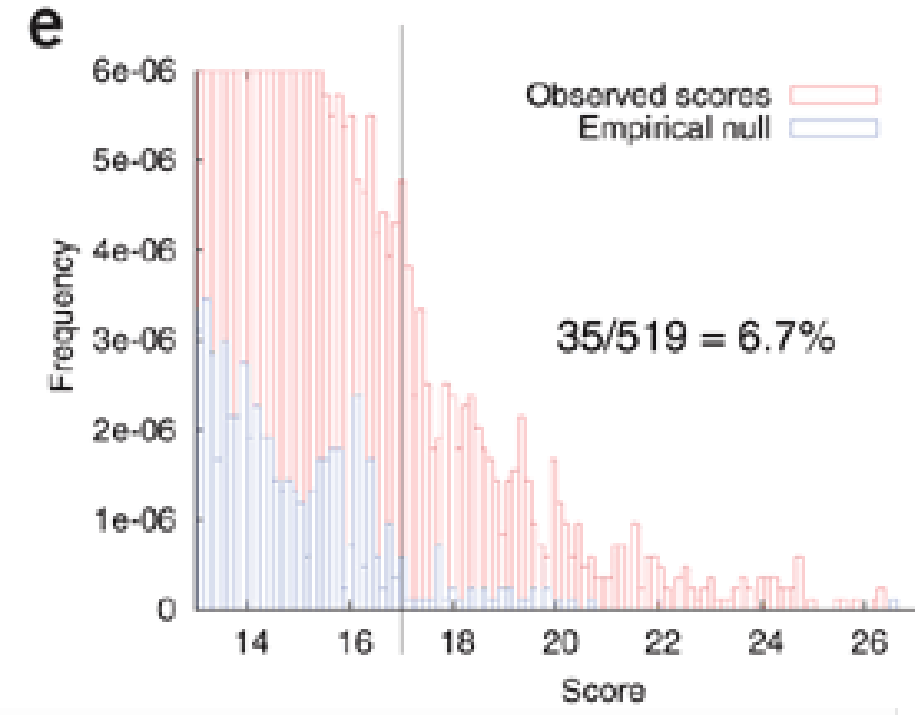
$$\mathrm{FDR} = \mathrm{E}(V/R). \tag{18.43}$$

The Elements of Statistical Learning

Controlling the False Discovery Rate (FDR)

- # null scores ≥ 17 (blue)
  - $s_{null1} = 35$
- # observed scores ≥ 17 (red)
  - $s_{obs1} = 519$

- $\frac{s_{null1}}{s_{obs1}} = 6.7\%1$
- This computes FDRs from scores
- Use Benjamini-Hochberg to compute FDR from p-values



Benjamini-Hochberg (BH)

Algorithm 18.2 *Benjamini–Hochberg (BH) Method.*

1. Fix the false discovery rate α and let $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(M)}$ denote the ordered p-values

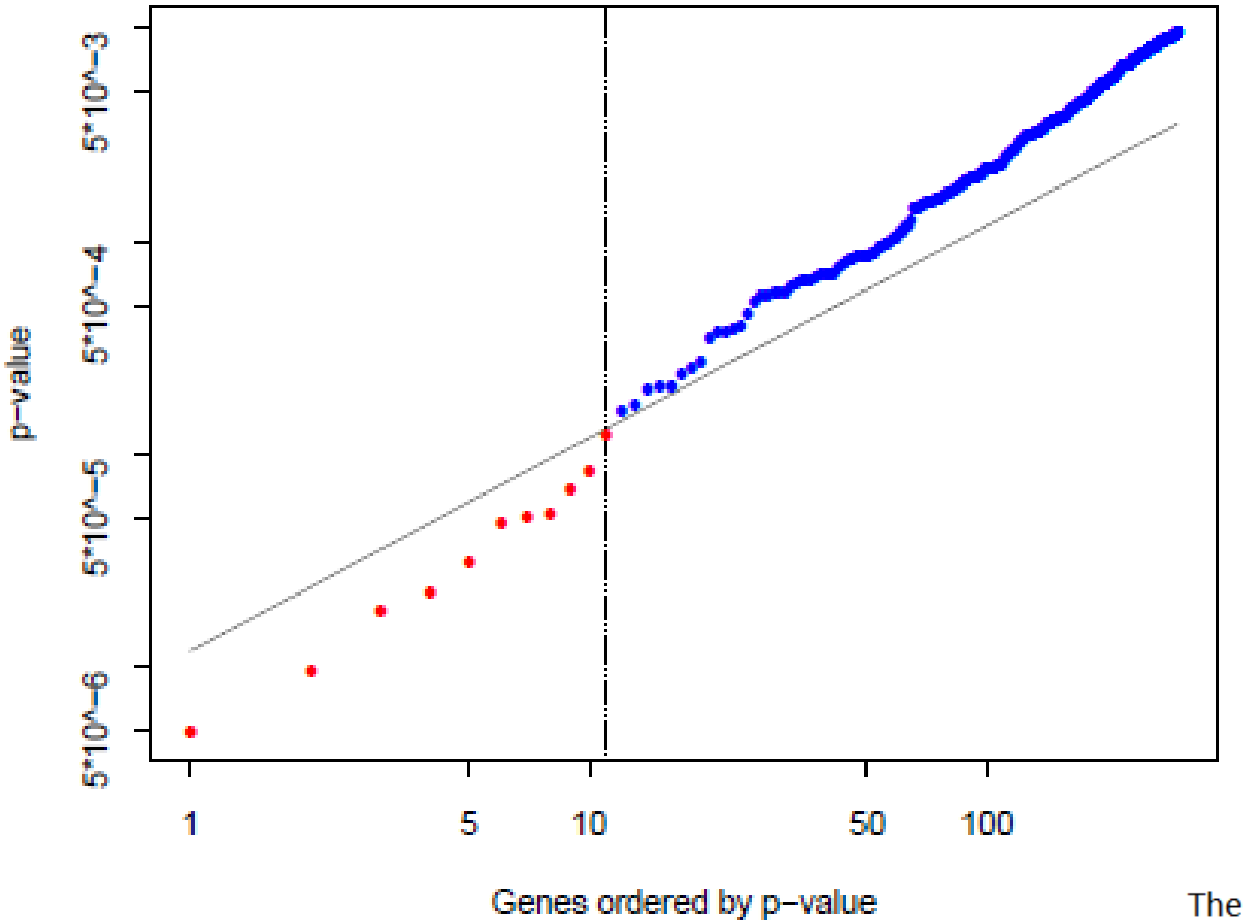
2. Define

$$L = \max\left\{j : p_{(j)} < \alpha \cdot \frac{j}{M}\right\}. \tag{18.44}$$

3. Reject all hypotheses H_{0j} for which $p_j \leq p_{(L)}$, the BH rejection threshold.

The Elements of Statistical Learning

Benjamini-Hochberg (BH)



Multiple testing problems in biology

- Massive scale of recent biology creates opportunities for spurious discoveries

- *Scanning a genome for occurrences of transcription factor binding sites*
- Searching a protein database for homologs of a query protein/BLAST search
- Identifying differentially expressed genes from microarray/RNA-seq
- Genome-wide association studies

Remember!

- Pset 1 posted – due Feb 20th (no AI problem)
- Pset 2 posted soon – Due Mar 13th
  - Programming problem
- Python tutorial – Feb 10th (Monday) 4-5pm.
- Project interests due – Feb 11th
  - Name, program, previous experience, interest in computational biology
  - We'll post these next week for you to find groups for project
- Office hours posted soon

7.91J / 20.490J / 20.390J / 7.36J / 6.802J / 6.874J / HST.506J Foundations of Computational and Systems Biology
Spring 2014