AMSC 663 midterm progress report

Application of the false discovery rate to microbial community comparison

james robert white
Fall 2007

Advisor: Mihai Pop, CBCB.

Outline

- Brief background in biology, "microbial census"
- Introduce problem of comparing communities
- Multiple hypothesis testing solution
- Validation
- Application
- Future work

Background

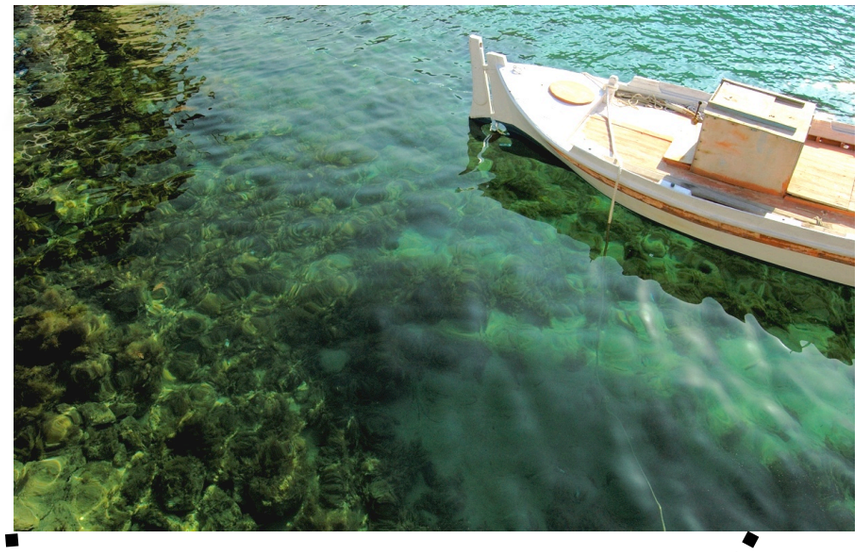
- Every microbe has a **conserved** *gene* called16S rRNA.
- Easy to recognize and exists in all known microbes.

Bacillus anthracis

E. coli

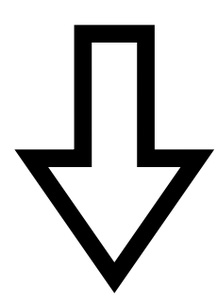
Mycobacterium tuberculosis

Metagenomics



16S gene

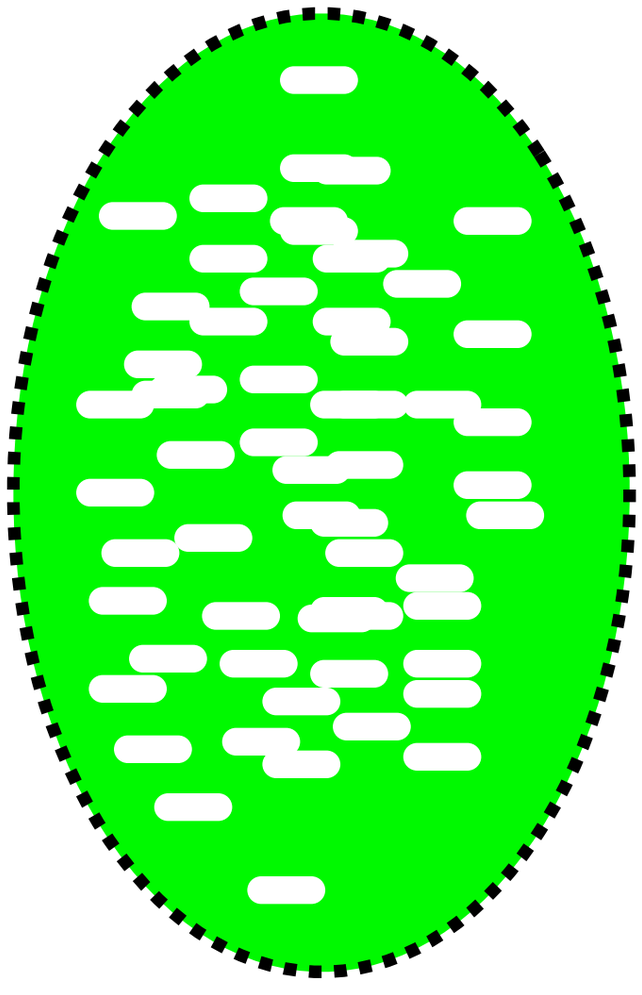
…TAGTCCATGACAG
TACCGTACAAAA …

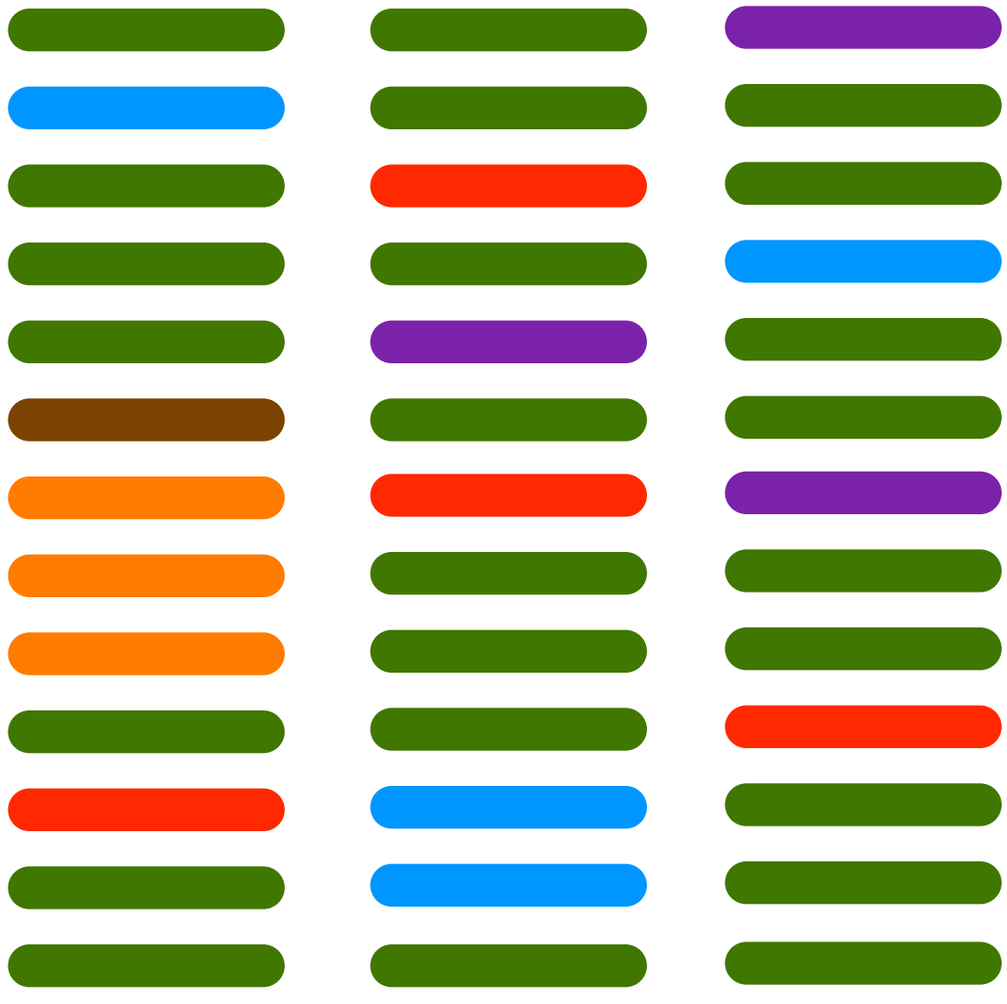


Prochlorococcus marinus

Background

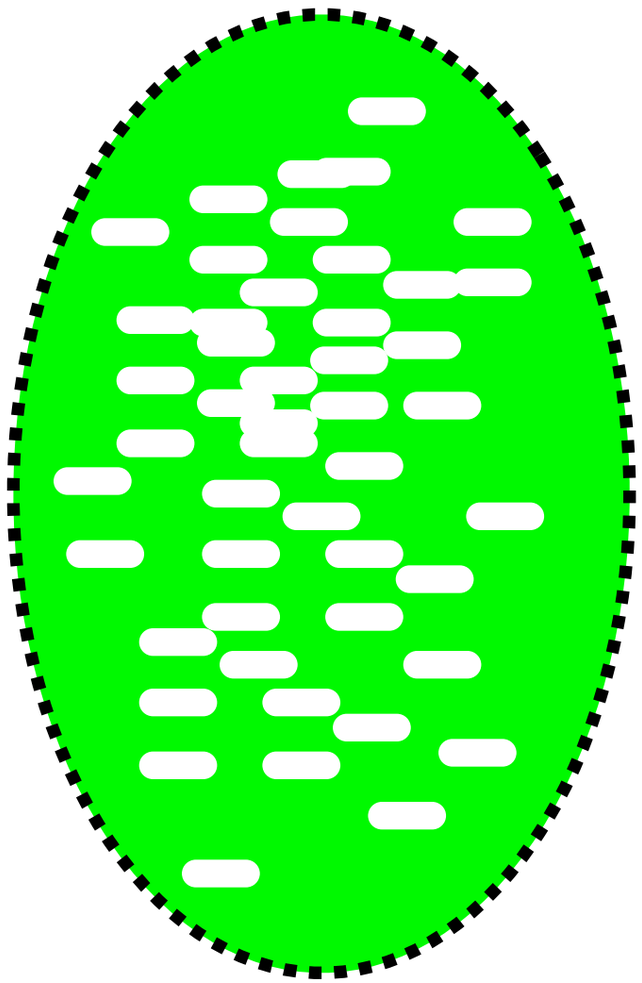
Environment
(radioactive waste)

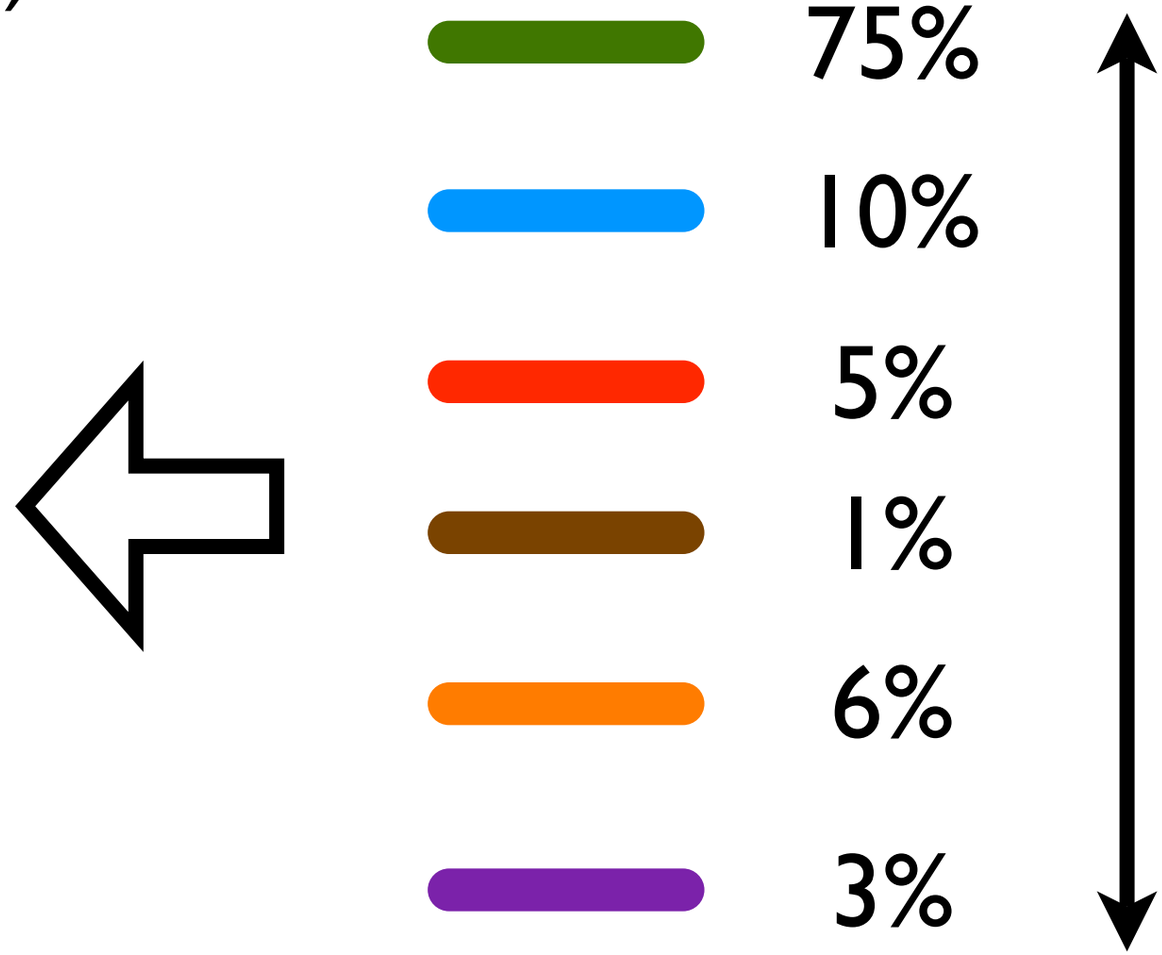




Background

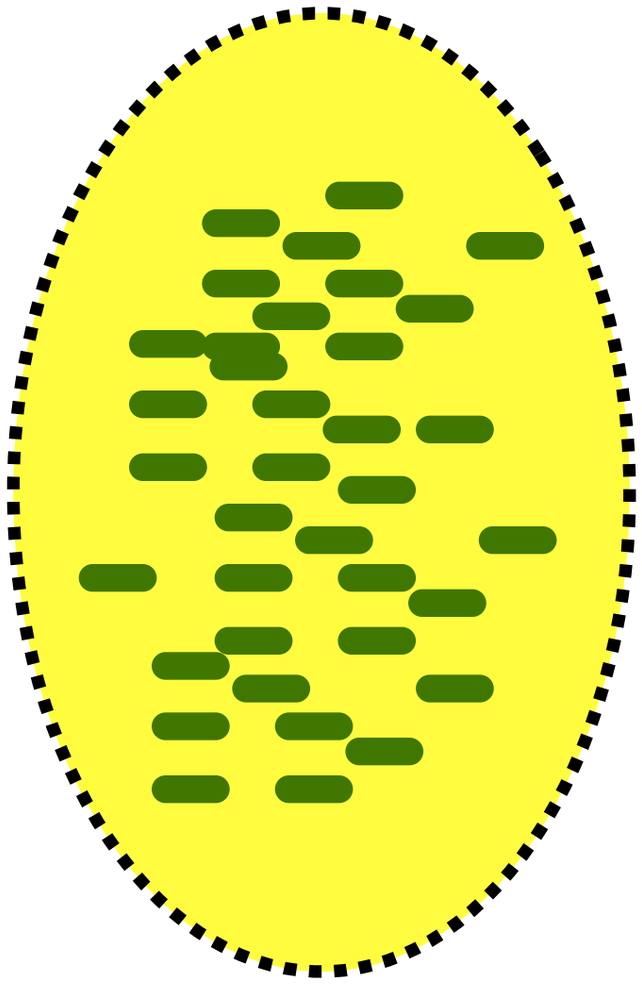
Environment
(radioactive waste)





The Problem

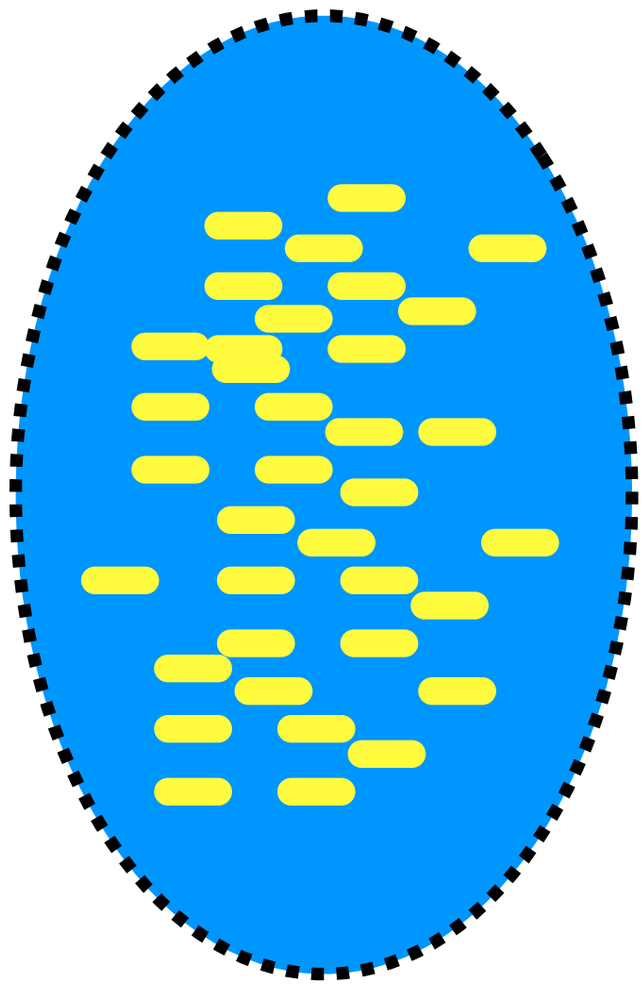
(Healthy ears)



How do two environments differ?

Which organisms are differentially abundant?

(Sick ears)



Taxa abundance matrix

	Healthy ears			Sick ears			
	p1	p2	p3	p4	p5	p6	p7
t1	243	300	120	0	43	21	66
t2	12	34	32	0	0	0	0
t3	0	3	10	200	140	134	70
t4	42	4	12	54	76	80	60
t5	2	0	10	4	6	0	0
t6	5	5	3	15	12	0	43

Differential abundance

- convert frequencies to relative **proportions**.
- compute sample means, variances.

	p1	p2	p3	p4	p5	p6	p7
t1	.37	.42	.35	0.0	.10	.05	.17

$$\bar{x}_{i1} = \frac{1}{n_1} \sum_{j \in \text{treatment 1}} a_{ij}$$

$$s_{i1}^2 = \frac{1}{n_1 - 1} \sum_{j \in \text{treatment 1}} \left(a_{ij} - x_{i1} \right)^2$$

$$t_i = \frac{\bar{x}_{i1} - \bar{x}_{i2}}{\sqrt{\frac{s_{i1}^2}{n_1} + \frac{s_{i2}^2}{n_2}}}.$$

Hypothesis tests

- So for each taxa, T_i, we perform a hypothesis test of proportions:
  - H_o: $\mu_{healthy} = \mu_{sick}$
  - H_A: $\mu_{healthy} != \mu_{sick}$
- We obtain a test statistic *ti*
- Reject or accept the null hypothesis?

Hypothesis tests

- suppose we perform *M* tests:

	accept null	reject null	total
null true	M_{aT}	M_{rT}	M_T
null false	M_{aF}	M_{rF}	M_F
total	M-M_r	M_r	M

Hypothesis tests

- criteria for rejecting H_o
  - *p*-value and *q*-value estimates of significance
  - choose a threshold α, and reject if *p* or *q* $\leq \alpha$
- same criteria for all tests

p vs. *q* values

- if you reject all H_o with $p \leq 0.05$, you expect 5% of all true H_o to be false positives.
- if you reject all H_o with $q \leq 0.05$, you expect 5% of all rejected H_o to be false positives.
- M = 10 tests? 1000 tests?

Project

- implement algorithms for calculating p and q values for hypothesis tests
- coded in R: free statistical software package with great visualization features.
- validation
- applied to real 16S data

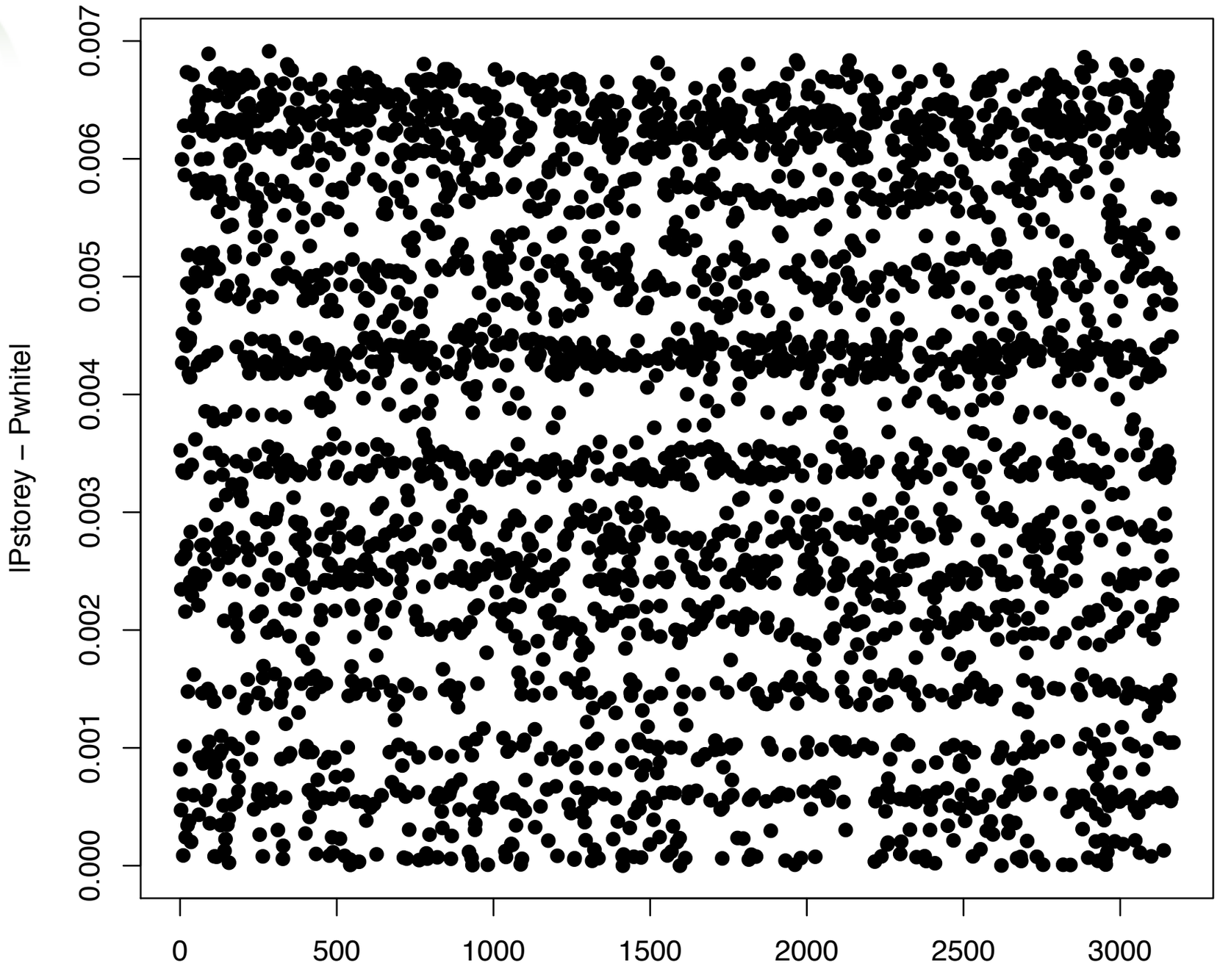
Validation

- Hedenfalk dataset, 2001, **NEJM**.
- microarray study of two forms of hereditary breast cancer (BRCA1 and BRCA2)
- looking for differentially active genes among 3,170 total genes.
- activity level of a gene <=> abundance level of a taxa.

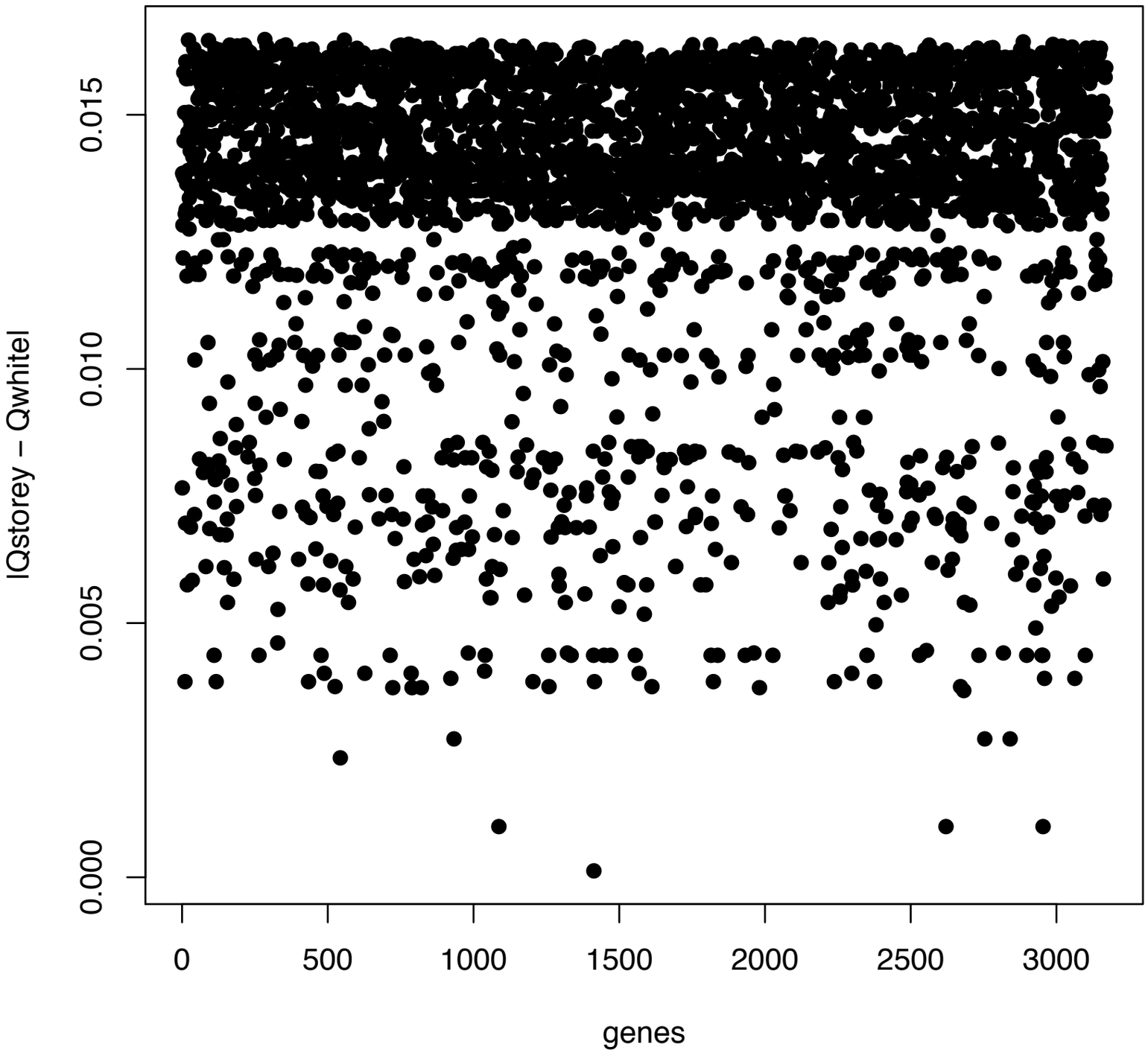
Validation

- Storey & Tibshirani, 2003, **PNAS,** calculated *q* and *p* values for all 3,170 genes.
- I computed my own *p*'s and *q*'s using my software.
- |Pstorey - Pwhite|, |Qstorey - Qwhite|
- rejected all H_o with $Q_{white} \leq 0.05$.

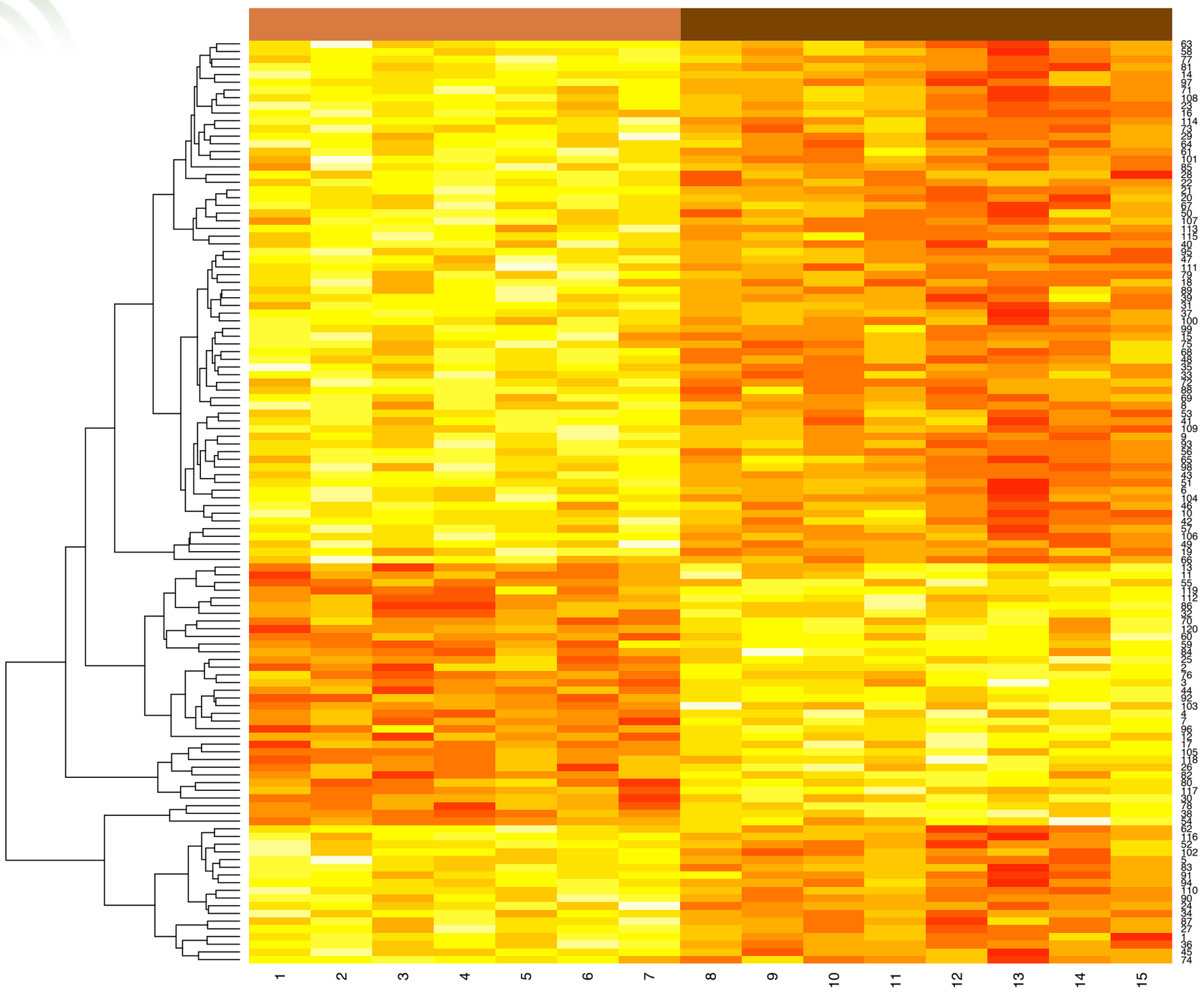
Hedenfalk p values



Hedenfalk *q* values



Differential expression

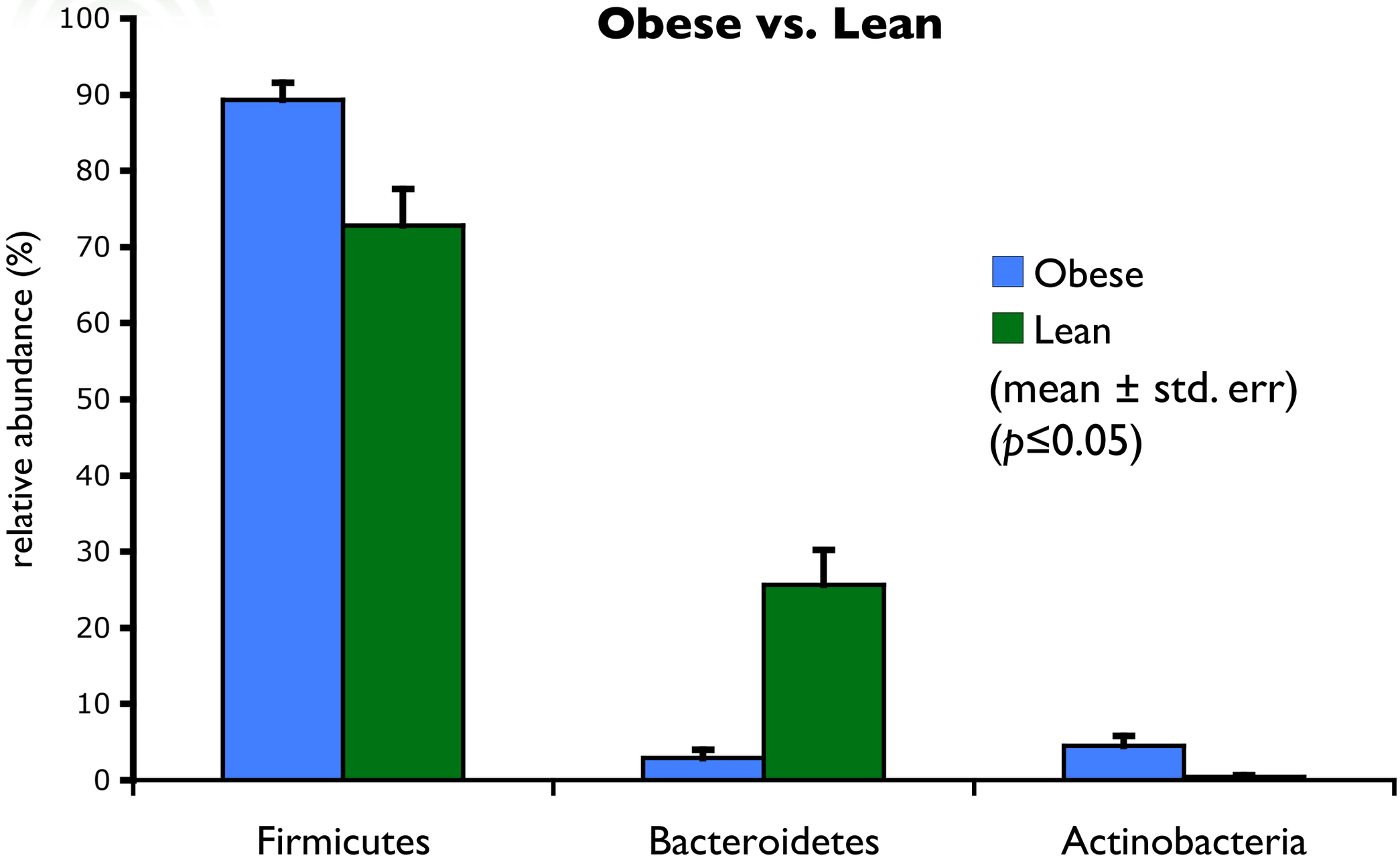


($q \leq 0.05$)

Real 16S data

- Ley *et al.* 2006, **Nature**
- metagenomic study of differentially abundant taxa between human guts of obese (12) and lean (5) people
- found significant differences between two high level taxa: *Bacteroidetes* and *Firmicutes*
- we generated taxa abundance matrices from original data and tried to replicate their results.

Ley *et al.* data



0	.01	.05	.1	.2	.3	.4	.5<

Phyla

g1 g2

p1						
p2						
p3						
p4						
p5						
p6						
p7						
p8						

Classes

Future work

- **December**
  - Consider statistical methodology given sampling issues.
  - Develop at least two methodologies to compare.
  - Design broad simulation to test q-values vs. p-values.
- **January**
  - Finish broad simulation.
  - Finalize statistical methodology.
  - Finish application of software to Ley data.
- **February**
  - Apply best method to additional metagenomic data.
  - Develop documentation for software.
- **April**
  - Complete final draft of report including edits from advisor.
  - Submit polished version of our software to BioConductor group.
- **May**
  - Deliver final report.
  - Final presentation.

Acknowledgments

- Mihai Pop, CBCB
- Frank Siewerdt, ANSC
- Ken Ryals, AMSC
- Paul Smith, STAT
- Radu Balan, CSCAMM
- Aleksey Zimin, IPST


Questions?