Statistical methods for microbial community comparison

james robert white
April 2008

Advisor: Mihai Pop, CBCB.

Outline

- Brief background in metagenomics
- Introduce my problem
- Methods
- Applications
- Future work

Biology!

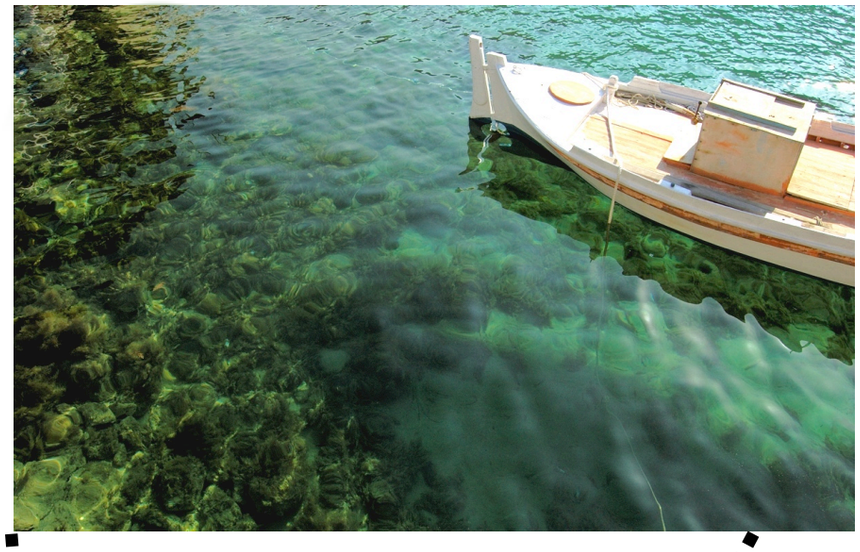
- Every microbe has a **conserved** *gene* called16S rRNA.
- Easy to recognize and exists in all known microbes.

Bacillus anthracis

E. coli

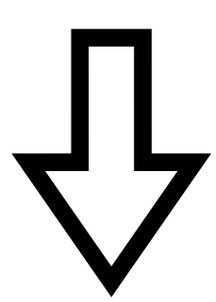
Mycobacterium tuberculosis

Metagenomics



16S gene

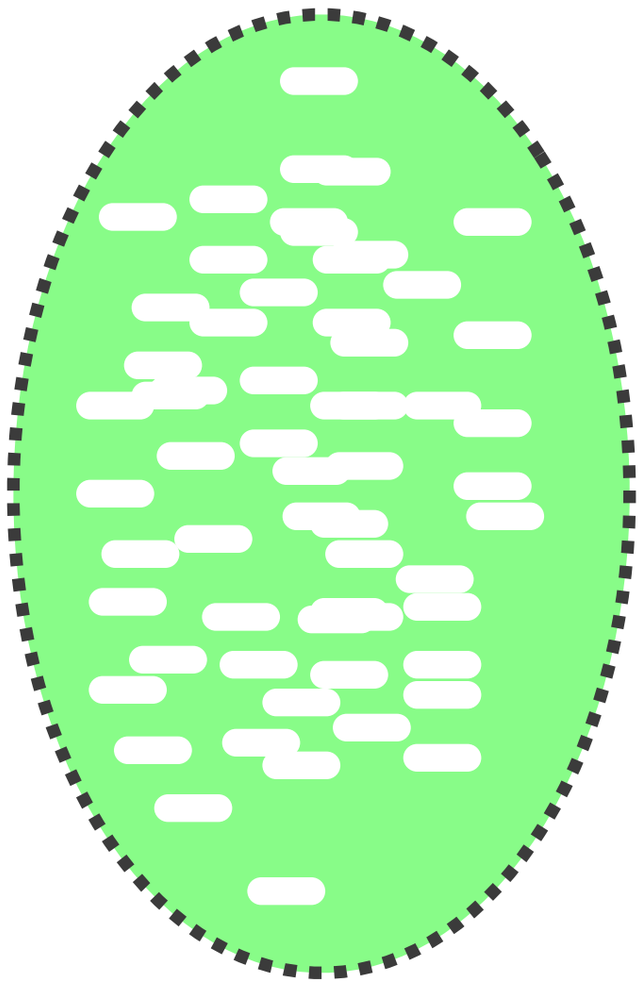
…TAGTCCATGACAG
TACCGTACAAAA …

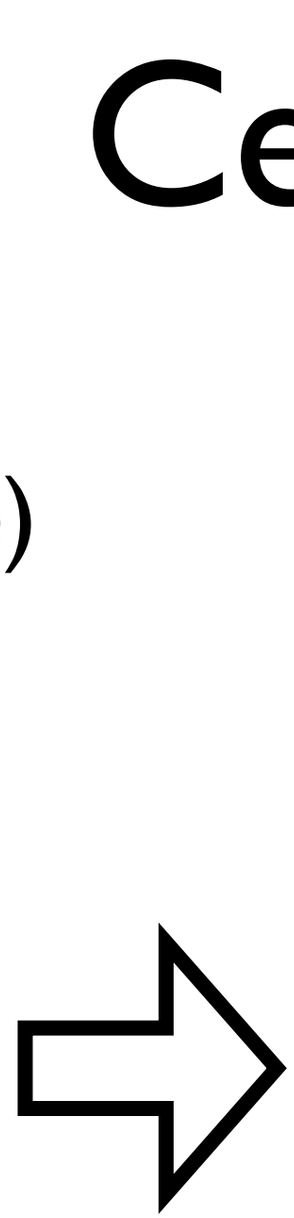


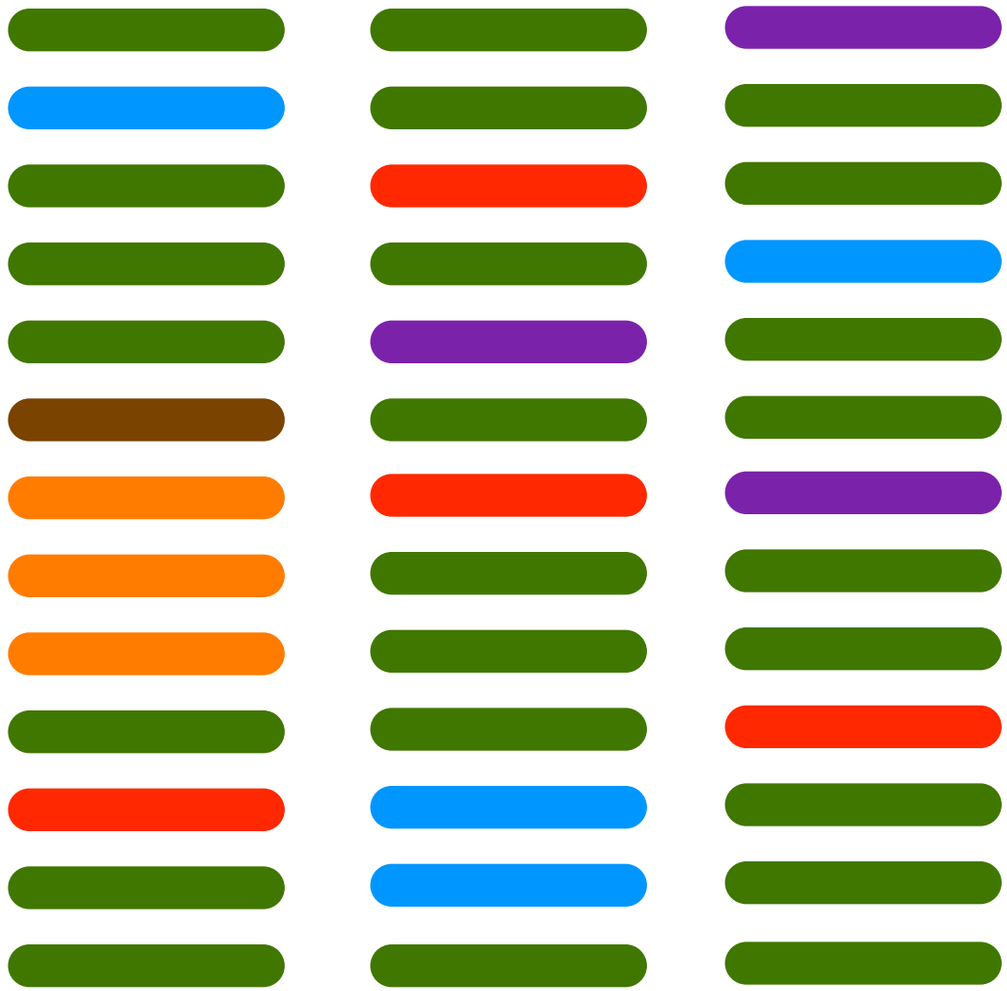
Prochlorococcus marinus

Census

Environment
(radioactive waste)

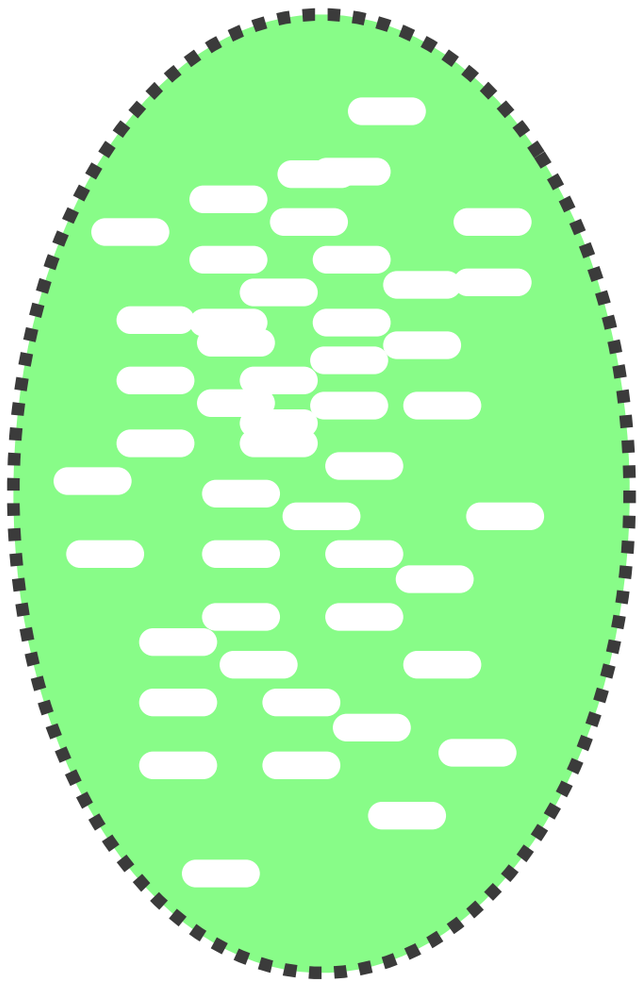


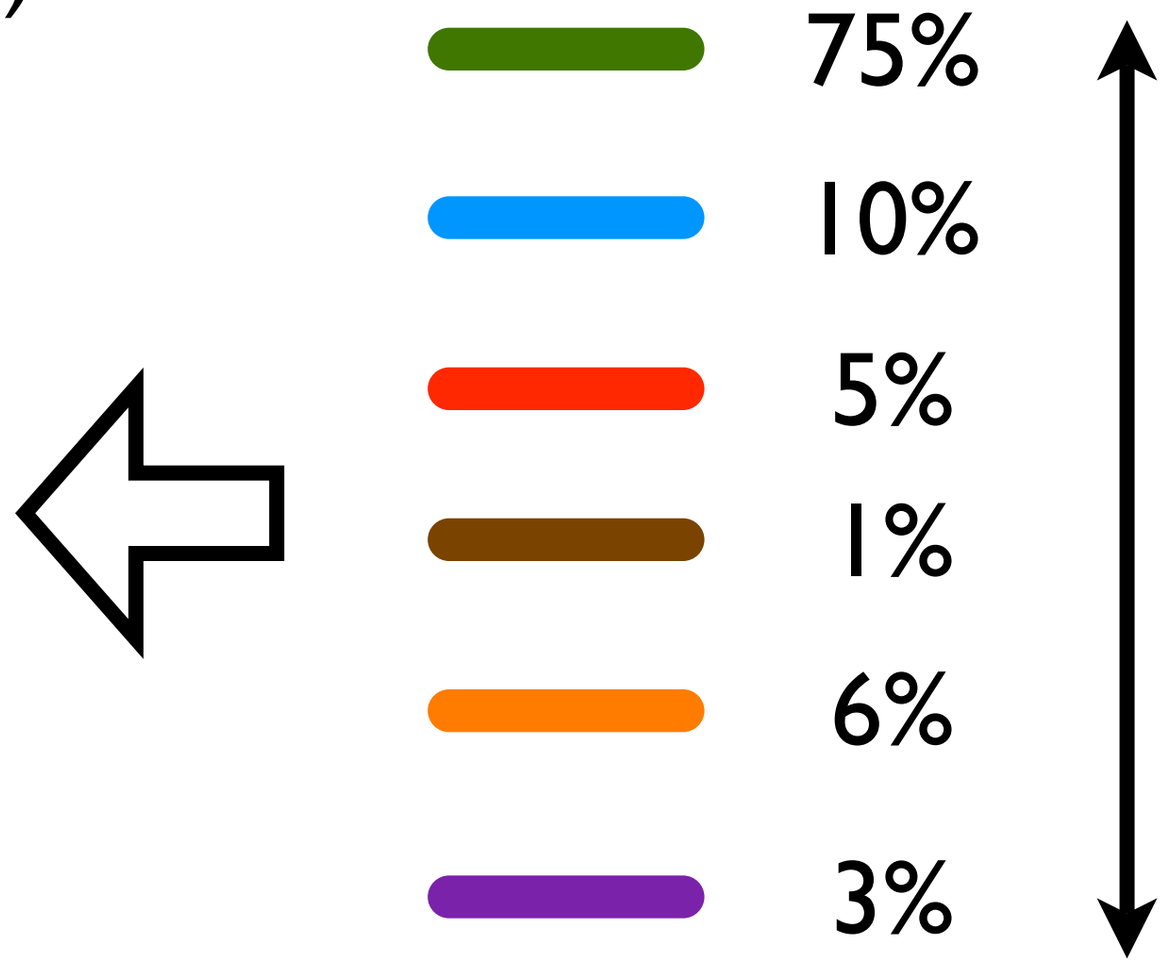




Census

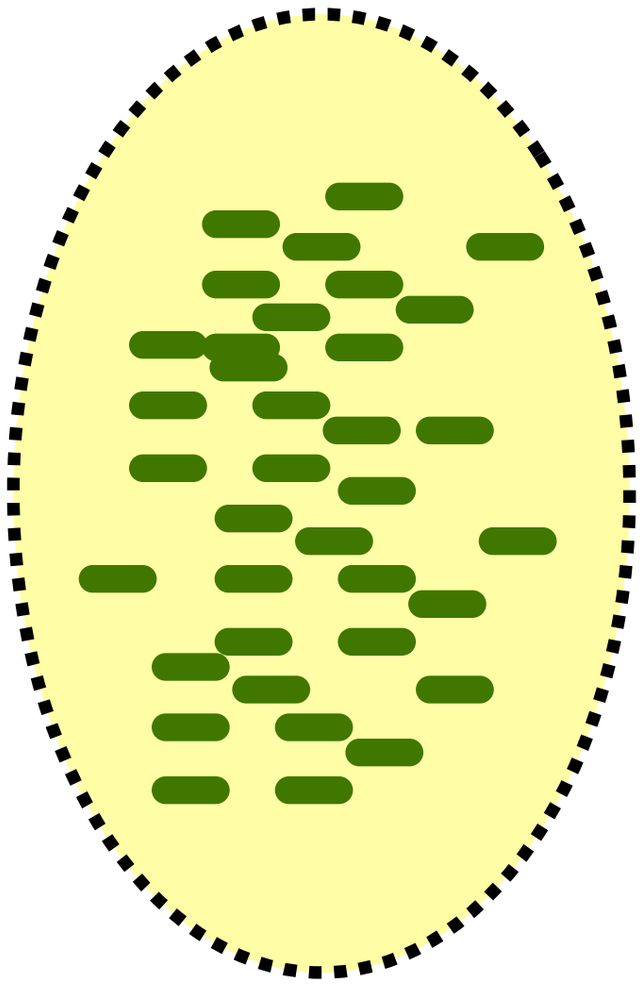
Environment
(radioactive waste)





The Problem

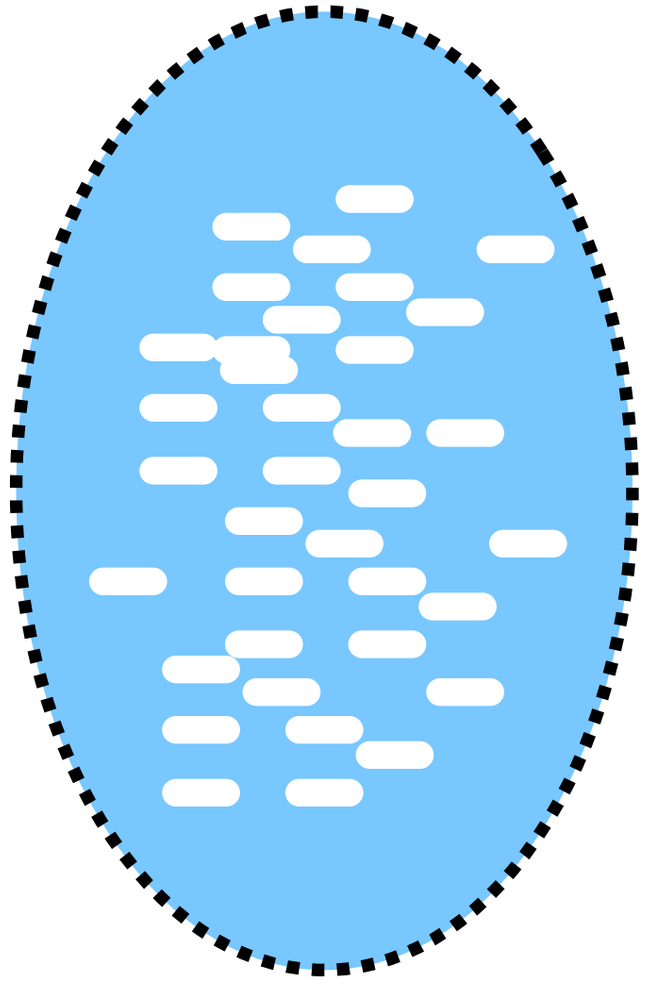
(Healthy colons)



How do two environments differ?

Which organisms are differentially abundant?

(Sick colons)



Taxa abundance matrix

	Healthy colons			Sick colons			
	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

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

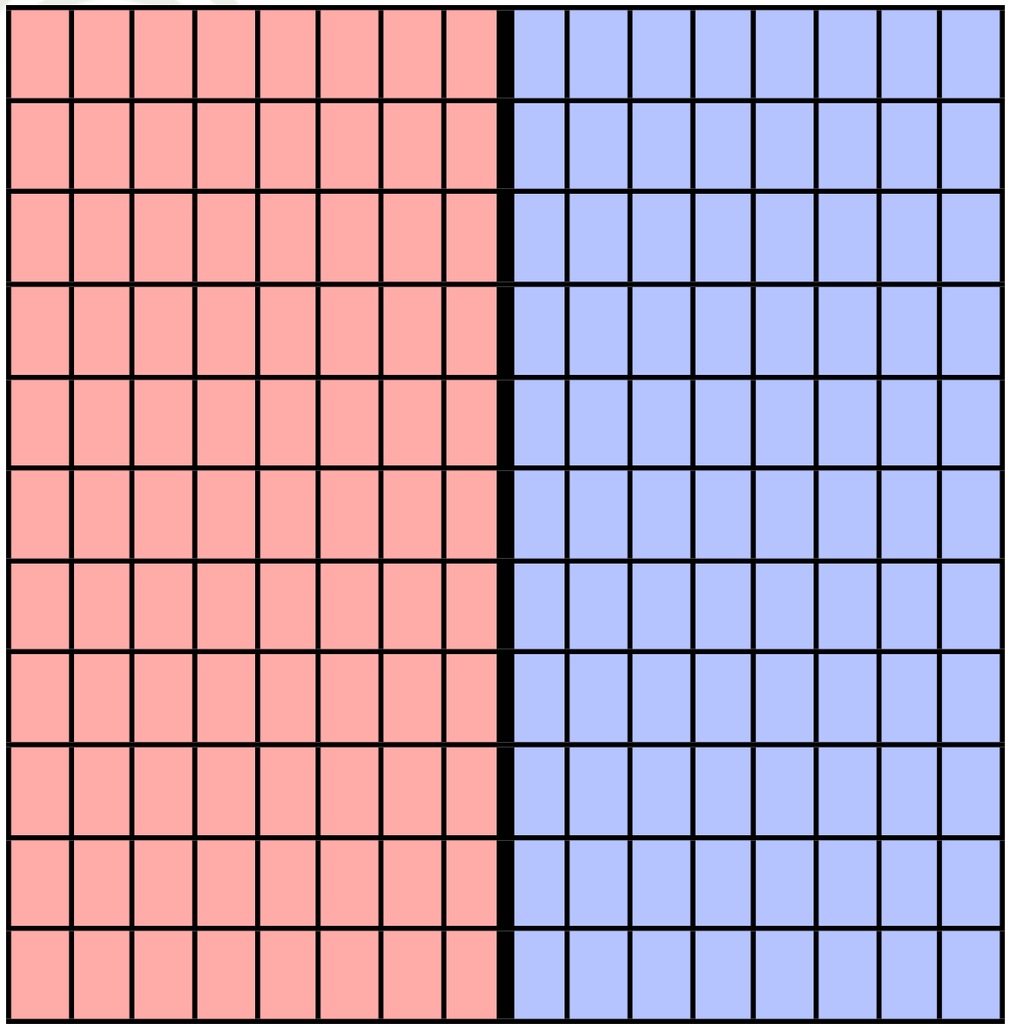
$$s_{i1}^2 = \frac{1}{n_1 - 1} \sum_{j \in \text{treatment 1}} \left(a_{ij} - x_{i1} \right)^2$$

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

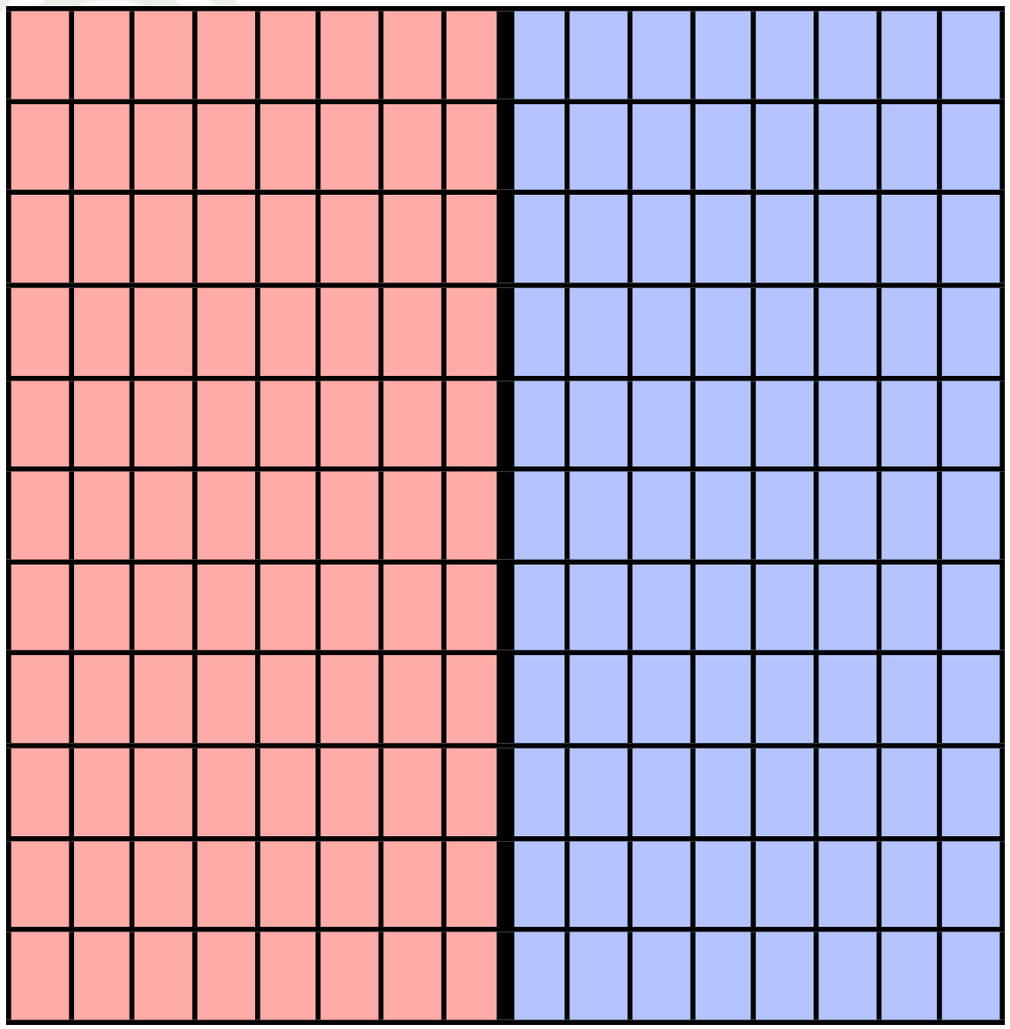
Hypothesis tests

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

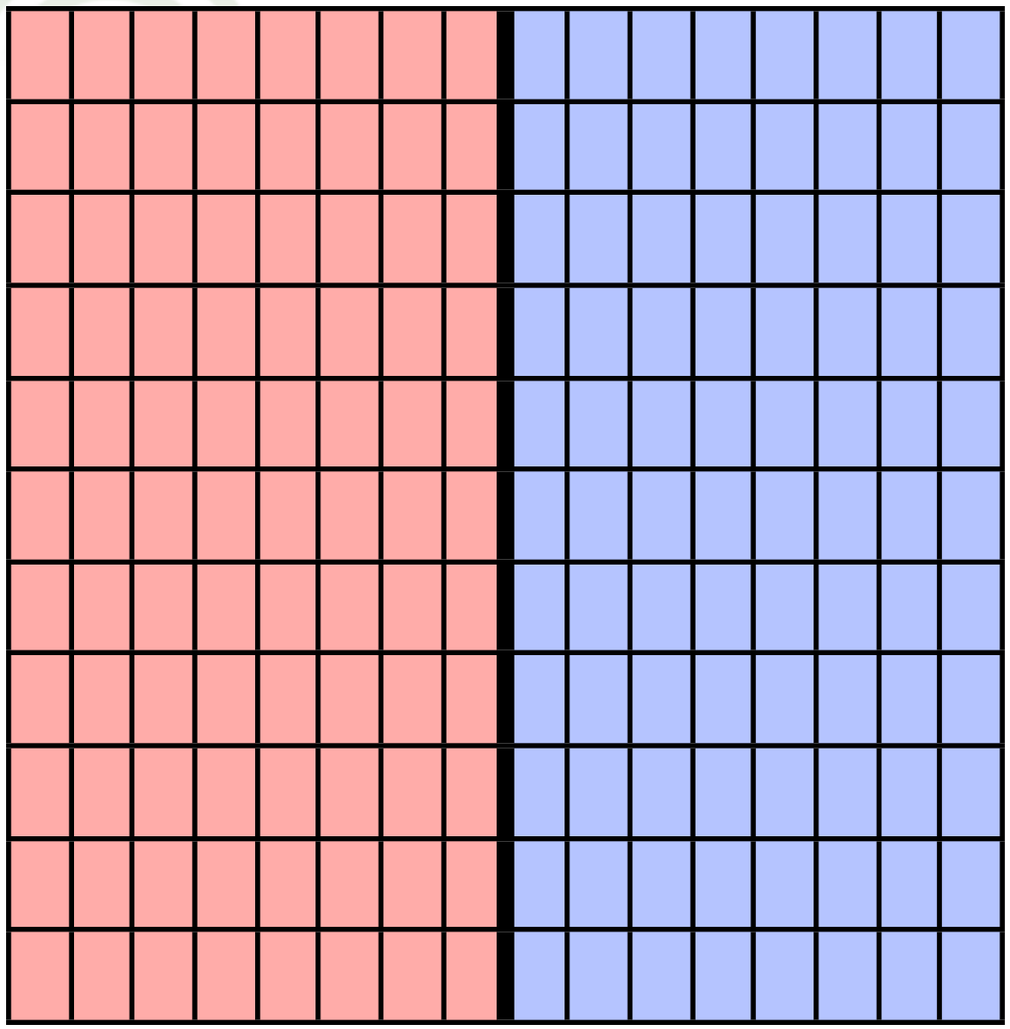
Permuted p-values



Permuted p-values

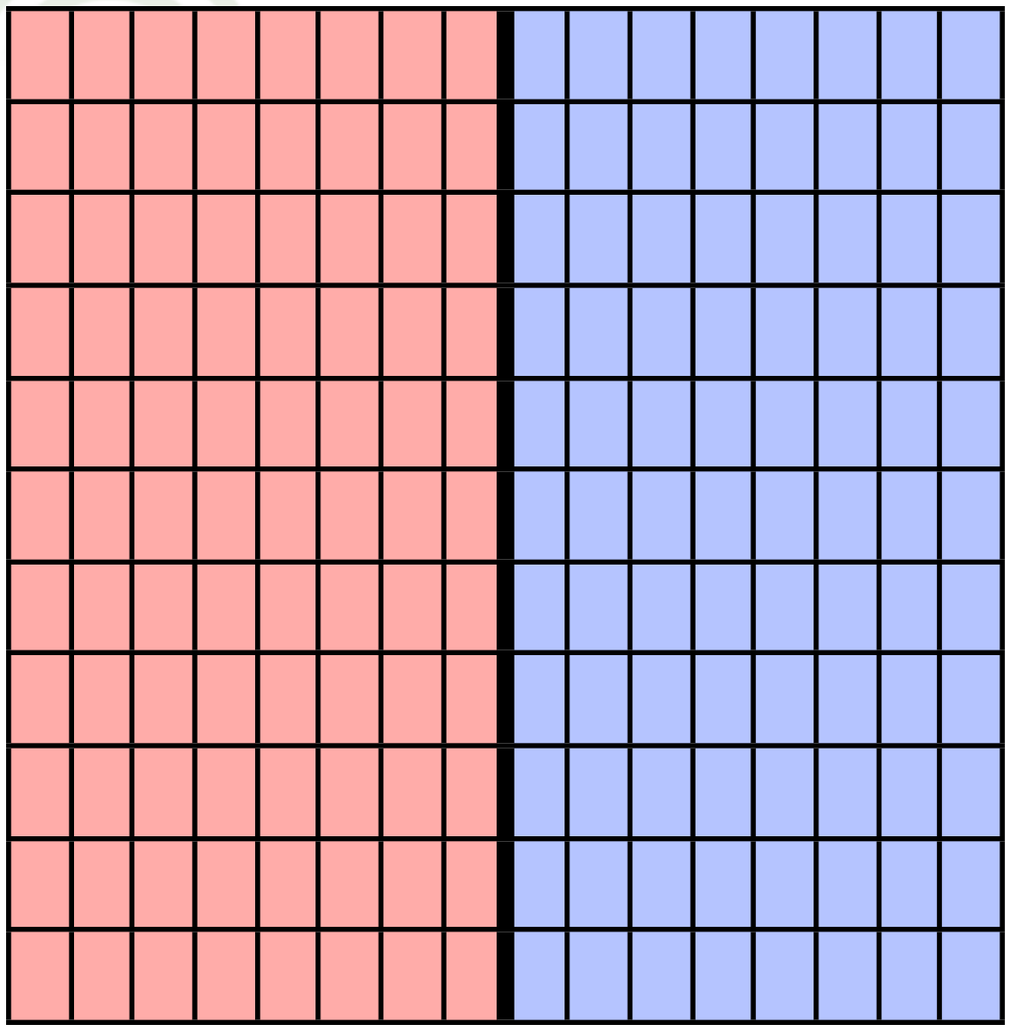


Permuted p-values

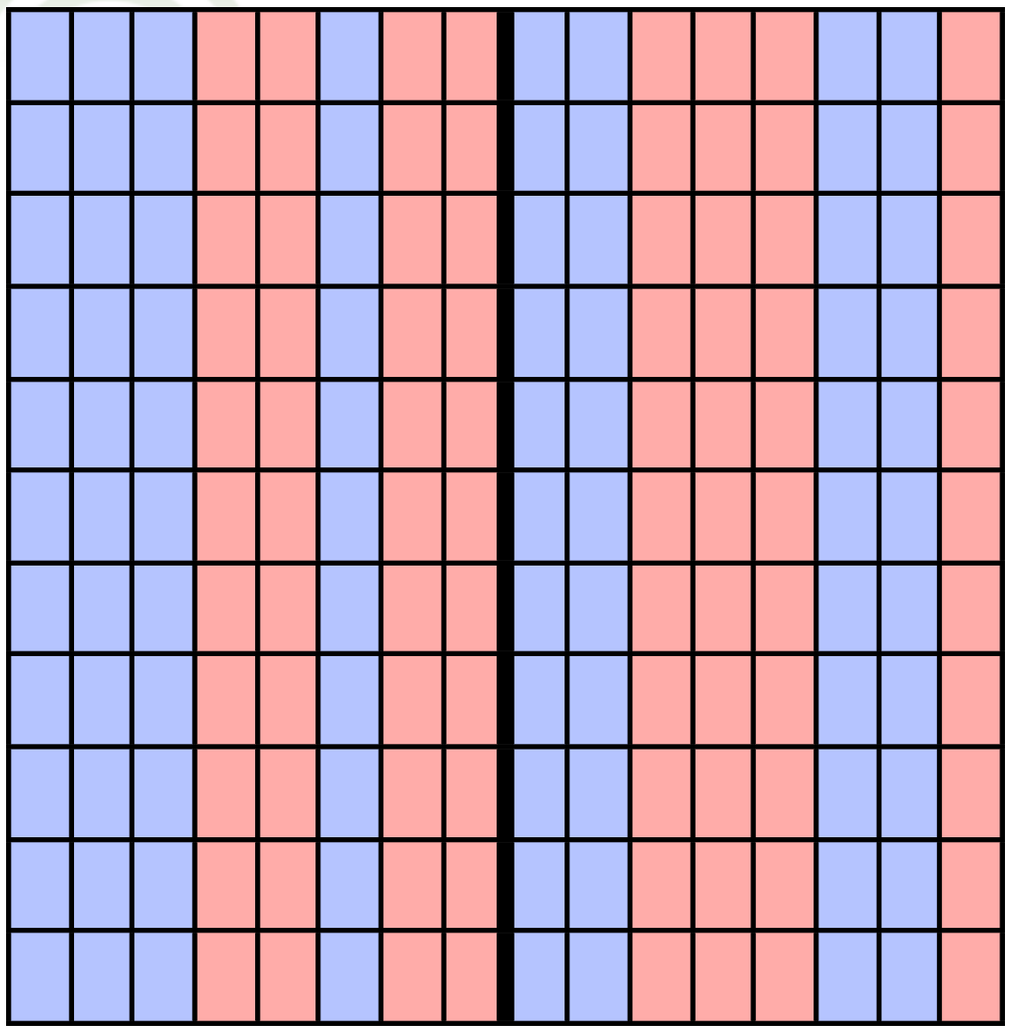


→ $t1 \ldots tM$

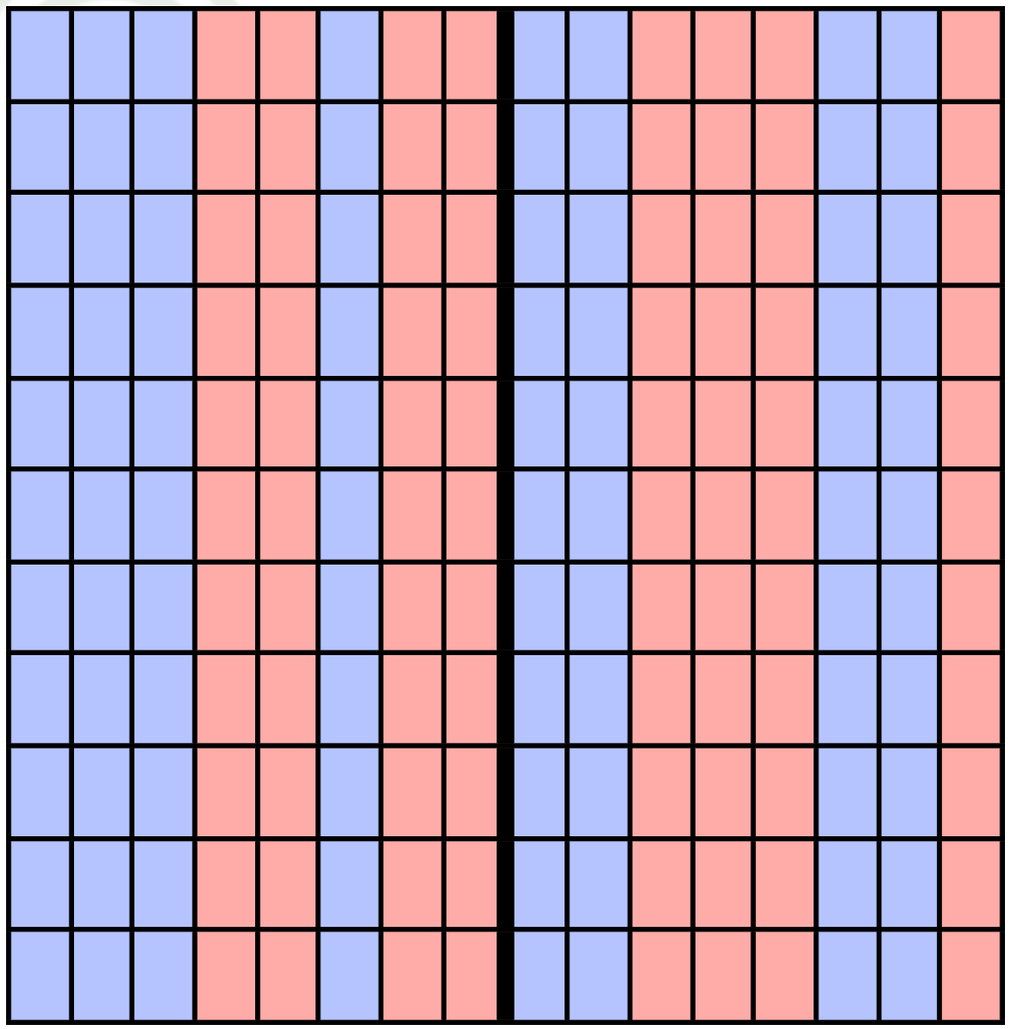
Permuted p-values



Permuted p-values

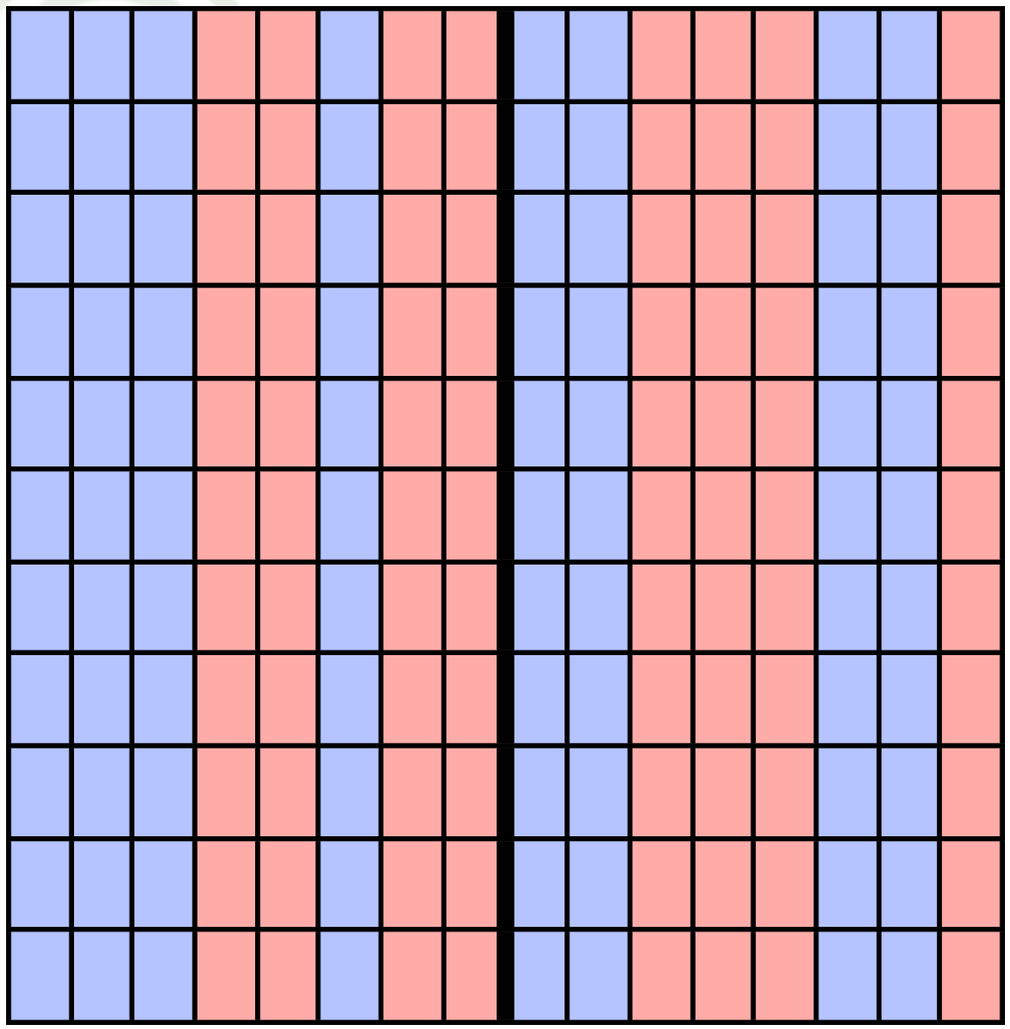


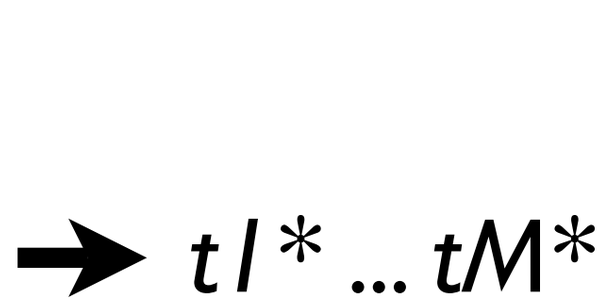
Permuted *p*-values



→ *t1** … *tM**

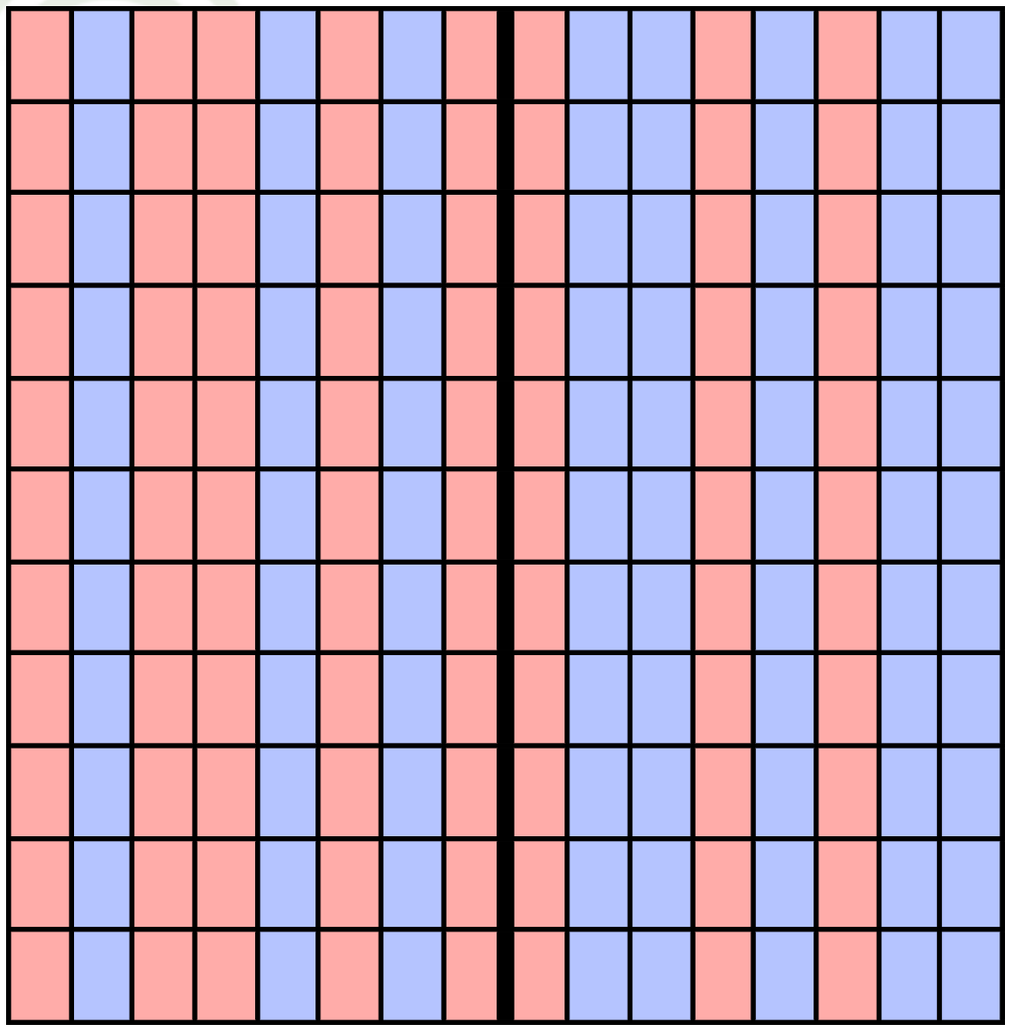
Permuted p-values





simulated t's

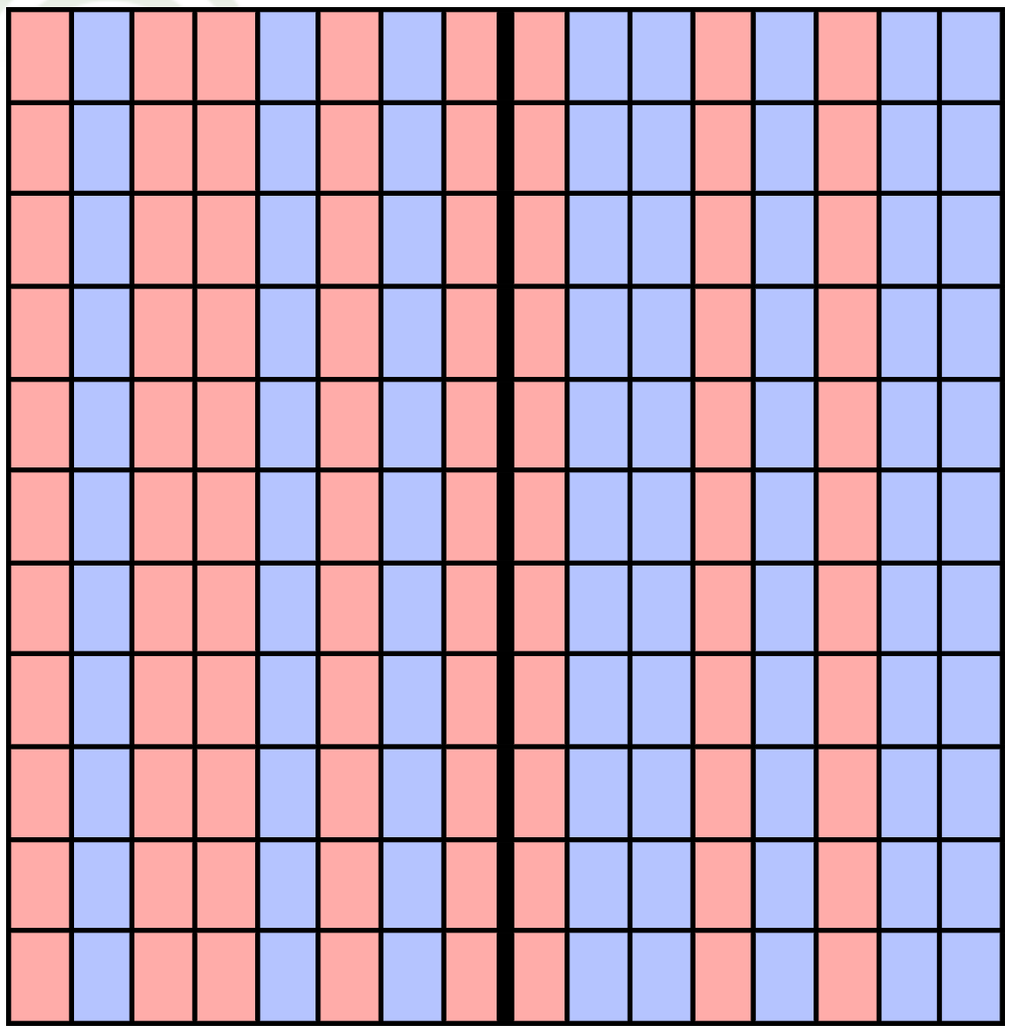
Permuted *p*-values



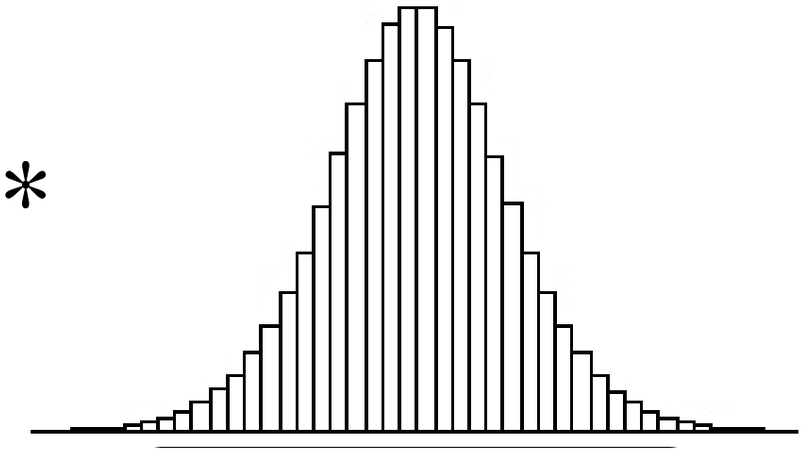
→ *t1** … *tM**

simulated *t*'s

Permuted p-values



$t1^* \ldots tM^*$



Multiple tests

- criteria for rejecting H_o
  - p-value estimates of significance
  - choose a threshold α, and reject if $p \leq \alpha$
  - if you reject all H_o with $p \leq 0.05$, you expect 5% of all **true** H_o to be false positives.
  - M = 10 tests? 1000 tests? 100,000 tests?
  - Bonferroni correction

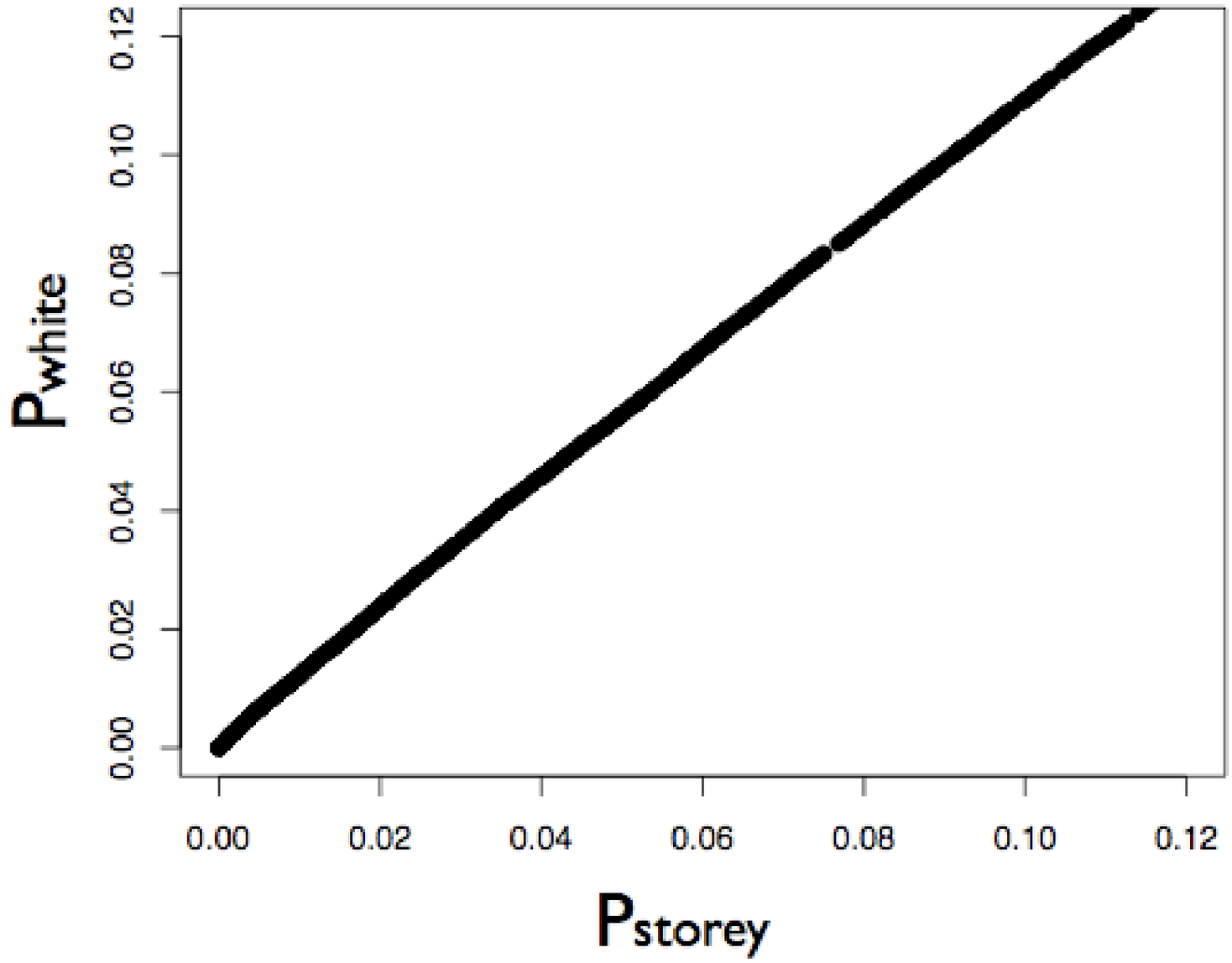
FDR alternative

- False Discovery Rate - "the rate of significant features that are truly null."
- Analog to *p*-value => *q*-value
- 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.
- e.g. 10,000 tests.

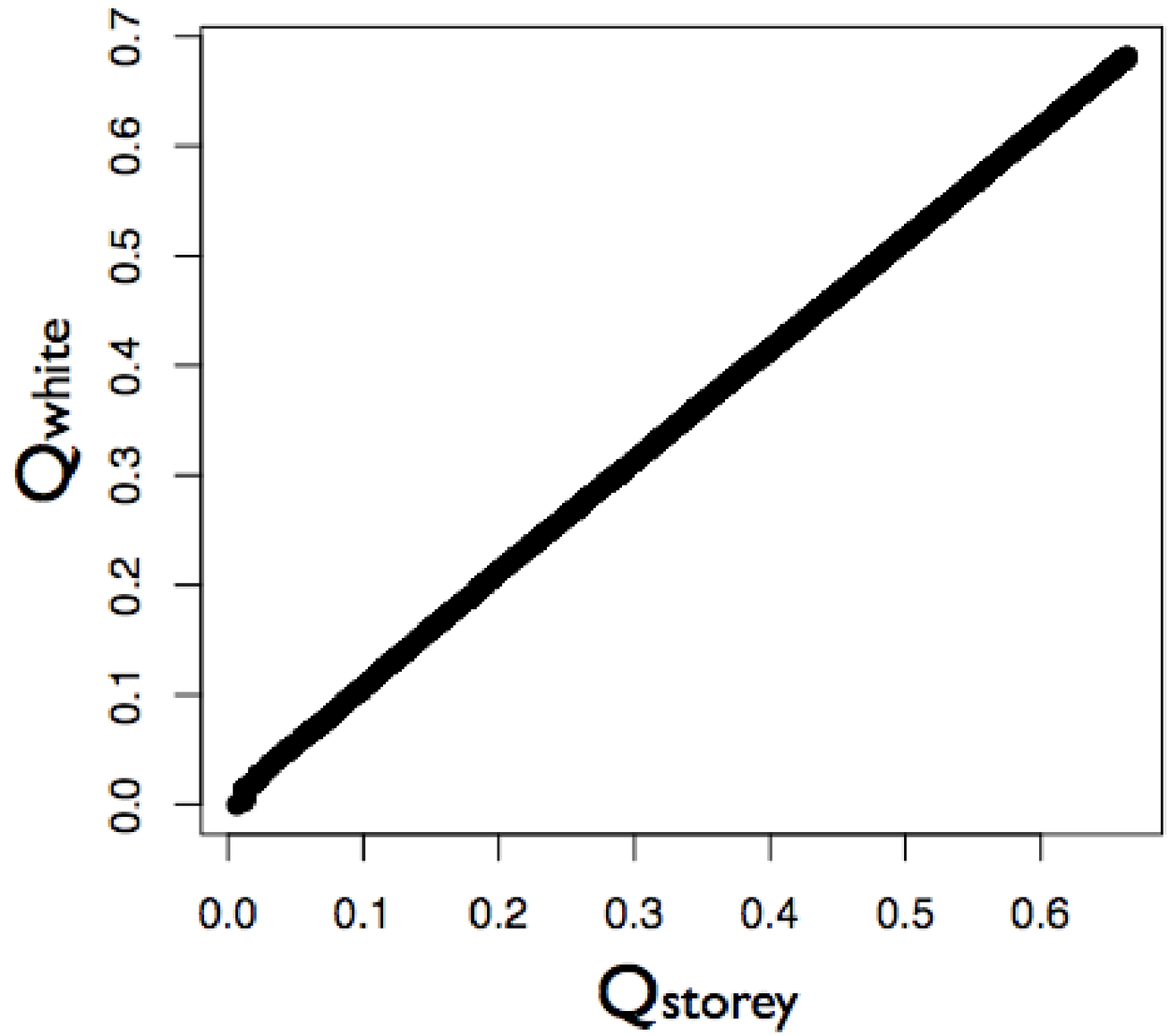
Multiple tests

	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

Hedenfalk p values



Hedenfalk *q* values



Additional issues

- Low frequency taxa.

	Healthy colons			Sick colons			
	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

heuristic

- N = total number of samples from treatment
- $N^*p \geq 25$ to use the *t* statistic
- $p \geq 25/N$
- $p \geq 25/5000 = 0.005$

small frequencies

- what about 25/N > p ?
- if p is this small, indicates small variance among subjects, so merge all samples into one large sample.
- use Fisher's exact test to find an appropriate p value.

e.g.

s1	s2	s3	s4	s5	s6	s7	s8	s9	s10
0	1	1	2	3	0	0	1	0	0
50	49	49	48	47	50	50	49	50	50

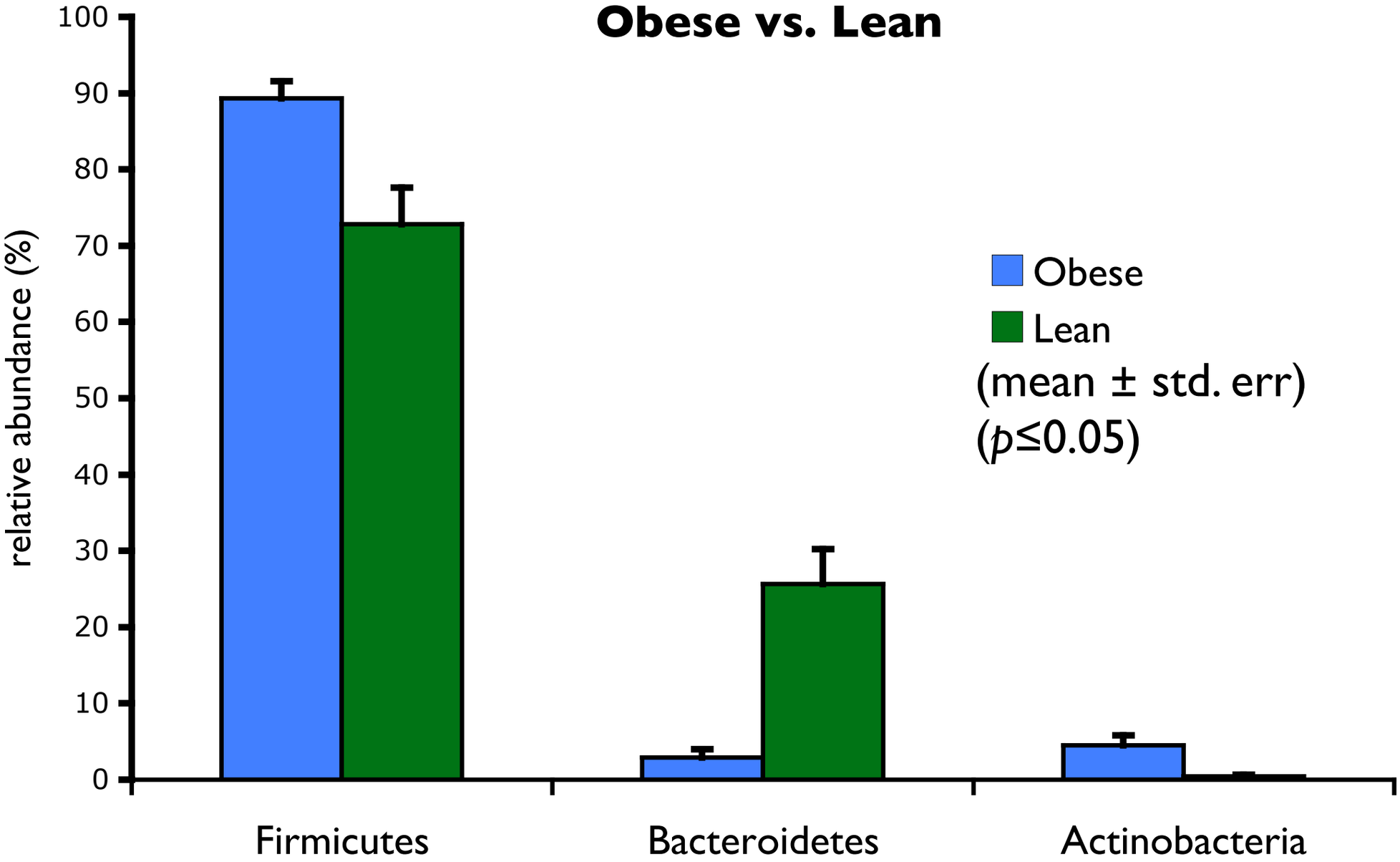
⇩

	g1	g2
S	7	1
F	243	249

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



human vs. mouse

- Two 16S distal gut studies:
  - 5 lean control humans (Ley *et al.*, 2006)
  - 12 lean control mice (Ley *et al.*, **PNAS**, 2005)
- 6,250 16S sequences.
- assigned using the RDP II Bayesian classifier

human vs. mouse

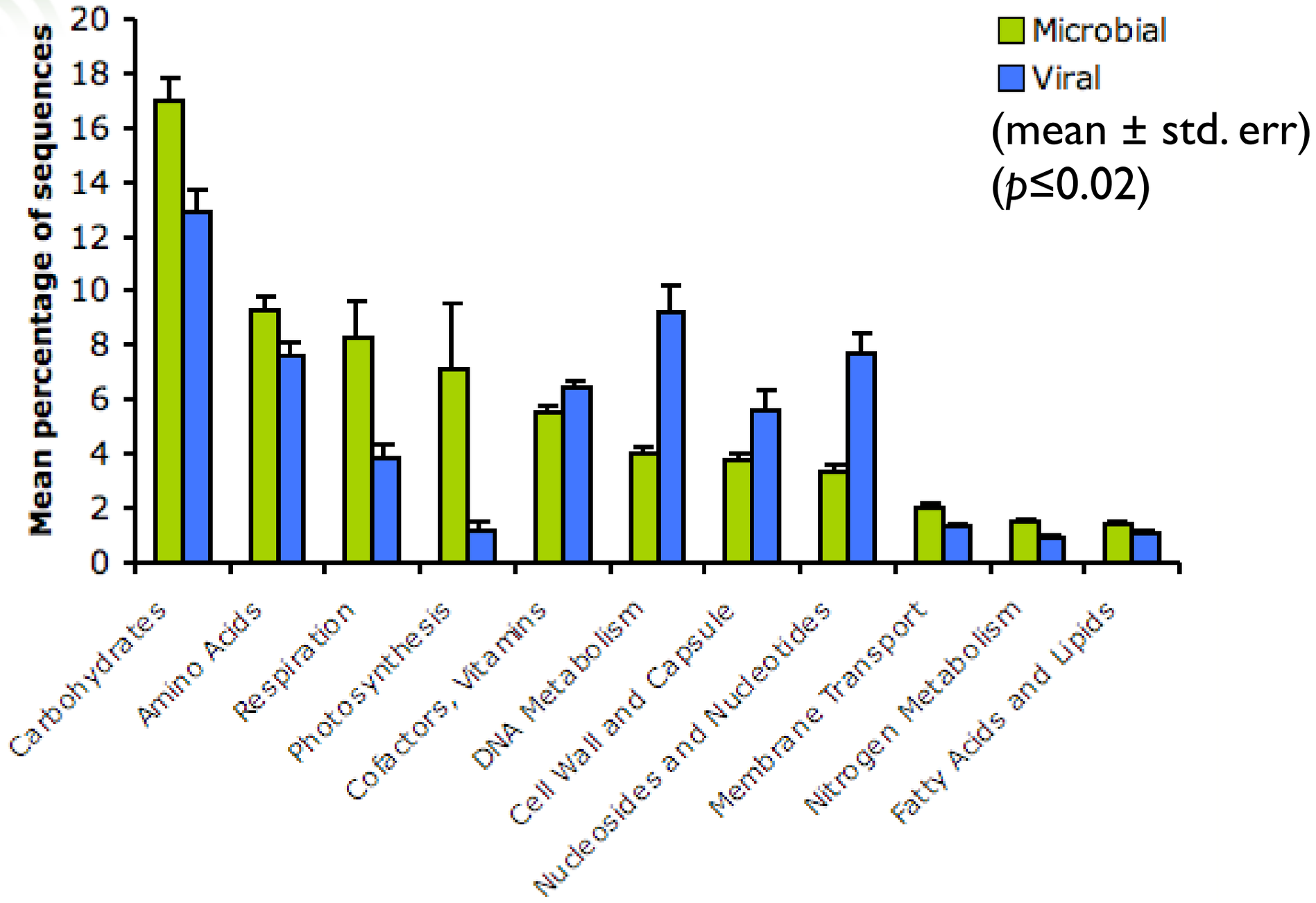
Class name	Human	Mouse	p-value
Clostridia	66.9 ± 5.8	49.1 ± 3.2	0.019
Bacilli	4.27 ± 0.97	12.1 ± 1.9	0.003
Actinobacteria (class)	0.447 ± 0.18	0.979 ± 0.17	0.041
Verrucomicrobiae	0.162 ± 0.14	0	0.006
Alphaproteobacteria	0.115 ± 0.12	0	0.026
Epsilonproteobacteria	0	0.261 ± 0.17	0.002
TM7 genera incertae sedis	0	0.220 ± 0.10	0.032

Table 1 Differentially abundant classes of organisms from human and mouse gut microbiota (p-values ≤ 0.05). Human and mouse columns display mean relative abundance (%) ± standard error. Cells containing '0' indicate that no observations of the taxa were found. Clostridia and Bacilli, two of the three most abundant classes observed were differentially abundant.

Metabolic profiles

- Dinsdale *et al.*, **Nature**, 2008.
- Collected 87 microbial and viral metagenomes.
- 15 million shotgun sequences!
- subterranean, coral reefs, hypersaline, freshwater, animal guts, mosquito viruses.

Metabolic profiles



statistical methods for metagenomics

Detect Differentially Abundant Taxa

Input frequency matrix and choose parameters:

Load a formatted matrix from disk: Choose File no file selected
a sample matrix is provided here.

Total number of
subjects from the 1st population:

Threshold by p or q values: p values

Significance level to threshold by: 0.05

Number of permutations to calculate p values: 1000

Prefix for output: tmp

Email address:

Submit Reset fields Help

Timeline

- **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
- **Beyond**
  - Submit paper.
  - New data.
  - Correlations between taxa.

Acknowledgments

- Mihai Pop, CBCB
- Andrea Ottesen, PLSC
- Frank Siewerdt, ANSC
- Paul Smith, STAT
- Radu Balan, CSCAMM
- Aleksey Zimin, IPST

Questions?