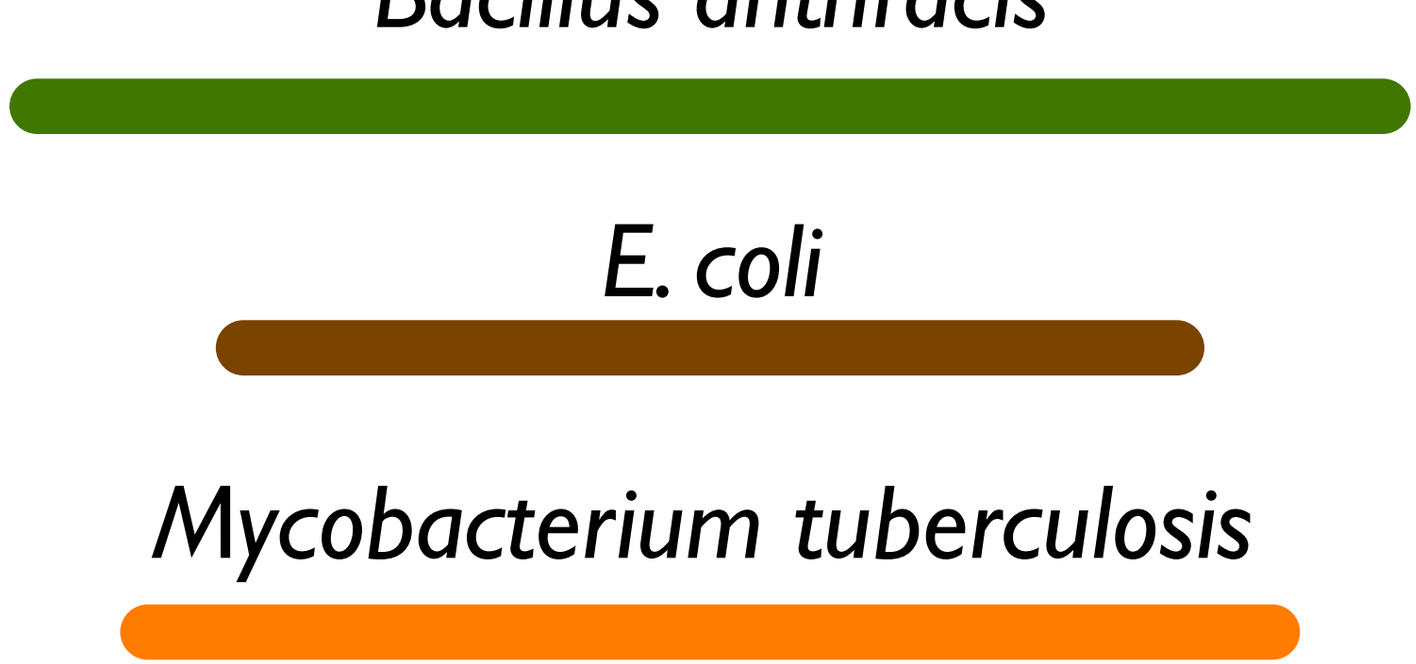
Methods for comparing microbial communities

james robert white
whitej@umd.edu

Advisor: Mihai Pop
Center for Bioinformatics and Computational Biology
University of Maryland - College Park

Background

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

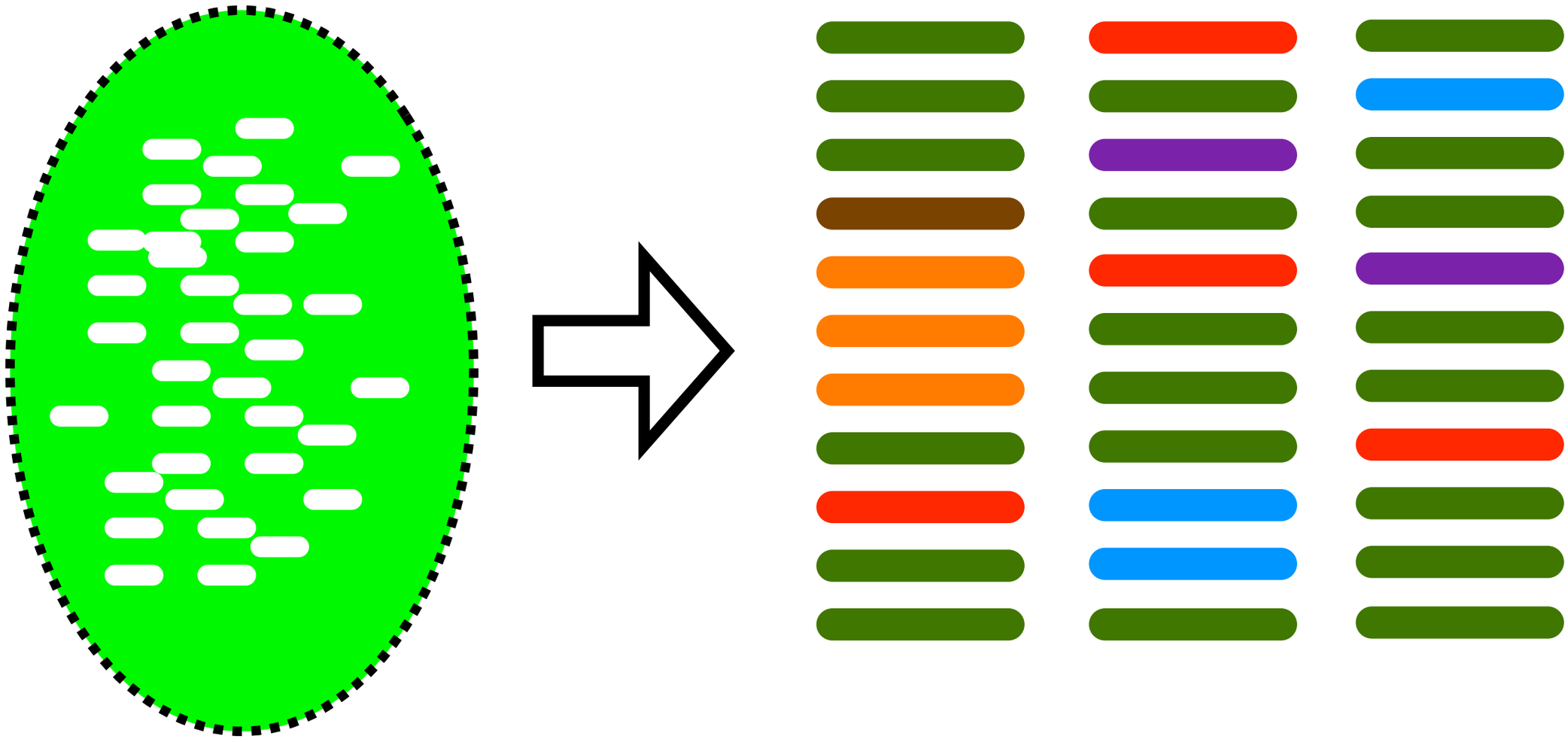


Background

- We have technology to take a sample from an environment, and read the 16S genes from every microbe we capture.
- This is how we can tell what's living in your gut, skin, eyes, mouth, ocean, desert, soil, sewage, ...

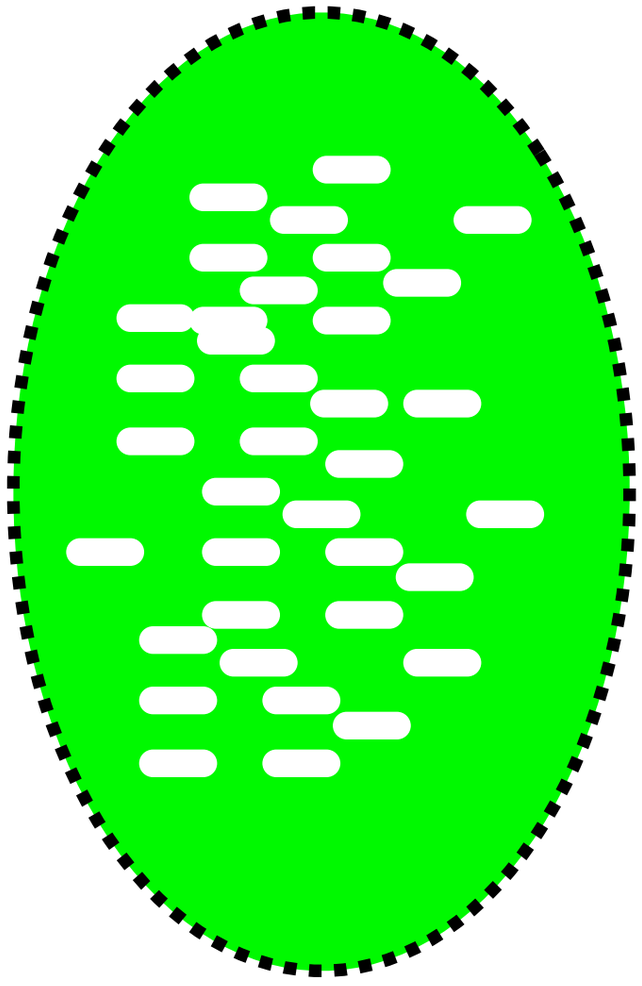
Background

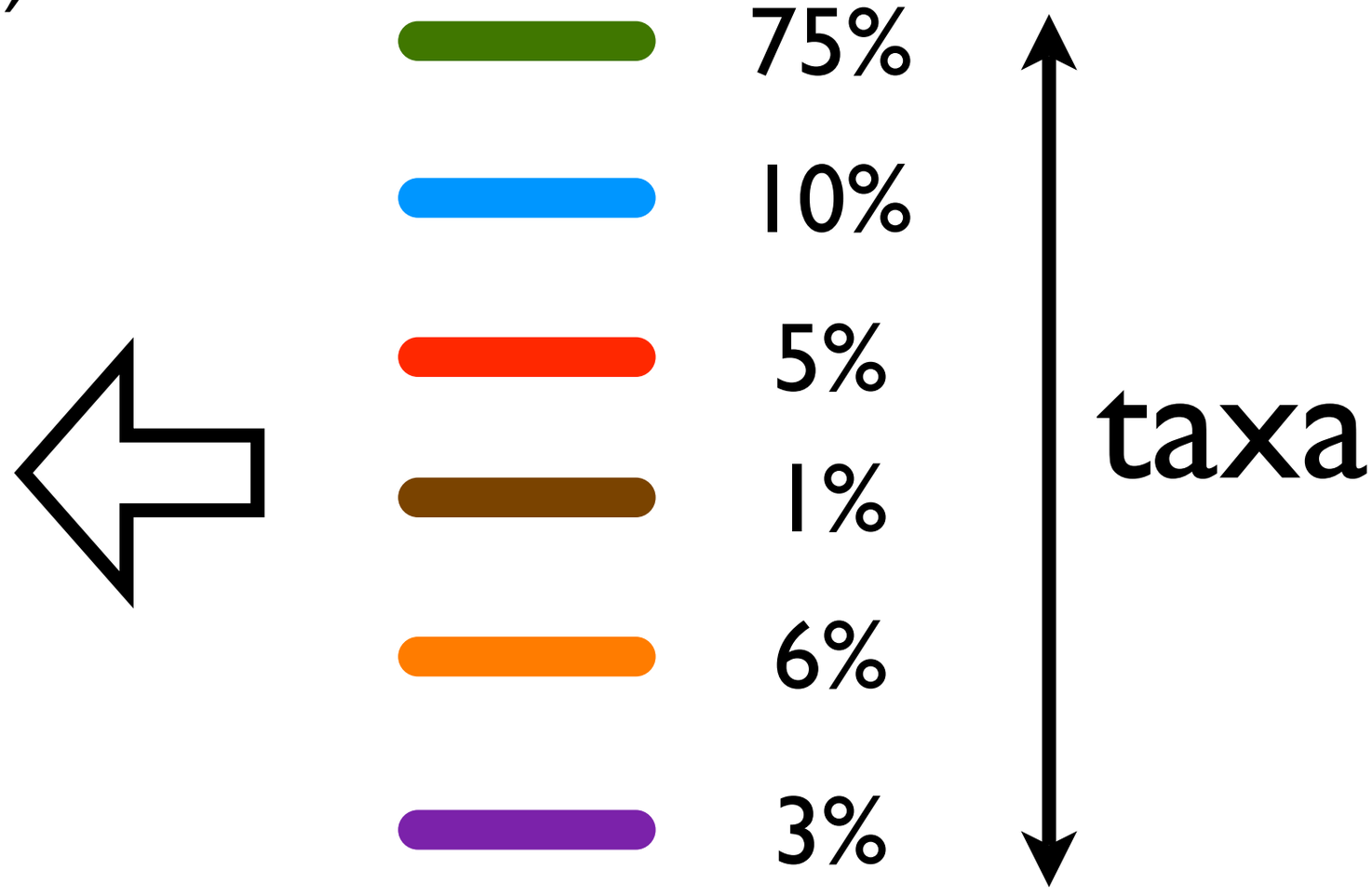
Environment
(radioactive waste)



Background

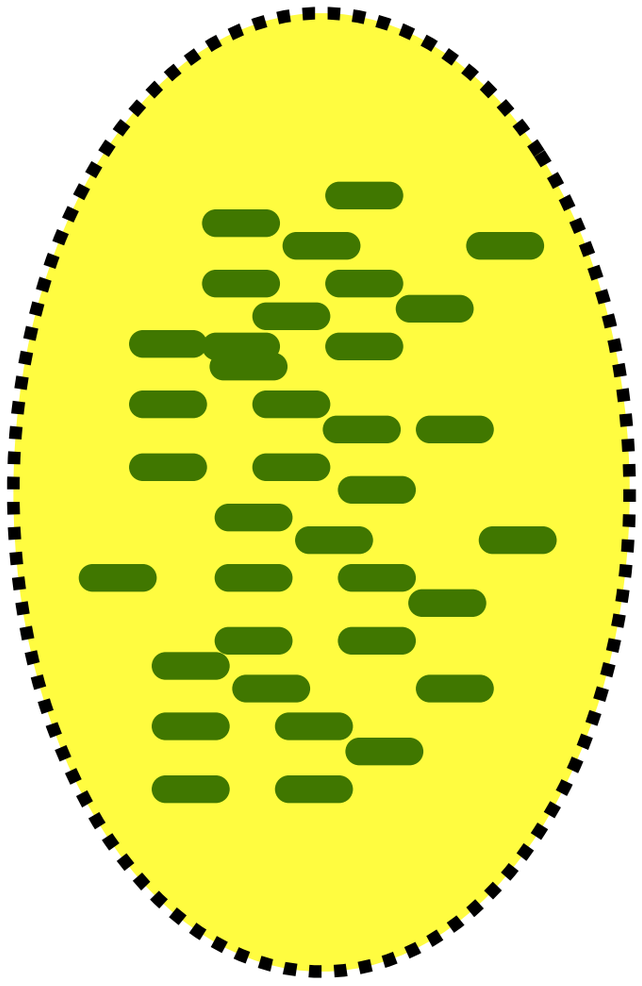
Environment
(radioactive waste)





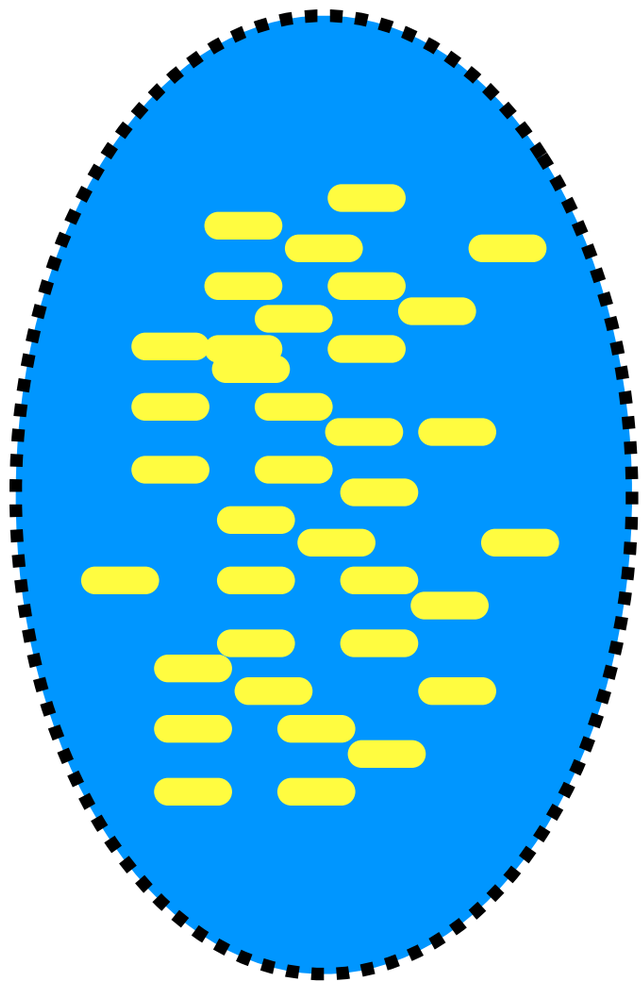
The problem

(Healthy ears)



How do two environments differ?

(Sick ears)



Differentially abundant organisms

- Strategy:
  - input species abundance matrix

	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

Differentially abundant organisms

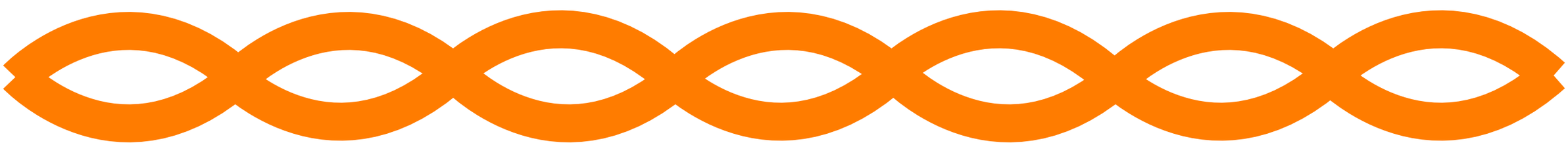
- Strategy:
  - change to proportions and normalize the data.
  - perform 2 sample t-test for each taxa.
  - for a particular taxa, what's the null hypothesis? alternative?

t-test

- Two populations: Healthy, Sick.
- For each taxa j:
- H_0: $\mu_{healthy} = \mu_{sick}$
- H_A: $\mu_{healthy} != \mu_{sick}$
- Two-tailed test

Differentially expressed genes

- Genes are portions of DNA that are **literally** decoded (*expressed*) into larger molecules which keep every function in our bodies going.
- A genetic disorder usually alters gene expression in some way for the worse.

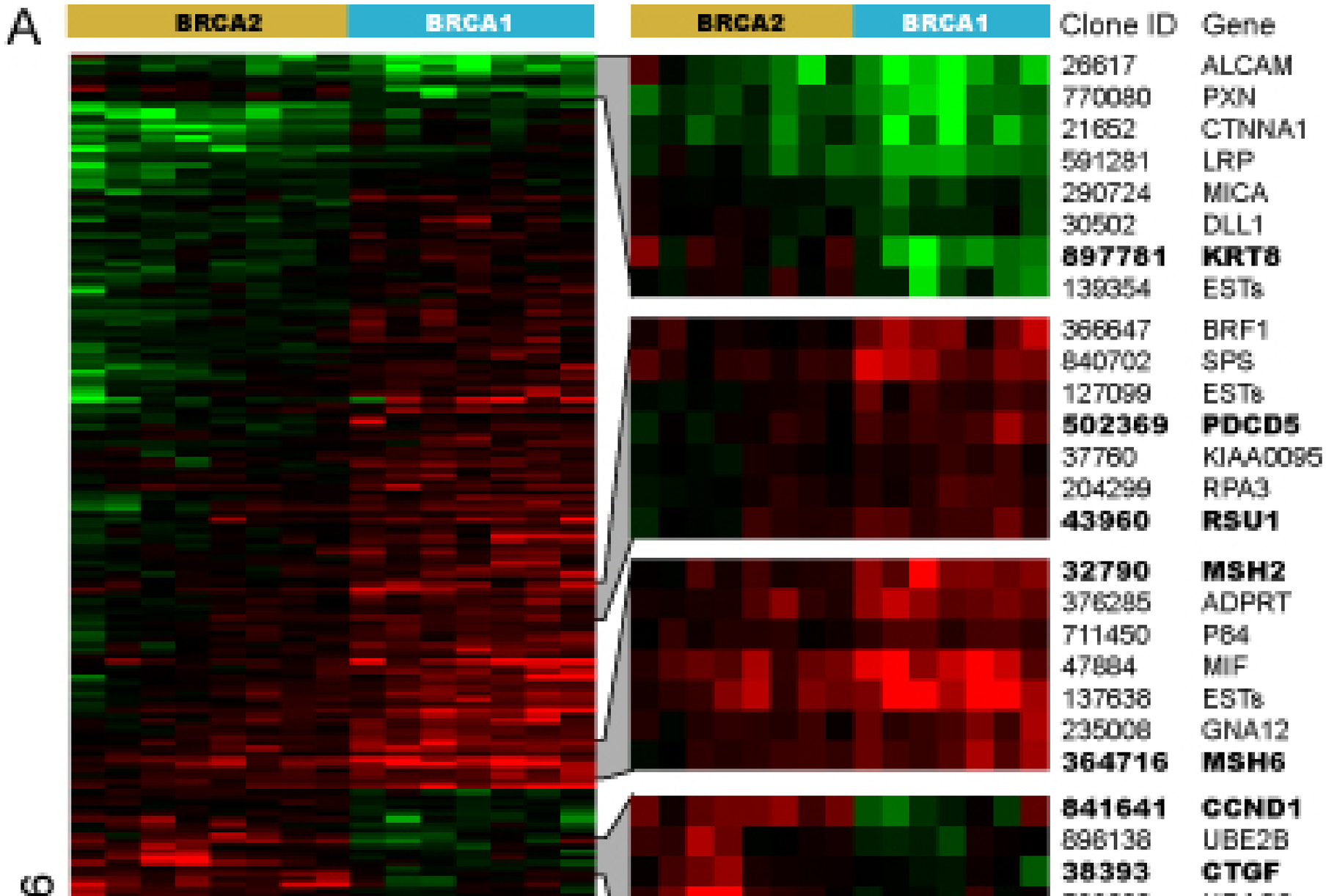


Differentially expressed genes

- When a sick population decodes a gene more or less often than a healthy population, this is differential expression.
- Someday your doctor will be able to test expression levels of thousands of your genes.
- thousands of genes = thousands of hypothesis tests.

Differentially expressed genes

(Hedenfalk, *PNAS*, **100**, 2001)



Multiplicity!

- Testing 1000 genes in humans
- I have alpha = 5% threshold for t-test
- expect 50 false positives!
- need to reduce false positives when dealing with multiple hypothesis tests

Multiplicity controls

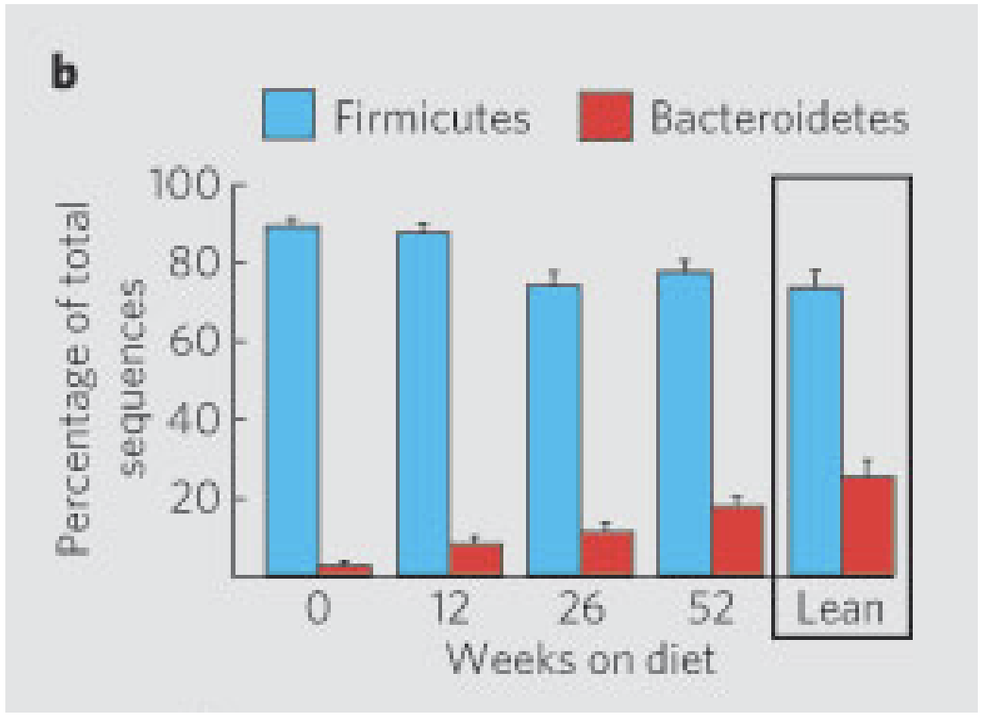
- *false discovery rate* - 1995.
  - expected proportion of **rejected null hypotheses** which are false positives.
- different than *false positive rate*:
  - expected proportion of **all significance tests** which are false positives.

Multiplicity controls

- p-value => individual false positive rate for a single test.
- q-value => individual false discovery rate for a single test.
- thresholding by q-values instead of p-values greatly reduces the number of false positives.

Cluster taxa

- cluster samples based on differentially abundant taxa.
- single, average, and complete linkage algorithms.
- cluster taxa into higher levels and repeat hypothesis testing.



(R. Ley, *Nature*, **444**, 2006)

Schedule

- Fall 07’ - functioning implementation of q-value algorithm with clustering algorithms in C++.
- Implement algorithms in R or Matlab, transition to C++.
- Spring 08’- validation and applications, address independent visualization and statistical concerns.

Resources

- CBCB servers: 2x and 4x Opterons - 8 and 32GB of RAM.
- Dell 2x 3.0 GHz, 2GB (Linux)
- Mac OS X 2.16 GHz, 2GB

Validation

- Validate statistical calculations using SAS, R.
- Classical dataset (Hedenfalk, 2001) - differentially expressed genes related to two independent forms of breast cancer.
- Final application to the human gut samples of obese and lean individuals.


Questions?