20.109
Laboratory Fundamentals in Biological Engineering

Module 1
Nucleic Acid Engineering
Lecture 8

Today

Donut Day

Finish Phylogenetics

Microbiome – other considerations

Basic Epidemiology

Assessing confidence

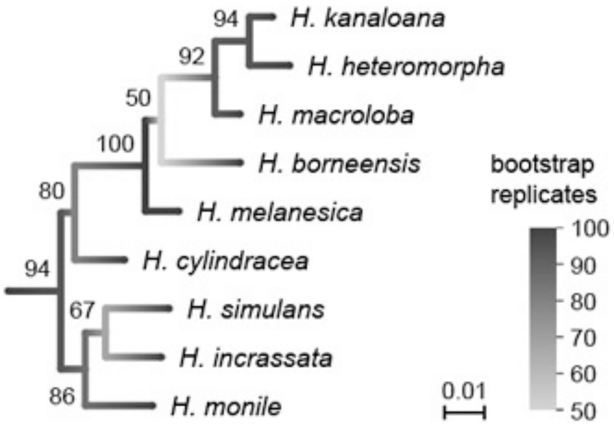
- Trees obtained by phylogenetics are subject to error like all other scientific hypotheses
- A tree will be generated regardless of whether there is a phylogenetic signal
- Need to quantify how strongly data supports each of the relationships in the tree
- What is the extent to which characters within a matrix contradict each other?

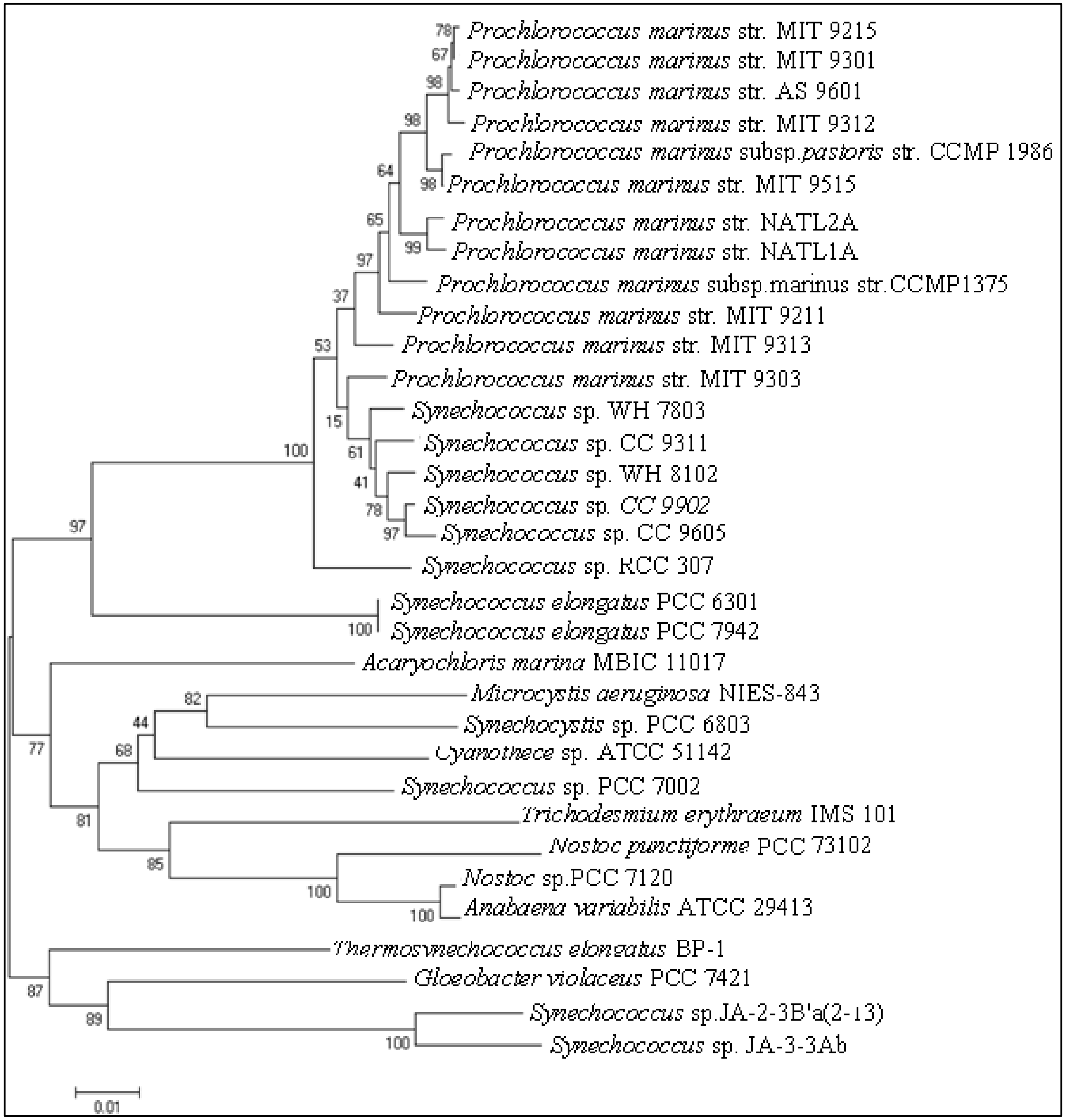
Bootstrapping

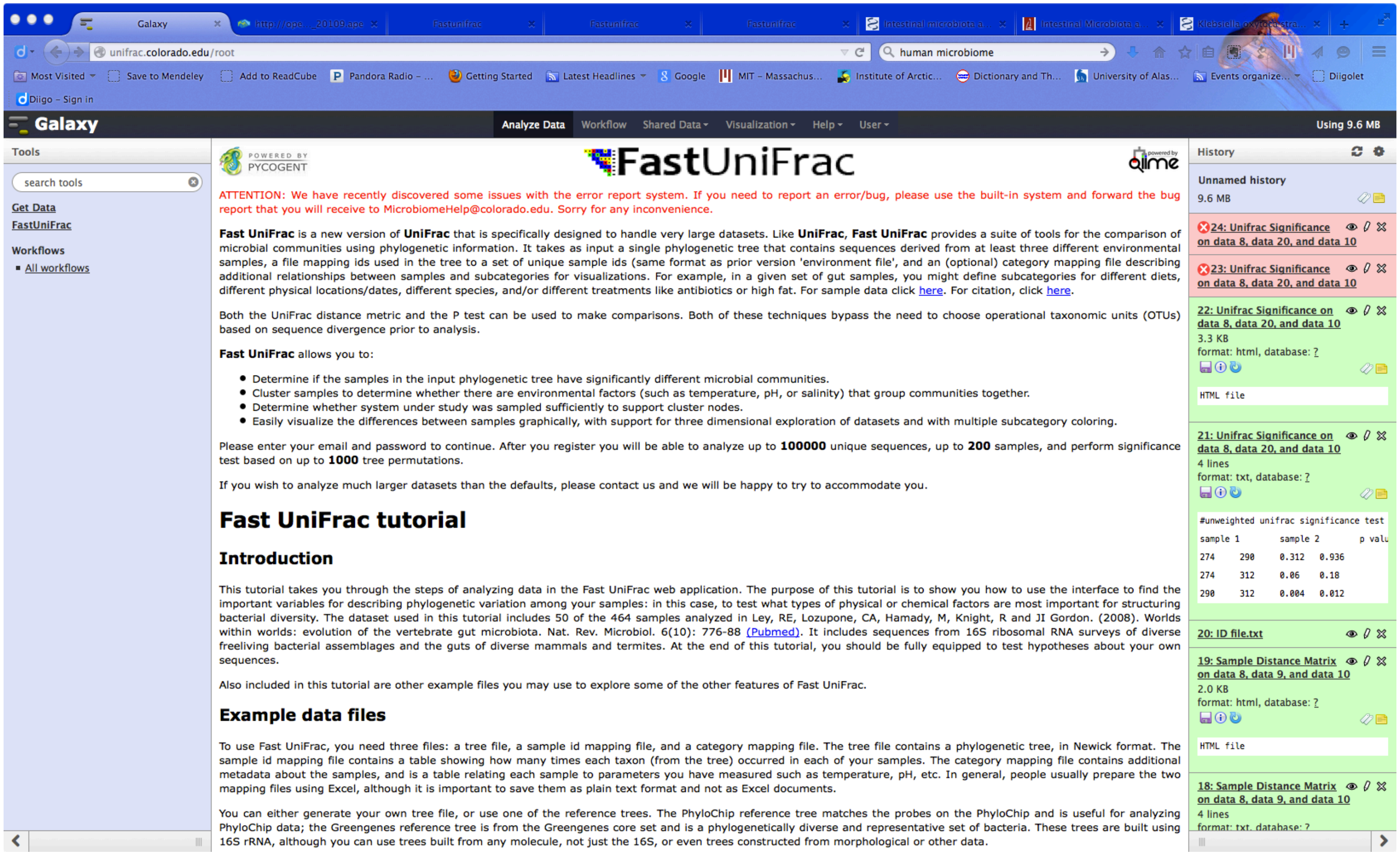
- Typically tackled with a statistical test called bootstrapping
- Assesses chances of recovering a particular clade again if we randomly re-sample our data
- Data matrix is sampled with replacement to produce pseudo-replicate datasets
- Measures which parts of the tree are weakly supported with a low bootstrap %

Bootstrap cut-offs

- Exact interpretation of bootstrap % is elusive
- Higher is better but what is a reasonable cut-off? 70%?
- Warning: bootstrapping predicts whether the same result would occur if more data were collected not whether the result is correct




78
Prochlorococcus marinus str. MIT 9215
67
Prochlorococcus marinus str. MIT 9301
98
Prochlorococcus marinus str. AS 9601
98
Prochlorococcus marinus str. MIT 9312
Prochlorococcus marinus subsp.pastoris str. CCMP 1986
64
98
Prochlorococcus marinus str. MIT 9515
65
Prochlorococcus marinus str. NATL2A
99
Prochlorococcus marinus str. NATL1A
97
Prochlorococcus marinus subsp.marinus str.CCMP1375
37
Prochlorococcus marinus str. MIT 9211
53
Prochlorococcus marinus str. MIT 9313
Prochlorococcus marinus str. MIT 9303
15
Synechococcus sp. WH 7803
100
61
Synechococcus sp. CC 9311
41
Synechococcus sp. WH 8102
78
Synechococcus sp. CC 9902
97
Synechococcus sp. CC 9605
97
Synechococcus sp. RCC 307
Synechococcus elongatus PCC 6301
100
Synechococcus elongatus PCC 7942
Acaryochloris marina MBIC 11017
82
Microcystis aeruginosa NIES-843
44
Synechocystis sp. PCC 6803
77
68
Cyanothece sp. ATCC 51142
Synechococcus sp. PCC 7002
81
Trichodesmium erythraeum IMS 101
85
Nostoc punctiforme PCC 73102
100
Nostoc sp.PCC 7120
100
Anabaena variabilis ATCC 29413
Thermosynechococcus elongatus BP-1
87
Gloeobacter violaceus PCC 7421
89
Synechococcus sp.JA-2-3B'a(2-13)
100
Synechococcus sp. JA-3-3Ab
0.01



ATTENTION: We have recently discovered some issues with the error report system. If you need to report an error/bug, please use the built-in system and forward the bug report that you will receive to MicrobiomeHelp@colorado.edu. Sorry for any inconvenience.

Fast UniFrac is a new version of **UniFrac** that is specifically designed to handle very large datasets. Like **UniFrac, Fast UniFrac** provides a suite of tools for the comparison of microbial communities using phylogenetic information. It takes as input a single phylogenetic tree that contains sequences derived from at least three different environmental samples, a file mapping ids used in the tree to a set of unique sample ids (same format as prior version 'environment file', and an (optional) category mapping file describing additional relationships between samples and subcategories for visualizations. For example, in a given set of gut samples, you might define subcategories for different diets, different physical locations/dates, different species, and/or different treatments like antibiotics or high fat. For sample data click here. For citation, click here.

Both the UniFrac distance metric and the P test can be used to make comparisons. Both of these techniques bypass the need to choose operational taxonomic units (OTUs) based on sequence divergence prior to analysis.

Fast UniFrac allows you to:

- Determine if the samples in the input phylogenetic tree have significantly different microbial communities.
- Cluster samples to determine whether there are environmental factors (such as temperature, pH, or salinity) that group communities together.
- Determine whether system under study was sampled sufficiently to support cluster nodes.
- Easily visualize the differences between samples graphically, with support for three dimensional exploration of datasets and with multiple subcategory coloring.

Please enter your email and password to continue. After you register you will be able to analyze up to **100000** unique sequences, up to **200** samples, and perform significance test based on up to **1000** tree permutations.

If you wish to analyze much larger datasets than the defaults, please contact us and we will be happy to try to accommodate you.

Fast UniFrac tutorial

Introduction

This tutorial takes you through the steps of analyzing data in the Fast UniFrac web application. The purpose of this tutorial is to show you how to use the interface to find the important variables for describing phylogenetic variation among your samples: in this case, to test what types of physical or chemical factors are most important for structuring bacterial diversity. The dataset used in this tutorial includes 50 of the 464 samples analyzed in Ley, RE, Lozupone, CA, Hamady, M, Knight, R and JI Gordon. (2008). Worlds within worlds: evolution of the vertebrate gut microbiota. Nat. Rev. Microbiol. 6(10): 776-88 (Pubmed). It includes sequences from 16S ribosomal RNA surveys of diverse freeliving bacterial assemblages and the guts of diverse mammals and termites. At the end of this tutorial, you should be fully equipped to test hypotheses about your own sequences.

Also included in this tutorial are other example files you may use to explore some of the other features of Fast UniFrac.

Example data files

To use Fast UniFrac, you need three files: a tree file, a sample id mapping file, and a category mapping file. The tree file contains a phylogenetic tree, in Newick format. The sample id mapping file contains a table showing how many times each taxon (from the tree) occurred in each of your samples. The category mapping file contains additional metadata about the samples, and is a table relating each sample to parameters you have measured such as temperature, pH, etc. In general, people usually prepare the two mapping files using Excel, although it is important to save them as plain text format and not as Excel documents.

You can either generate your own tree file, or use one of the reference trees. The PhyloChip reference tree matches the probes on the PhyloChip and is useful for analyzing PhyloChip data; the Greengenes reference tree is from the Greengenes core set and is a phylogenetically diverse and representative set of bacteria. These trees are built using 16S rRNA, although you can use trees built from any molecule, not just the 16S, or even trees constructed from morphological or other data.

Unnamed history
9.6 MB

24: Unifrac Significance on data 8, data 20, and data 10

23: Unifrac Significance on data 8, data 20, and data 10

22: Unifrac Significance on data 8, data 20, and data 10
3.3 KB
format: html, database: ?

```
HTML file
```

21: Unifrac Significance on data 8, data 20, and data 10
4 lines
format: txt, database: ?

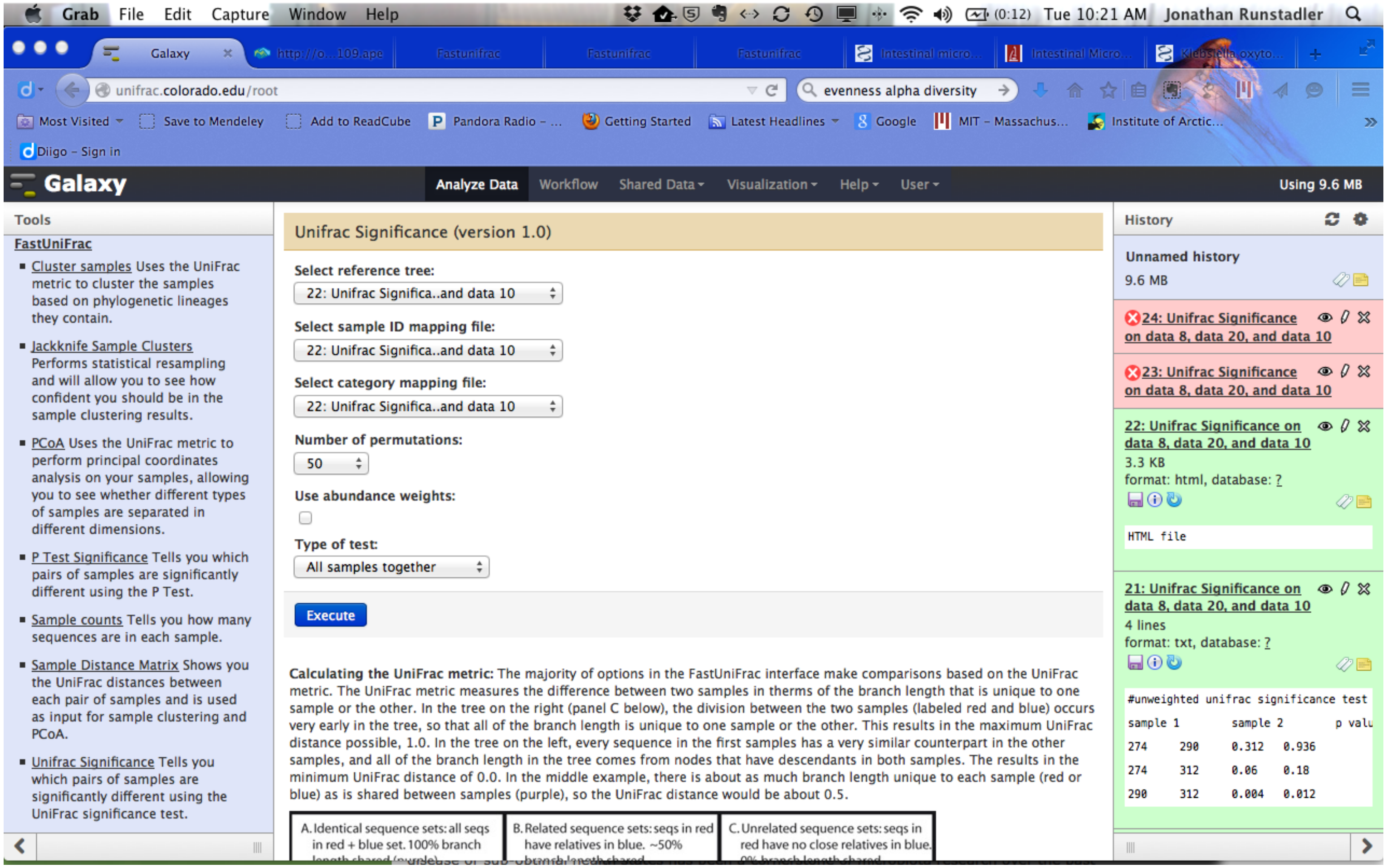
```
#unweighted unifrac significance test
sample 1        sample 2        p valu
274     290     0.312   0.936
274     312     0.06    0.18
290     312     0.004   0.012
```

20: ID file.txt

19: Sample Distance Matrix on data 8, data 9, and data 10
2.0 KB
format: html, database: ?

```
HTML file
```

18: Sample Distance Matrix on data 8, data 9, and data 10
4 lines
format: txt, database: ?



FastUniFrac

- Cluster samples Uses the UniFrac metric to cluster the samples based on phylogenetic lineages they contain.
- Jackknife Sample Clusters Performs statistical resampling and will allow you to see how confident you should be in the sample clustering results.
- PCoA Uses the UniFrac metric to perform principal coordinates analysis on your samples, allowing you to see whether different types of samples are separated in different dimensions.
- P Test Significance Tells you which pairs of samples are significantly different using the P Test.
- Sample counts Tells you how many sequences are in each sample.
- Sample Distance Matrix Shows you the UniFrac distances between each pair of samples and is used as input for sample clustering and PCoA.
- Unifrac Significance Tells you which pairs of samples are significantly different using the UniFrac significance test.

Unifrac Significance (version 1.0)

Select reference tree: 22: Unifrac Significa..and data 10

Select sample ID mapping file: 22: Unifrac Significa..and data 10

Select category mapping file: 22: Unifrac Significa..and data 10

Number of permutations: 50

Use abundance weights: □

Type of test: All samples together

Calculating the UniFrac metric: The majority of options in the FastUniFrac interface make comparisons based on the UniFrac metric. The UniFrac metric measures the difference between two samples in therms of the branch length that is unique to one sample or the other. In the tree on the right (panel C below), the division between the two samples (labeled red and blue) occurs very early in the tree, so that all of the branch length is unique to one sample or the other. This results in the maximum UniFrac distance possible, 1.0. In the tree on the left, every sequence in the first samples has a very similar counterpart in the other samples, and all of the branch length in the tree comes from nodes that have descendants in both samples. The results in the minimum UniFrac distance of 0.0. In the middle example, there is about as much branch length unique to each sample (red or blue) as is shared between samples (purple), so the UniFrac distance would be about 0.5.

A. Identical sequence sets: all seqs in red + blue set. 100% branch length shared (purple)

B. Related sequence sets: seqs in red have relatives in blue. ~50% branch length shared

C. Unrelated sequence sets: seqs in red have no close relatives in blue. 0% branch length shared

Unnamed history
9.6 MB

24: Unifrac Significance on data 8, data 20, and data 10

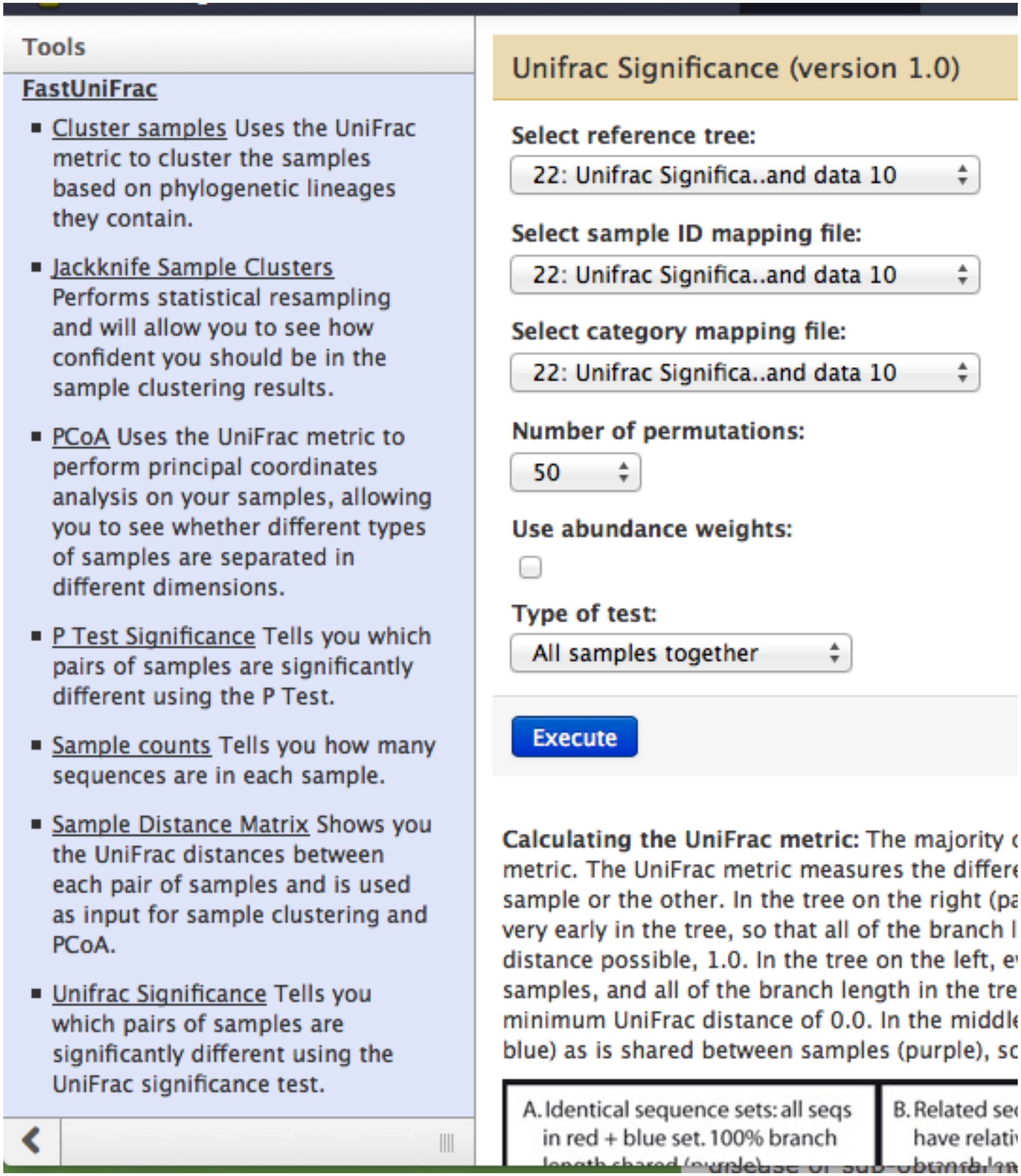
23: Unifrac Significance on data 8, data 20, and data 10

22: Unifrac Significance on data 8, data 20, and data 10
3.3 KB
format: html, database: ?

```
HTML file
```

21: Unifrac Significance on data 8, data 20, and data 10
4 lines
format: txt, database: ?

```
#unweighted unifrac significance test
sample 1        sample 2        p valu
274     290     0.312   0.936
274     312     0.06    0.18
290     312     0.004   0.012
```



Tools

FastUniFrac

- Cluster samples Uses the UniFrac metric to cluster the samples based on phylogenetic lineages they contain.
- Jackknife Sample Clusters Performs statistical resampling and will allow you to see how confident you should be in the sample clustering results.
- PCoA Uses the UniFrac metric to perform principal coordinates analysis on your samples, allowing you to see whether different types of samples are separated in different dimensions.
- P Test Significance Tells you which pairs of samples are significantly different using the P Test.
- Sample counts Tells you how many sequences are in each sample.
- Sample Distance Matrix Shows you the UniFrac distances between each pair of samples and is used as input for sample clustering and PCoA.
- Unifrac Significance Tells you which pairs of samples are significantly different using the UniFrac significance test.

Unifrac Significance (version 1.0)

Select reference tree: 22: Unifrac Significa..and data 10

Select sample ID mapping file: 22: Unifrac Significa..and data 10

Select category mapping file: 22: Unifrac Significa..and data 10

Number of permutations: 50

Use abundance weights: ☐

Type of test: All samples together

Calculating the UniFrac metric: The majority c
metric. The UniFrac metric measures the differe
sample or the other. In the tree on the right (pa
very early in the tree, so that all of the branch l
distance possible, 1.0. In the tree on the left, e
samples, and all of the branch length in the tre
minimum UniFrac distance of 0.0. In the middle
blue) as is shared between samples (purple), sc

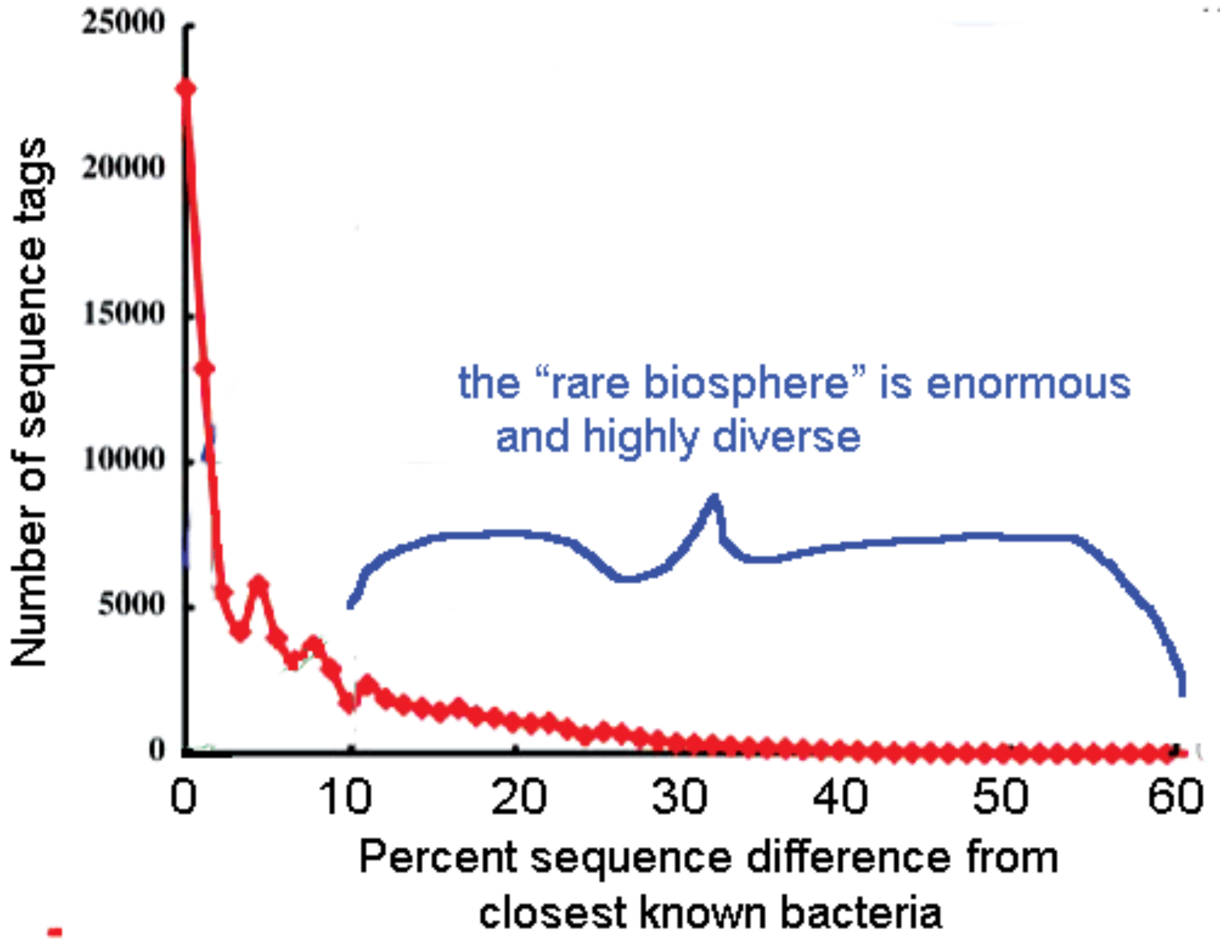
A. Identical sequence sets: all seqs in red + blue set. 100% branch length shared (purple)

B. Related se have relati

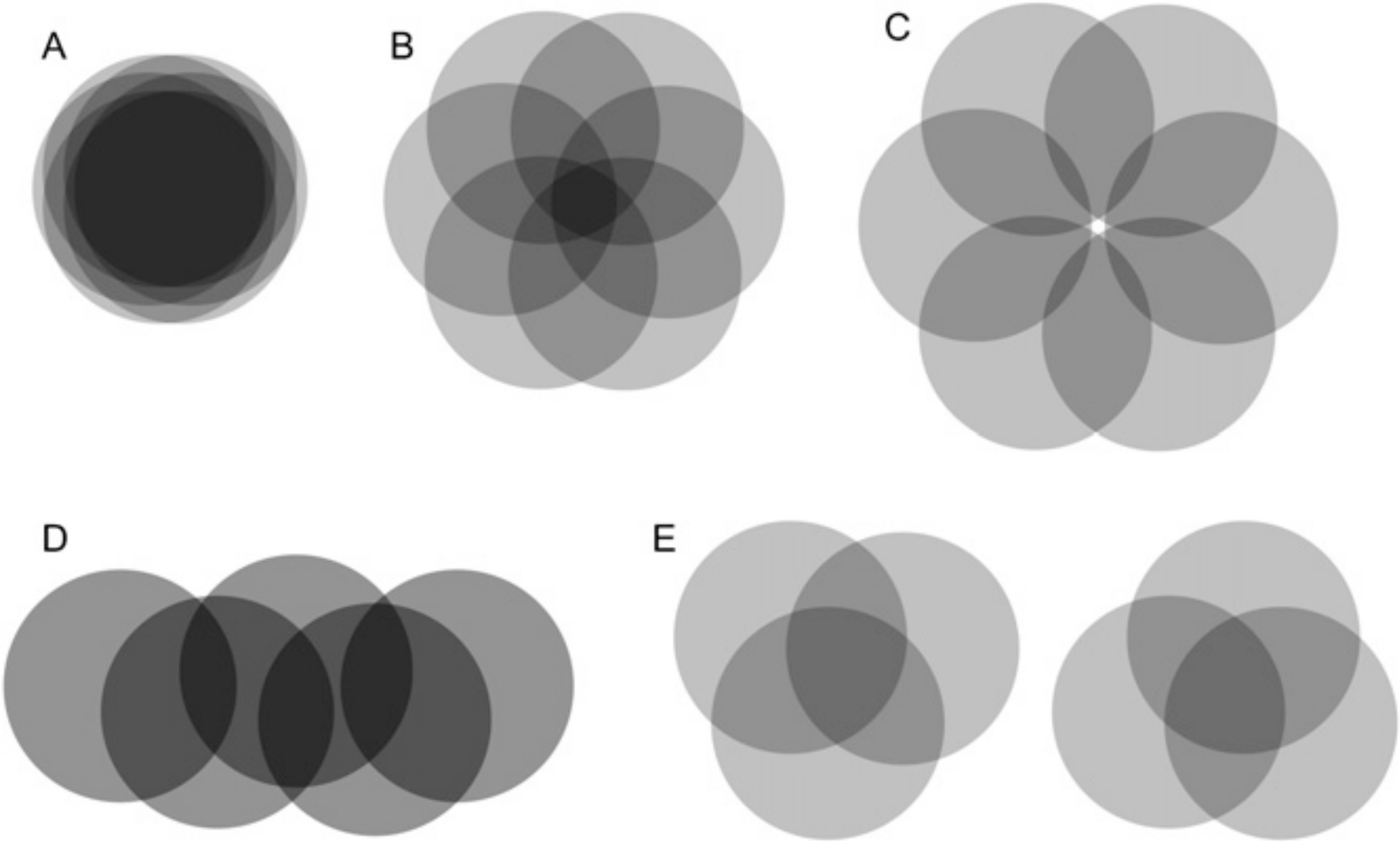
Back to the core questions

- Structure of the microbiome?
- Function of the microbiome?
- How can it be changed?

Rare biosphere



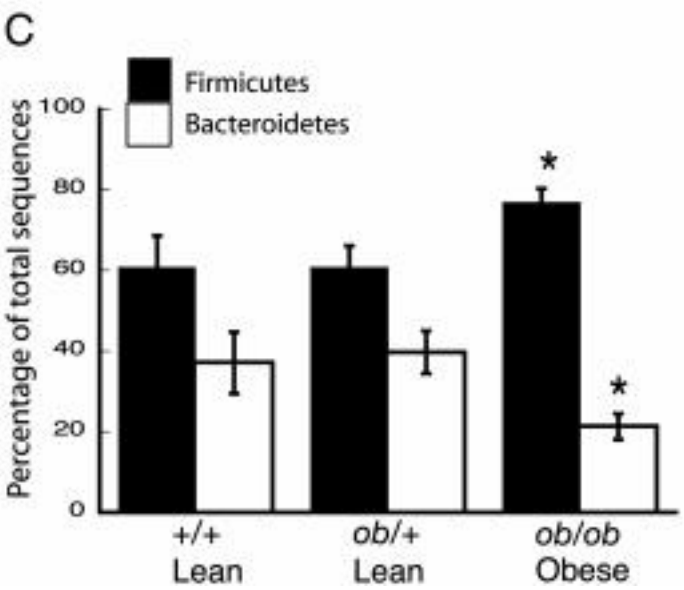
Models of a core microbiome



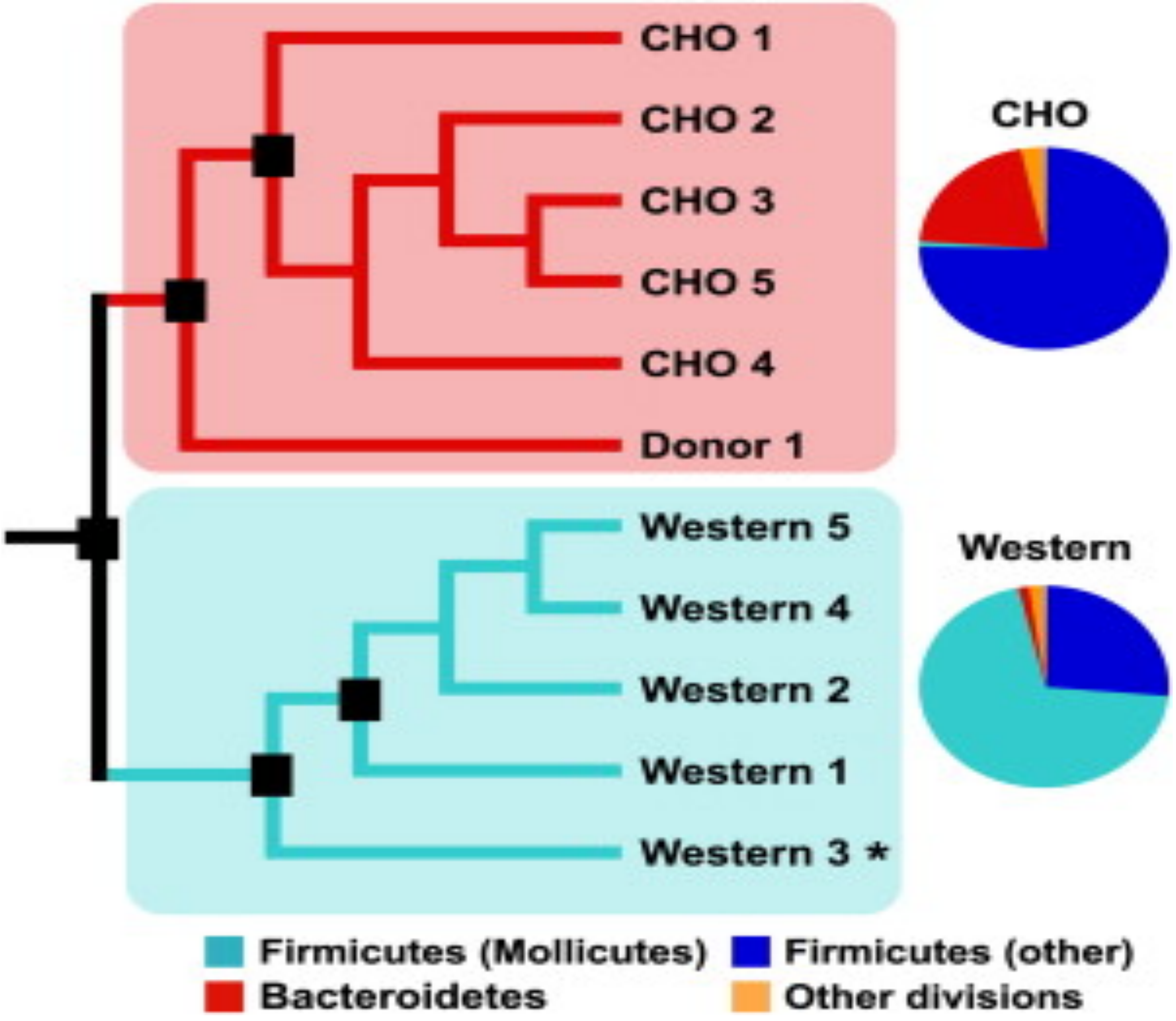
Hamady and Knight, 2009

Does diet affect microbial composition?

- Genetically Obese mice harbor a significantly different community than lean conventional mice

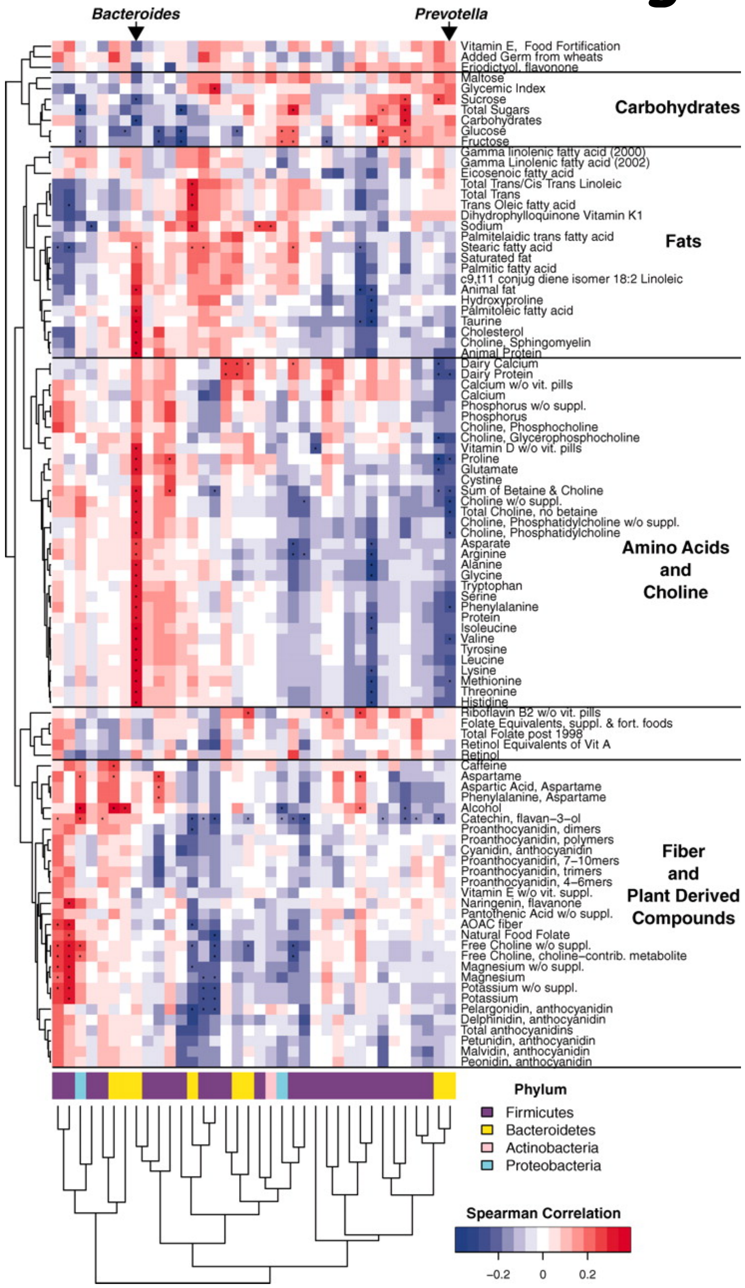


Diet affects microbial composition



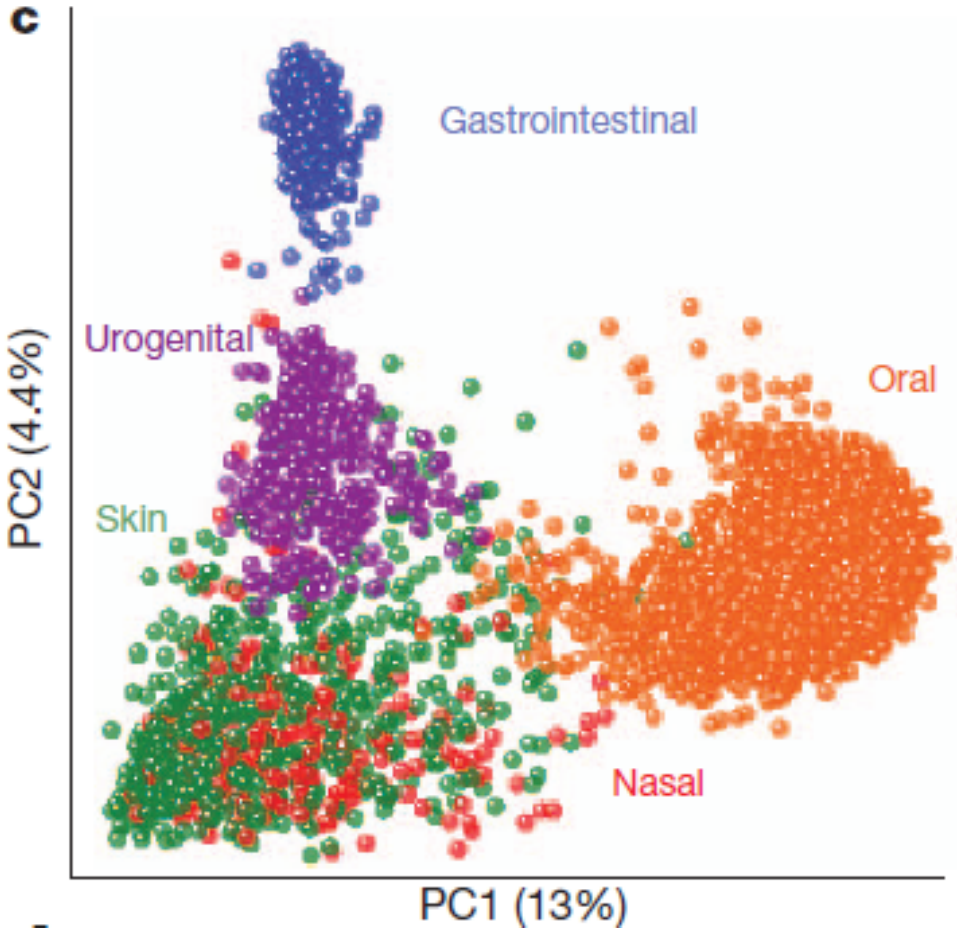
Turnbaugh et al, Cell Host & Microbe 2008 213 - 223

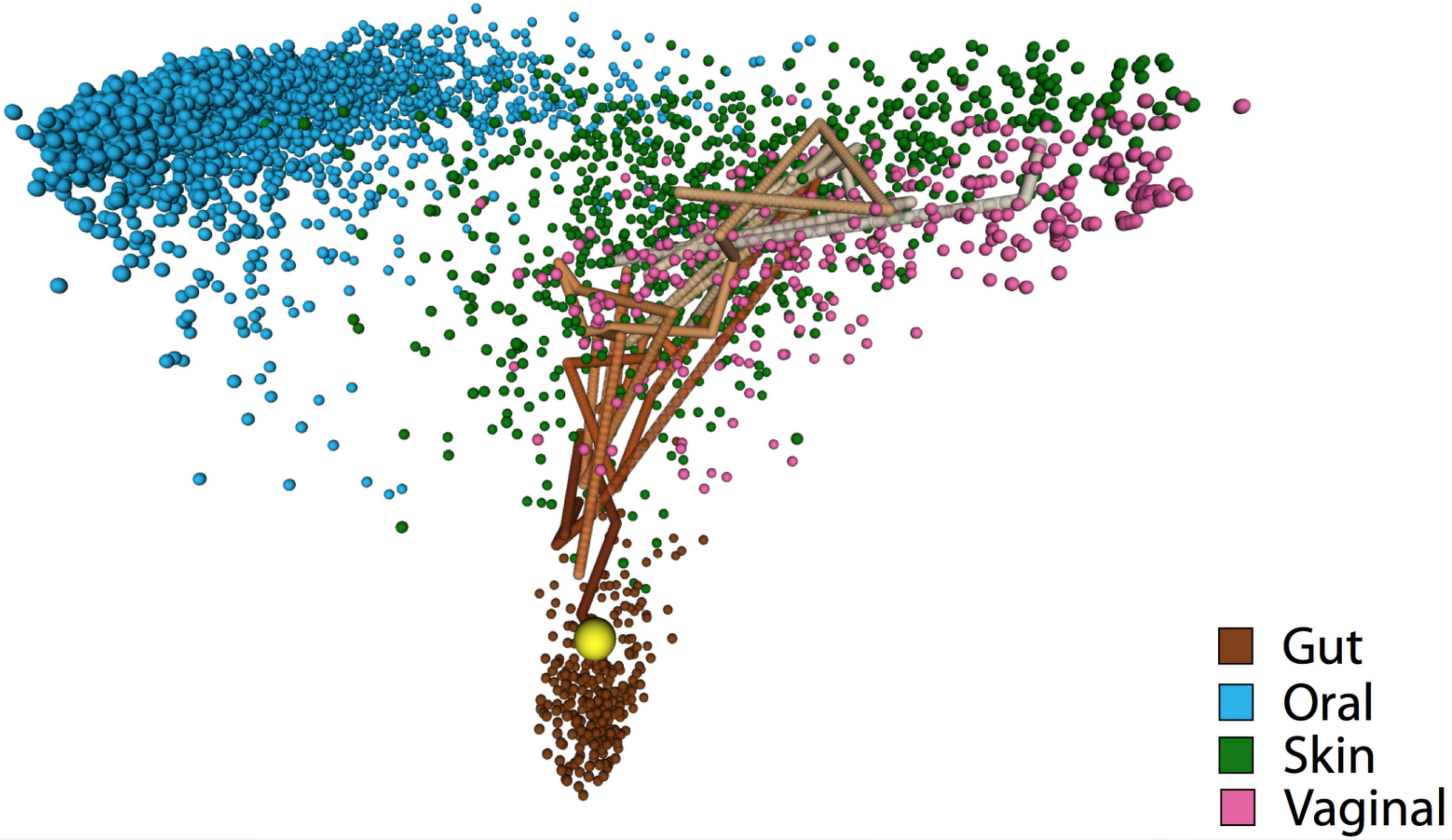
Correlation of diet and gut microbial taxa



G D Wu et al. Science 2011;334:105-108

The Human Microbiome

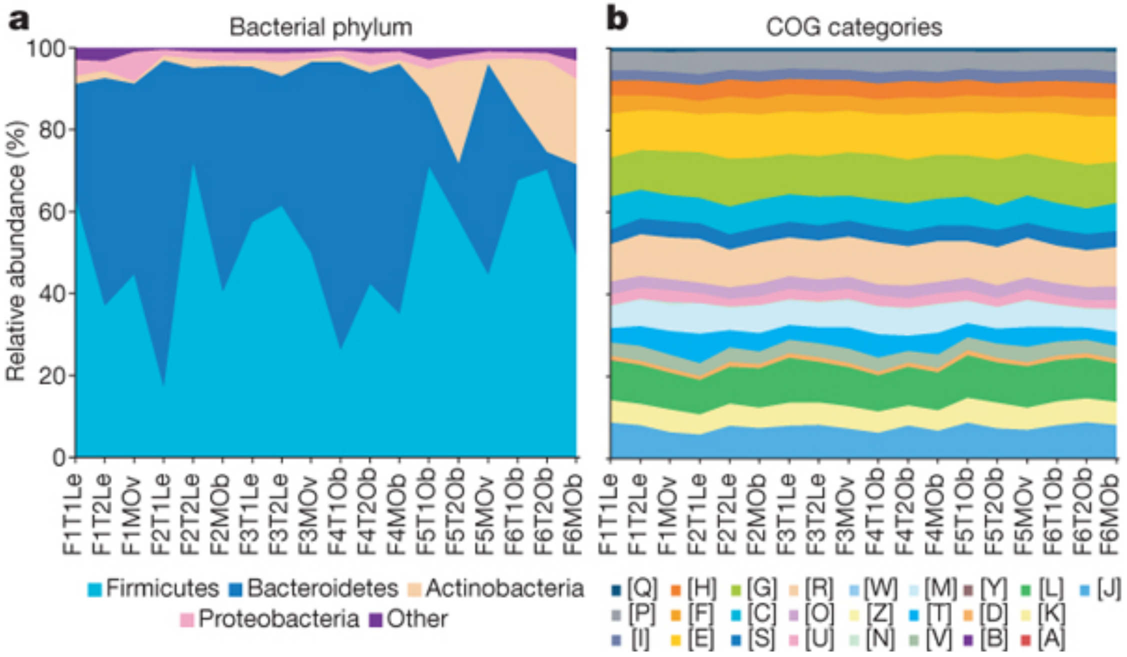



🟫 Gut
🟦 Oral
🟩 Skin
🟪 Vaginal

If taxonomy is not conserved, what does that mean for function?

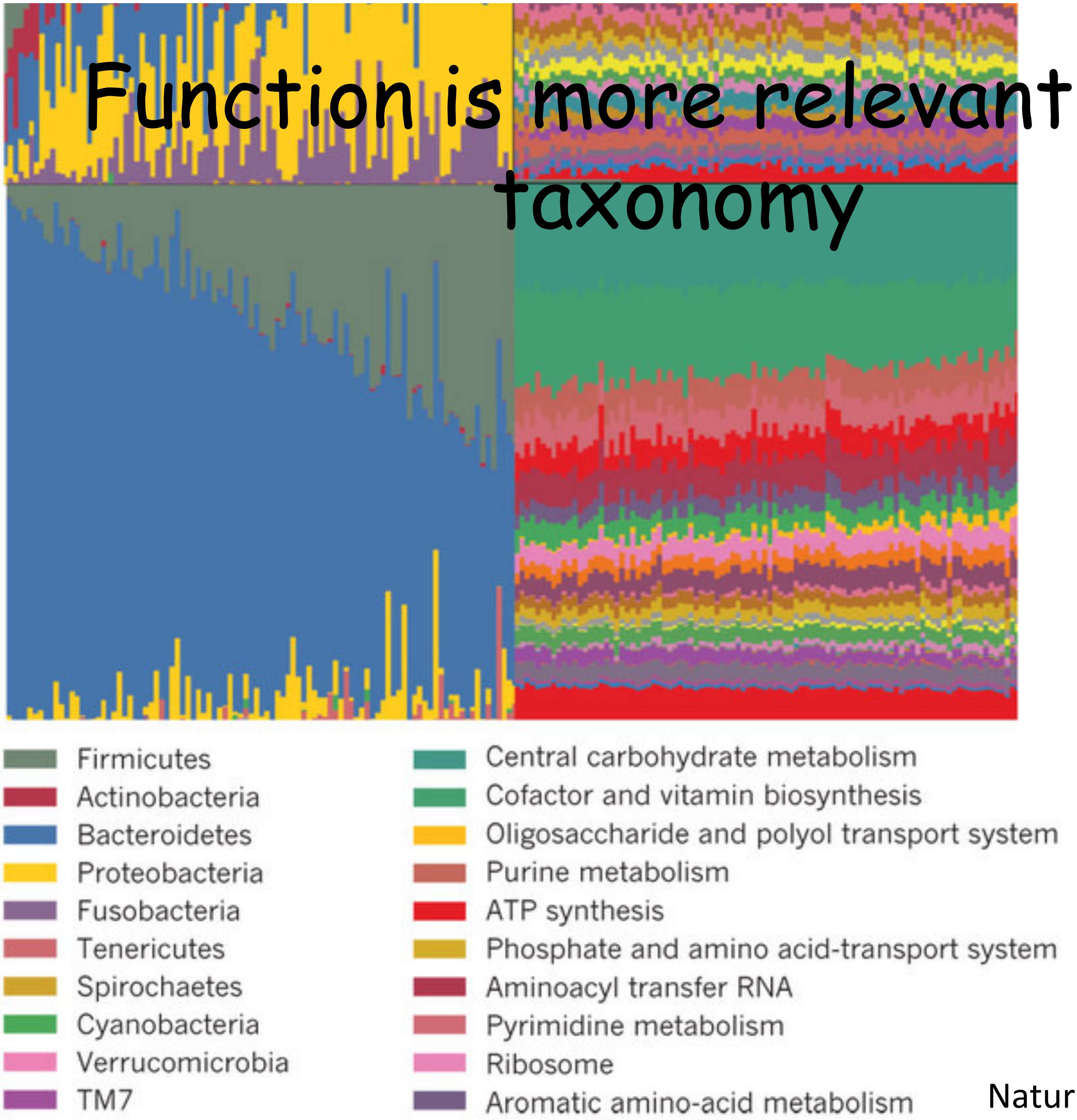
- Functional core?
- Interchangeable parts?

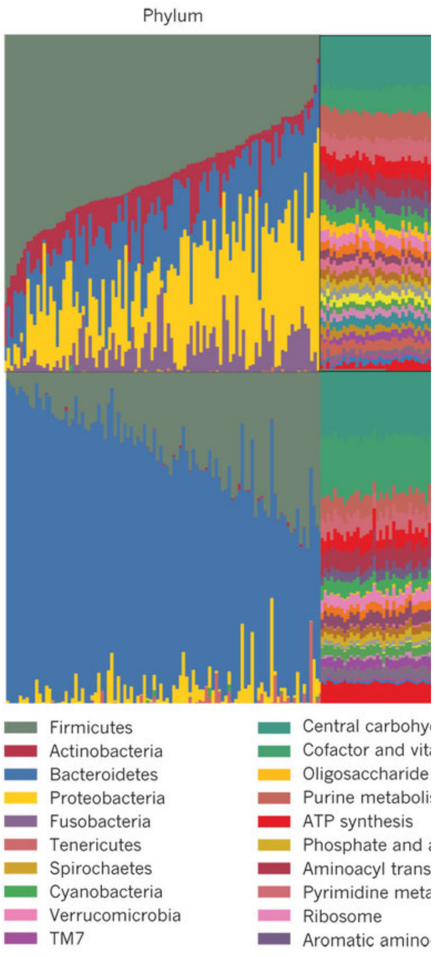
Comparison of taxonomic and functional variations



PJ Turnbaugh *et al. Nature* **000**, 1-5 (2008) doi:10.1038/nature07540

Function is more relevant than taxonomy

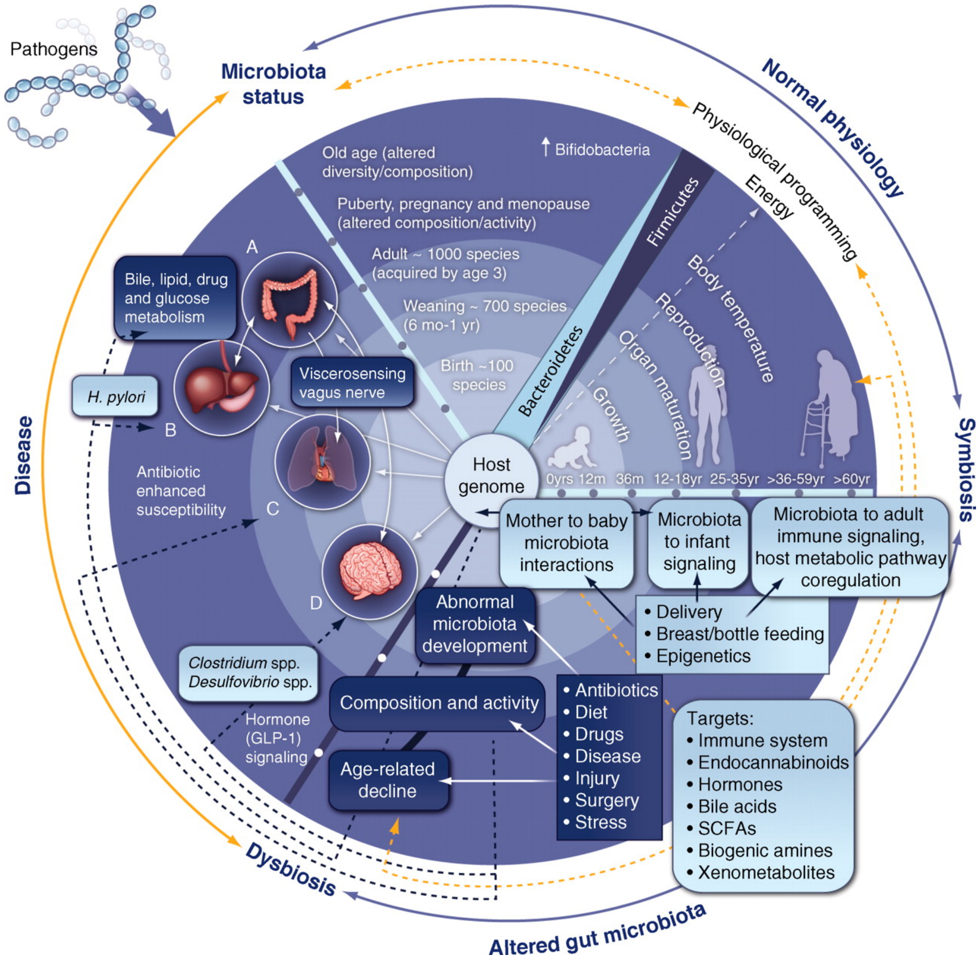




Nature 486 (2012)

Host-gut microbiota metabolic interactions

- After birth ~ 100 microbial spp.
- Env., nutrition, influence later development
- Microbiota influences normal development, physiology, immune system, etc, at all life stages
- Dysbiosis involved in a # of diseases:
  - IBD, IBS, colon cancer
  - ulcers, fatty liver, obesity
  - asthma, hypertension
  - Mood and behavior (GLP-1)



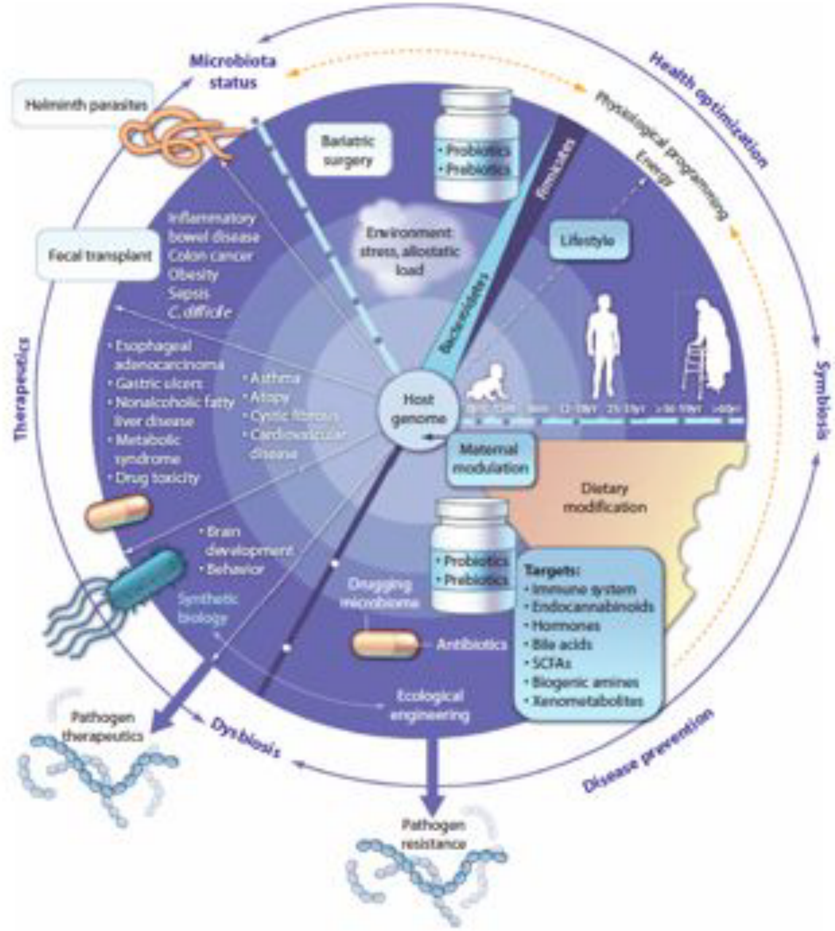
Pathogens
Microbiota status
Normal physiology
Physiological programming
Energy
Body temperature
Reproduction
Organ maturation
Growth
Symbiosis
Disease
Dysbiosis
Altered gut microbiota
Old age (altered diversity/composition)
↑ Bifidobacteria
Puberty, pregnancy and menopause (altered composition/activity)
Adult ~ 1000 species (acquired by age 3)
Weaning ~ 700 species (6 mo-1 yr)
Birth ~100 species
Firmicutes
Bacteroidetes
Host genome
0yrs 12m 36m 12-18yr 25-35yr >36-59yr >60yr
A
B
C
D
Bile, lipid, drug and glucose metabolism
H. pylori
Viscerosensing vagus nerve
Antibiotic enhanced susceptibility
Clostridium spp. Desulfovibrio spp.
Hormone (GLP-1) signaling
Abnormal microbiota development
Composition and activity
Age-related decline
Mother to baby microbiota interactions
Microbiota to infant signaling
Microbiota to adult immune signaling, host metabolic pathway coregulation
• Delivery
• Breast/bottle feeding
• Epigenetics
• Antibiotics
• Diet
• Drugs
• Disease
• Injury
• Surgery
• Stress
Targets:
• Immune system
• Endocannabinoids
• Hormones
• Bile acids
• SCFAs
• Biogenic amines
• Xenometabolites

J K Nicholson et al. Science 2012;336:1262-1267

Host-gut microbiota metabolic interactions

Is engineered homeostasis achievable?

- C. difficile transplants



J K Nicholson et al. Science 2012;336:1262-1267

Do you trust the microbiome?

5 questions:

1) Can experiments detect differences that matter?
2) Do studies show causation or just correlation?
3) What is the mechanism?
4) How much do experiments reflect reality?
5) Could anything else explain the results?

Evaluation of a diagnostic test

- Sensitivity

- Specificity

Calculating sensitivity and specificity

		True disease	
		+	-
Test	+	a 10	b 2
	-	c 5	d 83

Sensitivity = a/(a+c)   d/(b+d) = Specificity

10/15   83/85

Test Accuracy

		True disease	
		+	−
Test	+	a 20	b 2
	−	c 10	d 68

Accuracy =
88/100 = 88%

Prevalence =
30/100 = 30%

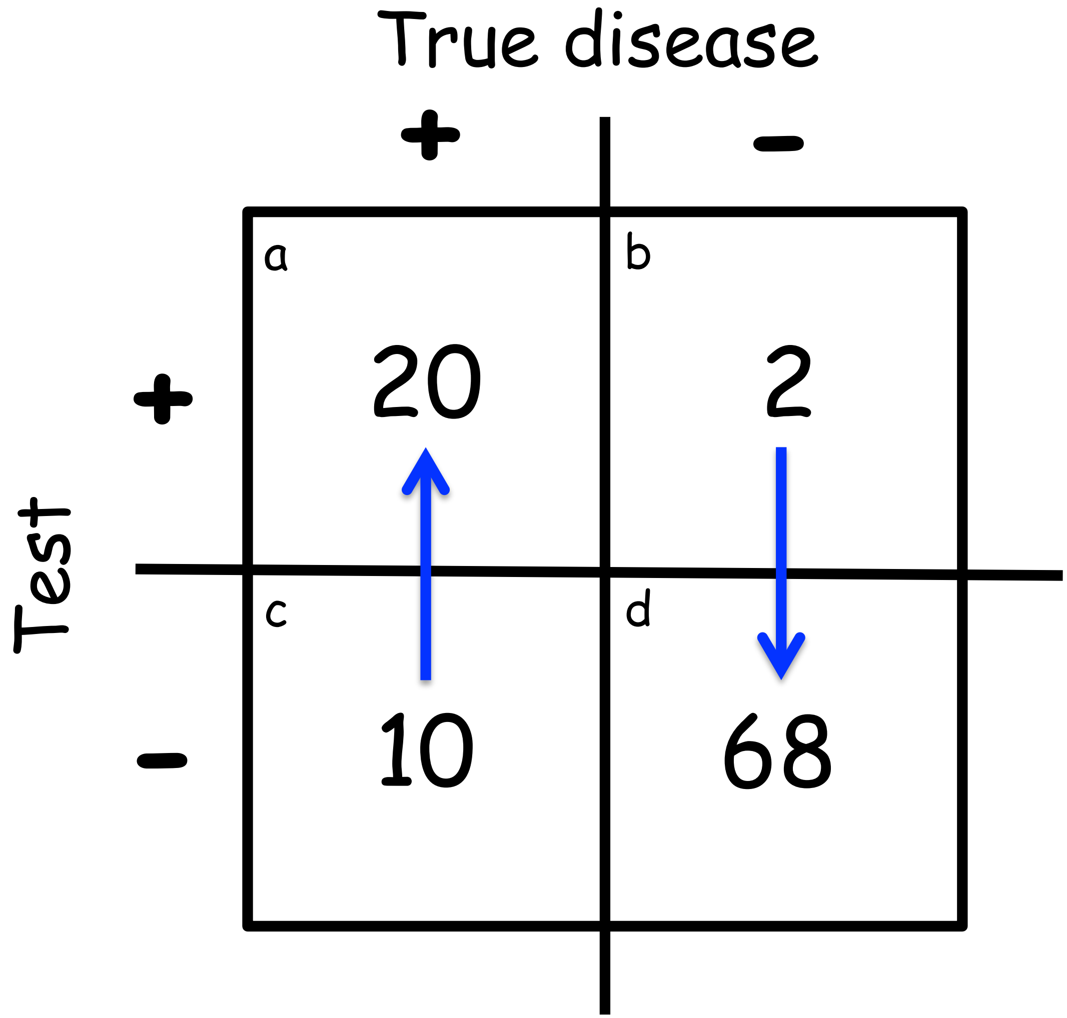
Savage Chickens

by Doug Savage

GOOD NEWS - IT'S JUST A CASE OF SLIGHTLY-MIFFED COW DISEASE
© 2007 BY DOUG SAVAGE

www.savagechickens.com

Sensitivity and Specificity and Predictive values



		True disease	
		+	−
Test	+	a 20	b 2
	−	c 10	d 68

Positive Predictive Value = a/(a+b)

20/22 ↑

10/12

Negative Predictive Value = d/(c+d)

68/78 ↓

83/88

Sensitivity = 67%

98% = Specificity

Liklihood ratios – diagnostic utility of a test

		True disease	
		+	-
Test	+	a 20	b 2
	-	c 10	d 68

Liklihood Ratio for a Positive Test= $\dfrac{a/a+c}{1-(d/b+d)}$

$$33.5 = \frac{20/30}{1-(68/70)}$$

Liklihood Ratio for a Negative Test= $\dfrac{1-(a/a+c)}{d/b+d}$

$$0.33 = \frac{1-(20/30)}{68/70}$$

Sensitivity = 67%

98% = Specificity

Comparing tests?

- When is a test with high sensitivity most useful?

- When is a test with high specificity most useful?