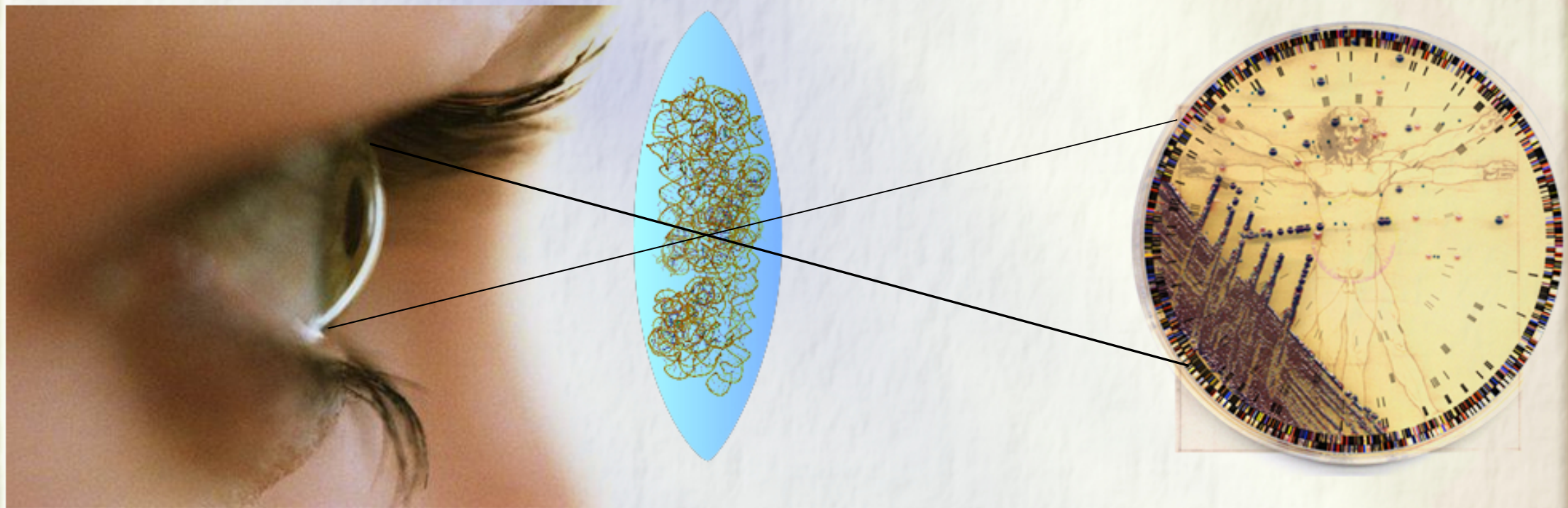


# QIIME: Quantitative Insights Into Microbial Ecology



**Rob Knight**

**HHMI, and Dept. Chemistry & Biochemistry  
University of Colorado, Boulder**



# Humans have always asked key questions about life



- What kinds of animals are there? (taxonomy)
- How many kinds of animals can I find where I hunt? (alpha diversity)
- How different is one place from another? (beta diversity)
- Who can I eat, and who can eat me? (interactions)

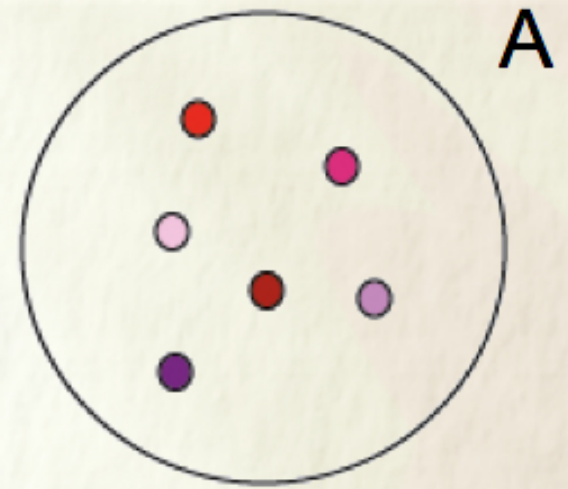


# How do we describe and compare diversity?

## $\alpha$ Diversity:

“How many species are in a sample?”  
(e.g. 6 colors in A and 6 in B)

e.g.: Are polluted environments  
less diverse than pristine?

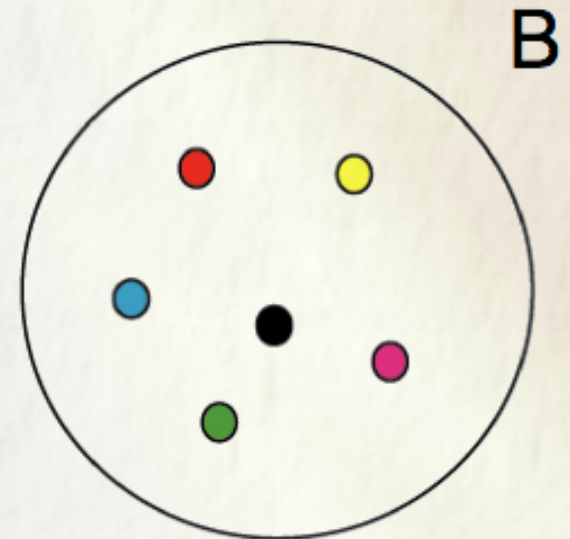


## $\beta$ Diversity:

“How many species are shared  
between samples?”

(e.g. 2 shared colors between A and B)

e.g.: Does the microbiota differ with  
different disease states?





# Qualitative versus quantitative measures

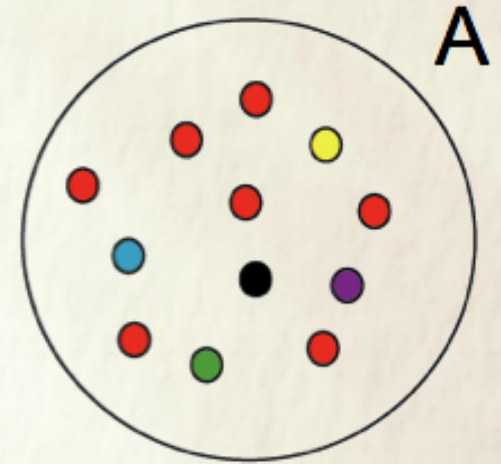
**Qualitative:** Considers presence/absence only

$\alpha$ : How many species are in a sample?

e.g.: 6 colors in both A and B.

$\beta$ : How many species shared among samples?

e.g.: A and B are identical because the same colors are present in both.

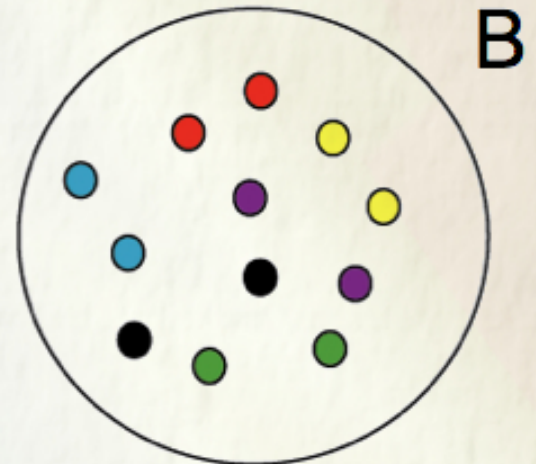


**Quantitative:** Also considers relative abundance.

$\alpha$ : Accounts for “evenness”:

e.g. B, where the population is evenly distributed across the 6 species, is more diverse than A, where all species are present but red dominates.

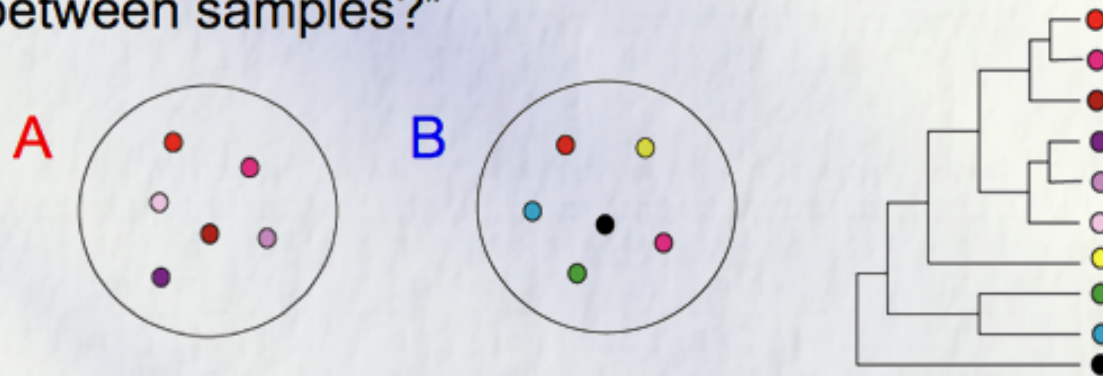
$\beta$ : Samples will be considered more similar if the same species are numerically dominant versus rare.  
e.g. B and A no longer look identical because of differences in abundance.



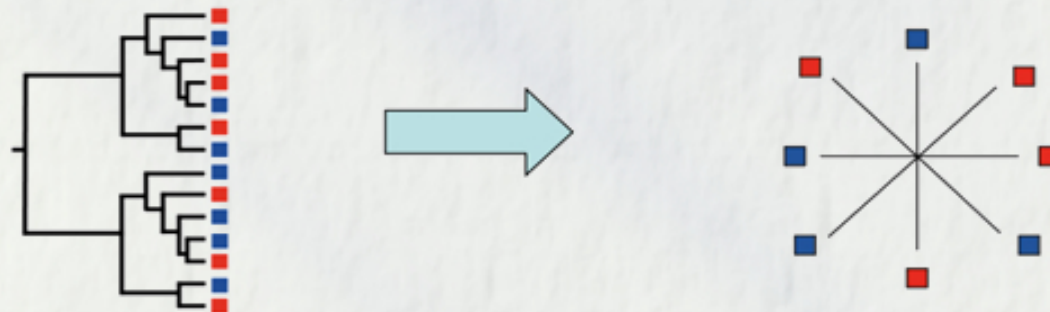


# Phylogenetic vs Taxonomic diversity measures

- Taxonomic: “How many species are shared between samples?”
- Phylogenetic: “How much phylogenetic distance is shared between samples?”



- Phylogenetically related organisms more likely to have similar roles in a community.
- Taxon-based methods assume a “star phylogeny,” where all relationships between taxa are ignored.





# Lots of diversity measures exist...

| name                      | formula                                                                                                                                                                                      | ref       |
|---------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------|
| bray-curtis               | $D_{ab} = \frac{\sum_j  A_{aj} - A_{bj} }{\sum_j (A_{aj} + A_{bj})}$                                                                                                                         | [1]       |
| cunbera                   | $D_{ab} = \frac{1}{Num_{a \cap b}} \sum_j  A_{aj} - A_{bj} $                                                                                                                                 | [1]       |
| $\chi^2$ (chi-squared)    | $D_{ab} = \sqrt{A_{++} \sum_j \left( \frac{1}{A_{+j}} \left( \frac{A_{aj}}{A_{+j}} - \frac{A_{bj}}{A_{b+}} \right)^2 \right)}$                                                               | [1]       |
| chord                     | $D_{ab} = \sqrt{\sum_j \left( \frac{A_{aj}}{\sqrt{\sum_j A_{aj}^2}} - \frac{A_{bj}}{\sqrt{\sum_j A_{bj}^2}} \right)^2}$                                                                      | [2]       |
| euclidean                 | $D_{ab} = \sqrt{\sum_j (A_{aj} - A_{bj})^2}$                                                                                                                                                 | [1]       |
| gower                     | $D_{ab} = \sum_j \frac{ A_{aj} - A_{bj} }{(\max_i A_{ij} + \max_i A_{bj})}$                                                                                                                  | [1]       |
| hellinger                 | $D_{ab} = \sqrt{\sum_j \left( \sqrt{\frac{A_{aj}}{A_{+j}}} - \sqrt{\frac{A_{bj}}{A_{b+}}} \right)^2}$                                                                                        | [3]       |
| kulczynski (quantitative) | $D_{ab} = 1 - \frac{1}{2} \left( \frac{\sum_j \min(A_{aj}, A_{bj})}{\sum_j A_{aj}} + \frac{\sum_j \min(A_{aj}, A_{bj})}{\sum_j A_{bj}} \right)$                                              | [1]       |
| manhattan                 | $D_{ab} = \sum_j  A_{aj} - A_{bj} $                                                                                                                                                          | [1]       |
| morisita-horn             | $\frac{2 \sum_j A_{aj} A_{bj}}{\left( \frac{\sum_j A_{aj}^2}{\sum_j A_{aj}} + \frac{\sum_j A_{bj}^2}{\sum_j A_{bj}} \right) \sum_j A_{aj} + \sum_j A_{bj}}$                                  | [4]       |
| pearson                   | $D_{ab} = 1 - \frac{\sum_j (A_{aj} - (\frac{A_{+j}}{M})) \sum_j (A_{bj} - (\frac{A_{b+}}{M}))}{\sqrt{\sum_j (A_{aj} - (\frac{A_{+j}}{M}))^2} \sqrt{\sum_j (A_{bj} - (\frac{A_{b+}}{M}))^2}}$ | [5]       |
| soergel                   | $D_{ab} = \frac{\sum_j  A_{aj} - A_{bj} }{\sum_j \max(A_{aj}, A_{bj})}$                                                                                                                      | [6]       |
| spearman                  | $D_{ab} = 1 - \frac{6 \sum_j ((\text{rank}(A_{aj}) - \text{rank}(A_{bj}))^2)}{M(M-1)}$                                                                                                       | [7]       |
| species profile           | $D_{ab} = \sqrt{\sum_j \left( \frac{A_{aj}}{A_{+j}} - \frac{A_{bj}}{A_{b+}} \right)^2}$                                                                                                      | [2]       |
| qualitative methods       |                                                                                                                                                                                              |           |
| $\chi^2$ (qualitative)    | see quantitative $\chi^2$                                                                                                                                                                    | see above |
| chord                     | see quantitative chord                                                                                                                                                                       | see above |
| euclidean                 | see quantitative euclidean                                                                                                                                                                   | see above |
| hamming                   | $D_{ab} = Num_a + Num_b - 2Num_{a \cap b}$ , see also manhattan                                                                                                                              | [8]       |
| jaccard                   | $D_{ab} = 1 - \left( \frac{Num_{a \cap b}}{Num_a + Num_b - Num_{a \cap b}} \right)$ , see also soergel                                                                                       | [9]       |
| lennon                    | $D_{ab} = 1 - \frac{Num_{a \cap b} + \min(Num_a - Num_{a \cap b}, Num_b - Num_{a \cap b})}{Num_{a \cap b}}$                                                                                  | [10]      |
| ochiai                    | $D_{ab} = 1 - \frac{Num_a + Num_b}{\sqrt{Num_a Num_b}}$                                                                                                                                      | [11]      |
| pearson                   | see quantitative pearson                                                                                                                                                                     | see above |
| dice                      | $D_{ab} = 1 - \frac{2Num_{a \cap b}}{Num_a + Num_b}$                                                                                                                                         | [12]      |



# Sequencing is getting much cheaper...

**Human Genome Project  
cost roughly \$3 billion,  
12 years (1991-2003)**

**Sequencing James  
Watson's genome in 2007  
cost \$2 million, 2 months**

**Today, you could get your  
genome sequenced for  
about \$2,500 and it would  
take a couple of weeks**

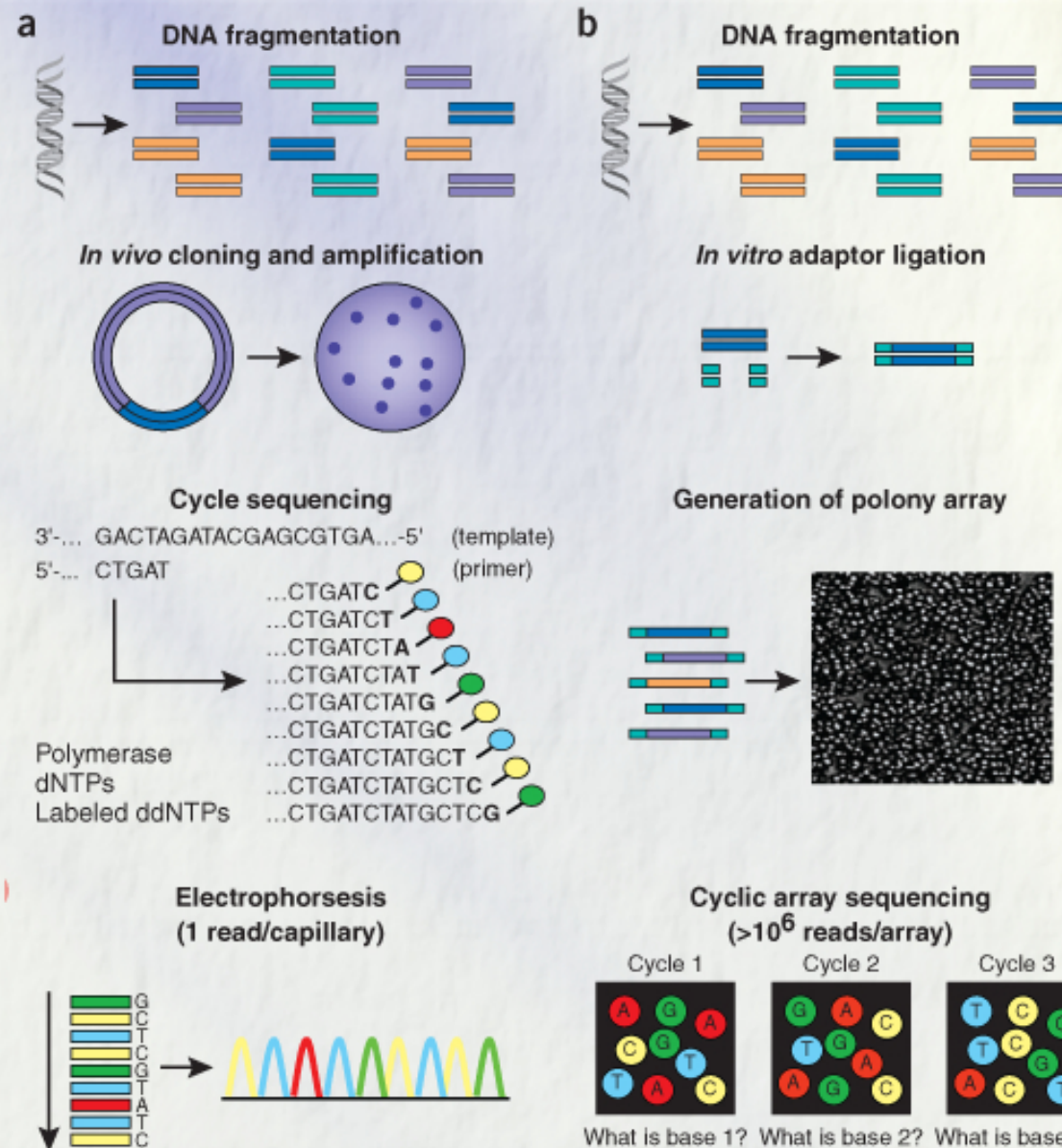
©\_Original Artist  
Reproduction rights obtainable from  
[www.CartoonStock.com](http://www.CartoonStock.com)



"Sequencing? No, this baby tells us  
how much we can charge for genome data."

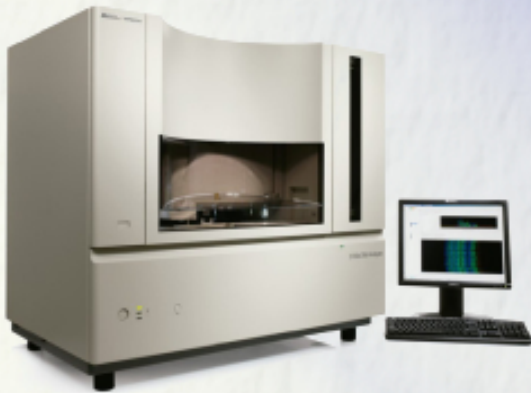


# Sanger vs Illumina sequencing





# Sequencing instruments



Applied Biosystems  
ABI 3730XL

Applied Biosystems  
SOLiD



Illumina/Solexa  
HiSeq

Helicos  
Heliscope



Roche 454  
Genome Sequencer FLX



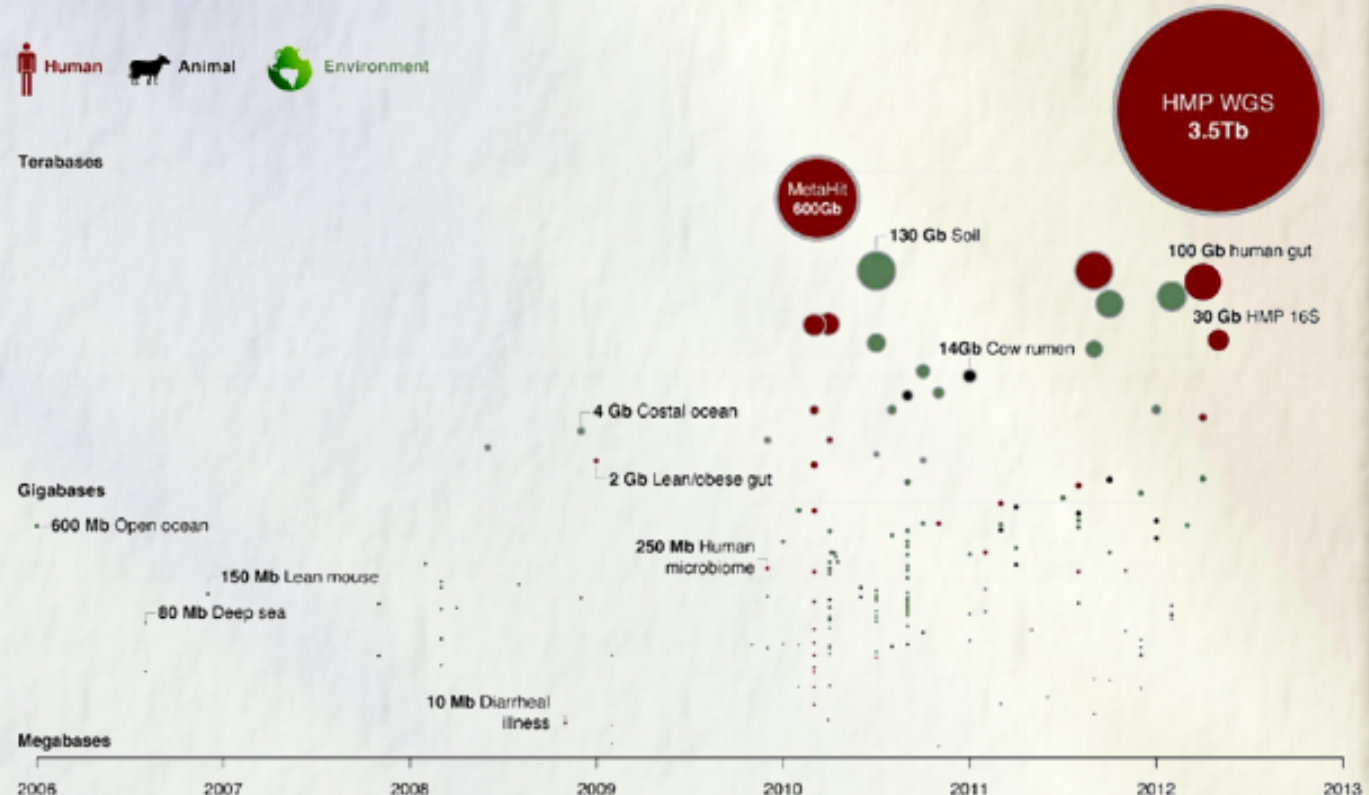
Illumina/Solexa  
Genome Analyzer II





# Sequencing technology rapidly advancing, enables big projects

|                      | AB 3730 | 454 Titanium | Illumina GAIIX | Illumina hiSeq2000 | MiSeq    | PacBio | IonTorrent |
|----------------------|---------|--------------|----------------|--------------------|----------|--------|------------|
| read length          | 800b    | 500b         | 76-101b        | 101-151b           | 101-151b | 1100b  | 100b       |
| runtime (hours/days) | 2h      | 9h           | 6-9d           | 9-15d              | 12-27h   | 1.5h   | 2h         |
| reads/run            | 96      | 1M           | 400M           | 1.6B               | 10M      | 35k    | 1.5M       |
| raw error rate       | 0.001%  | 1%           | <1%            | <1%                | <1%      | 15%    | 2%         |
| cost factor          | 100     | 1            | 0.1            | 0.025              | 1.5      | ?      | ?          |



Kuczynski et al. 2011 Nat Rev Genet;  
Gevers et al. 2012 PLoS Biol



# Sequencing technologies to watch out for!



# Sequencing technologies to watch out for!

**Emerald Isle, Inc.:**

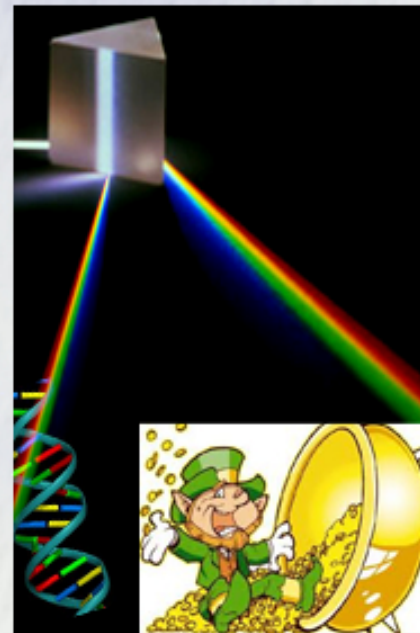
**Leprechaun-based technology uses half-silvered mirror on the rainbow, lets you choose sequences, pot of gold (no more grant writing!)**

## A New Breakthrough in DNA Sequencing

Published: January 4, 2011

DNA sequencing is proceeding at a breathtaking pace. But only one company has had the foresight to combine advanced technology with traditional folklore and wisdom dating back thousands of years.

[Enlarge This Image](#)



Leprechaun-based sequencing promises to cure cancer and to bring the pot of gold at the end of the rainbow into every laboratory.

Emerald Isle's leprechaun-based technology is based on spectral decomposition techniques that allow conversion of the pot of gold at the end of the rainbow into DNA sequences. The key breakthrough is that by using a half-silvered prism, they can achieve fractional conversion, thus leaving some of the gold intact so that the sequencing pays for itself (or, if lower coverage is needed, it is possible to turn a profit, reducing the need for time-consuming grant applications).

Recovery of the gold/sequences has thus far proved somewhat more challenging than anticipated in their last quarterly report, but all involved with the company see bright prospects for the future. ``We believe that this obstacle will not be insurmountable and are shipping the instrument now in the anticipation that our customers will be able to resolve this issue on their own". a source close to the company noted.

[E-MAIL](#)

[PRINT](#)

[SINGLE PAGE](#)

[REPRINTS](#)

[SHARE](#)





# Sequencing technologies to watch out for!

## Hail Mary Technologies:

Lies make the baby Jesus cry. By measuring the flow rate, the error rate can be recalibrated for unprecedented accuracy.

### DNA Sequencing Advancing at Rapid Pace

Published: January 6, 2011

Interest in the intersection of science and religion is at an all-time high point. Yet only one company has discovered how to use a well-established religious belief to improve the lives of those suffering from tragic genetic diseases.

[Enlarge This Image](#)



The error rate at each position in the DNA can be determined by how much the baby Jesus cries when the result is reported to the user. Regrettably, the production facility has experienced delays due to lightning strikes.

It is a fact well-known to everyone's Sunday school teacher that when you tell a lie, the baby Jesus cries. This phenomenon can be exploited to reduce the error rate common in many advanced DNA sequencing technologies.

A revolutionary new instrument from Hail Mary Technologies is extremely inaccurate on each pass, but depending on the magnitude of the lies being told about each nucleotide position (e.g. transitions vs. transversions), many of these errors can be corrected based on the tear flow rate.

The production facility has regrettably experienced an unusual number of lightning strikes, but sources close to the company are optimistic that this issue will be resolved in the near future.

 E-MAIL

 PRINT

 SINGLE PAGE

 REPRINTS

 SHARE





# Sequencing technologies to watch out for!

Petagenomics GmbH:

Throughput optimized to produce  $10^{15}$  sequences (each 1 base of Q17 data) and can sequence every rRNA in the human gut!

Taxonomy assignment methods are rapidly improving...

## Unprecedented Throughput in DNA Sequencing

Published: January 8, 2011

The human microbiome, the 100 trillion cells that inhabit the human body, have recently been found to cure persistent *Clostridium difficile* infections, assist weight loss, cure cancer, and balance auras. However, characterizing these beneficial microbes has remained a challenge, in large part because there are so many of them.

- E-MAIL
- PRINT
- SINGLE PAGE
- REPRINTS
- SHARE

[Enlarge This Image](#)



By sequencing one base of all the ribosomal RNA genes in the human gut, PetaGenomics GmbH will revolutionize our understanding of our microbial symbionts.

A radical new technology from a new upstart company provides an innovative resolution to this problem. The technology is capable of sequencing a petabase (1 followed by 15 zeroes) of sequence for a remarkably modest cost, thus providing coverage of all the bacteria in the human gut (typically estimated to number 100 to 1000-fold fewer than this).

"We can now produce more sequences than there are in the gut of even the largest human, which is increasingly important as the obesity epidemic worsens", said a spokesman for the company. Critics of the technology note that it is only able to sequence one nucleotide from each sequence, making it difficult to determine which species are found in each sample. The company asserts that these concerns are unfounded. "Microbiologists have discovered that you don't need the whole sequence to characterize a community, and often a short fragment is enough", a

consultant for the company noted. "Although we are not at species-level resolution, we are sure this technology will have many clinical applications and prevent cute puppies from dying of horrible diseases."





# Sequencing technologies to watch out for!

## Monoceros:

Unicorns are fabled for their ability to detect poisons, including viral and bacterial toxins and even the genes that encode them...

## DNA Sequencing Advances Find Pathogens Everywhere

Published: January 12, 2011

[Enlarge This Image](#)



Unicorns provide an innovative solution to the problem of pathogen detection, although the requisite virgins are easier to come by in the R&D facility than in the field.

The revolution in DNA sequencing promises to pervade every aspect of our lives, from tracking objects we have touched to identifying germs lurking where they have never been seen before. But which of these tiny invaders are harmful, and which can live with us peacefully?

It is an established scientific principle that unicorns are able to purify poison using their horns. Scientists at Monoceros recently discovered that this effect extends to microbial toxins, and even to the genes and pathogenicity islands that encode these toxins. By magnifying the sequences using an electron microscope and carefully monitoring small changes in unicorn physiology, they can therefore detect which strands of DNA are potentially harmful.

Due to the nature of the principles involved, this technology requires an isolated, temperature-controlled unicorn stable, which has caused issues at some sites due to the importance of preventing horn entanglement in the cables. An additional issue encountered in scaling from R&D to production was that only virgins are able to tame the unicorns: although this was not a problem in the R&D facility, it has led to certain issues with field deployment of the instruments, either because customers were insufficiently pure to begin with or were no longer virgins after being serviced by marketing and/or support teams. However, those close to the company believe these issues will shortly be resolved and are expecting to ship sufficient instruments to meet revenue targets.

[E-MAIL](#)

[PRINT](#)

[SINGLE PAGE](#)

[REPRINTS](#)

[SHARE](#)





# Sequencing technologies to watch out for!

## Zer0-Molecule Genomics:

Single-molecule biophysics is too hard for many users because of lack of reproducibility.

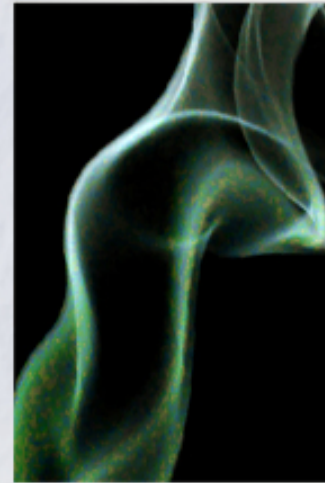
The next frontier, zero-molecule, fixes this issue.

Specificity excellent, sensitivity improving fast...

## DNA sequencing advances beyond single-molecule

Published: January 14, 2011

[Enlarge This Image](#)



ZoMG's zero-molecule technology has unparalleled reproducibility, allowing the same result to be obtained even in completely different experiments.

Single-molecule DNA sequencing has revolutionized the field of genomics, a remarkable feat given that the instruments that produce this task are not yet commercially available. Yet many potential customers are worried about the level of training required to interpret the data from the instruments, which is often highly variable.

- [E-MAIL](#)
- [PRINT](#)
- [SINGLE PAGE](#)
- [REPRINTS](#)
- [SHARE](#)



"Already my clients are concerned that what started off as a biology project has also turned into a sequence methods development project and a computer science project", said one core facility director who asked not to be named. "Now it's also a single-molecule biophysics project. I hear that the instrument is delayed because our institutional review board is still assessing the safety of the scalpel and

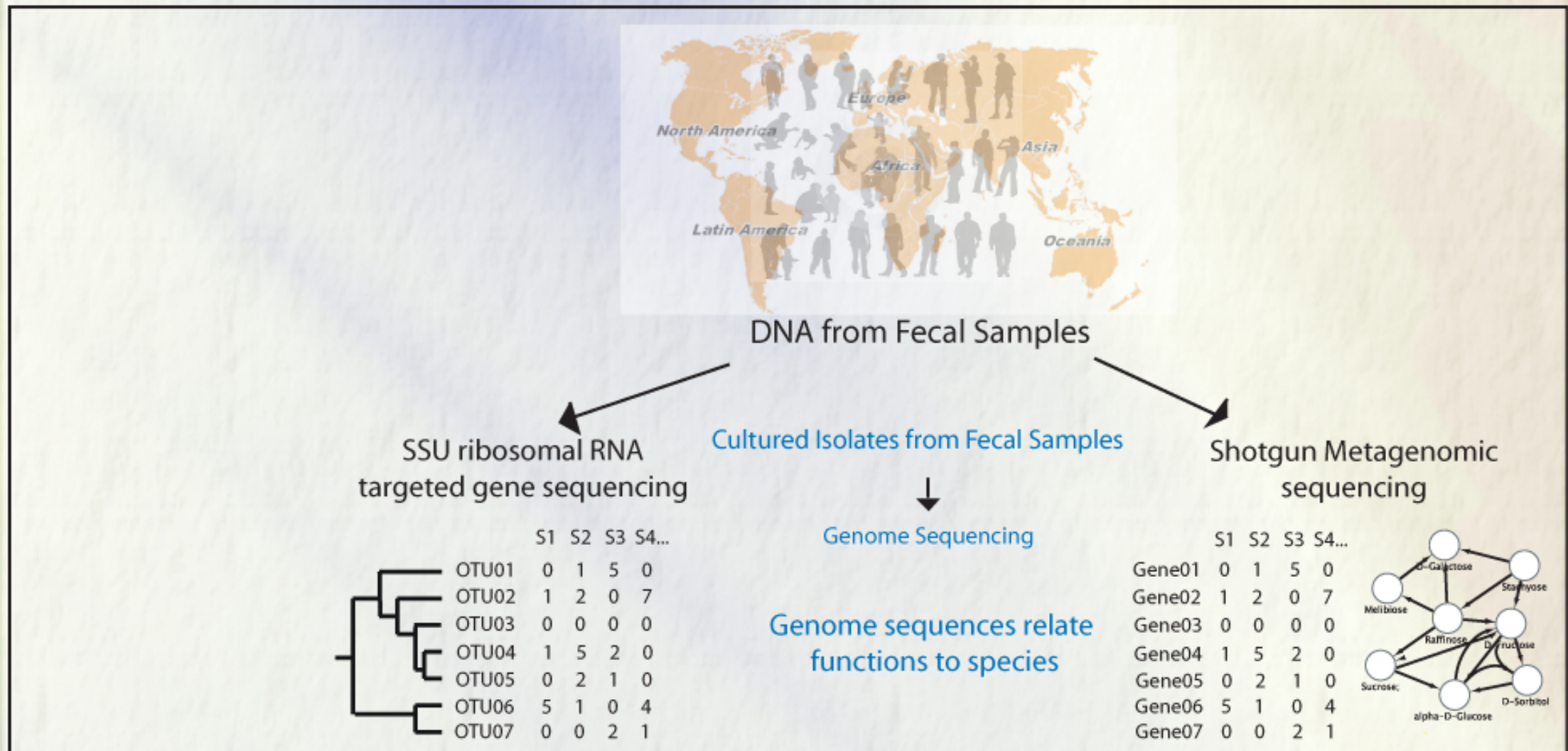
spaceship that it ships with, in order to accommodate the fact that it is also a brain surgery and a rocket science project".

Fortunately, new technology has provided a resolution to this conundrum. Zero-Molecule Genomics (ZoMG) has taken the next logical step: their zero-molecule sequencer operates without any molecules at all. The reproducibility of runs is excellent, and the false discovery rate is so low as to be undetectable.

"We have found that traditional ROC plots fail to fully capture the advantages of the device", a source close to the company said, "but by plotting only the specificity and omitting the sensitivity on a one-axis ROC plot, we can achieve excellent performance". This bioinformatics advance, together with the low run-to-run variability, will be invaluable for obtaining FDA approval for the instrument as a diagnostic device.



# Marker gene and shotgun studies produce abundance tables...





# ...that can be used for a suite of downstream analyses

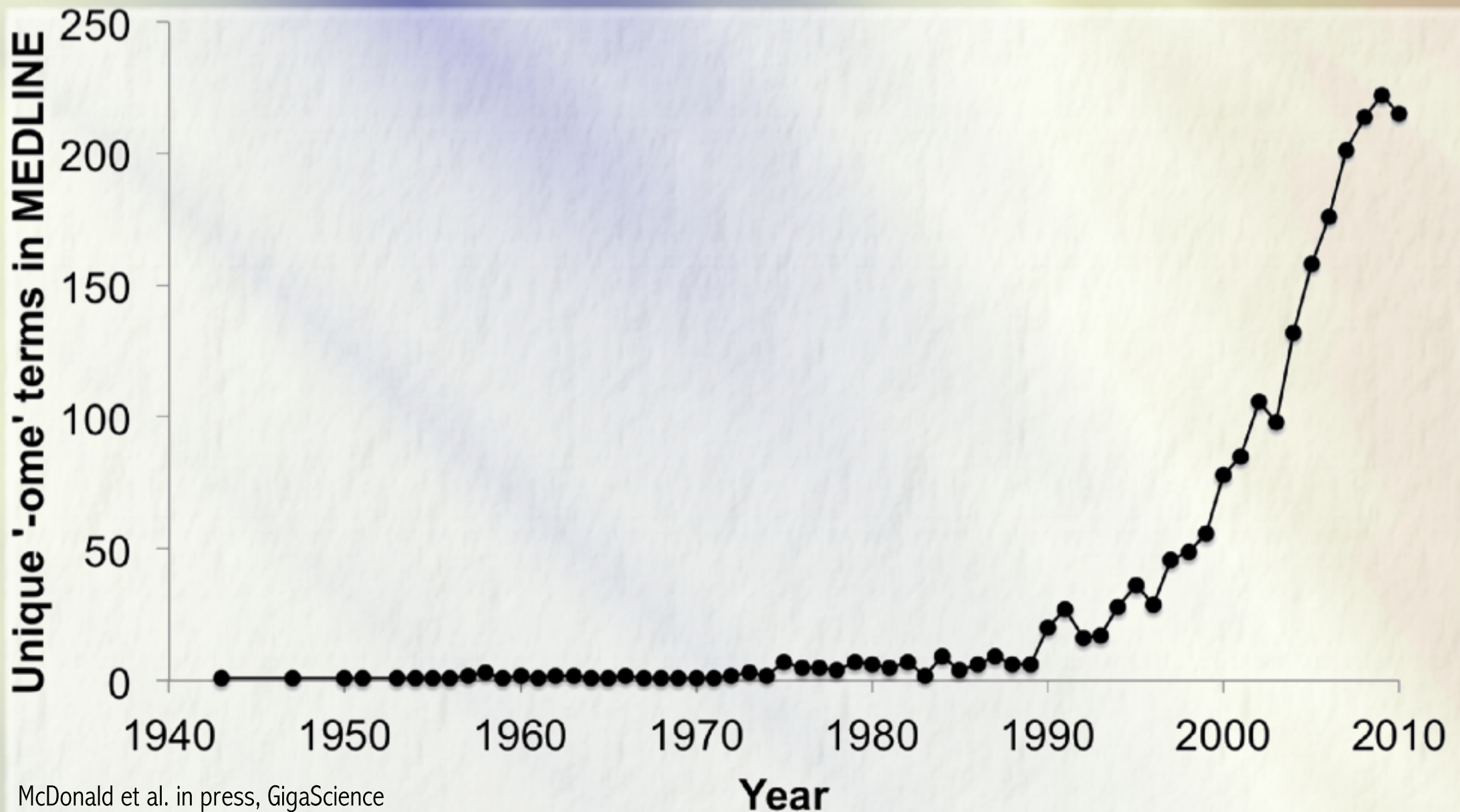
16S

Shotgun

|                                                                                                                          |                                                                                                                                                                                                                                                                                                                                                                                                                             |                                                                                                                                                                                                                                                                                                                          |                                                                                                                                                                                                                                                                                                                                                                                                         |
|--------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Alpha Diversity</b><br>              | <b>Which microbiota are the most compositionally diverse?</b><br><i>i.e.</i> have the most phylotypes or branch length?                                                                                                                                                                                                                                                                                                     | A scenario in which an increase in the number of taxa (compositional diversity) does not lead to an increase in the number of functions (functional diversity) can indicate functional redundancy.                                                                                                                       | <b>Which microbiota are the most functionally diverse?</b><br><i>i.e.</i> have the most gene families?                                                                                                                                                                                                                                                                                                  |
| <b>Beta Diversity</b><br>               | <b>Which factors correlate with differences in microbiota composition?</b><br><i>i.e.</i> do individuals of the same age, culture, or disease state have microbiota with more phylotypes or phylogenetic lineages in common?                                                                                                                                                                                                | Beta diversity plots based on taxonomic and functional composition are usually highly correlated. Indicating that functional genes have a high phylogenetic signal.                                                                                                                                                      | <b>Which factors correlate with differences in functional genes?</b><br><i>i.e.</i> do individuals of the same age, culture, or disease state have microbiota with a more similar representation of gene families?                                                                                                                                                                                      |
| <b>Machine Learning/ Classical statistics</b>                                                                            | <b>Which phylotypes differentiate groups?</b><br>Are there particular phylotypes or collections of phylotypes that discriminate/significantly differ between the microbiota of individuals in <i>e.g.</i> health and disease?                                                                                                                                                                                               | Functional differences are often largely reflective of taxonomic differences. Of interest are when selected functions cannot be explained by selected taxonomic groups; indicating potential functional convergence.                                                                                                     | <b>Which genes differentiate groups?</b><br>Are there particular gene families/ functions that discriminate/significantly differ between the microbiota of individuals in <i>e.g.</i> health and disease?                                                                                                                                                                                               |
| <b>Co-occurrence</b><br>              | <b>Which phylotypes co-distribute across people and what factors (e.g. diet or disease) explain distribution patterns of co-distributing groups?</b><br>Positive correlations can be driven by shared environmental preferences ( <i>e.g.</i> indicative of shared ecological attributes or symbiotic/syntrophic relationships). Negative correlations by divergent environmental preferences or competitive relationships. | Performing co-occurrence analyses on species with complete genomes can allow for further exploration of the driving factors of positive or negative correlations. For example genes/functions that are shared by phylogenetically unrelated bacteria that co-occur can indicate potential environmental driving factors. | <b>Which gene families co-distribute across people?</b><br>Not as useful as phylotype co distribution overall. Positive correlations likely driven by genes that co-occur in the same genome (phylotype).                                                                                                                                                                                               |
| <b>Metabolic Network Modeling</b><br> | <b>What metabolic activities would be predicted for a collection of phylotypes?</b><br>Metabolic predictions can in principal be made for collections of phylotypes if associated reference genome sequences are available.                                                                                                                                                                                                 | What metabolic activities would be predicted for a particular phylotype/species?                                                                                                                                                                                                                                         | <b>What overall metabolic activities would be predicted for a microbiota? What spectrum of metabolites would a microbiota produce from a defined diet?</b><br>A promising premise for designing personalized therapeutic diets in general, but challenging due to incomplete knowledge of which genes perform which enzymatic reactions, and which species express which enzymes in different contexts. |

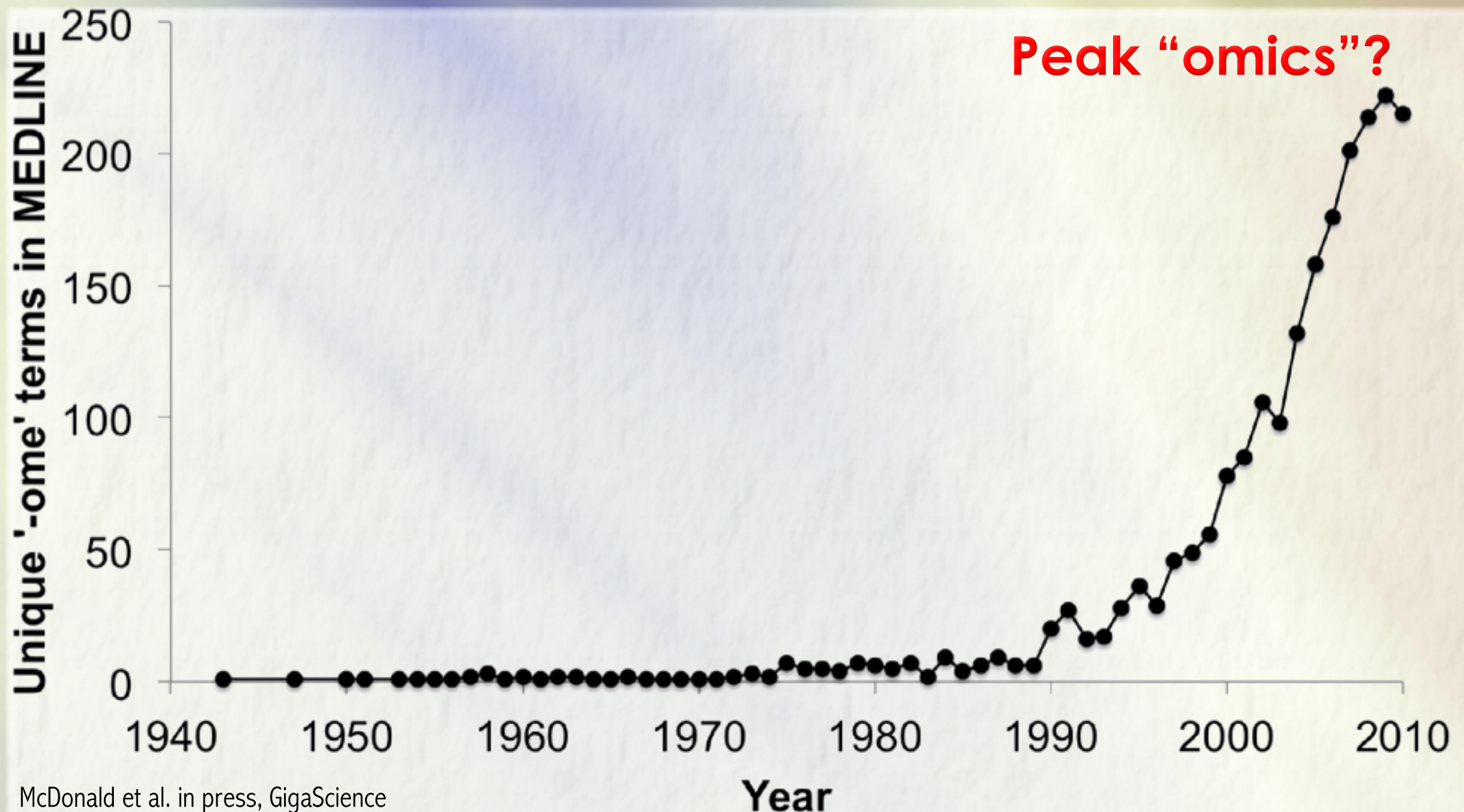


# It's important to recognize there are (too many?) other omics...





# It's important to recognize there are (too many?) other omics...

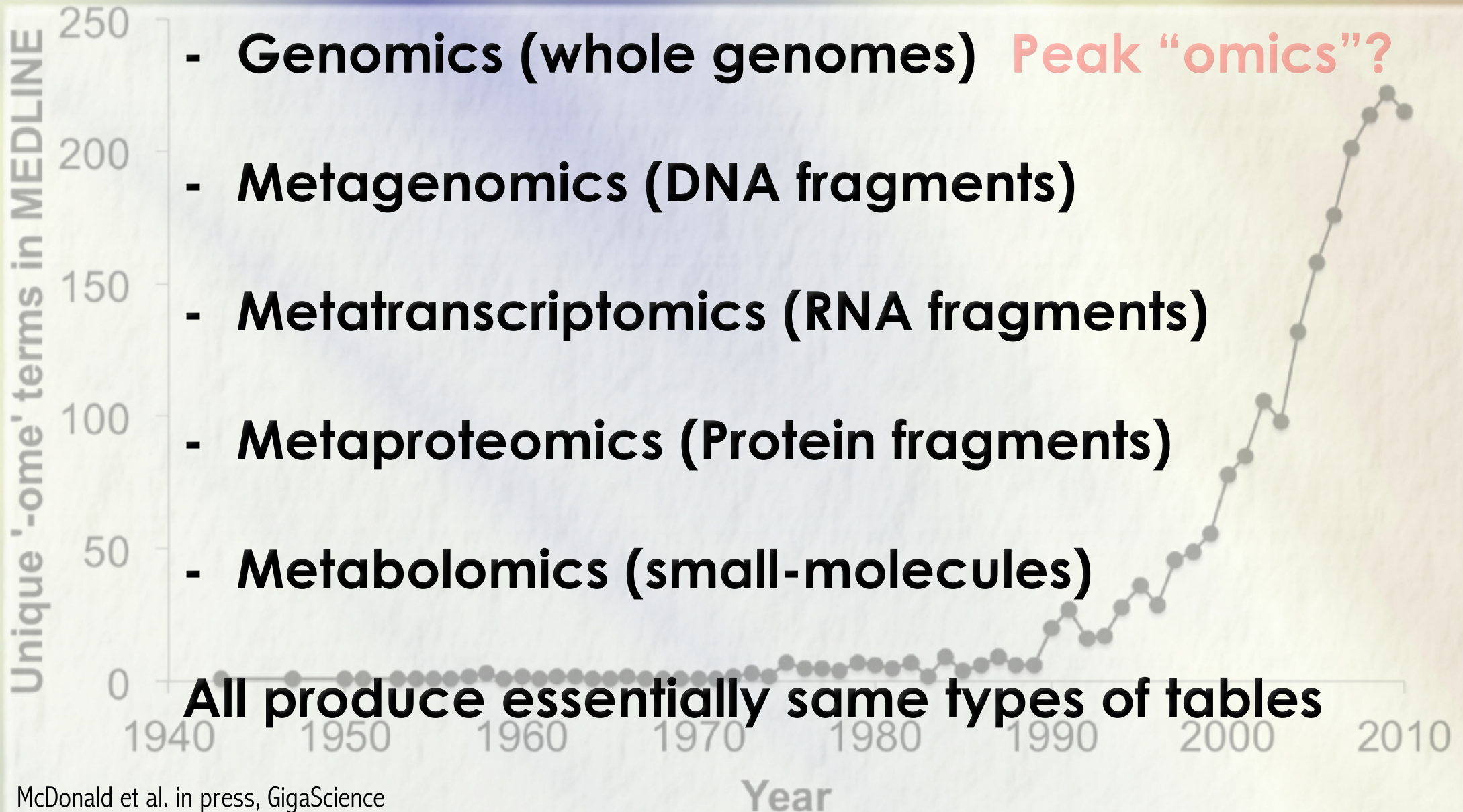




# ...but there are only a few “main” omics you need to worry about

- Genomics (whole genomes) **Peak “omics”?**
- Metagenomics (DNA fragments)
- Metatranscriptomics (RNA fragments)
- Metaproteomics (Protein fragments)
- Metabolomics (small-molecules)

**All produce essentially same types of tables**



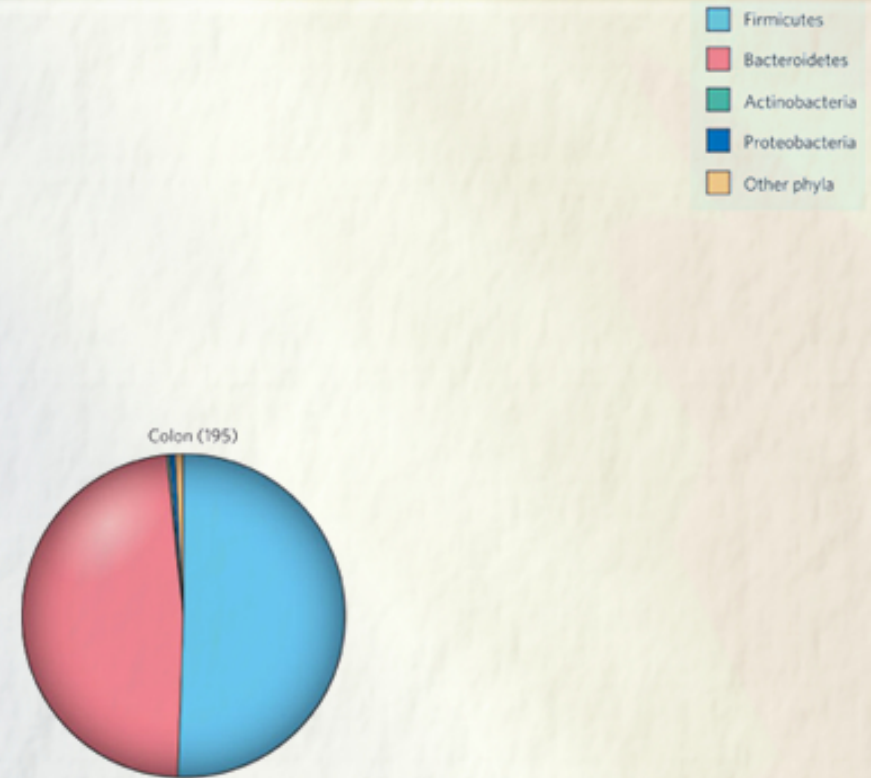


**Key challenge: relating different microbial communities**



# Key challenge: relating different microbial communities

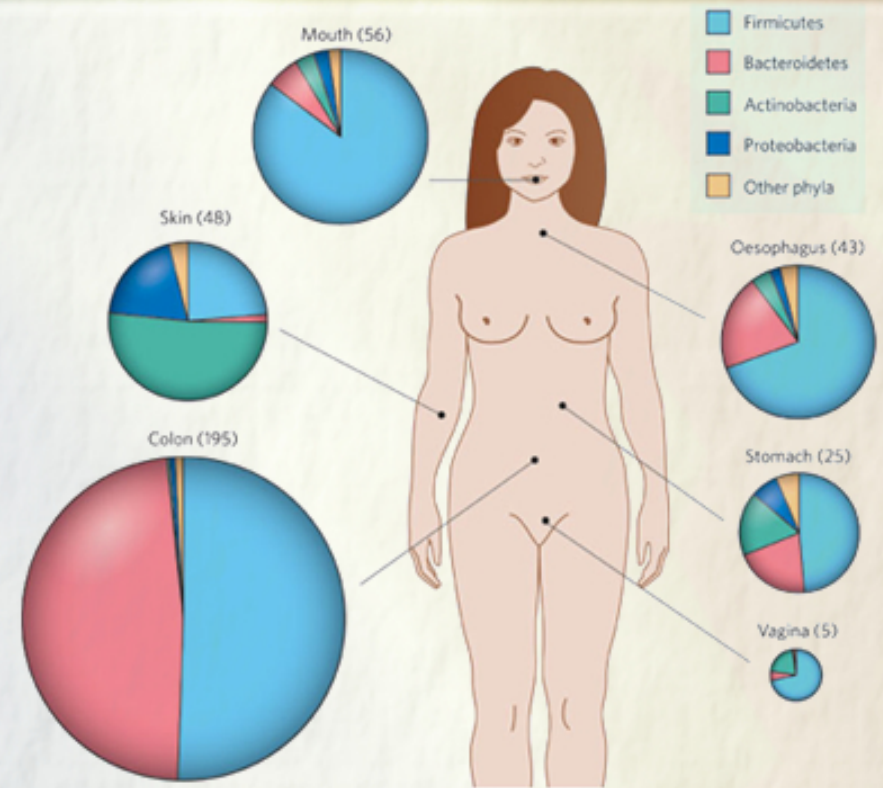
One community: can say something about taxa present, but not so informative...





# Key challenge: relating different microbial communities

...more informative to compare across communities, but pie chart approach doesn't scale...

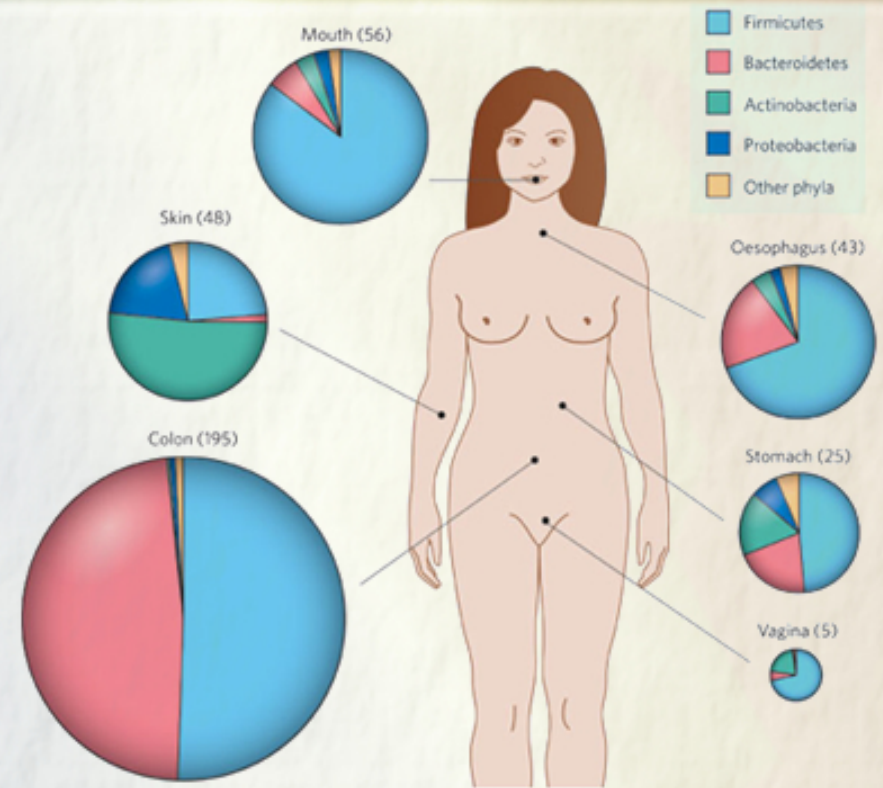




# Key challenge: relating different microbial communities

...more informative to compare across communities, but pie chart approach doesn't scale...

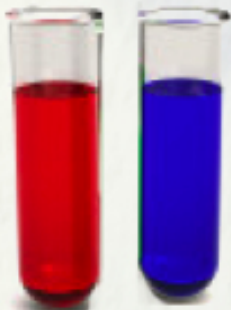
...and are the results different because of differences in the community, or because of technical biases? (in DNA extraction, PCR, analysis, etc.)





# Direct environmental sequencing sees the “other” 99% of microbes

1. Get samples  
and extract  
DNA



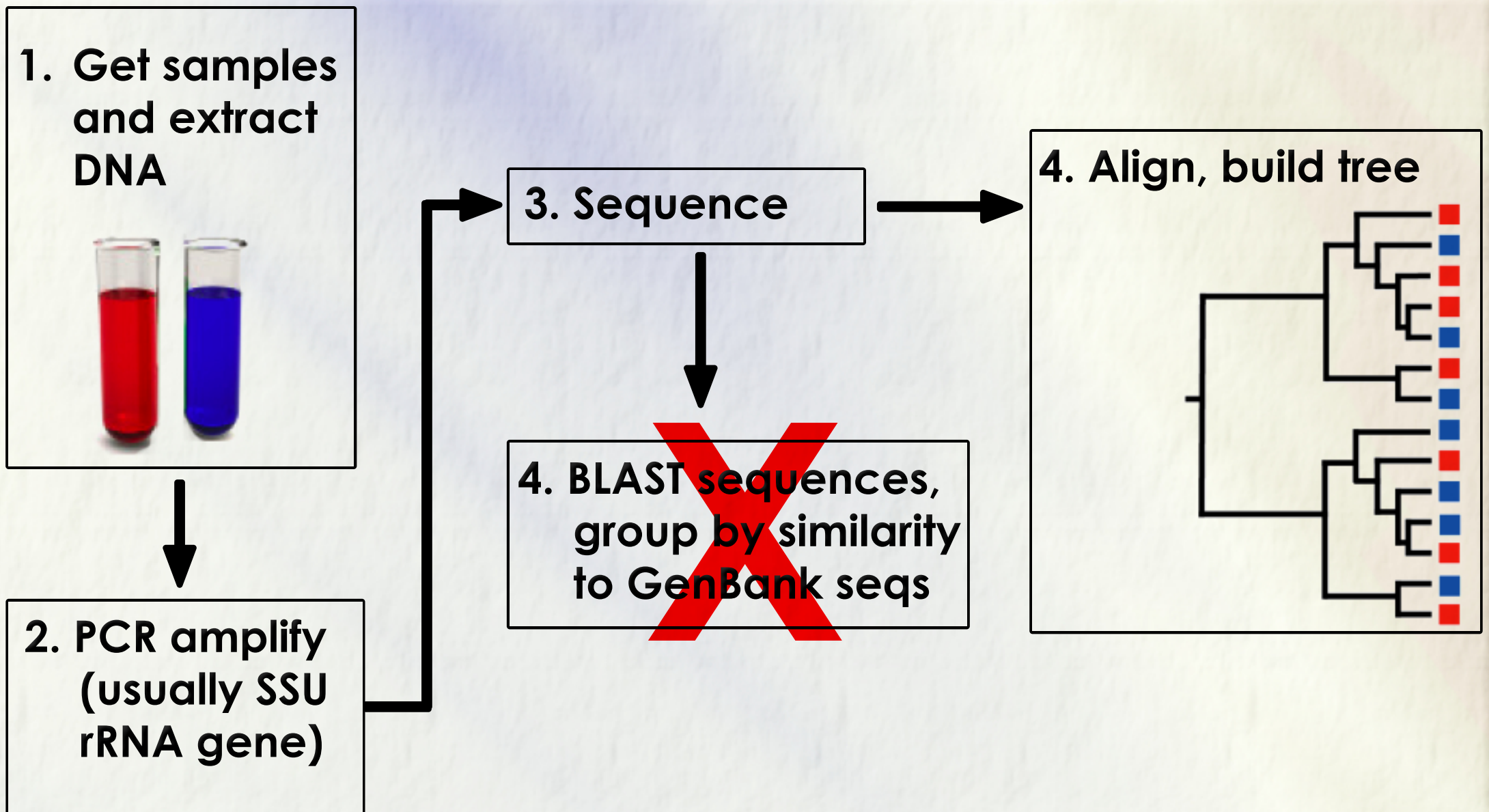
3. Sequence

4. BLAST sequences,  
group by similarity  
to GenBank seqs

2. PCR amplify  
(usually SSU  
rRNA gene)



# Direct environmental sequencing sees the “other” 99% of microbes

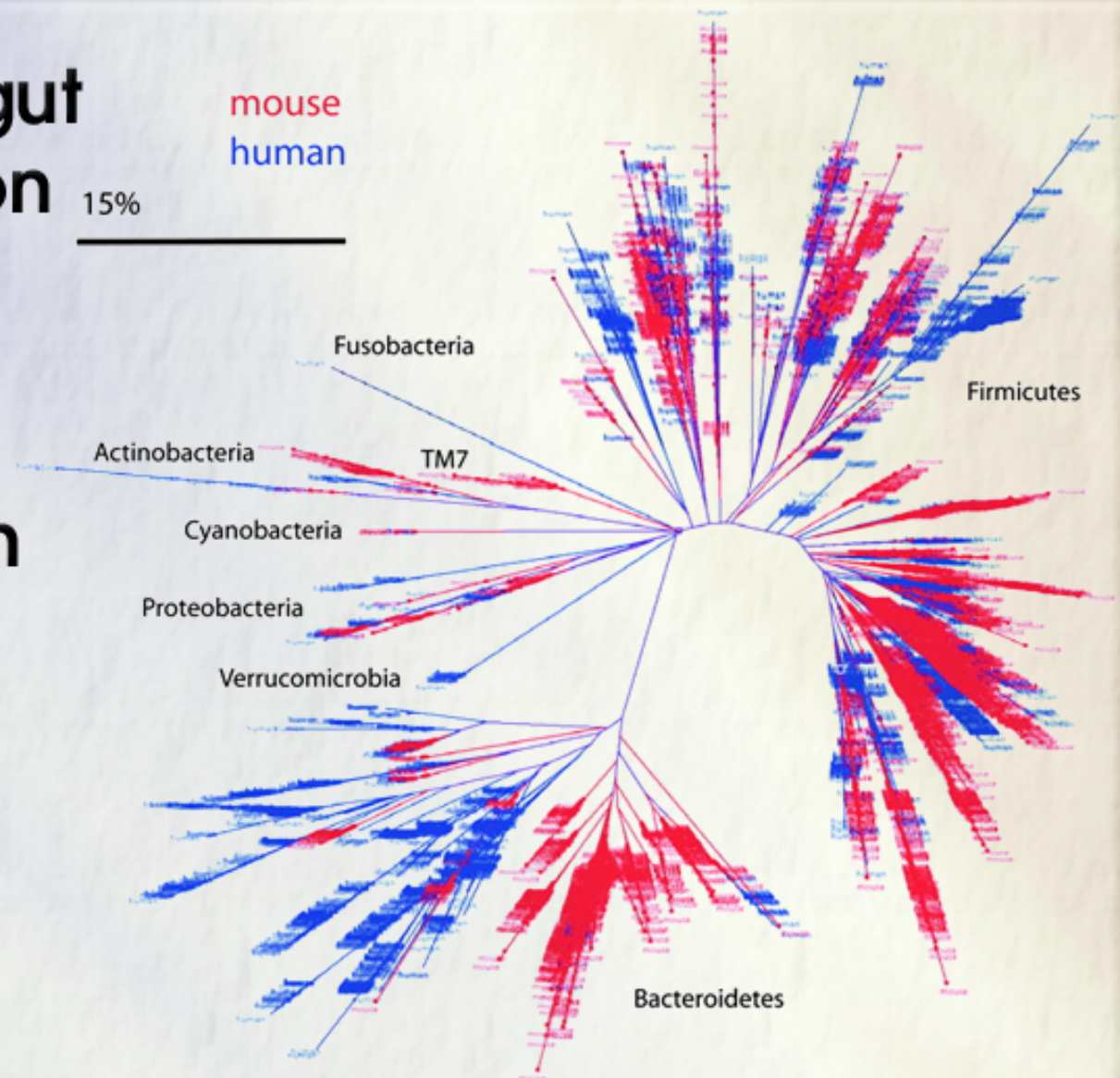




# Problem: big trees hard to understand and analyze

Example: 5088 mouse gut and 11,831 human colon bacterial sequences

- See many clusters of sequences from each sample
- Significance tests for differences, but no phylogenetic metric





# ...and we've all seen papers like this...



FEMS Microbiology Ecology 44 (2003) 231–242



[www.fems-microbiology.org](http://www.fems-microbiology.org)

## Molecular analysis of bacterial microbiota in the gut of the termite *Reticulitermes speratus* (Isoptera; Rhinotermitidae)

Yuichi Hongoh <sup>a,b,\*</sup>, Moriya Ohkuma <sup>a,b</sup>, Toshiaki Kudo <sup>a,b,c</sup>

<sup>a</sup> International Cooperative Research Project, Japan Science and Technology Corporation (JST-ICORP), Saitama, Japan

<sup>b</sup> Molecular Microbial Ecology Division, Bioscience Technology Center, RIKEN (The Institute of Physical and Chemical Research), Wako, Saitama 351-0198, Japan

<sup>c</sup> Division of Environmental Molecular Biology, Graduate School of Integrated Science, Yokohama City University, Kanagawa 230-0045, Japan

Received 5 August 2002; received in revised form 10 November 2002; accepted 29 December 2002

First published online 31 January 2003



# ...and we've all seen papers like this...



FEMS Microbiology Ecology 44 (2003) 231–242

**FEMS**  
MICROBIOLOGY  
Ecology

[www.fems-microbiology.org](http://www.fems-microbiology.org)

## Molecular analysis of bacterial microbiota in the gut of the termite *Reticulitermes speratus* (Isoptera; Rhinotermitidae)

Yuichi Hongoh<sup>a,b,\*</sup>, Moriya Ohkuma<sup>a,b</sup>, Toshiaki Kudo<sup>a,b,c</sup>

<sup>a</sup> International Cooperative Research Project, Japan Science and Technology Corporation (JST-ICORP), Saitama, Japan

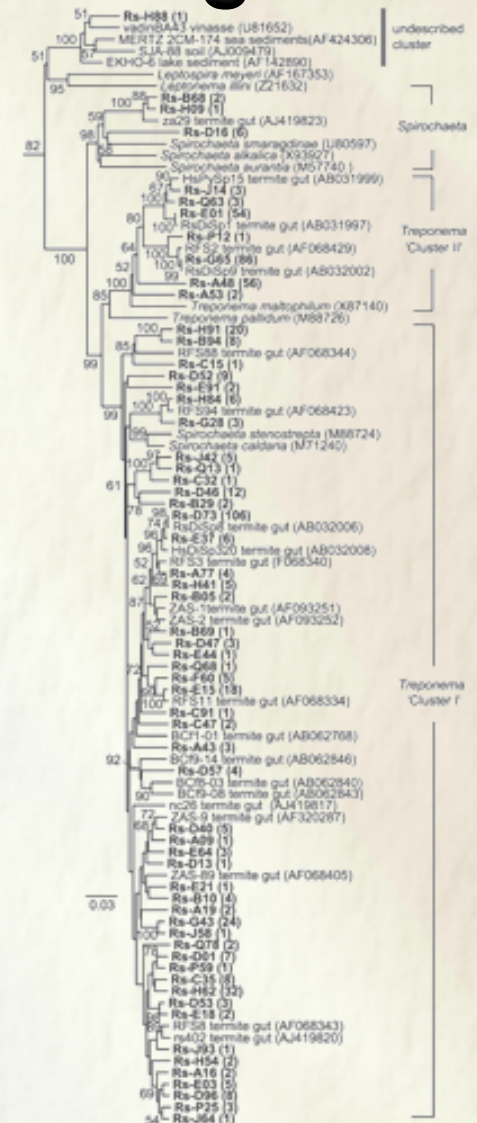
<sup>b</sup> Molecular Microbial Ecology Division, Bioscience Technology Center, RIKEN (The Institute of Physical and Chemical Research), Wako, Saitama 351-0198, Japan

<sup>c</sup> Division of Environmental Molecular Biology, Graduate School of Integrated Science, Yokohama City University, Kanagawa 230-0045, Japan

Received 5 August 2002; received in revised form 10 November 2002; accepted 29 December 2002

First published online 31 January 2003

**Fig. 3**





# ...and we've all seen papers like this...



FEMS Microbiology Ecology 44 (2003) 231–242

## Molecular analysis of bacterial microbiota in the gut of the termite *Reticulitermes speratus* (Isoptera; Rhinotermitidae)

Yuichi Hongoh<sup>a,b,\*</sup>, Moriya Ohkuma<sup>a,b</sup>, Toshiaki Kudo<sup>a,b,c</sup>

<sup>a</sup> International Cooperative Research Project, Japan Science and Technology Corporation (JST-ICORP), Saitama, Japan

<sup>b</sup> Molecular Microbial Ecology Division, Bioscience Technology Center, RIKEN (The Institute of Physical and Chemical Research), Wako, Saitama 351-0198, Japan

<sup>c</sup> Division of Environmental Molecular Biology, Graduate School of Integrated Science, Yokohama City University, Kanagawa 230-0045, Japan

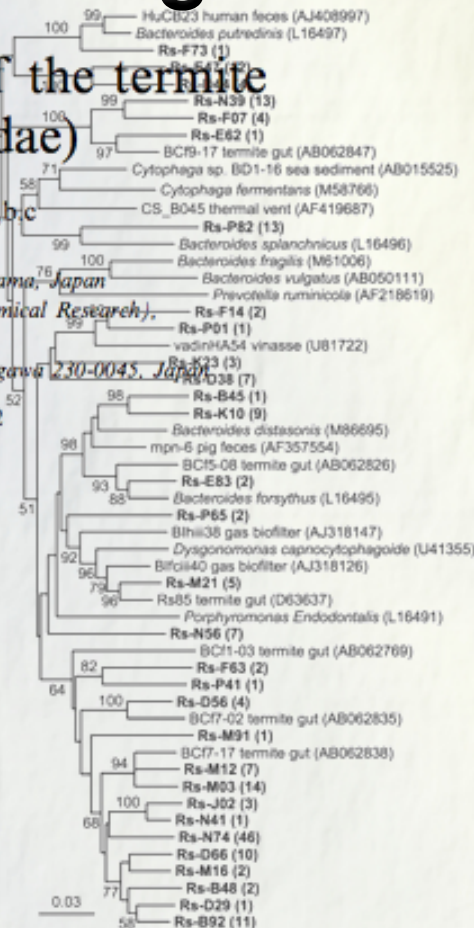
Received 5 August 2002; received in revised form 10 November 2002; accepted 29 December 2002

First published online 31 January 2003

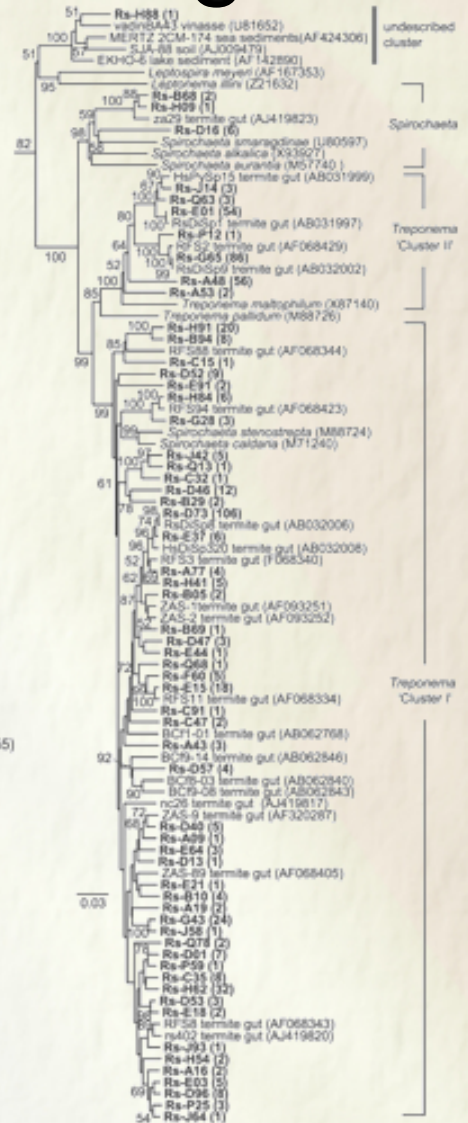
**FEMS**  
MICROBIOLOGY  
Ecology

[www.fems-microbiology.org](http://www.fems-microbiology.org)

**Fig. 4**



**Fig. 3**

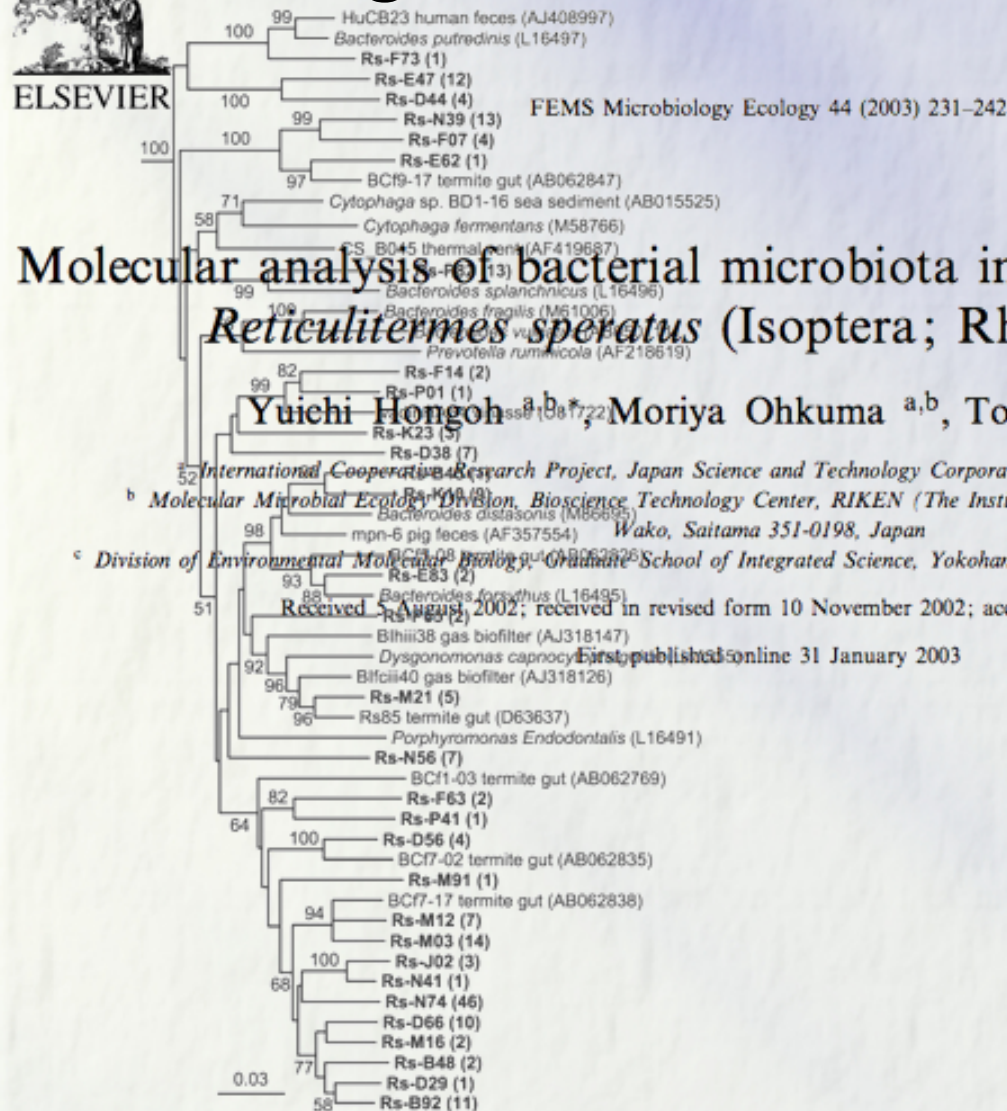




...and we've all seen papers like this....



Fig. 5



FEMS  
MICROBIOLOGY  
Ecology

www.fems-microbiology.org

Fig. 4

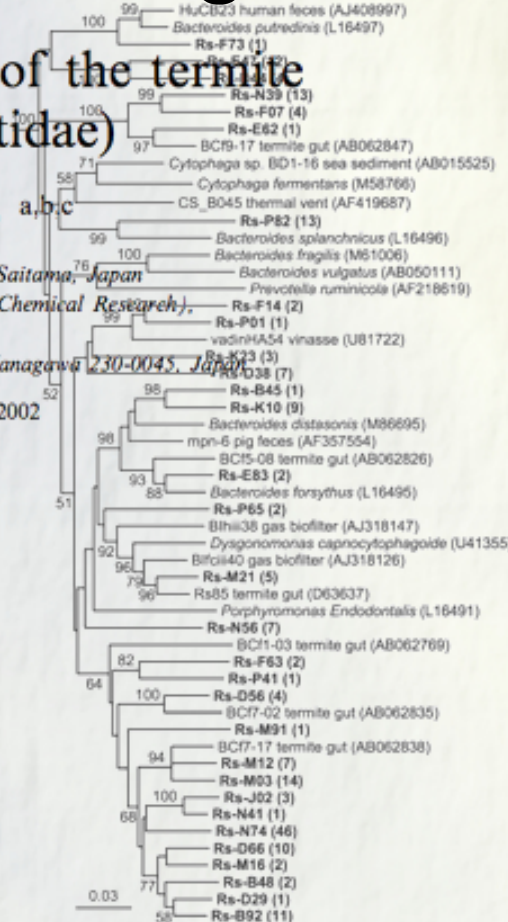


Fig. 3





...and we've all seen papers like this....



Fig. 5

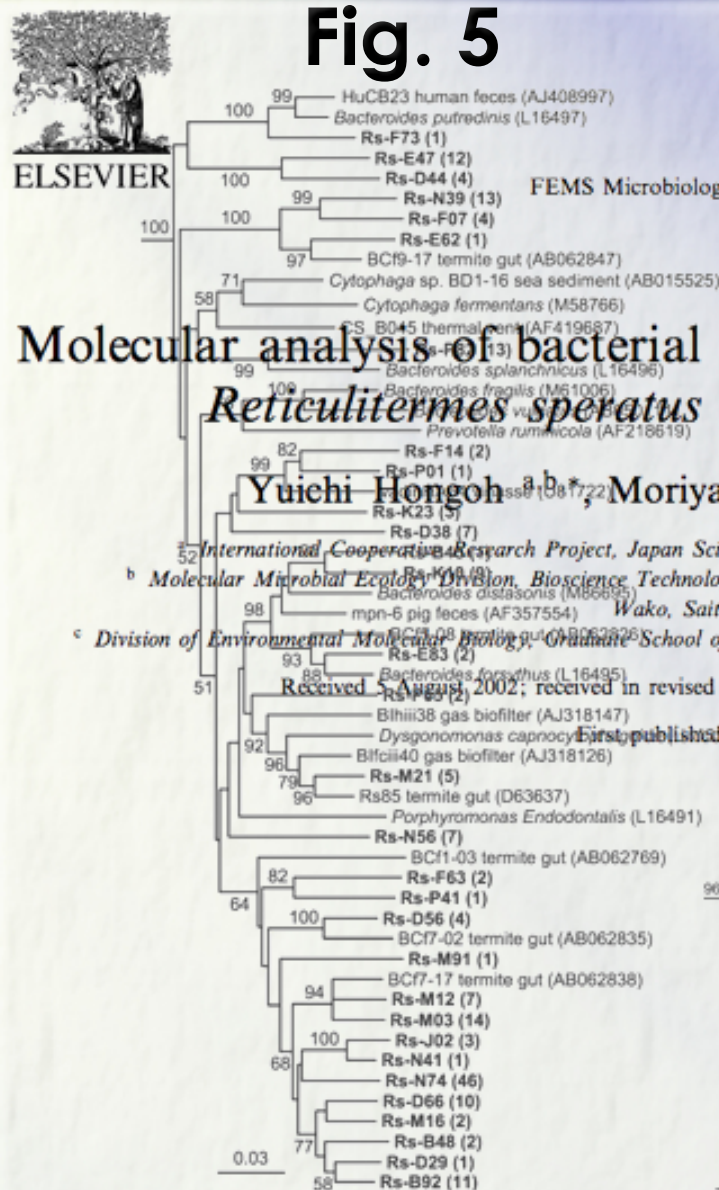


Fig. 6

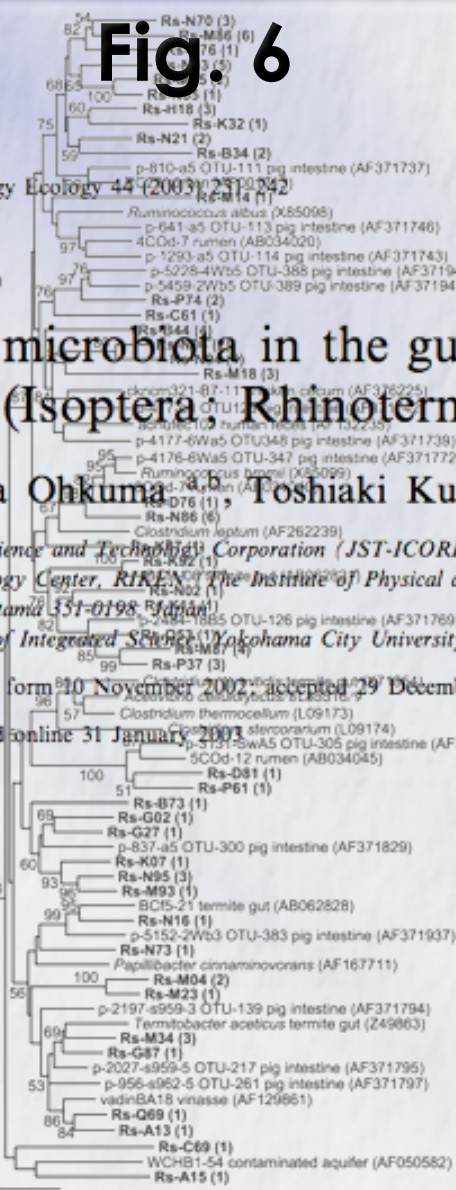


Fig. 4

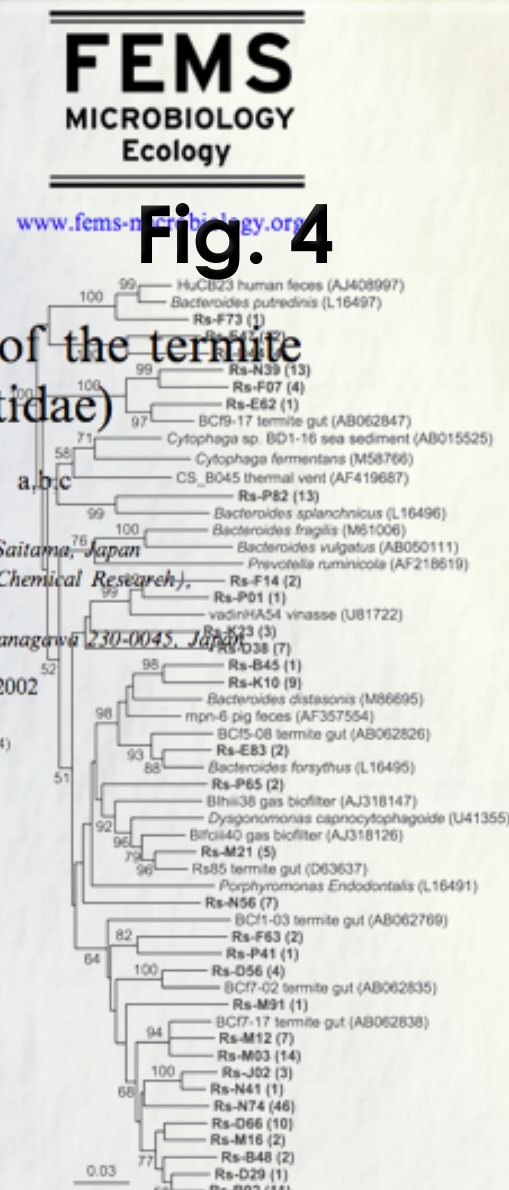


Fig. 3



Molecular analysis of bacterial microbiota in the gut of the termite *Reticulitermes speratus* (Isoptera: Rhinotermitidae)

Yuichi Hongoh<sup>a,b,\*</sup>, Moriya Ohkuma<sup>a,b</sup>, Toshiaki Kudo<sup>a,b,c</sup>

<sup>a</sup> International Cooperative Research Project, Japan Science and Technology Corporation (JST-ICORP), Saitama, Japan

<sup>b</sup> Molecular Microbial Ecology Division, Bioscience Technology Center, RIKEN, The Institute of Physical and Chemical Research, Wako, Saitama, Japan

<sup>c</sup> Division of Environmental Molecular Biology, Graduate School of Integrated Science, Yokohama City University, Kanagawa, Japan

Received 5 August 2002; received in revised form 10 November 2002; accepted 29 December 2002

First published online 31 January 2003



# Fig. 6

### Fig. 3

**FEMS**  
MICROBIOLOGY  
Ecology

**Fig. 4**

**Fig. 9**

# Fig. 12

# Fig. 10

Rs-H93 (1)

**vadin HB76 vinasse (U81754)**

*Citrobacterium capsulatum* D26  
p-2722-65A5 OTU-44 pg intestine (AF371680)  
Rs-461 (1)

www.pearsoned.com.au

ELSEVIER

35  
60  
91  
100

Rs-M603  
Rs-N35 (T)  
Rs-K75 (T)  
Rs-M72 (T)  
Desulfotomaculum  
Desulfotomaculum

**Molecule**

Phylogenetic tree showing the relationship between various bacterial strains. The tree is rooted at the top with 'Cam'. Below it, 'Helicob' is shown. A major branch leads to a cluster containing 'Rs-K54', 'BC16-0', 'Rhodoc', 'Azobir', 'Azarcu', 'Rs-87', 'Rs-A3', and 'Chrom'. Bootstrap values are indicated at the nodes: 100, 99, 76, 98, and 100.

Phylogenetic tree of the *rpoD* gene region. The tree shows relationships between various bacterial strains. Bootstrap values are indicated at the nodes. The strains are: D811-24, S10, D1435 (1), Pseudomonas, Escherichia, Rs-M74, Enterob, YNPRH718, Rs-H51 (1), Gluconacet, Rs-A44 (1), and M. luteus.

[illegible]

69  
94  
100  
76  
100  
68  
55  
93

*Ehrlichia*  
*Anaplasma*  
*Rickettsia*  
*Rickettsia*  
*Rickettsia*  
*Candidatus*  
*Holospirillum*  
*Rs-M*  
*Rs-116*  
*Rs-D39*  
*Symbiont*  
*Dechloromonas*

Phylogenetic tree of the Bcl2 gene family. The tree shows relationships between various Bcl2 family members. Bootstrap values are indicated at the nodes. A scale bar of 0.03 is shown at the bottom left.

- 94
  - 9
    - 100
      - 100B6
      - Bcl22 wa
    - Rs-M08
  - 67
    - Rs-H76
    - Rs-M75
  - 100
    - 88
      - Rs-B07
      - Rs-E51
      - Rs-D21

0.03

[illegible]

**Fig. 6**

**Fig. 8**

**Fig. 10**

**Fig. 12**

**Fig. 14**

**Fig. 16**

**Fig. 18**

**Fig. 20**

**Fig. 22**

**Fig. 24**

**Fig. 26**

**Fig. 28**

**Fig. 30**

**Fig. 32**

**Fig. 34**

**Fig. 36**

**Fig. 38**

**Fig. 40**

**Fig. 42**

**Fig. 44**

**Fig. 46**

**Fig. 48**

**Fig. 50**

**Fig. 52**

**Fig. 54**

**Fig. 56**

**Fig. 58**

**Fig. 60**

**Fig. 62**

**Fig. 64**

**Fig. 66**

**Fig. 68**

**Fig. 70**

**Fig. 72**

**Fig. 74**

**Fig. 76**

**Fig. 78**

**Fig. 80**

**Fig. 82**

**Fig. 84**

**Fig. 86**

**Fig. 88**

**Fig. 90**

**Fig. 92**

**Fig. 94**

**Fig. 96**

**Fig. 98**

**Fig. 100**

**Fig. 102**

**Fig. 104**

**Fig. 106**

**Fig. 108**

**Fig. 110**

**Fig. 112**

**Fig. 114**

**Fig. 116**

**Fig. 118**

**Fig. 120**

**Fig. 122**

**Fig. 124**

**Fig. 126**

**Fig. 128**

**Fig. 130**

**Fig. 132**

**Fig. 134**

**Fig. 136**

**Fig. 138**

**Fig. 140**

**Fig. 142**

**Fig. 144**

**Fig. 146**

**Fig. 148**

**Fig. 150**

**Fig. 152**

**Fig. 154**

**Fig. 156**

**Fig. 158**

**Fig. 160**

**Fig. 162**

**Fig. 164**

**Fig. 166**

**Fig. 168**

**Fig. 170**

**Fig. 172**

**Fig. 174**

**Fig. 176**

**Fig. 178**

**Fig. 180**

**Fig. 182**

**Fig. 184**

**Fig. 186**

**Fig. 188**

**Fig. 190**

**Fig. 192**

**Fig. 194**

**Fig. 196**

**Fig. 198**

**Fig. 200**

**Fig. 202**

**Fig. 204**

**Fig. 206**

**Fig. 208**

**Fig. 210**

**Fig. 212**

**Fig. 214**

**Fig. 216**

**Fig. 218**

**Fig. 220**

**Fig. 222**

**Fig. 224**

**Fig. 226**

**Fig. 228**

**Fig. 230**

**Fig. 232**

**Fig. 234**

**Fig. 236**

**Fig. 238**

**Fig. 240**

**Fig. 242**

**Fig. 244**

**Fig. 246**

**Fig. 248**

**Fig. 250**

**Fig. 252**

**Fig. 254**

**Fig. 256**

**Fig. 258**

**Fig. 260**

**Fig. 262**

**Fig. 264**

**Fig. 266**

**Fig. 268**

**Fig. 270**

**Fig. 272**

**Fig. 274**

**Fig. 276**

**Fig. 278**

**Fig. 280**

**Fig. 282**

**Fig. 284**

**Fig. 286**

**Fig. 288**

**Fig. 290**

**Fig. 292**

**Fig. 294**

**Fig. 296**

**Fig. 298**

**Fig. 300**

**Fig. 302**

**Fig. 304**

**Fig. 306**

**Fig. 308**

**Fig. 310**

**Fig. 312**

**Fig. 314**

**Fig. 316**

**Fig. 318**

**Fig. 320**

**Fig. 322**

**Fig. 324**

**Fig. 326**

**Fig. 328**

**Fig. 330**

**Fig. 332**

**Fig. 334**

**Fig. 336**

**Fig. 338**

**Fig. 340**

**Fig. 342**

**Fig. 344**

**Fig. 346**

**Fig. 348**

**Fig. 350**

**Fig. 352**

**Fig. 354**

**Fig. 356**

**Fig. 358**

**Fig. 360**

**Fig. 362**

**Fig. 364**

**Fig. 366**

**Fig. 368**

**Fig. 370**

**Fig. 372**

**Fig. 374**

**Fig. 376**

**Fig. 378**

**Fig. 380**

**Fig. 382**

**Fig. 384**

**Fig. 386**

**Fig. 388**

**Fig. 390**

**Fig. 392**

**Fig. 394**

**Fig. 396**

**Fig. 398**

**Fig. 400**

**Fig. 402**

**Fig. 404**

**Fig. 406**

**Fig. 408**

**Fig. 410**

**Fig. 412**

**Fig. 414**

**Fig. 416**

**Fig. 418**

**Fig. 420**

**Fig. 422**

**Fig. 424**

**Fig. 426**

**Fig. 428**

**Fig. 430**

**Fig. 432**

**Fig. 434**

**Fig. 436**

**Fig. 438**

**Fig. 440**

**Fig. 442**

**Fig. 444**

**Fig. 446**

**Fig. 448**

**Fig. 450**

**Fig. 452**

**Fig. 454**

**Fig. 456**

**Fig. 458**

**Fig. 460**

**Fig. 462**

**Fig. 464**

**Fig. 466**

**Fig. 468**

**Fig. 470**

**Fig. 472**

**Fig. 474**

**Fig. 476**

**Fig. 478**

**Fig. 480**

**Fig. 482**

**Fig. 484**

**Fig. 486**

**Fig. 488**

**Fig. 490**

**Fig. 492**

**Fig. 494**

**Fig. 496**

**Fig. 498**

**Fig. 500**

**Fig. 502**

**Fig. 504**

**Fig. 506**

**Fig. 508**

**Fig. 510**

**Fig. 512**

**Fig. 514**

**Fig. 516**

**Fig. 518**

**Fig. 520**

**Fig. 522**

**Fig. 524**

**Fig. 526**

**Fig. 528**

**Fig. 530**

**Fig. 532**

**Fig. 534**

**Fig. 536**

**Fig. 538**

**Fig. 540**

**Fig. 542**

**Fig. 544**

**Fig. 546**

**Fig. 548**

**Fig. 550**

**Fig. 552**

**Fig. 554**

**Fig. 556**

**Fig**

**Fig. 3**

**Fig. 4**

**Fig. 9**

**Fig. 12**

**Candidate division**

**Termites Group**

**unidentified bacterium**

**Acidobacteria group**

**Fig. 3**

**Fig. 4**

**Fig. 9**

**Fig. 12**

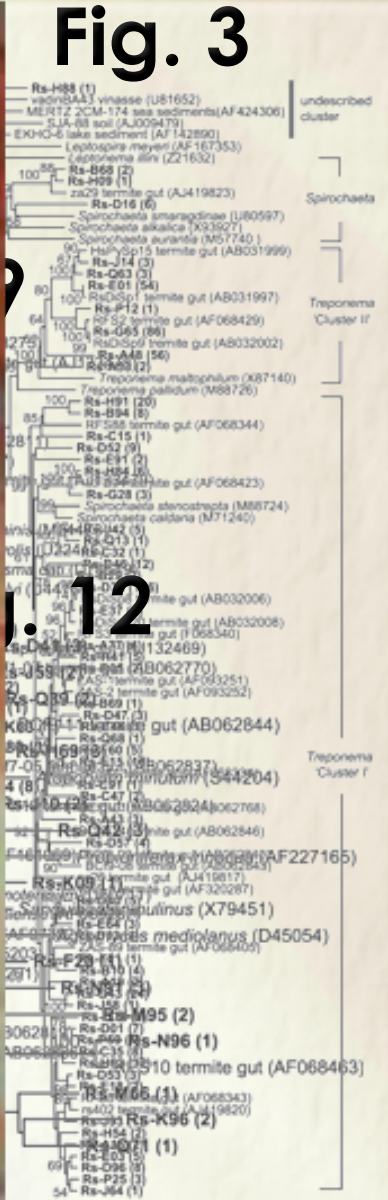
**Candidate division**

**Termites Group**

**unidentified bacterium**

**Acidobacteria group**



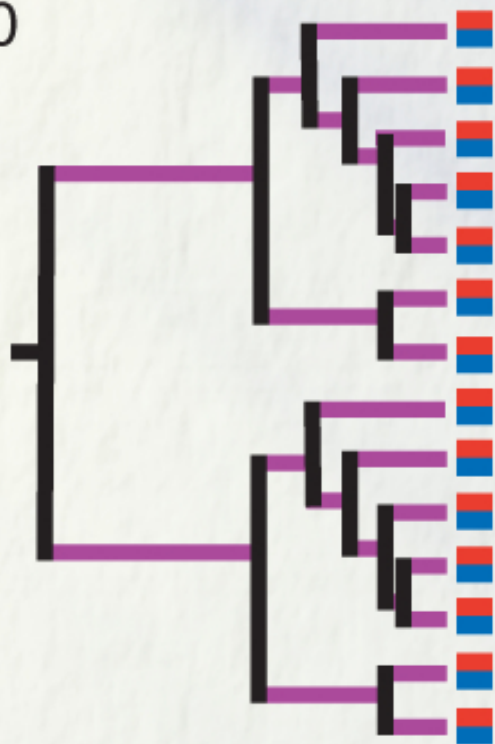




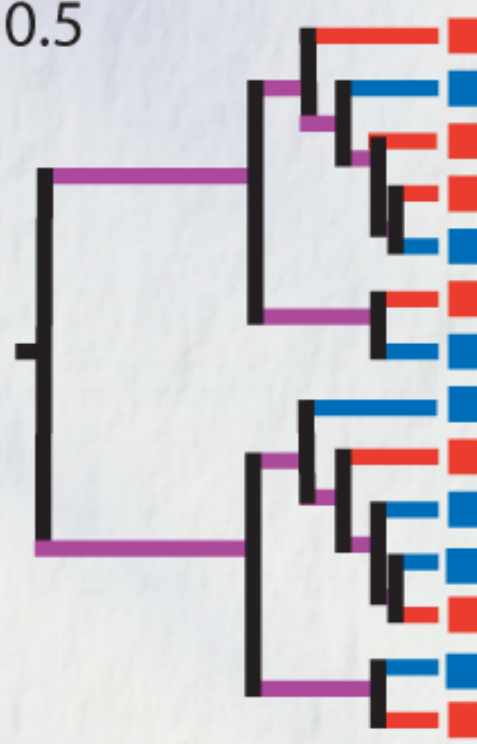
# Calculating a community distance metric

Distance = fraction of the total branch lengths that is unique to any particular environment

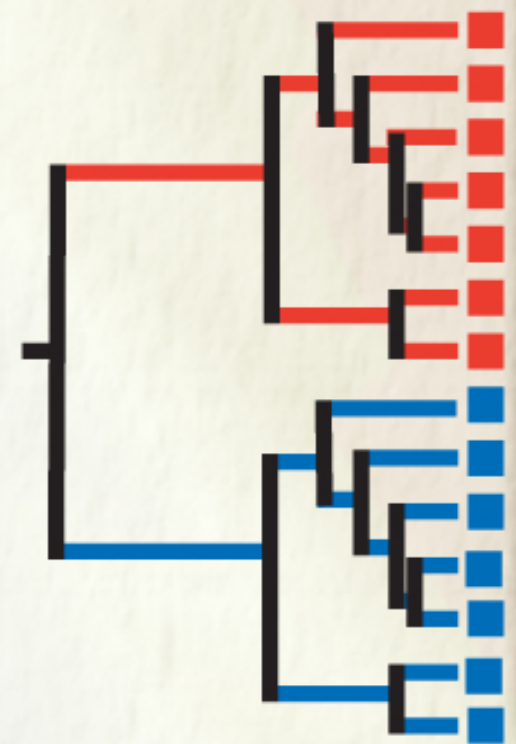
Identical communities:  
 $D = 0$



Related communities:  
 $D \sim 0.5$

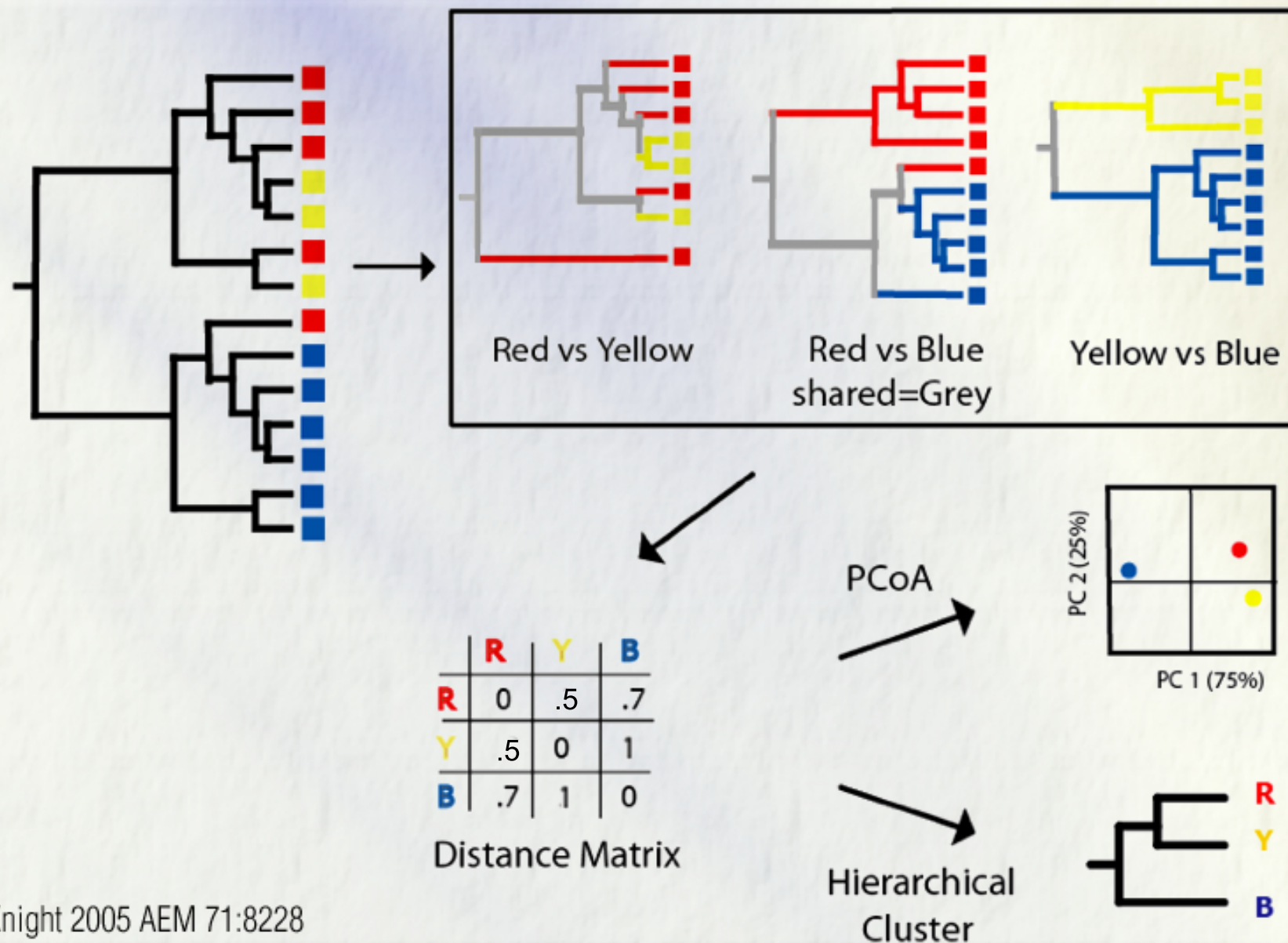


Unrelated communities:  
 $D = 1$



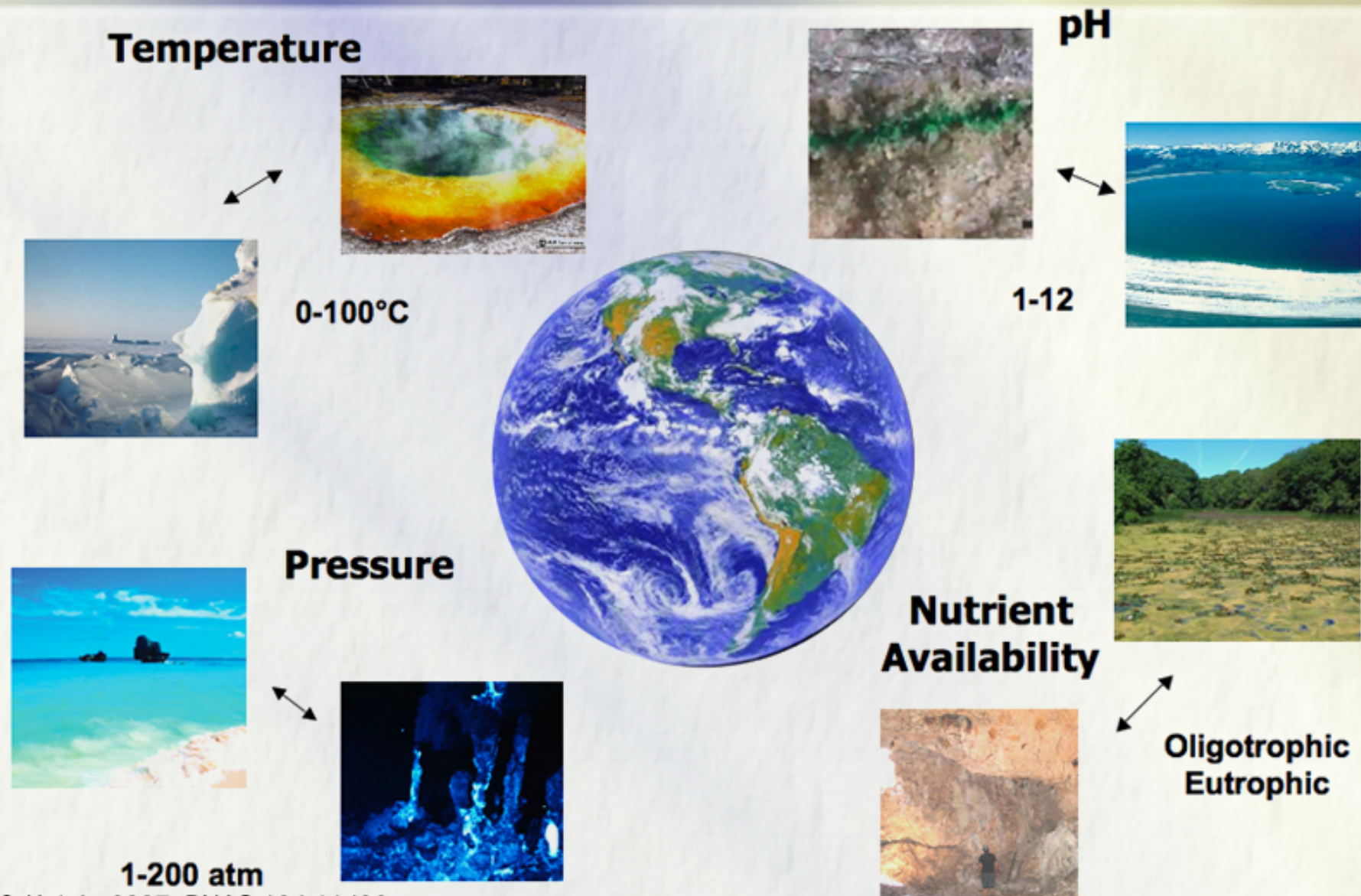


# Can relate many samples on one tree using phylogenetic info



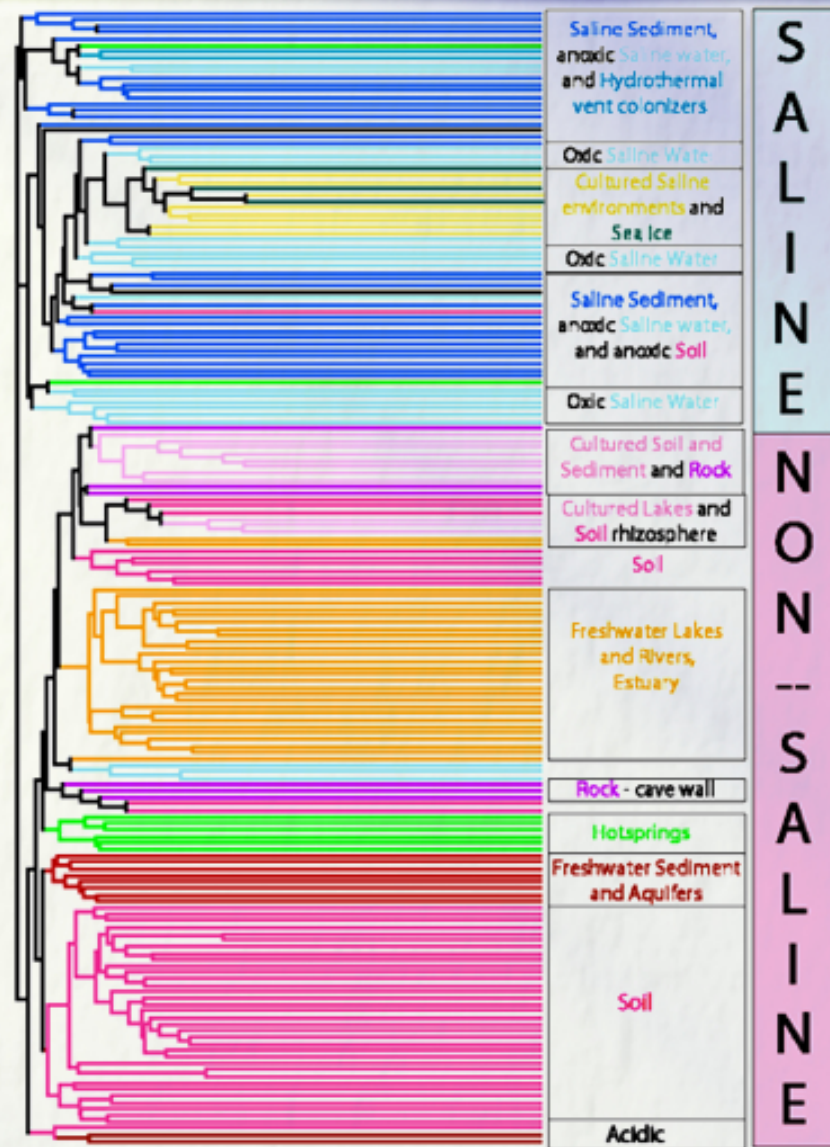


# Physical environments: 111 studies, 202 samples, 21,752 sequences

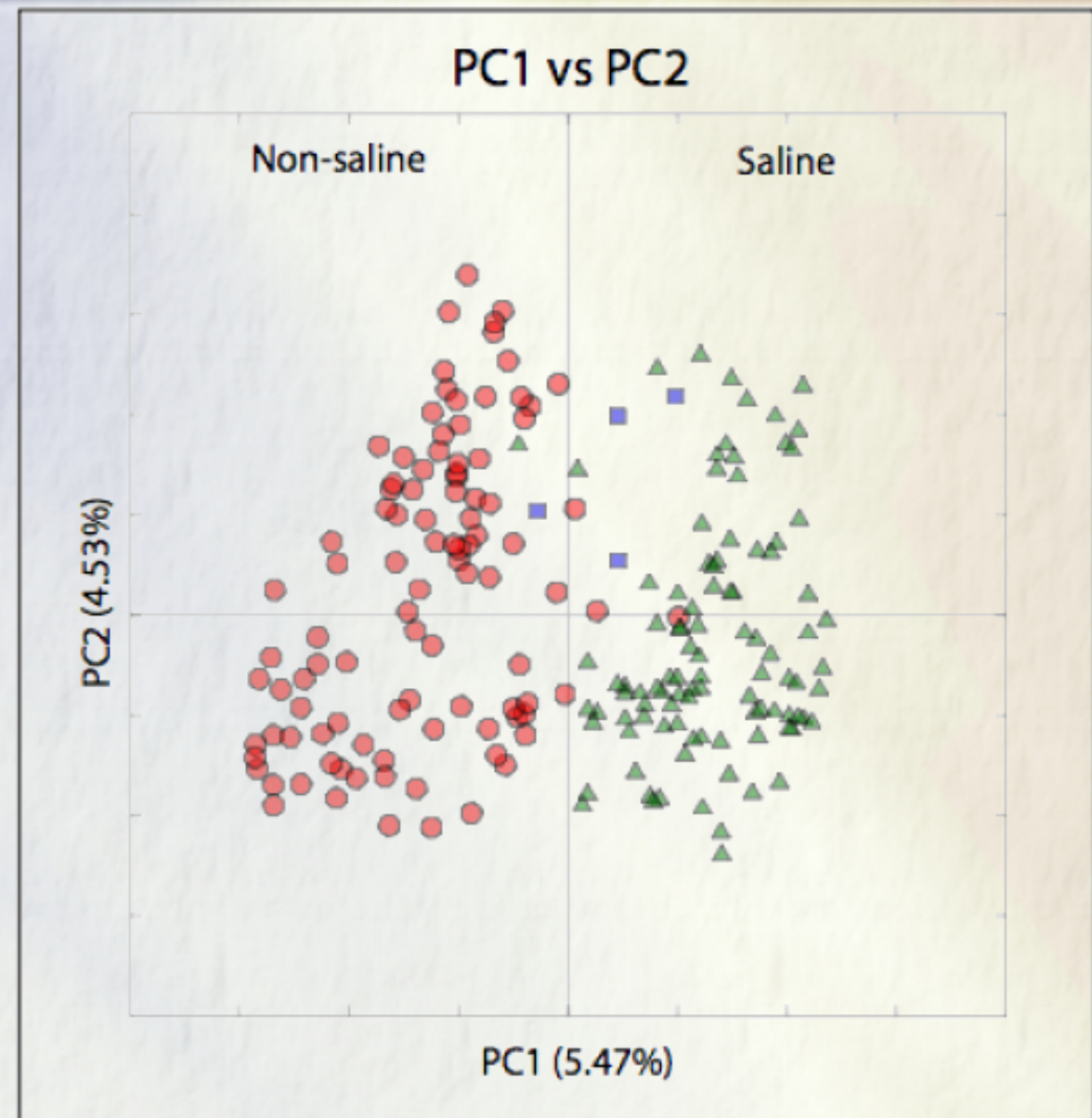




# Surprisingly, major split is saline/non-saline environment

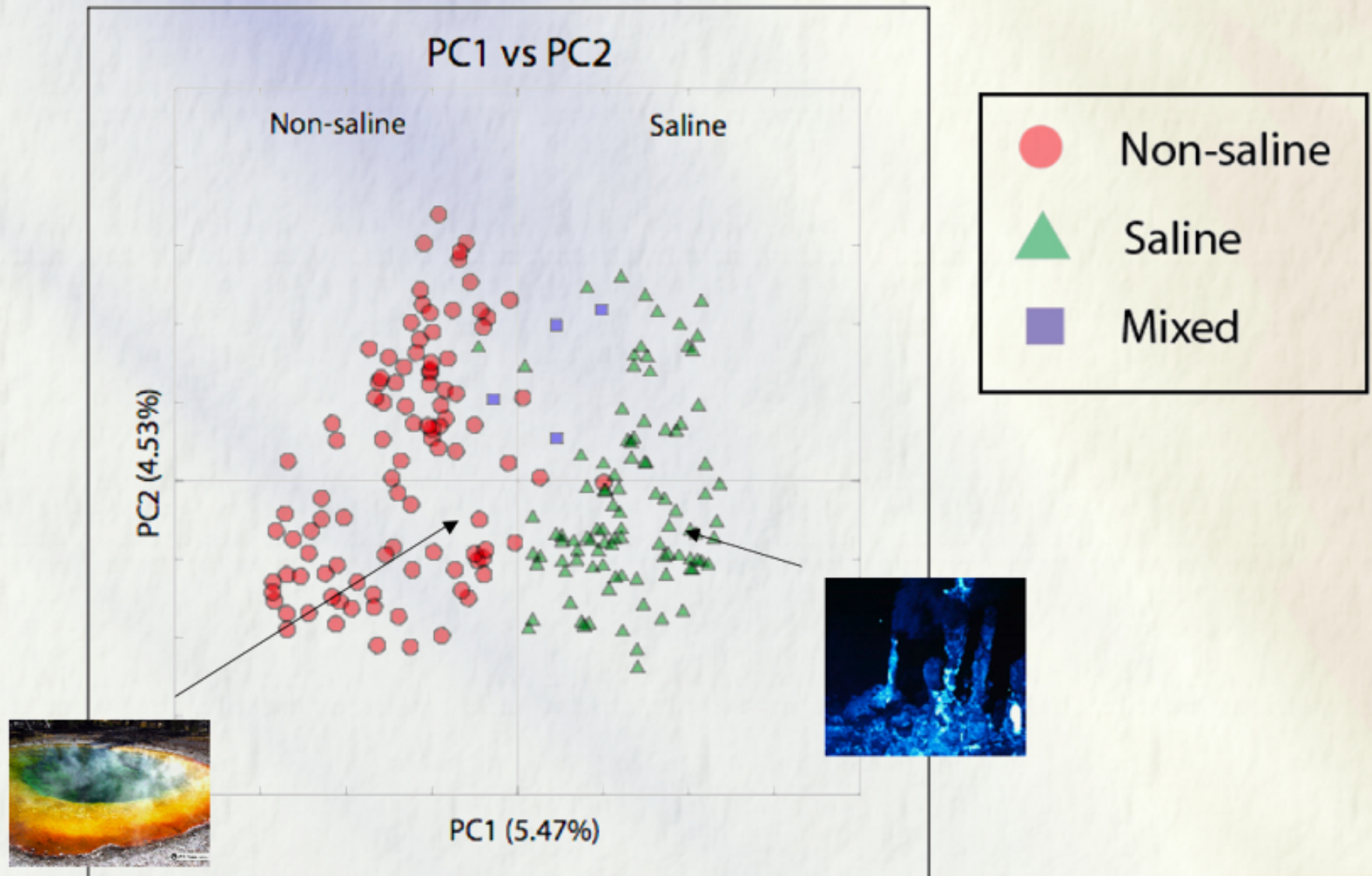


Lozupone & Knight 2007, PNAS 104:11436



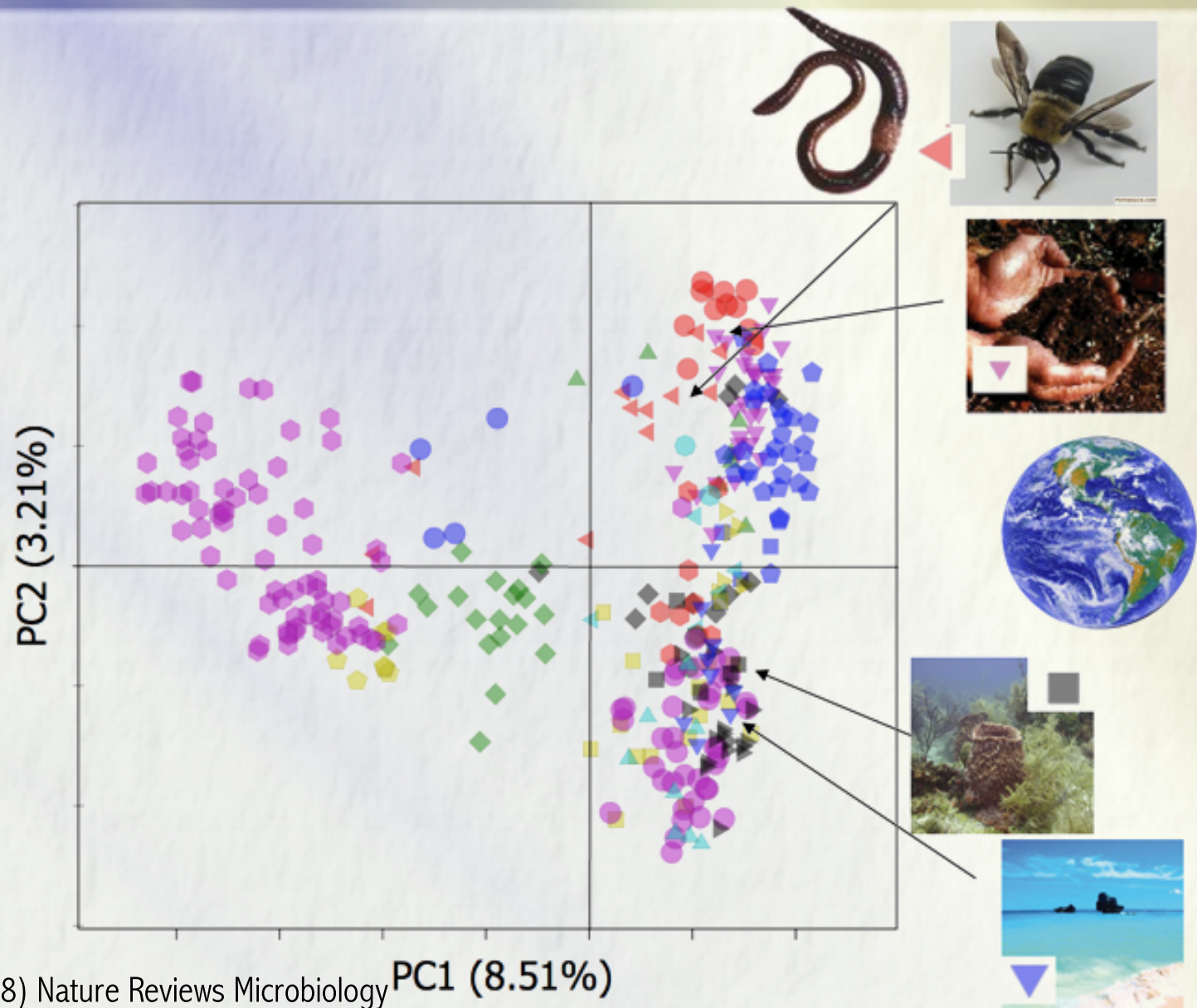


# Extreme environments are not outliers





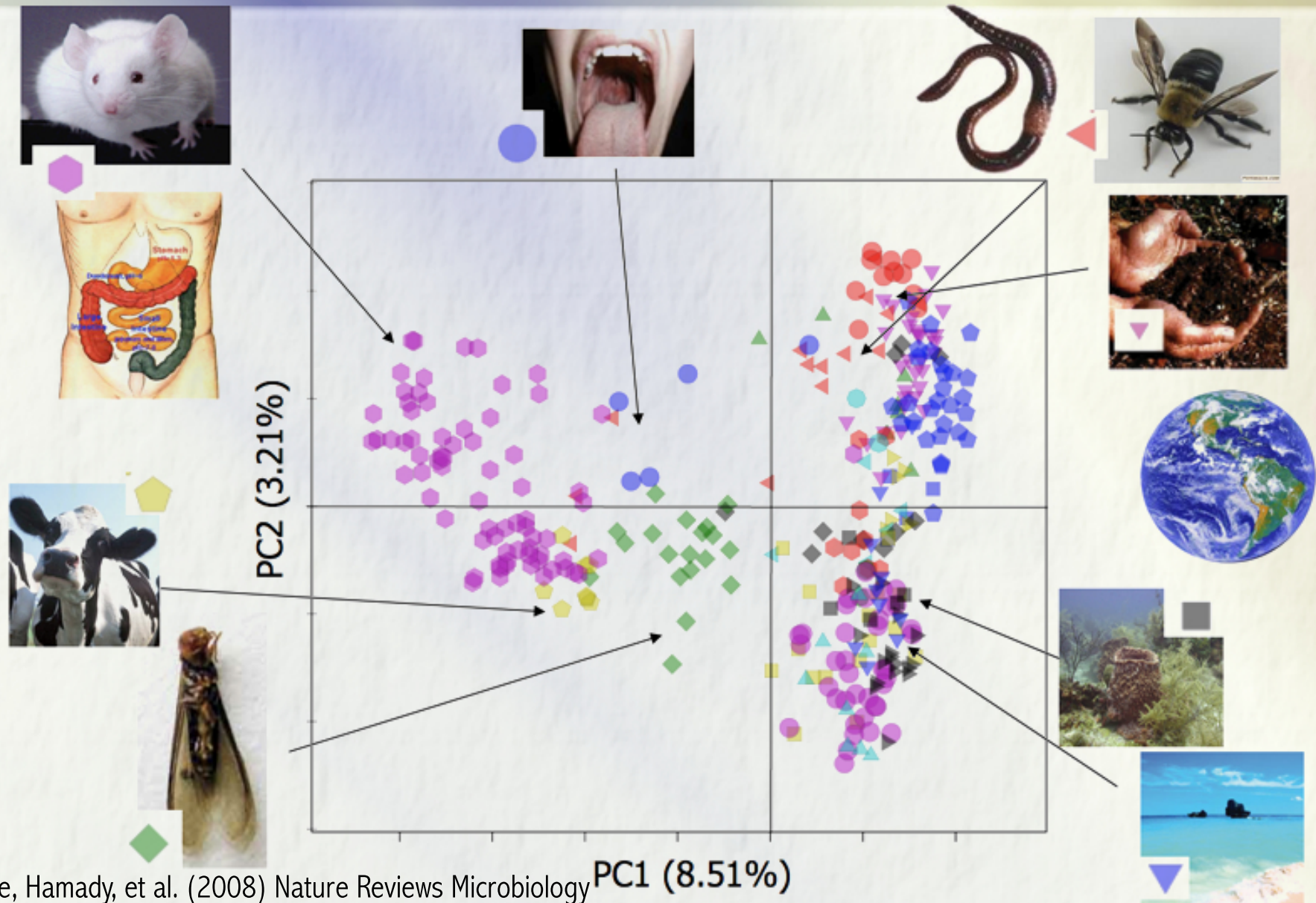
...but when we add samples we  
can find extreme environments...



Ley, Lozupone, Hamady, et al. (2008) Nature Reviews Microbiology



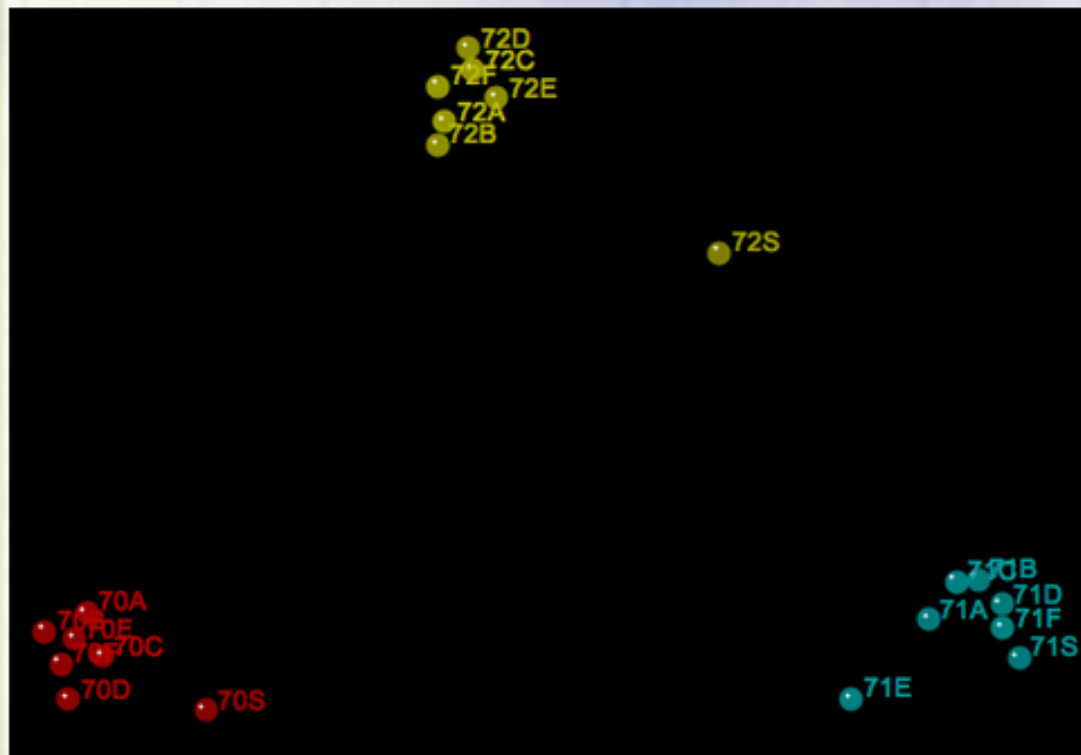
# ...most strikingly, in the vertebrate gut



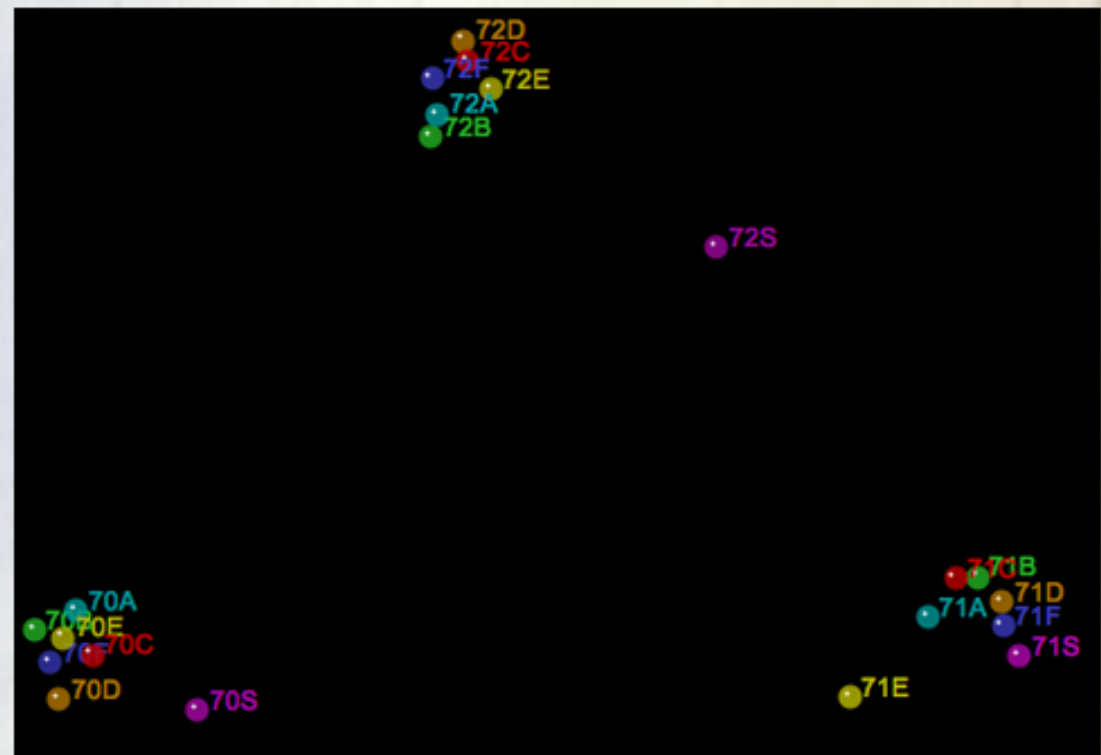
Ley, Lozupone, Hamady, et al. (2008) Nature Reviews Microbiology



# Are fecal samples OK for representing the gut?



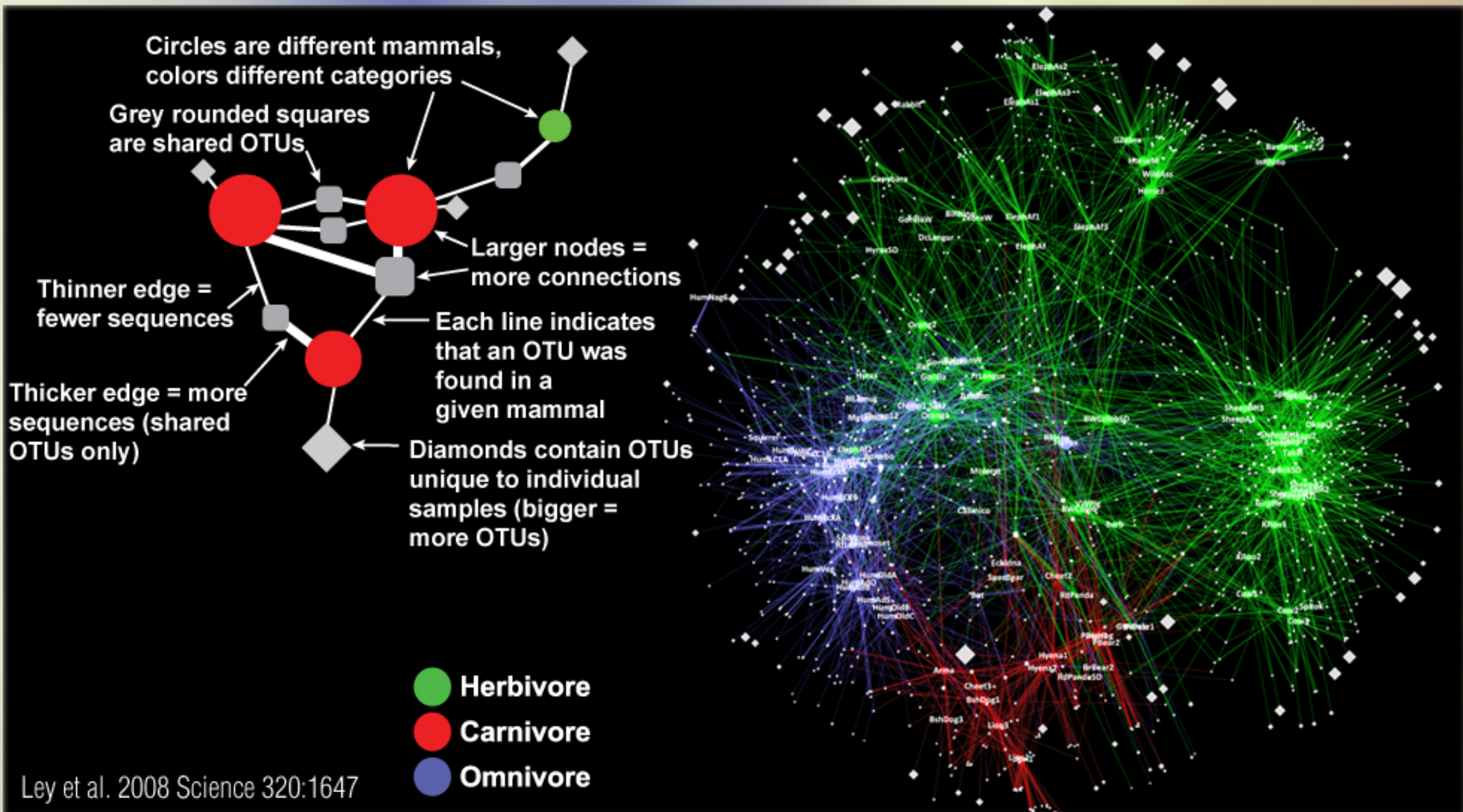
By subject



By site

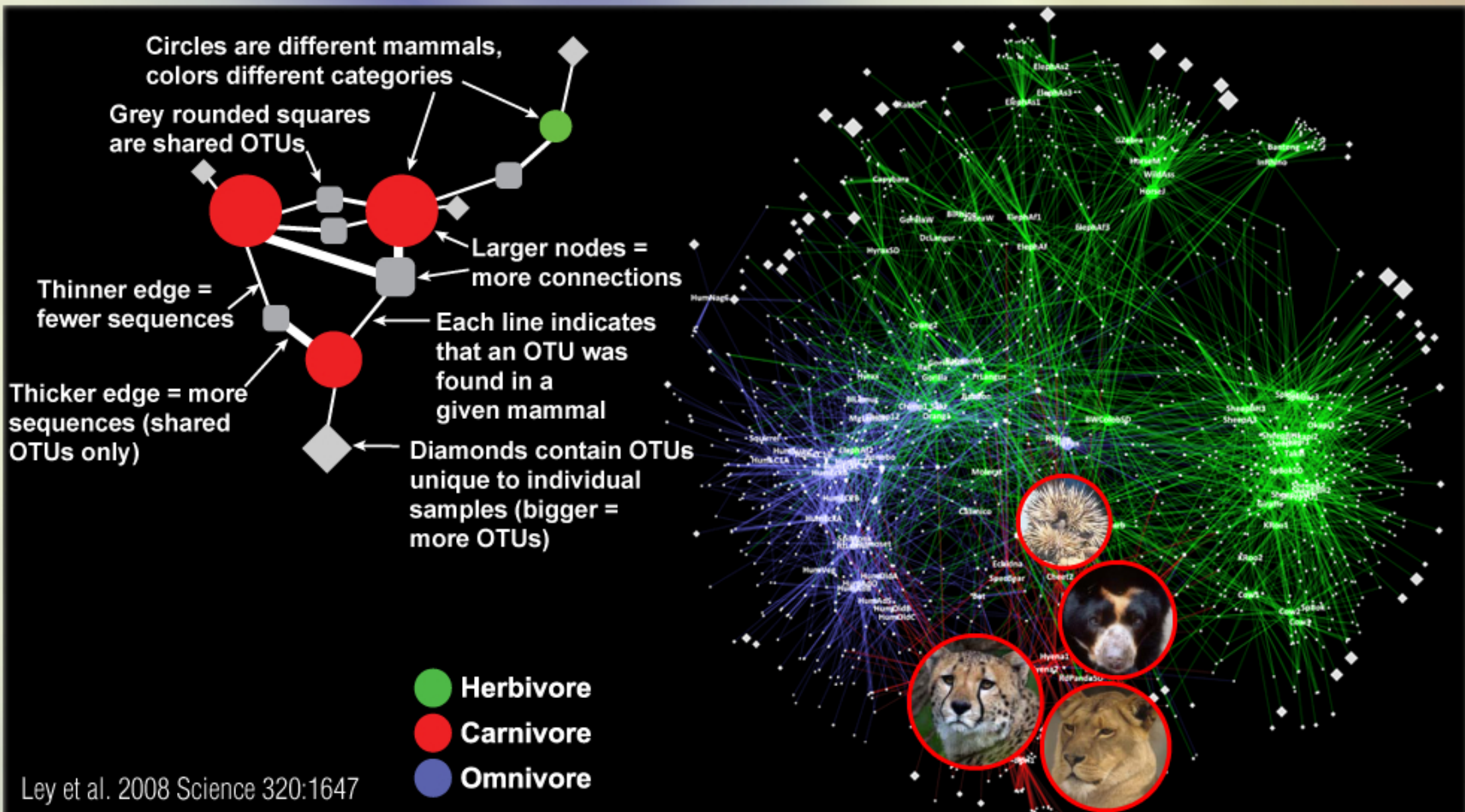


# Clear clustering by diet...



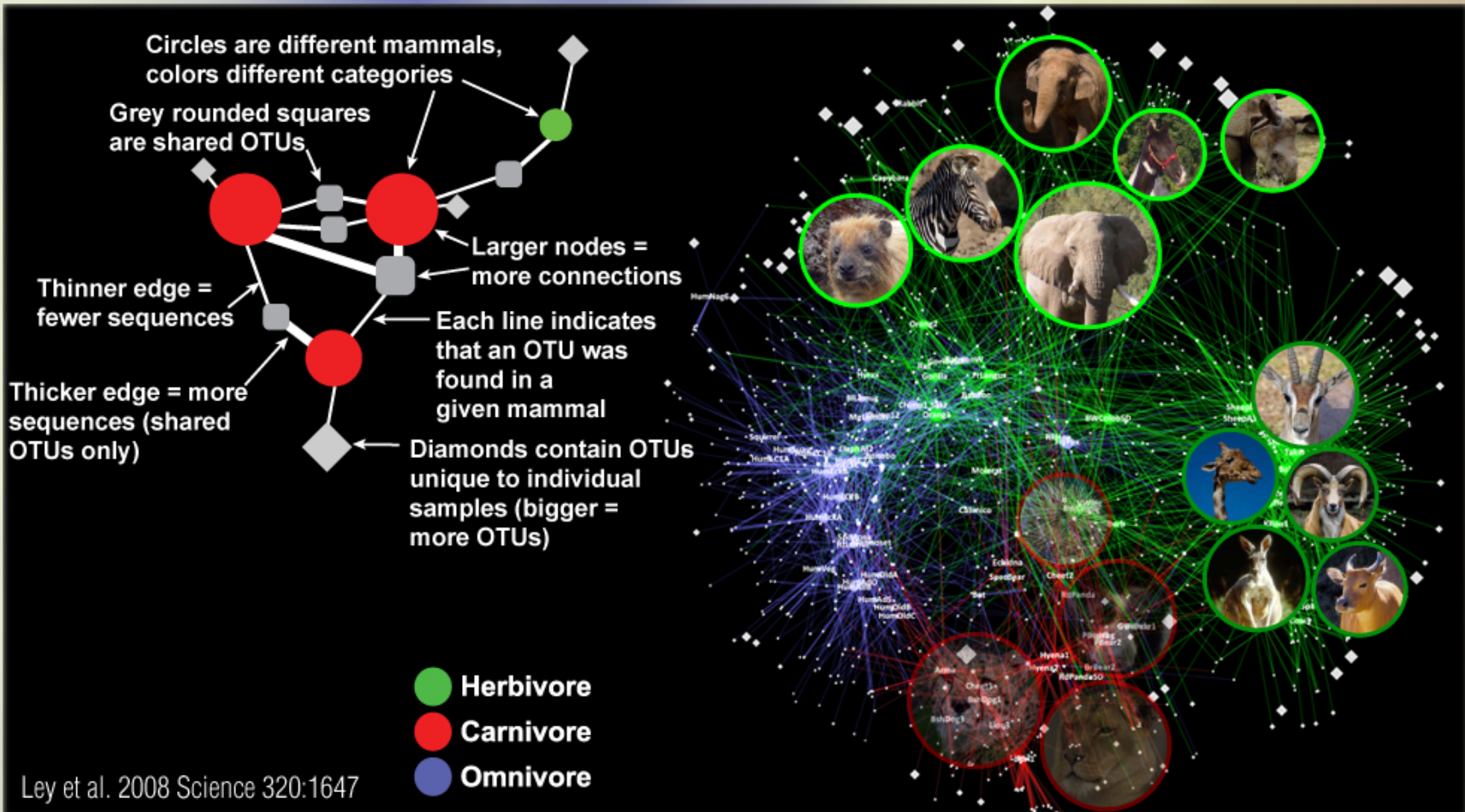


# ...with the carnivores clustering together...



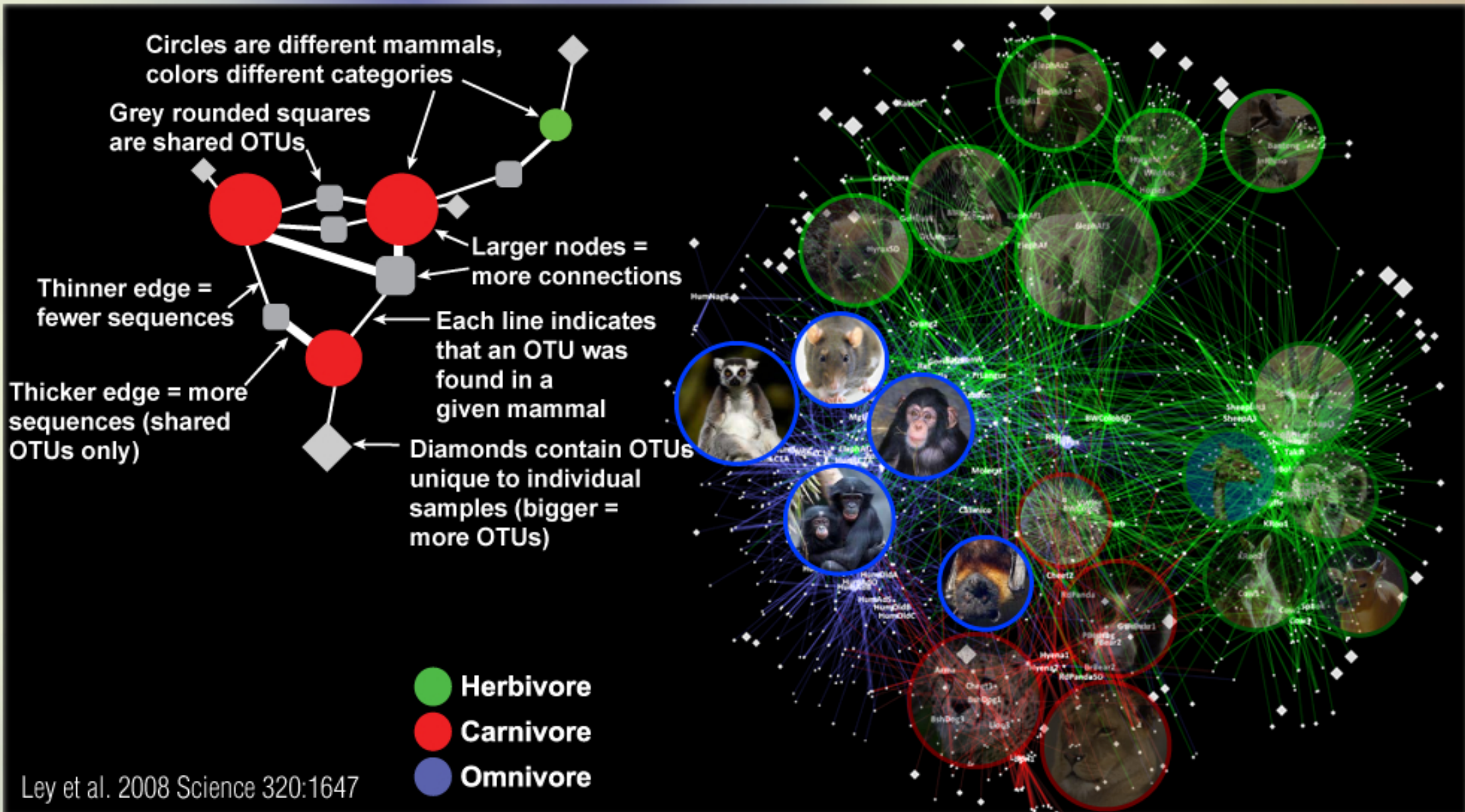


# ...two types of herbivores, the foregut and hindgut fermenters...



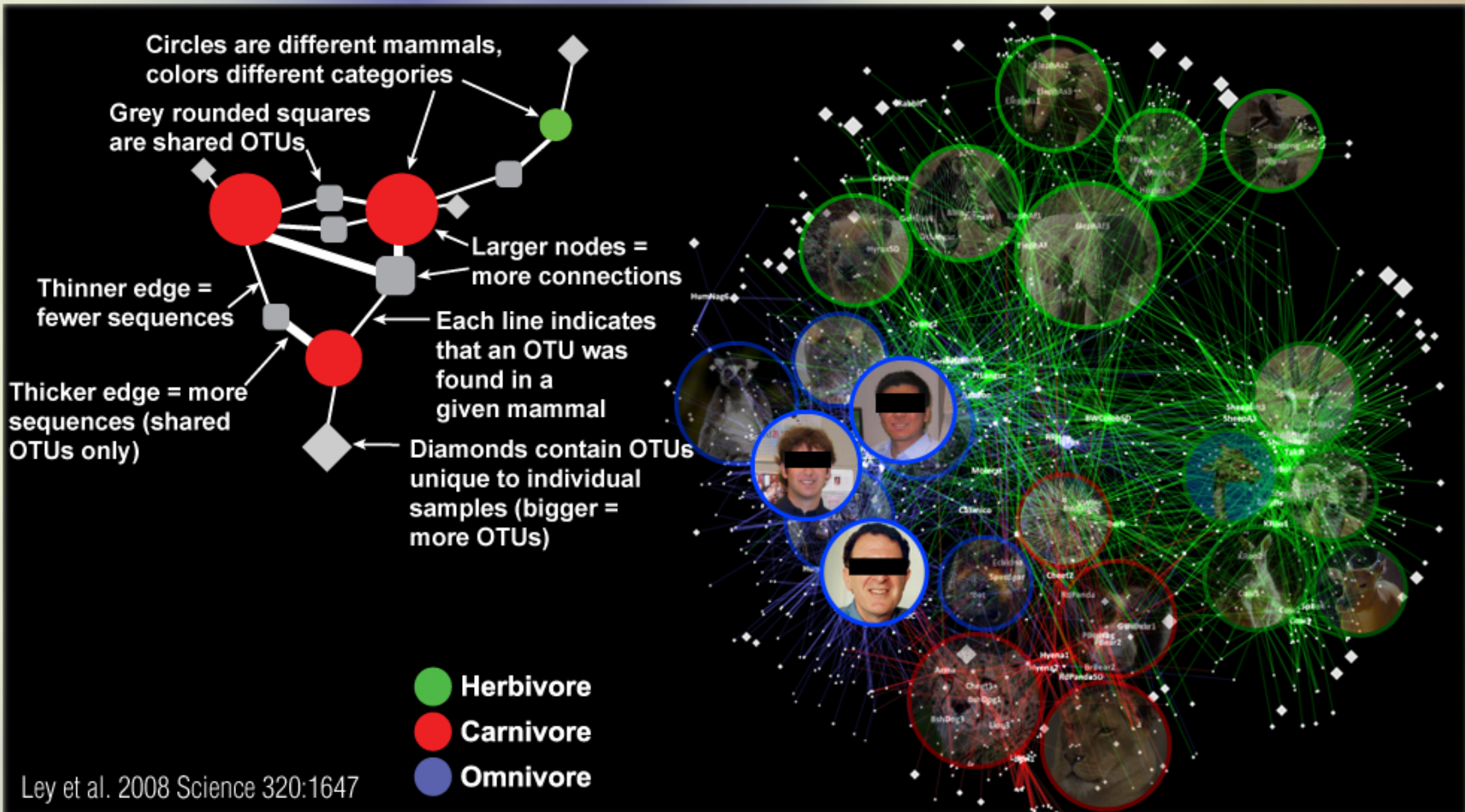


# ...and one group of omnivores...



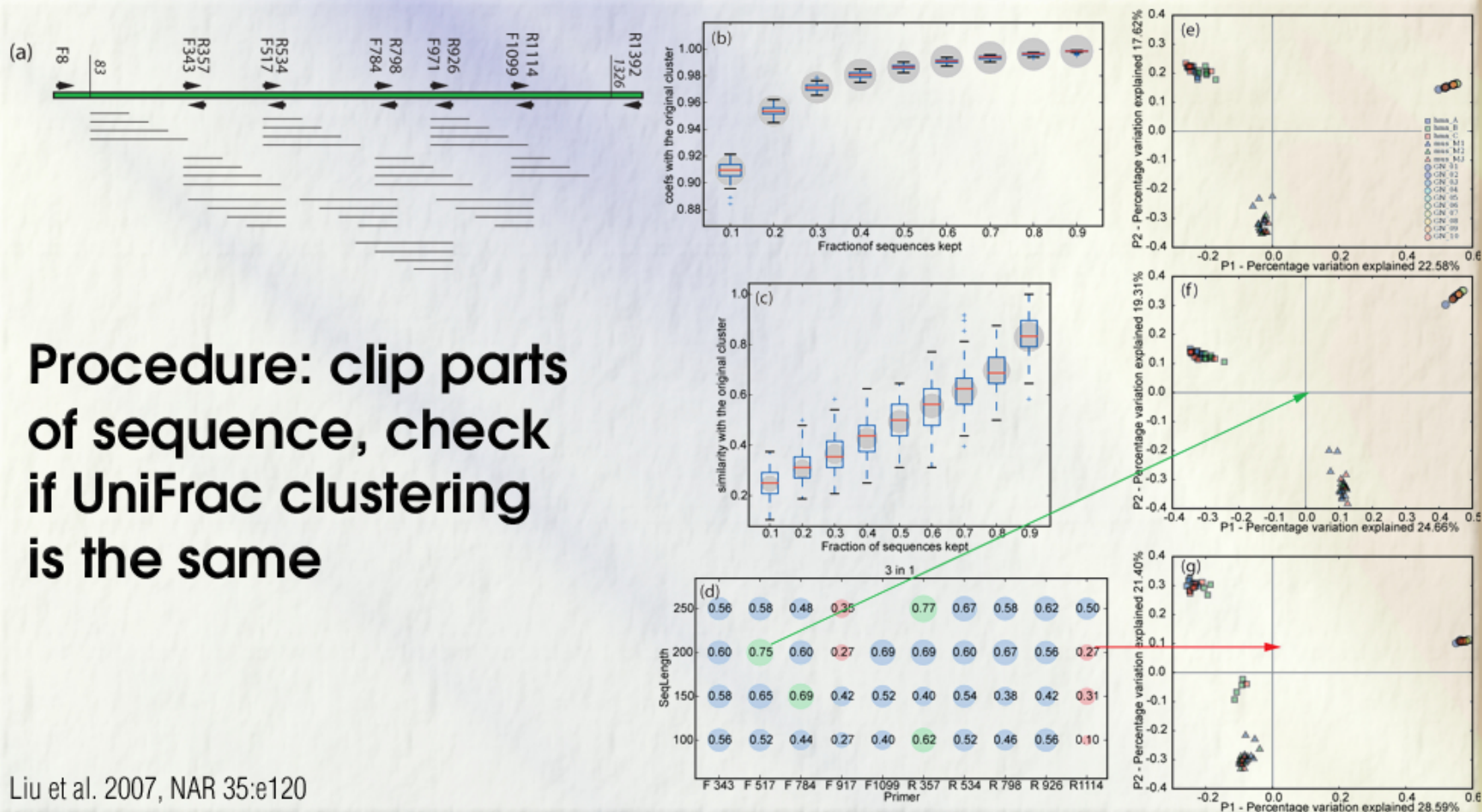


**...that includes humans, despite  
our diverse habitat and diet**



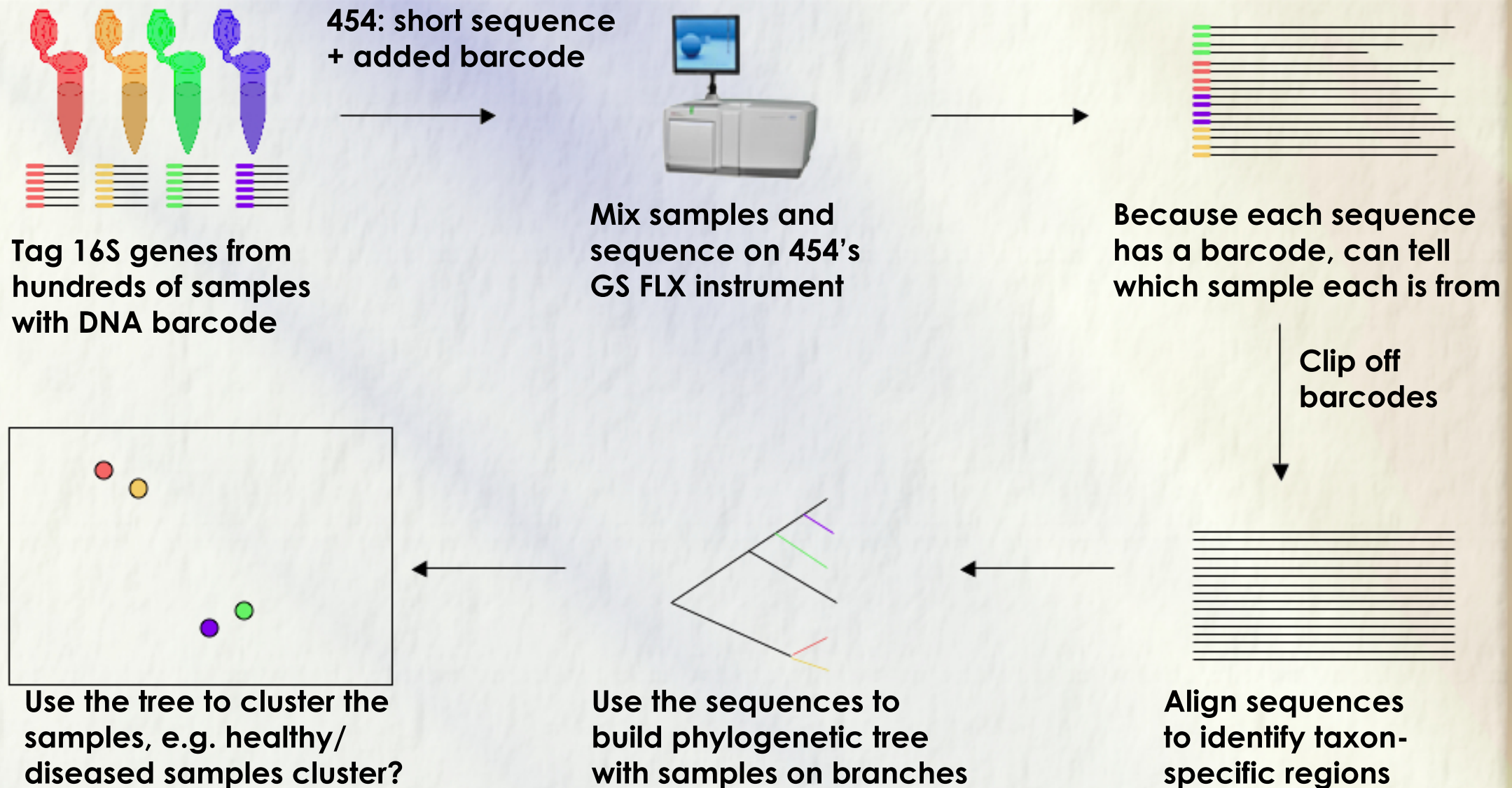


# Carefully chosen regions of 16S give same info as full sequence



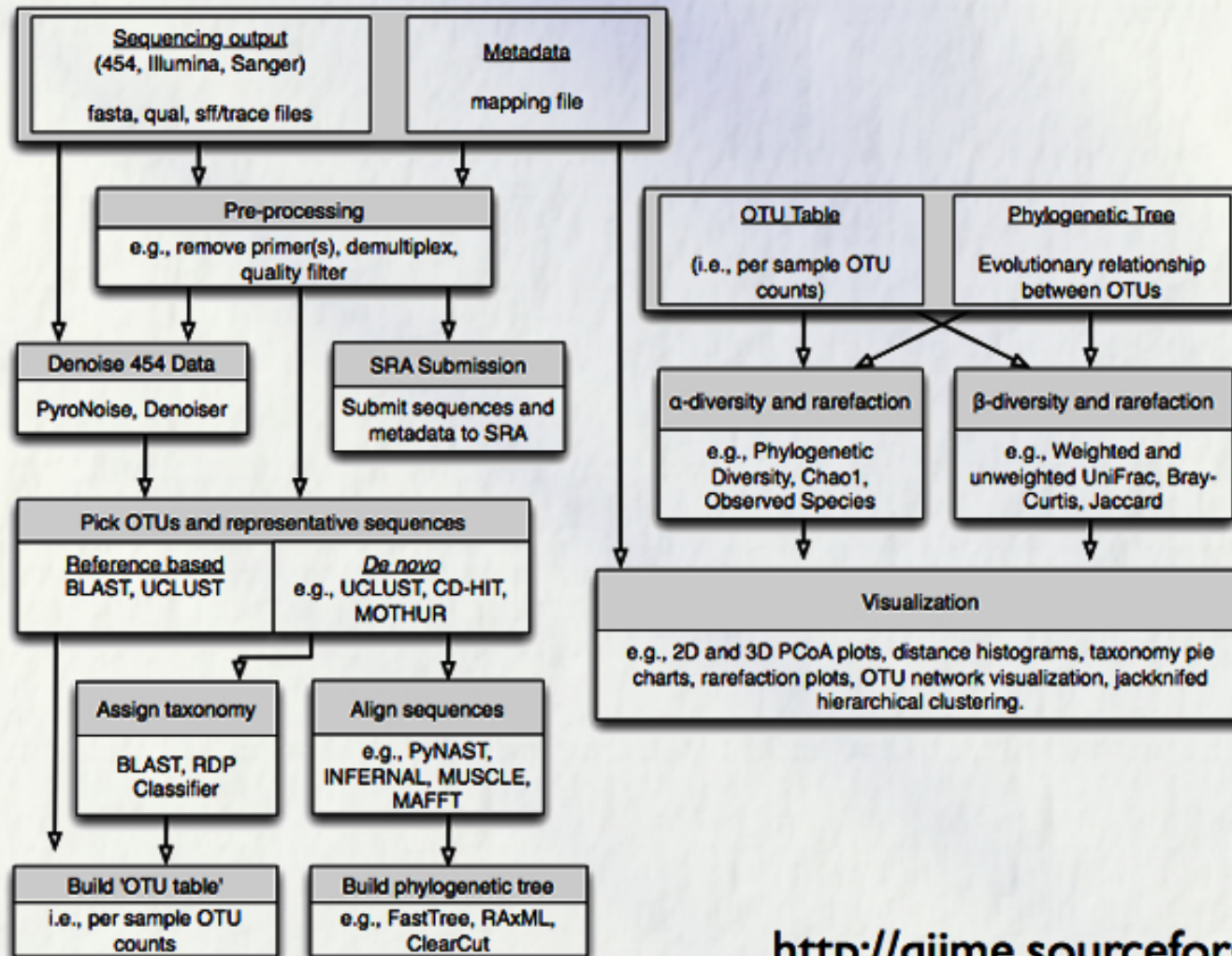


# QIIME: integrating analysis of hundreds of samples





# QIIME resolves technical challenges in dealing with next-gen data





# Our philosophy: wrap “name-brand” algorithms



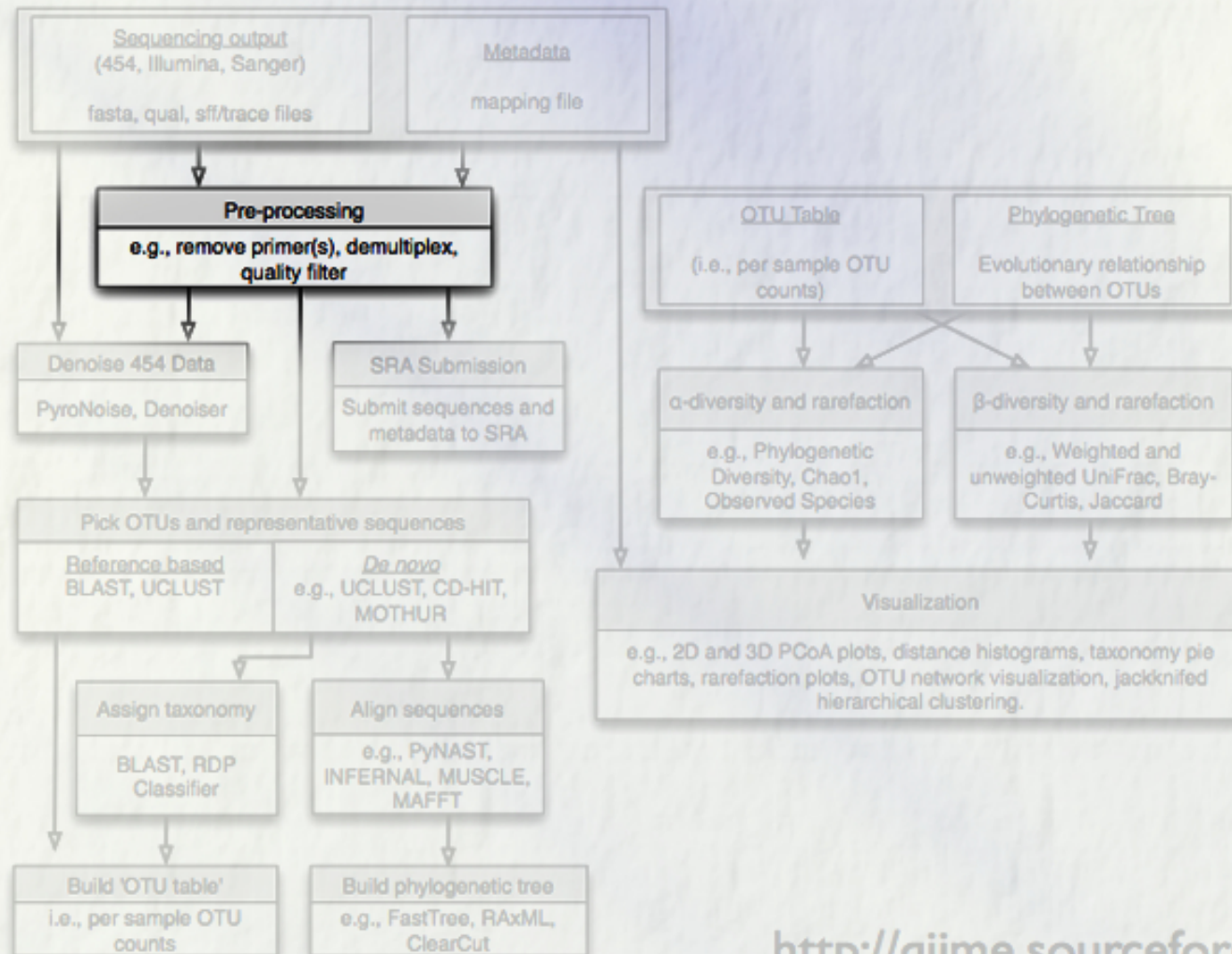
vs.



...with the generics, you don't know what you're getting!

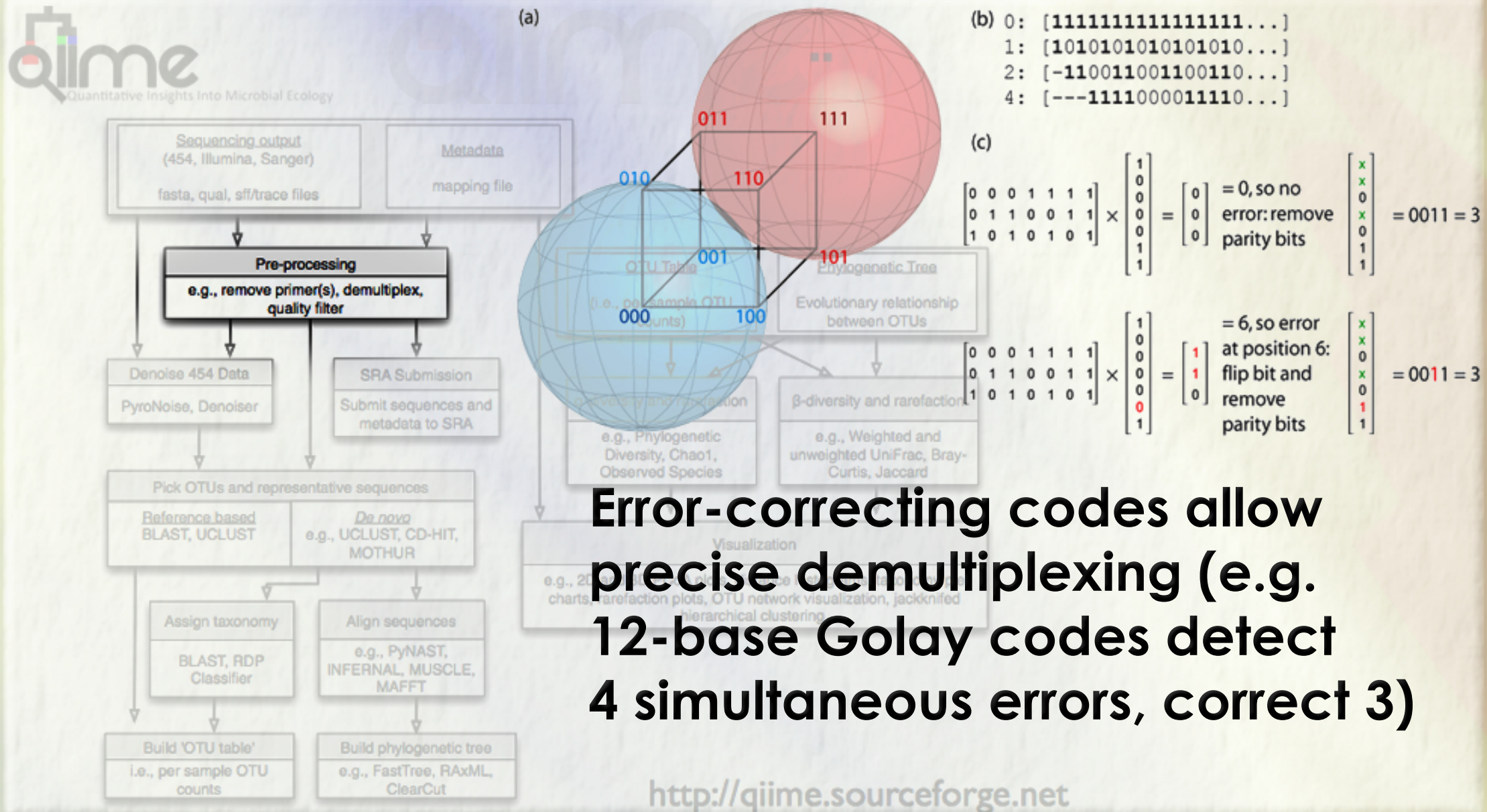


# QIIME resolves technical challenges in dealing with next-gen data



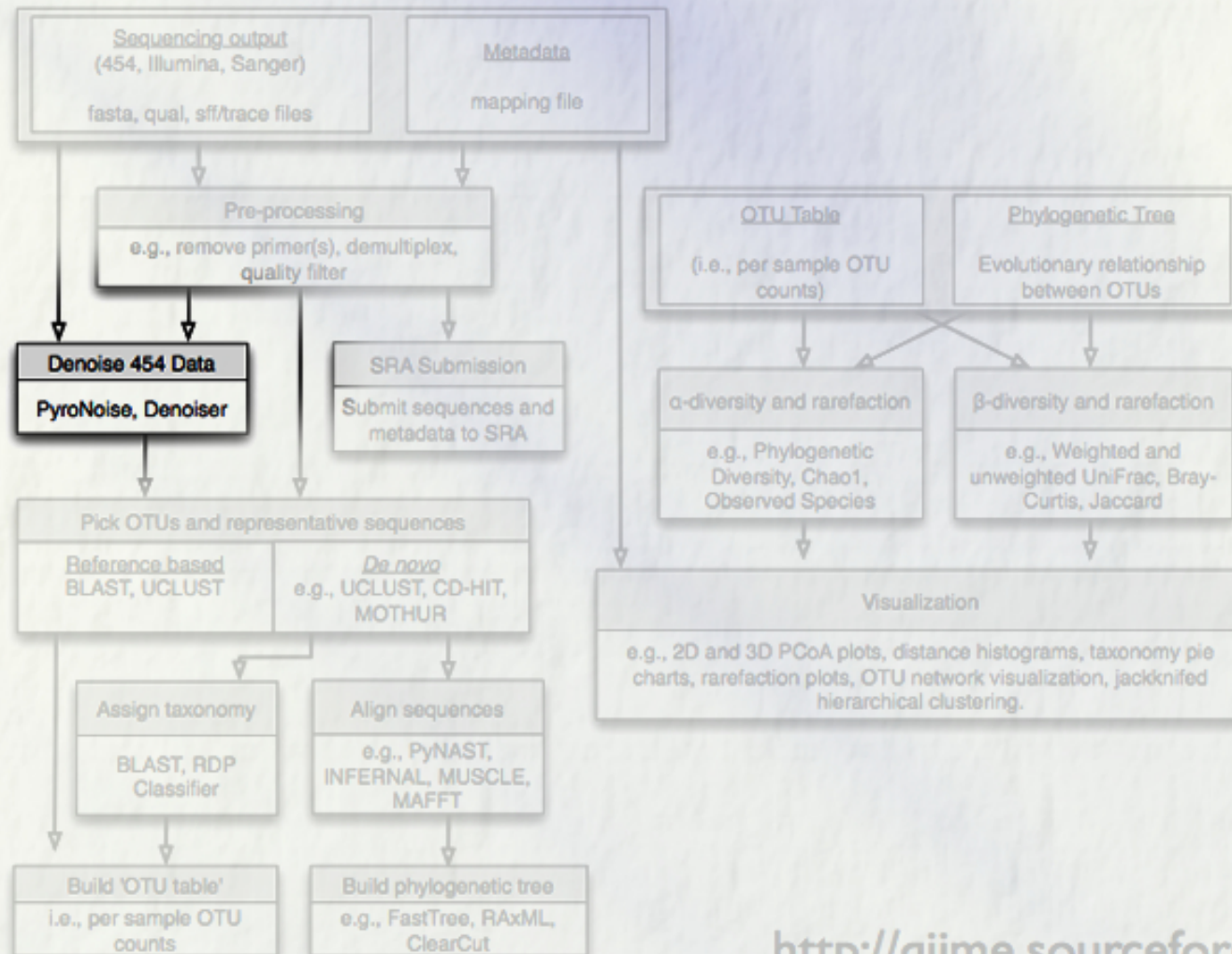


# QIIME resolves technical challenges in dealing with next-gen data





# QIIME resolves technical challenges in dealing with next-gen data



<http://qiime.sourceforge.net>



# QIIME resolves technical challenges in dealing with next-gen data

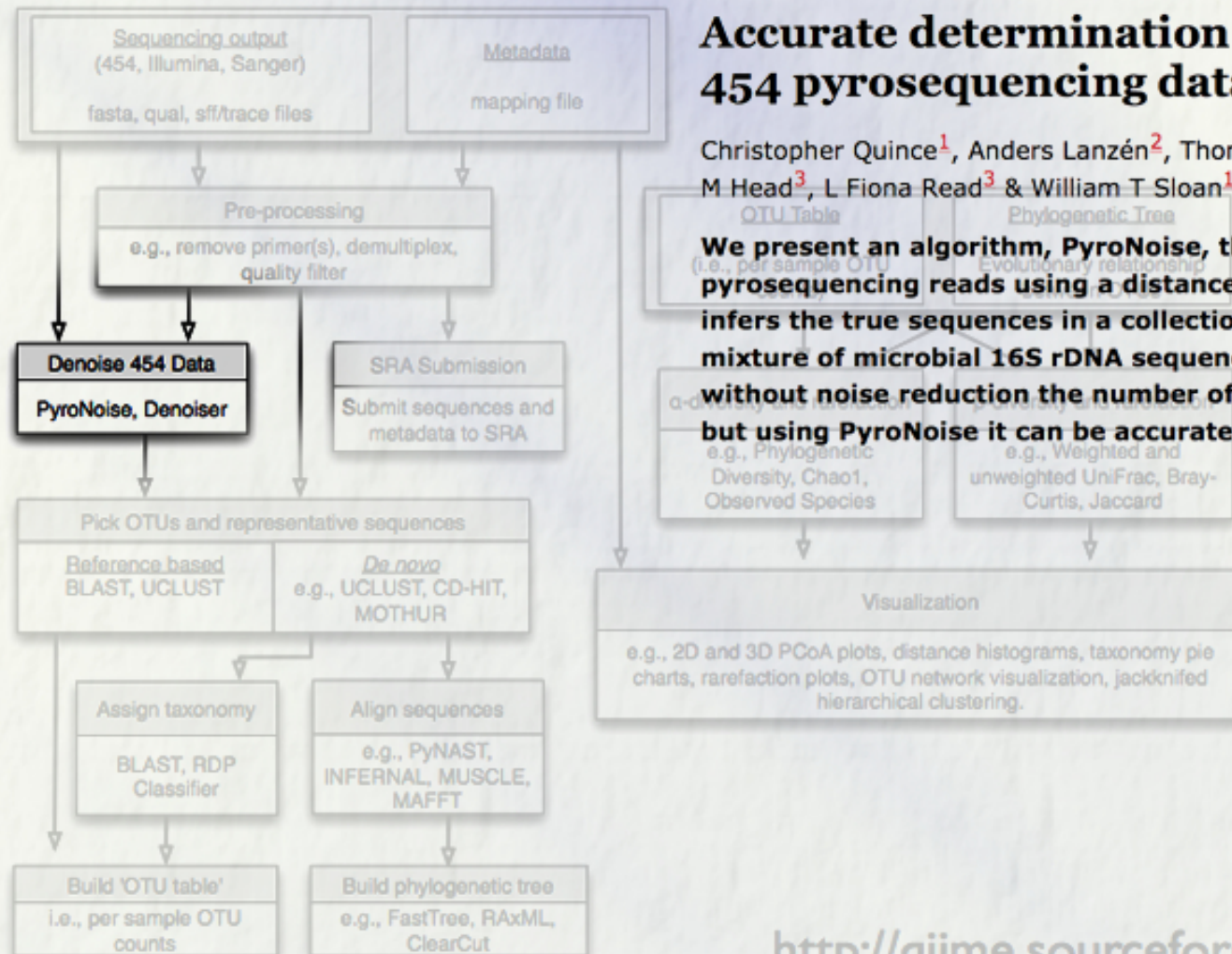


*Nature Methods* **6**, 639 - 641 (2009)  
Published online: 9 August 2009 | doi:10.1038/nmeth.1361

## Accurate determination of microbial diversity from 454 pyrosequencing data

Christopher Quince<sup>1</sup>, Anders Lanzén<sup>2</sup>, Thomas P Curtis<sup>3</sup>, Russell J Davenport<sup>3</sup>, Neil Hall<sup>4</sup>, Ian M Head<sup>3</sup>, L Fiona Read<sup>3</sup> & William T Sloan<sup>1</sup>

**We present an algorithm, PyroNoise, that clusters the flowgrams of 454 pyrosequencing reads using a distance measure that models sequencing noise. This infers the true sequences in a collection of amplicons. We pyrosequenced a known mixture of microbial 16S rDNA sequences extracted from a lake and found that without noise reduction the number of operational taxonomic units is overestimated but using PyroNoise it can be accurately calculated.**



<http://qiime.sourceforge.net>



# QIIME resolves technical challenges in dealing with next-gen data



## CORRESPONDENCE

*Nature Methods* **7**, 668 - 669 (2010)  
doi:10.1038/nmeth0910-668b

Rapidly denoising pyrosequencing amplicon reads by exploiting rank-abundance distributions

Jens Reeder<sup>1</sup> & Rob Knight<sup>1,2</sup>

1. Department of Chemistry & Biochemistry, University of Colorado at Boulder, Boulder, Colorado, USA.

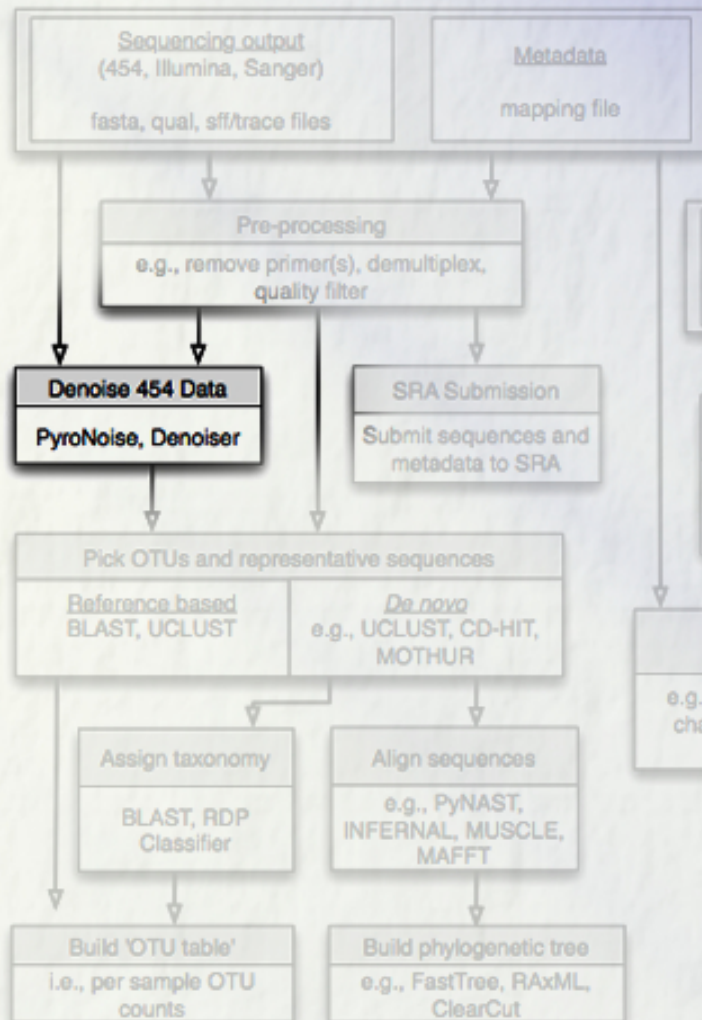
2. Howard Hughes Medical Institute, University of Colorado at Boulder, Boulder, Colorado, USA.

Correspondence to: Rob Knight<sup>1,2</sup> e-mail: [rob.knight@colorado.edu](mailto:rob.knight@colorado.edu)

e.g., 2D and 3D PCoA plots, distance histograms, taxonomy pie charts, rarefaction plots, OTU network visualization, jackknifed

Pyrosequencing has revolutionized microbial community analysis by allowing the simultaneous assessment of hundreds of microbial communities in multiplex with sufficient depth to resolve meaningful biological patterns. These techniques have been used to study microbial processes on scales ranging from continents<sup>1</sup> to within an individual's body<sup>2</sup>.

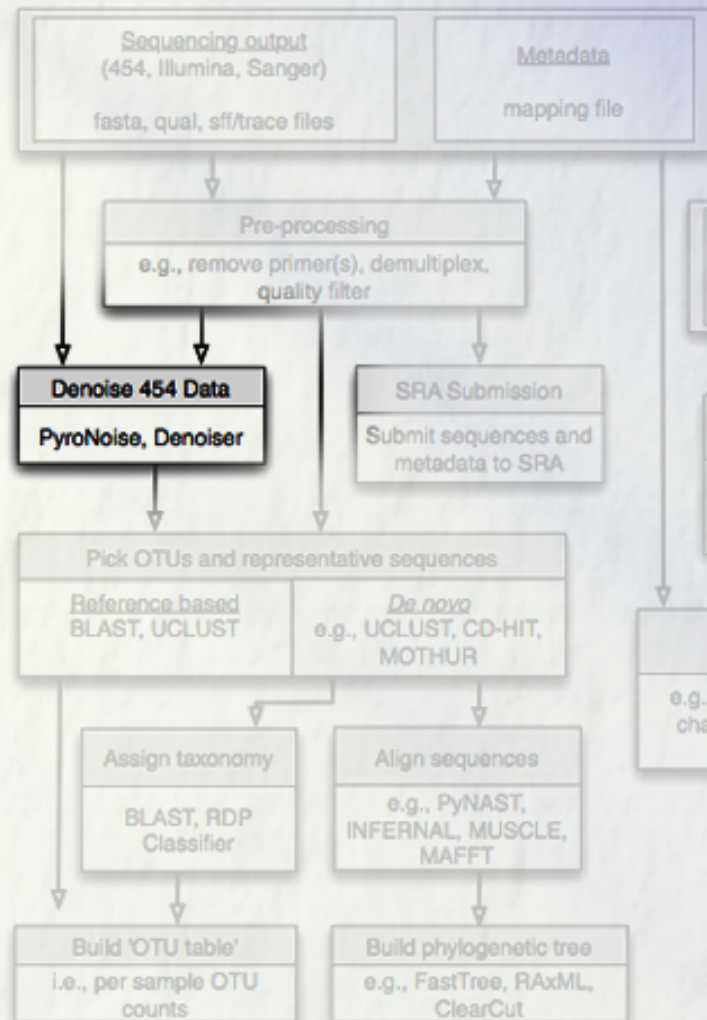
<http://qiime.sourceforge.net>





# QIIME resolves technical challenges in dealing with next-gen data

## The (mythical?) chimera



AAATCCTGGGC

+

TATTCCTATCC

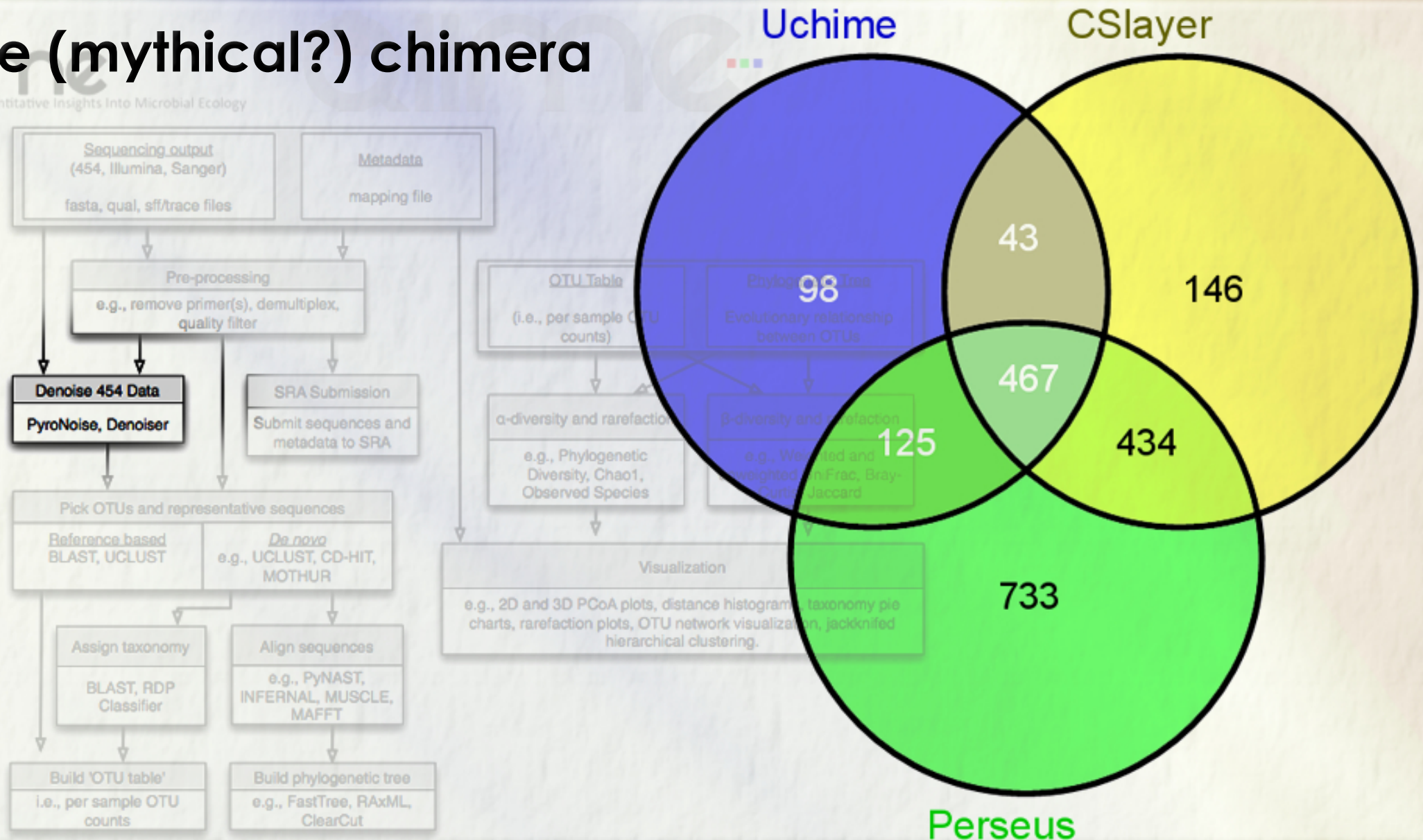
=

AAATCCTATCC



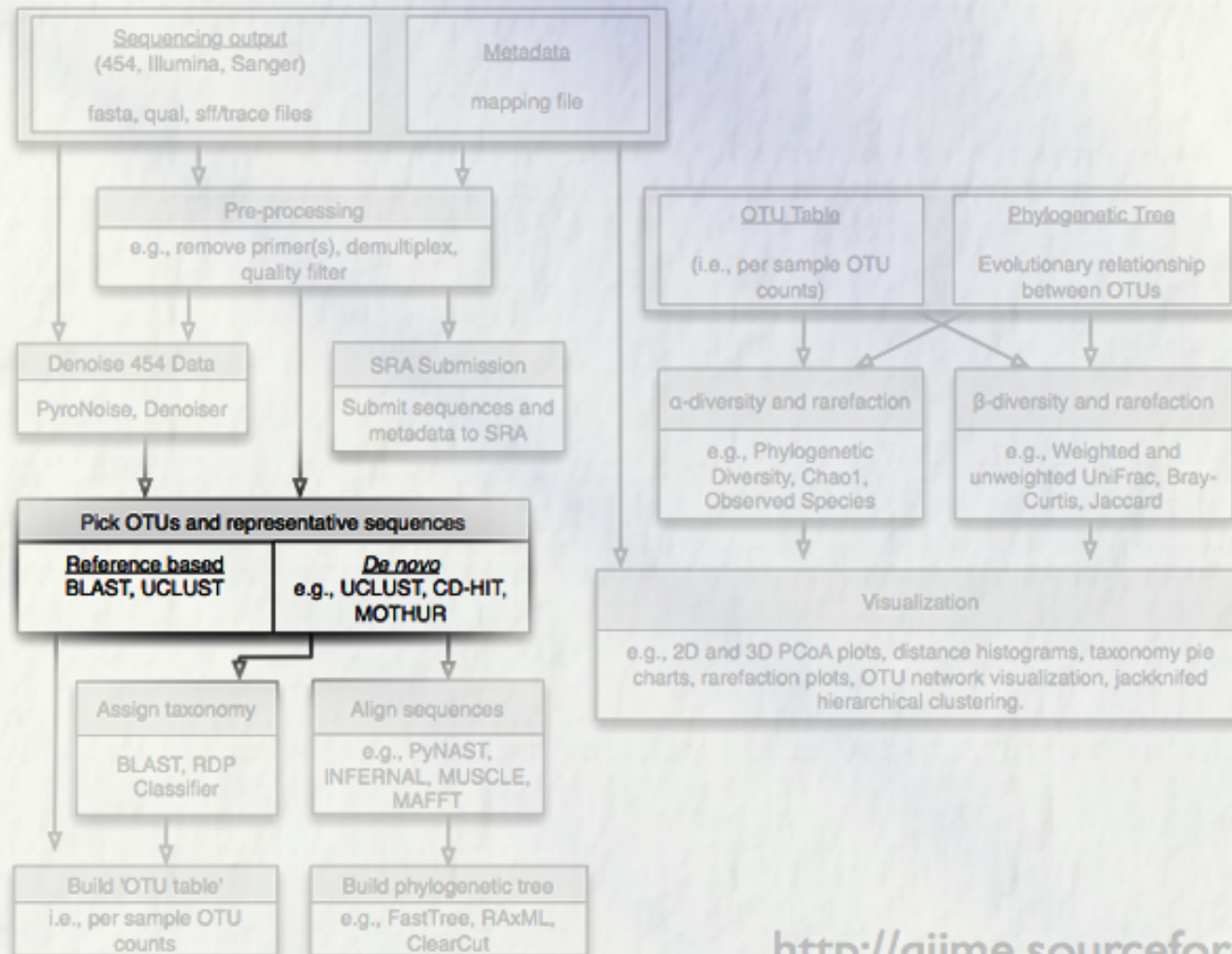
# QIIME resolves technical challenges in dealing with next-gen data

## The (mythical?) chimera





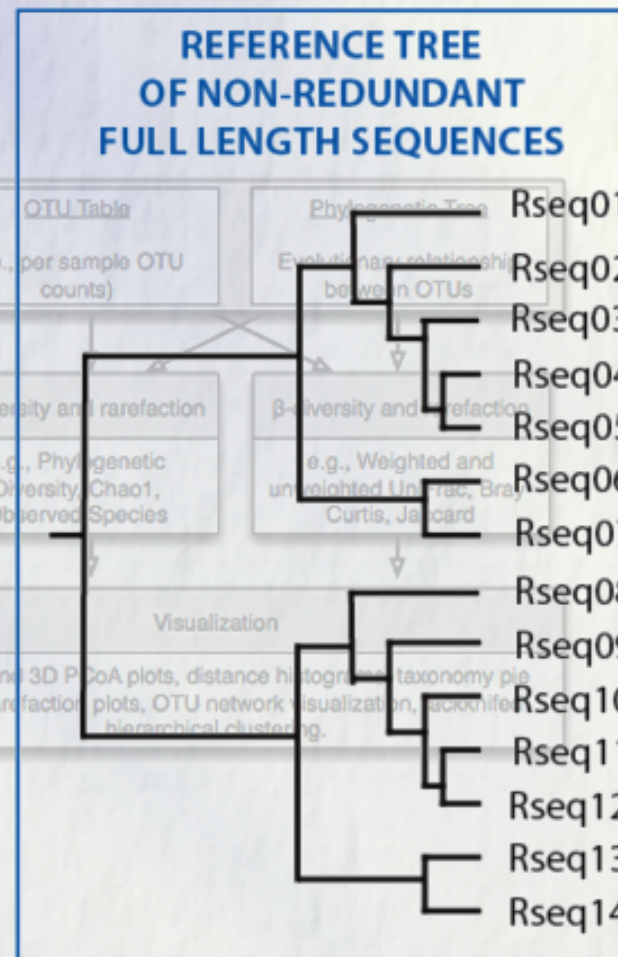
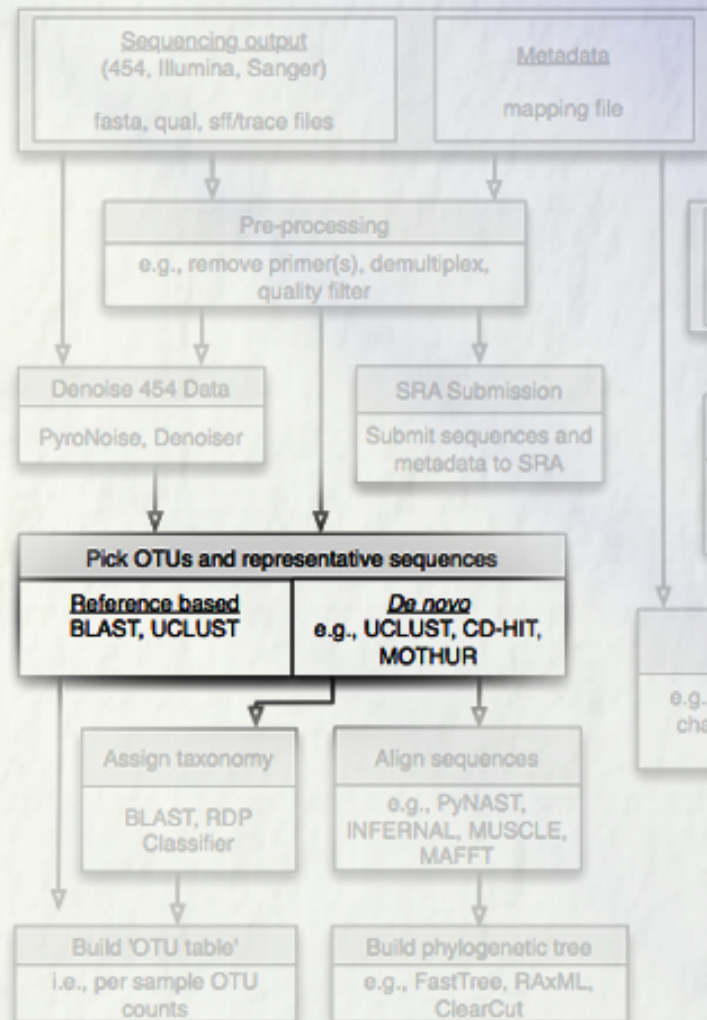
# QIIME resolves technical challenges in dealing with next-gen data



<http://qiime.sourceforge.net>



# QIIME resolves technical challenges in dealing with next-gen data



SEQUENCES FROM SAMPLES 1-5

BLAST

TOP HIT COUNTS

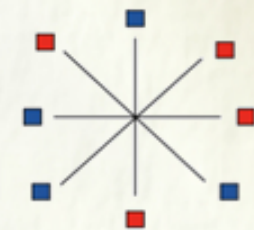
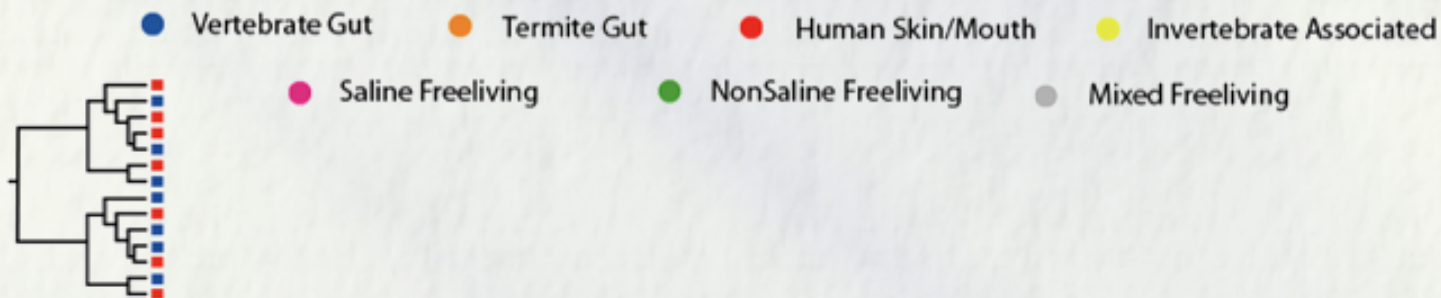
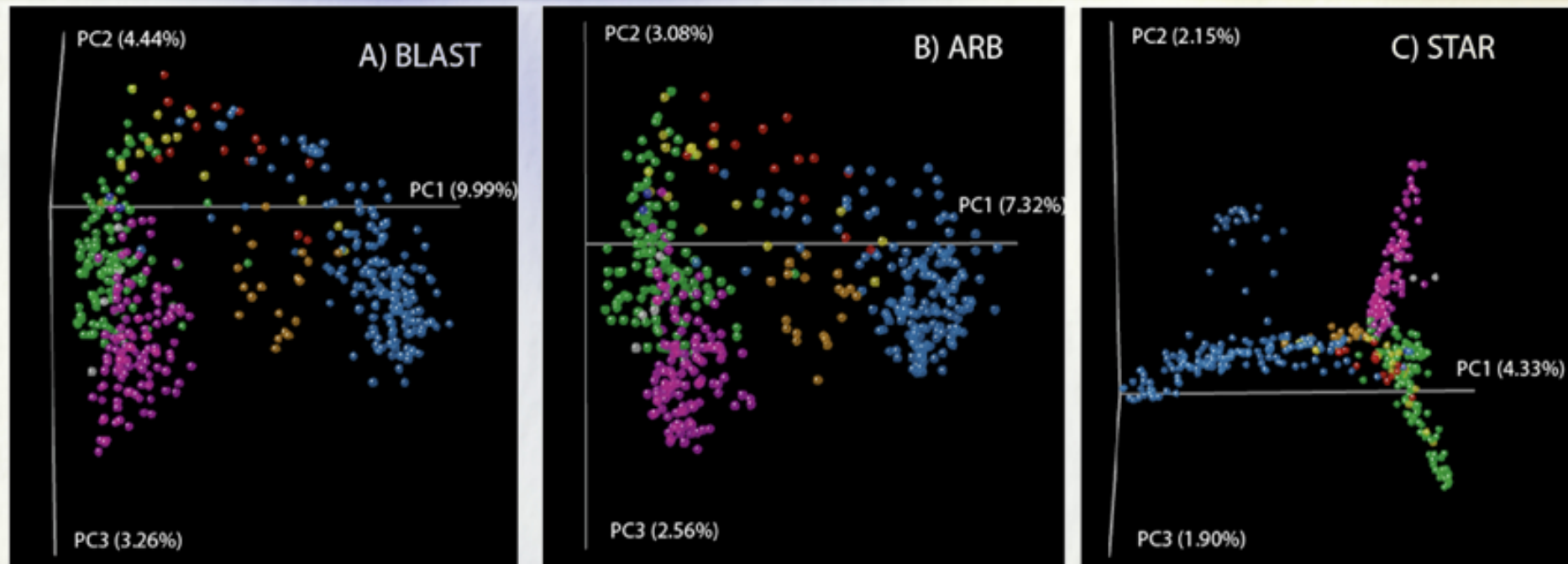
S1 S2 S3 S4 S5

|        |             |   |   |   |   |   |
|--------|-------------|---|---|---|---|---|
| Rseq01 | AGGCTAG...  | 0 | 1 | 5 | 0 | 2 |
| Rseq02 | GGACTAC...  | 1 | 2 | 0 | 7 | 5 |
| Rseq03 | CGTAGCA...  | 0 | 0 | 0 | 0 | 0 |
| Rseq04 | CGTAGCA...  | 1 | 5 | 2 | 0 | 0 |
| Rseq05 | TGATCGA...  | 0 | 2 | 1 | 0 | 1 |
| Rseq06 | AGGCGAG...  | 5 | 1 | 0 | 4 | 1 |
| Rseq07 | ATACTATC... | 0 | 0 | 2 | 1 | 0 |
| Rseq08 | GACTAGC...  | 2 | 3 | 1 | 0 | 1 |
| Rseq09 | AGAGCTG...  | 1 | 2 | 0 | 1 | 1 |
| Rseq10 | TAGCTAGT... | 5 | 1 | 0 | 0 | 1 |
| Rseq11 | GATCGATC... | 0 | 2 | 1 | 1 | 0 |
| Rseq12 | ACTAGCTA... | 2 | 3 | 0 | 1 | 1 |
| Rseq13 | TAGCTAGC... | 1 | 4 | 0 | 1 | 2 |
| Rseq14 | ACTAGCTG... | 4 | 0 | 1 | 2 | 0 |

<http://qiime.sourceforge.net>



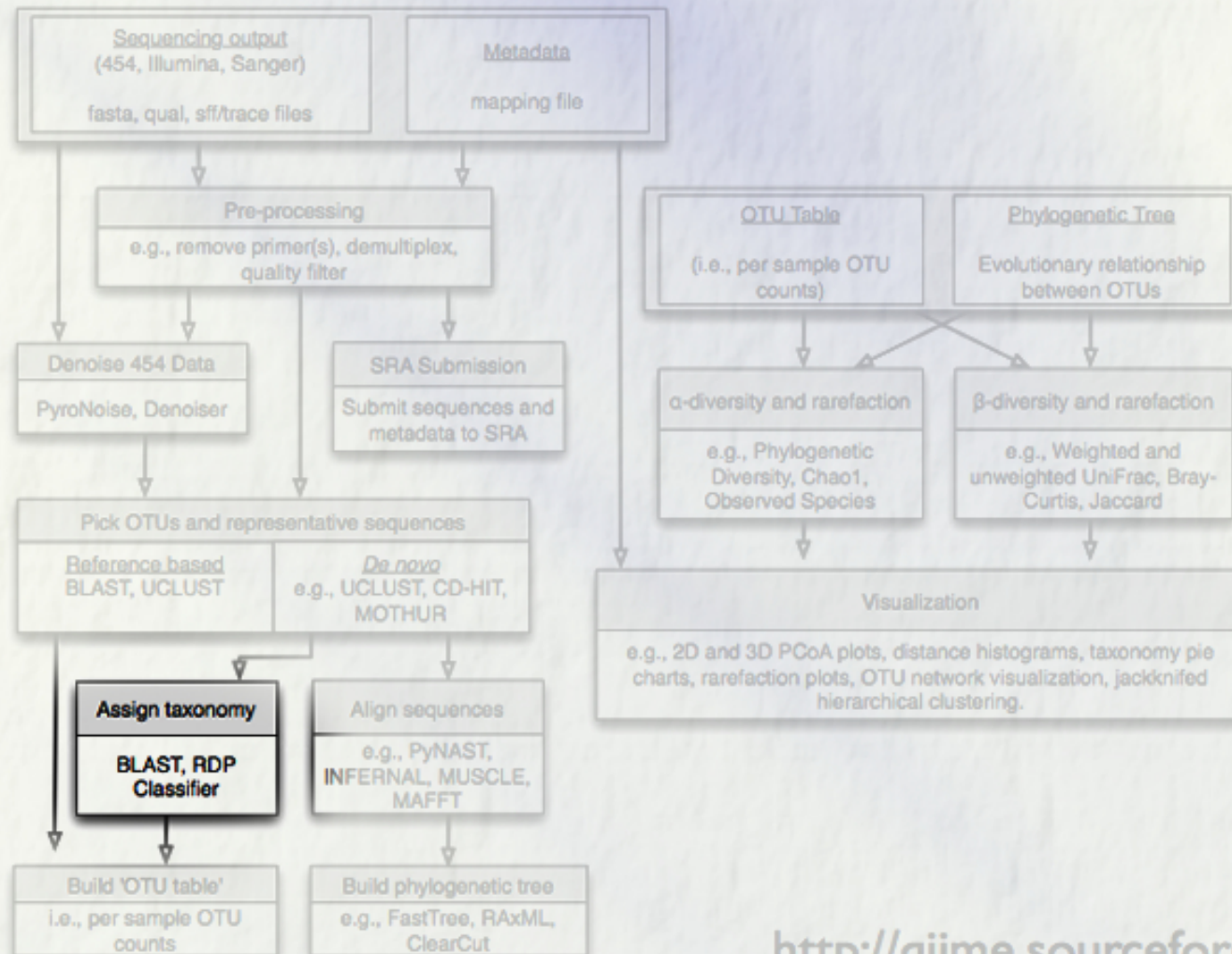
# BLAST to Reference Tree Protocol (global dataset reanalysis)



- BLAST to greengenes coresct tree replicates Arb parsimony insertion results when phylogenetic relationships matter

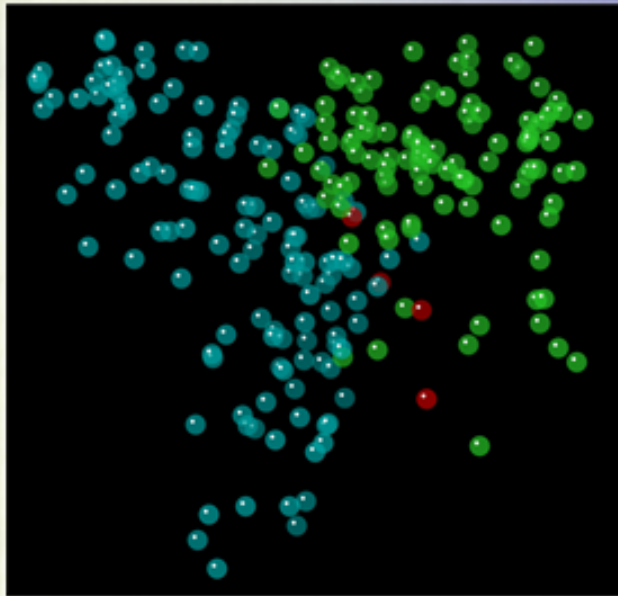


# QIIME resolves technical challenges in dealing with next-gen data

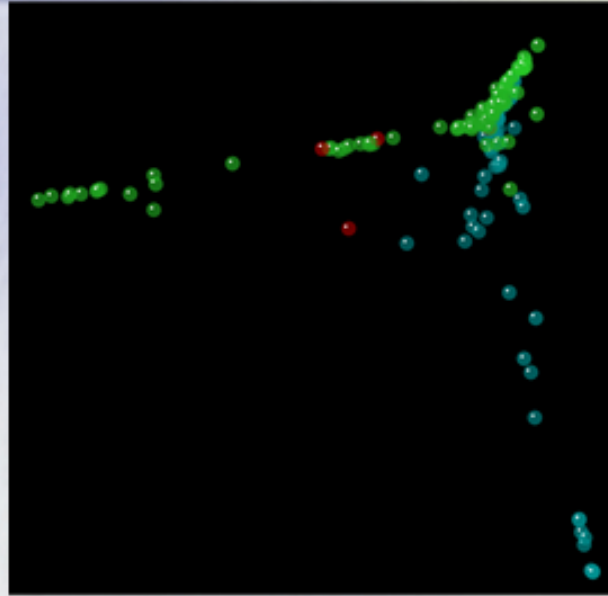




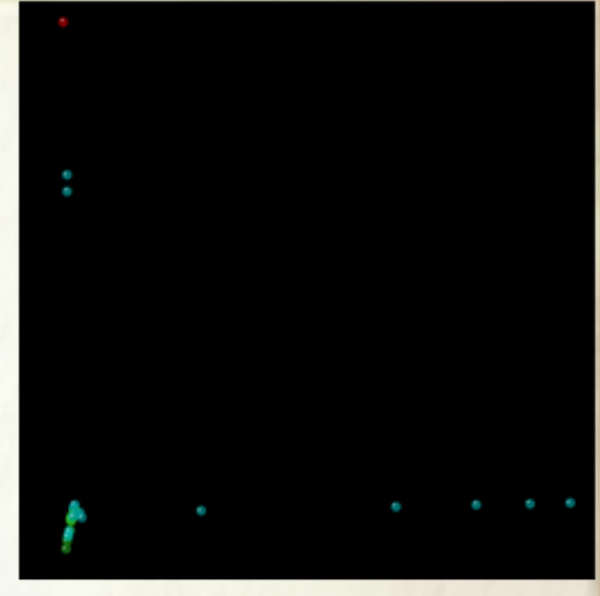
**...and the choices made in the pipeline can really matter!**



**UniFrac**



**Kulczynski**

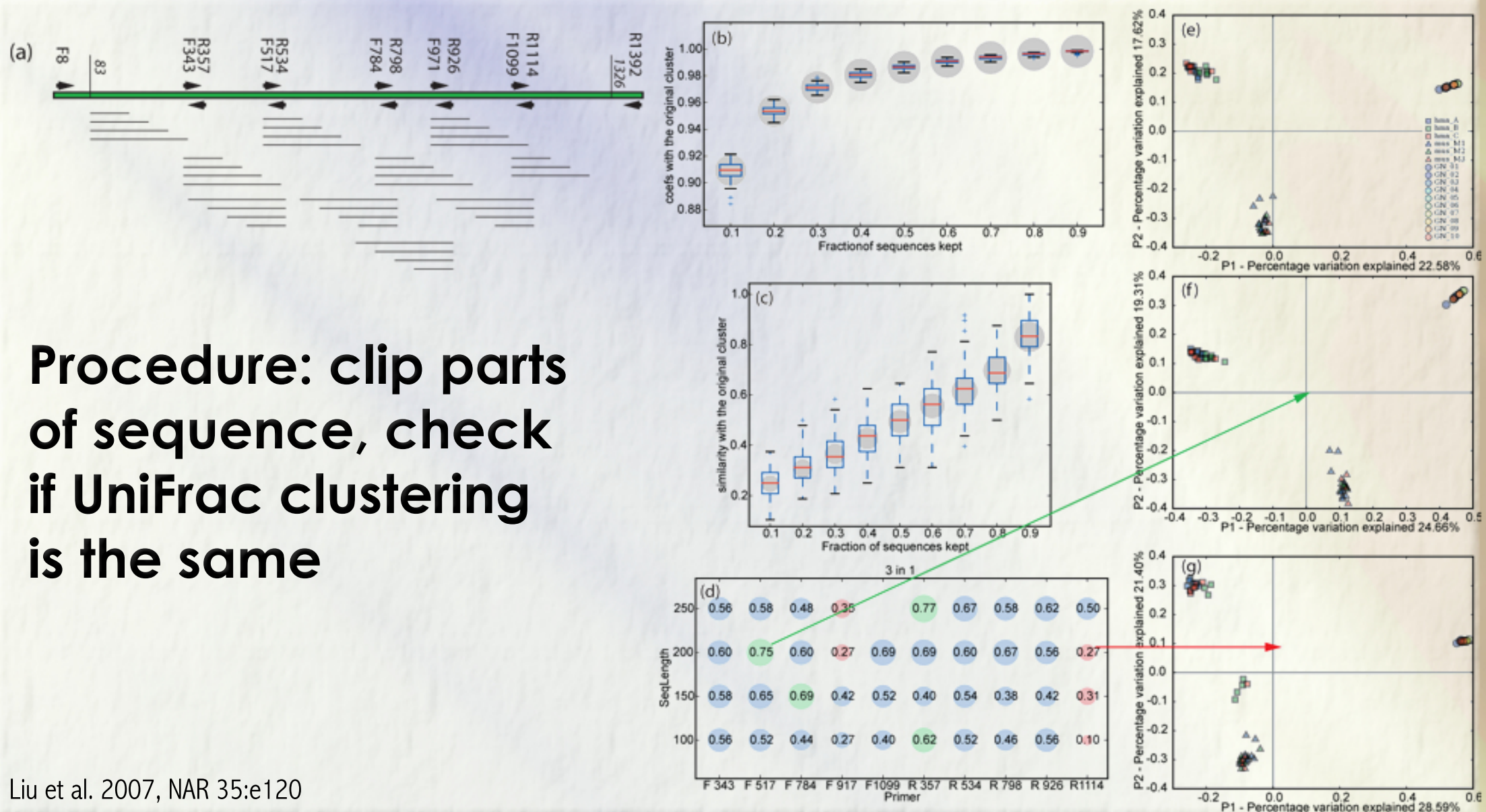


**Euclidean**

**Tree-based metric (UniFrac) reveals patterns that are obscured using taxon-based metrics because the tree better reconstructs evolutionary history**

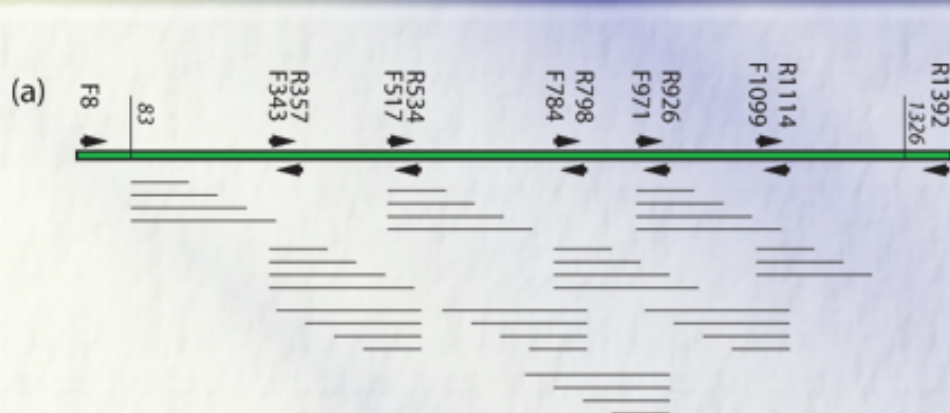


# The region of the 16S rRNA can have large effects on clustering...

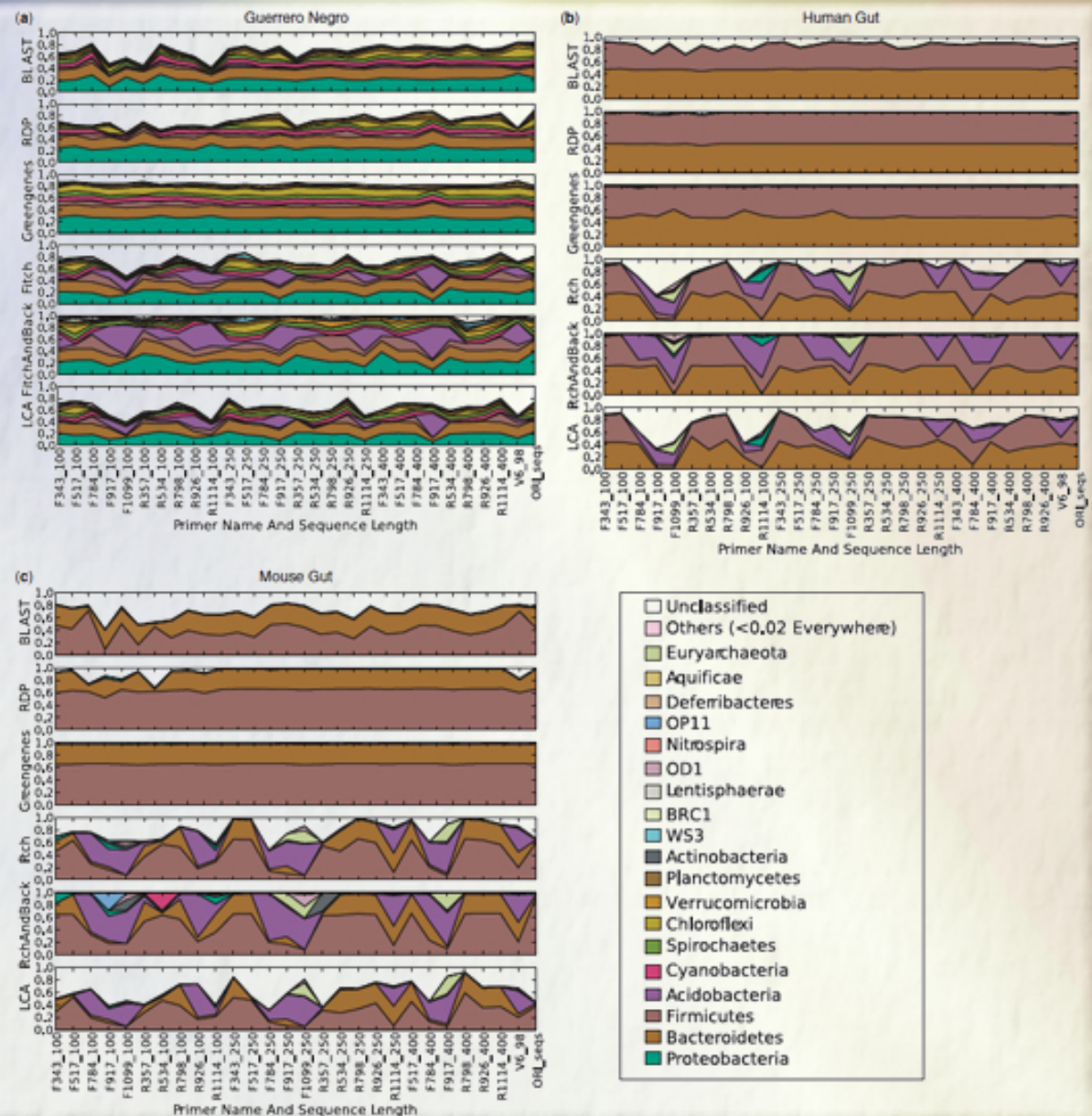




# ...and taxonomy: method not sample can dominate result!

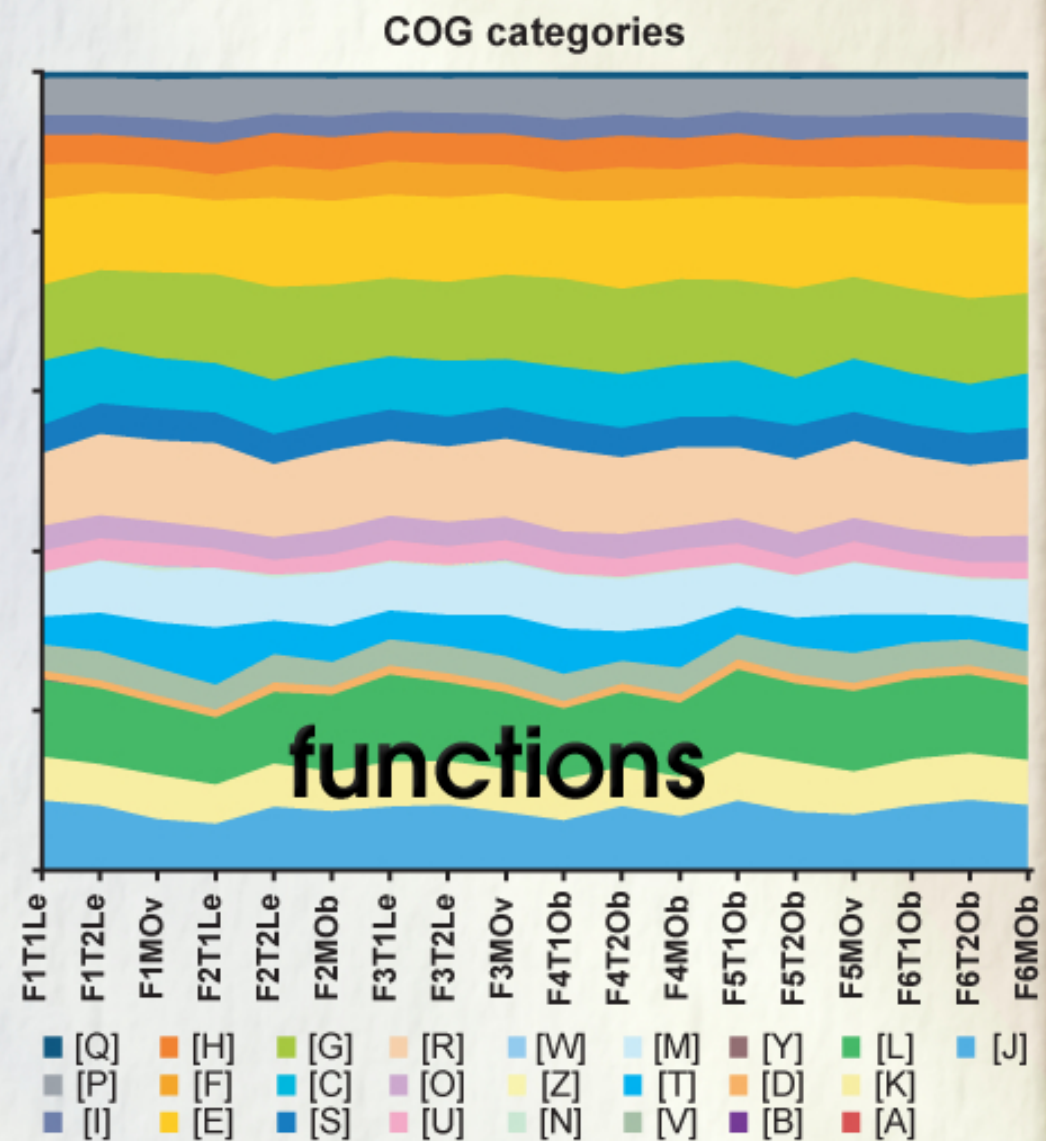
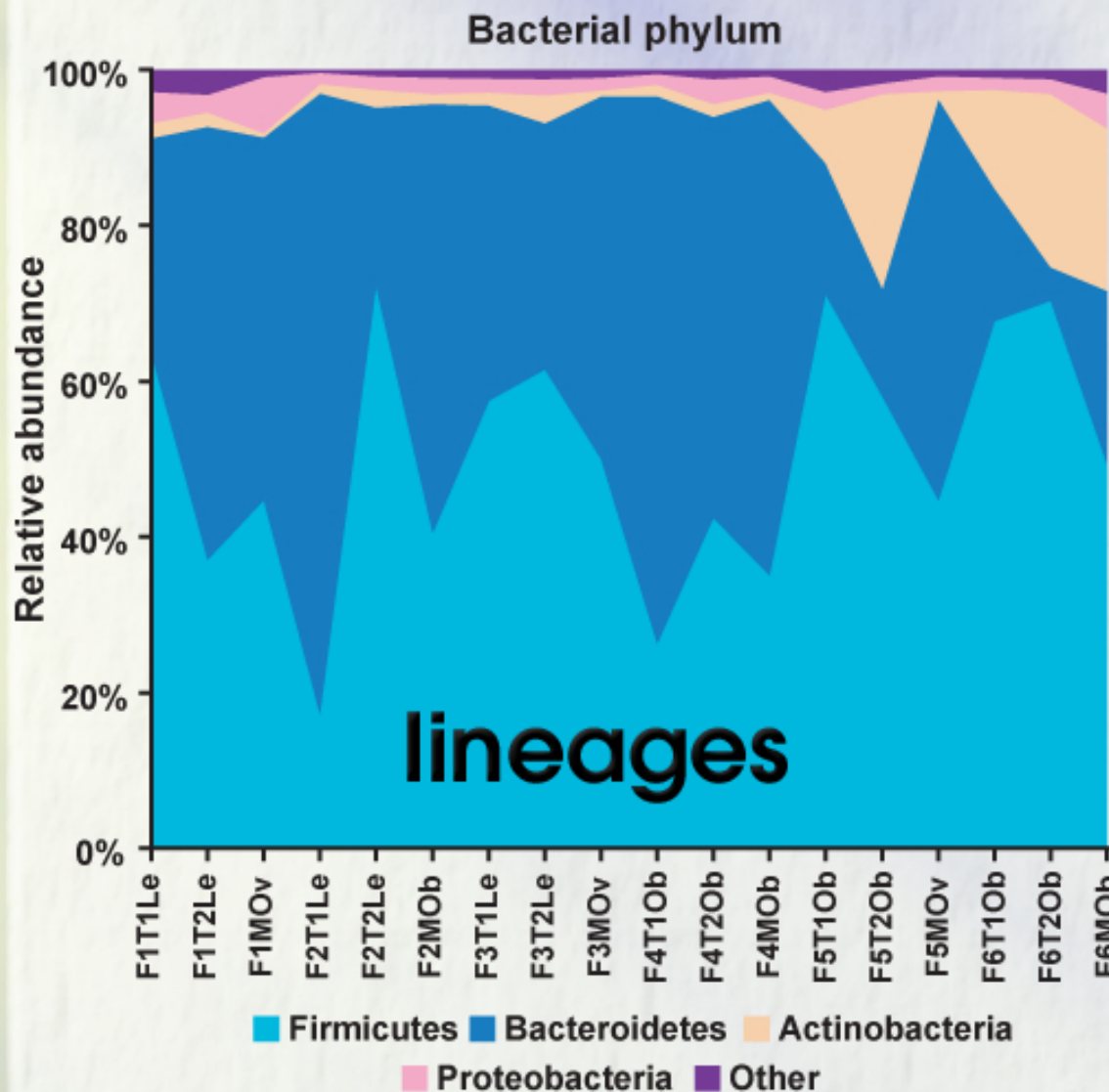


Procedure: clip parts of sequence, check if taxonomy assignment (using several methods) is the same





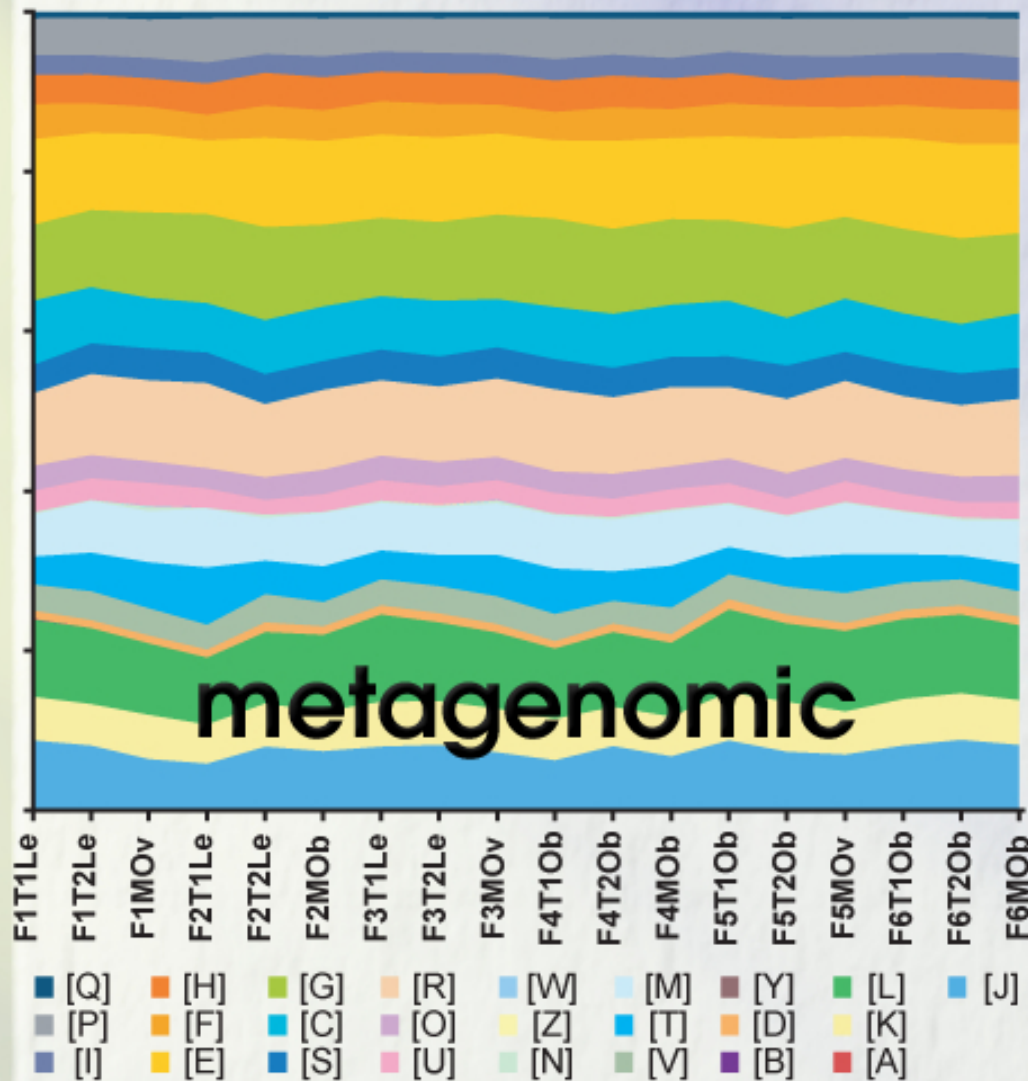
# Remarkably, totally different species lead to same functions...





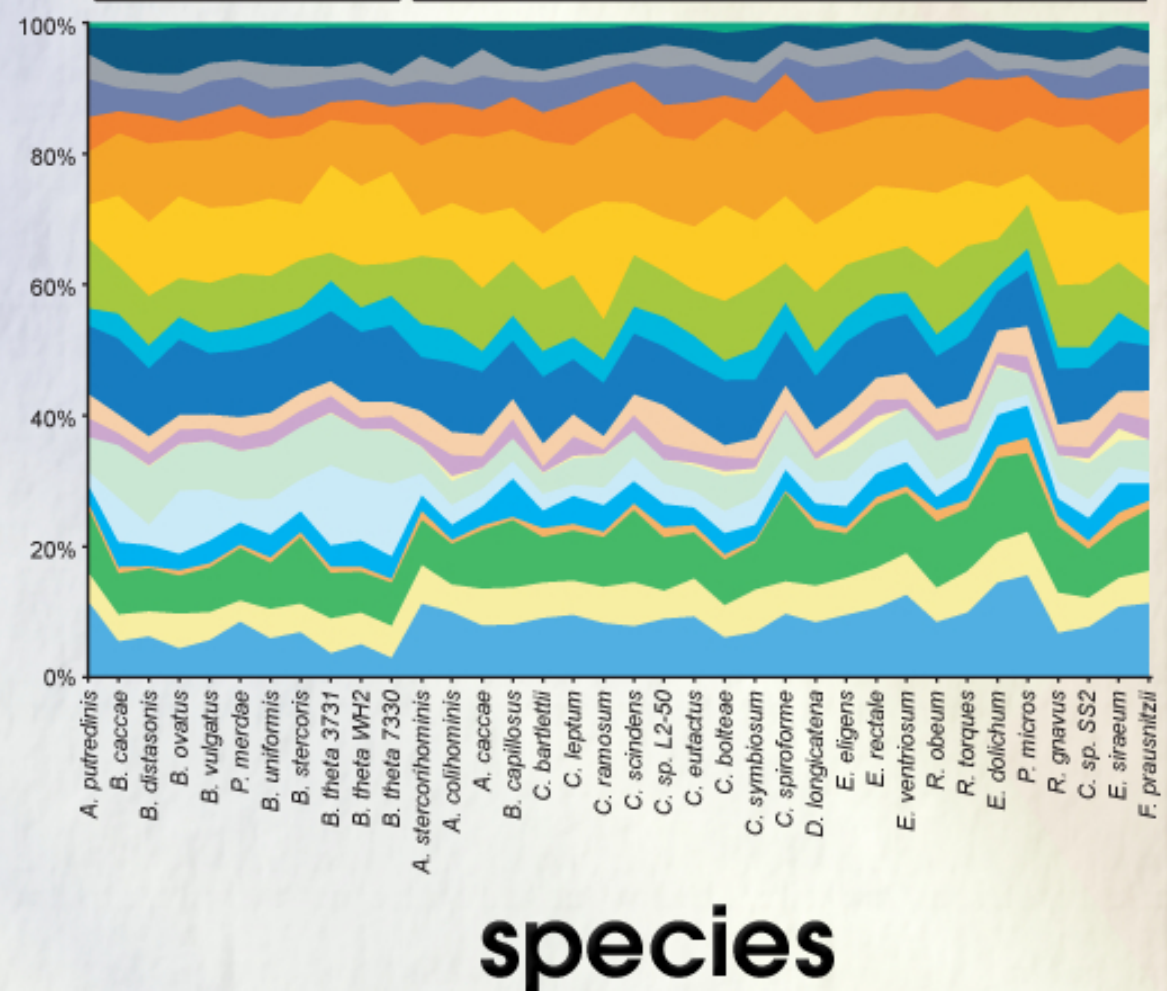
# ...in ways that can't be explained by the genomic gene contents

COG categories



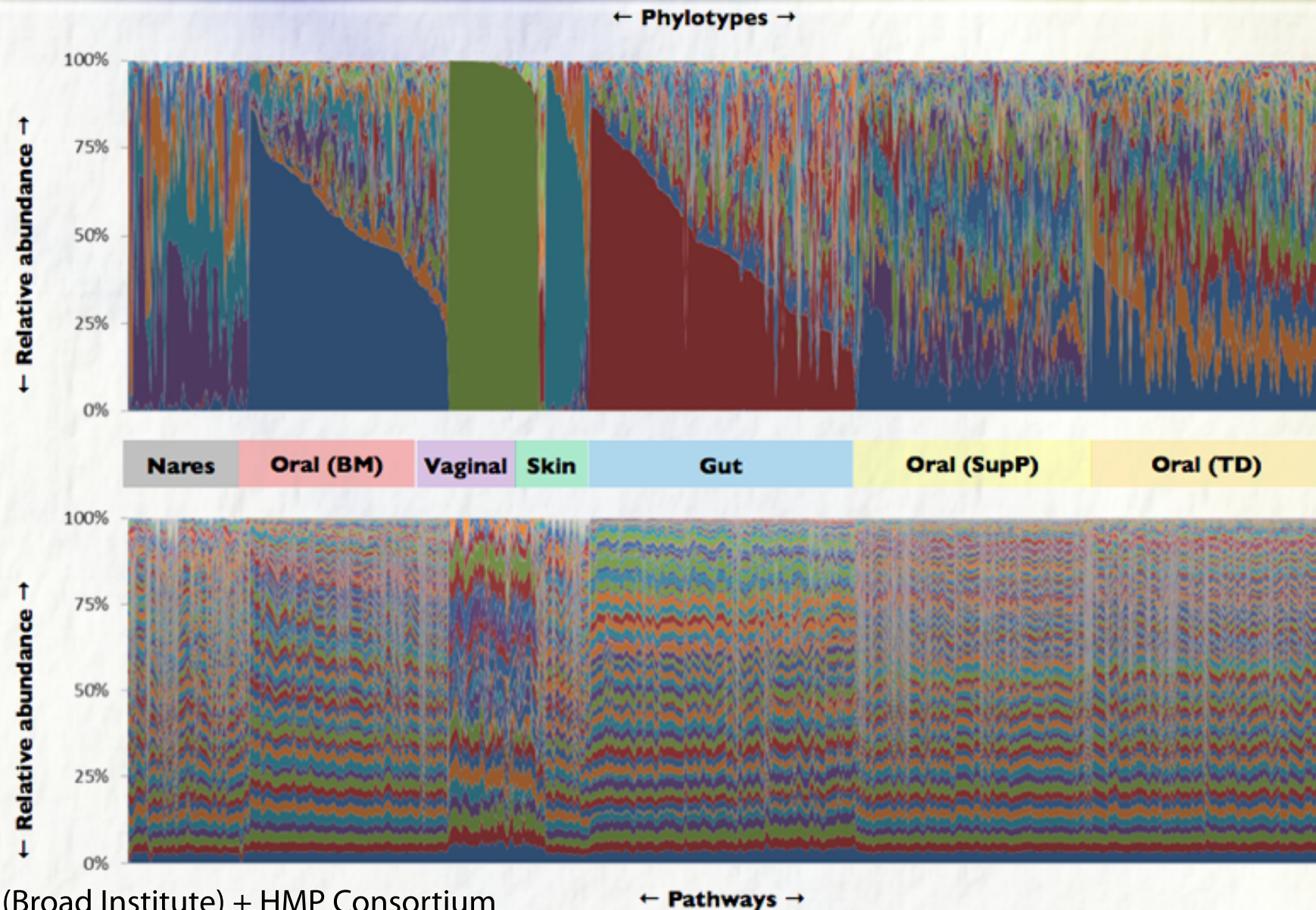
Bacteroidetes

Firmicutes



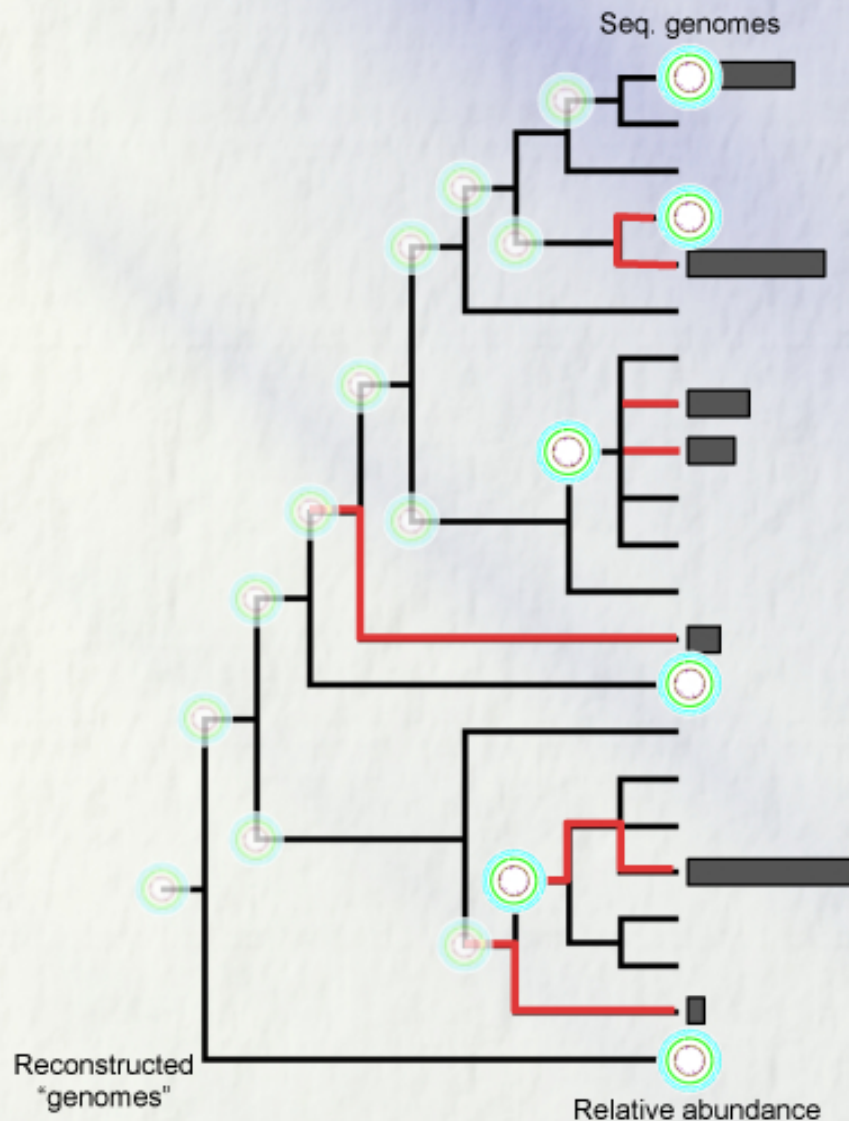


# HMP data confirms trend across body habitats



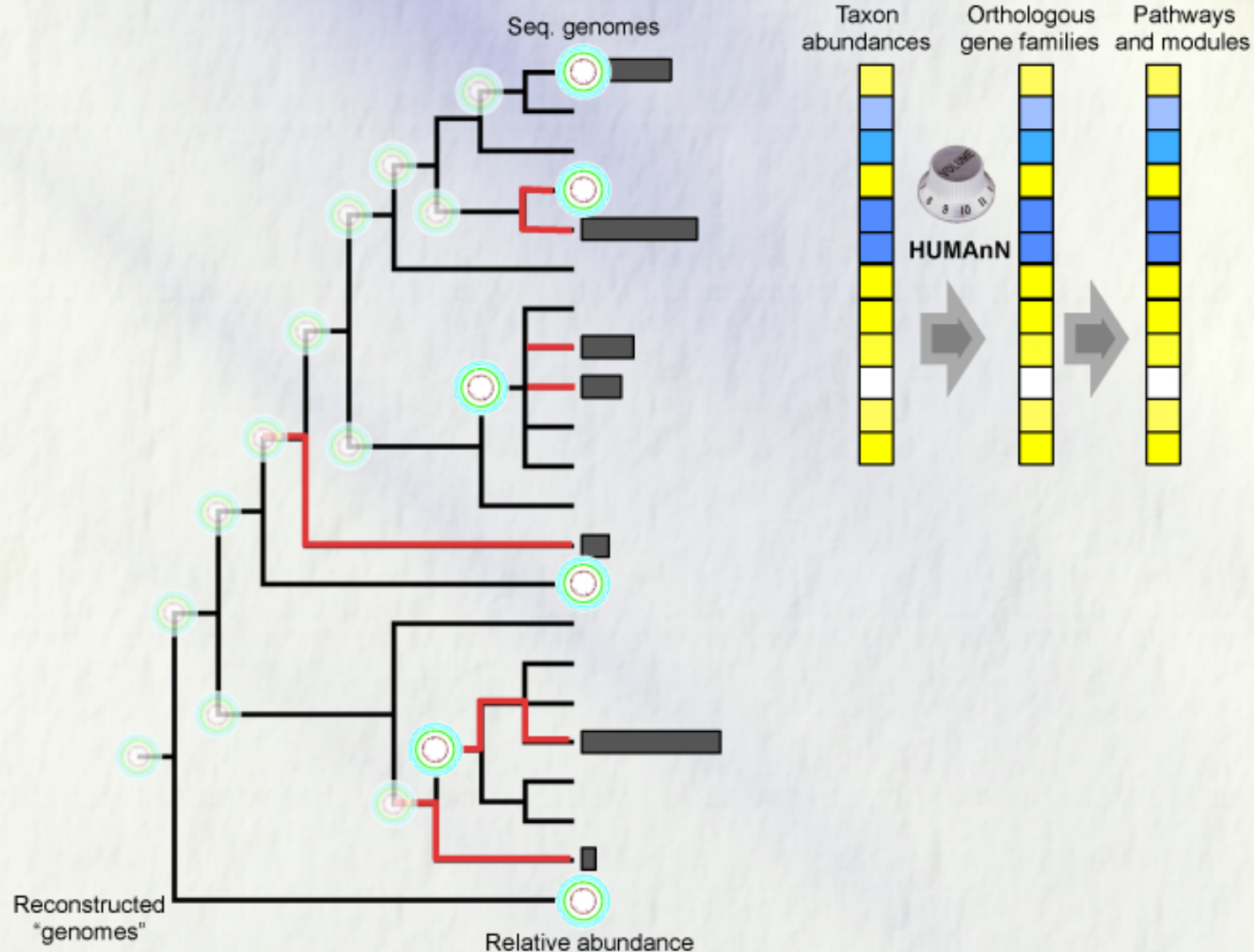


**...and of course 16S rRNA isn't the whole story...**



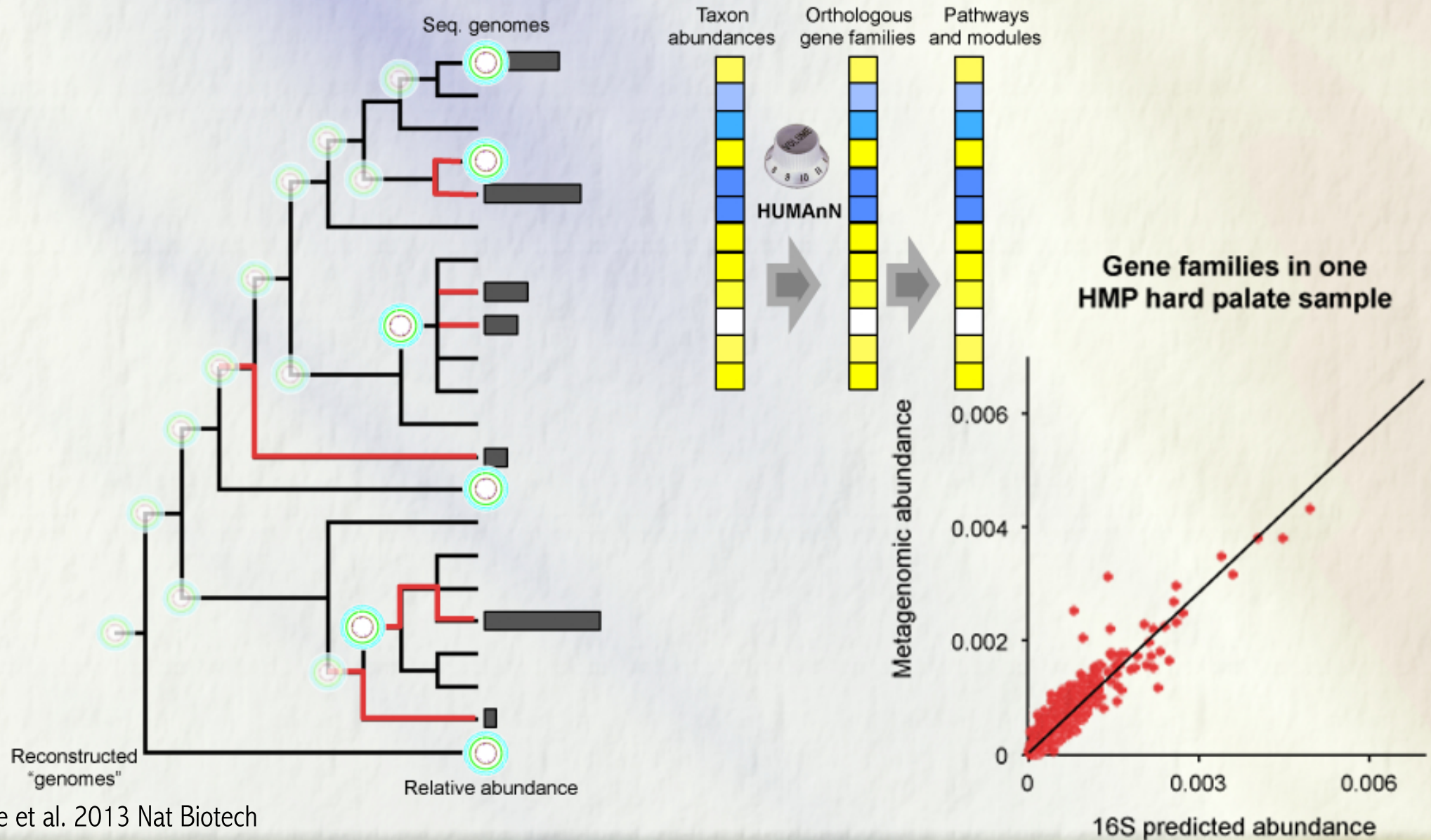


# ...and of course 16S rRNA isn't the whole story...



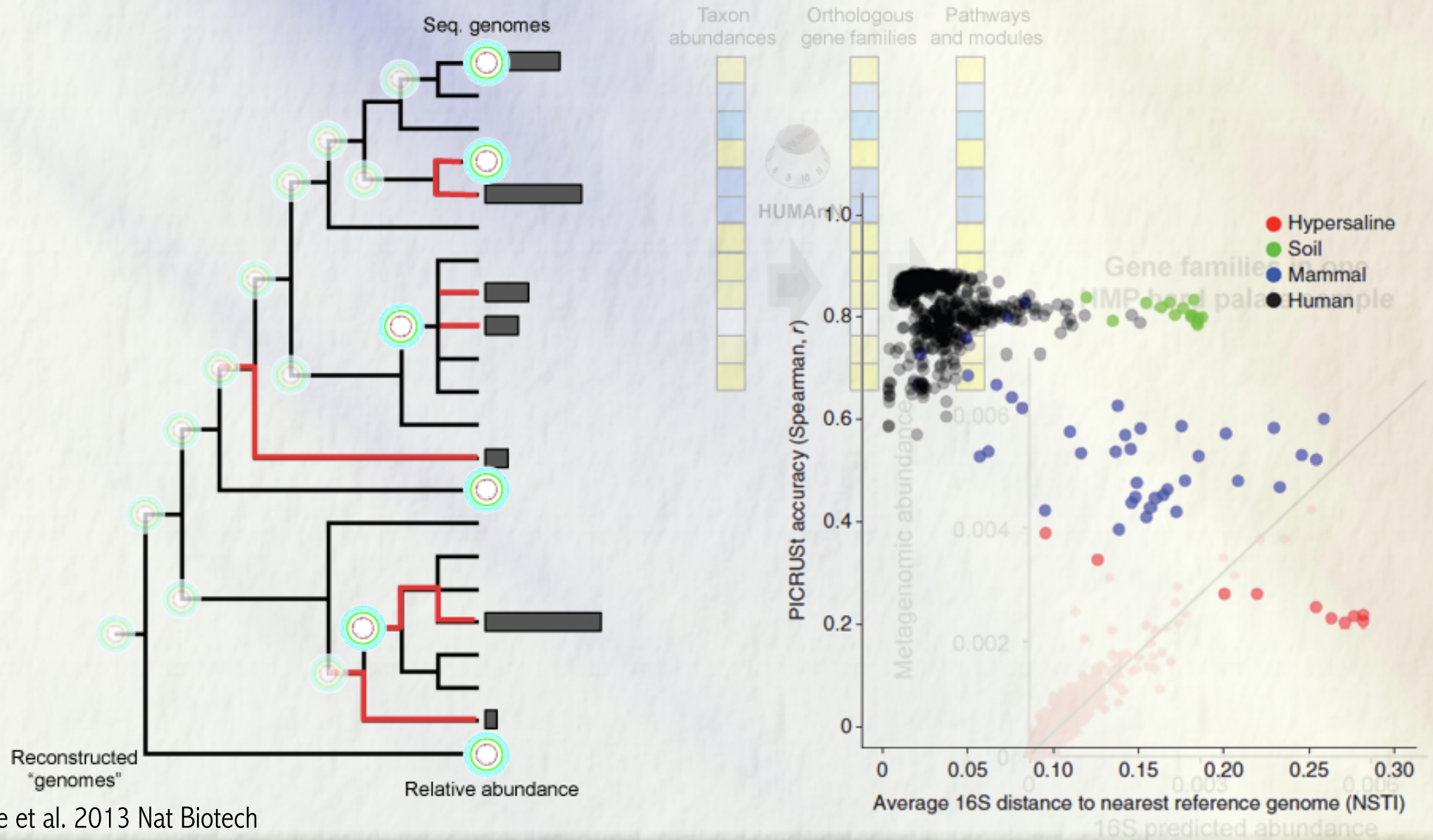


# ...though it's more than you might think...



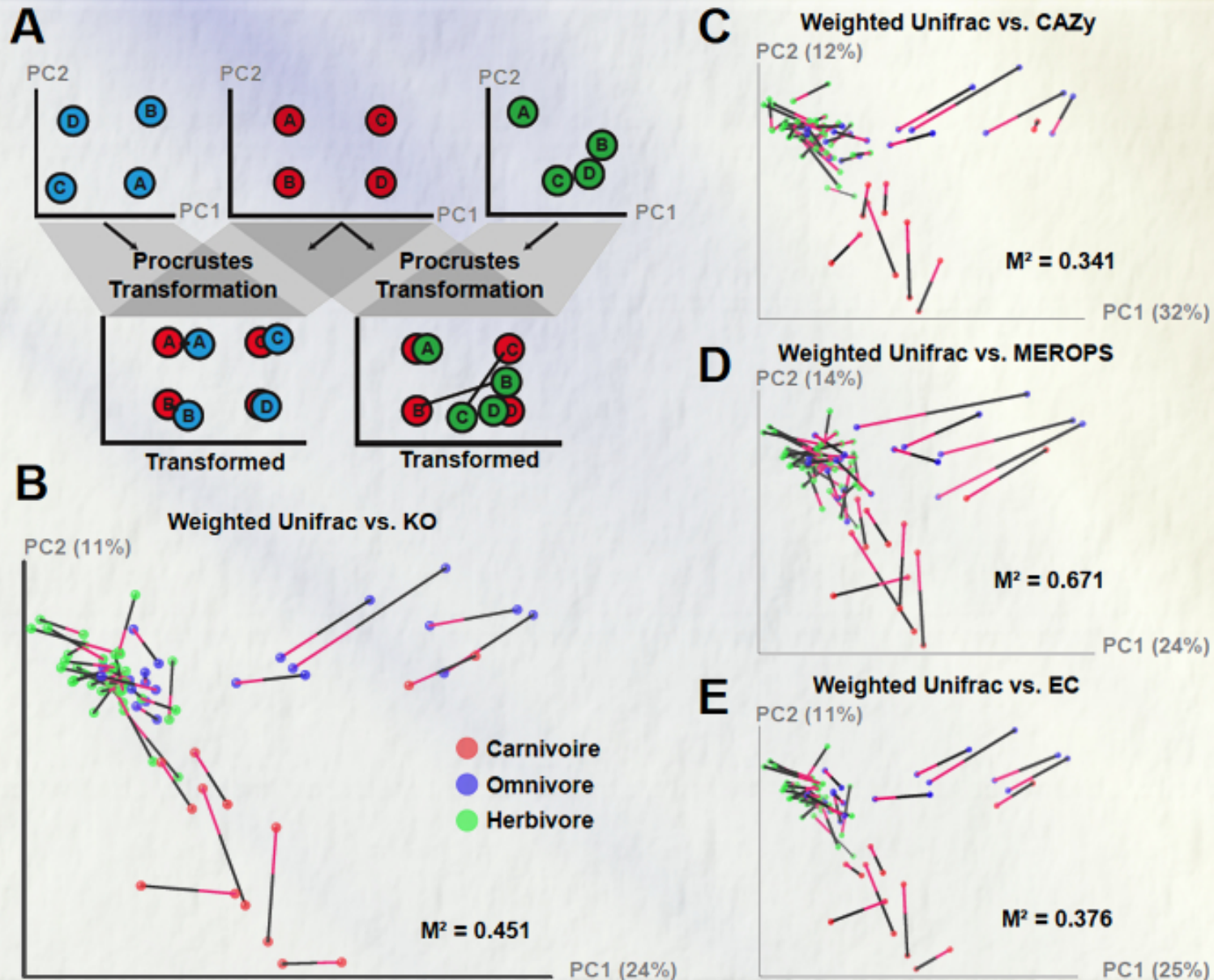


# ...though it's more than you might think...





# Can match multilevel data (e.g. 16S rRNA vs metagenomic) w/ Procrustes



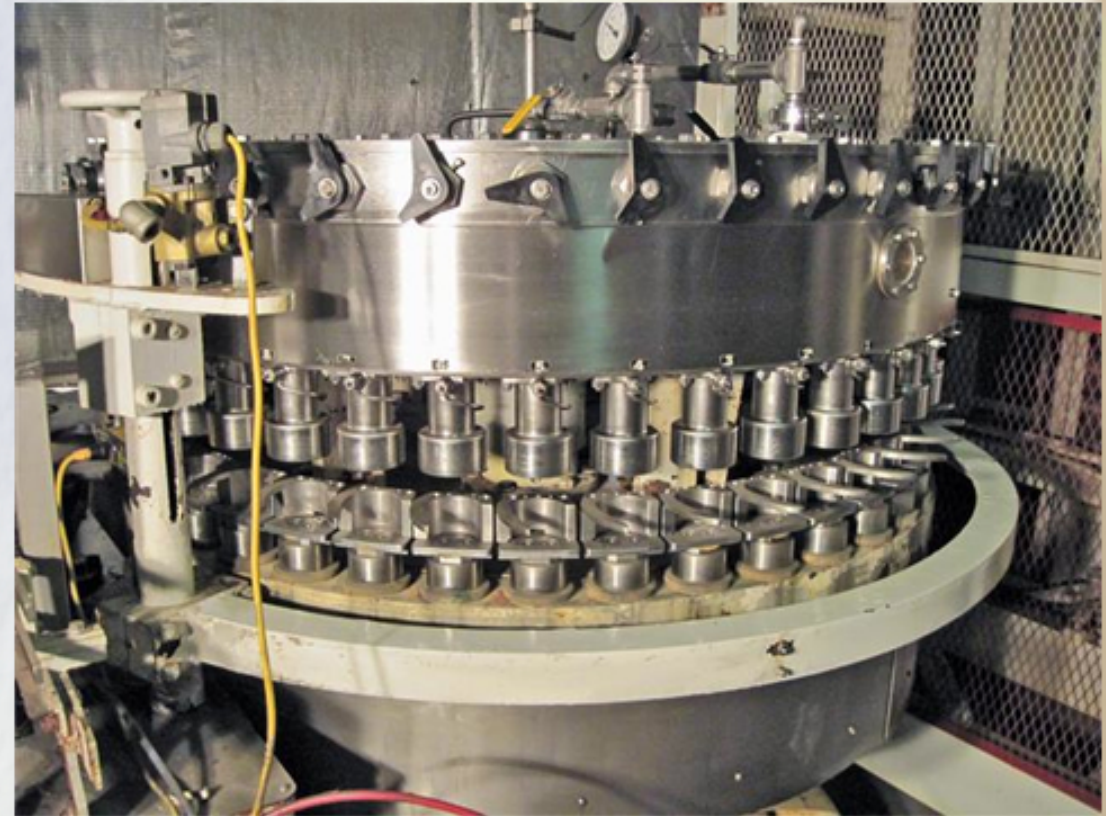


# As sequencing moves from a cottage industry...





**...to a real industry...**





# ...the labeling requirements also change





# Supervised Classification

Using a set of samples with known category labels, learn to put unlabeled samples in the correct category

E.g. identify cancer cells from microarrays

Differs from unsupervised classification (e.g. PCA), where we don't have predetermined categories

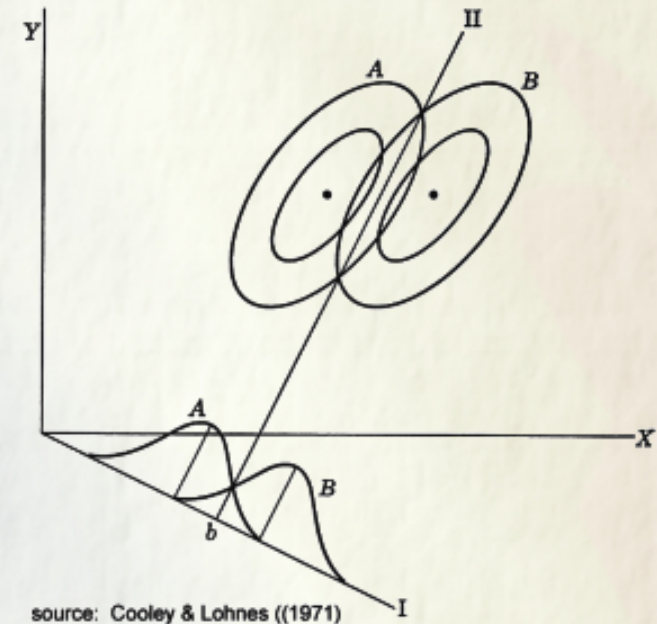
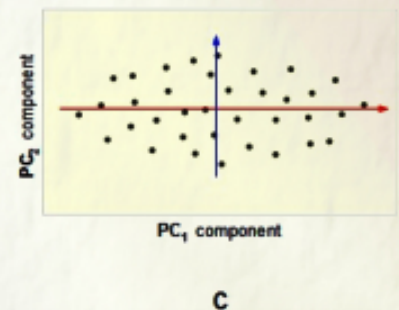
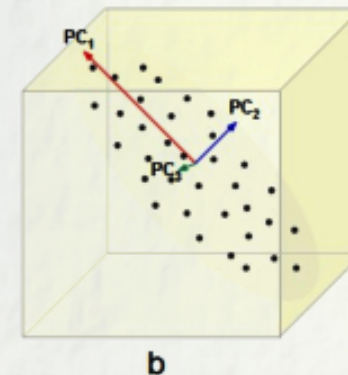
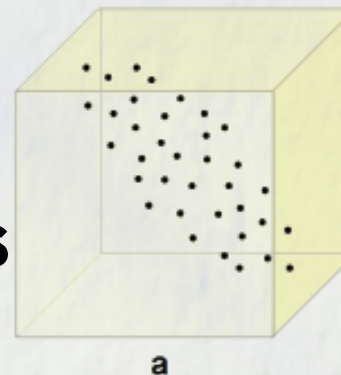


Figure 9.1.



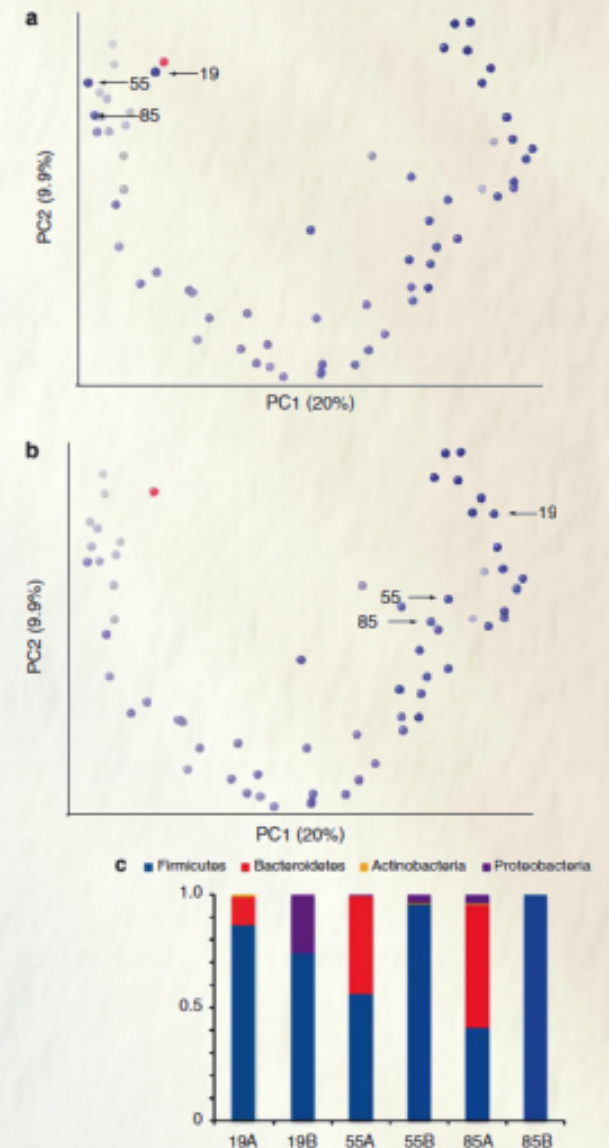


# Supervised Classification and mislabeling

Koenig et al. 2010: initially puzzling results disappear on resequencing the samples (Knights et al. 2010 ISME J)

When patterns are clear, trusting pattern is better than trusting asserted labels

-> use supervised learning to fix labels

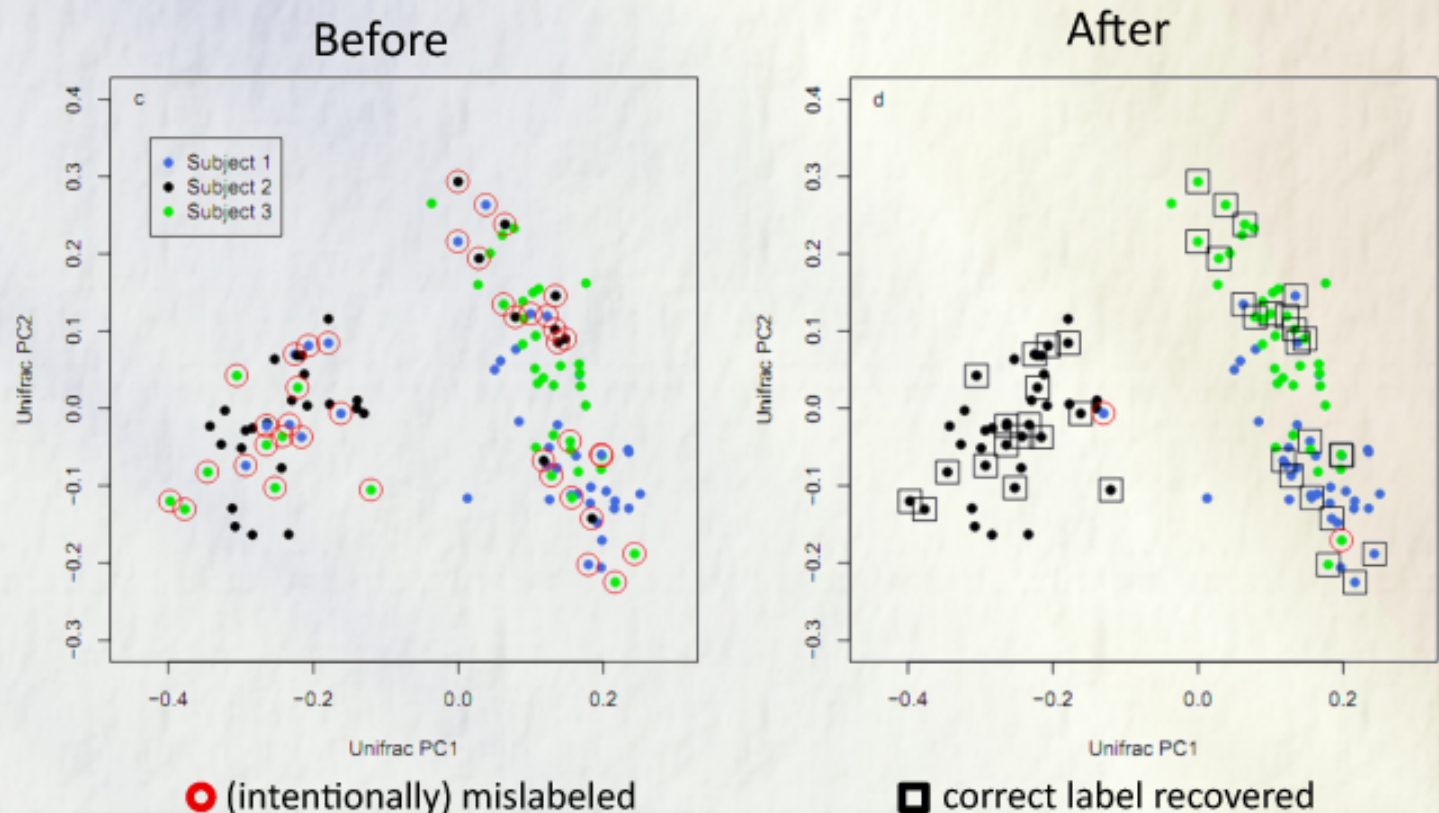




# Label correction works!

Using data from Fierer et al. 2010 PNAS and Costello et al. 2009 Science, randomly introduced labeling errors can be corrected by a random forests classifier

[What's a random forest? Decision trees trained on random sample of data and few variables, each tree votes once on new data so errors cancel]





# QIIME in the cloud lets anyone analyze their data...





...seriously, *anyone*...





# Data integration workflow



Upload and verify your  
MIMARKS-compliant metadata  
& RAW sequencing data

Reference pick OTUs

Download your and the HMP  
OTU reference based tables and  
mapping files



Create a QIIME cluster using  
StarCluster and upload your  
files

Merge your OTU tables and  
mapping files

Rarefy and filter your OTU table  
at convenience



Repeat

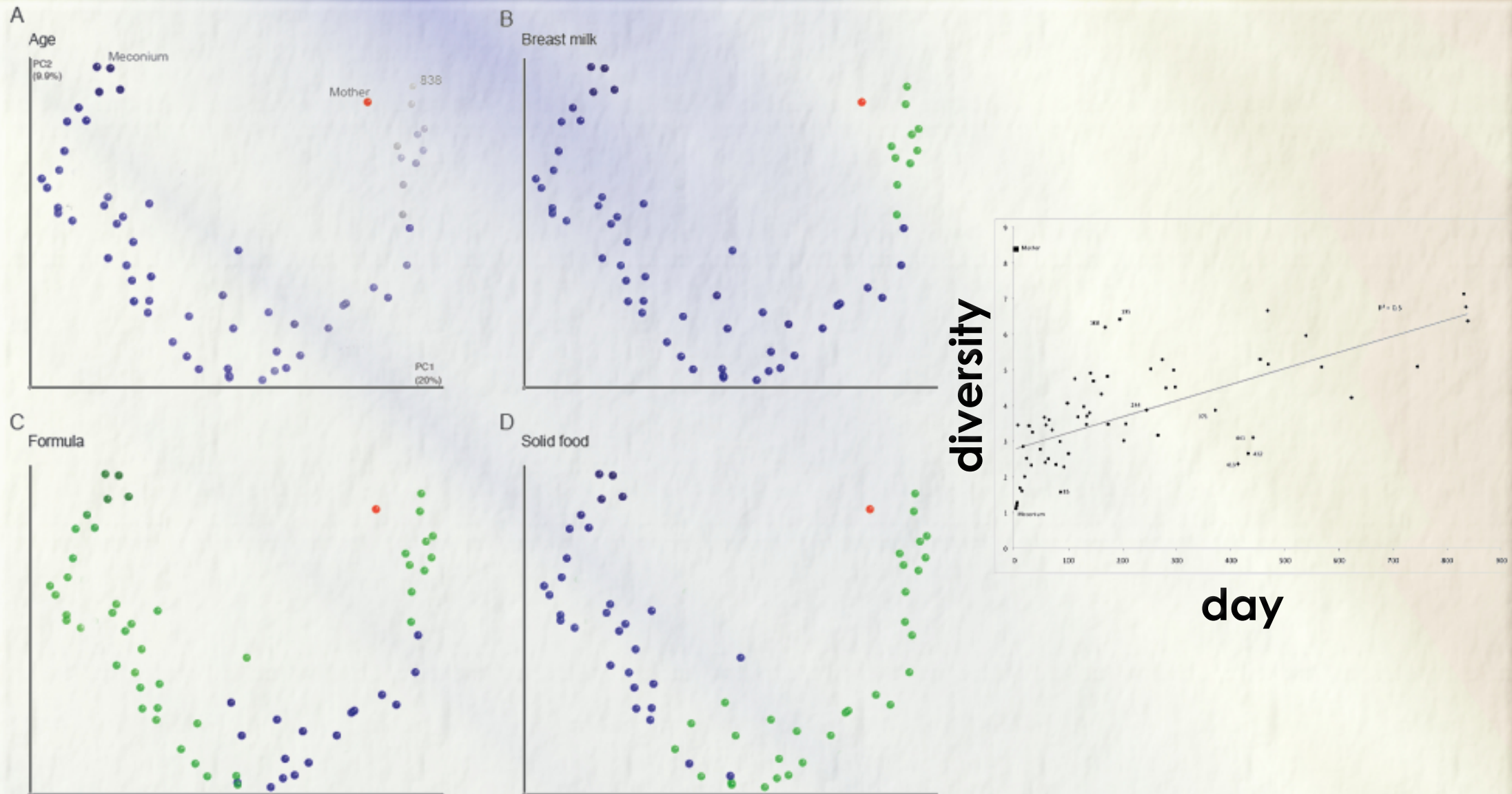
Filter at convenience

Analyses in parallel using QIIME

Antonio Gonzalez



# Infant gut microbiota changes radically: integrate with adults?





# Acknowledgments

## **Knight lab:**

Cathy Lozupone

Chris Lauber

Donna Berg-Lyons

Elizabeth Costello -> Stanford

Greg Caporaso -> NAU

Justin Kuczynski -> 2nd Genome

Jesse Stombaugh

Doug Wendel

Antonio Gonzalez

Jeremy Widmann

Meg Pirrung

Tony Walters

Daniel McDonald

Nigel Cook

Dan Knights

Jesse Zaneveld -> Oregon State

Jens Reeder -> Genentech

Julia Goodrich -> Cornell

Ryan Kennedy -> Penn

Zongzhi Liu -> Yale

Micah Hamady -> world travels

Jerry Kennedy

Greg Humphrey

Jose Clemente

Will van Treuren

Jess Metcalf

Laura Parfrey-Wagner

Bharath Prithviraj

Gail Ackermann

Luke Ursell

## **[a few of the] Collaborators using UniFrac/QIIME:**

Noah Fierer (CU, EEB), Jeff Gordon (Wash U), Ruth Ley (Cornell), Peter Turnbaugh (Harvard), Andrew Gewirtz (Emory), Maria Gloria Dominguez (UPR), Rick Bushman, Gary Wu, Jim Lewis (Penn), Scott Kelley (SDSU), Marty Blaser (NYU), Jacques Ravel (U Maryland), Allan Konopka (PNNL), David Relman (Stanford), Laura Katz (Smith), Norm Pace (CU), Andy Goodman (Yale), Jo Handelsman (Yale), Lars Angenent (Cornell), George Church (Harvard)

## **Gordon Lab:**

Tanya Yatsunenko, Brian Muegge, Michelle Smith, Vanessa Ridaura, Alejandro Reyes, J Faith, Henning Seedorf, Liz Hansen, Nate McNulty, Jill Manchester, Sabrina Wagoner, Jessica Hoisington-Lopez

## **PyCogent:**

Gavin Huttley (ANU) + many others

## **Methods:**

Joe Jones (USC), Eric Nawrocki (HHMI), Sean Eddy (HHMI), Phil Hugenholtz (U Syd), Henry Tufo (CU, CS), Manuel Lladser (CU, APPM), Andy Martin (CU, EEB), Alexis Stamatakis (Munich), Steven Brenner (Berkeley), Folker Meyer (ANL), Chris Quince (Glasgow), Pawel Gajer (U Maryland), Eytan Rupp (TAU)

## **HMP:**

DACC: Owen White, Jennifer Wortman, Todd Desantis, Gary Andersen, Nikos Kyrpides, Victor Markowitz; Centers: WashU, Broad, BCM, JCVI; Manuscript prep: Curtis Huttenhower, Dirk Gevers, HMP Consortium

## **ICDDR,B:**

Tahmeed Ahmed, Rashidul Haque, Dinesh Mondal, Mukul Sarker

## **EMP:**

Rick Stevens (Argonne), Jack Gilbert (Argonne), Folker Meyer (Argonne), Janet Jansson (LBNL), Jed Fuhrman (USC), Jonathan Eisen (UC Davis), many, many sample donors

## **Support (across all projects):**

HHMI, NAKFI, USAF, CCFA, Gates, NASA, NIDDK, NHGRI, NIGMS, NHLBI, BSF, USDA/NSF, DOE, SCR+Biophysics TGs, C2B2, Butcher Foundation, Sloan IE, NIJ, Keck, Templeton, Moore





# Acknowledgments

## Knight lab:

Cathy Lozupone

Chris Lauber

Donna Berg-Lyons

Elizabeth Costello -> Stanford

Greg Caporaso -> NAU

Justin Kuczynski -> 2nd Genome

Jesse Stombaugh

Doug Wendel

Antonio Gonzalez

Jeremy Widmann

Meg Pirrung

Tony Walters

Daniel McDonald

Nigel Cook

Dan Knights

Jesse Zaneveld -> Oregon State

Jens Reeder -> Genentech

Julia Goodrich -> Cornell

Ryan Kennedy -> Penn

Zongzhi Liu -> Yale

Micah Hamady -> world travels

Jerry Kennedy

Greg Humphrey

Jose Clemente

Will van Treuren

Jess Metcalf

Laura Parfrey-Wagner

Bharath Prithiviraj

Gail Ackermann

Luke Ursell

## [a few of the] Collaborators using UniFrac/QIIME:

Noah Fierer (CU, EEB), Jeff Gordon (Wash U), Ruth Ley (Cornell), Peter Turnbaugh (Harvard), Andrew Gerhart (Emory), Maria Gloria Dominguez (UPR), Rick Bushman, Gary Wu, John Lewis (Penn), Scott Kelley (SDSU), Marty Blaser (NYU), Jacques Ravel (U Maryland), Allan Konopka (PNNL), David Relman (Stanford), Laura Kitz (Harvard), Nori Tacke (Harvard), Andy Goodman (Yale), Jo Handelsman (Yale), Lars Angenent (Cornell), George Church (Harvard)

## Gordon Lab:

anyone, Jennifer, Bob, Meggie, Michelle Smith, Vanessa Ridaura, Alejandro Reyes, J. Paul, Henning Seedorf, Liz Hansen, Nate McNulty, Julie Monchesse, Sabrina Wagoner, Jessica Hoisington-Lopez

## PyCogent:

Govin Huttley (CU) + many others

## Methods:

Joe Jones (US), Eric Navrotsky (HHMI), Sean Eddy (HHMI), Phil Hugenholtz (U Syd), Henry Hoff (CU, CS), Manuel Lladser (CU, APPM), Andy Martin (CU, EEB), Alexis Sigrist (Munich), Steven Brenner (Berkeley), Folker Meyer (ANL), Chris Quince (Glasgow), Payal Gajer (U Maryland), Eytan Rupp (TAU)

## DACC:

Owen White, Jennifer Wortman, Todd Desantis, Gary Andersen, Nikos Kypides, Victor Markowitz, Centers: WashU, Broad, BCM, JCVI;

Manuscript prep: Curtis Huttenhower, David Gevo, HMP Consortium

## ICBP:

Tabinda Ahmed, Rishi Al-Jade, Dinesh Khanna, Mukul Sarker

## More:

Rick Stevens (Argonne), J. Paul (Argonne), Folker Meyer (Argonne), Janet Jansson (LBNL), Ted Fuhrman (USC), Jonathan Eisen (UC Davis), many, many sample donors

## Support (across all projects):

HHMI, NAKFI, USAF, CCFA, Gates IAS, NIDDK, NHGRI, NIGMS, NHLBI, BSF, USDA/NSF, DOE, SCR+Biophysics TGS, C2B2, Butcher Foundation, Sloan IE, NIJ, Keck, Templeton, Moore