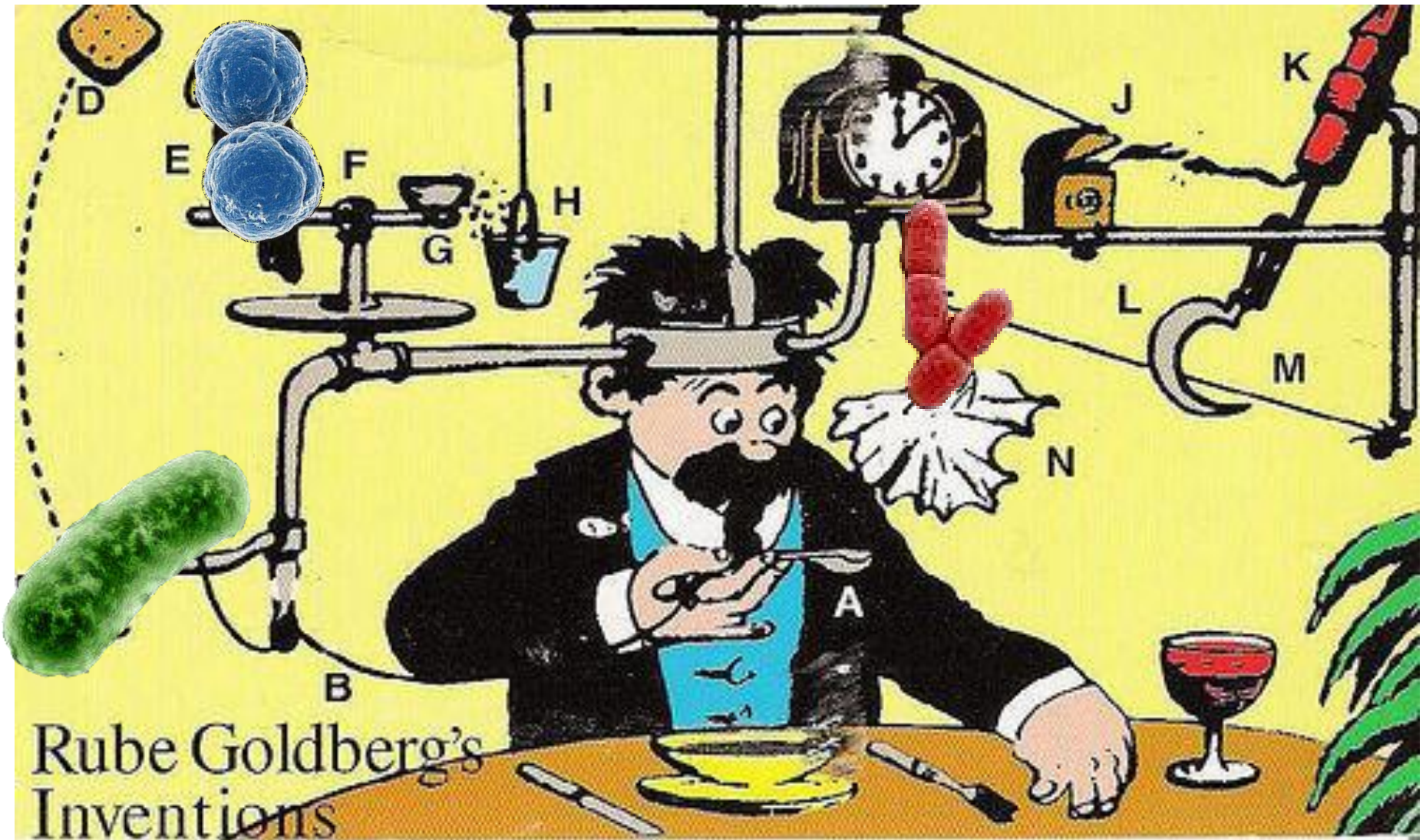


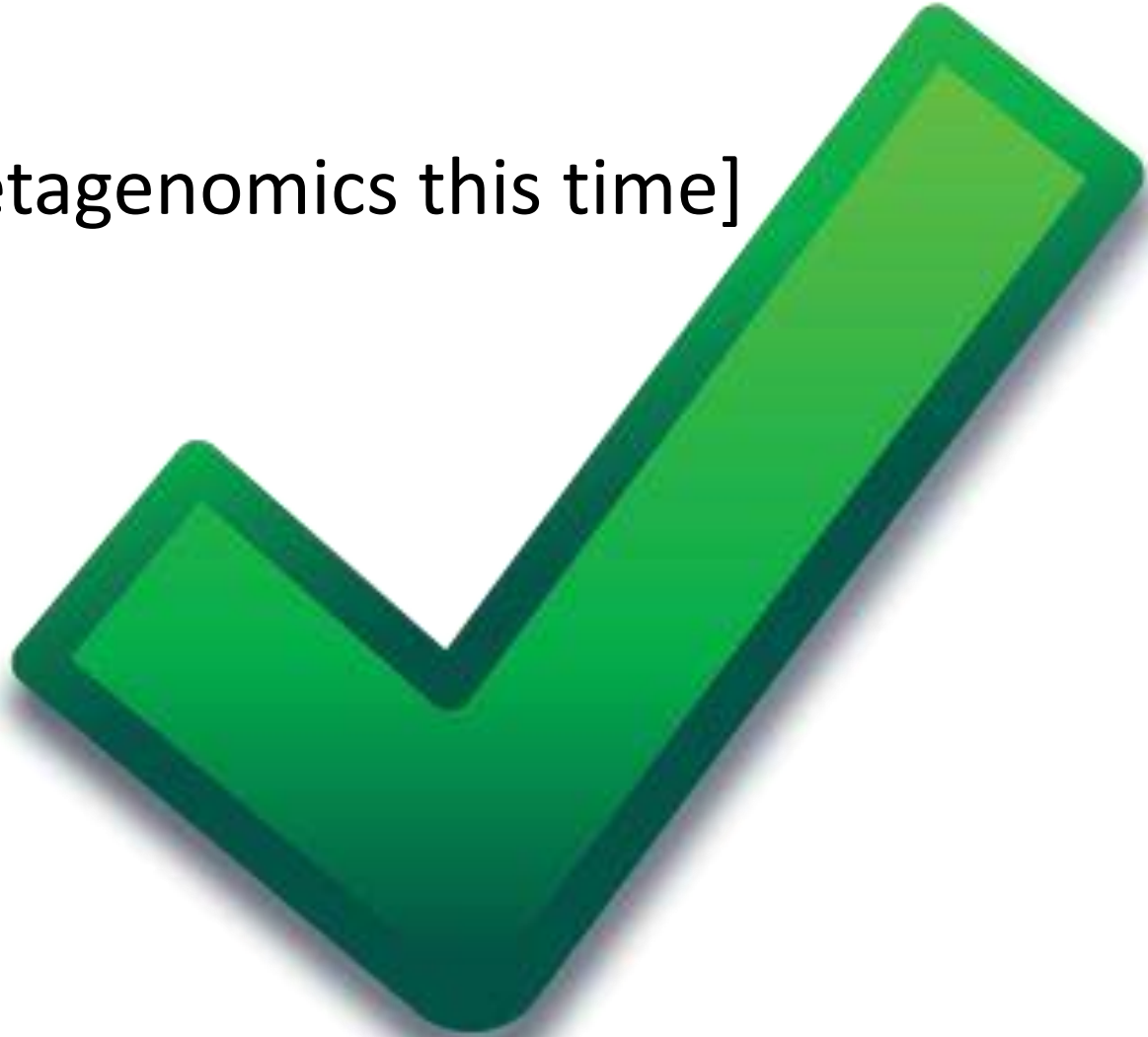
Metagenomics* II:

What are they doing?

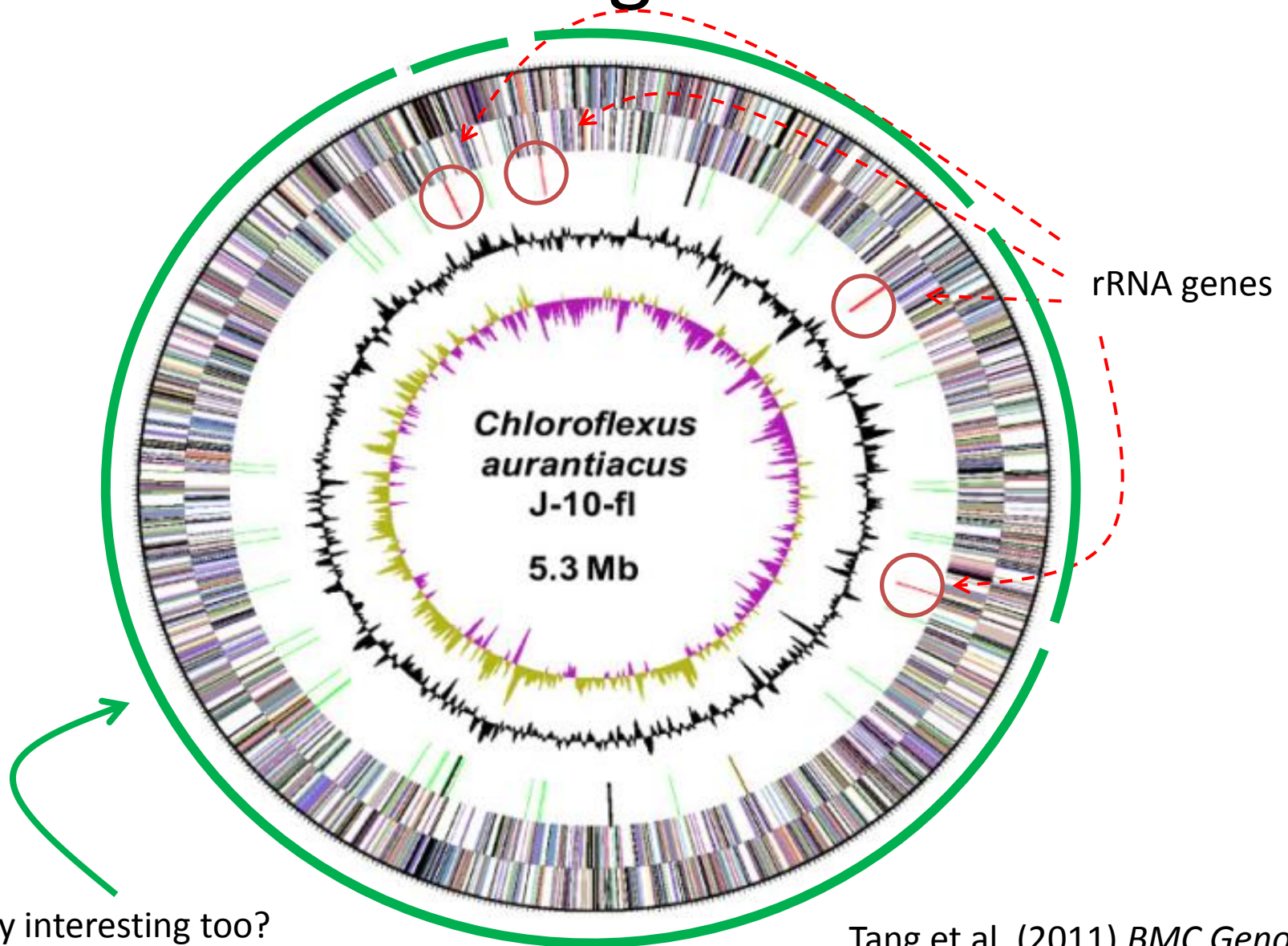


*

[actual metagenomics this time]



Ribosomal RNA genes in context



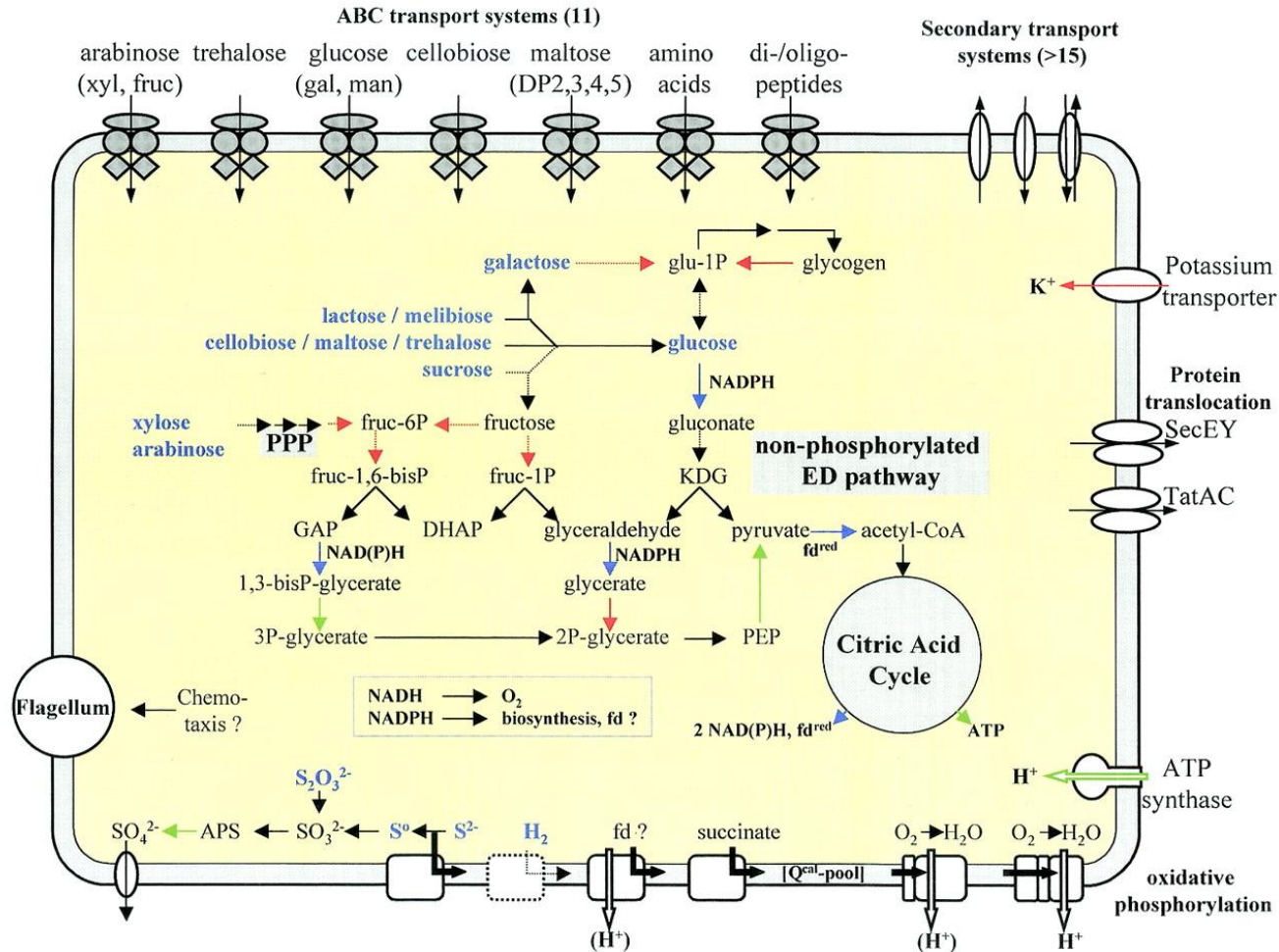
What a bug wants

What a bug needs

The functions we care about derive not from the
PHYLOGENY of organisms, but from the full
repertoire of GENES they possess

Vital questions about the workings of microbes
and microbial communities pertain to what
they can do

What is "function"?

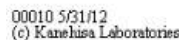


Sulfolobus solfataricus (80°C, pH 2-4)

Archaea; Crenarchaeota; Thermoprotei; Sulfolobales; Sulfolobaceae

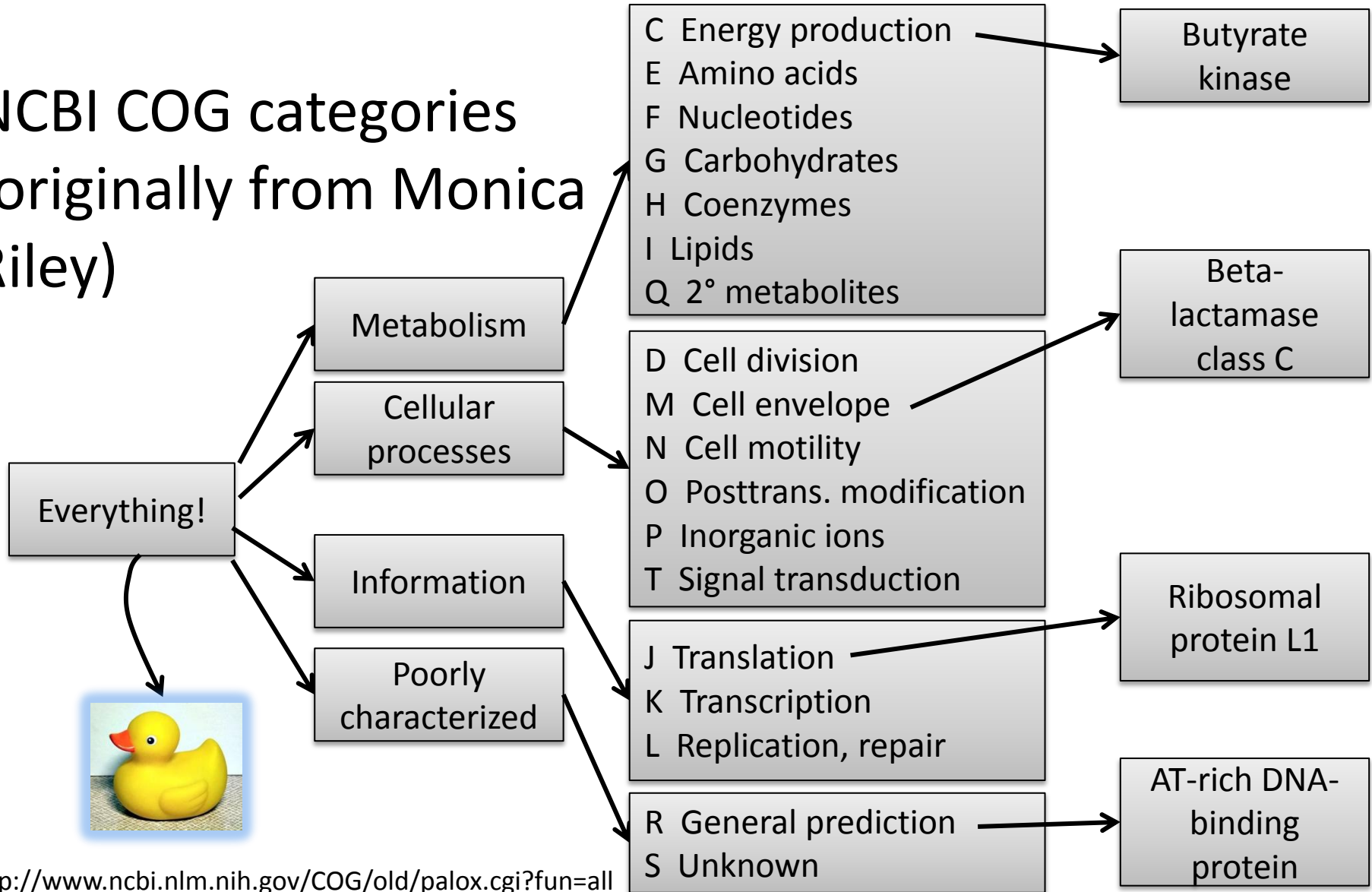
She Q et al. PNAS 2001;98:7835-7840

from KEGG (<http://www.genome.jp>)



Systems for classifying function

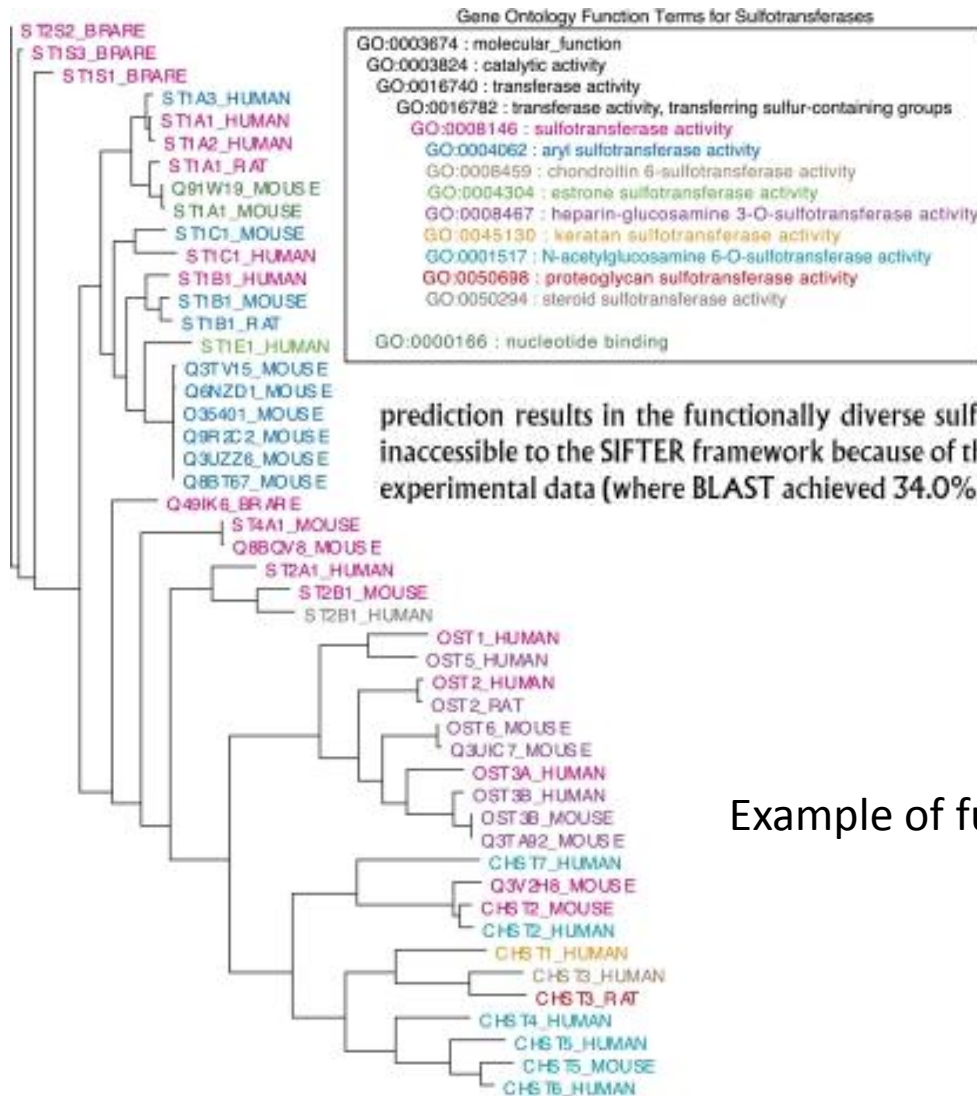
NCBI COG categories
(originally from Monica Riley)



Other systems

- JCVI Role Categories
- Enzyme Commission numbers
- Gene Ontology terms
- Gene names (arbitrary!)
- PFAM functions

Assigning function is (often) HARD



Genome-scale phylogenetic function annotation of large and diverse protein families

Barbara E. Engelhardt,^{1,4,6} Michael I. Jordan,^{1,2} John R. Srouji,^{3,5} and Steven E. Brenner³

¹Electrical Engineering and Computer Science Department, University of California, Berkeley, California 94720, USA; ²Statistics Department, University of California, Berkeley, California 94720, USA; ³Plant & Microbial Biology Department, University of California, Berkeley, California 94720, USA

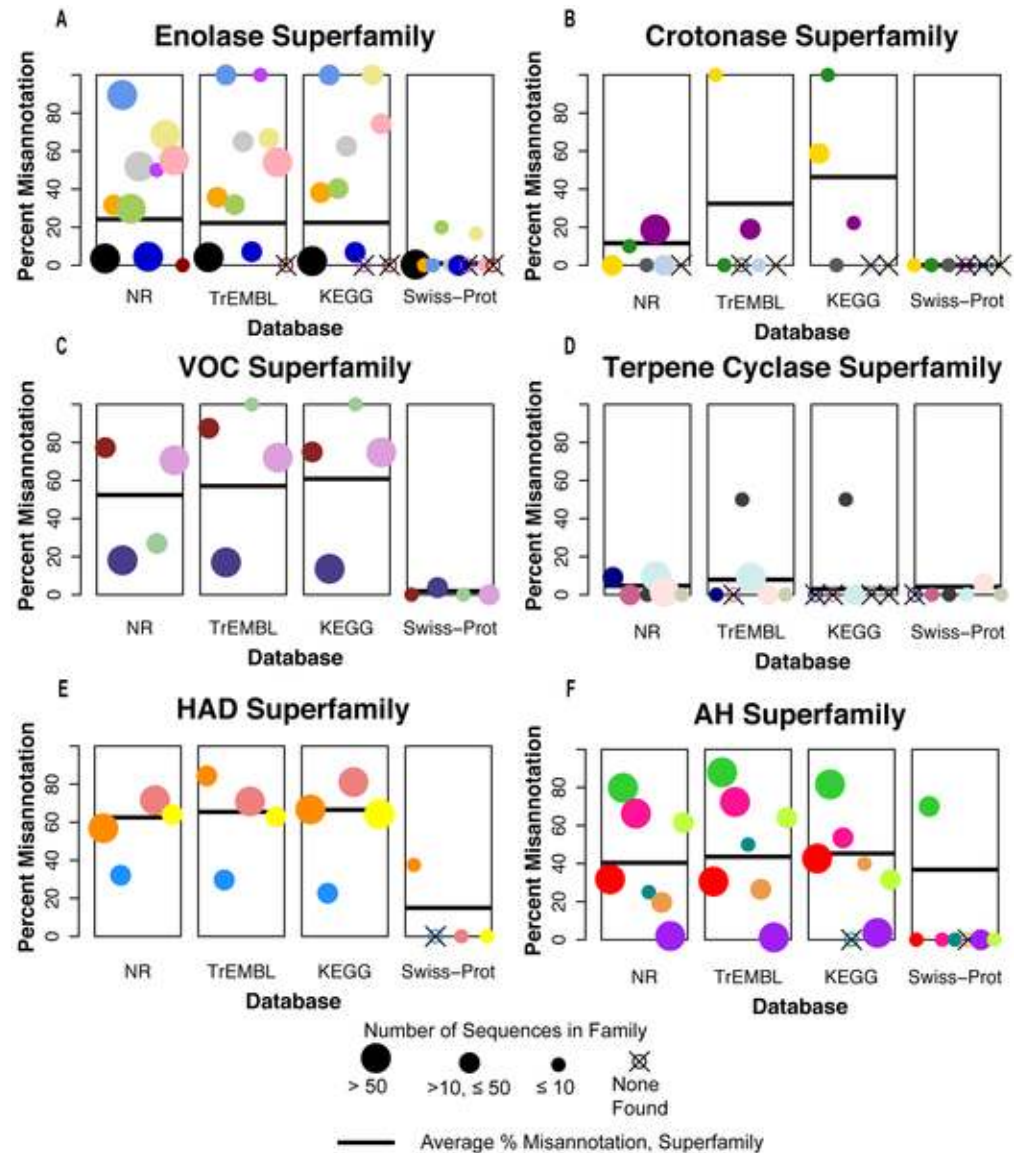
prediction results in the functionally diverse sulfotransferase protein family. On the Nudix protein family, previously inaccessible to the SIFTER framework because of the 66 possible molecular functions, SIFTER achieved 47.4% accuracy on experimental data (where BLAST achieved 34.0%). Finally, we used SIFTER to annotate all of the *Schizosaccharomyces pombe*

Example of functional shifts in homologous proteins

Accuracy in reference databases
is often poor

Example: % misannotation of
protein superfamilies in different
public databases

Relationship between prediction
method and accuracy



Rare validation and transitive annotation

"Multiple types of 'transitive annotation error' can occur during such propagation of putative function, including overly specific annotation, founder effects that obscure functional diversity in large families such as radical SAM, daisy-chain inference that passes through non-overlapping regions of a multidomain protein and faults from successive rounds of reinterpretation of an original protein name."



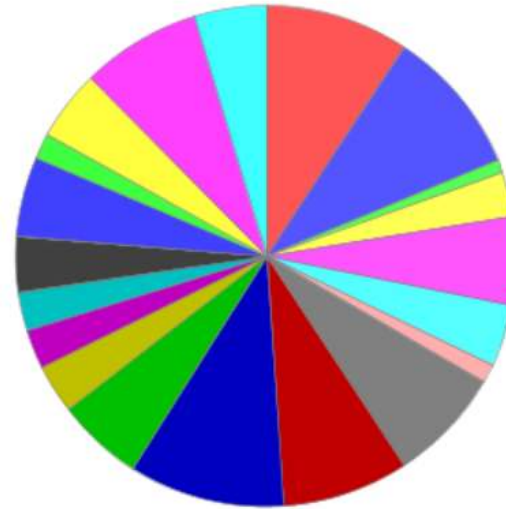
Experimentally
characterized

Madupu et al. (2012) *Nucleic Acids Res*

Proteins of UNKNOWN function

http://img.jgi.doe.gov/cgi-bin/w/main.cgi?section=TaxonDetail&page=cogs&cat=cat&taxon_oid=646311926

COG Categories	Gene Count	% of Total (3905)
Amino acid transport and metabolism	363	9.30
Carbohydrate transport and metabolism	368	9.42
Cell cycle control, cell division, chromosome partitioning	33	0.85
Cell motility	111	2.84
Cell wall/membrane/envelope biogenesis	226	5.79
Coenzyme transport and metabolism	153	3.92
Defense mechanisms	49	1.25
Energy production and conversion	287	7.35
Function unknown	318	8.14
General function prediction only	394	10.09
Inorganic ion transport and metabolism	217	5.56
Intracellular trafficking, secretion, and vesicular transport	124	3.18
Lipid transport and metabolism	100	2.56
Nucleotide transport and metabolism	97	2.48
Posttranslational modification, protein turnover, chaperones	138	3.53
Replication, recombination and repair	200	5.12
RNA processing and modification	2	0.05
Secondary metabolites biosynthesis, transport and catabolism	64	1.64
Signal transduction mechanisms	178	4.56
Transcription	301	7.71
Translation, ribosomal structure and biogenesis	182	4.66
Not in COGs	1000	22.24



E. Coli K-12 MG1655 – The Model Organism's Model Organism

318 "function unknown"

394 "general function prediction only"

1000 "not in COGs"

= 1712 unknown or poorly characterized

*This is **40.47%** of all predicted protein-coding genes!!*

How to predict protein function



Clues

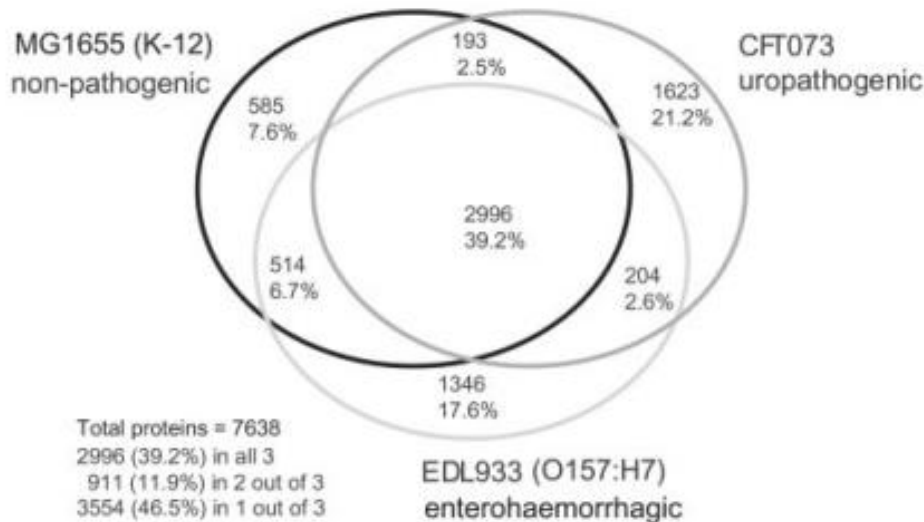
- Best-match homology
- Phylogeny
- Domains, motifs
- Structure
- Phylogenetic profiles
- Protein-protein interaction
- Operons
- Transmembrane predictions
- Pathway completion

The limitations of phylogeny / taxonomy

- Function obviously correlates to some degree with taxonomy
- Closely related groups share much in common, but this fades as deeper and deeper relationships are examined
- Even closely related organisms may differ in key properties of interest

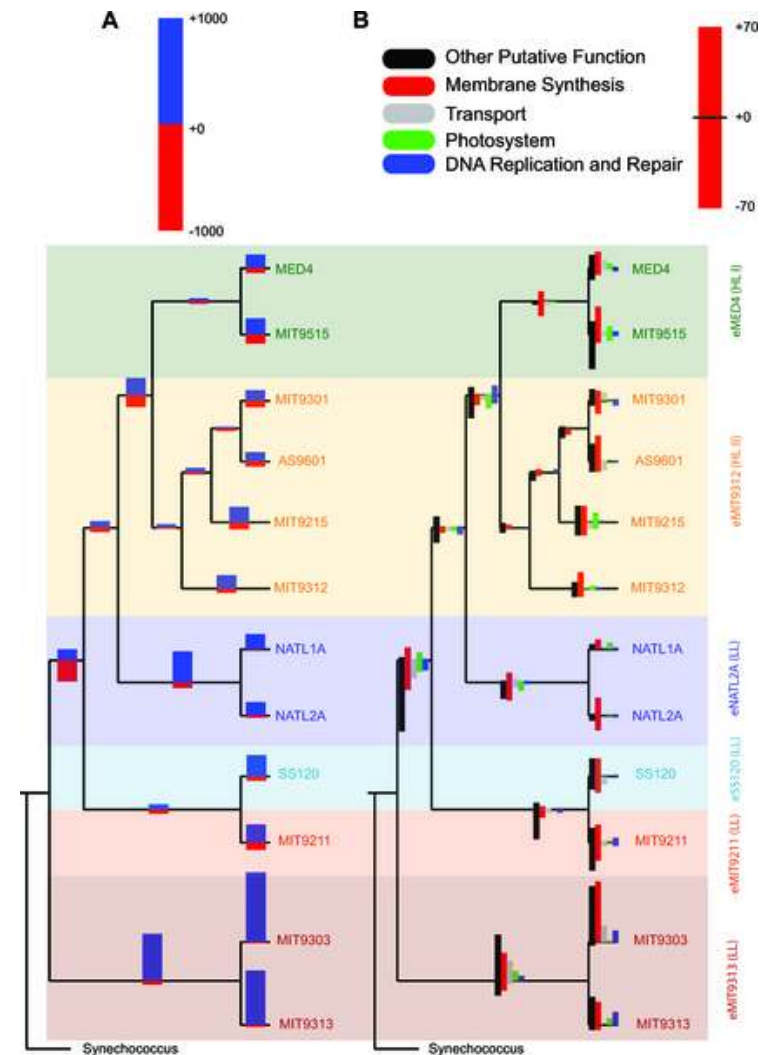
Escherichia

- Facultative anaerobe, rod-shaped
- "Enteric" except when they're not
 - Harmless commensal: *E. coli* K-12 MG1655
 - Human uropathogen: *E. coli* CFT073
 - Enterohaemorrhagic: *E. coli* EDL933



Prochlorococcus marinus

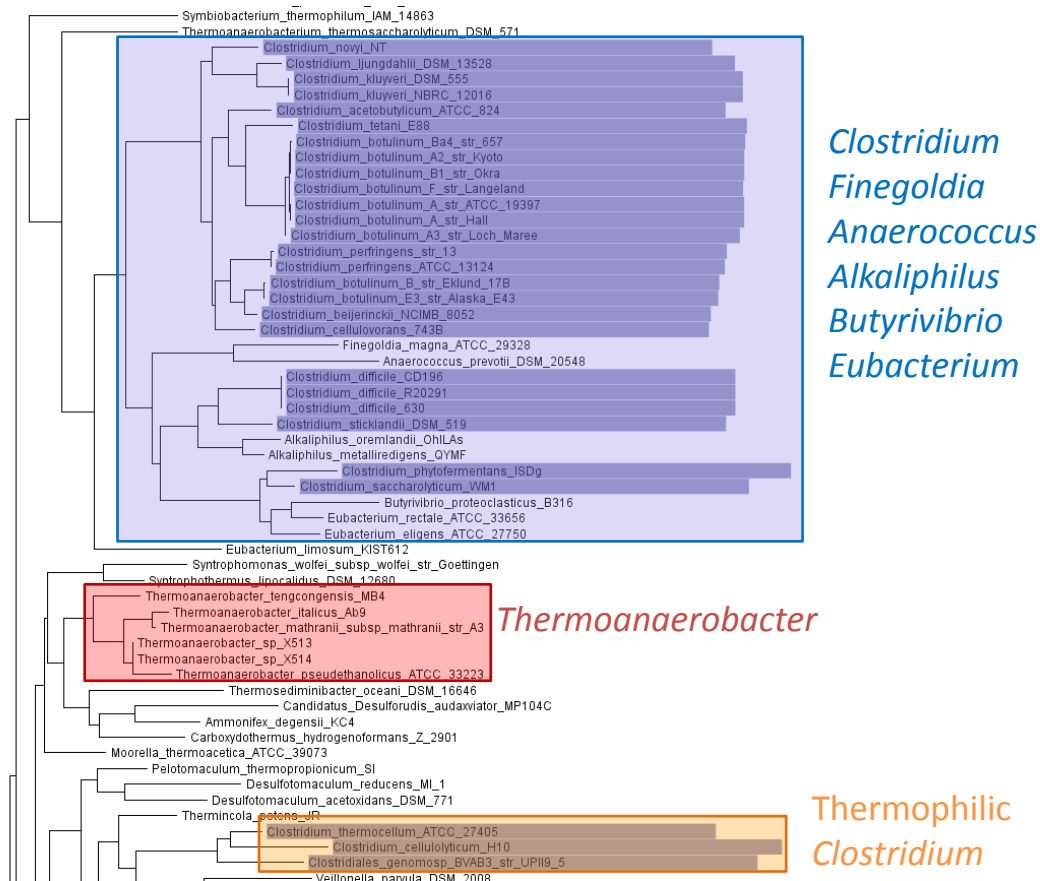
- Oxygenic phototrophs, divinyl chlorophyll pigments, chlorophyll-binding proteins
- High-light OR low-light adapted
- GC content range: 30-60%
- Genome size range: 1.7 – 2.4 Mb



Kettler et al. (2007) *PLoS Genet*

Clostridium

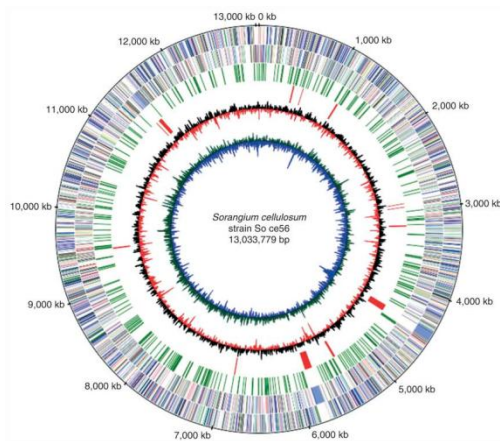
- Ha, ha, ha, ha



What's going on here?

- Taxonomy is messy to begin with. But evolutionary processes lead to divergence

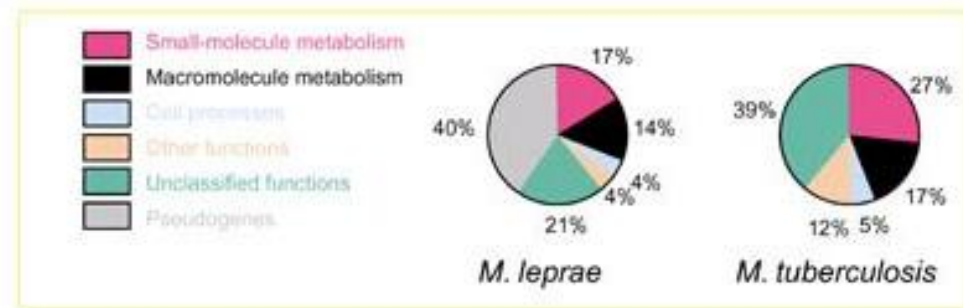
Gene GAIN through invention and duplication



Sorangium cellulosum: 13 Mb, ~10K genes
(~48% hypothetical)

Schneiker et al. (2007) *Nat Biotechnol*

Gene LOSS:
"use it or lose it"

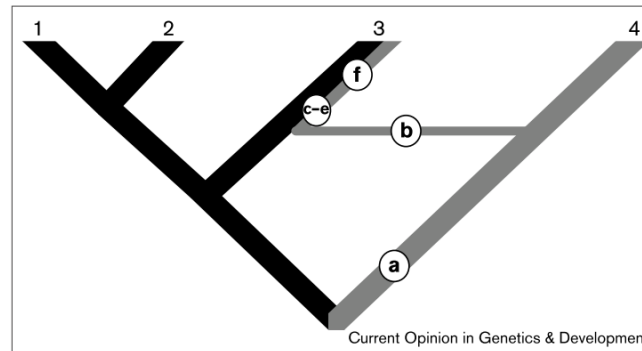


Genes ~2700 ~4000

M. leprae: unculturable for 140 years

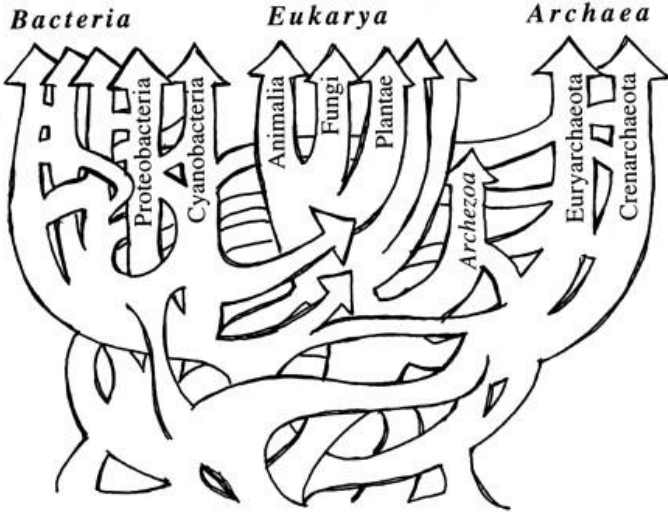
Cole et al. (2001) *Nature*

Lateral gene transfer

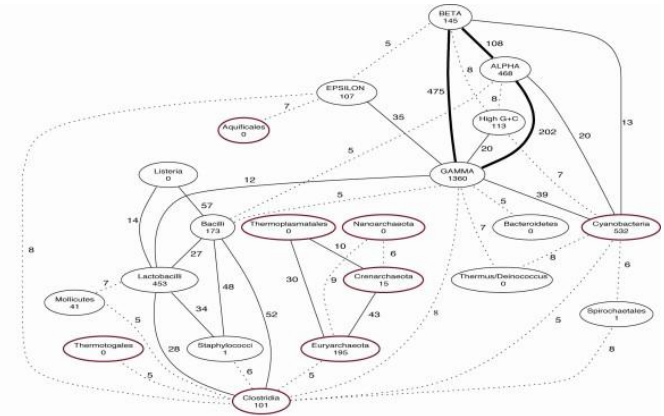
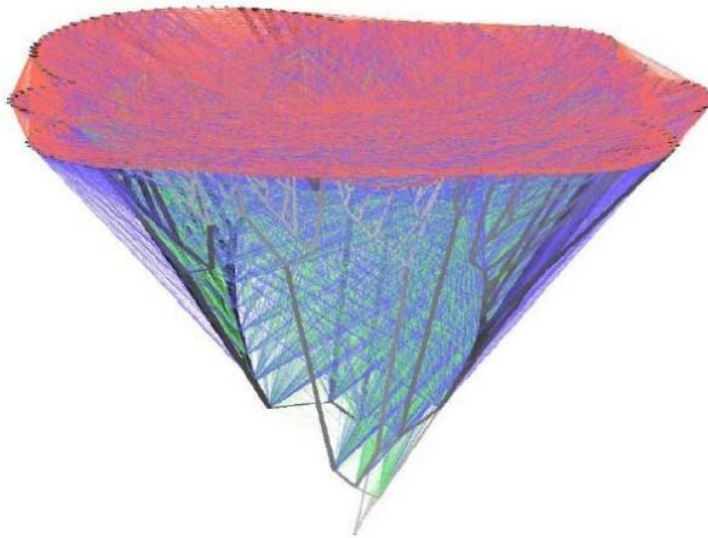


Eisen (2000) *Curr Opin Genes Dev*

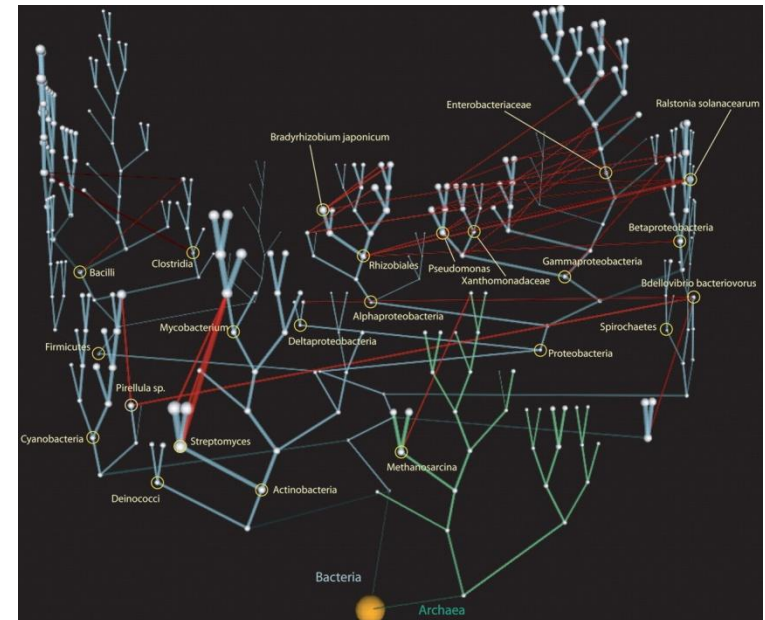
The net of life



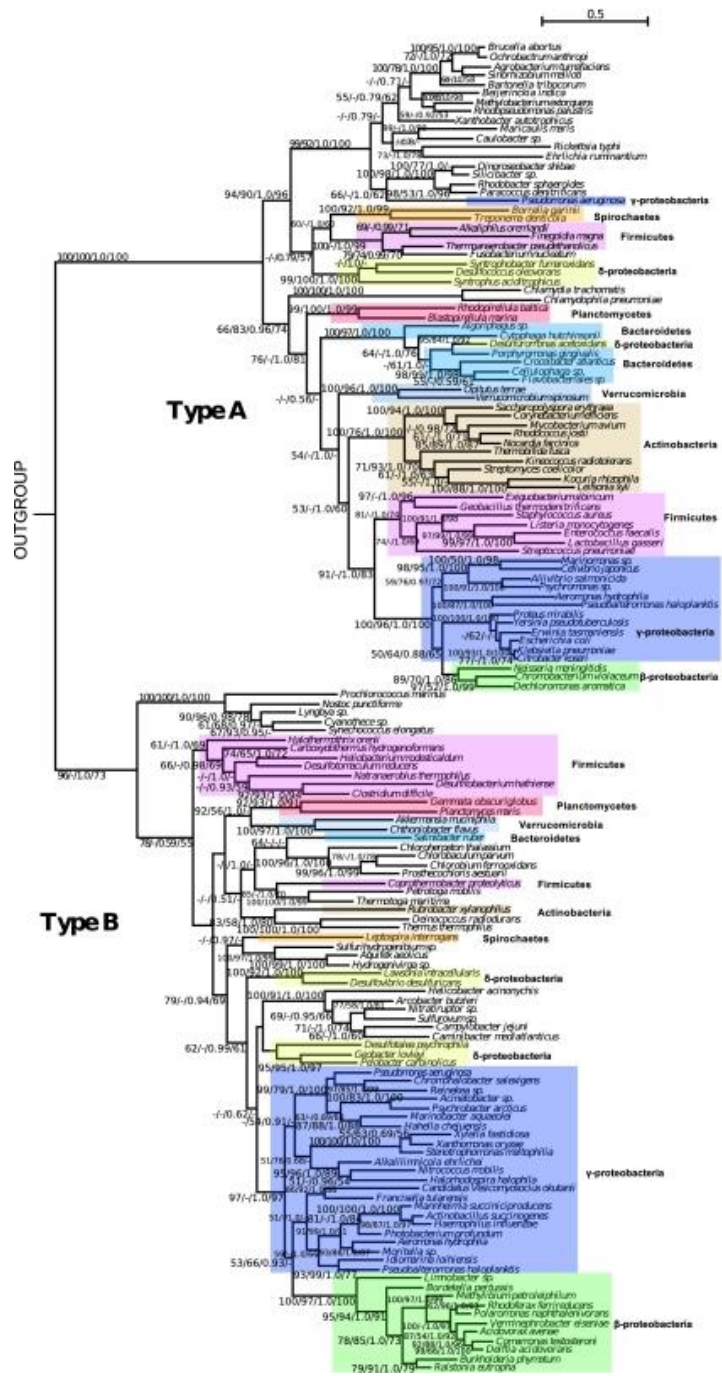
Doolittle (1999) *Science*

Beiko et al. (2005) *PNAS*

Dagan and Martin (2008) *PNAS*



Kunin et al. (2005) *Genome Res*

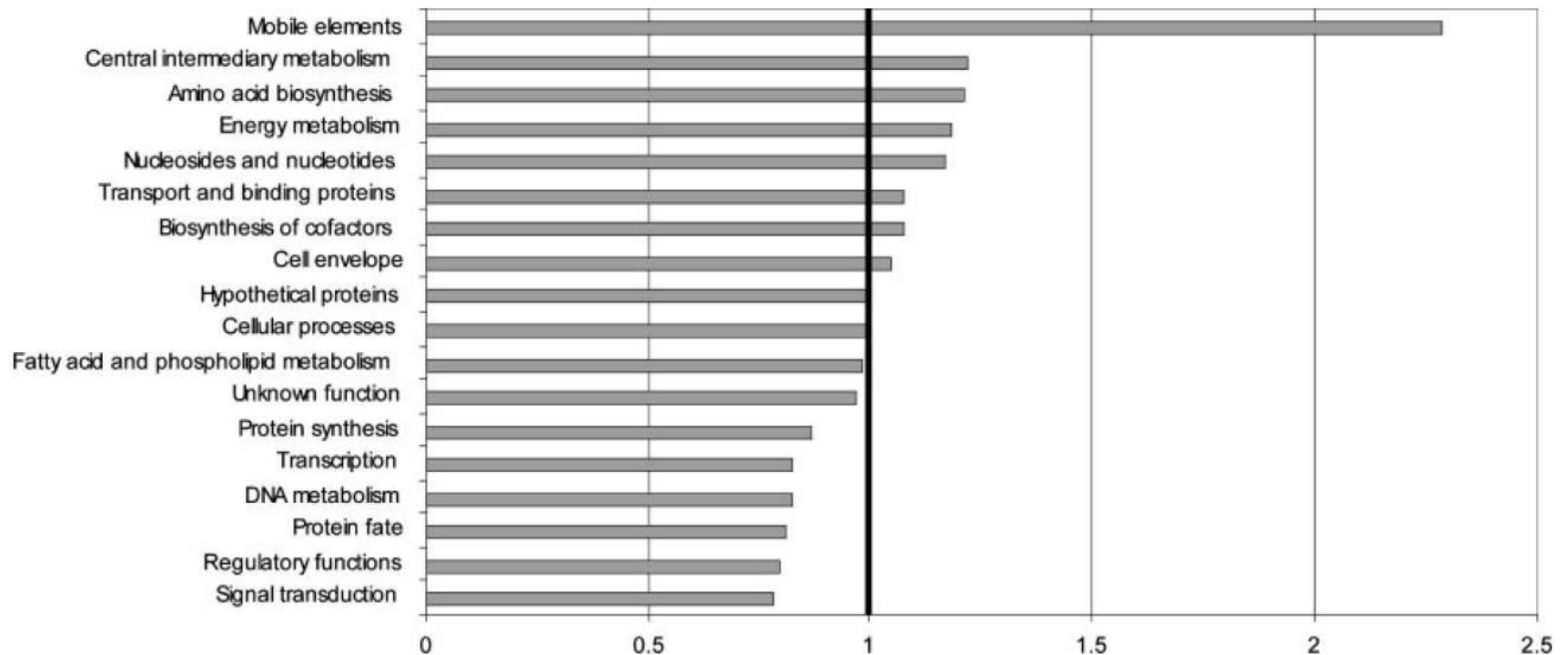


Tyrosyl-tRNA synthetases

Type A and Type B have nearly parallel phylogenies

Presence of both in one lineage is extremely rare

"Transferability" of different types of gene



aaaaaaaaand...

JOURNAL OF BACTERIOLOGY, Dec. 2003, p. 7241–7246
0021-9193/03/\$08.00+0 DOI: 10.1128/JB.185.24.7241-7246.2003
Copyright © 2003, American Society for Microbiology. All Rights Reserved.

Vol. 185, No. 24

Horizontal Transfer of Segments of the 16S rRNA Genes between Species of the *Streptococcus anginosus* Group

Leo M. Schouls,^{1*} Corrie S. Schot,¹ and Jan A. Jacobs²

Laboratory for Vaccine-Preventable Diseases, National Institute of Public Health and the Environment, Bilthoven,¹
and Department of Medical Microbiology, University Hospital of Maastricht, Maastricht,² The Netherlands

Catriona Harrington, Antonio Del Casale,
Jonathan Kennedy, Horst Neve,
Bernard E. Picton, Marlies J. Mooij,
Fergal O’Gara, Leonid A. Kulakov,
Michael J. Larkin, and Alan D. W. Dobson
**Evidence of bacteriophage-mediated
horizontal transfer of bacterial 16S
rRNA genes in the viral metagenome
of the marine sponge *Hymeniacidon
perlevis***

Microbiology November 2012
158:2789–2795; published ahead of print
August 17, 2012,

JOURNAL OF BACTERIOLOGY, Sept. 2005, p. 6106–6118
0021-9193/05/\$08.00+0 doi:10.1128/JB.187.17.6106-6118.2005
Copyright © 2005, American Society for Microbiology. All Rights Reserved.

Vol. 187, No. 17

Discordant 16S and 23S rRNA Gene Phylogenies for the Genus *Helicobacter*: Implications for Phylogenetic Inference and Systematics

Floyd E. Dewhirst,^{1,2†} Zeli Shen,^{3†} Michael S. Scimeca,³ Lauren N. Stokes,¹
Tahani Boumenna,¹ Tsute Chen,¹ Bruce J. Paster,^{1,2}
and James G. Fox^{3*}

Extremophiles (2000) 4:373–376

© Springer-Verlag 2000

ORIGINAL PAPER

Gudrun Amann · Karl O. Stetter · Enric Llobet-Brossa
Rudolf Amann · Josefa Antón

**Direct proof for the presence and expression of two 5% different 16S rRNA
genes in individual cells of *Haloarcula marismortui***

"Phylogeny" vs. function

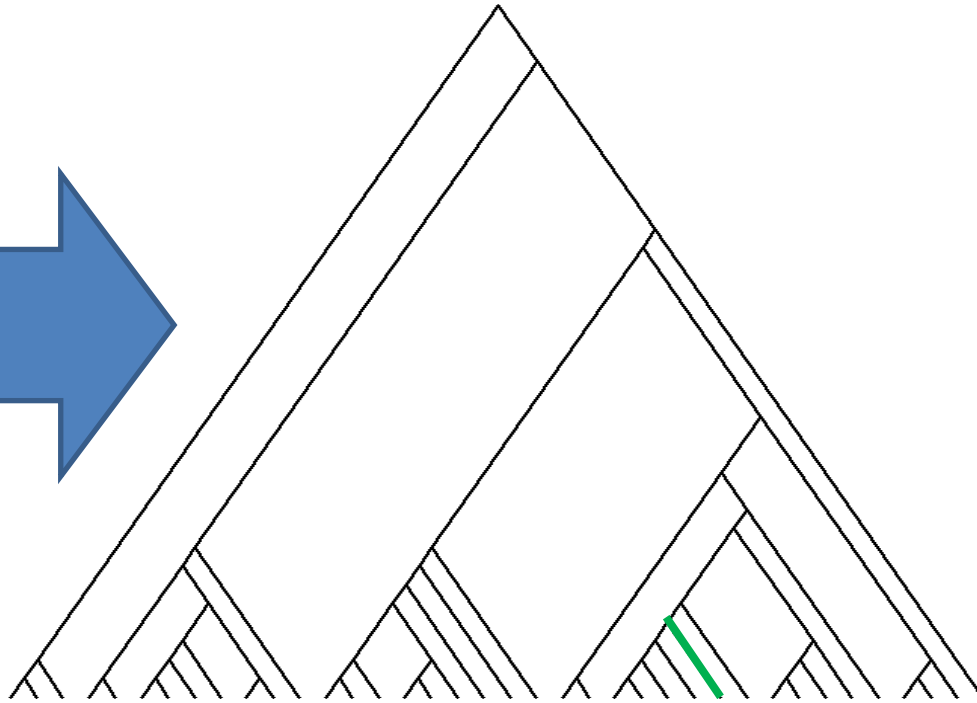
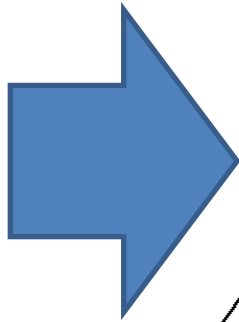
PICRUSt: Trying to do exactly this

- Starting with:
 - A **marker gene** sample (typically 16S)
 - A set of reference sequenced genomes, with identified **marker genes** and predicted **protein-coding genes**
 - A phylogeny of reference **marker genes**
- Try to predict **the metagenome**

Langille MGI*, Zaneveld J*, Caporaso JG, McDonald D, Knights D, Reyes JA, Clemente JC, Knight R, Beiko RG, Huttenhower C. Submitted to *Nature Biotechnology* (I think)

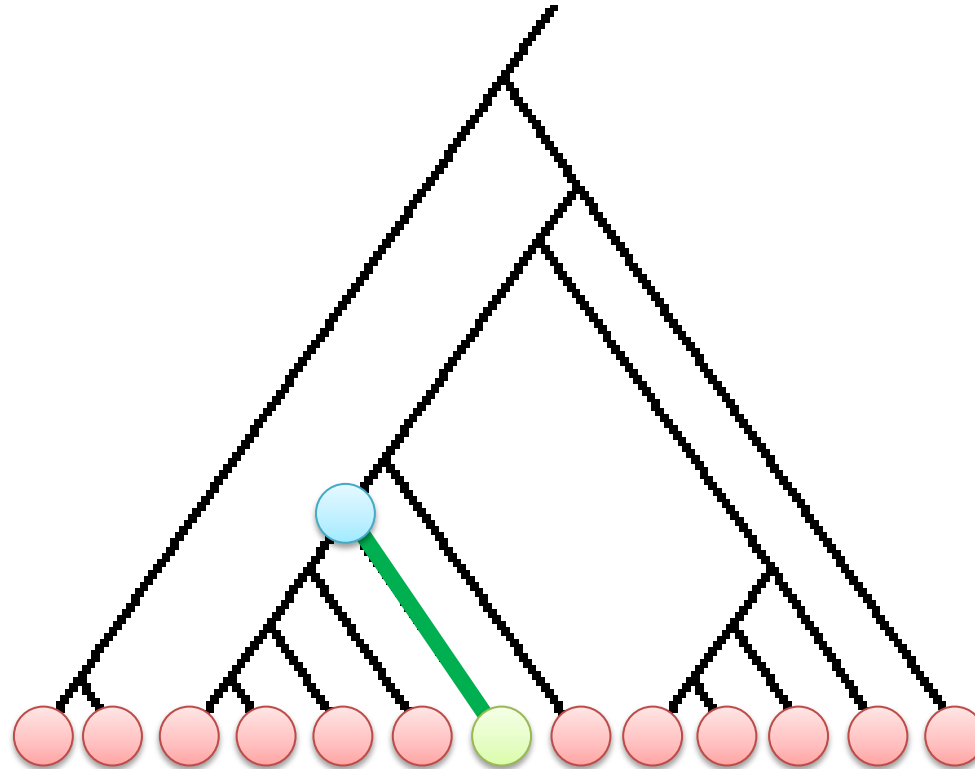
The idea

Reference
16S tree



Sampled
16S sequence

The idea

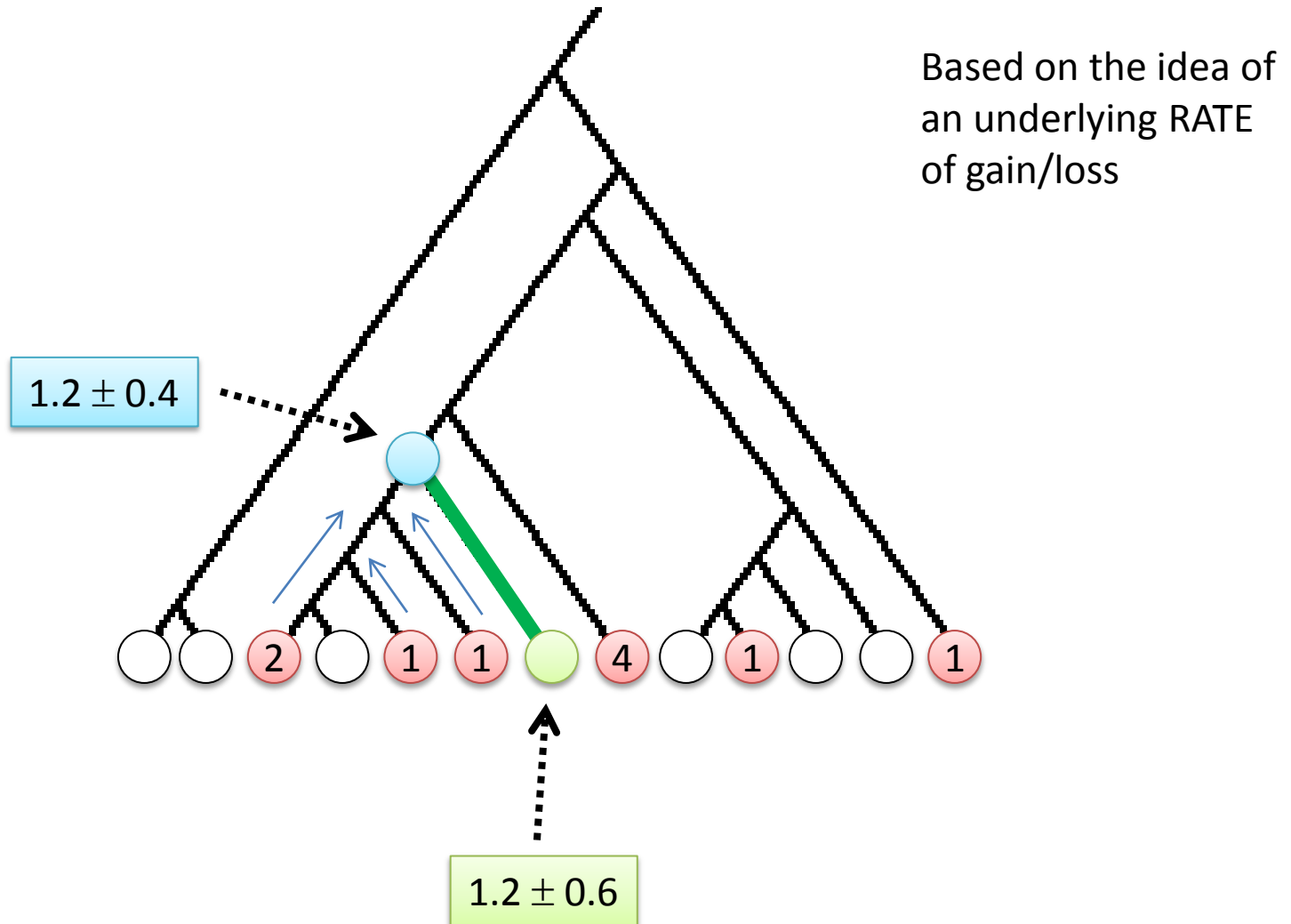


Use **KNOWN GENE CONTENTS**

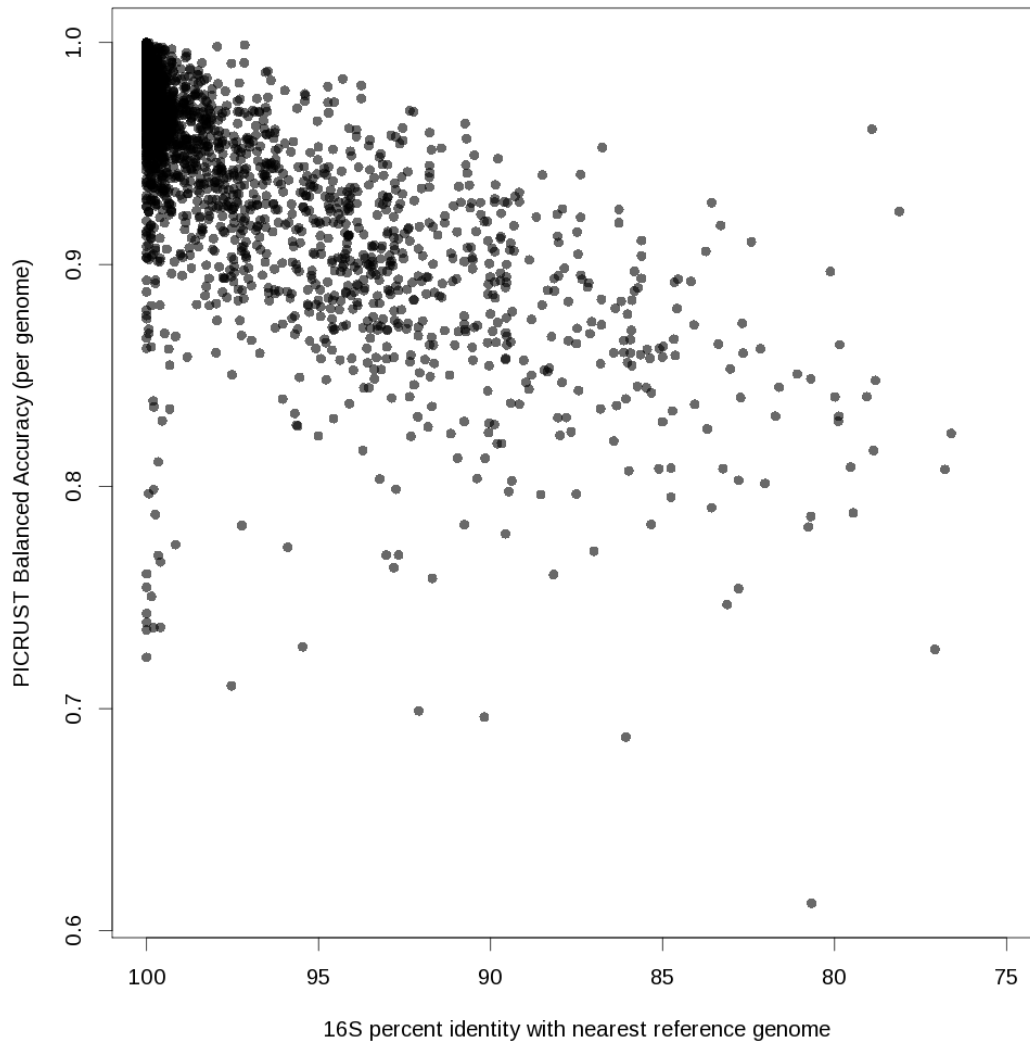
To infer **ANCESTRAL STATE**

In order to predict **FUNCTIONAL GENES** in the sampled organism

Predicting the abundance of a single functional gene



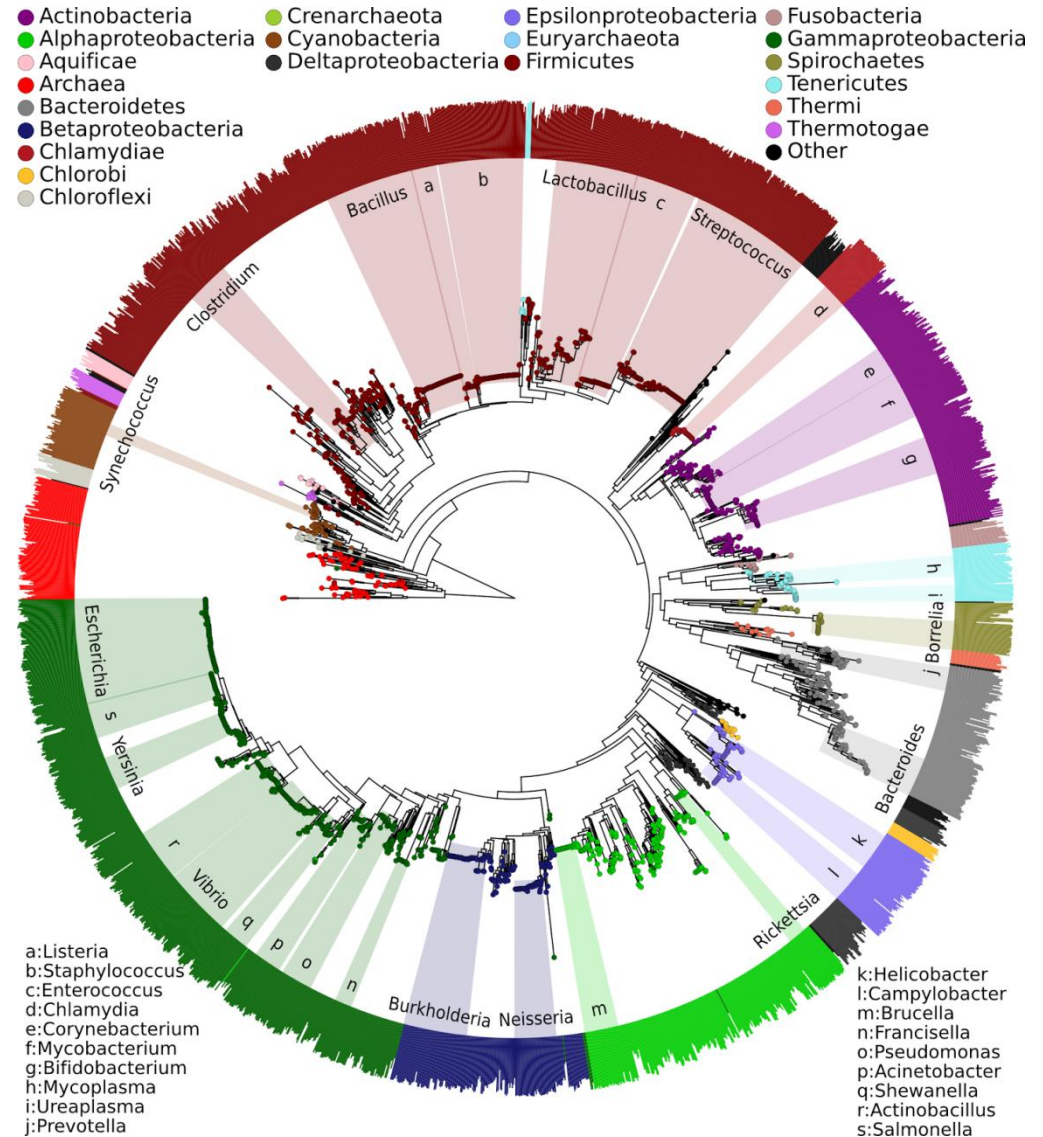
How well does PICRUSt work?



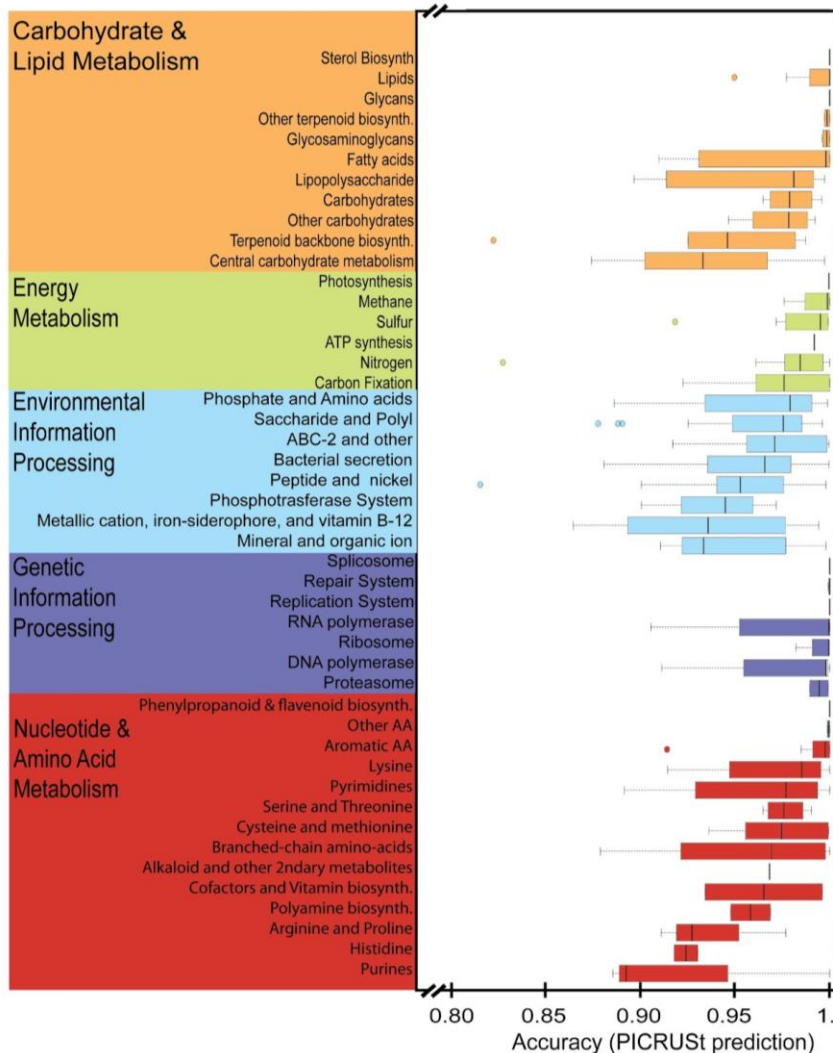
Predicting the content of sequenced genomes from that of every other sequenced genome

How well does PICRUSt work?

Accuracy is good across Bacteria and Archaea, 'weirdo' reduced genomes give worst accuracy



How well does PICRUSt work?



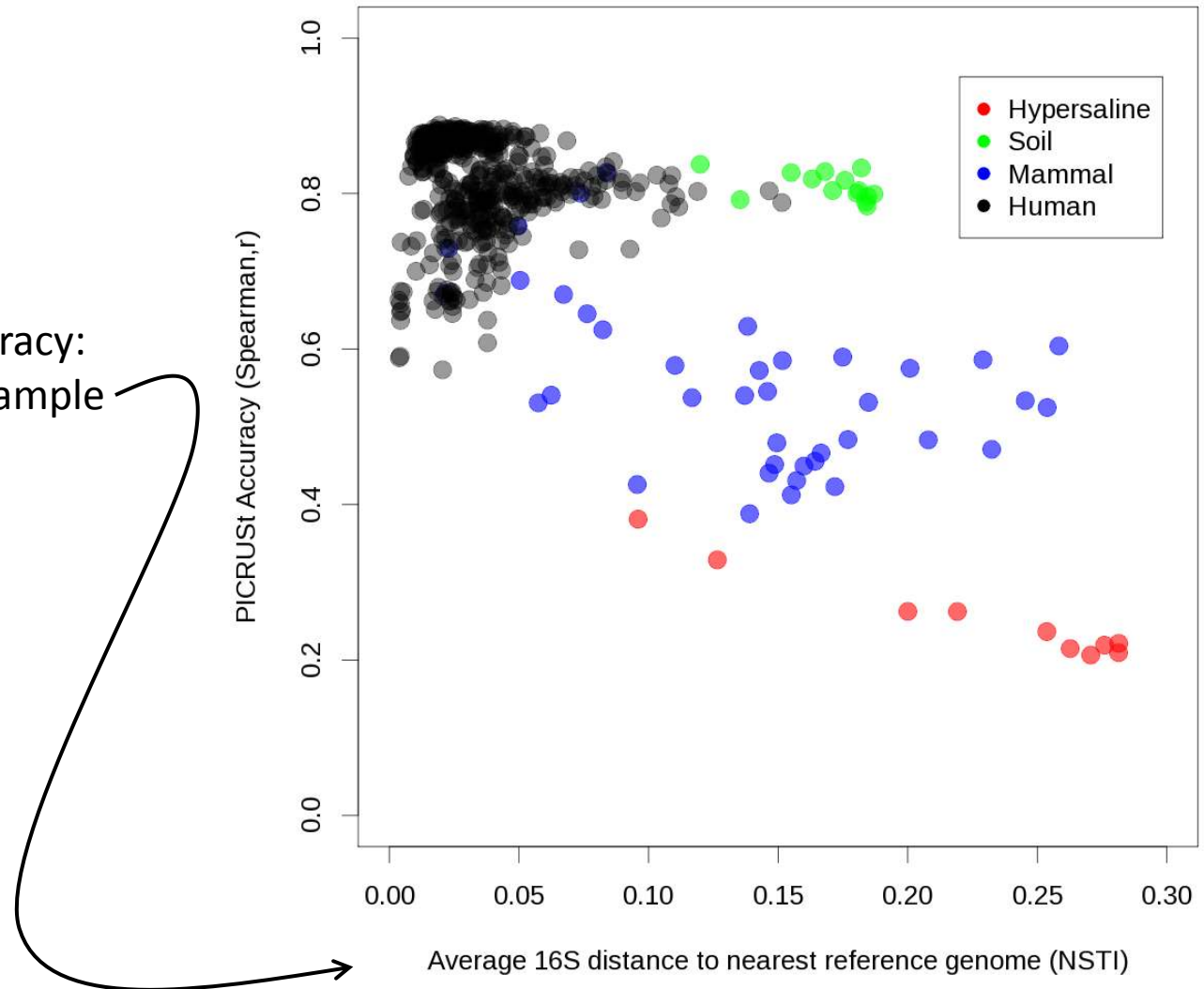
Accuracy by function

Worst accuracy is:

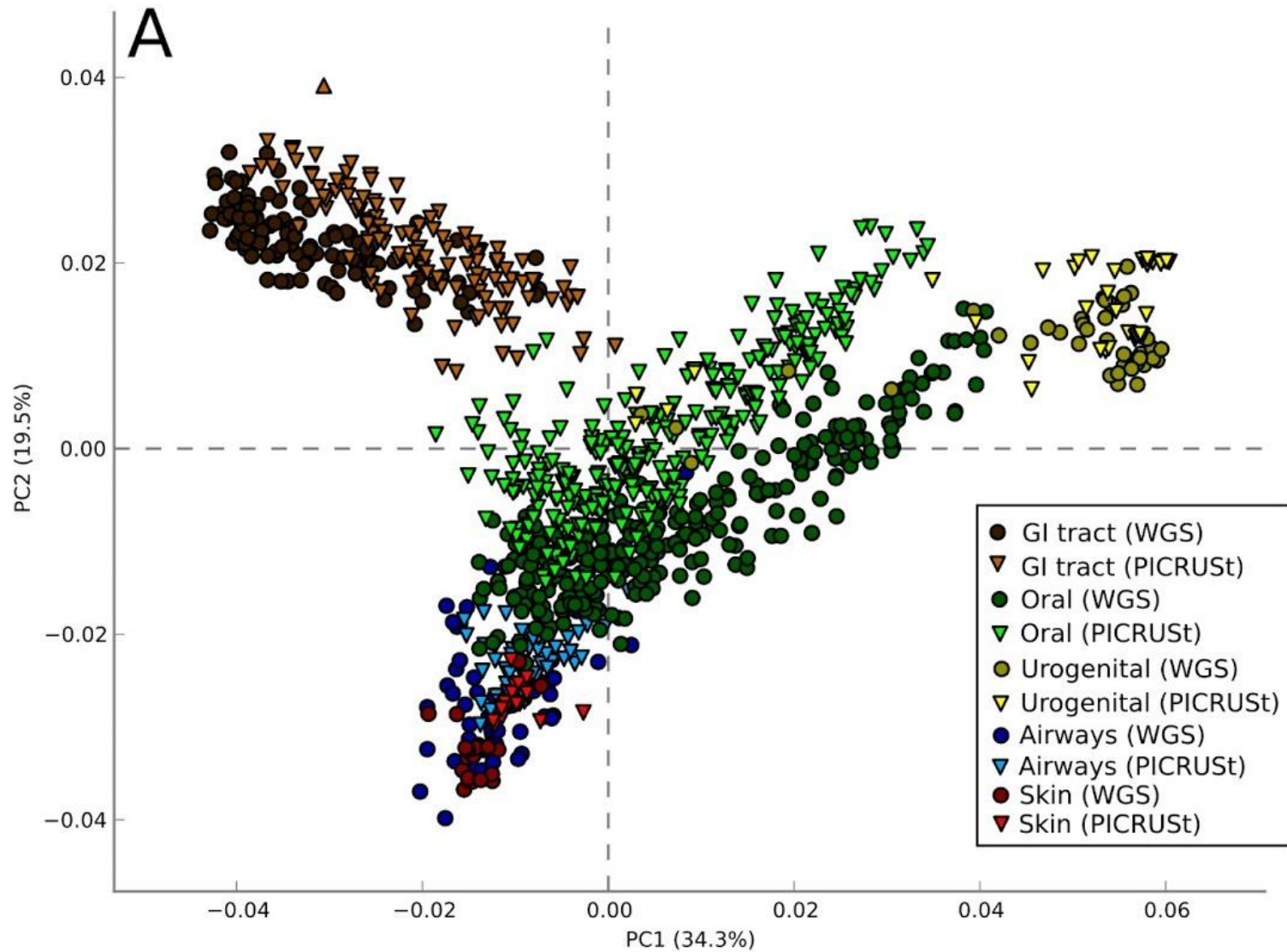
- Environmental information processing
- Central carbohydrate metabolism (?)
- Purines (??)

PICRUSt on metagenomes

Factors influencing accuracy:
-Taxonomic novelty of sample
- Sequencing depth



PICRUSt on metagenomes



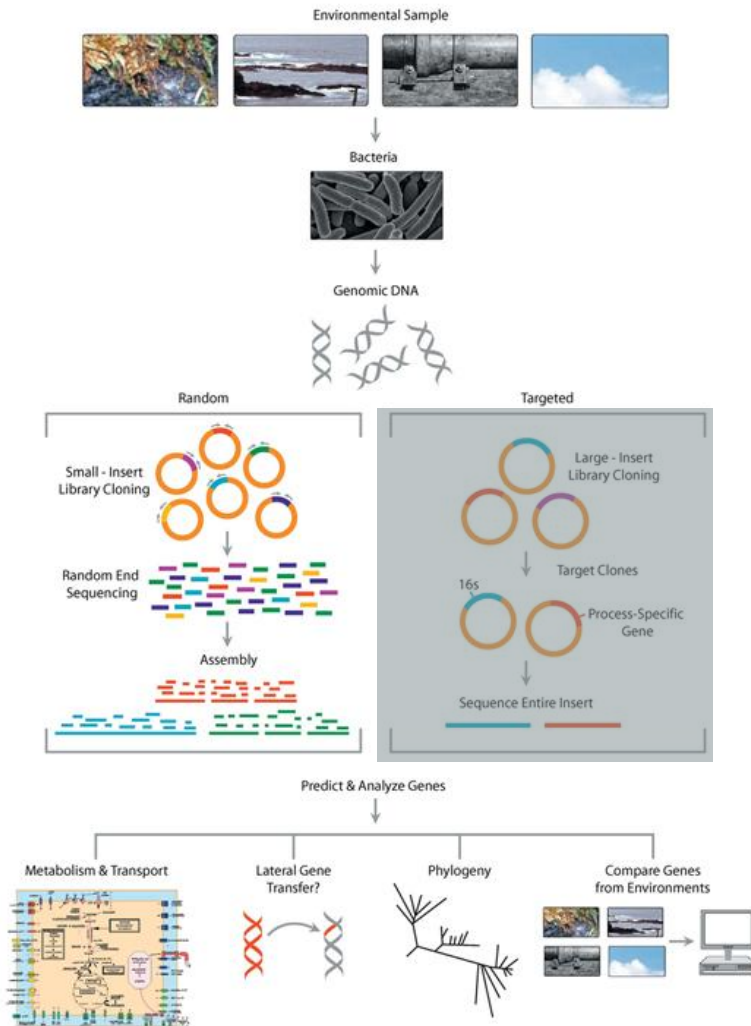
PICRUSt summary

- It works well enough to be useful (!), can recapture / discover information from metagenomic project
- Success depends on several factors, but it actually **outperforms** low-coverage WGS
- Not a replacement for actual WGS, but complementary

From "who is there" to "what are they doing"

- So we want to characterize the functional complement of a microbial "community"
- How can we do this?
 - Culture and characterize
 - Extrapolate (PICRUSt)
 - **Metagenomic sequencing**

Metagenomic sequencing



← Extract microorganisms

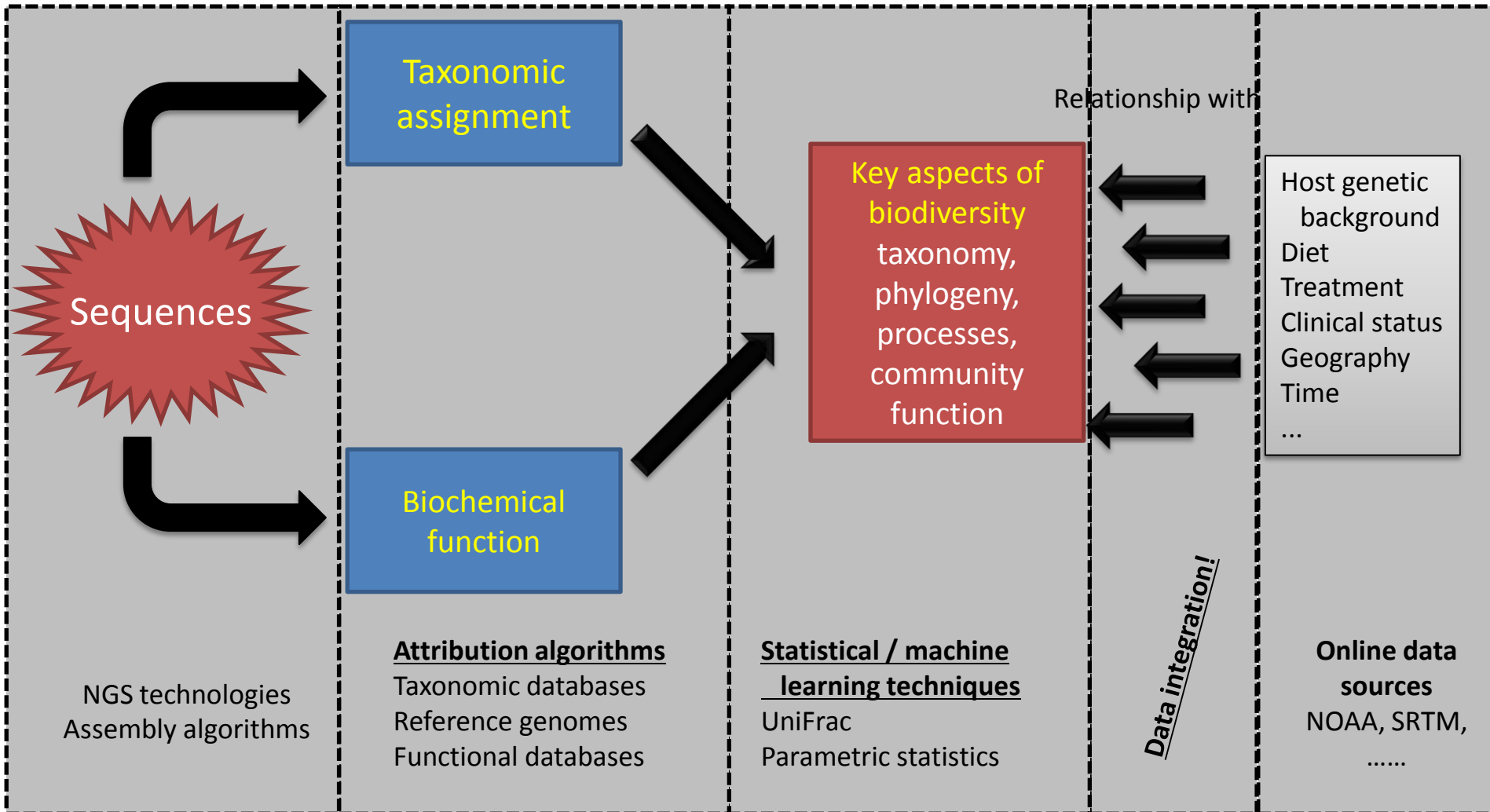
← Extract DNA

← Clone library construction

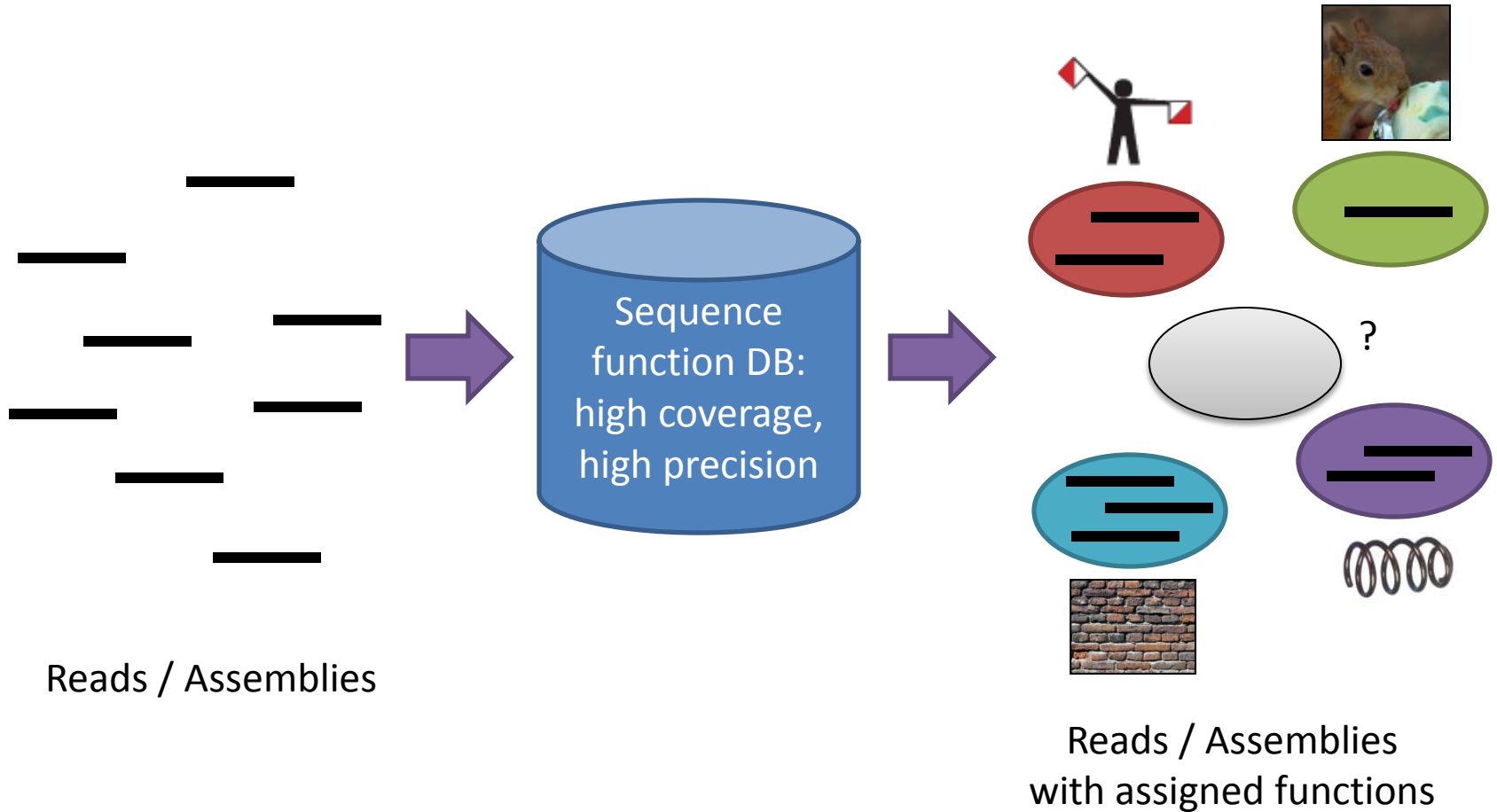
← Sequencing

Assembly: maybe!

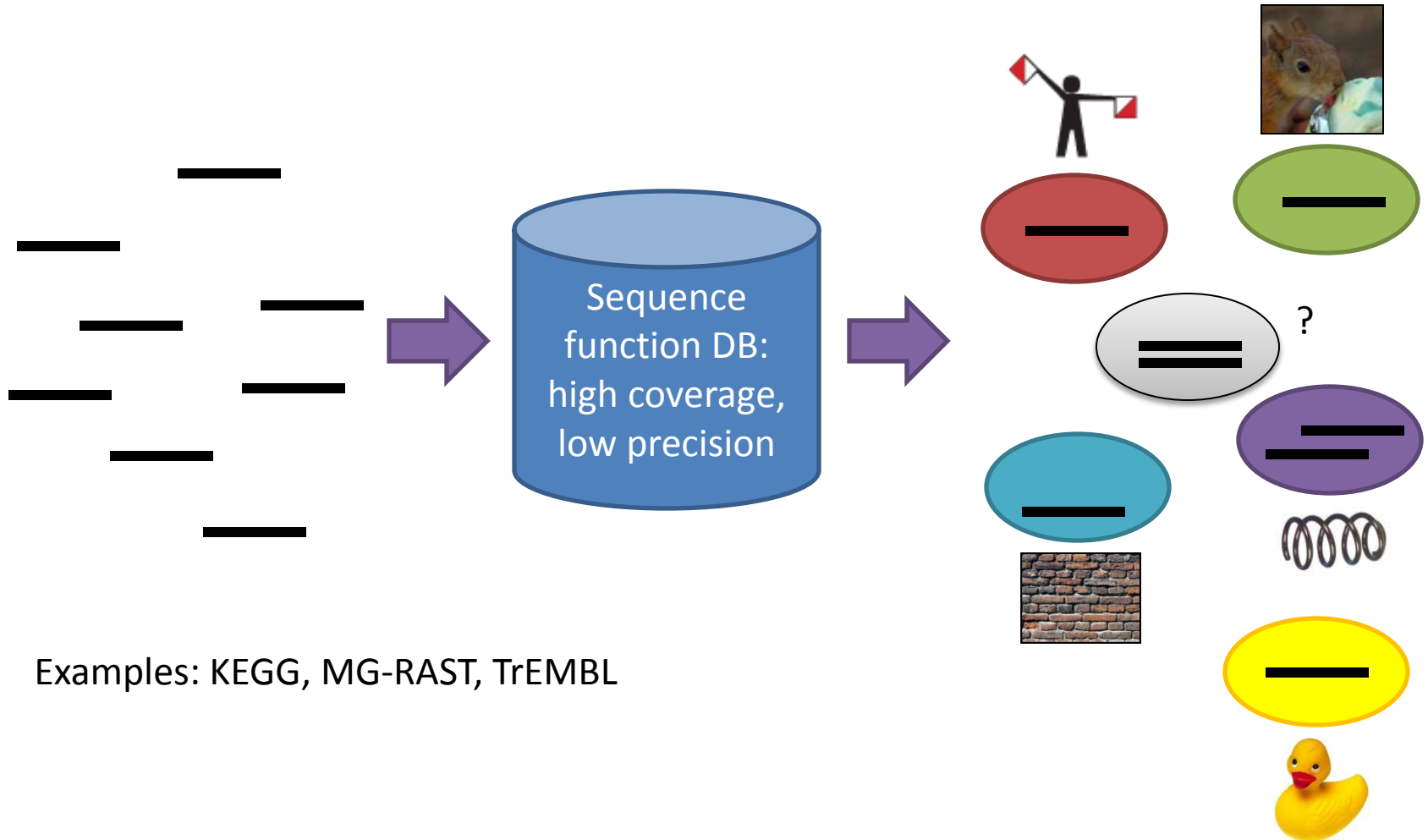
Metagenomic data analysis



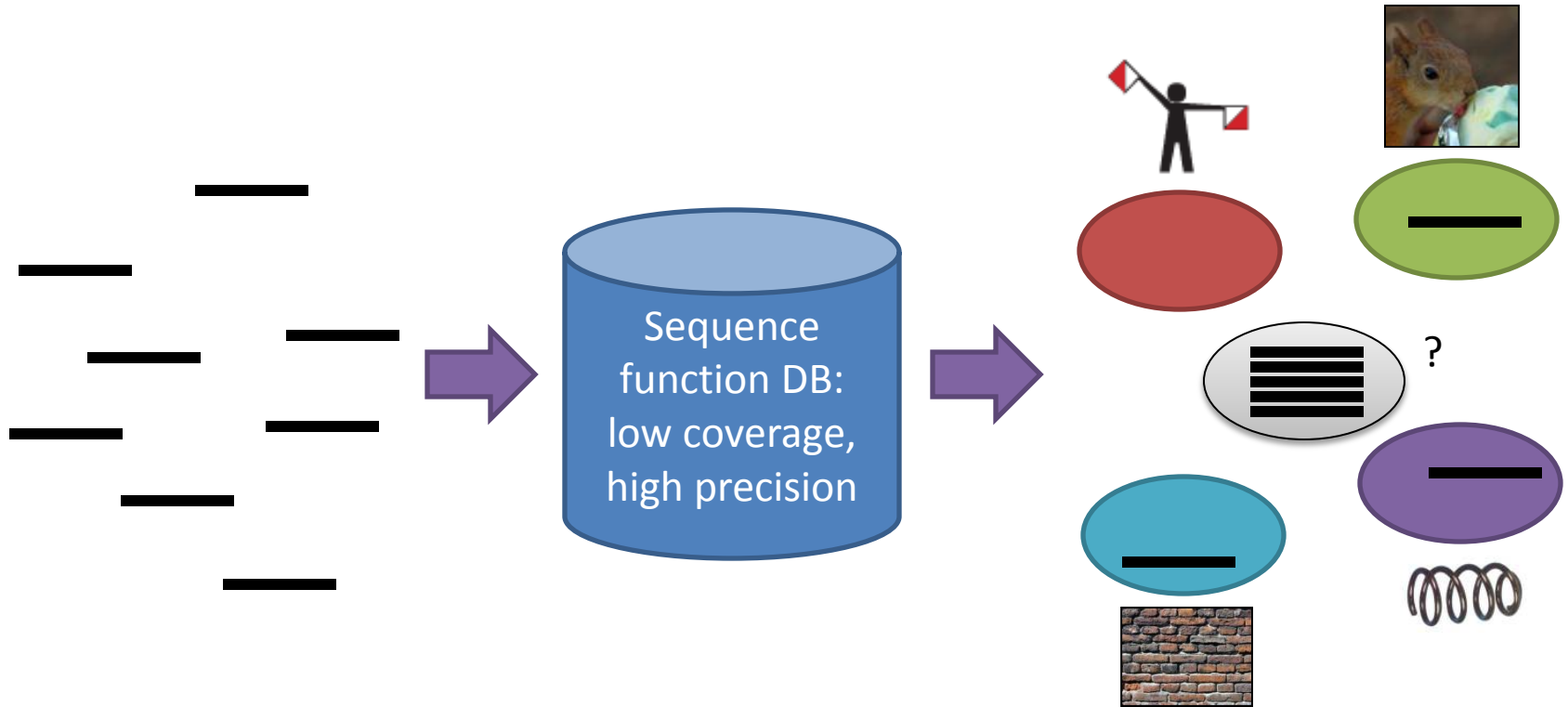
Functional Assignment: Objective



Realistic option #1

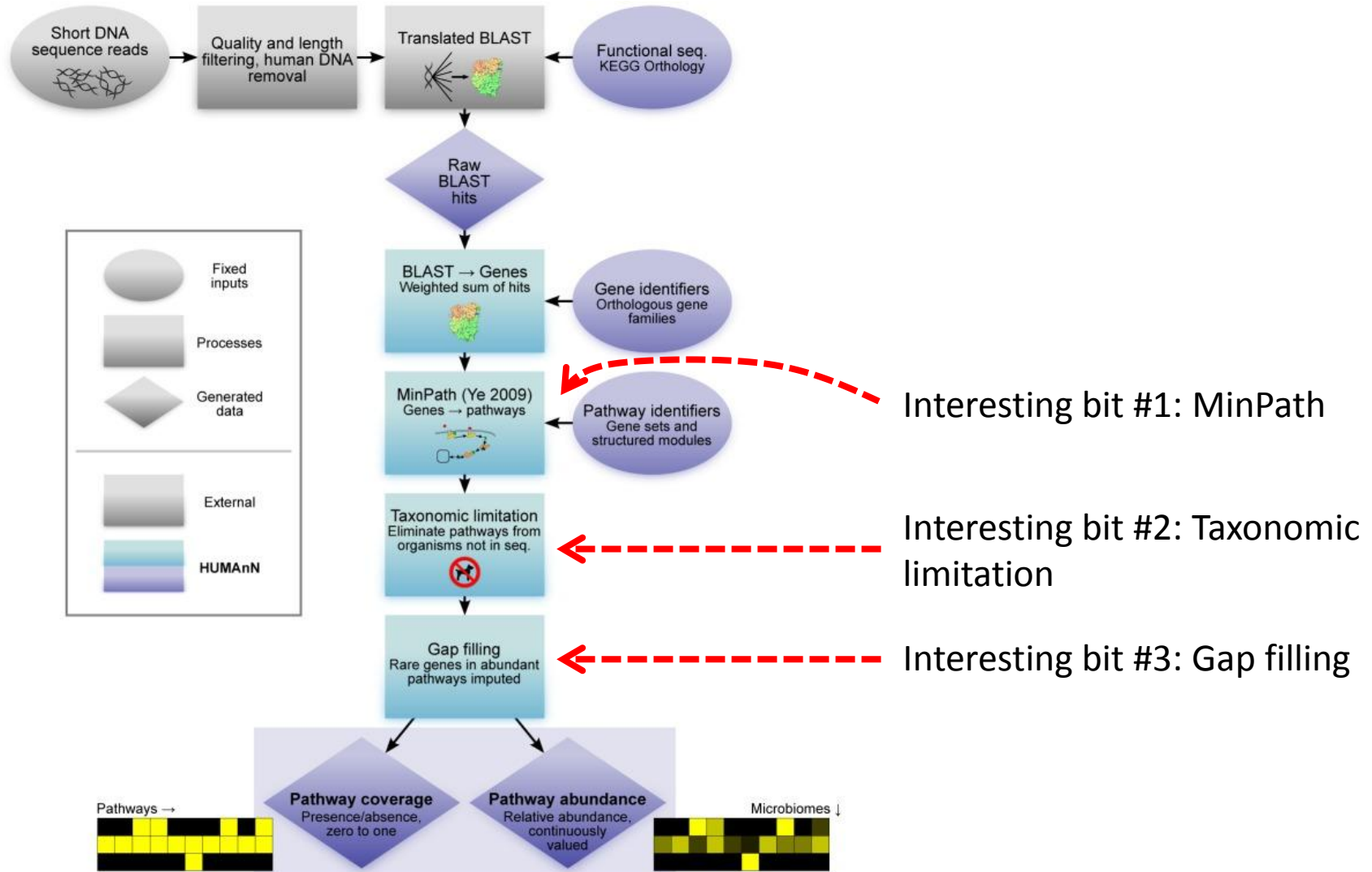


Realistic option #2



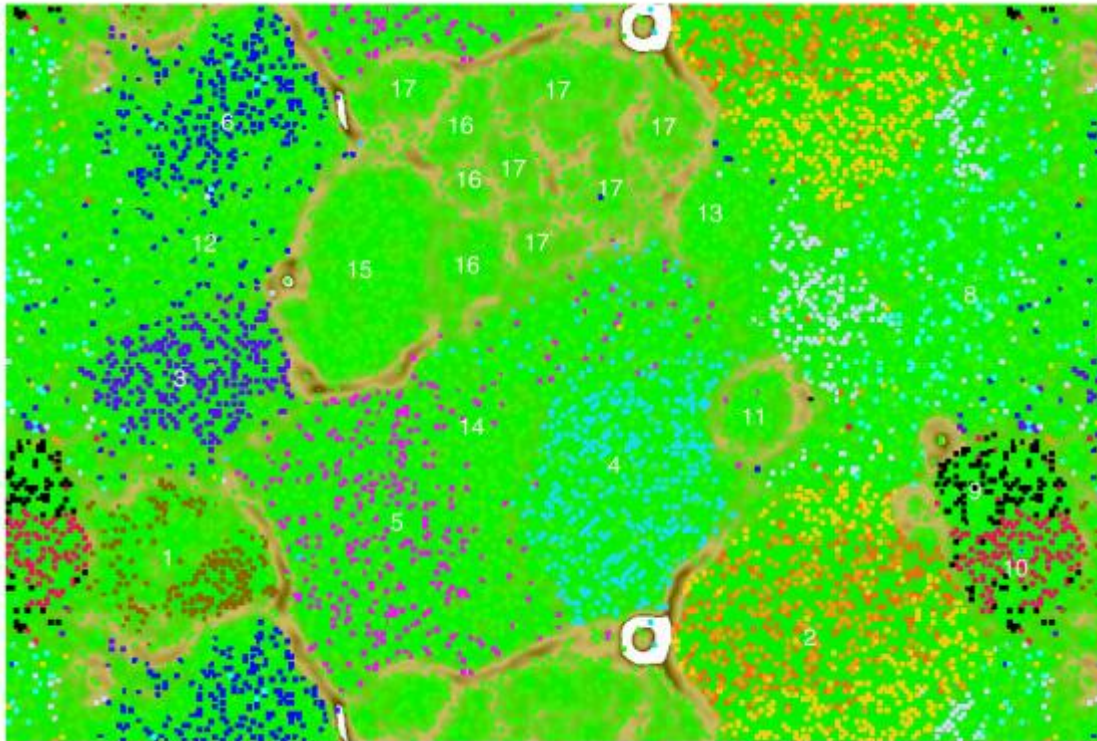
Examples: SWISS-PROT, *Cyc, BiGG

Example: HUMAnN



Taxonomy and Function – "Who is doing what"

- **Unsupervised:** "binning" using word frequencies

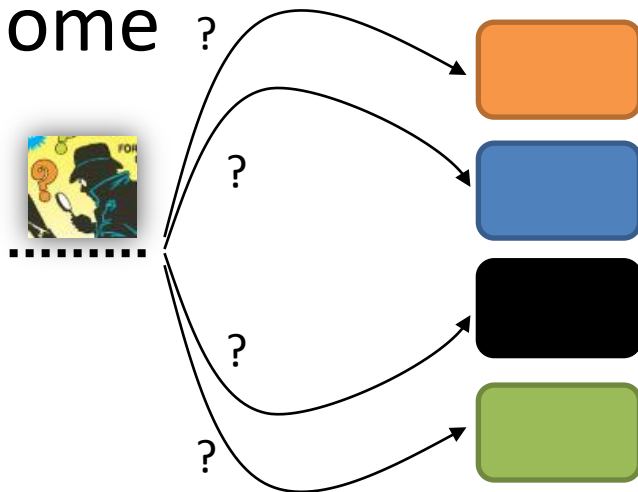


Self-organizing map:
Dick et al (2009)
Genome Biol

Taxonomy and Function – "Who is doing what"

Supervised: match to a reference database

- COMPOSITIONAL MODELS (k-mer, Markov models) for each reference genome

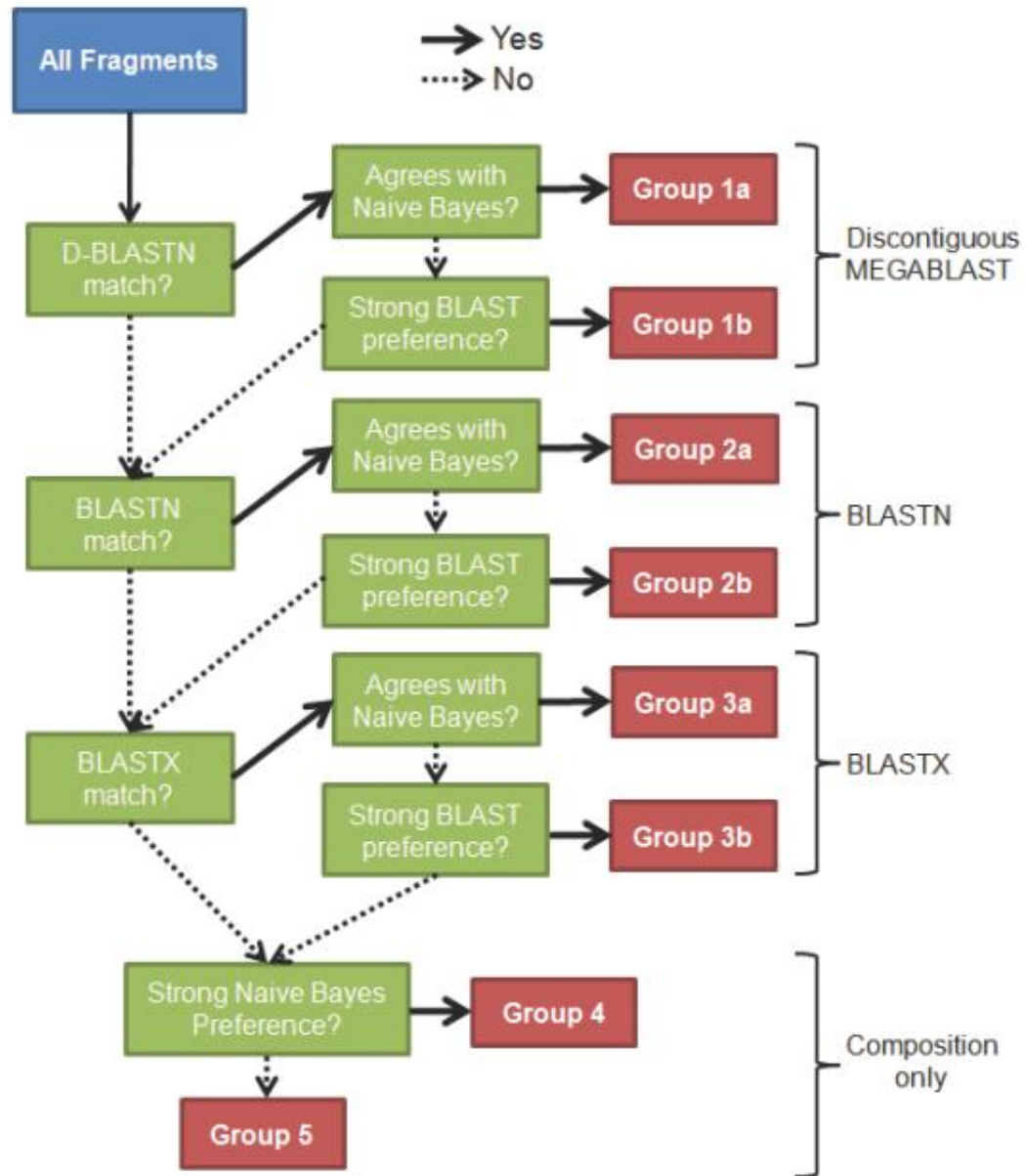


- Compare against a reference database using BLAST



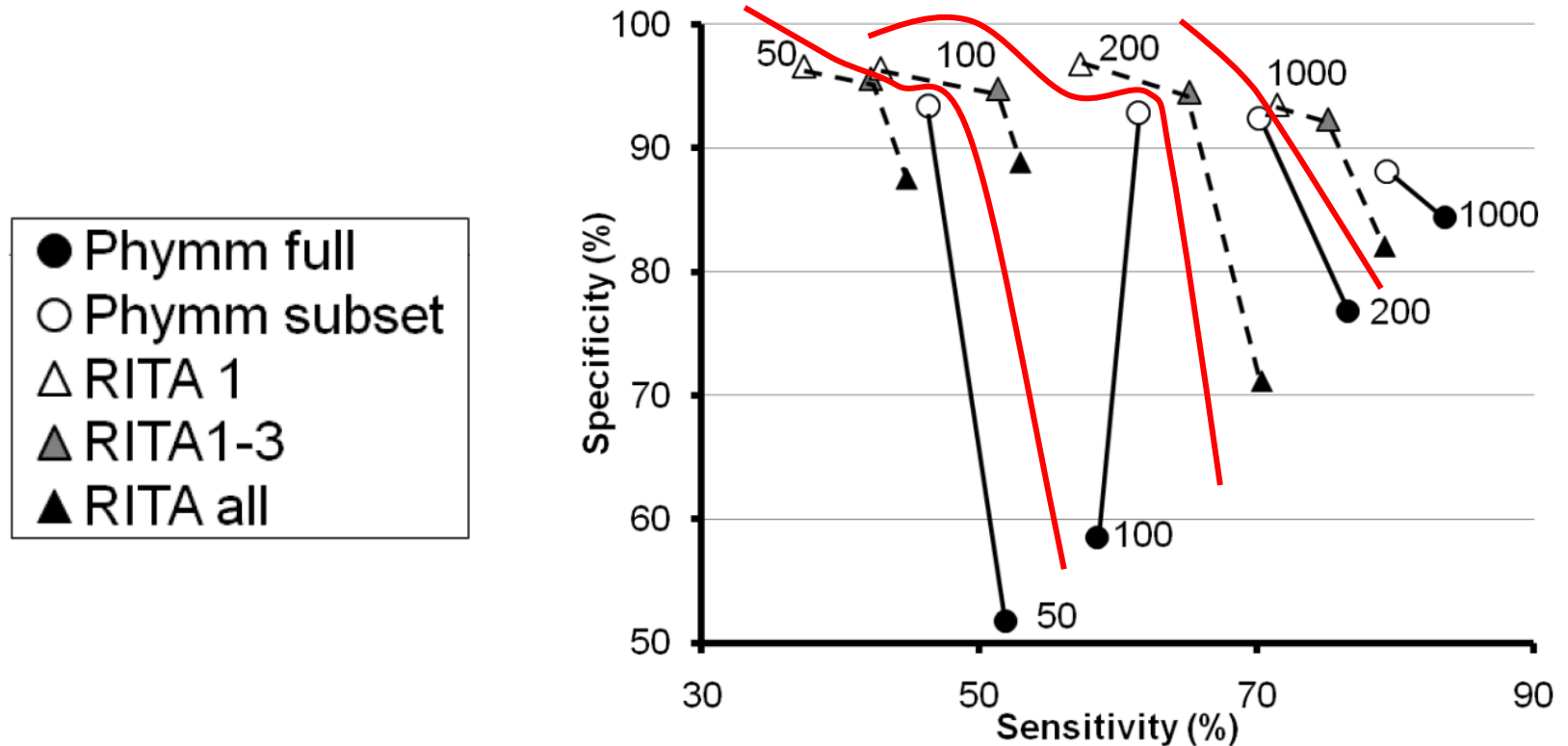
Hybrid classifiers (e.g. PhymmBL) –
combine predictions of both

RITA

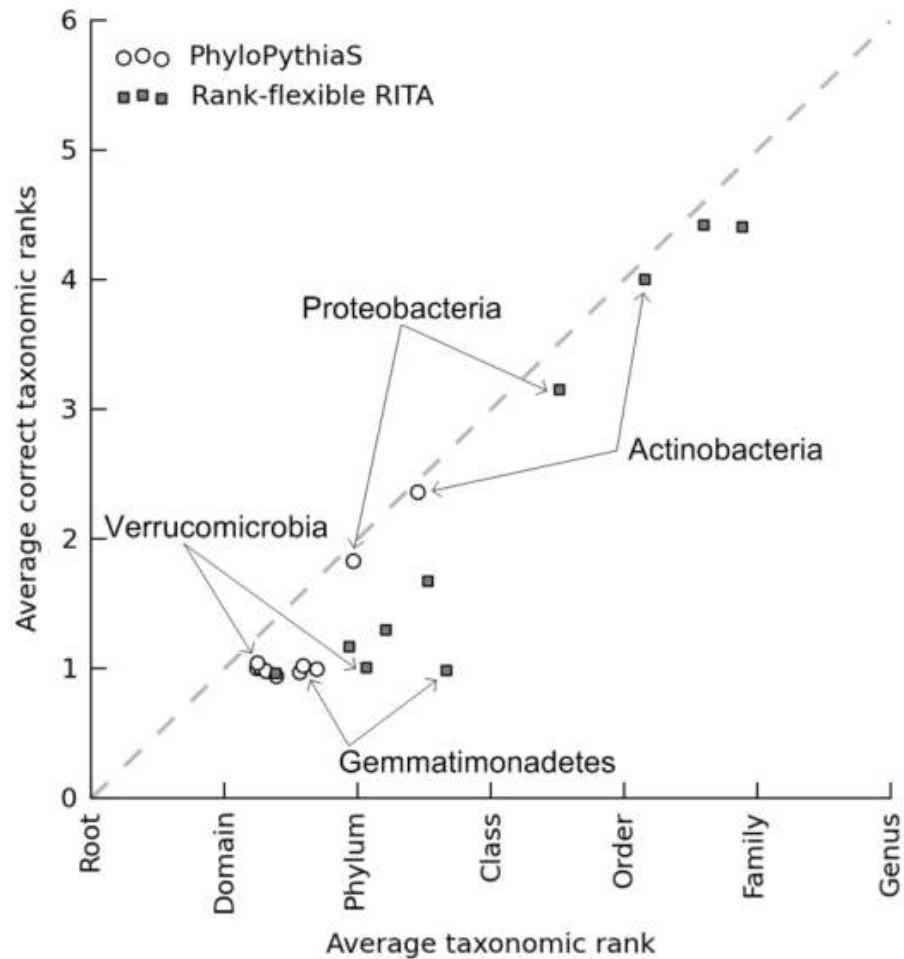


Rank-specific classification

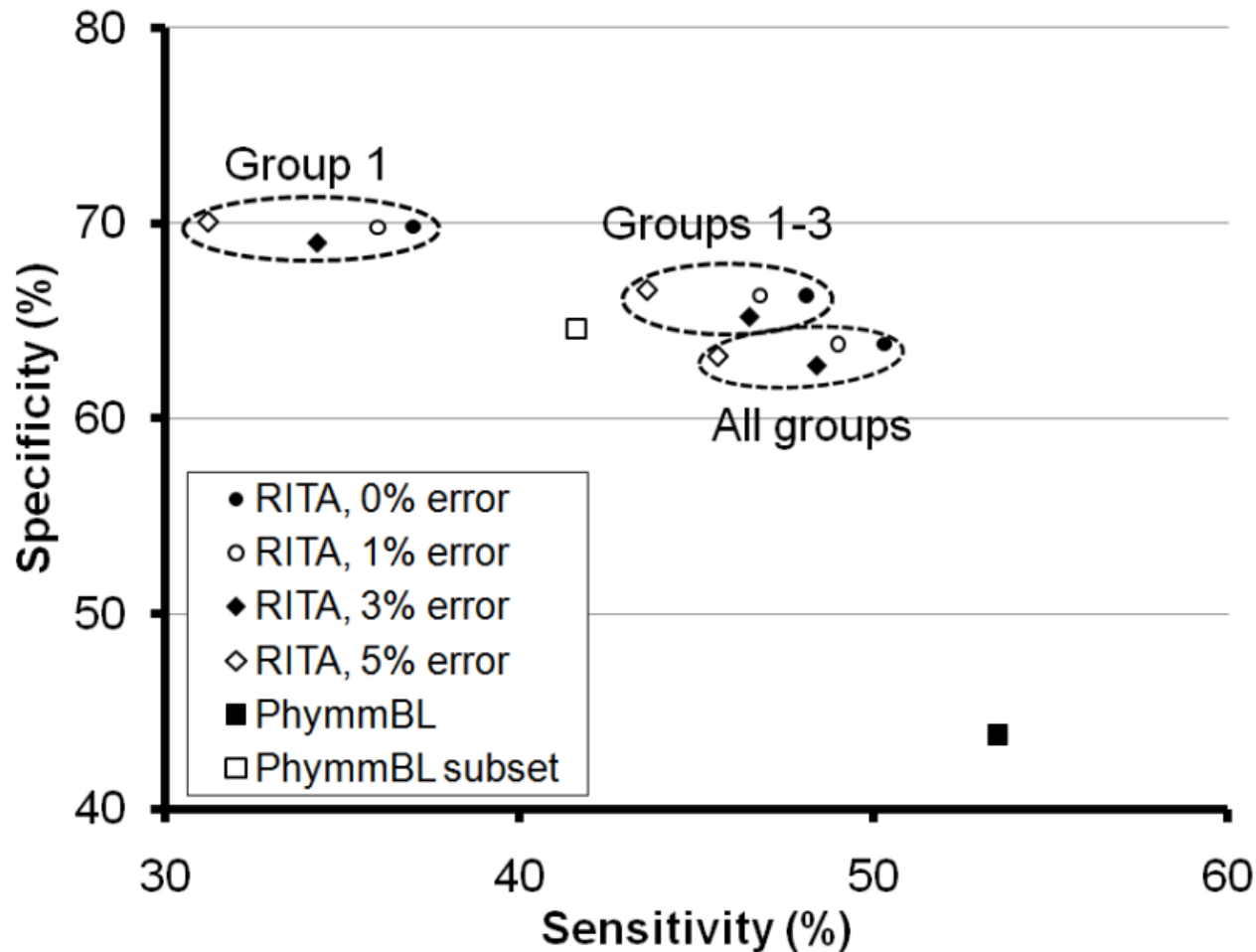
- Can we classify fragments from an isolate to the correct genus?



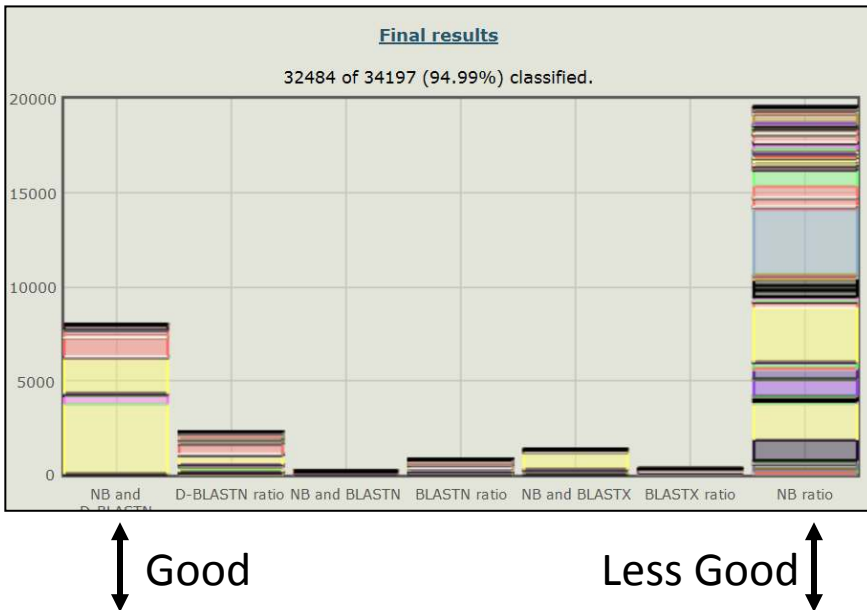
Rank-flexible



Performance on a real fake metagenome (read length ~230 nt)

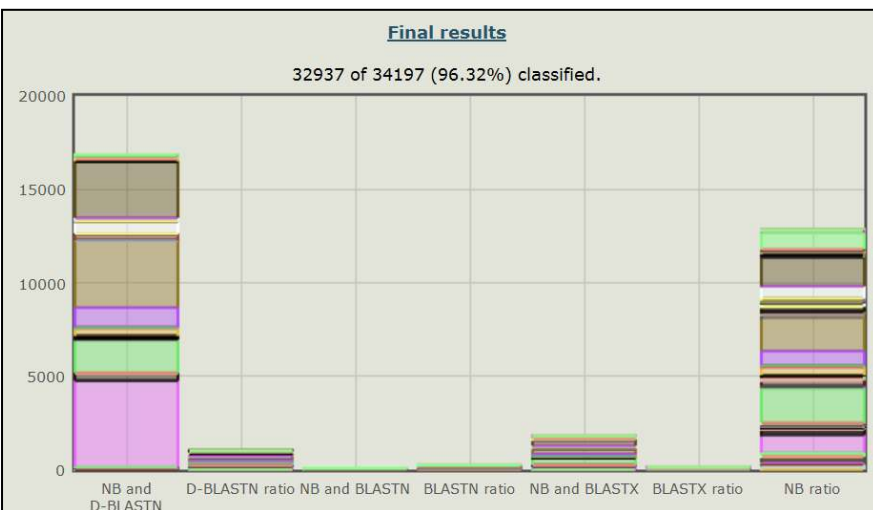


Obese twin gut metagenomes



Without HMP genomes:

Clostridium, *Bacteroides* and *Eubacterium*, but lots of low-confidence calls too



With HMP reference genomes:

Add *Ruminococcus*, *Faecalibacterium*, Lachnospiraceae

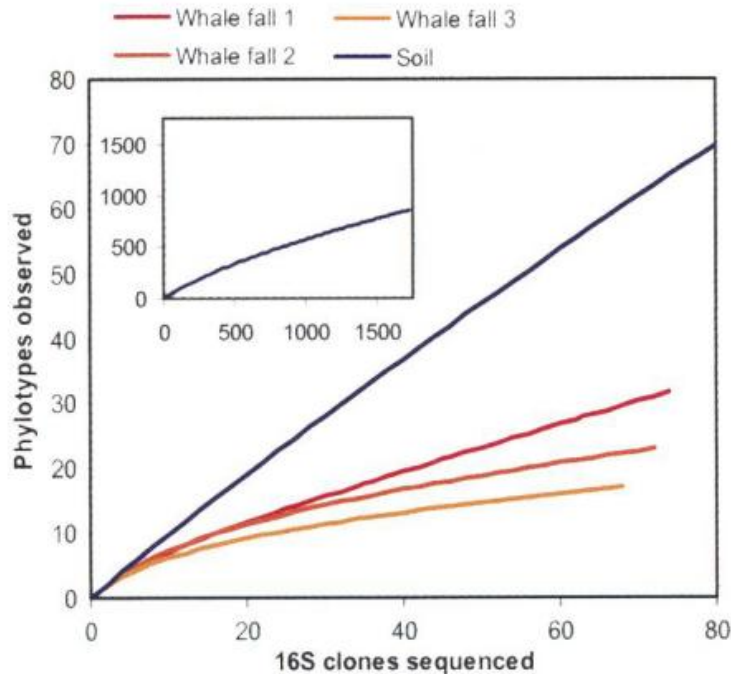
Metagenomics in action: Function and (sometimes) taxonomy

Comparisons between communities

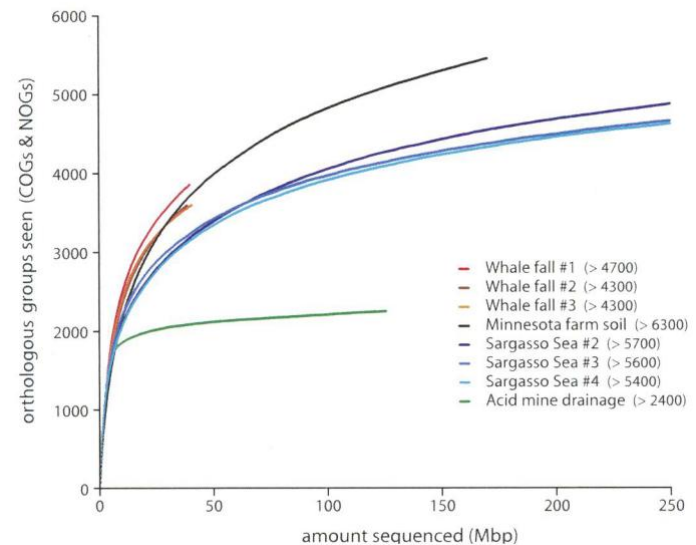


Comparisons between communities

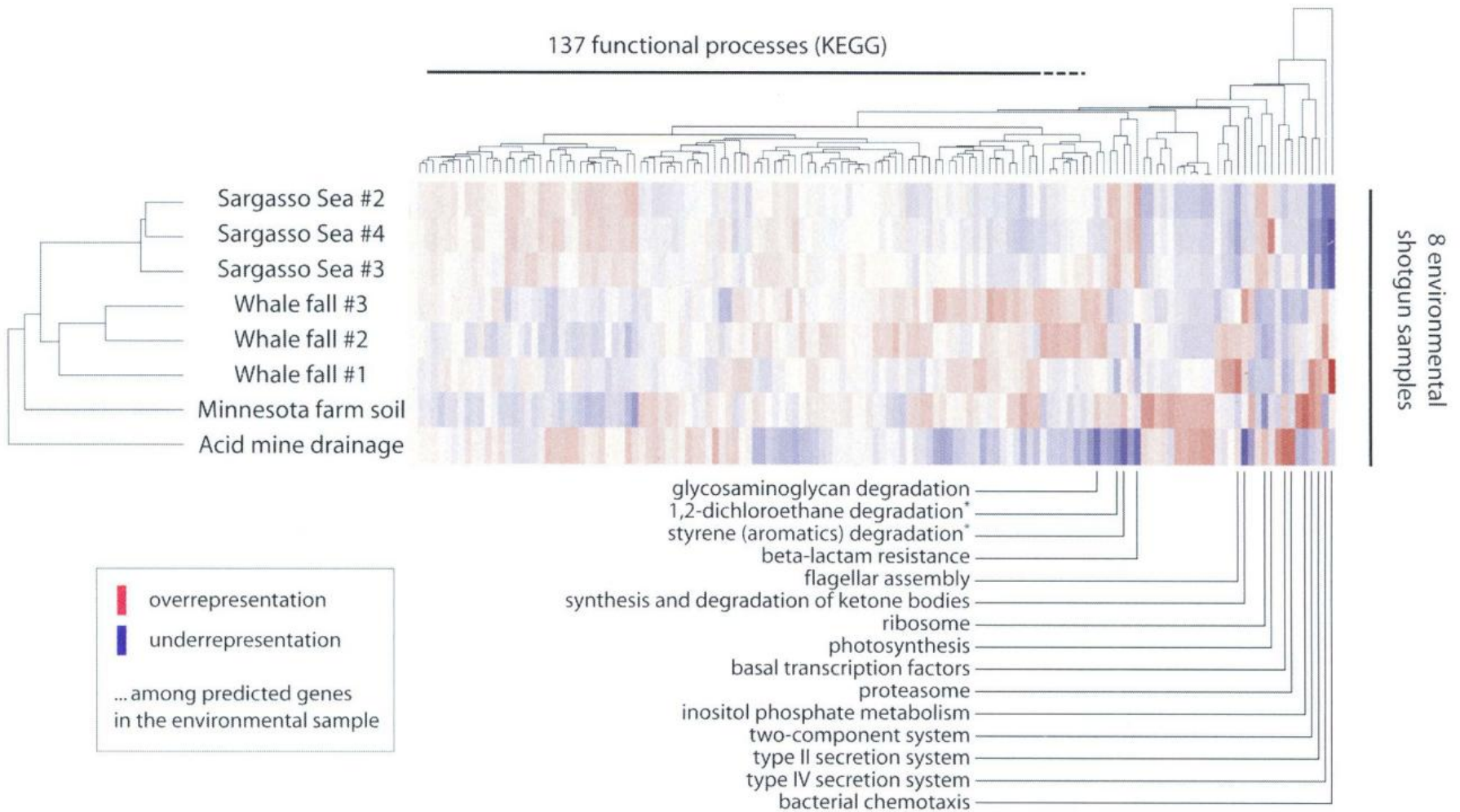
RAREFACTION – how diverse are they?



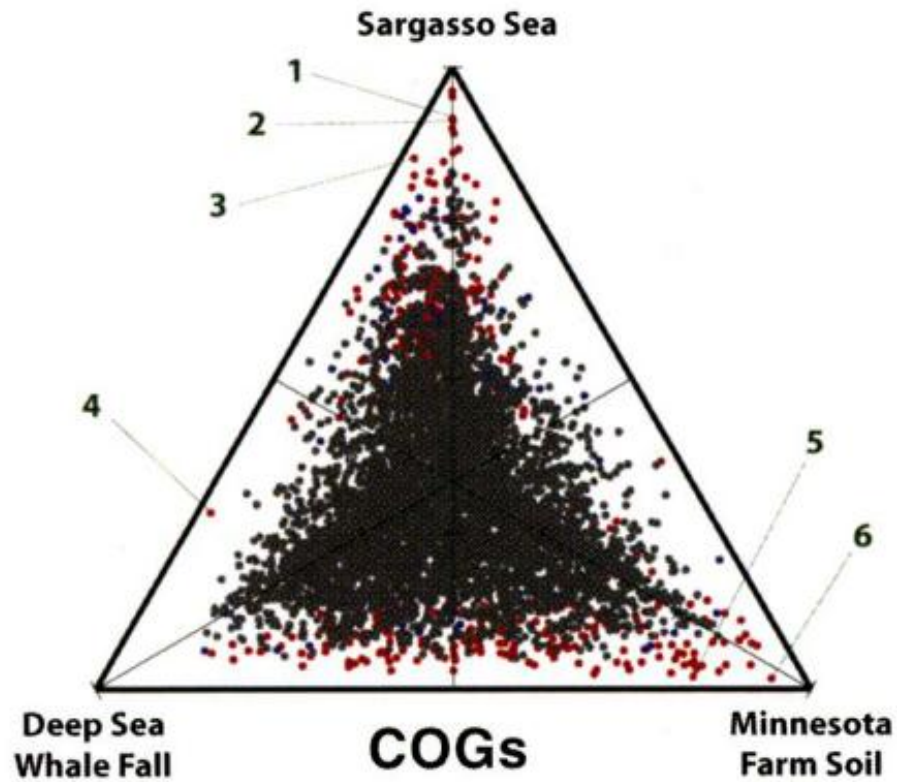
Marker genes



Metagenome samples



Clustering and over/underrepresentation

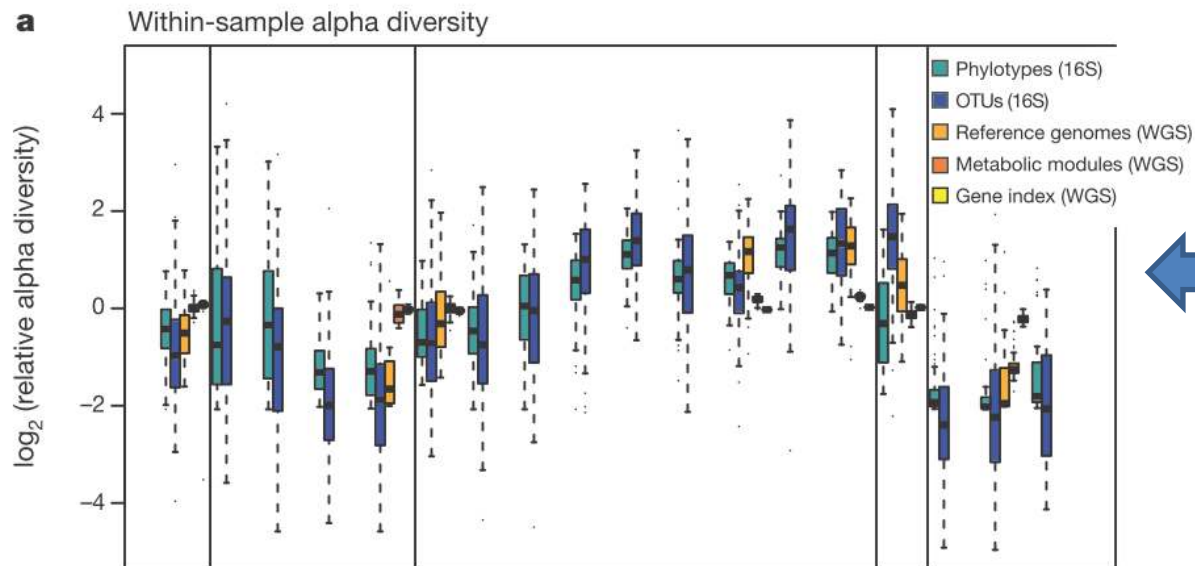


1: bacteriorhodopsin

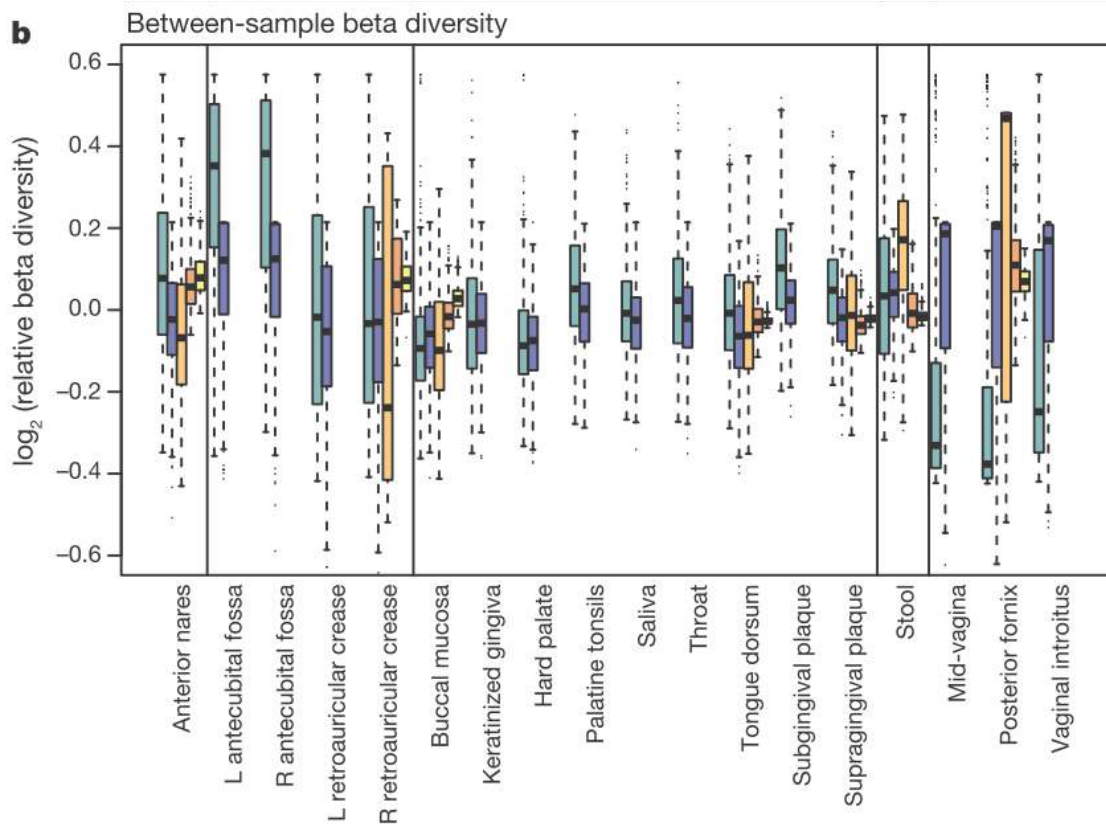
4: ???

5: cellobiose phosphorylase

The Human Microbiome Project

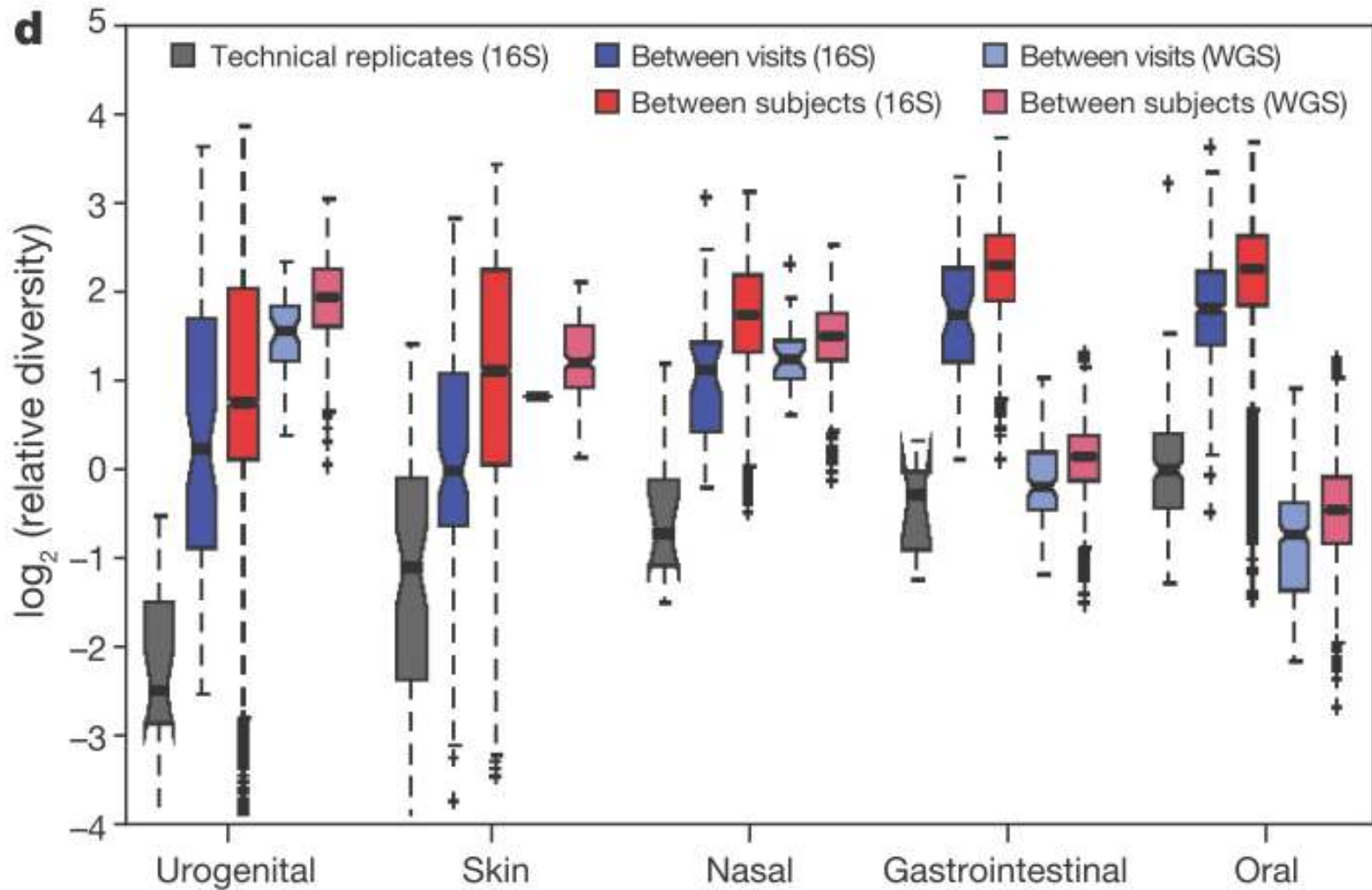


Alpha diversity, sliced
several ways



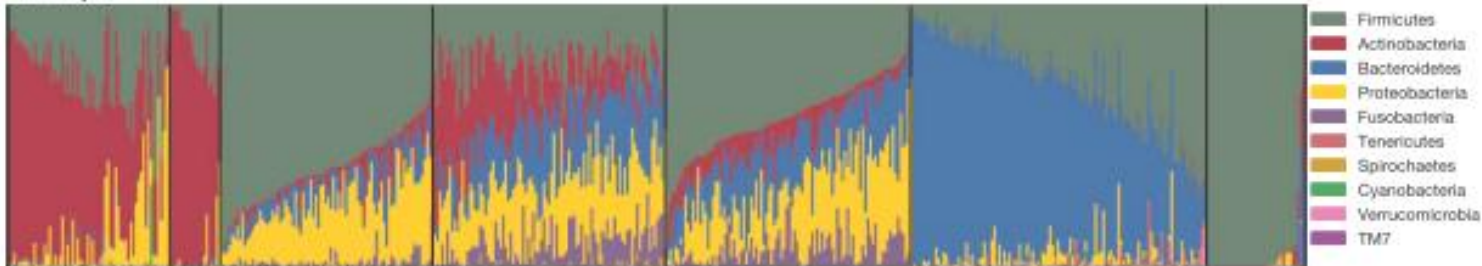
Beta diversity

Variation within replicates, across time points and between individuals

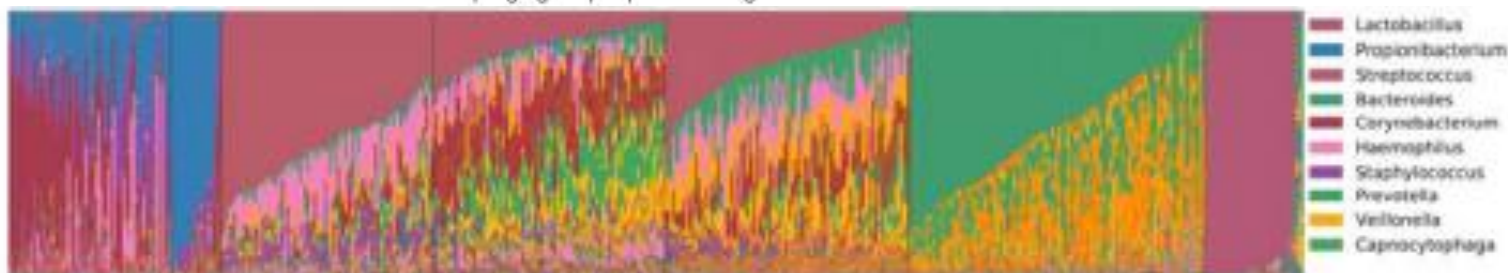
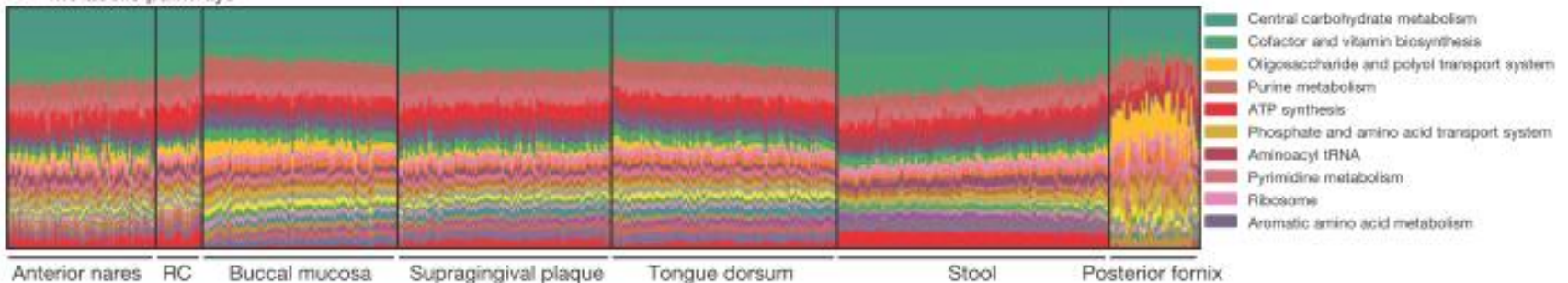


Taxonomic vs. functional diversity

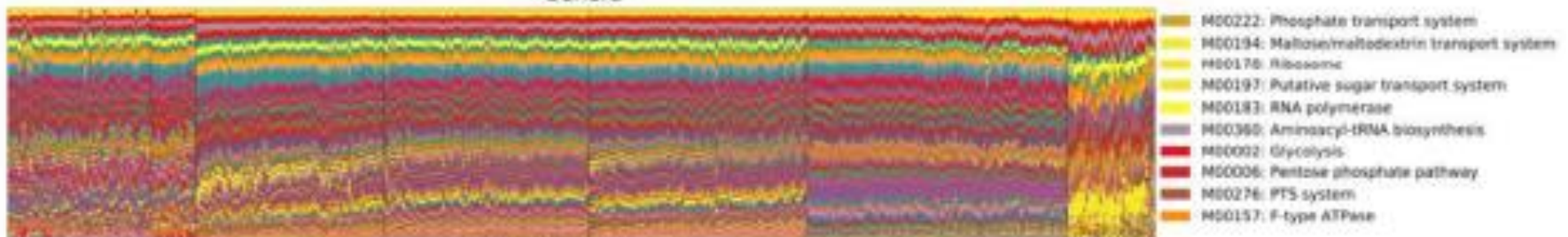
a Phyla



b Metabolic pathways

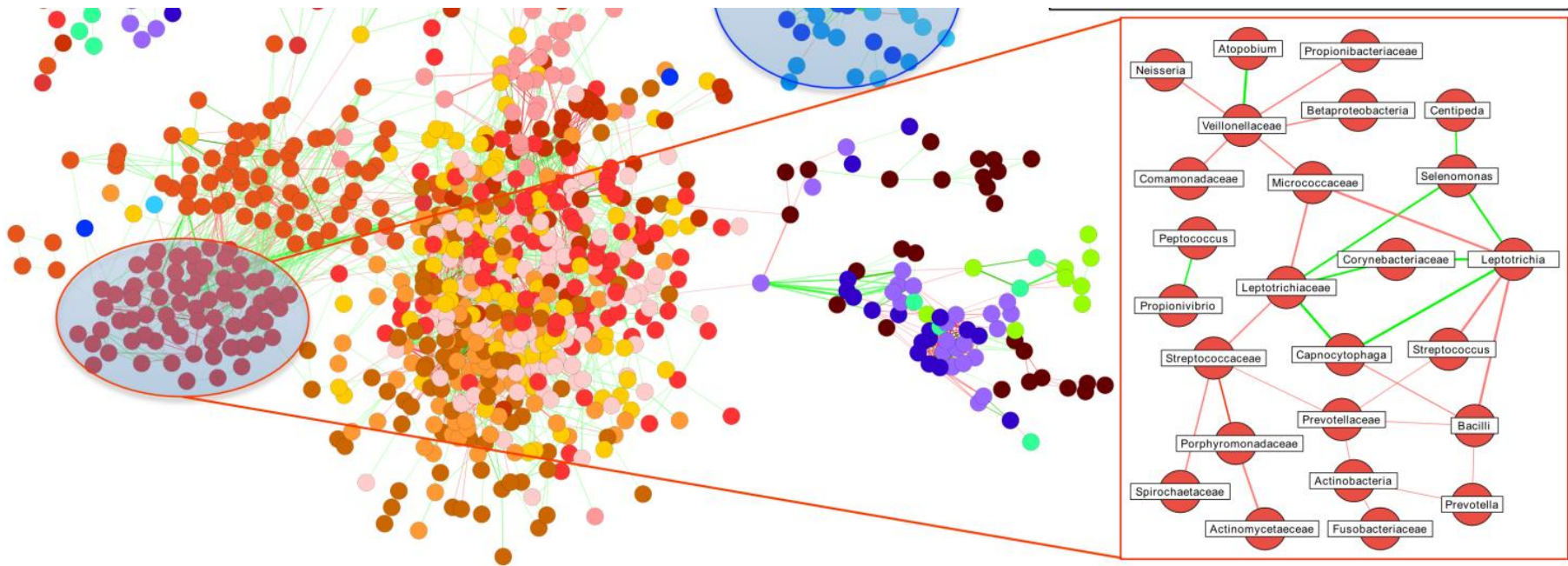


Genera

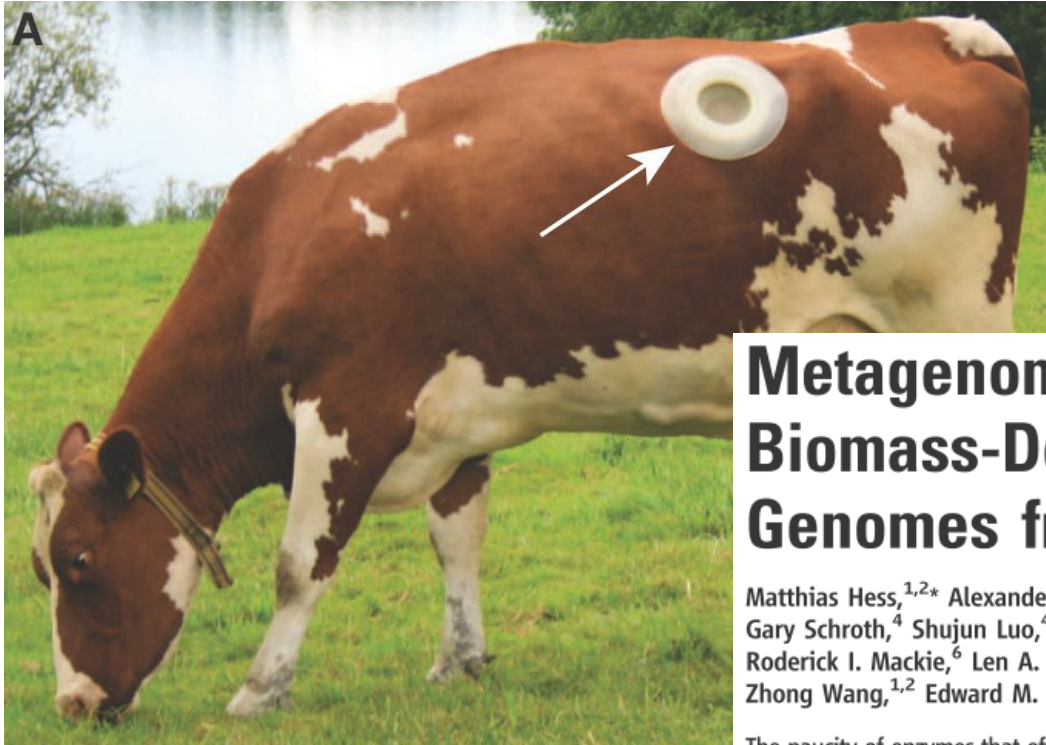


Metabolic modules

Co-occurrence networks – successional patterns in the dental plaque



Enzyme and pathway discovery (in the most exotic of places)



Metagenomic Discovery of Biomass-Degrading Genes and Genomes from Cow Rumen

Matthias Hess,^{1,2*} Alexander Sczyrba,^{1,2*} Rob Egan,^{1,2} Tae-Wan Kim,³ Harshal Chokhawala,³ Gary Schroth,⁴ Shujun Luo,⁴ Douglas S. Clark,^{3,5} Feng Chen,^{1,2} Tao Zhang,^{1,2} Roderick I. Mackie,⁶ Len A. Pennacchio,^{1,2} Susannah G. Tringe,^{1,2} Axel Visel,^{1,2} Tanja Woyke,^{1,2} Zhong Wang,^{1,2} Edward M. Rubin^{1,2†}

The paucity of enzymes that efficiently deconstruct plant polysaccharides represents a major bottleneck for industrial-scale conversion of cellulosic biomass into biofuels. Cow rumen microbes specialize in degradation of cellulosic plant material, but most members of this complex community resist cultivation. To characterize biomass-degrading genes and genomes, we sequenced and analyzed 268 gigabases of metagenomic DNA from microbes adherent to plant fiber incubated in cow rumen. From these data, we identified 27,755 putative carbohydrate-active genes and expressed 90 candidate proteins, of which 57% were enzymatically active against cellulosic substrates. We also assembled 15 uncultured microbial genomes, which were validated by complementary methods including single-cell genome sequencing. These data sets provide a substantially expanded catalog of genes and genomes participating in the deconstruction of cellulosic biomass.

Hess et al. (2011) *Science*

Do microbes form microbial "communities?"



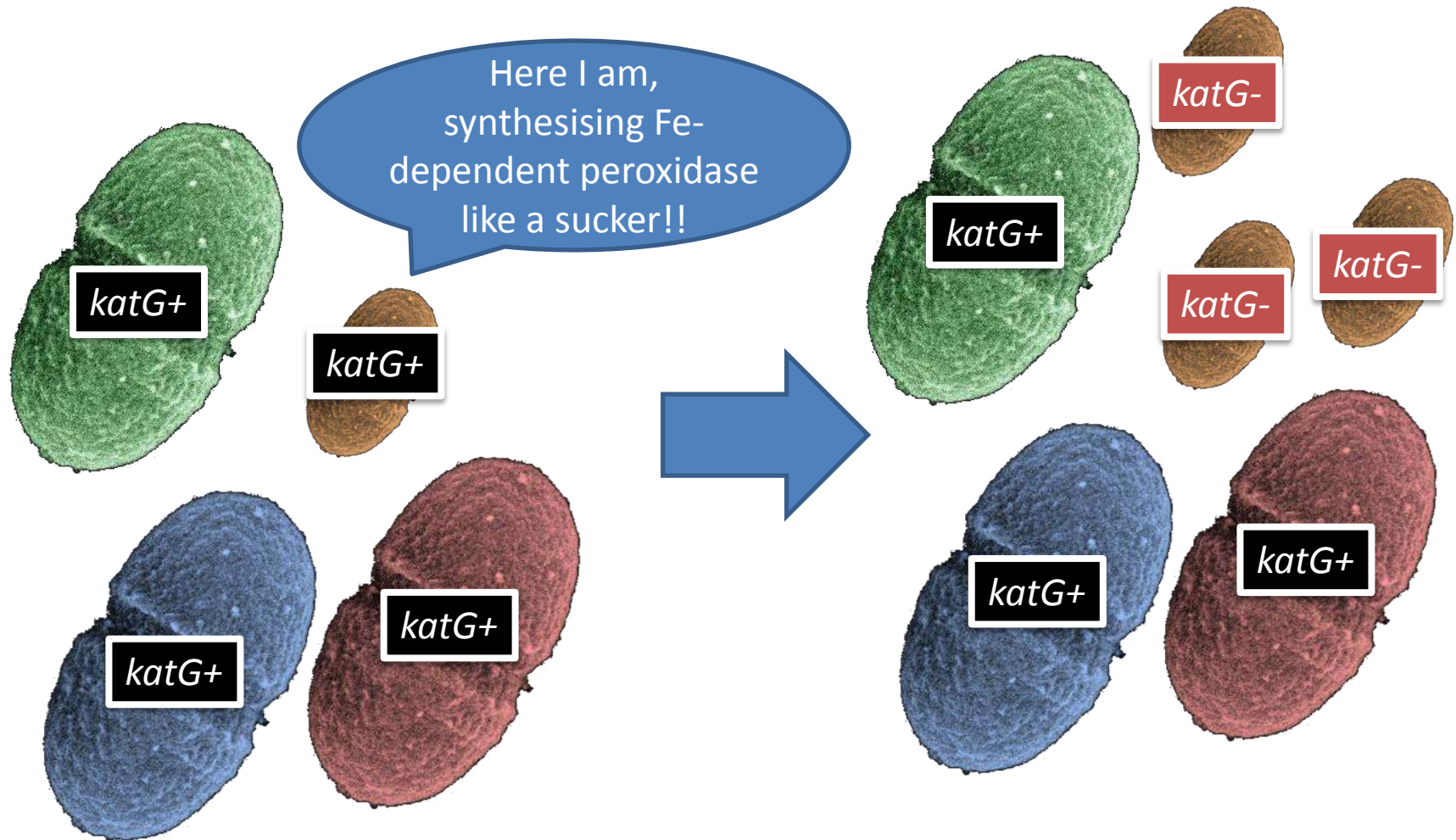
"Handoffs"

vs.



"Riding the elevator –
DON'T MAKE EYE CONTACT"

The Black Queen Hypothesis

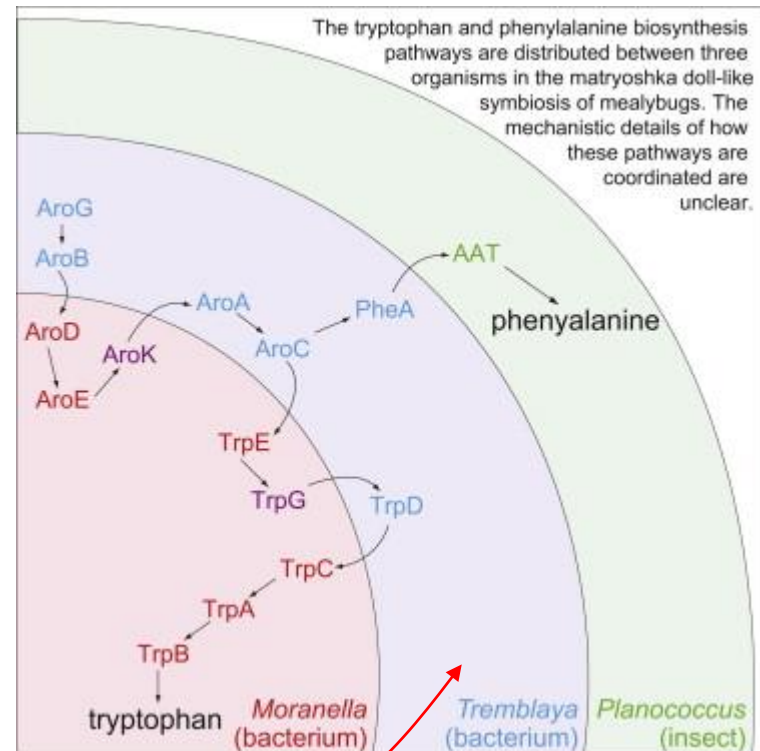


Insect bacteriomes

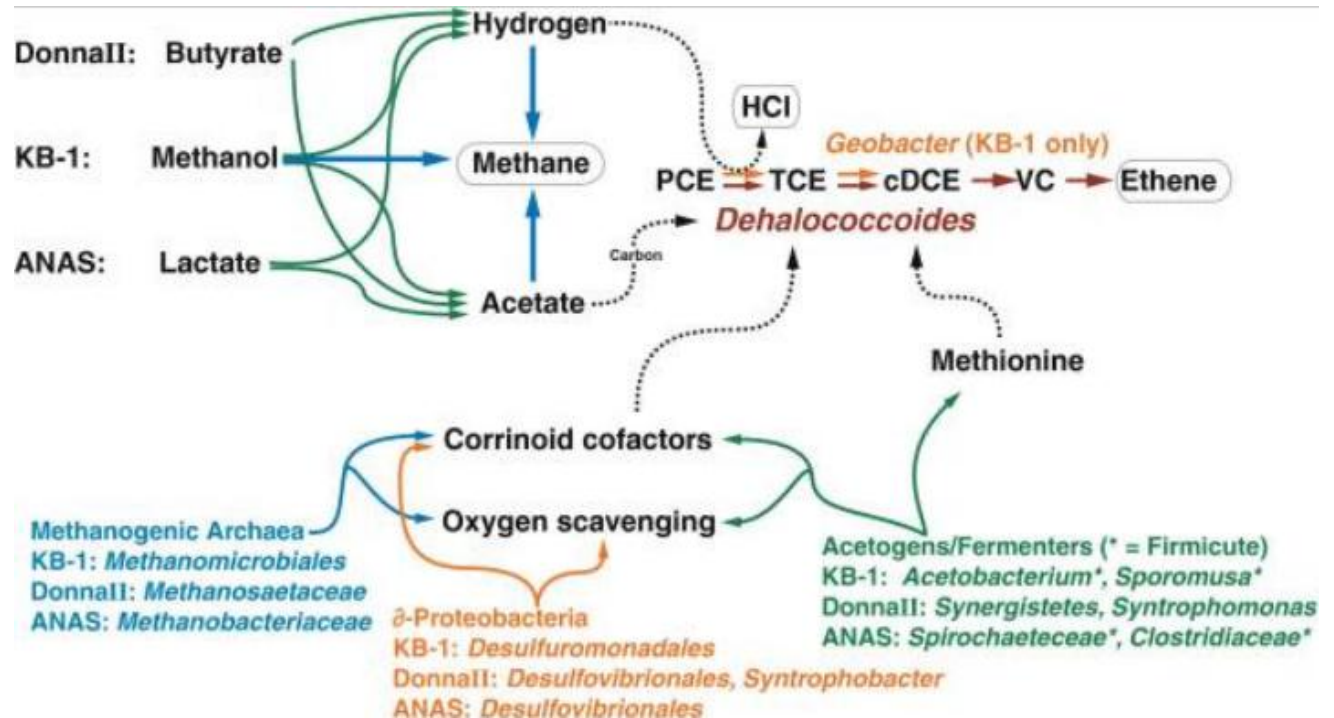
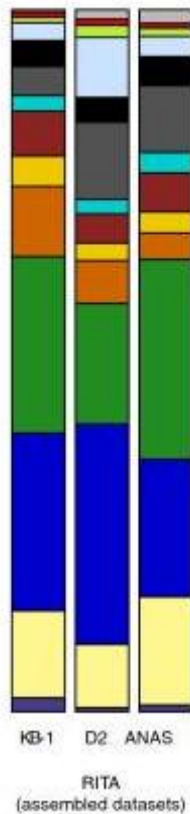
- The craziest system ever. No exceptions.



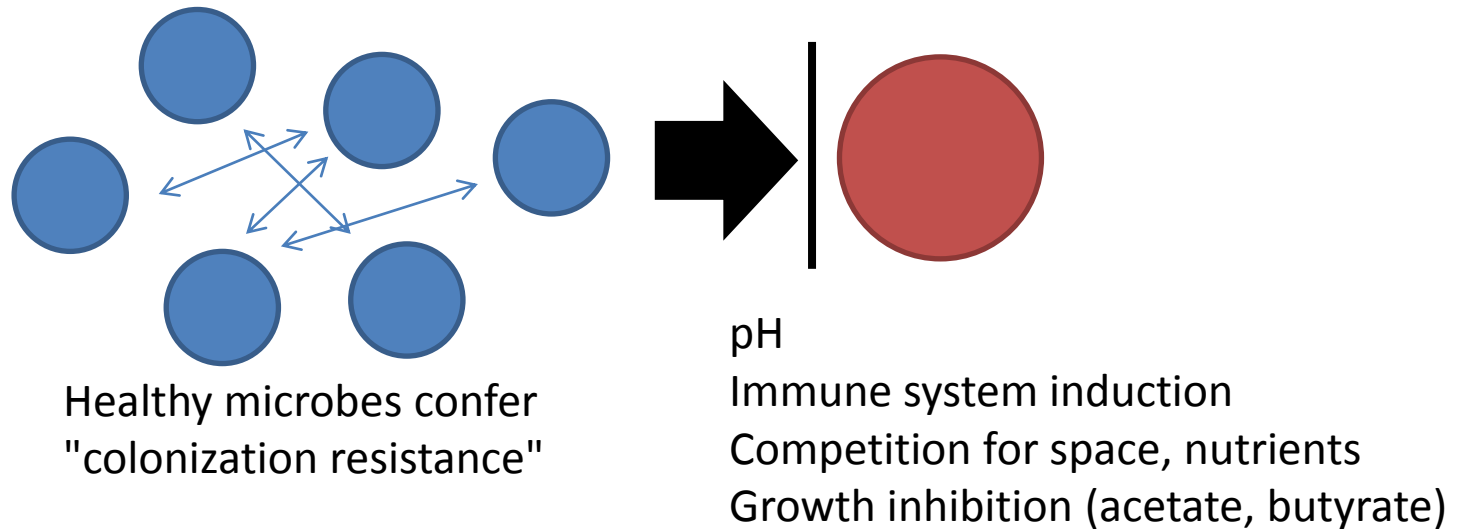
Tremblaya:
138,000 nt genome
140 genes
73% coding density



Dependencies



C. difficile?



Journal of Medical Microbiology (2011), 60, 1617–1625

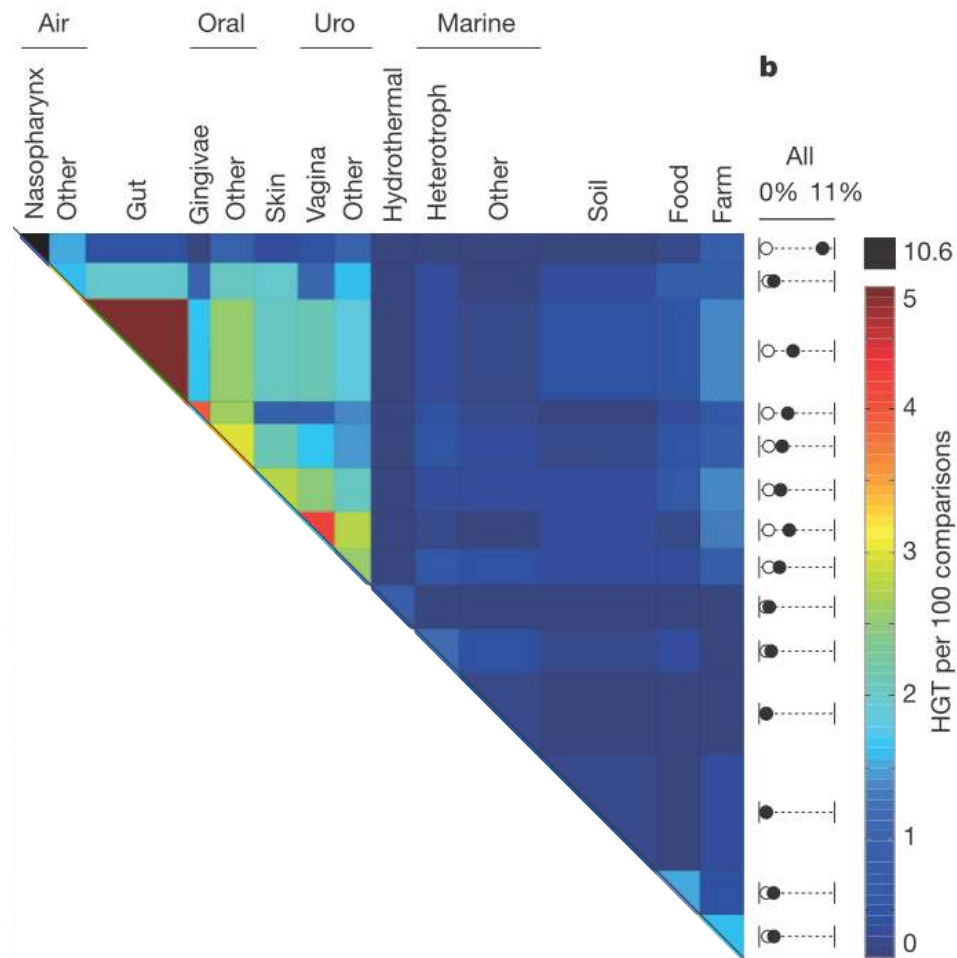
DOI 10.1099/jmm.0.033423-0

Inhibition of the cytotoxic effect of *Clostridium difficile* *in vitro* by *Clostridium butyricum* MIYAIRI 588 strain

Timothy D. H. Woo,¹ Kentaro Oka,² Motomichi Takahashi,² Fuhito Hojo,¹
Takako Osaki,¹ Tomoko Hanawa,¹ Satoshi Kurata,¹ Hideo Yonezawa¹
and Shigeru Kamiya¹

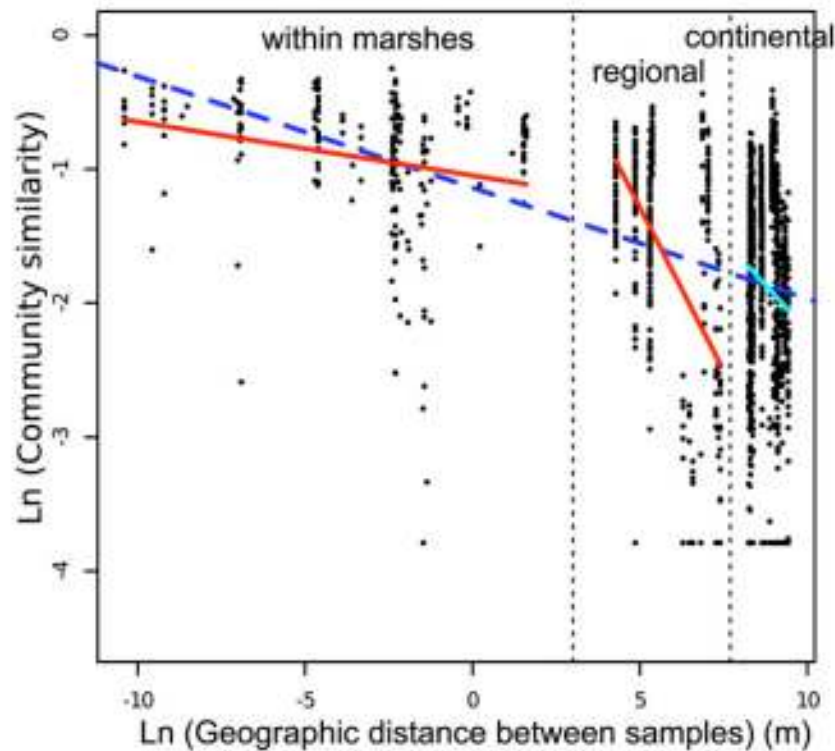
Big questions for the future:
communities and metagenomes

The role of lateral gene transfer in different settings

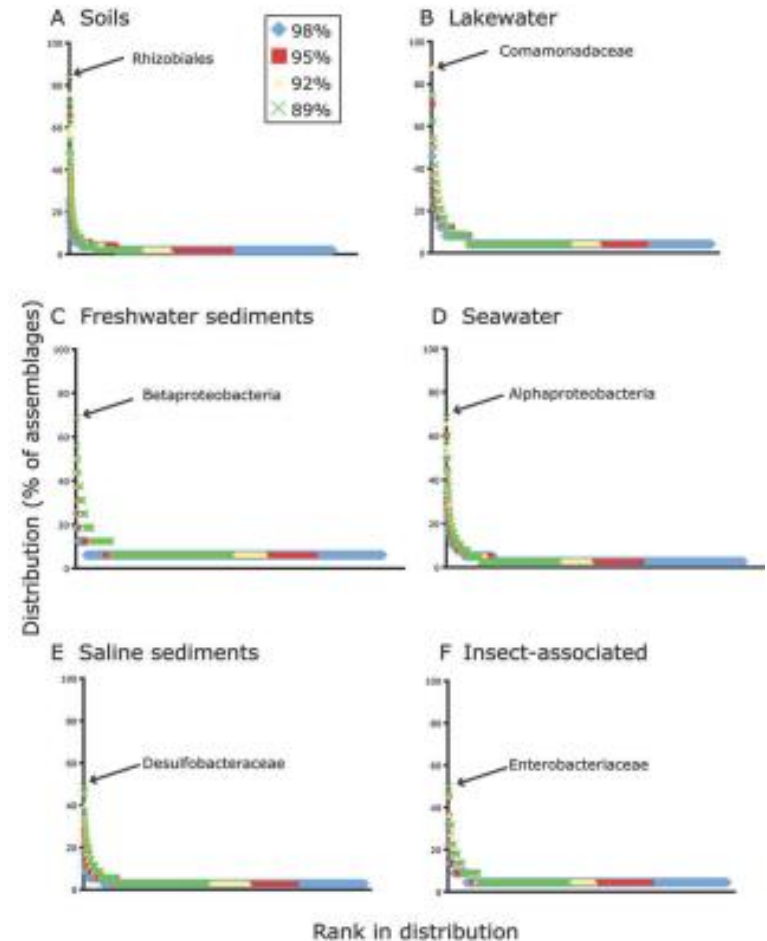


Biogeography

Is everything everywhere?



Distance-decay curves
Martiny et al. (2011) *PNAS*

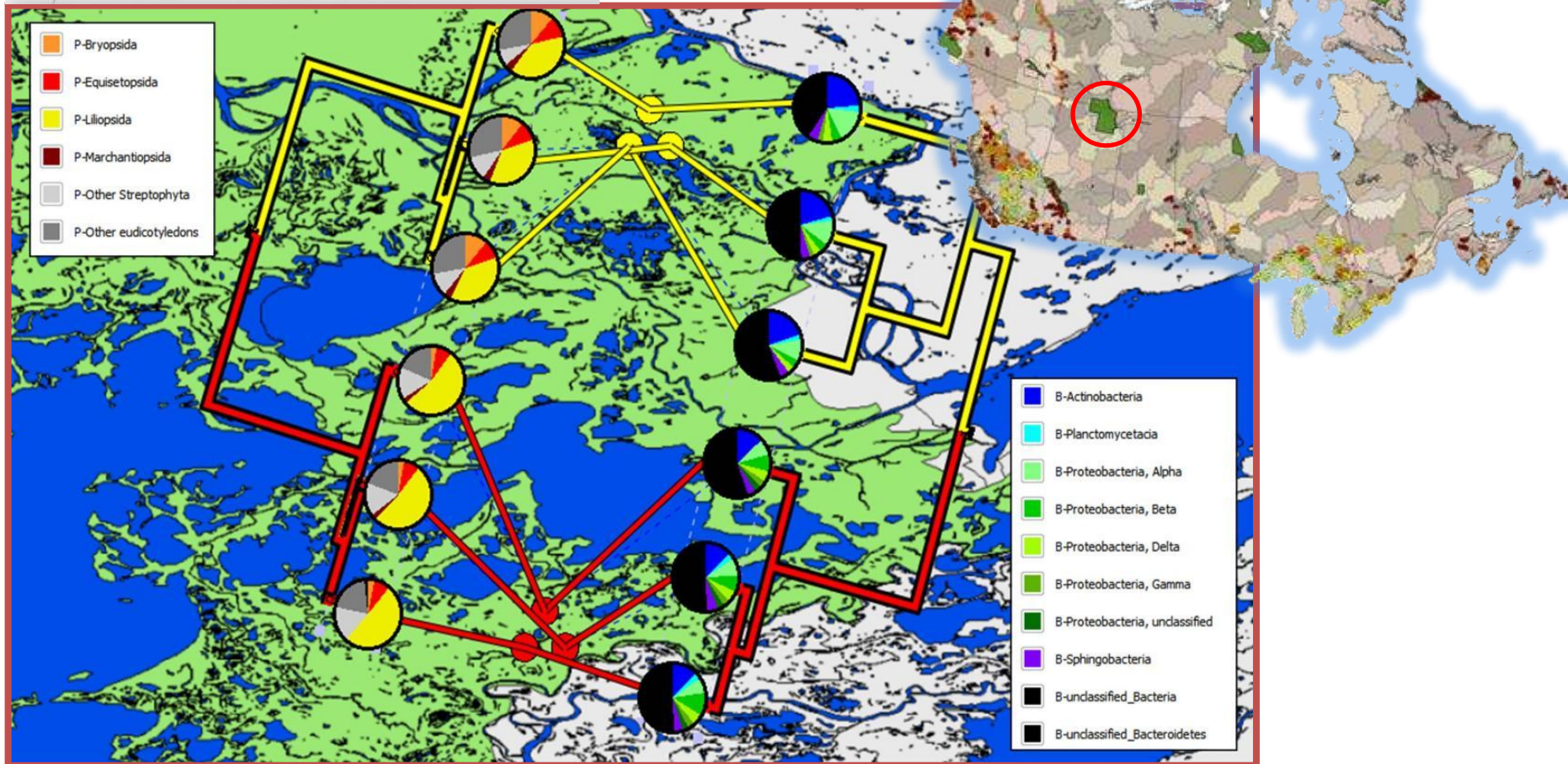


Nemergut et al. (2011) *Env Microbiol*

Environmental monitoring and response

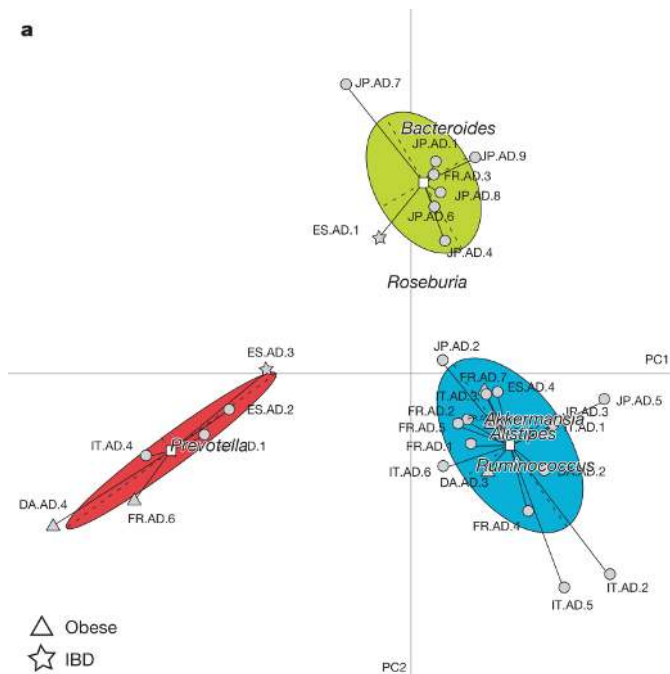


BIOMONITORING 2.0

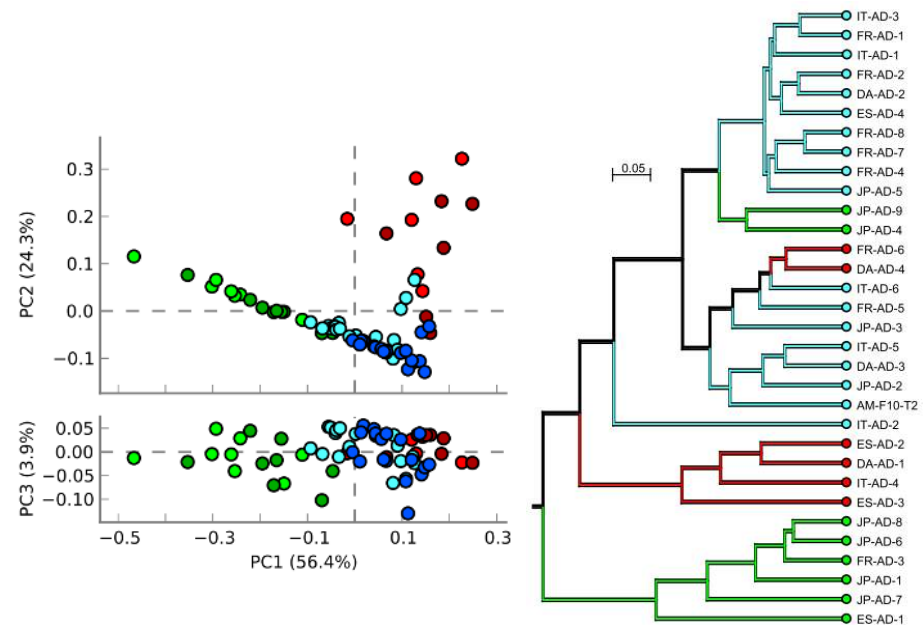


Are there distinct "types" of community?

- Whether there are stable points for communities, or gradients of diversity

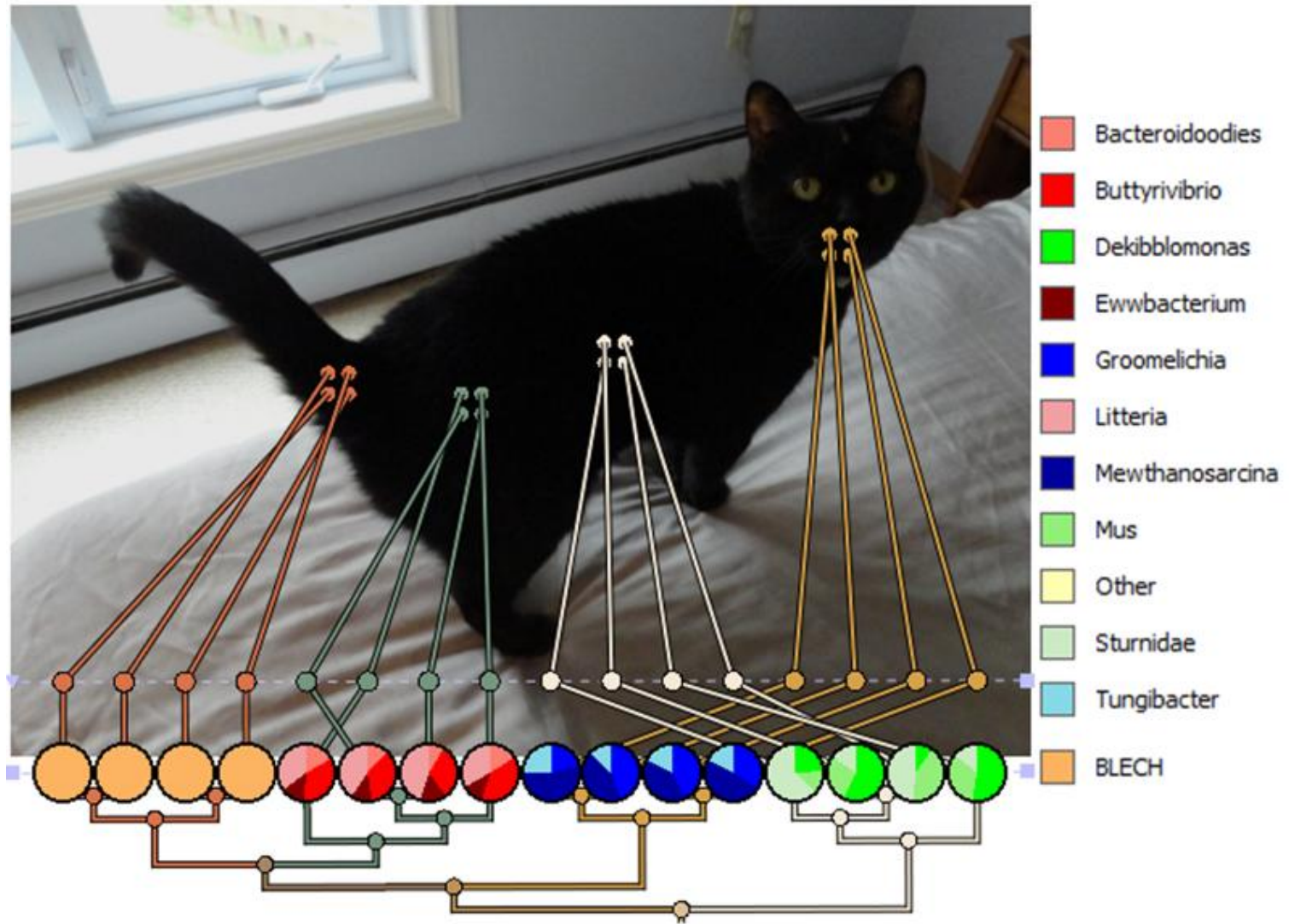


Arumugam et al. (2011) *Science*
"Structured" PCA



MacDonald et al. (2012) *Nucleic Acids Res*
"Plain old" PCA

The Kitten Microbiome Project: Not a real data slide



Fin