Shotgun sequencing of a simulated environmental sample

Jenna Morgan
02-05-08

Metagenomics

- "the application of modern genomics techniques to the study of communities of microbial organisms directly in their natural environments, bypassing the need for isolation and lab cultivation of individual species."

Chen and Pachter, 2005

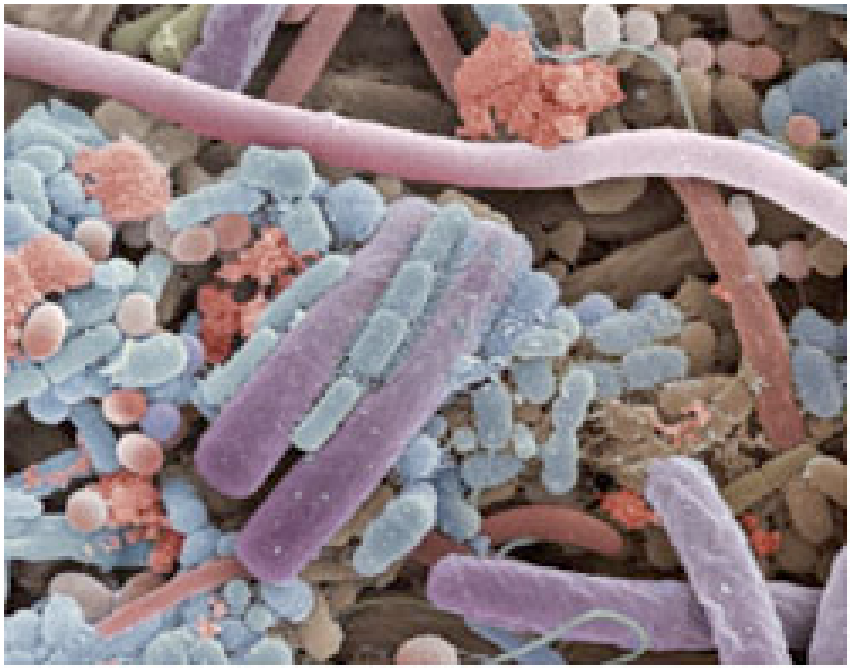
Bullet points

- Culture independent
- Microbial communities
- Modern genomics

Culture-independent

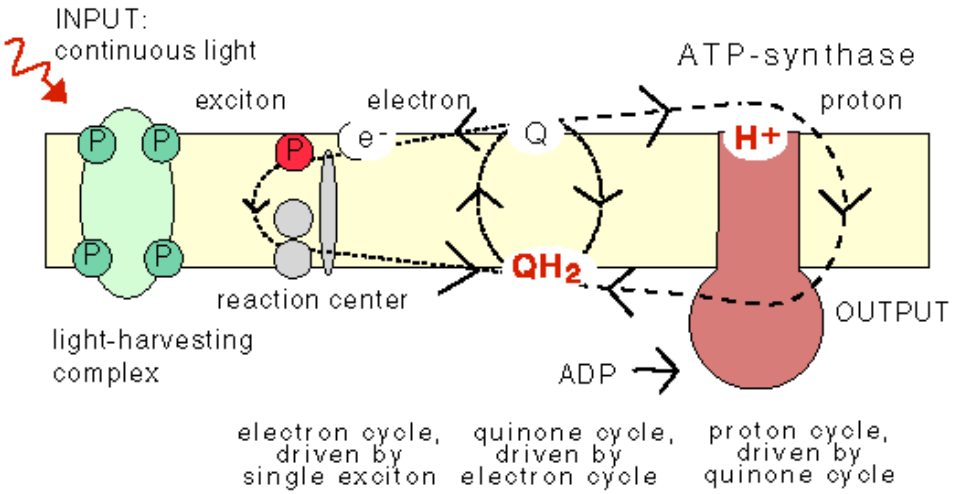
- Historically, to characterize the taxonomy, structure and function of a microorganism, they were first isolated from the environment and cultured in the lab
- A metagenomic approach bypasses this requirement by sequencing DNA that is extracted wholesale from the environment

The study of microbial communities



Who is there?

What are they doing?



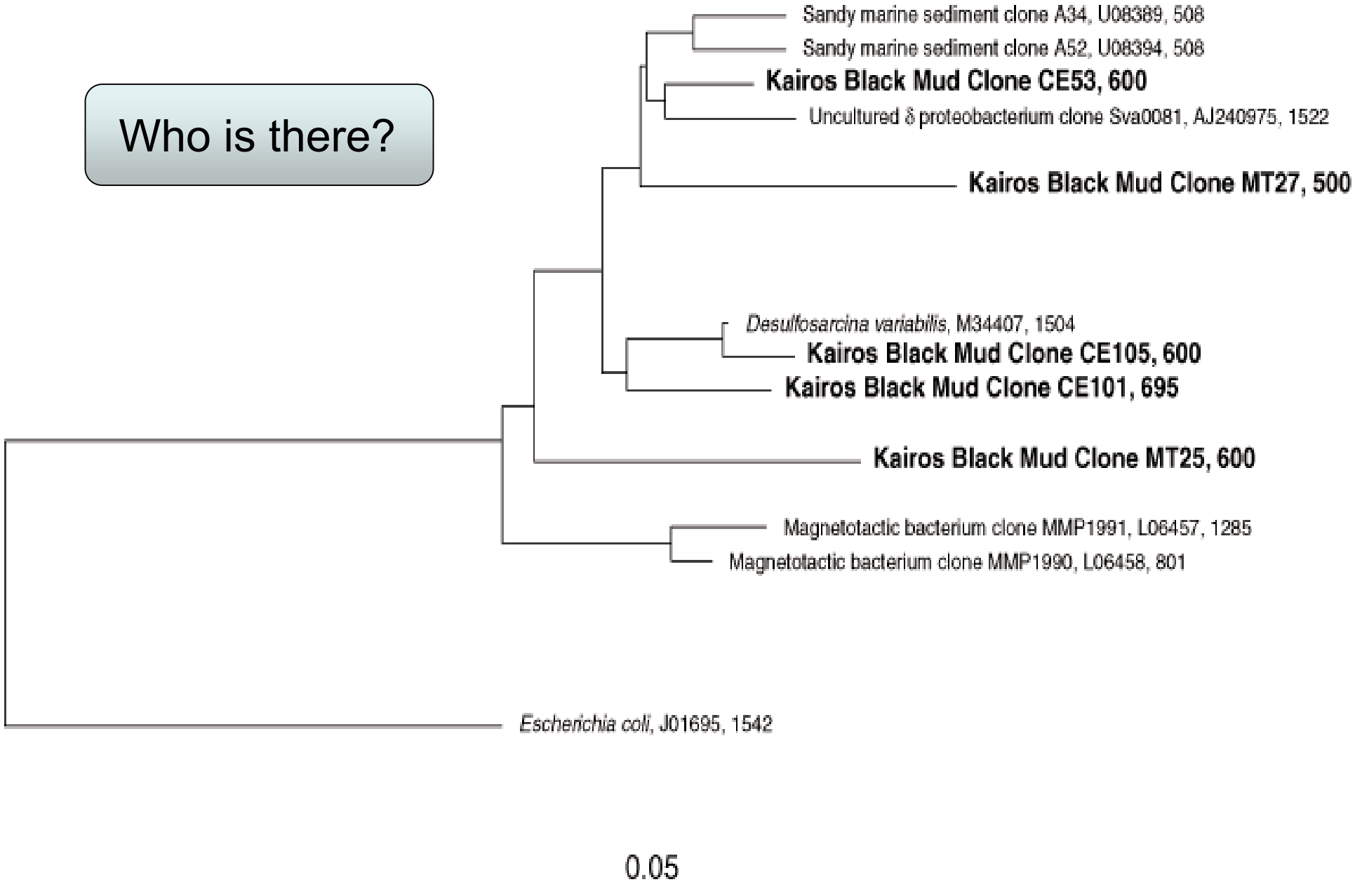
Modern genomic techniques

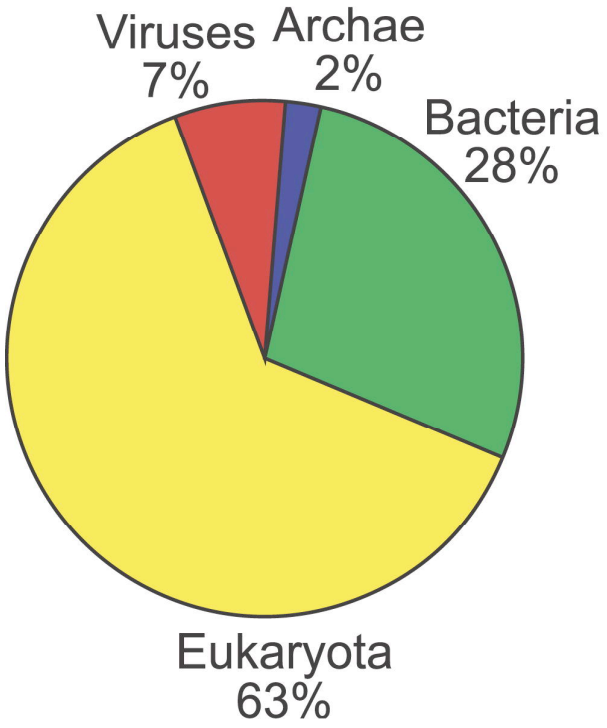
- Fragments of DNA from the environment are sequenced and then analyzed with bioinformatics
  - Assembly of fragments into longer contiguous sequences (or not)
  - Clustering based on sequence composition
  - Comparison to sequences from previously cultured organisms whose function is known

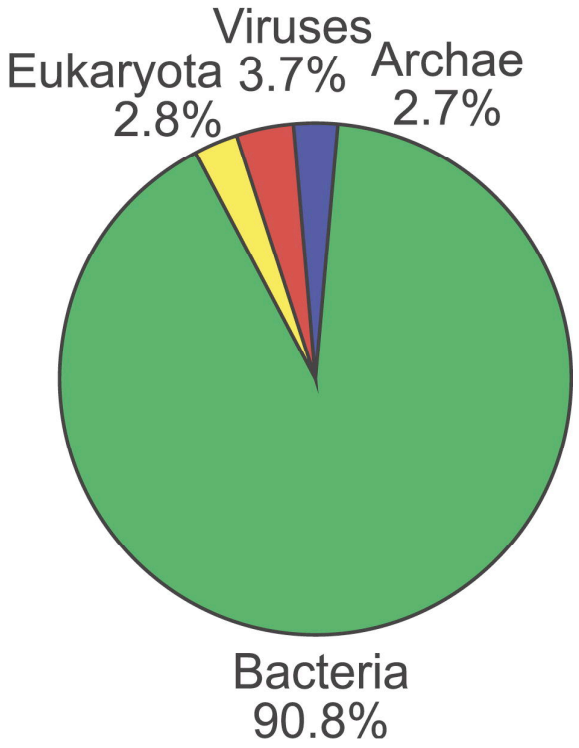
History of metagenomics

- Term coined by Handelsman in 1998
  - Cloned DNA from soil into expression vectors in E. coli - screened for biological activity
- Nearly complete genomes recovered from acid mine drainage by Tyson in 2004
- Explosion of metagenomic sequencing projects
  - Sargasso Sea, human intestine, termite hindgut, hot springs, GOS,

Who is there?


Sandy marine sediment clone A34, U08389, 508
Sandy marine sediment clone A52, U08394, 508
Kairos Black Mud Clone CE53, 600
Uncultured δ proteobacterium clone Sva0081, AJ240975, 1522
Kairos Black Mud Clone MT27, 500
Desulfosarcina variabilis, M34407, 1504
Kairos Black Mud Clone CE105, 600
Kairos Black Mud Clone CE101, 695
Kairos Black Mud Clone MT25, 600
Magnetotactic bacterium clone MMP1991, L06457, 1285
Magnetotactic bacterium clone MMP1990, L06458, 801
Escherichia coli, J01695, 1542
0.05

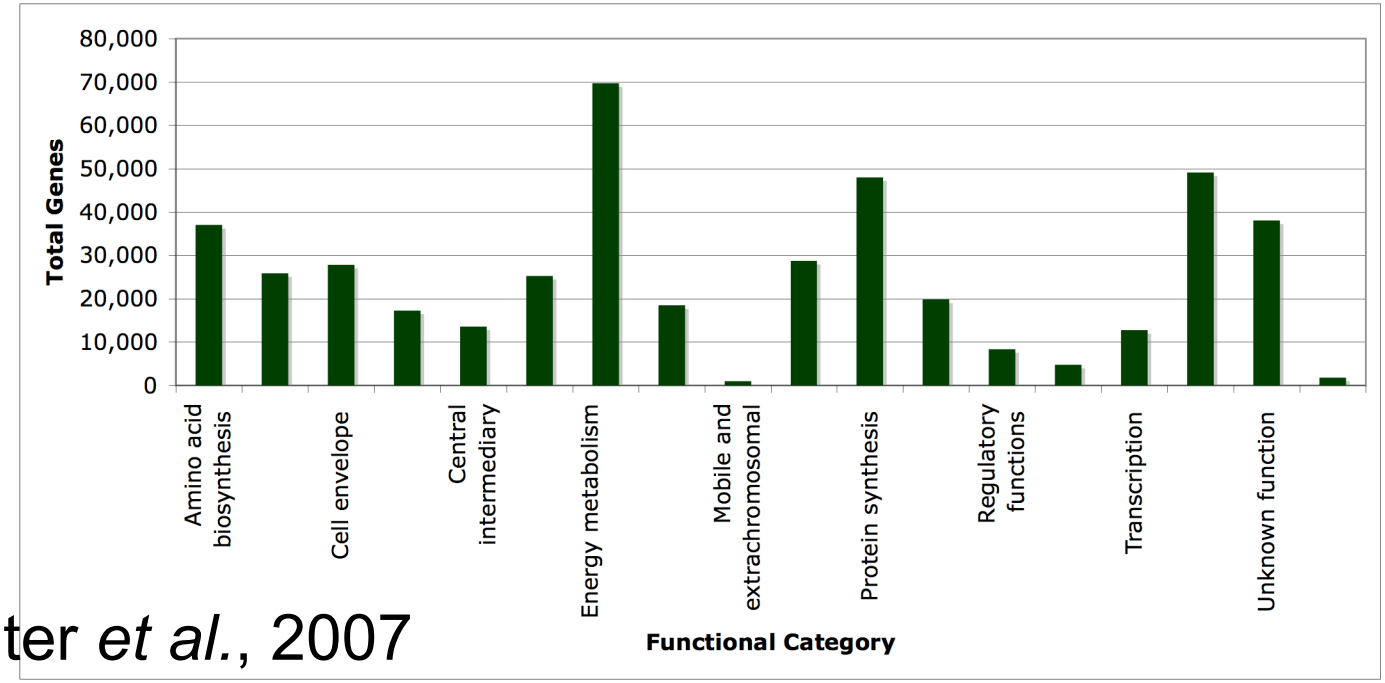




Yoosef *et al.*, 2007

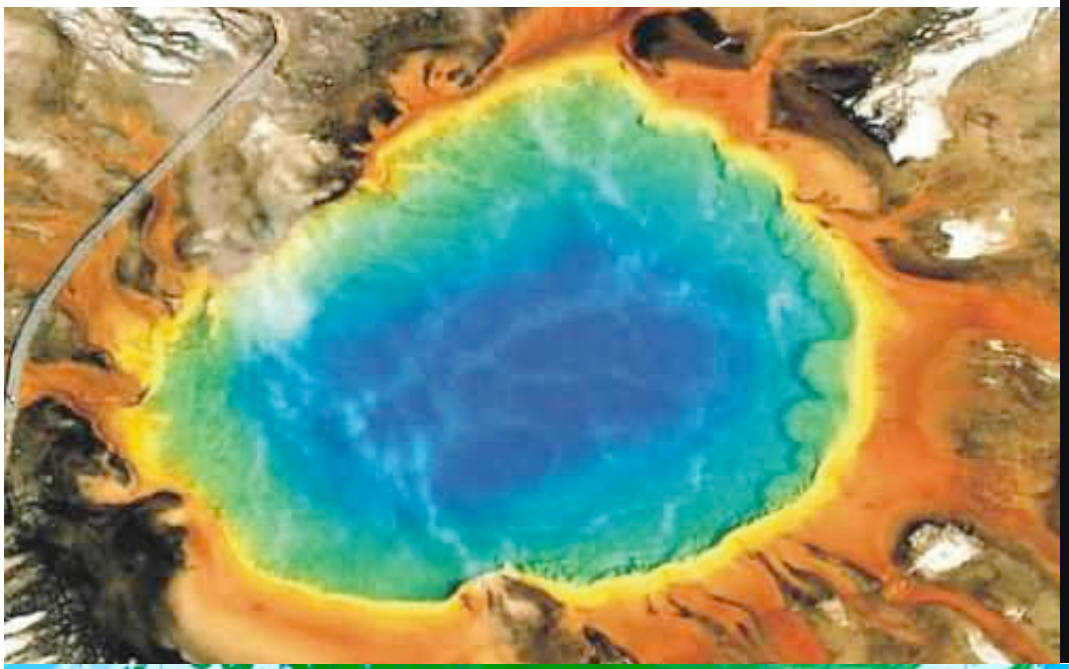
Who is there?

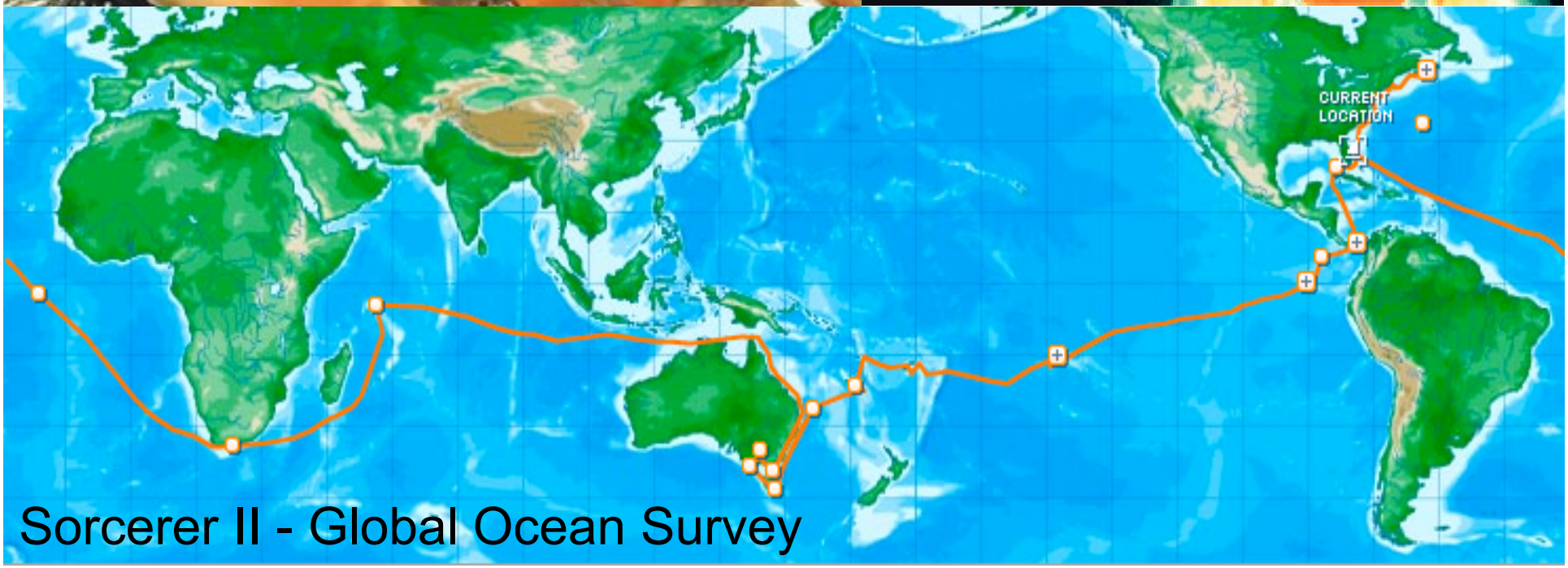
What are they doing?



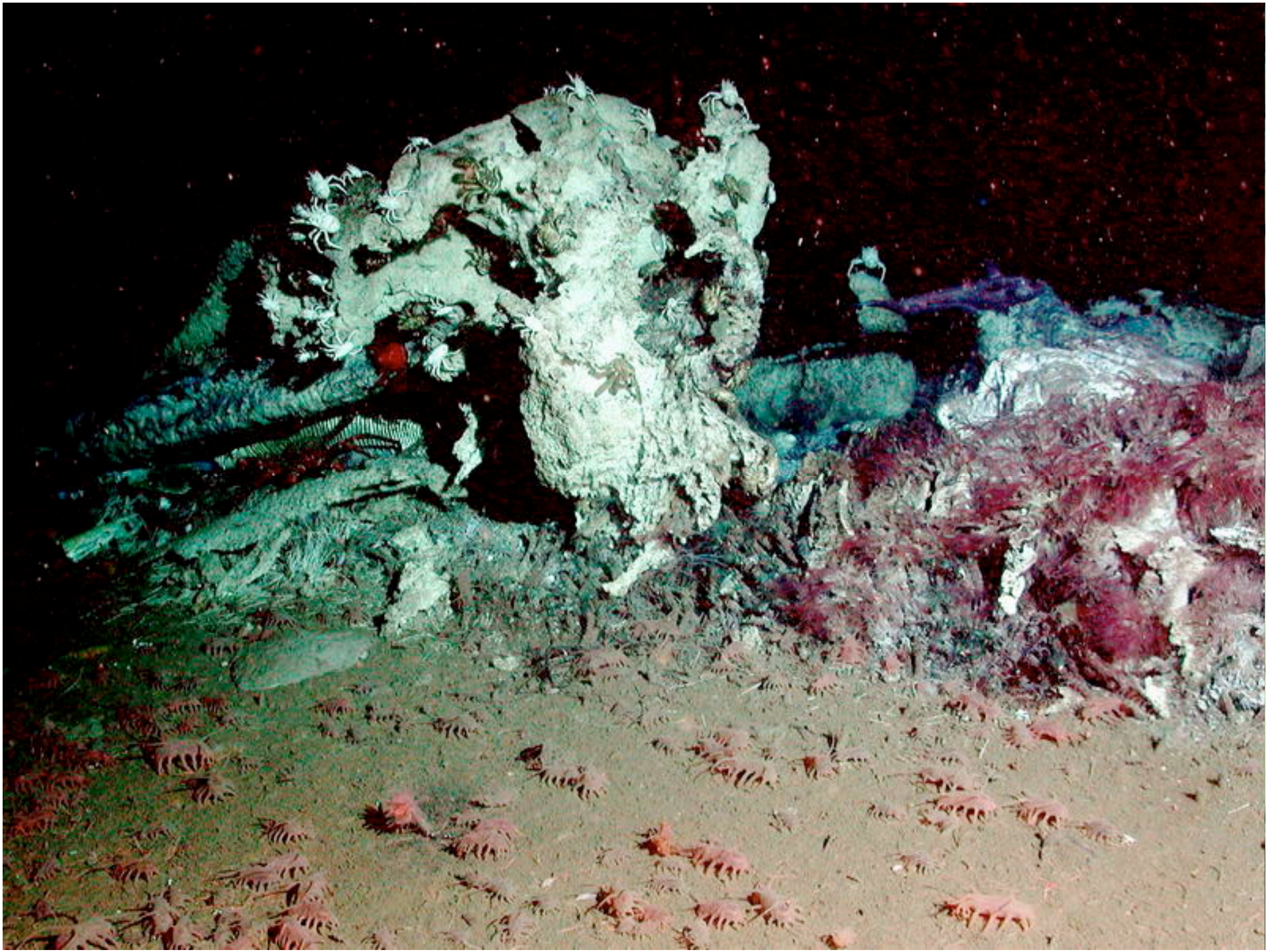
Adapted from Venter *et al.*, 2007

Grand Prismatic Spring, Yellowstone

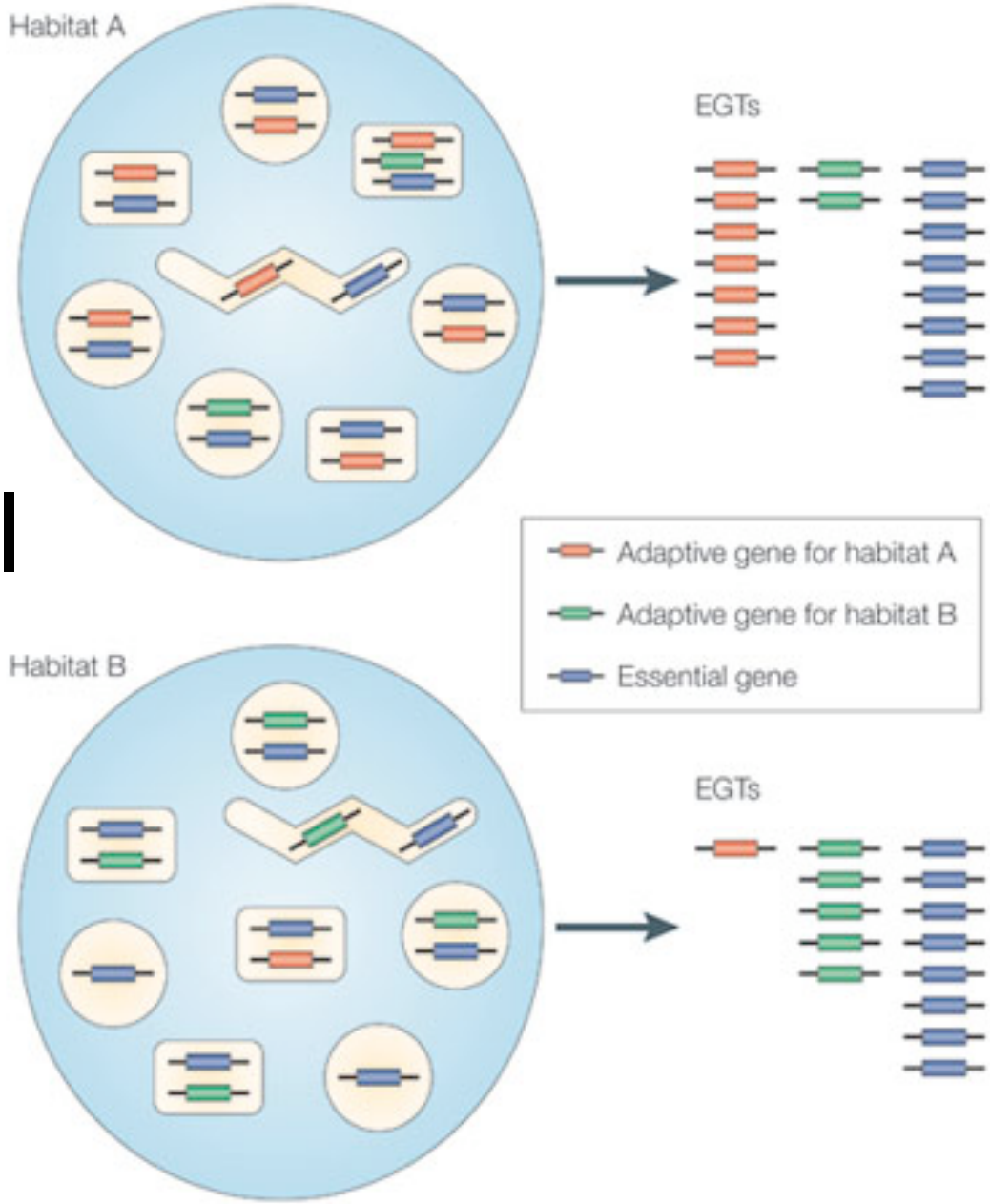




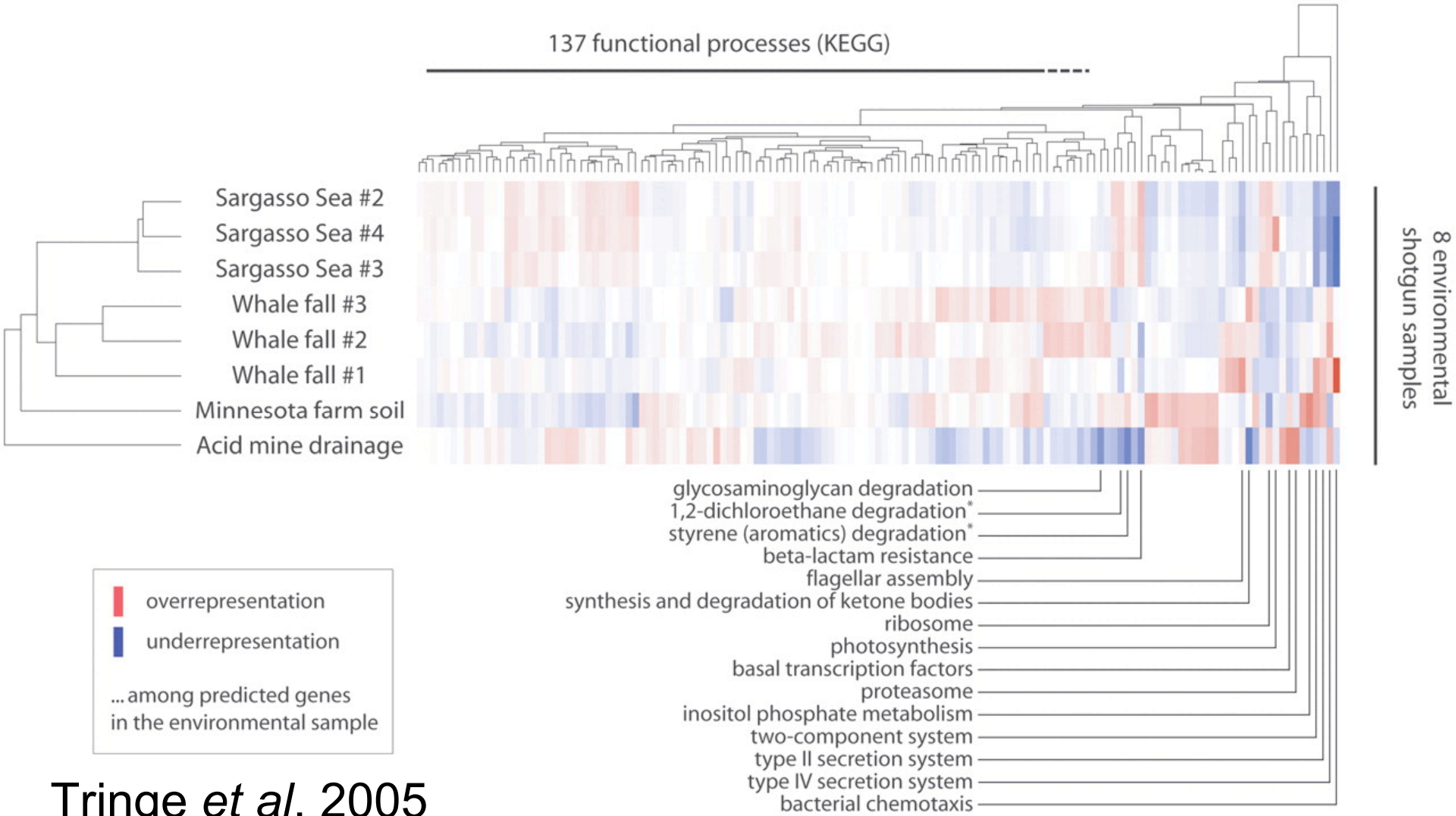
Sorcerer II - Global Ocean Survey



Environmental Gene Tags



Comparative metagenomics



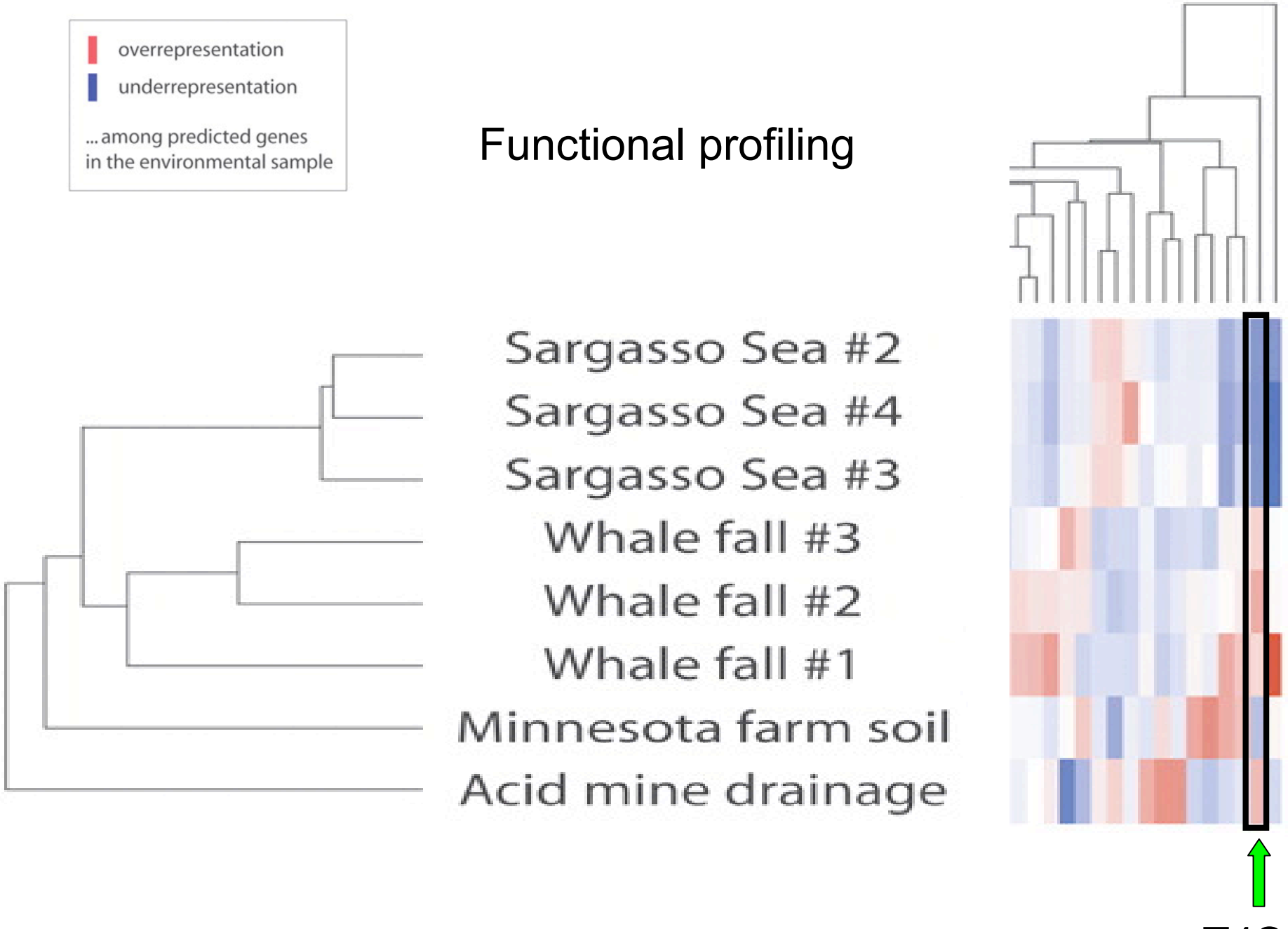
Tringe *et al*, 2005

overrepresentation

underrepresentation

... among predicted genes
in the environmental sample

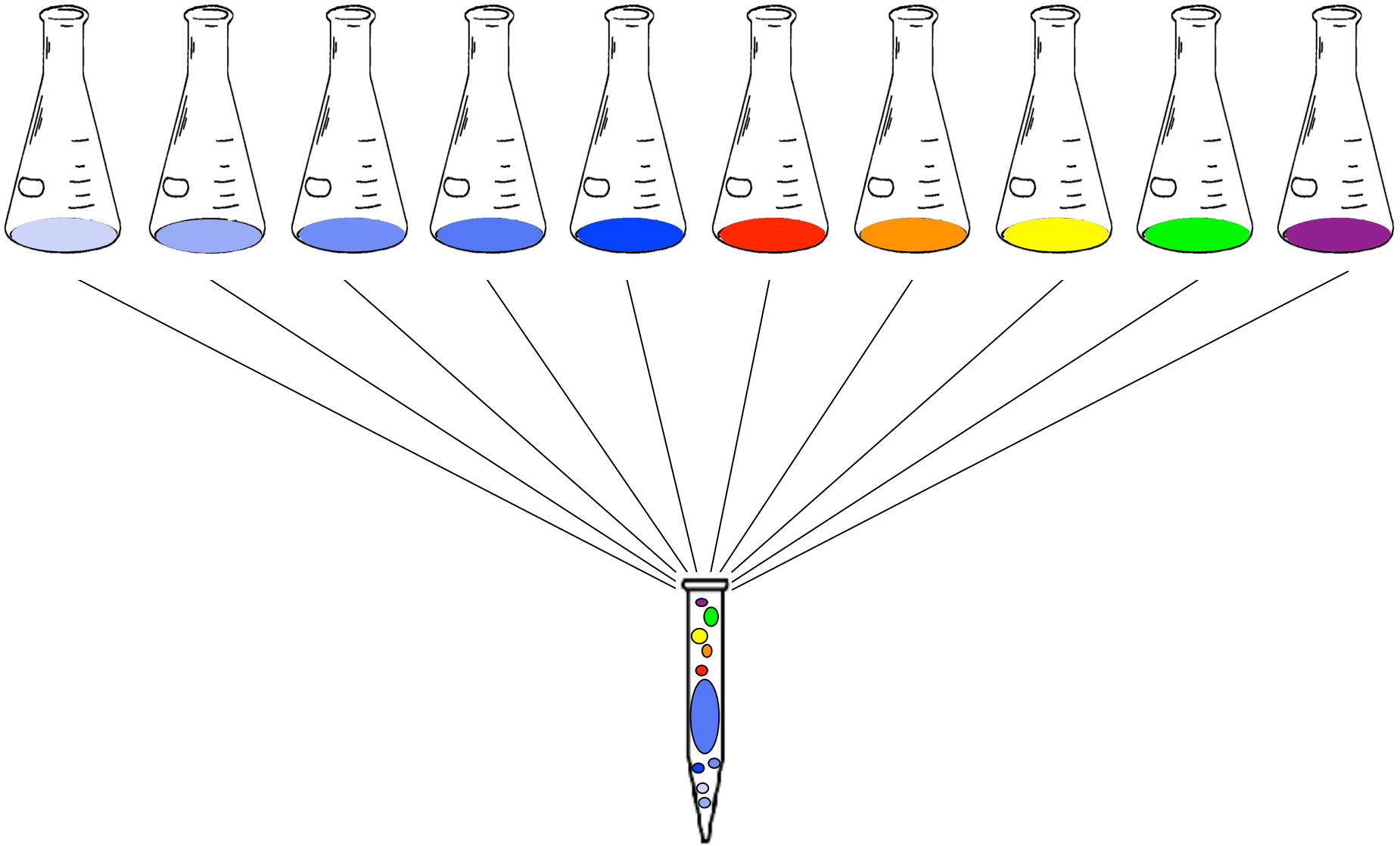
Functional profiling



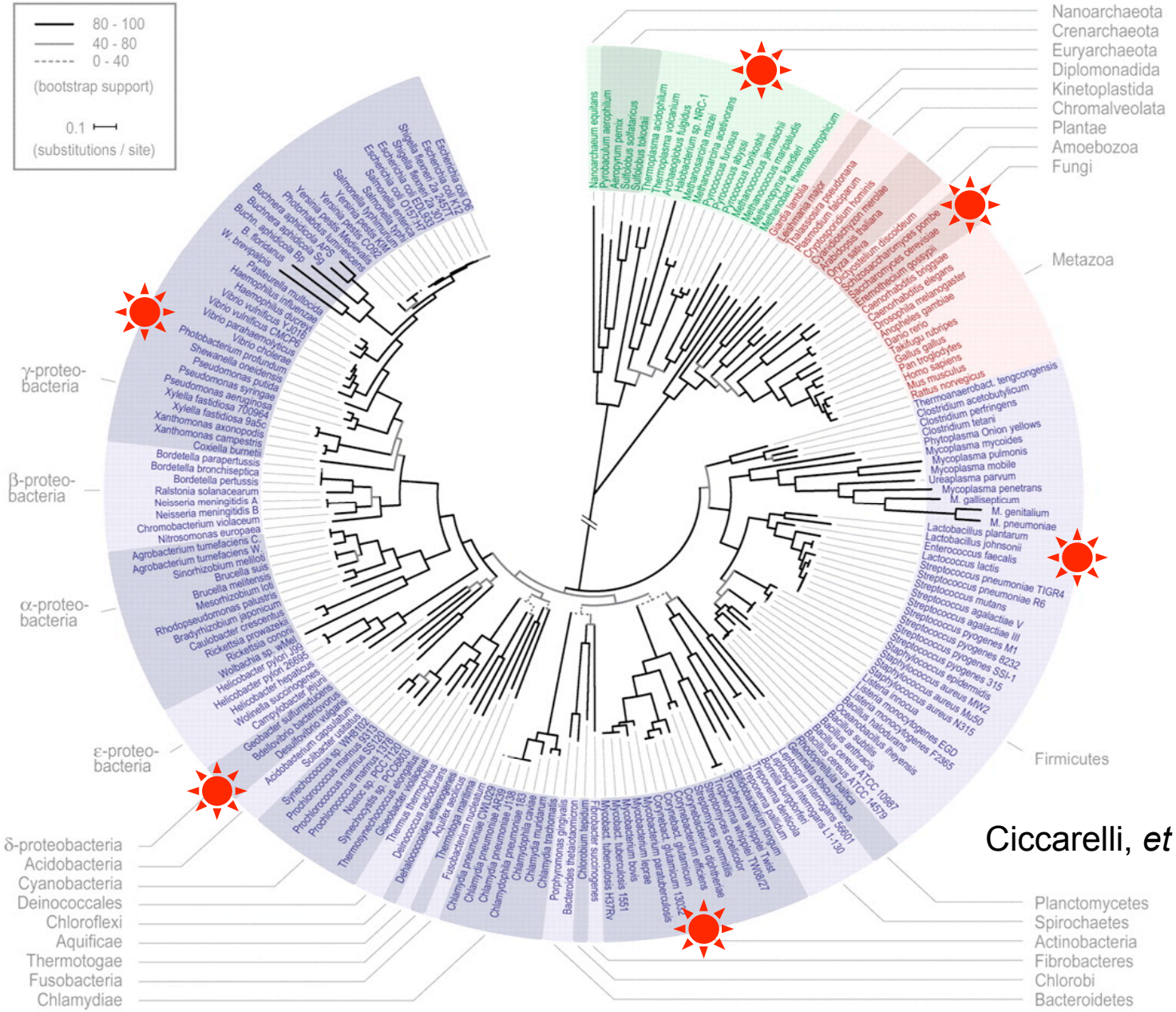
Assumptions

- All species are equally well represented in the extracted DNA
  - non-biased sampling of microbial diversity
- All of the extracted DNA is equally represented in the sequenced library
  - non-biased cloning and sequencing of the metagenomic DNA

Creating a “simulated” environmental sample



ORGANISMS USED FOR THIS STUDY:

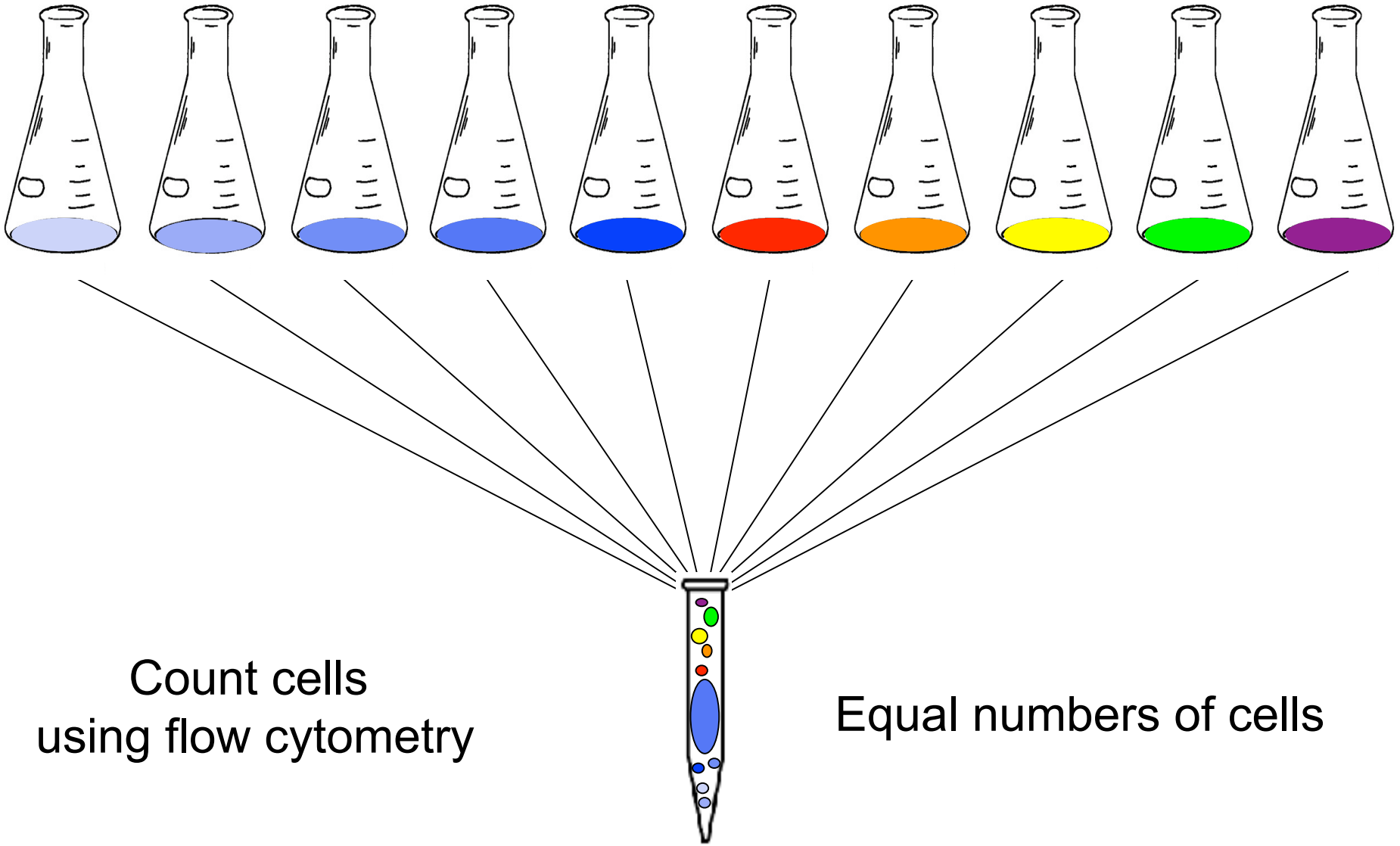


Ciccarelli, *et al.*, 2006

ORGANISMS USED FOR THIS STUDY:

Organism	GC(%)	Size (kb)
Lactococcus lactis cremoris IL1403	35.7	2529
Lactococcus lactis cremoris SK11	35.9	2560
Pediococcus pentosaceus	37.4	1800
Lactobacillus casei	46.6	2959
Lactobacillus brevis	46	2349
Shewanella amazonensis SB2B	53	4306
Myxococcus xanthus DK	68.9	9139
Acidothermus cellulolyticus 11B	66.9	2445
Saccharomyces cerevisiae S288C	38	12096
Halobacterium sp. NRC-1	65.9	2014

Creating a “simulated” environmental sample

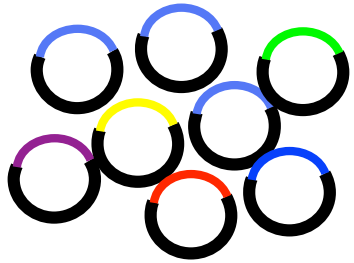


Count cells
using flow cytometry

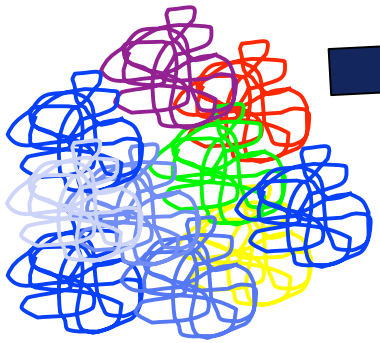
Equal numbers of cells

Binning methods are used to assign reads to organisms, and the inferred community is compared to the known community composition.

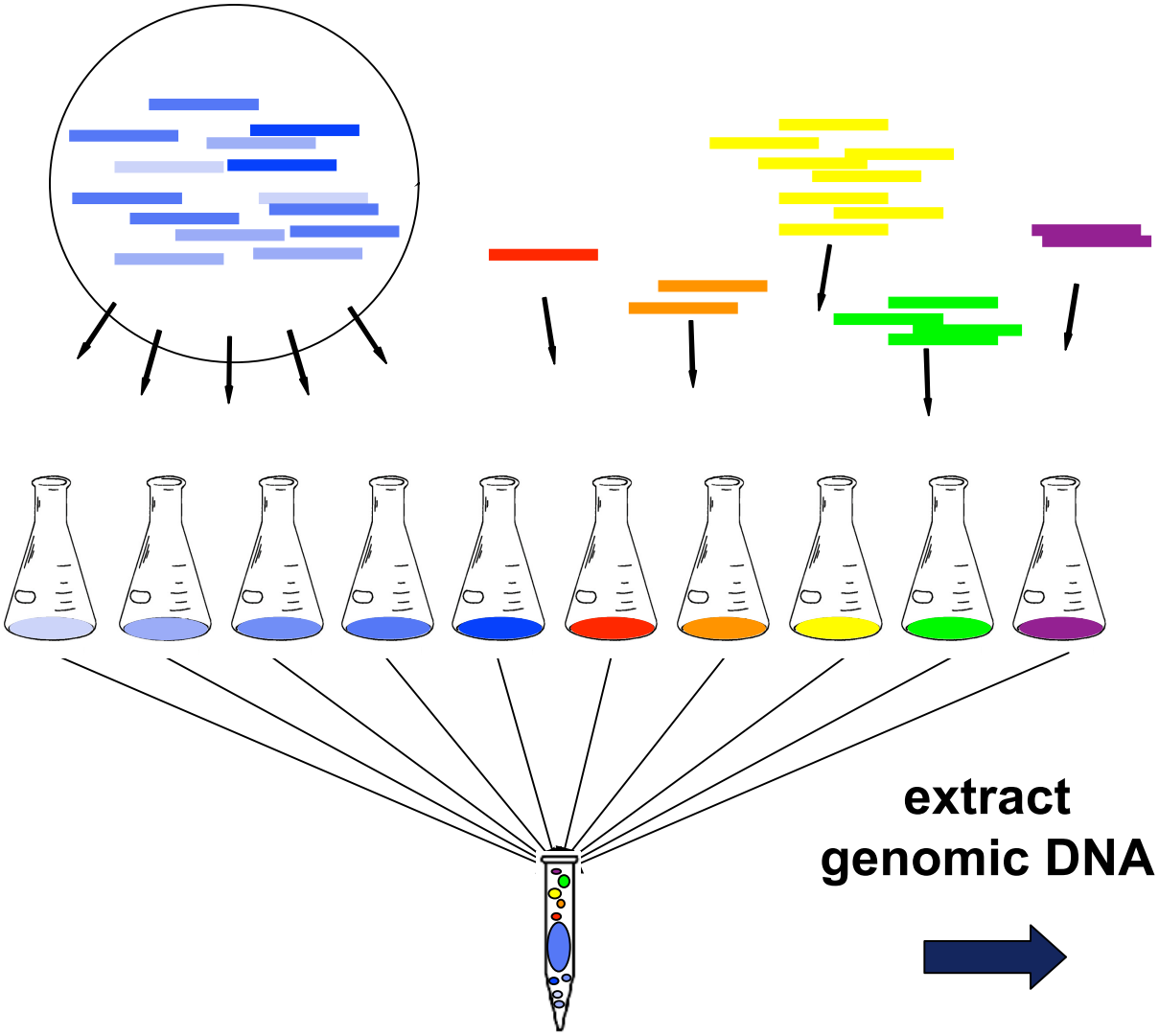
sequence clones



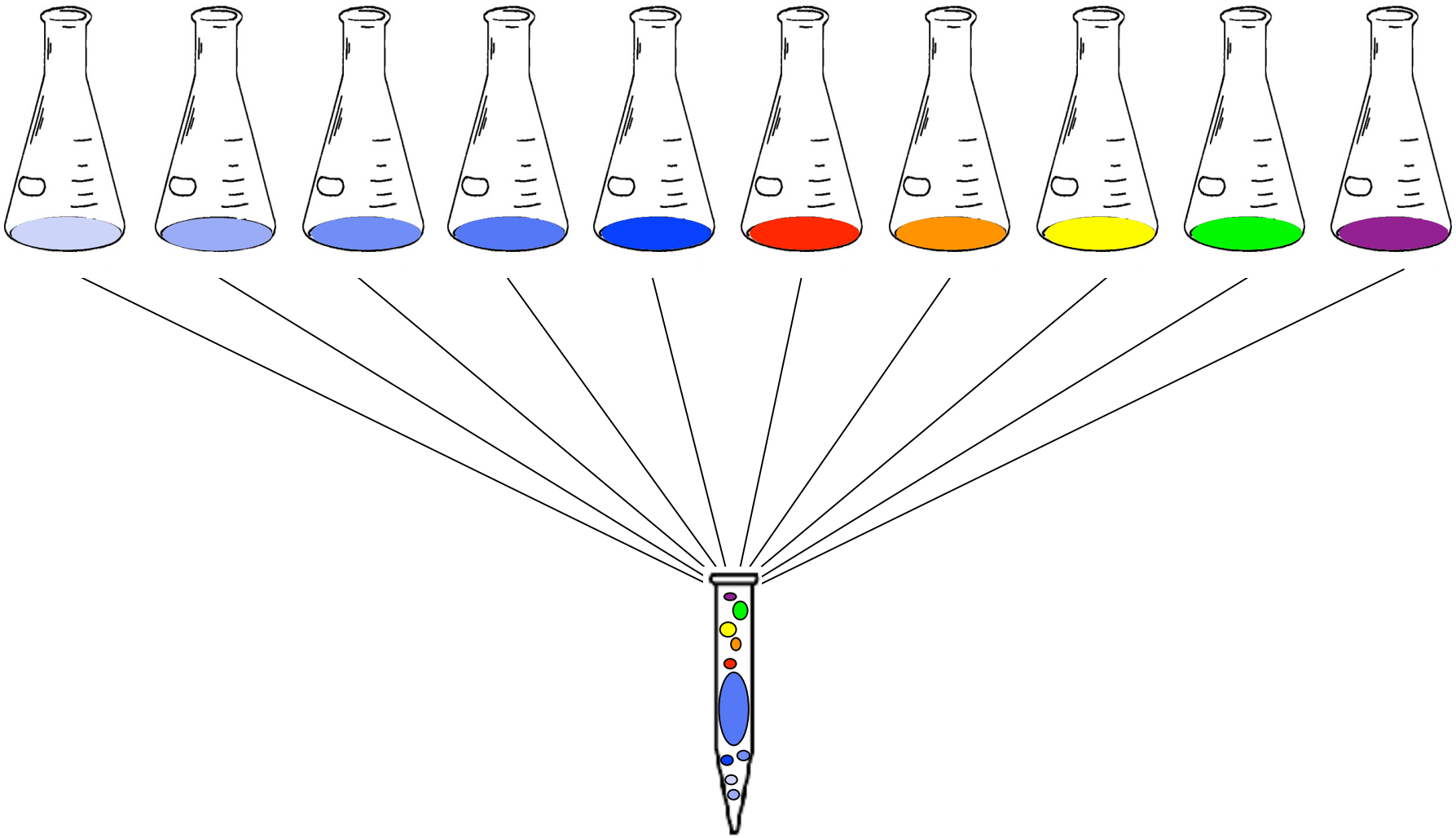
shotgun library

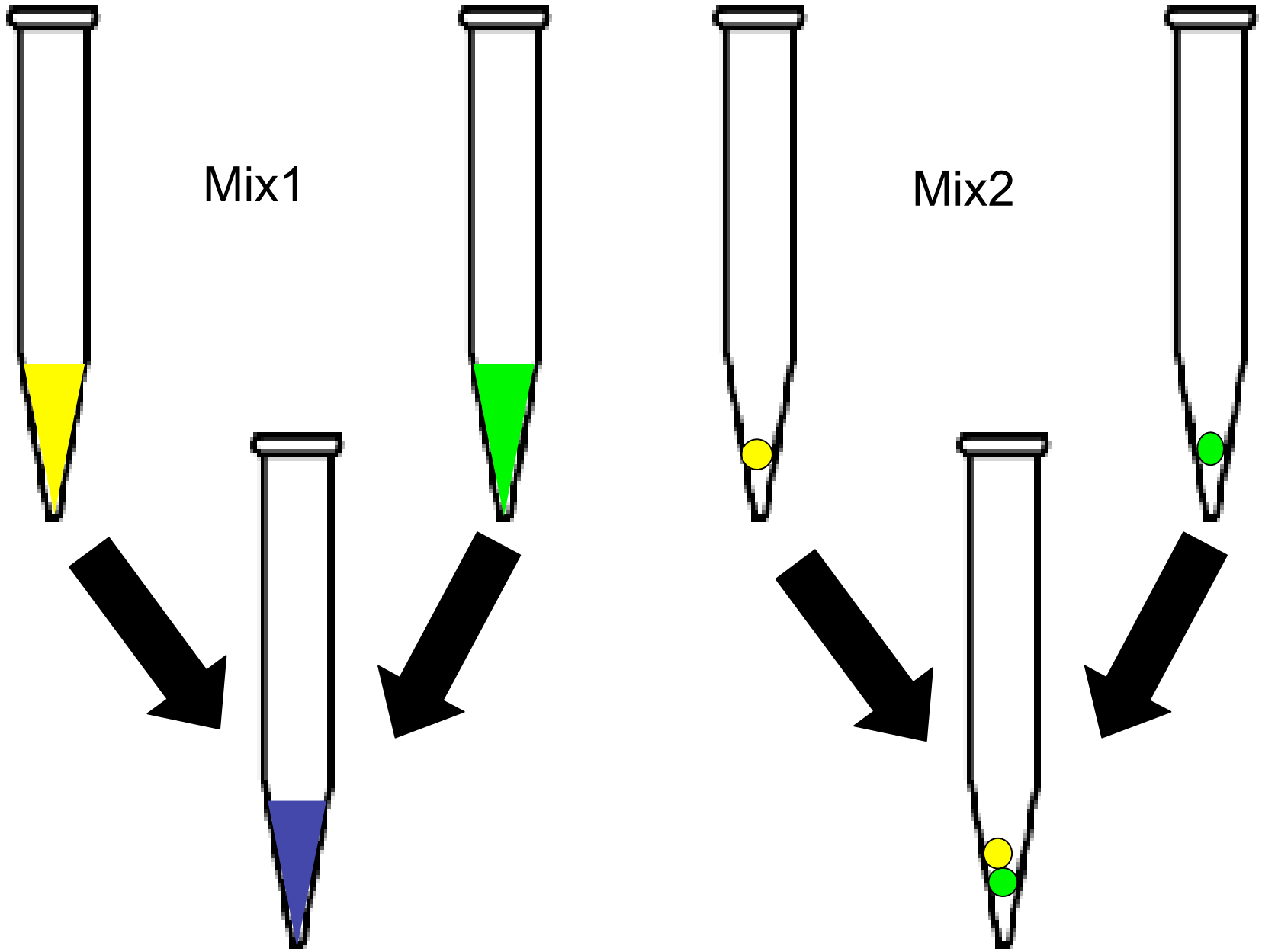


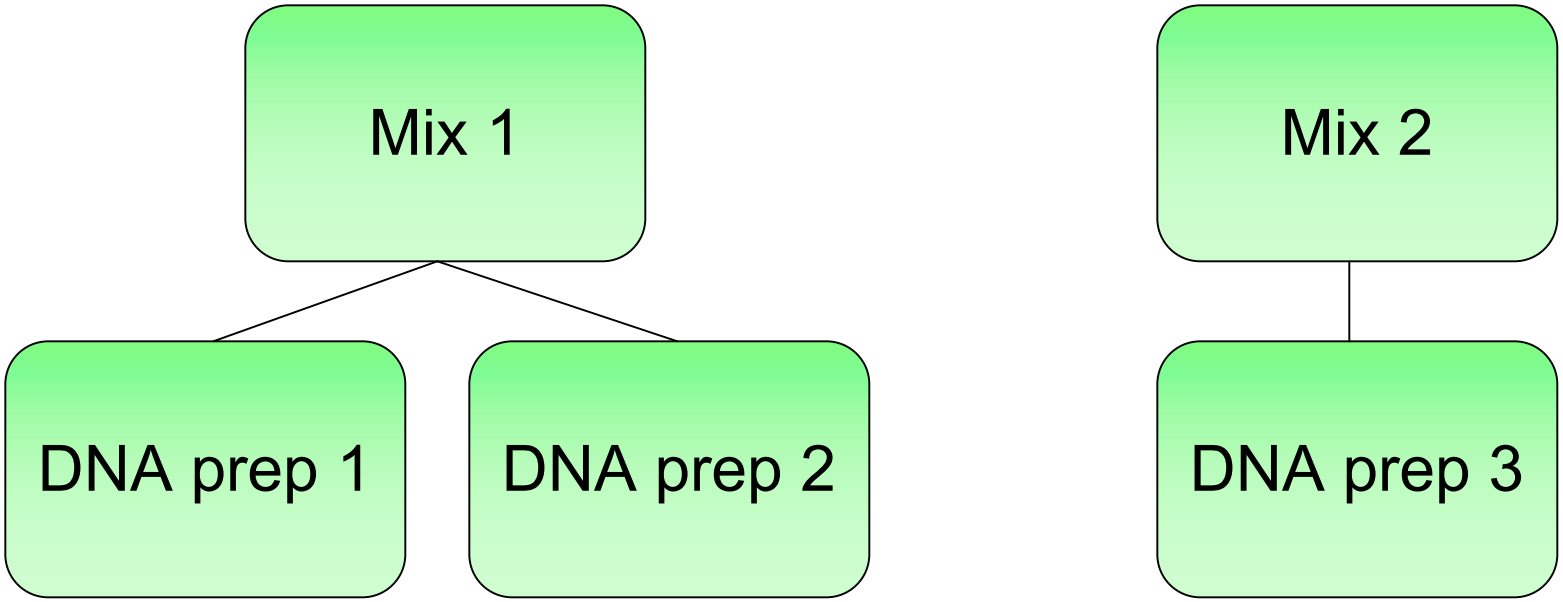
extract genomic DNA



Mixing




Mix1
Mix2

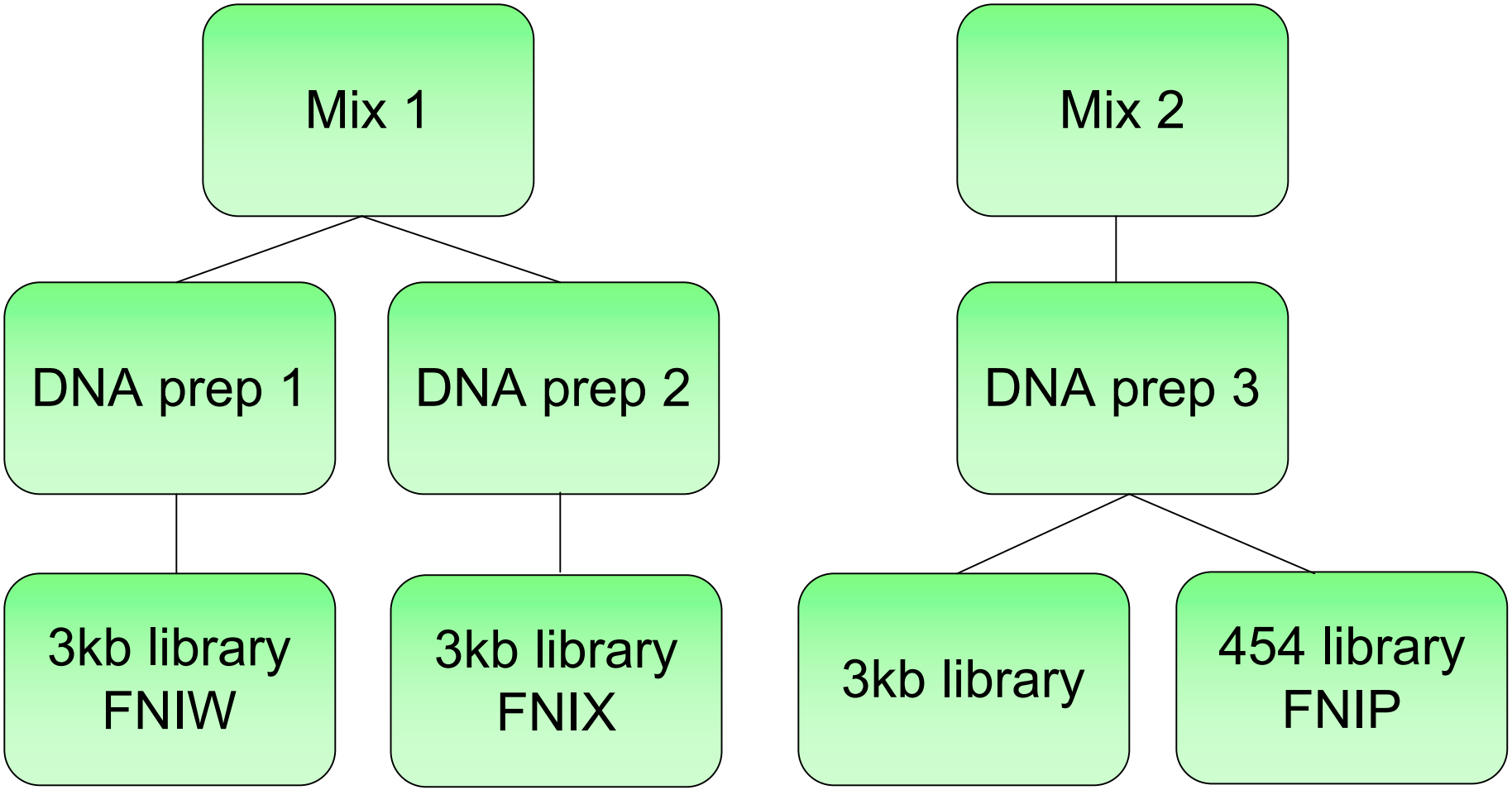

Mix 1
Mix 2
DNA prep 1
DNA prep 2
DNA prep 3

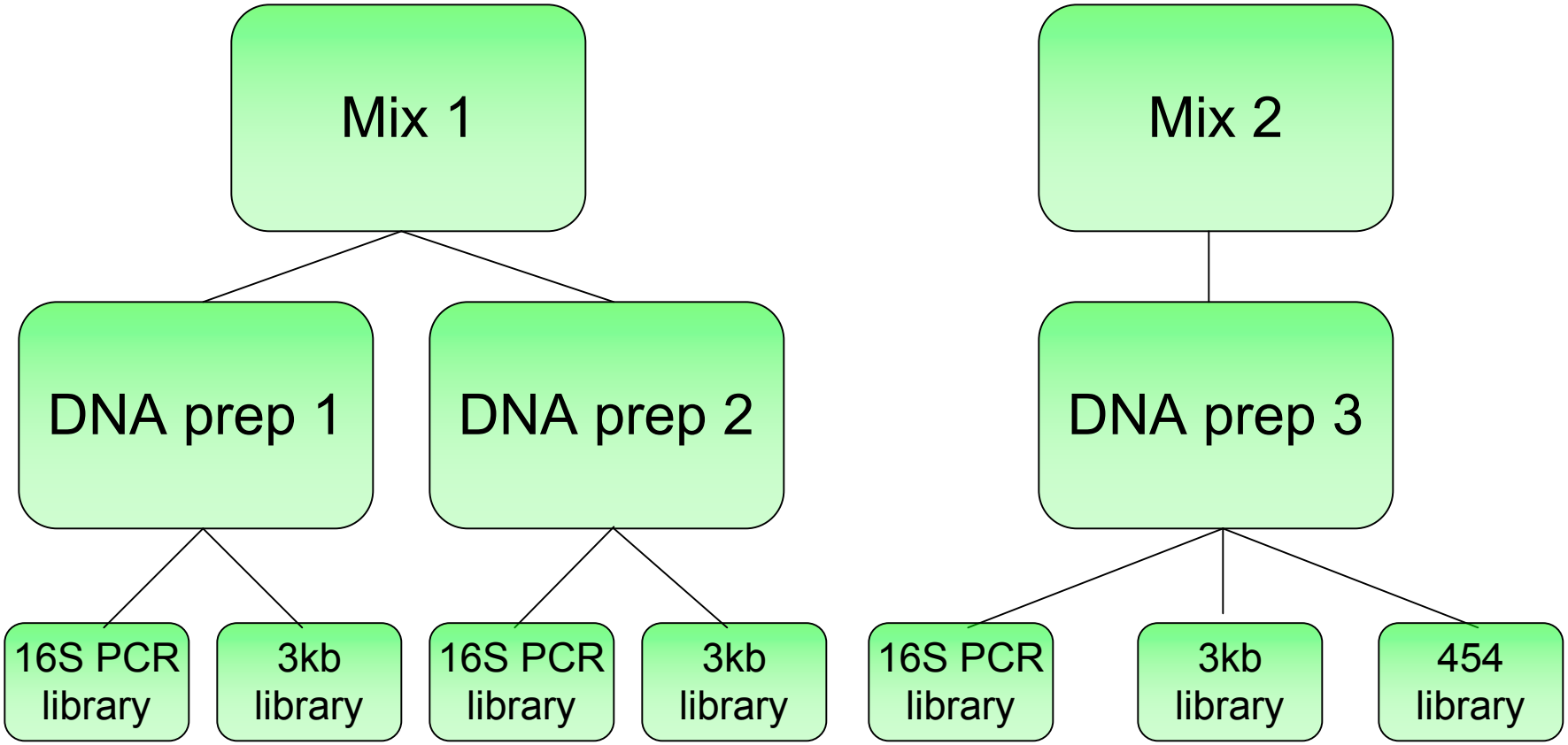
DNA prep 3

- **Enzymatic lysis**
  - Lysozyme
  - Proteinase-K
- **Mechanical disruption**
  - Bead beater
- **DNA extraction**
  - Phenol/cholorofm
  - Ethanol precipitation

DNA prep 2

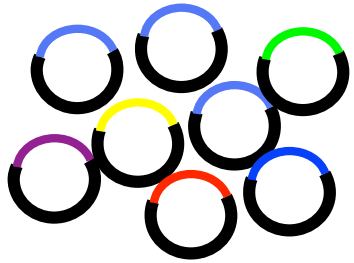
- **Qiagen Dneasy Kit**
  - Alkaline lysis
  - Bind DNA to column
  - Wash column
  - Elute DNA


Mix 1
Mix 2
DNA prep 1
DNA prep 2
DNA prep 3
3kb library FNIW
3kb library FNIX
3kb library
454 library FNIP

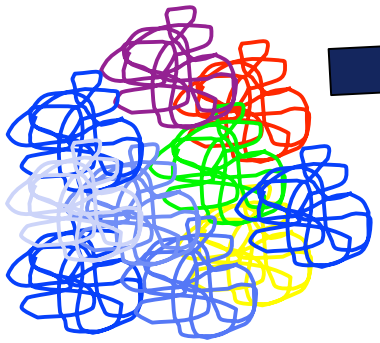

Mix 1
Mix 2
DNA prep 1
DNA prep 2
DNA prep 3
16S PCR library
3kb library
16S PCR library
3kb library
16S PCR library
3kb library
454 library

Binning methods are used to assign reads to organisms, and the inferred community is compared to the known community composition.

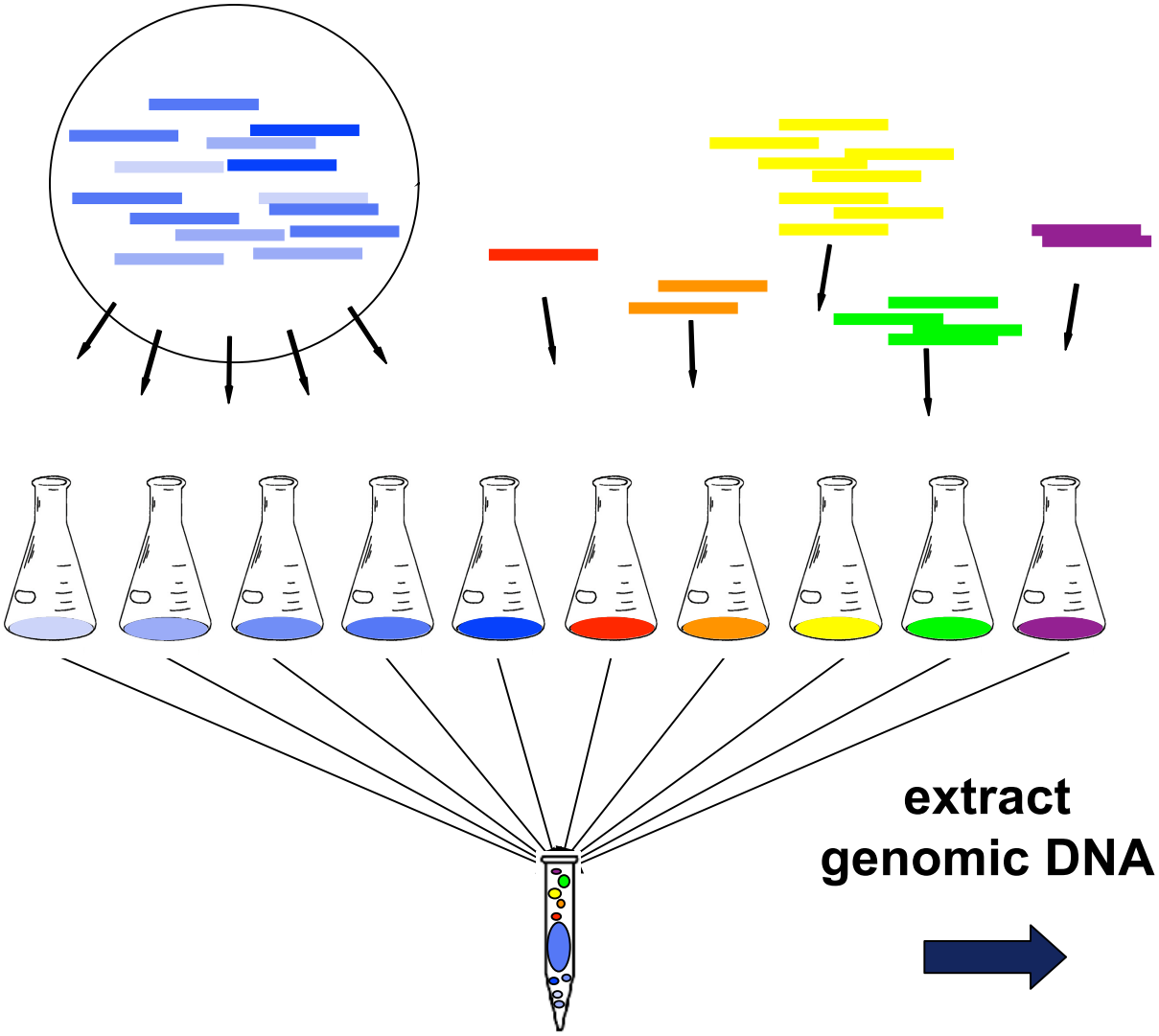
sequence clones



shotgun library



extract genomic DNA

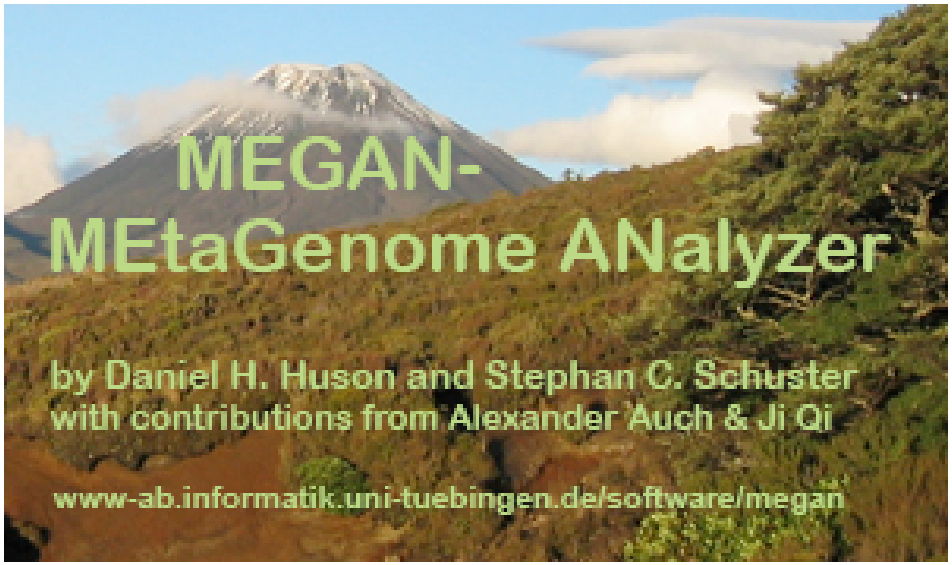


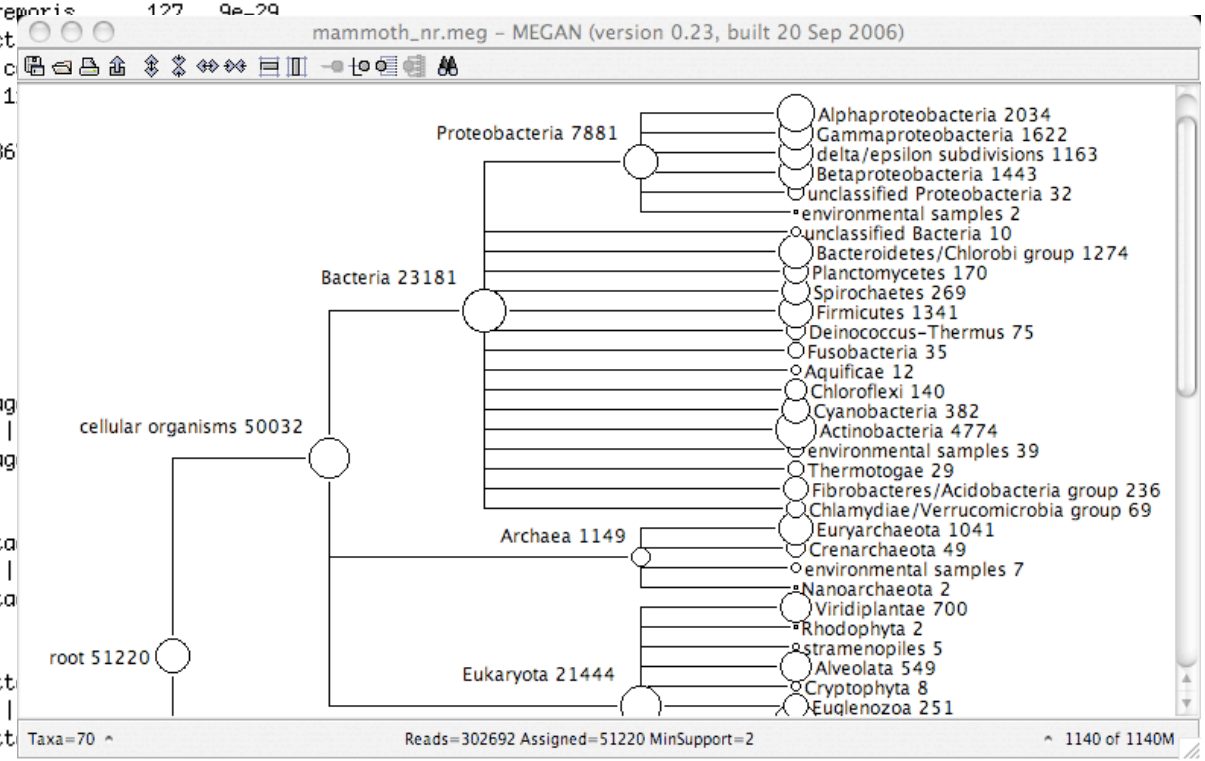
Best Case Scenario

- We know how many sequences we should obtain from each organism

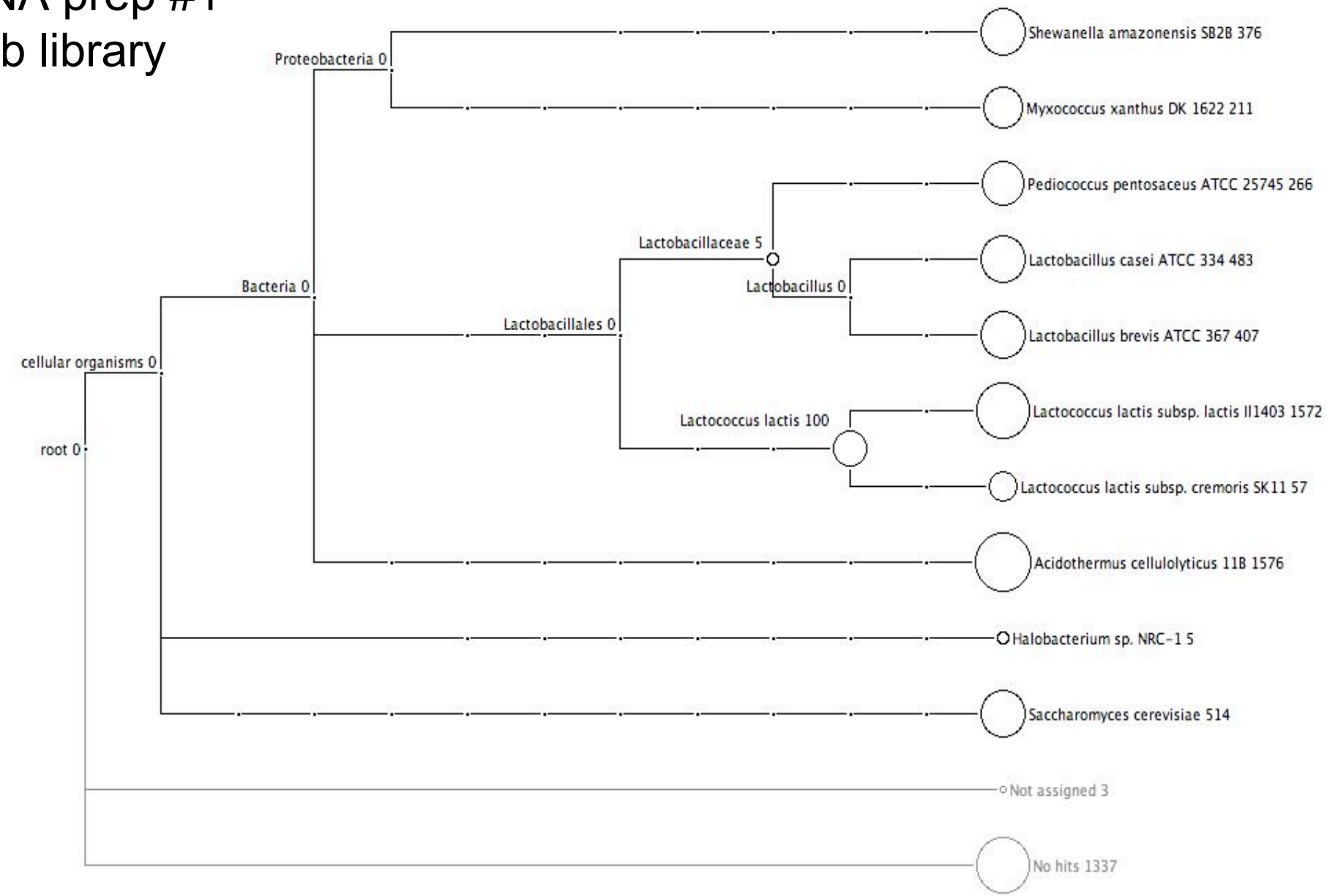
- We have the complete genome sequence of each organism

```
Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs",  Nucleic Acids Res. 25:3389-3402.

Query= FNIW780.g4  479      0    479  SCF
         (479 letters)

Database: blast_db/AllMembers
           37 sequences; 42,684,328 total letters

Searching.done

                                                                 Score    E
Sequences producing significant alignments:                      (bits) Value

gi|116332681|ref|NC_008497.1| Lactobacillus brevis ATCC 367, com...   706   0.0
gi|116491818|ref|NC_008525.1| Pediococcus pentosaceus ATCC 25745...   224   5e-58
gi|116493574|ref|NC_008526.1| Lactobacillus casei ATCC 334, comp...   170   7e-42
gi|116510843|ref|NC_008527.1| Lactococcus lactis subsp. cremoris      127   9e-29
gi|15671982|ref|NC_002662.1| Lactococcus lactis subsp. lact
gi|108756767|ref|NC_008095.1| Myxococcus xanthus DK 1622, c
gi|117927211|ref|NC_008578.1| Acidothermus cellulolyticus 1

>gi|116332681|ref|NC_008497.1| Lactobacillus brevis ATCC 36
             genome
          Length = 2291220

 Score =  706 bits (356), Expect = 0.0
 Identities = 368/372 (98%)
 Strand = Plus / Minus


Query: 24     ccttaggatactctcctcgactacctgtgtcggtttgcggtacag
               ||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 566660 ccttaggatactctcctcgactacctgtgtcggtttgcggtacag


Query: 84     ctagaagcttttctcggcagtgtgacatctggcgcttccctacta
               ||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 566600 ctagaagcttttctcggcagtgtgacatctggcgcttccctacta


Query: 144    cacgccttgtccttagcgataagcatttgactcatcaccagactt
               ||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 566540 cacgccttgtccttagcgataagcatttgactcatcaccagactt
```



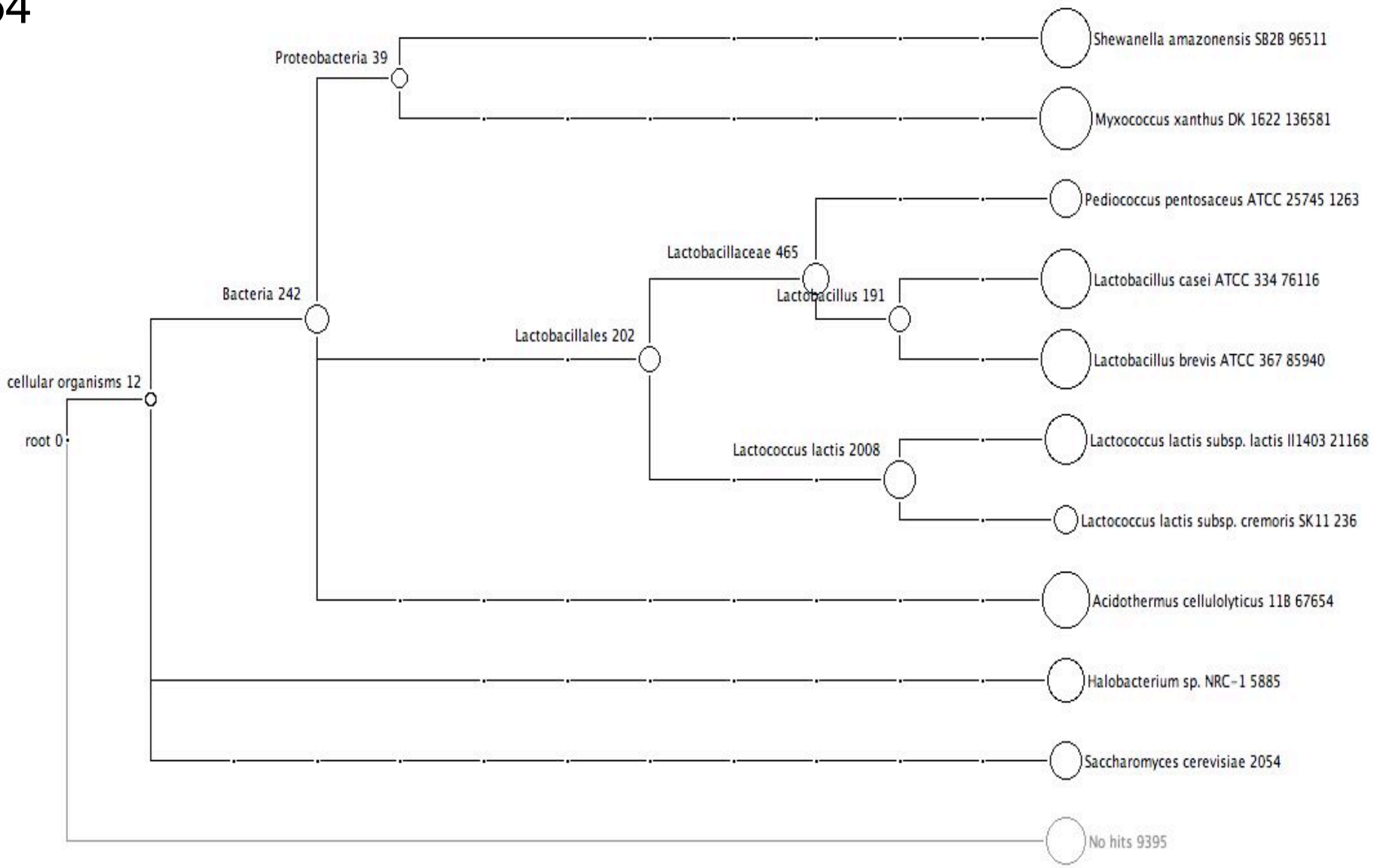


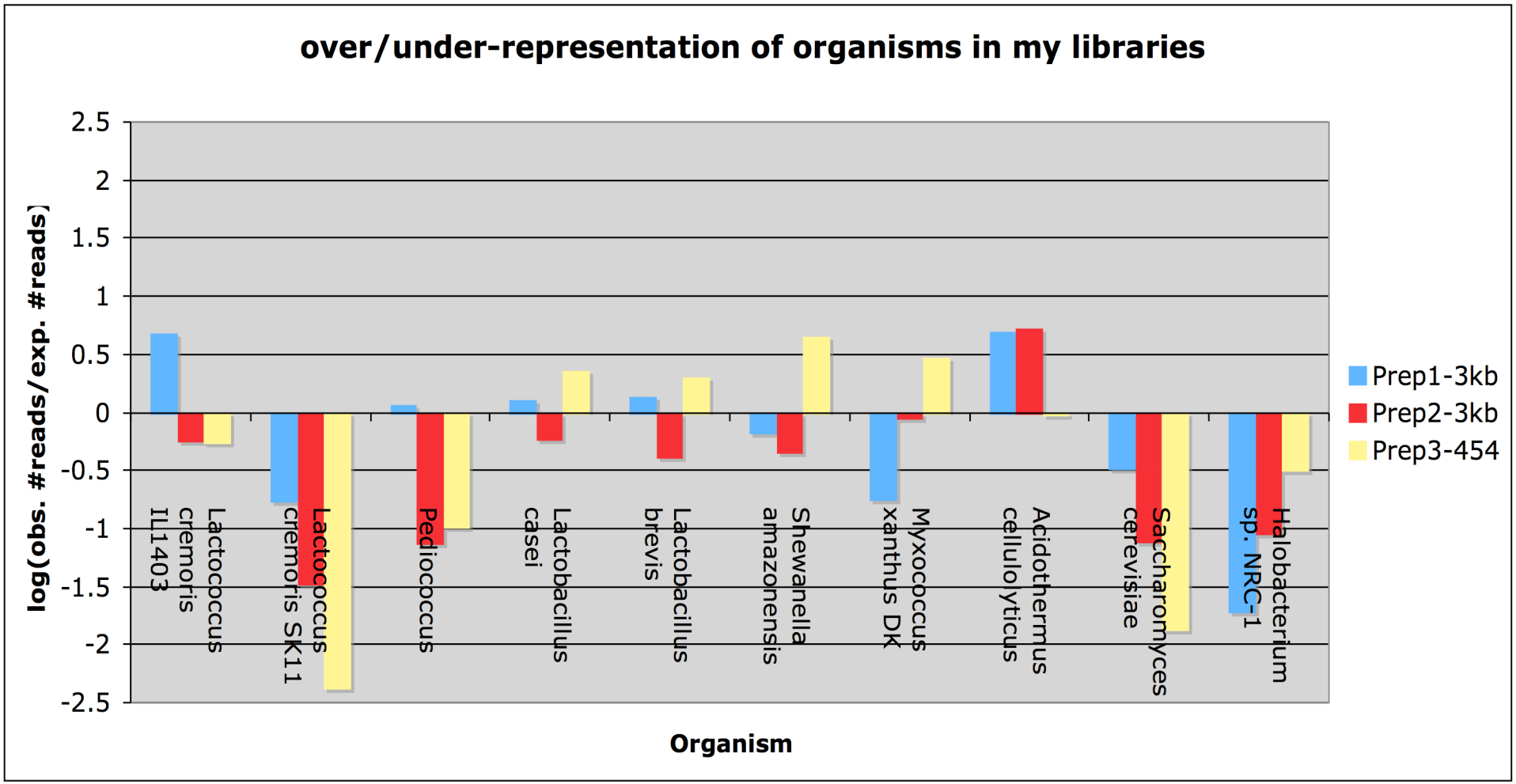
DNA prep #1
3kb library



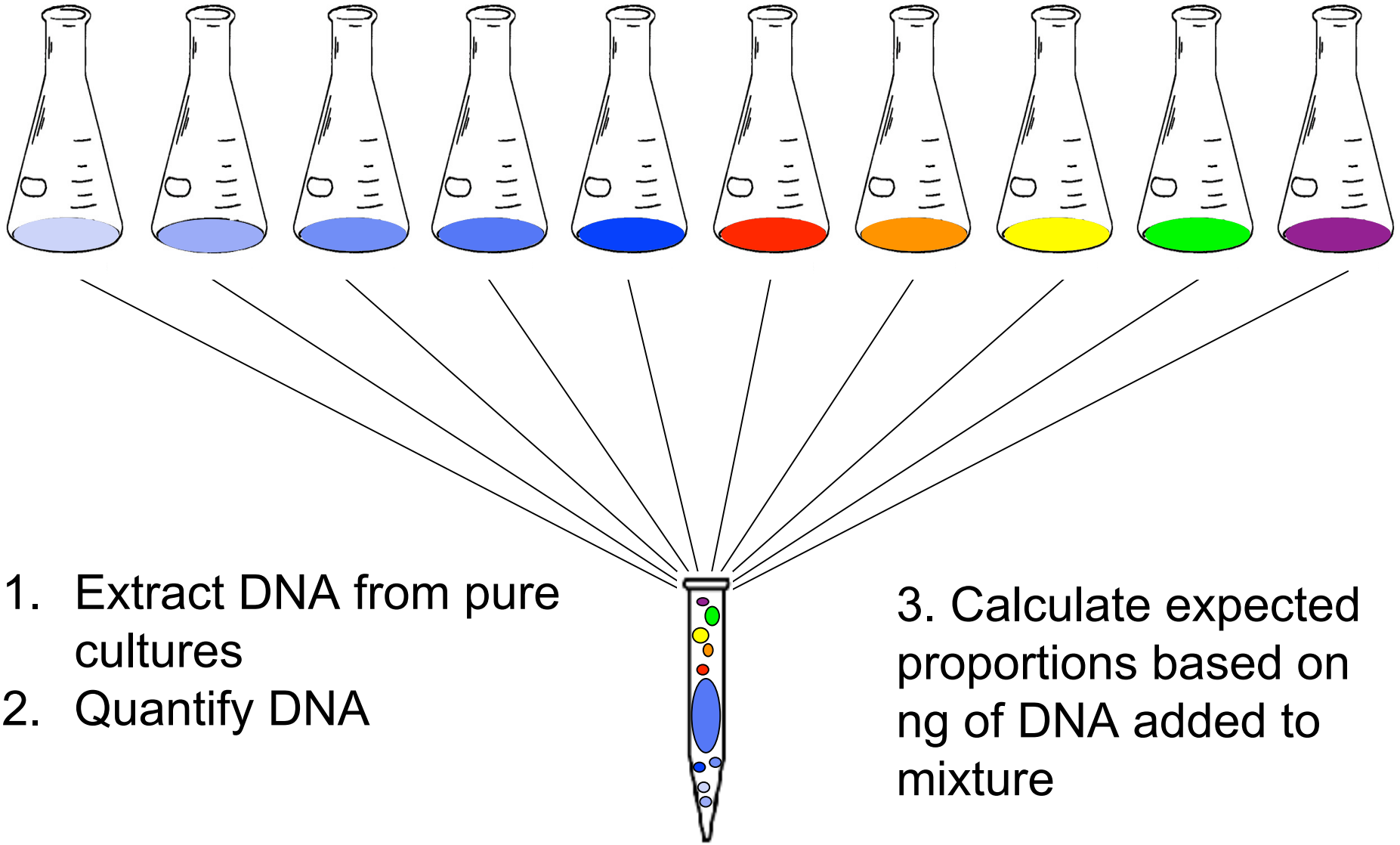
DNA prep #2
3kb library

DNA prep #3
454



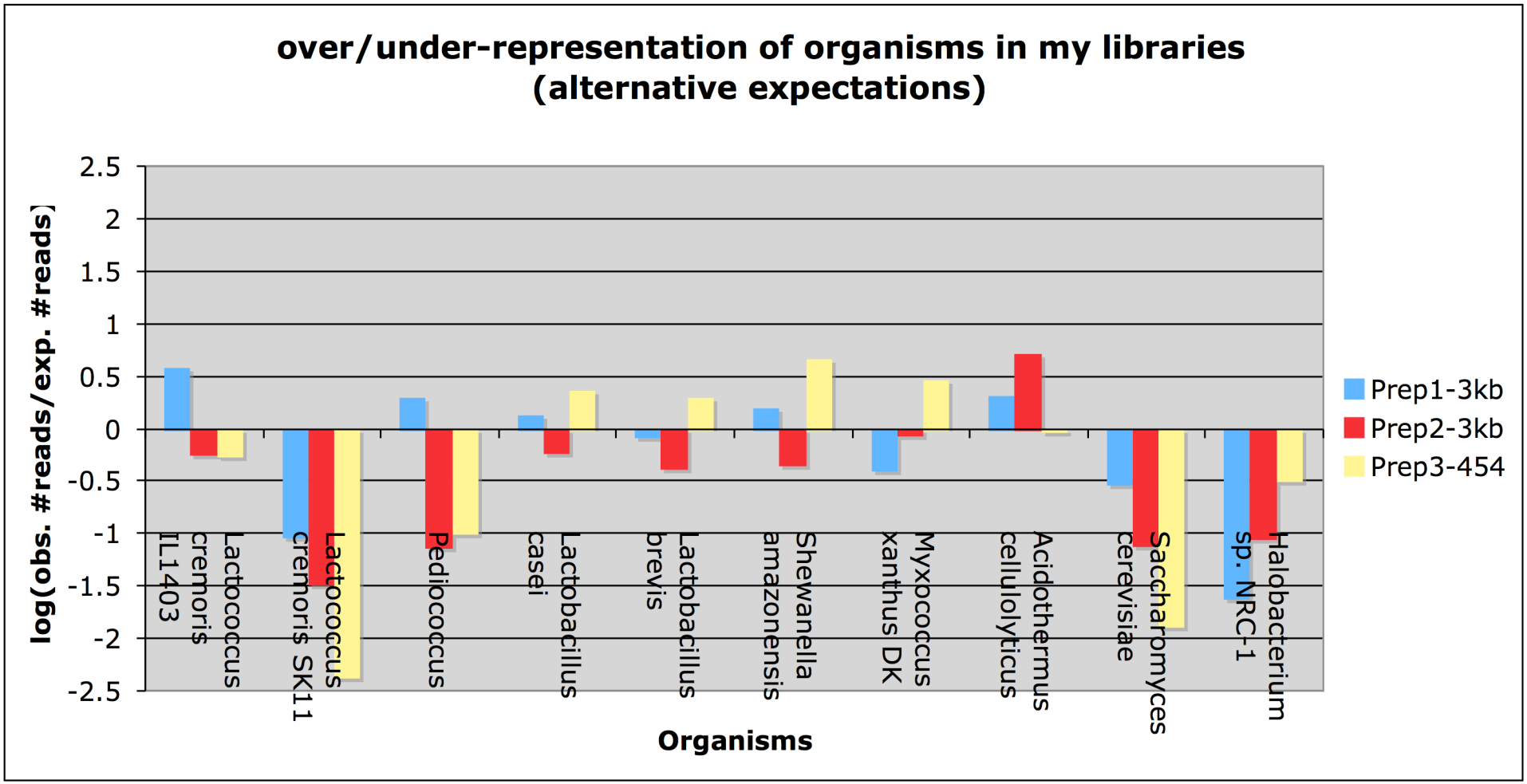

over/under-representation of organisms in my libraries
log(obs. #reads/exp. #reads)
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
Lactococcus cremoris IL1403
Lactococcus cremoris SK11
Pediococcus
Lactobacillus casei
Lactobacillus brevis
Shewanella amazonensis
Myxococcus xanthus DK
Acidothermus cellulolyticus
Saccharomyces cerevisiae
Halobacterium sp. NRC-1
Organism
Prep1-3kb
Prep2-3kb
Prep3-454

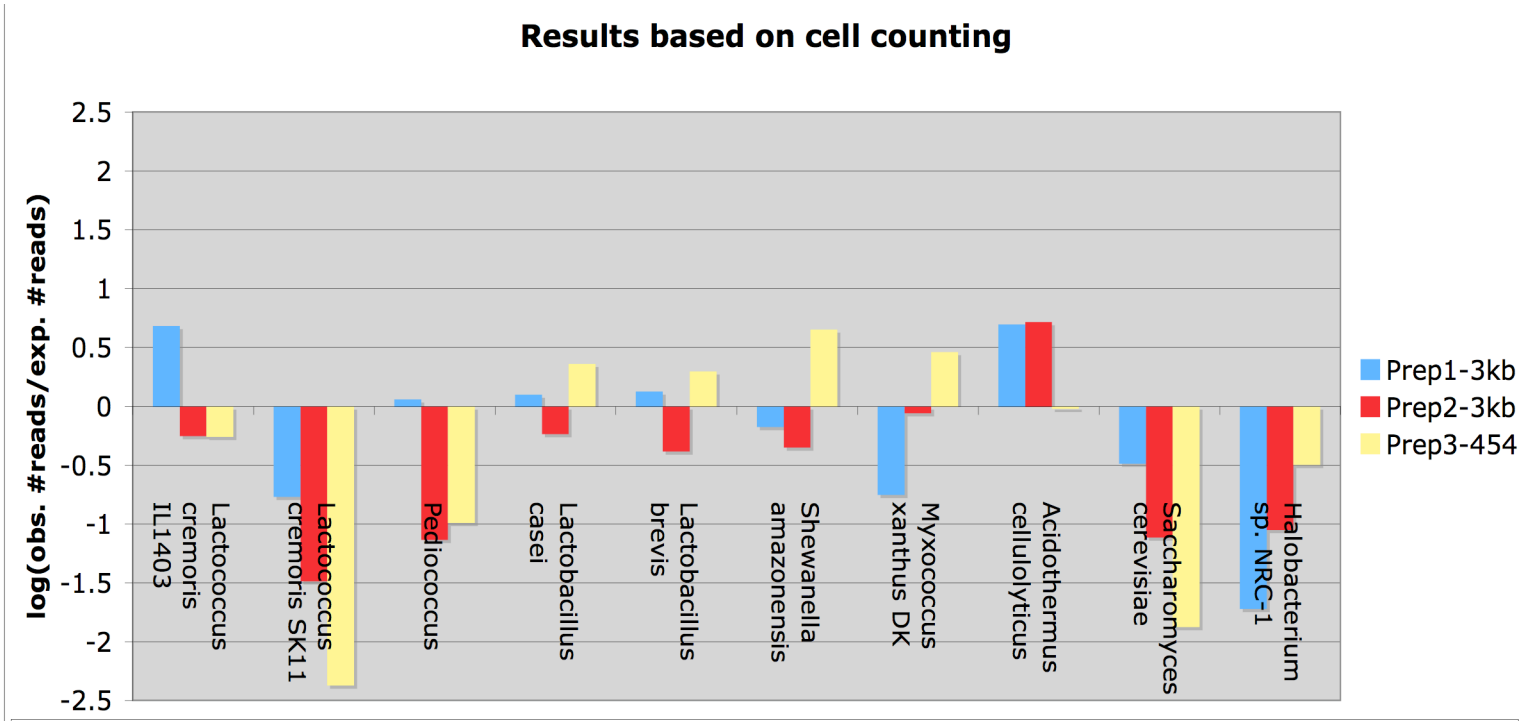
Alternative strategy for calculating expected proportions

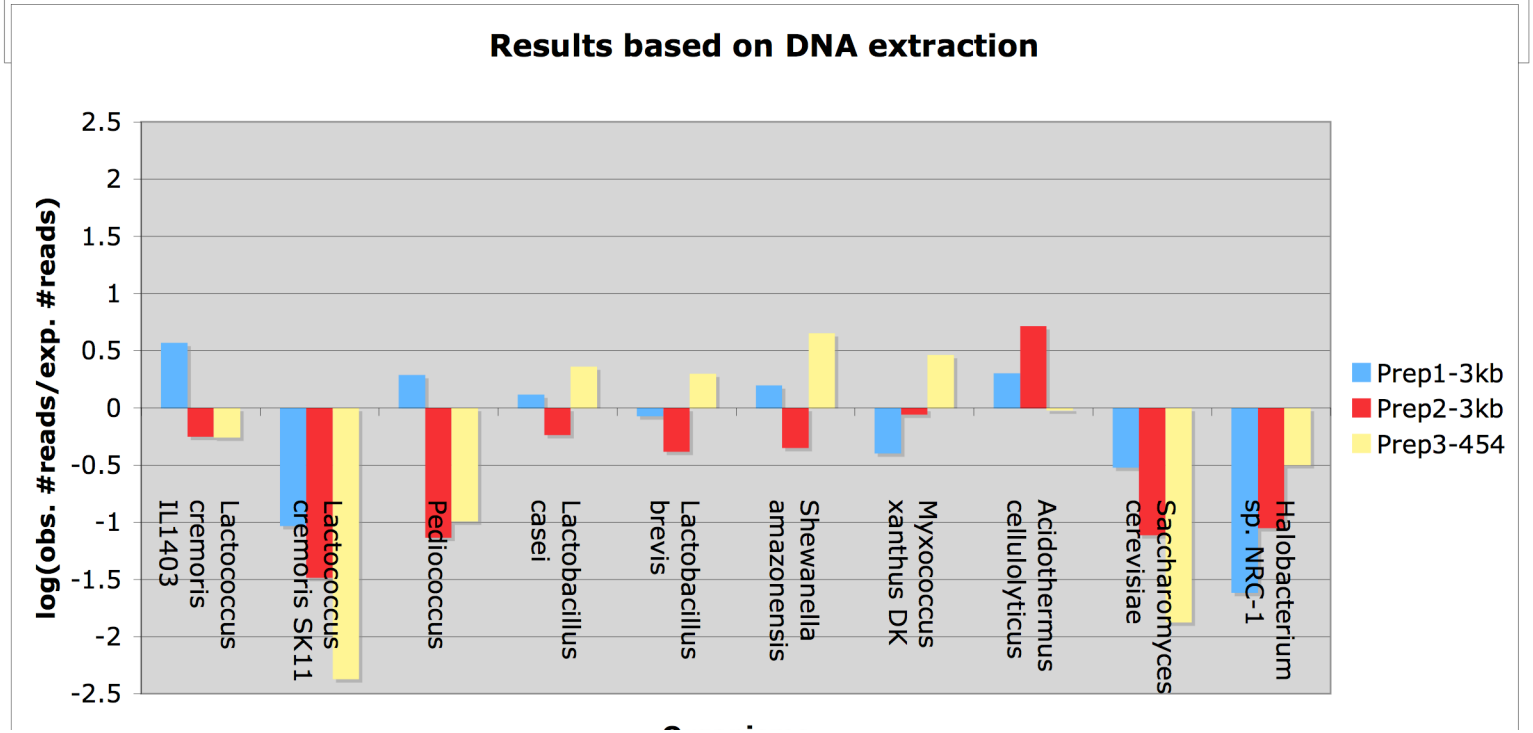


1. Extract DNA from pure cultures
2. Quantify DNA

3. Calculate expected proportions based on ng of DNA added to mixture


over/under-representation of organisms in my libraries
(alternative expectations)
log(obs. #reads/exp. #reads)
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
Lactococcus cremoris IL1403
Lactococcus cremoris SK11
Pediococcus
Lactobacillus casei
Lactobacillus brevis
Shewanella amazonensis
Myxococcus xanthus DK
Acidothermus cellulolyticus
Saccharomyces cerevisiae
Haldbacterium sp. NRC-1
Organisms
Prep1-3kb
Prep2-3kb
Prep3-454


Results based on cell counting
log(obs. #reads/exp. #reads)
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
Lactococcus cremoris IL1403
Lactococcus cremoris SK11
Pediococcus
Lactobacillus casei
Lactobacillus brevis
Shewanella amazonensis
Myxococcus xanthus DK
Acidothermus cellulolyticus
Saccharomyces cerevisiae
Halobacterium sp. NRC-1
Prep1-3kb
Prep2-3kb
Prep3-454


Results based on DNA extraction
log(obs. #reads/exp. #reads)
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
Lactococcus cremoris IL1403
Lactococcus cremoris SK11
Pediococcus
Lactobacillus casei
Lactobacillus brevis
Shewanella amazonensis
Myxococcus xanthus DK
Acidothermus cellulolyticus
Saccharomyces cerevisiae
Halobacterium sp. NRC-1
Prep1-3kb
Prep2-3kb
Prep3-454

Assumptions

- All species are equally well represented in the extracted DNA
  - non-biased sampling of microbial diversity
- All of the extracted DNA is equally represented in the sequenced library
  - non-biased cloning and sequencing of the metagenomic DNA

Assumptions

- All species are equally well represented in the extracted DNA
  - non-biased sampling of microbial diversity
- All of the extracted DNA is equally represented in the sequenced library
  - non-biased cloning and sequencing of the metagenomic DNA

Test for over/under-representation of functional categories

- Map reads to sequenced genomes
- Annotate reads based on which genes they overlap
- Expected distribution of functional categories = actual distribution in the genomes

Future

- Statistics to compare expected vs. observed for any given organism
- Investigate 599 “cellular organism” assignments for Prep #2
- How to do functional category analysis (or, is this even interesting?)
- Anything else?