

Introduction to Metagenomics

Gene Tyson
DeLong Lab (MIT)
gtysen@mit.edu

**Agouron Summer
Course 2008**

Introduction to Metagenomics

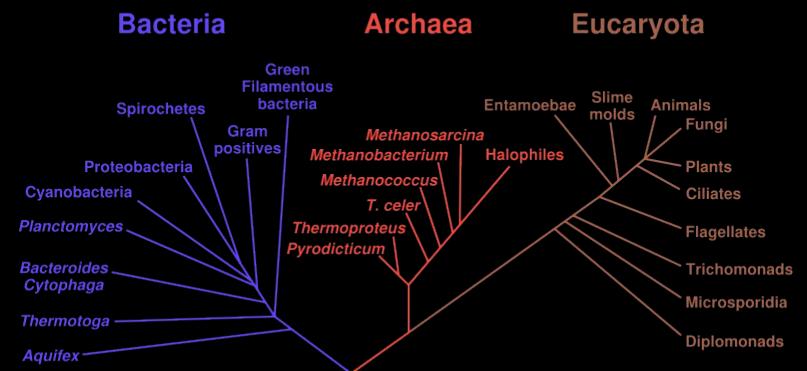
Outline:

- ❑ Putting metagenomics in perspective
- ❑ What can metagenomics tell us about microbial communities?
- ❑ Future of metagenomics
- ❑ Tutorial - Bioinformatic tools for metagenomic data analysis

Carl Woese

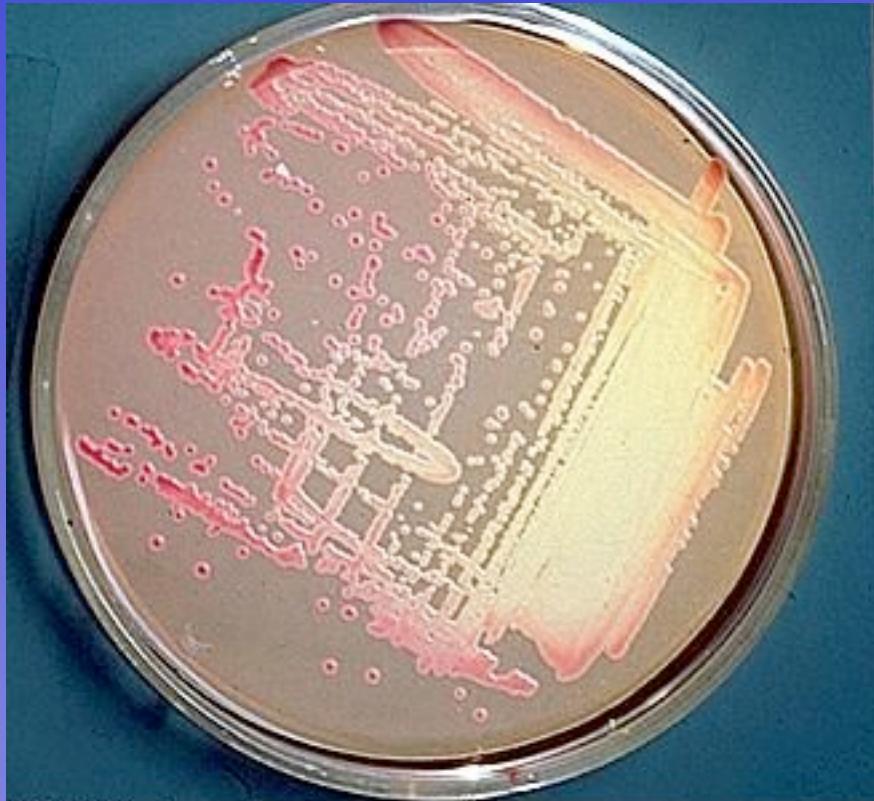
Woese C, Fox G (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms.
Proc Natl Acad Sci U S A 74 (11): 5088–90.

Woese C, Kandler O, Wheelis M (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya.
Proc Natl Acad Sci U S A 87 (12): 4576–9.

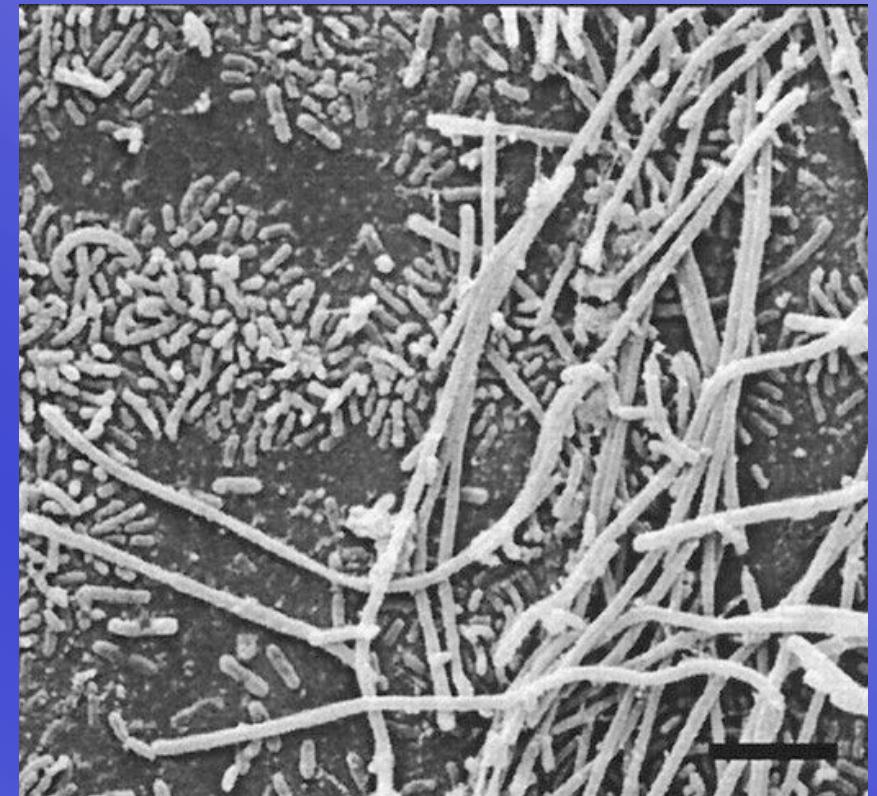


http://www.kva.se/KVA_Root/swe/_news/detail.asp?NewsId=291&br=ie&ver=4up

tame



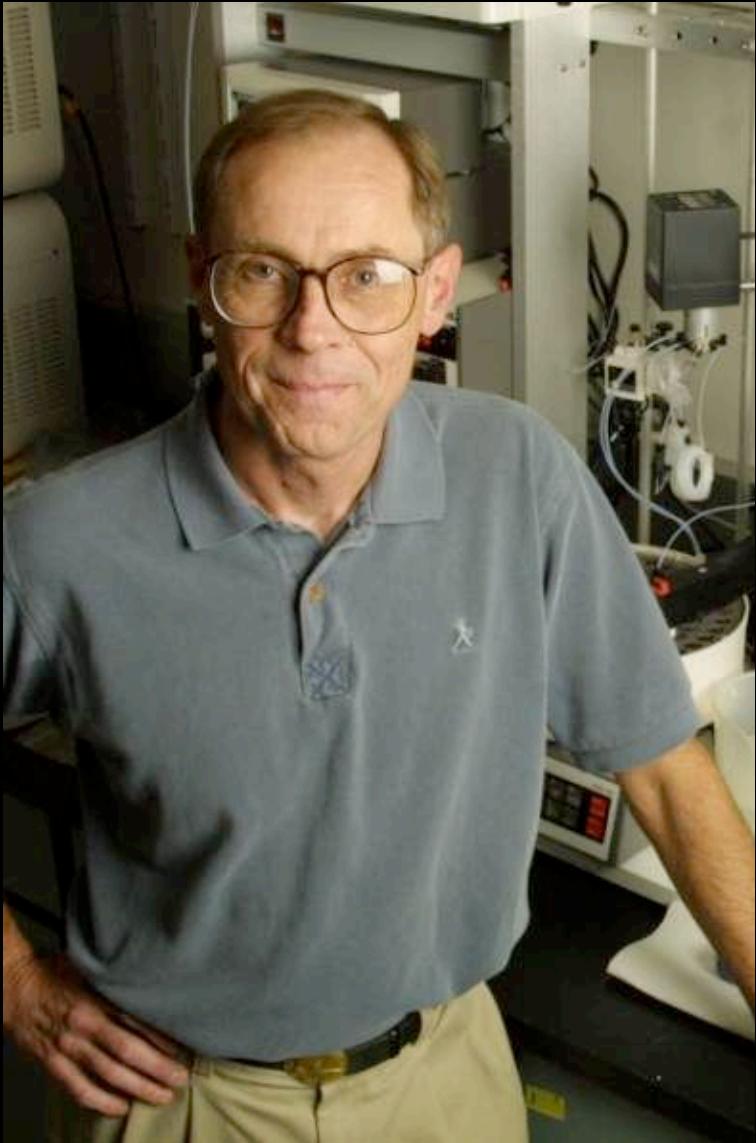
wild

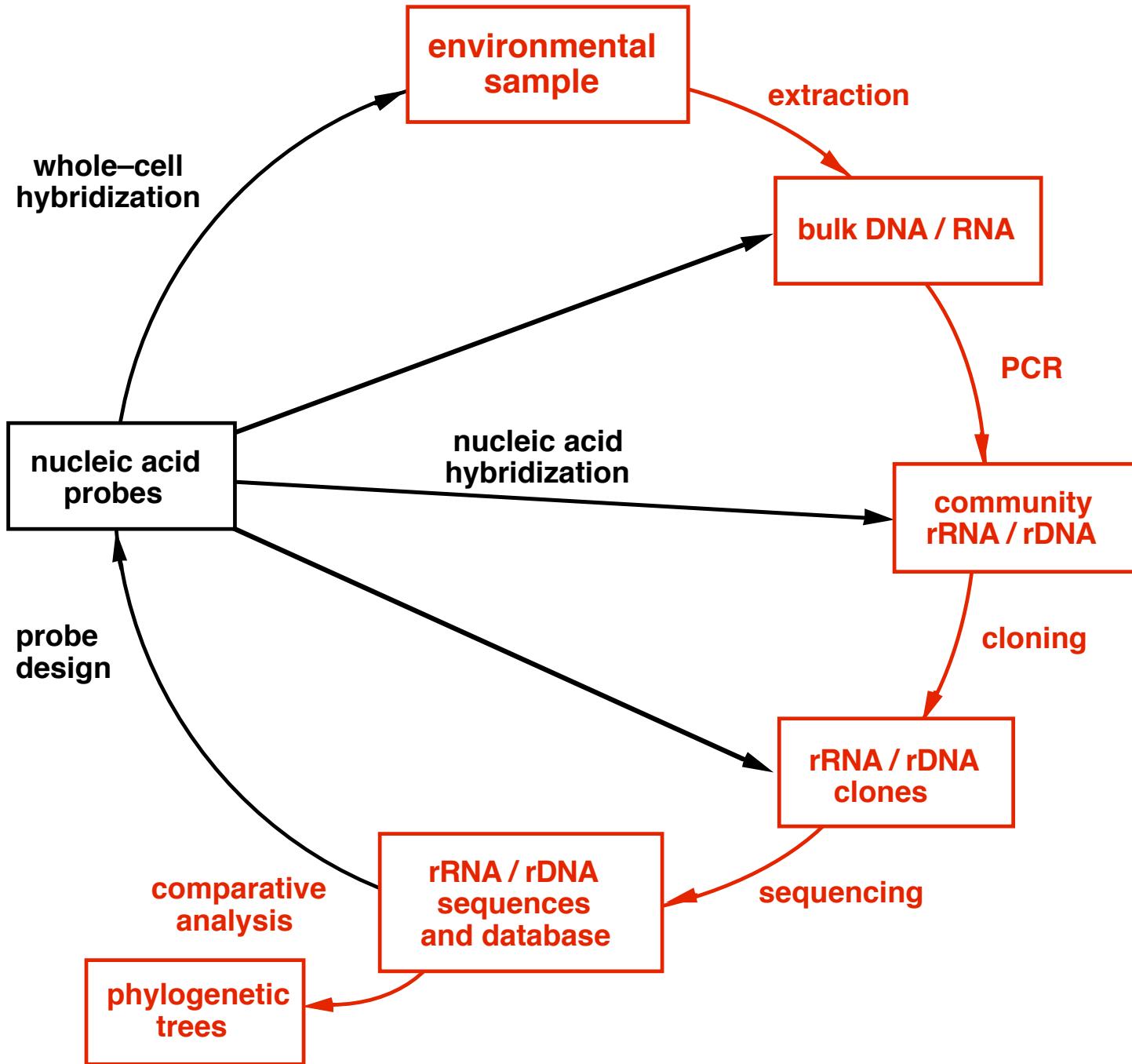


≠

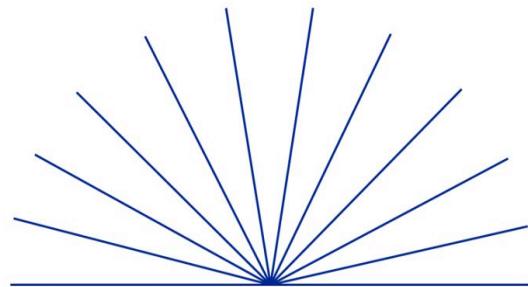
most bacteria don't grow on plates
“the great plate count anomaly”

Norm Pace (University of Colorado-Boulder)

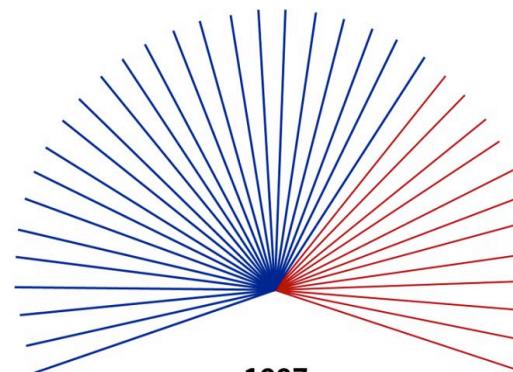




Known Bacterial Phylogenetic Divisions

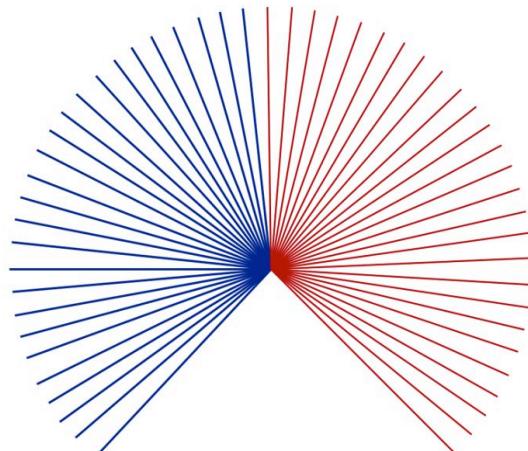


1987
(12 divisions; 12 cultured / 0 uncultured)

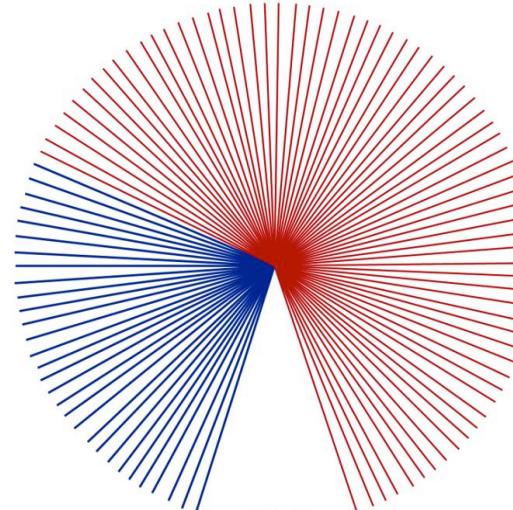


1997
(36 divisions; 24 cultured / 12 uncultured)

— Cultured
— Uncultured



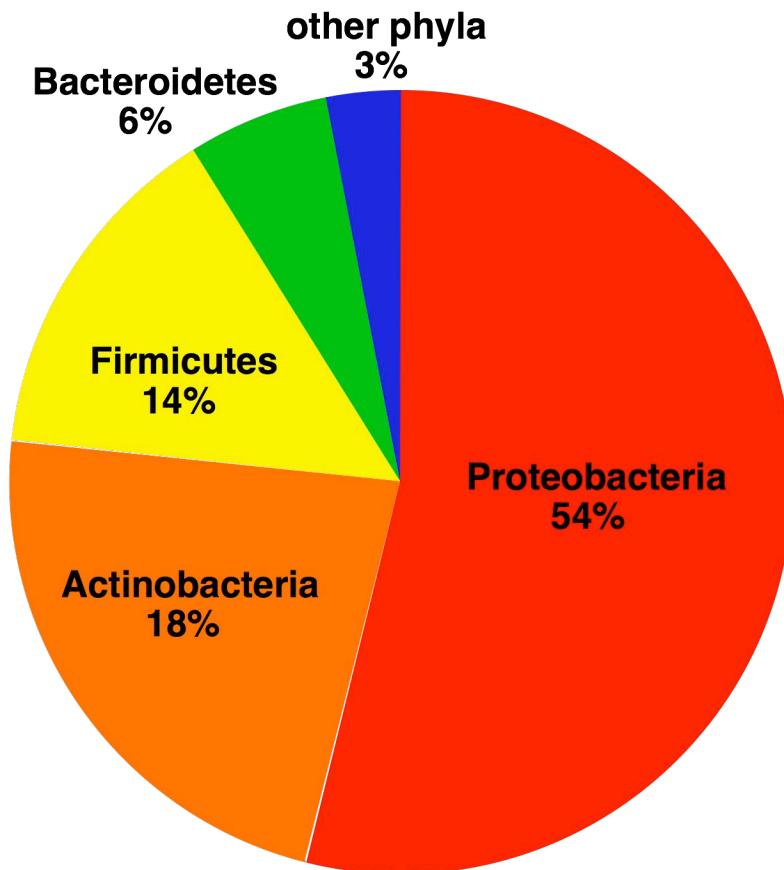
2003
(53 divisions; 26 cultured / 27 uncultured)



2006
(~100 divisions; 30 cultured / ~70 uncultured)

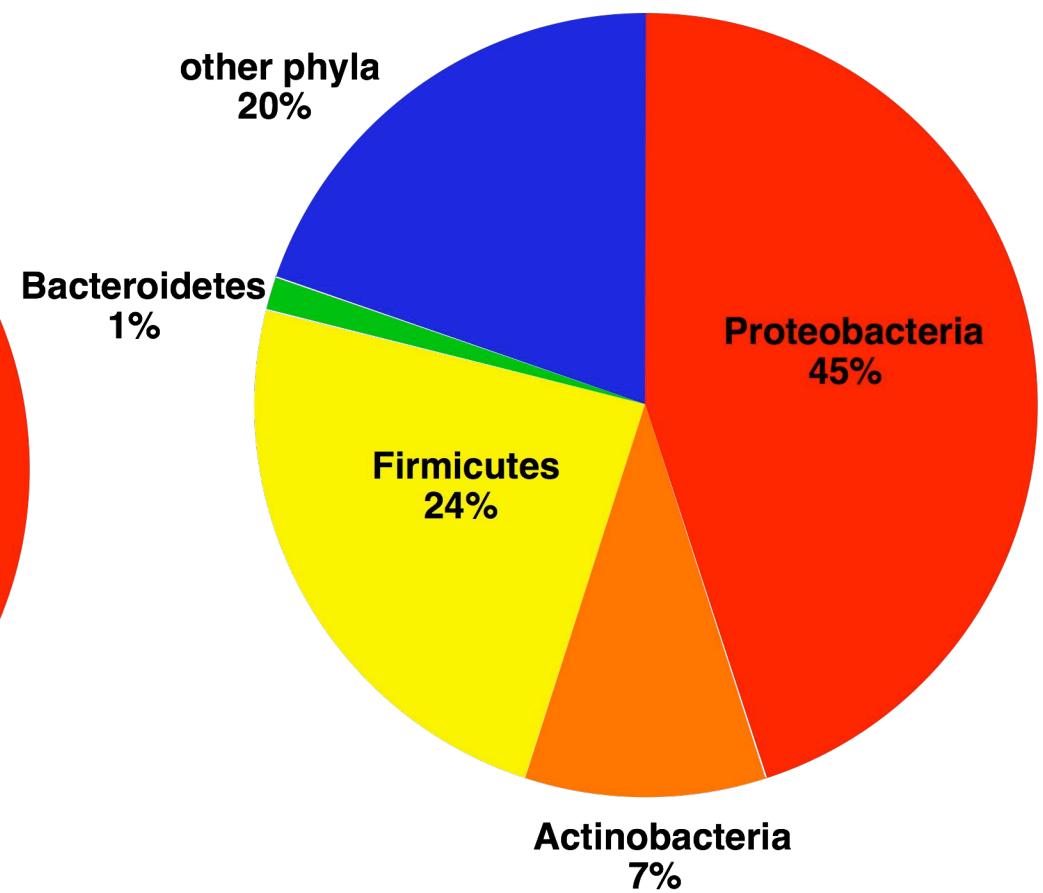
very skewed representation of Bacteria

culture collection (ACM)

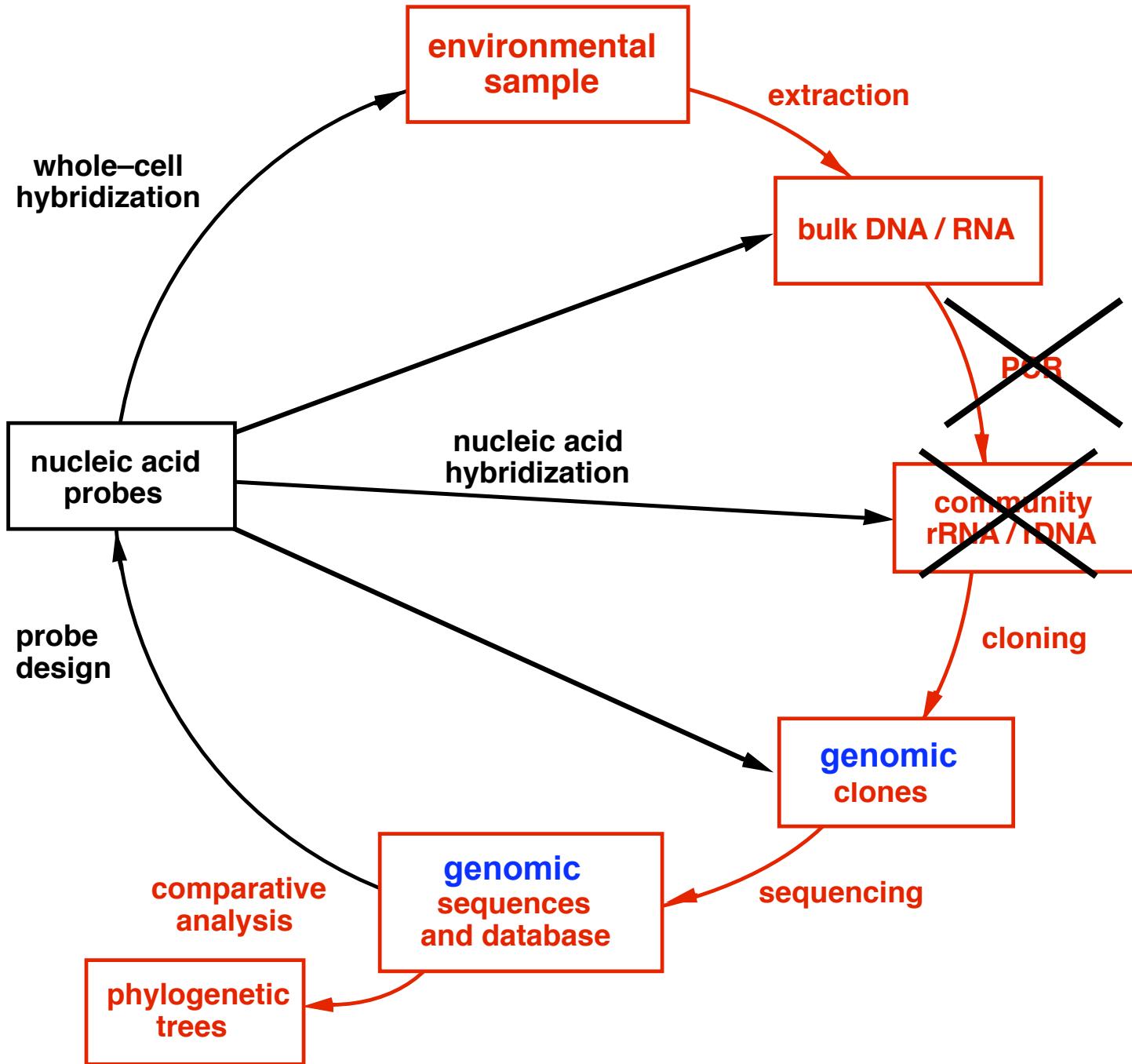


3760 bacterial cultures

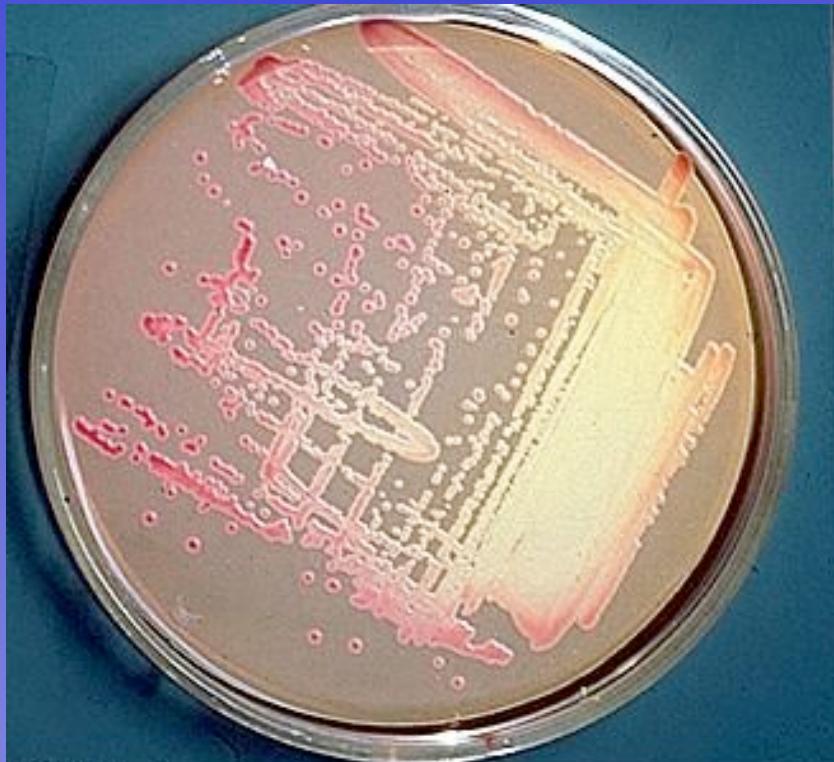
sequenced genomes



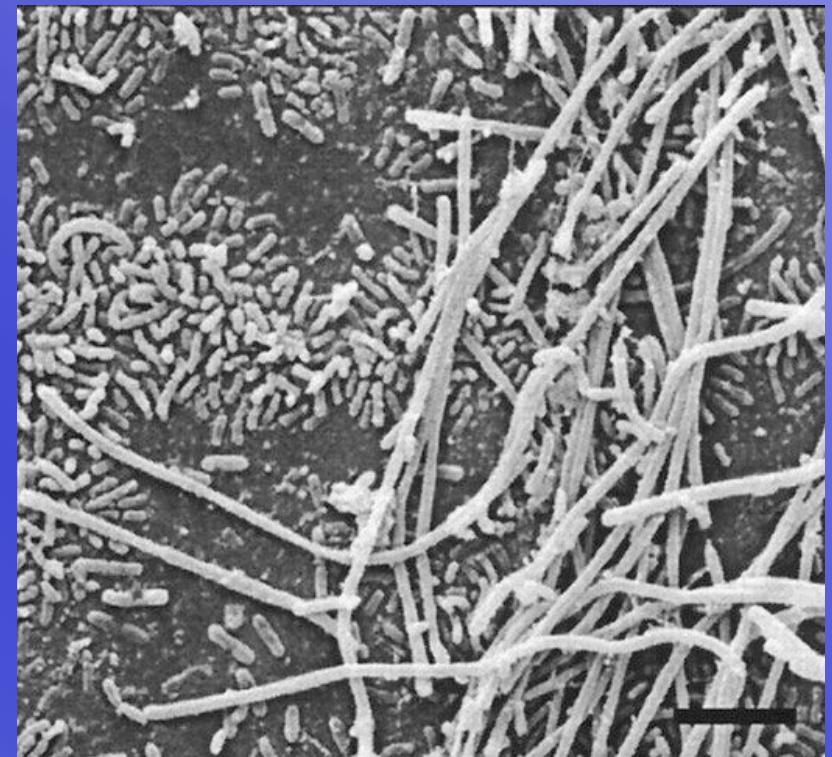
71 bacterial genomes



isolate



community

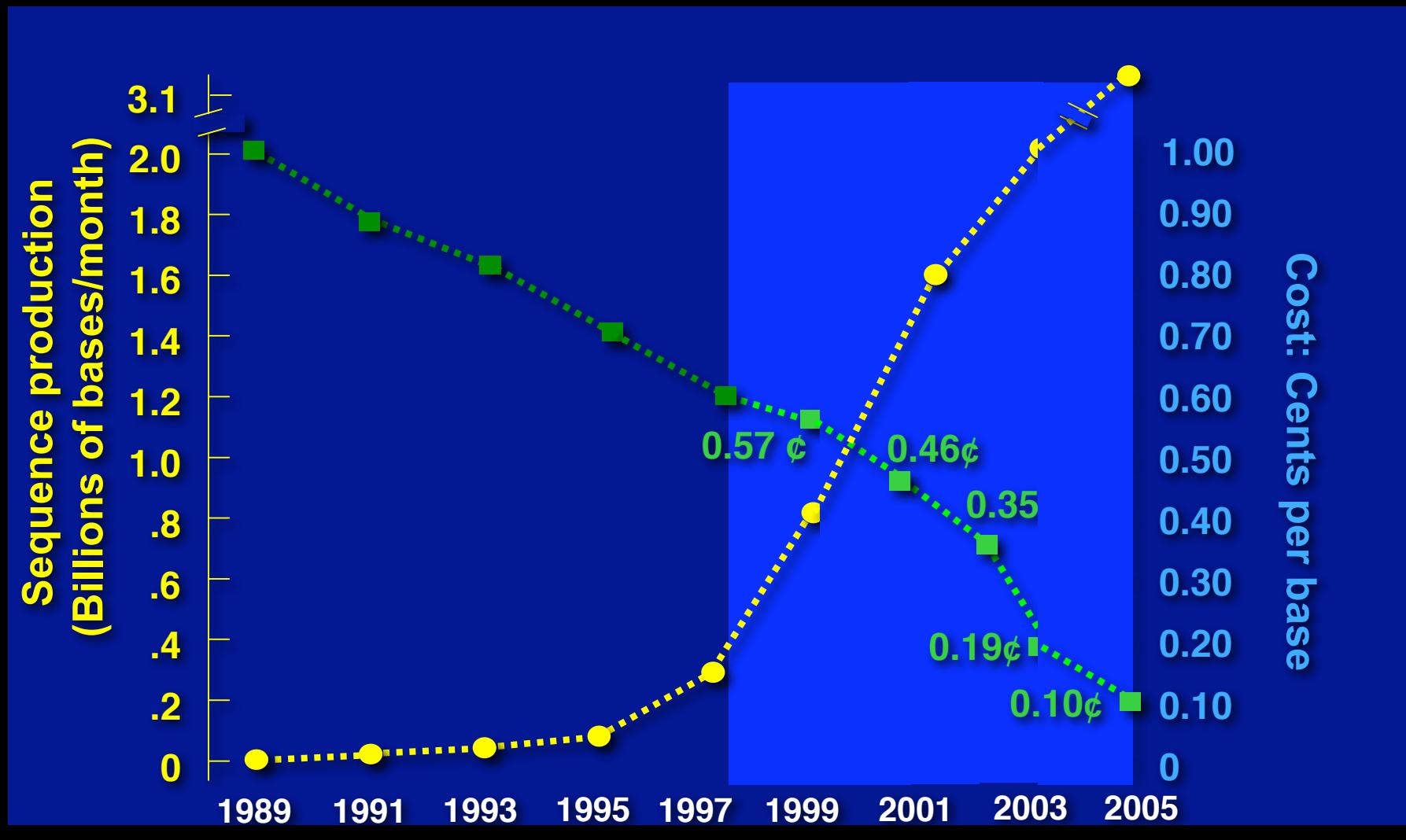


sequencing

Genomics

Metagenomics

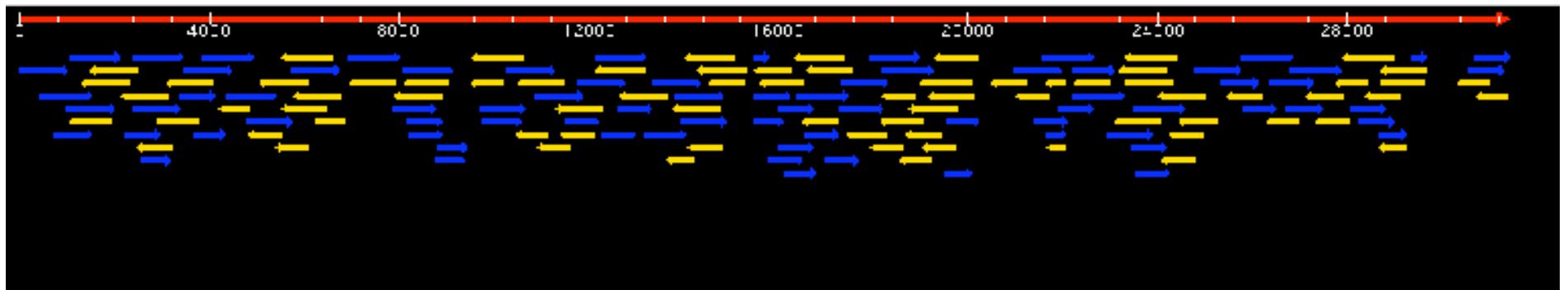
Sanger sequencing has become much cheaper Joint Genome Institute (JGI) statistics



Courtesy of Phil Hugenholtz (JGI)

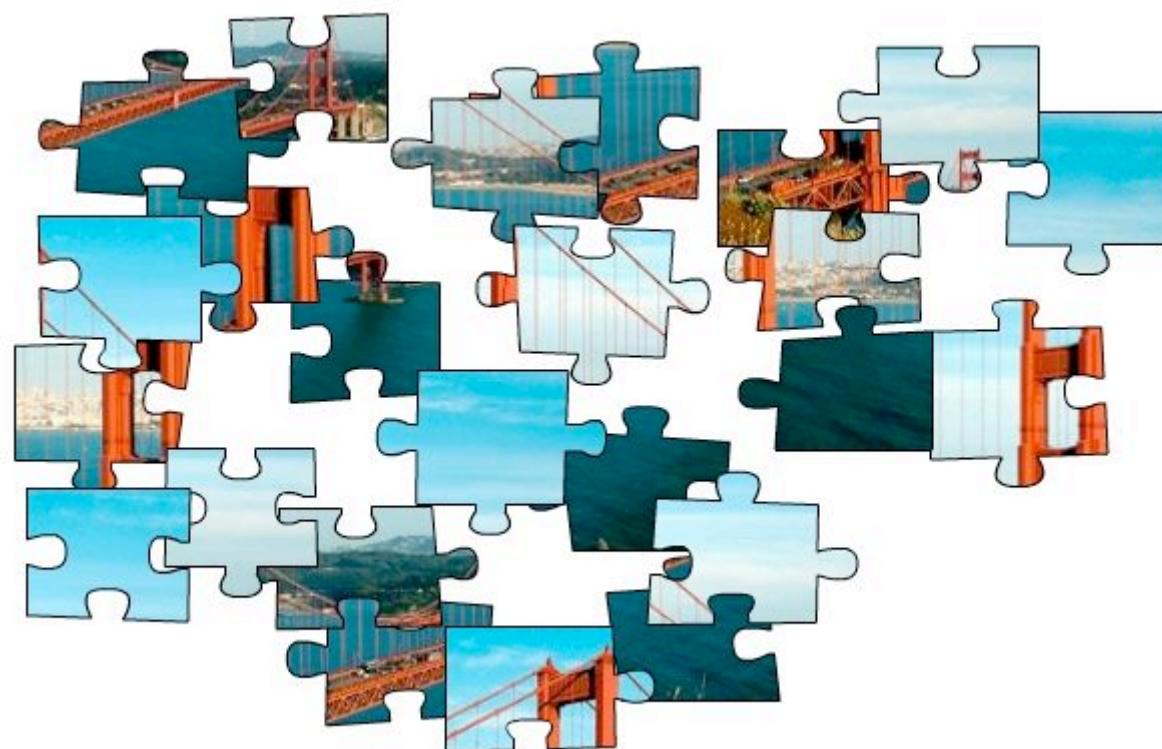
The catch.... it comes in small pieces

Average Sanger read length - 750 bases (bps)



**Must assemble the reads together
Giant jigsaw puzzles**

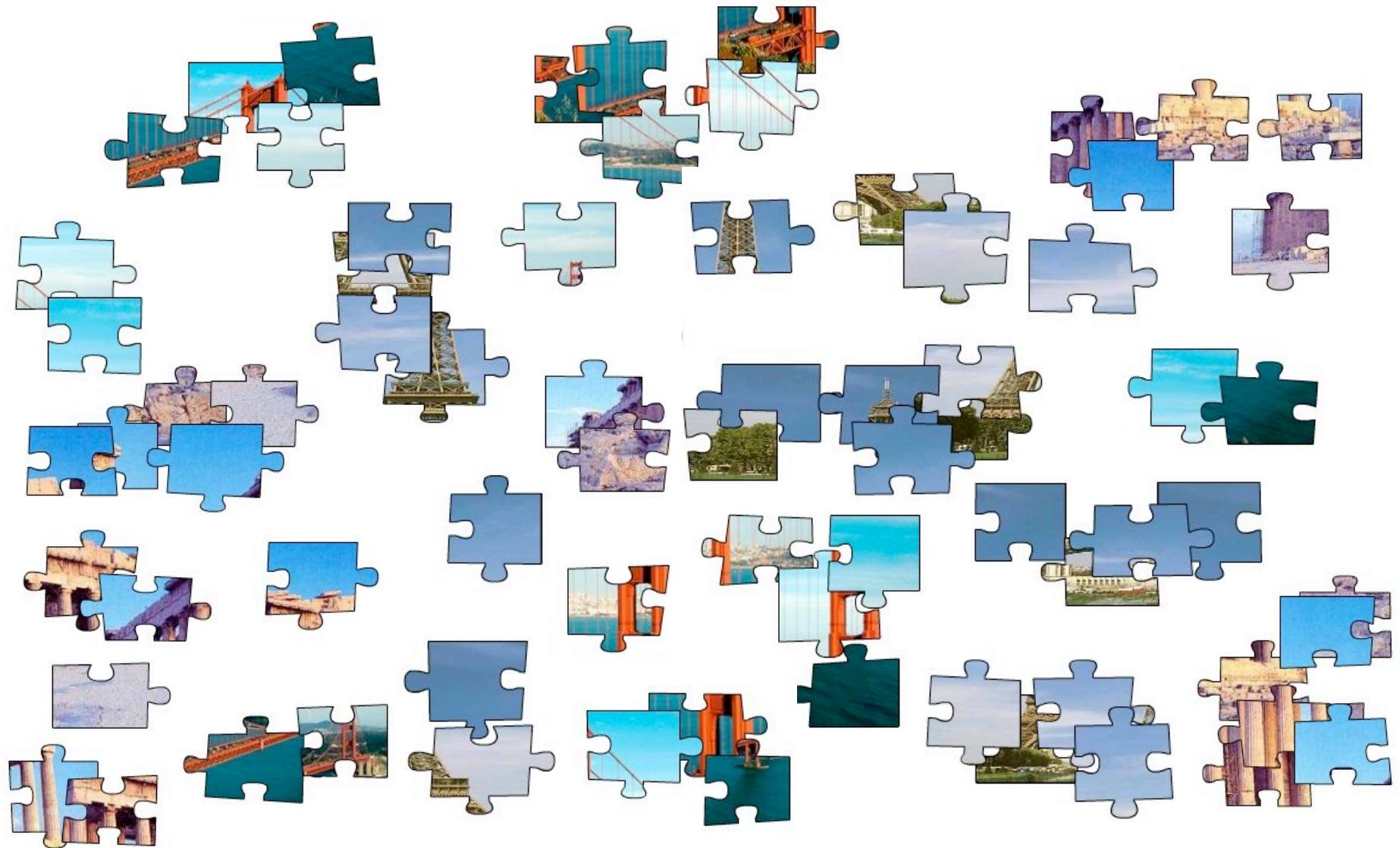
Genome assembly



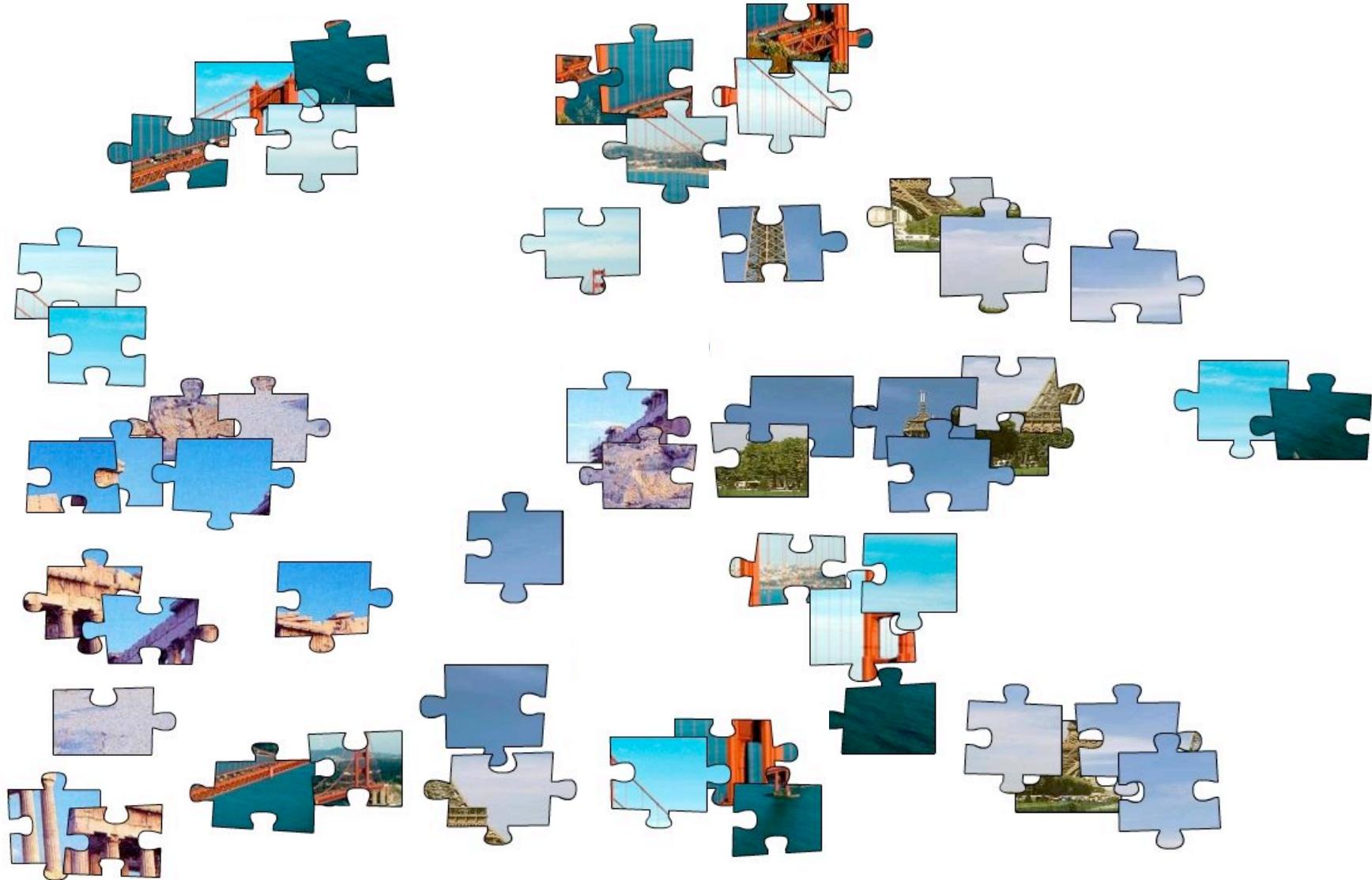
Genome assembly



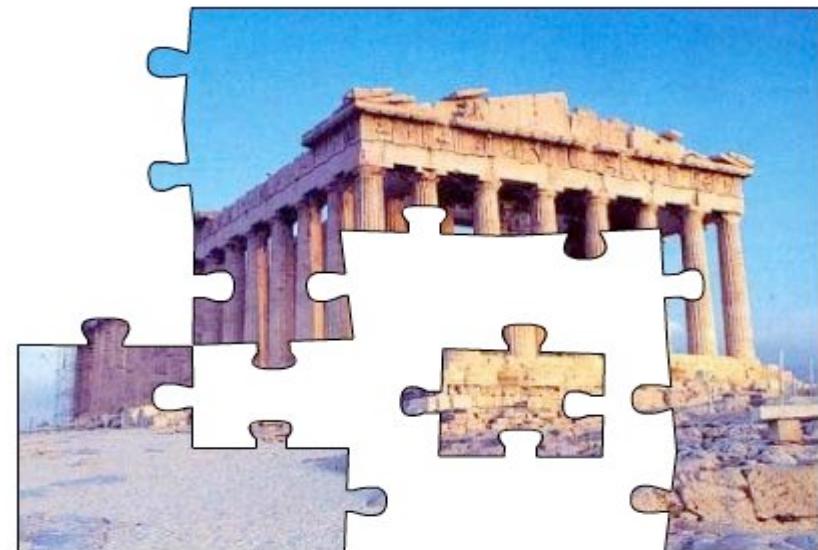
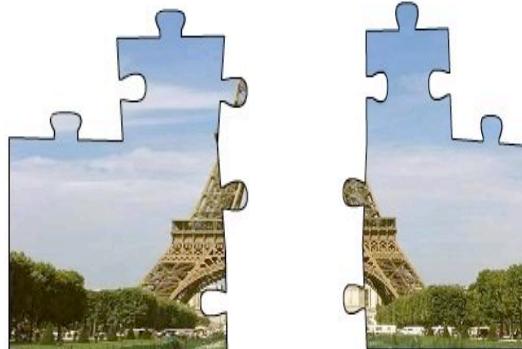
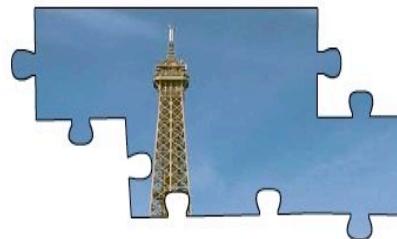
Metagenome assembly



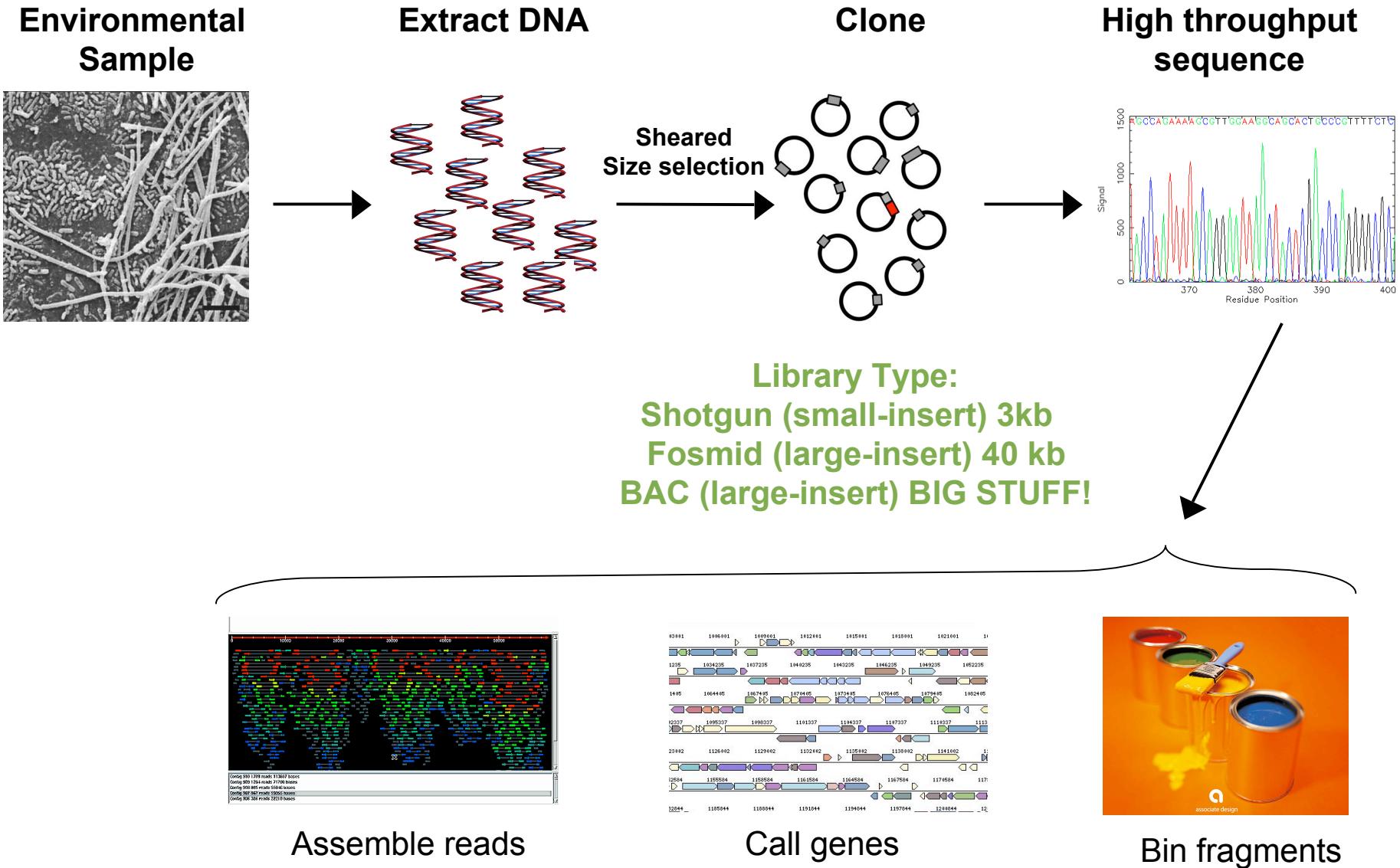
Metagenome assembly



Metagenome assembly



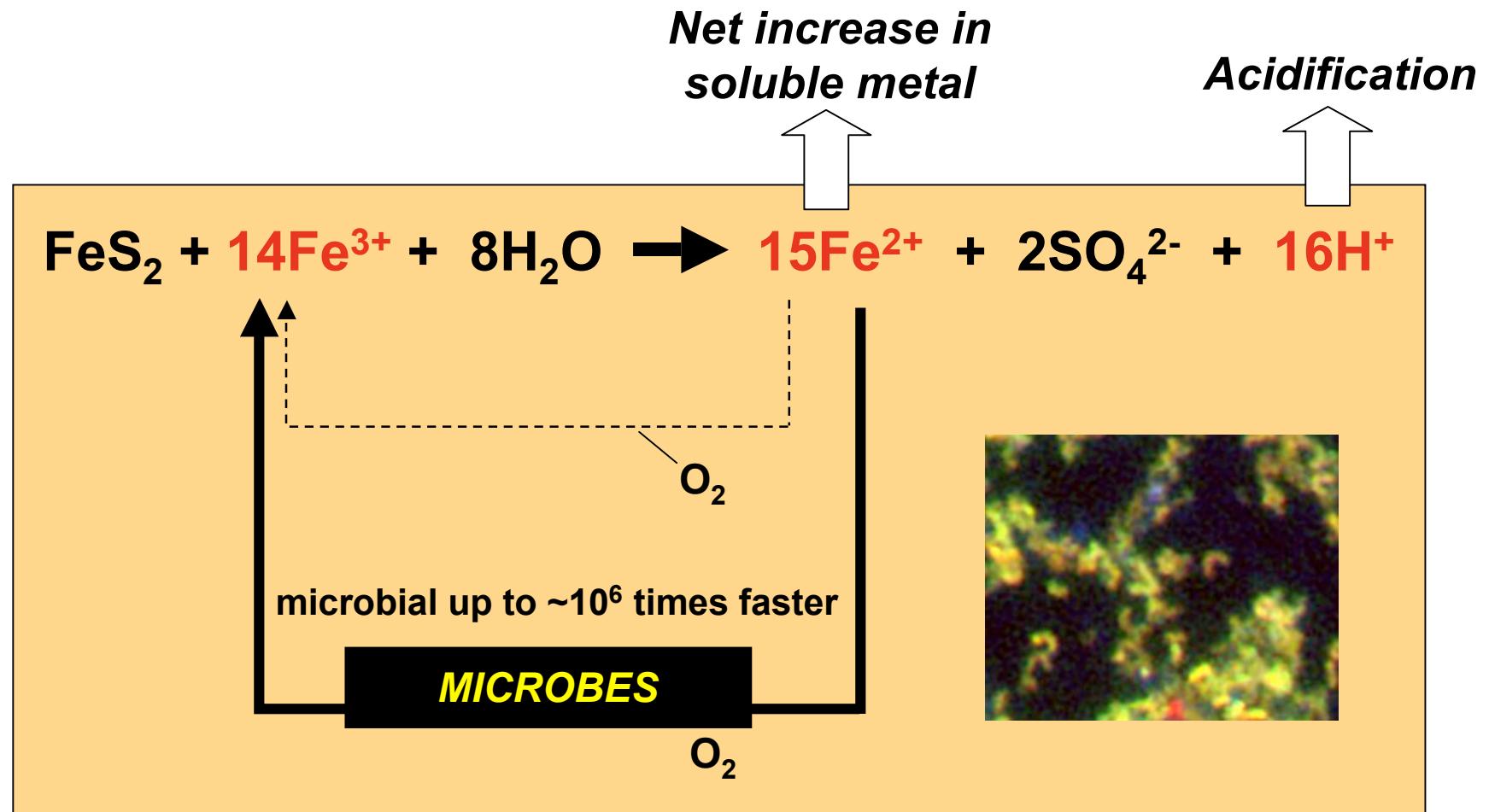
Decoding metagenomes



Metagenomics projects to date:

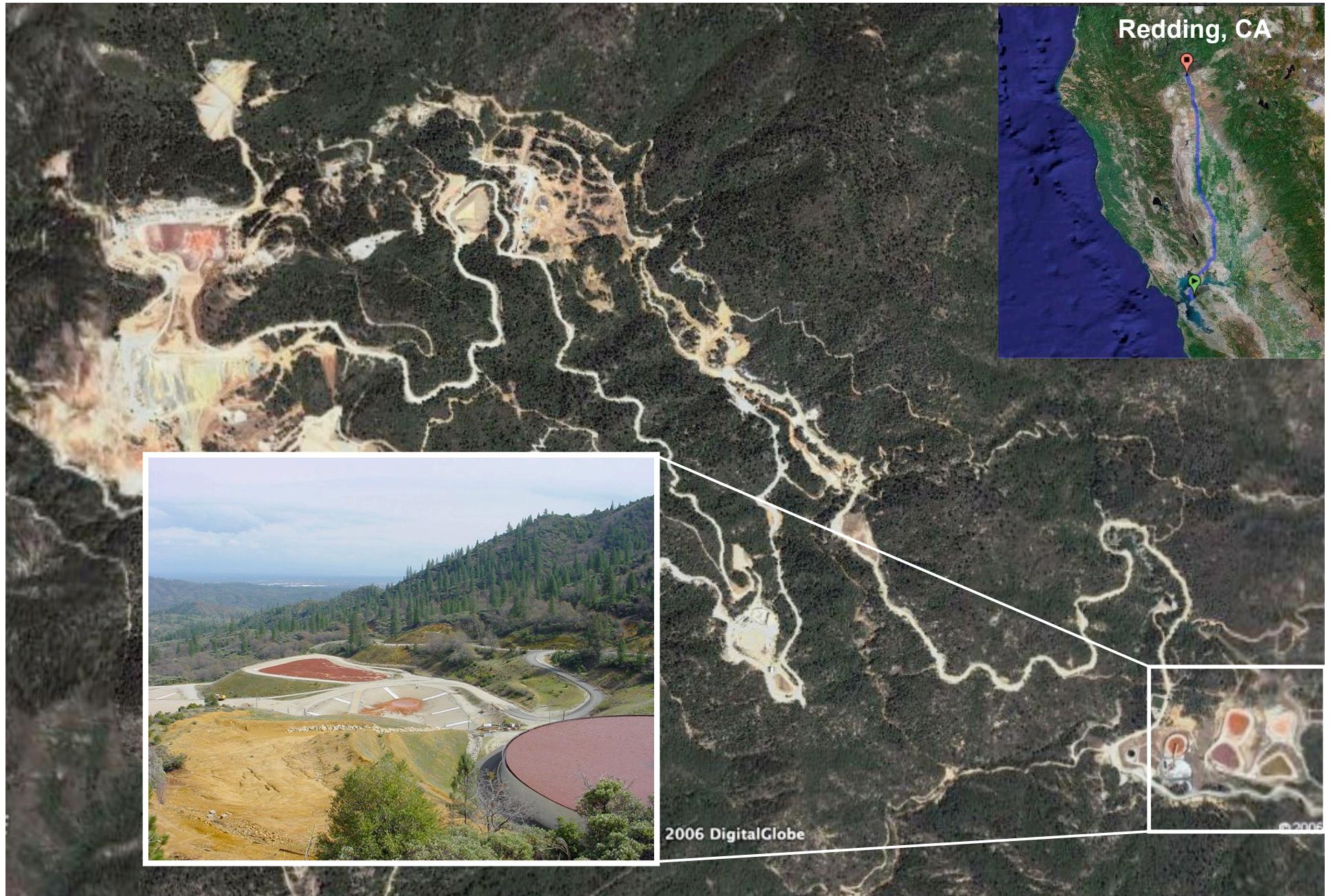
- **Metabolic profiling of environments without need for significant assembly or reference to the organism from which genes were derived (e.g. EGTs).**
- **Comprehensive analysis of the community that resolves gene complement of the dominant organisms and provides insight into population structure and evolution.**

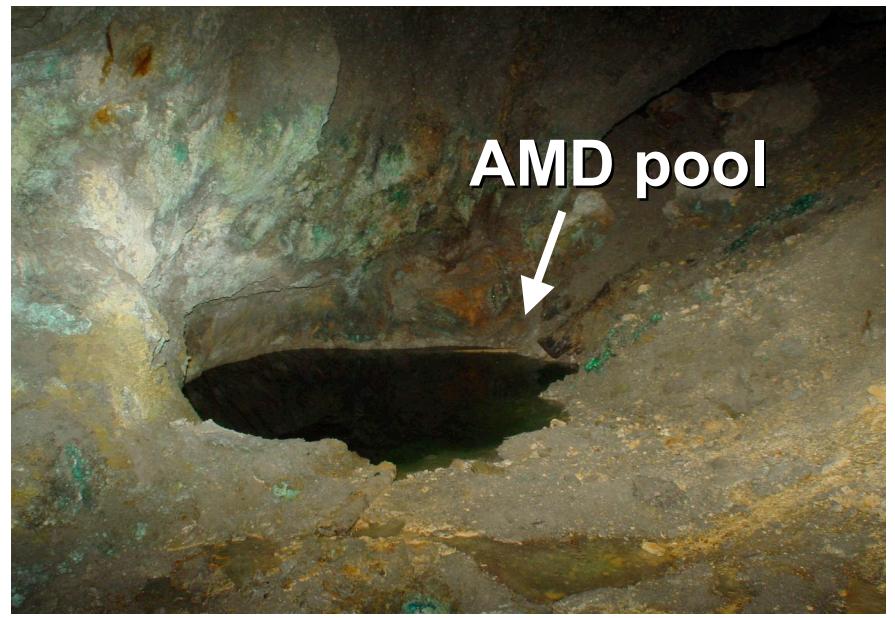
AMD: Coupling of microbial metabolism and mineral dissolution





Richmond Mine, Iron Mountain, CA





Richmond Mine Iron Mountain CA

95% pyrite (FeS_2) ore deposit

Extreme acidity ($\text{pH} < 1$)

Warm (30 - 50°C)

Very high ionic strength (g/l)

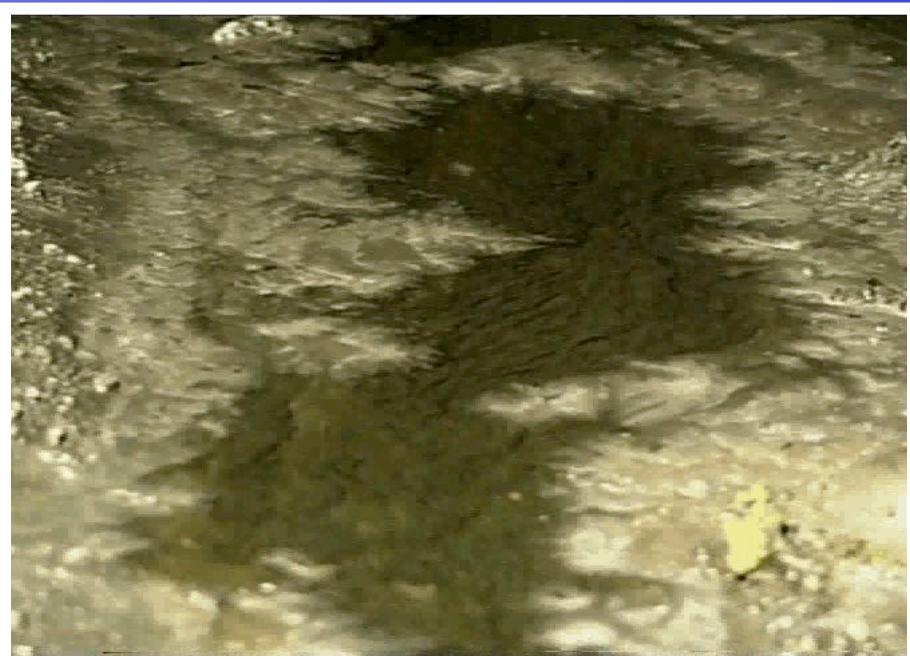
Abundant toxic metals

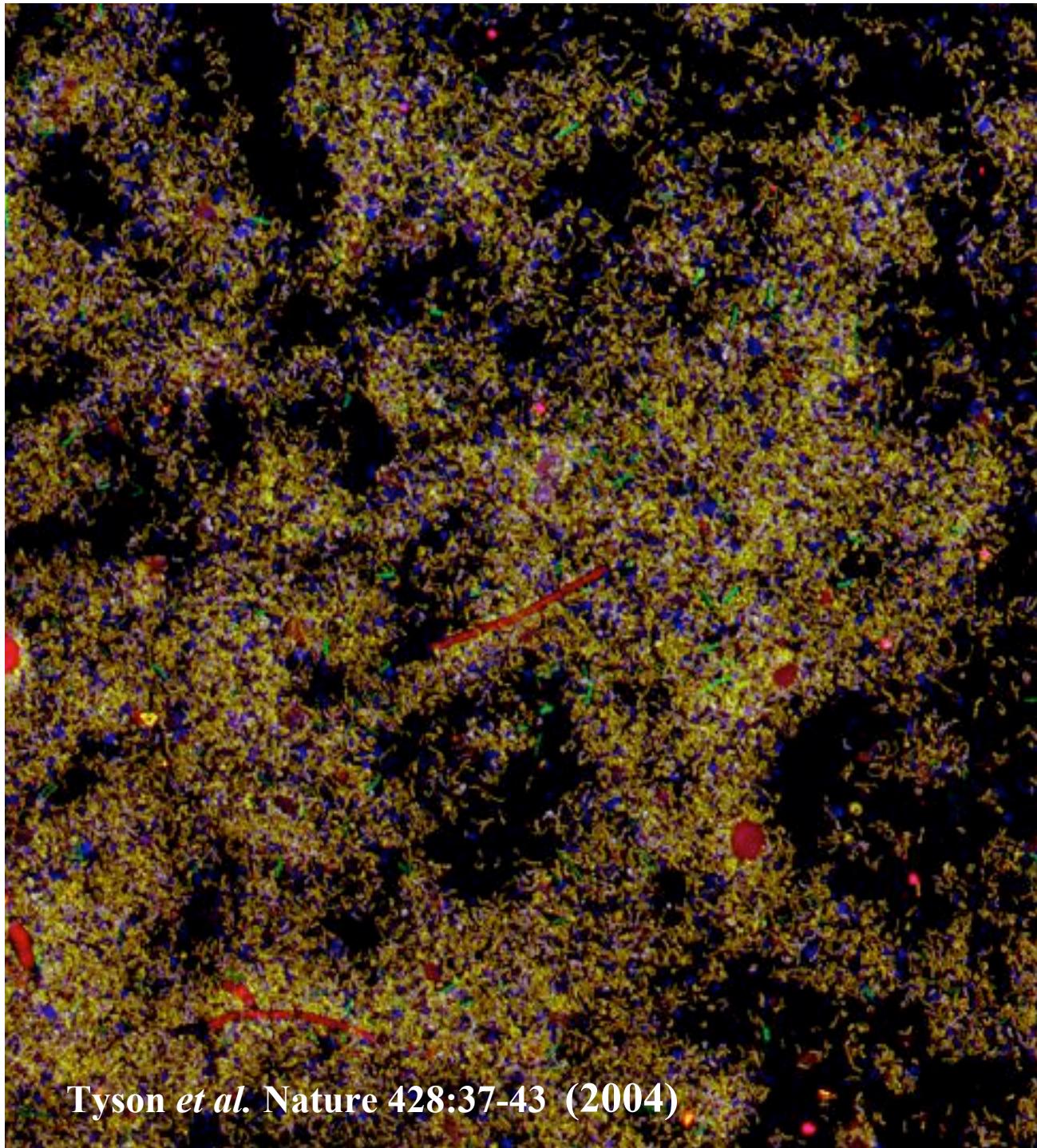
(molar Fe, mM Zn, Cu, As)

Iron oxidation- primary energy
source

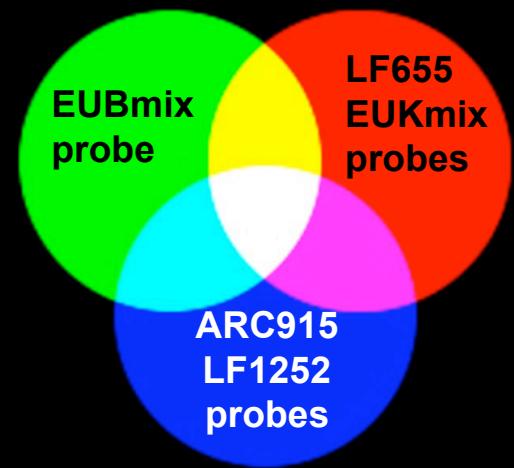
No sunlight

Limited external sources
of organic C and fixed N

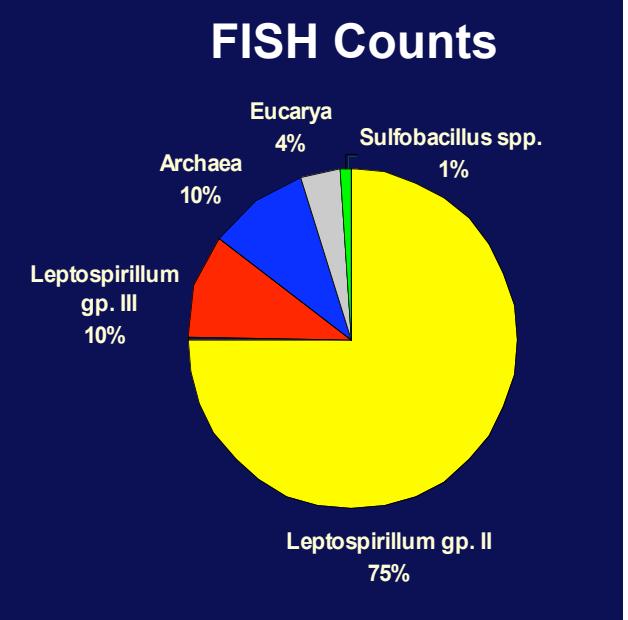




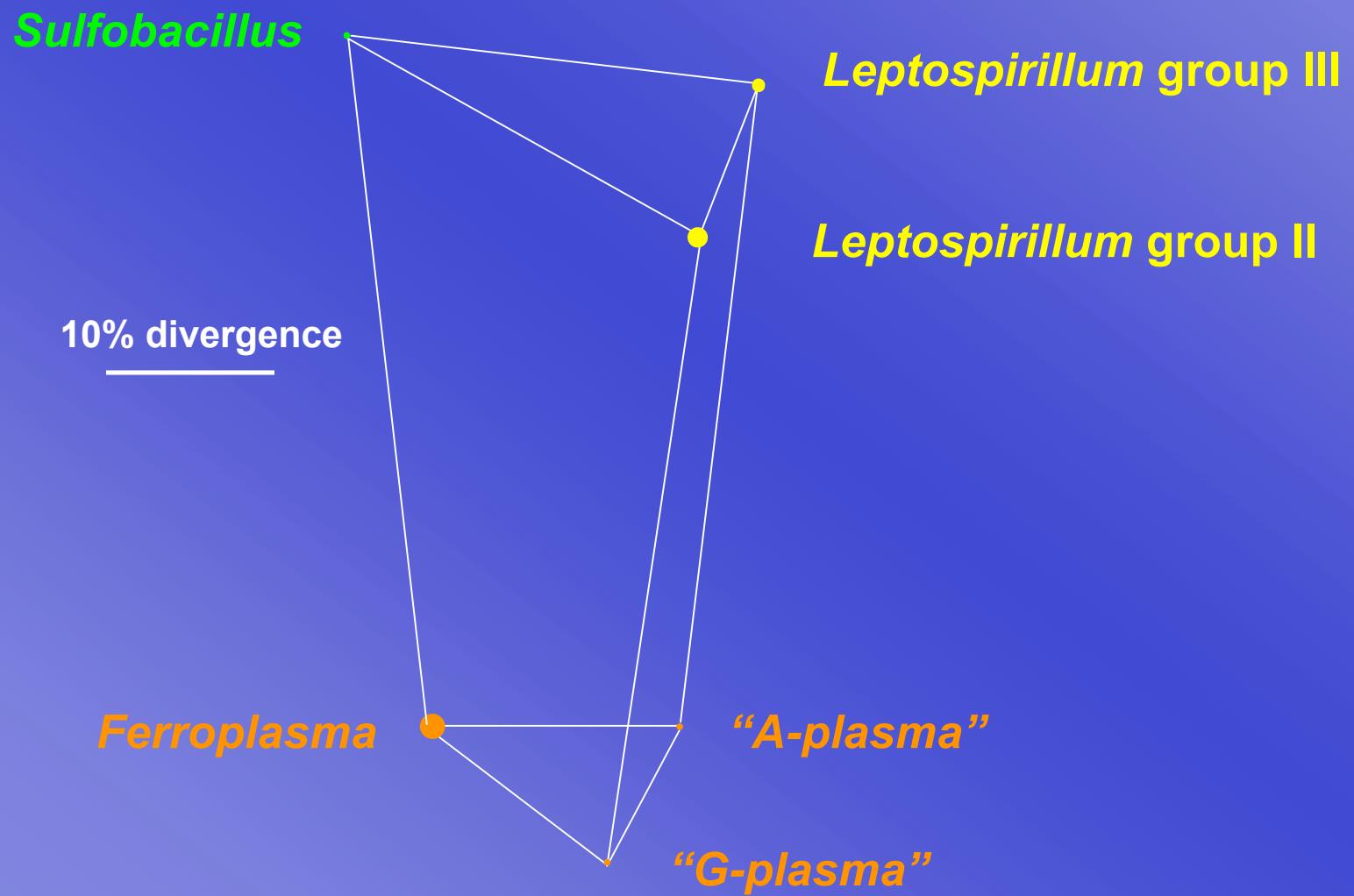
Tyson *et al.* Nature 428:37-43 (2004)



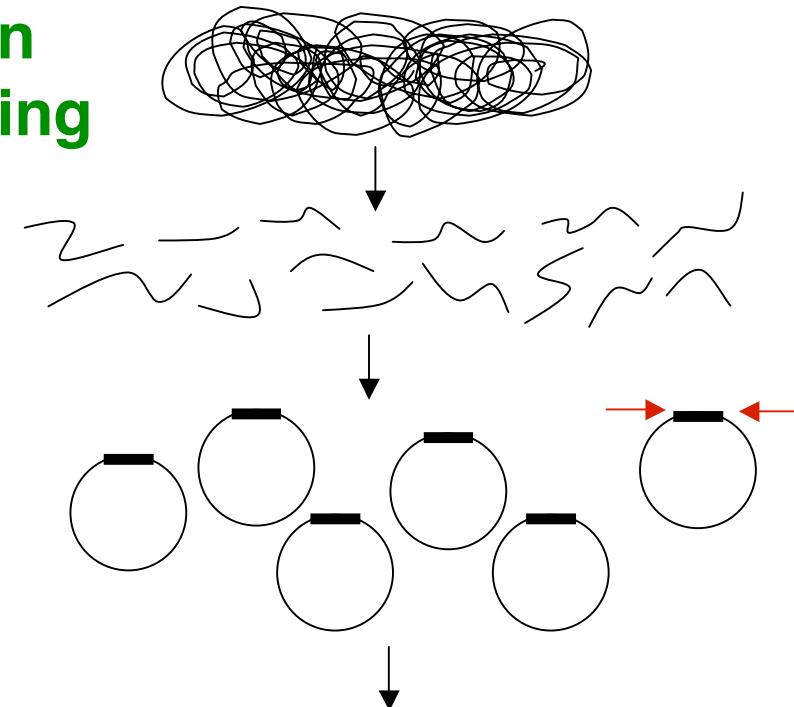
- Archaea
- *Leptospirillum* gp II
- *Leptospirillum* gp III
- *Sulfobacillus*
- Eukarya



16S rRNA gene PCR clone library analysis



Shotgun sequencing

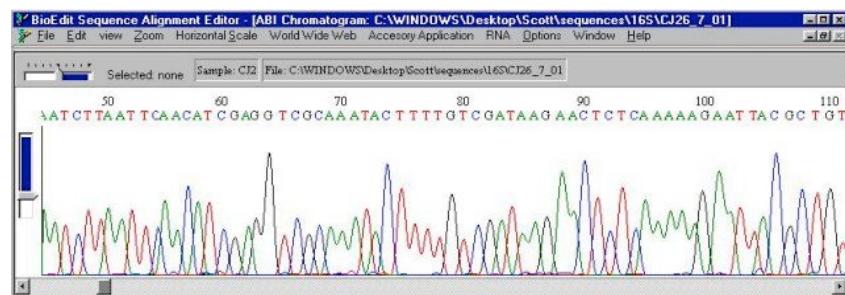


Bulk environmental DNA

sheared

3 – 4 kb shotgun library

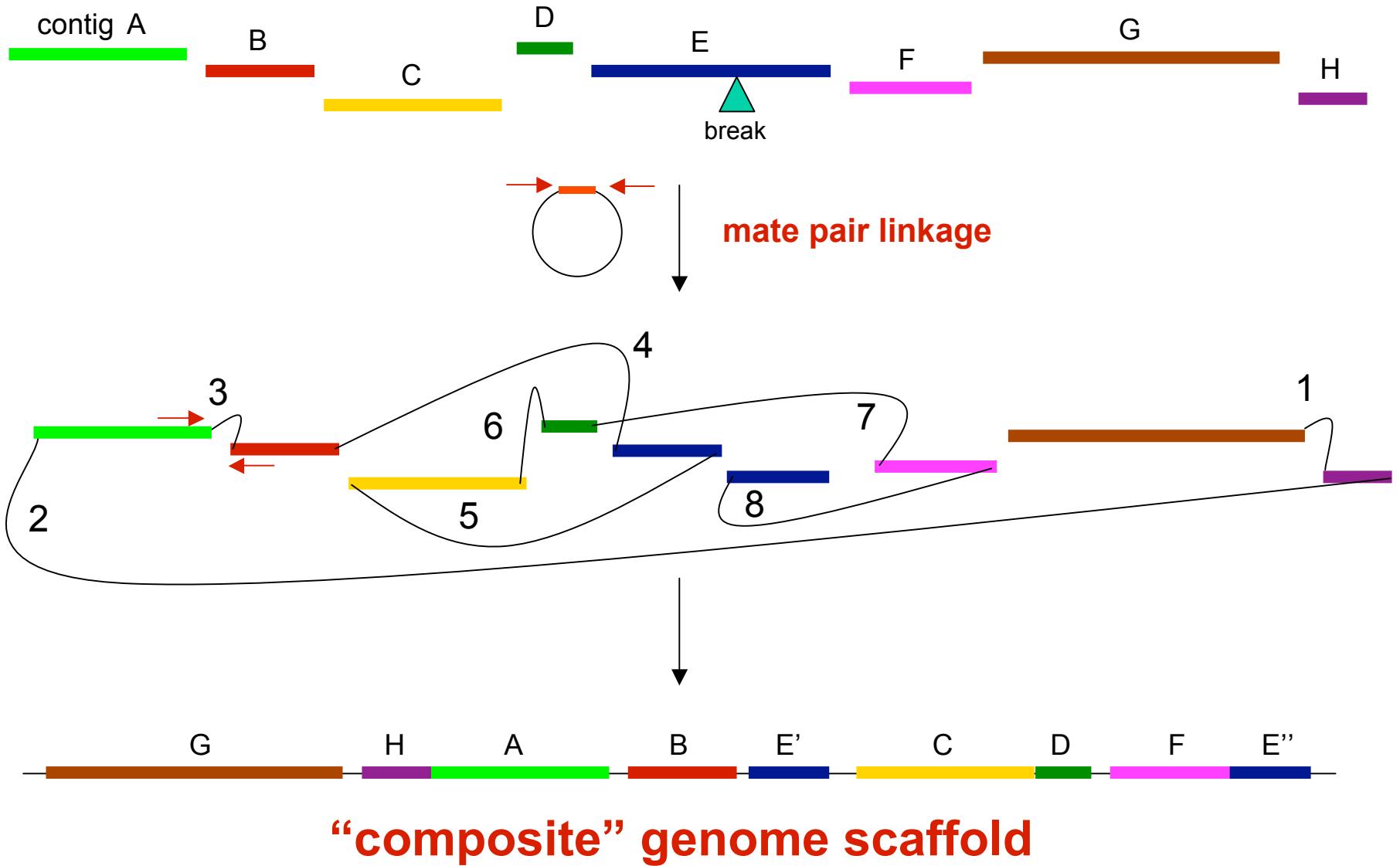
end sequence clones (f / r)



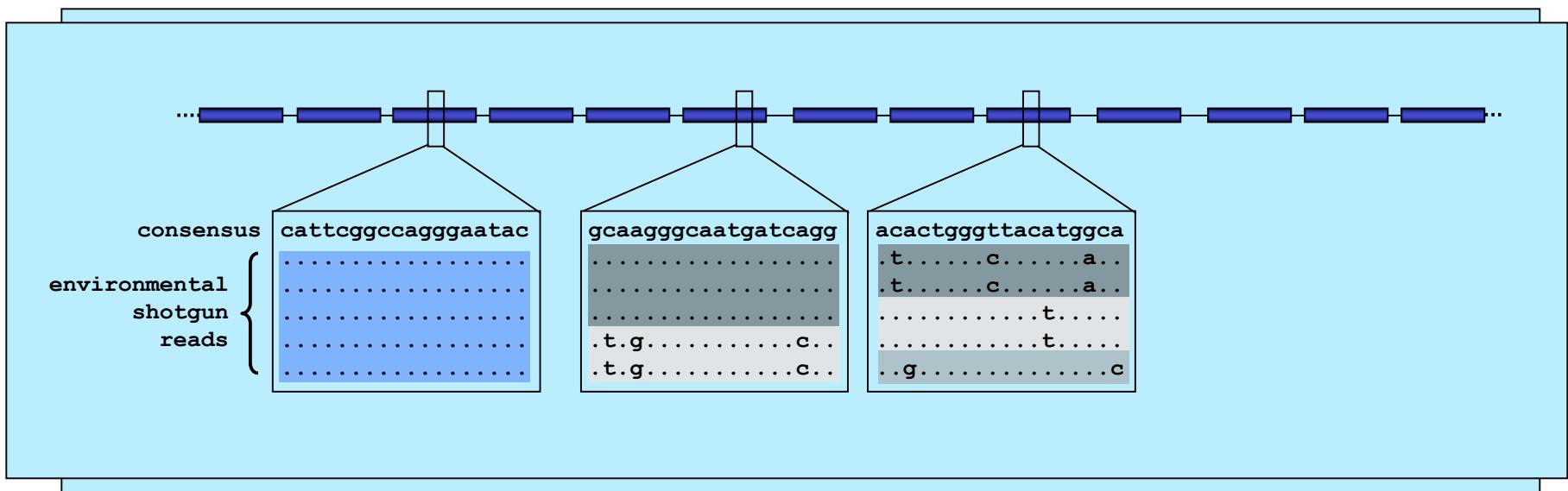
...ACGGCTGCGTTACATCGATCAT
ACATCGATCATTACGATACCATTG...

assemble reads by alignment identity

Genome Scaffolding



Assembling community genomic (‘metagenomic’) data



Other potential problems:

- repetitive and mobile elements
 - highly conserved genes

AMD community genome sequencing

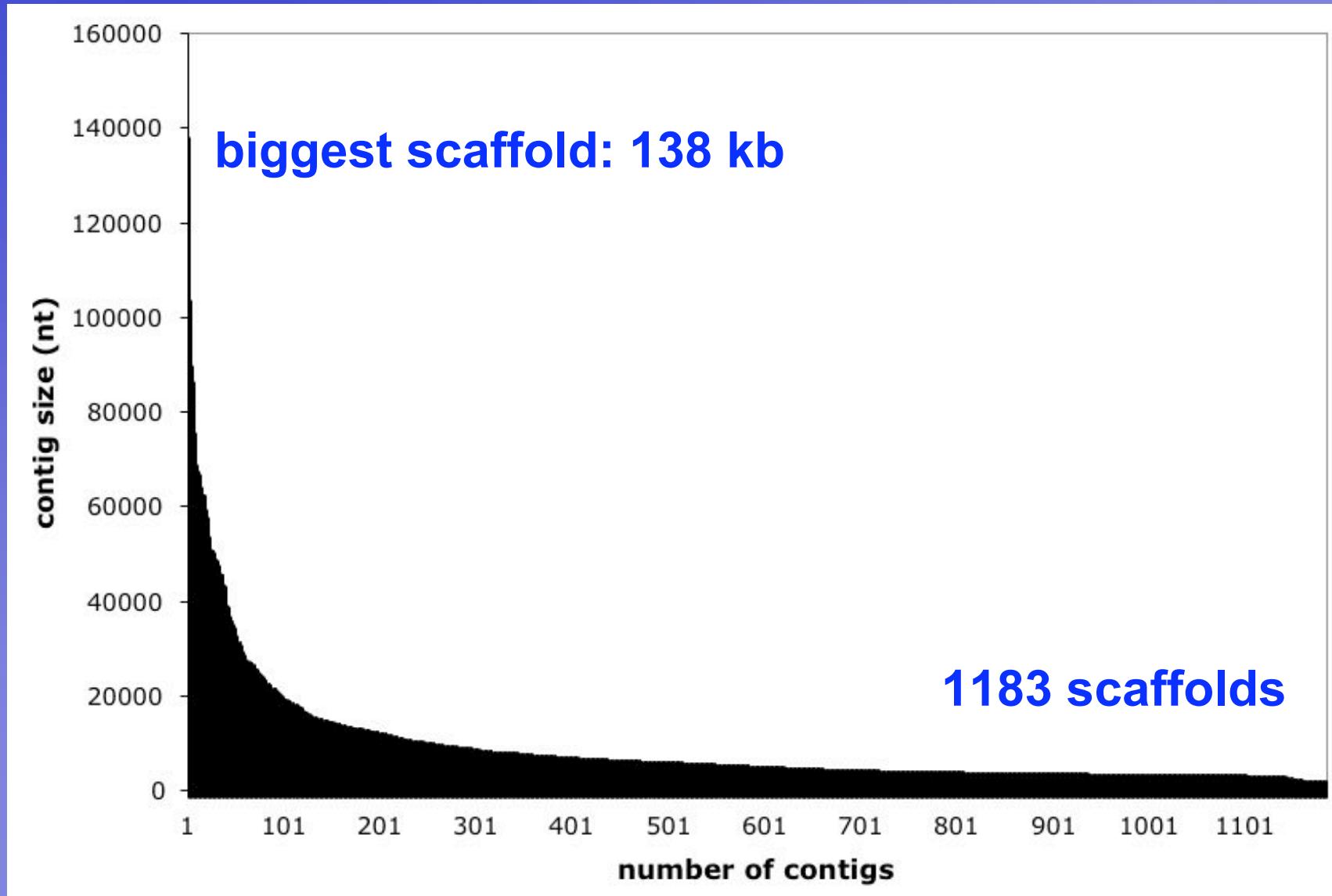
Ferroplasma acidarmanus isolate

- isolated in 1997
- 1.94 Mb genome of *F. acidarmanus* fer1
- annotation completed

AMD biofilm community

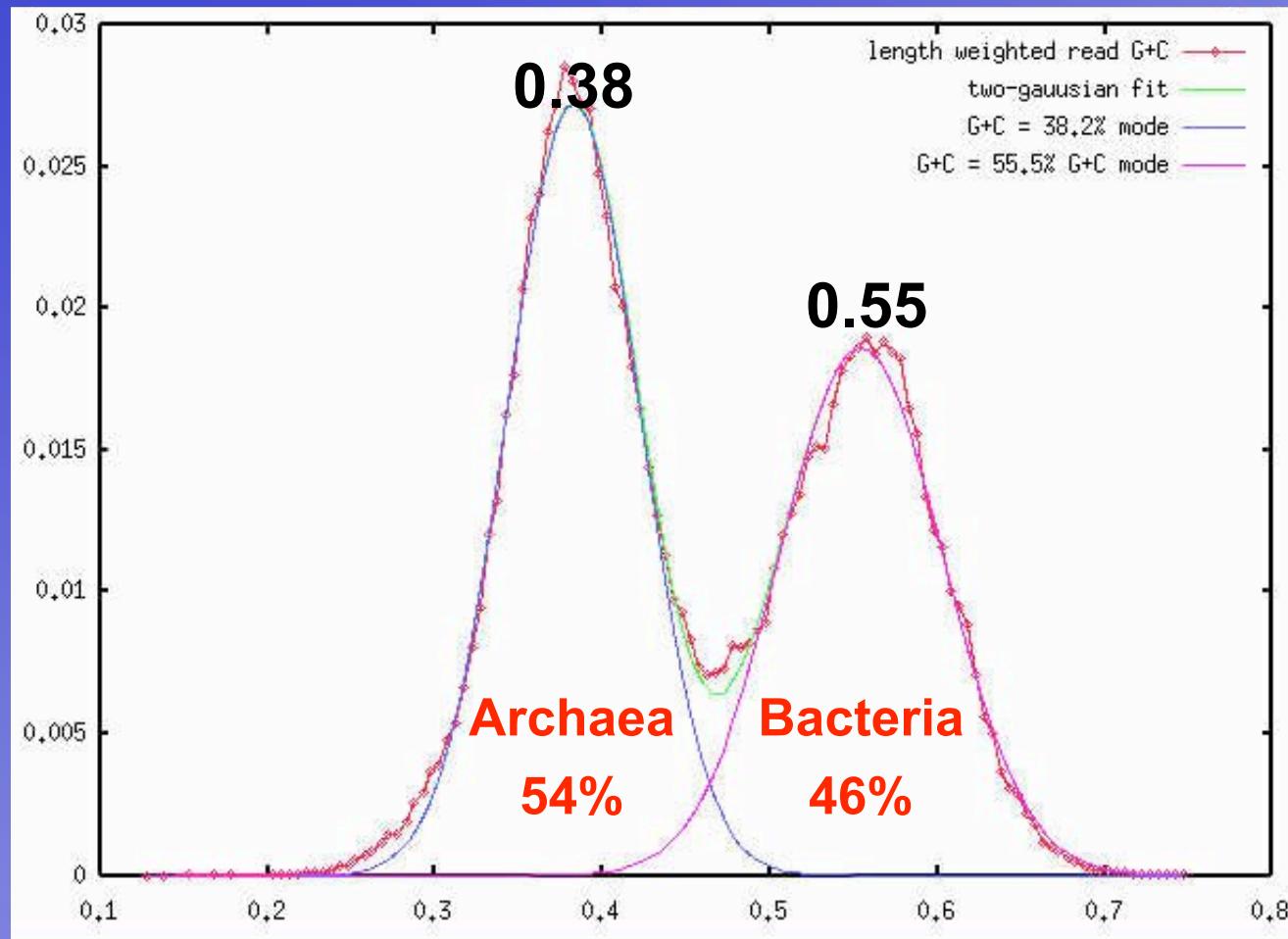
- 76 Mb shotgun library end-sequence data [133 Mb]
103,462 high quality reads (~737 bp/read) [189,000 reads]
- draft assembly 1,183 scaffolds (>2kb) totaling 10.83 Mb
- draft annotation completed

Assembly of the AMD community genomic data



G+C content of community genome scaffolds

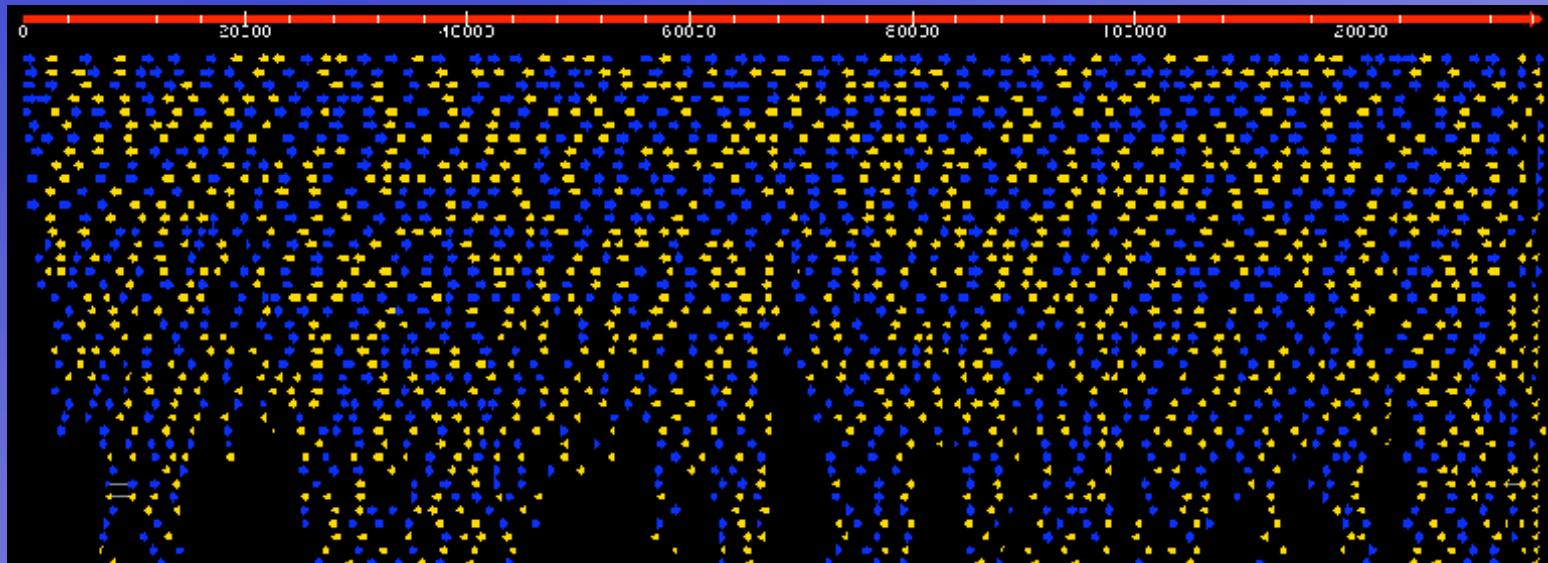
Fraction of total bases



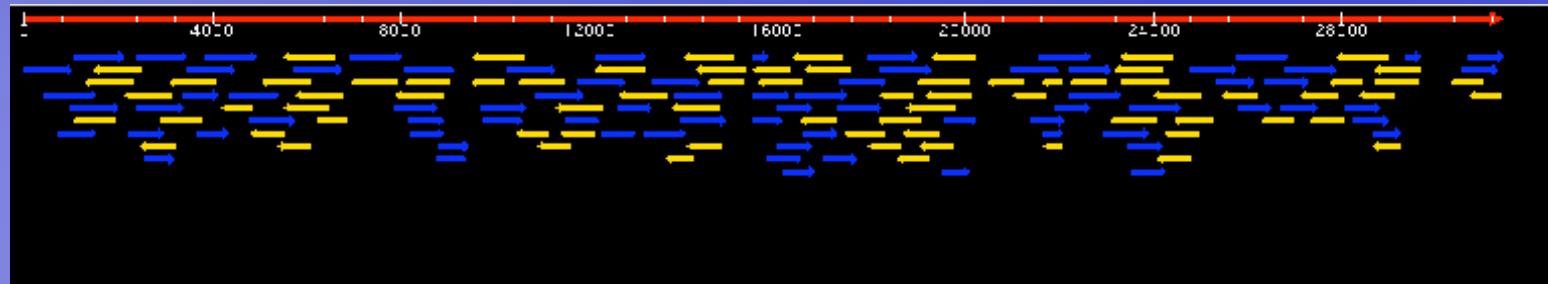
G+C content

Read depth / coverage

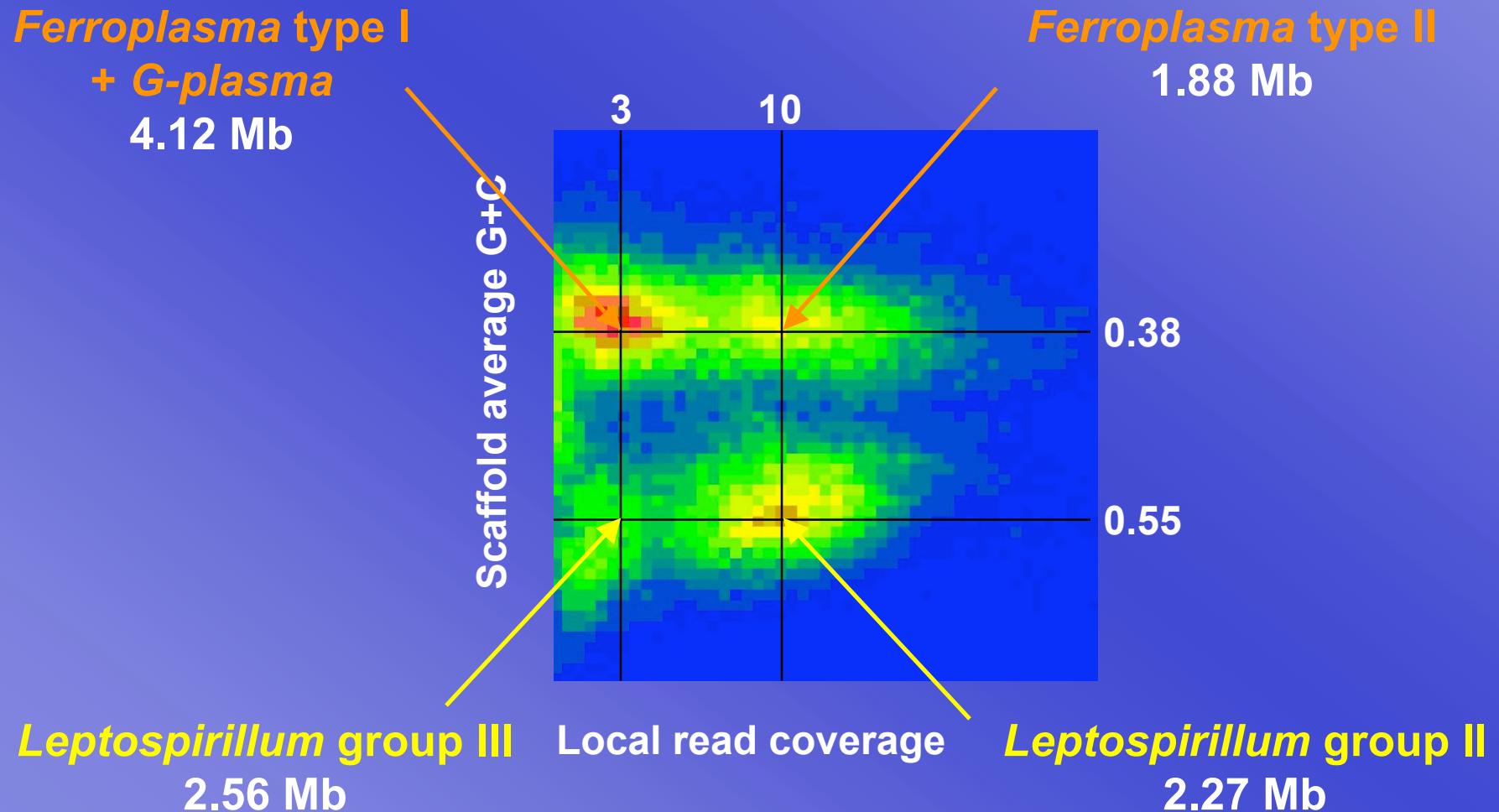
High depth scaffold - *Leptospirillum* group II



Low depth scaffold - *Leptospirillum* group III



Binning assembled community genome data by GC and depth



Poisson Calculations for Shotgun Sequencing (Lander Waterman)

Fold coverage	Percent of genome sequenced
0.25 x	22%
0.50 x	39%
0.75 x	53%
1 x	63%
2 x	88%
3 x	95%
4 x	98%
5 x	99.4%
6 x	99.75%
7 x	99.91%
8 x	99.97%
9 x	99.99%
10 x	99.995%

Community Genome Sequencing

-Near complete recovery of two genomes

-*Leptospirillum* group II

-*Ferroplasma* type II



*First sequenced member
of the Nitrospira phylum*

-Partial recovery of three other genomes

-*Leptospirillum* group III

-*Ferroplasma* type I

-G-plasma

*New species not recognized
by 16S analysis*

Uncultured microorganism

-Minor components sampled

-*Sulfobacillus thermosulfooxidans*

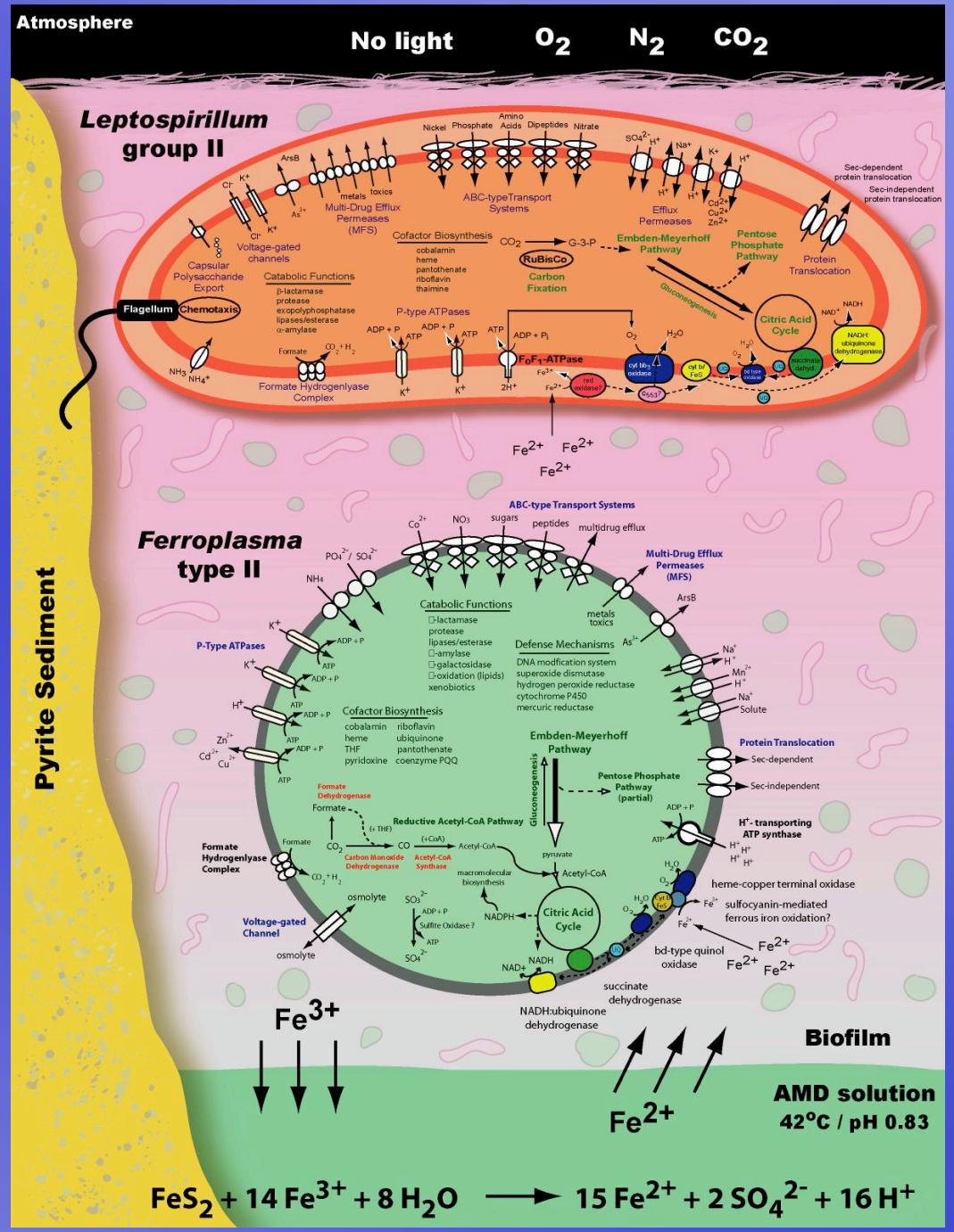
-A-plasma

Metagenomics provides:

- insight into the metabolism of organisms and overall community function

Metabolic Network

- EPS production (cellulose synthase)
- Motility (response to ferrous iron and oxygen gradients)
- Oxidative stress resistance
- Genes for resistance to copper, arsenite, mercury, zinc, silver, and cadmium (efflux pumps)
- Electron transport chain components and a number of novel cytochromes
- Partitioning of community essential roles
- C and N fixation



Metagenomics provides:

- **insight into the metabolism of organisms and overall community function**
- **sequences of co-habiting / co-evolving populations for comparative analyses**

Leptospirillum group II genomic variation

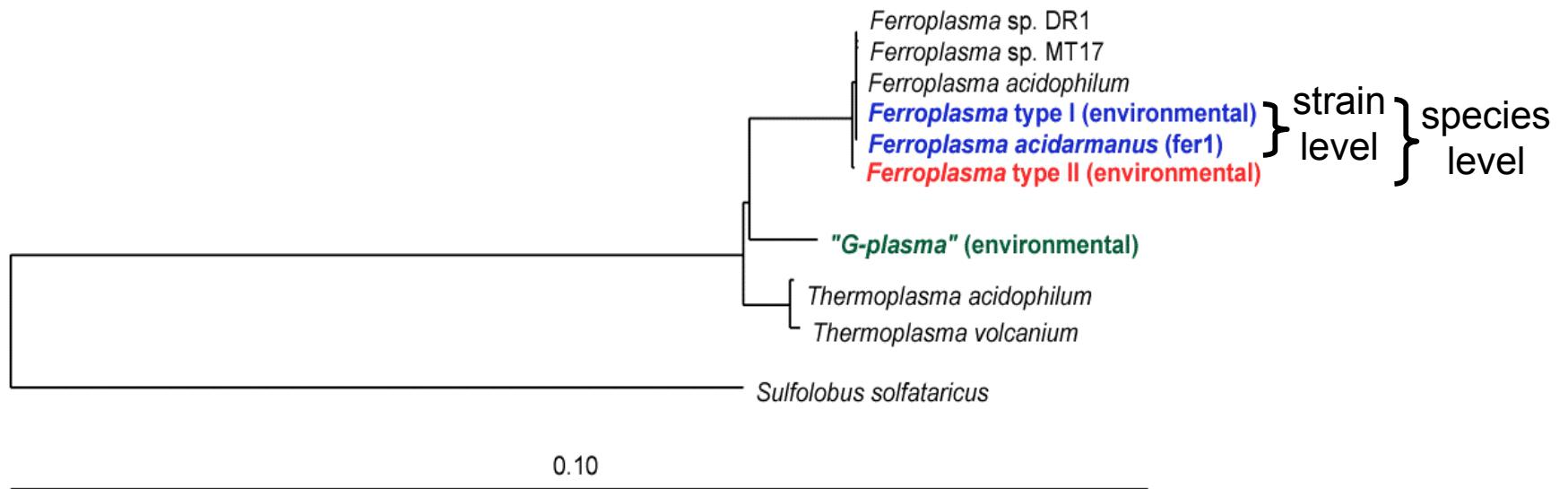
CONSENSUS	7245	cgagagagcttggcctcggtcggaatcgataccactgcatacagggtatctggta	7304
XYG29548.g1	396 a	455
XYG30059.g1	400	459
XYG2928.g2	203 a	262
XYD2524.g1	12	71
XYG33406.g1	112 a	171
XYG22598.g3	214	273
XYG52392.b1	295	354
XYG61369.b1	449	508

1 nt polymorphism / ~1,300 bases

1 strain

Recent selection event (sweep)?
Founder effect?

16S rRNA gene phylogeny of the *Thermoplasmatales*



fer1 : “fer1env” = 0.0% 16S divergence

fer1 : fer2 = 0.8% 16S divergence

CG
reads {

homogeneous region of fer1env population

ISOLATE	959	cattcgccagggaaatacactttgccttggtgaccacttgcaaggtagaaggcagaa
XYG23099.b1	164
XYD3165.b1	316
XYD2007.b1	242
XYD2300.g1	211
ISOLATE	1019	tagcactggaagaatttcaaaggctgttcctggattgaaaattgaaaaggccttcctg
XYG23099.b1	104
XYD3165.b1	376
XYD2007.b1	302
XYD2300.g1	271

cytochrome
P450

heterogeneous region of fer1env population – 2 variants

ISOLATE	361	gaatatggctgcaacacaattgcacatggtaacaggcaaggcaatgtcaggtaaga
XYG35496.g1	530
XYG3784.b2	266
XYG66165.g1	298	...g.....c.....c.....g
XYD6944.b1	224	...g.....c.....c.....g
ISOLATE	421	tttgaggttacaatgaaaggccctaatgaaatataatgtaatcgcacccgtaaggat
XYG35496.g1	490
XYG3784.b2	326
XYG66165.g1	238a....c.....a....a....
XYD6944.b1	164a....c.....a....

arginino-
succinate
synthase

only fer1env sequence type

ISOLATE	661	acactgggttacatggcatcctatgactatattttcacatggtatcttaacagaatg
XYG7614.g2	399t.....c.....c.....
XYG31767.g1	308t.....c.....c.....
XYD7766.b1	199t.....c.....c.....
XYG44653.b1	351t.....c.....
ISOLATE	721	agggaaccttcagatgggatataattgtcccataatggaaatgtcggttgcaggcact
XYG7614.g2	339g.....c.....a.....a
XYG31767.g1	248g.....c.....a.....a
XYD7766.b1	139g.....c.....a.....a
XYG44653.b1	411g.....c.....a.....a

glycerol-3-
phosphate-
dehydrog.
related

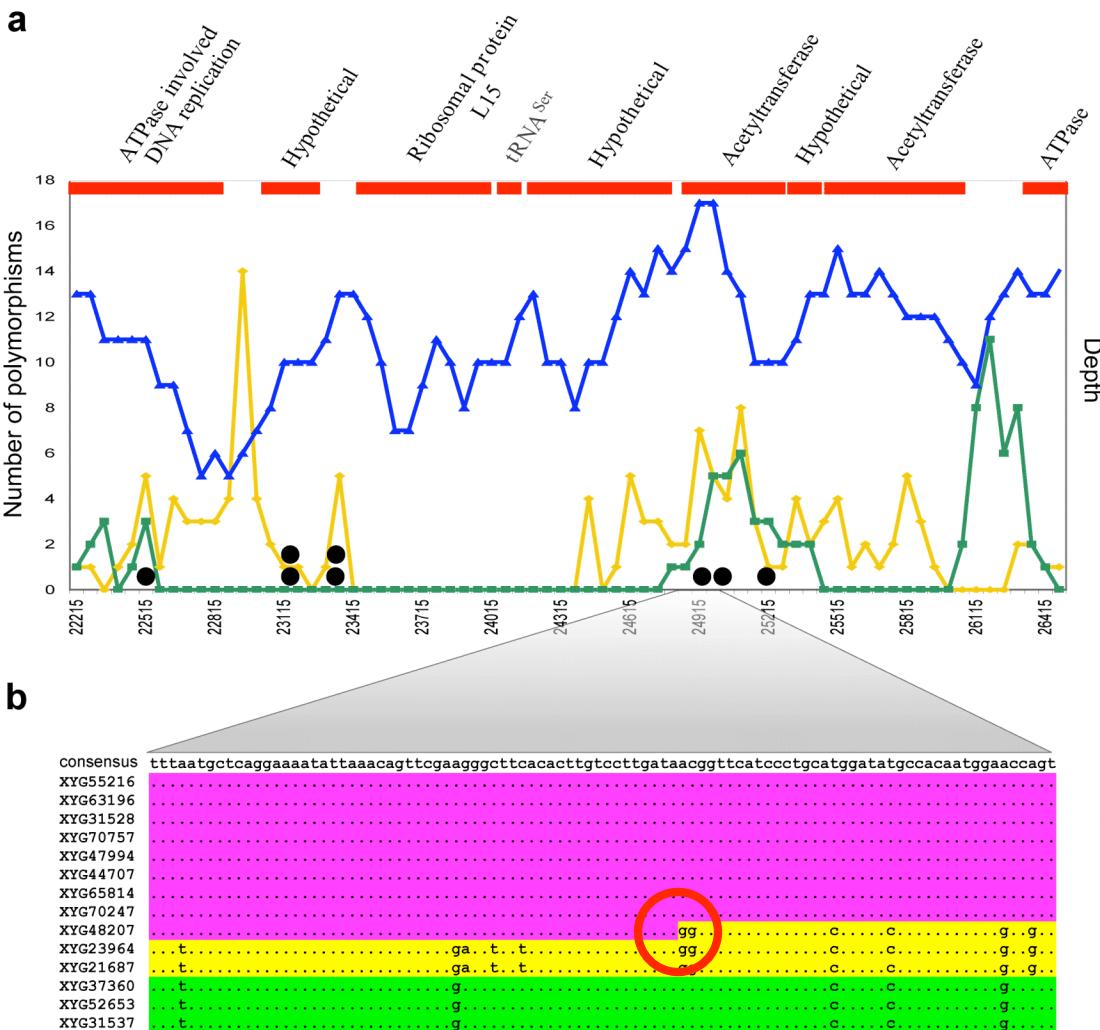
evidence for inter-strain recombination (glycosyltransferase)

QUERY	121	aagtcaacaactgagcaatcactttcacggcaattatacgaaataattattgttaaaaacttccggatacaatattgacgctttgca
fer1.144c	121
XYG7465.g2	215
XYD1690.g1	200
XYG64953.g1	198t.t..a.....a..c.....g.....aa.....t....c..g
XYG17515.g3	374t.t..a.....a..c.....g.....aa.....t....c..g

QUERY	211	gaagagtataatgtatataatattatagtaaaaatgaaacgctcttgaaagataatagaagcaatcggatcgccaggggaatatt
fer1.144c	211
XYG7465.g2	365
XYD1690.g1	250c.....a.....a.....c.....a.....g..c..g..t..t.....a.....
XYG64953.g1	348c.c.....t.....g.....c.....a.....a.....c.....a.....g..c..g..t..t.....a.....
XYG17515.g3	524c.c.....t.....g.....c.....a.....a.....

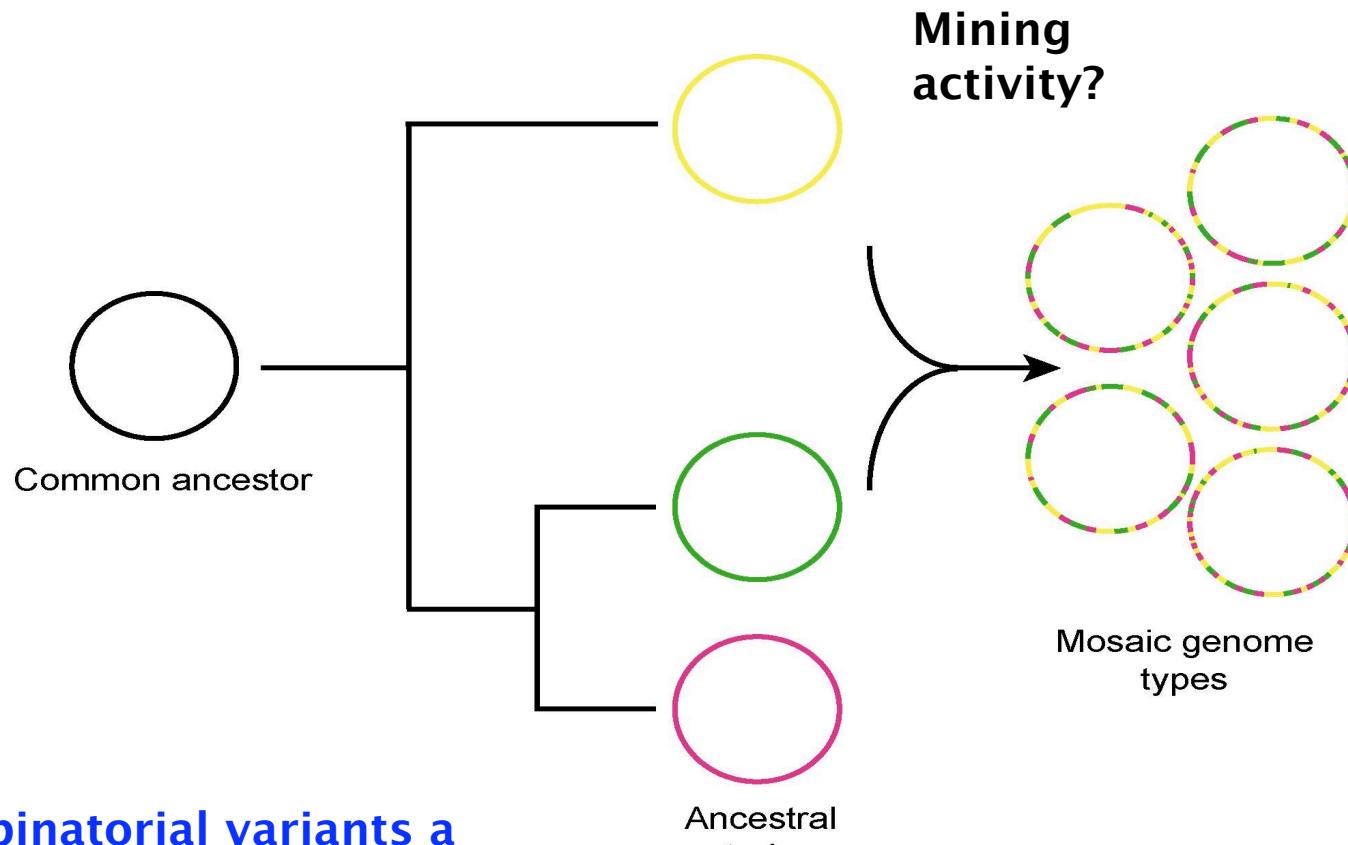
- observe linkage between isolate-type and non-isolate type sequence (mate-pairs)
- observe transitions within single reads...

Ferroplasma type II inter-strain variation



Tyson et al. *Nature*, 2004

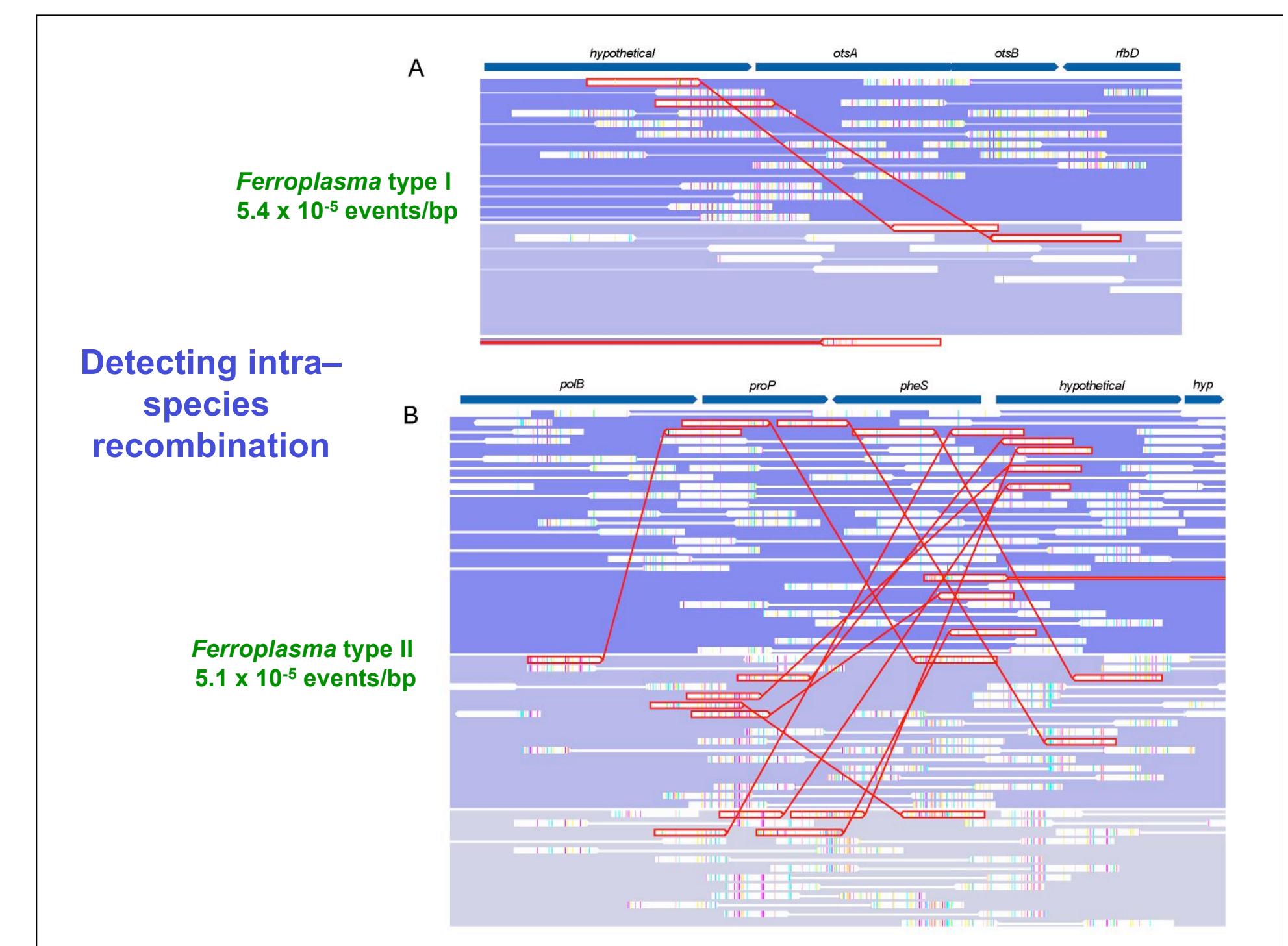
Mosaic genome structure

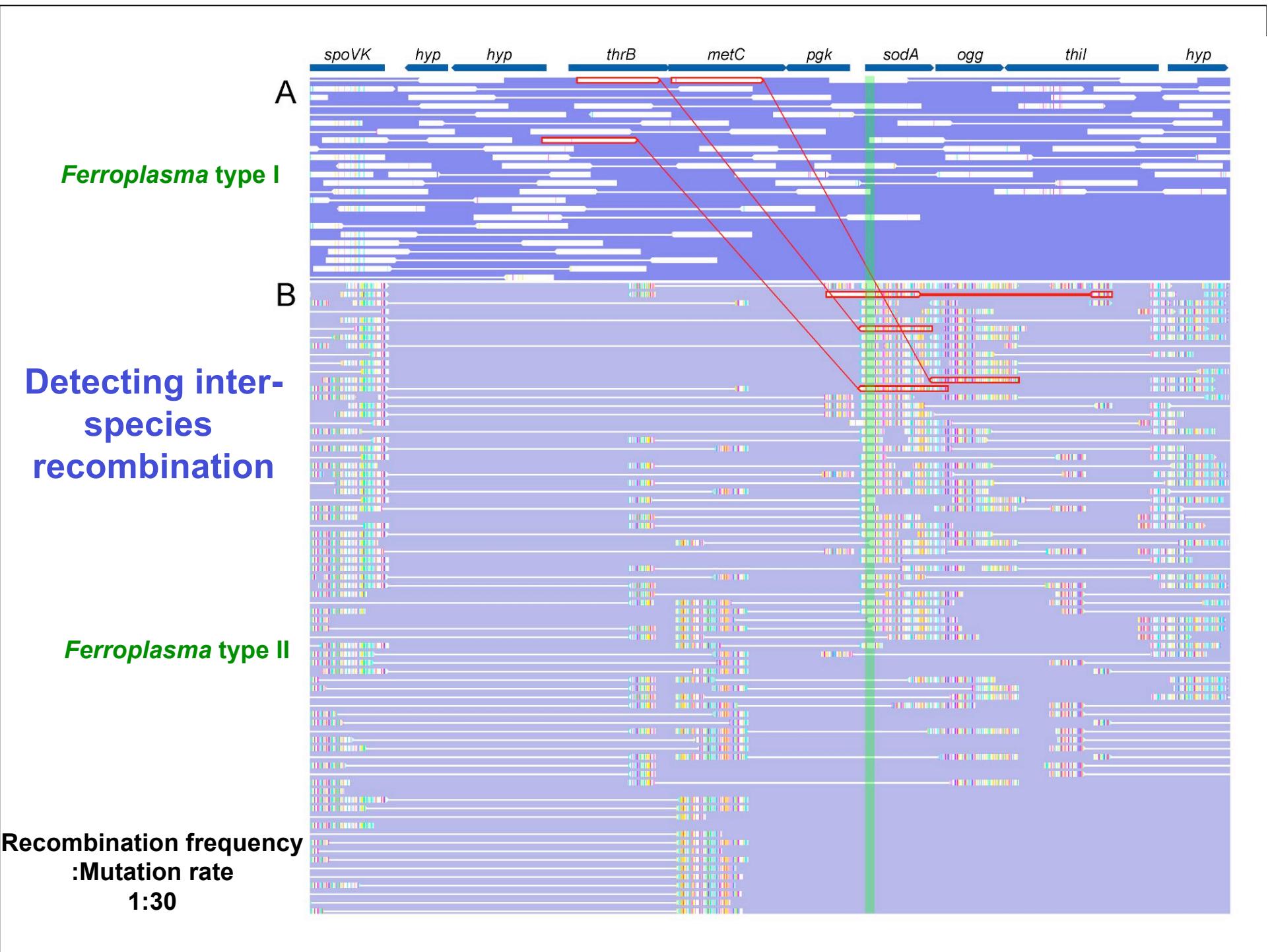


Are combinatorial variants a strategy for fine-tuning environmental optimization?

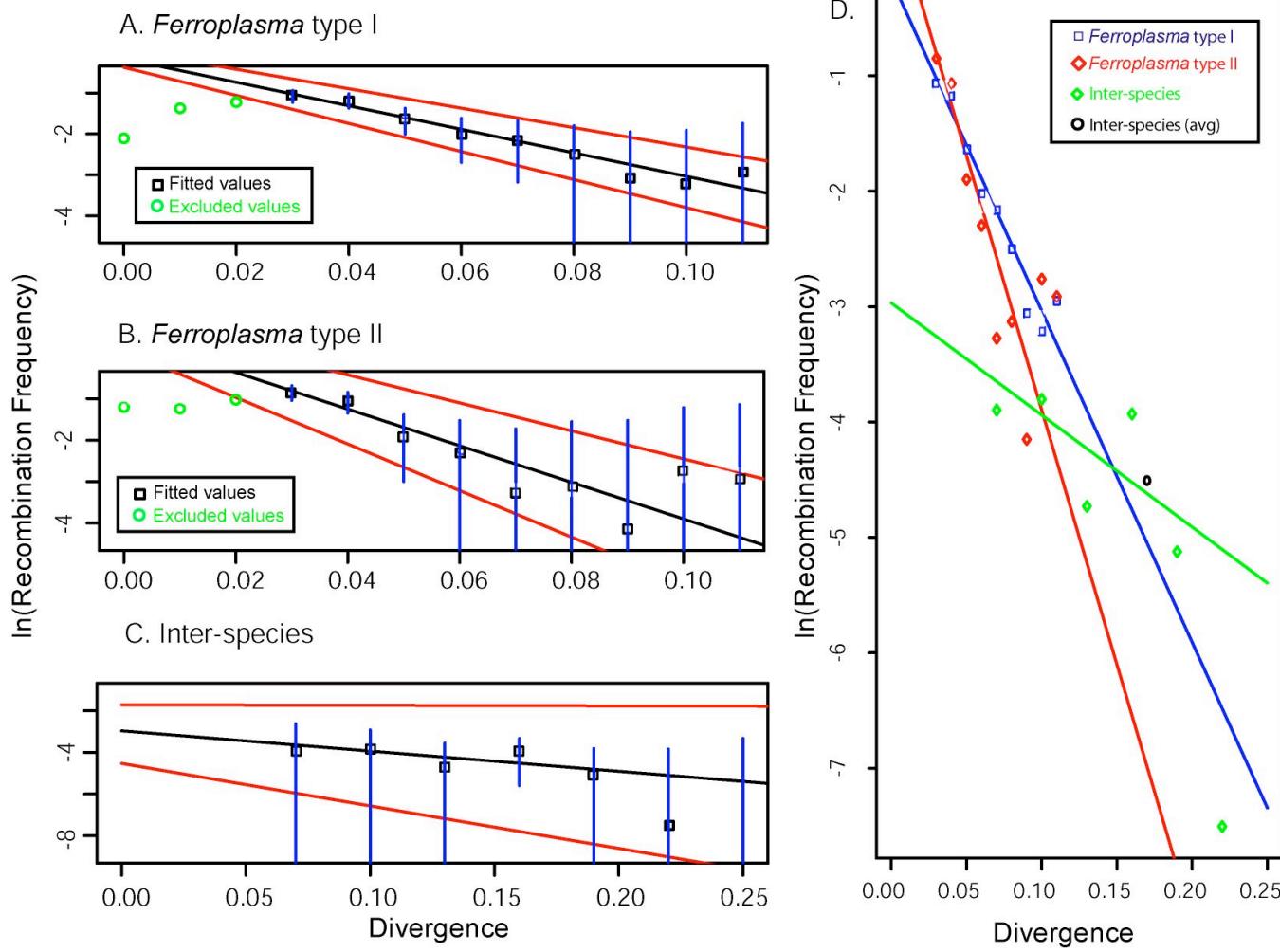
Can this (and other?) microbial species be defined like sexually reproducing organisms?

Detecting intra– species recombination





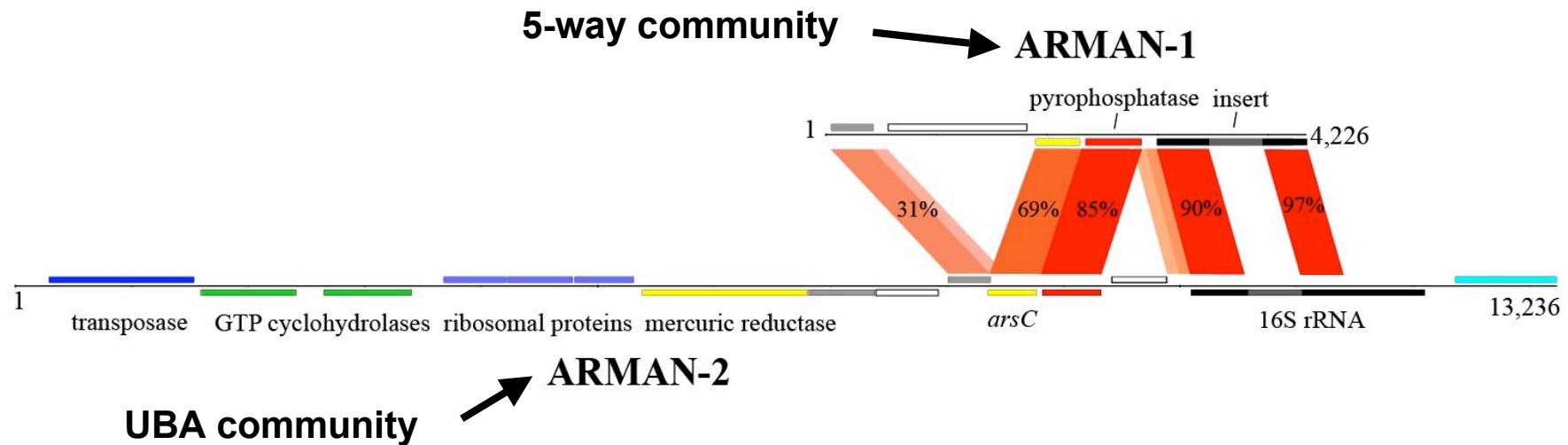
Recombination frequency versus sequence divergence



Metagenomics provides:

- **insight into the metabolism of organisms and overall community function**
- **sequences of co-habiting / co-evolving populations for comparative analyses**
- **possibility to detect organisms missed in 16S rRNA surveys**

Detection of Novel Diversity



Mismatches with commonly used 16S rRNA gene primers

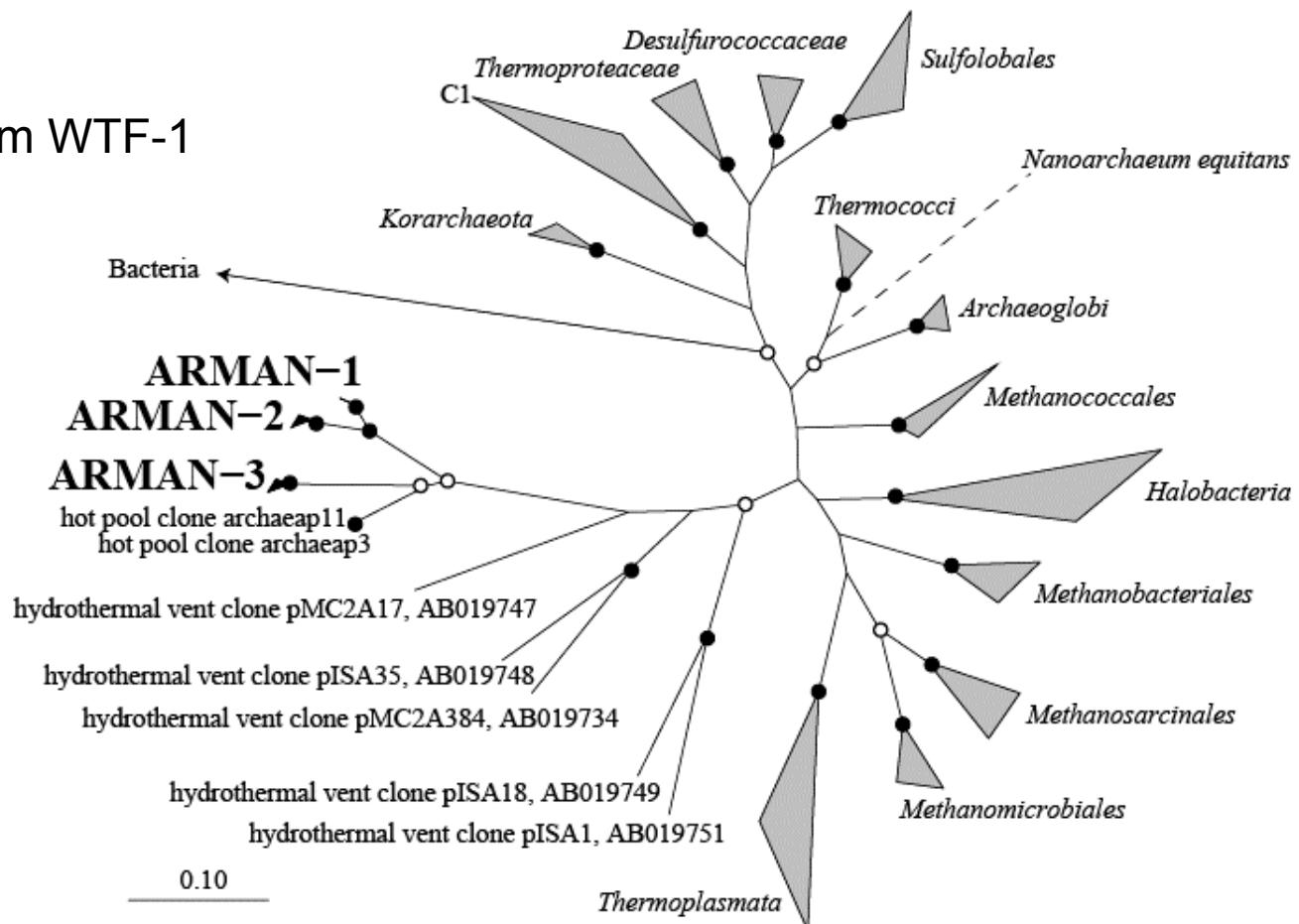
Primer	Primer sequence
21F	TT <u>CCG</u> GTT GATCCYGCC <u>GG</u> A
23F	T <u>CCG</u> GTT GATCCYGCC
1492R	GG <u>T</u> ACCTTGTAAUUC <u>GA</u> UUU <u>CT</u> T
1391R	GU <u>ACG</u> GGCGGTGWGTRCA
1525R	A <u>AGG</u> AGGTGAT <u>CC</u> <u>AG</u> UCC

16S rRNA Analysis

2 new groups:

8% and 17%
divergent from WTF-1

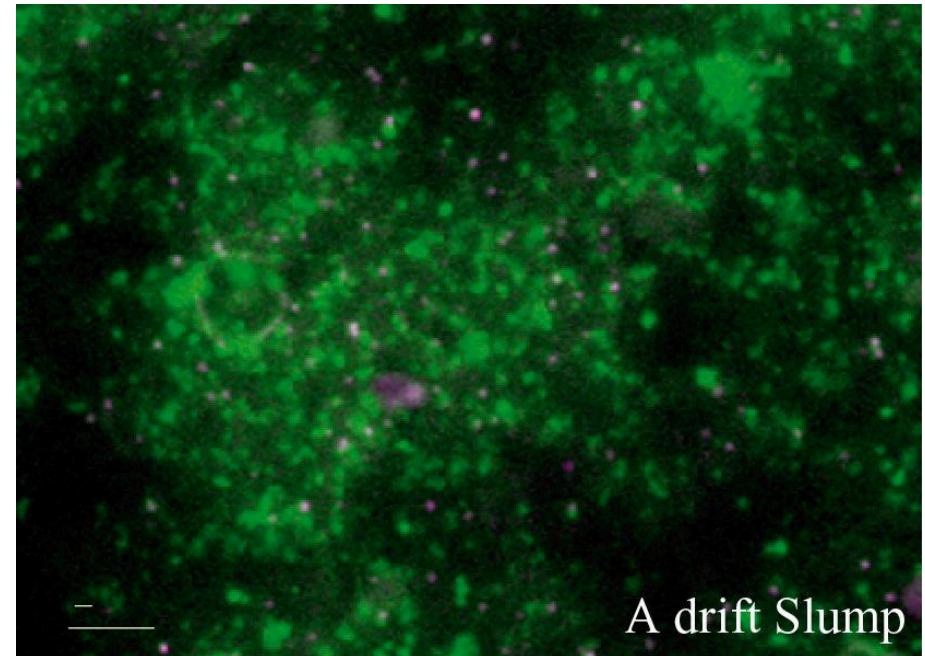
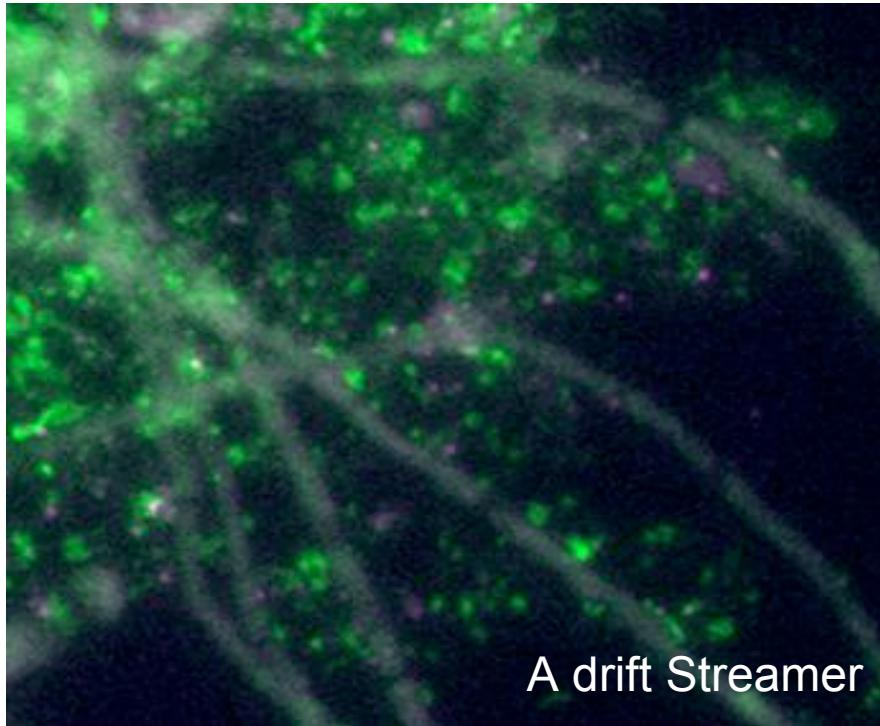
Crenarchaeota



Euryarchaeota

FISH (fluorescent *in situ* hybridization) analyses of ARMAN groups

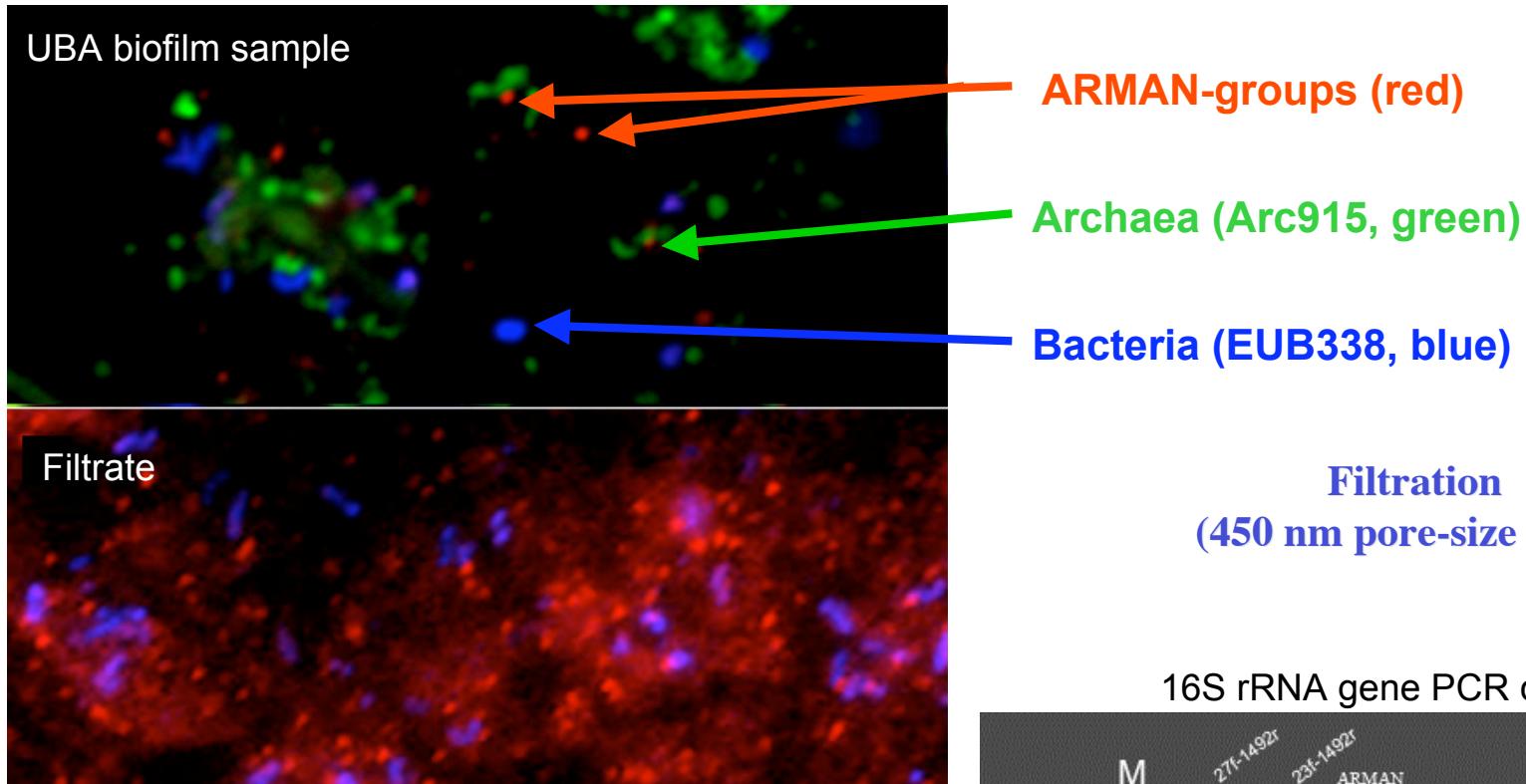
- They are present in low abundance, but seen in ALL samples in the mine
- MUCH smaller than other cells in the mine



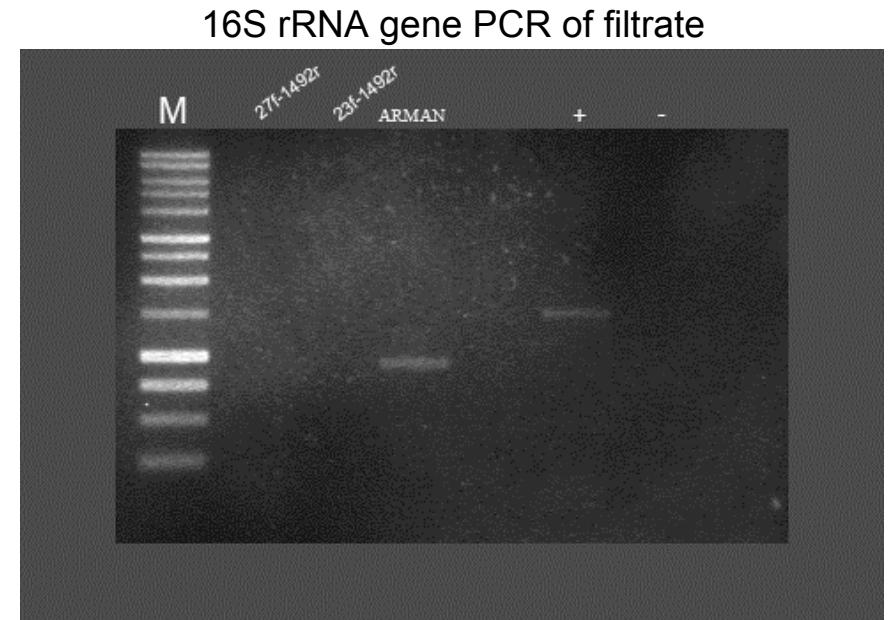
Archaea (Arc915, green)
ARMAN-groups (pink)

Since they are smaller than other cells we should be able to concentrate them by filtration...

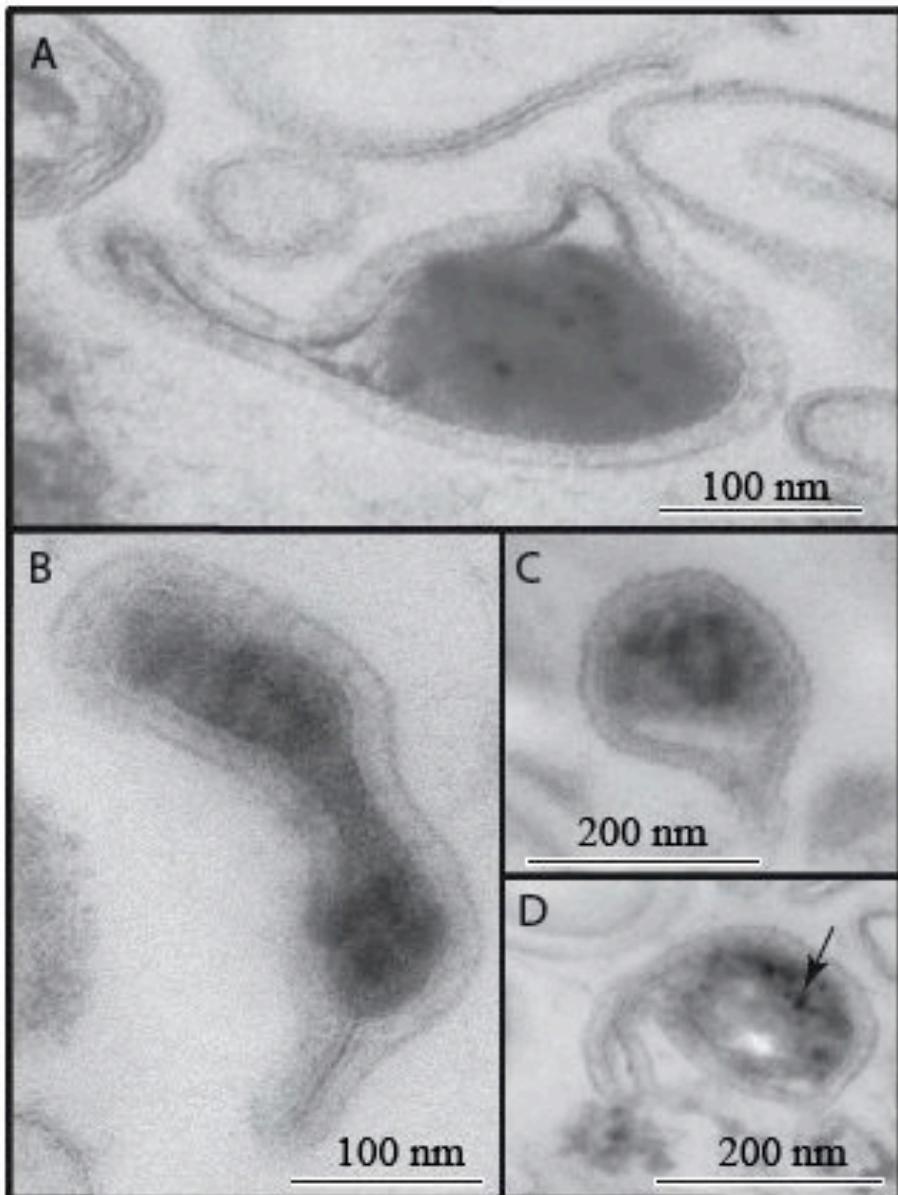
Concentration of ARMAN groups



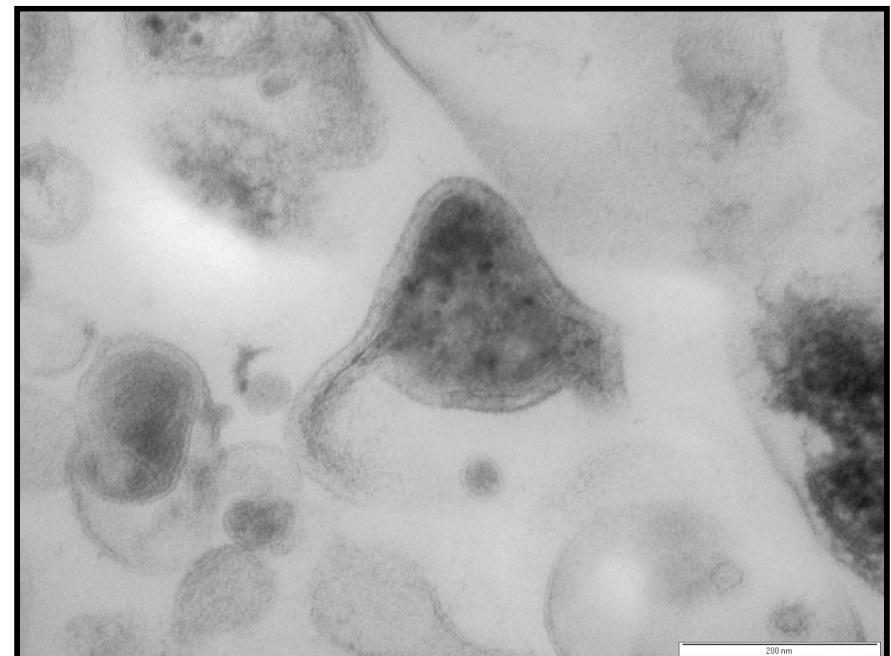
The ARMAN groups are highly enriched in the filtrate.



TEM characterization of the 450 nm filtrate



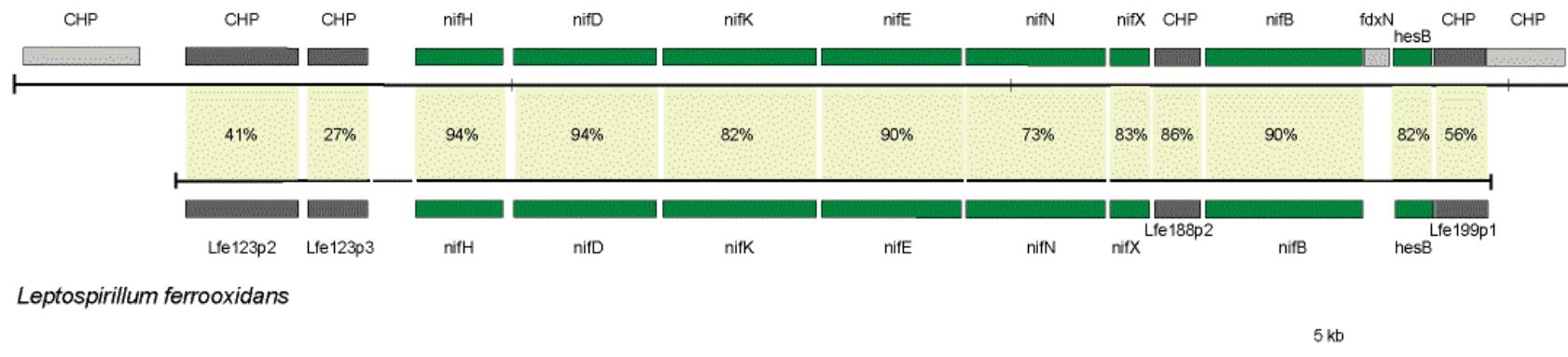
- The cells range in size from 170nm to 240nm (averaging 200 nm)
- The cell walls contain an S-layer
- The cytoplasm is densely packed
- 1 to 2 protrusions on each cell



Metagenomics provides:

- **insight into the metabolism of organisms and overall community function**
- **sequences of co-habiting / co-evolving populations for comparative analyses**
- **possibility to detect organisms missed in 16S surveys**
- **clues for cultivating uncultured organisms**

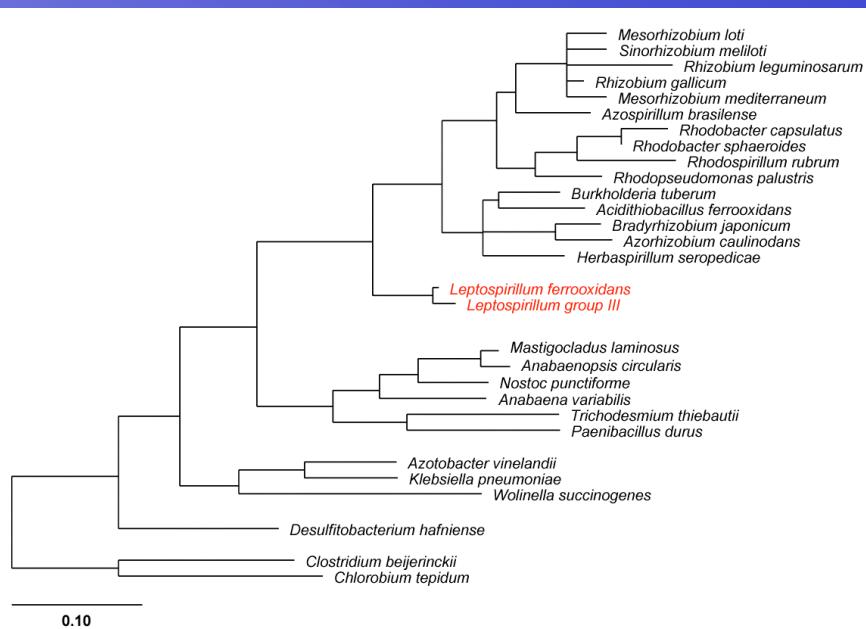
Leptospirillum group III



Leptospirillum ferrooxidans

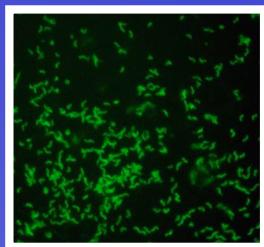
5 kb

*Environmental nifH
libraries confirmed the
presence of only one
nitrogen fixer*



Genome-directed isolation

Screened samples
using FISH



Serial diluted

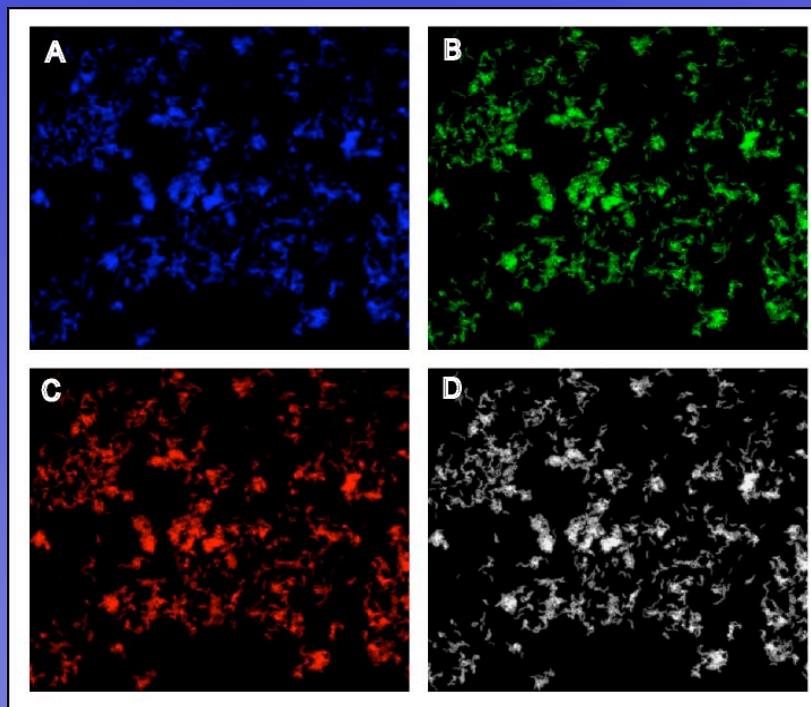


45 days at 37°C

Subcultured



EUB338
All Bacteria



LF655
Leptospirillum groups
I, II and III

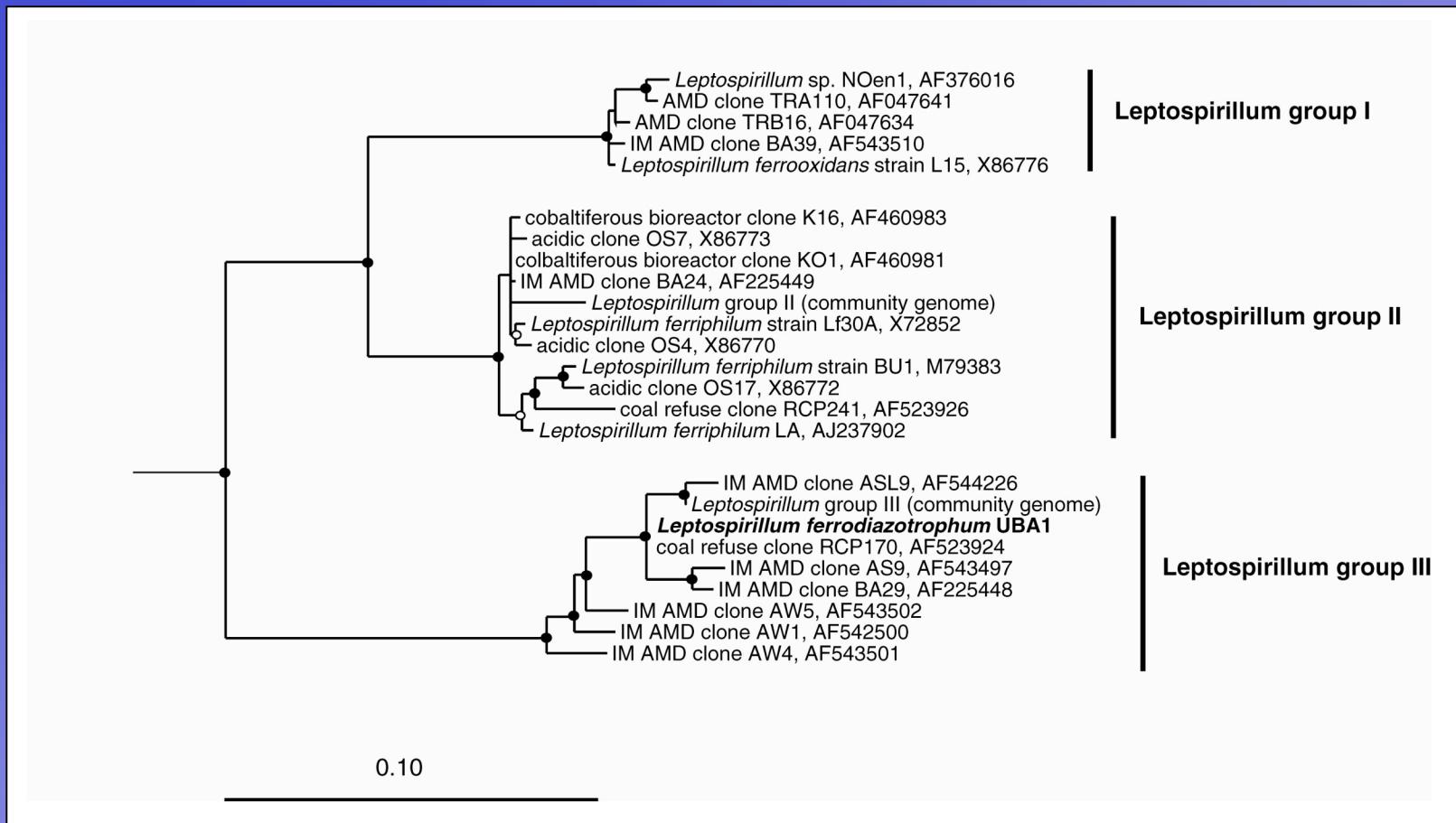
LF1252
Leptospirillum
group III only

Combined image showing
all three probes

“Genome-directed isolation of *Leptospirillum ferridiazotrophum*, the key nitrogen fixer
in acid mine drainage communities”

Gene W. Tyson, Ian Lo, Brett J. Baker, Eric E. Allen, Philip Hugenholtz & Jillian F. Banfield (AEM)

16S rRNA analysis of *Leptospirillum* group III culture



“Genome-directed isolation of *Leptospirillum ferrodiazotrophum*, the key nitrogen fixer in acid mine drainage communities”

G.W. Tyson *et al.*, (AEM)

Community genomics provides:

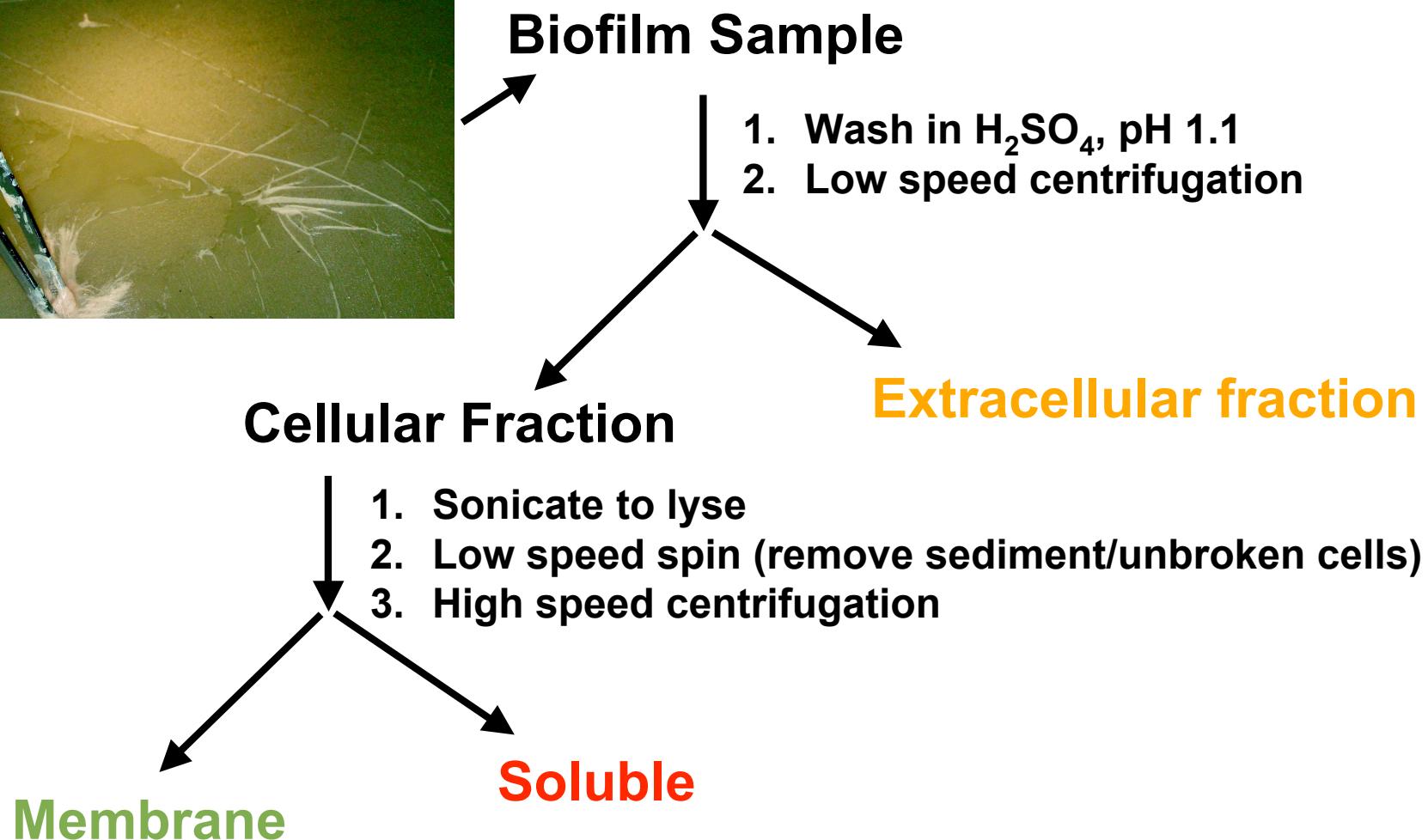
- insight into the metabolism of organisms and overall community function
- sequences of co-habiting / co-evolving populations for comparative analyses
- possibility to detect organisms missed in 16S surveys
- clues for cultivating uncultured organisms
- basis for proteomic analyses of ‘whole’ communities

Proteome Sample (FISH)

<u>Genome</u>	<u>Proteome</u>	
75%	83%	<i>Leptospirillum</i> group II
10%	9%	<i>Leptospirillum</i> group III
10%	8%	Archaea
1%	1%	<i>non-Leptospirillum</i> bacteria

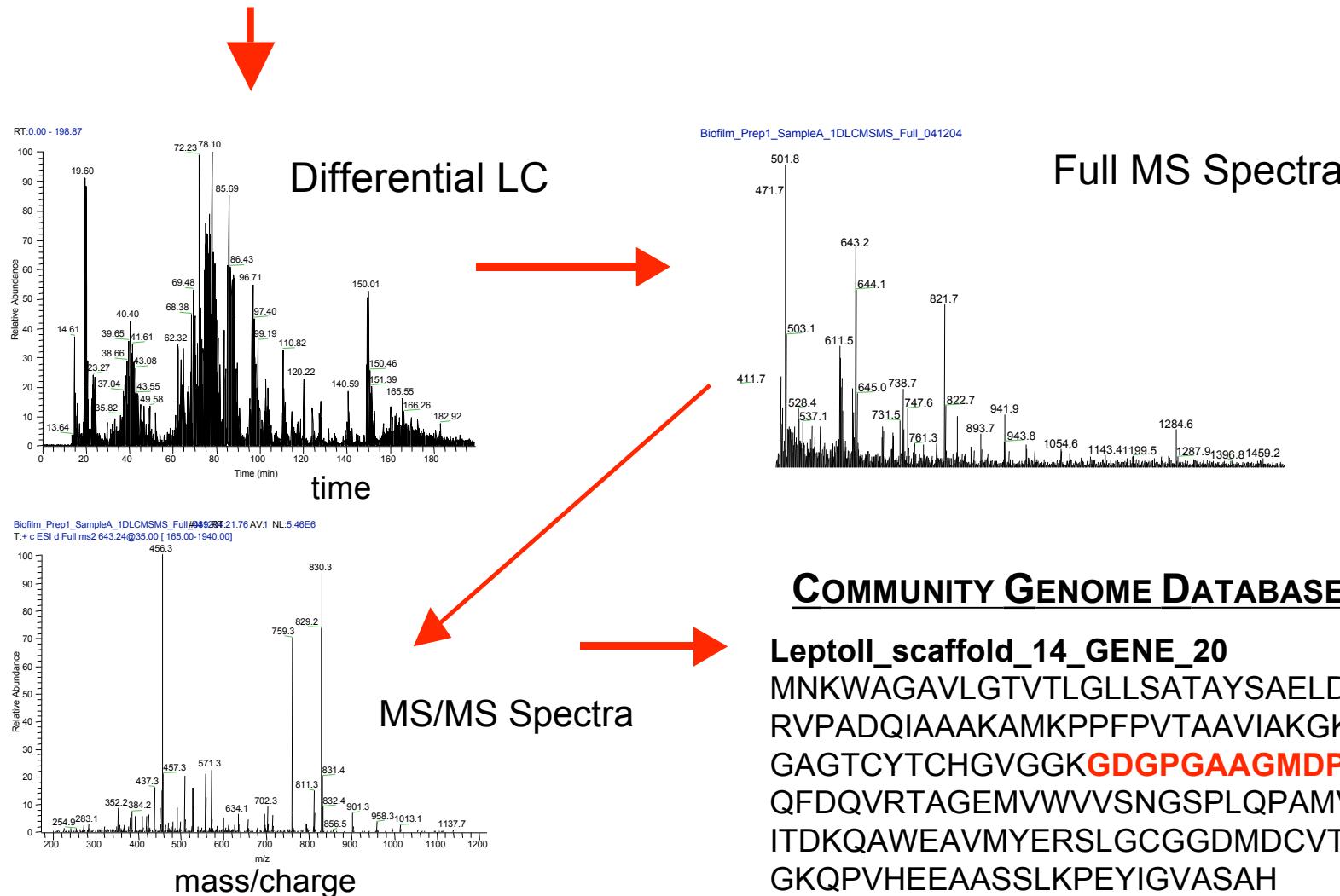
Ram et al. *Science* (2005)

Protein Extraction for Mass Spectrometry



Community Proteomics: 2D LC-MS/MS

Trypsin proteolysis of proteins in fractions



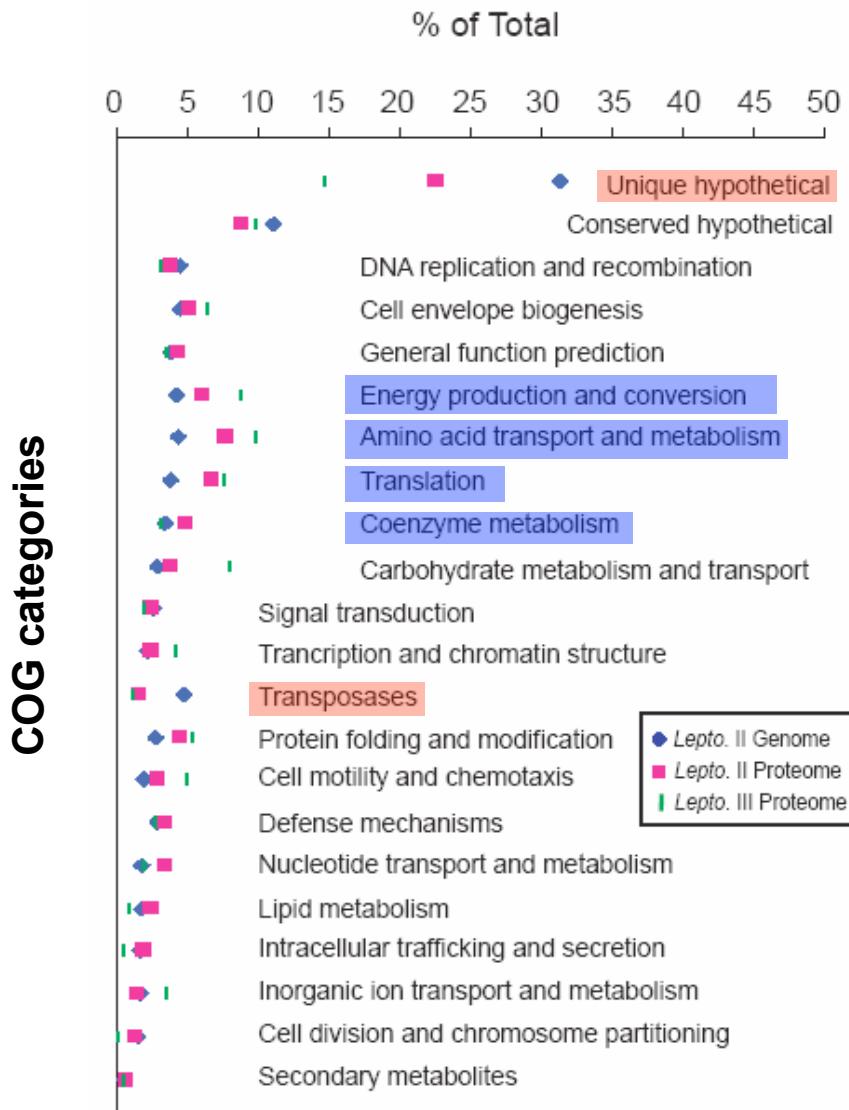
Proteins detected in the natural community



- 17% of predicted genes in the community
- 48% of proteins predicted from the *Leptospirillum* group II (1362 proteins)
- validation of 572 hypothetical proteins

Ram et al. *Science* (2005)

Functional Category Analysis of Expressed Proteins



prevalence of
hypothetical proteins

environmental context
clues to function

Importance of hypothetical proteins

31% of detected proteins are hypothetical

212 operons encoding proteins of ascribable function and expressed hypothetical proteins

e.g.: *operon encoding 8 flagellar proteins also encodes 4 detected hypothetical proteins*

Detected complete operons of hypothetical proteins

e.g.: *3 gene operon that is Leptospirillum group II*

Most abundant proteins overall

(compiled top 50 from each fraction)

17% hypothetical

13% ribosomal

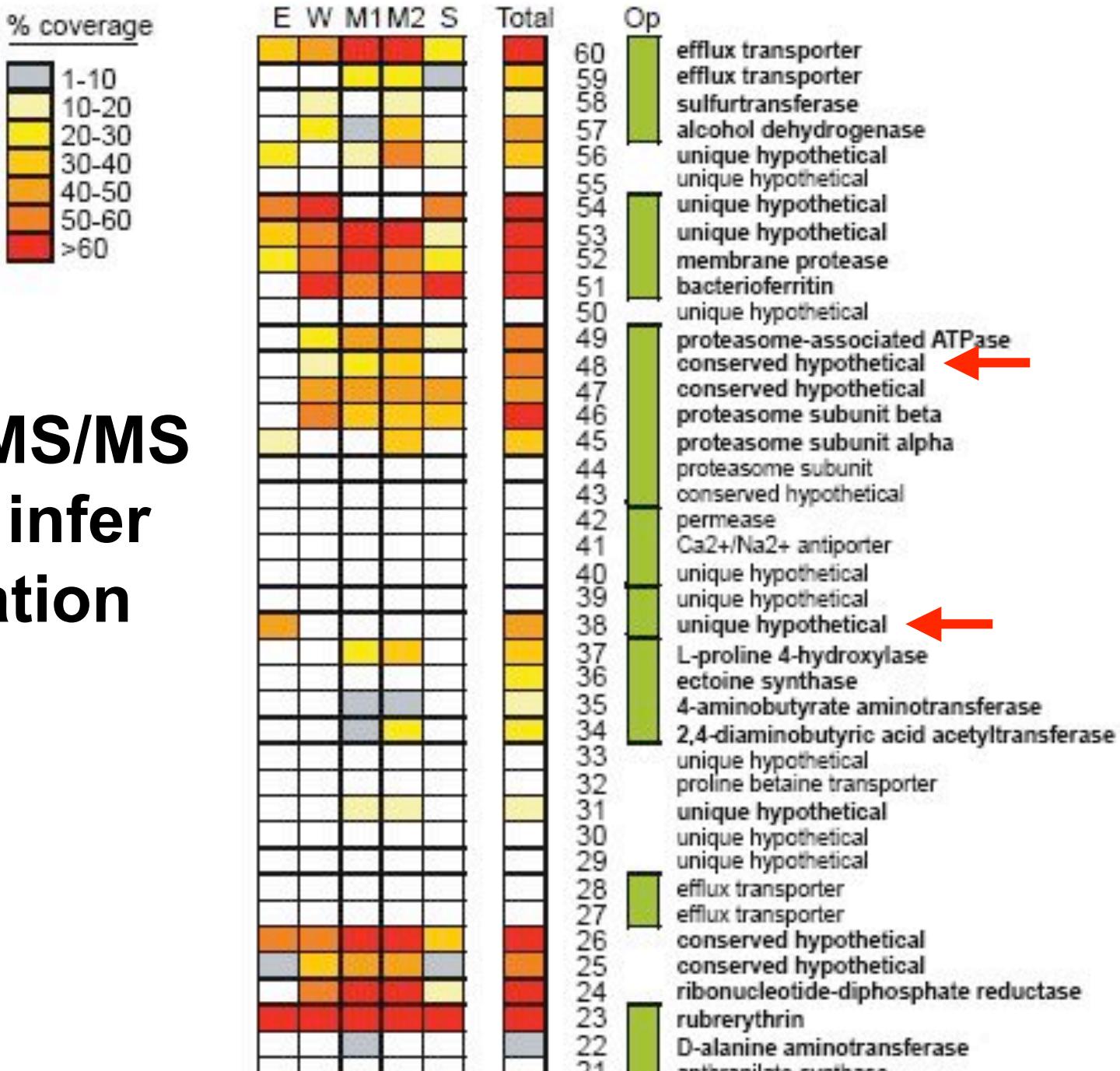
11% chaperones

9% thioredoxins

8% radical defense

***Protein folding
and radical defense***

Using MS/MS data to infer localization



Acid mine drainage



Sargasso Sea

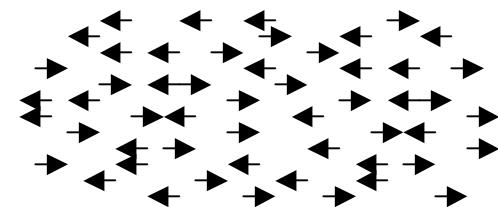
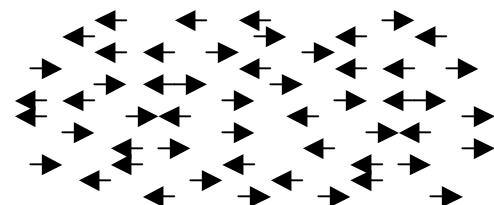
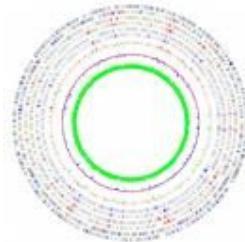
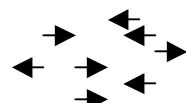
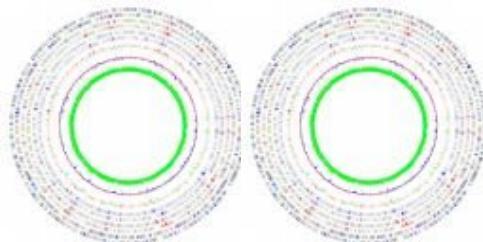


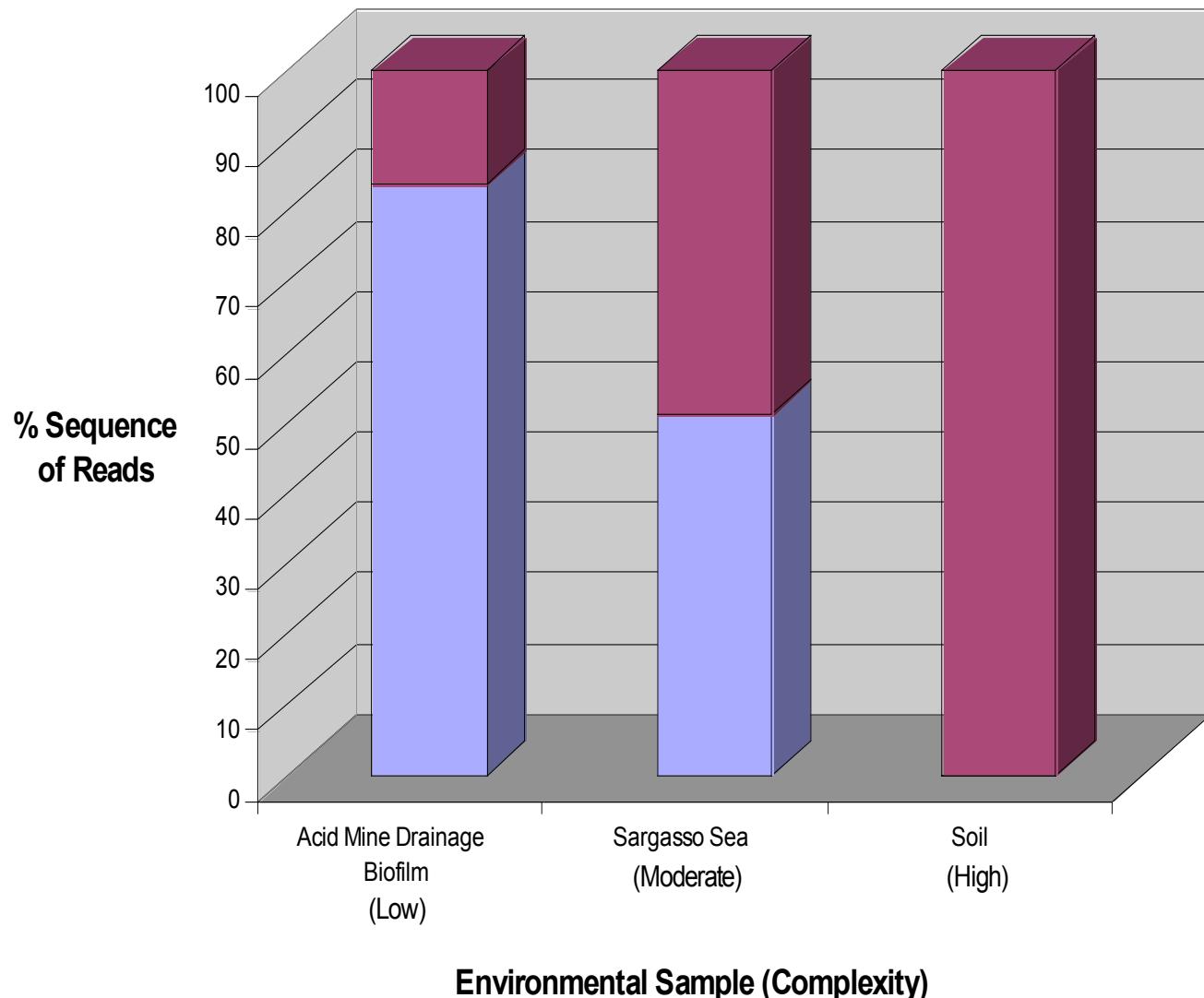
Soil



1 10 100 1000 10000

Species complexity →





Environmental Gene Tags (EGT)

Identify genes in sequence data (contigs to unassembled reads) from *multiple* environmental samples

Assign genes to their gene family, or higher level groupings

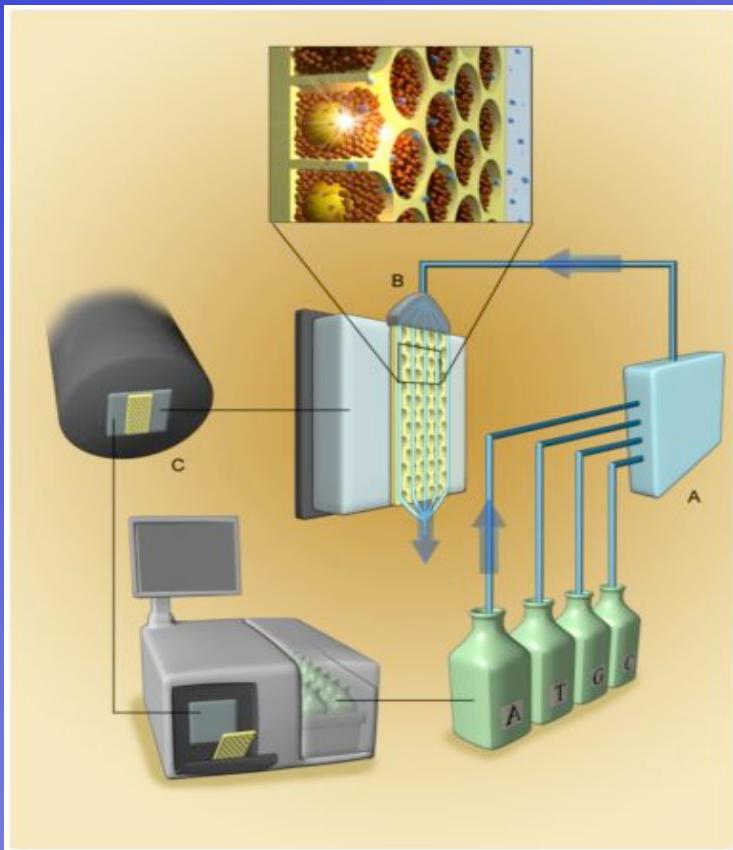
Compare relative abundance of different gene families according to habitat

Primary challenges for metagenomics in more complex environments

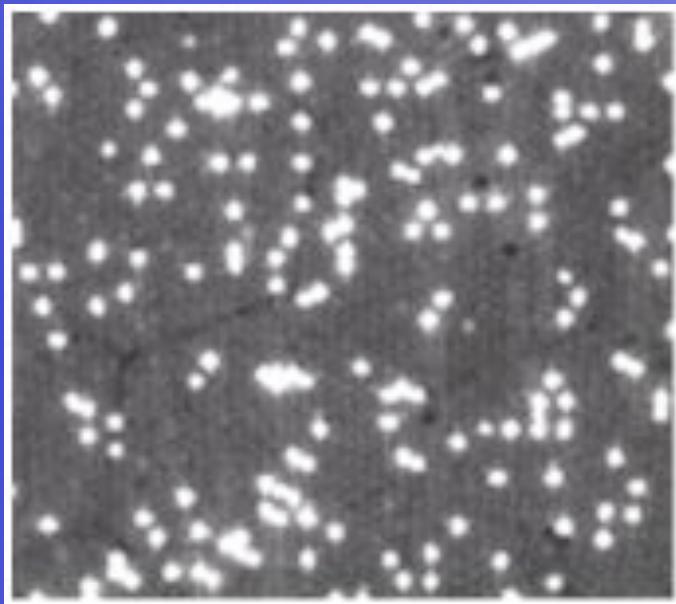
technology { Cost
Sequencing efficiency

Improved assembly algorithms/heuristics Resolution of strain heterogeneity { learn from simple communities

- Post-genomic databases
- Molecular evolution

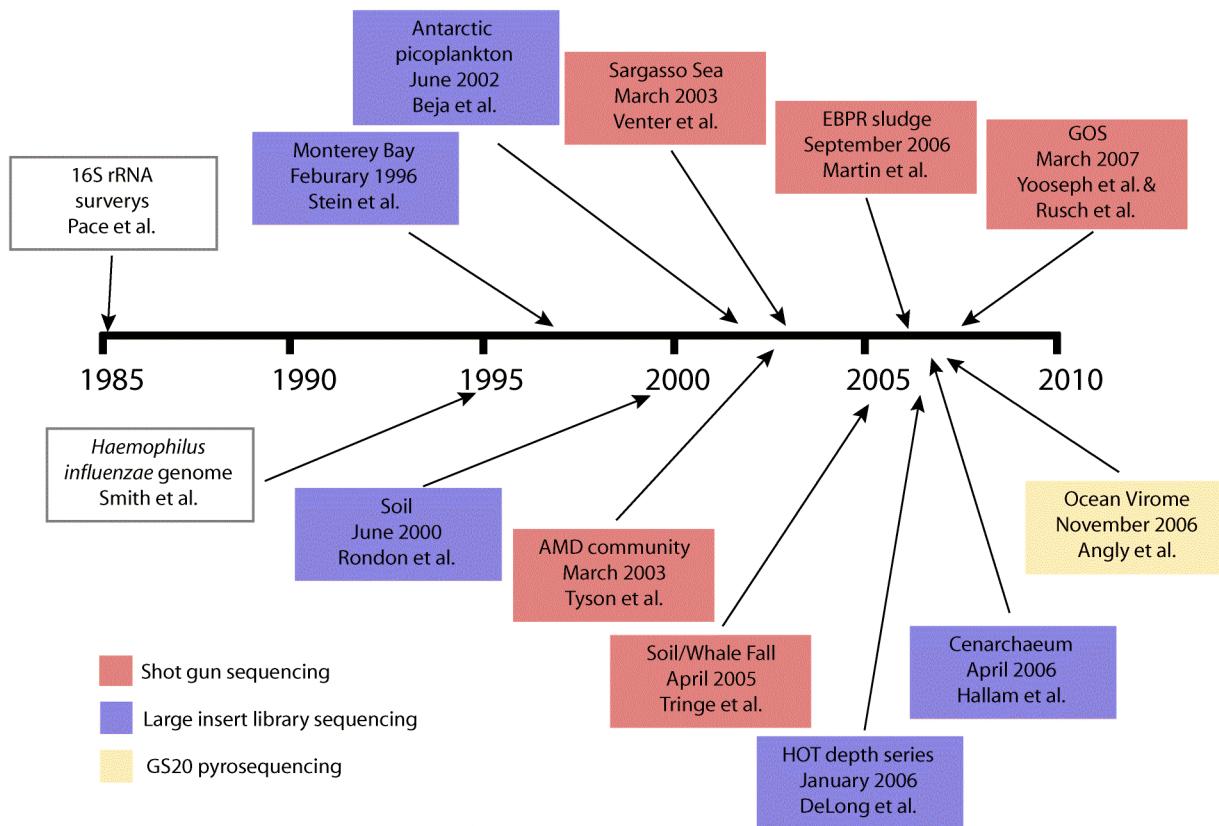


454 pyrosequencer



Sanger	GS20	FLX	XLR
0.7 Mbp	35 Mbp	100 Mbp	400 Mbp per run
700 bp	100 bp	200 bp	400 bp reads
0.1 ¢	0.03 ¢	0.01¢	0.003 ¢ per base

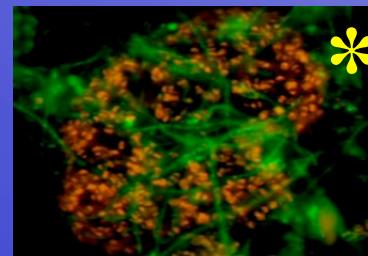
Metagenomics projects



23 completed and 130 ongoing metagenomic projects

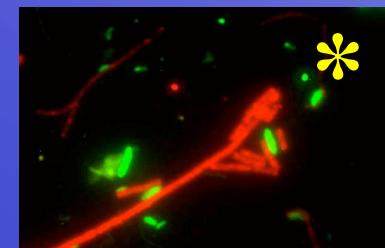
JGI metagenomic projects

2005

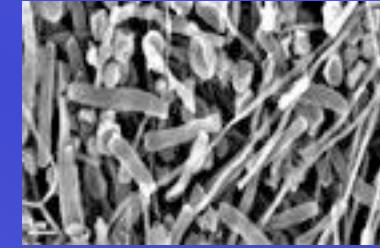


Gutless worm (MPI) planktonic archaea (MIT) EBPR sludge (UW/UQ) groundwater (ORNL)

2006

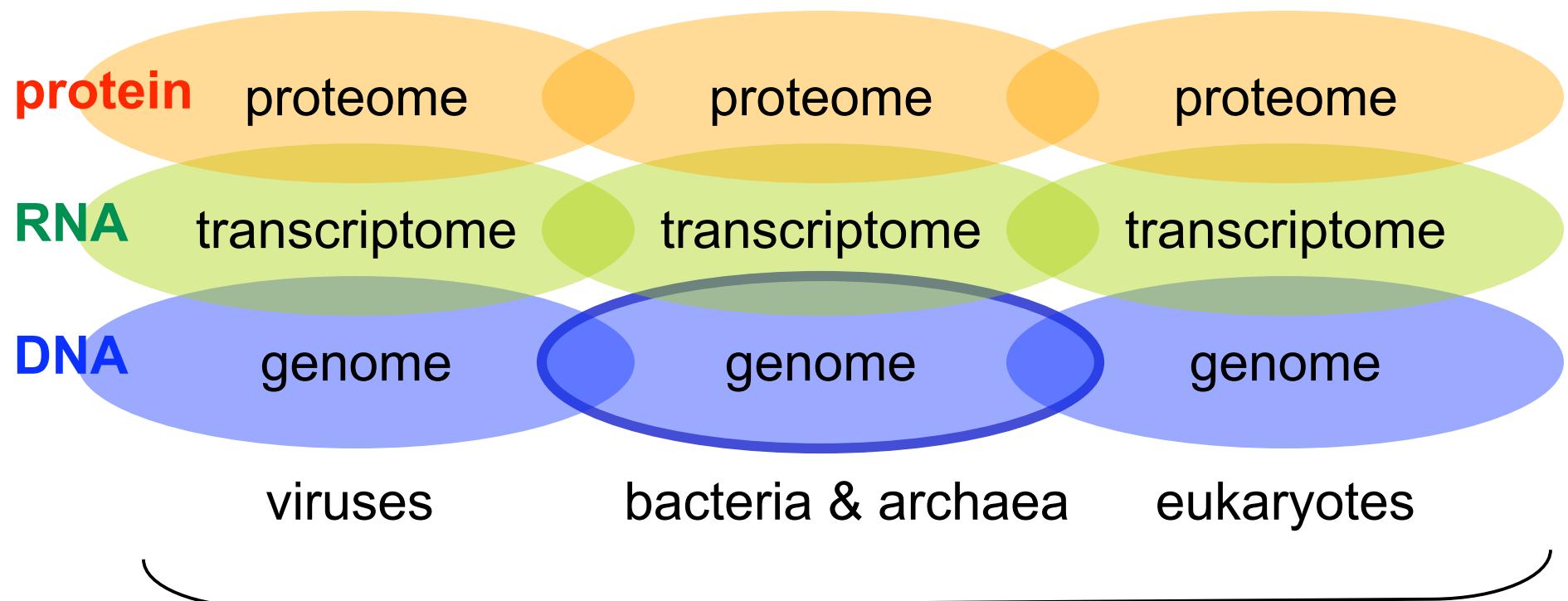


AMD nanoarchaea (UCB) Alaskan soil (UW) termite hindgut (CalTech) TA-degrading bioreactor (NUS)



Antarctic bacterioplankton (DRI) hypersaline mats (UCol) soil archaea (UW) Korarchaeota enrichment (Diversa)

Metagenomics is just the first level



microbial communities

Bioinformatic tools for metagenomic data analysis

MEGAN

- blast-based tool for exploring taxonomic content

MG-RAST (SEED, FIG)

- rapid annotation of metagenomic data, phylogenetic classification and metabolic reconstruction

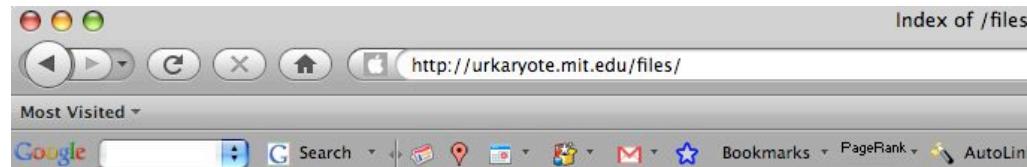
CAMERA (JCVI, Calit2, UCSD)

- metagenomic data repository and blast server

Getting started

<http://urkaryote.mit.edu/files/>

Download MEGAN and metagenomic datasets BLAST files:



The screenshot shows a Mac OS X desktop with a web browser window open to "http://urkaryote.mit.edu/files/". The title bar says "Index of /files". The toolbar includes standard icons for back, forward, search, and bookmarks. Below the toolbar is a menu bar with "Most Visited", "Google", "Search", "Bookmarks", "PageRank", and "AutoLink". The main content area displays a table of files and directories:

Name	Last modified	Size	Description
Parent Directory	04-Aug-2006 10:57	-	
41B09 PR.fna	25-Jun-2008 14:09	1k	
MEGAN_macos_2beta11.dmg	13-Jun-2008 14:21	15.5M	
MEGANmanual.pdf	13-Jun-2008 14:21	815k	
OMZ 20m allDNA.fna.5..>	15-Jun-2008 23:29	30.1M	
OMZ 20m allDNA.fna.5..>	24-Jun-2008 17:57	31.8M	
OMZ 45m allDNA.fna.5..>	15-Jun-2008 23:27	30.0M	
OMZ 60m allDNA.fna.5..>	15-Jun-2008 23:58	34.7M	

At the bottom of the page, the text "Apache/1.3.41 Server at urkaryote.mit.edu Port 16080" is visible.