Introduction to Metagenomics

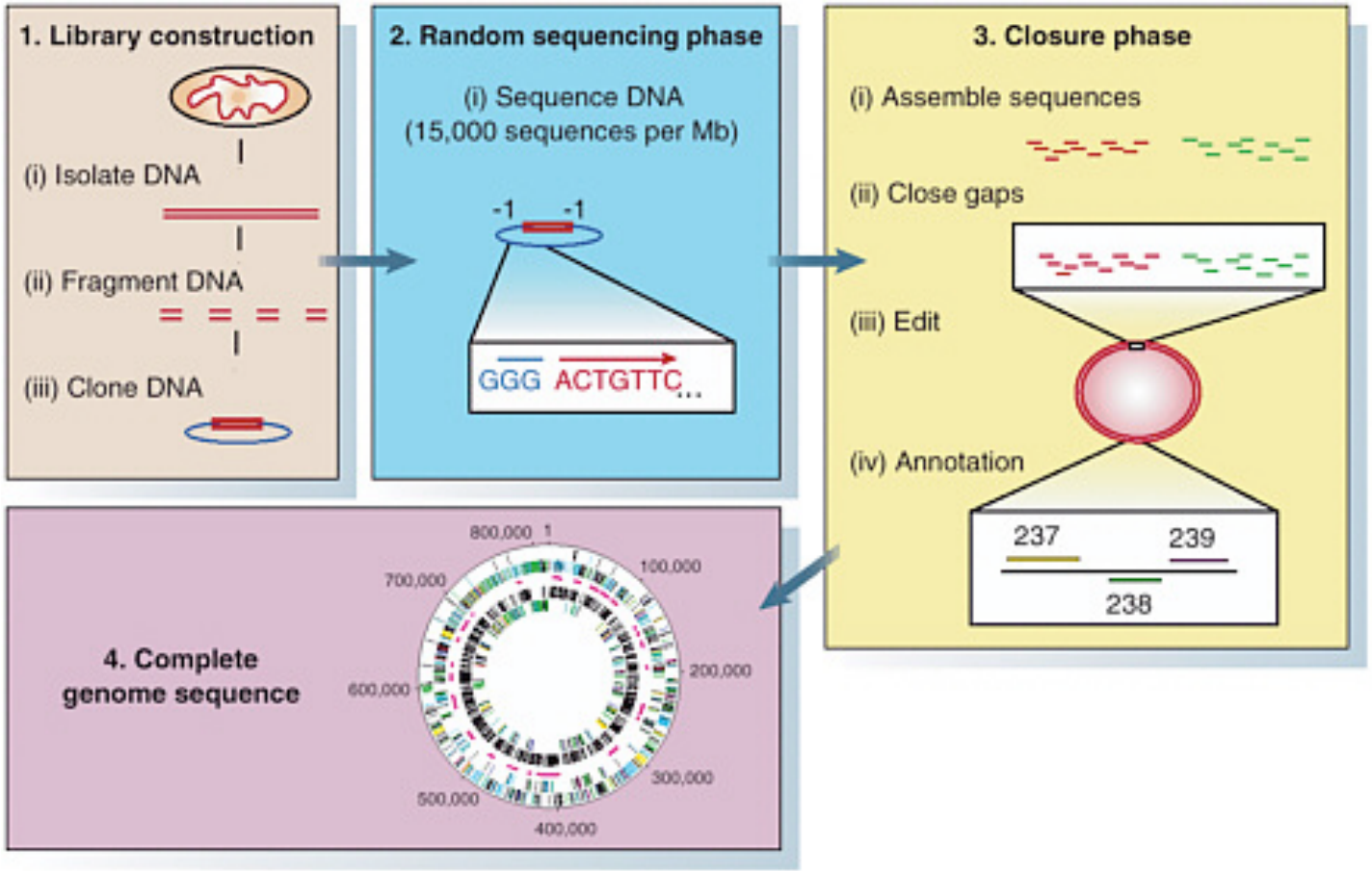
Yuzhen Ye (yye@indiana.edu)

I609 Bioinformatics Seminar I (Spring 2010)

School of Informatics and Computing

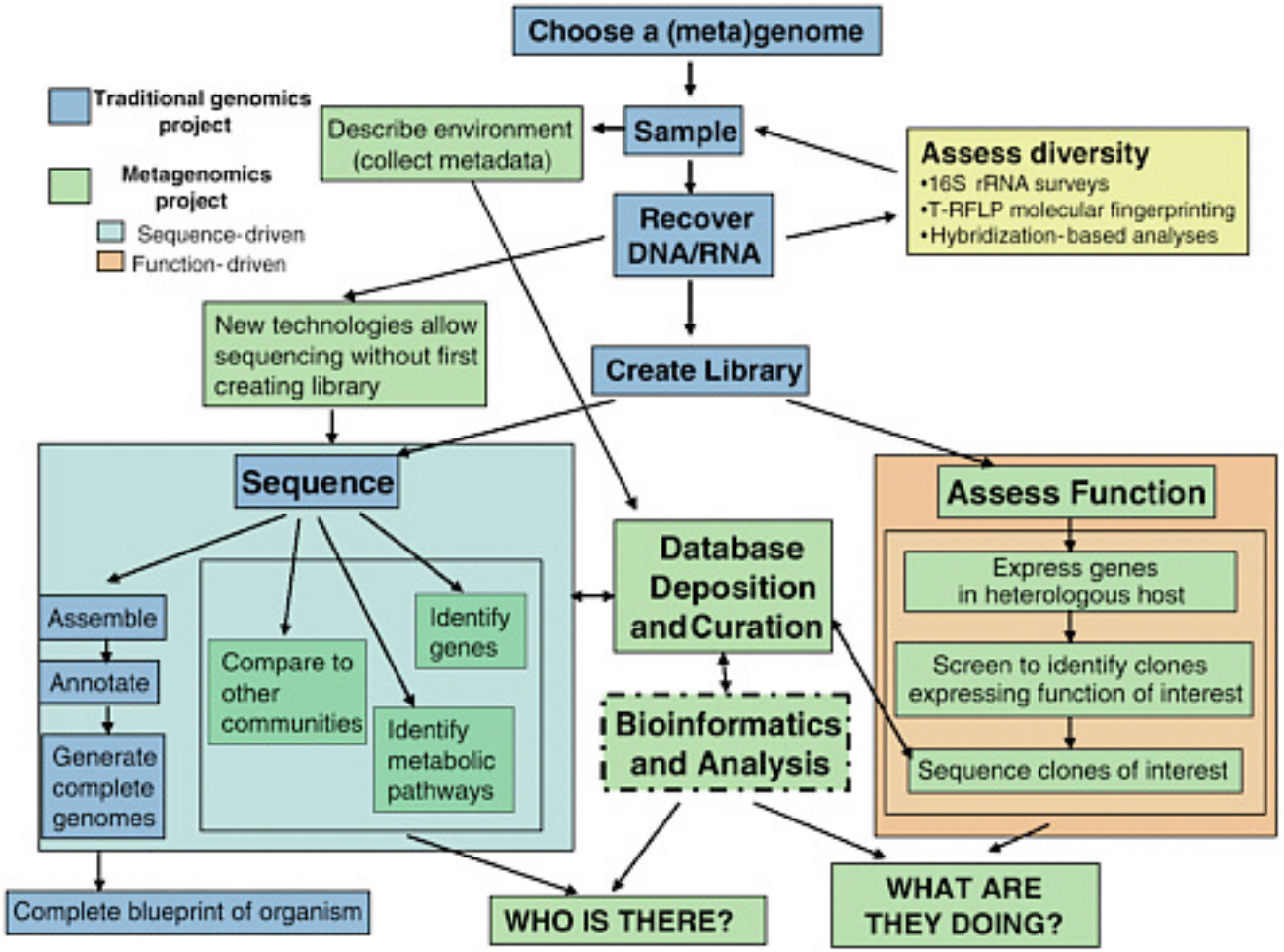
Indiana University

Traditional Microbial Genome Project



The new science of metagenomics: Figure 4-2

Metagenomics Differs in Many Ways



The new science of metagenomics: Figure 4-1

Why there is a boom of metagenomics?

- Most microbes cannot be cultured in the laboratory (nonculturable)
  - But we can still see them under microscope and can retrieve their DNA
- New sequencing techniques provide high-throughput and cheap sequences!

Metagenomics Step by Step (I)

- Habitat selection
  - The choice of the microbial community to study will be driven by the underlying scientific question being asked
  - The more information one has about the habitat -- physical, chemical, and ecological-- the more insight can be derived from the metagenomic data
  - The discovery of the ***keystone species*** (a community member whose significance to the community is larger than its relative abundance) replies on knowledge of the site

Metagenomics Step by Step (II)

- Sampling Strategy
  - Type, size, scale, number
    - The samples must be representative to the habitats
  - Timing
    - community's response to changing condition
    - Critical to understanding community structure, function, and robustness

Metagenomics Step by Step (III)

- Macromolecule recovery
  - The quality and completeness of data obtained from metagenomic analysis of any community will be only as good as the procedures used for the ***extraction of DNA*** from a sample
  - Cells from different species differ in their susceptibility to lysis under various conditions; even members of the same species may differ in their susceptibility to lysis in ***different physiological states;*** DNA from some members may be degraded
  - DNA from dead cells (may be important in drawing conclusions about the overall metabolic capabilities of a microbial community)

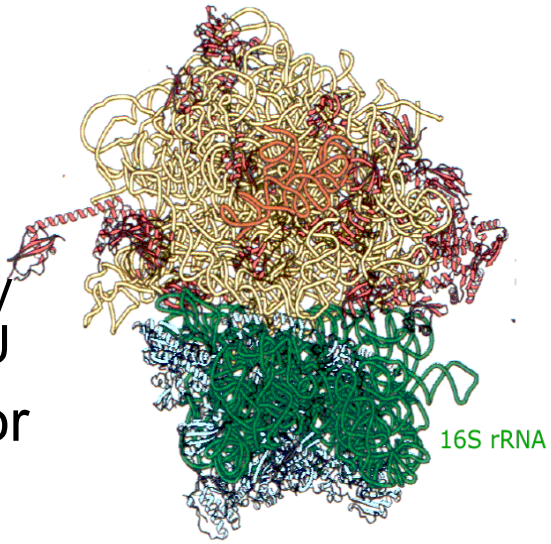
Types of Metagenomics Studies

- 16S rRNA-based surveys
  - 16S rRNA sequencing
  - 16S rRNA microarrays
  - PhyloChip: the PhyloChip detects on average twice as many taxa as 16S rRNA gene sequencing
  - Reveal microbial diversity and abundance
- Shotgun metagenomic sequencing
  - Reveal gene content of a community and its metabolic potential
- Targeted metagenomics -- Function-driven metagenomics analysis
  - Transcriptomics
  - Proteomics
  - Cell sorting

Get the most out of metagenomic projects!

1. 16S rRNA-Based Surveys

- Most commonly used molecular marker
  - essential function
  - Ubiquity
  - evolutionary properties
- OTU (operational taxonomic units) definition based on 16S rRNA gene
  - organisms displaying 97 to 98% identity in this gene to be part of the same OTU
- Rapid and cost-effective approaches for assessing bacterial diversity and abundance.
- Often serves as a first step in larger metagenomics projects to evaluate bacterial diversity in potential samples of interest



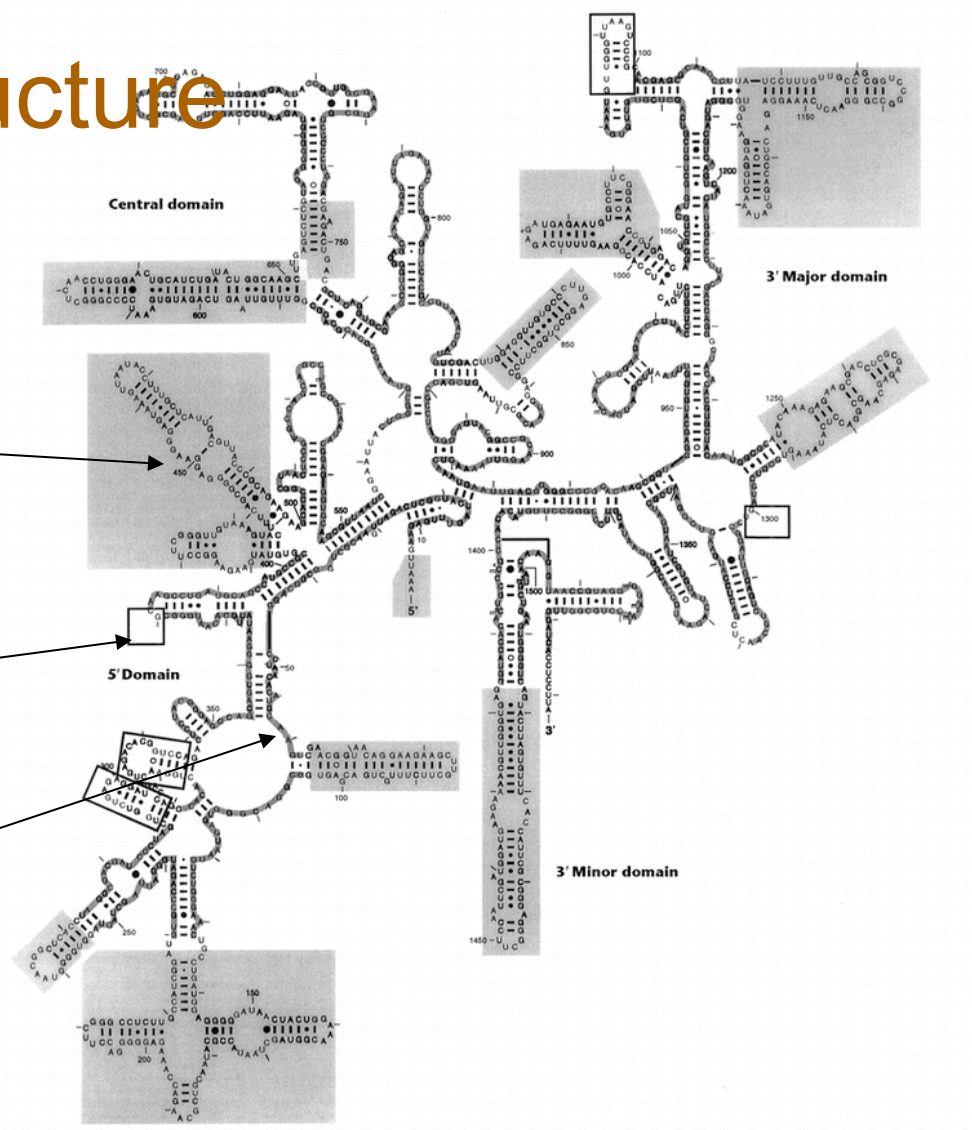
Bacterial ribosome (70S)

16s rRNA Structure

Vary in size from 50 nucleotides

Vary in size from 10 to 50 nucleotides

Universal core secondary structures

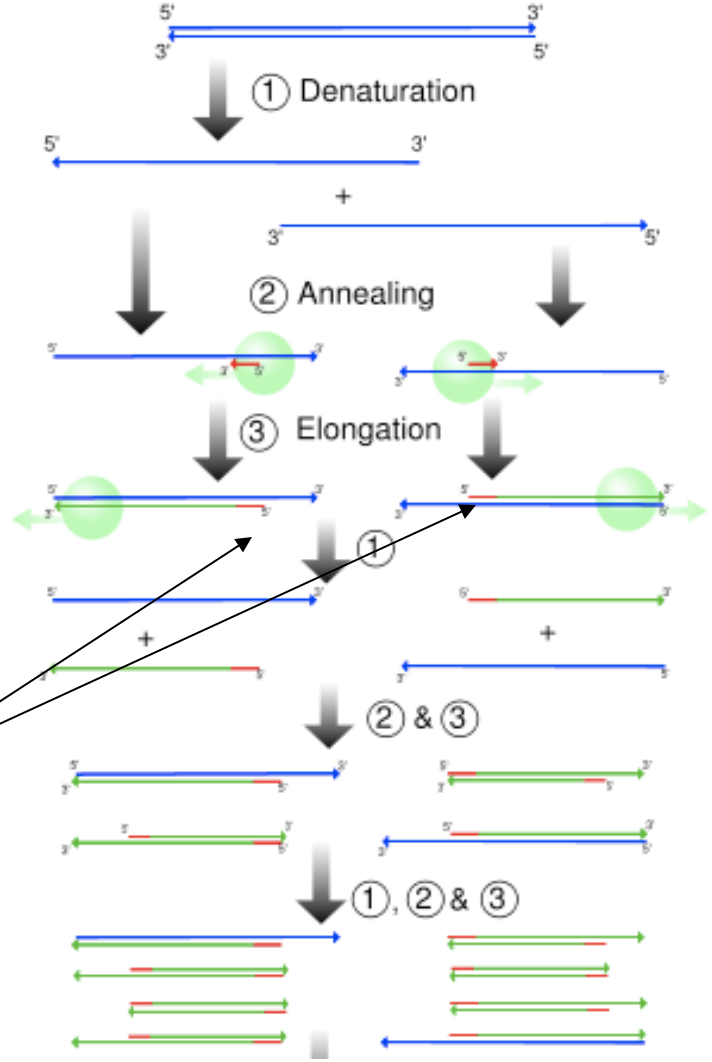


Design primers with different specificities

PCR (Polymerase Chain Reaction)

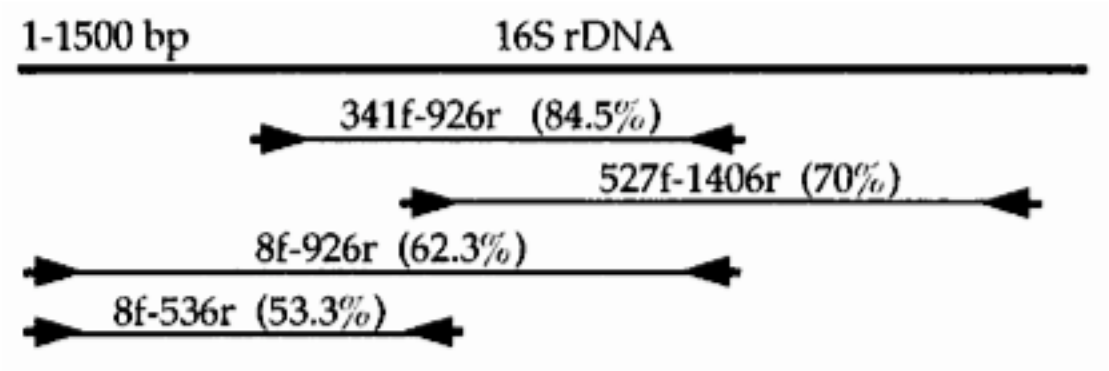
- Kary Mullis, 1983
- A technique widely used in molecular biology
- DNA polymerase; template; primers
- Selective amplification of a specific region of DNA

A pair of primers define the DNA region to be amplified



16s rRNA Primers

Primer	Specificity	Sequence (5′→3′)	Reference(s)
8f	Domain	AGAGTTTGATCCTGGCTCAG	1
341f	Domain	CCTACGGGAGGCAGCAG	21
536r	Universal	CAGCMGCCGCGGTAATWC	1, 11
527f	Universal	ACCGCGGCCKGCTGGC	This study
926r	Domain	CCGTCAATTCCTTTRAGTTT	20
1406r	Universal	ACGGGCGGTGTGTRC	1



(APPLIED AND ENVIRONMENTAL MICROBIOLOGY, Nov. 1997, p. 4516–4522)

16S rRNA Gene Sequencing

- PCR amplification with primers that ***hybridize to highly conserved regions*** in bacterial or archaeal 16S rRNA, followed by cloning and sequencing
- Phylogenetic analysis of 16S rRNA helps to reveal the species diversity in a community

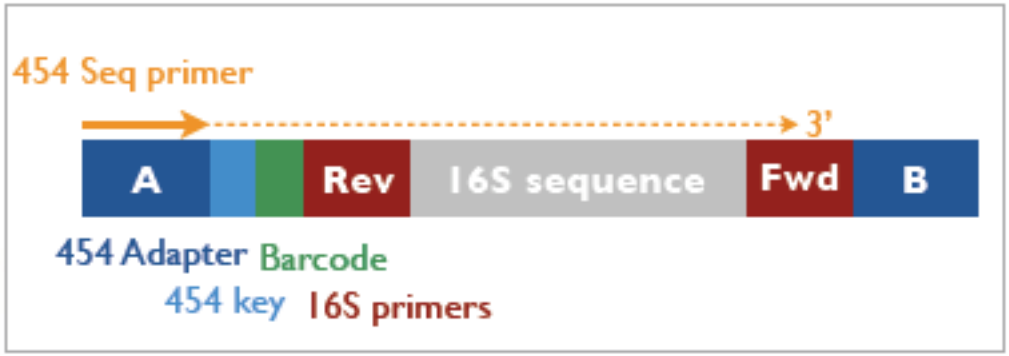
16S rRNA Sequencing Limitations

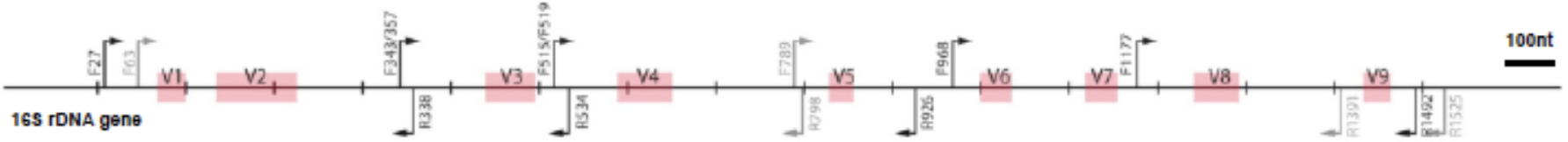
- Sampling challenges
  - Rich species vs sparse species
- Do not tell much about the functional abilities of a community
- PCR bias -- not all rRNA genes amplify equally well with the same “universal” primers
  - Discrepancies between 16S rRNA gene surveys and those derived from whole-genome shotgun data
- Multiple copies of 16S rRNA gene in some species (which may artificially lead to the overrepresentation of some species)
  - ***Single-copy phylogenetic markers***, recA and rpoB ??

Barcoded pyrosequencing

- “We constructed error-correcting DNA barcodes that allow one run of a massively parallel pyrosequencer to process up to 1,544 samples simultaneously. Using these barcodes we processed bacterial 16S rRNA gene sequences representing microbial communities in 286 environmental samples, corrected 92% of sample assignment errors, and thus characterized nearly as many 16S rRNA genes as have been sequenced to date by Sanger sequencing.” Nature Methods - 5, 235 - 237 (2008)

Barcoded pyrosequencing





Which regions? V1-V3, V3-V5

RecA as an Alternative Phylogenetic Marker

- recA is a multifunctional protein contributing to homologous recombination, DNA repair, etc
  - Housekeeping gene
- Used first to analyze V. cholerae strains
- Used to discriminate closely related species within the family Vibrionaceae
  - A correlation of 0·58 between recA and 16S rDNA pairwise similarities.
  - Strains of the same species have at least 94 % recA sequence similarity.
  - recA gene sequences are much more discriminatory than 16S rDNA. For 16S rDNA similarity values above 98 % there was a wide range of recA similarities, from 83 to 99 %

RpoB as an Alternative Phylogenetic Marker

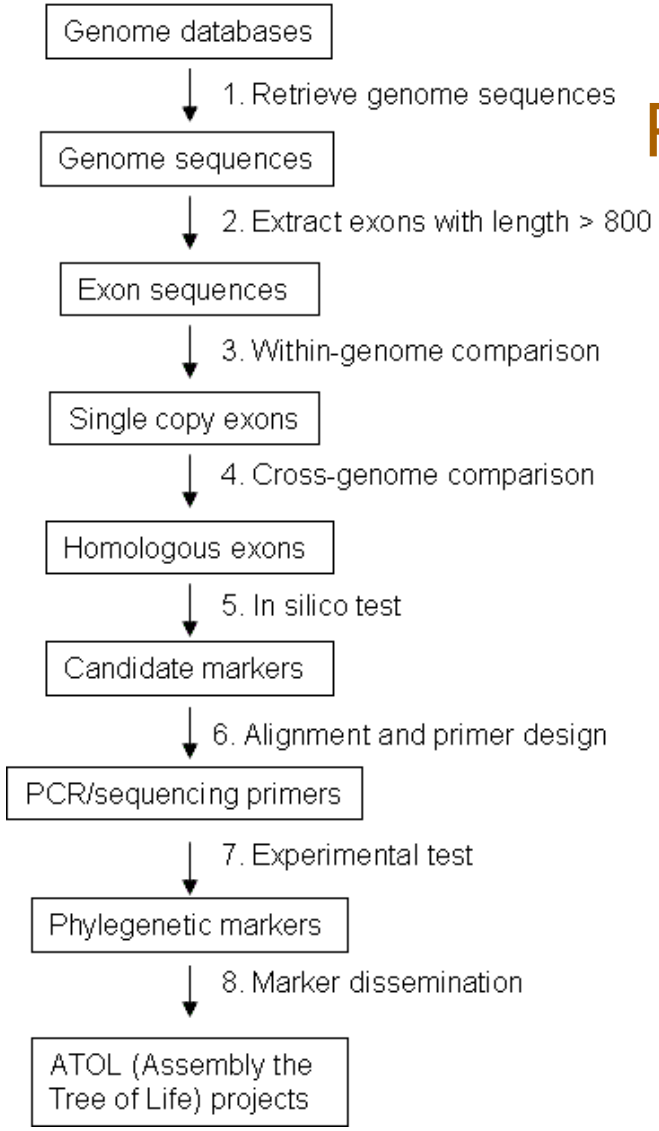
- rpoB: RNA polymerase ß subunit gene
- rpoB provided comparable phylogenetic resolution to that of the 16S rRNA gene at all taxonomic levels, except between closely related organisms (species and subspecies levels), for which it provided better resolution.

Applied and Environmental Microbiology, January 2007, p. 278-288, Vol. 73, No. 1

Comparison of the Trees

Applied and Environmental Microbiology, January 2007, p. 278-288, Vol. 73, No. 1

Bioinformatics Tools for New Phylogenetic Marker Discovery



Phylomarker candidates:

relatively well conserved, putatively single-copy gene fragments with long, uninterrupted exons

(BMC Evol Biol. 2007; 7: 44)

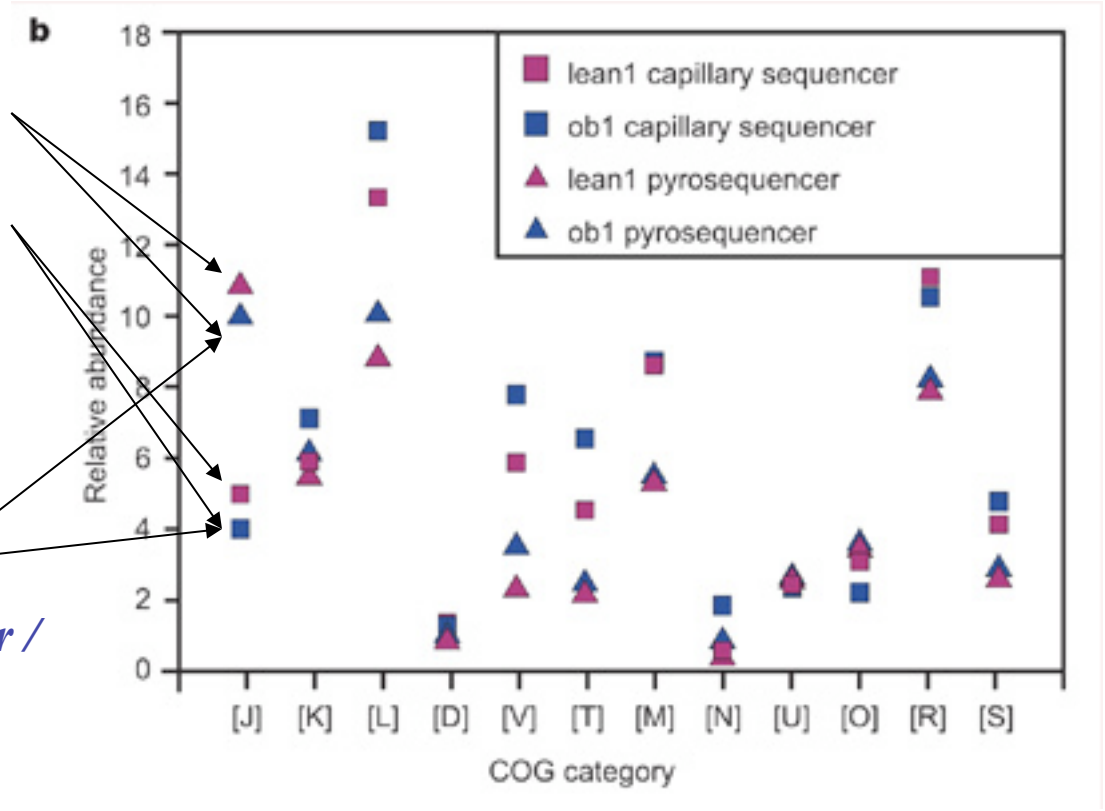
2. Shotgun metagenomic sequencing

- How the sequencing was done?
- It matters in many ways
  - Computational analysis
  - And maybe the results may be different (?)
- What's the cons and pros
  - Capillary sequencing:
    - More confident gene calling
    - Longer reads
    - Affected by cloing bias
  - Pyrosequencing
    - Higher sequence coverage
    - No cloing bias
    - Shorter reads
  - Illumina

Sequencer Caused the Difference?

Obese / lean

Capillary sequencer / pyrosequencer



"An obesity-associated gut microbiome with increased capacity for energy harvest" Nature 444, 1027-131(21 December 2006)

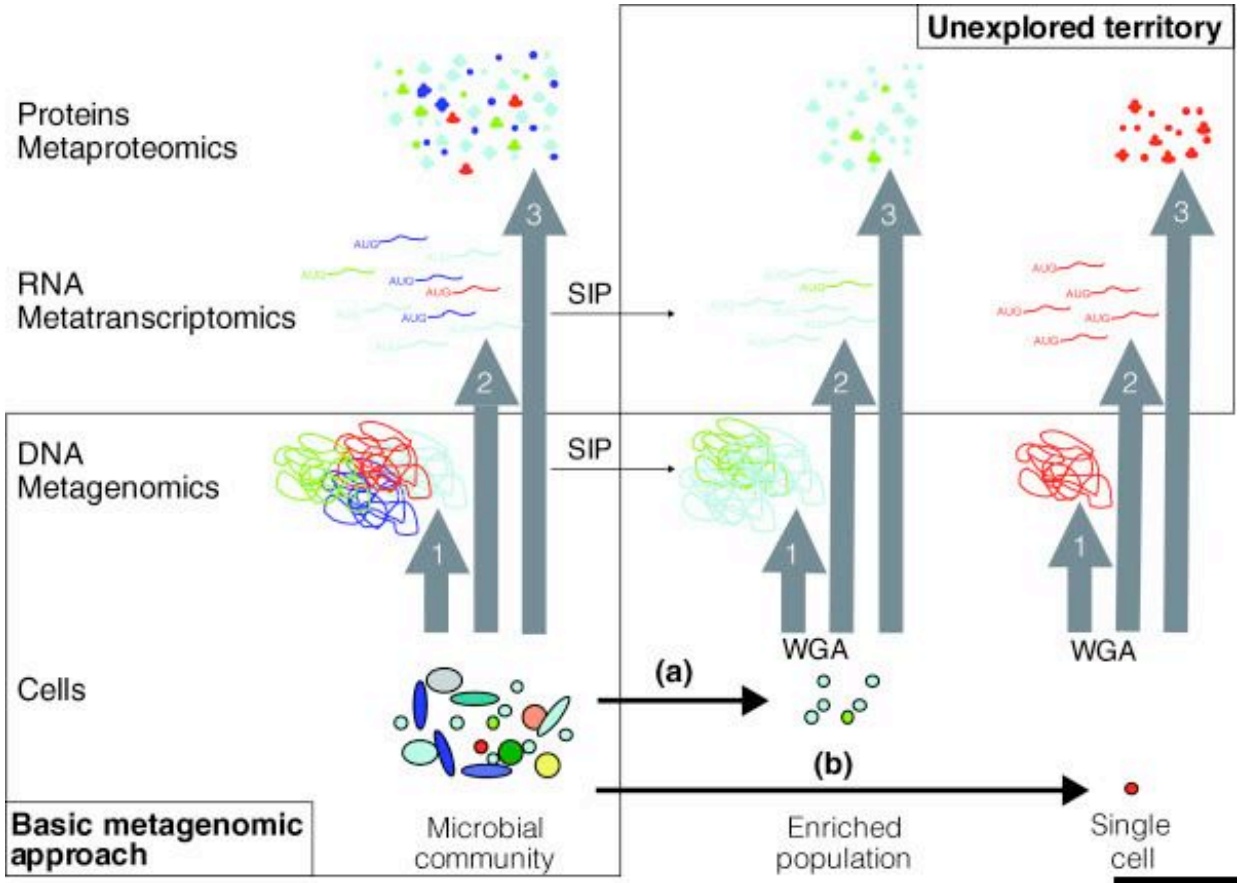
Genome-centric vs Gene-centric

- Genome-centric analyses
  - Requires to assemble whole genomes -- very difficult
- Gene-centric analyses
  - Environmental gene tags (EGTs): short DNA sequences that contain fragments of functional genes
  - EGTs “fingerprints” can be compared across multiple sites or habitats or over time in the same environment
  - Overrepresented or underrepresented EGTs can provide insights into unique metabolic capabilities associated with a particular environment even if it is not possible to assign a particular EGT to a particular environment.

Sequencing is Just One Kind of Metagenomics

- Sequencing provides information that is limited by what's in the databases and by the available algorithms for linking sequence to function
  - low resolution
  - the inability to classify short metagenomic fragments
  - the lack of functional verification
- To address "who is doing what" -- functional metagenomics
  - Screen the metagenomic libraries directly for expressed functions
  - Gene-expression systems: *E.coli; Streptomycetes; Bacillus subtilis*
  - Single-Cell Analyses
  - Culturing uncultured species

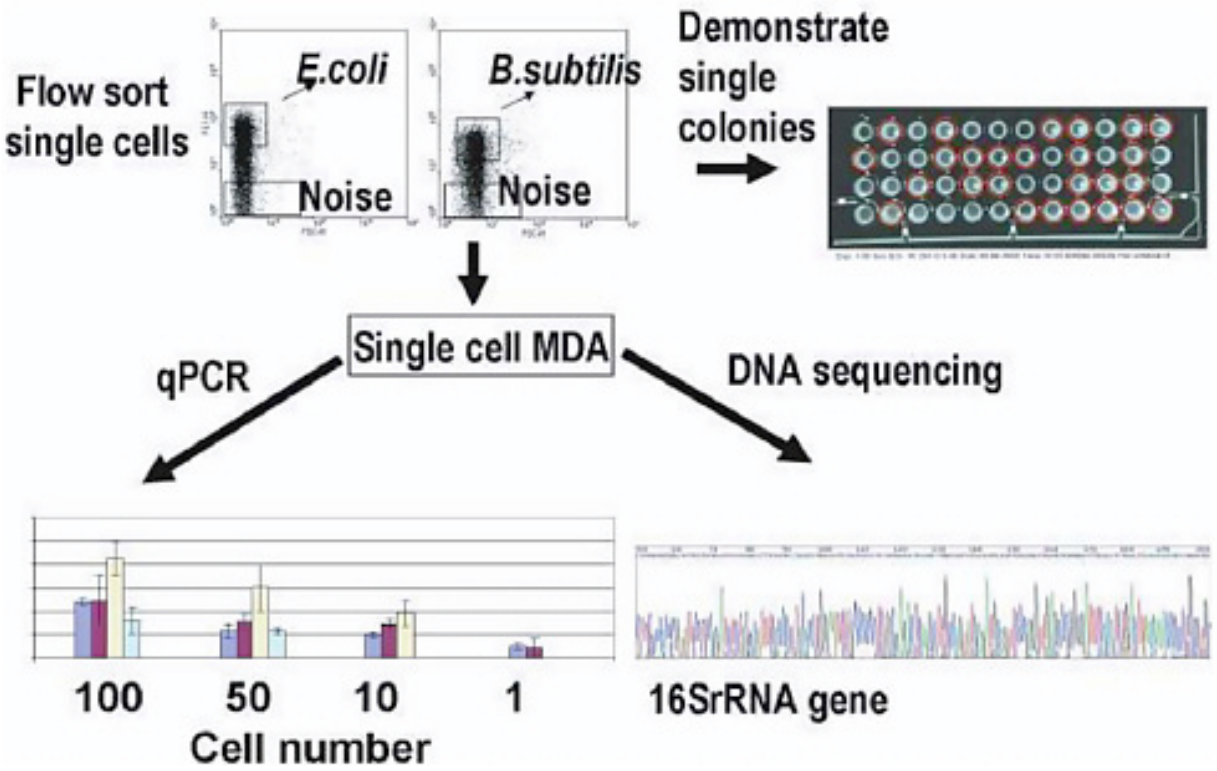
Enhancing the Basic Metagenomic Approach Through Complementary Technologies



Genome Biology 2007, 8:231

Single-Cell Analyses

"The ability to isolate single cells with microfluidics coupled with technology to amplify genomic DNA from single cells will revolutionize the study of unculturable species and the microheterogeneity within species.



The new science of metagenomics: Figure 4-4
See a review "Microfabrication meets microbiology" Nature Reviews Microbiology 5, 209-218, 2007

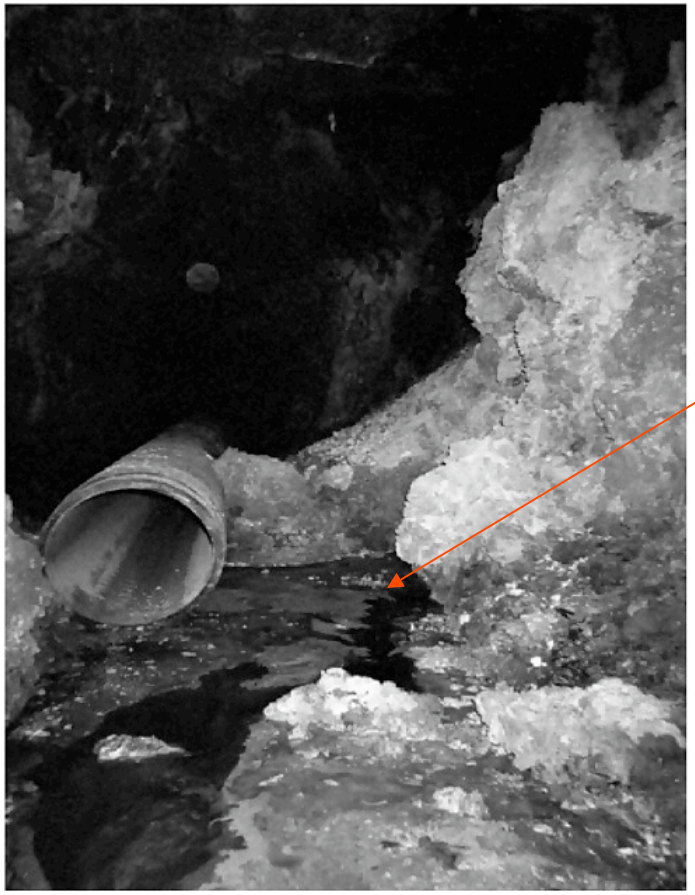
3. Targeted metagenomics

- Targeted metagenomics aims at acquiring sequence with specific protein functions, such as glycoside hydrolases, and bile salt hydrolase.
- Smaller scale, but more specific

Metagenomic Projects

The large-scale application of random shotgun sequencing to DNA extracted directly from environmental samples and resulting in at least 50 megabase pairs (Mbp) of sequence data
(Genome Biology 2007, 8:231)

The Acid Mine Drainage (AMD) Project



An acid mine drainage site

Biofilms growing on the surface of flowing AMD in the five-way region of the Richmond mine at Iron Mountain, California, were sampled in March 2000

Acid is produced by oxidation of sulfide minerals that are exposed to air as a result of mining activity

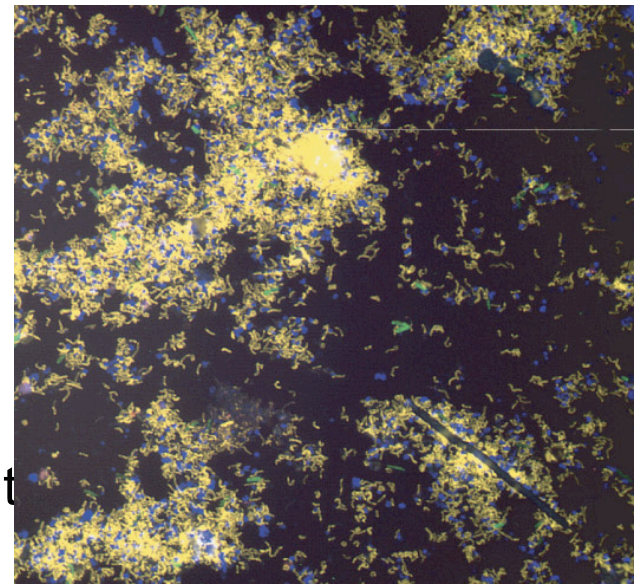
Why AMD Biofilm?

- Extreme acidic environment (self-contained ecosystem)
- Scientists are interested in the metabolic potential of such an environment: nitrogen fixation, sulfur oxidation, and iron oxidation
- To understand how the microbes tolerate the extremely acidic environment

- And it is a good pick -- low species complexity

Preliminaries

- Group-specific fluorescence in situ hybridization (FISH)
  - Results indicated the presence of mixtures of bacteria (*Leptospirillum*, *Sulfobacillus* and, in a few cases, *Acidimicrobium*) and archaea (*Ferroplasma* and other members of t *Thermoplasmatales*)

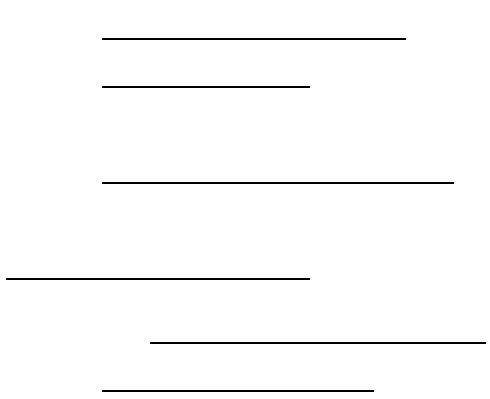


Yellow: *Leptospirillum* cells

- 16S ribosomal RNA gene clone library
  - 384 clones were end-sequenced
  - Results indicated the presence of three bacterial and three archaeal lineages. The most abundant clones are close relatives of *L. ferriphilum* and belong to *Leptospirillum* group II

DNA Sequencing

- A small insert plasmid library (average insert size 3.2 kb)
- Shotgun sequencing resulted 72.6 million bp; averaging 737 bp per read.



Reads could from ***different individuals***, ***different strains*** of the same species, and ***different species***

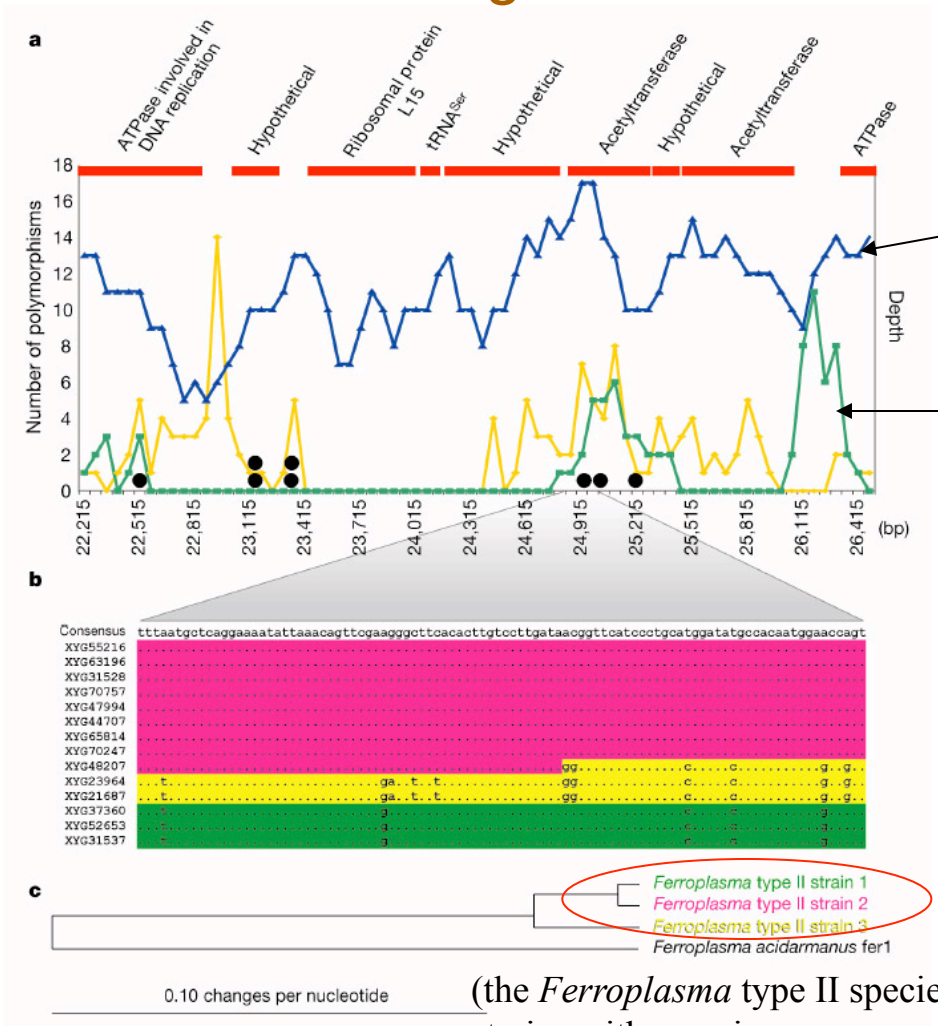
Whole-genome Assembly (for Metagenomic Data)

- Assembled using JAZZ (whole-genome shotgun assembler)
- Over 85% of the reads were assembled into 1,183 scaffolds longer than 2 kb; with 92.7% of end pairs from the same clone assembled with the appropriate orientation and separation.
- Assignment of scaffolds to organism type: separate scaffolds by ***G+C content*** (low G+C content, and G+C content bins); then subdivide them using ***read depth*** (coverage)

Species Discovery

- The high G+C scaffolds at ~10x coverage (70 in total) were identified by a single 16S rRNA gene as belonging to the genome of a *Leptospirillum* group II species
- The low G+C scaffolds (59 in total) at ~10x coverage represent a nearly complete genome of a ***previously unknown, uncultured Ferroplasma species distinct from Ferroplasma acidarmanus fer1***; designated as *Ferroplasma* type II
  - The single ***16S rRNA*** gene identified in these scaffolds is 99% identical to that of fer1
  - Alignment of the scaffolds to fer1 genome ***reveals 22% divergence*** at the nucleotide level
  - The ***total length*** of the scaffolds is similar to fer1 genome, the ***local gene order and content*** are highly conserved

An Genomic Segment of the *Ferroplasma* type II



coverage

polymorphisms

Reads are from different individuals, and may even from different strains

(the *Ferroplasma* type II species population seemed to be dominated by strains with mosaic genomes constructed by recombination of three closely related but distinct genome types -- species)

Highlights: Genome Assembly

- Reconstruction of ***near-complete*** genomes of *Leptospirillum* group II and *Ferroplasma* type II, and partial recovery of three other genomes.
  - Whole genome assembly sort of worked! (because of its low species diversity)
  - Binning procedures to assign contigs to different genomes on the basis of base composition, codon usage, etc.

Highlights: Keystone Species Discovery

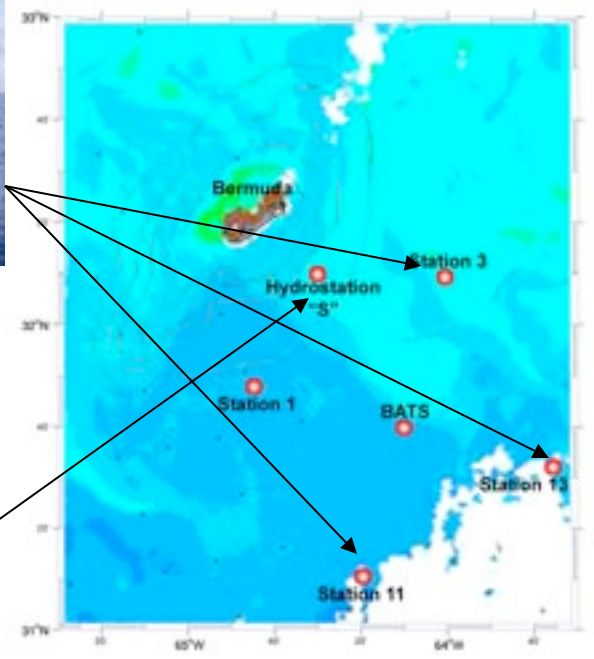
- Discovery of ***keystone species*** (a community member whose significance to the community is larger than its relative abundance)
  - the *Leptospirillum* group II (the dominant species) genome does not contain any genes for nitrogen fixation
  - One of its least abundant members of the community, *Leptospirillum* group III, has the nitrogen fixation operon with homology to that reported for a *L. ferrooxidans* strain
  - Metagenomic information guided the later cultivation of this species in pure form (by providing N_2 as the only nitrogen resource)

The Sargasso Sea Project

- Science 2004, 304:66-74
- A pilot project of Venter's global ocean voyage (with the ultimate goal of finding solutions to energy problems)
- Venter is not the first to sequence the genes of microbes from the ocean; but he is the first to do it on a large and ambitious scale
- Sargasso Sea takes its name from the sargassum seaweed that floats on its surface
- It is in the middle of the North Atlantic Ocean near Bermuda.

Where / How to Get the Water Samples

RV Weatherbird II



Sorcerer II

Venter's 95-foot sailboat converted into a research vessel equipped for an 18-month worldwide scientific expedition

Large-scale DNA Sequencing

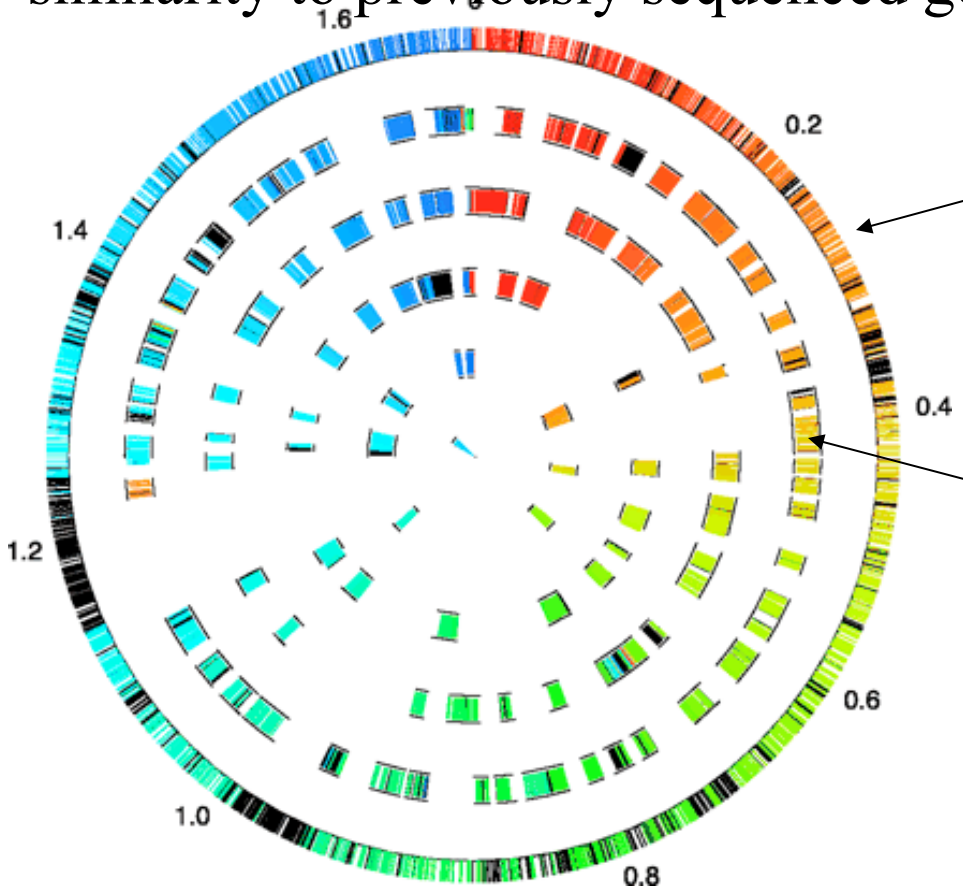
- Genomic libraries with insert sizes ranging from 2 to 6 kb; plasmid clones were sequenced from both ends to provide paired-end reads (mate pairs)
- Whole-genome random shotgun sequencing
- Weatherbird II samples
  - 1.66 million reads
  - averaging 818 bp in length
  - ~1.36 Gbp of DNA sequence.
- Sorcerer II samples
  - 325,561 reads
  - ~265 Mbp of DNA sequence.

Genome and Large Assemblies

- Celera Assembler was used
- The 1.66 million sequences from the Weatherbird II samples were assembled to provide a single master assembly for comparative purposes.
  - 64,398 scaffolds ranging in size from 826 bp to 2.1 Mbp, containing 256 Mbp of unique sequence and spanning 400 Mbp.
  - The Sorcerer II samples provided almost no assembly (with only mini-scaffolds)

Scaffolds Binning

Based on: depth of coverage, oligonucleotide frequencies, and similarity to previously sequenced genomes



Prochlorococcus marinus MED4

environmental genome segments with ***conserved gene order***

Highlights: Discrete Species

- ***Discrete species*** versus a population continuum
  - The most deeply covered of the scaffolds (21 scaffolds with over 14x coverage and 9.35 Mb of sequence), contain just over ***1 single nucleotide polymorphism (SNP) per 10,000 base pairs***
- But really, is species discrete?
  - The discrete units are helpful organizing tools but to truly understand biological diversity, the continuity should be considered

Highlights: Species Richness

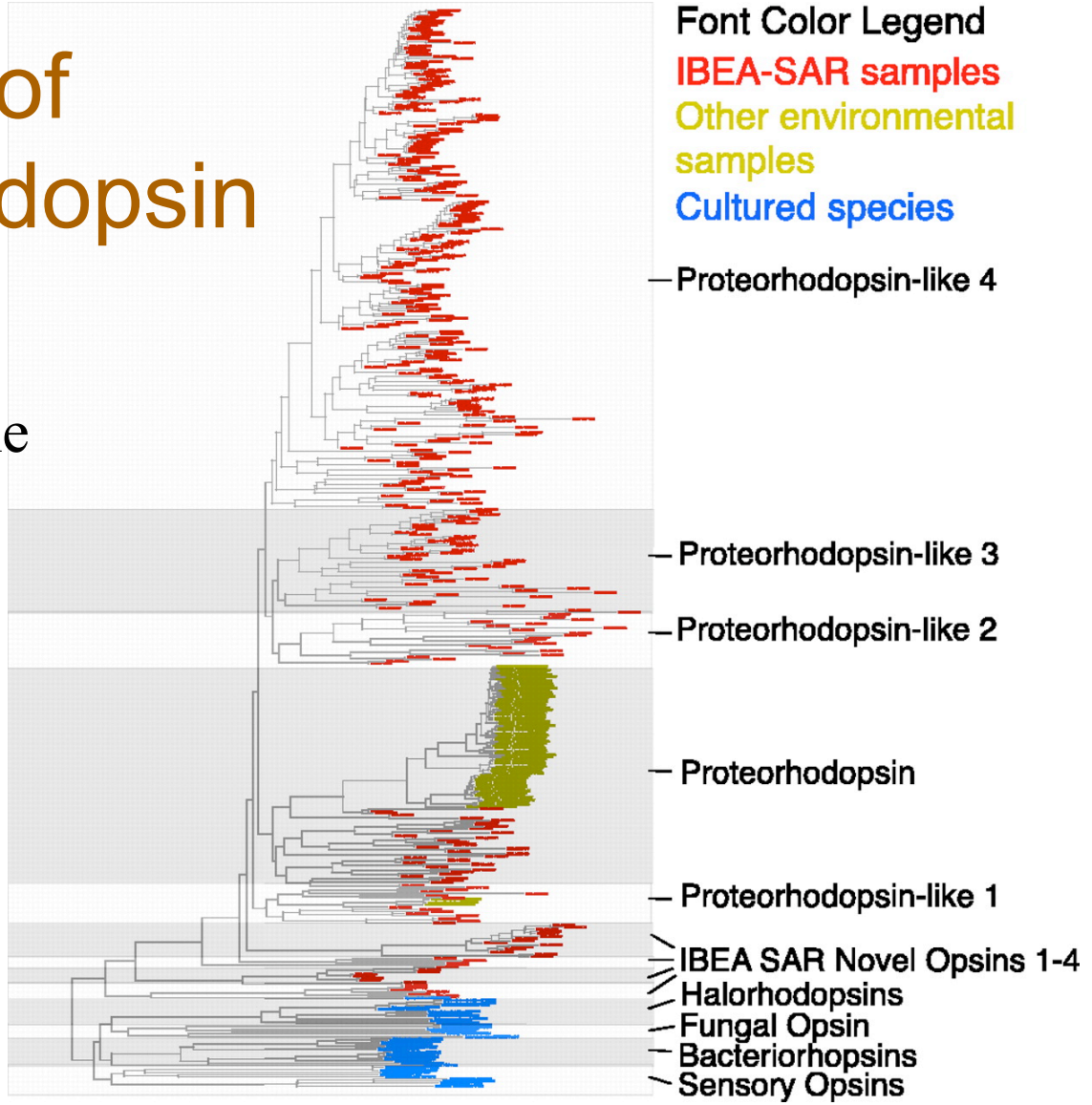
- Estimate there are ***at least 1800*** genomic species in the Sargasso Sea water based on sequence relatedness, including ***148 previously unknown*** bacterial phylotypes
- How these numbers are derived?

Highlights: Species Richness

- 16s rRNA-based survey
  - 1164 distinct small subunit rRNA genes in the Weatherbird II assemblies and 248 within the Sorcerer II reads.
  - 148 previously unknown phylotypes if using a 97% sequence similarity cutoff to distinguish unique phylotypes (against the RDP II database)
  - 643 if using a 99% similarity cutoff
- Alternative phylogenetic markers based
  - RecA/RadA, heat shock protein 70 (HSP70), elongation factor Tu (EF-Tu), and elongation factor G (EF-G)
  - Define "genomic" species as a clustering of assemblies or unassembled reads more than 94% identical on the nucleotide level (comparable to the 97% cutoff traditionally used for rRNA).
  - 451 (averaged over the six genes analyzed; range 341 to 569) -- this serves as the most conservative estimate of species richness.
- But random sampling may miss low abundance species -- how to get the real species diversity estimation
  - nonparametric methods for small sample corrections
  - parametric methods assuming a log-normal distribution of species abundance
  - fitting the observed depth of coverage to a theoretical model of assembly progress for a sample corresponding to a mixture of organisms at different abundances
  - ***at least 1800***

Tree of Proteorhodopsin

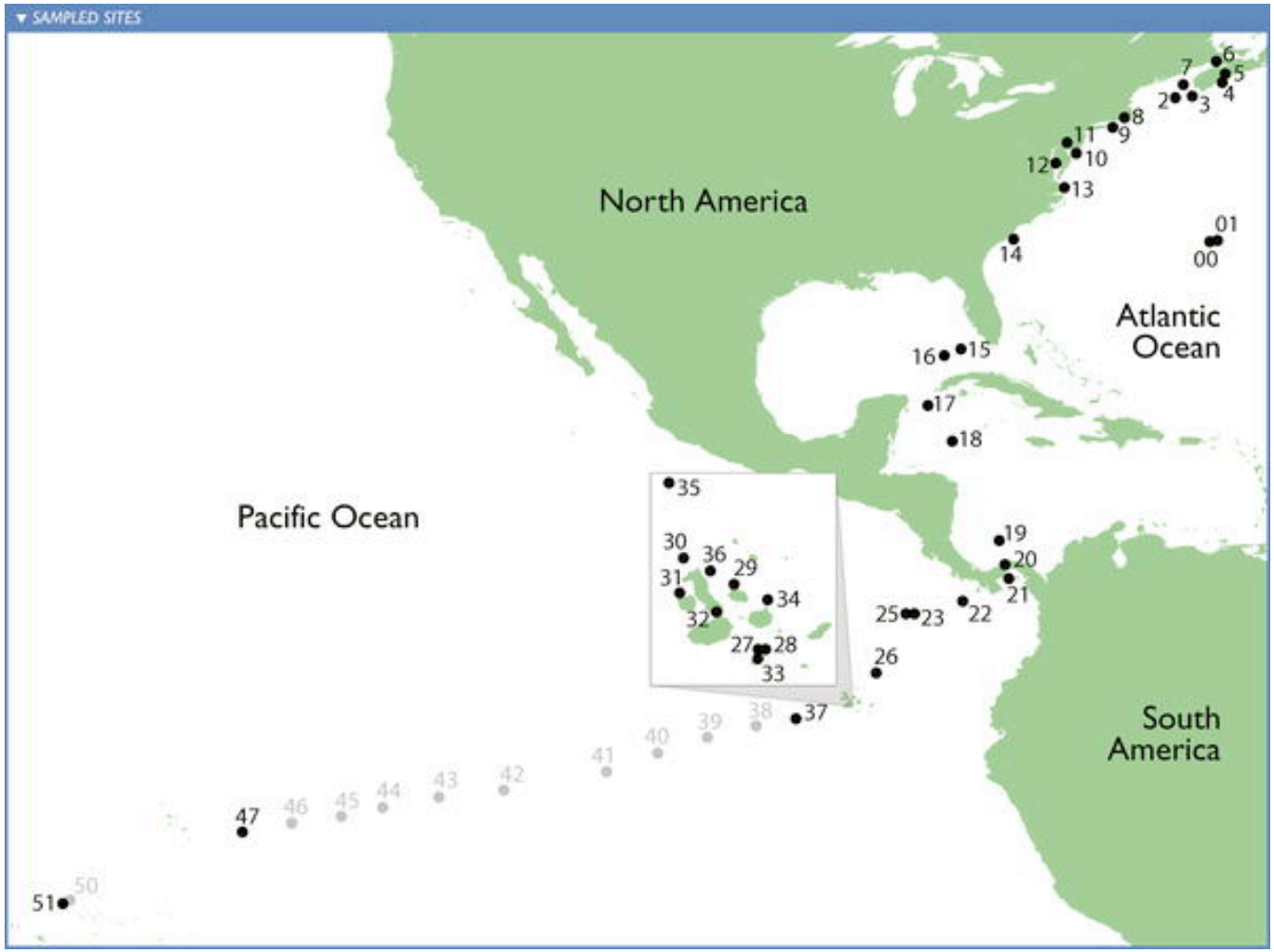
Problems (?)
The clustering of the proteorhodopsin proteins is ***mainly*** determined by the samples



The Sorcerer II Global Ocean Expedition

- PLoS Biol 2007, 5:e77
- Samples were collected from Northwest Atlantic through Eastern Tropical Pacific
- 7.7 million sequencing reads (6.3 billion bp)
  - Sargasso Sea: 1.36 Gbp + ~265 Mbp
  - GOS dataset

Sorcerer II Expedition Sampling Sites



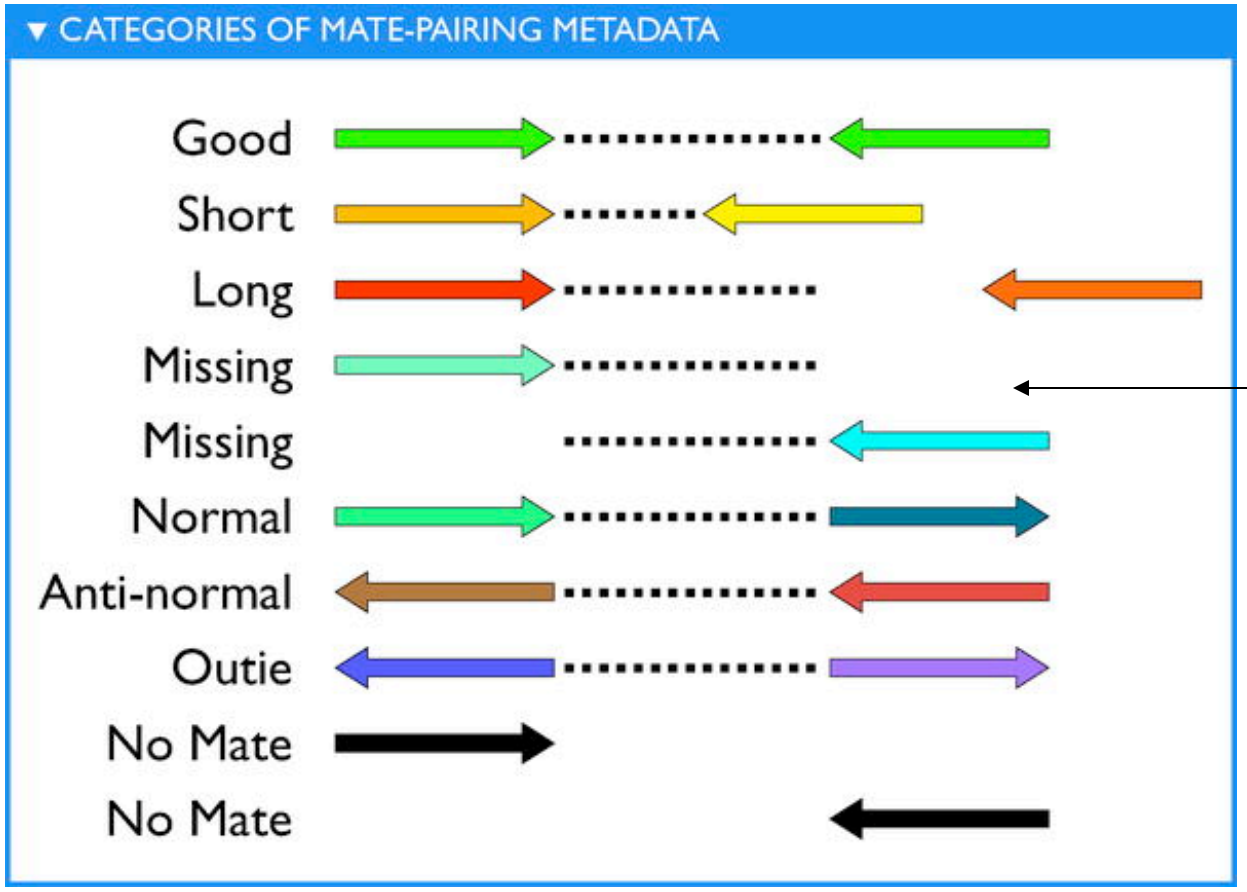
Assembly and Binning

- The primary assembly was ***strikingly fragmented***. Only 9% of sequencing reads went into scaffolds longer than 10 kbp. A majority (53%) of the sequencing reads remained unassembled singletons.
- Scaffolds longer than 50 kb totaled 20.7 Mbp; of these, >75% correspond to the *Burkholderia* or *Shewanella* assemblies
- ***"***These results highlight the unusual abundance of these two organisms in a single sample, ***which significantly affected our expectations regarding the current dataset."***
- New comparative genomic and assembly methods were developed
  - Fragment recruitment
  - Extreme assembly

Fragment Recruitment of Global Ocean Sampling Data to Finished Microbial Genome

- Direct comparison of the GOS sequencing data to the genomes of sequenced microbes by BLAST
- The amount and distribution of reads ***recruited*** to any given genome provides an indication of the abundance of closely related organisms

Fragment Recruitment and Genomic Structural Variation



"Missing" mates identify breaks in synteny between the environmental data and the reference sequence

Extreme Assembly

- Use an approach that aggressively resolves conflicts, which are known to disrupt whole-genome assemblers
- The more aggressive assemblies demonstrably suffered from higher rates of assembly artifacts, including chimerism and false consensus sequences

Agricultural Soil & “Whale fall” Carcasses

- Nutritional rich environment
  - But with very different nutrient sources (plant material for soil and lipid-rich bone for deep-sea whale fall samples).
- Comparative metagenomics
  - using largely unassembled sequence data
  - identification of environment-specific genes through a gene-centric comparative analysis (quantitative gene content analysis)

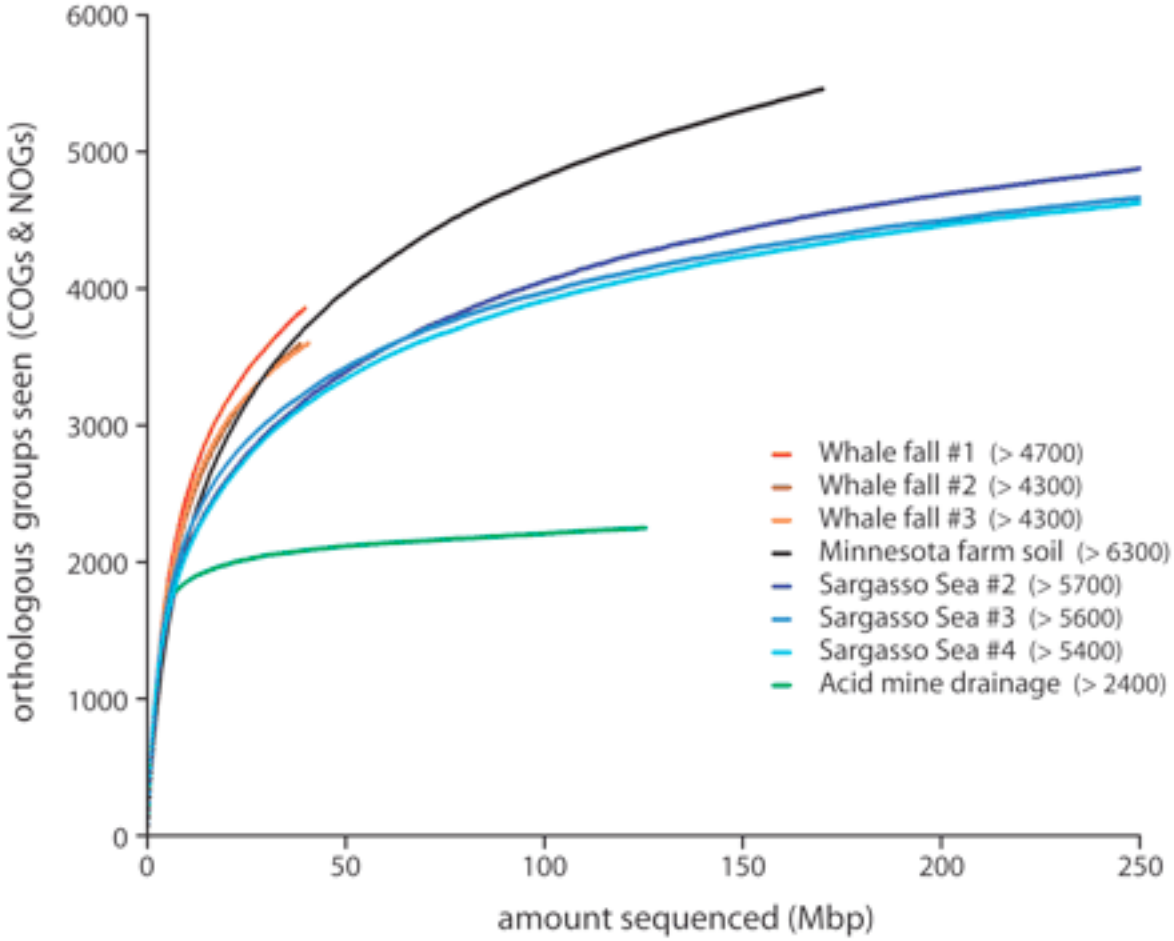
Sequencing and Assembly

- 100 million bp (Mbp) of sequence from the soil sample and 25 Mbp for each whale fall library
- Assembly was unsuccessful
- Projection: between two and five billion base pairs of sequence would be necessary to obtain the 8x coverage traditionally targeted for draft genome assemblies, even for the single most predominant genome in the soil community; between 100 and 700 Mbp of shotgun sequence data would be needed in order to generate a draft assembly for the most prevalent genome for the whale community
- Assembling genomes for low-abundance community members in any of these samples would clearly require significantly more sequence data
- So give up assembling

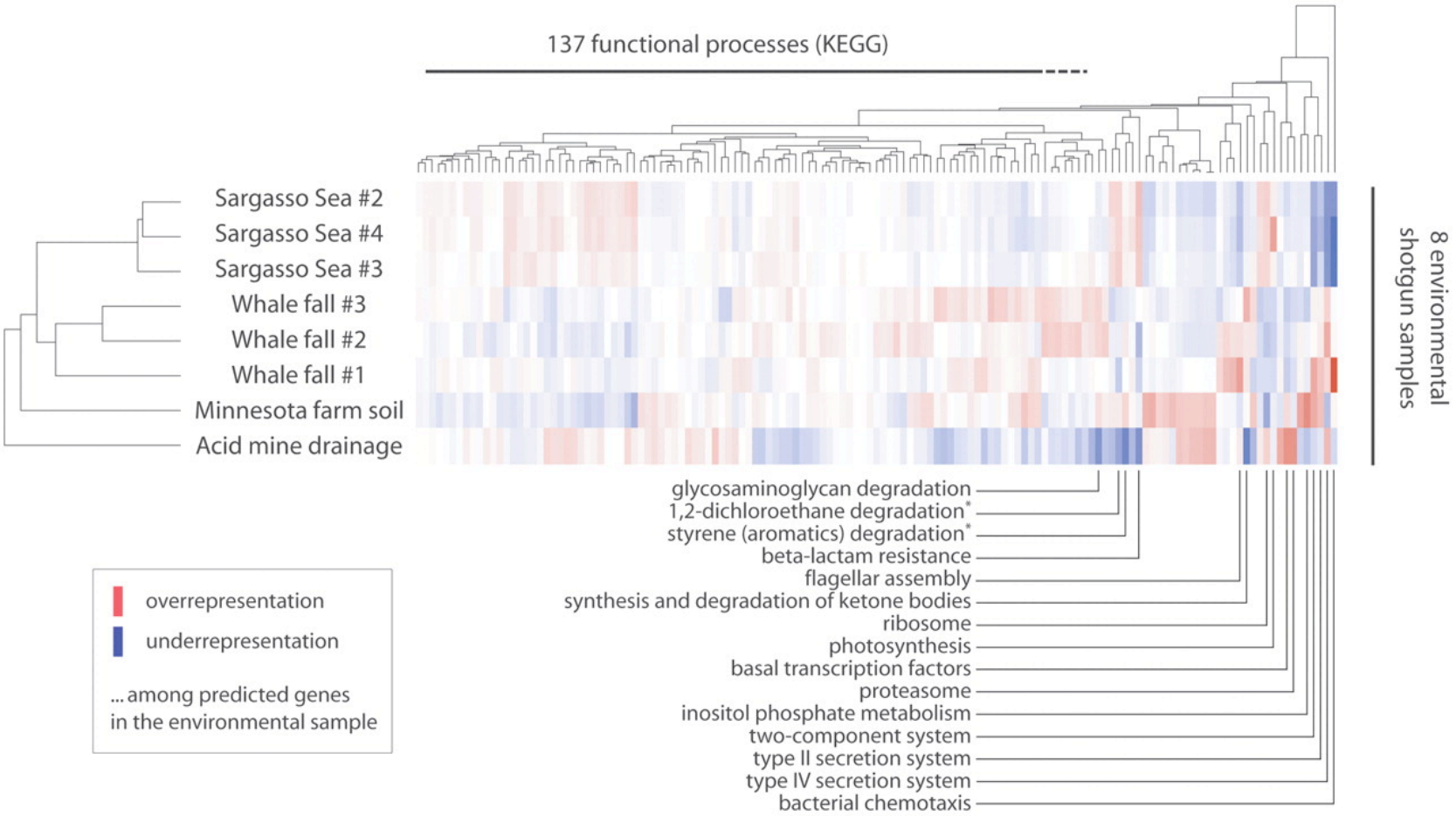
Environmental Gene Tag (EGT)

- To distinguish EGTs from the sequencing reads primarily used for the assembly of genomes
- Predict genes on unassembled sequences (EGTs)
- More than a third of the EGTs contained two or more predicted open reading frames

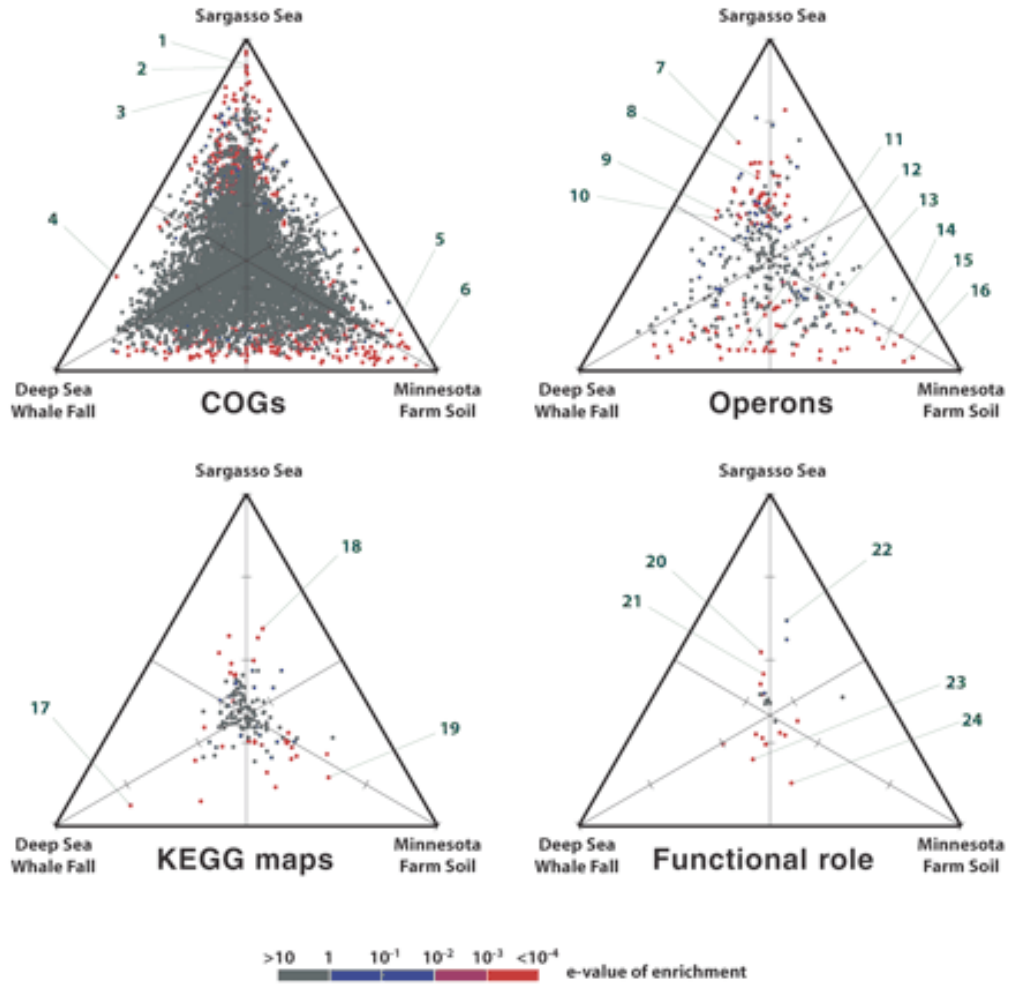
Searching for COGs in EGTs



Functional Profiling of Microbial Communities



Three-way Comparisons of Soil, Whale Fall, and Sargasso Sea Environments

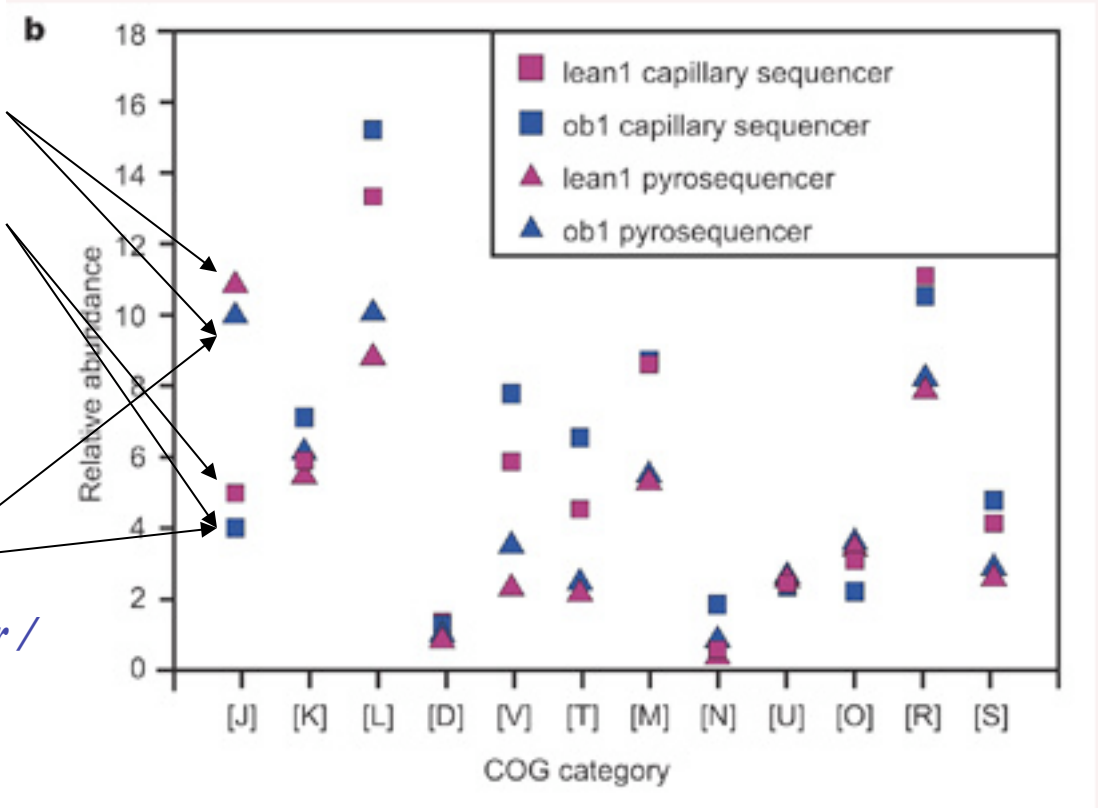


1, COG5524 bacteriorhodopsin;
5, COG3459 cellobiose phosphorylase;
7, ABC-type proline/glycine betaine transport system;
10, Na+-transporting NADH:ubiquinone reductase;
14, osmosensitive, active K+-transport system;
18, photosynthesis;
19, type I polyketide biosynthesis (antibiotics)

Obese/lean Mouse Gut Microbiome

Obese / lean

Capillary sequencer / pyrosequencer



"An obesity-associated gut microbiome with increased capacity for energy harvest" Nature 444, 1027-131(21 December 2006)

Shotgun Sequencing of Microbiomes

- Two sequencing methods: capillary sequencer & pyrosequencer
- Obese mouse vs lean mouse (ob/ob, +/+, ob/+)
- Read length
  - Capillary sequencer: 752 ±13.8 (s.e.m.)
  - GS20: 93.1 ±1.56 (s.e.m.)

Obese Microbiome

- Obesity is associated with changes in the relative abundance of the two dominant bacterial divisions, the Bacteroidetes and the Firmicutes
- Through metagenomic and biochemical analyses that these changes affect the metabolic potential of the mouse gut microbiota.
- The obese microbiome has an increased capacity to harvest energy from the diet.

More Metagenomics Projects

- The Human-microbiome project (*Science*, 2006)
- Enhanced biological phosphorus removal (EBPR) sludge communities (*Nat Biotechnol*, 2006)
- The marine viromes of four oceanic regions (*PLoS Biol* 2006)
- Termite hindgut (*Nature*, 2007)
- Nine biome (Nature, 2008)
- Twin human (Nature, 2009)
- And more
  - Habitats: soil, marine and lakes, human body
  - Showerhead (PNAS, 2009), windshield (Genome Res, 2009)

Computational and Statistical Tools

- *Assembly and Gene Prediction*
- *Tools for Characterizing Microbial Diversity Qualitatively and Quantitatively*
- *Function Prediction*
- *Comparative Metagenomics*
- *Statistical Tools for Metagenomics*
- Modeling Interactions Between Microbes and Their Environment (or Hosts)

How much can short reads tell us about a microbial community?